Data Integration in the Life Sciences

Sarah Cohen-Boulakia

Université Paris Sud, LRI CNRS UMR 8623

cohen@lri.fr

01 69 15 32 16

https://www.lri.fr/~cohen/BIGDATA/biodata-ami2b.html

Introduction

- Understanding Life Sciences

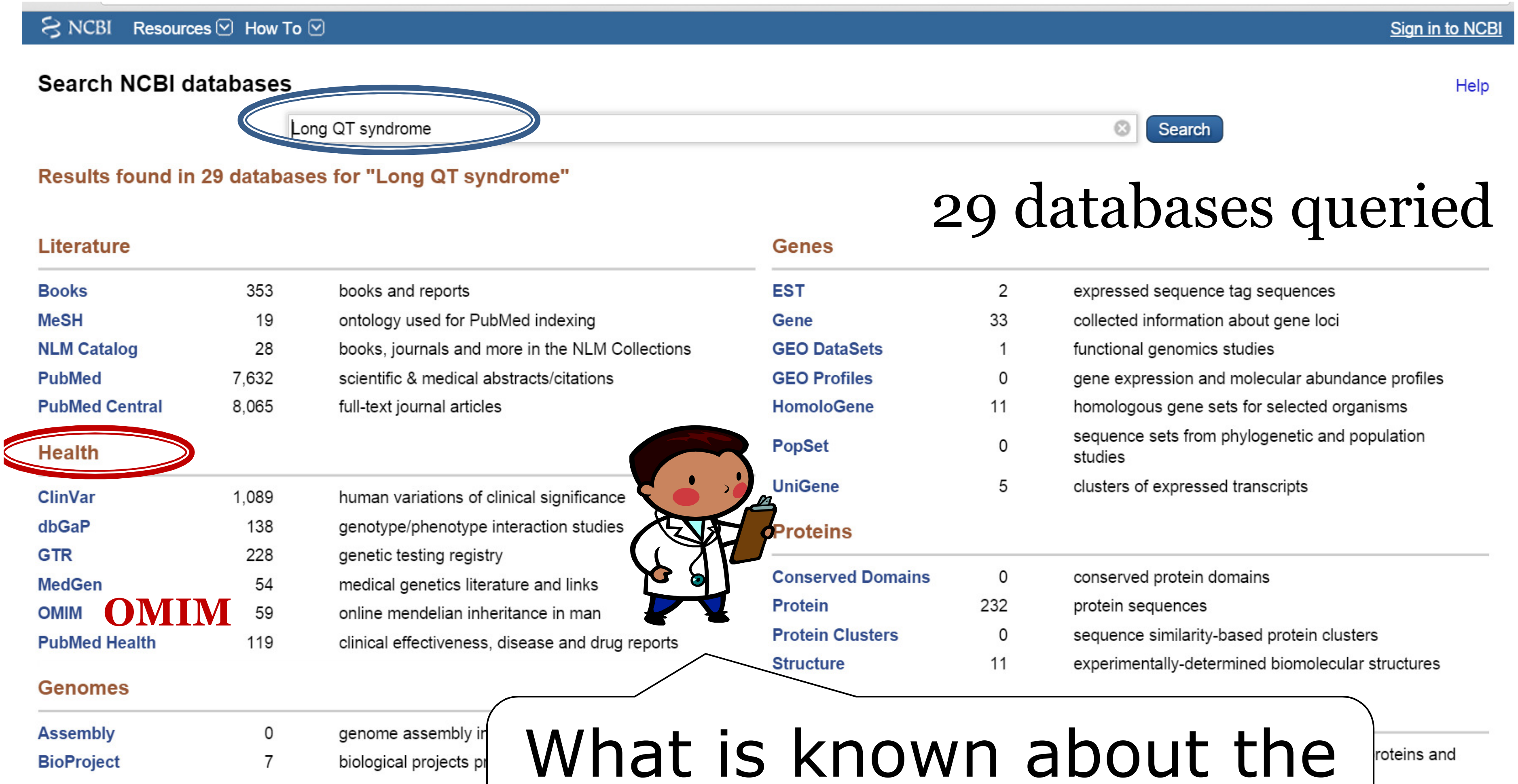
→ Progress in multiple domains: biology, chemistry, maths, computer science…

- Emergence of new technologies: Next generation sequencing,…
  - → Increasing volumes of raw data
  - → All stored in Web data sources
- Raw data are not sufficient
  - ➔ Data Annotated by experts
  - ➔ Bioinformatics analysis of data
  - ➔ New data sources
- Concrete example: Querying NCBI Entrez

  http://www.ncbi.nlm.nih.gov/gquery/

(« Gquery NCBI » on google ☺)

Querying (NCBI Portal)



Search NCBI databases

Long QT syndrome

Results found in 29 databases for "Long QT syndrome"

29 databases queried

Literature

Books	353	books and reports
MeSH	19	ontology used for PubMed indexing
NLM Catalog	28	books, journals and more in the NLM Collections
PubMed	7,632	scientific & medical abstracts/citations
PubMed Central	8,065	full-text journal articles

Health

ClinVar	1,089	human variations of clinical significance
dbGaP	138	genotype/phenotype interaction studies
GTR	228	genetic testing registry
MedGen	54	medical genetics literature and links
OMIM	59	online mendelian inheritance in man
PubMed Health	119	clinical effectiveness, disease and drug reports

OMIM

Genomes

Assembly	0	genome assembly i
BioProject	7	biological projects pr

Genes

EST	2	expressed sequence tag sequences
Gene	33	collected information about gene loci
GEO DataSets	1	functional genomics studies
GEO Profiles	0	gene expression and molecular abundance profiles
HomoloGene	11	homologous gene sets for selected organisms
PopSet	0	sequence sets from phylogenetic and population studies
UniGene	5	clusters of expressed transcripts

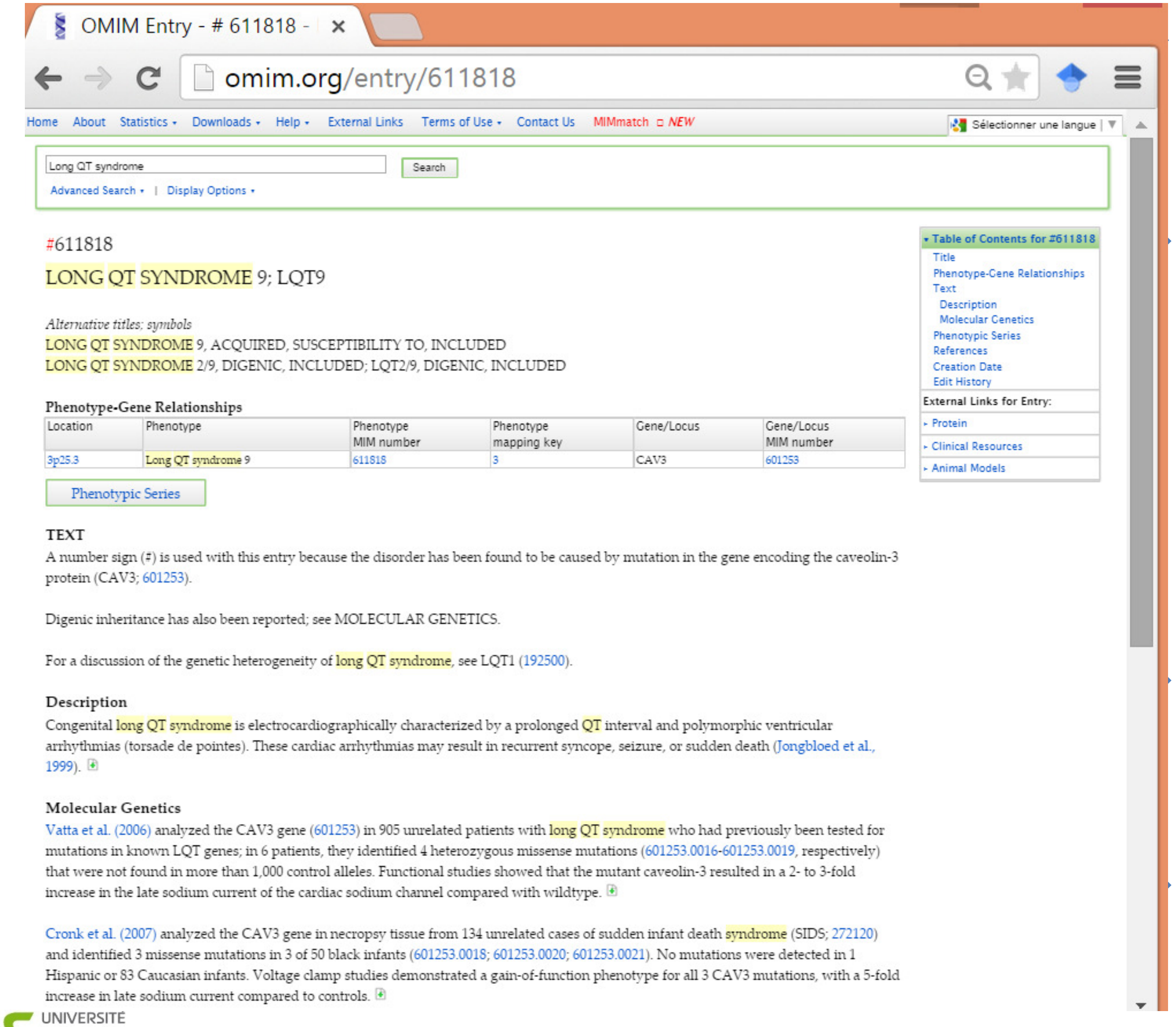
Proteins

Conserved Domains	0	conserved protein domains
Protein	232	protein sequences
Protein Clusters	0	sequence similarity-based protein clusters
Structure	11	experimentally-determined biomolecular structures

roteins and

What is known about the **Long QT syndrome**?

OMIM entry (Long QT)



#611818

LONG QT SYNDROME 9; LQT9

Alternative titles; symbols

LONG QT SYNDROME 9, ACQUIRED, SUSCEPTIBILITY TO, INCLUDED
LONG QT SYNDROME 2/9, DIGENIC, INCLUDED; LQT2/9, DIGENIC, INCLUDED

Phenotype-Gene Relationships

Location	Phenotype	Phenotype MIM number	Phenotype mapping key	Gene/Locus	Gene/Locus MIM number
3p25.3	Long QT syndrome 9	611818	3	CAV3	601253

TEXT

A number sign (#) is used with this entry because the disorder has been found to be caused by mutation in the gene encoding the caveolin-3 protein (CAV3; 601253).

Digenic inheritance has also been reported; see MOLECULAR GENETICS.

For a discussion of the genetic heterogeneity of long QT syndrome, see LQT1 (192500).

Description

Congenital long QT syndrome is electrocardiographically characterized by a prolonged QT interval and polymorphic ventricular arrhythmias (torsade de pointes). These cardiac arrhythmias may result in recurrent syncope, seizure, or sudden death (Jongbloed et al., 1999).

Molecular Genetics

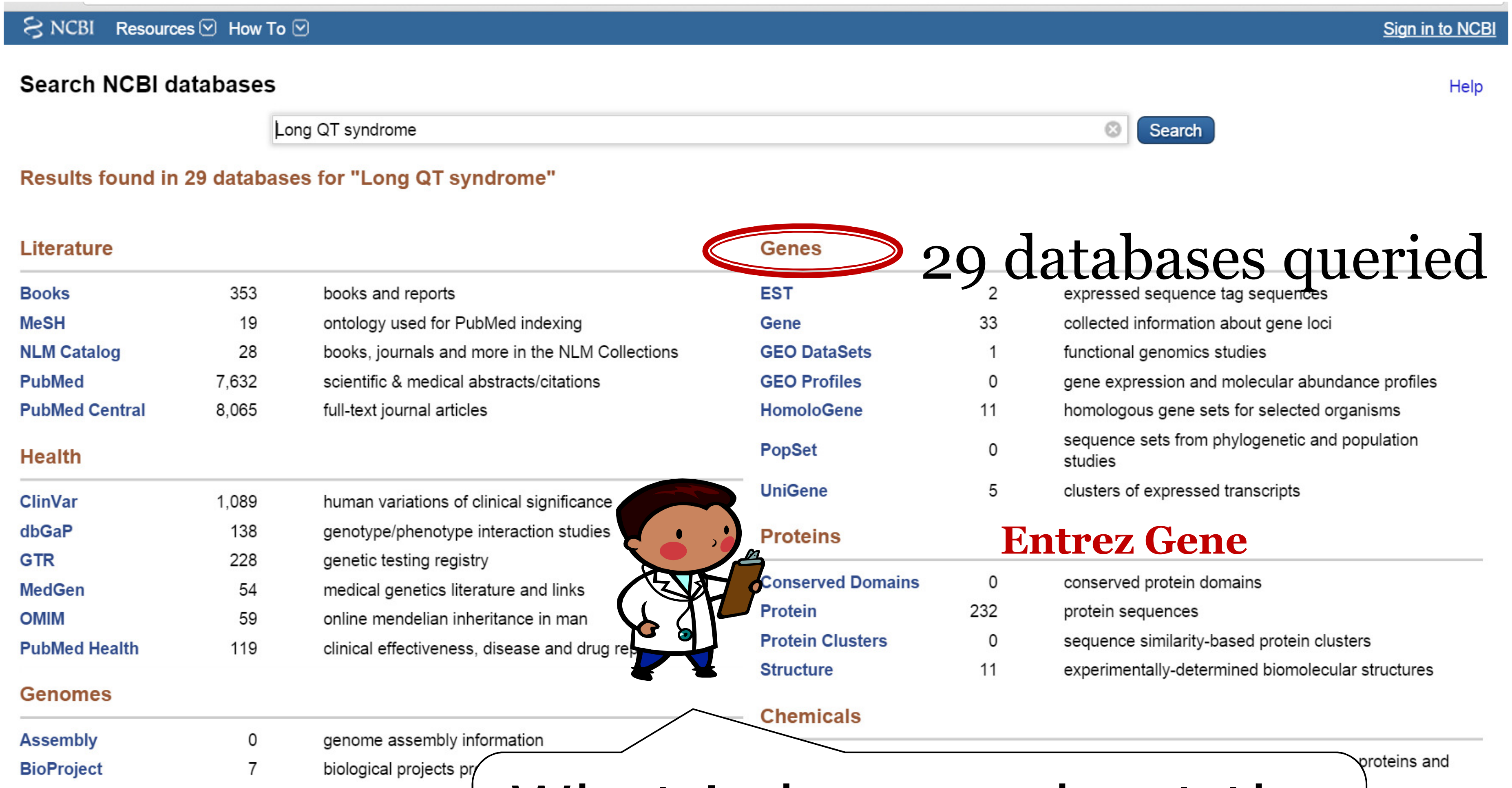
Vatta et al. (2006) analyzed the CAV3 gene (601253) in 905 unrelated patients with long QT syndrome who had previously been tested for mutations in known LQT genes; in 6 patients, they identified 4 heterozygous missense mutations (601253.0016-601253.0019, respectively) that were not found in more than 1,000 control alleles. Functional studies showed that the mutant caveolin-3 resulted in a 2- to 3-fold increase in the late sodium current of the cardiac sodium channel compared with wildtype.

Cronk et al. (2007) analyzed the CAV3 gene in necropsy tissue from 134 unrelated cases of sudden infant death syndrome (SIDS; 272120) and identified 3 missense mutations in 3 of 50 black infants (601253.0018; 601253.0020; 601253.0021). No mutations were detected in 1 Hispanic or 83 Caucasian infants. Voltage clamp studies demonstrated a gain-of-function phenotype for all 3 CAV3 mutations, with a 5-fold increase in late sodium current compared to controls.

http://omim.org/entry/611818

- Several pages of (structured) text describing the Long QT9 form of the disease
- Manual annotions only (few data)
- *Curated data (physicians)*

Querying (NCBI Portal)



Search NCBI databases

Long QT syndrome

Results found in 29 databases for "Long QT syndrome"

29 databases queried

Literature

Books	353	books and reports
MeSH	19	ontology used for PubMed indexing
NLM Catalog	28	books, journals and more in the NLM Collections
PubMed	7,632	scientific & medical abstracts/citations
PubMed Central	8,065	full-text journal articles

Health

ClinVar	1,089	human variations of clinical significance
dbGaP	138	genotype/phenotype interaction studies
GTR	228	genetic testing registry
MedGen	54	medical genetics literature and links
OMIM	59	online mendelian inheritance in man
PubMed Health	119	clinical effectiveness, disease and drug re

Genomes

Assembly	0	genome assembly information
BioProject	7	biological projects pr

Genes

EST	2	expressed sequence tag sequences
Gene	33	collected information about gene loci
GEO DataSets	1	functional genomics studies
GEO Profiles	0	gene expression and molecular abundance profiles
HomoloGene	11	homologous gene sets for selected organisms
PopSet	0	sequence sets from phylogenetic and population studies
UniGene	5	clusters of expressed transcripts

Proteins

Entrez Gene

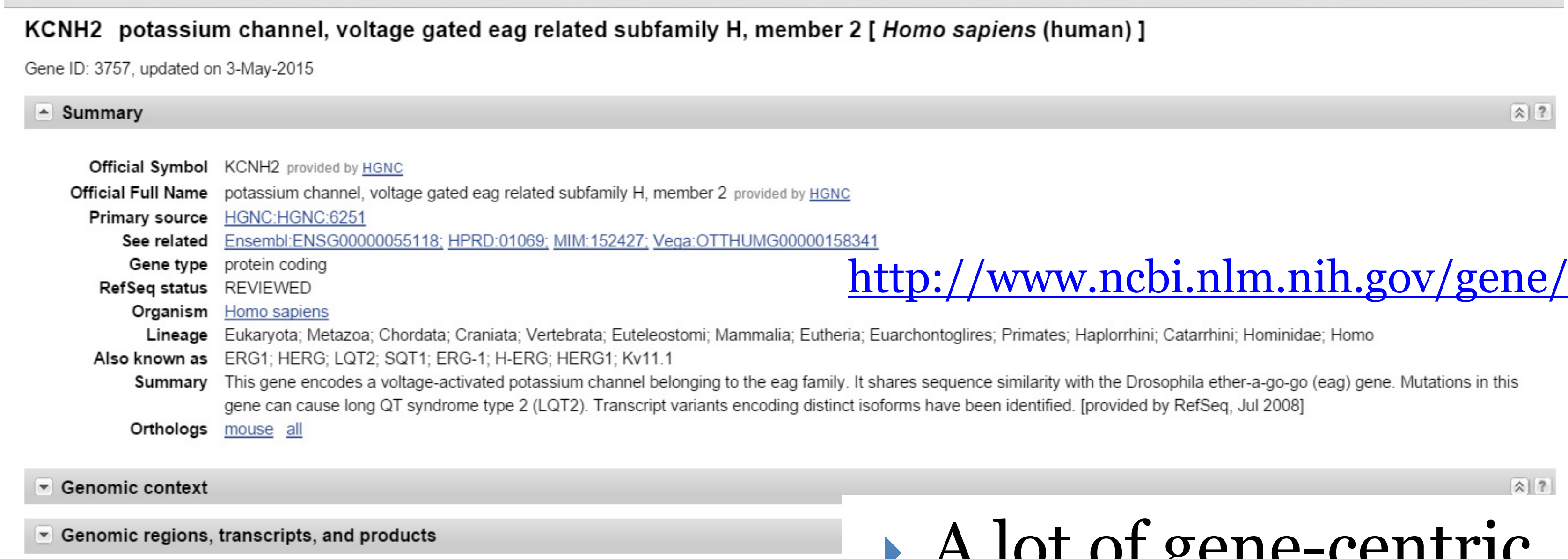
Conserved Domains	0	conserved protein domains
Protein	232	protein sequences
Protein Clusters	0	sequence similarity-based protein clusters
Structure	11	experimentally-determined biomolecular structures

Chemicals

proteins and

What is known about the **Long QT syndrome**?

One Entrez Gene entry (Long QT)



KCNH2 potassium channel, voltage gated eag related subfamily H, member 2 [*Homo sapiens* (human)]

Gene ID: 3757, updated on 3-May-2015

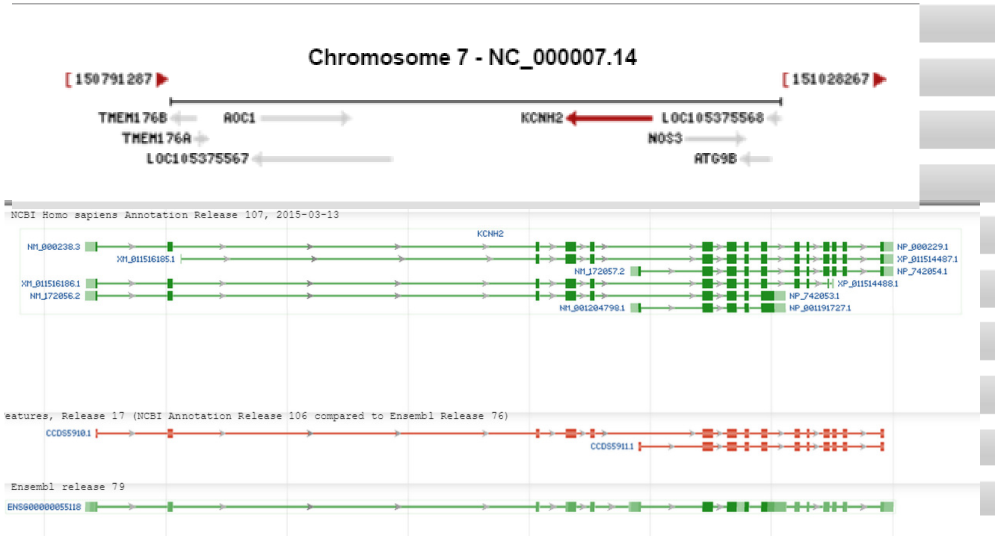
Summary

Official Symbol	KCNH2 provided by HGNC
Official Full Name	potassium channel, voltage gated eag related subfamily H, member 2 provided by HGNC
Primary source	HGNC:HGNC:6251
See related	Ensembl:ENSG00000055118; HPRD:01069; MIM:152427; Vega:OTTHUMG00000158341
Gene type	protein coding
RefSeq status	REVIEWED
Organism	Homo sapiens
Lineage	Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini; Catarrhini; Hominidae; Homo
Also known as	ERG1; HERG; LQT2; SQT1; ERG-1; H-ERG; HERG1; Kv11.1
Summary	This gene encodes a voltage-activated potassium channel belonging to the eag family. It shares sequence similarity with the Drosophila ether-a-go-go (eag) gene. Mutations in this gene can cause long QT syndrome type 2 (LQT2). Transcript variants encoding distinct isoforms have been identified. [provided by RefSeq, Jul 2008]
Orthologs	mouse all

Genomic context

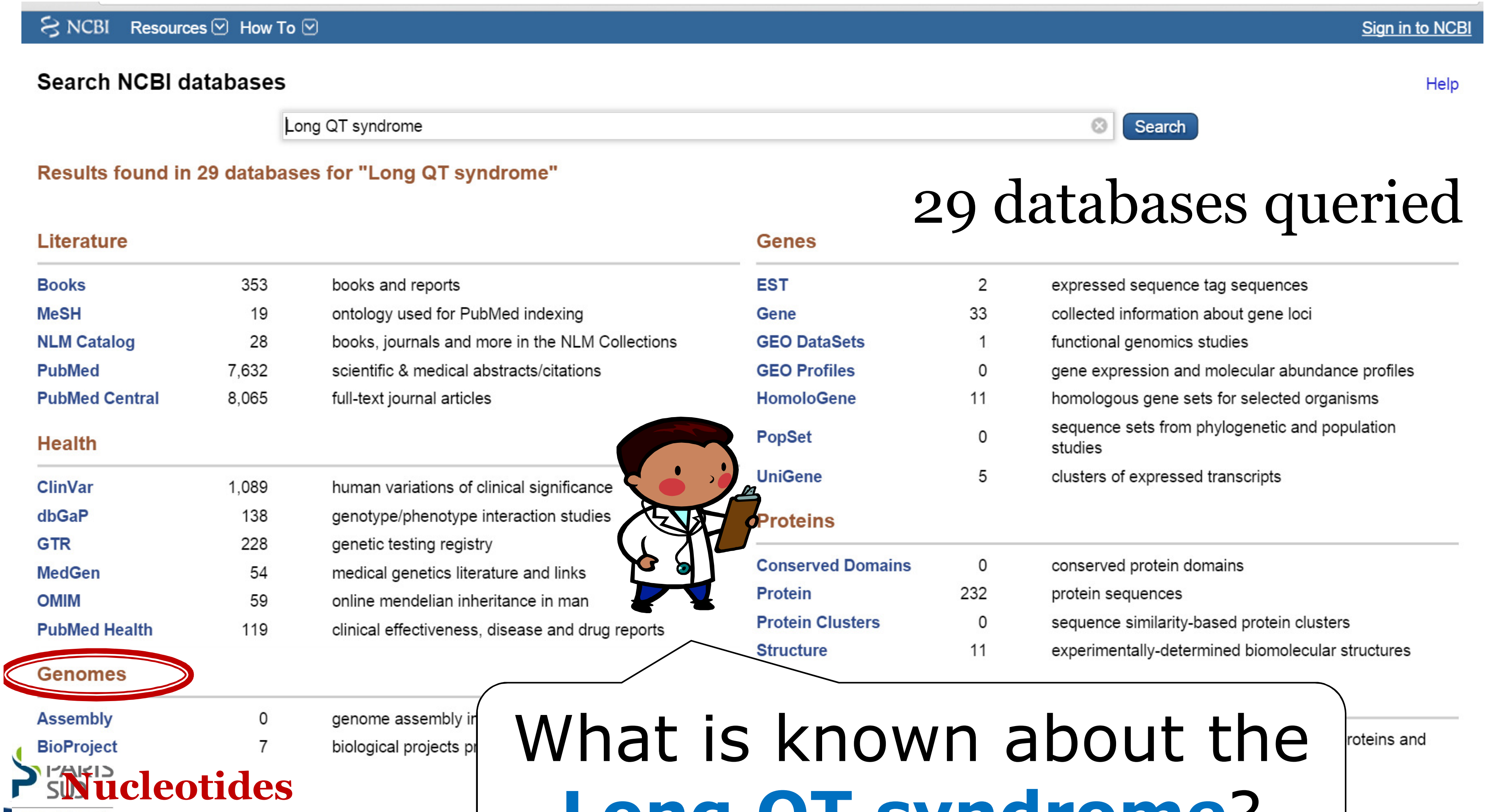
Genomic regions, transcripts, and products

http://www.ncbi.nlm.nih.gov/gene/3757



- A lot of gene-centric information
- Genomic context, genomic regions…
- *Gathering of data*

Querying (NCBI Portal)



Search NCBI databases

Long QT syndrome

Results found in 29 databases for "Long QT syndrome"

29 databases queried

Literature

Books	353	books and reports
MeSH	19	ontology used for PubMed indexing
NLM Catalog	28	books, journals and more in the NLM Collections
PubMed	7,632	scientific & medical abstracts/citations
PubMed Central	8,065	full-text journal articles

Health

ClinVar	1,089	human variations of clinical significance
dbGaP	138	genotype/phenotype interaction studies
GTR	228	genetic testing registry
MedGen	54	medical genetics literature and links
OMIM	59	online mendelian inheritance in man
PubMed Health	119	clinical effectiveness, disease and drug reports

Genomes

Assembly	0	genome assembly i
BioProject	7	biological projects pr

Nucleotides

Genes

EST	2	expressed sequence tag sequences
Gene	33	collected information about gene loci
GEO DataSets	1	functional genomics studies
GEO Profiles	0	gene expression and molecular abundance profiles
HomoloGene	11	homologous gene sets for selected organisms
PopSet	0	sequence sets from phylogenetic and population studies
UniGene	5	clusters of expressed transcripts

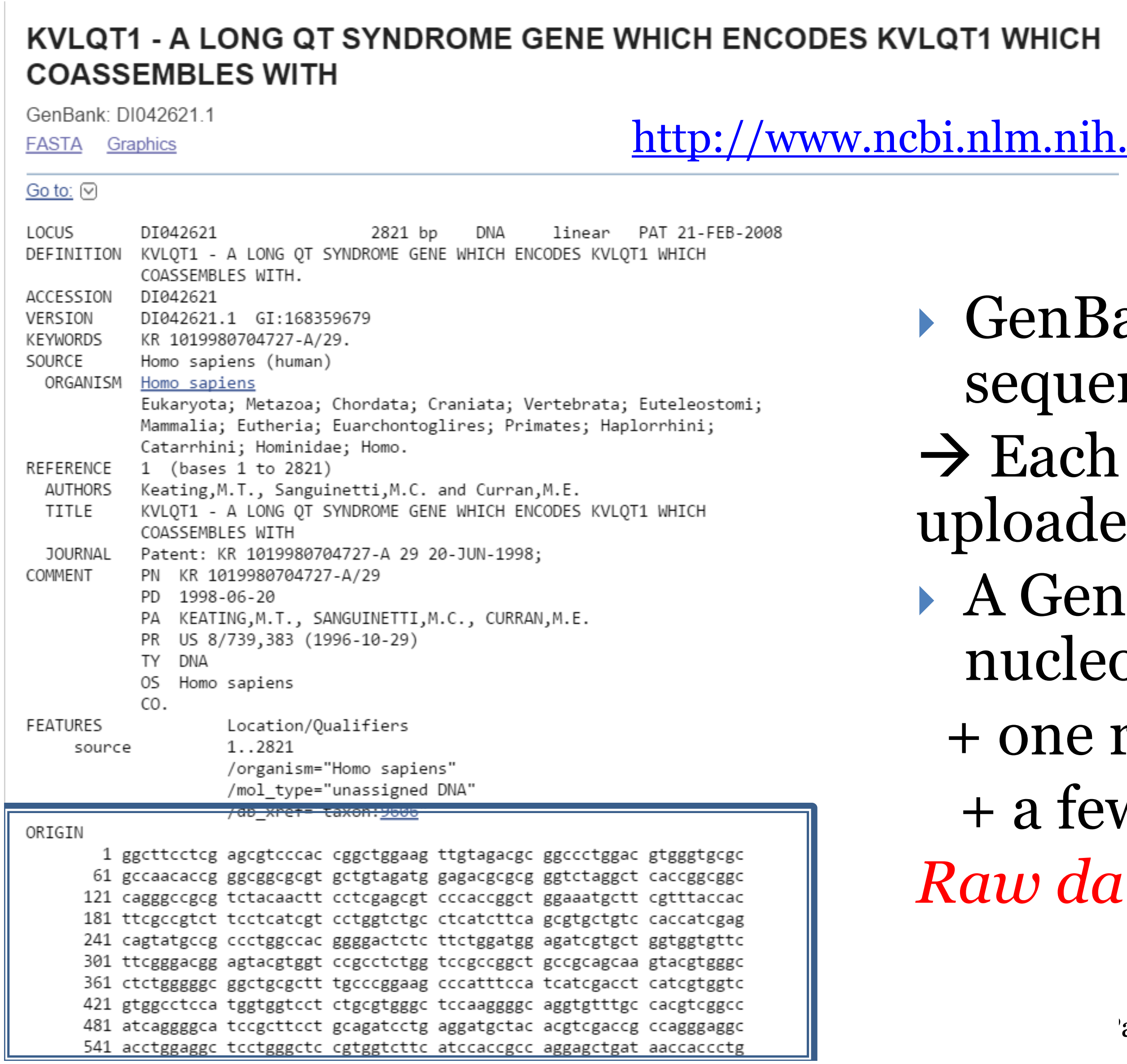
Proteins

Conserved Domains	0	conserved protein domains
Protein	232	protein sequences
Protein Clusters	0	sequence similarity-based protein clusters
Structure	11	experimentally-determined biomolecular structures

roteins and

What is known about the **Long QT syndrome**?

One GenBank entry (Long QT)



KVLQT1 - A LONG QT SYNDROME GENE WHICH ENCODES KVLQT1 WHICH COASSEMBLES WITH

GenBank: DI042621.1

FASTA Graphics

Go to: ⌄

```
LOCUS       DI042621                2821 bp    DNA     linear   PAT 21-FEB-2008
DEFINITION  KVLQT1 - A LONG QT SYNDROME GENE WHICH ENCODES KVLQT1 WHICH
            COASSEMBLES WITH.
ACCESSION   DI042621
VERSION     DI042621.1  GI:168359679
KEYWORDS    KR 1019980704727-A/29.
SOURCE      Homo sapiens (human)
  ORGANISM  Homo sapiens
            Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
            Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini;
            Catarrhini; Hominidae; Homo.
REFERENCE   1  (bases 1 to 2821)
  AUTHORS   Keating,M.T., Sanguinetti,M.C. and Curran,M.E.
  TITLE     KVLQT1 - A LONG QT SYNDROME GENE WHICH ENCODES KVLQT1 WHICH
            COASSEMBLES WITH
  JOURNAL   Patent: KR 1019980704727-A 29 20-JUN-1998;
COMMENT     PN  KR 1019980704727-A/29
            PD  1998-06-20
            PA  KEATING,M.T., SANGUINETTI,M.C., CURRAN,M.E.
            PR  US 8/739,383 (1996-10-29)
            TY  DNA
            OS  Homo sapiens
            CO.
FEATURES             Location/Qualifiers
     source          1..2821
                     /organism="Homo sapiens"
                     /mol_type="unassigned DNA"
                     /db_xref="taxon:9606"
ORIGIN
        1 ggcttcctcg agcgtcccac cggctggaag ttgtagacgc ggccctggac gtgggtgcgc
       61 gccaacaccg ggcggcgcgt gctgtagatg gagacgcgcg ggtctaggct caccggcggc
      121 cagggccgcg tctacaactt cctcgagcgt cccaccggct ggaaatgctt cgtttaccac
      181 ttcgccgtct tcctcatcgt cctggtctgc ctcatcttca gcgtgctgtc caccatcgag
      241 cagtatgccg ccctggccac ggggactctc ttctggatgg agatcgtgct ggtggtgttc
      301 ttcgggacgg agtacgtggt ccgcctctgg tccgccggct gccgcagcaa gtacgtgggc
      361 ctctgggggc ggctgcgctt tgcccggaag cccatttcca tcatcgacct catcgtggtc
      421 gtggcctcca tggtggtcct ctgcgtgggc tccaaggggc aggtgtttgc cacgtcggcc
      481 atcaggggca tccgcttcct gcagatcctg aggatgctac acgtcgaccg ccagggaggc
      541 acctggaggc tcctgggctc cgtggtcttc atccaccgcc aggagctgat aaccaccctg
```

GenBank id
http://www.ncbi.nlm.nih.gov/nuccore/DI010834.1

- GenBank is a deposit of sequences

→ Each sequence must be uploaded to GenBank

- A GenBank entry = nucleotide sequence

  + one reference

  + a few comments

Raw data

Wrap-up

- Even if scientists use a portal, querying biological databases is not easy…
- High heterogeneity of the sources
  - Very different kinds of contents
    - Free text (OMIM), semi-structured data (GenBank)…
    - From free text to controled vocabulary (free text to Ontologies)
- Diverses levels of data quality
  - From automatically obtained (EntrezGene) to manually annotated (OMIM)
- Different Biological entites
  - OMIM : Disease
  - Entrez Gene : Gene
  - GenBank : Nucleotides

➔ A bit of history…

Data Integration for the Life Sciences in 1994

- Robbins, R. J. (1994). "Report of the invitational DOE Workshop on **Genome Informatics I: Community Databases**." [Rob94a]
  - DOE funded large parts of the **Human Genome Project**

- "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, (...), 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 in 1994

- 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".*
  - Only needs 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

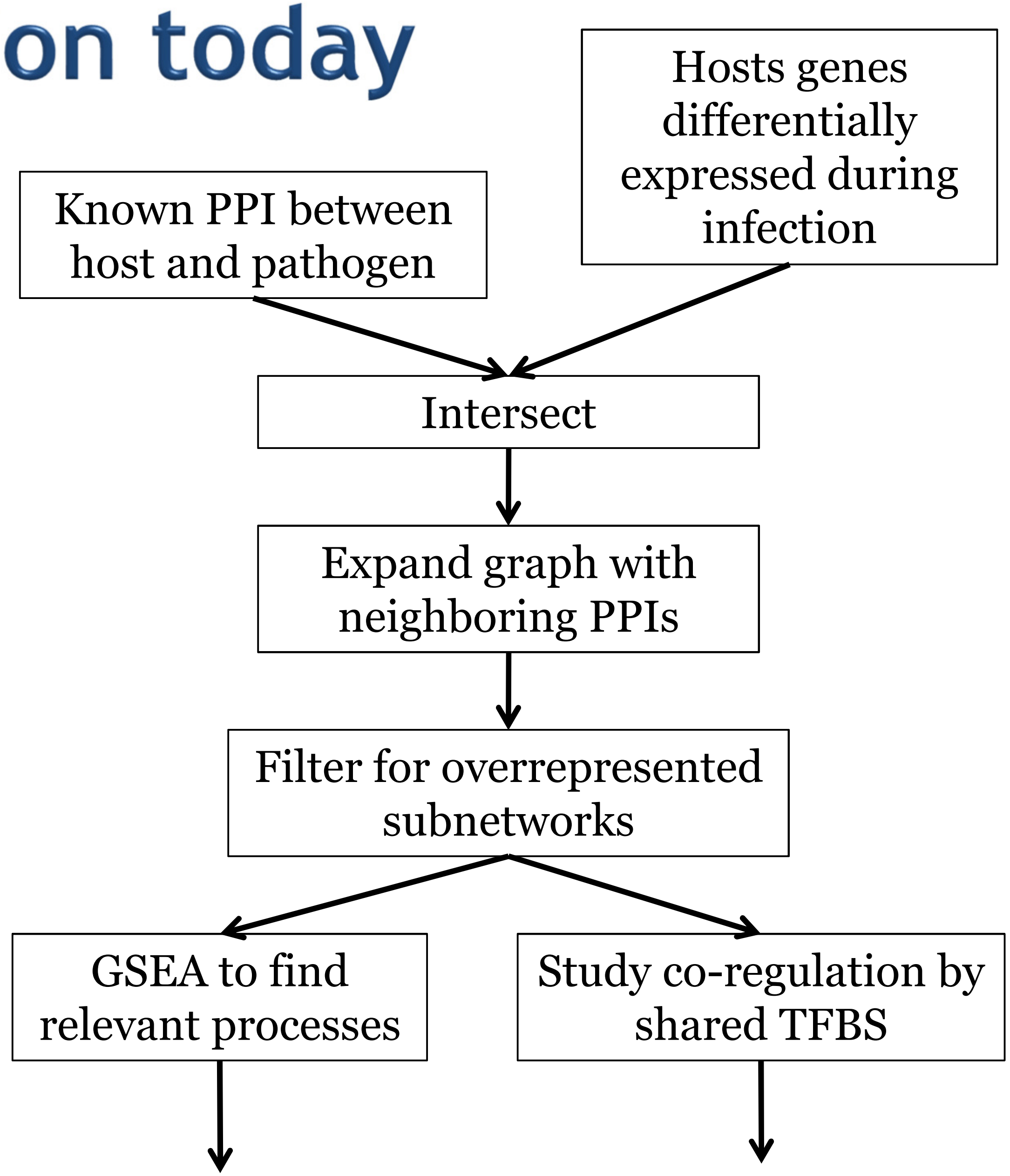
Take Home Message

- The classical problems are all there already

- Distributed information
- Semantic heterogeneity
- Scientific uncertainty and evolving concepts
- Naming conventions on the object level
- Naming conventions on the concept level
- Inclusion of non-standard processing

Data Integration today

- 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 PPI (e.g. transient signaling)


Known PPI between host and pathogen

Hosts genes differentially expressed during infection

Intersect

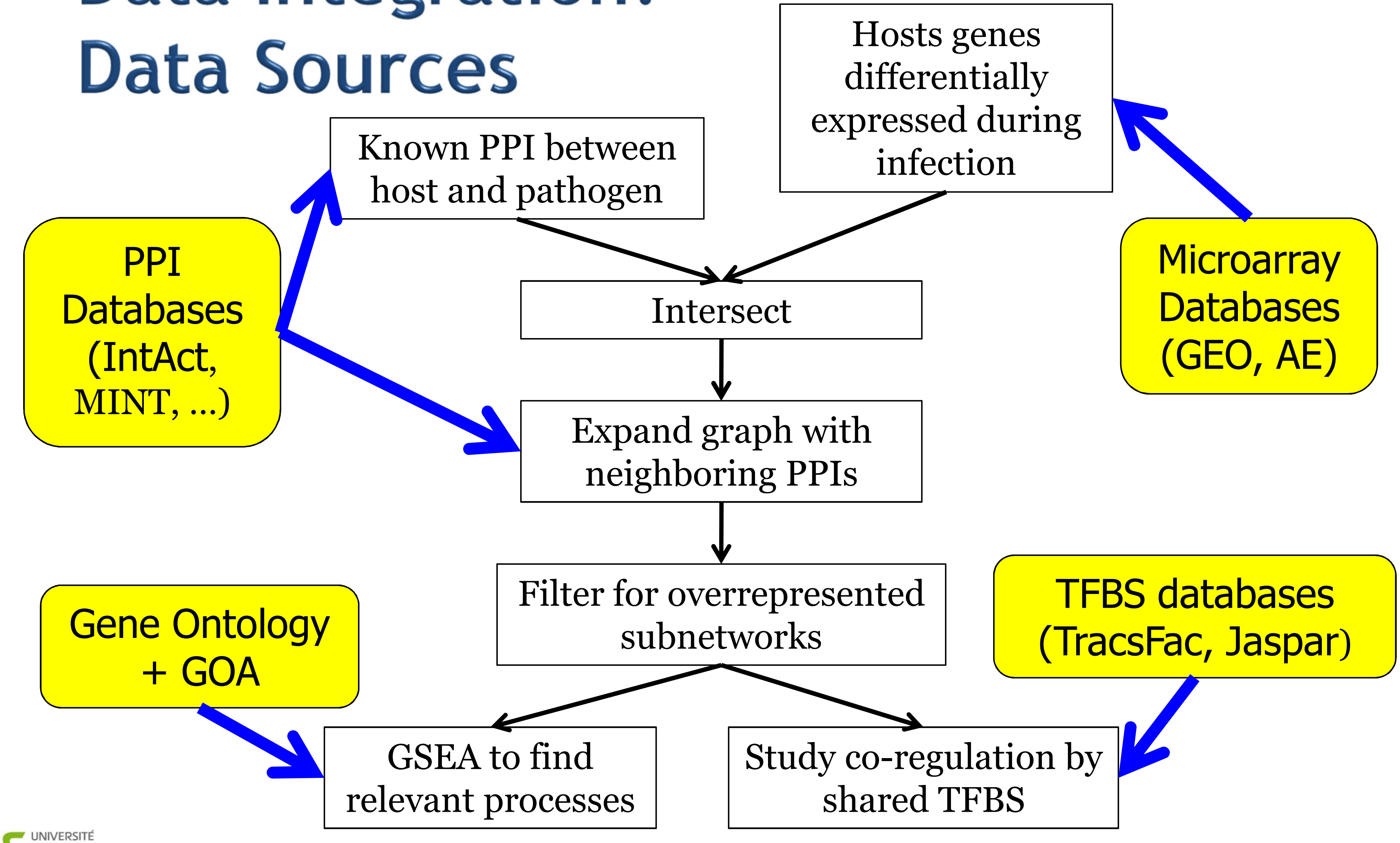
Expand graph with neighboring PPIs

Filter for overrepresented subnetworks

GSEA to find relevant processes

Study co-regulation by shared TFBS

Data Integration?
Data Sources



Known PPI between host and pathogen

Hosts genes differentially expressed during infection

PPI Databases (IntAct, MINT, ...)

Intersect

Microarray Databases (GEO, AE)

Expand graph with neighboring PPIs

Filter for overrepresented subnetworks

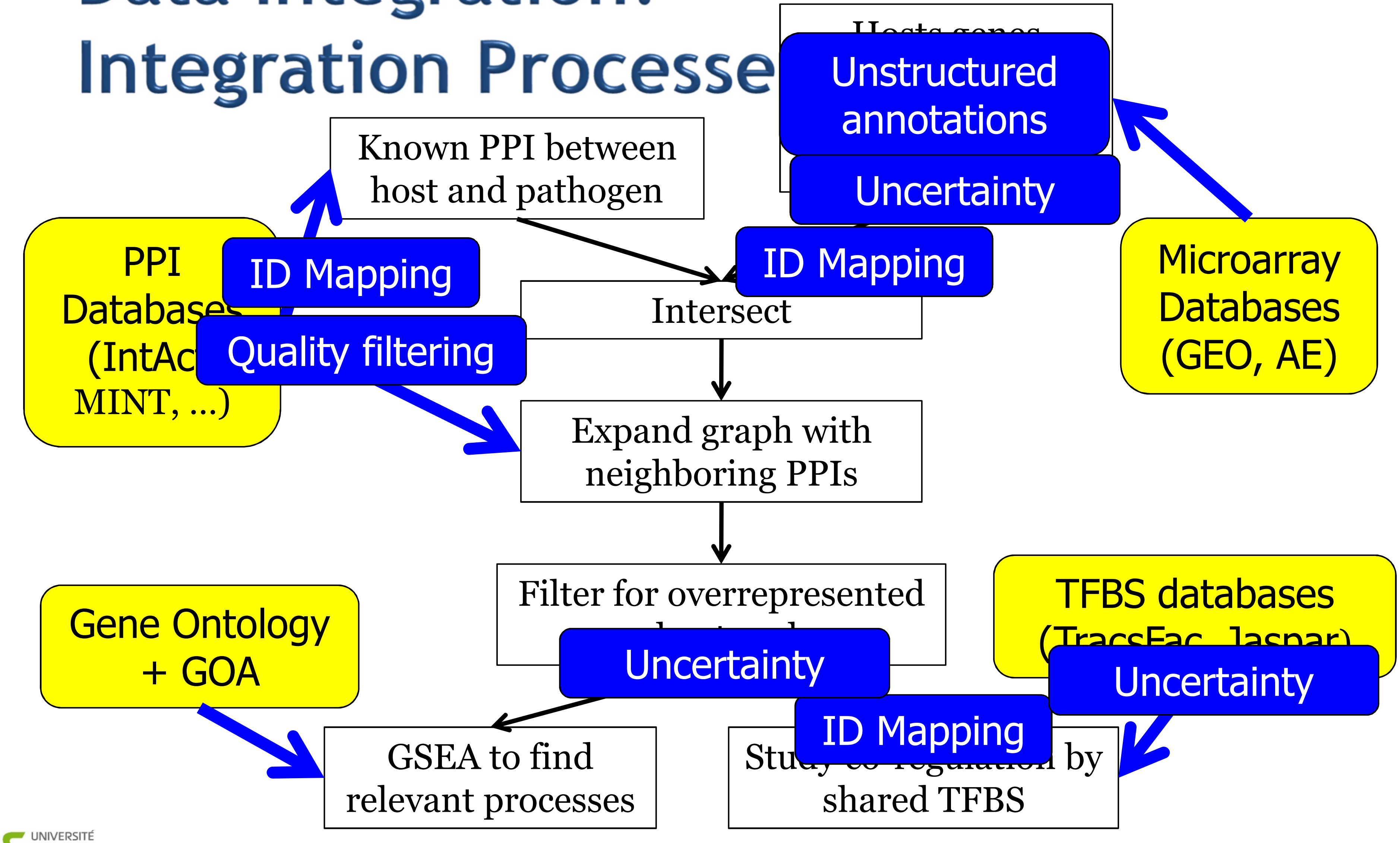
Gene Ontology + GOA

TFBS databases (TracsFac, Jaspar)

GSEA to find relevant processes

Study co-regulation by shared TFBS

Data Integration? Integration Processe



Hosts genes

Unstructured annotations

Uncertainty

Known PPI between host and pathogen

PPI Databases (IntAc MINT, …)

ID Mapping

Quality filtering

ID Mapping

Intersect

Microarray Databases (GEO, AE)

Expand graph with neighboring PPIs

Filter for overrepresented

Uncertainty

Gene Ontology + GOA

TFBS databases (TracsFac, Jaspar)

Uncertainty

ID Mapping

GSEA to find relevant processes

Stu by shared TFBS

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, development of ontologies

- The diversity of the sources has increased a lot
  - Inclusion of quality as a first-class citizen
  - Ranking of integrated search results

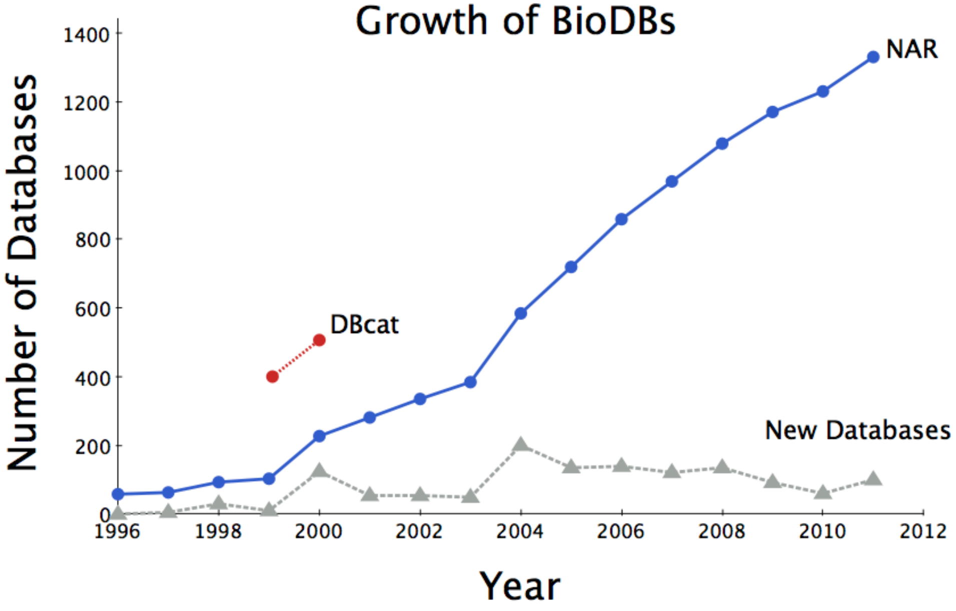
- The complexity of the questions to be answered has increased a lot
  - Integration requires analysis and analysis requires integration
  - Scientific workflows

This Tutorial

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

Are BDB Distributed?

- > 1,000 different databases
  - Plus many data sets that are not stored in a DB
  - e.g. Supplementary material
- Content is highly redundant
  - Replica (sequence databases)
  - Large unintentional overlaps (KEGG – Reactome)
  - Large intentional overlaps (species specific data)
  - Some databases mostly copy from other sources
- Content may be curated during copying
  - Inconsistencies



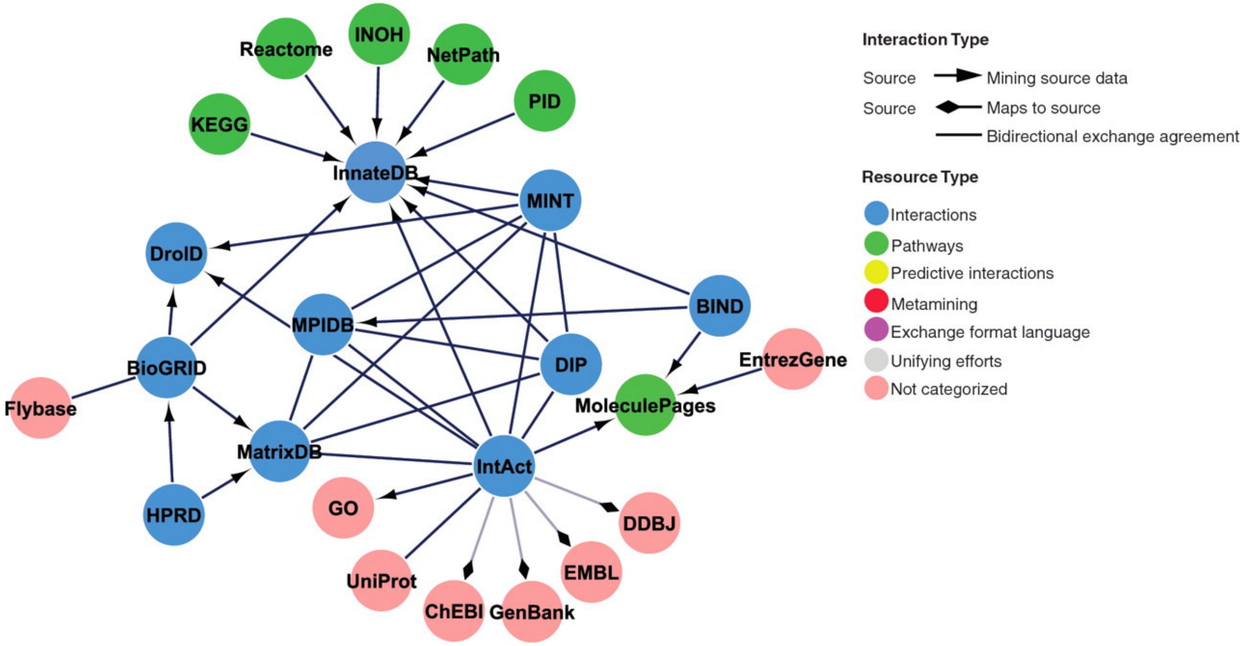
Number of existing (circles) and new databases (triangles) are plotted from 1996 to 2011. New databases are difference between the number of existing databases for each year. DBcat (red) is shown with NAR (blue) counts.

Copyright Geospiza 2011

Each year, the NAR (Nucleic Acid research) journal has a database issue, listing the databases available

Extreme Example: Protein-Protein-Interactions

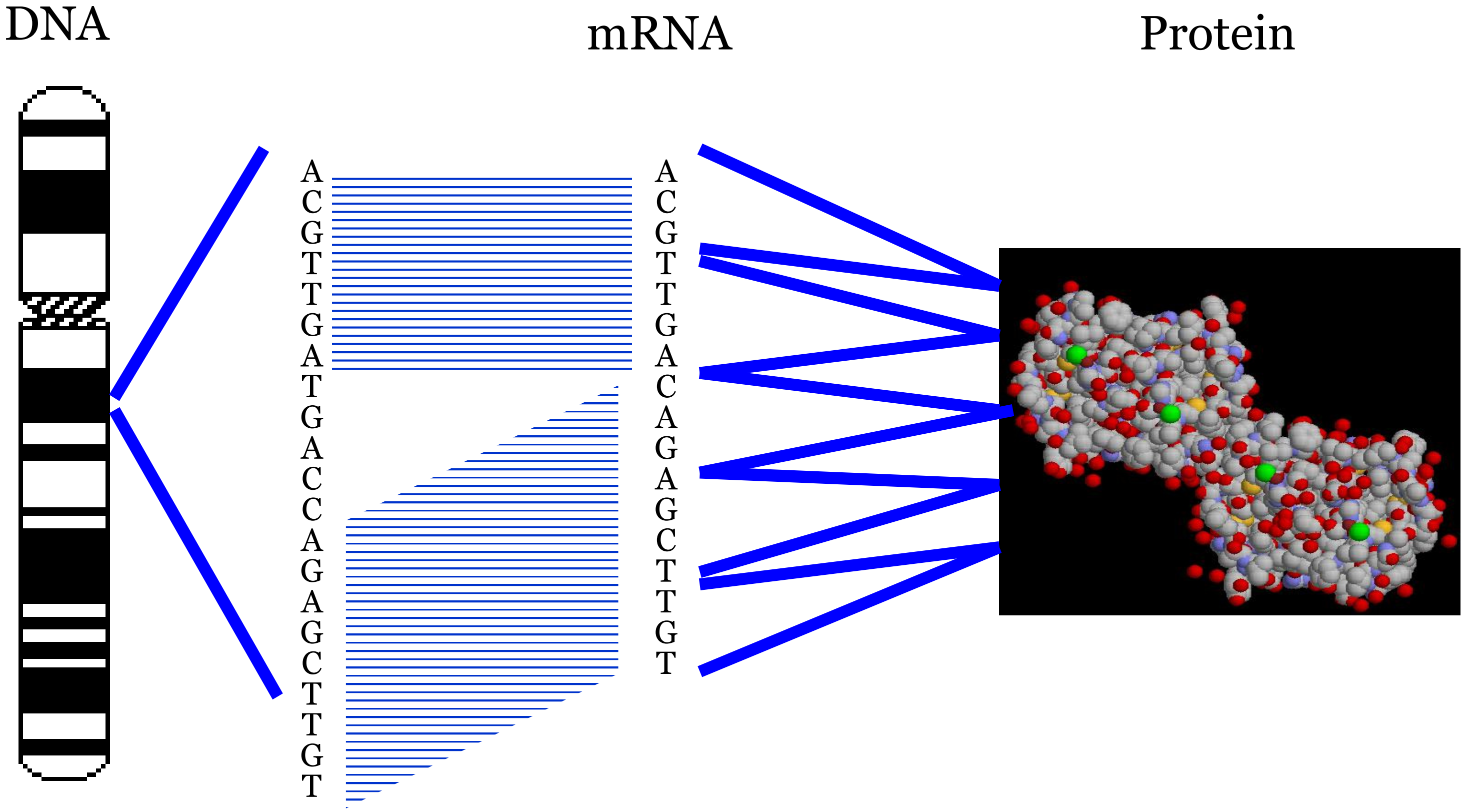
- There are >500 BDBs related to PPI and pathways
  - See http://www.pathguide.org
- Manually created "source" DBs



Are BDB Heterogeneous?

- Technical heterogeneity: a bit
  - Web services, HTML forms, …
- Syntactic heterogeneity: not 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 and states)
  - Definition of object types are heterogeneous, scientifically uncertain, and change over time
  - Schema element names are heterogeneous
  - Metadata often is not available in sufficient depth
- As usual – distribution creates (semantic) heterogeneity

What is a Gene (1)?


DNA
mRNA
Protein
A C G T T G A T G A C C A G A G C T T G T
A C G T T G A C A G A G C T T G T

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

What is a Gene (2)?

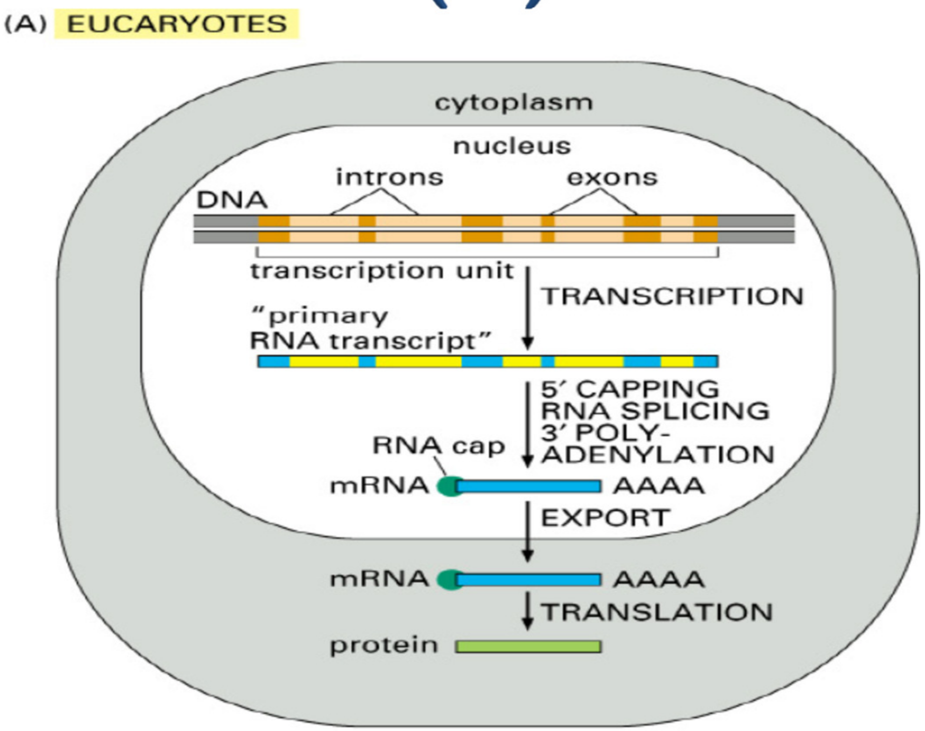
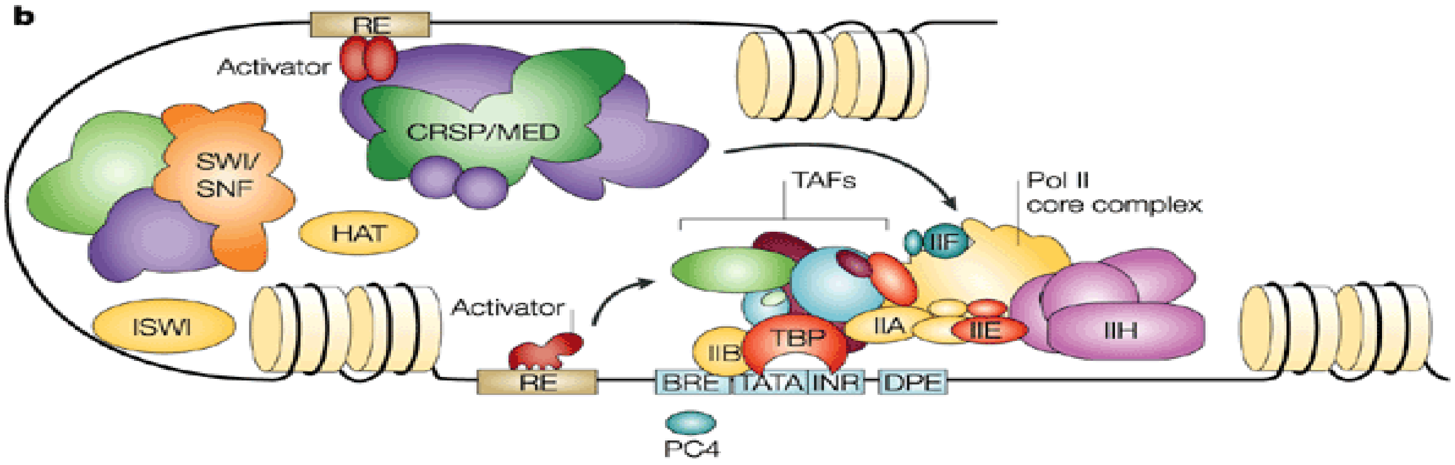


Figure 6–21 part 1 of 2. Molecular Biology of the Cell, 4th Edition.

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

What is a Gene (3)?



Nature Reviews | Molecular Cell Biology

- Like Def.2, plus parts of the sequence downstream that is necessary to regulate transcription of the gene

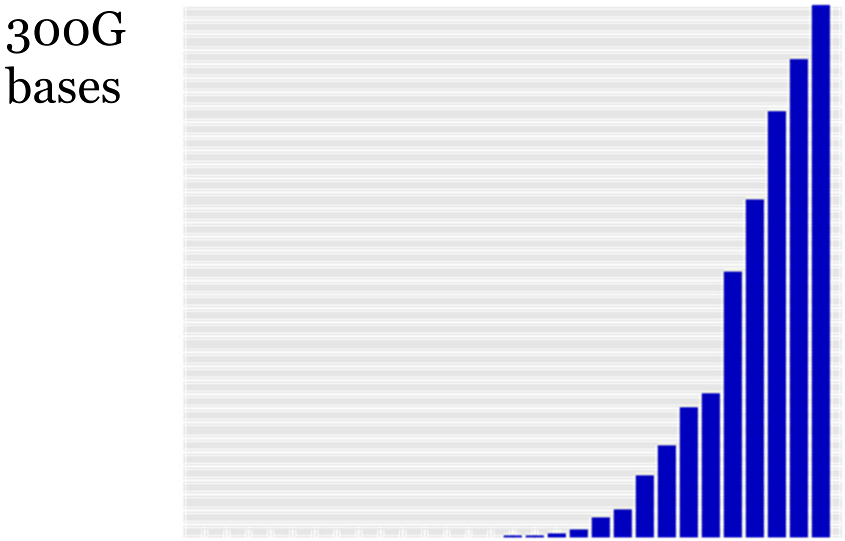
What is a Gene (4)? [GBR+07]

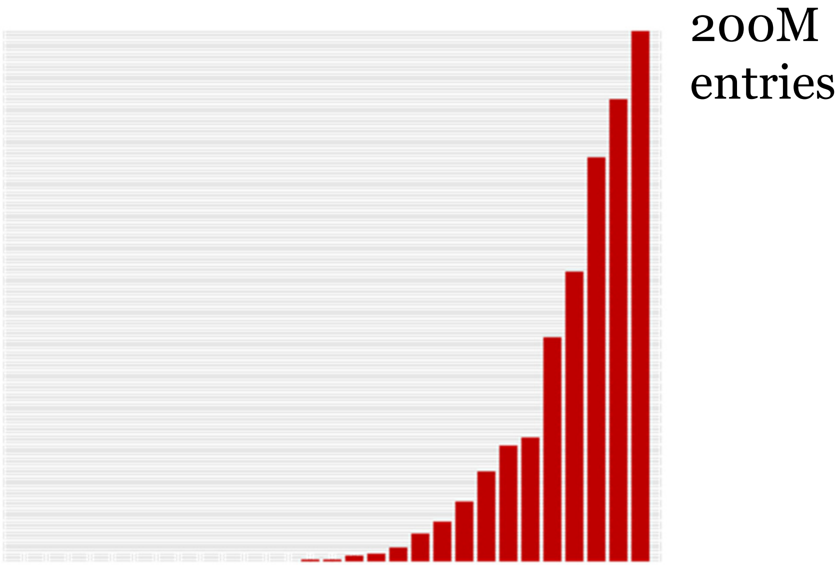
- The same gene?
  - Genes may generate different assemblies (differential splicing)
  - Gene duplications in a genome
  - The „same“ gene in another organism
  - Mutation of a gene
  - Genes with a different start site
- A gene?
  - Pseudo genes (never transcribed, yet highly similar)
  - Non-coding genes
  - miRNA (25 bases!)
- Gene definitions change(d) over centuries, decades, and … last years

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
  - Much copy-and-paste problems (predictions become reality)
- Recall: Most BDB are filled from (high-throughput) experiment
  - Experiments that are not perfect
  - Measurements that are highly context-dependent
  - Performing the same experiment again will produce different results
- Recall: Things change a lot over time
  - New techniques
  - New knowledge

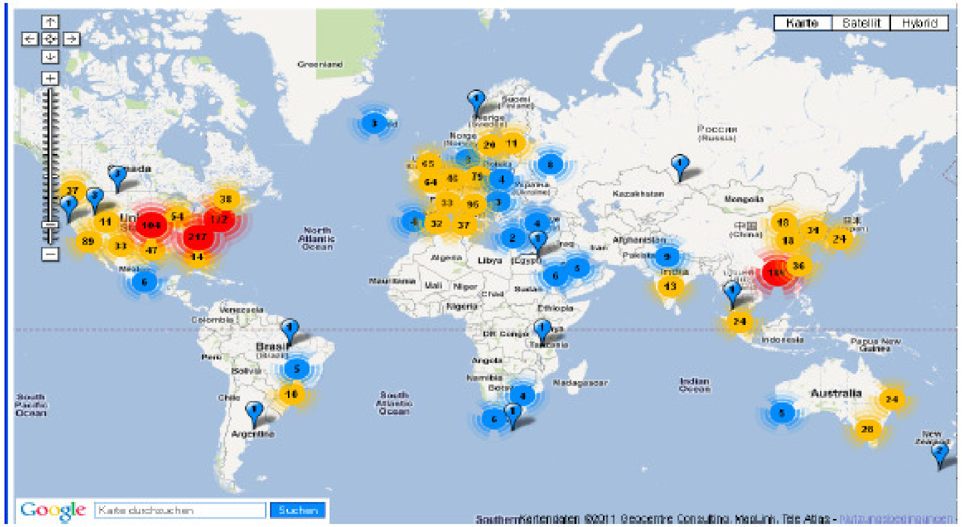
Are Data Volumes huge?





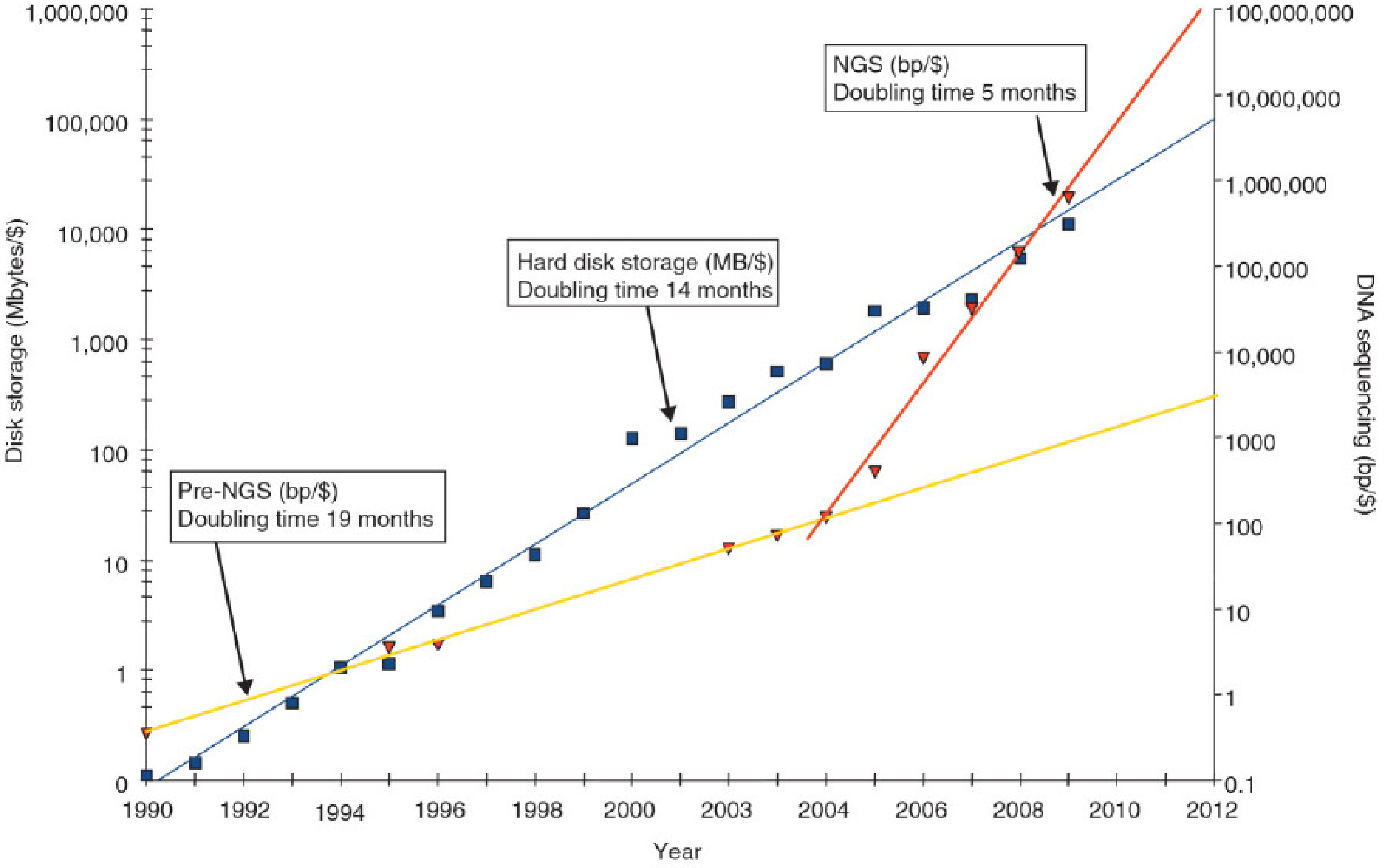
- All of EMBL now has ~150 TB (zipped), ENSEMBL has ~1TB (MySQL dump), UniProt has ~5GB (zipped)
- Probably 90% of the 1300 DB's in NAR have <1GB
- All secondary databases have "little" data
- Primary data explodes due to Next Generation Sequencing

Sequencing has become commodity



- Sequencing dozens of genomes/exomes feasible for any mid-size research project
- In 5 years: Hundreds of genomes
  - (Inter-)national projects: 100.000+ genomes
- Access to genomes is crucial: Bioinformatics goes medical
  - “Translational Bioinformatics”

Data Tsunami



Stein, L. D. (2010). *Genome Biol*

Is Reproducibiliy an Issue?

Is Reproducibiliy an Issue?
Studies on reproducibility

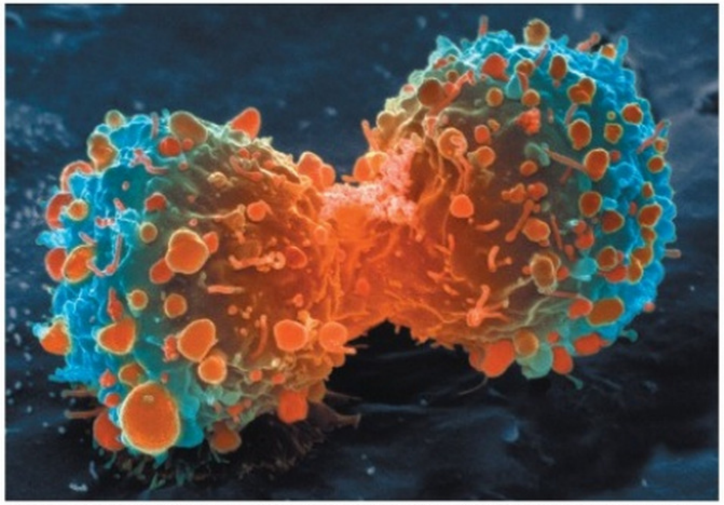
- Nekrutenko & Taylor, Nature Genetics (2012)
  - 50 papers published in 2011 using the Burrows-Wheeler Aligner for Mapping Illumina reads.
  - 31/50 (62%) provide no information
    - no version of the tool + no parameters used + no exact genomic reference seq.
  - 7/50 (14%) provide all the necessary details

Is Reproducibiliy an Issue?
Studies on reproducibility

- Nekrutenko & Taylor, Nature Genetics (2012)
  - 50 papers published in 2011 using the Burrows-Wheeler Aligner for Mapping Illumina reads.
  - 31/50 (62%) provide no information
    - no version of the tool + no parameters used + no exact genomic reference seq.
  - 7/50 (14%) provide all the necessary details
- Alsheikh-Ali et al, PLoS one (2011)
  - 10 papers in the top-50 IF journals → 500 papers (publishers)
    - 149 (30%) were not subject to any data availability policy (0% made their data available)
    - Of the remaining 351 papers
      - 208 papers (59%) did not adhere to the data availability instructions
      - 143 make a statement of *willingness* to share
      - 47 papers (9%) deposited full primary raw data online

Impacts of irreproducibility…

S. GSCHMEISSNER/SPL



Many landmark findings in preclinical oncology research are not reproducible, in part because of inadequate cell lines and animal models.

Raise standards for preclinical cancer research

C. Glenn Begley and Lee M. Ellis propose how methods, publications and incentives must change if patients are to benefit.

Efforts over the past decade to characterize the genetic alterations in human cancers have led to a better understanding of molecular drivers of this complex set of diseases. Although we in the cancer field hoped that this would lead to more effective drugs, historically, our ability trials in oncology have the highest failure rate compared with other therapeutic areas. Given the high unmet need in oncology, it is understandable that barriers to clinical development may be lower than for other disease areas, and a larger number of drugs with suboptimal preclinical validation will investigators must reassess their approac translating discovery research into gre clinical success and impact.

Many factors are responsible for the h failure rate, notwithstanding the inh ently difficult nature of this disease. C tainly, the limitations of preclinical tc

47/53 “landmark” publications could not be replicated

[Begley, Ellis Nature, 483, 2012]

Must try harder

Too many sloppy mistakes are creeping into scientific papers. at the data — and at themselves.

Error prone

Biologists must realize the pitfalls massive amounts of data.

If a job is worth doing, it is worth doing twice

Researchers and funding agencies need to put a premium on ensuring that results are reproducible, argues **Jonathan F. Russell**.

The case for open computer programs

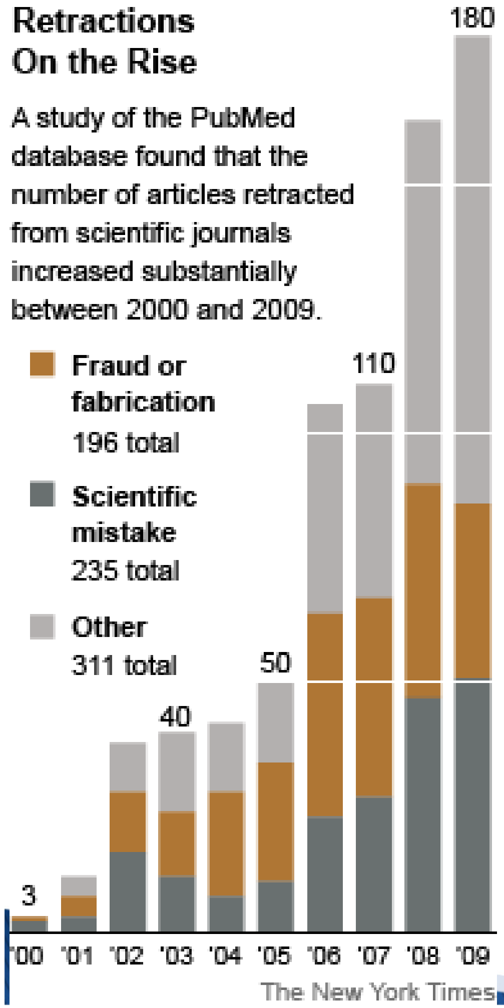
Six red flags for suspect work

C. Glenn Begley explains how to recognize the preclinical papers in which the data won’t stand up.

Know when your numbers are significant

Impacts of irreproducibility (cont.)

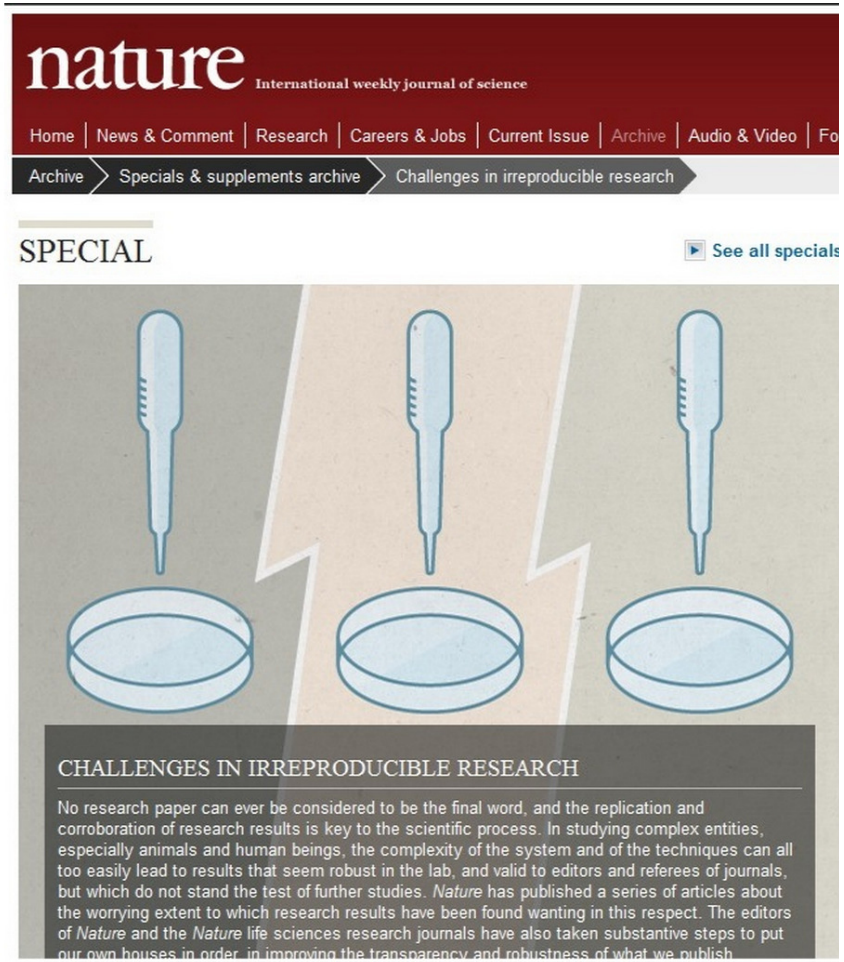
- Attacks on authors, editors, reviewers, publishers, funders…



Retractions On the Rise

A study of the PubMed database found that the number of articles retracted from scientific journals increased substantially between 2000 and 2009.

The New York Times



nature
International weekly journal of science
Home
News & Comment
Research
Careers & Jobs
Current Issue
Archive
Audio & Video
Fo
Archive
Specials & supplements archive
Challenges in irreproducible research
See all specials

SPECIAL

CHALLENGES IN IRREPRODUCIBLE RESEARCH

No research paper can ever be considered to be the final word, and the replication and corroboration of research results is key to the scientific process. In studying complex entities, especially animals and human beings, the complexity of the system and of the techniques can all too easily lead to results that seem robust in the lab, and valid to editors and referees of journals, but which do not stand the test of further studies. *Nature* has published a series of articles about the worrying extent to which research results have been found wanting in this respect. The editors of *Nature* and the *Nature* life sciences research journals have also taken substantive steps to put our own houses in order, in improving the transparency and robustness of what we publish

http://www.nature.com/nature/focus/reproducibility/index.html

→ *Nature* checklist

→ *Science* requirements for data and code availability

Wrap-up

- Integration more necessary than ever in the Life Sciences
- Biological data sources
  - Increasingly numerous, heterogeneous, distributed,…

→ Provenance is needed to understand and interpret data, ranking techniques has to be developed

- Breadth of scientific questions increases
- Reproducibility is a major issue

  → Scientific workflows

- Data sources contains errors
- Need standardization

  → Ontologies

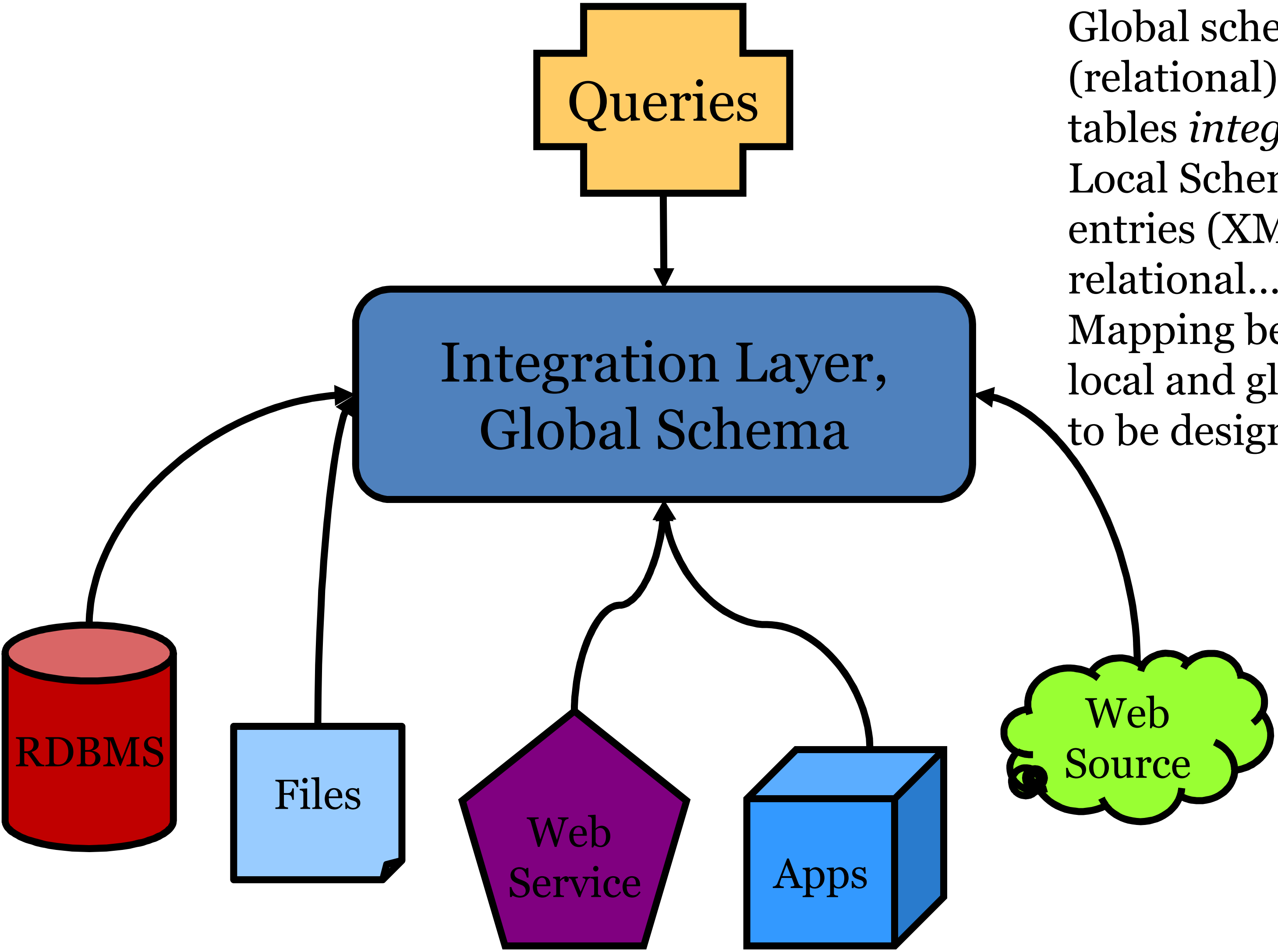
This Tutorial

- Part I – Data Integration for the Life Sciences
  - Biological data & biological databases
  - Some Myths, some Truths

- Part II – Integration -- Presence

- Part III – Current Trends and Conclusions

Integration -- Classical View

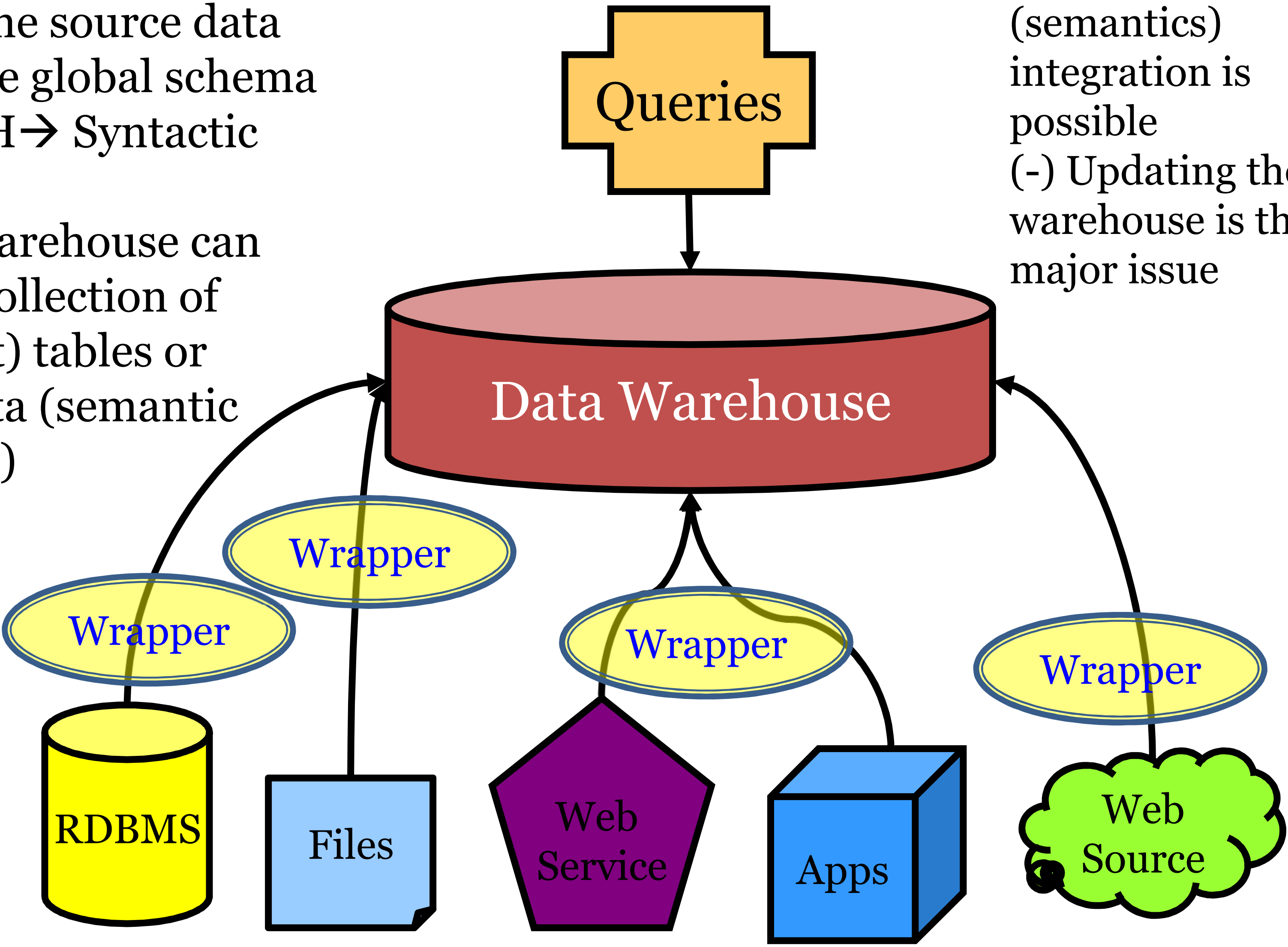


Global schema (relational) = set of tables *integrated*
Local Schema = set of entries (XML, relational…)
Mapping between the local and global schemas to be designed

Classical View - Data Warehouse

- Wrappers transform the format of the source data sets into the global schema of the DWH→ Syntactic integration
- The data warehouse can contain a collection of (redundant) tables or curated data (semantic integration)

(+) Fine (semantics) integration is possible
(-) Updating the warehouse is the major issue



The Presence

XML + Python + MySQL

- Or better

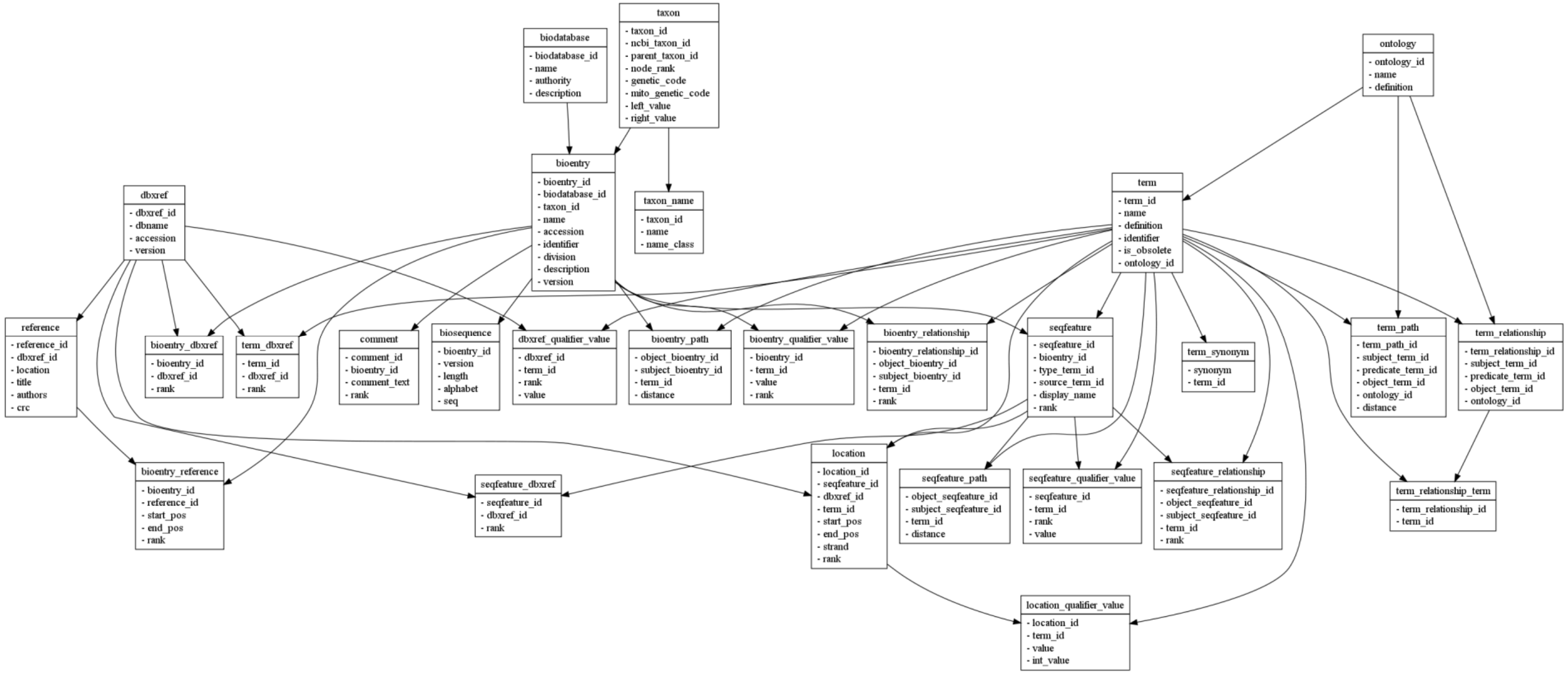
XML +
(Perl | Java | Python) +
(MySQL | Oracle | PostGreSql)

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

The Presence

- Architecture
  - Portals are used a lot but do not perform *tight* integration
  - Federated systems are mostly dead
    - Despite frequent papers stating the opposite
    - Survival in some niches: DAS, some mash-ups (no queries)
  - "Data Warehouses" approaches everywhere
- Semantic integration
  - No schema matching, little query rewriting
  - Performed manually (in custom-written wrappers)
- Several systems up-and-running integrating dozens of sources
  - Freshness in the presence of data cleansing remains a hard problem

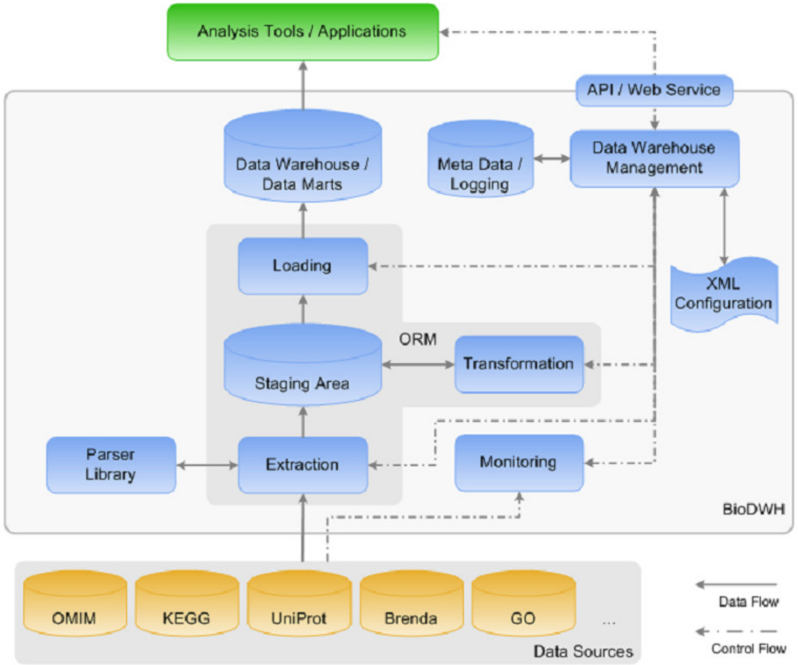
BioSQL [http://www.biosql.org/] and Bio*



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

BioWarehouse [LPW+06]

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



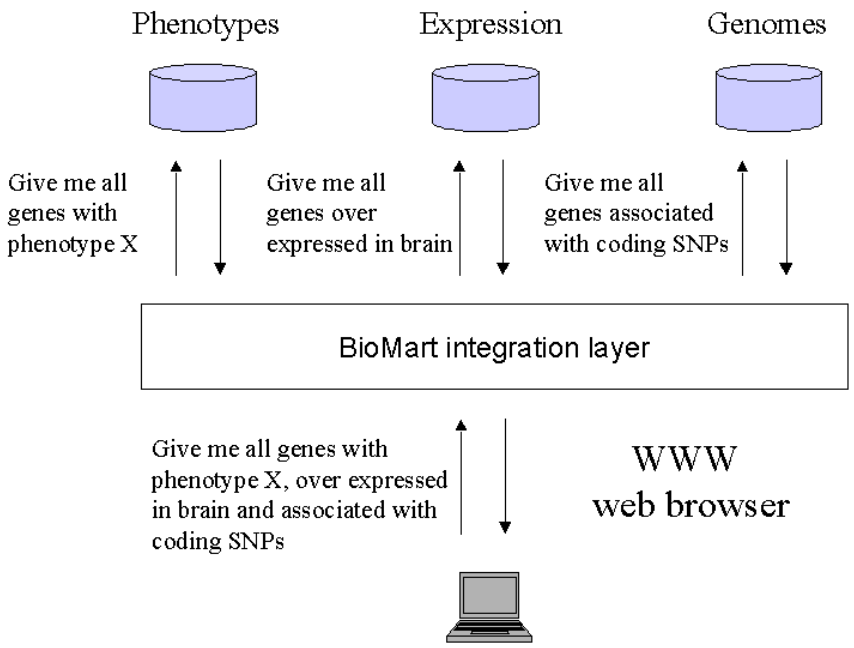
GMOD [SMS+02]

- "**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, … InterMine

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

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

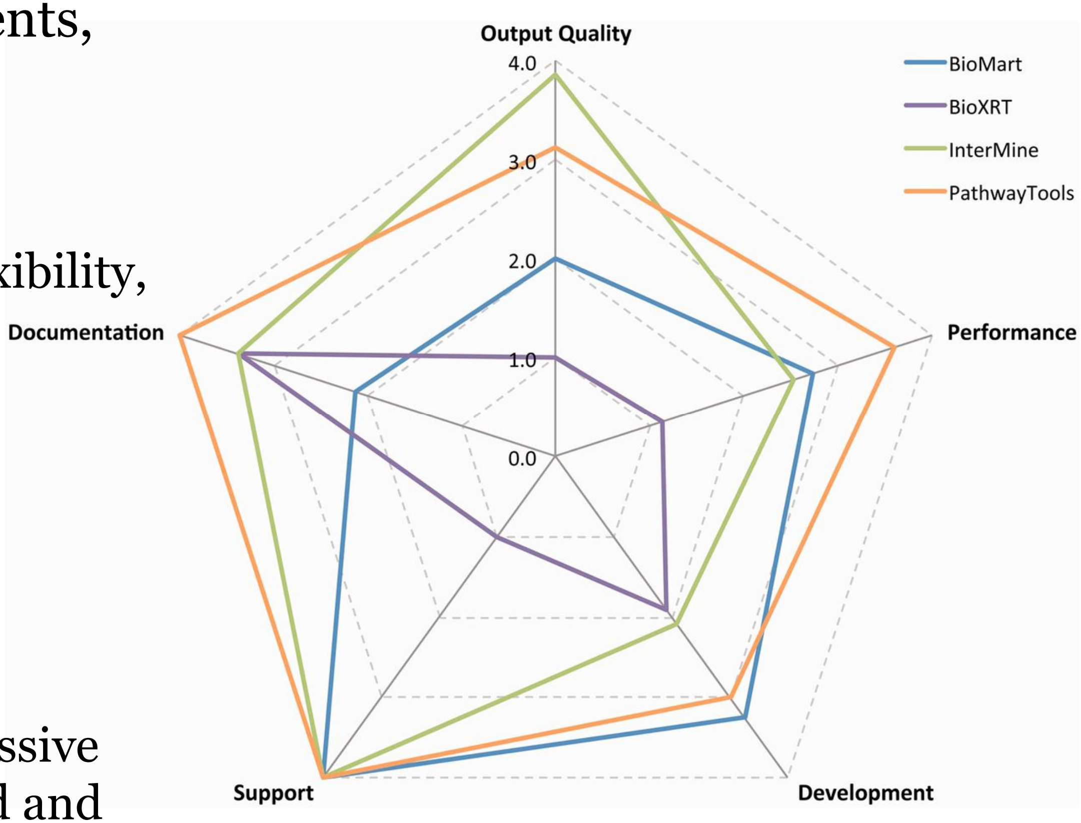

Phenotypes
Expression
Genomes
Give me all genes with phenotype X
Give me all genes over expressed in brain
Give me all genes associated with coding SNPs
BioMart integration layer
Give me all genes with phenotype X, over expressed in brain and associated with coding SNPs
WWW web browser

Comparision of 4 solutions [TB13]

- 11 queries, several environments, profiles, gold standards, benchmark…
- **InterMine**
  - (+) excellent results and flexibility,
  - (-) demanding in terms of development effort

→ for labs with IT resources.

- **PathwayTools**
  - (-) little customization
  - (+) easy-to-use, accurate
- **BioMart**
  - (-) not highly generic/expressive
  - (+) tight integration, unified and customizable interface; configured with minimal efforts.
- **BioXRT**
  - (-) not supported anymore



Thomas Triplet, and Gregory Butler Brief Bioinform 2013;bib.bbt031

Briefings in Bioinformatics

… 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]
- …

Wrap-Up

- Probably >95% of integration projects use materialization
- Successful systems implemented by domain scientists, with little participation of DR
- Very little semantic integration, very little query optimization, very little data fusion, very little schema matching / schema integration
- Full provenance information can/should be recorded

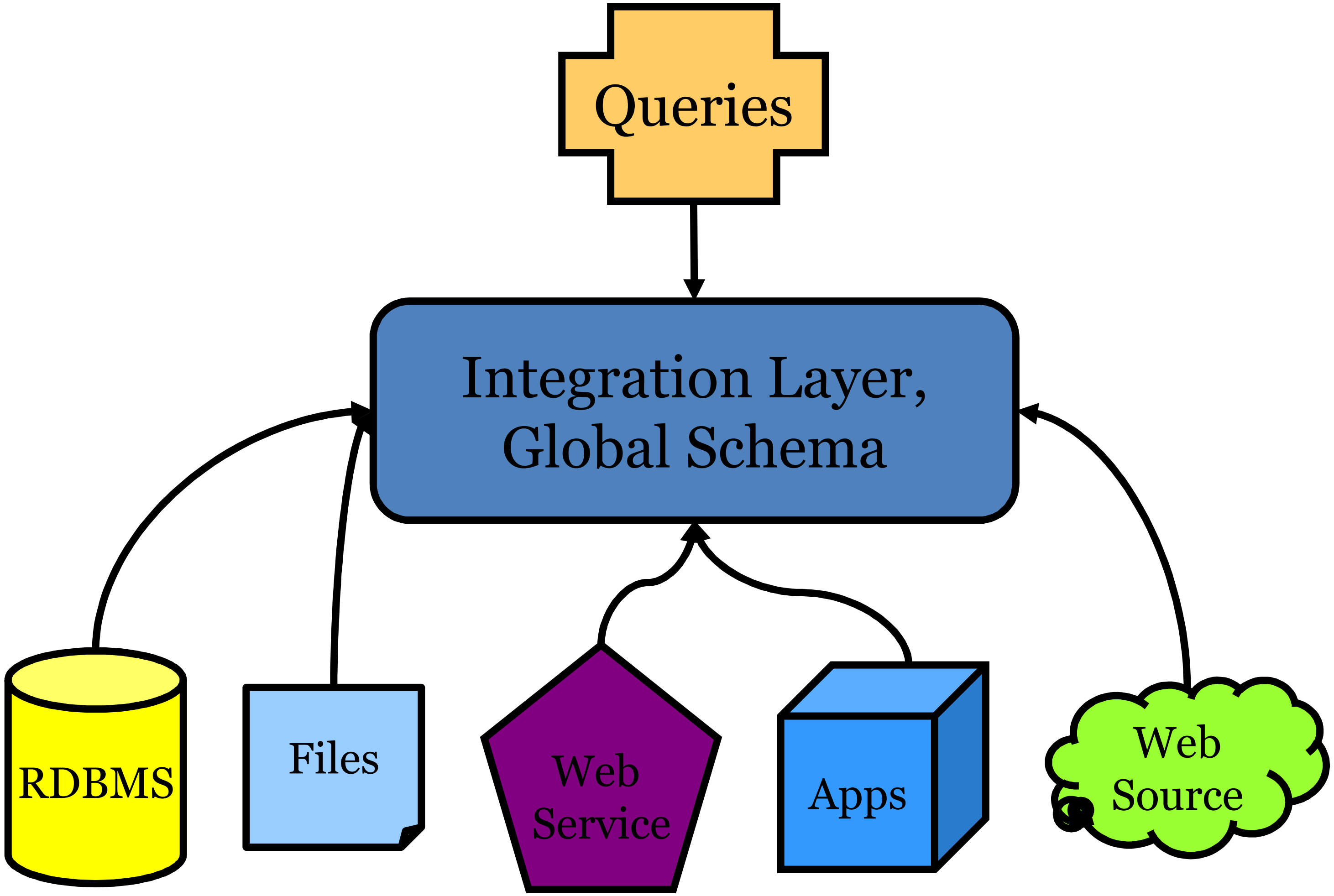
This Tutorial

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

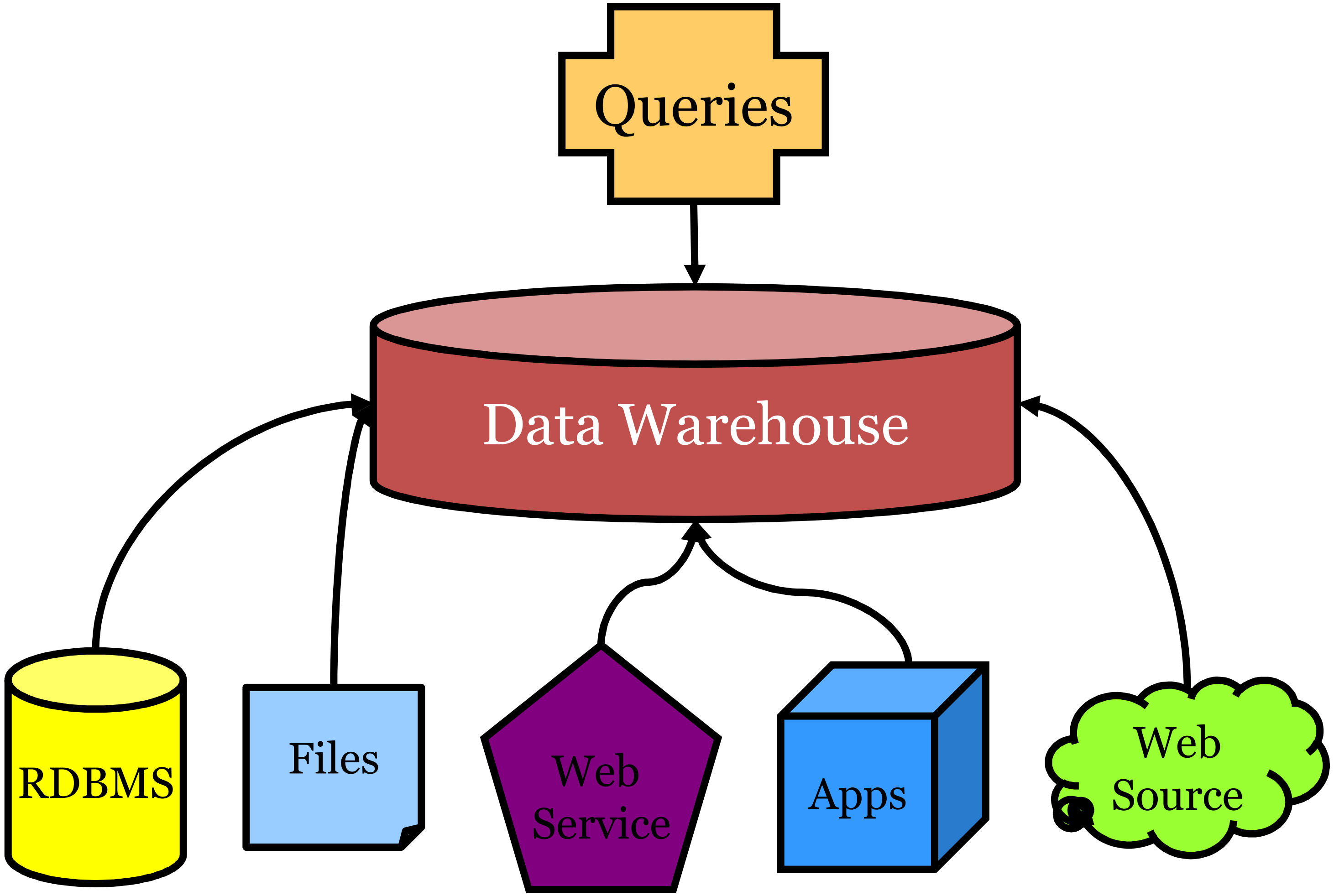
Trend 1

Analysis is integration and integration is analysis

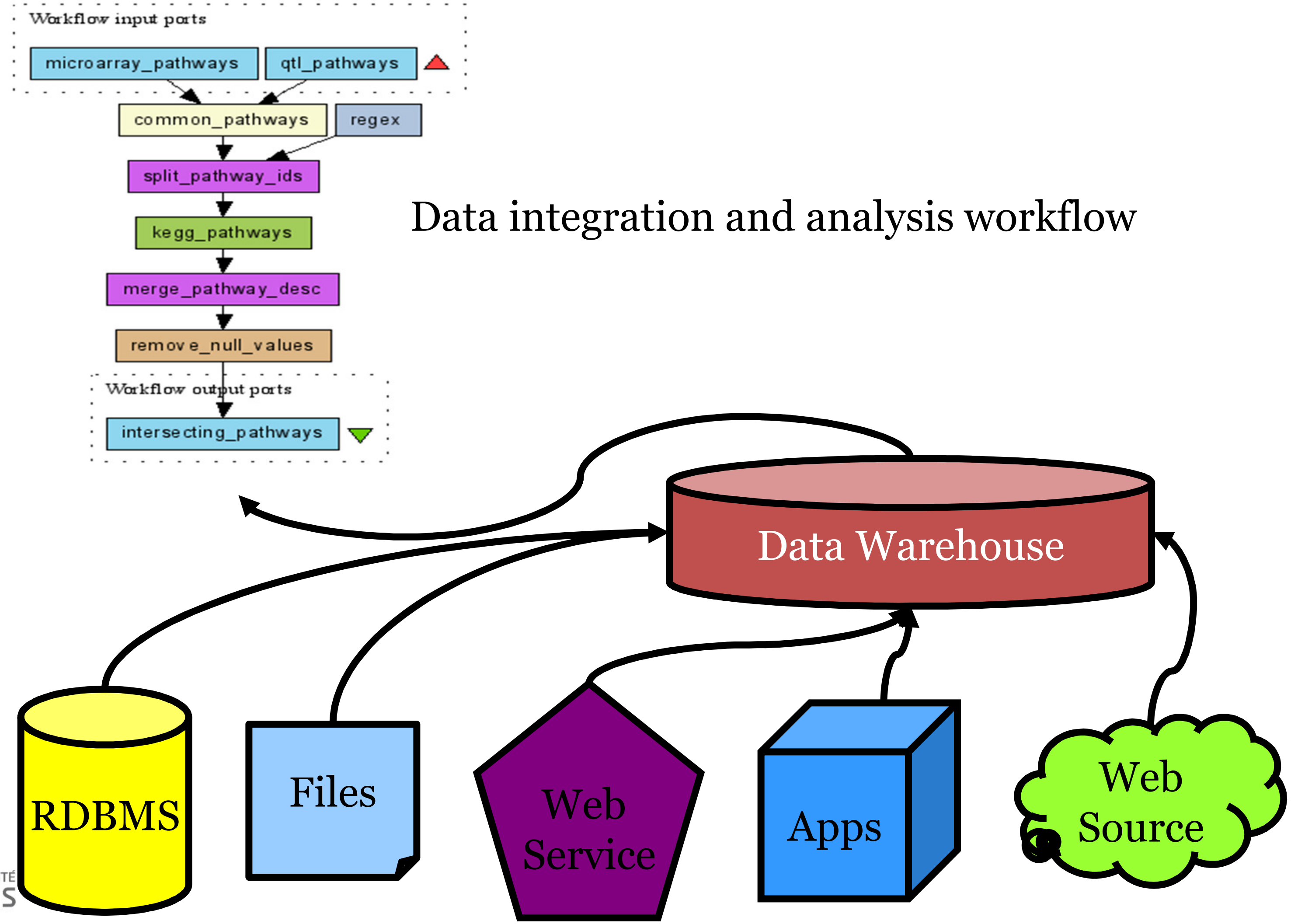
Integration Classical View (recall)



Classical View - DWH

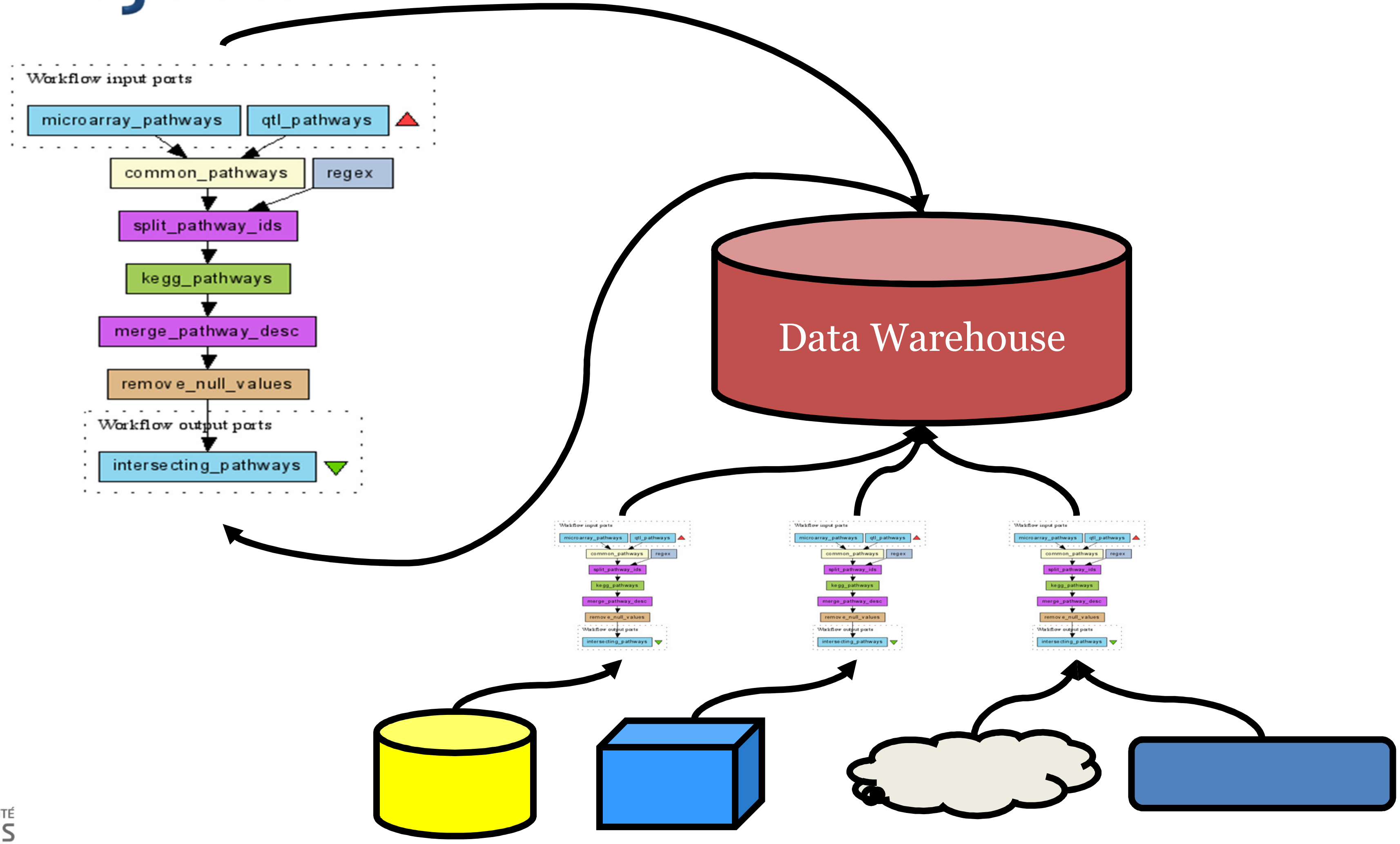


Classical View - Expanded

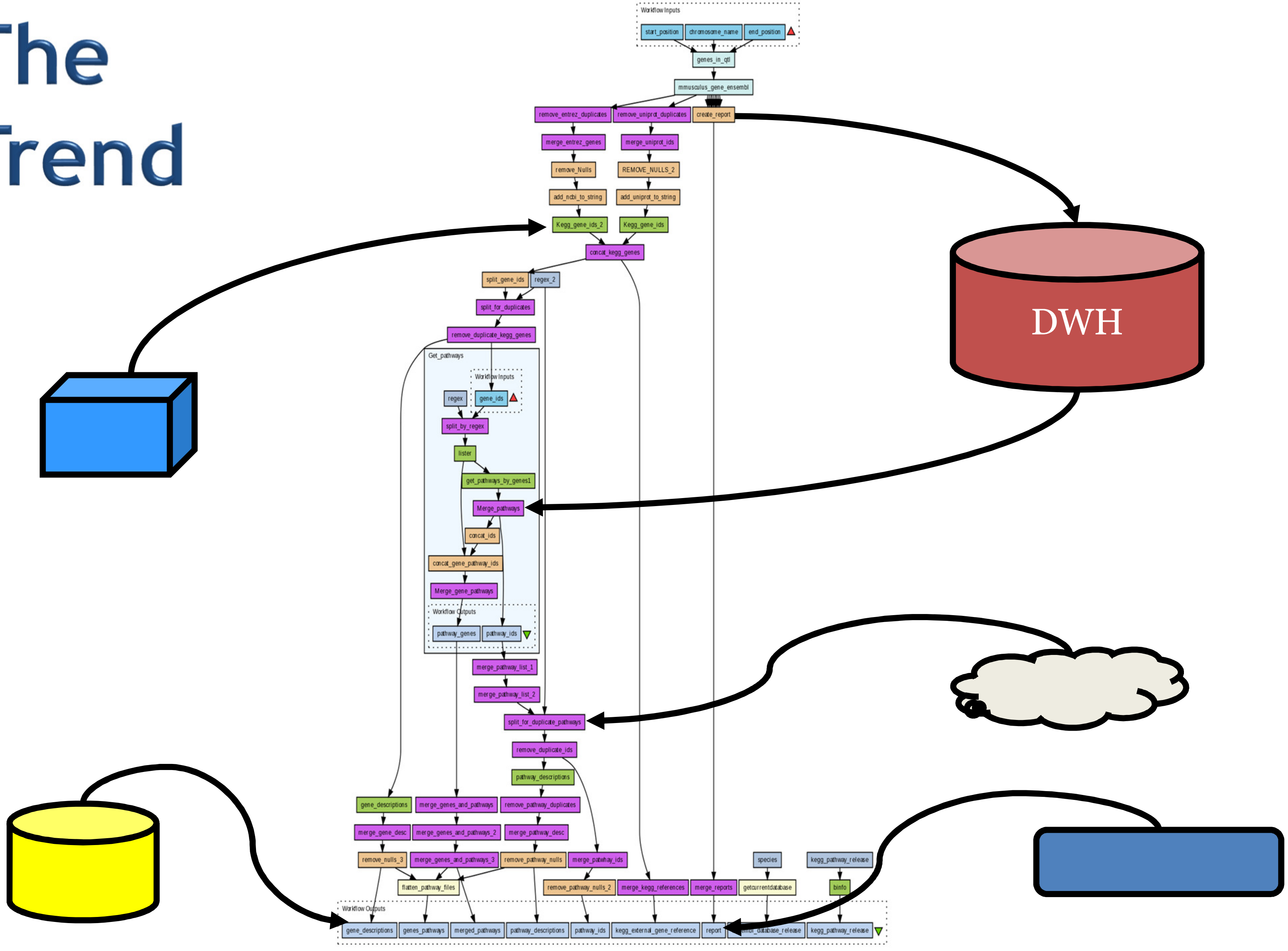


Data integration and analysis workflow

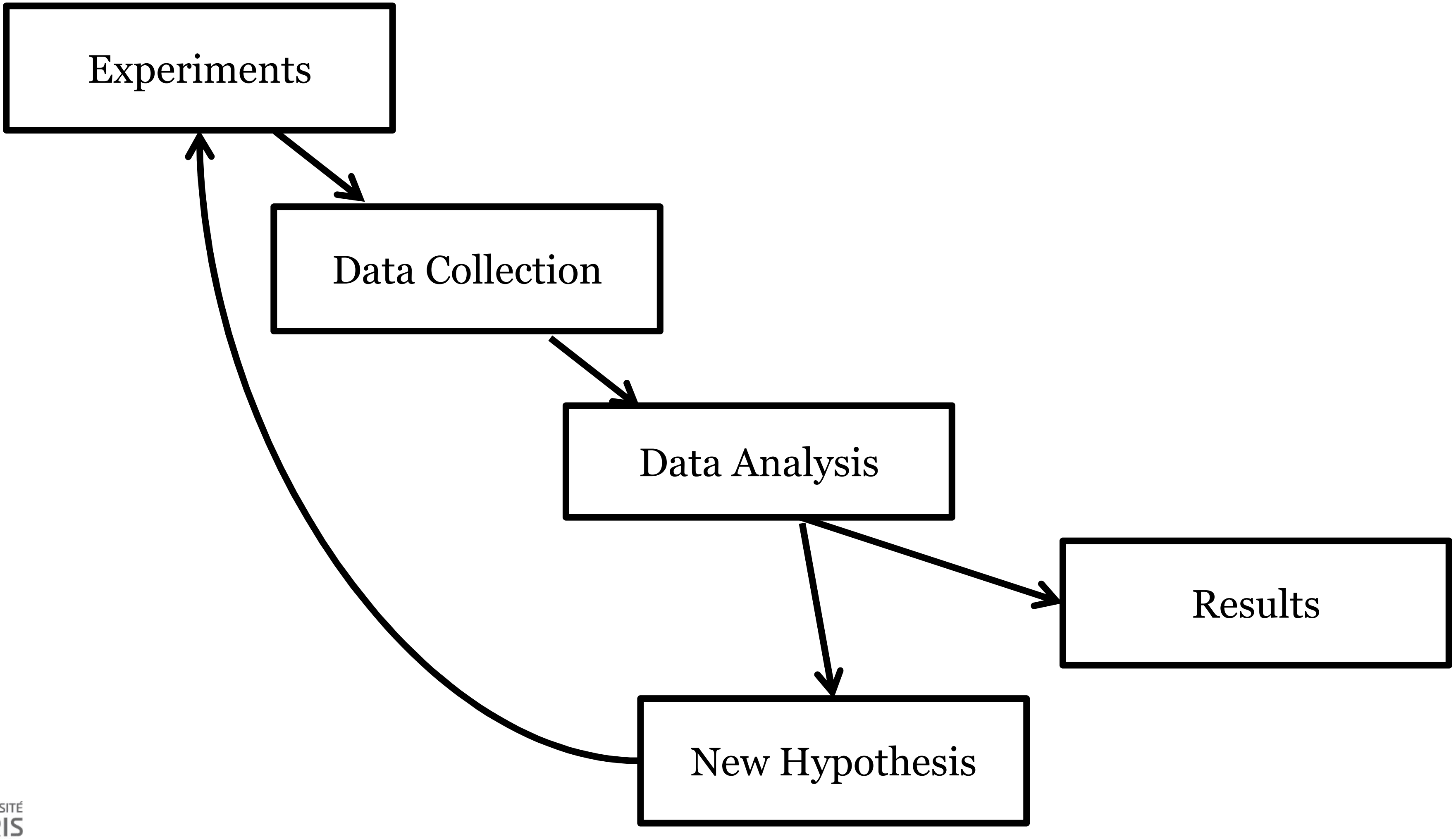
The True Architecture in Many Projects



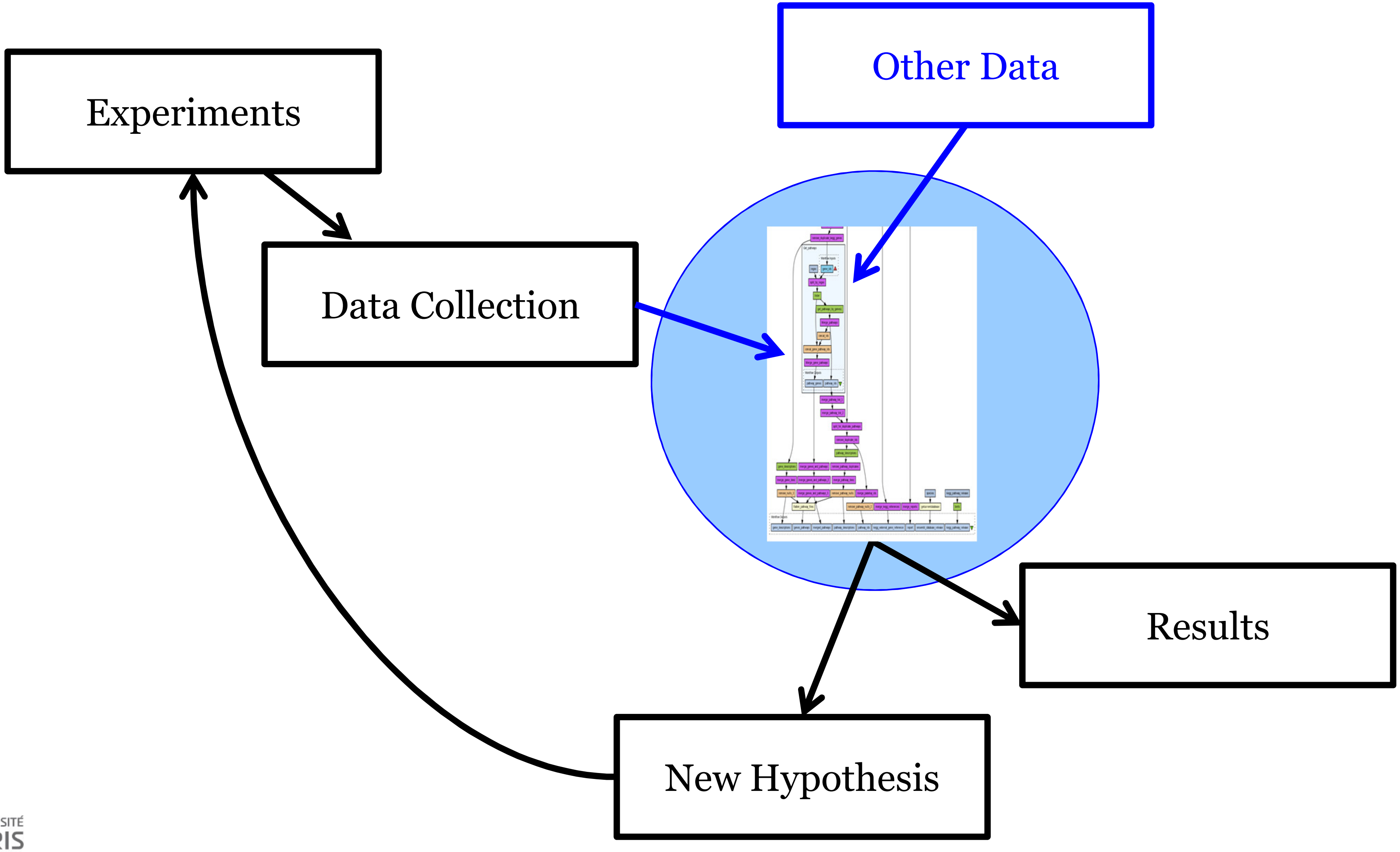
The Trend


Workflow Inputs
start_position
chromosome_name
end_position
genes_in_qtl
mmusculus_gene_ensembl
remove_entrez_duplicates
remove_uniprot_duplicates
create_report
merge_entrez_genes
merge_uniprot_ids
remove_Nulls
REMOVE_NULLS_2
add_ncbi_to_string
add_uniprot_to_string
Kegg_gene_ids_2
Kegg_gene_ids
concat_kegg_genes
split_gene_ids
regex_2
split_for_duplicates
remove_duplicate_kegg_genes
Get_pathways
Workflow Inputs
regex
gene_ids
split_by_regex
lister
get_pathways_by_genes1
Merge_pathways
concat_ids
concat_gene_pathway_ids
Merge_gene_pathways
Workflow Outputs
pathway_genes
pathway_ids
merge_pathway_list_1
merge_pathway_list_2
split_for_duplicate_pathways
remove_duplicate_ids
pathway_descriptions
gene_descriptions
merge_genes_and_pathways
remove_pathway_duplicates
merge_gene_desc
merge_genes_and_pathways_2
merge_pathway_desc
remove_nulls_3
merge_genes_and_pathways_3
remove_pathway_nulls
merge_patwhay_ids
species
kegg_pathway_release
flatten_pathway_files
remove_pathway_nulls_2
merge_kegg_references
merge_reports
getcurrentdatabase
binfo
Workflow Outputs
gene_descriptions
genes_pathways
merged_pathways
pathway_descriptions
pathway_ids
kegg_external_gene_reference
report
embl_database_release
kegg_pathway_release
DWH

Life Science Research Food Chain

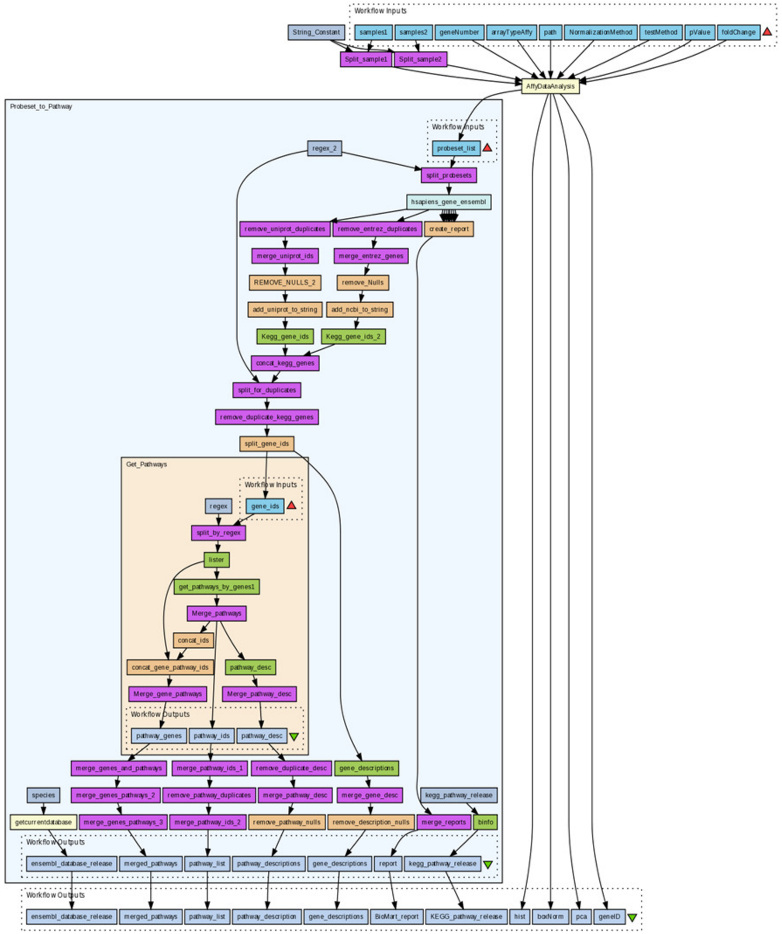


With DI Workflows



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
- SWFS manages scheduling, process control, logging, recovery, reproducibility, …
- Equipped with graphical workflow designer
- Several systems available (Galaxy, SnakeMake, Kepler, …)



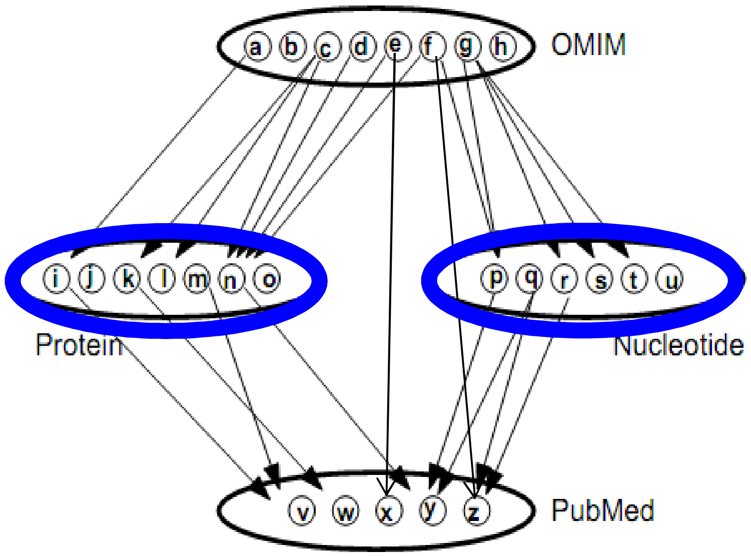
Trend 2

Data quality depends on provenance

Criteria for Relevance

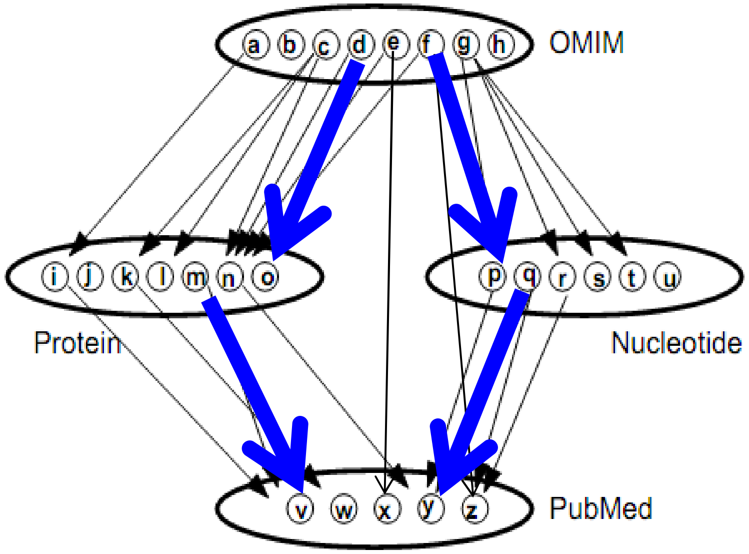
User provided	• Assessment of quality of used data sources • Assessment of quality of links • Currentness, completeness, trust, …
Query dependent	• Number of paths allowing to obtain a data item • Length of paths
Domain specific	• Similarity of linked sequences • Quality of matching leading to a link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

Example



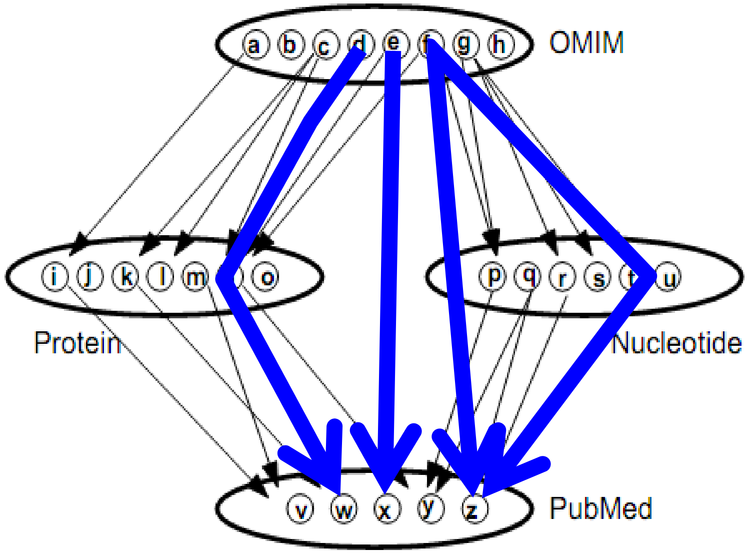
Which source is better?

Example



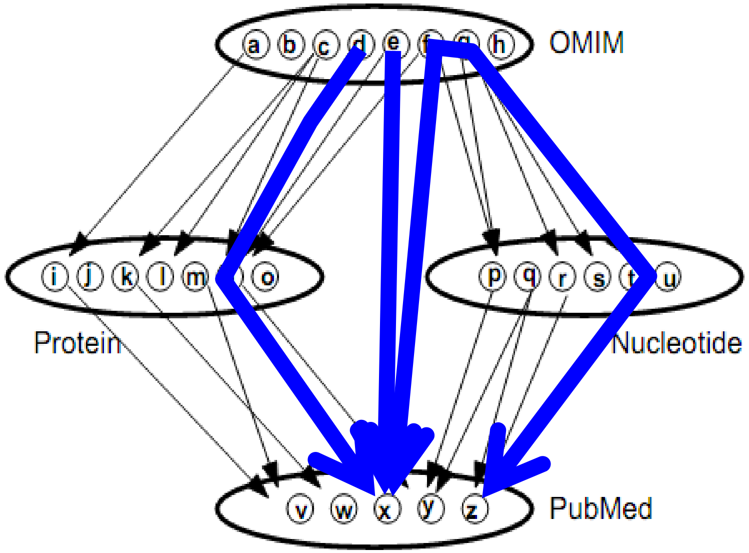
Which link is better?

Example



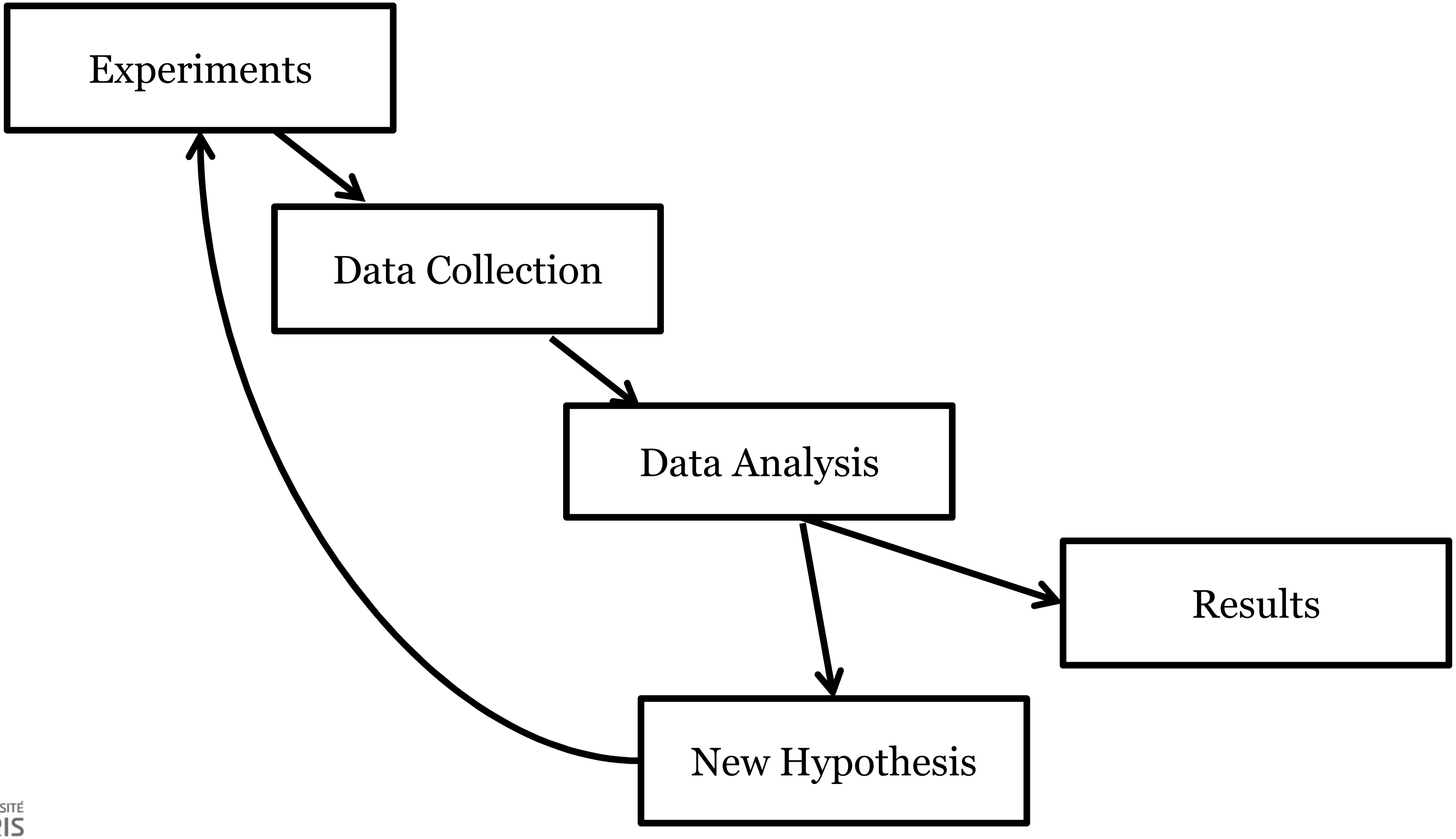
Which path is better?

Example

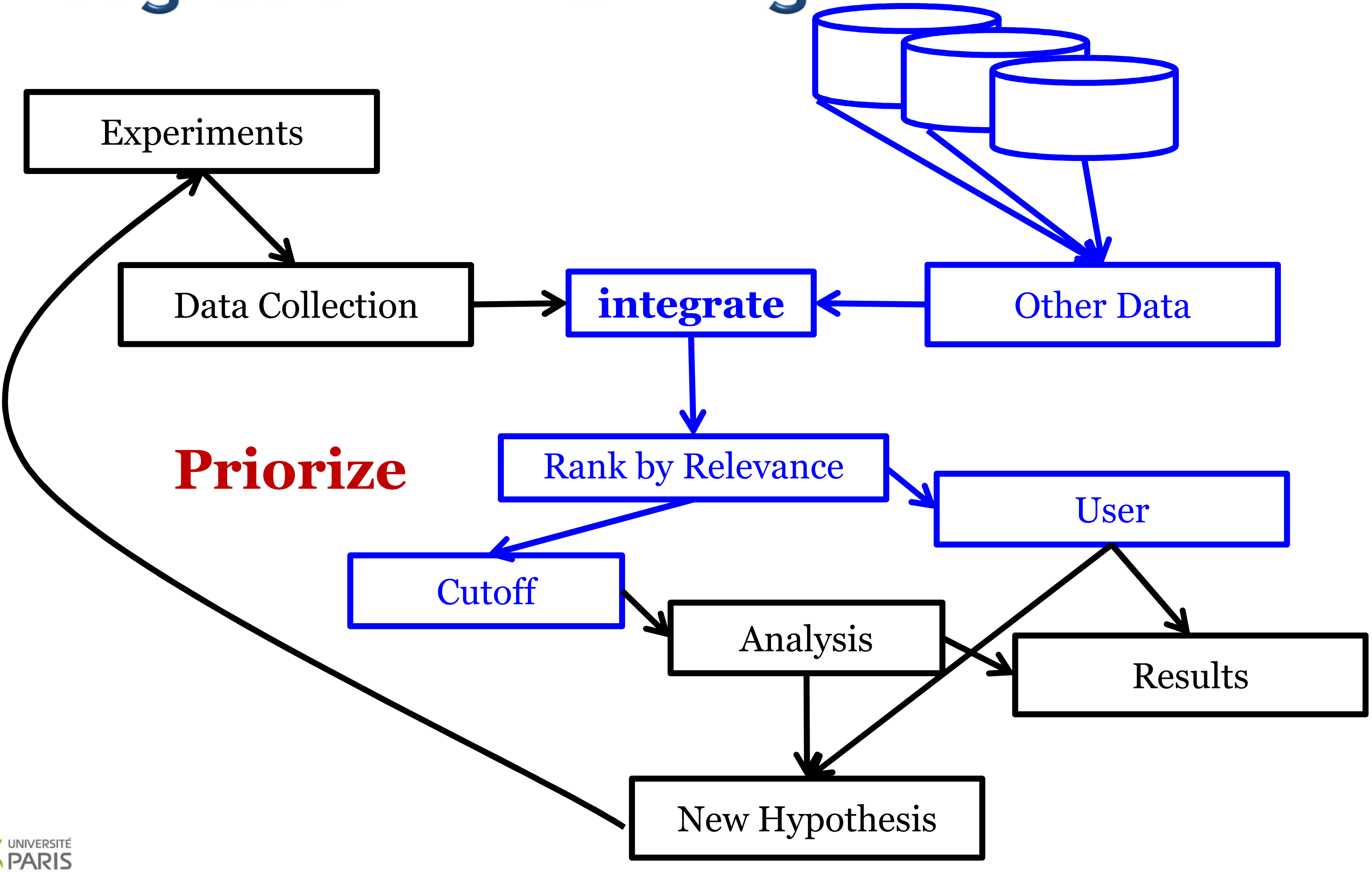


Which objects are reached by more paths?

Life Science Research Food Chain



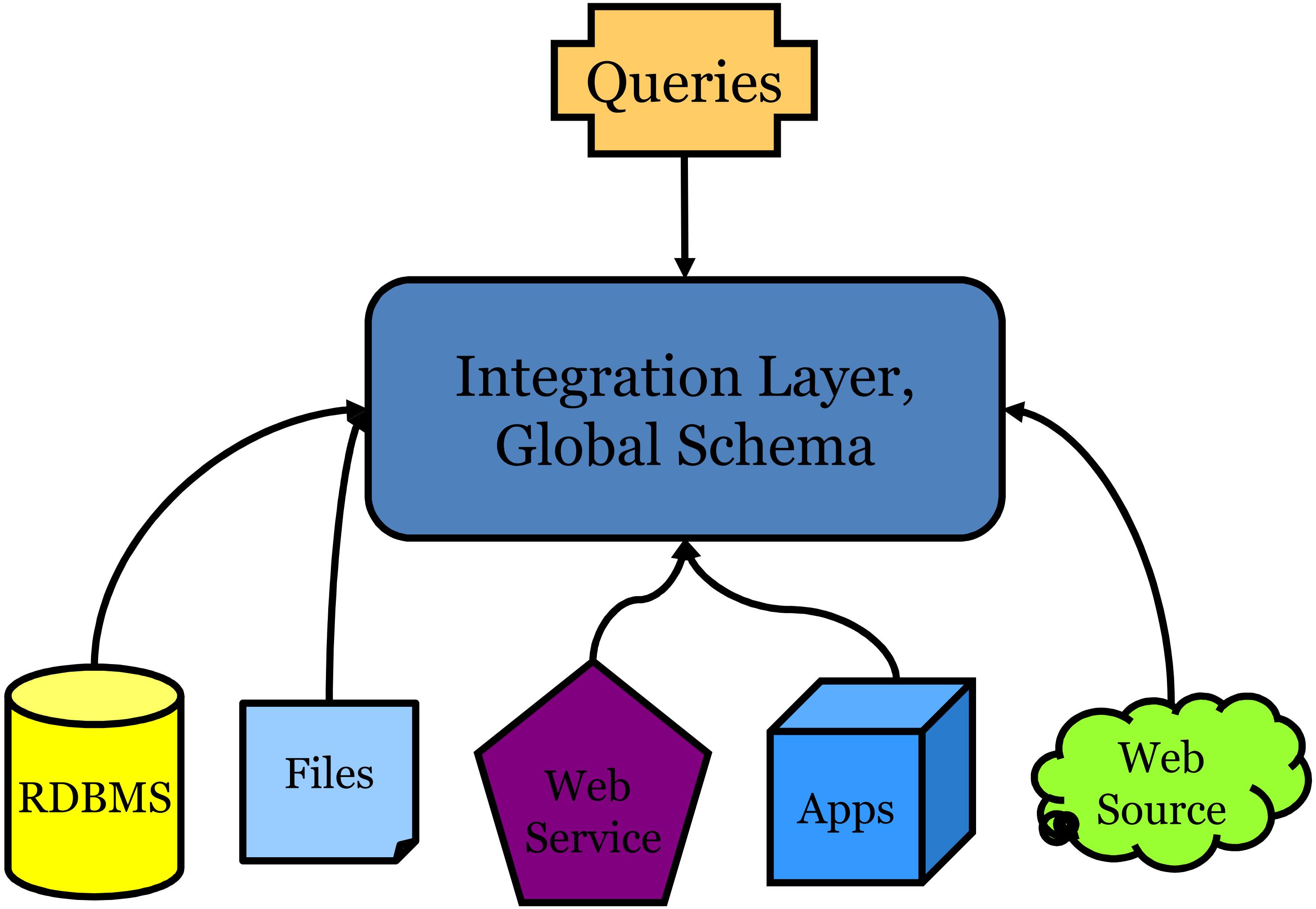
Integration + Ranking



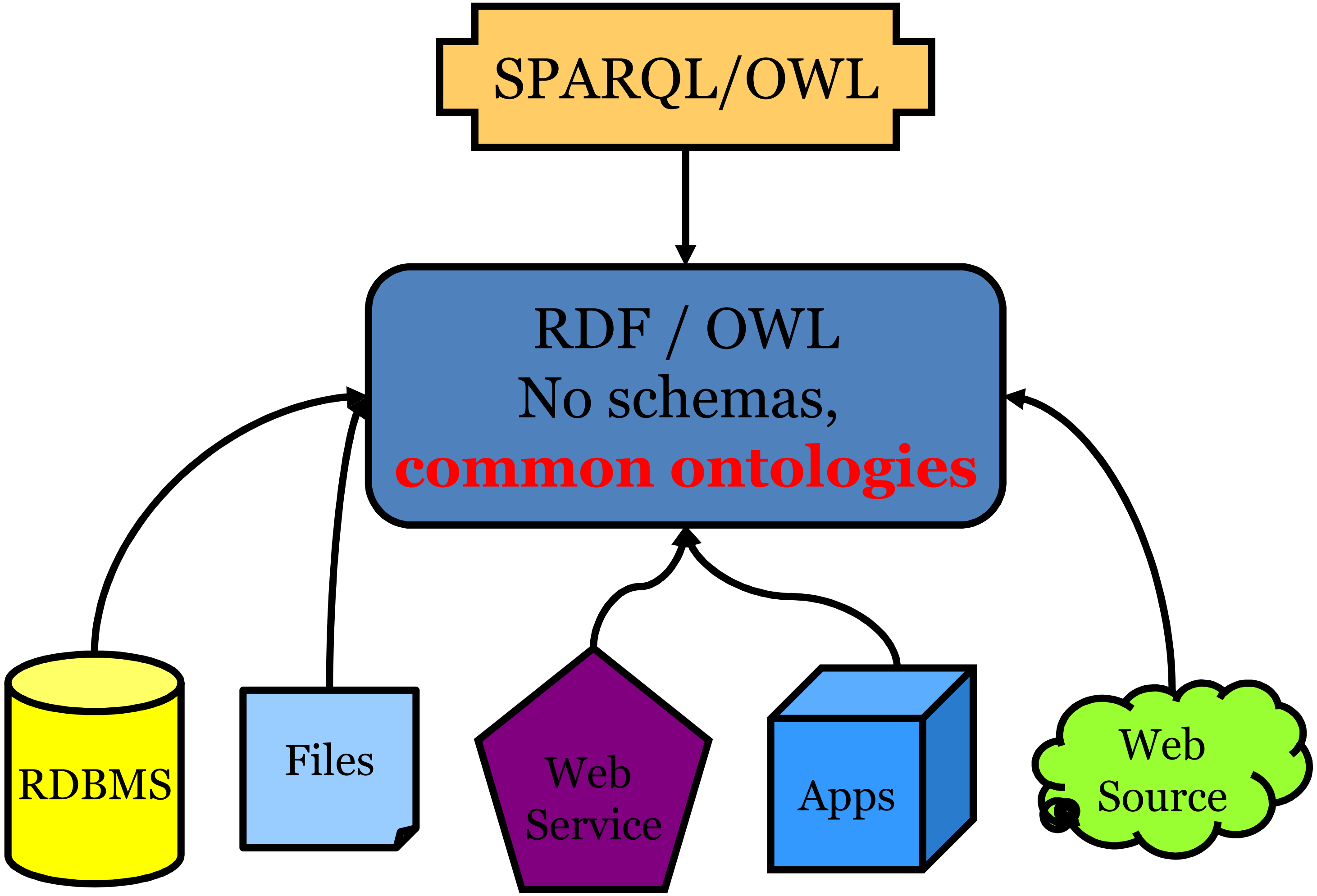
Trend 3

Semantic integration can be performed using ontologies (and Web semantics approaches)

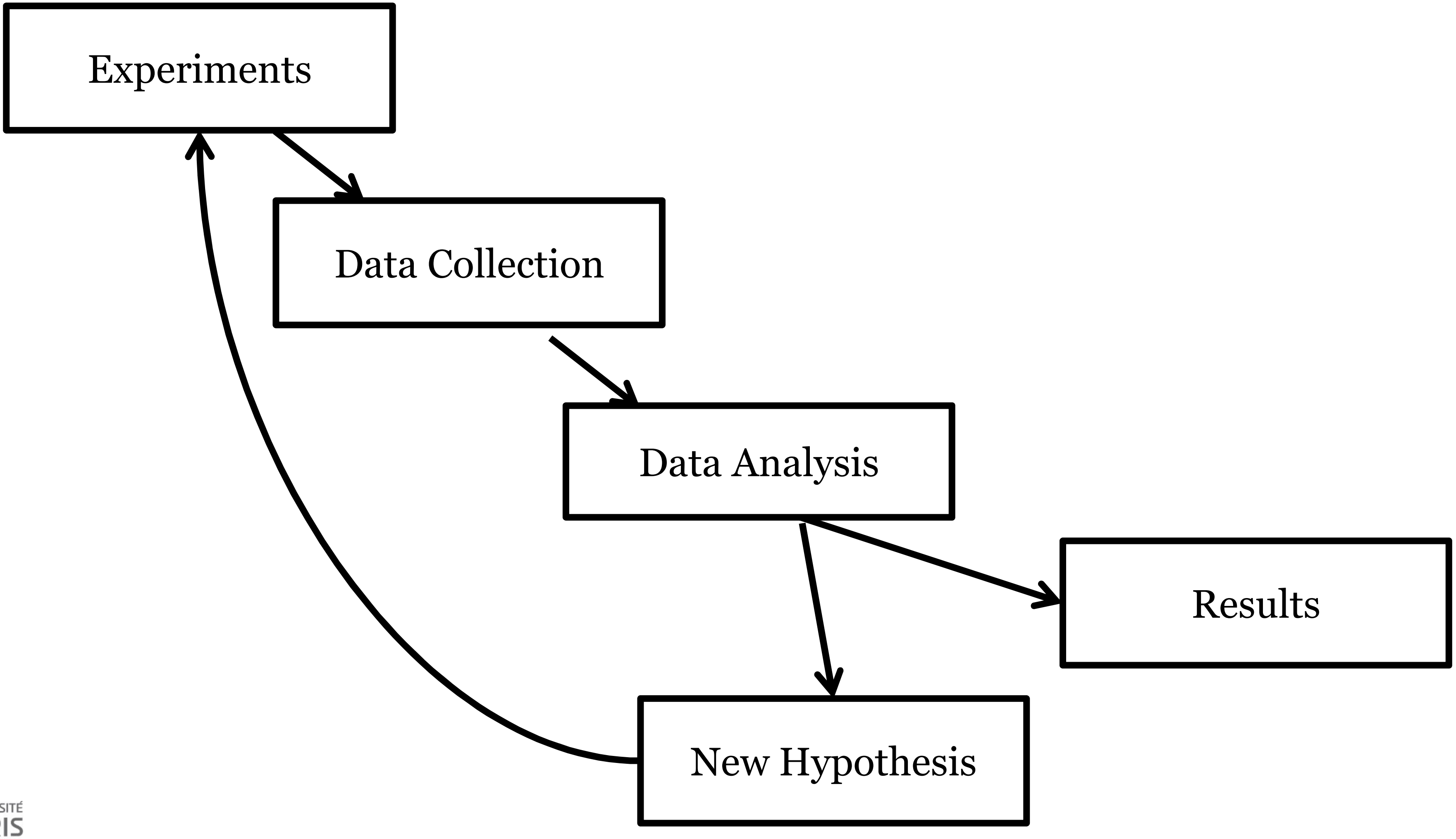
Classical View



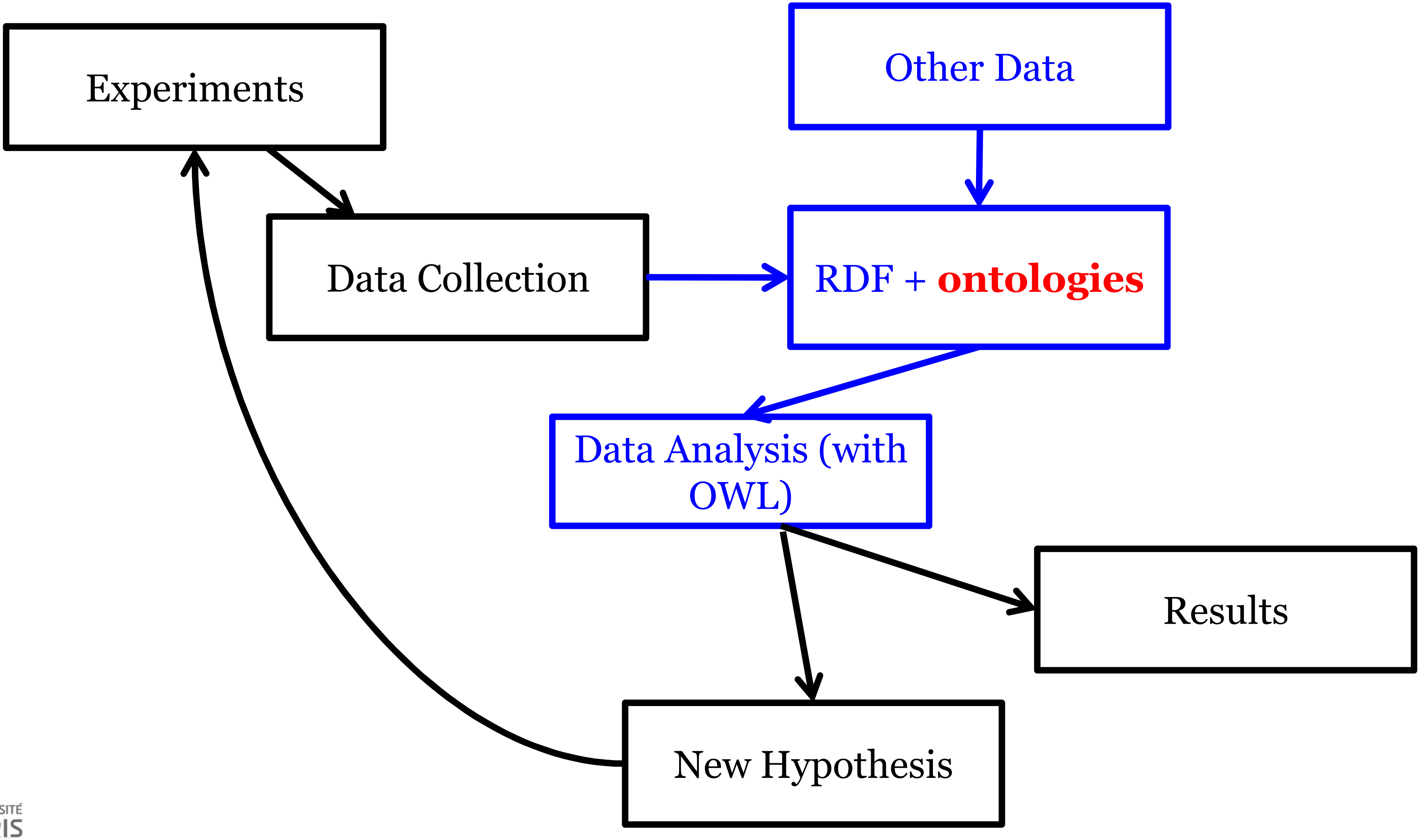
Semantic Web Approach



Life Science Research Food Chain



… using Semantic Web Techniques



Conclusions

- Data Integration in the Life Science (DILS) is more important than ever
- Portals perform syntactic integration and are frequently used
- Data warehouses are designed in several places. It remains the most frequently used in the Life Science community
- Faced with the increasing number of
  - data,
  - sources,
  - analytic tools,
  - and the increasing complexity of analysis pipelines…

… challenges are numerous…

Conclusions (cont.)

- The complexity of the questions to be answered has increased a lot
  - Integration requires analysis and analysis requires integration
  - **Scientific workflows**
- The diversity of the sources has increased a lot
  - Inclusion of quality as a first-class citizen
  - **Ranking** of integrated search results

- 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, development of **ontologies**

Data and Software Carpentry

- Initiatives worth looking at

DATA CARPENTRY

MAKING DATA SCIENCE MORE EFFICIENT

software carpentry

- ELIXIR European project (Infrastructure for bioinformatics)
  - Software and data carpentry (coordinator for the French Node)
  - Contact-me ☺ : cohen@lri.fr

elixir