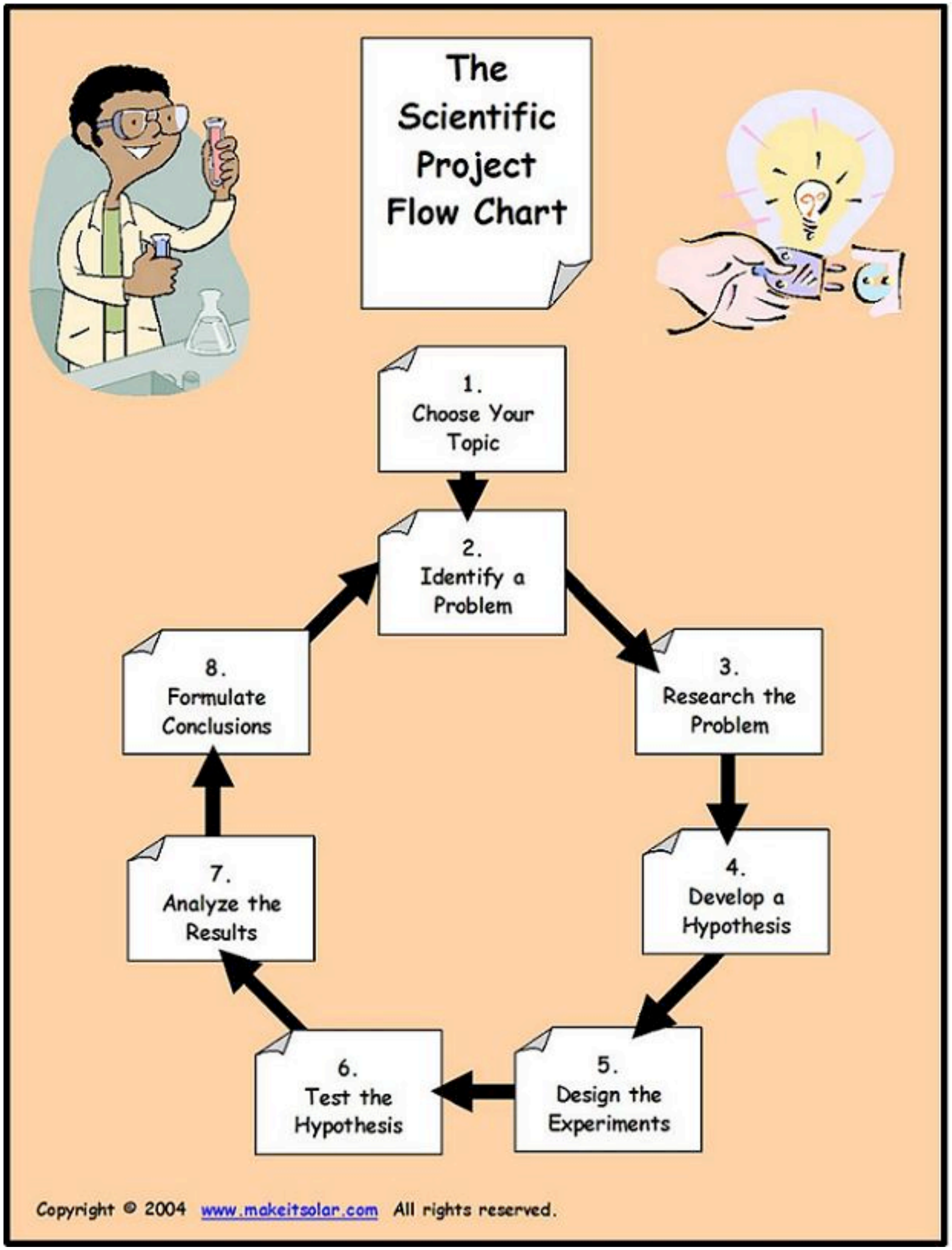
Services for science
Creating knowledge in the Internet age

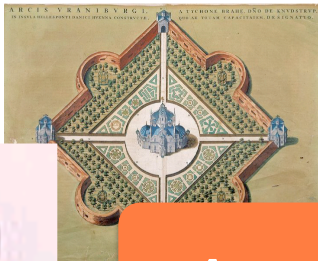
Ian Foster

Computation Institute

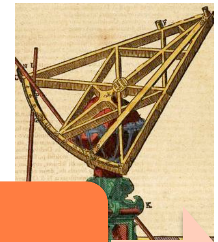
Argonne National Lab & University of Chicago

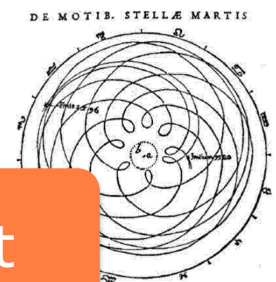

The Scientific Project Flow Chart
1. Choose Your Topic
2. Identify a Problem
3. Research the Problem
4. Develop a Hypothesis
5. Design the Experiments
6. Test the Hypothesis
7. Analyze the Results
8. Formulate Conclusions
Copyright © 2004 www.makeitsolar.com All rights reserved.

Knowledge generation in astronomy ~1600



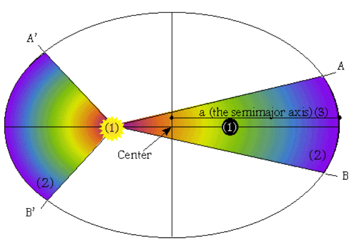
Acquire funding → Build apparatus → Collect data





? years

30 years



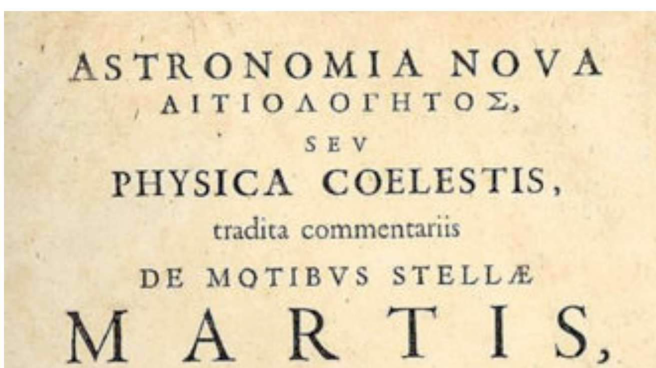
Publish ← Analyze data ← Acquire data

2 years

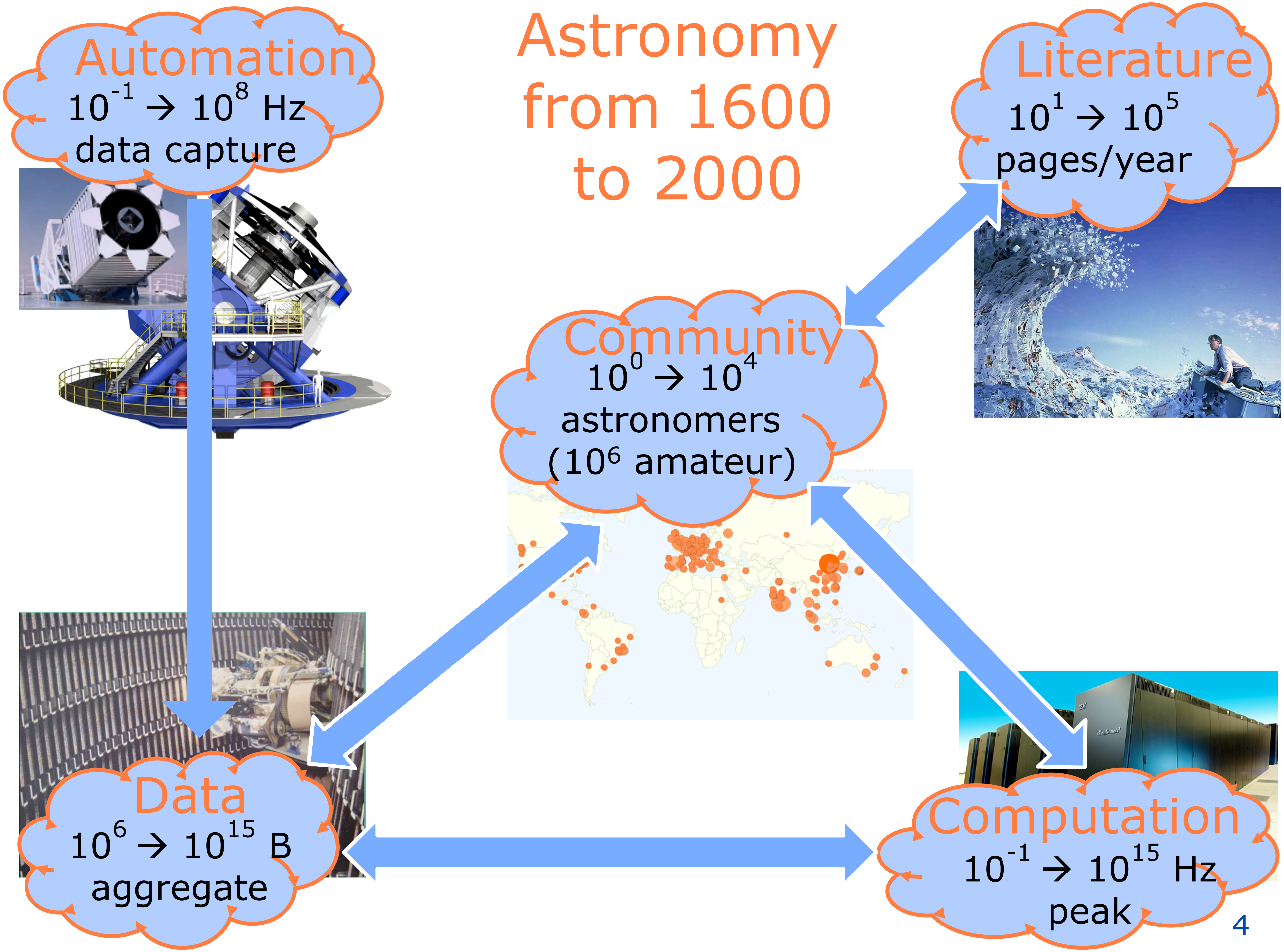
6 years

10 years

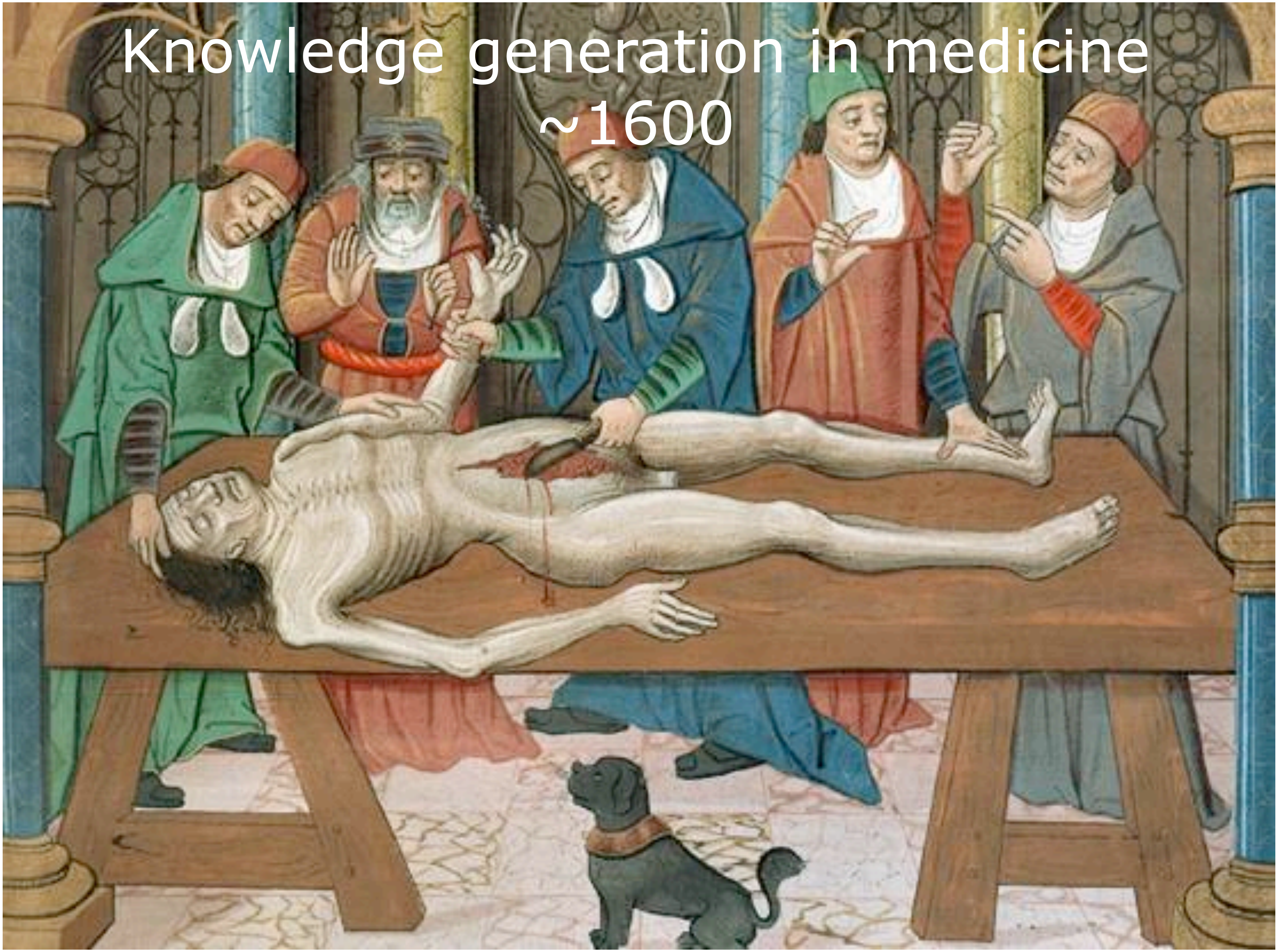
Steal data ← Poison advisor


ASTRONOMIA NOVA
ΑΙΤΙΟΛΟΓΗΤΟΣ,
SEV
PHYSICA COELESTIS,
tradita commentariis
DE MOTIBVS STELLÆ
MARTIS,

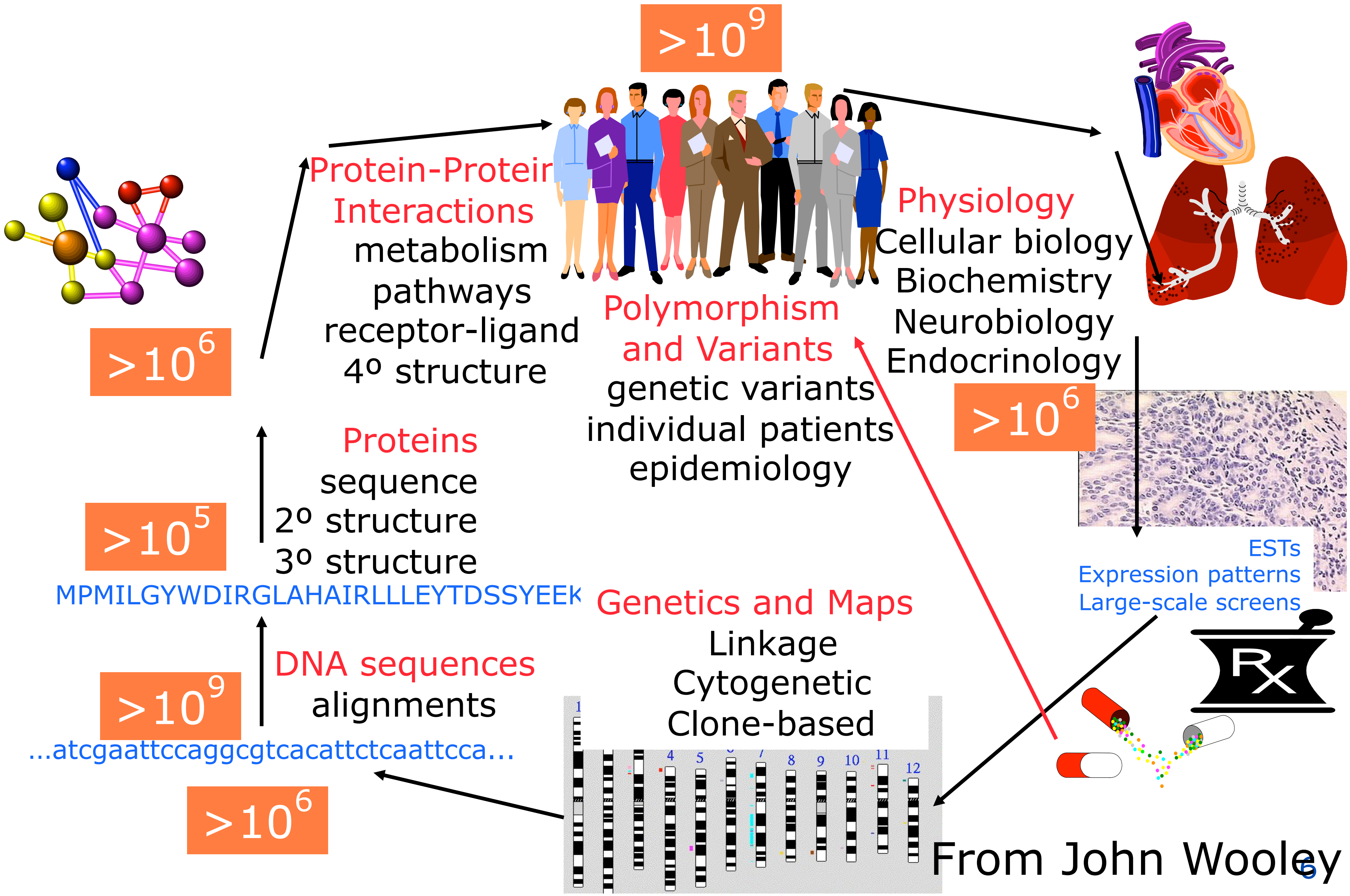
Astronomy from 1600 to 2000



Knowledge generation in medicine ~1600

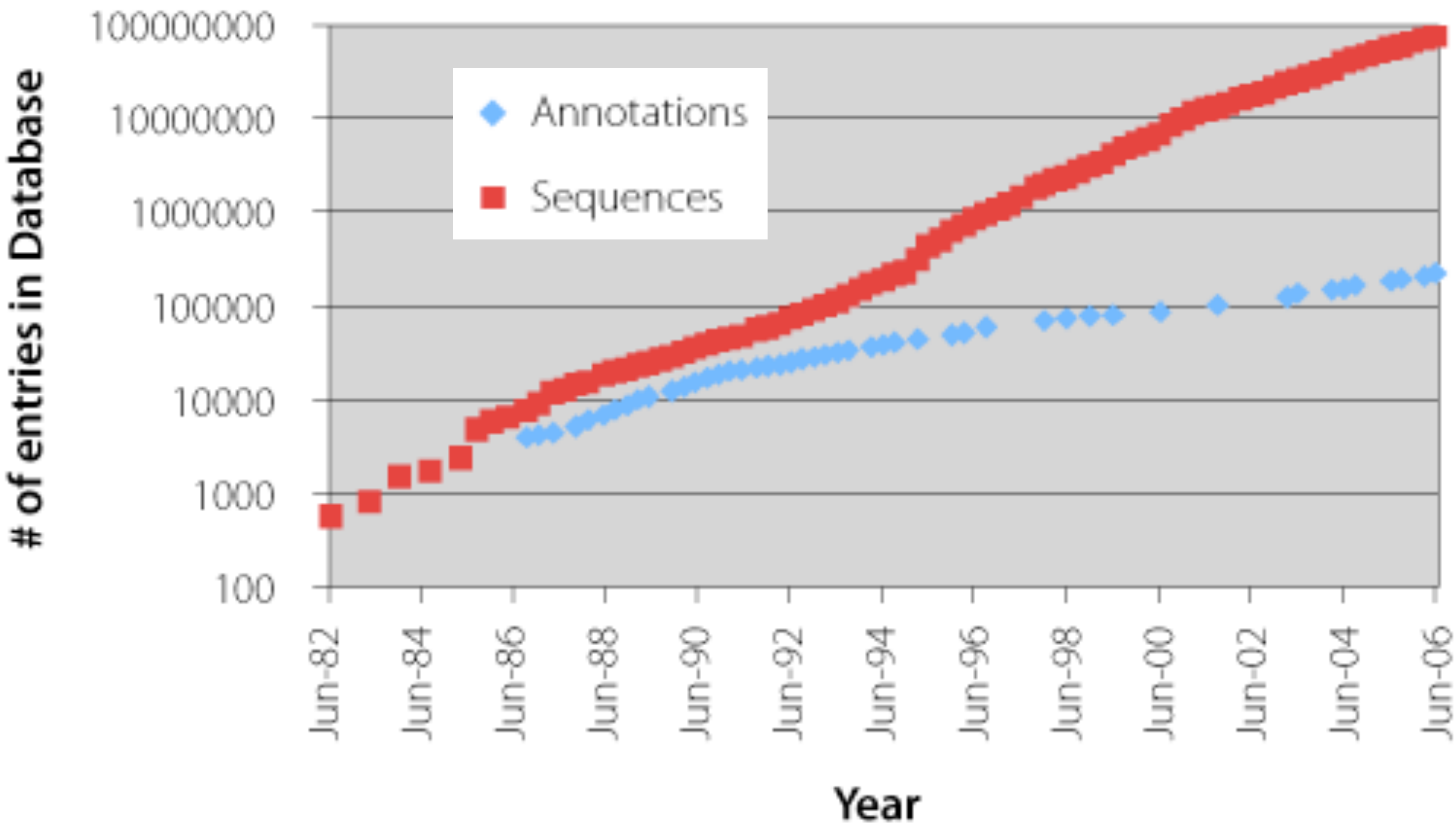


Biomedical research ~2000



From John Wooley

More data does not always mean more knowledge



Folker Meyer, Genome Sequencing vs. Moore's Law: Cyber Challenges for the Next Decade, **CTWatch**, August 2006.

Knowledge generation as a systems problem

- Many diverse actors
- Complex, often rapidly evolving processes
- Need for scalability in multiple dimensions

- With systemic properties
  - Rate of knowledge generation (throughput)
  - Time to answer questions (latency)
  - Completeness of exploration
  - Robustness to errors

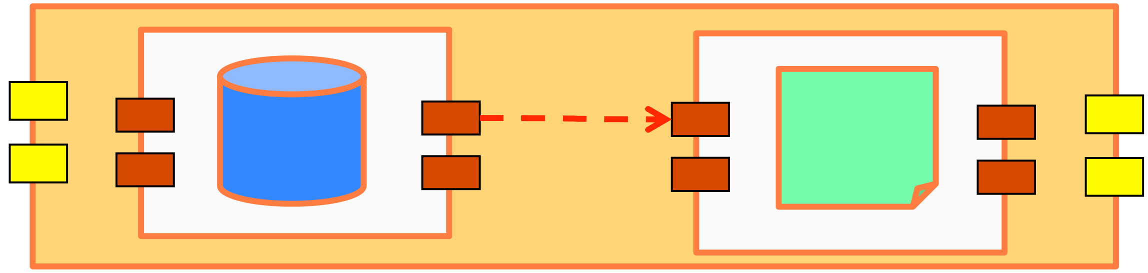
An incomplete list of process steps

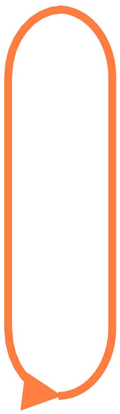
Discover
Access
Integrate
Analyze
Mine
Publish
Annotate
Validate
Curate
Share

Data
Analyses
Models
Experiments
Literature

Artisanal
↓
Industrial

SOA as an integrating framework?





We expose data and software as **services** …
which others **discover**, decide to use, …
and **compose** to create new functions …
which they **publish** as new services.

Technical … and **socio-technical** challenges

- Complexity
- Semantics
- Distribution
- Scale

- Incentives
- Policy, trust
- Reproducibility
- Life cycle

"Service-Oriented Science", *Science*, 2005

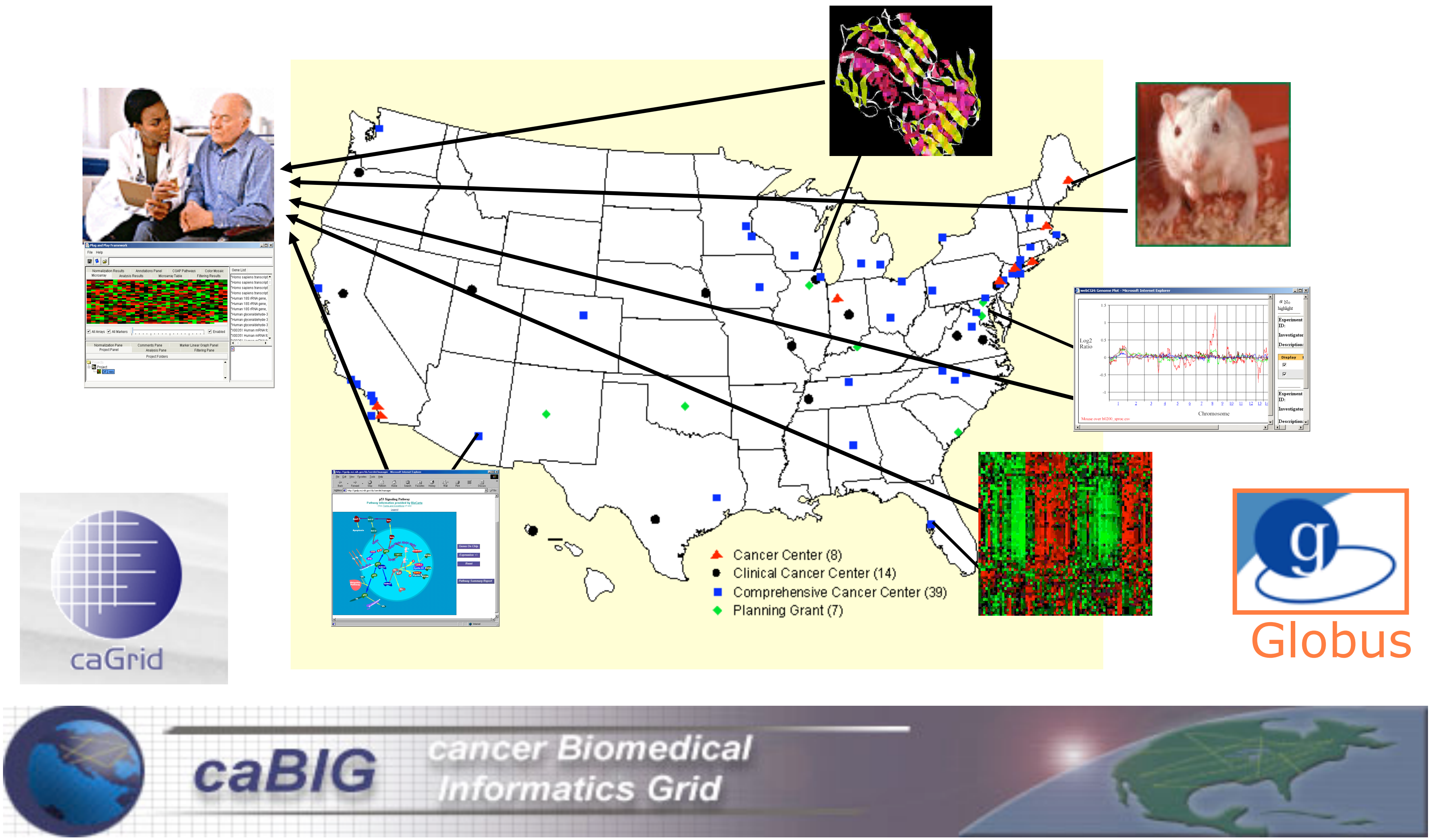
1070 molecular bio databases

Nucleic Acids Research Jan 2008
(96 in Jan 2001)

- Proteomics
- Genomics
- Transcriptomics
- Protein sequence prediction
- Phenotypic studies
- Phylogeny
- Sequence analysis
- Protein structure prediction
- Protein-protein interaction
- Metabolomics
- Model organism collections
- Systems biology
- Health epidemiology
- Organisms
- Disease ….

Slide: Carole Goble

The cancer Biomedical Informatics Grid



As of Sept 18, 2008: → 122 participants
81 services → 62 data / 19 analytical

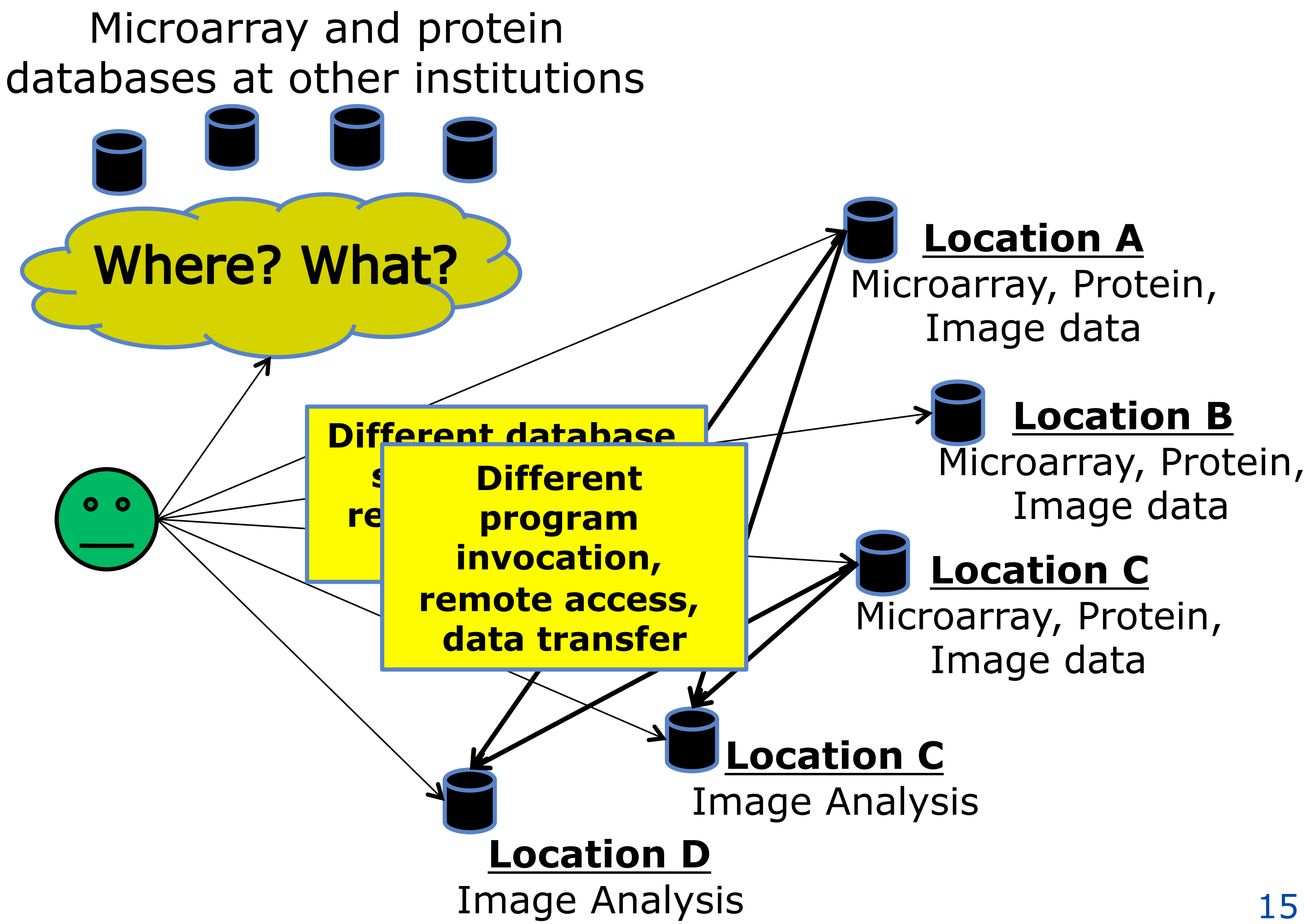
caBIG cancer Biomedical Informatics Grid

As of **Oct 19**, 2008: → 122 participants

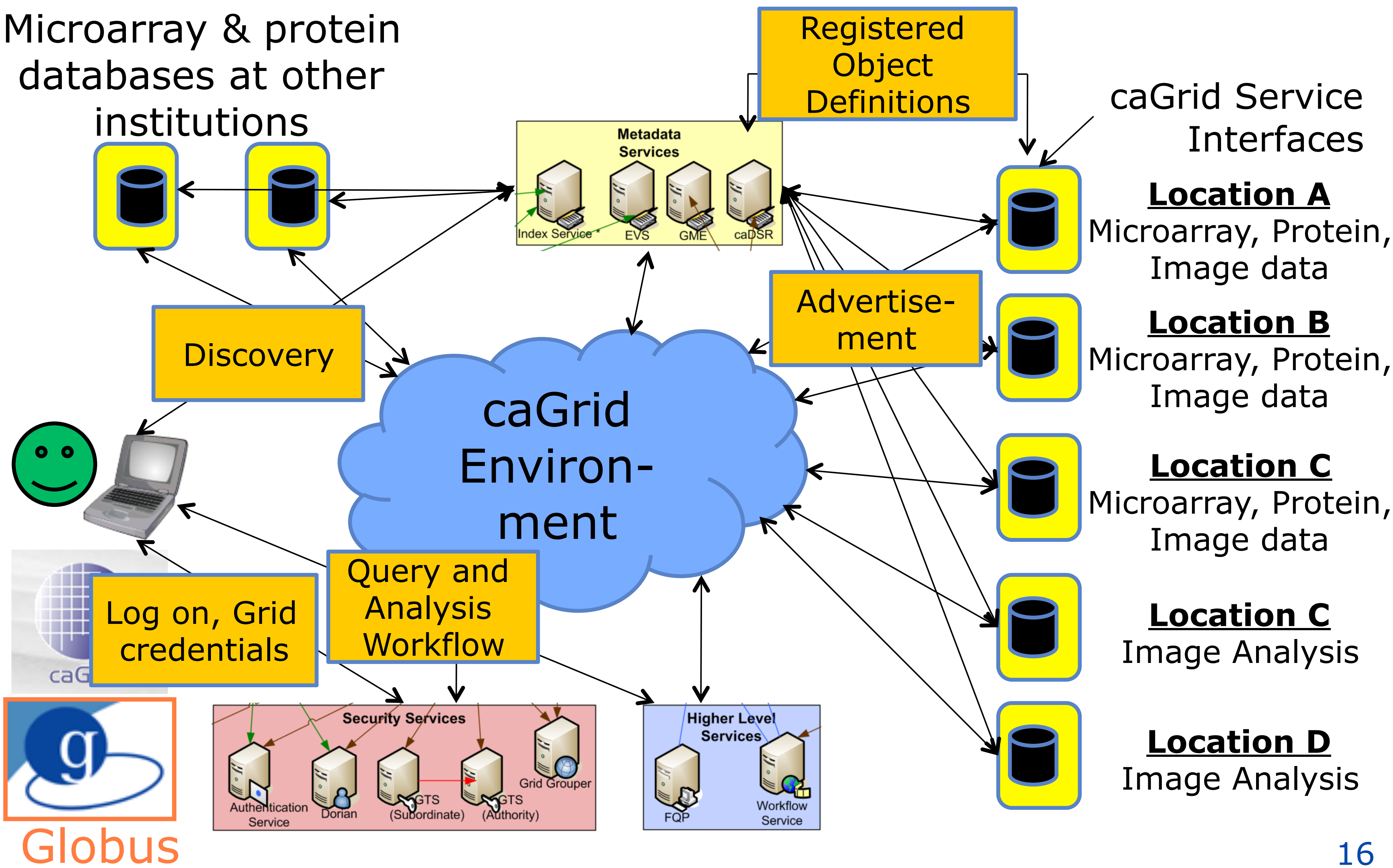
105 services → { **70** data, **35** analytical }

caBIG cancer Biomedical Informatics Grid

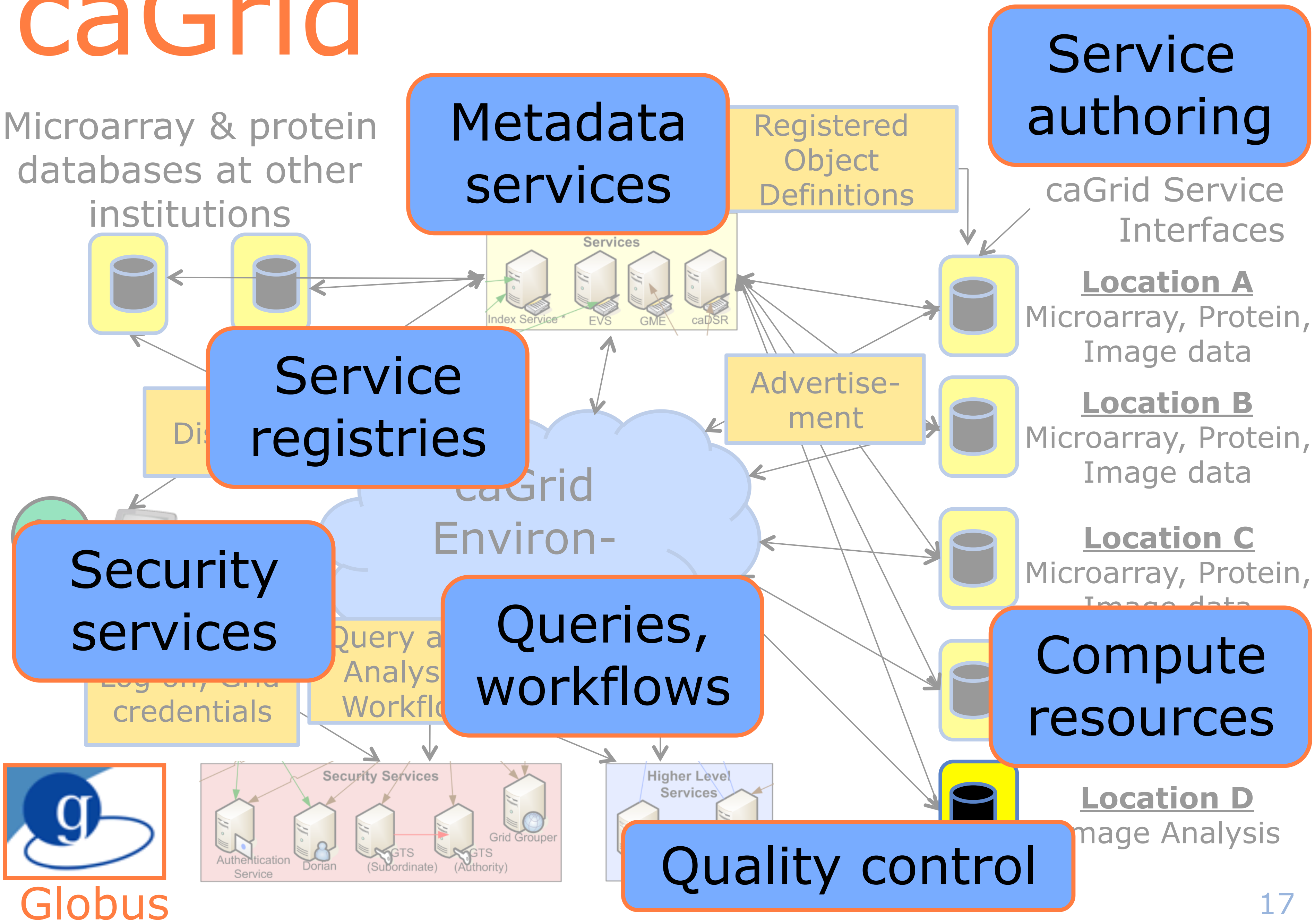
Automating the routine



Automating the routine

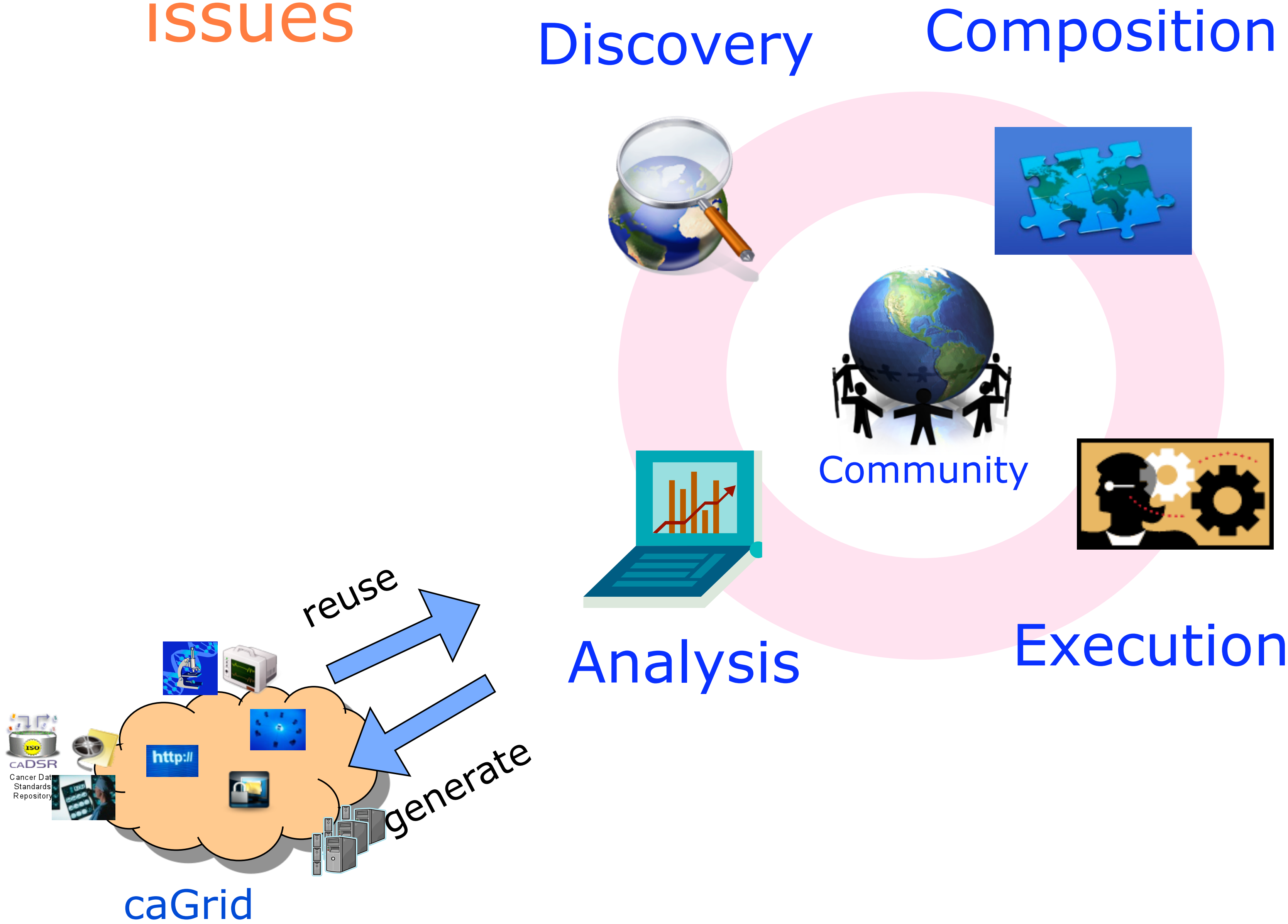


caGrid

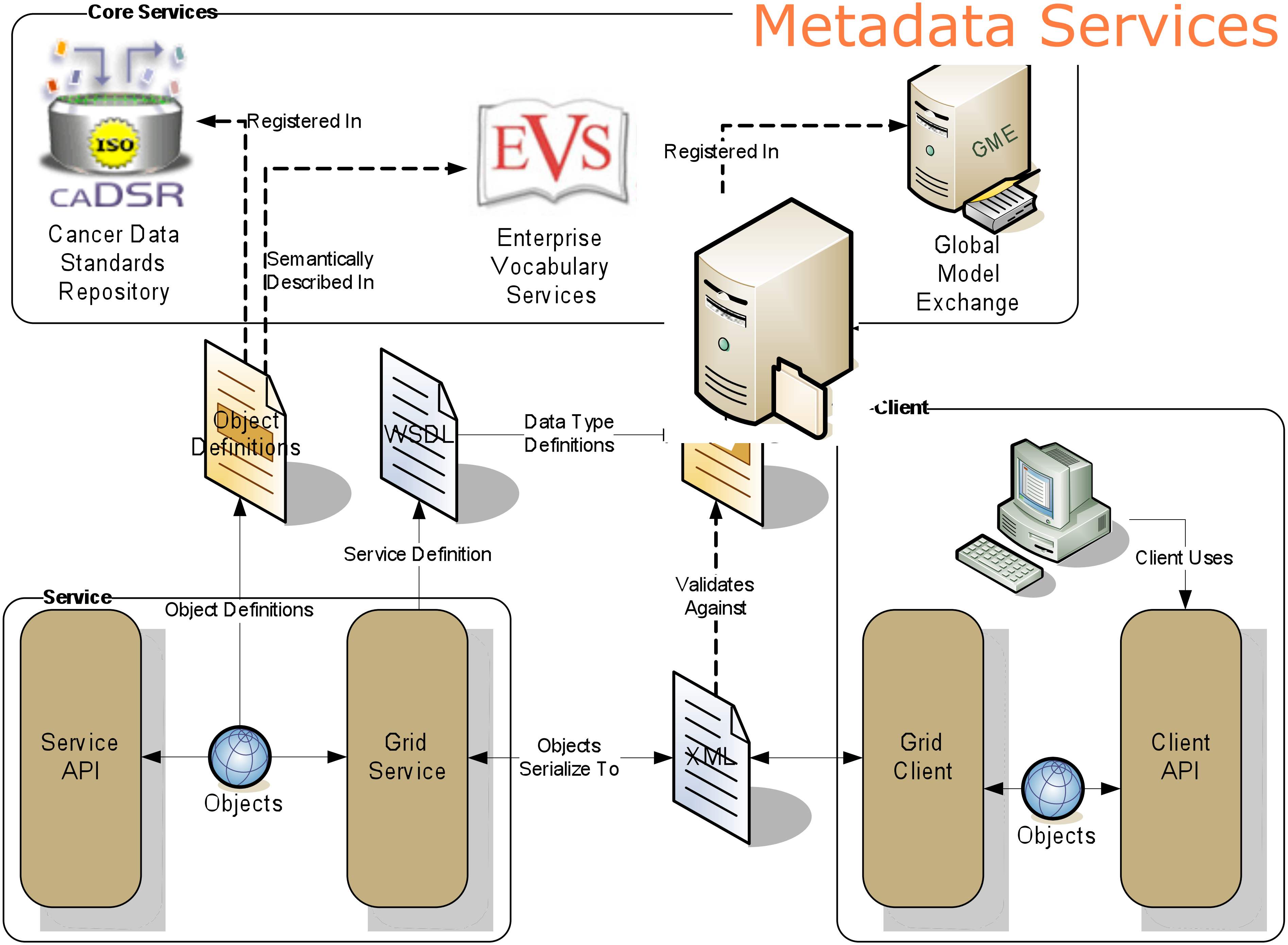


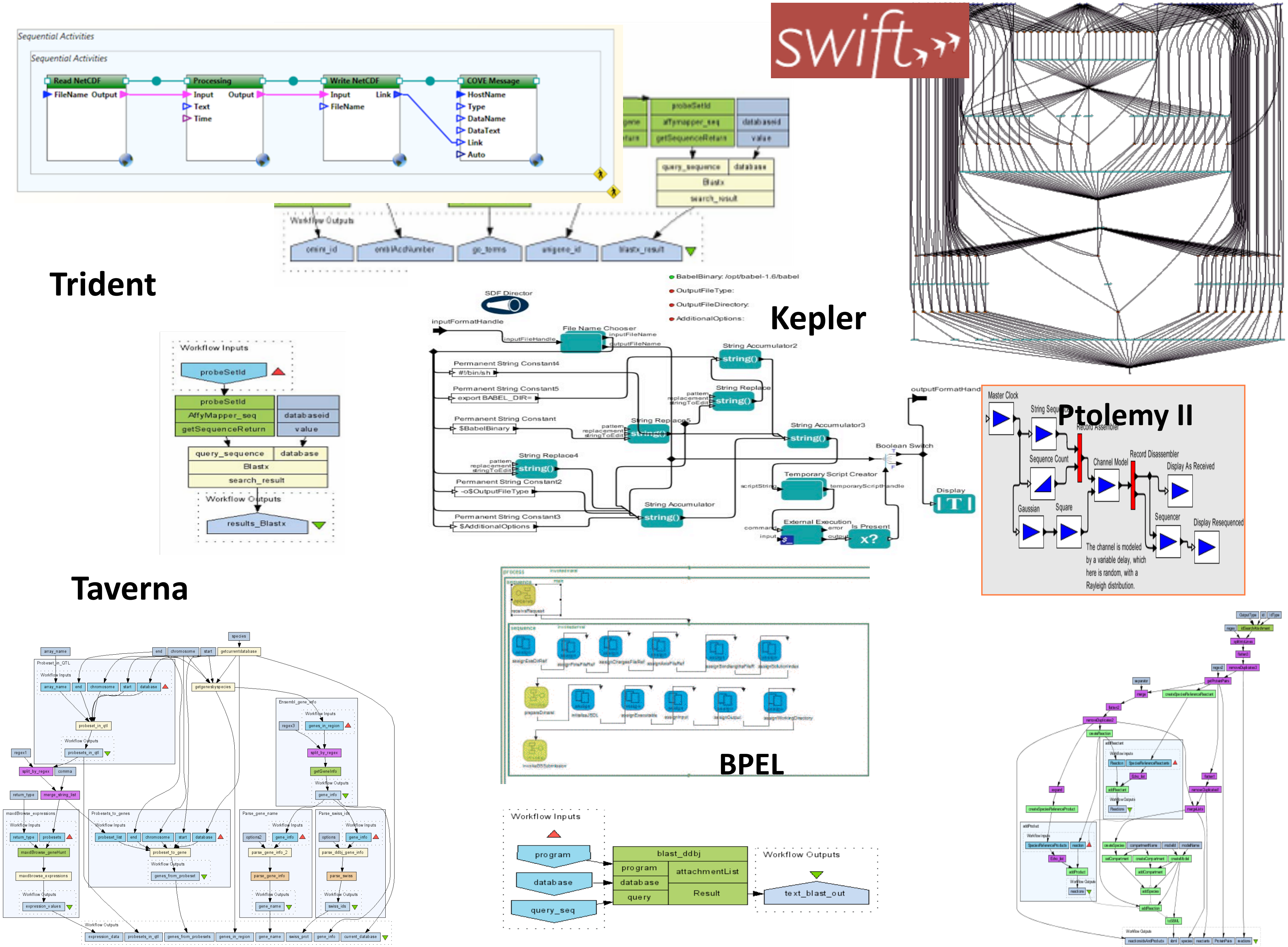
Microarray & protein databases at other institutions
Metadata services
Registered Object Definitions
Service authoring
caGrid Service Interfaces
Services
Index Service
EVS
GME
caDSR
Location A
Microarray, Protein, Image data
Location B
Microarray, Protein, Image data
Location C
Microarray, Protein,
Advertise-ment
Service registries
caGrid Environ-
Security services
credentials
Queries, workflows
Compute resources
Security Services
Authentication Service
Dorian
GTS (Subordinate)
GTS (Authority)
Grid Grouper
Higher Level Services
Location D
mage Analysis
Quality control
Globus

Lifecycle issues



Metadata Services



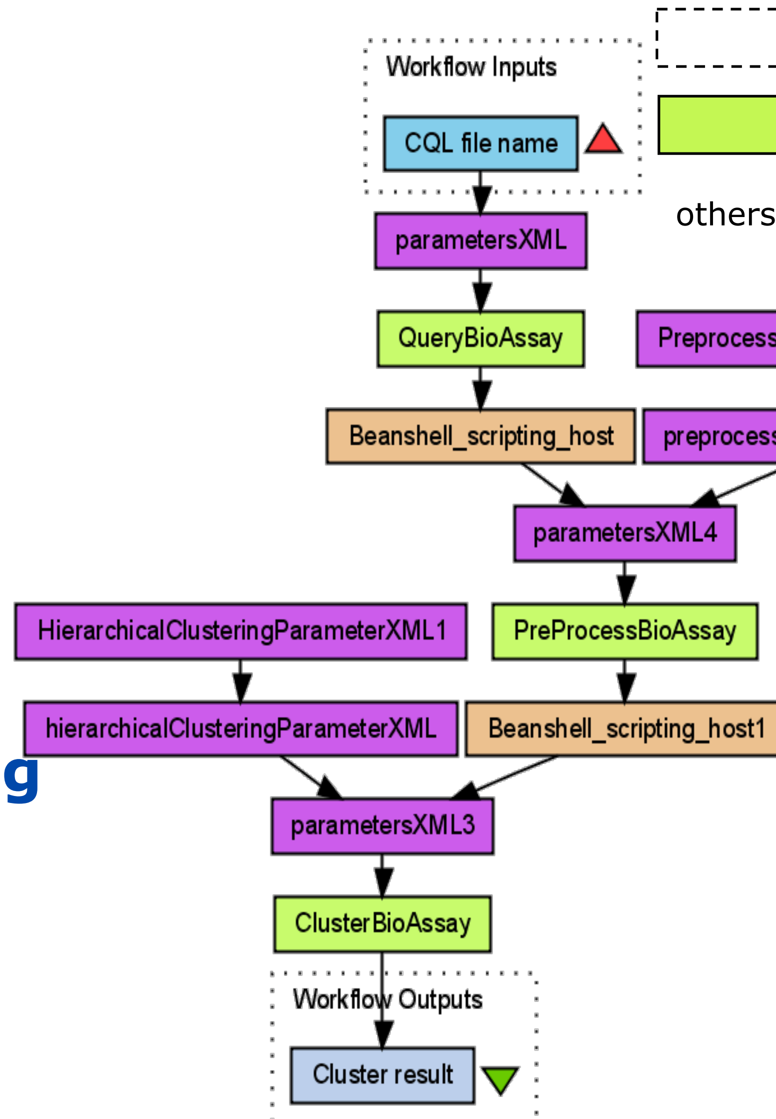

swift
Trident
Kepler
Ptolemy II
Taverna
BPEL

Microarray clustering in caBIG

1. **Query** and retrieve microarray data from a caArray data service: *cagridnode.c2b2.columbia.edu:8080/wsrf/services/cagrid/CaArrayScrub*

2. **Normalize** microarray data using GenePattern analytical service *node255.broad.mit.edu:6060/wsrf/services/cagrid/PreprocessDatasetMAGEService*

3. **Hierarchical clustering** using geWorkbench analytical service: *cagridnode.c2b2.columbia.edu:8080/wsrf/services/cagrid/HierarchicalClusteringMage*



(Taverna workflow)

Workflows as communication

Experimental method

Know-how

Standing operating procedures

Transparent science

Intellectual property

First class scientific assets

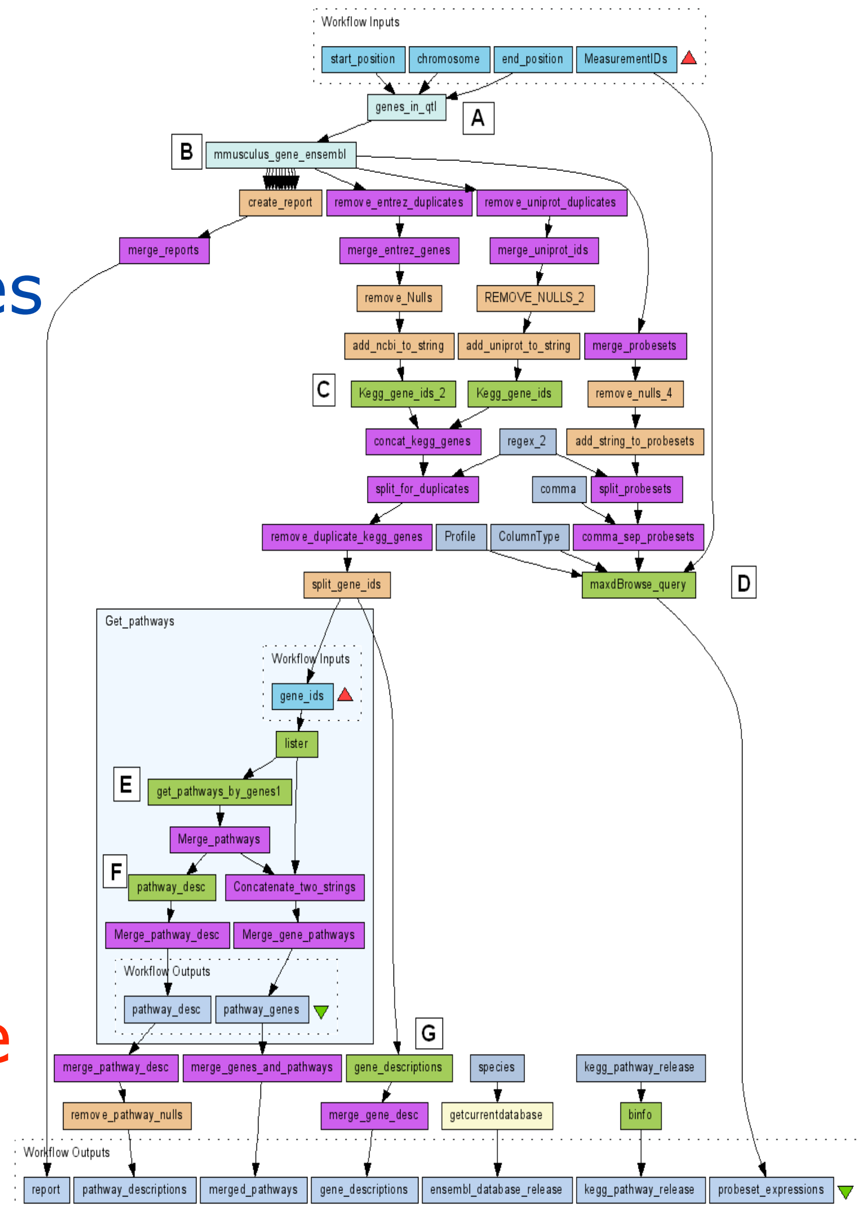
Memes

Variant design

To be reused and mashed up

Hard to design, esp. for reuse

Hard to reuse, esp. across discipline boundaries



Slide: Carole Goble

nature

www.nature.com/nature

Vol 442 | Issue no. 7098 | 6 July 2006

Illuminating the black box

Note to biologists: submissions to *Nature* should contain complete descriptions of materials and reagents used.

This journal aims to publish papers that are not only interesting and thought-provoking, but reproducible and useful. In order to do this, novel materials and reagents need to be carefully described and readily available to interested scientists.

established didn't want the author to reveal the sequences, as this would jeopardize its *raison d'être*. This kind of stalemate matters, because it prevents the replication of experiments and inhibits the selection of appropriate controls in subsequent work.

Reproducible science means

— context

— trust

— easy access to methods

from a company with a stake in the antibody in question.

But the problem is most acute in the case of new technologies, which sometimes experience a period of rapid development during

is unnecessary, because if appropriate controls are described other investigators will know how to control their experiments. This is a false premise: the controls for an experiment designed to test one

nature

www.nature.com/nature

Vol 442 | Issue no. 7098 | 6 July 2006

Illuminating the black box

Note to biologists: submissions to *Nature* should contain complete descriptions of materials and reagents used.

This journal aims to publish papers that are not only interesting and thought-provoking, but reproducible and useful. In order to do this, novel materials and reagents need to be carefully described and readily available to interested scientists.

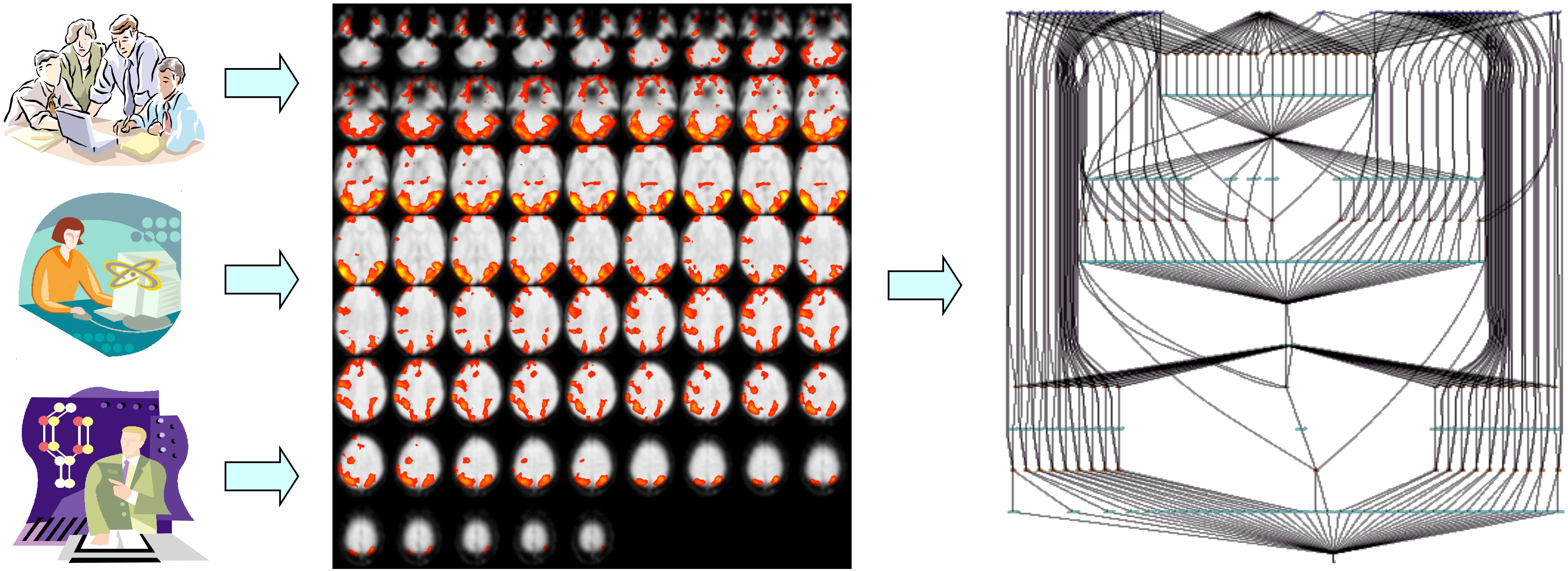
established didn't want the author to reveal the sequences, as this would jeopardize its *raison d'être*. This kind of stalemate matters, because it prevents the replication of experiments and inhibits the selection of appropriate controls in subsequent work.

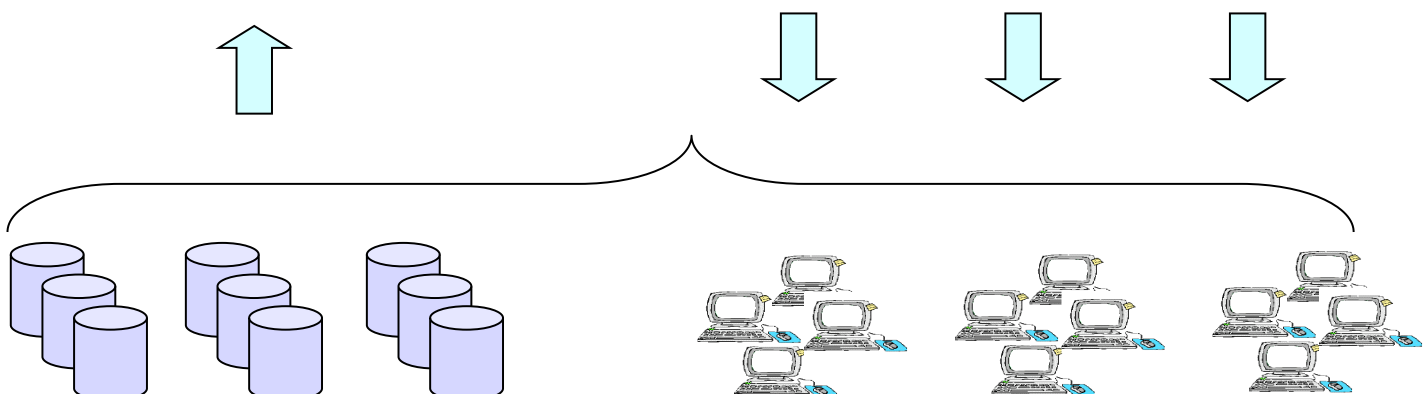
Workflows are another form of scholarly outcome to publish, curate and cite and archive along with data and publications

which sometimes experience a period of rapid development during

false premise: the controls for an experiment designed to test one

Functional Magnetic Resonance Imaging (fMRI)

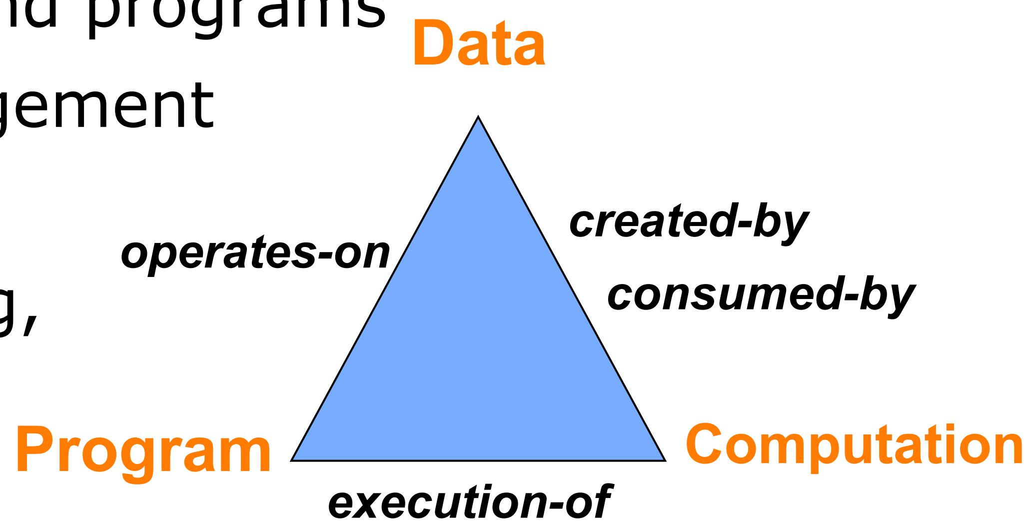




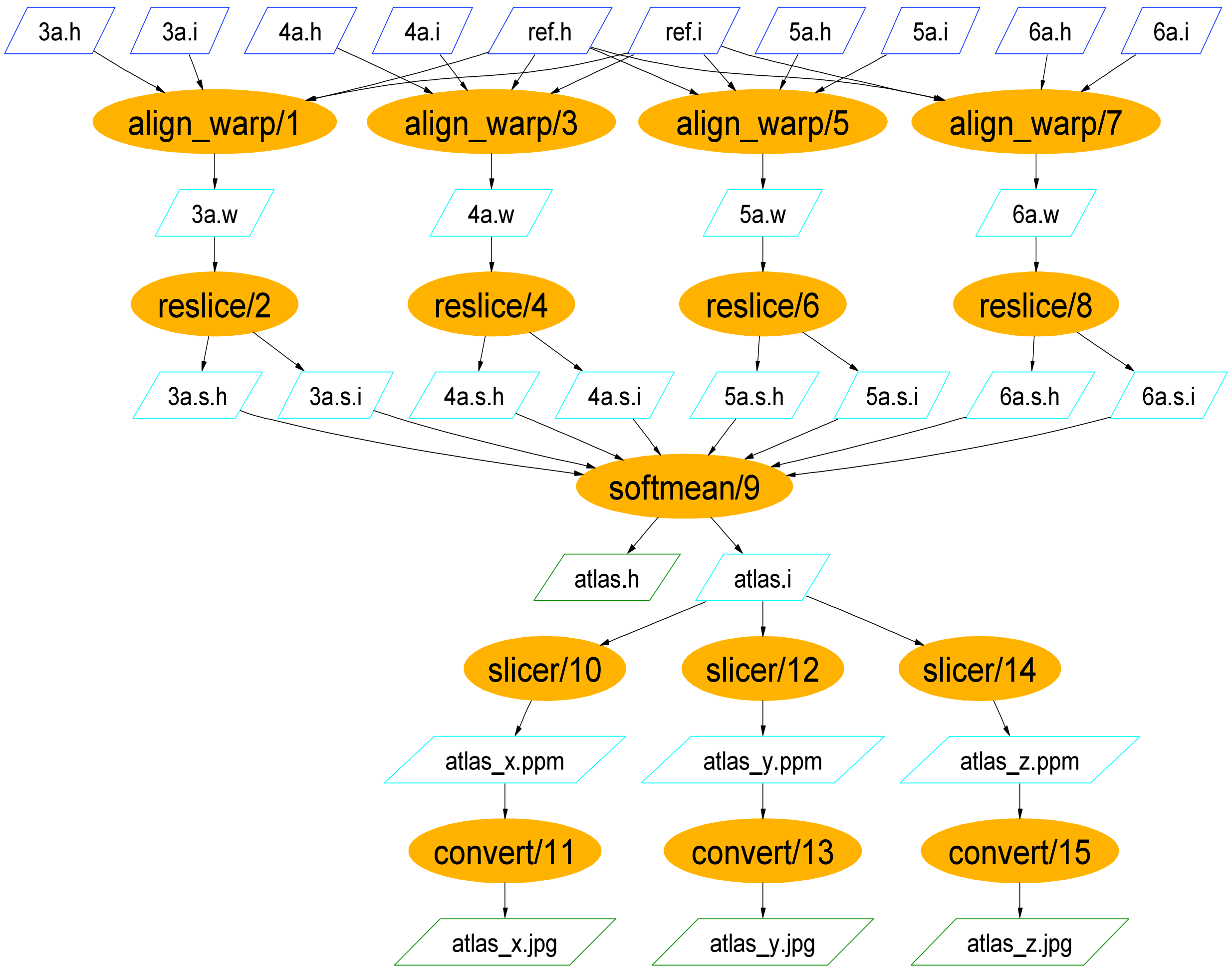
Mike Wilde

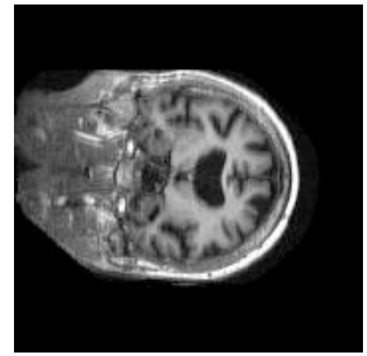
Computation as a first-class entity

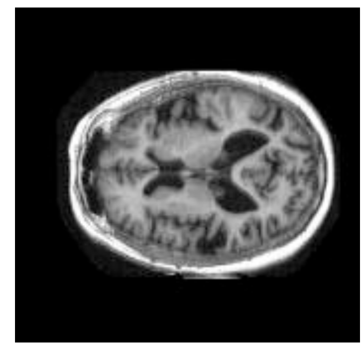
- Capture information about relationships among
  - Data (varying locations and representations)
  - Programs (& inputs, outputs, constraints)
  - Computations (& execution environments)
- Apply this information to:
  - Discovery of data and programs
  - Computation management
  - Provenance
  - Planning, scheduling, performance optimization

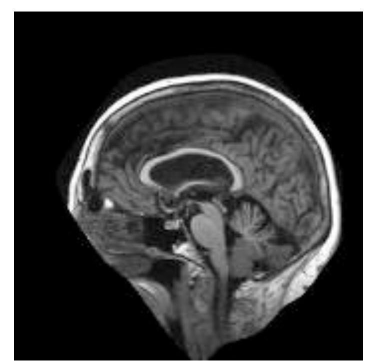


Example: fMRI analysis









First Provenance Challenge, http://twiki.ipaw.info/ [CCPE06]

Query examples

- Query by procedure signature
  - Show procedures that have inputs of type *subjectImage* and output types of *warp*
- Query by actual arguments
  - Show *align_warp* calls (including all arguments), with argument *model=rigid*
- Query by annotation
  - List anonymized subject images for young subjects:
    - Find datasets of type *subjectImage* , annotated with *privacy=anonymized* and *subjectType=young*
- Basic lineage graph queries
  - Find all datasets derived from dataset '5a'
- Graph pattern matching
  - Show me all output datasets of *softmean* calls that were aligned with *model=affine*

Challenges of scale

- Number of participants
- Volume of data
- Diversity of data
- Number of data producers
- Amount of computation

Hosting and provisioning

People **create** services (data or function) …
which others **discover**, decide to use, …
and **compose** to create a new function ...
which they **publish** as a new service.

!!

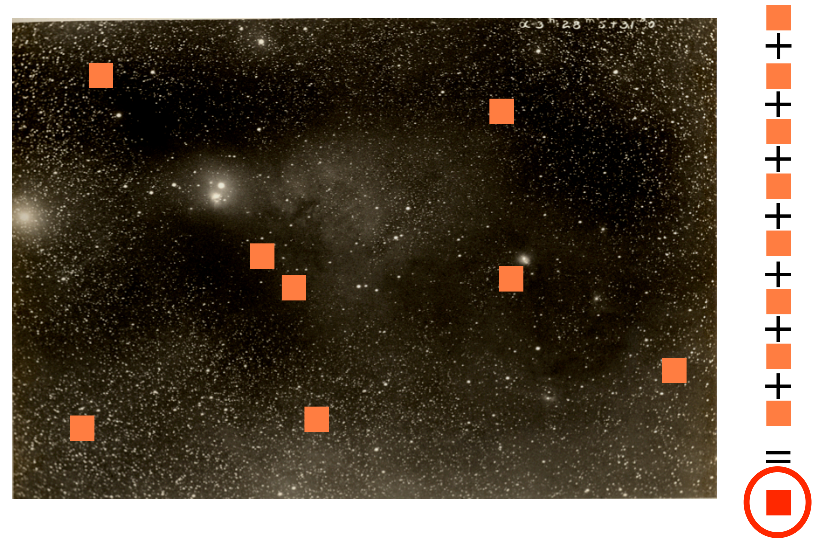
→ *I find "someone else" to **host** services, so I don't have to become an expert in operating services & computers!*

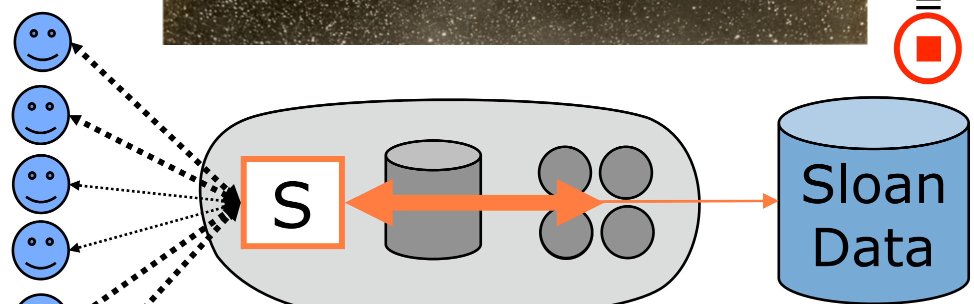
→ *I hope that this "someone else" can **manage** security, reliability, scalability, …*

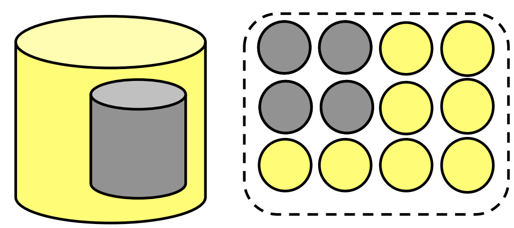
"Service-Oriented Science", *Science*, 2005

Provisioning for data-intensive workloads

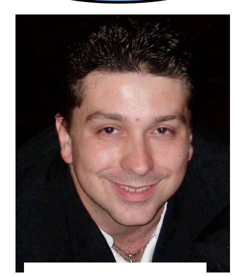
- Example: on-demand "stacking" of arbitrary locations within ~10TB sky survey
- Challenges
  - Random data access
  - Much computing
  - Time-varying load



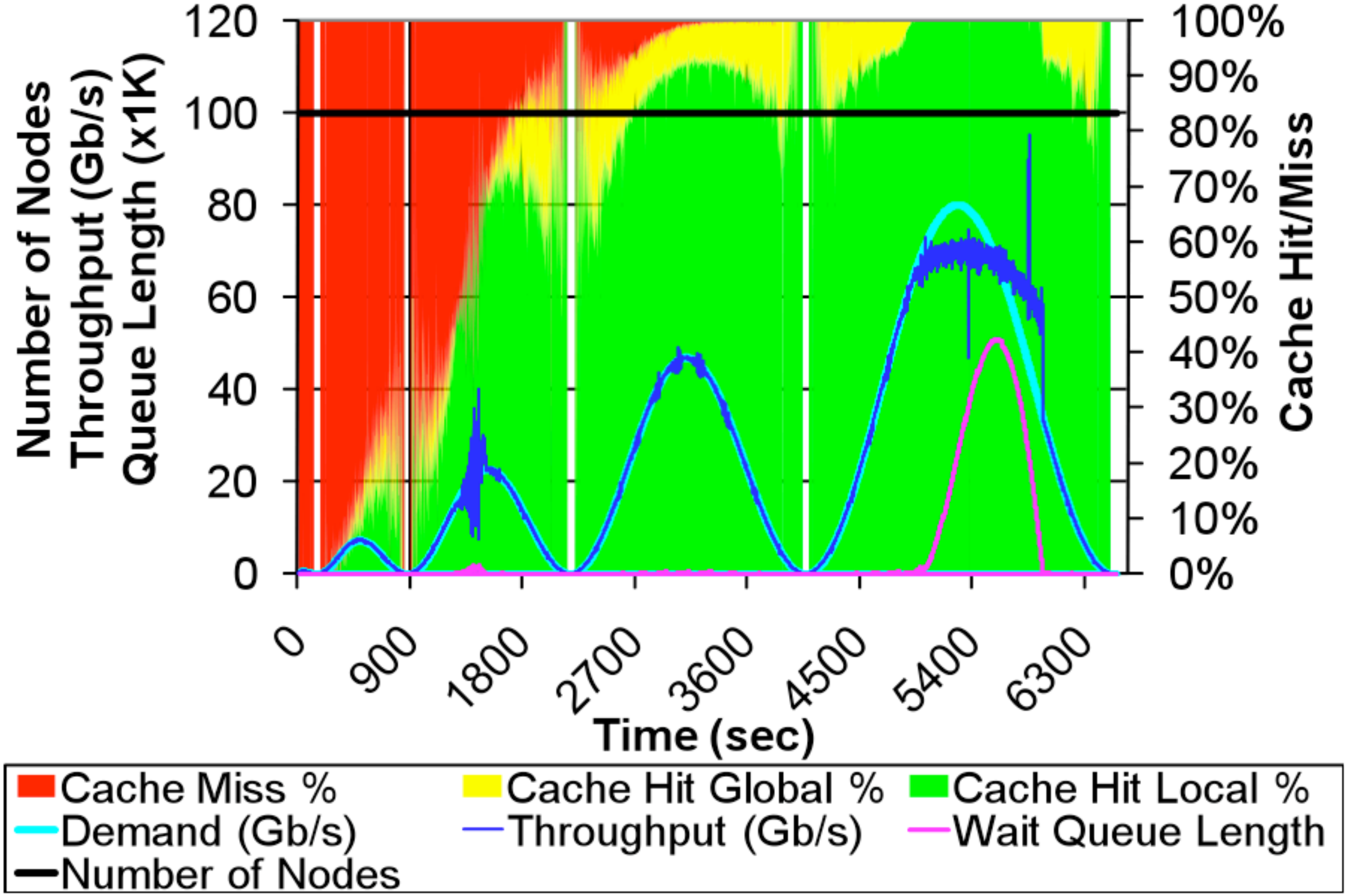




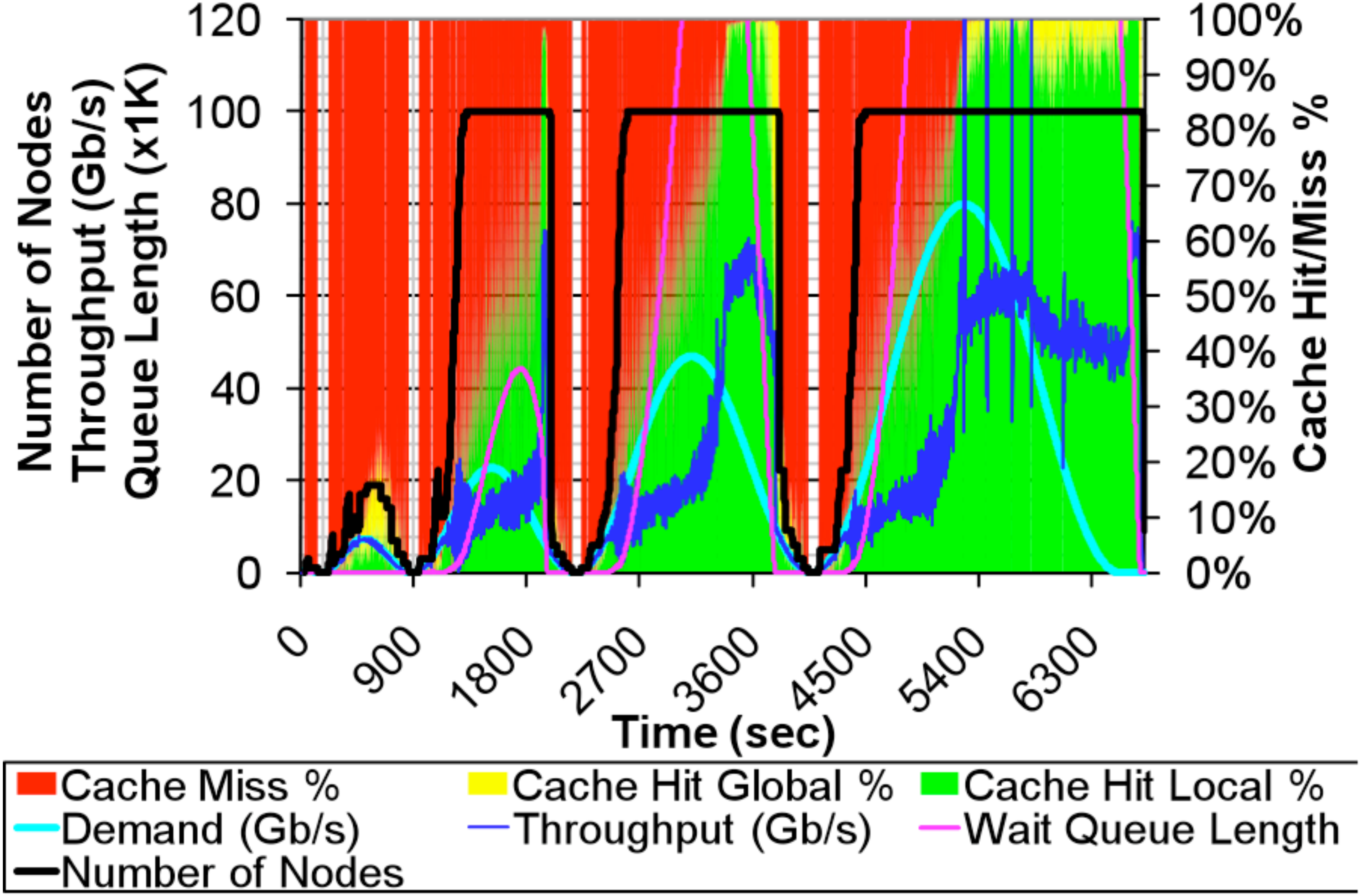
Data diffusion



Ioan Raicu

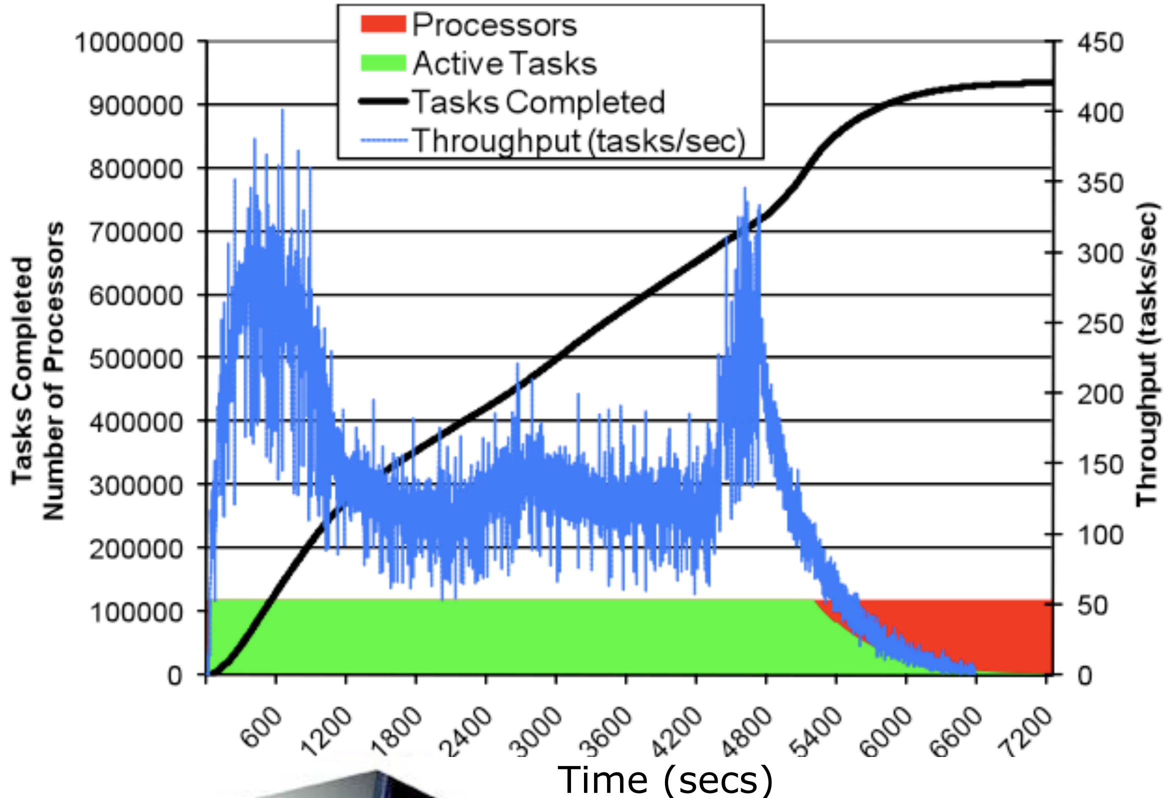


"Sine" workload, 2M tasks, 10MB:10ms ratio, 100 nodes, GCC policy, 50GB caches/node

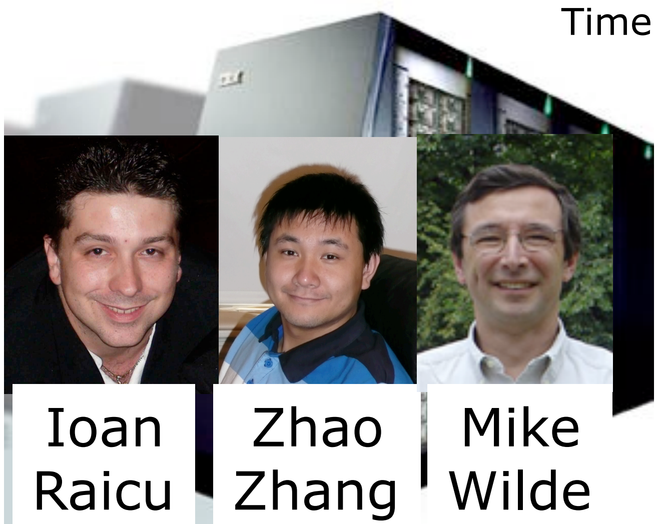


Same scenario, but with dynamic resource provisioning

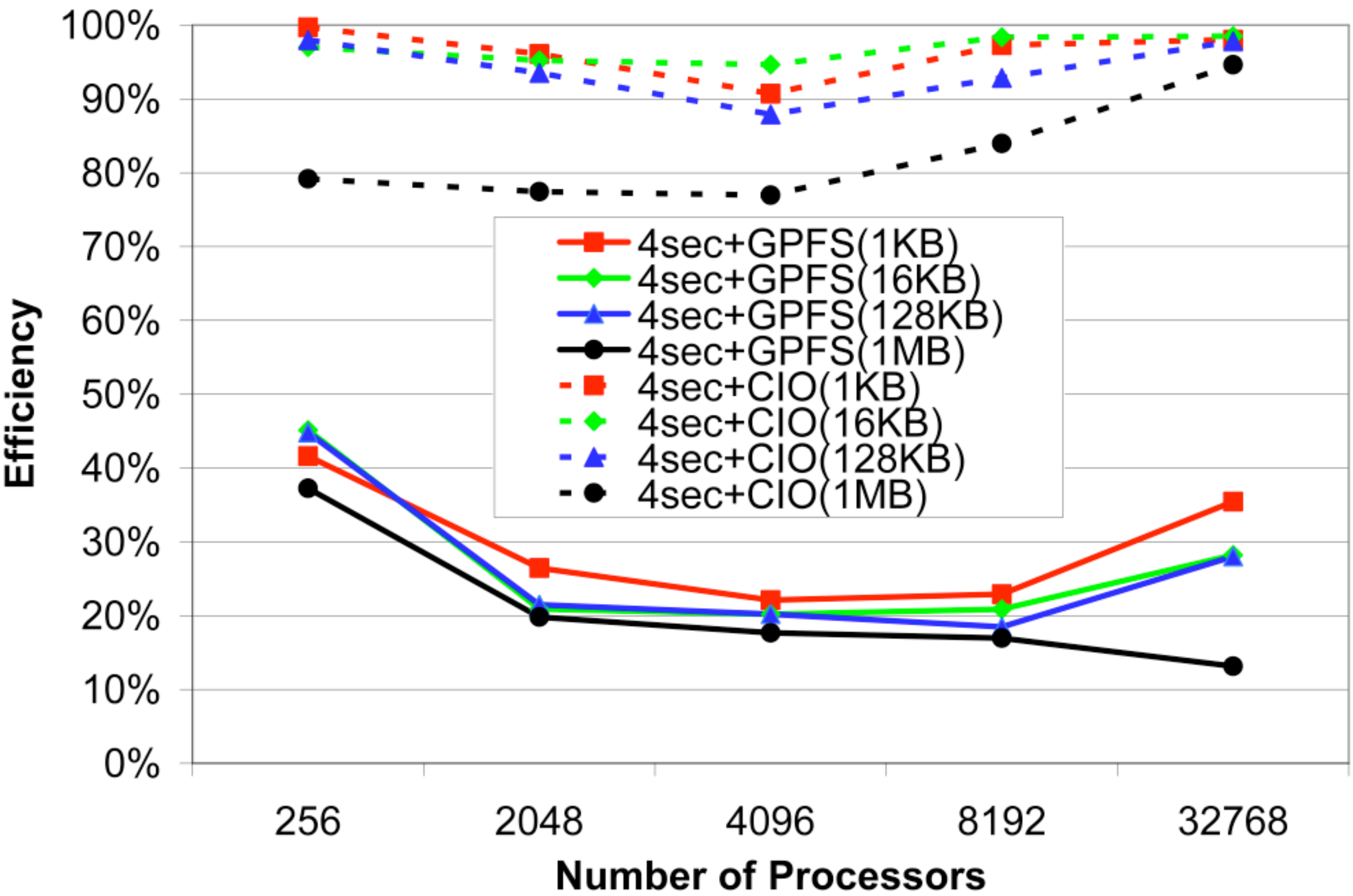
DOCK on BG/P: ~1M Tasks on 118,000 CPUs



- CPU cores: 118784
- Tasks: 934803
- Elapsed time: 7257 sec
- Compute time: 21.43 CPU years
- Average task time: 667 sec
- Relative Efficiency: 99.7%
- (from 16 to 32 racks)
- Utilization:
  - Sustained: 99.6%
  - Overall: 78.3%



- GPFS
  - 1 script (~5KB)
  - 2 file read (~10KB)
  - 1 file write (~10KB)
- RAM (cached from GPFS on first task per node)
  - 1 binary (~7MB)
  - Static input data (~45MB)



Efficiency **relative to no-I/O case** for 4 second tasks and varying data size (1KB to 1MB) for CIO and GPFS up to 32K processors


Text Mining

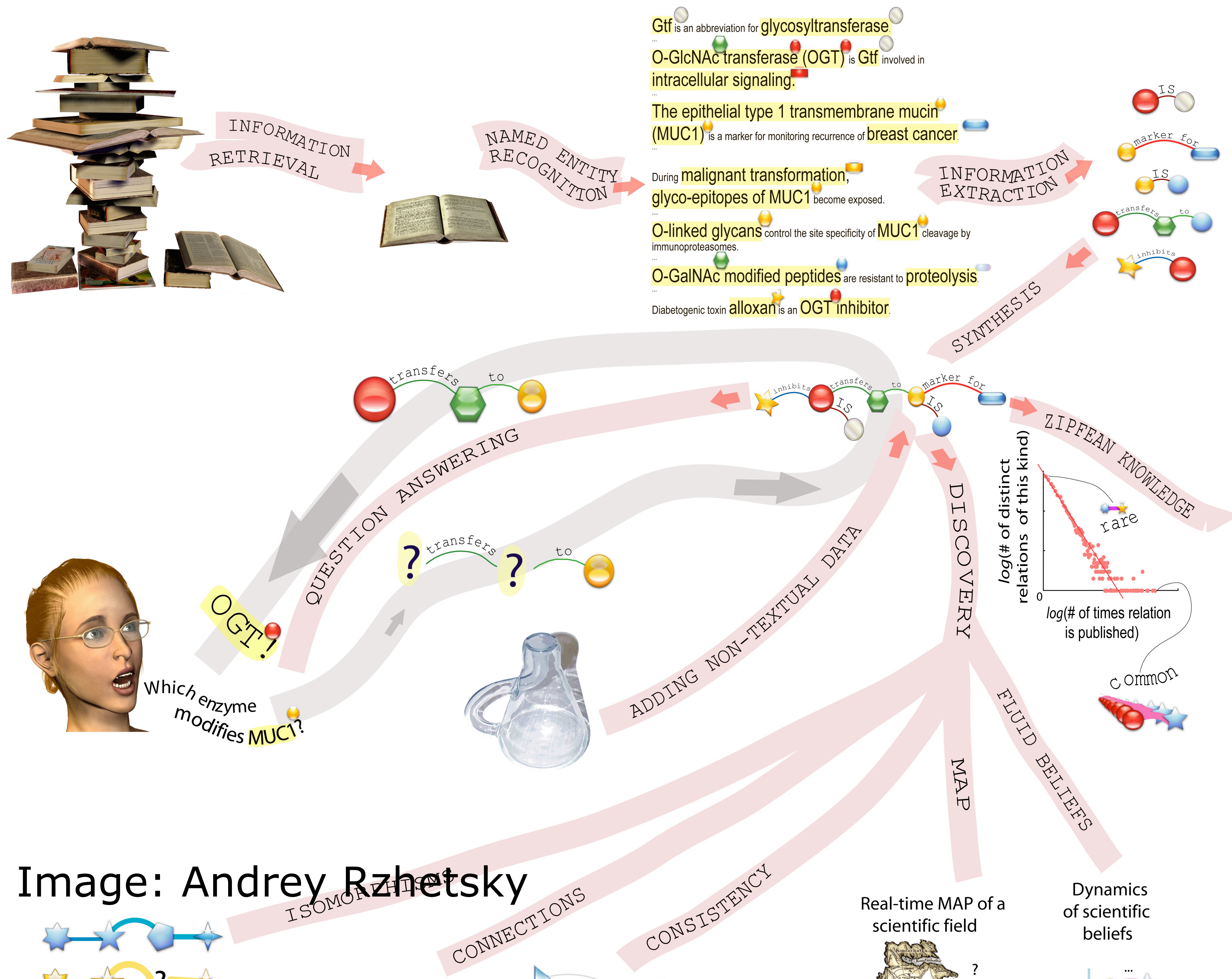
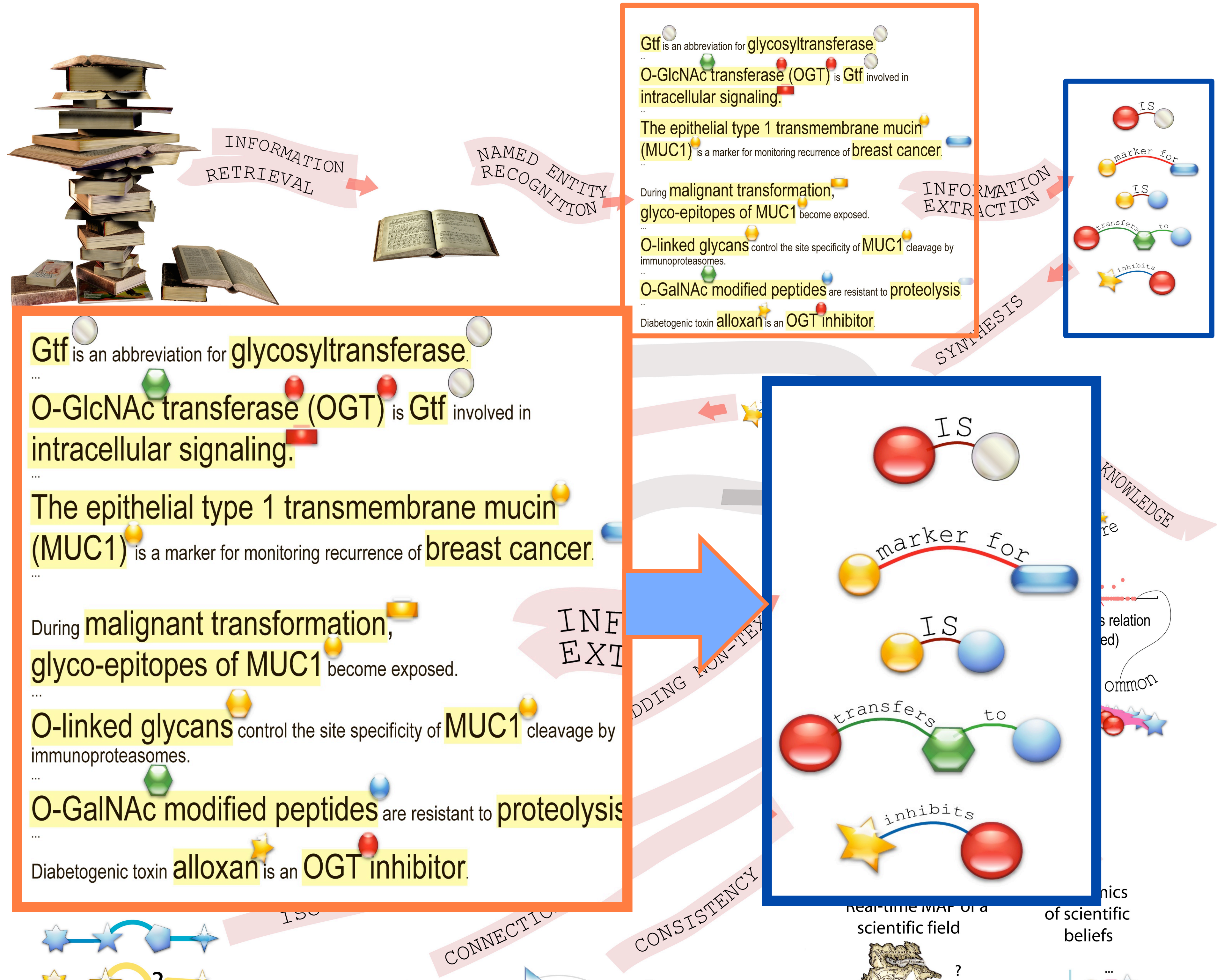


Image: Andrey Rzhetsky


INFORMATION RETRIEVAL
NAMED ENTITY RECOGNITION
INFORMATION EXTRACTION
SYNTHESIS
Gtf is an abbreviation for glycosyltransferase.
...
O-GlcNAc transferase (OGT) is Gtf involved in intracellular signaling.
...
The epithelial type 1 transmembrane mucin (MUC1) is a marker for monitoring recurrence of breast cancer.
...
During malignant transformation, glyco-epitopes of MUC1 become exposed.
...
O-linked glycans control the site specificity of MUC1 cleavage by immunoproteasomes.
...
O-GalNAc modified peptides are resistant to proteolysis.
...
Diabetogenic toxin alloxan is an OGT inhibitor.
IS
marker for
IS
transfers to
inhibits
KNOWLEDGE
CONSISTENCY
scientific field
of scientific beliefs

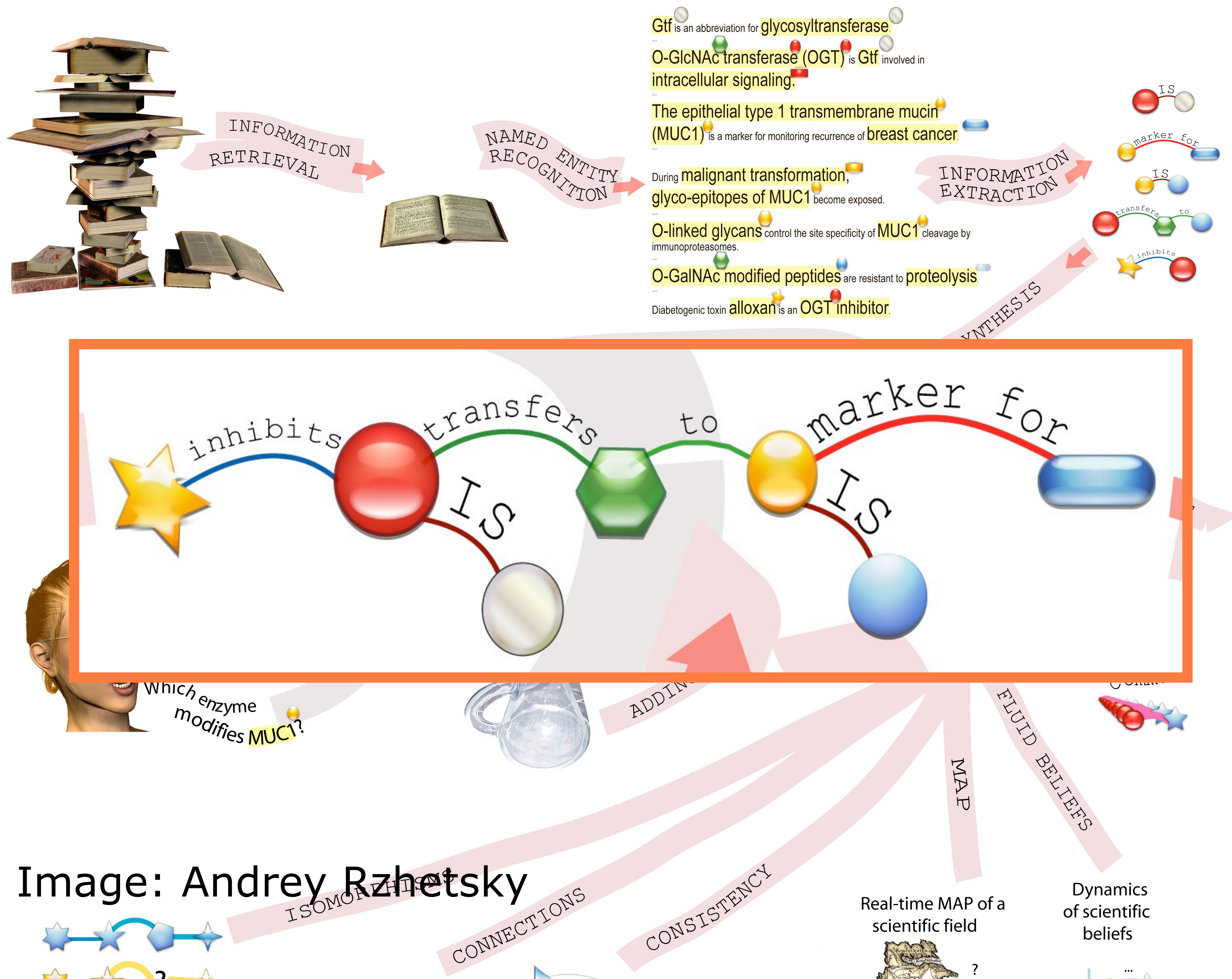


Image: Andrey Rzhetsky

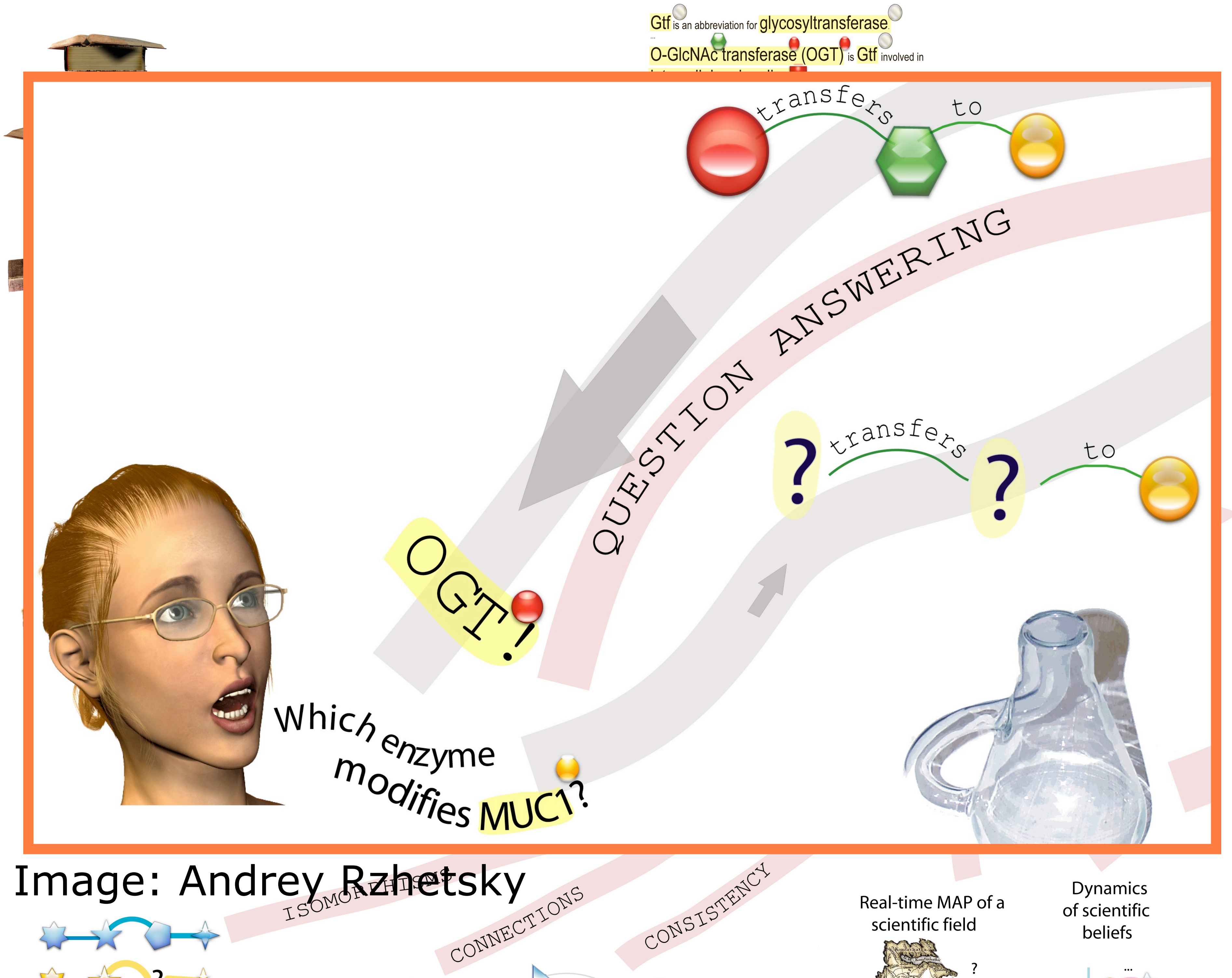


Image: Andrey Rzhetsky

Thanks!

- DOE Office of Science
- National Science Foundation
- National Institutes of Health
- Colleagues at Argonne, U.Chicago, USC/ISI, and elsewhere

Knowledge generation as a systems problem

- Many diverse actors
- Complex, often rapidly evolving processes
- Need for scalability in multiple dimensions
- With systemic properties
  - Rate of knowledge generation (throughput)
  - Time to answer questions (latency)
  - Completeness of exploration
  - Robustness to errors

- SOA as an integrating framework?

Service-oriented science

People **create** services (data or function) …

which others **discover**, decide to use, …

and **compose** to create a new function ...

which they **publish** as a new service.

!!

→ *I find "someone else" to **host** services, so I don't have to become an expert in operating services & computers!*

→ *I hope that this "someone else" can **manage** security, reliability, scalability, …*

"Service-Oriented Science", *Science*, 2005

Service-oriented science

People **create** services (data or function) …
which others **discover**, decide to use, …
and **compose** to create a new function ...
which they **publish** as a new service.

Profoundly revolutionary:
→ Accelerates the pace of enquiry
→ Introduces a new notion of "result"
→ Requires new reward structures, training, infrastructure

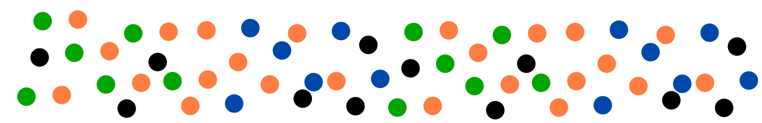
"Service-Oriented Science", *Science*, 2005

And big challenges …

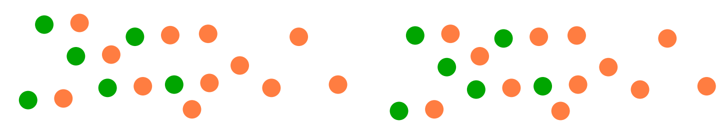
- Complexity and semantics
- Documentation of results
- Scaling in many dimensions
- Sociology and incentives

Service discovery and selection

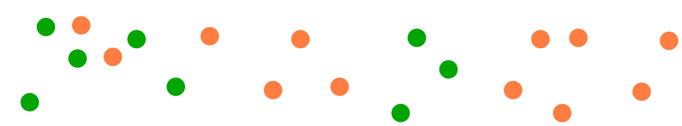
Assume success → Millions of services



Syntax, semantics → Types, ontologies

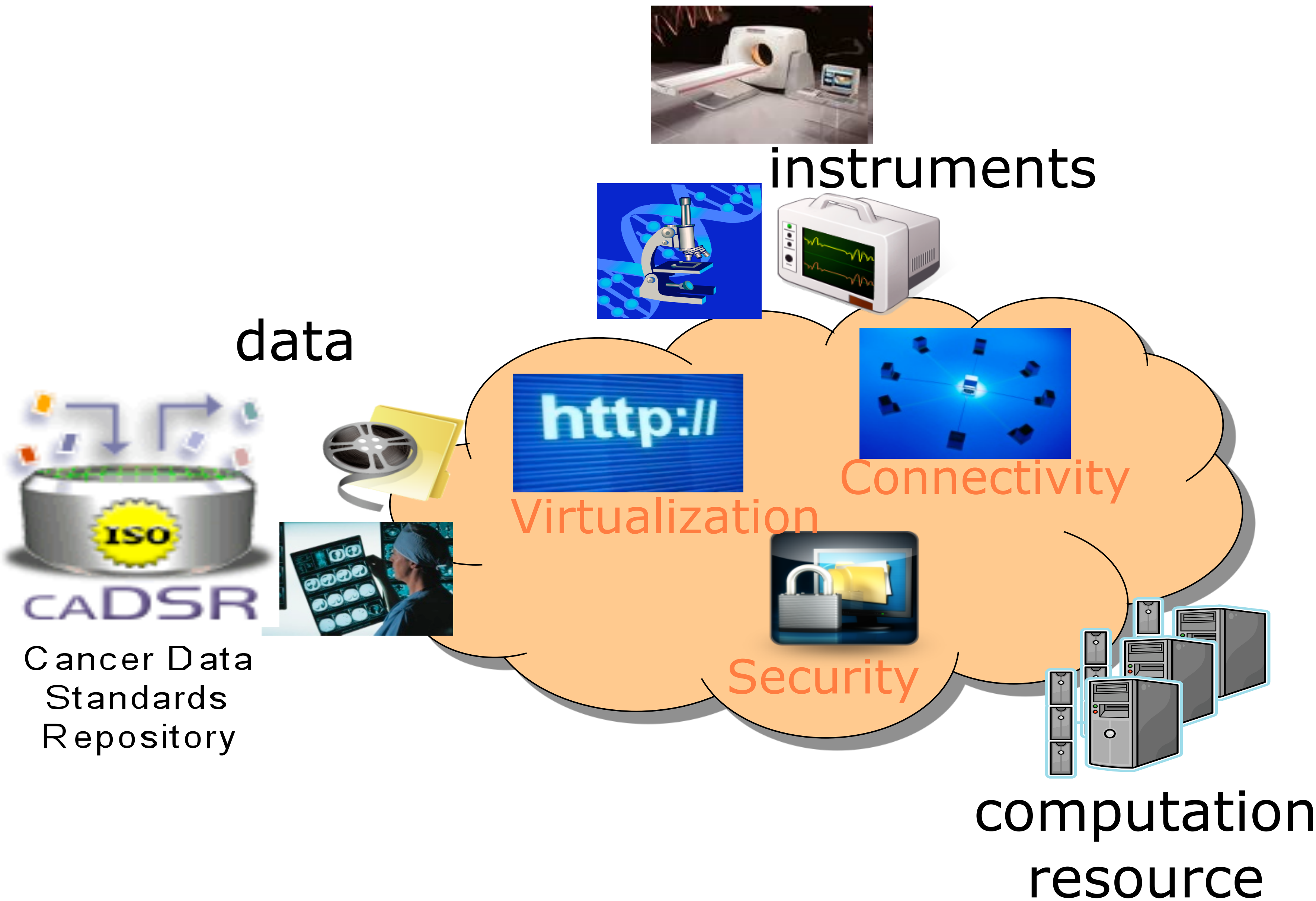


Permissions → Can I use it?



Reputation → The ultimate arbiter?

caGrid

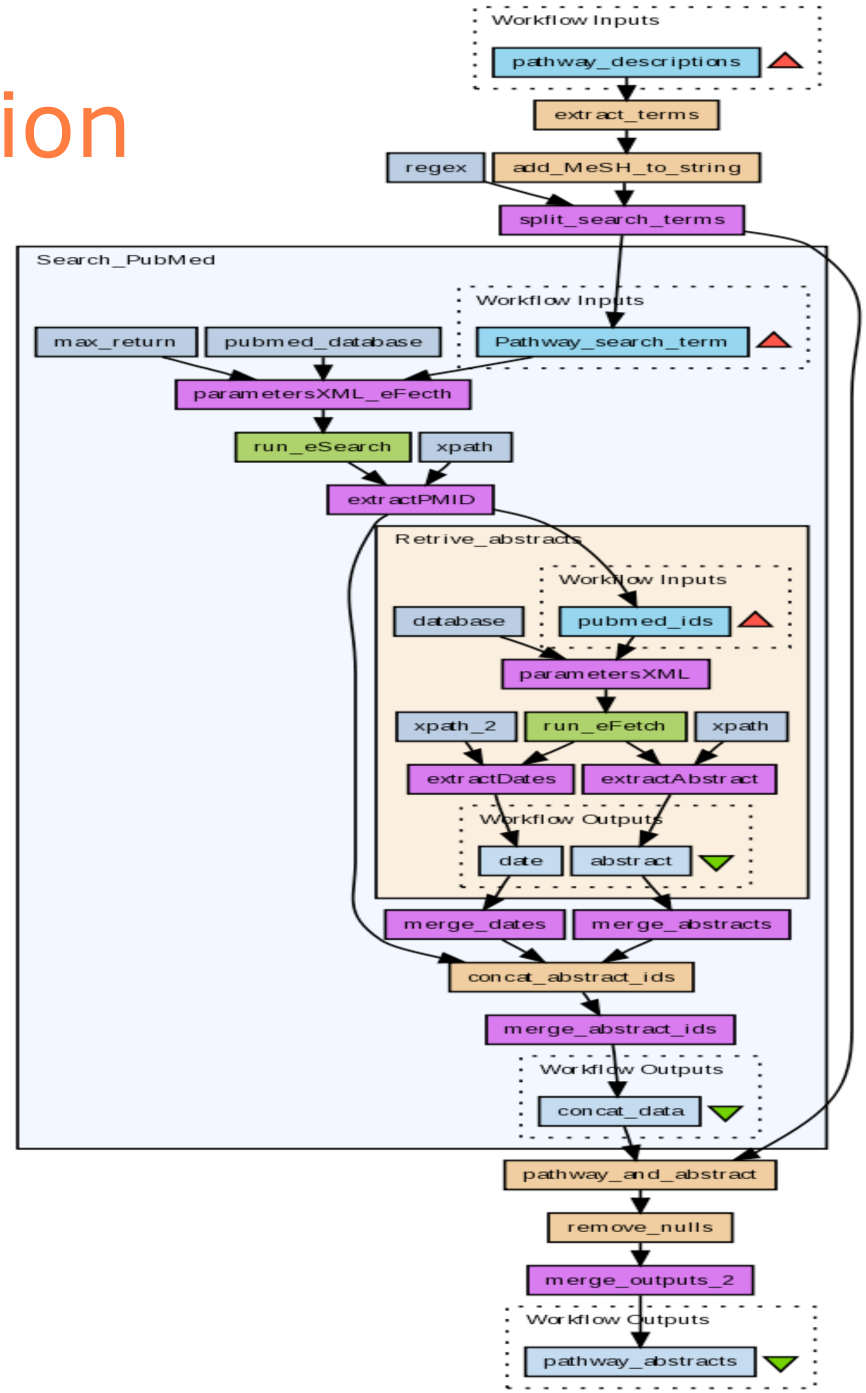

instruments
data
http://
Virtualization
Connectivity
Security
ISO
caDSR
Cancer Data
Standards
Repository
computation
resource

Service composition

Workflows are becoming a widespread mechanism for **coordinating** the execution of scientific services and **linking** scientific resources

Analytical and data processing pipelines

Industrialised Science
Data-intensive Science
Process-intensive Science



Slide: Carole Goble

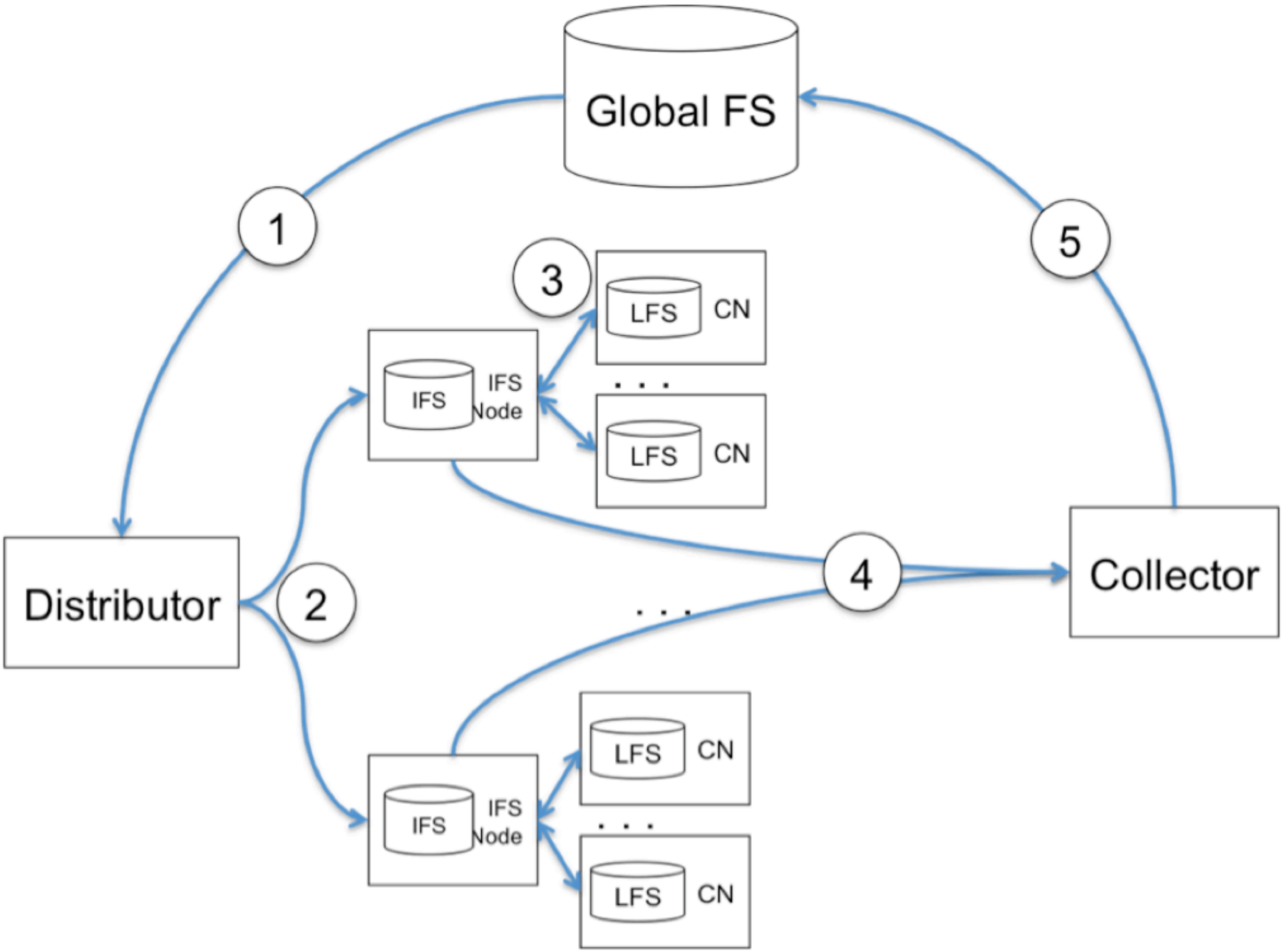
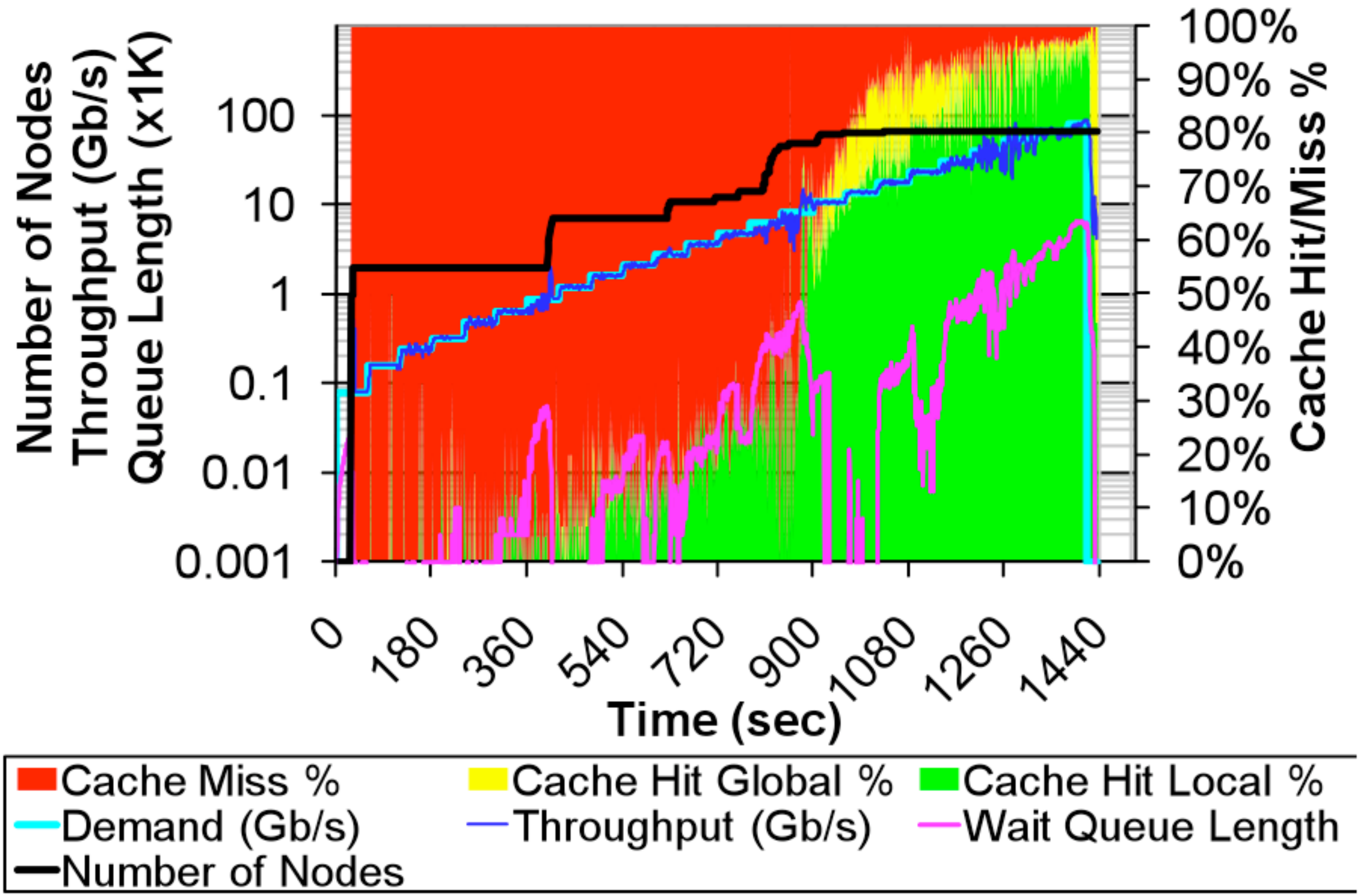


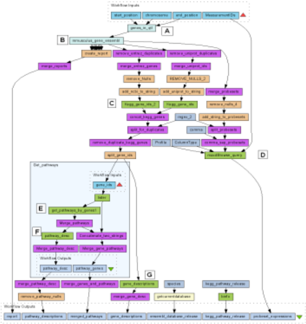
Figure 1: Logical Distributor/Collector Design



"MI" workload, 250K tasks, 10MB:10ms ratio, up to 64 nodes using DRP, GCC policy, 2GB caches/node

Reuse story that really happened

- Paul Fisher writes workflows for identifying biological pathways implicated in resistance to Trypanosomiasis in cattle
- Jo Pennock is investigating Whipworm in mouse
- Jo reuses one of Paul’s workflows without change
- Jo identifies the biological pathways involved in sex dependence in the mouse model, believed to be involved in the ability of mice to expel the parasite
- Previously a manual two year study by Jo had failed to do this



Slide: Carole Goble