SystemsX.ch
The Swiss Initiative in Systems Biology

SyBIT
SystemsX.ch
Biology IT

Enabling eScience

Challenges of Supporting the Digital Scientist

About Myself

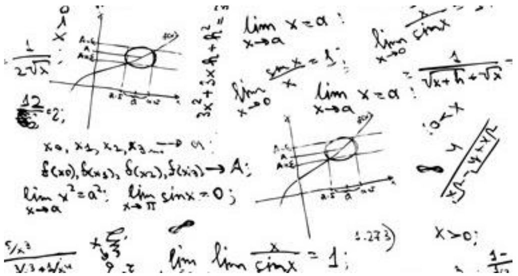
- PhD in theoretical physics – Bern
- Built large astrophysics science database – Sloan Digital Sky Survey, Johns Hopkins (3y)
- Headed data management tool development for LHC – CERN (5y)
- Built Swiss Tier2 at CSCS for LHC – CSCS (3y)
- Now leading SyBIT (5y)

Goals of Talk

- eScience – characteristics
  - What is eScience
  - Who is an eScientist
- eScience – 2 examples
  - Astronomy / Astrophysics
  - Life Sciences
- Infrastructure for eScience
- Challenges

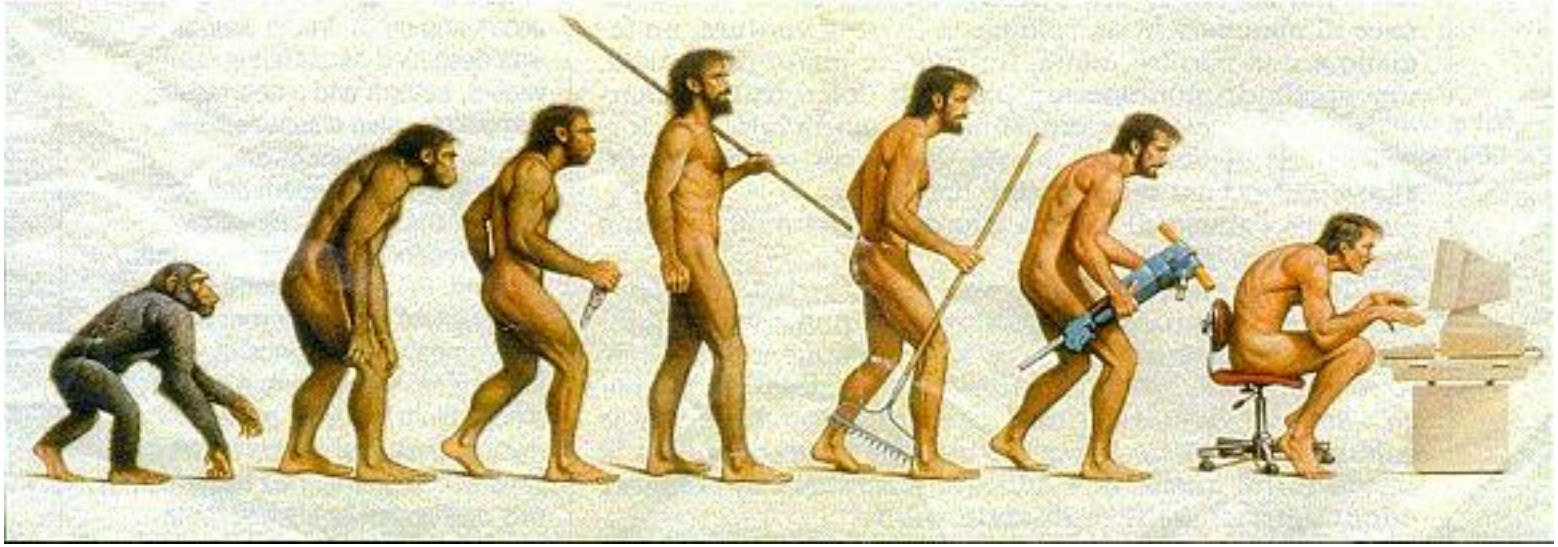
Evolution of Science

- 1000 years ago:
  - **Empirical Observation**
- Few 100 years ago:
  - **Theory – Experiment Cycle**
- Since 50 years:
  - **Computer Simulation of Complex Phenomena**
- Now added:
  - **Data Exploration: eScience**



Science Today

- Every measurement, experiment, simulation, etc produces **digital data**
- **eScience = Extract knowledge**



eScience : Unify it all

- Theory, Experiment, Simulation
- Data captured by instruments
- Data generated by simulation
- Data Processed by Software Pipelines
- Previous Information/Knowledge from Digital Libraries

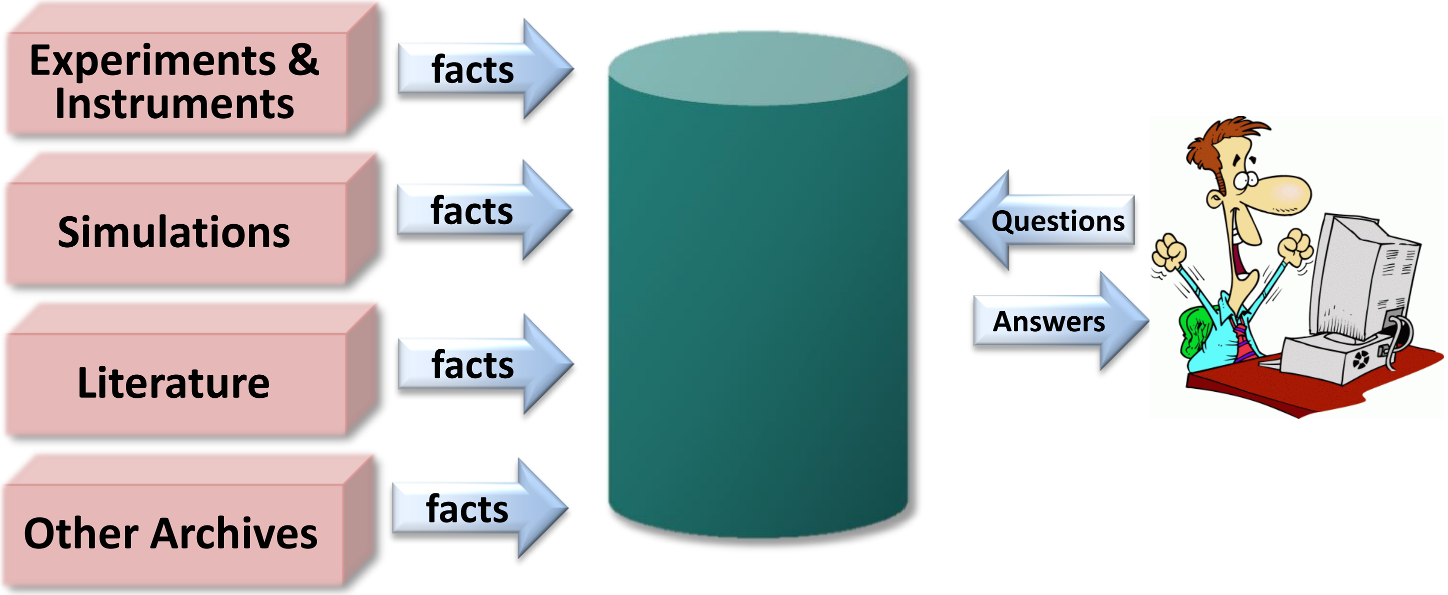
eScience Methods

- Data Collection
  - Digital Instrumentation
- Simulation
  - HPC, distributed computing, etc
- Data Management
  - Collect data so that it is found again
  - Make sure the data is available for reuse
- Data Analysis
  - Statistical analysis, mining, visualization
- Data Curation
  - Description, contextualization, annotation
  - Publication

DOMAIN-informatics and computational DOMAIN

Each discipline is evolving its own computational research and informatics (tooling and engineering)

How to codify and represent our knowledge is DOMAIN SPECIFIC



eScience People

- **Data Engineer / Data Manager**
  - Programmer / Database / NOSQL expert
  - Managing very large data volumes
  - Writing tools to collect data
  - Making sure data is coherent and valid
- **Algorithm Builder**
  - Programmer of simulations, workflow engineering
  - Models and algorithms
  - HPC or Grid specialist

eScience People

- **Data Analyst**
  - Build models based on data
  - Increasingly knowing the data
  - Ask right questions of the data
- **Data Steward**
  - Librarians, information specialists, ontologists
  - Make it discoverable, reusable, linked to other data
  - Well documented

+ Large Collaborations

- Increasing number of large / very large projects
- Dozens / hundreds of People
- Working together on the same project over years

Collaboration Examples

- Astronomy : Large Sky Surveys
- High Energy Physics : CERN
- Systems Biology : SystemsX.ch
- Neurobiology : Human Brain Project

- Every discipline has examples
  - Geography, Earth Observation
  - Medicine
  - Human medicine
  - History
  - …

Astronomy

- Used to be a single-person discipline
  - Data mining = the data is mine
  - No sharing of photographic plates

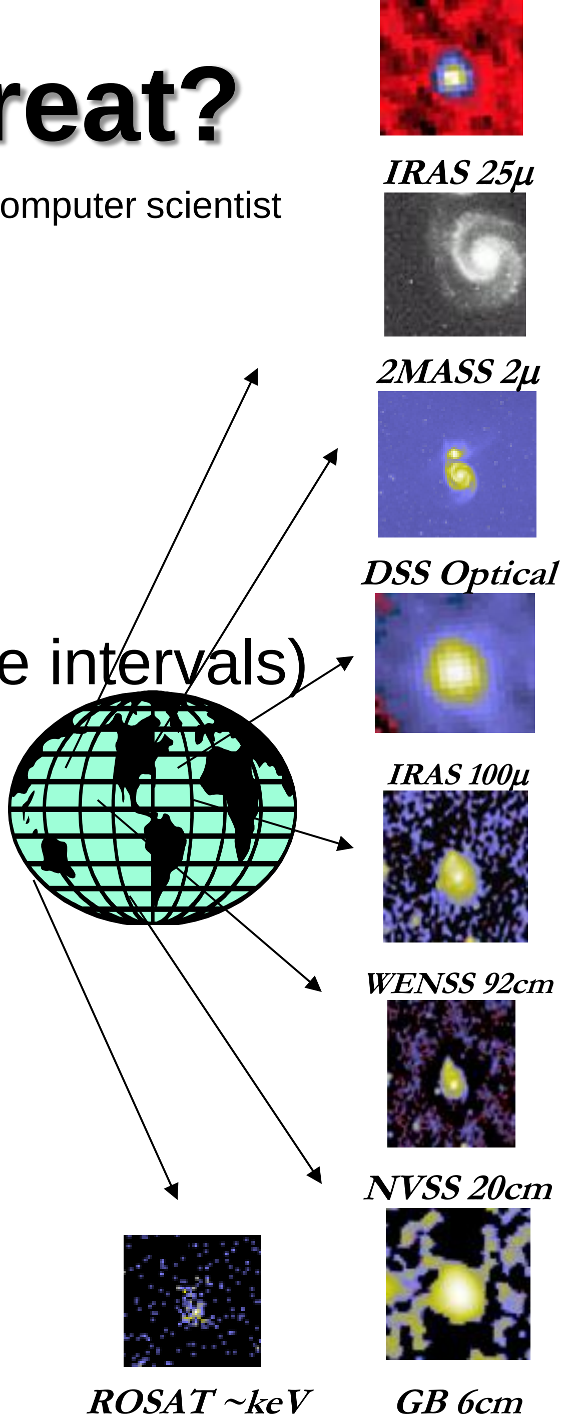
- But Large telescopes need a LOT of money
  - Large collaborations have emerged
  - Culture has completely changed

- Early engagement with computer science
  - Lots of method development
  - Lots of tooling
  - **Astronomers & Astrophysicists are good with computers**

Why is Astronomy Data great?

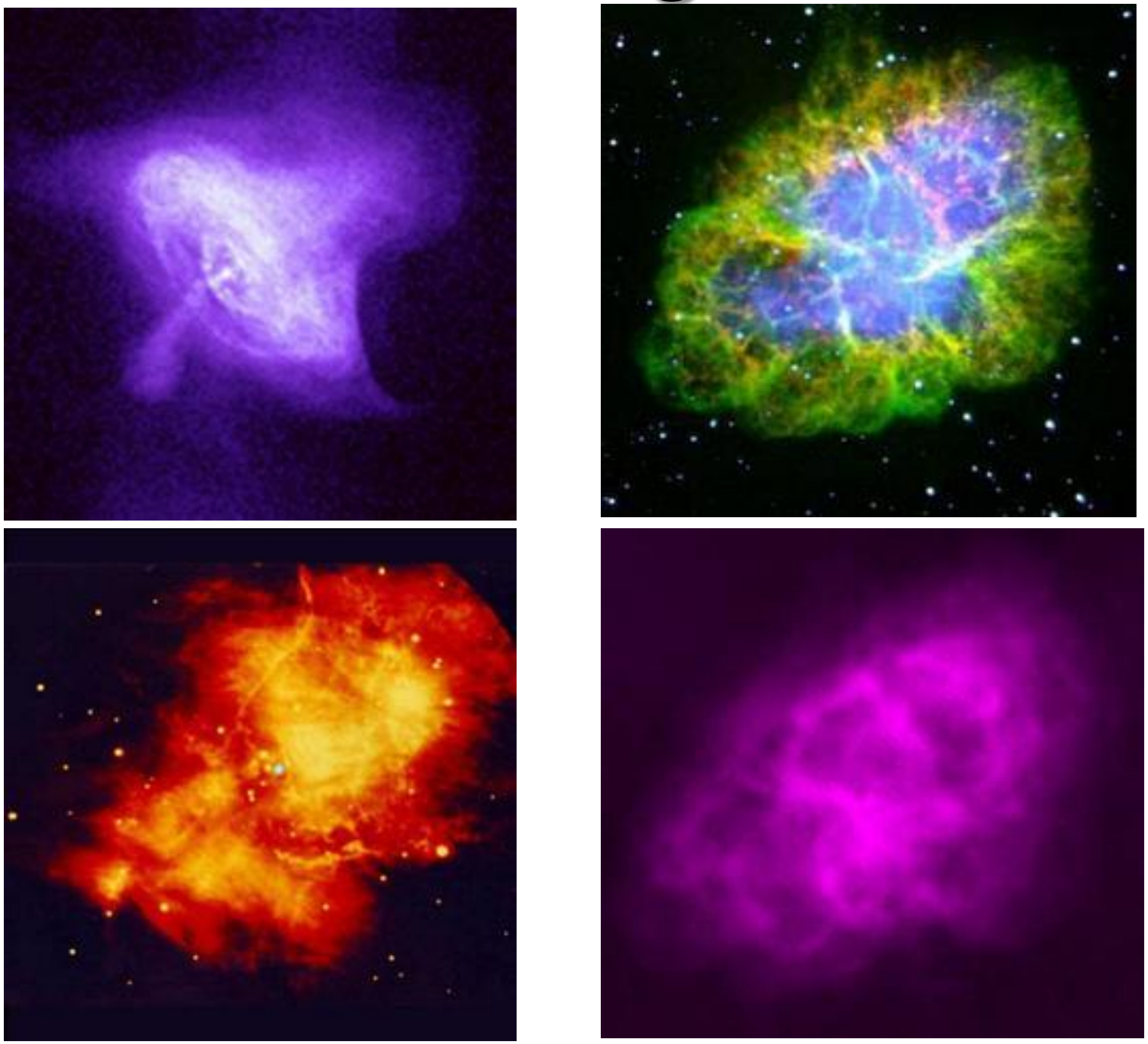
For a computer scientist

- **It has no commercial value**
  - No privacy concerns
  - Can freely share results with others
  - Great for experimenting with algorithms
- **It is real and well documented**
  - **High-dimensional data** (with confidence intervals)
  - **Spatial** data
  - **Temporal** data
- Many **different instruments** from many **different places** and many **different times**
- **Federation is a goal**
- There is a lot of it (petabytes)



Slide adapted from Jim Gray

Time and Spectral Dimensions
The Multiwavelength Crab Nebulae



Crab star
1053 AD

X-ray,

optical,

infrared, and

radio

views of the nearby Crab Nebula, which is now in a state of chaotic expansion after a supernova explosion first sighted in 1054 A.D. by Chinese Astronomers.

Slide courtesy of Robert Brunner @ CalTech.

SkyServer.SDSS.org

- A modern archive
  - Access to Sloan Digital Sky Survey Spectroscopic and Optical surveys
  - Raw Pixel data lives in file servers
  - Catalog data (derived objects) lives in Database
  - Online query to any and all
- Also used for education
  - Hundreds of hours of online Astro
  - Implicitly teaches data analysis
- Interesting things
  - Spatial data search
  - Client query interface via Java Applet
  - Query from Emacs, Python, ….
  - Cloned by other surveys (a template design)
  - Web services are core of it.

Welcome!!!

This website presents data from the Sloan Digital Sky Survey, a project to make a map of a large part of the universe. We would like to show you the beauty of the universe, and share with you our excitement as we build the largest map in the history of the world.

News

This site now contains data from the SDSS Data Release 1 (DR1).

More...

For Astronomers

A separate branch of this website for professional astronomers (English)

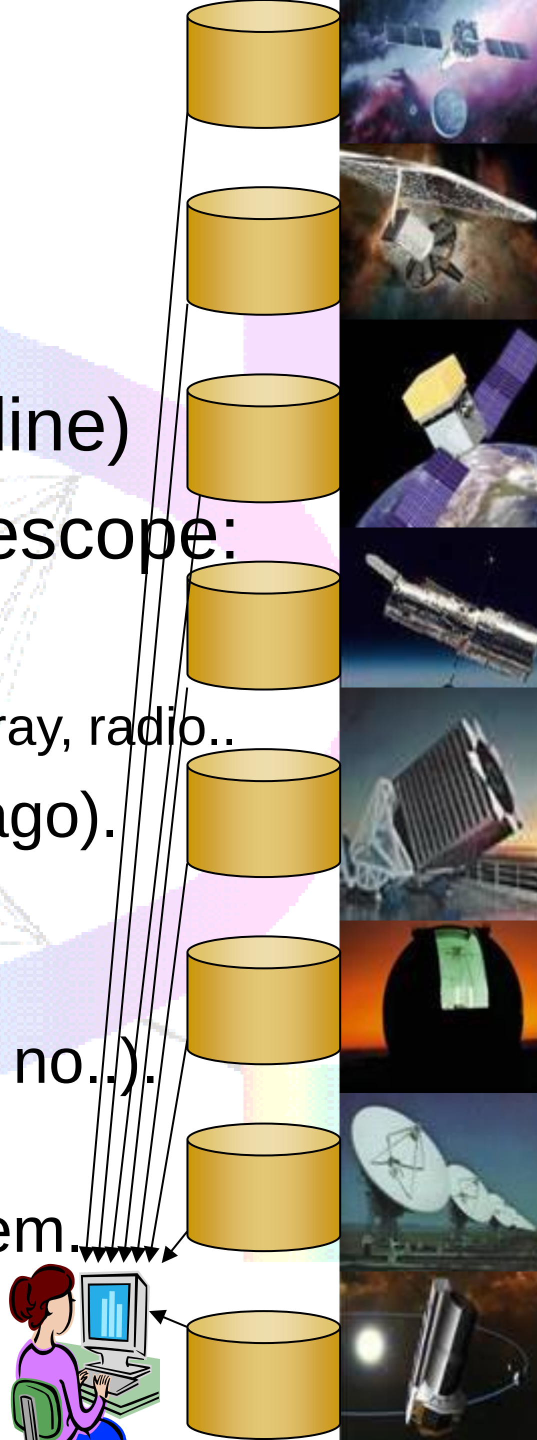
More...

Slide by Jim Gray

World Wide Telescope
Virtual Observatory

http://www.us-vo.org/ http://www.ivoa.net/

- Premise: Most data is (or could be online)
- So, the Internet is the world’s best telescope:
  - It has data on every part of the sky
  - In every measured spectral band: optical, x-ray, radio..
  - As deep as the best instruments (2 years ago).
  - It is up when you are up.
    The “seeing” is always great
    (no working at night, no clouds no moons no..).
  - It’s a smart telescope:
    links objects and data to literature on them.



Slide by Jim Gray

Characteristics of Astronomy

- Data is well known
- Schema is defined and optimized
- Indices are given
- Community has learned SQL!
  - It is faster to query the archive than to write a computer program

Astronomy is an eScience Discipline today

- Catalogs are omnipresent
- New projects are planned with data curation and integration in mind
- Communities are well organized
- Standards have been worked out and are in use
- People are being educated in these tools and methods, like SQL

How to get there?

1. Endorse domain as Data Intensive Science.
2. Plan computing infrastructure to be inherently distributed: “scale-out” architecture.
3. Bring computations to the data, rather than data to the computations wherever possible.
4. Design system with concrete questions to the data. Do not go overboard with what-if scenarios.
5. Plan for change, use an iterative, agile approach to everything.
6. Automation of everything possible

New Challenges

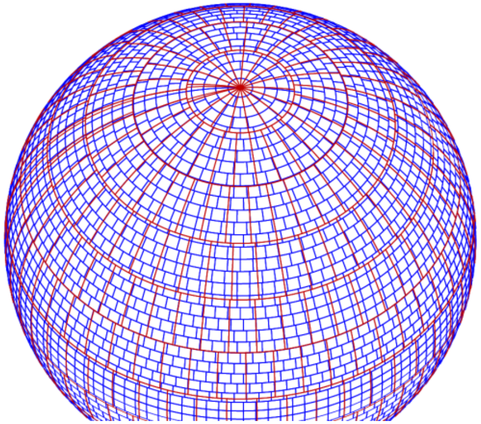
1. Needle in the Haystack, how to distinguish artifacts from real new objects / lost in automation
2. New instruments producing even more data, like LSST: full sky survey every 3 days
   - Monitor changing objects
   - Store light-curves for objects
   - Will start operations 2021
   - Hundreds of Petabytes of data

Address Challenge 1:

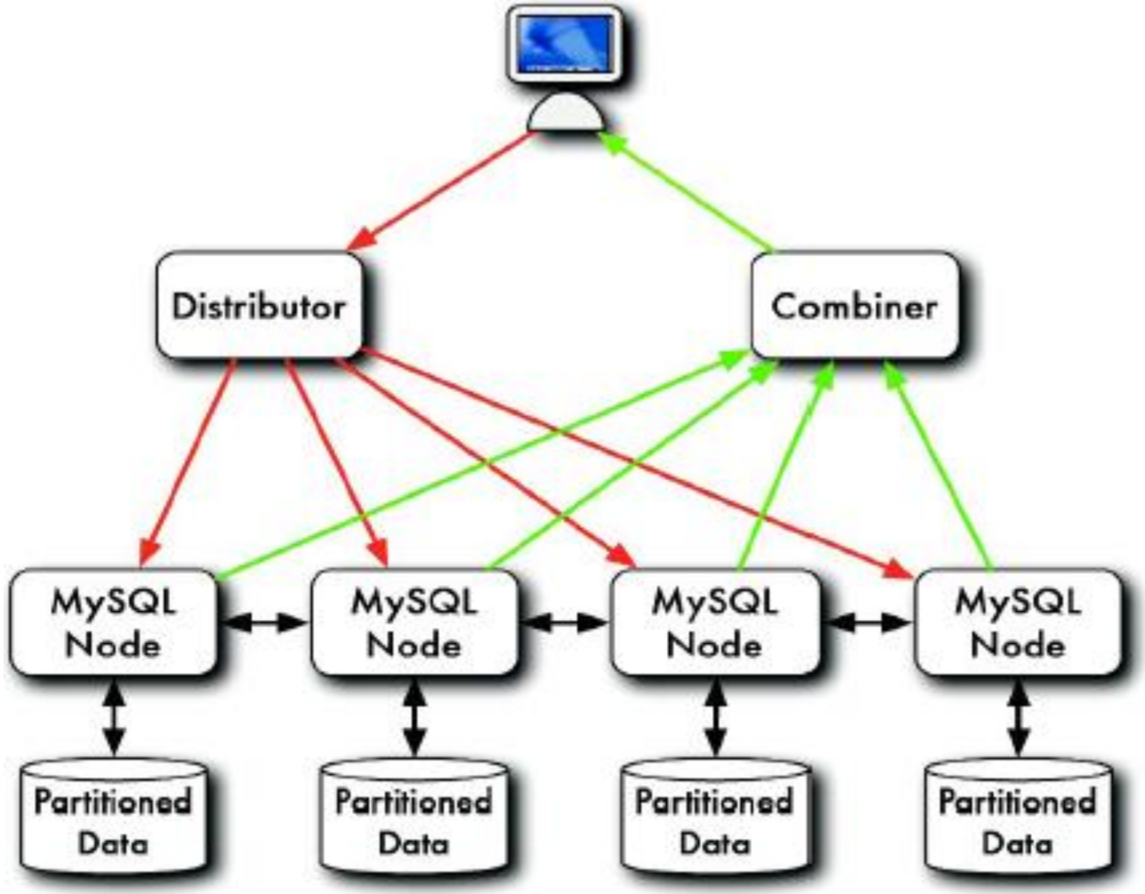
- Citizen Science site to classify Galaxies by looking at them www.galaxyzoo.org
- Data from SDSS and Hubble
- Thousands of users
- Several interesting objects have been found, leading to new science!

Addressing Challenge 2: LSST Database design

- Partitioning: distributed database architecture
- Shared-nothing data distribution
- Sharding of data and sharding of queries

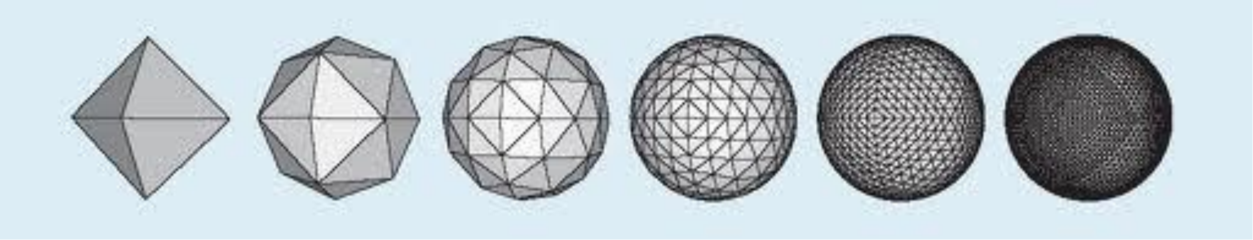


LSST Database design



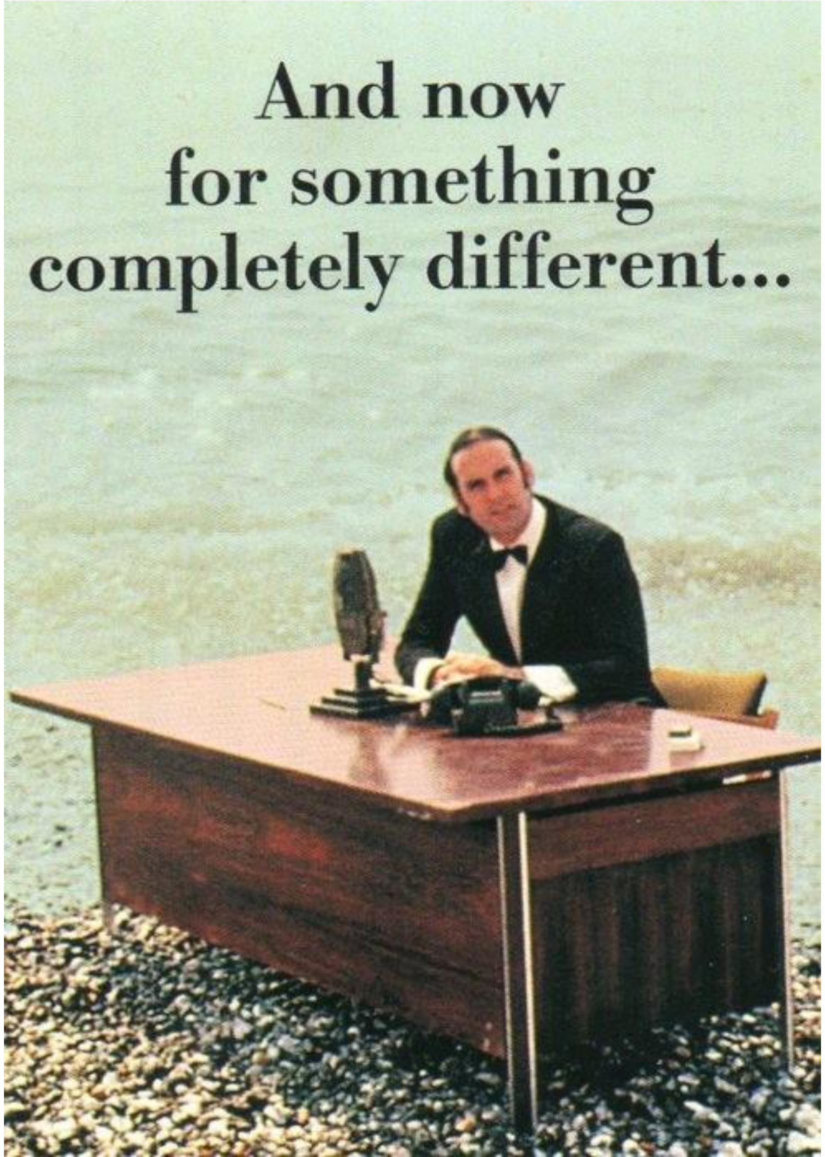
https://dev.lsstcorp.org/trac/wiki/db

LSST Database design



- Clustering of nearby objects using hierarchical triangular mesh
- Shared scanning – full table scans rather than indexed scans if more than just a few % of the data are being returned
  - Takes care of special cases where every object needs to be touched
  - Results to be stored in ‘myDB’

Systems Biology



Systems Biology

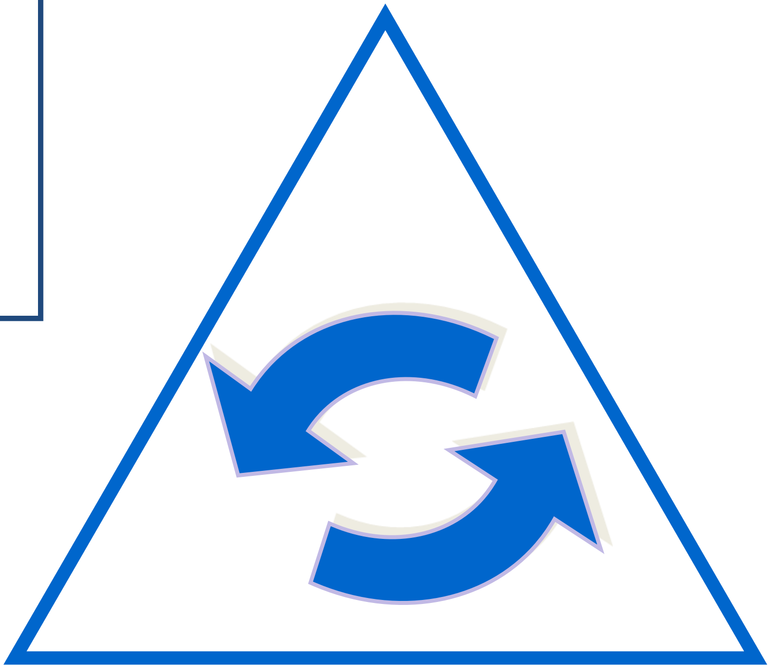
As understood by a physicist

- OK, so we have the Gene sequence (Genome) of many organisms, but what does it say? How does it all work?
- Cell Biology and Molecular Biology are producing more and more high-resolution and high-quality data to answer this question
- Bottom-up approach, understanding the cell from first principles is very difficult.
- Systems Biology Approach: Understand the available data top-down. Study the complex interaction of many levels of biological information to understand how they work together.

The Systems Biology Approach

Systems Biology:
- Biology
- Physics
- Chemistry
- Computer Science
- Mathematics
- Engineering
- Medicine

Biological System;
Qualitative,
Wet Lab Biology



Quantitative
Data sets of
Complete Systems;
Bio-Engineering

Network **Theory**;
Modeling,
Simulation

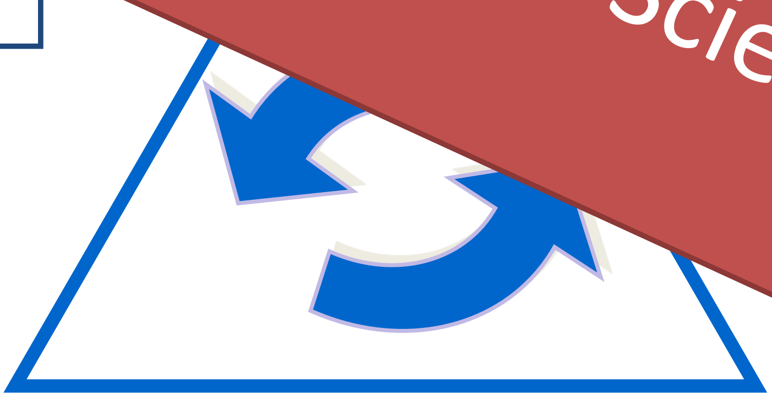
The Systems Biology Approach

SyBIT
SystemsX.ch
Biology IT

Systems B
- Biology
- Physic
- Chem
- Compu
- Mathematic
- Engineering
- Medicine

This is new for Life Sciences!

iological System;
tive,
logy



Quantitative
Data sets of
Complete Systems;
Bio-Engineering

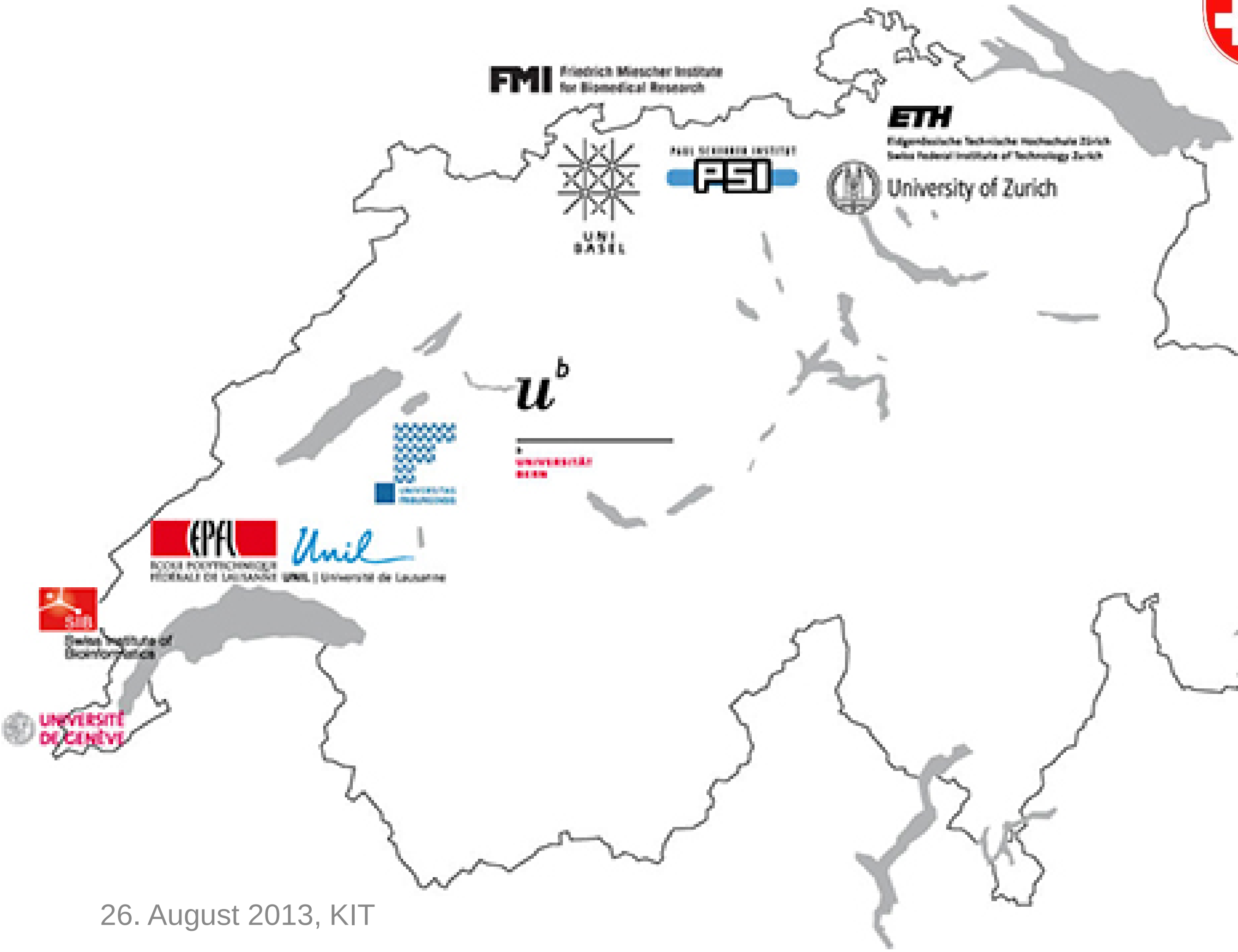
heory;
g,
Sim ion

Characteristics of Systems Biology

- Data is NOT well known
- Schema is NOT defined and optimized
- A lot of data is being searched for connections to ‘see’ anything by chance
  - Clustering, PCA, networks
- What is the question again? Lots of exploration of the unknown.
- Community is less computer savvy
  - Programming is not a skill biologists learn
  - Disconnect between modelers and experimentalists

SystemsX.ch

Largest Swiss national research effort to date


FMI Friedrich Miescher Institute for Biomedical Research
UNI BASEL
PAUL SCHERRER INSTITUT
PSI
ETH Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich
University of Zurich
u^b
UNIVERSITÄT BERN
UNIVERSITÄT FRIBOURG
EPFL
ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE
Unil
UNIL | Université de Lausanne
SIB
Swiss Institute of Bioinformatics
UNIVERSITÉ DE GENÈVE

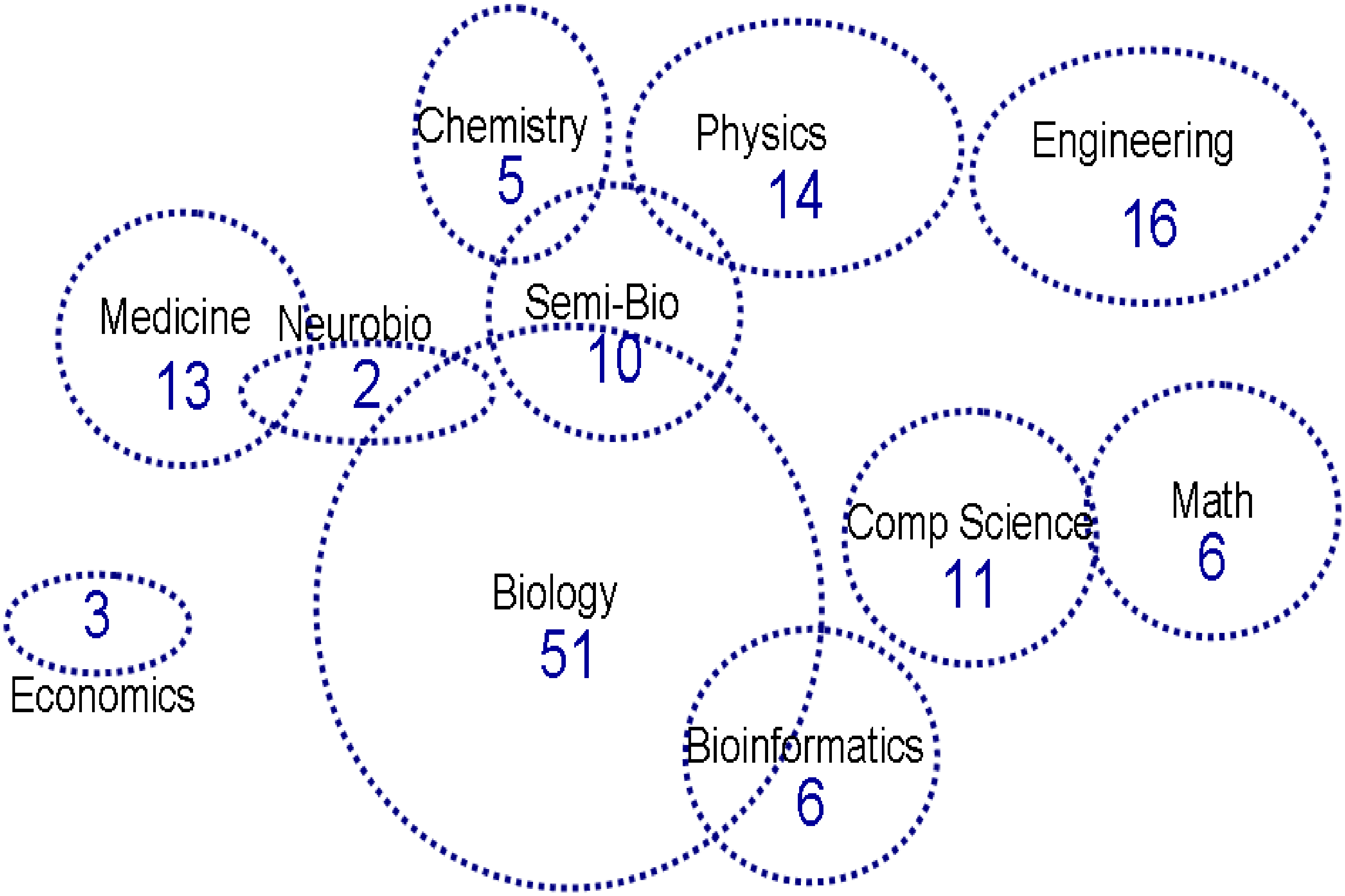
SCHWEIZERISCHE EIDGENOSSENSCHAFT
CONFÉDÉRATION SUISSE
CONFEDERAZIONE SVIZZERA
CONFEDERAZIUN SVIZRA

FNSNF
FONDS NATIONAL SUISSE
SCHWEIZERISCHER NATIONALFONDS
FONDO NAZIONALE SVIZZERO
SWISS NATIONAL SCIENCE FOUNDATION

ETH-RAT

SUK·CUS
Schweizerische Universitätskonferenz

Interdisciplinary Research



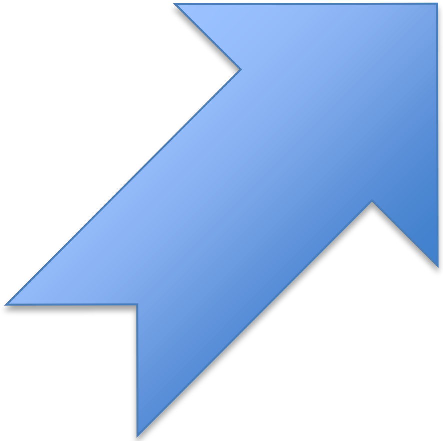
137 research groups

SystemsX.ch Projects

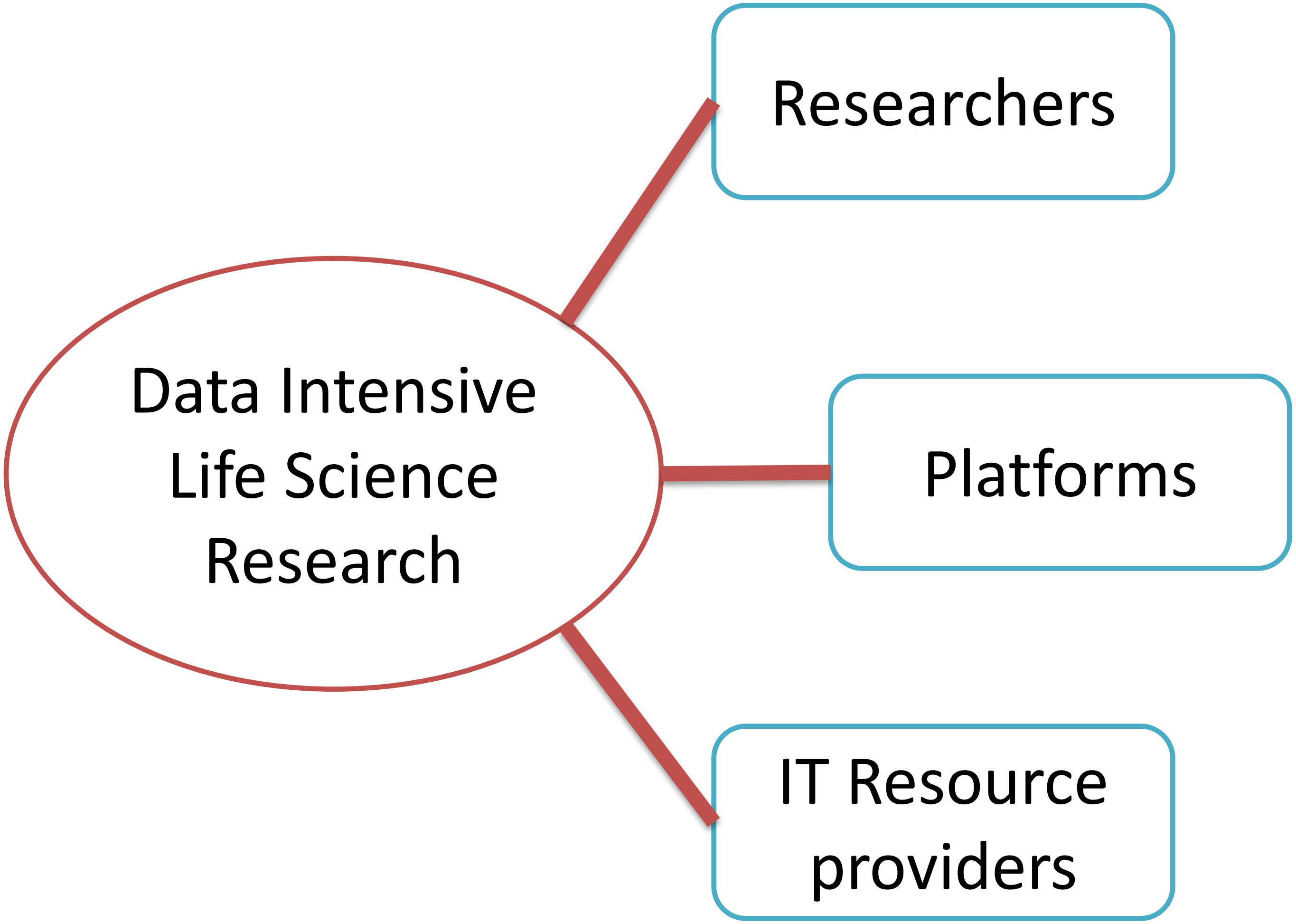
- Research Technology Development projects: large collaborations
  - 8 large projects have concluded
  - 17 ongoing projects
  - 10 more to start in 2014
- Industry collaboration projects – 10+
- Pilot ‘high risk’ projects – 10+
- Student, PhD projects – dozens

Challenge : Scale Up

- High Throughput Instruments
  - Much larger **data volumes**
  - Increased **data complexity**
- Large Collaborations
  - More people
  - More experiments and measurements



SystemsX.ch is Data Intensive

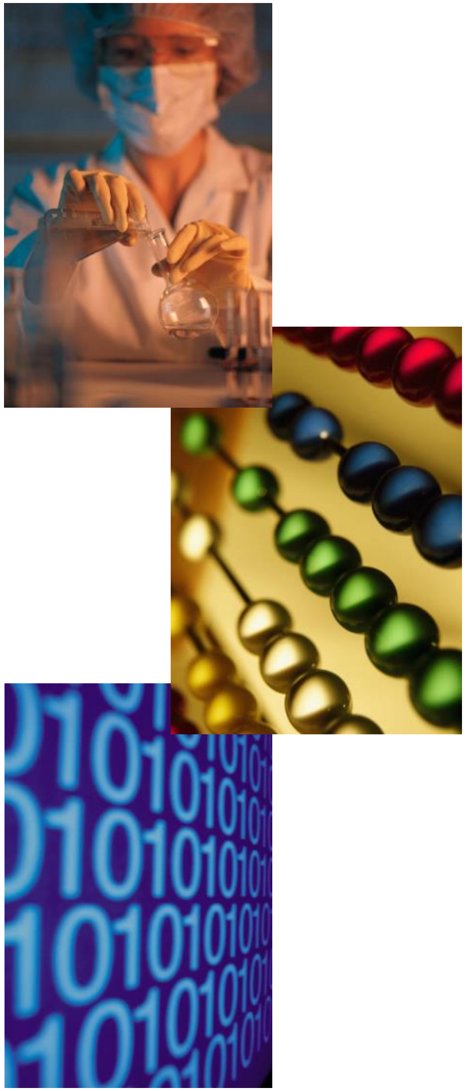


Data Intensive Life Science Research

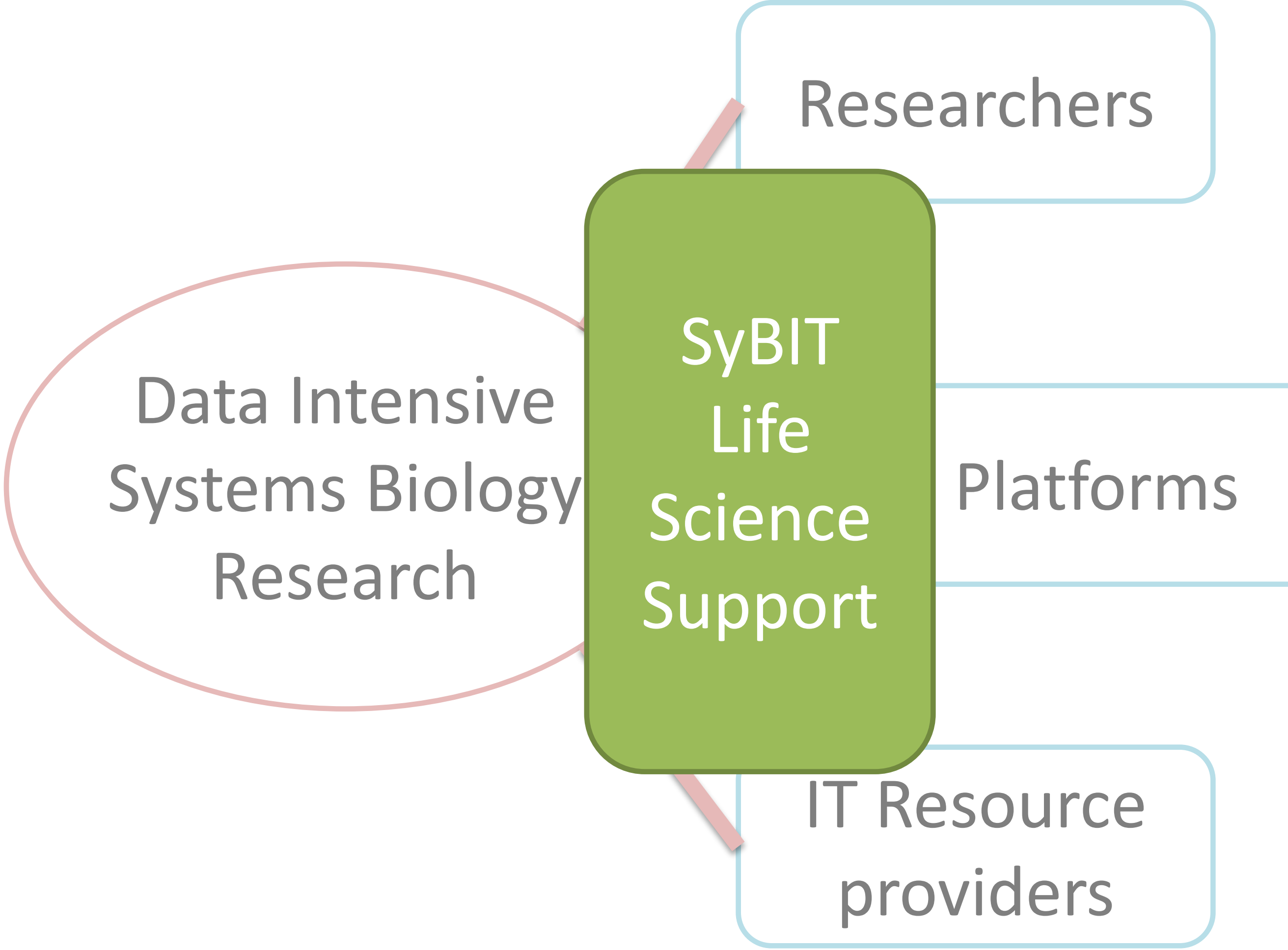
Researchers

Platforms

IT Resource providers



SyBIT: Build the Distributed Infrastructure and Tools



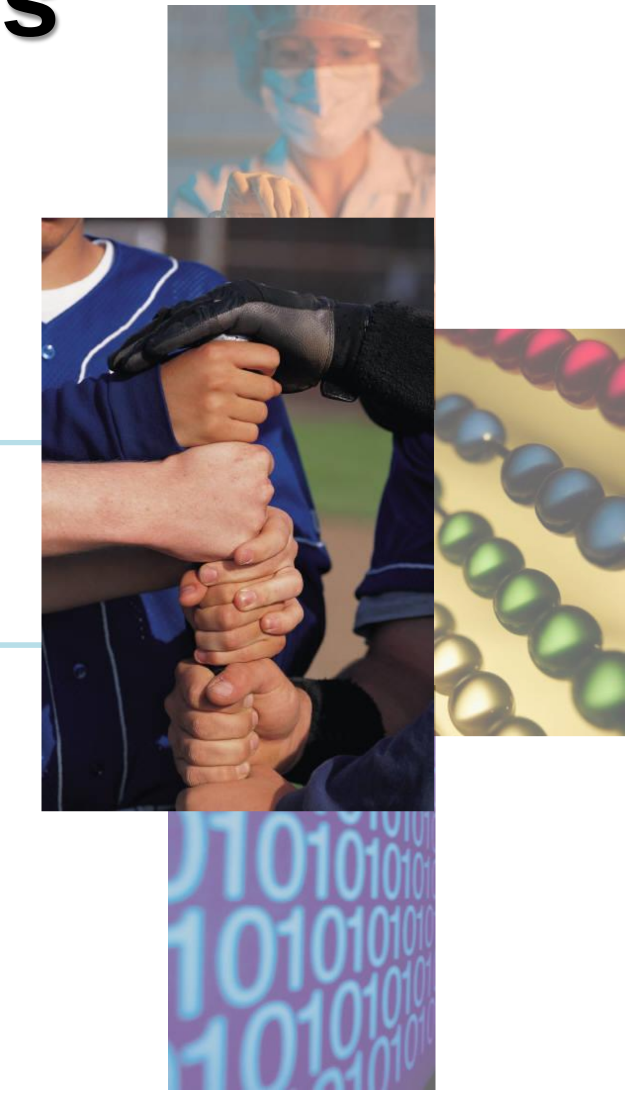
Data Intensive Systems Biology Research

SyBIT Life Science Support

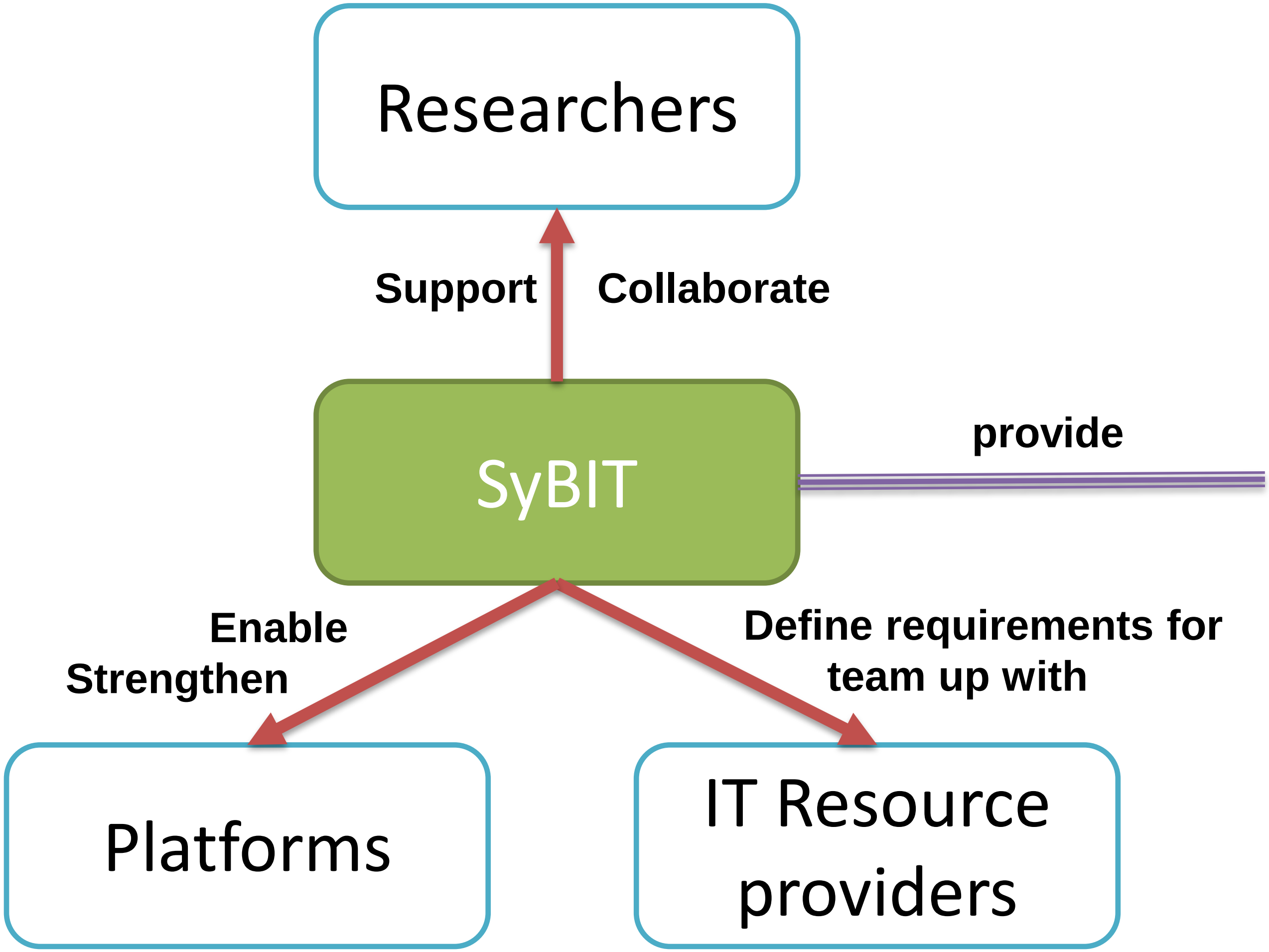
Researchers

Platforms

IT Resource providers



Fill the Gap



Researchers

Support Collaborate

SyBIT

provide

Enable
Strengthen

Define requirements for
team up with

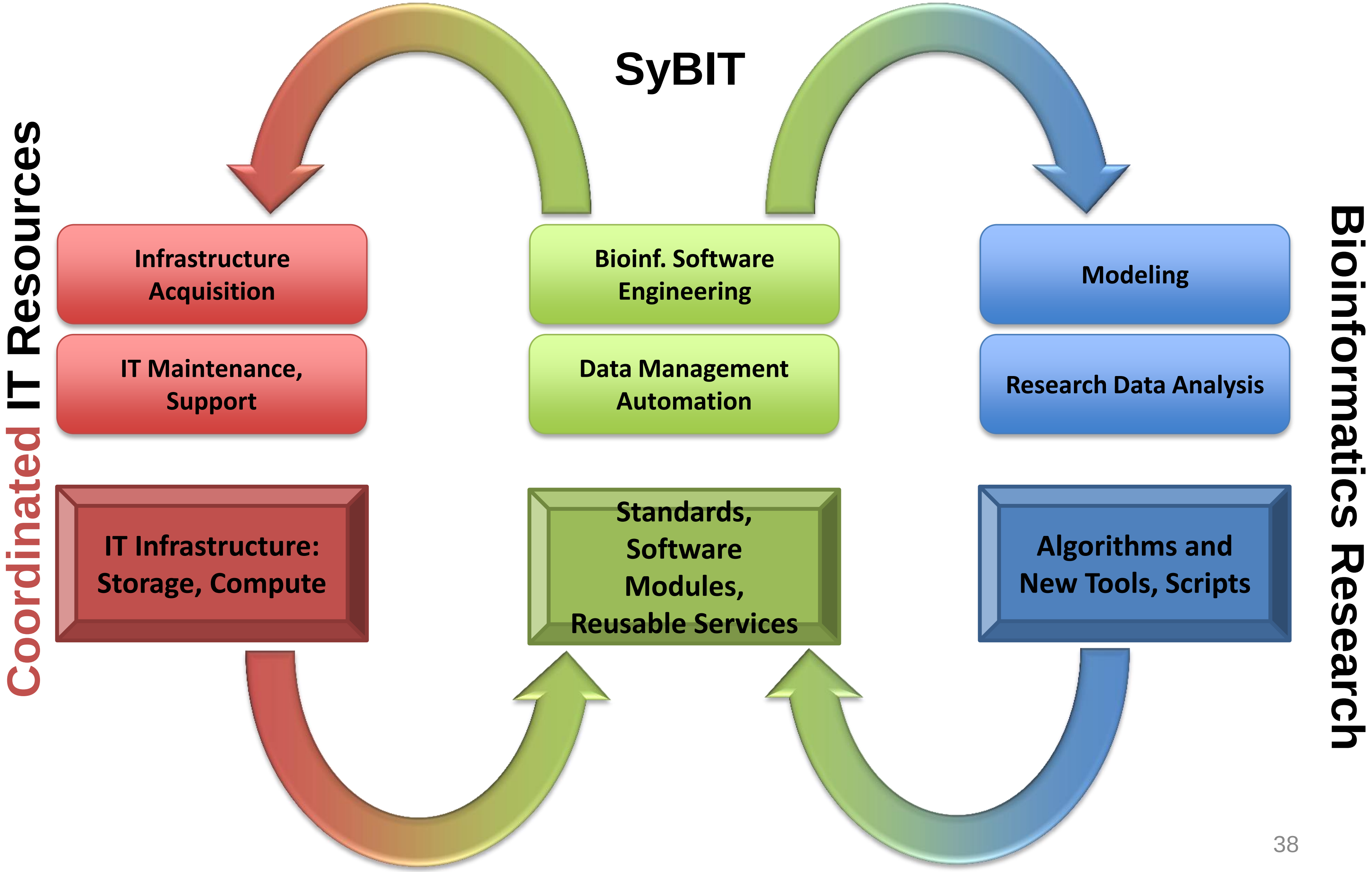
Platforms

IT Resource providers

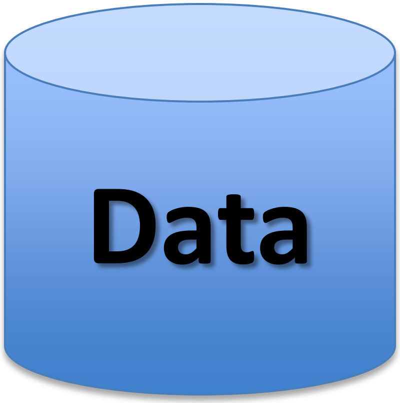
Tools
Bioinformatics
Services
Data Management
Standards, Integration
Software engineering
Legacy support
Documentation
Education
Collaboration support
Setup, Configuration
Versioning
Commercial software

Interactions in detail

SyBIT
SystemsX.ch
Biology IT

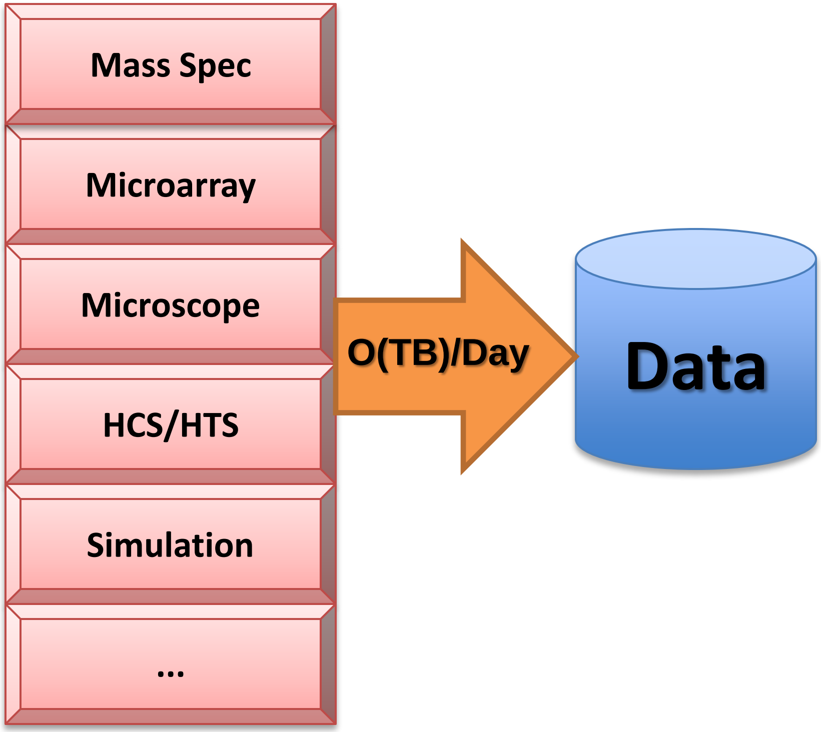

SyBIT
Coordinated IT Resources
Bioinformatics Research
Infrastructure Acquisition
IT Maintenance, Support
IT Infrastructure: Storage, Compute
Bioinf. Software Engineering
Data Management Automation
Standards, Software Modules, Reusable Services
Modeling
Research Data Analysis
Algorithms and New Tools, Scripts

Also here: Data at the Center of Work



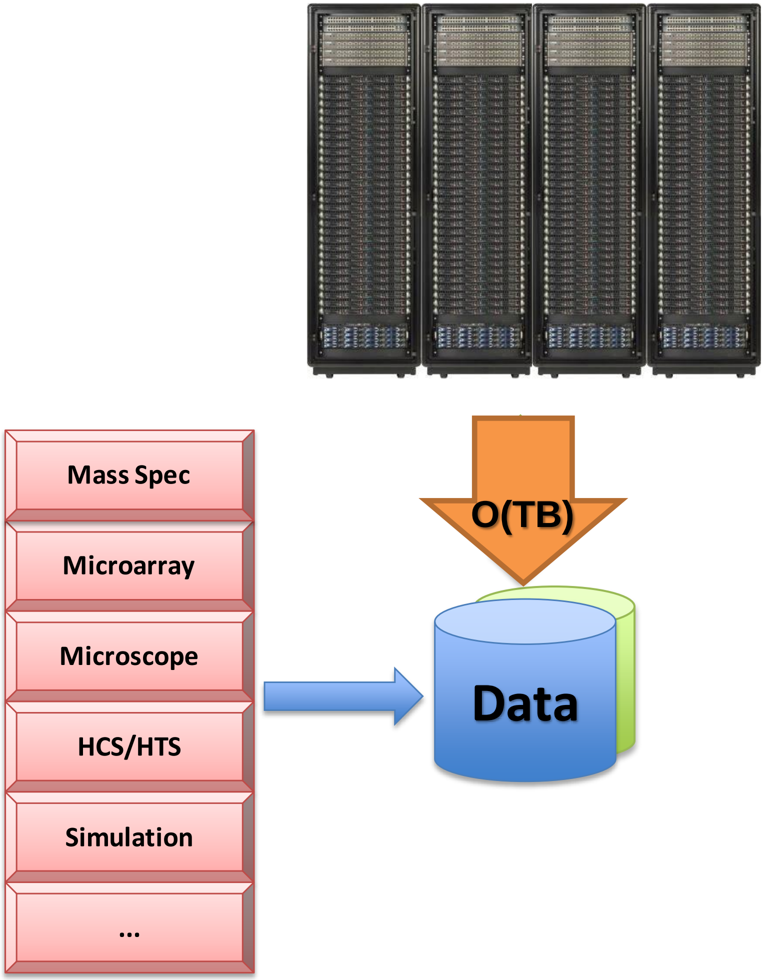
- Almost everything is data driven
- Many formats and access patterns

Data Production



- Many different kinds of instruments
- Different data types
- New instruments produce much more data
- Data volume increasing exponentially

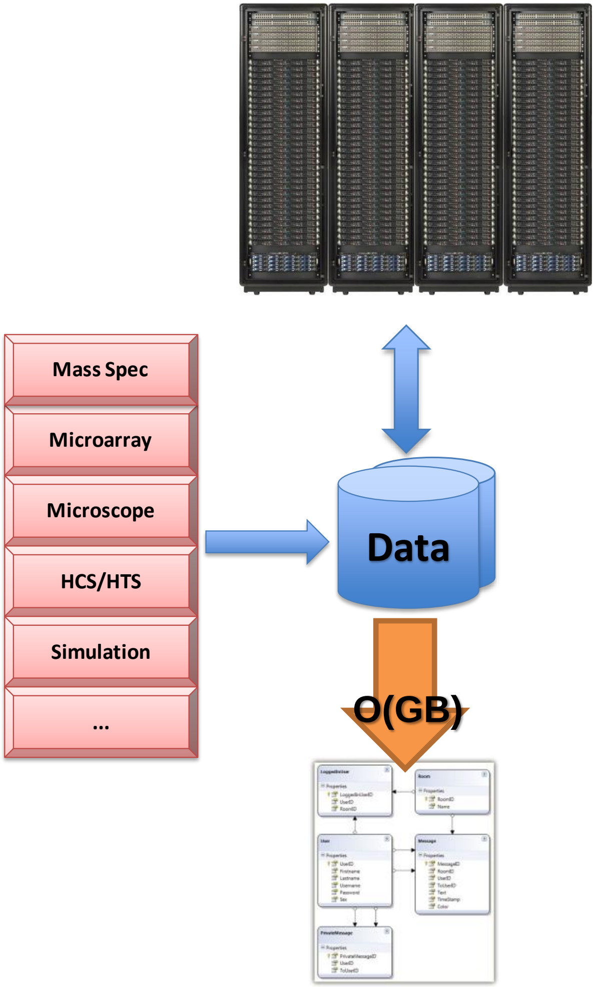
Data Validation and Filtering



- Validation, checks
- Conversion into standard formats
- Compression

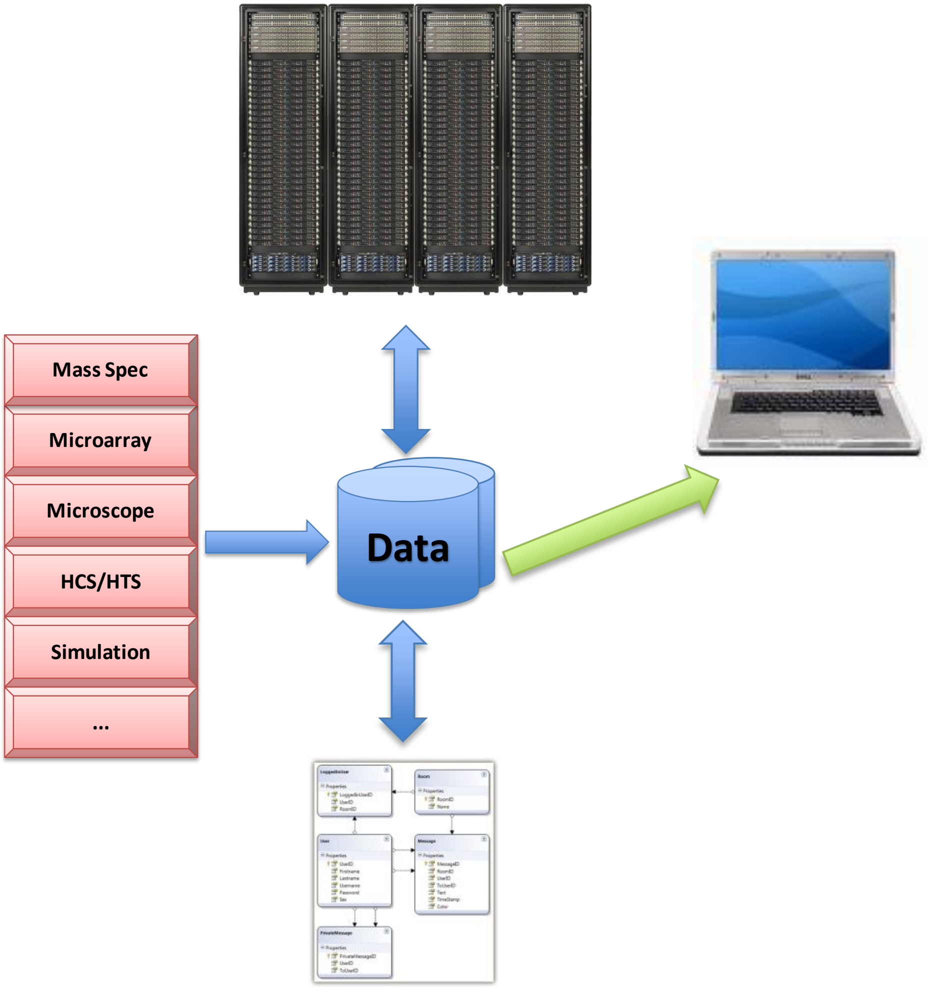
- This can be very compute intensive
- This can produce a lot of new data
- Needs clusters to do it in a timely fashion

Data Tracking and Metadata



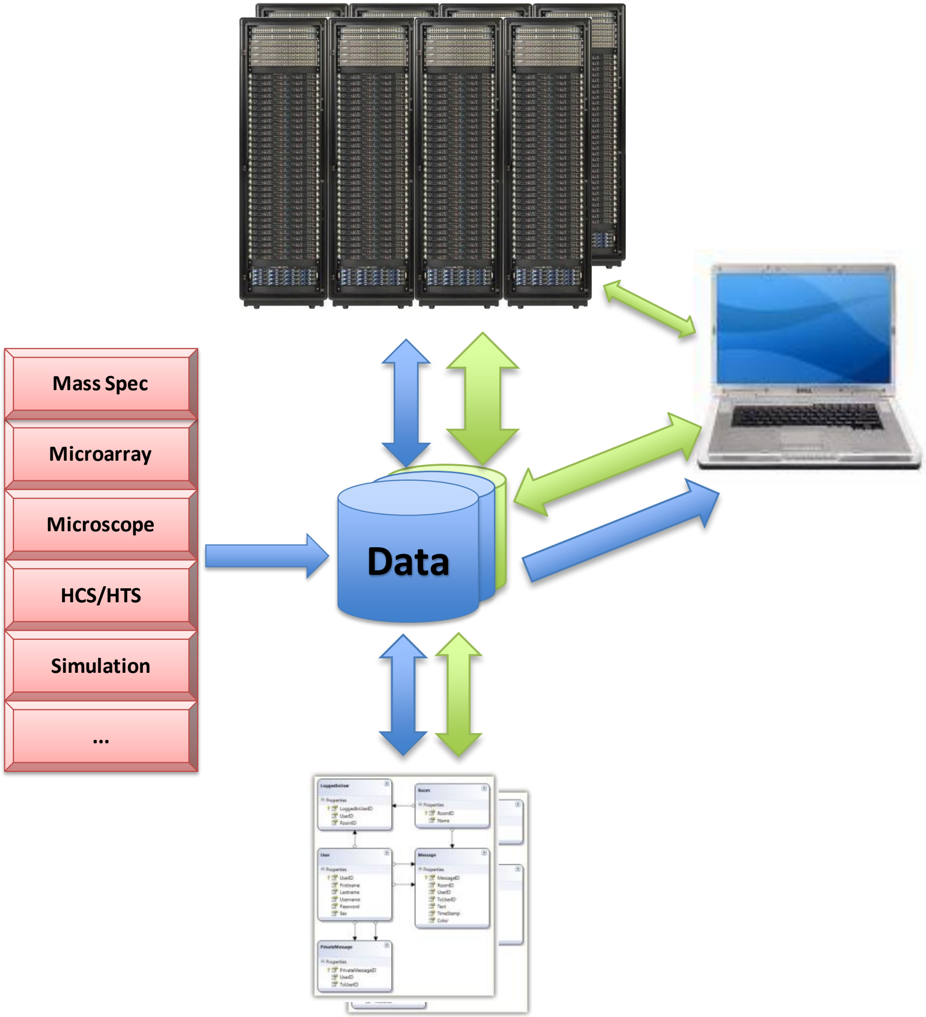
- Provenance metadata
- File catalogs
- Metadata on initial filtering

Data Exploration



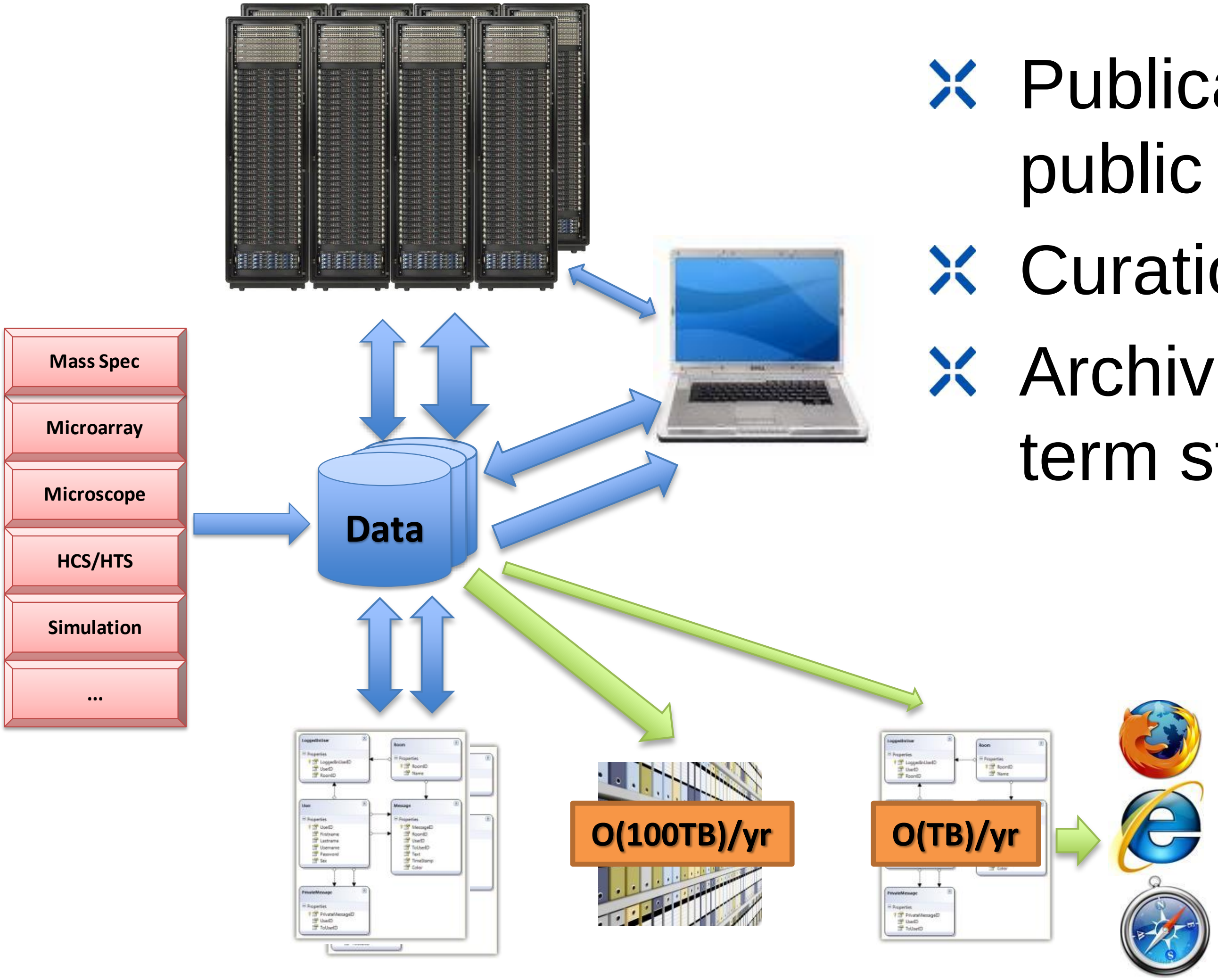
- Interactive exploration of data
- Small-scale analysis
- Planning of large-scale data analysis

Data Analysis and Modeling



- Large-scale analysis
  - On as much CPU power as possible
- Production of more data
  - Secondary datasets
  - Simulation, modeling
- Additional databases
  - Metadata
  - Result data

Publication and Archiving



Mass Spec

Microarray

Microscope

HCS/HTS

Simulation

...

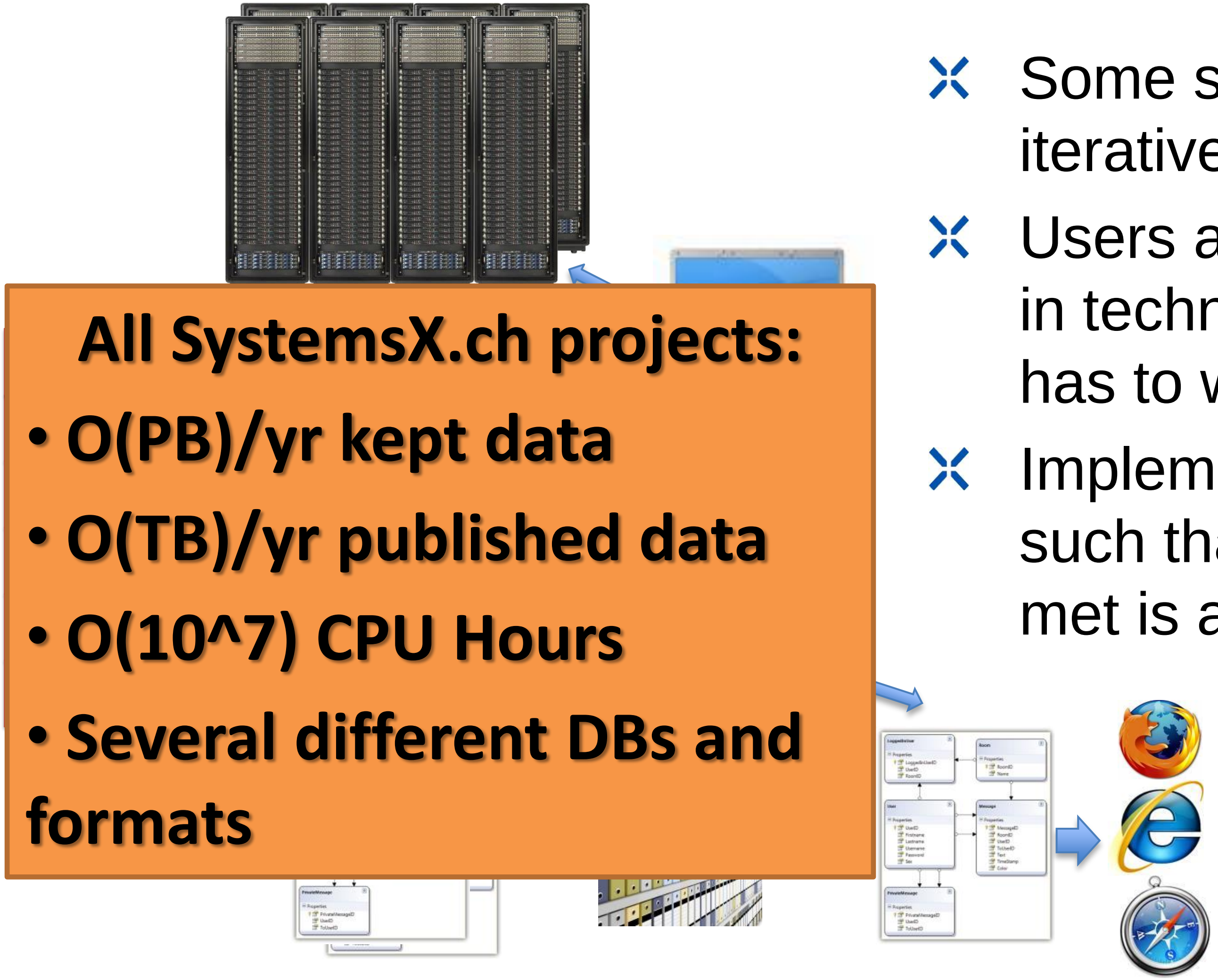
Data

O(100TB)/yr

O(TB)/yr

- Publication into public databases
- Curation
- Archiving for long-term storage

Data Lifecycle

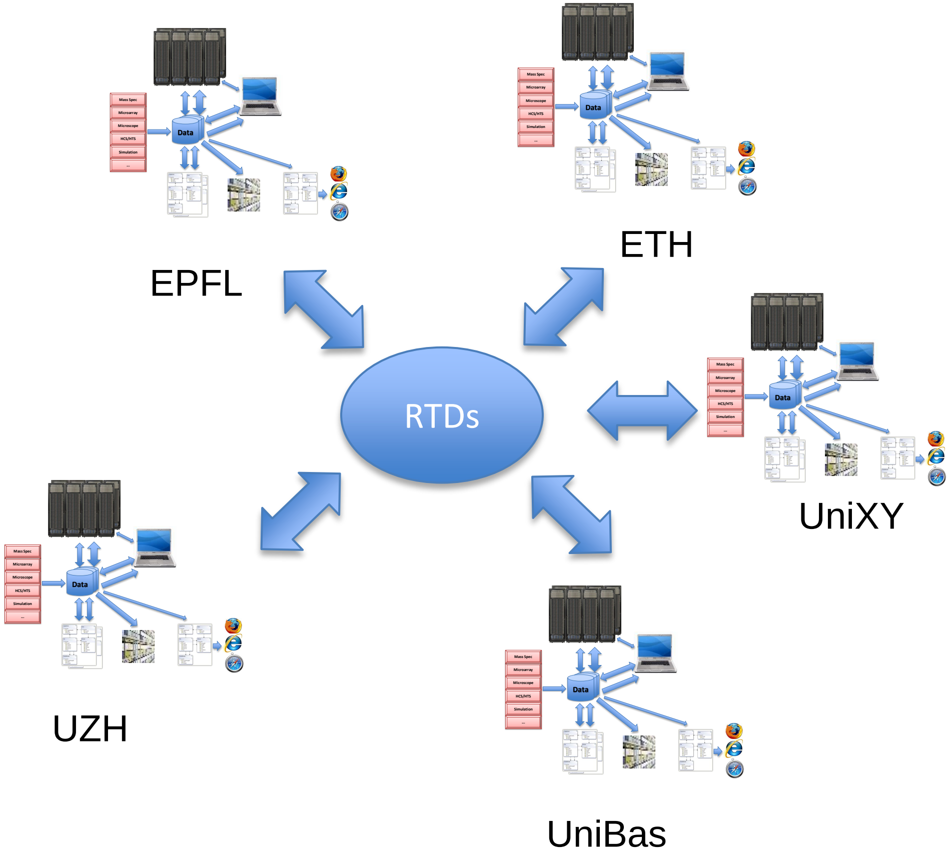


All SystemsX.ch projects:

- O(PB)/yr kept data
- O(TB)/yr published data
- O(10^7) CPU Hours
- Several different DBs and formats

- Some steps might be iterative
- Users are not interested in technology, it simply has to work
- Implementing ´Data´ such that all needs are met is a challenge

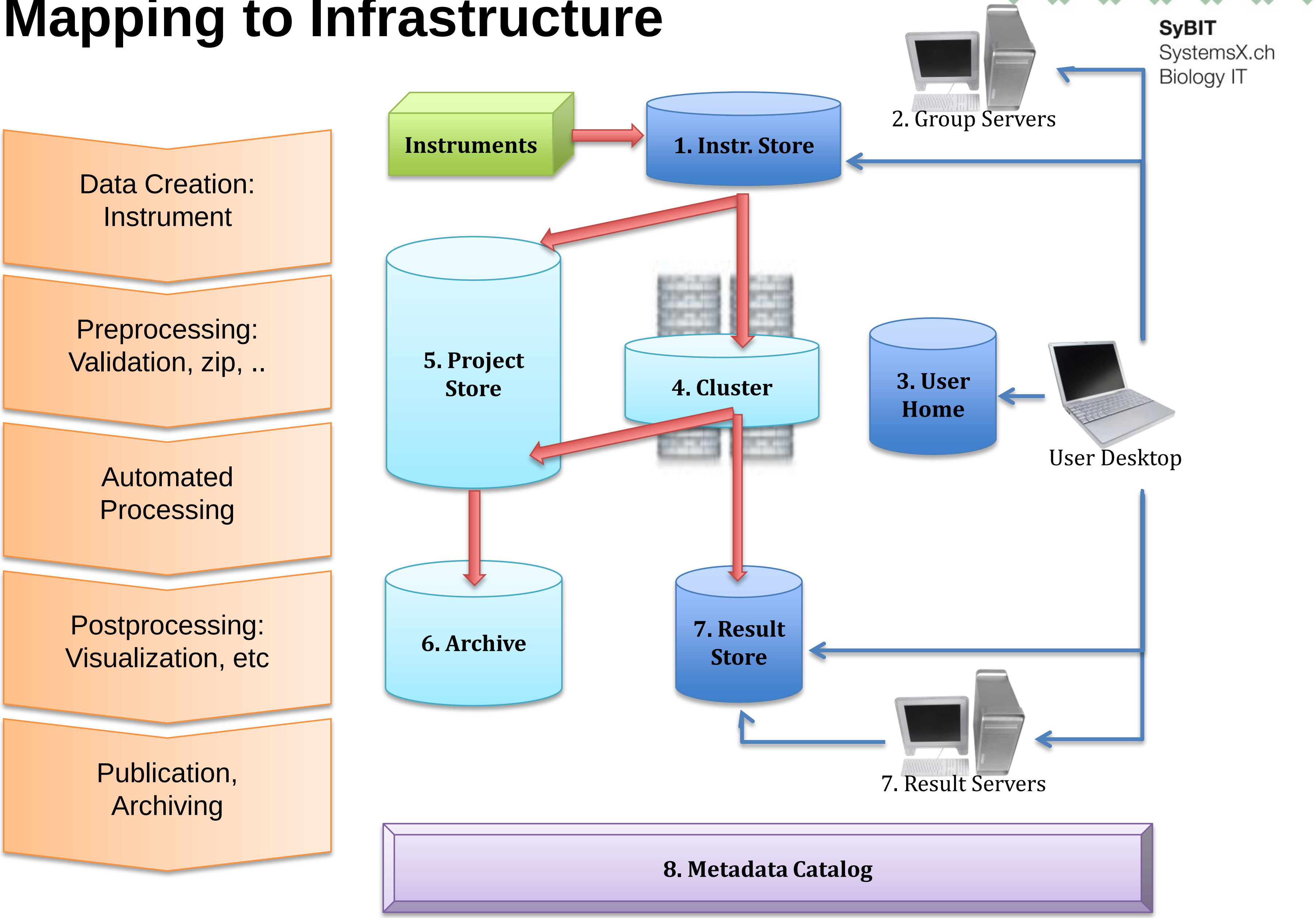
Repetitive problem



- Same lifecycle everywhere
  - Local Policies
  - Local Infrastructures
  - Local Services
  - Local Access control
- Nontrivial coordination effort for RTDs to share data and services

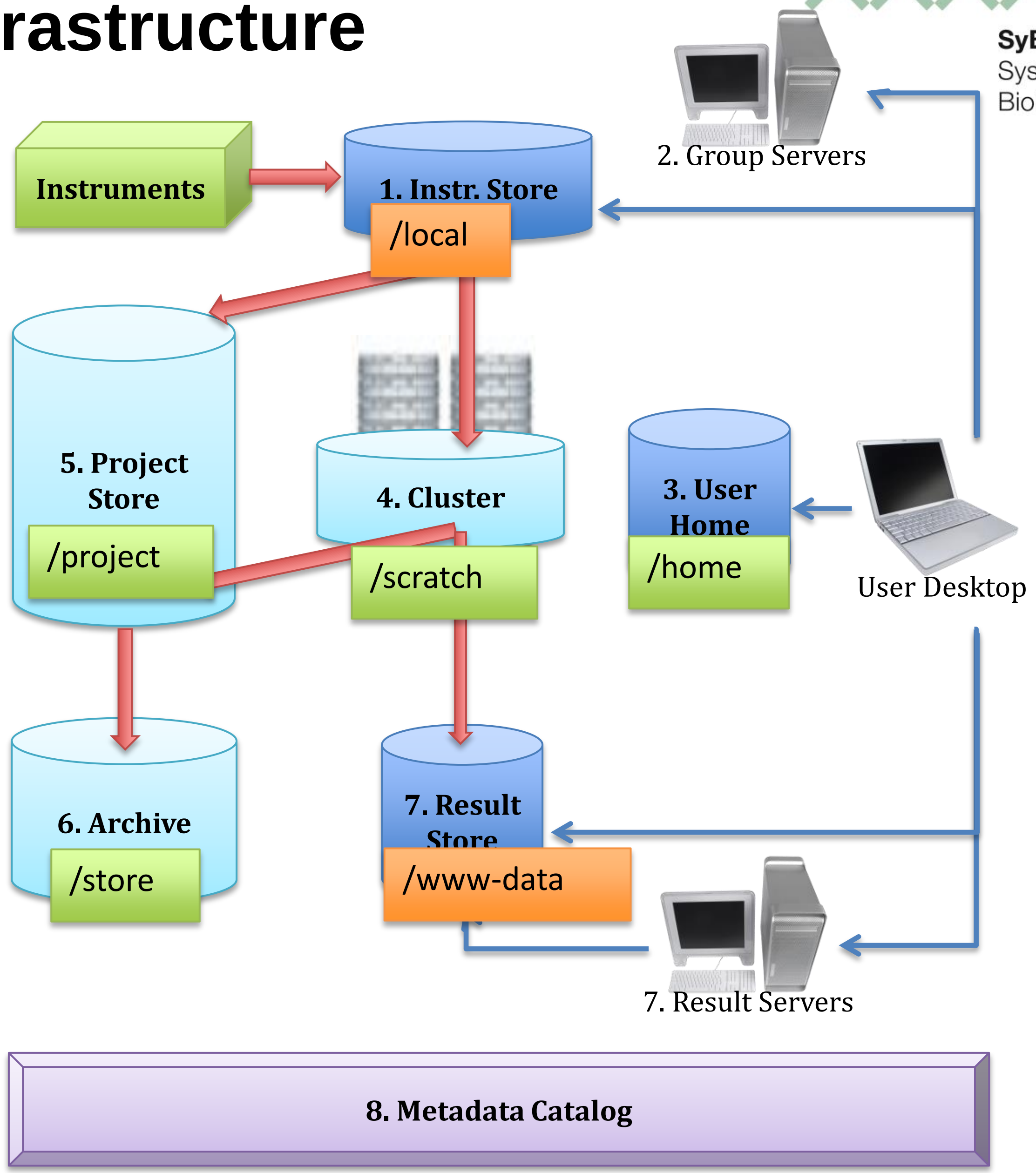
Mapping to Infrastructure

SyBIT
SystemsX.ch
Biology IT



Mapping to Infrastructure

- Data Creation: Instrument
- Preprocessing: Validation, zip, ..
- Automated Processing
- Postprocessing: Visualization, etc
- Publication, Archiving



Instruments

1. Instr. Store
/local

2. Group Servers

5. Project Store
/project

4. Cluster
/scratch

3. User Home
/home

User Desktop

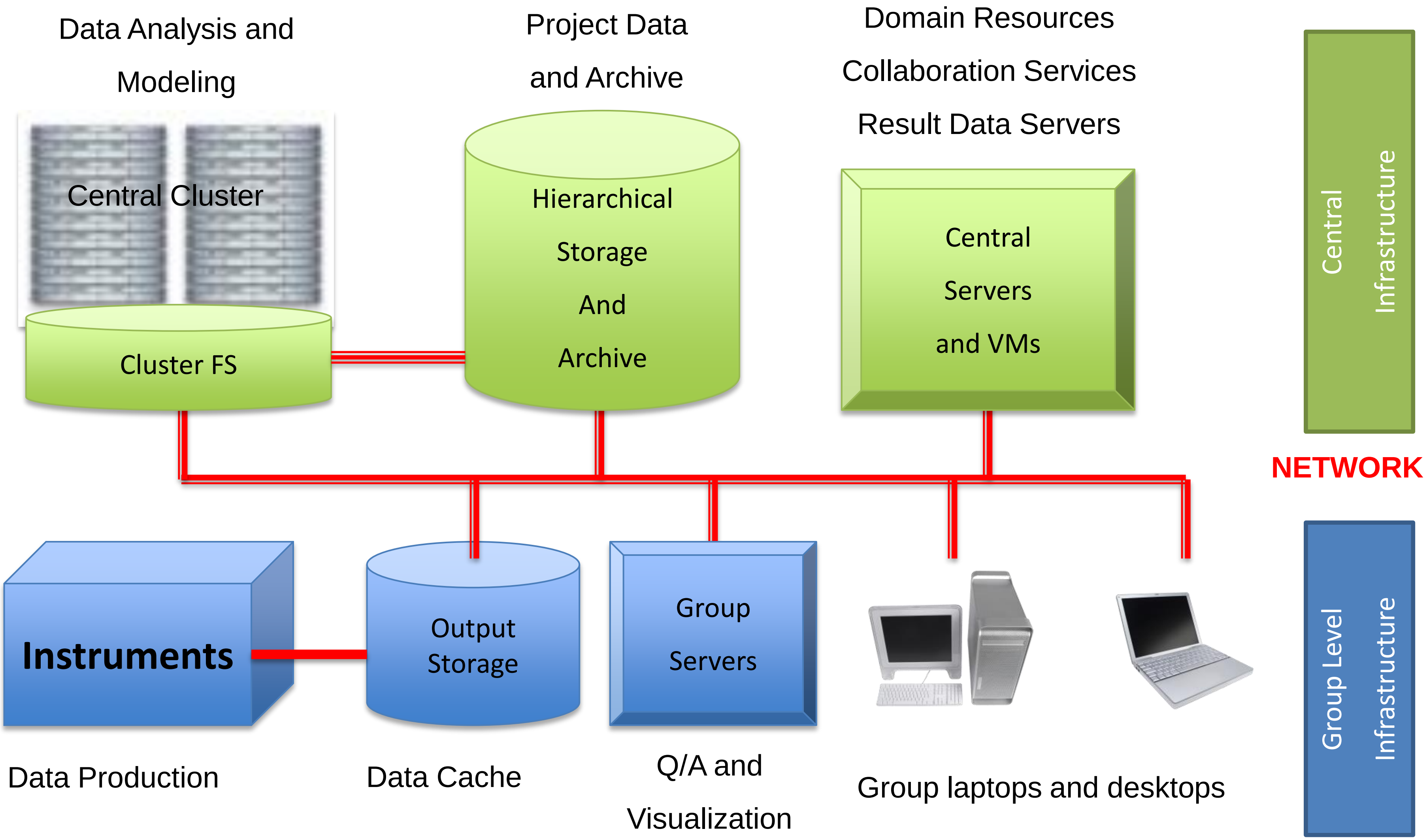
6. Archive
/store

7. Result Store
/www-data

7. Result Servers

8. Metadata Catalog

Central vs. Local Resources



So where are the problems?

- Well.. Where do we start

What is the data again?

- Data is not very large but there are many **different kinds**
- People don´t understand it yet very well
  - How to evolve, schemas, versioning
  - How to be efficient about navigating in data – people still use excel sheets for that
  - CONTROLLED VOCABULARIES help
- **Big Data in life science = More, faster, more diverse than you can handle**

Little motivation to annotate

- Just publish what is necessary and prescribed by the journals
- No recognition yet of producing ´good´ datasets
- Data quality as such is not yet very high, not much reuse (excel sheets limiting factor)

Many formats

- Not many standards exist, there are too many formats
  - Instrument vendors often introduce new formats or conventions
- No trust in central databanks
- People think their data is special and may have some value (in $)
- Not invented here effect is huge
  - Everyone builds their own submission sysem, LIMS, etc

Data Loss

- Student / postdoc leaves, nobody understands data anymore
- Who takes responsibility for data archives?
- No funding to build good archives
- Projects end, data is kept by chance only

SyBIT Divide and Conquer Approach

SyBIT
SystemsX.ch
Biology IT

- **Provide individual solutions for all projects**
  - Allow for several solutions initially
  - Consolidate over time and with new projects

Approach: Divide and Conquer

- **Define Data Policies and User Access**
  - Do not allow users to modify and move around large amounts of data themselves
  - No Filesystem Access for large data – only through data catalogs
  - Automation of data movement and storage

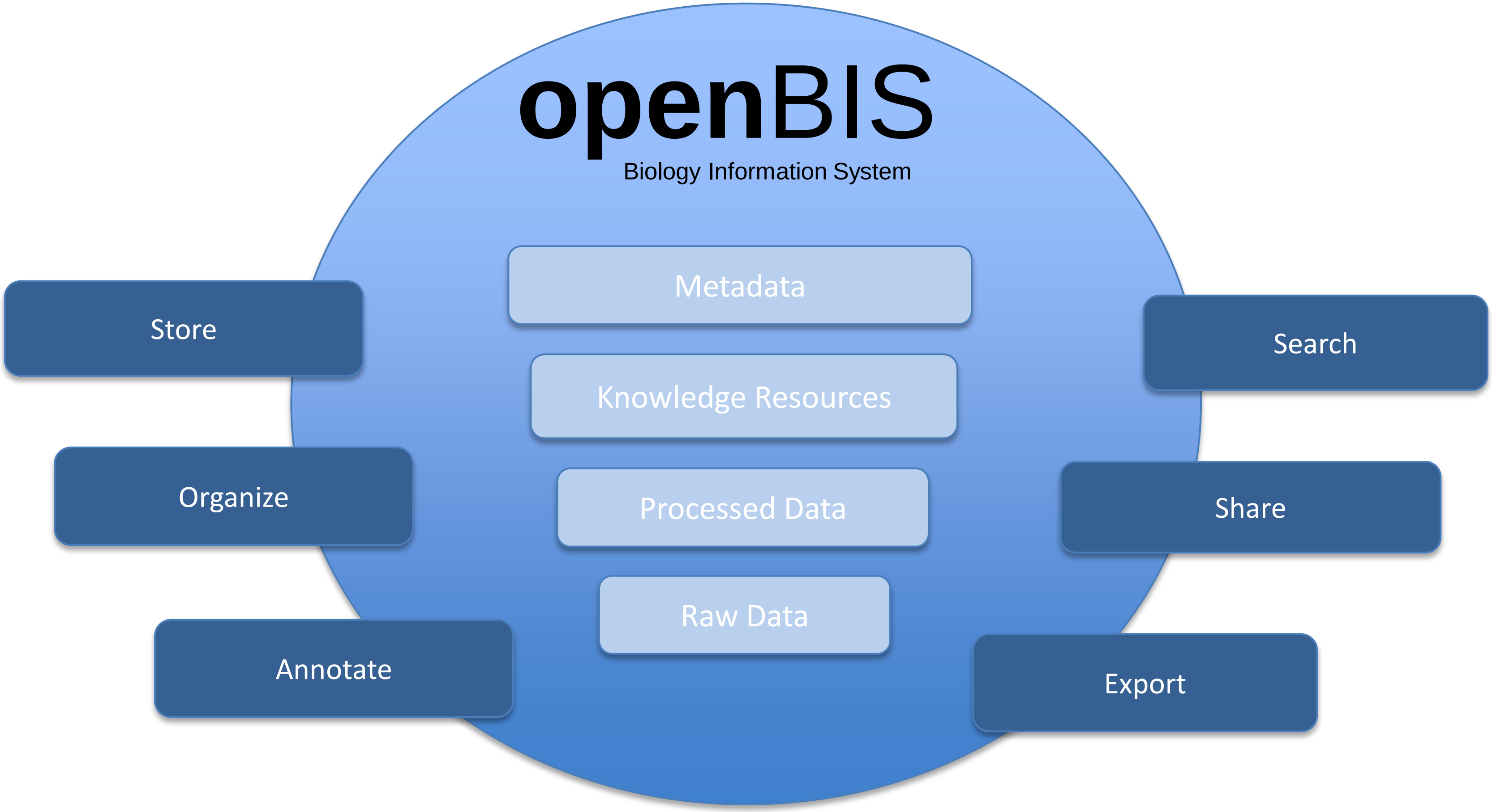
Local – Central infrastrucuture splitup

Several metadata management systems are in place

- **openBIS**
- **B-Fabric**
- **Dedicated databases**

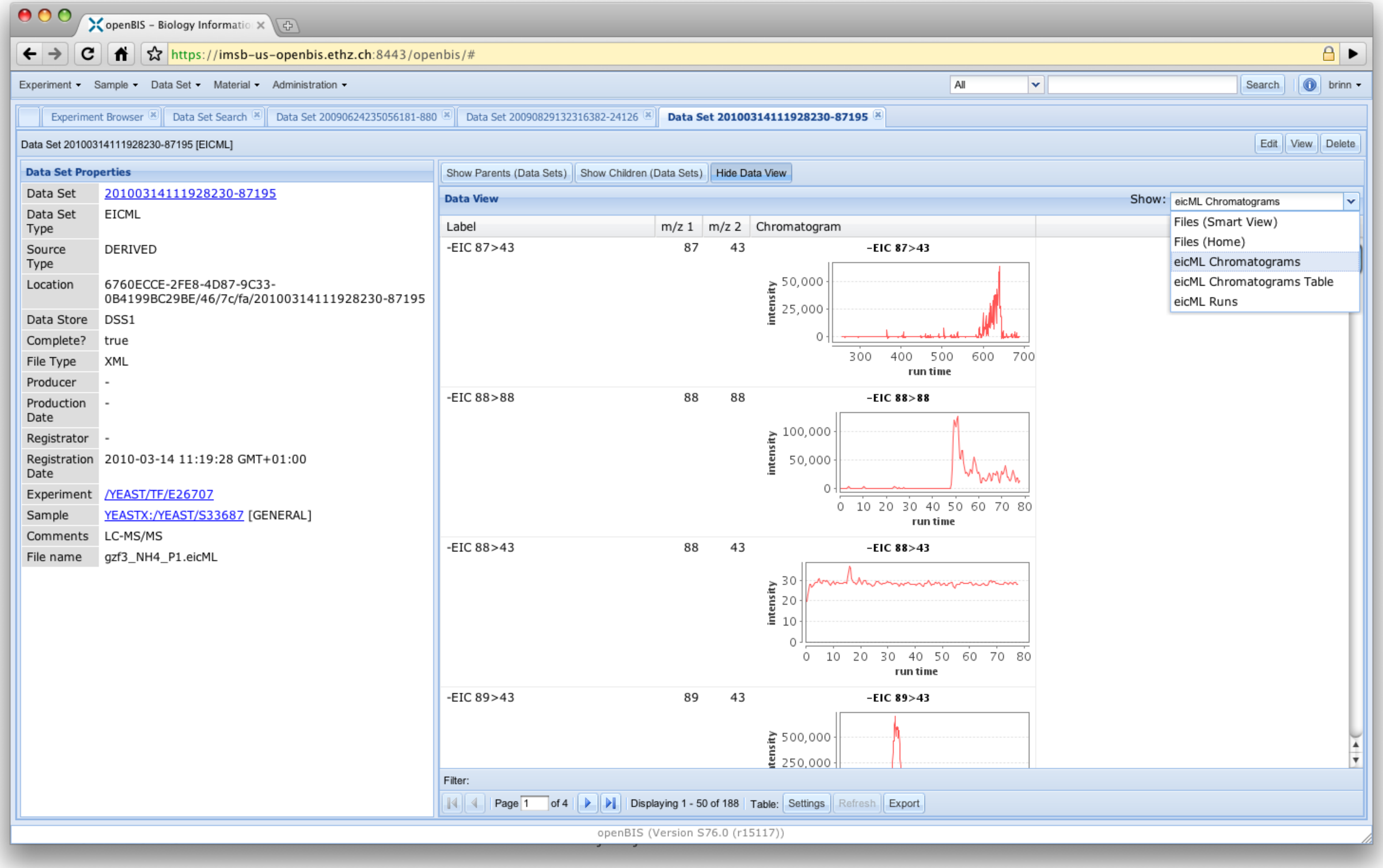
openBIS metadata store

SyBIT
SystemsX.ch
Biology IT



Online Interface ...

SyBIT
SystemsX.ch
Biology IT


openBIS – Biology Informatio
https://imsb-us-openbis.ethz.ch:8443/openbis/#
Experiment ▾ Sample ▾ Data Set ▾ Material ▾ Administration ▾
All
Search
brinn ▾
Experiment Browser
Data Set Search
Data Set 20090624235056181-880
Data Set 20090829132316382-24126
Data Set 20100314111928230-87195
Edit
View
Delete
Show Parents (Data Sets)
Show Children (Data Sets)
Hide Data View
Files (Smart View)
Files (Home)
eicML Chromatograms
eicML Chromatograms Table
eicML Runs
Filter:
Page 1 of 4
Displaying 1 - 50 of 188
Table:
Settings
Refresh
Export
openBIS (Version S76.0 (r15117))

Data Set 20100314111928230-87195 [EICML]

Data Set Properties

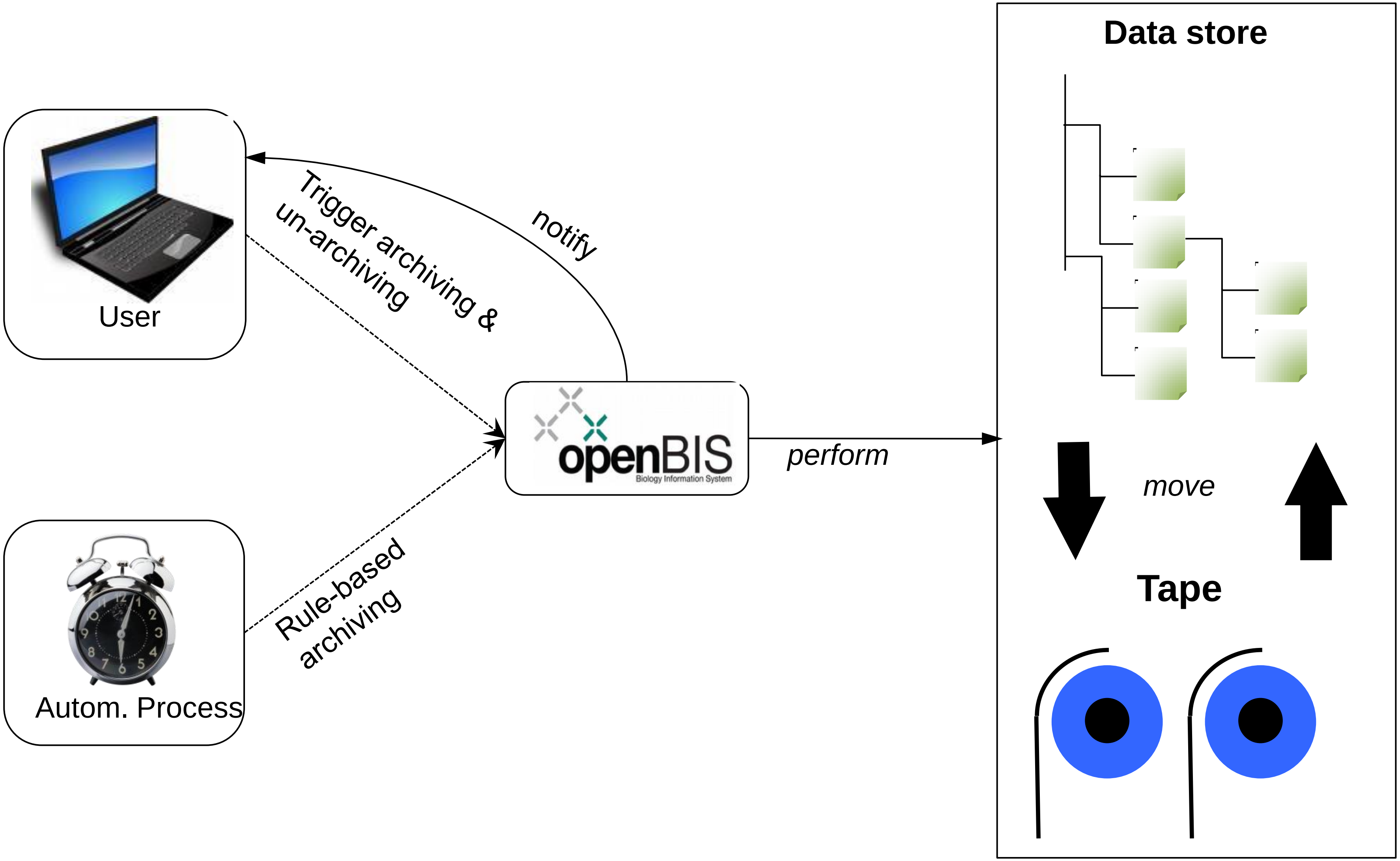
Data Set	20100314111928230-87195
Data Set Type	EICML
Source Type	DERIVED
Location	6760ECCE-2FE8-4D87-9C33-0B4199BC29BE/46/7c/fa/20100314111928230-87195
Data Store	DSS1
Complete?	true
File Type	XML
Producer	-
Production Date	-
Registrator	-
Registration Date	2010-03-14 11:19:28 GMT+01:00
Experiment	/YEAST/TF/E26707
Sample	YEASTX:/YEAST/S33687 [GENERAL]
Comments	LC-MS/MS
File name	gzf3_NH4_P1.eicML

Data View

Show: eicML Chromatograms

Label	m/z 1	m/z 2	Chromatogram
-EIC 87>43	87	43	
-EIC 88>88	88	88	
-EIC 88>43	88	43	
-EIC 89>43	89	43	

Automatic Data Management



Approach: Divide and Conquer

- **Define Pipelines for each instrument type**
  - Platform providers already exist and have a lot of experience we can build on

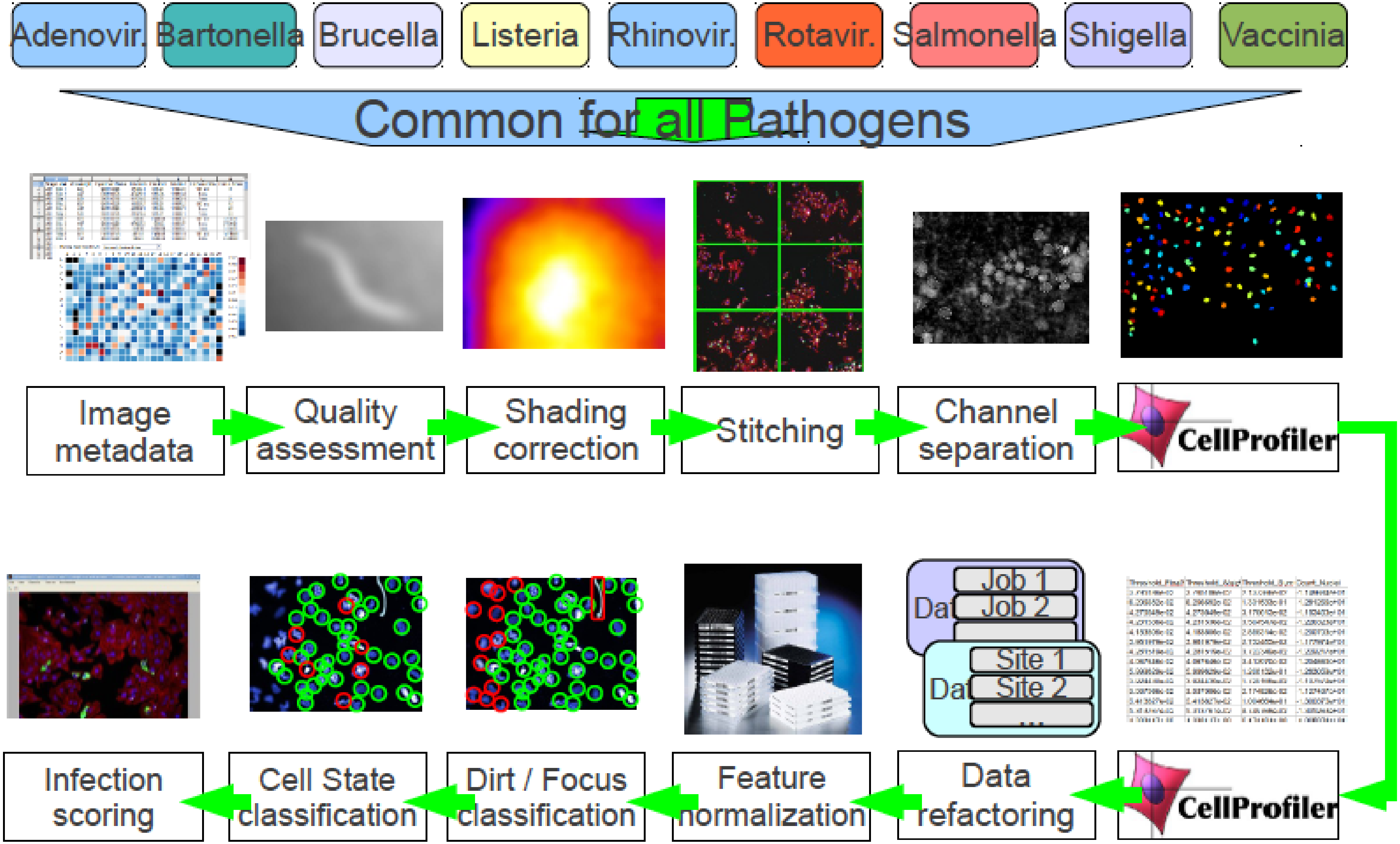
A lot of work was invested into pipelines and workflows: we are in a reasonably good shape

- **Proteomics**
- **Genomics**
- **Imaging**

We are improving these continuously

Image Analysis Workflow

SyBIT
SystemsX.ch
Biology IT



Adenovir. Bartonella Brucella Listeria Rhinovir. Rotavir. Salmonella Shigella Vaccinia

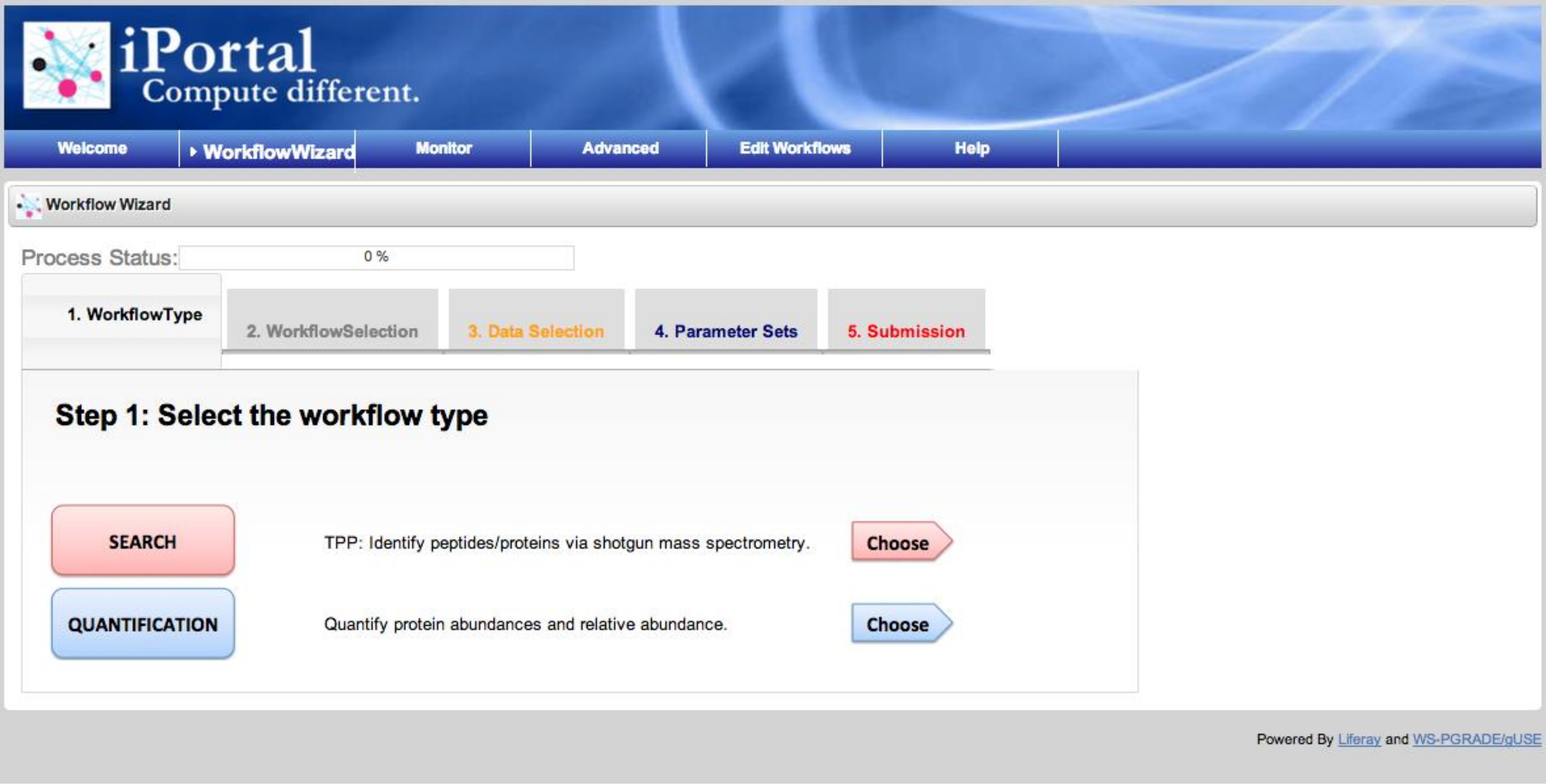
Common for all Pathogens

Image metadata → Quality assessment → Shading correction → Stitching → Channel separation → CellProfiler

CellProfiler → Data refactoring → Feature normalization → Dirt / Focus classification → Cell State classification → Infection scoring

iPortal Proteomics Portal

SyBIT
SystemsX.ch
Biology IT



Process Status: 0 %

Step 1: Select the workflow type

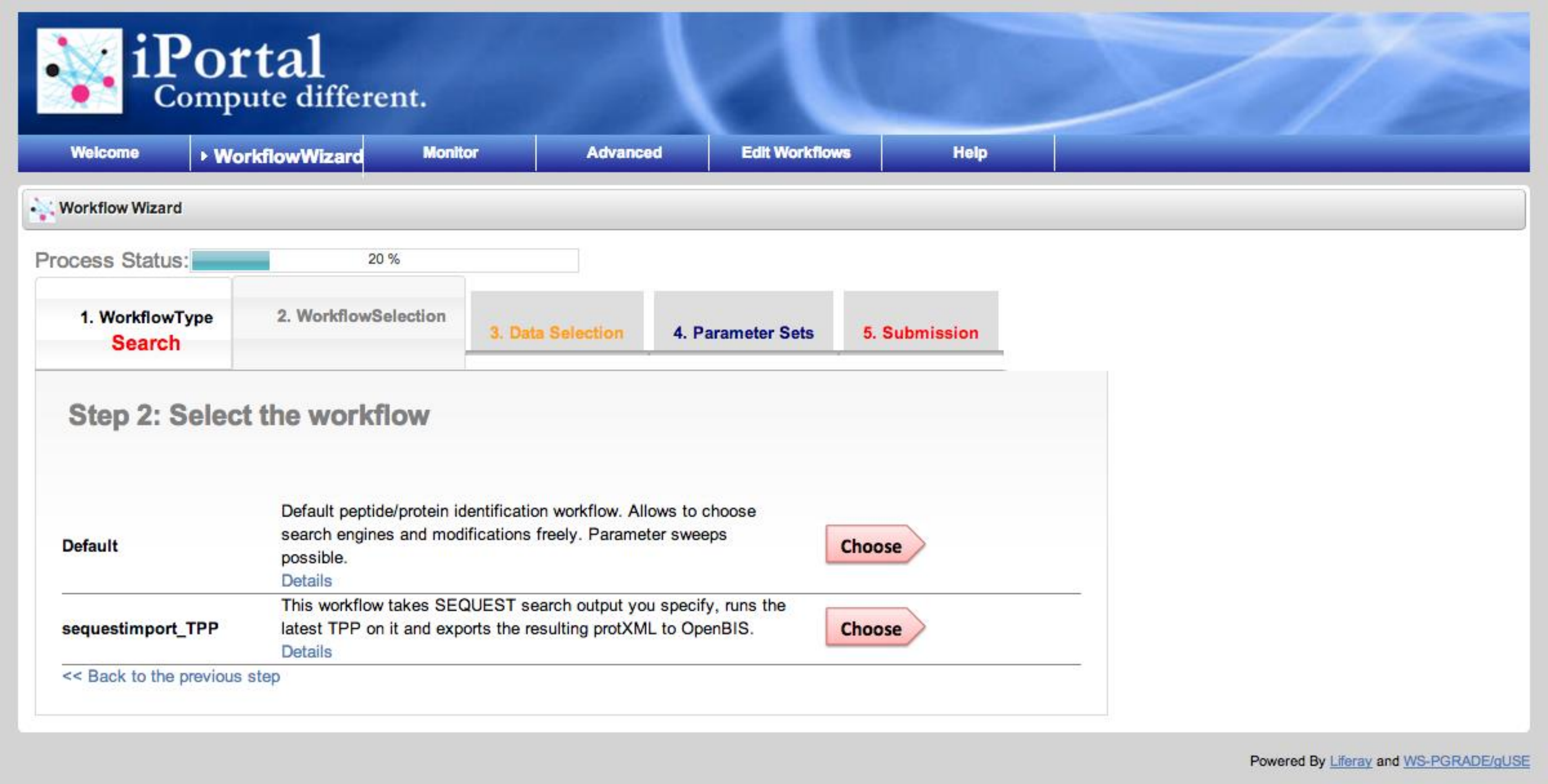
SEARCH — TPP: Identify peptides/proteins via shotgun mass spectrometry. — Choose

QUANTIFICATION — Quantify protein abundances and relative abundance. — Choose

Powered By Liferay and WS-PGRADE/gUSE

iPortal Proteomics Portal

SyBIT
SystemsX.ch
Biology IT



Step 2: Select the workflow

Default	Default peptide/protein identification workflow. Allows to choose search engines and modifications freely. Parameter sweeps possible. Details	Choose
sequestimport_TPP	This workflow takes SEQUEST search output you specify, runs the latest TPP on it and exports the resulting protXML to OpenBIS. Details	Choose

<< Back to the previous step

Powered By Liferay and WS-PGRADE/gUSE

Approach: Divide and Conquer

- **Define common storage types with storage providers**
  - What kinds of storage services do we define?
  - Which kind of service is offered where?

The process has started. Storage infrastructure is still an issue at most sites.

- **Local NAS, Central NAS solutions**
- **IBM Remote NAS project**

Approach: Divide and Conquer

- **Define common data retention semantics**
  - What kind of data is to be kept on what type of storage for how long
  - How are the costs covered

UNSOLVED PROBLEM TODAY

This is a policy issue. Few are willing to make and ENFORCE policy decisions.

Approach: Divide and Conquer

- **Define interfaces for data sharing**
  - Keep responsibility for the data where it was produced (ie. the people who know what it is)
  - Extract data into warehouses for each specific problem

This is what we do by default.

Each project has their dedicated metadata catalog interface. Most project chose openBIS. We are working on interfaces between systems, most already exist.

Approach: Divide and Conquer

- **Define final data repositories for public access**
  - Some public repositories already exist also abroad, agree which one to publish to for each project
  - Where no such repository exists, define and build one for SystemsX.ch

Approach: Divide and Conquer

SyBIT
SystemsX.ch
Biology IT

- Define final data repositories for public

MOSTLY UNSOLVED.

The public published data hosters (EBI, EMBL, NCBI) also struggle with data volumes.

Sending data is difficult, a lot of data curation is necessary – very time consuming process.

For others we set up dedicated openBIS instances.

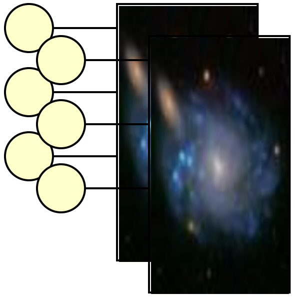
Projects cannot be responsible for long-term archiving!

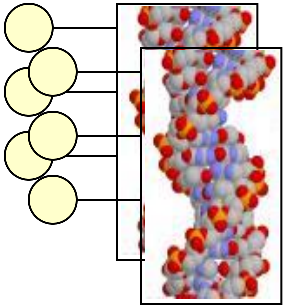
Long-term Data Archives : problem not only for LifeScience

- You have collected some data
  and want to publish science based on it.
- How do you publish the data
  so that others can read it and
  reproduce your results
  in 100 years?
  - Document collection process?
  - How document data processing
    (scrubbing & reducing the data)?
  - Where do you put it?

Generic Objectifying of Knowledge

- This requires agreement about
  - **Units**: cgs
  - **Measurements**: who/what/when/where/how
  - **CONCEPTS:**
    - What’s a planet, star, galaxy,…?
    - What’s a gene, protein, pathway…?
- **Need to objectify science:**
  - what are the objects?
  - what are the attributes?
  - What are the methods (in the object sense)?





Slide adapted from Jim Gray

Generic Objectifying of Knowledge

SyBIT
SystemsX.ch
Biology IT

- This requires agreement about
  - **Un**
  - **Me**
  - **CO**
    -
    -
- **Nee**
  - W
  - What are the attributes?
  - What are the methods (in the object sense)?

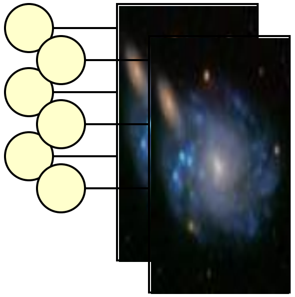
Agreements do not come easy. Need

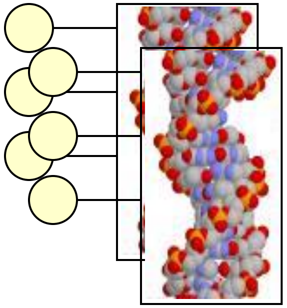
Ontologies

Controlled Vocabularies

Schemas

Domain Experts are still working it out!

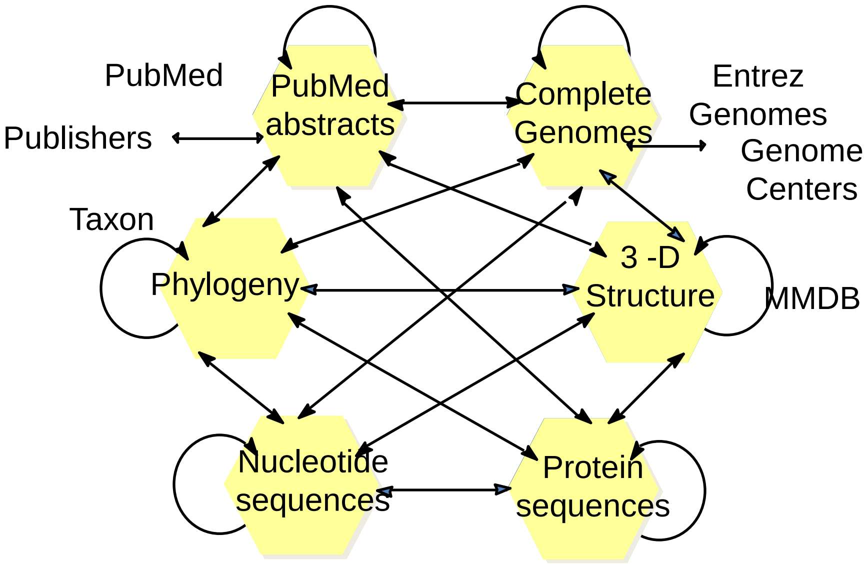




Slide adapted from Jim Gray

Working Example: Entrez-GenBank http://www.ncbi.nlm.nih.gov/

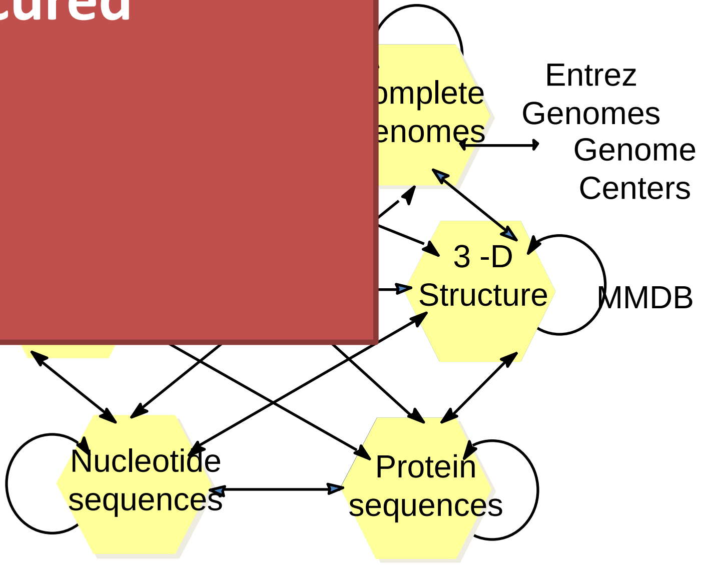
- Sequence data deposited with Genbank
- Literature references Genbank ID
- BLAST searches Genbank
- Entrez integrates and searches
  - PubMedCentral
  - PubChem
  - Genbank
  - Proteins, SNP,
  - Structure,..
  - Taxonomy…
  - Many more



Working Example: Entrez-GenBank
http://www.ncbi.nlm.nih.gov/

- Sequence data deposited with Genbank
- Lite
- BLA
- Ent
  - Pu
  - Pu
  - Ge
  - Pr
  - Structure,..
  - Taxonomy…
  - Many more

NCBI Funding is not secured



Data Sharing/Publishing Challenges

SyBIT
SystemsX.ch
Biology IT

- How to keep programs that access and work with data? Cloud App Store for science? Who maintains those?
- What is the business model (reward/career benefit)?
- Journals starting to adapt, see Scientific Data http://www.nature.com/scientificdata
- But, many kinds of data are still orphaned: where to send Microscopy images?

eScientists

Researchers

- Domain specific expertise
- Measured by publications and impact
- Develops
  - New algorithms
  - New models
- Find new connections
- Train new scientists

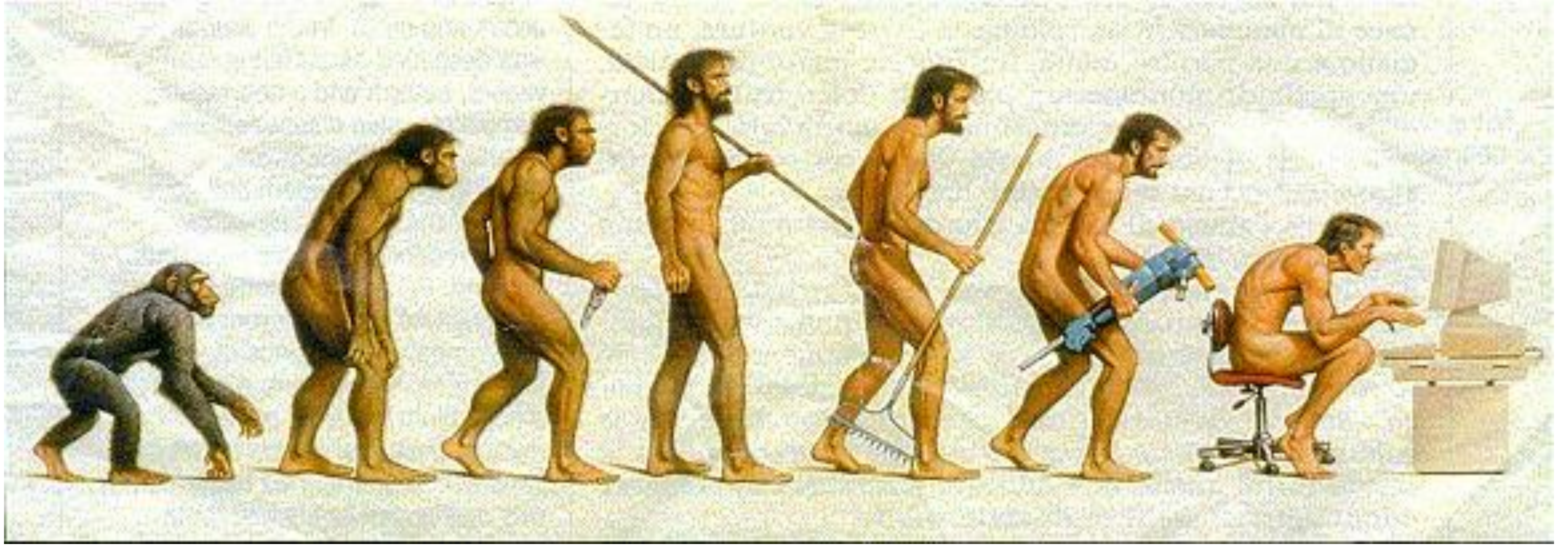
Supporters / Stewards

- Domain specific expertise
- Measured by number of people supported
- Maintains existing tools and services, operations
- Trains scientists in usage of tools
- Curation of results, documentation

Summary

- We need to manage the increased complexity (data and people)
- Need for specialized domain expert eScience supporters
- Scale-out distributed infrastructures enable us to grow: watch the clouds!
- Each scientific domain will address their challenges in their own way, no one size fits all.

Coming Up?



Thanks to

- Jim Gray and Tony Hey, Microsoft Research
- Alex Szalay, Johns Hopkins University
- Bernd Rinn, ETH Zurich

- skyserver.sdss.org
- www.lsst.org
- www.sybit.net
- http://www.nature.com/scientificdata/about/