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/teaching.html

Very large Data and Knowledge in Bioinformatics

UNIVERSITÉ PARIS SUD

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)

NCBI Resources How To Sign in to NCBI

Search NCBI databases Help

Long QT syndrome Search

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 ir
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

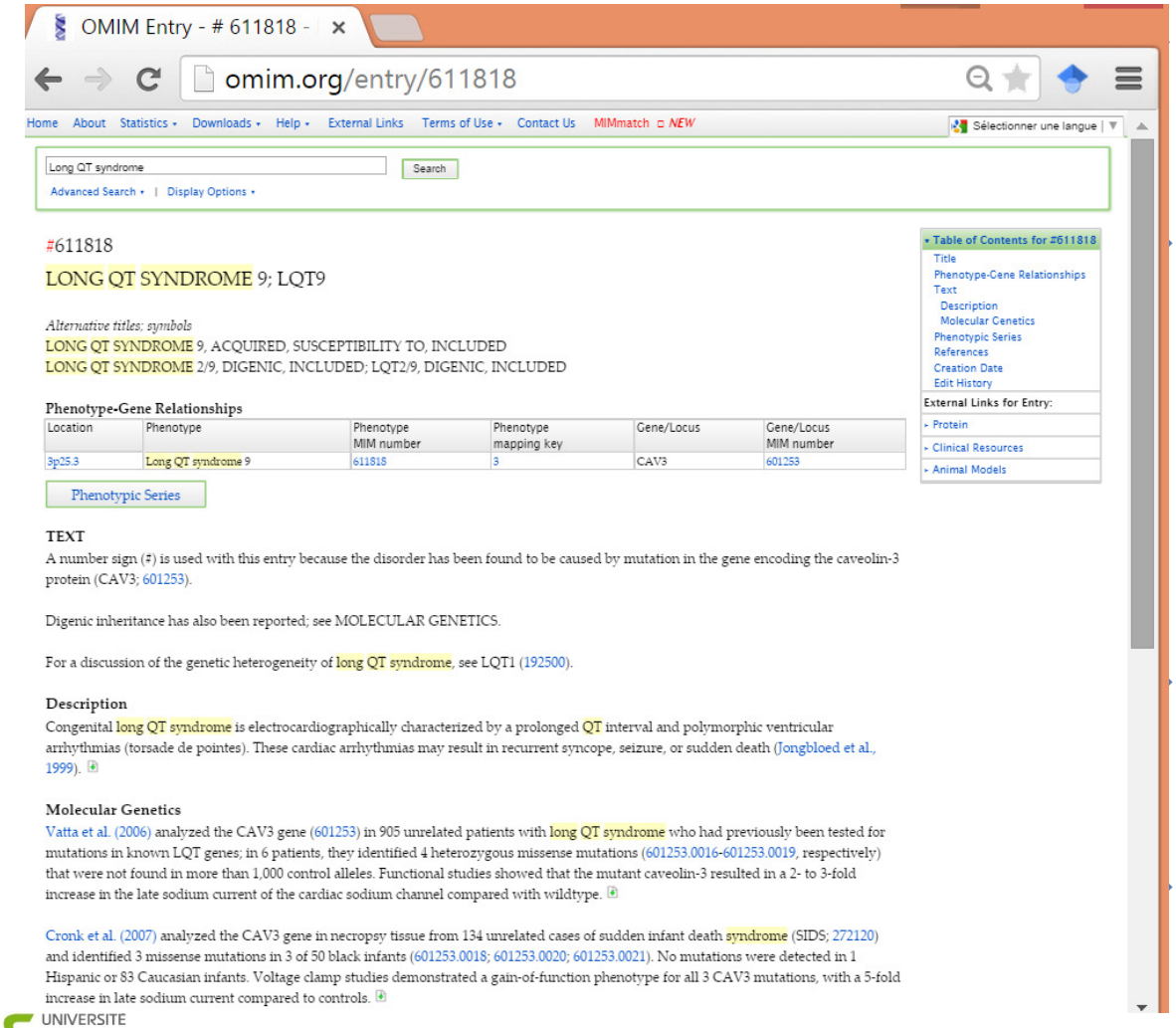
Proteins

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)



OMIM Entry - # 611818 -

omim.org/entry/611818

Home About Statistics Downloads Help External Links Terms of Use Contact Us MIMmatch NEW Sélectionner une langue

Long QT syndrome Search

Advanced Search | Display Options

#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

Phenotypic Series

Table of Contents for #611818

Title
Phenotype-Gene Relationships
Text
Description
Molecular Genetics
Phenotypic Series
References
Creation Date
Edit History

External Links for Entry:

Protein
Clinical Resources
Animal Models

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)

NCBI Resources How To Sign in to NCBI

Search NCBI databases Help

Long QT syndrome Search

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

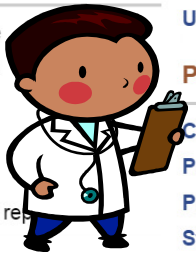
Genes		
EST	2	expressed sequence tag sequences
Gene	33	collected information about gene loci
GEO DataSets	1	functional genomics studies
GEO Profiles	0	gene expression and molecular abundance profiles
HomoloGene	11	homologous gene sets for selected organisms
PopSet	0	sequence sets from phylogenetic and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

Entrez Gene

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

Chemicals

proteins and



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

UNIVERSITÉ PARIS SUD
Comprendre le monde, construire l'avenir
université PARIS-SACLAY

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

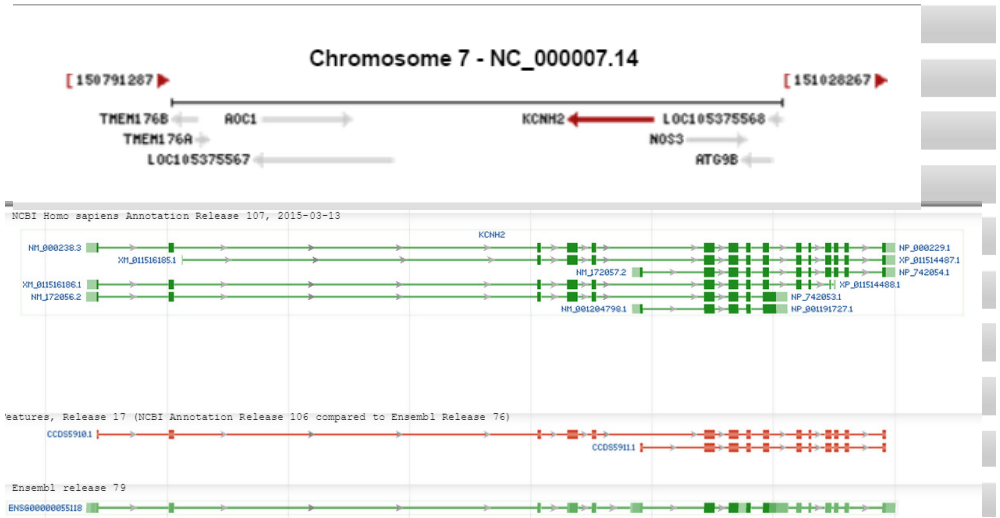
Summary

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

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

Genomic context

Genomic regions, transcripts, and products



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

Querying (NCBI Portal)

NCBI Resources How To Sign in to NCBI

Search NCBI databases Help

Long QT syndrome Search

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

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

Genomes

Assembly	0	genome assembly i
BioProject	7	biological projects p

roteins and

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

Nucleotides

One GenBank entry (Long QT)

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

GenBank: DI042621.1

FASTA Graphics

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

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 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

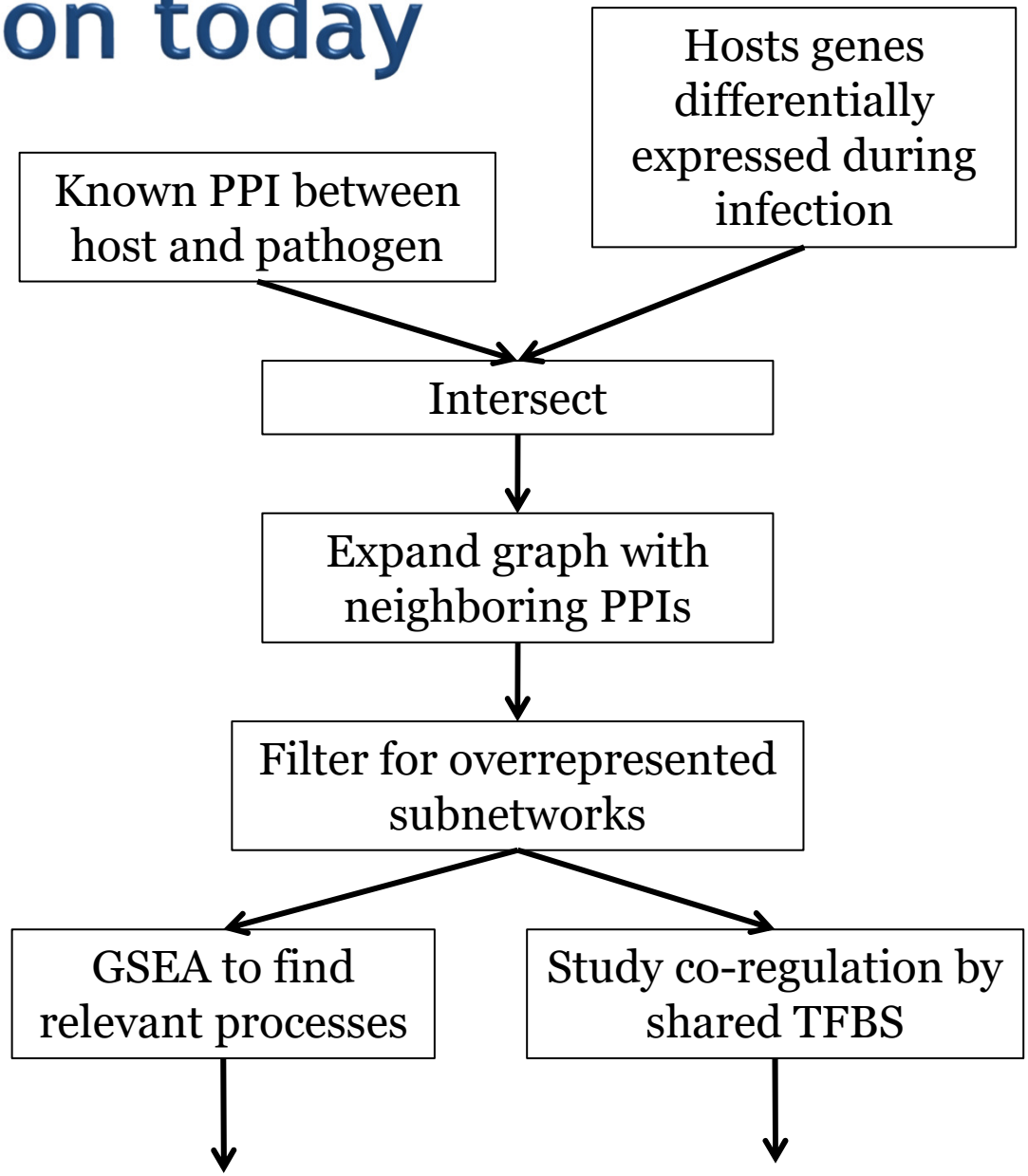
Take Home Message

- The classical problems are all there already

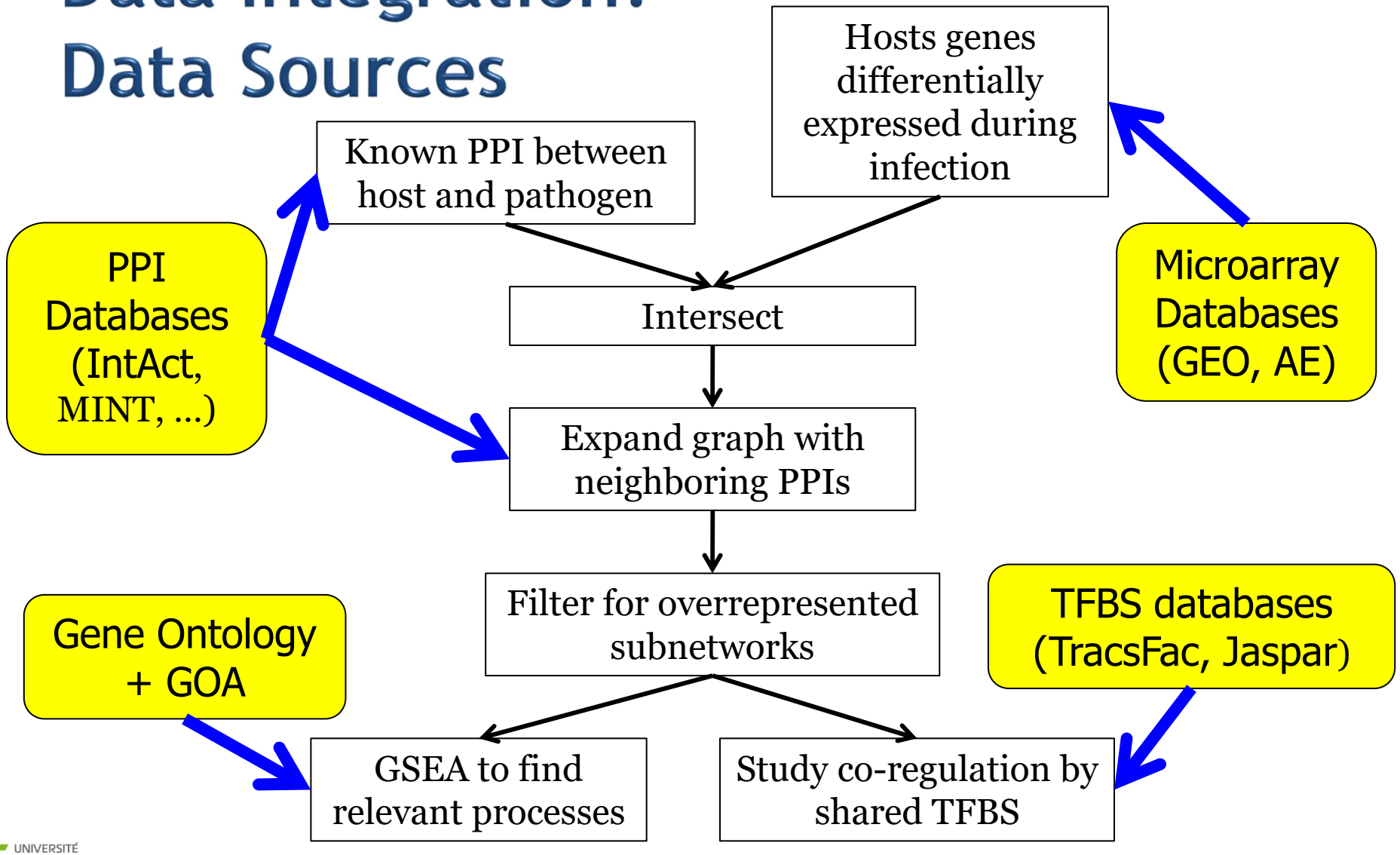
- Distributed information
- Semantic heterogeneity
  - Scientific uncertainty and evolving concepts
  - Naming conventions on the object level
  - Naming conventions on the concept level
  - Inclusion of non-standard processing

Data Integration today

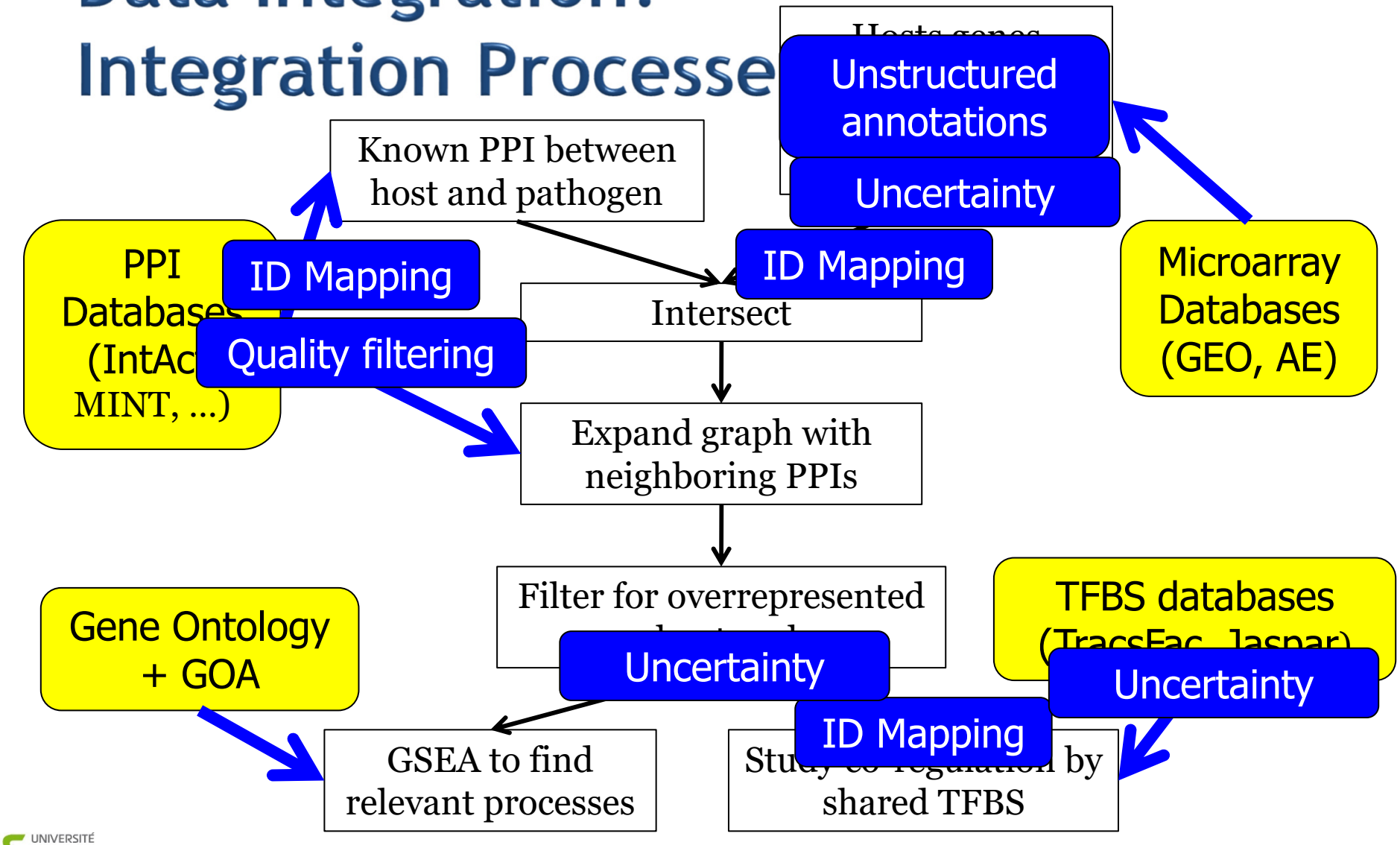
- Task: Find genes that play a central role in the response of a host to a pathogen
  - Bacteria / viruses must attach to cells to have an influence
  - Attachment is a physical binding of proteins
  - This binding provokes a reaction in the cell, transmitted by more PPI (e.g. transient signaling)



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, 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

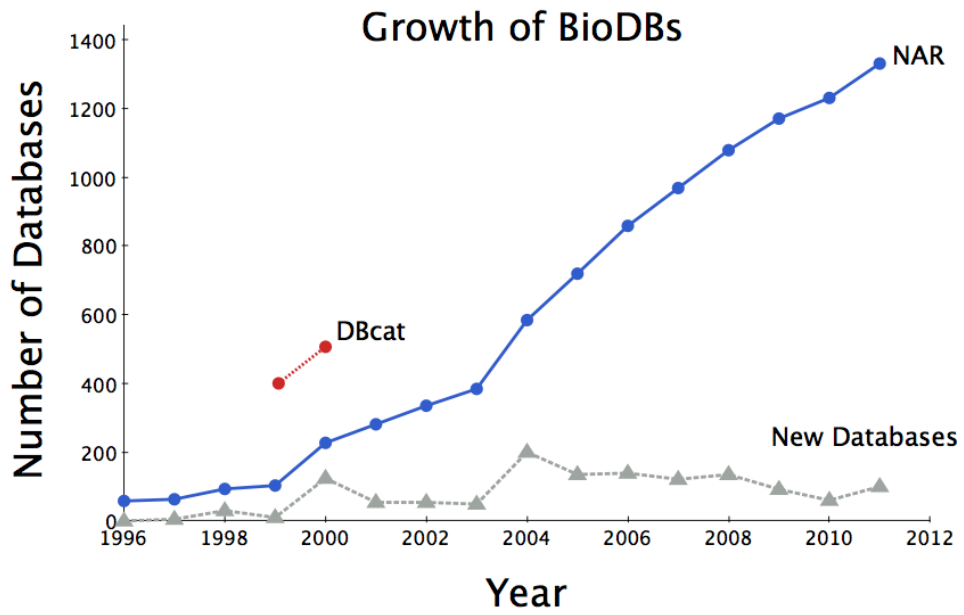
This Tutorial

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

- Part II – Scientific Workflows

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



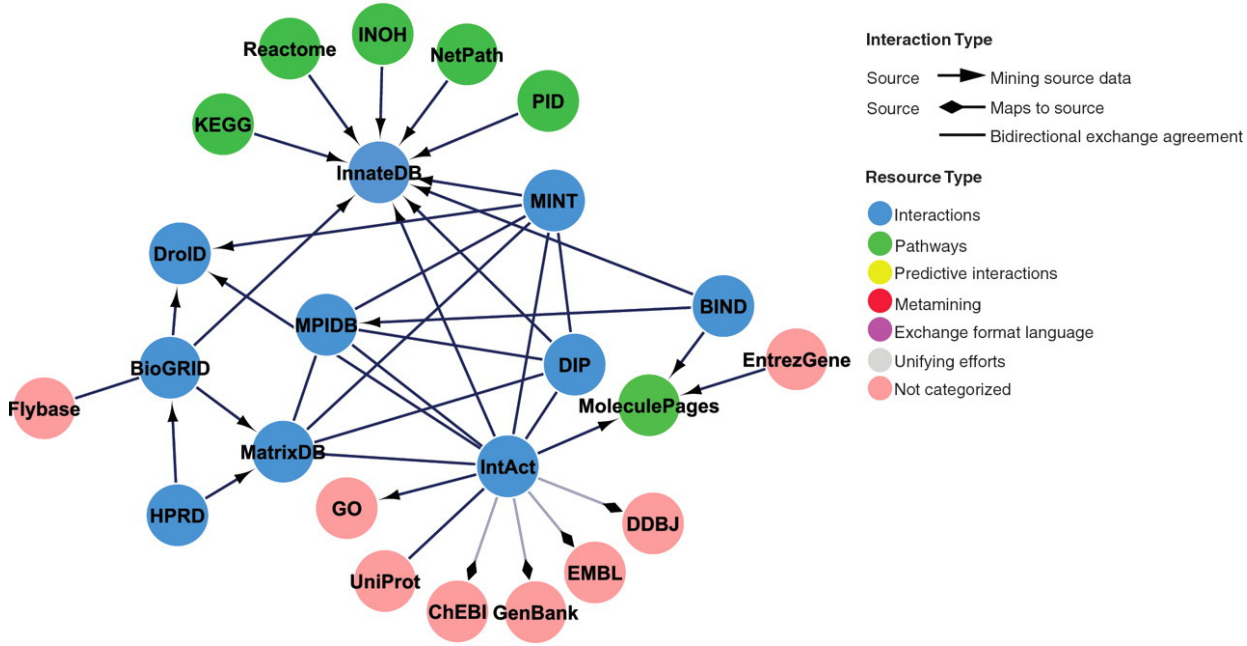
Number of existing (circles) and new databases (triangles) are plotted from 1996 to 2011. New databases are difference between the number of existing databases for each year. DBcat (red) is shown with NAR (blue) counts.

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

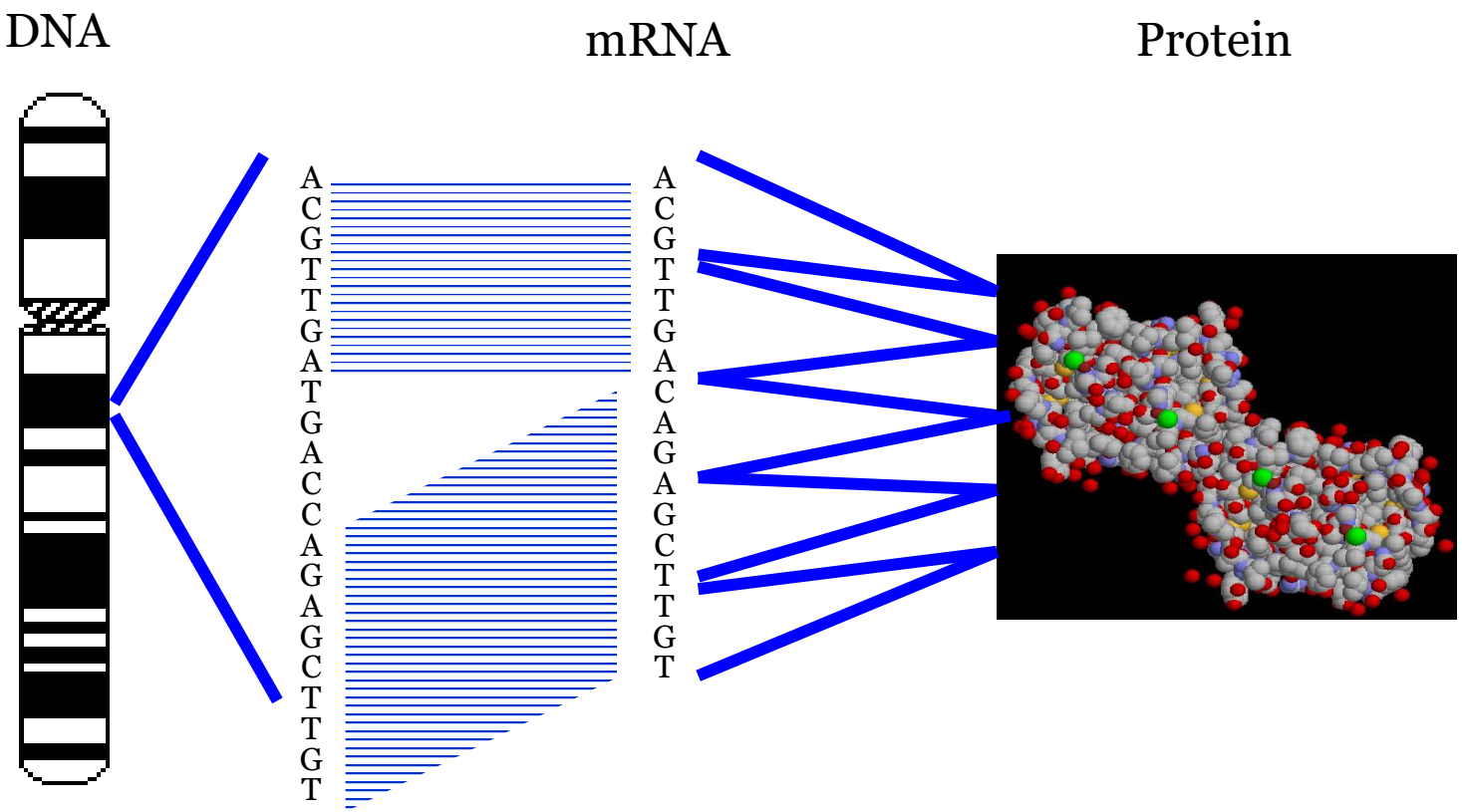
- Manually created "source" DBs



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)?



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

What is a Gene (2)?

(A) EUCARYOTES

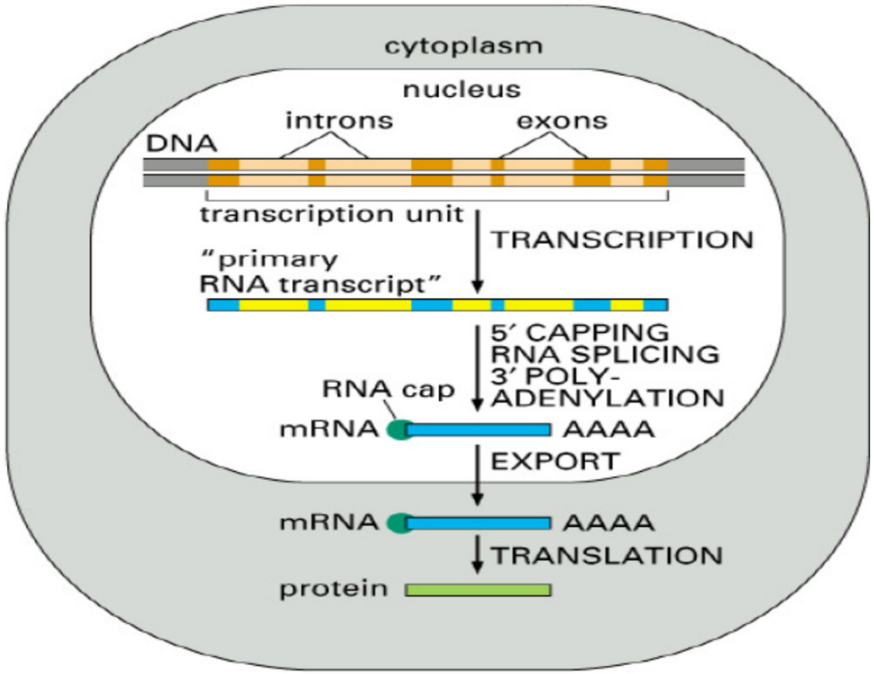
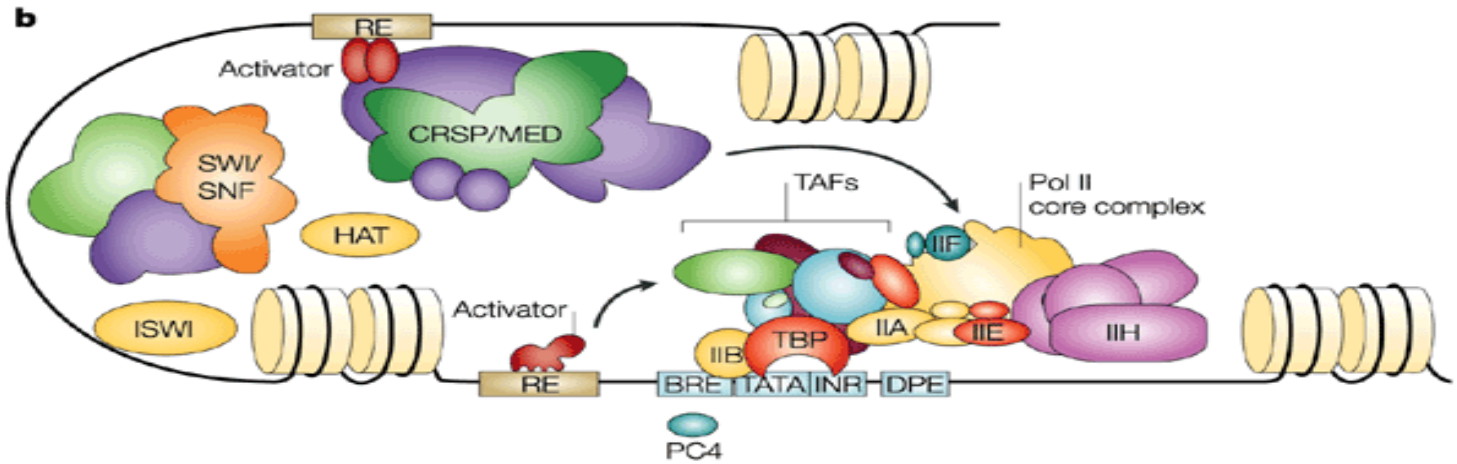


Figure 6–21 part 1 of 2. Molecular Biology of the Cell, 4th Edition.

- 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]

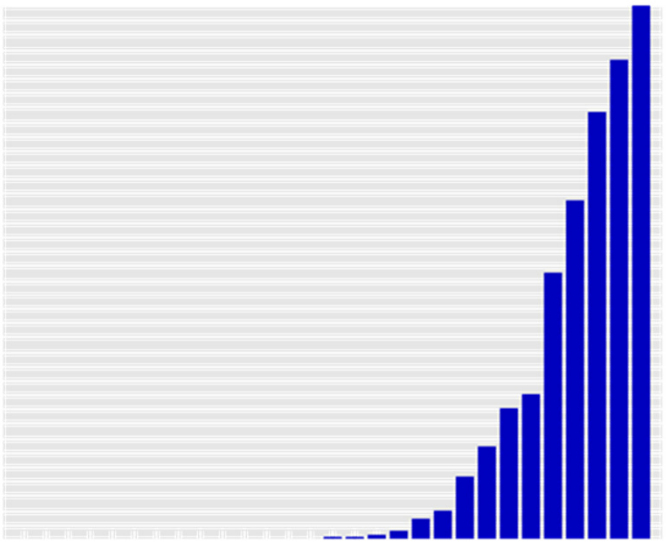
- The same gene?
  - Genes may generate different assemblies (differential splicing)
  - Gene duplications in a genome
  - The „same“ gene in another organism
  - Mutation of a gene
  - Genes with a different start site
- A gene?
  - Pseudo genes (never transcribed, yet highly similar)
  - Non-coding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

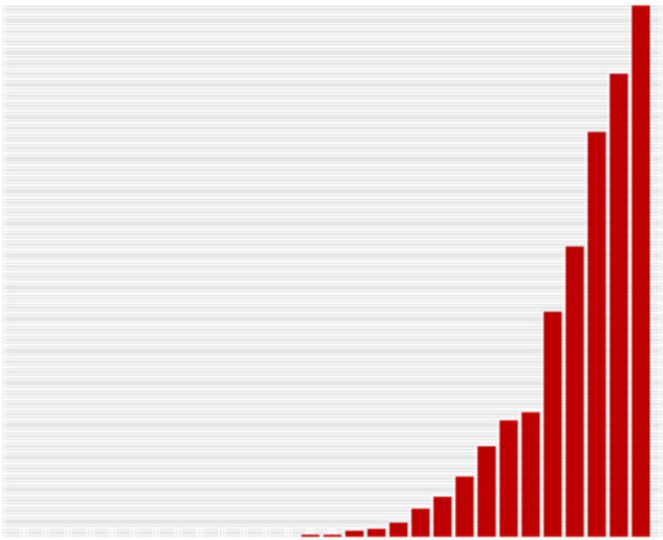
Is Data Quality an Issue in BDB?

- 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?

300G bases

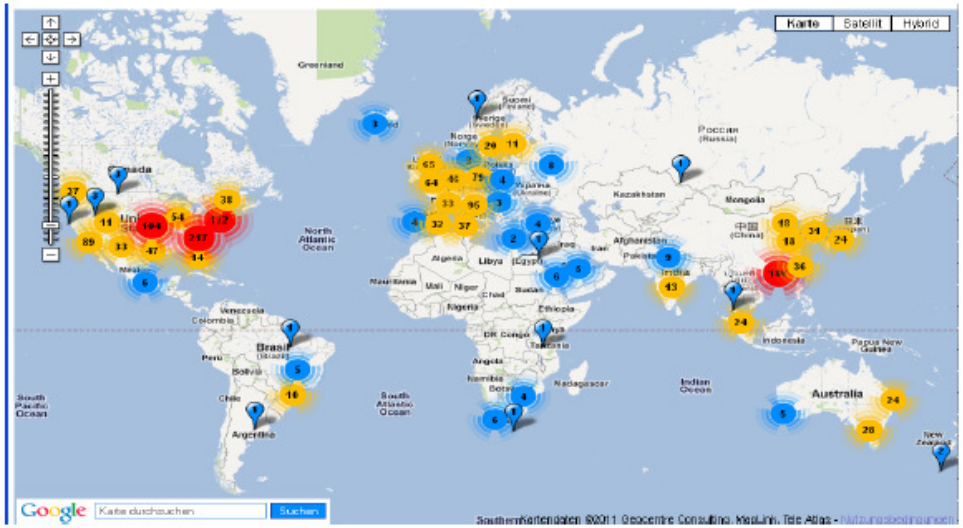




200M entries

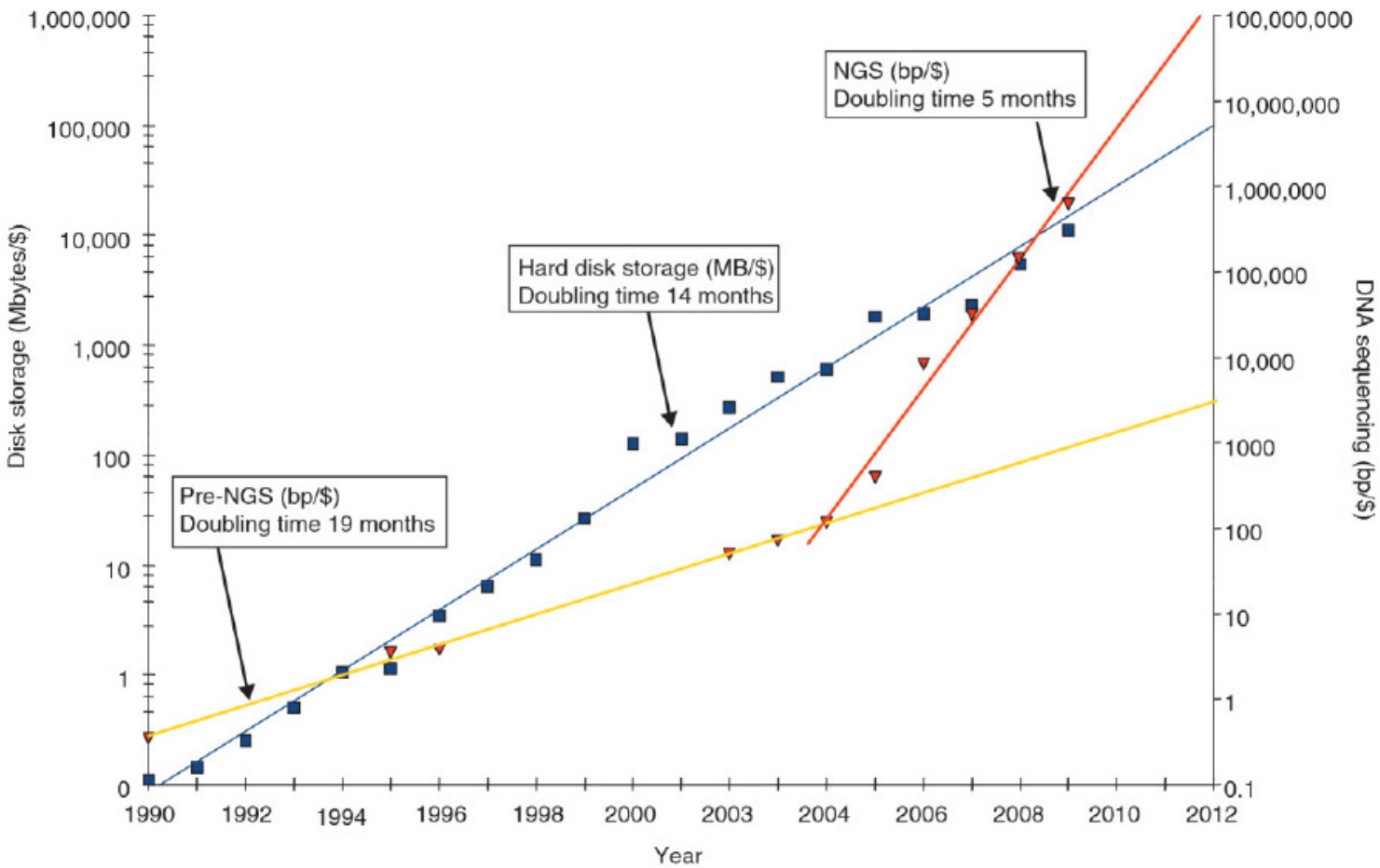
- All of EMBL now has ~150 TB (zipped), ENSEMBL has ~1TB (MySQL dump), UniProt has ~5GB (zipped)
- Probably 90% of the 1300 DB's in NAR have <1GB
- All secondary databases have "little" data
- Primary data explodes due to Next Generation Sequencing

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?

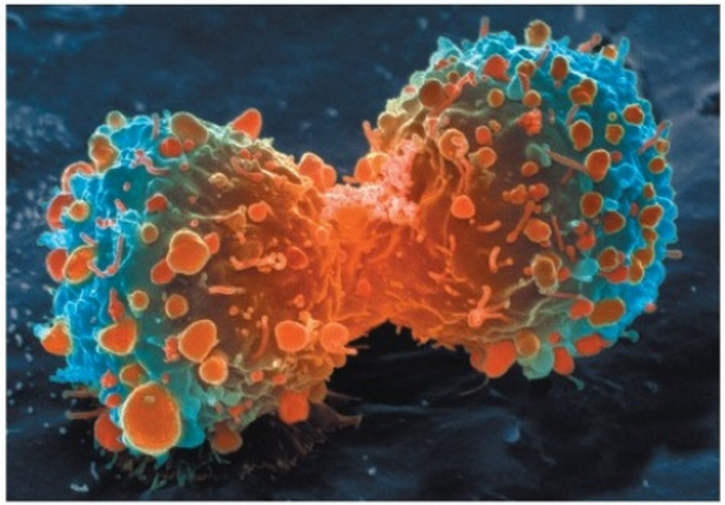
Is Reproducibiliy an Issue? Studies on reproducibility

- 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…



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 approach 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 to

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

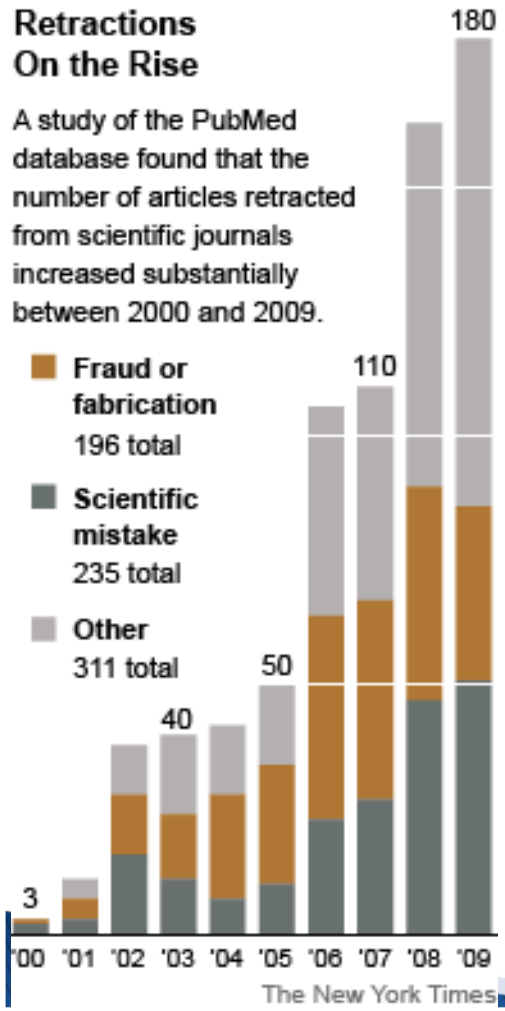
Six red flags for suspect work

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…



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 have to be developed

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

  → Scientific workflows
- Data sources contains errors
- Need standardization

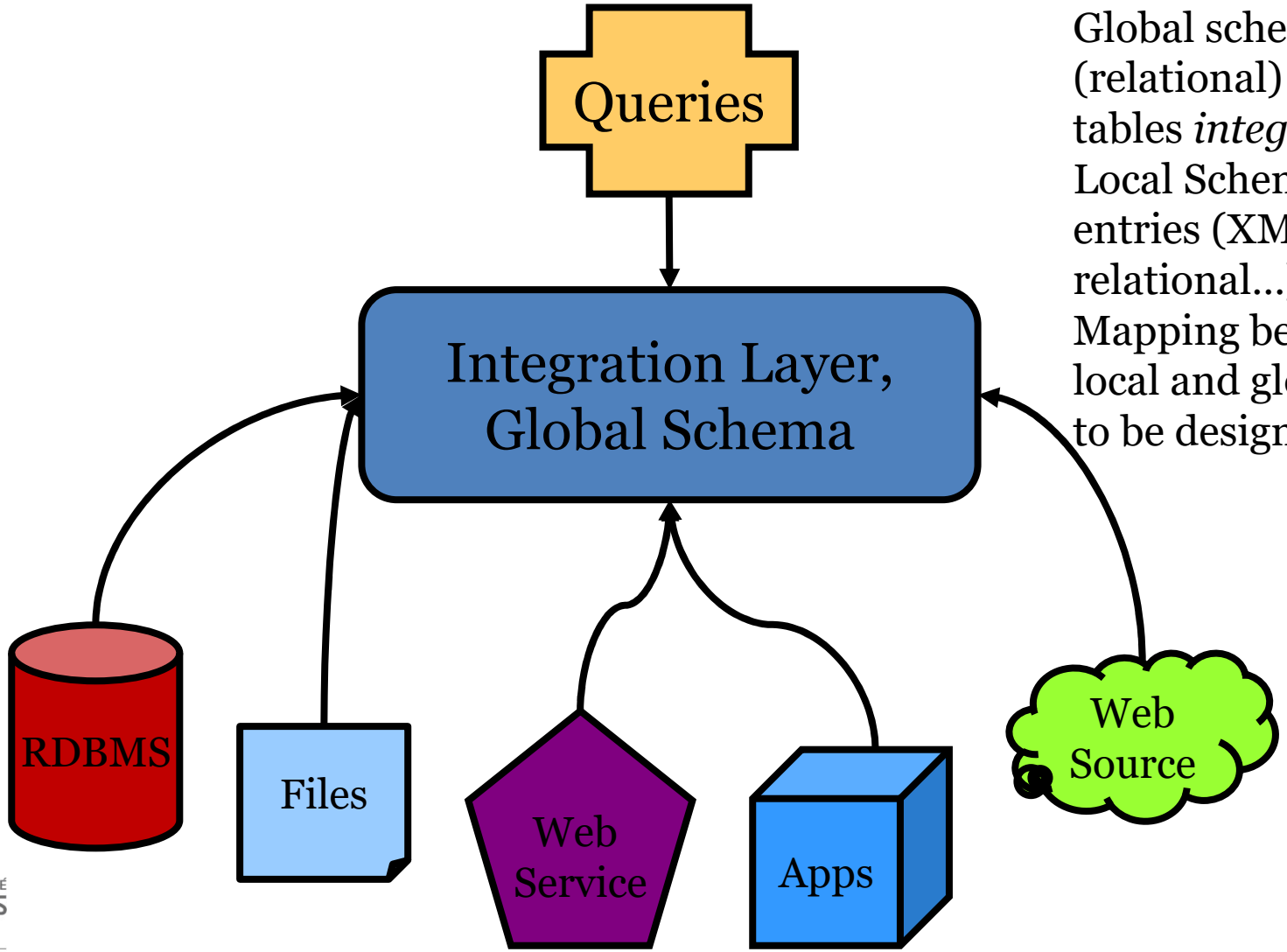
  → Ontologies

This Tutorial

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

- Part II – Scientific Workflows

Integration -- Classical View

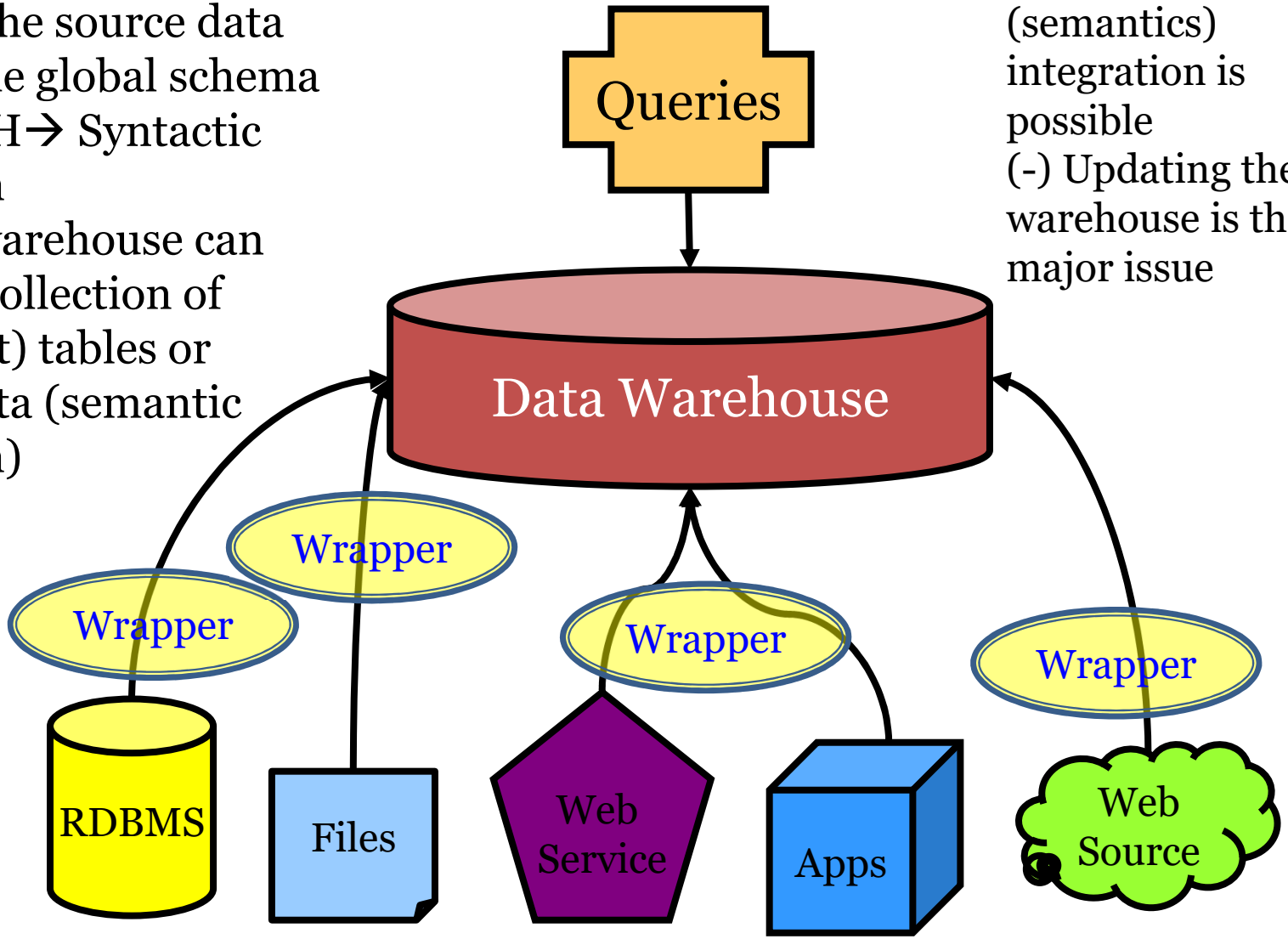


Global schema (relational) = set of tables *integrated*
Local Schema = set of entries (XML, relational…)
Mapping between the local and global schemas to be designed

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

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

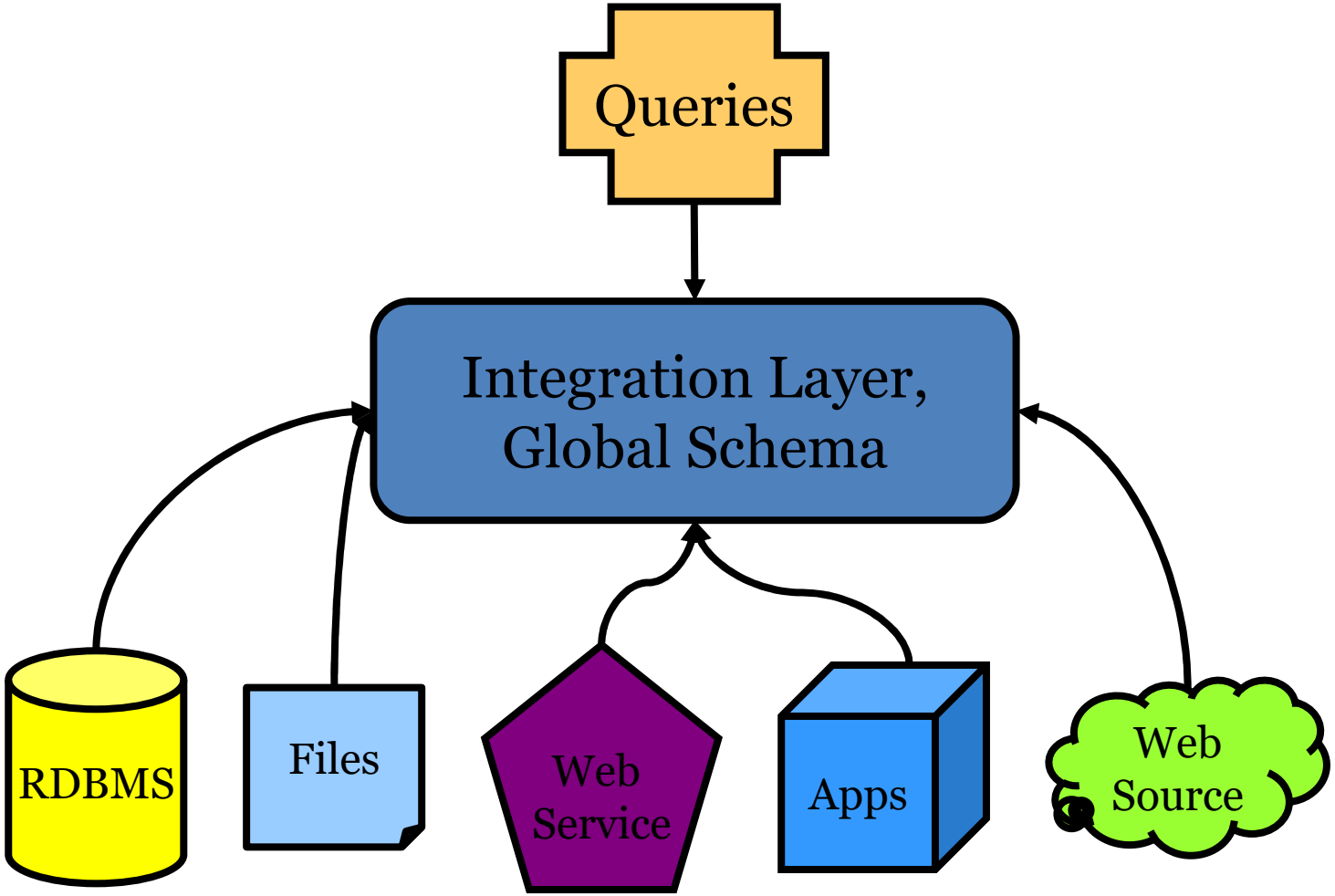
- Part I – Data Integration for the Life Sciences
  - Biological data & biological databases
  - Some Myths, some Truths
  - Presence

- Part II – Scientific Workflows

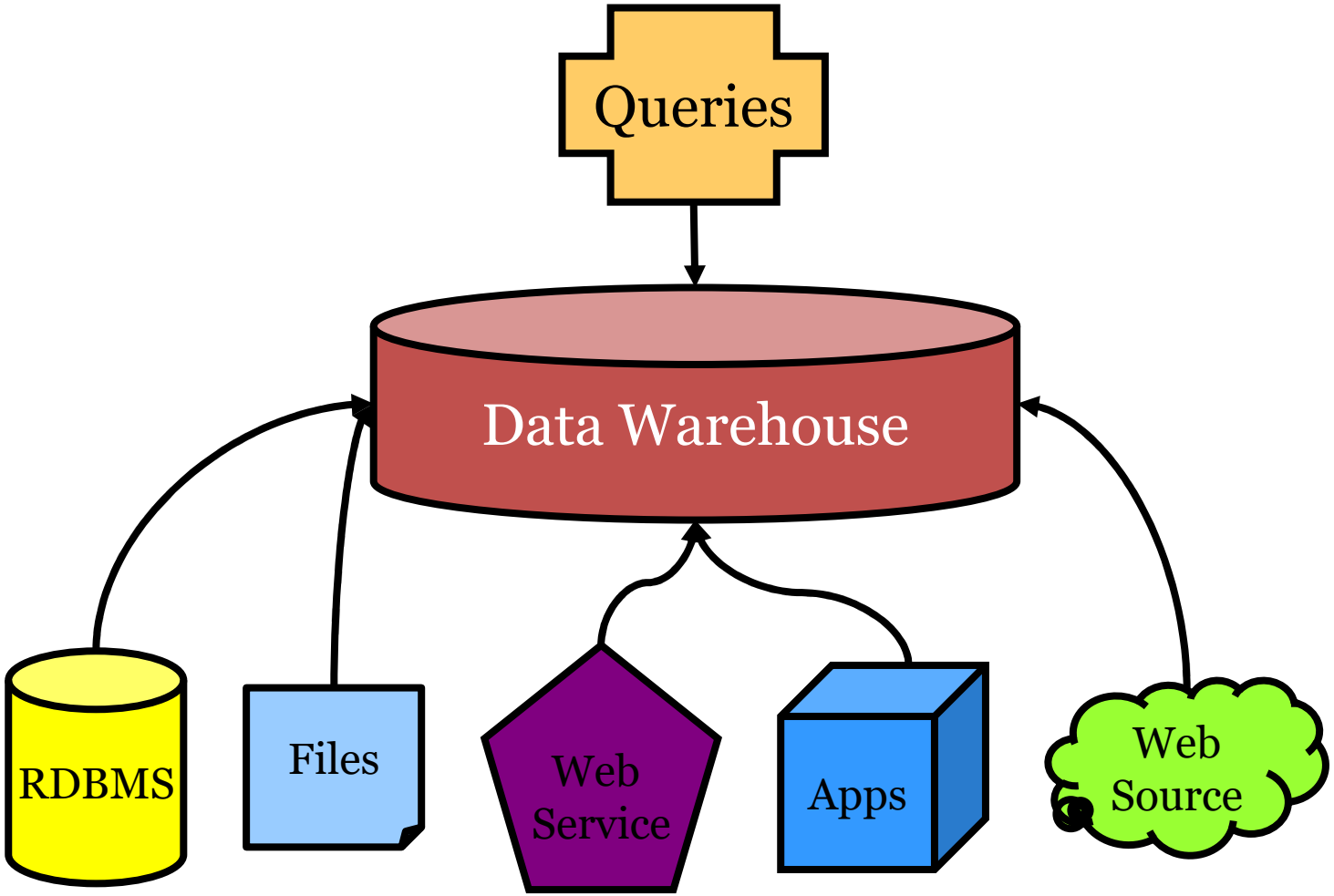
Trend

Analysis is integration and integration is analysis

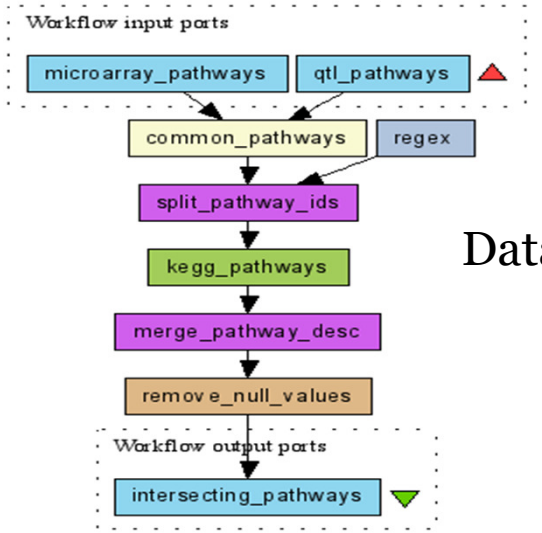
Integration Classical View (recall)



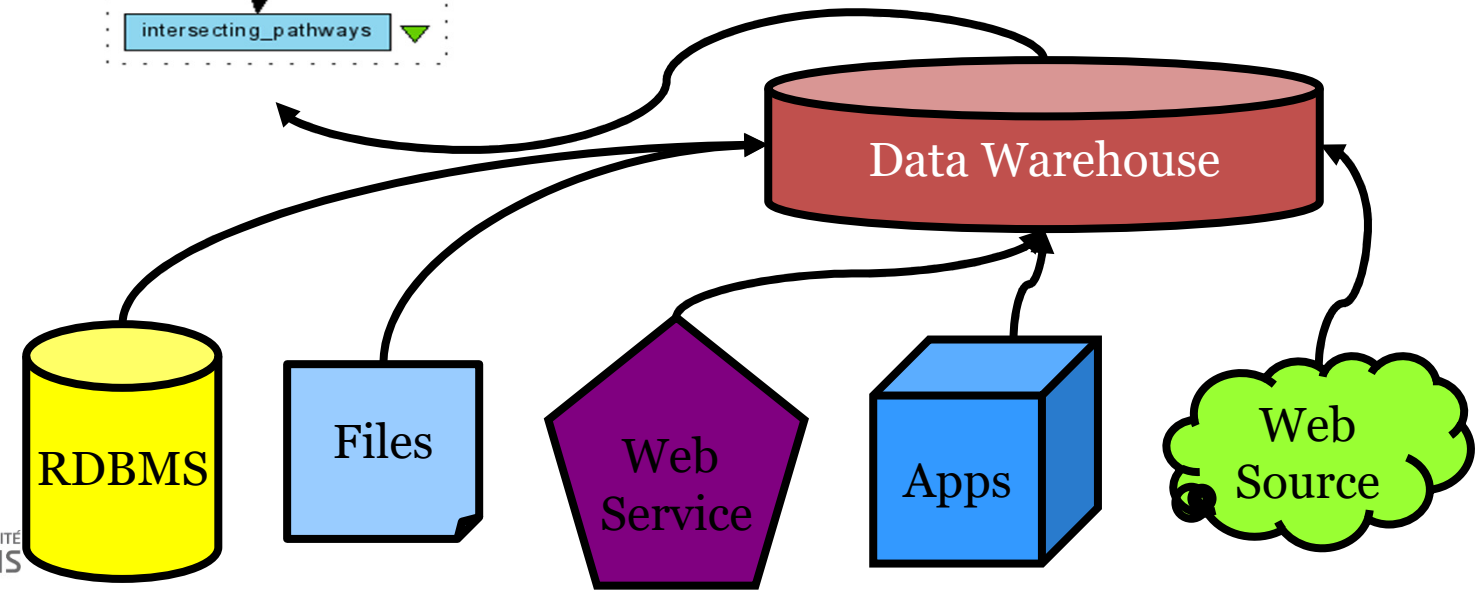
Classical View - DWH



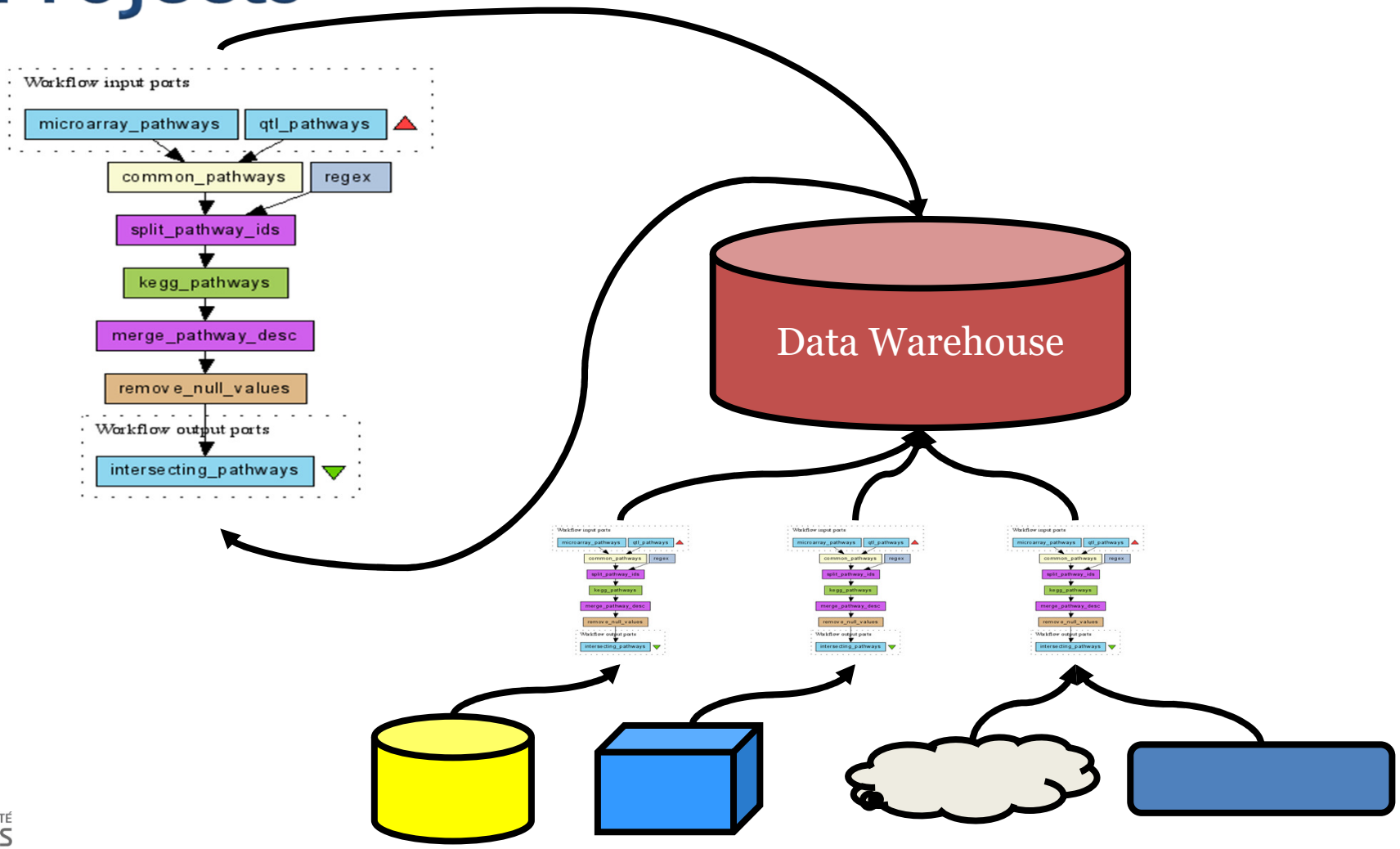
Classical View - Expanded



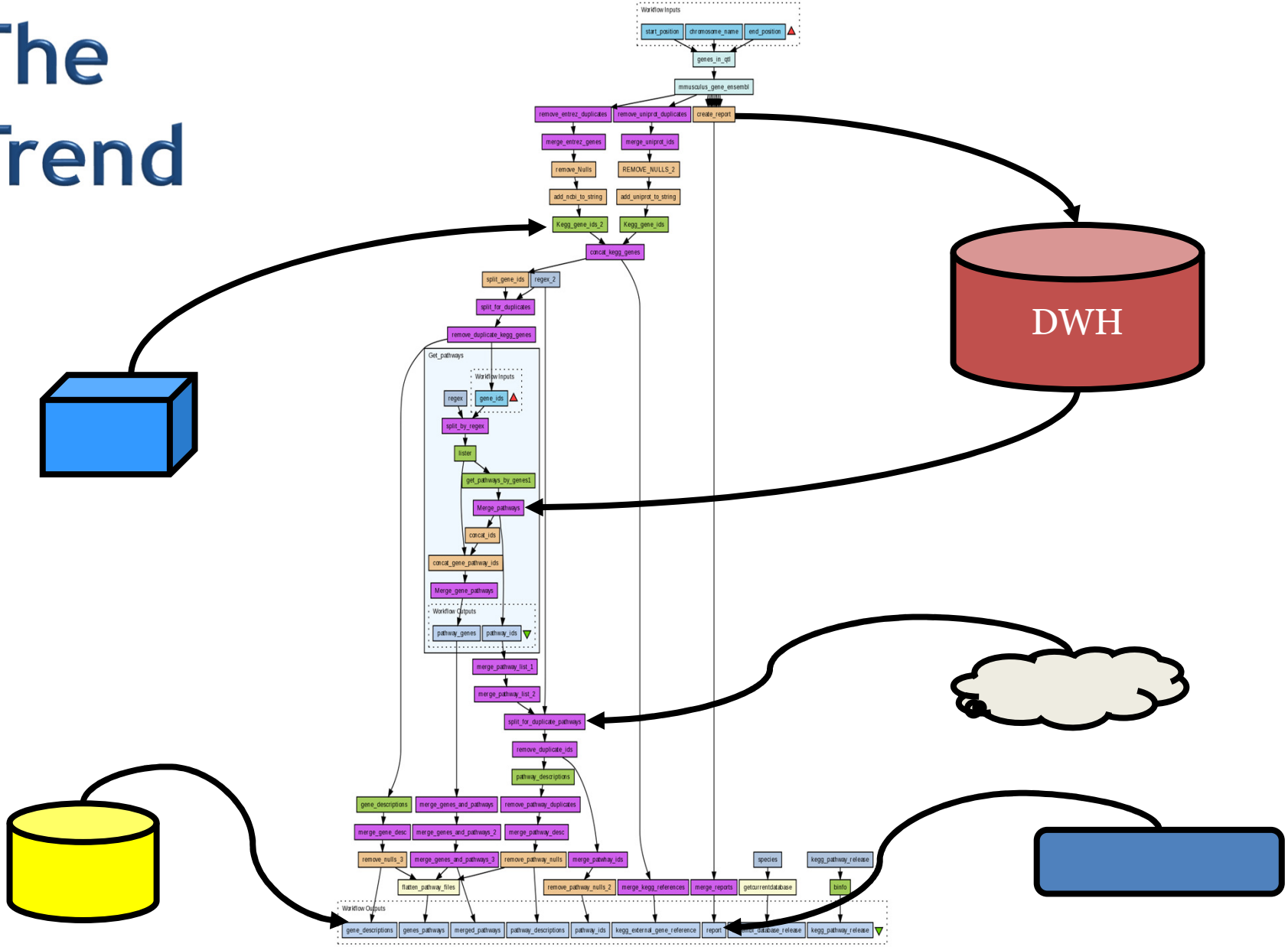
Data integration and analysis workflow



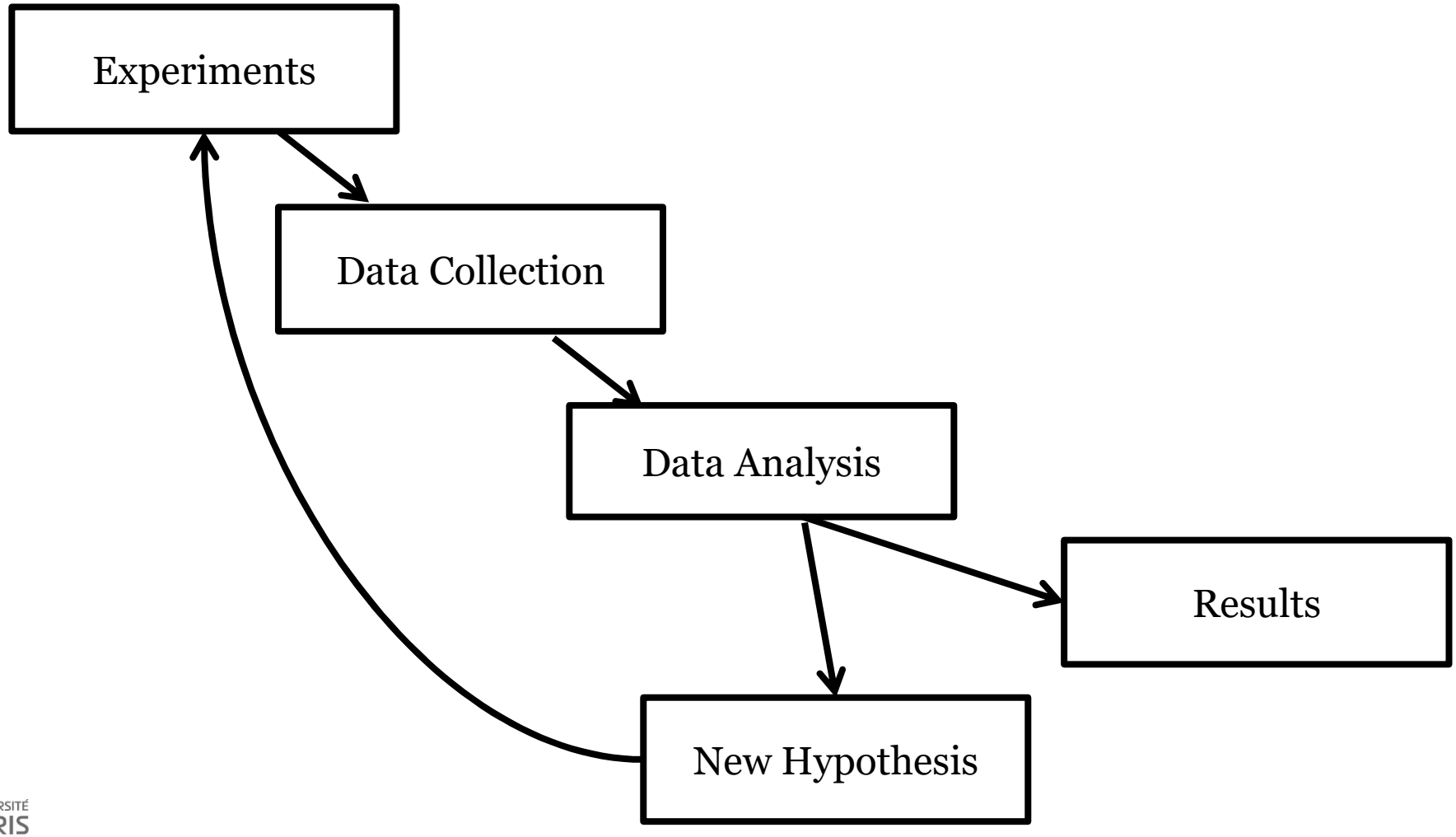
The True Architecture in Many Projects



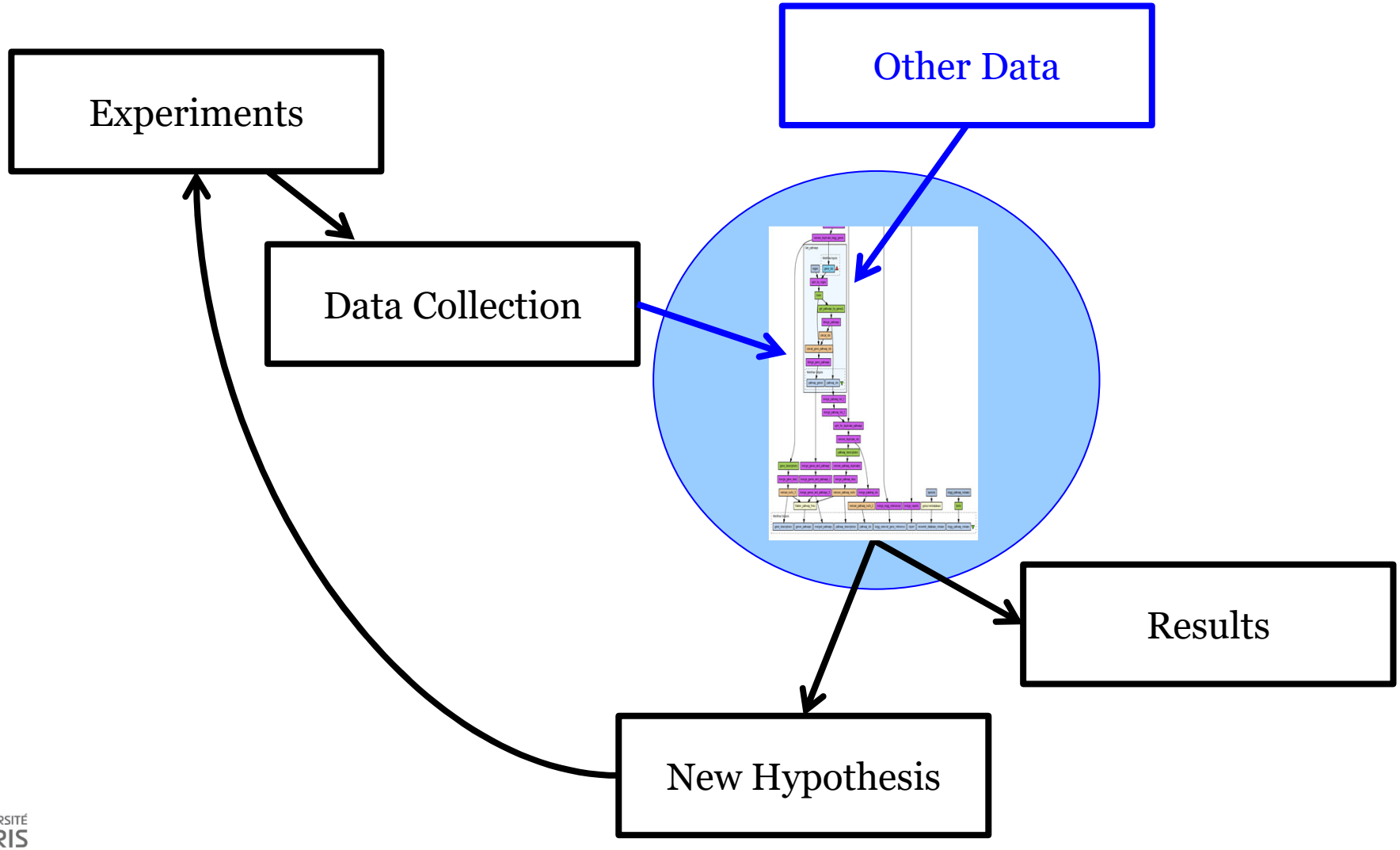
The Trend


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_pathway_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
kegg_database_release
kegg_pathway_release
DWH

Life Science Research Food Chain



With DI Workflows



Scientific workflow systems (specification)

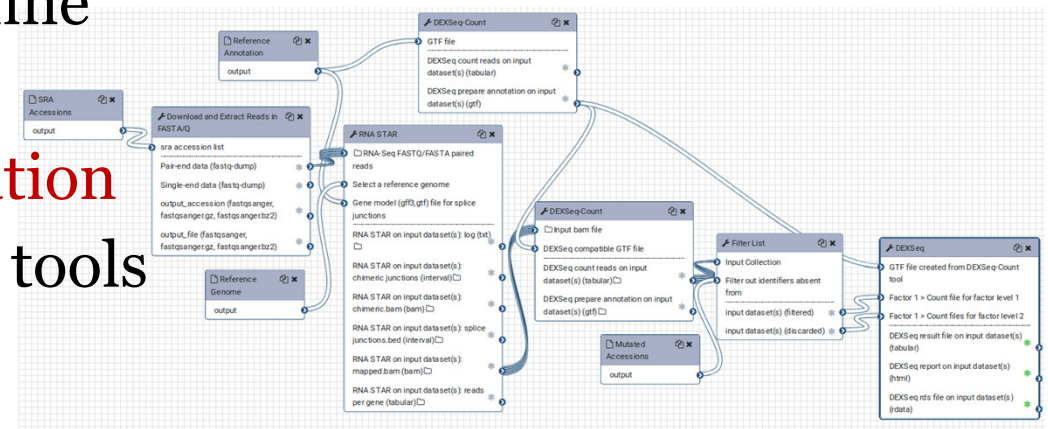
SWFS = “Data analysis pipeline”

Data flow driven

Encapsulation & Modularization

WF specification: connected tools

steps of the analysis



Encapsulation

Scripts are contained into boxes (steps)

Prog. Interface: input, parameters, output

Unified representation of steps

Modularization

Steps are independent of each others’

→ reusability

Scientific workflow systems (execution)

WF execution: data consumed/produced

Transparent, optimized, Traceable

SWFS scheduling, logging

Transparent

Able to run in any environments

Optimized

Able to run on different contexts (cluster, desktop, …)

Traceable

Keep track of the data consumed & produced during the execution

Provenance modules → *data management*

Scientific workflow systems (wrap-up)

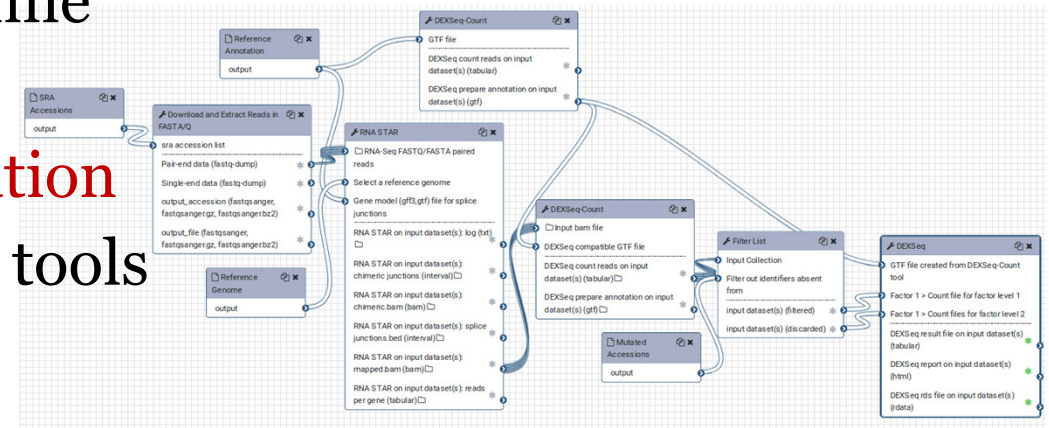
SWFS = “Data analysis pipeline”

Data flow driven

Encapsulation & Modularisation

WF specification: connected tools

steps of the analysis



WF execution: data consumed/produced

Transparent, optimized, Traceable

data management

Mature systems: Galaxy, NextFlow, SnakeMake…

The Galaxy Project

- Galaxy is an open source, web-based platform for data intensive biomedical research.
- The Galaxy Team is a part of
  - the Center for Comparative Genomics and Bioinformatics at Penn State,
  - the Department of Biology and at Johns Hopkins University.
- The Galaxy Project is supported in part by
  - NSF,
  - NHGRI,
  - The Huck Institutes of the Life Sciences,
  - The Institute for CyberScience at Penn State,
  - and Johns Hopkins University…
- Can be used with
  - the free public server (usegalaxy.org)
  - or other instances (several in France: Institut Curie, Institut Pasteur, Genouest, SouthGreen…)

Galaxy main concepts

https://wiki.galaxyproject.org/Learn

- Pages: documentation within Galaxy. To supplement publications or to present tutorials.

- Workflows: define the steps in an analysis process. Workflows are analyses that are intended to be executed (one ore more times) with different user-provided input Datasets.
  Steps come from the toolshed.

Workflow specification

- Histories are analyses records in Galaxy that show all input, intermediate, and final datasets, as well as every step in the process and the settings used with each job executed.
- Datasets represent individual files or jobs included within a History.
- Data Libraries are collections of Datasets accessible. Designed for sharing datasets in between users or groups.

Workflow execution

Other major workflow systems

- Taverna http://www.taverna.org.uk/
  - Pioneer, Univ. Manchester
  - Perfect to combine Web services

  → Not used anymore
- nextflow https://www.nextflow.io/
  - Programmation-oriented (no GUI)
  - Increasingly used
  - Able to represent the specification with arcs labelled with data files names
- Snakemake https://snakemake.readthedocs.io
  - Programmation-oriented (no GUI)
  - Need to understand make commands ;)
  - The workflow is described as a set of rules
  - Ability to visualize the execution graph

And many others…. !

- Kepler (https://kepler-project.org/, BioKepler)
- Pegasus (http://pegasus.isi.edu/, Cloud ++)
- Mobyle (http://mobyle.pasteur.fr/)
- OpenAlea (http://openalea.gforge.inria.fr, Plants ++)
- RapidMiner (https://rapidminer.com/)
- WINGS (http://www.wings-workflows.org/, semantics)
- KNIME (https://www.knime.org/)
- Cunieform (works on Hadoop YARN…)

Different systems for different users

- Snakemake & Nextflow
  + Excellent systems for programmers (prototyping)
  + Transparency, optimization of execution
  - Impossible to be used by end-users
  - Re-use, exchange /sharing

- Galaxy
  + Excellent system for end-users having admins ☺
  → 2 kinds of users: programmers(admins) and end-users
  + Provides toolsheds containing tools already encapsulated
  →end-users must use the tools available or ask admins
  + easy to share/exchange/reuse workflows within the same toolshed

This Tutorial

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

- Part II –Data Integration workflows
  - What are scientific workflow systems
  - Designing a workflow from scratch
  - Repositories of workflows and web services (reuse)
  - workflows and reproducibility
  - Current challenges

Scientific Workflow Repositories

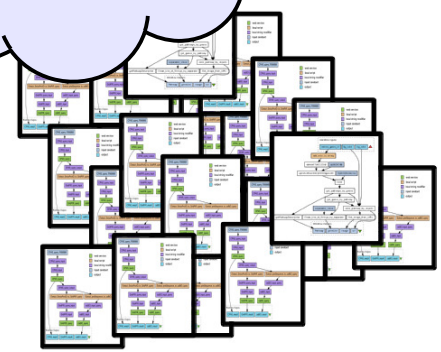
- Upload a scientific workflow
- Search, download & reuse existing scientific workflows
- Most specifically for single workflow system

Scientific Workflow Discovery

Pose keyword query

Search in textual annotations

?



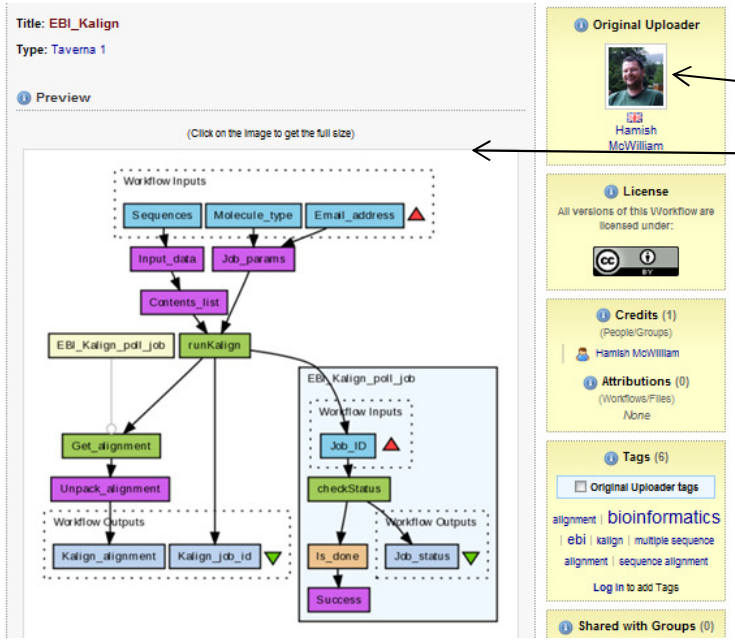
List of 10s or
100s of workflows

Find appropriate workflows

Reuse scientific workflow

myExperiment

- myExperiment.org
- Looking for workflows
  - By keywords
    - BioAID… workflow
    - Inspecting meta-data (author, favourited by, history…)
  - By authors
  - By group
  - …



- Conceptor
- Workflow
- Annotations
- …

Bio.tools (replaces BioCatalogue)

https://bio.tools/

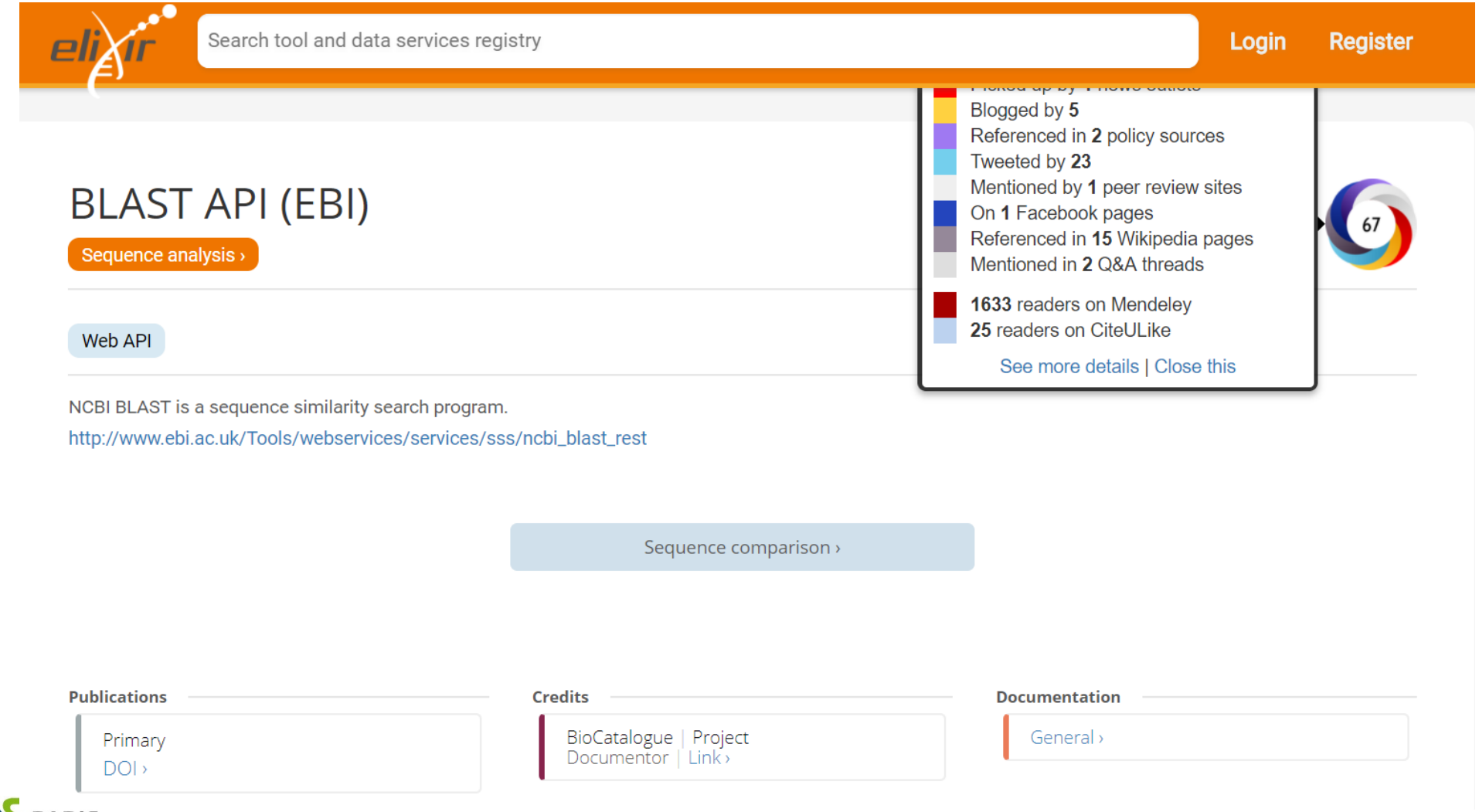
- Registry of Tools for the Life Sciences
  - find, understand, compare and select resources == **discovery**
  - use and connect them in workflows == **(inter)operability**

- Leaded by ELIXIR (European network of Excellence)

- Each tool must be described using biotoolsSchema
  - a formalized XML schema (XSD) which defines a description model for bioinformatics software (inputs, outputs and operations)
  - EDAM Ontology Terms are used

- EDAM Ontology
  - bioinformatics types of data including identifiers, data formats, operations and topics

Description of Tools in Bio.Tools



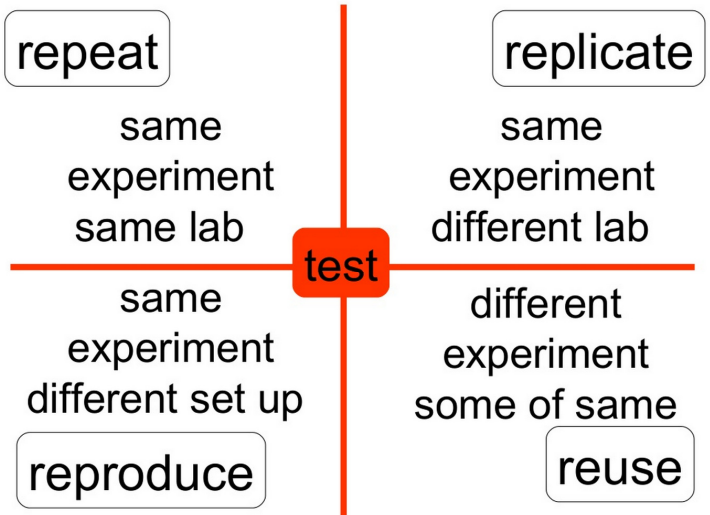
This Tutorial

- Part I – Data Integration workflows
  - What are scientific workflow systems
  - Designing a workflow from scratch
  - Repositories of workflows and web services (reuse)
  - Workflows and reproducibility
  - Current challenges

- Part II – Ranking Biological data
  - Ranking criteria
  - Introducing ranking into integration solutions
    - Data warehouses
    - Portals

- Part III – Conclusions

Repro with Workflows: ingredients and levels



Drummond C Replicability is not Reproducibility: Nor is it Good Science, online
Peng RD, Reproducible Research in Computational Science *Science 2 Dec 2011: 1226-1227.*

3 ingredients

Workflows Specification

Chained Tools

Workflow Execution

Input data and parameters

Workflow Environment

OS/librairies …

- Repeat
  - *Redo*: exact same context
  - Same workflow, execution setting, environement
  - Same *output*

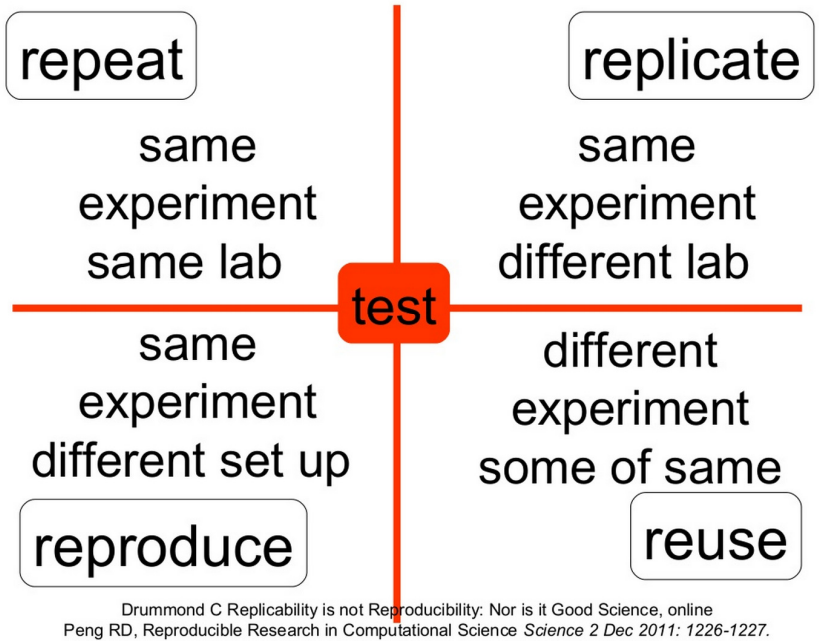
  →Aim = proof for reviewers ☺

- Replicate
  - Variation allowed in the workflows, execution setting, environement
  - Similar *output*

  → Aim = robustness

A continuum of possibilities

- Reproduce
  - Same *scientific result*
  - But the means used may be changed
  - Different workflows, execution setting, environment
  - Different output but in accordance with the result



- Reuse
  - Different scientific result
  - Use of tools/… designed in another context

Scientific workflow systems (reminder)

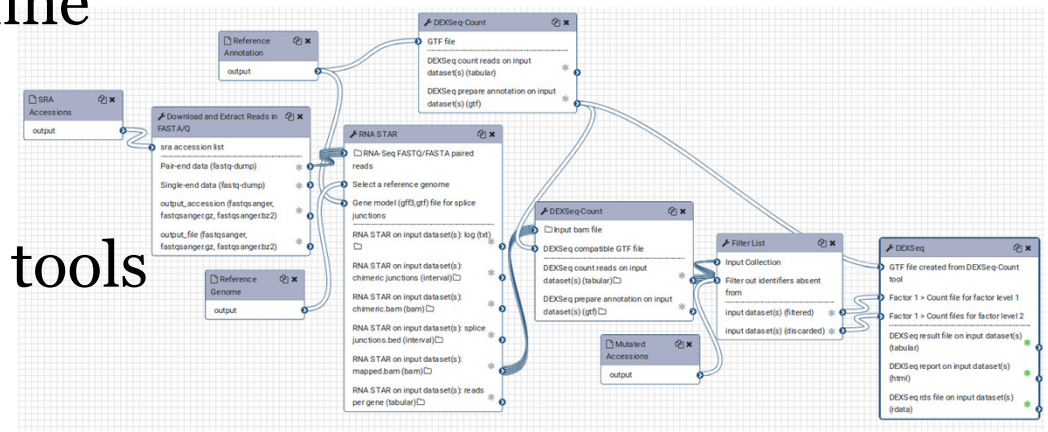
SWFS = “Data analysis pipeline”

Data flow driven

Encapsulation of scripts

WF specification: connected tools

steps of the analysis



WF execution: data consumed/produced

Provenance modules

data management

SWFS scheduling, logging,

May be equipped with GUI

Galaxy, NextFlow, SnakeMake…

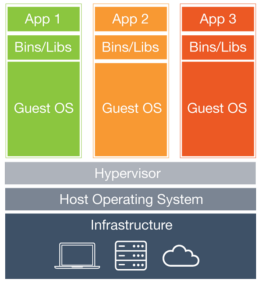
Capturing the programming environment

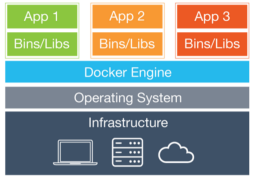
Ensuring your workflow has everything it needs to run

Libraries, dependencies… → *Transparent execution*

Virtual machines capture the programming environment

Container solutions





- package an application
  - with all of its dependencies
  - into a standardized unit for software development

include the application and its dependencies

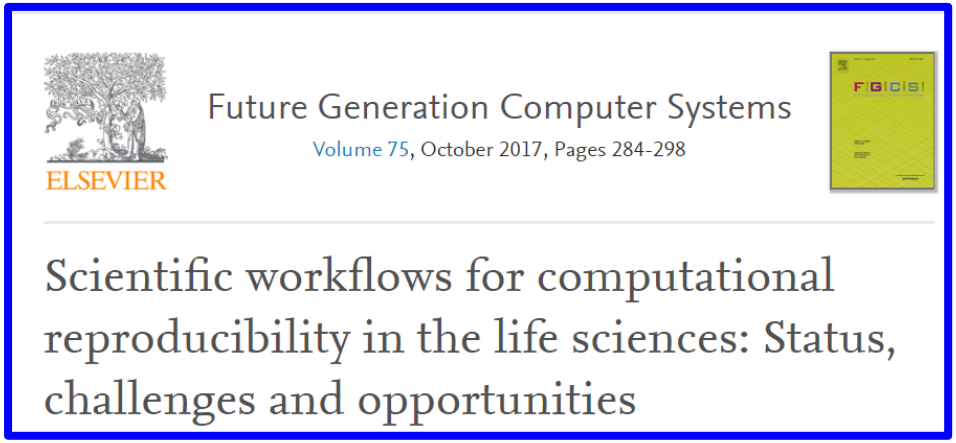
- but share the kernel with other containers
- They
  - are not tied to any specific infrastructure;
  - run on any computer, on any infrastructure and in any cloud

Lighter solution than classical VM

→ BioContainers: a registry of containers!

Reproducibility-friendly features

6 Systems: Galaxy, Nextflow, SnakeMake, VisTrails, OpenAlea, Taverna


ELSEVIER

Future Generation Computer Systems

Volume 75, October 2017, Pages 284-298

Scientific workflows for computational reproducibility in the life sciences: Status, challenges and opportunities

Specification

Language (XML, Python…)

Interoperability (CWL…)

Description of steps

- Remote services
- Command line
- Access to source code

Modularity (nested workflows?)

Annotation (tags, ontologies, myexperiment…)

Execution

Language and standard (PROV…,) → repeat … reuse

Presentation (interactivity with the results/provenance, notebooks) → replicate … reuse

Annotations → reuse

Environment

Ability to run workflows within a given environment

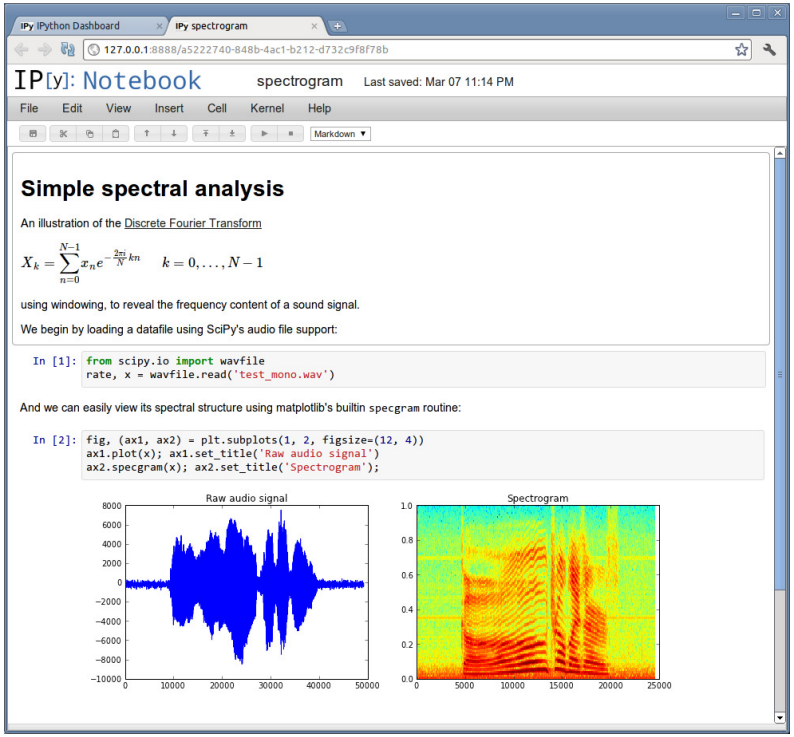
Virtual machines

- VMWare, KVM, VirtualBox, Vagran,…

Lighter solutions (containers)

- Docker, Rocket, OpenVZ, LXC, Conda

Another kind of systems: Notebooks



- Web-based interactive computational environment
- Combination of code execution, text, mathematics, plots and rich media into a single document
- Some systems export workflow execution as executable Jupyter papers…

Excellent mean to explain/present a scientific results obtained

Ten Simple Rules for Reproducible Computational Research (PlosOne)

- 1: For Every Result, Keep Track of How It Was Produced
- 2: Avoid Manual Data Manipulation Steps
- 3: Archive the Exact Versions of All External Programs Used
- 4: Version Control All Custom Scripts
- 5: Record All Intermediate Results, When Possible in Standardized Formats
- 6: For Analyses That Include Randomness, Note Underlying Random Seeds
- 7: Always Store Raw Data behind Plots
- 8: Generate Hierarchical Analysis Output, Allowing Layers of Increasing Detail to Be Inspected
- 9: Connect Textual Statements to Underlying Results
- 10: Provide Public Access to Scripts, Runs, and Results

→ Several ways to follow them

→ More or less complex (from manually to fully automatically)

→ More or less time-consuming (repeat, reproduce, …., reuse)

Wrap up

- Data Integration & Data Analysis
- Scientific workflows plays a major role to analyse bio data sets
- Major systems in place, large variety of solutions: Galaxy (GUI), SnakeMake/NextFlow (scripts)…
- Reproducibility and reuse is improved using such systems
  - Specification: which tools in what order
  - Execution: which data produced/consumed, which parameters
  - Environment: which OS, which librairies, …
- Notebooks are another very interesting solution (to expose/explain a scientific result)

This Tutorial

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

- Part II – Data Integration workflows
  - What are scientific workflow systems
  - Designing a workflow from scratch
  - Repositories of workflows and web services (reuse)
  - workflows and reproducibility
  - Latest results on workflows

  →Or How CS research may have direct impact on LS

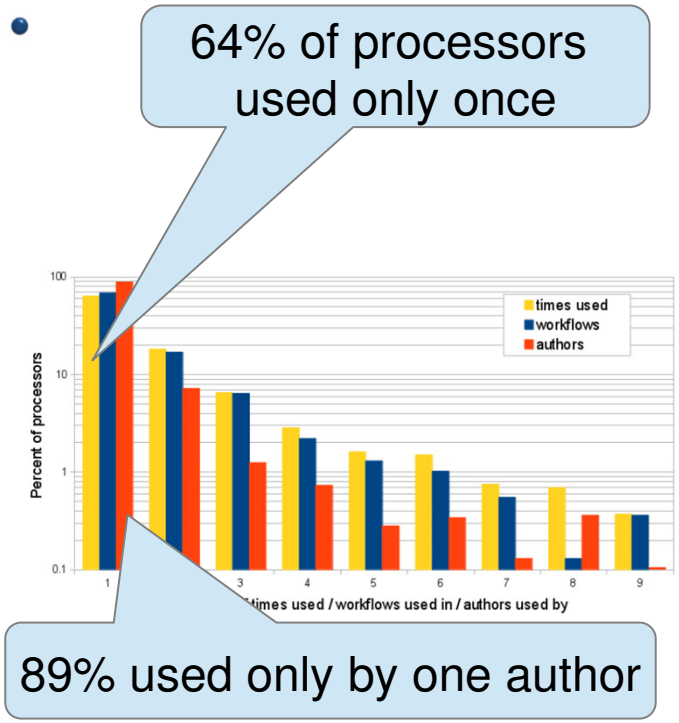
    Improving reuse

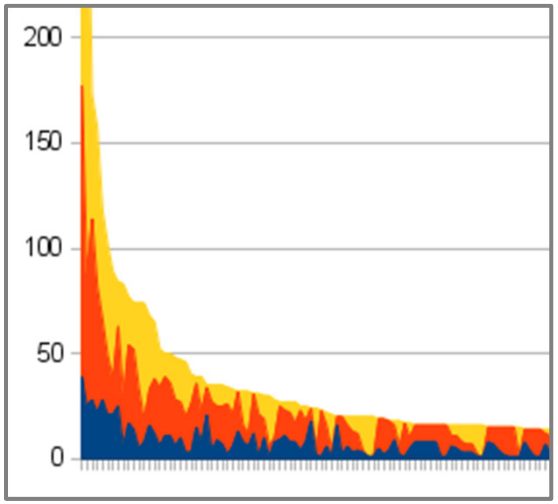
    Managing Provenance

    Comparing workflows executions

Study on workflow reuse….

- 36% of elements are re-used
  - These connect workflows quite densely
  - Can be exploited for repository IR
- Re-use rates have a Zipf-like distrib
  - Local : High re-use rates as-is
    - Web-Service : Authors have favorite services, unshared
    - Script & subworkflows : Authors have personal libraries
- True cross-author re-use is low: 3%
  - Authors have personal preferences & libraries
  - But don't use content from others

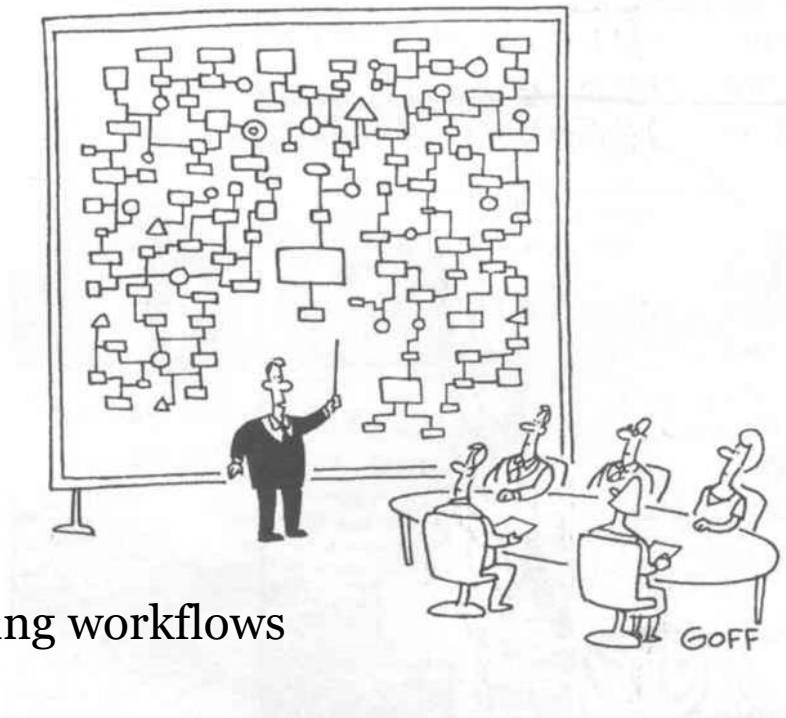




How to improve reuse?

Help finding *similar workflows*

Make workflow structures less complex!

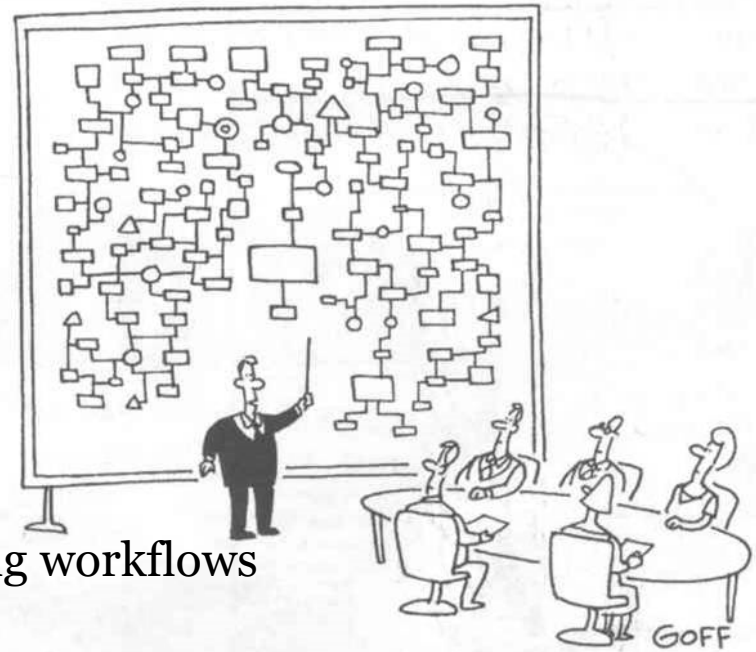


Plumbing workflows

How to improve reuse?

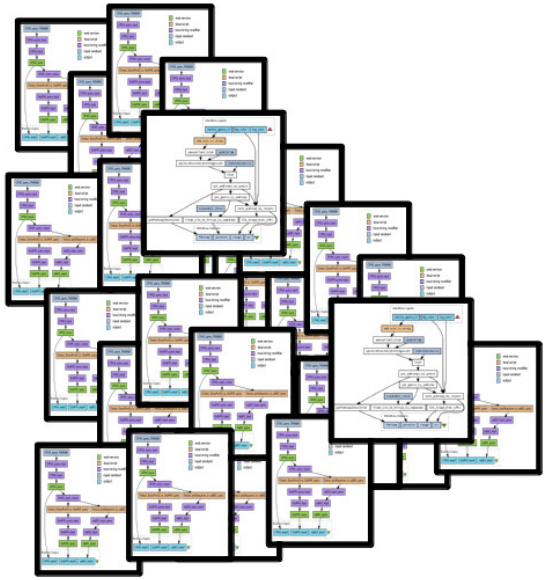
Help finding *similar workflows*

Make workflow structures less complex!



Plumbing workflows

Scientific Workflow Discovery Improvement



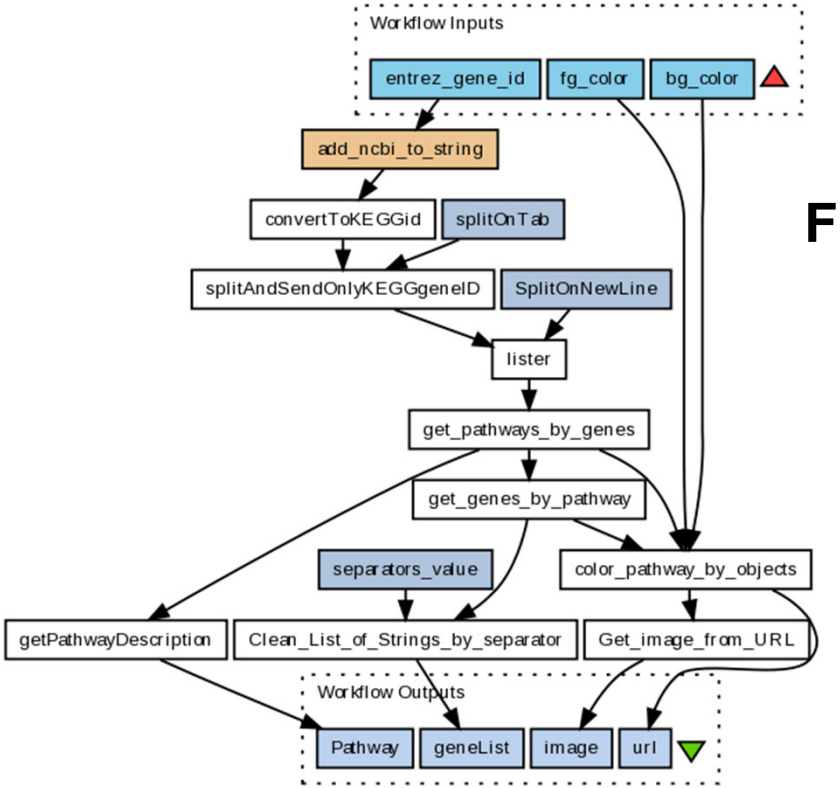
List of 10s or 100s of workflows

Goal

- Group results by similar workflows
- Search by sample workflow
- Provide recommendations
  - Similar workflows
  - Replacements
  - Extensions
- …

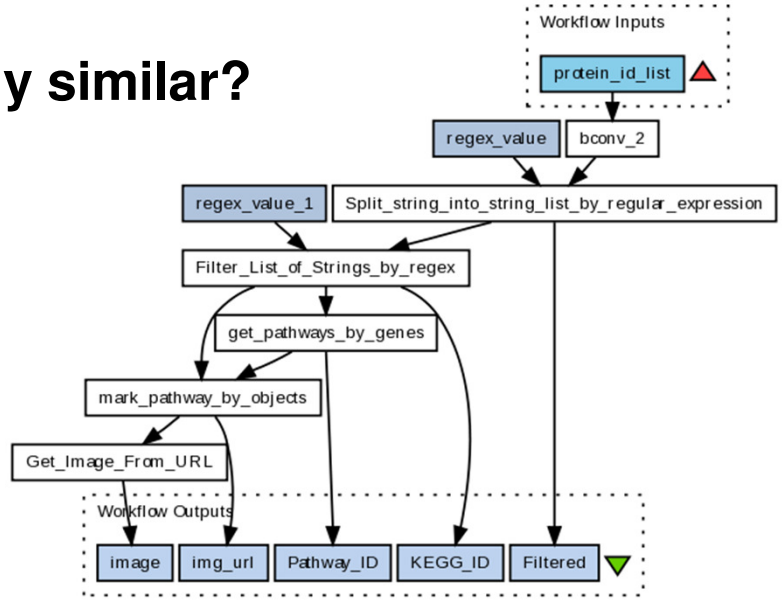
Need: **Similarity Measures**

The Central Question



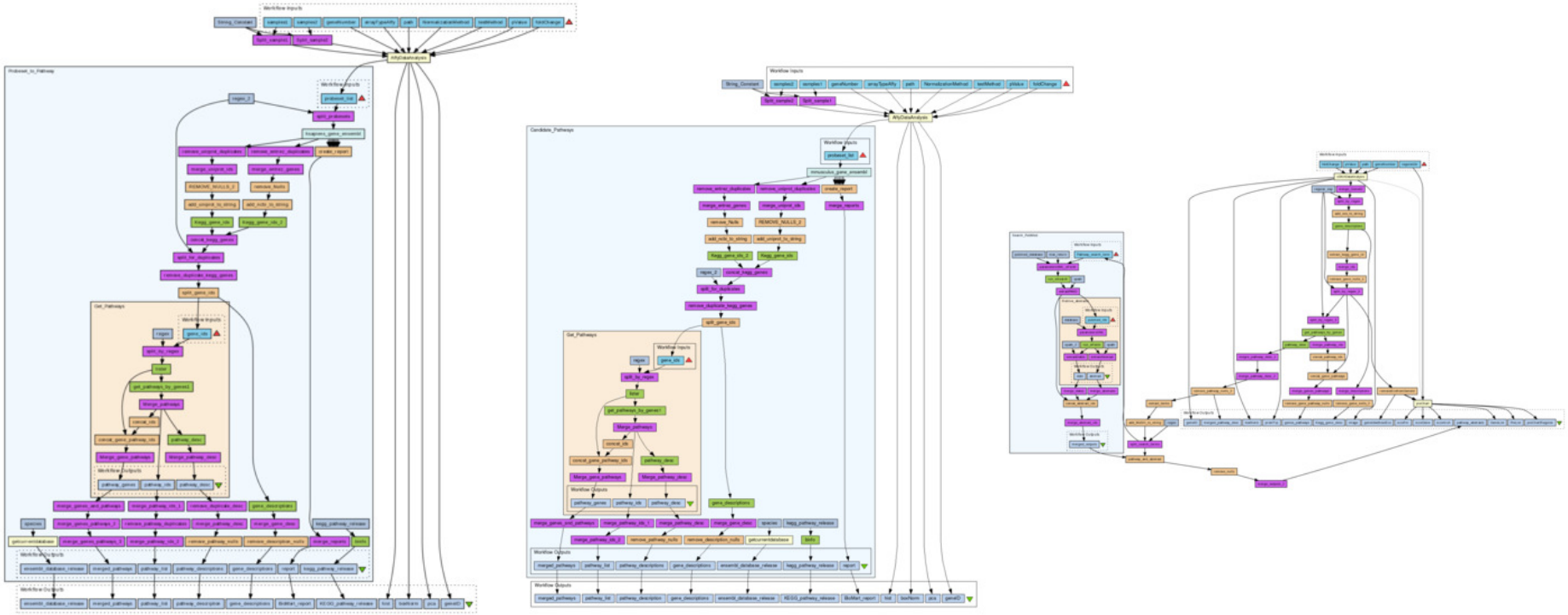
Functionally similar?

≈



Example

workflows perform microarray analysis integrating various sources (pathway DB, probe mapping, PubMed)

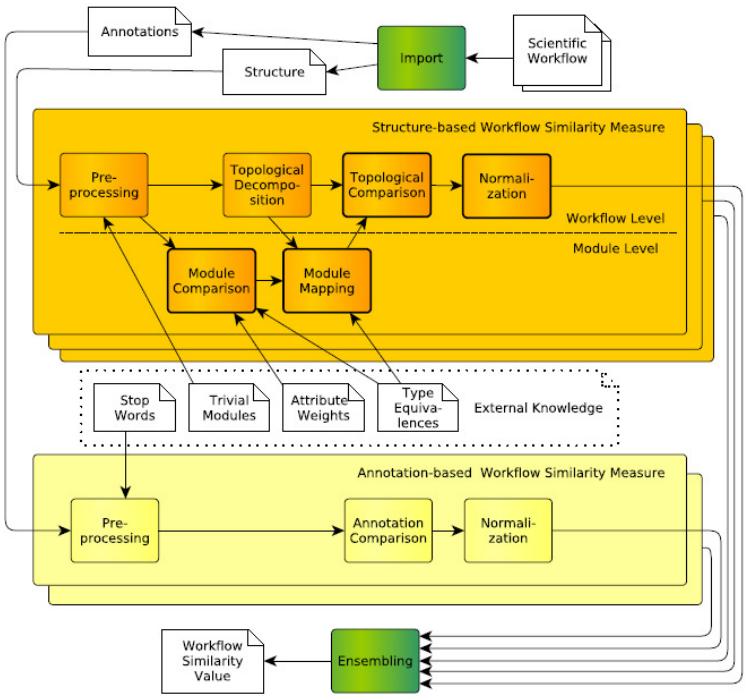


- All three workflows may be used
  - entirely (which fits best?) or partly (from probes to pathways)

Similarity search for scientific workflows

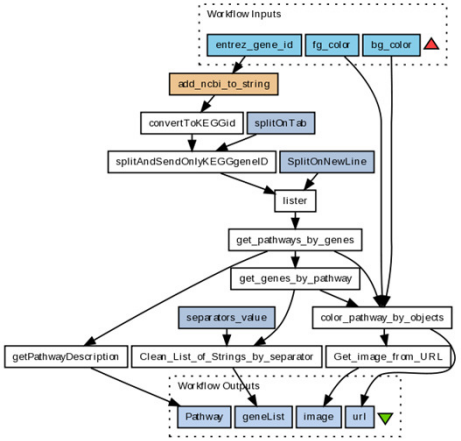
[VLDB 2014]

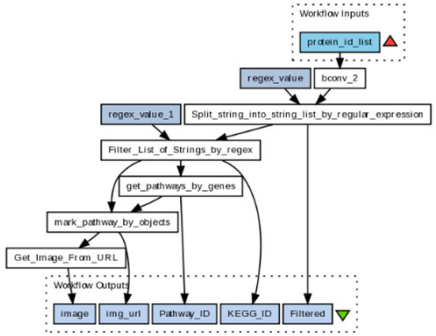
With Johannes Starlinger, Bryan Brancotte, Ulf Leser



- Framework
  - capture all the sim. search techniques
    - Structure-based
      - Graph struct. of the workflow
    - Annotation-based
      - Meta-data (description, tags…)
- Goal of the study
  - compare results obtained by all techniques
  - On various data sets
    - Taverna, Galaxy, VisTrails

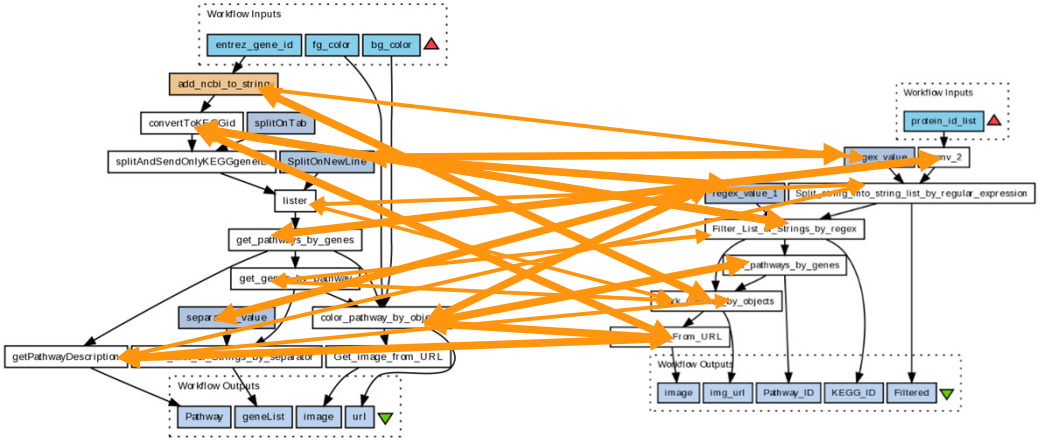
Subtasks of Scientific Workflow Comparison





Module Comparison

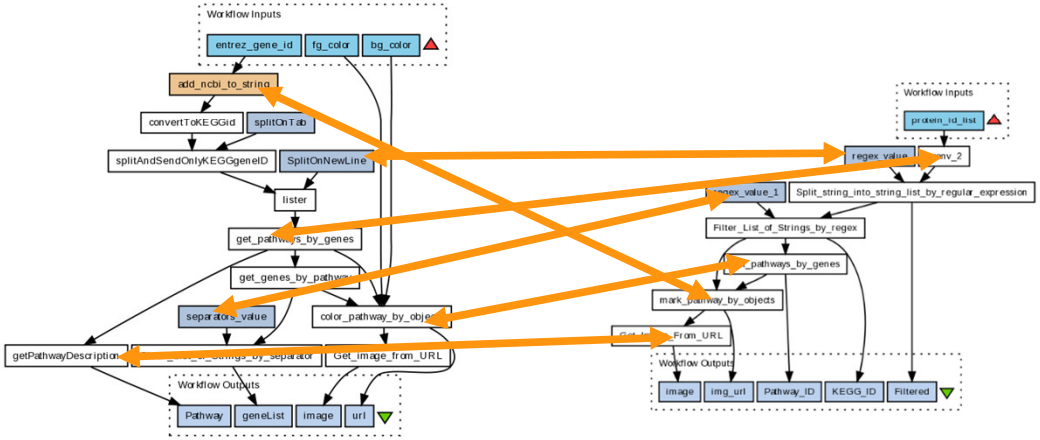
Subtasks of Scientific Workflow Comparison



Module Comparison

- Label
- Webservice Uri
- Scripts
- etc

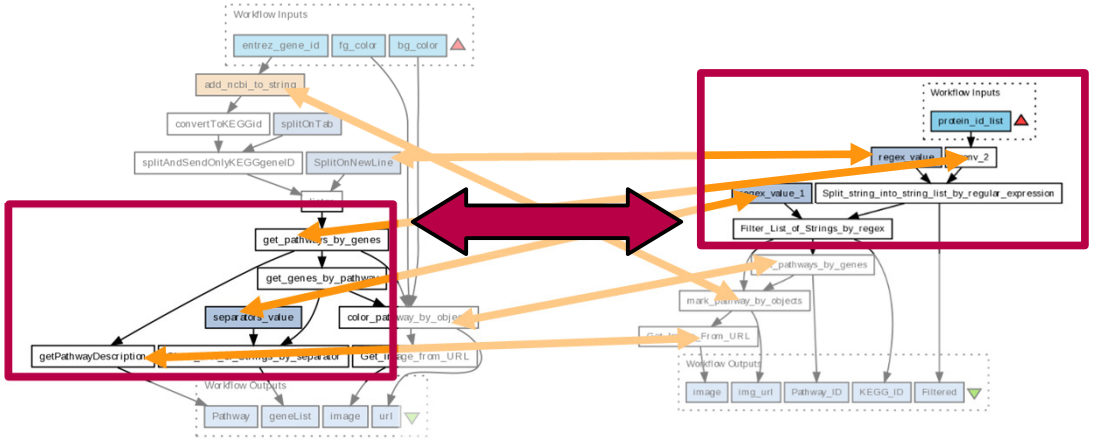
Subtasks of Scientific Workflow Comparison



Module Comparison → Module Mapping

- greedy
- maximum weight

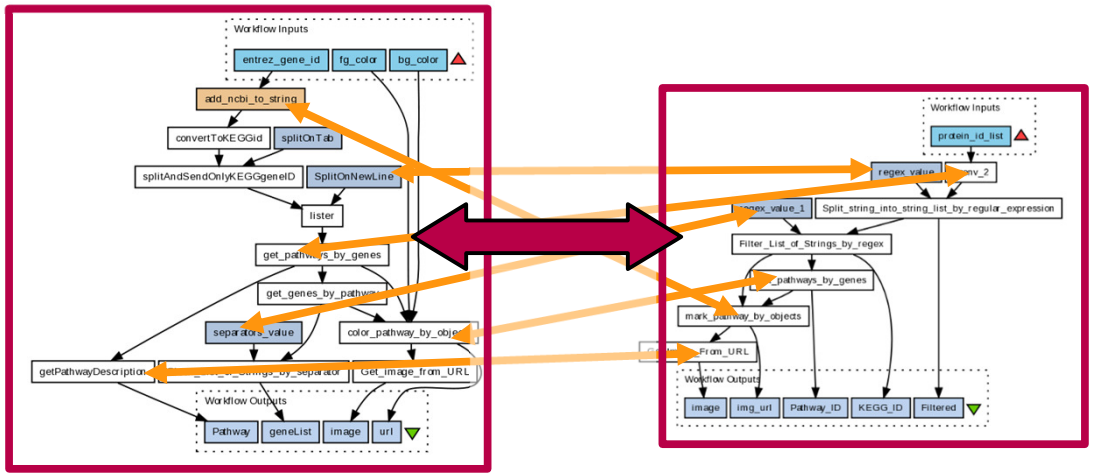
Subtasks of Scientific Workflow Comparison



Module Comparison → Module Mapping → Topological Comparison

- Set of Modules
- Substructures
- Full Structure

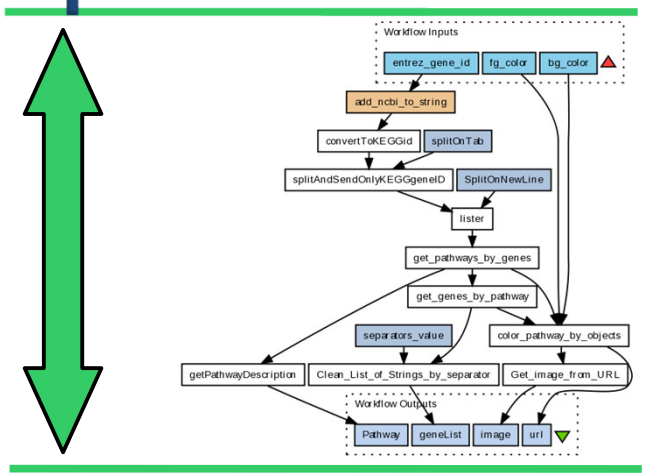
Subtasks of Scientific Workflow Comparison

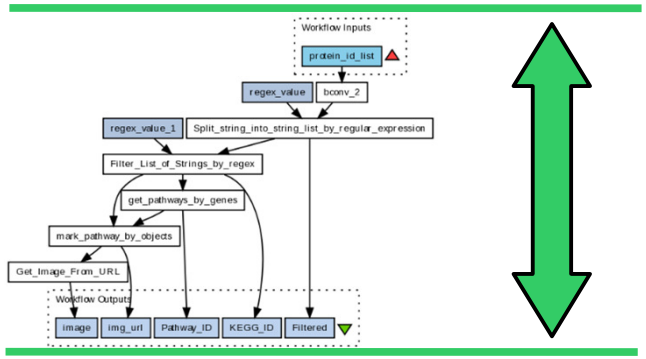


Module Comparison → Module Mapping → Topological Comparison

- Set of Modules
- Substructures
- Full Structure

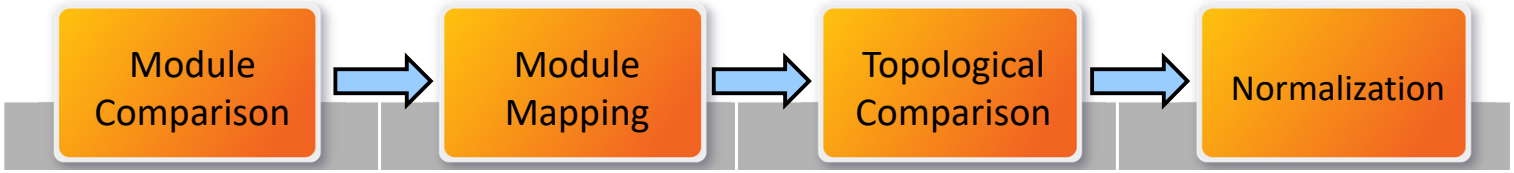
Subtasks of Scientific Workflow Comparison





Module Comparison → Module Mapping → Topological Comparison → Normalization

Existing Approaches



Stoyanovich et al.	
Silva et al.	
Bergmann et al.	
Santos et al.	
Goderis et al.	
Friesen et al.	
Xiang et al.	

MCS = Maximum Common Subgraph GED = Graph Edit Distance

Existing Approaches

	Module Comparison	Module Mapping	Topological Comparison	Normalization
Stoyanovich et al.	single attributes	-	modules	-
Silva et al.	multiple attributes	greedy	modules	|V| of smaller wf
Bergmann et al.	semantic annot.	max. weight	modules & edges	|V|+|E| of query wf
	label edit dist.	max. weight	modules & edges	|V|+|E| of query wf
Santos et al.	label matching	-	modules	-
	label matching	-	MCS	|V|+|E| of larger wf
Goderis et al.	label matching	-	MCS	-
	label matching	-	MCS	'workflow sizes'
Friesen et al.	type matching	-	modules	-
	type matching	-	MCS	-
	type matching	-	graph kernels	-
Xiang et al.	label matching	-	GED	-

MCS = Maximum Common Subgraph GED = Graph Edit Distance

Existing Approaches

	Module Comparison	Module Mapping	Topological Comparison	Normalization
Stoyanovich et al.	single attributes	-	modules	-
Silva et al.	multiple attri			r wf
Bergmann et al.	semantic ar			ery wf
	label edit di			ery wf
Santos et al.	label match			
	label match			ger wf
Goderis et al.	label match			
	label match			es'
Friesen et al.	type matchi			
	type matching			
	type matching	-	graph kernels	-
Xiang et al.	label matching	-	GED	-

What's best

At each step?

As a whole?

MCS = Maximum Common Subgraph GED = Graph Edit Distance

Expert Curated Similarity Corpus

FlowAlike – Scientific Workflow Similarity Evaluation starling@informatik.hu-berlin.de: dashboard | overview | help | logout

Reference workflow:

EBI_Kalign

Multiple sequence alignment using the Kalign tool. This workflow uses the EBI's WSKalign service (see http://www.ebi.ac.uk/Tools/webservices/services/kalign) to access the Kalign tool. The set of sequences to align and the molecule type (protein or nucleic acid) are the input, the other parameters for the search (see Job_params) are allowed to default.

Note: the WSKalign service used by this workflow is deprecated as of 21st September 2010 and should not be used in any new development. This service is will be retired during 2011. EBI's replacement Kalign services (REST or SOAP) should be used instead.

Are these 10 workflows similar to the reference?

EBI_NCBI_BLAST_with_prompts

Run a BLAST analysis using the EBI's WSNCBIBlast service (see http://www.ebi.ac.uk/Tools/webservices/services/ncbiblast). This workflow wraps the EBI_NCBI_BL… kflow to provide a basic

EBI_InterProScan

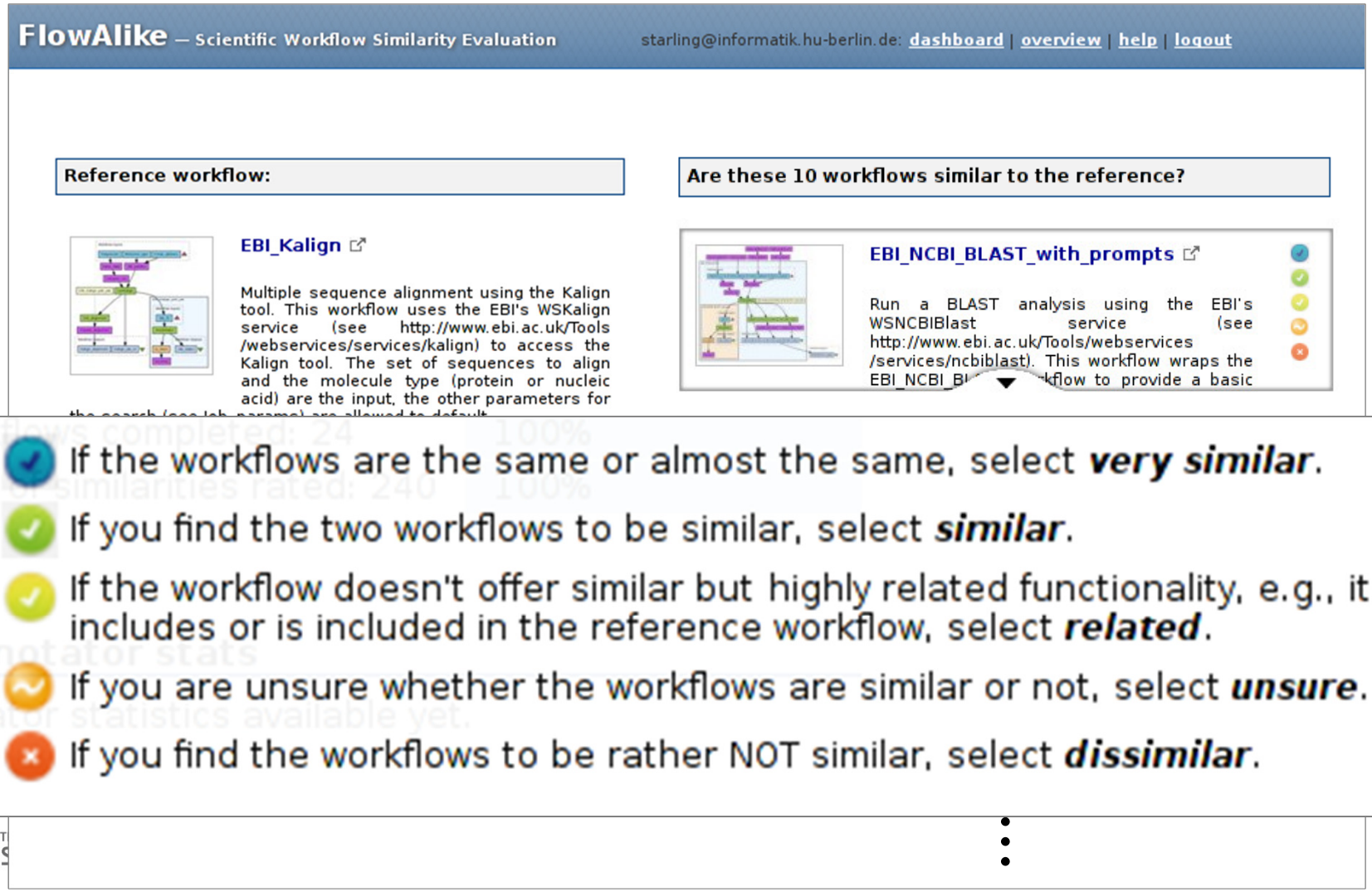
Note: the WSInterProScan web service used by this workflow is no longer available haveing been replaced by the EMBL-EBI's InterProScan (REST) (http://www.ebi.ac.uk/Tools/webs… ervices/pfa/iprscan_rest)

BioQuali asynchronous workflow

BioQuali: Network Compatibility and products variation inference in a biological network.

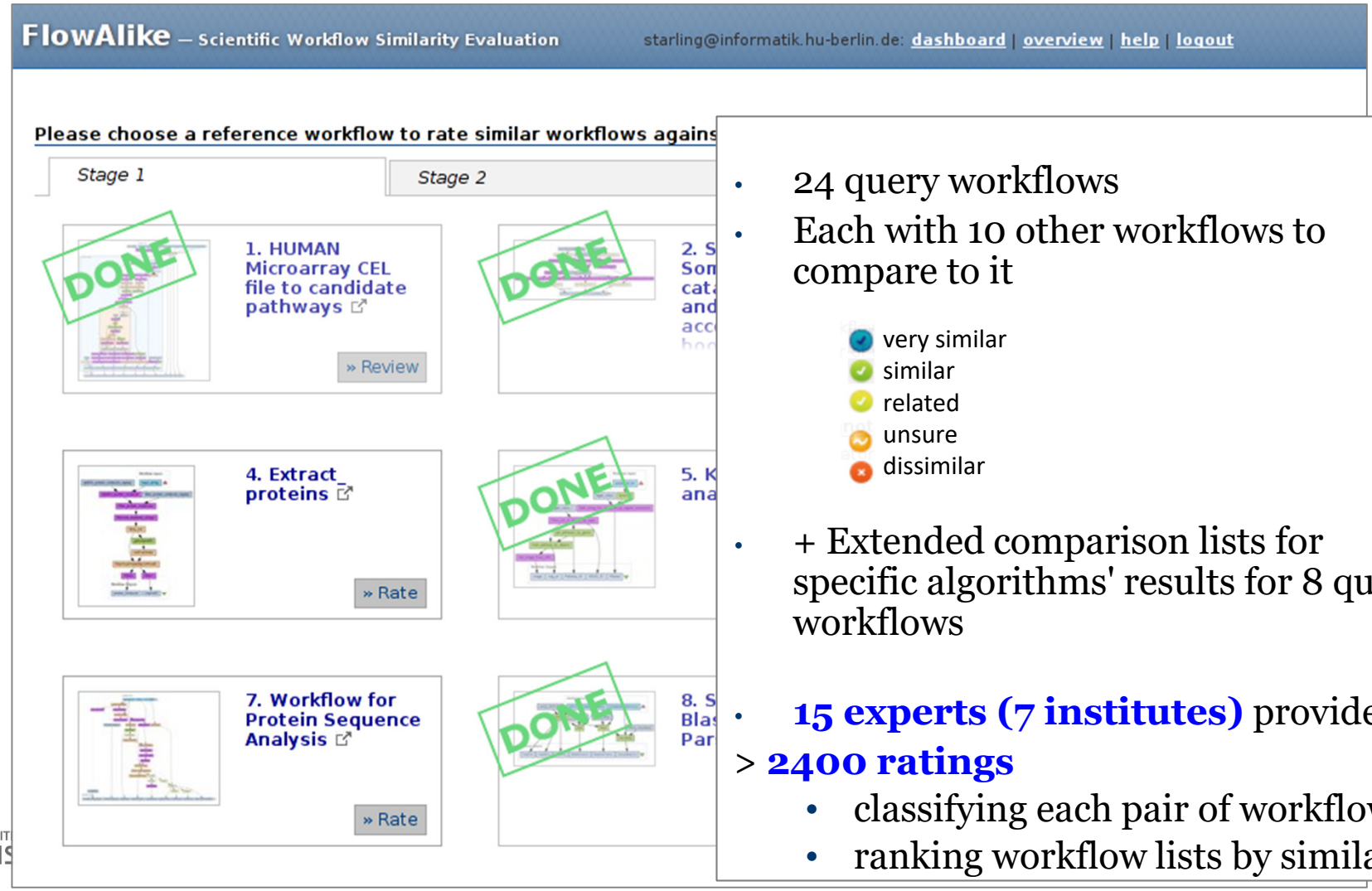
⋮

Expert Curated Similarity Corpus



- If the workflows are the same or almost the same, select ***very similar***.
- If you find the two workflows to be similar, select ***similar***.
- If the workflow doesn't offer similar but highly related functionality, e.g., it includes or is included in the reference workflow, select ***related***.
- If you are unsure whether the workflows are similar or not, select ***unsure***.
- If you find the workflows to be rather NOT similar, select ***dissimilar***.

Expert Curated Similarity Corpus



- 24 query workflows
- Each with 10 other workflows to compare to it

  - very similar
  - similar
  - related
  - unsure
  - dissimilar

- + Extended comparison lists for specific algorithms' results for 8 query workflows

- **15 experts (7 institutes)** provided

> **2400 ratings**

  - classifying each pair of workflows
  - ranking workflow lists by similarity

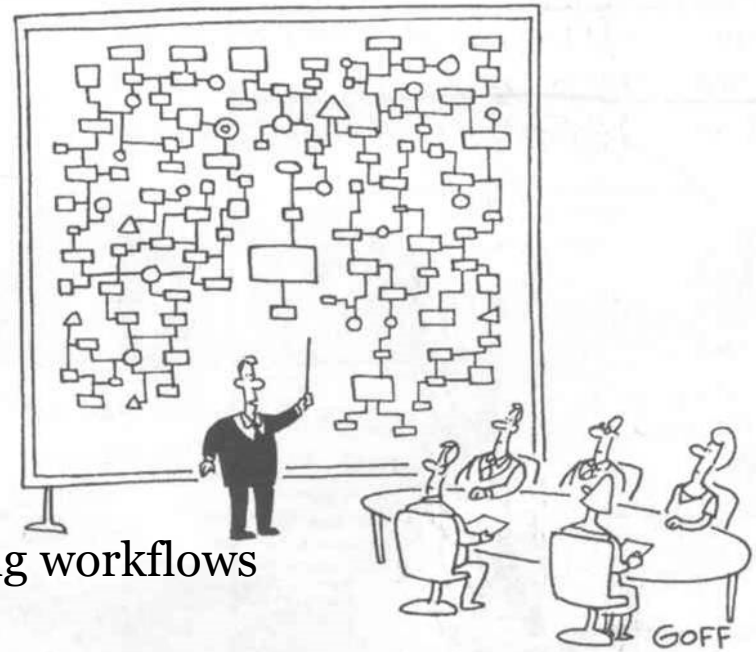
Results

- Experts agreed on the similarity of workflow pairs
- Annotation-based approaches
  - Provide best results
  - But only a few well-annotated workflows
- Structural approaches
  - Outperform annotation-based
    - Galaxy & VisTrails
  - Graph edit distance is too expensive
  - Module set provides good results
  - Room for solutions in between
    - LayerDecomposition **[eScience 2014]**
      with J. Starlinger, U. Leser, S. Davidson, S. Khanna
    - Usable in real environments (myExperiment)
      [Future Generation Computer System 2016]

How to improve reuse?

Help finding *similar workflows*

Make workflow structures less complex!

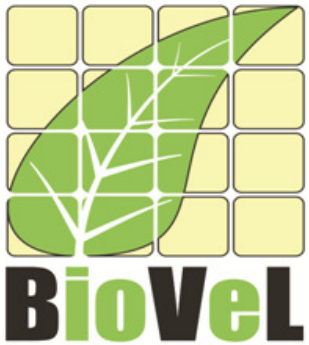


Plumbing workflows

DistillFlow

- Distilling workflow structures: Removing redundancy

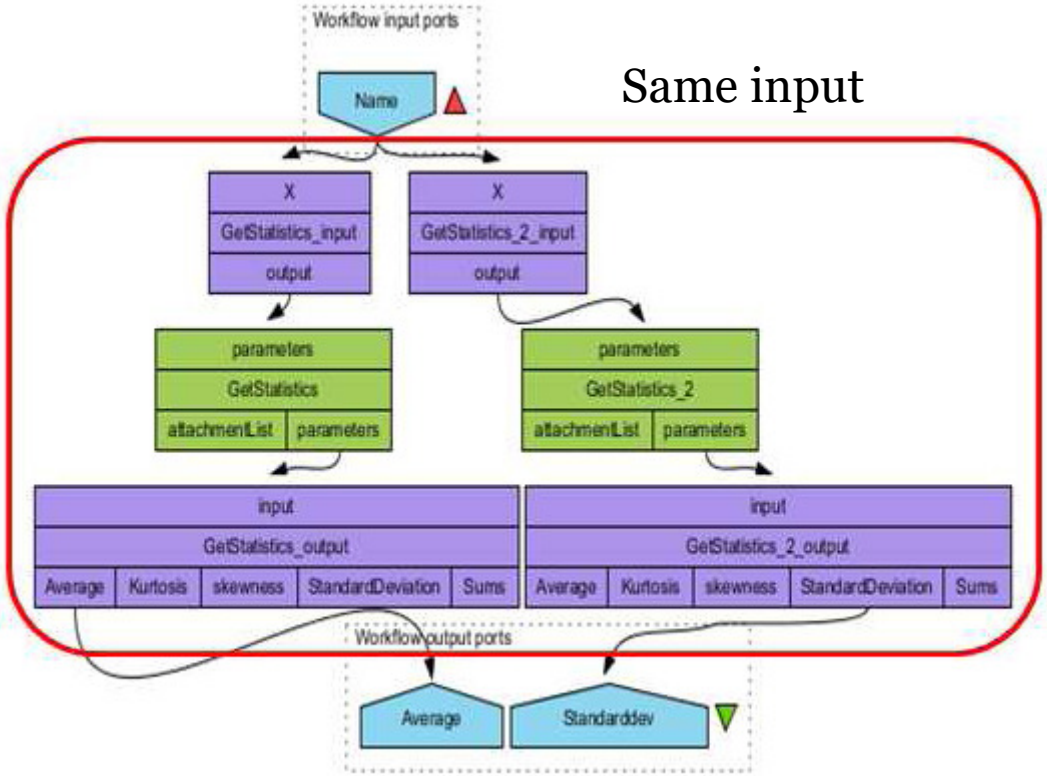
- Collaboration with Taverna & BioVel
- BioVel (FP7)



  - Virtual laboratory: Librairies of workflows for research on biodiversity
  - Consortium of 15 partners (9 countries)

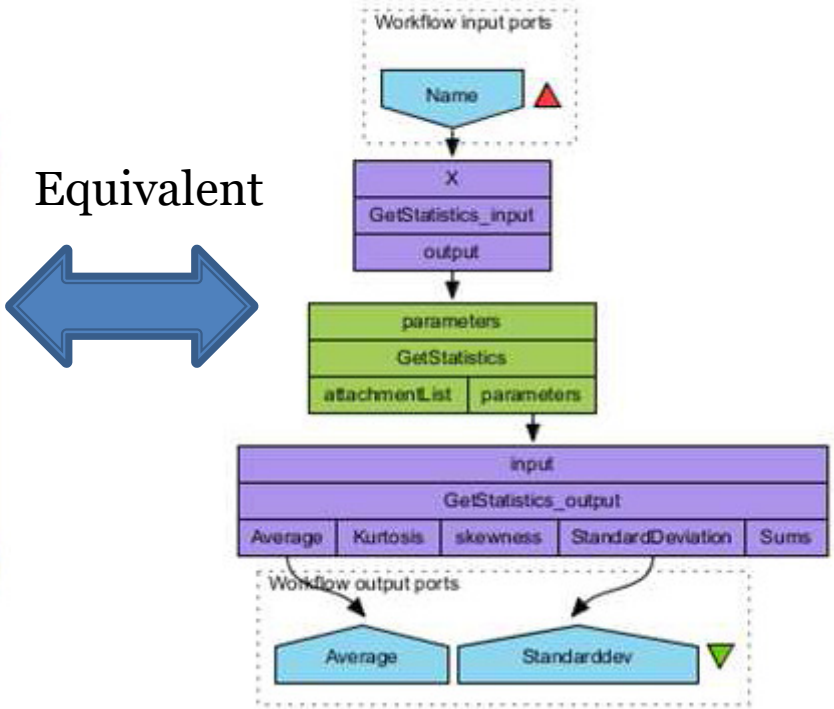
→ Improving reuse in BioVel
→ More generally: improving reuse in Taverna

Use case 1



Same input

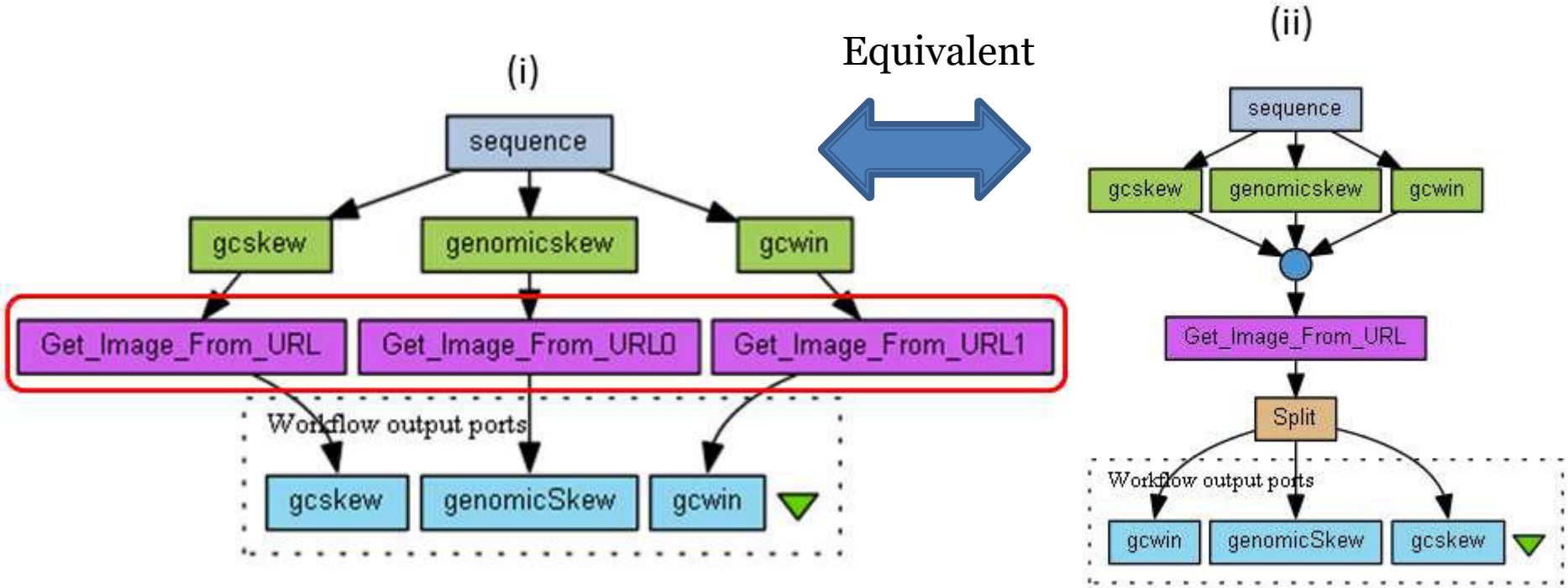
Equivalent



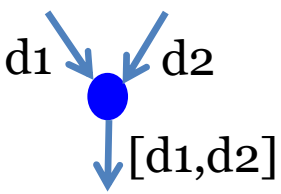
3 processors duplicated!
→ Pure redundancy

No redundancy

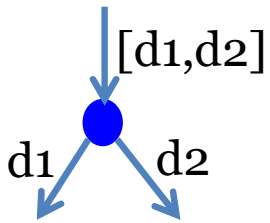
Use case 2



Workflow (ii) uses merge



split



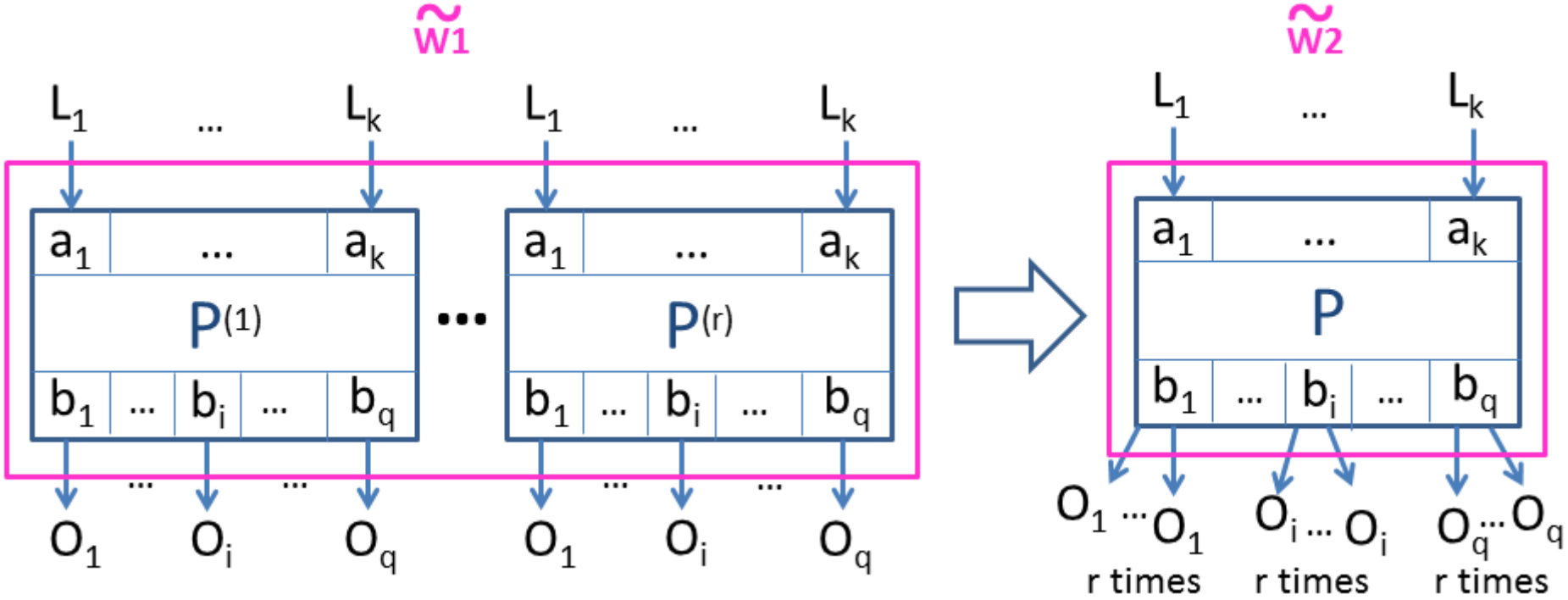
Rewriting workflows

- Exploting the implicit iteration feature of Taverna
  - List of items with merge/split instead of single items with duplication
- Assumptions before merging several copies of a processor
  - Only copies with the exact same code
  - Only copies that do not depend on each other
  - Only deterministic processors (same input → same output)

→ 2 anti-patterns and the corresponding rewriting

Anti-pattern (A)

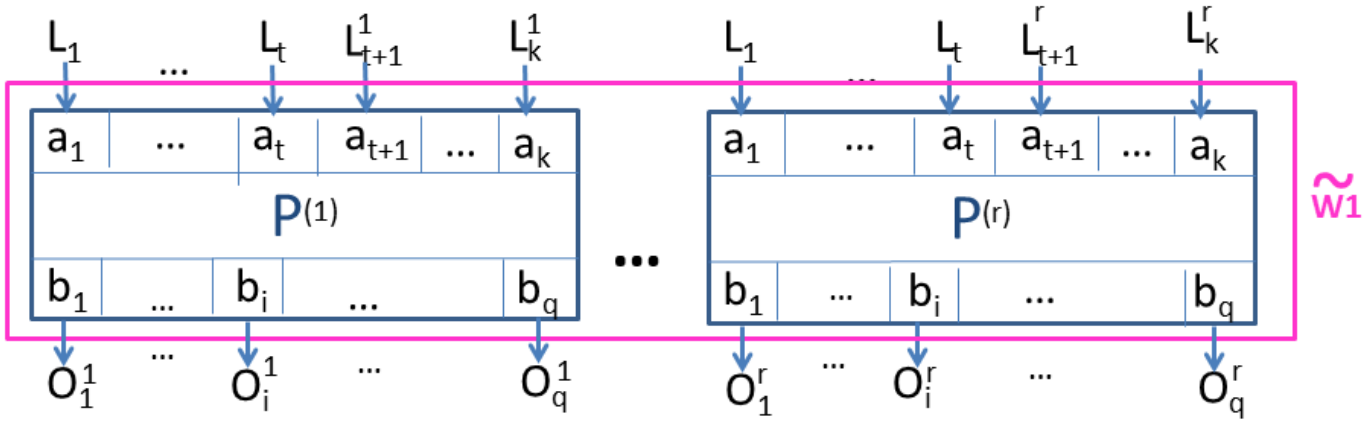
Corresponds to use case 1

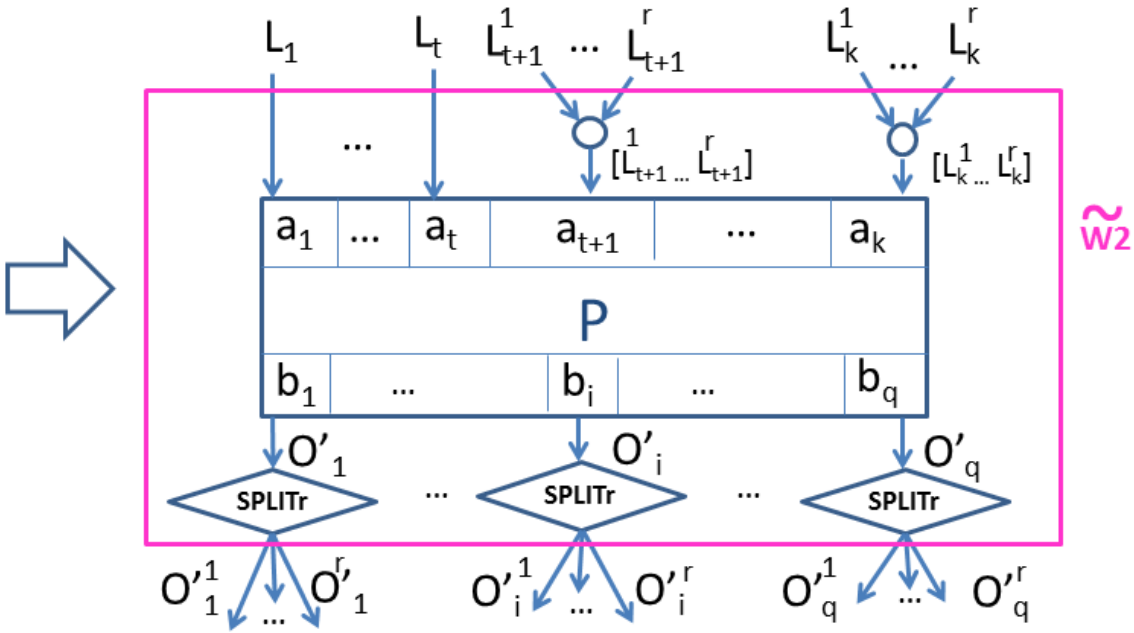


L_i can be one single value or a list of values

Anti-pattern (B)

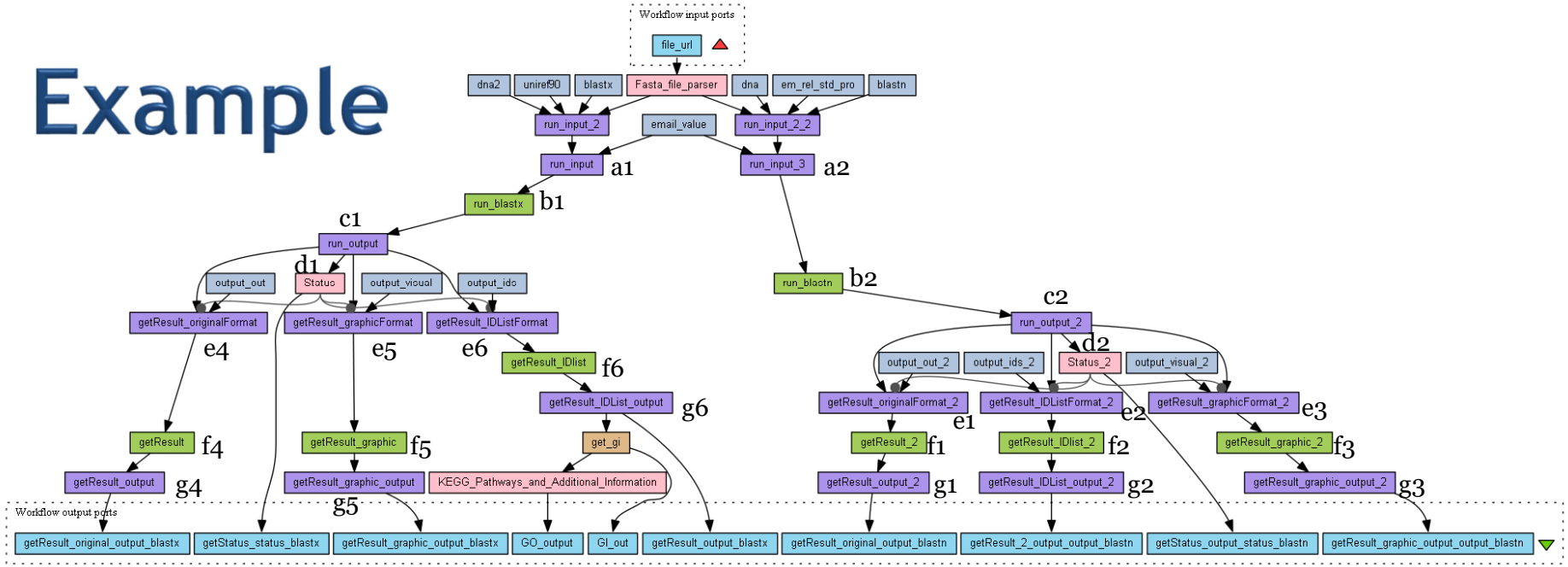
Corresponds to use case 2

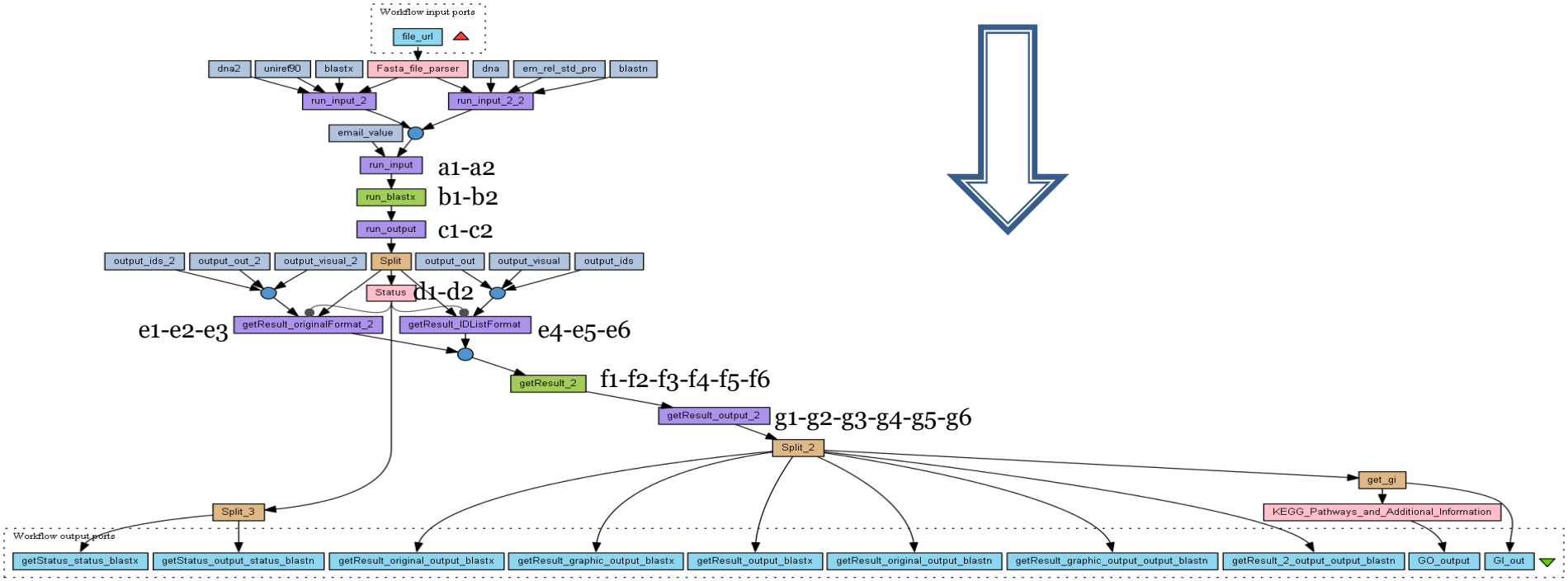




Processor P applies cross product to values on ports a_1 to a_t and dot product to values on ports a_{t+1} to a_k

Example


Workflow input ports
file_url
dna2
uniref90
blastx
Fasta_file_parser
dna
em_rel_std_pro
blastn
run_input_2
email_value
run_input_2_2
run_input
a1
run_input_3
a2
run_blastx
b1
c1
run_output
d1
output_out
Status
output_visual
output_ids
run_blastn
b2
c2
run_output_2
d2
output_out_2
output_ids_2
Status_2
output_visual_2
getResult_originalFormat
getResult_graphicFormat
getResult_IDListFormat
e4
e5
e6
getResult_IDlist
f6
getResult_IDList_output
g6
getResult_originalFormat_2
getResult_IDListFormat_2
getResult_graphicFormat_2
e1
e2
e3
getResult
f4
getResult_graphic
f5
get_gi
getResult_2
f1
getResult_IDlist_2
f2
getResult_graphic_2
f3
getResult_output
g4
getResult_graphic_output
KEGG_Pathways_and_Additional_Information
g5
getResult_output_2
g1
getResult_IDList_output_2
g2
getResult_graphic_output_2
g3
Workflow output ports
getResult_original_output_blastx
getStatus_status_blastx
getResult_graphic_output_blastx
GO_output
GI_out
getResult_output_blastx
getResult_original_output_blastn
getResult_2_output_output_blastn
getStatus_output_status_blastn
getResult_graphic_output_output_blastn


Workflow input ports
file_url
dna2
uniref90
blastx
Fasta_file_parser
dna
em_rel_std_pro
blastn
run_input_2
run_input_2_2
email_value
run_input
a1-a2
run_blastx
b1-b2
run_output
c1-c2
output_ids_2
output_out_2
output_visual_2
Split
output_out
output_visual
output_ids
Status
d1-d2
e1-e2-e3
getResult_originalFormat_2
getResult_IDListFormat
e4-e5-e6
getResult_2
f1-f2-f3-f4-f5-f6
getResult_output_2
g1-g2-g3-g4-g5-g6
Split_2
get_gi
KEGG_Pathways_and_Additional_Information
Split_3
Workflow output ports
getStatus_status_blastx
getStatus_output_status_blastn
getResult_original_output_blastx
getResult_graphic_output_blastx
getResult_output_blastx
getResult_original_output_blastn
getResult_graphic_output_output_blastn
getResult_2_output_output_blastn
GO_output
GI_out

This Tutorial

- Part II – Data Integration workflows
  - What are scientific workflow systems
  - Designing a workflow from scratch
  - Repositories of workflows and web services (reuse)
  - workflows and reproducibility
  - Latest results on workflows

  →Or How CS research may have direct impact on LS

  Improving reuse

  Managing Provenance

  Comparing workflows executions

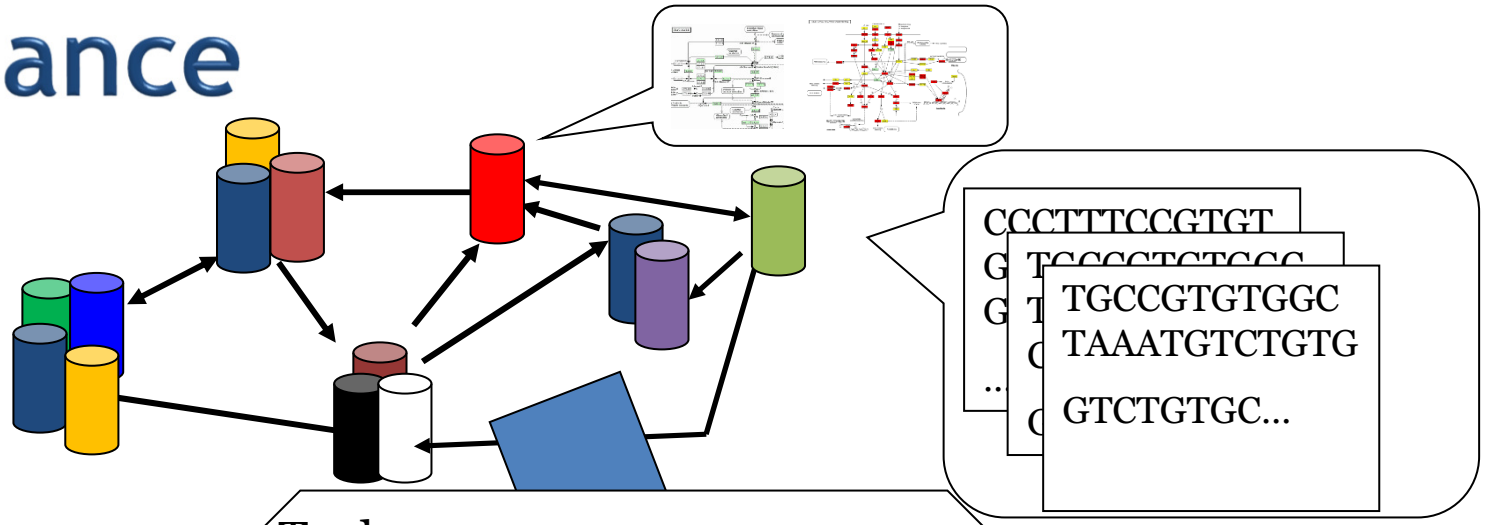
Provenance in scientific workflows

- Provenance is highly important for users to interpret any scientific result
- Workflow systems are now equipped of *Provenance Modules* capturing the exact set of data used and consumed by the execution of each workflow step
- Standards to represent provenance information are now defined (W3C)
- One of the major challenge lies in dealing with the huge amounts of information
  - Example of solution with ZOOM*userviews which use the composition to hide (part of) the data

Provenance

Public sources

- Distributed
- Heterogeneous
- Network



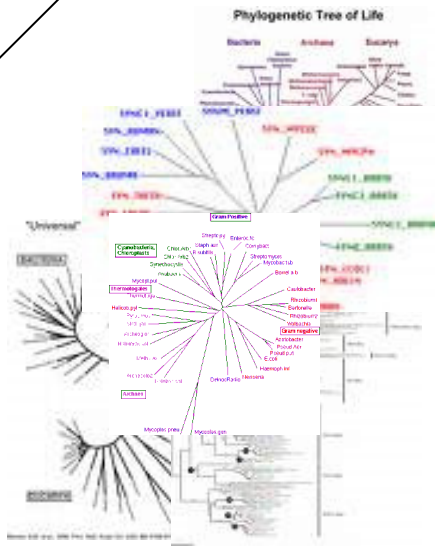
Tools
Scripts
Python
JAVA, Perl
Web services
...

Tools

- Distributed
- Heterogeneous
- Chained

How these data have been generated? With which input data? Which tools? Which parameters?

What is the difference between these two experiments?

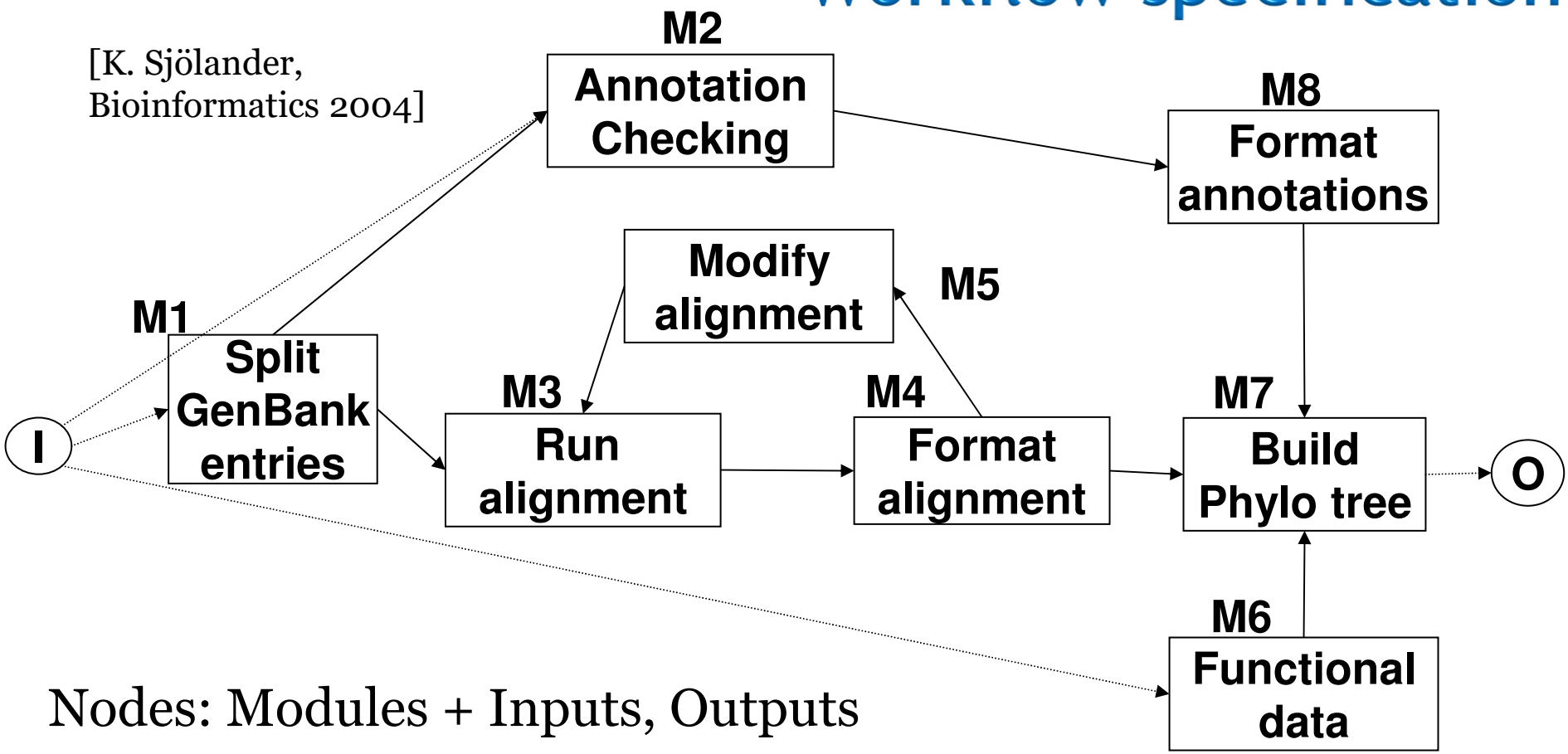


Workspace

Workflows are graphs

Workflow specification

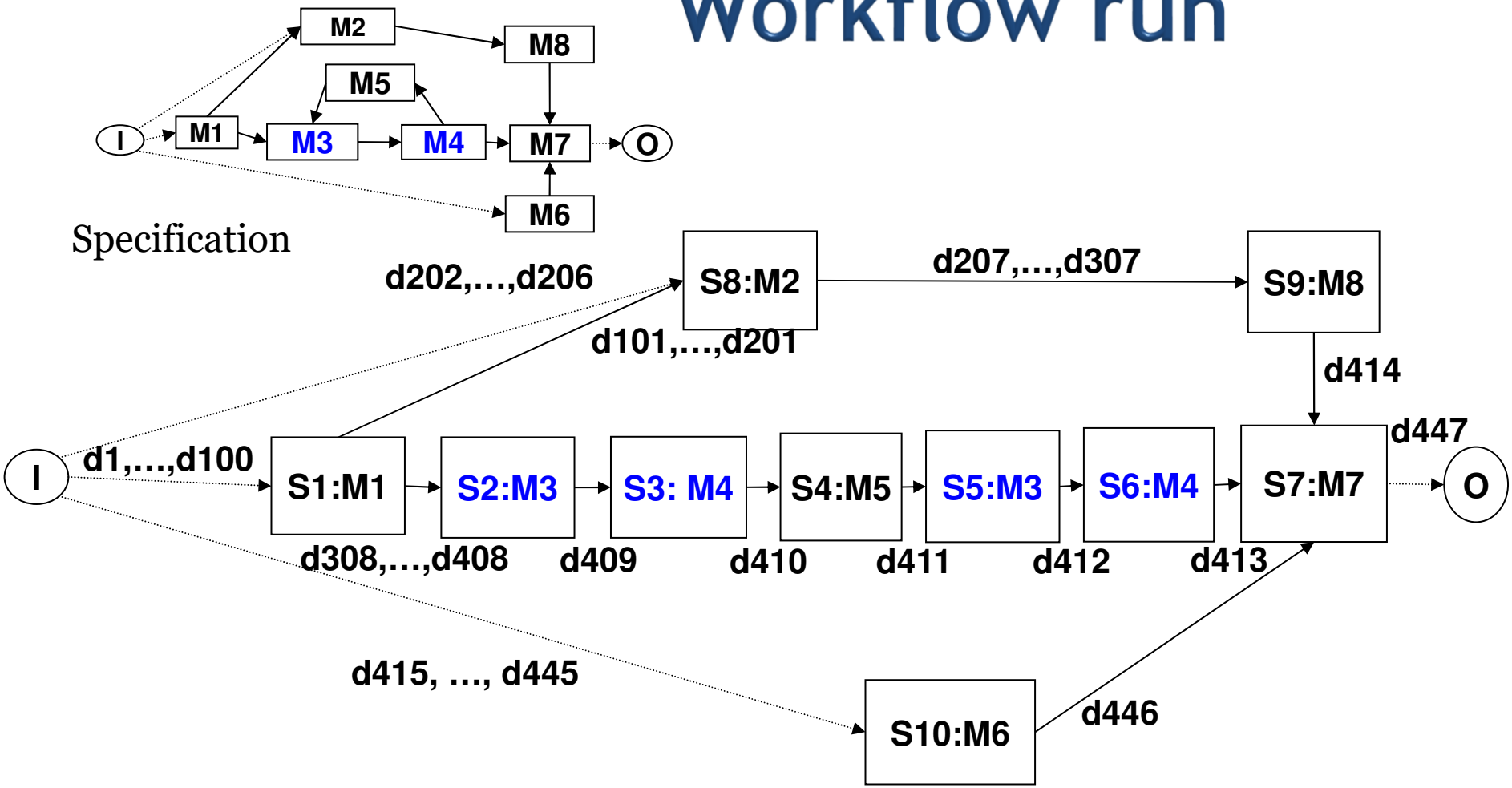
[K. Sjölander, Bioinformatics 2004]



Nodes: Modules + Inputs, Outputs

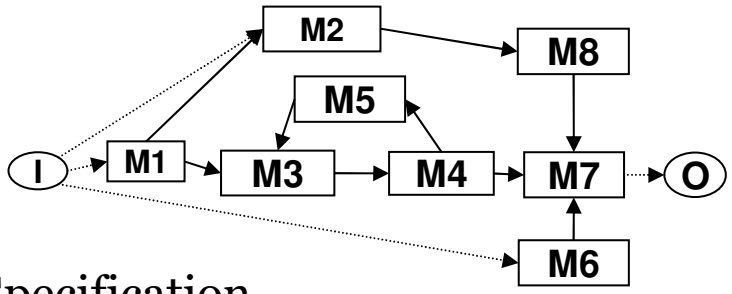
Edges: Possible dataflow

Workflow run



Nodes: Steps (executions of modules)

Edges: Actual dataflow (labelled with data object ids)

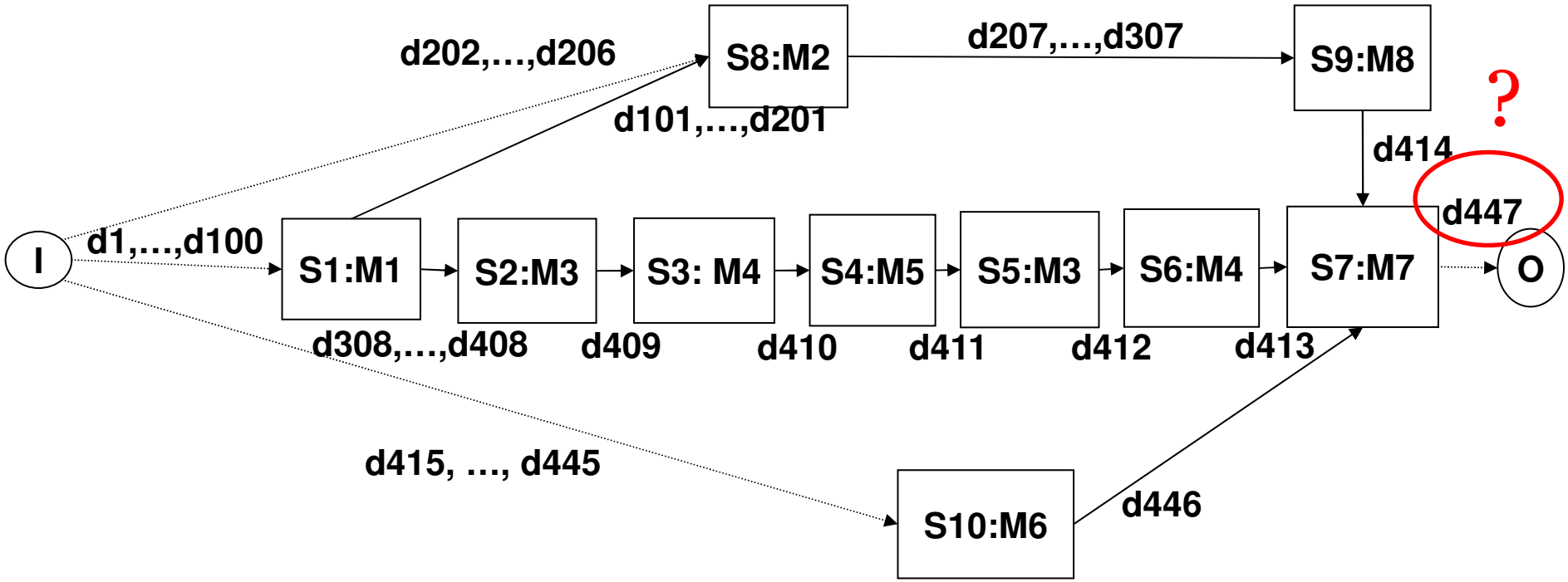


Specification

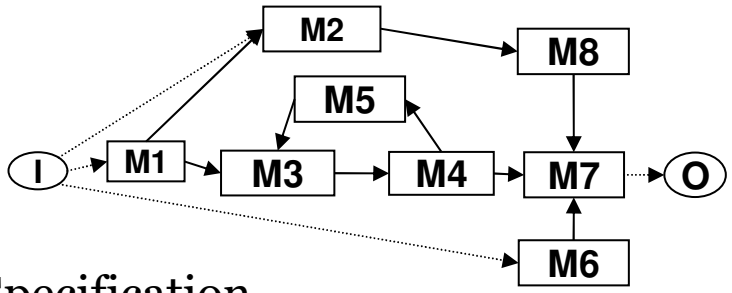
Workflow run:

Provenance of d447?

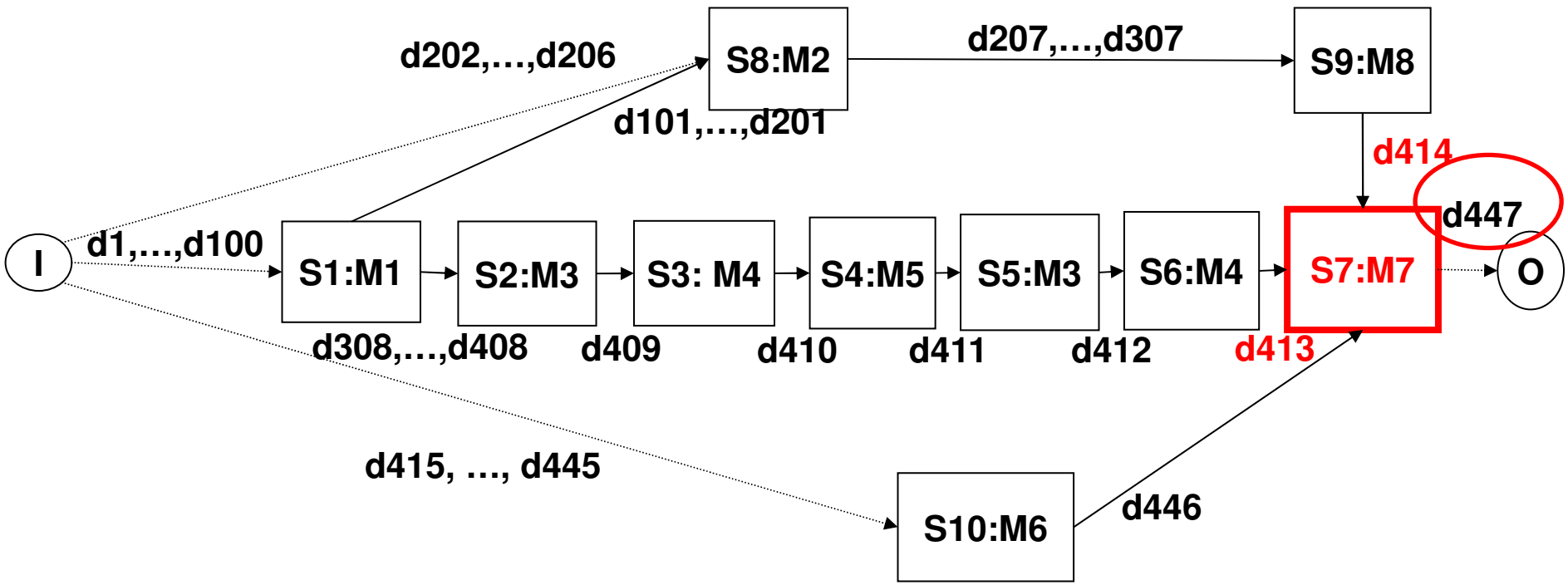
(tree generated)



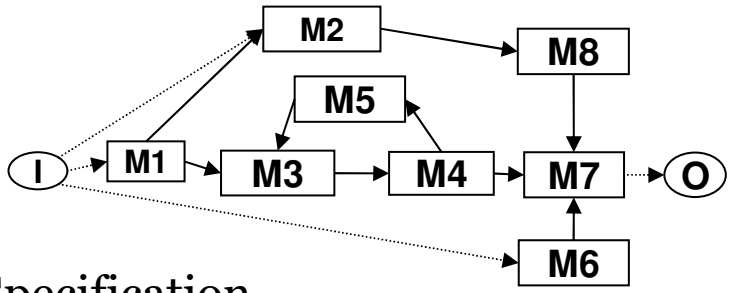
Workflow run:
Provenance of d447 (immediate)



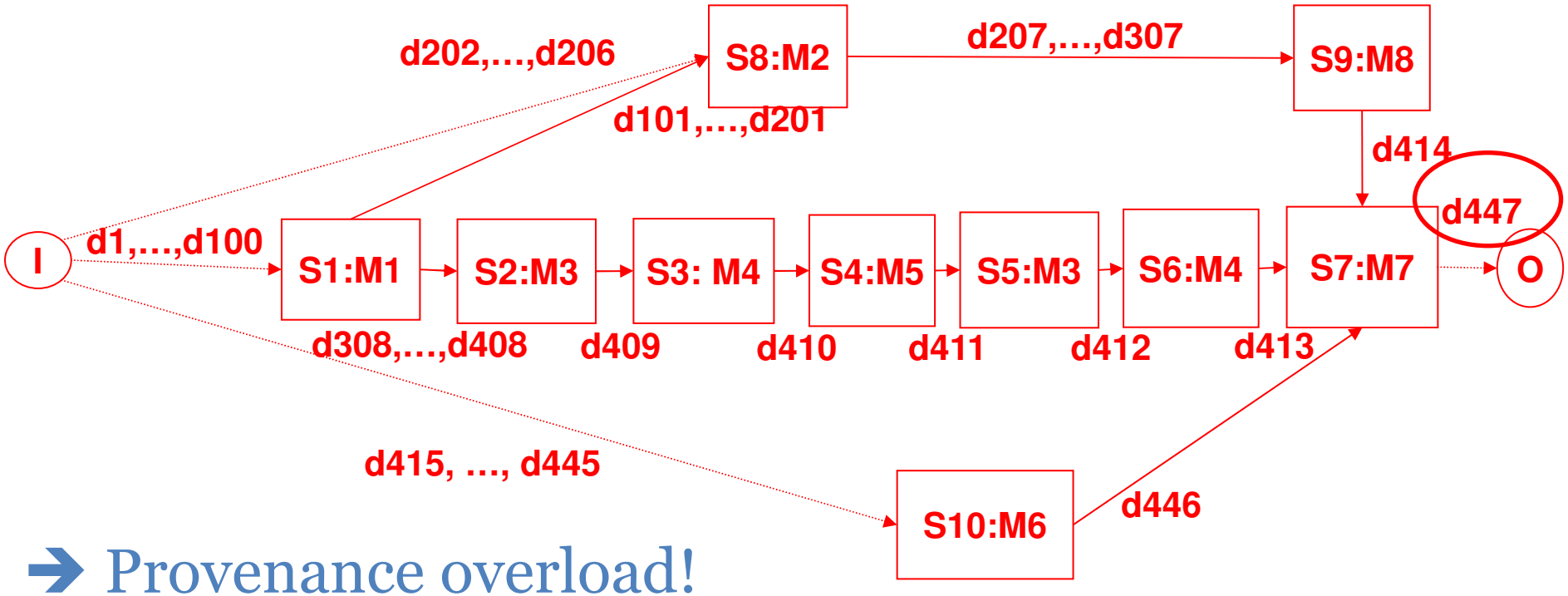
Specification



Workflow run:
Provenance of d447 (deep)



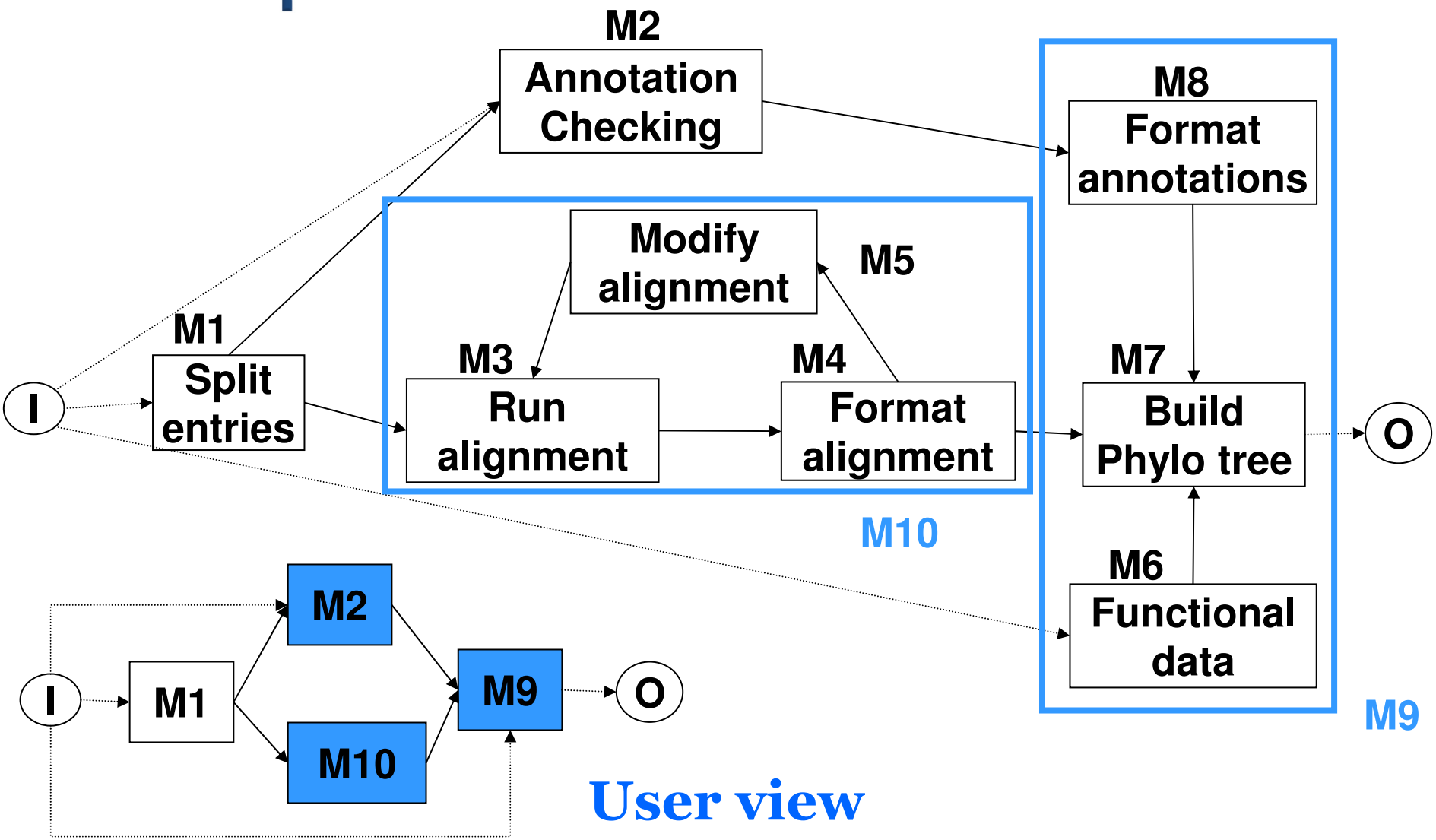
Specification



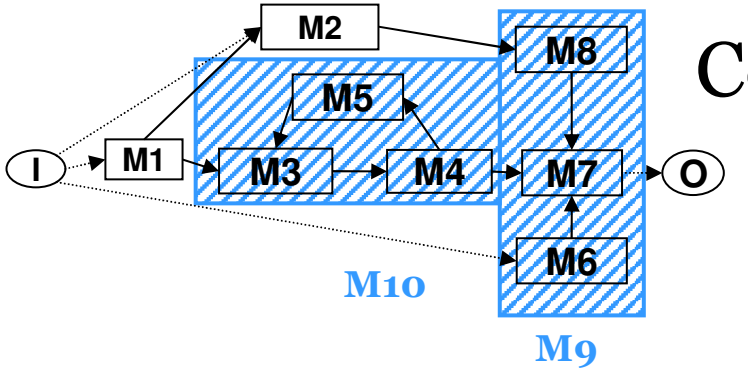
➔ Provenance overload!

➔ Need to focus on relevant information

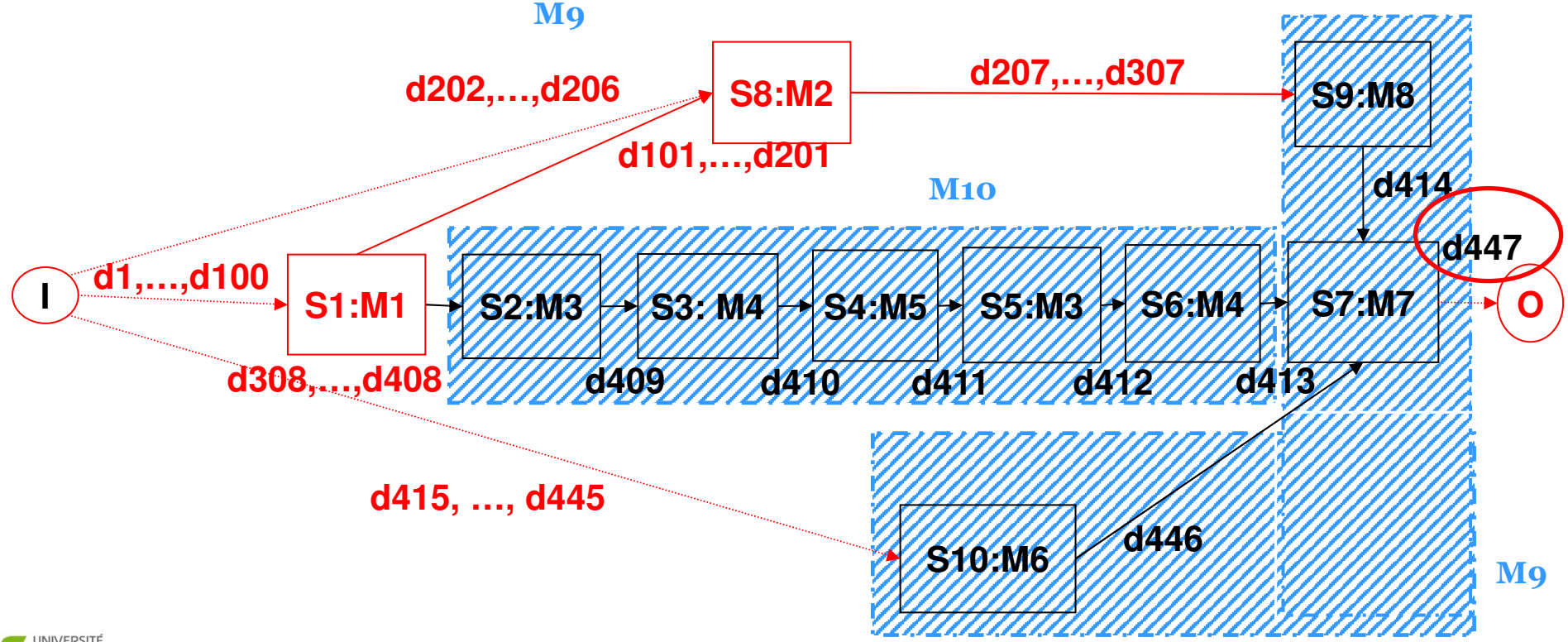
Composite modules



Composite modules



Composition simplifies provenance!



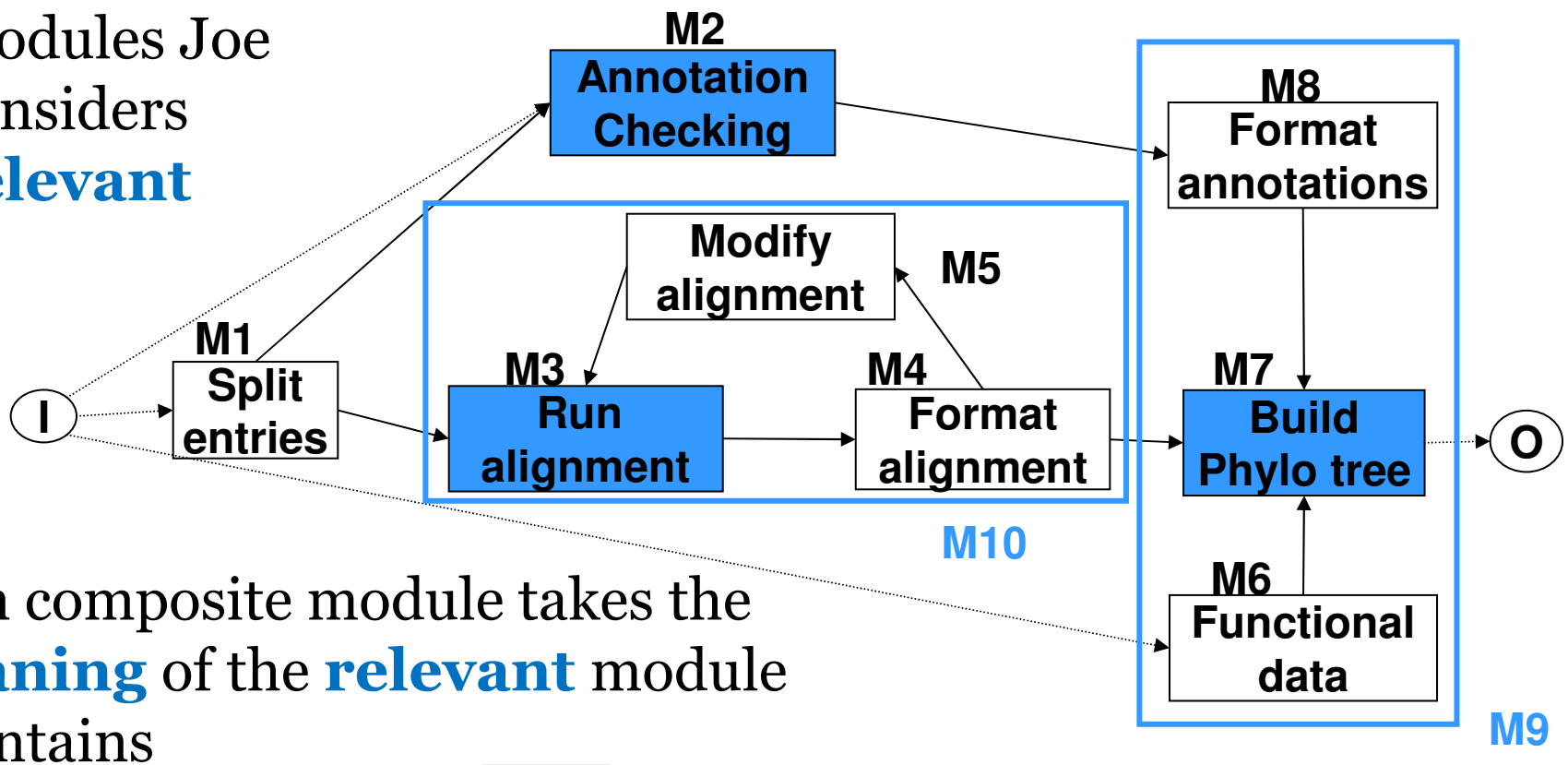
Designing composite modules

- Composite modules are typically defined **by the workflow designer** to
  - Enable **reuse** between workflows
  - **Simplify** the view of the workflow according to what modules the **designer thinks are relevant** in the workflow
- However, users may have **different interests**, i.e. have different relevant modules

➔ **Several user views** of a given workflow should thus be considered, constructed according to each user's interest

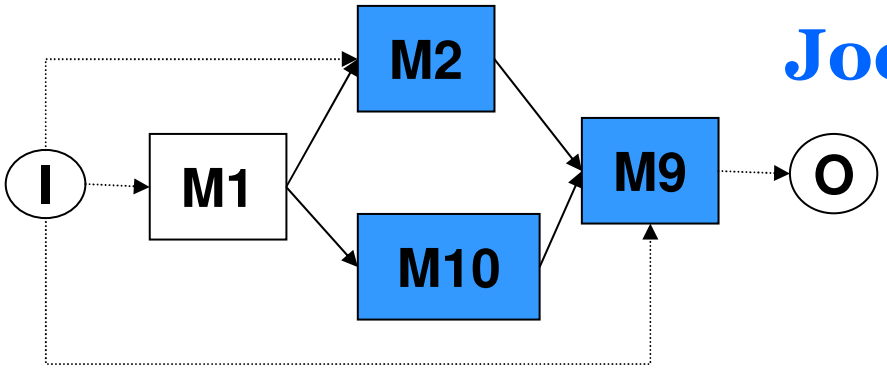
Relevant user view

Modules Joe considers **relevant**



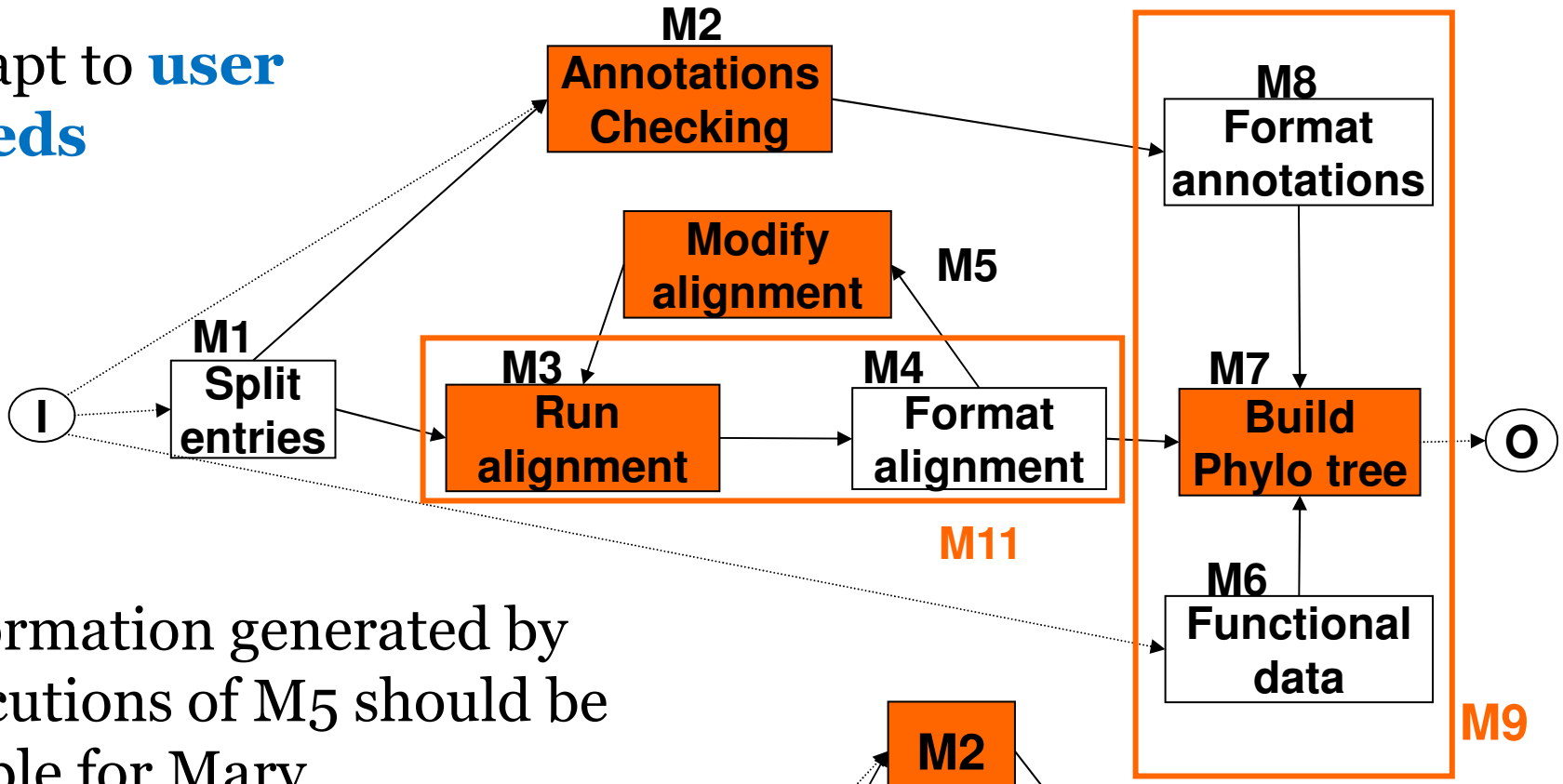
Each composite module takes the **meaning** of the **relevant** module it contains

Joe's view



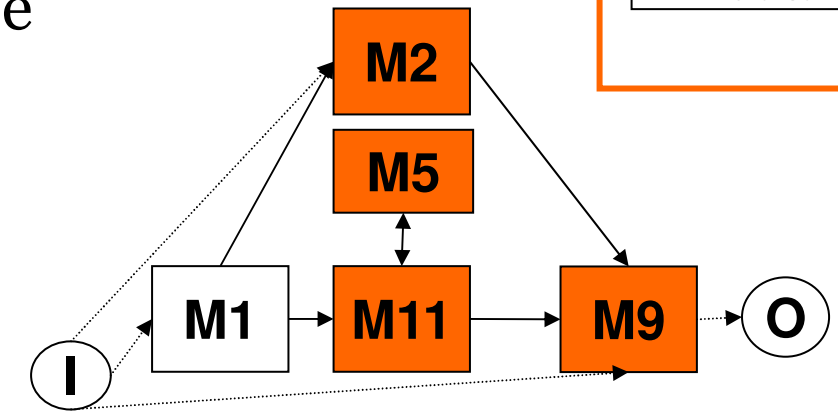
User views may differ

Adapt to **user needs**



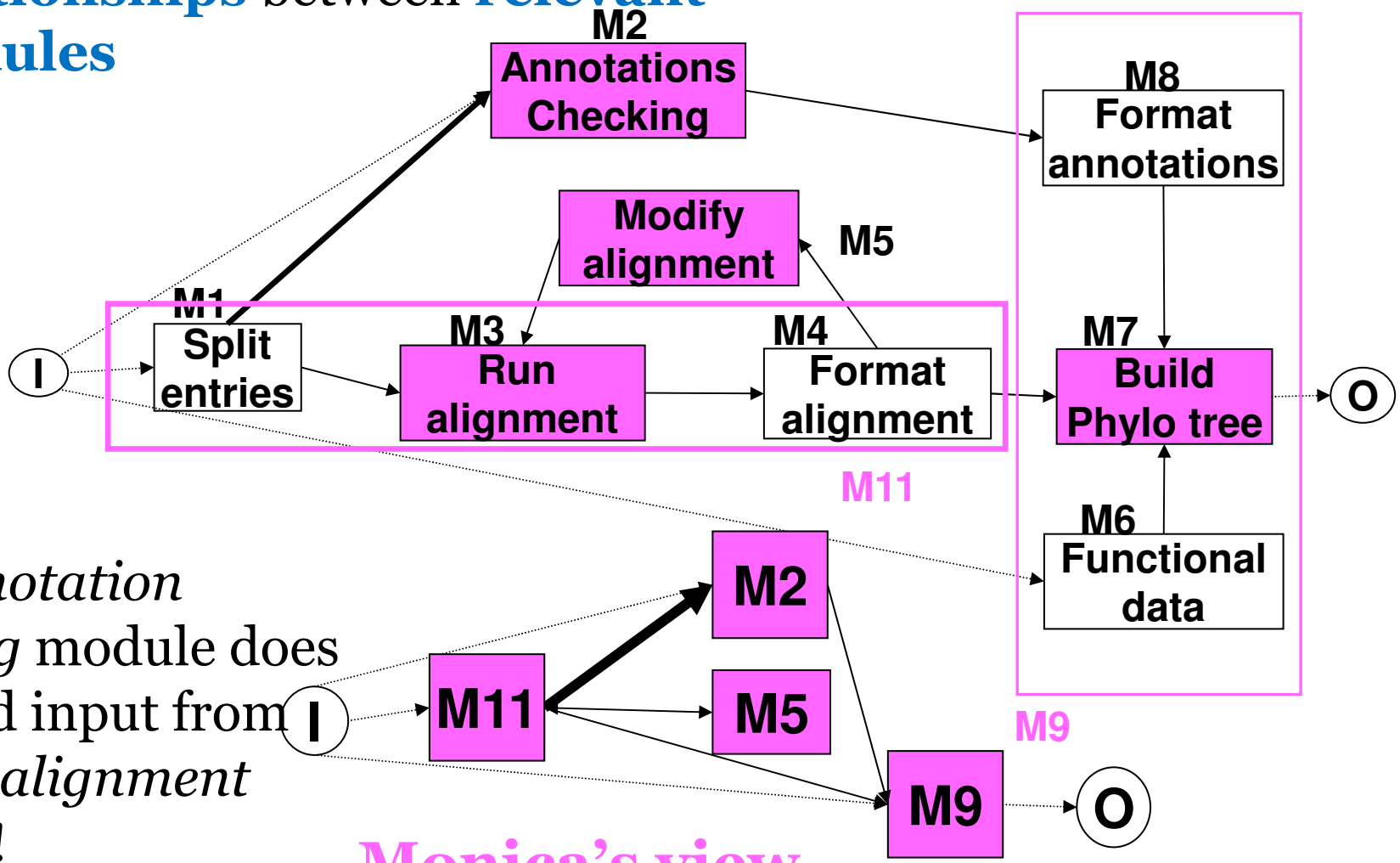
Information generated by executions of M5 should be visible for Mary

Mary's view



Grouping may be error-prone!

Grouping should **preserve the relationships** between **relevant modules**



The *annotation checking* module does not need input from the *run alignment* module!

Monica's view

ZOOM*UserViews

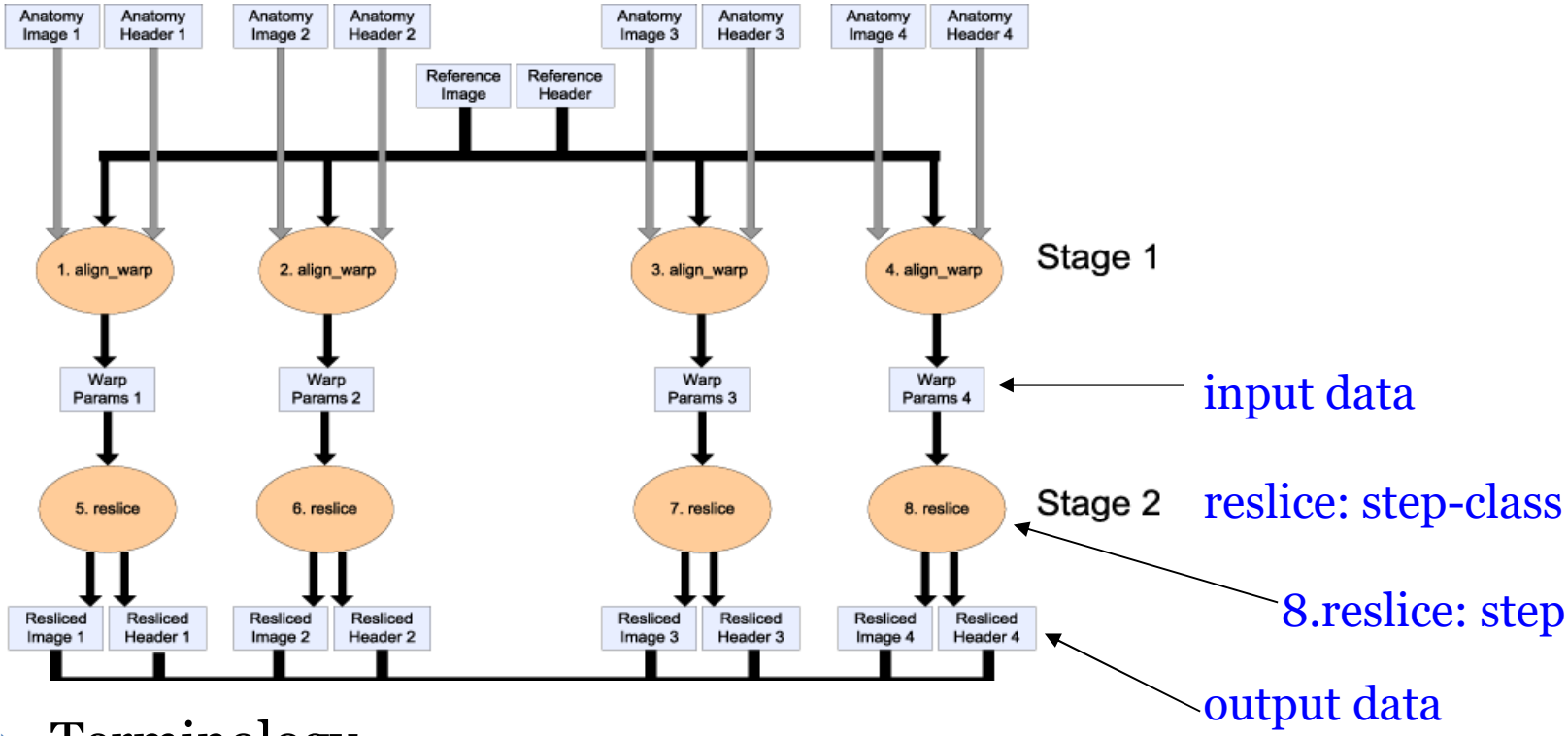
- **Goals**
  - Help user **construct relevant user views**
    - Preserving the **relationships** between **relevant modules**
  - Exploit ***user views*** to **reduce the provenance information** returned as answer to a query

- **Contributions**
  - **Model** for provenance and user views in scientific workflows
  - **Algorithm (polynomial)** for generating **relevant** user views according to the user's interests (minimal)
  - **Provenance Reasoning system:** Querying provenance **through user views**

Provenance challenge

- First Provenance Challenge (**twiki.ipaw.info**)
  - By S. Miles, M. Wilde, I. Foster and **L. Moreau,** at Washington DC, Sept. 2006
- **Aims:** Understanding the **capabilities** of provenance-related systems (17)
- The **challenge process**
  - **Workflow example (spec + run)** provided
  - **List of provenance queries** to be answered

Workflow Representation



- Terminology
  - Nodes are **step-classes** (static)
  - Edges capture the **flow of data** between step-classes
  - An **execution** of a workflow generates a partial order of steps (dynamic)
    - Instances of step classes
  - Each step has **input** and **output** data

This Tutorial

- Part II – Data Integration workflows
  - What are scientific workflow systems
  - Designing a workflow from scratch
  - Repositories of workflows and web services (reuse)
  - workflows and reproducibility
  - Latest results on workflows

  →Or How CS research may have direct impact on LS

  Improving reuse

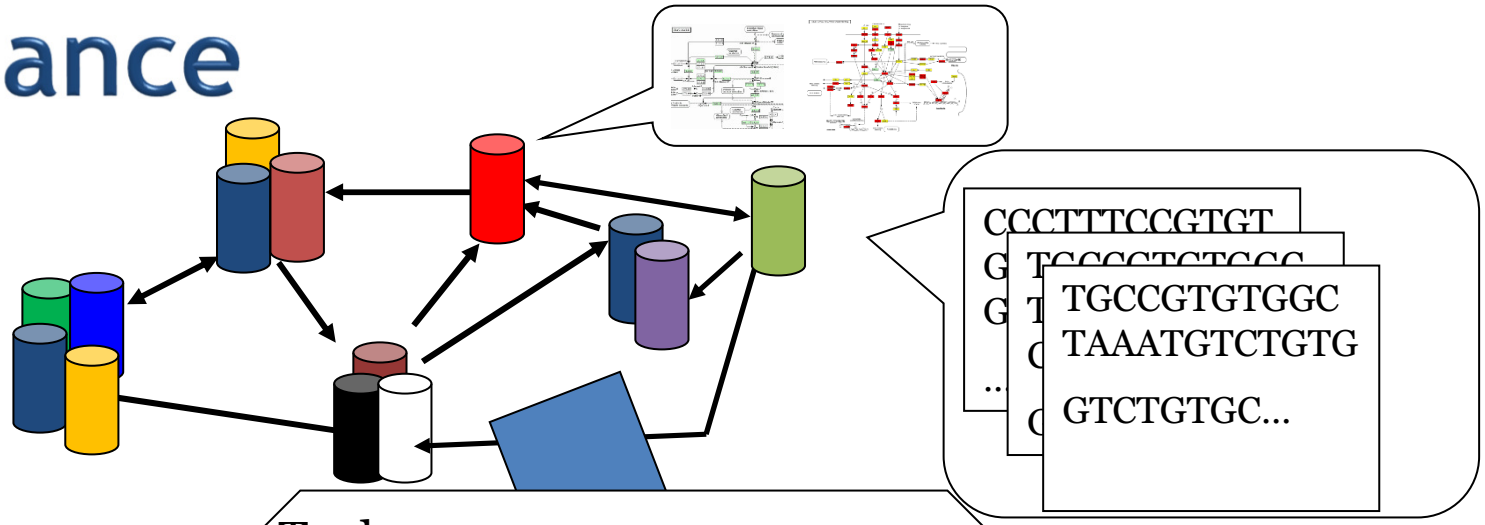
  Managing Provenance

  Comparing workflows executions

Provenance

Public sources

- Distributed
- Heterogeneous
- Network



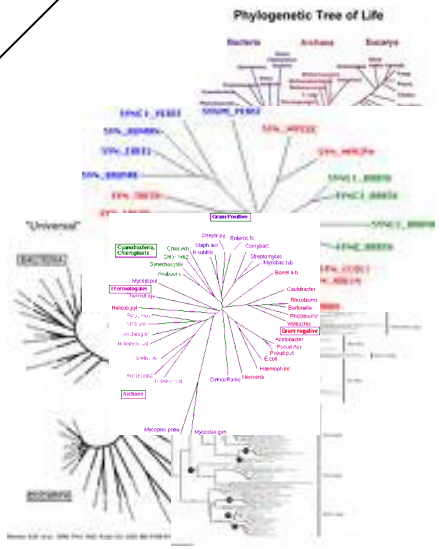
Tools
Scripts
Python
JAVA, Perl
Web services
...

Tools

- Distributed
- Heterogeneous
- Chained

How these data have been generated? With which input data? Which tools? Which parameters?

What is the difference between these two experiments?

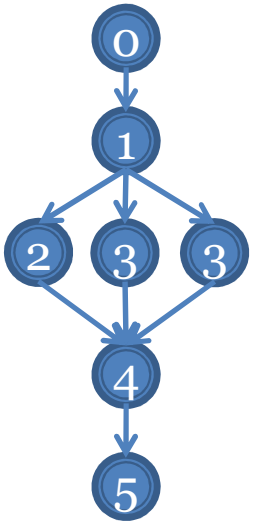


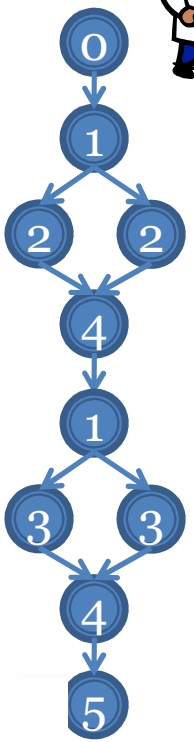
Workspace

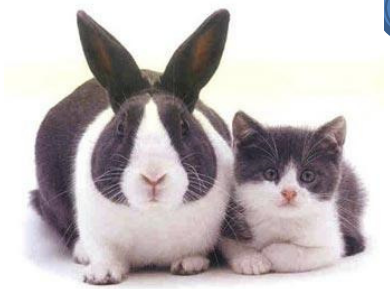
Workflow runs Difference Problem

What's the difference between these two runs of the same workflow?

Our problem is more than a "spot the difference" puzzle!



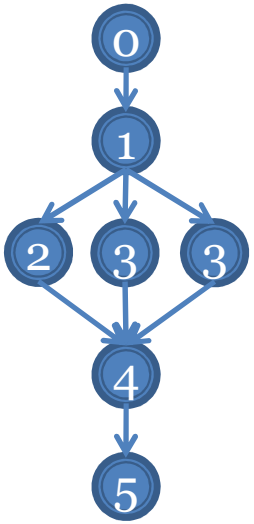


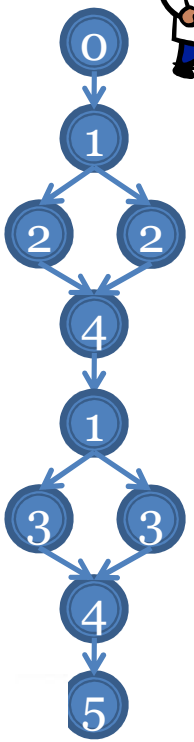


Workflow runs Difference Problem

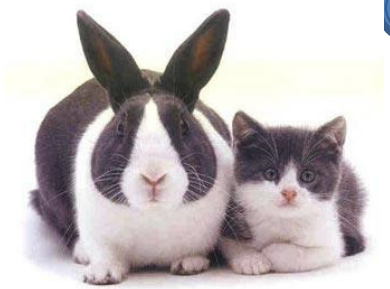
What's the difference between these two runs of the same workflow?

Our problem is more than a "spot the difference" puzzle!





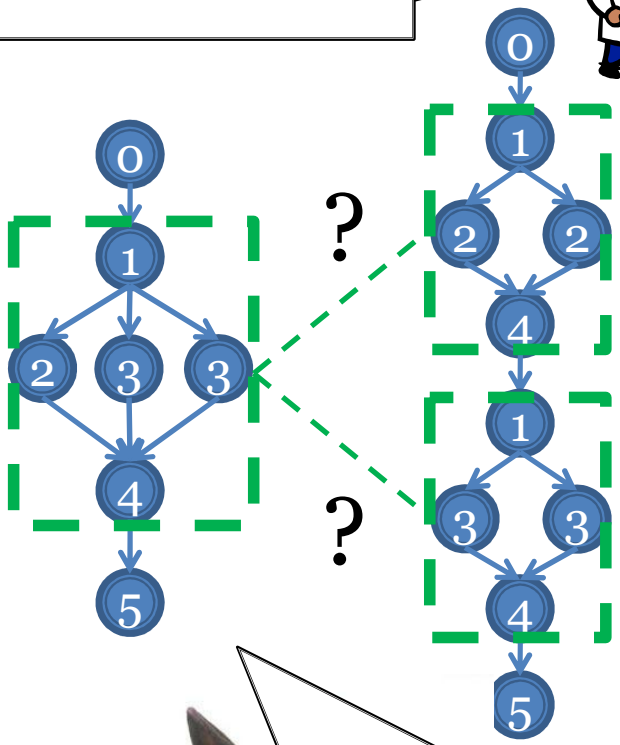
Mapping different objects in two figures is trivial.



Workflow runs Difference Problem

What's the difference between these two runs of the same workflow?

Our problem is more than a "spot the difference" puzzle!



Mapping different fork or loop copies in two runs is nontrivial!

Mapping different objects in two figures is trivial.

Workflow runs Difference Problem

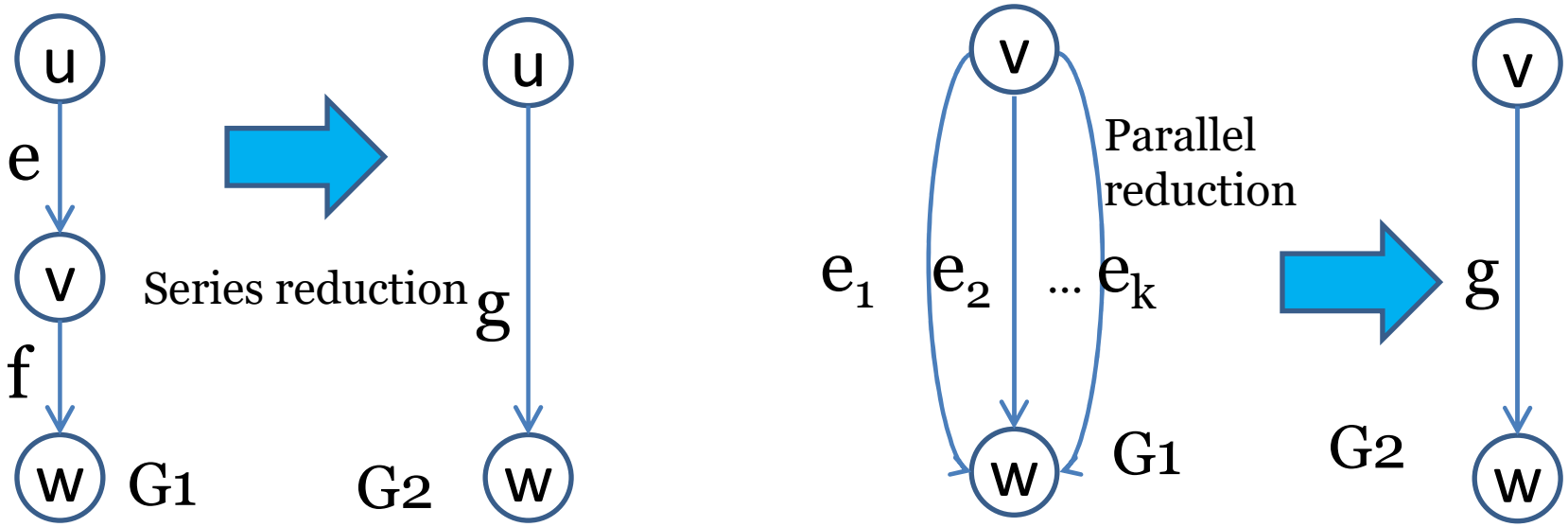
The problem of differencing runs is NP-hard on DAGs while polynomial time algorithms can be designed for Series-Parallel (SP) structures

→ Some approaches have considered such restrictions on workflow graph structures

Definition of SP-graphs

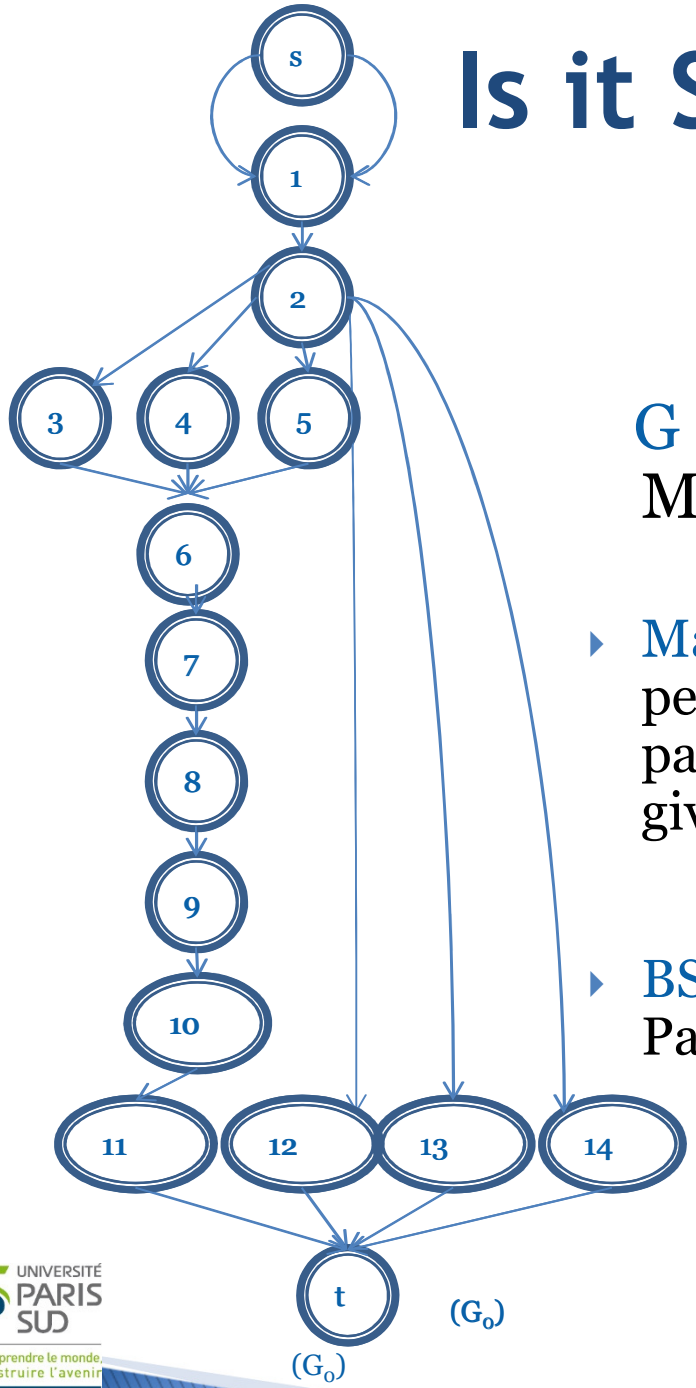
G is SP iff MaxRed(G) = BSP

- MaxRed(G): iteratively performs series and parallel reductions on a given graph G



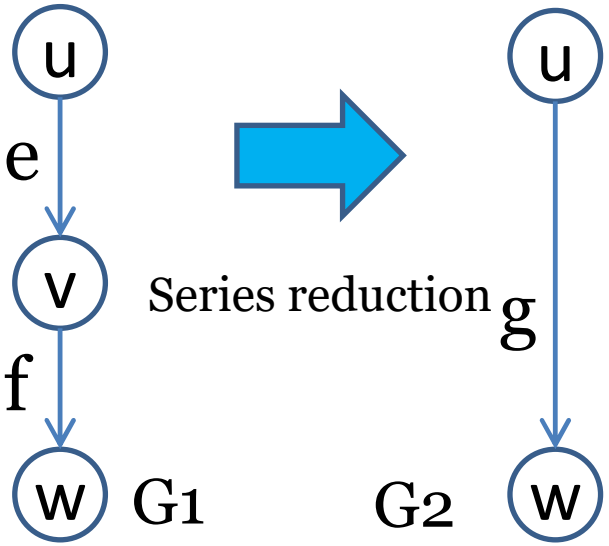
- BSP: Basic Series-Parallel

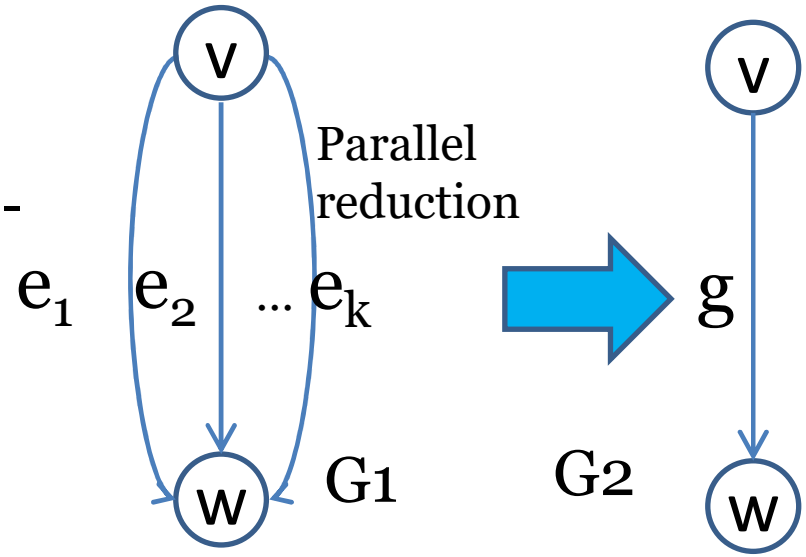
Is it Series-Parallel?

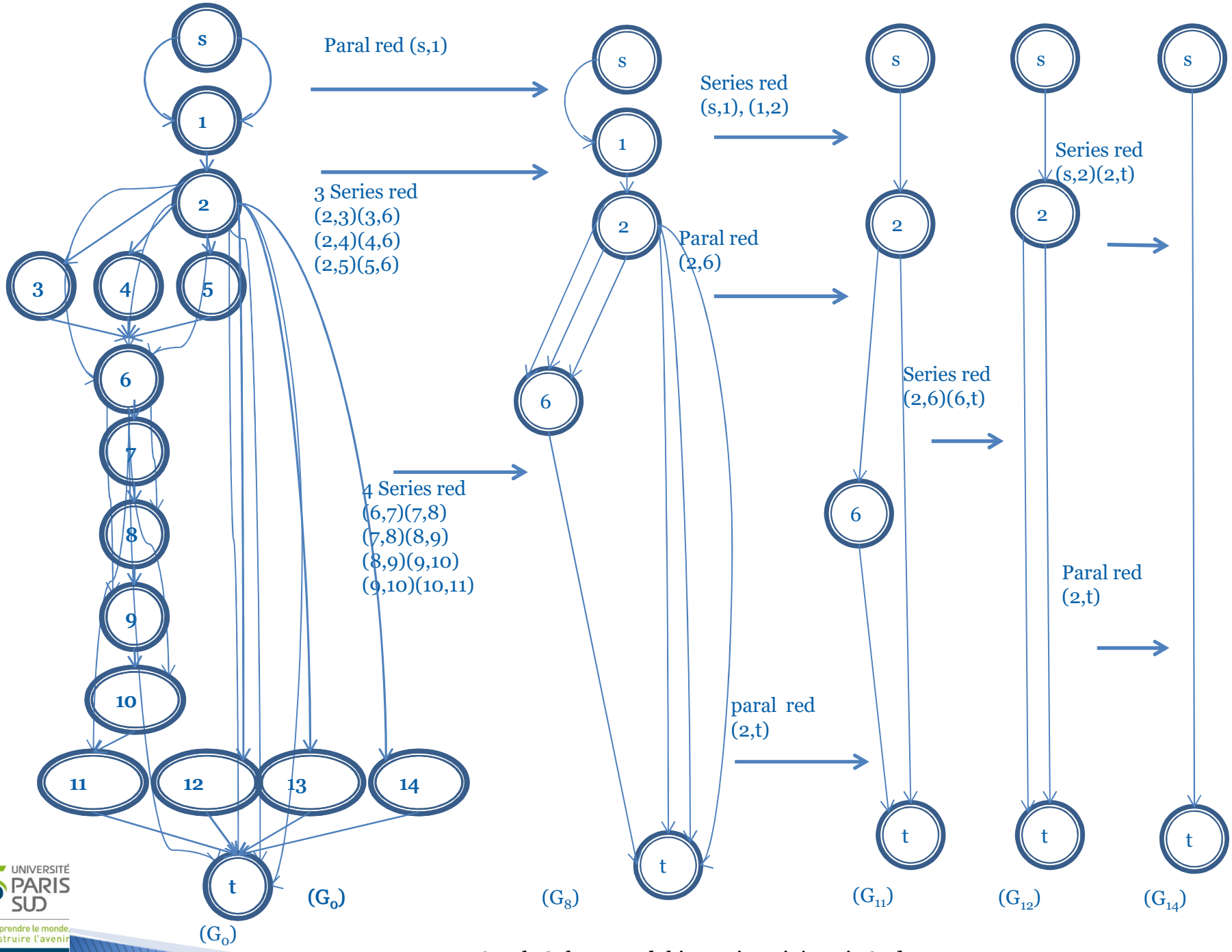


G is SP iff
MaxRed(G) = BSP

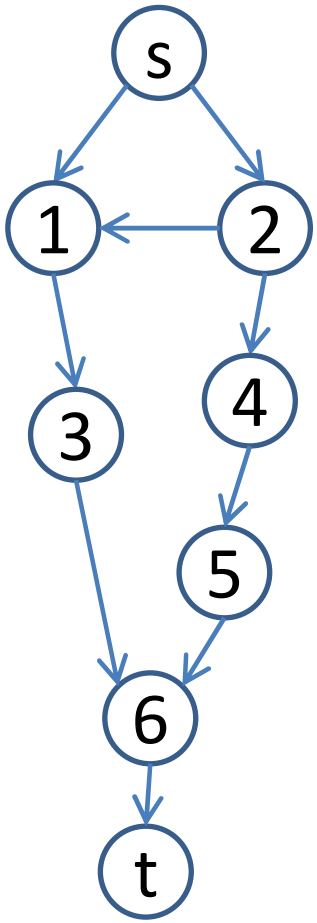
- MaxRed(G): iteratively performs series and parallel reductions on a given graph G
- BSP: Basic Series-Parallel





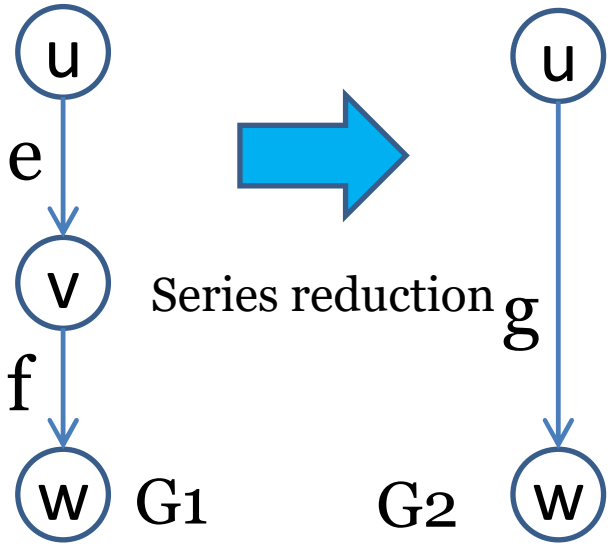

Paral red (s,1)
3 Series red
(2,3)(3,6)
(2,4)(4,6)
(2,5)(5,6)
4 Series red
(6,7)(7,8)
(7,8)(8,9)
(8,9)(9,10)
(9,10)(10,11)
Series red
(s,1), (1,2)
Paral red
(2,6)
paral red
(2,t)
Series red
(2,6)(6,t)
Series red
(s,2)(2,t)
Paral red
(2,t)
(G0)
(G8)
(G11)
(G12)
(G14)

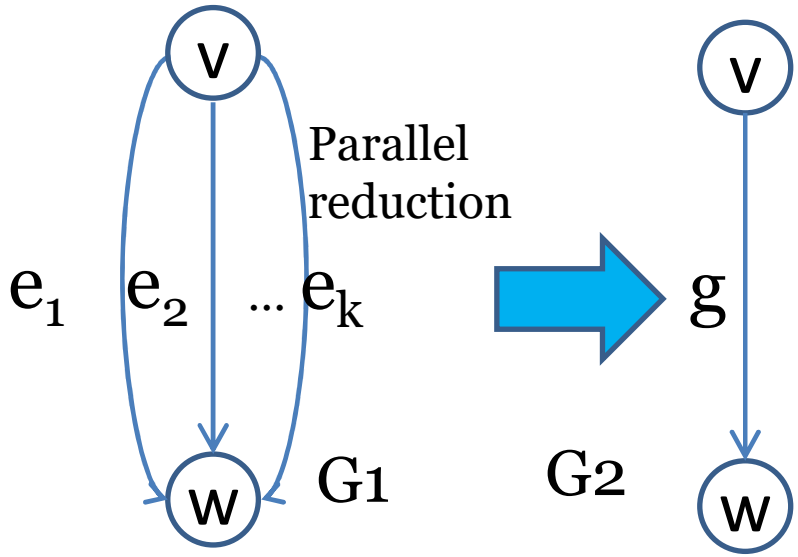
Is it Series-Parallel?



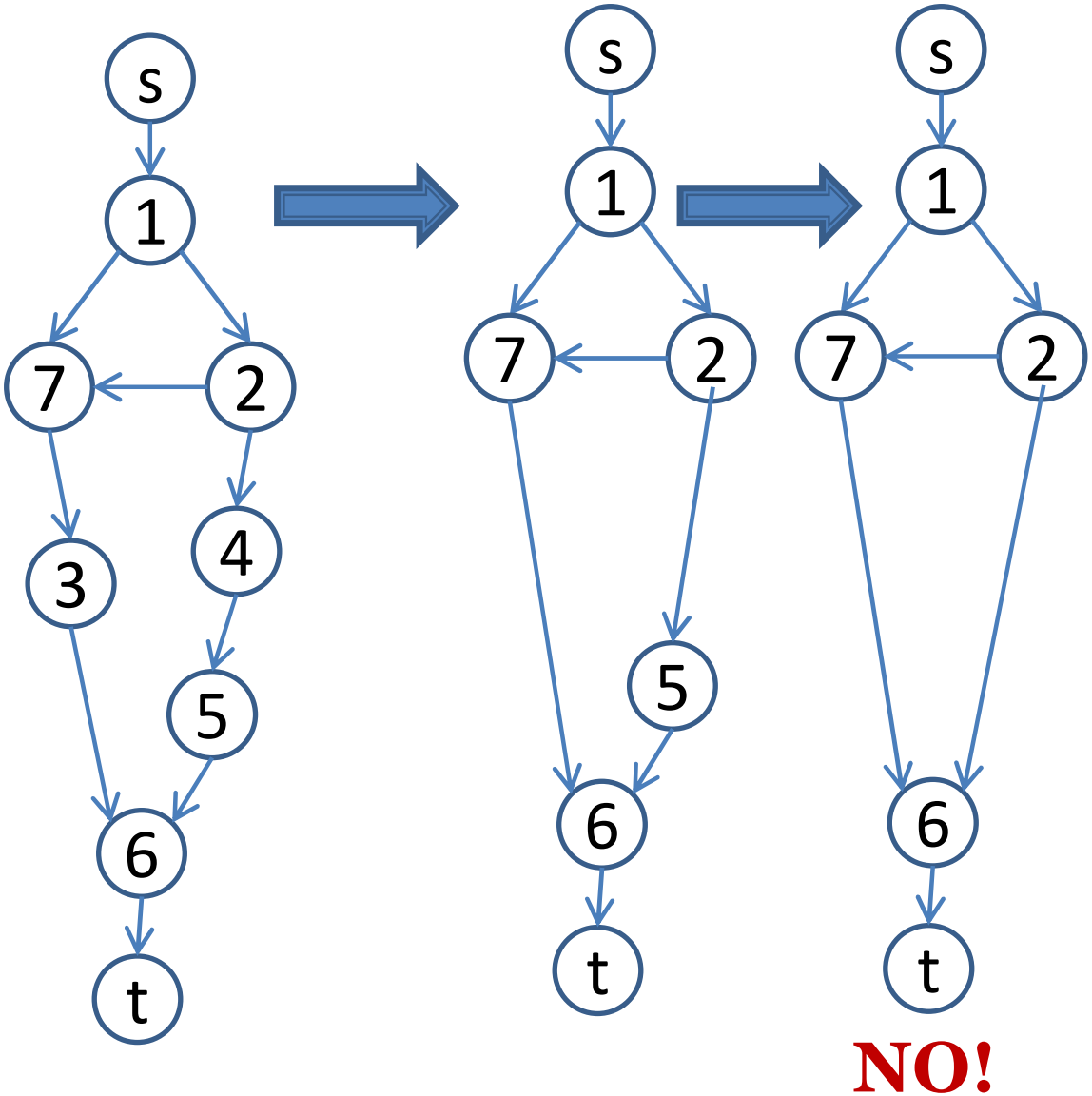
G is SP iff
MaxRed(G) = BSP

- MaxRed(G): iteratively performs series and parallel reductions on a given graph G
- BSP: Basic Series-Parallel





Is it Series-Parallel?

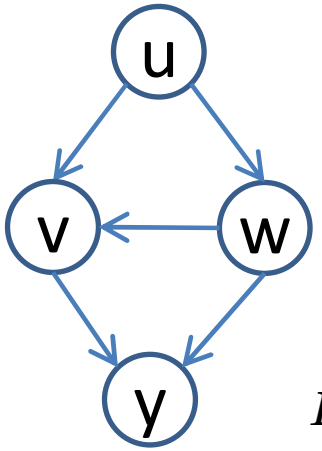


... Another definition of series-parallel graphs?

Another definition (Non SP-graphs)

G is non-SP iff MaxRed(G) contains $G_{forbidden}$

$G_{forbidden}$



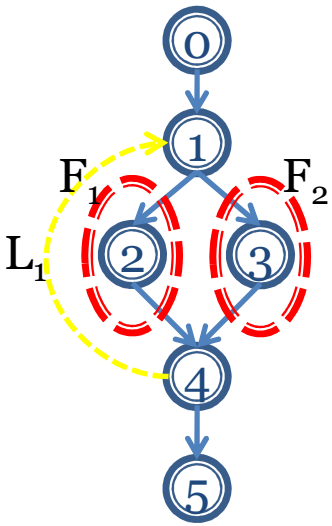
v and w are called reduction nodes

Intuitively, such graphs cannot be synchronized

Subgraph isomorphism is polynomial for SP graphs

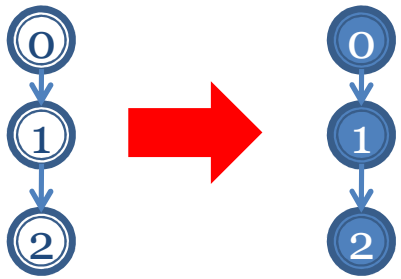
SPFL-Workflow Model (PDiffView)

- Workflow Specification
  - A series-parallel graph <u>overlaid with well-nested fork and loop subgraphs</u>
  - Four kinds of executions: series, parallel, fork and loop

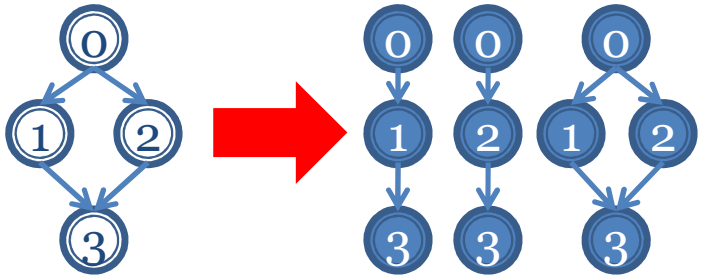


Spec

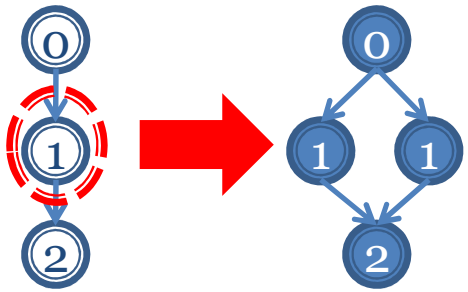
Series Execution



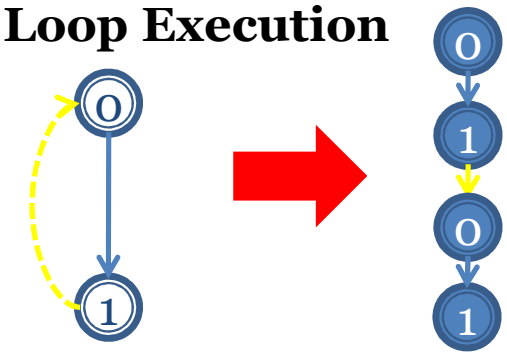
Parallel Execution



Fork Execution

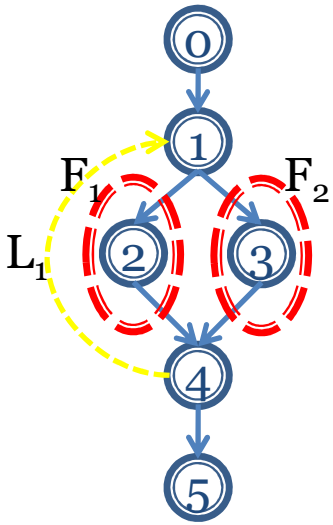


Loop Execution

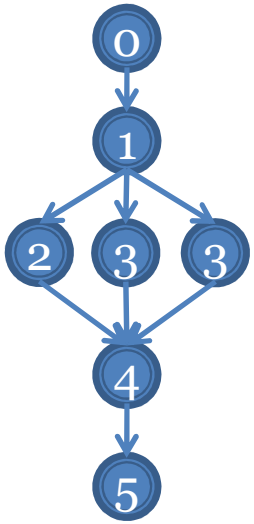


SPFL-Workflow Model

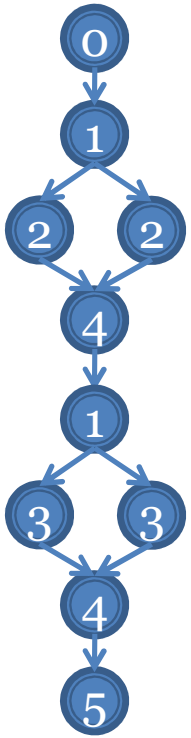
- Valid Runs
  - Derived from the specification by applying series, parallel, fork and loop executions recursively



Spec (G, F, L)



Valid run R_1



Valid run R_2

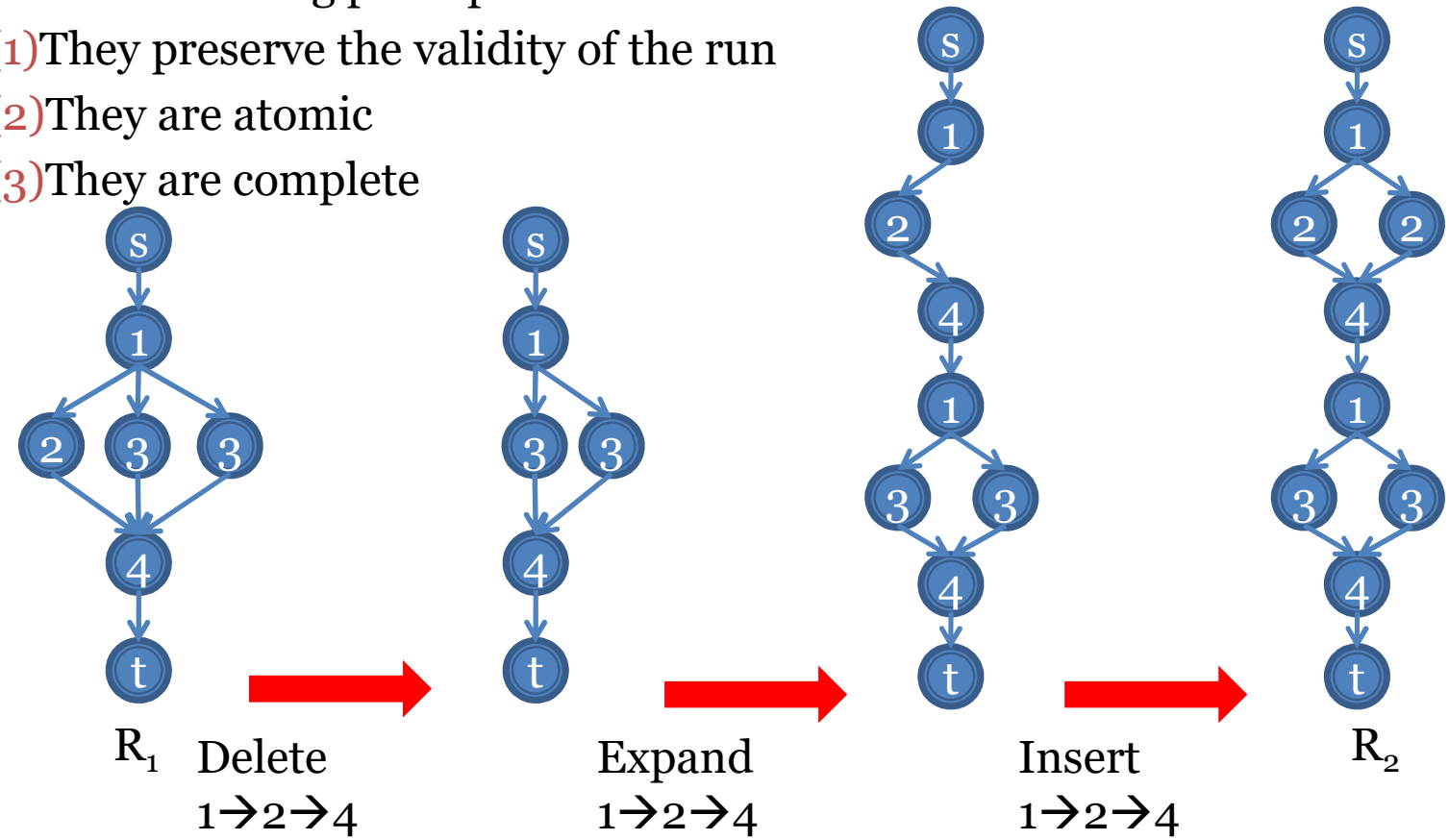
Edit Operations

- Path Insertion, Deletion, Expansion, Contraction
  - **Elementary path**: each internal vertex has exactly one incoming edge and one outgoing edge, and the resulting graph is still valid with respect to the specification.
  - Three motivating principles

    (1)They preserve the validity of the run

    (2)They are atomic

    (3)They are complete

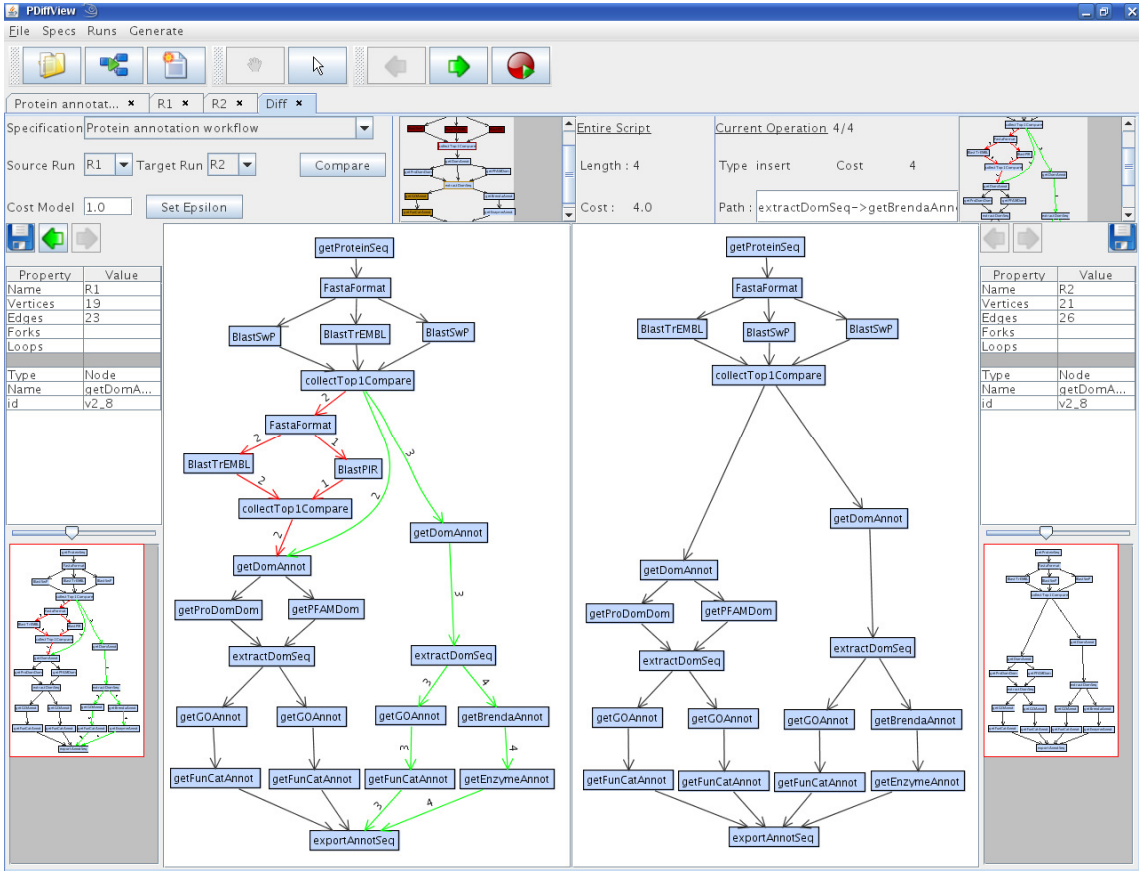


PDiffView

Polynomial-time algorithm designed in PDiffView for SPFL workflows

Problem statement

Given a pair of valid runs R_1 and R_2 of the same specification, and a cost function, compute a **minimum cost edit script** that transforms R_1 to R_2. The cost of this edit script is also known as the **edit distance** between R_1 and R_2



Conclusion on workflows

- Workflows plays a crucial role in biological data integration
- Various areas of computer sciences are involved
  - Databases (e.g., to query and store them)
  - Software engineering (e.g., to optimize or rewrite them)
  - Graph algorithmics (e.g., to query and compare them)
  - … and a lot of other optimization techniques
- Very large spectrum of challenges
  - From very theoretical (e.g., graph theory, equivalence of programs) to very technical and practical (user study, benchmarking on real data sets…)

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