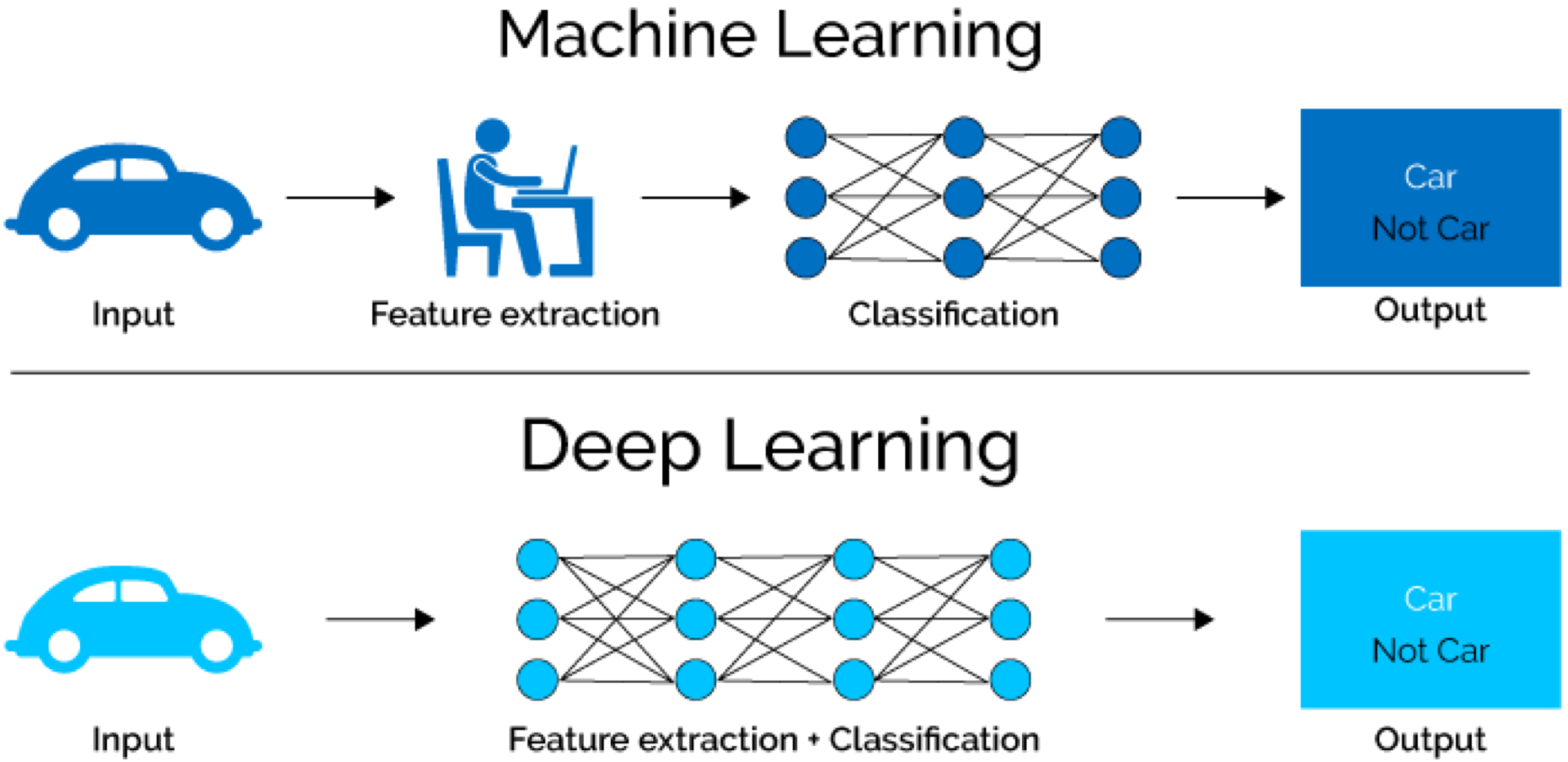
Deep Learning and Machine Learning in Hybrid Clouds

Davide Salomoni (davide@infn.it)

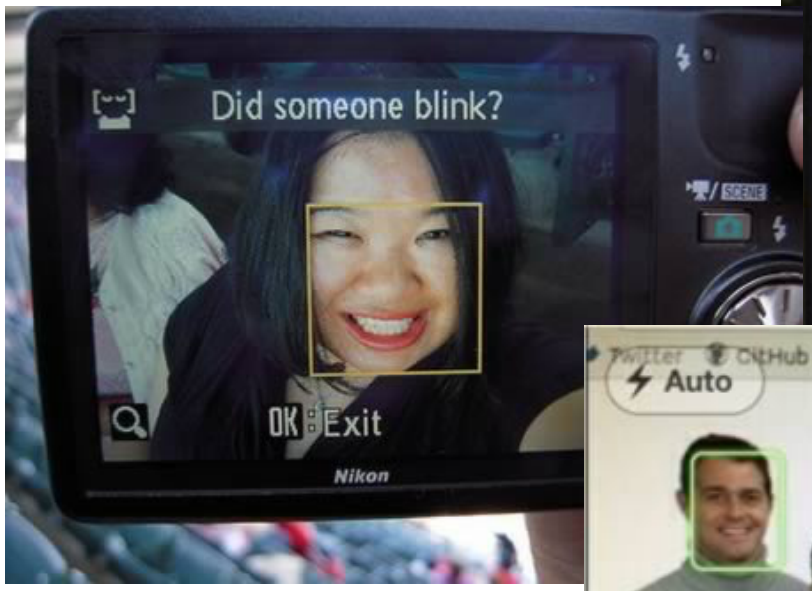
SOSC 2018

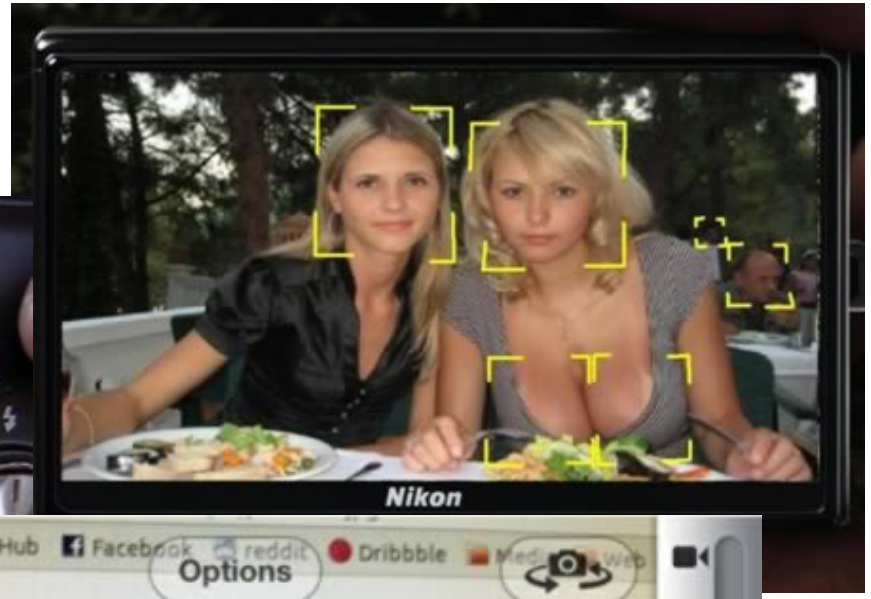
Perugia, Sep 20, 2018

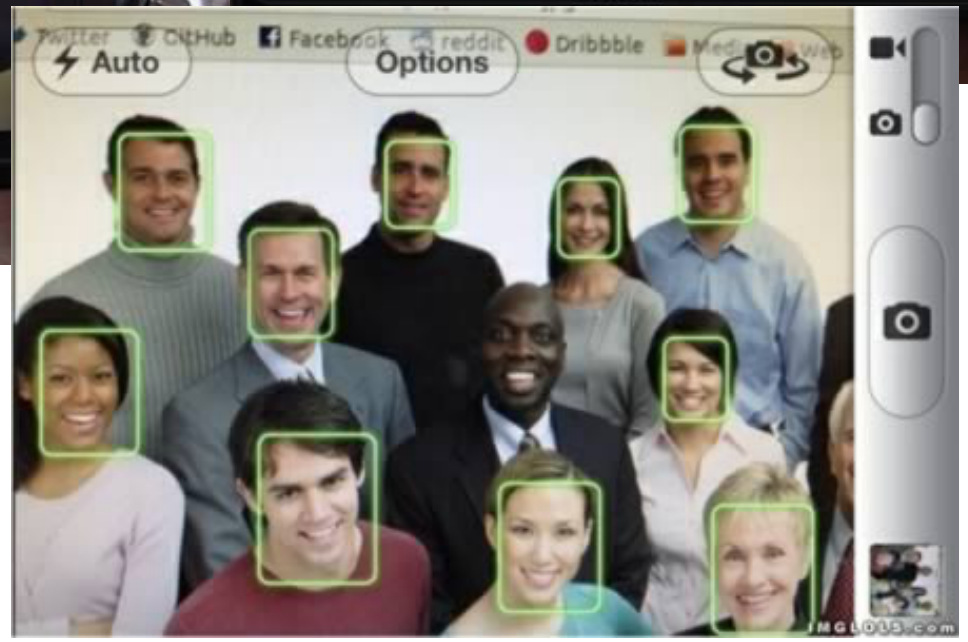
By now we all know what we mean by ML/DL…

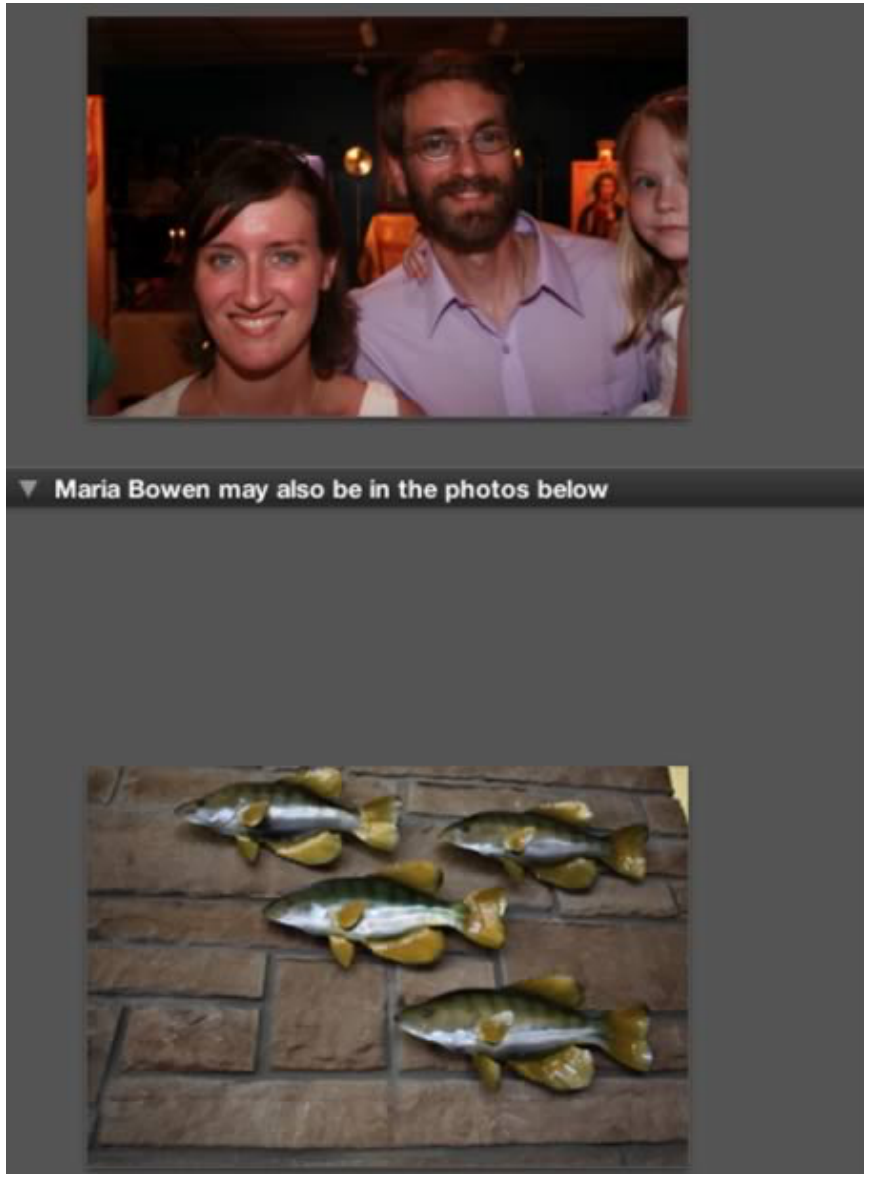


A word of caution about machine learning…









https://bit.ly/2D9ymgX

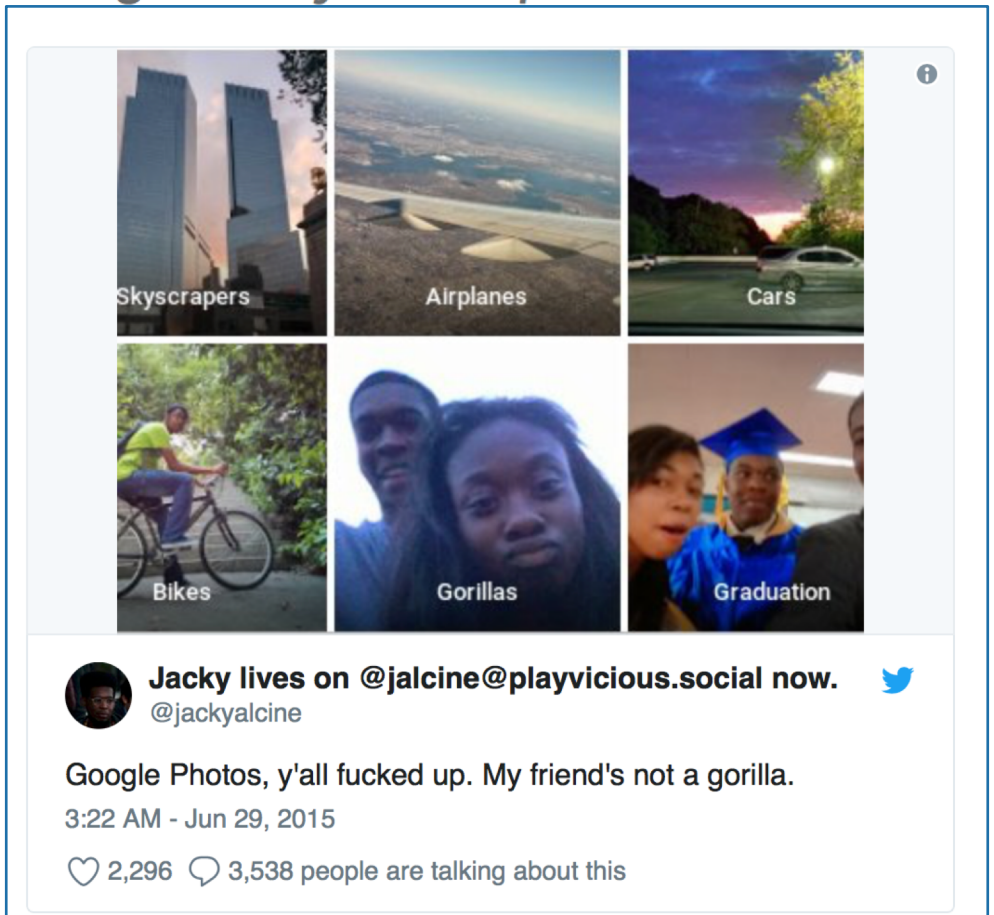
GOOGLE | TECH | ARTIFICIAL INTELLIGENCE

https://bit.ly/2D7W3ku

Google 'fixed' its racist algorithm by removing gorillas from its image-labeling tech

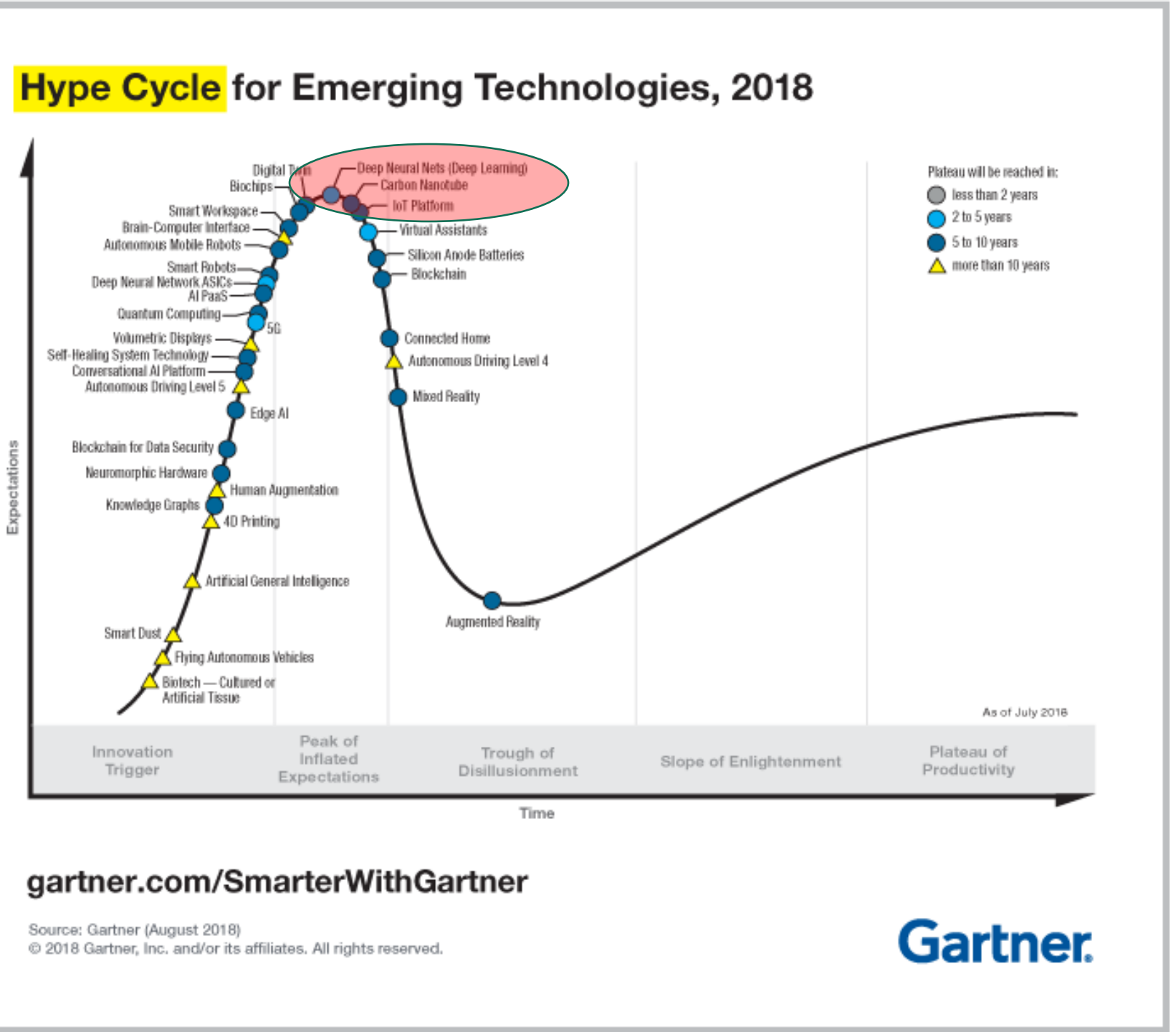
Nearly three years after the company was called out, it hasn't gone beyond a quick workaround

By James Vincent | @jjvincent | Jan 12, 2018, 10:35am EST



Whatever, ML/DL is here to stay, at least until the next buzzword kicks in

Davide Salomoni



So, you want to apply your ML/DL-related ideas

- But where do you find a suitable **ML/DL environment** that works…
  - … for you (and not against you)…
  - … configured the way you want (and not the way the resource owner thinks it's best)…
  - … quickly and for the time you need it (you don't want to wait until you're old to see whether your ideas worked)

usable

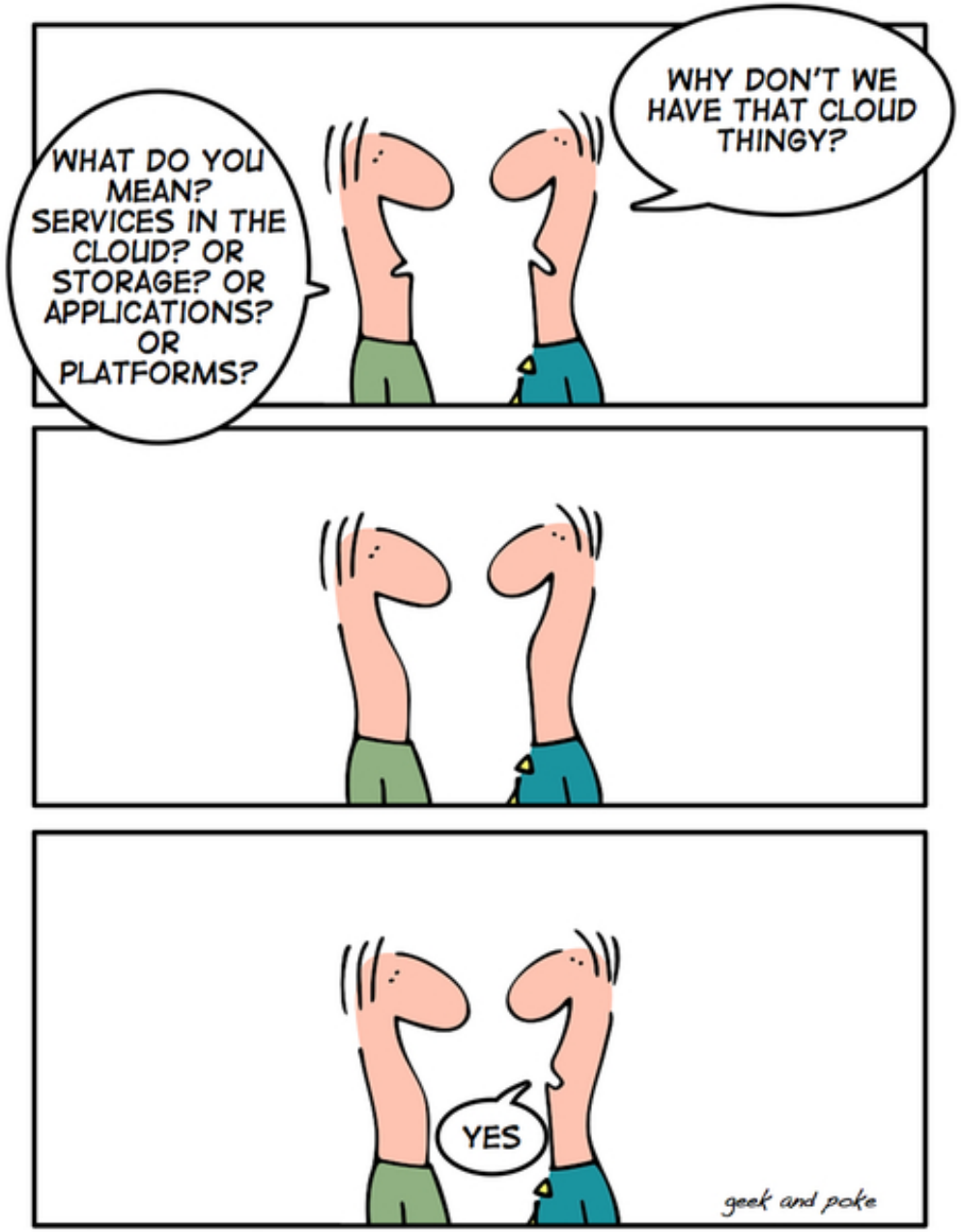
flexible

time-efficient, cost-effective

Let's add buzzwords

- We want to use «**The Cloud**» to find, self-provision and use the [compute, storage] resources we need.
- But what *is* "The Cloud"?
  - See http://goo.gl/eBGBk for the classical NIST definition. It is in practice a way to:

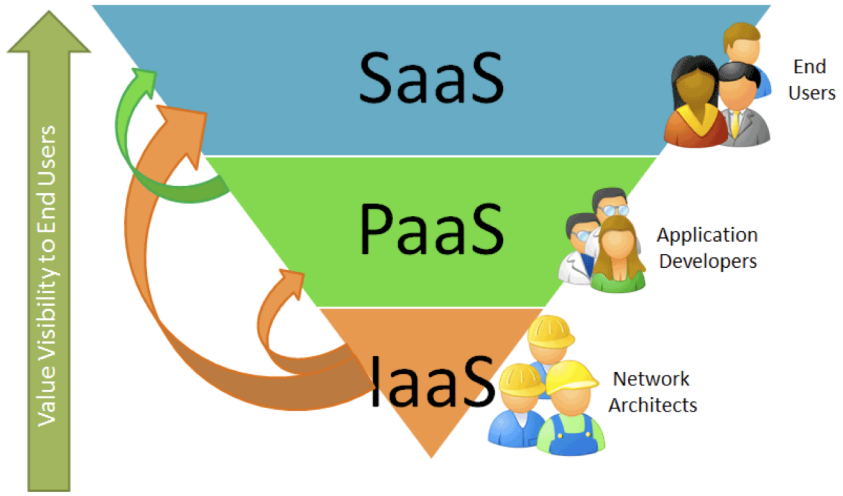
1. Supply
2. information and communication technologies
3. as a service



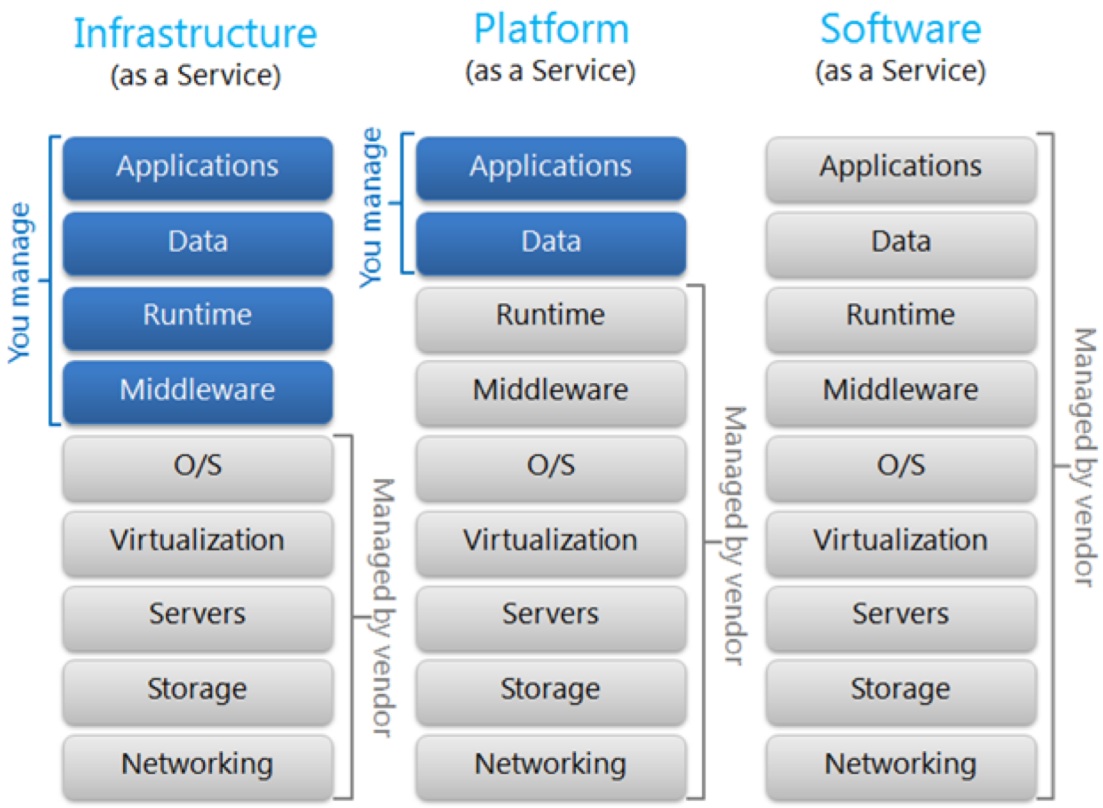
THE CLOUD THINGY

The 5 Cloud postulates

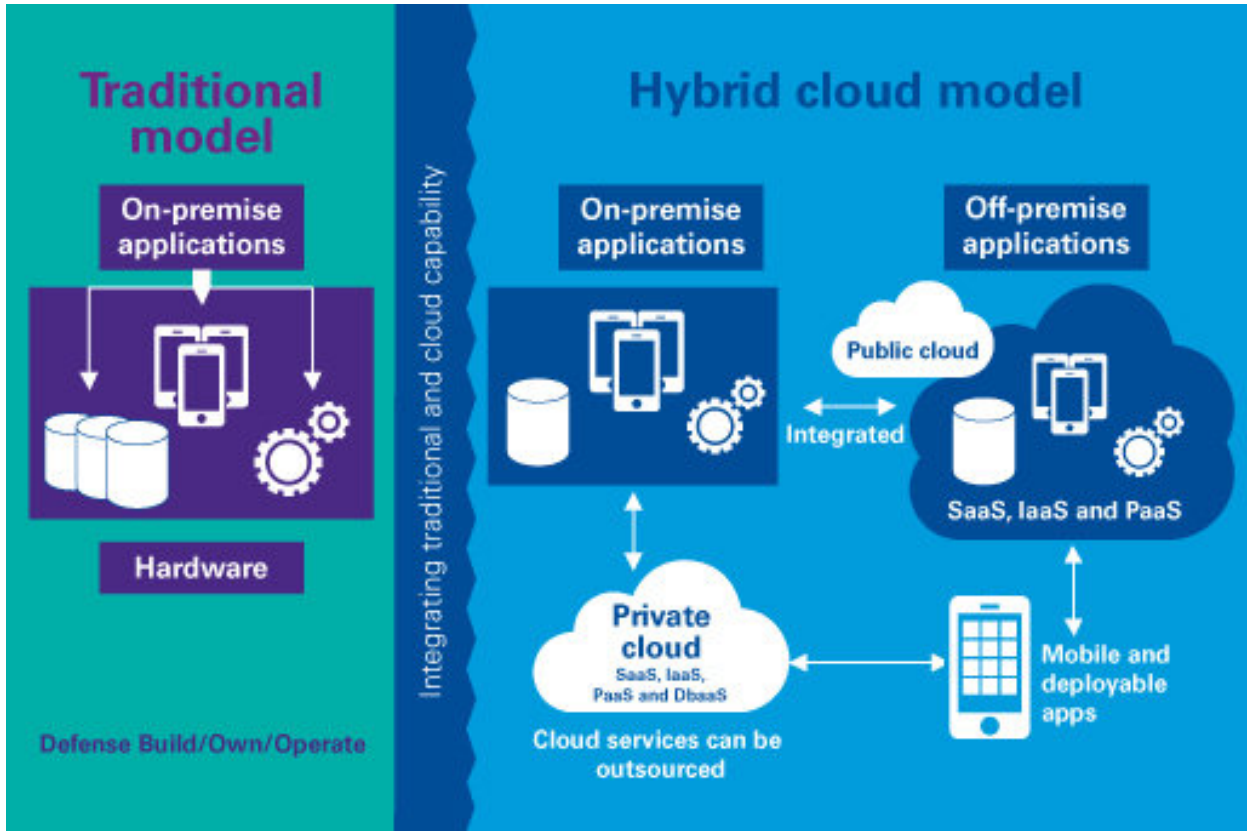
1. **Self-service, on-demand provisioning**
2. **Network-based access**
3. **Resource sharing**
4. **Elasticity (with *infinite resources*)**
5. **Pay-per-use**



What matters at the end
are the applications.



To provision resources (including those for ML/DL), we want to consider **Hybrid Clouds** (*)

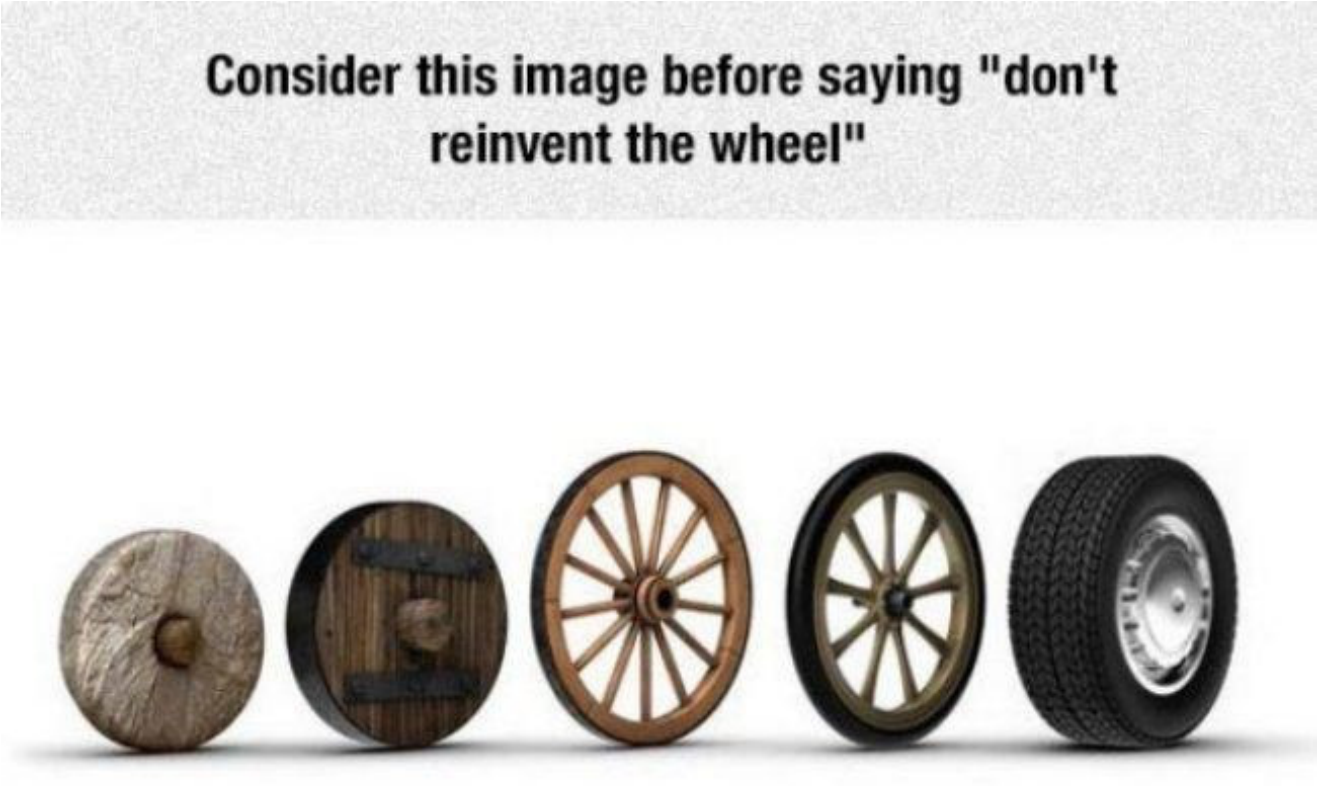


Source: KPMG international, 2016

(*) I won't discuss here why it is so, see Q&A or seek me offline

How?

- We certainly don’t want to reinvent the wheel
- We’d rather start from what is available and adapt / evolve it to satisfy our needs
- One can certainly do all of this by putting scores of PhD students to write ad-hoc, home-made and self-maintained solutions, but still..



INDIGO-DataCloud

- INDIGO-DataCloud (2015-2017) was a software development project that developed a comprehensive set of open source solutions to facilitate the exploitation of hybrid cloud infrastructures for scientific applications.
- **The ElectricIndigo Release (https://www.indigo-datacloud.eu/service-component):**
  - 47 open source modular components, distributed via 170 software packages, 50 ready-to-use Docker containers
  - Supported operating systems: CentOS 7, Ubuntu 16.04
  - Supported cloud frameworks: OpenStack Newton, OpenNebula 5.x (plus connection to Amazon, Azure, Google)
  - Download it from the INDIGO-DataCloud Software Repository: http://repo.indigo-datacloud.eu/index.html

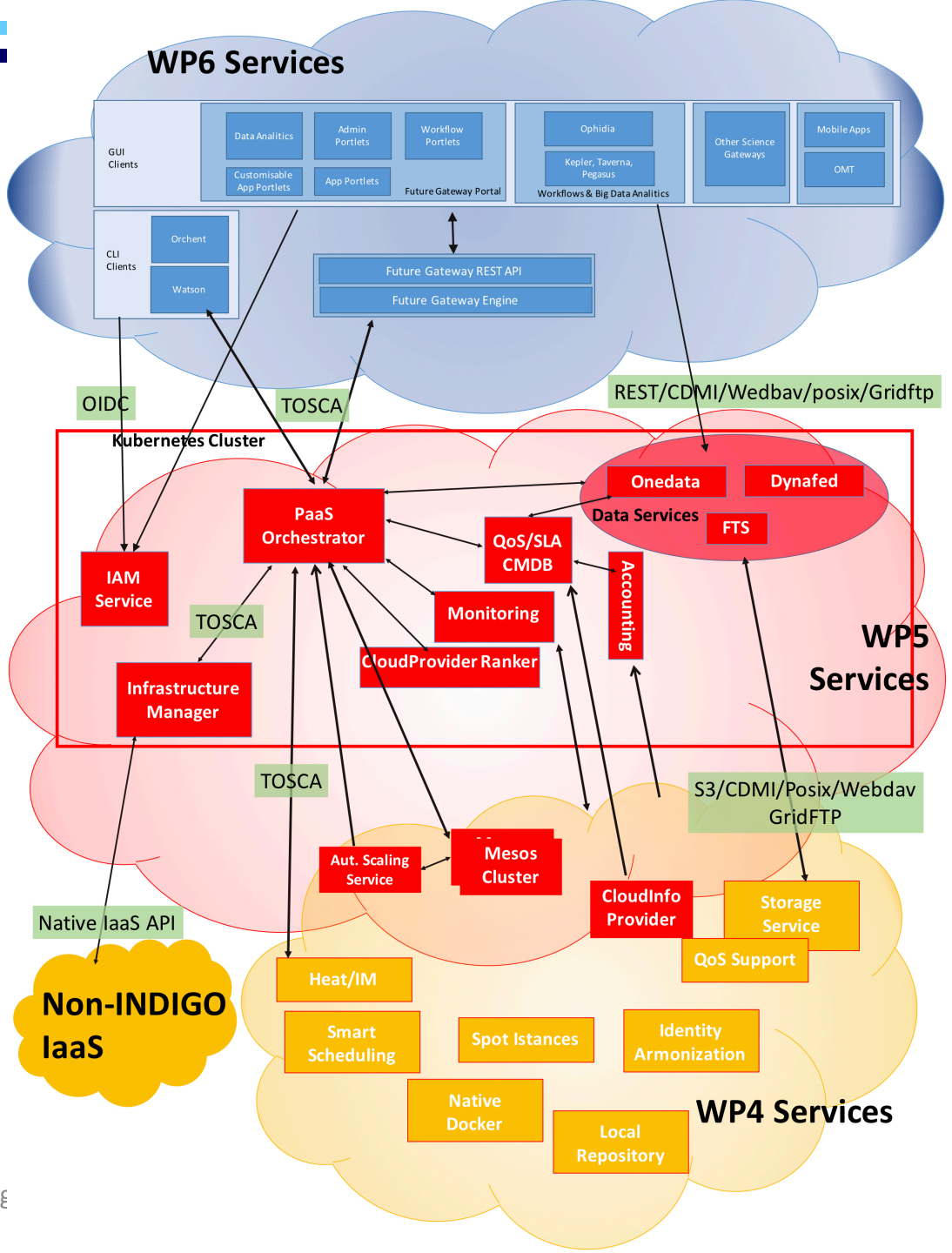
The ElectricIndigo Release

- The **ElectricIndigo modular software components** are organized around **5 areas**:

1. Application-level Interfaces to Cloud Providers and Automated Service Composition
   - For users **porting their apps to the Cloud**
2. Flexible Identity and Access Management
   - For users needing to **handle AAI**
3. Data Management and Data Analytics Solutions
   - For users **managing distributed [big] data**
4. Programmable Web Portals, Mobile Applications
   - For the **creation of front-ends**
5. Enhanced and Scalable Services for Data Centers and Resource Providers
   - For providers wishing to **optimize/enhance their service offerings**

The INDIGO-DataCloud Reference Architecture

- See details at http://arxiv.org/abs/1603.09536 and at https://doi.org/10.1007/s10723-018-9453-3



INDIGO-DataCloud & Sons

- INDIGO-DataCloud has now two ongoing follow-on European projects, complementary in nature:
  - **eXtreme-DataCloud** (http://www.extreme-datacloud.eu/), focusing on the development of **scalable technologies for federating storage resources and managing data** in highly distributed computing environments
  - **DEEP-HybridDataCloud** (https://deep-hybrid-datacloud.eu), focusing on intensive **computing services exploiting specialized hardware components**, such as GPUs, low-latency interconnects, and others resources.

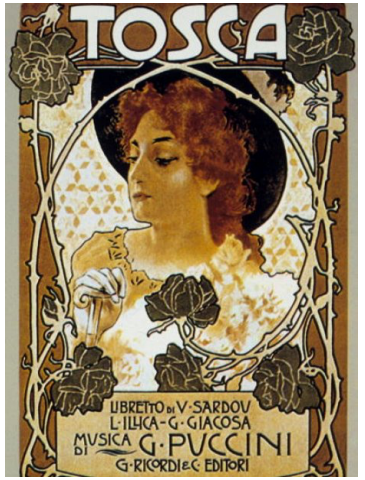
Moving beyond IaaS

How do you specify your *requirements*?

- **«Requirement»** : in general, what you would like to see to automagically appear for you in the Cloud
- For example, an auto-scalable cluster of GPU-equipped machines (physical machines, or containers), associated to some databases (maybe in an HA configuration), auto-configured to connect to some distributed file system, perhaps part of a Spark cluster, or of a virtual batch system.
- You don't want (trust me on this one) to instantiate, operate, manage the above manually.
- In INDIGO and follow-on projects, we use an ***open templating language*** to specify requirements, called **TOSCA**

TOSCA, in short

- **T**opology and **O**rchestration **S**pecification for **C**loud **A**pplications
- It standardizes the language to describe:
  - *The structure* of an IT service (its **topology** model)
  - *How to orchestrate operational behavior* (plans such as build, deploy, patch, shutdown, etc.)
- It is a declarative model that spans applications, virtual and physical infrastructures.
- INDIGO supports TOSCA declarations at the IaaS, PaaS and SaaS levels to automatize the definition & instantiation of services.



- An example of a **TOSCA topology template**
- For more examples, see https://github.com/indigo-dc/tosca-templates

```yaml
tosca_definitions_version: tosca_simple_yaml_1_0

description: Template for deploying a single server with predefined properties.

topology_template:
  inputs:
    cpus:
      type: integer
      description: Number of CPUs for the server.
      constraints:
        - valid_values: [ 1, 2, 4, 8 ]

  node_templates:
    my_server:
      type: tosca.nodes.Compute
      capabilities:
        # Host container properties
        host:
          properties:
            # Compute properties
            num_cpus: { get_input: cpus }
            mem_size: 2048  MB
            disk_size: 10 GB

  outputs:
    server_ip:
      description: The private IP address of the provisioned server.
      value: { get_attribute: [ my_server, private_address ] }
```

With TOSCA, you can create rather complex topologies

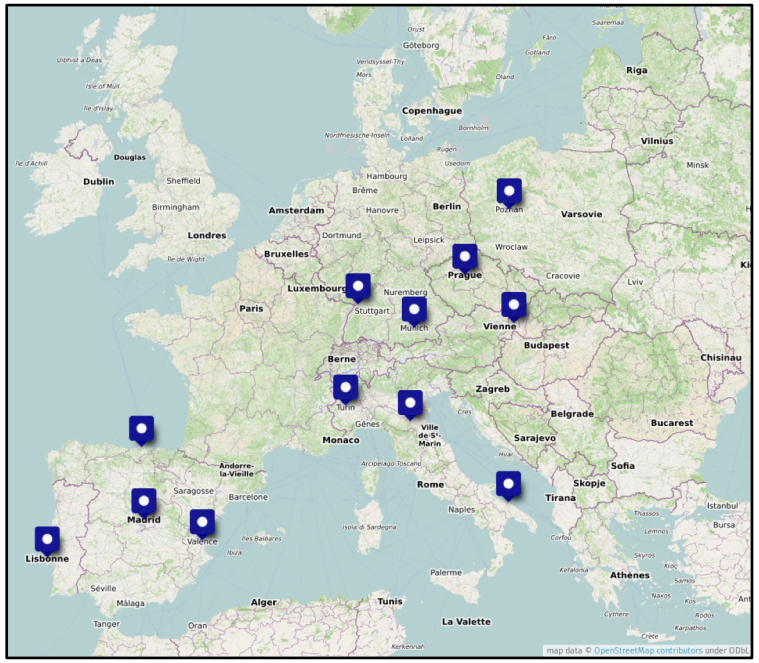
GitHub
INDIGO - DataCloud
TOSCA Documents and Dockerfiles per Use Case
1.b) Automated Build
INDIGO - DataCloud
INDIGO-DataCloud Docker Hub Organization
1.a.2) Dockerfile (commit)
1.a.1) build, push
Champion + JRA
User
2) Deploy TOSCA with Vanilla VM / Container
Future Gateway API Server
1) Stage Data
6)Access Web Portal
ONEDATA
Provider
Orchestrator
Other PaaS Core Services
IM
5) Mount
Virtual Elastic LRMS Cluster
Clues
WN
...
WN
WN
Galaxy
Public IP
Front-End
4) Install / Configure
IM
Cloud Sites
Heat
OpenNebula
OpenStack

Now on to ML/DL…

- With the PaaS technology shown above, we can e.g. **automate the instantiation of on-demand Spark clusters on hybrid Clouds**.
- However, we also want to support **computing techniques for the analysis of very large datasets, exploiting specialized hardware components:**
  - GPUs…
  - … low-latency interconnects…
  - … and other resources usually accessed as “bare metal”.
- Also ensuring interoperability with existing applications and infrastructures → these are the main objectives of the DEEP-Hybrid DataCloud project.

The DEEP consortium

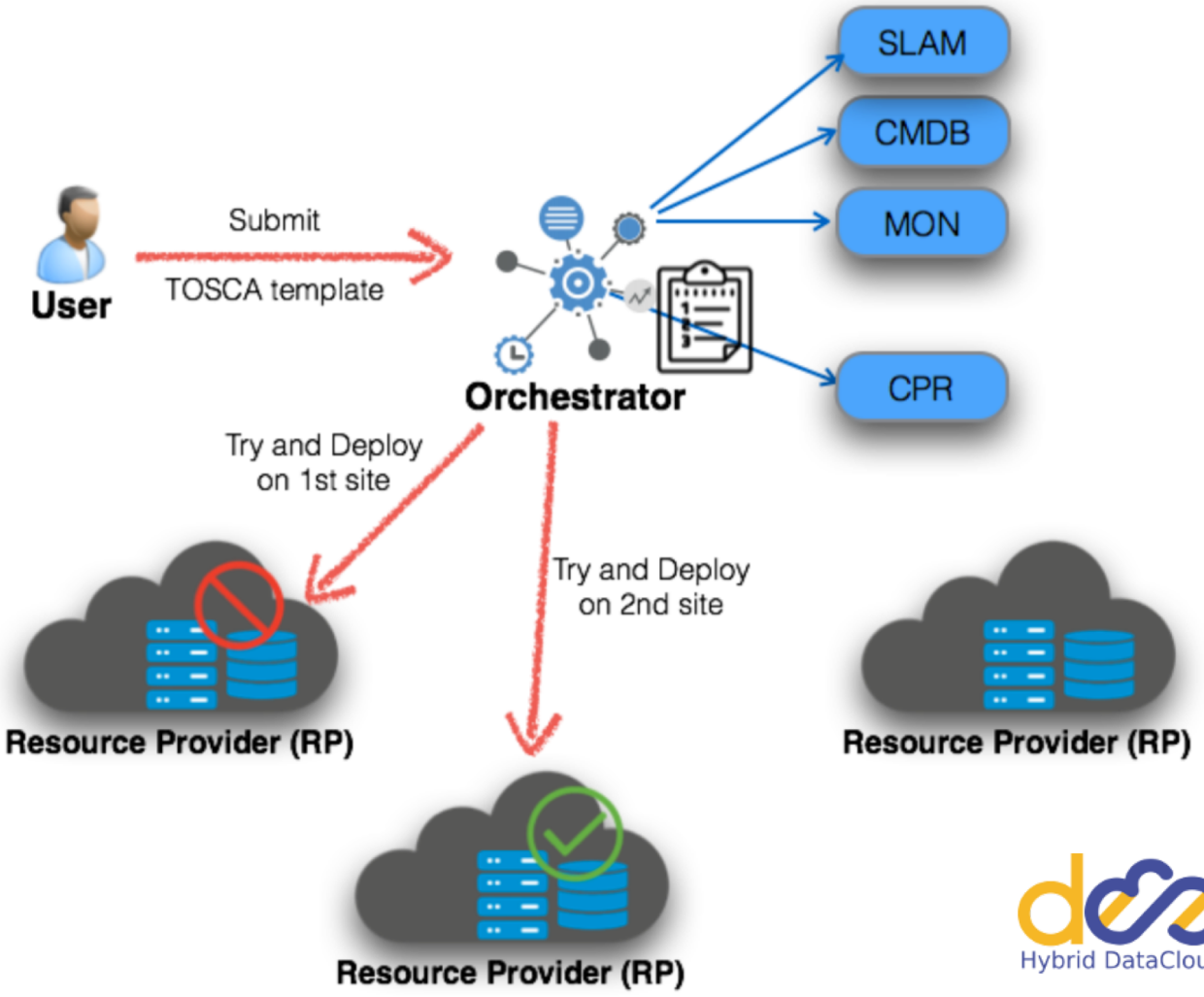
- A balanced set of partners
  - Strong technological background on development, implementation, deployment and operation of federated e-Infrastructures
- 9 academic partners
  - CSIC, LIP, INFN, PSNC, KIT, UPV, CESNET, IISAS, HMGU
- 1 industrial partner
  - Atos
- 6 countries
  - Spain, Italy, Poland, Germany, Czech Republic, Slovakia

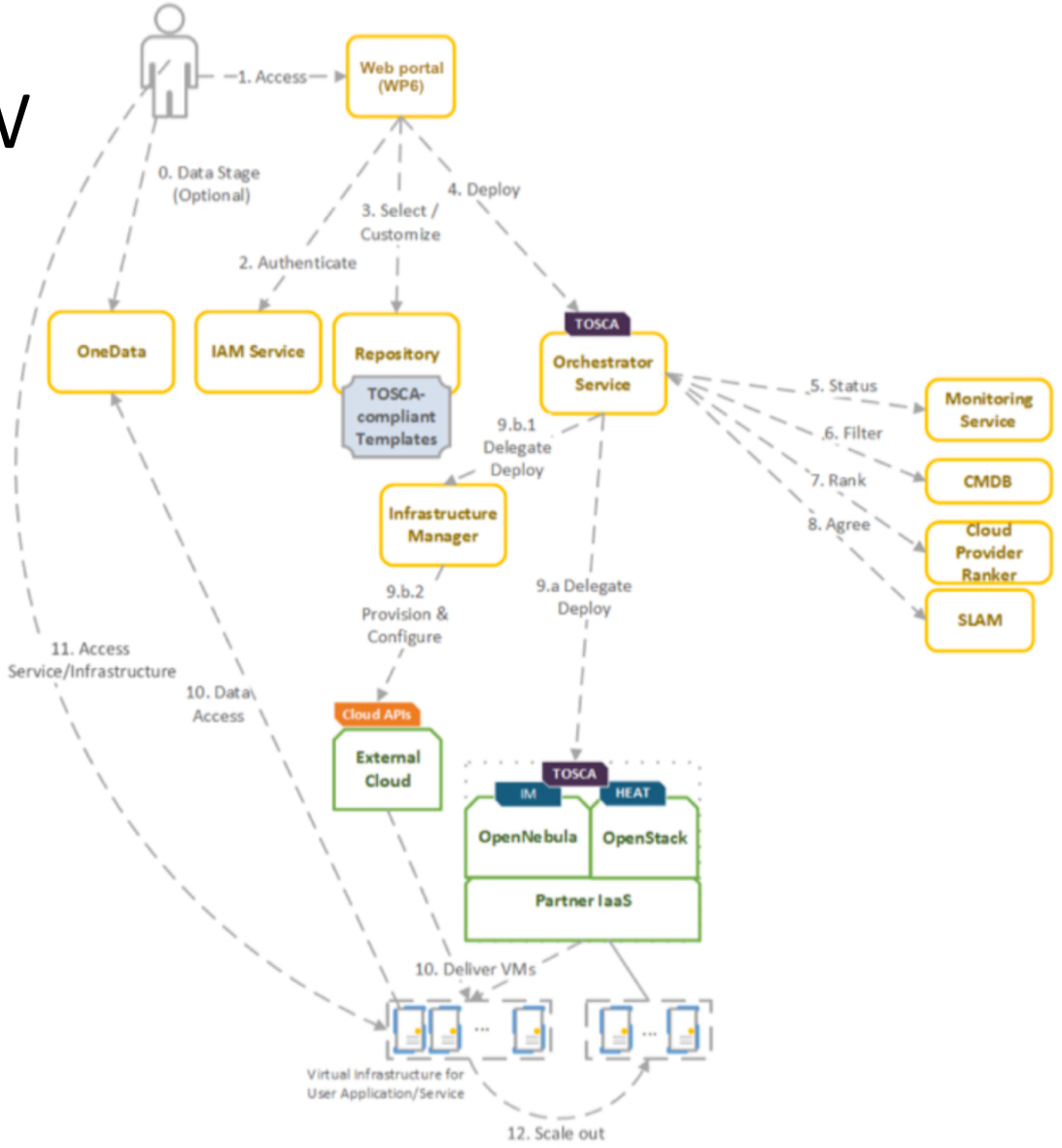


CSIC Consejo Superior de Investigaciones Científicas · LIP · INFN Istituto Nazionale di Fisica Nucleare · PSNC · KIT Karlsruhe Institute of Technology · Universitat Politècnica de València

cesnet · Ústav Informatiky Slovenská akadémia vied · Atos · Helmholtz Zentrum München German Research Center for Environmental Health

DEEP: a 10,000 foot view





Hybrid DataCloud

Some use cases, abstracted

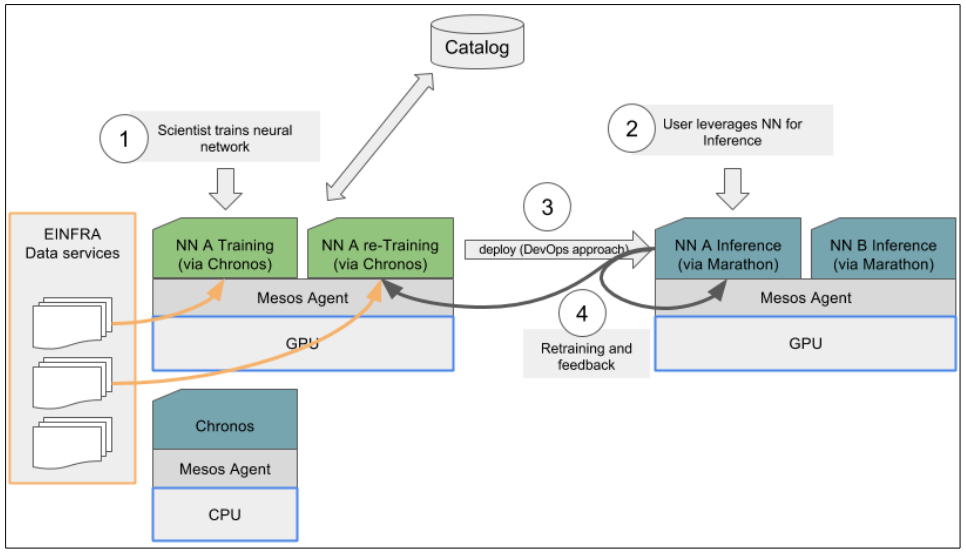


Figure 3: Initial architecture proposed to support a deep learning application using private cloud services.

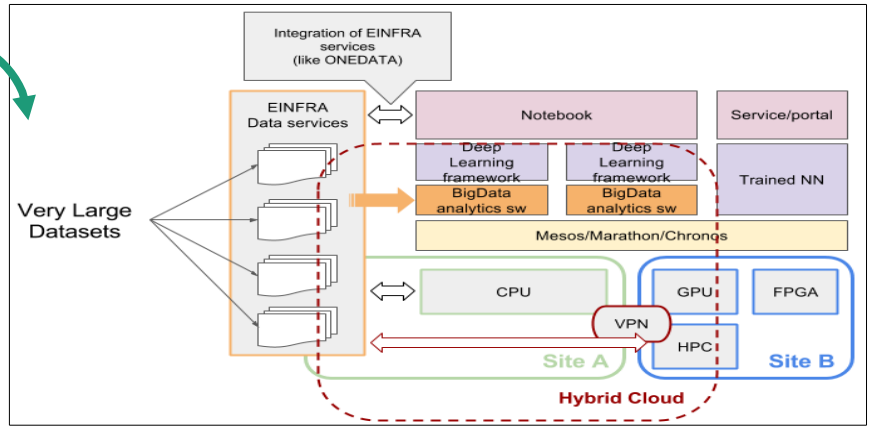


Figure 4: Initial architecture using a Hybrid Cloud approach.

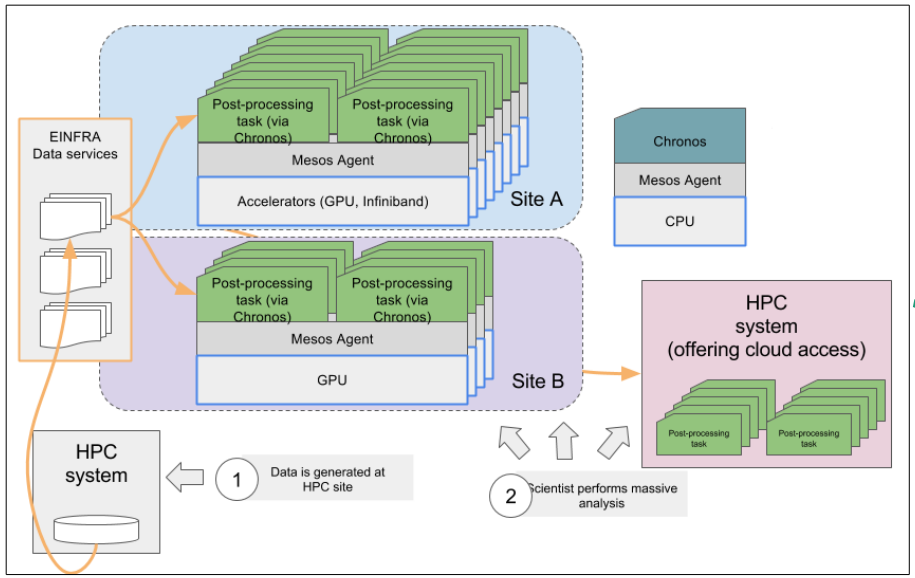


Figure 5: Massive post-processing of very large datasets using containers.

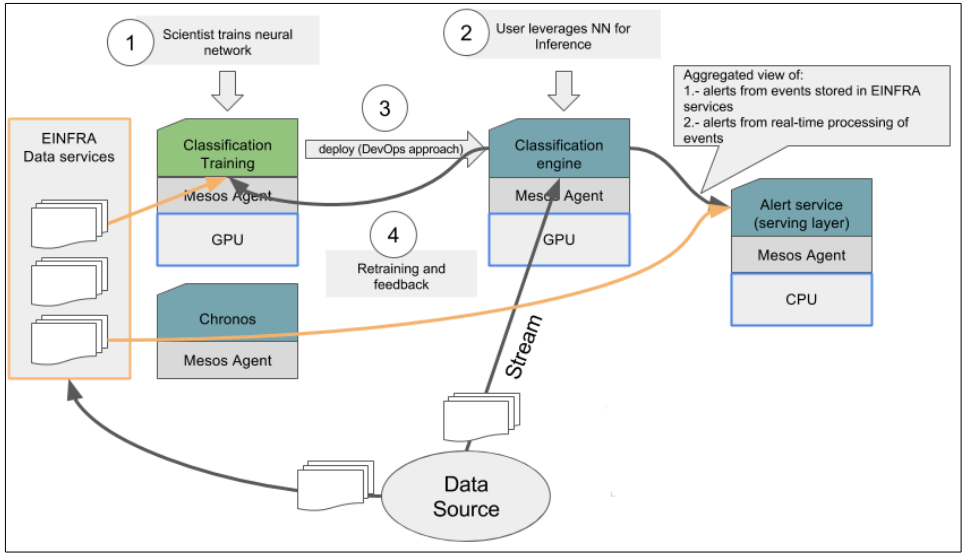
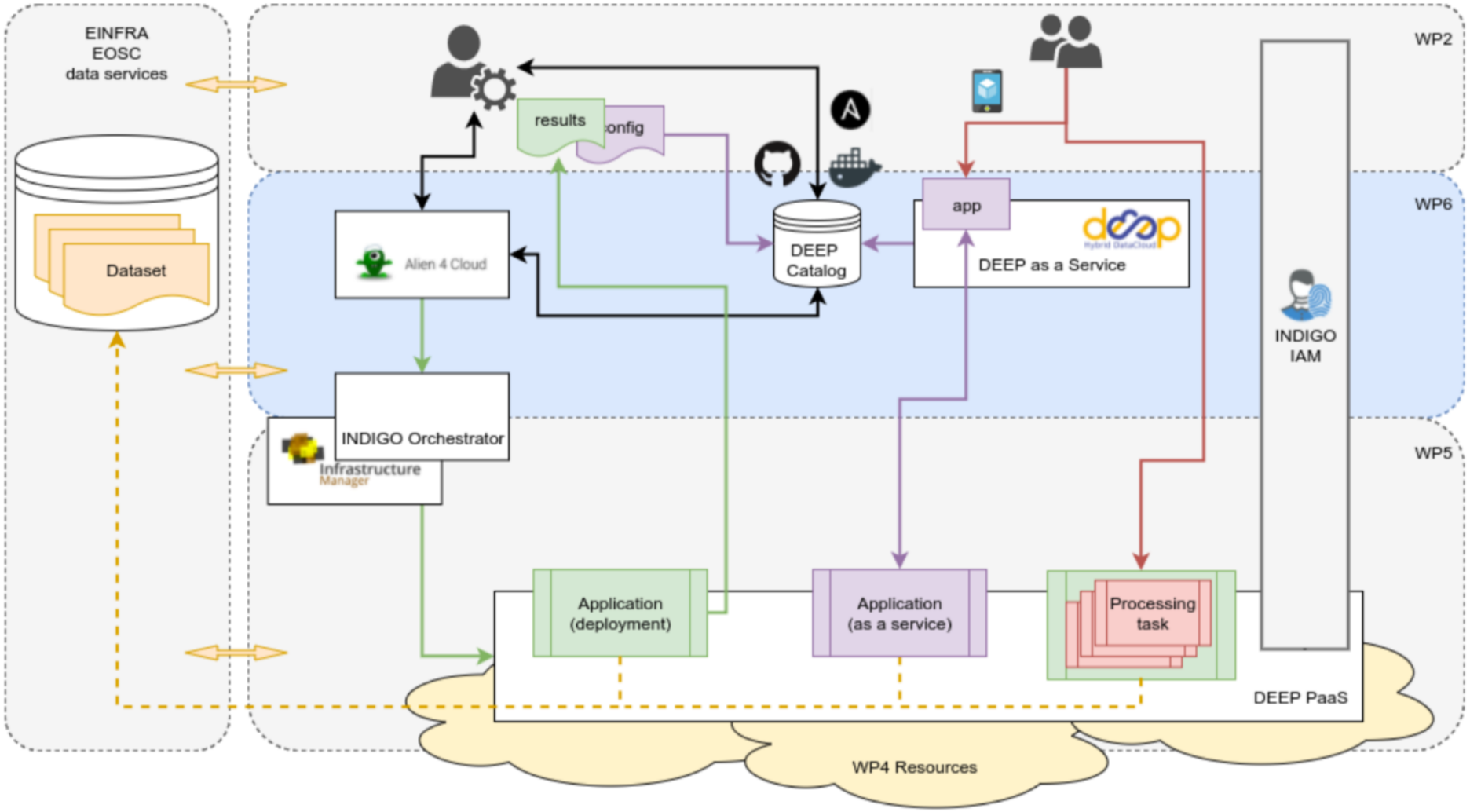


Figure 6: Analysis of massive real-time data streams following a lambda architecture.

Some use cases, concrete

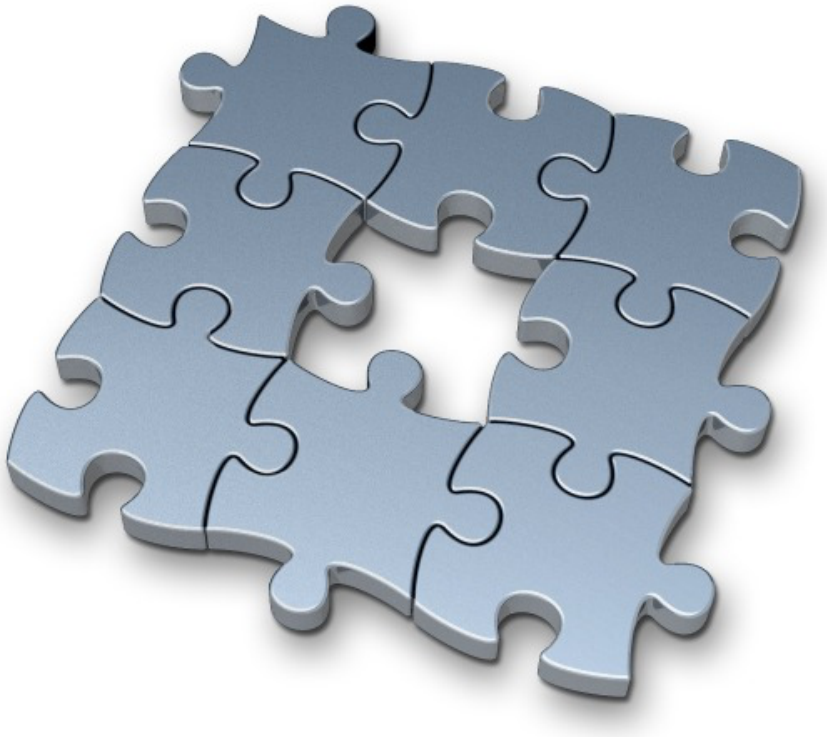
- **Retinopathy detection**: DL application for the classification of disease progression of diabetic retinopathy – classify stage and progression of the disease based on labelled image data taken from the back of the human eye. Dataset from https://www.kaggle.com/c/diabetic-retinopathy-detection.
- **Plant classification** with DL. Training data comes from images & metadata (author, data, url, species, id, etc.) from portals such as iNaturalist or Natusfera.
- **Pattern recognition in satellite images** (from Copernicus or NASA), associated to in situ measurements. Environmental data are used to validate the predictions.
- **Analysis of massive online data streams**, coming e.g. from monitoring of log files in big data centers.

DEEP, architectural diagram



What is missing?

- Application composition: beyond a text-based TOSCA template (see earlier), we want a tool capable of
  - graphically composing services to create a topology, and
  - deploying the created topology on some Clouds
- Software components that are able to properly parse TOSCA (and TOSCA extensions) through transparent hybrid multi-Cloud deployments, involving the use of specialized computing devices such as GPUs or low latency interconnects.



Two simple video demonstrations

1. How the INDIGO-DataCloud orchestrator can be used from the command line to deploy Jupyter notebooks over cloud infrastrucures. This is eventually realized with Docker containers with CUDA and Tensoflow through a Cloud-based Mesos cluster.
2. How the Alien4Cloud product is being extended by DEEP to graphically create a topology and the related TOSCA template, which will eventually be deployed on cloud infrastructures.

DEEP GPU deployment demo

Simple Jupyter deployment through the INDIGO Orchestrator

Álvaro López García
aloga@ifca.unican.es
Spanish National Research Council

DEEP-Hybrid-DataCloud has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 777435.

Visual Composition of TOSCA Templates with Alien4Cloud: JupyterHub + Kubernetes

EOSC-HUB Week
Málaga, Spain
April 2018

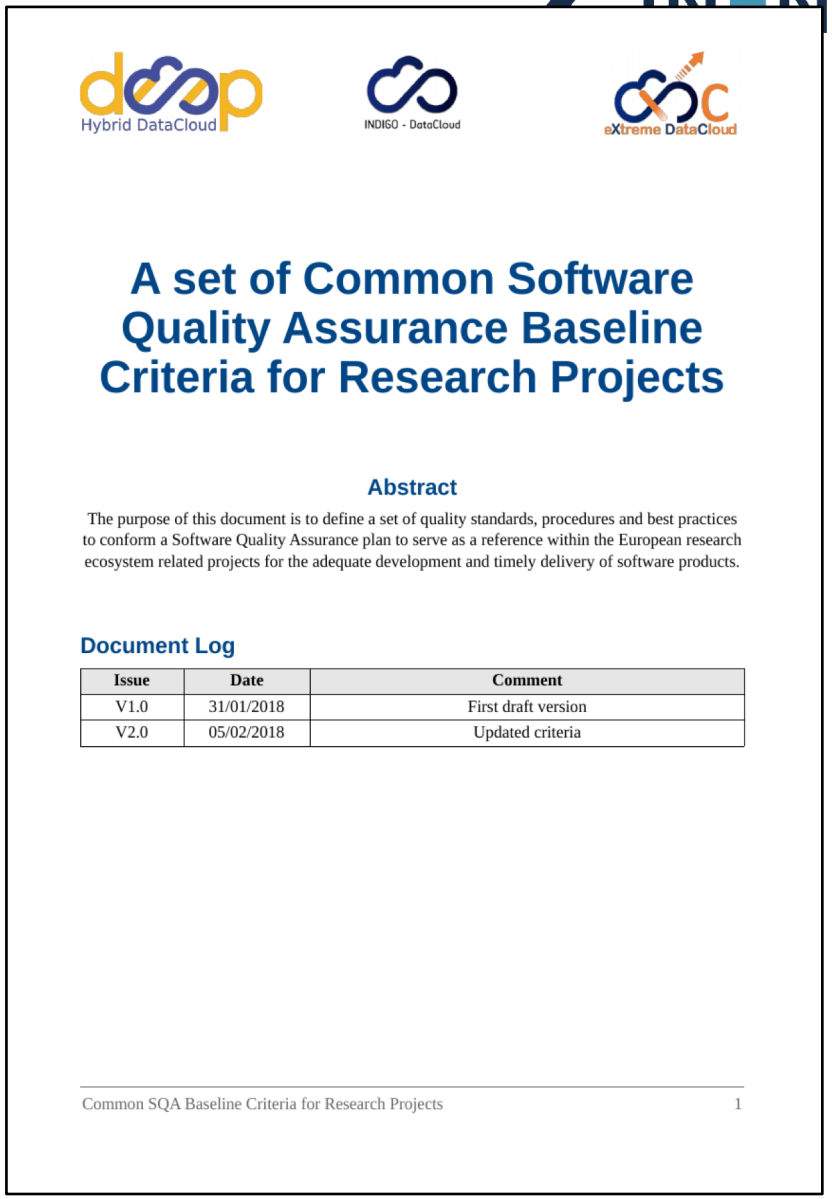
Germán Moltó, Miguel Caballer, Andy S. Alic
gmolto@dsic.upv.es, micafer1@upv.es, asalic@upv.es

Universitat Politècnica de València (UPV)

DEEP-Hybrid-DataCloud is funded by the Horizon 2020 Framework Programme of the European union under grant agreement number 777435

What about software quality?

- Remember that we said we don't want to reinvent the wheel… but how do we make sure that what we «invent» is solid enough for production use?
- INDIGO-DataCloud, eXtreme-DataCloud and DEEP-Hybrid DataCloud jointly delivered "**A set of common software quality assurance baseline criteria for research projects**", with the objective of delivering quality software
  - See http://hdl.handle.net/10261/160086
- Do not underestimate the importance of software QA!



Hybrid DataCloud · INDIGO - DataCloud · eXtreme DataCloud

A set of Common Software Quality Assurance Baseline Criteria for Research Projects

Abstract

The purpose of this document is to define a set of quality standards, procedures and best practices to conform a Software Quality Assurance plan to serve as a reference within the European research ecosystem related projects for the adequate development and timely delivery of software products.

Document Log

Issue	Date	Comment
V1.0	31/01/2018	First draft version
V2.0	05/02/2018	Updated criteria

Common SQA Baseline Criteria for Research Projects 1

The Elephant in the Room

- So far, I basically skipped one of the most important and most complex topics related to what we have been discussing.
- **What is it?**

Data Management Plans!

- If you don't know what a **Data Management Plan** is, I *strongly* advice you to go and look it up carefully.
  - See e.g. https://dmponline.dcc.ac.uk for some information and examples.
- From the all-powerful Wikipedia:
  - A data management plan or DMP is a formal document that outlines how data are to be handled both during a research project, and after the project is completed. The goal of a data management plan is to consider the many aspects of data management, metadata generation, data preservation, and analysis before the project begins; this ensures that data are well-managed in the present, and prepared for preservation in the future.
- Consider at least:
  - Information about data and data format (describe your data, how it will be acquired, when and where, how it will be processed, etc.)
  - Metadata content and format
  - Policies for access, sharing, re-use – can these policies change over time? How?
  - Long-term storage and data management
  - Budget!

Data management?

- This would require a *substantial* time to even just list the core issues, but I encourage you to think for example at the following topics:
  - How can you automate dataset distribution according to some «intelligence»?
  - Quality of service is important. Would you prefer to have fast or slow access to your data? How can you choose between the two
    - Hint: see what e.g. Amazon does with its Glacier vs S3 offerings
  - Are data access patterns important?
  - How can you access data that can sit in [perhaps multiple] Clouds transparently from a “client”? What if some Cloud sites offer object storage and other Posix storage?
  - Can you do some sort of “smart caching”?
    - Hint: think at what happens when you want to view a YouTube video.
  - How do you manage metadata?
  - What about sensitive data handling? Secure storage and encryption?

Conclusions

- We all appreciate how powerful can ML/DL techniques be. But we want also ways to effectively and efficiently get, configure and use resources for our ML/DL problems.
- Cloud-based resources are pervasive today, and *can* be extremely effective in reducing time and money to get the results we want.
- But in order to do so, without being locked in to proprietary, closed solutions, and without reinventing too many wheels, we still need some effort – so that we can streamline our experience.
- These are exciting times to be involved in this field! So, get involved and make a difference! (*)

(*) See also the talk on "Collaboration opportunities" on Friday

Thank you

“Stat rosa pristina nomine, nomina nuda tenemus”

Backup Slides

The INDIGO-DataCloud Consortium Members

- Software developers
- Industrial partners
- Research institutes
- Universities
- e-infrastructures
- Scientific communities

The European Open Science Cloud (EOSC)

- The EOSC: "**a model for the use of a cloud in the private and public sectors**" (European Parliament resolution on the European Cloud Initiative, Feb 2017)
- **Why the EOSC?**
  - To facilitate scientific developments and make the EU a center for global research
  - EO**S**C: "S" as in "Science", but with its user base to be extended to industry and governments
  - To foster the growth of the European Digital Economy → competiveness, global market positioning (esp. for SMEs)
  - To accelerate work on standards & interoperability, sharing of open data, creation of an open environment for storing, sharing and re-using scientific data and results, with the overall goal of removing fragmentation.

From the EOSC vision to action: the EOSC-hub project

- **The EOSC-hub project**, funded with about 30M€, mobilizes providers from 20 major digital infrastructures, EGI, EUDAT CDI and INDIGO-DataCloud, jointly offering services, software and data for advanced data-driven research and innovation through a **unified service catalogue.**
- Some facts:
  - 100 Partners, 75 funded beneficiaries
  - 3874 PMs, 108 FTEs, more than 150 technical and scientific staff involved
  - 36 months: Jan 2018 – Dec 2020

Some of the EOSC-hub services

Federation and Collaborative Services

- Identification, Authentication, Authorization and Attribute Management (EUDAT B2ACCESS, EGI CheckIn, INDIGO WaTTS)
- Marketplace and Order Management (Service Portfolio Management Tool, Data Project Management Tool)
- Integrated Business and Operations Support Systems (Configuration Management Data Base [GOCDB], EGI Operations Portal)
- Monitoring, Accounting, Messaging, Security Tools (ARGO, EGI Security monitoring tool, EUDAT accounting repository, APEL)
- Helpdesk Services and Tools (EGI & EUDAT)
- Application store, Software Repositories and other Collaboration Tools

Common Services: Integration and Maintenance

- Discovery and Access (INDIGO IAM, Onedata, CDMI storage services; EGI DataHub; EUDAT B2SHARE, B2FIND, B2STAGE, B2HANDLE, B2DROP; OpenStack Swift [external])
- Federated Compute (INDIGO advanced IAAS services [Tosca Heat translator, Cloud spot instances, udocker]; CREAM-CE, BDII, VOMS, STORM, CVMFS, ARGUS, DIRAC)
- Processing and Orchestration (INDIGO PaaS [TOSCA-based templates], FutureGateway, Open Mobile Toolkit)
- Data and Metadata management (EUDAT B2NOTE, B2SAFE, B2HANDLE)
- Preservation (Trusted Digital Repository [by DANS-KNAW])
- Sensitive Data (TSD [by SIGMA2], ePouta [by CSC])

The EOSC-hub Thematic Services and Competence Centers

INFN

Thematic Services

- **ECAS** – Climate Analytics Service (ECAS), provided by ENES
- **DARIAH** – Science gateway tailored for the digital arts and humanities communities
- **OPENCoastS** – On-demand Operational Coastal Circulation Forecast Service
- **GEOSS** – GEO DAB (Discovery and Access Broker), GEOSS portal
- **EO Pillar** – Earth observation services, coordinated by ESA
- **WeNMR** – Structural biology services
- **DODAS** – Dynamic On Demand Analysis Service (CMS & others)
- **LifeWatch** – Citizen science services, GBIF, Digital Knowledge preservation framework, remote monitoring and smart sensing
- **CMI** – The Component MetaData Infrastructure, including the Virtual Language Observatory and the Virtual Collection Registry, provided by CLARIN

Competence Centers

- **IFREMER**: unified data analytics platform for the marine science community wrt climatology and oceanography
- **LOFAR**: online platform for radioastronomy data storage and data analysis
- **Fusion/ITER**: Platform data storage and simulation & modelling for open data in fusion research
- **European Integrated Data Archive framework** (EIDA): tools for seismological data and services
- **ELIXIR**: dataset distribution service and tools for life science
- **EISCAT_3D**: data analysis for atmosphere and near-Earth space studies

INDIGO-DataCloud in the EOSC-hub Thematic Services

THEMATIC SERVICE	DESCRIPTION	INDIGO SOLUTIONS specified in the proposal
ECAS	Climate Analytics Service (ECAS) provided by ENES	Ophidia, KEPLER, FutureGateway
DARIAH SG	DARIAH science gateway tailored for the digital arts and humanities communities	OneDock, OpenStack Nova Docker, FutureGateway, Onedata
OPENCoastS	OpenCoastS: On-demand Operational Coastal Circulation Forecast Service	INDIGO udocker Infrastructure Manager Orchestration (TOSCA, HEAT)
WeNMR	Structural biology services DISVIS, POWERFIT, HADDOCK, GROMACS, AMPS-NMR, CS-ROSETTA, UNIO, FANTEN	IAM, PaaS Orchestrator, Infrastructure Manager FutureGateway, Onedata
DODAS	Dynamic On Demand Analysis Service	IAM, TTS, PaaS Orchestrator, Orchent, IM, TOSCA, Onedata, FutureGateway
LifeWatch	PAIRQURS, Citizen science services, GBIF, Digital Knowledge preservation framework, remote monitoring and smart sensing.	IAM, PaaS Orchestrator, Infrastructure Manager FutureGateway, Onedata
EO Pillar	Earth observation services coordinated by ESA. The tools are: MEA, EPOSAR, Sentinel playground, Datacube analytic service, Geohazards exploitation platform,OSS-X Sentinel service	Onedata. Analysis of INDIGO cloud software add-ons for OpenStack.
CMI	The Component MetaData Infrastructure Including the Virtual Language Observatory and the Virtual Collection Registry, provided by CLARIN	Providing interoperable metadata for (digital) humanities between both CLARIN and DARIAH.
GEOSS	GEO DAB (Discovery and Access Broker) GEOSS portal	-

INDIGO - DataCloud