ARCADA

YRKESHÖGSKOLAN ARCADA

Introduction to Analytics

Magnus.Westerlund @arcada.fi

Researcher & Programme Director
01.04.2017

ARCADA

Analytics related full-time Researchers at our Department

- **Ph.D Anton Akusok (data processing wizard)**
- **Ph.D Leonardo Espinosa (method master)**
- **D.Sc. Magnus Westerlund (autonomous agents (fintech))**
- D.Sc Shuhua Liu (NLP)
- Ph.D Jonny Karlsson (IOT)
- D.Sc Kaj-Mikael Björk (Optimization)

- **MSc Andrej Shcherbakov (lecturer)**
- **Stig Blomqvist (Administration and Payments)**

BDA Specialisation study programme schedule			[as of 27.8.2018]
06.09.2018	13:00 - 18:00	Introduction to Analytics	B518
07.09.2018	13:00 - 18:00	Introduction to Analytics	F249 (Lilla auditoriet)
20.09.2018	13:00 - 18:00	Introduction to Analytics	F249 (Lilla auditoriet)
21.09.2018	13:00 - 18:00	Introduction to Analytics	F249 (Lilla auditoriet)
04.10.2018	13:00 - 18:00	Introduction to Analytics	F249 (Lilla auditoriet)
05.10.2018	13:00 - 18:00	Introduction to Analytics	F249 (Lilla auditoriet)
18.10.2018	13:00 - 18:00	Machine Learning for Predictive Problems	F249 (Lilla auditoriet)
19.10.2018	13:00 - 18:00	Machine Learning for Predictive Problems	F249 (Lilla auditoriet)
01.11.2018	13:00 - 18:00	Machine Learning for Predictive Problems	B518
02.11.2018	13:00 - 18:00	Machine Learning for Predictive Problems	F143 (Stora auditoriet)
15.11.2018	13:00 - 18:00	Machine Learning for Predictive Problems	F249 (Lilla auditoriet)
16.11.2018	13:00 - 18:00	Machine Learning for Predictive Problems	F249 (Lilla auditoriet)
29.11.2018	13:00 - 18:00	Visual Analytics	B518
30.11.2018	13:00 - 18:00	Visual Analytics	F249 (Lilla auditoriet)
13.12.2018	13:00 - 18:00	Visual Analytics	F143 (Stora auditoriet)
14.12.2018	13:00 - 18:00	Visual Analytics	F249 (Lilla auditoriet)
17.01.2019	13:00 - 18:00	Visual Analytics	D173
18.01.2019	13:00 - 18:00	Visual Analytics	F143 (Stora auditoriet)
31.01.2019	13:00 - 18:00	Machine Learning for Descriptive Problems	F249 (Lilla auditoriet)
01.02.2019	13:00 - 18:00	Machine Learning for Descriptive Problems	F249 (Lilla auditoriet)
14.02.2019	13:00 - 18:00	Machine Learning for Descriptive Problems	F249 (Lilla auditoriet)
15.02.2019	13:00 - 18:00	Machine Learning for Descriptive Problems	F249 (Lilla auditoriet)
28.02.2019	13:00 - 18:00	Machine Learning for Descriptive Problems	F249 (Lilla auditoriet)
01.03.2019	13:00 - 18:00	Machine Learning for Descriptive Problems	F249 (Lilla auditoriet)
14.03.2019	13:00 - 18:00	Big Data Analytics	F249 (Lilla auditoriet)
15.03.2019	13:00 - 18:00	Big Data Analytics	F249 (Lilla auditoriet)
28.03.2019	13:00 - 18:00	Big Data Analytics	F249 (Lilla auditoriet)
29.03.2019	13:00 - 18:00	Big Data Analytics	F143 (Stora auditoriet)
11.04.2019	13:00 - 18:00	Big Data Analytics	F249 (Lilla auditoriet)
12.04.2019	13:00 - 18:00	Big Data Analytics	F249 (Lilla auditoriet)
25.04.2019	13:00 - 18:00	Analytical Service Development	F143 (Stora auditoriet)
26.04.2019	13:00 - 18:00	Analytical Service Development	F249 (Lilla auditoriet)
09.05.2019	13:00 - 18:00	Analytical Service Development	F249 (Lilla auditoriet)
10.05.2019	13:00 - 18:00	Analytical Service Development	F249 (Lilla auditoriet)
23.05.2019	13:00 - 18:00	Analytical Service Development	F249 (Lilla auditoriet)
24.05.2019	13:00 - 18:00	Analytical Service Development	F249 (Lilla auditoriet)

Intro to Analytics - Course Schedule

- Week 1
  - 6.9: Intro to Analytics, Machine Learning, and AI
  - 7.9: Feature engineering, Pandas
- Week 2
  - 20.9: Time series processing, linear modeling and setting targets/labels
  - 21.9: Time series data visualization and regression
- Week 3
  - 4.10: Understanding model output, and going from output to decision
  - 5.10: Open discussion, creating decisions, finalizing project

Week 1 - Objectives

- You can get hold of lecture slides from ItsLearning and Slack
- We have prepared a Jupyter guide to Analytics in Python
  - After downloading you can run it and modify the examples
- The first (individual) **assignment**, based on the guide, **is due 20.9**
  - Please note, that you need to have access to a Python environment to do the assignment.
- We prepared an extra notebook for those that want a challenge
  - We will set up a competition for who gets the best “repeatable” score.
  - Prize award ceremony will be held first week of 2nd course.

Todays Agenda

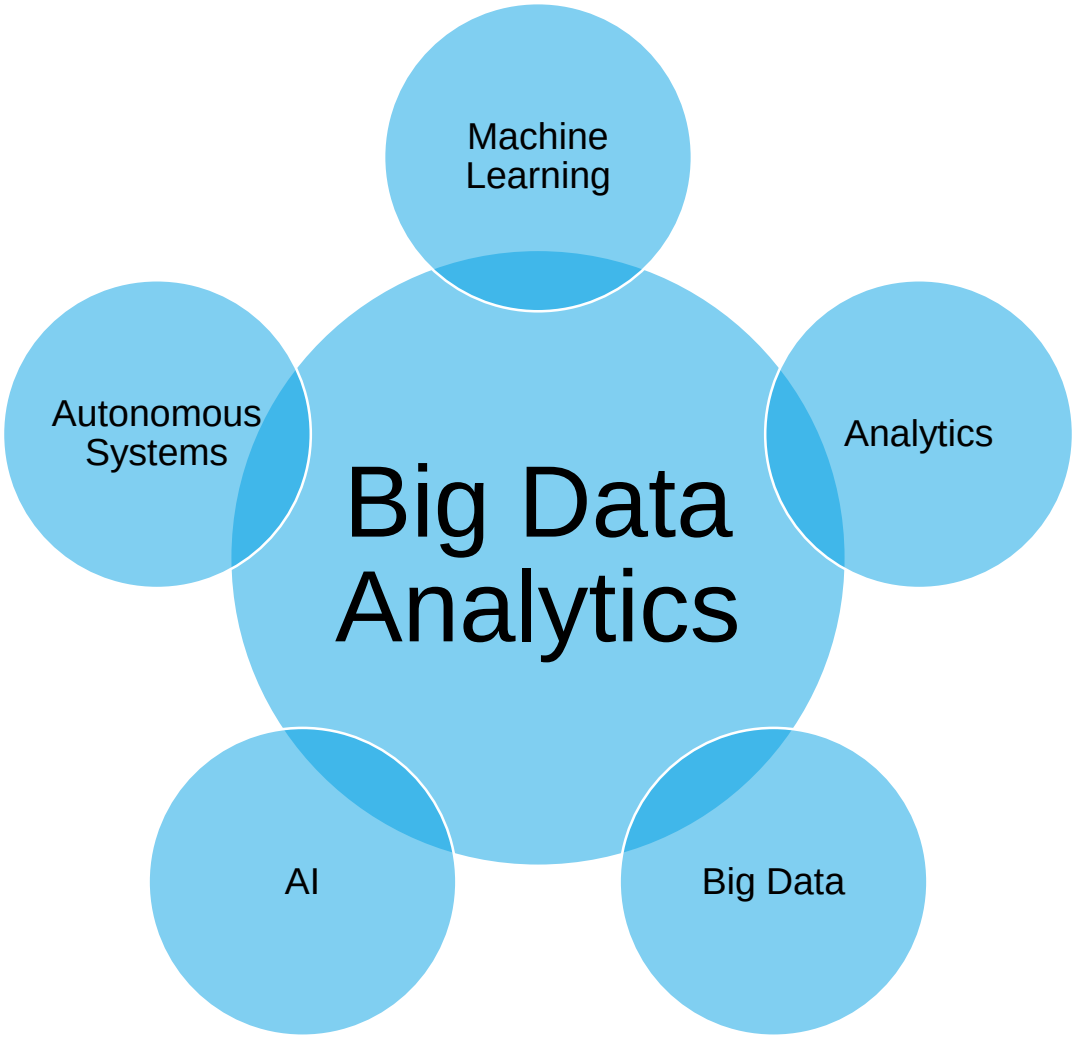
- Let’s define Big Data Analytics
- Use cases: How will intelligent algorithms change every sector it comes in contact with
  - Intelligent Algorithms, Fad or Truth?
  - Big Data Analytics
  - Machine Learning
- What is quality from a management perspective?
- IT-tools and services at Arcada
  - **Tomorrow (7.9) 11-13, walk in studio in F368 BDA Lab** (you will need a key)
  - Andrej Scherbakov (advisor), he will monitor the channel next week
  - Use Slack for all (non-personal) communication

ARCADA

Definitions

Big Data Analytics (BDA)

The field and some important terms



Defining Analytics

- INFORMS:
  - The scientific process of **transforming data into insights** for the purpose of **making better decisions**.
  - Analytics is always an **action-driven approach** and a decision is to be made when we look at doing analytics.
  - Data scientists love to analyze data just for the sake of analyzing it. However, it is important to **ensure our analysis is driving business action**.
  - We want analytics to empower an organization's vision.

Defining Artificial Intelligence (AI)

- “Artificial intelligence is that activity devoted to making machines intelligent, and intelligence is that quality that enables an entity to function appropriately and with foresight in its environment.”
  - Nils J. Nilsson, *The Quest for Artificial Intelligence: A History of Ideas and Achievements*
- To date human intelligence has no match in the biological and artificial worlds for sheer versatility, with the abilities “to reason, achieve goals, understand and generate language, perceive and respond to sensory inputs, prove mathematical theorems, play challenging games, synthesize and summarize information, create art and music, and even write histories.” - Nilsson
- AI research trends: https://ai100.stanford.edu/2016-report/section-i-what-artificial-intelligence/ai-research-trends

Defining Machine Learning (ML)

- Machine Learning gives "computers the ability to learn without being explicitly programmed.“ – Samuel
- Examples:
  - The **self-driving** Tesla/Google car? The essence of machine learning.
  - Online **recommendation** offers such as those from Amazon and Netflix? Machine learning applications for everyday life.
  - **Knowing** what customers are saying **about** you on Twitter? Machine learning combined with linguistic rule creation.
  - Fraud **detection**? One of the more obvious, important uses in our world today.

The FOUR V's of Big Data

From traffic patterns and music downloads to web history and medical records, data is recorded, stored, and analyzed to enable the technology and services that the world relies on every day. But what exactly is big data, and how can these massive amounts of data be used?

As a leader in the sector, IBM data scientists break big data into four dimensions: **Volume, Velocity, Variety and Veracity**

Depending on the industry and organization, big data encompasses information from multiple internal and external sources such as transactions, social media, enterprise content, sensors and mobile devices. Companies can leverage data to adapt their products and services to better meet customer needs, optimize operations and infrastructure, and find new sources of revenue.

By 2015

4.4 MILLION IT JOBS

will be created globally to support big data, with 1.9 million in the United States

Volume

SCALE OF DATA

40 ZETTABYTES

[43 TRILLION GIGABYTES]

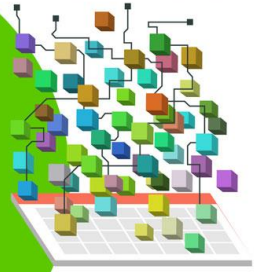
of data will be created by 2020, an increase of 300 times from 2005

It's estimated that

2.5 QUINTILLION BYTES

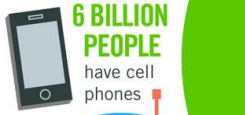
[2.3 TRILLION GIGABYTES]

of data are created each day



6 BILLION PEOPLE

have cell phones



WORLD POPULATION: 7 BILLION

Most companies in the U.S. have at least

100 TERABYTES

[100,000 GIGABYTES]

of data stored

Variety

DIFFERENT FORMS OF DATA

As of 2011, the global size of data in healthcare was estimated to be

150 EXABYTES

[161 BILLION GIGABYTES]

30 BILLION PIECES OF CONTENT

are shared on Facebook every month

By 2014, it's anticipated there will be

420 MILLION WEARABLE, WIRELESS HEALTH MONITORS

4 BILLION+ HOURS OF VIDEO

are watched on YouTube each month

400 MILLION TWEETS

are sent per day by about 200 million monthly active users

Velocity

ANALYSIS OF STREAMING DATA

The New York Stock Exchange captures

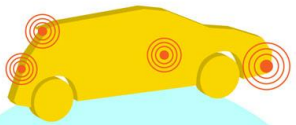
1 TB OF TRADE INFORMATION

during each trading session

Modern cars have close to

100 SENSORS

that monitor items such as fuel level and tire pressure



By 2016, it is projected there will be

18.9 BILLION NETWORK CONNECTIONS

– almost 2.5 connections per person on earth

Veracity

UNCERTAINTY OF DATA

1 IN 3 BUSINESS LEADERS

don't trust the information they use to make decisions

27% OF RESPONDENTS

in one survey were unsure of how much of their data was inaccurate

Poor data quality costs the US economy around

$3.1 TRILLION A YEAR

Sources: McKinsey Global Institute, Twitter, Cisco, Gartner, EMC, SAS, IBM, MEPTEC, QAS

IBM

Autonomous Systems

- Often a “node” that exists and acts in a distributed network, examples can be found from:
  - Vehicles
  - Robotics
  - Blockchain
  - Intelligent agents
  - IoT
- The node should independently be able to react to data from its surrounding environment, in a sort of feedback loop.
  - Observe -> detect possible actions -> react

What is Big Data Analytics?

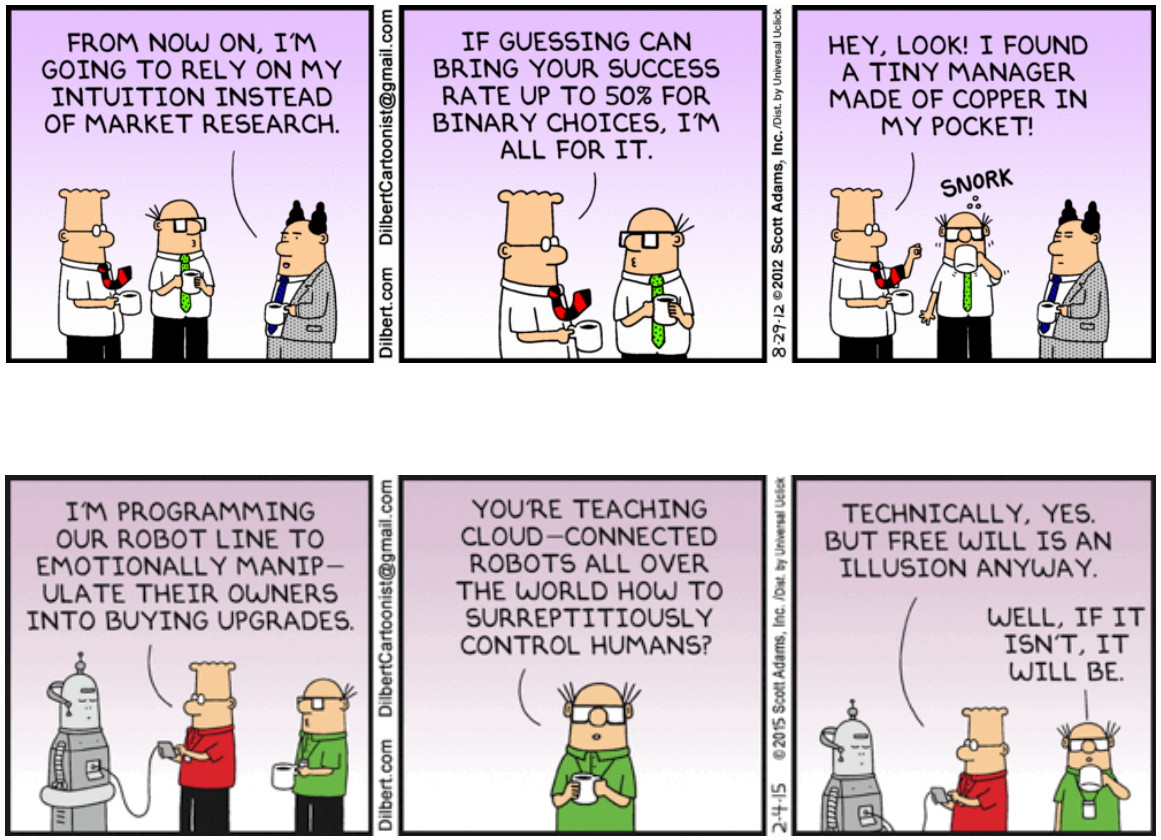
- IBM:

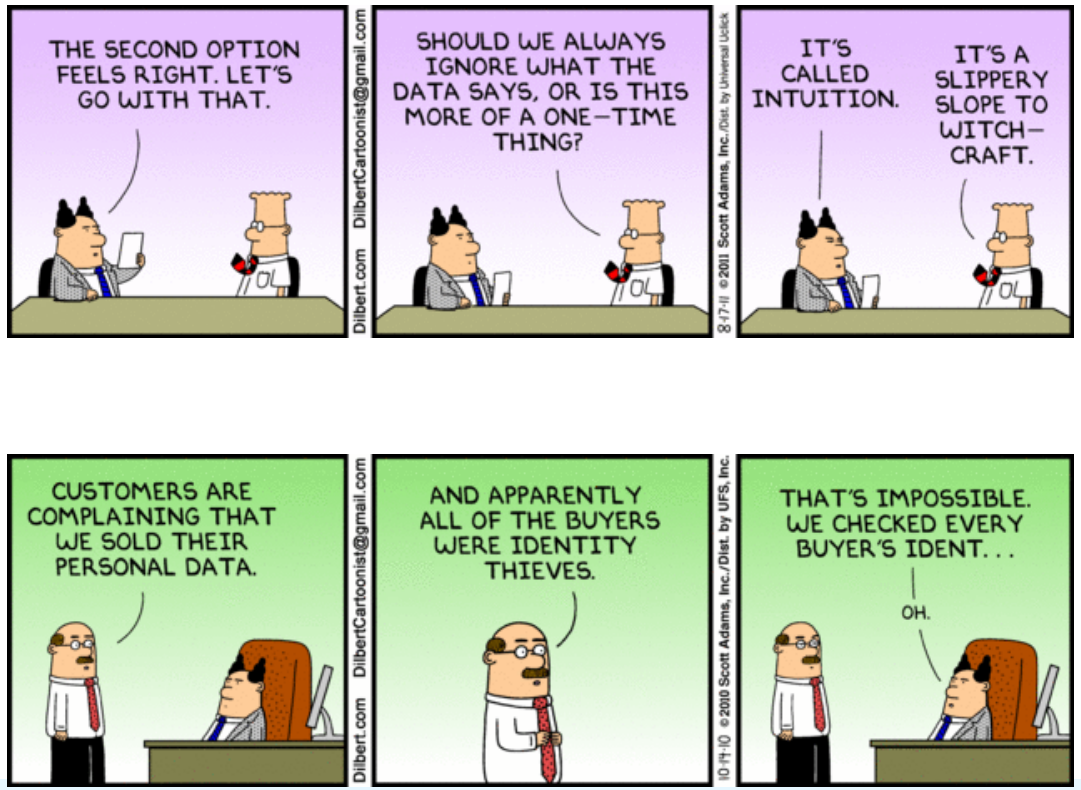
  Big data analytics is the use of advanced analytic techniques against very large, diverse data sets that include different types such as structured/unstructured and streaming/batch, and different sizes from terabytes to zettabytes.

- SAS:

  Big data analytics examines large amounts of data to uncover hidden patterns, correlations and other insights.

Mgmt intuition vs. Data insights





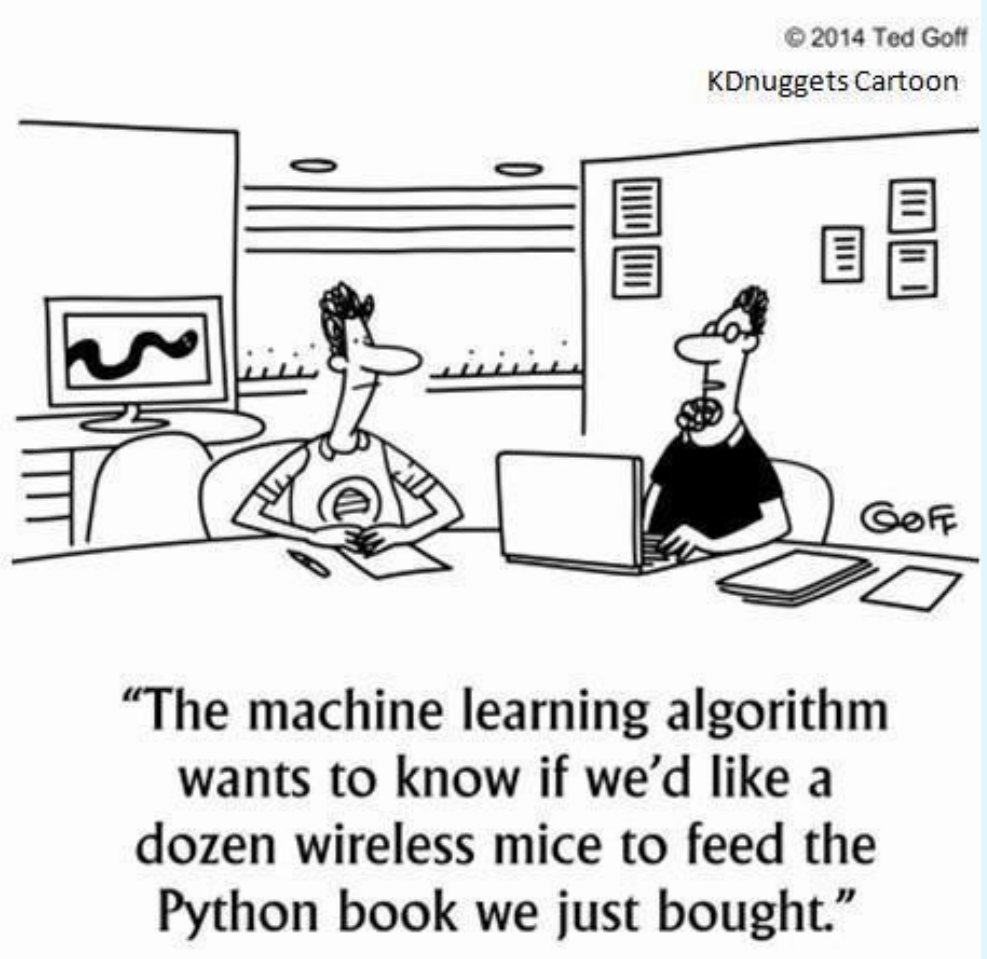
Accenture's view, intuition + insight

- The hype around data and analytics may lead many executives to assume that the answers to difficult questions—how much to raise prices, where to site a new retail store, whether one product will cannibalize another—can be found through aggressive number crunching. But does that mean that expert judgment and managerial intuition are obsolete? Hardly.
- In fact, the **full potential of quantitative analytics can be unlocked only when combined with business intuition**.

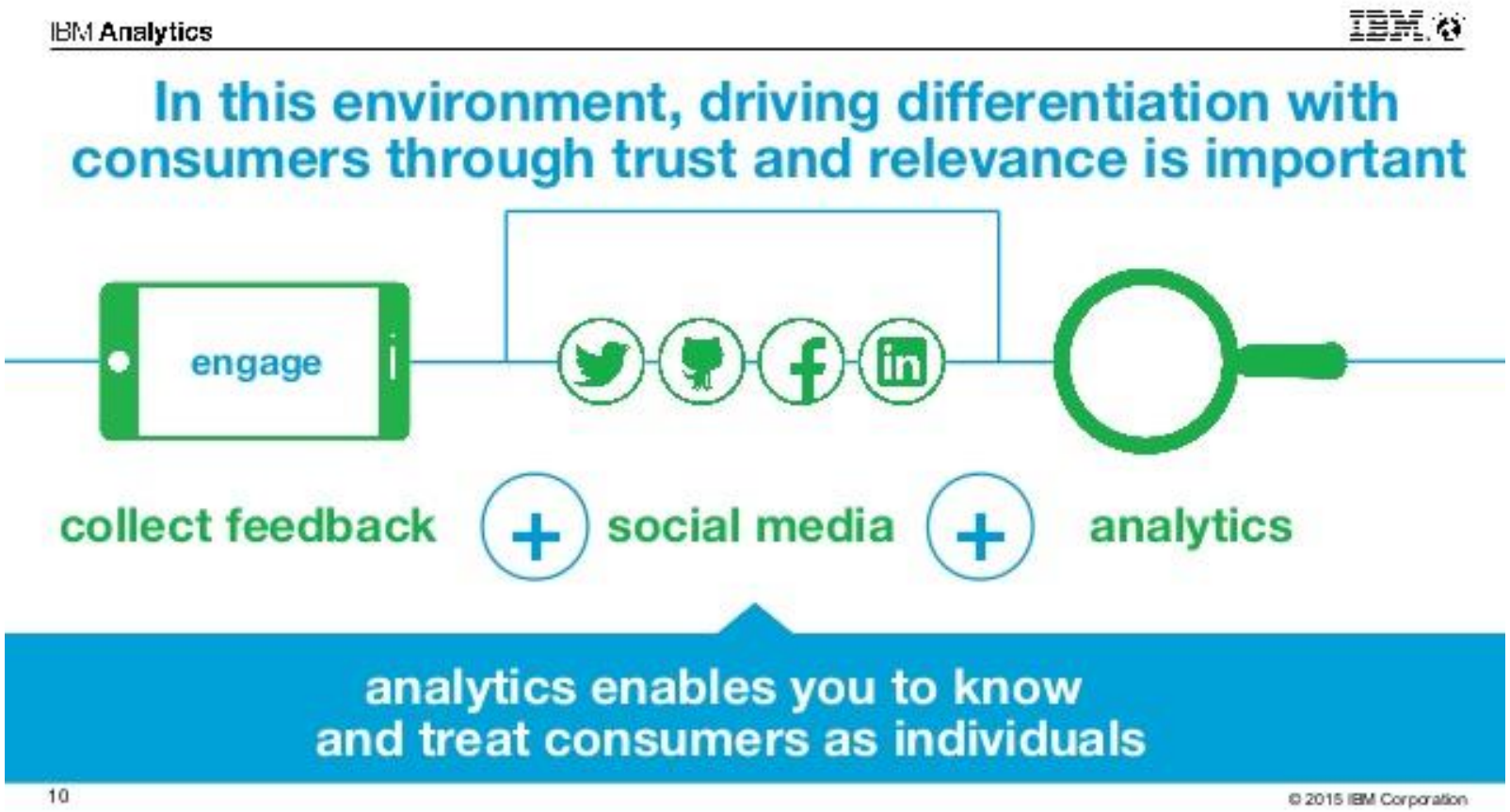
ARCADA

Its all about quality data!

Intelligent Services: A few examples

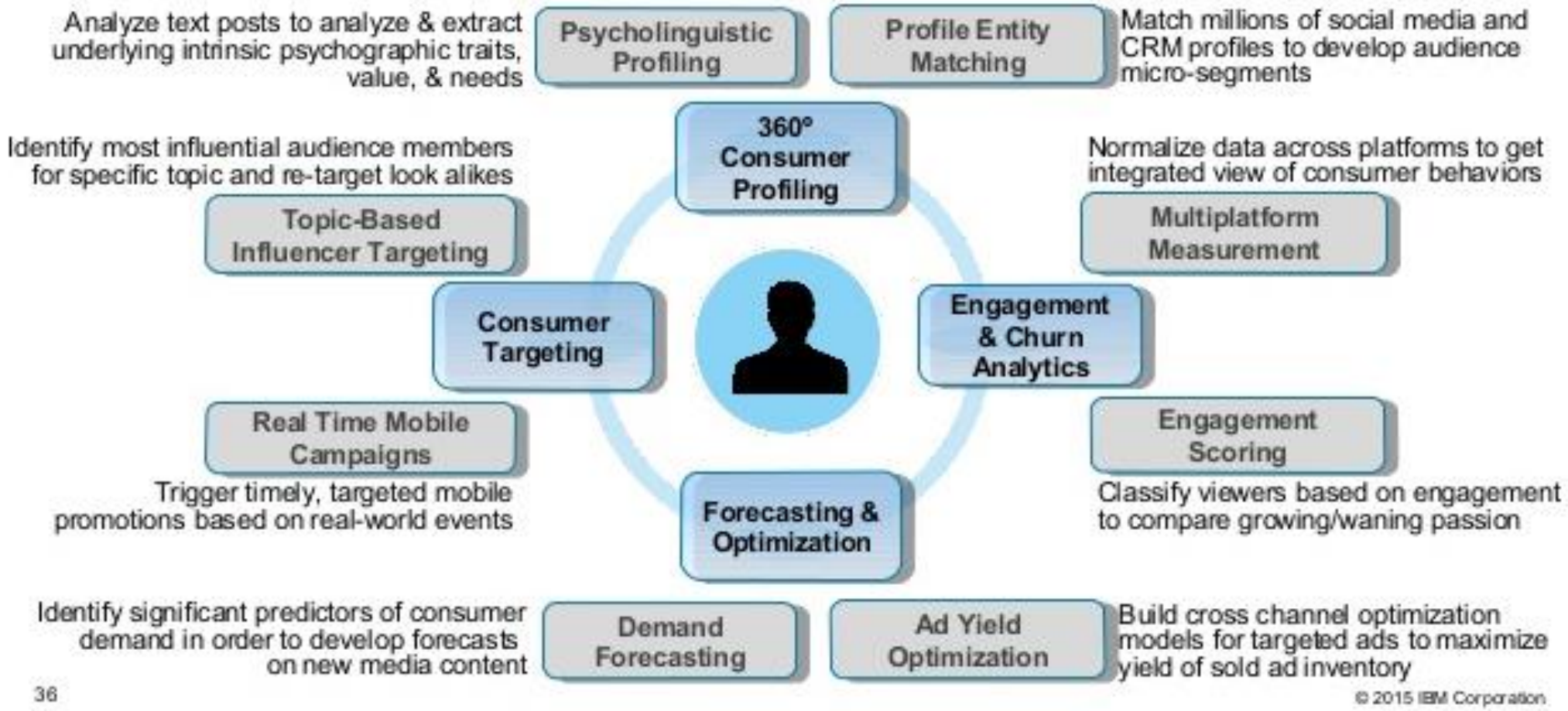


The Big Data Revolution in Consumer Behavior



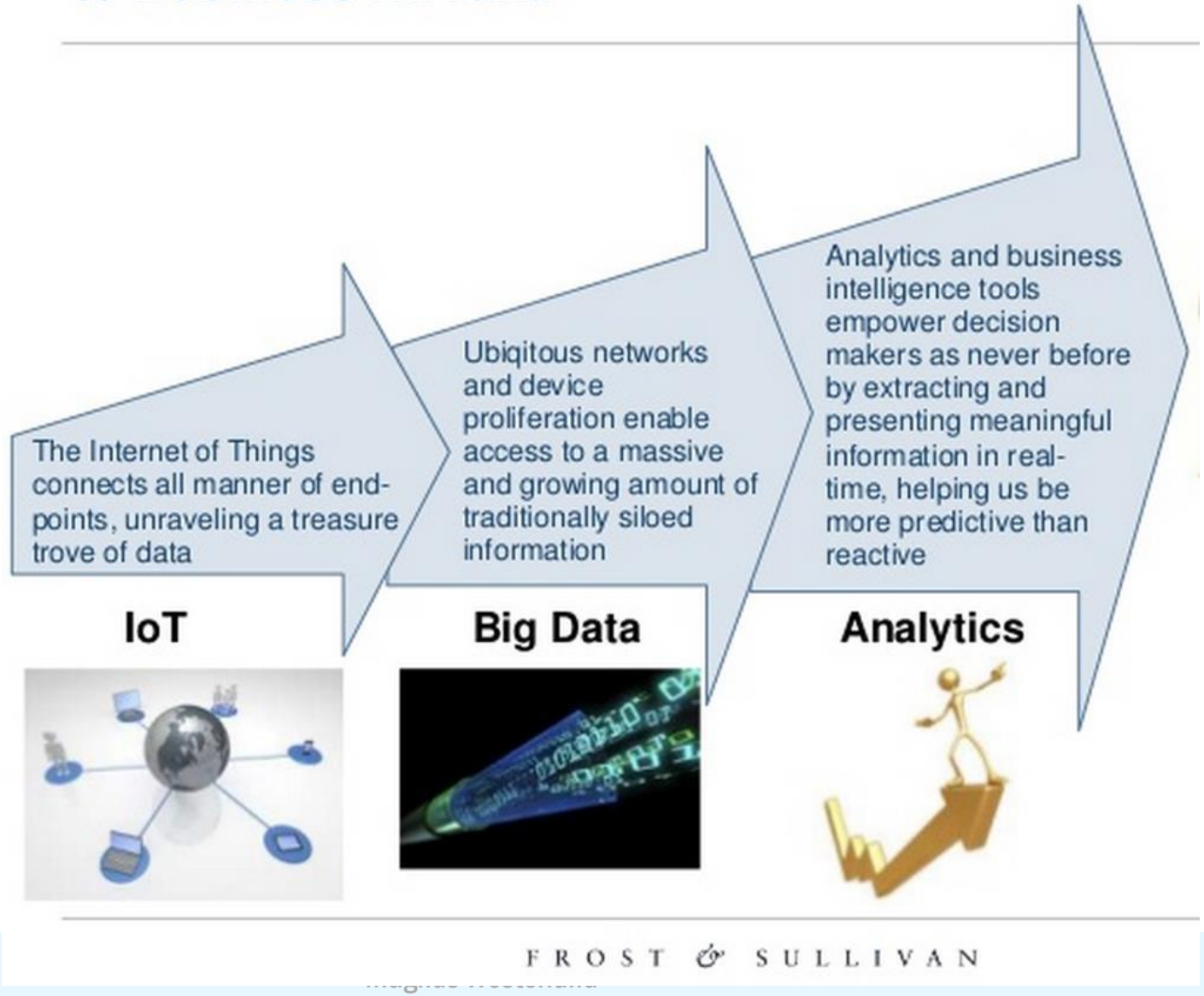
Value creation in B2C

A consumer-centric view in value creation for B2C

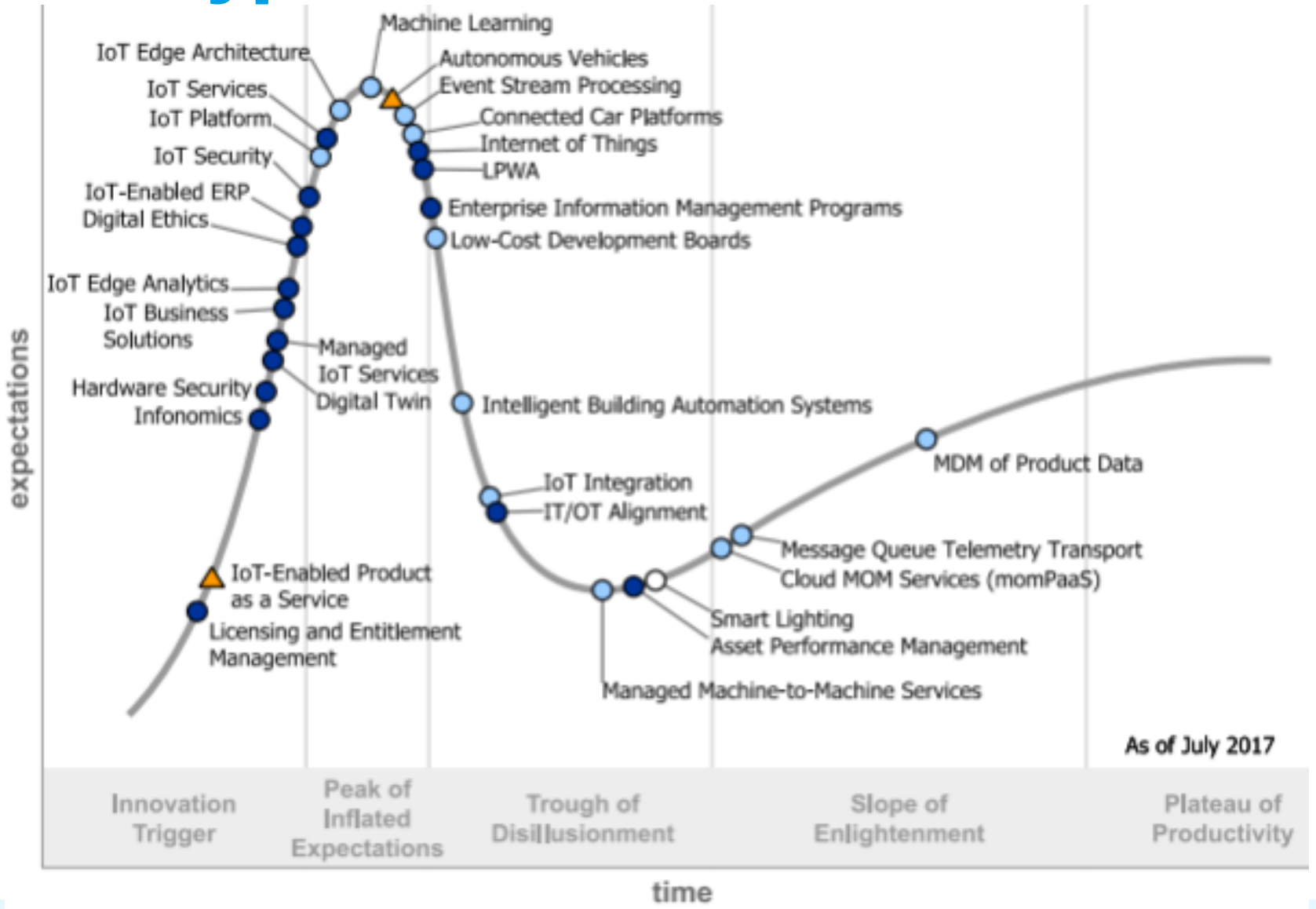


36

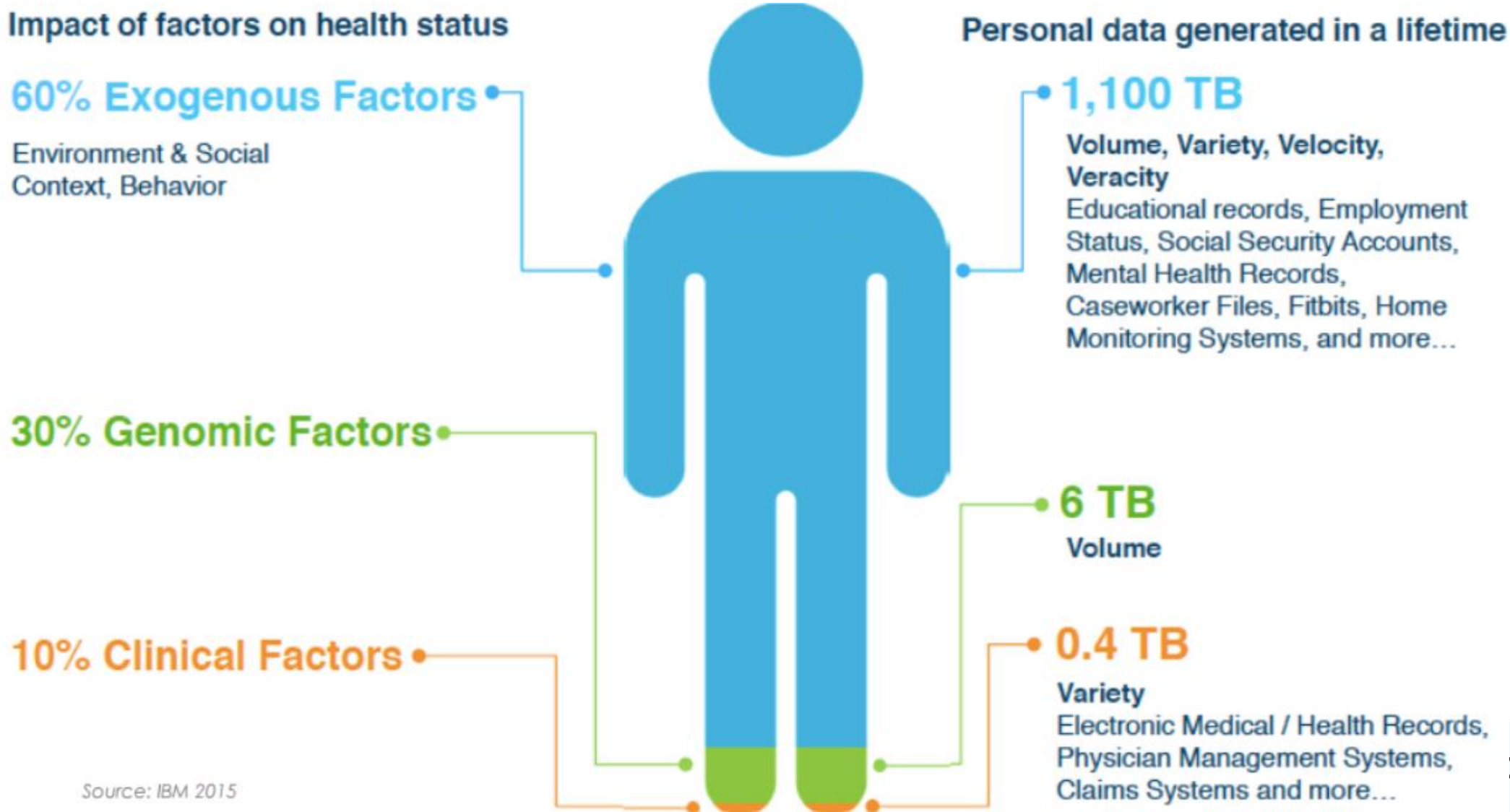
Convergence of technologies, enabled by IoT



Gartner Hype Curve



The Big Data Revolution in Healthcare



Source: IBM 2015

Example Areas in Health Care

Diagnostic

- Remote vital monitoring
- Sleep/pulmonary monitoring
- Neuromonitoring
- Clinical decision support

Wellness

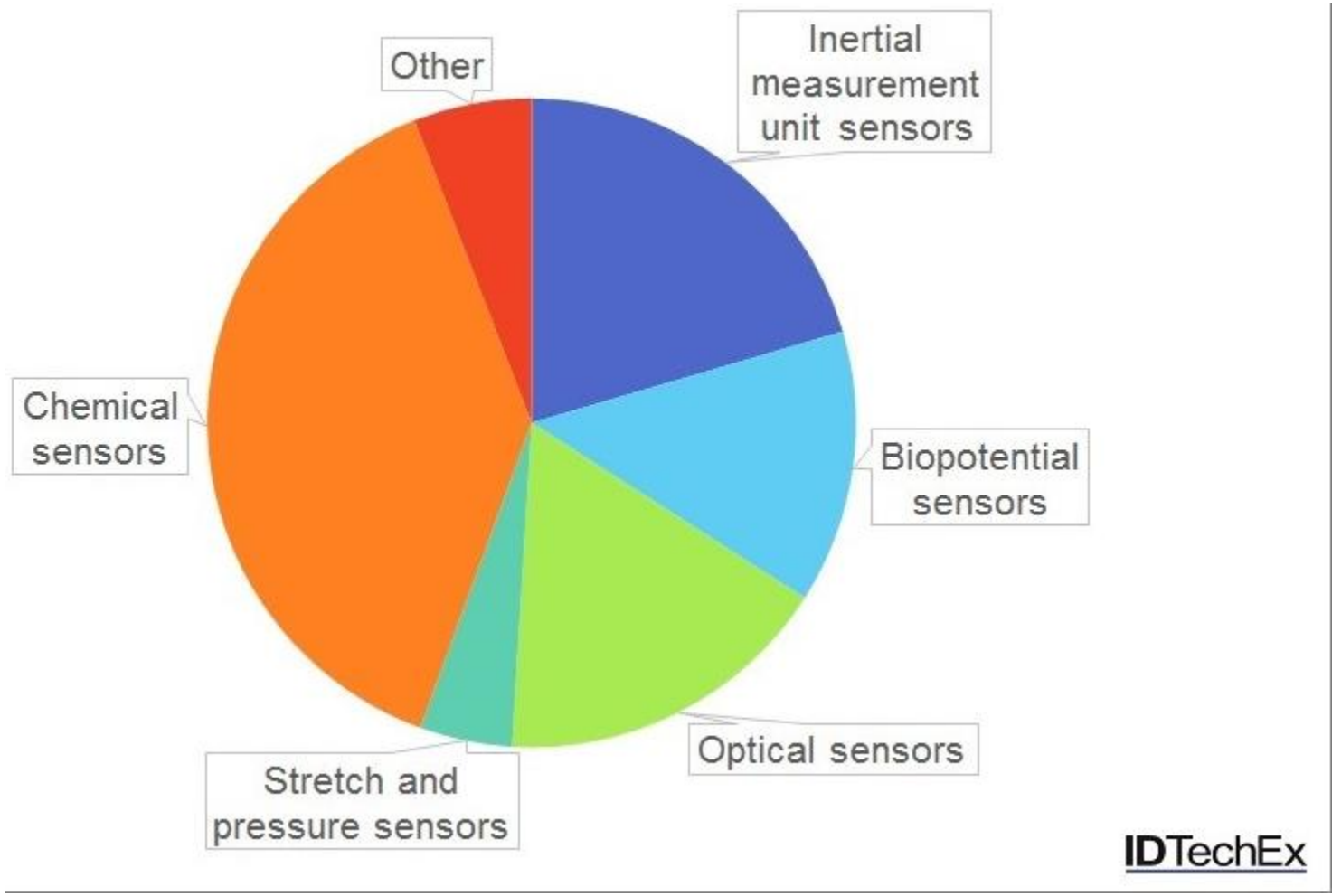
- Fitness management
- Preventative care
- Aging in Place

Therapeutic

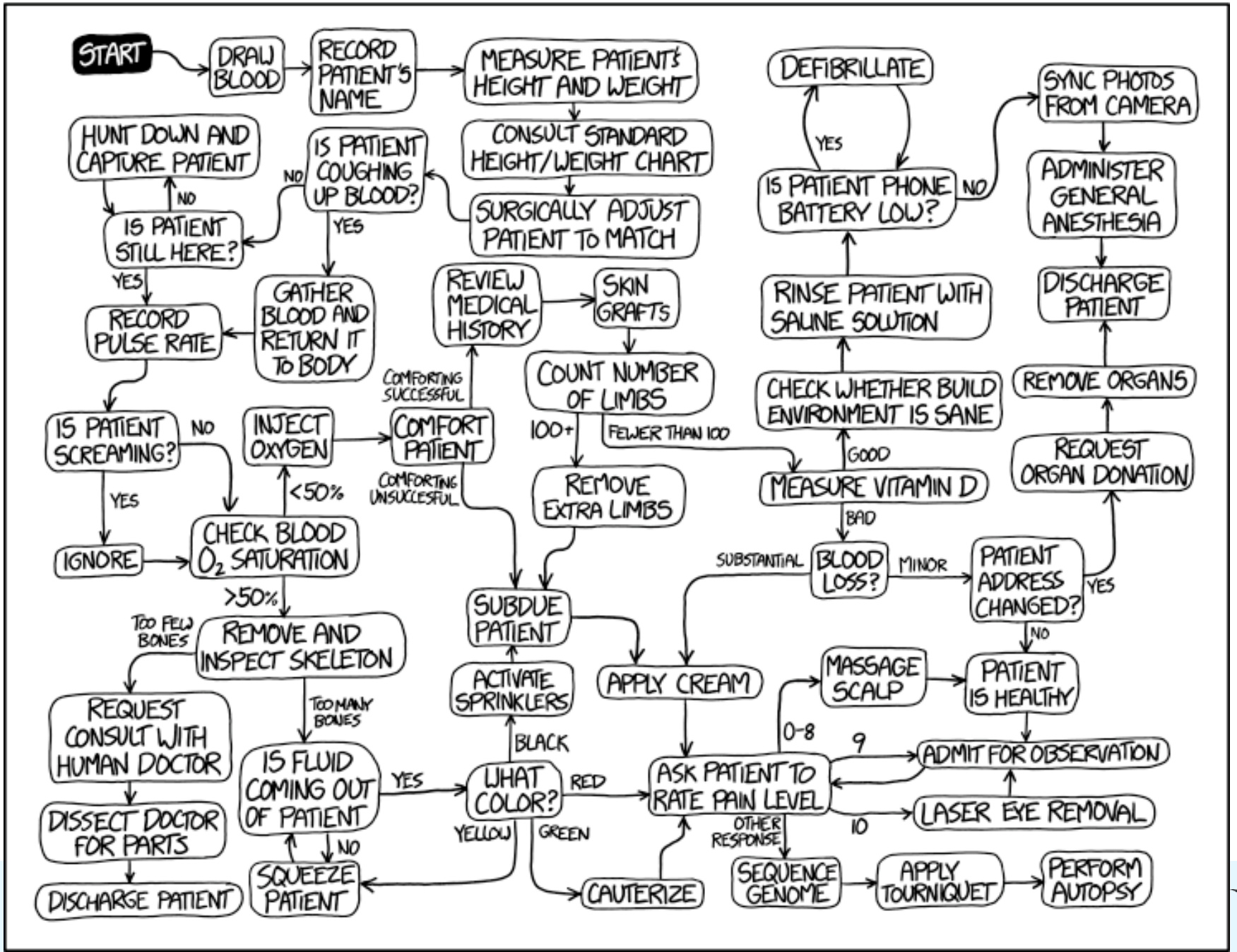
- Pain management
- Drug delivery
- Optical assist
- Surgical assist
- Pulmonary assist
- Care delivery

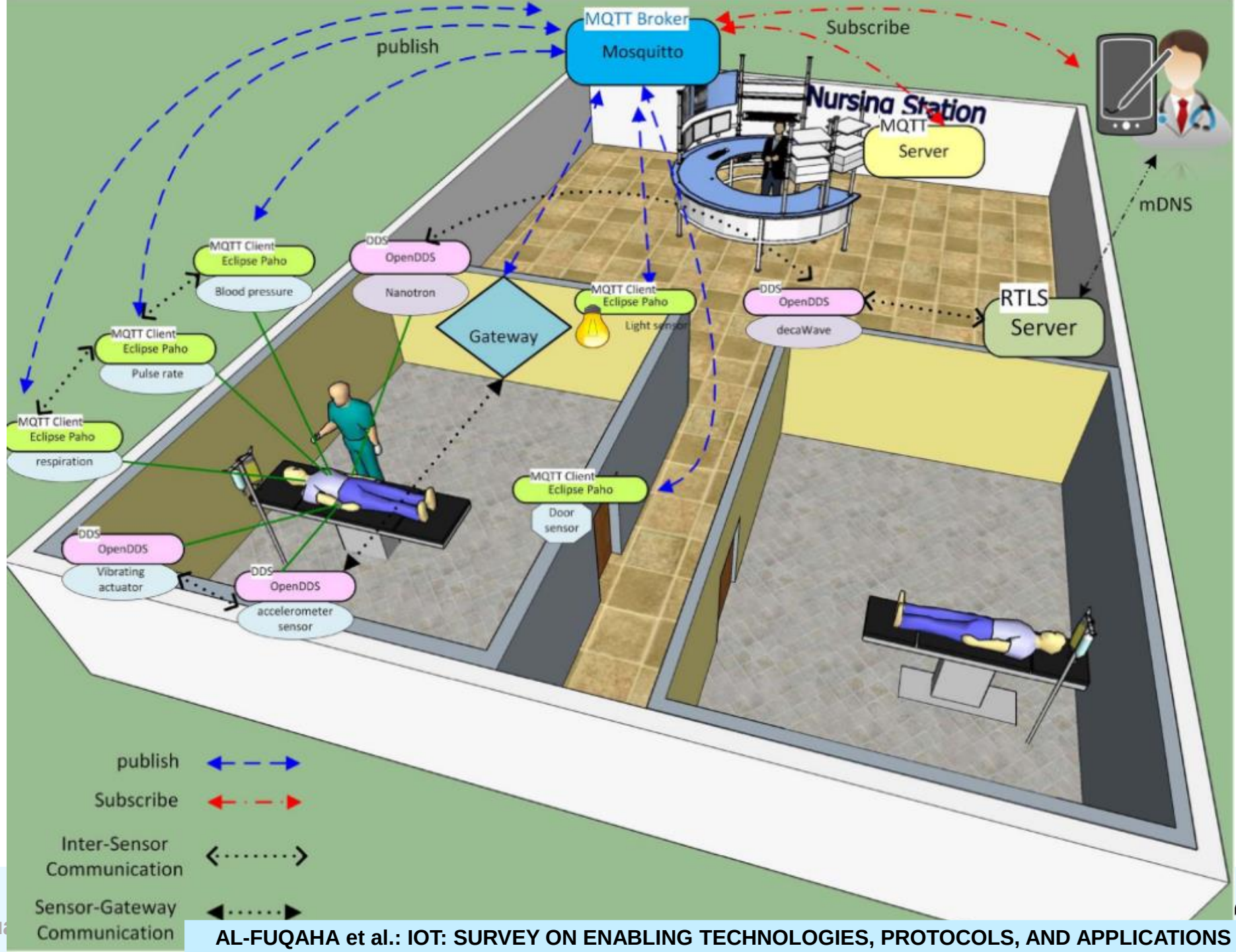
Operational

- Security, access control
- Pharmacy management
- Inventory management
- Training delivery

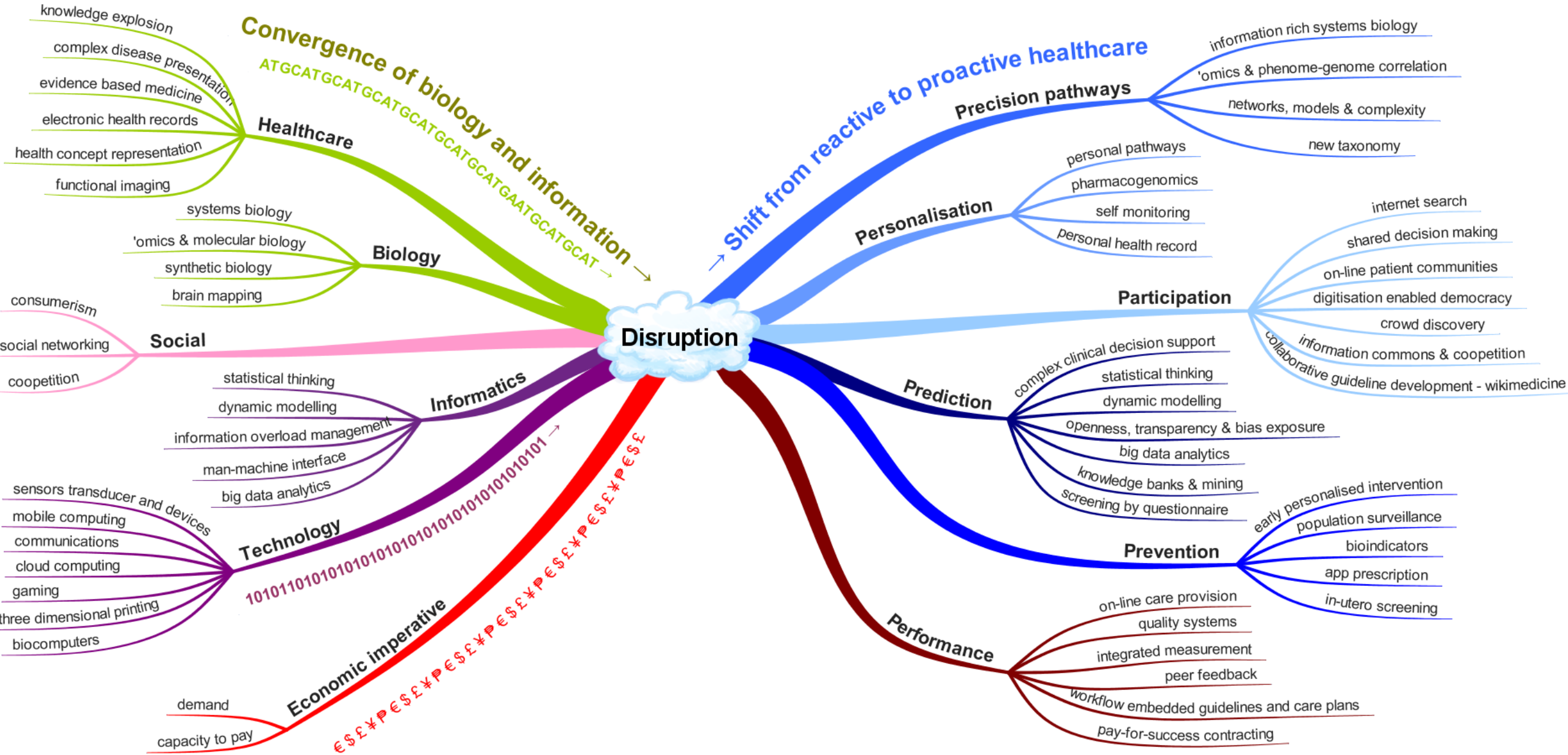


A GUIDE TO THE MEDICAL DIAGNOSTIC AND TREATMENT ALGORITHM USED BY IBM'S WATSON COMPUTER SYSTEM




MQTT Broker
Mosquitto
publish
Subscribe
Nursing Station
MQTT
Server
mDNS
MQTT Client
Eclipse Paho
Blood pressure
DDS
OpenDDS
Nanotron
Gateway
MQTT Client
Eclipse Paho
Light sensor
DDS
OpenDDS
decaWave
RTLS
Server
MQTT Client
Eclipse Paho
Pulse rate
MQTT Client
Eclipse Paho
respiration
MQTT Client
Eclipse Paho
Door
sensor
DDS
OpenDDS
Vibrating
actuator
DDS
OpenDDS
accelerometer
sensor
publish
Subscribe
Inter-Sensor
Communication
Sensor-Gateway
Communication

Transforming Healthcare with Analytics



ARCADA

Building intelligent platforms and services for eHealth and identifying their challenges

Break

ARCADA

Building intelligent platforms and services for a sector and identifying their challenges

Challenges with growing data sets

What is Big Data Analytics?

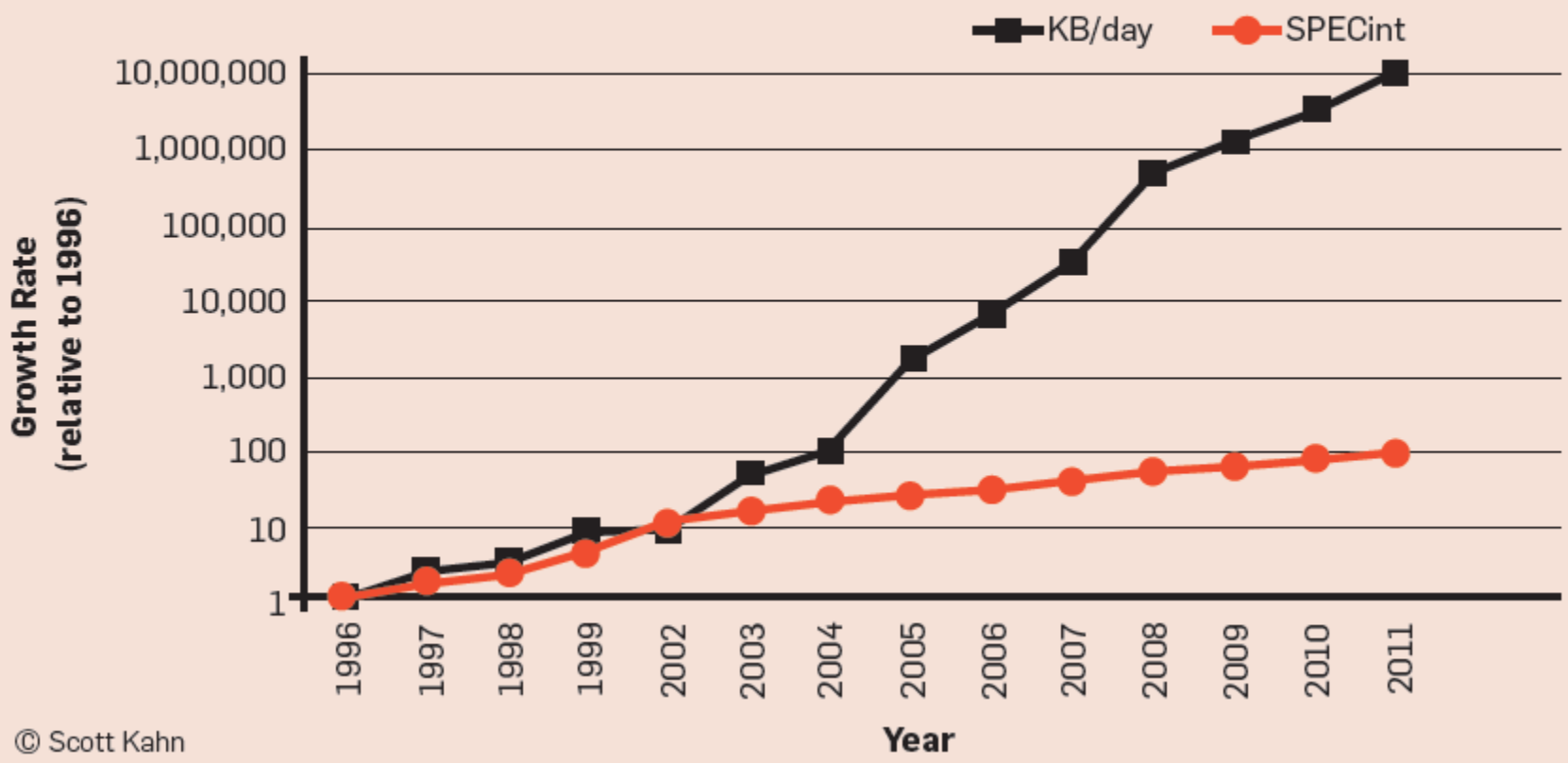
- *Its about high-volume, high-velocity and highly varied information assets that demand cost-effective, innovative forms of information processing to support enhanced insight and decision making.*

  (Gartner)

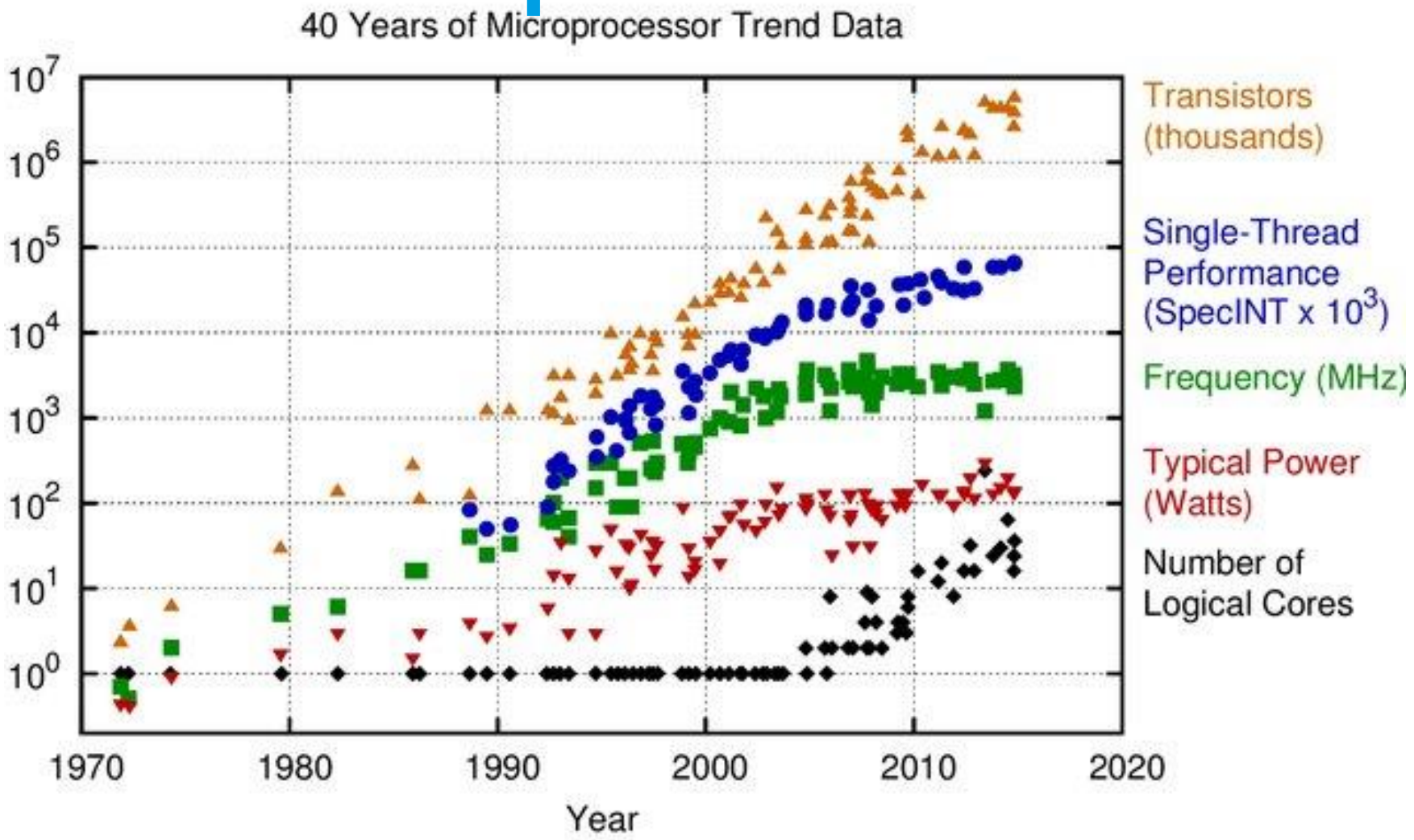
  My personal view:
  - Making sense of potentially large amounts of data
  - Employing complex real-time data (streams) to create analytical services
  - Models learn from existing data to make predictions/decisions with new data; ultimately models learn continuously from new data
  - Being able to process on incomplete data
  - Maintaining these big data analytics services and platforms

Data growth vs. Processing capacity

Figure 1. Next-gen sequence data size compared to SPECint.



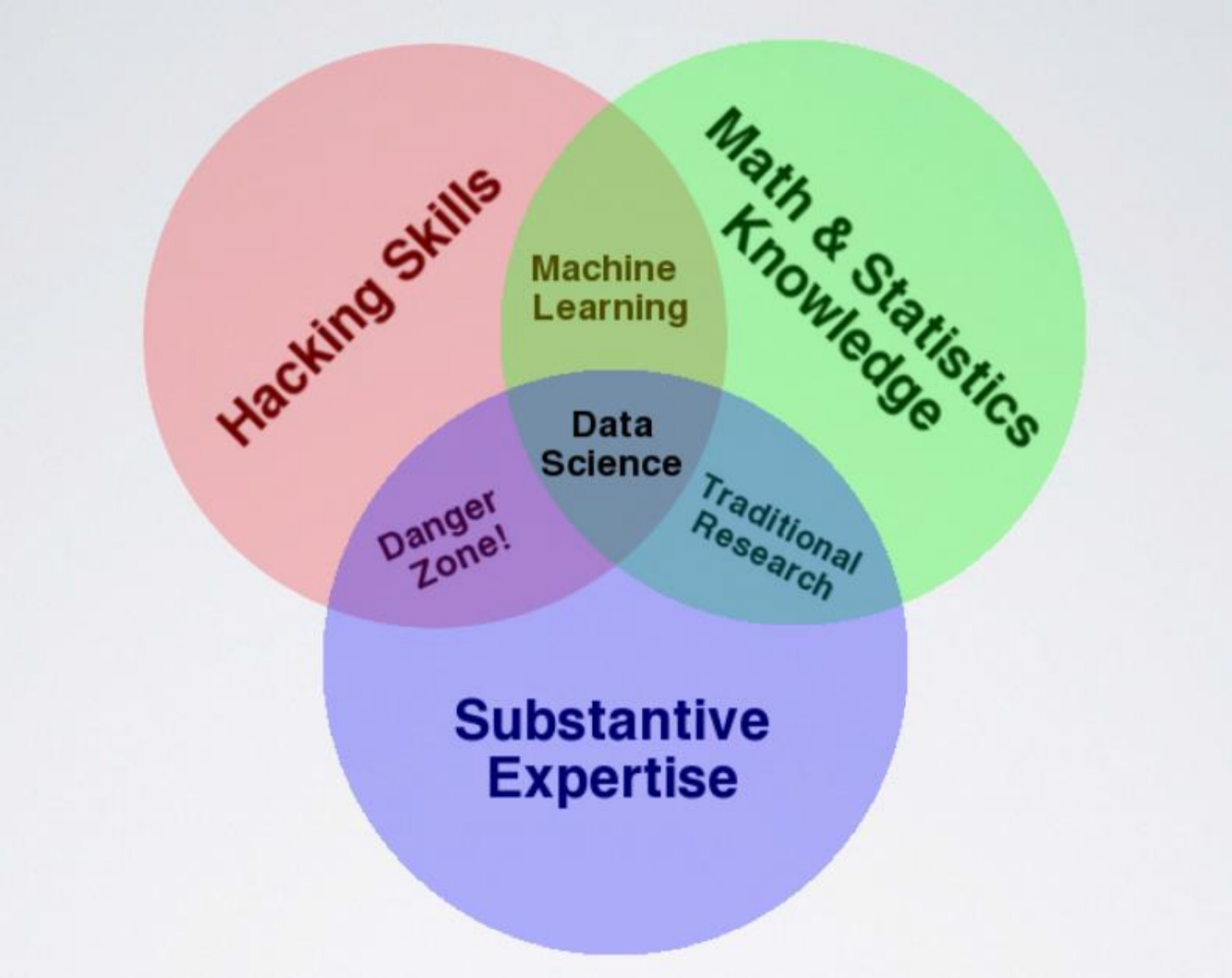
Processor development



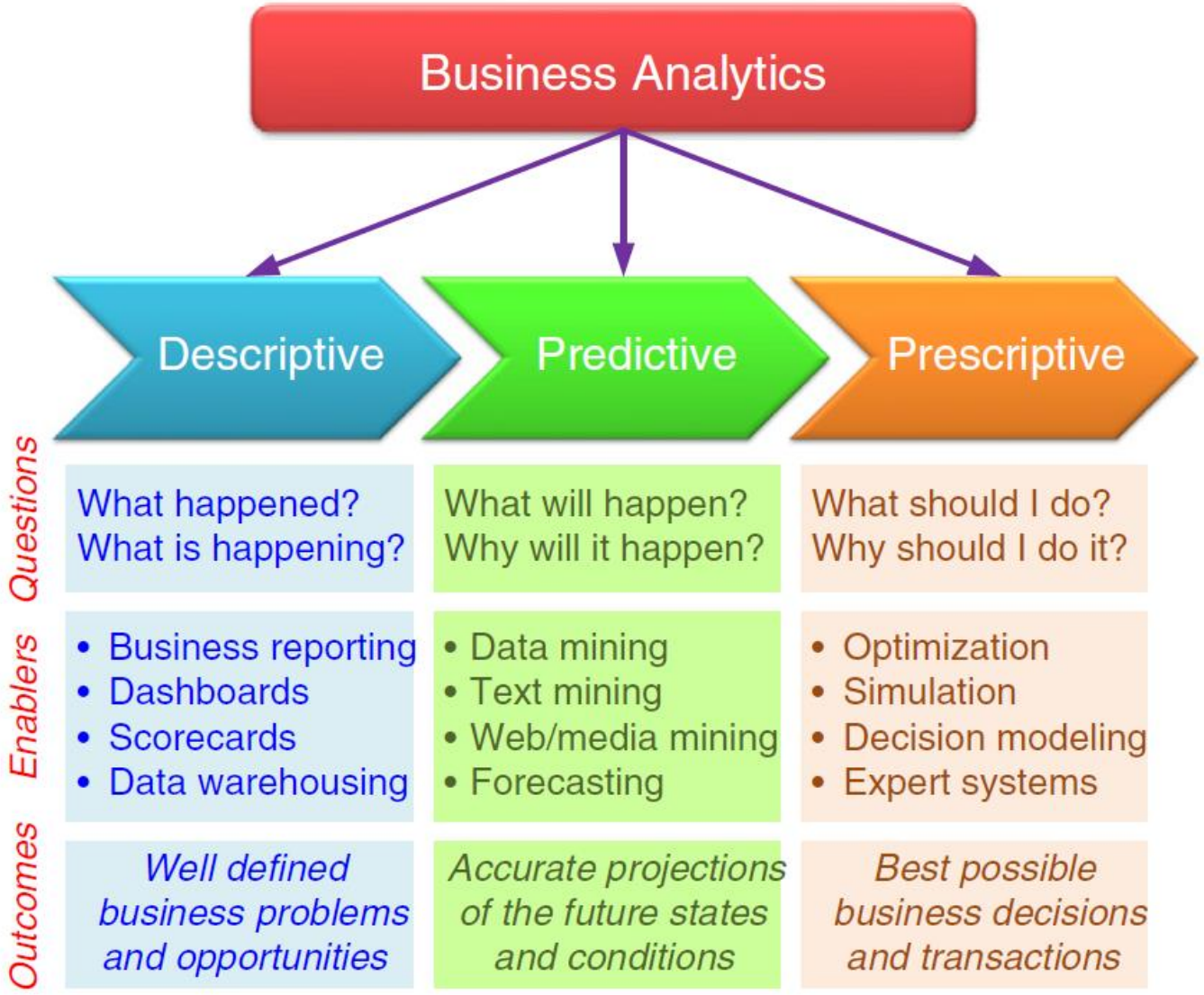
Original data up to the year 2010 collected and plotted by M. Horowitz, F. Labonte, O. Shacham, K. Olukotun, L. Hammond, and C. Batten
New plot and data collected for 2010-2015 by K. Rupp

Working smarter - what is Data Science?

- Drew Convey's
- Venn Diagram

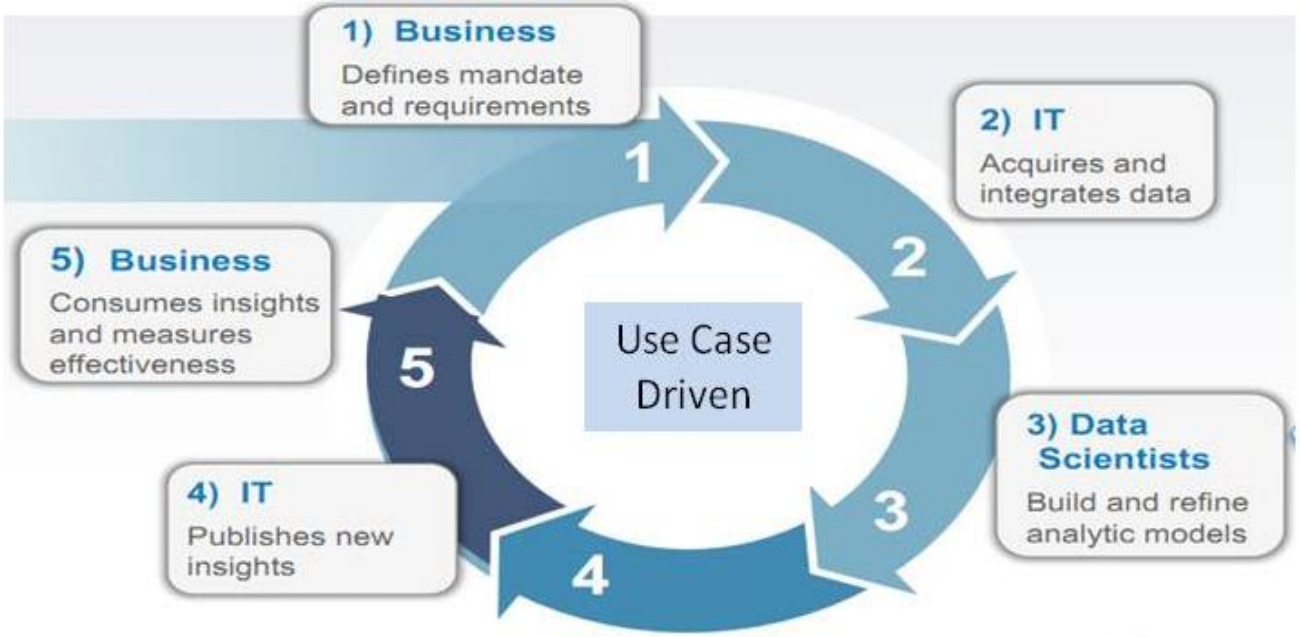


Data science in business - Analytics taxonomy



D. Delen, H. Demirkan, Data, information and analytics as services, Decis. Support Syst. (2012)

What do consultancies sell?



Focus on an organization's **Strategic Business Initiatives** in order to be relevant to the business

- Public statement of business intent
- Delivers compelling business value
- Cross-functional
- Championed by a senior executive
- Has measurable goals
- Has well-defined delivery timeframe

Use "**visioning methodology**" to identify "where" and "how" Big Data Analytics can power an organization's Strategic Business Initiatives

ARCADA

Building intelligent platforms and services for a sector and identifying their challenges

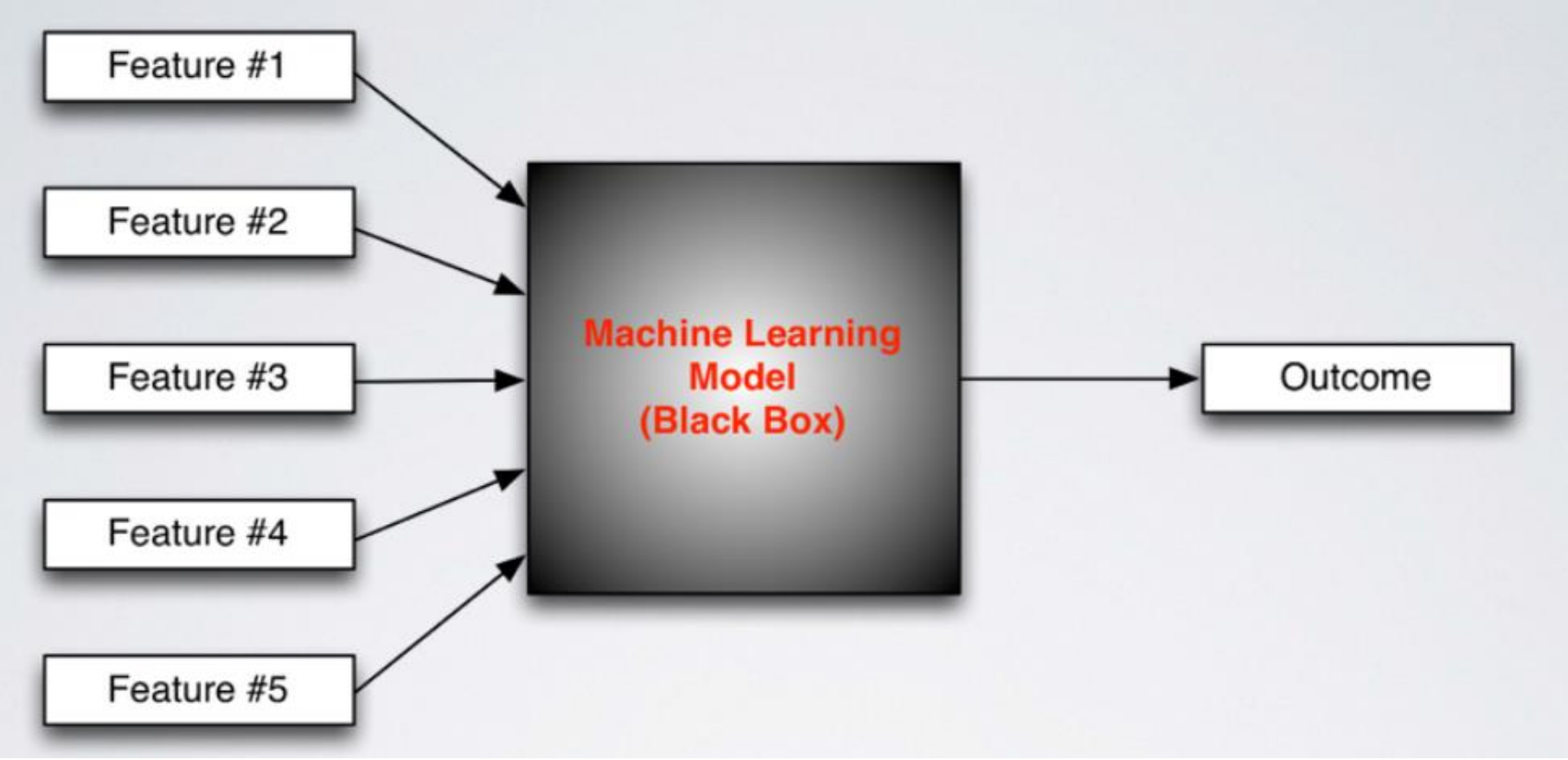
Machine Learning

What is Machine Learning?

Defining the aim of modeling:

- **Clustering:** Group records together that have similar field values. Often used for recommendation systems. (e.g. group customers with similar buying habits)
- **Regression:** Learn to predict a numeric outcome field, based on all of the other fields present in each record. (e.g. predict a student's graduating GPA)
- **Classification:** Learn to predict a non-numeric outcome field. (e.g. predict the field of a student's first job after graduation)

What is ML model?



Titanic - Example Application

- Predict the outcome:
  - Survived
  - Perished
- From passenger features:
  - Gender
  - Name
  - Passenger class
  - Age
  - Family members present
  - Port of embarkation
  - Cabin
  - Ticket

The original labels (data)

Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	3	Owen Harris	male	22	1	0	A/5 21171	7.25		S
1	1	riggs Thayer)	female	38	1	0	PC 17599	71.2833	C85	C
1	3	ı, Miss. Laina	female	26	0	0	O2. 3101282	7.925		S
1	1	ily May Peel)	female	35	1	0	113803	53.1	C123	S
0	3	/illiam Henry	male	35	0	0	373450	8.05		S
0	3	n, Mr. James	male		0	0	330877	8.4583		Q
0	1	Mr. Timothy J	male	54	0	0	17463	51.8625	E46	S
0	3	ɔsta Leonard	male	2	3	1	349909	21.075		S
1	3	elmina Berg)	female	27	0	2	347742	11.1333		S
1	2	dele Achem)	female	14	1	0	237736	30.0708		C
1	3	ırguerite Rut	female	4	1	1	PP 9549	16.7	G6	S
1	1	iss. Elizabeth	female	58	0	0	113783	26.55	C103	S
0	3	/illiam Henry	male	20	0	0	A/5. 2151	8.05		S
0	3	ınders Johan	male	39	1	5	347082	31.275		S
0	3	nda Adolfina	female	14	0	0	350406	7.8542		S

Creating an input feature vector

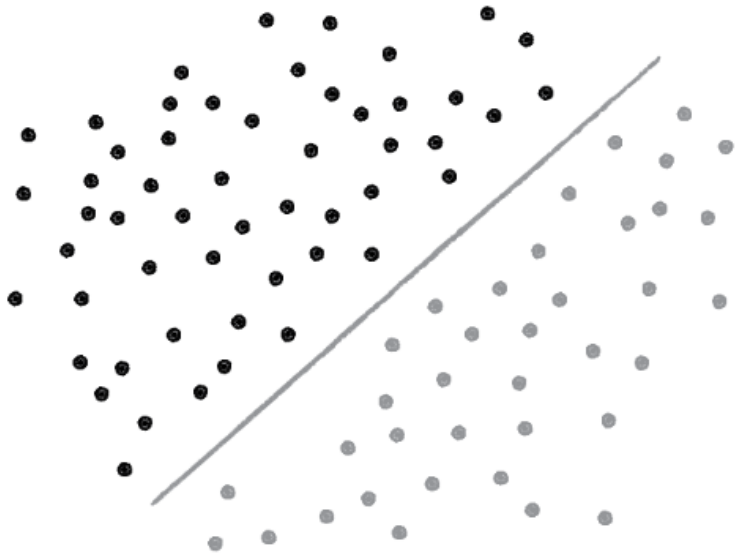
- **Age**: The interpolated age normalized to -1 to 1.
- **Sex-male**: The gender normalized to -1 for female, 1 for male.
- **Pclass**: The passenger class [1-3] normalized to -1 to 1.
- **Sibsp**: Value from the original data set normalized to -1 to 1.
- **Parch**: Value from the original data set normalized to -1 to 1.
- **Fare**: The interpolated fare normalized to -1 to 1.
- **Embarked-c**: The value 1 if the passenger embarked from Cherbourg, -1 otherwise.
- **Embarked-q**: The value 1 if the passenger embarked from Queenstown, -1 otherwise.
- **Embarked-s**: The value 1 if the passenger embarked from Southampton, -1 otherwise.
- **Name-mil:** The value 1 if passenger had a military prefix, -1 otherwise.
- **Name-nobility**: The value 1 if passenger had a noble prefix, -1 otherwise.
- **Name-Dr.**: The value 1 if passenger had a doctor prefix, -1 otherwise.
- **Name-clergy**: The value 1 if passenger had a clergy prefix, -1 otherwise.

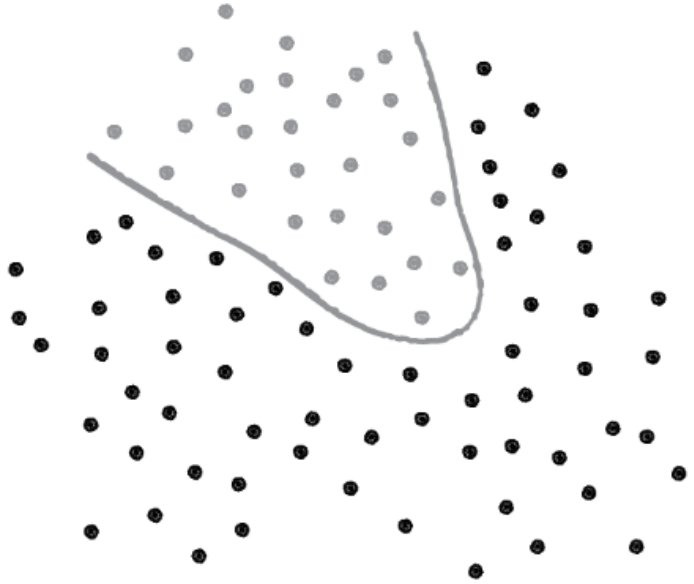
Understanding data through descriptive statistics: completeness and representativness

	#	Survived	Male Survived	Female Survied	Avg Age
Master	76	58%	58%		
Mr.	915	16%	16%		
Miss.	332			71%	21.8
Mrs	235			79%	36.9
Military	10	40%	40%		36.9
Clergy	12	0%	0%		41.3
Nobility	10	60%	33%	100%	41.2
Doctor	13	46%	36%	100%	43.6

What happens inside a ML model?

- Data points are projected into a multi-dimensional space
- Separation of data points, ex. by measuring distance
- The boundary is “learned” by the machine, linearly or non-linearly depending on model used





ARCADA

Tools in use

IT environment

Activate your Arcada IT account

- Start reading here:
  - https://start.arcada.fi/en/it-support/it-accounts-and-passwords
  - Click the link on the page to go to AAS

- You need to either verify using your Finnish bank online credentials or another Finnish University account (that belongs HAKA)
- If you do not have either, then contact it-support@arcada.fi

IT-systems at Arcada

- The first place to look for information and IT-tools:
  - https://start.arcada.fi/en
  - Look under “**tools**”
- Email and Office:
  - We have Office 365 licenses for you
  - Please configure your Arcada email to your mobile, I will send you info there at times, also course results.
- LMS - ItsLearning
  - Lecture slides and project submissions will come through here.

IT-systems at Arcada (cont.)

- Asta: student admin system
- Arbs: room bookings
- Research papers,
  - https://arcada.finna.fi/?lng=en-gb
  - https://scholar.google.fi/
- Authorization
  - We use two different systems to log in, psw are synced
  - ADFS for MSFT services
  - Luckan for others

Security

- Please note that no one will ever ask for your log-in details
- Also, note that we receive a great deal of **phishing emails**

- We are part of HAKA and Eduroam, so you can use your Arcada credentials at many Universities the world over.
  - Use: login@arcada.fi and your regular psw
  - Be careful though, the certificate should be valid!

Python Software Environment

- These tools will be needed:
  - Installing Anaconda, select Python 3.6 version: https://www.continuum.io/downloads
  - Jupyter Quick Start guide: https://jupyter-notebook-beginner-guide.readthedocs.io/en/latest/index.html
  - Jupyter Notebook - Basics: http://nbviewer.jupyter.org/github/jupyter/notebook/blob/master/docs/source/examples/Notebook/Notebook%20Basics.ipynb
  - A good Python IDE is Pycharm Community edition: https://www.jetbrains.com/pycharm/download/
  - An intro to the language can be found in these videos: https://www.youtube.com/playlist?list=PLQVvvaa0QuDe8XSftW-RAxdo6OmaeL85M

Questions?

- The end..