961701 Section 001
Everything Starts with Data

Dr. Pree Thiengburanathum

Announcements

- We have 22 students enrolled (Plan A)
- 1 International student?

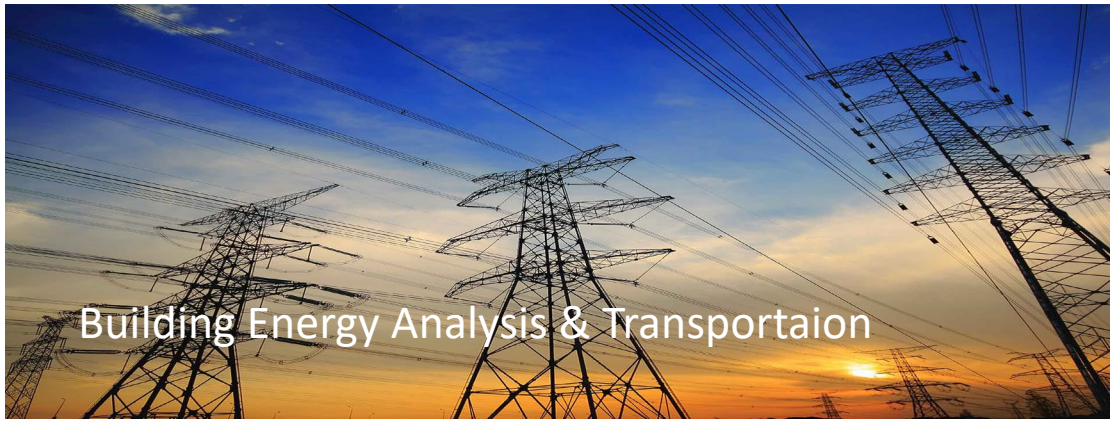
Agenda

- Getting to a bit know of me and you.
- Course outline
- Pre-test
- History Industrials to data-driven
- Roles in Data Science world
- Workshop 1

Education

- **Ph.D. in (Computing and Informatic) Bournemouth University,** *2013 – Mar 2018, Faculty of Science and Technology, Department of Computing and Informatics, Bournemouth University, UK.*
- **ERASMUS MUNDUS Research fellow,** *Feb 1, 2010- Dec 1, 2010, Universite De Lyon 2, France*
- **Master of Science in Computer Science**, Fall 2008, *University of Colorado at Denver, Denver, Colorado, USA.*
- **Bachelor of Science in Computer Science**, 2005, *Colorado State University*, *Fort Collins, Colorado, USA.*

Area of interest



Getting to know you

- Your demographic and such
- Experience in Data science
- Go to **www.menti.com** and use the code **23 44 26 4**

Course outline

Course Title: CTDS 701 (961701) Everything Starts with Data Location: Sec 801:

Location 2nd floor meeting room RTT building, Sundays 08:30 – 11:30am

- **Course Description**

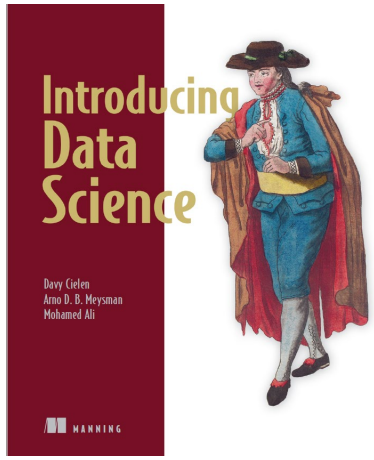
Theory and discussion of data analytics, foundation of data scientists, data analytic process, data analytic tools, data set,data quality, data analytic project management, case study

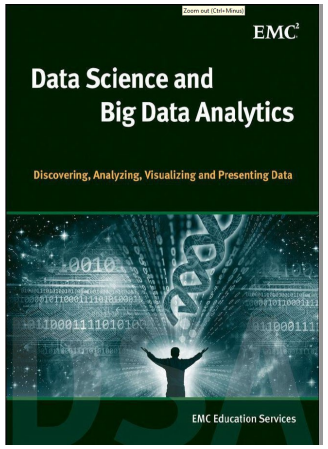
Course objectives

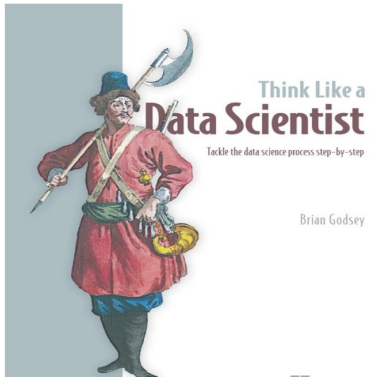
- 1. Student will be able to apply data science methodologies with problem.
- 2. Students will be able to explain the principle of data-driven mythologies.

Texts

- Cielen, D., Meysman, A., & Ali, M. (2016). *Introducing data science: big data, machine learning, and more, using Python tools*. Manning Publications Co.
- EMC Education Services. (2015). *Data Science and Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data.* Wiley.
- Godsey, B. (2017). *Think like a data scientist*. Manning Publications Co.







Course work

Lectures (3 hours per week)

Term project

Paper-based midterm and final examinations

Grading System

The semester grade is calculated as follows:

Attendance, quiz and other class activities • Score will be checked via Google classroom	***15%***
Term Project(s)	***15%***
Midterm Examination	***35%***
Final Examination	***35%***
***T*otal**	**100%**

Communication

- **MS team:** (announcement)
- **Google Classrom:** Please use this code to join “ansdfnz”
- **Study room:** https://discord.gg/QG8NPEAW

- **Email:** pree.t@cmu.ac.th

Course contents

<u>Note</u>: Some topics of the contents might be subject to change or add. The notice will be announced in the course website.

i^{th} week	Topics
1	Roles in the Data Science World
2	Basic data analytics – Knowledge Discovery and Data Mining (KDD)
3	Basic data analytics – From Data to Data Product
4	What is Data?
5	Dataset
6	Dataset and Data Quality
7	Data Science Process – The Historical view from KDD to Modern Processes
8	Data Science Process in Action
9	Midterm Examination
10	Introduction to Data Science Project Management
11	Introduction to Machine Learning – Part I
12	Introduction to Machine Learning – Part II
13	Basic Model Deployment
14	Case Study - I
15	Case Study - II
16	Closing
17	Final Examination

Pre-Test time

(test your background, basic stat, coding, database, language, etc.)

- A bit of history and revolution ☺

First Industrial Revolution

- 1760-1840s began in Great Britain
- Water wheel/steam engine
- From hand production methods to machines
- Products: textiles, coal, and iron
- Replace the traditional agriculture with industry

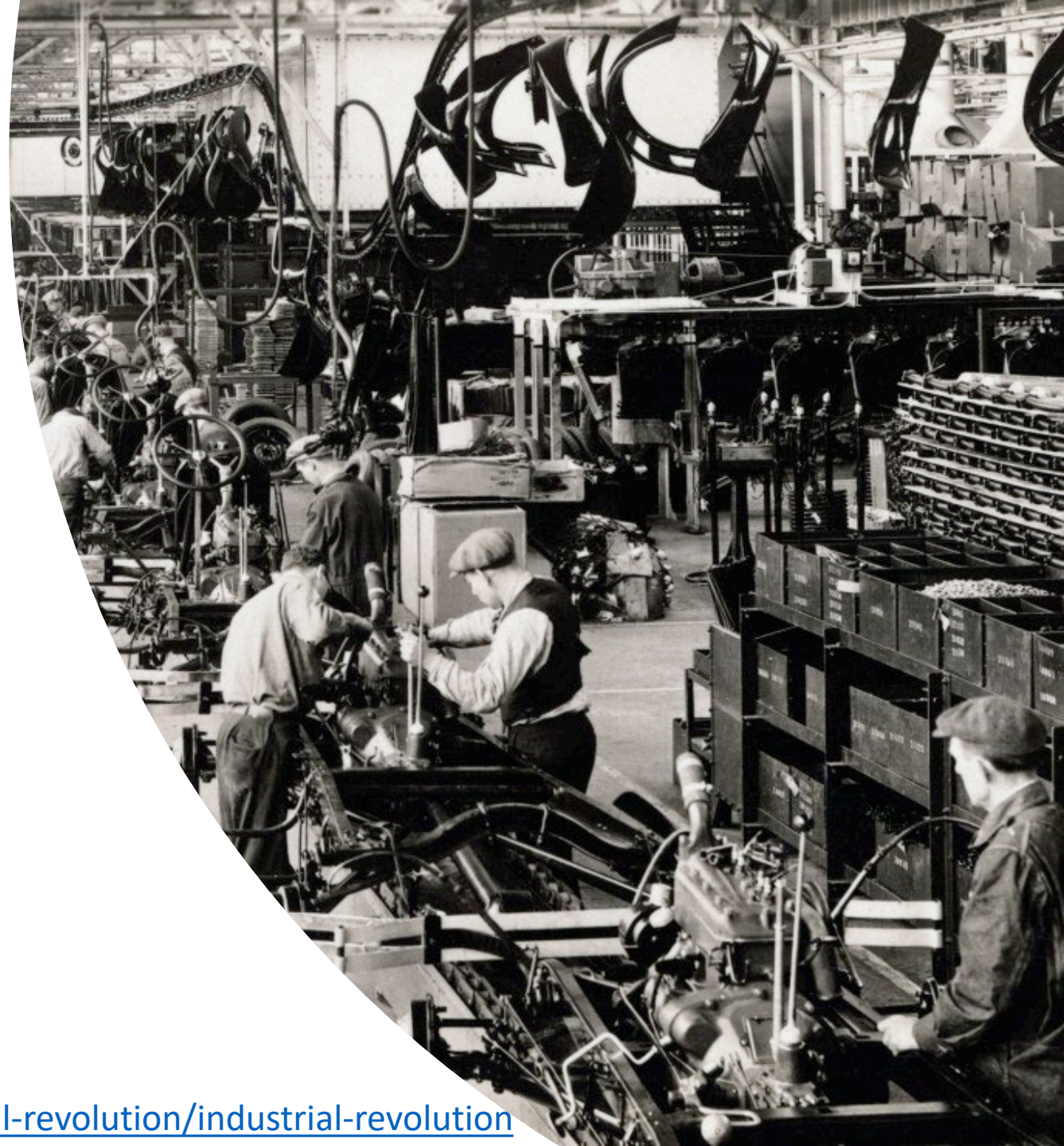
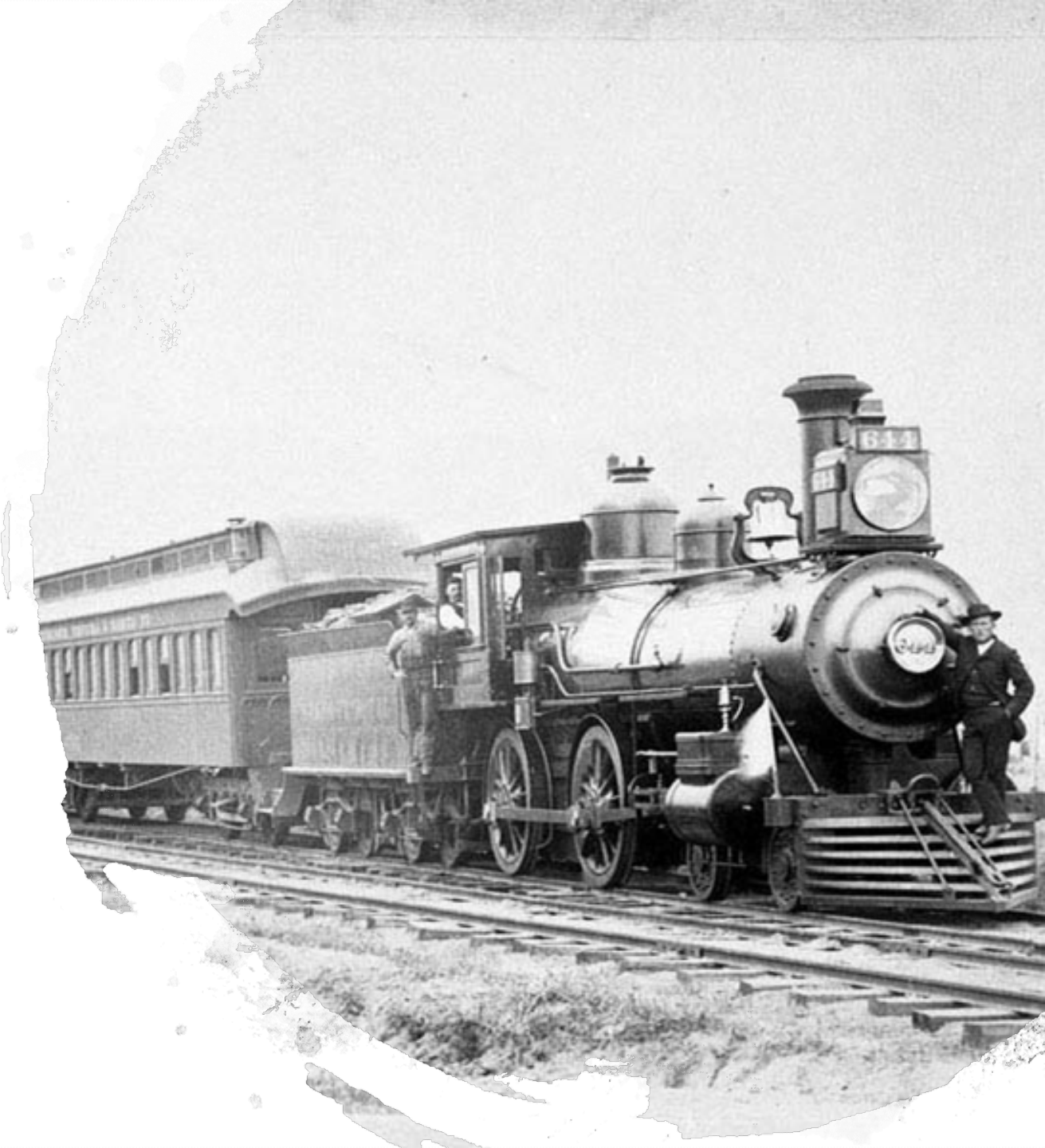


Image source:https://www.history.com/topics/industrial-revolution/industrial-revolution

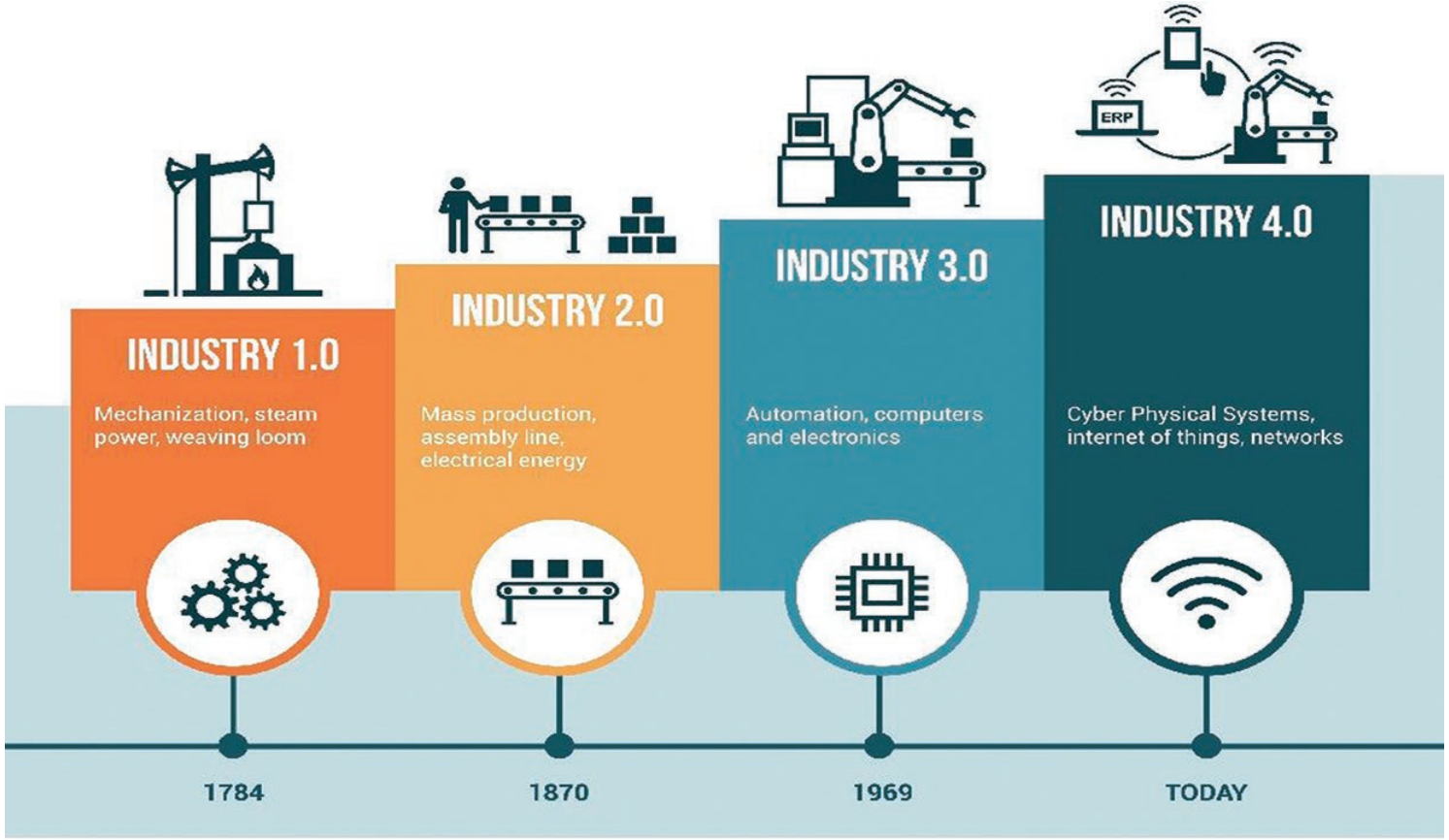
Second industry revolution

- 1870-1914 century
- A.K.A Technology revolution in electricity, chemical, petroleum
- Product of : Telegraph, railroad networks, gas, water supply and sewage systems
- Big mover is the electric power and telegraph.



Third industry revolution

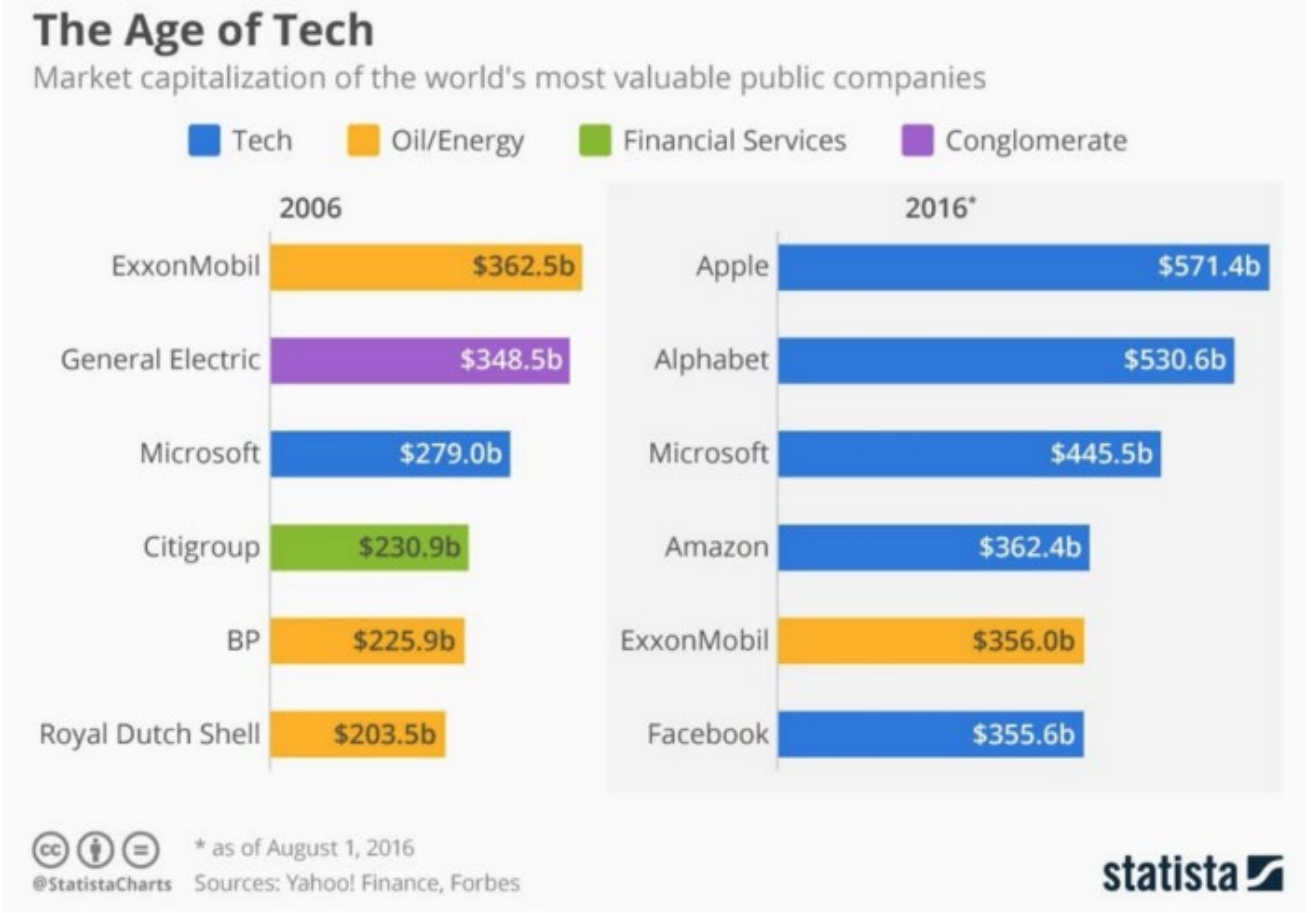
- 1995- now
- Rise of the Internet
- Computer and automation ruled the industrial scene

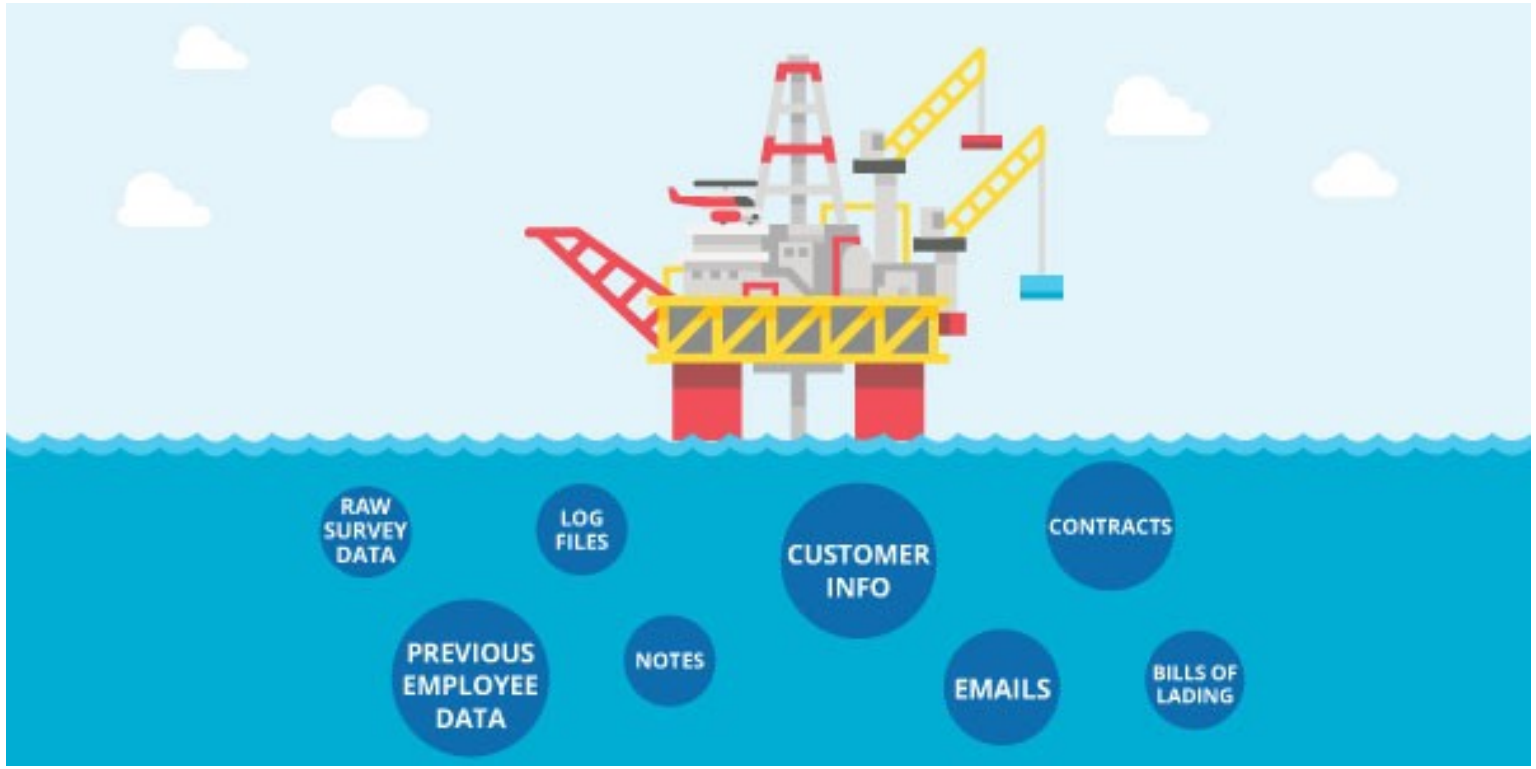


Fourth industry revolution

- Wireless enhanced
- IoT, robotic, VR, AI
- Smart devices, Smart cities
- Heavily related to the data
- Predictive model
- Fusion between physical, digital, biological
- Automated complex task.

Data is a new oil?





Big data and Data
Crude oil and oil refinery
-Essential raw resource -> econ
-knowledge economy

- Oil vs Info extraction
- Data flow <-> oil flow

What is data?

Unit	Value	Size
bit (b)	0 or 1	1/8 of a byte
byte (B)	8 bits	1 byte
kilobyte (KB)	1000^1 bytes	1,000 bytes
megabyte (MB)	1000^2 bytes	1,000,000 bytes
gigabyte (GB)	1000^3 bytes	1,000,000.000 bytes
terabyte (TB)	1000^4 bytes	1,000,000,000,000 bytes
petabyte (PB)	1000^5 bytes	1,000,000,000,000,000 bytes
exabyte (EB)	1000^6 bytes	1,000,000,000,000,000,000 bytes
zettabyte (ZB)	1000^7 bytes	1,000,000,000,000,000,000,000 bytes
yottabyte (YB)	1000^8 bytes	1,000,000,000,000,000,000,000,000 bytes

Structured vs Unstructured data

	Indicator ID	Dimension List	Timeframe	Numeric Value	Missing Value Flag	Confidence Inte
2	214390830	Total (Age-adjusted)	2008	74.6%		73.8%
3	214390833	Aged 18-44 years	2008	59.4%		58.0%
4	214390831	Aged 18-24 years	2008	37.4%		34.6%
5	214390832	Aged 25-44 years	2008	66.9%		65.5%
6	214390836	Aged 45-64 years	2008	88.6%		87.7%
7	214390834	Aged 45-54 years	2008	86.3%		85.1%
8	214390835	Aged 55-64 years	2008	91.5%		90.4%
9	214390840	Aged 65 years and over	2008	94.6%		93.8%
10	214390837	Aged 65-74 years	2008	93.6%		92.4%
11	214390838	Aged 75-84 years	2008	95.6%		94.4%
12	214390839	Aged 85 years and over	2008	96.0%		94.0%
13	214390841	Male (Age-adjusted)	2008	72.2%		71.1%
14	214390842	Female (Age-adjusted)	2008	76.8%		75.9%
15	214390843	White only (Age-adjusted)	2008	73.8%		72.9%
16	214390844	Black or African American only (Age-adjusted)	2008	77.0%		75.0%
17	214390845	American Indian or Alaska Native only (Age-adjusted)	2008	66.5%		57.1%
18	214390846	Asian only (Age-adjusted)	2008	80.5%		77.7%
19	214390847	Native Hawaiian or Other Pacific Islander only (Age-adjusted)	2008	DSU		
20	214390848	2 or more races (Age-adjusted)	2008	75.6%		69.6%

Figure 1.1 An Excel table is an example of structured data.

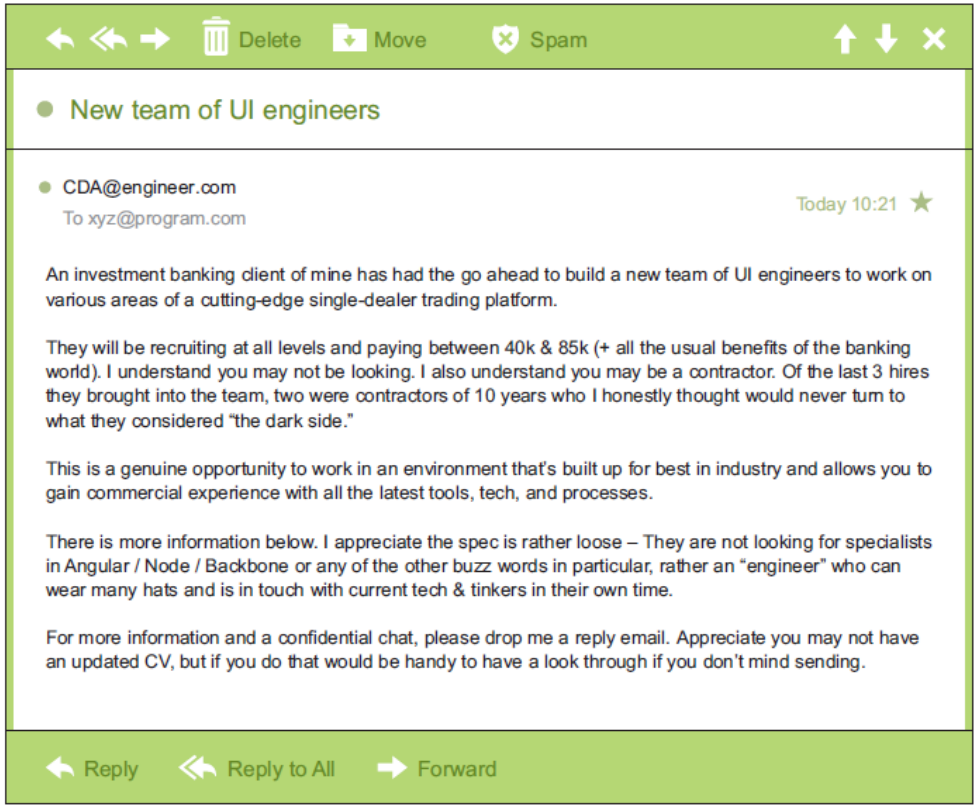


Figure 1.2 Email is simultaneously an example of unstructured data and natural language data.

Images source: Course text book 1

Hierarchy of data (cont.)

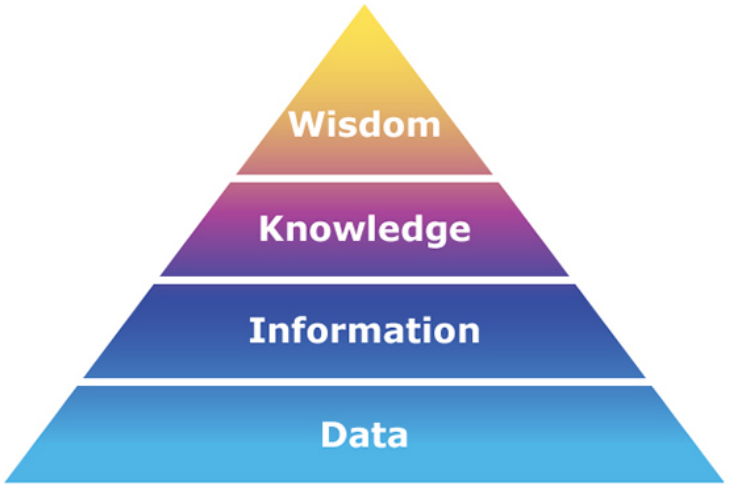
- **Bit** – a bit is the smallest unit of data representation (value of a bit may be 0 or 1)
- **Byte** – a unit of digital information that most commonly consists of eight bits to encode a single character.
- **Field** – a filed consists of a grouping of characters. A data field represents an attribute (a characteristic or quality) of some entity (object, person, place, or event)

Hierarchy of data (cont.)

- **Record** – a record represents a collection of attributes that describe an entity
- **File** – a group of related records.
- **Database** – is an integrated collection of logically related records or files.

Data vs Information vs Knowledge

- **Data:** can be any facts, numbers, text that can be processed by a computer.
- **Information** is a data that have been processed in such way as to increate the knowledge of the person who uses the data. (i.e, information is data in context)
- **Knowledge:** The concept of understand information based of finding pattern, to gain insight to information.



The DIKW-pyramid: Data - Information - Knowledge - Wisdom: Data can be structured to information. Data and information can be disseminated. Knowledge and understanding must be built in each person or within a group. Wisdom - or the classical Greek expression: *Phronesis:* Sufficient and adequate Knowledge and insight to apply it in problem solving and making the right decisions.

History of Database technology

1950's and early 1960's:

- Data processing using magnetic tapes for storage
  - Tapes provided only sequential access
- Punched cards for input

Late 1960's and 1970's:

- Hard disks allowed direct access to data
- Network and hierarchical data models in widespread use
- Ted Codd defines the relational data model
  - Would win the ACM Turing Award for this work
  - IBM Research begins System R prototype
  - UC Berkeley begins Ingres prototype
- High-performance (for the era) transaction processing

History of Database technology

1980's:

- Research relational prototypes evolve into commercial systems
- SQL becomes industrial standard
- Parallel and distributed database systems
- Object-oriented database systems

1990's:

- Large decision support and data-mining applications
- Large multi-terabyte data warehouses
- Emergence of Web commerce

Early 2000's:

- XML and XQuery standards
- Automated database administration

Later 2000's: **NoSQL and NewSQL (RDMS enhanced)**

- Giant data storage systems
- Google BigTable, Yahoo PNuts, Amazon,etc.

Driver of big data

- Moore’s law 1965
- “The density of transistors in integrated circuits would continue double every 18 months.”

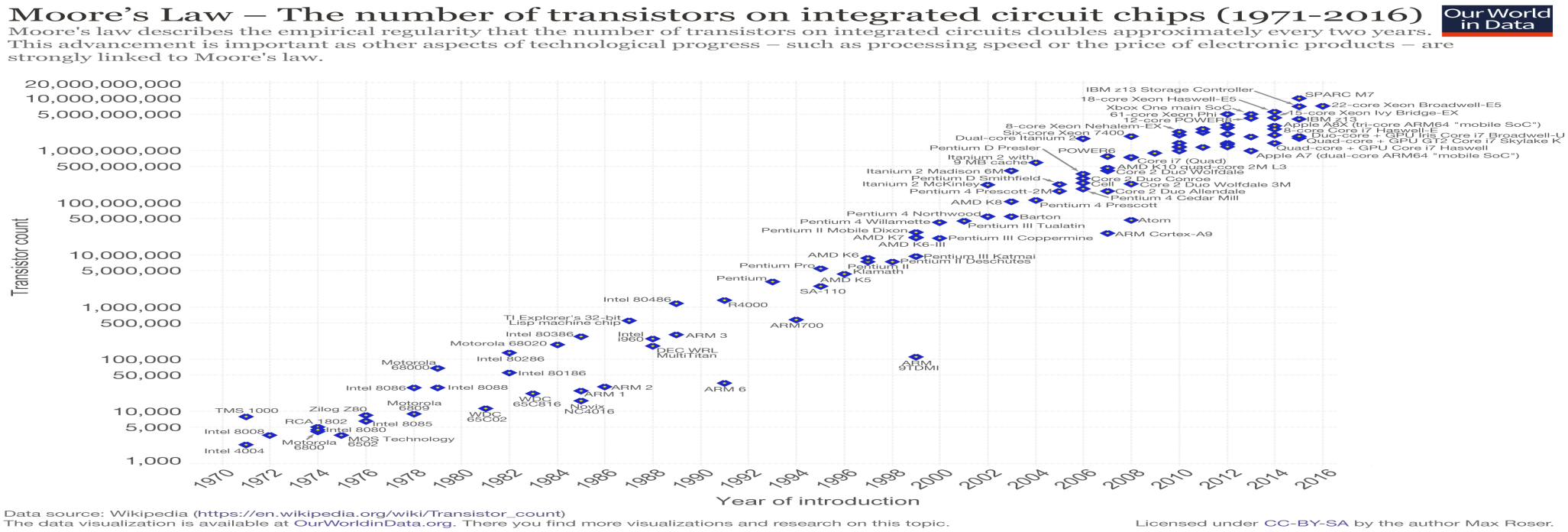
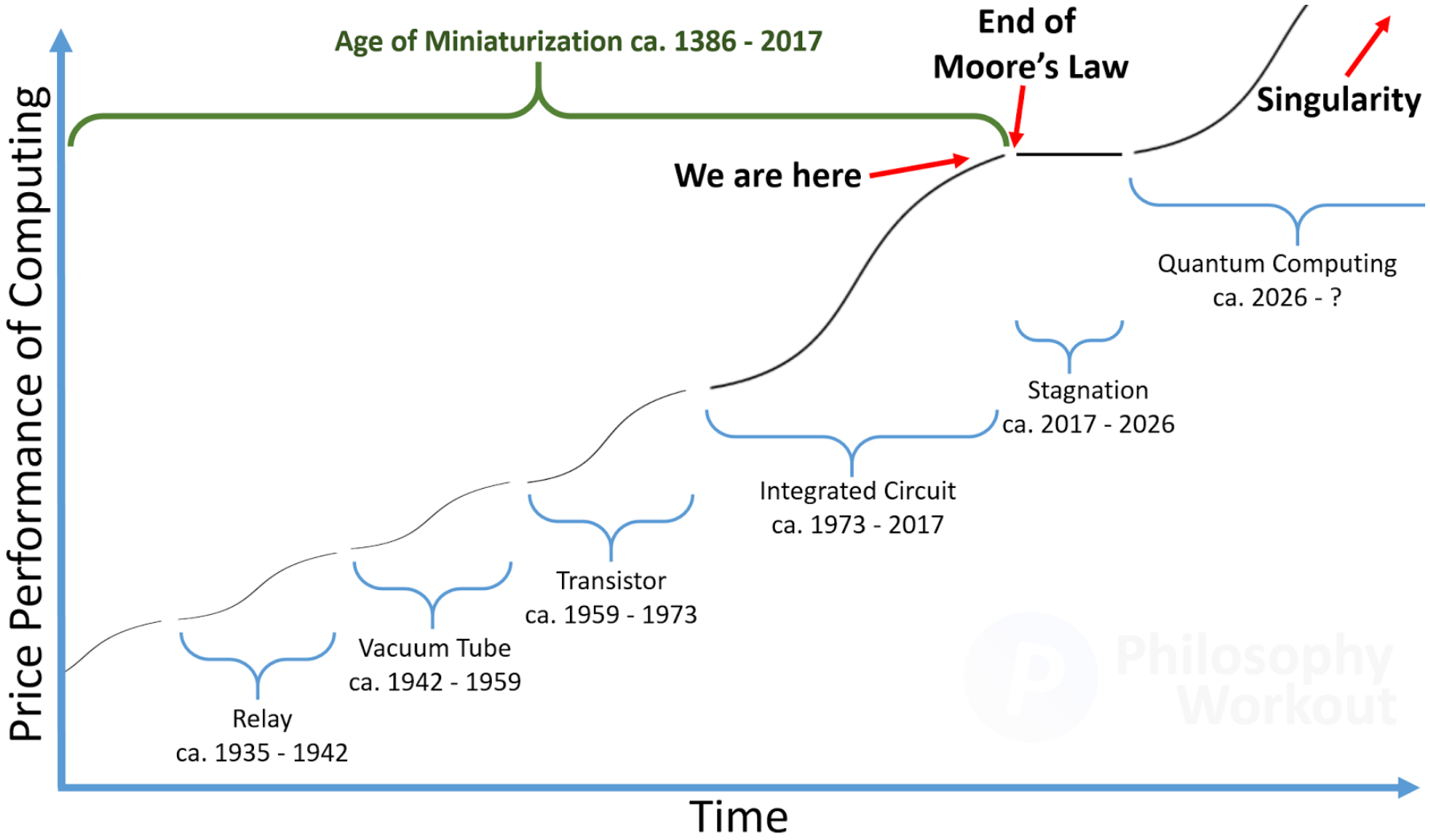


Image source: https://www.futuristgerd.com/2017/02/vanishing-point-the-rise-of-the-invisible-computer-and-moores-law/

The reality


Age of Miniaturization ca. 1386 - 2017
End of
Moore's Law
Singularity
We are here
Quantum Computing
ca. 2026 - ?
Stagnation
ca. 2017 - 2026
Integrated Circuit
ca. 1973 - 2017
Transistor
ca. 1959 - 1973
Vacuum Tube
ca. 1942 - 1959
Relay
ca. 1935 - 1942
Price Performance of Computing
Time
Philosophy
Workout

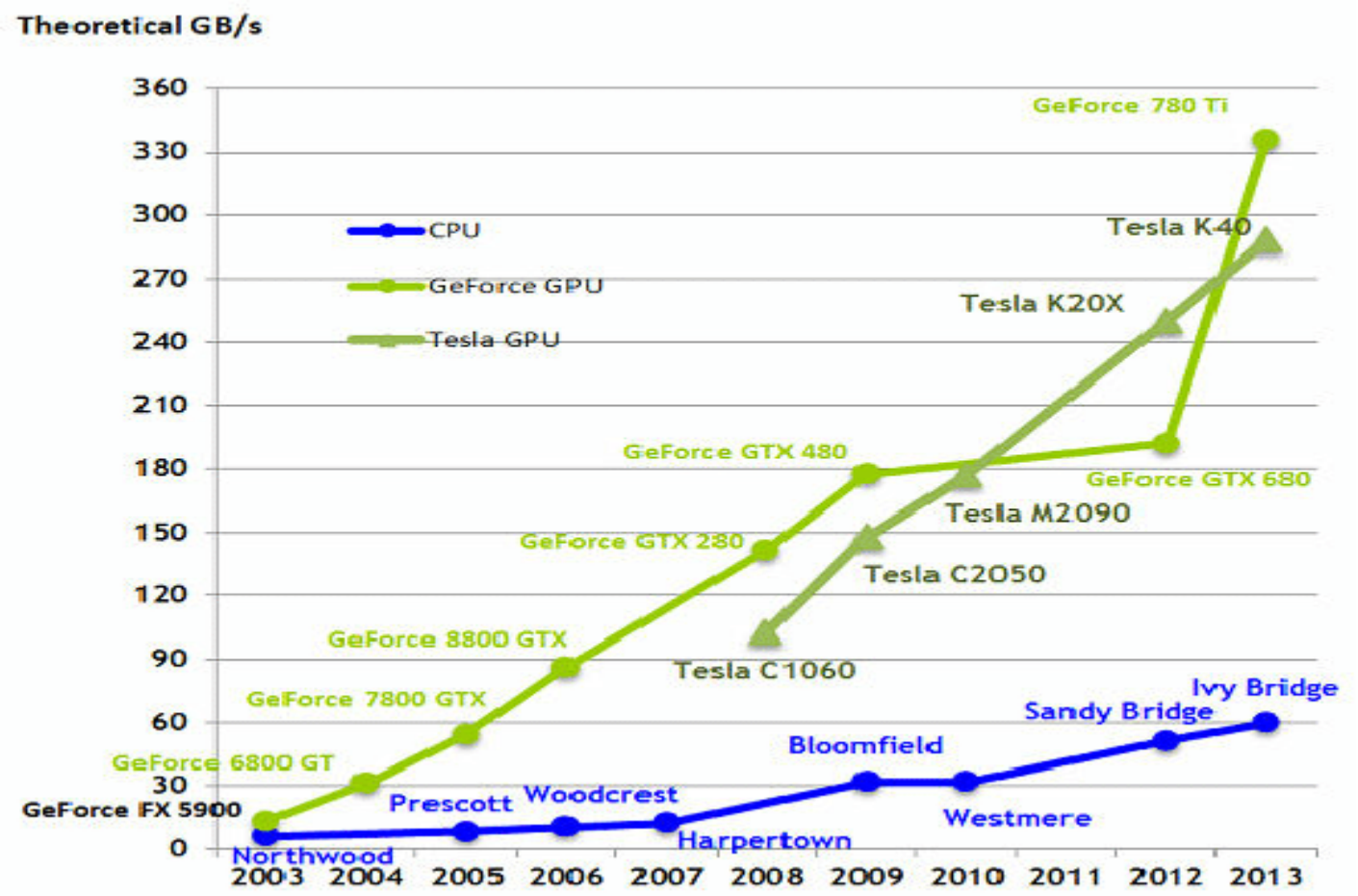
Driver of the Big Data (cont.)

The algorithm also gains improvement.

- In 1988, a benchmark production planning model would have taken 82 years to solve.
- In 2003, the same model would taken only 1 minutes.
- Improved by 43 * 10^6.
- Some parts come from increased processor speed (factor of 1,000) and some parts come from algorithms (factor of 43,000).

- Source:Kurzweil, R. (2013).*How to create a mind: The secret of human thought revealed*. Penguin.

Driver of the Big data



- GPU becomes a part of modern super computing
- Parallel computing

Potential domain applications in Thailand

- Tourism
- Logistic and Transport
- Medical heath
- Agriculture
- Energy
- Financial

- **Let’s discuss here (idea)**

Potential domain applications (cont.)

- Market segmentation – Identify the common characteristic of customers
- Fraud detection – Identify transactions that are most likely to be fraudulent.
- Activity recognition – identify which activity base on sensors historical records.
- Lie Detection – bringing out the truth from a criminal, data collected from previous
- Destination Recommendation System

Question and discussion

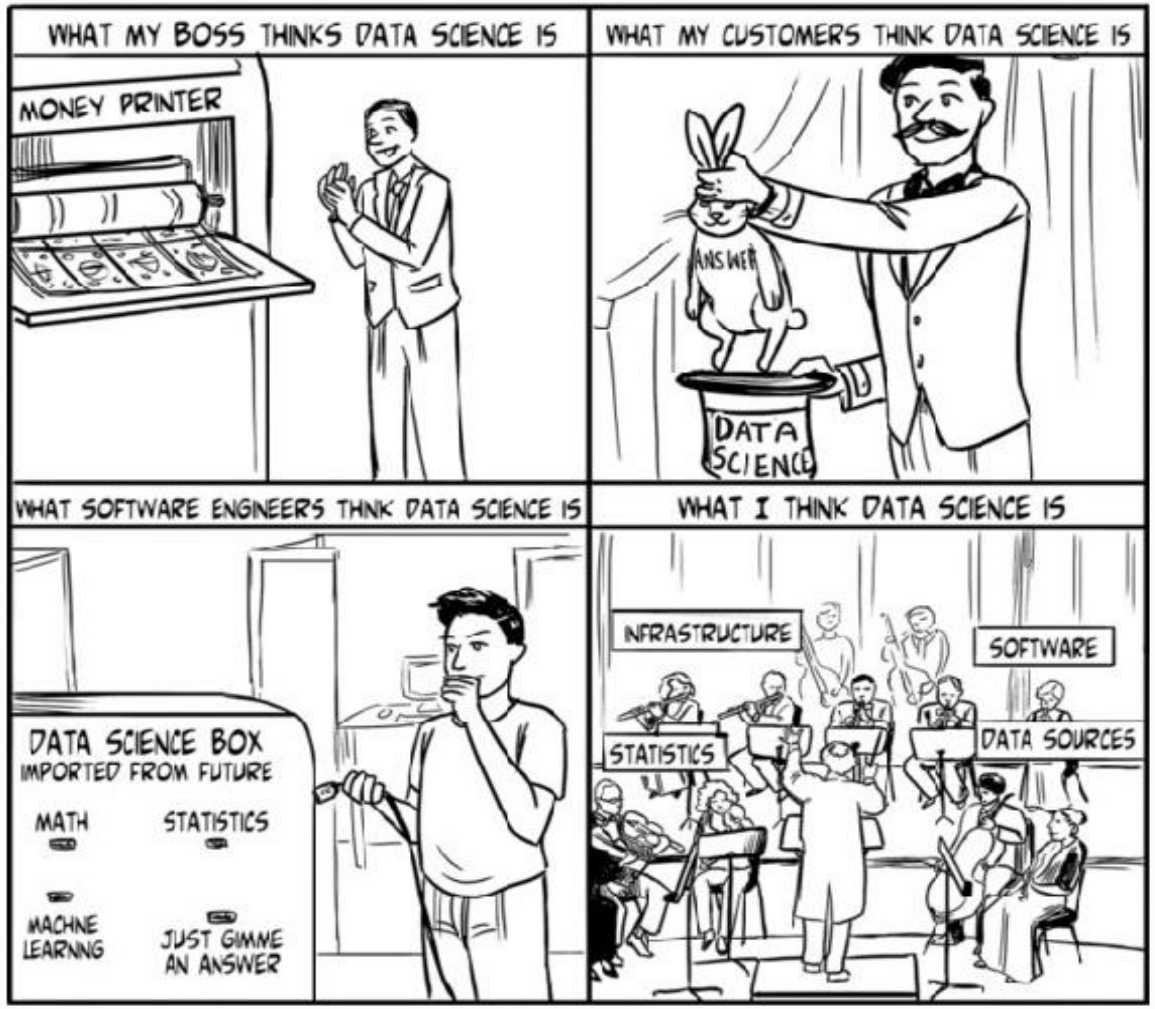
Coffee Break

Data Scientist Preface

- In 2012, Harvard business review
- The sexiest job of the 21st century
- Most of the books explained how to use the latest tools and tech.
- Where do we start?

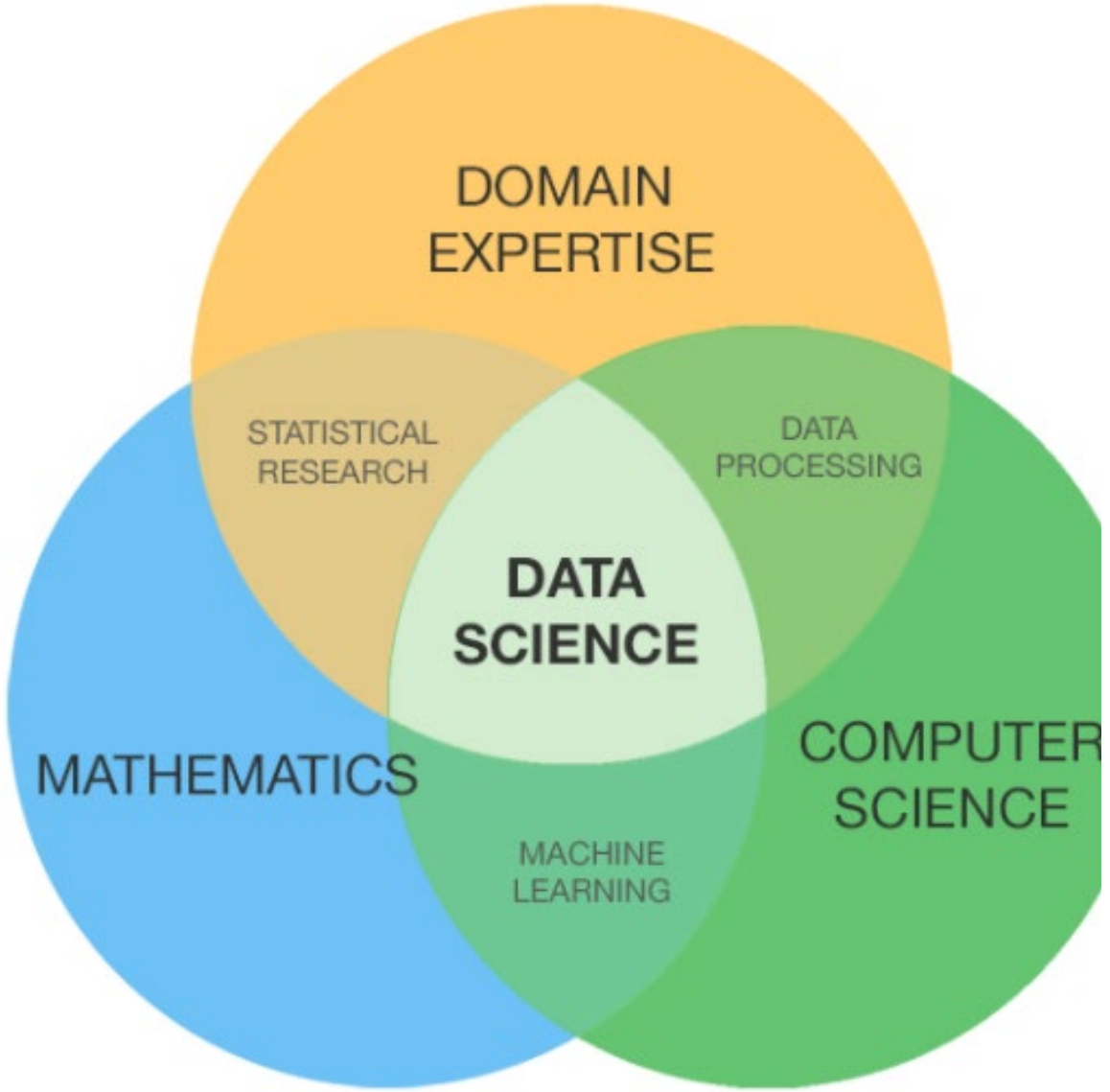
The role of data scientist

Figure 1.1. Some stereotypical perspectives on data science



Origins of data sciences

- Multi-disciplines (see figure)
- In practical: Statistics + software development + domain knowledge
- “A good data scientist can switch domains and begin contributing relatively soon.”



Definition of data science

"The systematic study of digital data using scientific techniques of observation, theory development, systematic analysis, hypothesis testing, and rigorous validation"

The National Consortium for Data Science

Developer vs Data scientist

- Somewhat common (good at designing and building complex system, with tools and frameworks)
- Software dev. -> well-defined components
- Data Science -> work on component isn’t well defined (i.e., data pre-processing, analysis)
- Data Science Data Science -> create system that rely on statically results

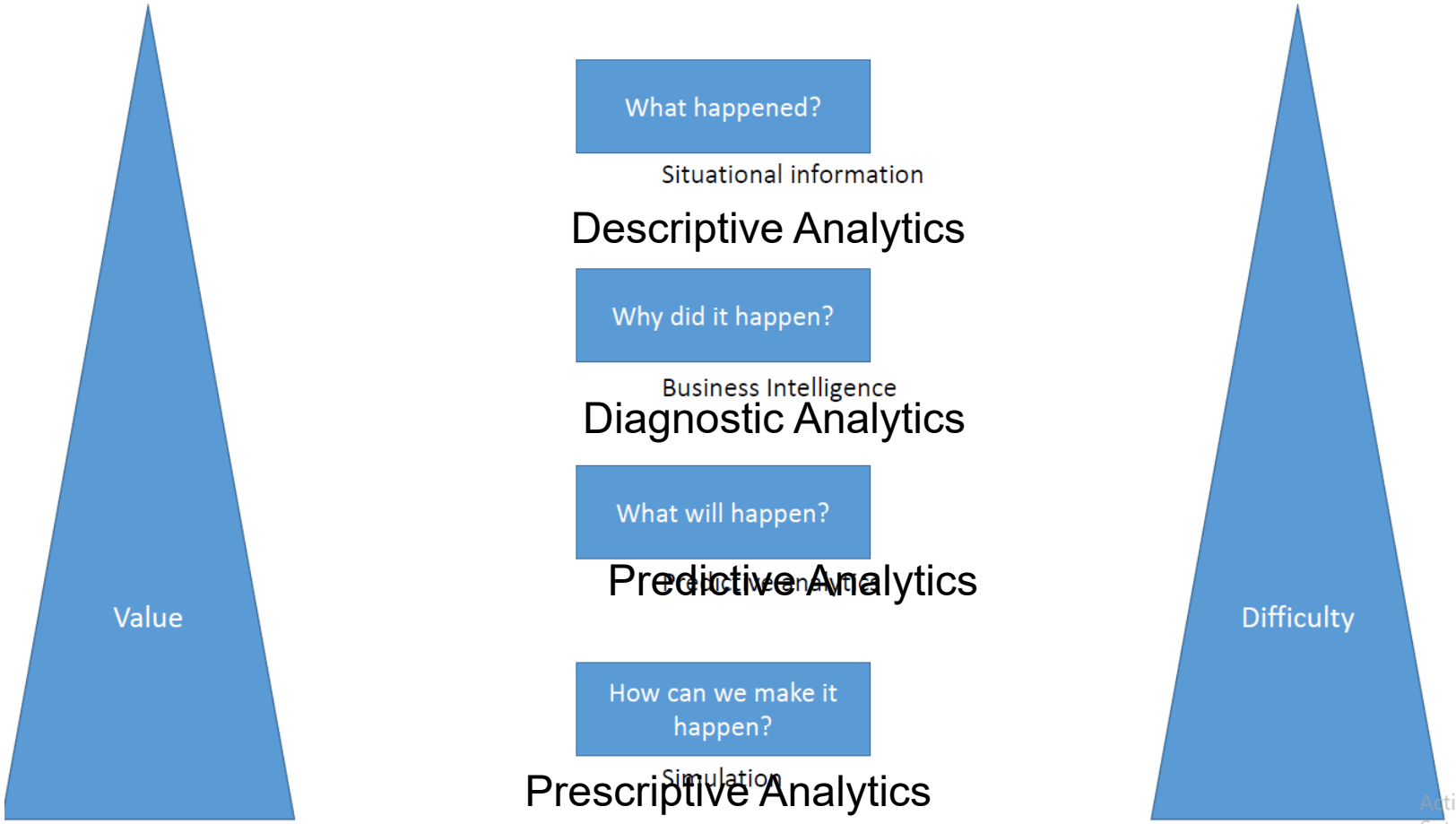
Developer vs Data scientist (cont.)

Dealing with uncertainty is often what separates the role of a data scientist from that of a software developer.

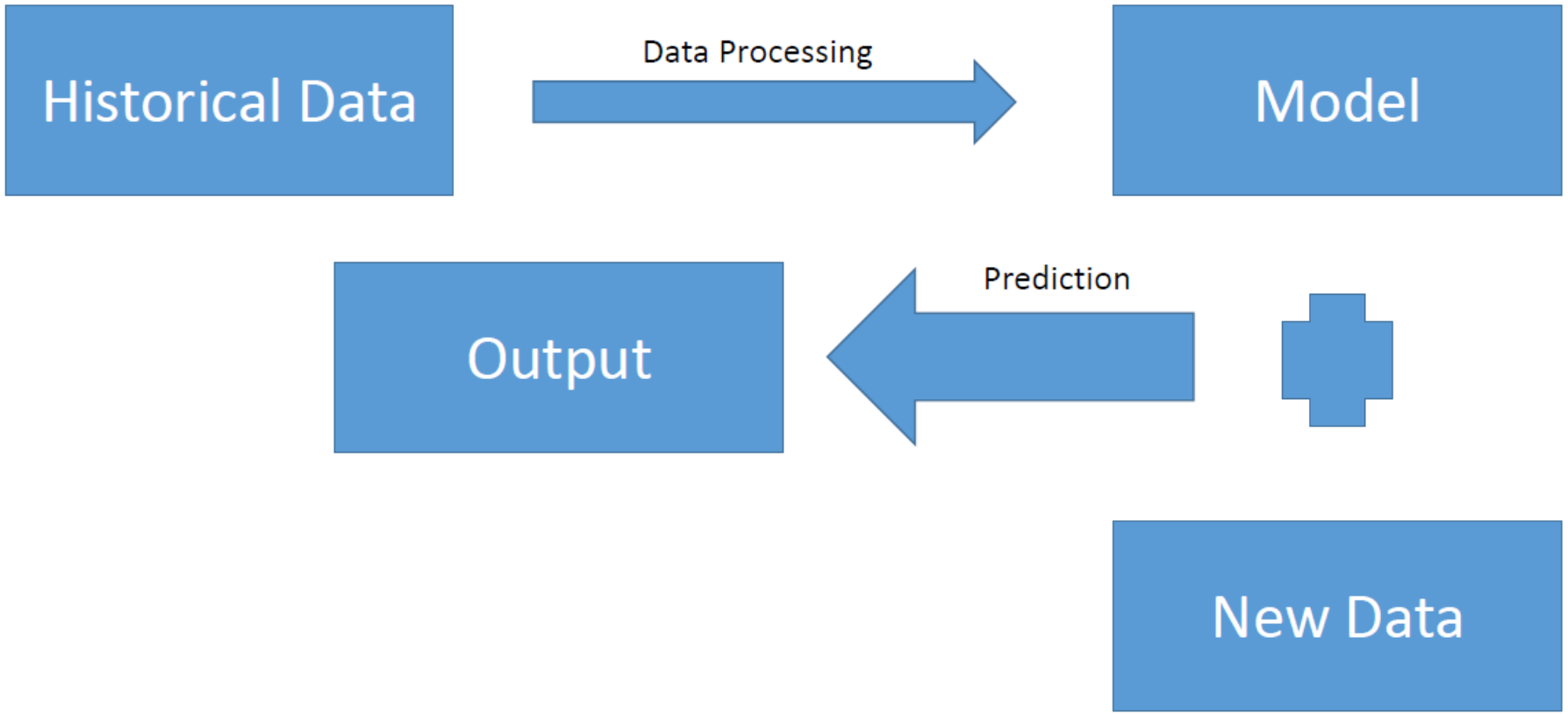
Goal of Data science

- “Turn data to data product”

The data analytic



Data Science operation (naive)



Think like data science

1

Knowledge first— Get to know your problem, your data, your approach, and your goal before you do anything else, and keep those at the forefront of your mind.

2

Technology second— Software is a tool that serves you. It both enables and constrains you. It shouldn't dictate your approach to the problem except in extenuating circumstances.

3

Opinions third— Opinions, intuition, and wishful thinking are to be used only as guides toward theories that can be proven correct and not as the focus of any project.

Workshop 1 (Simple weather data set)

- A simple data set of temperature for each city

```
-2 -2  2  8 14 19 21 20 16 10  4 -1
20 20 22 23 26 27 28 28 27 26 23 20
-1  1  6 13 18 23 26 25 21 15  7  1
11 13 16 20 23 27 27 27 26 21 16 12
-8 -5  0  7 14 19 22 20 16 10  2 -5
10 12 16 20 23 27 28 28 26 21 16 12
12 14 16 21 26 31 33 32 30 23 16 12
 5  6  7 10 13 16 18 18 16 12  8  6
10 12 12 13 14 15 15 16 17 16 14 11
13 14 15 16 17 19 21 22 21 19 16 14
```

Workshop 1 (Simple weather data set)

- **Problem 1**: find an average temperature for each city.
- **Problem 2:** Assume we have three seasons in year Find an average, minimum, and maximum temperature for each season.

- Use any programming language, MS-Excel, or calculate by hand on sheet.
- Submit your source code, sheet to the workshop 1 section

References:

- Godsey, B. (2017). *Think like a data scientist*. Manning Publications Co. (chapter 1)