Big Data Management

Master in Data Science & Erasmus Mundus BDMA

Introduction to Big Data

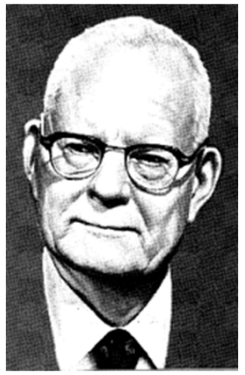
Knowledge Objectives

1. Recognise the relevance of data driven decision making
2. Identify the three high level categories of analytical tools OLAP, Querying and reporting, DM and ML
3. Identify the two main sources of Big Data
4. Give a definition of Big Data
5. Compare traditional data warehousing against Big Data management
6. Distinguish descriptive, predictive and prescriptive analytics
7. Explain the novelty of Cloud Computing
8. Justify the benefits of Cloud Computing
9. Explain the link between Big Data and Cloud Computing
10. Distinguish the main four service levels in Cloud Computing

Data driven decision making

The relevance of data

- "Without data you are just another person with an opinion."
  - William Edwards Deming (American engineer, statistician, professor and consultant



- "It is a capital mistake to theorize before one has data."
  - Sherlock Holmes (A Study in Scarlet)

We live in a data-driven society

Collect, store, combine and analyze any relevant data to gain competitive advantage

- Decision making:

  "To identify and choose alternatives based on values, preferences and beliefs of the decision-maker … every decision-making process produces a final choice." Wikipedia

- 90% of the world's data has been generated in the last two years
  - Data-driven decision making

Marr

Data as the New Cornerstone

- We have witnessed the bloom of a new business model based on data analytics: *Data is not a passive but an active asset*
  - «*Data is the new oil*!» - Clive Humby, 2006
  - «*No! Data is the new soil*» - David McCandless, 2010
- Confluence of three major socio-economic and technological trends makes data-driven innovation a new phenomenon today:
  - The **exponential growth** in data generated and collected,
  - the **widespread** use of **data analytics** including start-ups and small and medium enterprises (SMEs), and
  - the emergence of a **paradigm shift in knowledge**
- Organizations must adapt infrastructures to leverage the data deluge (digital data doubling every 18 months)

International Data Corp.'s (IDC)

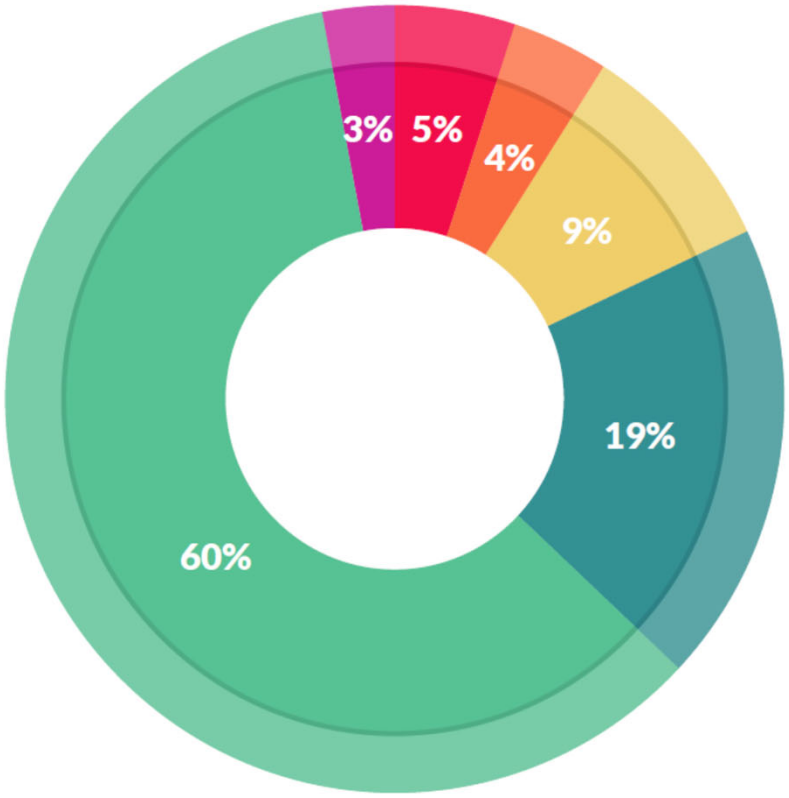
The Data Science Cake

Ingredients:
50g statistics
120g linear algebra
200g programming
1kg visualisation
300g software engineering

Additional skills:
creativity
out of the box thinking
grit
team spirit

Sonia Bergamaschi

Data Science Labor



What data scientists spend the most time doing

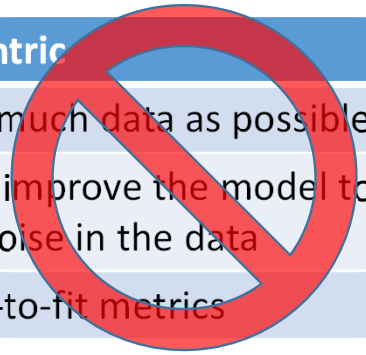
- *Building training sets: 3%*
- *Cleaning and organizing data: 60%*
- *Collecting data sets; 19%*
- *Mining data for patterns: 9%*
- *Refining algorithms: 4%*
- *Other: 5%*

https://visit.figure-eight.com/rs/416-ZBE-142/images/CrowdFlower_DataScienceReport_2016.pdf

Model-centric vs Data-centric

- People do not focus on improving the quality of data …
- … Models do not work well on real data

Model-centric	Data-centric
Collect as much data as possible	Hold the model fixed
Iteratively improve the model to deal with the noise in the data	Iteratively improve the quality of the data to obtain better results
Goodness-to-fit metrics	Goodness-of-data metrics

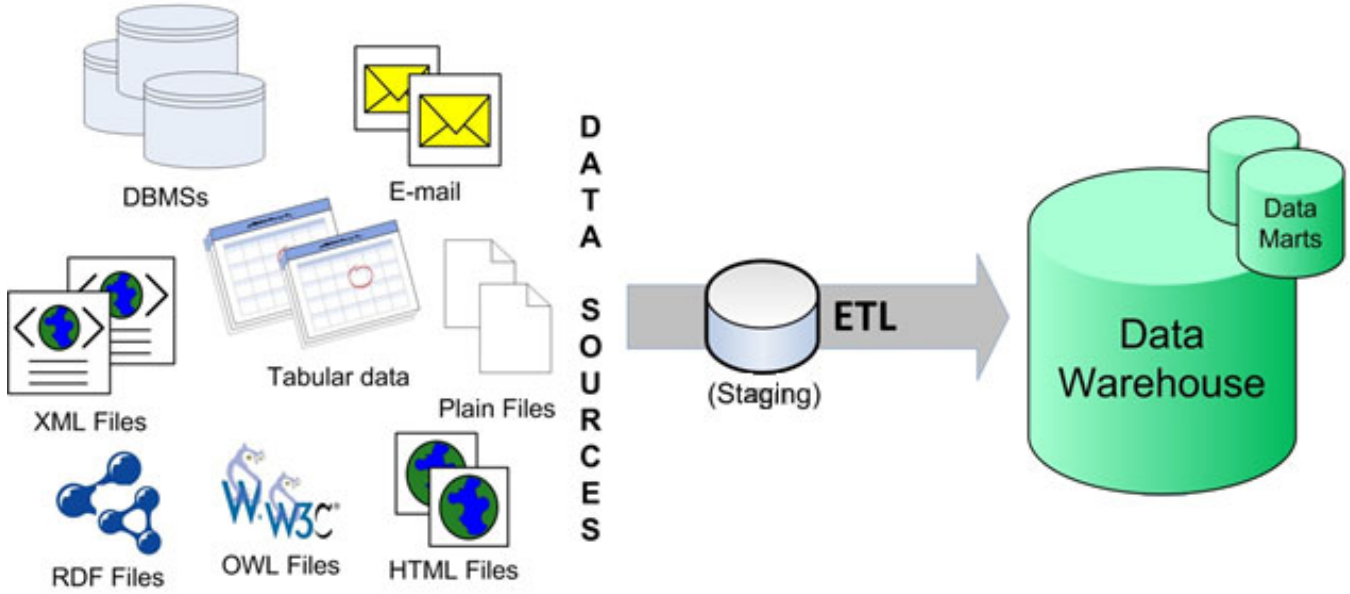


- Systematic improvement of data quality on a basic model is better than using the state-of-the-art models with low-quality data

N. Sambasivan, S. Kapania, H. Highfill, D. Akrong, P. Paritosh, and L. M Aroyo.
Everyone wants to do the model work, not the data work. Data Cascades in High-Stakes AI.
Conference on Human Factors in Computing Systems (CHI). ACM, 2021

Business Intelligence: Data Management

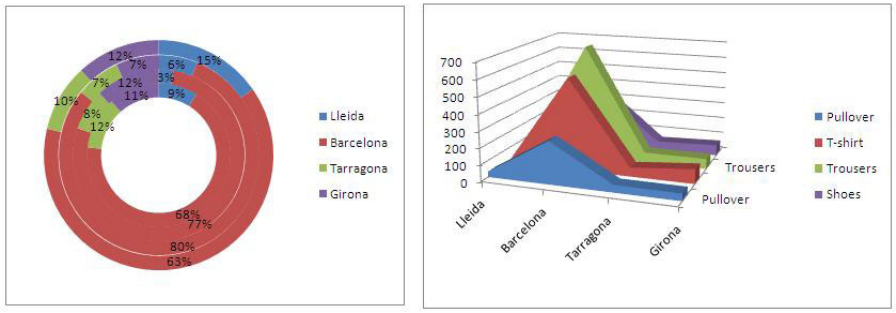
- Well-established *de facto* standards:
  - Architecture: Corpotare Information Factory
  - Data Modeling: Multidimensional model

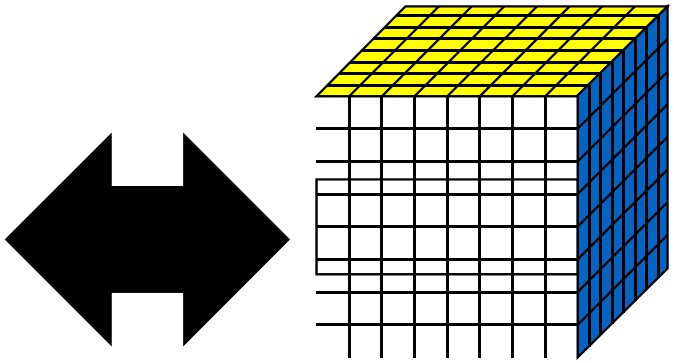


Business Intelligence: Analytics

- Three different levels of detail
  - Querying & Reporting: Static report generation
  - OLAP: Dynamic navigation of data
  - Data Mining and Machine Learning: Inference of hidden patterns or trends

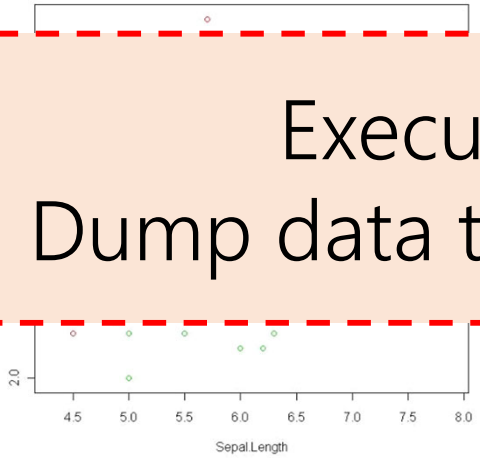
	Lleida	Barcelona	Tarragona	Girona
Pullover	34	260	45	44
T-shirt	20	564	56	89
Trousers	55	700	63	62
Shoes	87	360	54	67





Business Intelligence: Analytics

- Three different levels of detail
  - Querying & Reporting: Static report generation
  - OLAP: Dynamic summarizations of data
  - Data Mining and Machine Learning: Inference of hidden patterns or trends

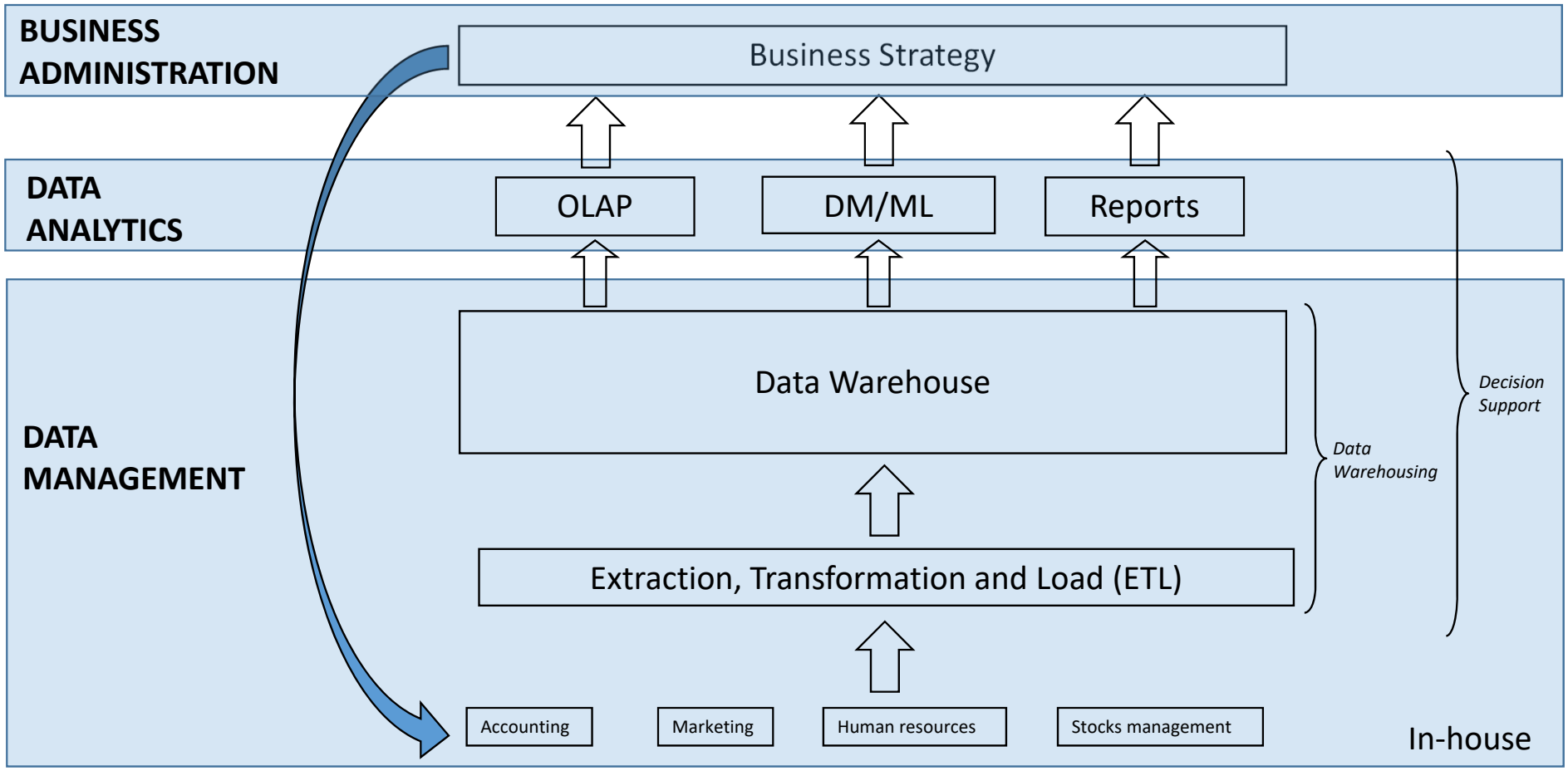



```
> (kc <- kmeans(newiris, 3))
K-means clustering with 3 clusters of sizes 38, 50, 62

Cluster means:
  Sepal.Length Sepal.Width Petal.Length Petal.Width
```

Executed outside the DW
Dump data to specialized tools (R/SAS)

```
Within cluster sum of squares by cluster:
[1] 23.87947 15.15100 39.82097

Available components:
[1] "cluster"  "centers"  "withinss" "size"
```

The Business Intelligence (BI) Cycle



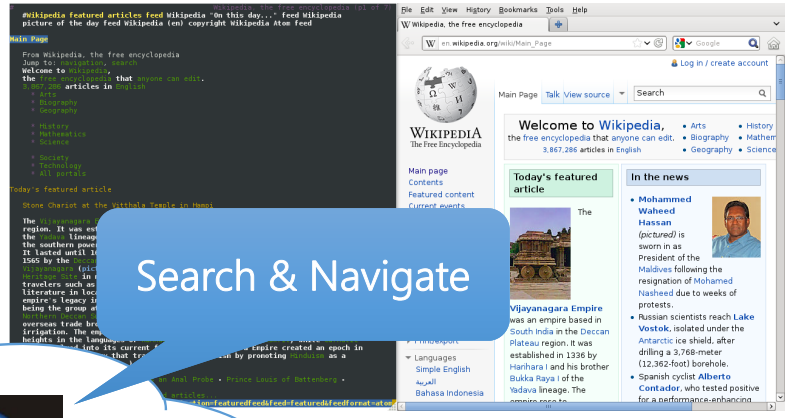
Big Data

Opportunity

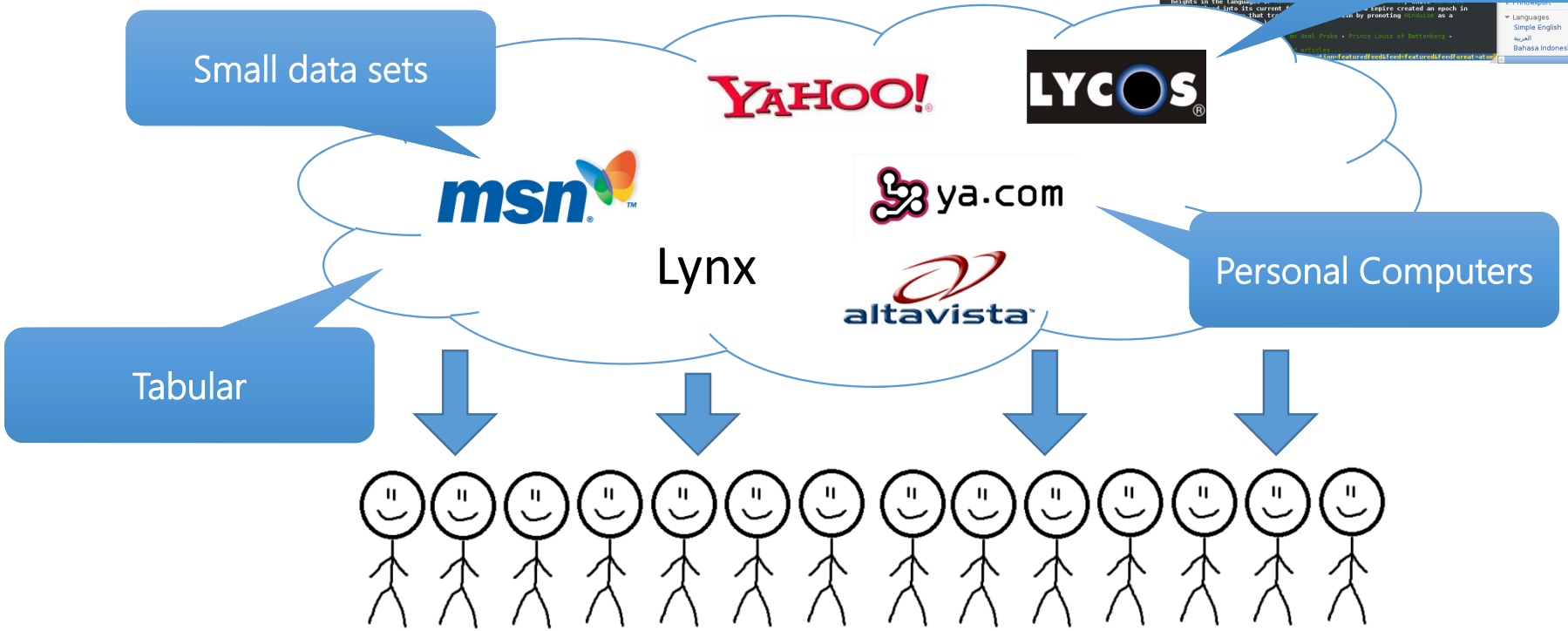
Means

The end of an architectural era

WEB 1.0 – Read Era

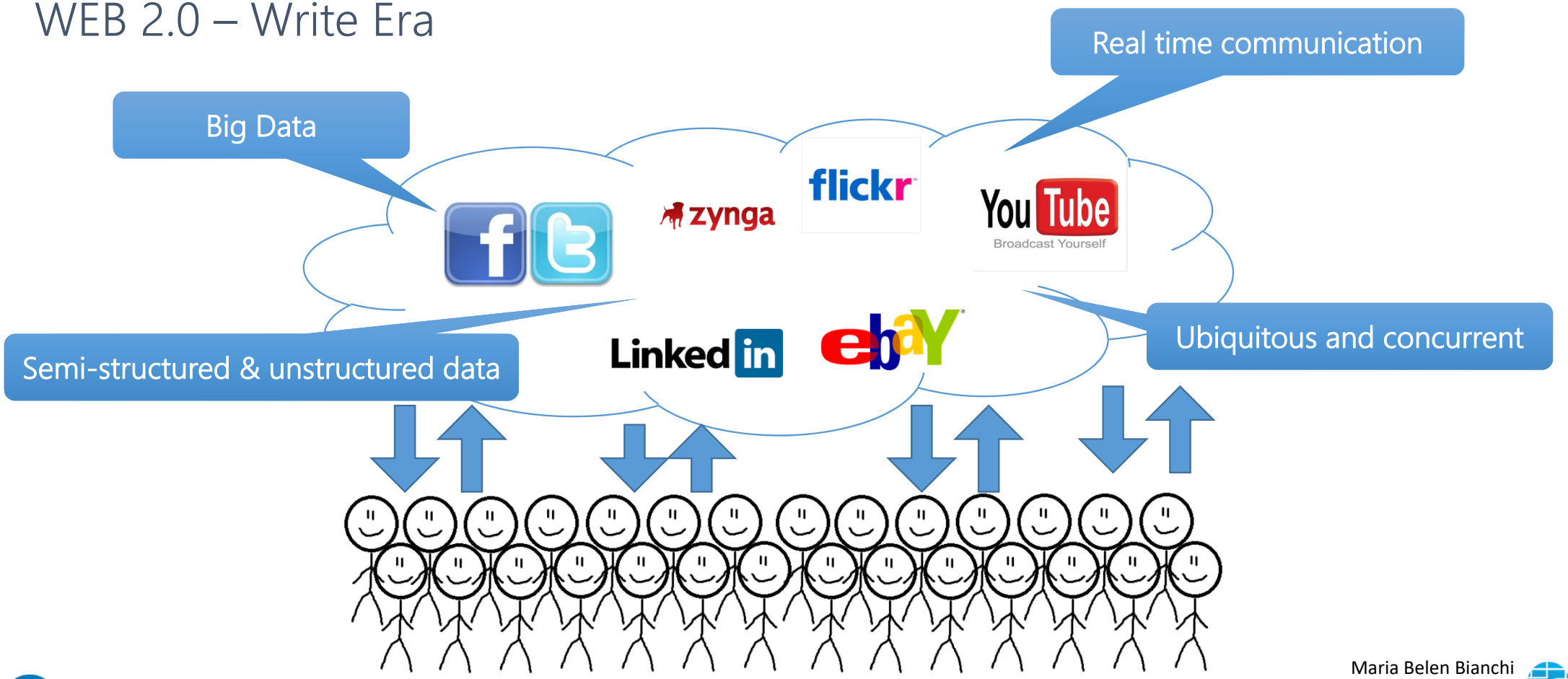


Wikipedia



The end of an architectural era

WEB 2.0 – Write Era



Futbol Club Barcelona fans information

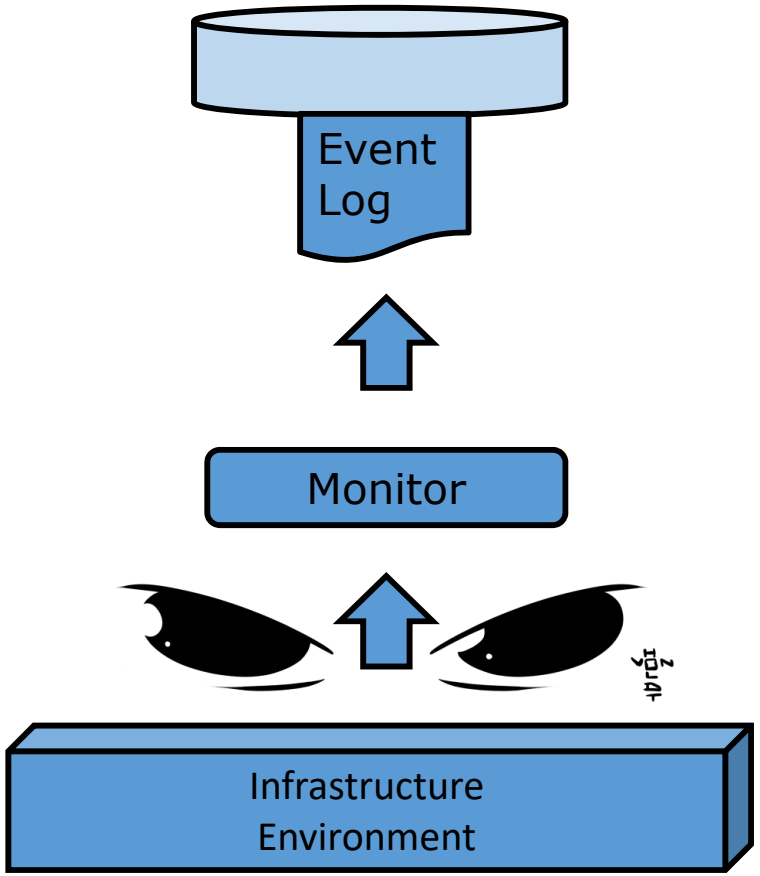
- Socio-demographic data
  - Contents: 140K rows and 6 columns (one per fan)
  - Source: Web page of the club
- Operational data
  - Source: Both online and physical shopping of tickets and merchandising
- Likes
  - Contents: ~36M rows and 4 columns (18K different likes)
  - Source: Facebook (web page and wifi autentications)
- Interests
  - Contents: ~20M rows and 6 columns (~371K fans)
  - Source: Third-party (Xeerpa)
- Personality insights
  - Contents: ~7M rows and 4 columns (~140K fans)
  - Source: Third-party (Xeerpa)

ID	Fan ID	Like	Date
1	500	Messi	11-21-2013
2	500	FC Barcelona	11-21-2013
3	500	Shakira	01-03-2013
4	500	Spotify	07-28-2010
5	501	BlackBerry	07-12-2013
6	501	TV3	03-31-2017
7	501	Ucrania	02-01-2012
8	502	Rolling Stones	12-10-2011

ID	Fan ID	Interest1	Interest2	Interest3	Score
1	500	Food-Beverage	Alcoholic	Whisky	7
2	500	Food-Beverage	Alcoholic	Rum	8
3	500	Sports	Sports	Futbol	5
4	500	Sports	Sports	Baloncesto	4
5	501	Food-Beverage	Alcoholic	Rum	6
6	501	Sports	Sports	Handball	7
7	501	Sports	Sports	Futbol	8
8	502	Food-Beverage	Alcoholic	Cerveza	2

ID	User ID	Trait	Percent
1	5ce47dbc36f95c42afeb	Openness	0.56
2	5ce47dbc36f95c42afeb	Adventurousness	0.65
3	5ce47dbc36f95c42afeb	Artistic interests	0.41
4	5ce47dbc36f95c42afeb	Emotionality	0.80
5	5ce47dbc36f95c42afeb	Imagination	0.25
6	5ce47dbc36f95c42afeb	Intellect	0.72
7	5ce47dbc36f95c42afeb	Authority-challenging	0.34
8	5ce47dbc36f95c42afeb	Conscientiousness	0.43

Monitoring the infrastructure



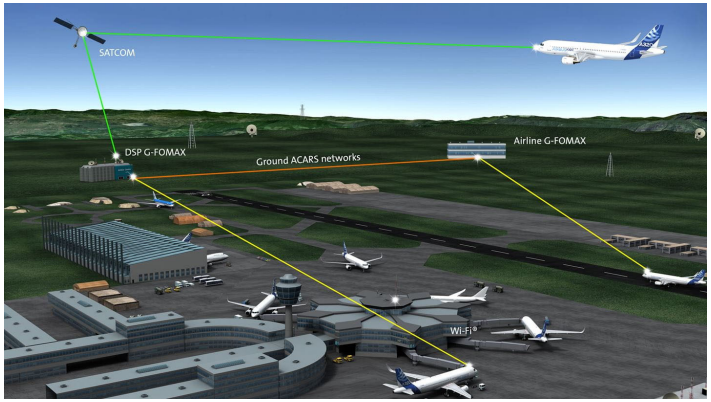
Danish wind turbines

- One park:
  - 100+ turbines
- One turbine:
  - 500 sensors
  - More than 2500 derived data streams
- One sensor:
  - 8 bytes sampled at 100+Hz

100 turbines*2500 streams*100 samples/sec = $25 \cdot 10^6$ samples/second
8bytes*$25 \cdot 10^6$ samples/second*3600seconds/hour*24hours/day = 17.5TB/day
17.5TB/day*365 days/year = 6+ PB/year/park

Having thousands of parks and storing 20+ years of history …

Aerospace corporation



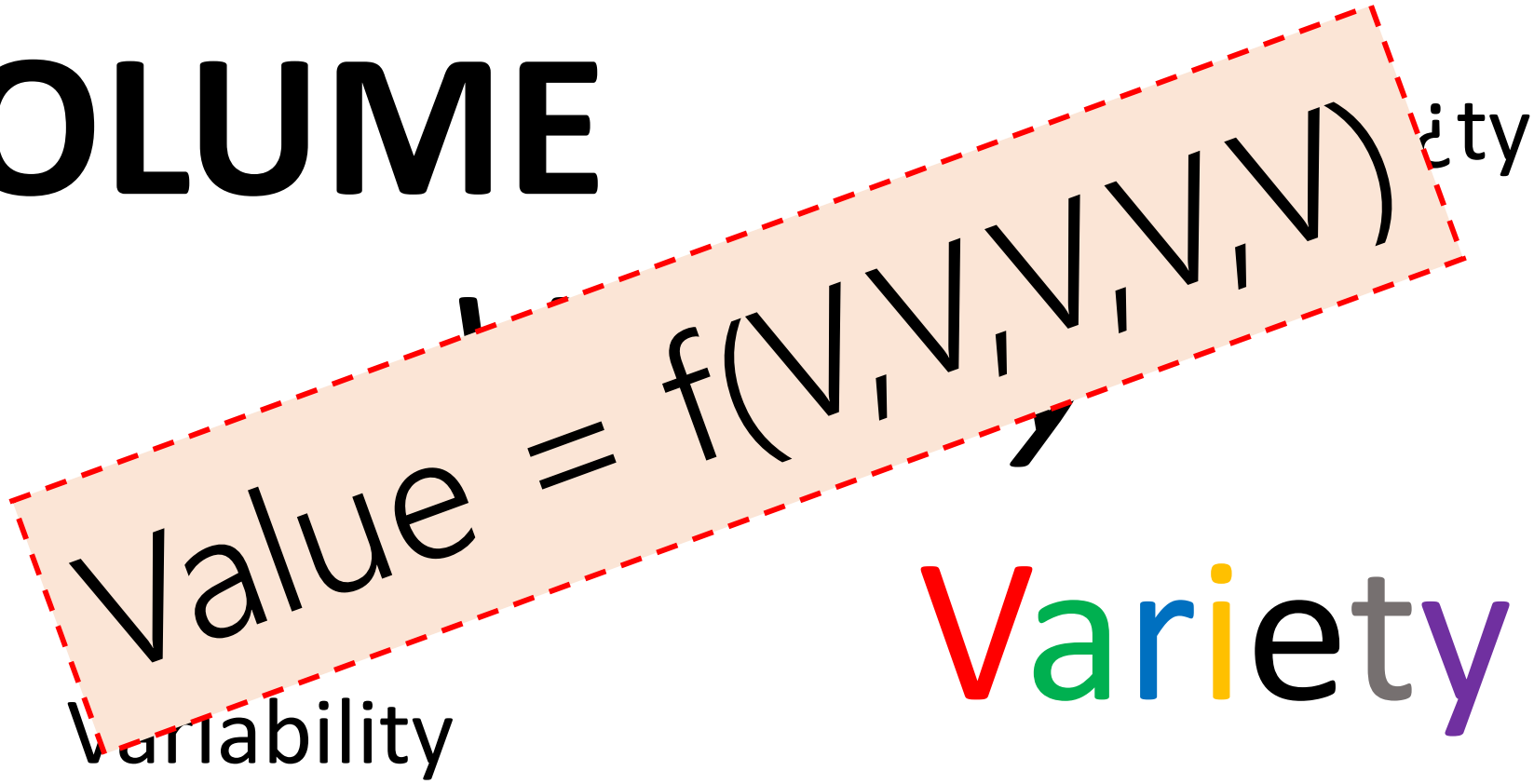
Collins Corp.

- One (not big) airline
  - 125 planes
- One plane:
  - 24.000 sensors (Flight Operations & Maintenance Exchanger, FOMAX)
  - 10 hours/day
- One sensor
  - 8 bytes sampled at 20+Hz

125 planes*24.000 sensors*20 samples/sec = $60 \cdot 10^6$ samples/second
8 bytes*$60 \cdot 10^6$ samples/second*3600seconds/hour = 1.73TB/hour
1.73TB/hour*10 hours/day*365 days/year = 6+ PB/year

Having tens of airlines and storing 10+ years of history …

New challenges for data management



Big Data facets

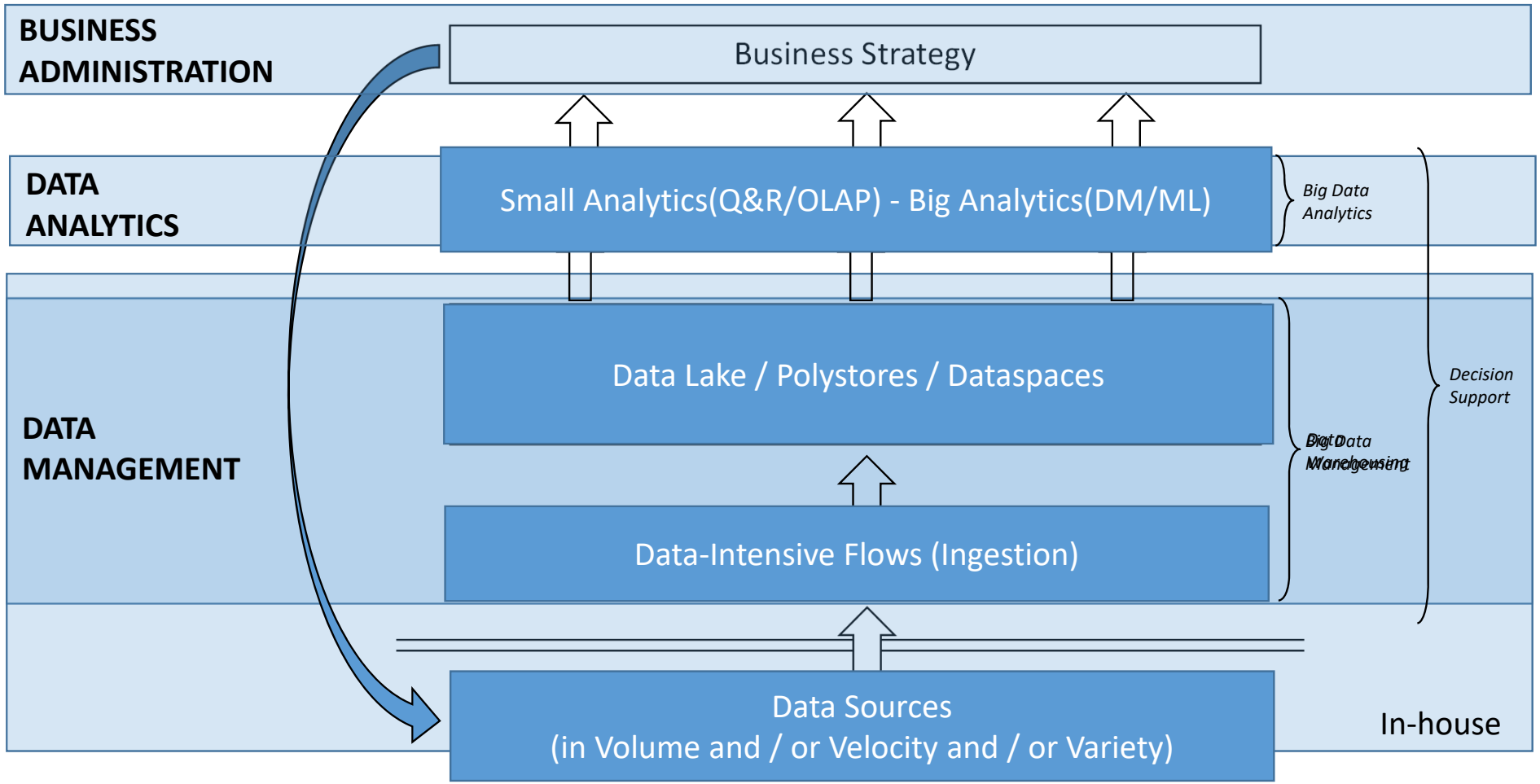
a) The Original
b) as Technology
c) as Data Distinctions
d) as Signals
e) as Opportunity
f) as Metaphor
g) as New Term for Old Stuff

Timo Elliott

1963 (CRAM): Each CRAM deck of 256 cards recorded about 5.5 MB.

https://www.computerhistory.org/timeline/memory-storage

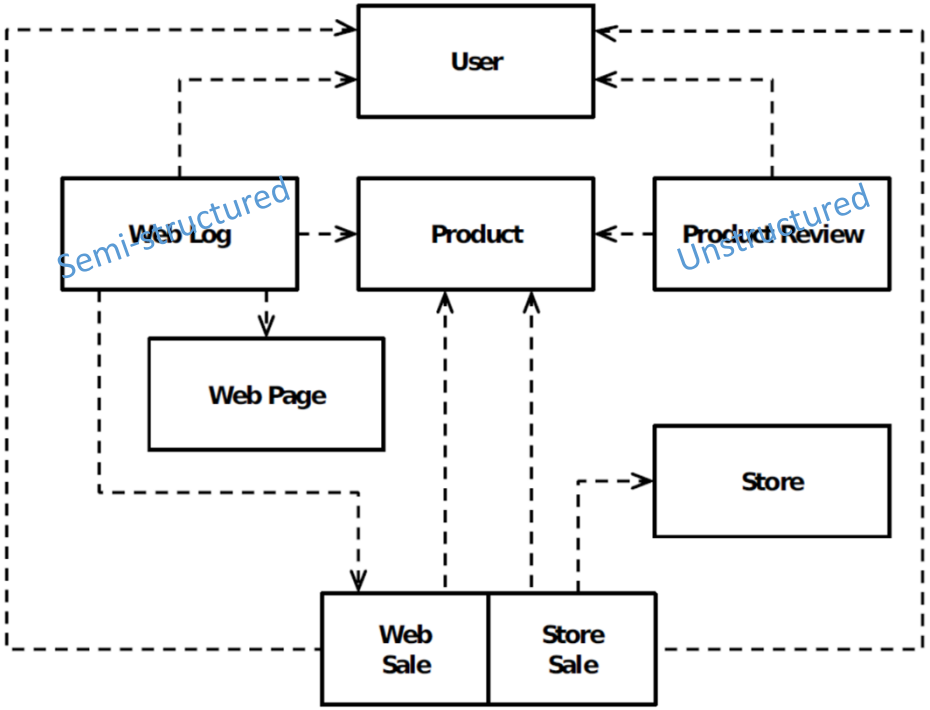
The Big Data Cycle



Big Data related areas

- Volume and Velocity
  - Distributed processing
  - Parallelism
  - Declarative querying
  - Query optimization
- Variety and Variability
  - Information retrieval
  - Web and text mining
  - Schema evolution
  - Data integration
- Veracity/Validity
  - Data quality
  - Uncertainty
  - Statistical reasoning
  - Data lineage and provenance
- Value
  - Analytics (ML)
  - Biology, Linguistics, Sports

Big Bench v2



Data\Factor	1	100	1000	10000
webpage	26	26	26	26
product	1,000	1,900	4,063	10,900
user	10,900	109,900	1,009,900	10,009,900
store	100	105	150	600
web sale	143,880	1,450,680	13,330,680	132,130,680
store sale	59,950	604,450	5,554,450	55,054,450
product review	163,863	1,652,163	15,182,163	150,482,163
weblog	23,000,000	236,000,000	2,200,000,000	21,500,000,000

Business Category	**BigBench V2**	
	No. of queries	Percentage
Marketing	20	69.0%
Merchandising	3	10.3%
Operations	2	6.9%
Supply chain	1	3.3%
New business models	4	13.8%
Query Type	**BigBench V2**	
	No. of queries	Percentage
Declarative	7	24.1%
Procedural	4	13.3%
Declarative & Procedural	19	65.6%
Data Source	**BigBench V2**	
	No. of queries	Percentage
Structured	5	16.7%
Semi-Structured	20	66.7%
Unstructured	5	16.7%

A. Ghazal et al. *BigBench v2: The New and Improved BigBench*. ICDE'17

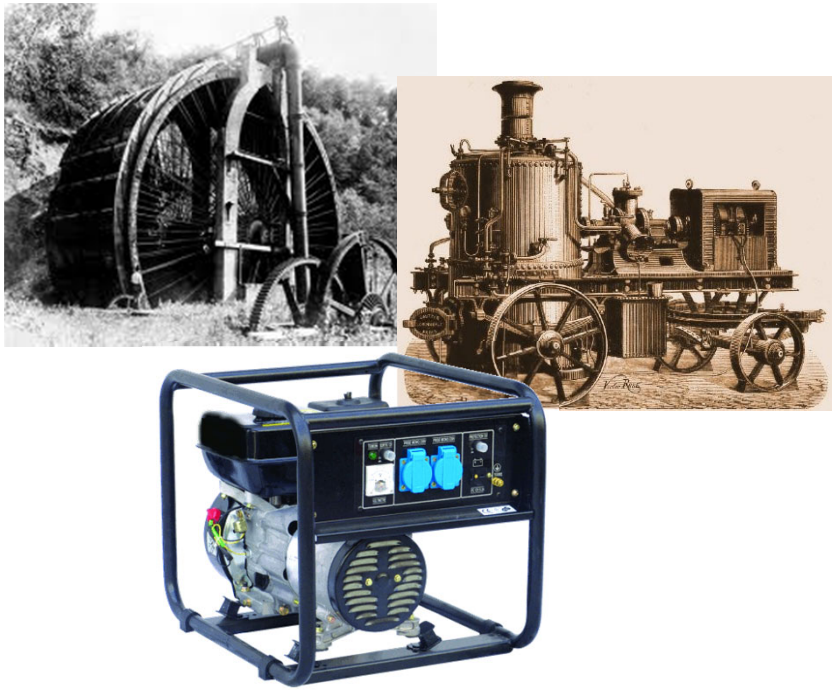
An orthogonal classification: kinds of data analytics

- Descriptive: Deterministically compute summarizations
  - Count, sum, average, min, max, etc.
  - Typical OLAP operations
- Predictive: Probabilistic by nature, try to forecast what may happen according to what have happened
  - Linear and non-linear regression,
  - Classification,
  - Clustering,
  - Association rules, etc.
- Prescriptive: Given the prediction(s) of a (several) model(s), understand why something is happening and undertake automatic action(s)
  - Examples:
    - Stock market (buy/sell shares)
    - Set Price (automatically increase/decrease)

Cloud Computing

Providing access to infrastructure

Analogy: Electricity as a Utility



Own production

Pay-per-use

Computation as a Utility

Private Data Centre

(*"Own production"*)

Public/Private Cloud

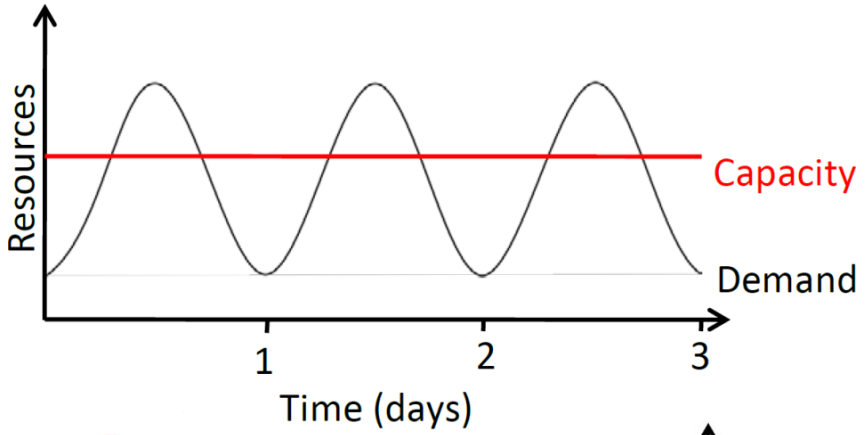
(Pay-per-use)

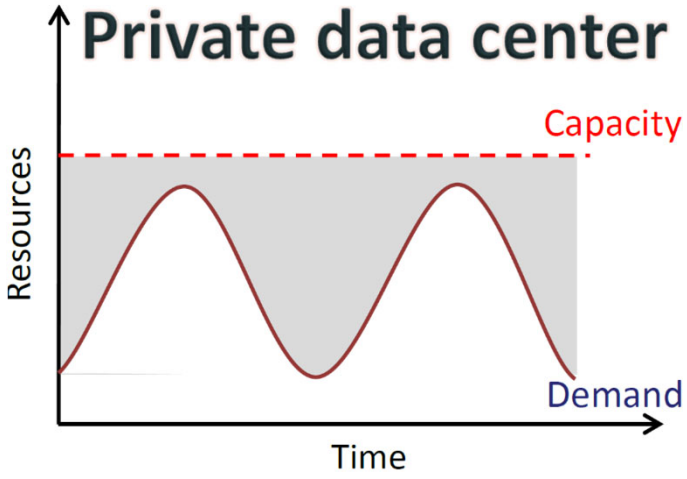
Cloud Computing definition

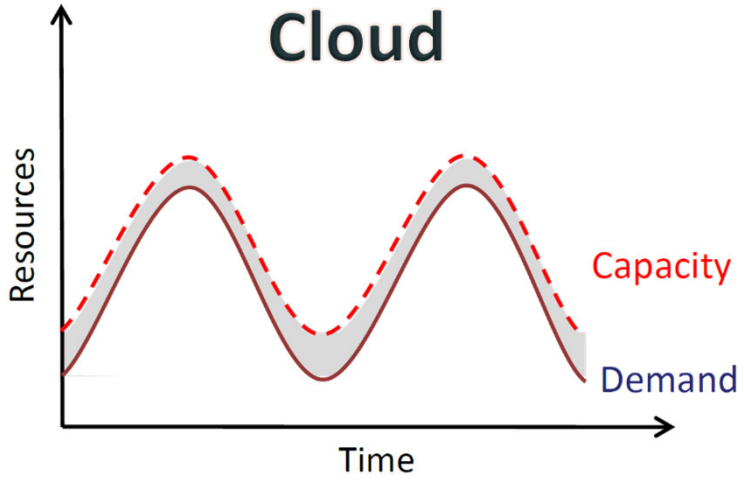
"Cloud computing is a model for enabling convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction."

NIST (National Institute of Standards and Technology)

Undercapacity Risk







D. Abadi

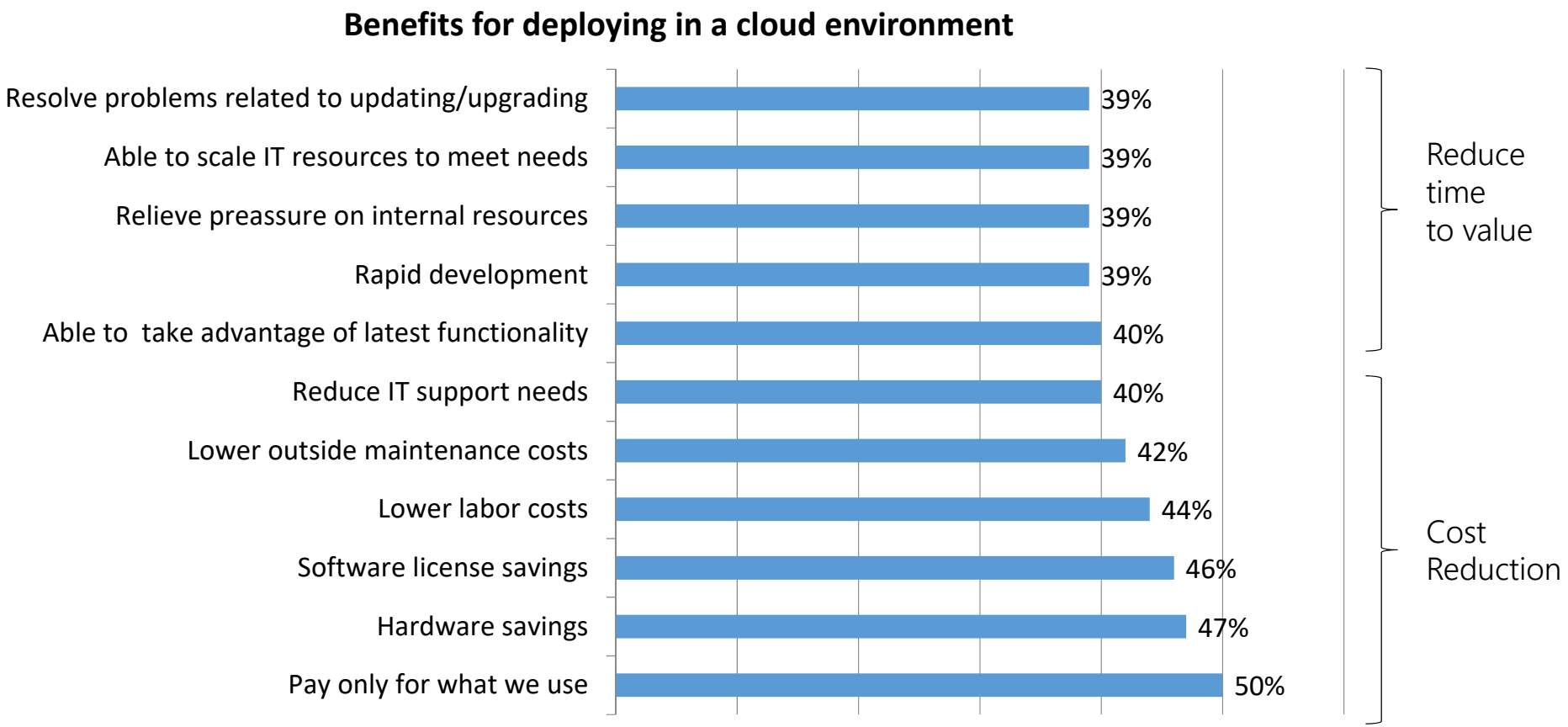
Novelty of cloud computing

- Elimination of up-front commitment
- Illusion of infinite resources
- Pay-per-use (elasticity)
  - Cost is 5-7 times cheaper than in-house computing
- Service Level Agreements
  - E.g., Availability=uptime/(uptime+downtime)
    - Measured in terms of nines (99.99…9%)

Benefits of Cloud computing

- Reduce costs
  - Economy of scale in software development
  - Energetic efficiency
- Agility
- Flexibility
- Easier manageme
- Superior safety
- Better upgradeability
- More business

Benefits of cloud computing

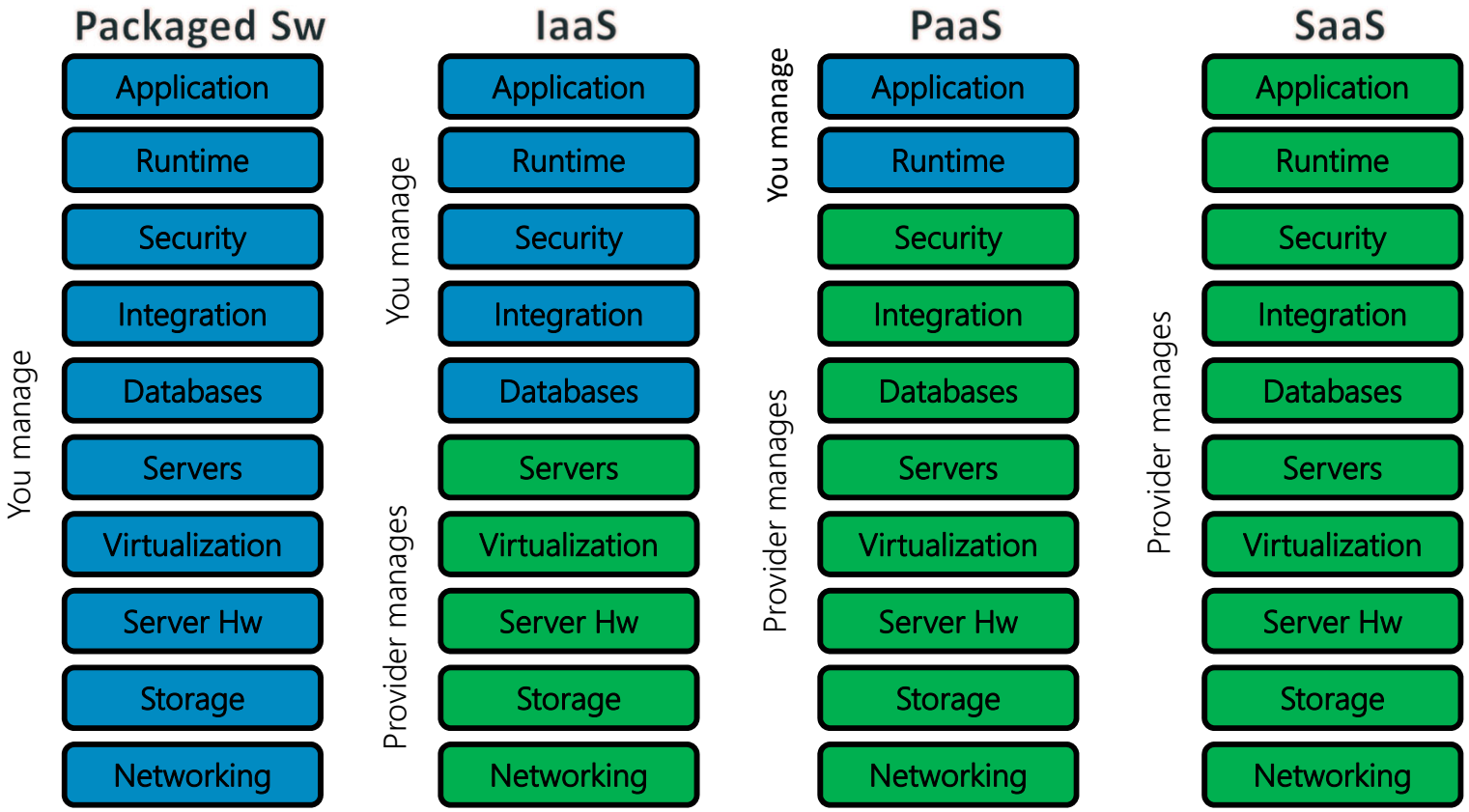


IBM global survey of IT and line-of-business decision makers 2012

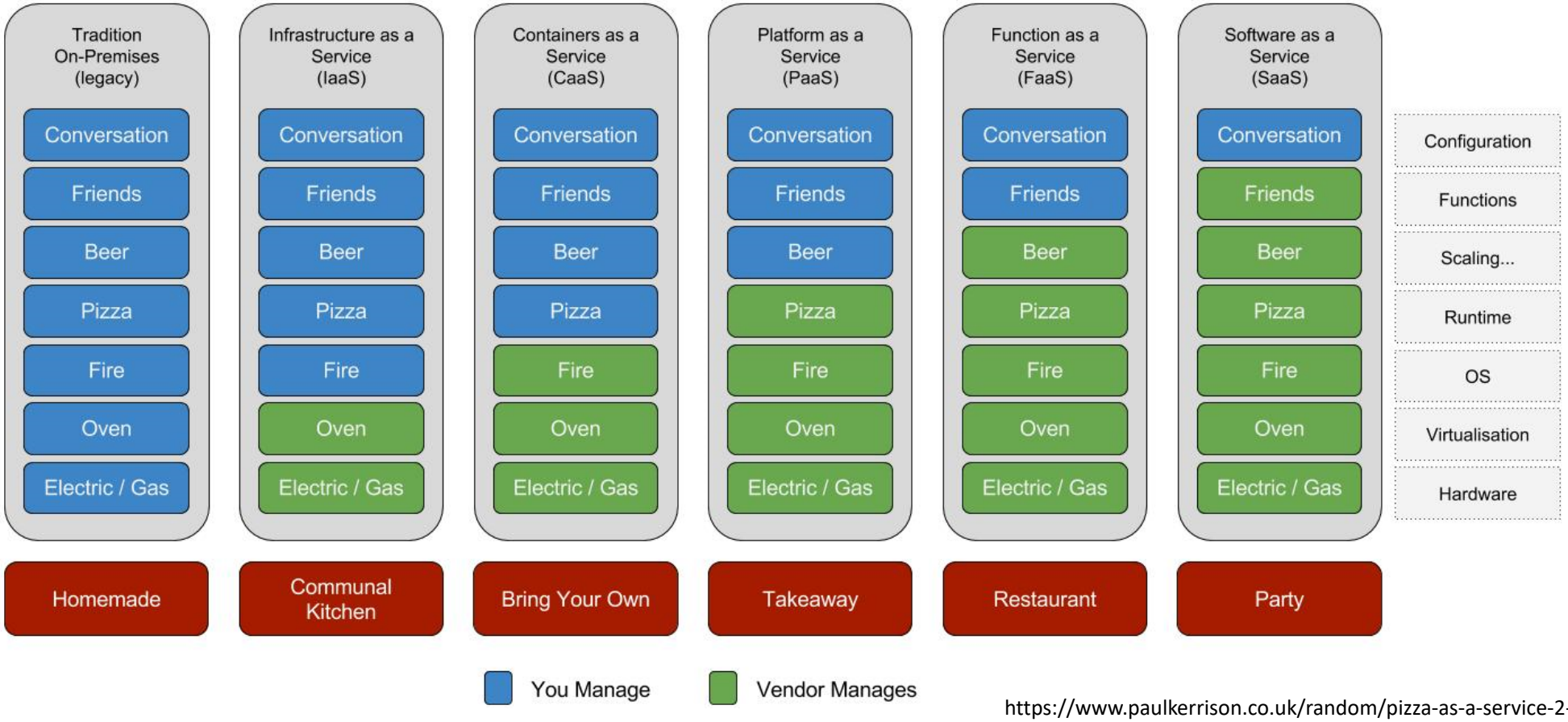
Levels of Service

- The company outsources some responsibility to the service provider
  - Business as a Service (BaaS)
    - A whole business process is outsourced (e.g., Paypal, Amadeus)
  - Software as a Service (SaaS)
    - Software is there, ready to be used (e.g., Google Docs, Dropbox)
  - Platform as a Service (PaaS)
    - You get software modules needed to run applications (e.g., databases, web servers, security)
  - Infrastructure as a Service (IaaS)
    - You get a server to connect through remote connection protocols (e.g., VPN, SSH, FTP)
    - Typically it covers the hardware (e.g., computers, network, virtualization)
- Levels are incremental: SaaS entails PaaS, and PaaS entails IaaS

Share of responsibility



Pizza as a Service



https://www.paulkerrison.co.uk/random/pizza-as-a-service-2-0

Service providers

- Some of the strongest players in the market
  - Amazon Web Services (AWS)
  - Google Cloud
  - Microsoft Azure
  - IBM Cloud
  - Rackspace
  - Digital Ocean

Closing

Summary

- Big Data definition
  - From a data management perspective
  - From a data analysis perspective
- Cloud computing needs and contribution

References

- D. Abadi. *Data management in the cloud: Limitations and opportunities*. IEEE Data Engineering Bulletin 32(1), 2009
- C. Baun et al. *Cloud Computing*. Springer, 2011
- P. Mell and T. Grance. *The NIST Definition of Cloud Computing*. Special Publication 800-145, National Institute of Standards and Technology (September 2011)
- M. Madsen. *Cloud Computing Models for Data Warehousing*. Third Nature Technology White Paper, 2012
- A. Ghazal et al. *BigBench v2: The New and Improved BigBench*. ICDE'17
- N. Sambasivan, S. Kapania, H. Highfill, D. Akrong, P. Paritosh, and L. M Aroyo. *Everyone wants to do the model work, not the data work. Data Cascades in High-Stakes AI*. Conference on Human Factors in Computing Systems (CHI). ACM, 2021
- NIST Cloud Computing Program, http://www.nist.gov/itl/cloud
- Gartner Reports. G00232650, G00175593, and G00219131