

Data Science & Society

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



October 23, 2018
Wienand Omta



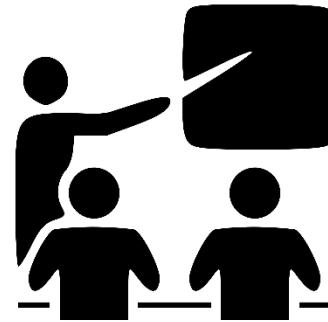
Utrecht University



Core Life Analytics

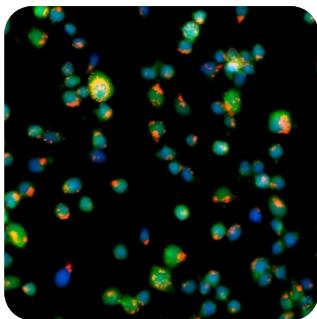
Agenda Data Analytics Lecture

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



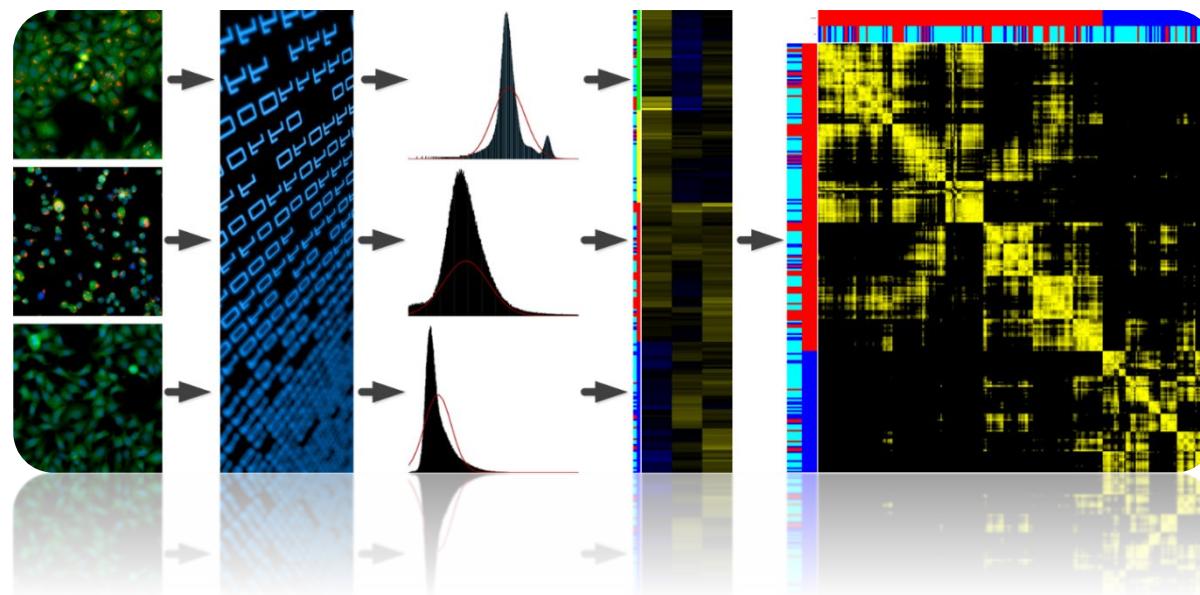
Who we are and what we do

- 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

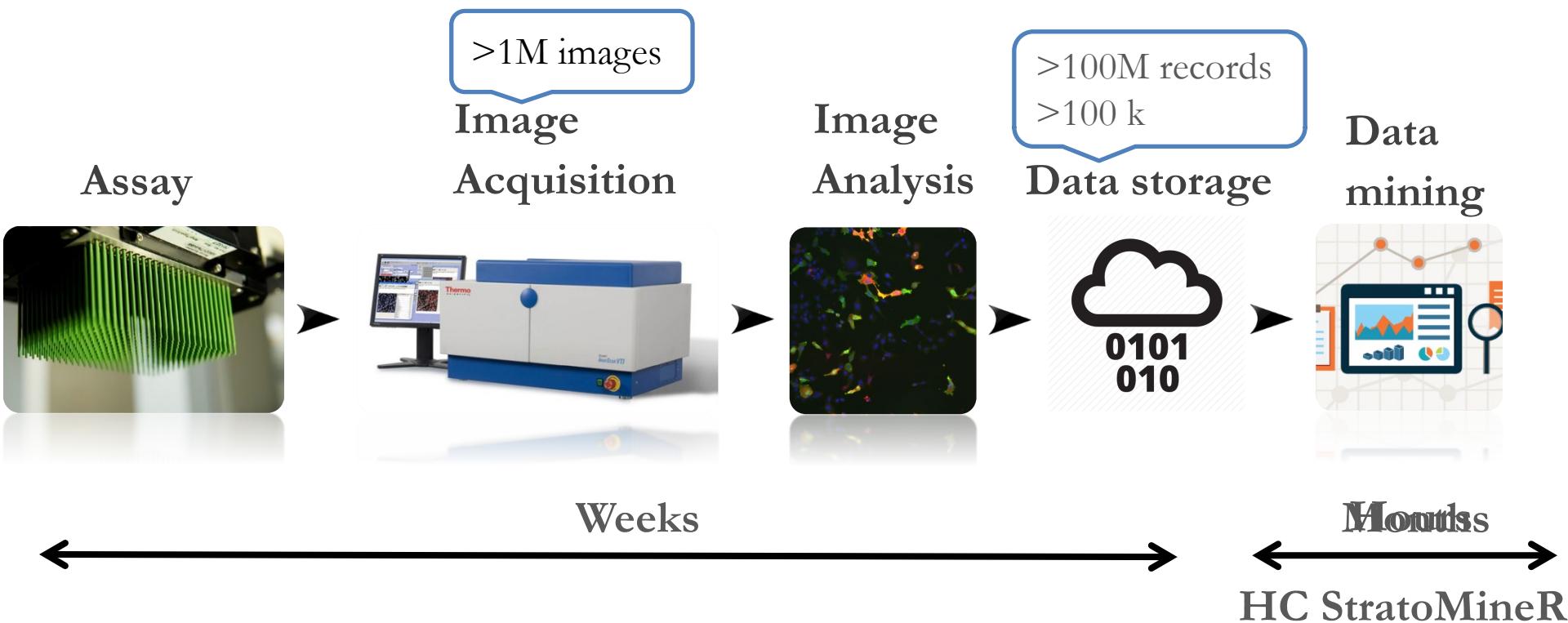


Core Life Analytics

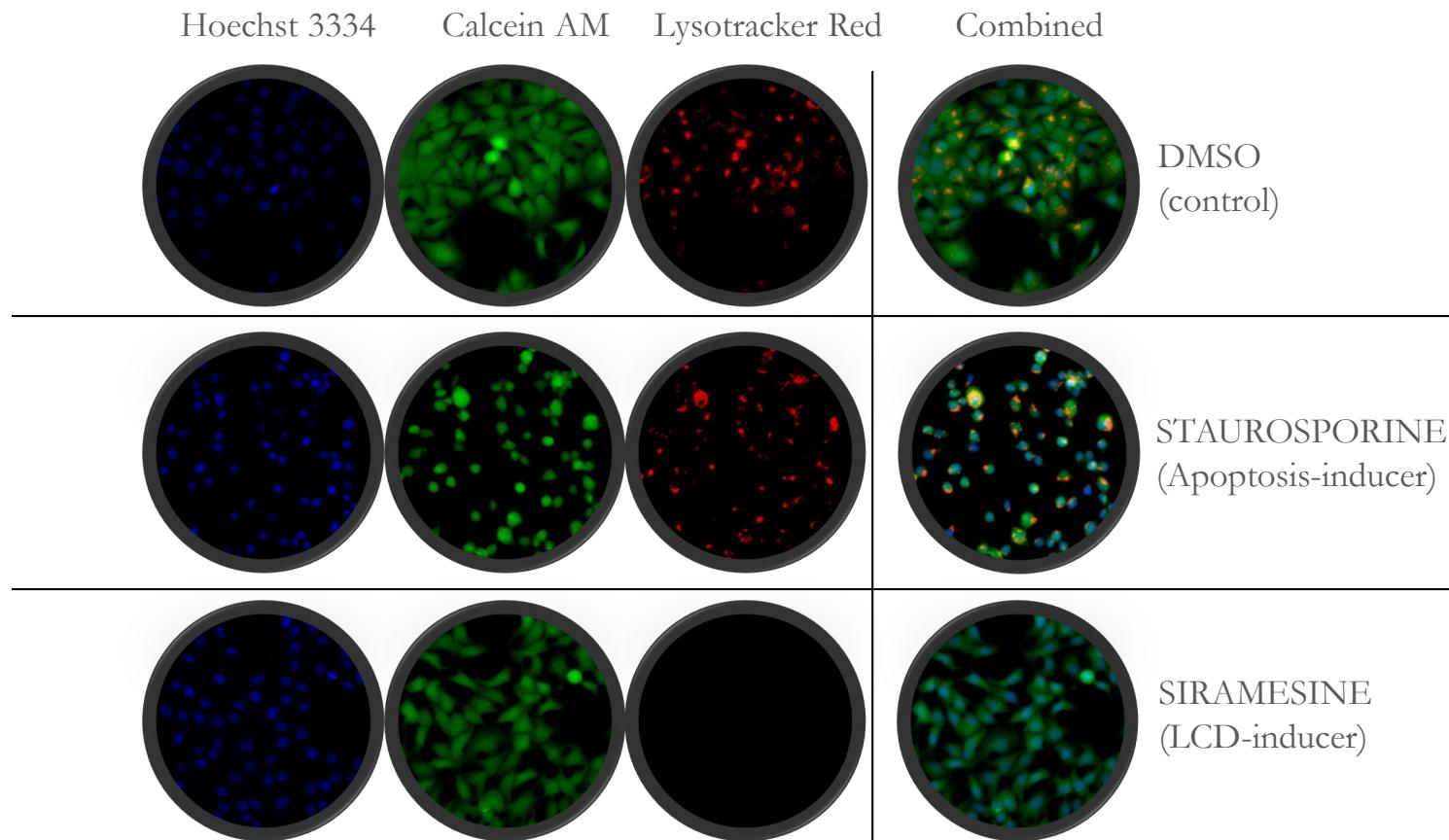
What is HCS?



- 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)



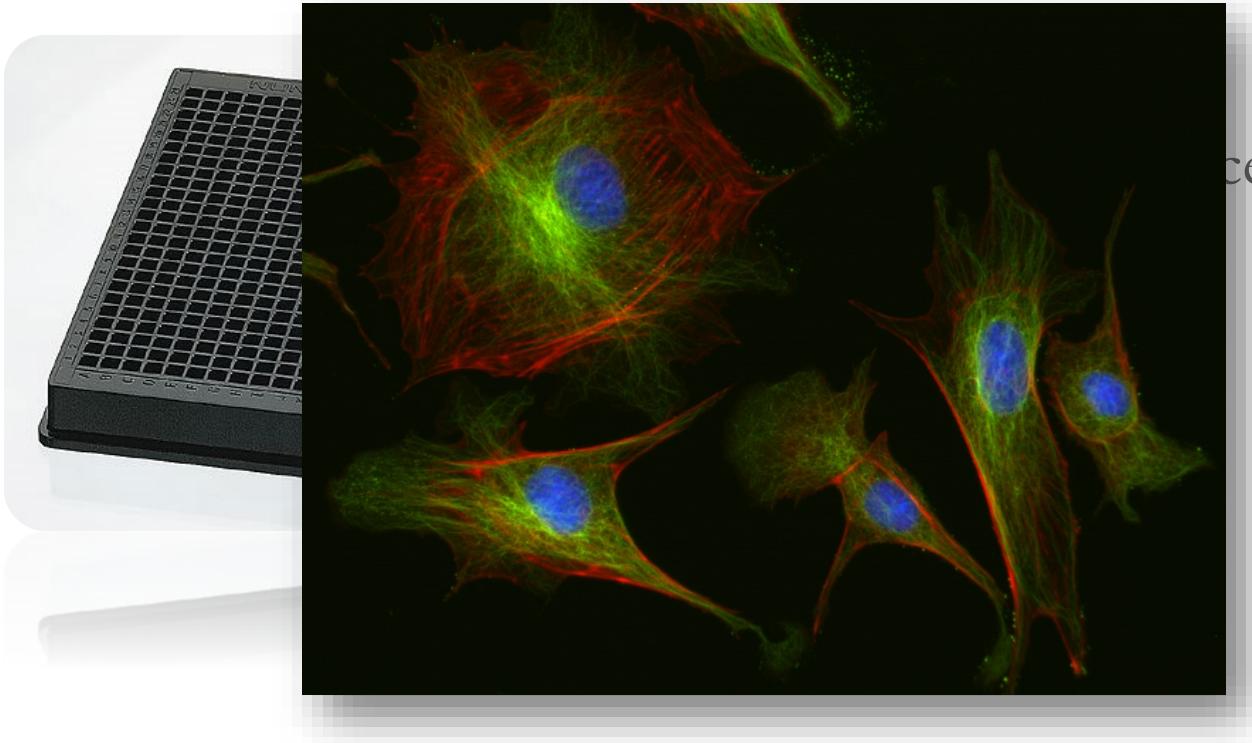
The Power of High Content Screening



Case study (Example High Content Screen)

Problem (I)

Case study

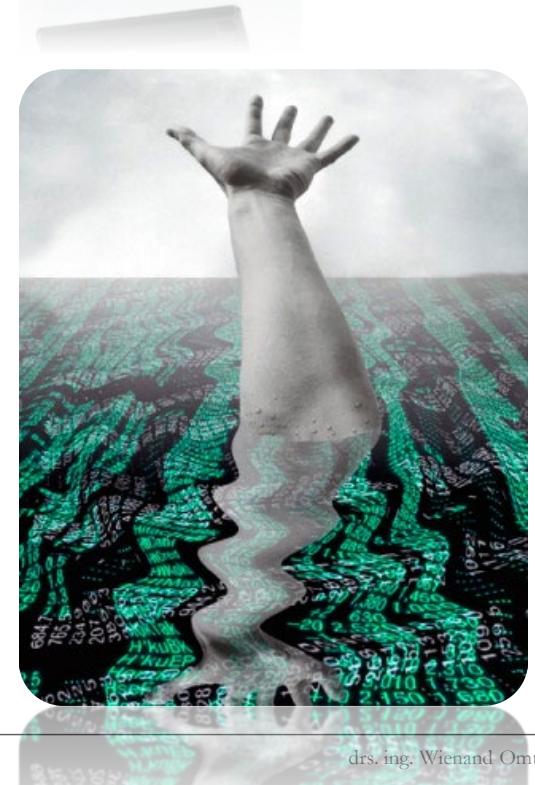


cells per well

Problem (II)

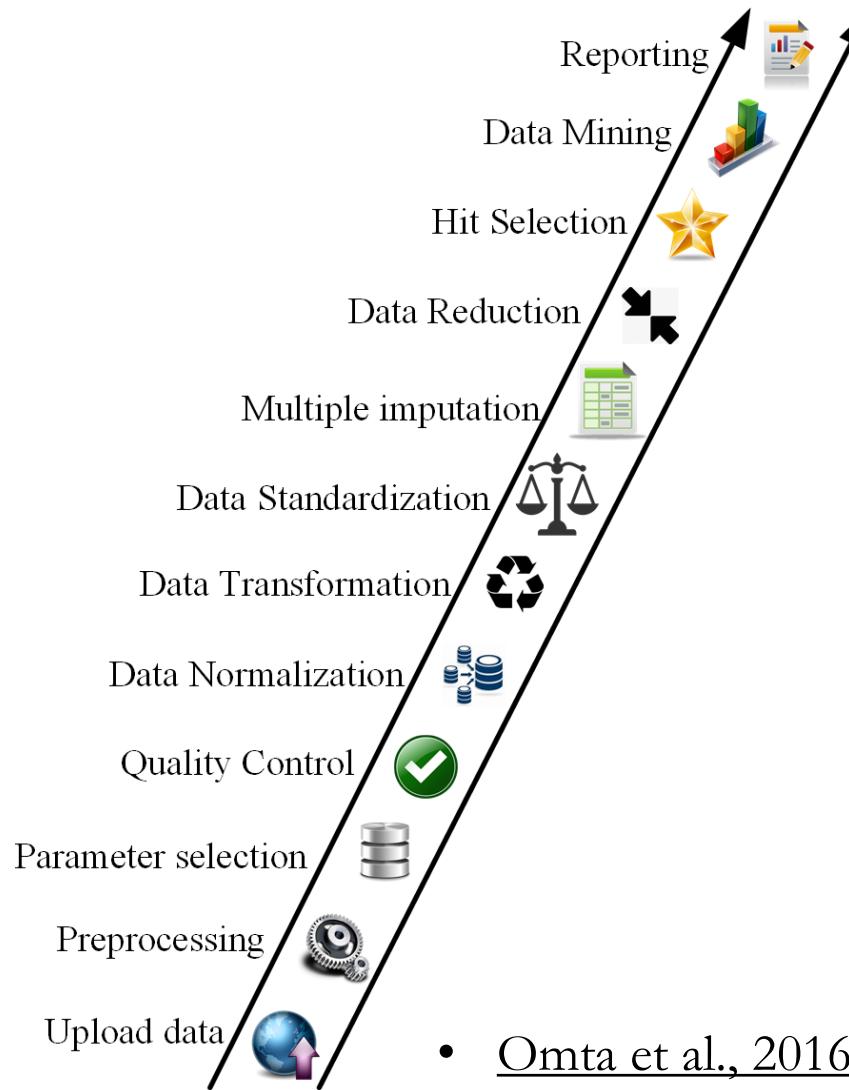
Case study

- 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.)
- $162 \times 384 \times 2 \times 1000$, 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)

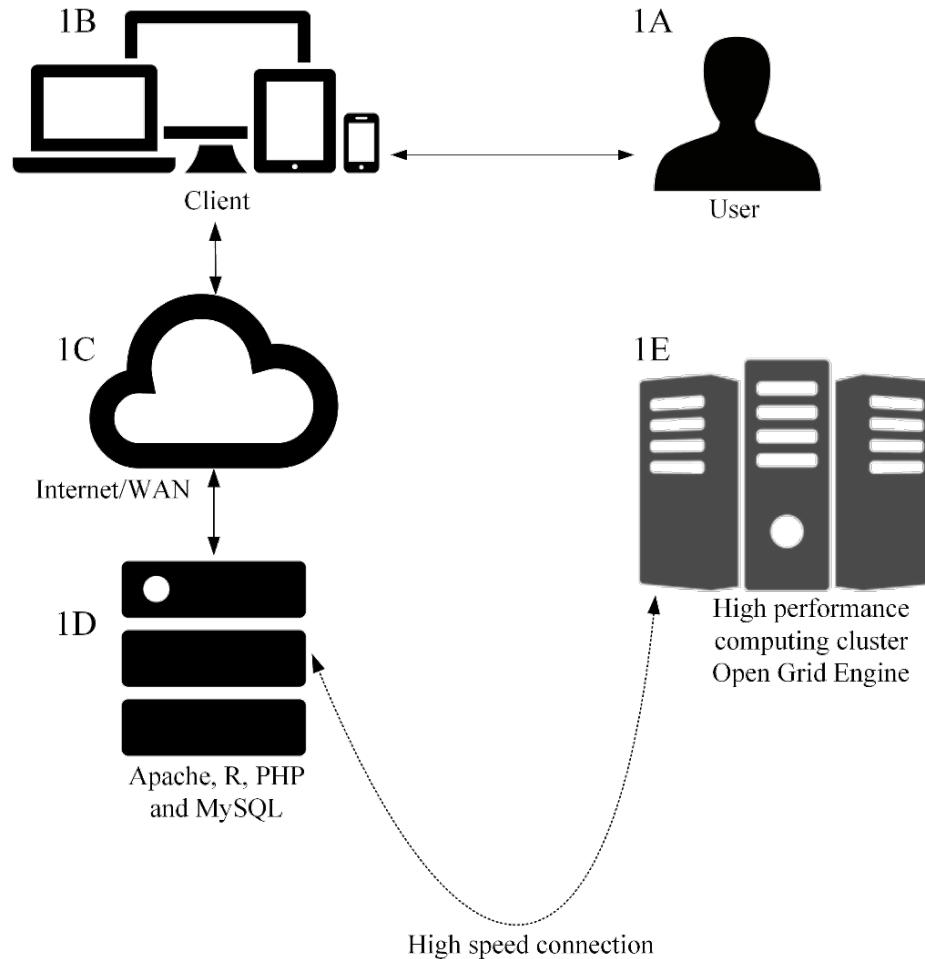


'Drowning in data and starving for knowledge'
Singh, Carpenter and Genovesio (2014)

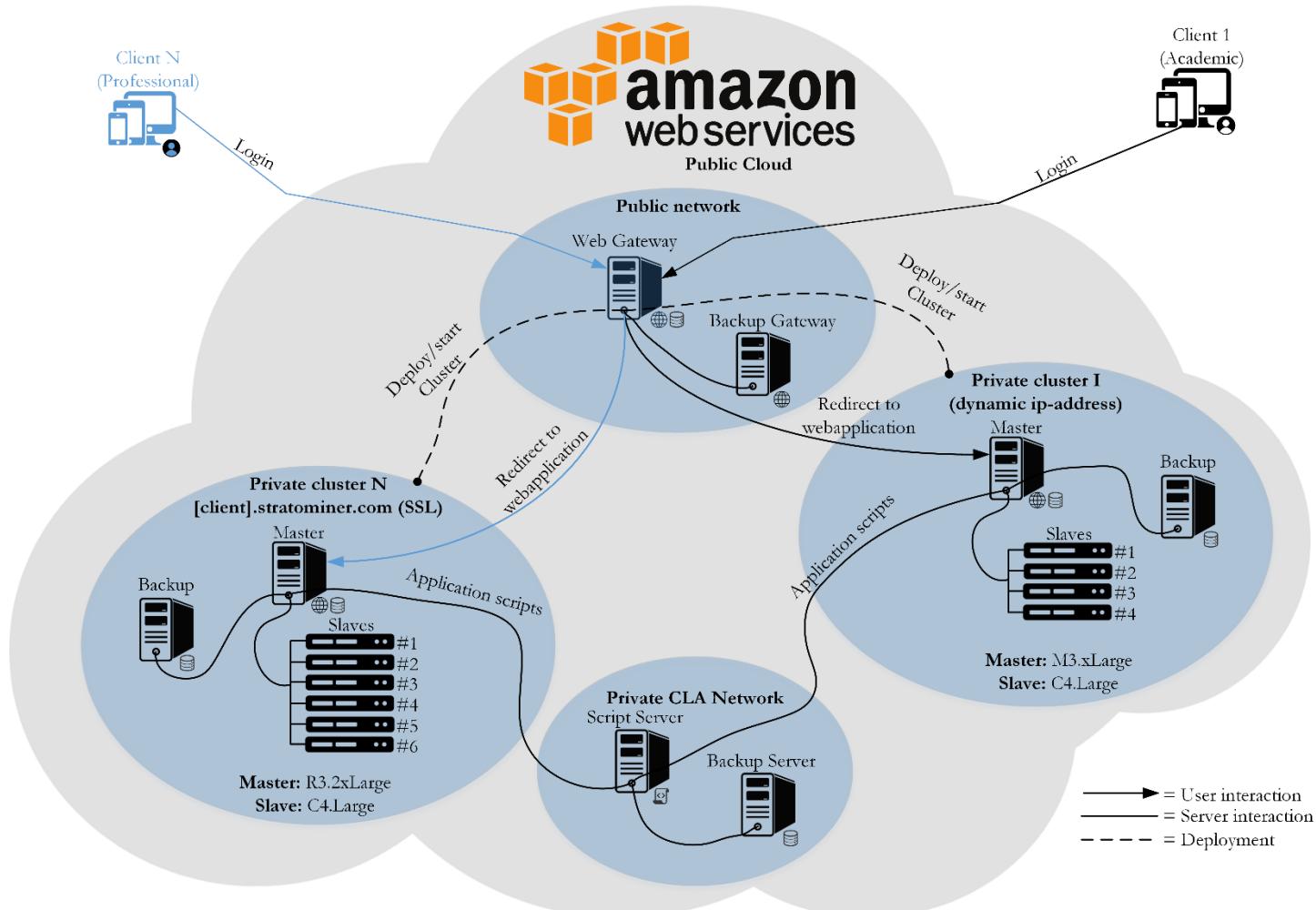
HC StratoMineR workflow



Academic HC StratoMineR architecture



Commercial HC StratoMineR architecture



HC StratoMineR Workflow Demonstration



Core Life Analytics

HC StratoMineR Screenshots

The image shows a complex software interface for bioinformatics analysis, specifically HC StratoMineR. The main window displays a 'Report' for 'Screenshots' from 2015-06-09. The report includes sections for 'Basic statistics' and 'QC & Controls'. The 'Basic statistics' section shows a histogram for 'CellAvgIntenCh2' with a mean of 48.2411, median of 28.0615, and standard deviation of 82.1642. It also lists analytical and deleted parameters. The 'QC & Controls' section is a grid where rows A-E and columns 1-24 represent different samples or conditions, with some entries colored green, red, or black.

Report

Screenshots
2015-06-09

Download dump database Download hitlist (combined replicates)

Basic statistics

Analytical parameters: (43)
CellMemberAvgTotalIntensityCh2
CellMemberAvgAvgIntensityCh2
CellMemberAvgConvexHullAreaRatioCh2
CellMemberAvgConvexHullPerimRatioCh2
CellMemberAvgEqCrcDiamCh2
CellMemberAvgEqEllipsLVRCh2
CellMemberObjectAreaRatioCh2
CellMemberObjectAreaDiffCh2
CellTotalIntensityCh2
CellAvgIntensityCh2

Deleted parameters: (18)
CellEventTypeStatus
CellEventTypeStatus
CellEventType2Status
CellEventTypeProfile
CellMemberCountCh2Status
CellMemberInCountCh2
CellMemberInCountCh2Status
CellObjectWidthCh2Status
CellProcessAvgLengthCh1
CellProcessAvgLengthCh2Status

Keyparameter: CellAvgIntensityCh2

Mean: 48.2411
Median: 28.0615
Modus: 17
Standard deviation: 82.1642

QC & Controls

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
A	empty	empty	A02	A04	A05	A06	A07	A08	A09	A10	A11	A12	A13	A14	A15	A16	A17	A18	A19	A20	A21	A22	empty	empty
B	empty	empty	B02	B04	B05	B06	B07	B08	B09	B10	B11	B12	B13	B14	B15	B16	B17	B18	B19	B20	B21	B22	empty	empty
C	C01 +	C02 +	C02	C04	C05	C06	C07	C08	C09	C10	C11	C12	C13	C14	C15	C16	C17	C18	C19	C20	C21	C22	C23 +	C24 +
D	D01 +	D02 +	D02	D04	D05	D06	D07	D08	D09	D10	D11	D12	D13	D14	D15	D16	D17	D18	D19	D20	D21	D22	D23 +	D24 +
E	E01 -	E02 -	E02	E04	E05	E06	E07	E08	E09	E10	E11	E12	E13	E14	E15	E16	E17	E18	E19	E20	E21	E22	E23 -	E24



Validation of HC StratoMineR

Prof. Rene Medema



NKI RIVM



high science

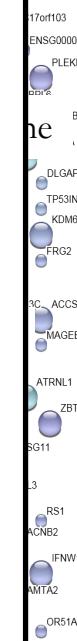


Galápagos



"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



Core Life Analytics

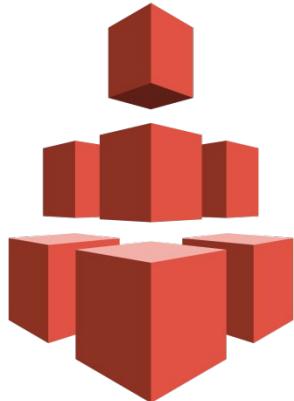
Using the Cloud

- 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)



Amazon EC2



Core Life Analytics

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



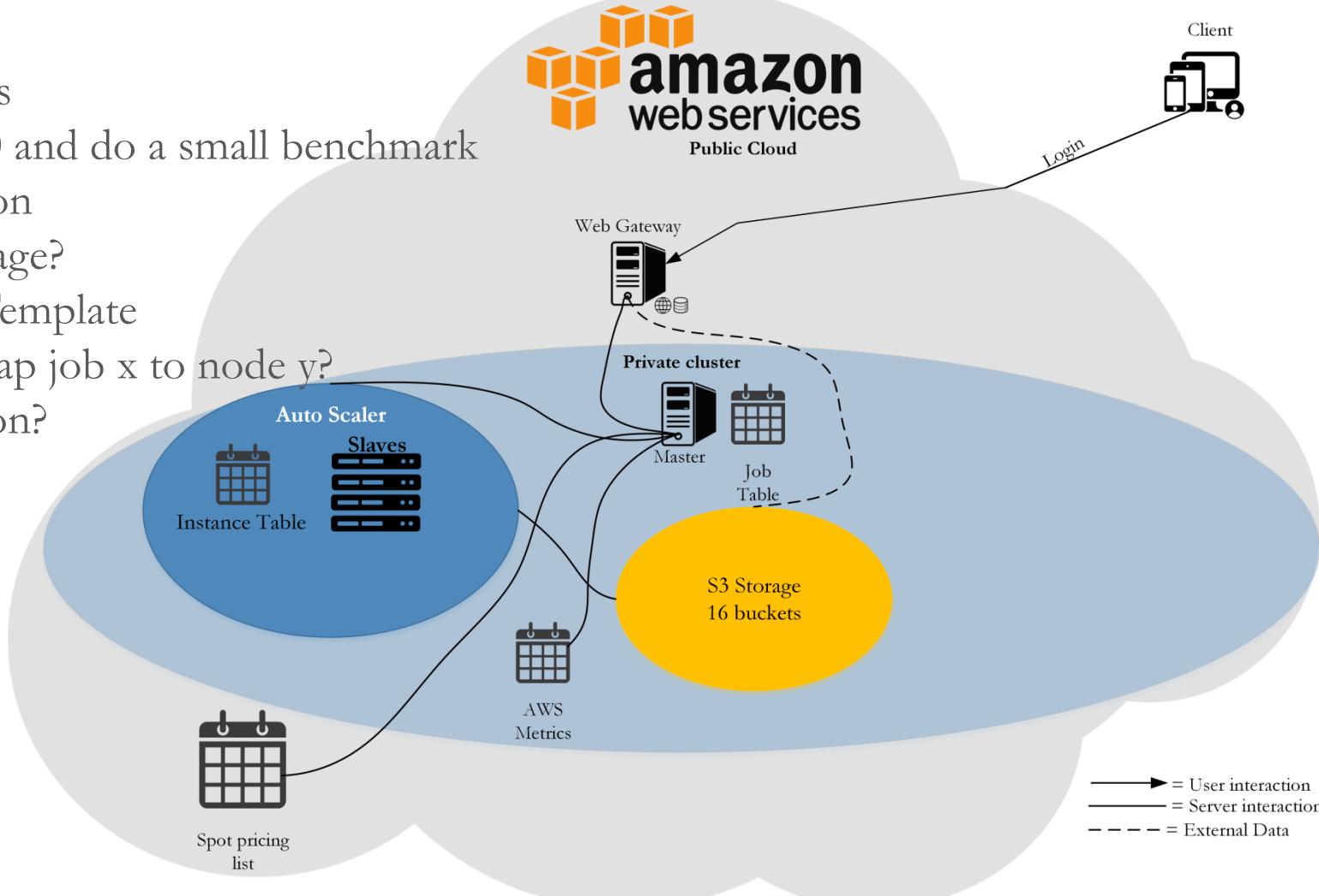
Storage

- 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?? → $16*800*100 = 1.28\text{Tb/s}$ (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?



→ = User interaction
— = Server interaction
- - - = External Data



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