Data Science & Society

The cures of tomorrow are in the data that you can't analyze today

October 23, 2018
Wienand Omta

UMC Utrecht

Utrecht University

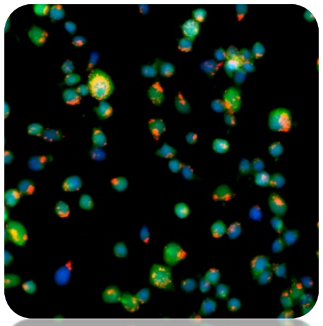
Core Life Analytics

Agenda Data Analytics Lecture

- Introduction High Content Screening
- Case study
- Machine Learning
- Scalable Cloud Architecture

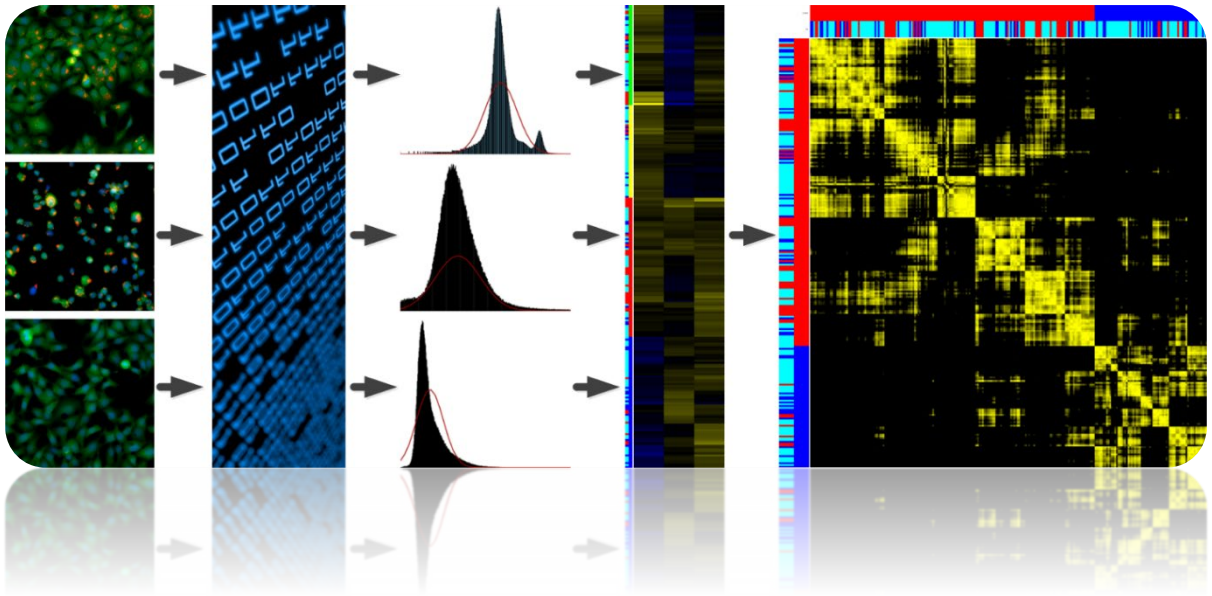
Who we are and what we do

- Core Life Analytics, Cell Screening / Core Dept. of Cell Biology, CMC
- Image based readout, automated microscopy
- Automated image analysis
- Data at single cell level, analyze subpopulations
- Multiple features
- Profiles vs. single numbers
- Complex phenotypes, unexpected phenotypes
- Insight into mechanism of action



Mission

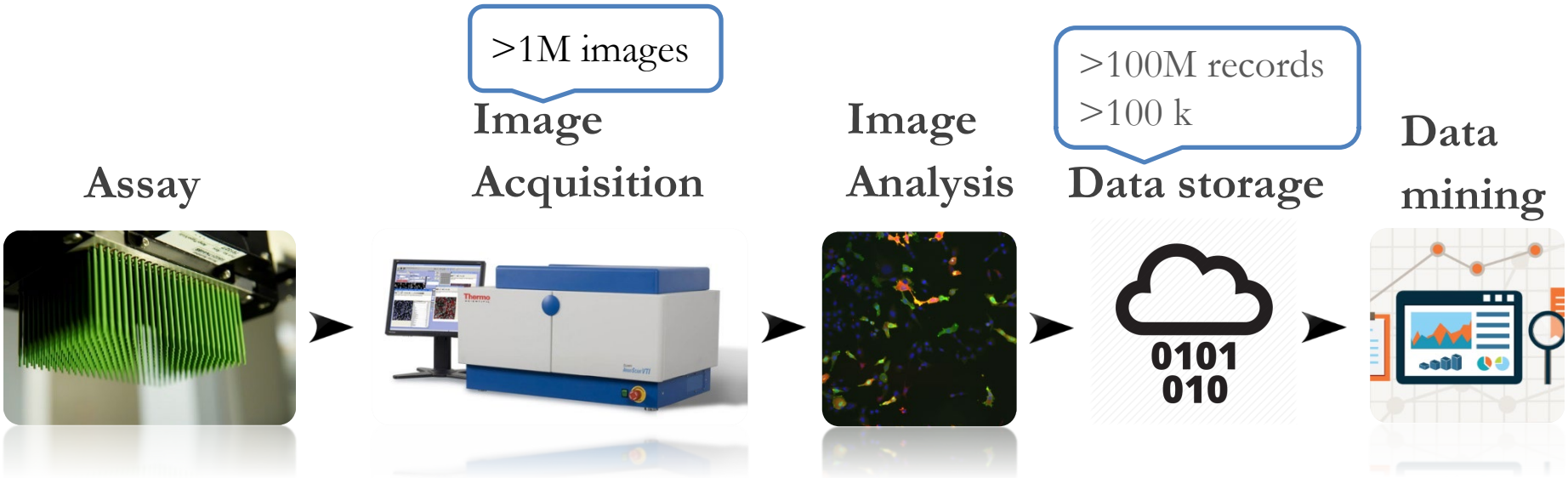
Provide scientists the ability to access technology for high throughput functional genomics and chemical biology experiments, with a focus on high content analysis.



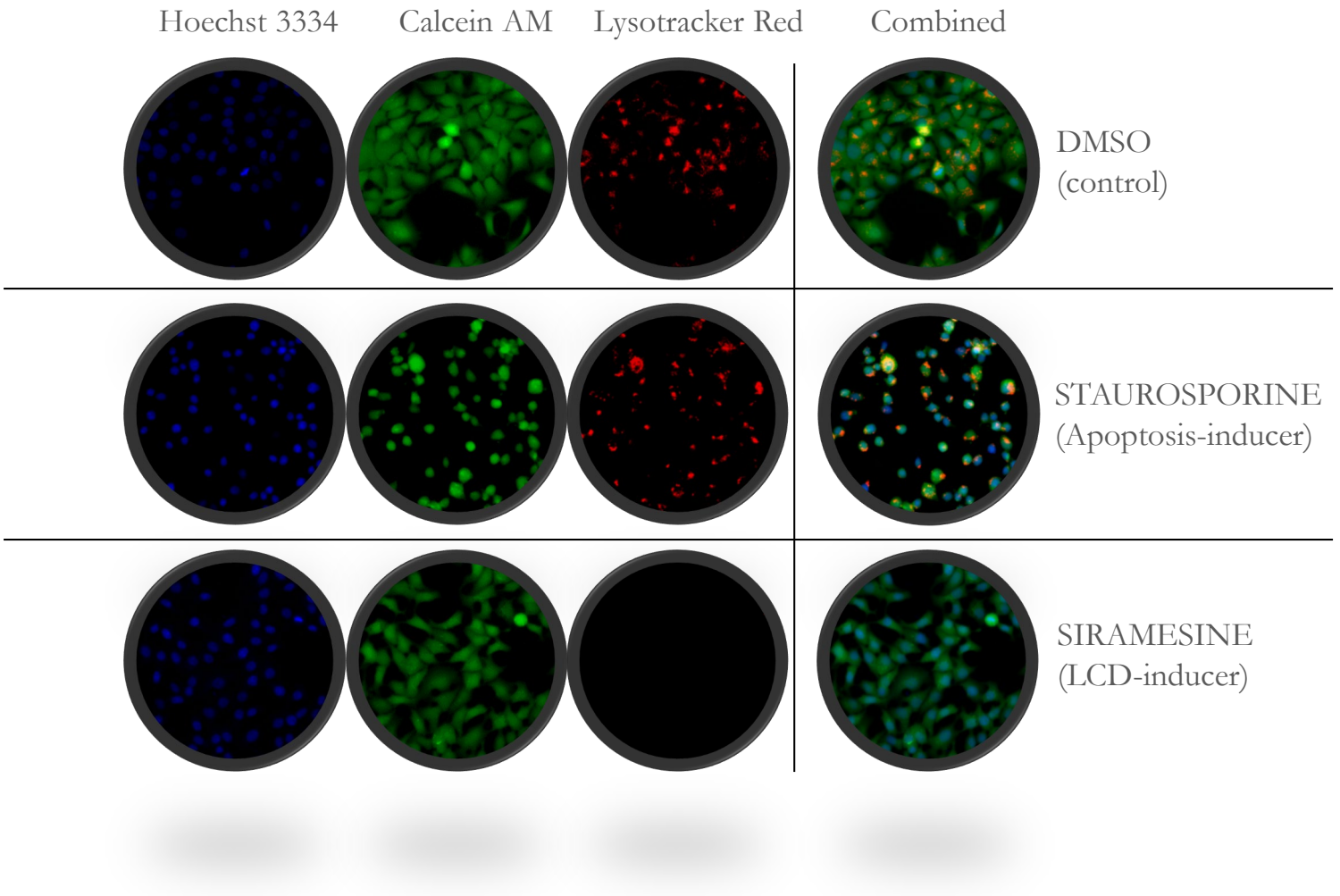
Introduction High Content Screening

What is HCS?

- A subset of High Throughput Screening (HTS)
- High Throughput Screening (HTS) is a process in which large libraries of chemical or biological reagents can be tested for activity in assays using automated methods (Allan et al., 2012)



The Power of High Content Screening



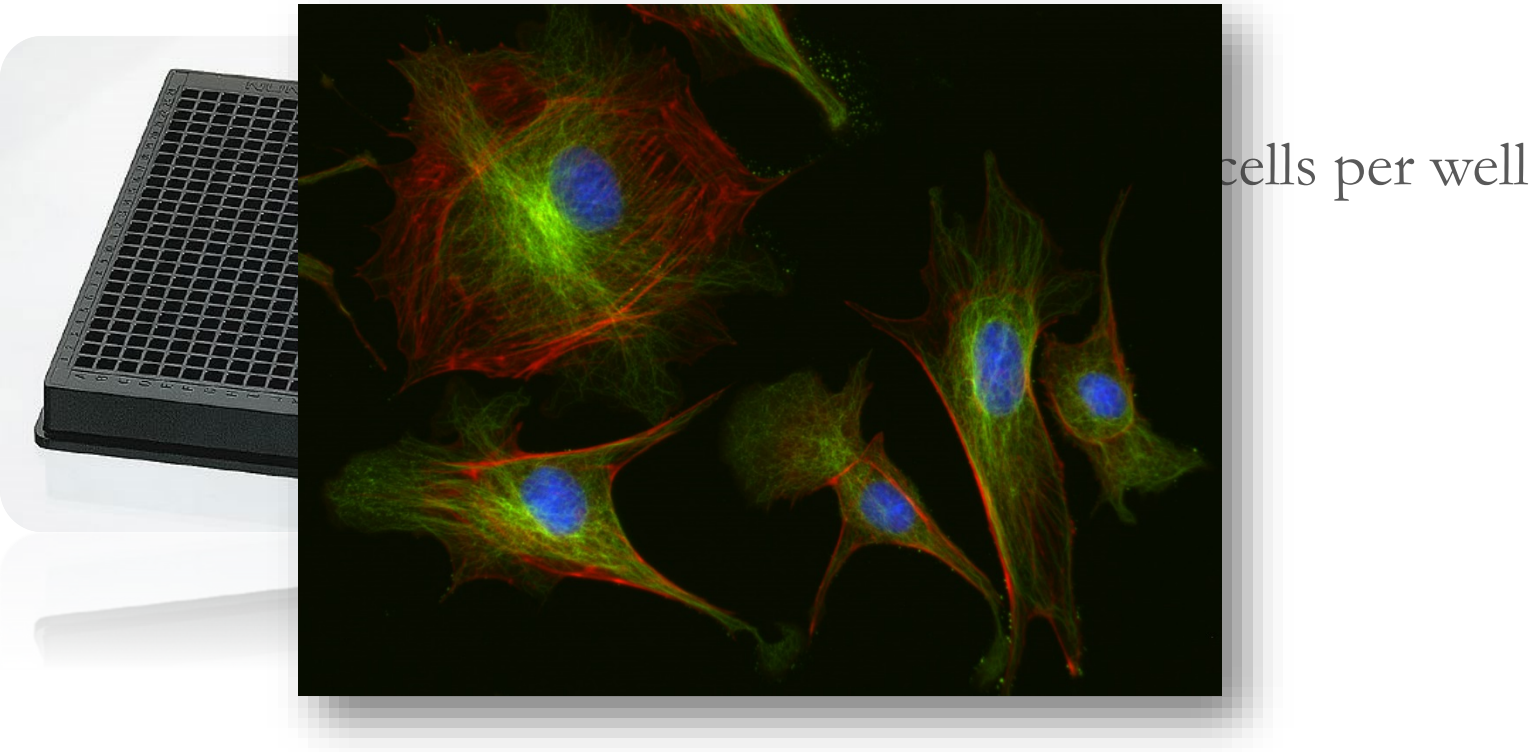
Case study
(Example High Content Screen)

UMC Utrecht

Utrecht University

Core Life Analytics

Problem (I)
Case study



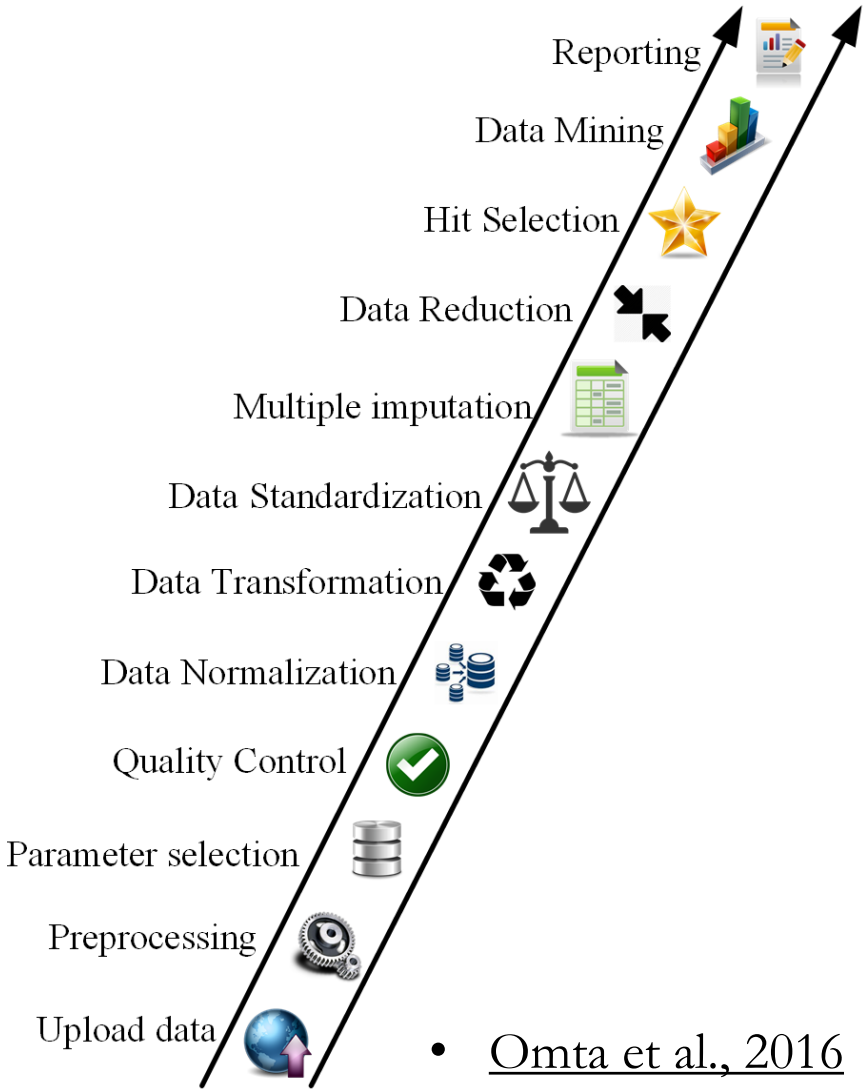
cells per well

Problem (II)
Case study

- 121 (SPECS) + 37 (FSM) + 4 (PRESTWICK) (162) plates of 384 wells in duplicates (326 plates)
- ~1000 cells per well
- 3 fluorescent labels, staining for DAPI, PY and GHR
- 258 variables/metrics per individual cell (Bioapplication CellHealthProfiling V4.)
- 162*384*2*1000, in total ~ 124 million records (cells) and 258 variables/metrics
  - 124 million* 258 ~ 32 billion fields
    ~ 146 GB numeric data
- ~50,000 chemicals/drugs/natural products + 10,500 controls (duplicates ~ 121,000 wells)

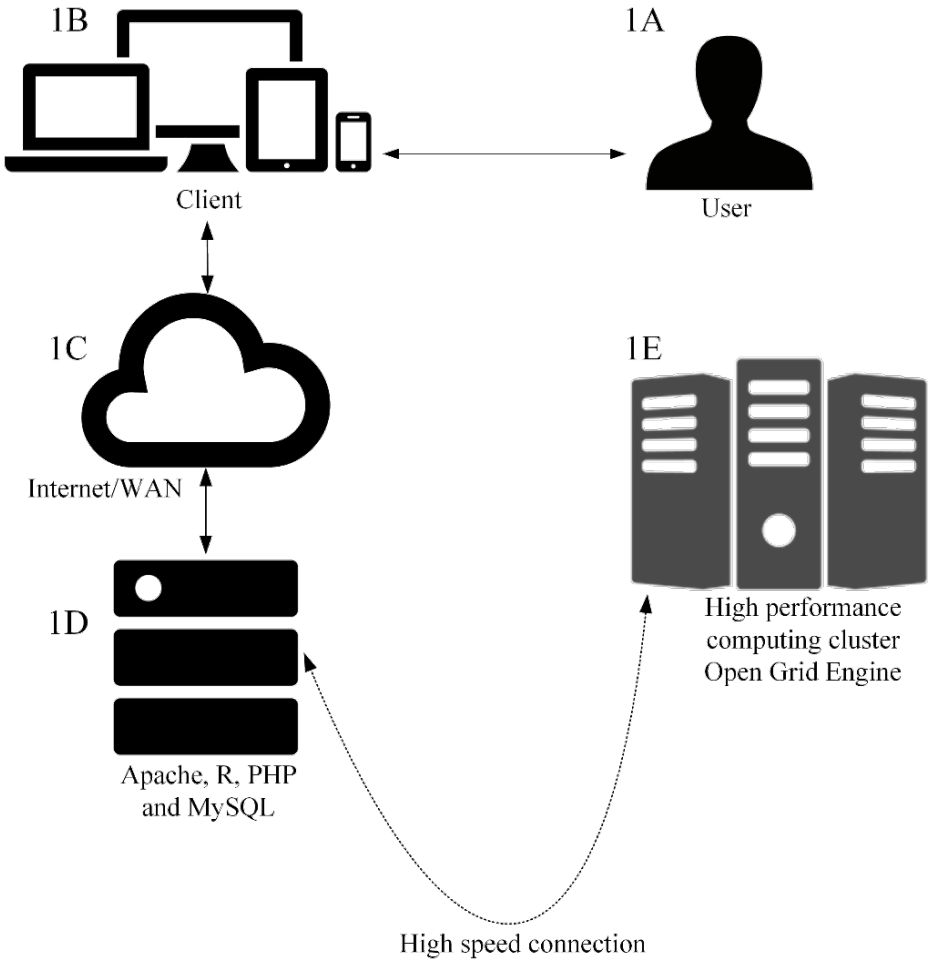
'Drowning in data and starving for knowledge'
Singh, Carpenter and Genovesio (2014)

HC StratoMineR workflow

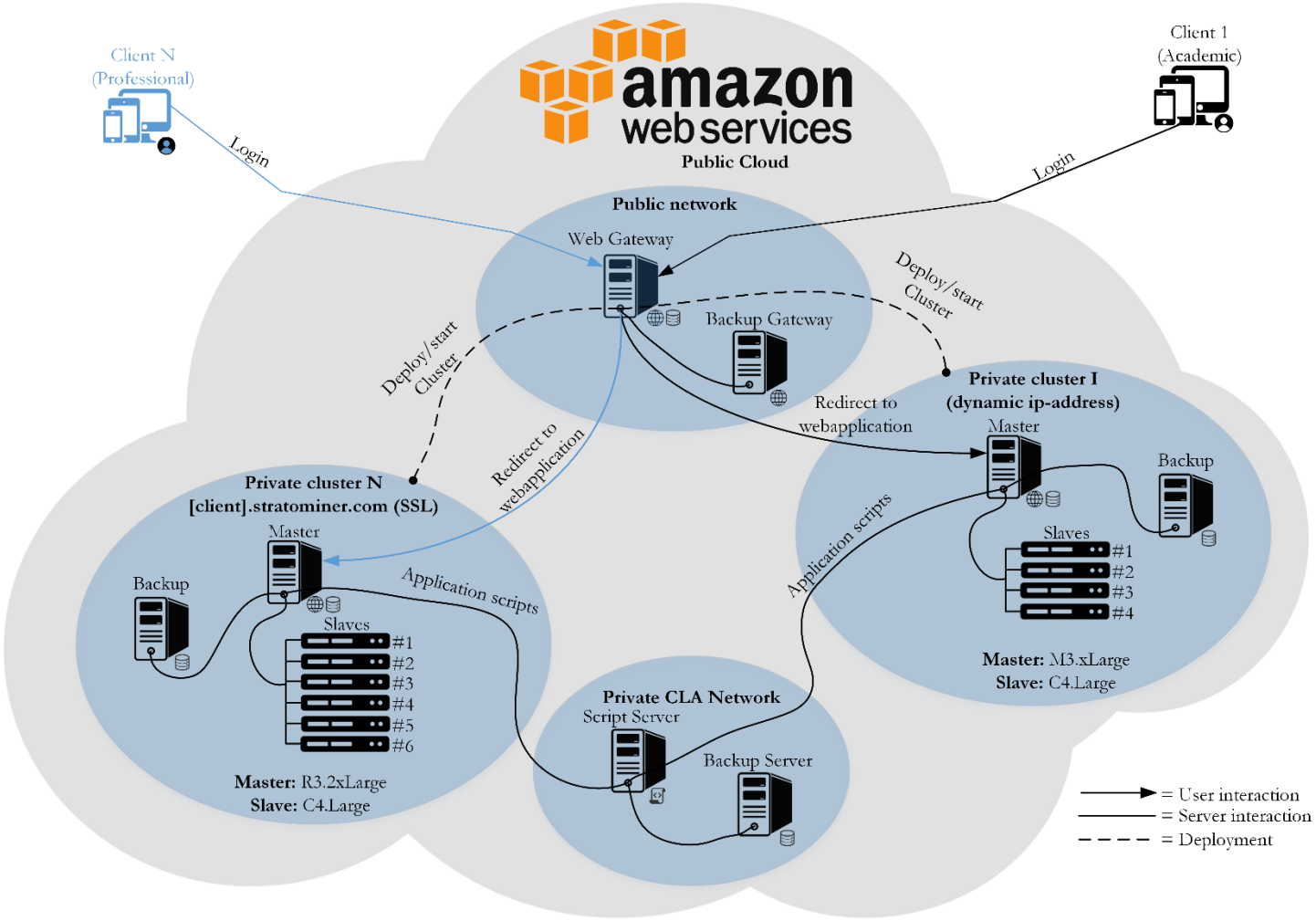


- Omta et al., 2016

Academic HC StratoMineR architecture

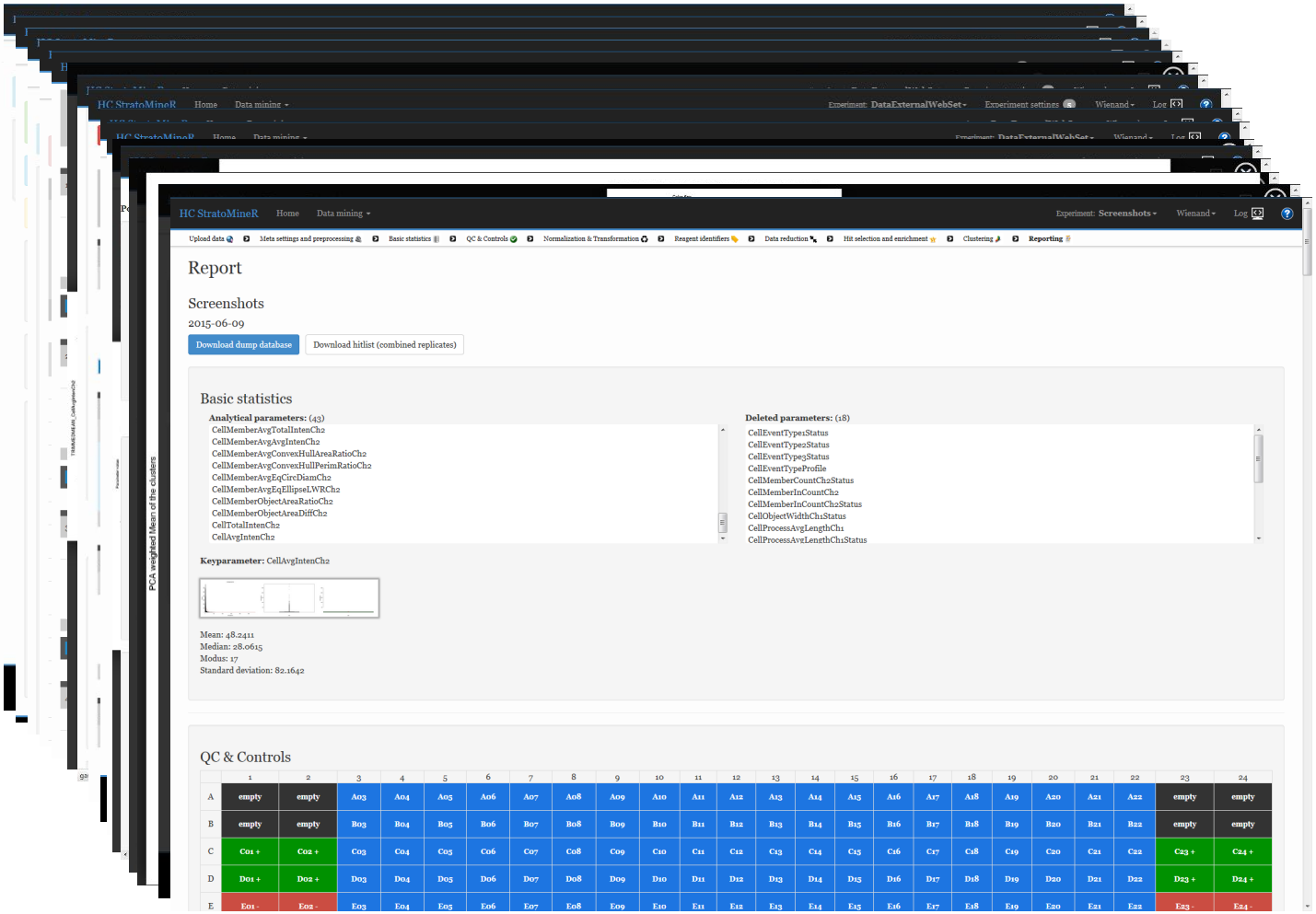


Commercial HC StratoMineR architecture



HC StratoMineR Workflow Demonstration

HC StratoMineR Screenshots



Core Life Analytics

Validation of HC StratoMineR

Prof. Rene Medema

Galápagos

"In **a matter of minutes** HC Stratominer generated **new insights** for our high content screening efforts. It's **smooth looking interface** and abundance of statistical analysis are inviting to **delve deeper into datasets**. I expect that we will make use of it in the future to upgrade our high content efforts to **true multi variable analysis**."

Edo Elstak Ph.D.

Core Life Analytics

Scalable Cloud Architecture
Cost-efficient

Core Life Analytics

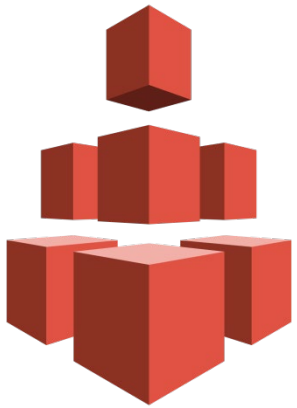
Cloud Providers

Using the Cloud

- **Processing**
- **Storage**
- Security
- Focus (Speed, Cost Efficiency, security, reliability)
- Programming language

Processing

- Lambda (Serverless)
- S3 hosting (plain web pages)
- EC2 instances (nodes)
- AWS services (saas)



Processing using EC2 instances/nodes

- Types
  - GPU
  - # cores
  - Type of Local storage (NVMe, EBS, SSD, Cold storage)
  - Orientation (fast processor, bulk storage, cheap instances for cluster)
- Costs
  - Server types
  - Duration
  - Dedicated/Virtual/Shared/Spot pricing

Storage

- EBS (inflexible, not scalable)
- EFS (very fast, very expensive)
- S3
  - Standard
  - Glacier

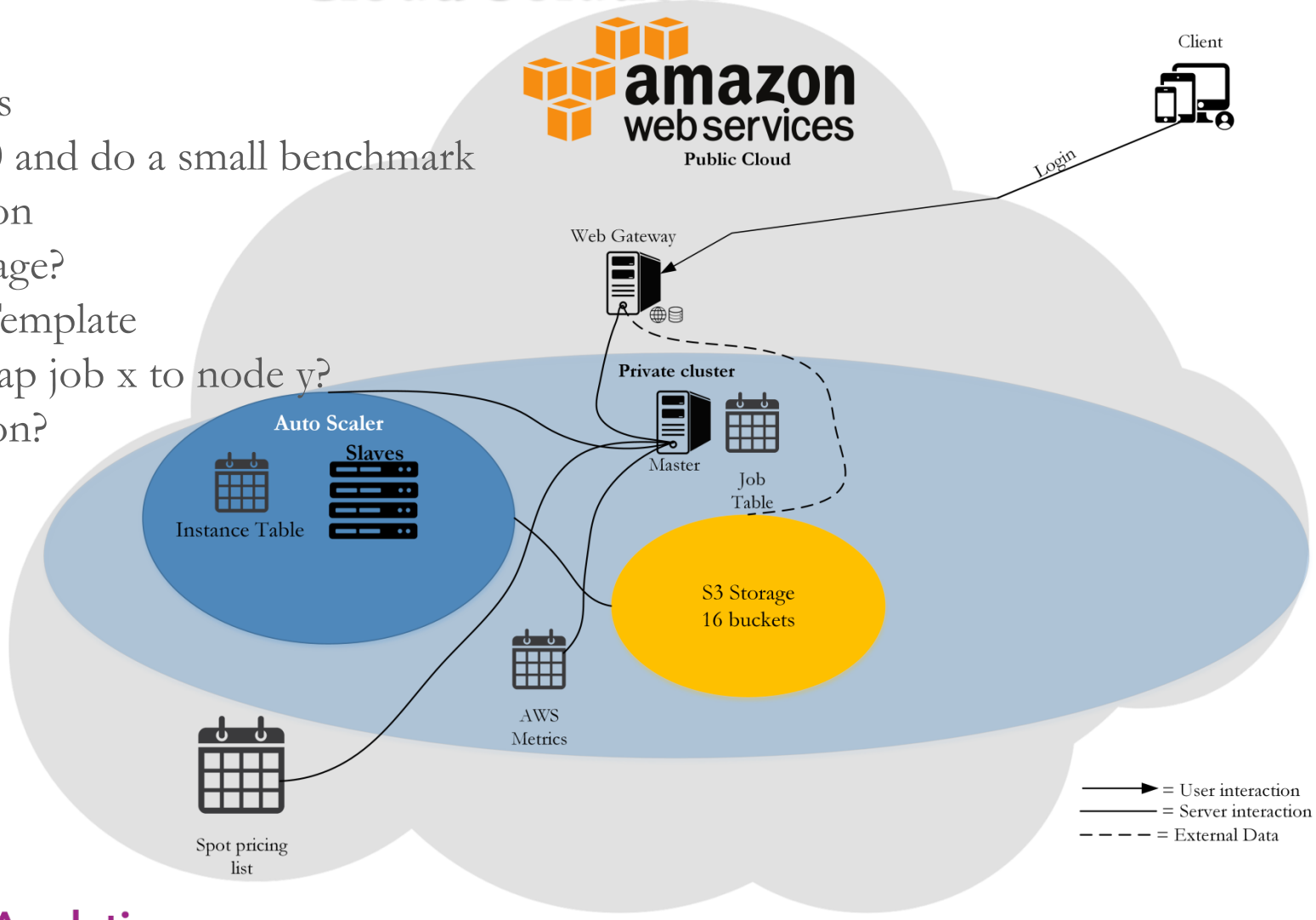
Storage

- Bucket
- Max of 800 connections of 100Mbit/s
- How to scale up to get maximum result?

- 800 connections limit
- md5 every file name to process/store
- Setup 16 buckets a-f & 0-9
- Auto distribute files to 16 buckets
- Scale-up?? → 16*800*100 = 1.28Tb/s (0.16 Terabyte per second)

Cloud Solution

- T-instances
- Spin up 10 and do a small benchmark
- Interruption
- Local storage?
- Instance Template
- How to map job x to node y?
- Aggregation?



Core Life Analytics

Cloud Solution

- **Scalable:** N TB of storage, N nodes for processing
- **Flexible:** Any job, any node
- **Cost-Effective:** Spot pricing
- **Cloud Provider:** AWS
- **Processing:** various EC2 instances
- **Storage:** S3 + NVMe SSD storage
- **Security:** SSL web access, VPC, firewalled, SSH passwordless public + private key
- **Focus (Speed, Cost Efficiency, security, reliability):** Flexible
- **Programming language:** Flexible

Acknowledgements

UMCU Cell Biology

- Prof. Judith Klumperman
- Dr. David Egan
- ing. Jacob de Nobel
- ing. Desmond Robers

- **UMCU Julius Centrum**
- Prof. Rene Eijkemans

- **UU Decision support systems**
- Prof. Linda van der Gaag

NKI

- Prof. René Medema
- Dr. Roy van Heesbeen

- **UU Algorithmic data analysis**
- Prof. Arno Siebes
- Dr. Ad Feelders

UU Software Systems

- Prof. Sjaak Brinkkemper
- Dr. Marco Spruit
- Drs. Ian Shen

UMC Utrecht

Utrecht University

Core Life Analytics