Big Data Insights for Networks

Françoise Soulié Fogelman 苏丽叶

Agenda

- What is Big Data?
- The Big Data process
- Impacts for networks ?

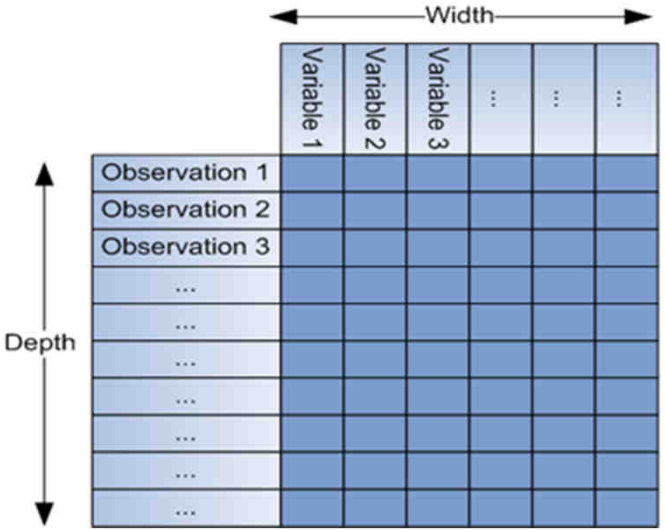
What is Big Data?

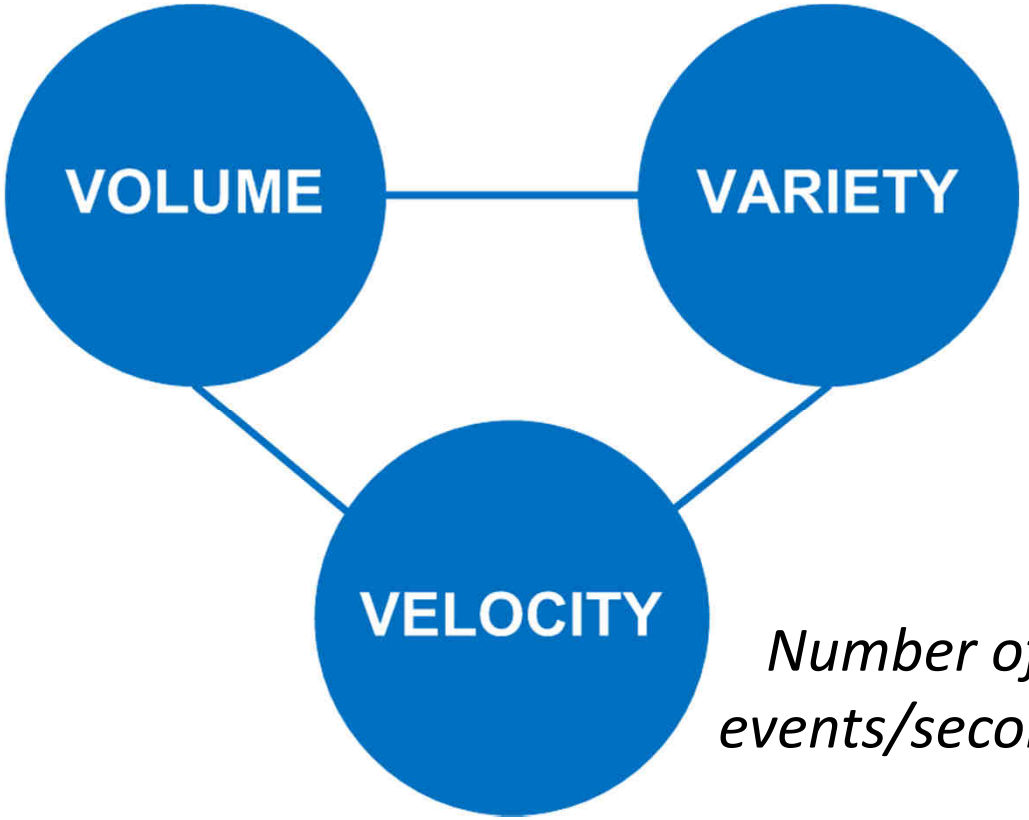
Big Data

- The concept was introduced in 2001
  - Doug Laney (Meta Group/Gartner)

Number of variables / sources

Number of observations x Number of variables





Number of events/second

http://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf

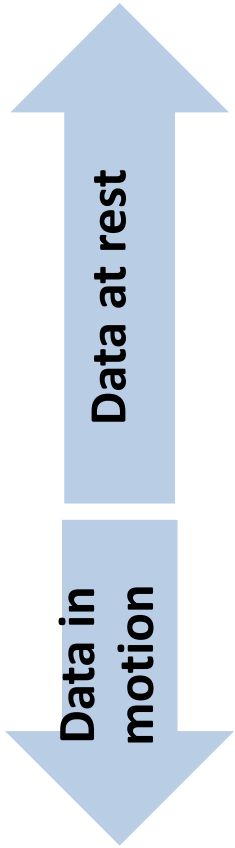
Data Variety

- Data collected is increasingly varied & non-structured
  → **Large Variety**

Structured Data	• Data bases • Data files
Unstructured Data	• Text • Tags
Sensor Data	• RFID • Temperature, pressure, acceleration, GPS … • Network sensors
New data types	• Video, photo, image, Voice, audio … • Social

Data Velocity

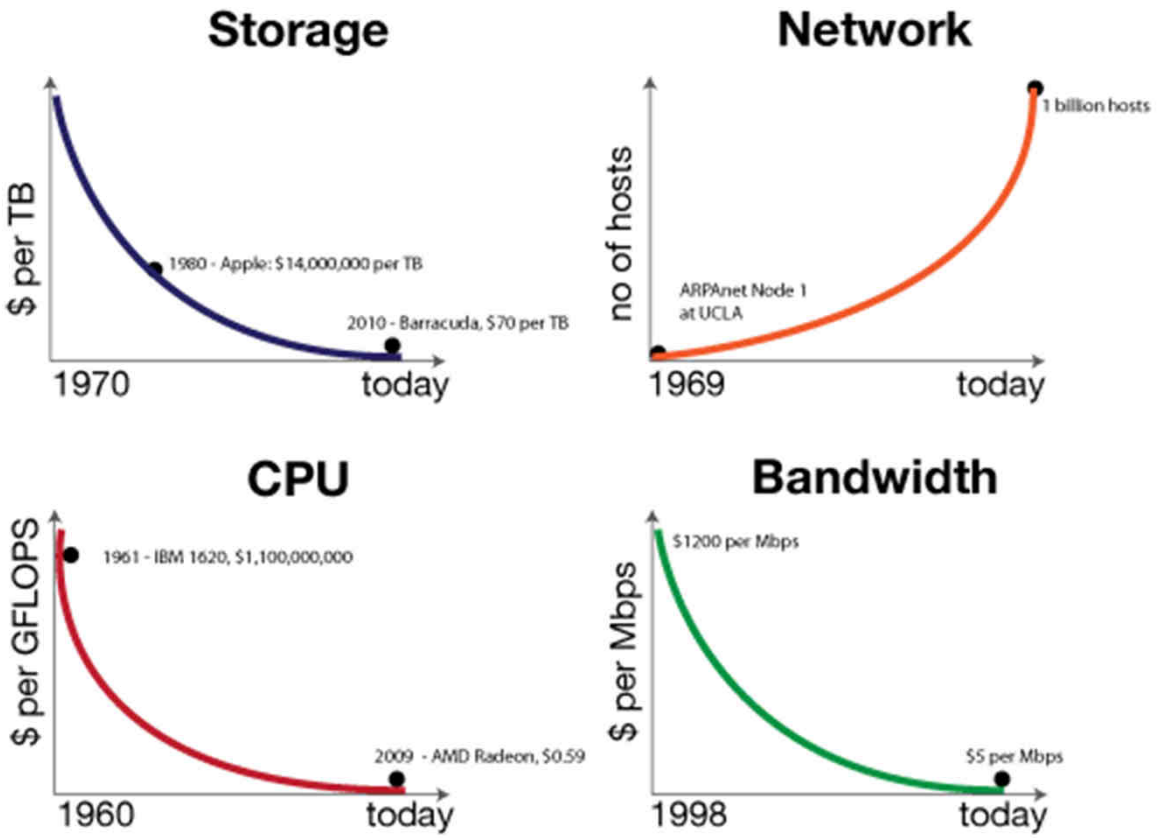
Data produced is coming increasingly fast → **Large Velocity**



Batch Data	• Data collected in bulk from relatively static data sources
Near time Data	• Data collected with delay from dynamic data sources
Real time Data	• Data collected & processed from dynamic data sources
Streams	• Data processed from very dynamic data sources. Can hardly be collected

How can we handle Big Data?

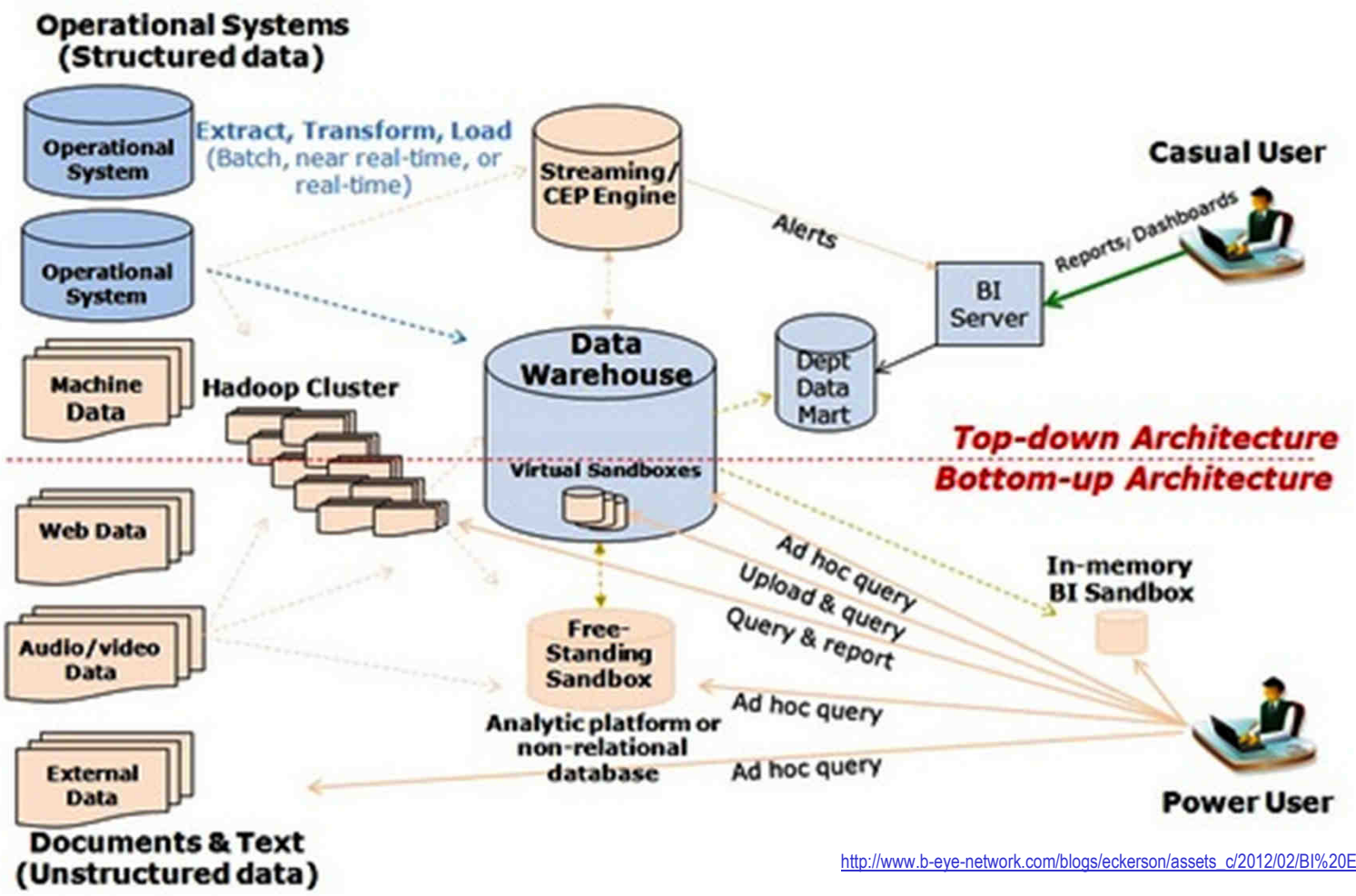
Because of exponential improvements in hardware …



http://radar.oreilly.com/2011/08/building-data-startups.html

How can we handle Big Data?

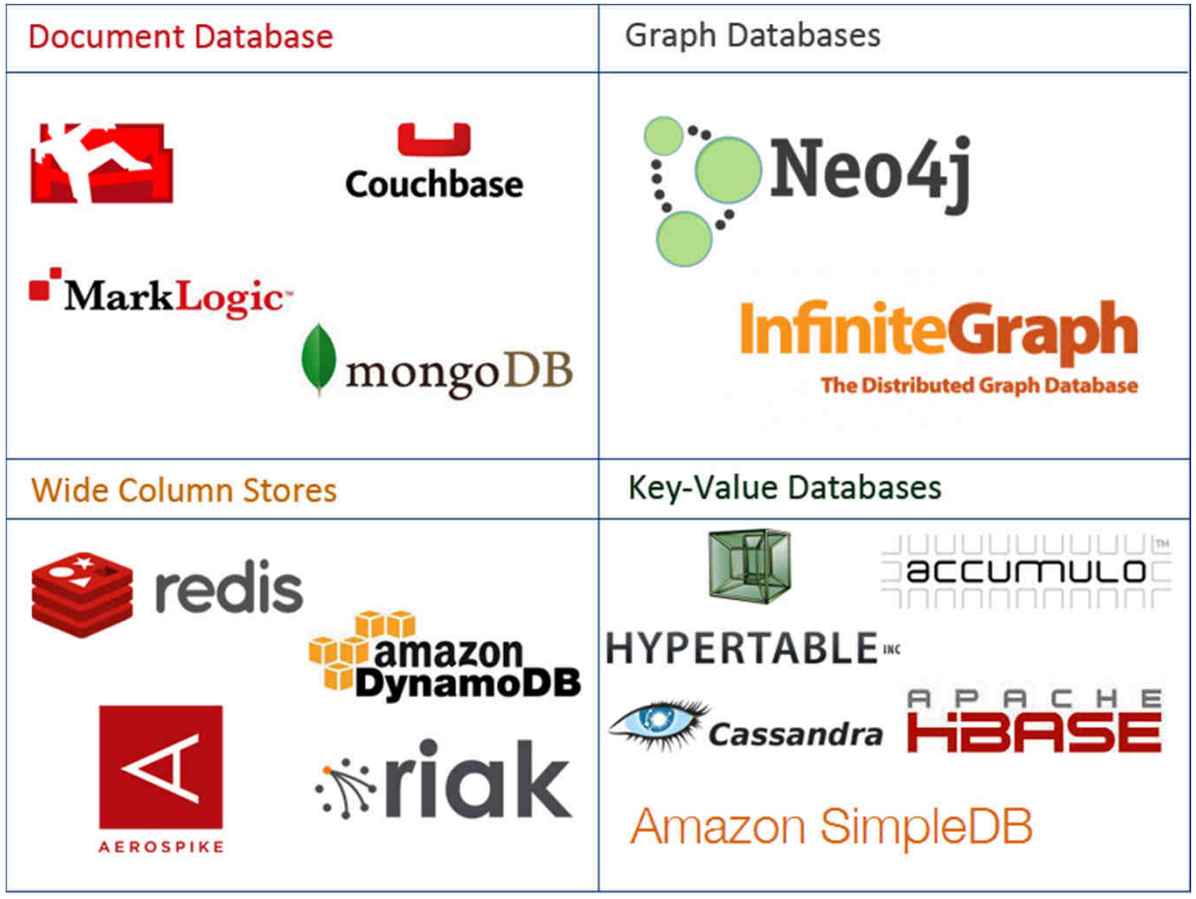
… new IT architectures



http://www.b-eye-network.com/blogs/eckerson/assets_c/2012/02/BI%20Ecosystem-474.php

How can we handle Big Data?

… new Data Bases : the NoSQL family



http://aryannava.com/2014/04/06/nosql-databases-family/

How can we handle Big Data?

… new data architectures

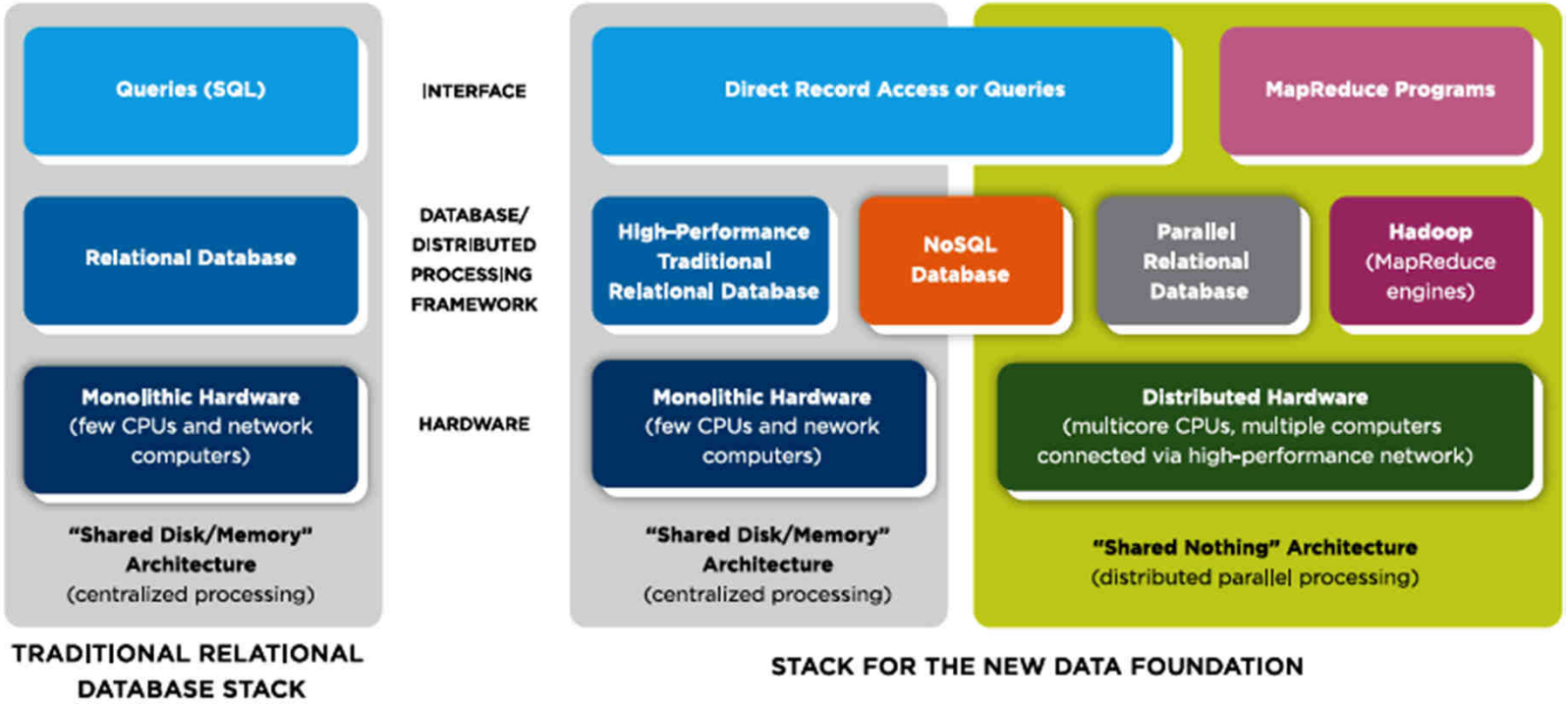


Figure 10 The traditional relational database stack, a staple of enterprise computing for decades, has evolved to a more varied data foundation.

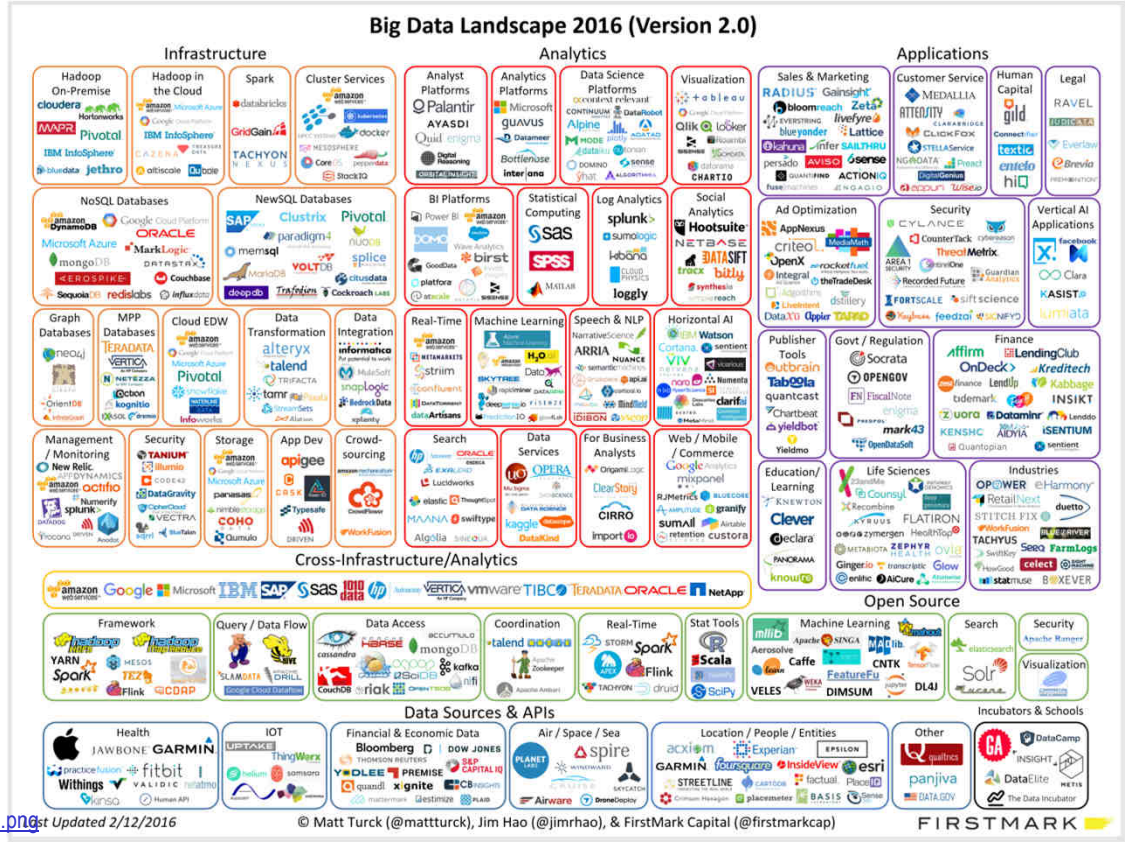
Source: CSC

http://assets1.csc.com/innovation/downloads/LEF_2011Data_rEvolution.pdf

How can we handle Big Data?

… many Big Data software tools

- Distributed computing (Hadoop, Spark)
- Data mining (scalable & distributed on hadoop / Spark)
- Social Network Analysis (large graphs)



http://mattturck.com/wp-content/uploads/2016/02/matt_turck_big_data_landscape_v11.png

Big Data … is not always so big

- Volume
  - For example in a Churn detection problem, the number of observations is the number of Customers
- Velocity
  - Very few applications require small response time
    - Except certainly in network monitoring
- Variety
  - But it is possible – and highly recommended – to increase it
    - We'll see how

    → that will increase volume

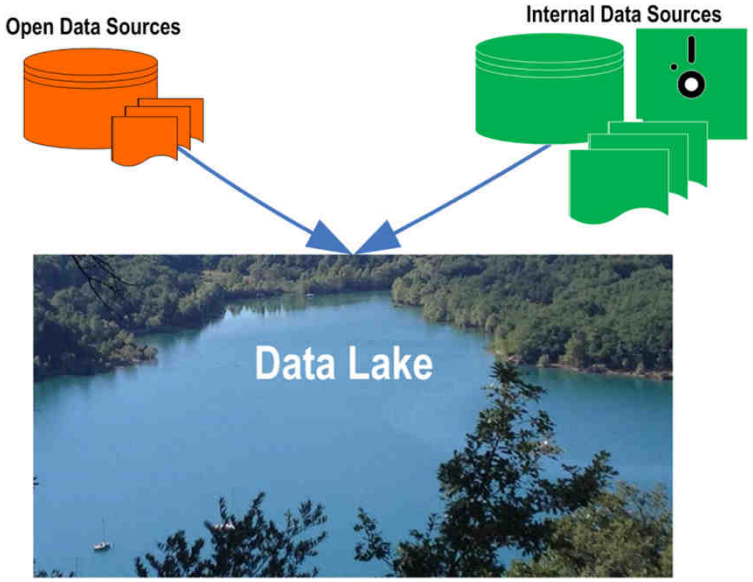
- Many projects actually use Big Data techniques on relatively "small" data

The Big Data process

Big Data process

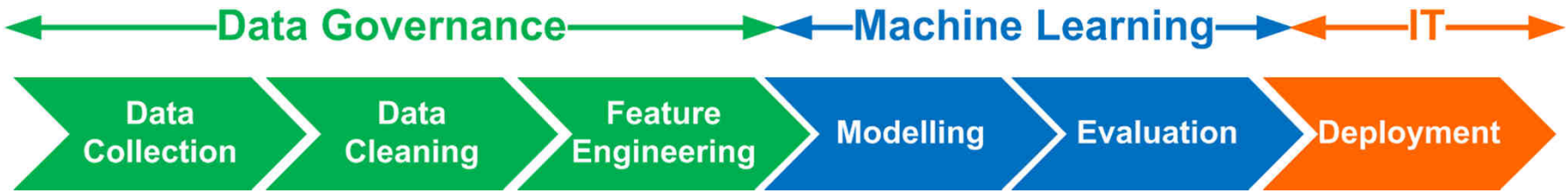
Implementing Big Data means

- A **big data collection problem**
  - Collect internal data sources & add external sources …
  - It is an always-active process (the more data the merrier)
- A **big migration problem**
  - Many internal databases / datawarehouses
  - A data lake might be considered
    - Store as-is
    - Avoid “integral” reconciliation
  - Move to distributed elastic storage



Big Data process

- Projects will get data from the data lake as needed
- A Big Data project has 3 main phases
  - **Data Governance** deals with collecting, cleaning, storing data & producing new features
  - **Machine Learning** deals with the analysis of the data to produce predictive models
    - Machine Learning is the tool to extract Value
    - Algorithms are not new (1980-1990)
  - **IT** involves the deployment of the application in the IT architecture



What is important ?

Not so important issue: algorithms

- Many well known algorithms
  - Linear/ logistic regression
  - Decision trees
  - Random forests
  - K-nn
  - Naïve Bayes
  - Neural networks
  - Support vector machine (SVM)
  - Deep learning …

Important issues

- **Data: Variety**
- Algorithms
  - Scalability
  - Explicability / predictability
  - Resistance to noise/ missing data (sparsity) / correlated data
  - Computing time (Build / Apply)
  - Performance
- Security / data protection
- Automation / productivity

"Invariably, simple models and a lot of data trump more elaborate models based on less data" https://static.googleusercontent.com/media/research.google.com/en//pubs/archive/35179.pdf

The model with best performance is not necessarily the best for deployment (ex: Netflix prize)

Increasing Variety

- **Variety** is the critical factor in Big Data
  - More varied sources
  - More variables

"At the end of the day, some machine learning projects succeed and some fail. What makes the difference? Easily the most important factor is the features used".

- **Feature engineering**
  - It aims at generating new features computed from the raw data
  - So far Feature engineering largely is a manual process, with domain-dependent expertise required, using up to 70% of a project effort

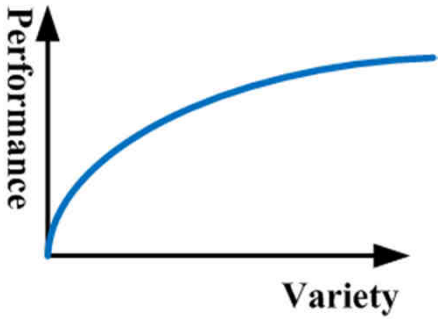
"One of the holy grails of machine learning is to automate more and more of the feature engineering process"

https://homes.cs.washington.edu/~pedrod/papers/cacm12.pdf

Increasing Variety

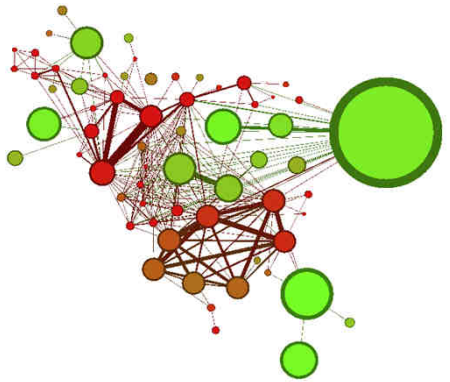
An example in Credit Card fraud detection on Internet

- Progressively increase Variety to increase performance
  - Produce new features from initial variables
    - Aggregates
    - Social variables
    - Scores
    - Segment …
  - 37 → 997



  - **Precision is multiplied by 5**

Model	Precision
Baseline	8,18%
Baseline + Agg.	19,00%
Baseline + Agg. + Social	40,58%



Challenges

- Organize data collection & storage
  - Architecture for storage : almost always a distributed cluster / cloud
- For building a Machine Learning model
  - A server with large RAM (2-4 TB) is ideal (in-memory analytics)
    - Computing features is costly: having all data/computations in RAM is fast
  - Distributed cluster + Spark may be necessary
  - Use open-source libraries
    - Scikit-learn (python, in-memory), MLLib (spark)
- For deploying a Machine Learning model
  - Balance the cost of using/maintaining many features in the model with the gain in performance
- The **hardest challenge**
  - Change mindset : learn how to phrase the question for Machine Learning

Impacts for networks ?

The power of Variety

Session

- **Secure Multi-party Based Cloud Computing Framework for Statistical Data Analysis of Encrypted Data**
  - *Harsha S Gardiyawasam Pussewalage (University of Agder, Norway)*
- **Data I/O Provision for Spark Applications in a Mesos Cluster**
  - *Nam Hoai Do, Tien Van Do and Xuan Tran (Budapest University of Technology and Economics, Hungary); Lorant Farkas and Csaba Rotter (Nokia Networks, Hungary)*
- **Subjective Perception Scoring**
  - *Jörg Niemöller (Ericsson, Sweden); Nina Washington (Ericsson Research, Sweden)*
- **Fighting Fire with Fire: Survey of Strategies for Counteracting The Complexity of Future Networks Management**
  - *Anne-Marie C. Bosneag, Sidath Handurukande, MingXue Wang (Ericsson Research Centre, Ireland)*

Thank you!

#ICIN2016