

Lecture 1: Introduction

Pilsung Kang

School of Industrial Management Engineering
Korea University

Course Information

- Lecturer

- ✓ Pilsung Kang, 801A, Innovation Hall
- ✓ pilsung_kang@korea.ac.kr

- Course homepage

- ✓ <http://github.com/pilsung-kang/multivariate-data-analysis>

The screenshot shows the GitHub profile page for the user 'pilsung-kang'. The profile picture is a portrait of a man with dark hair. The bio section includes the name 'Pilsung Kang' and the handle 'pilsung-kang'. It also mentions 'Associate Professor School of Industrial Management Engineering Korea University'. Below the bio, there are links to 'Korea University', 'Seoul, South Korea', and a website 'http://dsba.korea.ac.kr/'. The 'Overview' tab is selected, showing statistics: 13 repositories, 8 stars, 29 followers, and 3 following. The 'Pinned repositories' section contains five entries: 'text-mining', 'multivariate-data-analysis' (which is highlighted with a red box), 'machine-learning-with-R', 'R-for-Data-Analytics', and 'Business-Analytics'. The 'multivariate-data-analysis' repository is described as 'Multivariate data analysis @Korea University (Undergraduate)' and has 7 stars, 13 forks, and is in R. The 'Contributions' section at the bottom shows a heatmap of activity over the last year, with the highest density of contributions occurring in March, April, and May. A legend indicates that darker shades of green represent more contributions.

AGENDA

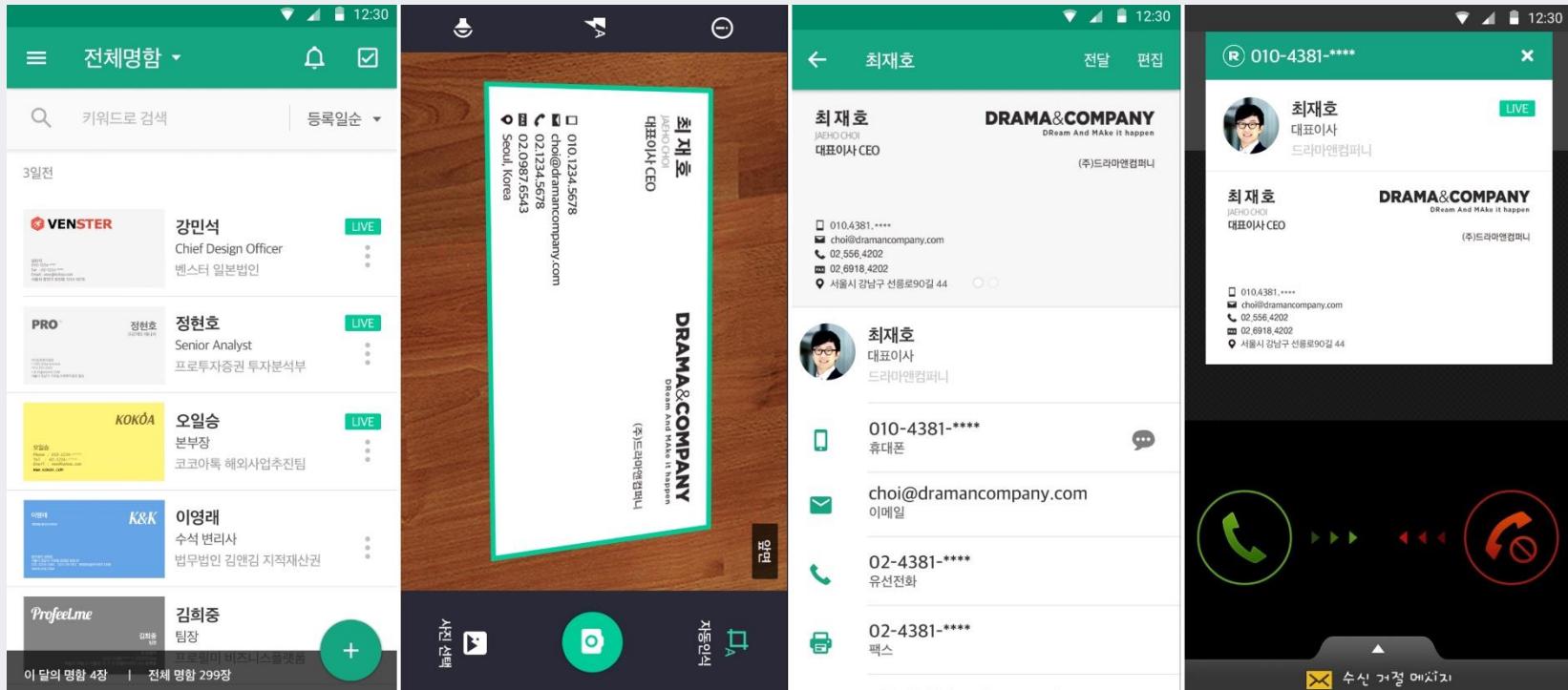
- 01** **Introduction to Data Science**
- 02** **Data Science Applications**
- 03** **Multivariate Data Analysis in Data Science**
- 04** **Data Science Procedure**

Future Human AI



Is Data a Tool or Purpose of Business?

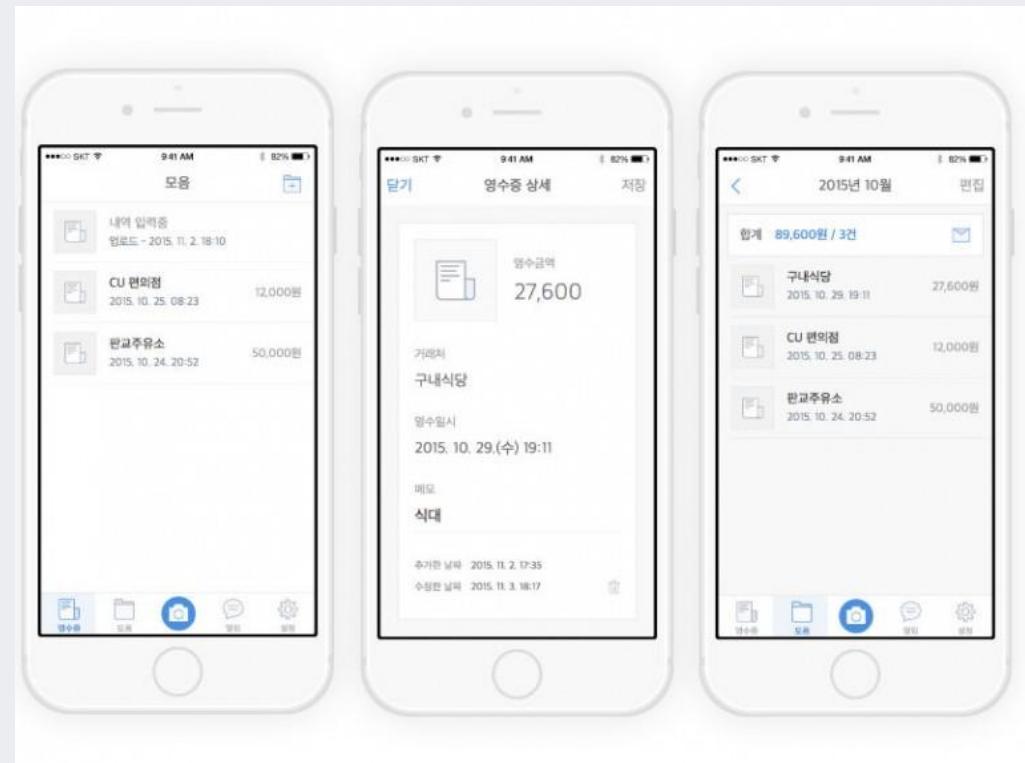
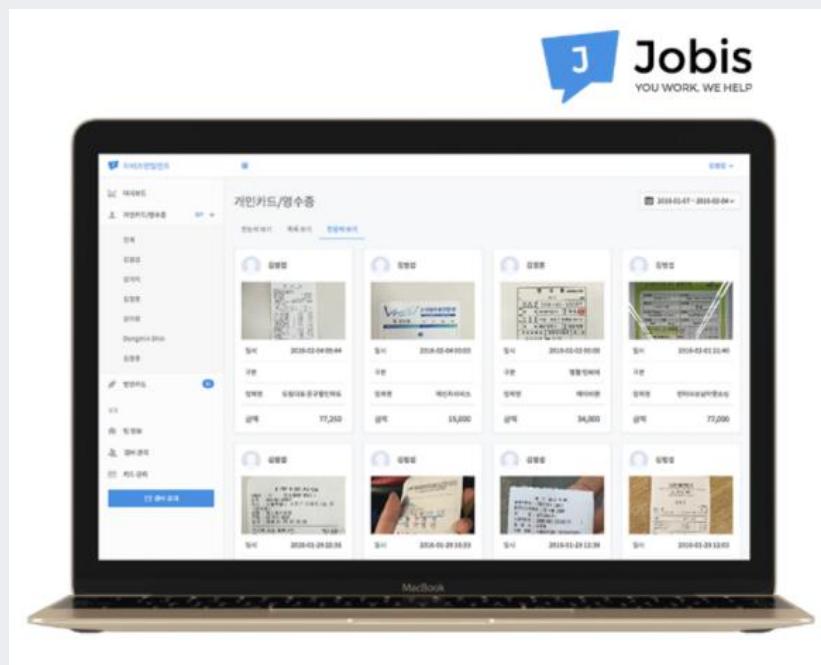
- Sell products or services to collect data?



✓ Acquired by Line+ and Naver on 2017-12-21

Is Data a Tool or Purpose of Business?

- Alpha Go? No! Alba-Go!



Data-driven Decision Making

- What we want to know



Data-driven Decision Making

- Descriptive vs. Predictive vs. Prescriptive Analytics

Understanding analytics			
	Descriptive	Predictive	Prescriptive
What the user needs to DO	What HAS happened?	What COULD happen?	What SHOULD happen?
What the user needs to DO	<ul style="list-style-type: none">Increase asset reliabilityReduce labor and inventory costs	<ul style="list-style-type: none">Predict infrastructure failuresForecast facilities space demands	<ul style="list-style-type: none">Increase asset utilizationOptimize resource schedules
What the user needs to KNOW	<ul style="list-style-type: none">The number and types of asset failuresWhy maintenance costs are highThe value of the materials inventory	<ul style="list-style-type: none">How to anticipate failures for specific asset typesWhen to consolidate underutilized facilitiesHow to determine costs to improve service levels	<ul style="list-style-type: none">How to increase asset productionWhere to optimally route service techniciansWhich strategic facilities plan provides the highest long-term utilization
How analytics gets ANSWERS	<ul style="list-style-type: none">Standard reporting - What happened?Query/drill down - Where exactly is the problem?Ad hoc reporting - How many, how often, where?	<ul style="list-style-type: none">Predictive modeling - What will happen next?Forecasting - What if these trends continue?Simulation - What could happen?Alerts - What actions are needed?	<ul style="list-style-type: none">Optimization - What is the best possible outcome?Random variable optimization - What is the best outcome given the variability in specified areas?
What makes this analysis POSSIBLE	<ul style="list-style-type: none">Alerts, reports, dashboards, business intelligence	<ul style="list-style-type: none">Predictive models, forecasts, statistical analysis, scoring	<ul style="list-style-type: none">Business rules, organization models, comparisons, optimization

Machine Learning

- Definition
 - ✓ A computer program is said to **learn** from experience E with respect to some class of tasks T and performance measure P, if its performance at task in T, as measured by P, improves with experience E,” – Mitchell (1997)

Supervised Learning

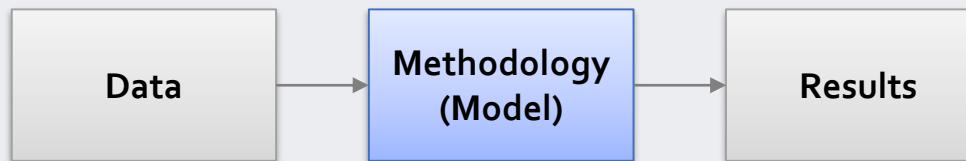
- Predict a single “target” or “outcome” variable
- Finds relations between X and Y.
- Train (learn) data where target value is known
- Score data where target value is not known

Unsupervised Learning

- Explores intrinsic characteristics.
- Estimates underlying distribution
- Segment data into meaningful groups or detect patterns
- There is no target (outcome) variable to predict or classify

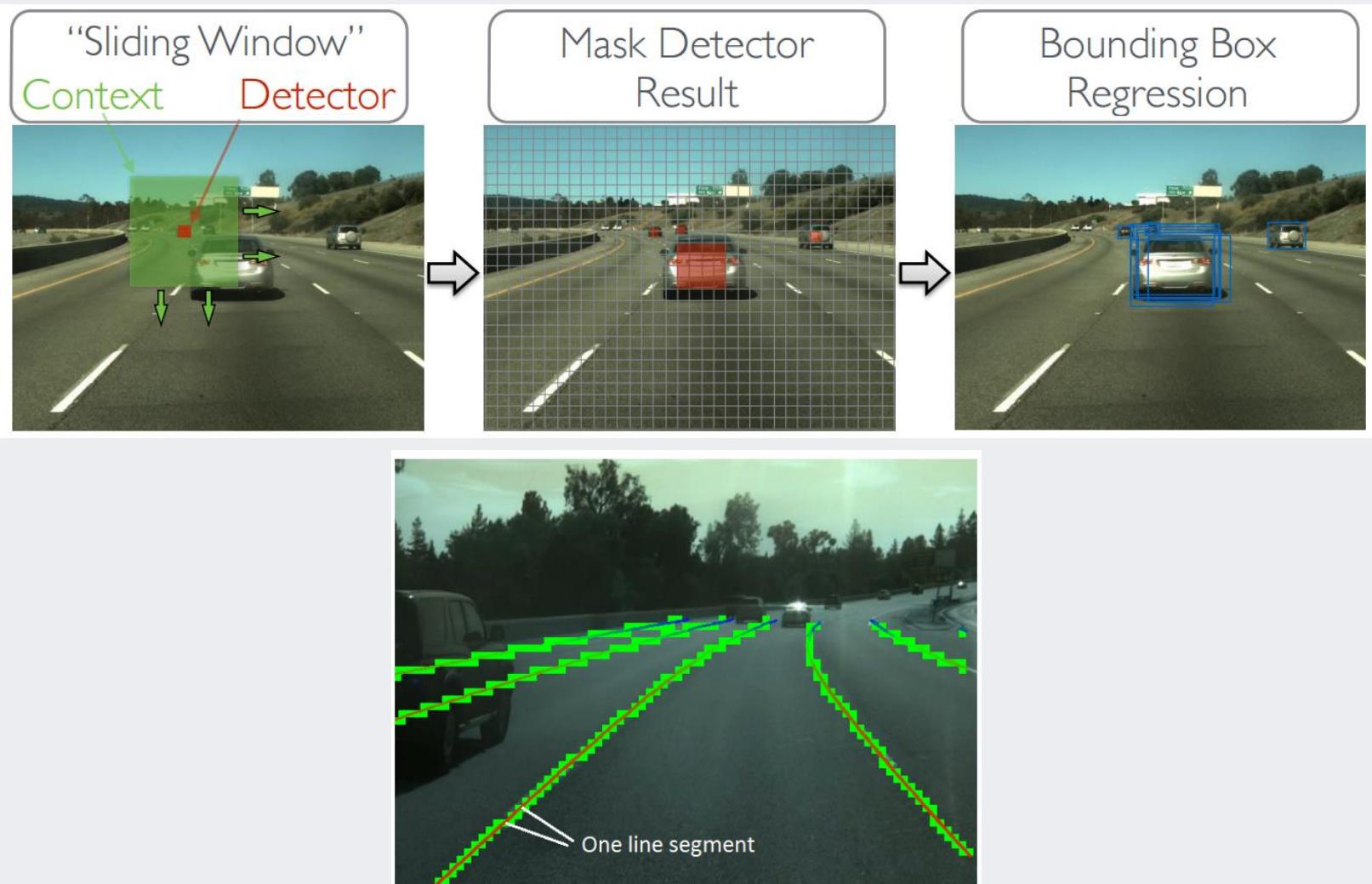
Machine Learning

- Definition
 - ✓ A computer program is said to **learn** from experience E with respect to some class of tasks T and performance measure P, if its performance at task in T, as measured by P, improves with experience E,” – Mitchell (1997)



Machine Learning

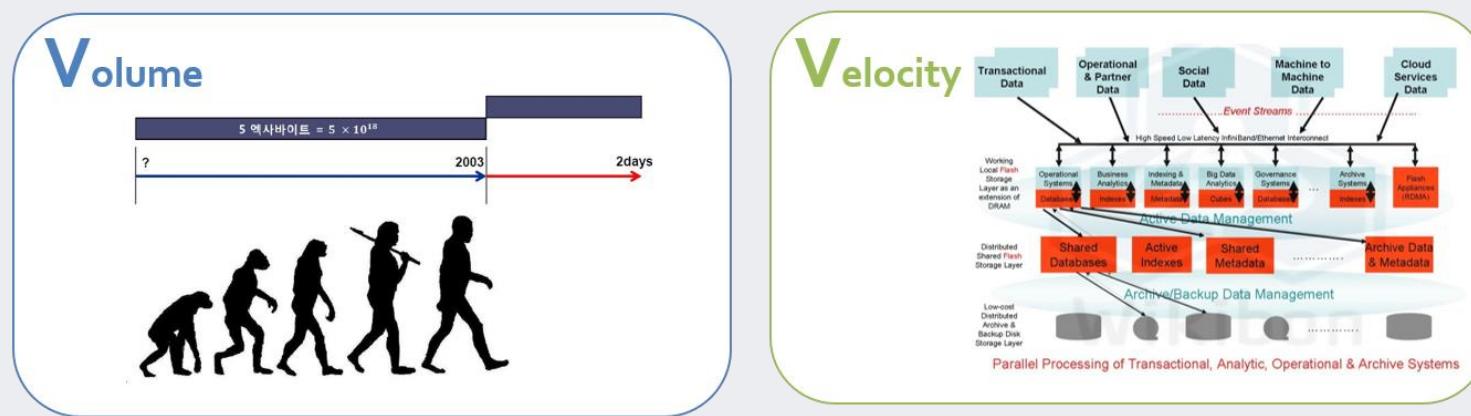
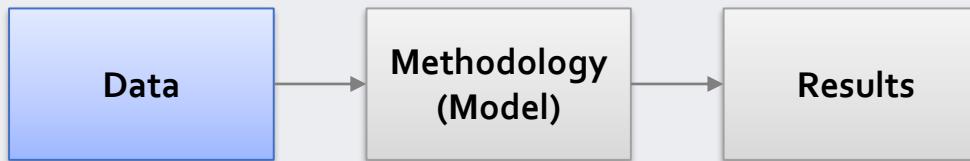
- Machine learning models in Self-driving cars



Big Data

- 4Vs in Big Data

✓ Volume, Velocity, Variety, Value



자료: McKinsey (2011.05)

Big Data

- Big Data itself is valuable without any complicated analytics methods

Type	Institution	Forecast
Economic Feasibility in Industries	Economist (2010)	<input type="checkbox"/> “ Data are becoming the new raw material of business : an economic input almost on a par with capital and labour.”
	Gartner (2011)	<input type="checkbox"/> “Intelligence about Information is the Oil of the 21st Century .” Future competitive advantage depends on data. <input type="checkbox"/> Winning organizations understand the stage of the data economy and overcome information silos through effective information sharing.
	McKinsey (2011)	<input type="checkbox"/> Big Data is the next frontier for innovation, competitiveness and productivity <input type="checkbox"/> Big Data will create values worth more than \$600 billion in 5 areas including medicine and public administration
National Competitiveness	US PCAST	<input type="checkbox"/> Advisors emphasize that US government organizations should focus on the strategy for transformation of data into knowledge, and of knowledge into action .
	Singapore	<input type="checkbox"/> Singapore looks to evaluate threatening risks and detect environmental changes based on data

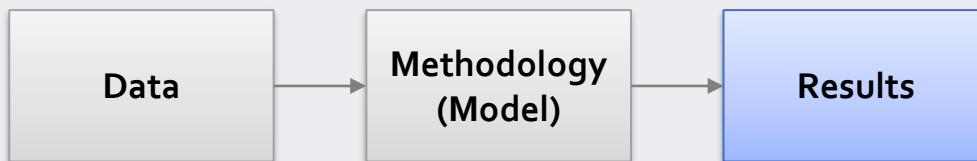
Big Data

- Case Study: Seoul



Data Mining

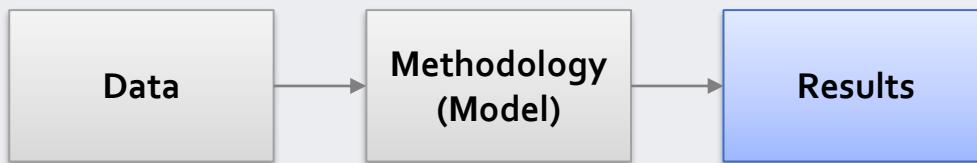
- Definitions



- ✓ Extracting useful information from large datasets. (Hand et al., 2001)
- ✓ The process of exploration and analysis, by automatic or semi-automatic means, of large quantities of data in order to discover meaningful patterns and rules. (Berry and Linoff, 1997, 2000)
- ✓ The process of discovering meaningful new correlations, patterns and trends by sifting through large amount data stored in repositories, using pattern recognition technologies as well as statistical and mathematical techniques. (Gartner Group, 2004)

Data Mining

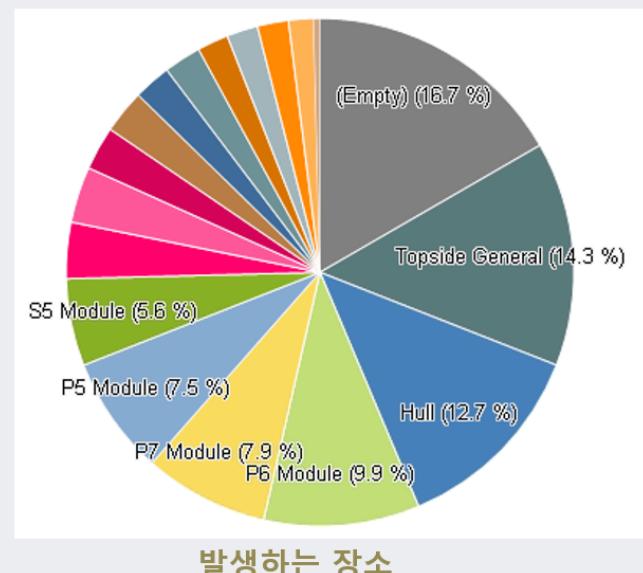
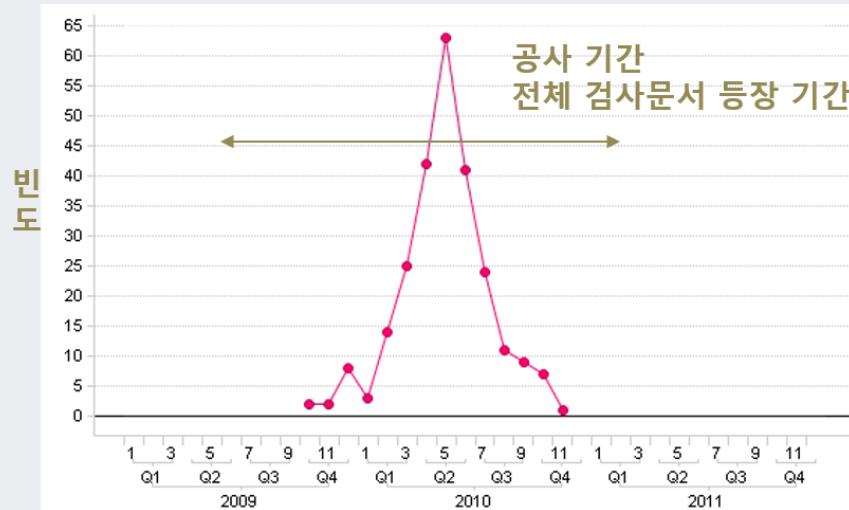
- Definitions



“파이프(pipe)가 흔들리니(shake), 지주(support)를 추가(add)하라”

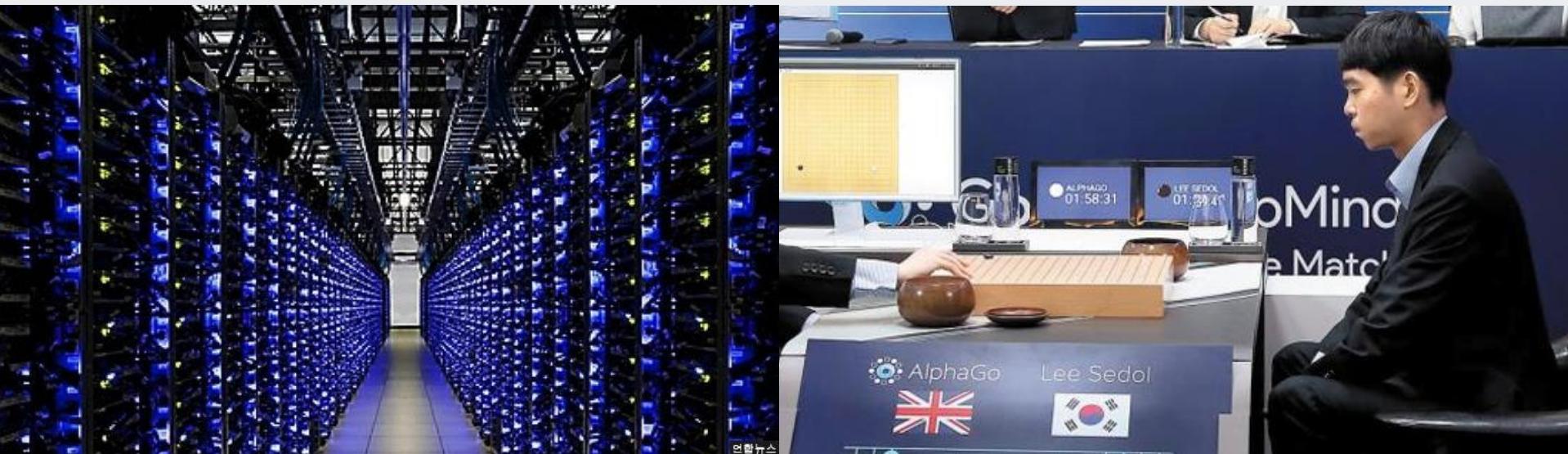
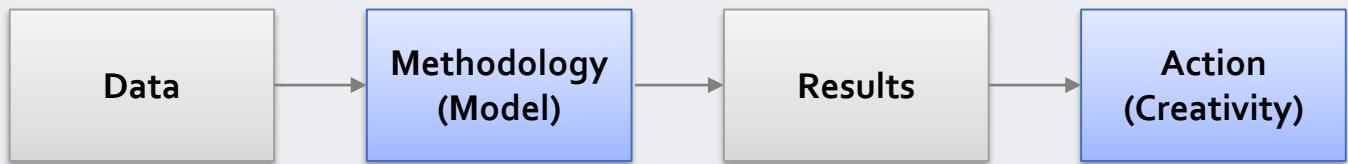
언제, 어디서?

“공사 중반, Topside General, Hull, P5,6,7 Module 등에서 주로 발생한다”



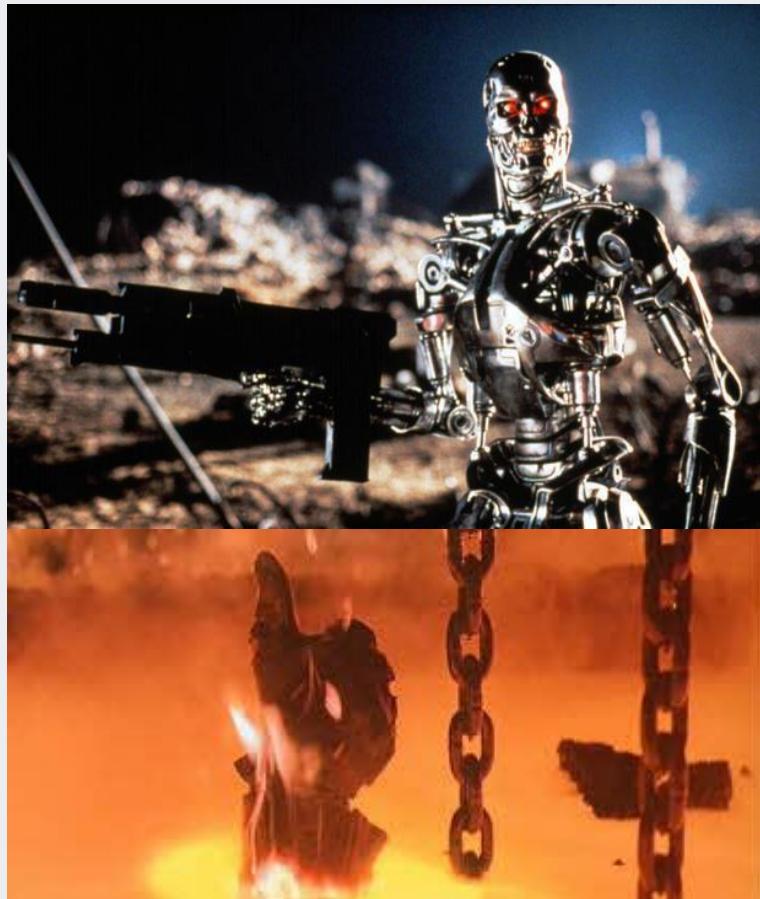
Artificial Intelligence

- Definition
 - ✓ Computers and computer software that are capable of intelligent behavior
 - ✓ Intelligent agent perceives its environment and takes actions that maximize its chance of success



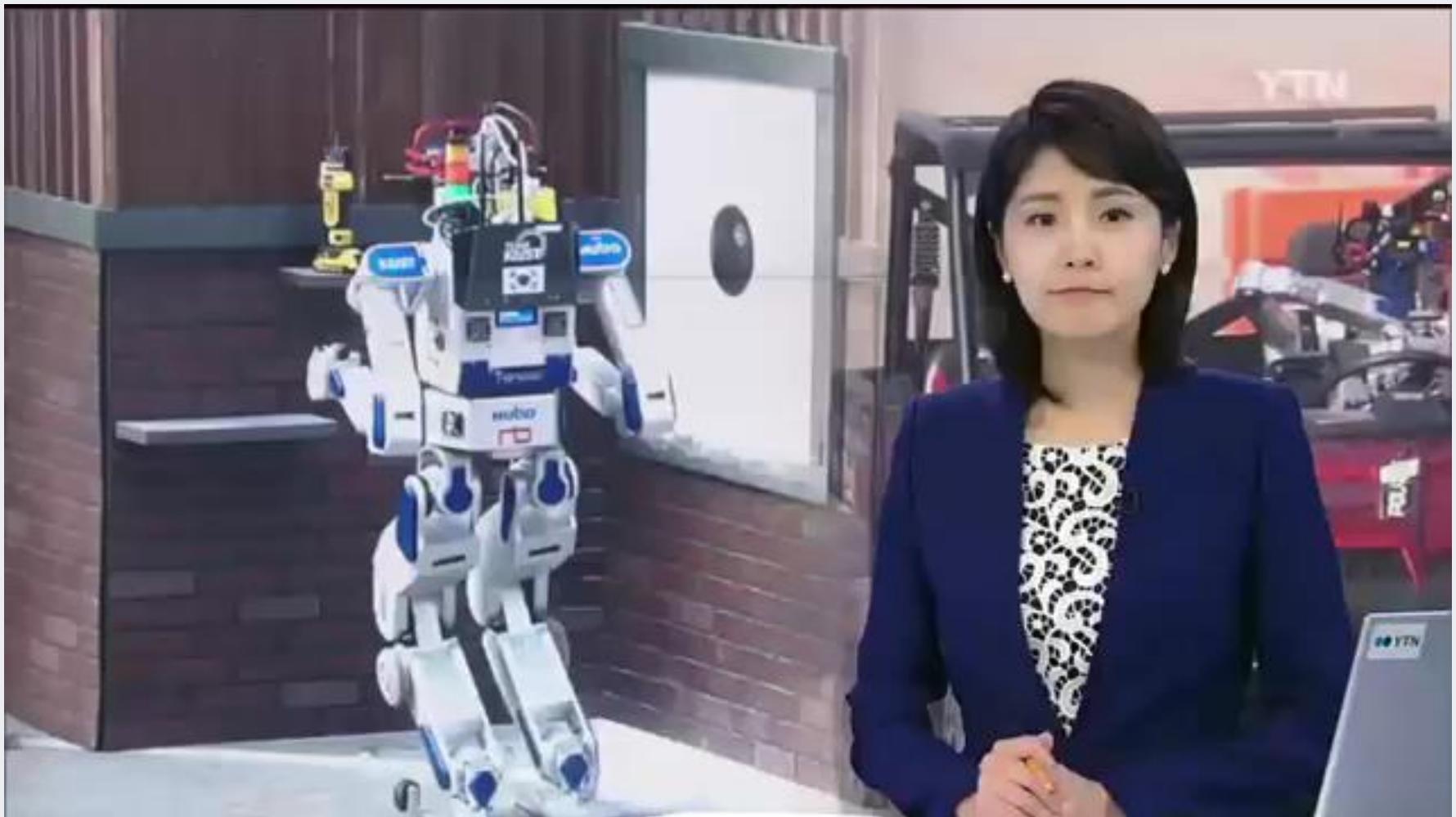
Artificial Intelligence

- AI should be...



Artificial Intelligence

- AI in nowadays...



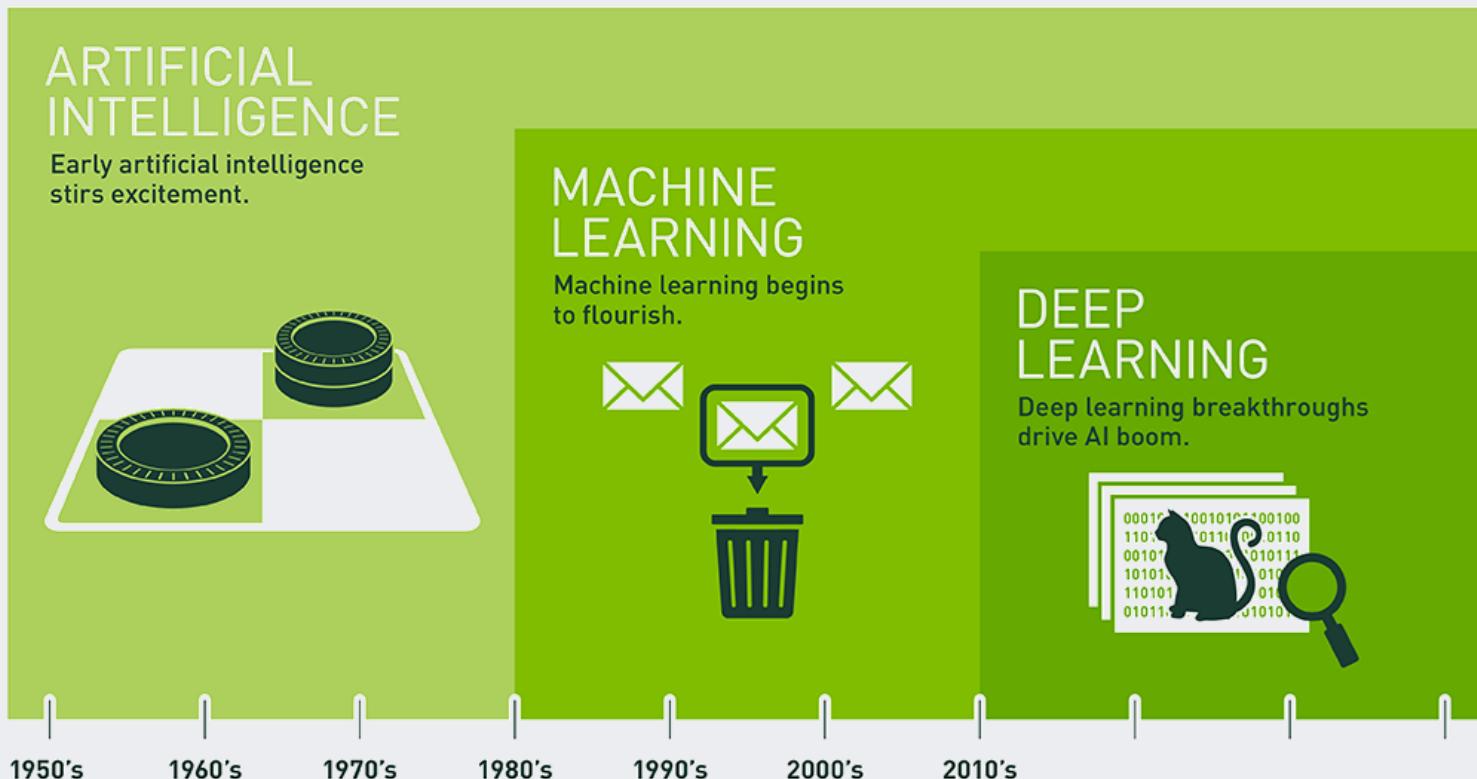
Artificial Intelligence

- AI in nowadays...



Artificial Intelligence

- AI vs. Machine Learning vs. Deep 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.

<https://blogs.nvidia.com/blog/2016/07/29/whats-difference-artificial-intelligence-machine-learning-deep-learning-ai/>

AGENDA

01 Introduction to Data Science

02 Data Science Applications

03 Multivariate Data Analysis in Data Science

04 Data Science Procedure

Data Science Applications

1

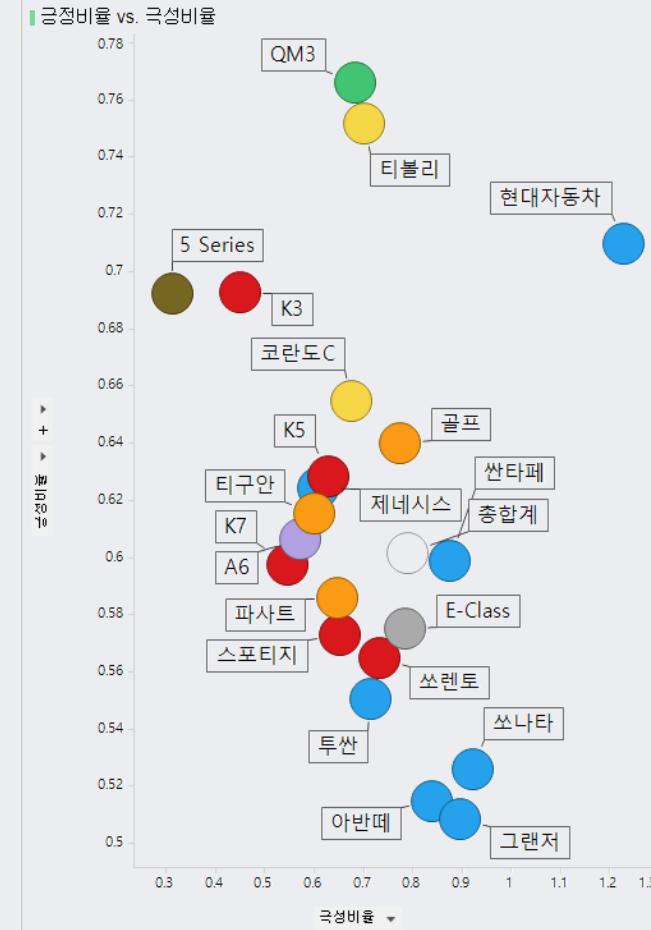
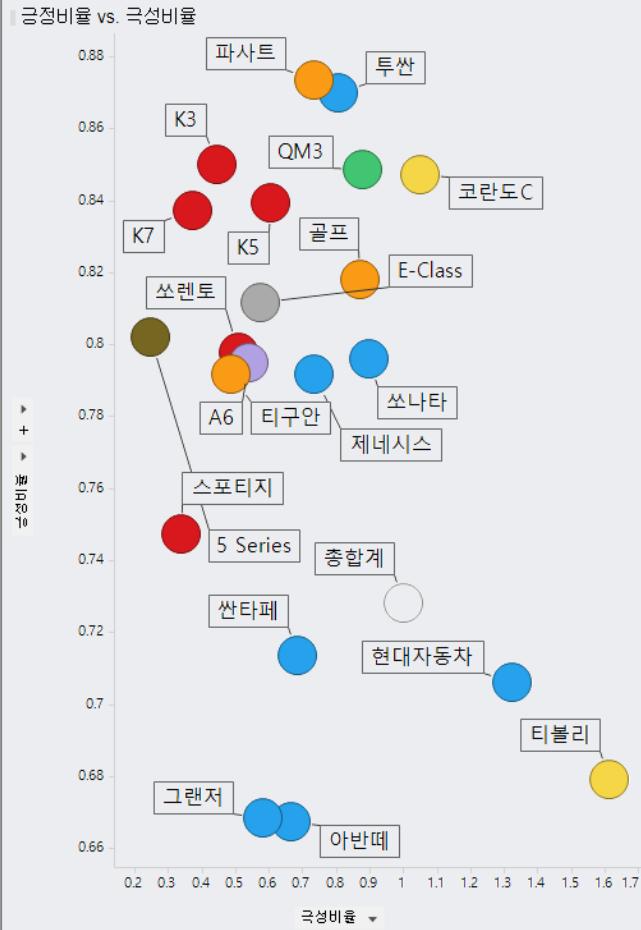
Visualization for intuitive understanding



Data Science Applications

1

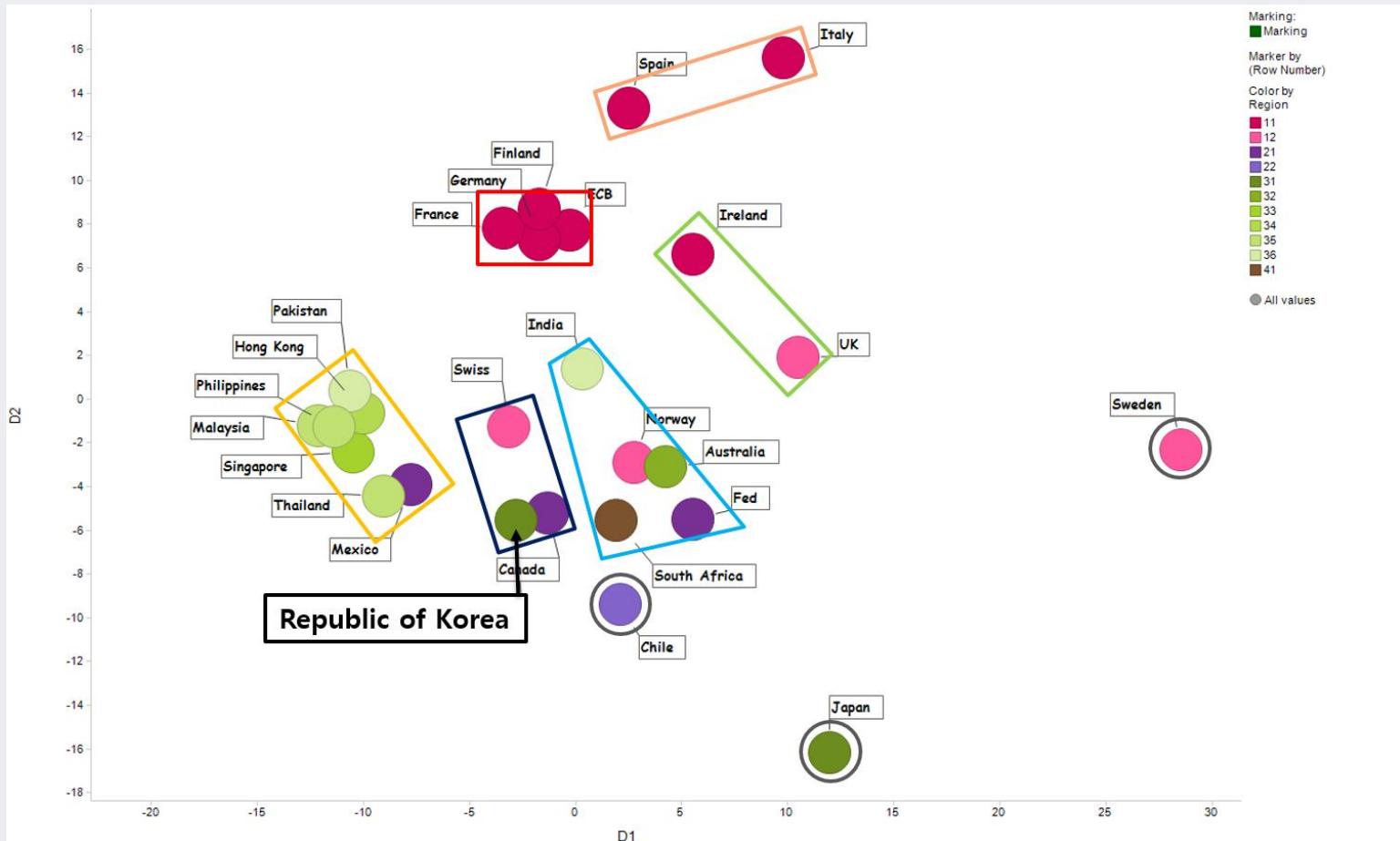
Visualization for intuitive understanding



Data Science Applications

1

Visualization for intuitive understanding



Data Science Applications

1

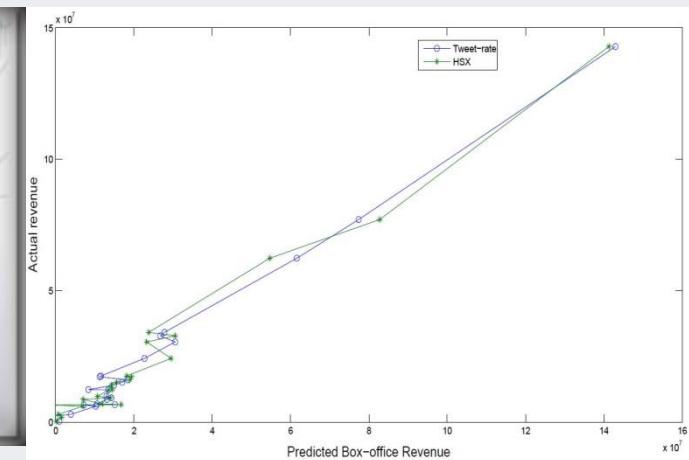
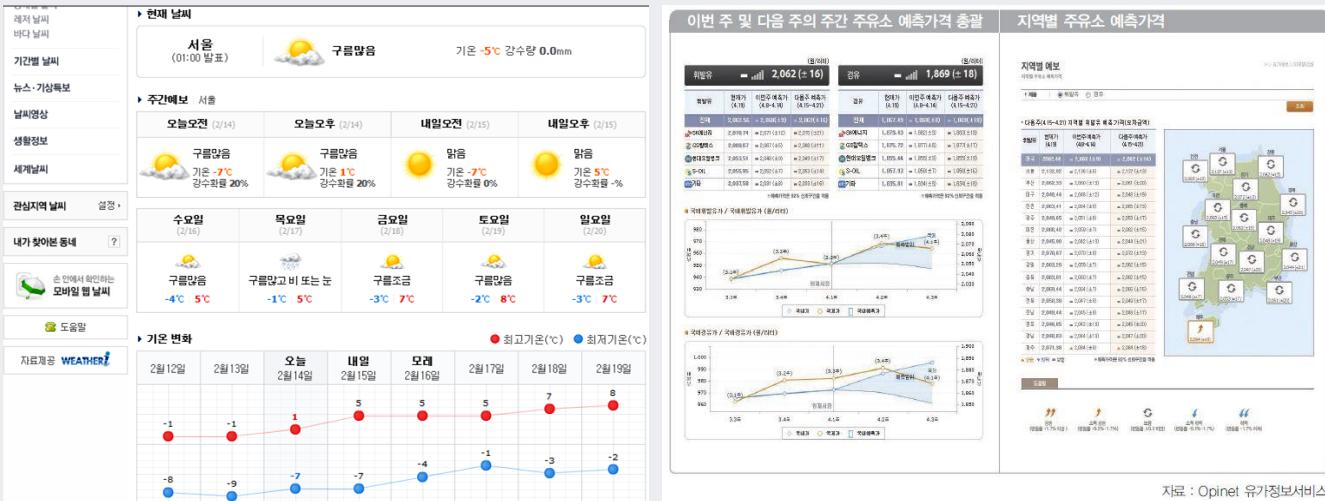
Visualization for intuitive understanding

Year	Common Concern	Federal Reserve System	European Central Bank	Deutsche Bundesbank	Bank of England	Bank of Japan
2004	Sustainability Credibility	Expansion Imports	Parliaments Cooperation	Retirement Ages Working Hours	Household-Spending	QE Deflation
2005	China Inflation	Deficits Competitive	Financial Integration	Global Imbalance	Households Future Inflation	QE Recession
2006	Competitive Global Imbalance	Incentives Risk Taking	Administered Price Indirect Taxes	Inflation	China / India Commodities	Domestic and-External Demand
2007	Subprime-Mortgage	Subprime-Mortgage	Price Stability Turmoil	Banking Supervision Disclosure	Credit	Subprime-Mortgage
2008	Financial Turmoil Commodity Prices	Financial Turmoil Funding Markets	Financial Turmoil Liquidity	Financial Turmoil Subprime	Commodity Prices Housing Market	Securitized Product
2009	Financial Crisis Lehman Brothers	Financial Crisis ABS	Non-standard-Measure	Financial Crisis Rescue	Asset Purchase Recovery	Credit Bubble Financial Crisis
2010	Recovery Reform	Recovery Recession	ESRB/ FSB Deficits	Microprudential Macroprudential	Recovery/ QE VAT/ TAX	Deflation
2011	Sovereign Debt Basel III	Dodd-Franc Act Recovery	Sovereign Debt EFSF	Debt Crisis Basel III	Commodity Prices Basel III	Asset Purchase ETFs / REITs
2012	Europe Deleveraging	Recovery (has been) Labor Market	OMT/ ESM/ SSM Fragmentation	Banking Union Taxpayer	Investment-Banking	European Debt Deleveraging
2013	Real Economy Price Stability	(At least as long as) Unemployment	SRM/ SSM/ OMT	SRM / SSM Banking Union	Prudential-Regulation	Price Stability QE

Data Science Applications

2

Predict, Diagnosis, and Detection



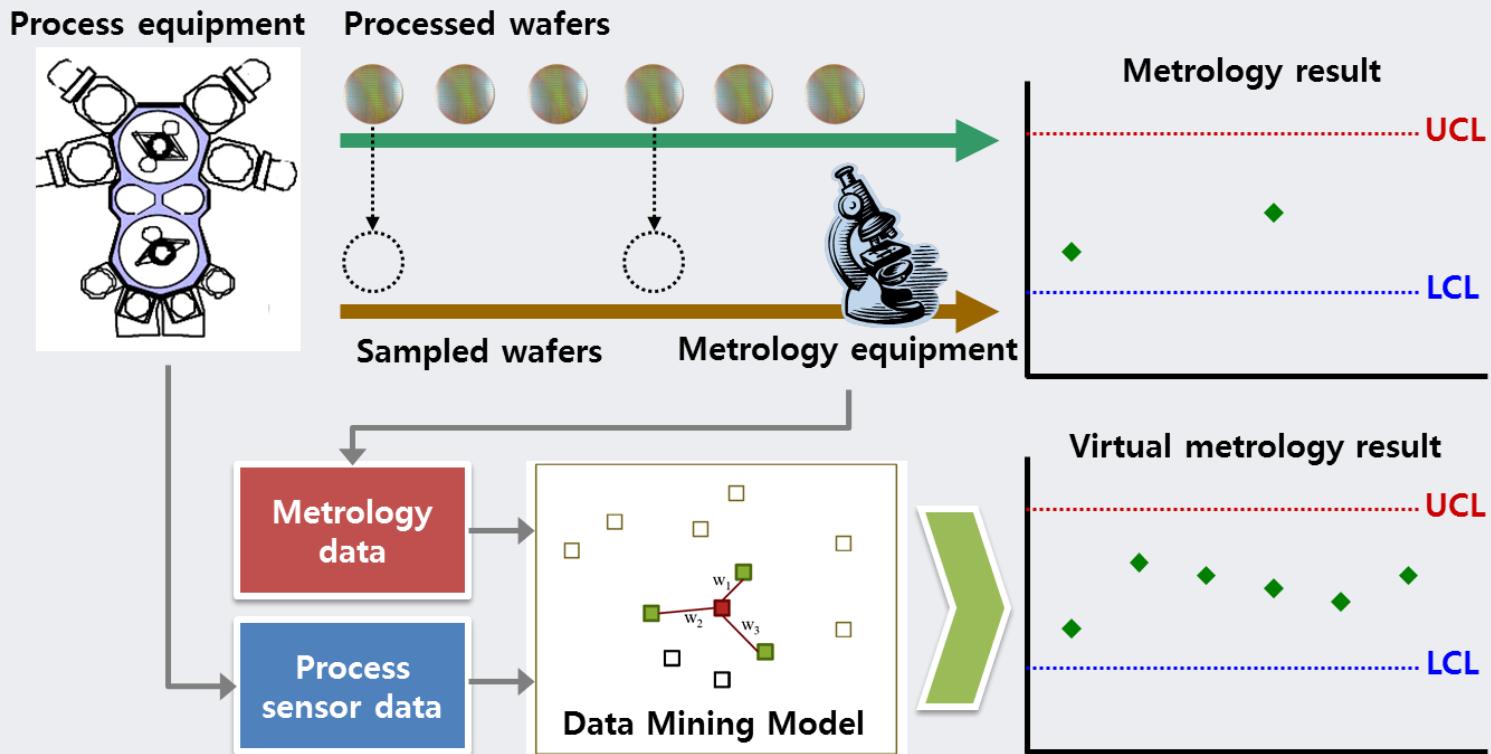
Asur and Huberman (2010) Predicting the Future with Social Media, WI-IAT10: 492-499

Data Science Applications

Predict, Diagnosis, and Detection

2

Virtual Metrology in Semiconductor Manufacturing



Data Science Applications

Support decision making in everyday life

3

Pilsung's Amazon.com | Today's Deals | Gift Cards | Help

kindle fire HD from \$199

Hello, Pilsung Your Account | Join Prime | Cart

Shop by Department | Books | Advanced Search | New Releases | Best Sellers | The New York Times® Best Sellers | Children's Books | Textbooks | Sell Your Books | Best Books of the Month

Click to LOOK INSIDE!

Data Mining for Business Intelligence: Concepts, Techniques, and Applications in Microsoft Office Excel with XLMiner [Hardcover]

Galit Shmueli (Author), Nitin R. Patel (Author), Peter C. Bruce (Author)

★★★★★ (11 customer reviews)

Buy New \$89.67 & eligible for FREE Super Saver Shipping. [Details](#)

Rent \$57.50 & this item ships for FREE with Super Saver Shipping. [Details](#)

In Stock. Ships from and sold by Amazon.com. Gift-wrap available.

In Stock. Rented by apex_media and Fulfilled by Amazon.

Want it Friday, March 8? Order within 17 hrs 57 mins, and choose One-Day Shipping at checkout. [Details](#)

49 new from \$65.00 48 used from \$37.70

FREE TWO-DAY SHIPPING FOR COLLEGE STUDENTS

amazonstudent

Formats Amazon price New from Used from

Kindle Edition	\$70.75	--	--
Hardcover	\$89.67	\$65.00	\$37.70

Rent \$57.50
Buy Used \$41.69
Buy New \$89.67

Quantity: 1

Yes, I want FREE Two-Day Shipping with Amazon Prime

Add to Cart

Sign in to turn on 1-Click ordering

Add to Wish List

Sell Us Your Item
For a \$31.29 Gift Card
Trade in

Learn more

More Buying Choices
97 used & new from \$37.70
Have one to sell? [Sell on Amazon](#)

Share

Data Science Applications

Support decision making in everyday life

3

Frequently Bought Together



Price for both: \$121.17

Add both to Cart

Add both to Wish List

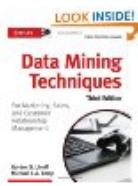
Show availability and shipping details

This item: Data Mining for Business Intelligence: Concepts, Techniques, and Applications in Microsoft Office Excel with XLMiner by Galit Shmueli Hardcover \$89.67

Data Mining Techniques: For Marketing, Sales, and Customer Relationship Management by Gordon S. Linoff Paperback \$31.50

Customers Who Bought This Item Also Bought

Page 1 of 25



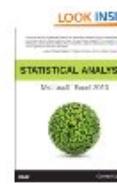
Data Mining Techniques:
For Marketing, Sales, ...
Gordon S. Linoff

★★★★☆ (34)
Paperback
\$31.50



Visual Data Mining: The
VisMiner Approach
Russell K. Anderson

Hardcover
\$62.36



Statistical Analysis:
Microsoft Excel 2010
Conrad Carlberg

★★★★★ (10)
Paperback
\$22.53



Customer Relationship
Management: A ...
V. Kumar

★★★★★ (2)
Paperback
\$93.75



AGENDA

- 01 Introduction to Data Science
- 02 Data Science Applications
- 03 Multivariate Data Analysis in Data Science
- 04 Data Science Procedure

Multivariate Data Analysis for Data Science

Data Reduction/
Structural
Simplification

Sorting and
Grouping

Investigation of the
dependence among
variables

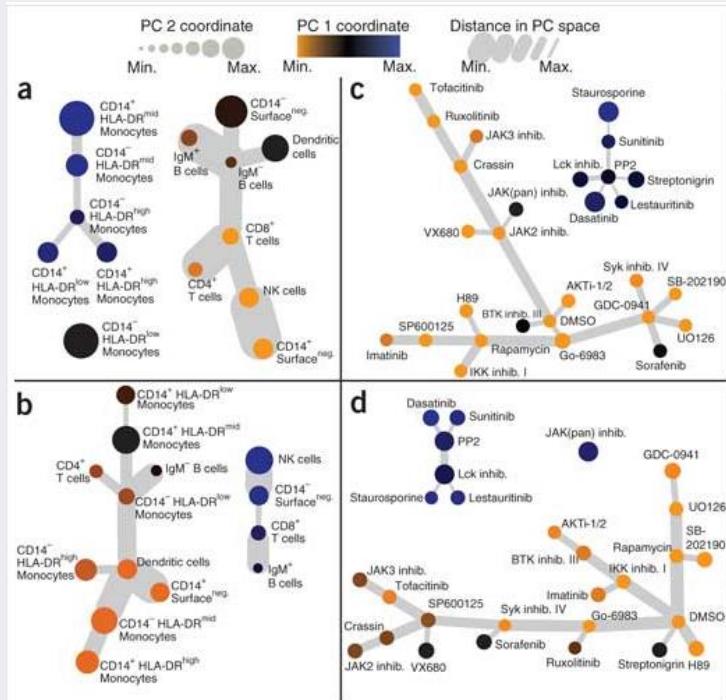
Prediction

Hypothesis
construction and
testing

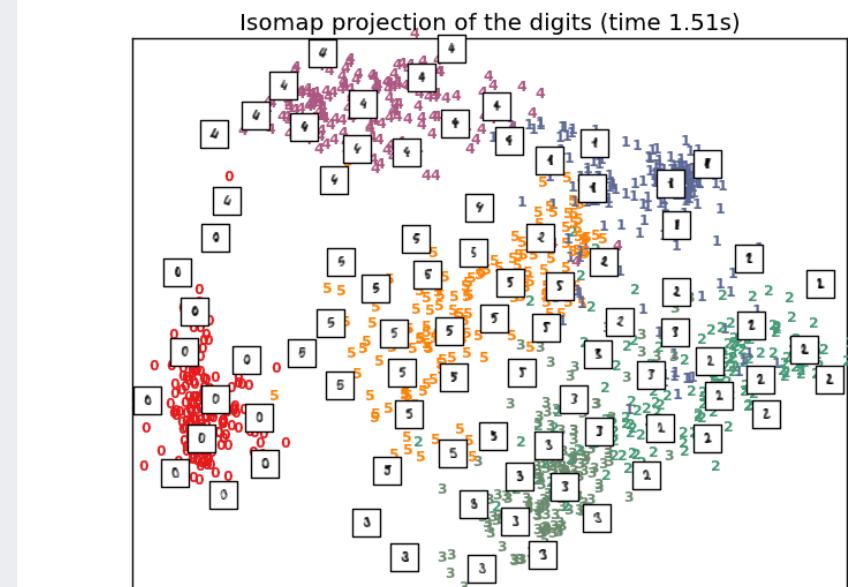
The phenomenon being studied is represented as simply as possible without sacrificing valuable information.

It is hoped that this will make interpretation easier.

Principal Component Analysis



Variable Reduction



Multivariate Data Analysis for Data Science

Data Reduction/
Structural
Simplification

Sorting and
Grouping

Investigation of the
dependence among
variables

Prediction

Hypothesis
construction and
testing

The phenomenon being studied is represented as simply as possible without sacrificing valuable information.

It is hoped that this will make interpretation easier.

- Applications

- ✓ Using data on several variables related to cancer patient responses to radio-therapy, a simple measure of patient response to radiotherapy was constructed
- ✓ Track records from many nations were used to develop an index of performance for both male and female athletes
- ✓ Multispectral image data collected by a high-altitude scanner were reduced to a form that could be viewed as images (pictures) of a shoreline in two dimensions
- ✓ Data on several variables relating to yield and protein content were used to create an index to select parents of subsequent generations of improved bean plants

Multivariate Data Analysis for Data Science

Data Reduction/
Structural
Simplification

Sorting and
Grouping

Investigation of the
dependence among
variables

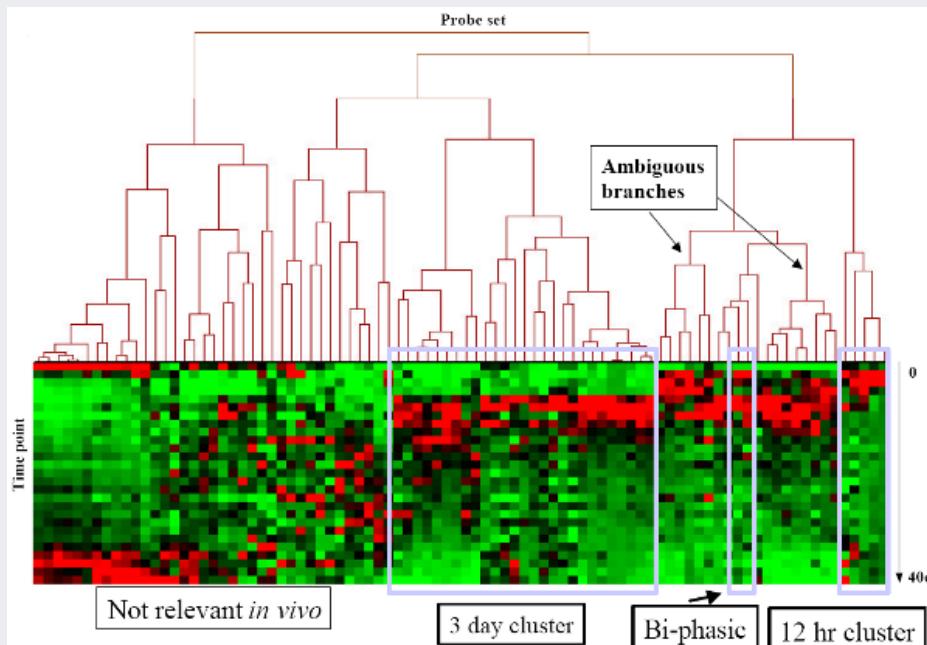
Prediction

Hypothesis
construction and
testing

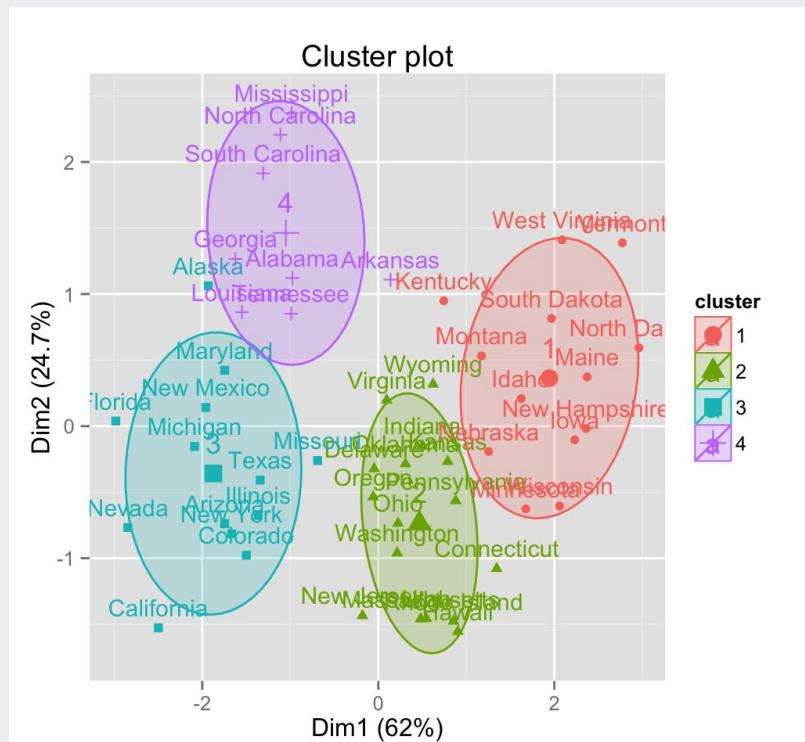
Groups of “similar” objects or variables are created, based upon measured characteristics.

Alternatively, rules for classifying objects into well-defined groups may be required.

Clustering: Hierarchical Clustering



Clustering: K-Means Clustering



Multivariate Data Analysis for Data Science

Data Reduction/
Structural
Simplification

Sorting and
Grouping

Investigation of the
dependence among
variables

Prediction

Hypothesis
construction and
testing

Groups of “similar” objects or variables are created, based upon measured characteristics.

Alternatively, rules for classifying objects into well-defined groups may be required.

- Applications

- ✓ Data in several variables related to computer use were employed to create clusters of categories of computer jobs that allow a better determination of existing computer utilization
- ✓ Measurements of several physiological variables were used to develop a screening procedure that discriminates alcoholics from nonalcoholics
- ✓ Data related to responses to visual stimuli were used to develop a rule for separating people suffering from a multiple-sclerosis-caused visual pathology from those not suffering from the disease

Multivariate Data Analysis for Data Science

Data Reduction/
Structural
Simplification

Sorting and
Grouping

Investigation of the
dependence among
variables

Prediction

Hypothesis
construction and
testing

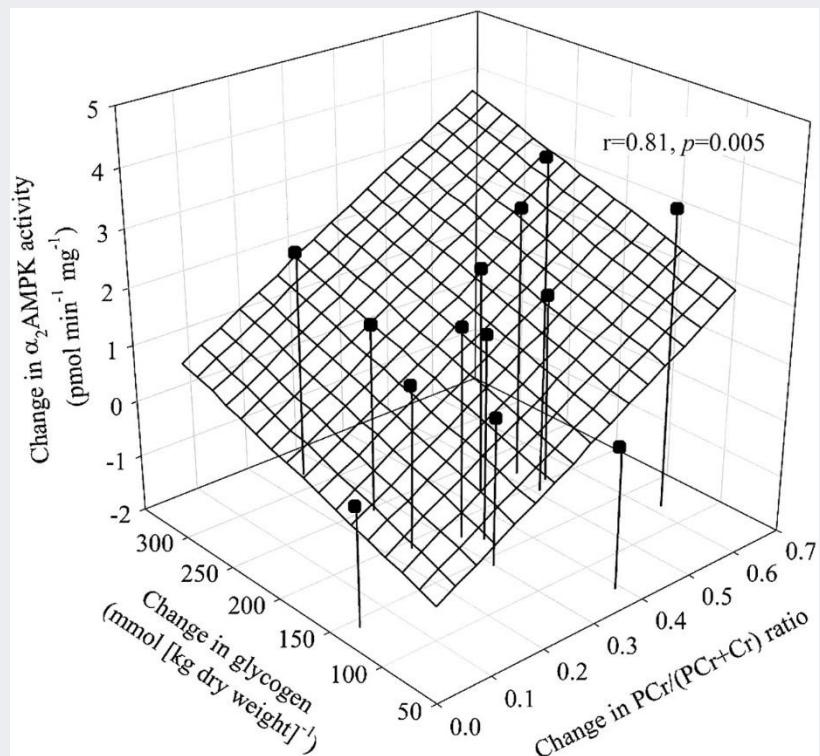
The nature of the relationships among variables is of interest

Are all the variables mutually independent or are one or more variables dependent on the others? If so, how?

Association Rule Mining



Factor Analysis



Multivariate Data Analysis for Data Science

Data Reduction/
Structural
Simplification

Sorting and
Grouping

Investigation of the
dependence among
variables

Prediction

Hypothesis
construction and
testing

The nature of the relationships among variables is of interest

Are all the variables mutually independent or are one or more variables dependent on the others? If so, how?

- Applications

- ✓ Data on several variables were used to identify factors that were responsible for client success in hiring external consultants.
- ✓ Measurements of variables related to innovation, on the one hand, and variables related to the business environment and business organization, on the other hand, were used to discover why some firms are product innovators and some firms are not.
- ✓ The associations between measures of risk-taking propensity and measures of socioeconomic characteristics for top-level business executives were used to assess the relation between risk-taking behavior and performance

Multivariate Data Analysis for Data Science

Data Reduction/
Structural
Simplification

Sorting and
Grouping

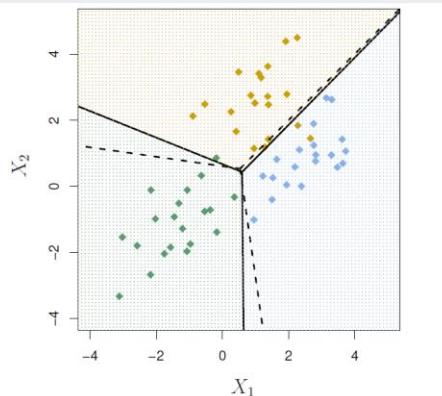
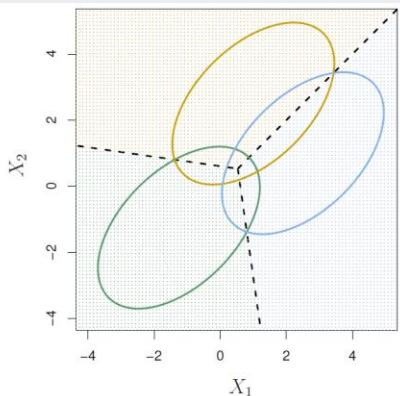
Investigation of the
dependence among
variables

Prediction

Hypothesis
construction and
testing

Relationships between variables must be determined for the purpose of predicting the value of one or more variables on the basis of observations on the other variables.

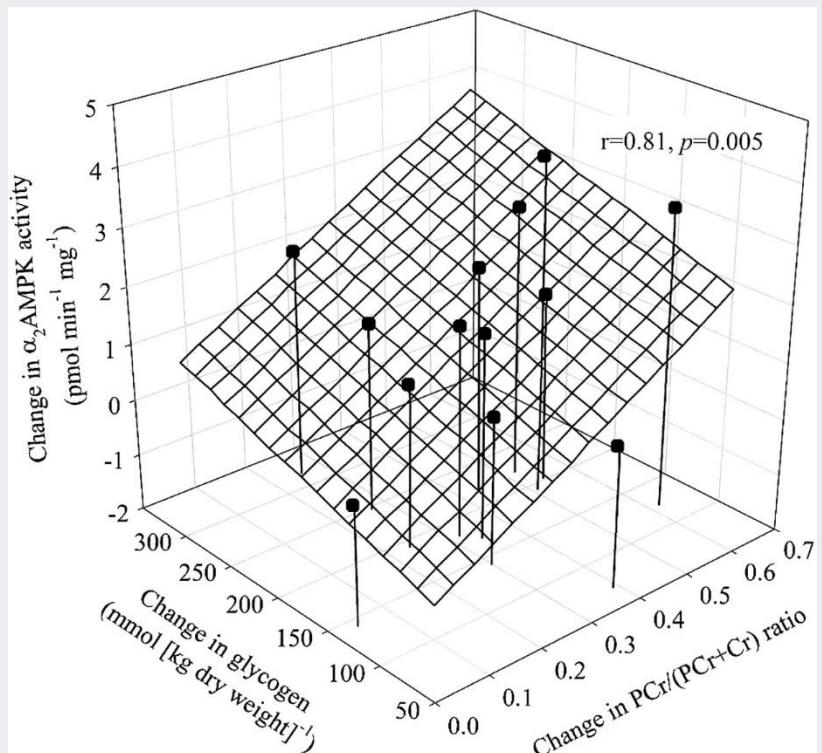
Discrimination and Classification



Here $\pi_1 = \pi_2 = \pi_3 = 1/3$.

The dashed lines are known as the *Bayes decision boundaries*. Were they known, they would yield the fewest misclassification errors, among all possible classifiers.

Multivariate Linear Regression



Multivariate Data Analysis for Data Science

Data Reduction/
Structural
Simplification

Sorting and
Grouping

Investigation of the
dependence among
variables

Prediction

Hypothesis
construction and
testing

Relationships between variables must be determined for the purpose of predicting the value of one or more variables on the basis of observations on the other variables.

- Applications

- ✓ The associations between test scores, and several high school performance variables, and several college performance variables were used to develop predictors of success in college
- ✓ Data on several variables related to the size distribution of sediments were used to develop rules for predicting different depositional environments
- ✓ Measurements on several accounting and financial variables were used to develop a method for identifying potentially insolvent property-liability insurers

Multivariate Data Analysis for Data Science

Data Reduction/
Structural
Simplification

Sorting and
Grouping

Investigation of the
dependence among
variables

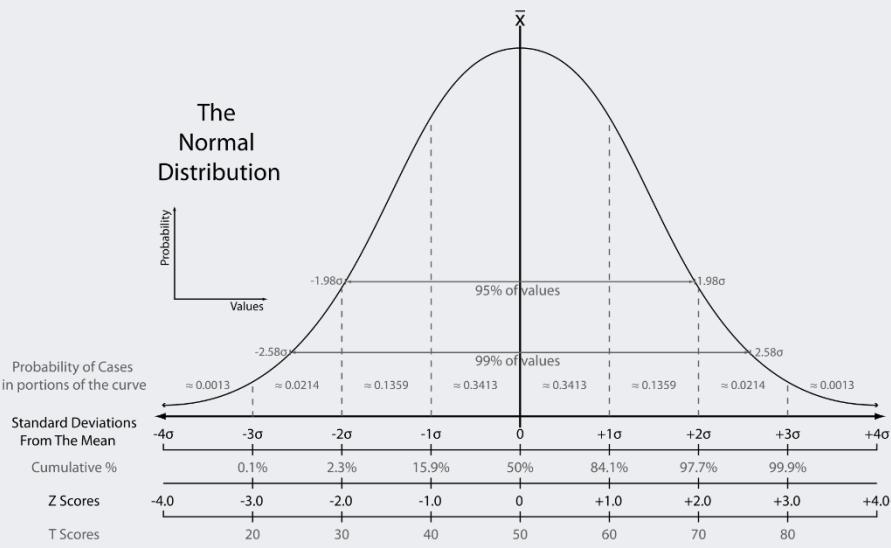
Prediction

Hypothesis
construction and
testing

Specific statistical hypotheses, formulated in terms of the parameters of multivariate populations, are tested.

This may be done to validate assumptions or to reinforce prior convictions.

Inferences about a Mean Vector



Comparisons of Several Multivariate Means

		Treatment		
		1	2	...
Subject	1	$\mathbf{Y}_{11} = \begin{pmatrix} Y_{111} \\ Y_{112} \\ \vdots \\ Y_{11p} \end{pmatrix}$	$\mathbf{Y}_{21} = \begin{pmatrix} Y_{211} \\ Y_{212} \\ \vdots \\ Y_{21p} \end{pmatrix}$	\dots
		$\mathbf{Y}_{12} = \begin{pmatrix} Y_{121} \\ Y_{122} \\ \vdots \\ Y_{12p} \end{pmatrix}$	$\mathbf{Y}_{22} = \begin{pmatrix} Y_{221} \\ Y_{222} \\ \vdots \\ Y_{22p} \end{pmatrix}$	\dots
		\vdots	\vdots	\vdots
n_1	$\mathbf{Y}_{1n_1} = \begin{pmatrix} Y_{1n_11} \\ Y_{1n_12} \\ \vdots \\ Y_{1n_1p} \end{pmatrix}$	$\mathbf{Y}_{2n_1} = \begin{pmatrix} Y_{2n_11} \\ Y_{2n_12} \\ \vdots \\ Y_{2n_1p} \end{pmatrix}$	\dots	$\mathbf{Y}_{gn_g} = \begin{pmatrix} Y_{gn_g1} \\ Y_{gn_g2} \\ \vdots \\ Y_{gn_gp} \end{pmatrix}$
		\vdots	\vdots	\vdots

Multivariate Data Analysis for Data Science

Data Reduction/
Structural
Simplification

Sorting and
Grouping

Investigation of the
dependence among
variables

Prediction

Hypothesis
construction and
testing

Specific statistical hypotheses, formulated in terms of the parameters of multivariate populations, are tested.

This may be done to validate assumptions or to reinforce prior convictions.

- Applications

- ✓ Several pollution-related variables were measured to determine whether levels for a large metropolitan area were roughly constant throughout the week, or whether there was a noticeable difference between weekdays and weekends
- ✓ Experimental data on several variables were used to see whether the nature of the instructions makes any difference in perceived risks, as quantified by test scores
- ✓ Data on many variables were used to investigate the differences in structure of American occupations to determine the support for one of two competing sociological theories

AGENDA

01 Introduction to Data Science

02 Data Science Applications

03 Multivariate Data Analysis in Data Science

04 Data Science Procedure

Data Science Procedure

Ask an interesting question

- ▶ 풀고자 하는 문제가 무엇인가?
- ▶ 만약 관련해 모든 데이터를 보유하고 있다면 무엇을 할 것인가?
- ▶ 무엇을 예측하고 추정하기를 원하는가?

Get the data

- ▶ 데이터는 어떻게 샘플링할 것인가?
- ▶ 어떤 데이터와 정보가 우리 목표와 관련이 있는가?
- ▶ 프라이버시나 개인정보 이슈는 없는가?

Explore the data

- ▶ 데이터를 그려보며 데이터의 속성과 구조를 알아보기
- ▶ 데이터에서 이상한 점은 없는가?
- ▶ 데이터에 어떠한 패턴이 존재하는가?

Model the data

- ▶ 모델 수립
- ▶ 모델 적합화
- ▶ 모델 검증

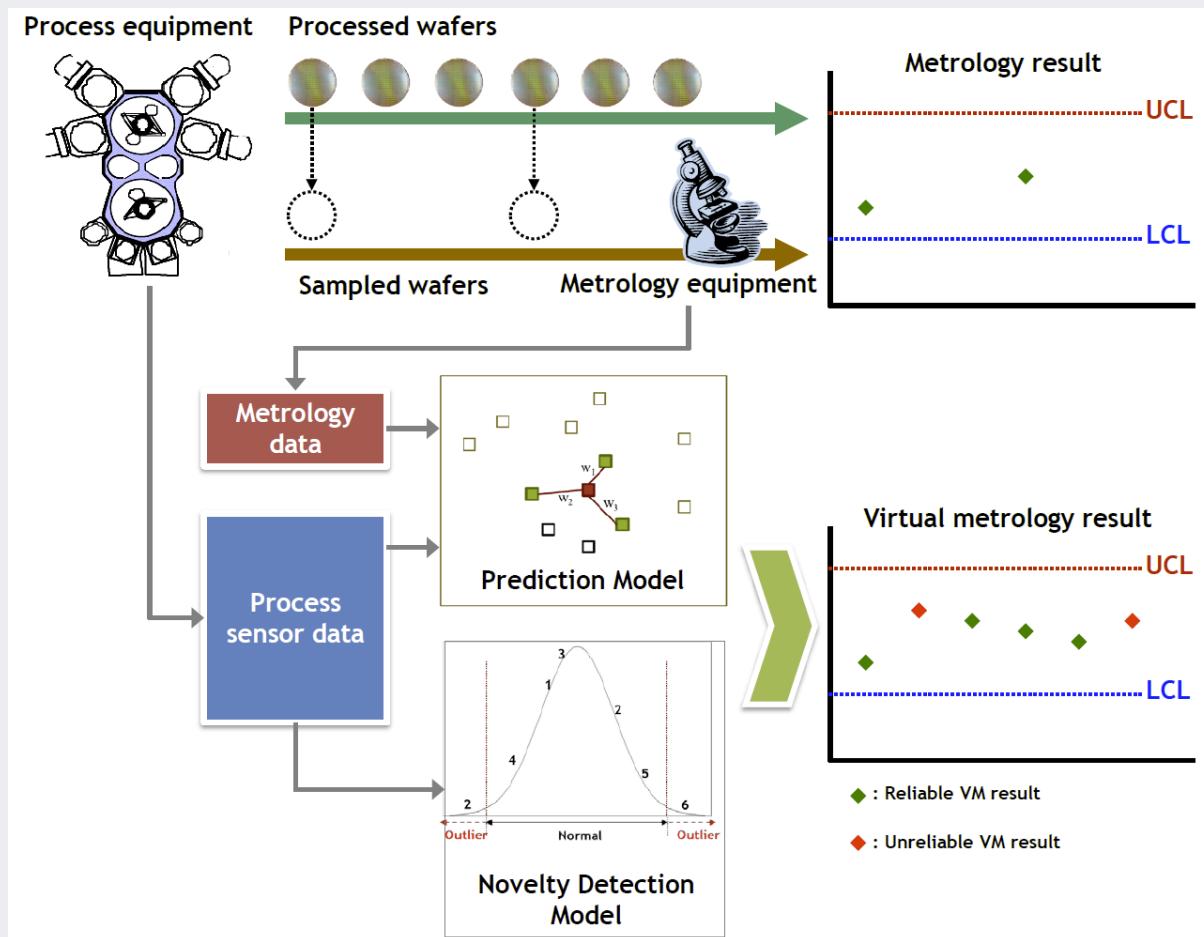
Communicate and visualize the results

- ▶ 결과 요약 및 시사점 분석
- ▶ 결과가 타당한가?
- ▶ 스토리를 말할 수 있는가? (전략 수립)

Ask an Interesting Question

- Step 1: Ask an interesting question

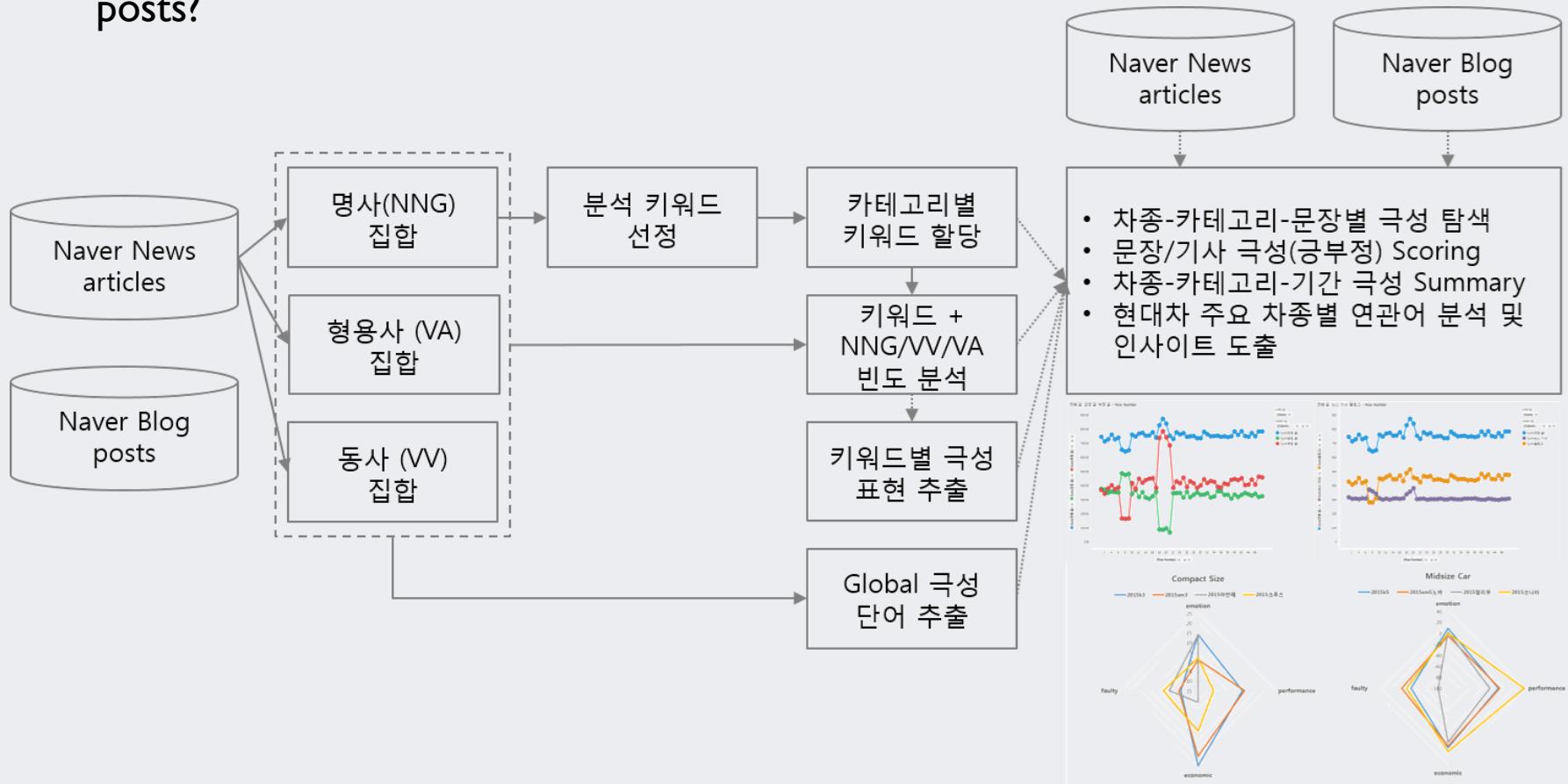
- ✓ Can we predict the product quality based on the sensor data collected from equipment?



Ask an Interesting Question

- Step 1: Ask an interesting question

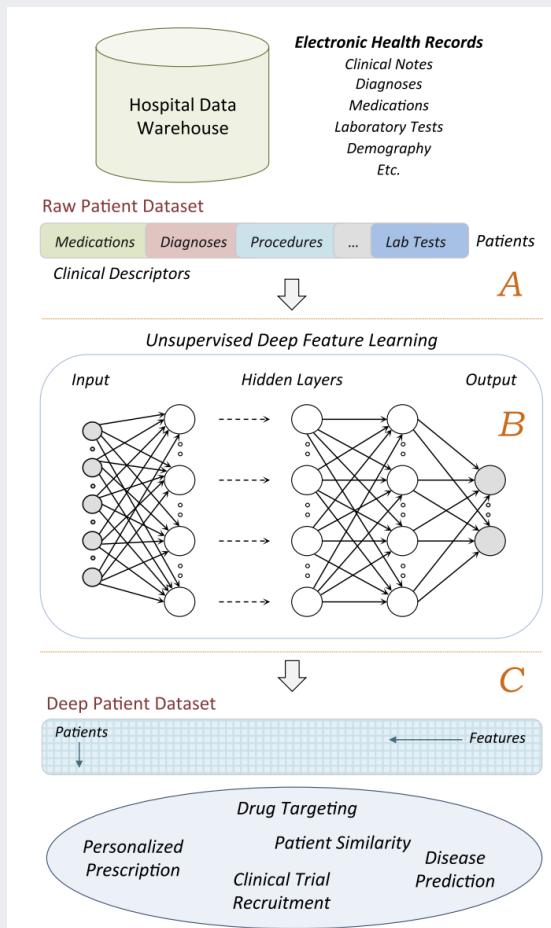
- ✓ Is it possible to understand customers' preference based on news articles and blog posts?



Ask an Interesting Question

- Step 1: Ask an interesting question

✓ Can we predict various diseases based on the electric health records (EHR)?

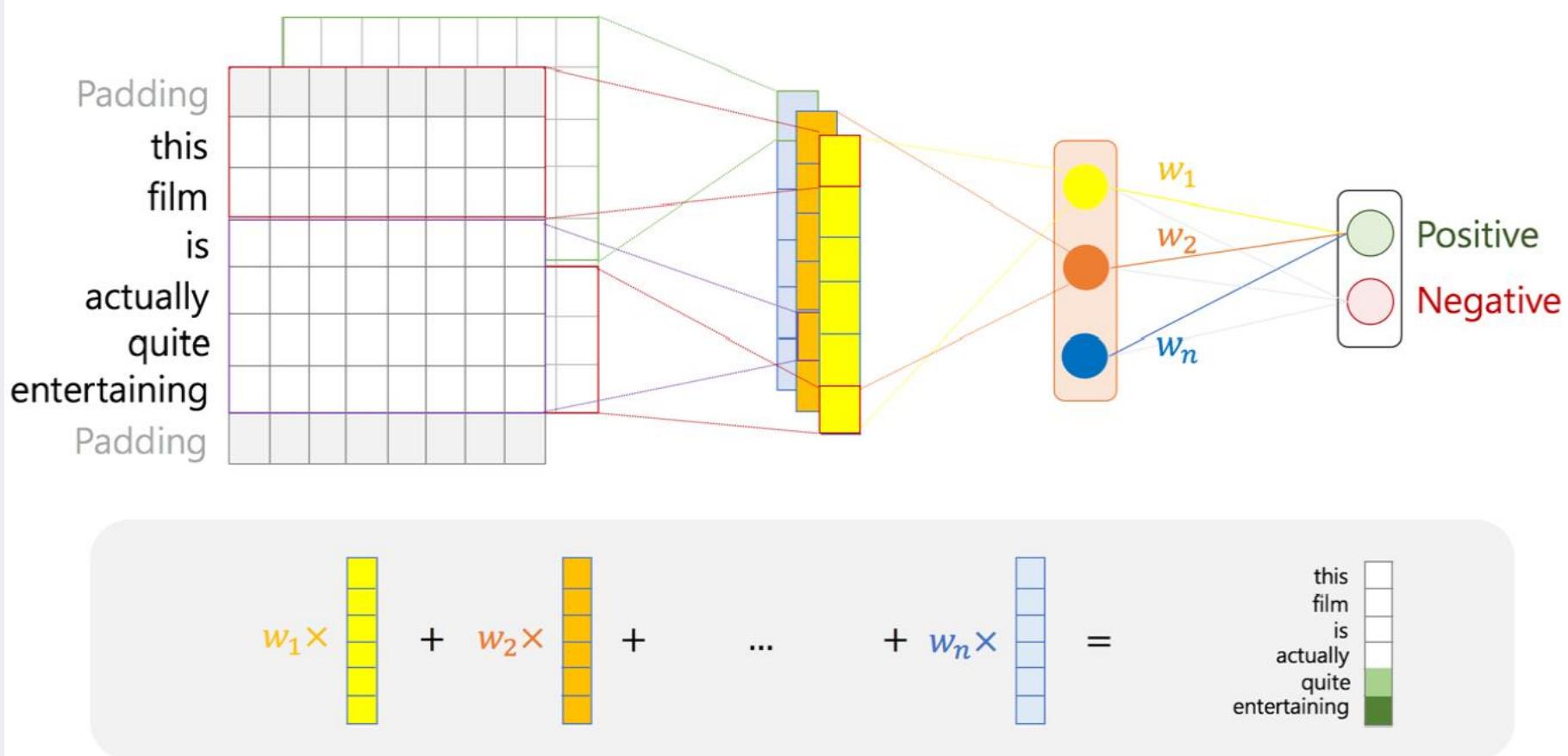


Disease	Area under the ROC curve		
	RawFeat	PCA	DeepPatient
Diabetes mellitus with complications	0.794	0.861	0.907
Cancer of rectum and anus	0.863	0.821	0.887
Cancer of liver and intrahepatic bile duct	0.830	0.867	0.886
Regional enteritis and ulcerative colitis	0.814	0.843	0.870
Congestive heart failure (non-hypertensive)	0.808	0.808	0.865
Attention-deficit and disruptive behavior disorders	0.730	0.797	0.863
Cancer of prostate	0.692	0.820	0.859
Schizophrenia	0.791	0.788	0.853
Multiple myeloma	0.783	0.739	0.849
Acute myocardial infarction	0.771	0.775	0.847
Personality disorders	0.787	0.788	0.846
Inflammatory conditions of male genital organs	0.659	0.825	0.841
Endometriosis	0.697	0.765	0.839
Inflammatory diseases of female pelvic organs	0.714	0.799	0.830
Cancer of ovary	0.646	0.788	0.824
Sickle cell anemia	0.567	0.689	0.822
Nephritis, nephrosis and renal sclerosis	0.763	0.775	0.821
Cancer of bladder	0.711	0.744	0.818
Chronic kidney disease	0.764	0.758	0.814
Cancer of testis	0.508	0.771	0.811
Menopausal disorders	0.681	0.772	0.808
Delirium, dementia and amnestic (and other cognitive disorders)	0.728	0.720	0.803
Peritonitis and intestinal abscess	0.689	0.747	0.801
Cardiac arrest and ventricular fibrillation	0.711	0.747	0.799
Developmental disorders	0.705	0.737	0.798

Ask an Interesting Question

- Step 1: Ask an interesting question

✓ Can we find the emotional expressions automatically from review texts?



Ask an Interesting Question

- Step 1: Ask an interesting question
 - ✓ Can we find the emotional expressions automatically from review texts?

Method	Sentence
Raw text	One of the funniest most romantic and most musical movies ever; definitely worth renting/buying especially if you have a taste for older style of cinematography. The animals and the songs alone will make you smile while watching the movie. A definite must for Madonna fans. :o) (10 / 10 points)
Rand	One of the the funniest most romantic and musical movies ever definitely worth renting buying especially if you have a taste for older style cinematography The animals songs alone will make smile while watching movie A definite must Madonna fans Positive
Static	One of the funniest most romantic and musical movies ever definitely worth renting buying especially if yo u have a taste for older style cinematography The animals songs alone will make smile while watching movi e A definite must Madonna fans Positive
NStatic	One of the funniest most romantic and musical movies ever definitely worth renting buying especially if yo u have a taste for older style cinematography The animals songs alone will make smile while watching movi e A definite must Madonna fans Positive
2ch	One of the funniest most romantic and musical movies ever definitely worth renting buying especially if yo u have a taste for older style cinematography The animals songs alone will make smile while watching movi e A definite must Madonna fans Positive

Ask an Interesting Question

- Step 1: Ask an interesting question

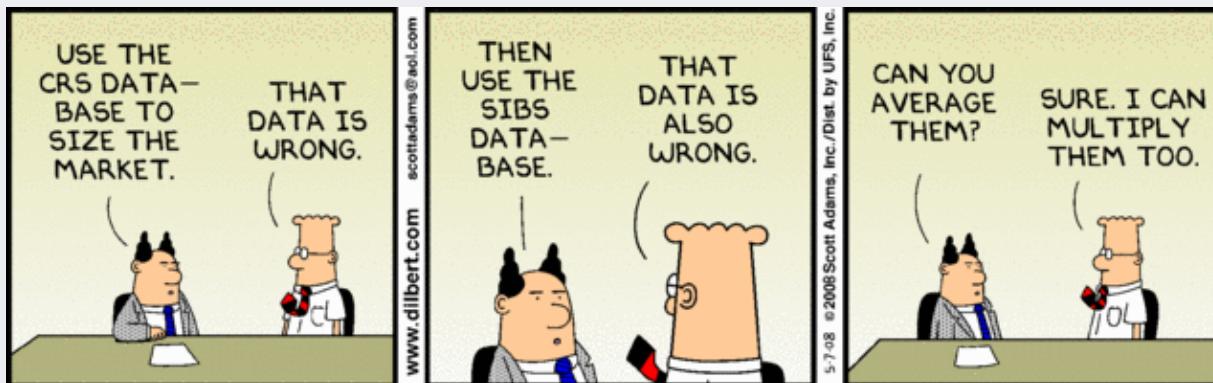
✓ Can we find the emotional expressions automatically from review texts?

Method	Sentence
Raw text	This is one of the most boring films I've ever seen. The three main cast members just didn't seem to click well. Giovanni Ribisi's character was quite annoying. For some reason he seems to like repeating what he says. If he was the Rain Man it would've been fine but he's not. (3 / 10 points)
Rand	This is one of the most boring films I've ever seen. The three main cast members just didn't seem to click well. Giovanni Ribisi's character was quite annoying. For some reason he seems to like repeating what he says. If Rain Man it would be fine but he's not Negative
Static	This is one of the most boring films I've ever seen. The three main cast members just didn't seem to click well. Giovanni Ribisi's character was quite annoying. For some reason he seems to like repeating what he says. If Rain Man it would be fine but he's not Negative
NStatic	This is one of the most boring films I've ever seen. The three main cast members just didn't seem to click well. Giovanni Ribisi's character was quite annoying. For some reason he seems to like repeating what he says. If Rain Man it would be fine but not Negative
2ch	This is one of the most boring films I've ever seen. The three main cast members just didn't seem to click well. Giovanni Ribisi's character was quite annoying. For some reason he seems to like repeating what he says. If Rain Man it would be fine but not Negative
4ch	This is one of the most boring films I've ever seen. The three main cast members just didn't seem to click well. Giovanni Ribisi's character was quite annoying. For some reason he seems to like repeating what he says. If Rain Man it would be fine but not Negative

Get the Data

- Step 2: Get the Data

✓ Garbage in, garbage out



✓ The larger, the better



“We don’t have better algorithms than anyone else. We just have more data.”

Get the Data

- Step 2: Get the Data

- ✓ Use domain knowledge from experts (especially when making answer sets)

The screenshot shows a medical research article from JAMA. The title is "Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs". The authors listed are Varun Gulshan, PhD; Lily Peng, MD, PhD; Marc Coram, PhD; Martin C. Stumpe, PhD; Derek Wu, BS; Arunachalam Narayanaswamy, PhD; Subhashini Venugopalan, MS; Kasumi Widner, MS; Tom Madams, MEng; Jorge Cuadros, OD, PhD; Ramasamy Kim, OD, DNB; Rajiv Raman, MS, DNB; Philip C. Nelson, BS; Jessica L. Mega, MD, MPH; Dale R. Webster, PhD. The abstract highlights the importance of deep learning in medical imaging, the objective of creating an algorithm for automated detection of diabetic retinopathy and macular edema, and the design and setting of the study using a deep convolutional neural network trained on 128,175 retinal images. The article is categorized under Research and Original Investigation.

Research

JAMA | Original Investigation | INNOVATIONS IN HEALTH CARE DELIVERY

Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs

Varun Gulshan, PhD; Lily Peng, MD, PhD; Marc Coram, PhD; Martin C. Stumpe, PhD; Derek Wu, BS; Arunachalam Narayanaswamy, PhD; Subhashini Venugopalan, MS; Kasumi Widner, MS; Tom Madams, MEng; Jorge Cuadros, OD, PhD; Ramasamy Kim, OD, DNB; Rajiv Raman, MS, DNB; Philip C. Nelson, BS; Jessica L. Mega, MD, MPH; Dale R. Webster, PhD

IMPORTANCE Deep learning is a family of computational methods that allow an algorithm to program itself by learning from a large set of examples that demonstrate the desired behavior, removing the need to specify rules explicitly. Application of these methods to medical imaging requires further assessment and validation.

OBJECTIVE To apply deep learning to create an algorithm for automated detection of diabetic retinopathy and diabetic macular edema in retinal fundus photographs.

DESIGN AND SETTING A specific type of neural network optimized for image classification called a deep convolutional neural network was trained using a retrospective development data set of 128 175 retinal images, which were graded 3 to 7 times for diabetic retinopathy, diabetic macular edema, and image gradability by a panel of 54 US licensed ophthalmologists and ophthalmology senior residents between May and December 2015. The resultant algorithm was validated in January and February 2016 using 2 separate data sets, both graded by at least 7 US board-certified ophthalmologists with high intragrader consistency.

← Editorial

+ Supplemental content

Get the Data

Table. Baseline Characteristics^a

Characteristics	Development Data Set	EyePACS-1 Validation Data Set	Messidor-2 Validation Data Set
No. of images	128 175	9963	1748
No. of ophthalmologists	54	8	7
No. of grades per image	3-7	8	7
Grades per ophthalmologist, median (interquartile range)	2021 (304-8366)	8906 (8744-9360)	1745 (1742-1748)

Patient demographics



^a Summary of image characteristics and available demographic information in the development and clinical validation data sets (EyePACS-1 and Messidor-2). Abnormal images were oversampled for the development set for algorithm training. The clinical validation sets were not enriched for abnormal images.

^b Unique patient codes (deidentified) were available for 89.3% of the development set ($n = 114\,398$ images).

^c Individual-level data including age and sex were available for 66.1% of the development set ($n = 84\,734$ images).

^d Image quality was assessed for a subset of the development set.

^e Referable diabetic retinopathy, defined as the presence of moderate and worse diabetic retinopathy and/or referable diabetic macular edema according to the International Clinical Diabetic Retinopathy Scale,¹⁴ was calculated for each ophthalmologist before combining them using a majority decision. The 5-point grades represent the grade that received the highest number of votes for diabetic retinopathy alone. Hence, the sum of moderate, severe, and proliferative diabetic retinopathy for the 5-point grade differs slightly from the count of referable diabetic retinopathy images.

Explore the Data

- Step 3: Explore the data before modeling

This dashboard includes a choropleth map of the United States showing homelessness rates by state, a bar chart of the top 50 states and territories, and a scatter plot of homelessness rates versus population.

Analysis of U.S. Homelessness

INTERACTIVE

This dashboard features a world map where countries are colored based on their life expectancy rates, accompanied by a bar chart of life expectancy by country.

World Facts Analysis

INTERACTIVE

This dashboard contains two scatter plots showing the relationship between market value and 1-year relative returns for various sectors.

Performance Attribution Analytics

INTERACTIVE

This dashboard includes a world map showing energy consumption by country and a bar chart of energy consumption in million tonnes.

World Energy Survey Analysis

INTERACTIVE

This dashboard displays a campaign overview with a bar chart of total revenue, a line graph of daily revenue over time, and a table of campaign details.

Marketing Campaign Performance Analytics

INTERACTIVE

This dashboard features a horizontal bar chart of weather alert types and a choropleth map of California showing警报 distribution.

US National Weather Service Alert Analysis

INTERACTIVE **VIDEO**

Explore the Data

- Step 3: Explore the data before modeling

✓ Data visualization software can be helpful

The image displays four separate web-based data visualization platforms side-by-side:

- TIBCO Spotfire:** Shows a dashboard titled "Laws of Attrition" featuring a heatmap and a bar chart.
- Qlik:** Shows a dashboard titled "Salesforce" with a laptop displaying a QlikView interface.
- Tableau:** Shows a map titled "Boris Bikes by Station During the July 2015 Tube Strike in London" with many colored dots representing bike stations.
- QlikView Demos:** Shows a grid of various QlikView dashboards under categories like Business, Life, and Tags for Business.

Model the Data

- Step 4: Model the data

- ✓ Select an appropriate algorithm based on the purpose of task, data characteristics, requirement for interpretability.



Communicate and Visualize the Results

- Step 5: Communicate and Visualize the Results
 - ✓ System implementation, A/B test, model updates, etc.



<https://adparlor.zendesk.com/hc/en-us/articles/202933459-How-To-Efficiently-A-B-Test-with-AdParlor-on-Facebook>



ANY
questions?