Data Science and its role in Big Data analytics

WE ALL CAN BE DATA SCIENTISTS NOW!

I. Properties of Data: What is a Big Data?
II. What is Data Science?
III. Data Science Process
IV. Why Data Science is in a sudden boom?
V. Who is a Data Scientist?

Kyungwon Kim

Assistant Professor
Department of International Trade
College of Global Political Science and Economics
Incheon National University

WE ALL CAN BE DATA SCIENTISTS NOW!

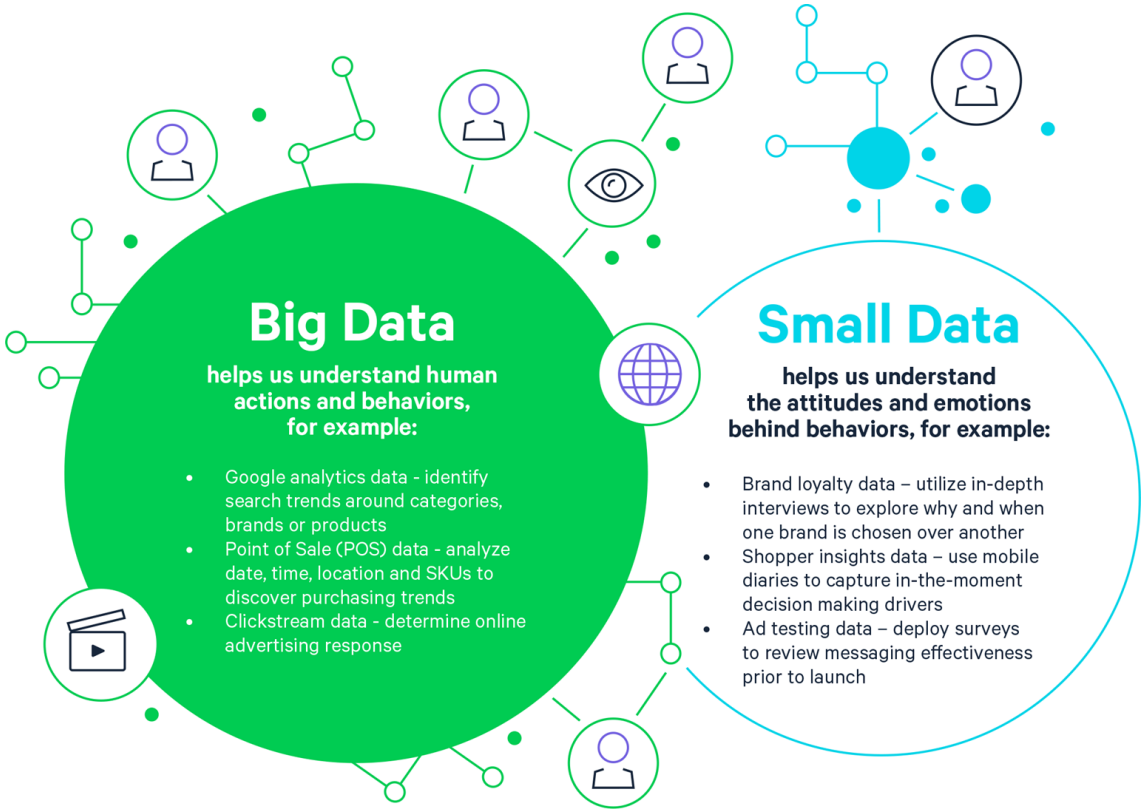
I. Properties of Data: What is a Big Data?
II. What is Data Science?
III. Data Science Process
IV. Why Data Science is in a sudden boom?
V. Who is a Data Scientist?

Properties of Data: What is Big Data?

➢ Big Data vs. Small Data

- **Small Data:** Getting machines to do what humans are good at.
- **Big Data:** Feeding an algorithm data to learn and predict something.

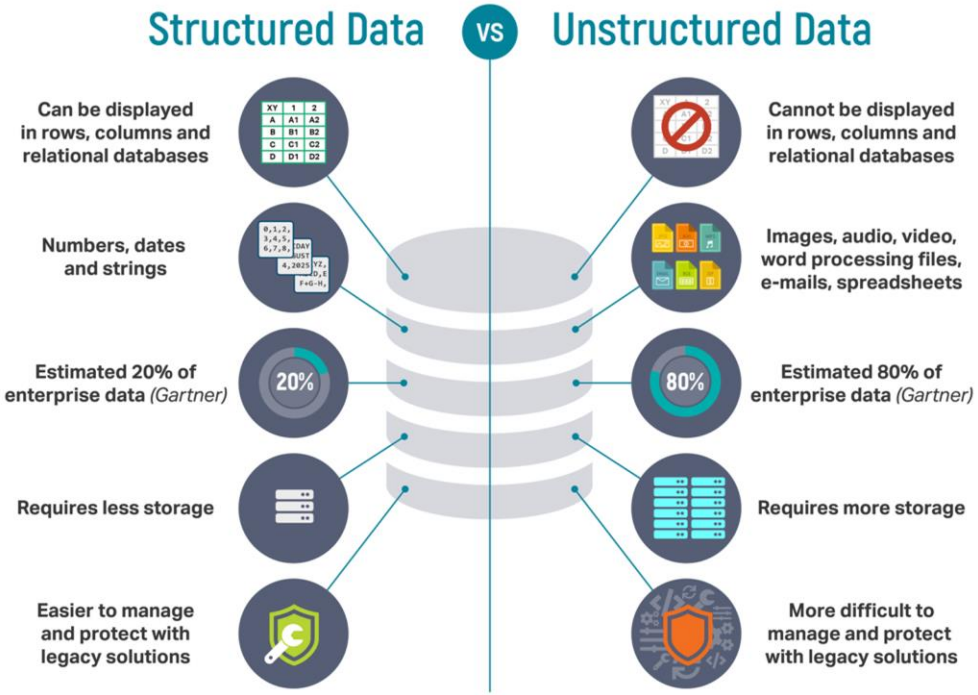
=> A Powerful Equation: Real Human Insight = Big Data + Small Data



Properties of Data: What is a Big Data?

➢ Structured vs. Unstructured Data

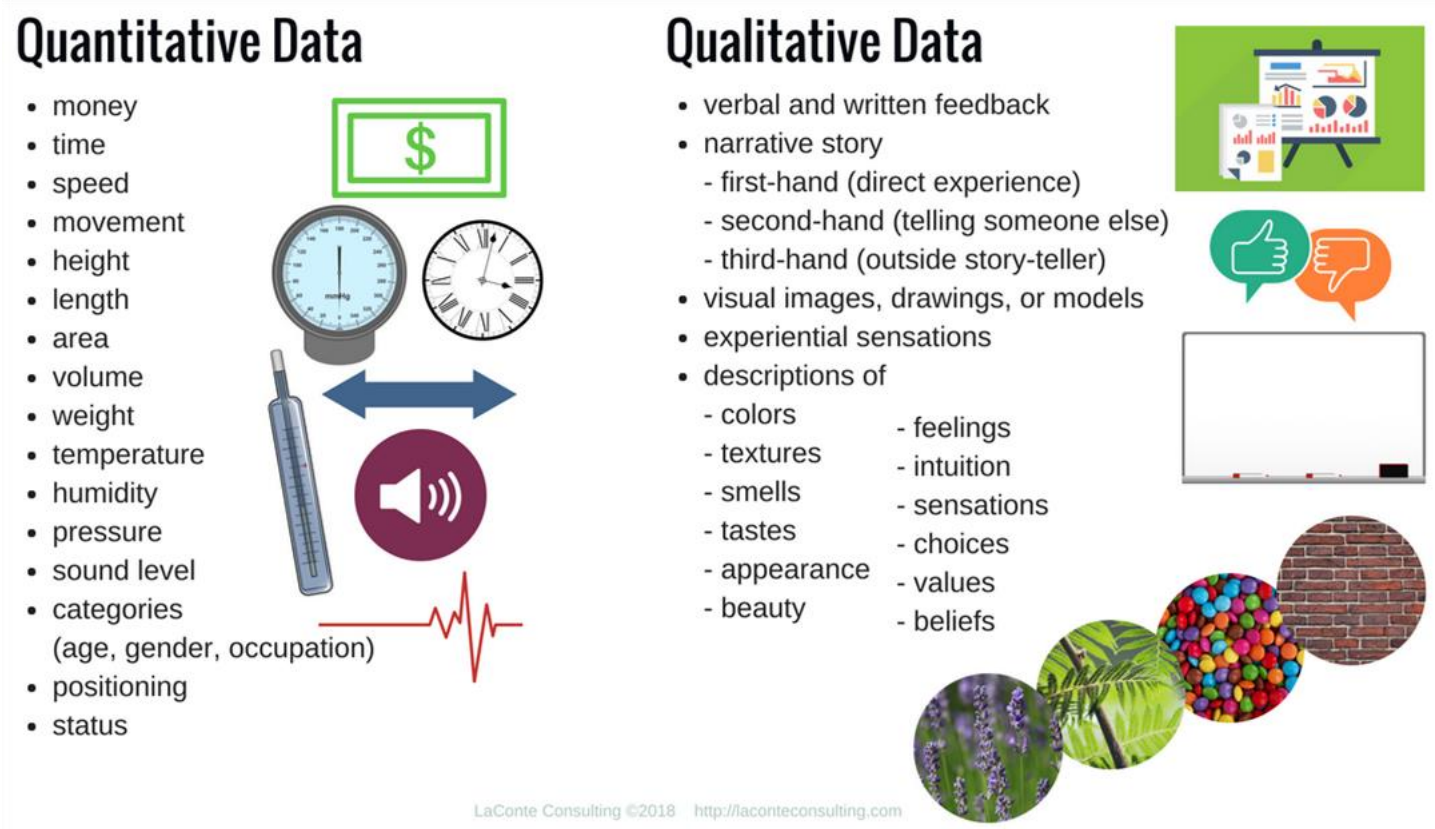
- **Structured Data:** the type of data that fits nicely into a relational database. It's highly organized and easily analyzed. Most IT staff are used to working with structured data.
- **Unstructured Data:** It doesn't fit nicely into a spreadsheet or database. It can be textual or non-textual. It can be human- or machine-generated.



Properties of Data: What is a Big Data?

➢ Quantitative vs. Qualitative Data

- **Quantitative Data:** Numerical calculations and measurements.
- **Unstructured Data:** Sensations, feelings, and experiences.



Properties of Data: What is a Big Data?

➢ Output (Y): Labelled vs Unlabelled

Lets say we want to **Classify Houses by Size**

Given Features or Feature Set

FullBath	HalfBath	Bedrooms	Home Age	Size
1	0	2	56	M
1	1	3	59	L
2	1	3	20	M
2	1	3	19	S

← Label

Supervised Learning

Use the labels to build a model. Model used to classify new house size based ONLY on the known feature set.

Unsupervised

SIZE is missing! We need to look for similarities in the data and group them into clusters.

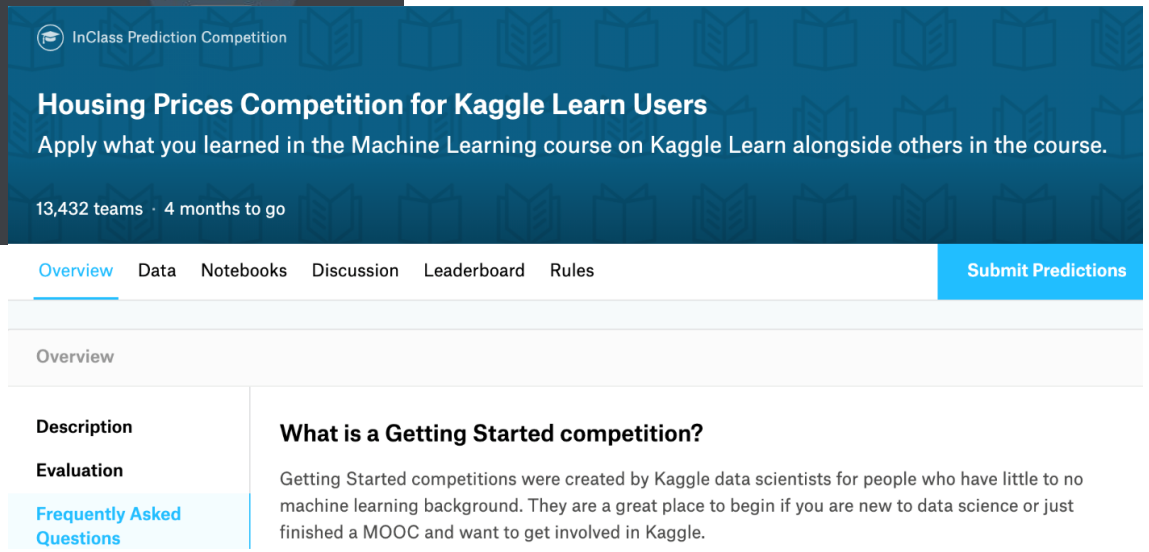
WE ALL CAN BE DATA SCIENTISTS NOW!

I. Properties of Data: What is a Big Data?
II. What is Data Science?
III. Data Science Process
IV. Why Data Science is in a sudden boom?
V. Who is a Data Scientist?

What is Data Science?

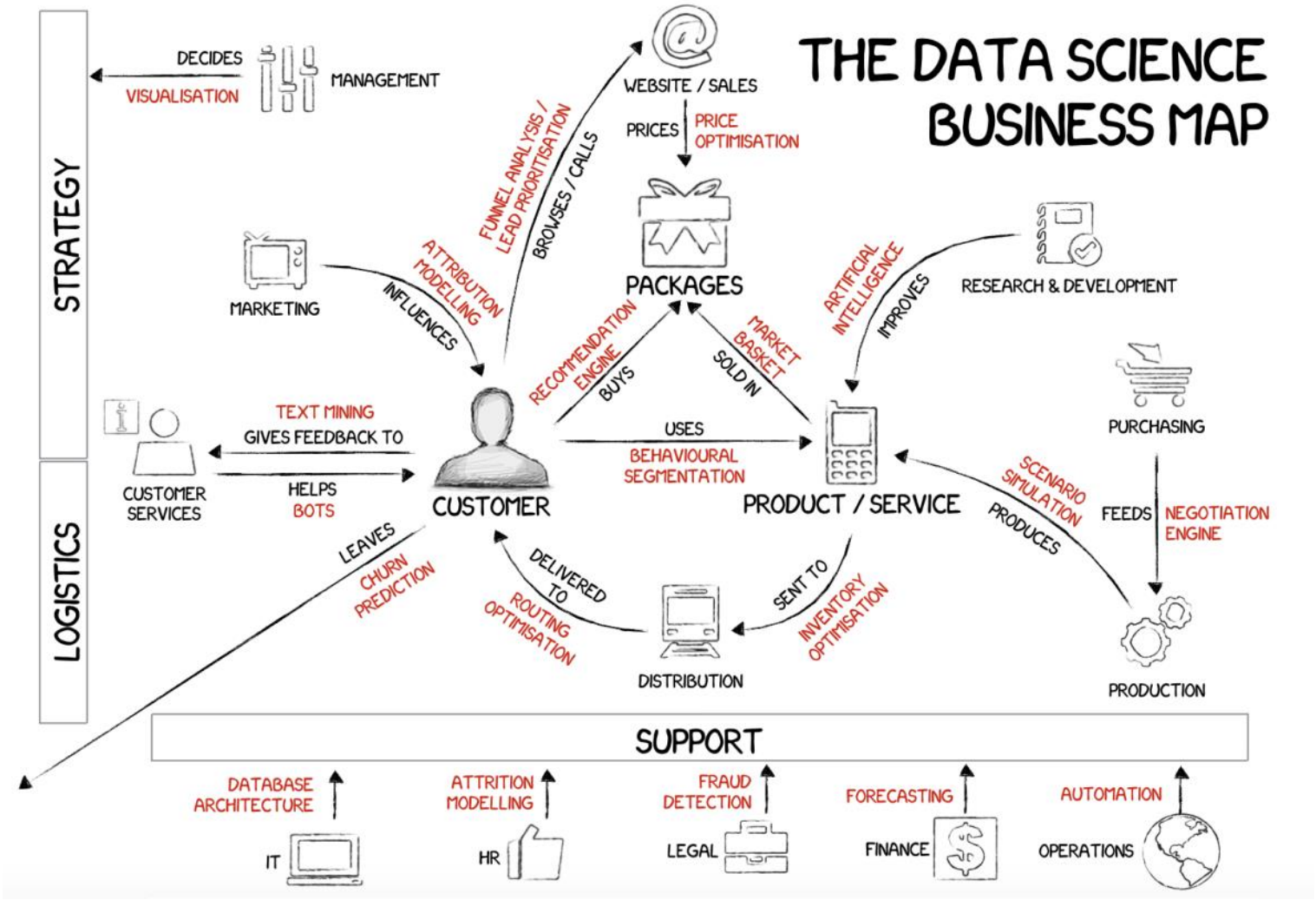
➢ The Real World Data Science is not a Kaggle Competition

- It can be worthwhile to step back a little and realize what exactly your **ultimate goal** is.
- The **best performance** might not be equivalent to a model yielding the **best score** in real.



What is Data Science?

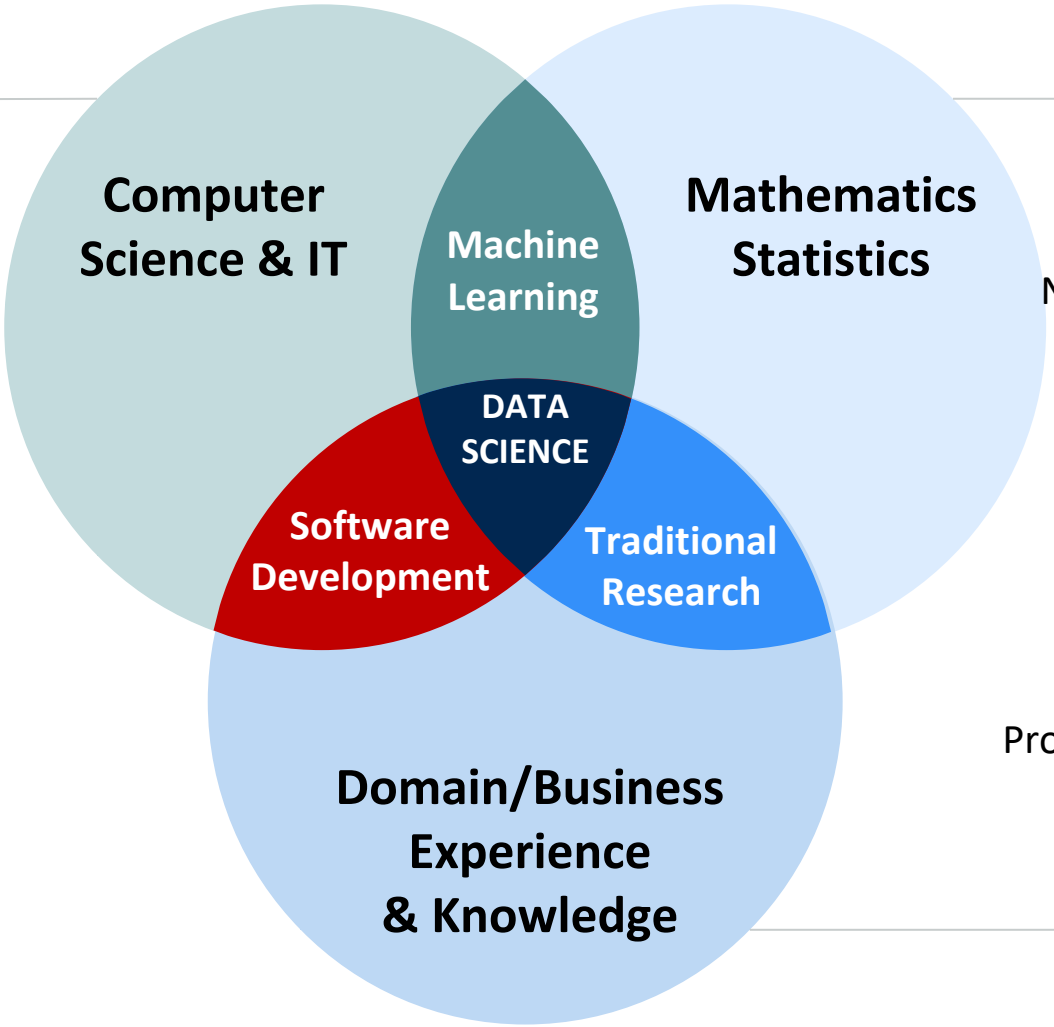
➢ The Real World Data Science



What is Data Science?

➢ Data Science: Solving Problems with Data (Real Human Insight)

Computer Science,
Data Engineering,
Data Warehouse,
Pattern Recognition,
High Performance



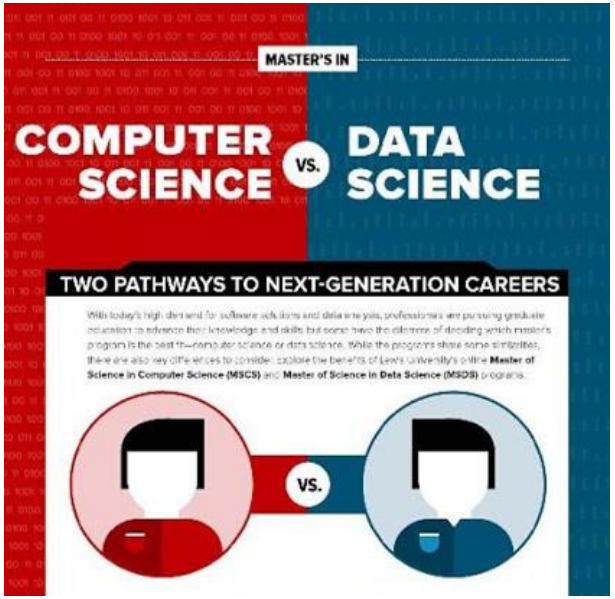
Statistical Learning,
Machine Learning,
Probability,
Numerical Techniques
to derive Insights

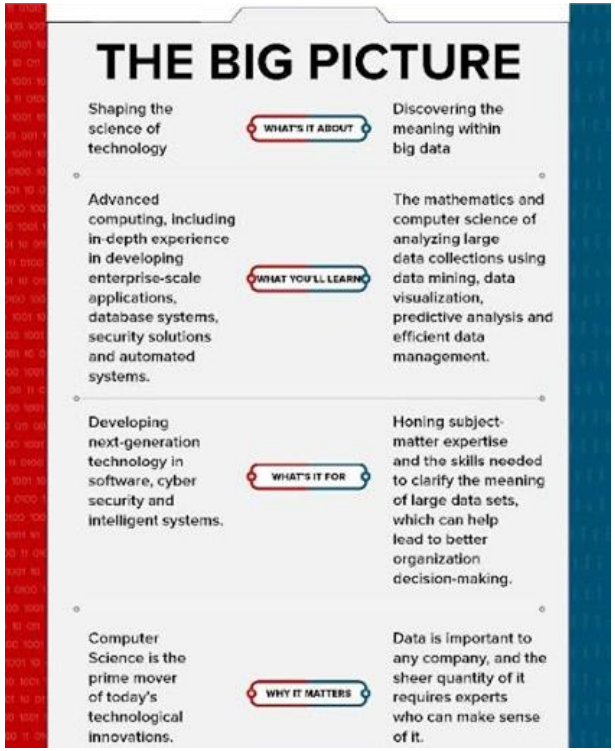
Problem and Objectives,
Domain Knowledge,
Business Experience,
Value to the Business

What is Data Science?

➢ Computer Science, Science, and Data Science

- **Computer Science:** The study of the theory and practice of how computers work.
- **Science:** Focusing on solving problems through the lens of the domain's scientific principles.
- **Data Science:** Interdisciplinary field involving computer science and statistics.



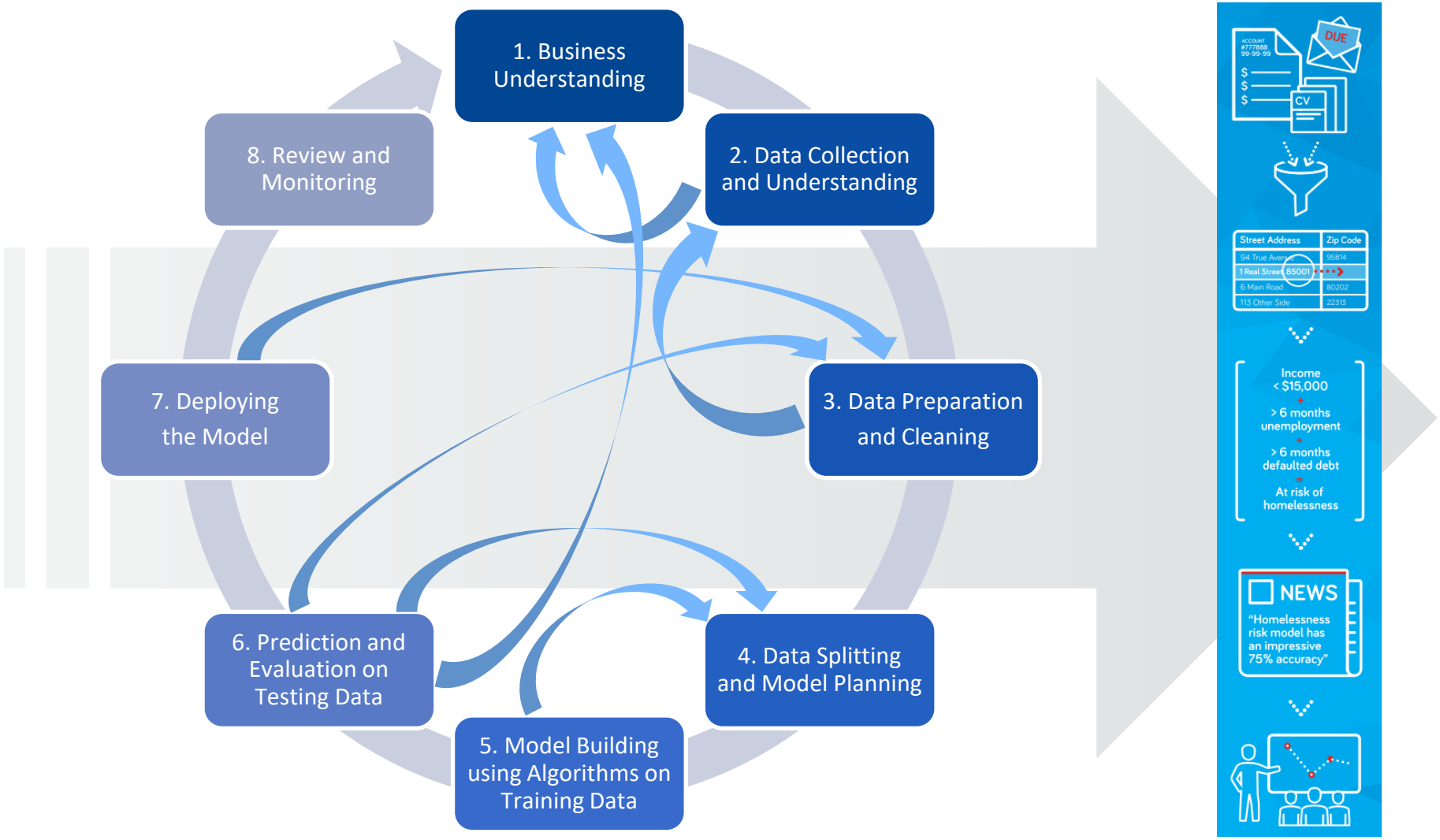


WE ALL CAN BE DATA SCIENTISTS NOW!

I. Properties of Data: What is a Big Data?
II. What is Data Science?
III. Data Science Process
IV. Why Data Science is in a sudden boom?
V. Who is a Data Scientist?

Data Science Process

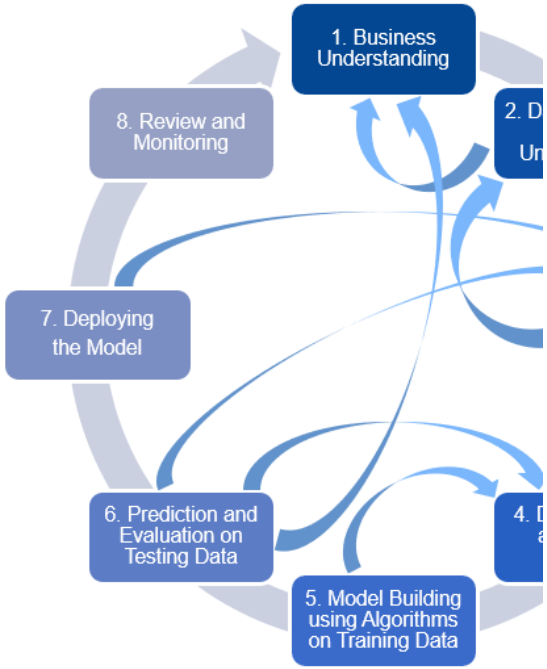
➢ Getting from Raw Data to Outcomes



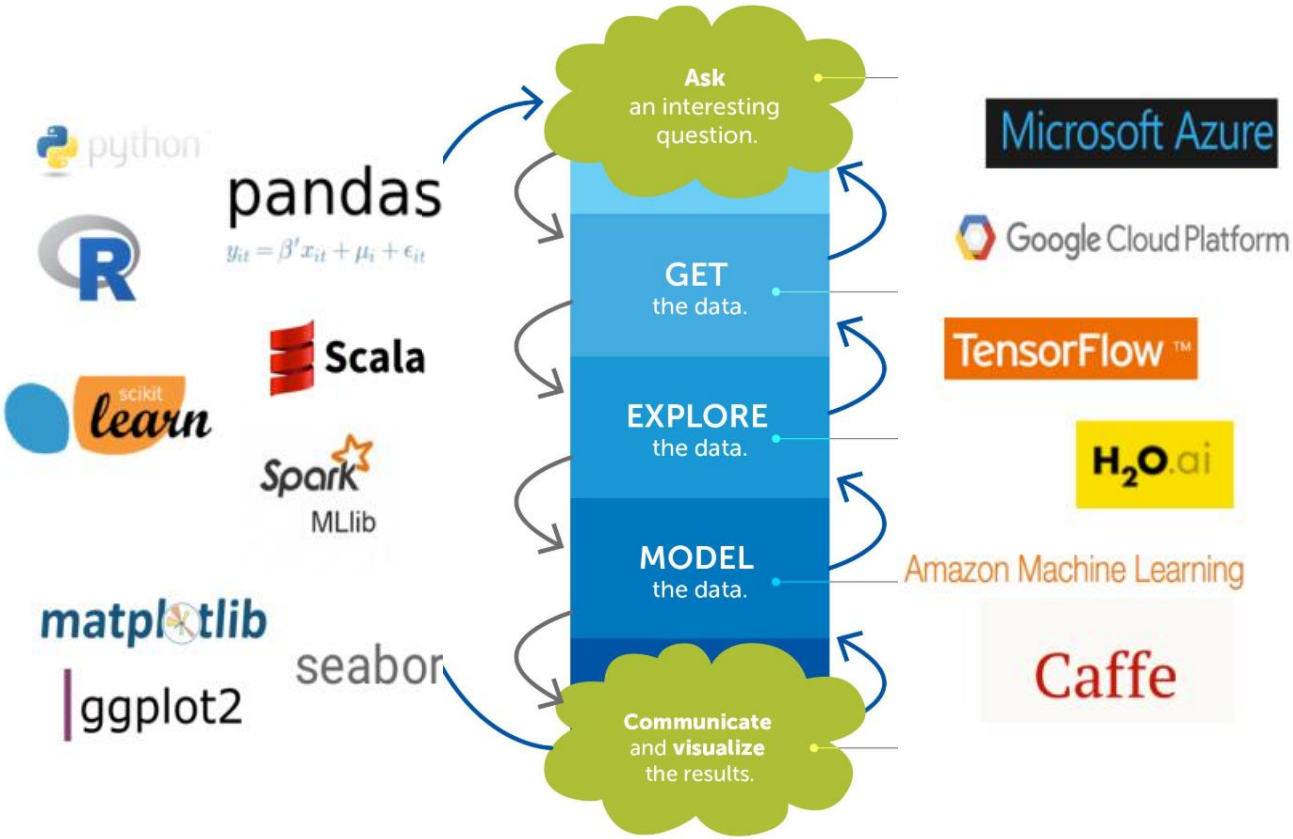
Data Science Process

➢ Getting from Standard Framework to Data Science

“Standard Framework for Data Mining”



“The Data Science Workflow”



WE ALL CAN BE DATA SCIENTISTS NOW!

I. Properties of Data: What is a Big Data?
II. What is Data Science?
III. Data Science Process
IV. Why Data Science is in a sudden boom?
V. Who is a Data Scientist?

Why Data Science is in a sudden boom?

➢ Why use Machine Learning instead of Traditional Statistics?

Traditional Statistics — **Machine Learning**

White-box modelling
simpler computation, emphasis on introspection, form, causal effects and processes, finding a 'correct' model

Black-box modelling
high computational complexity, emphasis on speed and quality of prediction, finding a 'performant' model

- **Bayes Theorem:**

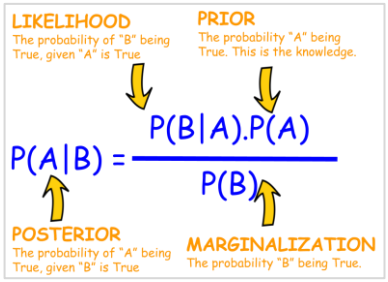
Thomas Bayes mid 1700's

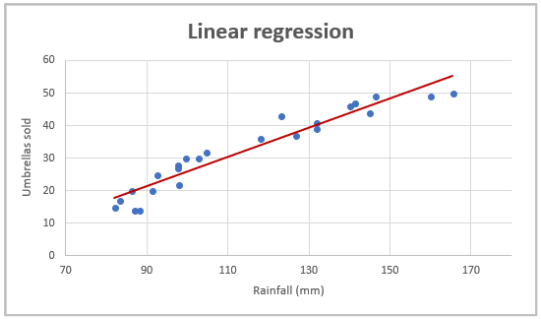
- **Regression:**

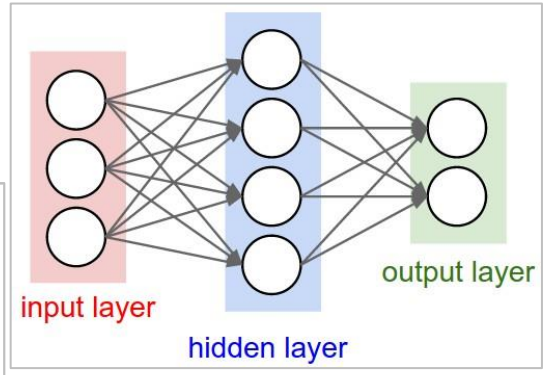
Legendre, Gauss and Galton early 1800's

- **Neural Networks:**

McCulloch and Pitts early 1940s

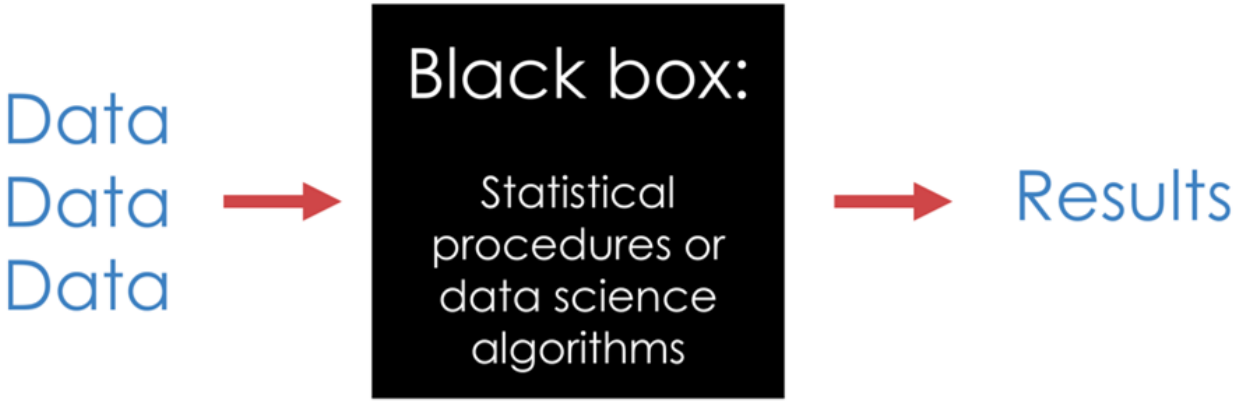




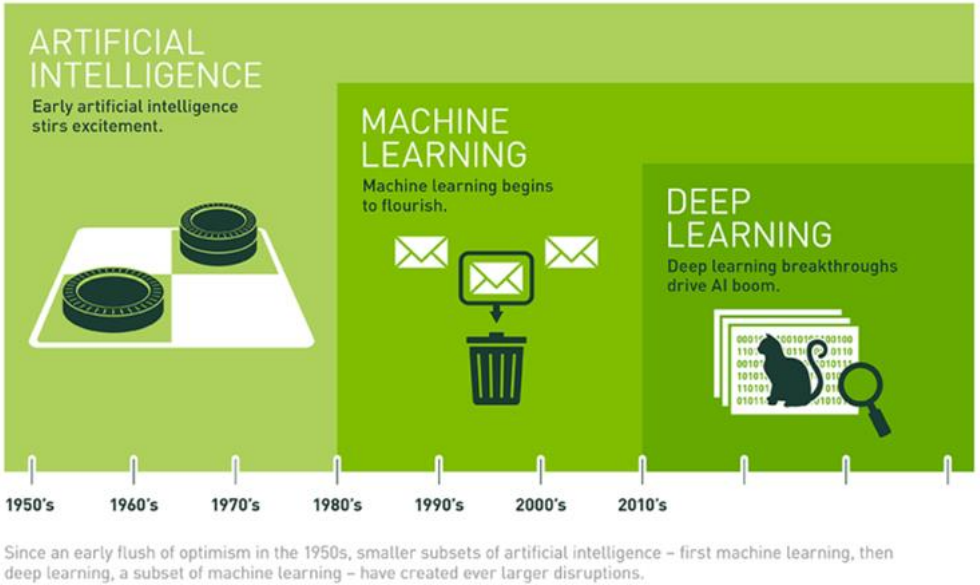


Why Data Science is in a sudden boom?

➢ Why use Machine Learning instead of Traditional Statistics?



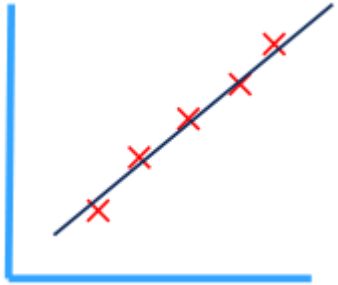
- **AI:** Getting machines to do what humans are good at
- **Machine Learning:** Feeding an algorithm data to learn and predict something
- **Deep Learning:** A type of machine learning



Why Data Science is in a sudden boom?

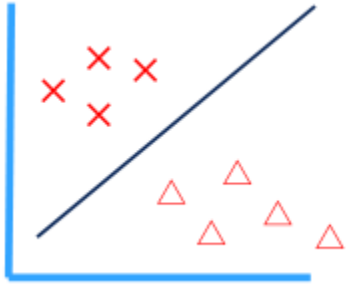
➢ Solution directions to the black box problem

- How much is the stock of Samsung Electronics tomorrow?



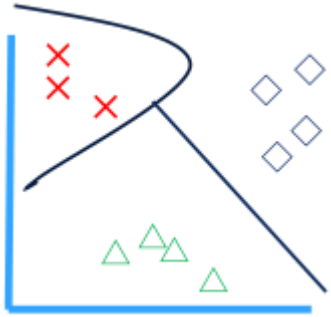
Regression – Looking for a statistical relationship across variables that may give us an estimate of a particular outcome. (Supervised)

- Will Samsung Electronics' stocks rise or fall tomorrow?



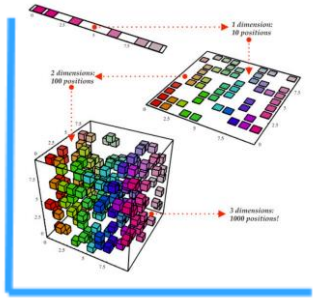
Classification – Similar to regression but looking for separations in the data given predefined classes. (Supervised)

- Are Samsung Electronics and Naver similar business companies?



Clustering – Do not have predefined classes but trying to find groups or sets based upon data at hand. (Unsupervised)

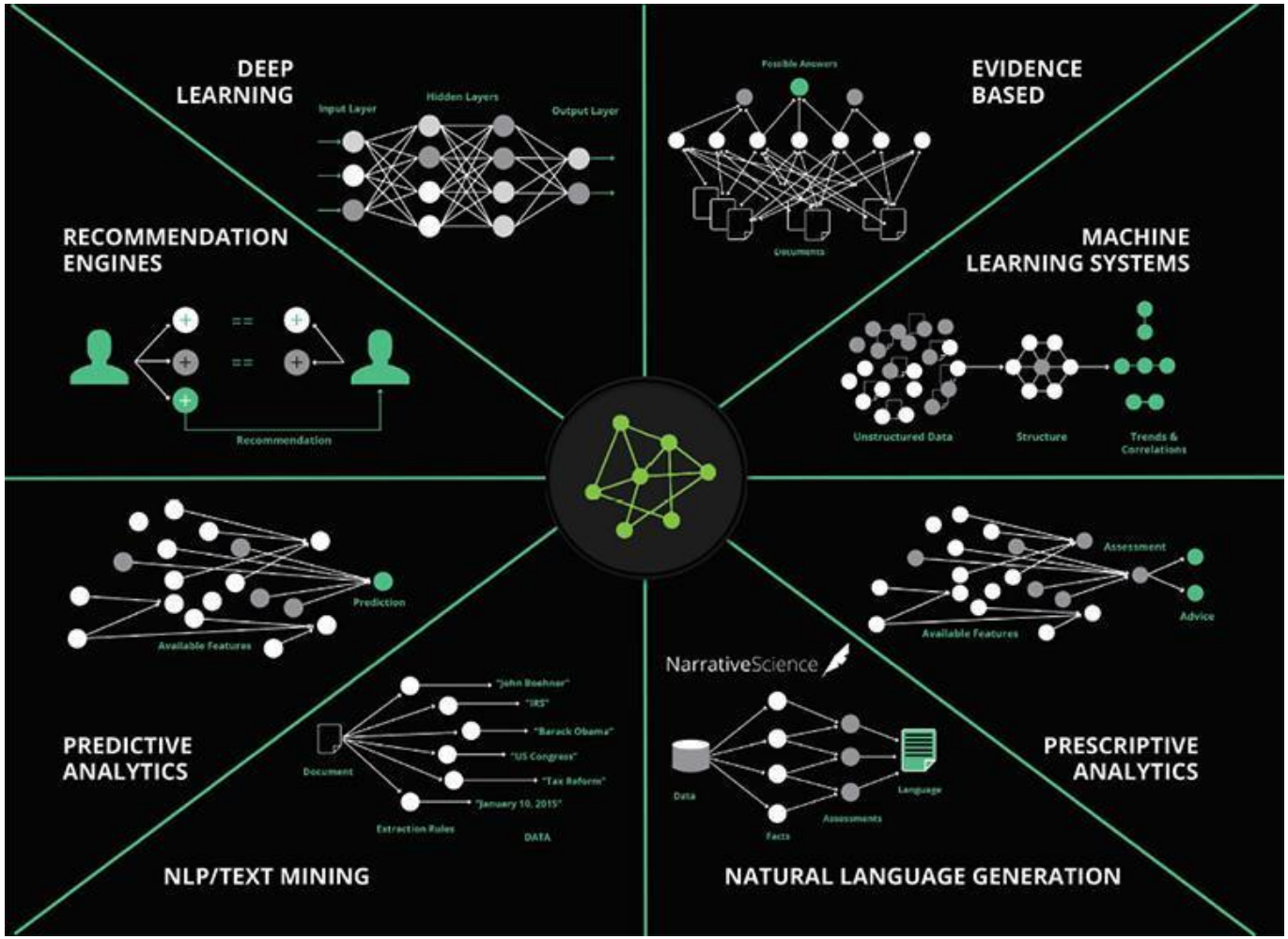
- What are the representatives among all stocks in the KOSPI?



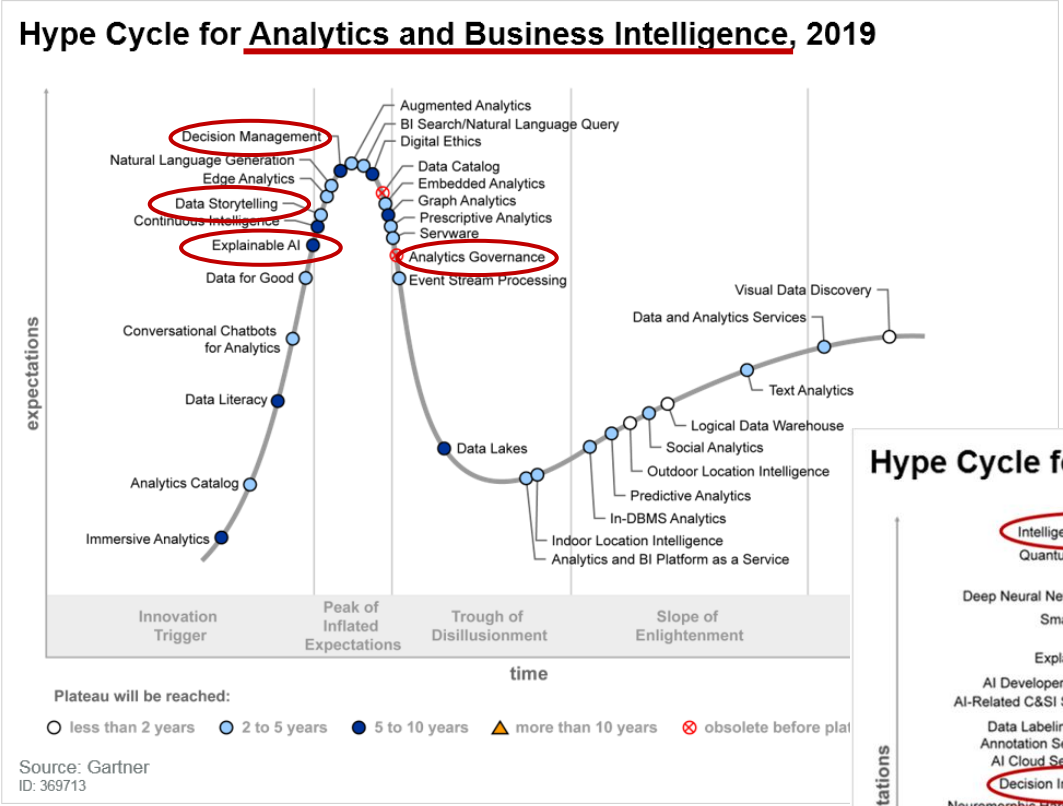
Dimensionality Reduction – Transformation of data from high-dimensional into a low-dimensional space so that it retains some meaningful properties of the origin data. (Unsupervised)

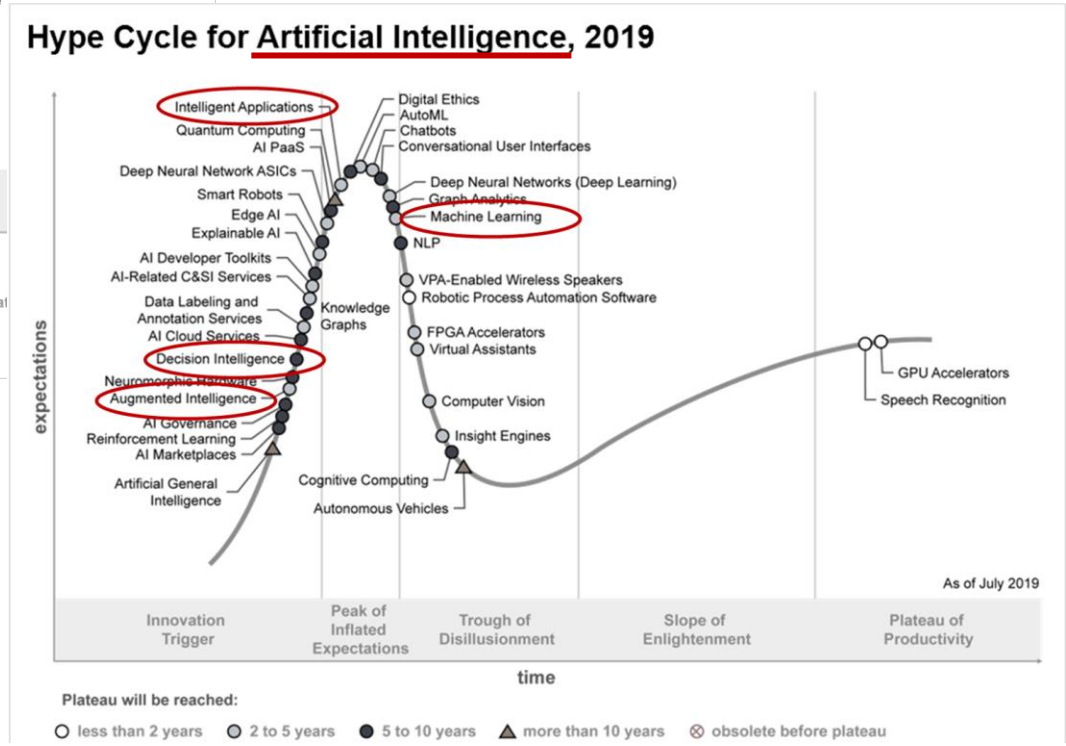
Why Data Science is in a sudden boom?

➢ Different types of Machine Learning algorithms explained



Global “Technology” Trends





WE ALL CAN BE DATA SCIENTISTS NOW!

I. Properties of Data: What is a Big Data?
II. What is Data Science?
III. Data Science Process
IV. Why Data Science is in a sudden boom?
V. Who is a Data Scientist?

Who is a Data Scientist?

- **What should we do?**

Who is a Data Scientist?

➢ Traditional Specialist of Data Science Team

Data Analyst (DA)

: Assist DS with domain understanding, data preprocessing and problem defining.

Data Scientist (DS)

: Prepares data, engineers features, most valuable skill: training models.

Data Engineer (DE)

: Data acquisition focus. Build data pipelines. Not uncommon to have 5:1 ratio DE:DS

Data Application Architect (DAA)

: Design complete solution; deploy and maintain models in production

Who is a Data Scientist?

➢ Traditional Specialist of Data Science Team

Data Analyst	**Analyst**	**(정통) 기획 및 전략**	- 기획서, 보고서, 기초분석, 시각화
	Data Analyst	**데이터분석 설계**	- 데이터 표준 확립, 구조 및 품질관리, 데이터 분석
Data Scientist	**Statistician**	**(정통) 통계기반 연구**	- 결과가 나온 이유를 연구하는 방향
	Data Scientist	**데이터분석 전반 연구**	- Why보다 Result 마련/대응/전략/의사결정 방향
	ML / DL Engineer	**응용 및 구현 연구**	- 연구결과를 실질적 서비스 및 비즈니스 집중화
Data Engineer	**Database Administrator**	**데이터베이스 관리자**	- 데이터베이스 운영, 관리, 설계
	Back-end Engineer	**(정통) 백엔드 개발자**	- 서버 개발, 데이터베이스 시스템 구현 및 관리 - RDB/NOSQL/Hbase/Spark 등
	Infra Engineer	**인프라 엔지니어**	- 데이터 파이프라인 구축 및 운영 - Cloud/Spark/Hadoop 등

Who is a Data Scientist?

➢ Typical Collaboration of Data Science Project

- Makes data science teams more productive
- Broad support for open source libraries in various languages

Understand Business Objectives

ID Mapping Procure Training Data

Prepare Data and Build Features

Train, Tune, and Test Models

Deploy and Operationalize Models

Update Models

Who is a Data Scientist?

➢ What is a Project of Data Science?

- **TASKS**

: In addition to advanced analytic skills, this individuals are also proficient at integrating and preparing large, varied datasets, architecting specialized database and computing environments, and communicating results.

- **MISSION**

: A data scientist may or may not have specialized industry knowledge to aid in modeling business problems and with understanding and preparing data.

- **TALENT**

: Creating value from data requires a range of talents from data integration and preparation, to architecting specialized computing/database environments, to data mining and intelligent algorithms.

- **RESPONSIBILITY**

: An individual responsible for modeling complex business problems, discovering business insights and identifying opportunities through the use of statistical, algorithmic, mining and visualization techniques.

Data Science Roadmap

➢ We can be the best Data Science team

Data Analysis	Data Analysis Cycle	●	●	●	●
	Data Visualization and Communication	●	●	○	○
	Data Wrangling and Intuition	●	●	●	○
Mathematics	Linear Algebra	○	●	○	○
	Numerical Analysis	○	●	○	○
	Optimization	○	●	○	○
	Multivariate Calculus	○	●	○	○
Statistics	Probability and Statistics	○	●	○	○
	Experimental Design	○	●	○	○
	Statistical Thinking and Algorithms	○	●	○	○
Artificial Intelligence	Machine Learning	○	●	○	○
	Deep Learning	○	●	○	○
Computation	Databases and Distributed Systems	○	○	○	●
	Programming Tools	○	●	●	○
	Algorithmic and Programming	○	○	●	●
	Software Engineering	○	○	●	●
	Platform Understanding	○	○	○	●

○ Not that important ○ Somewhat important ● Very important

Data Science Curriculum

➢ Related 44 Lectures (132 Credits)

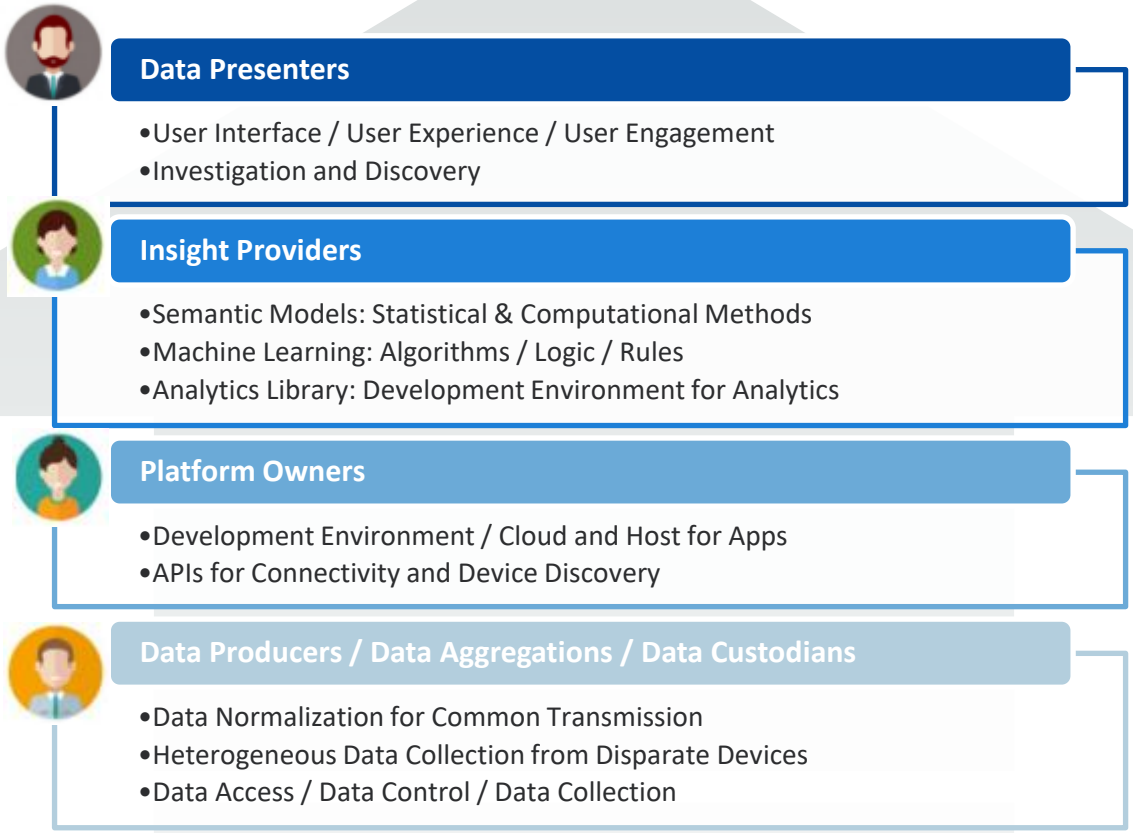
초급 중급 고급

Data Analysis (6)	Mathematics (6)	Statistics (6)	Artificial Intelligence (6)	Computation (10)	Professionalism (10)
• 데이터사이언스 개론 • 데이터분석 언어 • 데이터모델링 • 데이터시각화 • 데이터분석 응용 • 데이터사이언스 응용 및 활용	• 미적분학 기초 • 선형대수 기초 • 수치해석 기초 • 다변량 미적분학 • 인공지능 수치해석 • 최적화	• 확률 및 통계 기초 • 실험계획법 • 통계적사고 및 알고리즘 • 다변량 자료분석 • 시계열분석 • 시뮬레이션	• 데이터마이닝 개론 • 기계학습 • 딥러닝 • 인공지능 • 자연어처리 • 시각처리	• 데이터베이스 시스템 • 자료구조 및 알고리즘 • 서버시스템 및 클라우드 • 빅데이터 분산처리 • 소프트웨어 공학 • 웹마이닝 • 미디어콘텐츠 처리 • 컴퓨터비전 • 정보보안 특론 • 빅데이터 플랫폼 특론	• Communication • Teamwork • Ethical Issues • Privacy and Confidentiality • Legal Considerations • Economic Considerations • Intellectual Property • Change Management • Continuing Professional Development • On Automation

Who is a Data Scientist?

➢ The Insights Revolution?

- A **Data Economy** is a global digital ecosystem in which data is gathered, organized, and exchanged by a network of vendors for the purpose of deriving value from the accumulated information.



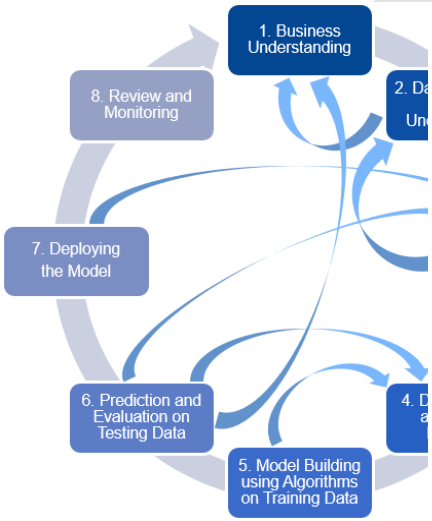
Who is a Data Scientist?

➢ The Insights Revolution?

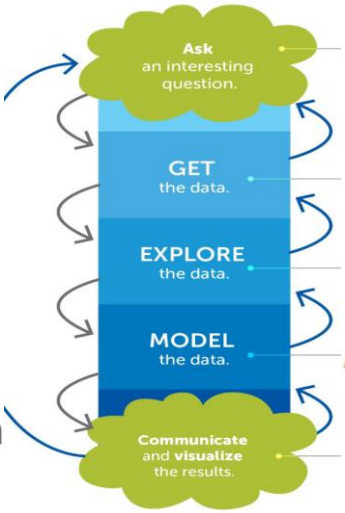
Data Presenters

- User Interface / User Experience / User Engagement
- Investigation and Discovery

"Standard Framework for Data Mining"



"The Data Science Workflow"



Insight Providers

- Semantic Models: Statistical
- Machine Learning: Algorithms
- Analytics Library: Development Environment for Analytics

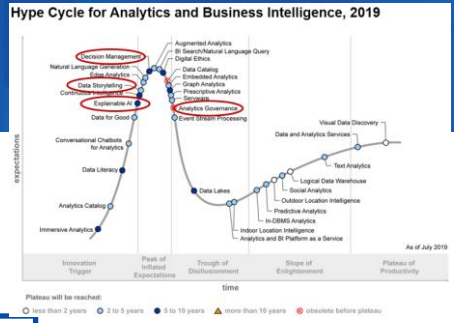
Platform Owners

- Development Environment
- APIs for Connectivity

Data Producers / Data Aggregations / Data Custodians

- Data Normalization for Common Transmission
- Heterogeneous Data Collection from Disparate Devices
- Data Access / Data Control / Data Collection

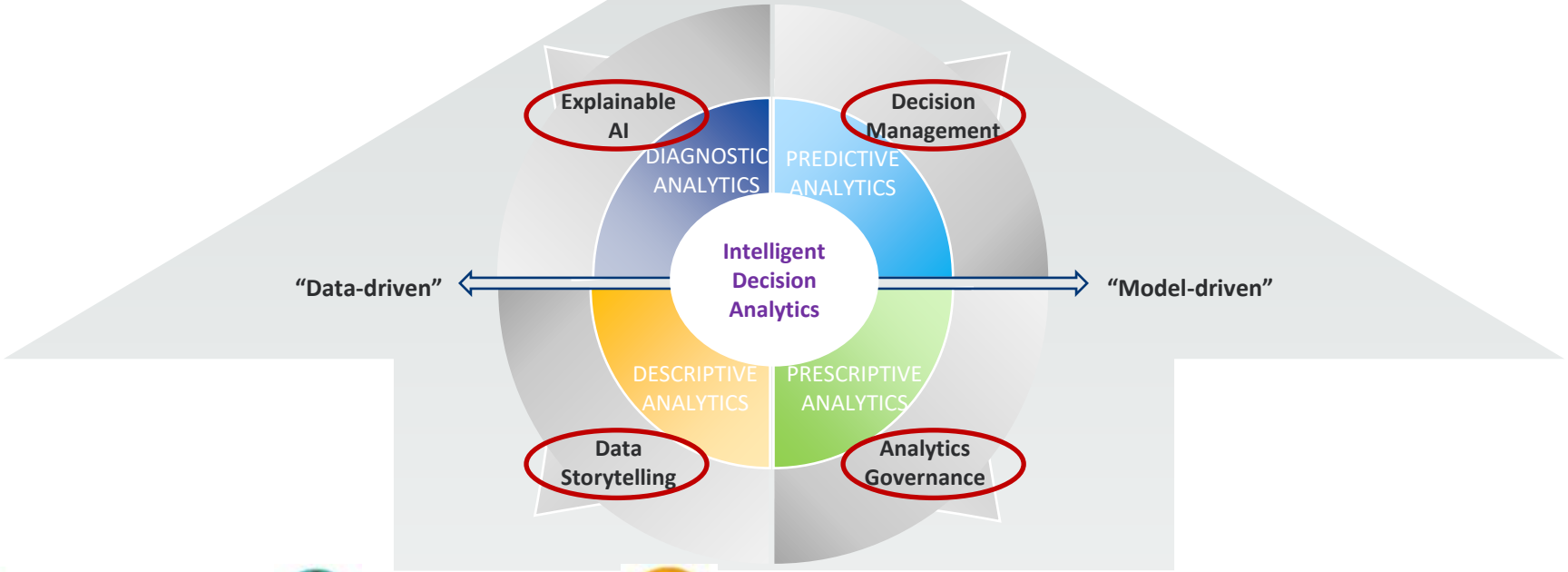
Who is a Data Scientist?



➢ The Insights Revolution?

Data Presenters

- User Interface / User Experience / User Engagement
- Investigation and Discovery



Insight Providers

- Semantic Models: Statistical
- Machine Learning: Algorithms
- Analytics Library: Development Environment for Analytics

Platform Owners

- Development Environment
- APIs for Connectivity and Devi

Data Producers / Data Aggregations / Data Custodians

- Data Normalization for Common Transmission
- Heterogeneous Data Collection from Disparate Devices
- Data Access / Data Control / Data Collection

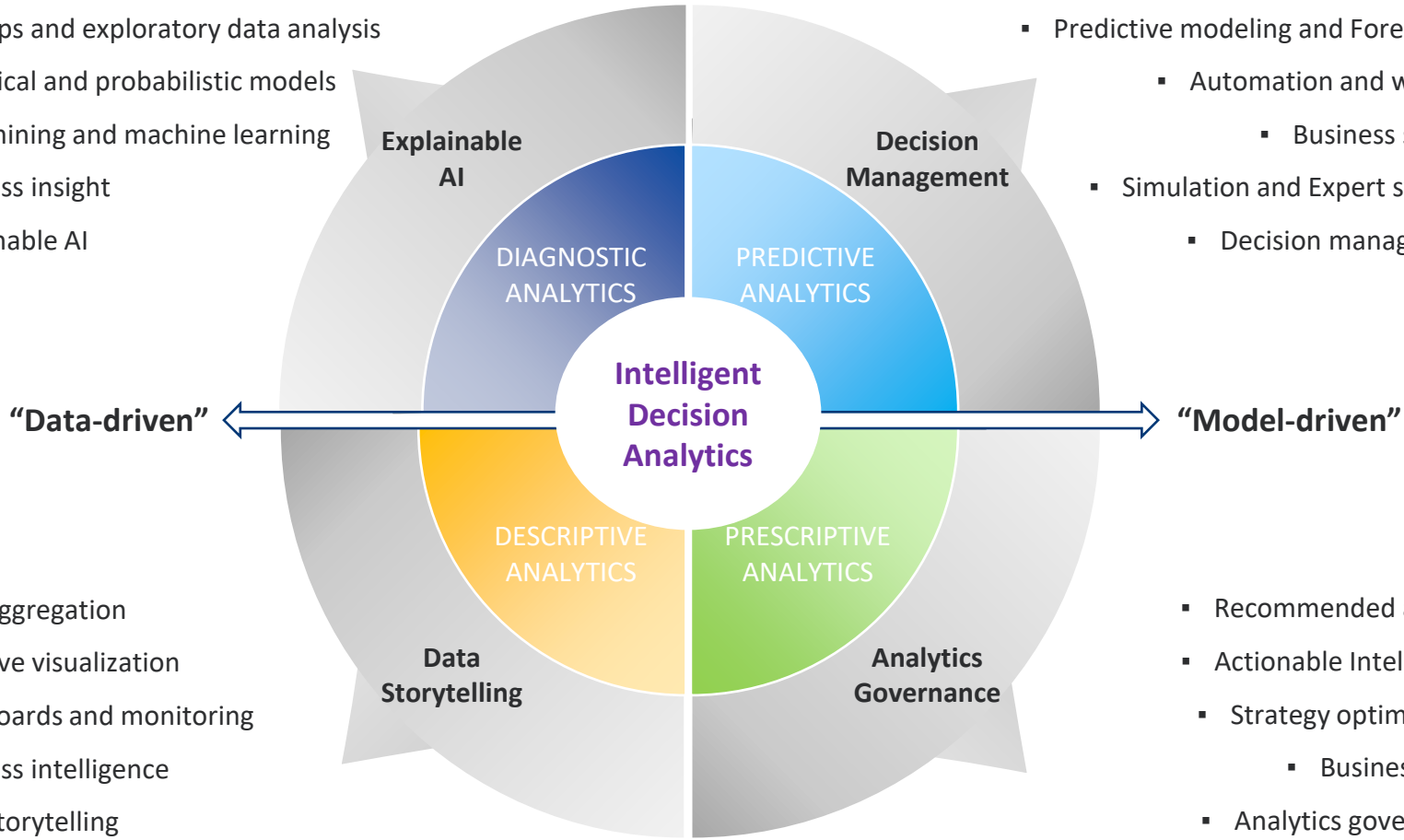
Who is a Data Scientist?

➢ Why DO it happen?

- DataOps and exploratory data analysis
- Statistical and probabilistic models
- Data mining and machine learning
- Business insight
- Explainable AI

➢ What COULD happen?

- Predictive modeling and Forecasting
- Automation and warning
- Business scoring
- Simulation and Expert systems
- Decision management



- Data aggregation
- Effective visualization
- Dashboards and monitoring
- Business intelligence
- Data storytelling

➢ What HAS happened?

- Recommended actions
- Actionable Intelligence
- Strategy optimization
- Business rules
- Analytics governance

➢ What SHOULD happened?

THANK YOU