



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!

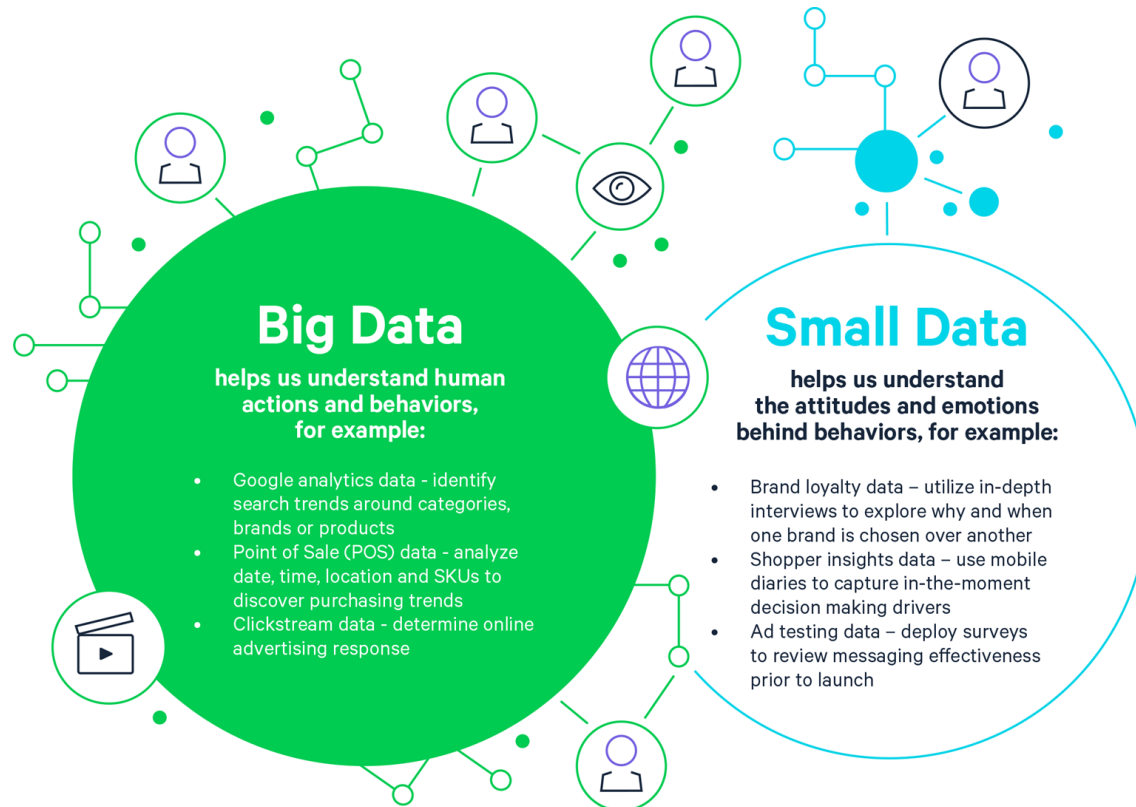
- I. Properties of Data: What is a Big Data?
- II. What is Data Science?
- III. Data Science Process
- IV. Why Data Science is in a sudden boom?
- V. Who is a Data Scientist?

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.

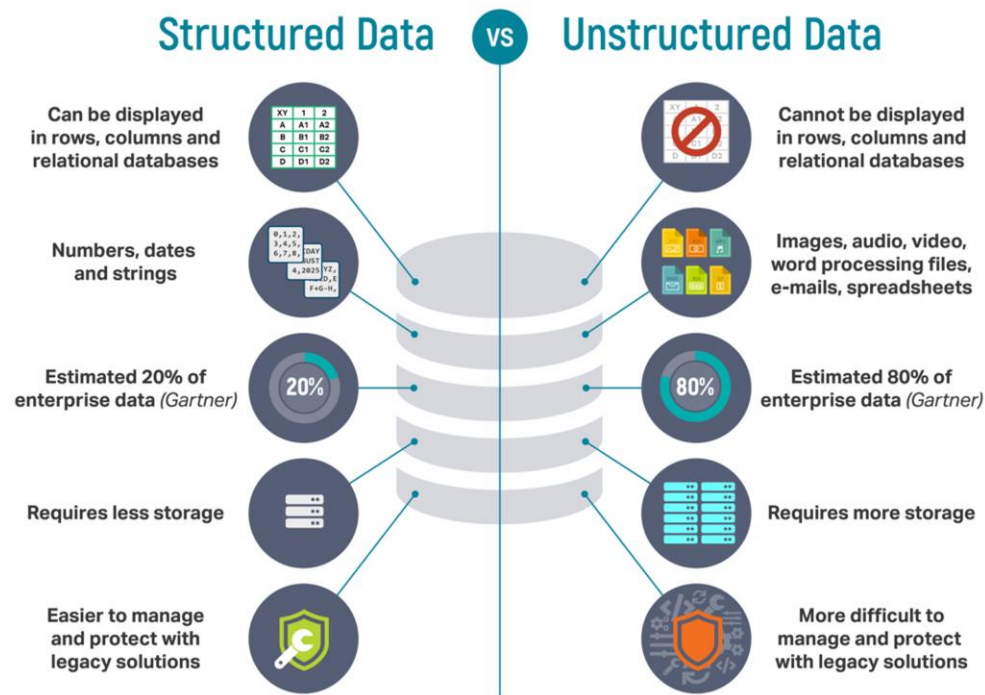
=> A Powerful Equation: Real Human Insight = Big Data + Small Data



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.



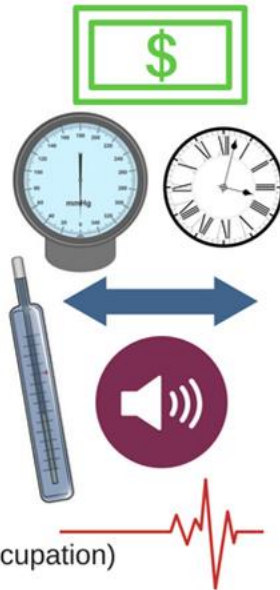
Properties of Data: What is a Big Data?

➤ Quantitative vs. Qualitative Data

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

Quantitative Data

- money
- time
- speed
- movement
- height
- length
- area
- volume
- weight
- temperature
- humidity
- pressure
- sound level
- categories
(age, gender, occupation)
- positioning
- status



Qualitative Data

- verbal and written feedback
 - first-hand (direct experience)
 - second-hand (telling someone else)
 - third-hand (outside story-teller)
- visual images, drawings, or models
- experiential sensations
- descriptions of
 - colors
 - textures
 - smells
 - tastes
 - appearance
 - beauty
 - feelings
 - intuition
 - sensations
 - choices
 - values
 - beliefs



LaConte Consulting ©2018 <http://laconteconsulting.com>

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.

The image shows a composite of two screenshots from the Kaggle website. The top screenshot is the 'Welcome to Kaggle Competitions' landing page, which features three main options: 'New to Data Science?' (with a tutorial on Titanic), 'Build a Model' (with a guide on using data and tools), and 'InClass Prediction Competition' (highlighted). The bottom screenshot is the 'Housing Prices Competition for Kaggle Learn Users' page, showing the competition title, a description, the number of teams (13,432), and the duration (4 months). The page includes a navigation bar with links for Overview, Data, Notebooks, Discussion, Leaderboard, Rules, and a Submit Predictions button. The Overview section is expanded, showing a description of the competition and a frequently asked question about getting started.

Welcome to Kaggle Competitions
Challenge yourself with real-world machine learning problems

New to Data Science?
Get started with a tutorial on our most popular competition for beginners, [Titanic: Machine Learning from Disaster](#).

Build a Model
Get the data & use whatever tools or methods you prefer to make predictions.

InClass Prediction Competition

Housing Prices Competition for Kaggle Learn Users
Apply what you learned in the Machine Learning course on Kaggle Learn alongside others in the course.

13,432 teams · 4 months to go

[Overview](#) [Data](#) [Notebooks](#) [Discussion](#) [Leaderboard](#) [Rules](#) [Submit Predictions](#)

Overview

Description

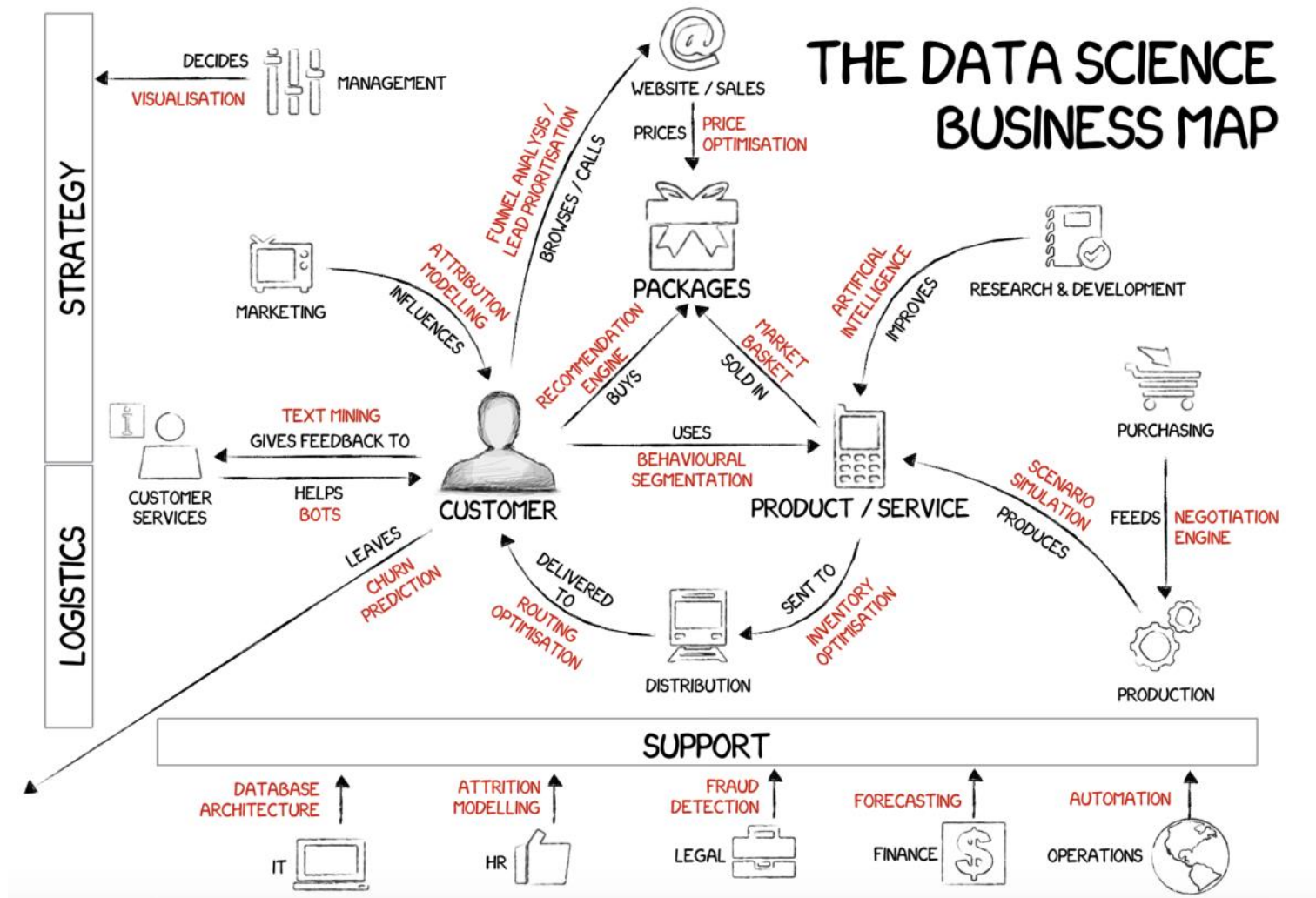
Evaluation

[Frequently Asked Questions](#)

What is a Getting Started competition?
Getting Started competitions were created by Kaggle data scientists for people who have little to no machine learning background. They are a great place to begin if you are new to data science or just finished a MOOC and want to get involved in Kaggle.

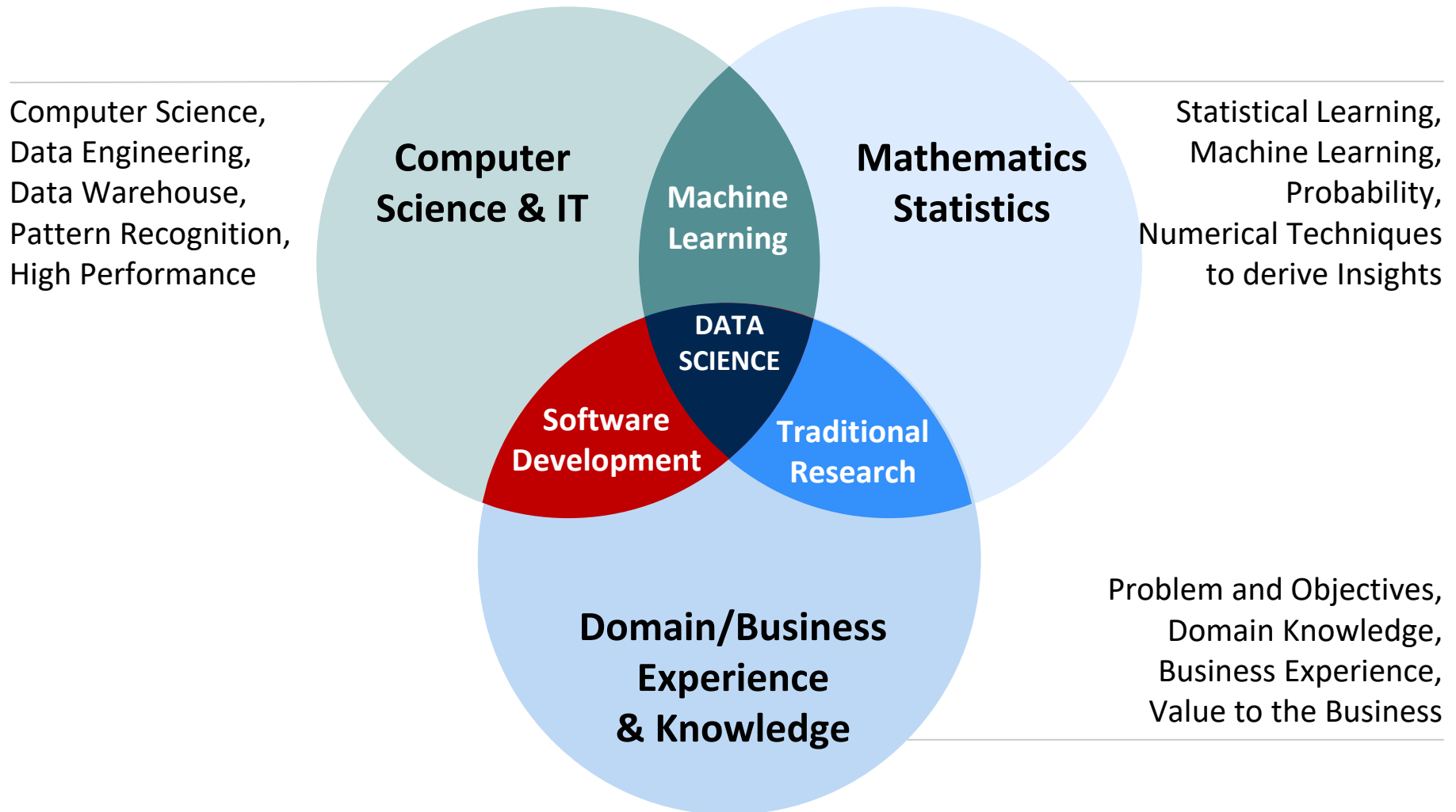
What is Data Science?

➤ The Real World Data Science



What is Data Science?

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



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.

MASTER'S IN

COMPUTER SCIENCE VS. DATA SCIENCE

TWO PATHWAYS TO NEXT-GENERATION CAREERS

With today's high demand for software solutions and data science, professionals are pursuing graduate education to advance their knowledge and skills. But some have the dilemma of deciding which program is the best fit for their career goals. While the programs share some similarities, there are also key differences to consider. Explore the benefits of Levens University's online **Master of Science in Computer Science (MSCS)** and **Master of Science in Data Science (MSDS)** programs.

VS.

The infographic features two stylized human figures, one red and one blue, representing the two fields. The red figure is associated with Computer Science and the blue figure with Data Science.

THE BIG PICTURE

Shaping the science of technology

WHAT'S IT ABOUT

Discovering the meaning within big data

Advanced computing, including in-depth experience in developing enterprise-scale applications, database systems, security solutions and automated systems.

WHAT YOU'LL LEARN

The mathematics and computer science of analyzing large data collections using data mining, data visualization, predictive analysis and efficient data management.

Developing next-generation technology in software, cyber security and intelligent systems.

WHAT'S IT FOR

Honing subject-matter expertise and the skills needed to clarify the meaning of large data sets, which can help lead to better organization decision-making.

Computer Science is the prime mover of today's technological innovations.

WHY IT MATTERS

Data is important to any company, and the sheer quantity of it requires experts who can make sense of it.

CAREERS

SELECTED CAREERS

- Application Developer
- Computer Engineer
- Computer Programmer
- Database Architect
- Database Developer
- Data Center Manager
- IT Engineer
- Mobile Specialist
- Network Administrator
- Network Architect
- Network Engineer
- Software Engineer
- Systems Architect
- Systems Programmer
- Web Developer

VS.

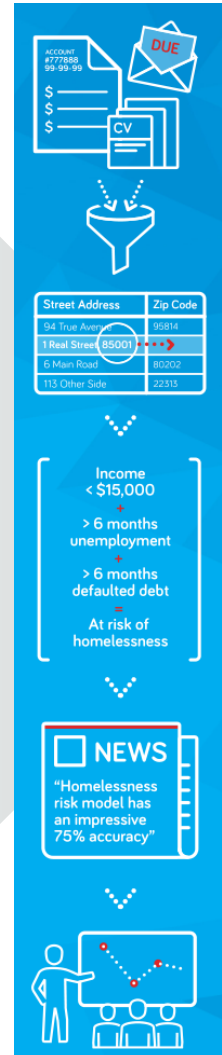
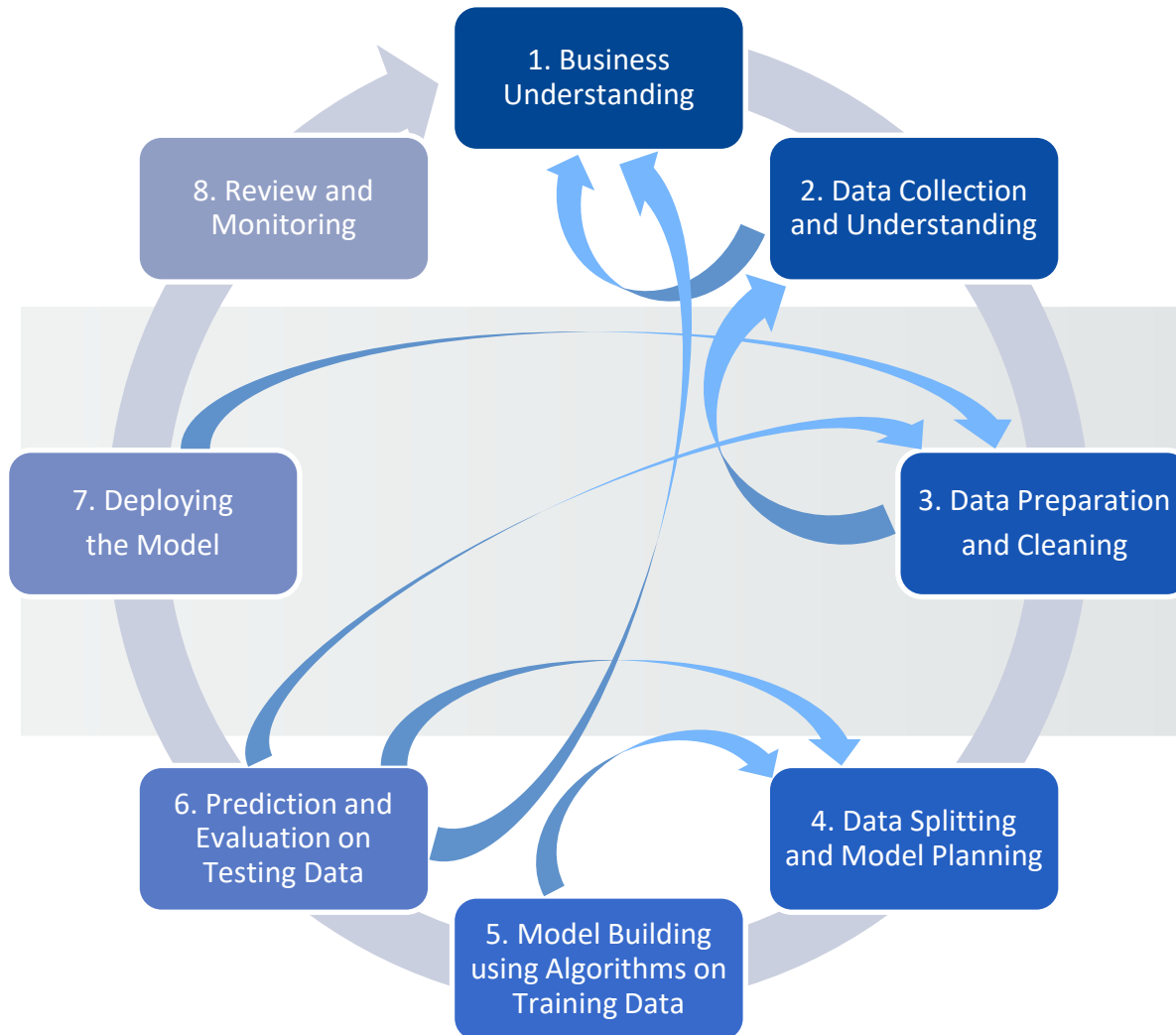
- Business Intelligence Manager
- Business Systems Analyst
- Clinical Researcher
- Computational Biologist
- Data Analyst
- Database Developer
- Data Scientist
- Data Strategist
- Health Informatics Analyst
- Financial Analyst
- Marketing Analyst
- Predictive Modeler
- Researcher
- Research Analyst
- Risk Analyst
- Statistician

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

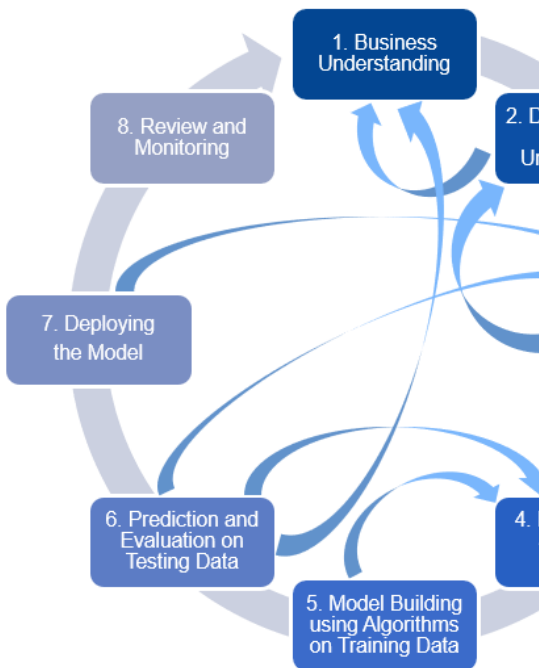
➤ Getting from Raw Data to Outcomes



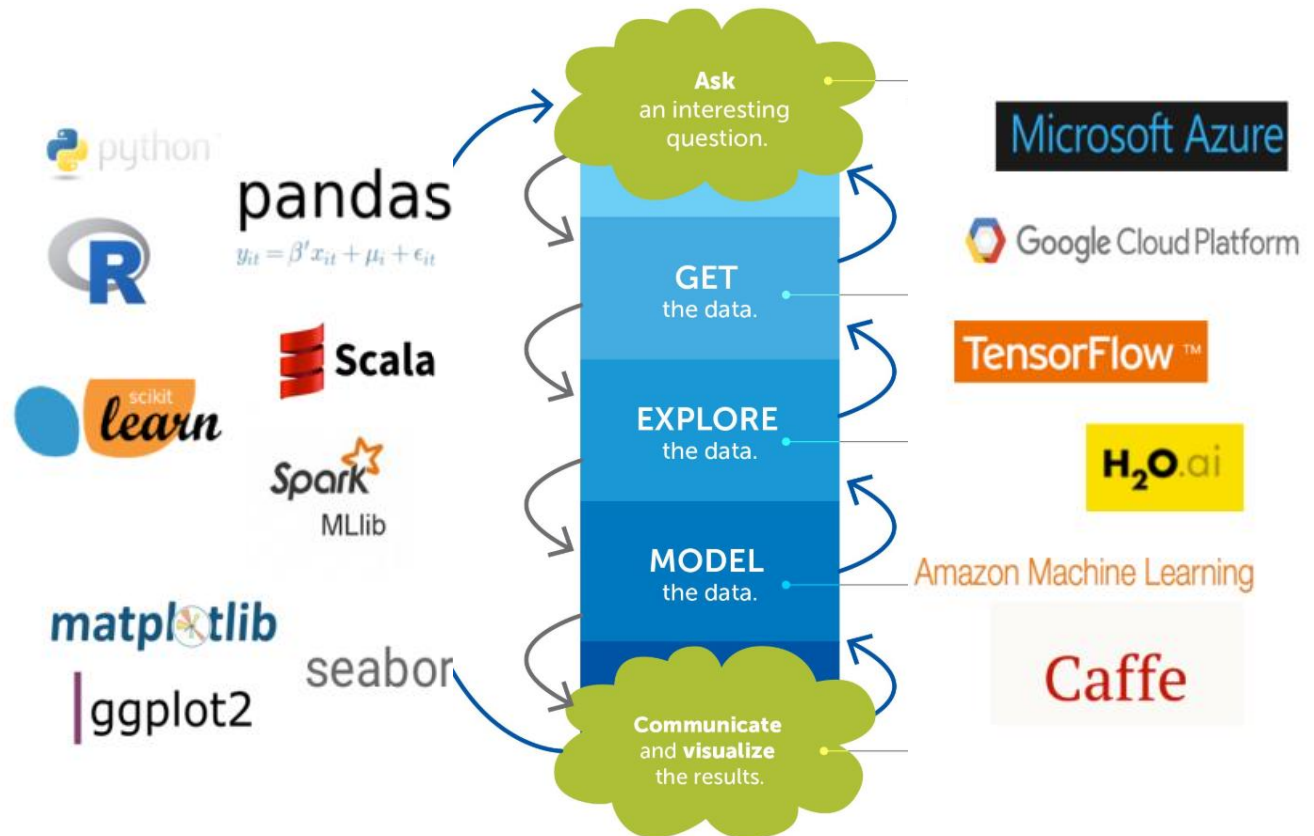
Data Science Process

➤ Getting from Standard Framework to Data Science

“Standard Framework for Data Mining”



“The Data Science Workflow”

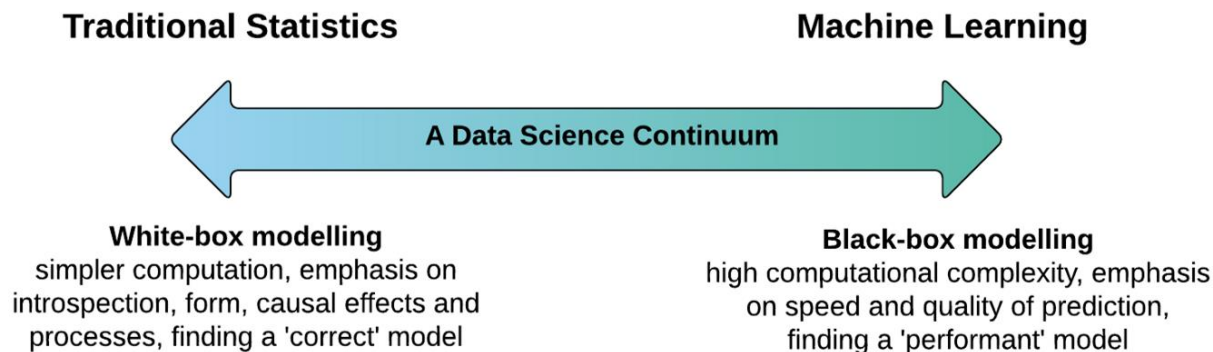


WE ALL CAN BE DATA SCIENTISTS NOW!

- 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?



- **Bayes Theorem:**

Thomas Bayes mid 1700's

- **Regression:**

Legendre, Gauss and Galton early 1800's

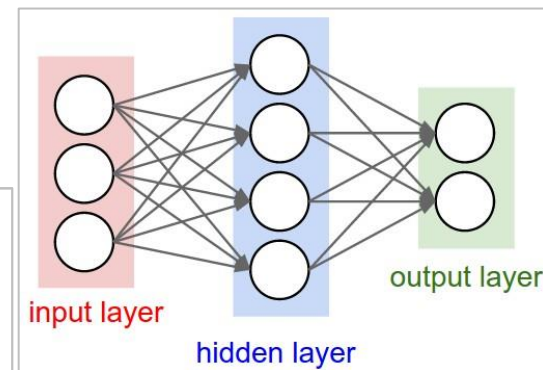
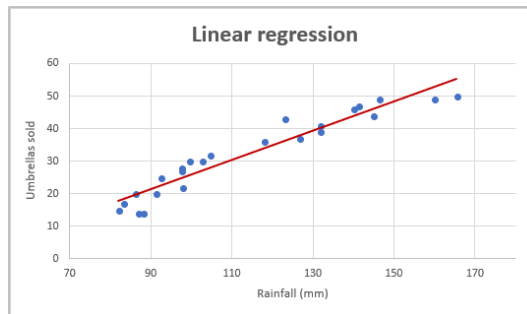
- **Neural Networks:**

McCulloch and Pitts early 1940s

Diagram illustrating Bayes Theorem with labels:

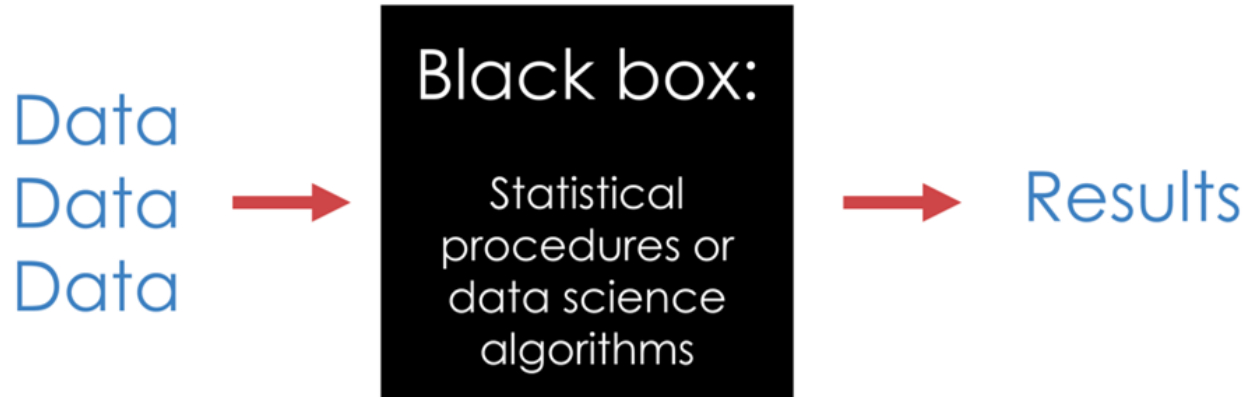
- LIKELIHOOD**: The probability of "B" being True, given "A" is True.
- PRIOR**: The probability "A" being True. This is the knowledge.
- POSTERIOR**: The probability of "A" being True, given "B" is True.
- MARGINALIZATION**: The probability "B" being True.

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)}$$

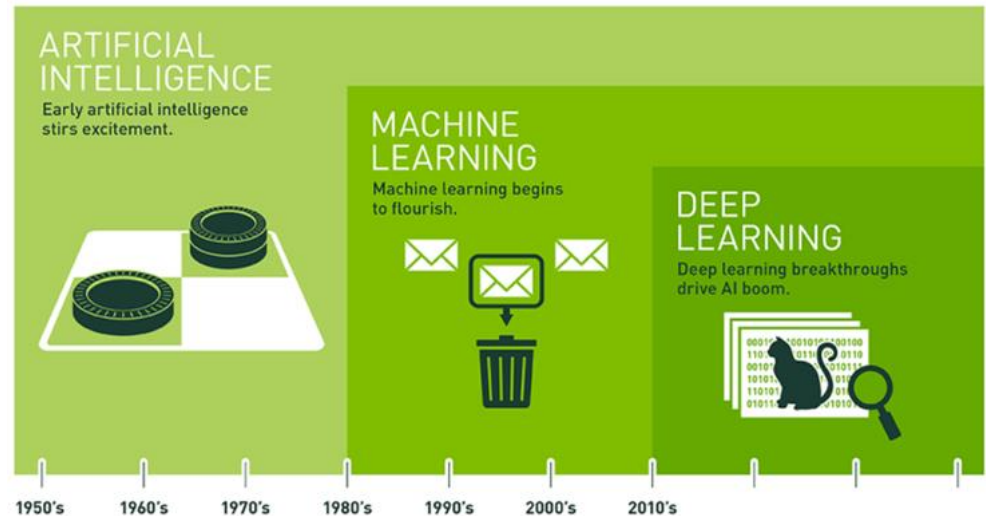


Why Data Science is in a sudden boom?

➤ 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

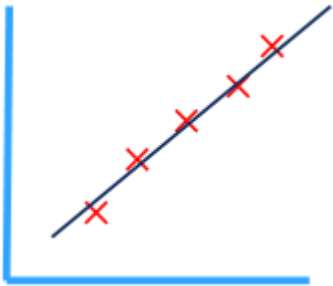


Since an early flush of optimism in the 1950s, smaller subsets of artificial intelligence – first machine learning, then deep learning, a subset of machine learning – have created ever larger disruptions.

Why Data Science is in a sudden boom?

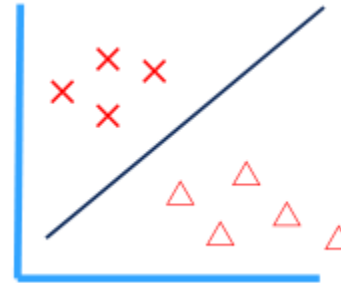
➤ Solution directions to the black box problem

- How much is the stock of Samsung Electronics tomorrow?



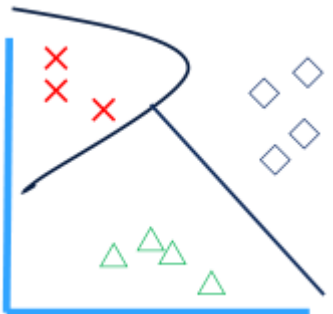
Regression – Looking for a statistical relationship across variables that may give us an estimate of a particular outcome. (Supervised)

- Will Samsung Electronics' stocks rise or fall tomorrow?



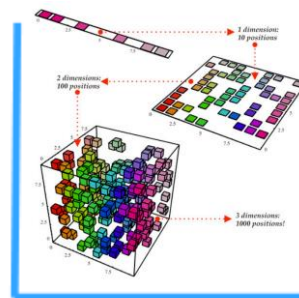
Classification – Similar to regression but looking for separations in the data given predefined classes. (Supervised)

- 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)

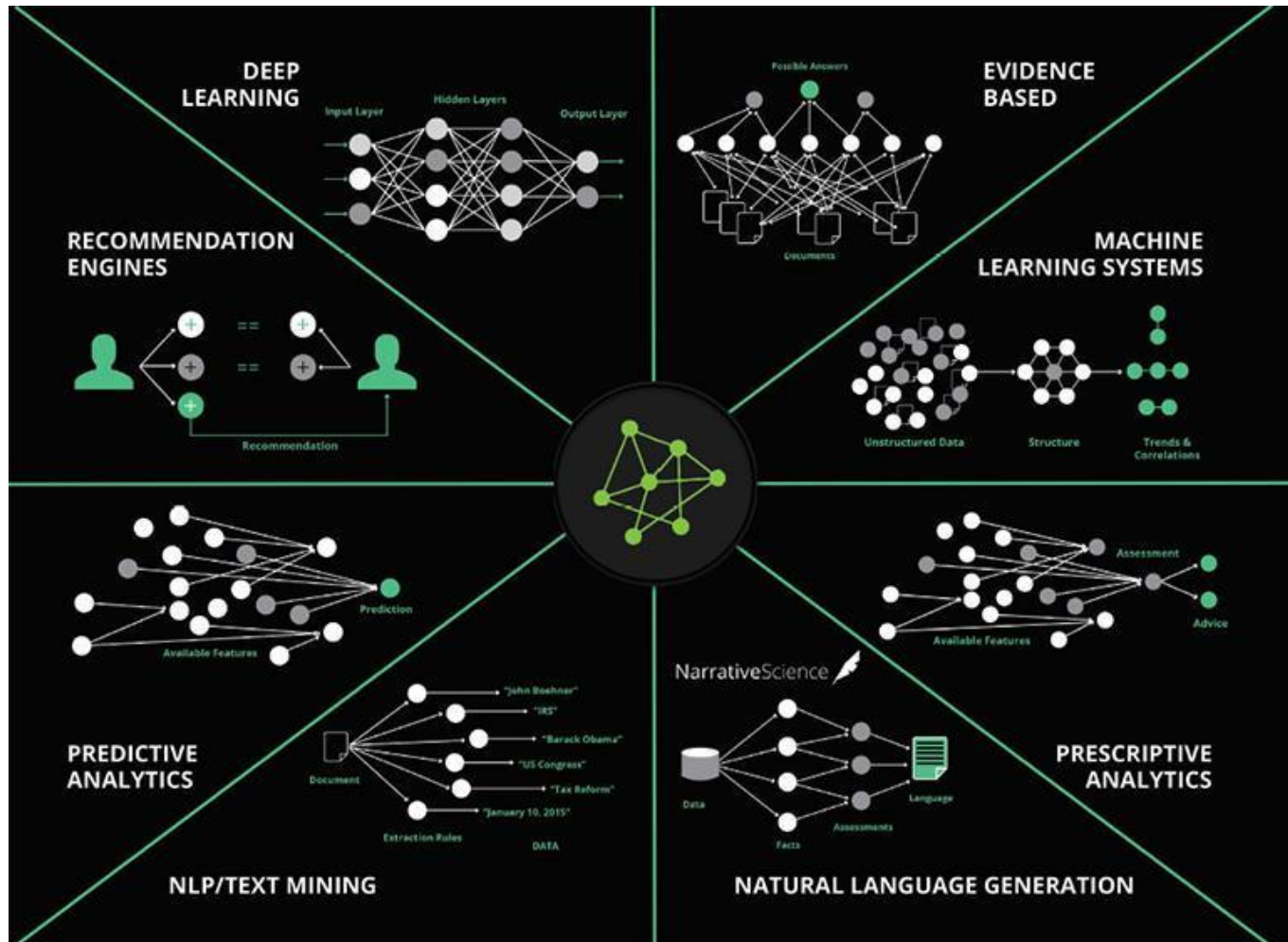
- What are the representatives among all stocks in the KOSPI?



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)

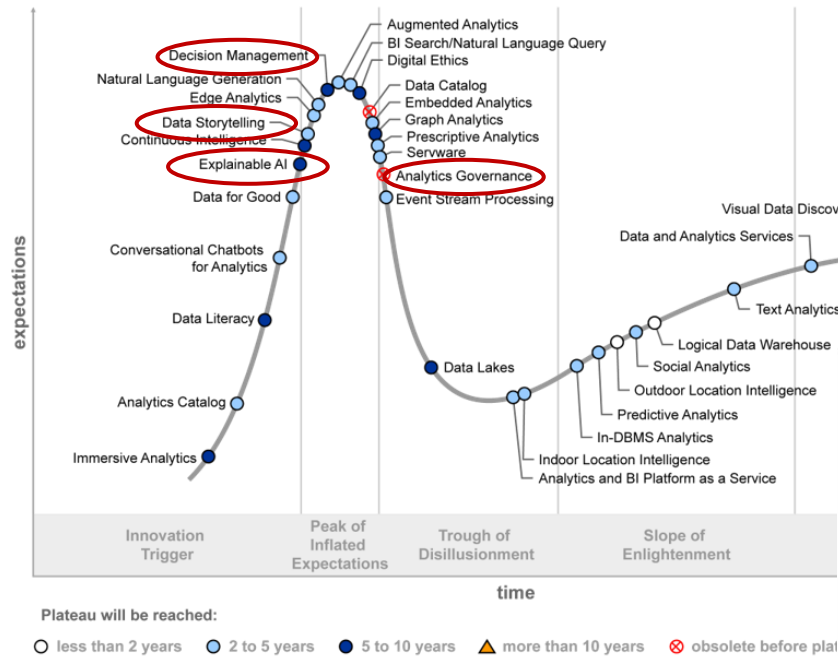
Why Data Science is in a sudden boom?

➤ Different types of Machine Learning algorithms explained

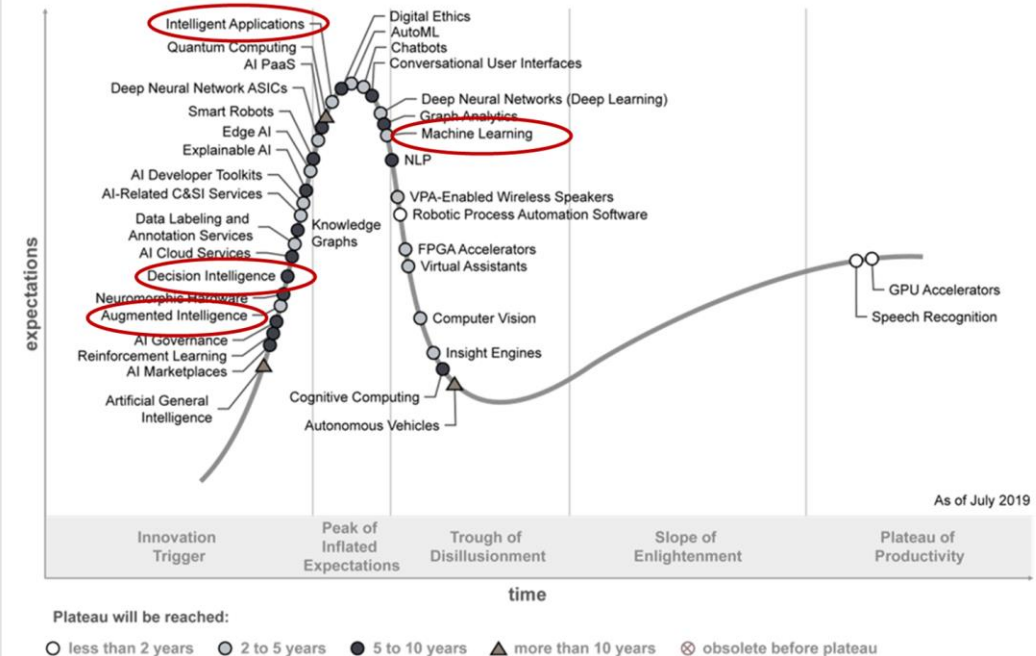


Global “Technology” Trends

Hype Cycle for Analytics and Business Intelligence, 2019



Hype Cycle for Artificial Intelligence, 2019



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?



Data Scientist, the sexiest job of 21st century requires a mixture of multidisciplinary skills ranging from an intersection of mathematics, statistics, computer science, communication and business. Finding a data scientist is hard. Finding people who understand who a data scientist is, is equally hard. So here is a little cheat sheet on who the modern data scientist really is.



MATH & STATISTICS <ul style="list-style-type: none">☆ Machine learning☆ Statistical modeling☆ Experiment design☆ Bayesian inference☆ Supervised learning: decision trees, random forests, logistic regression☆ Unsupervised learning: clustering, dimensionality reduction☆ Optimization: gradient descent and variants	PROGRAMMING & DATABASE <ul style="list-style-type: none">☆ Computer science fundamentals☆ Scripting language e.g. Python☆ Statistical computing package e.g. R☆ Databases SQL and NoSQL☆ Relational algebra☆ Parallel databases and parallel query processing☆ MapReduce concepts☆ Hadoop and Hive/Pig☆ Custom reducers☆ Experience with xaaS like AWS
DOMAIN KNOWLEDGE & SOFT SKILLS <ul style="list-style-type: none">☆ Passionate about the business☆ Curious about data☆ Influence without authority☆ Hacker mindset☆ Problem solver☆ Strategic, proactive, creative, innovative and collaborative	COMMUNICATION & VISUALIZATION <ul style="list-style-type: none">☆ Able to engage with senior management☆ Story telling skills☆ Translate data-driven insights into decisions and actions☆ Visual art design☆ R packages like ggplot or lattice☆ Knowledge of any of visualization tools e.g. Flare, D3.js, Tableau

MarketingDistillery.com is a group of practitioners in the area of e-commerce marketing. Our fields of expertise include marketing strategy and optimization, customer tracking and on-site analytics, predictive analytics and econometrics, data warehousing and big data systems, marketing channel insights in Paid Search, SEO, Social, CRM and brand.

Marketing
DISTILLERY

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, these 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)

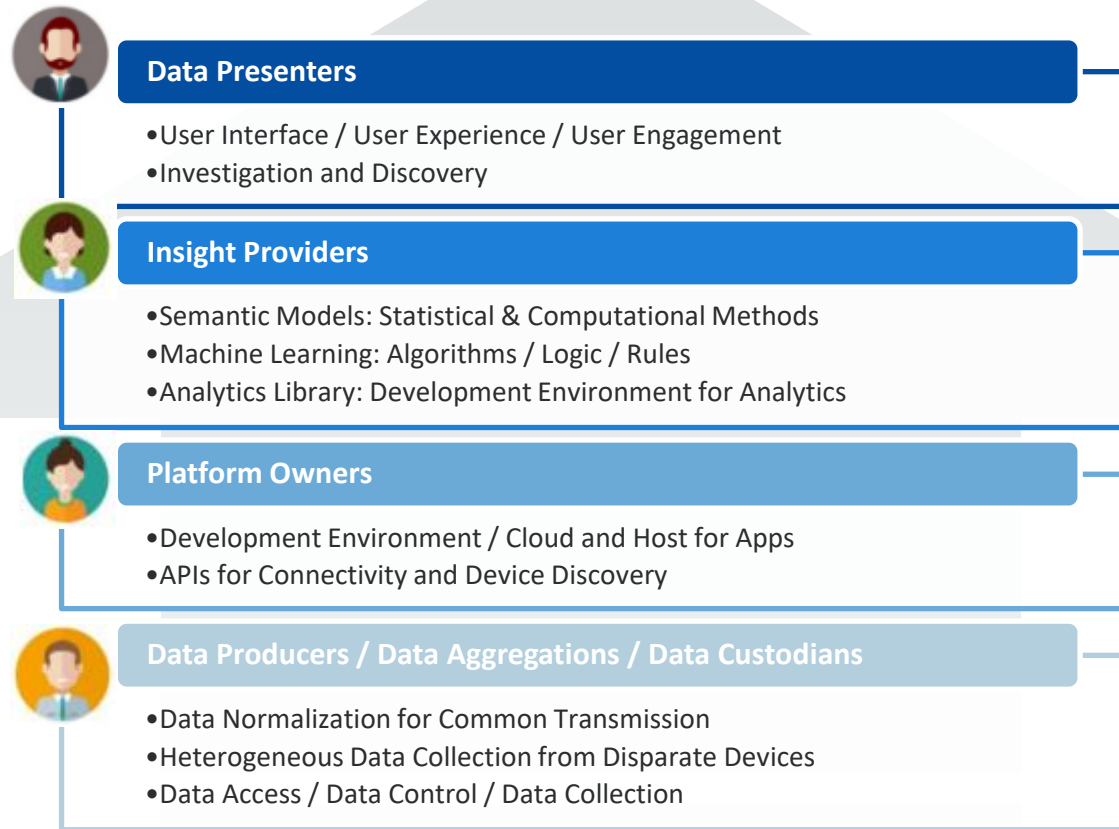
초급 중급 고급

Data Analysis (6)	Mathematics (6)	Statistics (6)	Artificial Intelligence (6)	Computation (10)	Professionalism (10)
<ul style="list-style-type: none"> 데이터사이언스 개론 데이터분석 언어 데이터모델링 데이터시각화 데이터분석 응용 데이터사이언스 응용 및 활용 	<ul style="list-style-type: none"> 미적분학 기초 선형대수 기초 수치해석 기초 다변량 미적분학 인공지능 수치해석 최적화 	<ul style="list-style-type: none"> 확률 및 통계 기초 실험계획법 통계적사고 및 알고리즘 다변량 자료분석 시계열분석 시뮬레이션 	<ul style="list-style-type: none"> 데이터마이닝 개론 기계학습 딥러닝 인공지능 자연어처리 시각처리 	<ul style="list-style-type: none"> 데이터베이스 시스템 자료구조 및 알고리즘 서버시스템 및 클라우드 빅데이터 분산처리 소프트웨어 공학 웹마이닝 미디어콘텐츠 처리 컴퓨터비전 정보보안 특론 빅데이터 플랫폼 특론 	<ul style="list-style-type: none"> 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?

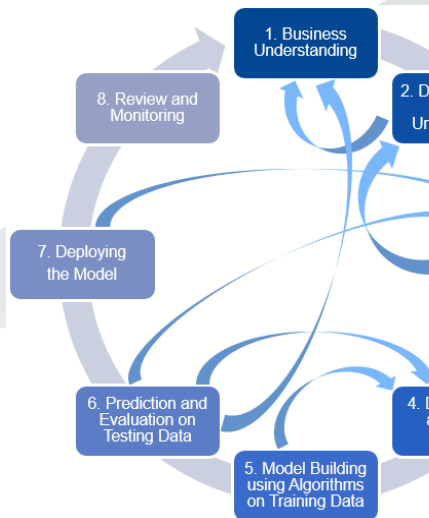
➤ The Insights Revolution?



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



Platform Owners

- Development Environment / APIs for Connectivity and Development

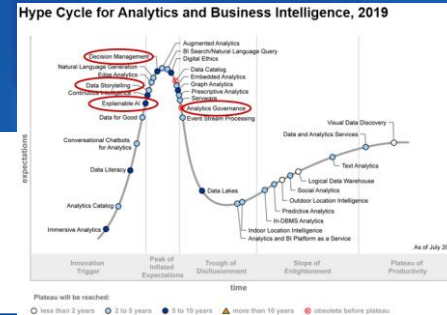


Data Producers / Data Aggregations / Data Custodians

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

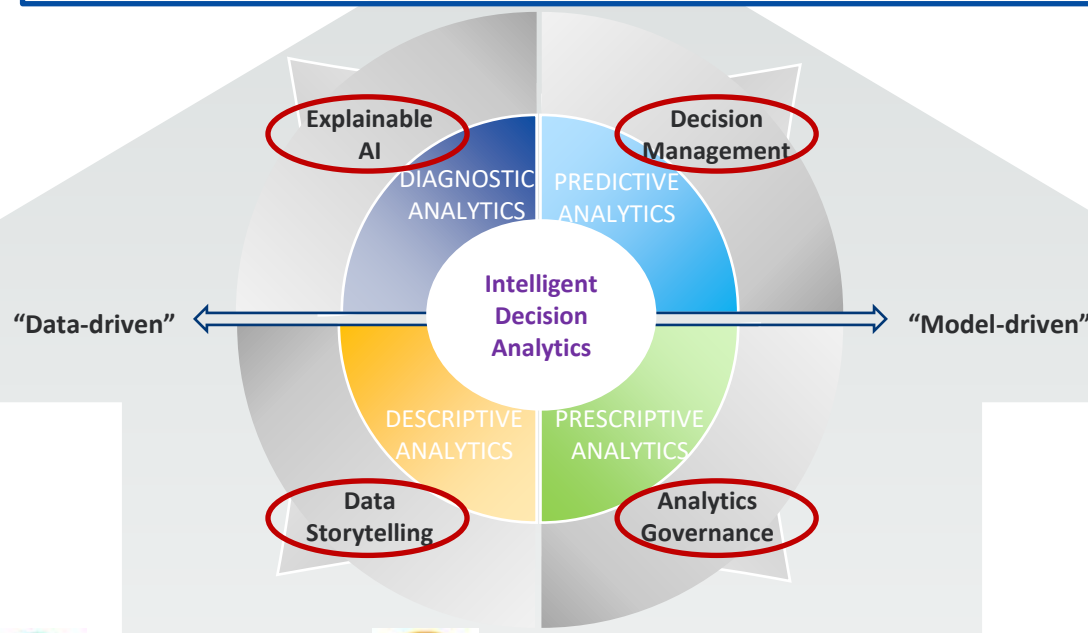
Who is a Data Scientist?

➤ The Insights Revolution?



Data Presenters

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



Insight Providers

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



Platform Owners

- Development Environment / APIs for Connectivity and Development



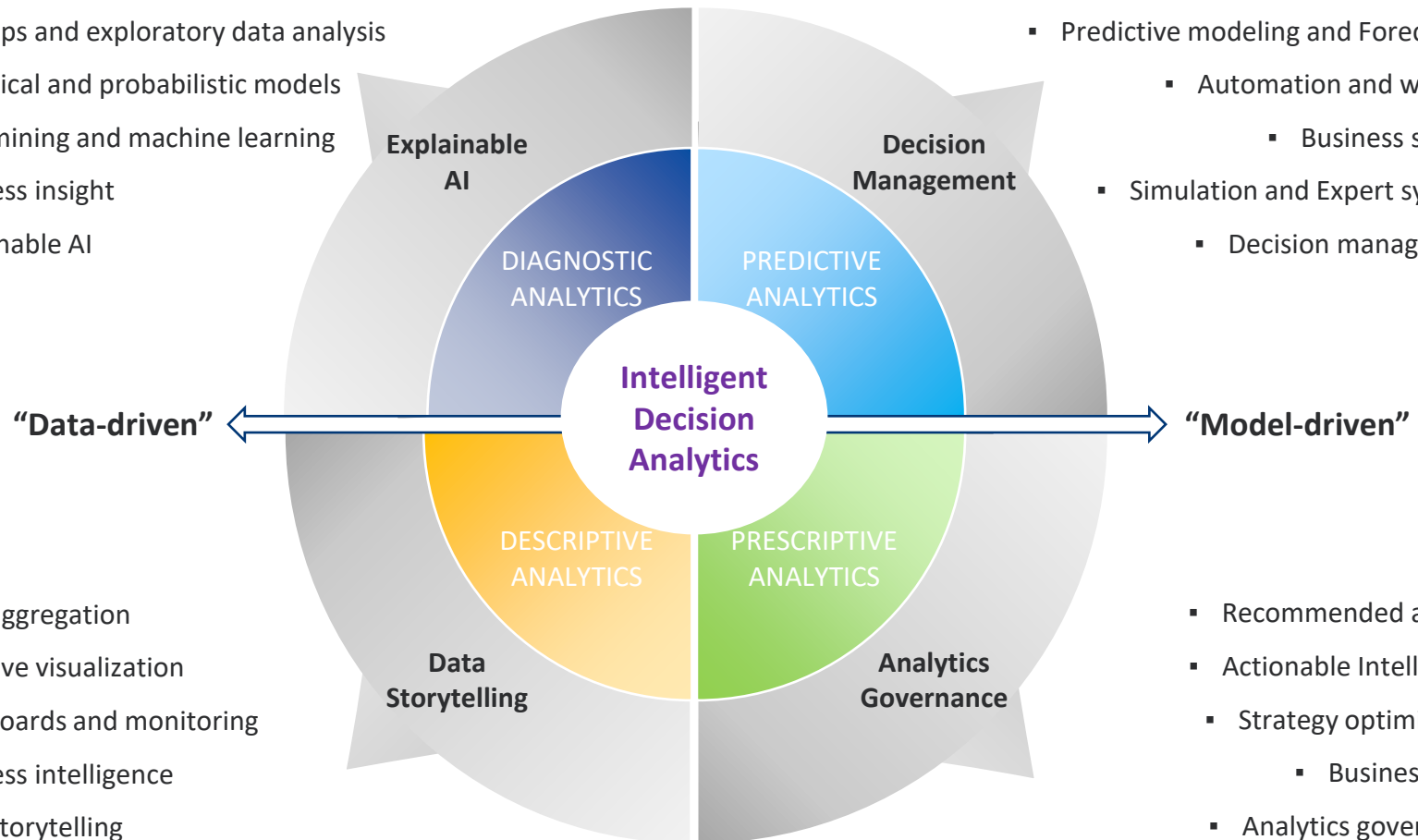
Data Producers / Data Aggregations / Data Custodians

- 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



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

➤ What COULD happen?

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

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

➤ What HAS happened?

➤ What SHOULD happened?

THANK YOU

