enterome
bioscience

Agile pipelines with Nextflow: how to go from development to production without pain

Francesco Strozzi

Head of Bioinformatics

Contents

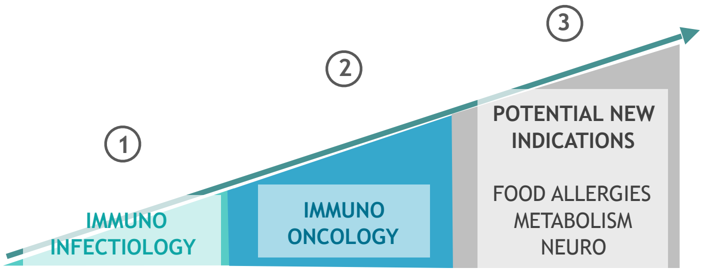
- Enterome core activities
- Our experience from zero to full Nextflow in production on AWS
- Lessons learned
- Spoiler alert: there have been some pain here and there

A clinical stage microbiome company supported by recognised partners and shareholders

enterome
bioscience

- 50 FTES
- Inception 2012
- Pioneer in the development of microbiome-related therapeutic solutions
- 2 complementary platforms identifying gut microbiome changes involved in disease development

Therapeutic potential



Strategic partnerships

IBD

Johnson & Johnson — Dec. 2015

Takeda — Dec. 2015

IMMUNO-ONCOLOGY

Bristol-Myers Squibb — Nov. 2016

MICROBIOME DIAGNOSTIC PARTNERS

enterome, NestléHealthScience (Where Nutrition Becomes Therapy), PROMETHEUS Therapeutics & Diagnostics

Jul. 2017

Shareholders

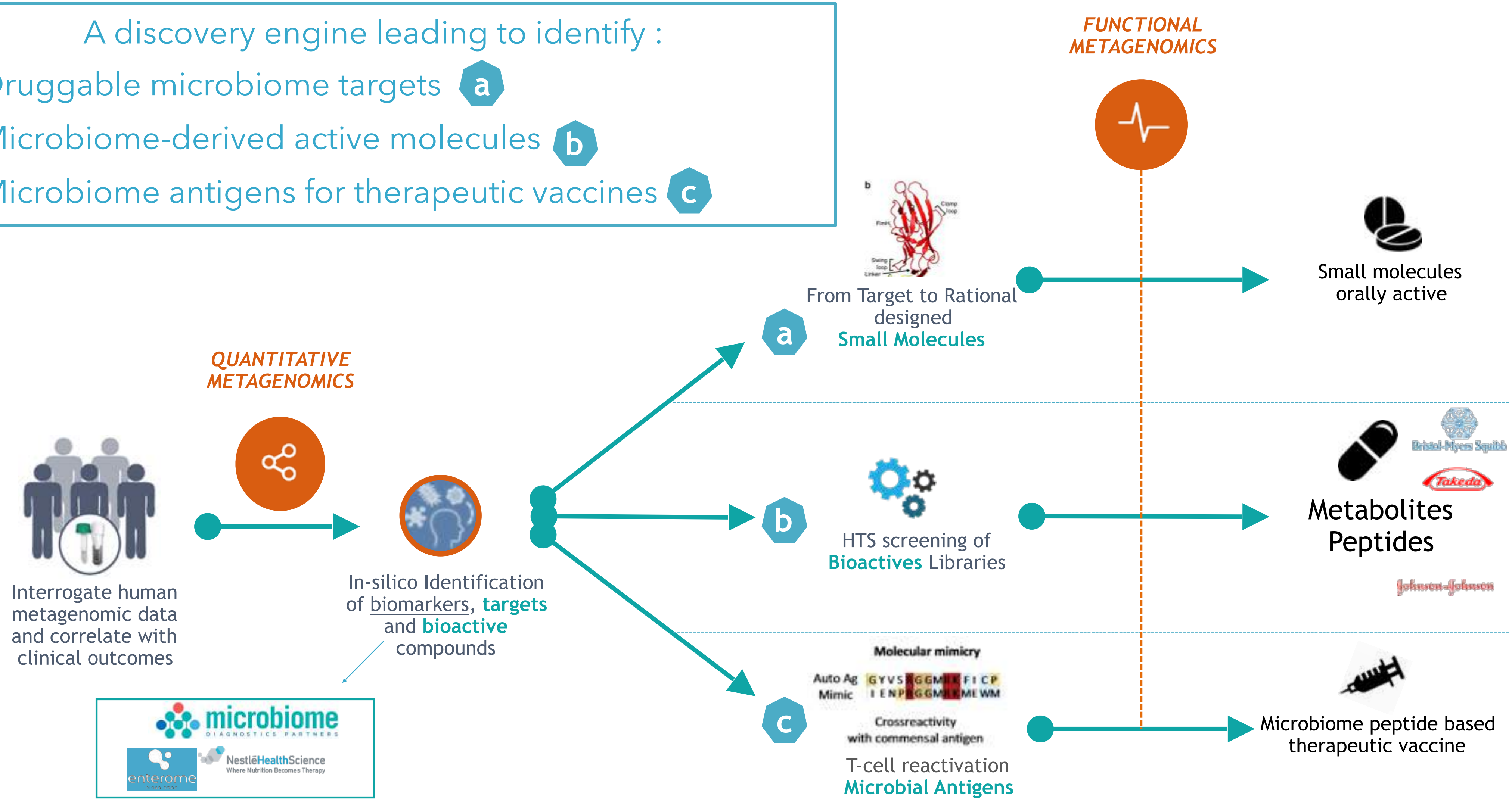
seventure, Health For Life, Lundbeckfonden, NestléHealthScience (Where Nutrition Becomes Therapy), PRINCIPIA SGR, OMNES CAPITAL, Shire, Bristol-Myers Squibb

Developing microbiome related therapies
leveraging on 20+ years of pioneering gut microbiome research INRA

An Integrated Platform : From Correlation to Drug Discovery

A discovery engine leading to identify :

1. Druggable microbiome targets a
2. Microbiome-derived active molecules b
3. Microbiome antigens for therapeutic vaccines c



FUNCTIONAL METAGENOMICS

QUANTITATIVE METAGENOMICS

Interrogate human metagenomic data and correlate with clinical outcomes

In-silico Identification of biomarkers, **targets** and **bioactive** compounds

microbiome DIAGNOSTICS PARTNERS

enterome

NestléHealthScience Where Nutrition Becomes Therapy

a From Target to Rational designed **Small Molecules**

Small molecules orally active

b HTS screening of **Bioactives** Libraries

Metabolites Peptides

Bristol-Myers Squibb

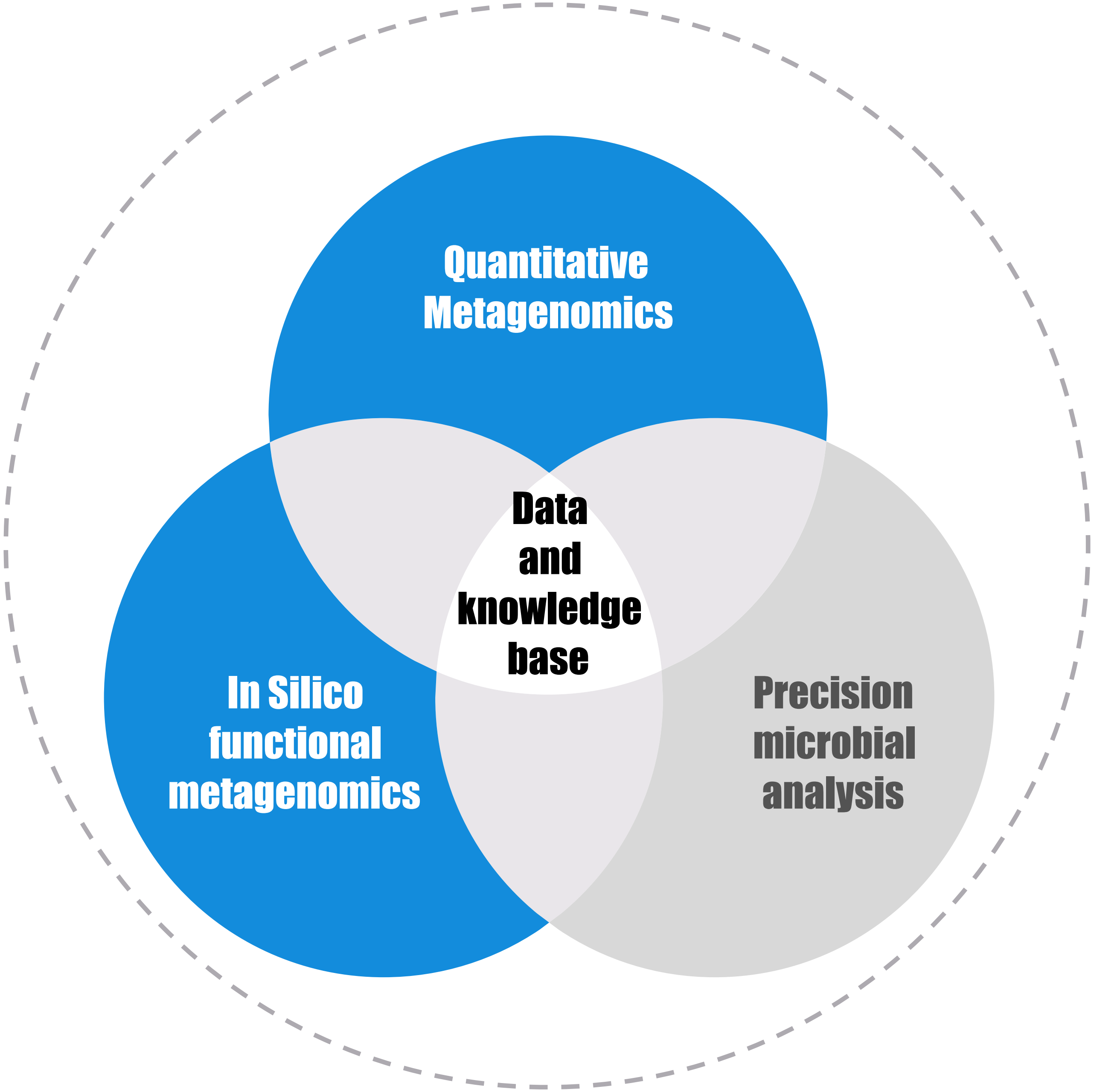
Takeda

Johnson-Johnson

c T-cell reactivation **Microbial Antigens**

Microbiome peptide based therapeutic vaccine

Bioinformatics @ Enterome


Quantitative Metagenomics
In Silico functional metagenomics
Precision microbial analysis
Data and knowledge base

Bioinformatics @ Enterome

- We have several production bioinformatics pipelines

- Many of these pipelines have been qualified and are compliant with ISO13485

- All our pipelines have been translated in Nextflow since 2017

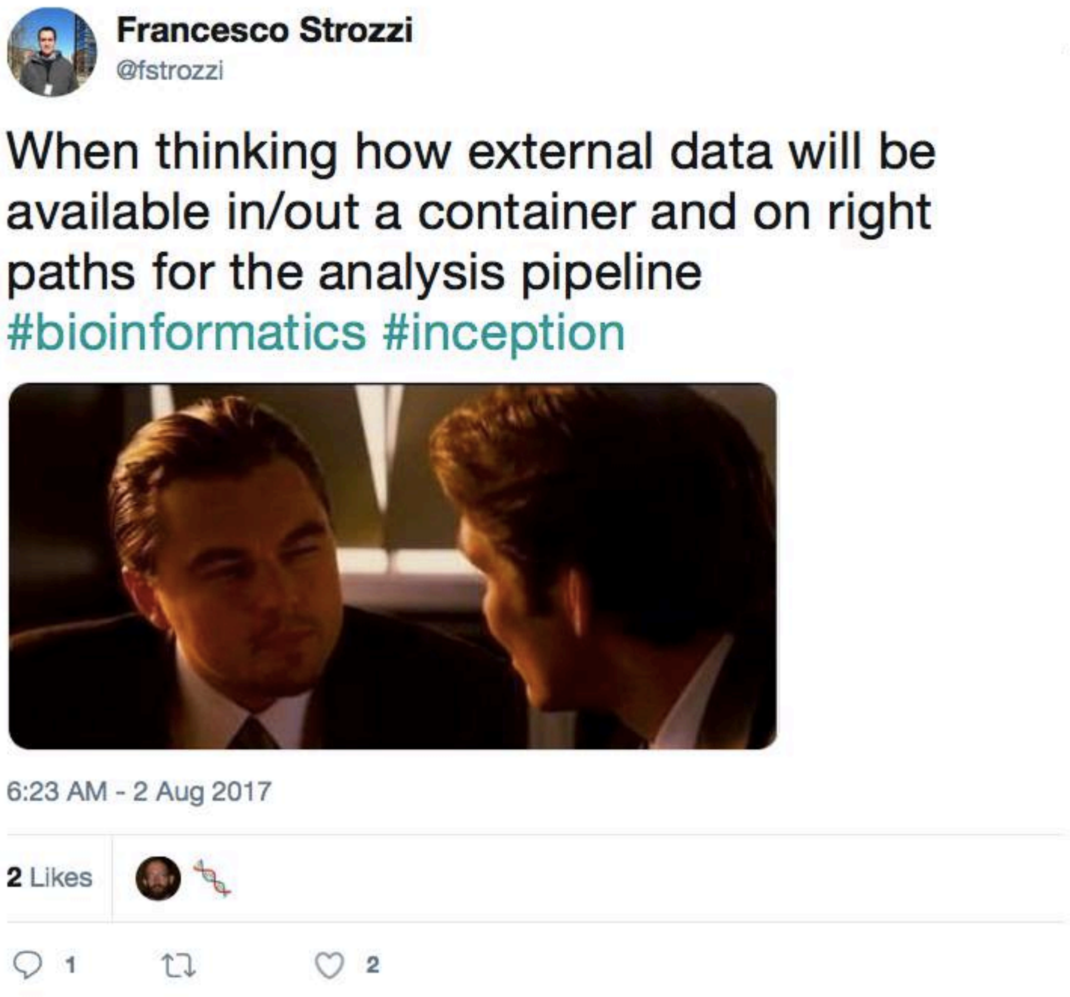
- We are 100% on cloud computing (AWS Batch)
  - ➡ We routinely run large workflows involving thousands of jobs
  - ➡ One workflow can easily consume at its peak usage around 5k-10k CPUs

- Our pipelines are principally focused on
  - ➡ Microbiome profiling from multiple cohort of patients to identify signatures and develop predictive biomarkers
  - ➡ Functional metagenomics analysis of the human gut microbiome to identify new candidates and targets for the drug discovery programs

A year and a half ago: why Nextflow ?

- We needed an effective way to mange our analysis and to make them reproducible across multiple users
- We needed a way to describe workflows that was simple to read and powerful
- We didn't want just a language or a specification to describe workflows, but a functioning framework that could unite description AND execution under the same roof
- We were already using pipeline managers, but they were very limited in both workflows description and execution engines supported
- We needed a framework that could support multiple platforms, especially the cloud

Scaling on the cloud: can be hard



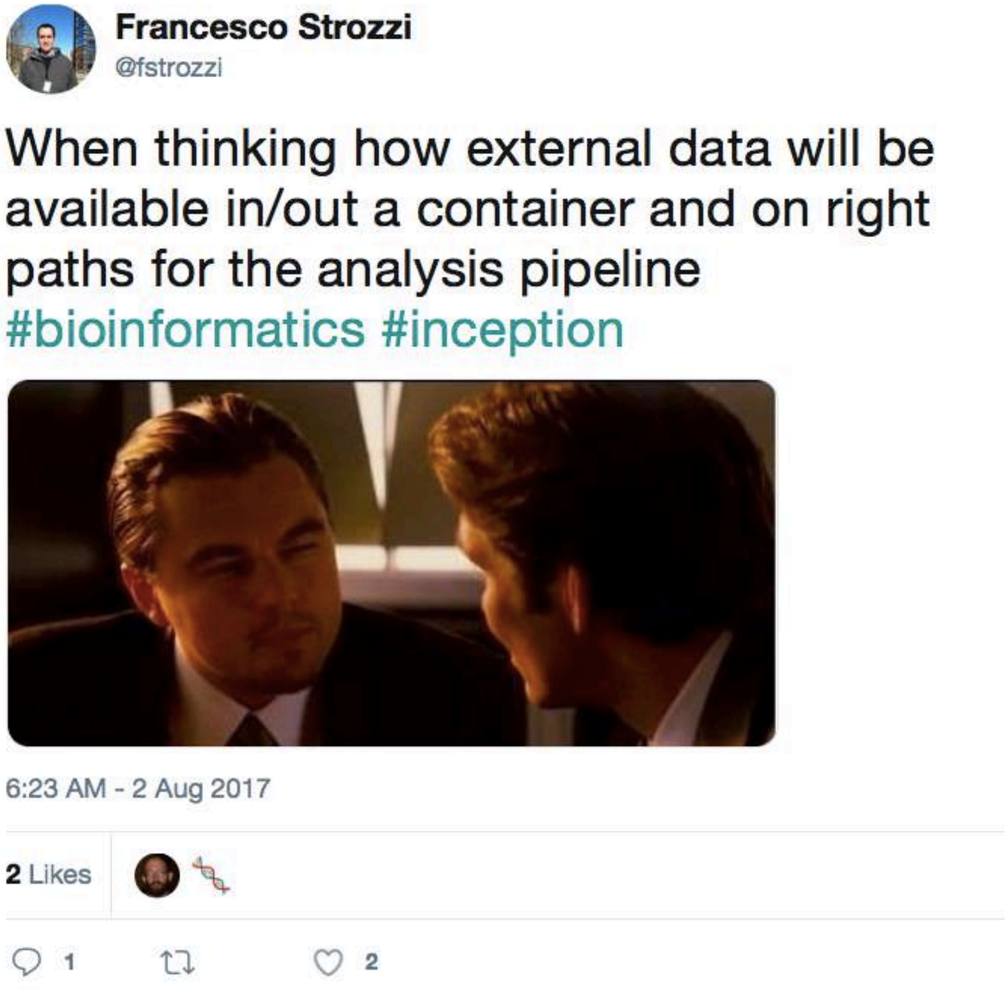
Francesco Strozzi
@fstrozzi

When thinking how external data will be available in/out a container and on right paths for the analysis pipeline #bioinformatics #inception

6:23 AM - 2 Aug 2017

2 Likes

1 2



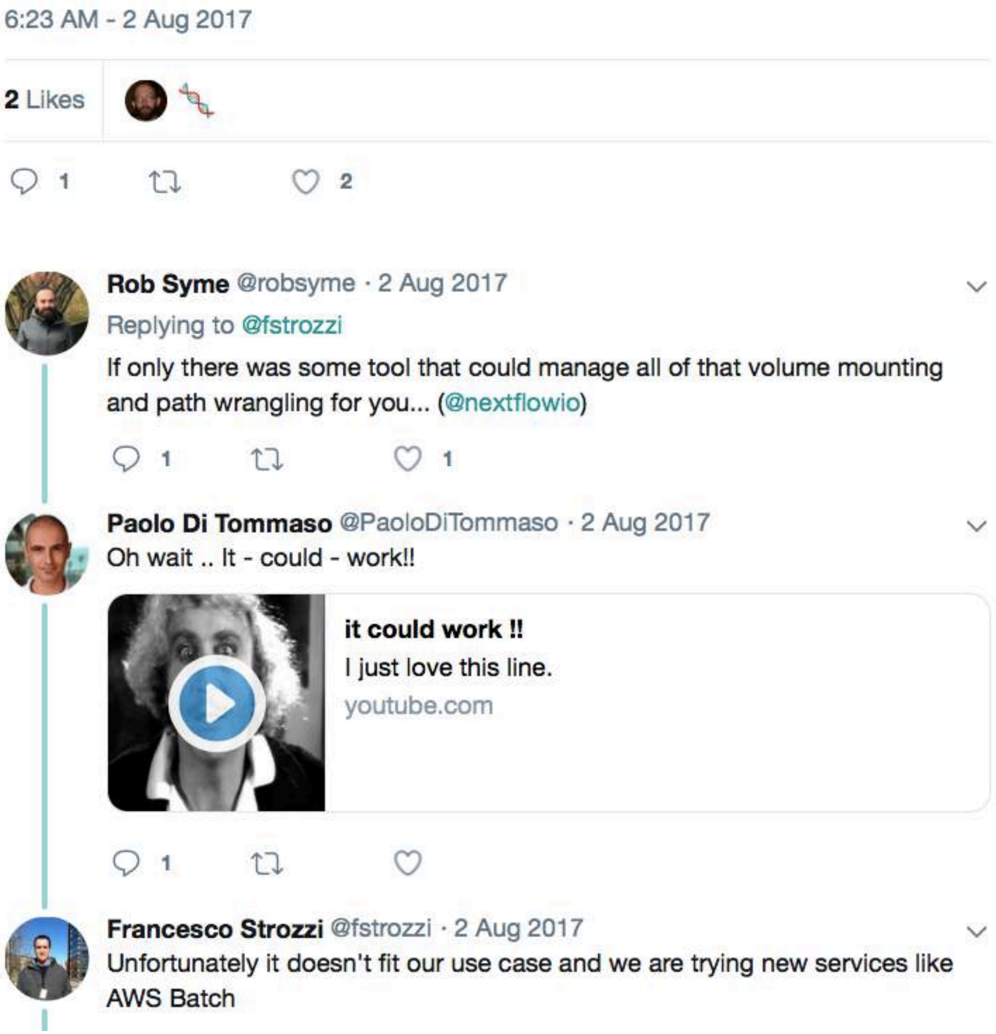
Francesco Strozzi
@fstrozzi

When thinking how external data will be available in/out a container and on right paths for the analysis pipeline #bioinformatics #inception

6:23 AM - 2 Aug 2017

2 Likes

💬 1 ⟲ ♡ 2



6:23 AM - 2 Aug 2017

2 Likes

💬 1 ⟲ ♡ 2

Rob Syme @robsyme · 2 Aug 2017
Replying to @fstrozzi
If only there was some tool that could manage all of that volume mounting and path wrangling for you... (@nextflowio)

💬 1 ⟲ ♡ 1

Paolo Di Tommaso @PaoloDiTommaso · 2 Aug 2017
Oh wait .. It - could - work!!

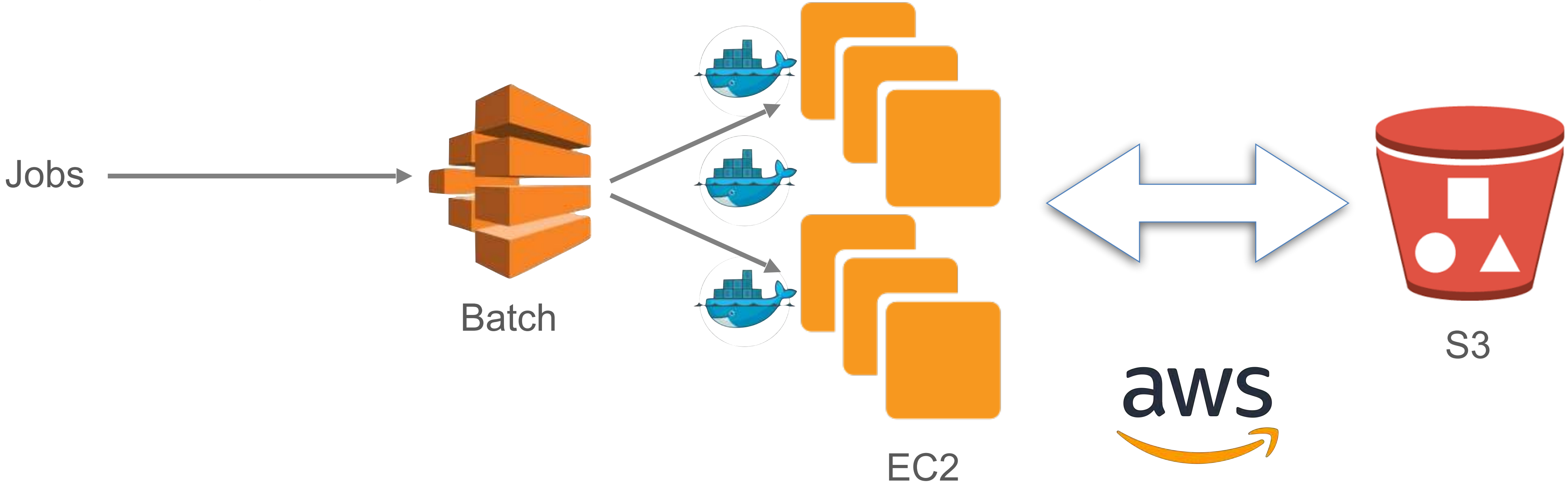
it could work !!
I just love this line.
youtube.com

💬 1 ⟲ ♡

Francesco Strozzi @fstrozzi · 2 Aug 2017
Unfortunately it doesn't fit our use case and we are trying new services like AWS Batch

AWS BATCH

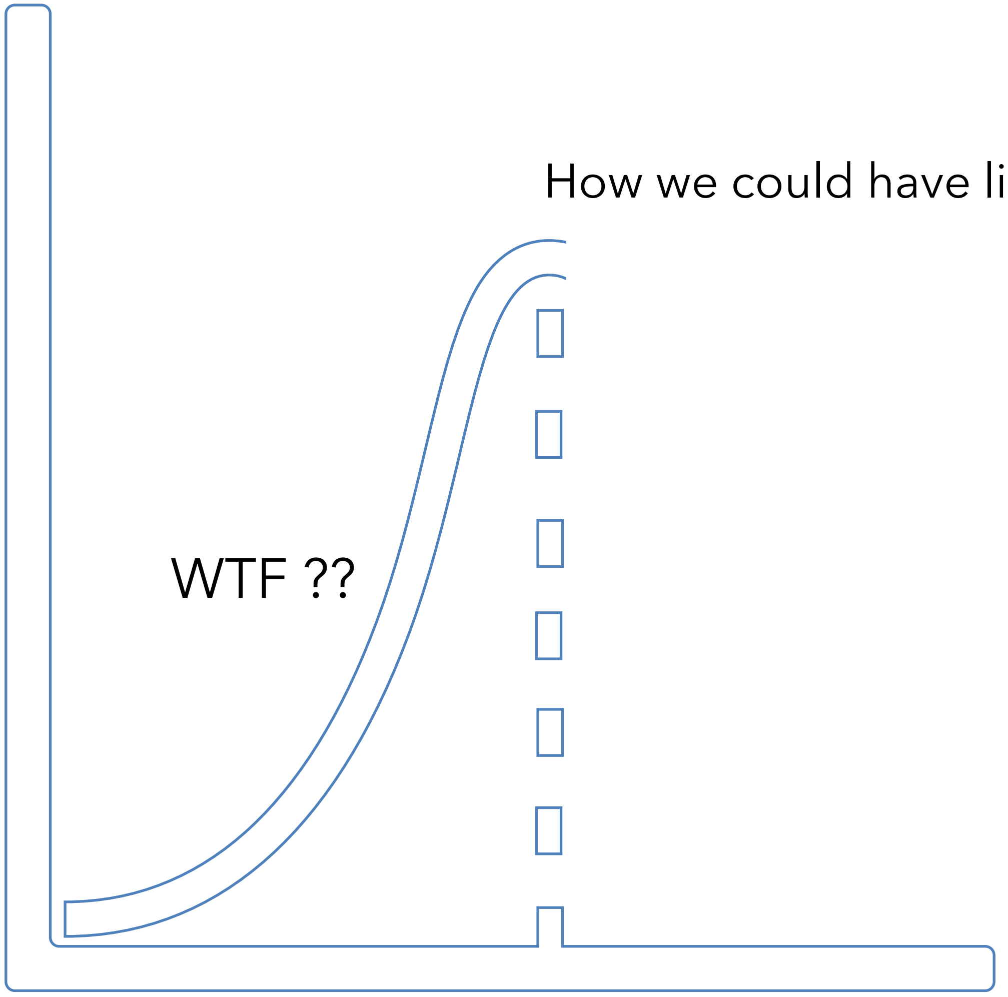
- Automates dynamic computing clusters creation
- Based on Docker
- Automatically optimise the EC2 instance types used depending on jobs requirements
- Using the spot instances and with the new per second billing, it has a dramatic impact on the costs for data analysis



AWS BATCH

- Available in Europe since June 2017
- First NF support introduced in September 2017, gone into stable release in November 2017
- Game changer for deployment and to run workflows at scale
- You get a full scale HPC infrastructure without having to think about the infrastructure
- Definitely one of the key component of our “agile” approach to pipelines
- Similar learning curve as Nextflow

Learning curve



How we could have lived without it before ?

WTF ??

From zero to agile pipelines

EDGE OF TOMORROW
LIVE. DIE. REPEAT.

From zero to agile pipelines

Agile techniques :

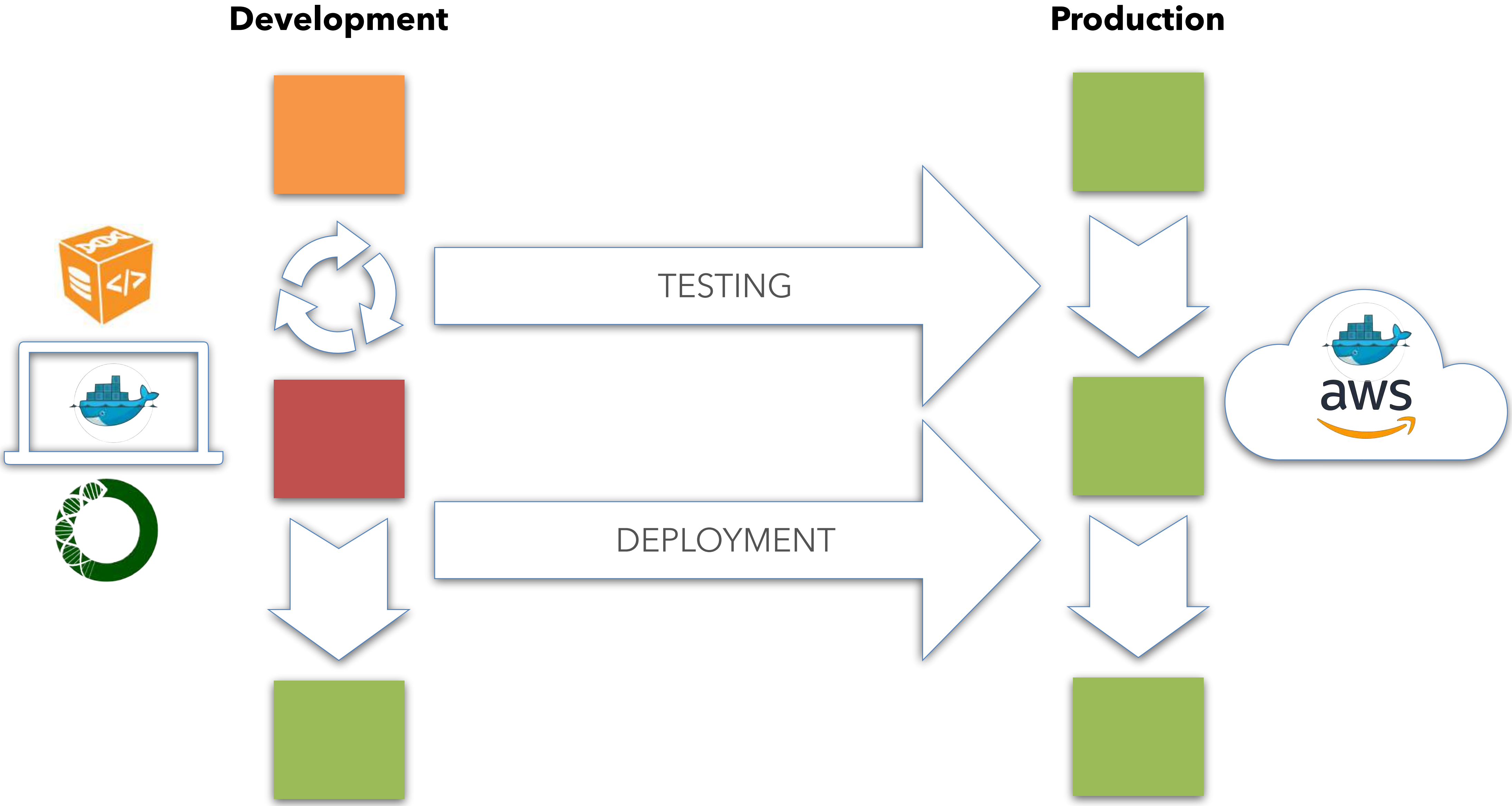
- ➡ development in small controllable units
- ➡ testing until no errors are detected
- ➡ when happy, move on

Nextflow allows to do exactly that:

- ➡ processes are small controllable units
- ➡ we can control what gets in and what gets out
- ➡ we can check failures, fix, change and then restart the workflow
- ➡ when happy we can move on and add a new process

We have found ourselves in the same exact habit as for software development

From zero to agile pipelines



From zero to agile pipelines

Some lessons learned along the way:

- ➡ we started with monolithic containers and then we decoupled things progressively
- ➡ pipelines became more and more general, so that we could re-use components
- ➡ the number of unnecessary NF hacks dropped down with time

The good

- Nextflow process isolation

- Portability and deployment
  - ➡ Our development cycle starts locally on our laptops using Docker
  - ➡ And evolve naturally on AWS to scale out the analysis
  - ➡ This transition is one of the least painful activities we have experienced so far

- Caching
  - ➡ It works well also on AWS and S3

- Cloud computing support
  - ➡ Using S3 as a local file system
  - ➡ All the plumbing is automatically managed by AWS and NF

The weird

- Caching
  - ➡ Sometimes successfully terminated process are re-run and we do not know why
  - ➡ No changes in the process, no changes in the input
  - ➡ On very large analysis with thousands of jobs, running through cache can take time

- Managing multiple channels with different elements can be non-intuitive sometimes

- S3 can be slow, especially when you have a lot (thousands) of files to move around

- You need to get past some initial complicate moment when learning NF and especially when configuring AWS and Batch for the first time

The (almost) missing

- Enhanced workflows "composability"
  - ➡ We found ourselves doing a lot of copy and paste of processes from one working workflow to a new development one
  - ➡ Ideally, one should have a set of general workflows from which more specified workflows could inherit processes and logic
  - ➡ Of course this is highly dependent on the application domain
  - ➡ Flowcraft is definitely going into this direction

- Workflows unit tests
  - ➡ We check our workflows with test data to ensure results are what we expect
  - ➡ NF-core introduced pipeline testing with CI
  - ➡ More fine grained tests could be useful
  - ➡ A common unit testing framework for NF would be just amazing

Wrap up - Our experience so far

- We went from zero to full production usage with Nextflow in just a few months
- The combination of Nextflow and AWS Batch pushed us naturally into a more agile development of workflows
- We literally do not care any more about workflows execution, analysis scaling out and infrastructure
- We just focus on data analysis and the development of the best fit-for-the-purpose pipelines
- So in the end, way less pain

- Everyone in Enterome and particularly the bioinformatics team who followed me into this migration journey
- The Nextflow community and Paolo, who is always very receptive when discussing new implementations
- All Nextflow workshop organisers and sponsors

Thank you ! Any question ?

@fstrozzi

✉ fstrozzi@enterome.com

The aim of the wise is not to secure pleasure, but to avoid (unnecessary) pain

Aristotle