Data and (semantic) standards in clinical genomics / bioinformatics

Martin Kerick

Bioinformatic standards – first try

Minimum standarts for bioinformatic command line tools

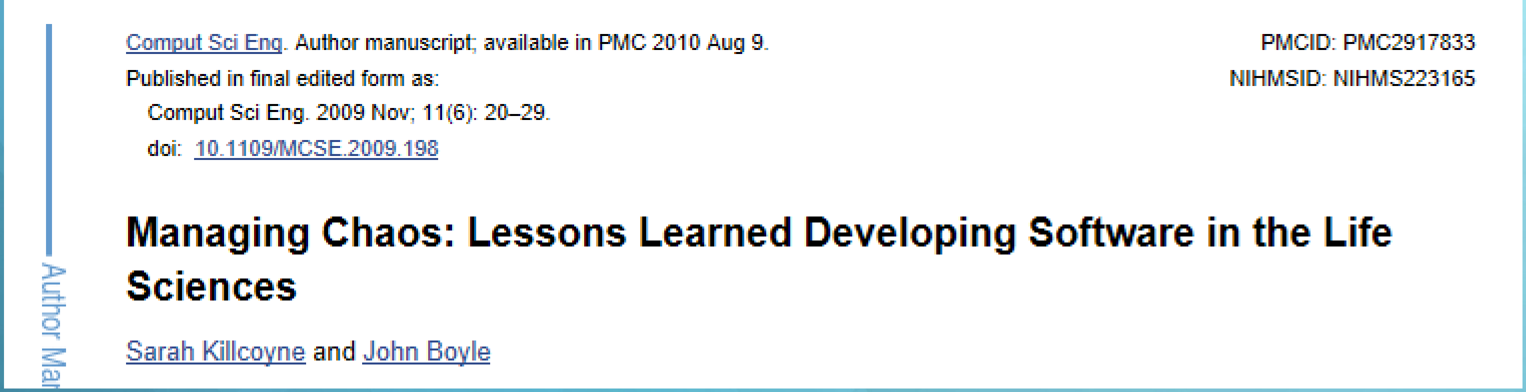
Always print something if no parameters are supplied
Always have a “-h” or “--help” switch
Have a “-v” or “--version” switch
Use stderr for messages and errors
Validate your parameters
Don’t hardcode any paths
Don’t pollute the command-line namespace
Don’t distribute bare JAR files
Check that your dependencies are installed
Be strict if you are a Perl tragic like me

The Genome Factory

Bioinformatics tips, tricks, tools and commentary with a microbial genomics bent.
Torsten Seemann from Melbourne, Australia.

http://thegenomefactory.blogspot.com.es/2013/08/minimum-standards-for-bioinformatics.html

Chaos in science - but it works



Comput Sci Eng. Author manuscript; available in PMC 2010 Aug 9.

Published in final edited form as:

Comput Sci Eng. 2009 Nov; 11(6): 20–29.

doi: 10.1109/MCSE.2009.198

PMCID: PMC2917833

NIHMSID: NIHMS223165

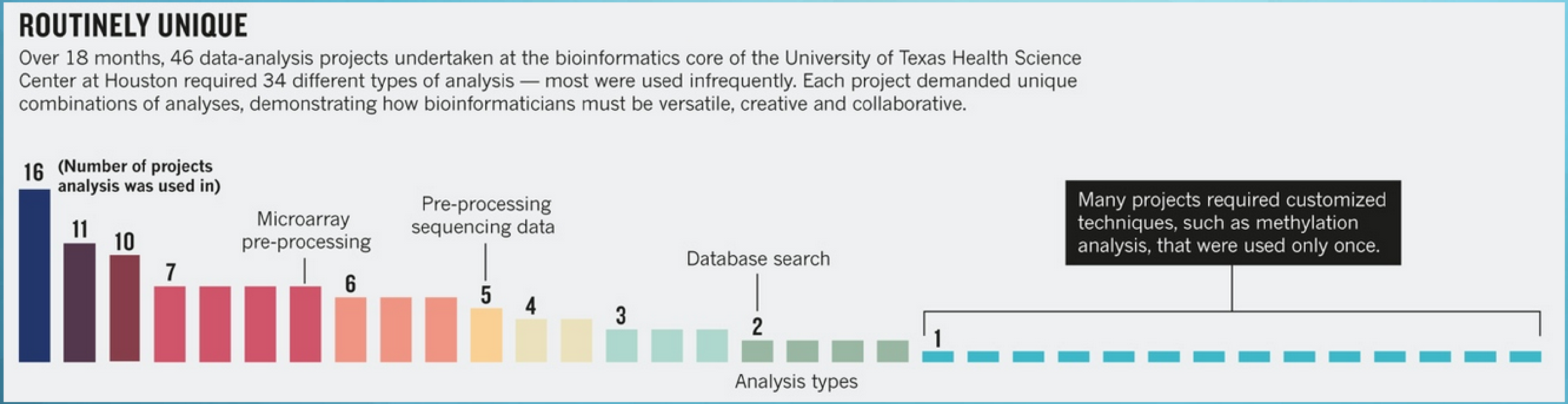
Managing Chaos: Lessons Learned Developing Software in the Life Sciences

Sarah Killcoyne and John Boyle

Life sciences research is, by nature, borderline chaotic. Scientists tend to work in small, isolated, and focused groups, collaborating only loosely with others. The process of testing and refining (or discarding) hypotheses leads to a multitude of elaborate experiments—each of which differs, using a unique mix of techniques, technologies, and analyses. Research mechanisms constantly change; researchers are continually introducing new technologies and refining older technologies. Experimental results can lead to myriad conclusions, some of which are contradictory and others of which are ignored. This constantly shifting landscape means that scientific discovery can sometimes be perceived as a manic foraging exercise rather than a rational, hypothesis-driven process. One of the most confusing elements of science is that this jumble of experiments leads to the development of ideas that directly advance our understanding of living systems. That is, the system works, and works well.

..in science!

Bioinformatics is largely driven by singular projects



ROUTINELY UNIQUE

Over 18 months, 46 data-analysis projects undertaken at the bioinformatics core of the University of Texas Health Science Center at Houston required 34 different types of analysis — most were used infrequently. Each project demanded unique combinations of analyses, demonstrating how bioinformaticians must be versatile, creative and collaborative.

Many projects required customized techniques, such as methylation analysis, that were used only once.

16
(Number of projects analysis was used in)
11
10
7
Microarray pre-processing
6
Pre-processing sequencing data
5
4
3
Database search
2
1
Analysis types

most analysis are hardly reused

Data and (semantic) standards
in clinical genomics / bioinformatics - second try

data standards

semantic
standards

clinical genomics

bioinformatics

Data and (semantic) standards
in clinical genomics / bioinformatics

~~semantic standards~~

data standards

controlled vocabulary

~~clinical genomics~~

clinical diagnosis
utilizing genomics

bioinformatics

Data and (semantic) standards in clinical genomics / bioinformatics

data standards

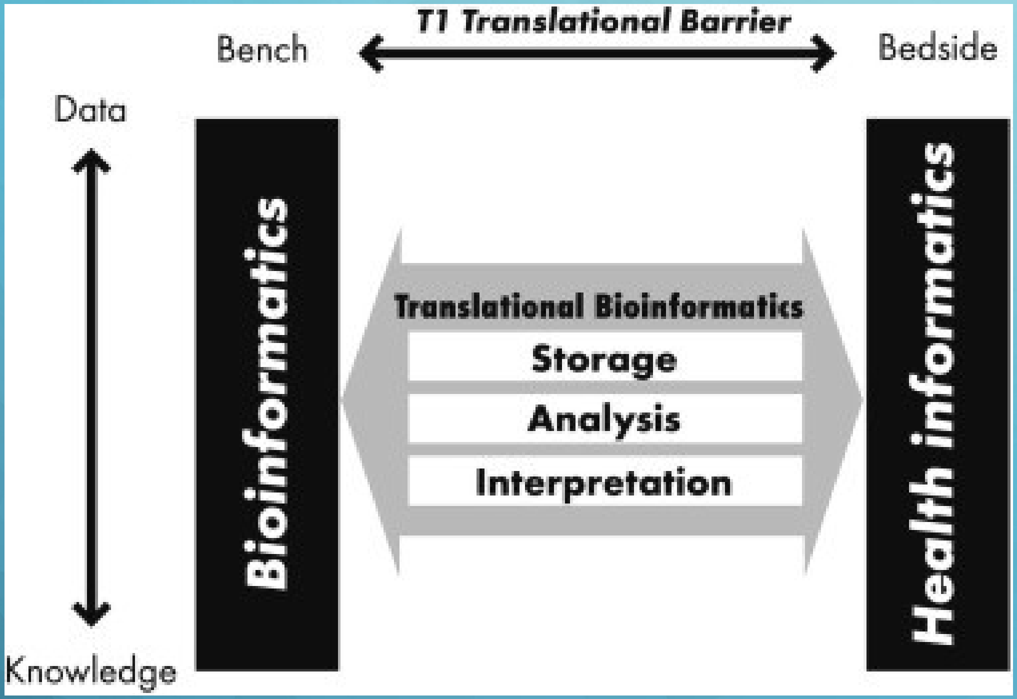
controlled vocabulary

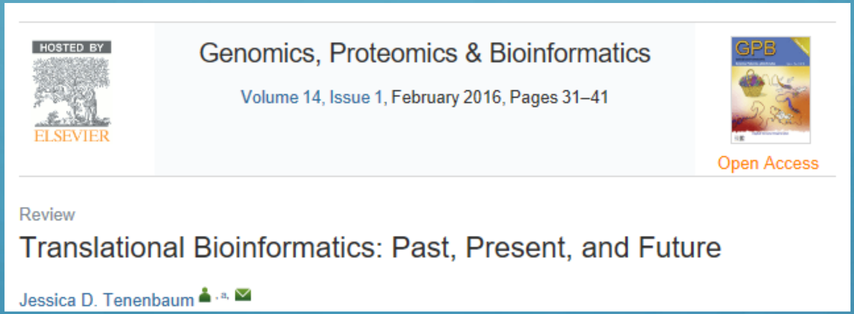
translational bioinformatics

clinical diagnosis
utilizing genomics

bioinformatics

Translational bioinformatics





Genomics, Proteomics & Bioinformatics

Volume 14, Issue 1, February 2016, Pages 31–41

Open Access

Review

Translational Bioinformatics: Past, Present, and Future

Jessica D. Tenenbaum [a]

Basic science **Vs** clinical use of genomics

I found ?/%^!? !

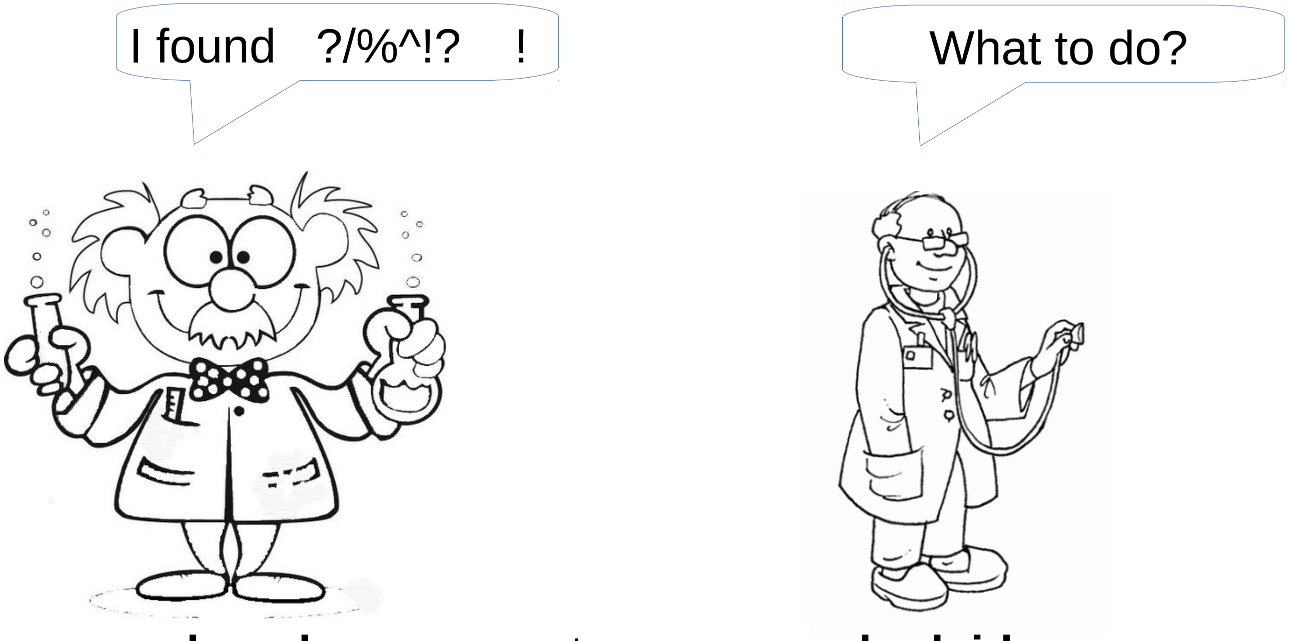
bench

to

What to do?

bedside

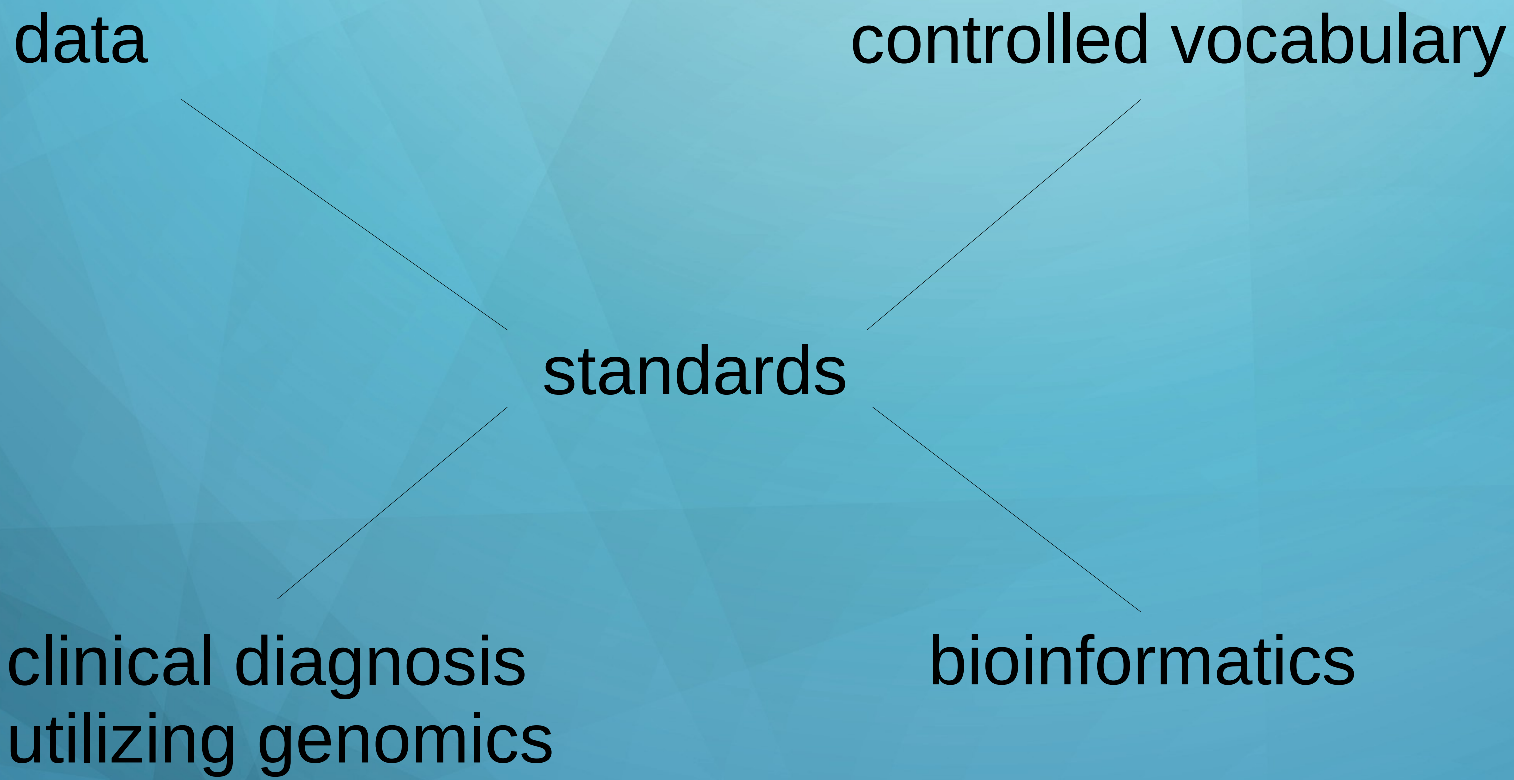
Basic science Vs clinical use of genomics



bench to **bedside**

information	action
panorama	focus
whole genome	gene panels
ranked list	diagnosis
tomorrow	now

Translational bioinformatics



data

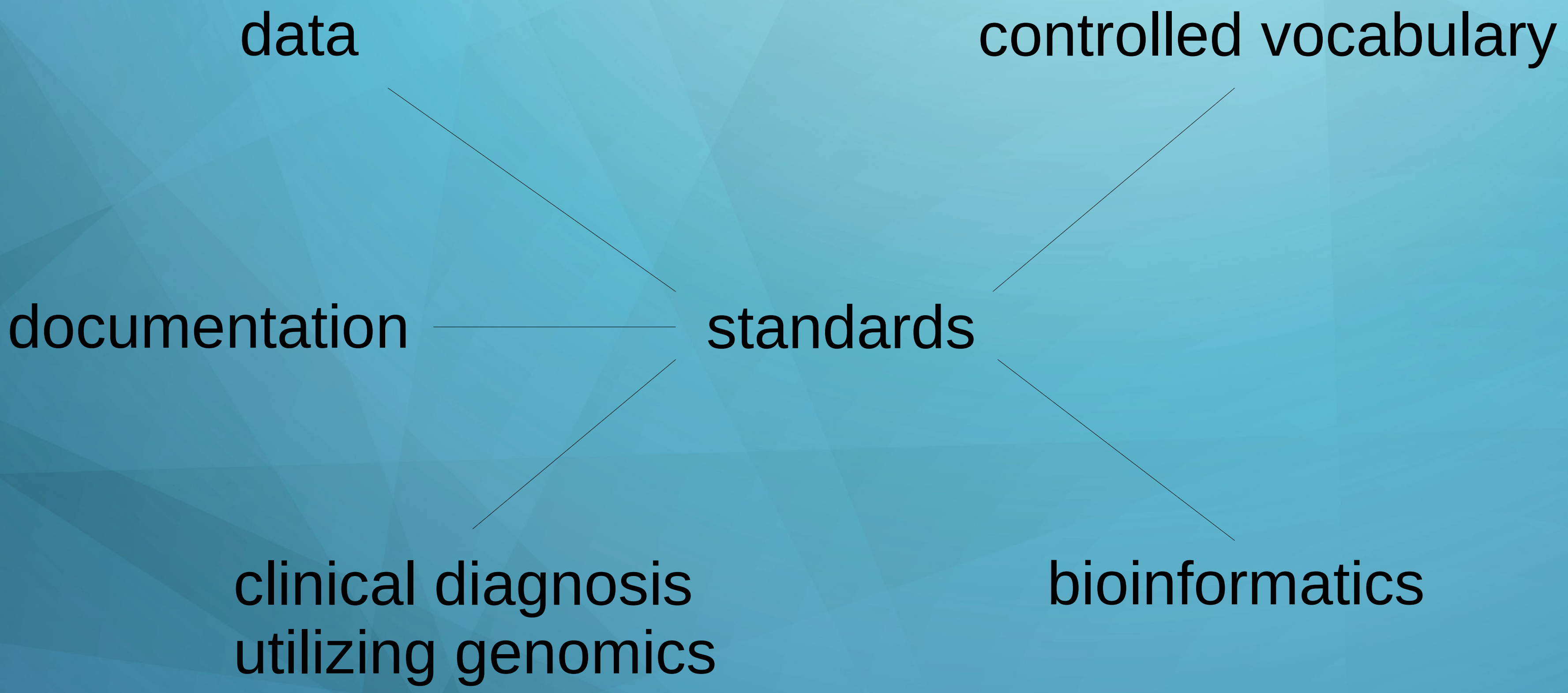
controlled vocabulary

standards

clinical diagnosis
utilizing genomics

bioinformatics

Translational bioinformatics



data

controlled vocabulary

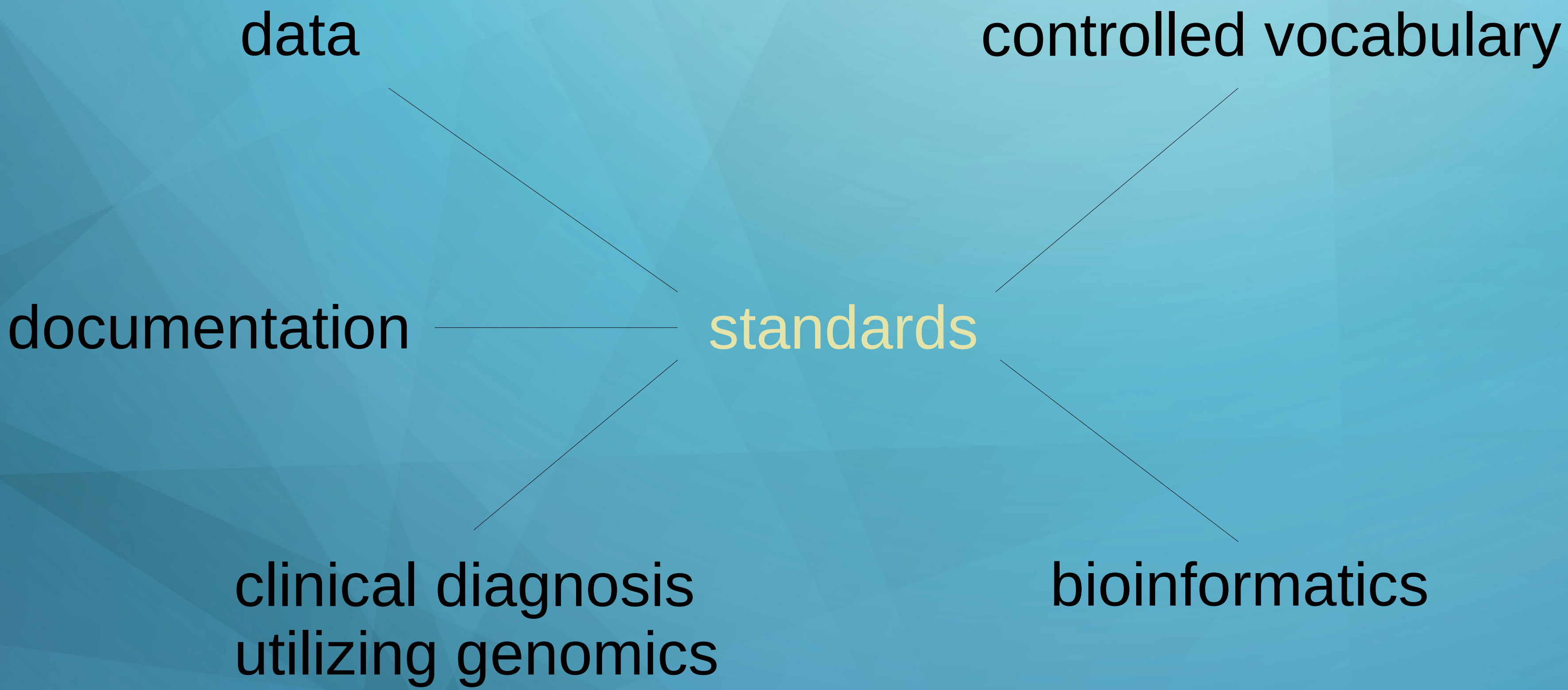
documentation

standards

clinical diagnosis
utilizing genomics

bioinformatics

Translational bioinformatic



data

controlled vocabulary

documentation

standards

clinical diagnosis
utilizing genomics

bioinformatics

...enable reproducibility

Reproducibility

needs controlled

- input data set (machine, protocol, tissue type, disease type)
- software environment (operation system)
- software
- parameters
- documentation
- interpretation

Reproducibility

needs controlled

- **input data set** (machine, protocol, tissue type, disease type)
- software environment (operation system)
- software
- parameters
- documentation
- interpretation

OMICS data types

Binary or
continuous data values?

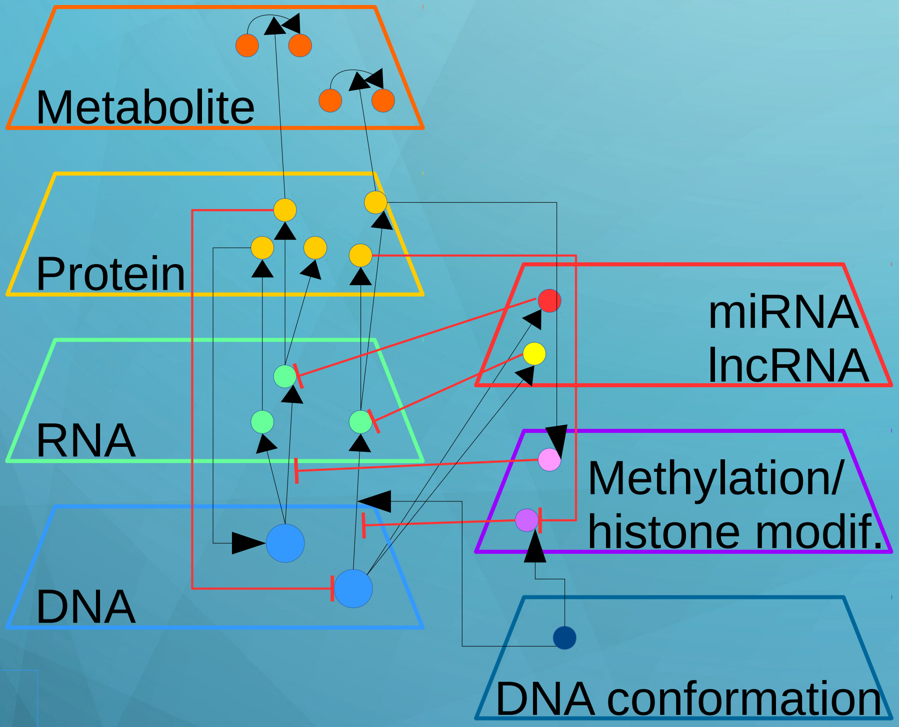
genomewide?
or
whole exome?
or
selected gene panel?

abundance
difference
activity

abundance
difference
modification
activity

abundance
difference
modification
(splicing)

SNPs, indels
CNV
rearrangements



abundance
difference

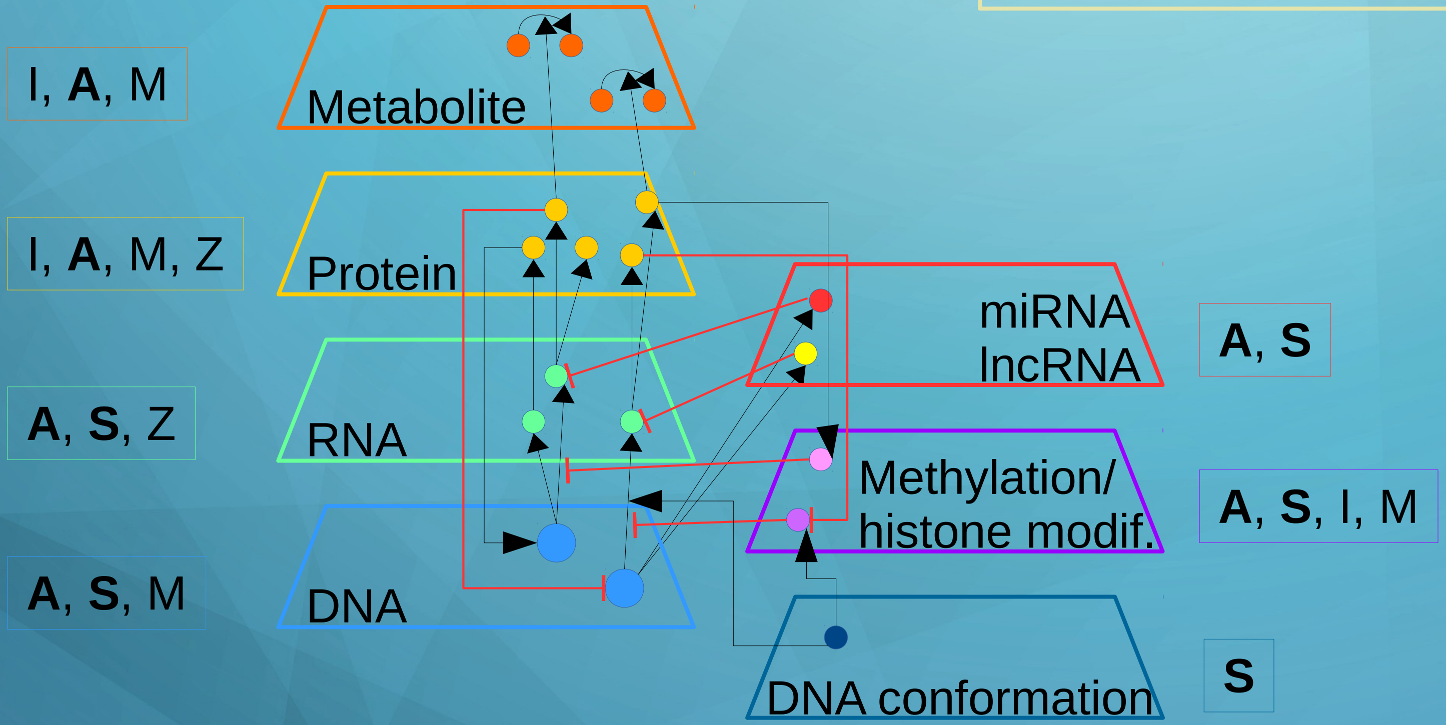
abundance
difference

conformation

figure adapted from Kristian Unger PMC3901372

OMICS methods

I	= immuno precipitation
A	= microarray
S	= sequencing
M	= mass spectrometry
Z	= specialized assay



I, **A**, M

Metabolite

I, **A**, M, Z

Protein

A, **S**, Z

RNA

A, **S**, M

DNA

miRNA
lncRNA

A, **S**

Methylation/
histone modif.

A, **S**, I, M

DNA conformation

S

OMICS data formats

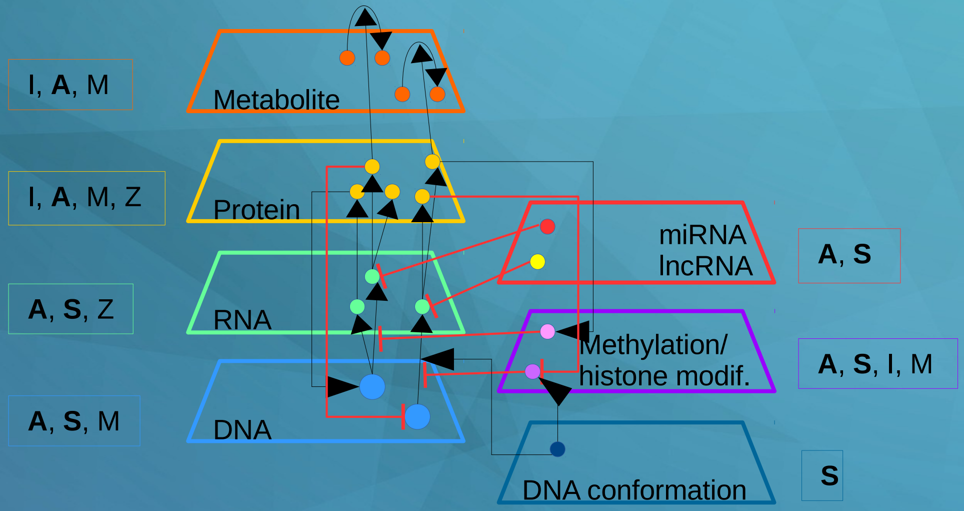
I = immuno precipitation
A = microarray
S = sequencing
M = mass spectrometry
Z = specialized assay

plain text

.csv
.tsv
.bed
.fasta
.fastq
.map
.ped
.json
.xml
.vcf 4.2
.sam
.gtf = .gff "2.5"
.gff 3.0
.owl
.tabix

binary

.bam
.cram
.bcf
.2bit
.bed(plink)
.bigBed
.bigWig
.RObject
.Rdata
.IDAT
.cel



bold = will likely stay for longer

https://software.broadinstitute.org/software/igv/FileFormats

https://genome-euro.ucsc.edu/goldenPath/help/customTrack.html#format

Example: The Cancer Genome Atlas (TCGA)

34 different cancer types - 11,077 samples

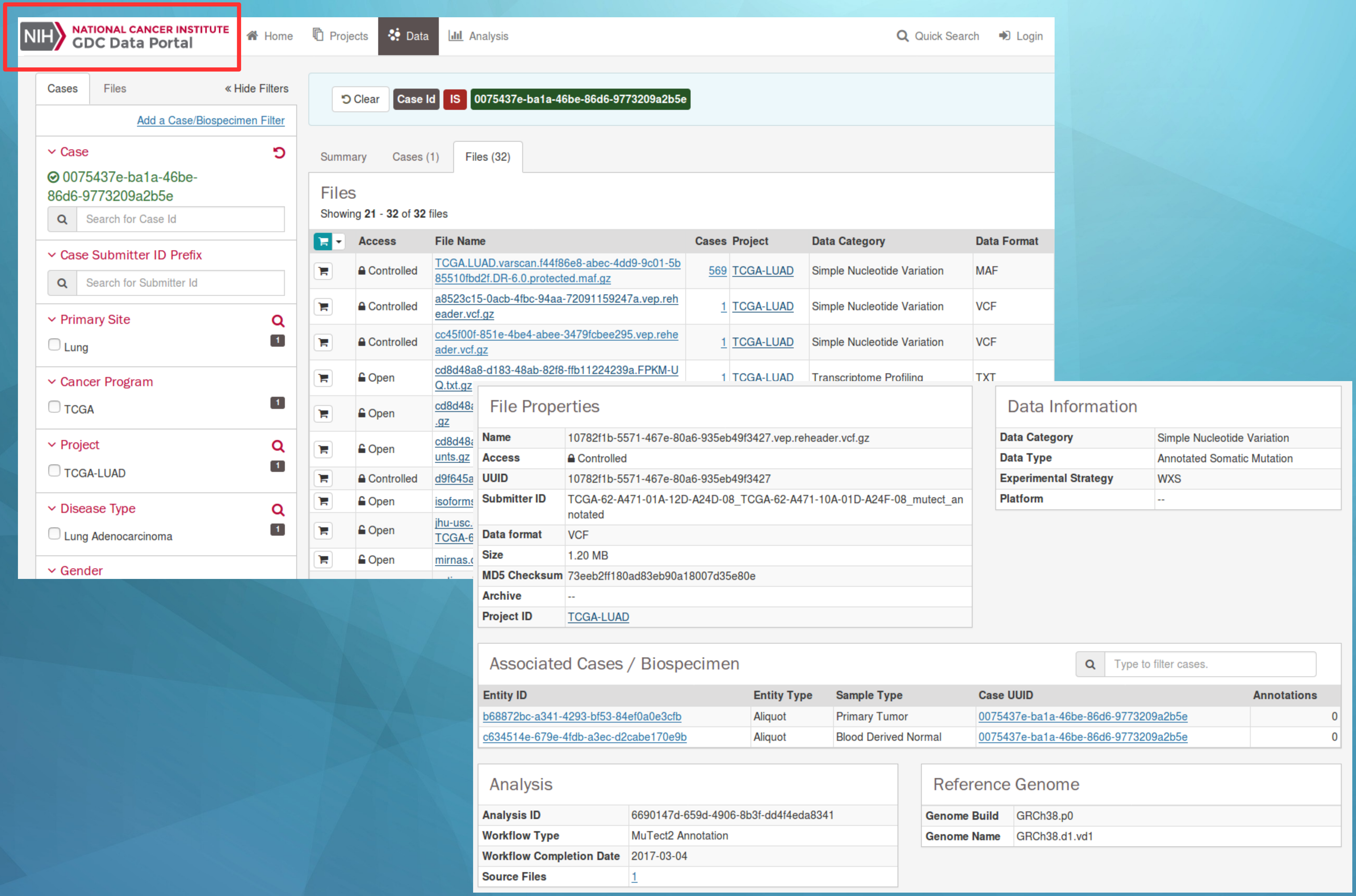
Lung Cancer:

Biospecimen: Primary tumor & Blood derived Normal

Techniques: whole Exome sequencing
Genotyping Array Affymetrix SNP 6.0
RNA Seq
miRNA Seq
Methylation Array Illumina 450K

Data: Raw data, BAM/CEL/IDAT
SNPs, somatic mutations, VCF, 1.2Mb
Somatic CNVs, TXT, 51Kb
Gene Expression values (FPKM), TXT, 519Kb
miRNA Expression values, TSV, 286Kb
Beta methylation values, TXT, 141Mb

Example: The Cancer Genome Atlas (TCGA)



Case
0075437e-ba1a-46be-86d6-9773209a2b5e

Case Submitter ID Prefix

Primary Site
□ Lung 1

Cancer Program
□ TCGA 1

Project
□ TCGA-LUAD 1

Disease Type
□ Lung Adenocarcinoma 1

Gender

Files

Showing **21** - **32** of **32** files

Access	File Name	Cases	Project	Data Category	Data Format
Controlled	TCGA.LUAD.varscan.f44f86e8-abec-4dd9-9c01-5b85510fbd2f.DR-6.0.protected.maf.gz	569	TCGA-LUAD	Simple Nucleotide Variation	MAF
Controlled	a8523c15-0acb-4fbc-94aa-72091159247a.vep.reheader.vcf.gz	1	TCGA-LUAD	Simple Nucleotide Variation	VCF
Controlled	cc45f00f-851e-4be4-abee-3479fcbee295.vep.reheader.vcf.gz	1	TCGA-LUAD	Simple Nucleotide Variation	VCF
Open	cd8d48a8-d183-48ab-82f8-ffb11224239a.FPKM-UQ.txt.gz	1	TCGA-LUAD	Transcriptome Profiling	TXT
Open	cd8d48a .gz				
Open	cd8d48a unts.gz				
Controlled	d9f645a				
Open	isoforms				
Open	jhu-usc. TCGA-6				
Open	mirnas.c				

File Properties

Name	10782f1b-5571-467e-80a6-935eb49f3427.vep.reheader.vcf.gz
Access	Controlled
UUID	10782f1b-5571-467e-80a6-935eb49f3427
Submitter ID	TCGA-62-A471-01A-12D-A24D-08_TCGA-62-A471-10A-01D-A24F-08_mutect_annotated
Data format	VCF
Size	1.20 MB
MD5 Checksum	73eeb2ff180ad83eb90a18007d35e80e
Archive	--
Project ID	TCGA-LUAD

Data Information

Data Category	Simple Nucleotide Variation
Data Type	Annotated Somatic Mutation
Experimental Strategy	WXS
Platform	--

Associated Cases / Biospecimen

Entity ID	Entity Type	Sample Type	Case UUID	Annotations
b68872bc-a341-4293-bf53-84ef0a0e3cfb	Aliquot	Primary Tumor	0075437e-ba1a-46be-86d6-9773209a2b5e	0
c634514e-679e-4fdb-a3ec-d2cabe170e9b	Aliquot	Blood Derived Normal	0075437e-ba1a-46be-86d6-9773209a2b5e	0

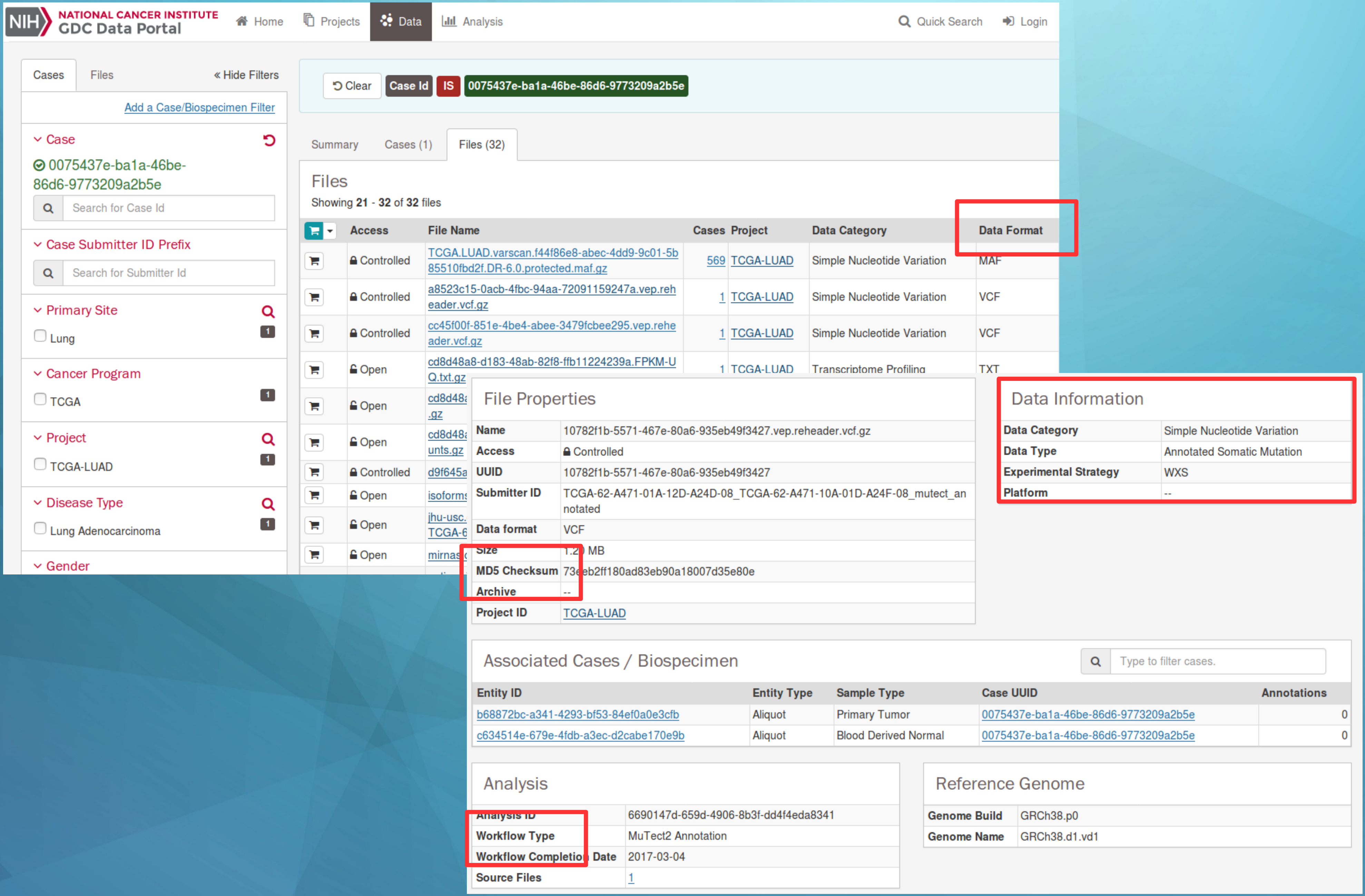
Analysis

Analysis ID	6690147d-659d-4906-8b3f-dd4f4eda8341
Workflow Type	MuTect2 Annotation
Workflow Completion Date	2017-03-04
Source Files	1

Reference Genome

Genome Build	GRCh38.p0
Genome Name	GRCh38.d1.vd1

Example: The Cancer Genome Atlas (TCGA)



Case

0075437e-ba1a-46be-86d6-9773209a2b5e

Search for Case Id

Case Submitter ID Prefix

Search for Submitter Id

Primary Site

□ Lung 1

Cancer Program

□ TCGA 1

Project

□ TCGA-LUAD 1

Disease Type

□ Lung Adenocarcinoma 1

Gender

Files

Showing **21** - **32** of **32** files

	Access	File Name	Cases	Project	Data Category	Data Format
	Controlled	TCGA.LUAD.varscan.f44f86e8-abec-4dd9-9c01-5b85510fbd2f.DR-6.0.protected.maf.gz	569	TCGA-LUAD	Simple Nucleotide Variation	MAF
	Controlled	a8523c15-0acb-4fbc-94aa-72091159247a.vep.reheader.vcf.gz	1	TCGA-LUAD	Simple Nucleotide Variation	VCF
	Controlled	cc45f00f-851e-4be4-abee-3479fcbee295.vep.reheader.vcf.gz	1	TCGA-LUAD	Simple Nucleotide Variation	VCF
	Open	cd8d48a8-d183-48ab-82f8-ffb11224239a.FPKM-UQ.txt.gz	1	TCGA-LUAD	Transcriptome Profiling	TXT
	Open	cd8d48a .gz				
	Open	cd8d48a unts.gz				
	Controlled	d9f645a				
	Open	isoforms				
	Open	jhu-usc. TCGA-6				
	Open	mirnas				

File Properties

Name	10782f1b-5571-467e-80a6-935eb49f3427.vep.reheader.vcf.gz
Access	Controlled
UUID	10782f1b-5571-467e-80a6-935eb49f3427
Submitter ID	TCGA-62-A471-01A-12D-A24D-08_TCGA-62-A471-10A-01D-A24F-08_mutect_annotated
Data format	VCF
Size	1.20 MB
MD5 Checksum	73eeb2ff180ad83eb90a18007d35e80e
Archive	--
Project ID	TCGA-LUAD

Data Information

Data Category	Simple Nucleotide Variation
Data Type	Annotated Somatic Mutation
Experimental Strategy	WXS
Platform	--

Associated Cases / Biospecimen

Type to filter cases.

Entity ID	Entity Type	Sample Type	Case UUID	Annotations
b68872bc-a341-4293-bf53-84ef0a0e3cfb	Aliquot	Primary Tumor	0075437e-ba1a-46be-86d6-9773209a2b5e	0
c634514e-679e-4fdb-a3ec-d2cabe170e9b	Aliquot	Blood Derived Normal	0075437e-ba1a-46be-86d6-9773209a2b5e	0

Analysis

Analysis ID	6690147d-659d-4906-8b3f-dd4f4eda8341
Workflow Type	MuTect2 Annotation
Workflow Completion Date	2017-03-04
Source Files	1

Reference Genome

Genome Build	GRCh38.p0
Genome Name	GRCh38.d1.vd1

Example: PRECISESADS

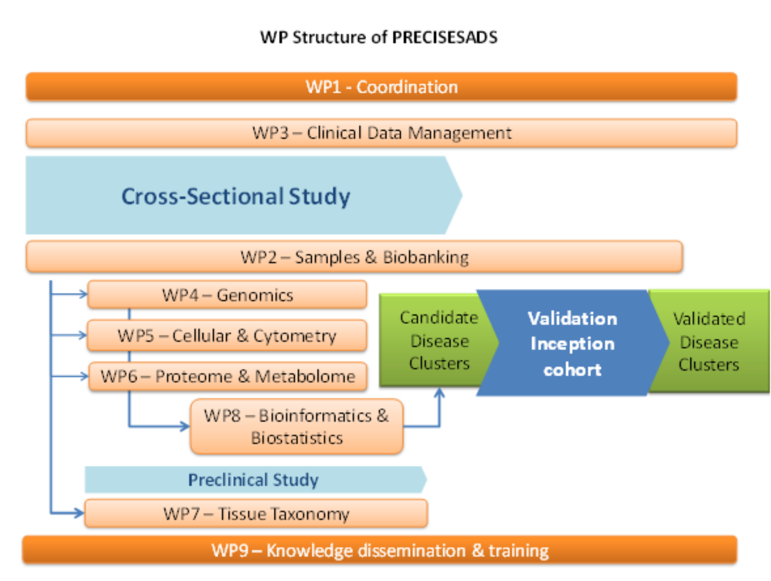
4+3 different autoimmune diseases ~ 2,600 samples

Rheumatoide arthritis:

Biospecimen: whole blood, selected cell populations

Techniques: Genotyping Array Illumina Human Core 360 K
(Expression Array Human HT12v4)
RNA Seq
Methylation Array Illumina 450K
8 color flow cytometry of 8 Antibody panels
Mass-spectrometry of Plasma Metabolites
Cytokines & Auto-antibodies (Luminex Assay)
Imaging Analysis

Data: Raw data, BAM/IDAT/TXT
SNPs, Bed/Bim/Fam
Germline CNVs, TXT
Gene Expression values (FPKM),TXT
TXT
miRNA Expression values, TXT
Beta methylation values, TXT
Metabolomic peaks, TXT
Cytokine levels, TXT
Auto-antibody-levels, TXT
Flow Cytometry, ?



WP Structure of PRECISESADS

WP1 - Coordination
WP3 – Clinical Data Management
Cross-Sectional Study
WP2 – Samples & Biobanking
WP4 – Genomics
WP5 – Cellular & Cytometry
WP6 – Proteome & Metabolome
WP8 – Bioinformatics & Biostatistics
Candidate Disease Clusters
Validation Inception cohort
Validated Disease Clusters
Preclinical Study
WP7 – Tissue Taxonomy
WP9 – Knowledge dissemination & training

Reproducibility

needs controlled

input data set (machine, protocol, tissue type, disease type)

software environment (operation system)

software

parameters

interpretation

documentation

Linux.

Reproducibility

needs controlled

- input data set (machine, protocol, tissue type, disease type)
- software environment (operation system)
- **software**
- parameters
- interpretation
- documentation

Bioinformatic “standard” Software

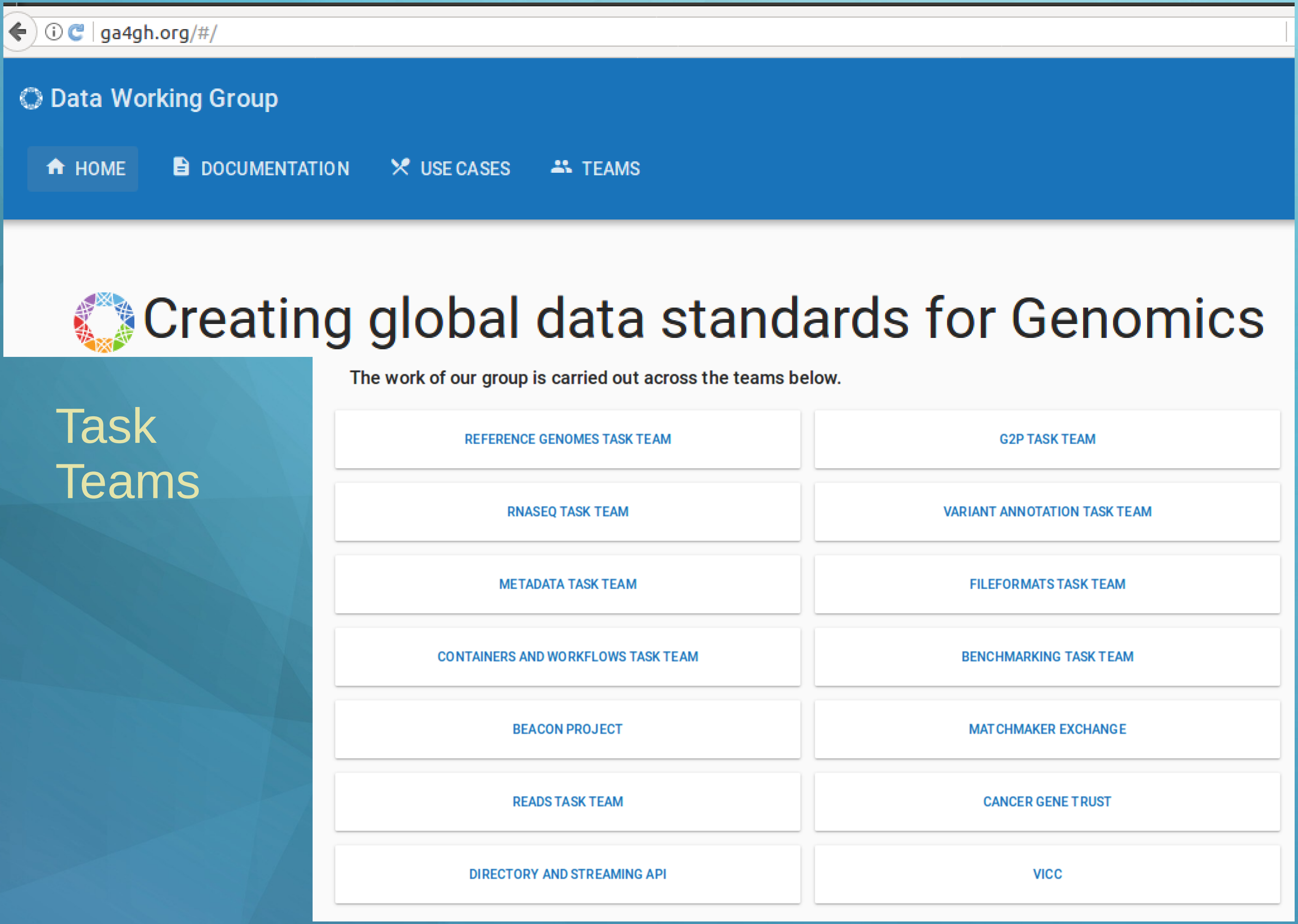
as found by majority vote

Task/Person	Pedro	Edu	Axel	Sven	Martin	Carlos
QC	**fastqc**, qualimap	**fastqc**	**fastqc**, bseqc, rseqc	**fastqc**	**fastqc**	**fastqc**
Sequence Trimming	-	cut adapt, reaper, minion, fastqx	flexbar, seqtrimnext	trimmomatic, superdeduper, prinseq	-	cut adapt, fastqx
Alignment DNA	**bwa**, **bowtie2**	**bwa**, **bowtie2**	**bwa**	**bwa**, **bowtie2**	**bwa**	**bwa, bowtie2**, blasr, dazzler
Alignment RNA	**star**	tuxedo, **star**, bowtie1, miarma-Seq	**star**, kallisto, salmon	hisat2	**star**	tophat, gmap, **star**, blasr
Alignment Bisulphite DNA	bwa	rubio-seq	**bismark**, **bsmap**	**bismark, bsmap**	**bismark**, **bsmap**	bwa-meth, **bismark**
SNP/Indel detection	**samtools**, **gatk**, varscan	**samtools**, bcftools	**gatk**, freebayes	**gatk**	**gatk**, **samtools**	**samtools**, bcftools
CNV detection	-	gistic	delly	bedtools coverage, cnver	dnacopy, penncnv	-
Expression analysis	**deseq2**, rsem	**limma**, **edger**, **deseq2**, noiseq	**limma**, **edger, deseq2**	stringtie, ballgown	**limma**, **edger**	**limma**, **edger**, **deseq2**, noiseq, sqanti
Methylation analysis	rnbeads, minfi	wanderer, lumi	methylkit	qsea, mcall	minfi, qsea	methylkit, bsseq
Gene enrichment	genecodis, **gsea**, **david**	**gsea**, goseq, **david**, **ingenuity**	**gsea**, clusterprofiler	**gsea**, **david**, fisher.test	**gsea**, **david**, **ingenuity**, consensuspath db	**david**, **ingenuity**, blast2go
Clustering	**hclust**, kmeans, som, pca, ica	venny, dendrogram, heatmap2	pca, kmeans, random forest	**hclust**, pca	gmm, **hclust**, nmf	dendrogram, **hclust**, heatmap2

Bioinformatic “standard” Software

as found by “Expert vote”

ga4gh.org



ga4gh.org/#/

Data Working Group

HOME DOCUMENTATION USE CASES TEAMS

Creating global data standards for Genomics

The work of our group is carried out across the teams below.

REFERENCE GENOMES TASK TEAM	G2P TASK TEAM
RNASEQ TASK TEAM	VARIANT ANNOTATION TASK TEAM
METADATA TASK TEAM	FILEFORMATS TASK TEAM
CONTAINERS AND WORKFLOWS TASK TEAM	BENCHMARKING TASK TEAM
BEACON PROJECT	MATCHMAKER EXCHANGE
READS TASK TEAM	CANCER GENE TRUST
DIRECTORY AND STREAMING API	VICC

Task Teams

Reproducibility

needs controlled

- input data set (machine, protocol, tissue type, disease type)
- software environment (operation system)
- software
- **parameters**
- interpretation
- documentation

Reproducibility

data

soft
ware

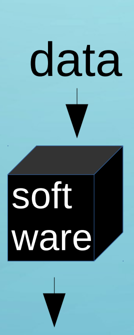
result = **apple spritzer**

Reproducibility

data

soft ware

result = **apple spritzer**

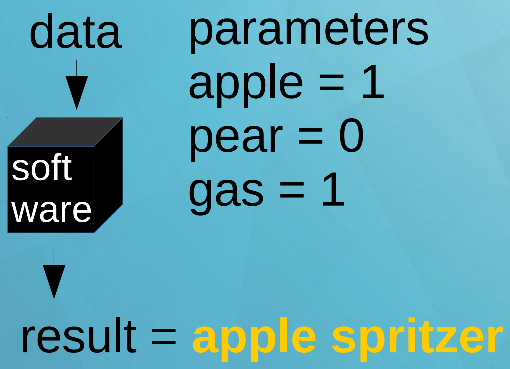


data

soft ware

result = **bad pear juice**

Reproducibility

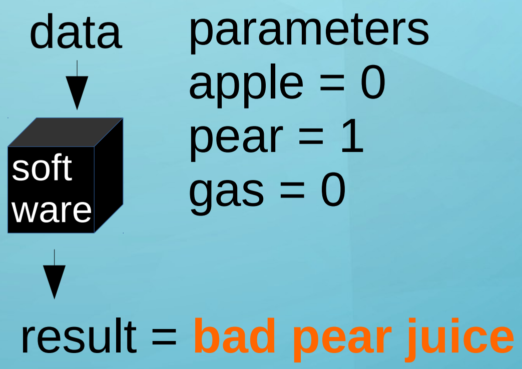


data

parameters
apple = 1
pear = 0
gas = 1

soft
ware

result = **apple spritzer**



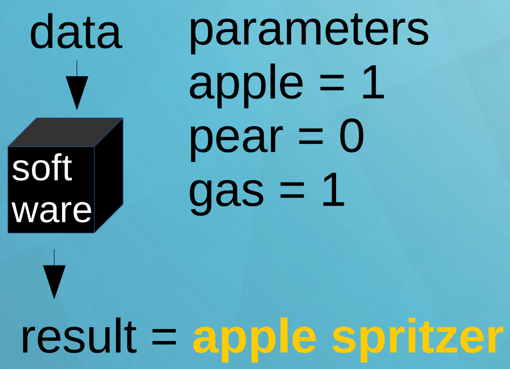
data

parameters
apple = 0
pear = 1
gas = 0

soft
ware

result = **bad pear juice**

Reproducibility

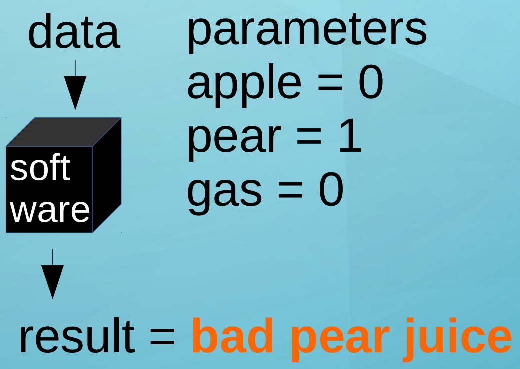


data

parameters
apple = 1
pear = 0
gas = 1

soft
ware

result = **apple spritzer**



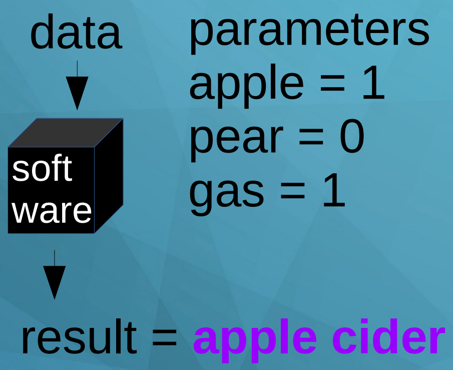
data

parameters
apple = 0
pear = 1
gas = 0

soft
ware

result = **bad pear juice**

a year passes by ...

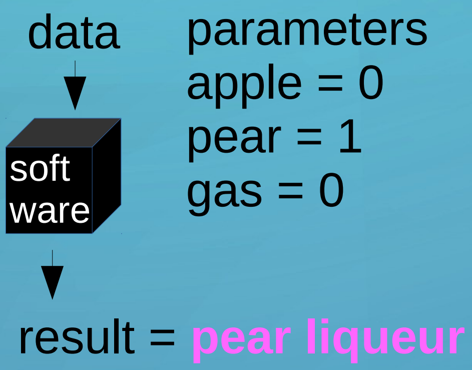


data

parameters
apple = 1
pear = 0
gas = 1

soft
ware

result = **apple cider**



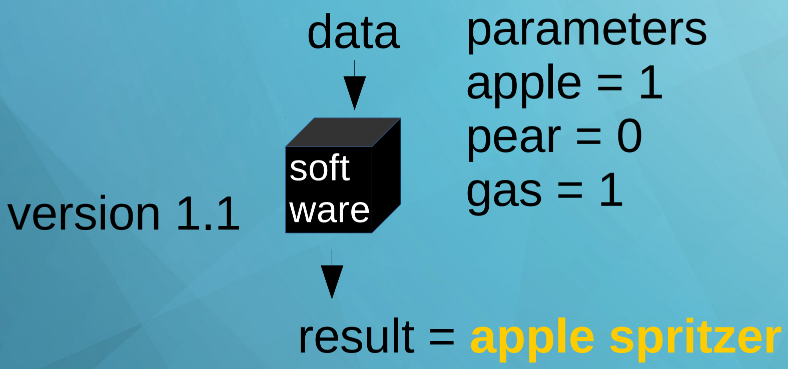
data

parameters
apple = 0
pear = 1
gas = 0

soft
ware

result = **pear liqueur**

Reproducibility

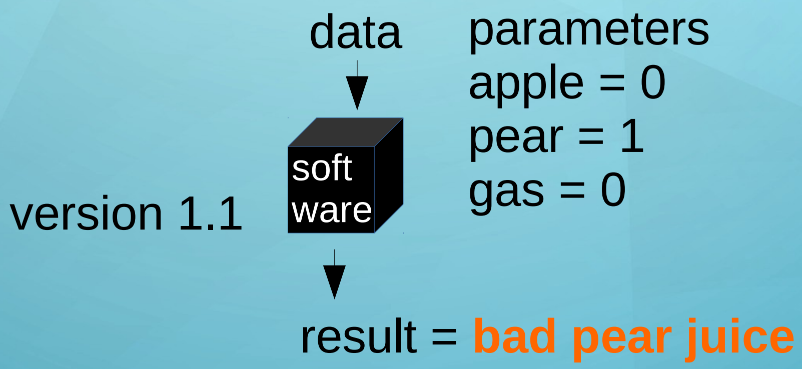

data

parameters
apple = 1
pear = 0
gas = 1

version 1.1

soft
ware

result = **apple spritzer**


data

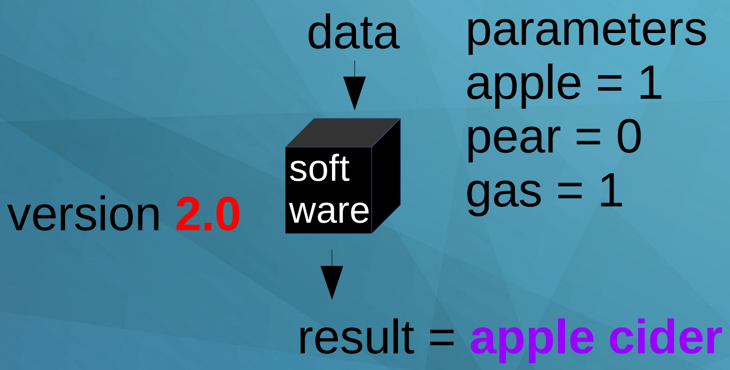
parameters
apple = 0
pear = 1
gas = 0

version 1.1

soft
ware

result = **bad pear juice**

a year passes by ...

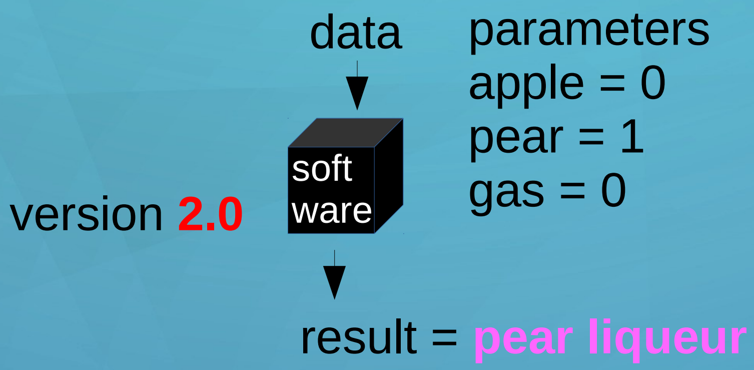

data

parameters
apple = 1
pear = 0
gas = 1

version **2.0**

soft
ware

result = **apple cider**


data

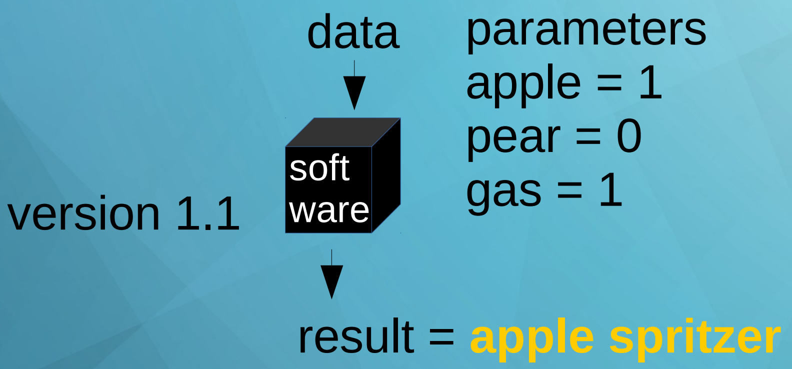
parameters
apple = 0
pear = 1
gas = 0

version **2.0**

soft
ware

result = **pear liqueur**

Reproducibility



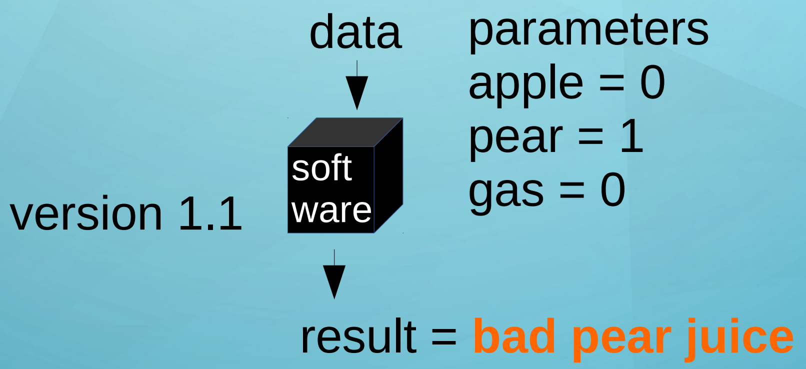
data

parameters
apple = 1
pear = 0
gas = 1

version 1.1

soft
ware

result = **apple spritzer**



data

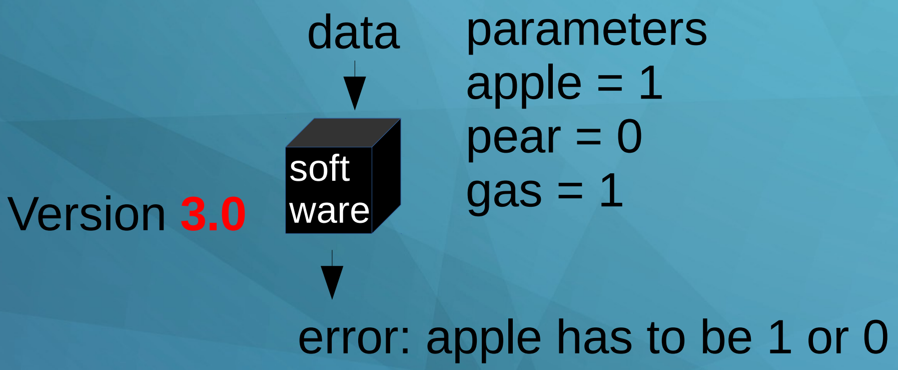
parameters
apple = 0
pear = 1
gas = 0

version 1.1

soft
ware

result = **bad pear juice**

three years pass by ...



data

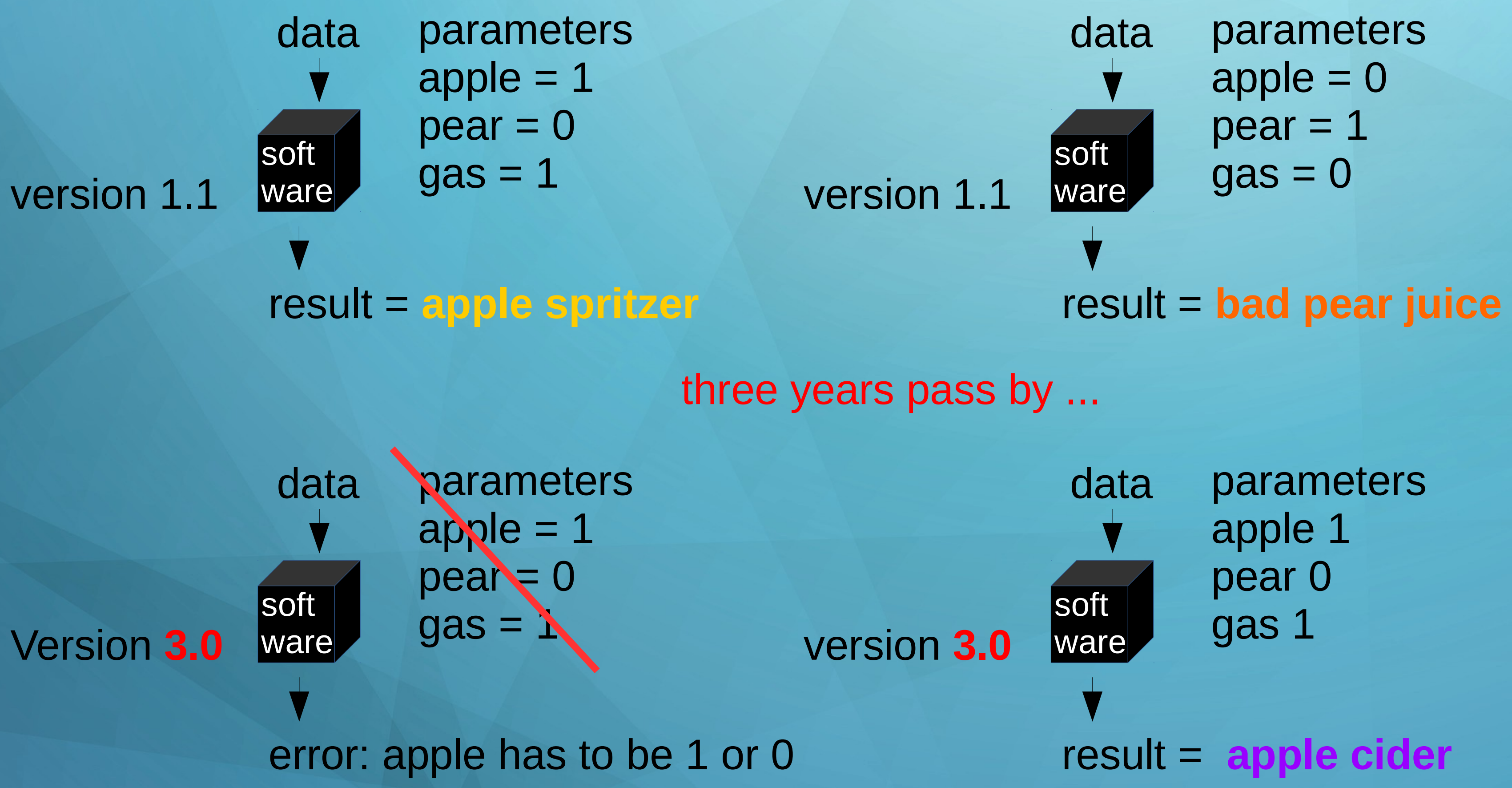
parameters
apple = 1
pear = 0
gas = 1

Version **3.0**

soft
ware

error: apple has to be 1 or 0

Reproducibility



data

parameters
apple = 1
pear = 0
gas = 1

version 1.1

soft
ware

result = **apple spritzer**

data

parameters
apple = 0
pear = 1
gas = 0

version 1.1

soft
ware

result = **bad pear juice**

three years pass by ...

data

~~parameters~~
~~apple = 1~~
~~pear = 0~~
~~gas = 1~~

Version **3.0**

soft
ware

error: apple has to be 1 or 0

data

parameters
apple 1
pear 0
gas 1

version **3.0**

soft
ware

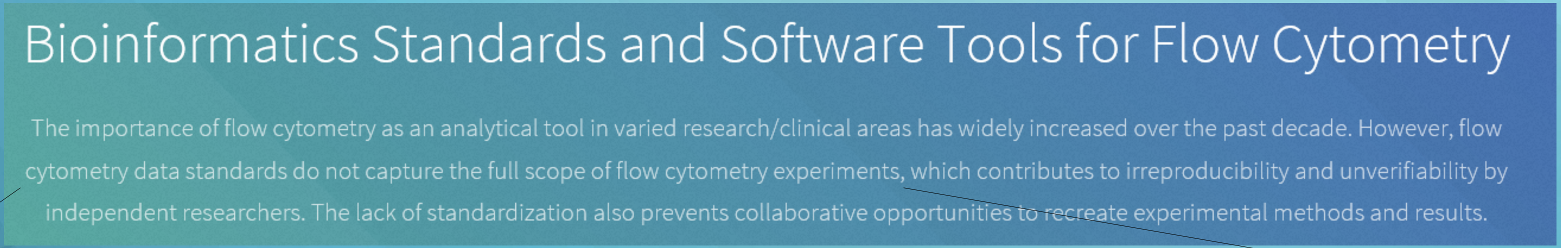
result = **apple cider**

Reproducibility

needs controlled

- input data set (machine, protocol, tissue type, disease type)
- software environment (operation system)
- software (version)
- parameters
- **documentation**
- interpretation

Documentation



Bioinformatics Standards and Software Tools for Flow Cytometry

The importance of flow cytometry as an analytical tool in varied research/clinical areas has widely increased over the past decade. However, flow cytometry data standards do not capture the full scope of flow cytometry experiments, which contributes to irreproducibility and unverifiability by independent researchers. The lack of standardization also prevents collaborative opportunities to recreate experimental methods and results.

Cytometry data standards do not capture the full scope of flow cytometry experiments

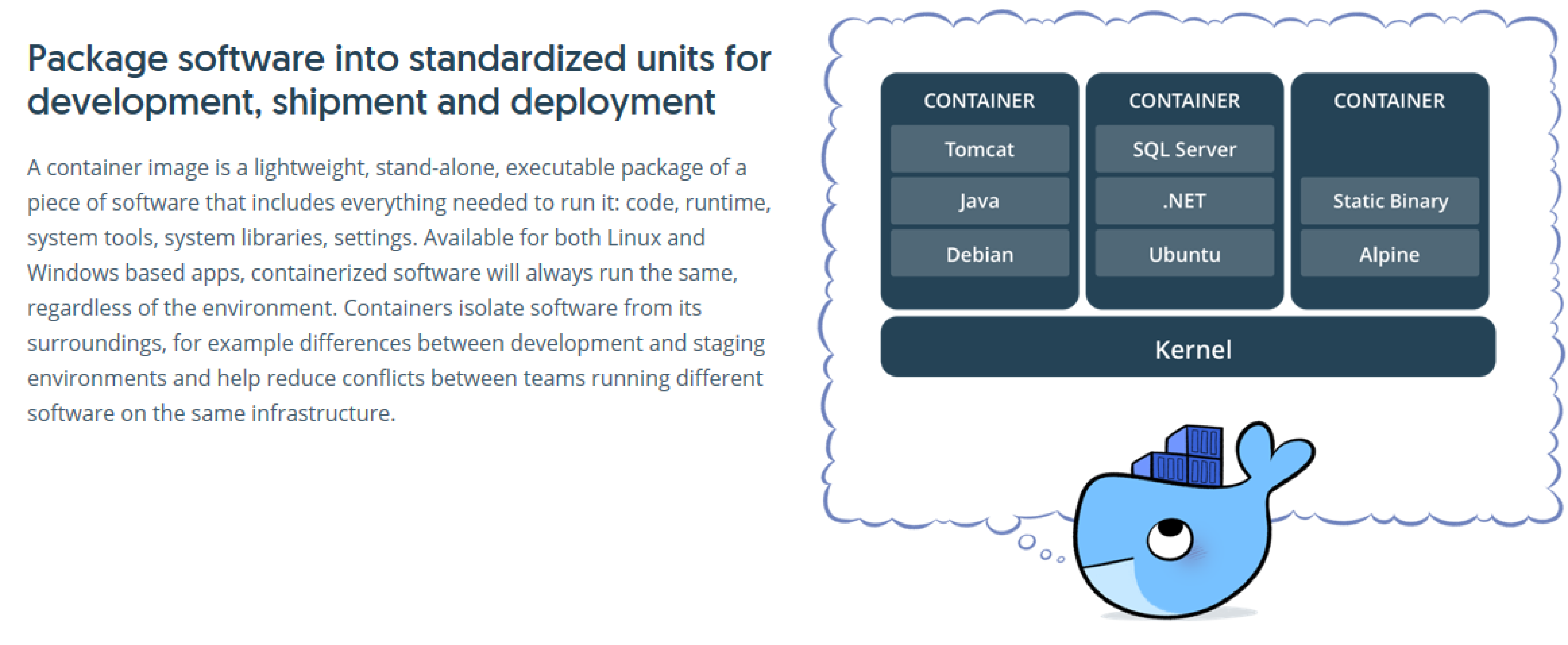
my personal documentation:

(commented) Perl code
(commented) R code
(sometimes) README files

Documentation from a data analyst point of view

Sweave, **Knitr** integrate **Rcode** with **LaTeX** into a “executable” and “readable” **pdf**

Docker saves your stable version of tools/pipelines within the changing software environment

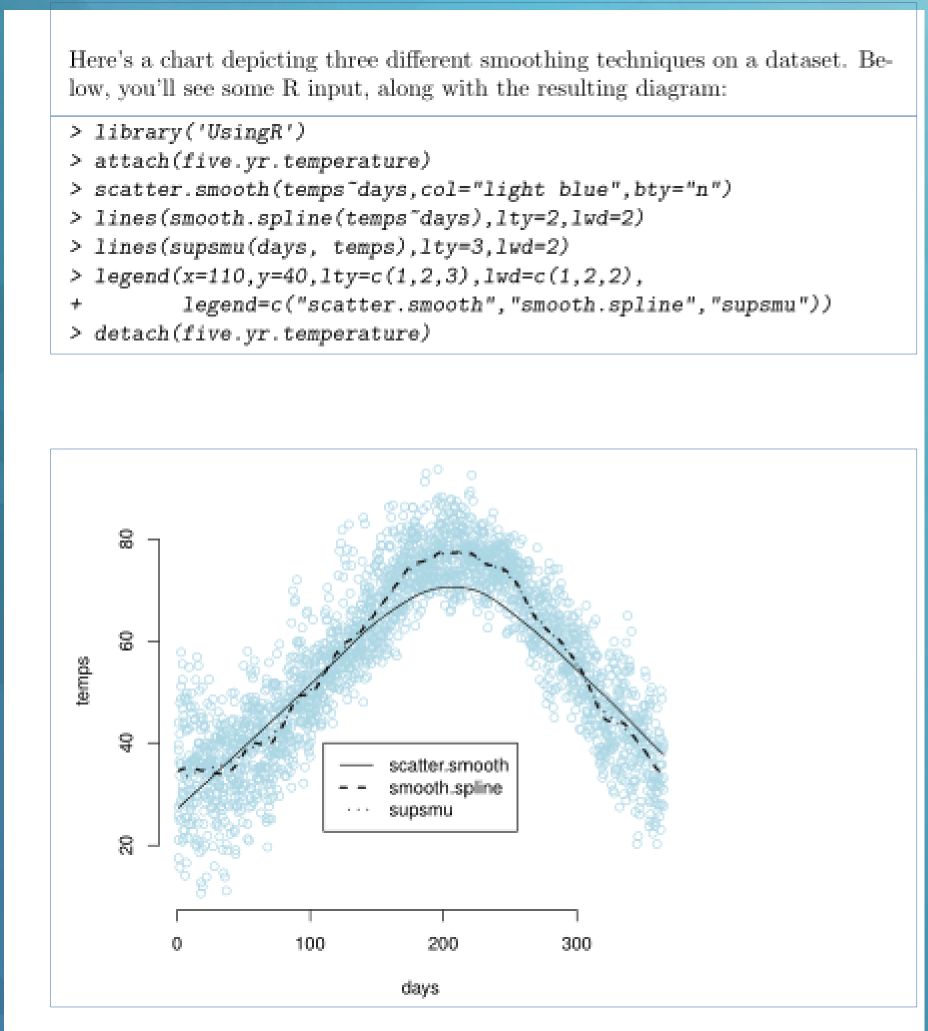


Package software into standardized units for development, shipment and deployment

A container image is a lightweight, stand-alone, executable package of a piece of software that includes everything needed to run it: code, runtime, system tools, system libraries, settings. Available for both Linux and Windows based apps, containerized software will always run the same, regardless of the environment. Containers isolate software from its surroundings, for example differences between development and staging environments and help reduce conflicts between teams running different software on the same infrastructure.

Documentation from a data analyst point of view

Sweave, **Knitr, Rcode** example
produces a **pdf**



Here's a chart depicting three different smoothing techniques on a dataset. Below, you'll see some R input, along with the resulting diagram:

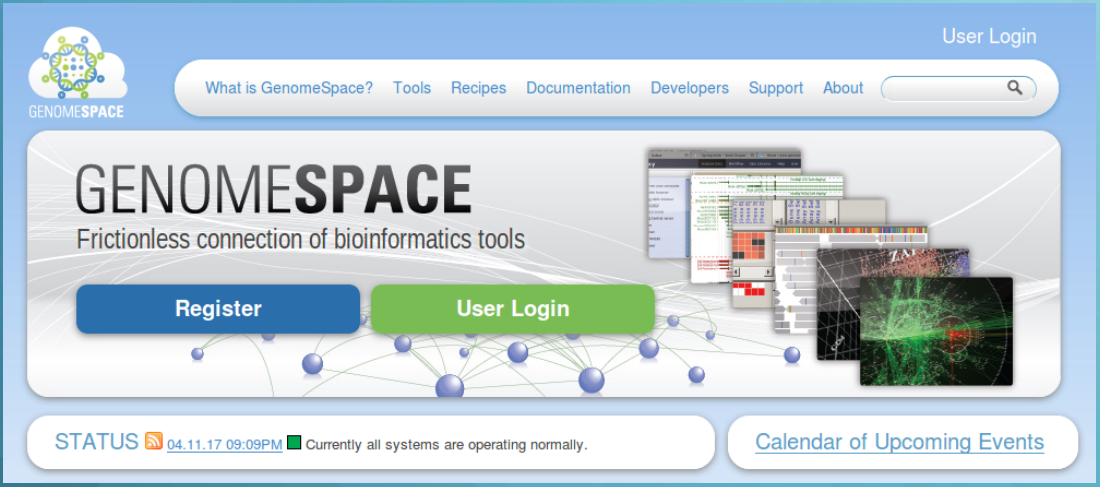
```
> library('UsingR')
> attach(five.yr.temperature)
> scatter.smooth(temps~days,col="light blue",bty="n")
> lines(smooth.spline(temps~days),lty=2,lwd=2)
> lines(supsmu(days, temps),lty=3,lwd=2)
> legend(x=110,y=40,lty=c(1,2,3),lwd=c(1,2,2),
+        legend=c("scatter.smooth","smooth.spline","supsmu"))
> detach(five.yr.temperature)
```

a commentary text
to your analysis

the R code producing
the result below

the result

Documentation
..integrated with user-defined pipelines



User Login

GENOMESPACE

What is GenomeSpace? Tools Recipes Documentation Developers Support About

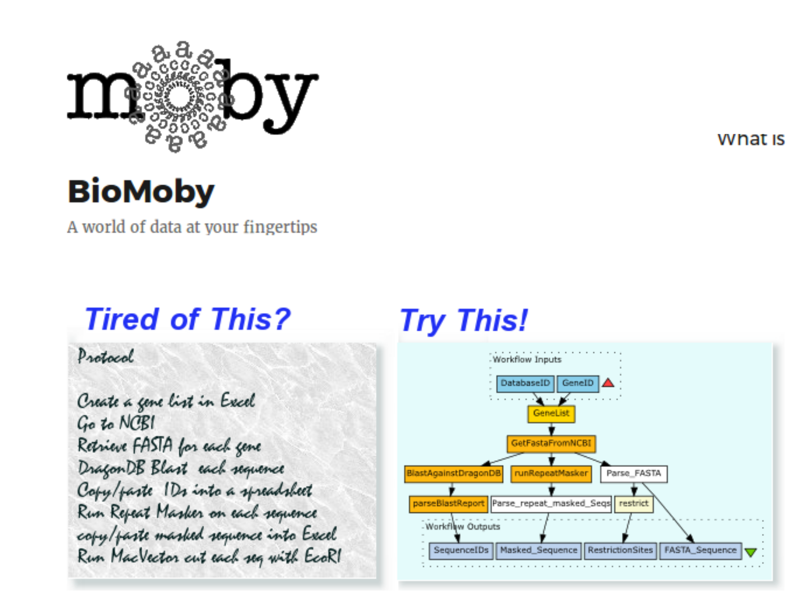
GENOMESPACE

Frictionless connection of bioinformatics tools

Register | User Login

STATUS 04.11.17 09:09PM 🟩 Currently all systems are operating normally.

Calendar of Upcoming Events



moby

What is

BioMoby

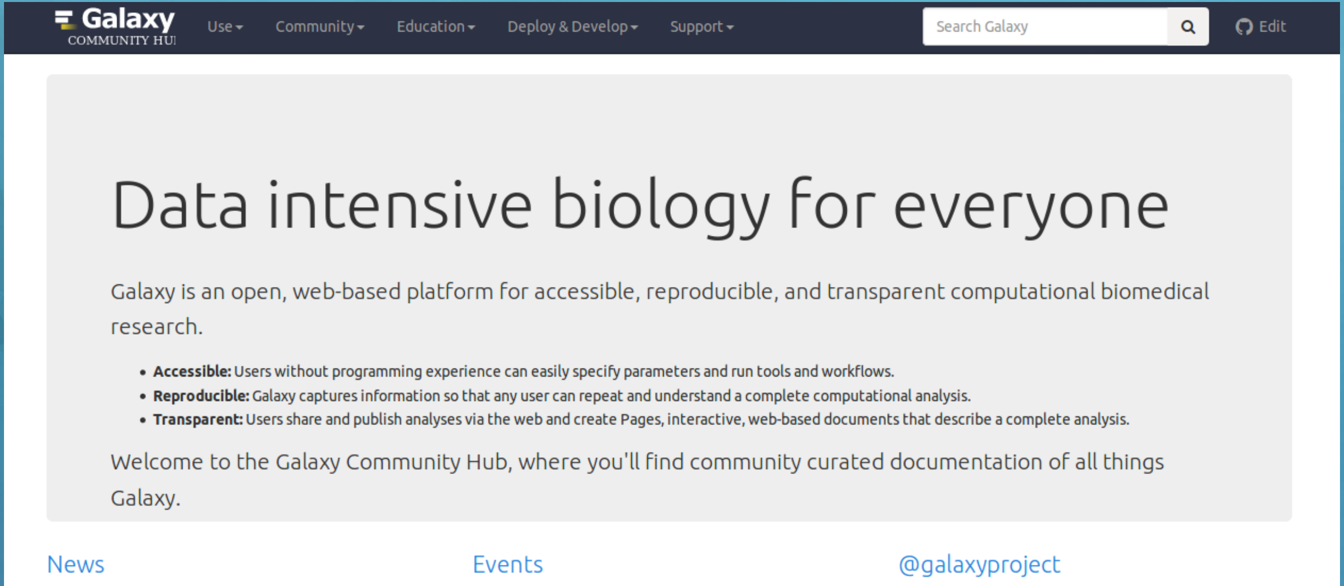
A world of data at your fingertips

Tired of This?

Protocol

Create a gene list in Excel
Go to NCBI
Retrieve FASTA for each gene
DragonDB Blast each sequence
Copy/paste IDs into a spreadsheet
Run Repeat Masker on each sequence
copy/paste masked sequence into Excel
Run MacVector cut each seq with EcoRI

Try This!



Galaxy COMMUNITY HUB

Use▾ Community▾ Education▾ Deploy & Develop▾ Support▾ Search Galaxy Edit

Data intensive biology for everyone

Galaxy is an open, web-based platform for accessible, reproducible, and transparent computational biomedical research.

- **Accessible:** Users without programming experience can easily specify parameters and run tools and workflows.
- **Reproducible:** Galaxy captures information so that any user can repeat and understand a complete computational analysis.
- **Transparent:** Users share and publish analyses via the web and create Pages, interactive, web-based documents that describe a complete analysis.

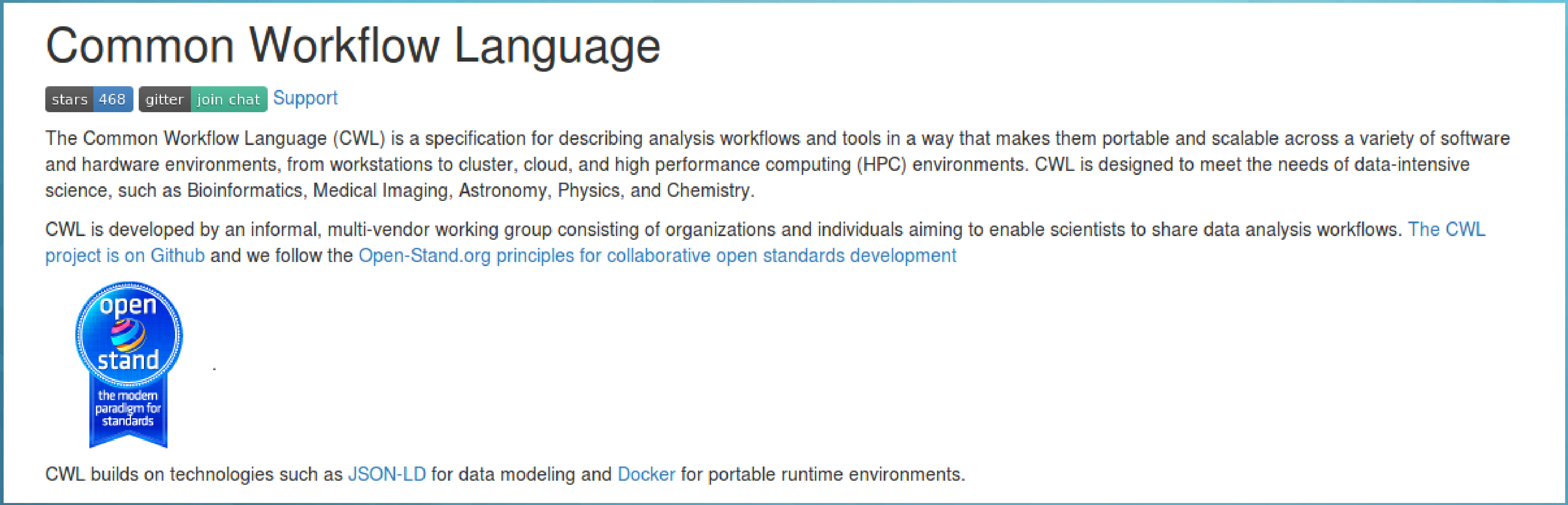
Welcome to the Galaxy Community Hub, where you'll find community curated documentation of all things Galaxy.

News Events @galaxyproject

analysis recipes
can be published

Documentation – Workflows

http://www.commonwl.org



Common Workflow Language

stars 468 gitter join chat Support

The Common Workflow Language (CWL) is a specification for describing analysis workflows and tools in a way that makes them portable and scalable across a variety of software and hardware environments, from workstations to cluster, cloud, and high performance computing (HPC) environments. CWL is designed to meet the needs of data-intensive science, such as Bioinformatics, Medical Imaging, Astronomy, Physics, and Chemistry.

CWL is developed by an informal, multi-vendor working group consisting of organizations and individuals aiming to enable scientists to share data analysis workflows. The CWL project is on Github and we follow the Open-Stand.org principles for collaborative open standards development

CWL builds on technologies such as JSON-LD for data modeling and Docker for portable runtime environments.

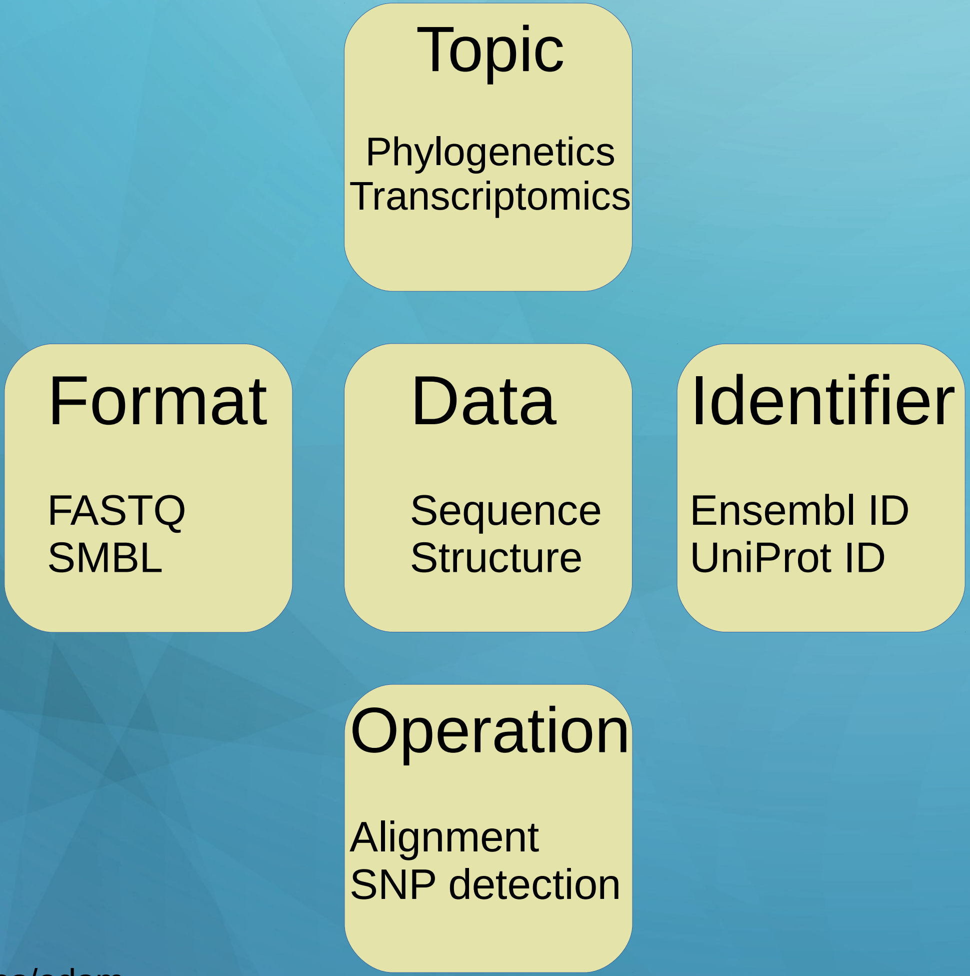
strategic break

http://www.toggo.de

Documentation – Ontology

EDAM Ontology

EDAM provides a set of concepts with preferred terms and synonyms, definitions, and some additional information - organized into an intuitive hierarchy



Topic

Phylogenetics
Transcriptomics

Format

FASTQ
SMBL

Data

Sequence
Structure

Identifier

Ensembl ID
UniProt ID

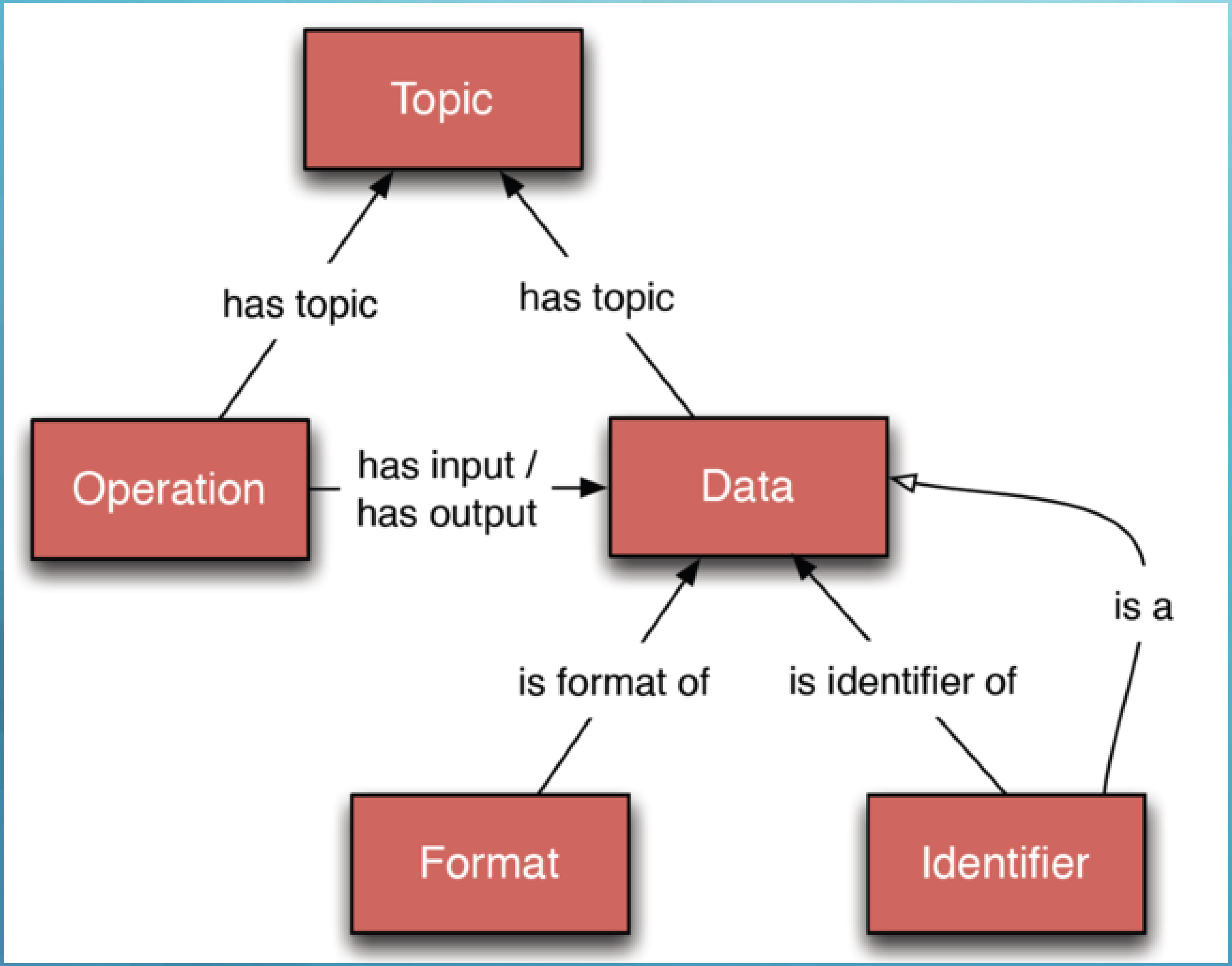
Operation

Alignment
SNP detection

http://edamontology.org
http://www.ebi.ac.uk/ols/ontologies/edam
http://bioportal.bioontology.org/ontologies/EDAM?p=classes

Documentation – Ontology

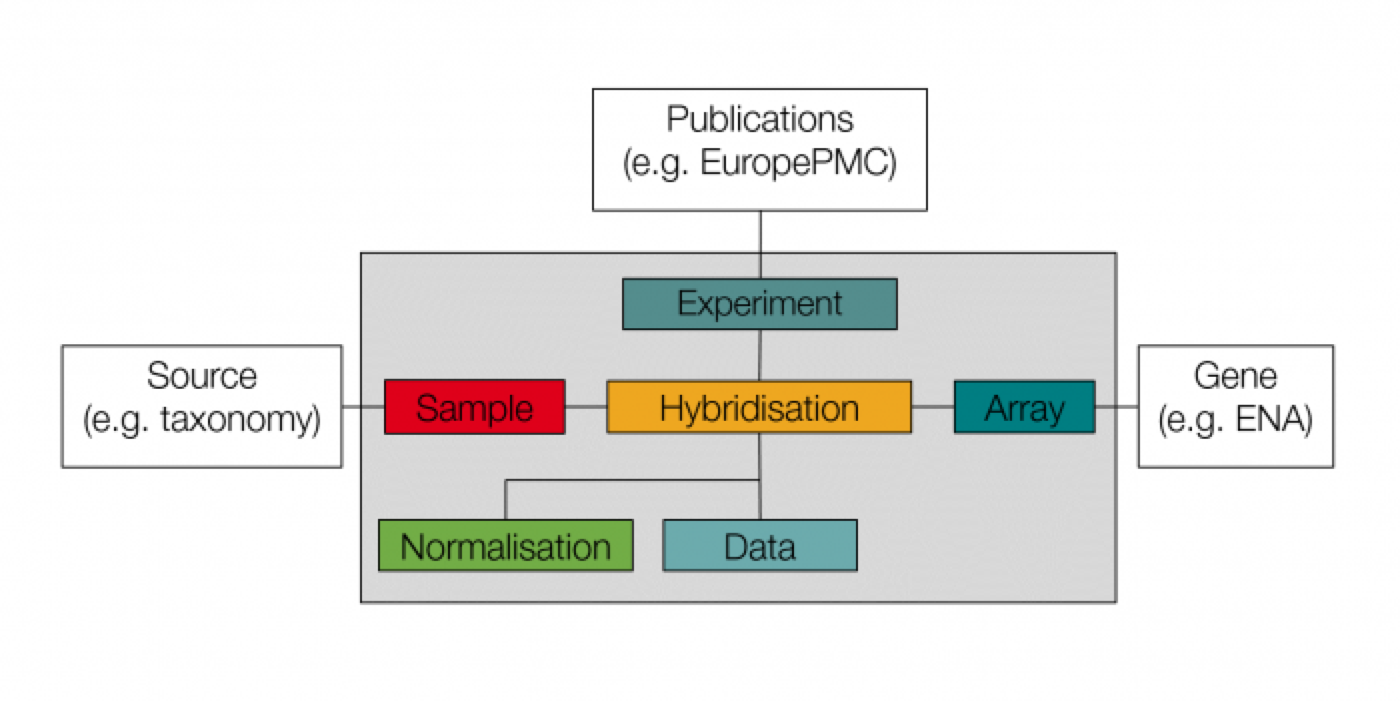
EDAM Ontology structure



Documentation – minimum information standards

https://en.wikipedia.org/wiki/Minimum_Information_Standards

https://www.ncbi.nlm.nih.gov/geo/info/MIAME.html



https://www.ebi.ac.uk/training/online/course/bioinformatics-terrified/minimum-information-standards

Reproducibility

needs controlled

input data set (machine, protocol, tissue type, disease type)

software environment (operation system)

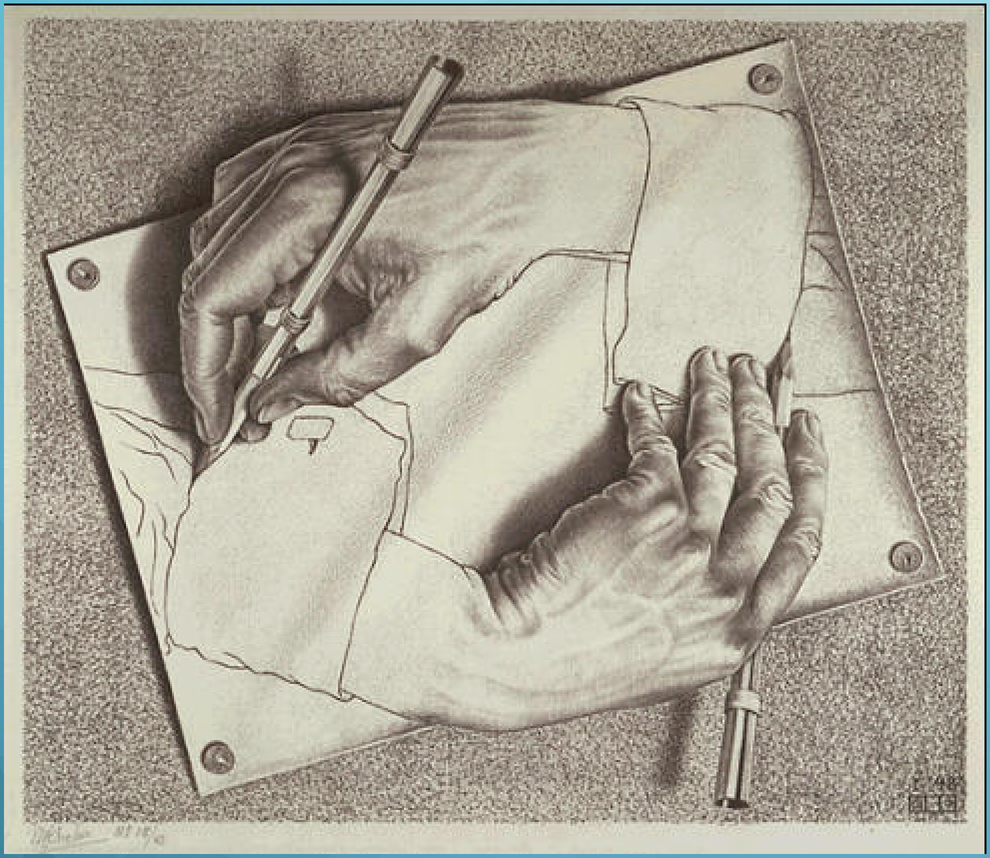
software (version)

parameters

documentation

interpretation

Interpretation



The data interpretation in a clinical context is constantly changing

Interpretation

I favor a data structure that distinguishes between

“findings”

its supporting data

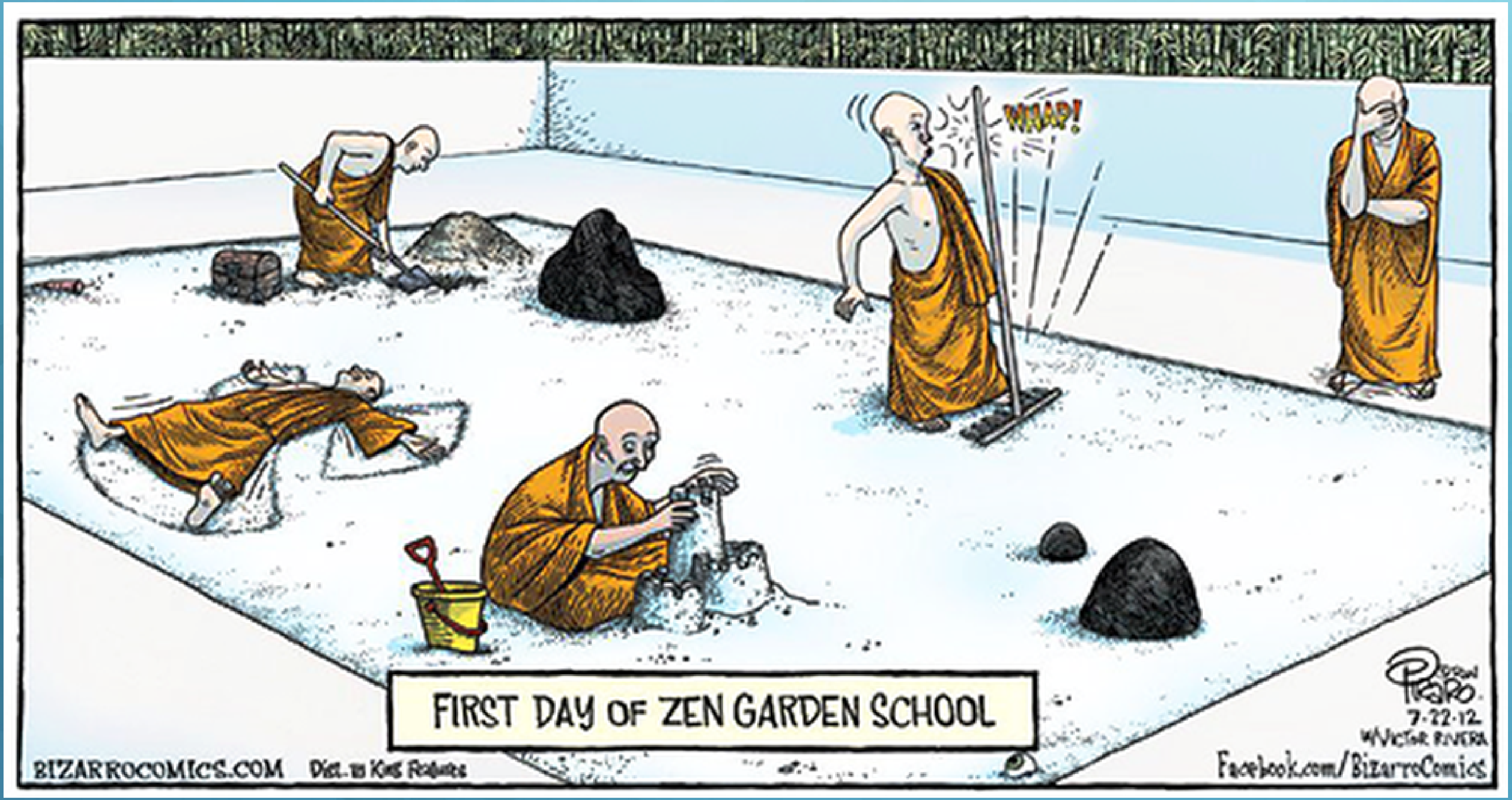
and additional data

“finding” = actionable fact

findings could be grouped by molecular class

findings could be grouped to create meta findings

“Data is a zen garden”



Thank you

Questions?