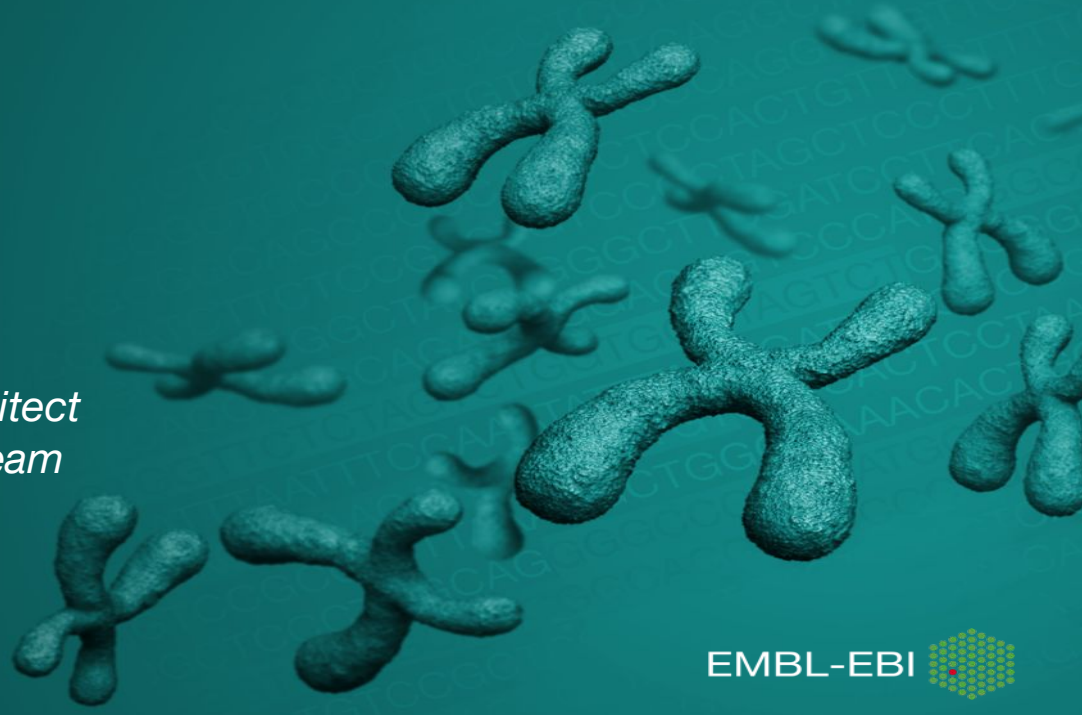
Hybrid Cloud

Integration Challenges of Big Data Science

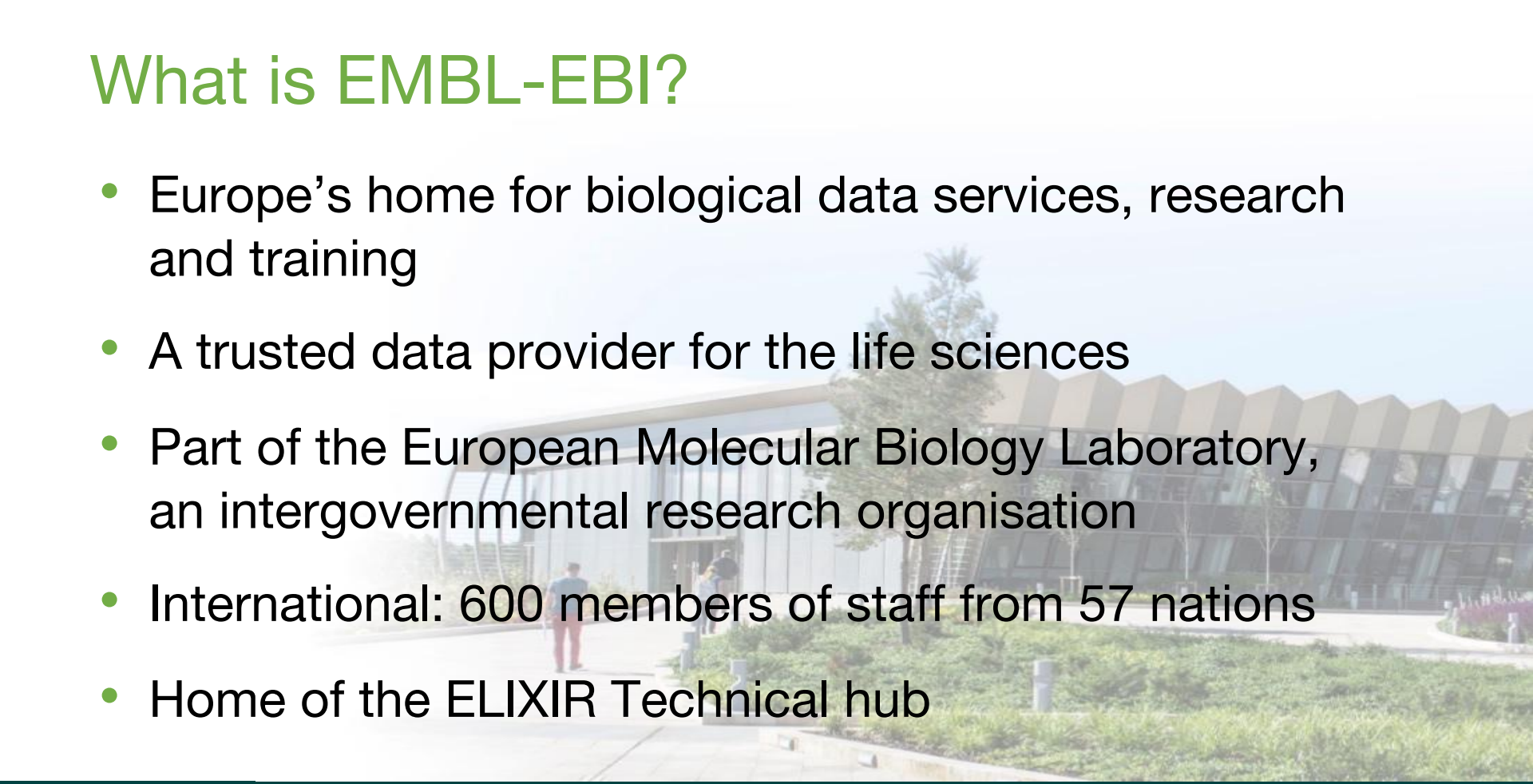
Dario Vianello (@vianello_d)
Cloud Bioinformatics Application Architect
Technology and Science Integration team
EMBL-EBI



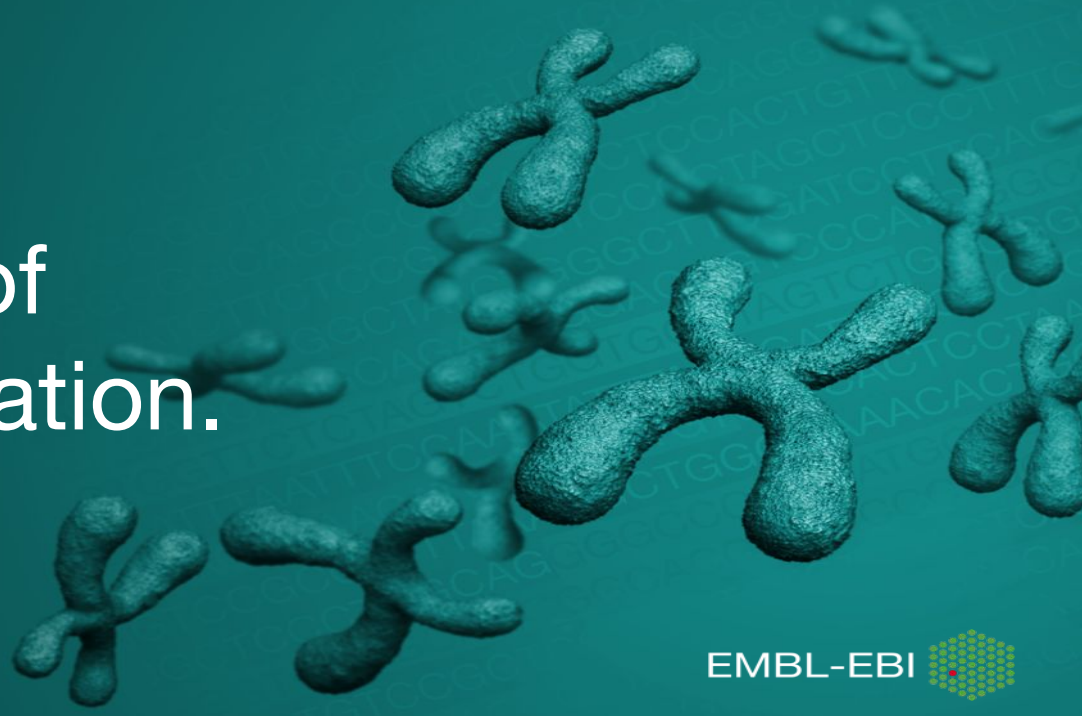
EMBL-EBI

What is EMBL-EBI?

- Europe's home for biological data services, research and training
- A trusted data provider for the life sciences
- Part of the European Molecular Biology Laboratory, an intergovernmental research organisation
- International: 600 members of staff from 57 nations
- Home of the ELIXIR Technical hub



Bioinformatics is
the science of storing,
retrieving and
analysing
large amounts of
biological information.



EMBL-EBI

So, from a **very** low point of view, EMBL-EBI is…

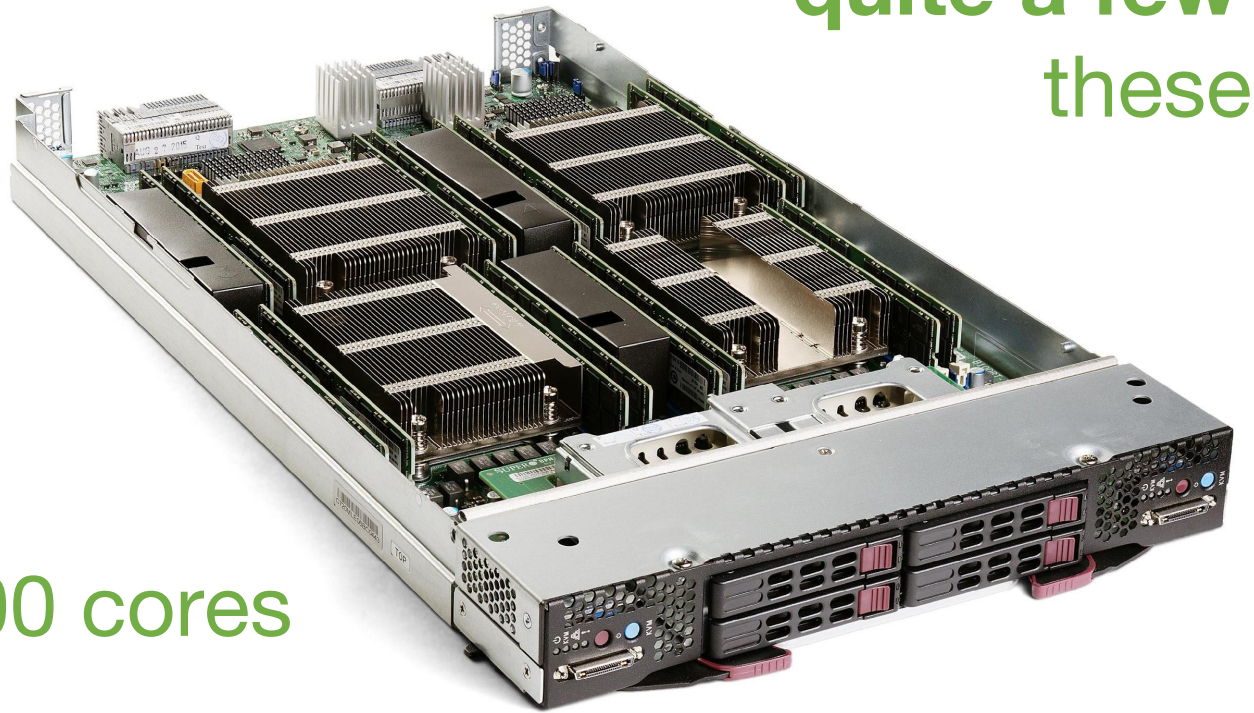
A **massive** number of these…

(~ 120 PB)

Packed into a reasonable number of ***those***…

(~200 racks)

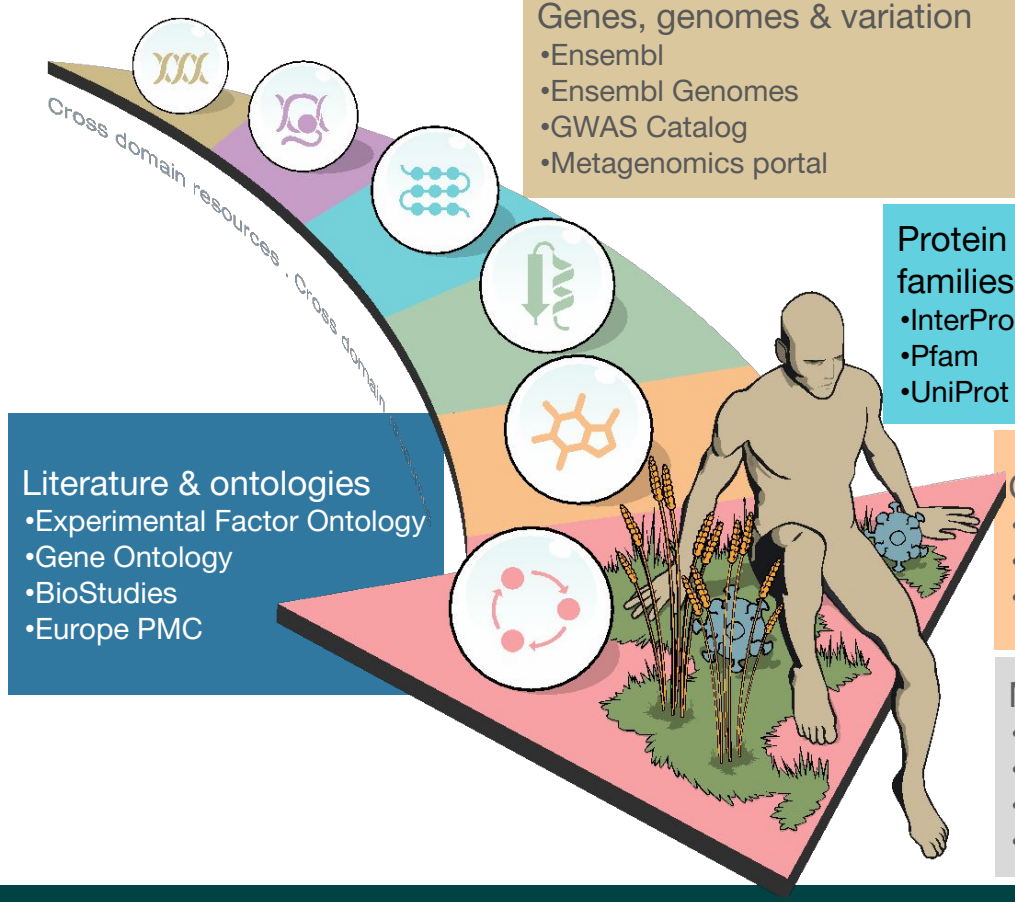
And accessed by **quite a few** of these…



~40000 cores

Not so easy, though...

Data sources at EMBL-EBI



Genes, genomes & variation
- Ensembl
- Ensembl Genomes
- GWAS Catalog
- Metagenomics portal

Gene, protein & metabolite expression
- Expression Atlas
- Metabolights
- PRIDE
- RNA Central

Protein sequences, families & motifs
- InterPro
- Pfam
- UniProt

Molecular structures
- Protein Data Bank in Europe
- Electron Microscopy Data Bank

Literature & ontologies
- Experimental Factor Ontology
- Gene Ontology
- BioStudies
- Europe PMC

Chemical biology
- ChEBI
- ChEMBL
- SureChEMBL

Systems
- BioModels
- BioSamples
- Enzyme Portal
- IntAct
- Reactome

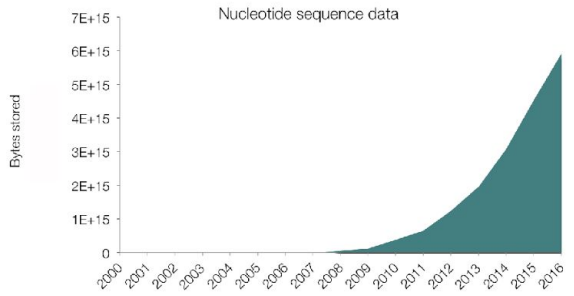
Molecular Archives
- European Nucleotide Archive
- European Variation Archive
- European Genome-phenome Archive
- ArrayExpress

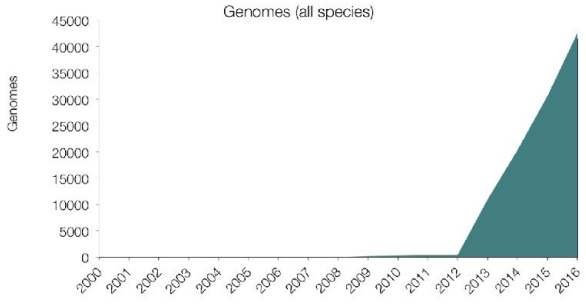
The EMBL-EBI Clouds

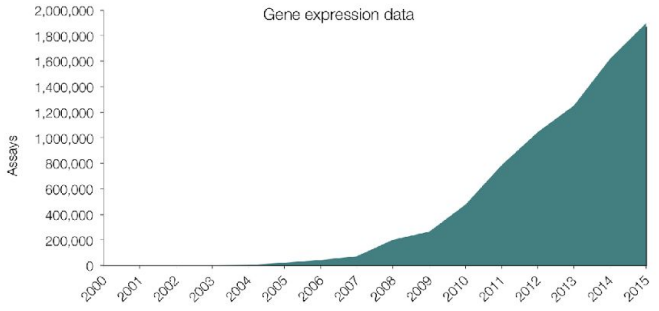
- vmware® cloud(s)
  - 40 Hypervisors
  - 2.4K VMs
- EMBASSY cloud
  - Powered by openstack®
  - 170 compute nodes
  - 11,000 vCPUs, 4GB RAM per vCPU
  - 2x10Gb network, 40Gb to storage networks
  - 3.5PB NFS Isilon, 3.5PB Object Storage (S3)

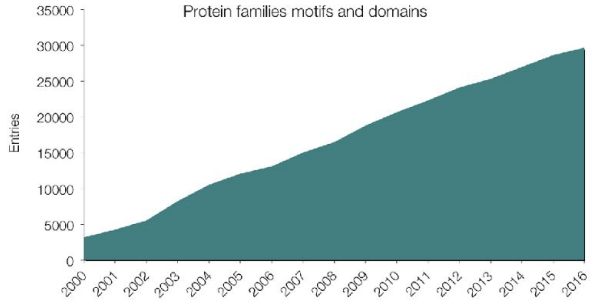
Not so easy, though...

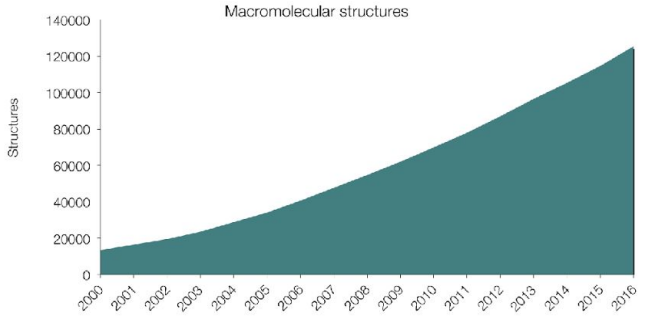
The Big Data problem



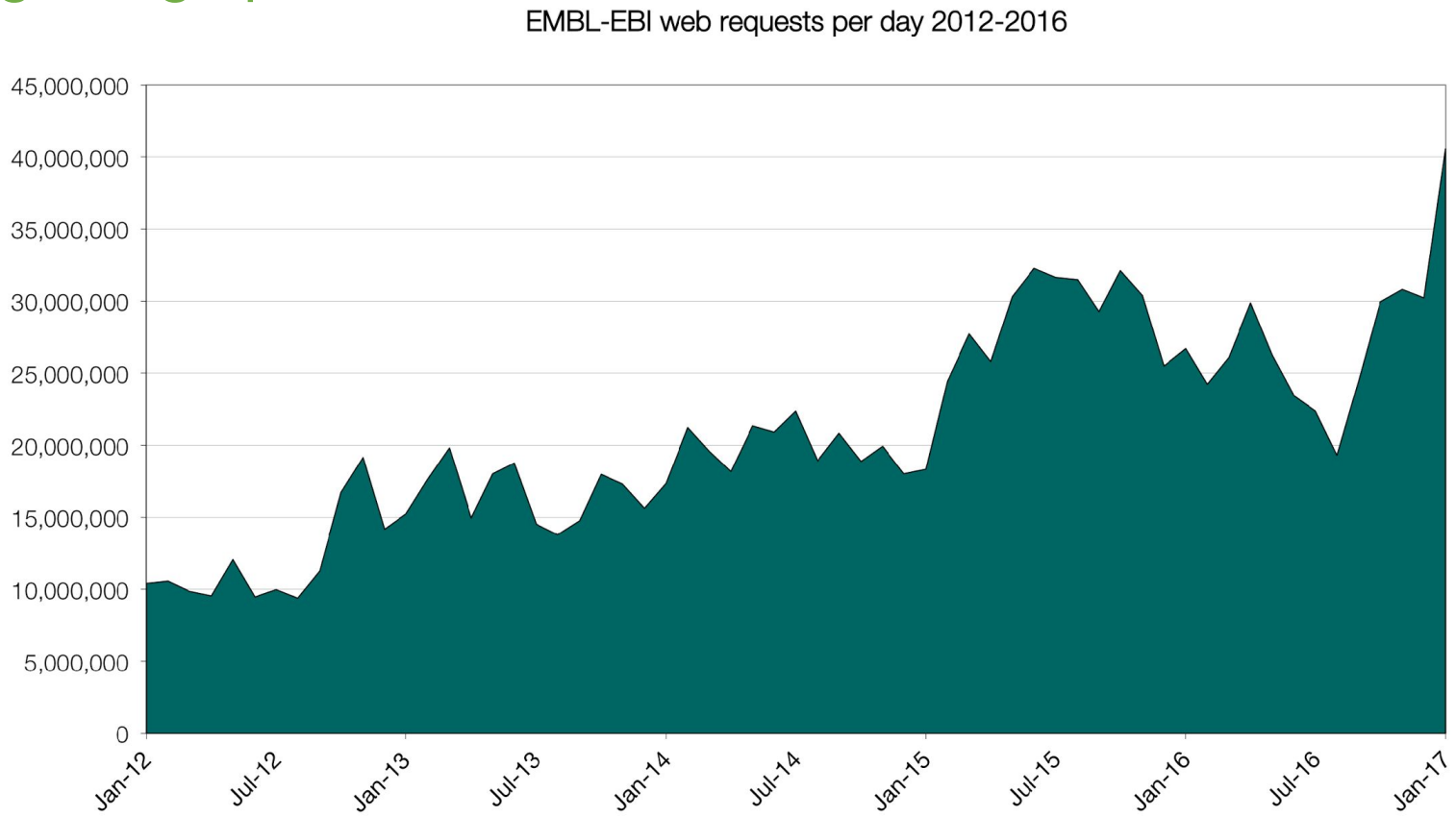








The Big Usage problem



Something needs to change...

Something needs to change...

Here
comes
“The Cloud”

Why EMBL-EBI is looking at Clouds?

- As data increases, have the community bring their compute to data
  But not all their compute to our data centre!
- Need to push relevant data sets and services out to cloud providers
  EMBL-EBI Embassy Cloud → ELIXIR → EOSC → AWS/GCP/MSA → ?
- Hybrid cloud an approach to optimise EMBL-EBI CapEx vs. OpEx
  Allow CapEx to lag demand & use OpEx to manage peaks

How to make data and workloads fly to the clouds?

The multi-cloud hybrid approach

Why?

- Long-term storage in the cloud can be expensive
- EMBL-EBI needs to keep a custodial copy anyway
- Avoid lock-in with a single cloud vendor
- Move workloads where it's cheaper to run them

Key factor:

Being able to run the **same** workload on-prem & in the cloud

The EMBL-EBI Cloud procurement for 2016/17

3 **lots**:

- OpenStack-compatible
- Scale Out
- VMware-compatible

Several projects:

- How to run **Science** in the cloud(s)
- How to **extend** our DC in the cloud(s)
- How to do **disaster recovery** in the cloud(s)

The pain points

1. Science will (likely) need to **adapt**

2. Data **movement** is challenging

3. Procurement is **lengthy** and **complex**

1. Science will (likely) need to adapt

- **Lifted** the the Marine Metagenomics Pipeline
  - Deployed **batch** system in the cloud
  - **Adapted** the pipeline to run in clouds and to pull/push from/to FTP
- **Scalability** testing in Embassy Cloud
- **Ran** in 4 cloud(s)
- **Compared** running times and costs

MMG, Running times

ENA study	Size	Cluster (min)	Embassy (min)	GCP (min)	Azure (min)
ERP000554	22 MB, 1 run	99	89	96	96
SRP000664	61 MB, 4 runs	98	89	92	91
SRP005784	160 MB, 10 runs	99	92	93	94
ERP005409	6.3 GB, 49 runs	274	729	629	579
ERP006630	14 GB, 10 runs	1063	N/A	4583	N/A

MMG, the bill(s)

		UKCloud		GCP		Azure		Embassy	
ENA Code	vCPUs	Running time (min)	Cost (£)	Running time (min)	Cost (£)	Running Time (min)	Cost (£)	Running time (min)	Cost (£)
ERP000554	40	95	3.64	96	3.74	96	8.69	94	0.77
SRP000664	40	266	9.1	203	7.22	200	11.18	226	1.85
SRP005784	80	338	18.34	292	15.17	281	22.5	300	4.30

MMG, the (estimated) Spot bill

		GCP			Embassy
Run	vCPUs	Running time (min)	Normal VMs (£)	Preemptible VMs (£)	Normal VMs (£)
ERP000554	40	96	3.74	1.39	0.77
SRP000664	40	203	7.22	2.41	1.85
SRP005784	80	292	15.17	6.48	4.30
ERP005409	640	626	274	65.34	/
ERP006630	640	4583	2113	485.96	/

So far so good, but...

Pipelines must be deployable **on-demand**, as **compute**

So we need *DevOps*, but for **Research** (*ResOps*)

Since:

- Clouds will **move** (as the real ones!)
- We'll likely be in a hybrid **multi-cloud** environment
- We might need to re-deploy short notice somewhere else

2. Data movement is challenging

Moving PBs isn’t easy

Requires:

- (a lot of) **Bandwidth**
- (a lot of) **Time**
- (a lot of) **Money** (*egress* charges)
- (a lot of) **Authorisations** to do so

2. Data movement is challenging

- **What** data do we store?
  - *Personal, Scientific Research, Administrative, Professional, Private*
- **How** sensitive is the data?
  - *Controlled, Confidential, Restricted, Public*
- What are the **storage options**?
  - *'Vault', Managed, Standard, Any Cloud, EU Cloud, Hosting*

End up with a matrix describing
what can go where!

3. Procurement is complex

- **Benchmarking** is key
  - *Not all clouds are born equal*

- **Price** comparison is complex
  - *Different pricing models*
  - *Different contractual models*

- Post-award contract **negotiation**
  - ***Brace*** yourself
  - *~6 months in some cases*

Let’s start to address those issues!

1. Science will need to adapt
   - Training (**ResOps**)
   - Try to lower those barriers (HNSciCloud, EOSCpilot)

2. Data transfer is challenging
   - Rethink the way we do it

3. Procurement is complex
   - Update purchasing models

At the European Level...

A ***pre-commercial*** procurement

Procurers: CERN, CNRS, DESY, EMBL-EBI, ESRF, IFAE, INFN, KIT, STFC, SURFSara
Experts: Trust-IT & EGI.eu

Total procurement budget >5M€

The group of procurers have committed:

- Procurement **funds**
- **Manpower** for testing/evaluation
- **Use-cases** with applications & data
- **In-house** IT **resources**

Resulting services will be made available to end-users from many research communities

Co-funded via H2020 Grant Agreement 687614

HNSciCloud - The (hopefully) real impact

Key challenges:

- *Data transparency layer*
- *Federated AAI*
- *Procurement Frameworks*

They **do** sound a bit familiar, *don't they?*

Take home messages

- Institutional cloud use
  *Both a **legal** and a **technical** problem*

- Need to develop a shared language and understanding
  Legal may want the machines labelled that the contain EMBL data!

- Update purchasing model
  *Purchase order doesn't bode well with **pay-per-use***

Thank you! ;-)