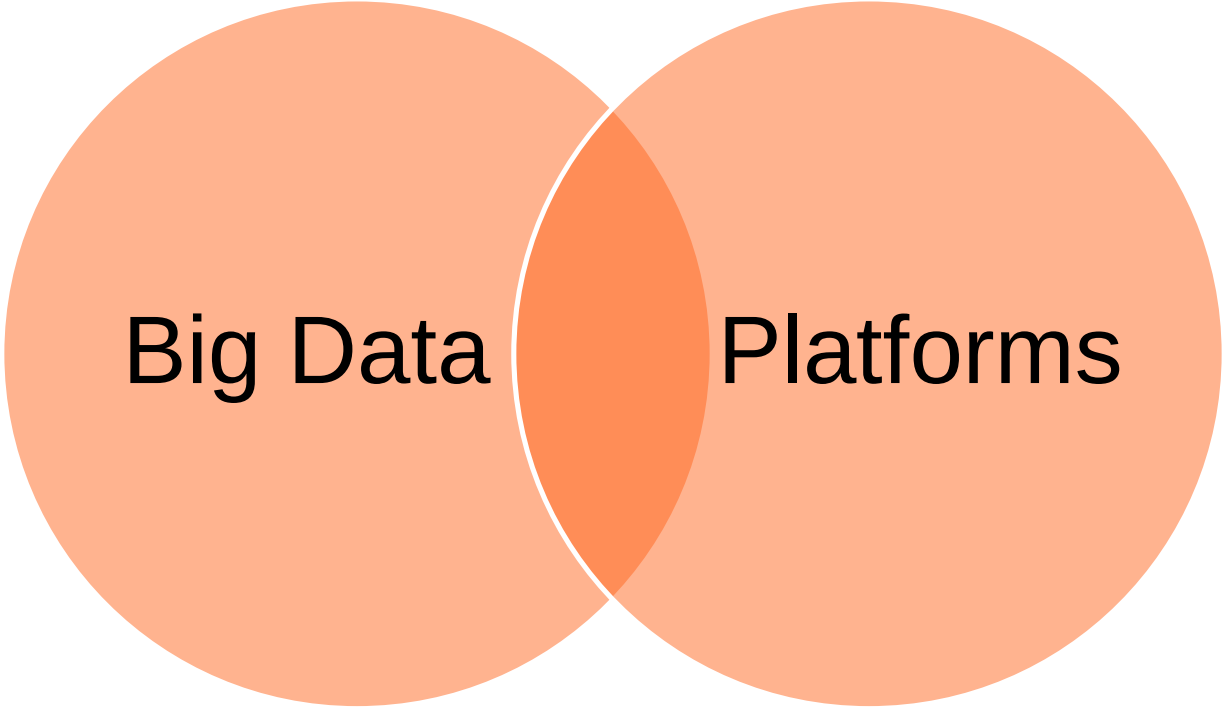
Aalto University
School of Science

Introduction to Big Data Platforms

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

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

Sentinel Data Access Annual Report 2017

18 May 2018

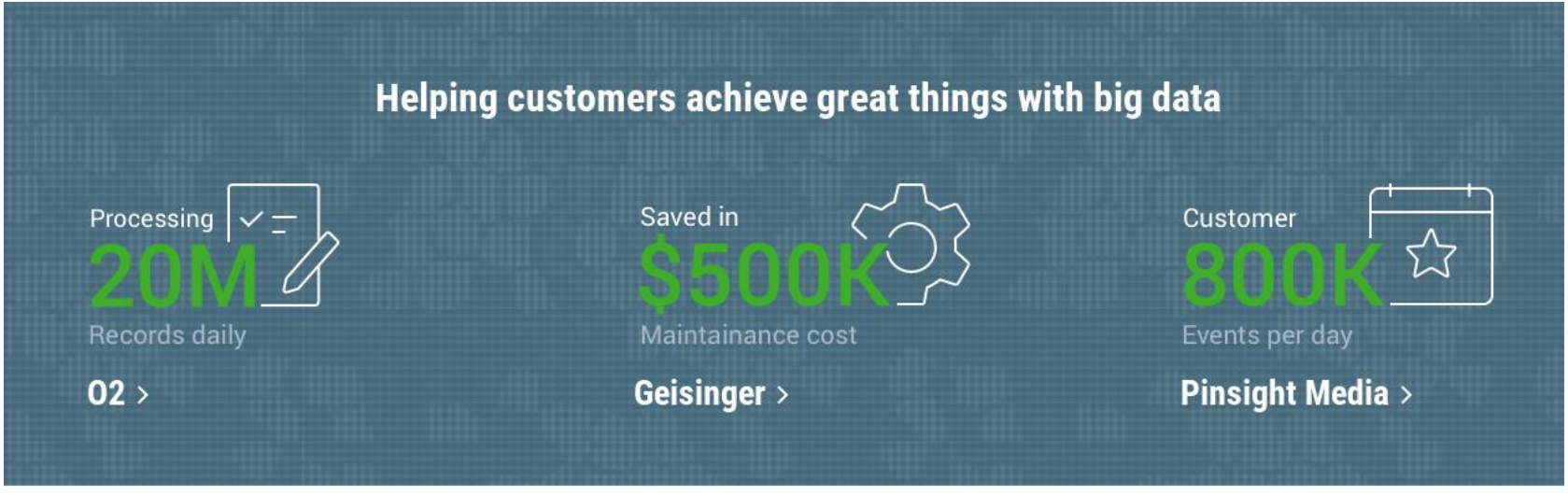
We are pleased to publish the third Sentinel Data Access Annual Report. This 2017 report takes up from where the 2016 report left off, and analyses the uptake of Copernicus Sentinel data and the performance of the Sentinel Data Access System during the period 01 December 2016 to 30 November 2017.

By the end of the reporting period, the Sentinel Data Access System was supporting 109,690 registered users, a daily publication rate of 10.04TB/day, and an average daily download volume of 93.5TB. A total of 50,964,670 products had been downloaded by users since the start of data access operations, with a total data volume of 41.35 PB.

Source: https://sentinel.esa.int/web/sentinel/news/-/article/sentinel-data-access-annual-report-2017

Is it big?

From a company doing business in big data



Snapshot from https://hortonworks.com/, 12.07.2019

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

Big Data

- **Extremely large, complex data sets**
  - they need to be handled with new techniques
- **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:**
  - complexity of different formats and types of data
- **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 in the Internet**
  - Facebook, Twitter, Instagram, etc.
- **Internet of Things (IoT)/Machine-to-Machine (M2M)**
  - data generated from monitoring of devices, infrastructures and environment
- **Advanced sciences data generated by advanced instruments**
  - e.g. earth observation from Sentinel satellites
- **Personal information (e.g., healthcare)**
- **Open (and transparent) government**
- **Asset Management of cars, homes, etc.**
- **Software systems**

Why do we need to care?

- **Because of the values of data!**
- **Top-down: Data economy**
  - More data → more evidences → more business successes
- **Bottom-up**
  - Optimizing → saving cost/creating new values
- **"The Unreasonable Effectiveness of Data" (Alon Halevy, Peter Norvig, and Fernando Pereira) → with more data, the same algorithm performs much more 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

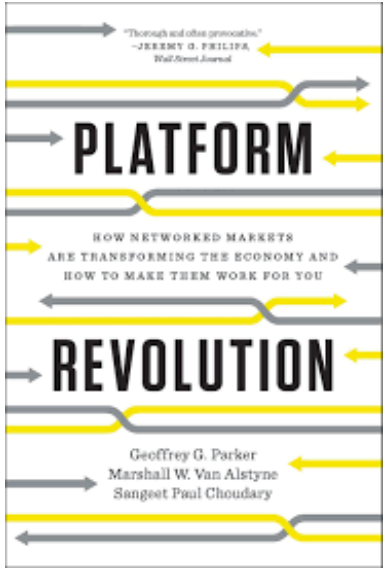
Take a vote:

https://bit.ly/2mbo1sV

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?

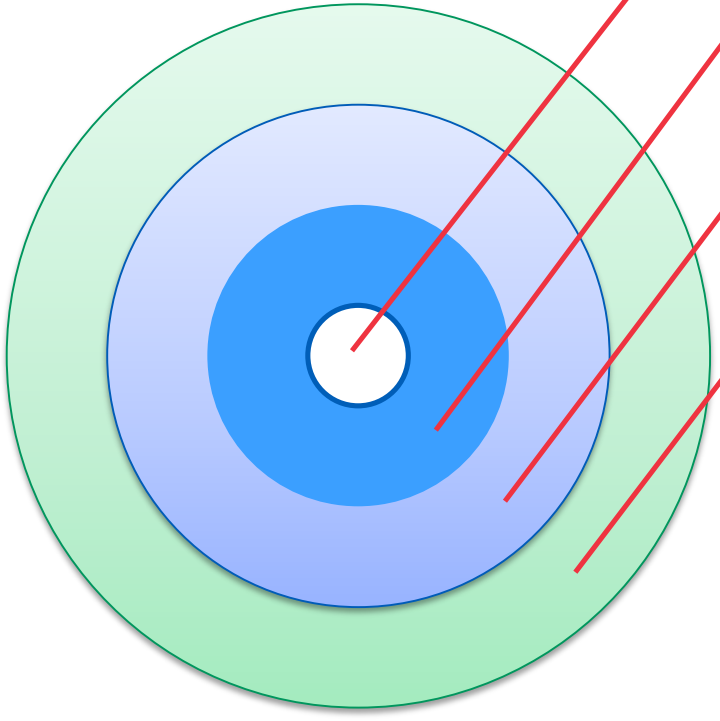
Our interpretation of platforms for big data

- **Digital platforms, e.g.**
  - On-demand computing platforms
  - On-demand analytics service platforms
  - On-demand data platforms
- **Enabling interactions between big data producers and big data consumers**
  - Integration, Management, Analysis, Optimization
- **Facilitating the exchange of big data and services centered around data**
- **Not just a database or data marketplace (even very big!)**

Big data platforms

- **for data-intensive applications**
  - Data: a lot of data and added continuously
  - Complex technological infrastructures
- **extensible 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



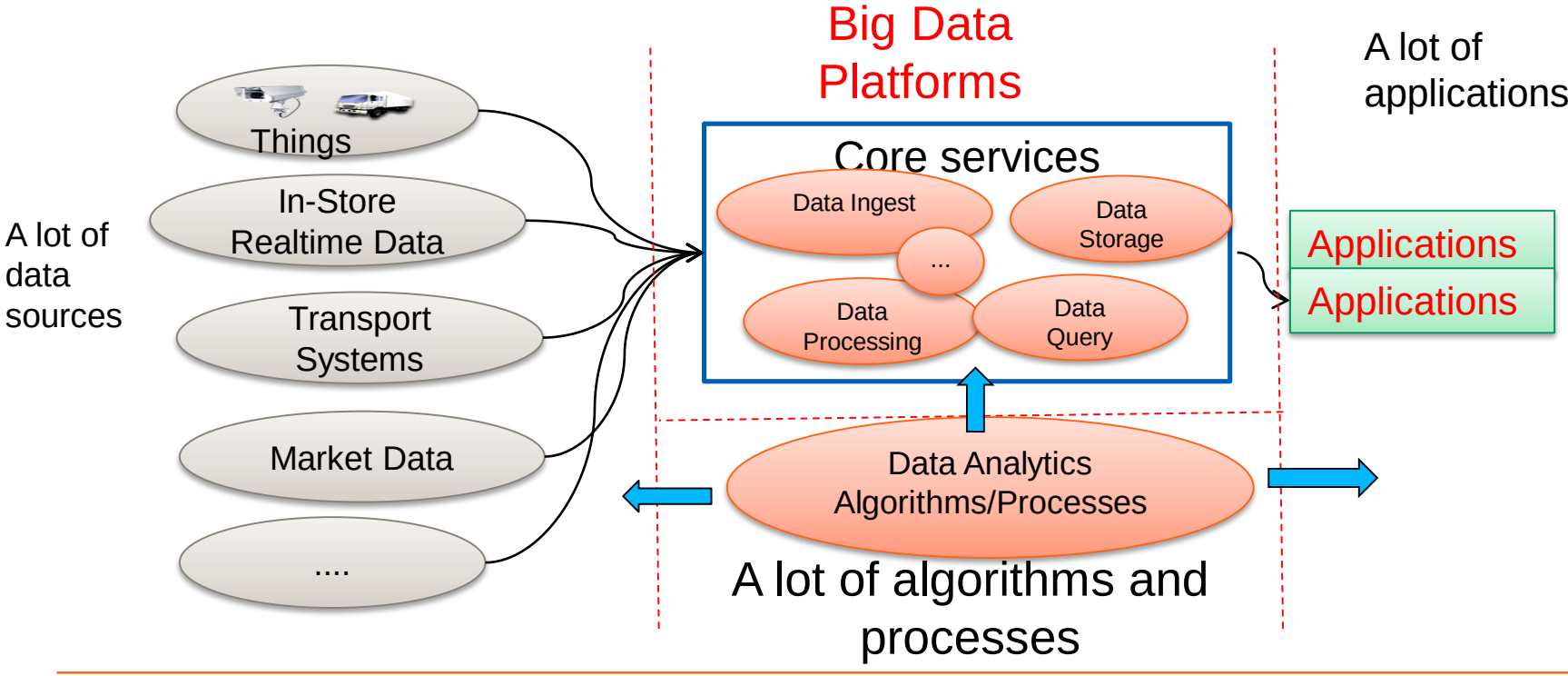
Data-centric Virtualized Infrastructures

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

Big data services & applications

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

A bird view of big data platforms

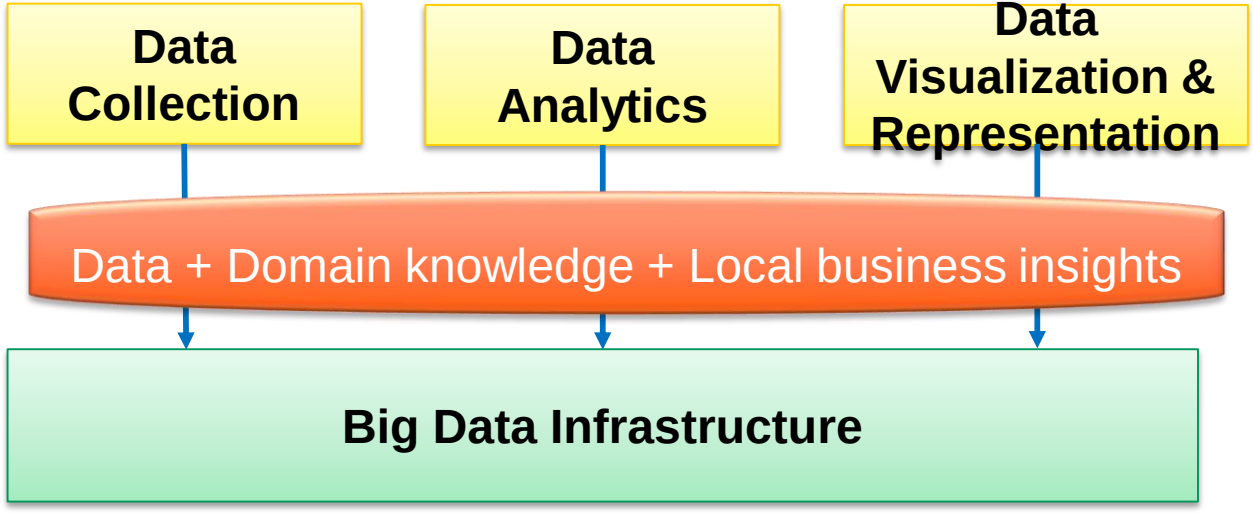


Why are big data platforms important?

- **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

Example: relationship between data science and big data platforms

Big data platform provide scalable computing infrastructures, data management, middleware, and tools

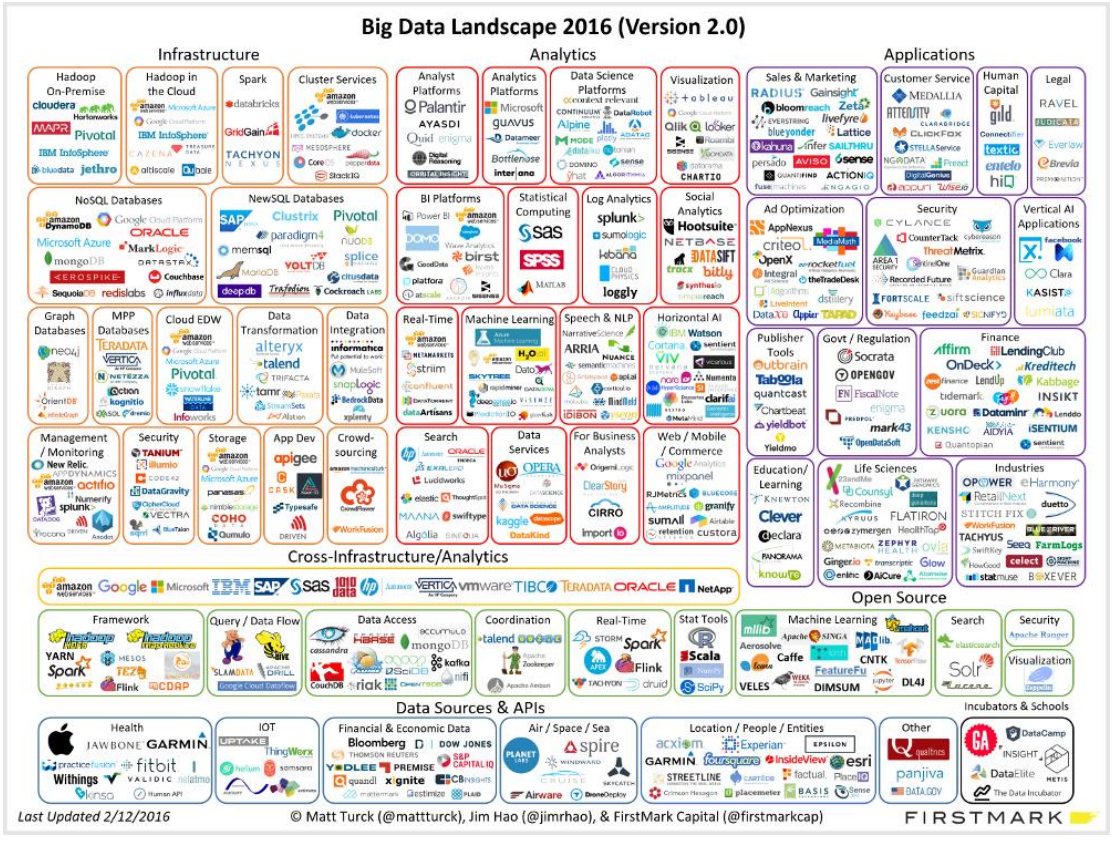


Highly relevant to industries

- **Wide range application domains**
  - Business, Healthcare, Technology, and Science
- **Examples of key industries for infrastructural and platform services for big data**
  - Amazon, Microsoft, Google, IBM, 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**

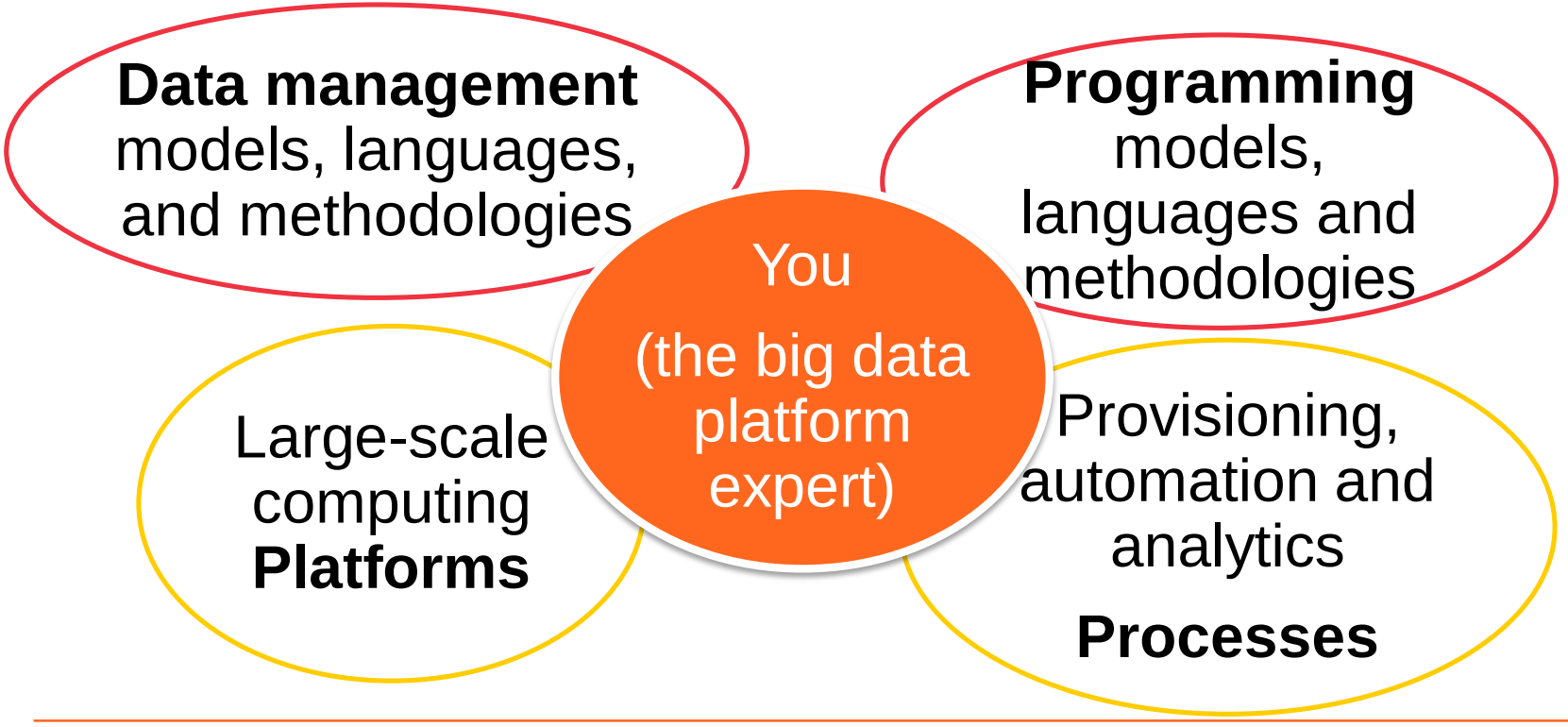
The landscape is complex

We must <u>balance</u> between concepts & implementation!



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

Big data platforms need a right skill set



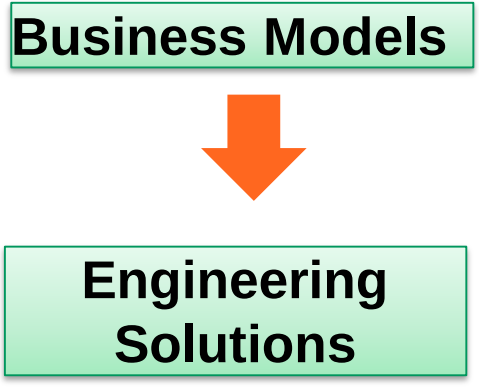
Focuses in studying big data platforms

- **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 - full Devs and Ops roles

- 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

Service/Platform: Business Models vs Engineering



- **Distinguish between Engineering and Service Models**
  - Engineering: design, operation, governance
  - Business Models: business aspects
- **Mainly focus on engineering**
  - It is nice if you can have business aspects and reflect them int the engineering to show their relationships

Concepts/Techniques vs Technologies

- **Concepts/Techniques versus Technologies**
  - Concepts: e.g., data models and existing NoSQL 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*

Course content

- **Lectures**

https://mycourses.aalto.fi/course/view.php?id=24363&section=1

- **Tutorials**

https://mycourses.aalto.fi/course/view.php?id=24363&section=3

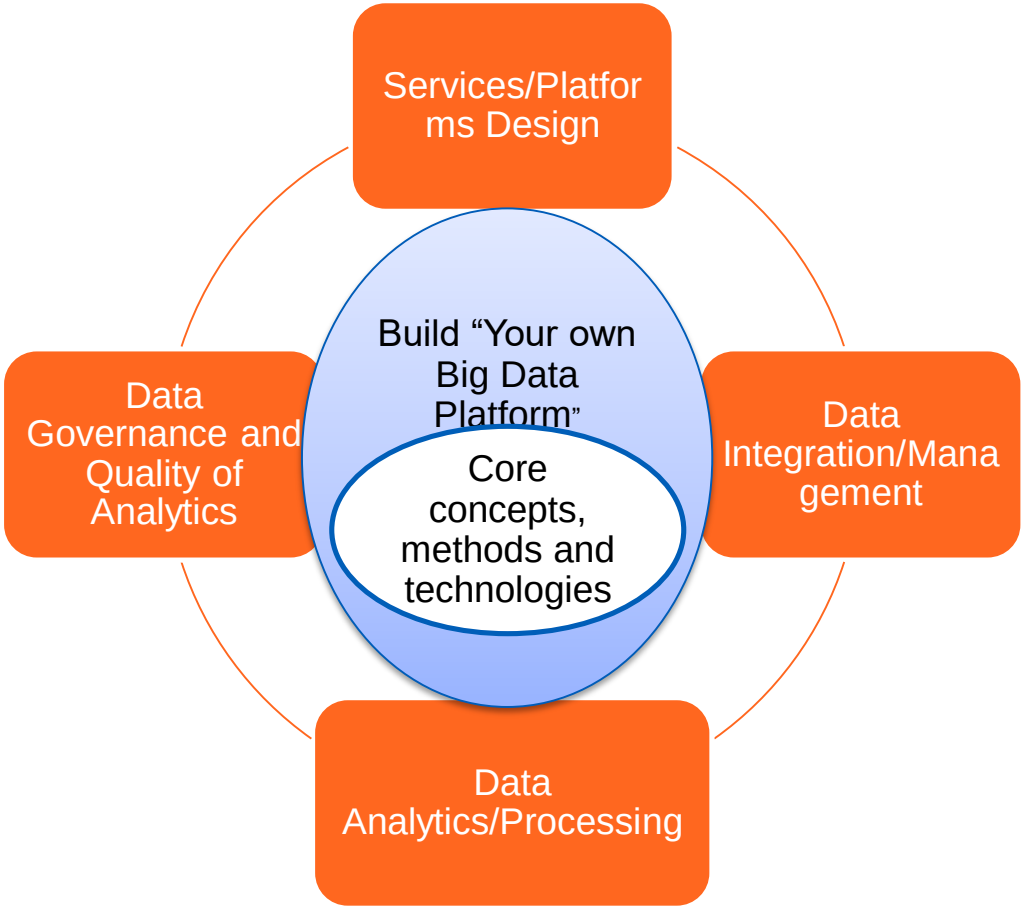
“I don’t take computer science major! ”

- **Not all of you need everything**
  - Just want to learn analytics atop big data platforms?
    - *E.g., too much “systems” in Big Data Platforms!*

- **So what would be the best strategy to learn this course?**

Build your story

Focus on foundations and explore your strengths/interests



Assignment topics reflect activities of different stakeholders in the lifecycle of a big data platform

But I cannot tell you the content now (why should we know the content now? Let us move on with our story and see)

Related courses

- **Cloud computing and Virtualization**
  - OS, Concurrent Programming, Mobile Cloud Computing, Application protocols
- **Databases and data management**
  - Data modeling, ETL
- **Algorithms and programming Models**
  - Parallel programming, concurrent programming
- **Service and software engineering:**
  - Topics: microservices, serverless, DevOps, testing

Feedback for today

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

https://mycourses.aalto.fi/mod/questionnaire/view.php?id=480269

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

https://mycourses.aalto.fi/mod/questionnaire/view.php?id=480267

Thanks!

Hong-Linh Truong
Department of Computer Science

rdsea.github.io