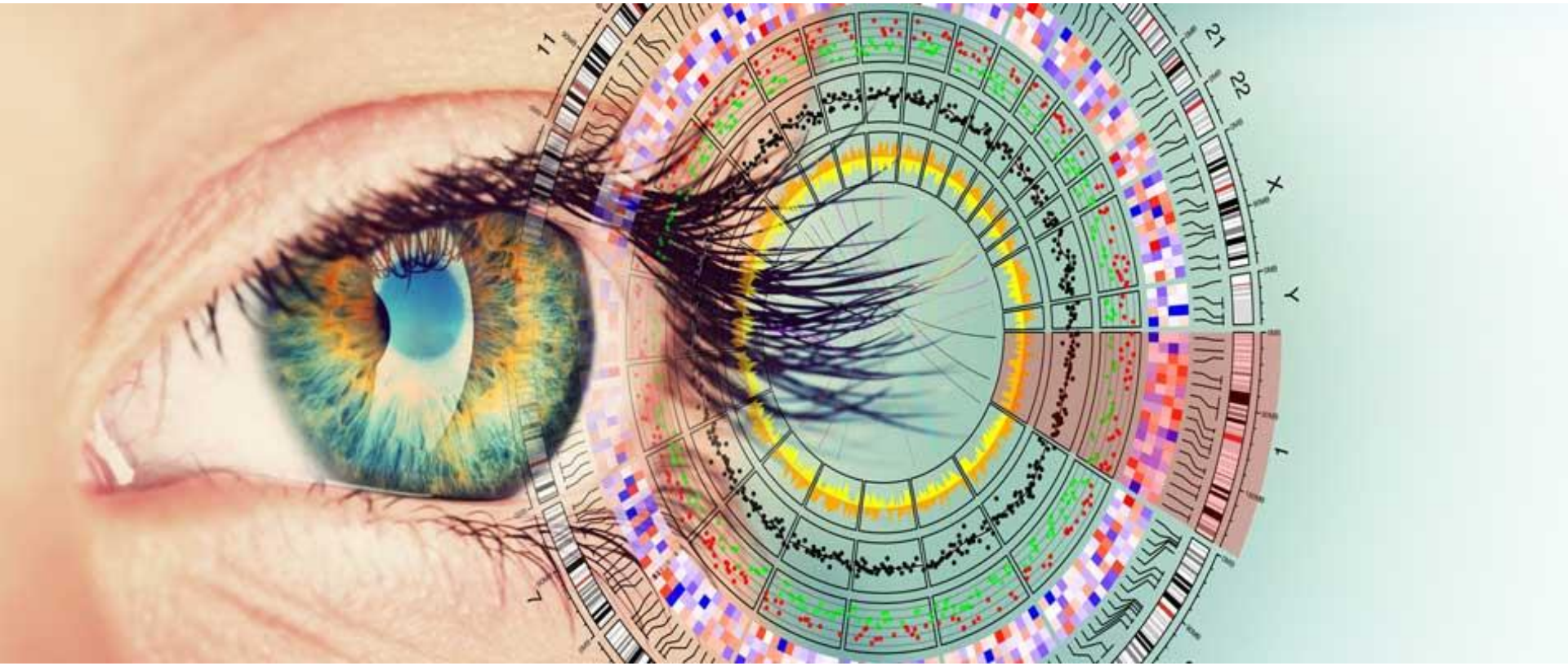
Large-scale Cancer Genomics Analysis in the Cloud with Butler

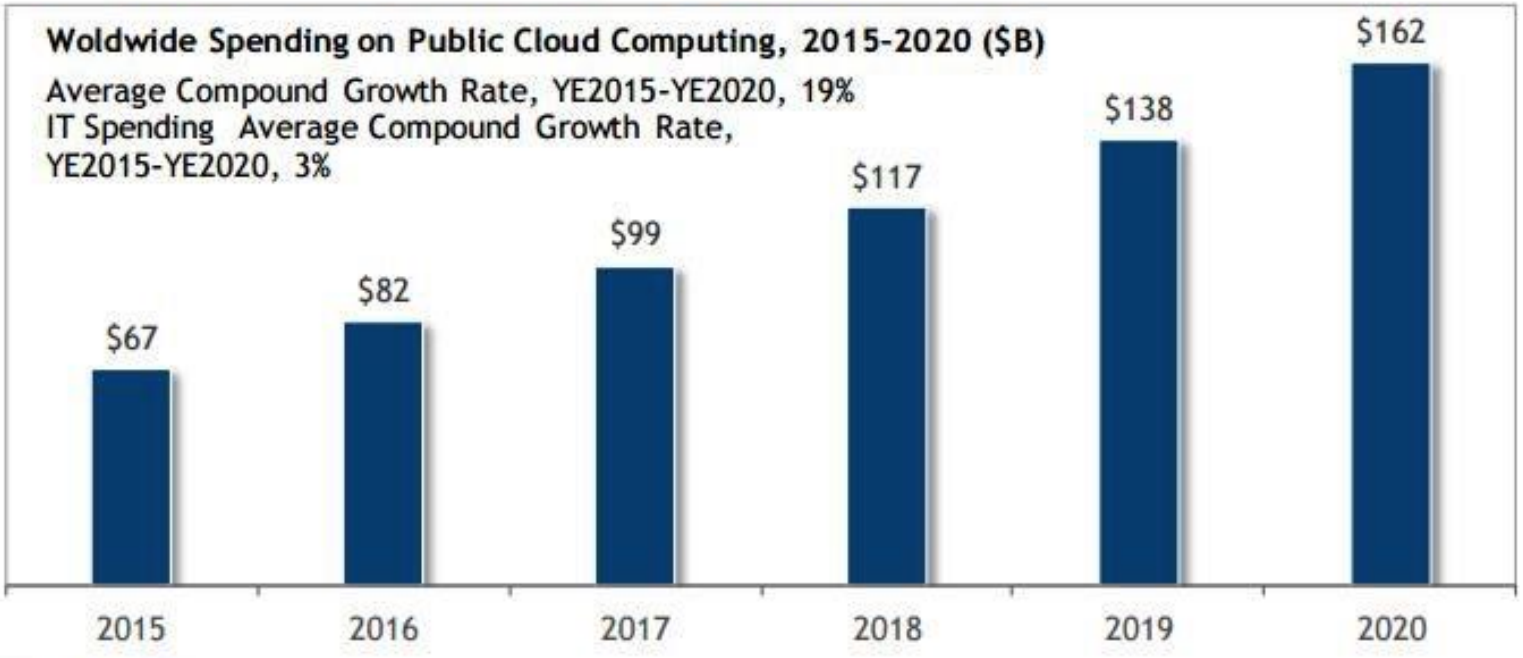


Sergei Yakneen
European Molecular Biology Laboratory (EMBL)
December 2017

EMBL

Growth of Cloud Computing

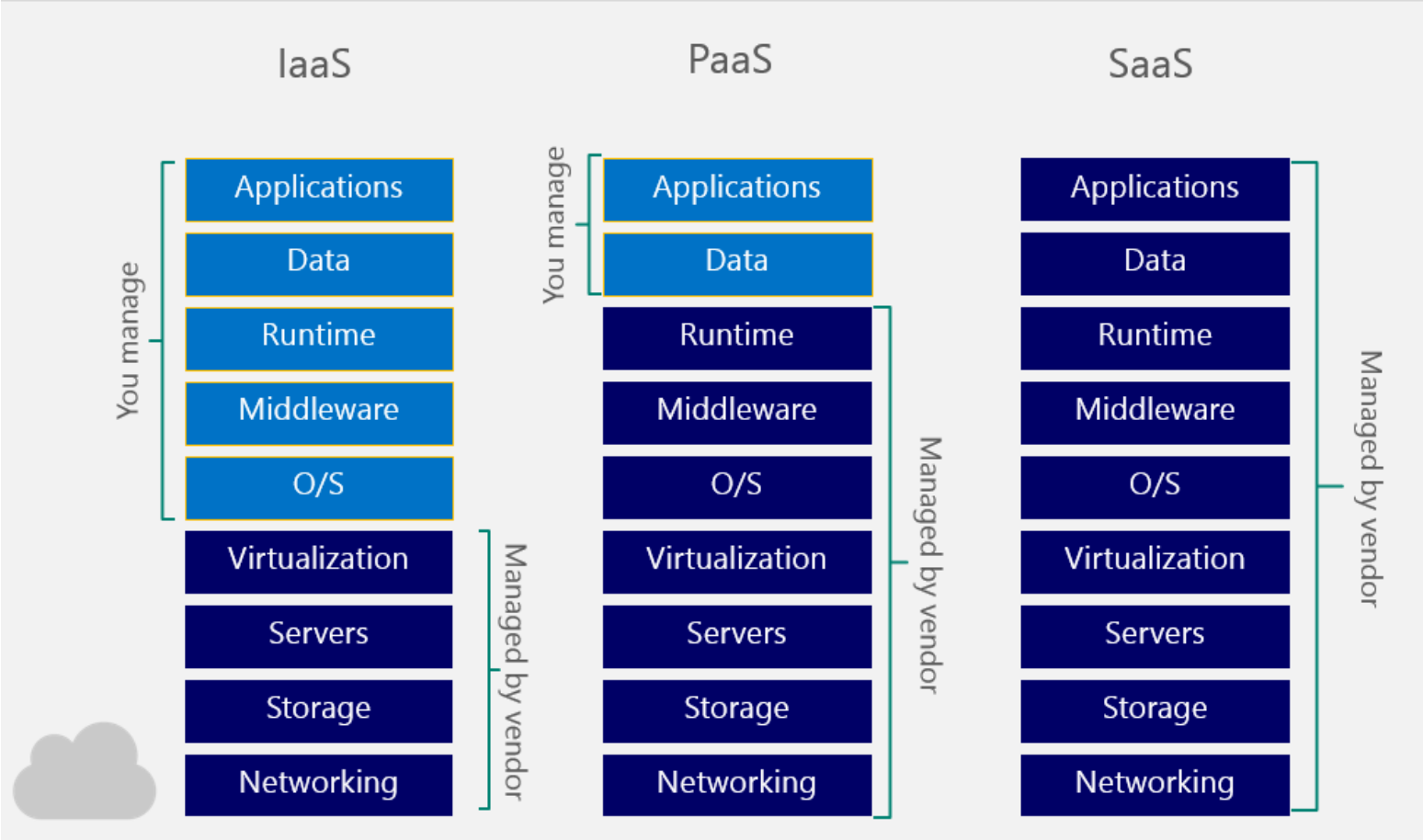
The Rapid Growth of Cloud Computing, 2015-2020



Source: IDC, 2016

EMBL

Cloud Computing Models



http://inspira.co.in/

Cloud Computing in bioinformatics

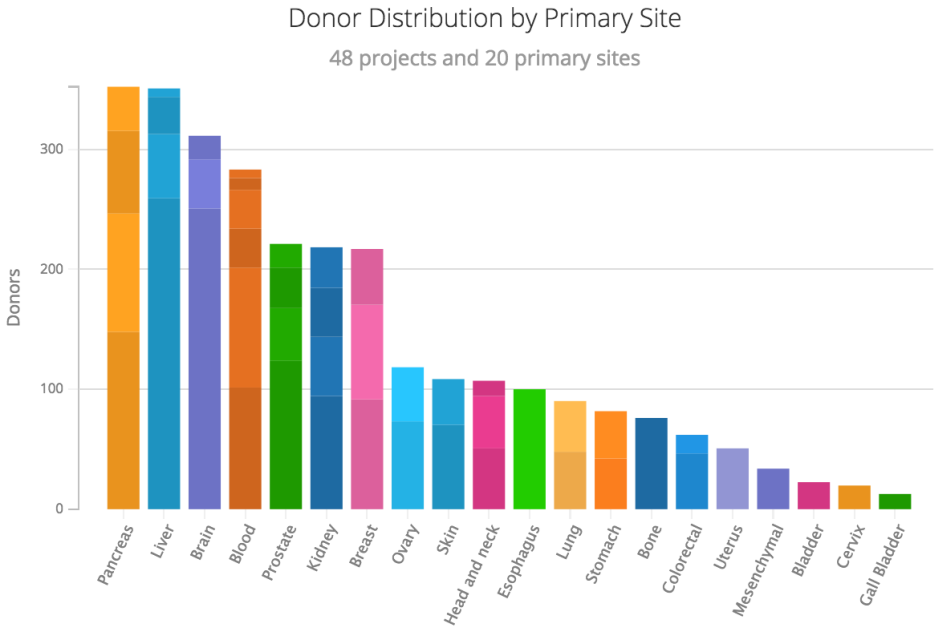
Computing environments

- AWS/GCP/Azure
- EMBL-EBI Embassy Cloud
- Open Science Data Cloud
- de.NBI
- European Open Science Cloud
- DNANexus, Seven Bridges Genomics, etc.

Best-practices and standards

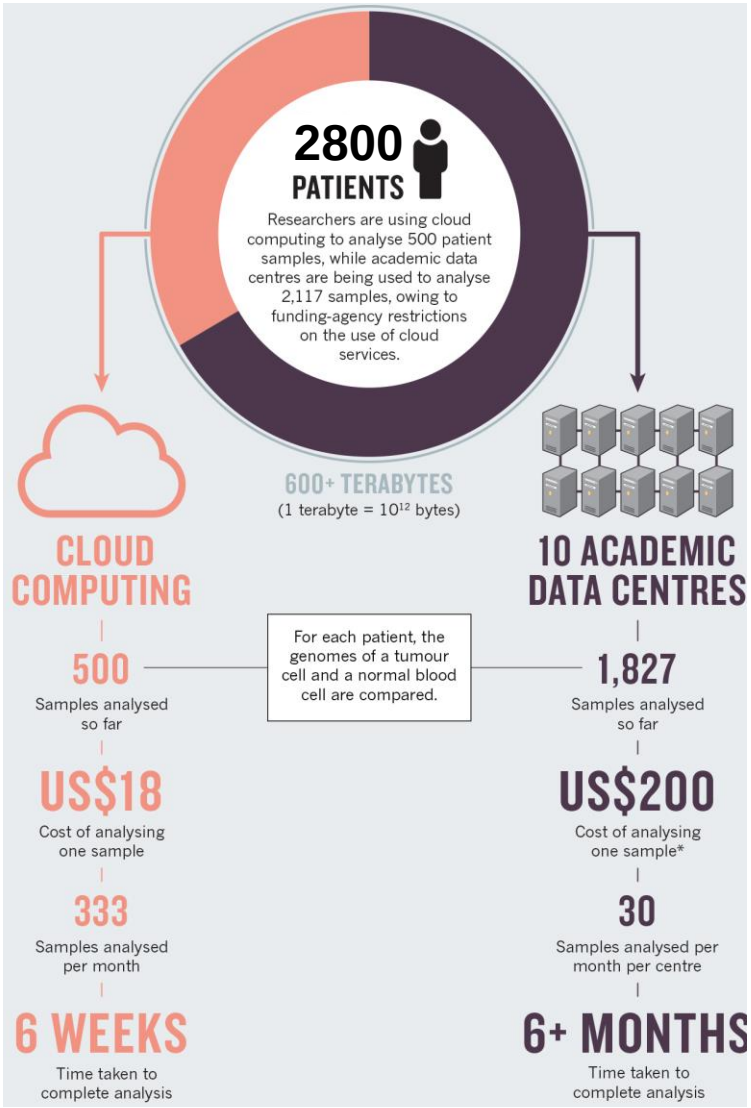
- Docker ecosystem
- Global Alliance for Genomics and Health

PCAWG



2,834 Donors | **70,313** Files | **729.09 TB**

Data Type	# Donors	# Files	Format	Size
SGV	2,834	8,865	VCF	502.33 GB
StGV	2,834	5,908	VCF	7.06 GB
Aligned Reads	2,834	8,721	BAM	728.42 TB
Simple Somatic Mutations	2,834	26,165	VCF	184.47 GB
Copy Number Somatic Mutations	2,834	5,911	VCF	131.74 MB
Structural Somatic Mutations	2,834	14,743	VCF	1.58 GB

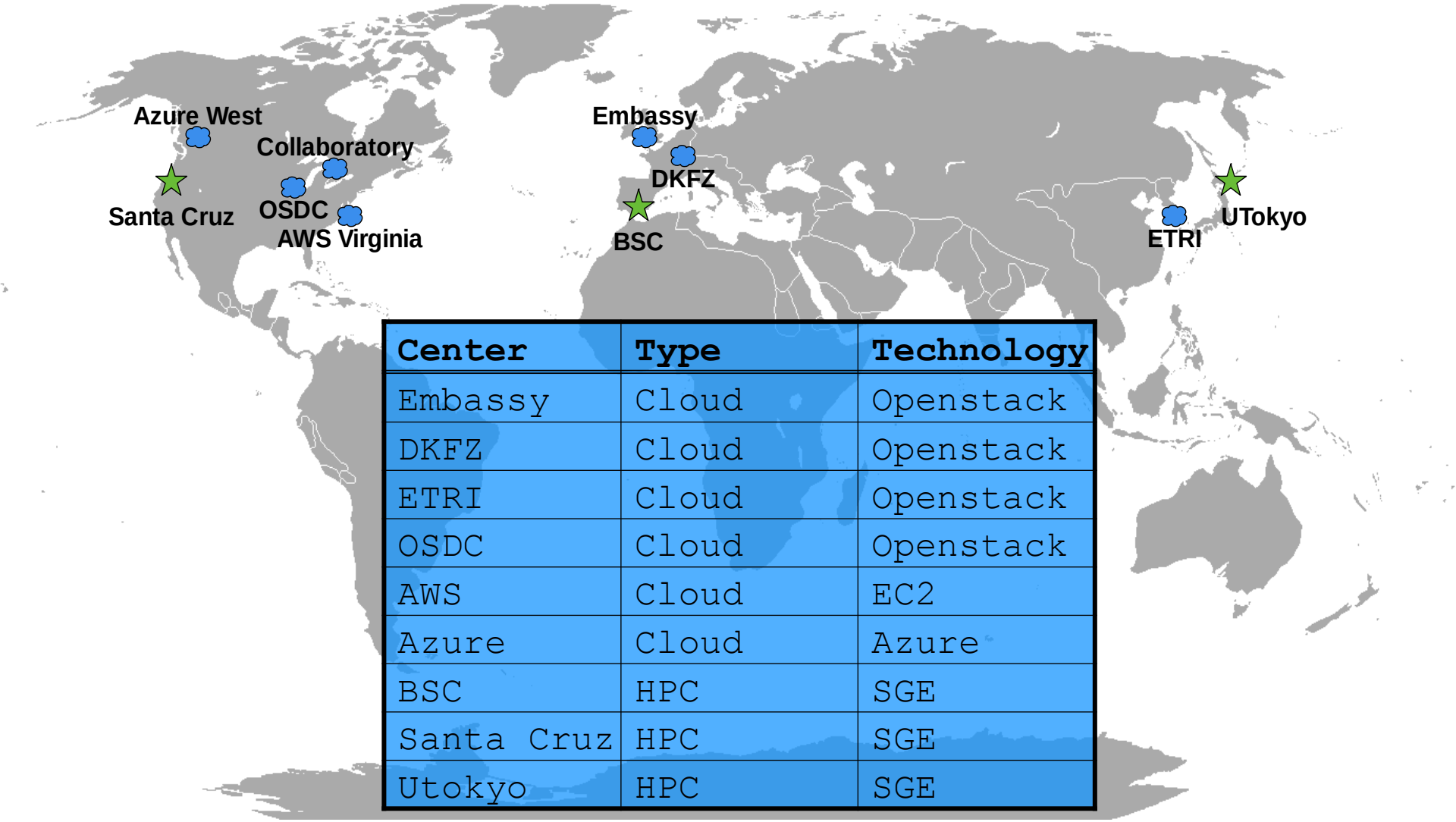


The International Cancer Genome Consortium (ICGC)

Stein, Knoppers, Campbell, Getz & Korbel, *Nature* 2015

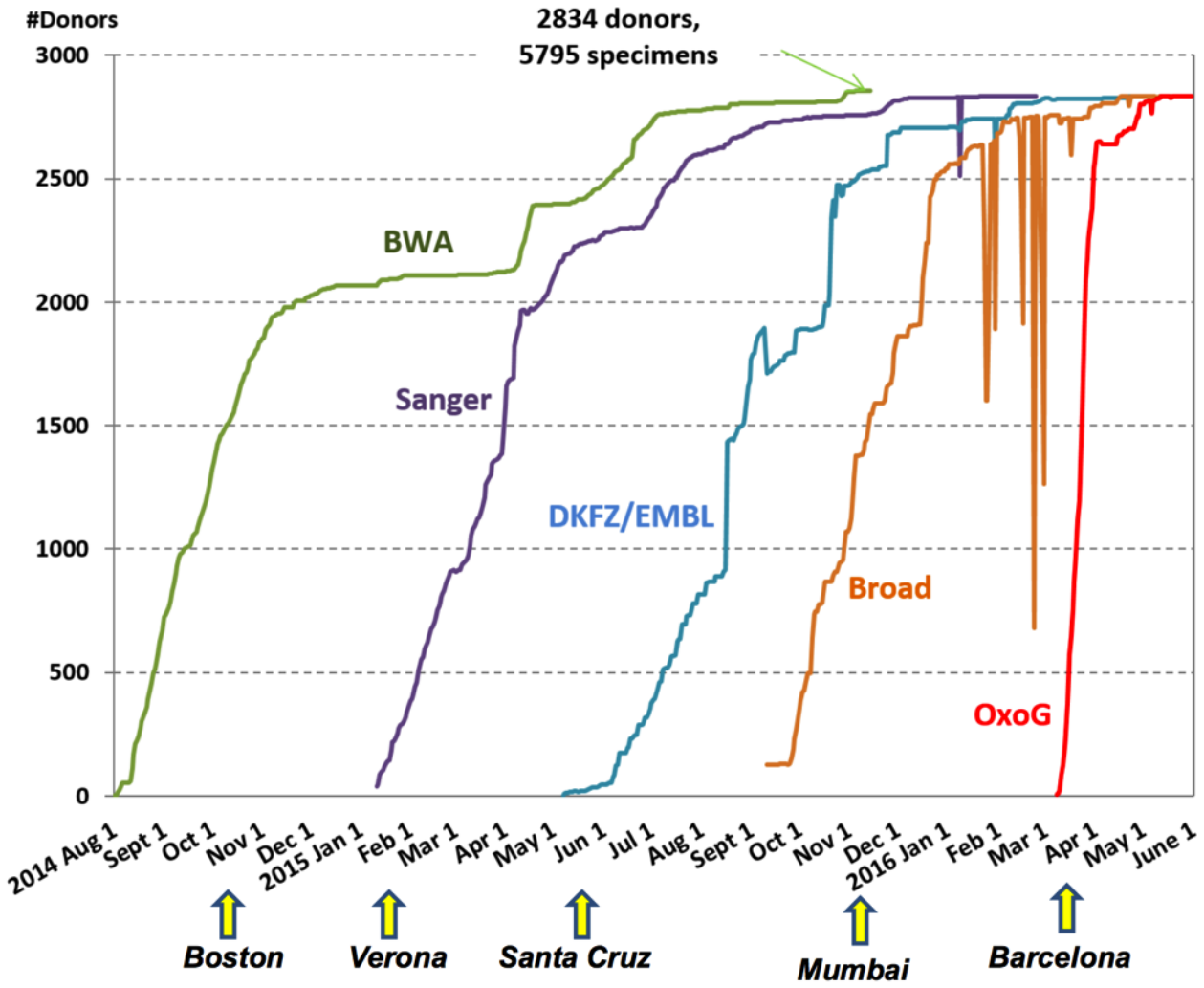
PCAWG Data Centers

VMs	Cores	RAM
638	16,000	60 TB



Center	Type	Technology
Embassy	Cloud	Openstack
DKFZ	Cloud	Openstack
ETRI	Cloud	Openstack
OSDC	Cloud	Openstack
AWS	Cloud	EC2
Azure	Cloud	Azure
BSC	HPC	SGE
Santa Cruz	HPC	SGE
Utokyo	HPC	SGE

PCAWG Data Processing



L. Stein

EMBL

Lessons Learned

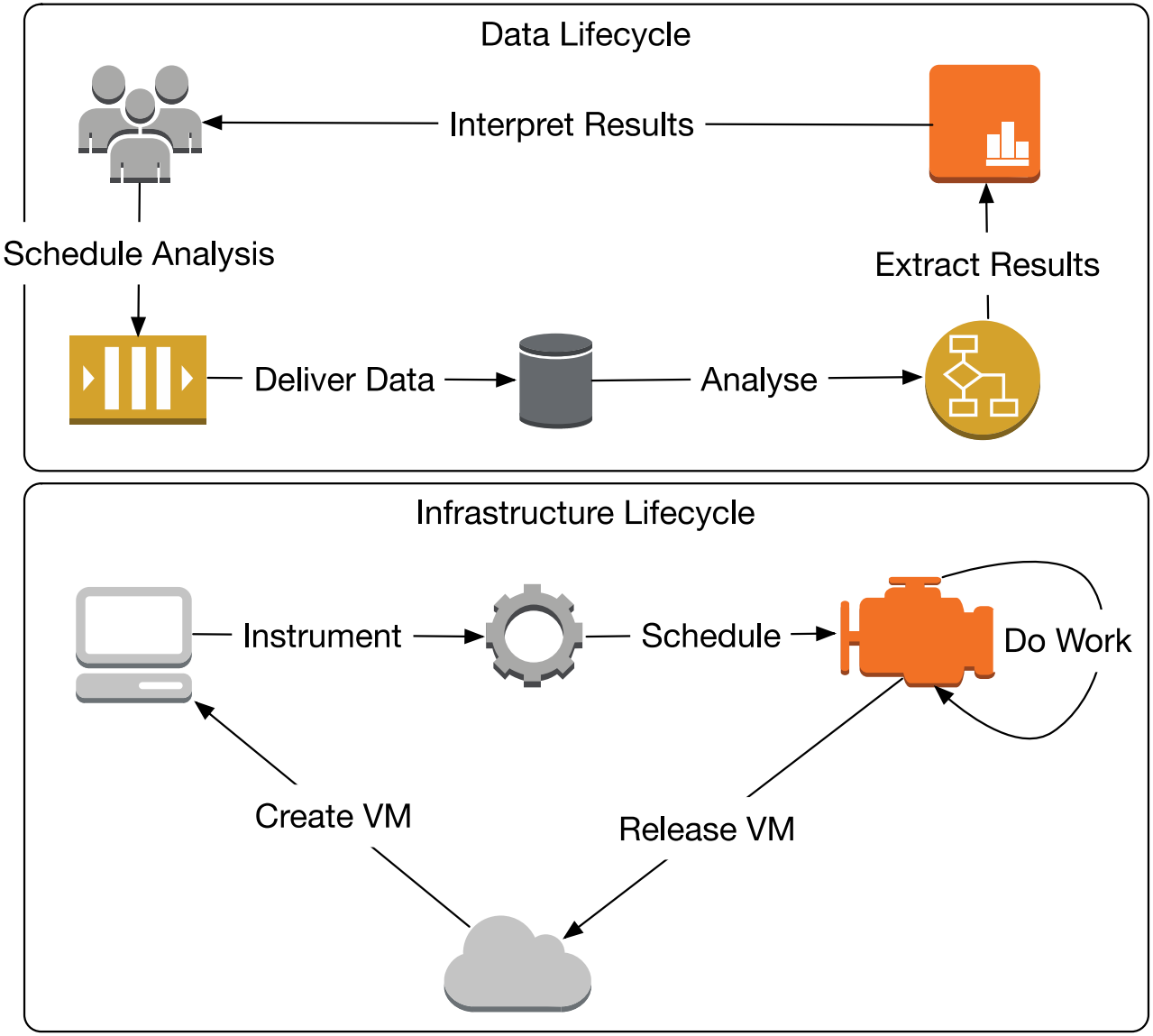
- Infrastructure fails

- Bioinformatics tools fail

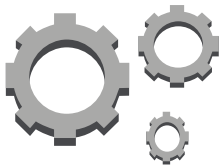
- End user must be self-reliant

- Manual investigation =

Analysis Lifecycle



Key Needs

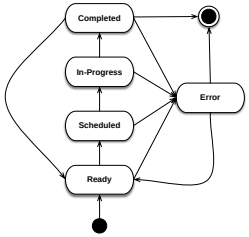


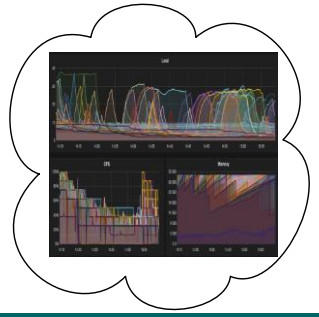
Provisioning

Configuration Management

Workflow

Operations Management





EMBL

Enabling rapid cloud-based analysis of thousands of human genomes via Butler

Sergei Yakneen, *Sebastian Waszak, Michael Gertz,* ***Jan O. Korbel***, *PCAWG Germline Cancer Genome Working Group, PCAWG Technical Working Group, ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Network*

https://doi.org/10.1101/185736

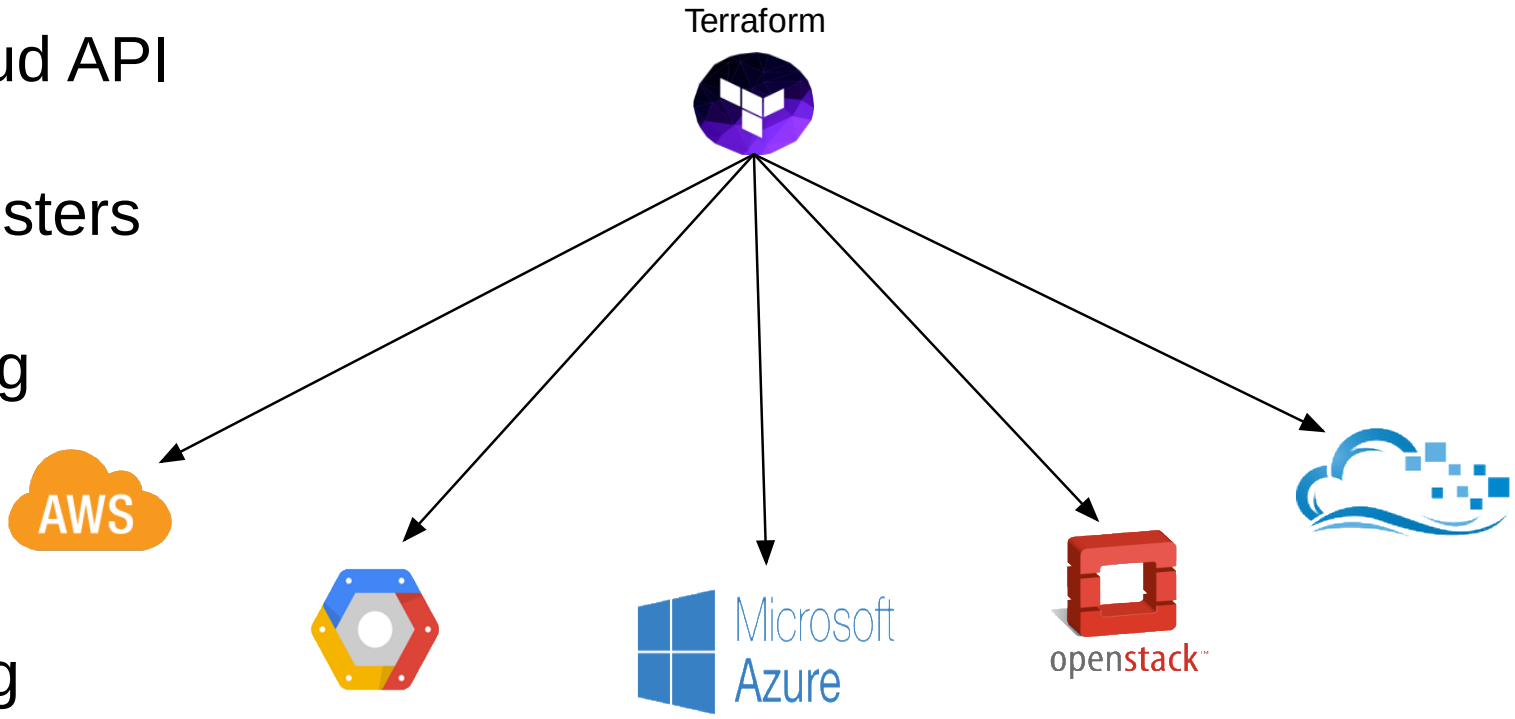
2017

Butler

https://github.com/llevar/butler

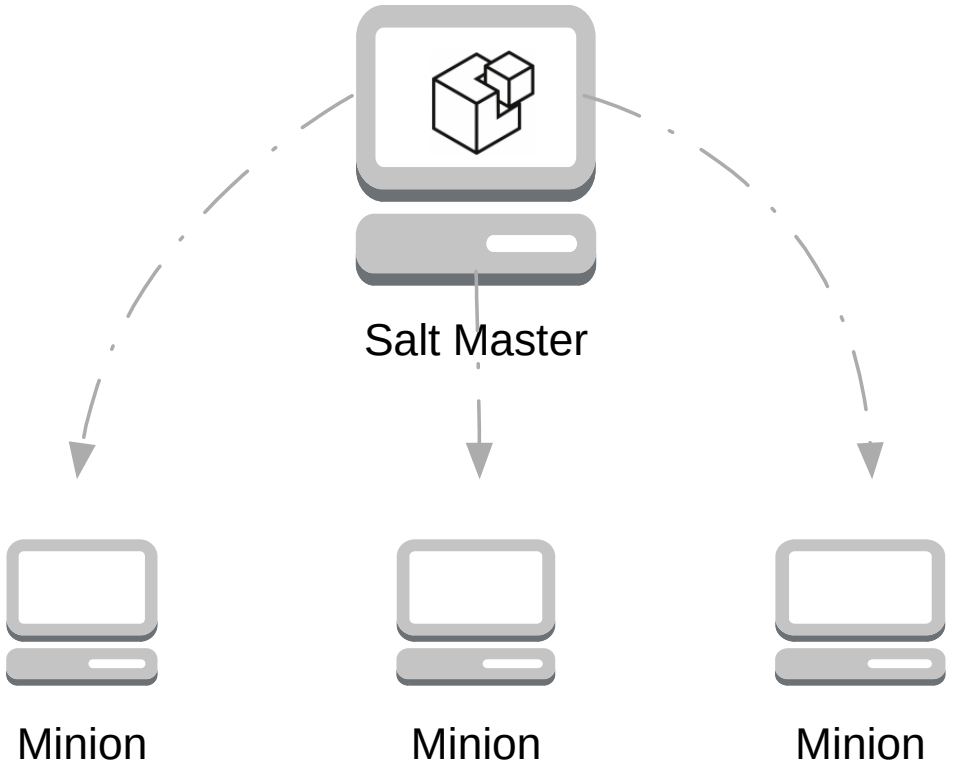
Provisioning with Terraform

- Cross-cloud API
- Create clusters
- Networking
- Security
- Templating
- Infrastructure-as-code



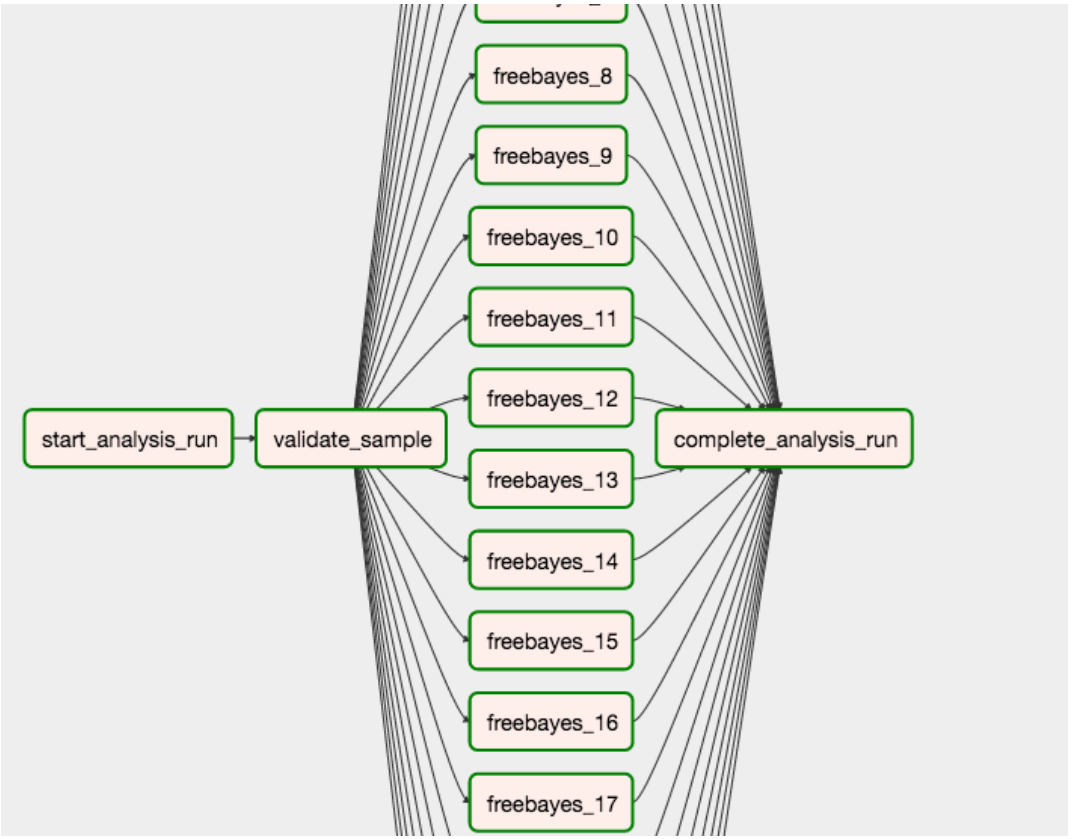
Configuration Management with Saltstack

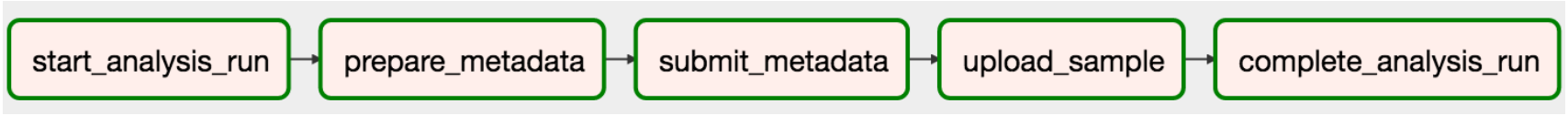
- Manage clusters up to 10,000 nodes
- Operating System agnostic
- YAML-based
- Large and active OSS project
  - ~8500 stars
  - 1580 contributors
- Configuration as code



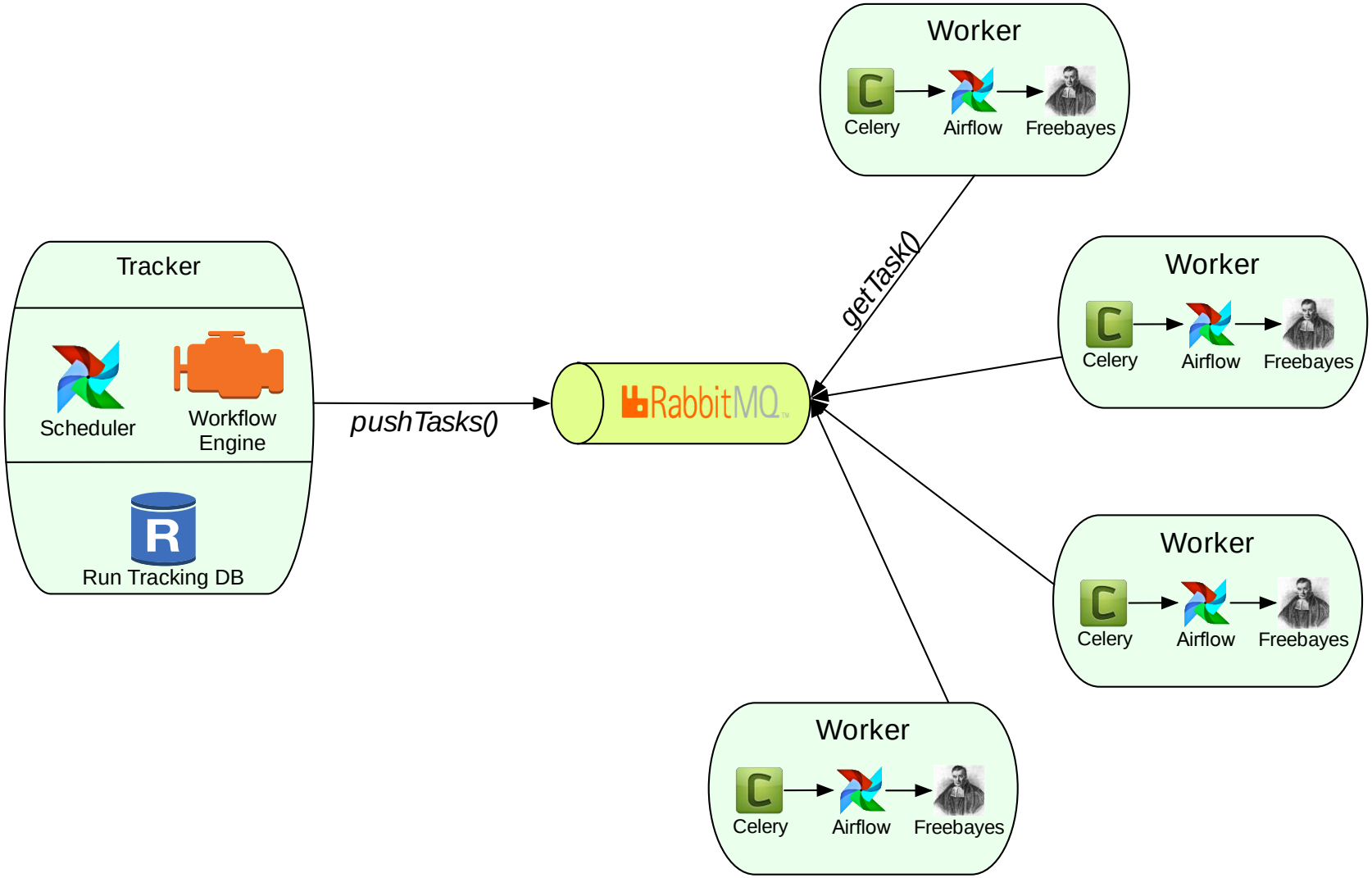
Workflows with Airflow

- Directed Acyclic Graph.
- Python program.
- Run on a random machine.
- Integrate Spark, Hive, HDFS
- Docker
- CWL



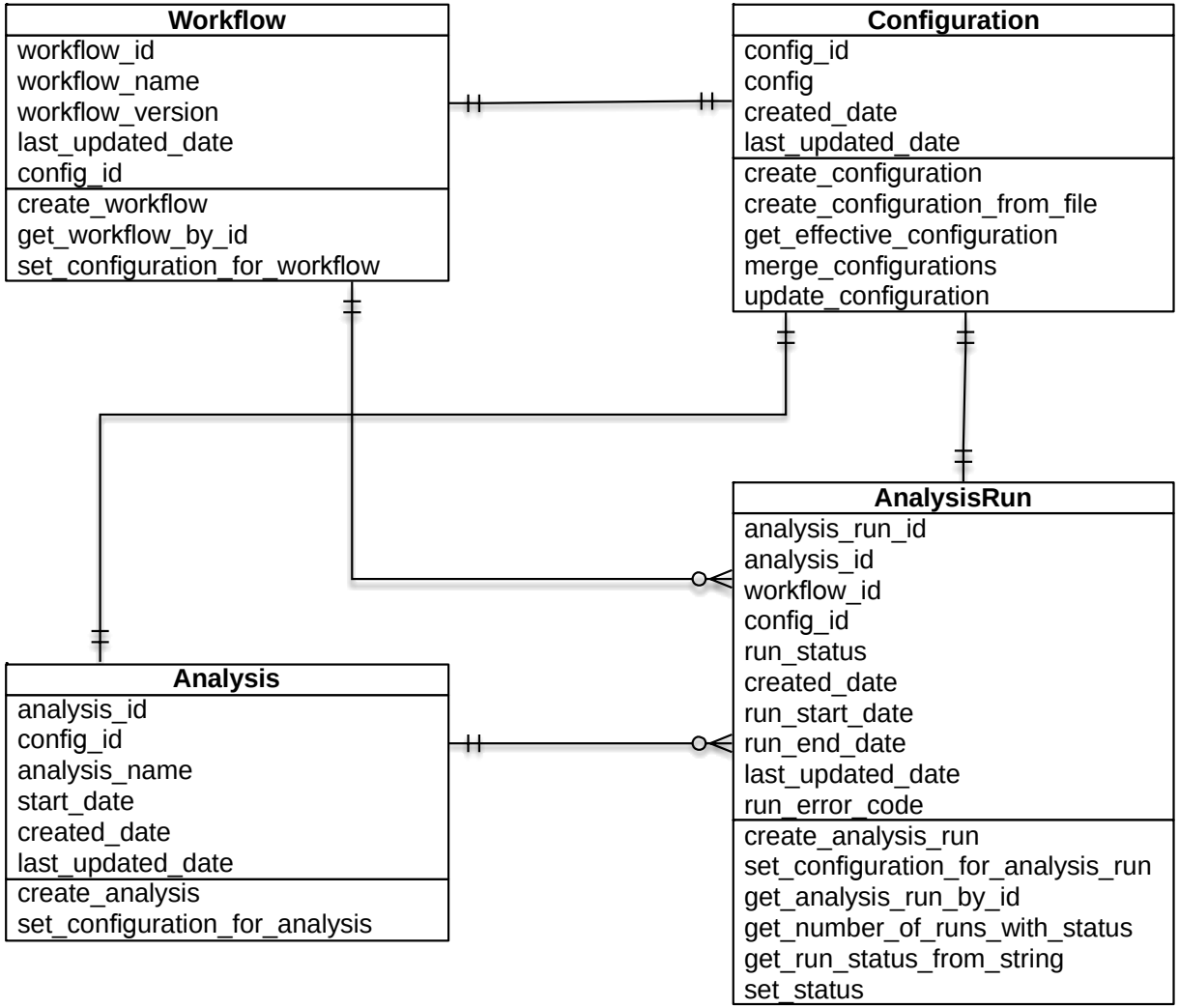


Workflow Execution Engine

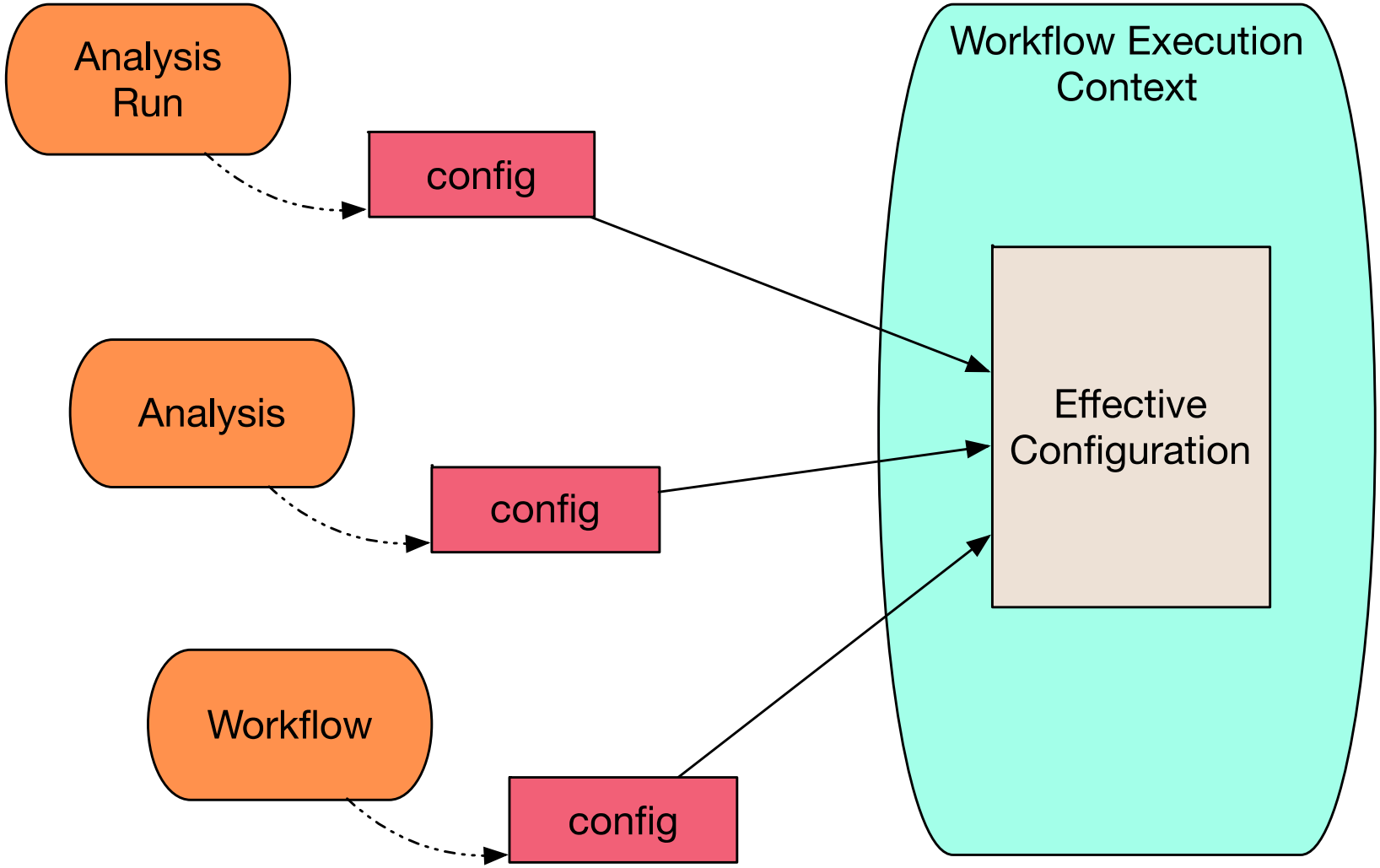


EMBL

Keeping Track of Analyses



Workflow Configuration



Workflow Runtime Management

DAGs

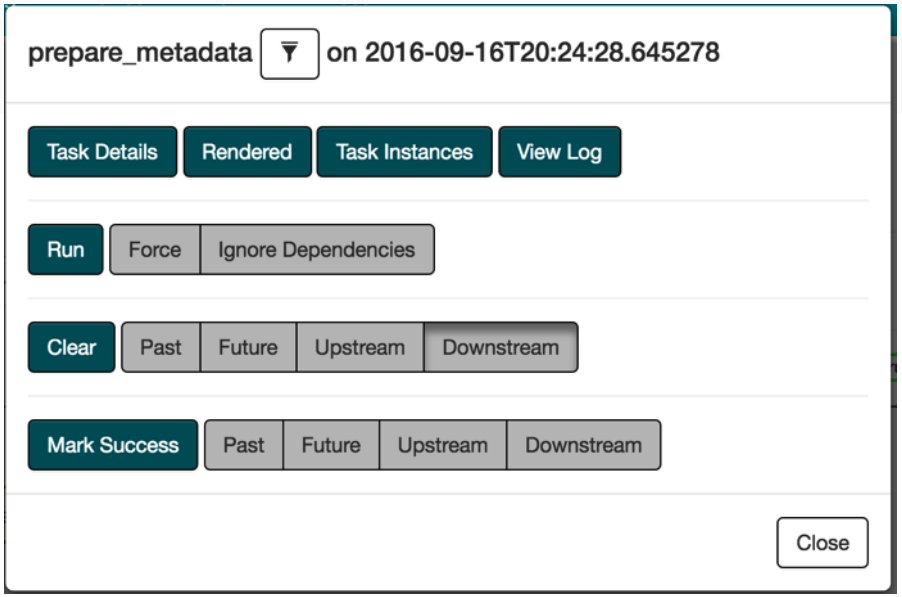
Show 10 entries

Search:

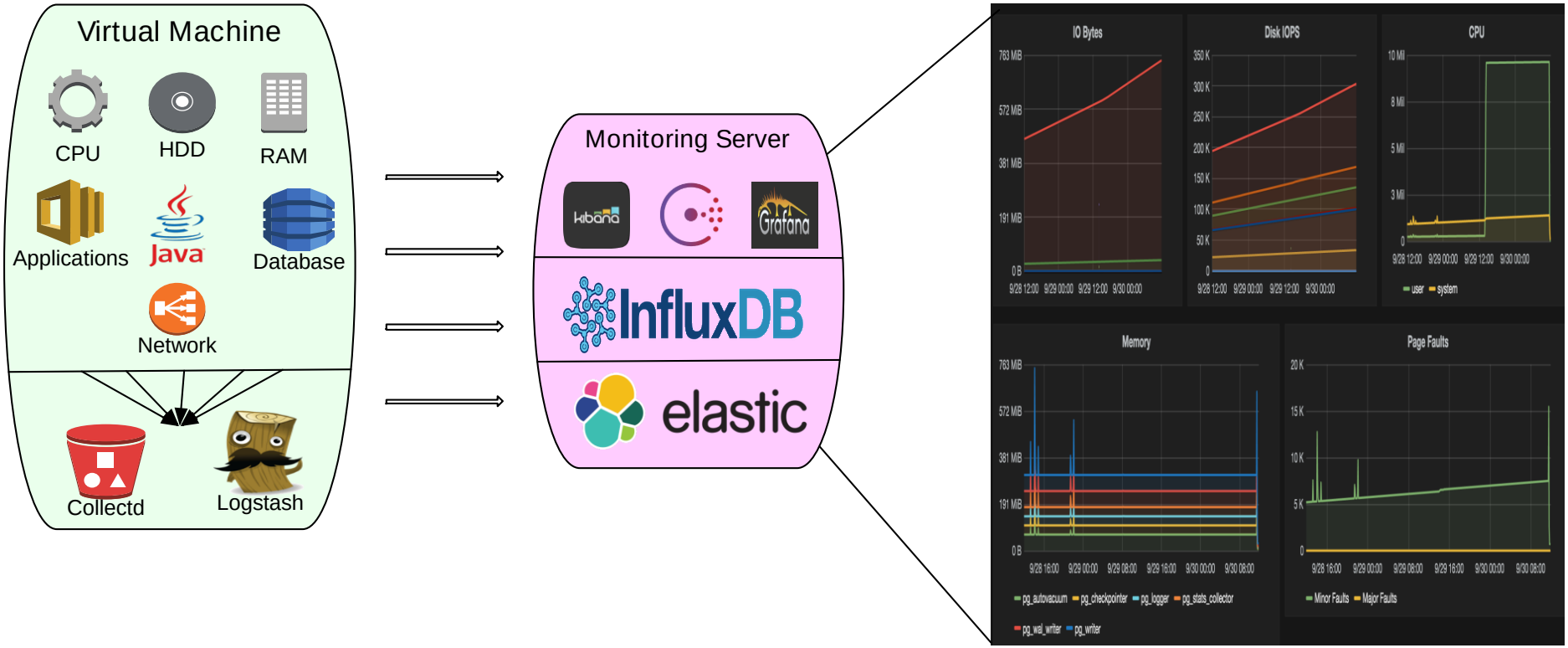
		DAG	Schedule	Owner	Statuses	Links
ℹ	On	delly	None	airflow	11374 865 1 1	
ℹ	On	freebayes	None	airflow	67749 1	

Showing 1 to 2 of 2 entries

Previous 1 Next

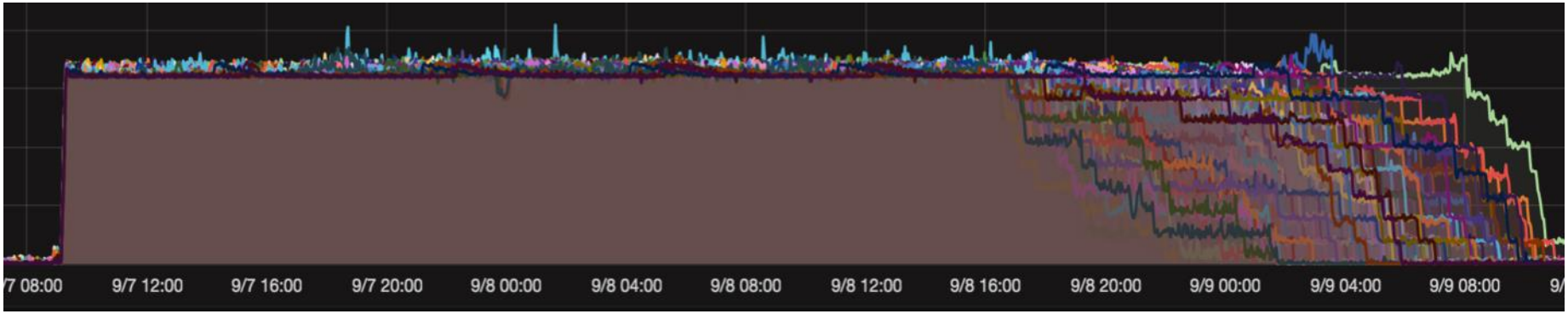


Operational Management

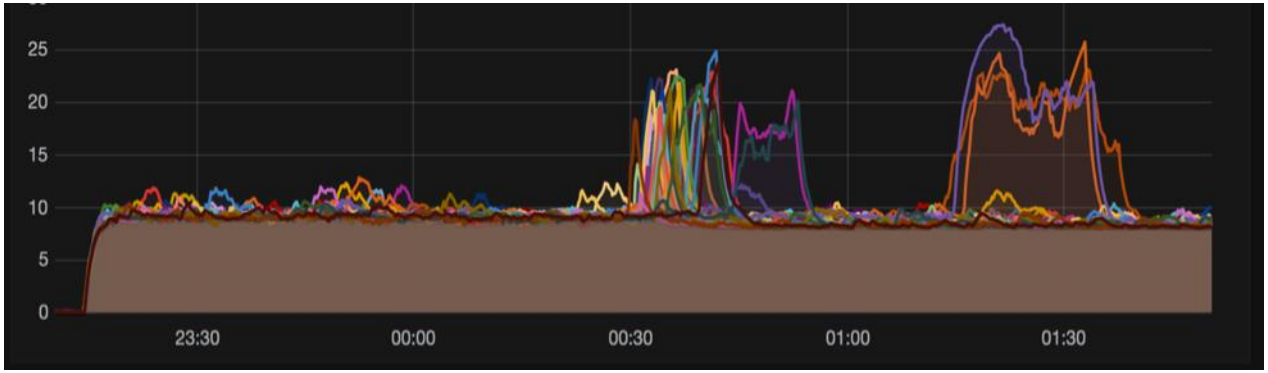


EMBL

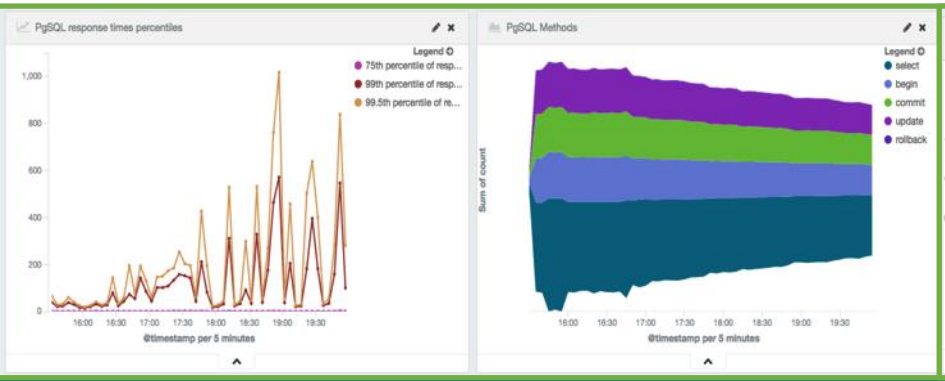
Normal Operation

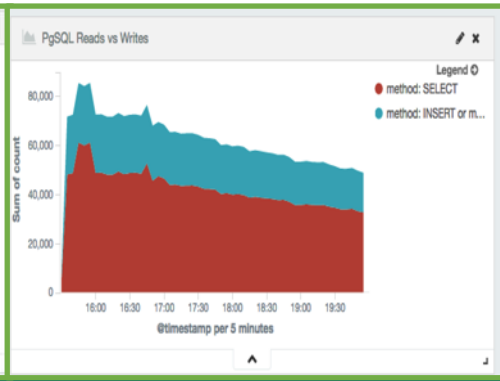


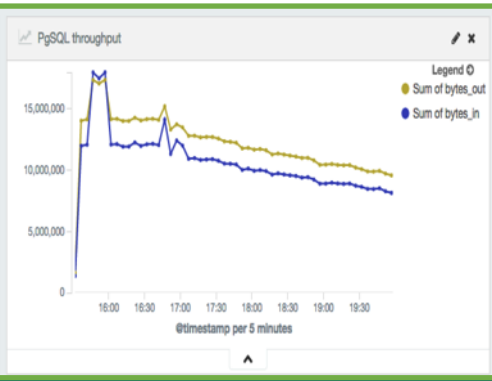
Infrastructure “event”



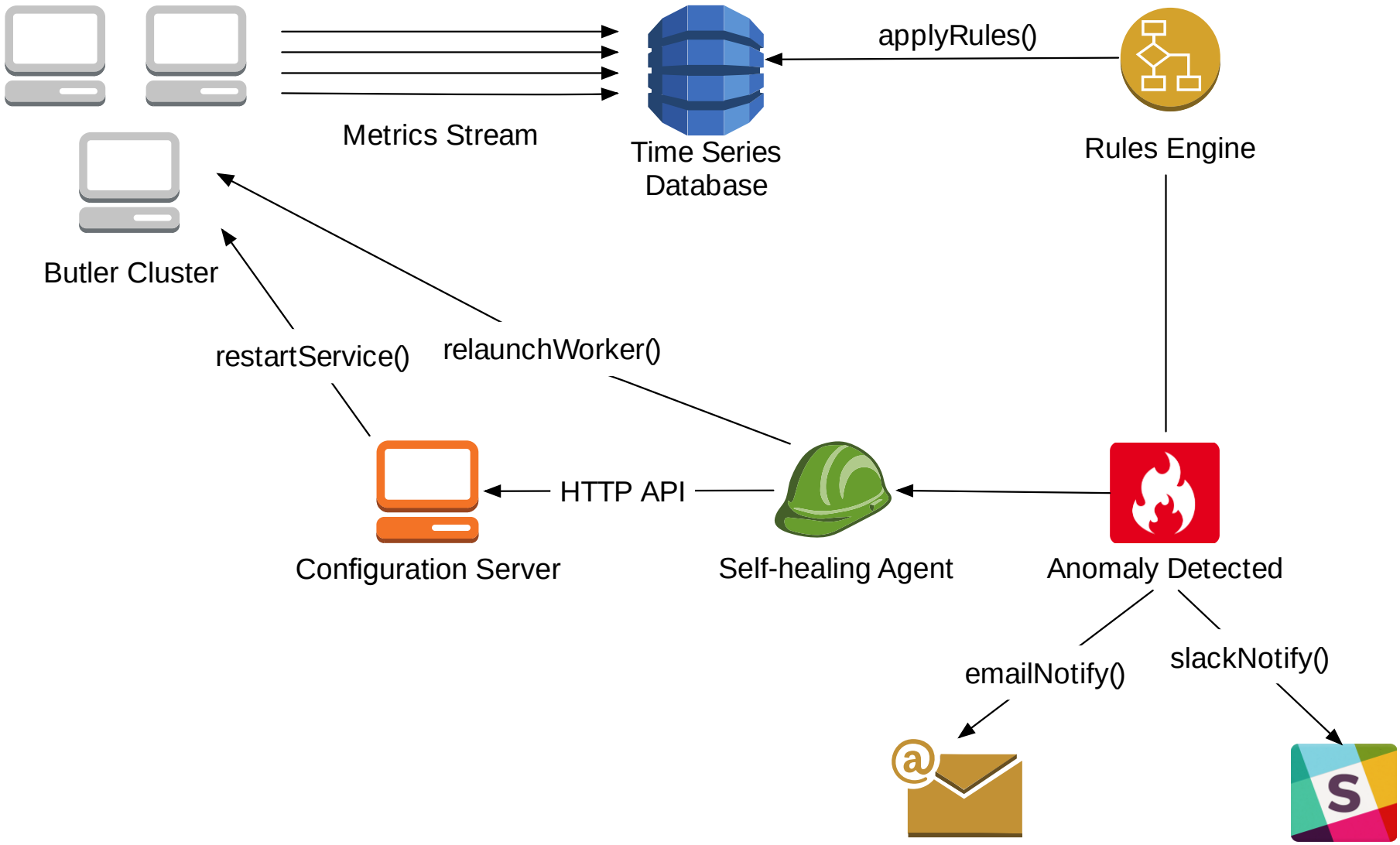
Database metrics harvested from logs

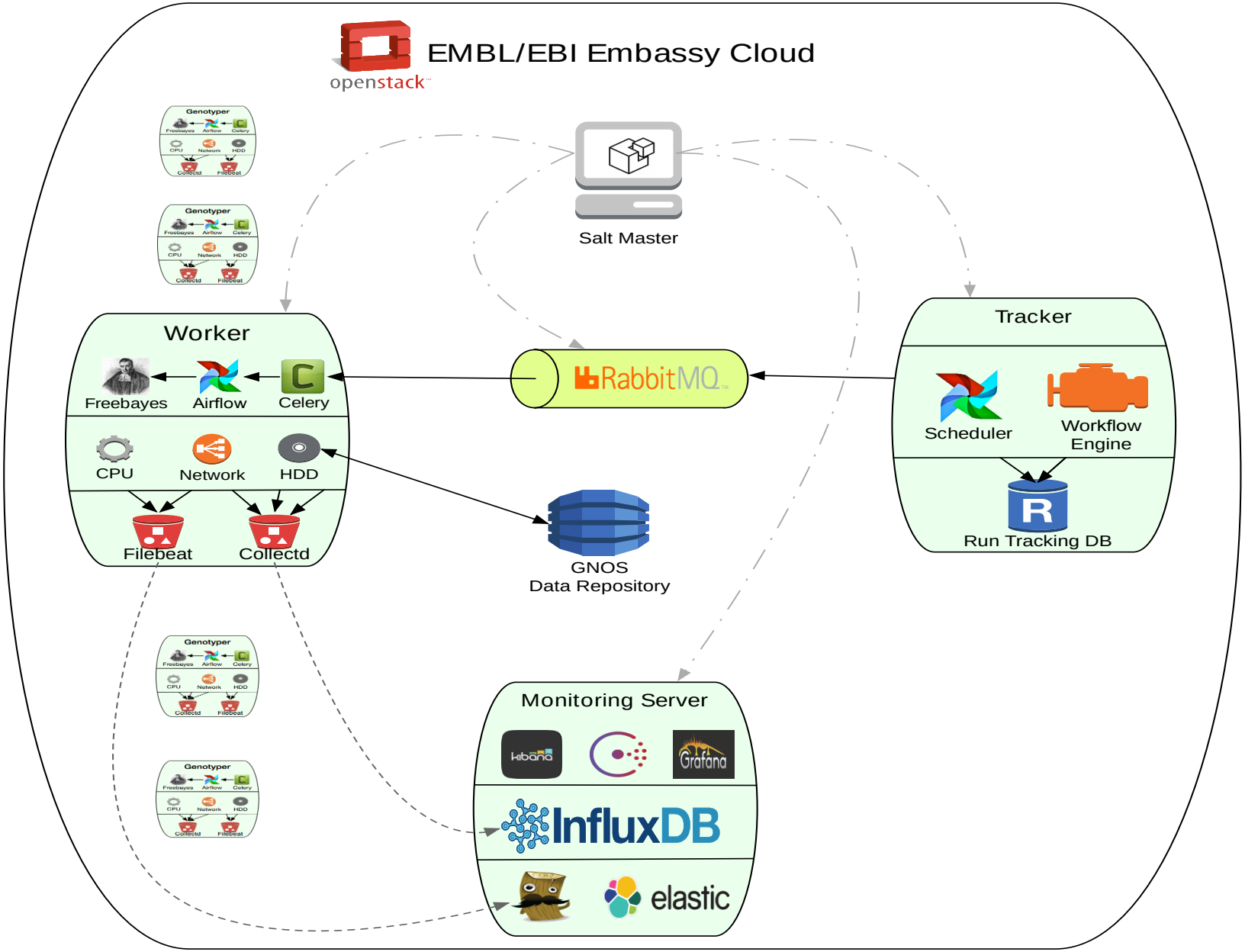






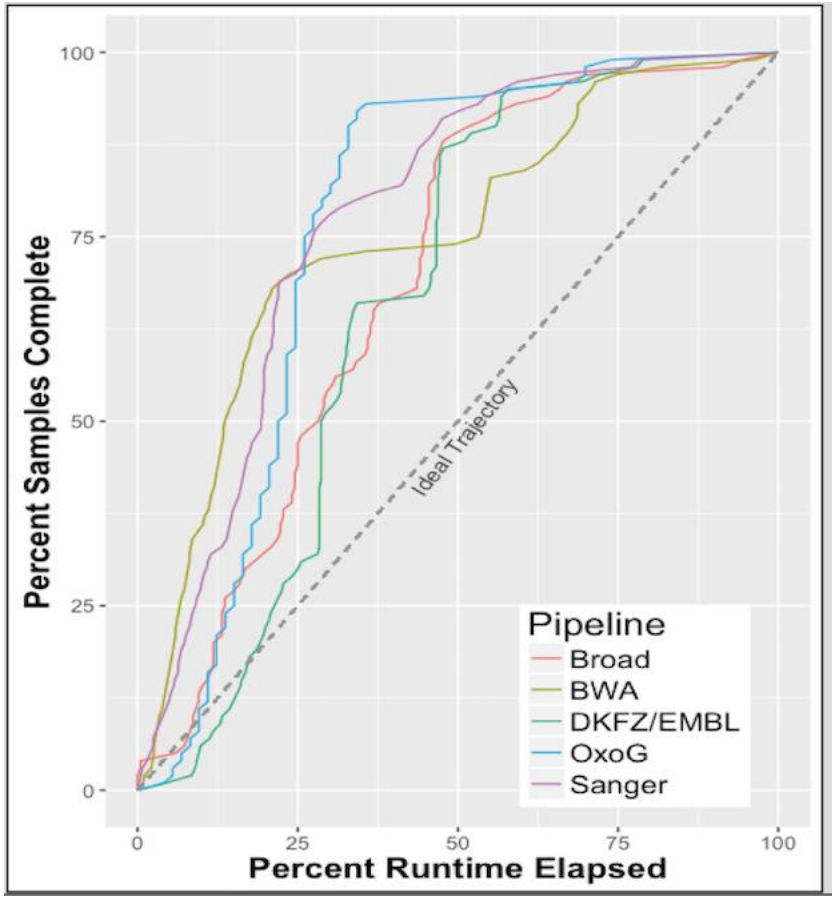
Self-Healing

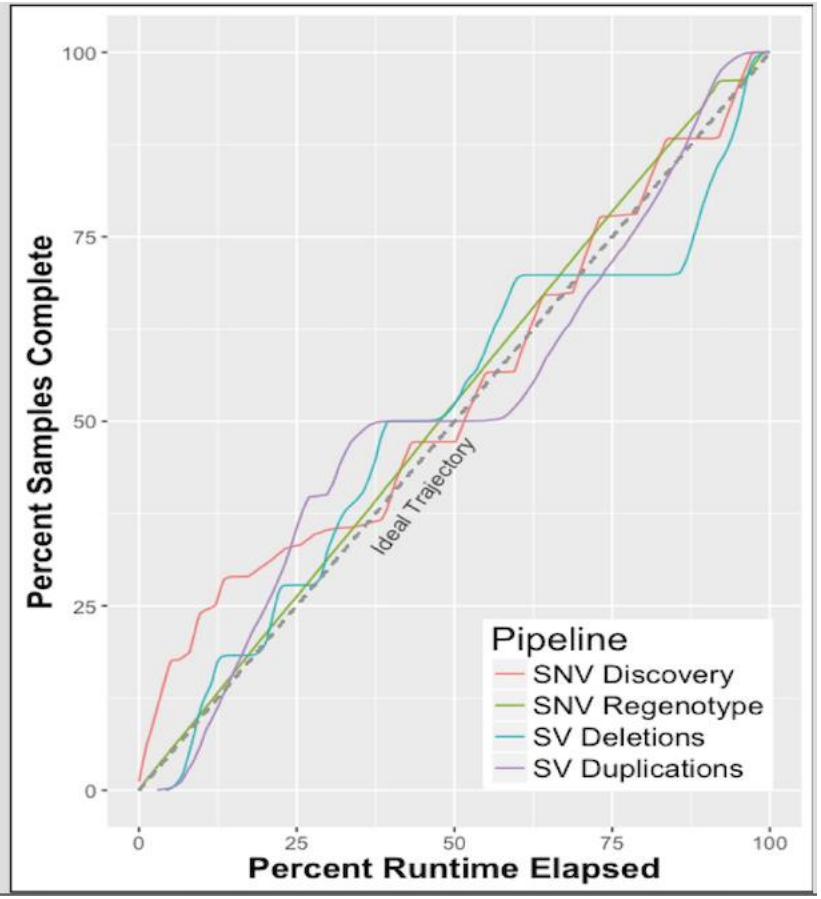



openstack
EMBL/EBI Embassy Cloud
Genotyper
Salt Master
Worker
Freebayes
Airflow
Celery
CPU
Network
HDD
Filebeat
Collectd
RabbitMQ
Tracker
Scheduler
Workflow Engine
R
Run Tracking DB
GNOS
Data Repository
Monitoring Server
kibana
Grafana
InfluxDB
elastic

Butler usage in PCAWG

- 1500 cores, 5.5 TB RAM, 750 TB disk
- Discovery and genotyping of **90M** variants across **2834** high-coverage WGS T/N **sample pairs** – 6x 750 TB.





	EMBL/EBI Embassy	Compute Canada	Cyfronet
vCPU	1000	1000	700
RAM	4 TB	4 TB	2.6 TB
Disk	1 PB	150 TB	200 TB
Data	448 samples from 224 prostate cancer donors	422 samples from 211 pediatric brain tumour donors	2081 samples from 1000 Genomes Project
	71 TB raw data	62 TB raw data	50 TB raw data
Status	**Alignment complete Variant calling in-progress**	**Alignment in-progress**	**Alignment in-progress**

Recap

- Clouds facilitate **sharing** of data and methods.
- Bioinformaticians need a wide array of **technical skills** to operate effectively on clouds.
- Butler provides functionality in four key areas – **infrastructure** provisioning, **configuration** management, **workflow**, **operations** management.
- Running real large-scale analyses shows that **operational management** tools have the **biggest impact** on how long it takes to finish the job.

Acknowledgements

EMBL

Jan Korbel
Sebastian Waszak
Tobias Rausch

EMBL/EBI

Andy Cafferkey
Rich Boyce
David Ocana
Charles Short
Steven Newhouse
Alvis Brazma
Dario Vianello
Erik van den Bergh

ComputeCanada

Greg Newby
Michael Cave

Cyfronet

Bartosz Kryza
Lukasz Dutka
Marek Magrys