

Introduction to Data Science & related concepts

DR. PHAM QUOC TRUNG



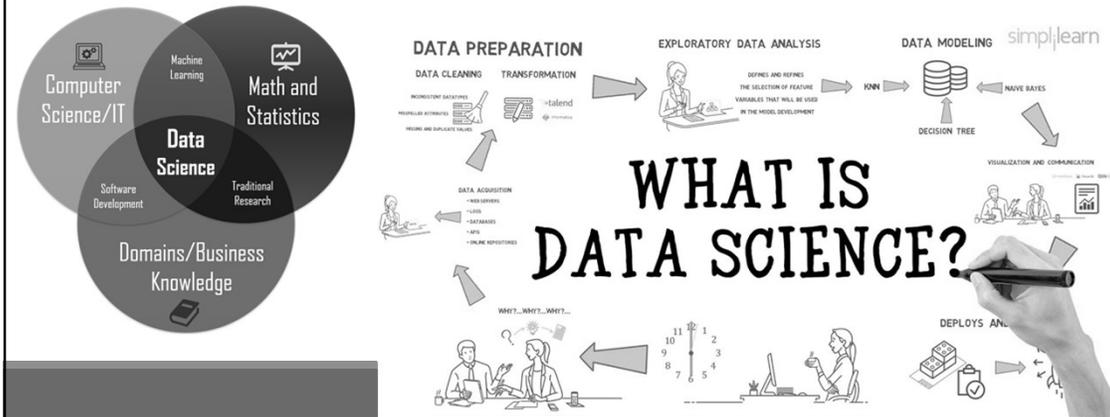
Main Contents

1. Data science
2. Data analytics
3. Business intelligence & Knowledge management
4. Big Data & Data mining
5. Data visualization
6. Artificial Intelligence & Machine Learning
7. Industry 4.0 & future technologies: cloud computing, IOT, VR, AR, Blockchain, FinTech, InsurTech...

1. Data Science (1)

Data Science

an interdisciplinary field about scientific methods, processes, and systems to extract knowledge or insights from data in various forms, either structured or unstructured, similar to data mining.



1. Data Science (2)

THE DATA SCIENCE INDUSTRY WHO DOES WHAT

A LOOK AT THE KEY ROLES IN DATA SCIENCE

DATA SCIENTIST

SO RARE AS UNICORNS

Languages
R, SAS, Python, Matlab, SQL, Hive, Pig, Spark

Skills & Talents

- ✓ Distributed computing
- ✓ Predictive modeling
- ✓ Story-telling and visualizing
- ✓ Math, Stats, Machine Learning

Role
Cleans, massages and organizes (big) data

Mindset
Curious data wizard

HIRED BY
Google, Microsoft, Amazon

DATA ENGINEER

SOFTWARE ENGINEERS BY TRADE

Languages
SQL, Hive, Pig, R, Matlab, SAS, SPSS, Python, Java, Ruby, C++, Perl

Skills & Talents

- ✓ Database systems (SQL & NO SQL based)
- ✓ Data modeling & ETL tools
- ✓ Data APIs
- ✓ Data warehousing solutions

Role
Develops, constructs, tests and maintains architectures (such as databases and large-scale processing systems)

Mindset
All-purpose everyman

HIRED BY
Spotify, Facebook, Amazon

STATISTICIAN

RHETORIC LEADERS OF DATA

Languages
R, SAS, SPSS, Matlab, Stata, Python, Perl, Hive, Pig, Spark, SQL

Skills & Talents

- ✓ Statistical theories & methodology
- ✓ Data mining & machine learning
- ✓ Distributed Computing (Hadoop)
- ✓ Database systems (SQL and NO SQL based)
- ✓ Cloud tools

Role
Collects, analyzes and interprets qualitative as well as quantitative data with statistical theories and methods

Mindset
Logical and enthusiastic stats genius

HIRED BY
LinkedIn, Johnson & Johnson, Pepsico

DATABASE ADMINISTRATOR

DATABASE CARETAKER

Languages
SQL, Java, Ruby on Rails, XML, C#, Python

Skills & Talents

- ✓ Backup & recovery
- ✓ Data modeling and design
- ✓ Distributed Computing (Hadoop)
- ✓ Database systems (SQL and NO SQL based)
- ✓ Data security
- ✓ ERP & business knowledge

Role
Ensures that the database is available to all relevant users, is performing properly and is being kept safe

Mindset
Master of Disaster Prevention

HIRED BY
Tableau, Reddit, Twitter

1. Data Science (3)

THE DATA SCIENCE INDUSTRY WHO DOES WHAT

A LOOK AT THE KEY ROLES IN DATA SCIENCE

DATA ANALYST

DATA DETECTIVE

Role: Collects, processes and performs statistical data analyses

Mindset: Intuitive data junkie with high "figure-it-out" quotient

Languages: R, Python, HTML, Javascript, C/C++, SQL

Skills & Talents:

- ✓ Spreadsheet tools (e.g. Excel)
- ✓ Database systems (SQL and NO SQL based)
- ✓ Communication & visualization
- ✓ Math, Stats, Machine Learning

HIRED BY: IBM, DHL

DATA ARCHITECT

THE CONTEMPORARY DATA MODELLER

Role: Creates blueprints for data management systems to integrate, centralize, protect and maintain data sources

Mindset: Inquiring ninja with a love for data architecture design patterns

Languages: SQL, XML, Hive, Pig, Spark

Skills & Talents:

- ✓ Data warehousing solutions
- ✓ In-depth knowledge of database architecture
- ✓ Extraction Transformation and Load (ETL), spreadsheet and BI tools
- ✓ Data modeling
- ✓ Systems development

HIRED BY: VISA, Coca-Cola, logitech

BUSINESS ANALYST

CHANGE AGENT

Role: Improves business processes as intermediary between business and IT

Mindset: Resilient project juggler

Languages: SQL

Skills & Talents:

- ✓ Basic tools (e.g. MS Office)
- ✓ Data visualization tools (e.g. Tableau)
- ✓ Conscious listening and storytelling
- ✓ Business Intelligence understanding
- ✓ Data modeling

HIRED BY: UBER, DELL, ORACLE

DATA AND ANALYTICS MANAGER

DATA SCIENCE TEAM LEADER

Role: Manages a team of analysts and data scientists

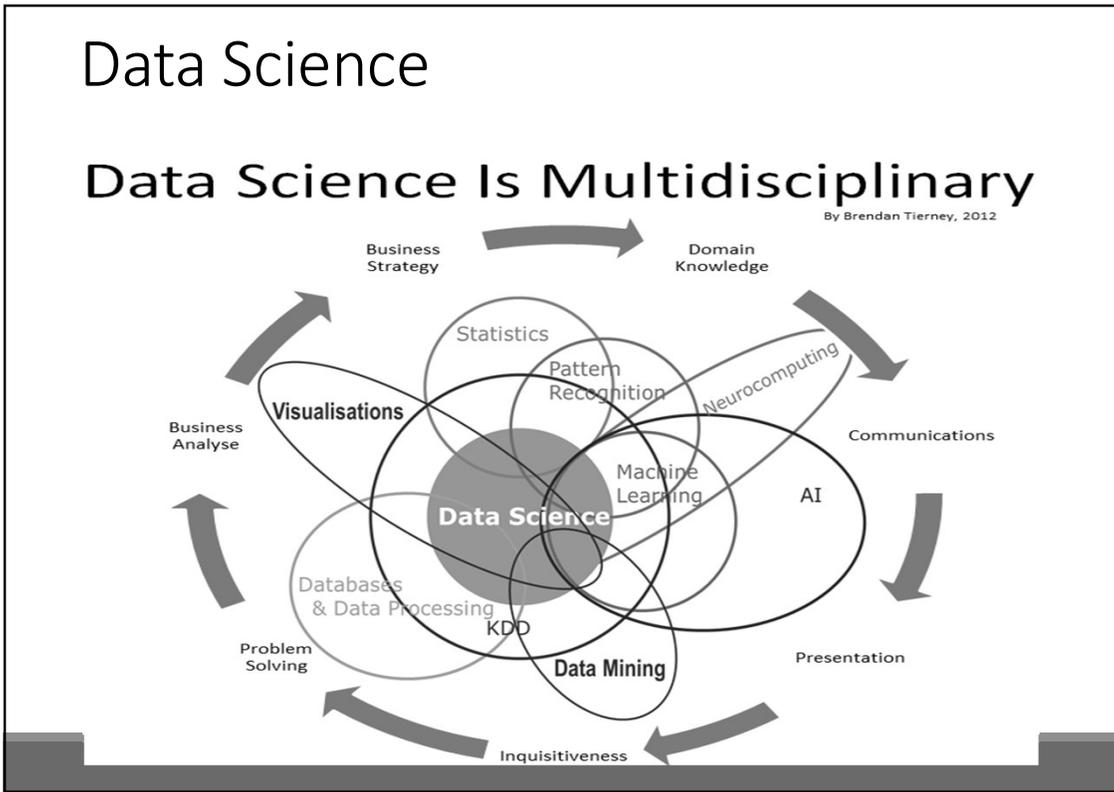
Mindset: Data Wizards' Cheerleader

Languages: SQL, R, SAS, Python, Matlab, Java

Skills & Talents:

- ✓ Database systems (SQL and NO SQL based)
- ✓ Leadership & project management
- ✓ Interpersonal communication
- ✓ Data mining & predictive modeling

HIRED BY: COURSERA, slack, MOTOROLA SOLUTIONS



Concentration in Data Science

Mathematics and Applied Mathematics

Applied Statistics/Data Analysis

Solid Programming Skills (R, Python, Julia, SQL)

Data Mining

Data Base Storage and Management

Machine Learning and discovery

2. Data Analytics (1)

Data Analytics (Data Analysis):

*qualitative and quantitative
techniques and processes*

*used to enhance productivity
and business gain*

*data is extracted and
categorized to identify and
analyze behavioral data and
patterns*

*techniques vary according to
organizational requirements*

4 types of Data Analytics

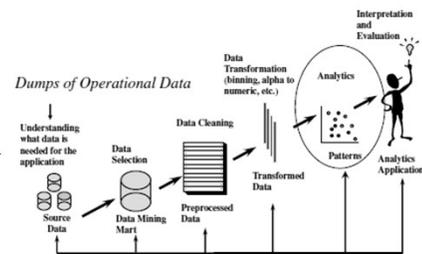
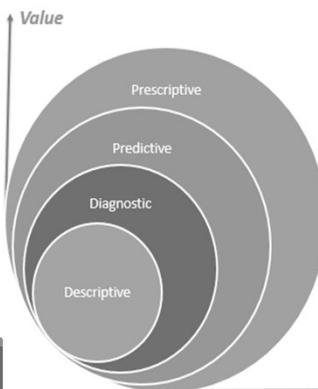


Figure 1.2 The Analytics Process Model

What is the data telling you?

Descriptive: *What's happening in my business?*

- Comprehensive, accurate and live data
- Effective visualisation

Diagnostic: *Why is it happening?*

- Ability to drill down to the root-cause
- Ability to isolate all confounding information

Predictive: *What's likely to happen?*

- Business strategies have remained fairly consistent over time
- Historical patterns being used to predict specific outcomes using algorithms
- Decisions are automated using algorithms and technology

Prescriptive: *What do I need to do?*

- Recommended actions and strategies based on champion / challenger testing strategy outcomes
- Applying advanced analytical techniques to make specific recommendations

2. Data Analytics (2)

Descriptive Analytics: What is happening?

In business, it provides the analyst with a view of key metrics and measures within the company.

An example of this could be a monthly profit and loss statement. Understanding demographic information on their customers would be categorized as "descriptive analytics".

Utilizing useful visualization tools enhances the message of descriptive analytics.



2. Data Analytics (3)

Diagnostic Analytics: Why is it happening?

On the assessment of the descriptive data, diagnostic analytical tools will empower an analyst to drill down and in so doing isolate the root-cause of a problem.

Well-designed business information (BI) dashboards incorporating reading of time-series data and featuring filters and drill down capability allow for such analysis.



	A	B	C	D	E	F	G	H	I	J
1										
2										
3	Store Sales Net					Store Type				
4	USA	Store Size	Store City	Product Family	Deluxe Super	Gourmet Supermar	Mid-Size Groce	Small Grocery	Supermarket	Grand Total
5		CA	Beverly Hills	Drink	\$2,992.83					\$2,992.83
6				Food	\$20,028.18					\$20,028.18
7				Non-Consumable	\$3,964.79					\$3,964.79
8				Beverly Hills Total	\$27,485.80					\$27,485.80
9			Los Angeles	Drink				\$2,493.33	\$2,493.33	\$2,493.33
10				Food				\$23,598.28	\$23,598.28	\$23,598.28
11				Non-Consumable				\$8,305.14	\$8,305.14	\$8,305.14
12			Los Angeles Total					\$32,773.74	\$32,773.74	\$32,773.74
13			San Diego	Drink				\$3,050.43	\$3,050.43	\$3,050.43
14				Food				\$23,627.83	\$23,627.83	\$23,627.83
15				Non-Consumable				\$6,109.94	\$6,109.94	\$6,109.94
16			San Diego Total					\$32,777.61	\$32,777.61	\$32,777.61
17			San Francisco	Drink				\$227.38	\$227.38	\$227.38
18				Food				\$1,960.53	\$1,960.53	\$1,960.53
19				Non-Consumable				\$474.35	\$474.35	\$474.35
20			San Francisco Total					\$2,662.26	\$2,662.26	\$2,662.26
21			CA Total					\$27,485.80	\$2,662.26	\$65,491.35
22			OR	Drink	\$4,438.49			\$2,982.45	\$7,200.94	\$7,200.94
23				Food	\$32,778.35			\$23,818.87	\$61,149.22	\$61,149.22
24				Non-Consumable	\$18,177.89			\$8,428.53	\$18,606.41	\$18,606.41
25			OR Total		\$55,294.72			\$33,109.84	\$85,556.57	\$85,556.57
26			WA	Drink	\$3,680.56		\$1,409.50	\$498.51	\$7,968.57	\$13,517.07
27				Food	\$32,497.76		\$10,262.19	\$4,148.19	\$87,915.89	\$114,994.85
28				Non-Consumable	\$3,796.36		\$2,813.73	\$1,900.54	\$17,116.28	\$29,999.89
29			WA Total		\$44,884.68		\$14,115.42	\$5,668.24	\$83,300.57	\$158,488.91
30			USA Total		\$87,979.48		\$14,115.42	\$8,128.51	\$191,901.71	\$279,810.96
31			Grand Total		\$97,479.40		\$27,483.80	\$14,115.42	\$8,330.51	\$191,901.71

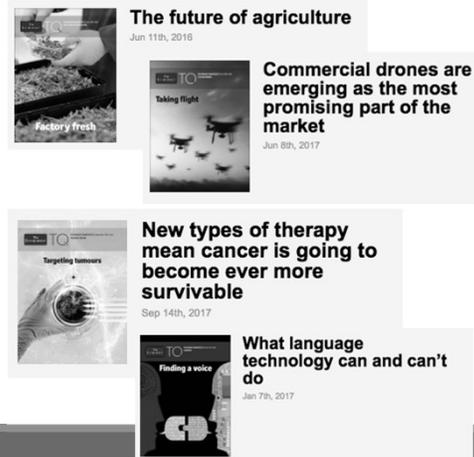
2. Data Analytics (4)

Predictive Analytics: What is likely to happen?

Predictive analytics is all about forecasting.

Predictive models typically utilize a variety of variable data to make the prediction. The variability of the component data will have a relationship with what it is likely to predict. These data are then compiled together into a score or prediction.

In a world of significant uncertainty, being able to predict allows one to make better decisions.



2. Data Analytics (5)

Prescriptive Analytics: What do I need to do?

The prescriptive model utilizes an understanding of what has happened, why it has happened and a variety of "what-might-happen" analysis to help the user determine the best course of action to take.

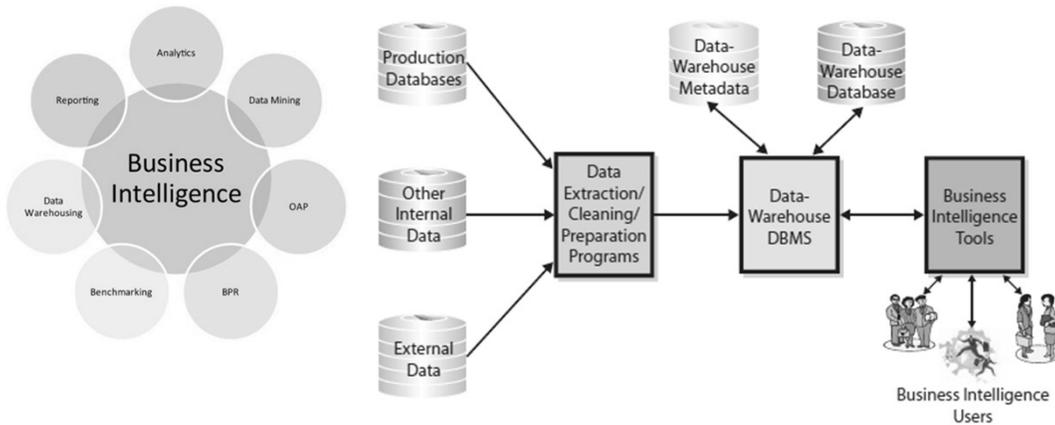
An example of this is a traffic application helping you choose the best route home and taking into account the distance of each route, the speed at which one can travel on each road and, the current traffic constraints.



3. Business intelligence & KMS (1)

Business intelligence

an umbrella term - applications, infrastructure and tools, and best practices that enable access to and analysis of information to improve and optimize decisions and performance.

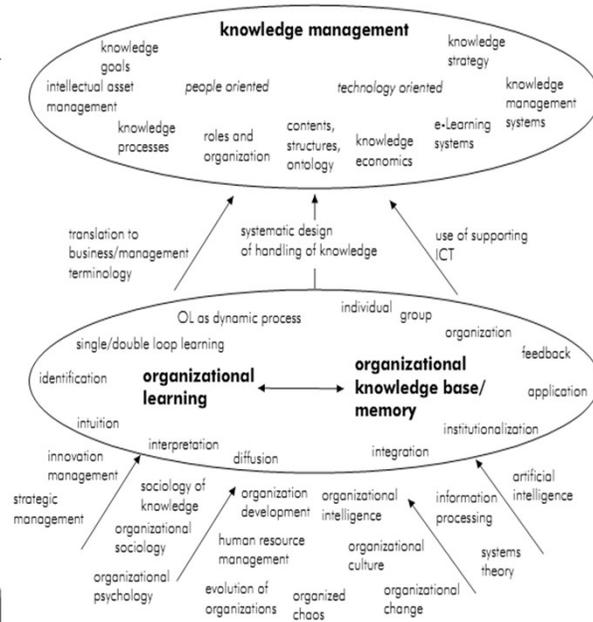


3. Business intelligence & KMS (2)

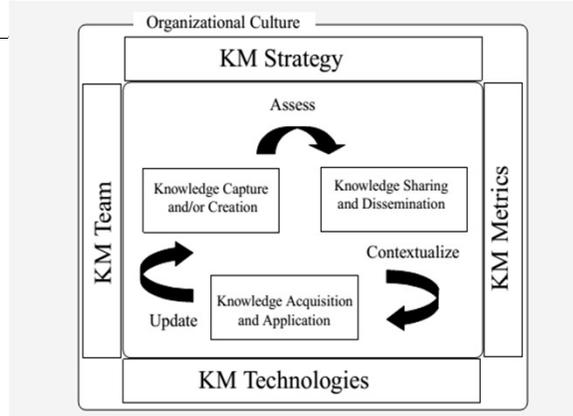
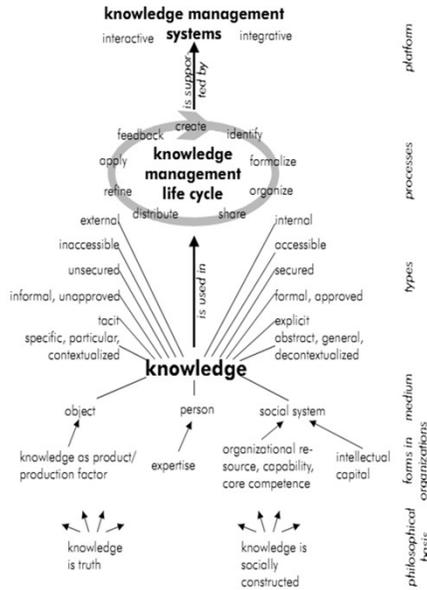
Knowledge management

a systematic approach for turning knowledge assets (tacit & explicit) into values.

KMS is a combination of technologies and management practices for identifying, creating, presenting, sharing, & applying knowledge in organization.



3. Business intelligence & KMS (3)



Knowledge management cycle

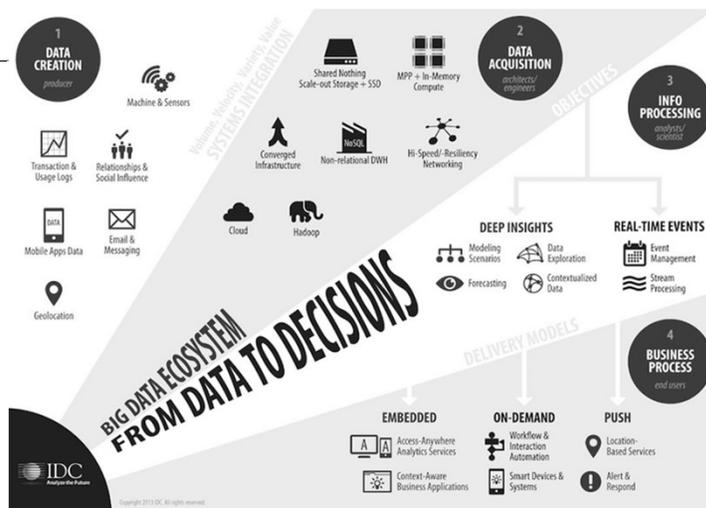
4. Big Data & Data mining (1)

Big Data:

high-volume, high-velocity, high-variety information assets

demand cost-effective, innovative forms of information processing

enable enhanced insight, decision making, and process automation



<https://www.i-scoop.eu/big-data-action-value-context/>

Data All Around

Lots of data is being collected and warehoused

- Web data, e-commerce
- Financial transactions, bank/credit transactions
- Online trading and purchasing
- Social Network



How Much Data Do We have?

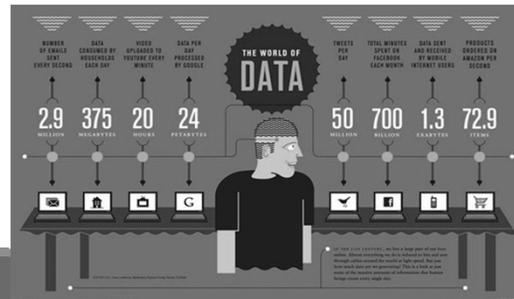
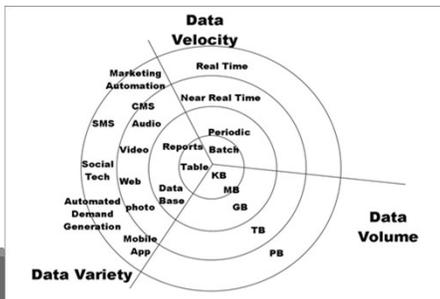
Google processes 20 PB a day (2008)

Facebook has 60 TB of daily logs

eBay has 6.5 PB of user data + 50 TB/day (5/2009)

1000 genomes project: 200 TB

- Cost of 1 TB of disk: \$35
- Time to read 1 TB disk: 3 hrs (100 MB/s)



Types of Data We Have

Relational Data (Tables/Transaction/Legacy Data)

Text Data (Web)

Semi-structured Data (XML)

Graph Data

Social Network, Semantic Web (RDF), ...

Streaming Data

You can afford to scan the data once

What To Do With These Data?

Aggregation and Statistics

- Data warehousing and OLAP

Indexing, Searching, and Querying

- Keyword based search
- Pattern matching (XML/RDF)

Knowledge discovery

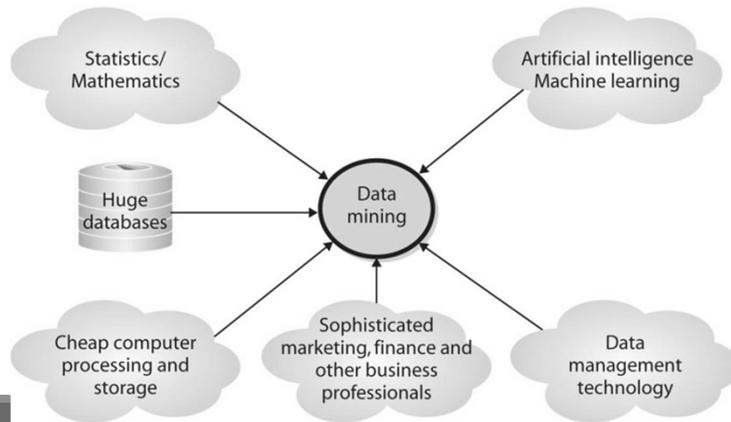
- Data Mining
- Statistical Modeling



4. Big Data & Data mining (2)

Data mining

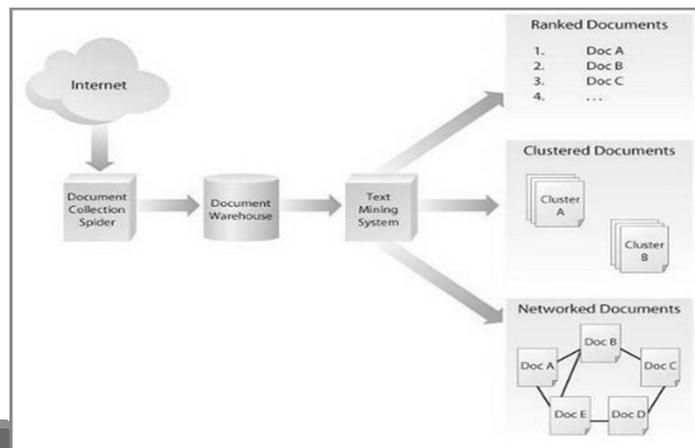
- use sophisticated statistical techniques, regression analysis and decision tree analysis
- used to discover hidden patterns and relationships
- market-basket analysis.



4. Big Data & Data mining (3)

Text mining

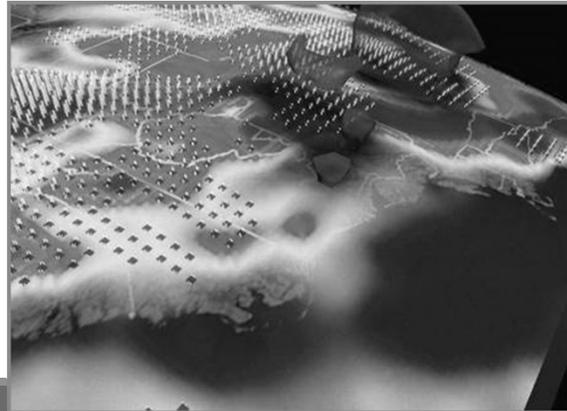
A specific form of data mining focusing on the extraction of information from textual documents. Ex: Web crawlers used to extract information from Internet



5. Data Visualization

Data Visualization

Tools or techniques to display of complex data relationships using graphical methods

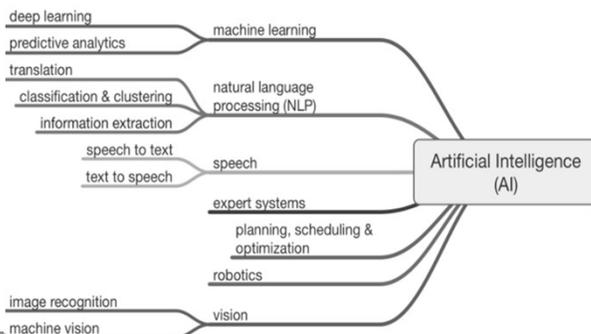


Visualization of a weather system

6. Artificial Intelligence & Machine Learning(1)

Artificial Intelligence

We seek to understand intelligence as manifest in living systems, build artificial systems capable of intelligent reasoning, perception, and behavior, and build principled models of reasoning and thinking applicable to a wide variety of real-world problems.



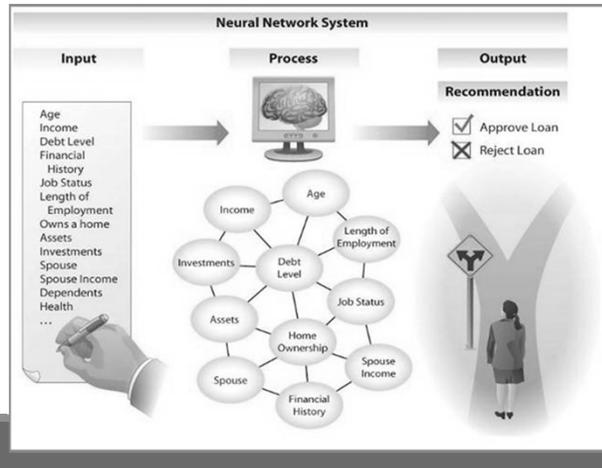
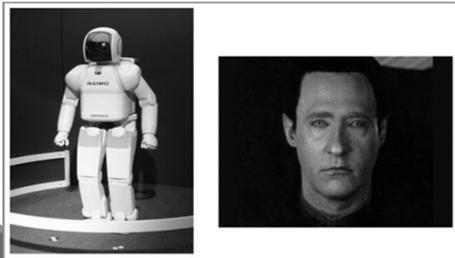
6. Artificial Intelligence & Machine Learning(2)

- **Intelligent system**

- Sensors, software and computers
- Emulate and enhance human capabilities

- **Three types**

- Expert systems
- Neural networks
- Intelligent agents



6. Artificial Intelligence & Machine Learning(3)

Machine Learning

Machine Learning is a subset of artificial intelligence. It could be defined as:

- *method of data analysis that automates analytical model building*
- *algorithms that iteratively learn from data, to find hidden insights*
- *science of getting computers to act without being explicitly programmed*

There are various types of machine learning, the three most common types being: supervised, unsupervised and reinforcement learning.

➤ In **supervised learning**, training of the algorithms needs to take place. Some applications: speech recognition, medical diagnosis, fraudulent detection, etc.

➤ In **unsupervised learning**, to detect what is being provided as input and explore the data to find structure and patterns. Unsupervised learning is used in content or product recommendations, like Netflix or Amazon.

➤ With **reinforcement learning**, a computer needs a continual input of data to improve continually through trial and error approach. Mostly used in developing driverless cars, dynamic pricing strategy or supply chain risk management.



7. Industry 4.0 & future technologies (1)

Industry 4.0

IoT Analytics – Quantifying the connected world

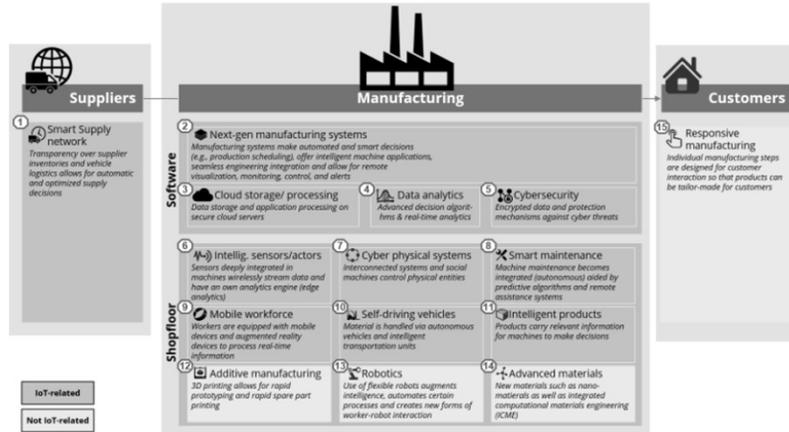
15 components of the smart factory of the future

First – steam, machines

Second - electricity, assembly line, mass production

Third - computers, beginnings of automation, robots and machines replace human workers on assembly lines.

Fourth – “**smart factory**” in which cyber-physical systems monitor the physical processes of the factory and make decentralized decisions.



7. Industry 4.0 & future technologies (2)

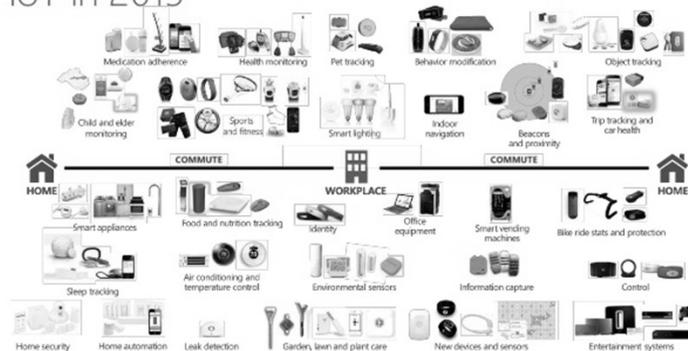
IoT

In 1999 Kevin Ashton coined the term 'Internet of Things'

flow between:

- BAN (body area network): wearables,
- LAN (local area network): smart home,
- WAN (wide area network): connected car,
- and
- VWAN (very wide area network): the smart city

IoT in 2015



7. Industry 4.0 & future technologies (3)

Cloud computing

storing and accessing data and programs over the Internet instead of your computer's hard drive

cloud is also a metaphor for the Internet

Top services or applications moving to the cloud:*

Small business	Medium business	Large business	Federal govt.
<ol style="list-style-type: none"> 1. Storage (40%) 2. Conferencing & collaboration (37%) 3. Messaging (36%) 	<ol style="list-style-type: none"> 1. Storage (35%) 2. Messaging (33%) 3. Office & productivity suites (32%) 	<ol style="list-style-type: none"> 1. Conferencing & collaboration (40%) 2. Storage/business process apps (35%) 3. Messaging/compute power (34%) 	<ol style="list-style-type: none"> 1. Conferencing & collaboration (39%) 2. Messaging (37%) 3. Business process apps (31%)
State/local govt.	Healthcare	Higher education	K-12
<ol style="list-style-type: none"> 1. Storage (19%) 2. Conferencing & collaboration (17%) 3. Messaging/business process apps/compute power (15%) 	<ol style="list-style-type: none"> 1. Conferencing & collaboration (29%) 2. Compute power (26%) 3. Office & productivity suites (22%) 	<ol style="list-style-type: none"> 1. Storage (31%) 2. Messaging/conferencing & collaboration (29%) 3. Compute power (25%) 	<ol style="list-style-type: none"> 1. Storage (40%) 2. Conferencing & collaboration (36%) 3. Office & productivity suites (33%)

*Those who are migrating or have migrated

7. Industry 4.0 & future technologies (4)

Virtual Reality

a 3-dimensional, computer generated environment which can be explored and interacted with by a person

one becomes part of the virtual world or is immersed within this environment and whilst there, is able to manipulate objects or perform a series of actions.



The Most Active AR/VR Investors 2011 - 2015

Investor	Investments
BURNBERG VENTURES	[Logos of companies invested in]
bCOSTVC	[Logos of companies invested in]
G/	[Logos of companies invested in]
QUALCOMM VENTURES	[Logos of companies invested in]
ANDRIENSON HOROWITZ	[Logos of companies invested in]
PARTECH	[Logos of companies invested in]
CBINSIGHTS	[Logos of companies invested in]



Explore Second Life

7. Industry 4.0 & future technologies (5)

Augmented Reality

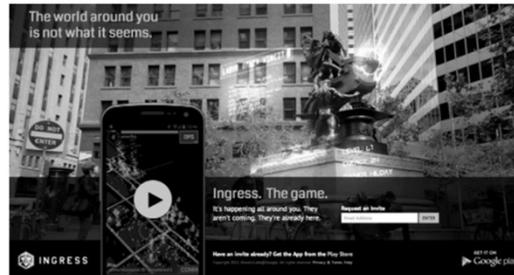
technology that layers computer-generated enhancements atop an existing reality

developed into apps and used on mobile devices to blend digital components into the real world



1. Ingress

This App is good news to every gamer in the world. It proves that there is more to the world than what you usually see; if anything, this App is the definition of seeing the world through different eyes.



The Ingress Augmented Reality App from Google turns your real life surrounding into capturable objectives in-game portals. Landmarks and points of interest are captured on your phone after it uses GPS technology access your location and gives you a virtual alternative to reality.

7. Industry 4.0 & future technologies (6)

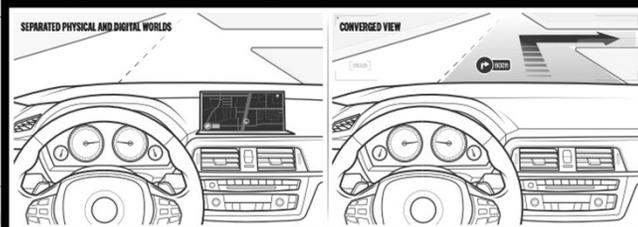
WHY EVERY ORGANIZATION NEEDS AN AUGMENTED REALITY STRATEGY

BY MICHAEL E. PORTER AND JAMES E. HEPPELMANN

There is a fundamental disconnect between the wealth of digital data available to us and the physical world in which we apply it. While reality is three-dimensional, the rich data we now have to inform our decisions and actions remains trapped on two-dimensional pages and screens. This gulf between the real and digital worlds limits our ability to take advantage of the torrent of information and insights produced by billions of smart, connected products (SCPs) worldwide.

CONVERGING PHYSICAL AND DIGITAL

Augmented reality reduces the mental effort needed to connect digital information about the physical world with the context it applies to.



Mentally transposing GPS images onto the road ahead is demanding and prone to errors.

AR superimposes digital data directly on the real world.



VISUALIZE
An AR glasses demo developed by Microsoft and Intel provides an X-ray view of a car's engine and undercarriage.

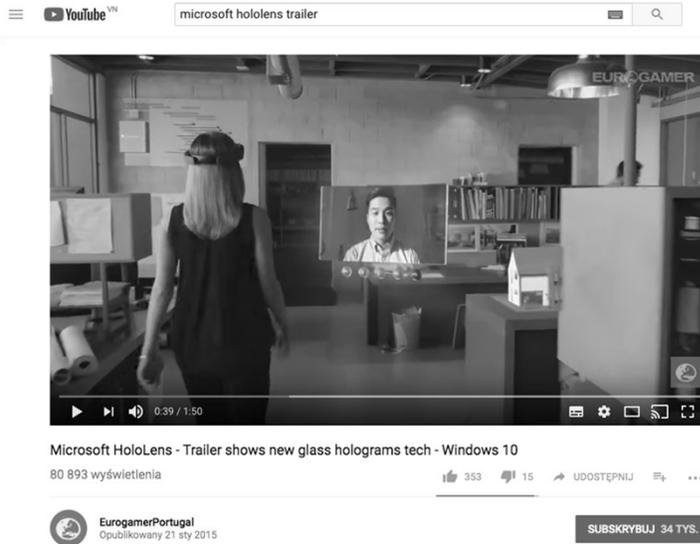
7. Industry 4.0 & future technologies (7)

Mixed Reality

takes AR to the next level

not just a layer on top of the world we see every day

ability to mix digitally rendered objects into our real environment



<https://www.youtube.com/watch?v=zEtLqtunD6g>

7. Industry 4.0 & future technologies (8)

Crypto currency

digital asset designed to work as a medium of exchange using cryptography to secure the transactions and to control the creation of additional units of the currency

bitcoins and other cryptocurrencies work on the blockchain technology

Bitcoin's price bubble will burst under government pressure
Kenneth Rogoff

The cryptocurrency is up 1,600% in two years - but state efforts to remove its near-anonymity will undermine its popularity



Bitcoin's price is up 600% over the past 12 months Photograph: Alexander Demianchuk/TASS

Is the cryptocurrency bitcoin the biggest bubble in the world today, or a great investment bet on the cutting edge of new-age financial technology? My best guess is that in the long run, the technology will thrive, but that the price of bitcoin will collapse.

If you haven't been following the bitcoin story, its price is up 600% over the past 12 months, and 1,600% in the past 24 months. At over \$4,200 (as of 5 October), a single unit of the virtual currency is now worth more than three times an ounce of gold. Some bitcoin evangelists see it going far higher in the next few years.

Warnings grow louder over cryptocurrency as valuations soar

With bitcoin and Ethereum gathering momentum among investors, some experts fear a bubble could soon burst



It's a game, and it looks very much like a bubble, says one expert. Photograph: Artyom Korostayev/TASS

Joe Kennedy, patriarch of the Kennedy clan, said he knew it was time to exit the stock market after a shoeshine boy gave him stock tips. If everyone thinks it's time to buy, it's time to sell, reasoned Kennedy. Then came the great crash of 1929 to prove him right. Perhaps some of that thinking could be applied today to the digital currency bonanza.

In recent months, warning voices have grown louder as the digital assets known as cryptocurrencies have attained record valuations. The price of bitcoin, the most famous cryptocurrency, has soared this year, from \$969 to more than \$5,000 in September; rival Ethereum began the year at \$8 and has traded as high as \$400 - while new coins or tokens are issued weekly, often attached to tech startups as a way to raise venture capital.

7. Industry 4.0 & future technologies (9)

Blockchain

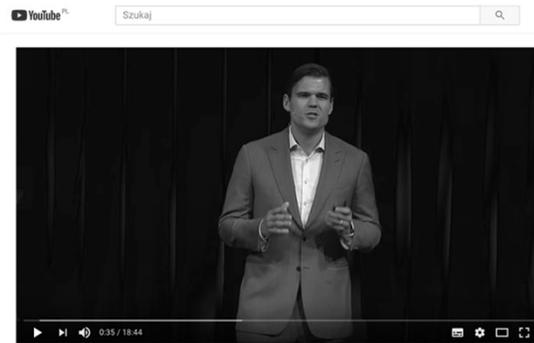
blockchain technology, which is a digital, distributed transaction ledger with identical copies maintained on each of the network's members' computers

all parties can review previous entries and record new ones

transactions are grouped in blocks, recorded one after the other in a chain of blocks (the 'blockchain')

the links between blocks and their content are protected by cryptography, so previous transactions cannot be destroyed or forged

the ledger and the transaction network are trusted without a central authority – a 'middleman'.



Blockchain is Eating Wall Street | Alex Tapscott | TEDxSanFrancisco

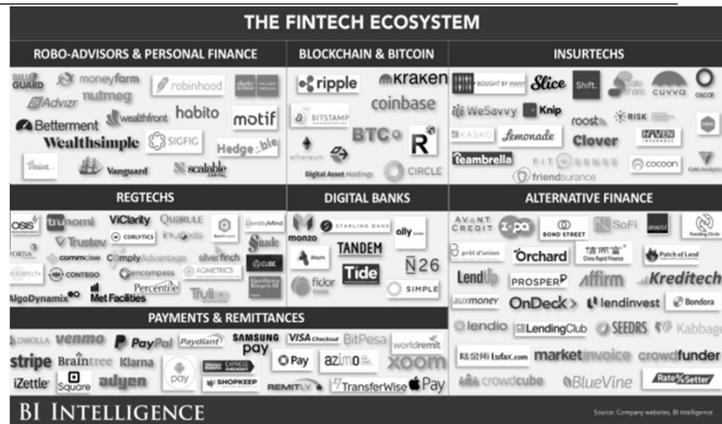
<https://www.youtube.com/watch?v=WnEYakUxsHU>

7. Industry 4.0 & future technologies (10)

Fintech / Insuretech

emerging financial services sector in the 21st century

technological innovation in the financial sector, including innovations i.e. in financial literacy, retail banking, insurance, investment and crypto-currencies



Thank you for listening!

Discussion

Q & A

