Data Science Research Infrastructure

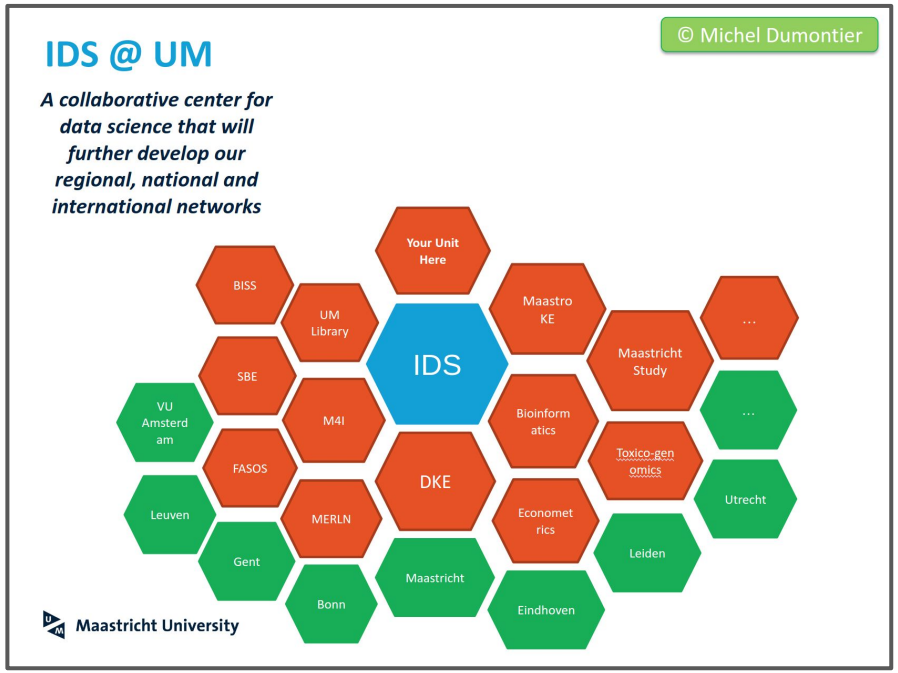
University Meeting Eindhoven - July 2nd 2019

Alexander Malic - Institute Of Data Science @ Maastricht University

Partners

Introduction

- Self taught university drop out
- 18 years of professional experience (8 years consulting, 10 years in Research-IT at multinational company)
- Simple solutions for complex problems
- Automate repetitive tasks both for myself and my customers (internal/external)
- Shared vision with Michel

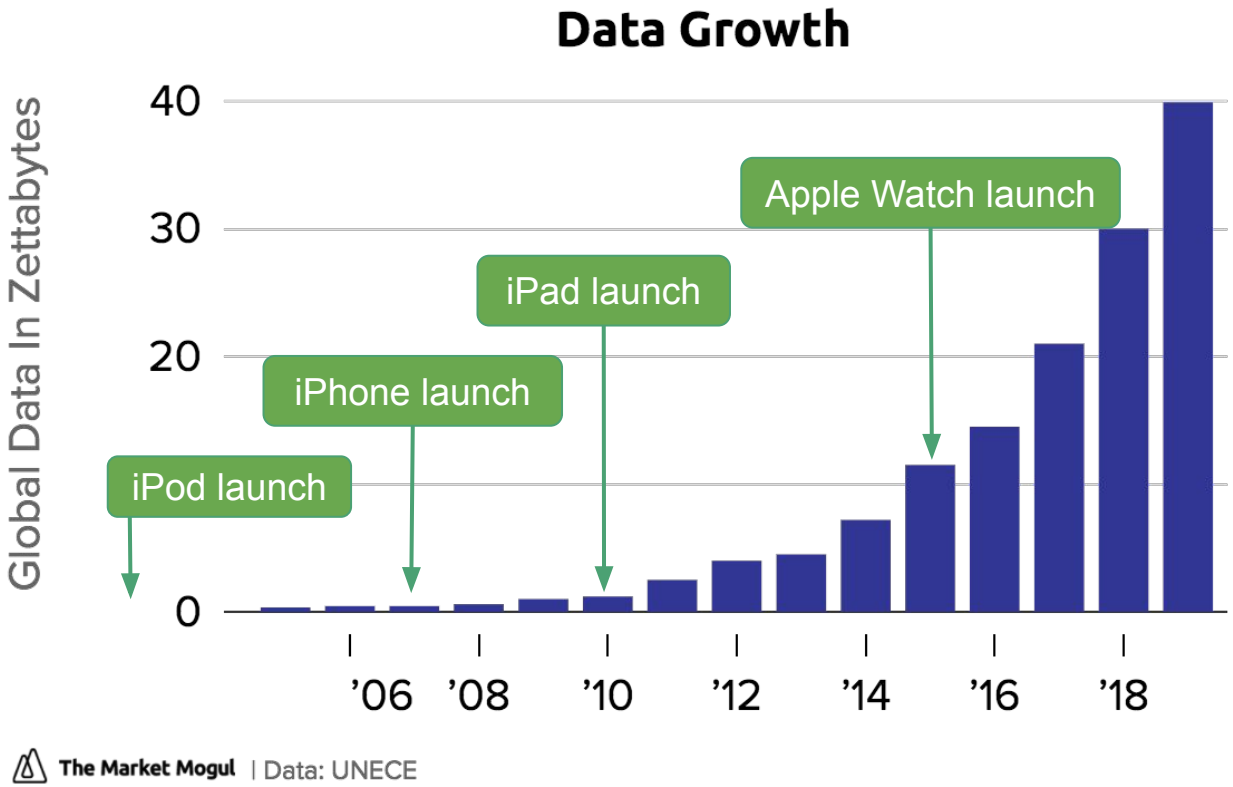


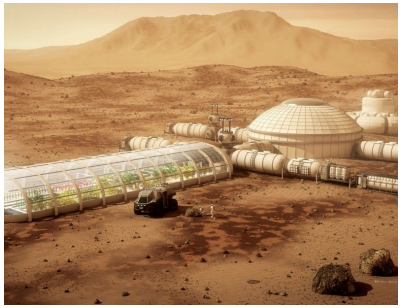
How does Google do it?

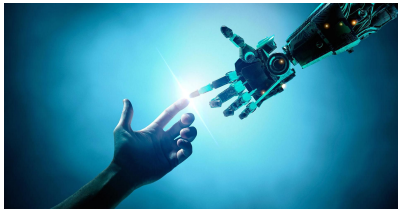
- Scalable & distributed File-System:
  → The Google File system [1]
  → Apache Hadoop
- Scalable & distributed data-base:
  → Google BigTable [2]
  → Apache HBase database
- Scalable computing:
  → Google Borg [3]
  → Docker/Kubernetes

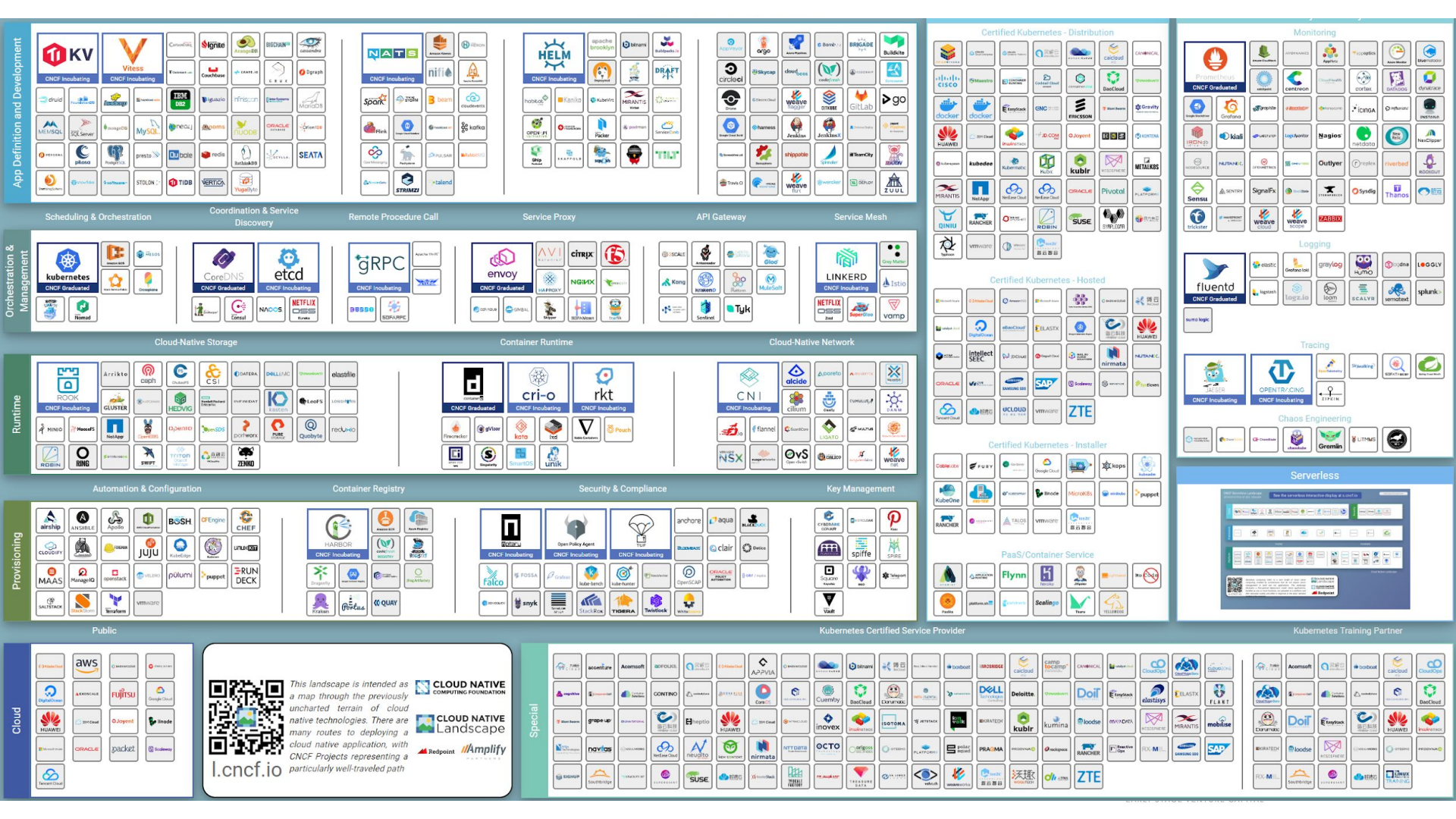
[1] The Google File System - https://ai.google/research/pubs/pub51
[2] Bigtable: A Distributed Storage System for Structured Data - https://ai.google/research/pubs/pub27898
[3] Large-scale cluster management at Google with Borg - https://ai.google/research/pubs/pub43438

Data growth problem







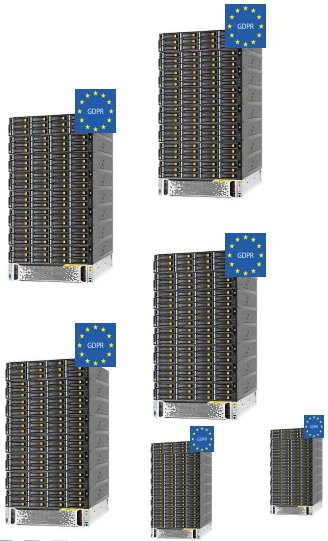

App Definition and Development
Scheduling & Orchestration
Coordination & Service Discovery
Remote Procedure Call
Service Proxy
API Gateway
Service Mesh
Orchestration & Management
Cloud-Native Storage
Container Runtime
Cloud-Native Network
Runtime
Automation & Configuration
Container Registry
Security & Compliance
Key Management
Provisioning
Public
Certified Kubernetes - Distribution
Certified Kubernetes - Hosted
Certified Kubernetes - Installer
PaaS/Container Service
Kubernetes Certified Service Provider
Monitoring
Logging
Tracing
Chaos Engineering
Kubernetes Training Partner
Serverless
Cloud
Special
l.cncf.io
This landscape is intended as a map through the previously uncharted terrain of cloud native technologies. There are many routes to deploying a cloud native application, with CNCF Projects representing a particularly well-traveled path
CLOUD NATIVE COMPUTING FOUNDATION
CLOUD NATIVE Landscape
Redpoint
Amplify

Flexible computing

Current

Data

Results



CDDI Project

State of the art

Code

Code

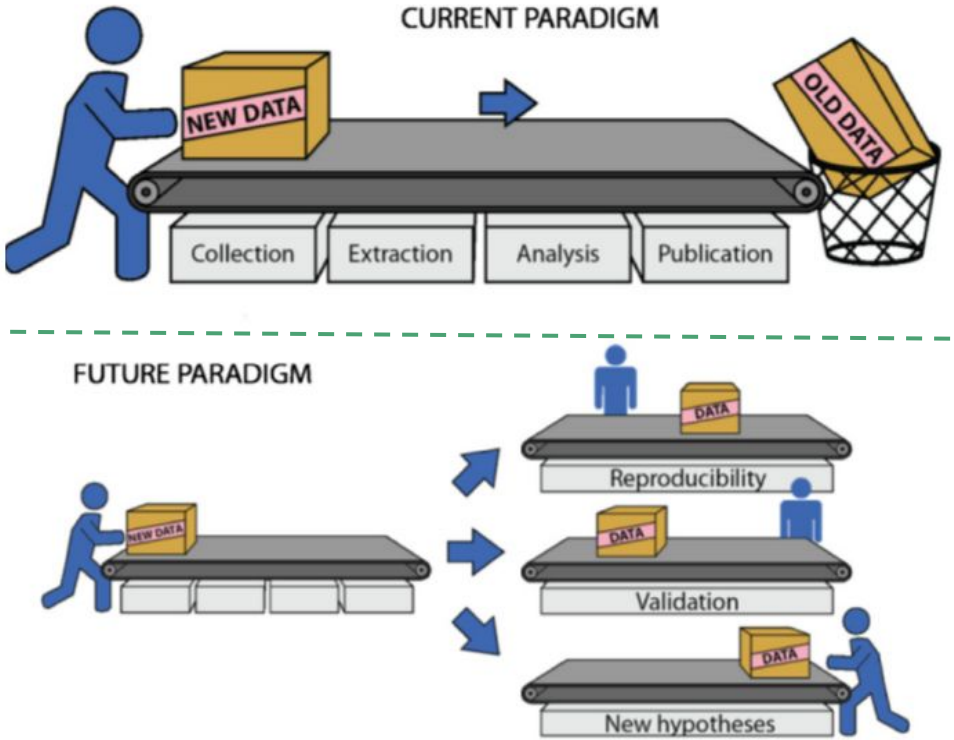
Results & Data

Reproducible research

Current

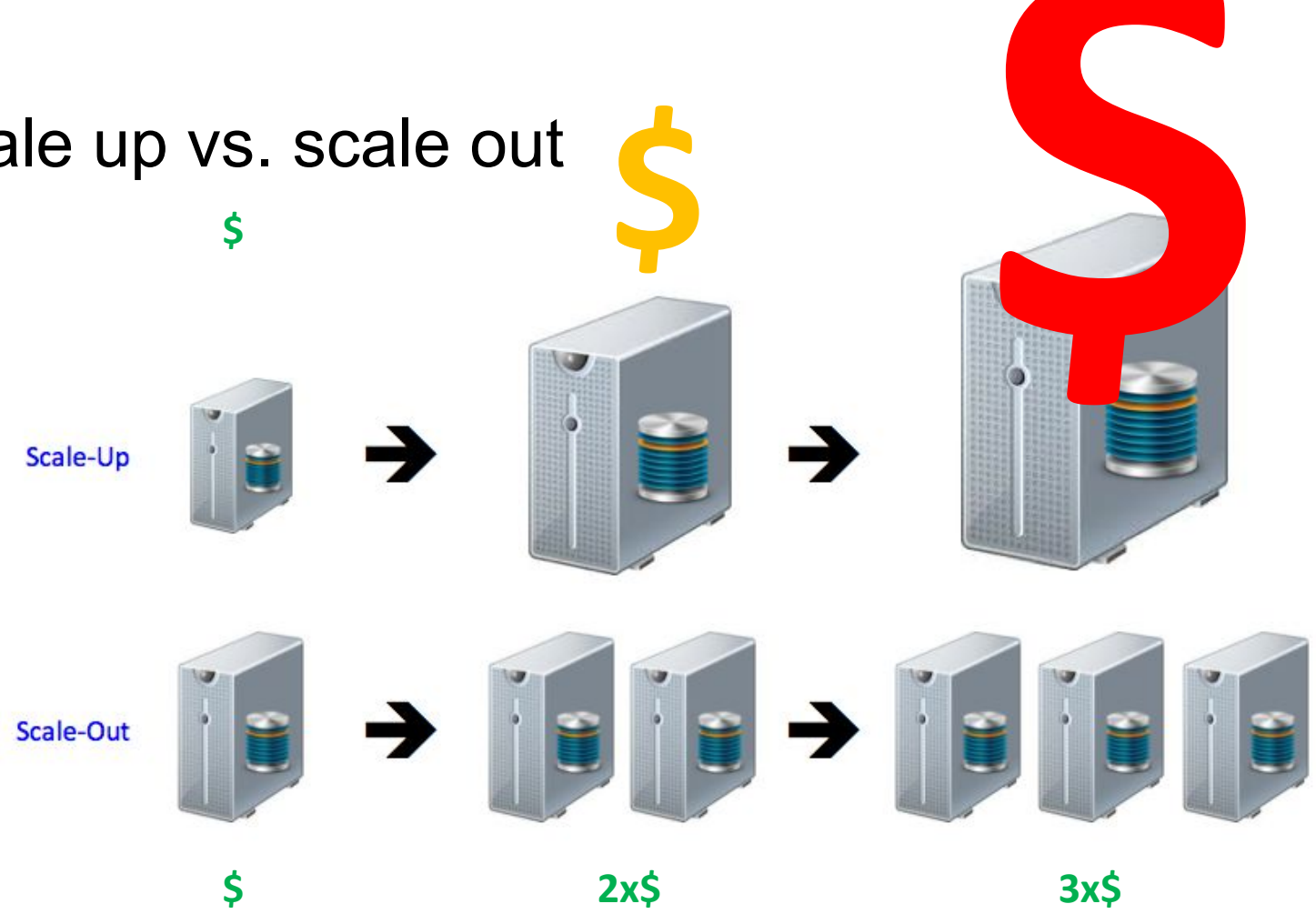
CDDI Project

State of the art



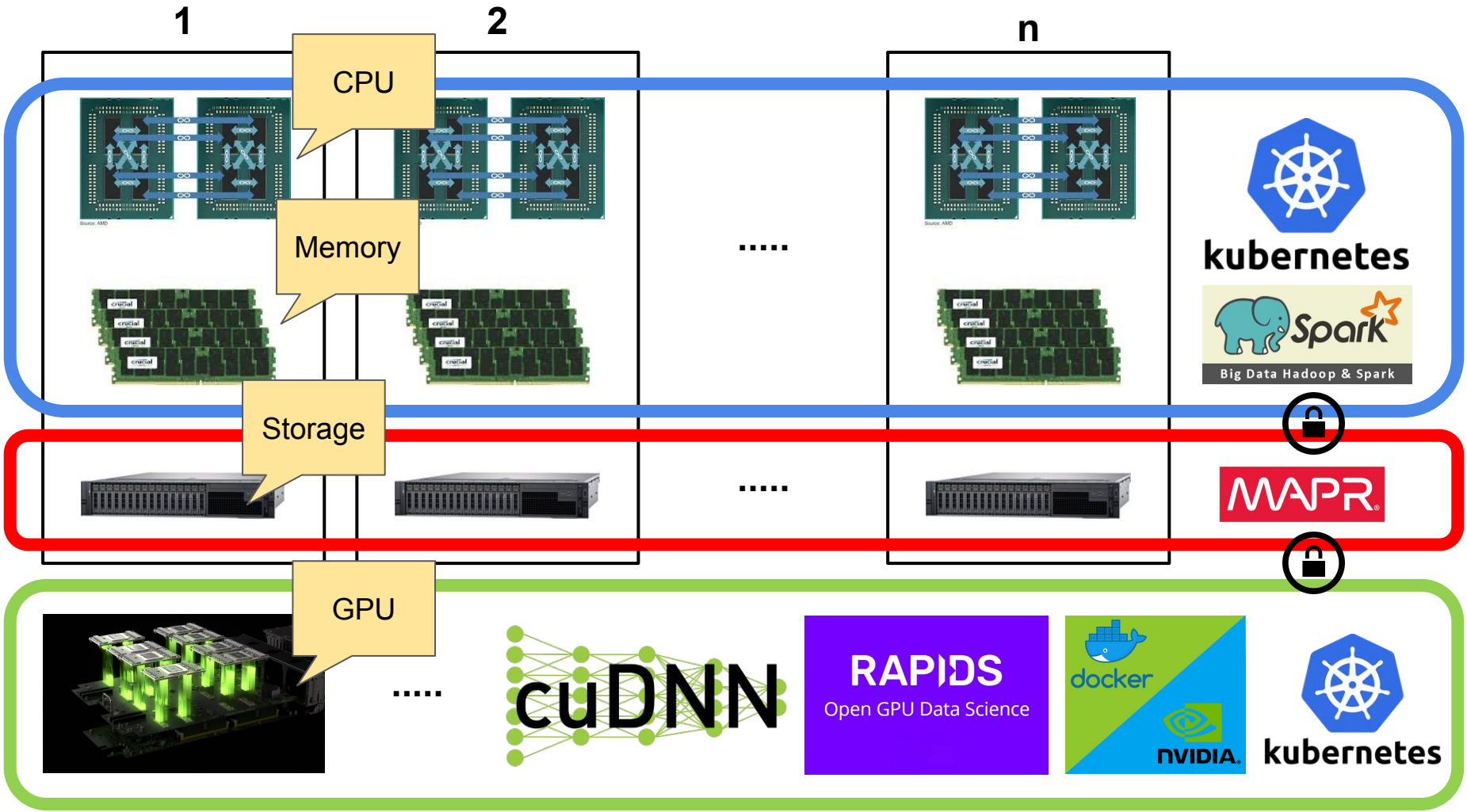
2013 - Philippe Lambin - https://doi.org/10.1016/j.radonc.2013.07.007

Scale up vs. scale out



What is DSRI?

- Infrastructure
  - shared pool of resources
  - high availability
  - collaborative
- Methods
  - reusable components (algorithms & code, data)
  - reusable data
  - reproducible methods
- Partnerships
  - Win-Win situation based
  - Dell, NVIDIA, AMD and MapR



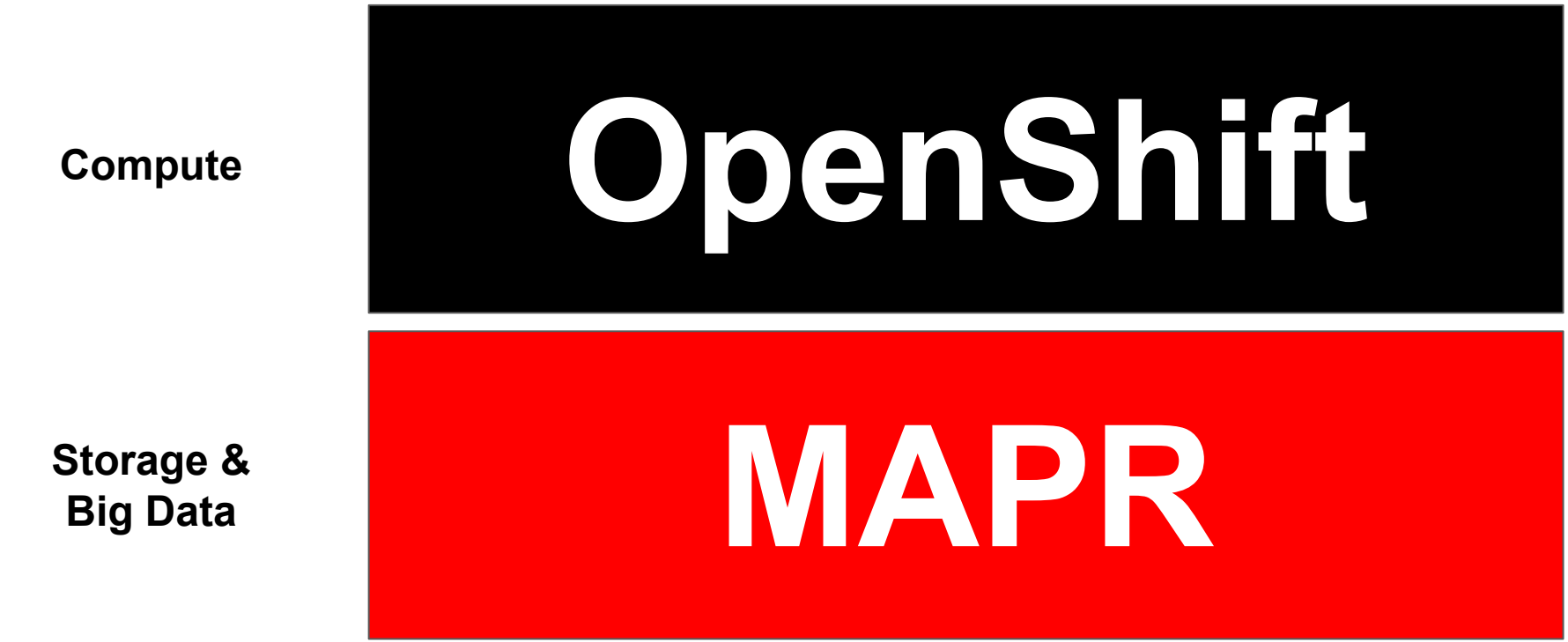
DSRI project

Performance indicators - numbers

	Base	Expanded
64 cpu cores per server	576 CPU cores	**960 CPU cores**
128 simultaneous threads per server	1152 threads	**1920 threads**
512 GB per server	4608 GB	**7680 GB**
120 TB net storage per server	1080 TB - 1350 TB*	**1800 TB - 2250 TB***
DGX 1 - (8x Tesla V100 - 32GB GPU)	**1 x NVIDIA DGX 1**	

* estimated because of MAPR file system compression (assumed compression factor 2 - 2.5)

MAPR & OKD (Origin Kubernetes Distribution)



Why MAPR?

- Multi-Tenant
- High Available
- Scalable
- Secure
- Global Namespace
- Multi Location
- Multi Temperature
- Kubernetes integration
- NFS
- Open Source compliant

Why OpenShift?

- Multi Tenant
- Open Source version of OpenShift
- Management interface that enables self service
- Kubernetes and Docker

What can we do with it?

- IT view
  - Multiple servers running the same software
  - High available (20% of the cluster can be out of service, without interruptions or data loss)
  - Flexible (rolling upgrades, automated failover)
- Research view
  - Scalable (Hardware contributions through grants and investments)
  - Low maintenance (Managed through Research and IT)
  - Easy to use (Containers, Methods, Best practices and blueprints, support)
  - Easy to contribute
  - Easy to share and reproduce

CDDI pilot project - 500 TB MRI data

CDDI Pilot Project - Auditory Cortex

Alexander Malic

Pilot project charter:

The goal of this project is to implement a Research data workflow which enables Researchers to share MRI images with other Researchers across and outside Maastricht University through a platform called XNAT.

Scope:

In scope:

- Focus only on FPN MRI images for this pilot
- Set up a XNAT platform for sharing MRI data
- Develop a process which converts proprietary DICOM data images into open BIDS data

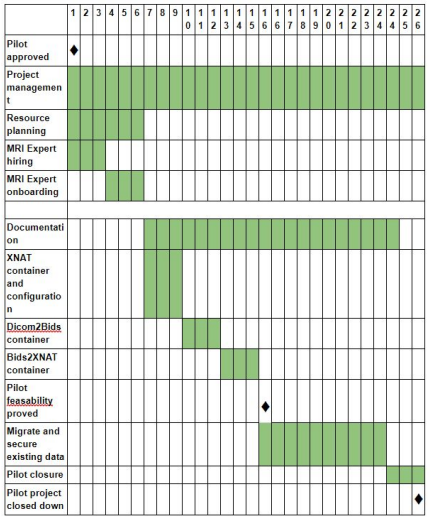
Out of scope:

- Fully functional XNAT workflow for all UM (Idea is that extending the use case should be easily possible but might require changes and additional ethical approvals.

Requirements:

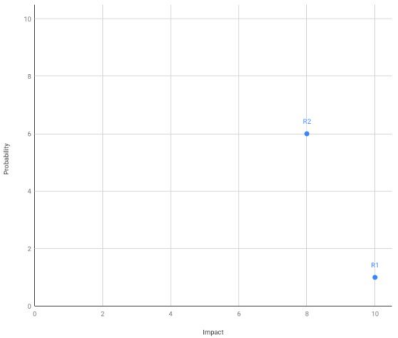
- DSRI Infrastructure in place
- 0.5 FTE for 6 months for Data engineer with MRI and XNAT expertise and ideally already familiar with BIDS data format (Jan 2019 to June 2019 → ~128 effective working days → 0.5 * 128 * 8 = ~512hr)
- Approval and steering by CDDI

Timelines:

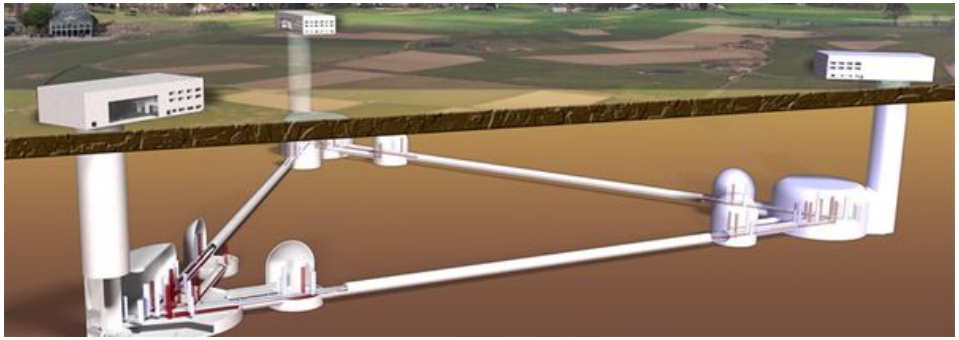


Risks

Risk	Impact	Probability	Description
R1	10	1	DSRI infrastructure not in place. Project will not start until infrastructure is available. Work on this pilot can start as soon as DSRI is online and does not need to wait for the DSRI project to be finished.
R2	8	6	MRI Expert with XNAT knowledge not available. If a MRI expert is not available for hiring, the project team will look into building up a colleague who is on PostDoc level. Project timelines need to be reevaluated in such a case.



Einstein Telescope on DSRI as IoT device



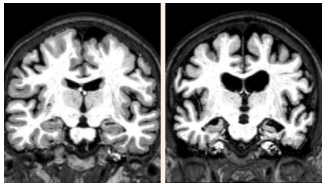
Edge Cluster A | Edge Cluster B | Edge Cluster C

Topic 1 / Topic 2 / Topic 3 → Stream

DSRI
HA storage / scalable CPU & GPU

BReIN project
Maastricht - Backup Facility - FZ Jülich

Randwyck



MRI data

```
@read1
AGCTTATCCTCTGCTCACCCCCGGGTTAGCGCACTTGATGTATTCACAGC
+
BA1@CC7CBCCC9C8;B2@>C?B@B@B3=9?@B1:AB7B?B8B?B6B.7.
@read2
TTGGGCGGGATCTCCAGAAGCATATGGATGTGATCCACACAGCATTCTGC
+
?>?B@)<?@,AA7A@C<C?=@@B;+)?B5*@2=@+=BB,=B6C>AB@B24
@read3
TATGCTCAAGAAGGGGCTGATGAGTTGGTGTTTTACGATATCACTGCCTC
+
A3AB:B1:B;9/0BBBCBB<BB@AA0?BB9:BB<A@BB@7@6@<A@@@<3
```

Sequencing data

	A	B	C	D	E
1	**Use In Big Data analysis?**	**Study_ID**	**Location**	**Dataset**	**Project**
2	Yes	Study01	dIXa data warehouse	DIXA-002	CarcinoGenomics
3	Yes	Study02	dIXa data warehouse	DIXA-002	CarcinoGenomics
4	Yes	Study03	dIXa data warehouse	DIXA-002	CarcinoGenomics
5	Yes	Study04	dIXa data warehouse	DIXA-002	CarcinoGenomics
6	*No*				*CarcinoGenomics_blinc*
7	Yes	Study05	/ngs-data/data_storage/transcriptomics/m		CarcinoGenomics_extr
8	Yes	Study06	ftp://ftp.biosciencedbc.jp/ar	DIXA-006	TG-GATEs
9	Yes	Study07	/ngs-data/data_storage/tran	GSE28878	STW_Magkoufopoulou
10	Yes	Study08	dIXa data warehouse	DIXA-028	Predictomics
11	Yes	Study09	dIXa data warehouse	DIXA-028	Predictomics
12	Yes	Study10	dIXa data warehouse	DIXA-029	Esnats
13	Yes	Study11	/ngs-data/data_storage/tran	DIXA-083	NTC_WP4.1.3_E02
14	*No*			GSE51952	*NTC_Van den Hof*
15	Yes	Study12	/ngs-data/data_storage/tran	E-MEXP-2458	Aanjaag_Jennen
16	Yes	Study13	/ngs-data/data_storage/tran	GSE58235	Deferme
17	Yes	Study14	/ngs-data/data_storage/tran	GSE39291	Deferme
18					

Meta and other data

Daalhof

BReIN servers

Heerlen

Backup facility

Jülich

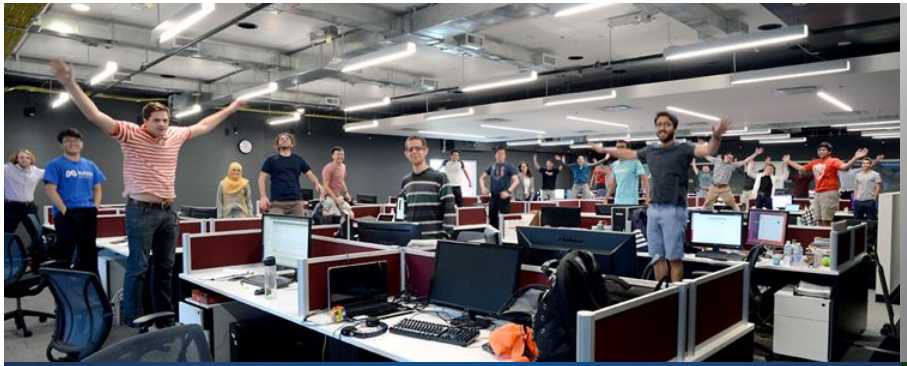
HPC

BReIN - server specifications

Item	Description	Quantity
Compute and Storage node	Dell PowerEdge 7425 servers, 2x 32 core AMD Epyc (2x 64 core w. Rome?), 512 GB - 1TB Ram, 15x 12TB HDD	16
GPU node	1x NVidia DGX 1 (8x NVidia Tesla-V100 32GB), 512GB Ram	1

Item	Description	Quantity
Backup storage node	Baseline: 45-drives Storinator XL60 (max specs), 60x 12TB HDD	5

5. Vision



Gitlab for DEVOPS → SCIOPS

- Made for DEVOPS by DEVOPS
- Case and Project management
- Continuous Integration platform

- Container Registry/Cache
- NPM & Maven Repository/Cache
- One-Click deployment of code into containers

Liferay

Horizontal Portal

Content Management

Collaboration and Social

DAM and EFSS

Mobile Experience Platform

Forms, Workflow, Business Process Automation

Developer Tools

Open edX overview:

Built to showcase the latest in learning sciences and instructional design, the Open edX learner platform is driven by our community of developers, technology partners, research teams, and users.

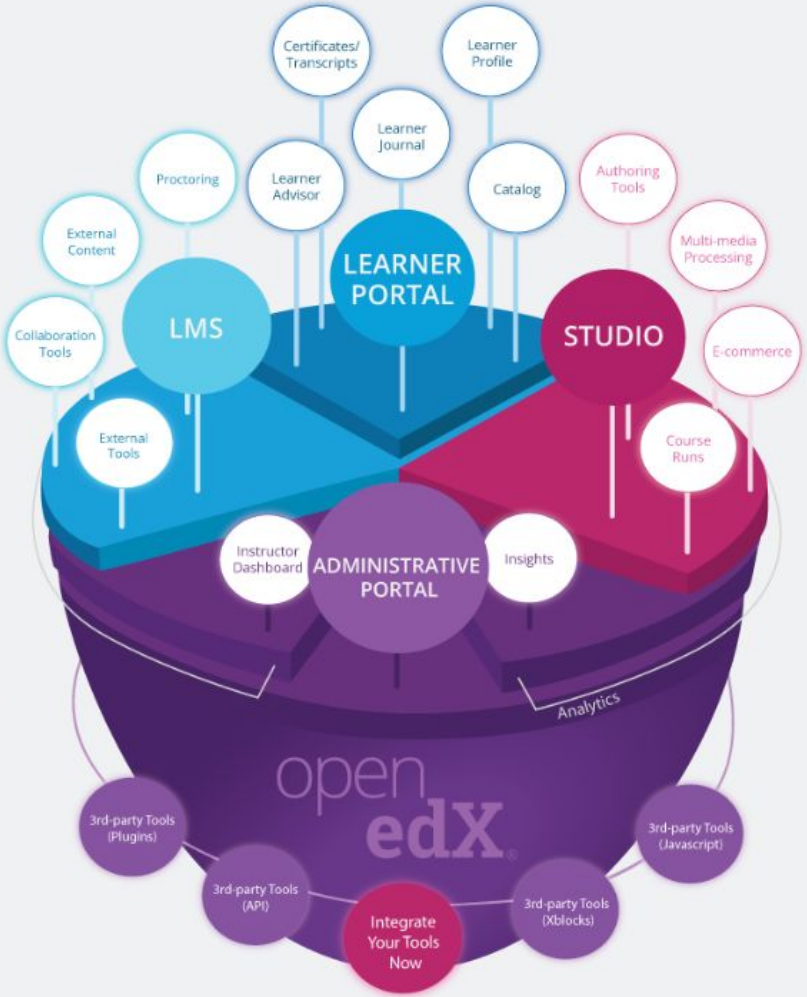
BUILT WITH

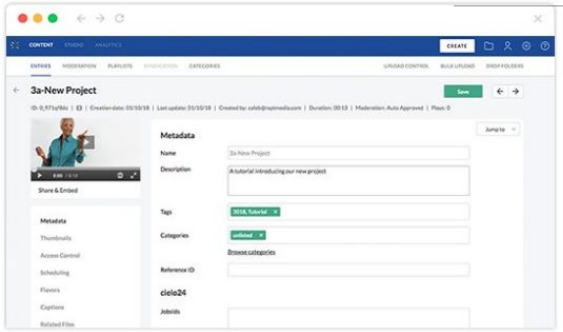
ubuntu
docker
django
React

RUNS ON

CLOUD
amazon web services
Google Cloud
IBM Cloud
Microsoft Azure
openstack.

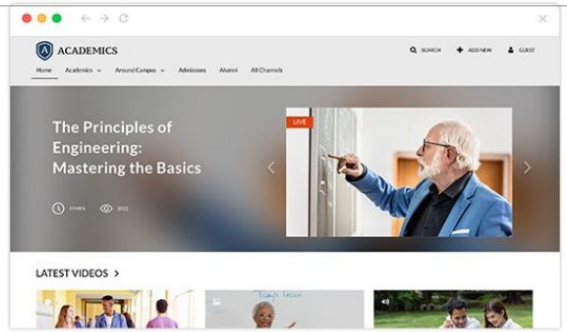
ON-PREM





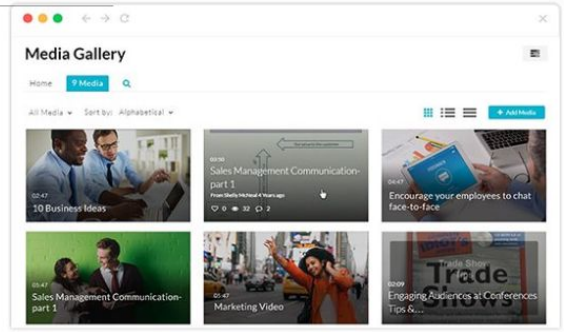
Kaltura Video Platform

Manage and publish your media through one intuitive interface



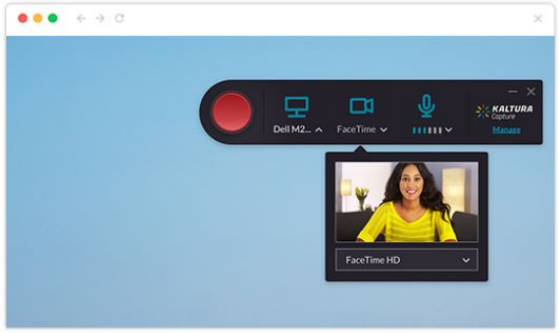
Kaltura MediaSpace Video Portal

Create your own private YouTube-like video portal



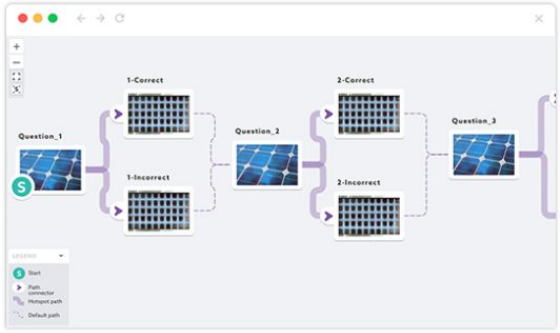
LMS Video Plugins

Add a full layer of video to your learning management system



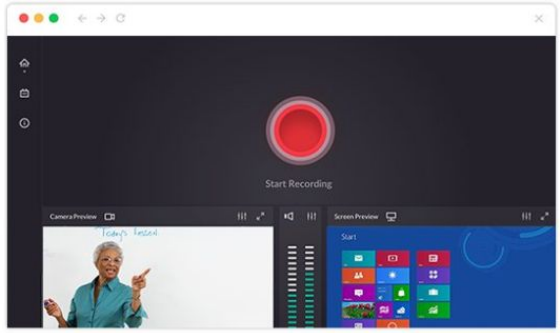
Personal Capture

Give your end users powerful tools to easily create engaging videos from their own desktops



Interactive Video

Create immersive video experiences for user engagement



Lecture Capture

Manage all your lecture captures from any recording device


Thank You
DSRI