Aalto University
School of Science

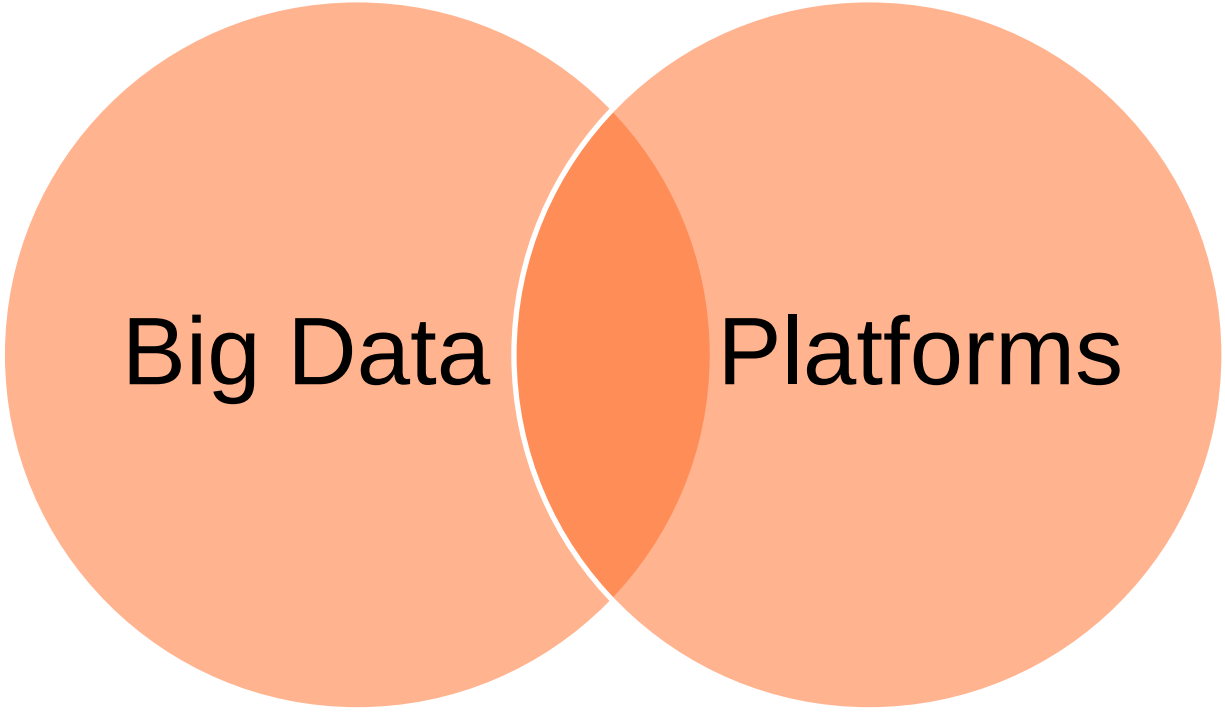
Introduction to Big Data Platforms

Hong-Linh Truong
Department of Computer Science
linh.truong@aalto.fi, https://rdsea.github.io

Learning objectives

- Understand “big data” and “platforms” in big data platforms
- Capture high-level views of big data platforms and understand the role of big data platforms
- Understand key aspects in studying big data platforms

What are they?



Big Data

Data: facts, responses, events, measurement, etc.


```
{
"station_id":"1160629000",
"datapoint_id":122,
"alarm_id":310,
"event_time":"2016-09-
17T02:05:54.000Z",
"isActive":false,
"value":6,
"valueThreshold":10
}
```


Source: https://toddwschneider.com/posts/analyzing-1-1-billion-nyc-taxi-and-uber-trips-with-a-vengeance

Is it big?

From a network infrastructure monitoring

5M sensors/monitoring points with
~1.4B events/day~ 72GB/day

Is it big ?

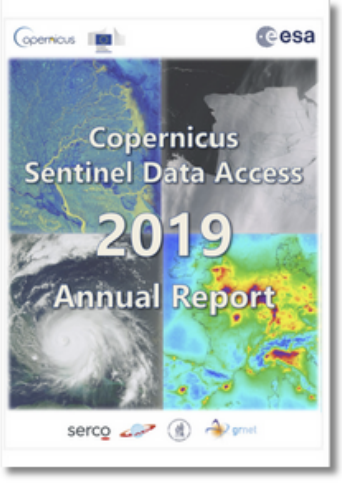
From earth observation/remote sensing

Annual Report 2019 — annual reports archive

This report for 2019 follows directly on from the 2018 report, and analyses the uptake of Copernicus Sentinel data and the performance of the Sentinel Data Access System during the period 1 December 2018 to 30 November 2019 (referred to as Y2019).

By the end of the reporting period, the Sentinel Data Access System was supporting over 280,000 registered users, a daily publication rate of over 30,500 products/day, and an average daily download volume of 214 TiB. A total of 254 million products had been downloaded by users since the start of data access operations, consisting of a total data volume of 158.4 PiB. Over half of these downloads - 128 million - occurred during Y2019 alone. The report provides the detailed statistics behind these numbers, as well as examining the demographics of users, the status of agreements with collaborative and international partners, the challenges and solutions found by the Data Access Operations team in publishing and disseminating such huge volumes of data and evolving the System to cope with them, and the outlook for the future.

The 2019 Copernicus Sentinel Data Access Annual Report is available here .



Source: https://scihub.copernicus.eu/twiki/do/view/SciHubWebPortal/AnnualReport2019

Now you know what does it mean “big data”?

Big Data

- **Extremely large, complex data sets**
  - they need to be handled with new techniques
- **Individual data items can be small or big**
  - e.g., simple sensor events versus high quality satellite images
- **Often characterized by V***
  - e.g., Volume, Variety, Velocity, and Veracity

Characterize big data with V*

- **Volume:**
  - big size, large data set, massive of small data
- **Variety:**
  - complex, different formats, types of data and their links
- **Velocity:**
  - generating speed, data movement speed
- **Veracity:**
  - quality is very different (timeliness, accuracy, etc.)

Why do we have big data now?

- **Social media data generated by human activities**
  - Facebook, Twitter, Instagram, etc.
- **Internet of Things (IoT)/Machine-to-Machine (M2M)**
  - data generated from monitoring of equipment, infrastructures and environments
- **Advanced sciences data generated by advanced instruments**
  - Earth observation from Sentinel satellites
- **Personal and disease information (e.g., healthcare)**
- **Business-related customer data**
- **Asset management and lodging (e.g., cars, homes)**
- **Software systems (e.g., logs and test results)**

Why do we need to care?

- **Because of the values of data!**
- **Top-down: Data economy**
  - more data → more insights → better decision making → more business successes
- **Bottom-up**
  - understanding → optimizing → saving cost/creating new values
- **"The Unreasonable Effectiveness of Data" principle → with more data, the same algorithm performs much better!**

Alon Halevy, Peter Norvig, and Fernando Pereira. 2009. The Unreasonable Effectiveness of Data. IEEE Intelligent Systems 24, 2 (March 2009). http://static.googleusercontent.com/media/research.google.com/en//pubs/archive/35179.pdf

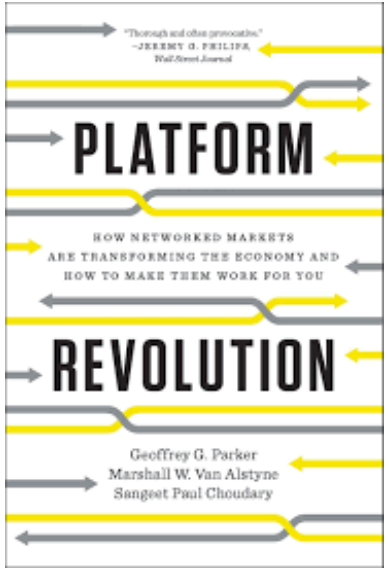
Take a vote:

https://presemo.aalto.fi/bdp

What are platforms?

Example of platforms

Let us see from the business viewpoint from “Platform Revolution”:



Disruptive platforms: Airbnb, Amazon, Uber, Alibaba, Instagram, Facebook, Youtube, etc.

https://www.amazon.com/Platform-Revolution-Networked-Markets-Transforming/

The “Platform Revolution”s definition of a platform (from a business viewpoint)

“A platform is a business based on enabling value-creating **interactions** between external producers and consumers. The platform provides **an open, participative infrastructure** for these interactions and sets **governance conditions** for them. The platform’s overarching purpose: to consummate matches among users and **facilitate the exchange of** goods, services, or social currency, thereby enabling value creation for all participation”

Source: Geoffrey G. Parker, Van Alstyne, Marshall W. Van Alstyne , Sangeet Paul Choudary, *Platform Revolution: How Networked Markets Are Transforming the Economy - and How to Make Them Work for You,* March 28, 2016

What are in a platform for big data?

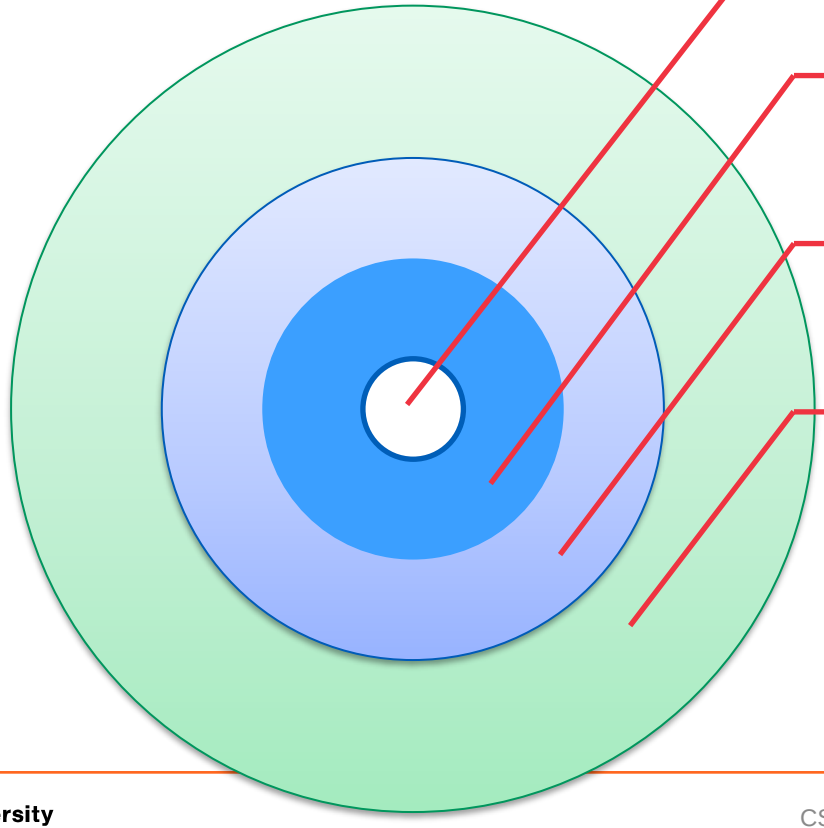
Our interpretation of platforms for big data

- **Being large-scale service platforms, e.g.**
  - On-demand computing platforms for data-centric products
  - On-demand analytics service platforms
  - On-demand data management platforms
- **Enabling interactions between big data producers and big data consumers**
  - Integration, management, analysis, optimization
- **Facilitating the exchange of big data and products centered around data**
- **Not just a database or data marketplace (even they are big!)**

Big data platforms

- **Data-centric services**
  - Data: a lot of data with different types and added continuously
  - Complex technological infrastructures
- **Extensibility: allowing new services, components to be added and integrated**
- **With diverse types of stakeholders**
  - Data consumers, data providers, and data integrators
  - Service consumers, service providers and service integrators
  - Regulators/auditors, etc.

Big data platforms: Onion architecture



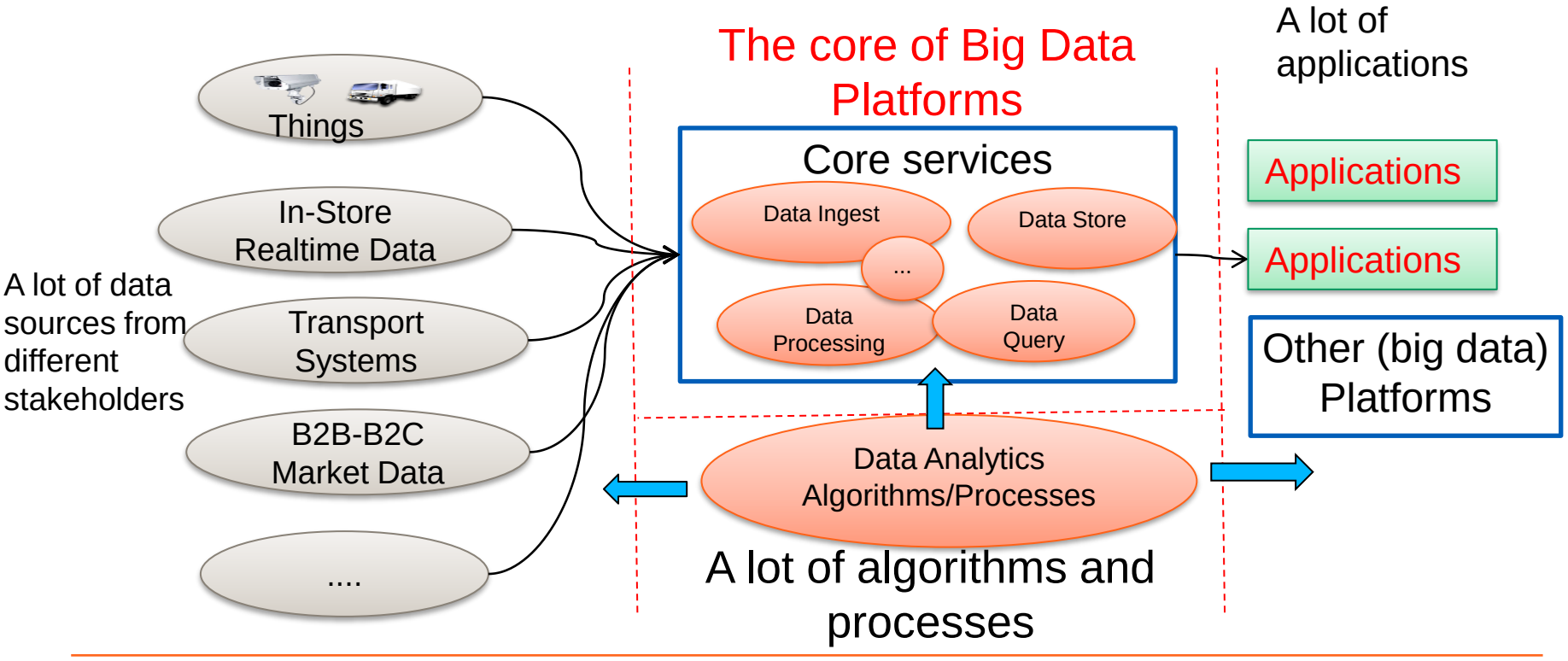
Data-centric Virtualized Infrastructures

Middleware Platforms: building, deploying, operating and scaling reliable big data services and applications

Big data services & applications

Big data platforms consumers and producers: Sensors, Things, Equipment, Industrial Processes, People

Big data platforms: system of systems view



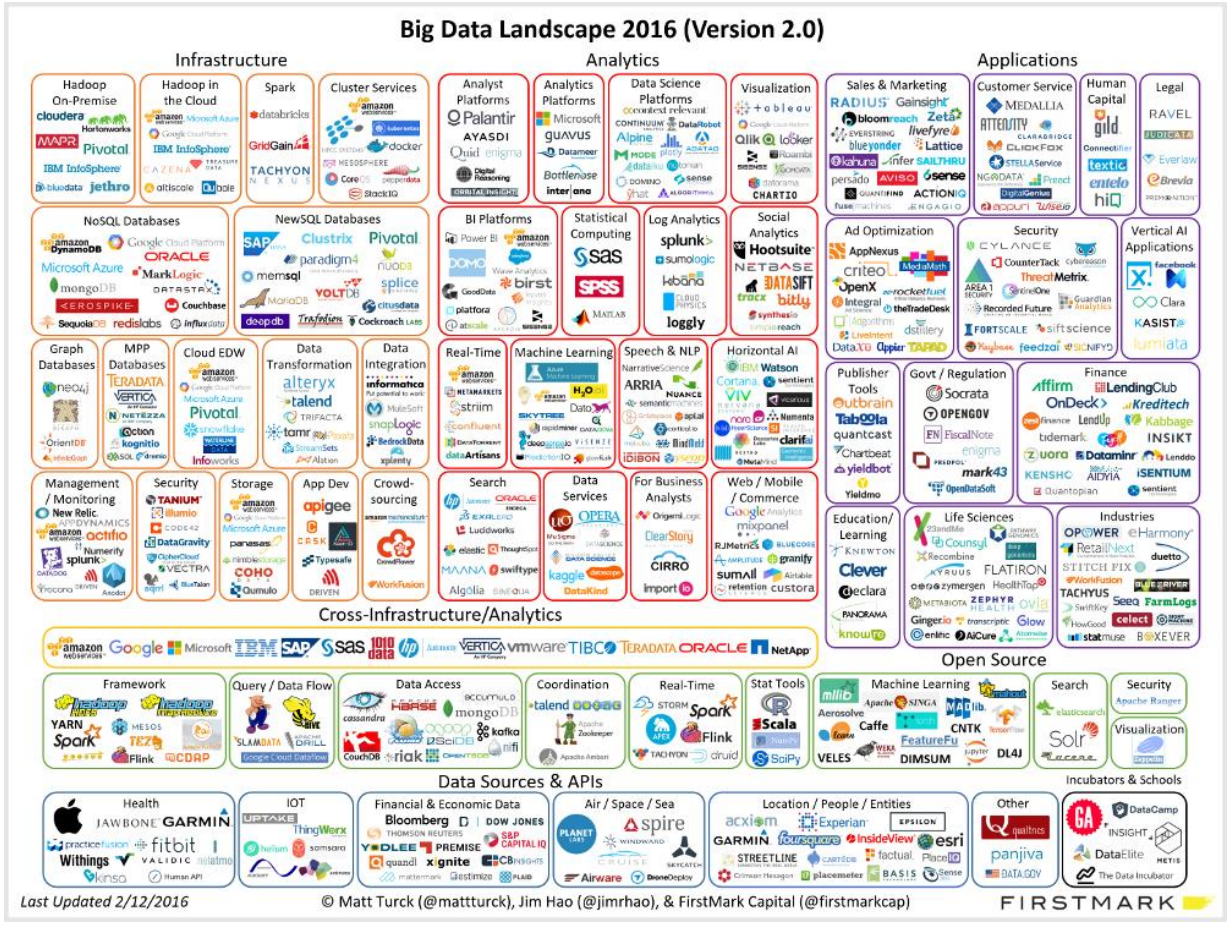
Why are big data platforms important?

- **Foundations and backbones for various "hot" areas, especially in data-driven economy**
  - Machine Learning/AI
    - *Data management, pipelines, …*
  - Data Science
    - *Data and data management, computational models, statistics and algorithm*
  - Enterprise computing
    - *360-degree analytics of customers*
  - Industrial IoT/Manufacturing/Predictive maintenance
    - *Monitor machines and optimizing machines through real-time and predictions*
  - Smart cities

Highly relevant to industries & businesses

- **Wide range application domains**
  - Business, healthcare, manufacturing, and science
- **Examples of key industries for infrastructural and platform services for big data**
  - Amazon, Microsoft, Google, IBM, Alibaba, Huawei, etc.
  - Gartner Data Management Solutions for Analytics "https://www.gartner.com/reviews/market/data-warehouse-solutions"
- **Big data and large-scale data analytics are fundamental in most companies/organizations doing digital business**

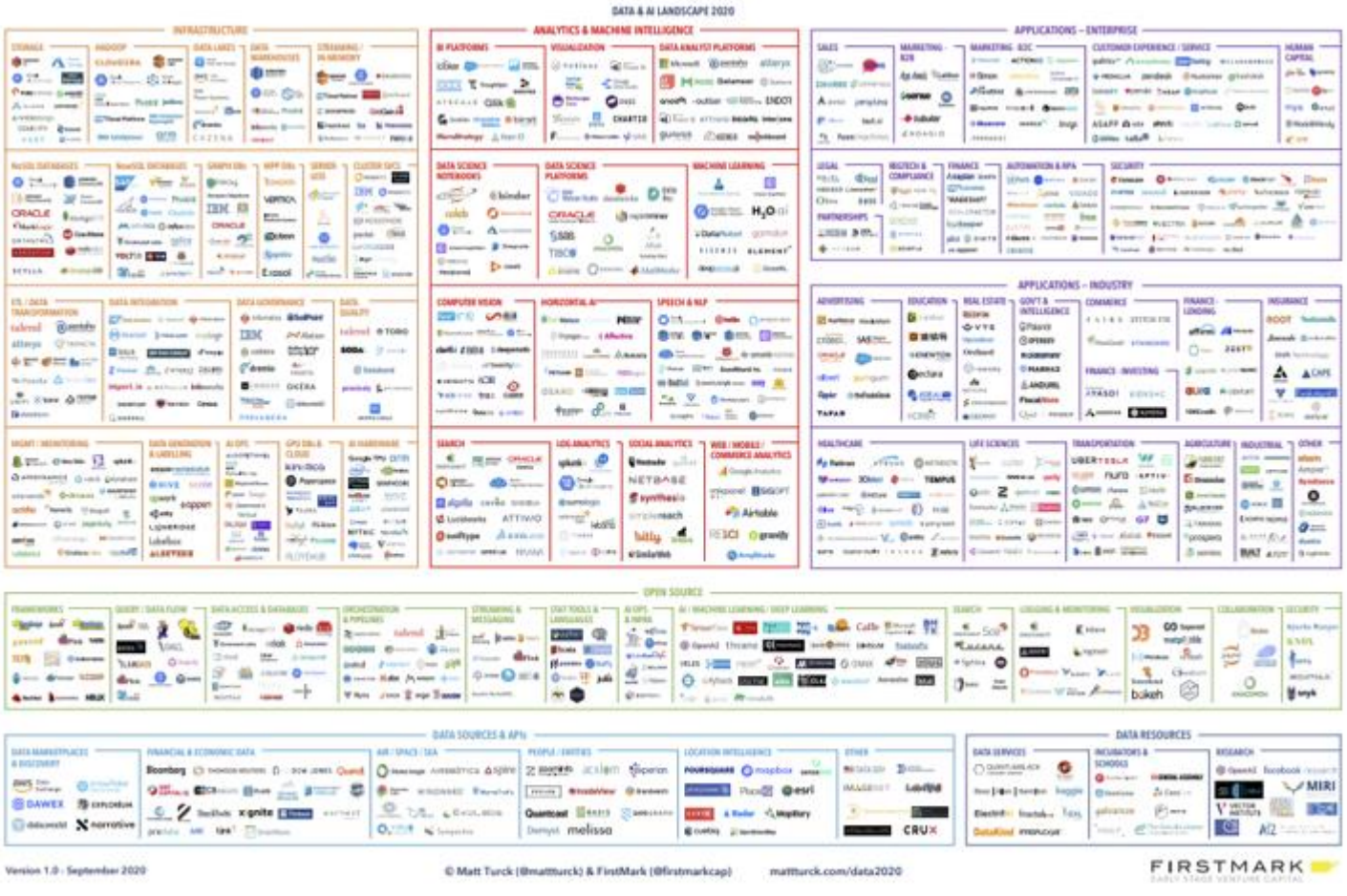
The landscape is complex for our study



Source: http://www.datameer.com/wp-content/uploads/2016/06/matt_turck_big_data_landscape_v11r.png

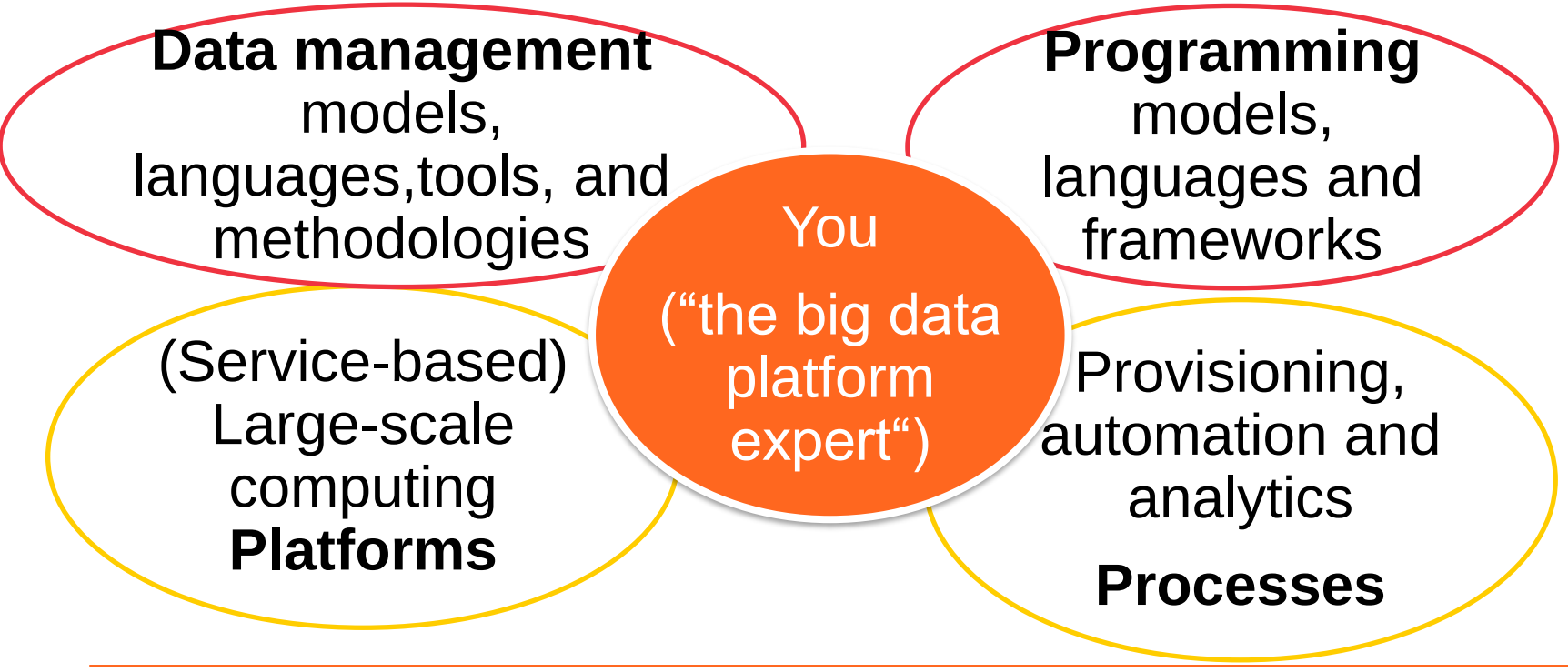
The landscape is complex for our study

the study must <u>balance</u> between concepts & practices!



Source: http://mattturck.com/wp-content/uploads/2020/09/2020-Data-and-AI-Landscape-Matt-Turck-at-FirstMark-v1.pdf

Core principles/techniques from Compute Science for Big Data Platforms



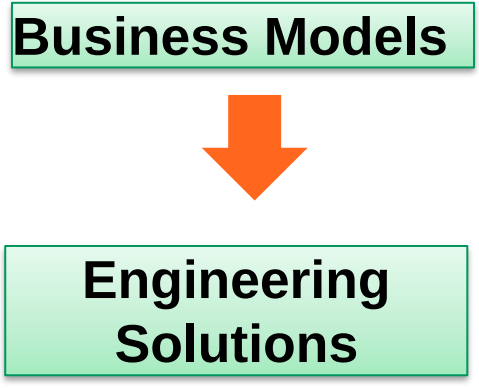
Focuses in studying big data platforms

- **Design/Development vs Operation**
- **Data-centric vs Service-centric vs Platform-centric activities**
- **High-level analytics vs low-level programming models and processes**
- **Quality and governance**

Target goals for the study

- As a user: able to program atop big data platforms
- As a provider: able to operate big data platforms
- As a designer/architect: able to design new big data platforms
- As a developer: able to develop services/applications in big data platforms

Business models vs platform engineering



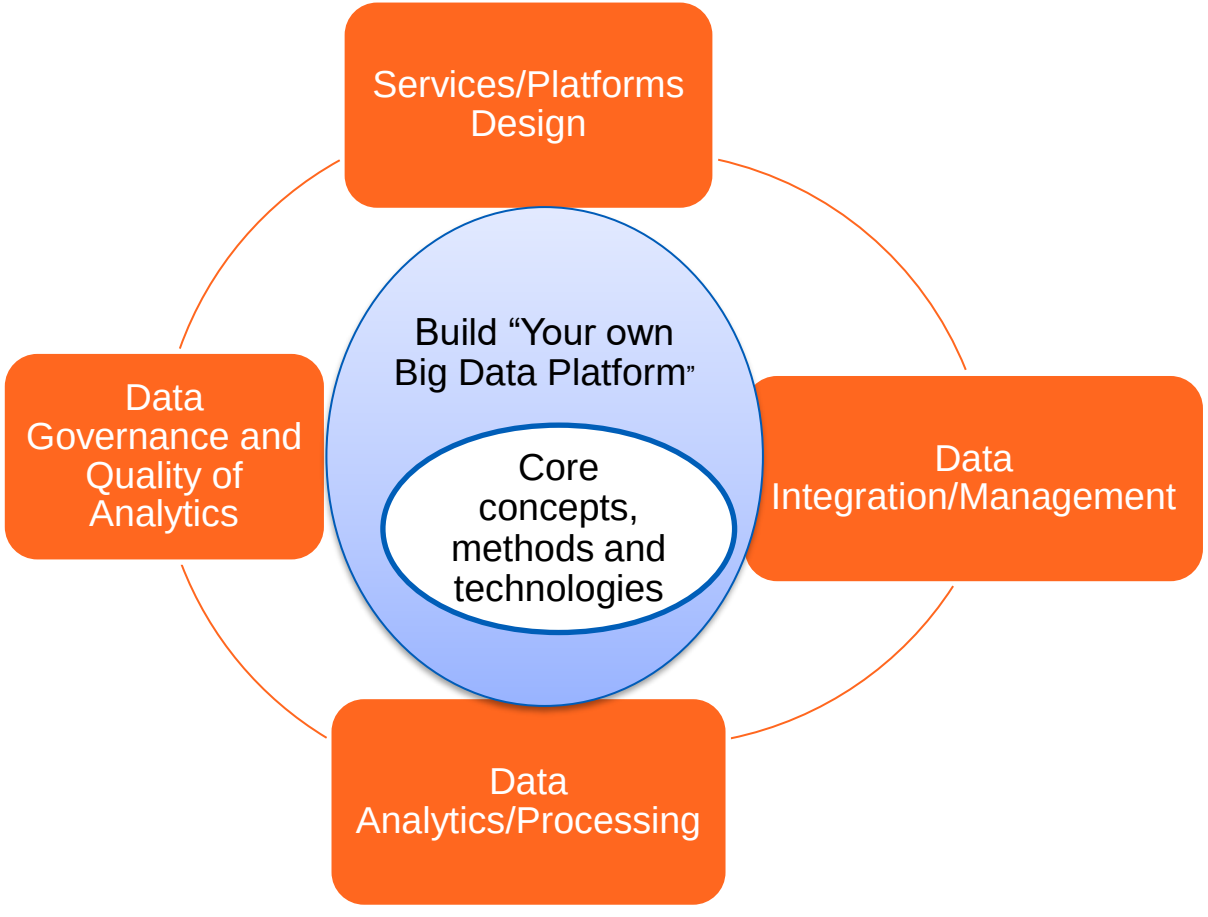
- **Distinguish between engineering and business models**
  - Engineering: design, operation and governance
  - Business models: stakeholder management, pay-per-use, pricing, tenant models
- **We mainly focus on engineering**
  - Business aspects are reflected in requirements for the engineering

Concepts/Techniques vs Technologies

- **Concepts/Techniques versus Technologies**
  - Concepts: e.g., NoSQL data models and scalable datastores
  - Technologies: e.g., Cassandra, Hadoop, Apache Spark, Google Cloud
- **We still focus mainly on concepts/techniques**
  - Technologies can be very complex or "everything is behind an API"
  - But don't forget key concepts and techniques
    - *Implement key concepts with state-of-the-art technology in a limited but realistic scenarios*

Build your story

Focus on foundations & explore your strengths/interests



Design and implementation mainly reflect activities of platform developers/providers in the lifecycle of a big data platform based on real-world data sets and scenarios

Related concepts/techniques

- **Distributed systems and cloud computing**
  - Virtualized environments and cloud deployment, concurrency, consistency/availability/fault management, application protocols
- **Databases and data management**
  - Data modeling, ETL/data pipeline, data partitioning, databases
- **Algorithms and programming models**
  - Parallel/concurrent programming, workflows, streaming processing
- **Service and software engineering:**
  - Service engineering & microservices

Feedback for today

- **Kindly do the survey, if you have not done so**

https://mycourses.aalto.fi/mod/questionnaire/view.php?id=595261

- **Kindly provide the feedback for the Introduction here**

https://mycourses.aalto.fi/mod/questionnaire/view.php?id=595262

Thanks!

Hong-Linh Truong
Department of Computer Science

rdsea.github.io