MAX PLANCK INSTITUTE FOR **BIOLOGY OF AGEING**

Bioinformatics Business Services in 2015: culture, good practices, and structure

http://mpg-age-bioinformatics.github.io

http://tinyurl.com/boucas-zagreb

Jorge.Boucas@age.mpg.de

Bioinformatics Services

Misunderstandings in 2015

Hi there!
I heard you do NMR, MS, dissect flies,
operate the heart of pigs, protein modeling, microbiology, operate the cool HiSeq, apply the latest machine learning / deep learning algorithms to model million of daily publications in an integrative way while genotyping hundreds of mice!
Cool!

IT guy

Lab guy

Common misunderstandings in 2015

Bioinformatics Services

Misunderstandings in 2015

3 data sets

analyze each data set	develop general pipeline
time: 90 hours	90 h for development
turnover time: 30 hours / dataset	turnover time: 1 hour / dataset
collaboration partner sees this as: *"lots of work"*	collaboration partner sees this as: *"all you do is running scripts"*

Solution: transform the job into supplying the code

Bioinformatics Services

Misunderstandings in 2015

Guided analytics templates for self-service analytics

Solution: transform the job into supplying the code

Bioinformatics Services

What, Who, How

Challenge:

1. High speed of development
2. Continuous increase in data types and sizes
3. Limited personnel (worldwide)
4. Know-how abysm between wet-bench and dry-analytics (and vice versa)

Solution:

A. Biology-driven analysis

B. Reward for code environment

C. Increase communication/interaction

D. High-level team work

Bioinformatics Services

What, Who, How

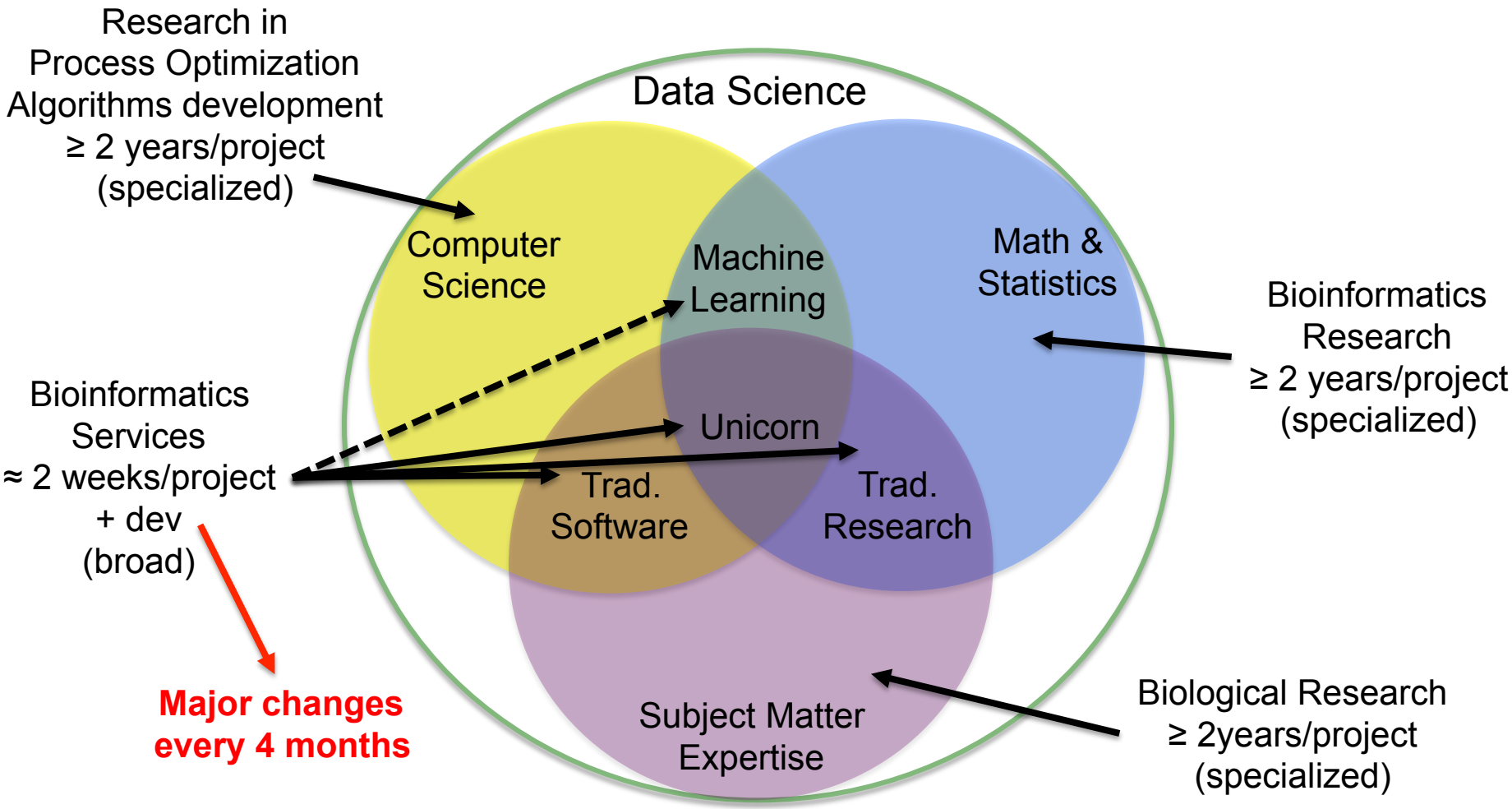
Bioinformatics is an interdisciplinary field that develops methods and software tools for understanding biological data.

As an interdisciplinary field of science, bioinformatics combines **computer science, statistics, mathematics, and engineering** to analyze and interpret **biological data**.

Source: Wikipedia

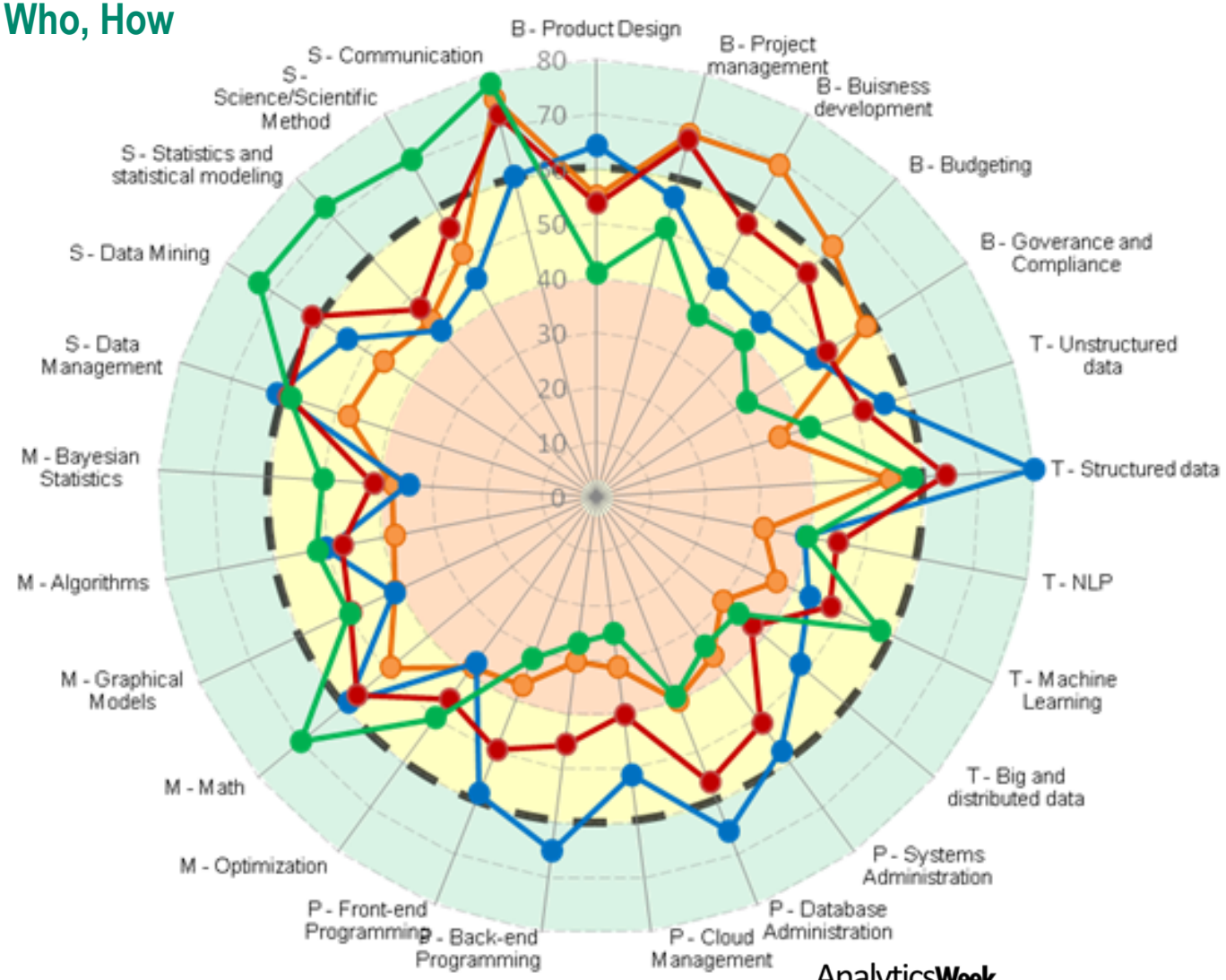
Bioinformatics Services

Data Science



Bioinformatics Services

What, Who, How

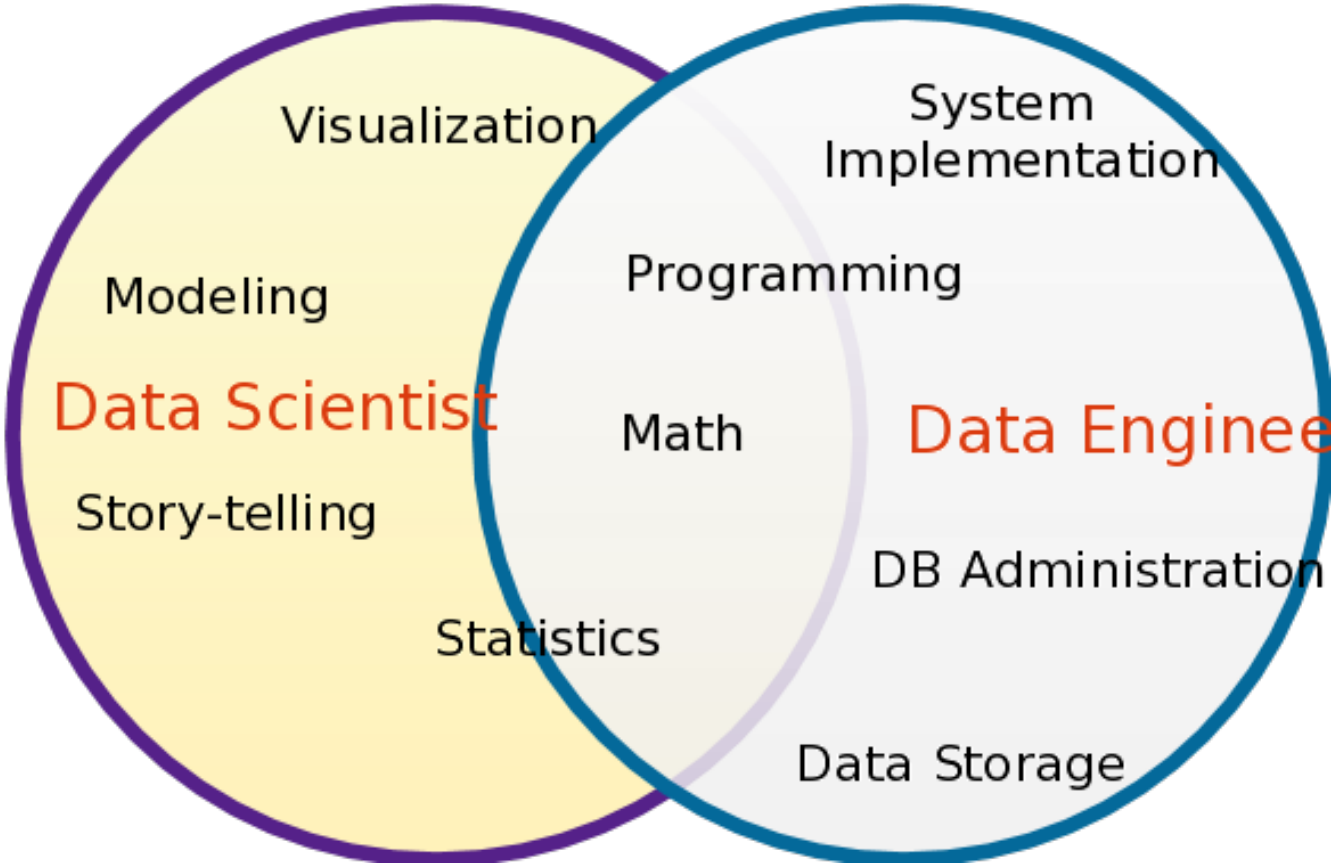


AnalyticsWeek

BUSINESS BROADWAY

Bioinformatics Services

What, Who, How

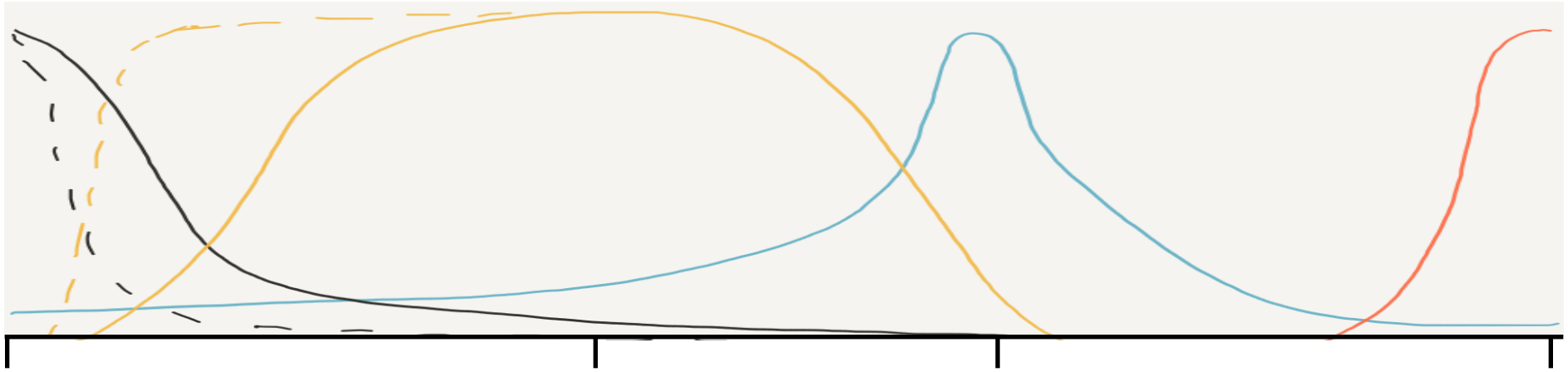


Yuki Katoh
Senior data scientist at mbr targeting
Cologne, 18.09.2015

http://101.datascience.community/2014/07/08/data-scientist-vs-data-engineer/

Bioinformatics Services

A linear perspective and the hole between Biology and Bioinformatics



Biology
2 years/proj.
(specialized)

Bioinformatics
Services
2 weeks/proj.
(broad)

Bioinformatics
Research
2 years/proj.
(specialized)

Computer
Science
2 years/proj.
(specialized)

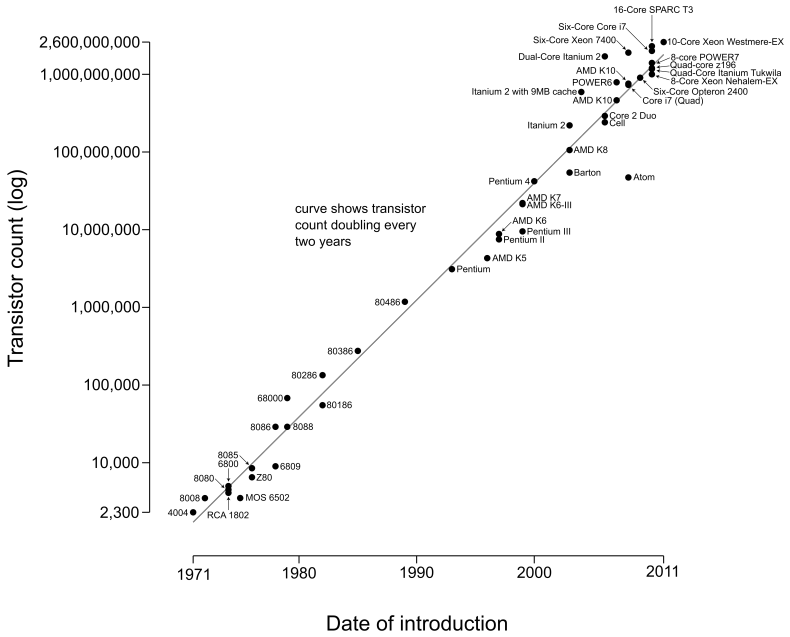
If you want your organization to progress you have to help all users!

At all levels!

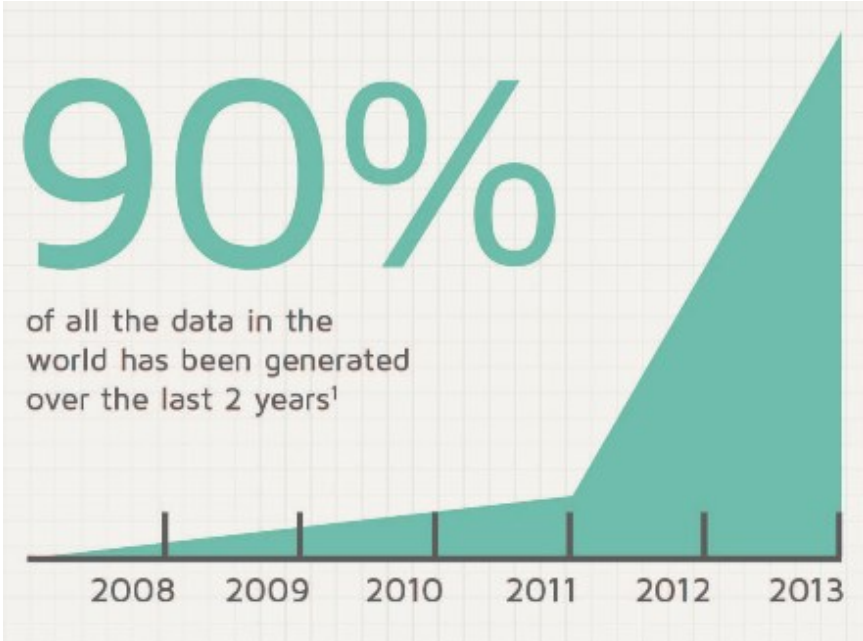
Bioinformatics Services

The origin of the hole

Microprocessor Transistor Counts 1971-2011 & Moore's Law



"Number of transistors per square inch on integrated circuits had doubled every year."
Source: Wikipedia

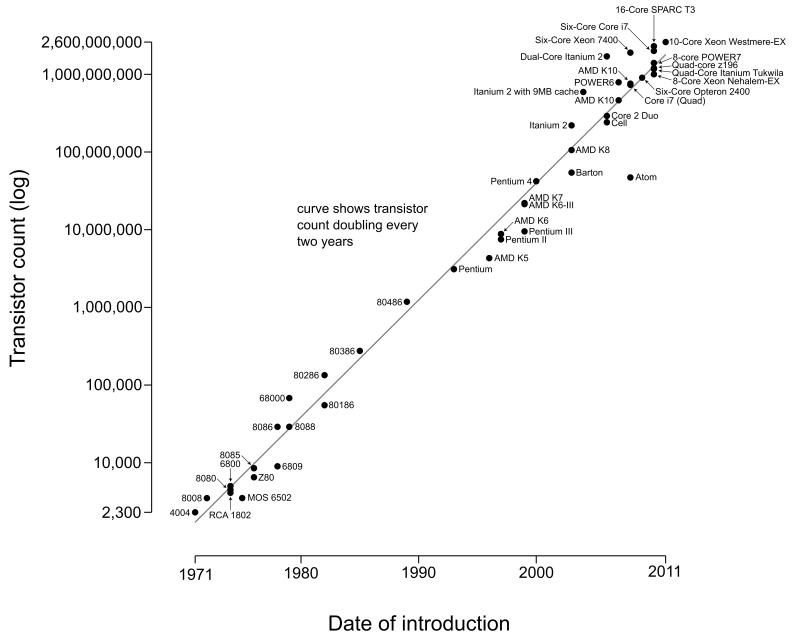


Source: www.projects.ac

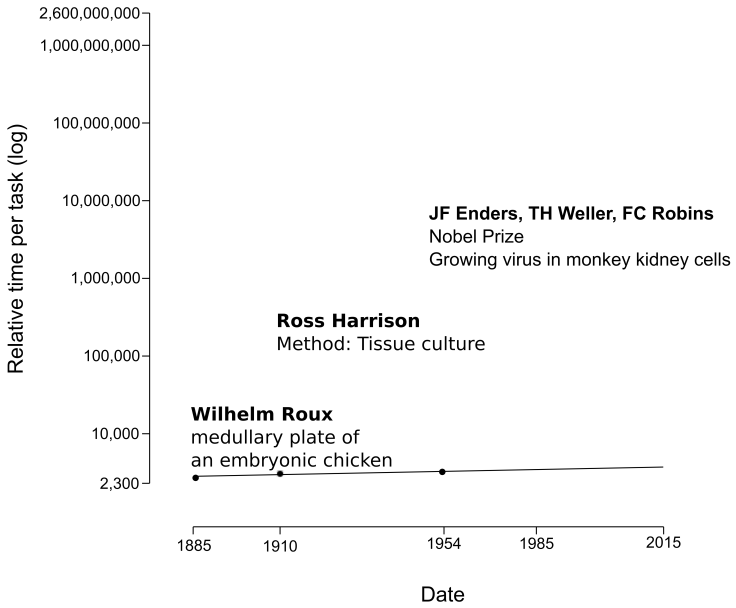
Bioinformatics Services

The origin of the hole

Microprocessor Transistor Counts 1971-2011 & Moore's Law



Cell culture 1885-2015 (Fictitious figure)



"Number of transistors per square inch on integrated circuits had doubled every year."
Source: Wikipedia

> 80 % of academic labs → **> 80 % deficit in automation**

Users can't find the time to keep up with dev. in Bioinformatics !

Bioinformatics Services

How to fix it

Do !

Think like a developer.

No “hard coding”.

Your focus is results that go on papers not beautiful software.

Expose !

Share your code with the user.

Share your code with the public.

Project tracker for you and the user.

Support!

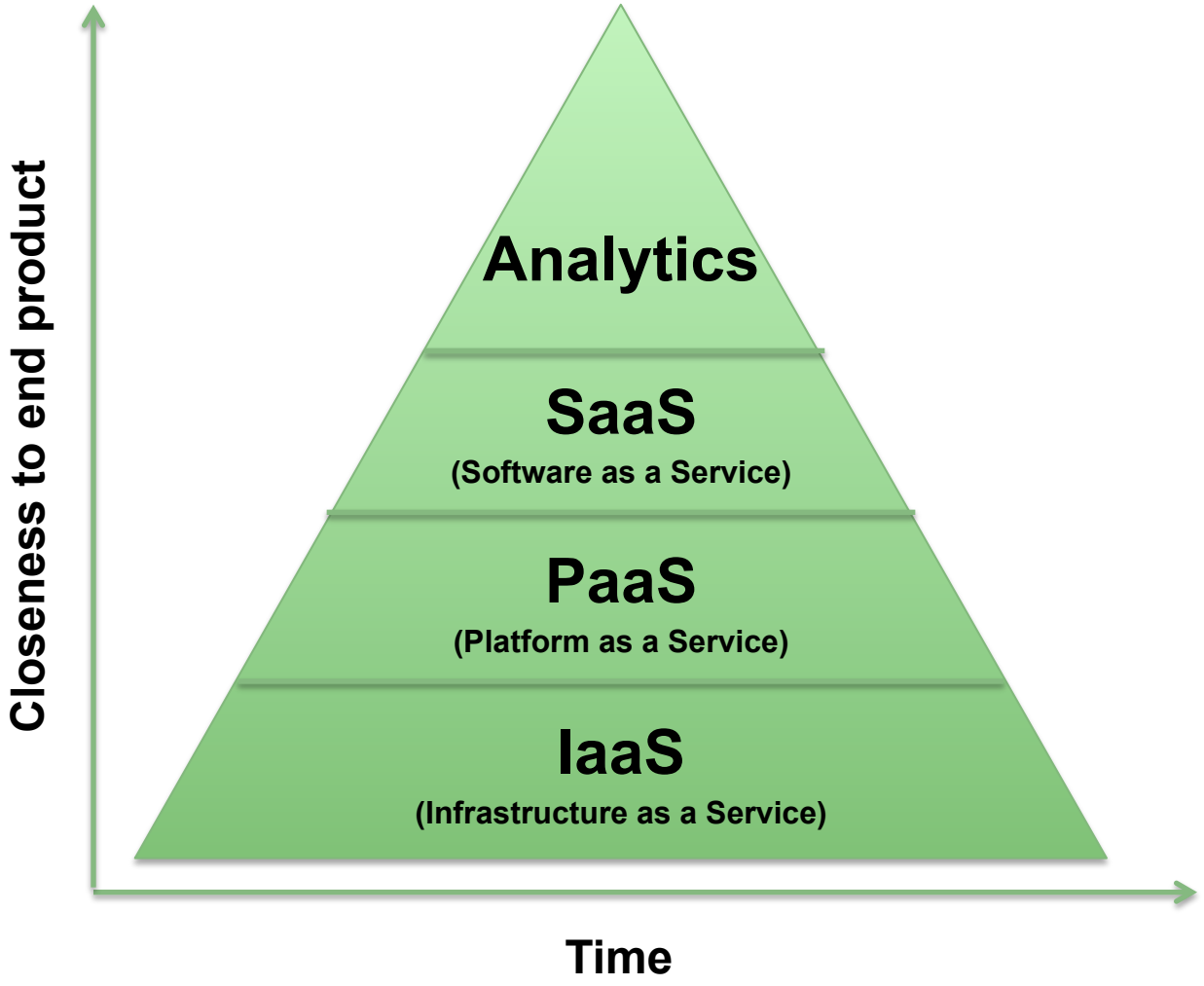
Open door culture.

2 guest terminals in bioinformatics office

Weekly
2 hours sessions
open to all users
with
hands-on support

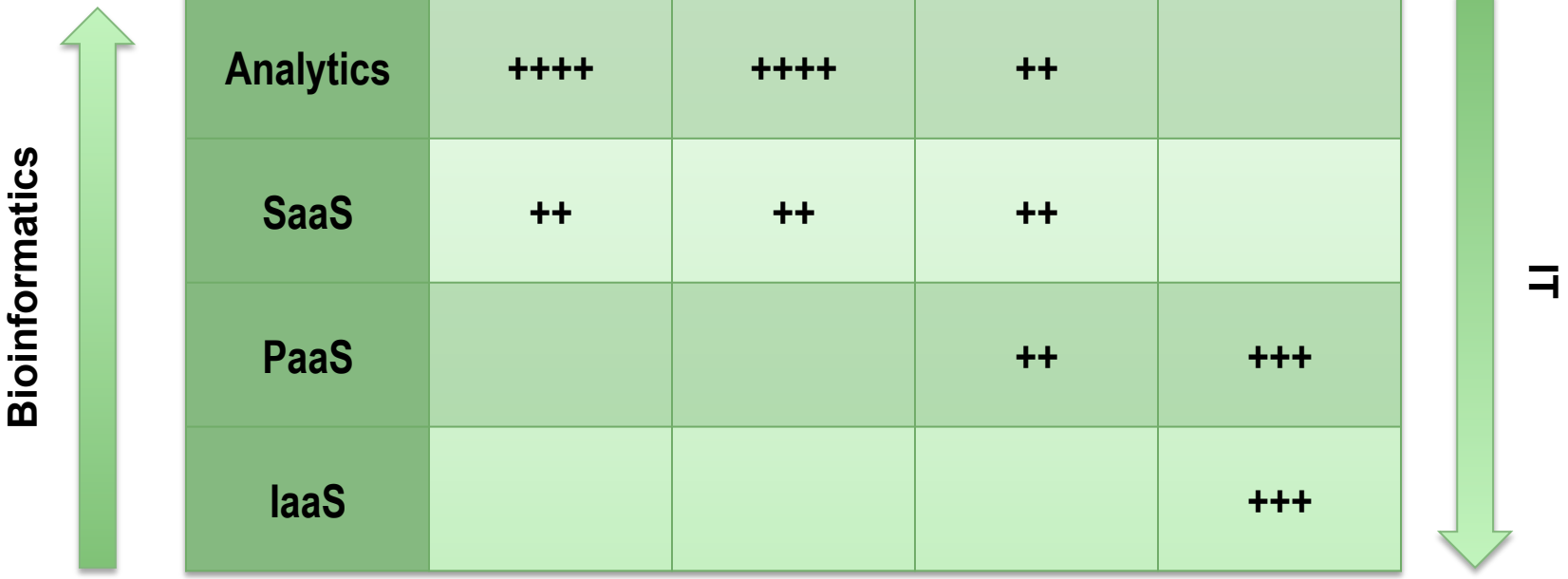
Do!

What


Analytics
SaaS
(Software as a Service)
PaaS
(Platform as a Service)
IaaS
(Infrastructure as a Service)
Closeness to end product
Time

Do!

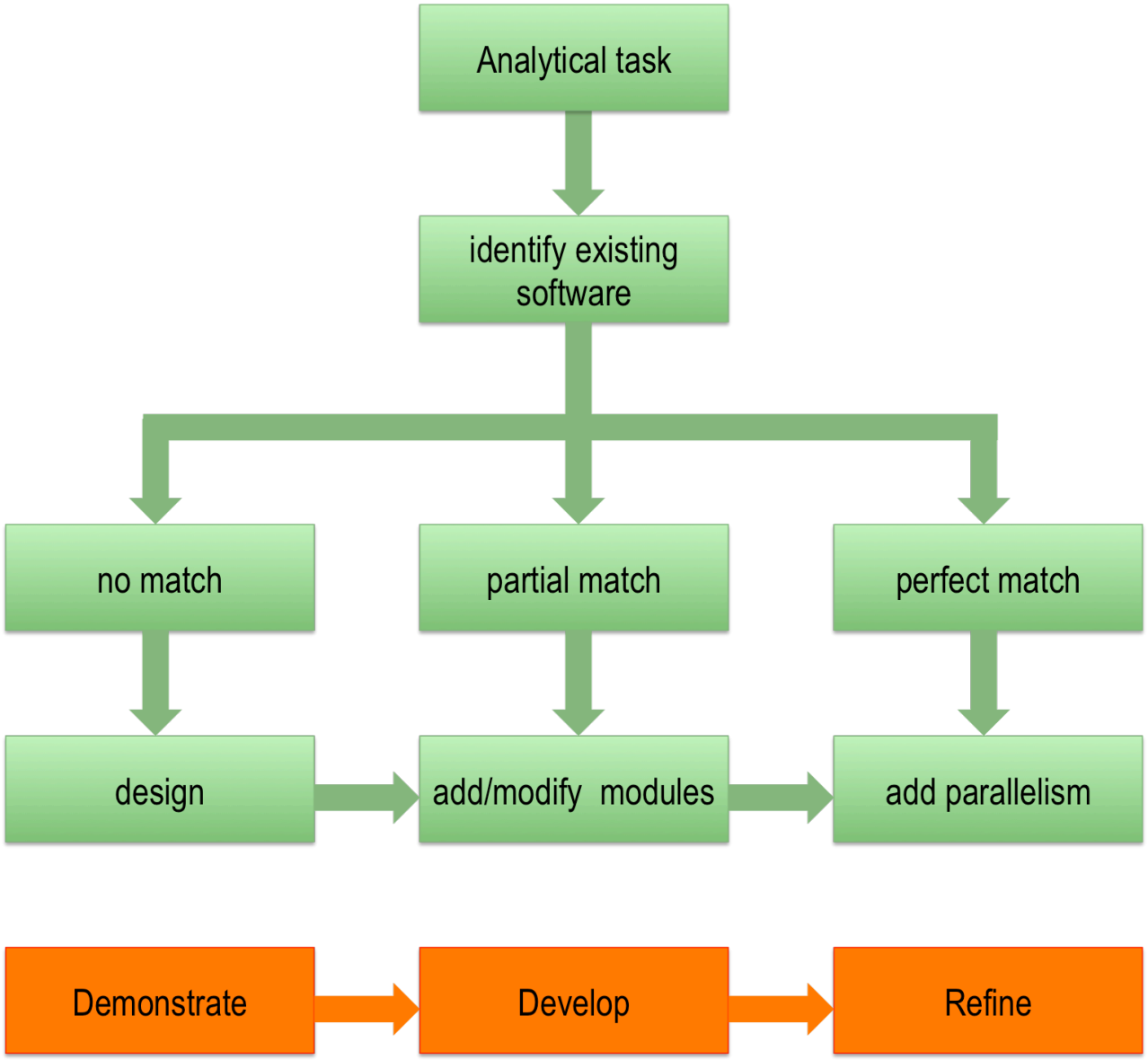
Who



	John	Joe	Jim	Jay
Analytics	++++	++++	++	
SaaS	++	++	++	
PaaS			++	+++
IaaS				+++

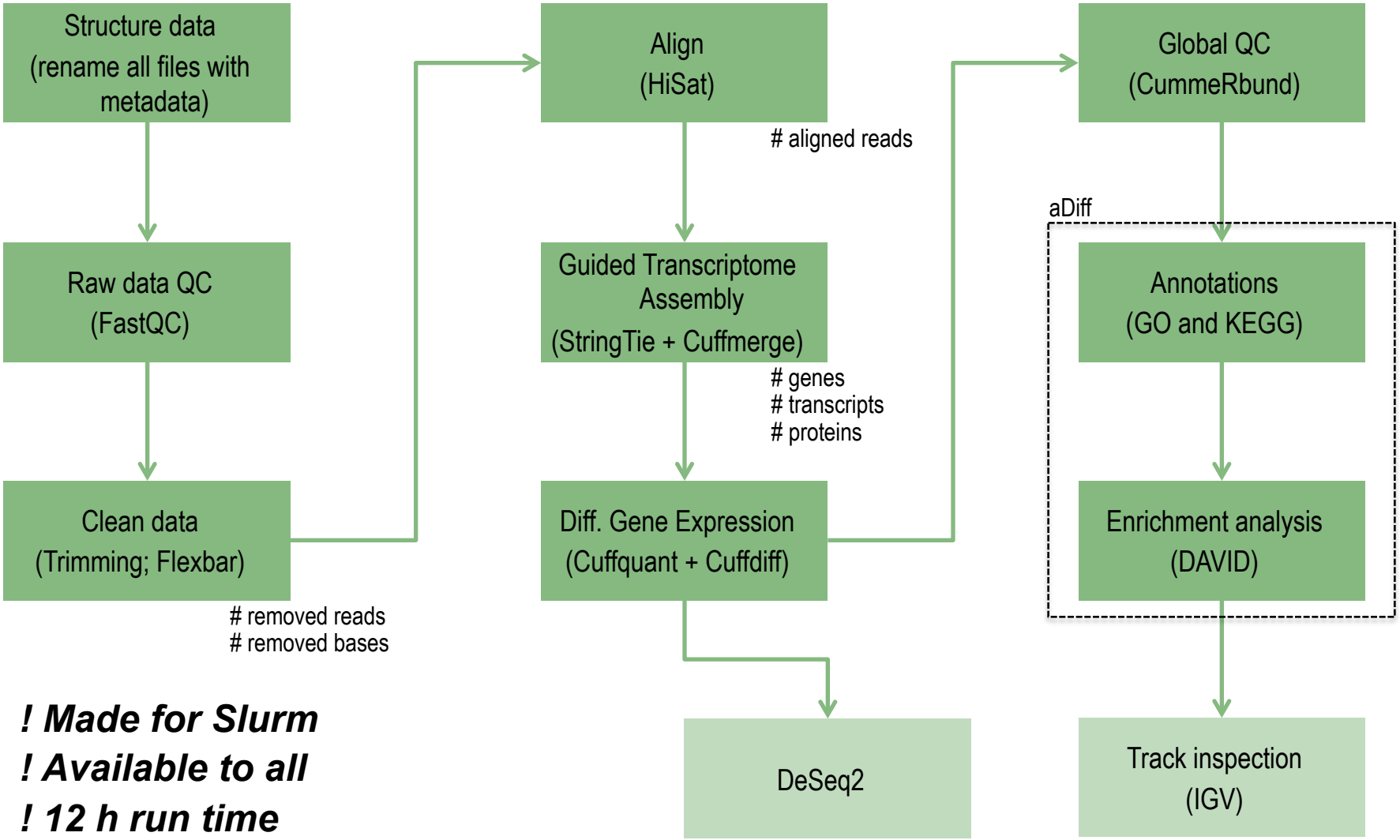
Do!

How


Analytical task
identify existing software
no match
partial match
perfect match
design
add/modify modules
add parallelism
Demonstrate
Develop
Refine

Do!

Analytics



! Made for Slurm
! Available to all
! 12 h run time

Do!

Structure data and projects

Data lake friendly!

Folders structure

```
Projects
├─ department
│  ├─ project_name
│  │  ├─ raw_data
│  │  ├─ code
│  │  ├─ tmp
│  │  ├─ software1_output
│  │  └─ software2_output
```

```
Projects
├─ John_Doe
│  ├─ JD_miRNA_banana
│  │  ├─ raw_data
│  │  ├─ scripts
│  │  ├─ tmp
│  │  ├─ tophat_output
│  │  └─ cufflinks_output
```

Do!

Structure data and projects

Data lake friendly!

raw_data structure (ln –s)

S_XXX-F_XXXX-**L_XXXXX**-XXX-**XXXX**-REP_X-**READ_x**.fastq.gz

Sample_serial – Project– **genotype** – Time_point – **treatment** – REPlicate – **READ**

S_001-F_HaTS-**L____N2**-__0-____-REP_1-**READ_1**.fastq.gz

Do!

Step wise development!

- No dataset is a perfect standard.
- Standards change.
- First automated pipeline after the 3rd data set.
- Results > Development
- Your incoherent analysis might be the user's Nature paper.
- Academic Post-doc half-life: 2 years

Do!

Agile!

Deliver working software / results frequently, from a couple of weeks to a couple of months, with a preference to the shorter timescale.

Modified from: Principles behind the Agile Manifesto, agilemanifesto.org

Do!

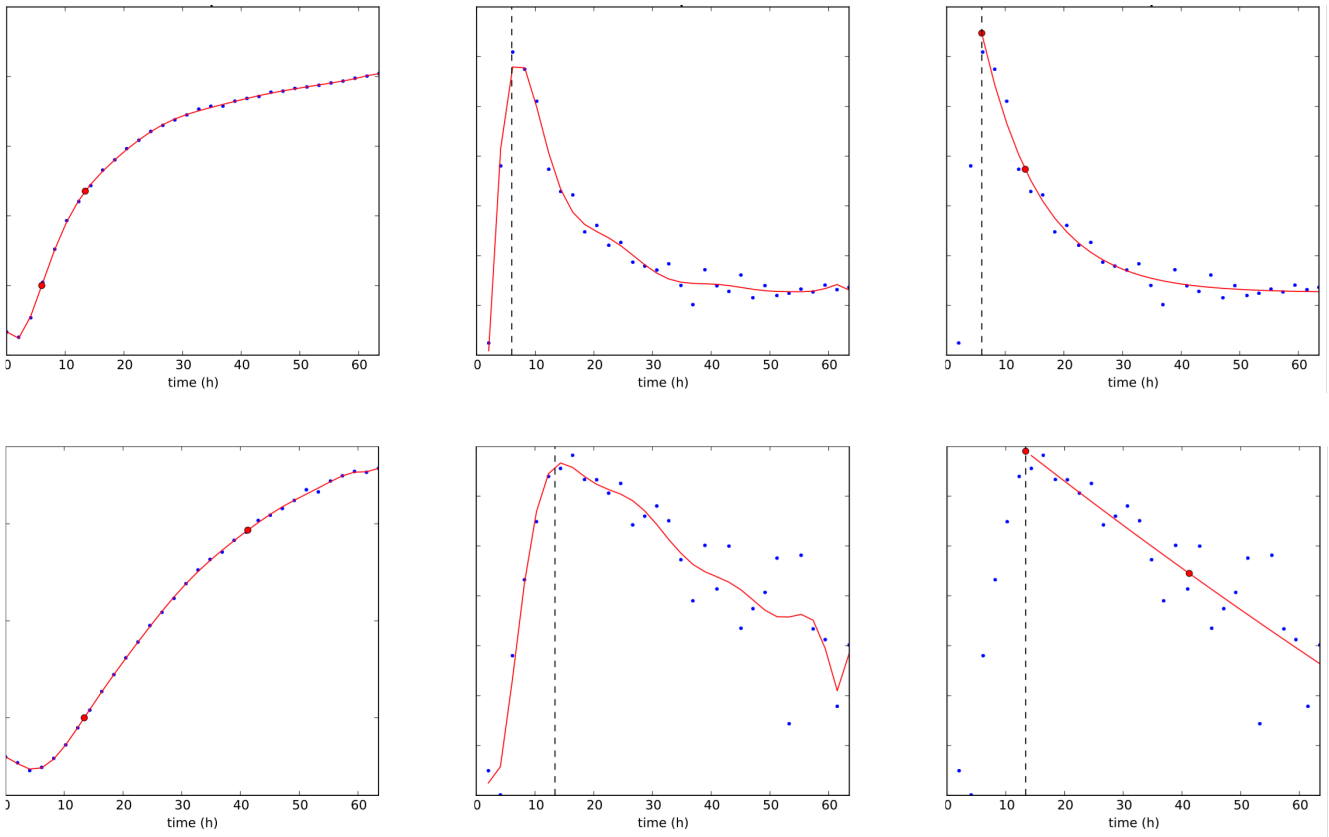
Software as a Service

384 data points every 2.5 minutes

txt files with raw data and time stamps

Curve fitting, extrapolation, calculation, plotting

Traceable report in Excel, PDF, PNG, SVG



! Locally in any computer
! 3 min run time

! Python
! Galaxy Friendly

Do!

Agile!

Business people and developers

must work together

daily throughout the project.

Principles behind the Agile Manifesto, agilemanifesto.org

Do!

Software as a Service

Software customization

vs

Specialized software

Do!

Platform as a Service

OPEN SOURCE	REDMINE flexible project management	http://hostedredmine.com (partially free online) Project Management; Wiki; Forums; Virtual Journal Club
OPEN SOURCE	GitLab	http://gitlab.com (fully free online) Versioning & Sharing of Code
OPEN SOURCE	Galaxy	http://galaxyproject.org (partially free online) Code Free Programming; Sharing of Pipelines; HPC integration
OPEN SOURCE	RStudio	http://www.rstudio.com (as server) Statistical computing; HPC integration
COSTLY	HPC	ssh UName@cluster Centralized software – modules; Databases; Genomes; Indexes; Systems administration
OPEN SOURCE	ownCloud	provided by IT (test online: owncube.com) data delivery with temporary links over email – NO EDITING! Don‘t delete (2-3 years) / storage is cheap, confusion not!

Do!

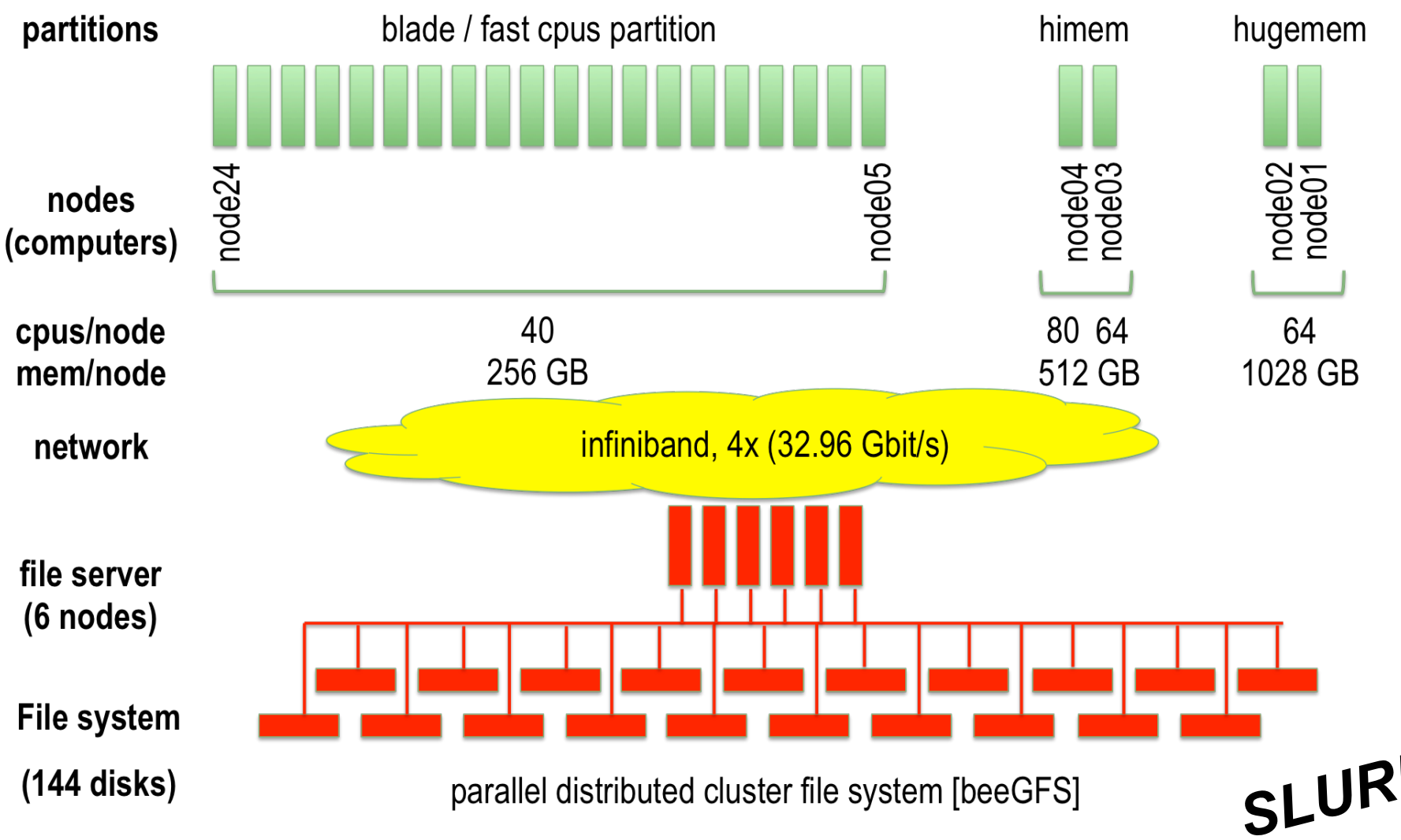
Platform as a Service

Bioinformaticians as Galaxy users

Bioinformaticians as Galaxy dev.

Do!

Infrastructure as a Service



Do!

Infrastructure as a Service

Backups:

1. raw_data folder
    1. Cluster file system (FS1)
    2. Alternative/backup file system (FS2)
    3. User

2. scripts folder
    1. Cluster file system (FS1)
    2. GIT (FS3)
    3. GIT backup (FS2)

Do!

Infrastructure as a Service

Cluster on-demand:

1. once upload and download of large data is efficient

2. once cluster can be easily set on-demand

http://gc3-uzh-ch.github.io/elasticluster/

http://star.mit.edu/cluster/

Year: 2022 ?

Alternative:

Centralized cluster resources: eg. national / regional

Expose!

Code & Project!

Every project gets:

1. an entry - JD_miRNA_banana - in Redmine (proj. management) with relevant proj. information

2. a GitLab repository - JD_miRNA_banana - with all code

All updated and shared daily with the user.

(code can be open at end of project - eg. Github)

Expose!

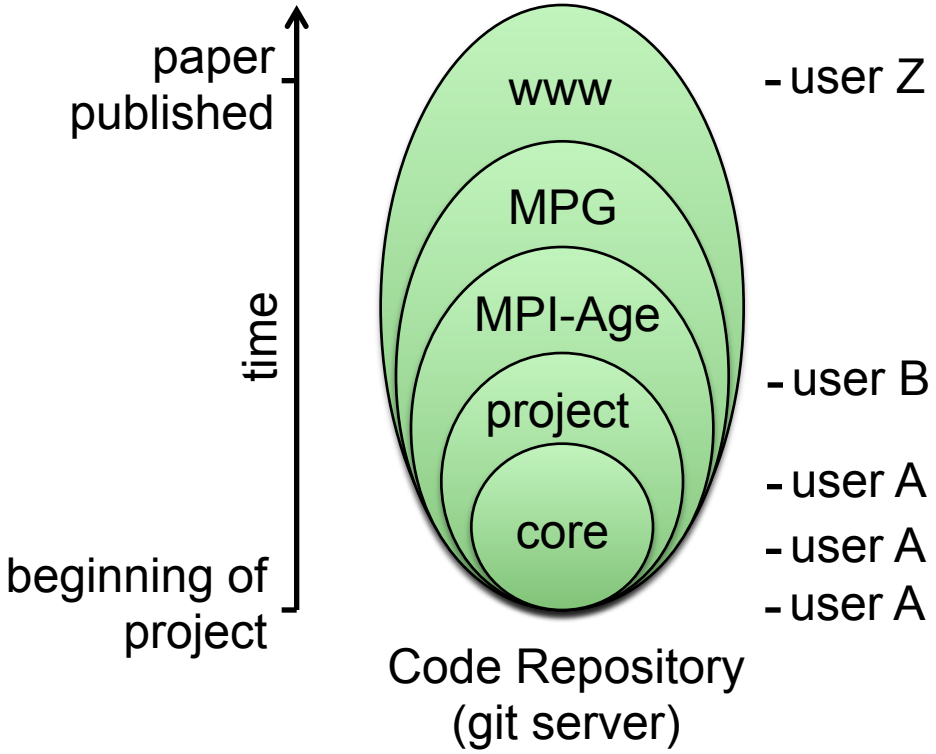
Code & Project!

Give away know-how?

Not good enough for sharing?

Loss of intellectual property?

Misusage / overwhelming bug reporting?



Traceable centralized repository

Authorships

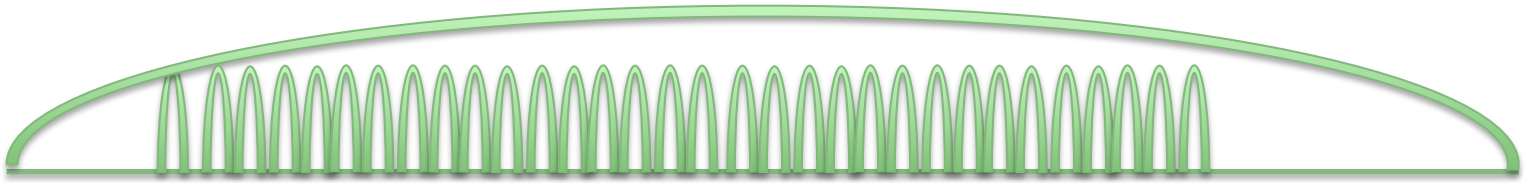
the no authorship policy

You don’t need a computer
to do great research!

You can do great research with
great computers!

Authorships

the no authorship policy



Wet-bench biology

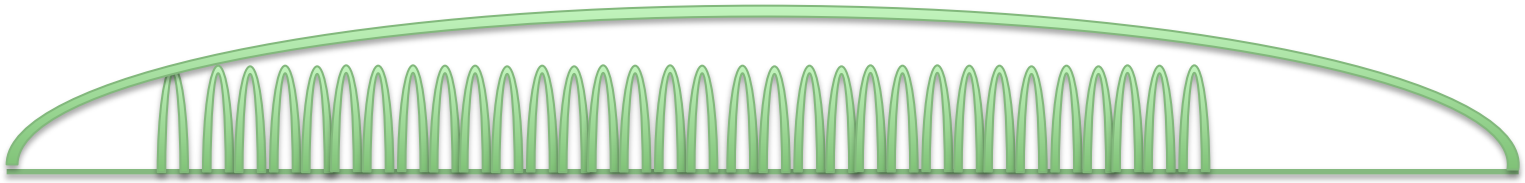
Dry-bench bioinformatics

Work anytime anywhere in a large variety of projects involving a wide range of technologies.

Work at home at 3:00 AM on a project of no interest to the analyst.

Authorships

the no authorship policy



Wet-bench biology

Dry-bench bioinformatics

Users have different needs.

We supply a **service** for each need.

We don't drive our own research.

We **support** the research of others.

We don't set project priority based on own scientific interest.

We don't choose our users.

≠

Research

Scientific

Career

Authorships

if you want an academic career

- 4-6 projects / 2 years / group member
- 2 grants / 2 years / group member
- Decide/enforce wet bench experimental design
- Enforce first/shared authorship for bioinformatics group members
- Enforce last/shared authorship for bioinformatics group leader

Bioinformatics Services

sustainability

- Friday afternoon it's developer's time!
- Push your work on a weekly basis to open git repositories (traceable)
- 4 international meetings / group member / year
- 1-2 technical and/or career development events / year
- Space for guest bioinformaticians (eg. *associated with research groups*)
- >= 1 central development project
- 2 days home-office / month (get that work done!)

Summary

A. Biology-driven analysis

B. Reward for code environment - **active Git**

C. Increase communication/interaction - **Redmine, Git, 2 seating spaces, ..**

D. High-level team work - **Structure, HPC cluster, Git, Redmine**

---

1. Code sharing promotes better coding practices
2. Active git repositories are synonyms of productivity, efficiency, and quality
3. Active git repositories push the value of an institution and of the developers
4. Active git repositories attract and promote highly motivated individuals
5. Open source publication of fixes and add-ons promotes further development of these items by the community with subsequent in-house profit
6. Code associated with papers improves the paper's weight