Next Generation Data Integration for the Life Sciences

Sarah Cohen-Boulakia
Université Paris Sud 11 - Orsay

Ulf Leser
Humboldt-Universität zu Berlin

Data Integration for the Life Sciences, 1993

- Robbins, R. J. (1994). "Report of the invitational DOE Workshop on Genome Informatics I: Community Databases." [Rob94a]
  - DOE funded large parts of the HGP starting end of the 80ties
- "*Continued HGP progress will depend in part upon the ability of genome databases to answer increasingly complex queries that span multiple community databases. Some examples of such queries are given in this appendix.*"
- "*Note, however, until a fully atomized sequence database is available (i.e., no data stored in ASCII text fields), none of the queries in this appendix can be answered. The current emphasis of GenBank seems to be providing human-readable annotation for sequence information. Restricting such information to human-readable form is totally inadequate for users who require a different point of view, namely one in which the sequence is an annotation for a computer-searchable set of feature information.*"

Twelve Queries Unanswerable (1993)

- 1. Return all sequences which map 'close' to marker M on chrom. 19, are put. members of the olfactory receptor family, and have been mapped on a contig
  - Multidatabase: Chromosome maps from GDB, sequence-contig in GenBank, annotation from elsewhere
- 3. Return the map location, where known, of all *alu* elements having homology greater than "h" with the *alu* sequence "S".
  - GenBank and a similarity search
- 4. Return all h. gene sequences for which a putative functional homologue has been identified in a non-vertebrate organism
  - Human: GenBank, non-vertebrates: species databases; how to describe function?
- 8. Return the number and a list of the distinct human genes that have been sequenced
  - What is a gene? Semantic heterogeneity and scientific uncertainty
- 11. Return all publications from the last two years about my favorite gene, accession number X####.
  - Synonyms & homonyms; naming conventions, disambiguation

The Classical Problems are all there already

- 1. Return all sequences which map 'close' members of the olfactory receptor family,
  - Multidatabase: Chromosome maps from GD annotation from elsewhere

  Distributed information
- 3. Return the map location, where known, greater than "h" with the *alu* sequence "S"
  - GenBank and a similarity search

  Non-standard processing
- 4. Return all h. gene sequences for which been identified in a non-vertebrate organis
  - Human: GenBank, non-vertebrates: species

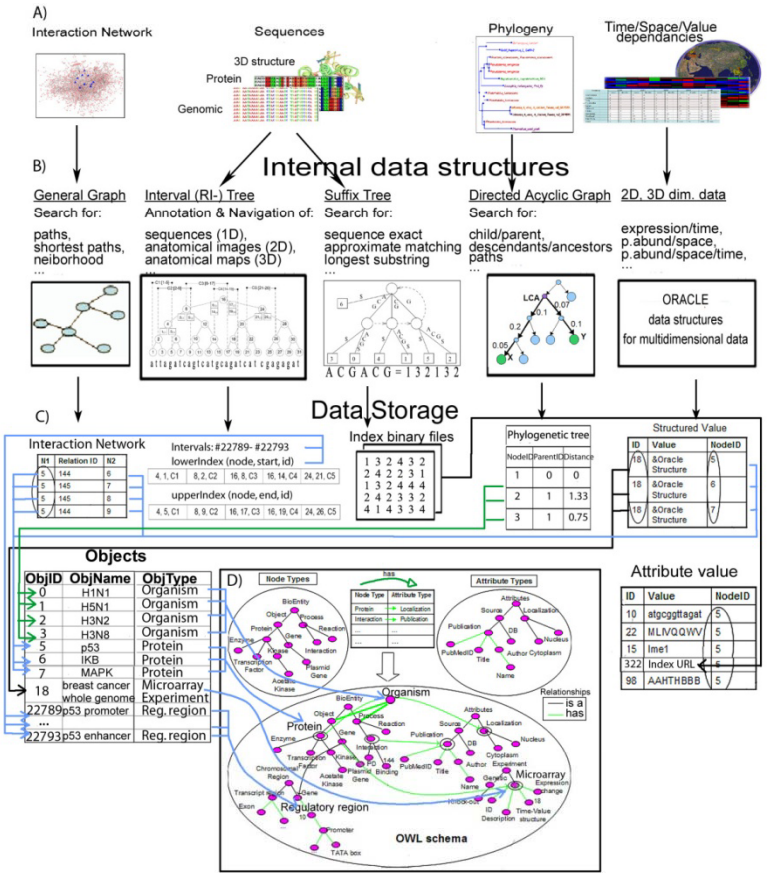
  Scientific uncertainty and evolving concepts
- 8. Return the number and a list of the dist sequenced
  - What is a gene? Semantic heterogeneity an

  Semantic heterogeneity
- 11. Return all publications from the last tw accession number X####.
  - Synonyms & homonyms; naming conventio

  Naming ambiguity

Data Integration, 2011

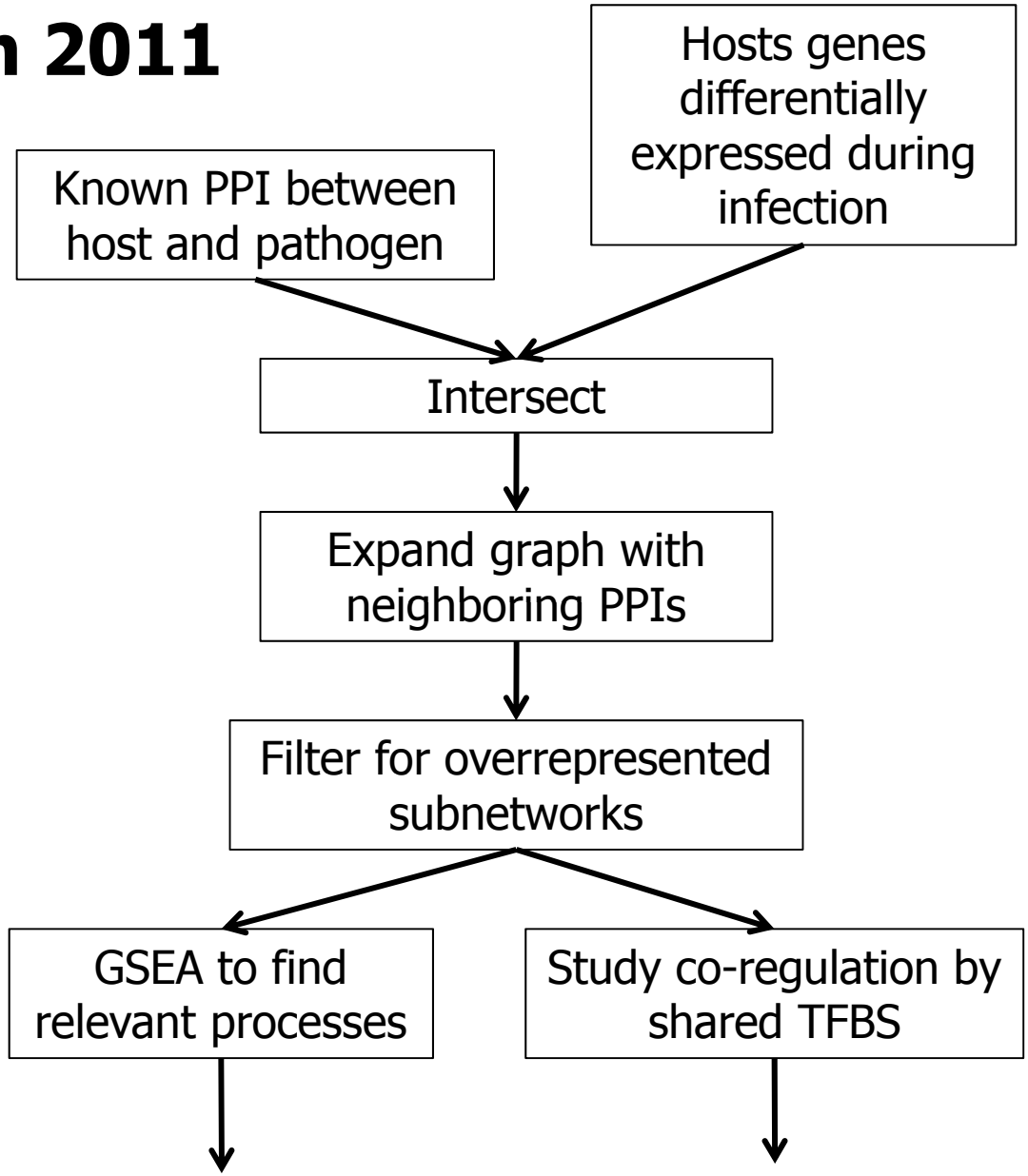
- BiologicalNetworks
- Integrated system using four primitive data types
  - Sequences
  - (Phylogenetic) trees
  - Histograms (arrays)
  - Graphs (networks)
- Pre-integrates a large set of public databases and ontologies
- Integration of further, specific data sets possible



- Adapted from [KSD+11]

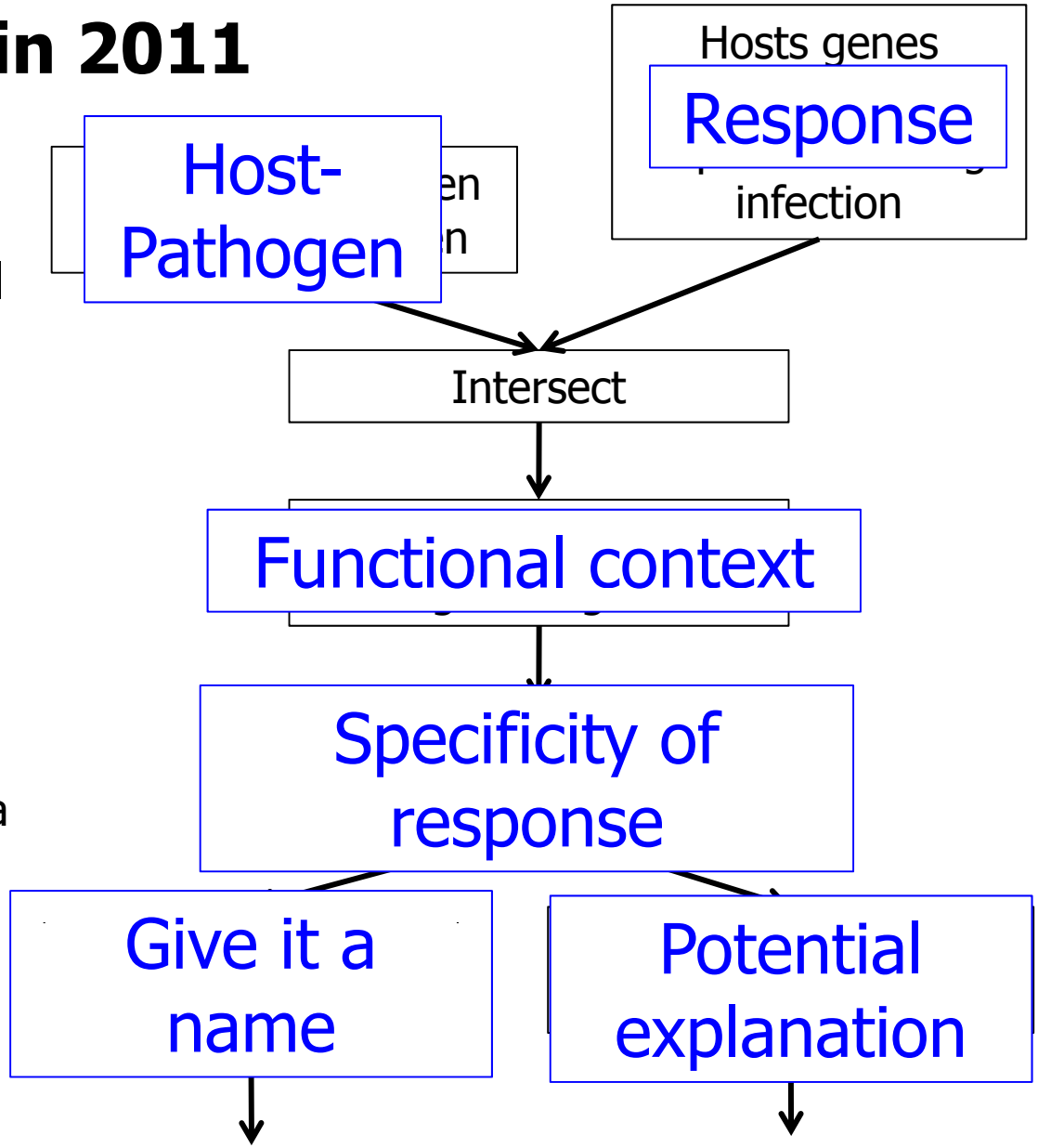
Data Integration in 2011

- Example task: Find genes that play a central role in the response of a host to a pathogen
  - Bacteria / viruses must attach to cells to have an influence
  - Attachment is a physical binding of proteins
  - This binding provokes a reaction in the cell, transmitted by more protein-protein interactions (e.g. signaling)

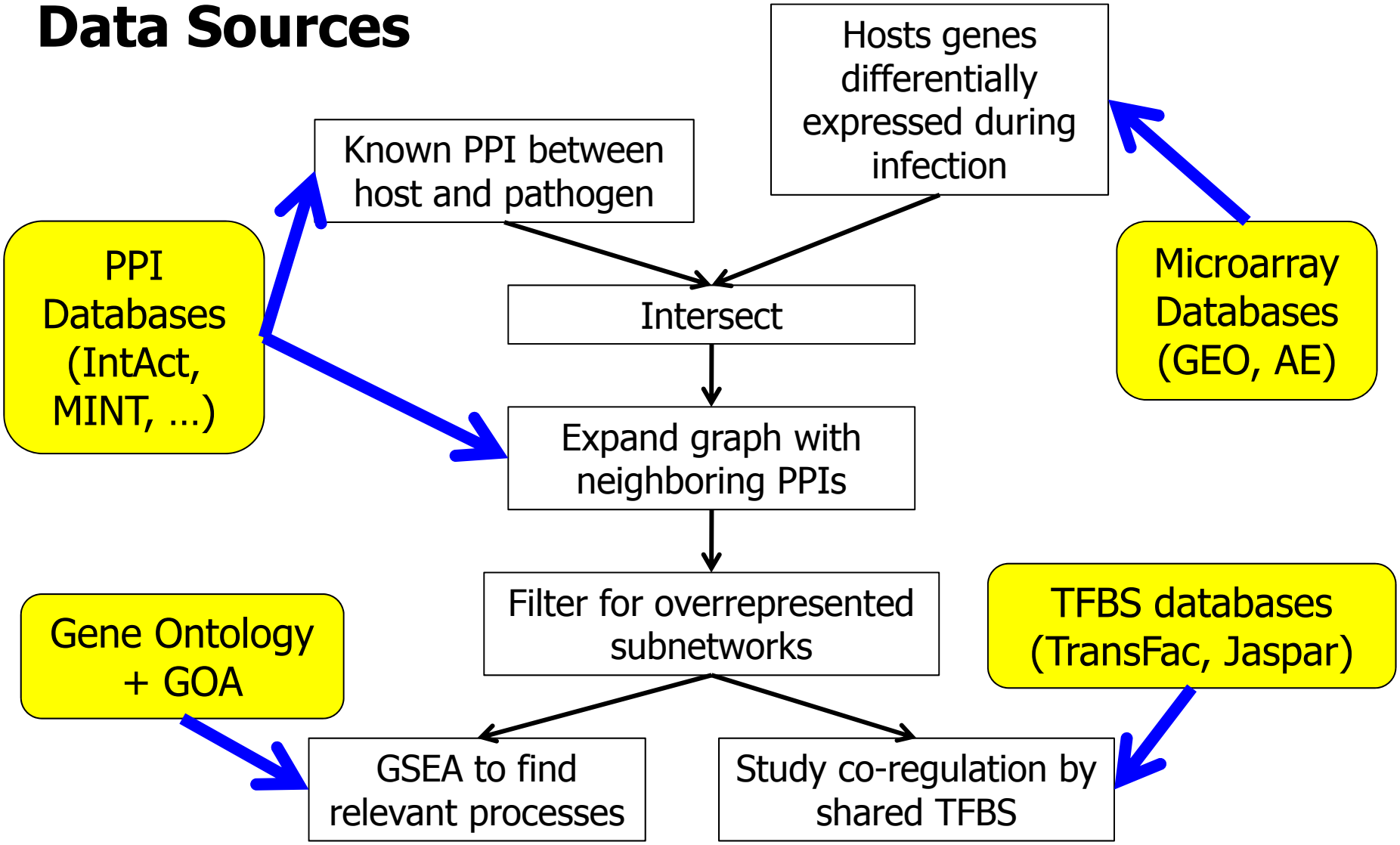


Data Integration in 2011

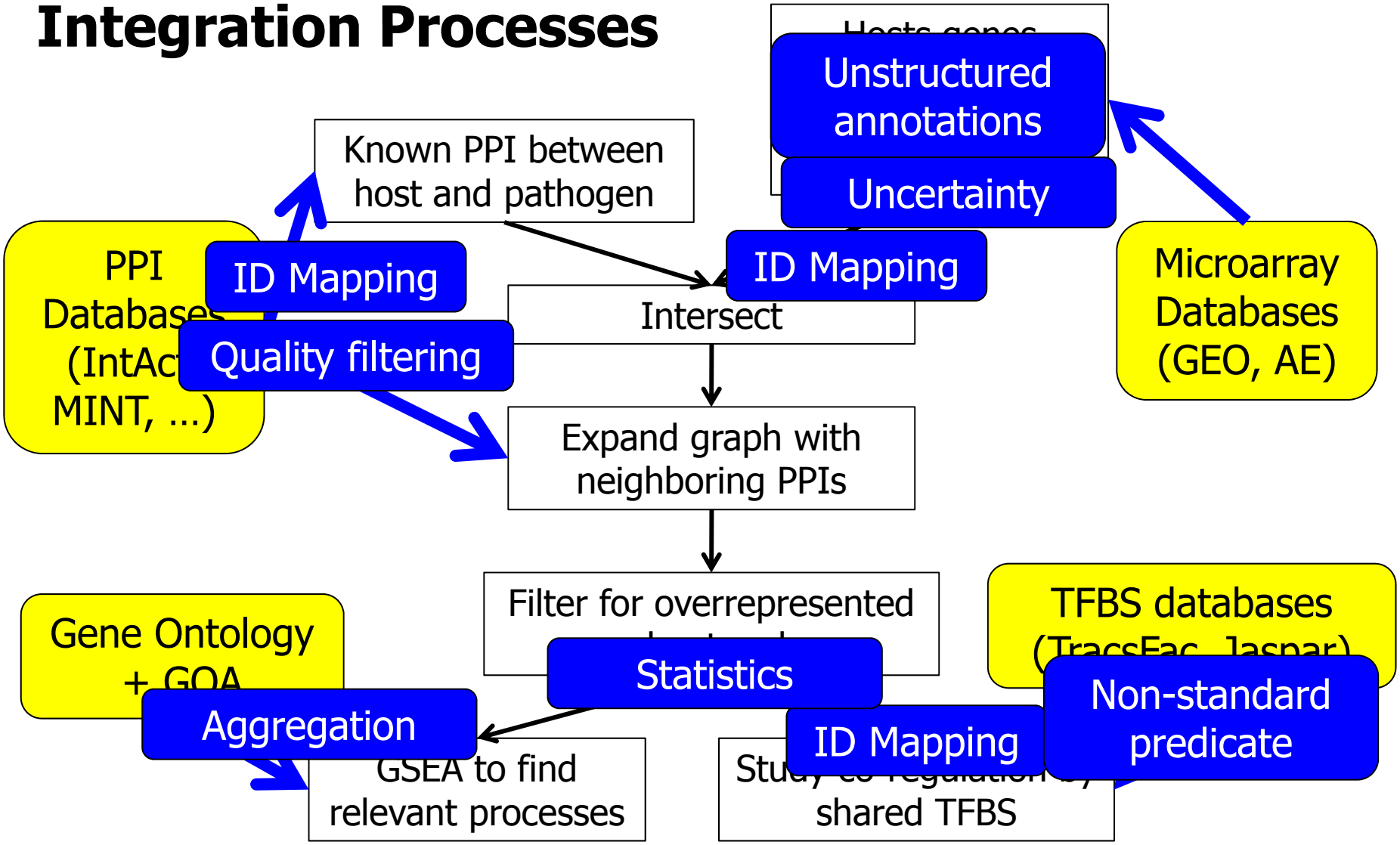
- Example task: Find genes that play a central role in the response of a host to a pathogen
  - Bacteria / viruses must attach to cells to have an influence
  - Attachment is a physical binding of proteins
  - This binding provokes a reaction in the cell, transmitted by more protein-protein interactions (e.g. signaling)



Data Integration?
Data Sources



Data Integration? Integration Processes



Take Home Message

- The number of sources to be used has increased a lot
- The diversity of the sources has increased a lot
- The complexity of the questions to be answered has increased a lot

Emergence of New Trends

- The number of sources to be used has increased a lot
  - ➢Scalability of integration in number of sources
  - ➢One major goal of the Semantic Web
- The diversity of the sources has increased a lot
  - ➢Inclusion of quality as a first-class citizen into models
  - ➢Ranking of integrated search results
- The complexity of the questions to be answered has increased a lot
  - ➢Integration requires analysis and analysis requires integration
  - ➢Scientific workflows

This Tutorial

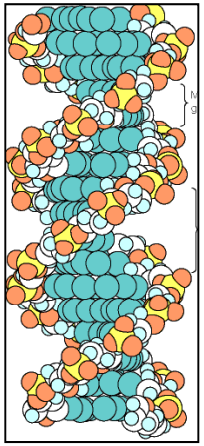
- Part I – Data Integration for the Life Sciences (45 min)
  - Biological Data & Biological Databases
  - Data Integration
  - Some Myths, some Truths
- Part II – Past and Presence (35 min)
- Part III – Current Trends (85 min)
- Part IV – Conclusions (5 min)

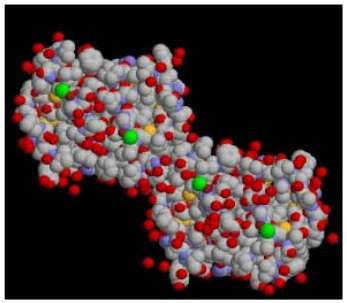
Scope: What are the Life Sciences?

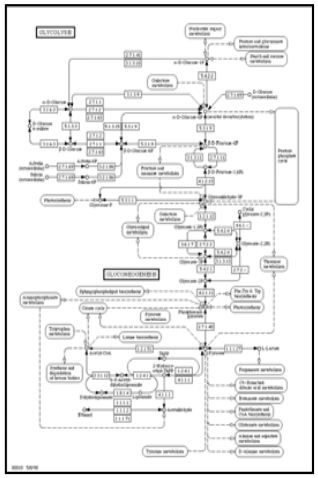
- Molecular Biology (Biophysics, Biochemistry)
- Systems Biology
- Molecular medicine
- Translational medicine

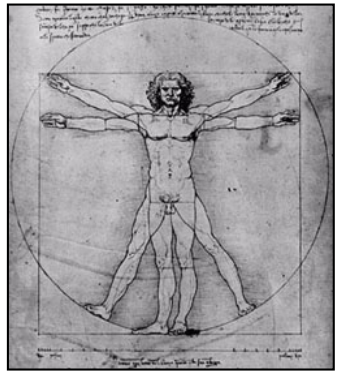
- A zillion species (human, animals, bacteria, virus, plants, …)

Computational Biology









Genomics

Sequencing
Gene prediction
Phylogeny
Regul. elements
RNA and miRNA
…

Proteomics

Structure prediction
Structure comp.
Motives, domains
Docking
PP- Interaction
…

Systems Biology

Pathway analysis
Pathway simulation
Gene regulation
Signaling
Metabolism
…

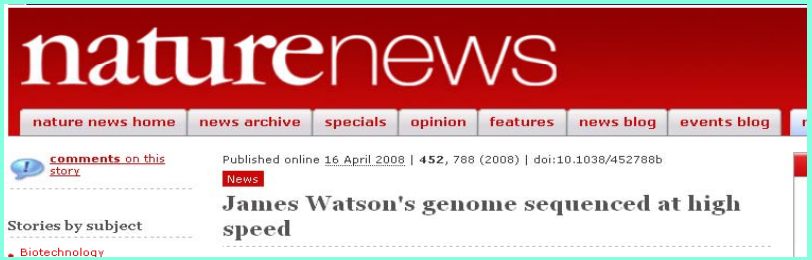
Medicine

Phenotype – genotype
Mutations and risk
Population genetics
Drug-drug interact.
…

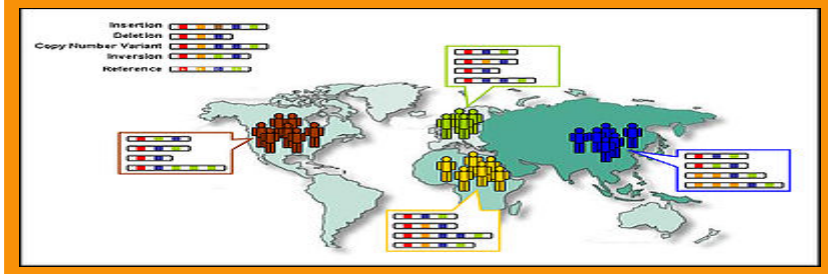
Enourmous Speed

1953
Double helix structure of DNA,
Watson/Crick

2003
First human genome sequenced
Took ~14 years, ~3 billion USD

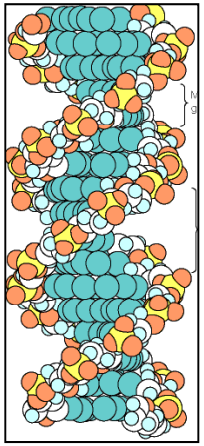


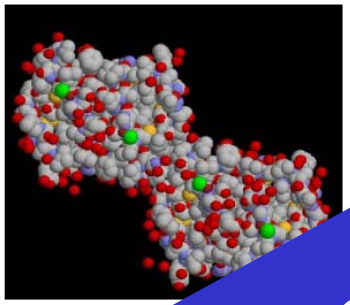
2008
Genome of J. Watson finished
4 Months, 1.5 Million USD

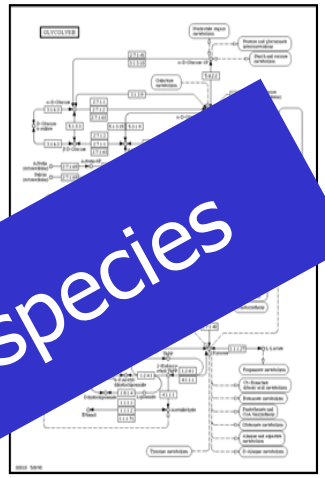


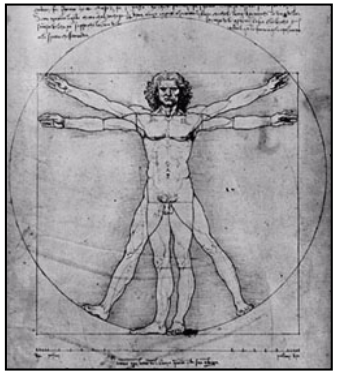
2010
1000 Genomes Project releases
first results

Database Perspective









In many, many different species

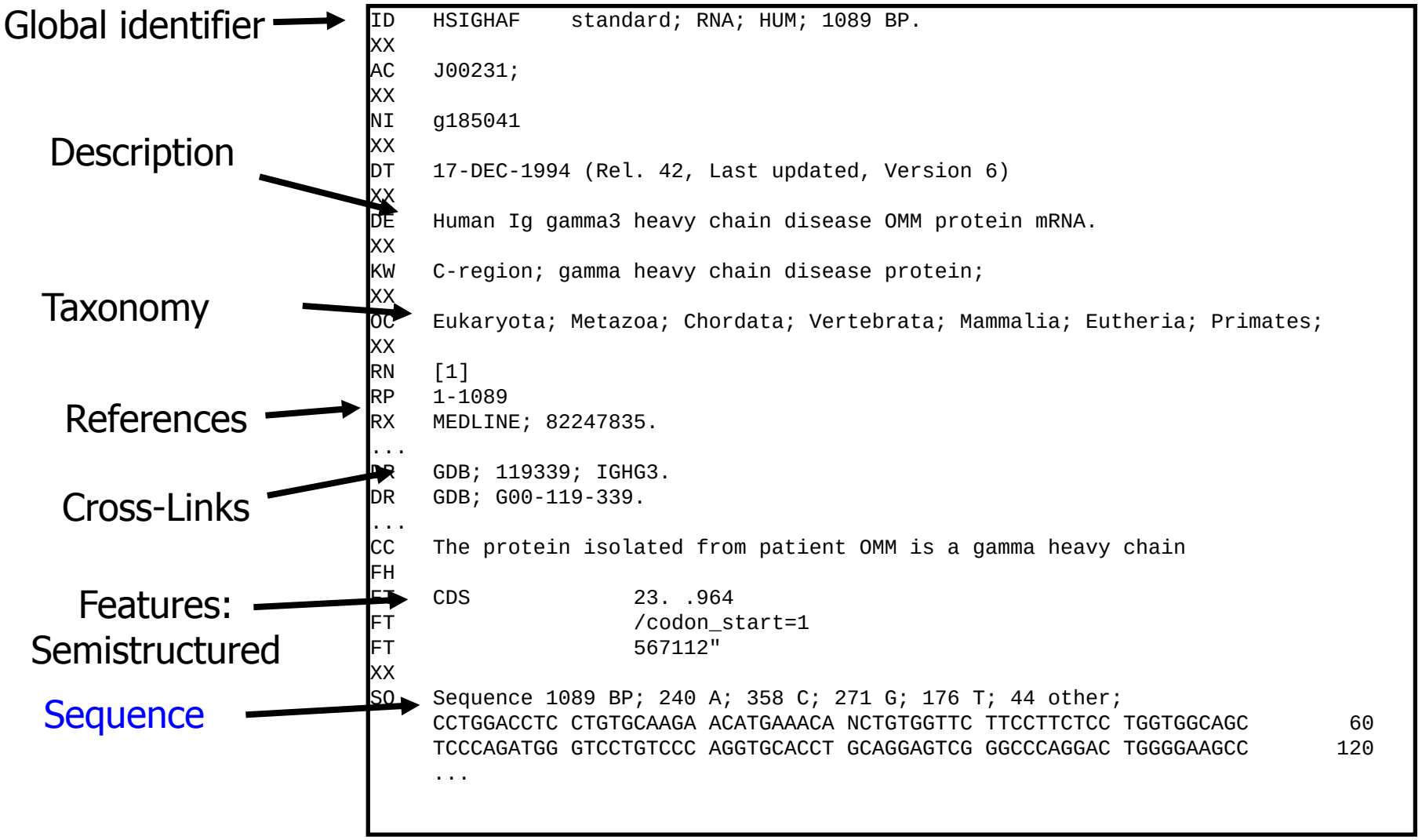
Genomics	...nics	Systems Biology	Medicine
Sequ...	...cture DBs	Pathway DBs	Patient DBs
...bs	Protein DBs	Regulation DBs	Biobanks
... DBs	Small molecule DBs	Signaling DBs	Drug DBs
...igenetic DBs	Motive DBs	Metabolic DBs	Study DBs
miRNA DBs	PPI DBs	Model DBs	Population DBs
mRNA DBS	...	Kinetic DBs	...
...		...	

Different Species are Important

- We are >96% genetically identical to chimps, orang-utan, mice, …
- We share 2000-3000 genes with E.coli
- We perform many experiments with mice / E.coli we cannot technically perform and do not want to perform with humans
- Genetics of bacteria / viruses is essential for fighting infectious disease
- Most things we eat has lived before

- Probably most of what we know about humans was learned from mice

A Biological Database (GenBank)



A Biological Database (GenBank)

Annotation

Real data

```
ID   HSIGHAF    standard; RNA; HUM; 1089 BP.
XX
AC   J00231;
XX
NI   g185041
XX
DT   17-DEC-1994 (Rel. 42, Last updated, Version 6)
XX
DE   Human Ig gamma3 heavy chain disease OMM protein mRNA.
XX
KW   C-region; gamma heavy chain disease protein;
XX
OC   Eukaryota; Metazoa; Chordata; Vertebrata; Mammalia; Eutheria; Primates;
XX
RN   [1]
RP   1-1089
RX   MEDLINE; 82247835.
...
DR   GDB; 119339; IGHG3.
DR   GDB; G00-119-339.
...
CC   The protein isolated from patient OMM is a gamma heavy chain
FH
FT   CDS             23. .964
FT                   /codon_start=1
FT                   567112"
XX
SQ   Sequence 1089 BP; 240 A; 358 C; 271 G; 176 T; 44 other;
     CCTGGACCTC CTGTGCAAGA ACATGAAACA NCTGTGGTTC TTCCTTCTCC TGGTGGCAGC        60
     TCCCAGATGG GTCCTGTCCC AGGTGCACCT GCAGGAGTCG GGCCCAGGAC TGGGGAAGCC       120
     ...
```

Properties

Micro-Syntax (non 1st normal form)

Line codes (pre-XML)

Free text fields

Layout influences syntax / semantics

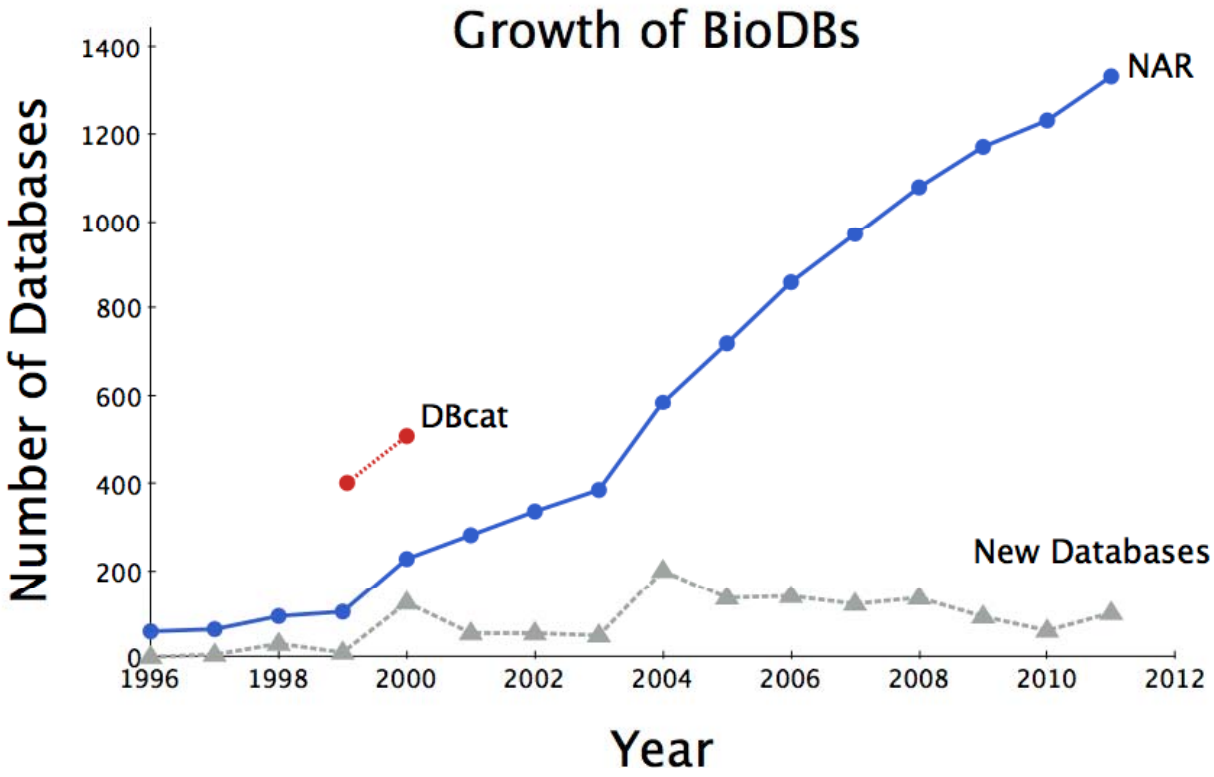
(Un)controlled vocabularies

```
ID   HSIGHAF    standard; RNA; HUM; 1089 BP.
XX
AC   J00231;
XX
NI   g185041
XX
DT   17-DEC-1994 (Rel. 42, Last updated, Version 6)
XX
DE   Human Ig gamma3 heavy chain disease OMM protein mRNA.
XX
KW   C-region; gamma heavy chain disease protein;
XX
OC   Eukaryota; Metazoa; Chordata; Vertebrata; Mammalia; Eutheria; Primates;
XX
RN   [1]
RP   1-1089
RX   MEDLINE; 82247835.
...
DR   GDB; 119339; IGHG3.
DR   GDB; G00-119-339.
...
CC   The protein isolated from patient OMM is a gamma heavy chain
FH
FT   CDS             23. .964
FT                   /codon_start=1
FT                   567112"
XX
SQ   Sequence 1089 BP; 240 A; 358 C; 271 G; 176 T; 44 other;
     CCTGGACCTC CTGTGCAAGA ACATGAAACA NCTGTGGTTC TTCCTTCTCC TGGTGGCAGC        60
     TCCCAGATGG GTCCTGTCCC AGGTGCACCT GCAGGAGTCG GGCCCAGGAC TGGGGAAGCC       120
     ...
```

Biological Databases Today

- This "flatfile horror" mostly has gone
  - Much XML for exchange (considerable standardization)
  - Flat files only for export / exchange
- Exotic techniques did exist – not any more
  - Almost all BDB today are maintained in relational systems
- "Read-only", no transactions
  - Very few BDB accept user submissions
- Web-based user interfaces
  - Very very few direct SQL accesses (but dump files for own use)
  - Simplicity rules: IR-style queries
- Most BDB are available entirely for download
  - New releases every X months
- Not big (changing rapidly)

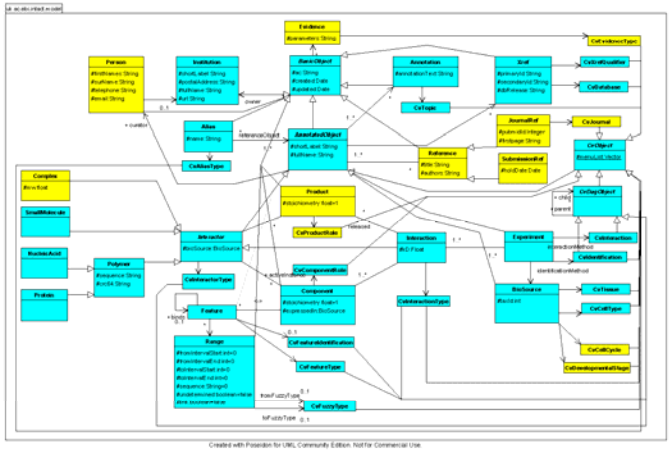
There are 100reds of Them



Number of existing (circles) and new databases (triangles) are plotted from 1996 to 2011. New databases are difference between the number of existing databases for each year. DBcat (red) is shown with NAR (blue) counts.

Data Model

- The classical DBs all started as books
  - One object – one page
    - Sequence, gene, protein, diseases, …
    - Object may be an abstract concept (gene) or a measurement (sequence)
  - Entry-based: A primary object and its (nested) annotations
  - Perfectly suited for XML and flatfiles
- More recent, none-archival (integrated) databases often are more "multi-object"
  - Multiple primary objects
  - Links between objects
  - More "normal" database

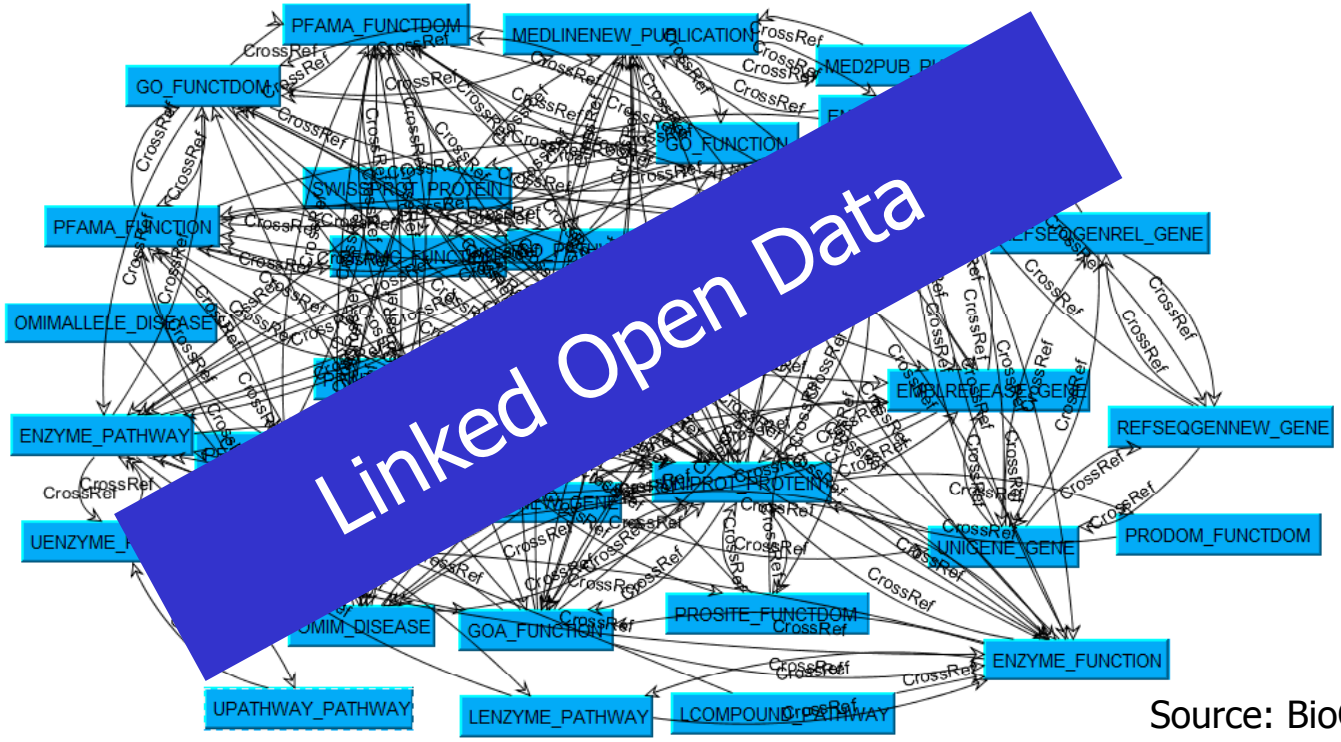


Classes of Biological Databases

- Primary – secondary – tertiary - …
  - Primary BDB for experimental data (sequences)
  - Secondary BDB for conclusions drawn from experiments (genes)
  - Relatively few primary (20?), many secondary (100reds)
- Species–specific – type-specific
  - All stuff on one species (MGD), all on one topic across species (GenBank)
- Curated or not
  - Most secondary databases are created and maintained manually
  - Many of them by reading and summarizing (curation)
  - Issues: Consistency, completeness, quality assurance, objectivity, …
- Some primary databases are international de-facto standard
  - Sequences: Genbank, proteins: UniProt, structures: PDB, …

Links

- BDB maintain links to many other BDBs
  - Instance level - external IDs, web browsing support
- No central authority for ID or links
- No consistency – "link hell"



Source: BioGuide

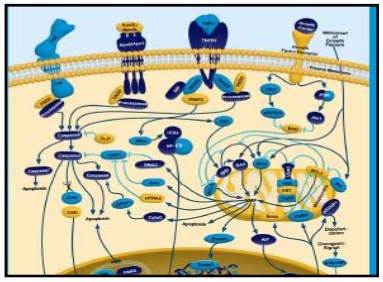
Different Cultures

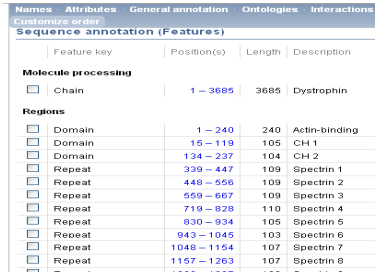
- BDB developers often are more similar to BDB users (from LS) than to database researchers (DR)
- DR publishes methods, LS publishes results
  - 1.300 databases = 1.300+ LS papers
  - 1.300 databases = ~10 DR papers
- DR: Often little willingness to become domain-specific
  - Building a BDB usually is not considered CS research (no papers, no PhD)
- DR: Often little willingness to consider CS as science
  - Too abstract, no concrete results on physical objects

- A VLDB paper on a BDB is by no means certainly a contribution to LS
- A NAR paper on a BDB is by no means certainly interesting for a DR

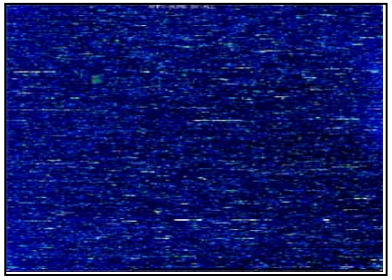
Types of "Data"

[Biocarta]





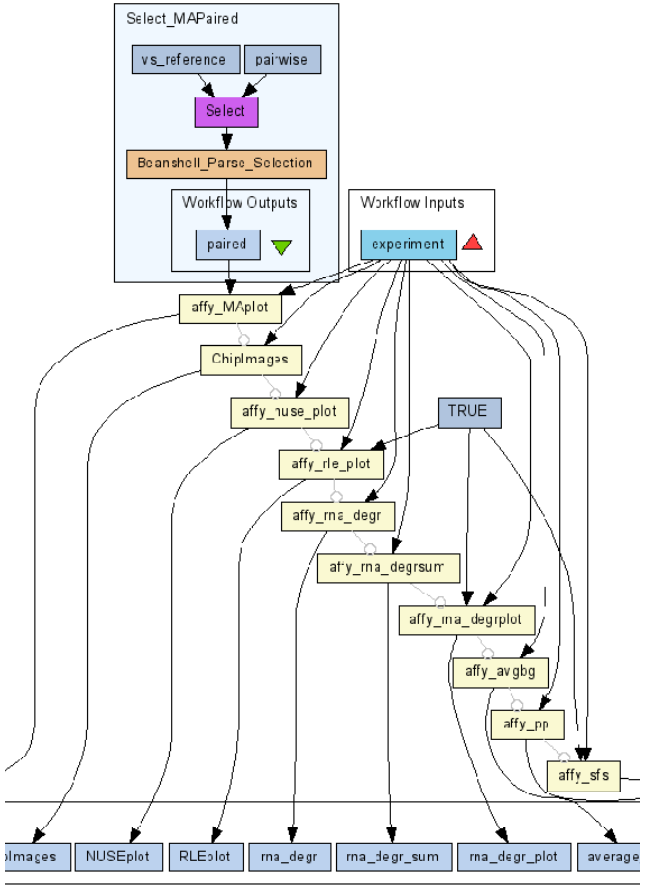
[Affymetrics]



- Knowledge
  - Confirmed, abstract, condensed
  - Text, graphics
  - Publications
- Information
  - Interpreted, filtered
  - Objects, annotations
  - BDB – secondary databases
- Data
  - Measured - raw, noisy, context-free
  - Numbers, sequences, metadata
  - BDB – primary databases

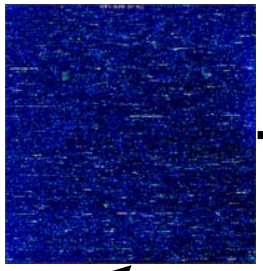
Data and Analysis Workflow

- High-throughput experiments require a multi-step analysis pipeline
- Many different suggestions for each step and for their composition into a process
- User only interested in result: Which genes are over-expressed in acute lymphoma?
- Data (information) may be integrated at various levels
  - Resulting in very different final results
- Rule-of-thumb: The later, the less comparable numerically
  - You may write a survey after mentally aggregating the results, but you cannot compute further with them



[myExperiment]

Data to Information



Scanning

Image

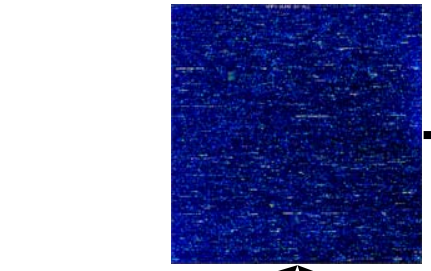
Spot detection & assignment

Raw data

Background correction spot-to-gene, …

[Affymetrics]

Data to Information



Scanning

Image

Spot detection & assignment

Raw data

Background correction spot-to-gene, …

Experiment-level normalization

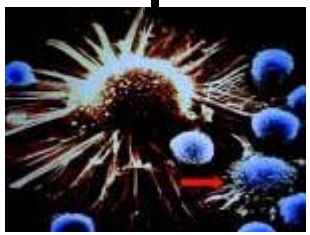
Differentially expressed genes

Signatures

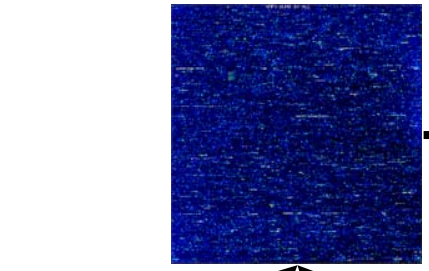
Clustering

Functional analysis

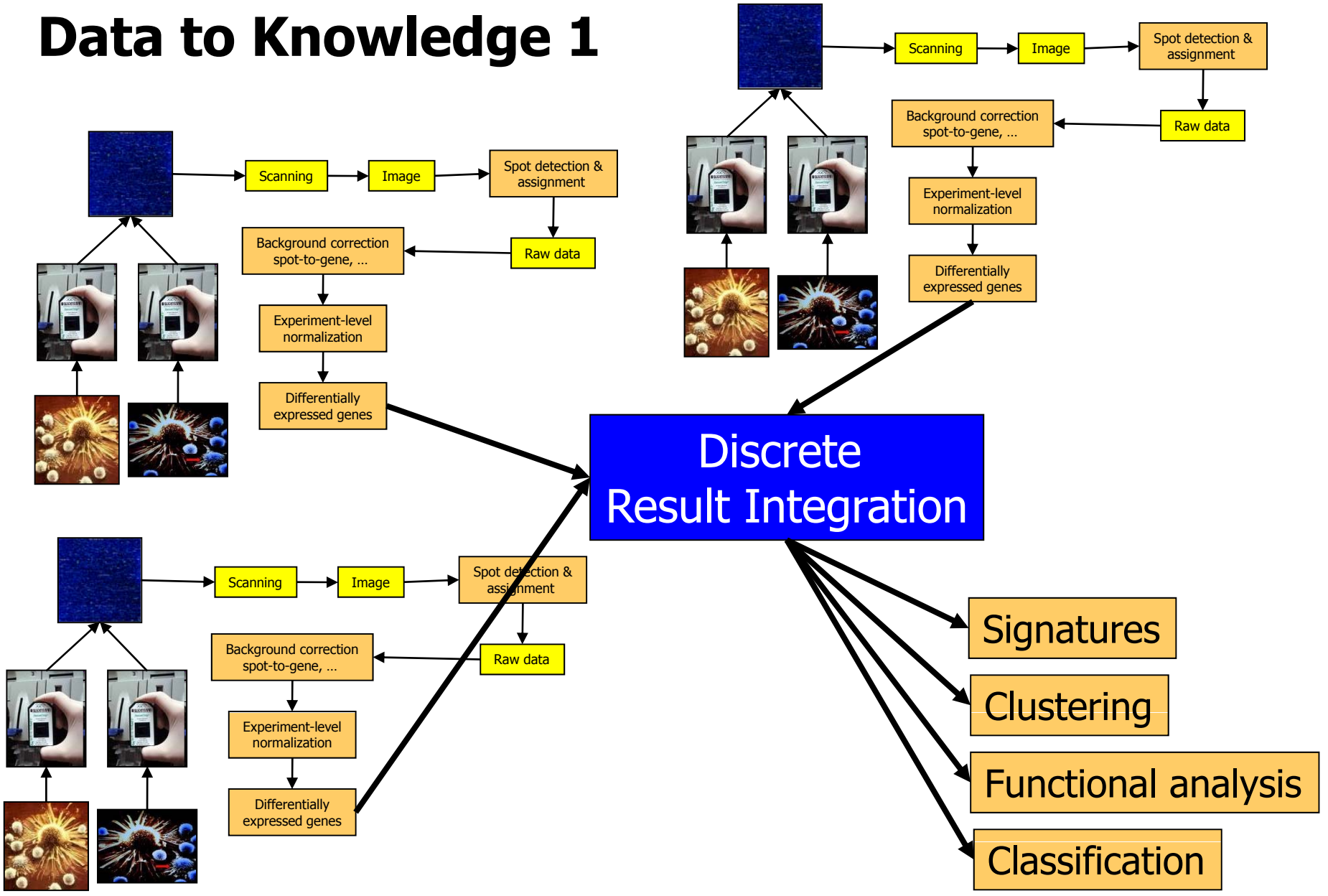
Classification



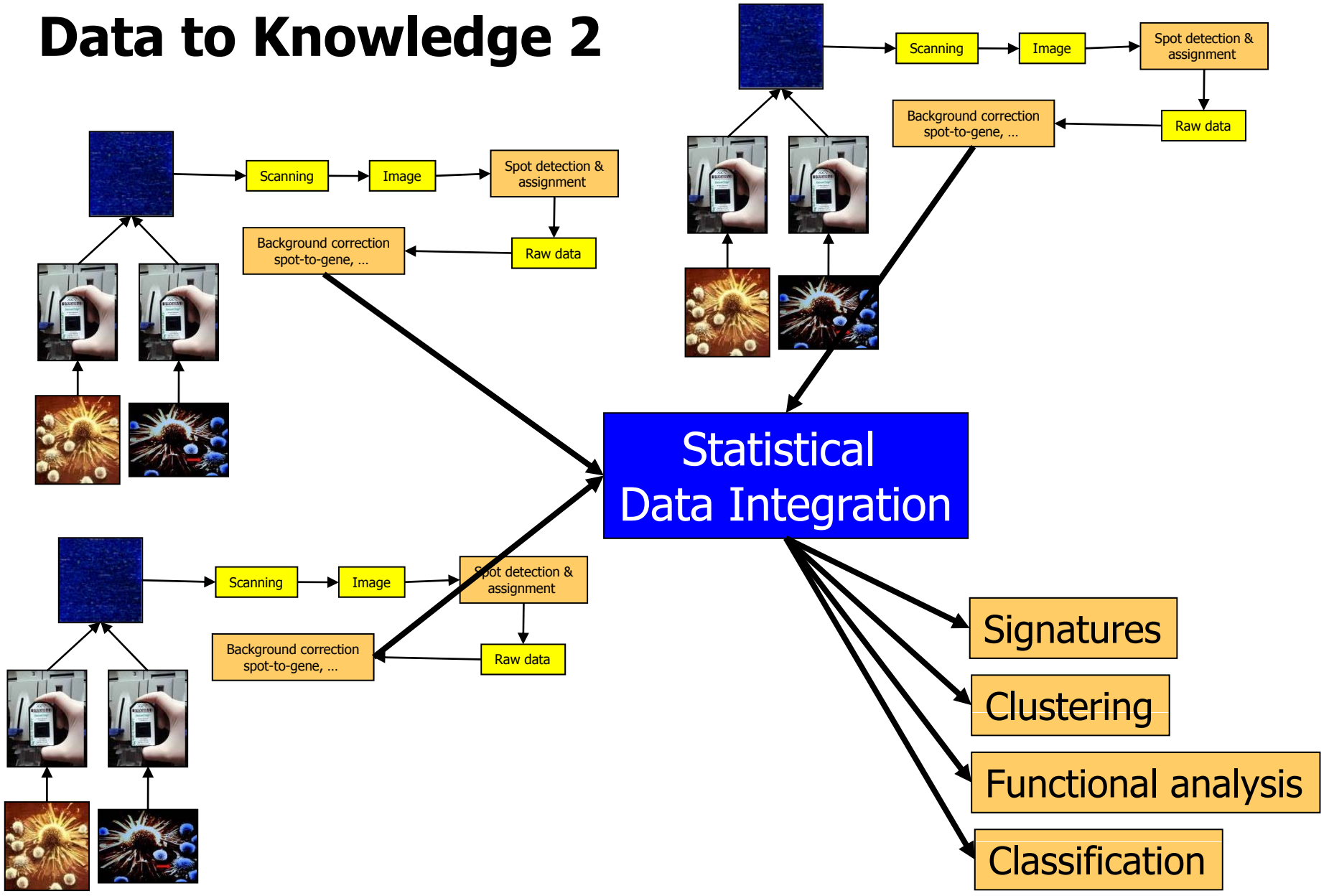
Steps with a wide Choice of Methods



Data to Knowledge 1



Data to Knowledge 2



This Tutorial

- Part I – Data Integration for the Life Sciences
  - Biological Data & Biological Databases
  - Data Integration
  - Some Truths, some Myths
- Part II – Past and Presence
- Part III – Current Trends
- Part IV – Conclusions

Because digital data are so easily shared and replicated and so recombinable, they present tremendous reuse opportunities, accelerating investigations already under way and taking advantage of past investments in science."

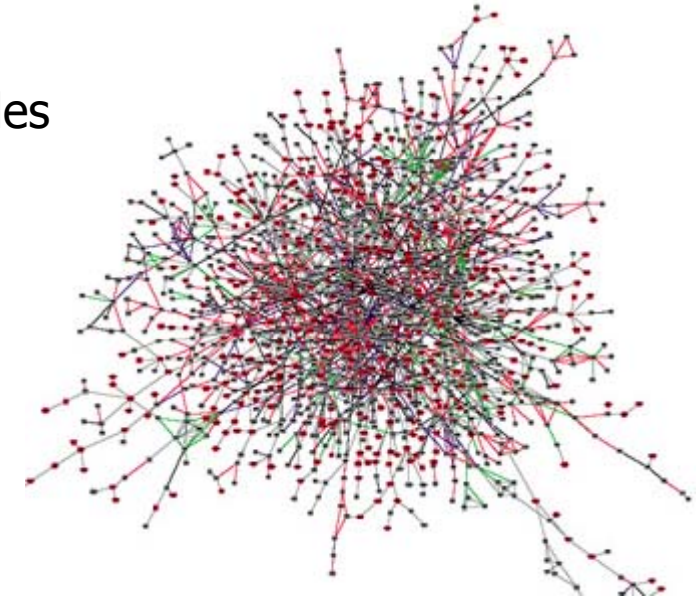
(Clifford Lynch, Nature 2008)

Why Integration?

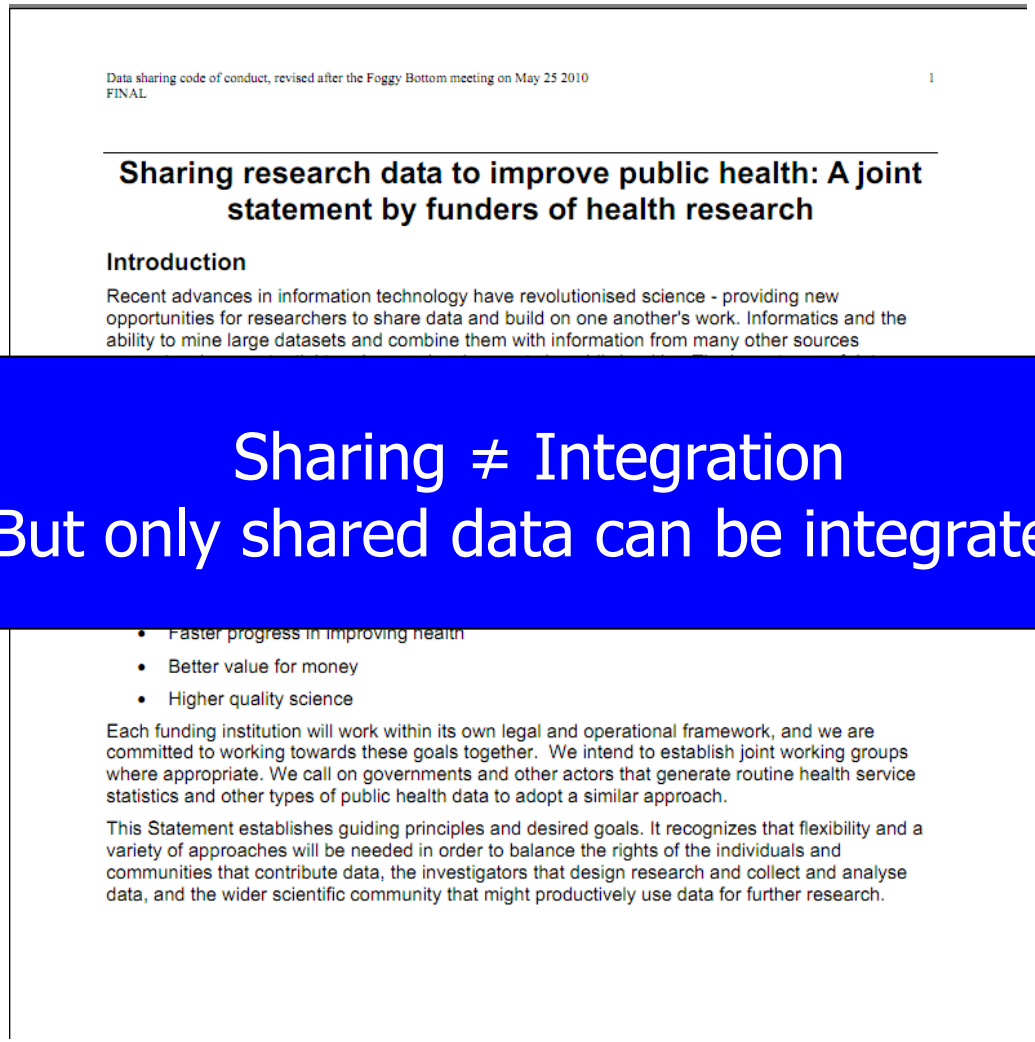
- Cost savings: Avoid duplication of experiments
- Quality control: Compare your result with that of others
- Complementation: Use additional data to strengthen your results
- Credibility: Let others redo your analysis
- Synergy: Combine data to produce stronger results
- Uniqueness: Some experiments are (almost) irreproducible
- Higher utility: Let others produce new results with your data

Success Stories

- Biological research is full of data-sharing success stories
  - International reference databases (Genbank, PDB, GEO, …)
  - Data published as Supplementary Material
  - Bioinformatics very much depends on sharing
- Numerous findings support benefits of integrated data
- For instance, integrated PPI datasets …
  - are more complete
  - yield better results in function prediction
  - yield better results in finding functional modules
  - allow more stringent quality filtering
  - help to identify false positives more easily
  - help to find disease genes more accurately
  - allow more accurate inference on evolutionary relationships
  - …



Political Will



Data sharing code of conduct, revised after the Foggy Bottom meeting on May 25 2010 1
FINAL

Sharing research data to improve public health: A joint statement by funders of health research

Introduction

Recent advances in information technology have revolutionised science - providing new opportunities for researchers to share data and build on one another's work. Informatics and the ability to mine large datasets and combine them with information from many other sources

Sharing ≠ Integration
But only shared data can be integrated

- Faster progress in improving health
- Better value for money
- Higher quality science

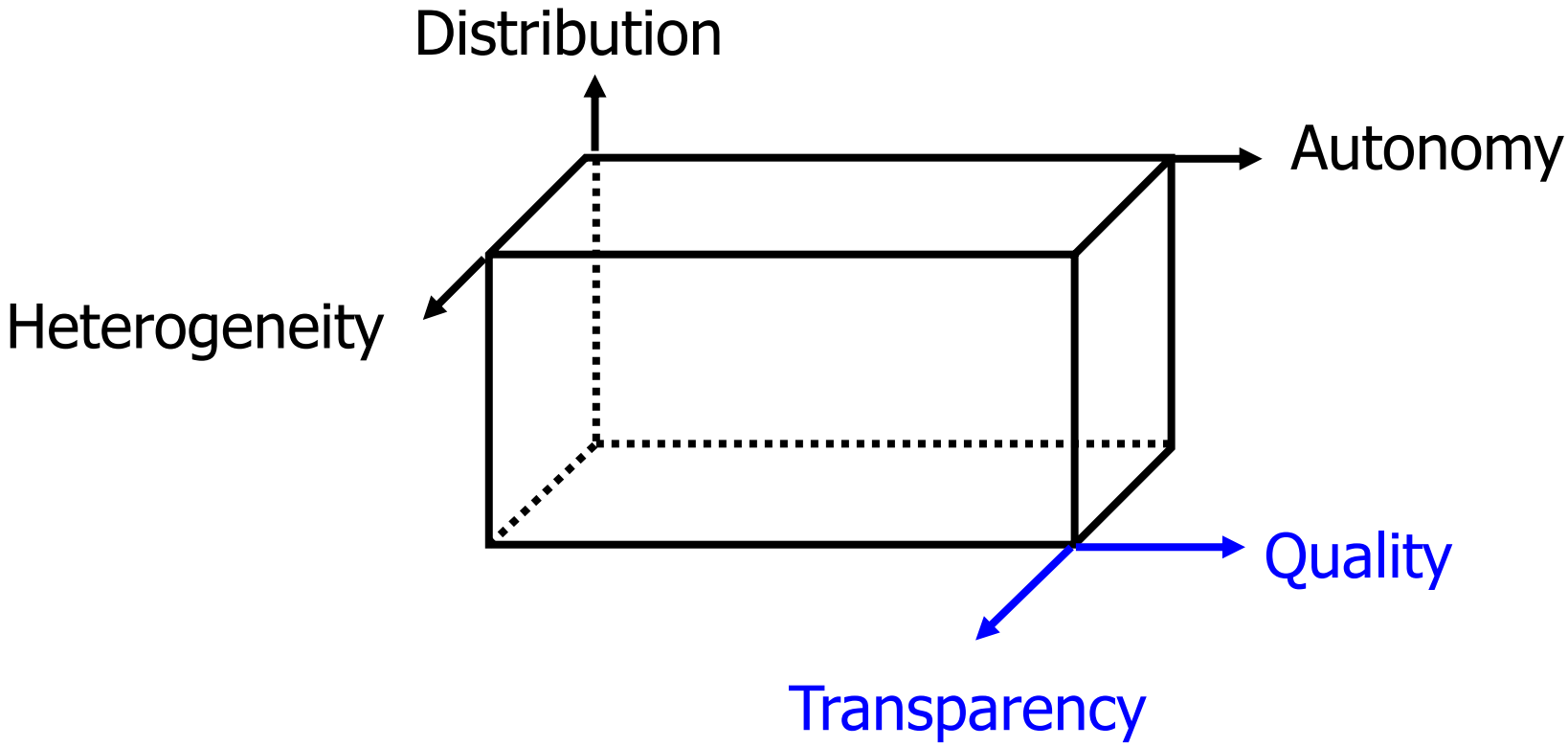
Each funding institution will work within its own legal and operational framework, and we are committed to working towards these goals together. We intend to establish joint working groups where appropriate. We call on governments and other actors that generate routine health service statistics and other types of public health data to adopt a similar approach.

This Statement establishes guiding principles and desired goals. It recognizes that flexibility and a variety of approaches will be needed in order to balance the rights of the individuals and communities that contribute data, the investigators that design research and collect and analyse data, and the wider scientific community that might productively use data for further research.

The Problem

- Research in molecular biology
  - is performed world-wide in thousands of labs – mostly in competition
  - has a multi-scale, highly complex target: Life in all its variants
  - is driven by dozens of experimental techniques to reveal different properties of genes, cells, organisms, diseases, …
  - produces results that are highly context-dependent – integration always has to face inconsistencies, noise, large error margins, …
  - works with concepts that are in constant evolution – class names change their meaning with time
  - more and more requires consideration of many different experimental techniques, scientific approaches, and interdisciplinary teams

Classical Dimensions of Distributed DBs [OV99] (and two not-so-classical ones)



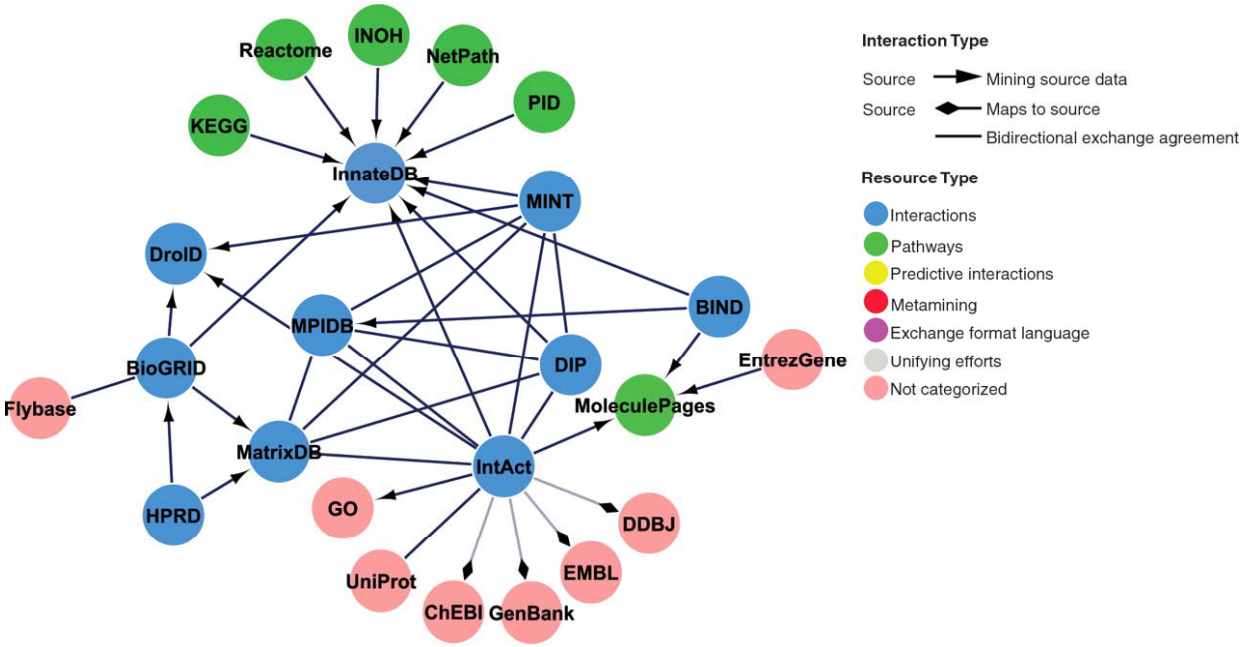
Are BDB Distributed?

- > 1000 different databases
  - Plus many data sets that are not stored in a DB
  - E.g. Supplementary material
- Content is highly redundant
  - Replica (sequence / microarray databases)
  - Large unintentional overlaps (UniProt – PIR, KEGG – Reactome)
  - Large intentional overlaps (selection of species-specific data)
  - Some databases mostly copy from other sources (Ensembl)
- Content may be changed (curated) during copying
  - Inconsistencies

Example: Protein-Protein-Interactions

- There are >300 BDBs related to PPI and pathways
  - See http://www.pathguide.org

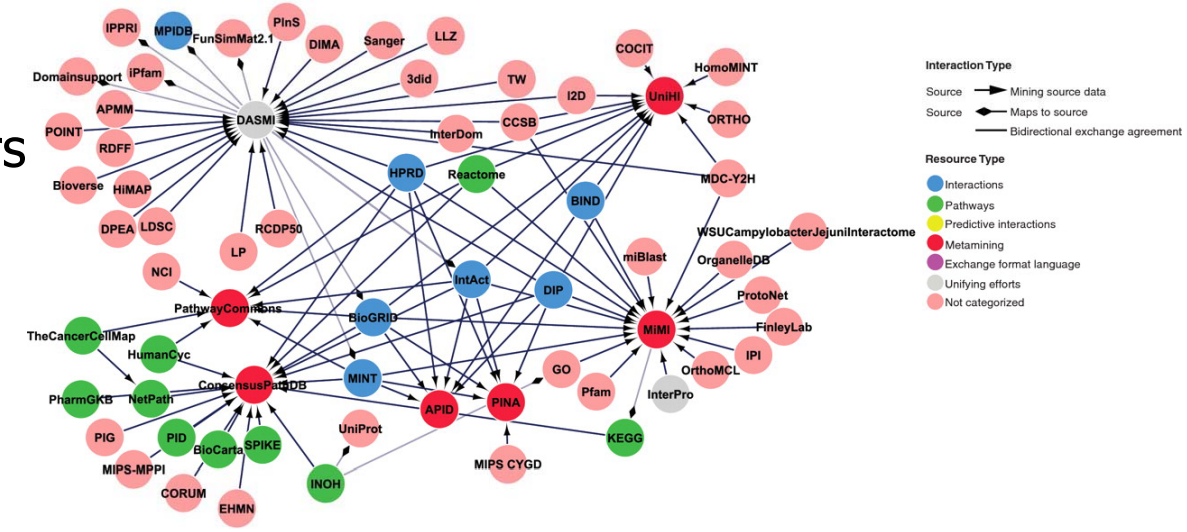
- Manually created "source" DBs



Example: Protein-Protein-Interactions

- There are >300 BDBs related to PPI and pathways
  - See http://www.pathguide.org

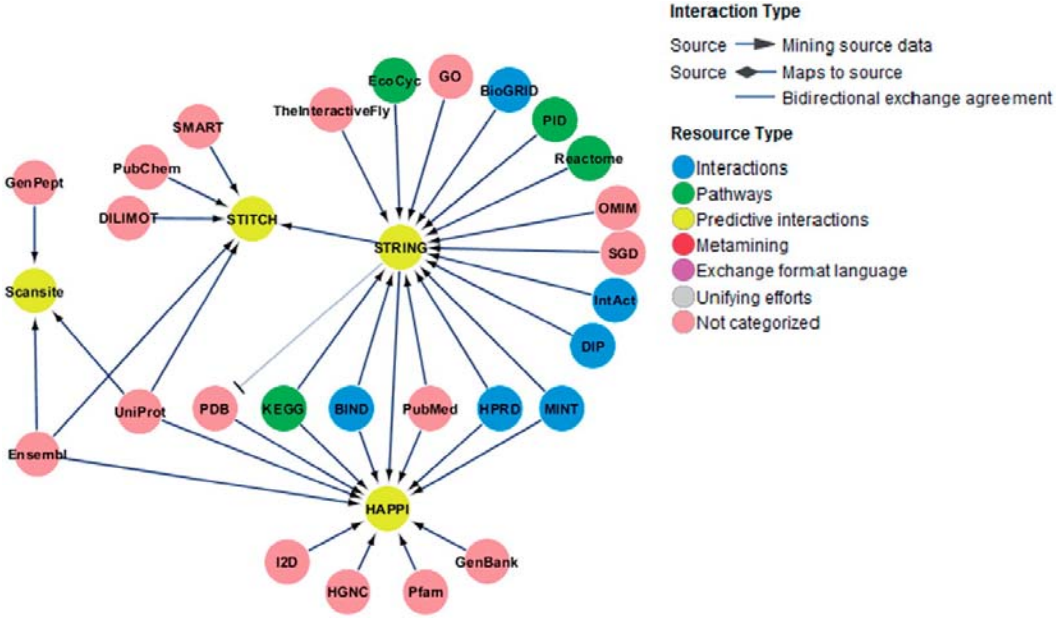
- Manually created "source" DBs
- DBs integrating others and HT data sets



Extreme Example: Protein-Protein-Interactions

- There are >300 BDBs related to PPI and pathways
  - See http://www.pathguide.org

- Manually created "source" DBs
- DBs integrating others and HT data sets
- Predicted interactions

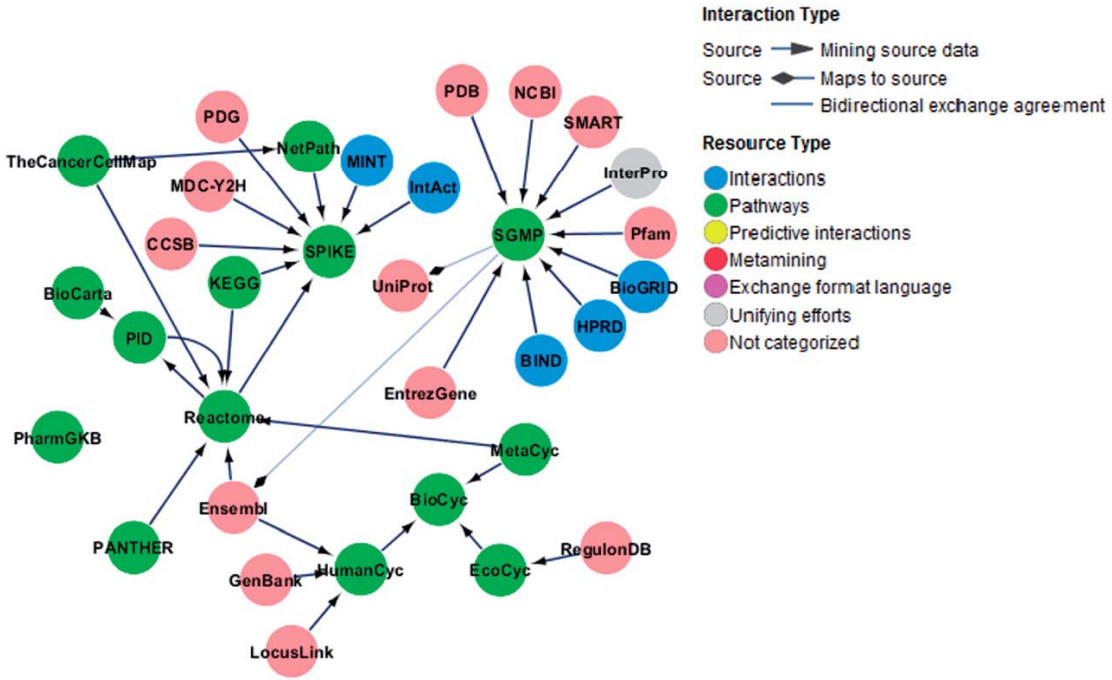


Extreme Example: Protein-Protein-Interactions

- There are >300 BDBs related to PPI and pathways
  - See http://www.pathguide.org

- Manually created "source" DBs
- DBs integrating others and HT data sets
- Predicted interactions
- Pathway DBs (consisting of PPI)

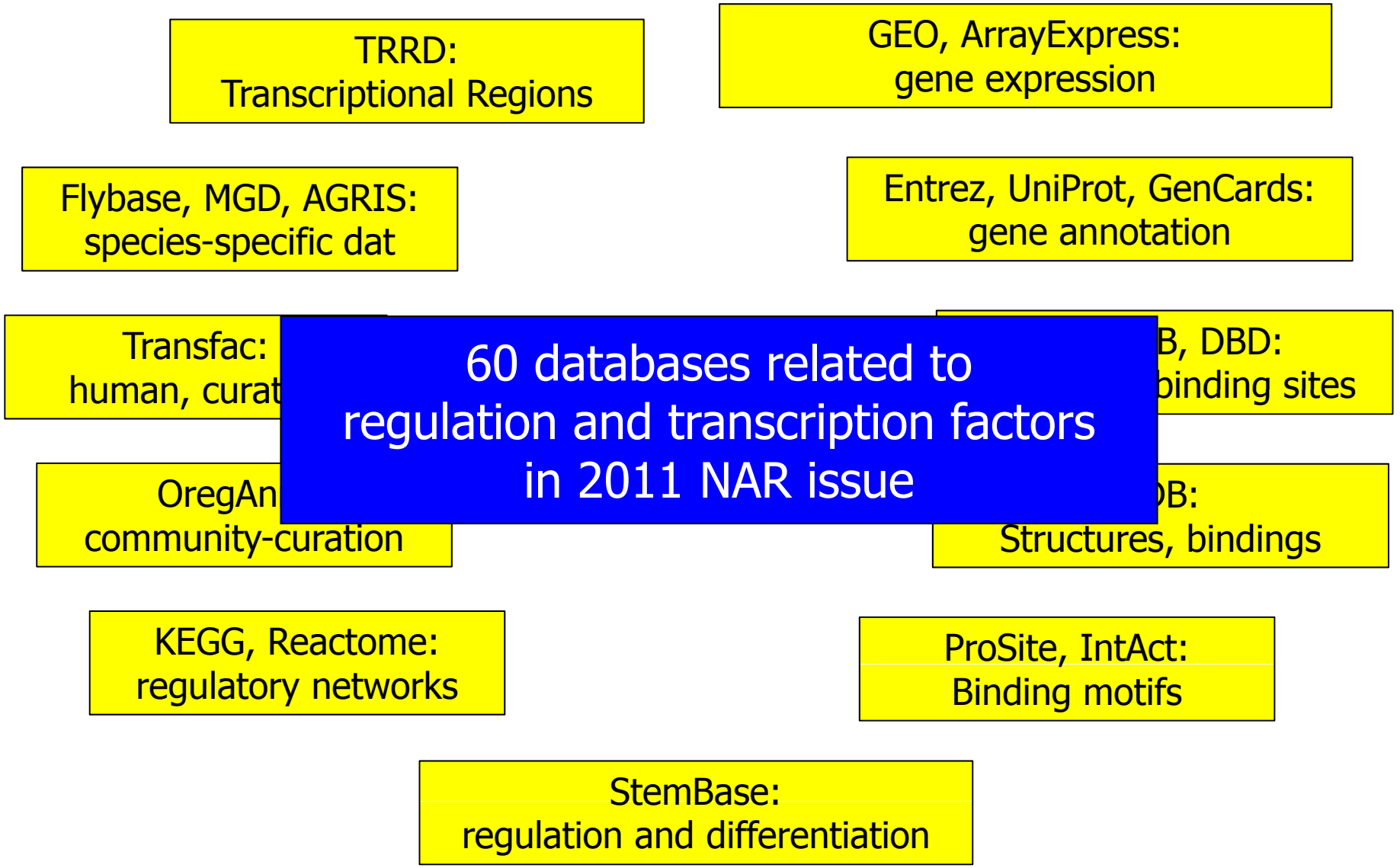
- [KP10]



A Mess [KP10]

- Inconsistent understanding of what a PPI actually is
  - Binary, physical interaction
  - Complexes
  - Transient, functional association
- Some integrated DBs have imported more data than there is in the sources
- Source databases overlap to varying degrees
  - Effort to sort things out in IMex consortium
- Largely different reliability of content
  - Literature, high-throughput experiments, transferred from orthologs, …
- Literature-curated DBs do not exhibit higher quality than HT [CYS08]
  - Re-annotation reveals inconsistencies, subjective judgments, errors in gene name assignment, …

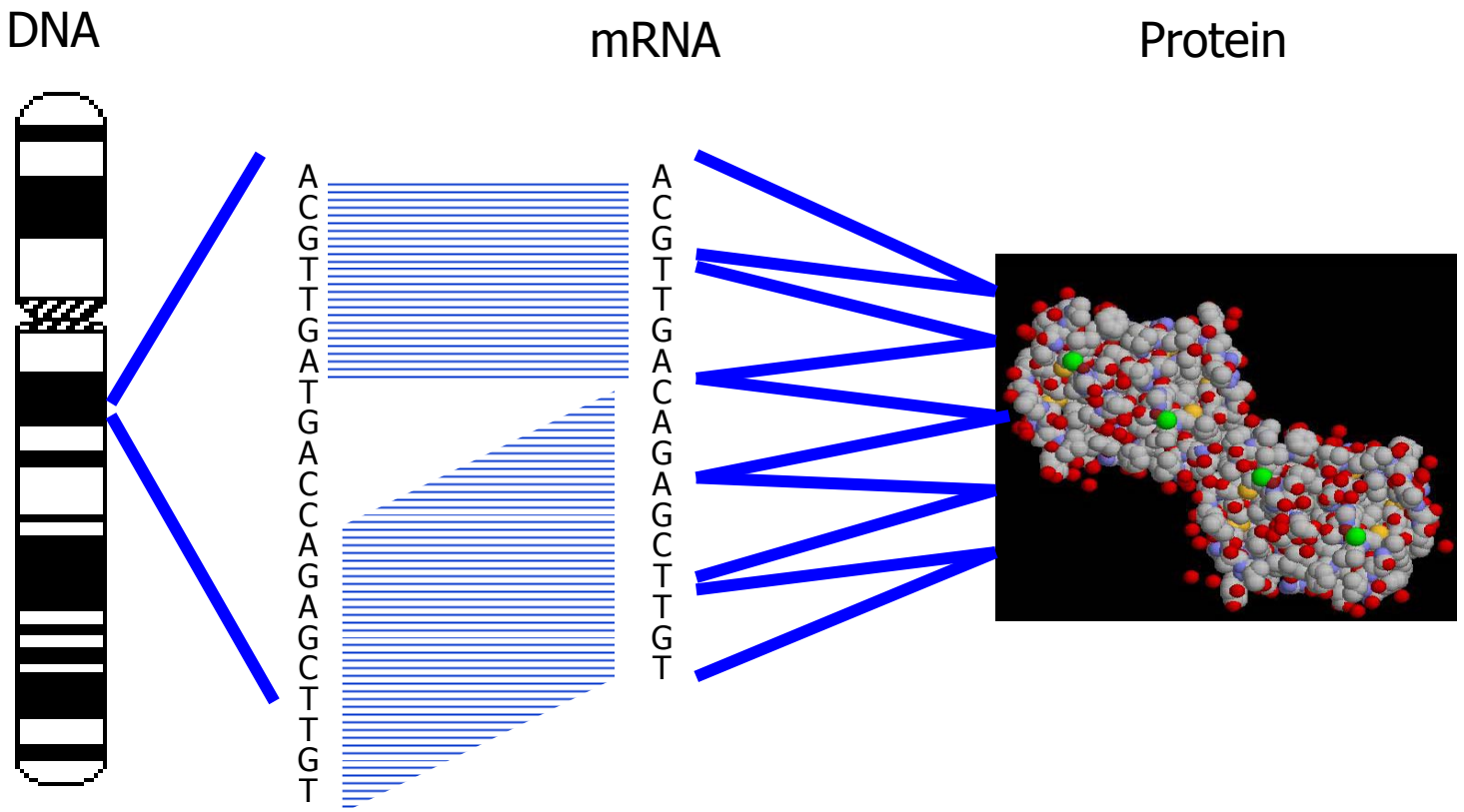
Vertical Distribution



Are BDB Heterogeneous?

- Technical heterogeneity: not that much
  - RDBMS, web services, HTML forms, …
- Syntactic heterogeneity: not too much of a problem any more
  - XML exchange, flatfiles
  - Many ready-to-use parsers are available
- Semantic heterogeneity: terrible
  - Objects have several names and IDs (and versions, states, orthologs, …)
  - Meaning of schema elements are heterogeneous, scientifically uncertain, and change over time
  - Metadata often is not available in sufficient detail
- As usual – distribution creates (semantic) heterogeneity

What is a Gene (1)?



- A stretch of DNA (with holes) on a chromosome that at some stage gets translated into a protein

What is a Gene (2)?

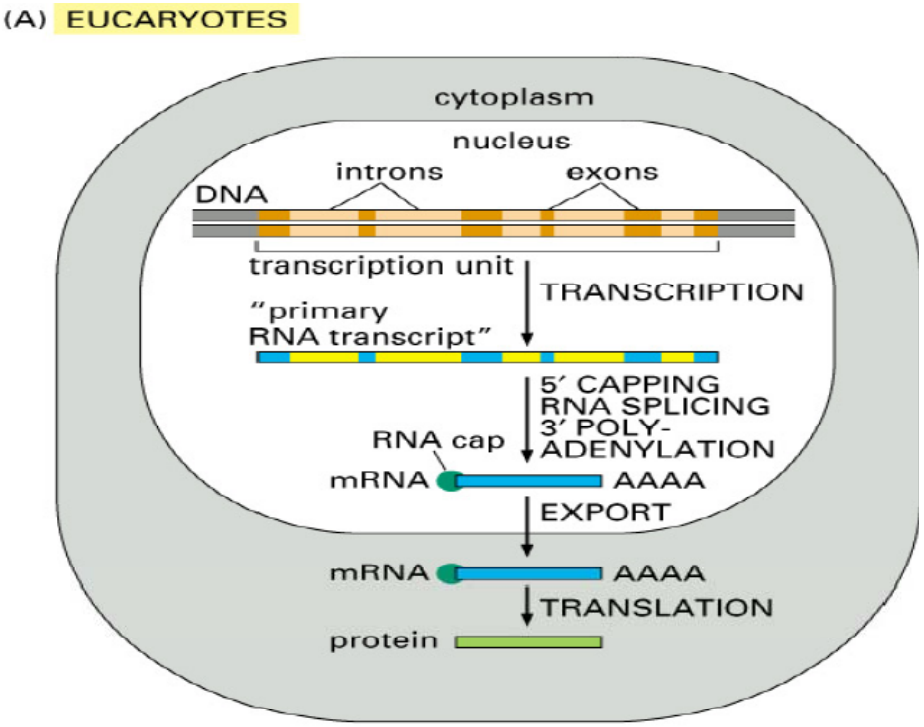
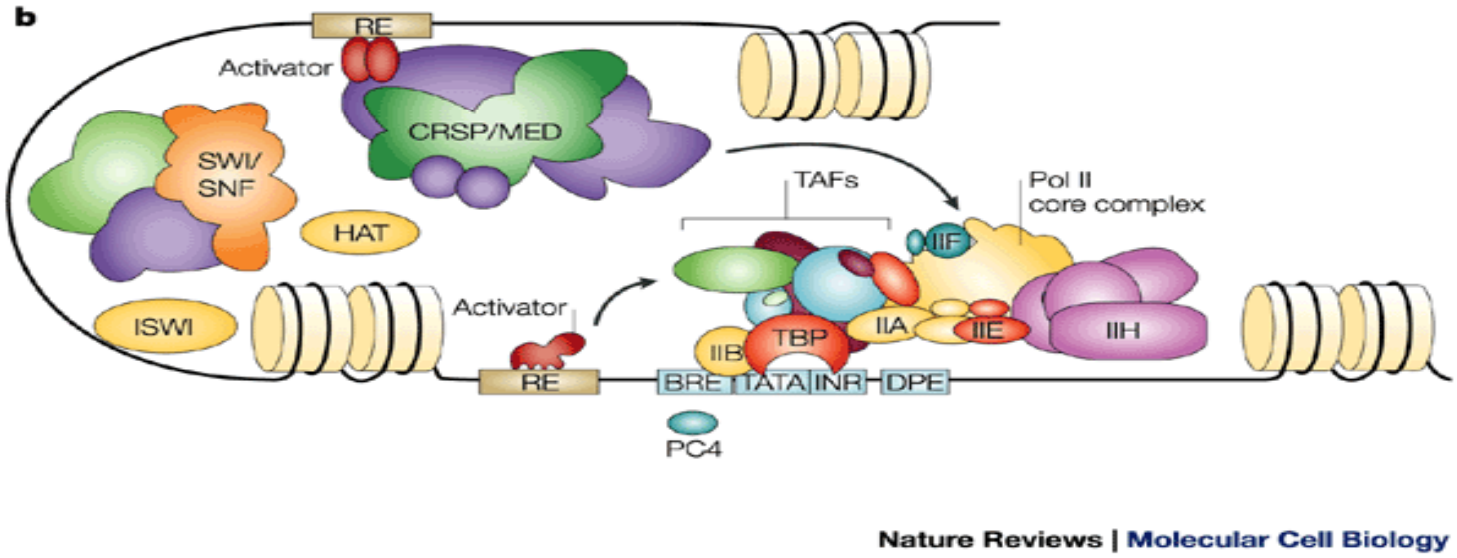


Figure 6–21 part 1 of 2. Molecular Biology of the Cell, 4th Edition.

- A re-assembly of stretches of DNA that are transcribed together plus some further editing on the mRNA level

What is a Gene (3)?



- A re-assembly of stretches of DNA that are transcribed together plus some further editing on the mRNA level plus parts of the sequence downstream that is necessary to regulate transcription of the gene

What is a Gene (4)? [GBR+07]

- The same gene?
  - Genes may generate different assemblies (differential splicing)
  - Genes may have interspersed genes
  - Gene have duplicates in the same genome
  - The „same" gene in another organism
  - Mutated genes
  - Common variations of a gene
  - …
- A gene?
  - Pseudo genes (never transcribed, yet highly similar)
  - Non-coding genes
  - miRNA (25 bases!)
- Gene definitions change(d) over centuries, decades, and last year

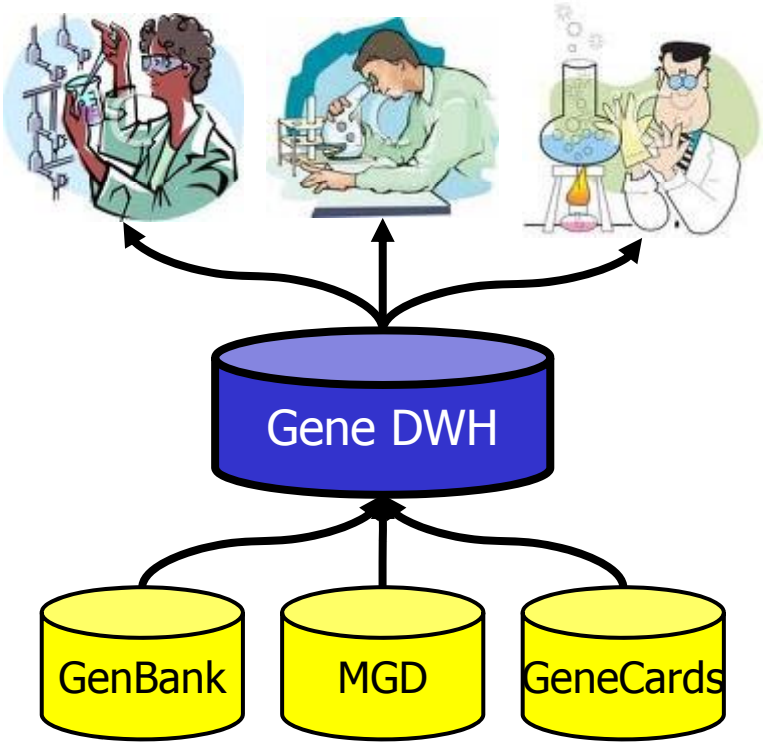
Does it Matter?

- Sometimes yes
  - E.g. to study differential splicing
  - E.g. to study regulatory relationships
- Sometimes no
  - E.g. to study gene function (without too many details)
  - E.g. to study gene interactions (without too many details)

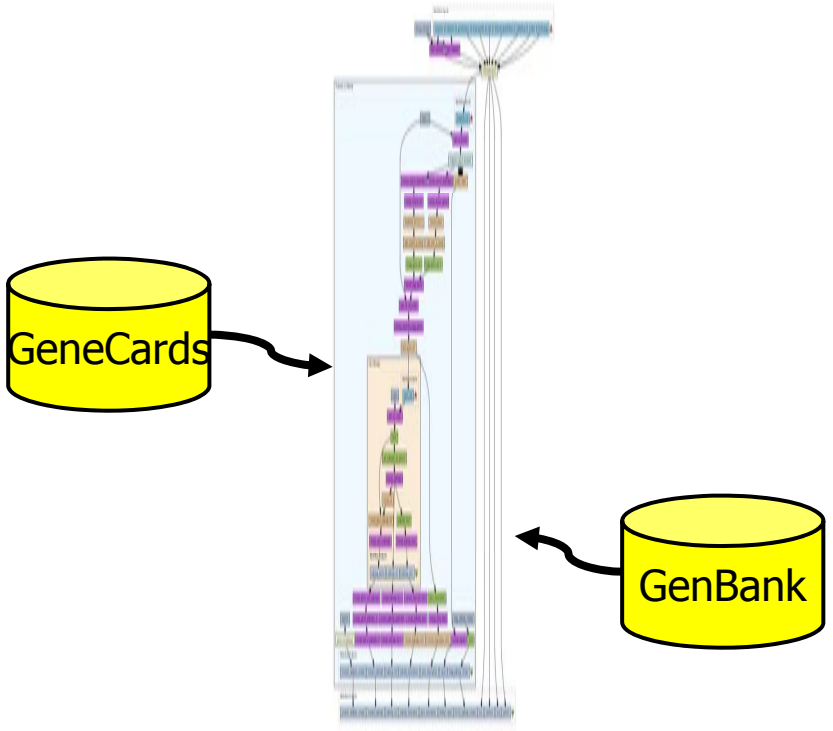
- Most studies today are carried out "without too much detail"
- E.g., detailed knowledge on splice variants and their functional differences is still almost non-existing
- Researchers know they are doing wrong, but it is the best they can do (now)

Is this a Problem?

Yes, if you plan to create a stable, precise, comprehensive integrated gene database



No, if you are pursuing a specific study taking into account your selection of genes



Is Data Quality an Issue in BDB?

- Most important quality aspects: Completeness and error-freeness
- BDB have terrible problems in both aspects
  - Complete collections exist nowhere (maybe except PDB and GenBank)
  - All BDB have a severe level of all kinds of errors
  - Many copy-and-paste problems (predictions become reality)
  - Most of the errors are statistical in nature (noise)
- Why? Recall: most BDB are filled from (high-throughput) experiments
  - Experiments that are not perfect
  - Measurements that are highly context-dependent
  - Performing the same experiment again will produce different results

Are BDB Autonomous?

- The big ones are maintained by specialized institutions
  - EBI, NCBI, EMBL, …
- Few of them have continuous, secured funding
  - Need to comply to calls from funding agencies
- It is (bad) tradition to reinvent the wheel all over again
  - 80 different software systems for microarray storage and analysis – with largely overlapping functionality [KZTL11]
- De-facto standards for some subfields
  - Gene Ontology, NCBI taxonomy, BioPax, SMBL, PSI-MI
  - Annotation guidelines: Minimum information about … (MIA*)

This Tutorial

- Part I – Data Integration for the Life Sciences
  - Biological Data & Biological Databases
  - Data Integration
  - Some Truths, some Myths
- Part II – Past and Presence
- Part III – Current Trends
- Part IV – Conclusions

Myth: Integration of 100dres of Databases

- There are hundreds of BDB
- Integrating 15-25 of them today is common practice
- But no project (we know of) needs to integrate >30-40 BDB

- Why not?

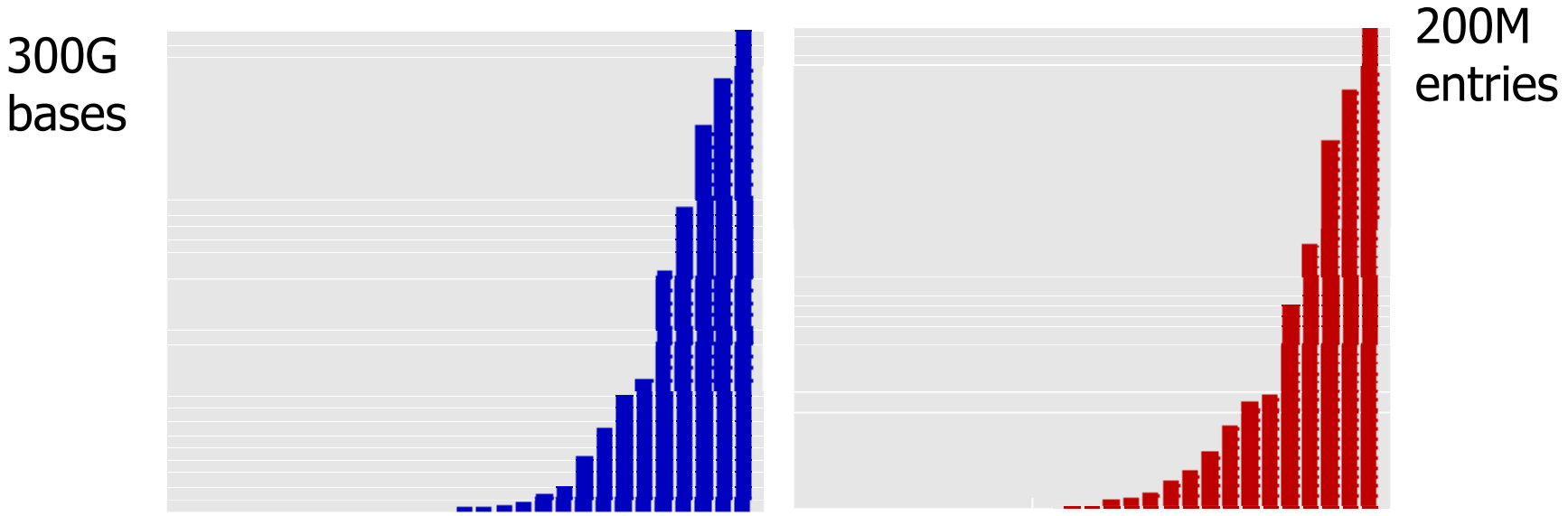
- Noise accumulates – the more joins over erroneous data, the larger the resulting error
- Nobody has such broad knowledge to pose any meaningful queries
- Nobody could review the papers stemming from the results ☺

Myth: Users do not Know Where to Search

- Transparency is nothing
- Provenance is everything

- User mostly know their favorite data sources very well
  - But pointing to alternatives can be helpful
- New databases have a very hard time before getting accepted
  - Unless created by big shots
- A piece of data without knowledge where it comes from is meaningless for most researches
  - How produced it? Which method? How many replications? Has it been confirmed? Where was it published? How paid for the study?

Was Myth: Data Volumes are so huge that Virtual Integration is Necessary



- All of EMBL now has ~150 TB (zipped), ENSEMBL hast ~1TB (MySQL dump), UniProt has ~5GB (zipped)
- Probably 90% of the 1300 DB's in NAR have <1GB
- Sequence data explodes due to Next Generation Sequencing
- Not many images (yet)

Truth: Every Piece of Information lives in Many Places – with Many Different Values

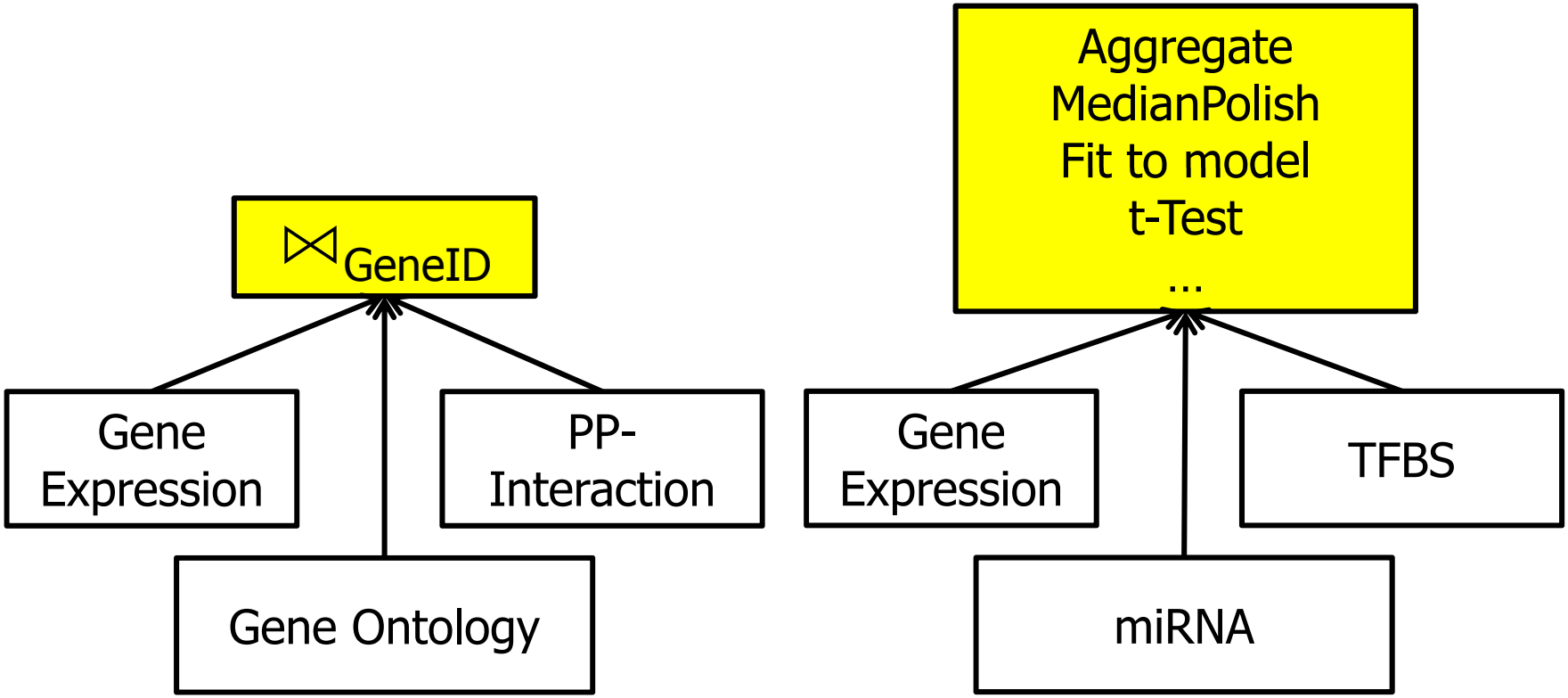
- For most classes of objects, there are more than one database that covers them
- Values often are contradicting
  - Different context, different conclusions, different facts
- Copy & paste errors

- Often, there is no true value
- Integrating different measurements usually is treated as a statistical problem
  - No majority voting etc.

Truth: Integration is Vital for Many Projects in the Life Sciences

Discrete Integration
Join by ID

Statistical Integration
Aggregate and test



This Tutorial

- Part I – Data Integration for the Life Sciences
- Part II – Past and Presence
  - Early Approaches (<2000)
  - State-of-the-art
- Part III – Current Trends
- Part IV – Conclusions

The Early Days (to the Best of our Knowledge)

- Large-scale digitization of biological data started in the early 90ties
  - Data volumes grew too big to be handled manually
  - Mostly sequences (DNA, proteins)
- Human Genome Project: Designed as a collaborative effort
  - Data sharing and integration considered crucial for project success
- First calls for data integration infrastructures in the early 90ties
  - "If the informatics is not handled well, the HGI could spend billions of dollars and researchers might still find it easier to obtain data by repeating experiments than by querying the database. If this happens, someone blew it". Robbins, NSF program director, 1991
- First functional systems: 1993 – 1995
- Database people jumped in: 1994-
- First workshops: 1994/1995
- Explosion of papers / prototypes: 1995-

First Systems

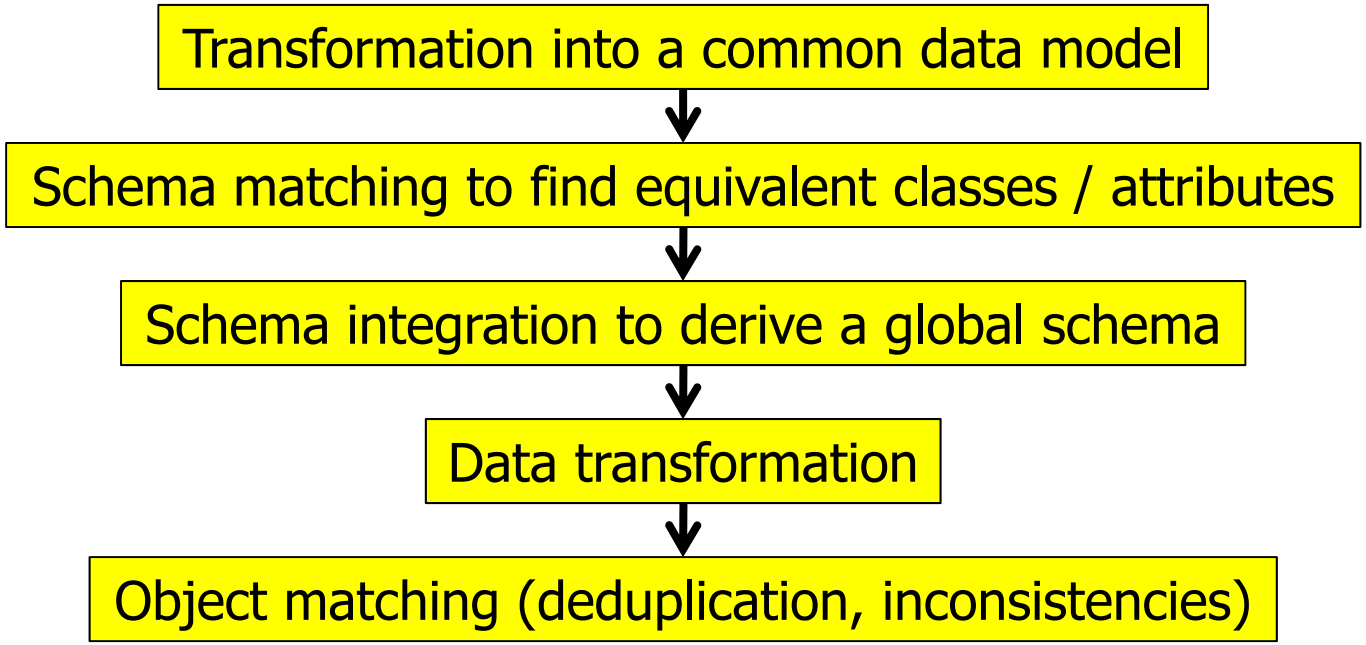
- First functional systems emerged around sequence databases
  - Etzold, T. and Argos, P. (1993). "SRS - an indexing and retrieval tool for flat file data libraries." CABIOS
    - Still working and probably still the most popular system to date
  - Akiyama, Y., Goto, S., Uchiyama, I. and Kanehisa, M. (1995). "WebDBGET: an integrated DB retrieval system which provides hyper-links among related database entries". 2nd Meeting on IMDB
    - Still working
  - Ritter, O. (1994). The Integrated Genomic Database (IGD). In Suhai, S. (ed). Book "Computational Methods in Genome Research". Plenum Press
    - Built on proprietary technique and failed quickly
- From the start, database entries were link-rich
  - IDs from other databases
  - Became HTTP-links on the web

First Contributions

- Citations
  - Karp, P. D. Ed. (1994). "Report of the Workshop on Interconnection of Molecular Biology Databases", SRI, Stanford, California
    - http://www.ai.sri.com/pkarp/mimbd/94/abstracts.html
  - Karp, P. D., Ed. (1995). "2nd Meeting on Interconnection of Molecular Biology Databases". Cambridge, UK
    - http://www.ai.sri.com/pkarp/mimbd/95/abstracts.html
- Early input from DB people (examples)
  - Davidson / Buneman: Semistructured data; declarative transformations
  - Wiederhold: Mediators; semantic heterogeneity
  - Goodmann: Modularization, standardization, software design
  - Spaccapietra: Schema integration, schema mappings (correspondences)
  - Kemp: Functional data model
  - Wong / Kosky: Distributed query optimization

Influential Paper

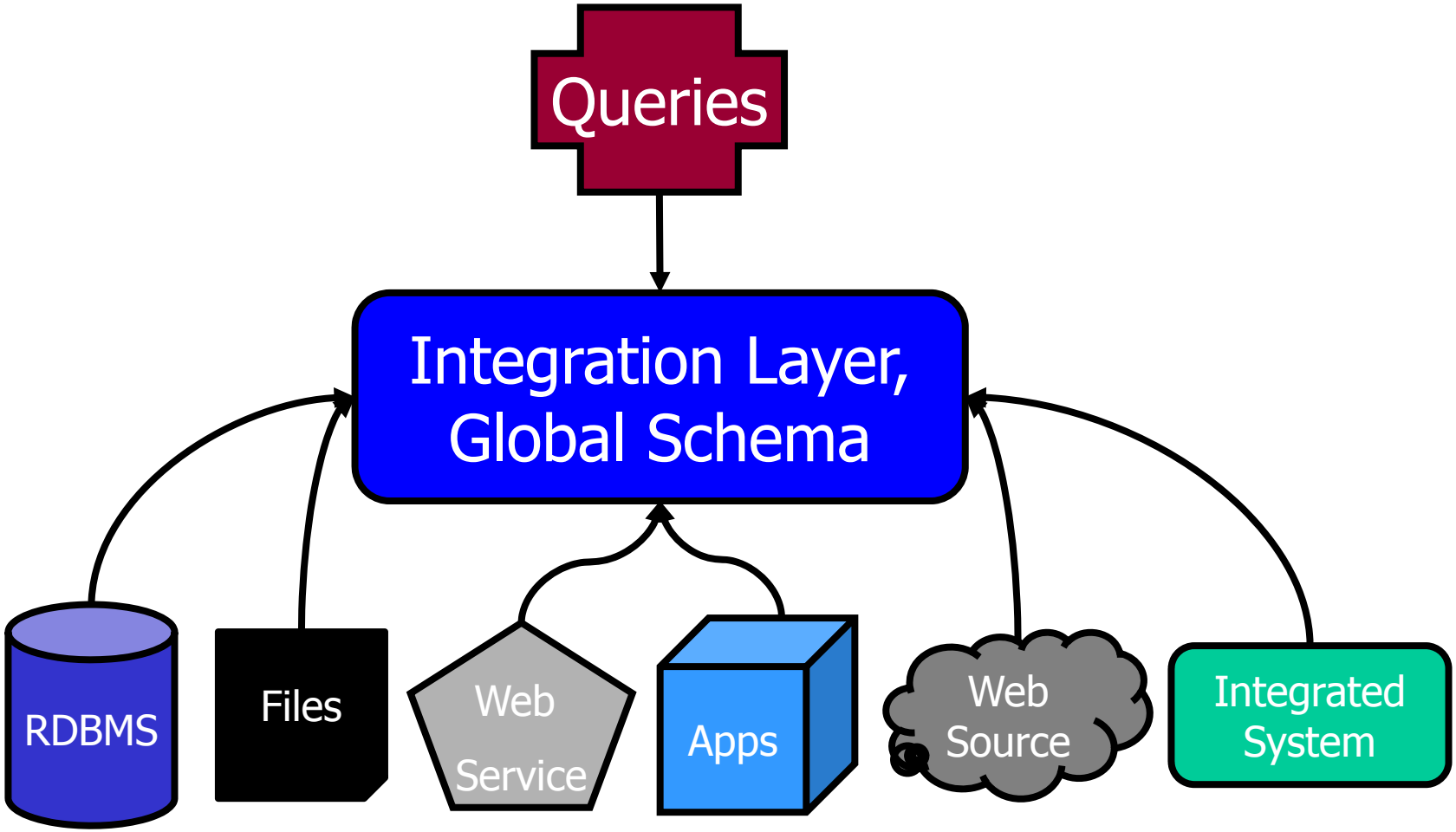
- Defined the framework on what would happend in the next years
  - Davidson, S., Overton, G. C. and Buneman, P. (1995). "Challenges in Integrating Biological Data Sources." *Journal of Computational Biology*
- Two classes of systems: Federated or materialized
  - Autonomy and currentness versus performance and reliability



„Classical“ Systems

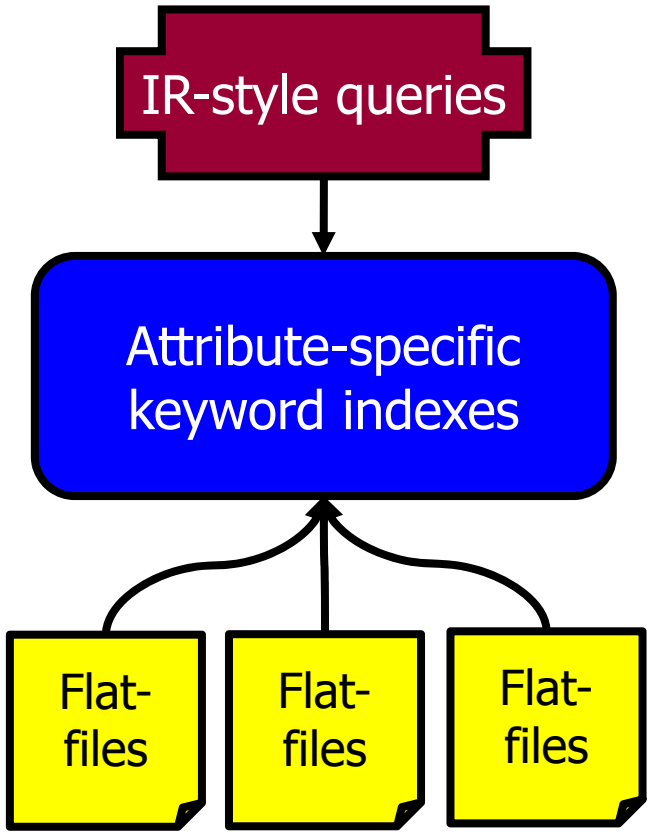
- SRS: Flat file indexing
  - Also representing DBGet, Entrez, Atlas, …
- Kleisli: Multi-database query language
  - Also representing OPM, P/FDM, …
- DiscoveryLink: Federated database
  - Also representing BioMediator, caGRID, …
- TAMBIS: Ontology-based integration
  - Several follow-ups in the early 2000: SEMEDA, BACIIS, …

Common Ground



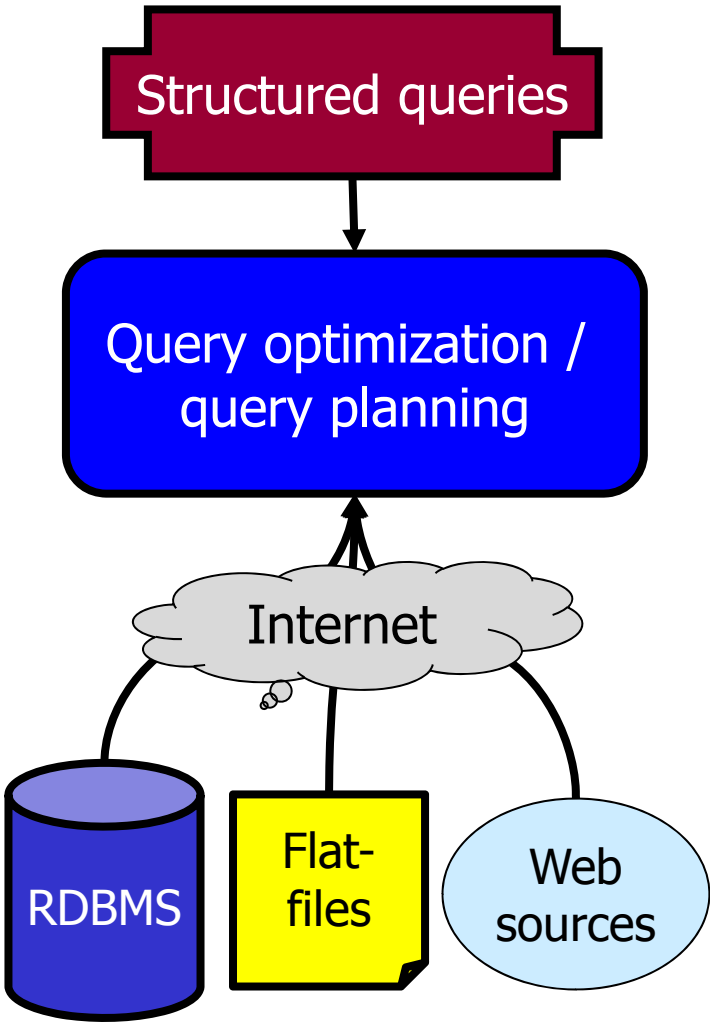
SRS

- Architecture
  - Flat-files are parsed into a semi-structured model
  - Per-attribute textual indexes
  - IR-style queries
- Features
  - Semantic free: No semantic integration, no deduplication, no schema matching, no data fusion
  - No distributed access
  - Only simple types of structured queries
  - Joins following links on instance level
- Extremely successful
- Largely ignored by DB community



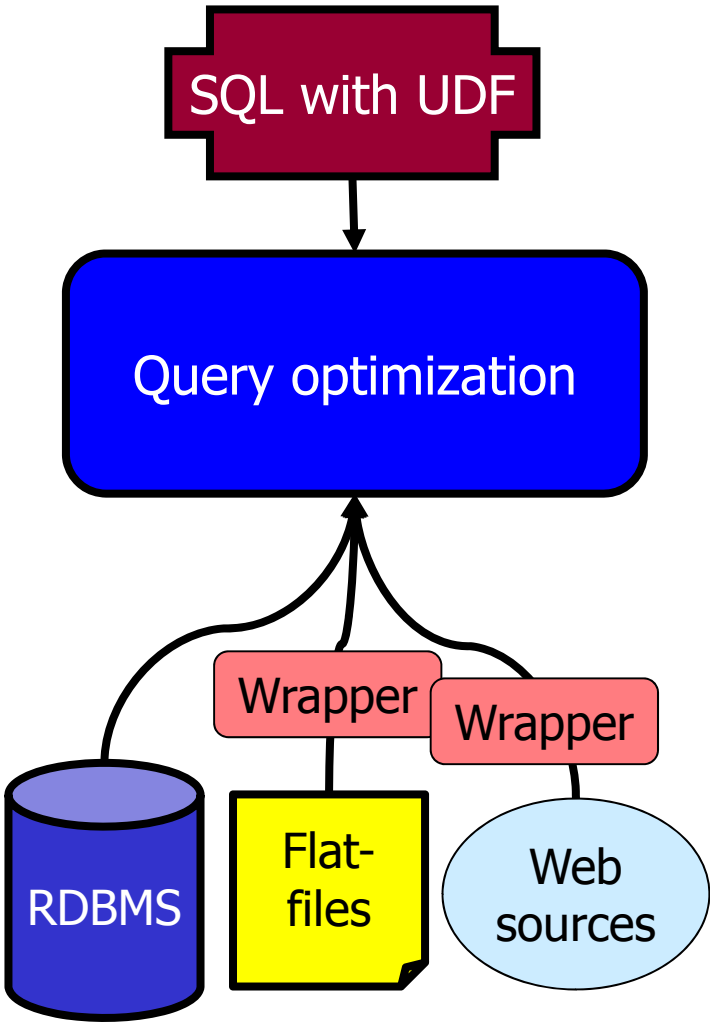
Kleisli / BioKleisli / K2

- Architecture
  - Structured multi-database query language
  - Emphasis on distributed query optimization
- Features
  - Semantic integration must be achieved by users through complex queries
  - No object deduplication, no data fusion
  - Sources may be distributed
- Popular in DB community
- BioKleisli resulted in commercial system for some years



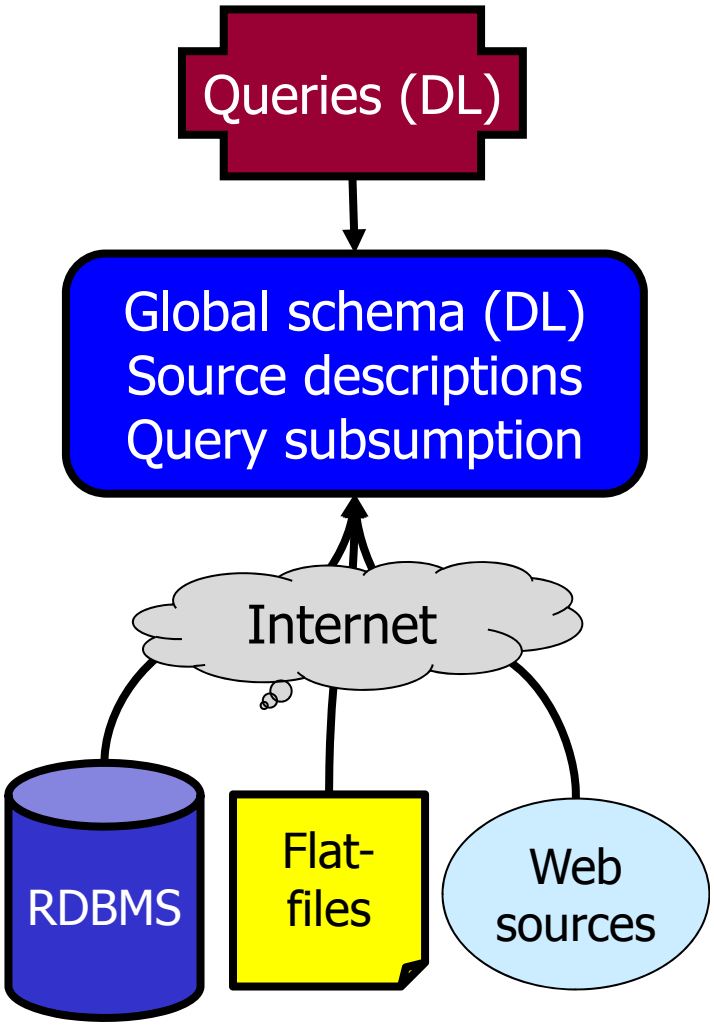
DiscoveryLink

- Architecture
  - Federated database
  - Queries over non-relational sources
- Features
  - Semantic integration must be achieved by defining proper (relational) views
  - No object deduplication, no data fusion
  - Sources may be outside DB (and can be distributed)
- Very popular in DB community
- Started as Garlic, commercially marketed as DiscoveryLink, stopped very soon



TAMBIS

- Architecture
  - Mediator-based architecture
  - Emphasis on semantic integration by ontology-based query rewriting
- Features
  - Source descriptions in Description Logic
  - Query planning as subsumption
  - Full semantic integration on schema level
  - No object deduplication, no data fusion
  - Sources may be distributed (Kleisli)
- Builds on previous work in DB community (SIMS, Kleisli)
- Only prototype



Summary

	SRS	Kleisli	Discoverylink	TAMBIS
Global schema	No	No (queries)	No (views)	Yes
Distributed data	No (later added)	Yes	Not in focus	Yes (Kleisli)
Virtual	Somehow	Yes	Yes	Yes
Global data model	-	Nested collections	Relational	DL
Data handling	No	No	No	No
Process integration	Limited	No	No (UDF)	No

Impact in the Life Sciences

- Except SRS / Entrez, systems were essentially ignored in the LS community

- Many citations (from DR) but negligible practical impact
- None of the DB-drive systems still in use today (maybe K2?)
- Many have never been used in practice

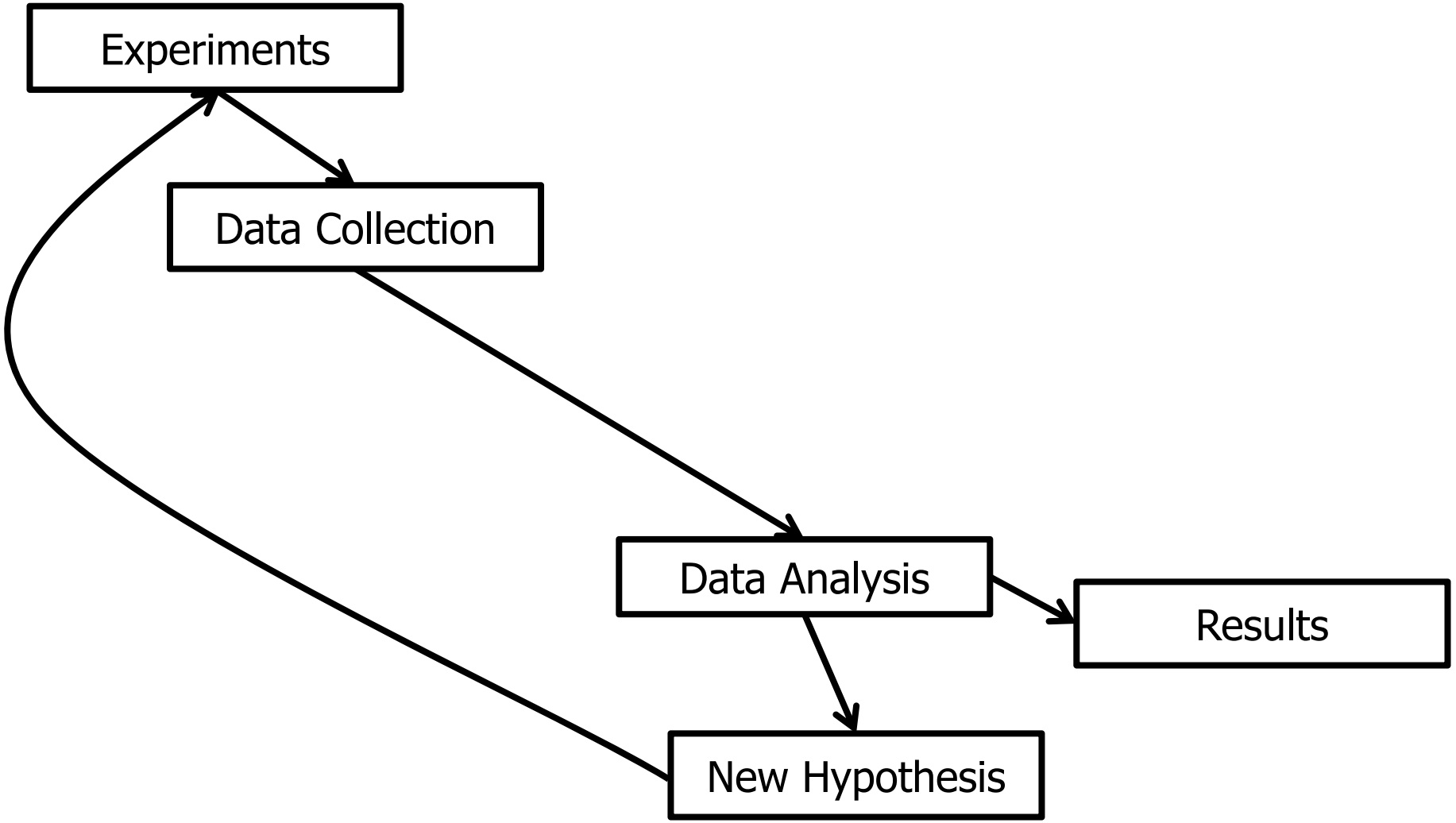
Essence of the Approaches from DR

- Fix your set of sources to be integrated
- Build one schema (GS) embracing all others
  - Global (TAMBIS) or user-defined (Garlic etc.)
  - Goal: Non-redundant, minimal, comprehensible
  - Semantic integration – homonymy, hyponymy, partonomy, …
- Wait for users to pose queries against GS
  - Use schema mappings for query rewriting

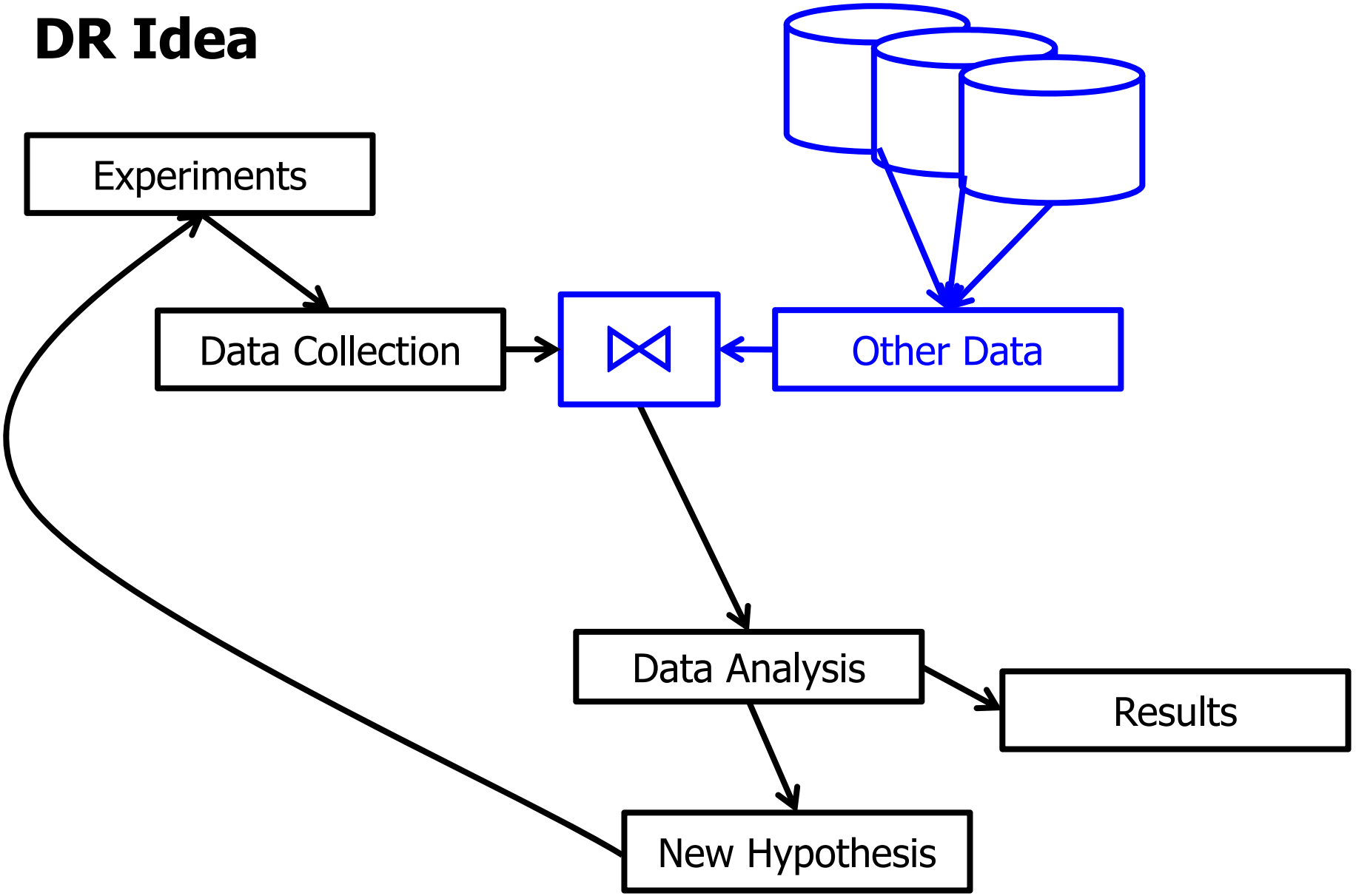
Possible Explanations

- Focused on schemas, while biologists focus on data
  - Content is king
- Virtual integration prevents changing the data
  - Statistical integration often needs to manipulate data
- Transparency hides provenance as indicator for quality
- Approaches tried to remain domain-independent
  - Genes cannot be compared with the same methods as person names – different error models, different primary data, different additional data, different types of "equality"
- DR target discrete integration, while LS thinks in statistical integration
  - Schema, queries, mappings, …
  - Sequence alignment, normal distribution, error models, …

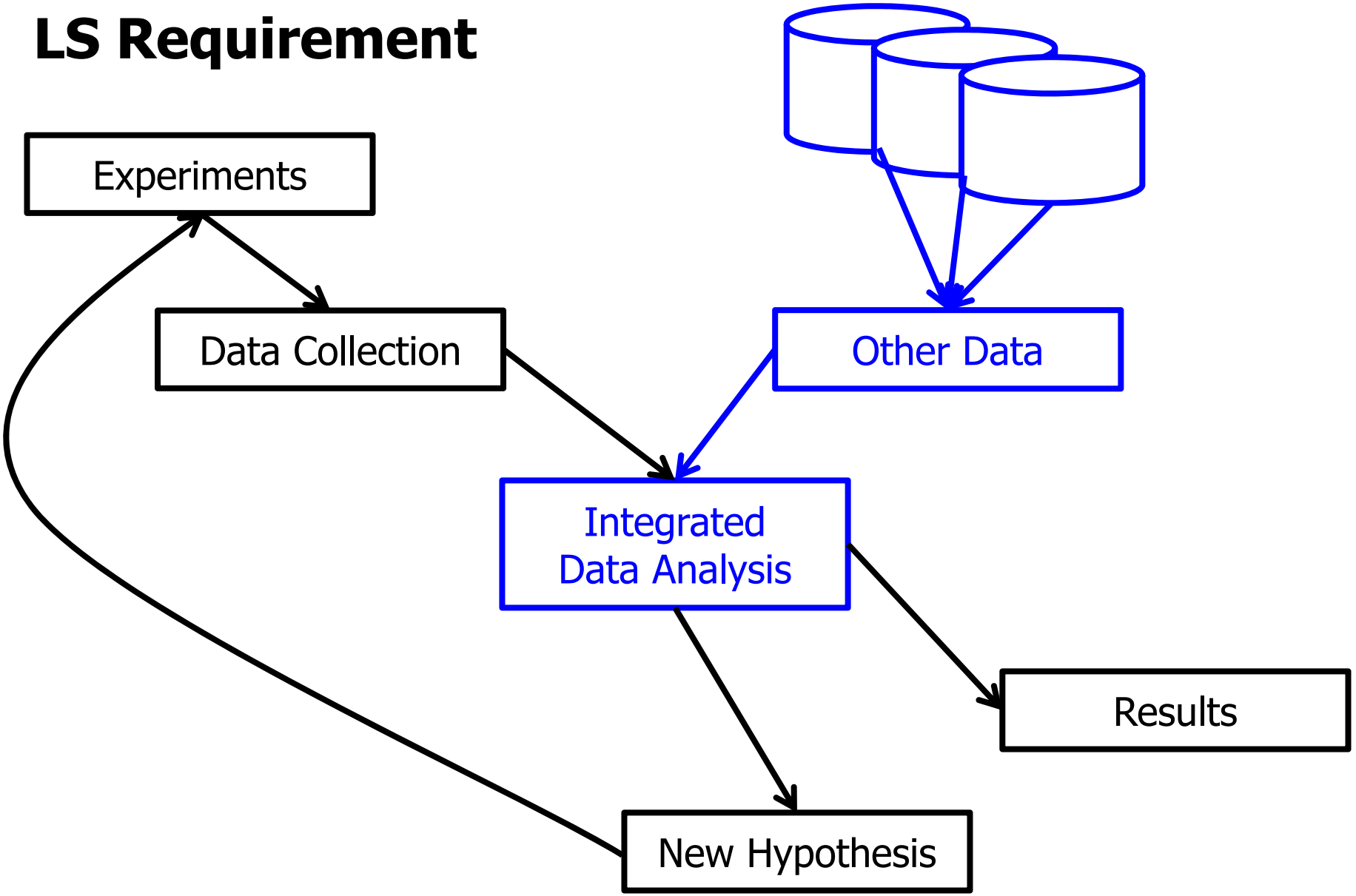
Life Science Research Food Chain



DR Idea


Experiments
Data Collection
⋈
Other Data
Data Analysis
Results
New Hypothesis

LS Requirement



Success Stories

- General DR research
  - Relational technology
  - Idea of modeling data
  - Importance of versioned data

- Integration technologies
  - Controlled vocabularies (ontologies)
  - Data Warehouse architecture (ETL – little OLAP)
  - XML (for data exchange)

Other Way round: Influence of LS-DI on DR

- XML after UnQL which cites ACeDB as a main motivation
- Still, many DR-DI papers use LS requirements as motivation
- Motivation for research in topics such as
  - Information integration in general
  - Quality-based source selection
  - Integration of string search capabilities in DBMS
  - Wrapper development (query–to–parser)
  - Integration of web sources
  - Semistructured data
  - Coping with limited source capabilities (web integration)
  - Data fusion
- One reason: All these BDB are really available – for free

This Tutorial

- Part I – Data Integration for the Life Sciences
- Part II – Past and Presence
  - Early Approaches (<2000)
  - State-of-the-art
- Part III – Current Trends
- Part IV – Conclusions

The Presence

XML + Perl + MySQL

- Or better

XML +
(Perl | Java | Python) +
(MySQL | Oracle | PostGreSql)

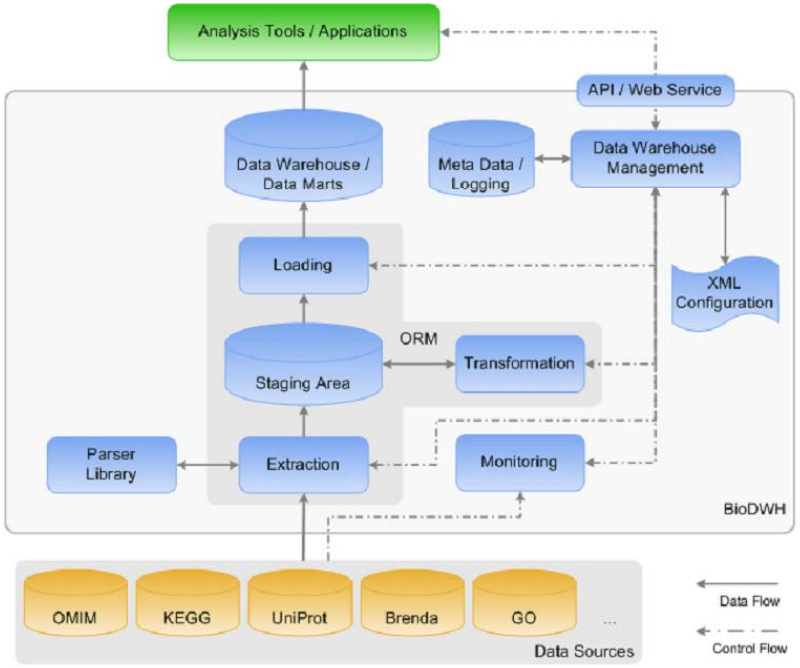
- Big role of open source libraries and frameworks
- Ontologies are common practice

The Presence

- "Data Warehouses" approaches everywhere
  - Virtual integration is mostly dead
    - Despite frequent papers stating the opposite
  - Survival in some niches: DAS, some mash-ups
- Semantic integration performed manually (wrappers)
  - No schema matching, little query rewriting
- Several systems up-and-running integrating dozens of sources
  - Freshness in the presence of data cleansing remains a hard problem

BioWarehouse [LPW+06]

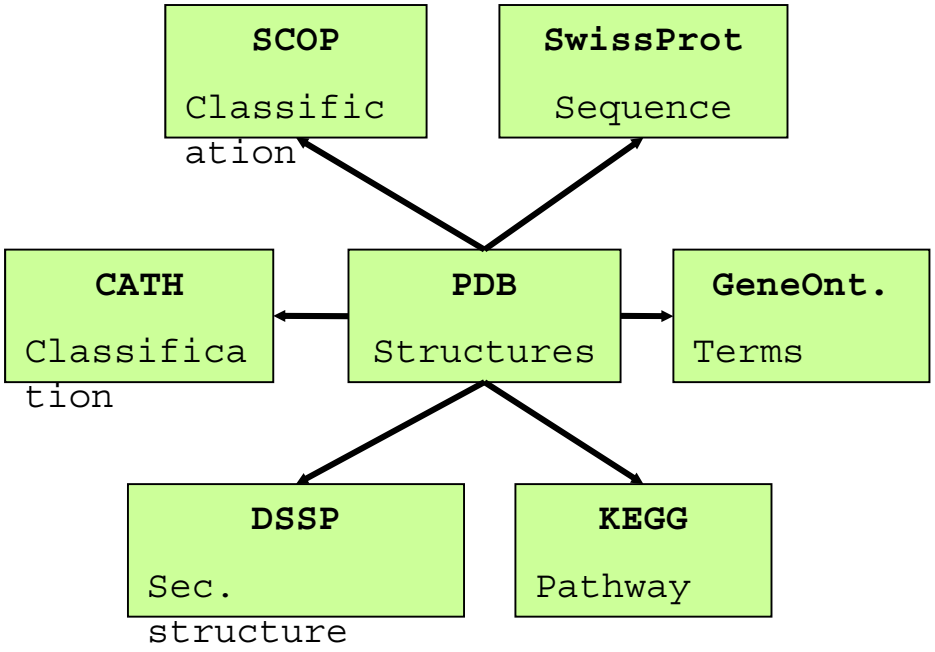
- Standard ETL design
- Unified schema defined manually
  - Leads to semantic differences within tables
  - No cleansing or de-duplication
  - Mappings are programmed in the „loaders"
- Loader for 14 sources
- Full provenance information
- Ships with JAVA lib and GUI



[LPW+06]

Columba [TRM+05]

- Integrates 12 sources describing aspects of protein structure
- Standard DWH approach
  - Custom-made wrappers
  - Reuse of open source tools
- Multidimensional integration
  - Each source builds its own domain
  - Semantic overlaps are not resolved
  - Provenance information attached to the dimension table



EnsMart/BioMart [KKS04]

- Multidimensional access to the Ensembl database
  - Reverse star schema
  - Frequent changes to the APIs
  - Perl, web services, Taverna, …
- Highly complex & un-documented creation process
- Full-fledged web interface
- Very successful
  - Dozens of installations around the world
  - Used by many for accessing genomics data (R/Bioconductor)

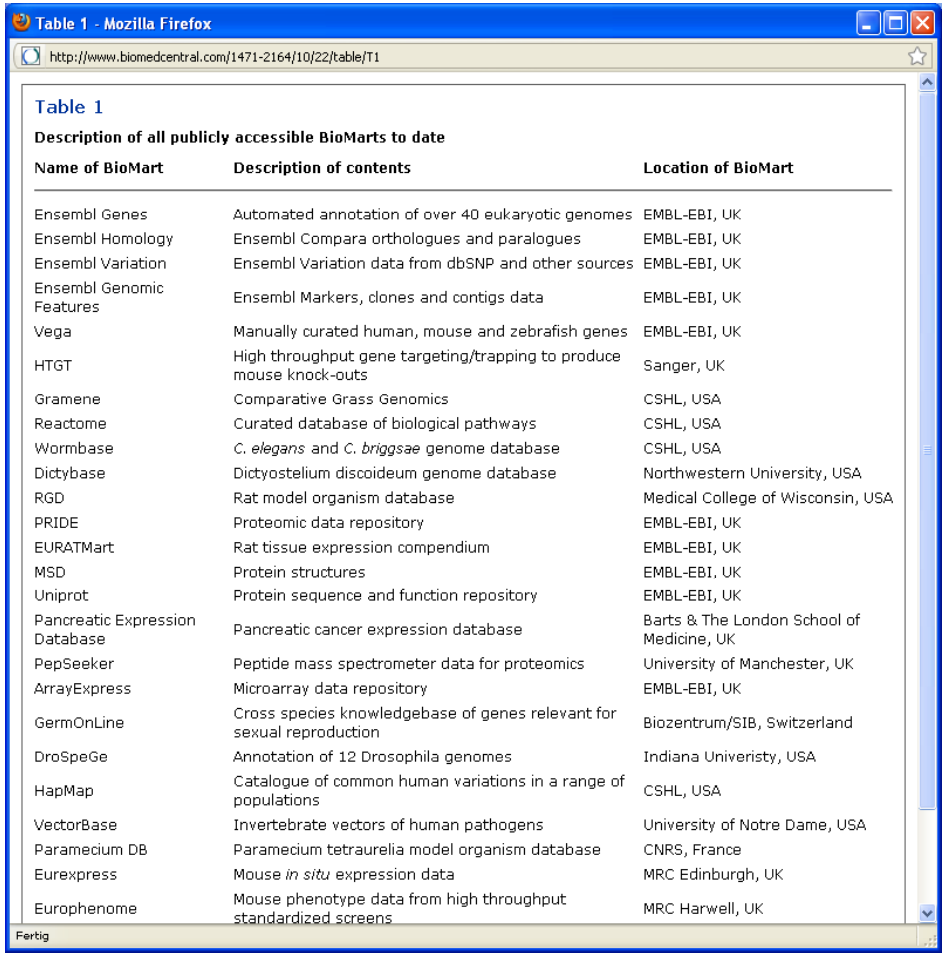
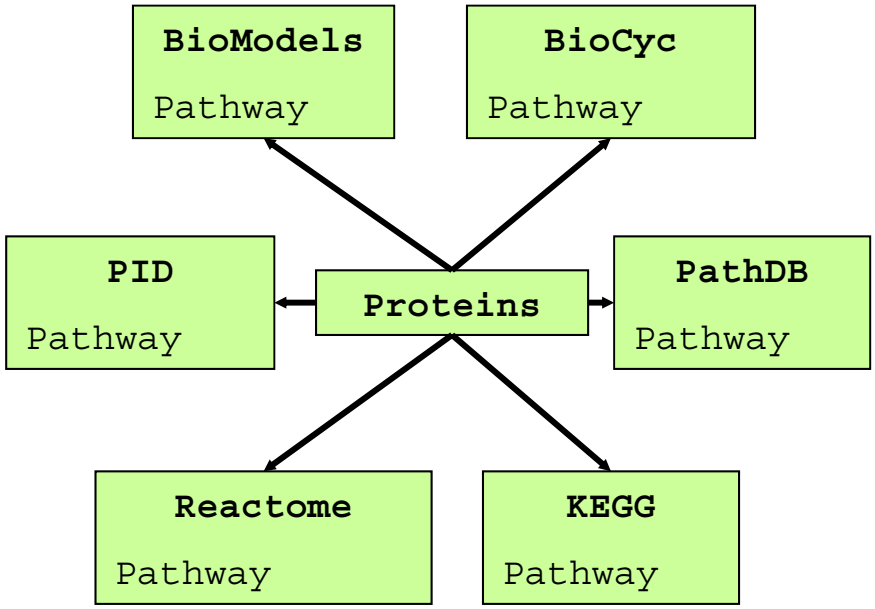


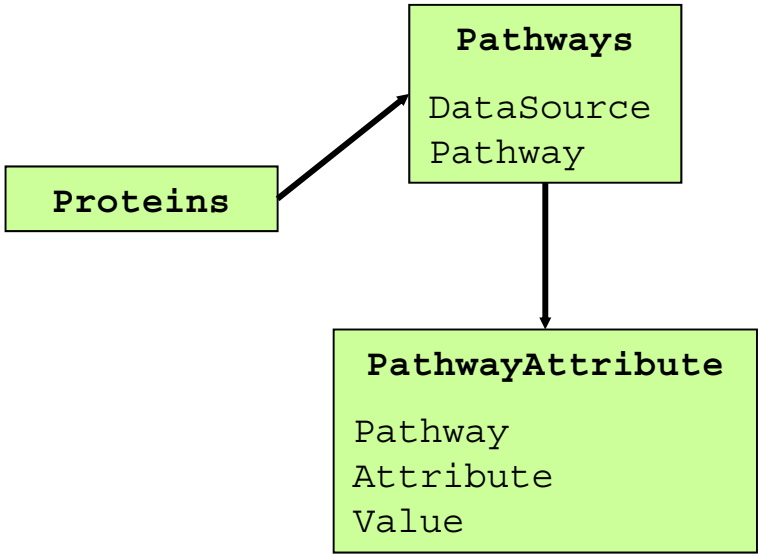
Table 1

Description of all publicly accessible BioMarts to date

Name of BioMart	Description of contents	Location of BioMart
Ensembl Genes	Automated annotation of over 40 eukaryotic genomes	EMBL-EBI, UK
Ensembl Homology	Ensembl Compara orthologues and paralogues	EMBL-EBI, UK
Ensembl Variation	Ensembl Variation data from dbSNP and other sources	EMBL-EBI, UK
Ensembl Genomic Features	Ensembl Markers, clones and contigs data	EMBL-EBI, UK
Vega	Manually curated human, mouse and zebrafish genes	EMBL-EBI, UK
HTGT	High throughput gene targeting/trapping to produce mouse knock-outs	Sanger, UK
Gramene	Comparative Grass Genomics	CSHL, USA
Reactome	Curated database of biological pathways	CSHL, USA
Wormbase	*C. elegans* and *C. briggsae* genome database	CSHL, USA
Dictybase	Dictyostelium discoideum genome database	Northwestern University, USA
RGD	Rat model organism database	Medical College of Wisconsin, USA
PRIDE	Proteomic data repository	EMBL-EBI, UK
EURATMart	Rat tissue expression compendium	EMBL-EBI, UK
MSD	Protein structures	EMBL-EBI, UK
Uniprot	Protein sequence and function repository	EMBL-EBI, UK
Pancreatic Expression Database	Pancreatic cancer expression database	Barts & The London School of Medicine, UK
PepSeeker	Peptide mass spectrometer data for proteomics	University of Manchester, UK
ArrayExpress	Microarray data repository	EMBL-EBI, UK
GermOnLine	Cross species knowledgebase of genes relevant for sexual reproduction	Biozentrum/SIB, Switzerland
DroSpeGe	Annotation of 12 Drosophila genomes	Indiana Univeristy, USA
HapMap	Catalogue of common human variations in a range of populations	CSHL, USA
VectorBase	Invertebrate vectors of human pathogens	University of Notre Dame, USA
Paramecium DB	Paramecium tetraurelia model organism database	CNRS, France
Eurexpress	Mouse *in situ* expression data	MRC Edinburgh, UK
Europhenome	Mouse phenotype data from high throughput standardized screens	MRC Harwell, UK

Comparison





- Source-specific sub-schemata
- Provenance encoded in tables
- Difficult UNION, simple value-based selection

- Generic pathway table
- Will contain data with different semantics
- Simple UNION, difficult value-based selection

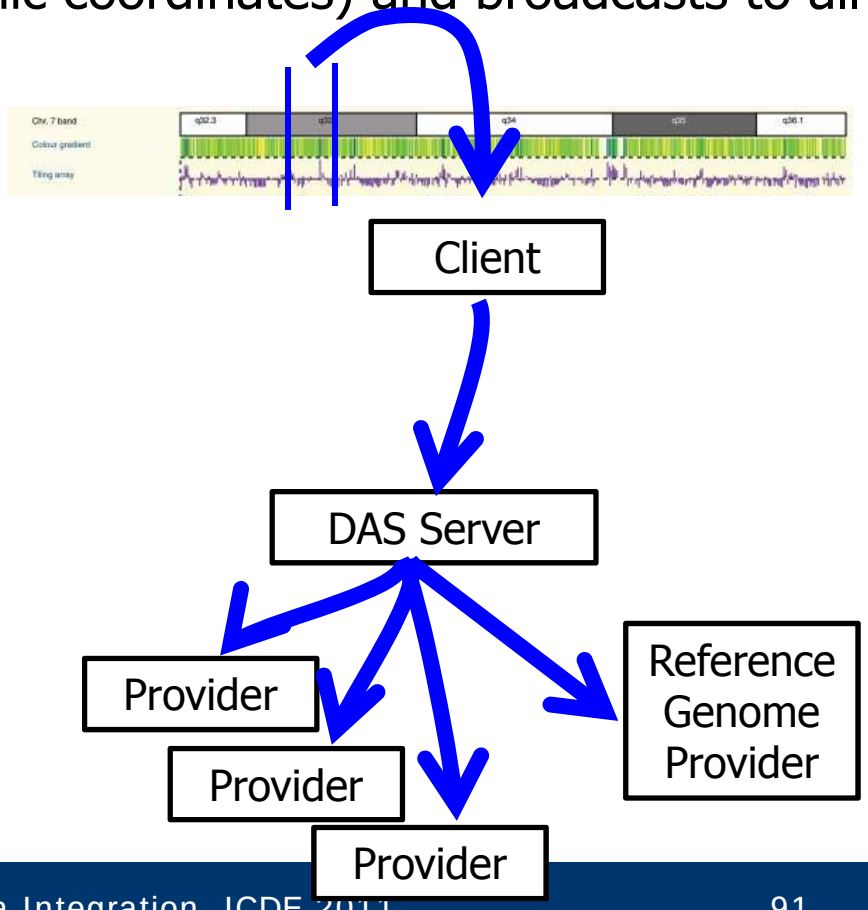
… and many more …

- All following the „DWH"-approach

- GUS [DCB+01]
- IMG [MKP+05]
- ArrayExpress [SPLO05]
- Atlas [SHX+05]
- Biozon [BY06]
- GeWare [RKL07]
- GenoQuery [LLF08]
- …

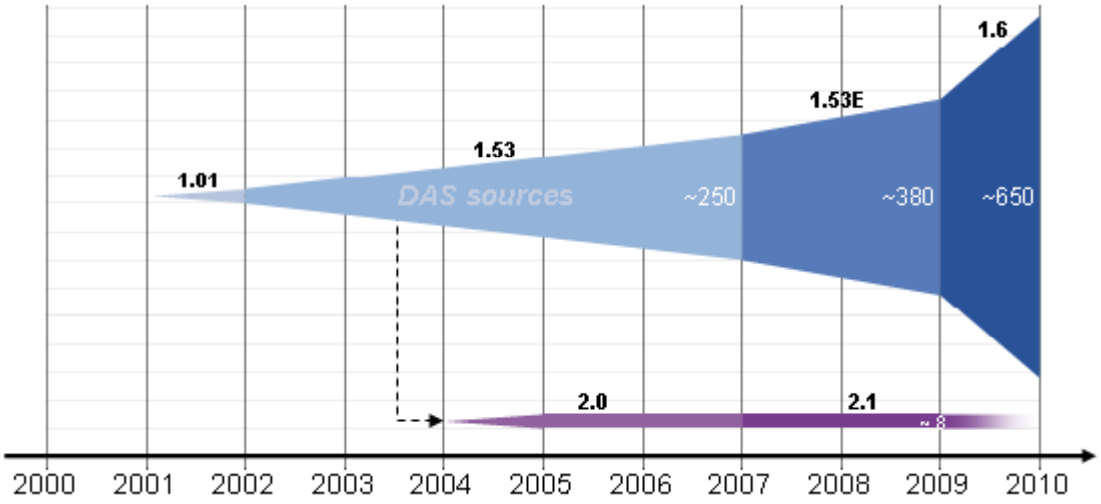
Notable Exception 1: Distributed Annotation System [JAB+08]

- Federated system serving a single type of information
  - Genomic annotation
- DAS server receives query (genomic coordinates) and broadcasts to all DAS providers
- Results are bundled and reported
- No semantic integration,
  no annotation types,
  simple XML format,
  very simple protocol,
- Highly successful



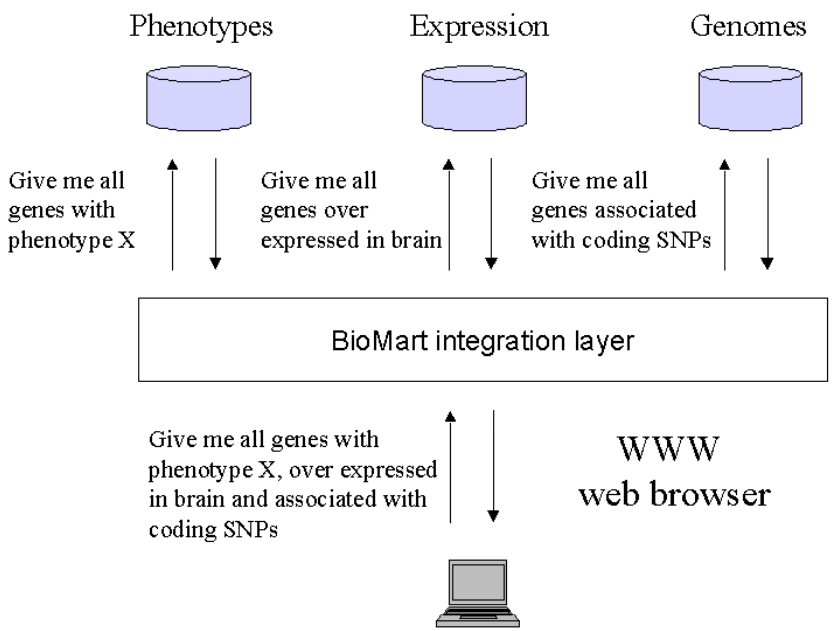
Notable Exception 1: Distributed Annotation System [JAB+08]

- Federated system serving a single type of information
  - Genomic annotation
- DAS server receives query (genomic coordinates) and broadcasts to all DAS providers
- Results are chained and reported
- No semantic integration,
  no annotation types,
  simple XML format,
  very simple protocol,
- Highly successful



Notable Exception 2: BioMart

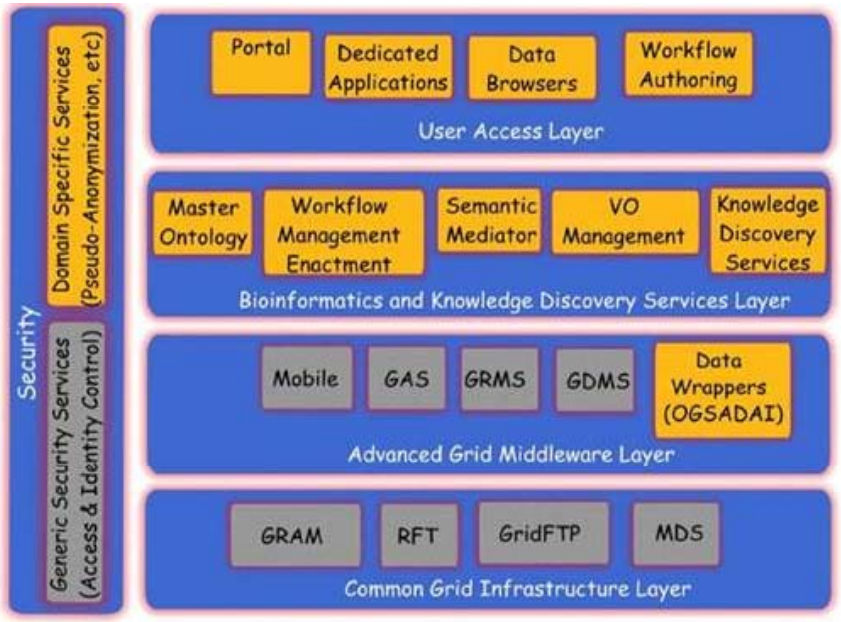
- BioMart actually is capable of accessing distributed data sources
- Source schemas must comply to BioMart layout and naming conventions
- Links and schemas have to be declared and configured in the middleware
- No semantic integration, no query optimization / rewriting
- BioMart Portal: >100 databases
- Full provenance information
  - You query a source, not a relation
- Highly successful



Notable Exception 3: caBIG

caBIG™
cancer Biomedical Informatics Grid™

- Heavy-weight, full fledged data sharing and analysis middleware
  - Model-driven architecture, XML, Semantic Web, SOA, Grid, …
- Top-down development (heavily criticized), funded by the NCI
- Four levels of compatibility to caBIG standards
- Slow adoption
- "Critics of the massive project say it's inaccessible. Champions say the payoff requires embracing the new language, and culture, of bioinformatics." [JNCI news, 2/2010]

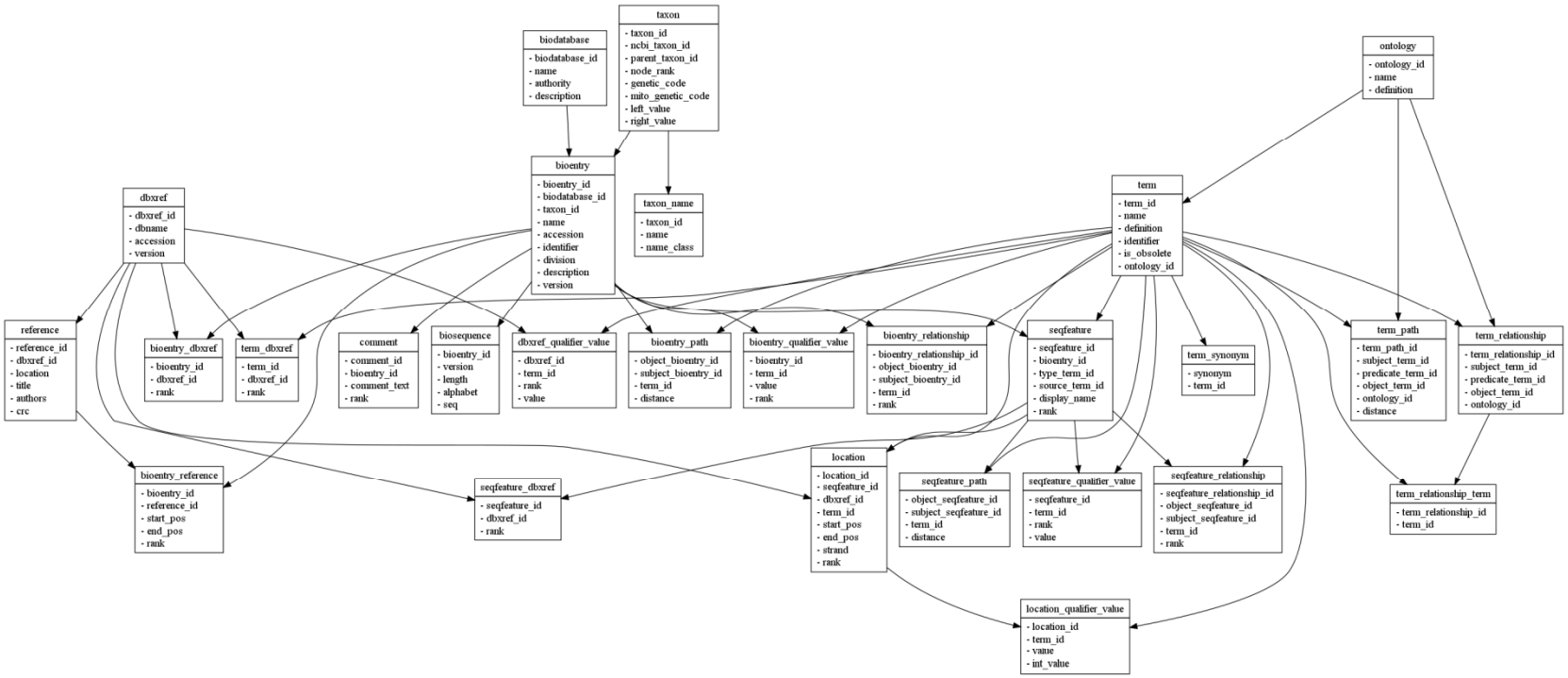


[SOH+06]

Wrap-Up

- Probably >95% of integration projects use materialization
- Successful systems implemented by domain scientists, with little participation of DR
  - Exception: caBIG
- Little automatic semantic integration, very little distributed query optimization, very little data fusion, very little schema matching / schema integration
- Full provenance information
- Exceptions only support canned queries and require standardized schemas

BioSQL [http://www.biosql.org/]



- Generic relational schema for representing sequences and features
- Standard storage layer for BioPerl, BioPython, BioJava
- Ready-made parsers from Genbank, UniProt, NCBI Taxonomy, …

GMOD [SMS+02]

- "**GMOD** is the **G**eneric **M**odel **O**rganism **D**atabase project, a collection of open source software tools for creating and managing genome-scale biological databases"
- Developed by app. 20 organizations
- Ships with schema (Chado), genome browser, annotation pipeline, exchange middleware, web-app development tool, …

- Essentially everything that many small/midsize genome projects need
- Of course: Integrating several GMOD databases is fairly simple

Ontologies

- Ontologies definitely are a success story in LS since ~2000
  - OBO hosts ~130 ontologies, BioPortal ~200
  - Most famous: Gene Ontology, ~30.000 concepts, used world-wide
- Almost all are simply DAGs of ISA relationships (and PART_OF)
- Usage as structured, controlled vocabulary
  - Speaking about the same thing
  - Function prediction, semantic similarity, Text Mining, …
- Very little usage of logical inference: no constraints, roles, axioms, …
- Remove semantic heterogeneity in data integration upfront
  - At the instance / value level
  - Take the role of standards

This Tutorial

- Part I – Data Integration for the Life Sciences
- Part II – Past and Presence
- Part III – Current Trends
  - Data Integration Workflows
  - Semantic Web
  - Ranking in Integrated Datasets
- Part IV – Conclusions

Lesson's Learned - Observations

- Semantic and technical heterogeneity, data distribution, redundancy and inconsistencies are real problems in the Life Sciences

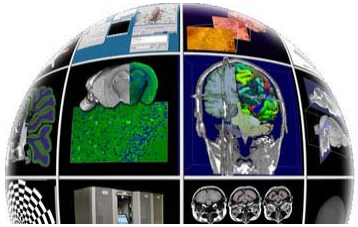
- Data volume is not much of an issue, nor is it up-to-date'ness
- Materialization possible and viable
  - Faster, data cleansing, more robust, easier to build and maintain
- Virtual integration is only pursued under very specific conditions
  - Restricted queries, semantic heterogeneity removed up-front
- General technique to raise the level of automation did not find much uptake

Increasing Need

- Integration is more necessary than ever
  - Holistic, comprehensive, genome-wide, data-driven … everywhere
  - Systems biology, translational medicine, biodiversity, personal medicine, …
- Encompasses data integration and information integration
  - Ever growing number and diversity of available data sources
  - Ever growing repertoire of high-throughout techniques
    - Most of the raw data is statistical in nature
- Breadth of scientific questions increases
  - Calling for an integrated view on data from many fields
  - Example personalized medicine: genome, pedigree, environment, intoxication, medical status, …

Example Large-Scale Projects

- Large-scale EU framework 7 project
  - „To construct and operate a sustainal… ture for biological information in Europe t… ce research and its translation t… ."
  - 32 organizations, 27…
- Biomedical Inform… work
  - "… is a nati… data-sharing infrastr… techniques, and advisory se…
  - …ons, 14 million / year
- C…
  - … Biomedical Informatics Grid® initiative… to …data and knowledge, simplify collaboration, speed research … realize the potential of Personalized Medicine."
  - >50 centers, 20 million / year



Note: The successful DI projects we know of were grass-root style

Open Challenges

Effort	Integrating dozens of data sources still requires considerable effort
Analysis	Interesting (from a LS perspective) DI problems require complex analysis processes
Provenance	Users want to know exactly where each piece of data comes from
Quality	Finding the right answer, not „finding any answer" or "finding all answers"

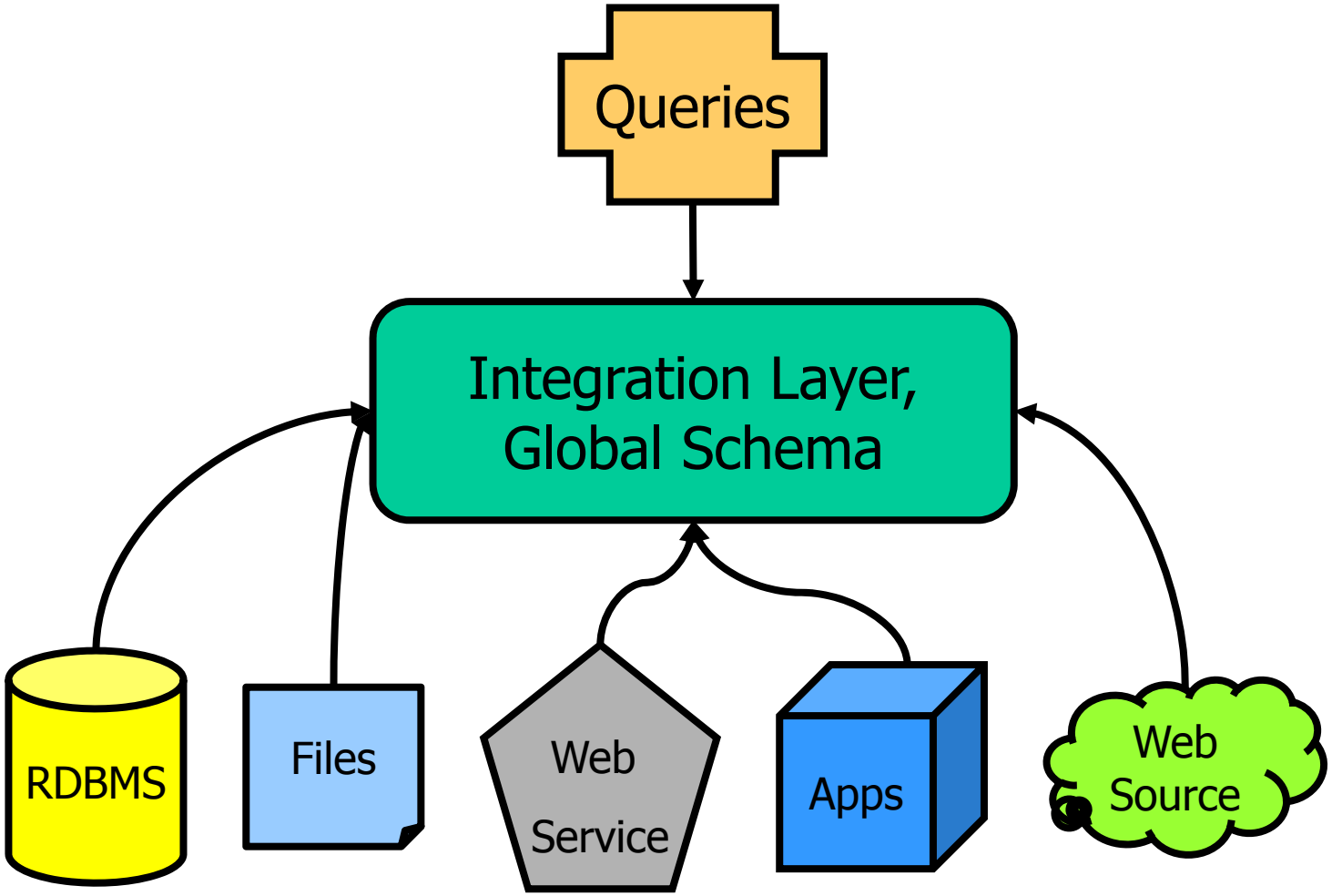
Three Trends

Data Integration Workflows	• Integration means analysis, and analysis means integration • No schemas, no explicit semantics • Scientific workflow systems	Effort Analysis Provenance Quality
Ranking	• Report results in a biologically meaningful order • Stays with queries, adds ranking • Requires a DI system in place	Effort Analysis Provenance Quality
Semantic Web	• Reduce upfront cost of DI • No schemas, explicit semantics • Semantic Web tech. (RDF, SPARQL)	Effort Analysis Provenance Quality

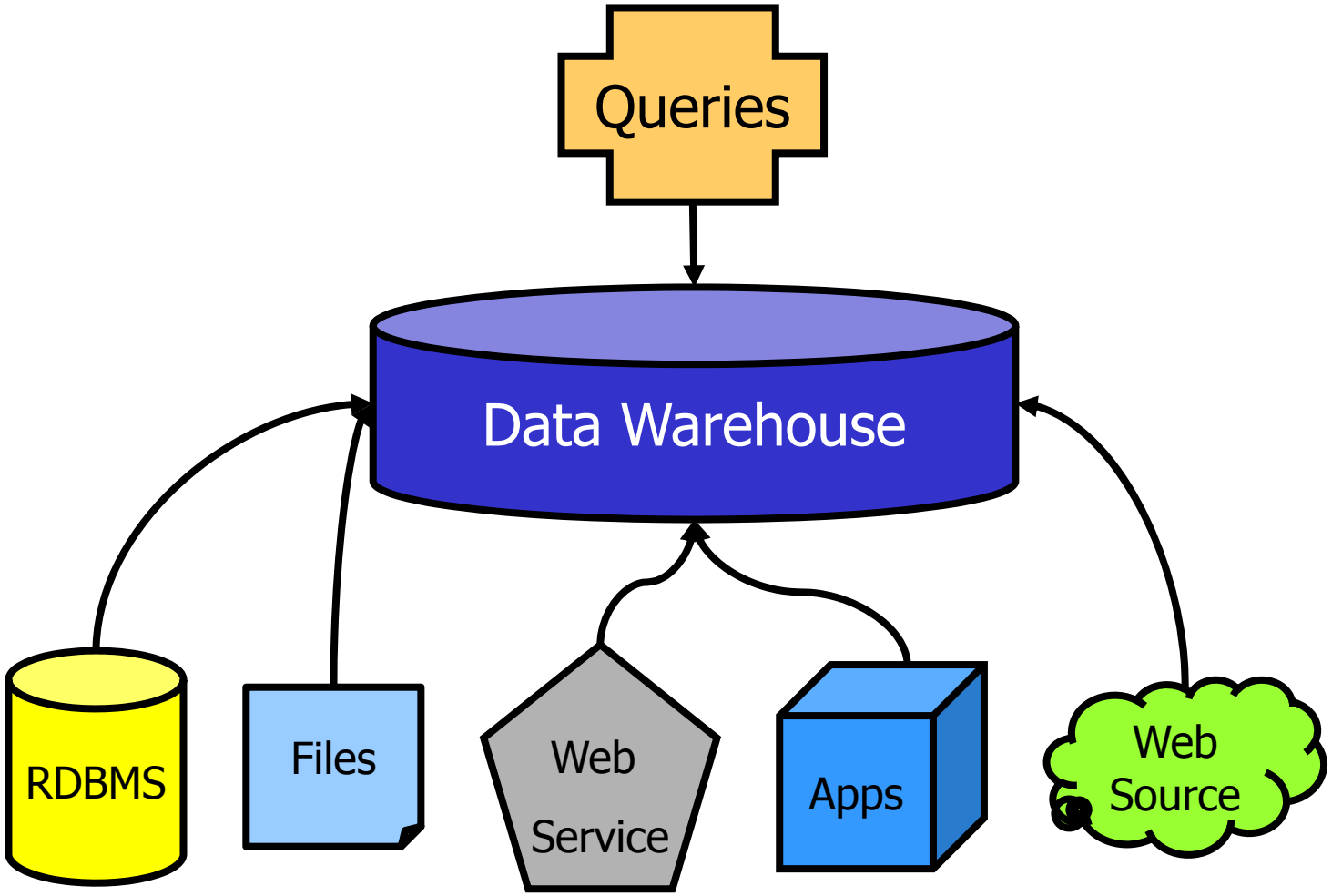
This Tutorial

- Part I – Data Integration for the Life Sciences
- Part II – Past and Presence
- Part III – Current Trends
  - Data Integration Workflows
  - Semantic Web
  - Ranking in Integrated Datasets
- Part IV – Conclusions

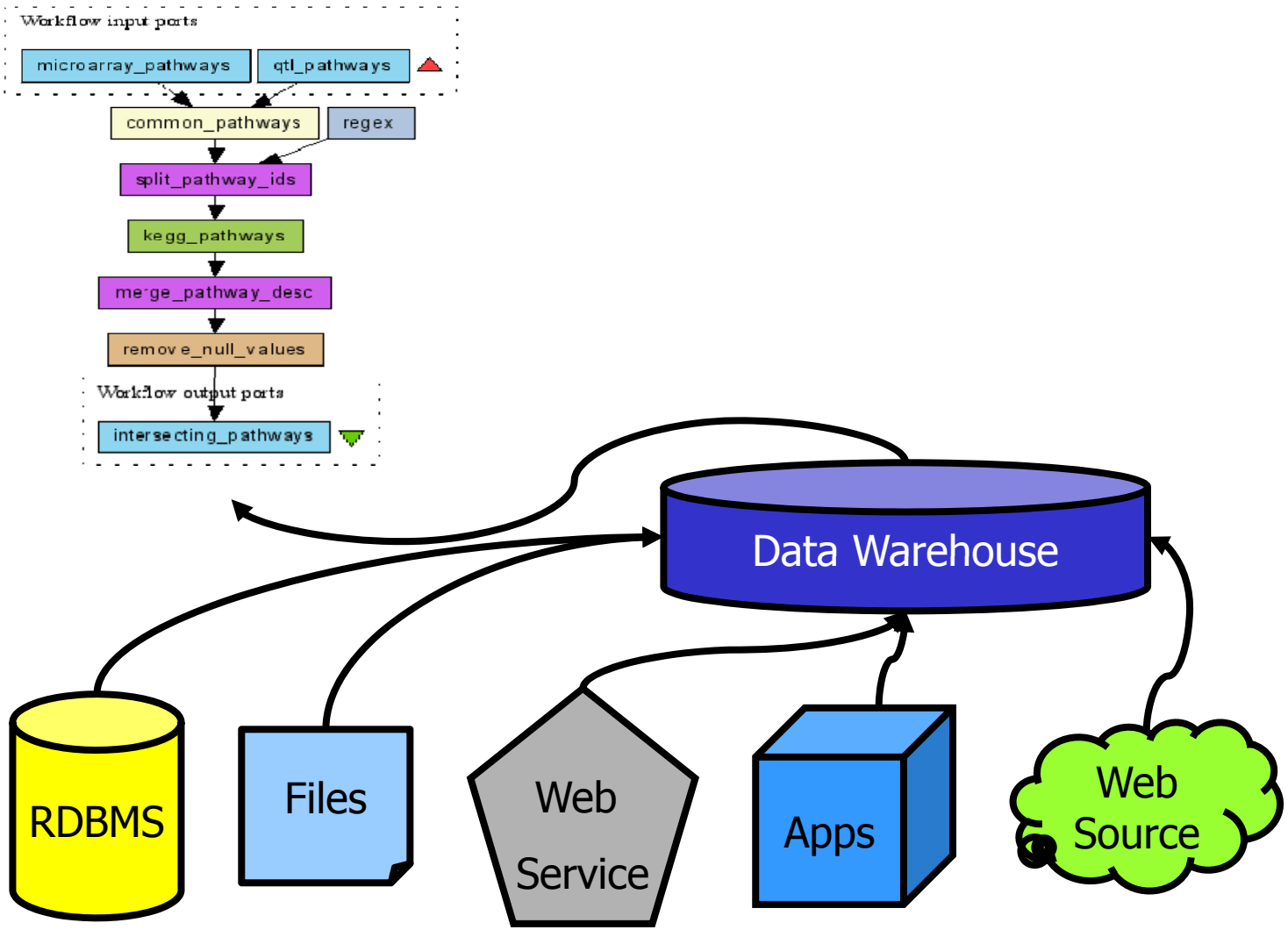
Classical View



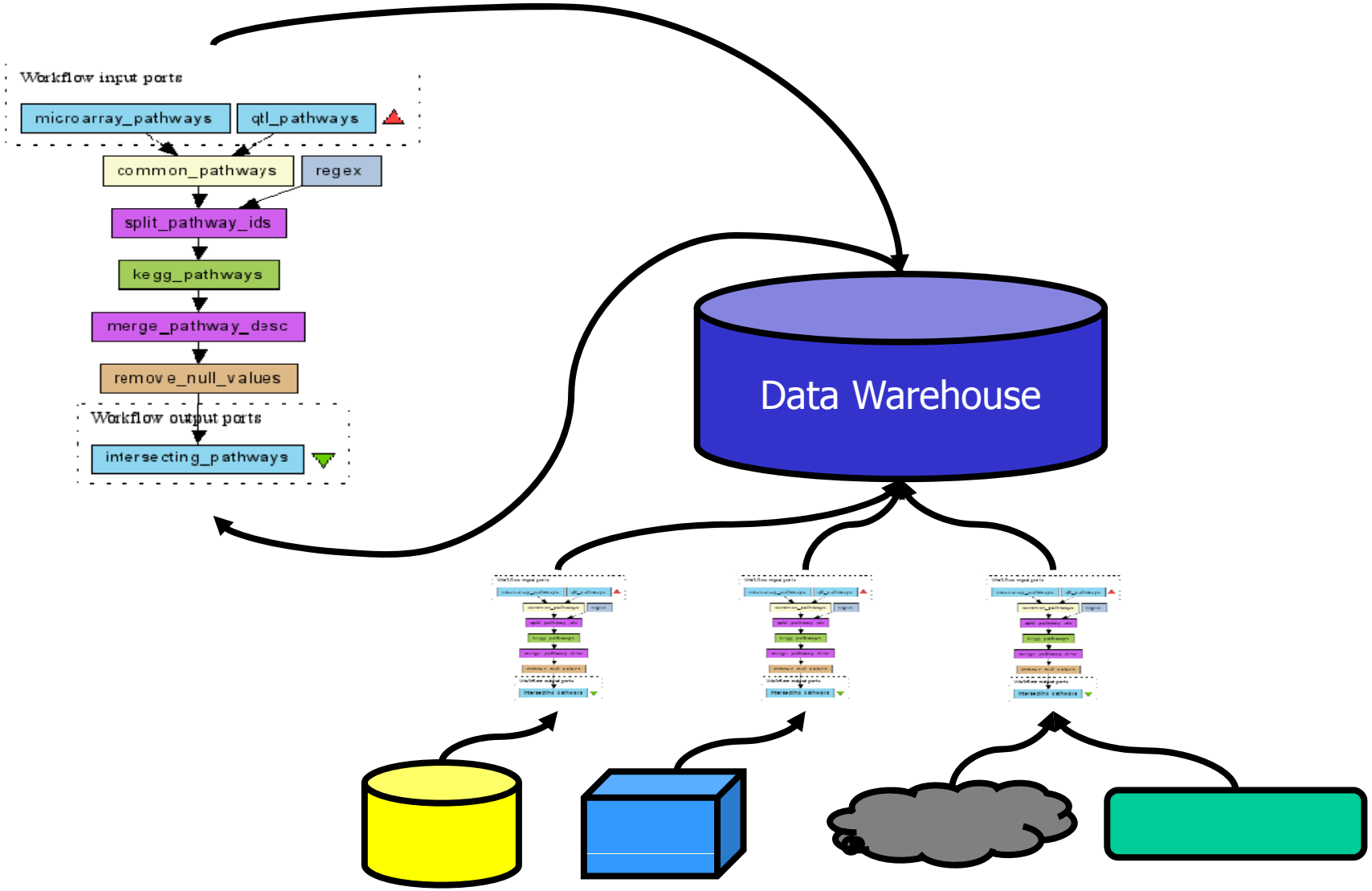
Classical View - DWH



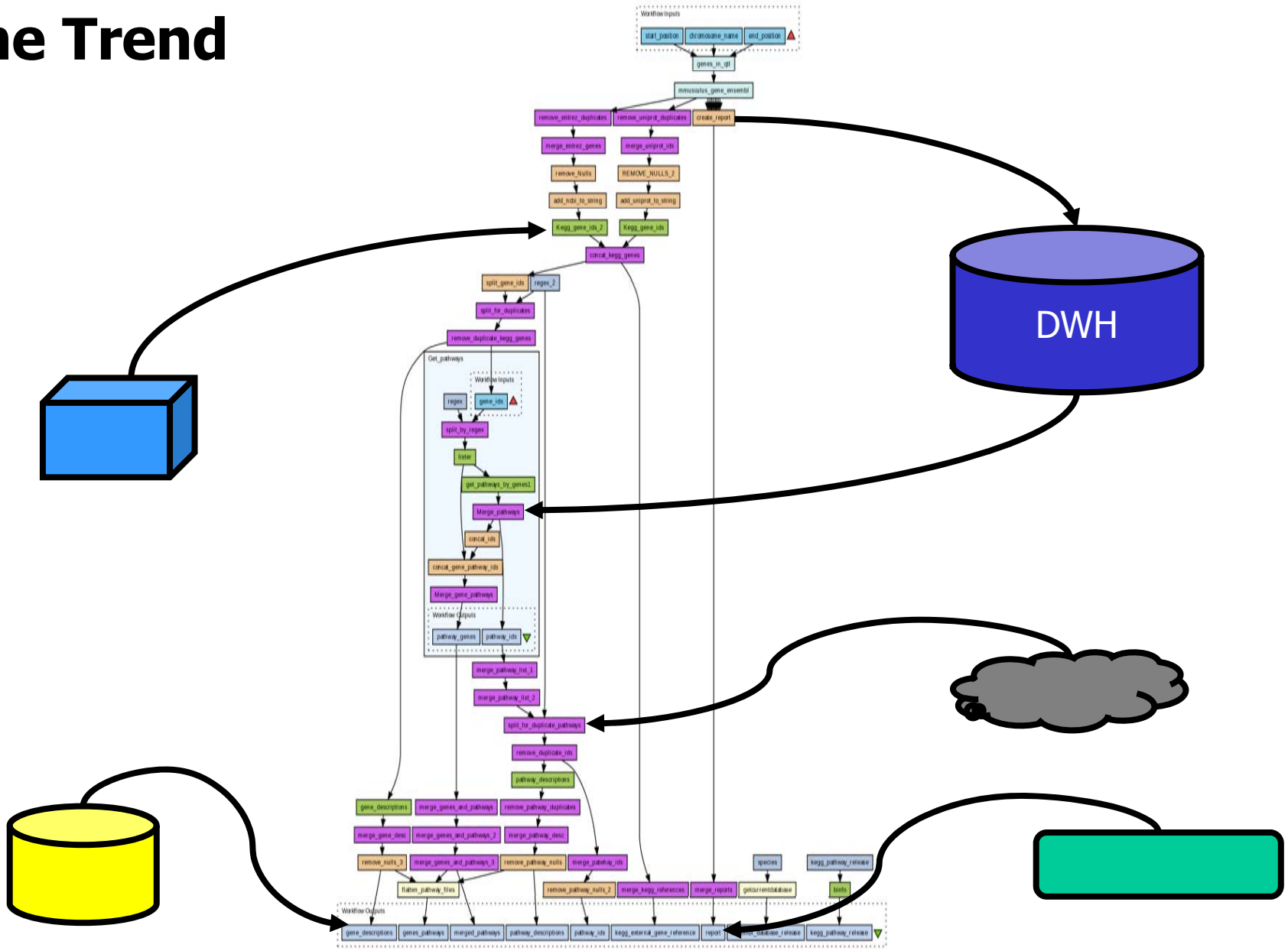
Classical View – Expanded



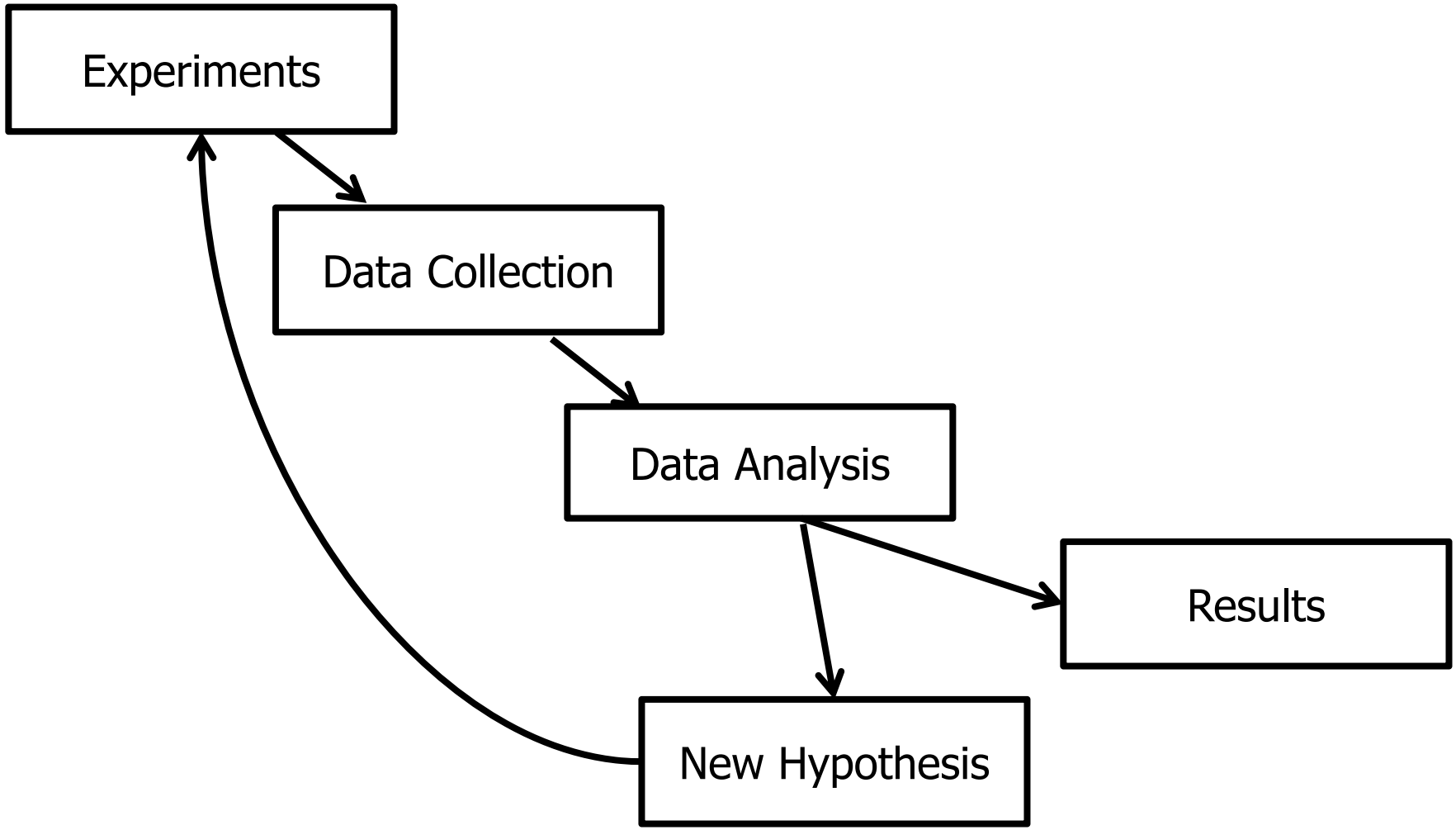
True Architectures



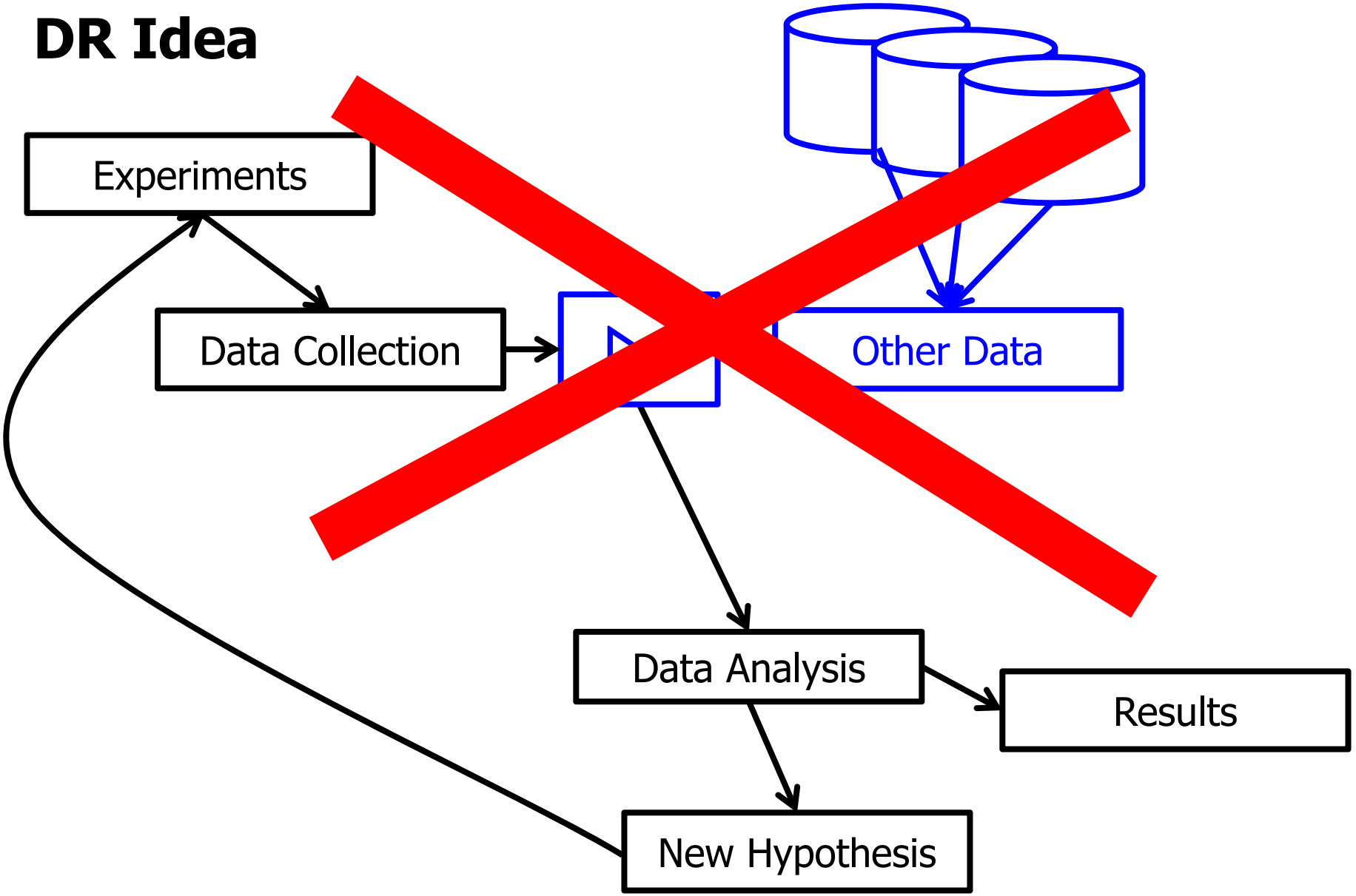
The Trend



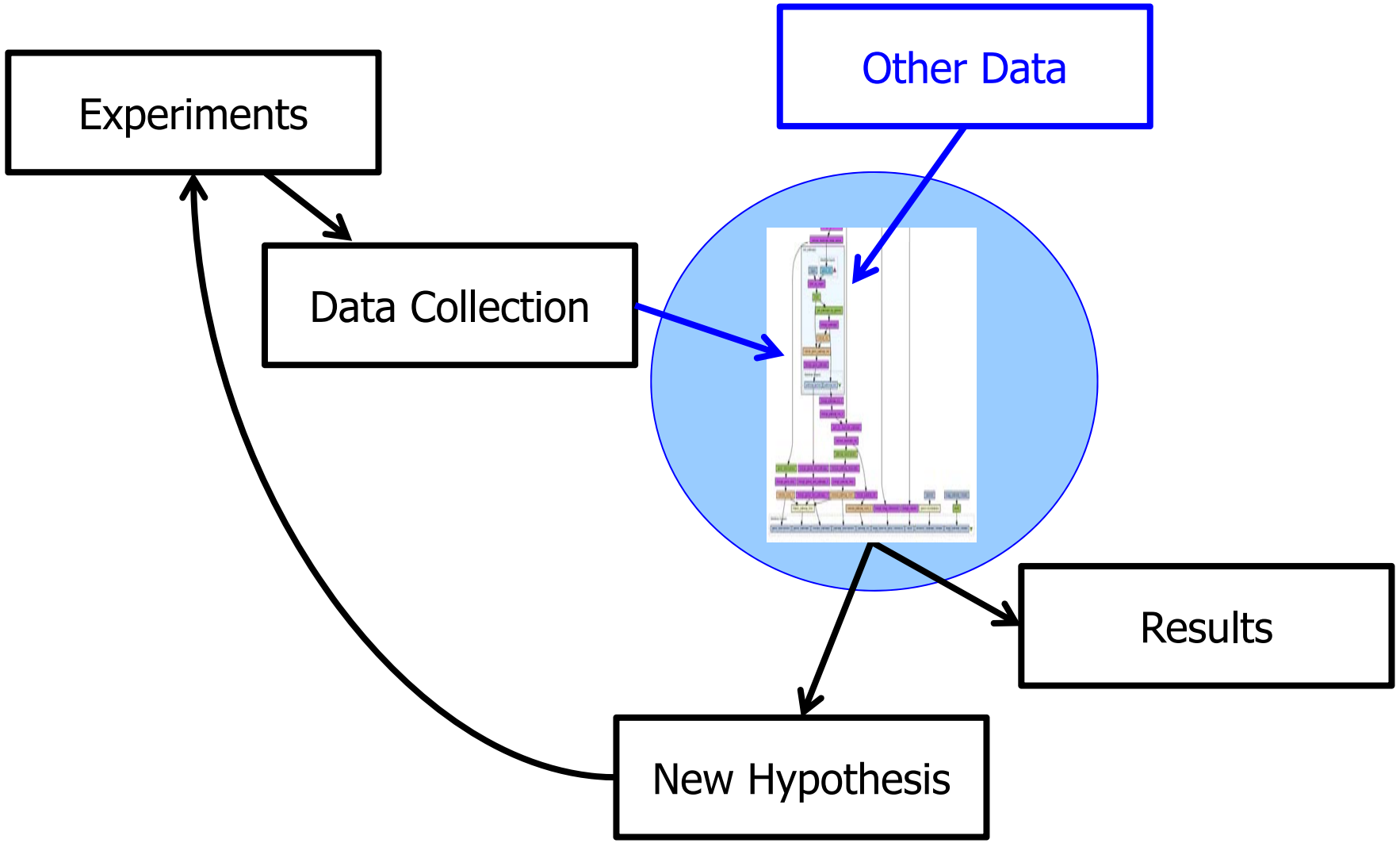
Life Science Research Food Chain



DR Idea


Experiments
Data Collection
Other Data
Data Analysis
Results
New Hypothesis

With DI Workflows

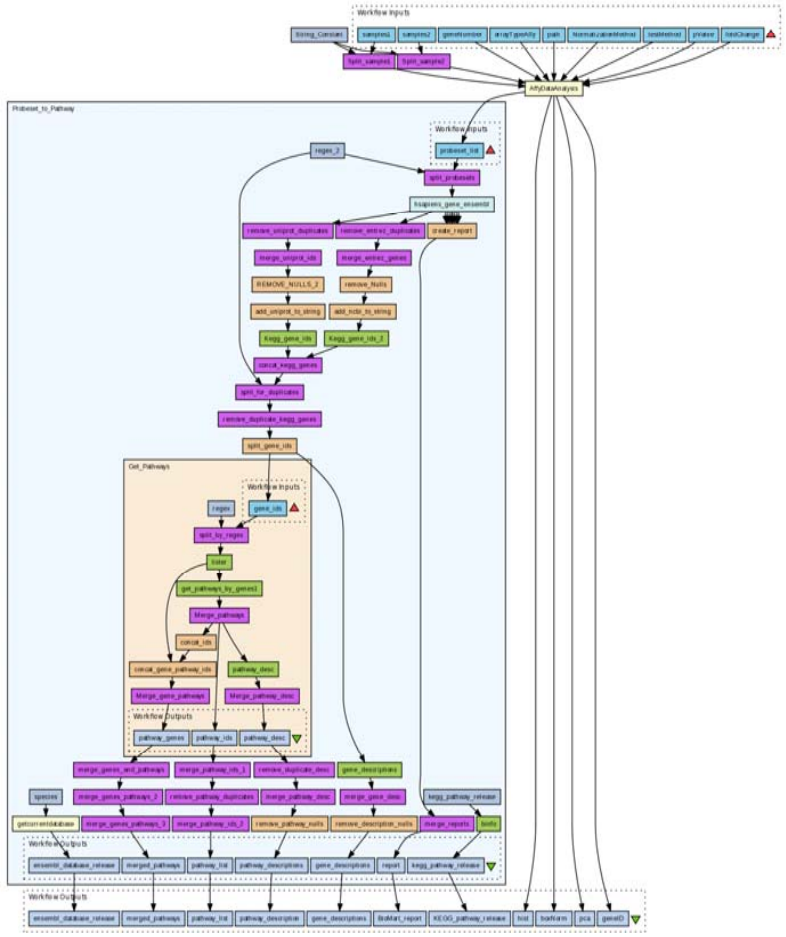


Data Integration Workflows

- No separation between integration and analysis
- Scientific Workflow Management System (SWFS) to encode integration and analysis process
- Integrated (cleansed) data sets can be a by-product
- Tasks and sub-workflows may be shared across workflows
  - Data access, parser and filter, data normalization, tests, …
- Uses materialization and virtual access –whatever is best

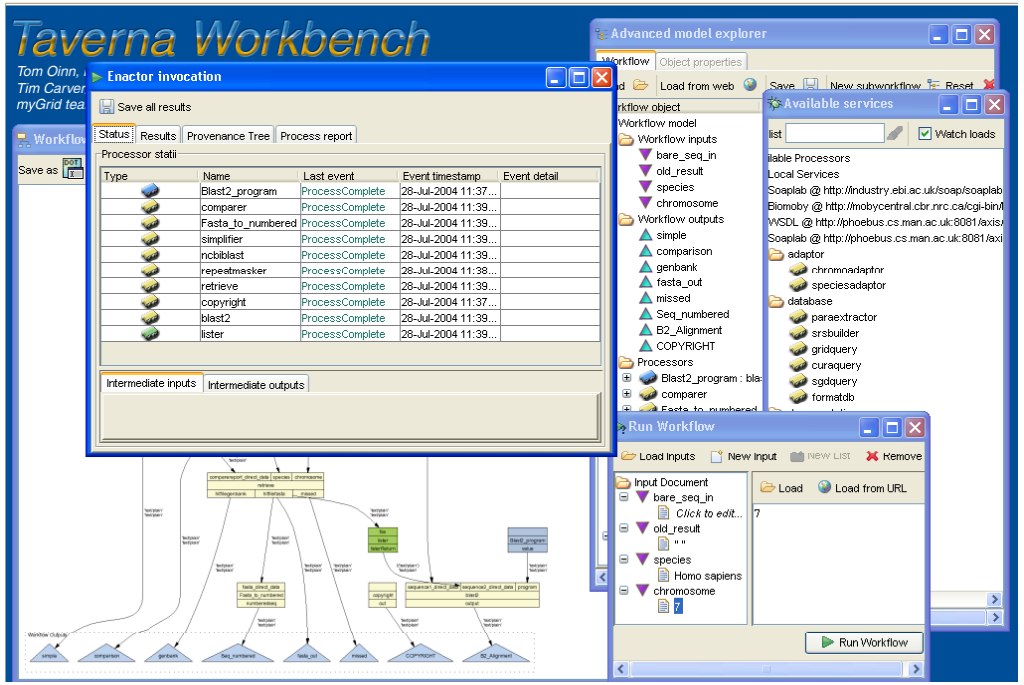
Scientific Workflow Management System

- SWFS = WFS for scientific tasks
  - "Data analysis pipeline"
  - Complex pipelines are broken into tasks and their connection
  - Data flow driven
- Tasks can be executed locally or distributed (web services)
- SWFS manages scheduling, process control, logging, recovery, reproducibility, …
- Often equipped with graphical workflow designer
- Several systems available (Taverna, Kepler, Triana, …)



Example: Taverna

- SWFS developed at U Manchester for ~10 years (myGrid)
- Full fledged, production-level system
- Integrates hundreds of bioinformatics resources and services
  - Ontology-based service lookup
- SCUFL: Simple Conceptual Unified Flow Language
- Hundreds of users, some reports on real projects



[MSR+10]

Data Integration Workflows

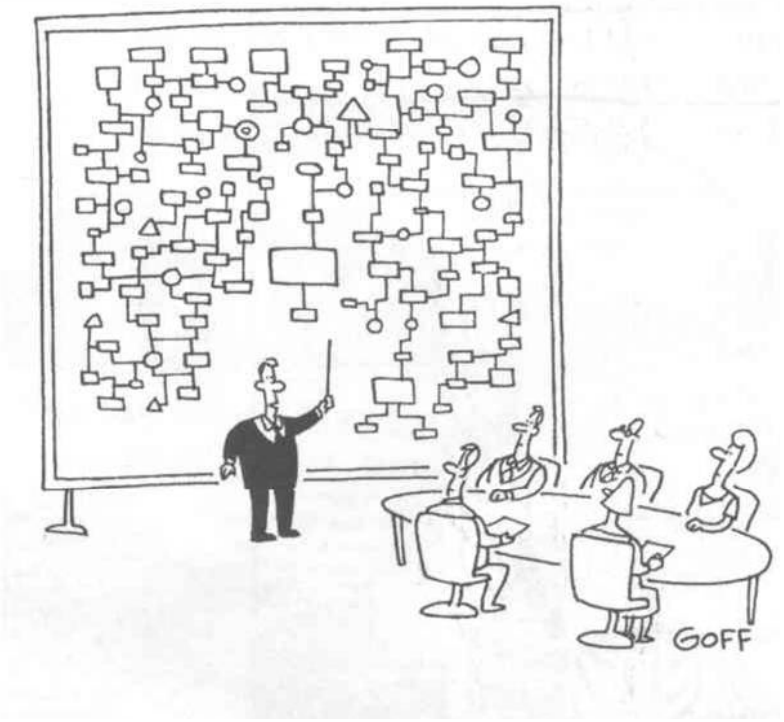
- There is no clear separation between a (scientific) data integration workflow and an ordinary scientific workflow
- Scientific workflows embody integration tasks
  - Data access (remote or local)
  - Parsing
  - Transformation (values, structure)
  - Filtering (selection, projection)
  - Discrete merging (union, join, difference)
  - Statistical aggregation (mean, median, testing, …)
  - User-defined predicates
  - …
- SWFS treat integration tasks the same as any other tasks

Problems Tackeled

- Get away from "once for ever" idea of classical information integration
- Complex integration processes become first-class citizen
  - Exposes what is done rather than hiding it
  - Data quality issues can (and must) be considered (selection, filtering, …)
- Produces results that are immediately interesting for the researcher
  - No queries
- Requires deep understanding of the domain
  - Integration is only one ingredient to the solution

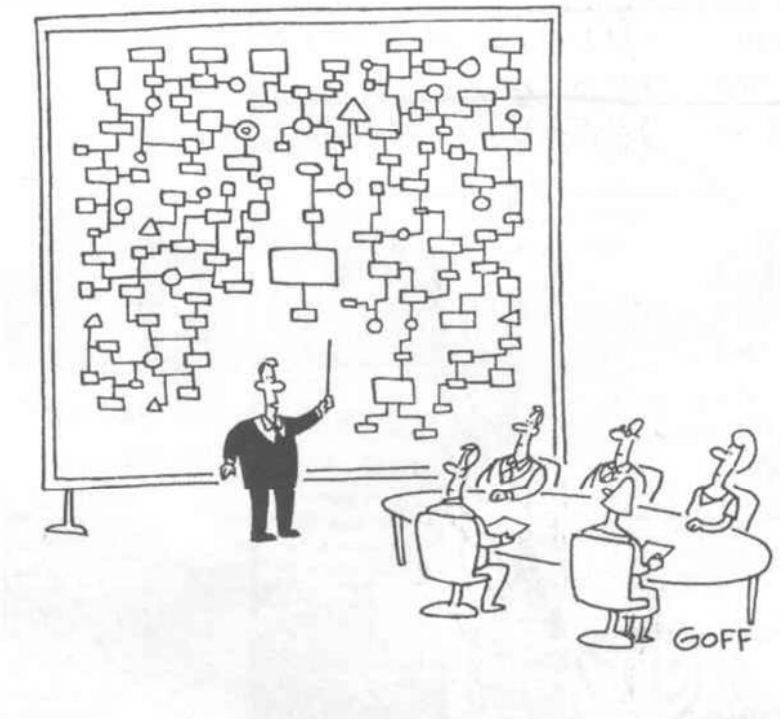
But …

- What do we save compared to Perl?
  - No support for semantic integration
  - Potentially, everything must be programmed anew every time
  - Workflows are not easier to read than Perl programs



But …

- What do we save compared to Perl?
  - No support for semantic integration
  - Potentially, everything must be programmed anew every time
  - Workflows are not easier to read than Perl programs



- But Perl doesn't do
  - Automatic logging of all steps
    - Reproducibility, credibility
  - Automatic scheduling on available hardware
  - Automatic restart in case of failure
  - …

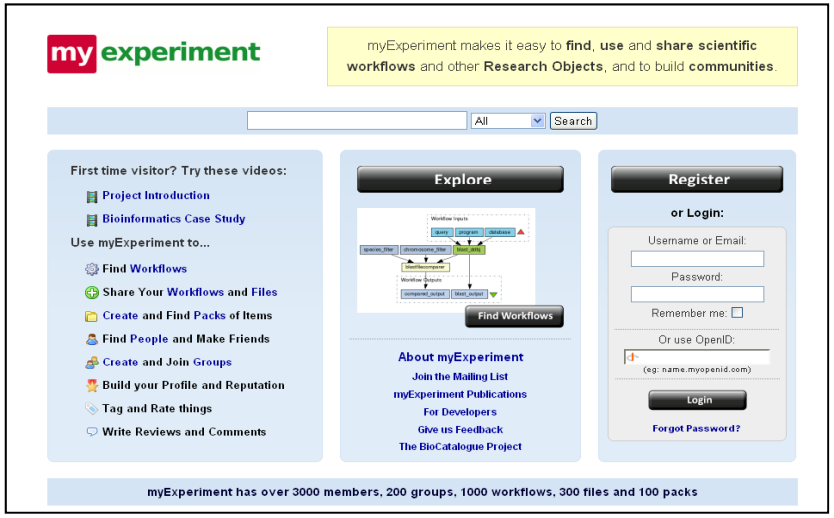
Less Obvious Advantage: Sharing (Sub-)Workflows

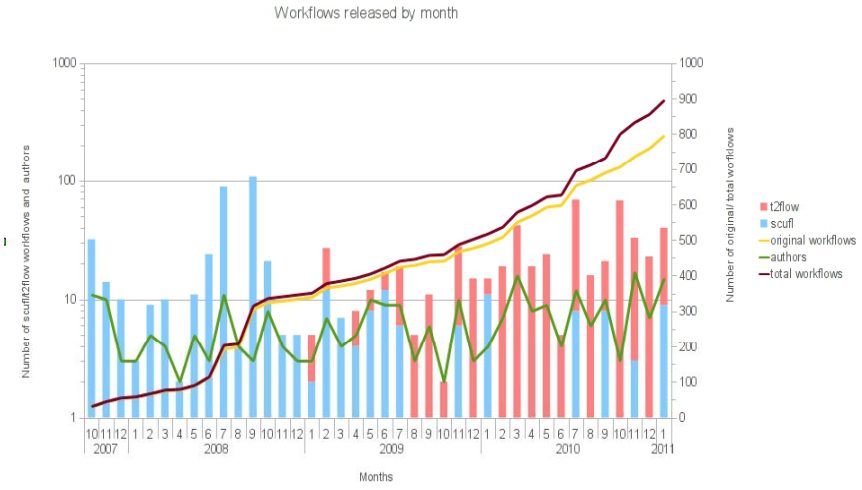
- Existing tasks and sub-workflows are available in SWFS repositories
- These can be searched, downloaded, and reused

- Sharing tasks
  - Generic parser is shared, specific filter is developed
- Sharing sub-workflows
  - Performing some complex processes producing a defined result
  - Partly relief from the infamous "shims and glue" trap
- Parameterization increases reusability
  - Which filtering / selection?
  - Where is the data source / service to use?

myexperiment

- > 1300 workflows available for immediate download
- Cross-system: Taverna, Triana, Kepler
- Social functionality: Tagging, rating, usage statistics
- Reuse features could be improved





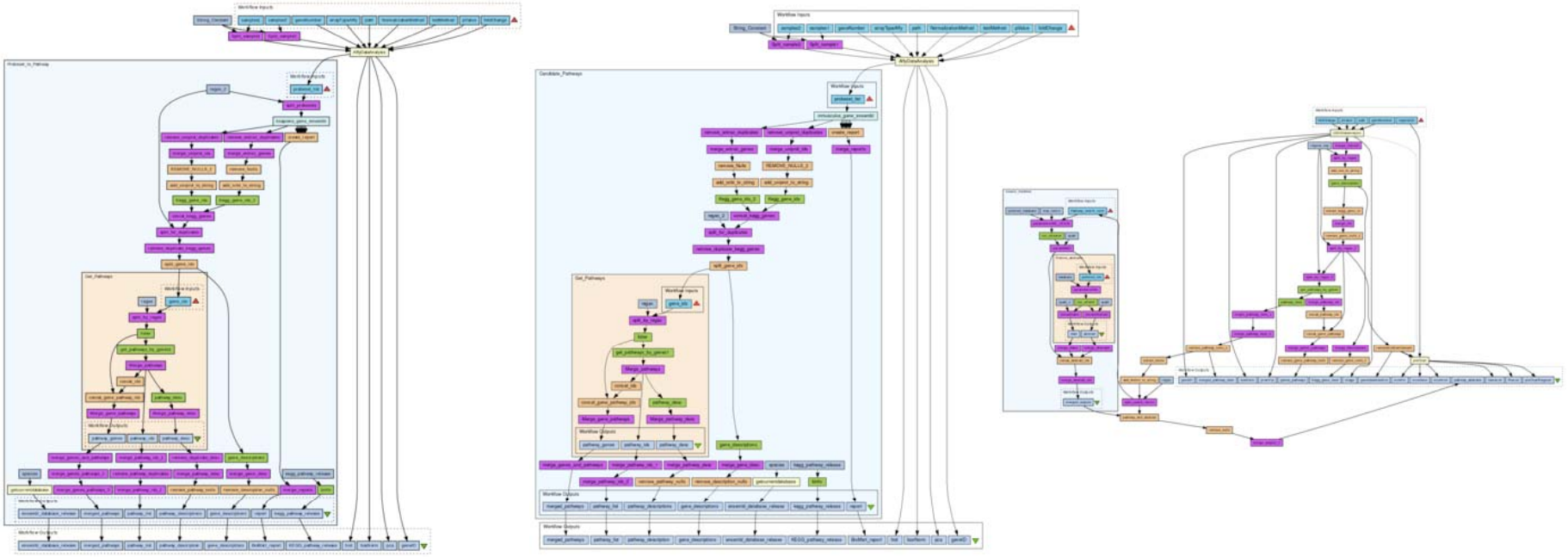
Opportunities (and Untackeled Problems)

1. Improve support for workflow sharing
   - Beyond searching (missing) documentation
2. Supporting typical integration tasks
   - Reducing amount of repeated work
3. Distributed data access
   - Data to service or service to data?
4. Adaptive execution environments
   - Many tasks require special environments – not portable per-se

1. Finding the Right Workflow

- Currently only IR-style queries on metadata / documentation
- Open question: Querying workflow repositories
  - Given a high-level description of a (integration) task – a sketch
  - Given a input and/or and output format/type
  - Given a workflow
  - …
  - Find workflows (global similarity) or sub-workflows (local similarity)
- Core of the problem: Workflow similarity
  - Metadata similarity, topological similarity, semantic similarity
- Becomes a practical topic only now: Large repositories are available
- Complication: Search across workflow models

Example



- All three workflows perform microarray analysis integrating various sources (pathway DB, probe mapping, PubMed)
- May be re-used entirely (which fits best?) or partly (from probes to genes? Differently expressed genes? From DE to pathways?)

Work on Workflow Queries

- Using only topological properties [GLG06]
  - Ignoring WF metadata and task descriptions
- Topological similarity in serial-parallel graphs [ZCBD+09]
  - Captures a large class of workflow graphs
  - Can be solved in polynomial time
- Query languages from the business workflow community
  - BPQL, BPMN-Q [AS10], BP-QL [BEKM08], …
  - Do not include notion of similarity not local (sub-workflow) matches
  - Bound to workflow specification languages (BPEL)
- Query languages for repositories of workflow runs [KSB10, MPB10]
  - Querying the log of a workflow execution to find, e.g., the lineage of a specific result /trace
- Queries for filtering workflow runs [BCB+08]
  - Definition of views to filter relevant from irrelevant

2. Supporting DI Tasks

- Integration tasks are typically data-intensive and time-consuming
- Especially during WF development, such tasks need to be executed again and again
- Storing and reusing intermediate results can be of high benefit
  - Transparent materialization and reuse (caching)
- Open problem: Savepoints in SWFS
  - How to define (language, graphical)?
  - Who places them into a WF (manual or automatic)?
  - Mapping of results to workflow steps?
  - Efficiently storing and reusing the data

- Note: Results depend on concrete data, workflows do not
- Note: Storing input together with results also enhances reproducibility

Work on "Smart Recomputation"

- Caching
- Strong Links [KSB+10]
  - Mapping of files using signature of "upstream" workflow
  - Support for post-WF analysis (which runs used this file?)
- Smart re-computation [LAF+06]
  - Moves responsibility to the file system
  - Requires tight integration with SWFS

- Also see "Managing Scientific Data", CACM 2010, [AKD10]

To the Extreme: Global Analysis Repository

- Savepoint data could even be exchanged globally
  - Analysis on a particular data set is performed once and then re-used all over the world
- This has predecessors – storing processed and raw data
  - Sequence database: DNA sequences and trace files
  - Proteomics: Identified proteins and 2D-Page Gels
  - Transcriptomics: CEL files and CCD images

- Repositories would have to store runs, data sets, results, and intermediate results

3. Distributed Data Access

- Intensive usage of web services, though attractive from a reuse point-of-view, has a cost
  - Data must be shipped back-and-forth
  - Reliability of entire WF decreases with every additional service
- In the LS, data files typically are not terribly large, but analysis requires many steps
- Open question: Reducing round-trip cost
  - Services must become "location-aware"
  - Data could be passed by reference (if servers are used for many tasks)
  - Code could be moved (many tasks are R modules anyway)
  - Decision should be based on estimations about runtime
  - Alternative: Data transfer as first-class citizen
    - Allows users to influence behavior in optimal manner

4. Execution Environement

- If tasks are not executed remotely, they need to run locally
- But tasks may require certain software to be pre-installed
  - Programming languages, runtime libraries, infrastructure services, …
- Open question: Make SWFS infrastructure-aware
  - Problem is well studied in operating systems / middleware
    - Linux package loader, …
  - Tasks need to specify dependencies
  - Needs a "module concept" for downloading and installing missing pieces
  - SWFS need to communicate with operating system
  - Essentially, one needs OGSi for SWFMS

- Low hanging fruit with potential for large impact

Towards Ultimate Credibility

- A major advantage of SWFS is reproducibility (hence: credibility)
- Publish workflow together with its results (and input data)
- Everybody can reproduce analysis
  - If input is available – in the correct versions
  - If workflows runs an machine
  - If all services are available in the correct versions

- Stronger: Also publish WF traces and intermediate results
  - Every step of the analysis becomes visible and can be checked
- Calls (again) for global repositories of entire analysis's

A Vision: The Executable Paper

- "The Executable Paper Grand Challenge" (Elsevier)
  - How can we develop a model for executable files that is compatible with the user's operating system and architecture and adaptable to future systems?
  - How do we manage very large file sizes?
  - How do we validate data and code, and decrease the reviewer's workload?
  - How to support registering and tracking of actions taken on the executable paper?
  - [http://www.executablepapers.com/]

- The other way round: Make pipelines citable
  - Get credits for your pipeline, not for your paper

DI Workflows and ETL

- DI workflows are similar to ETL processes
- But there are important differences [Alb09]
  - ETL mostly consists of relational operators, SWFS mostly uses user-defined predicates
  - ETL mostly runs on relational data, SWFS on any data
  - ETL are proprietary and bound to companies, SWFS mostly use public data
  - ETL always runs on different data, SWFS often repeated on the same data
  - ETL are a business asset and not shared, SWFS are a scientific achievement and shared (until now, mostly by papers)

DI Workflows and ETL

- DI workflows are similar to ETL processes
- But there are important differences [Alb09]
  - ETL mostly consists of relational operators, SWFS mostly uses user-defined predicates
  - ETL mostly runs on relational data, SWFS on any data
  - ETL are proprietary and bound to companies, SWFS mostly use public data
  - ETL always runs on different data, SWFS often repeated on the same data
  - ETL are a business asset and not shared, SWFS are a scientific achievement and shared (until now, mostly by papers)

Gives sharing of SWF a better perspective than sharing in ETL/business

DI Workflows and Data Flow Languages

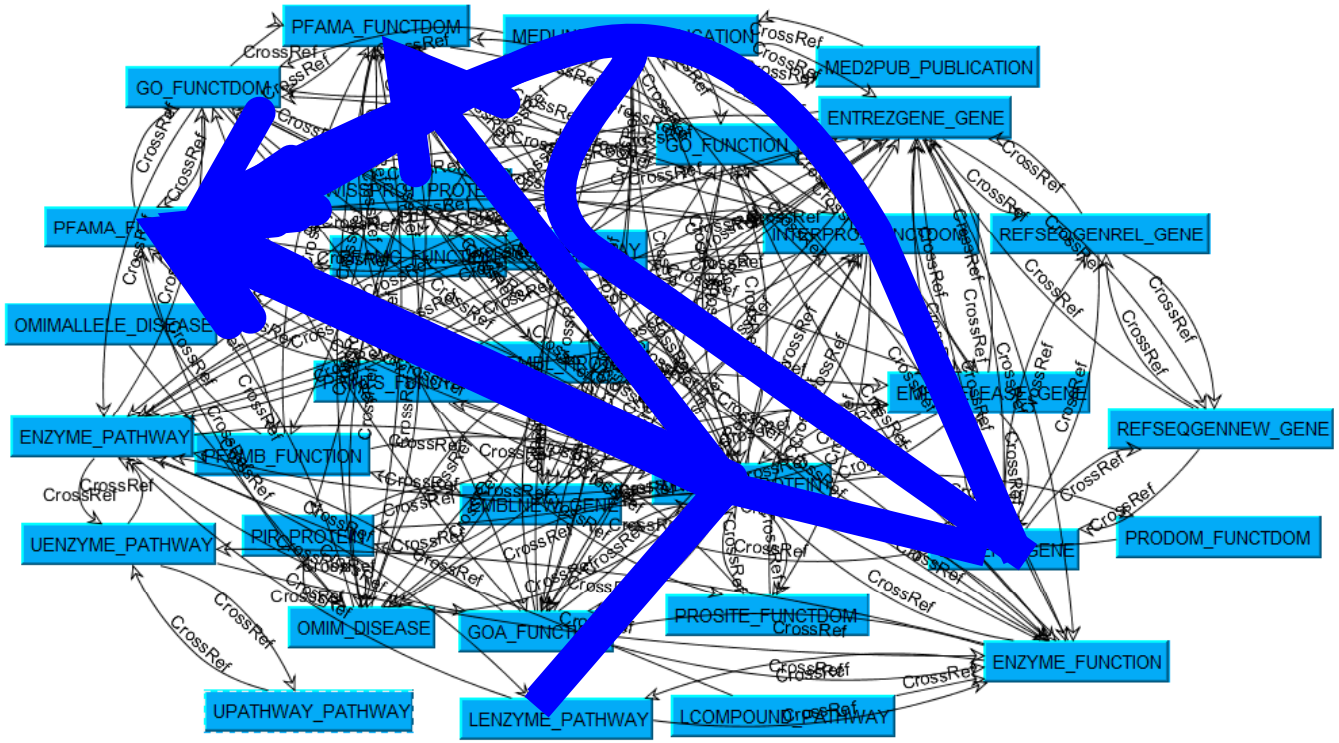
- Data flow languages recently became en vogue
  - iFuice [RTA+05], PIG-Latin [ORS+08], DryadLinq [YIF+08], …
- Invented to analyze terabytes of data
- Often focusing on scalability (parallelization, Map&Reduce)
- Typically declarative (to a certain degree) and less expressive than a typical SWFS language

- Certainly worth exploring:
  Similarity and differences between SWFS and data flow languages

Three Trends

Data Integration Workflows	• Integration means analysis, and analysis means integration • No schemas, no explicit semantics • Scientific workflow systems	Effort Analysis Provenance Quality
Ranking	• Report results in a biologically meaningful order • Stays with queries, adds ranking • Requires a DI system in place	Effort Analysis Provenance Quality
Semantic Web	• Reduce upfront cost of DI • No schemas, explicit semantics • Semantic Web tech. (RDF, SPARQL)	Effort Analysis Provenance Quality

Recall: Links



Report all GO annotation for a given protein

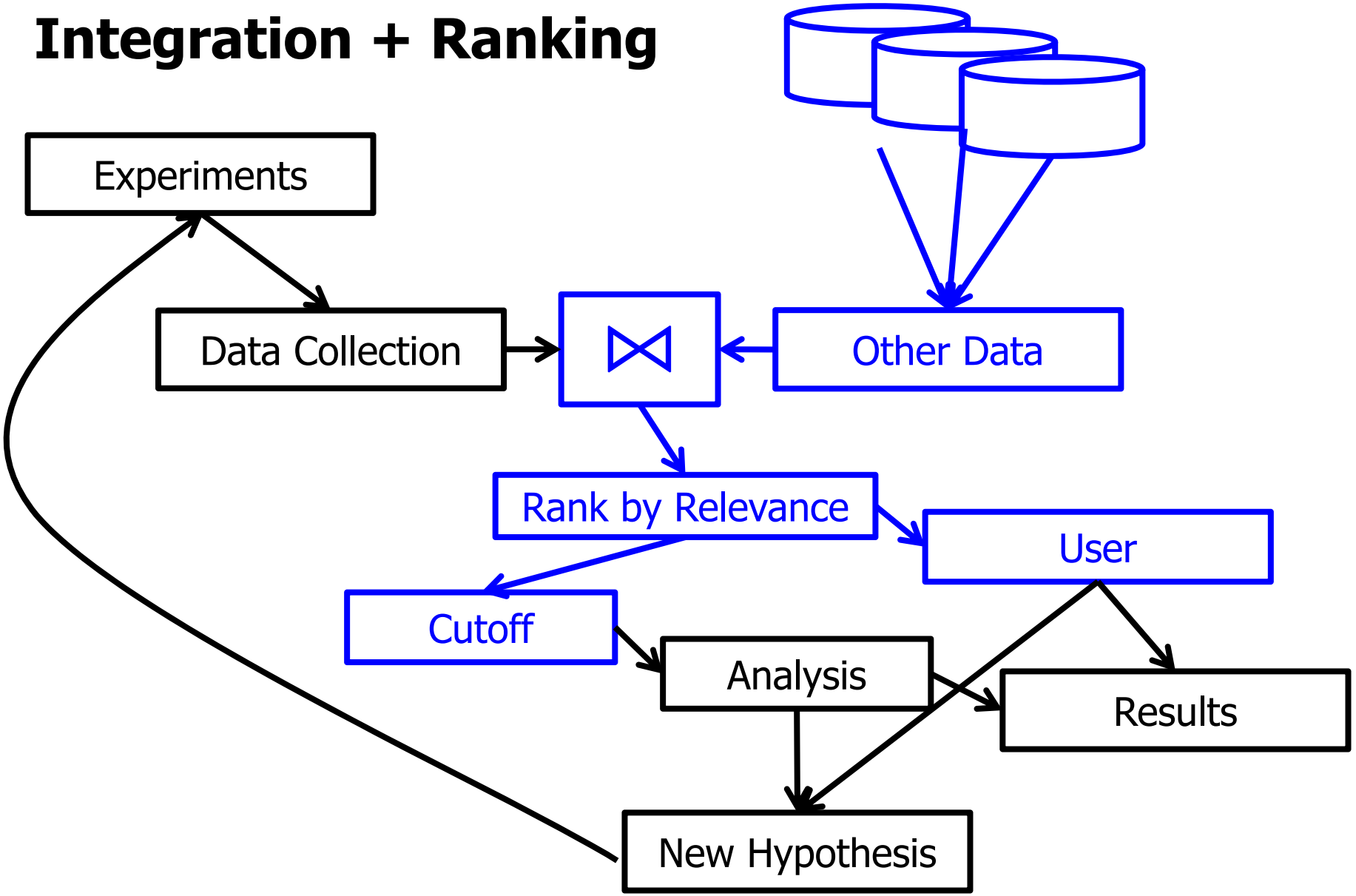
Ranking of Search Results

- Recall: Most types of objects are represented in multiple sources
- Recall: Sources link to each other (extensively yet unsystematically)

- As a consequence, for a query X->*->Y there usually exist multiple paths producing an excessive number of results
  - Which results are the best
  - Which results have the highest relevance to the query?

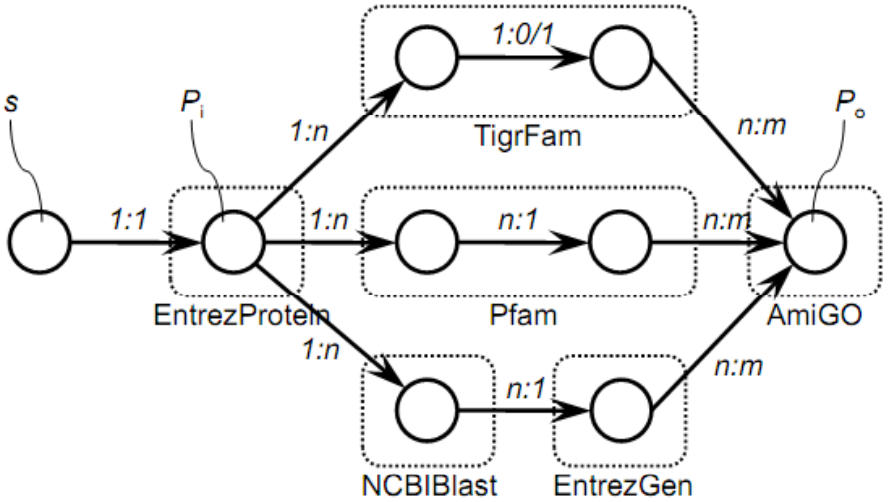
- This is not data fusion: No consensus, but present best choice(s)

Integration + Ranking



Typical Queries

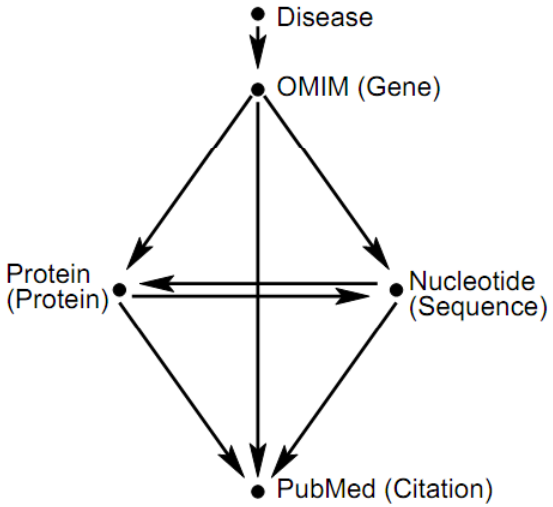
- For an object s of a source relation S, rank all objects t from a target relation T that are reachable from s
  - Through a path of joins/links
- Possibly augmented with various types of constraints
  - Attribute values of source / target / intermediate objects
  - Minimum quality of links
  - Maximal length of a path
  - …

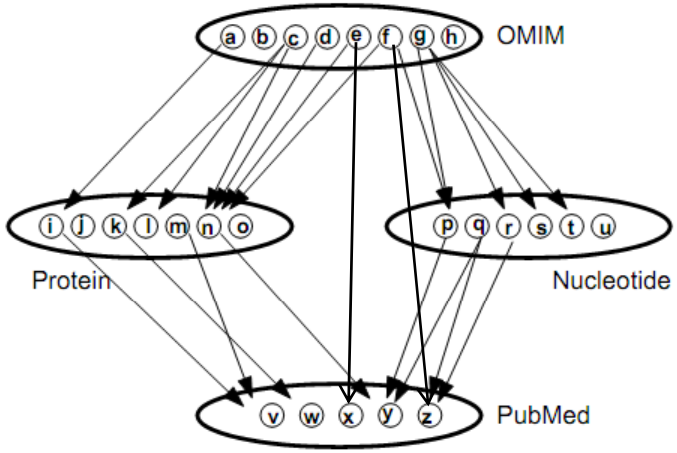


Rank all annotations from GO reachable from an entry in EntrezProtein [DGL+09]

Common Approach

- Execute query and map result into a graph
  - Compute and follow all (or some) paths
  - Collect intermediate objects on each path
  - Build data graph (objects and links)
- Compute ranks based on graph



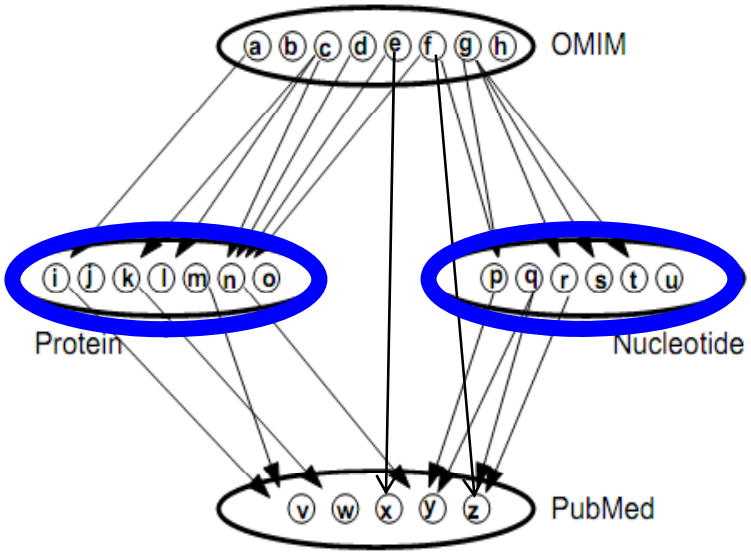


[BLM+04]

Relevant for Relevance

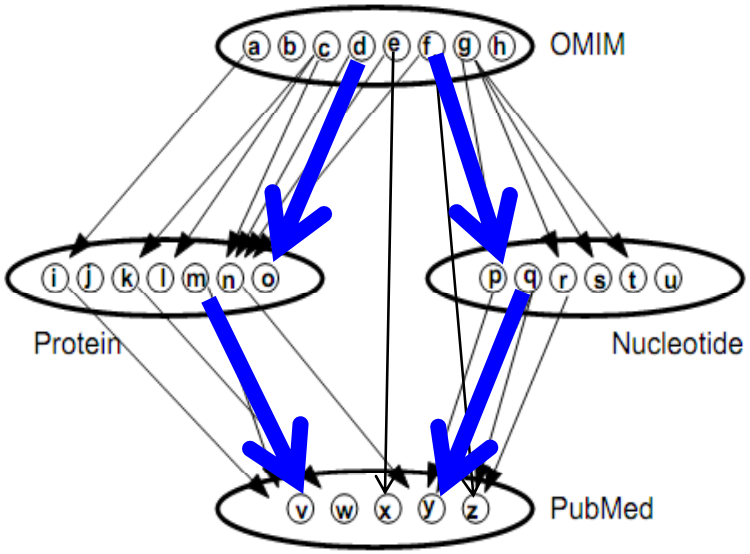
User provided	• Assessment of quality of data sources • Assessment of quality of links • Currentness, completeness, trust, …
Query dependent	• Number of paths • Length of paths • Overlap in paths
Domain specific	• Similarity of linked sequences • Quality of matching leading to a link • Many more
Graph intrinsic	• Density of the graph • Topology of the graph
Technical issues	• Execution time (joins, distributed query optimization) • Budget-based optimization • Best-effort optimization

Example



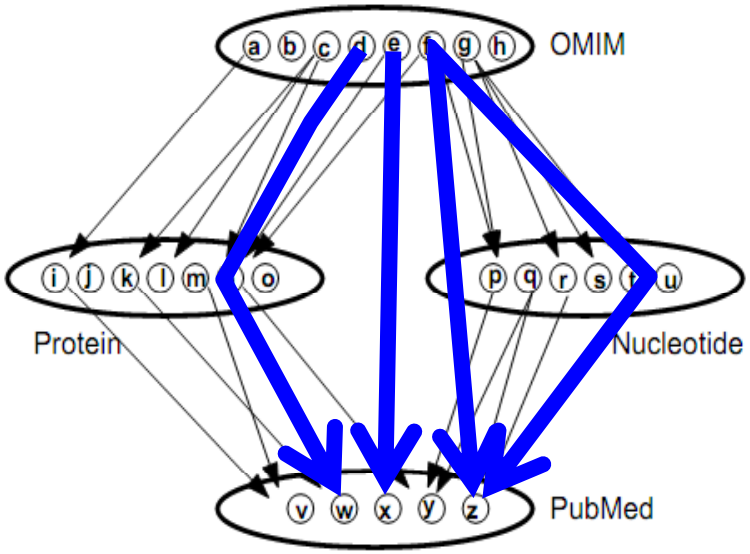
Which source is better?

Example



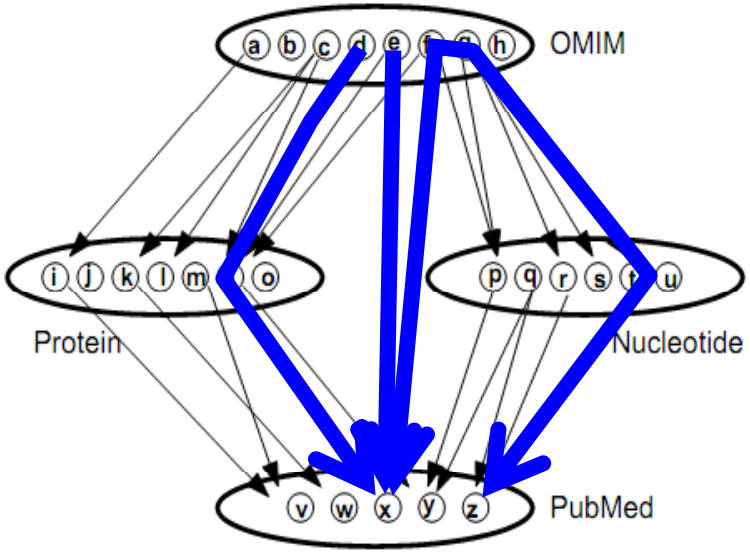
Which link is better?

Example



Which path is longer?

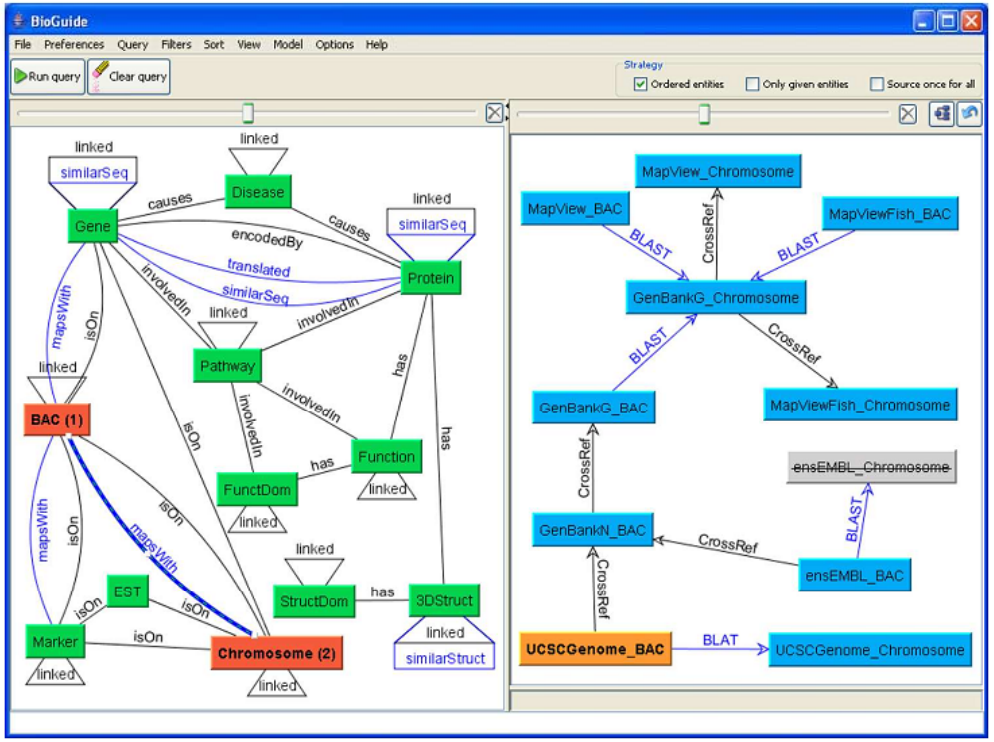
Example



Which objects are reached by more paths?

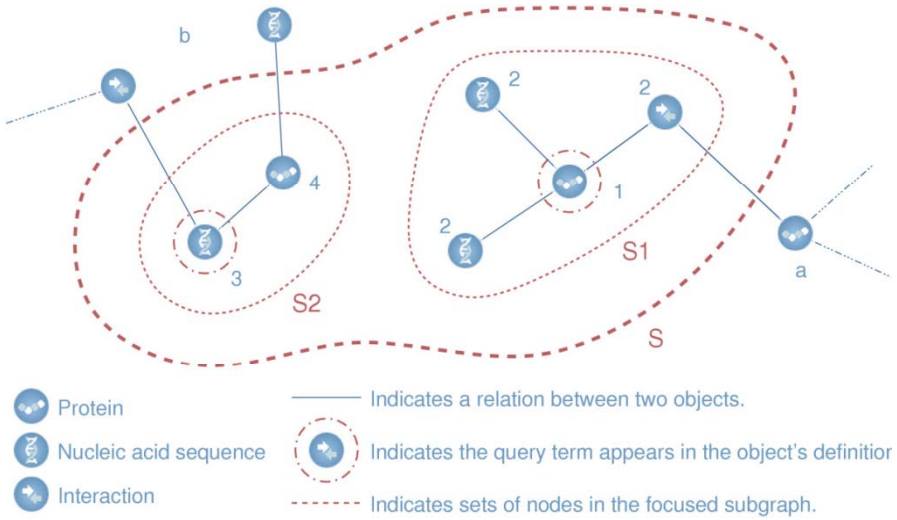
Example: BioGuide [CBD+06]

- Conceptual three-level entity model
  - Entity types (genes), source entities (EntrezGene), objects (DMD)
- Fully implemented and functioning system
- Query execution using SRS
- Ranking based on
  - User-provided assessments
  - Computed links-quality
  - Query-dependent criteria
- Various restrictions on path structure possible
  - Direction of links?
  - Paths with loops inside the entity source graph?



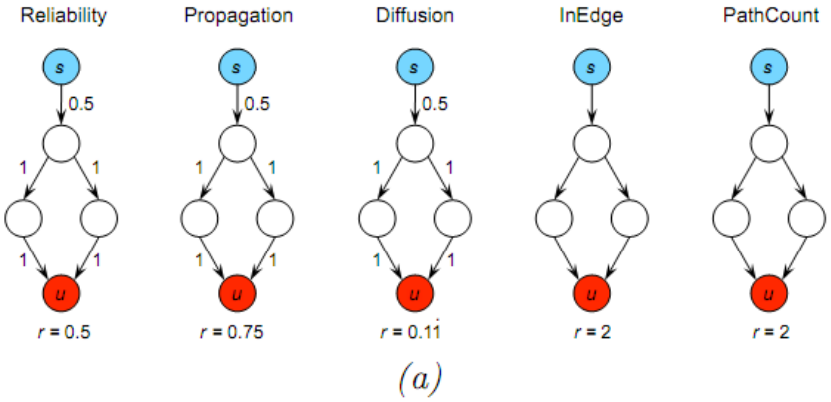
Systems: BioZon [SIY06]

- DWH of ~30 different sources (sequence, protein)
- Pre-computation of additional links
  - Sequence homology, structural similarity, …
- Four "prominence" models
  - Eigenvalue centrality
  - PageRank
  - Hubs & Authorities
  - Katz's Status
- Either computed on entire database or on query-dependent subgraph

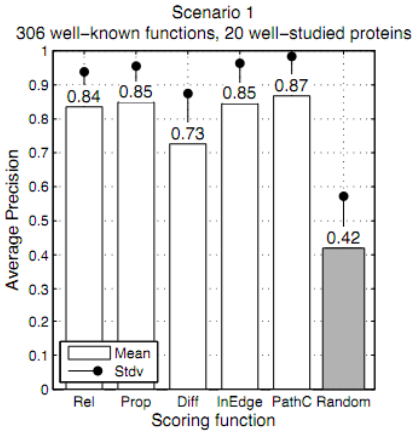


- Tendency: PageRank best, global scheme better than local ones
  - Yet yery difficult evaluation

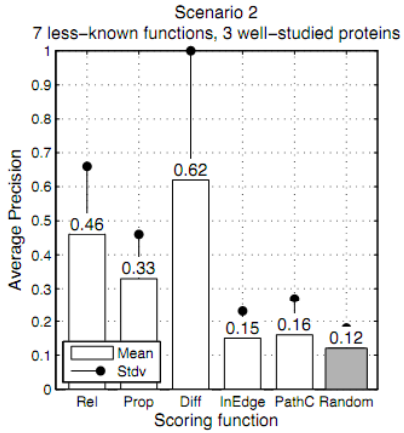
BioRank [DGL+09]



(a)



(a)

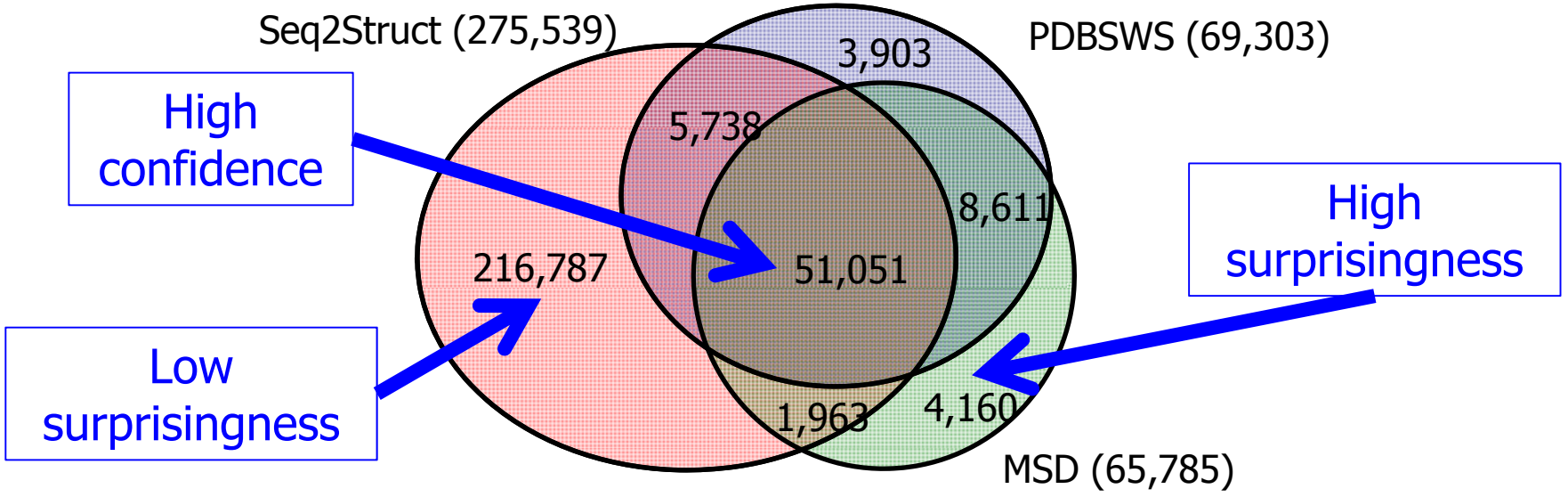


(b)

- Built upon BioMediator [SMB+04]
- Several graph-based ranking schemes
  - Network reliability (approx), diffusion, pathcount, inEdge, propagation
- Evaluation using gold standard annotation and expert opinion
- No clear results
  - Simple measures work well for well-known proteins (publication bias?)
  - Complex measures work better for less-known proteins

Columba [HTL07]

- Ranking built for the Columba DWH
- Probabilistic model to rank results by confidence or by surprisingness
  - What's new, what's certain?
  - Considering size of data sources, size of link sources, and mutual overlaps
- But: Restricted data model, does not work in general graphs



Summary

- Systems require underlying data access (integration) infrastructure
  - Can be central (BioZon, Columba) or distributed (BioGuide, BioRank)
- Systems require different degrees of human intervention
  - Fully automatic (BioZon, Columba) to human-driven (BioGuide, BioRank)
- Ranking considerations may have an influence on query execution
  - Prune paths/subplans if expected quality too low
  - Not here, but other projects, e.g. [BLM+04, NLF99]
- To date, no in-use DI system implements a decent ranking method

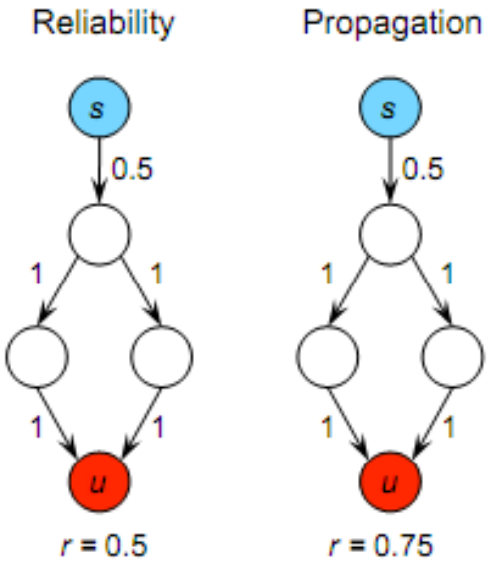
Opportunities & Challenges

1. Exploiting different semantics of links
2. Obtaining confidence scores
3. Considering incompleteness of links
4. Integrating matches with textual data
5. Comparable, objective evaluation strategies

1. Link Semantic

- Frequent interpretation: Equality (unweighted, symmetric, transitive)
- But not all links are equal
  - Similarity: Weighted, (usually) symmetric, intransitive
  - Specialization: Weighted or not, asymmetric, transitive
  - Part-of: Unweighted, asymmetric, (usually) transitive
  - Associative: ?, ?, ?
  - …
  - Context-dependent links: Sometimes true, sometimes wrong
- Propagation schemes make different assumptions
  - About link semantics
  - About independence of links
  - About semantic of network

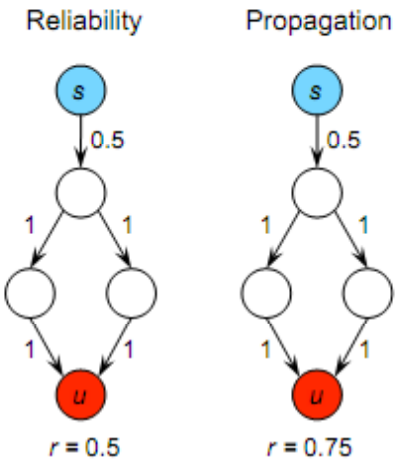
Example

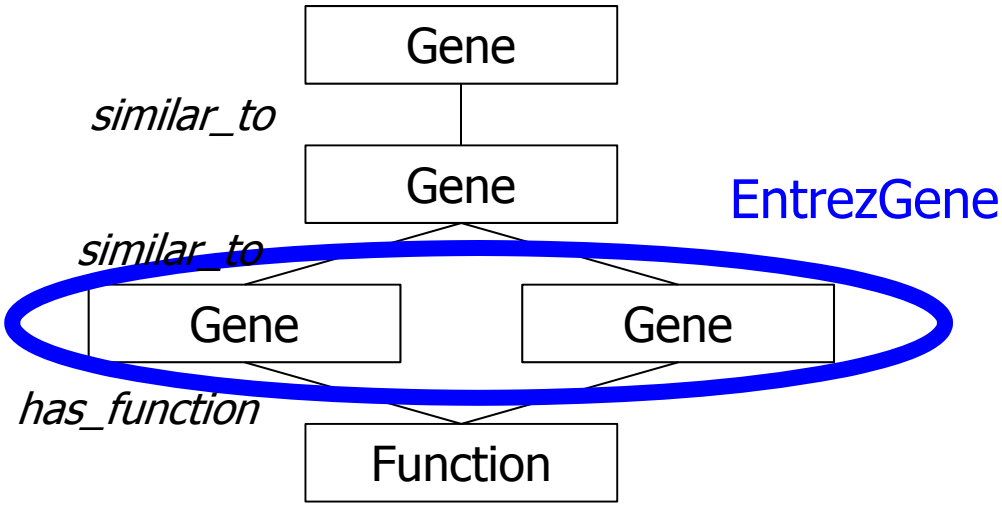


[DGL+09]

- Reliability: Probability of being equal based on ground truth
  - Probability of getting a signal from source to sink
- Propagation: Probability of being equal based on local evidences
  - Strength of evidence in a Bayesian sense

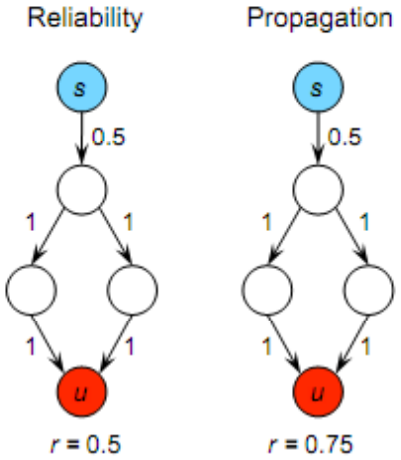
Example

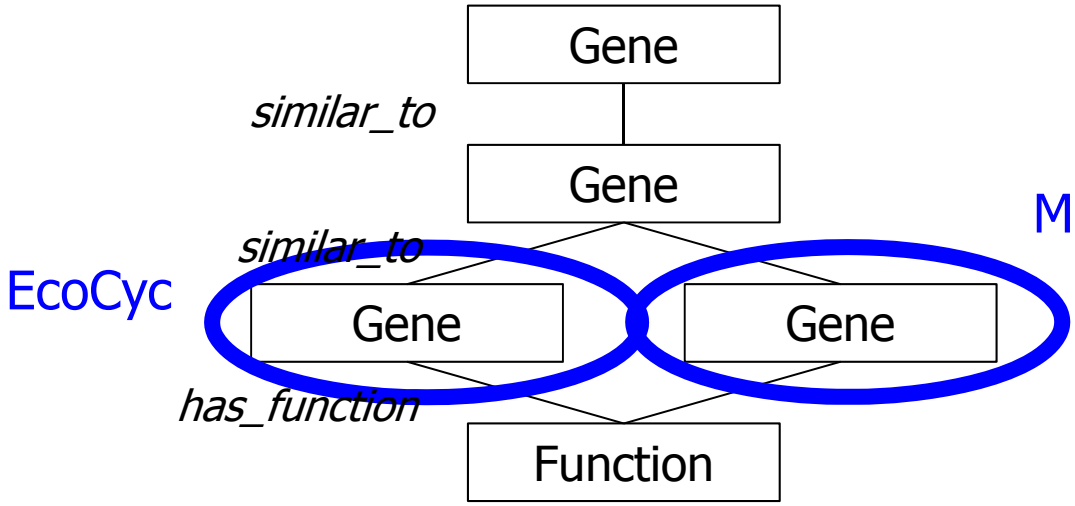




- Probably dependent evidence – little (no) increase in confidence
- Reliability is a proper model

Example

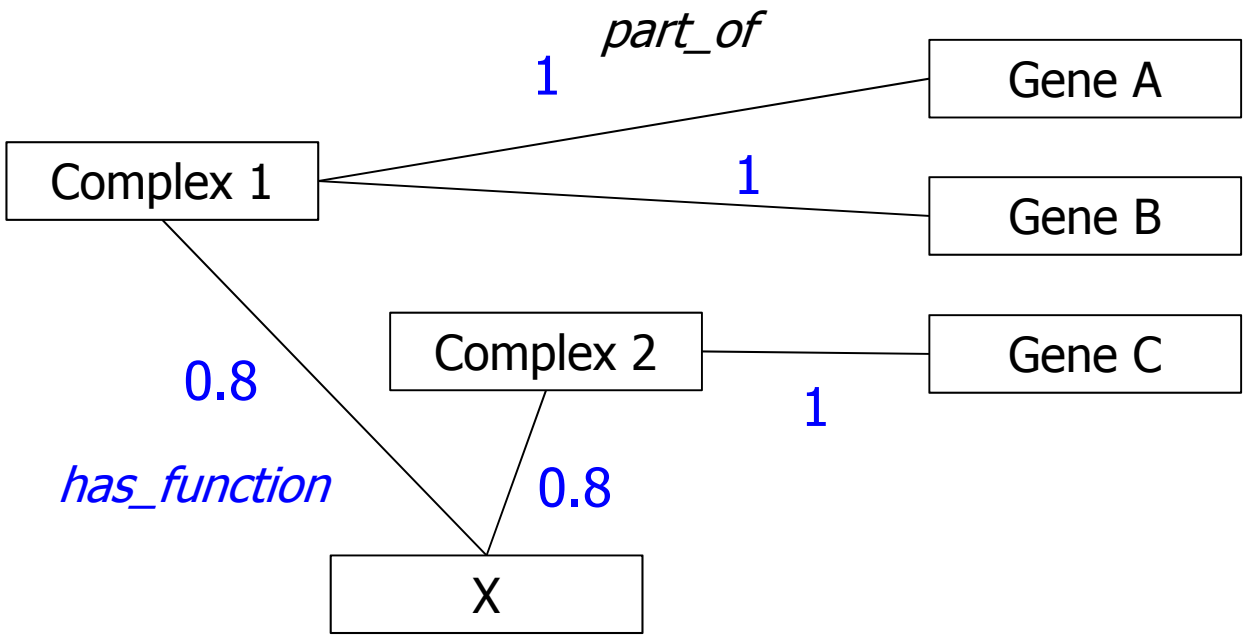




Mouse Genome Database

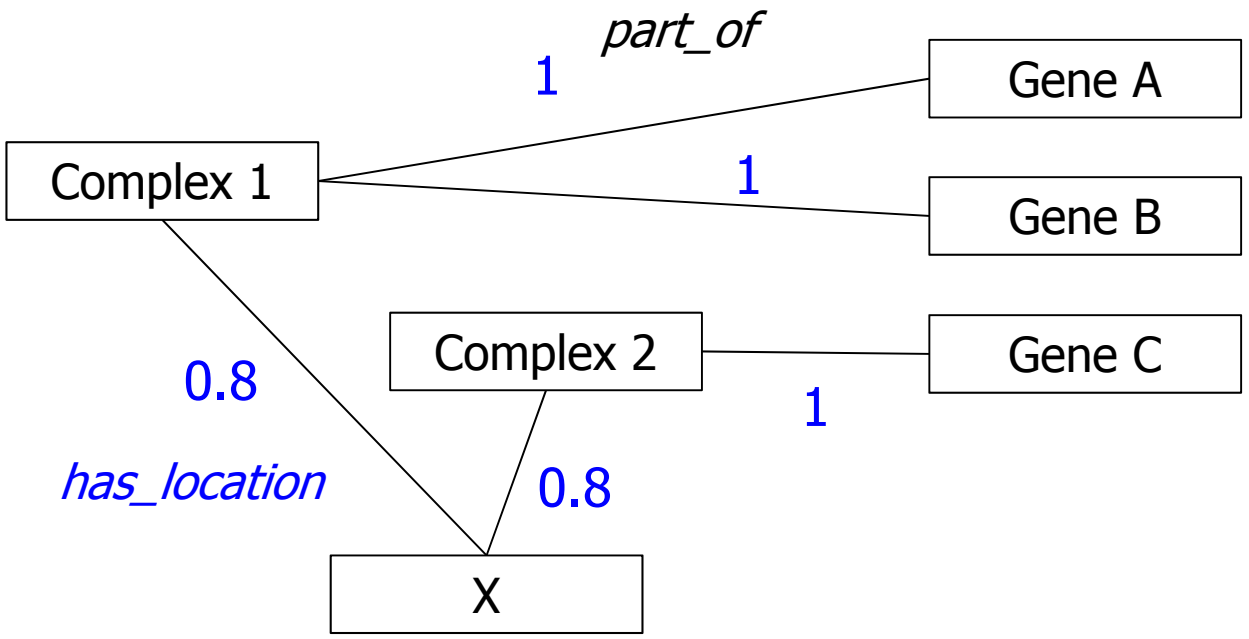
- Probably independent evidence – strong increase in confidence
- Propagation is a proper model

Second Example



- Which genes have function X?
- Ranking: C > A = B

Second Example



- Which gene are active in subcellular location X?
- Ranking: C = A = B

Opportunities

- What are appropriate classes of links?
- How can we classify links (annotation, automatic)?
- How does link type influence the interpretation of assigned weights?
- How to consider context for the interpretation of weights?
  - Gene has function X only in certain location or in certain stage of the cell
- How can ranking algorithms be aware of the semantic of links?

2. Obtaining Scores

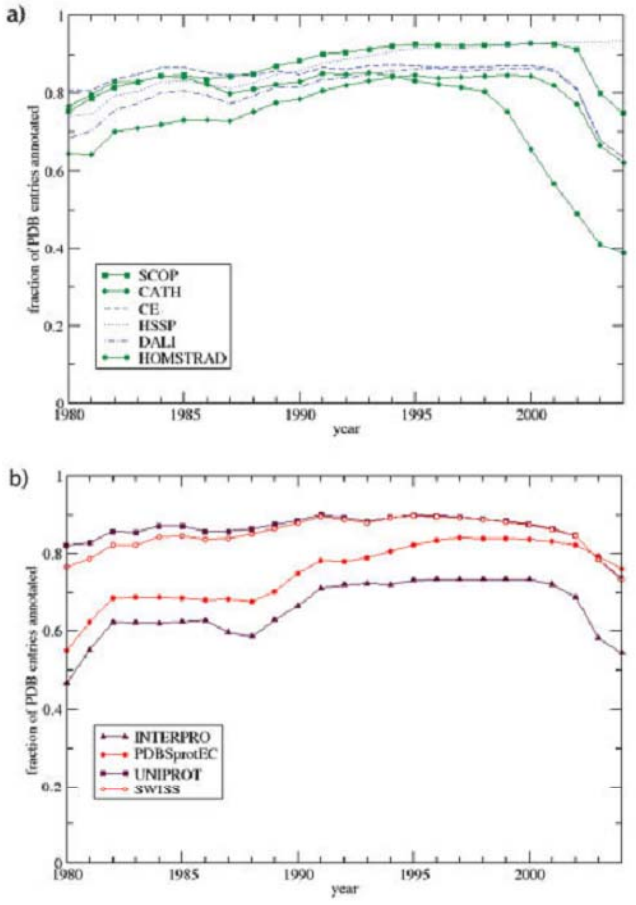
- We need (confidence, probabilistic) scores for data sources, link sources, objects, links
- Quality of biological databases is a much discussed issue, but difficult to map into a single value
- Computed scores usually cannot be used directly
  - Sequence similarity of 40% in proteins -> very likely same function
  - Sequence similarity of 40% in genes -> no statement about function
- User-defined preferences are hard to specify and to obtain

Work in this Direction

- Quality of biological databases [BCF+07, BBF+01, MNF03]
  - Often completeness / currentness
  - Measuring „degree of truth" is notoriously difficult – different experiments, different results
- Quality criteria / user preferences [NLF99, BFL+04]
- Learning user preferences from relevance feedback [TJM+08]
  - Based on BioGuide system
- Robustness of ranking [DGL+09]
  - With respect to small derivations in preference scores
- Interactive search processes are under-researched

3. Link Bias

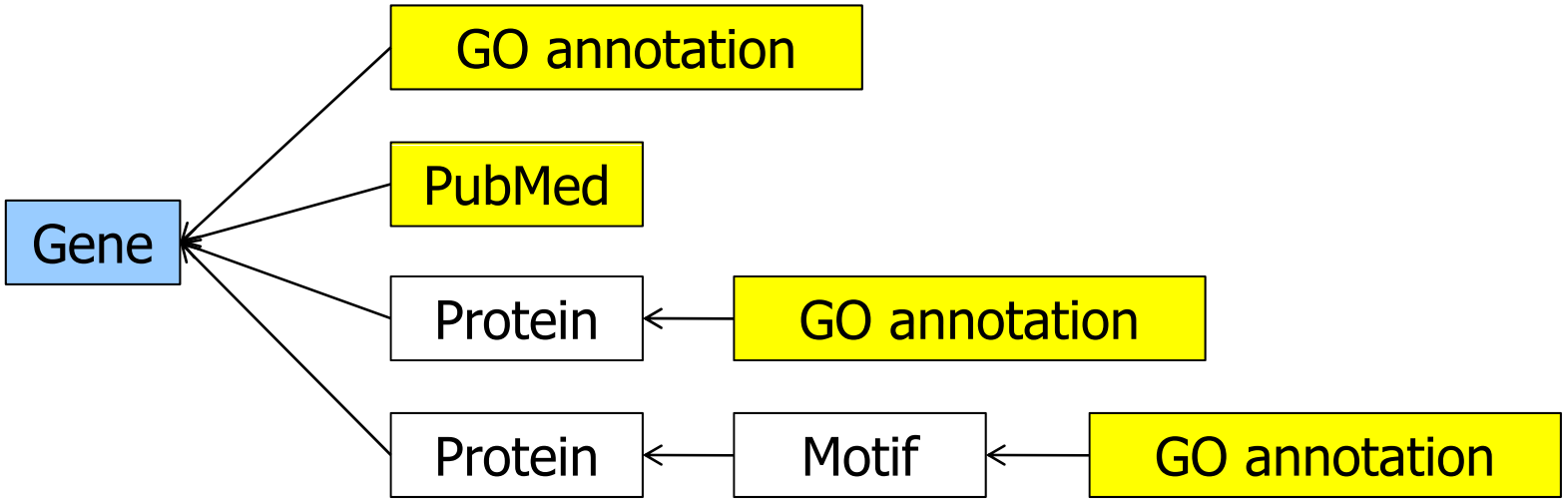
- Link sets are incomplete
- Incompleteness is not a random process
  - Popular objects receive more research – more curation – more links
  - Objects discovered more recently have less links
  - Highly linked objects are found more often – are linked more often
- Opportunity: Consider this fact for ranking



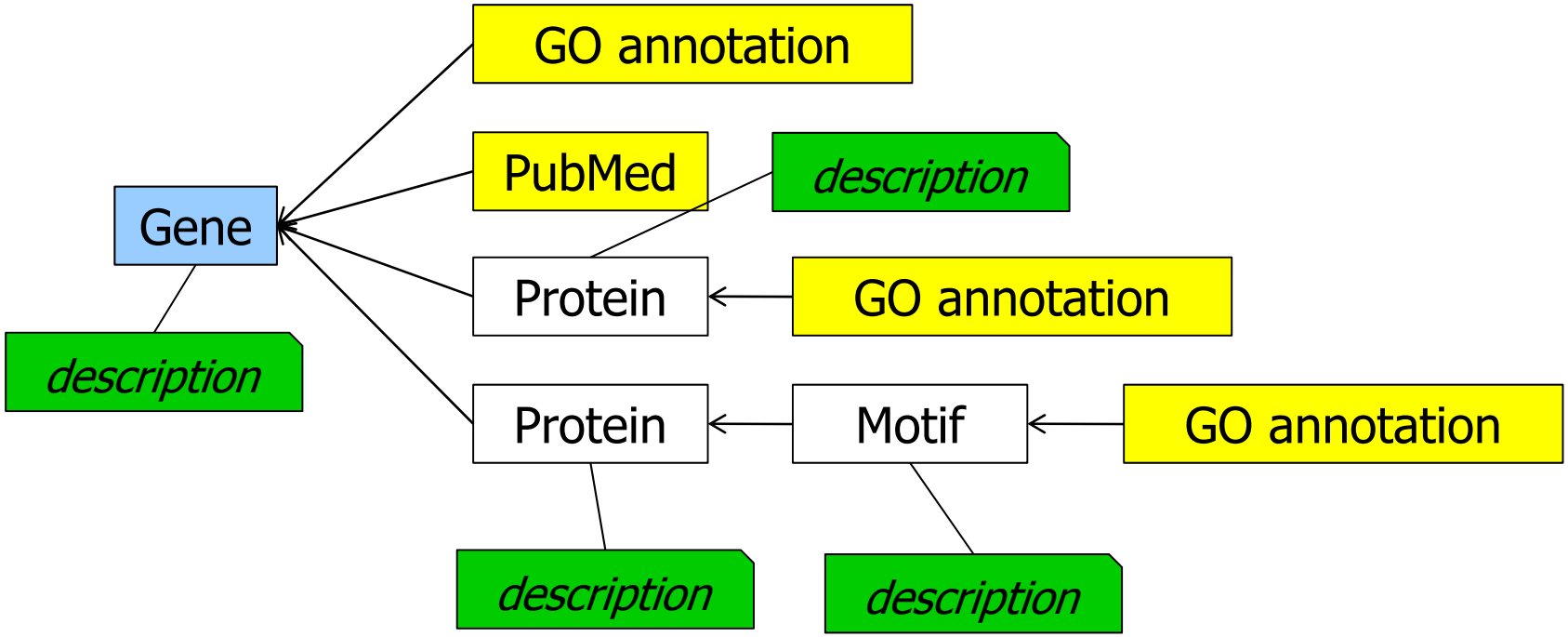
[RML05]

4. Textual Attributes

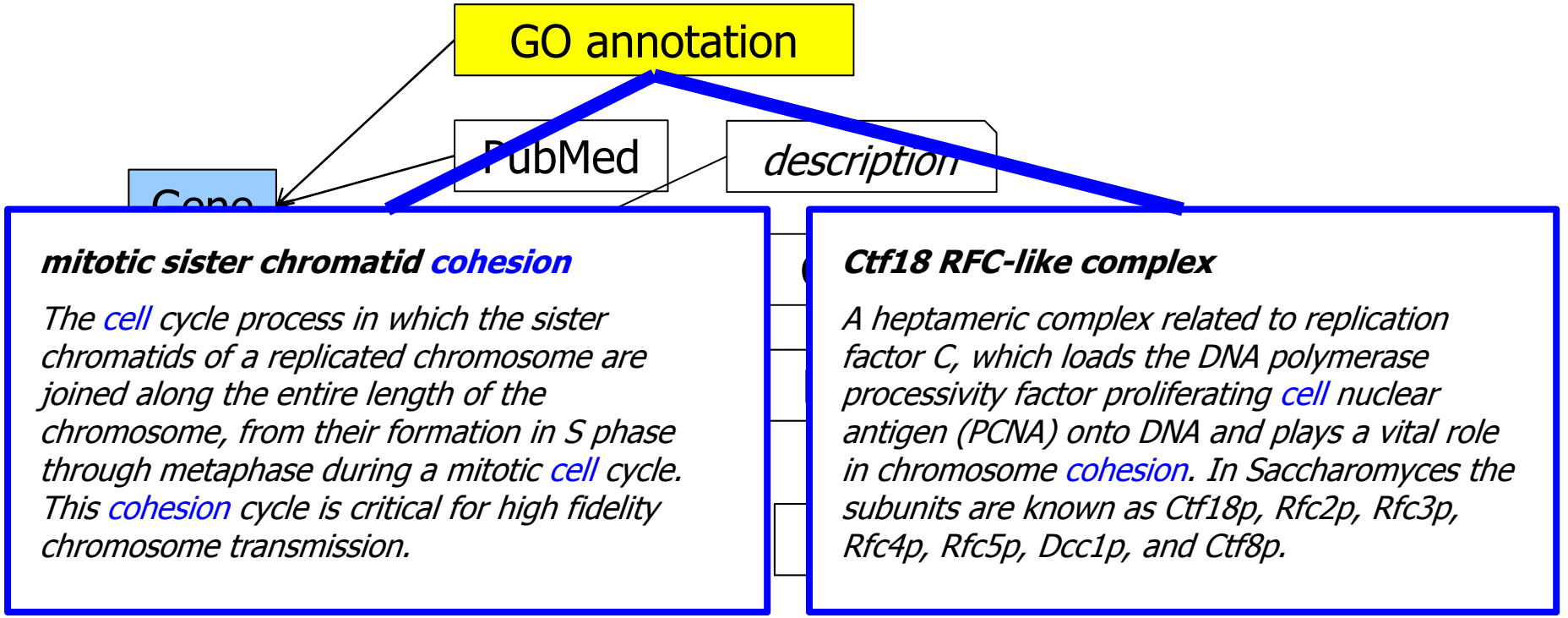
- Many queries are not based on strict criteria but use keywords
- Keywords match fields like descriptions, annotations, explanations, abstract, summary
- Accordingly, nodes on a path are matched to different degrees
  - Example: "Search genes involved in cell cohesion"



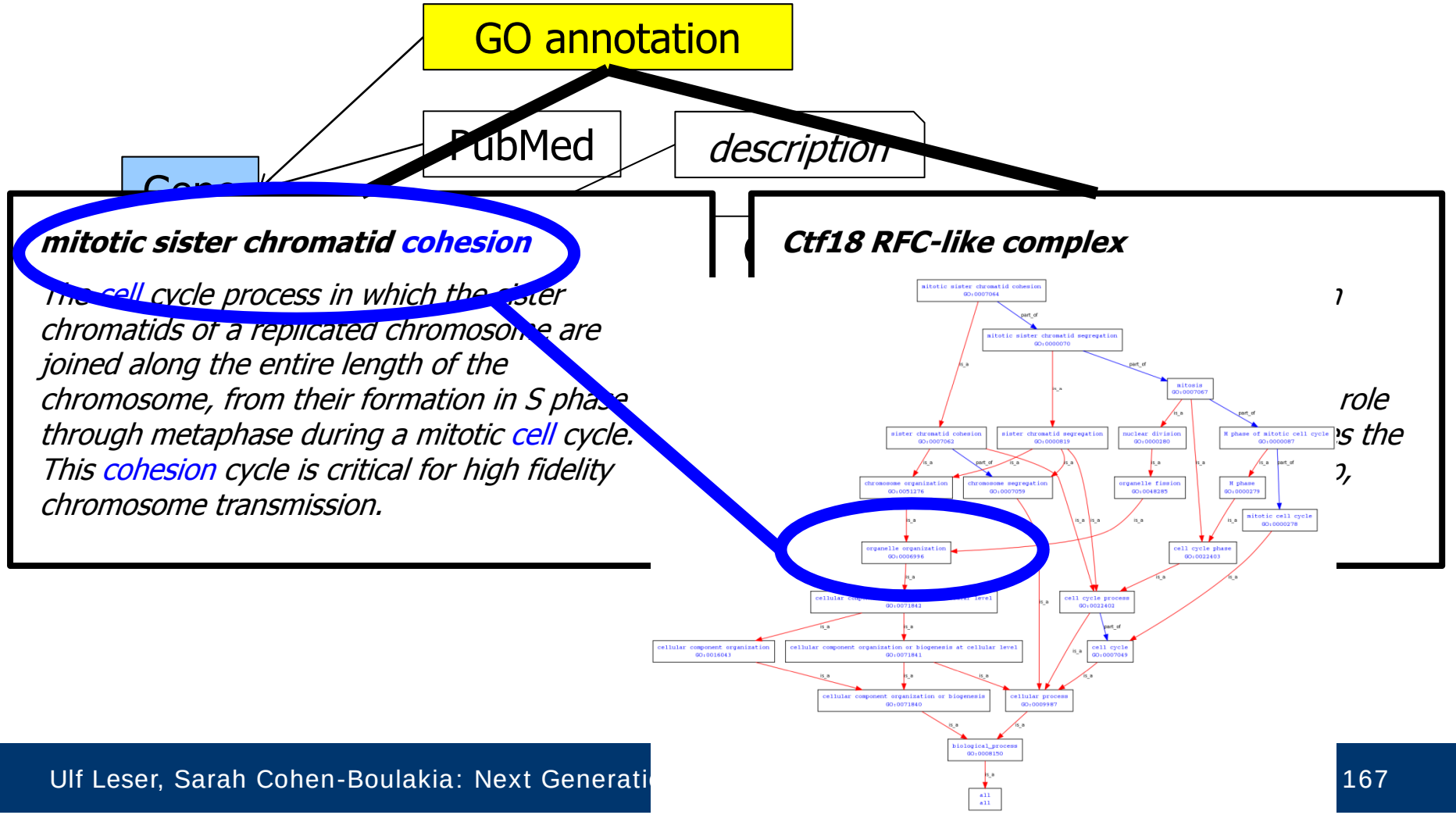
Example Continued



Example Continued



Example Continued



5. Evaluation

- Probably the hardest problem
- Problem: To what and how should results be compared?

- How: Choice of metrics
  - Precision at k, average precision, ROC, …

- "To what" option 1: Expert opinion
  - Favors the certain, ignores the surprising
  - Subjective (inter-annotator agreement?)
  - Not scalable
- "To what" option 2: Gold standard data sets
  - No generally accepted gold standards exist - everybody uses its own

More Problems with Ranking

- Computation and comparison of ranking under multiple criteria is a hard problem relevant for many domains
- Many results apply to the LS as well
- Note that in LS result sets are often different, i.e., the overlap of ranked results is small

- Comparing rankings, e.g. [FKM+06]
- Computing consensus ranking, e.g. [Ail10]
- Top-K query optimization, e.g. [IBS08]

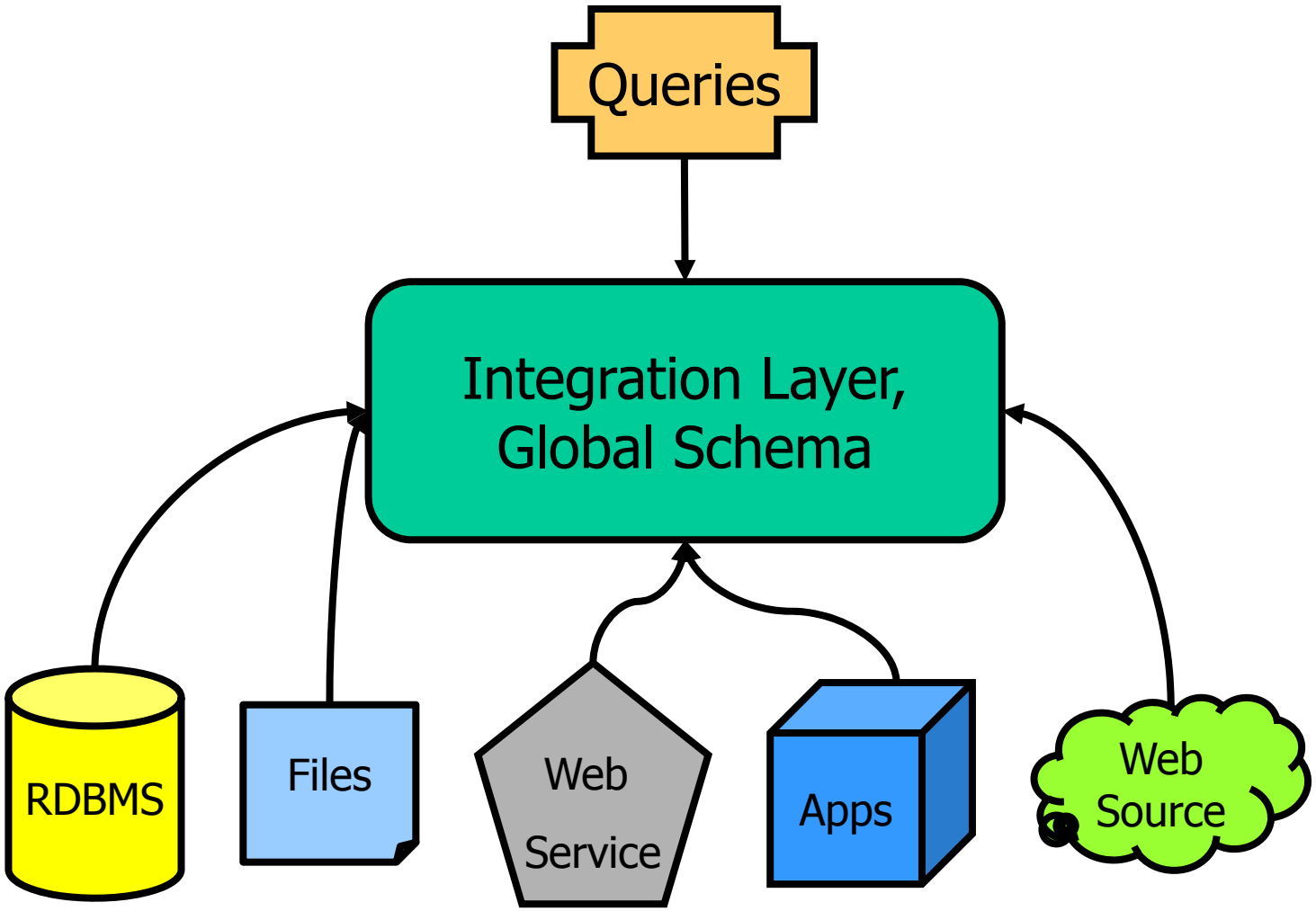
Related Work

- Ranking in IR, especially on the web
  - Also combine textual with topological evidence
  - But: Unstructured entities, no entity classes, semantic-free links, different query types (no paths)
- Keyword searches in relational databases
  - Also consider paths through a data graph
  - Also may use class information
  - But: Different query types (subgraphs)
- Long tradition in AI research
  - Bayesian networks, fuzzy logic, Dempster-Shafer Theory of Believe, …
- Probabilistic databases
  - Highly similar setting
  - Also research on different semantics of uncertainty and on different methods for uncertainty propagation through a query network

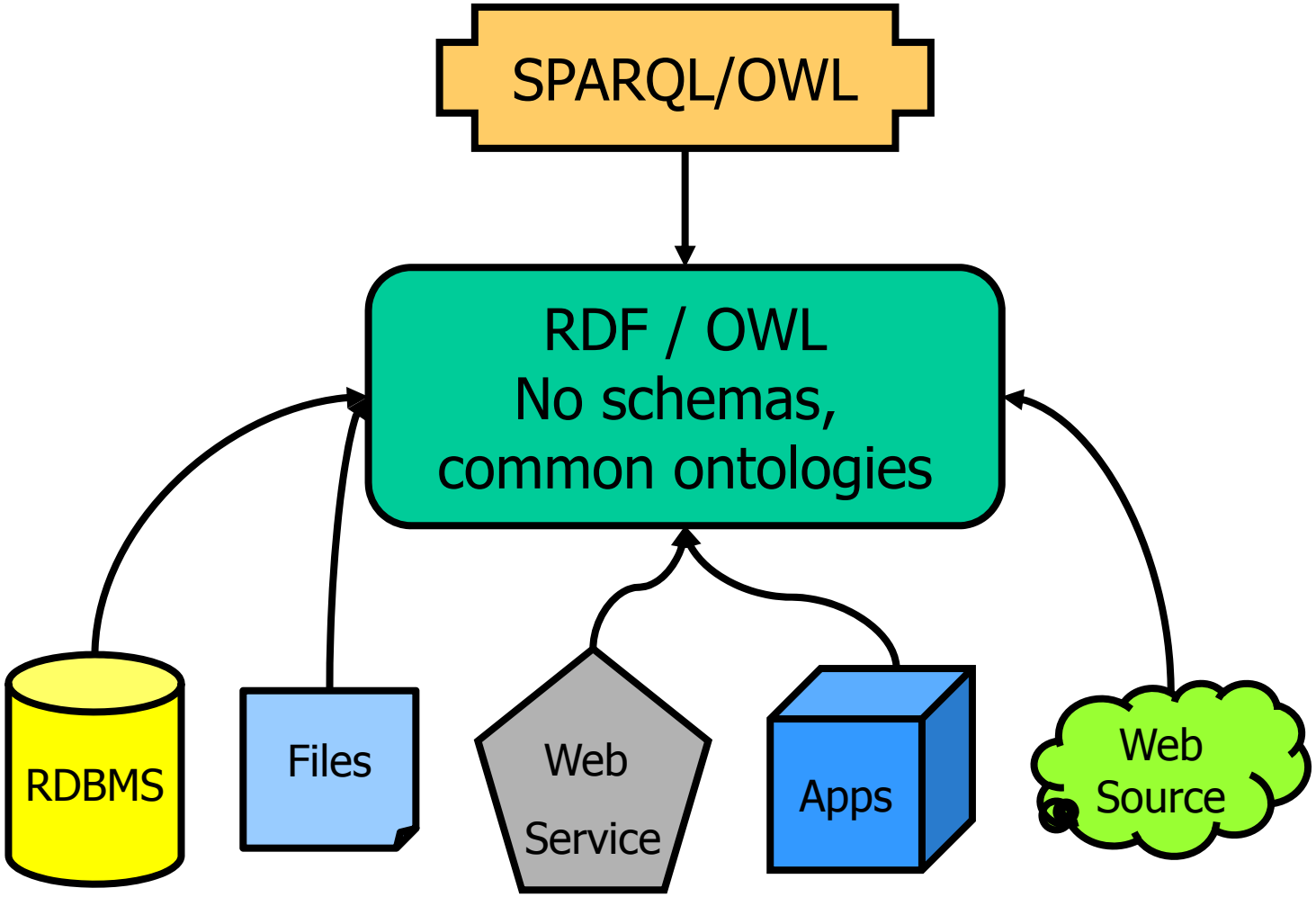
Three Trends

Data Integration Workflows	• Integration means analysis, and analysis means integration • No schemas, no explicit semantics • Scientific workflow systems	Effort Analysis Provenance Quality
Ranking	• Report results in a biologically meaningful order • Stays with queries, adds ranking • Requires a DI system in place	Effort Analysis Provenance Quality
Semantic Web	• Reduce upfront cost of DI • No schemas, explicit semantics • Semantic Web tech. (RDF, SPARQL)	Effort Analysis Provenance Quality

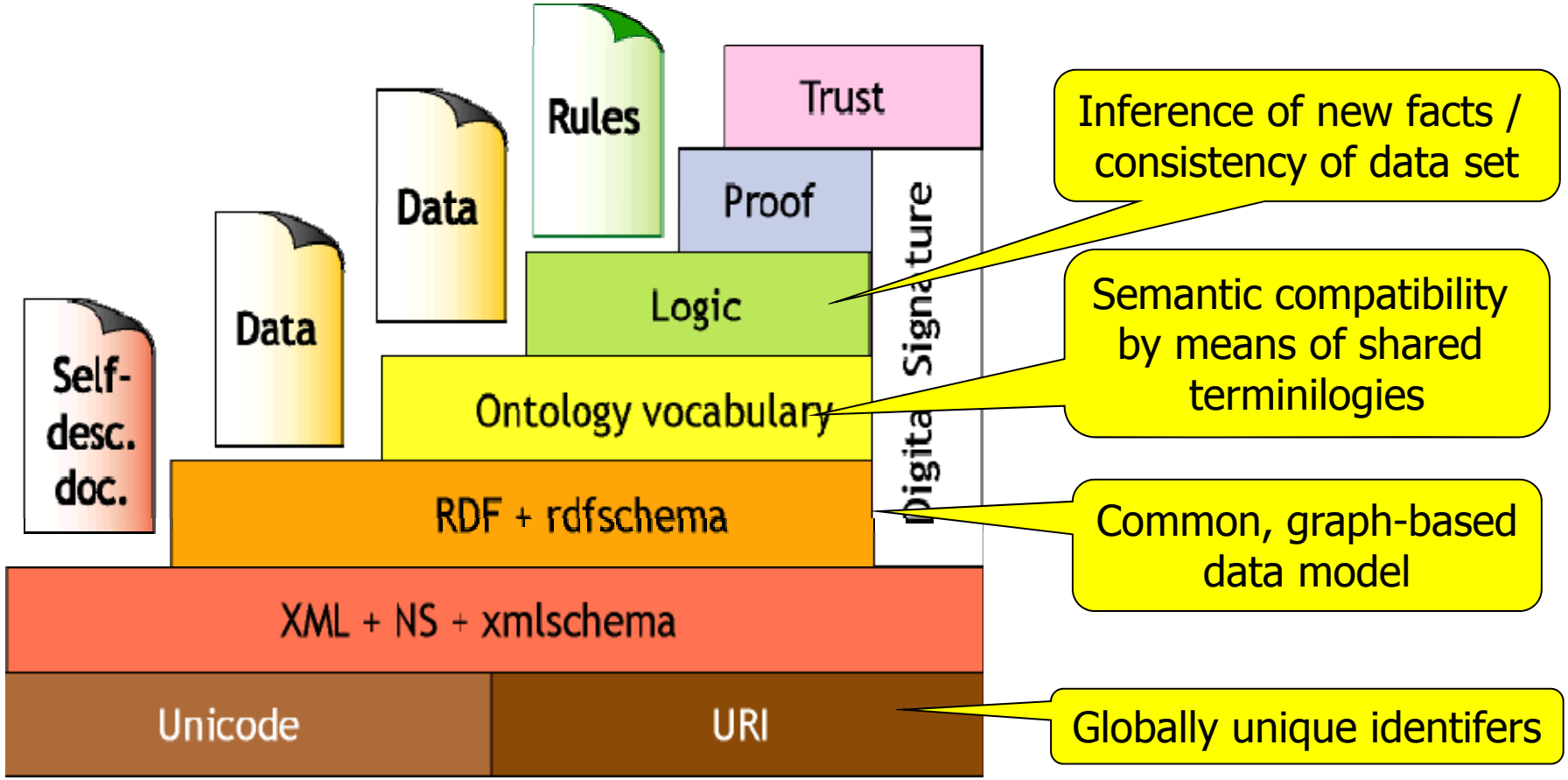
Classical View



Semantic Web Approach

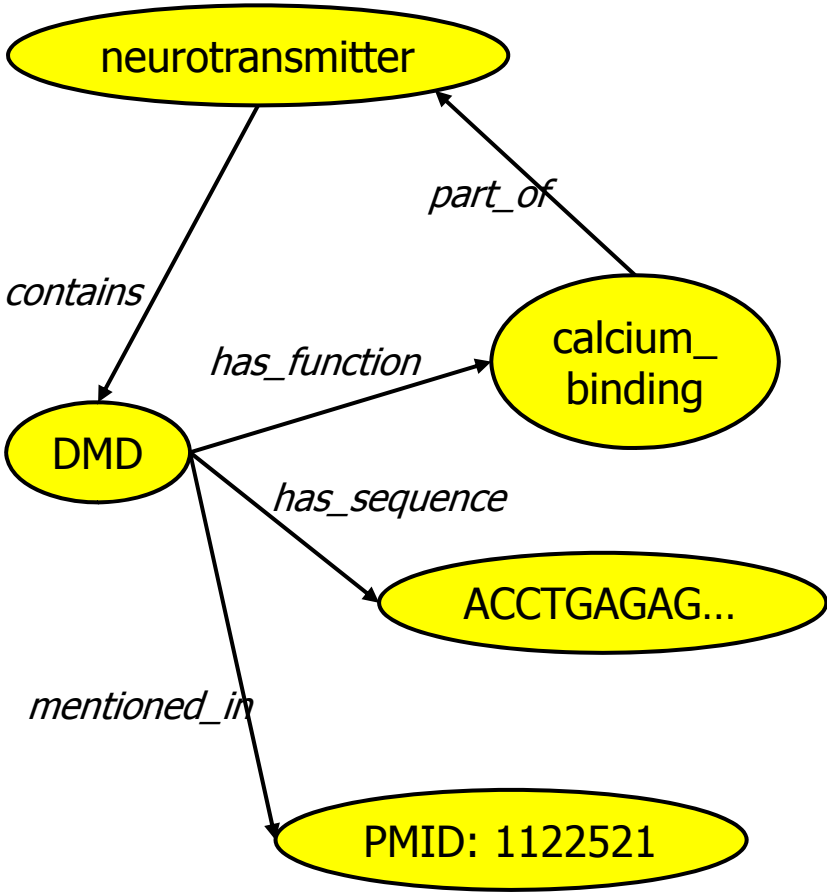


Semantic Web „Layer Cake“ [BHL01]



Resource Description Framework

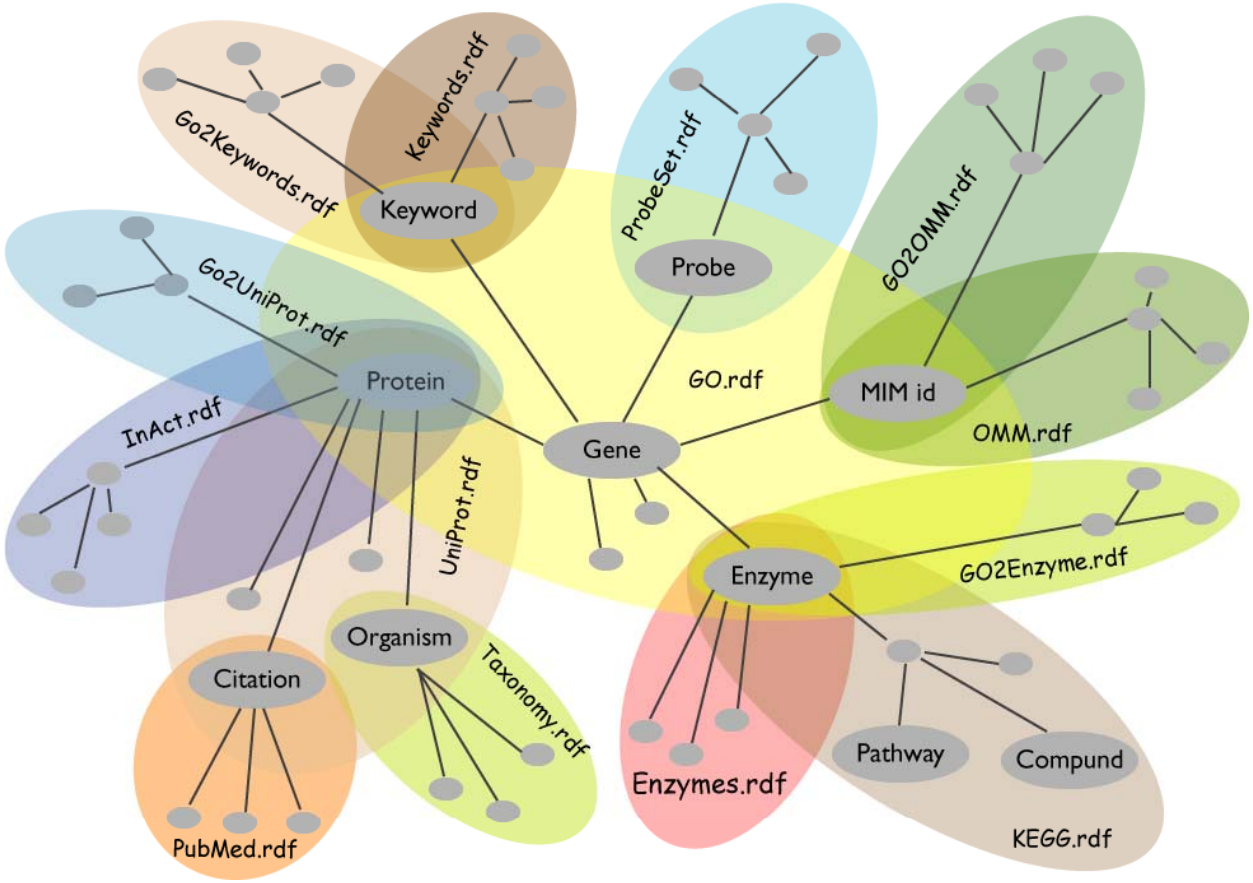
- Simple, graph-based data model
- Triples: Subject, predicate, object
  - n-ary relationships through blank nodes
  - Reification: Statements about statements
- Several syntactic representations
- RDF database: Set of RDF triple
  - Several systems available
- SPARQL: W3C standard for querying a RDF database



Semantic Web for Data Integration

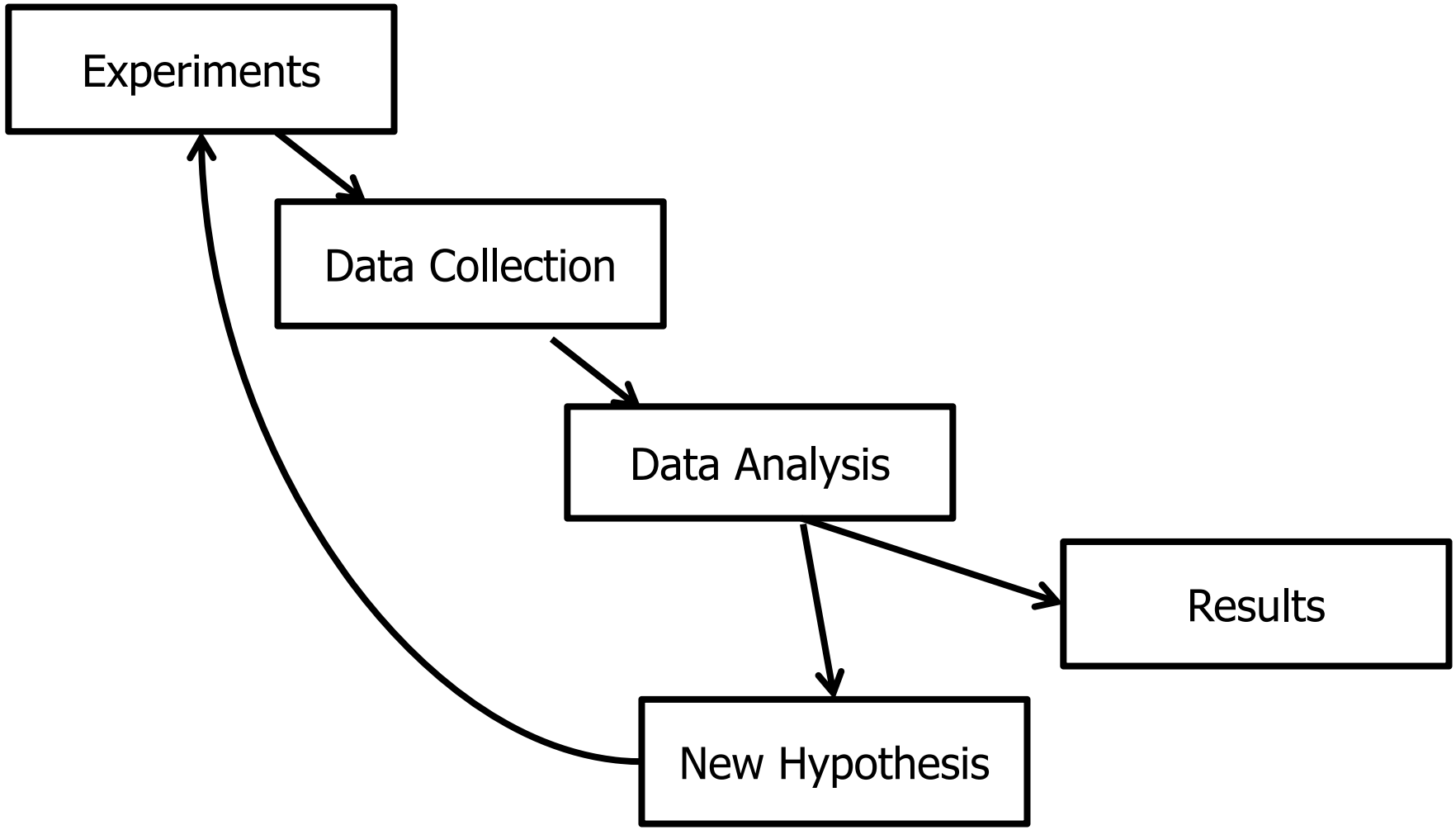
- Focus on semantic problems and upfront integration cost
- Usual approach
  - RDFify everything
    - RDF as common data model (not as global schema)
  - Trust on the usage of ontologies to cope with semantic heterogeneity at the instance level
  - Trust on the existence of ontologies to cope with semantic heterogeneity on the schema level
  - Use SPARQL as language to pose queries across data sources
- Sometimes: Use OWL for inferencing
  - Especially consistency of data sets, inference of new triples
  - Almost exclusively used: class, subclassOf, sameAs

RDF as Common Data Model

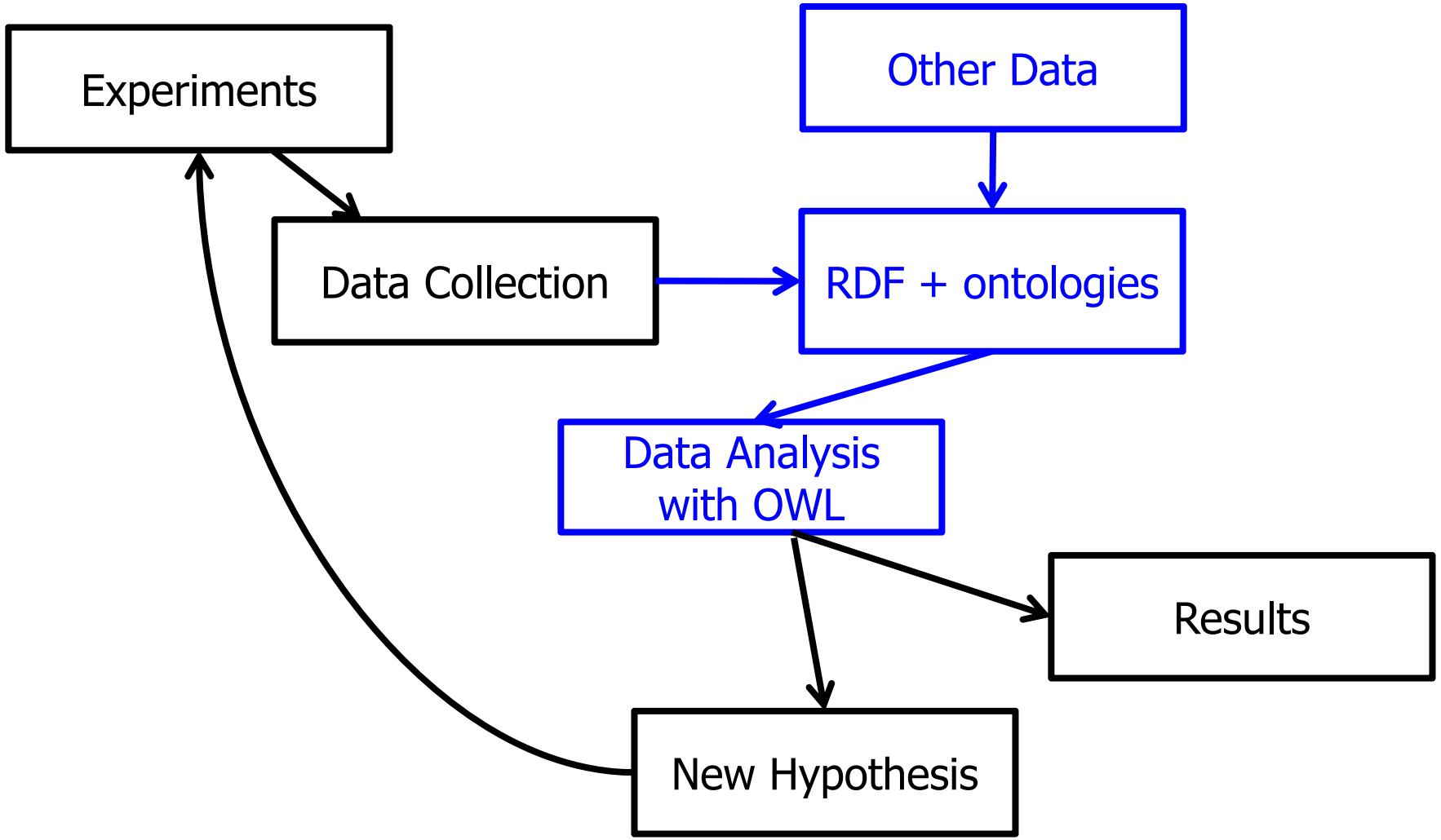


Hermann, W3C, 2007

Life Science Research Food Chain



… using Semantic Web Techniques



Some Examples

- BioDash [NQ06]
  - Drug discovery, focus on browsing, „semantic lenses" as views on a RDF
- Semantic web-enabled data integration (SWEDI) [PRM+07]
  - Transcription factors, lack of schema-level ontologies
- SemWeb for translational research [RCB+07]
  - OWL performance, lack of rules and axioms, data cleansing and ranking
- BioGateway [ABE+08]
  - General purpose DI system, problem of missing transitivity in SPARQL
- Rio2RDF [BNT+08]
  - Large-scale transformation of biological databases into RDF
- Chem2Bio2RDF [CDJ+10]
  - Chemoinformatics, no semantic integration, issue of de-duplication

Much Uptake

- Some of the largest RDF data sets come from the LS
  - LinkedLifeData: 6 billion triples (PubMed: 1.5B; UniProt: ~2B), 23 sources
  - Biuo2RDF: 40 sources, 30B triples

- B. Chen, et al., *Chem2Bio2RDF: a semantic framework for linking and data mining chemogenomic and systems chemical biology data. BMC Bioinformatics, 2010. 11:*
- H. Oliver, et al., *A user-centred evaluation framework for the Sealife semantic web browsers. BMC Bioinformatics, 2009. 10 Suppl 10: p. S14.*
- K.H. Cheung, et al., *A journey to Semantic Web query federation in the life sciences. BMC Bioinformatics, 2009. 10 Suppl 10: p. S10.*
- T. Slater, C. Bouton, and E.S. Huang, *Beyond data integration. Drug Discov Today, 2008. 13(13-14): p. 584-9.*
- J.A. Sagotsky, L. Zhang, Z. Wang, S. Martin, and T.S. Deisboeck, *Life Sciences and the web: a new era for collaboration. Mol Syst Biol, 2008. 4: p. 201.*
- C. Pasquier, *Biological data integration using Semantic Web technologies. Biochimie, 2008. 90(4): p. 584-94.*
- A. Newman, J. Hunter, Y.F. Li, C. Bouton, and M. Davis. *A scale-out RDF molecule store for distributed processing of biomedical data. in Workshop on Semantic Web*
- N. Kobayashi and T. Toyoda, *Statistical search on the Semantic Web. Bioinformatics, 2008. 24(7): p. 1002-10.*
- C. Goble and R. Stevens, *State of the nation in data integration for bioinformatics. J Biomed Inform, 2008. 41(5): p. 687-93.*
- H.F. Deus, et al., *A Semantic Web management model for integrative biomedical informatics. PLoS One, 2008. 3(8): p. e2946.*
- F. Belleau, M.-A. Nolin, N. Tourigny, P. Rigault, and J. Morissette, *Bio2RDF: Towards a mashup to build bioinformatics knowledge systems. Journal of Biomedical*
- E. Antezana, et al. *Structuring the life science resourceome for semantic systems biology: lessons from the BioGateway project. in Workhop on Semantic Web*
- S. Sahoo, O. Bodenreider, K. Zeng, and A. Sheth. *An Experiment in Integrating Large Biomedical Knowledge Resources with RDF: Application to Associating*
- A. Ruttenberg, et al., *Advancing translational research with the Semantic Web. BMC Bioinformatics, 2007. 8 Suppl 3: p. S2.*
- L.J. Post, M. Roos, M.S. Marshall, R. van Driel, and T.M. Breit, *A semantic web approach applied to integrative bioinformatics experimentation: a biological use case*
- R.C. Gudivada, X.A. Qu , A.G. Jegga, E.K. Neumann , and B.J. Aronow. *A Genome - Phenome Integrated Approach for Mining Disease-Causal Genes using Semantic*
- E.K. Neumann and D. Quan. *BioDash: a Semantic Web dashboard for drug development. in Pac Symp Biocomput. 2006. Hawai, US.*
- T. Kazic. *Putting Semantics into the Semantic Web: How Well Can It Capture Biology? in Pacific Symposium on Biocomputing. 2006.*
- B.M. Good and M.D. Wilkinson, *The Life Sciences Semantic Web is full of creeps! Brief Bioinform, 2006. 7(3): p. 275-86.*
- S. Mukherjea, *Information retrieval and knowledge discovery utilising a biomedical Semantic Web. Brief Bioinform, 2005. 6(3): p. 252-62.*
- …

Much Uptake?

- Most papers promise success (if X, Y, Z)
- Fewer papers report on successful applications

- B. Chen, et al., *Chem2Bio2RDF: a semantic framework for linking and data mining chemogenomic and systems chemical biology data. BMC Bioinformatics, 2010. 11:*
- H. Oliver, et al., *A user-centred evaluation framework for the Sealife semantic web browsers. BMC Bioinformatics, 2009. 10 Suppl 10: p. S14.*
- K.H. Cheung, et al., *A journey to Semantic Web query federation in the life sciences. BMC Bioinformatics, 2009. 10 Suppl 10: p. S10.*
- T. Slater, C. Bouton, and E.S. Huang, *Beyond data integration. Drug Discov Today, 2008. 13(13-14): p. 584-9.*
- J.A. Sagotsky, L. Zhang, Z. Wang, S. Martin, and T.S. Deisboeck, *Life Sciences and the web: a new era for collaboration. Mol Syst Biol, 2008. 4: p. 201.*
- C. Pasquier, *Biological data integration using Semantic Web technologies. Biochimie, 2008. 90(4): p. 584-94.*
- A. Newman, J. Hunter, Y.F. Li, C. Bouton, and M. Davis. *A scale-out RDF molecule store for distributed processing of biomedical data. in Workshop on Semantic Web*
- N. Kobayashi and T. Toyoda, *Statistical search on the Semantic Web. Bioinformatics, 2008. 24(7): p. 1002-10.*
- C. Goble and R. Stevens, *State of the nation in data integration for bioinformatics. J Biomed Inform, 2008. 41(5): p. 687-93.*
- H.F. Deus, et al., *A Semantic Web management model for integrative biomedical informatics. PLoS One, 2008. 3(8): p. e2946.*
- F. Belleau, M.-A. Nolin, N. Tourigny, P. Rigault, and J. Morissette, *Bio2RDF: Towards a mashup to build bioinformatics knowledge systems. Journal of Biomedical*
- E. Antezana, et al. *Structuring the life science resourceome for semantic systems biology: lessons from the BioGateway project. in Workhop on Semantic Web*
- S. Sahoo, O. Bodenreider, K. Zeng, and A. Sheth. *An Experiment in Integrating Large Biomedical Knowledge Resources with RDF: Application to Associating*
- A. Ruttenberg, et al., *Advancing translational research with the Semantic Web. BMC Bioinformatics, 2007. 8 Suppl 3: p. S2.*
- L.J. Post, M. Roos, M.S. Marshall, R. van Driel, and T.M. Breit, *A semantic web approach applied to integrative bioinformatics experimentation: a biological use case*
- R.C. Gudivada, X.A. Qu , A.G. Jegga, E.K. Neumann , and B.J. Aronow. *A Genome - Phenome Integrated Approach for Mining Disease-Causal Genes using Semantic*
- E.K. Neumann and D. Quan. *BioDash: a Semantic Web dashboard for drug development. in Pac Symp Biocomput. 2006. Hawai, US.*
- T. Kazic. *Putting Semantics into the Semantic Web: How Well Can It Capture Biology? in Pacific Symposium on Biocomputing. 2006.*
- B.M. Good and M.D. Wilkinson, *The Life Sciences Semantic Web is full of creeps! Brief Bioinform, 2006. 7(3): p. 275-86.*
- S. Mukherjea, *Information retrieval and knowledge discovery utilising a biomedical Semantic Web. Brief Bioinform, 2005. 6(3): p. 252-62.*
- …

Unusual : NCBO Resource Index [JLP+10]

- Recognizes and tags concepts from 200+ ontologies in flat-file representation of 20+ BDB
- Concept-based keyword queries over multiple databases
  - No data integration, but uniform, "semantic" search
- Scalability problem:
  Tagging gigabytes of texts with ~4 million terms, each consisting of multiple tokens and allowing for errors [SBJ+09]
- Not sure if this is a Semantic Web application …

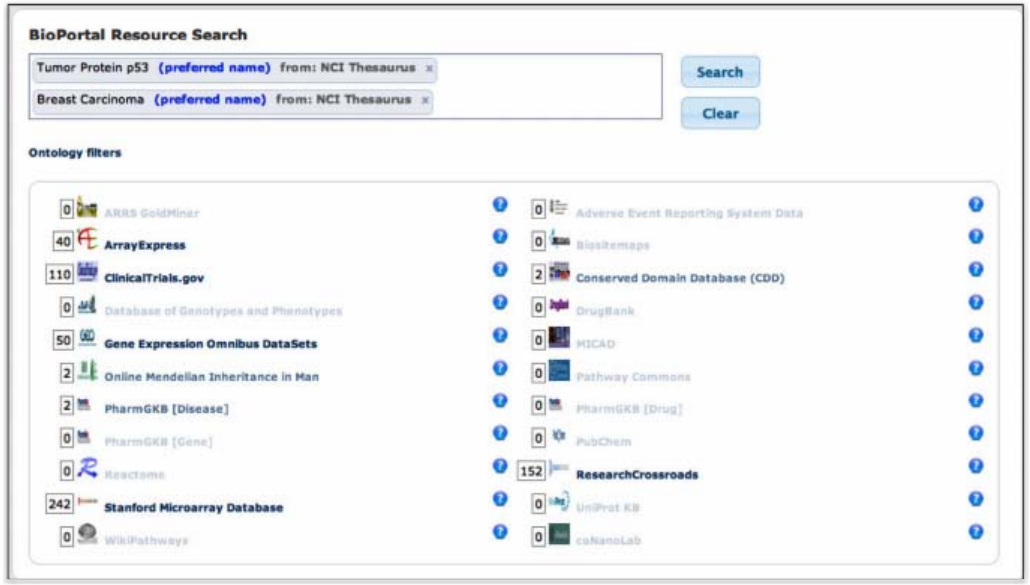


Fig. 2. The search for resources that contain both "Tumor Protein p53" AND "Breast Carcinoma."

Problems Tackeled

- Low upfront cost
  - RDFifying is simple, many DBs are available in RDF
  - Very flexible model, no schemas
  - No semantic reconciliation in first phase
  - Allows quick and uniform access to data from many sources
- Inference over equality-links supported by sameAs (OWL)
- Data provenance is exposed (usually in namespaces)
- Exploitation of the fact that BDB are highly interlinked at instance level
  - Perfect model for Linked Open Data (LOD)

Opportunities

1. Dealing with semantic heterogeneity
2. RDF as data model
3. Extensions to SPARQL

1. No Semantic Integration (without Ontologies)

Dealing with Semantic Heterogeneity

- RDF'fization does not solve semantic heterogeneity, but postpones it
  - Data is only available in a common data model
  - Predicate names are not unified, but added
  - Objects are not unified, but added
- URIs do not enforce common IDs for common objects
  - Everybody may invent arbitrary URIs
  - sameAs is not enough – not all links are equal
- Existence of ontologies are a prerequisite for using SemWeb technologies, not a consequence of doing so
  - But: Ontologies may contradict each other – new problems

SemWeb Ontologies ≠ LS Ontologies

- Ontologies are extraordinary successful in the LS
- Almost all successful LS-ontologies are informal
  - No axioms, roles, attributes, formulas; only ISA and PART-OF
- LS-ontologies are used only for annotation
  - Controlled vocabularies on the instance level
- The field does slowly solve the problem of inconsistent terminologies on the instance level
- Little work on the schema level

Work in This Direction

- BioPortal [NSW+09]
  - Common access to 200+ biological ontologies
- OBO foundry [SAR+08]
  - De-facto standard for design of biological ontologies
- Ontology matching [ES08, KTR07]
  - Instance level: Power of links between objects and ontology terms
  - Increased complexity if OWL predicates should be considered
- De-duplication in RDF [IPSN10]
- Extensions to SPARQL [KJ07]
  - RegExp for predicate names
- Ontology bootstrapping from text [BL09]
  - Recognition of concepts and ISA relationships

2. RDF as Common Data Model

- RDF actually was meant to be a model for representing metadata
  - Discrete, certain facts
  - Geared towards logical inference
  - Numerical values not considered as such (no data types)
- But: LS data is dirty
  - Dealing with uncertainty, contradictions, noise, …
- But: LS data can be voluminous
  - Not terrible large, but much larger than typical metadata sets
  - Experimental data
  - Need for hybrid approach
    - RDF for representing information (derived facts)
    - Links to original data sets in other format

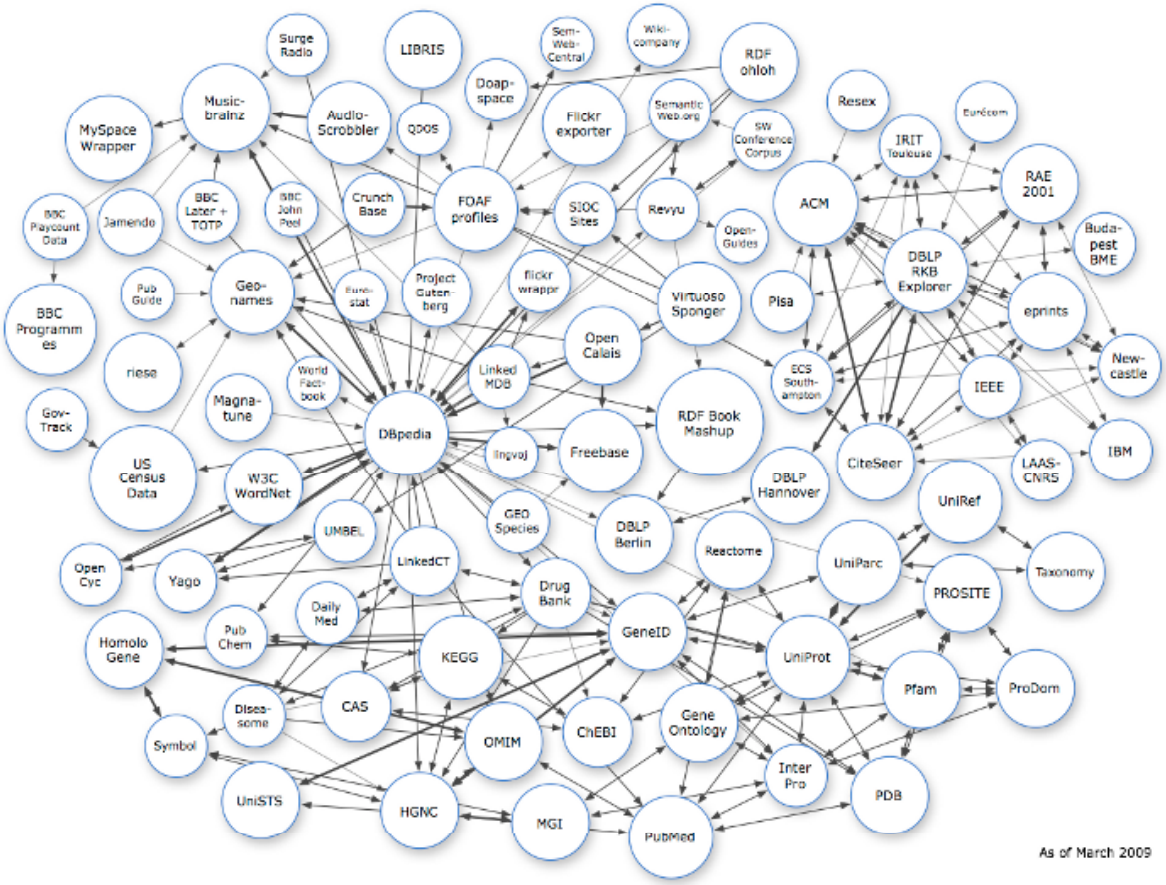
3. Extensions to SPARQL

- Given the level of heterogeneity in merged RDF data sets, a powerful query language is a pre-requisite for comprehensive analysis
- However, SPARQL lacks
  - … grouping and aggregation for in-query de-duplication and data fusion
  - … user-defined predicates for implementing non-standard DI functions
  - … an understanding of class hierarchies to exploit semantic structures
  - … general transitive predicates to cope with heterogeneous schemas
  - … a sensible way to access multiple distributed RDF databases
  - … methods to cope with confidence / probabilities

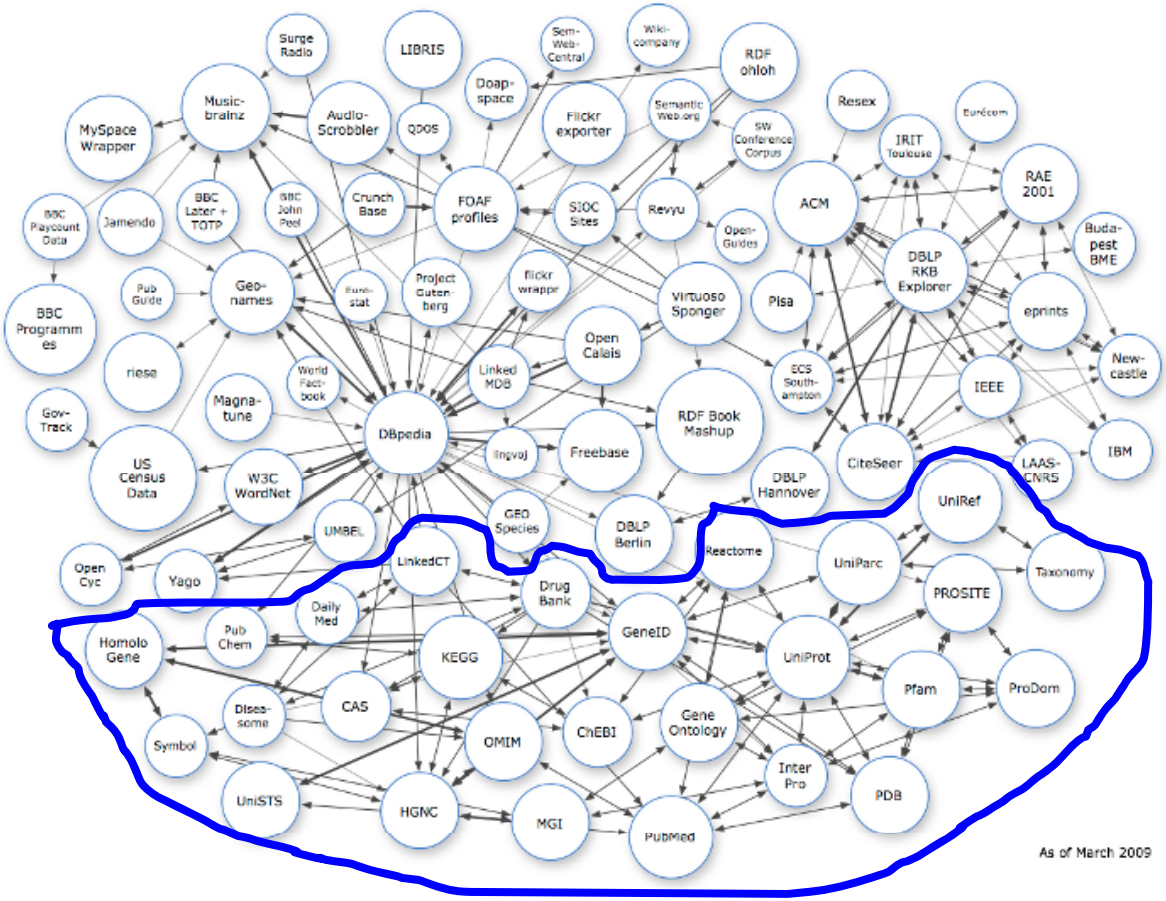
Work in these Directions

- Distributed SPARQL optimization
  - DARQ: Query rewriting based on predicate mappings [QL08]
  - Avalanche: SPARQL over Linked Open Data [BA10]
- Statistical aggregation in SPARQL [KT08]
  - Ad-hoc syntactic extension to SPARQL
- Using ontology mappings in query processing
  - Query rewriting using graph pattern rewriting [CSM+10]
  - SPARQL query rewriting using (relational) views [CWWM07]
- Transitive predicates for SPARQL [KAC+02, KJ07]
- OWL for query rewriting
  - Not scalable [ZAV+07]

Linked Open Data

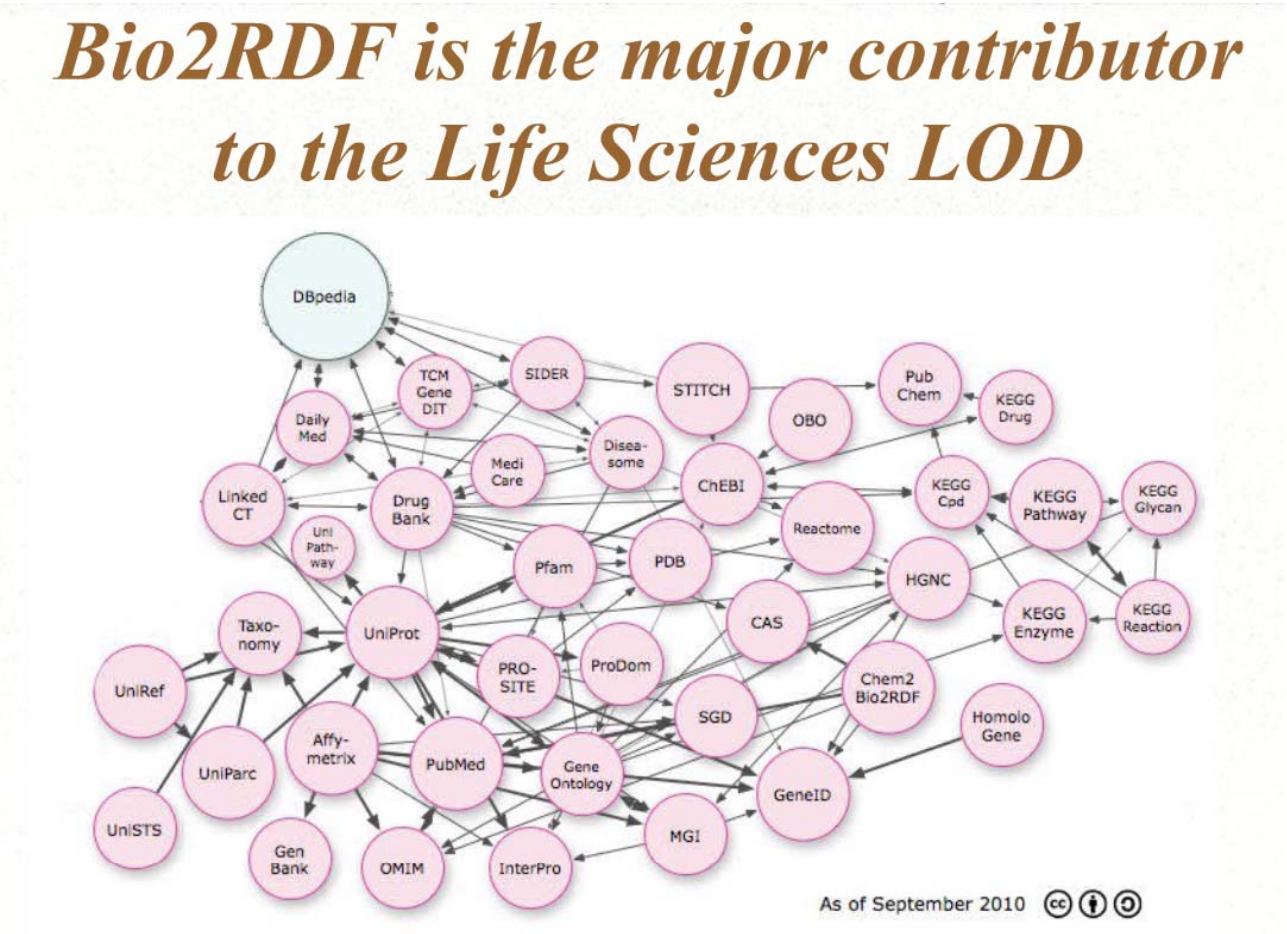


Linked Open Life Science Data



Life Sciences are a major contributor to LOD

Life Sciences in General?



[Bio2RDF; BioCuration 2010]

This Tutorial

- Part I – Data Integration for the Life Sciences
- Part II – Past and Presence
- Part III – Current Trends
- Part IV – Conclusions

Wrap-Up

- Integration in CS research mostly means logical information integration
  - Schemas first, discrete attributes, schema matching, queries rewriting
  - "Data not important"
- Integration in LS firstly requires statistical data integration
  - Noisy experimental data, statistical aggregation
  - "Schema not important"
- $II \cap DI \neq \emptyset$
  - DI requires the data to be at hand
  - II may use instance data
  - DI techniques may depend on the origin of the data
  - DI and II require reconciliation of objects and object IDs

Three Trends

- DI workflows emphasize data analysis and may support DI by sharing
  - But may be inefficient if results are to be re-used a lot
- Ranking focuses on providing meaningful answers despite questionable data quality
  - But falls short in further processing the ranked data
- Semantic Web strive for cost reduction for initial DI phases
  - But do not yet provide mature tools for a tighter integration or integrated analysis

Three Trends

- DI workflows emphasize data analysis and may support DI by sharing
  - But may be inefficient if results are to be re-used a lot
  - Probably most appealing to LS researchers
- Ranking focuses on providing meaningful answers despite questionable data quality
  - But falls short in providing clues on how to further process the ranked data
  - Probably most appealing to database researchers
- Semantic Web approaches strive for cost reduction for initial DI phases
  - But do not yet provide mature tools for a tighter integration or integrated analysis

Further Trend: Standardization

- With proper standards in place, II becomes simple
  - Vocabularies (ontologies)
  - Schemas (GMOD, DAS, BioMart)
  - Required information (MIA* standards)
  - Exchange formats (BioPax, GFF, MAGE-TAB, …)

- Essentially, semantic integration is performed upfront in the sources

Example: Int. Cancer Genome Cons.

- Large-scale, international endeavor
- Planned for 50 different cancer types
- Cancer types are assigned to countries
- Distributed BioMar… infrastructure
- First federat… to a large int… project [HAA+…

50 different cancer types, 500 samples per type, always control + cancer
> 50.000 genomes

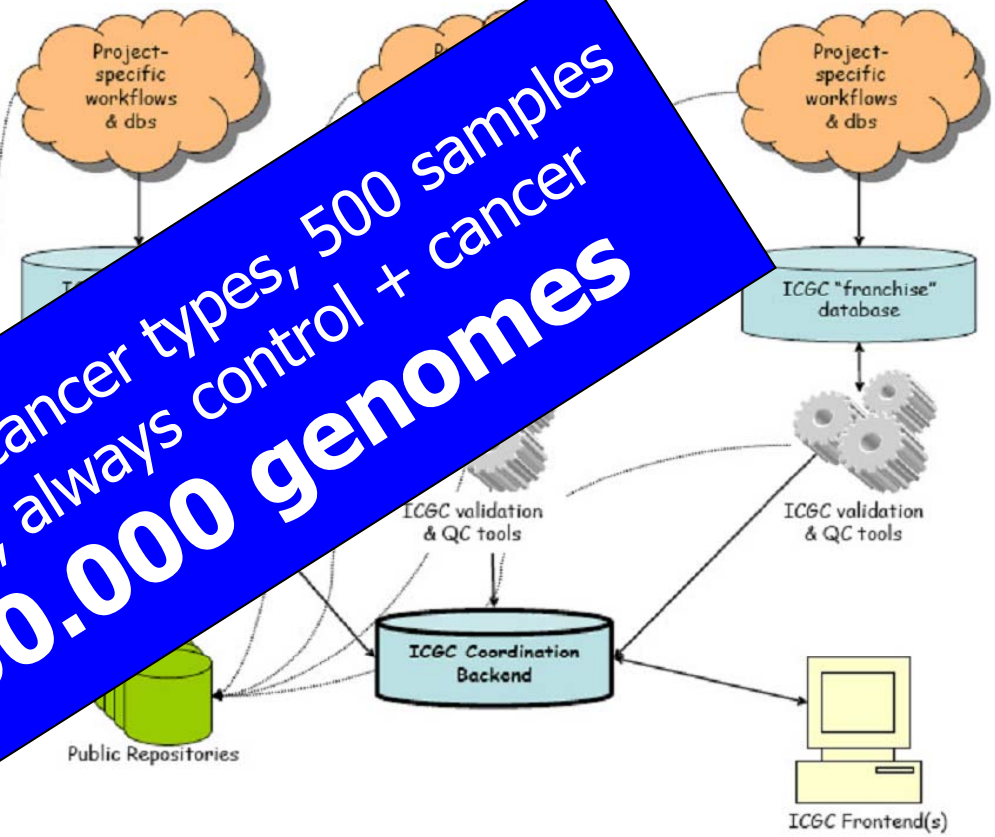


Figure 2: ICGC data coordination as a franchise system

- Standard schema, restricted functionality, distributed architecture

Further Issue: Sharing

„We requested data from ten investigators who had published in either PLoS Medicine or PLoS Clinical Trials. All responses were carefully documented. In the event that we were refused data, we reminded authors of the journal's data sharing guidelines. If we did not receive a response to our initial request, a second request was made. Following the ten requests for raw data, three investigators did not respond, four authors responded and refused to share their data, two email addresses were no longer valid, and one author requested further details. A reminder of PLoS's explicit requirement that authors share data did not change the reply from the four authors who initially refused. Only one author sent an original data set.” [SV09]

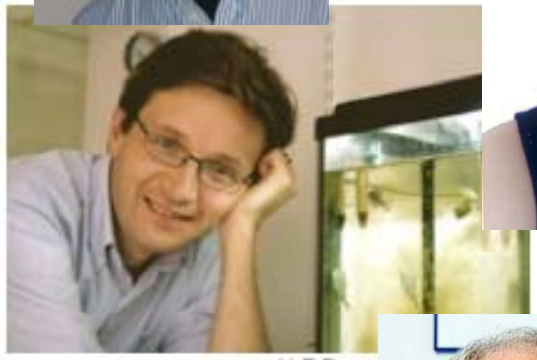
Most integration projects fail for social reasons, not for technical ones

What the Heck …

- Apparently, LS researchers do not like DR
- Two cultures
  - Publish a database or a method for building databases
  - Publish in conferences or in journals
  - Find a new fact about the physical world or a new method for an abstract problem
  - Know 100 genes and 3 methods, or know 10 methods and 1 gene
- So why should you care?
  - Data integration is a pressing, real, ubiquitous problem in LS

Life Sciences are changing the (your) world

Acknowledgements



Bayer HealthCare
Bayer Schering Pharma

Deutsche Forschungsgemeinschaft DFG

Bundesministerium für Bildung und Forschung

Berlin.de

Surveys

[DOB95] Davidson, S., Overton, G. C. and Buneman, P. (1995). "Challenges in Integrating Biological Data Sources." *Journal of Computational Biology* ***2(4): 557-572.***

[GS08] Goble, C. and Stevens, R. (2008). "State of the nation in data integration for bioinformatics." *J Biomed Inform* ***41(5): 687-93.***

[HK04] Hernandez, T. and Kambhampati, S. (2004). "Integration of Biological Sources: Current Systems and Challenges Ahead." *SIGMOD Record* ***33(5): 51-60.***

[Karp94] Karp, P. D. (1994). "Report of the Workshop on Interconnection of Molecular Biology Databases", SRI International Artificial Intelligence Center, Stanford, California.

[LC03] Lacroix, Z. and Critchlow, T., Eds. (2003). "Bioinformatics - Managing Scientific Data". San Francisco, California, Morgan Kaufmann Publishers.

[Sea05] Searls, D. B. (2005). "Data Integration: Challenges for Drug Discovery." *Nature Reviews Genetics* ***4: 45-58.***

[Ste03] Stein, L. D. (2003). "Integrating biological databases." *Nat Rev Genet* ***4(5): 337-45.***

[Ste08] Stein, L. D. (2008). "Towards a cyberinfrastructure for the biological sciences: progress, visions and challenges." *Nature Reviews Genetics* ***9(9): 678-88.***

References

[AKD10] Ailamaki, A., Kantere, V. and Dash, D. (2010). "Managing scientific data." *Communications of the ACM* ***53(6): 68-78.***

[Ail10] Ailon, N. (2010). "Aggregation of Partial Rankings, p-Ratings and Top-m Lists." *Algorithmica* ***57(2): 284-300.***

[Alb09] Albrecht, A. (2009). "METL: Managing and Integrating ETL Processes". VLDB PhD workshop.

[ABE+08] Antezana, E., Blondé, W., Egana, M., Rutherford, A., Stevens, R., De Baets, B., Mironov, V. and Kuiper, M. (2008). "Structuring the life science resourceome for semantic systems biology: lessons from the BioGateway project". Workhop on Semantic Web Applications and Tools for Life Sciences, Edinburgh, UK.

[AS10] Awad, A. and Sakr, S. (2010). "Querying Graph-Based Repositories of Business Process Models ". Database Systems for Advanced Applications. pp 33-44.

[ZCBD+09] Bao, Z., Cohen-Boulakia, S., Davidson, S. B., Eyal, A. and Khanna, S. (2009). "Differencing Provenance in Scientific Workflows". EEE Int. Conf. on Data Engineering, IEEE Computer Society.

[BA10] Basca, C. and Bernstein, A. (2010). "Avalanche: Putting the Spirit of the Web back into Semantic Web Querying". Workshop on Scalable Semantic Web Knowledge Base Systems.

[BCF+07] Baumgartner Jr, W. A., Cohen, K. B., Fox, L. M., Acquaah-Mensah, G. and Hunter, L. (2007). "Manual curation is not sufficient for annotation of genomic databases." *Bioinformatics* ***23(13): i41.***

[BEKM08] Beeri, C., Eyal, A., Kamenkovich, S. and Milo, T. (2008). "Querying business processes with BP-QL." *Information Systems* ***33(6): 477-507.***

[BNT+08] Belleau, F., Nolin, M.-A., Tourigny, N., Rigault, P. and Morissette, J. (2008). "Bio2RDF: Towards a mashup to build bioinformatics knowledge systems." *Journal of Biomedical Informatics* ***41(5): 706-716.***

[BHL01] Berners-Lee, T., Hendler, J. and Lassila, O. (2001). "The Semantic Web." *Scientific American* ***284: 34-43.***

[BBF+01] Bhat, T. N., Bourne, P., Feng, Z., Gilliland, G., Jain, S., Ravichandran, V., Schneider, B., Schneider, K., Thanki, N., Weissig, H.*, et al. (2001). "The PDB data uniformity project." Nucleic Acids Res* ***29(1): 214-8.***

References

[BY06] Birkland, A. and Yona, G. (2006). "BIOZON: a system for unification, management and analysis of heterogeneous biological data." *BMC Bioinformatics* ***7: 70.***

[BCB+08] Biton, O., Cohen-Boulakia, S., Davidson, S. B. and Hara, C. S. (2008). "Querying and Managing Provenance through User Views in Scientific Workflows". 24th Int. Conf. on Data Engineering, IEEE Computer Society.

[BLM+04] Bleiholder, J., Lacroix, Z., Murthy, H., Naumann, F., Raschid, L. and Vidal, M.-E. (2004). "BioFast: Challenges in Exploring Linked Life Science Sources." *SIGMOD Record* ***33(2).***

[BL09] Böhm, C. and Leser, U. (2009). "Graph-Based Ontology Construction from Heterogeneous Evidences". Int. Semantic Web Conference (ISWC), Washington, US.

[CDJ+10] Chen, B., Dong, X., Jiao, D., Wang, H., Zhu, Q., Ding, Y. and Wild, D. J. (2010). "Chem2Bio2RDF: a semantic framework for linking and data mining chemogenomic and systems chemical biology data." *BMC Bioinformatics* ***11: 255.***

[CWWM07] Chen, H., Wu, Z., Wang, H. and Mao, Y. (2006). "RDF/RDFS-based Relational Database Integration". International Conference on Data Engineeringing (ICDE), Atlanta, USA. pp 94.

[CBD+06] Cohen-Boulakia, S., Davidson, S. B., Froidevaux, C., Lacroix, Z. and Vidal, M.-E. (2006). "Path-based systems to guide scientists in the maze of biological data sources." *Journal of Bioinformatics and Computational Biology* ***4(5): 1069-1095.***

[BFL+04] Cohen-Boulakia, S., Froidevaux, C., Lair, S., Stransky, N., Radvanyi, F., Graziani, S. and Barillot, E. (2004). "Selecting Biomedical Data Sources According To User Preferences". Int. Conference on Intelligent Systems in Molecular Biology (ISMB/ECCB), Glasgow, UK.

[CSM+10] Correndo, G., Salvadores, M., Millard, I., Glaser, H. and Shadbolt, N. (2010). "SPARQL query rewriting for implementing data integration over linked data". EDBT Workshops, ACM. pp 1-11.

[CYS+08] Cusick, M. E., Yu, H., Smolyar, A., Venkatesan, K., Carvunis, A. R., Simonis, N., Rual, J. F., Borick, H., Braun, P., Dreze, M.*, et al. (2009). "Literature-curated protein interaction datasets." Nat Methods* ***6(1): 39-46.***

References

[DCB+01] Davidson, S., Crabtree, J., Brunk, B. P., Schug, J., Tannen, V., Overton, G. C. and Stoecker Jr., C. J. (2001). "K2/Kleisli and GUS: Experiments in Integrated Access to Genomic Data Sources." *IBM Systems Journal* ***40(2): 512-531.***

[DGL+09] Detwiler, L., Gatterbauer, W., Louie, B., Suciu, D. and Tarczy-Hornoch, P. (2009). "Integrating and ranking uncertain scientific data". Int. Conf. on Data Engineering, Shanghai, CN, IEEE. pp 1235-1238.

[ES08] Euzenat, J. and Shvaiko, P. (2007). "Ontology matching". Heidelberg, Springer.

[FKM+06] Fagin, R., Kumar, R., Mahdian, M., Sivakumar, D. and Vee, E. (2006). "Comparing partial rankings." *SIAM J. Discrete Mathematics* ***20(3): 628-648.***

[GBR+07] Gerstein, M. B., Bruce, C., Rozowsky, J. S., Zheng, D., Du, J., Korbel, J. O., Emanuelsson, O., Zhang, Z. D., Weissman, S. and Snyder, M. (2007). "What is a gene, post-ENCODE? History and updated definition." *Genome Res* ***17(6): 669-81.***

[GLG06] Goderis, A., Li, P. and Goble, C. (2006). "Workflow discovery: the problem, a case study from e-Science and a graph-based solution". Int. Conf. on Web Services, Chicago, Illinois.

[HBF+09] Hedeler, C., Belhajjame, K., Fernandes, A., Embury, S. and Paton, N. (2009). "Dimensions of dataspaces". 26th British National Conference on Databases. pp 55-66.

[HAA+09] Hudson, T. J., Anderson, W., Artez, A., Barker, A. D., Bell, C., Bernabe, R. R., Bhan, M. K., Calvo, F., Eerola, I., Gerhard, D. S.*, et al. (2010). "International network of cancer genome projects." Nature* ***464(7291): 993-8.***

[HTL07] Hussels, P., Trissl, S. and Leser, U. (2007). "What's new? What's certain? Scoring Search Results in the Presence of Overlapping Data Sources". 4th Workshop on Data Integration in the Life Sciences, Philadelphia, US, Springer, Lecture Notes in Computer Science. pp 231-246.

[IBS08] Ilyas, I. F., Beskales, G. and Soliman, M. A. (2008). "A Survey of Top-k Query Processing Techniques in Relational Database Systems." *ACM Computing Surveys* ***40(4).***

References

[IPSN10] Ioannou, E., Papapetrou, O., Skoutas, D. and Nejdl, W. (2010). "Efficient Semantic-Aware Detection of Near Duplicate Resources". Extended Semantic Web Conference, pp 136-150.

[JAB+08] Jenkinson, A. M., Albrecht, M., Birney, E., Blankenburg, H., Down, T., Finn, R. D., Hermjakob, H., Hubbard, T. J., Jimenez, R. C., Jones, P.*, et al. (2008). "Integrating Biological Data – The Distributed Annotation System ". Data Integration for the Life Sciences, Springer LNBI 5109, Evry, France.*

[JLP+10] Jonquet, C., LePendu, P., Falconer, S. M., Coulet, A., Noy, N. F., Musen, M. A. and Shah, N. H. (2010). "NCBO Resource Index: Ontology-Based Search and Mining of Biomedical Resources". Semantic Web Challenge, at ISWC10, Shanghai, CN.

[KAC+02] Karvounarakis, G., Alexaki, S., Christophides, V., Plexousakis, D. and Scholl, M. (2002). "RQL: a declarative query language for RDF". World Wide Web Conference, Honolulu, Hawaii, USA. pp 592-603.

[KKS04] Kasprzyk, A., Keefe, D., Smedley, D., London, D., Spooner, W., Melsopp, C., Hammond, M., Rocca-Serra, P., Cox, T. and Birney, E. (2004). "EnsMart: a generic system for fast and flexible access to biological data." *Genome Res* ***14(1): 160-9.***

[KTR07] Kirsten, T., Thor, A. and Rahm, E. (2007). "Instance-based matching of large life science ontologies". 4th Int. Conf. on Data Integration in the Life Sciences Philadelphia, USA. pp 172-187

[KP10] Klingstrom, T. and Plewczynski, D. (2010). "Protein-protein interaction and pathway databases, a graphical review." *Brief Bioinform.*

[KT08] Kobayashi, N. and Toyoda, T. (2008). "Statistical search on the Semantic Web." *Bioinformatics* ***24(7): 1002-10.***

[KJ07] Kochut, K. and Janik, M. (2007). "SPARQLeR: Extended Sparql for Semantic Association Discovery". European Conference on the Semantic Web, Innsbruch, Austria.

[KSD+11] Kozhenkov, S., Sedova, M., Dubinina, Y., Gupta, A., Ray, A., Ponomarenko, J. and Baitaluk, M. (2011). "BiologicalNetworks - tools enabling the integration of multi-scale data for the host-pathogen studies." *BMC Syst Biol* ***5: 7.***

References

[KSB10] Kumar, A. M., Shawn, B. and Bertram, L. (2010). "Techniques for efficiently querying scientific workflow provenance graphs". 13th Int. Conf. on Extending Database Technology. Lausanne, Switzerland, ACM.

[LPW+06] Lee, T. J., Pouliot, Y., Wagner, V., Gupta, P., Stringer-Calvert, D. W., Tenenbaum, J. D. and Karp, P. D. (2006). "BioWarehouse: a bioinformatics database warehouse toolkit." *BMC Bioinformatics* ***7: 170.***

[LLF08] Lemoine, F., Labedan, B. and Froidevaux, C. (2008). "GenoQuery: a new querying module for functional annotation in a genomic warehouse." *Bioinformatics* ***24(13): i322-9.***

[LAF+06] Liu, D. T., Abdulla, G. M., Franklin, M. J., Garlick, J. and Miller, M. (2006). "Data-preservation in scientific workflow middleware". Scientific and Statistical Database Management, Vienna, AU, IEEE. pp 49-58.

[MKP+05] Markowitz, V. M., Korzeniewski, F., Palaniappan, K., Szeto, E., Ivanova, N. and Kyrpides, N. C. (2005). "The integrated microbial genomes (IMG) system: a case study in biological data management". 31st Conference on Very Large Databases (VLDB), Trondheim, Norway.

[MPB10] Missier, P., Paton, N. W. and Belhajjame, K. (2010). "Fine-grained and efficient lineage querying of collection-based workflow provenance". Int. Conf. on Extending Database Technology, Lausanne, CH, ACM. pp 299-310.

[MSR+10] Missier, P., Soiland-Reyes, S., Owen, S., Tan, W., Nenadic, A., Dunlop, I., Williams, A., Oinn, T. and Goble, C. (2010). "Taverna, Reloaded". Scientific and Statistical Database Management Systems, Heidelberg, Germany.

[MNF03] Müller, H., Naumann, F. and Freytag, J.-C. (2003). "Data Quality in Genome Databases". Conference on Information Quality, Boston, US.

[NLF99] Naumann, F., Leser, U. and Freytag, J. C. (1999). "Quality-driven Integration of Heterogeneous Information Systems". 25th Conference on Very Large Database Systems, Edinburgh, UK. pp 447-458.

[NQ06] Neumann, E. K. and Quan, D. (2006). "BioDash: a Semantic Web dashboard for drug development". Pac Symp Biocomput, Hawai, US. pp 176-87.

[NSW+09] Noy, N. F., Shah, N. H., Whetzel, P. L., Dai, B., Dorf, M., Griffith, N., Jonquet, C., Rubin, D. L., Storey, M. A., Chute, C. G.*, et al. (2009). "BioPortal: ontologies and integrated data resources at the click of a mouse."* *Nucleic Acids Res* ***37(Web Server issue): W170-3.***

References

[OV99] Oezsu, M. T. and Valduriez, P. (1999). "Principles of Distributed Database Systems". New Jersey, Prentice Hall, Inc.

[ORS+08] Olston, C., Reed, B., Srivastava, U., Kumar, R. and Tomkins, A. (2008). "Pig latin: a not-so-foreign language for data processing". SIGMOD Conference, Vancouver, CD, ACM. pp 1099-1110.

[PRM+07] Post, L. J., Roos, M., Marshall, M. S., van Driel, R. and Breit, T. M. (2007). "A semantic web approach applied to integrative bioinformatics experimentation: a biological use case with genomics data." *Bioinformatics* ***23(22): 3080-7.***

[QL08] Quilitz, B. and Leser, U. (2008). "Querying Distributed RDF Data Sources with SPARQL". European Semantic Web Conference (ESWC), Teneriffa, Spain.

[RKL07] Rahm, E., Kirsten, T. and Lange, J. (2007). "The GeWare data warehouse platform for the analysis of molecular-biological and clinical data." *Journal of Integrative Bioinformatics* ***4(1): 47.***

[RTA+05] Rahm, E., Thor, A., Aumueller, D., Do, H. H., Golovin, N. and Kirsten, T. (2005). "iFuice - Information Fusion utilizing Instance Correspondences and Peer Mappings". WebDB, Baltimore, USA. pp 7-12.

[Rob94a] Robbins, R. J. (1994). "Report of the invitational DOE Workshop on Genome Informatics I: Community Databases." *Journal of Computational Biology* ***3: 173-190.***

[RML05] Rother, K., Michalsky, E. and Leser, U. (2005). "How well are protein structures annotated in annotation databases?" *PROTEINS* ***60(4): 571-576.***

[RCB+07] Ruttenberg, A., Clark, T., Bug, W., Samwald, M., Bodenreider, O., Chen, H., Doherty, D., Forsberg, K., Gao, Y., Kashyap, V.*, et al. (2007). "Advancing translational research with the Semantic Web." BMC Bioinformatics* ***8 Suppl 3: S2.***

[SPLO05] Sarkans, U., Parkinson, H., Lara, G. G., Oezcimen, A., Sharma, A., Abeygunawardena, N., Contrino, S., Holloway, E., Rocca-Serra, P., Mukherjee, G.*, et al. (2005). "The ArrayExpress gene expression database: a software engineering and implementation perspective." Bioinformatics* ***21(8): 1495-501.***

References

[SV09] Savage, C. J. and Vickers, A. J. (2009). "Empirical study of data sharing by authors publishing in PLoS journals." *PLoS One* ***4(9): e7078.***

[SIY06] Shafer, P., Isganitis, T. and Yona, G. (2006). "Hubs of knowledge: using the functional link structure in Biozon to mine for biologically significant entities." *BMC Bioinformatics* ***7: 71.***

[SBJ+09] Shah, N. H., Bhatia, N., Jonquet, C., Rubin, D., Chiang, A. P. and Musen, M. A. (2009). "Comparison of concept recognizers for building the Open Biomedical Annotator." *BMC Bioinformatics* ***10 Suppl 9: S14.***

[SHX+05] Shah, S. P., Huang, Y., Xu, T., Yuen, M. M., Ling, J. and Ouellette, B. F. (2005). "Atlas - a data warehouse for integrative bioinformatics." *BMC Bioinformatics* ***6(1): 34.***

[SMB+04] Shaker, R., Mork, P., Brockenbrough, J. S., Donelson, L. and Tarczy-Hornoch, P. (2004). "The BioMediator System as a Tool for Integrating Biologic Databases on the Web". Information Integration on the Web.

[SAR+08] Smith, B., Ashburner, M., Rosse, C., Bard, J., Bug, W., Ceusters, W., Goldberg, L. J., Eilbeck, K., Ireland, A., Mungall, C. J.*, et al. (2007). "The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration." Nat Biotechnol* ***25(11): 1251-5.***

[SMS+02] Stein, L. D., Mungall, C., Shu, S., Caudy, M., Mangone, M., Day, A., Nickerson, E., Stajich, J. E., Harris, T. W., Arva, A.*, et al. (2002). "The generic genome browser: a building block for a model organism system database." Genome Res* ***12(10): 1599-610.***

[TJM+08] Talukdar, P. P., Jacob, M., Mehmood, M. S., Crammer, K., Ives, Z. G., Pereira, F. and Guha, S. (2008). "Learning to create data-integrating queries." *Proceedings of the VLDB Endowment* ***1(1): 785-796.***

[TRM+05] Trissl, S., Rother, K., Müller, H., Koch, I., Steinke, T., Preissner, R., Frömmel, C. and Leser, U. (2005). "Columba: An Integrated Database of Proteins, Structures, and Annotations." *BMC Bioinformatics* ***6:81.***

[YIF+08] Yu, Y., Isard, M., Fetterly, D., Budiu, M., Erlingsson, Ú., Gunda, P. K. and Currey, J. (2008). "DryadLINQ: A System for General-Purpose Distributed Data-Parallel Computing Using a High-Level Language". Symposium on Operating Systems Design and Implementation.

[ZAV+07] Zacharias, V., Abecker, A., Vrandecic, D., Borgi, I., Braun, S. and Schmidt, A. (2007). "Mind the Web". Workshop on New Forms of Reasoning for the Semantic Web, Busan, Korea.