Hasso Plattner Institut
Digital Engineering · Universität Potsdam

Data Management for Precision Oncology

Borchert, Dr. Schapranow

Data Management for Digital Health
Winter 2023

Agenda
Pillars of the Lecture

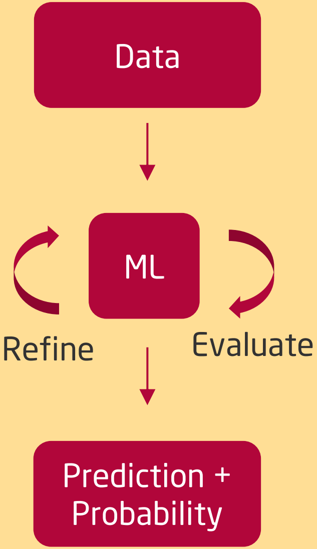
Medical Use Cases

- Biology Recap
- Oncology
- Nephrology
- Infectious Diseases

Technology Foundation

- Data Sources
- Data Formats
- Processing and Analysis
- Software Architectures

Machine Learning



Data

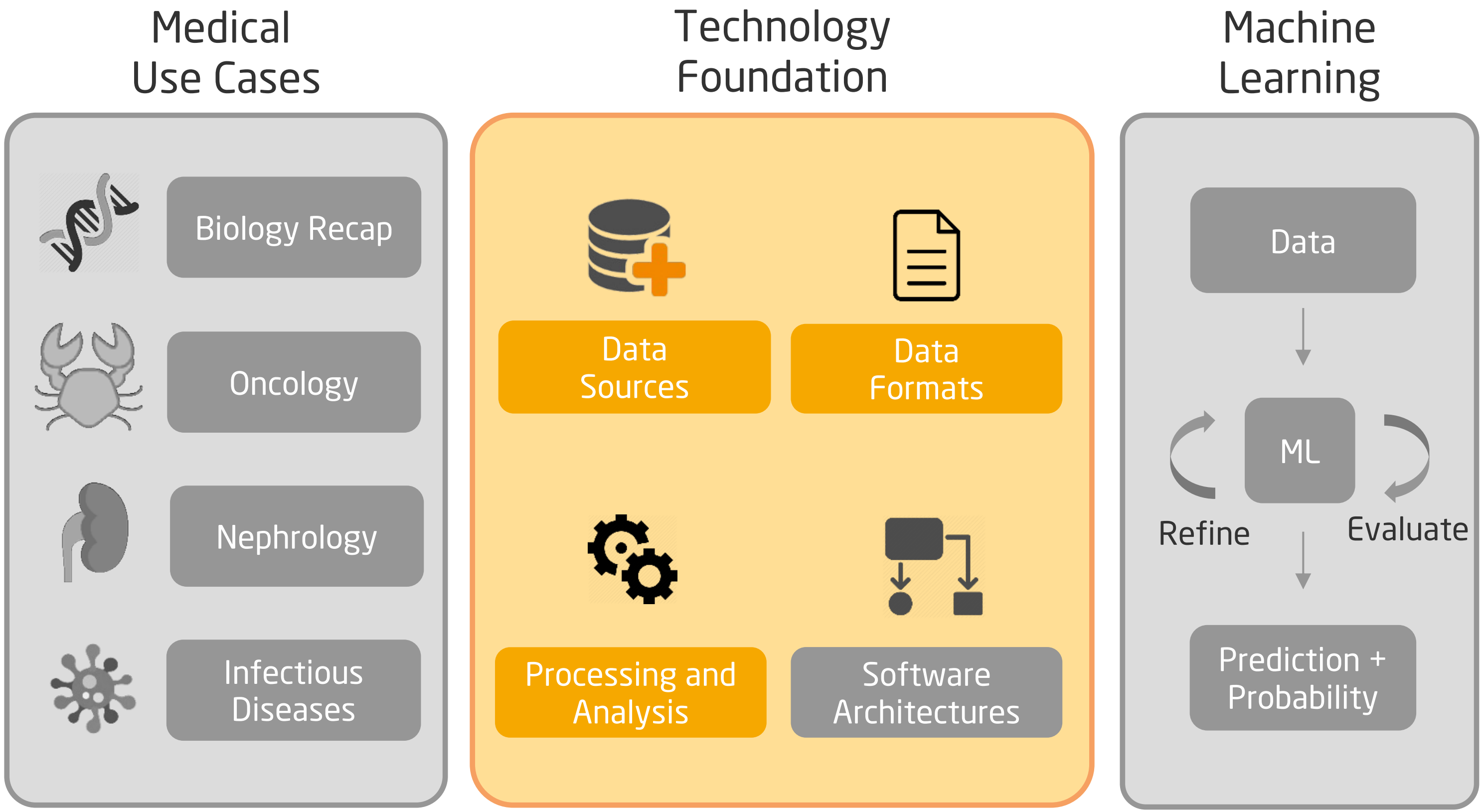
ML

Refine

Evaluate

Prediction + Probability

Agenda
Pillars of the Lecture



Medical Use Cases

Biology Recap

Oncology

Nephrology

Infectious Diseases

Technology Foundation

Data Sources

Data Formats

Processing and Analysis

Software Architectures

Machine Learning

Data

ML

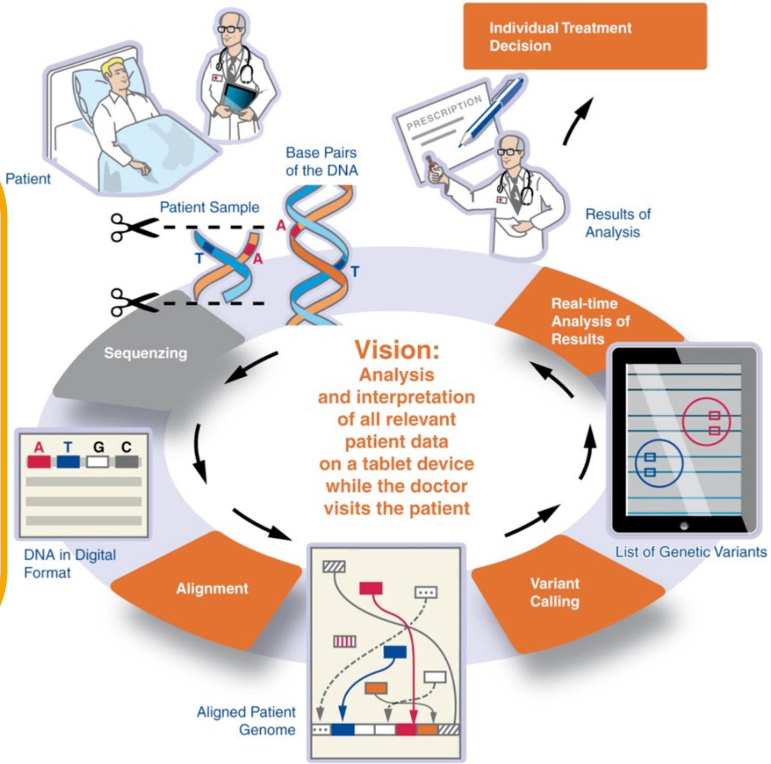
Refine

Evaluate

Prediction + Probability

From Raw Genome Data to Analysis

- **Sequencing:** Acquire digital DNA data (FASTQ)
- **Alignment**: Reconstruction of complete genome with snippets (SAM,BAM)
- **Variant Calling**: Identification of genetic variants (VCF)
- **Data Annotation**: Linking genetic variants with research findings



Science Crisis
<< QUIZ >>

- What do you think is the most challenging aspect in science today?

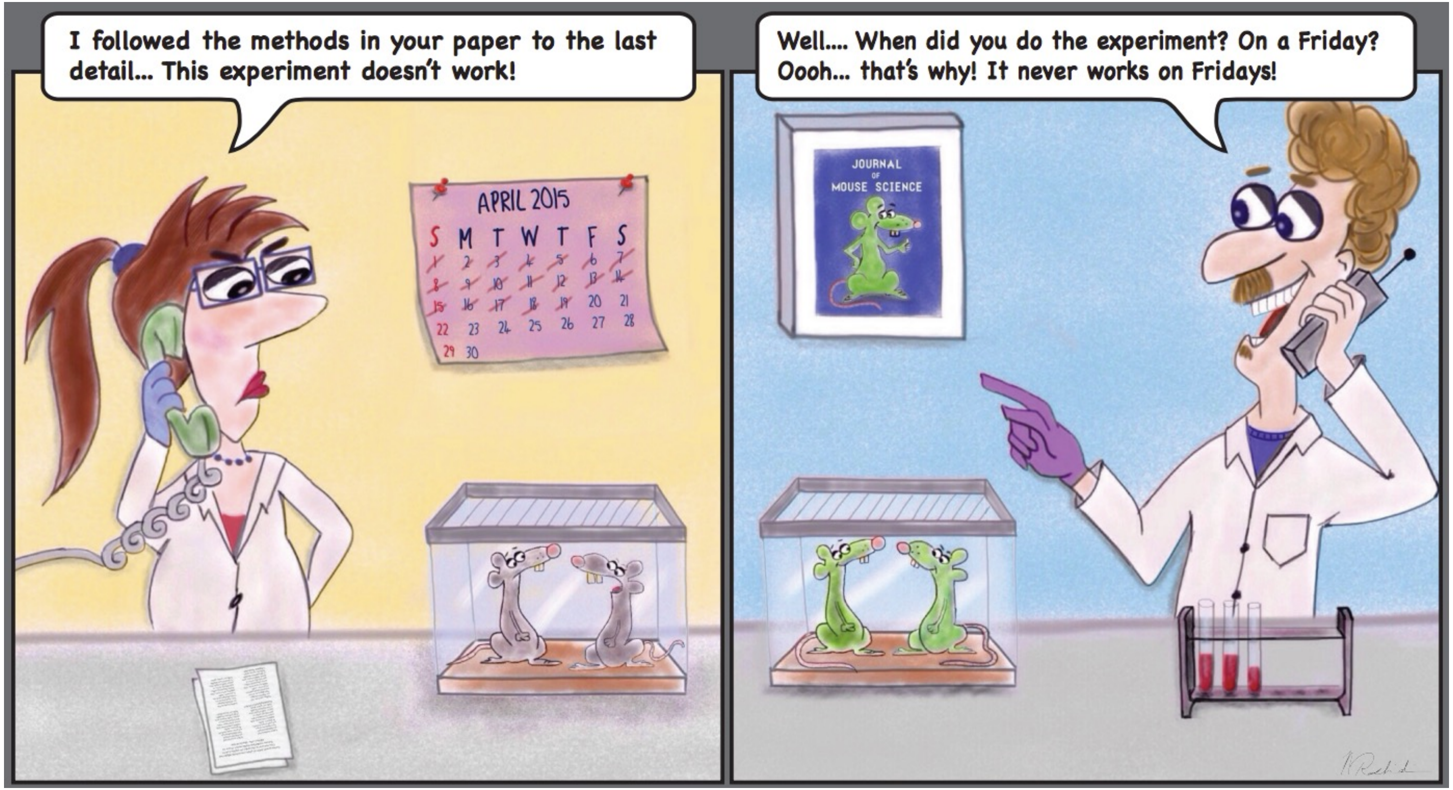
A. Effects of COVID-19 pandemic

B. Salary of experts

C. Reproducibility of results

D. Missing subject-matter experts

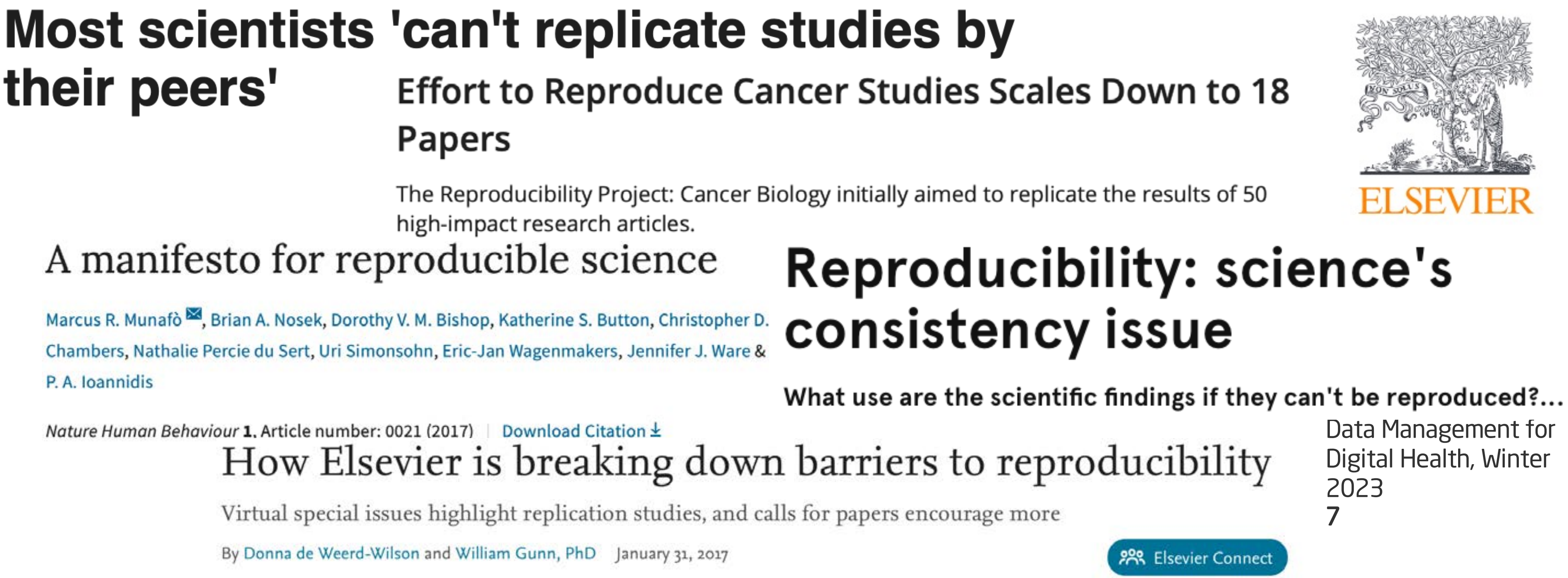
Challenge: Reproducibility Crisis of Science


I followed the methods in your paper to the last detail... This experiment doesn't work!
APRIL 2015
S M T W T F S
1 2 3 4 5 6 7
8 9 10 11 12 13 14
15 16 17 18 19 20 21
22 23 24 25 26 27 28
29 30
Well.... When did you do the experiment? On a Friday? Oooh... that's why! It never works on Fridays!
JOURNAL OF MOUSE SCIENCE

https://www.digital-science.com/blog/2015/03/digital-science-doodles-data-reproducibility/

Challenge: Reproducibility Crisis of Science

- Reproducibility crisis is named to have its roots in the early 2010s
- Still an ongoing issue with latest initiatives addressing it



Most scientists 'can't replicate studies by their peers'

Effort to Reproduce Cancer Studies Scales Down to 18 Papers

The Reproducibility Project: Cancer Biology initially aimed to replicate the results of 50 high-impact research articles.

A manifesto for reproducible science

Marcus R. Munafò, Brian A. Nosek, Dorothy V. M. Bishop, Katherine S. Button, Christopher D. Chambers, Nathalie Percie du Sert, Uri Simonsohn, Eric-Jan Wagenmakers, Jennifer J. Ware & P. A. Ioannidis

Nature Human Behaviour **1**, Article number: 0021 (2017) | Download Citation ⤓

Reproducibility: science's consistency issue

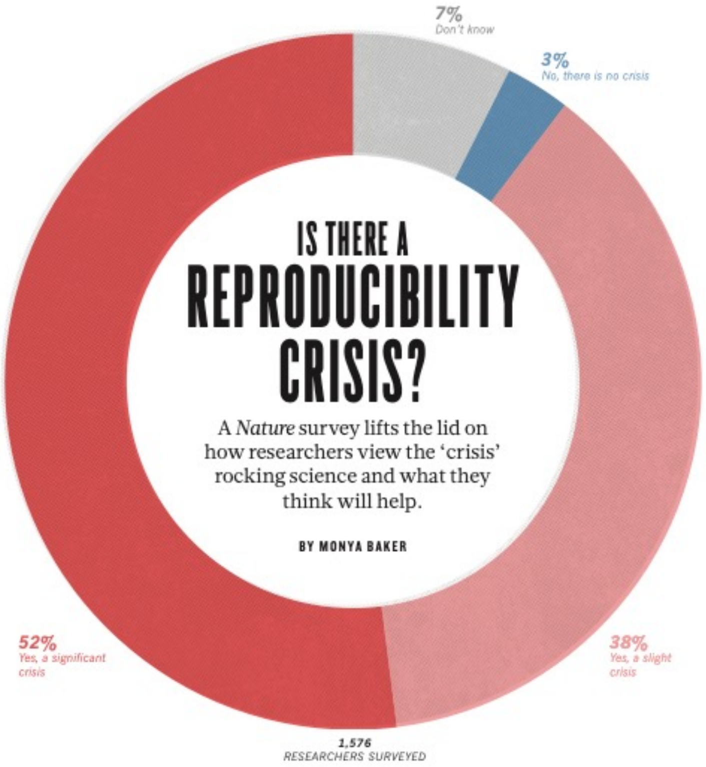
What use are the scientific findings if they can't be reproduced?...

How Elsevier is breaking down barriers to reproducibility

Virtual special issues highlight replication studies, and calls for papers encourage more

By Donna de Weerd-Wilson and William Gunn, PhD January 31, 2017

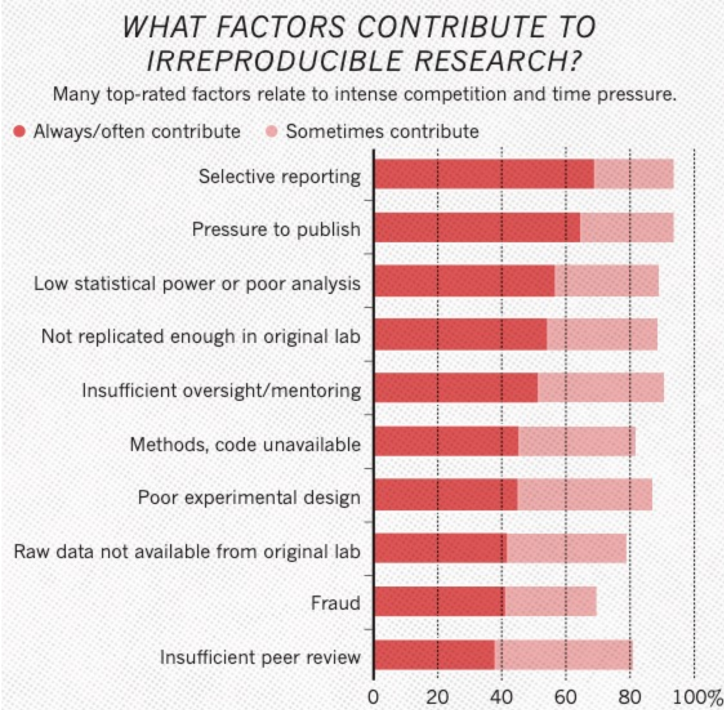
Challenge: Reproducibility Crisis of Science
1,500 Scientists Lift the Lid on Reproducibility



IS THERE A REPRODUCIBILITY CRISIS?

A *Nature* survey lifts the lid on how researchers view the 'crisis' rocking science and what they think will help.

BY MONYA BAKER

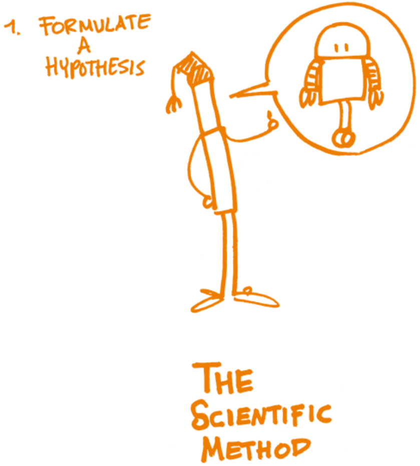


WHAT FACTORS CONTRIBUTE TO IRREPRODUCIBLE RESEARCH?

Many top-rated factors relate to intense competition and time pressure.

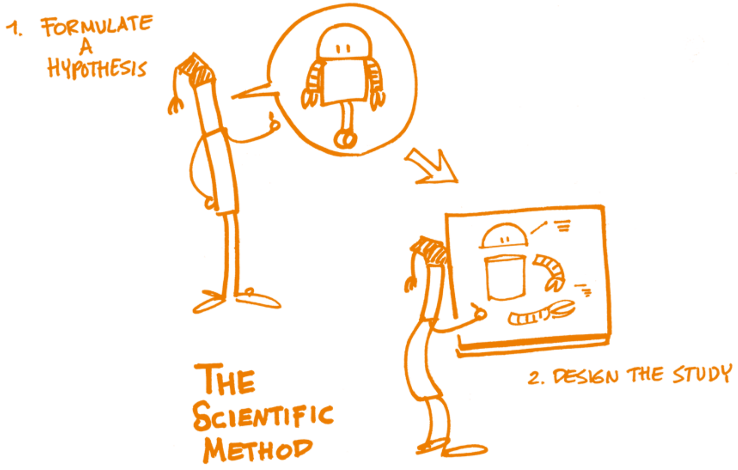
Baker, M. 1,500 scientists lift the lid on reproducibility. *Nature* 533, 452–454 (2016). https://doi.org/10.1038/533452a

The Science Loop



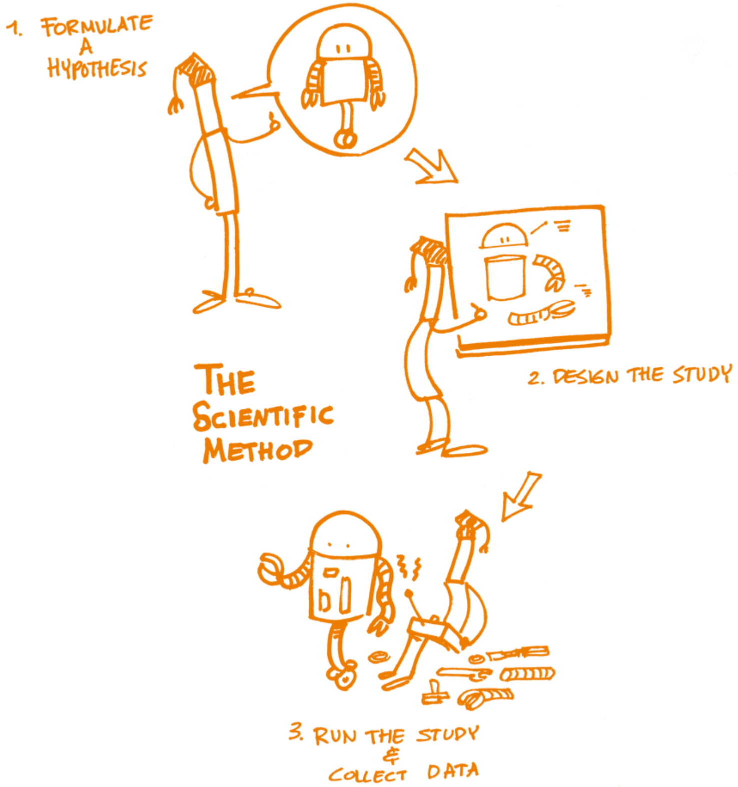
https://open-science-training-handbook.gitbook.io/book/open-science-basics/reproducible-research-and-data-analysis

The Science Loop (cont'd)



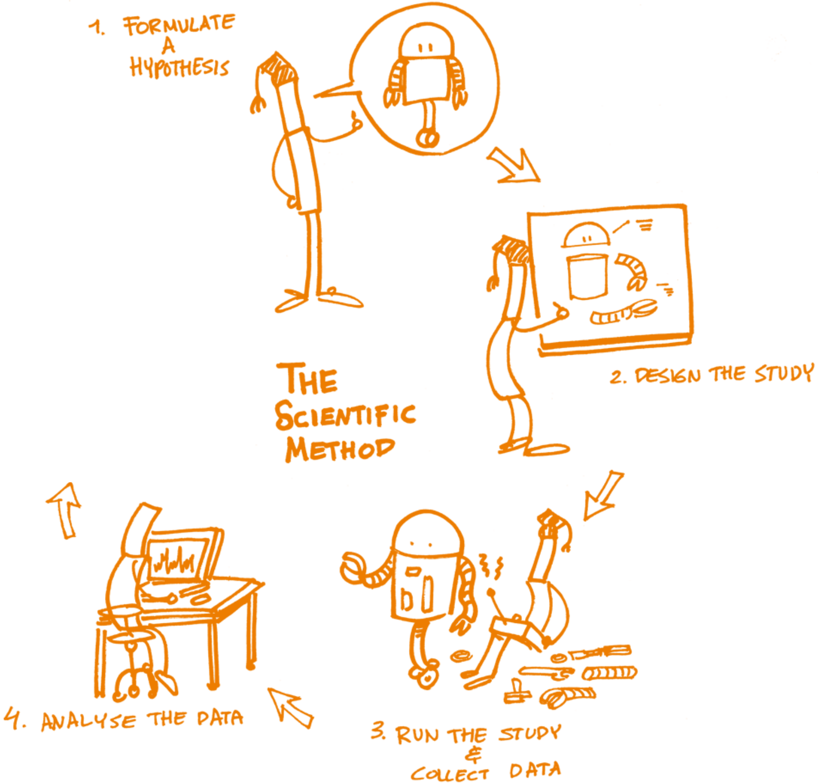
https://open-science-training-handbook.gitbook.io/book/open-science-basics/reproducible-research-and-data-analysis

The Science Loop (cont'd)



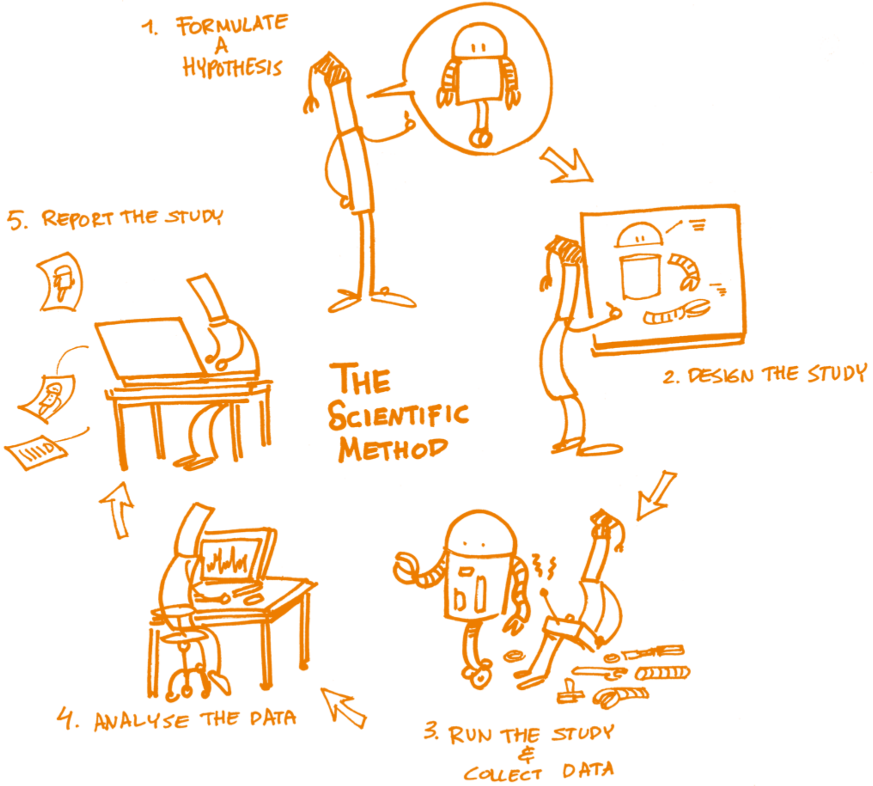
https://open-science-training-handbook.gitbook.io/book/open-science-basics/reproducible-research-and-data-analysis

The Science Loop (cont'd)



https://open-science-training-handbook.gitbook.io/book/open-science-basics/reproducible-research-and-data-analysis

The Science Loop (cont'd)



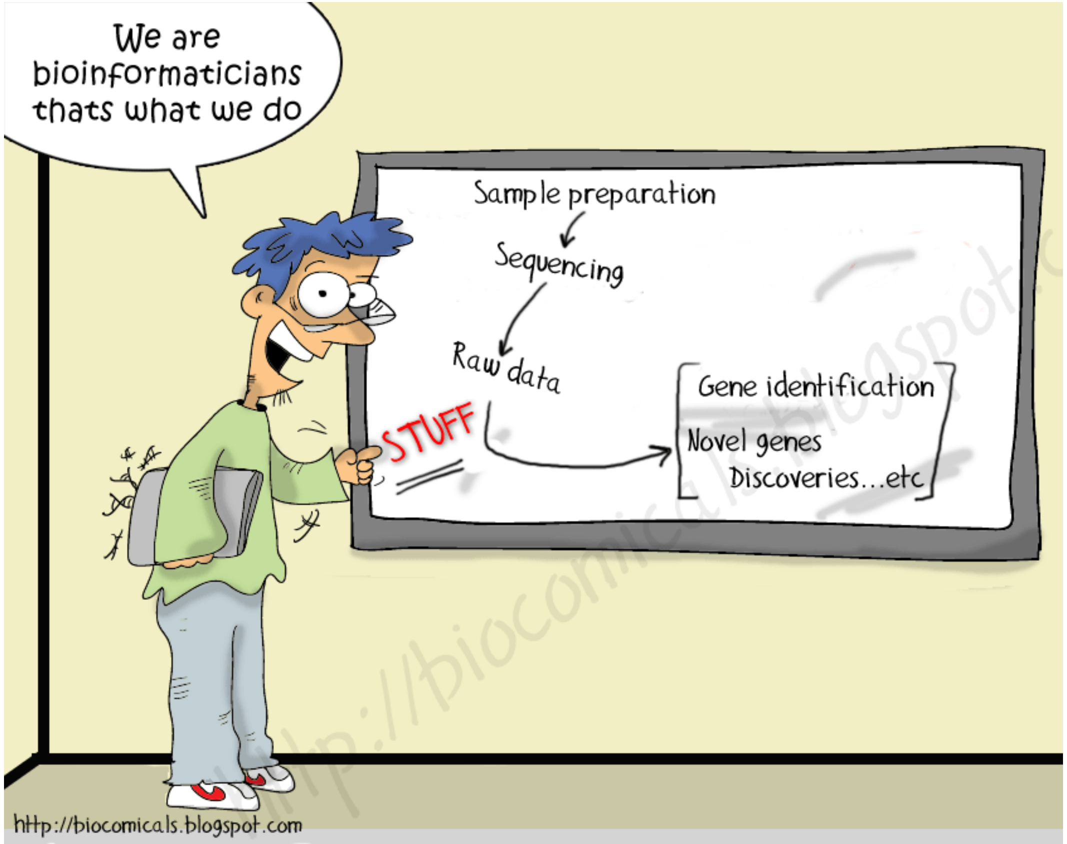
https://open-science-training-handbook.gitbook.io/book/open-science-basics/reproducible-research-and-data-analysis

Good Scientific Practicies

Approach	Description
Open Data	Share results + underlying data with other scientists in an open way
Collaboration	Working with other research groups, both formally and informally
Automation	Use technology to standardize processing, thereby reducing the probability for human error
Open Methods	Publicly publishing the detail of a study protocol
Post-publication Review	Continuing discussion of a study in a public forum after it has been published (most are reviewed before publication only)
Reporting Guidelines	Guidelines and checklists that help researchers to meet certain criteria for publishing results

Adapted from: Reproducibility and reliability of biomedical research: Improving research practice. Symposium report, October 2015. https://acmedsci.ac.uk/viewFile/56314e40aac61.pdf

How to Find a Solution?
Call a Bioinformatician



Motivation: Ensure Reproducible through Genome Data Processing Pipelines

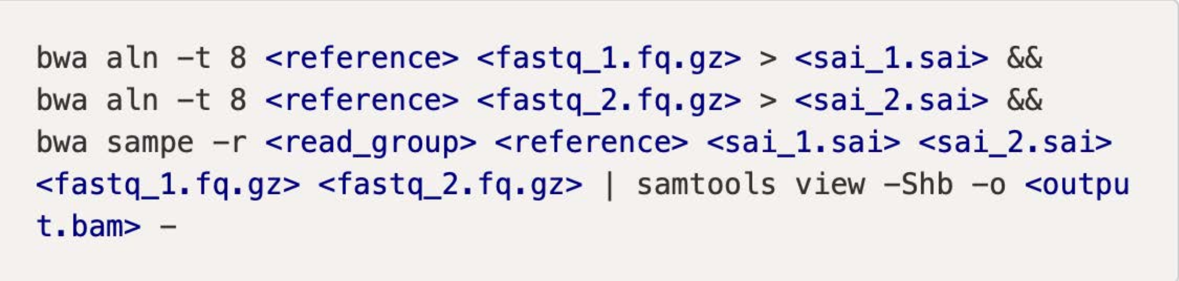
- **Genome Data Processing Pipelines (GDPPs)** := Structured processing of raw genome data, e.g. FASTQ, to provide meaningful insights, e.g. variants, annotations

- Objective: GDPPs should be…
  - □ Human- and machine-readable to support understanding, e.g. graphical modeling
  - □ Reproducible, i.e. generation of identical output for identical input
  - □ Exchangeable, i.e. other sites should be able to create identical results
  - □ Understandable, e.g. by non-IT experts, lab staff, etc.

Genome Data Processing Pipelines: State of the Art

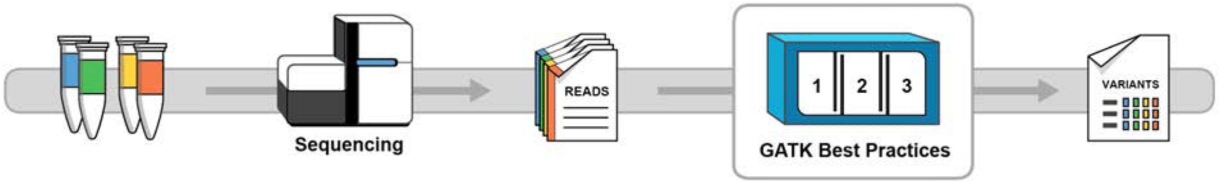
bwa aln hg19.fa sample.fastq | bwa samse hg19.fa – sample.fastq | samtools view -Su - | samtools sort …

- Concatenation of command line tools reading/writing files from/to hard disk
- Requires dedicated expertise for
  - □ Setup and configuration,
  - □ Error handling, and
  - □ Scalable processing
- Lack of
  - □ Standardization and exchangeability,
  - □ Understandability,
  - □ Maintainability, and
  - □ Reproducibility.

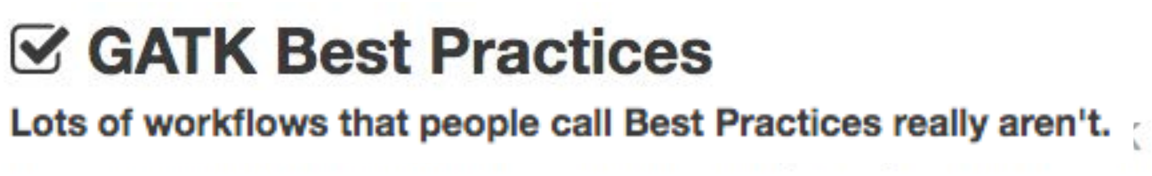


```
bwa aln -t 8 <reference> <fastq_1.fq.gz> > <sai_1.sai> &&
bwa aln -t 8 <reference> <fastq_2.fq.gz> > <sai_2.sai> &&
bwa sampe -r <read_group> <reference> <sai_1.sai> <sai_2.sai>
<fastq_1.fq.gz> <fastq_2.fq.gz> | samtools view -Shb -o <outpu
t.bam> -
```

https://docs.gdc.cancer.gov/Data/Bioinformatics_Pipelines/DNA_Seq_Variant_Calling_Pipeline/



- The BROAD is a joint institute of MIT and Harvard established 2004 in Cambridge, MA
- Genome Analysis Toolkit (GATK) focuses on variant detection
- Open-source tools and shared best-practices

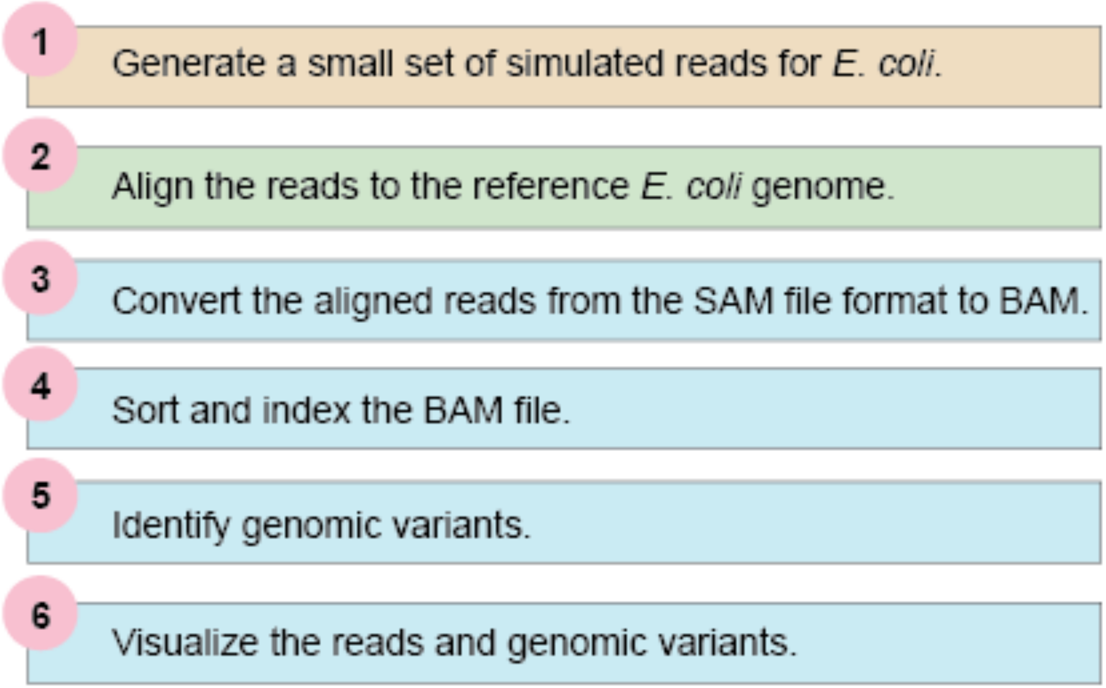


☑ GATK Best Practices

Lots of workflows that people call Best Practices really aren't.

https://software.broadinstitute.org/gatk/best-practices/

Best Practices I:
Let us Make a Workflow!



1. Generate a small set of simulated reads for *E. coli*.
2. Align the reads to the reference *E. coli* genome.
3. Convert the aligned reads from the SAM file format to BAM.
4. Sort and index the BAM file.
5. Identify genomic variants.
6. Visualize the reads and genomic variants.

wgsim (optional step, intermediate files available for download)

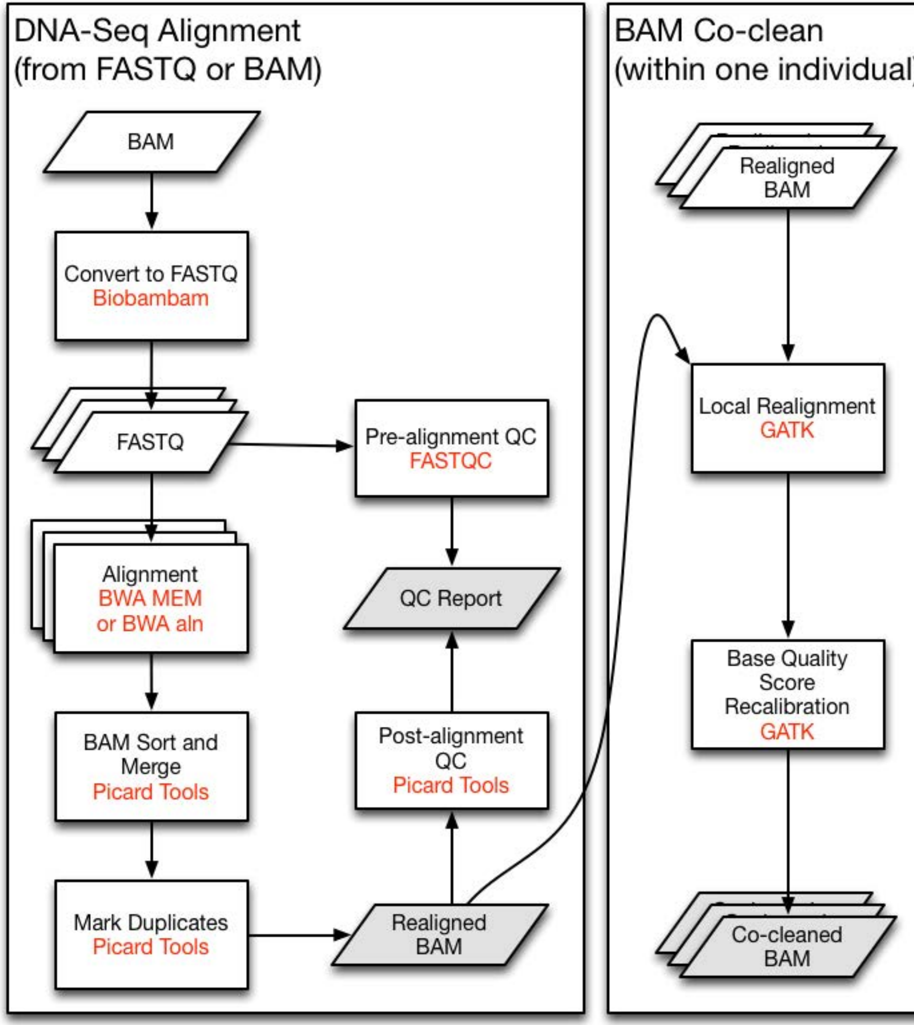
bowtie2 (optional step, intermediate files available for download)

samtools

http://biobits.org/samtools_primer.html

Best Practices II: Modeling

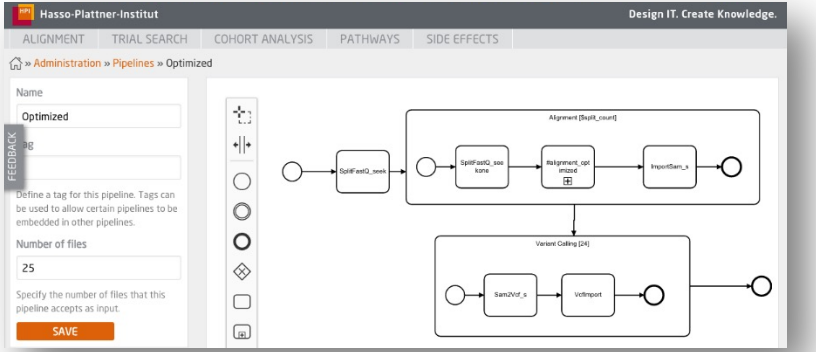
- Incorporate graphical modeling techniques to document and share knowledge
- Build on existing methods to benefit from existing tools or communities



https://docs.gdc.cancer.gov/Data/Bioinformatics_Pipelines/DNA_Seq_Variant_Calling_Pipeline

Reproducibility at we.analyzegenomes.com: Modeling of Data Analysis Pipelines

1. Design time (researcher, process expert)
   - □ Definition of parameterized process model
   - □ Uses graphical editor and jobs from repository
2. Configuration time (researcher, lab assistant)
   - □ Select model and specify parameters, e.g. aln opts
   - □ Results in model instance stored in repository
3. Execution time (researcher)
   - □ Select model instance
   - □ Specify execution parameters, e.g. input files



Name

Optimized

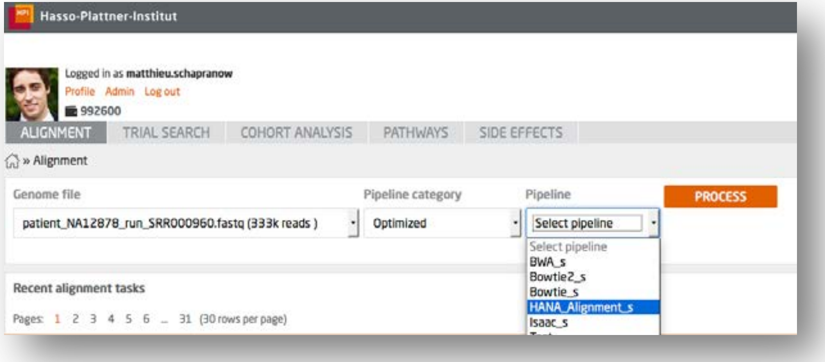
ag

Define a tag for this pipeline. Tags can be used to allow certain pipelines to be embedded in other pipelines.

Number of files

25

Specify the number of files that this pipeline accepts as input.



Logged in as **matthieu.schapranow**

992600

Genome file: patient_NA12878_run_SRR000960.fastq (333k reads)

Pipeline category: Optimized

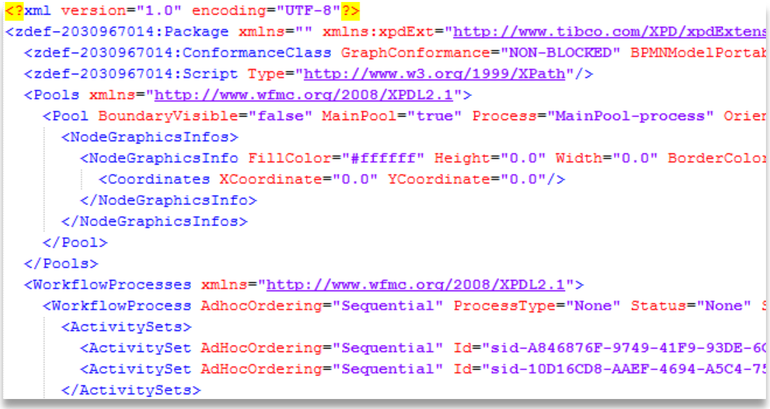
Pipeline: Select pipeline

Recent alignment tasks

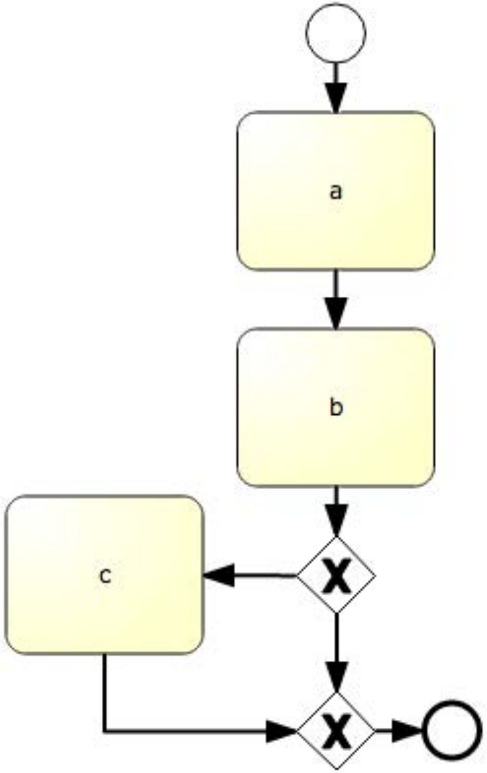
Pages: 1 2 3 4 5 6 ... 31 (30 rows per page)

Business Process Modeling and Notation (BPMN) 2.0

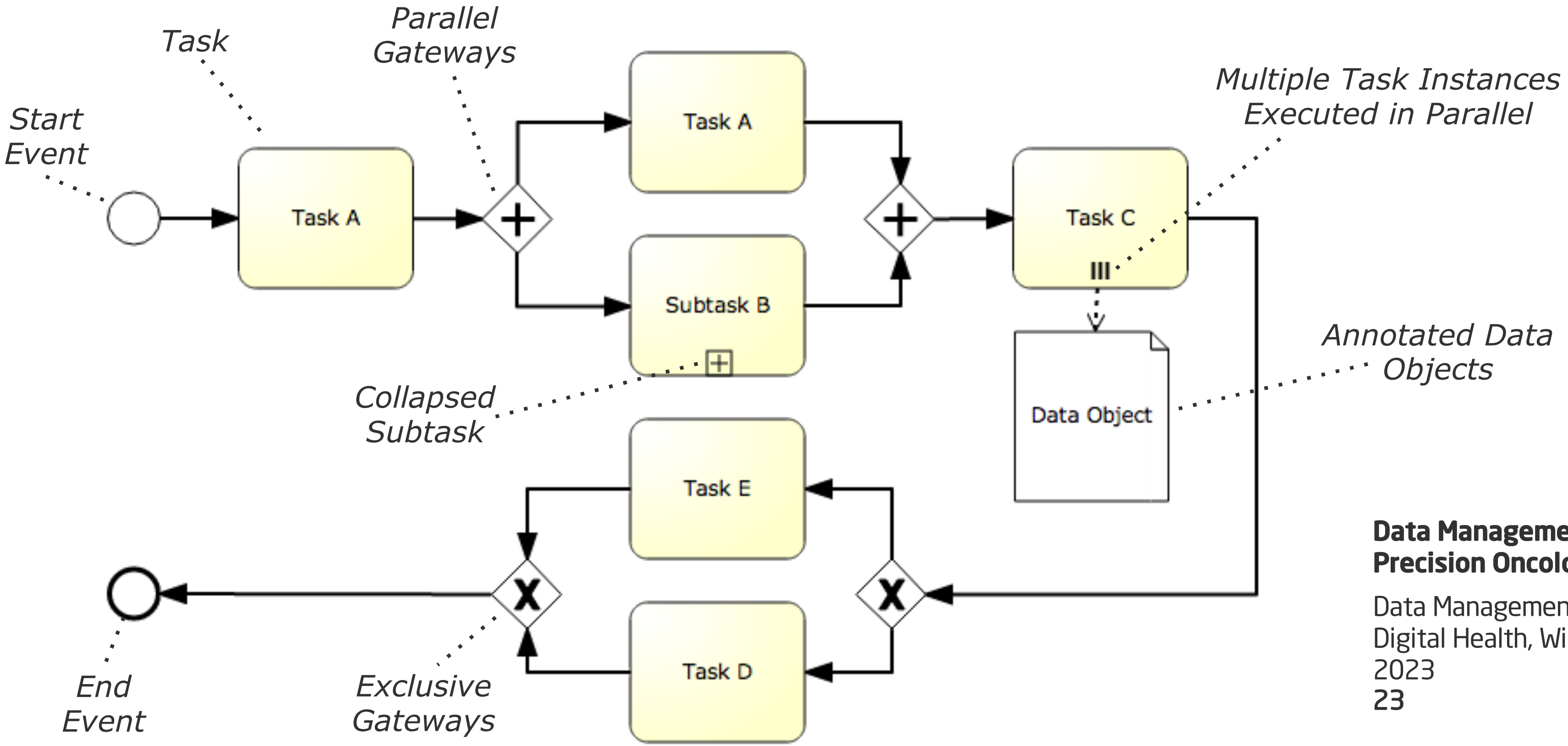
- Used for functional modeling of business processes and workflows
- Graphical notation addresses business and technical users → intuitive modeling and understanding
- Can be serialized and exchanged using XML Process Definition Language (XPDL)



```
<?xml version="1.0" encoding="UTF-8"?>
<zdef-2030967014:Package xmlns="" xmlns:xpdExt="http://www.tibco.com/XPD/xpdExtens
  <zdef-2030967014:ConformanceClass GraphConformance="NON-BLOCKED" BPMNModelPortab
  <zdef-2030967014:Script Type="http://www.w3.org/1999/XPath"/>
  <Pools xmlns="http://www.wfmc.org/2008/XPDL2.1">
    <Pool BoundaryVisible="false" MainPool="true" Process="MainPool-process" Orien
      <NodeGraphicsInfos>
        <NodeGraphicsInfo FillColor="#ffffff" Height="0.0" Width="0.0" BorderColo
          <Coordinates XCoordinate="0.0" YCoordinate="0.0"/>
        </NodeGraphicsInfo>
      </NodeGraphicsInfos>
    </Pool>
  </Pools>
  <WorkflowProcesses xmlns="http://www.wfmc.org/2008/XPDL2.1">
    <WorkflowProcess AdhocOrdering="Sequential" ProcessType="None" Status="None" S
      <ActivitySets>
        <ActivitySet AdHocOrdering="Sequential" Id="sid-A846876F-9749-41F9-93DE-6C
        <ActivitySet AdHocOrdering="Sequential" Id="sid-10D16CD8-AAEF-4694-A5C4-75
      </ActivitySets>
```

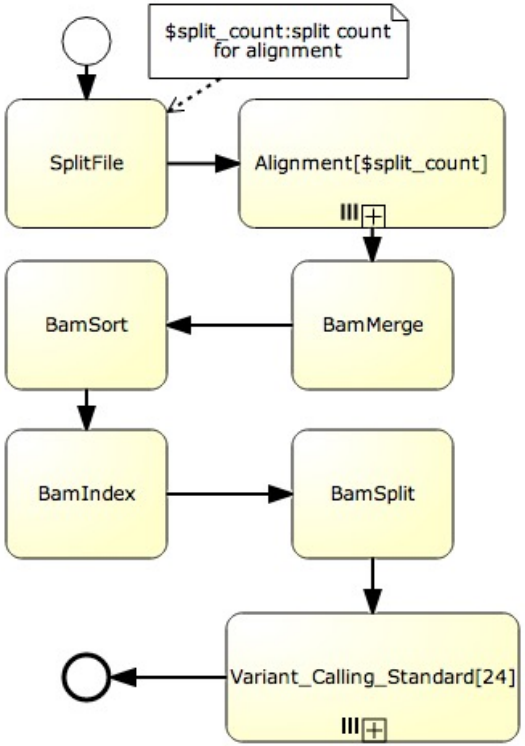


BPMN 2.0: Cheat Sheet
Basic Notation Overview



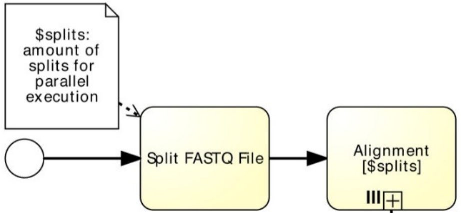
Graphical Modeling of Genome Data Processing Pipelines

- Graphical modeling notation extends BPMN 2.0:
  - □ Modular structure
  - □ Parallelization annotations
  - □ Parameters and variables

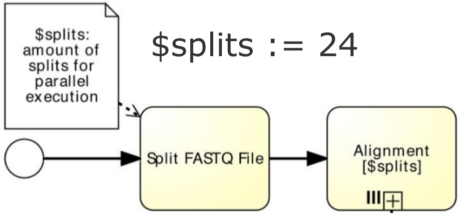

$split_count:split count for alignment
SplitFile
Alignment[$split_count]
BamSort
BamMerge
BamIndex
BamSplit
Variant_Calling_Standard[24]

Model vs. Model Instance

- **Model** := template for multiple instances, e.g. general description of all alignment processes

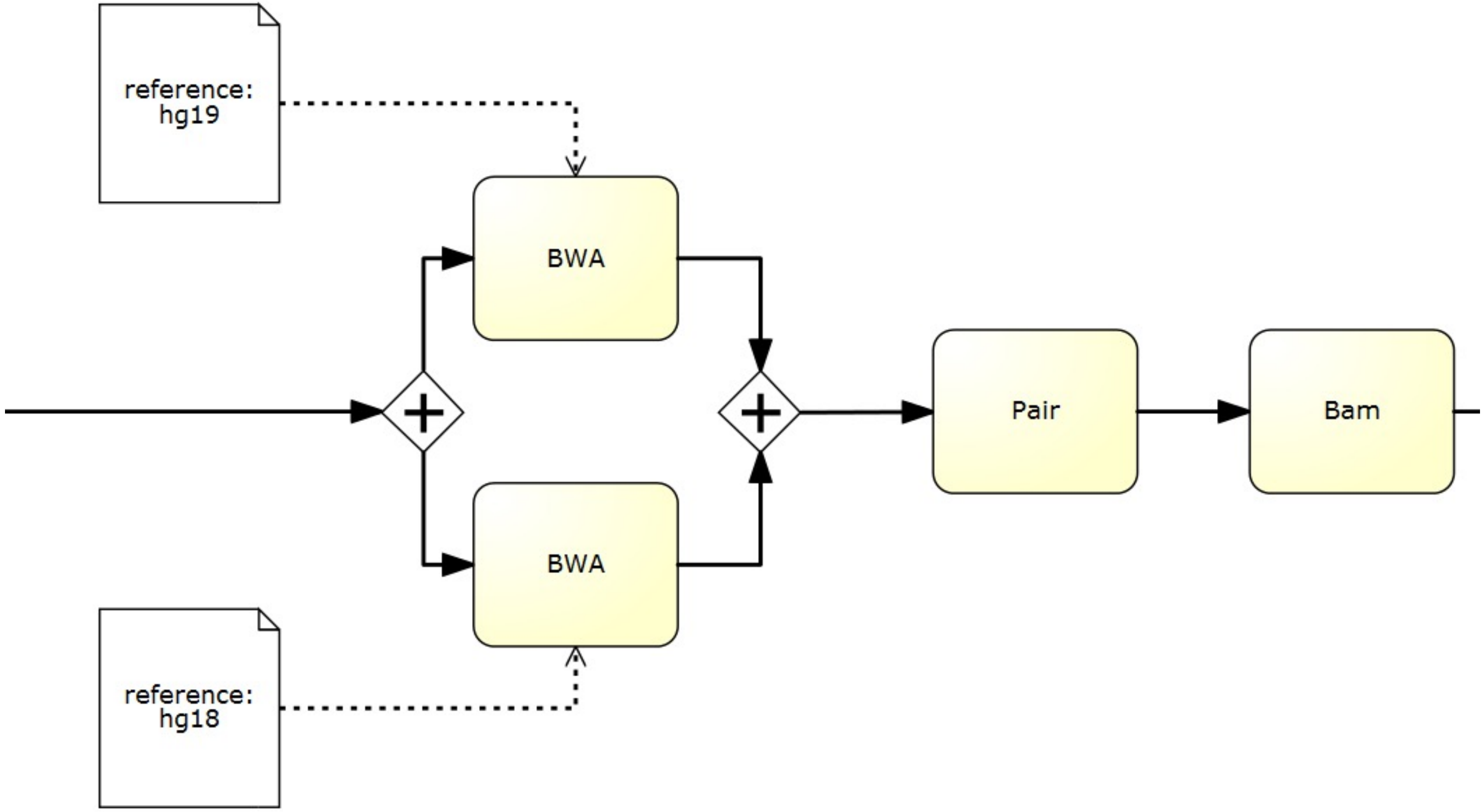


- **Model instance** := specific instance of a model, e.g. configured for a set of specific runs.

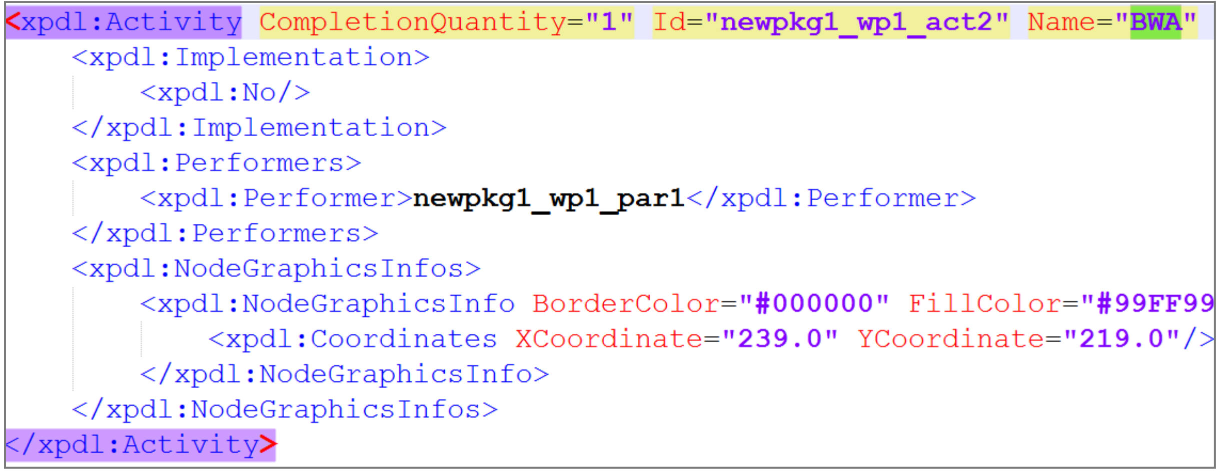


- Models and model instances are stored within the IMDB
- Model instances are translated into graph structure and executed by a dedicated runtime environment

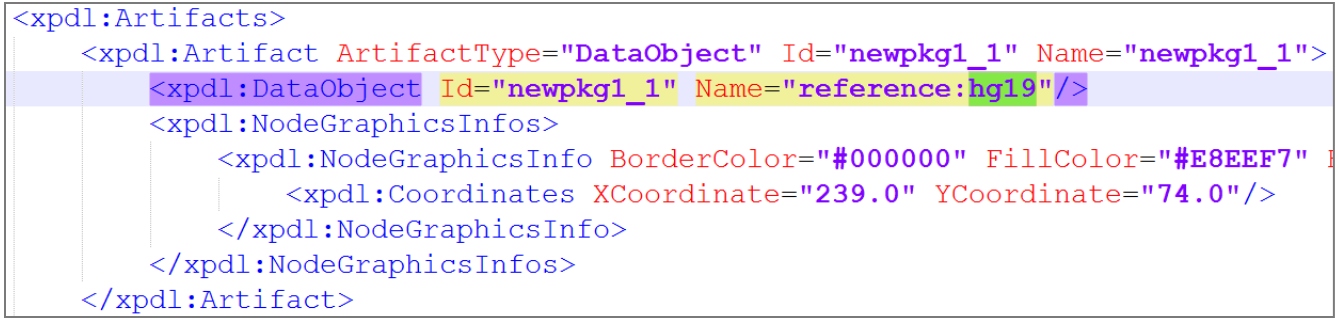
BPMN Example



Persisting Pipelines
XML Process Definition Language



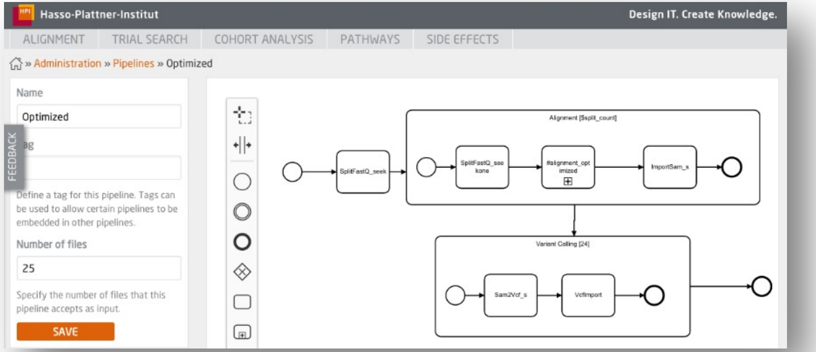
```
<xpdl:Activity CompletionQuantity="1" Id="newpkg1_wp1_act2" Name="BWA"
    <xpdl:Implementation>
        <xpdl:No/>
    </xpdl:Implementation>
    <xpdl:Performers>
        <xpdl:Performer>newpkg1_wp1_par1</xpdl:Performer>
    </xpdl:Performers>
    <xpdl:NodeGraphicsInfos>
        <xpdl:NodeGraphicsInfo BorderColor="#000000" FillColor="#99FF99
            <xpdl:Coordinates XCoordinate="239.0" YCoordinate="219.0"/>
        </xpdl:NodeGraphicsInfo>
    </xpdl:NodeGraphicsInfos>
</xpdl:Activity>
```



```
<xpdl:Artifacts>
    <xpdl:Artifact ArtifactType="DataObject" Id="newpkg1_1" Name="newpkg1_1">
        <xpdl:DataObject Id="newpkg1_1" Name="reference:hg19"/>
        <xpdl:NodeGraphicsInfos>
            <xpdl:NodeGraphicsInfo BorderColor="#000000" FillColor="#E8EEF7"
                <xpdl:Coordinates XCoordinate="239.0" YCoordinate="74.0"/>
            </xpdl:NodeGraphicsInfo>
        </xpdl:NodeGraphicsInfos>
    </xpdl:Artifact>
```

What to take home?

- Use of standardized modeling tools supports implementation / exchange
- Graphical modeling facilitates understanding (also for non-professionals)
- Process modeling is the foundation for reproducibility of results and scalable use



» Administration » Pipelines » Optimized

Name: Optimized

Tag:

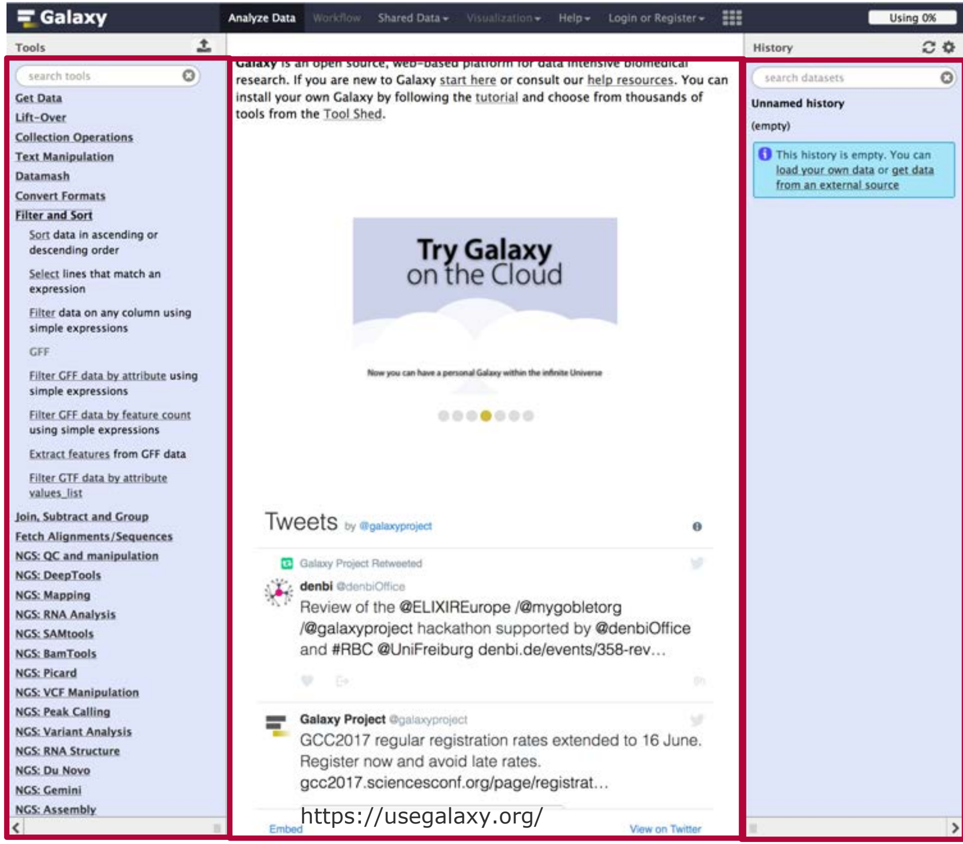
Define a tag for this pipeline. Tags can be used to allow certain pipelines to be embedded in other pipelines.

Number of files: 25

Specify the number of files that this pipeline accepts as input.

Galaxy Workbench

- Open-source, web-based platform
- Supports data-intensive research
- Focuses on process automation and high-throughput sequencing



Get Data
Lift-Over
Collection Operations
Text Manipulation
Datamash
Convert Formats
Filter and Sort
Sort data in ascending or descending order
Select lines that match an expression
Filter data on any column using simple expressions
GFF
Filter GFF data by attribute using simple expressions
Filter GFF data by feature count using simple expressions
Extract features from GFF data
Filter GTF data by attribute values_list
Join, Subtract and Group
Fetch Alignments/Sequences
NGS: QC and manipulation
NGS: DeepTools
NGS: Mapping
NGS: RNA Analysis
NGS: SAMtools
NGS: BamTools
NGS: Picard
NGS: VCF Manipulation
NGS: Peak Calling
NGS: Variant Analysis
NGS: RNA Structure
NGS: Du Novo
NGS: Gemini
NGS: Assembly

Galaxy is an open source, web-based platform for data intensive biomedical research. If you are new to Galaxy start here or consult our help resources. You can install your own Galaxy by following the tutorial and choose from thousands of tools from the Tool Shed.

Try Galaxy on the Cloud

Now you can have a personal Galaxy within the infinite Universe

Tweets by @galaxyproject

Galaxy Project Retweeted

denbi @denbiOffice
Review of the @ELIXIREurope /@mygobletorg /@galaxyproject hackathon supported by @denbiOffice and #RBC @UniFreiburg denbi.de/events/358-rev…

Galaxy Project @galaxyproject
GCC2017 regular registration rates extended to 16 June. Register now and avoid late rates. gcc2017.sciencesconf.org/page/registrat…

Embed
View on Twitter

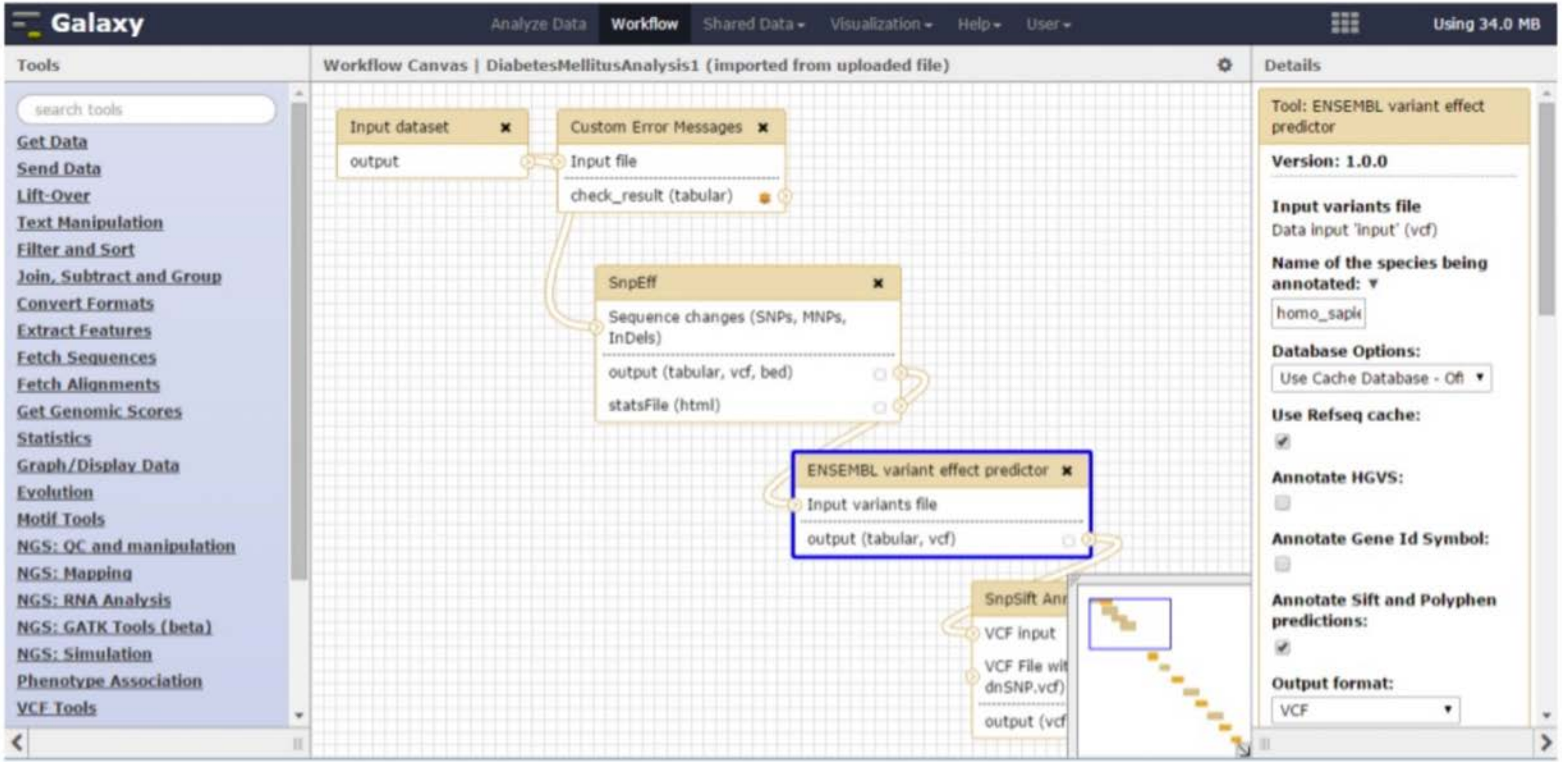
Unnamed history

(empty)

This history is empty. You can load your own data or get data from an external source

https://usegalaxy.org/

Galaxy
Workflow Modeling



Tool: ENSEMBL variant effect predictor

Version: 1.0.0

Input variants file
Data input 'input' (vcf)

Name of the species being annotated: ▾
homo_sapie

Database Options:
Use Cache Database - Off ▾

Use Refseq cache:
☑

Annotate HGVS:
□

Annotate Gene Id Symbol:
□

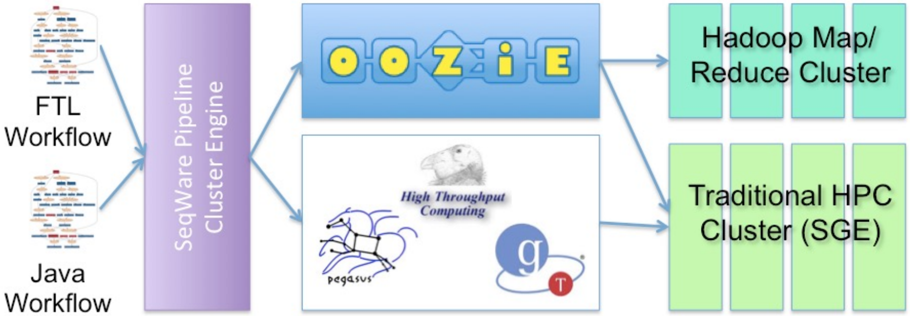
Annotate Sift and Polyphen predictions:
☑

Output format:
VCF ▾

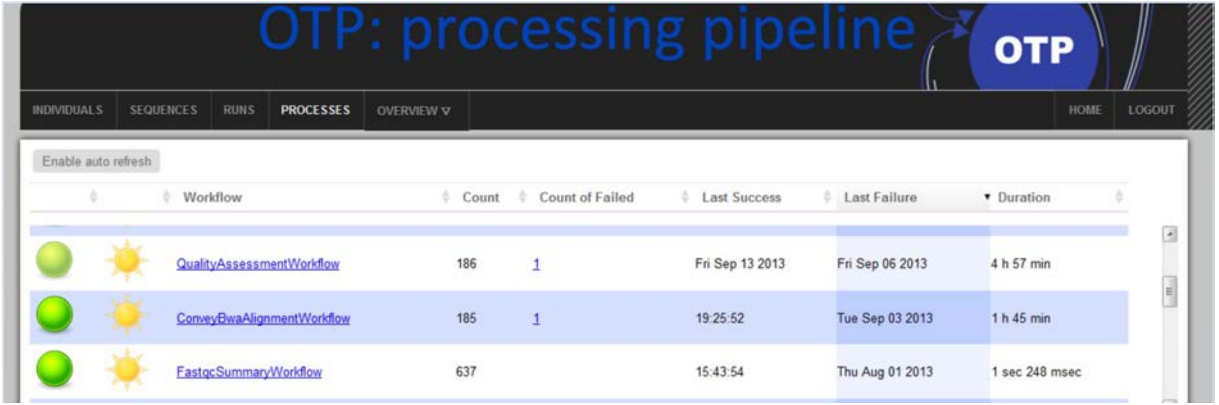
https://usegalaxy.org/

DKFZ One Touch Pipeline

- IT process automation at DKFZ, HD
- Builds upon OpenStack to reduce setup time
- Workflow managed by SeqWare Pipeline Manager
- Special-purpose developed for DKFZ requirements


FTL Workflow
Java Workflow
SeqWare Pipeline Cluster Engine
OOZIE
High Throughput Computing
pegasus
Hadoop Map/ Reduce Cluster
Traditional HPC Cluster (SGE)

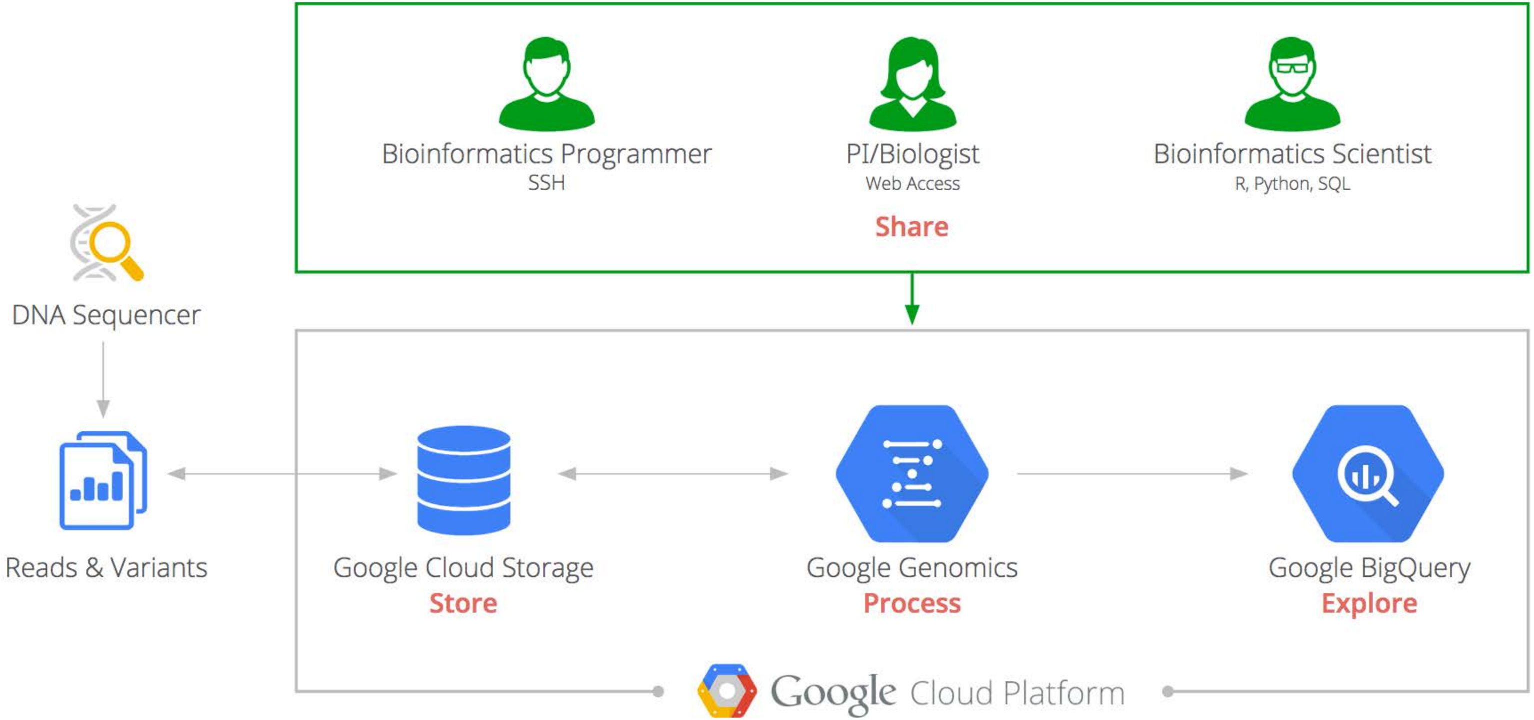
https://seqware.github.io/docs/6-pipeline/



		Workflow	Count	Count of Failed	Last Success	Last Failure	Duration
●	☀	QualityAssessmentWorkflow	186	1	Fri Sep 13 2013	Fri Sep 06 2013	4 h 57 min
●	☀	ConveyBwaAlignmentWorkflow	185	1	19:25:52	Tue Sep 03 2013	1 h 45 min
●	☀	FastqcSummaryWorkflow	637		15:43:54	Thu Aug 01 2013	1 sec 248 msec

Google Genomics

- Integration of existing Google services to genome data processing



https://cloud.google.com/genomics/resources/google-genomics-whitepaper.pdf

The Challenge
Distributed Heterogeneous Data Sources

Human genome/biological data
600GB per full genome
15PB+ in databases of leading institutes

Human proteome
160M data points (2.4GB) per sample
>3TB raw proteome data in ProteomicsDB

Hospital information systems
Often more than 50GB

PubMed database
>23M articles

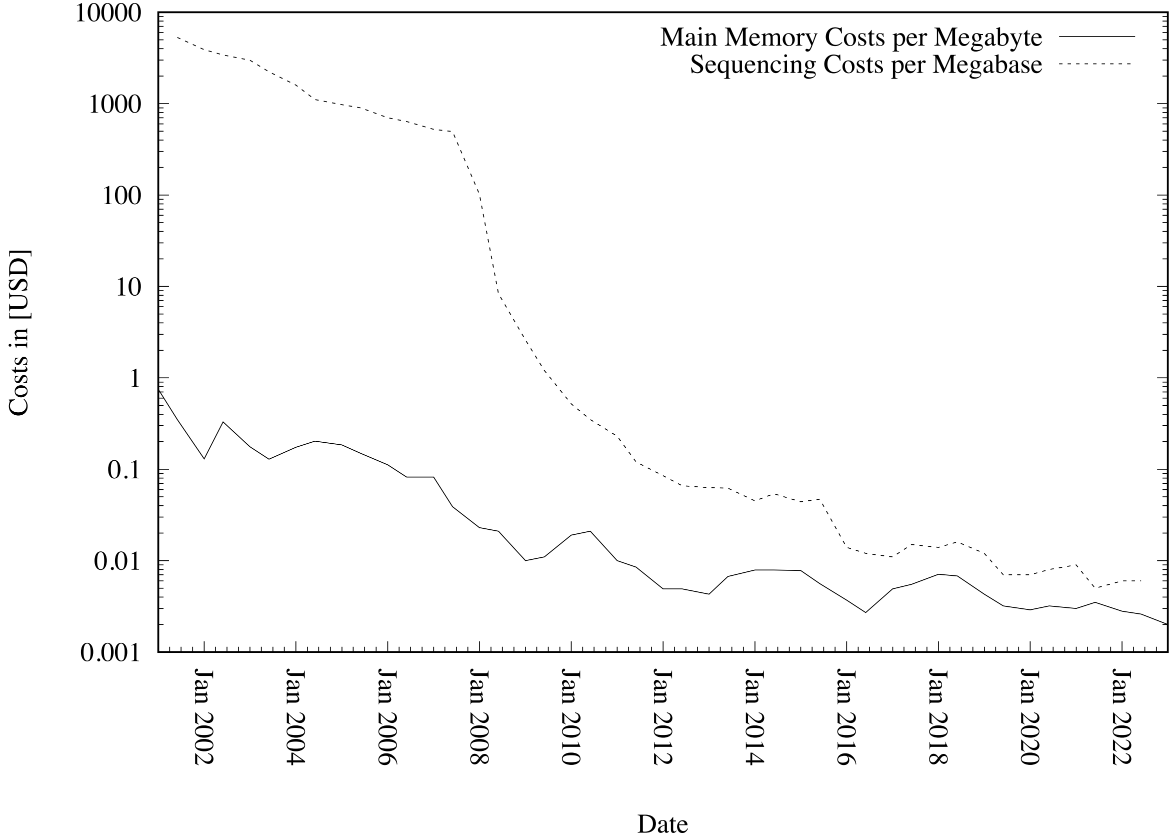
Cancer patient records
>160k records at NCT

Medical sensor data
Scan of a single organ in 1s
creates 10GB of raw data

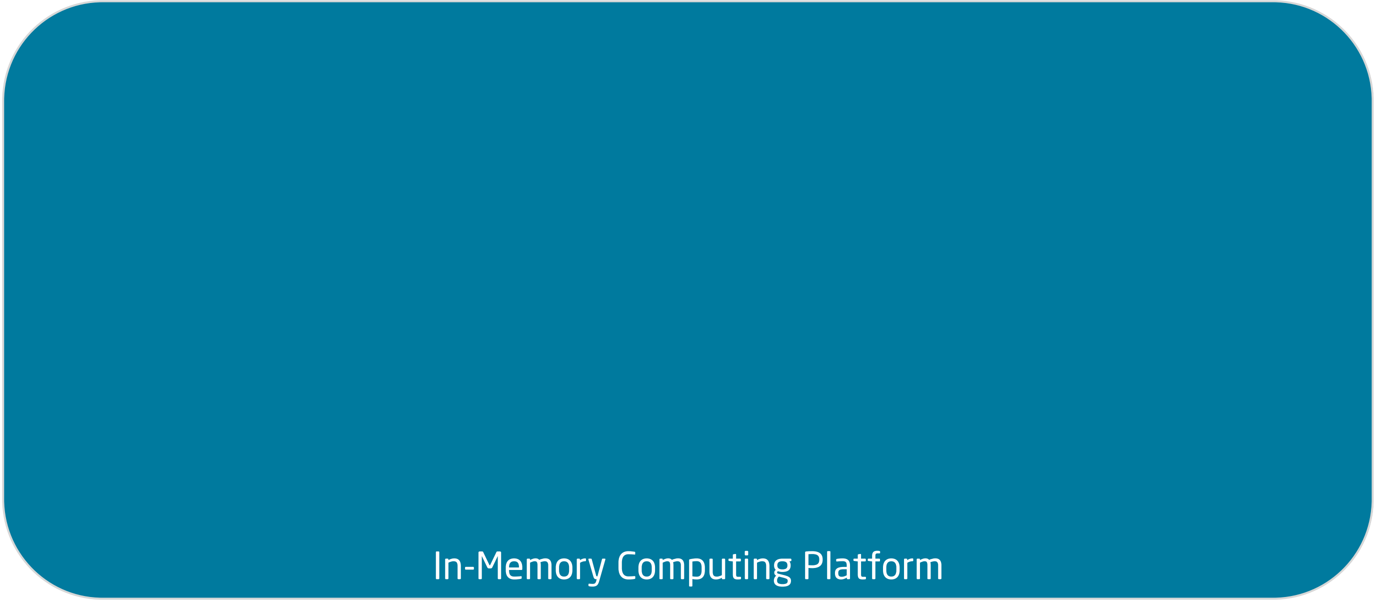
Prescription data
1.5B records from 10,000 doctors and
10M Patients (100 GB)

Clinical trials
Currently more than 30k
recruiting on ClinicalTrials.gov

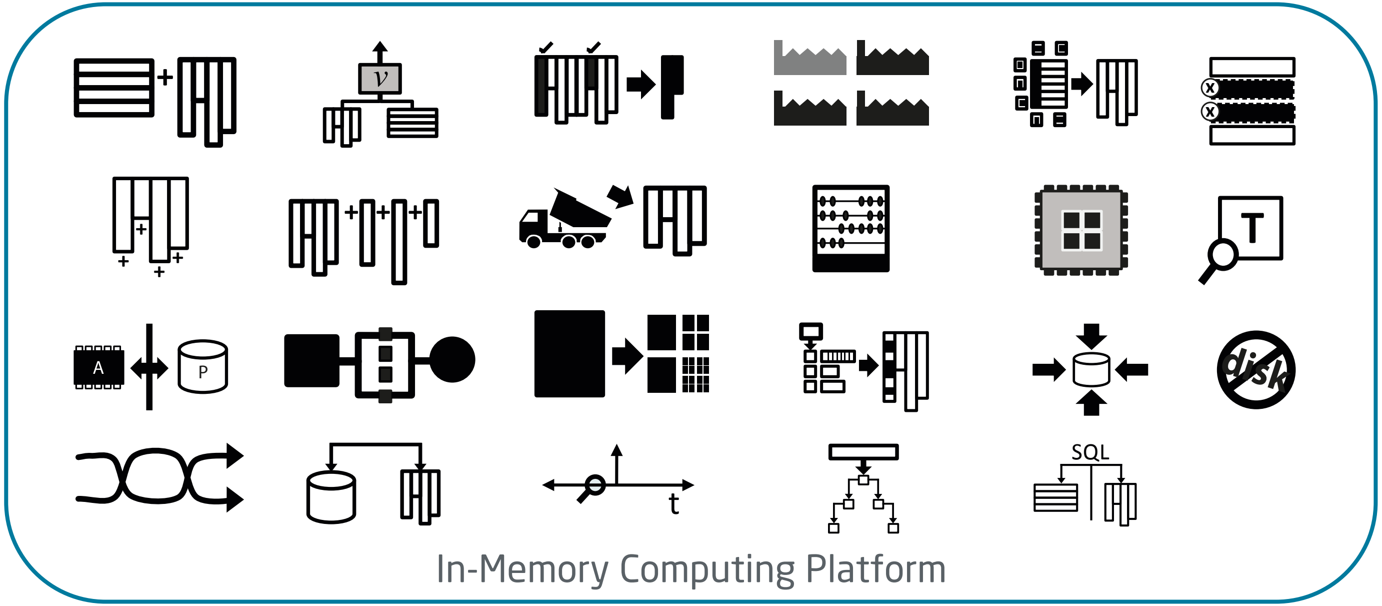
Numbers You Should Know
Comparison of Costs



Our Approach: AnalyzeGenomes.com
In-Memory Computing Platform for Big Medical Data

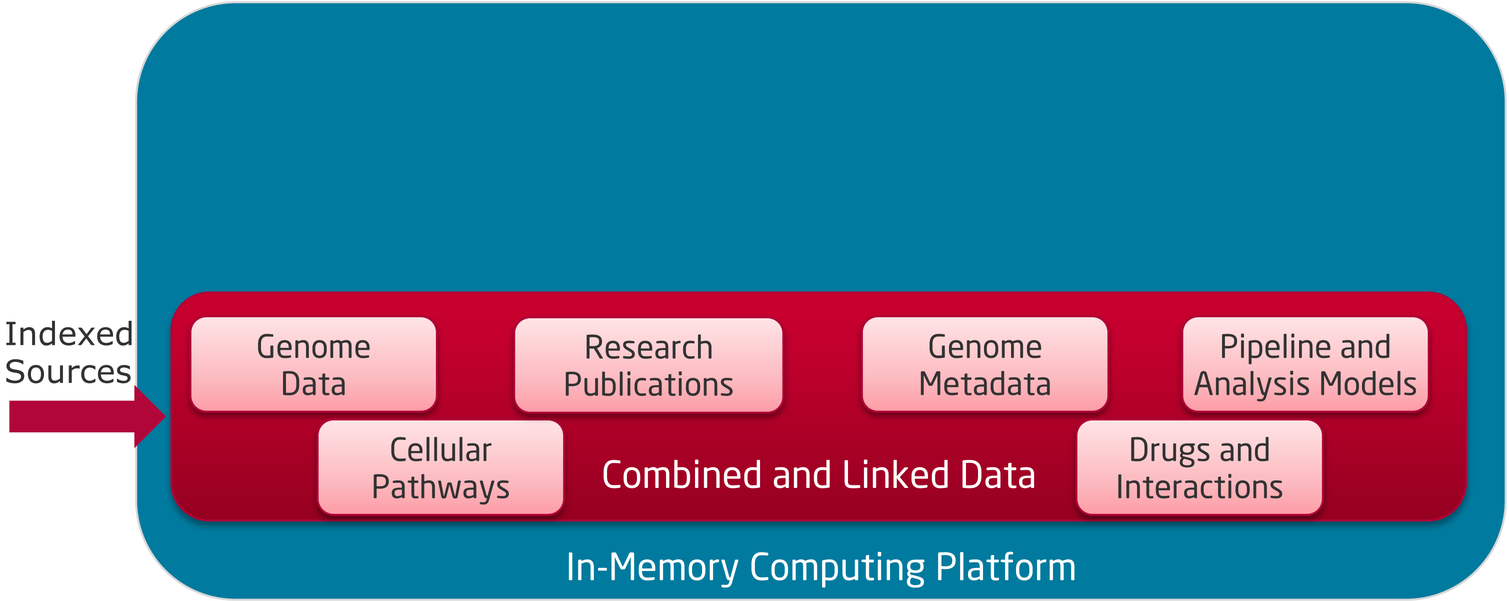


Our Approach: AnalyzeGenomes.com
In-Memory Computing Platform for Big Medical Data

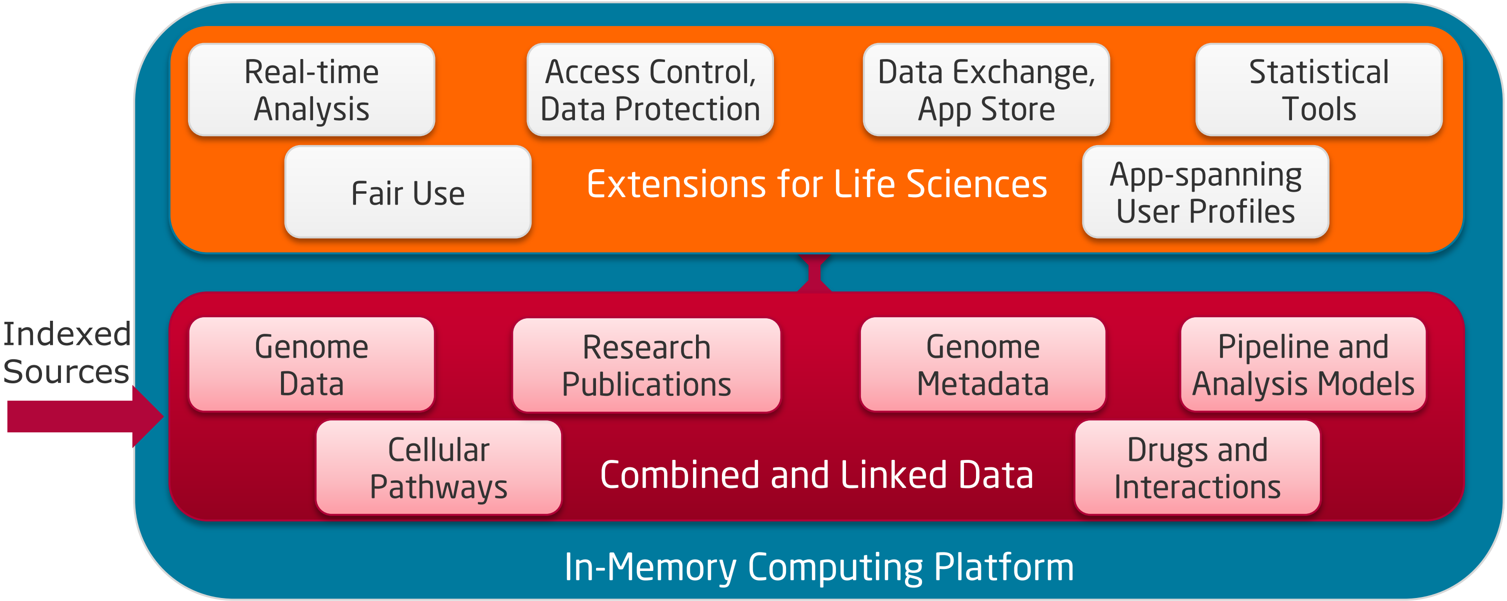


In-Memory Computing Platform

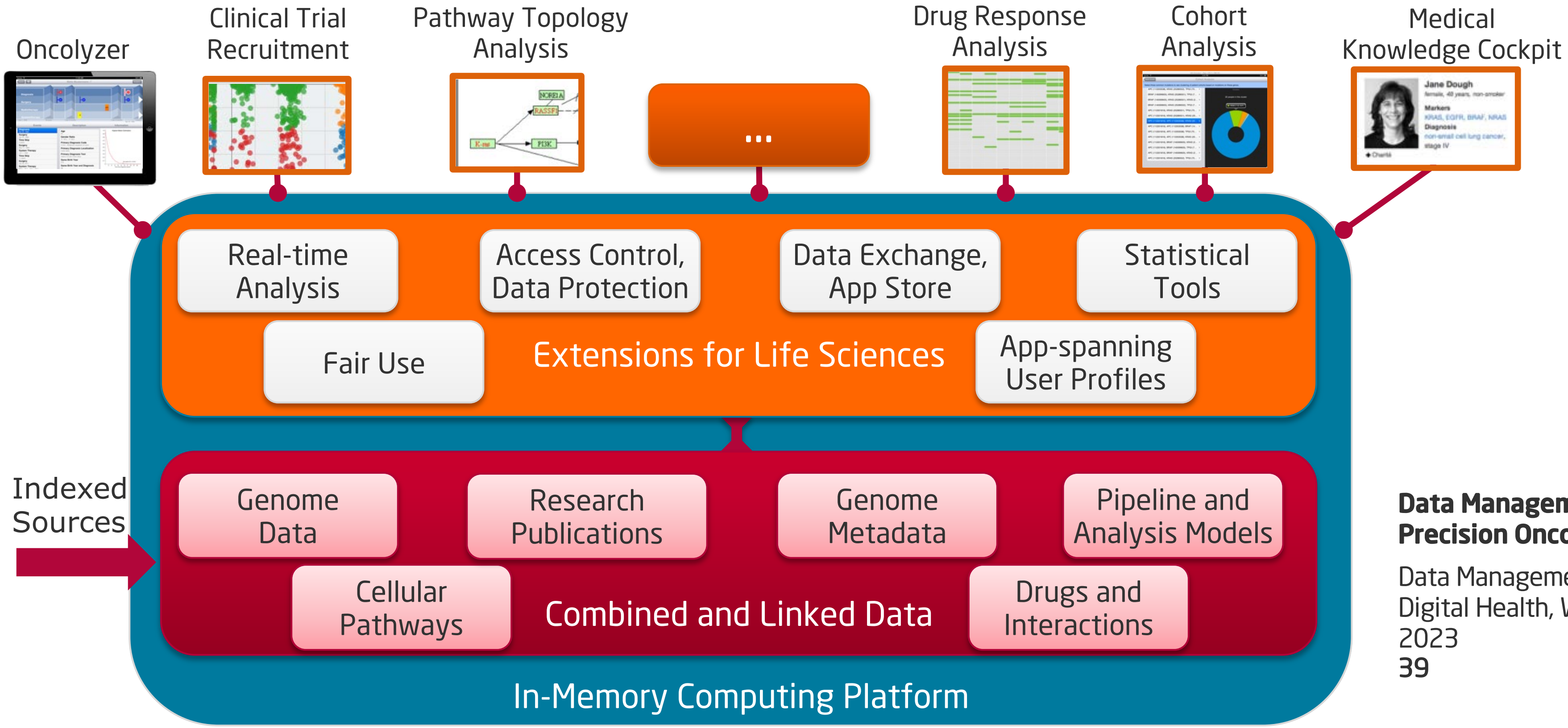
Our Approach: AnalyzeGenomes.com
In-Memory Computing Platform for Big Medical Data



Our Approach: AnalyzeGenomes.com In-Memory Computing Platform for Big Medical Data



Our Approach: AnalyzeGenomes.com
In-Memory Computing Platform for Big Medical Data



Oncolyzer

Clinical Trial Recruitment

Pathway Topology Analysis

...

Drug Response Analysis

Cohort Analysis

Medical Knowledge Cockpit

Real-time Analysis

Access Control, Data Protection

Data Exchange, App Store

Statistical Tools

Fair Use

Extensions for Life Sciences

App-spanning User Profiles

Indexed Sources

Genome Data

Research Publications

Genome Metadata

Pipeline and Analysis Models

Cellular Pathways

Combined and Linked Data

Drugs and Interactions

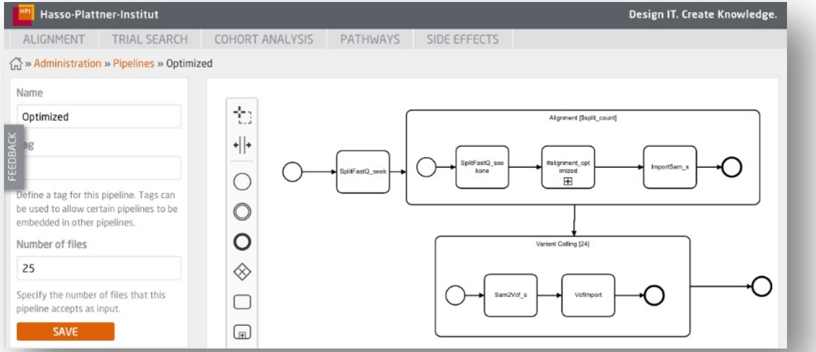
In-Memory Computing Platform

From Model to Execution

1. Design time (researcher, process expert)
   - □ Definition of parameterized process model
   - □ Uses graphical editor and jobs from repository

2. Configuration time (researcher, lab assistant)
   - □ Select model and specify parameters, e.g. aln opts
   - □ Results in model instance stored in repository

3. Execution time (researcher)
   - □ Select model instance
   - □ Specify execution parameters, e.g. input files



» Administration » Pipelines » Optimized

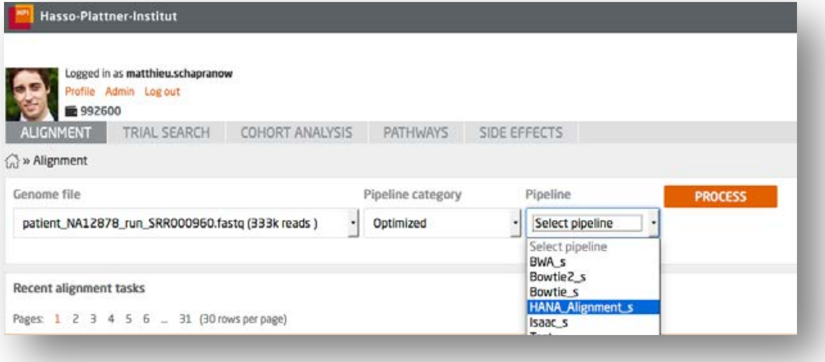
Name: Optimized

ag:

Define a tag for this pipeline. Tags can be used to allow certain pipelines to be embedded in other pipelines.

Number of files: 25

Specify the number of files that this pipeline accepts as input.



Logged in as **matthieu.schapranow**

992600

» Alignment

Genome file: patient_NA12878_run_SRR000960.fastq (333k reads)

Pipeline category: Optimized

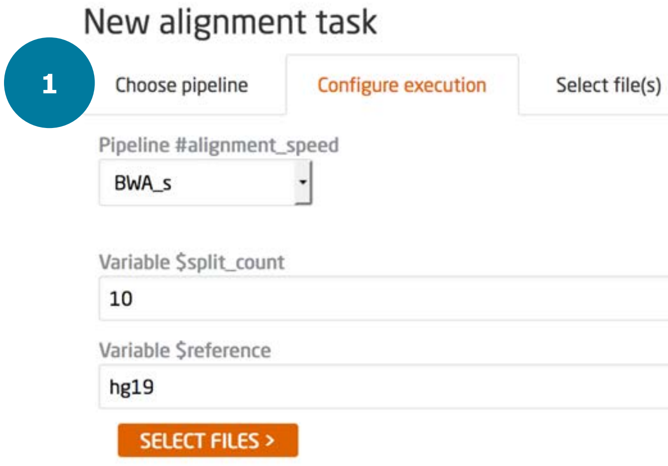
Pipeline: Select pipeline

Recent alignment tasks

Pages: 1 2 3 4 5 6 ... 31 (30 rows per page)

Execution of Genome Data Processing Pipelines

- Uses workflow, which is…
  - Predefined by a subject-matter expert
  - Preconfigured for a specific run or set of experiments
- Requires only minimal configuration whilst enabling reproducibility

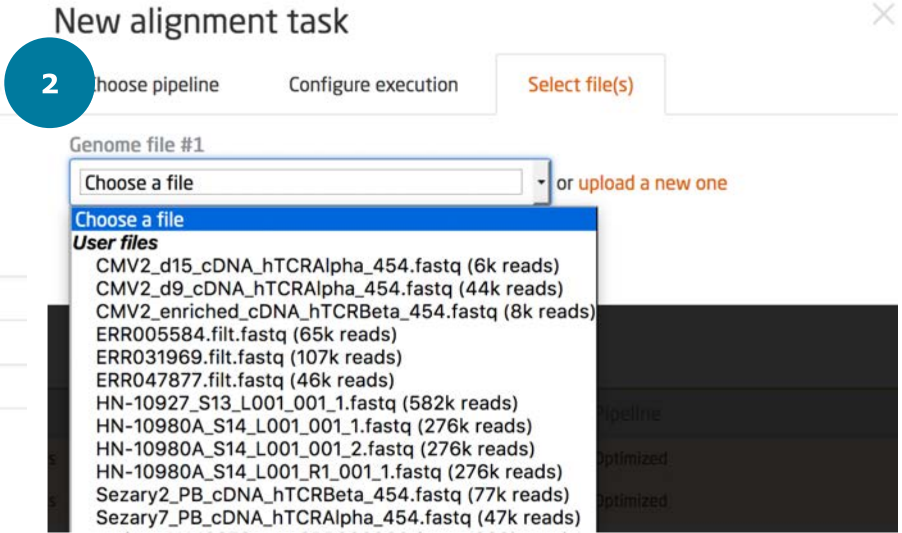


New alignment task

Pipeline #alignment_speed: BWA_s

Variable $split_count: 10

Variable $reference: hg19



New alignment task

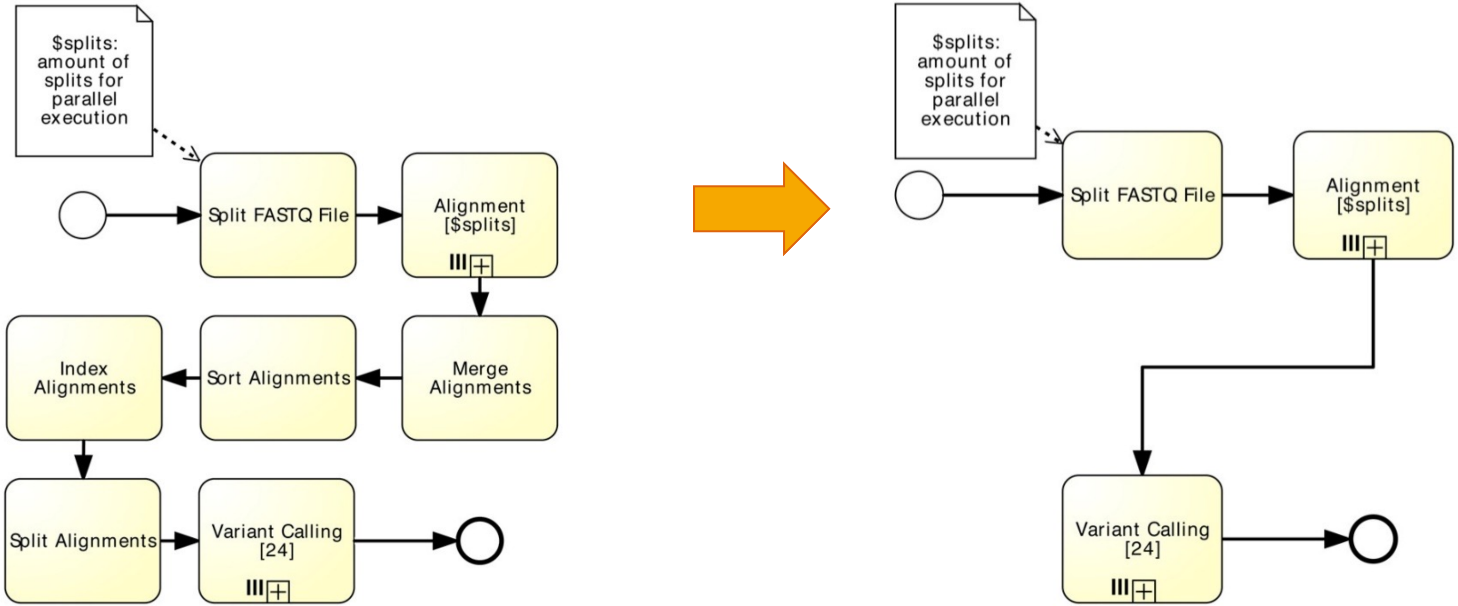
Genome file #1: Choose a file or upload a new one

Choose a file
User files
CMV2_d15_cDNA_hTCRAlpha_454.fastq (6k reads)
CMV2_d9_cDNA_hTCRAlpha_454.fastq (44k reads)
CMV2_enriched_cDNA_hTCRBeta_454.fastq (8k reads)
ERR005584.filt.fastq (65k reads)
ERR031969.filt.fastq (107k reads)
ERR047877.filt.fastq (46k reads)
HN-10927_S13_L001_001_1.fastq (582k reads)
HN-10980A_S14_L001_001_1.fastq (276k reads)
HN-10980A_S14_L001_001_2.fastq (276k reads)
HN-10980A_S14_L001_R1_001_1.fastq (276k reads)
Sezary2_PB_cDNA_hTCRBeta_454.fastq (77k reads)
Sezary7_PB_cDNA_hTCRAlpha_454.fastq (47k reads)

Data Management for Precision Oncology

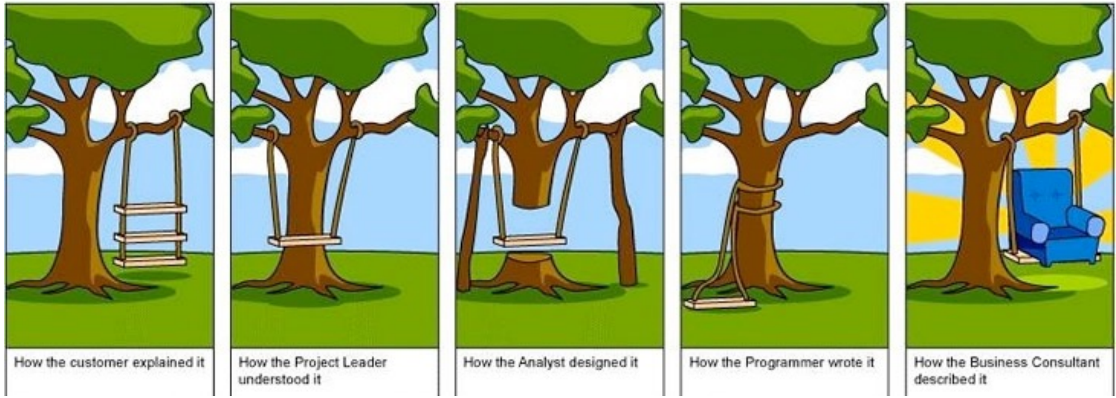
Bridging Gaps: From File-based to IMDB-based GDPPs

- Processing is performed and results are kept within IMDB
- Eliminated media breaks and time-intensive file I/O
- Optimization reduced execution time by >**50%**



Execution Environment for GDPPs
Software Requirements in Life Sciences

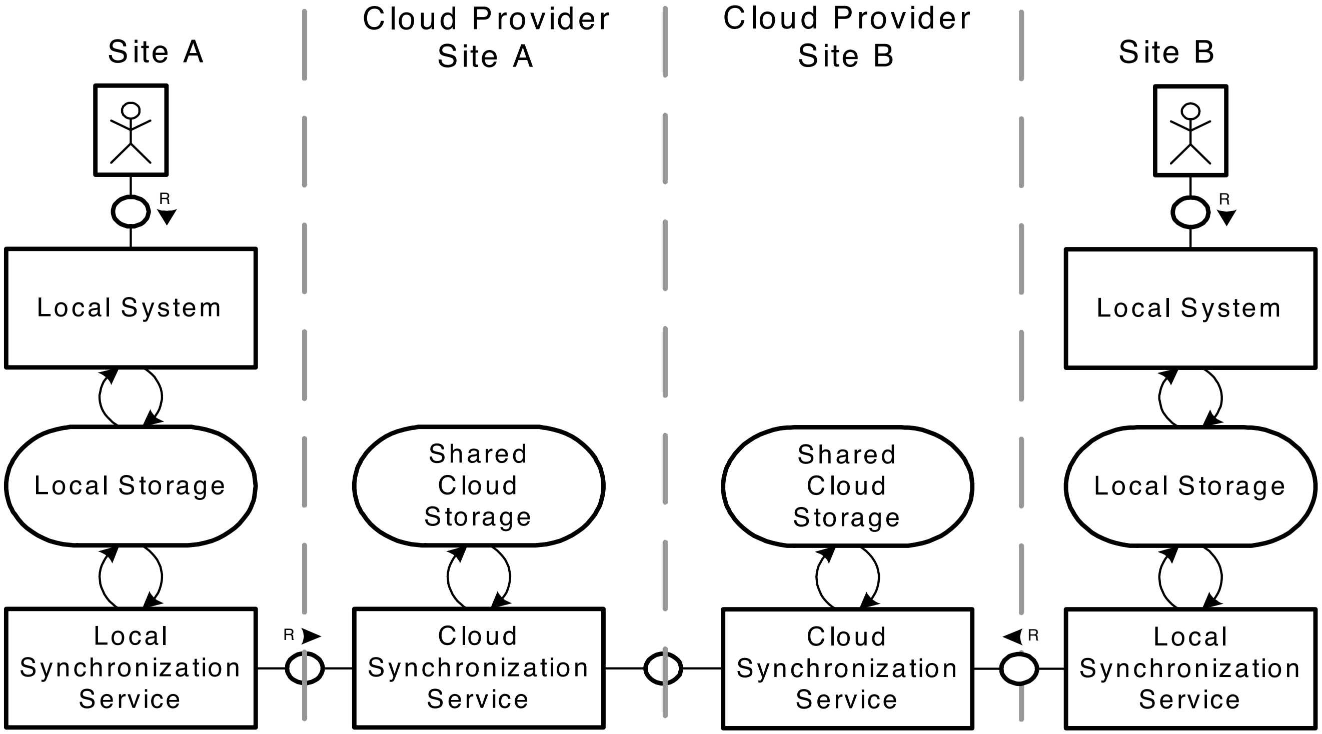
- Requirements
  - □ Managed services
  - □ Reproducibility
  - □ Real-time data analysis of big data
- Restrictions
  - □ Data privacy
  - □ Data locality
  - □ Volume of big medical data
- Solution?
  - □ Federated In-Memory Database System vs. Cloud Computing



How the customer explained it
How the Project Leader understood it
How the Analyst designed it
How the Programmer wrote it
How the Business Consultant described it

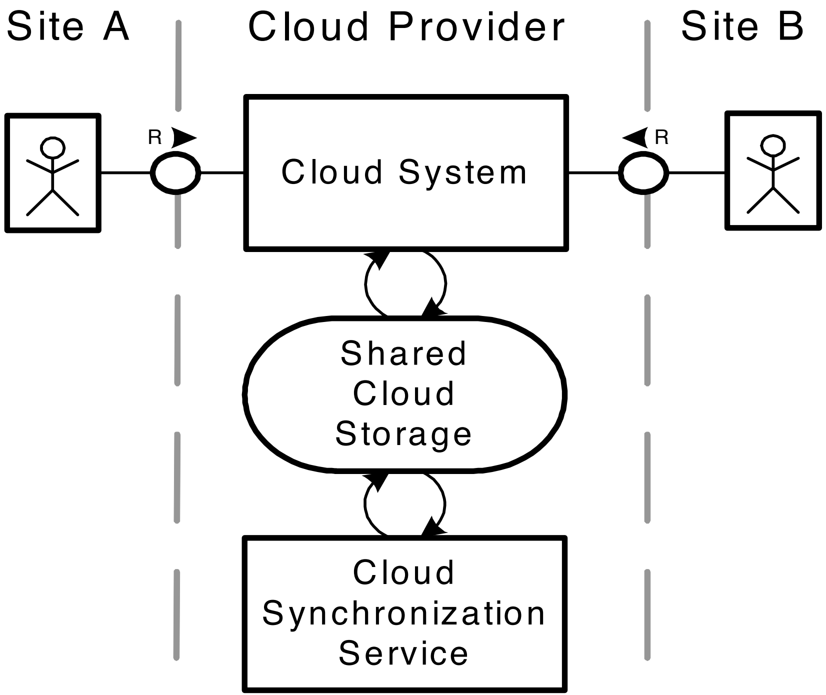
http://stevedempsen.blogspot.de/2013/08/agile-software-requirements-comic.html

Multiple Cloud Service Providers



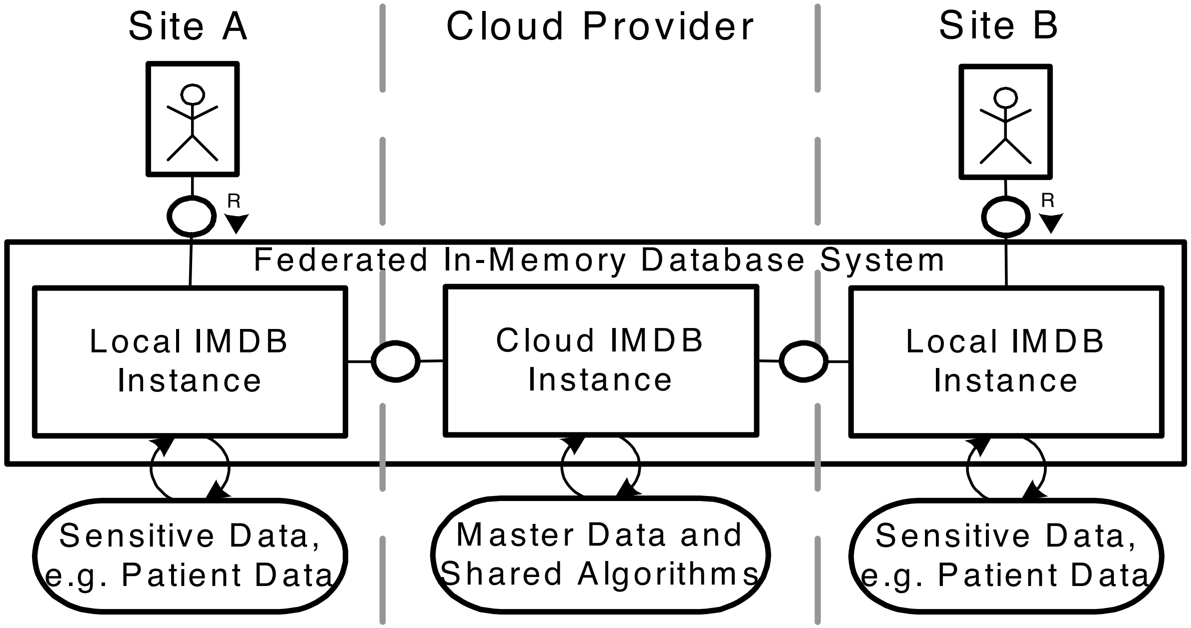
Schapranow, M.-P. et al.: A Federated In-memory Database System for Life Sciences. In: Real-Time Business Intelligence and Analytics. BIRTE 2015, BIRTE 2016, BIRTE 2017. Springer, Cham (2019).

A Single Service Provider



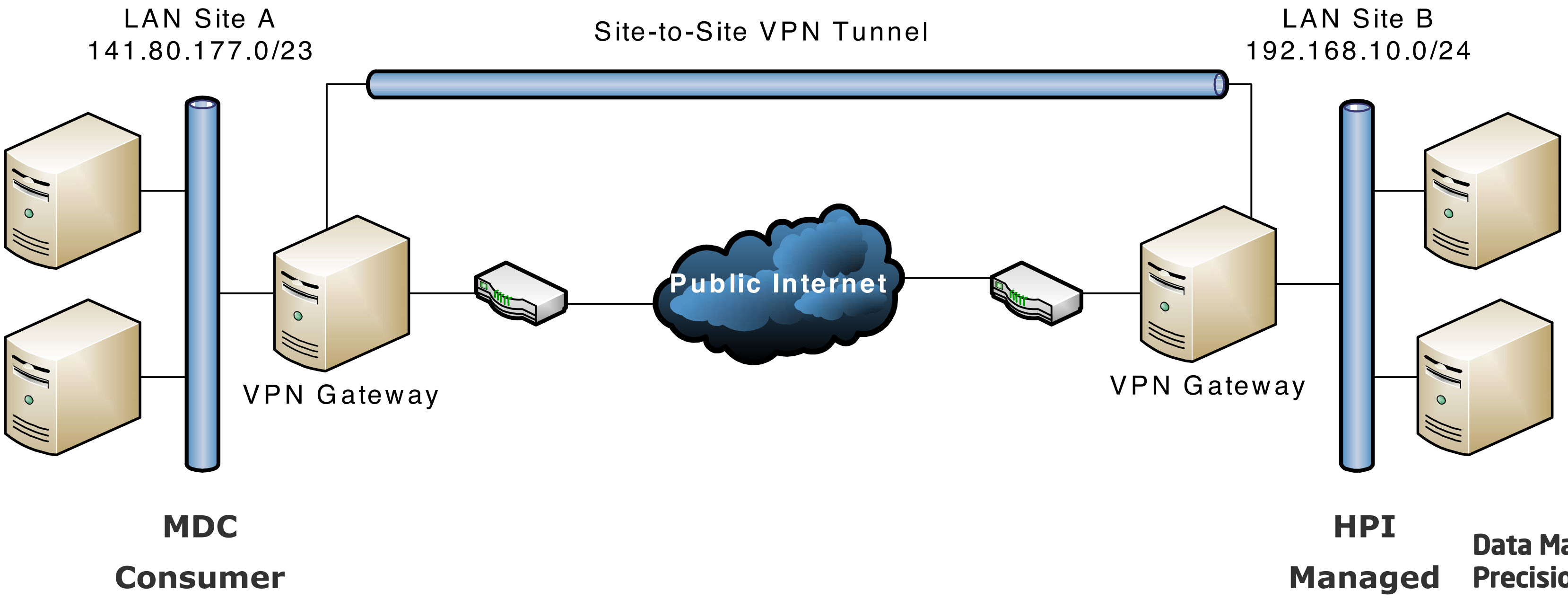
Schapranow, M.-P. et al.: A Federated In-memory Database System for Life Sciences. In: Real-Time Business Intelligence and Analytics. BIRTE 2015, BIRTE 2016, BIRTE 2017. Springer, Cham (2019).

Multiple Sites Forming the Federated In-Memory Database System


Site A
Cloud Provider
Site B
R
R
Federated In-Memory Database System
Local IMDB Instance
Cloud IMDB Instance
Local IMDB Instance
Sensitive Data, e.g. Patient Data
Master Data and Shared Algorithms
Sensitive Data, e.g. Patient Data

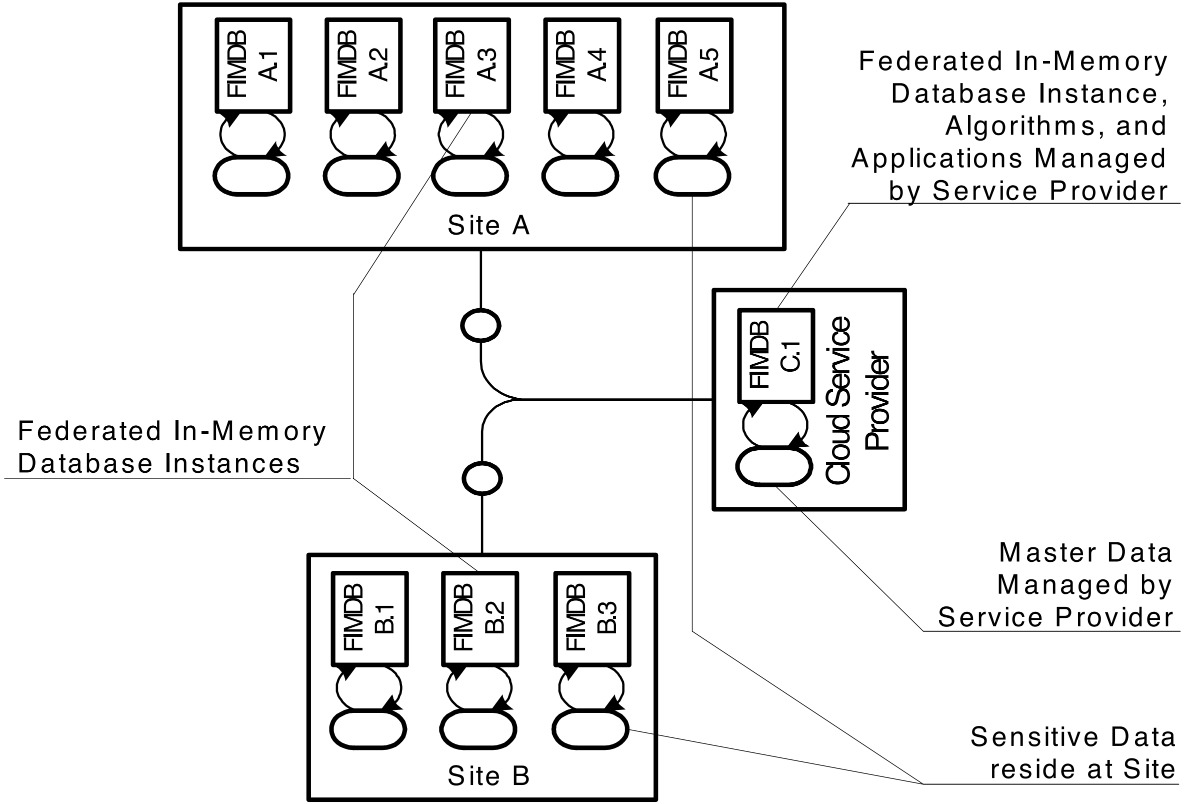
Schapranow, M.-P. et al.: A Federated In-memory Database System for Life Sciences. In: Real-Time Business Intelligence and Analytics. BIRTE 2015, BIRTE 2016, BIRTE 2017. Springer, Cham (2019).

Network Setup
Site-to-Site VPN



Schapranow, M.-P. et al.: A Federated In-memory Database System for Life Sciences. In: Real-Time Business Intelligence and Analytics. BIRTE 2015, BIRTE 2016, BIRTE 2017. Springer, Cham (2019).

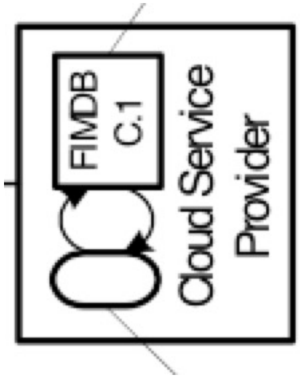
Federated In-Memory Database (FIMDB) Incorporating Local Compute Resources


FIMDB A.1
FIMDB A.2
FIMDB A.3
FIMDB A.4
FIMDB A.5
Site A
Federated In-Memory Database Instance, Algorithms, and Applications Managed by Service Provider
FIMDB C.1
Cloud Service Provider
Federated In-Memory Database Instances
Master Data Managed by Service Provider
FIMDB B.1
FIMDB B.2
FIMDB B.3
Site B
Sensitive Data reside at Site

Schapranow, M.-P. et al.: A Federated In-memory Database System for Life Sciences. In: Real-Time Business Intelligence and Analytics. BIRTE 2015, BIRTE 2016, BIRTE 2017. Springer, Cham (2019).

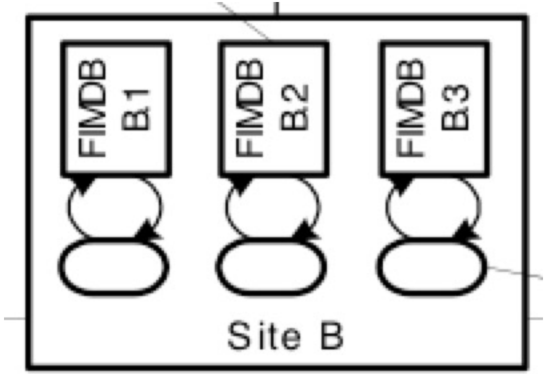
Provided by the Cloud Service Provider

- File System
  - Managed services directory
  - OS binaries statically compiled for individual platforms
- Database
  - In-memory database landscape
  - Stored procedures and database algorithms
  - Master application data



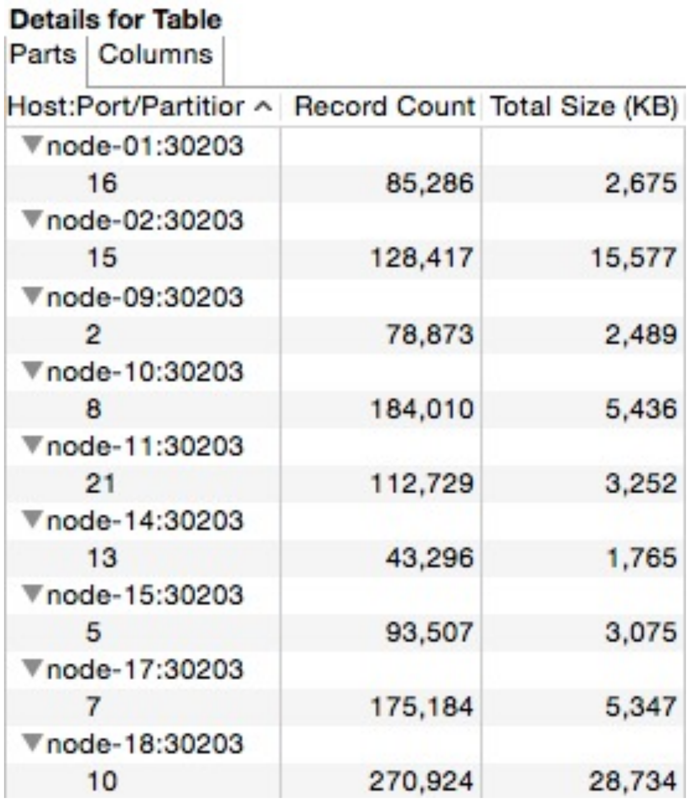
Setup of a New Client

1. Establish site-to-site VPN connection b/w site and cloud service provider
2. Mount remote services directory
3. Install and configure local IMDB instance from services directory
4. Subscribe to and configure selected managed service



Data Partitioning

- Supports parallel query execution
- Protects sensitive data
- Brings algorithms to data



Details for Table

Parts | Columns

Host:Port/Partitior ^	Record Count	Total Size (KB)
▼node-01:30203		
16	85,286	2,675
▼node-02:30203		
15	128,417	15,577
▼node-09:30203		
2	78,873	2,489
▼node-10:30203		
8	184,010	5,436
▼node-11:30203		
21	112,729	3,252
▼node-14:30203		
13	43,296	1,765
▼node-15:30203		
5	93,507	3,075
▼node-17:30203		
7	175,184	5,347
▼node-18:30203		
10	270,924	28,734

NephroCAGE: German-Canadian Consortium on AI for Improved Kidney Transplantation Outcome

- Applying AI technology for improved donor-recipient matching of kidney transplants
- Initial funding period: 2021-2023
- Funding: > 1.5 MEUR
- German partners supported by the German Federal Ministry for Economic Affairs and Climate Action

NephroCAGE:
Nephrology Disease Cooperation between Canada and Germany for Applied AI

Real-world Demonstrator

CHARITÉ UNIVERSITÄTSMEDIZIN BERLIN
UBC
PIRCHE

Learning Systems and Federated Learning

HPI Hasso Plattner Institut
KIT Karlsruhe Institute of Technology

Data Providers and Clinical Experts

CHARITÉ UNIVERSITÄTSMEDIZIN BERLIN
GenomeCanada
GenomeQuébec
McGill University Health Centre
CHUM
UBC
Genome BritishColumbia

NEPHRO CAGE

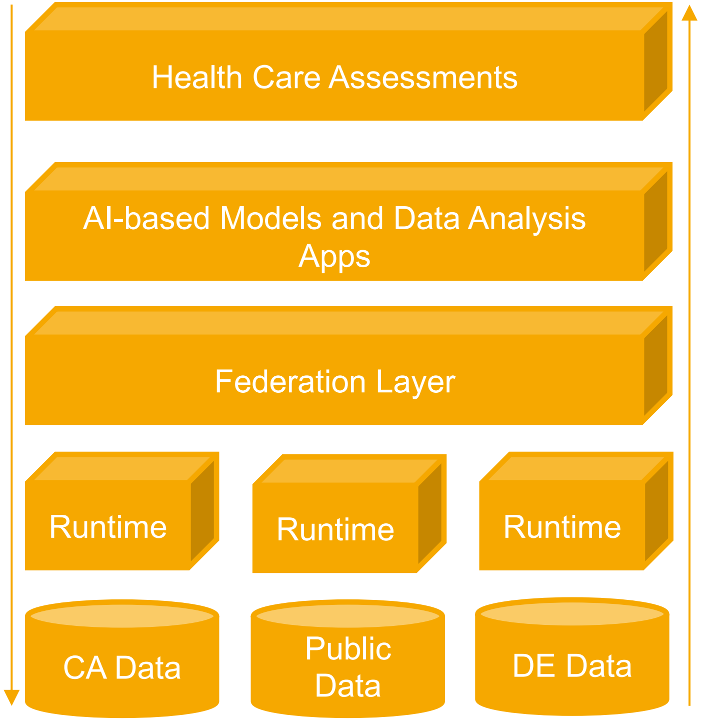
Supported by:

Federal Ministry for Economic Affairs and Climate Action

on the basis of a decision by the German Bundestag

NephroCAGE Federated Learning Software Architecture

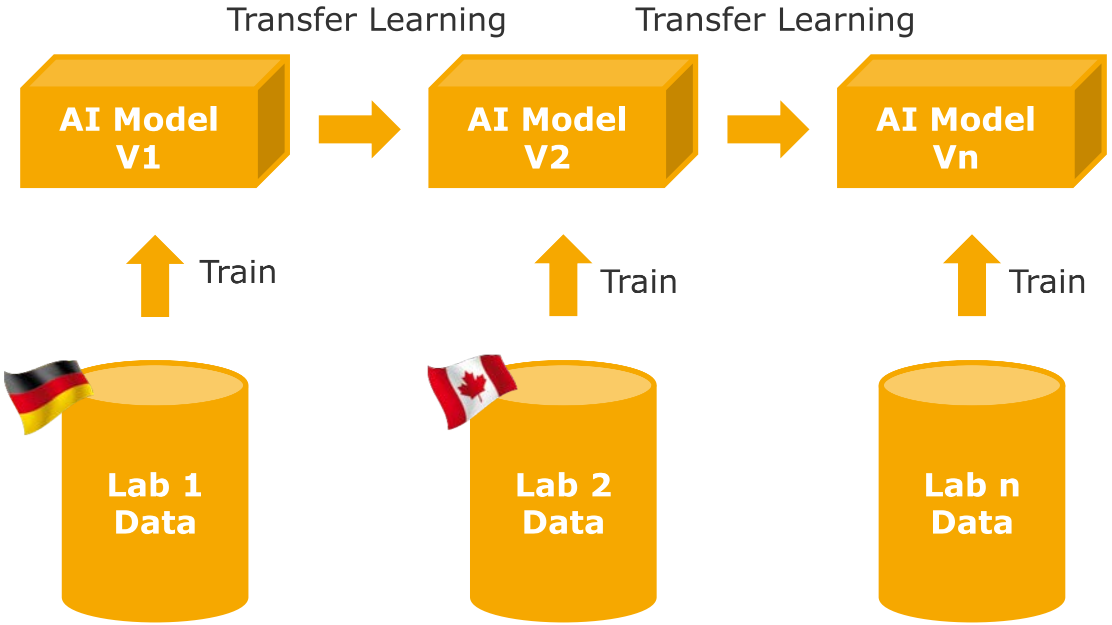
- Assess real-world transplant data from German and Canadian medical centers
- Access to 10yrs+ transplant data
- Healthcare data remains protected
- AI algorithms travel to data
- Federated learning enables data analysis whilst keeping data protected



NephroCAGE Federated Learning Software Architecture (cont'd)

Hasso Plattner Institut

NEPHRO CAGE



https://nephrocage.org

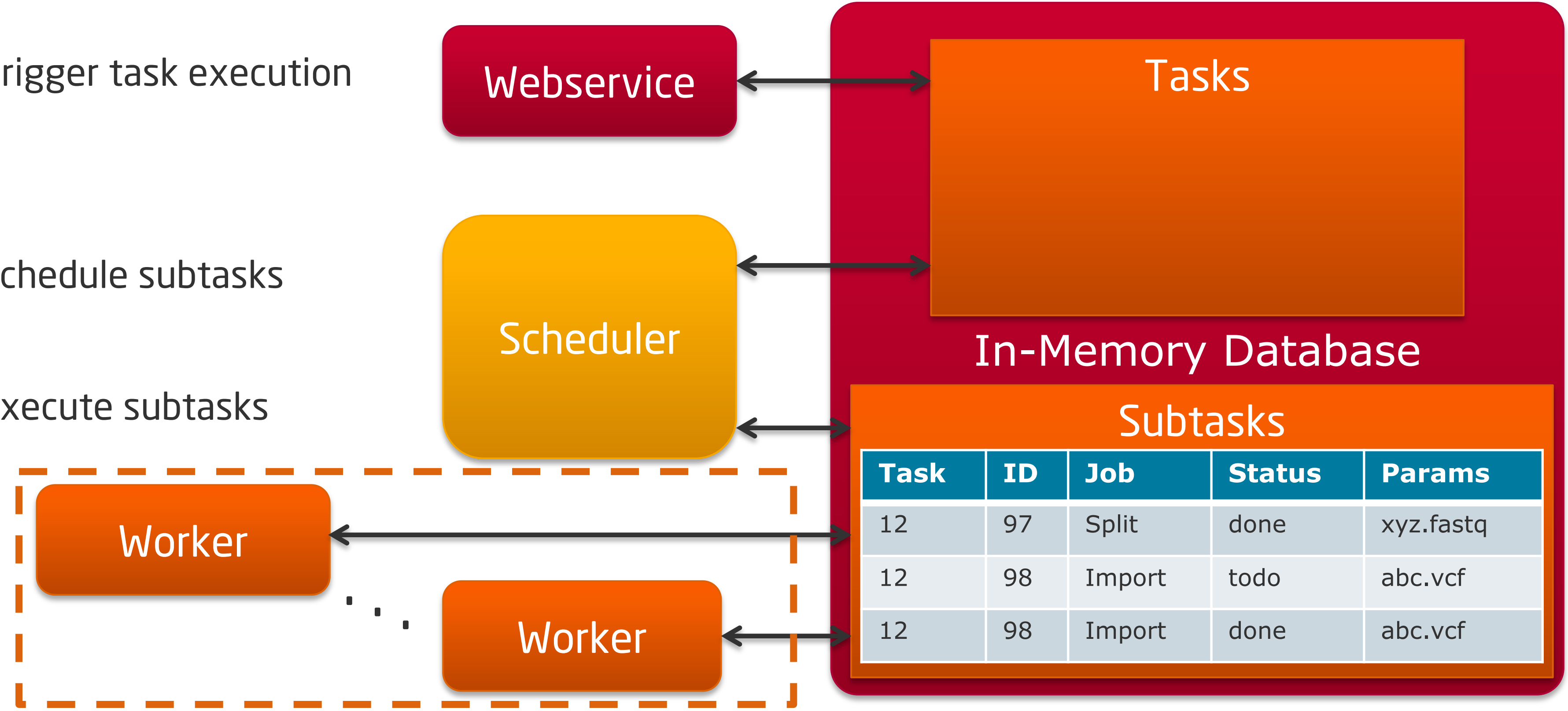
Photo by Louis Reed

Runtime Environment
Scheduling and Execution of GDPPs

1. Trigger task execution

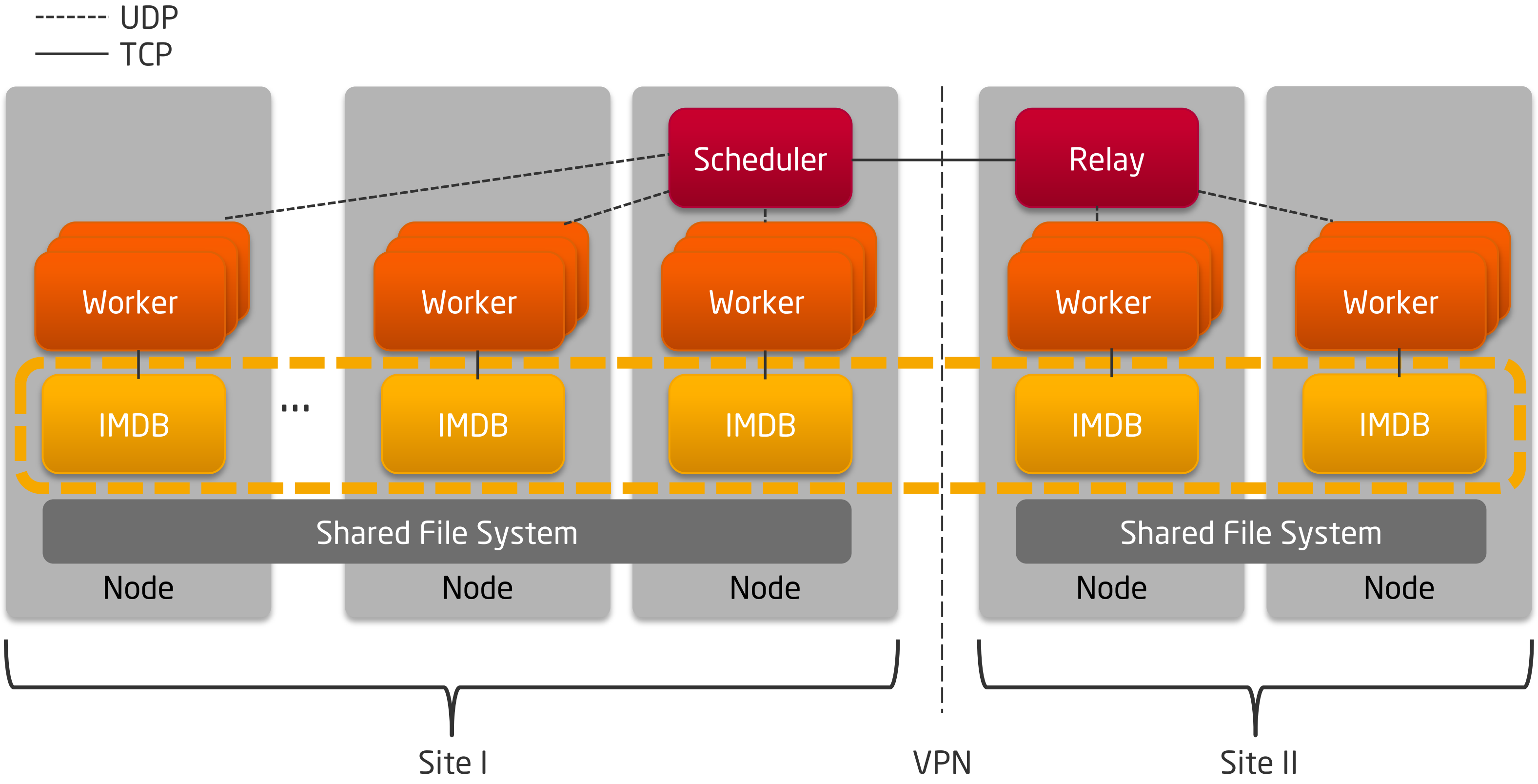
2. Schedule subtasks

3. Execute subtasks



Task	ID	Job	Status	Params
12	97	Split	done	xyz.fastq
12	98	Import	todo	abc.vcf
12	98	Import	done	abc.vcf

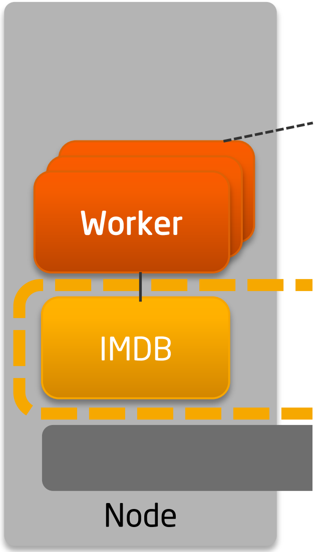
Runtime Environment
Software Components and Communication



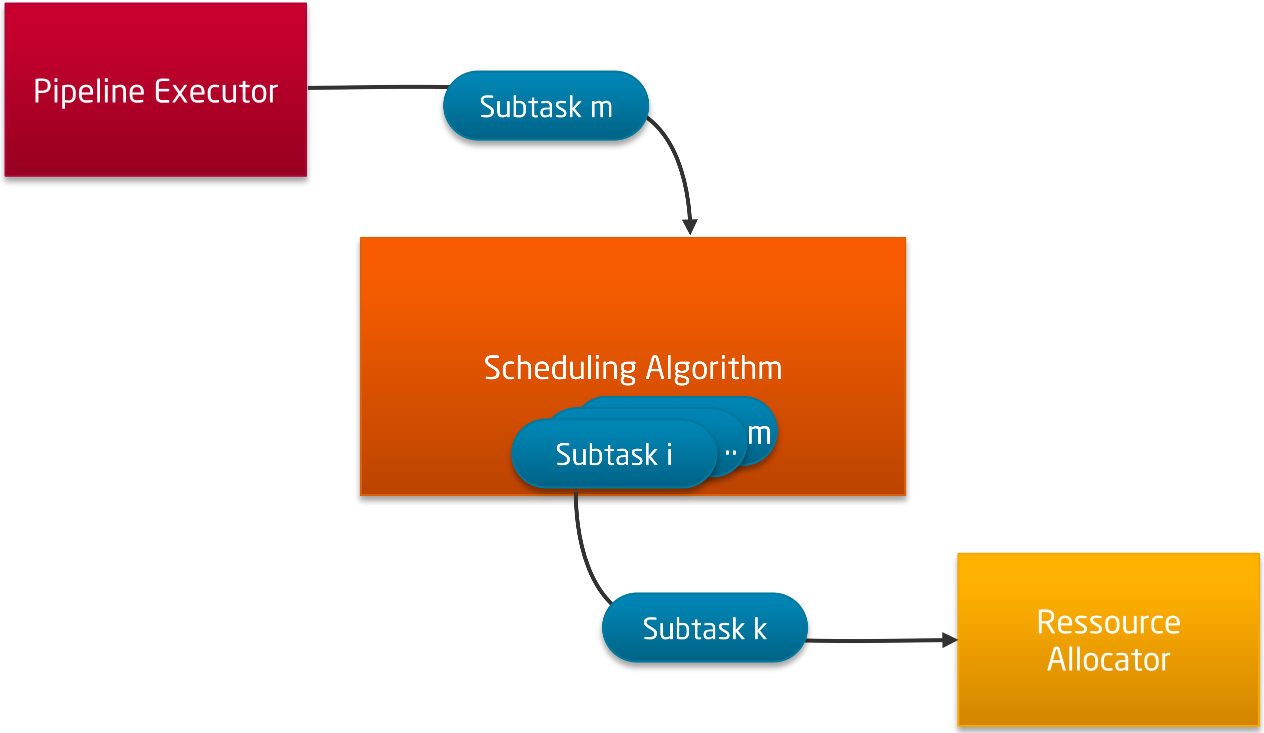
Runtime Environment
Workers

- Workers execute jobs one by one
- Subtask execution status in IMDB:
  - Ready (0),
  - In Progress (1),
  - Done (2), or
  - Erroneous (3).

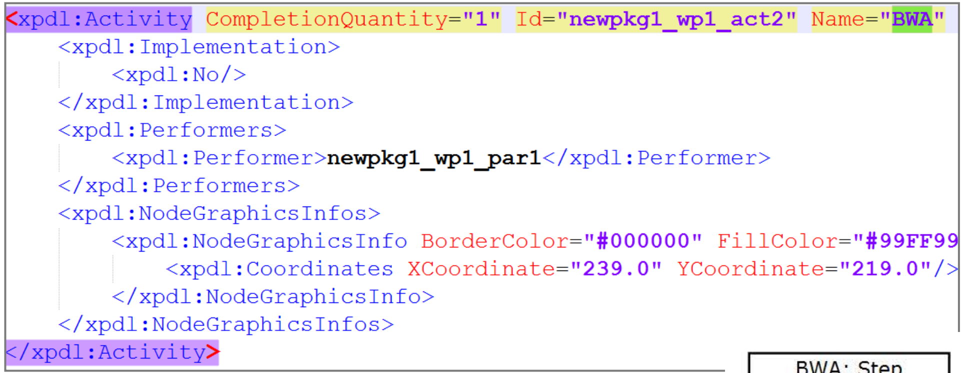
- Jobs implemented as Python modules/classes
  - Can contain arbitrary code
  - Have access to IMDB
  - Can read/write to shared working directory



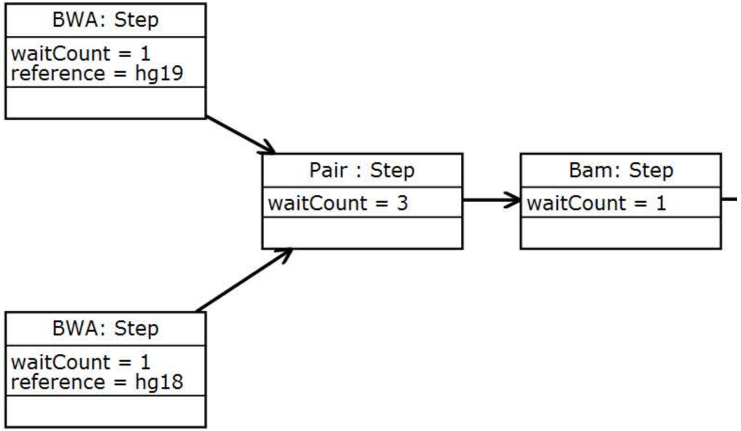
Runtime Environment
Task Scheduler



Runtime Environment
Pipeline Execution

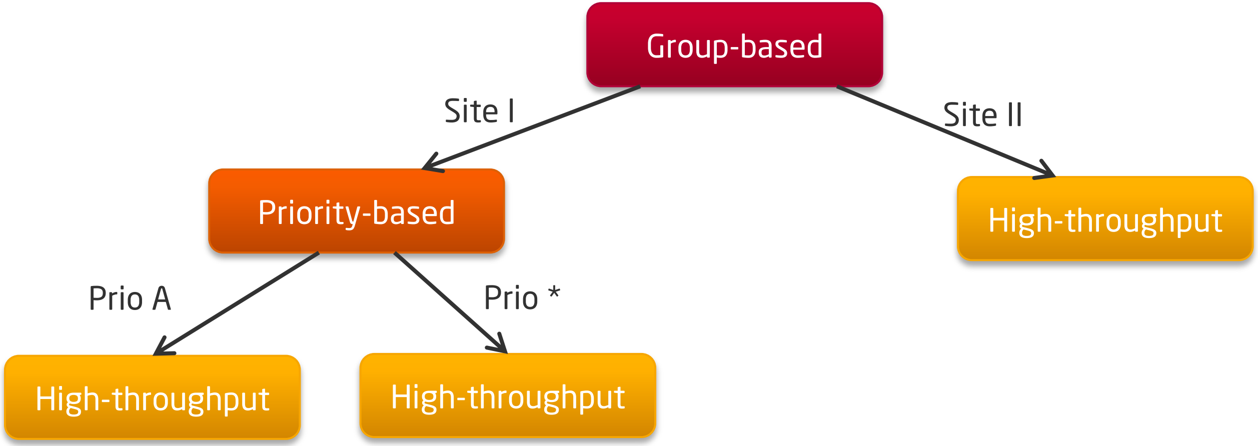


```
<xpdl:Activity CompletionQuantity="1" Id="newpkg1_wp1_act2" Name="BWA"
    <xpdl:Implementation>
        <xpdl:No/>
    </xpdl:Implementation>
    <xpdl:Performers>
        <xpdl:Performer>newpkg1_wp1_par1</xpdl:Performer>
    </xpdl:Performers>
    <xpdl:NodeGraphicsInfos>
        <xpdl:NodeGraphicsInfo BorderColor="#000000" FillColor="#99FF99
            <xpdl:Coordinates XCoordinate="239.0" YCoordinate="219.0"/>
        </xpdl:NodeGraphicsInfo>
    </xpdl:NodeGraphicsInfos>
</xpdl:Activity>
```



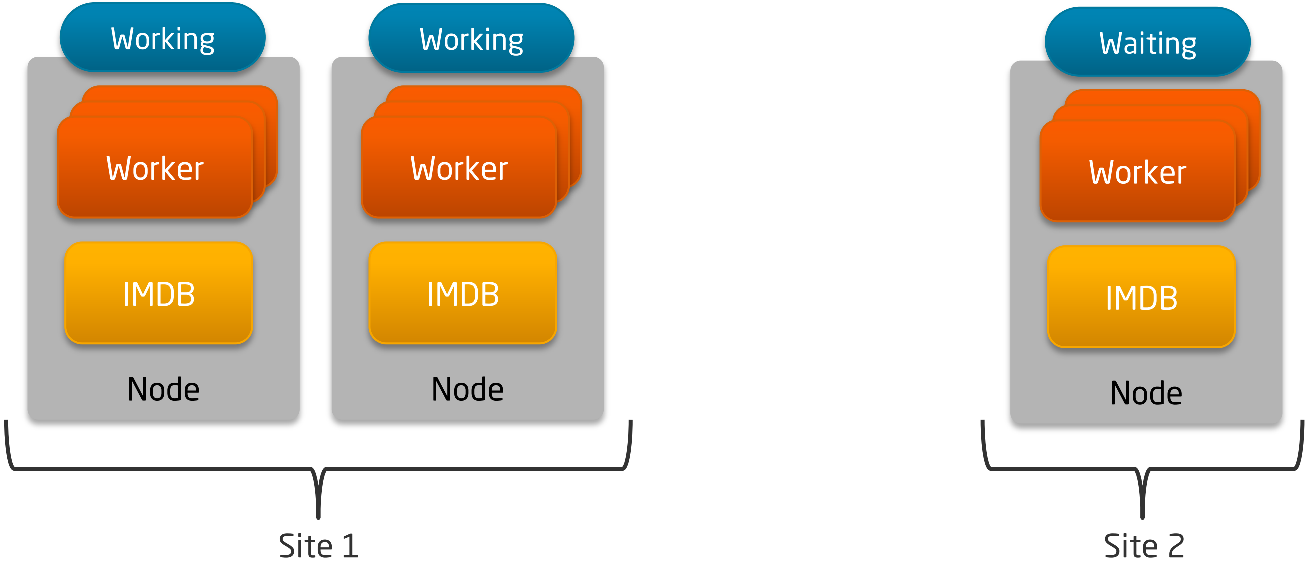
Runtime Environment
Scheduling Algorithms

- Scheduling algorithms are plug-in software modules
  - □ "User-/Group-based" to let users execute their tasks on their local site only
  - □ "Priority First" to prefer important users
  - □ "High Throughput", i.e. "shortest task first" to deal with high load
- Scheduling algorithms can also be composed hierarchically



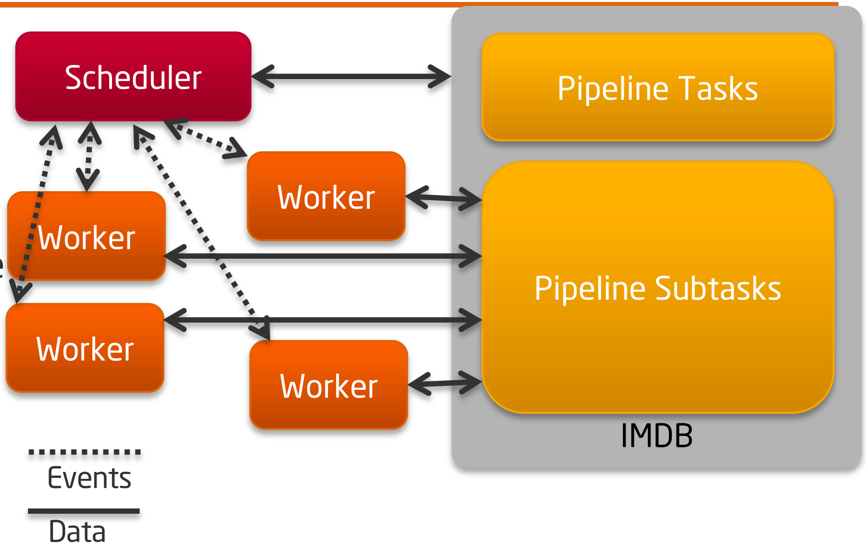
Runtime Environment
Resource Allocator

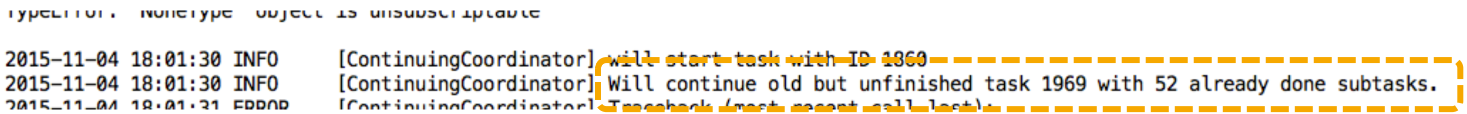
- Maintains lists of running and idle nodes
- Idle worker requests new sub task for its assigned groups
- If there is no matching sub task, it sleeps until a new sub task gets ready



Runtime Environment
Recoverability

- All execution data is stored in IMDB
- Temporary files on a shared file system
- In case of any failure, the system-wide state can be restored





```
TypeError: 'NoneType' object is unsubscriptable

2015-11-04 18:01:30 INFO     [ContinuingCoordinator] will start task with ID 1969
2015-11-04 18:01:30 INFO     [ContinuingCoordinator] Will continue old but unfinished task 1969 with 52 already done subtasks.
2015-11-04 18:01:31 ERROR    [ContinuingCoordinator] Traceback (most recent call last):
```

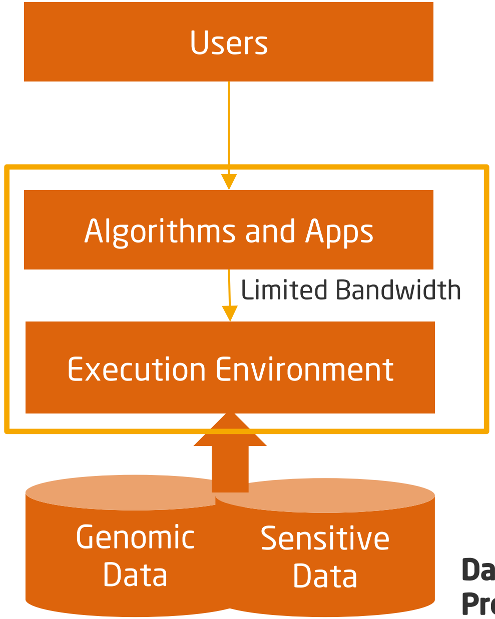
Federated Data Processing Comparison

- (Smaller) algorithms travel to (larger) data sets
- Forms a single virtual database (FIDMB) across sites and locations
- Master data managed by service provider whilst sensitive data resides locally

Pros	Cons
Single database license	Complex operation
Easy to consume services	Single setup required
Query propagation by IMDB	

What to take home?

- Modeling is key for reproducible research
- Use of standards can accelerate adoption
- Federation of data/algorithms facilitates data protection
- Move algorithms to data if size_of(data) > size_of(algorithms) by far
- Decentralized execution environment req., e.g., for execution of algorithms and results assembly
- Build on existing knowledge, e.g., resource scheduling


Users
Algorithms and Apps
Limited Bandwidth
Execution Environment
Genomic Data
Sensitive Data