Data Analytics in Cloud

Dr. Nishant Saurabh

Utrecht University

About me

- Current position
  - **Assistant Professor (Software Division)**
  - **Information and Computing Sciences**
  - **Utrecht University**
- Research
  - **Cloud, Edge and Quantum computing systems**
  - **Distributed Computing and Storage**
  - **Performance Management**
- Teaching
  - **Cloud and Edge computing (coordinator)**
  - **Software architecture**
  - **Information security**
- More about my research:
  - **Web:** https://www.uu.nl/staff/NSaurabh
  - **Google scholar:** https://scholar.google.co.in/citations?user=UI10EQQAAAAJ&hl=en

Utrecht University

Agenda for today

- Data to Big data landscape
- Cloud deployment and service models
- Cloud and Big data
- Challenges to realising Big data analytics
- Take-away message

Utrecht University

What is Data?

Utrecht University

What is Data?

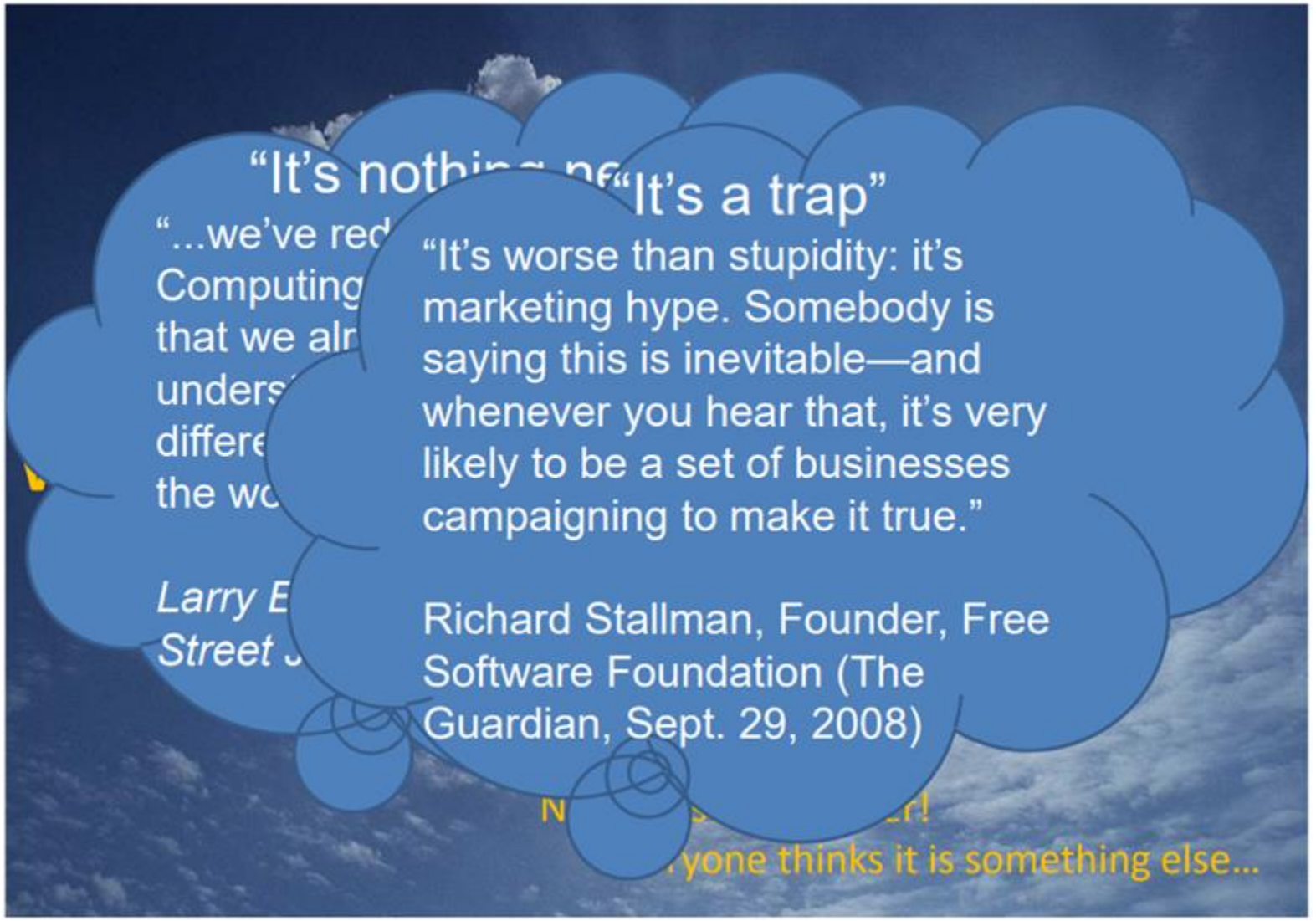


Figure taken from Prof. Anthony D. Joseph's lecture at RWTH Aachen.

Utrecht University

Data, data, data! Everywhere!

Large Hadron Collider generates 40TB data per second

Boeing Jet Engine creates 10TB information every 30 minutes

Google

Crawls more than 20B web pages a single day

NETFLIX

695,000 hours of content watched every minute

463 Exabytes of data generation per day by 2025!!!! (2.5 Exabytes in 2012 ?)

Utrecht University

Data, data, data! Everywhere!

“640K ought to be enough for anybody.”
- Bill Gates (1981)

Utrecht University

Era of Big Data!

Utrecht University

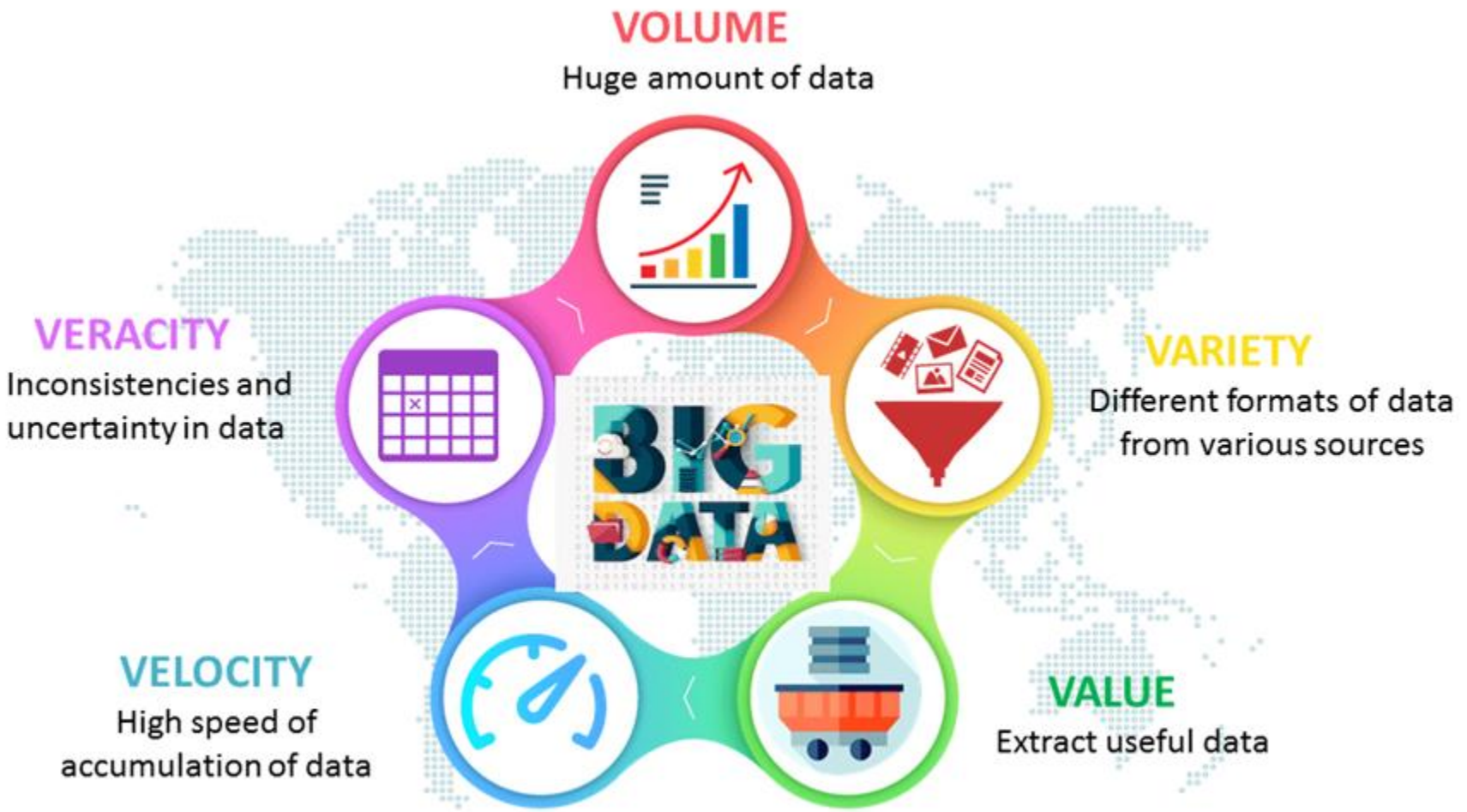
What is Big Data?

Big Data is similar to Small Data, but
Just a bit BIGGERRR!

"Big Data is a term for datasets that are so large or complex that traditional data processing techniques are inadequate to deal with them. Challenges include analysis, capture, integration, data curation, search, sharing, storage, transfer, visualization, querying, updating, security and privacy"

Utrecht University

Big data characteristics



5 Vs of Big Data

Utrecht University

Big data characteristics

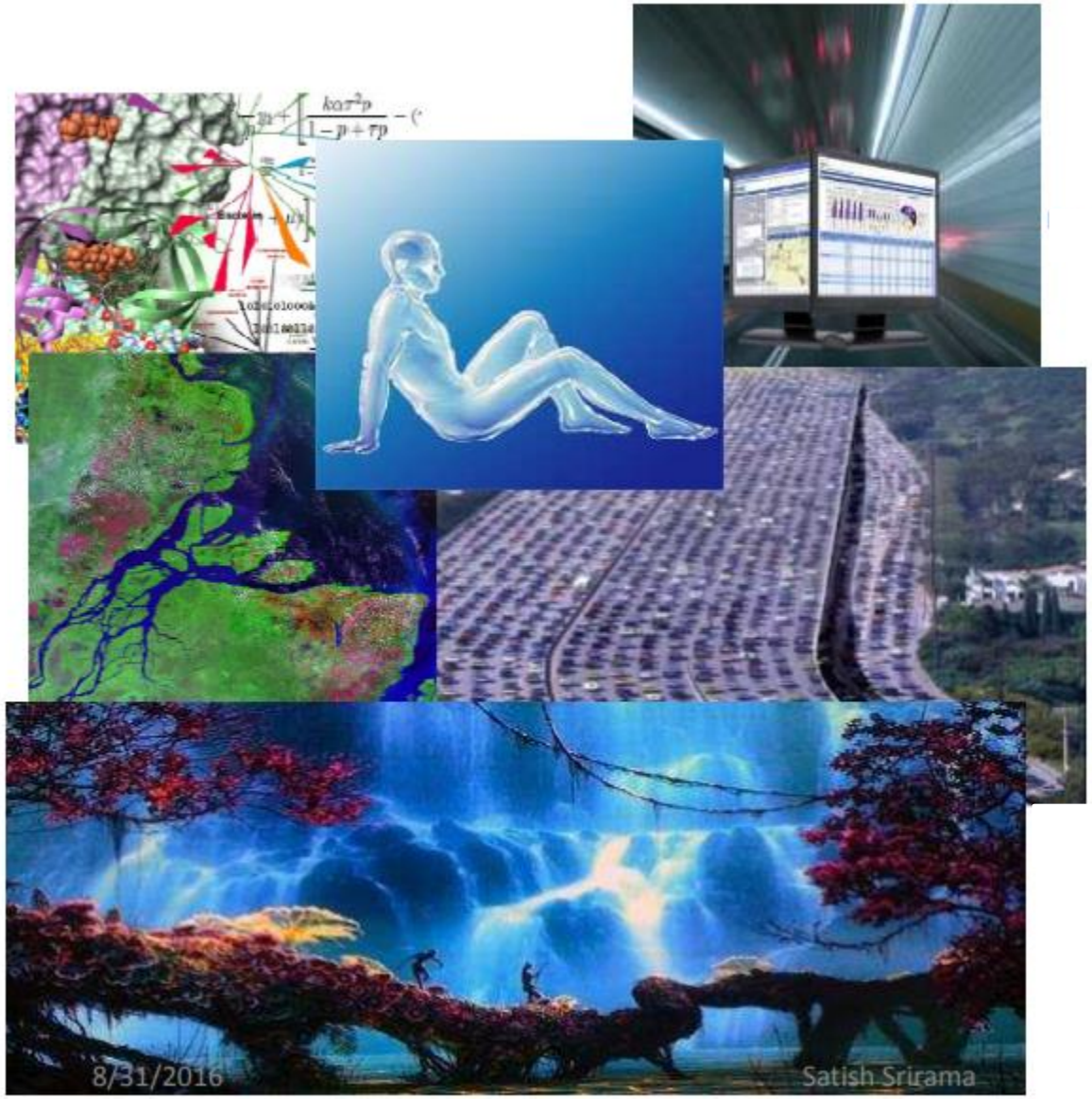
- V5: V1 for Volume
  - **Rapidly increasing size of data to be processed**
  - **More storage capacity, more computation …**
- V5: V2 for Variety
  - **Various formats, types and structures,**
  - **Text, numerical, images, audio, video …**
  - **Static data v/s streaming data**
  - **Extracting knowledge requires all data types to be linked together**
- V5: V3 for Velocity
  - **Data is generated fast, hence needs to be processed fast**
  - **Even a delay of 10 ns (nano seconds) delay is too much in some cases**
- V5: V4 for Veracity
  - **Consistency, accuracy, quality and trustworthiness**
  - **Inaccurate data may lead to biased knowledge**
- V5: V5 for value
  - **Useful data in form of unique insights**
  - **Value for decision making as an end goal**

Due to these 5Vs, sometimes

Data is also called Gold Dust in 21st Century

Utrecht University

What can we do with this wealth?



- Scientific breakthroughs
- Business process efficiency
- Improve quality of life
  - **Transportation**
  - **Healthcare**
  - **Disaster management**
  - **Day to day life activities**

- Can we do more ?

Utrecht University

*How can we process **MASSIVE** Amount of **DATA**?*

Utrecht University

Cloud Comes to the Rescue!

Utrecht University

You have house to rent

- What does the tenant want?

  An independent house

- What can you offer ?

Utrecht University

You have house to rent

- What does the tenant want?
- What can you offer ?

- Is it affordable ?
- Is it spacious?
- Will I be disturbed by outsiders?
- Will energy cost be billed separately?

Utrecht University

You have a Computer to Rent ?

- What does the tenant want?
  Their own Computer
- What can you offer and How?
- What does the tenant look for?
  Is it affordable to rent?
  Is there enough CPU/Memory/Disk capacity?
  Is network connection efficient enough?
  Do I pay for what I use?

Utrecht University

What is Cloud Computing?

NIST Definition of Cloud

A model for enabling ubiquitous, convenient, on-demand network access to a shared pool of configurable computing resources (e.g. networks, servers, storage, applications and services) that can rapidly provisioned and released with minimal management effort or service provider interaction.

Utrecht University

What is Cloud Computing?

"Simply put, cloud computing is the delivery of computing services – servers, storage, databases, networking, software, analytics and more – over the Internet ("the cloud"). Companies offering these computing services are called cloud providers and typically charge for cloud computing services based on usage, similar to how you're billed for gas or electricity at home."

https://azure.microsoft.com/en-gb/overview/what-is-cloud-computing/

Utrecht University

Cloud Computing

- Cloud for ALL

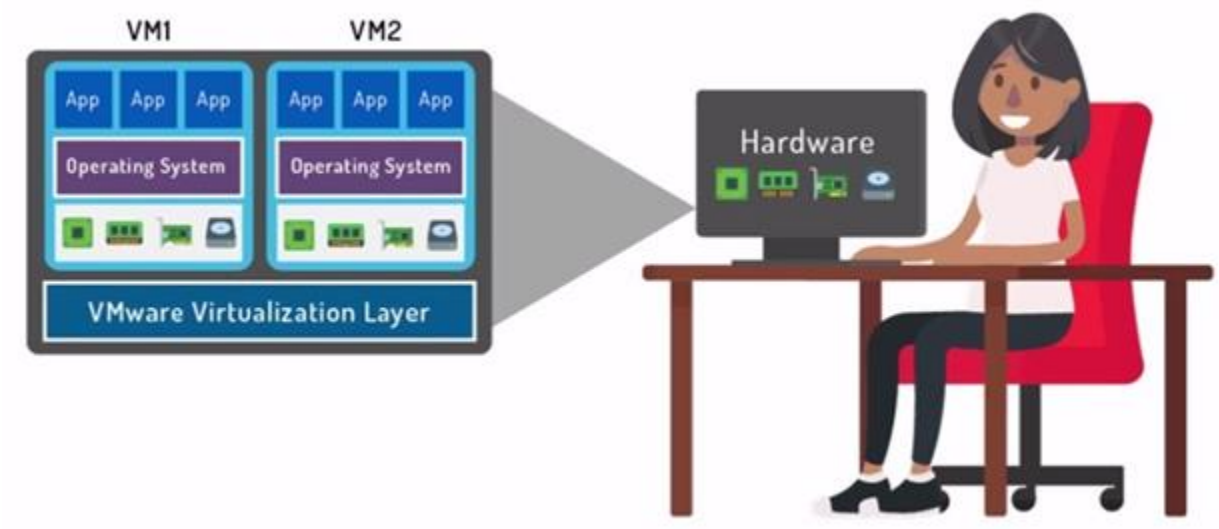
  Model for enabling convenient, on-demand network access to a shared pool of configurable computing resources

- Minimal management effort

  Network, servers, storage, applications, services rapidly provisioned, and released with minimal management effort or service provider interactions

- Key enabler

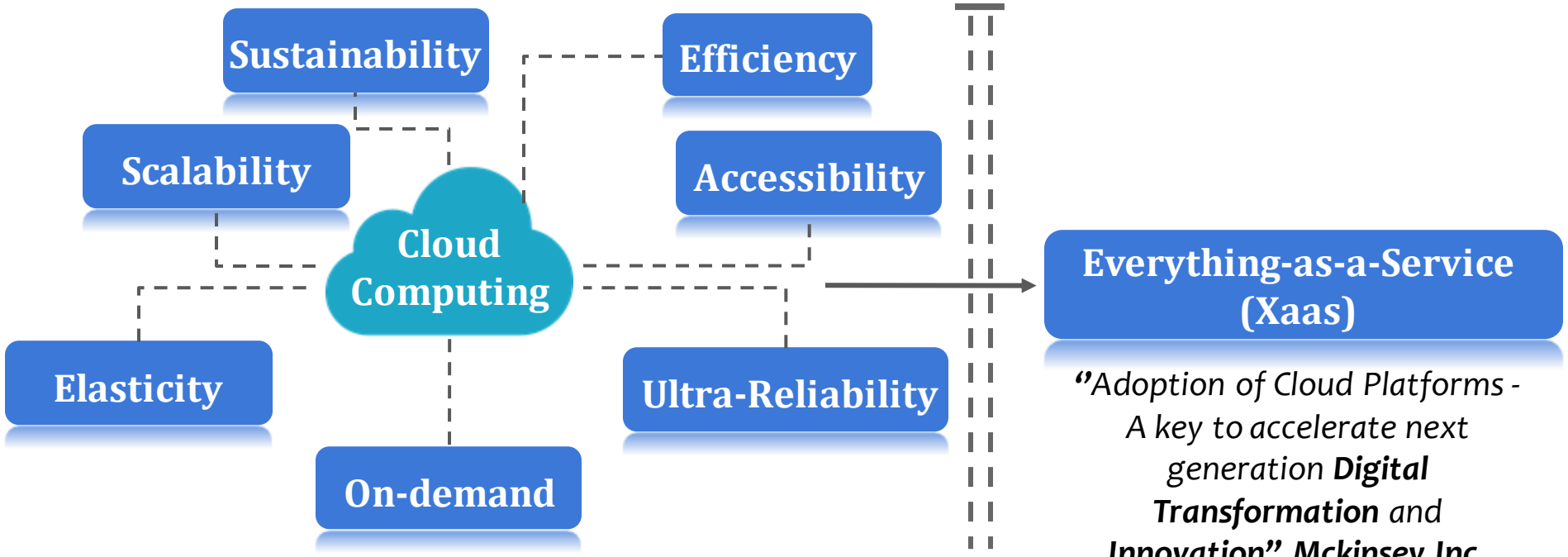
  Virtualization



Utrecht University

What makes Cloud Ubiquitous?

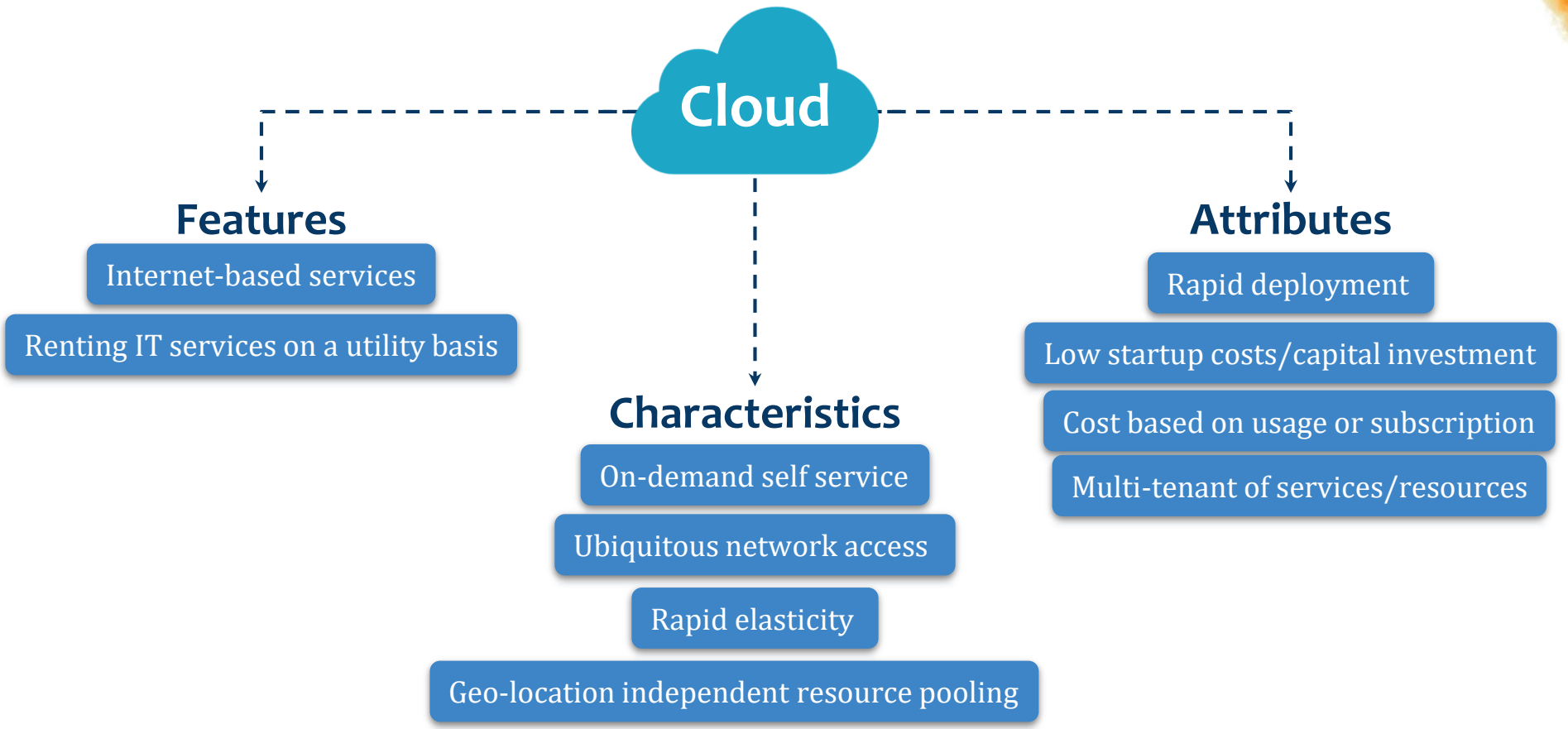
Seven Characteristic Pillars



"Adoption of Cloud Platforms - A key to accelerate next generation ***Digital Transformation*** *and* ***Innovation" Mckinsey Inc, 2020.***

Utrecht University

What makes Cloud Ubiquitous?

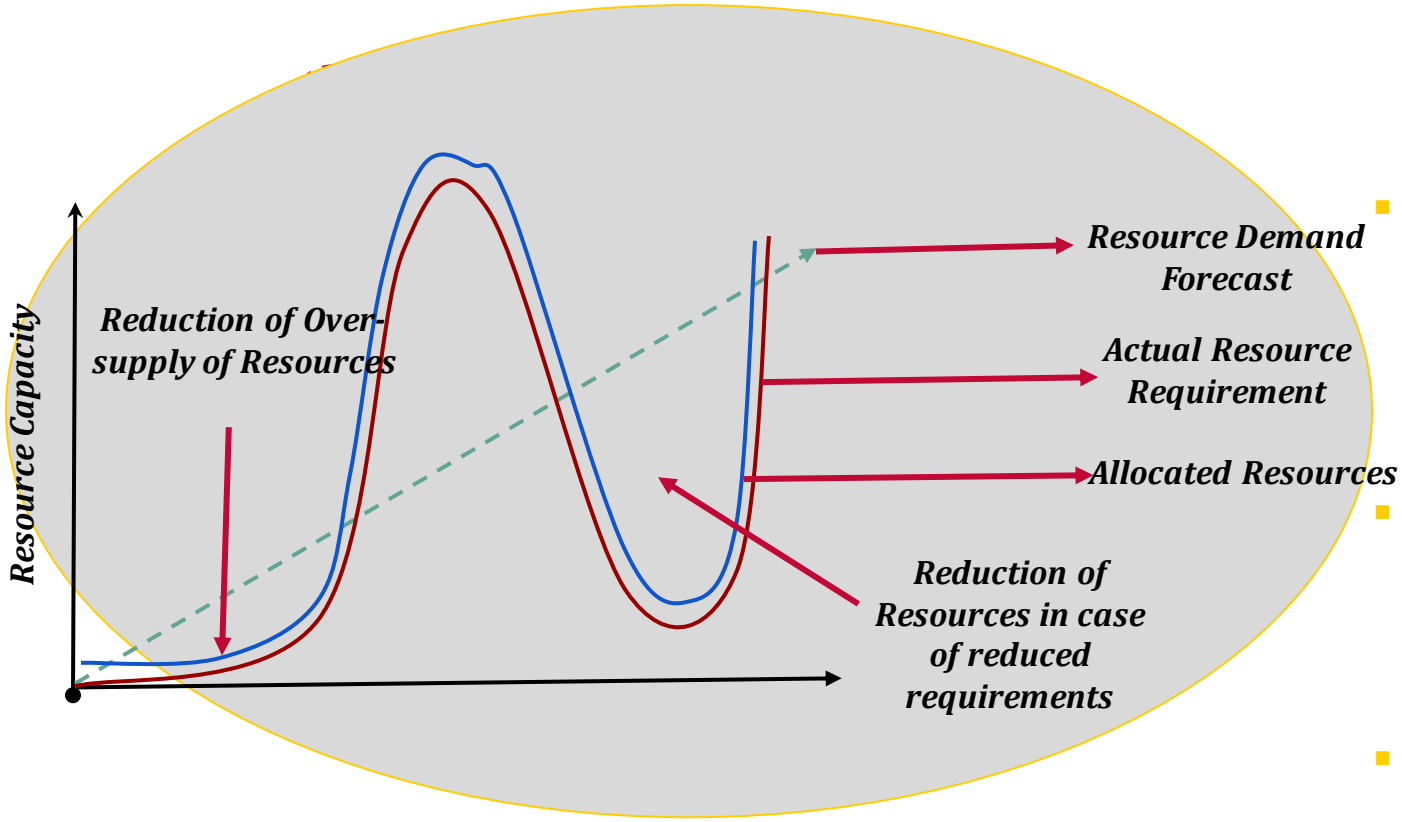


Redefining Cloud Computing

Cloud computing is a compilation of technologies, packaged within a infrastructure paradigm that offers improved ***scalability, elasticity, business agility, faster startup time, reduced management costs*** and ***just-in-time availability of resources.***

Utrecht University

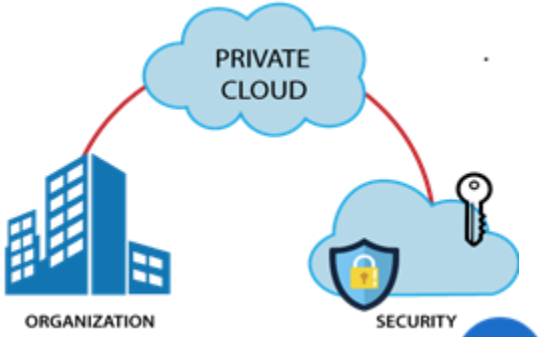
What makes Cloud Ubiquitous?

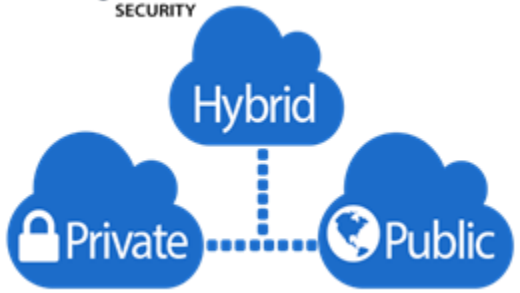


- On-demand self service
  - **Unilaterally provision computing capabilities without requiring human interactions**
- Broad network access
  - **Available over the network**
  - **Heterogenous thin or thick client platforms (mobile phones, laptops)**
- Resource pooling
  - **Compute resources serve multiple users**
  - **Multi-tenant model**
- Metering capability
- Rapid elasticity
  - **Scale out**
  - **Scale in**

Utrecht University

Cloud deployment models





- Public Cloud

  Cloud infrastructure available to all, owned by large organisations responsible for control of data and operations in Cloud
  Example: Amazon, Google Cloud

- Private Cloud

  Cloud infrastructure for and managed by the organization or a third party (On-premise, Off-premise)
  Example: OpenStack

- Hybrid Cloud

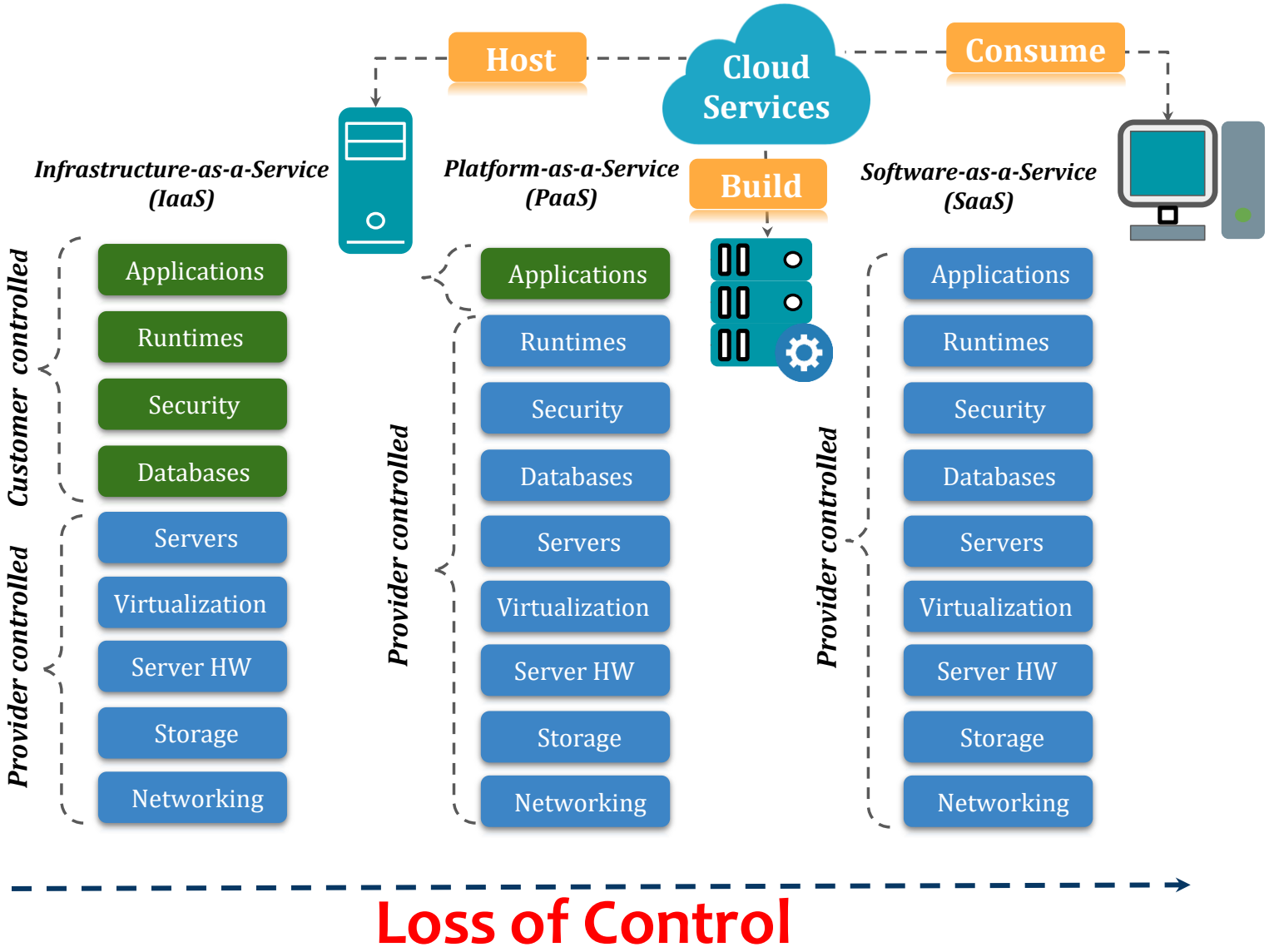
  A combination of two or more public and private Clouds federated by standardized protocol
  Example: Rackspace Cloud

- Community Cloud

  Cloud infrastructure shared by organisations with similar business goals
  Example: Salesforce

Utrecht University

Cloud delivery models



- Software-as-a-Service
- Platform-as-a-Service
- Infrastructure-as-a-Service

Utrecht University

Cloud Datacenters

CISCO

GOOGLE

Microsoft datacenter: 11.5 times the size of a football field.
More than 100, 000+ servers

MICROSOFT

NSA

Utrecht University

Cloud Datacenters

"I think there is a world market for maybe five computers."

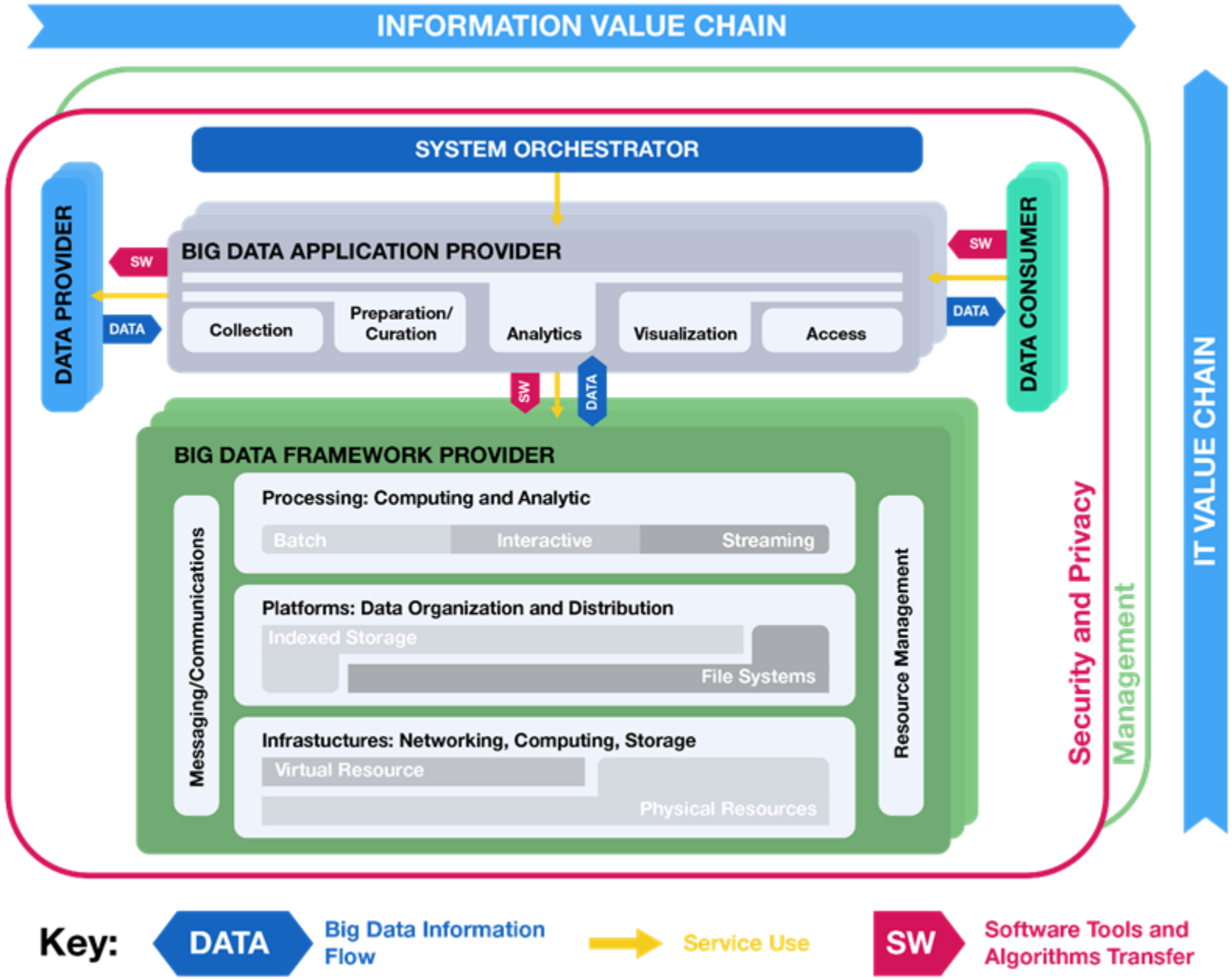
- Thomas Watson, Head of IBM (1943)

Utrecht University

*Now that we have **massive computing power**,*
What's next ?

Utrecht University

NIST Big Data Architecture



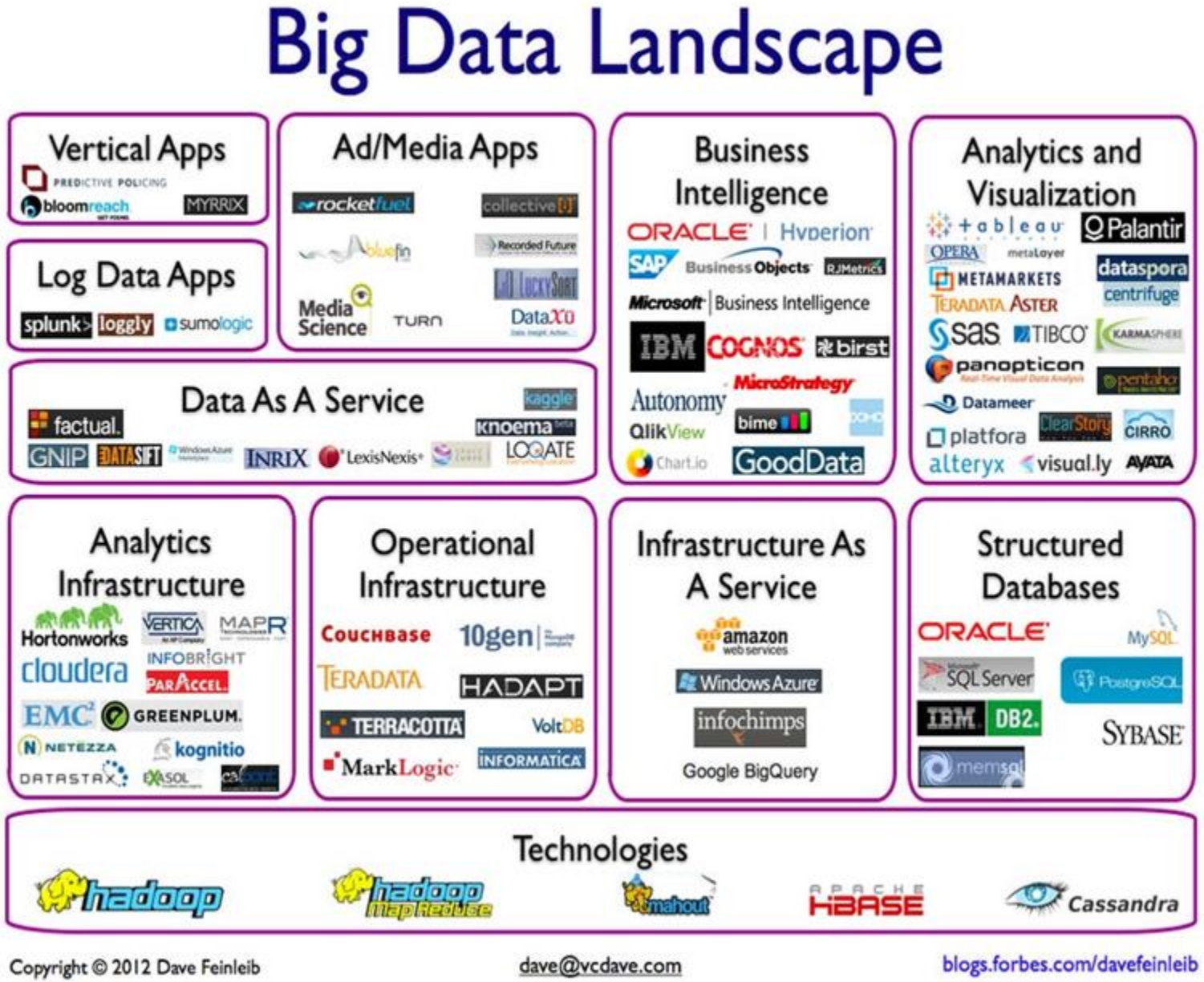
Utrecht University

Big Data-As-A-Service

Outsource Big Data Management and Analytics to Third parties such as Cloud.

Utrecht University

Big Data-As-A-Service



Utrecht University

Big Data service models

- BI-as-a-Service
  Extracting data, data warehousing, interactive front-end
  Example: AWS Quicksight, PowerBI
- ML-as-a-Service
  Predictive analytics, deep learning, visualization services
  Example: TensorFlow
- Database-as-a-Service
  Database management system with fast querying
  Example: Amazon DynamoDB
- Computing/Storage-as-a-Service
  Access to powerful machines and disk for computing and storage
  Example: Amazon VMs/Containers, S3 object storage

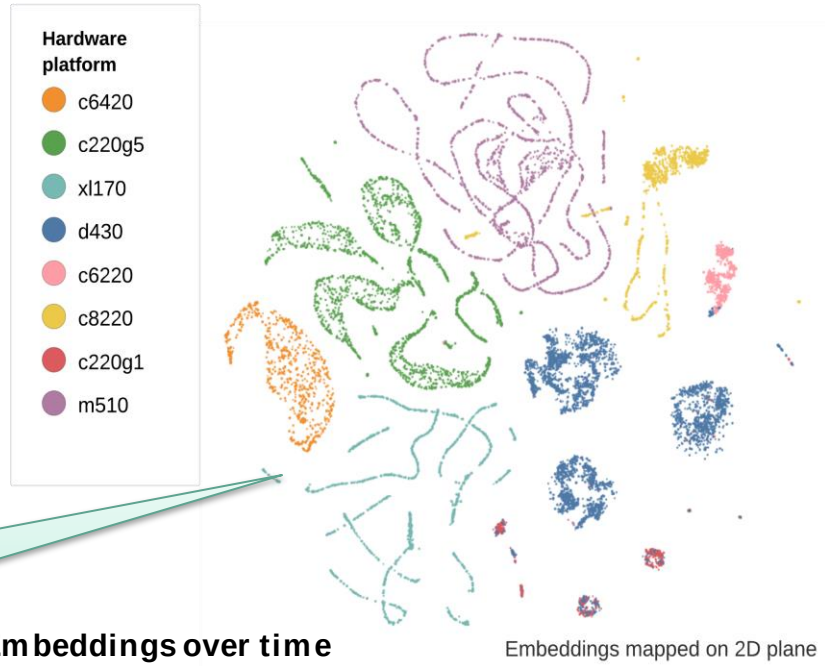
Utrecht University

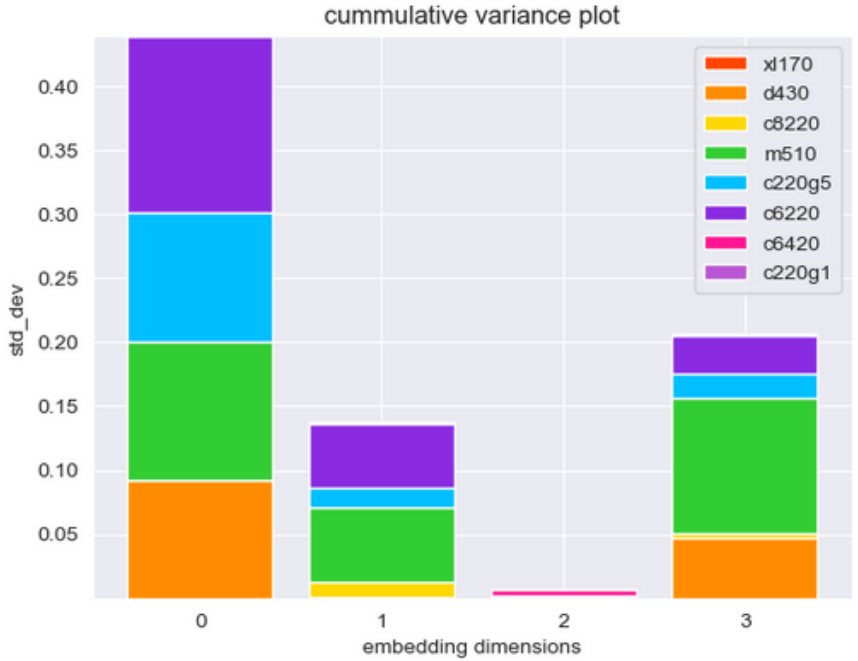
Can Data Analytics and Cloud Computing walk hand in hand ?

Utrecht University

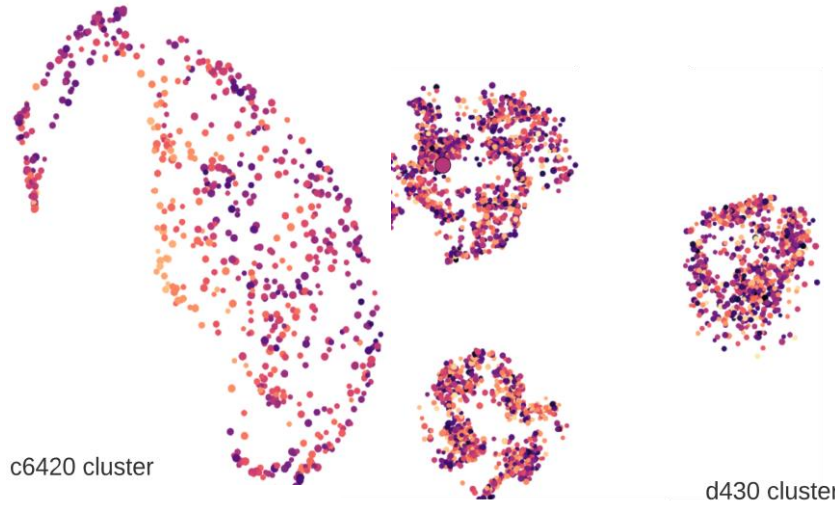
From colleagues of your program!

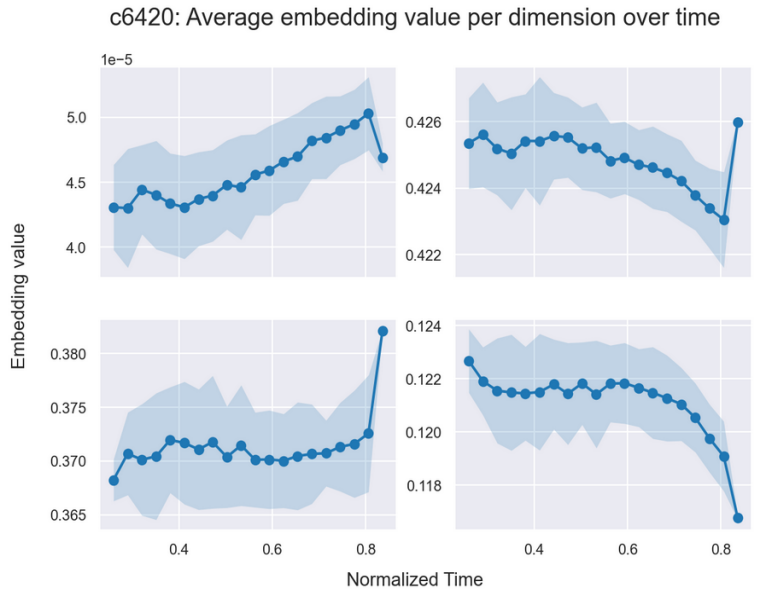
Fingerprinting the Cloud infrastructure





Embeddings over time





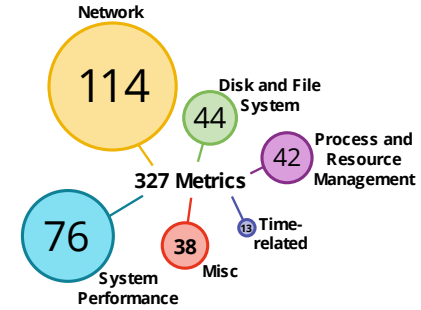
From colleagues of your program!

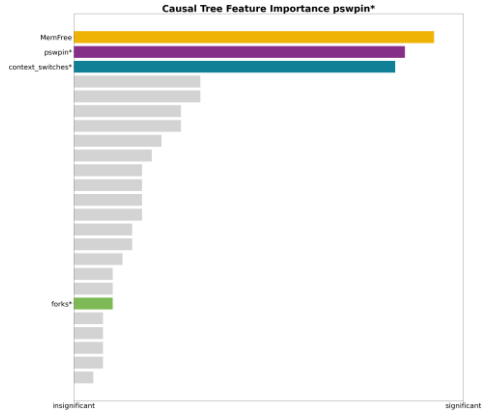
Utrecht University

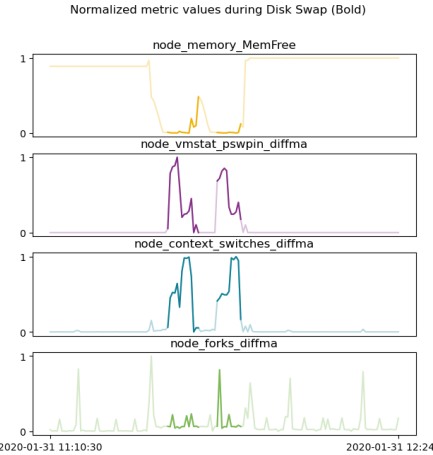
1. Time-lagged cross correlation function
2. Recursive Decision Trees
3. DAGMA

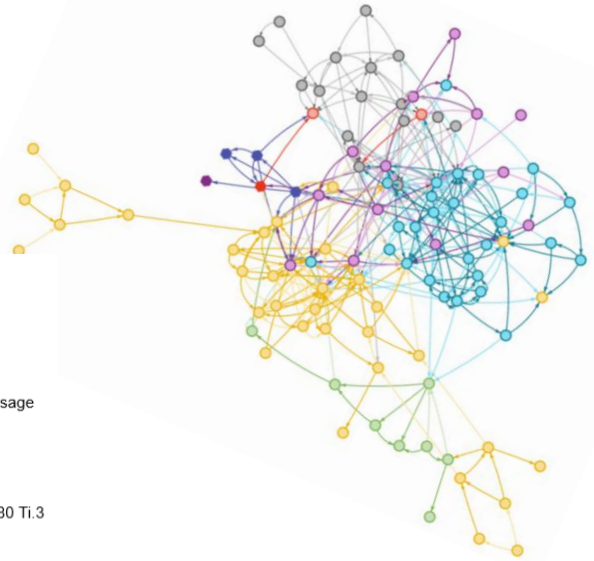
Class *CausalCloudScape(metric_data)*:
  def *feature_engineering(n_lags)*:
    for lag in range (n_lags):
      return pearsson, kendalltau, spearmanr
  def *feature_extraction(threshold)*:
    until threshold:
      rdtf(most important features)
  def *causal_discovery(MLP_dims, lambda)*:
    run DAGMA, return output graph as g_n

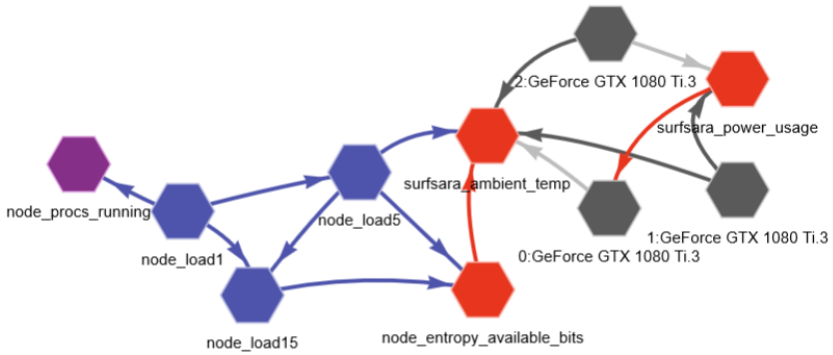
Causal discovery cloud performance metrics

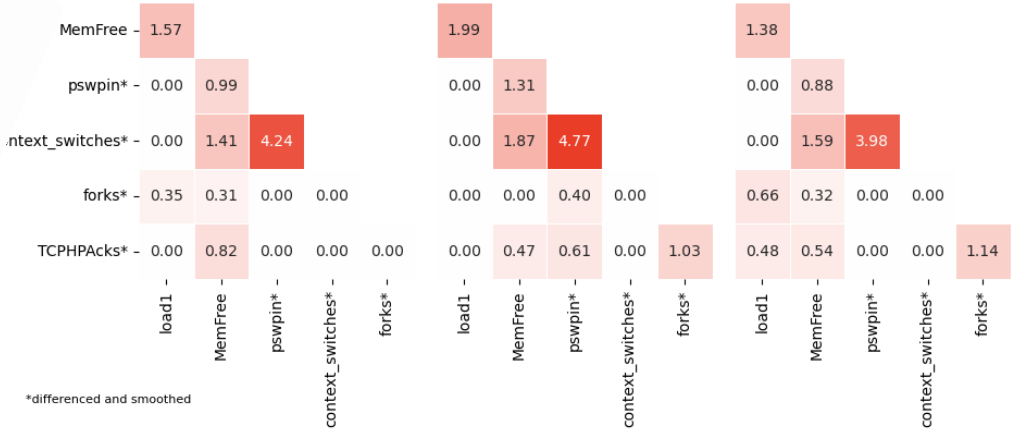








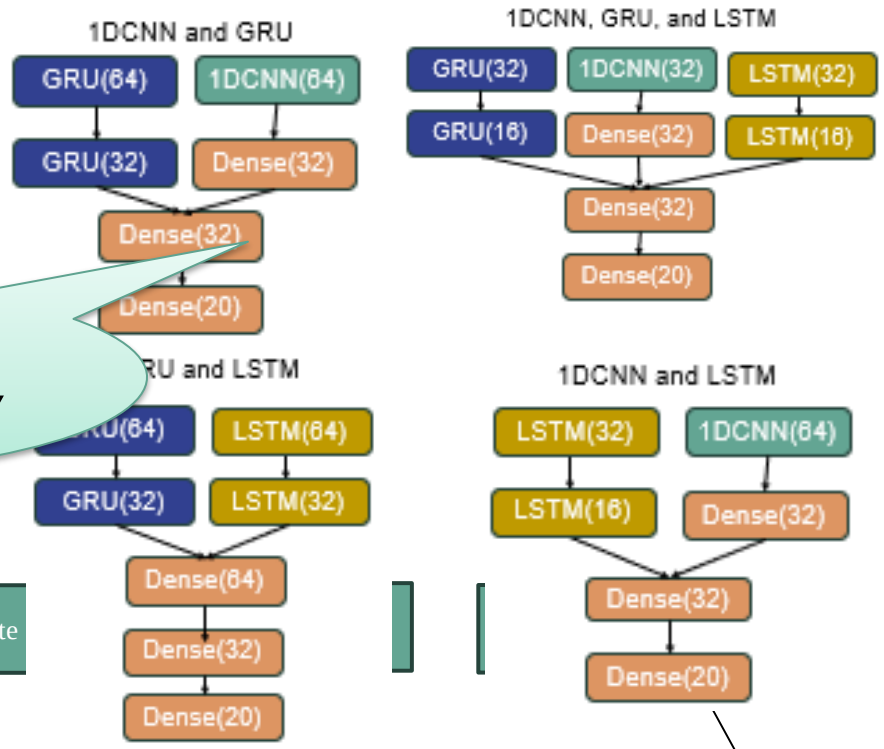


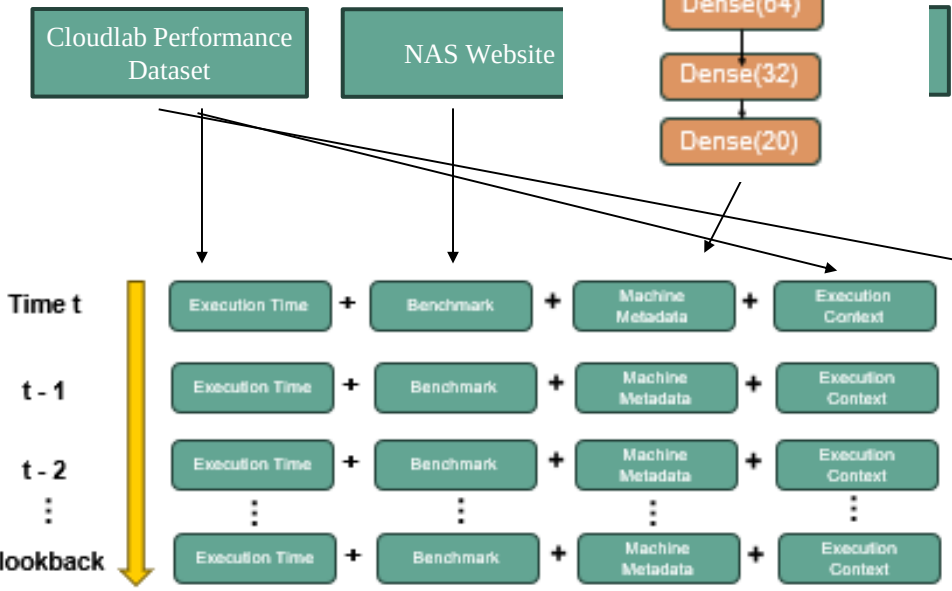


From colleagues of your program!

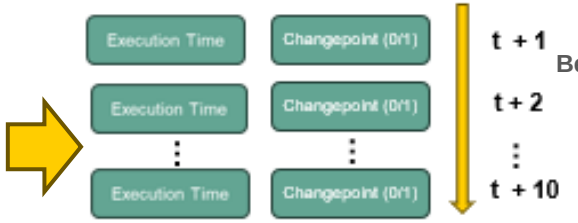
Utrecht University

Forecasting the Cloud performance variability

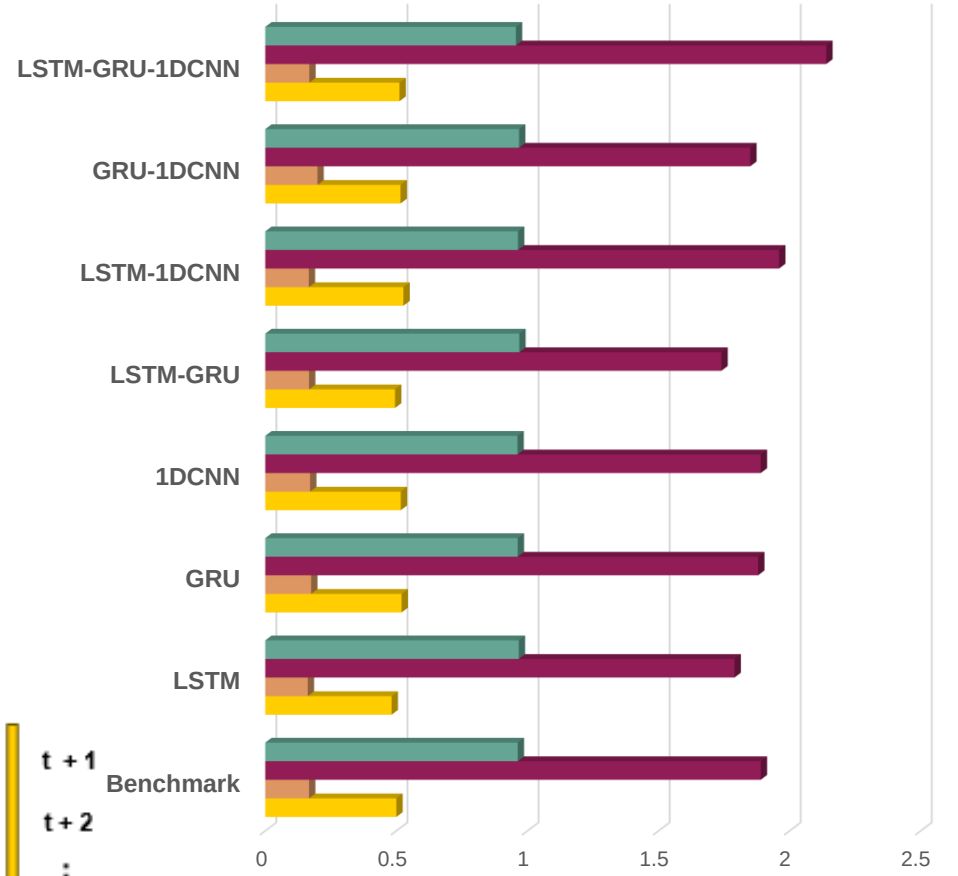




Restructured Input



Restructured Output



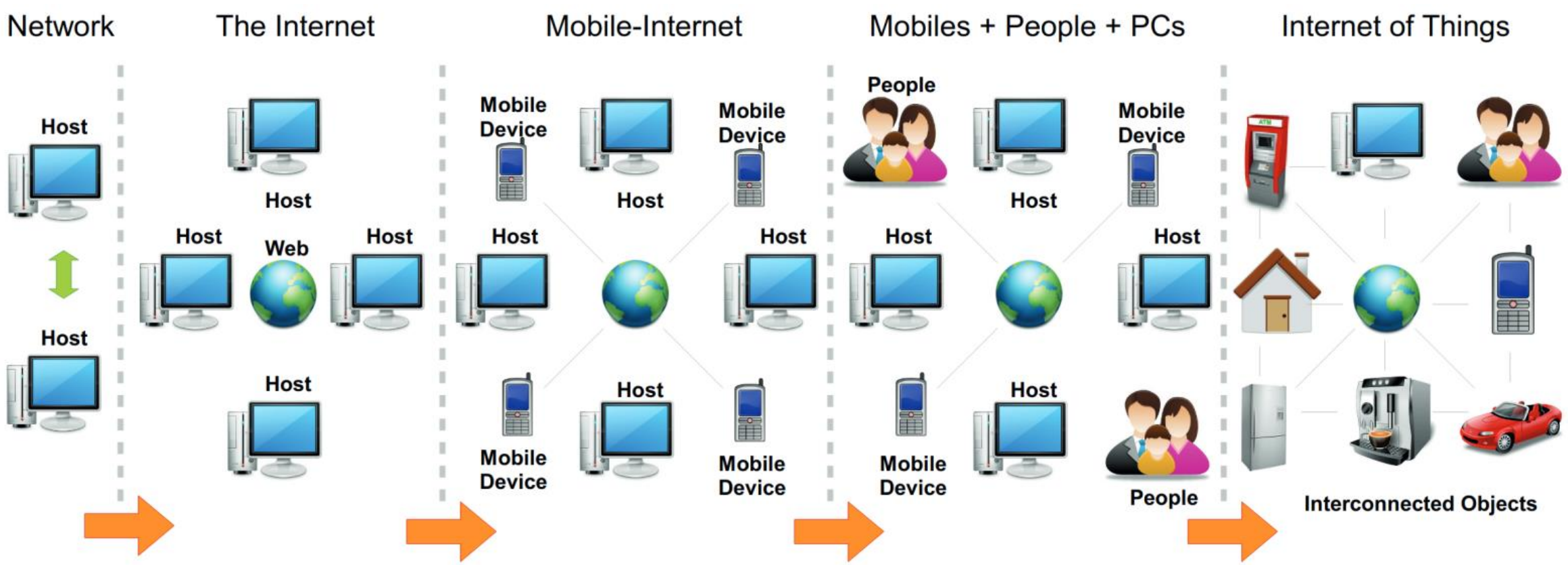
	Bench mark	LSTM	GRU	1DCNN	LSTM-GRU	LSTM-1DCNN	GRU-1DCNN	LSTM-GRU-1DCNN
Execution Time - R^2	0.963	0.967	0.963	0.962	0.969	0.964	0.968	0.956
Execution Time - RMSE	1.89	1.79	1.88	1.89	1.74	1.96	1.85	2.14
Changepoint-Custom	0.167	0.162	0.175	0.171	0.167	0.166	0.199	0.168
Changepoint - ROCAUC	0.5	0.482	0.52	0.517	0.495	0.527	0.516	0.512

Utrecht University

Are Cloud-based Big Data Service Models good enough to tackle future data processing challenges ?

Utrecht University

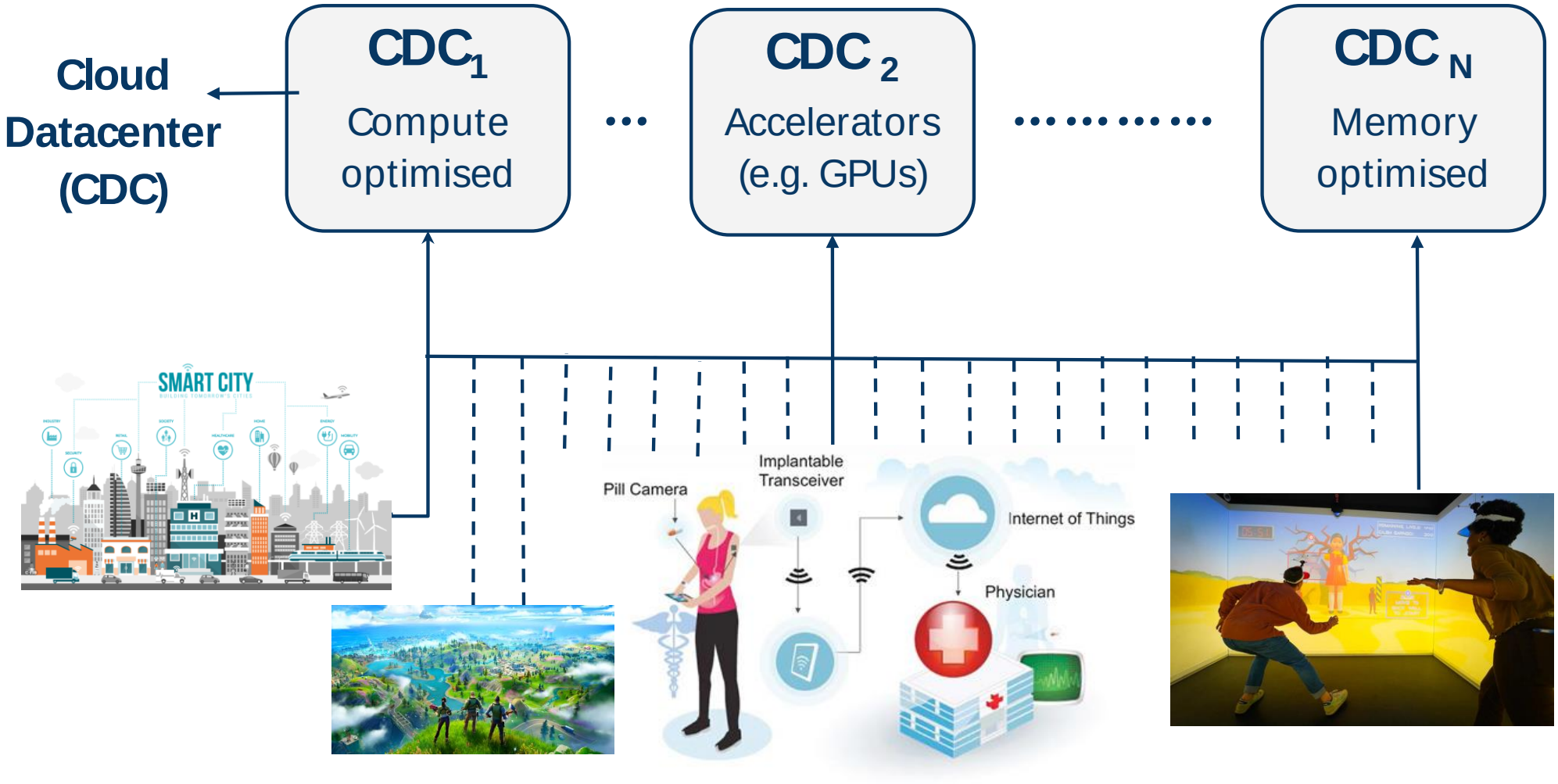
Evolving Internet



Perera et al. Context-aware Internet of Things: A survey. IEEE communications surveys and tutorials, 16(1), 414-454.

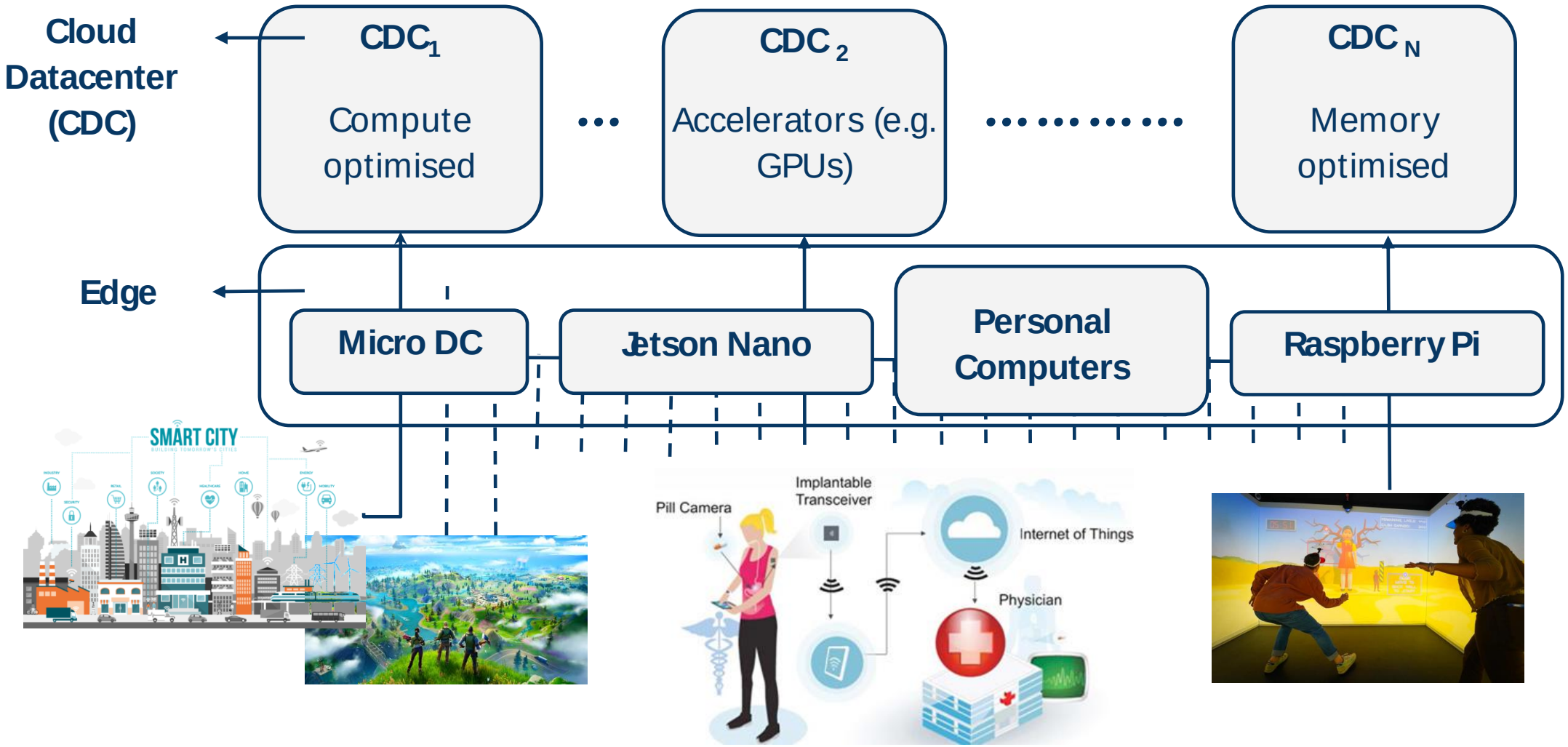
Utrecht University

Cloud Computing Challenges

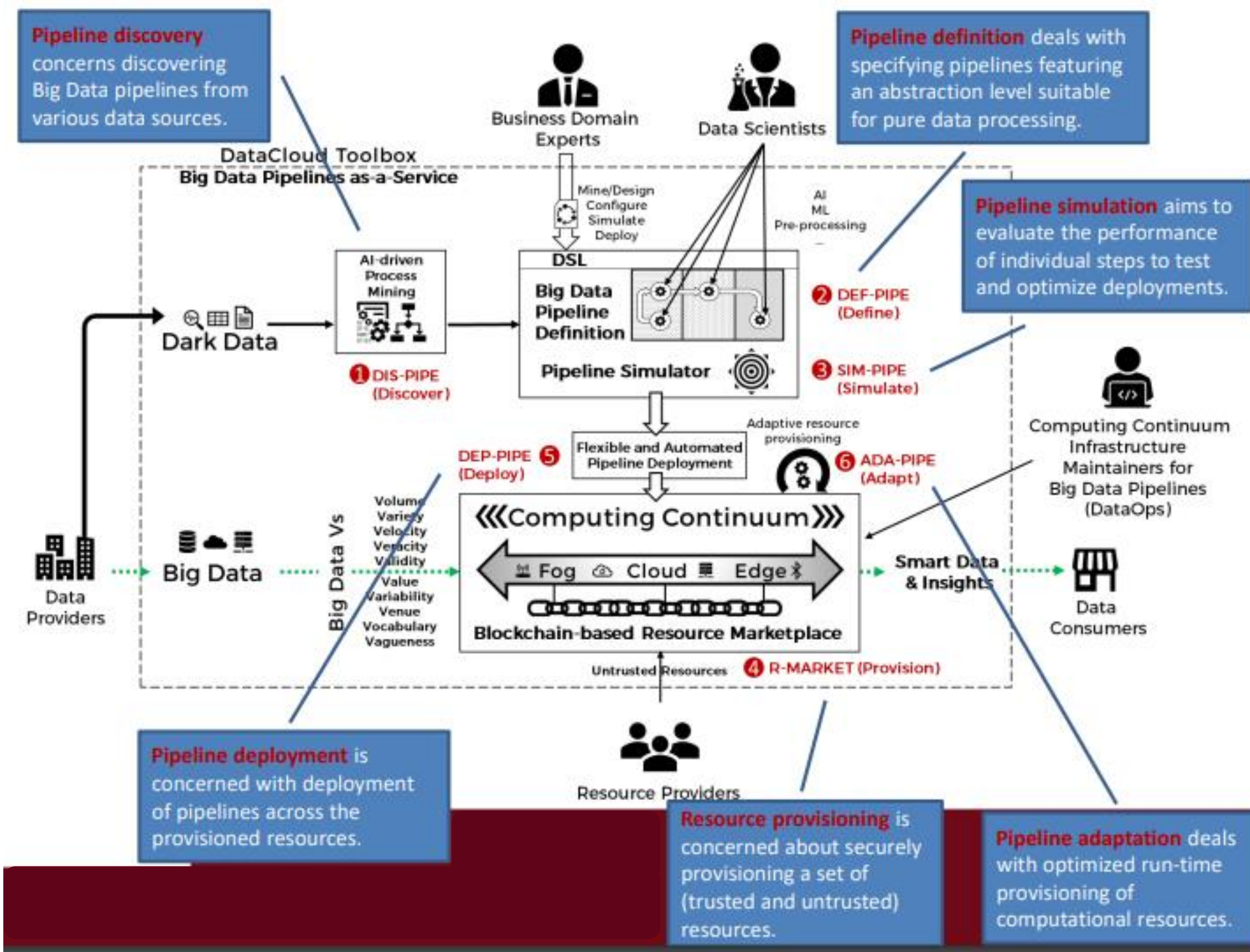


Utrecht University

Moving beyond Cloud to the Edge



Utrecht University



Utrecht University

Additional challenges: Evolution towards Computing Continuum

Confidentiality

- Will the sensitive data stored on a cloud remain confidential?
- Will Cloud compromise or leak confidential data?

Integrity

- Is Cloud provider doing the computations correctly?
- Is Cloud provider storing the data without tampering it?

Availability

- What is a Cloud provider is attacked in a denial of service attack (DoS)?
- Would Cloud scale enough? Multi-cloud ? Interoperability?

Privacy

- Massive data mining to get large information on clients?

Legal compliance and transitive trust

- Who is responsible for complying with regulations?
- If a Cloud provider subcontracts to a third party Edge providers, will data be still secure?

Utrecht University

Take-home Message

In a Data-driven world as of Today, Data Analytics involves more than just applying some eminent data operations!

Utrecht University

DISCLAIMER

The information in this presentation has been compiled with the utmost care, but no rights can be derived from its contents.