Projected sales of main products in 2013
Distribution of market share among the major industry players
Share of market activity
Changes in the activity of the active and passive market is uncertain. Established positive trends in various market segments.
Projected sales of main products in 2013
25%
75%
Passive market share
350
300
250
200
150
100
50
0

Lecture 1: Data Analytics

Pilsung Kang

School of Industrial Management Engineering

Korea University

AGENDA

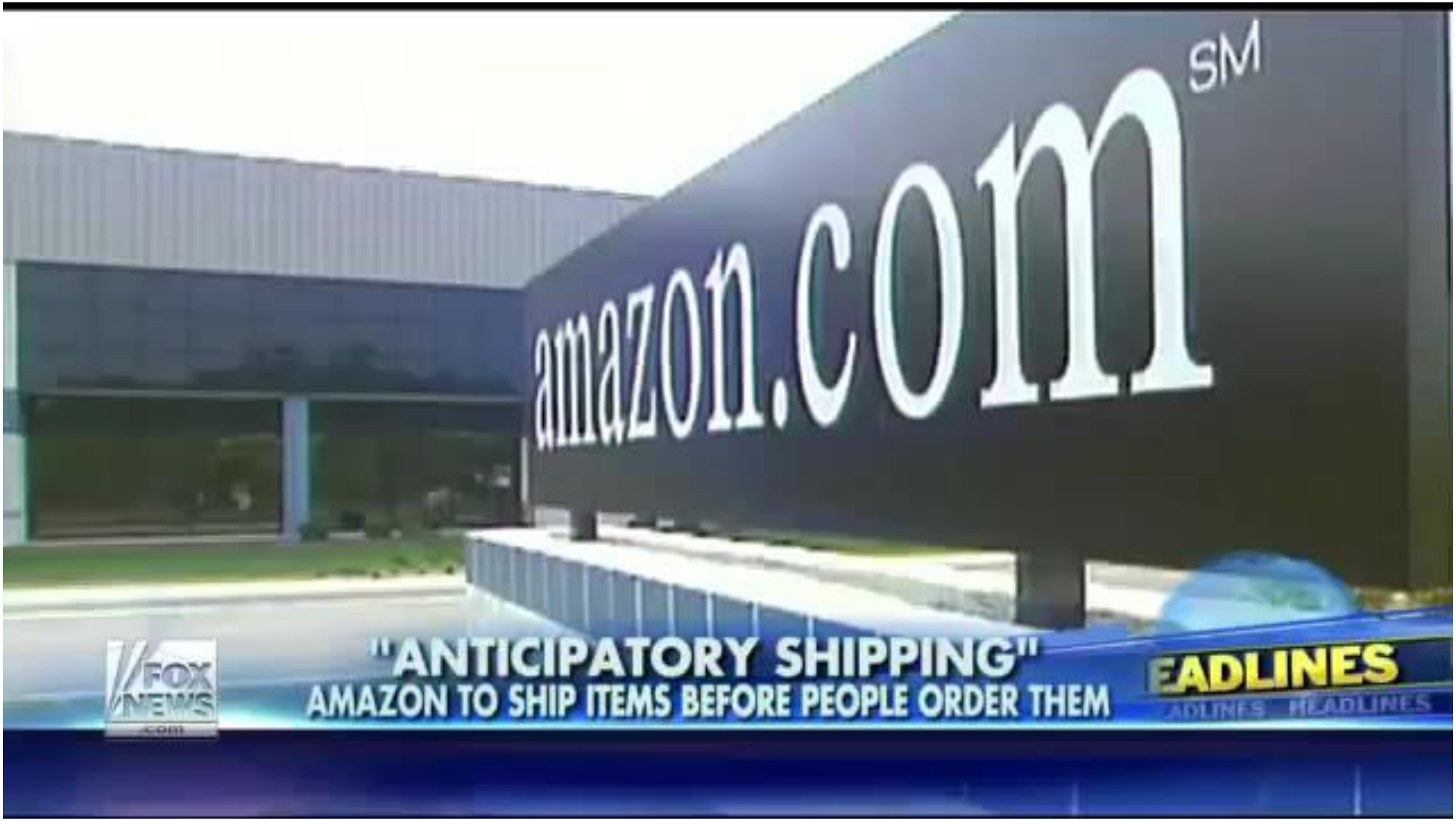
01 What is Data Analytics?

02 What Can We Do with Data Analytics?

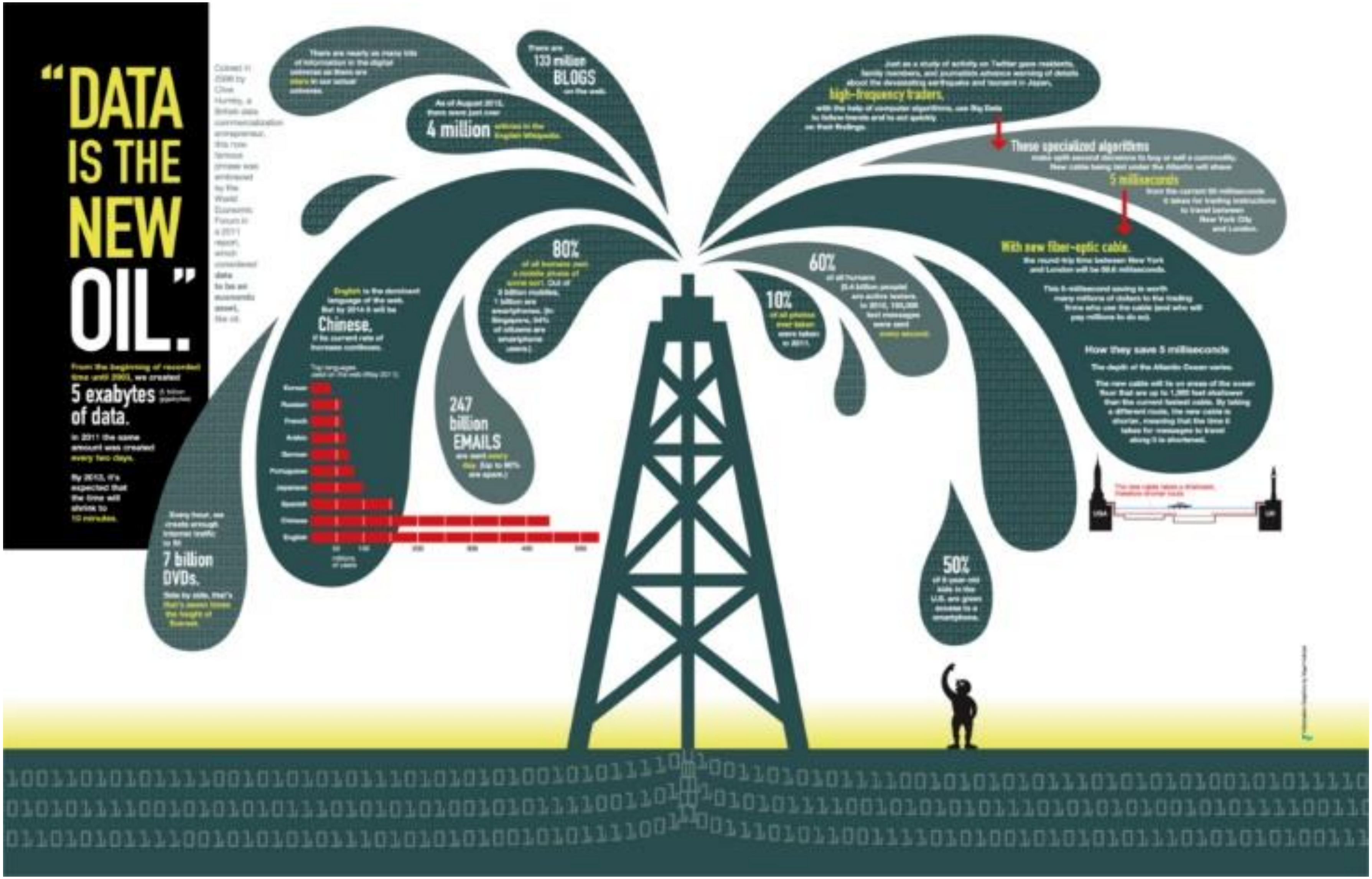
03 Data Analytics Languages: R & Python

Data Analytics

- Data Analytics by Amazon



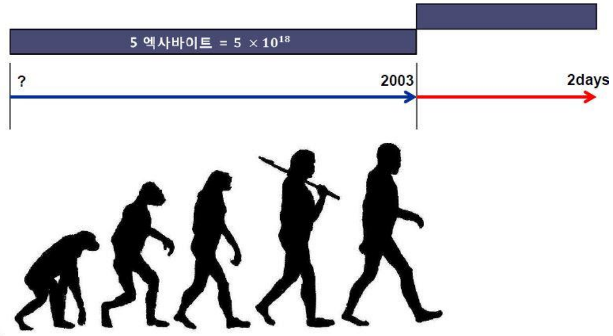
Data Analytics: The Era of "Big Data"


"DATA IS THE NEW OIL."
Coined in 2006 by Clive Humby, a British data commercialization entrepreneur, this now famous phrase was embraced by the World Economic Forum in a 2011 report, which considered data to be an economic asset, like oil.
From the beginning of recorded time until 2003, we created
5 exabytes of data.
In 2011 the same amount was created every two days.
By 2013, it's expected that the time will shrink to 10 minutes.
There are nearly as many bits of information in the digital universe as there are stars in our actual universe.
There are 133 million BLOGS on the web.
As of August 2012, there were just over 4 million articles in the English Wikipedia.
English is the dominant language of the web. But by 2014 it will be Chinese, if the current rate of increase continues.
Top languages used on the web
Korean
Russian
French
Arabic
German
Portuguese
Japanese
Spanish
Chinese
English
millions of users
80% of all humans own a mobile phone of some sort. Out of 5 billion mobiles, 1 billion are smartphones.
247 billion EMAILS are sent every day.
Every hour, we create enough Internet traffic to fill 7 billion DVDs.
Just as a study of activity on Twitter gave residents, family members, and journalists advance warning of details about the devastating earthquake and tsunami in Japan,
high-frequency traders,
with the help of computer algorithms, use Big Data to follow trends and to act quickly on their findings.
These specialized algorithms
make split-second decisions to buy or sell a commodity.
5 milliseconds
With new fiber-optic cable,
the round-trip time between New York and London will be 59.6 milliseconds.
10% of all photos ever taken were taken in 2011.
60% of all humans are active texters.
How they save 5 milliseconds
The depth of the Atlantic Ocean varies.
USA
UK
50% of 8-year-old kids in the U.S. are given access to a smartphone.
1001101010101110010101010101110101010100101010111

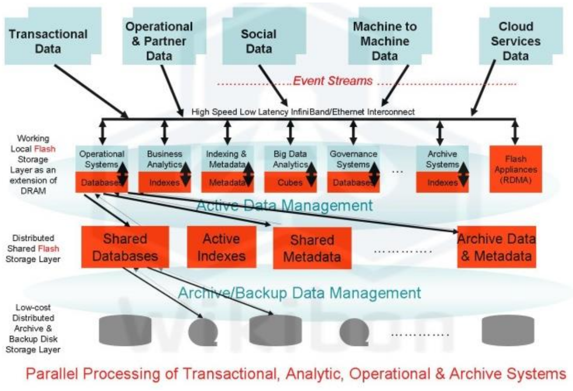
Data Analytics: The Era of "Big Data"

- 4Vs in Big Data

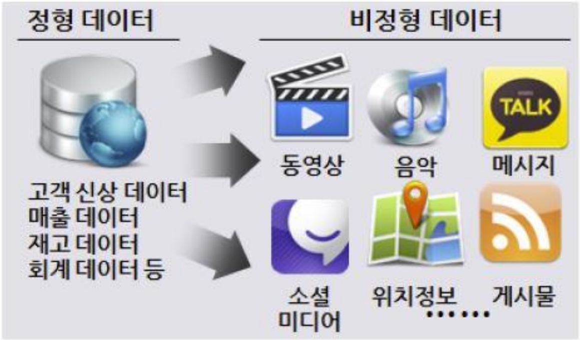
Volume



Velocity



Variety



Value

의료 건강(美)
- 매년 $3,300억 가치
- 年 ~0.7%의 생산성 증가

공공, 행정 부문(유럽)
- 매년 €2,500억 가치
- 年 ~0.5%의 생산성 증가

개인 위치 정보(글로벌)
- 서비스 공급자 매출은 $1,000억 이상
- 사용자 혜택은 $7,000억

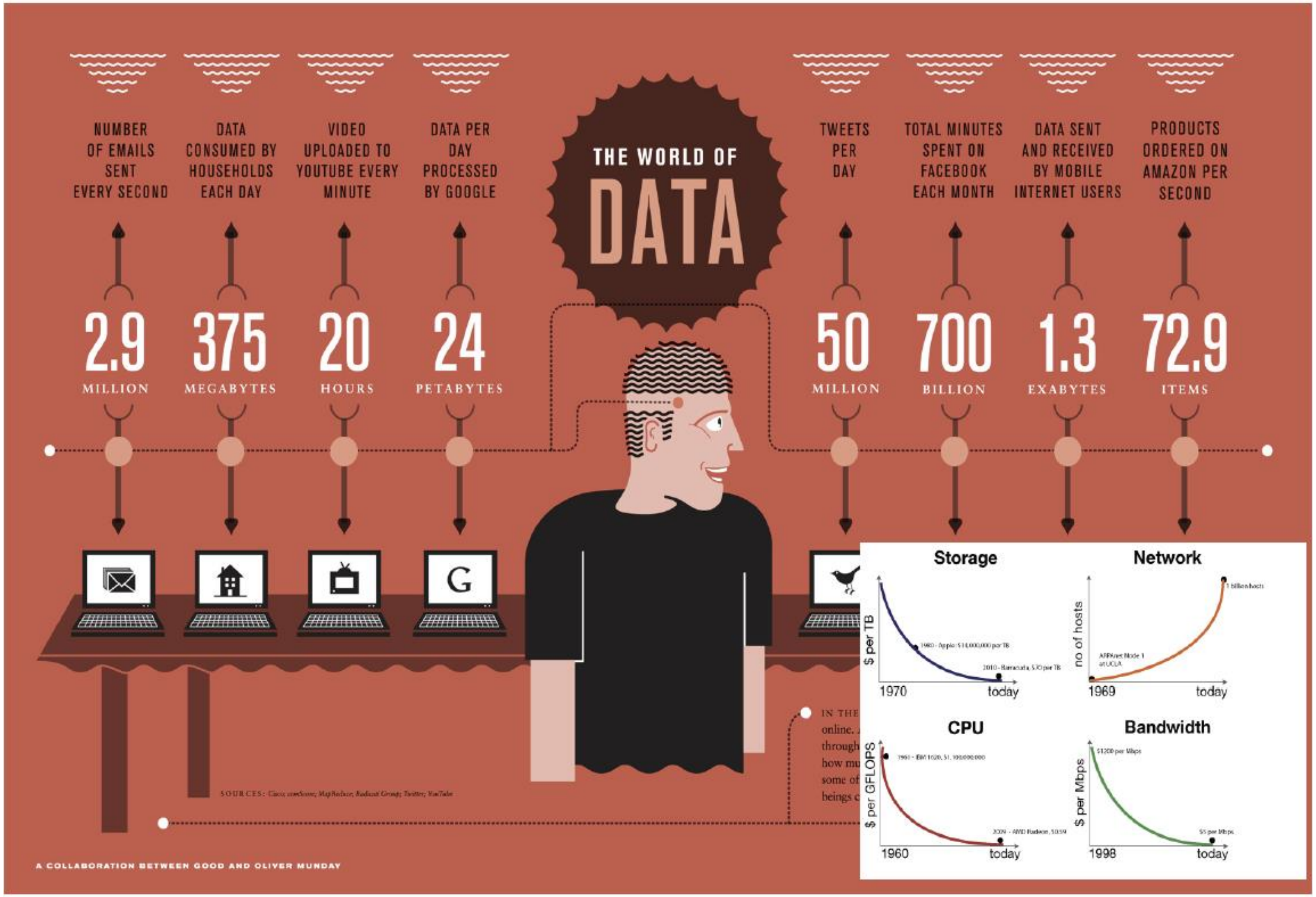
소매업(美)
- 이윤 60% 증가 가능
- 年 0.5~1% 생산성 증가

제조업
- 제품개발비 50% 감소
- 운전자본 7% 절감 가능

자료: McKinsey (2011.05)

Data Analytics: The Era of “Big Data”

- Volume of Big Data

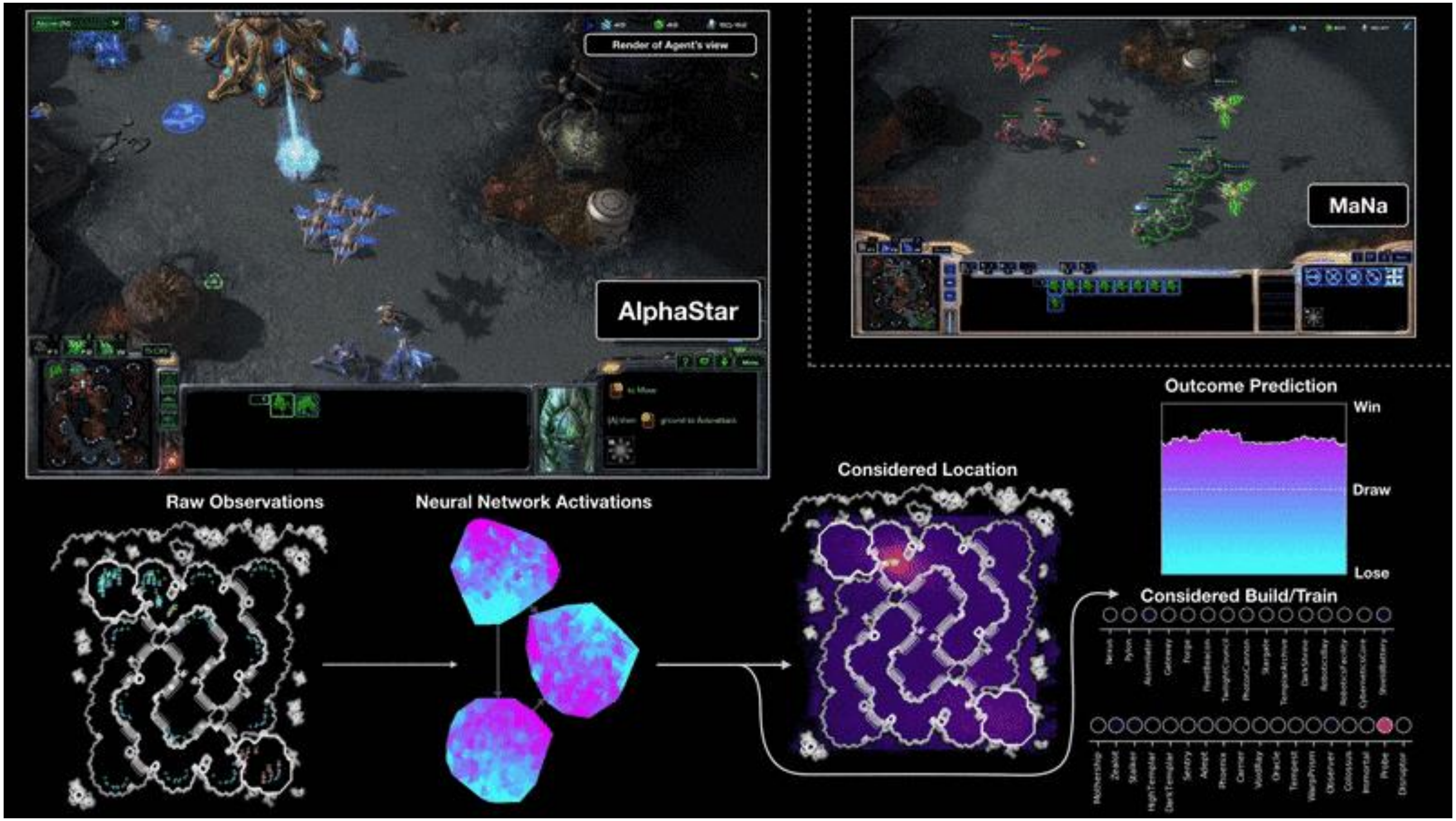


SOURCES: Cisco; comScore; MapReduce; Radicati Group; Twitter; YouTube

A COLLABORATION BETWEEN GOOD AND OLIVER MUNDAY

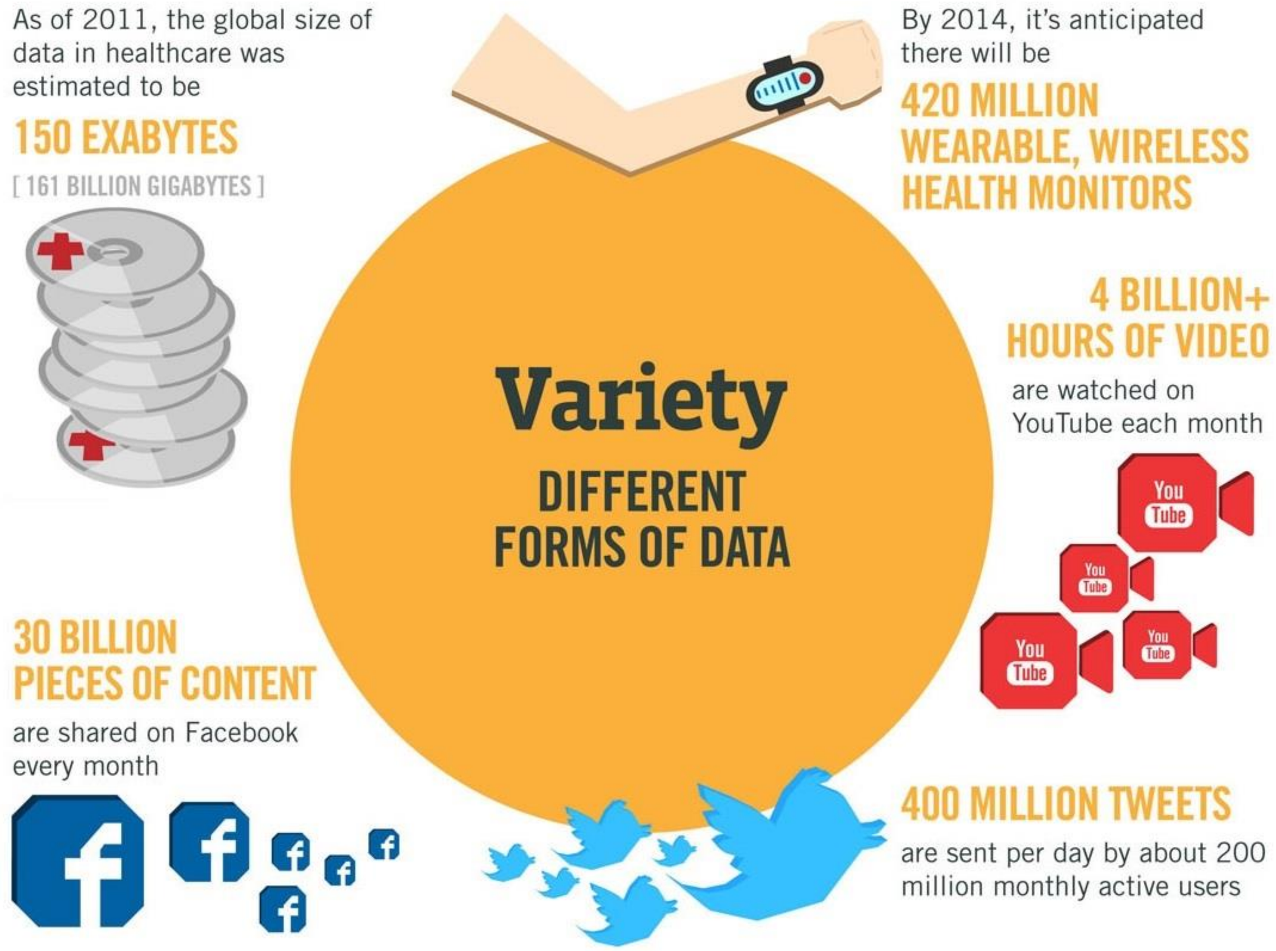
Data Analytics: The Era of "Big Data"

- Velocity of Big Data



Data Analytics: The Era of "Big Data"

- Variety in Big Data



As of 2011, the global size of data in healthcare was estimated to be

150 EXABYTES

[161 BILLION GIGABYTES]

By 2014, it's anticipated there will be

420 MILLION WEARABLE, WIRELESS HEALTH MONITORS

Variety

DIFFERENT FORMS OF DATA

4 BILLION+ HOURS OF VIDEO

are watched on YouTube each month

30 BILLION PIECES OF CONTENT

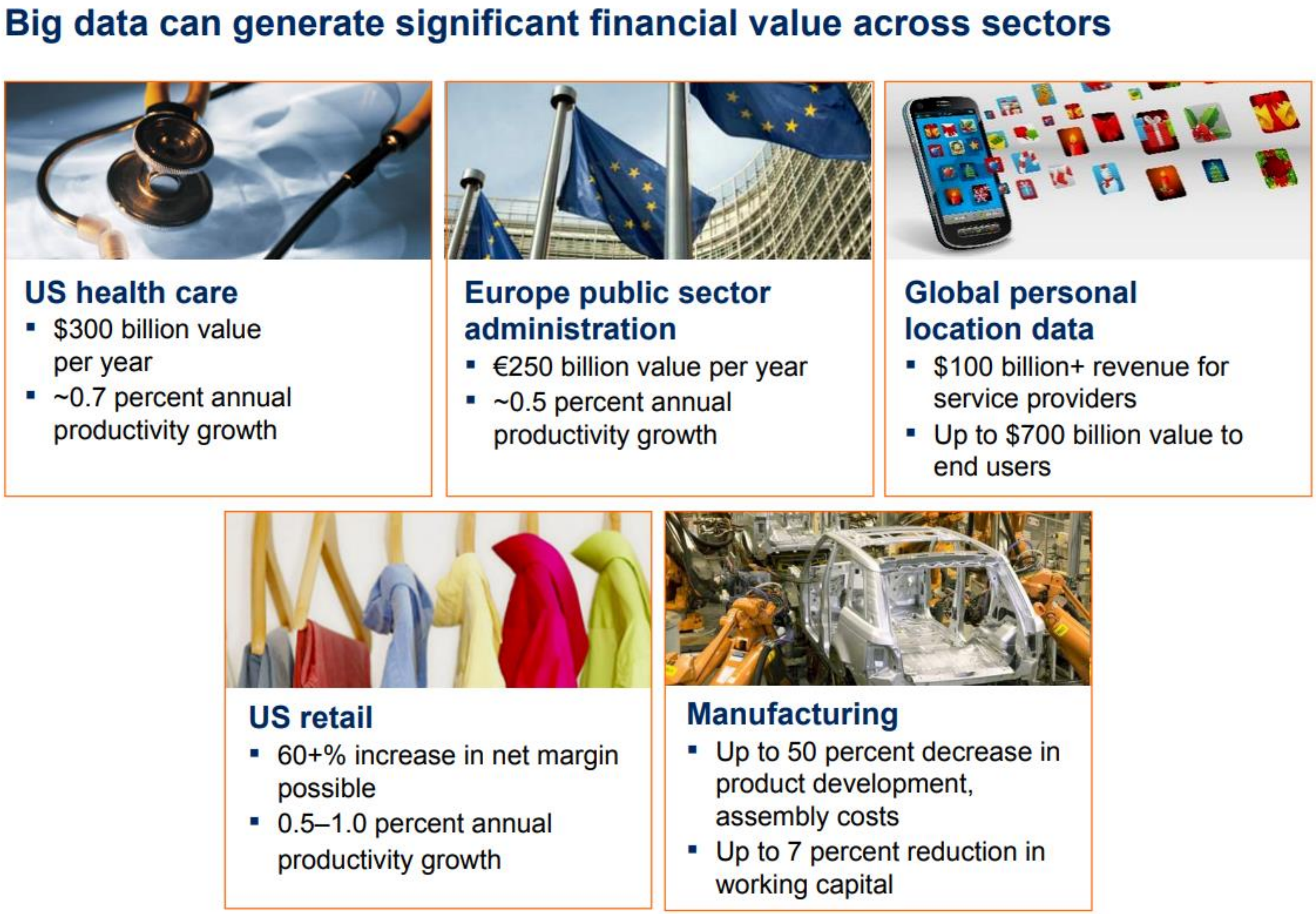
are shared on Facebook every month

400 MILLION TWEETS

are sent per day by about 200 million monthly active users

Data Analytics: The Era of "Big Data"

- Value in Big Data



Big data can generate significant financial value across sectors

US health care

- $300 billion value per year
- ~0.7 percent annual productivity growth

Europe public sector administration

- €250 billion value per year
- ~0.5 percent annual productivity growth

Global personal location data

- $100 billion+ revenue for service providers
- Up to $700 billion value to end users

US retail

- 60+% increase in net margin possible
- 0.5–1.0 percent annual productivity growth

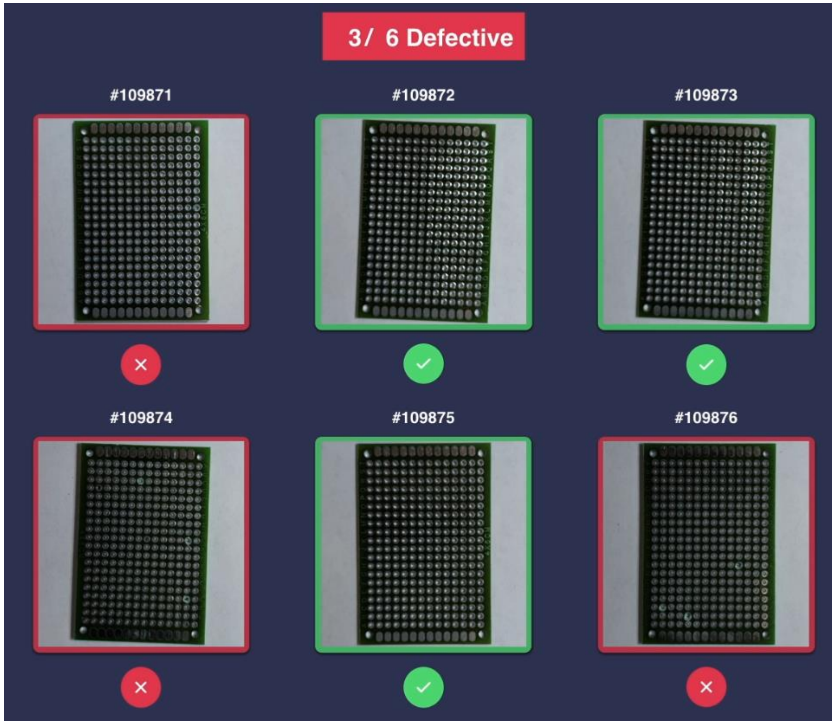
Manufacturing

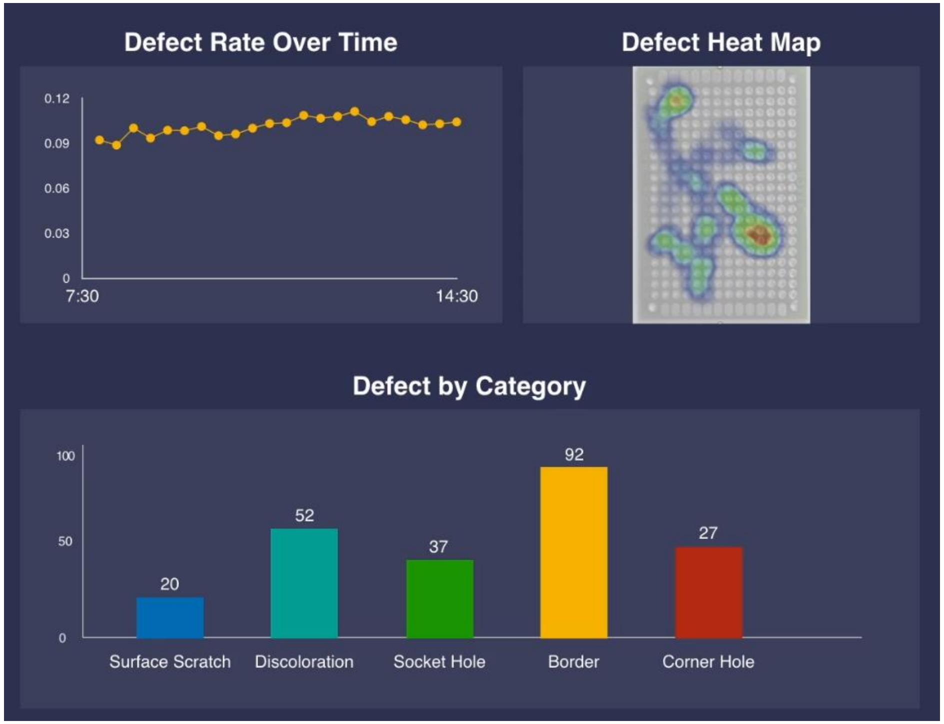
- Up to 50 percent decrease in product development, assembly costs
- Up to 7 percent reduction in working capital

SOURCE: McKinsey Global Institute analysis

Data Analytics

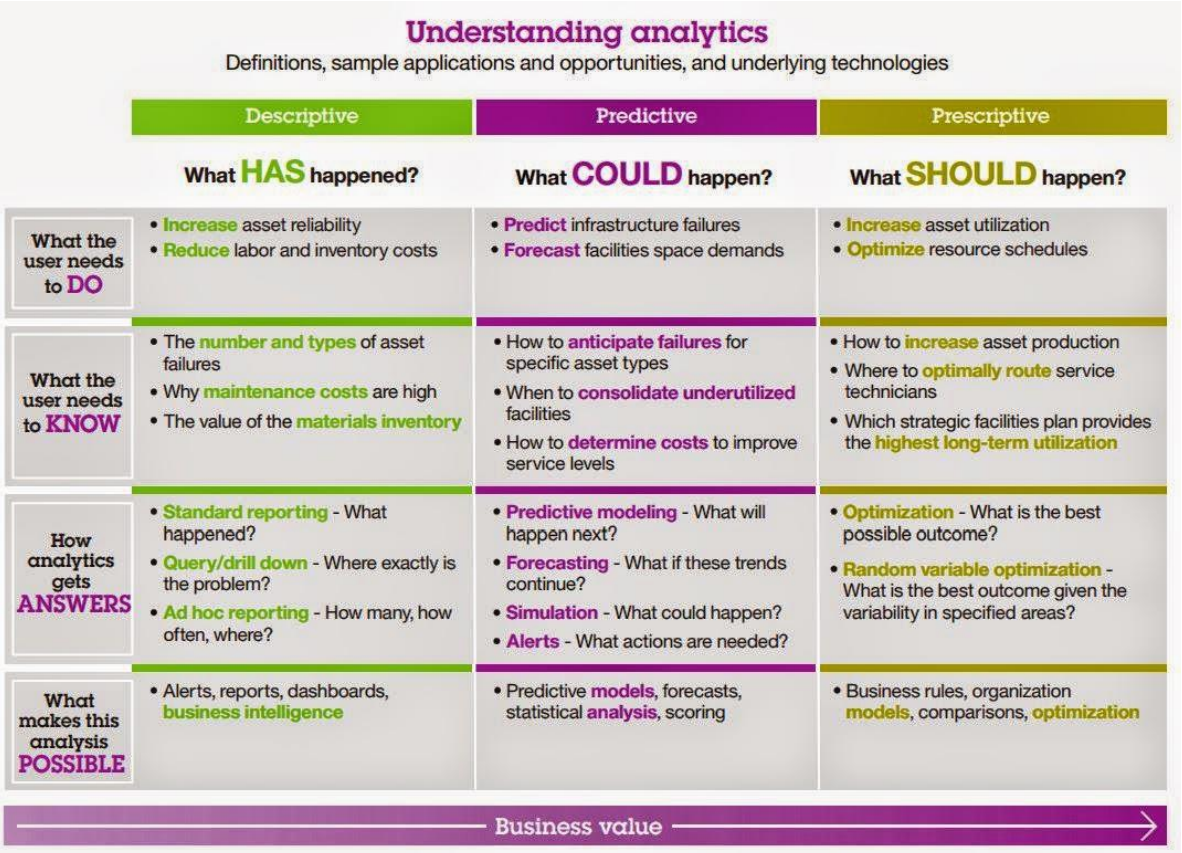
- Data Analytics in Industrial Engineering (Manufacturing)
  - ✓ Landing.ai: AI Startup found by Andrew Ng
  - ✓ Provide various image/video analytics solutions for fault detection, leak defect detection, etc.





Understanding Analytics

- Descriptive vs. Predictive vs. Prescriptive Analytics



Understanding analytics

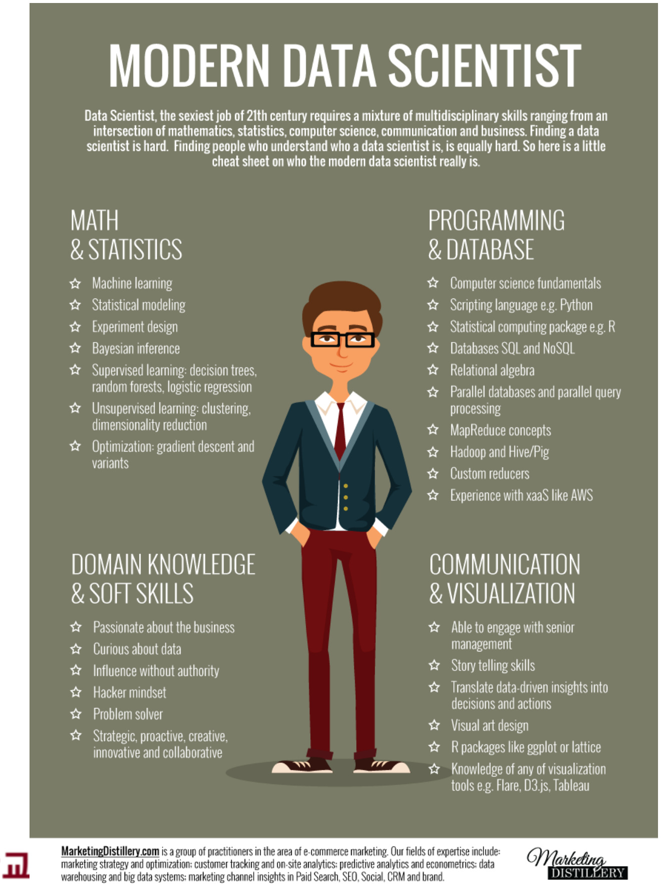
Definitions, sample applications and opportunities, and underlying technologies

	Descriptive	Predictive	Prescriptive
	What HAS happened?	What COULD happen?	What SHOULD happen?
What the user needs to DO	• Increase asset reliability • Reduce labor and inventory costs	• Predict infrastructure failures • Forecast facilities space demands	• Increase asset utilization • Optimize resource schedules
What the user needs to KNOW	• The number and types of asset failures • Why maintenance costs are high • The value of the materials inventory	• How to anticipate failures for specific asset types • When to consolidate underutilized facilities • How to determine costs to improve service levels	• How to increase asset production • Where to optimally route service technicians • Which strategic facilities plan provides the highest long-term utilization
How analytics gets ANSWERS	• Standard reporting - What happened? • Query/drill down - Where exactly is the problem? • Ad hoc reporting - How many, how often, where?	• Predictive modeling - What will happen next? • Forecasting - What if these trends continue? • Simulation - What could happen? • Alerts - What actions are needed?	• 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	• Alerts, reports, dashboards, business intelligence	• Predictive models, forecasts, statistical analysis, scoring	• Business rules, organization models, comparisons, optimization

데이터분석을 위한 프로그래밍 언어 수리통계 및 실습 자료구조 및 알고리즘 응용통계 및 실습	데이터마이닝 다변량분석 예측애널리틱스 영상정보시스템	OR-I 및 실습 OR-II 및 실습 최적화이론 최적화응용 메타휴리스틱

Data Scientist

- Data Scientist: The Sexiest Job of the 21st Century
  - ✓ Harvard Business Review
  - ✓ https://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century



MODERN DATA SCIENTIST

Data Scientist, the sexiest job of 21th 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

- ☆ 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

- ☆ 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

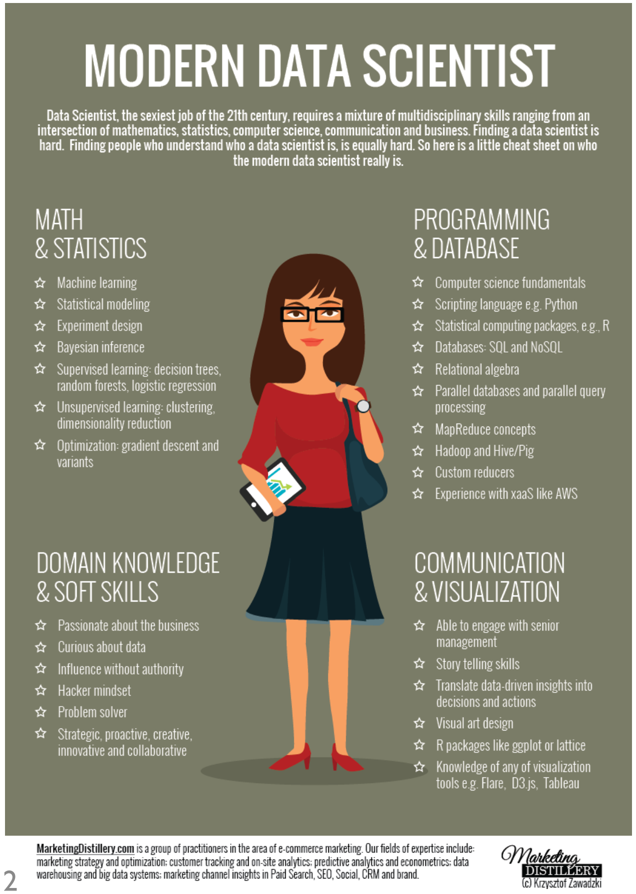
- ☆ Passionate about the business
- ☆ Curious about data
- ☆ Influence without authority
- ☆ Hacker mindset
- ☆ Problem solver
- ☆ Strategic, proactive, creative, innovative and collaborative

COMMUNICATION & VISUALIZATION

- ☆ 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



MODERN DATA SCIENTIST

Data Scientist, the sexiest job of the 21th 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

- ☆ 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

- ☆ Computer science fundamentals
- ☆ Scripting language e.g. Python
- ☆ Statistical computing packages, 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

- ☆ Passionate about the business
- ☆ Curious about data
- ☆ Influence without authority
- ☆ Hacker mindset
- ☆ Problem solver
- ☆ Strategic, proactive, creative, innovative and collaborative

COMMUNICATION & VISUALIZATION

- ☆ 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

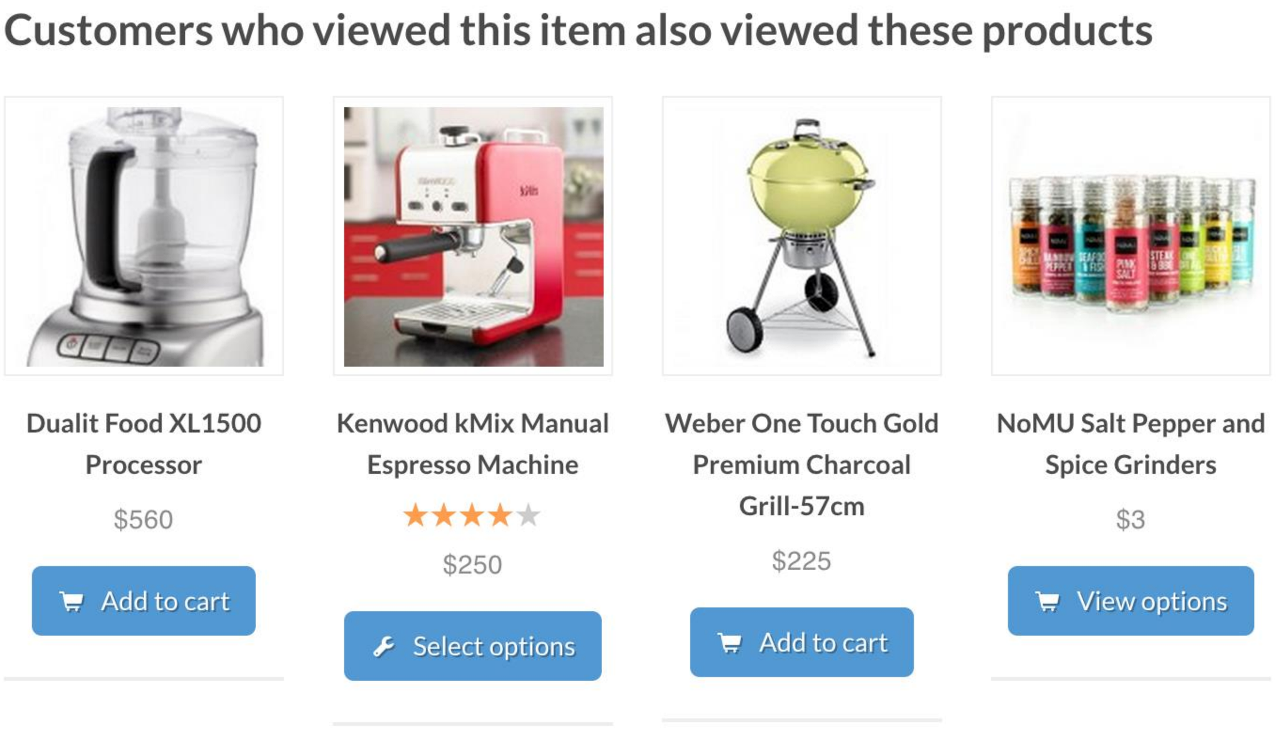
MarketingDistillery.com is a group of practitioners in the area of e-commerce marketing. Our fields of expertise include: marketing strategy and optimization; customer tracking and on-site analytics; predictive analytics and econometrics; data warehousing and big data systems; marketing channel insights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

Marketing DISTILLERY

(c) Krzysztof Zawadzki

Data Science Tasks

- Data Mining
  - ✓ 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)



Customers who viewed this item also viewed these products

Dualit Food XL1500 Processor

$560

Kenwood kMix Manual Espresso Machine

★★★★☆

$250

Weber One Touch Gold Premium Charcoal Grill-57cm

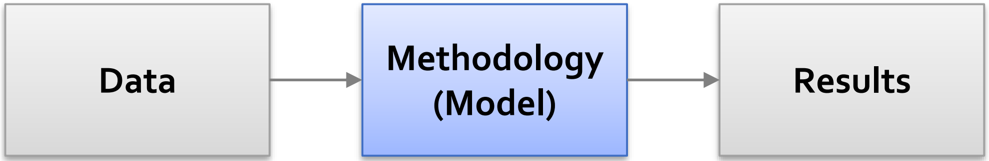
$225

NoMU Salt Pepper and Spice Grinders

$3

Data Science Tasks

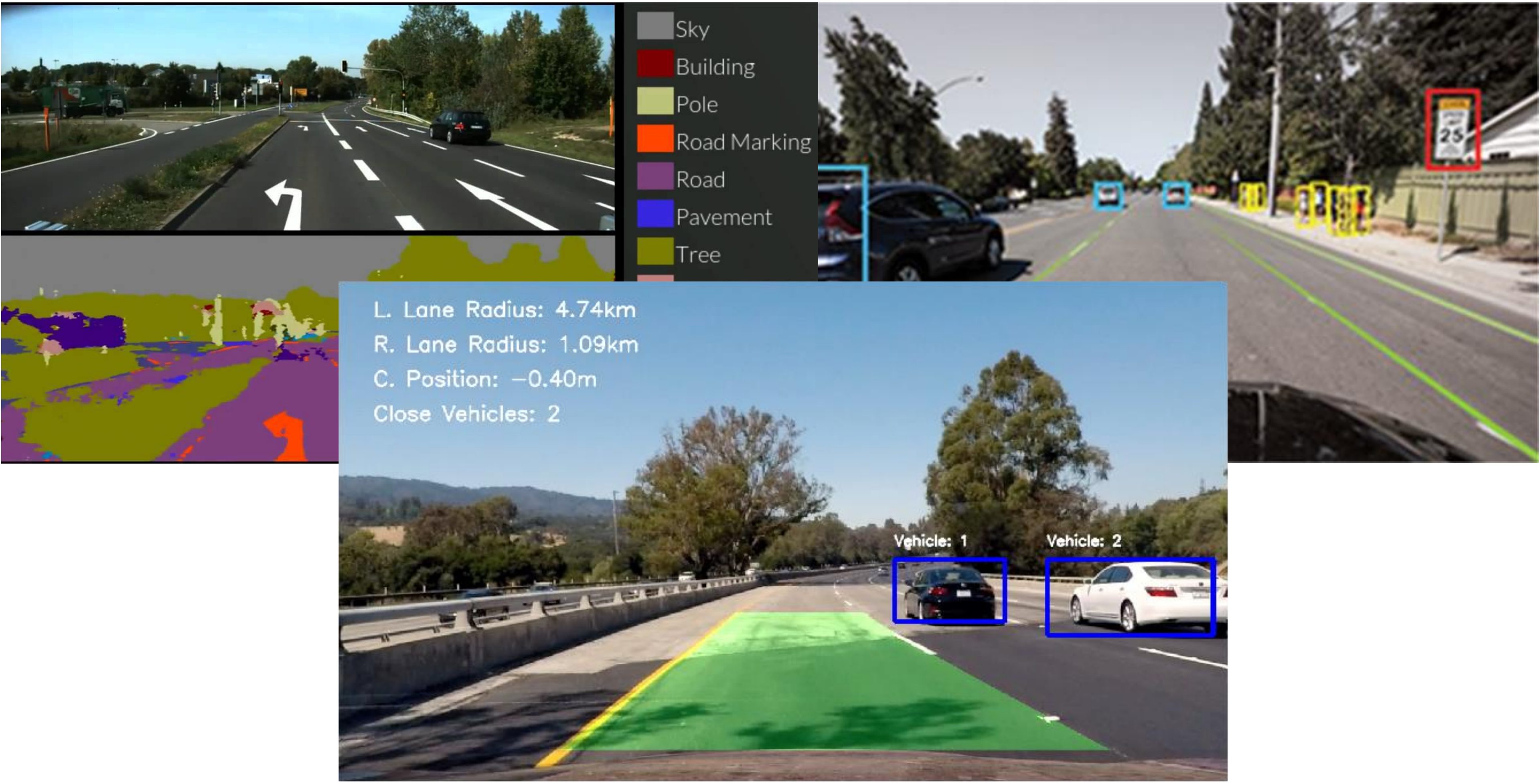
- Machine Learning
  - ✓ A computer program is said to learn from <u>experience</u> <u>**E**</u> with respect to some class of <u>tasks</u> <u>**T**</u> and <u>performance measure</u> <u>**P**</u>, if its performance at task in T, as measured by P, improves with experience E," – Mitchell et al. (2013)



Data → Methodology (Model) → Results

Data Science Tasks

- Machine Learning Models in Self-Driving Cars



Data Science Tasks

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

AGENDA

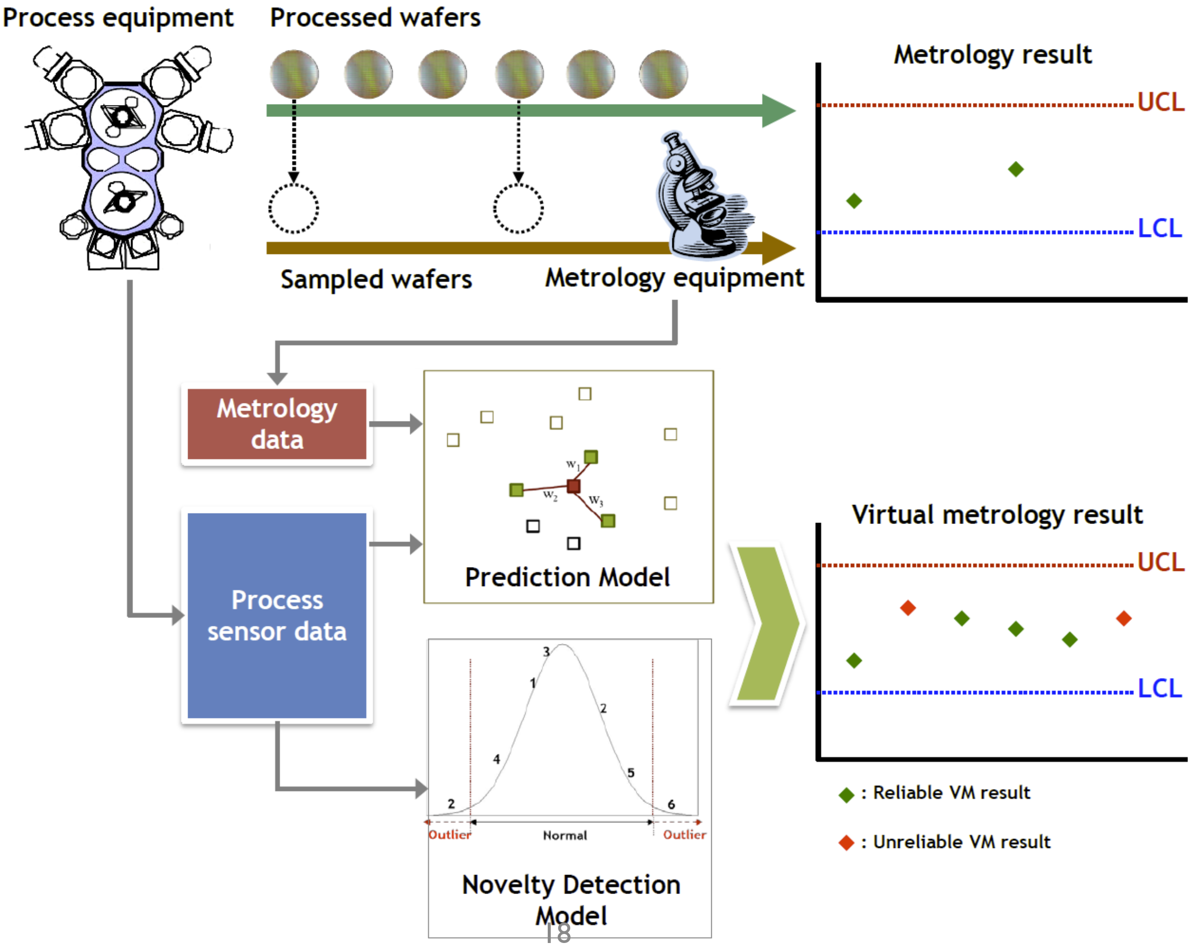
01 What is Data Analytics?

02 What Can We Do with Data Analytics?

03 Data Analytics Languages: R & Python

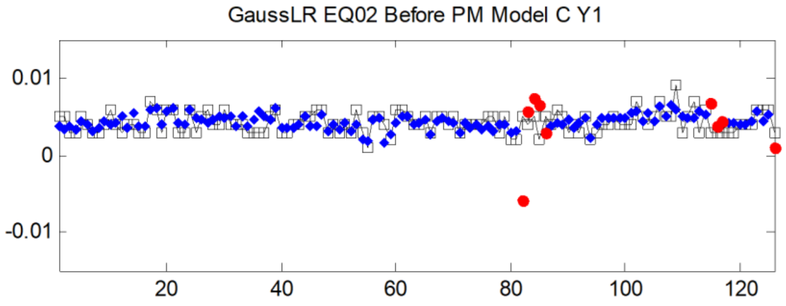
Process Monitoring & Control

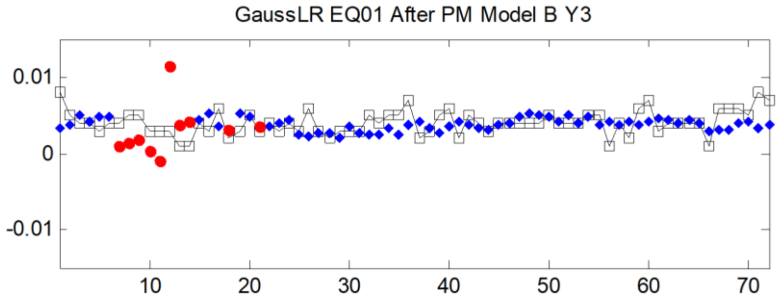
Input Data	Algorithms	Output
Sensor Data from Production Equipment	Regression + Novelty Detection	Metrological Values With Reliability Scores

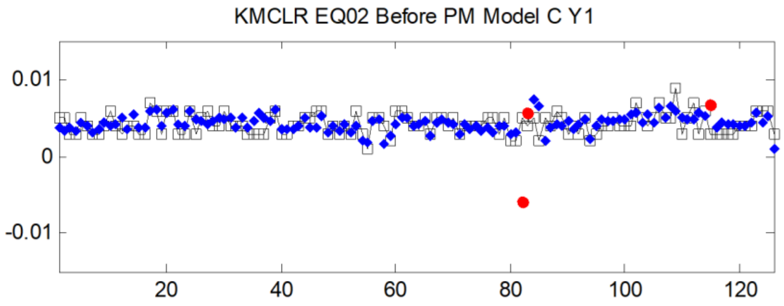


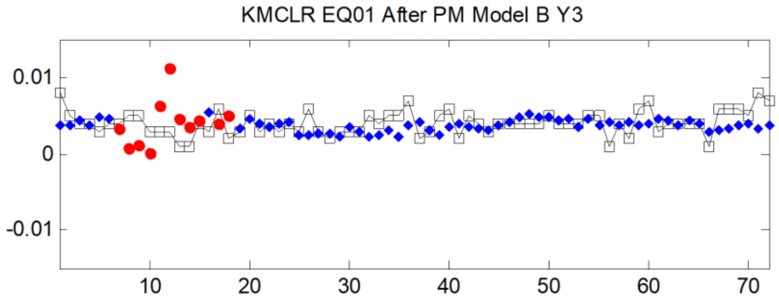
Process Monitoring & Control

Input Data	Algorithms	Output
Sensor Data from Production Equipment	Regression + Novelty Detection	Metrological Values With Reliability Scores


GaussLR EQ02 Before PM Model C Y1


GaussLR EQ01 After PM Model B Y3


KMCLR EQ02 Before PM Model C Y1

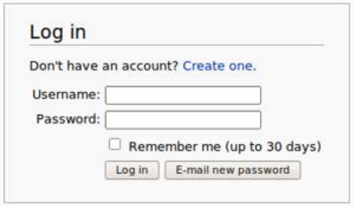

KMCLR EQ01 After PM Model B Y3

Are You a Valid User?

Input Data	Algorithms	Output
Time stamps collected during typing	Novelty Detection	Valid user score

Through various input devices

In any stages



Log in

Don't have an account? Create one.

Username:

Password:

☐ Remember me (up to 30 days)

Log in

E-mail new password

오후 5:54

그룹채팅 (6명)

와 어디서 저녁먹어요?

박준형: 여기가 어디여?

이민영: 맛있겠당

오늘 오픈한 중국식당에 왔어요. 매운갈비찜을 시켰습니다. :-)

김은희

오후 6:13

박준형

박준형: 경치 좋지?

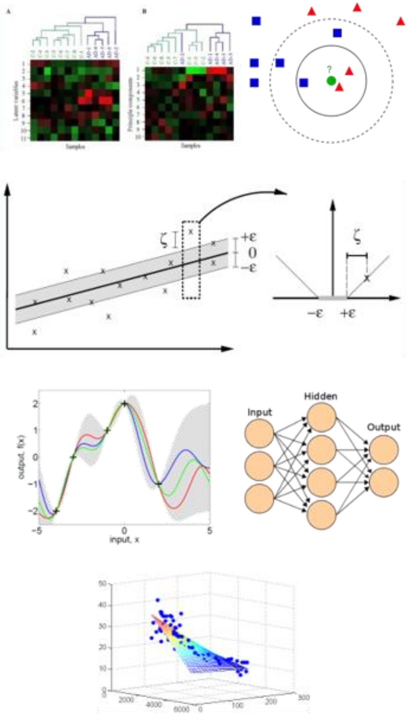
박준형: 여기를 가보셈

새롬이 연락처 알고 있으면 보내주셈

박준형: 한새롬

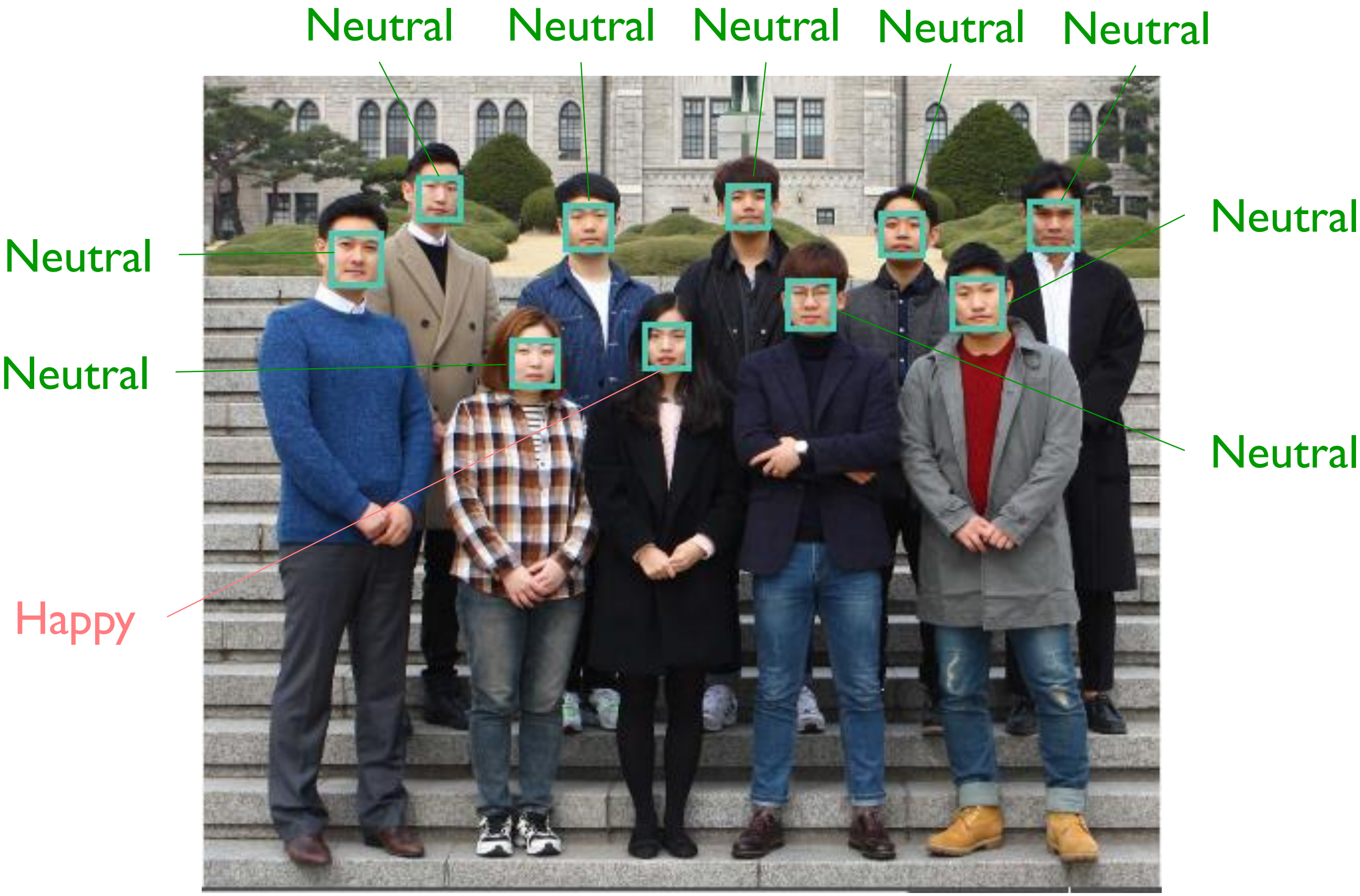
기동이 연락처도 부탁~~ :-)

Customized (real-time) authenticator



Who Looks Happy?

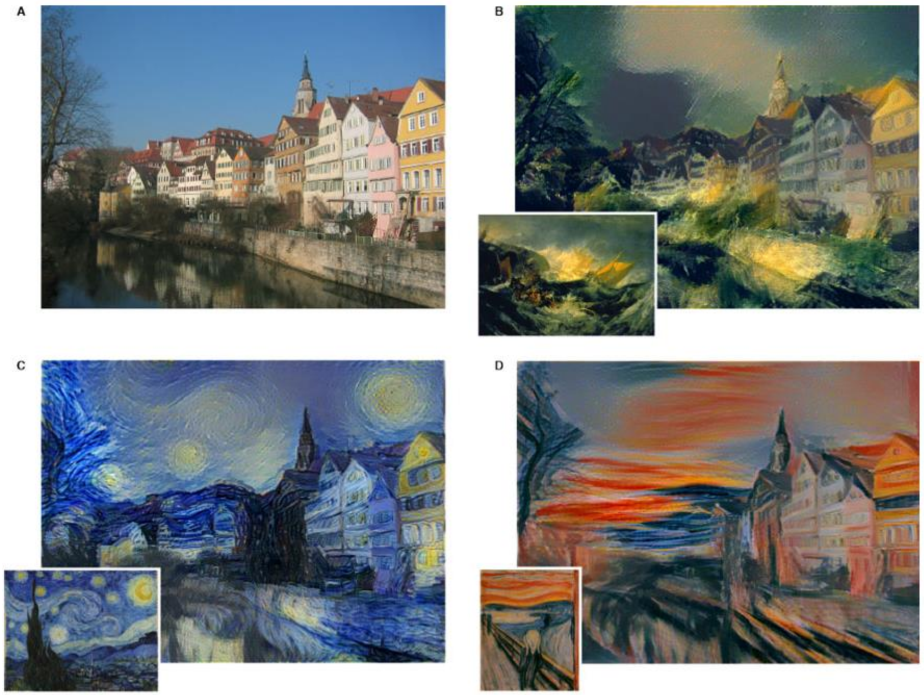
Input Data	Algorithms	Output
Image	Convolutional Neural Networks	Emotions Recognized

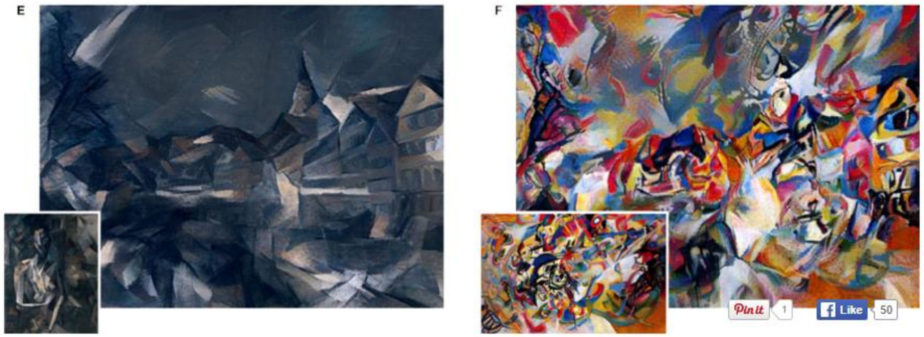


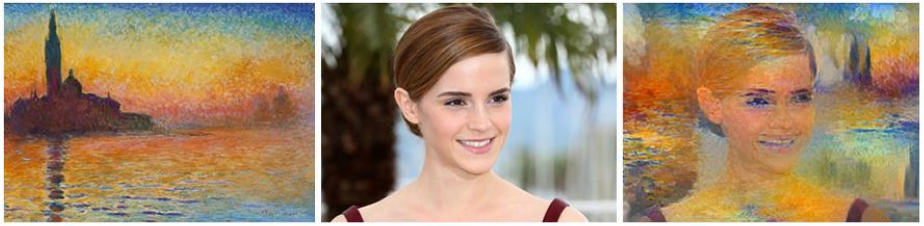
https://www.microsoft.com/cognitive-services/en-us/emotion-api

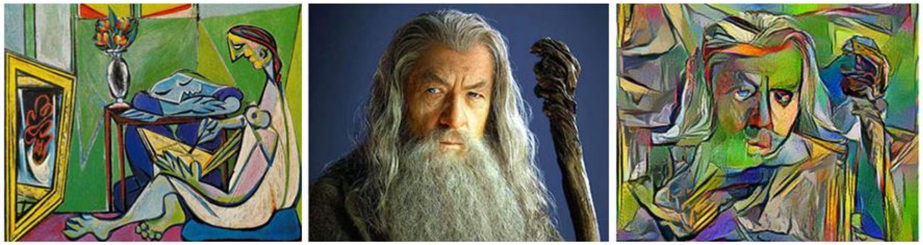
Favorite Artistic Style

Input Data	Algorithms	Output
Image	Convolutional Neural Networks	Image with Preferred Style





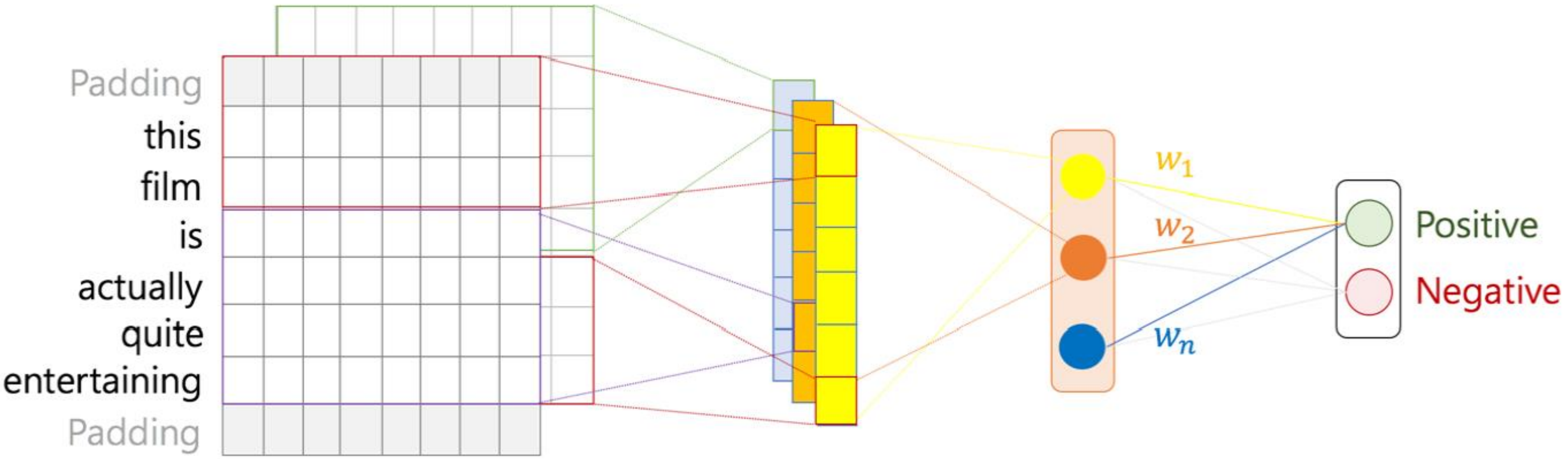


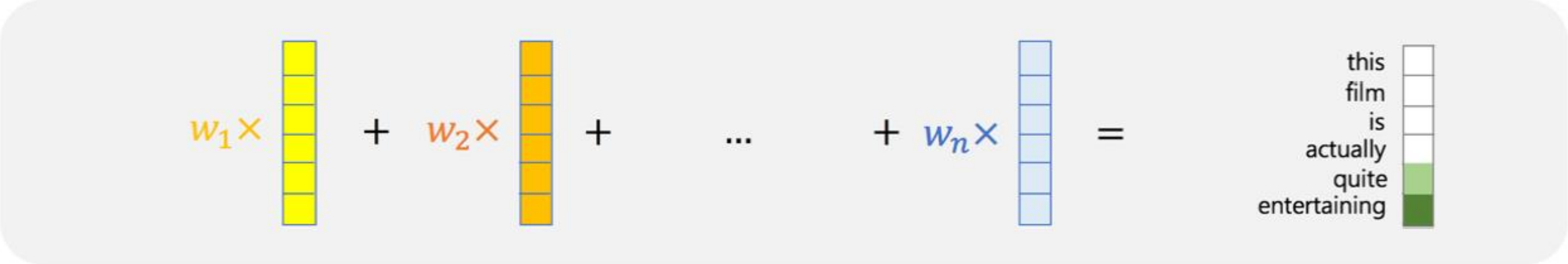


Gatys, L. A., Ecker, A. S., & Bethge, M. (2015). A neural algorithm of artistic style. arXiv preprint arXiv:1508.06576.

Understanding Emotions

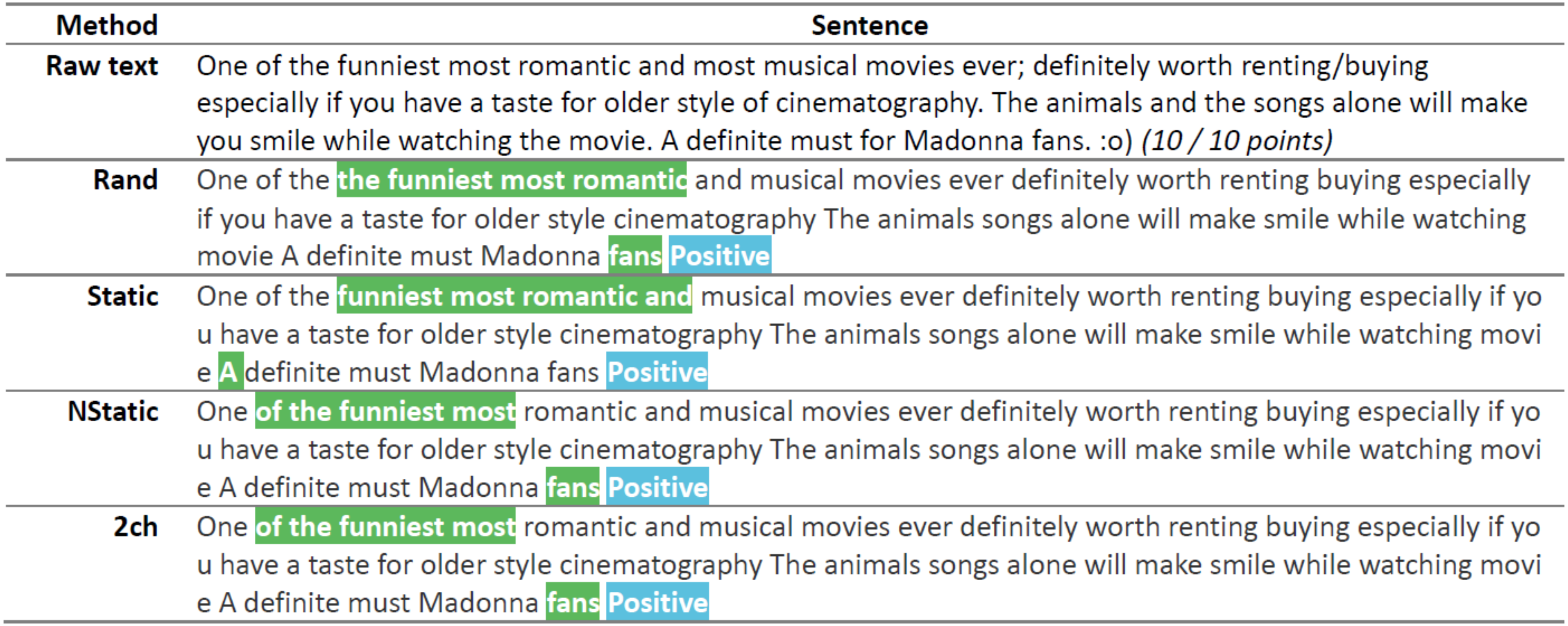
Input Data	Algorithms	Output
Movie Review Text	Convolutional Neural Networks	Sentiment (Pos/Neg) & Keyword attention





Understanding Emotions

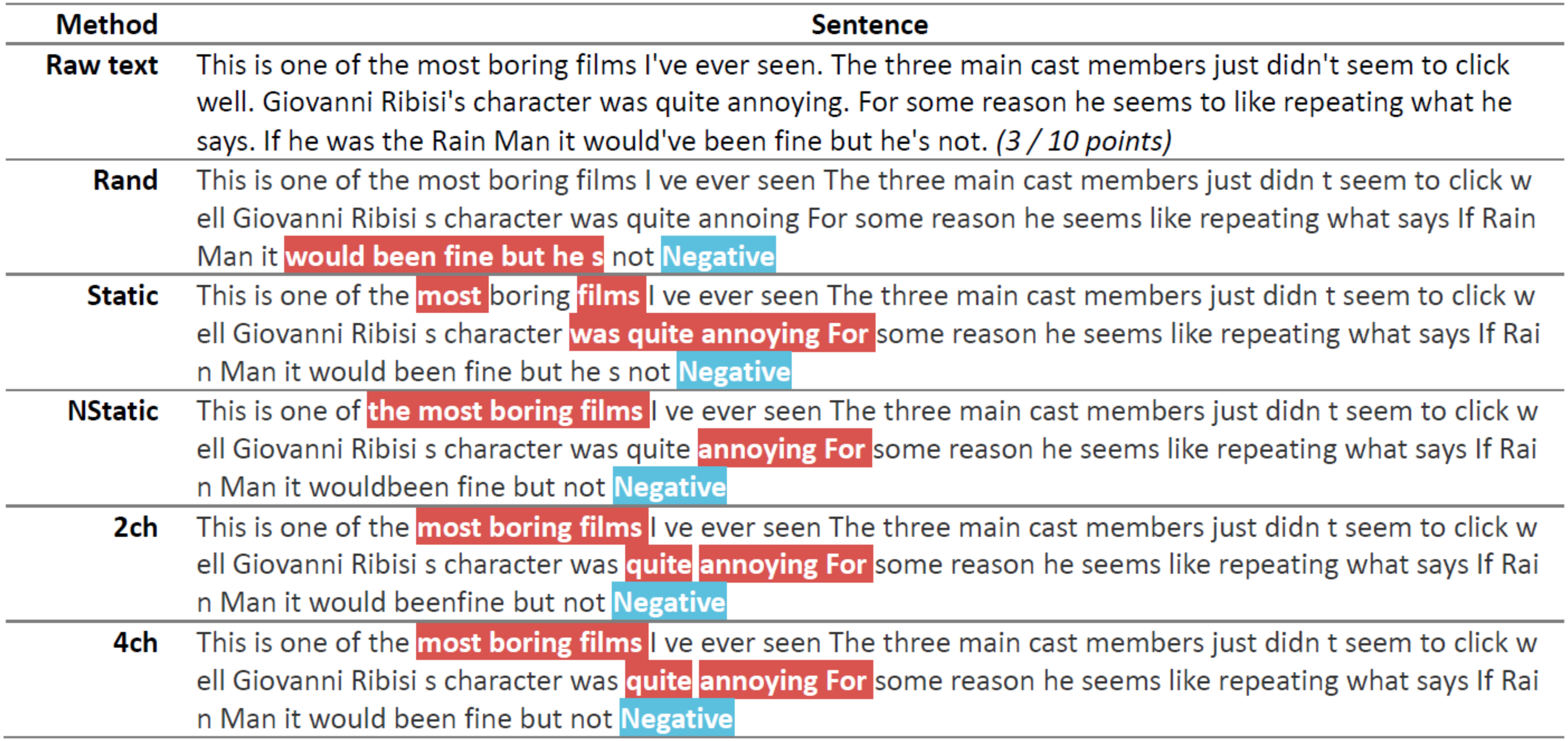
Input Data	Algorithms	Output
Movie Review Text	Convolutional Neural Networks	Sentiment (Pos/Neg) & Keyword attention



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 you have a taste for older style cinematography The animals songs alone will make smile while watching movie **A** definite must Madonna fans **Positive**
NStatic	One **of 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**
2ch	One **of 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**

Understanding Emotions

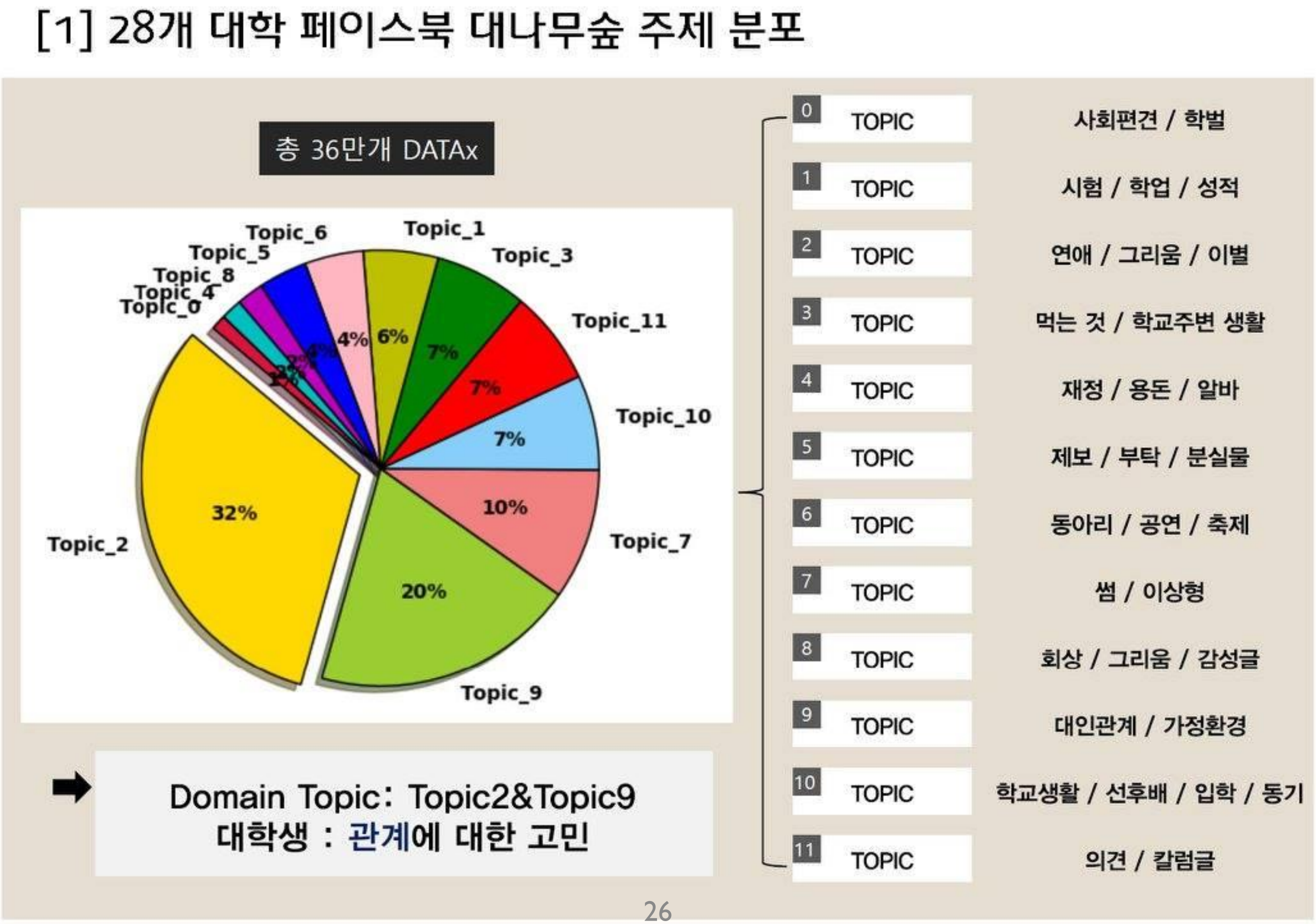
Input Data	Algorithms	Output
Movie Review Text	Convolutional Neural Networks	Sentiment (Pos/Neg) & Keyword attention



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 w ell Giovanni Ribisi s character was quite annoing For some reason he seems like repeating what says If Rain Man it **would been 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 w ell Giovanni Ribisi s character **was quite annoying For** some reason he seems like repeating what says If Rai n Man it would been 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 w ell Giovanni Ribisi s character was quite **annoying For** some reason he seems like repeating what says If Rai n Man it wouldbeen 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 w ell Giovanni Ribisi s character was **quite** **annoying For** some reason he seems like repeating what says If Rai n Man it would beenfine 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 w ell Giovanni Ribisi s character was **quite** **annoying For** some reason he seems like repeating what says If Rai n Man it would been fine but not **Negative**

What are the Main Topics in Anonymous Posts of University Students?

Input Data	Algorithms	Output
Facebook Posts (Bamboo forest)	Latent Dirichlet Allocation (Topic Modeling)	Topic Distribution



[1] 28개 대학 페이스북 대나무숲 주제 분포

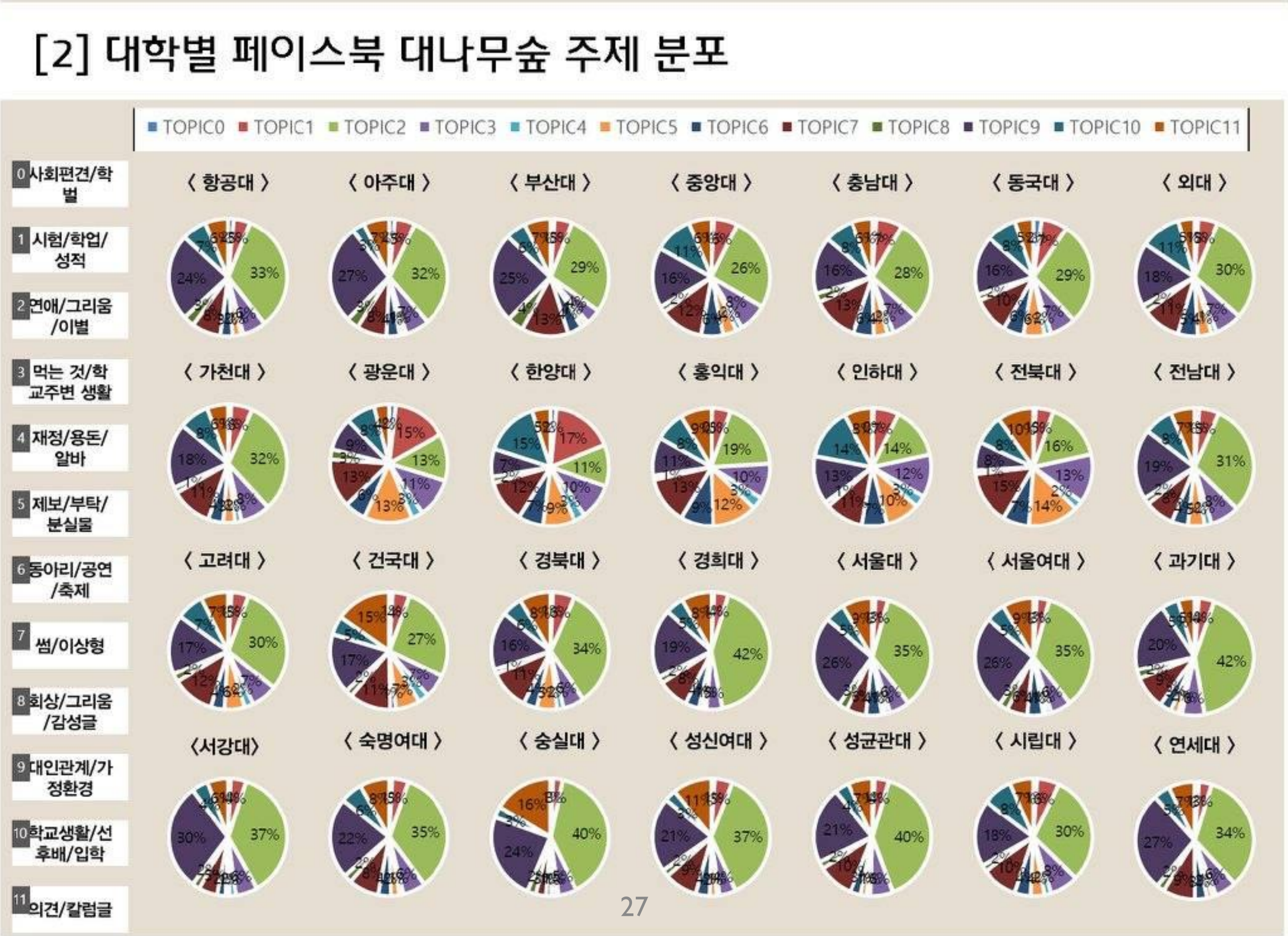
총 36만개 DATAx

0	TOPIC	사회편견 / 학벌
1	TOPIC	시험 / 학업 / 성적
2	TOPIC	연애 / 그리움 / 이별
3	TOPIC	먹는 것 / 학교주변 생활
4	TOPIC	재정 / 용돈 / 알바
5	TOPIC	제보 / 부탁 / 분실물
6	TOPIC	동아리 / 공연 / 축제
7	TOPIC	썸 / 이상형
8	TOPIC	회상 / 그리움 / 감성글
9	TOPIC	대인관계 / 가정환경
10	TOPIC	학교생활 / 선후배 / 입학 / 동기
11	TOPIC	의견 / 칼럼글

➡ Domain Topic: Topic2&Topic9
대학생 : 관계에 대한 고민

What are the Main Topics in Anonymous Posts of University Students?

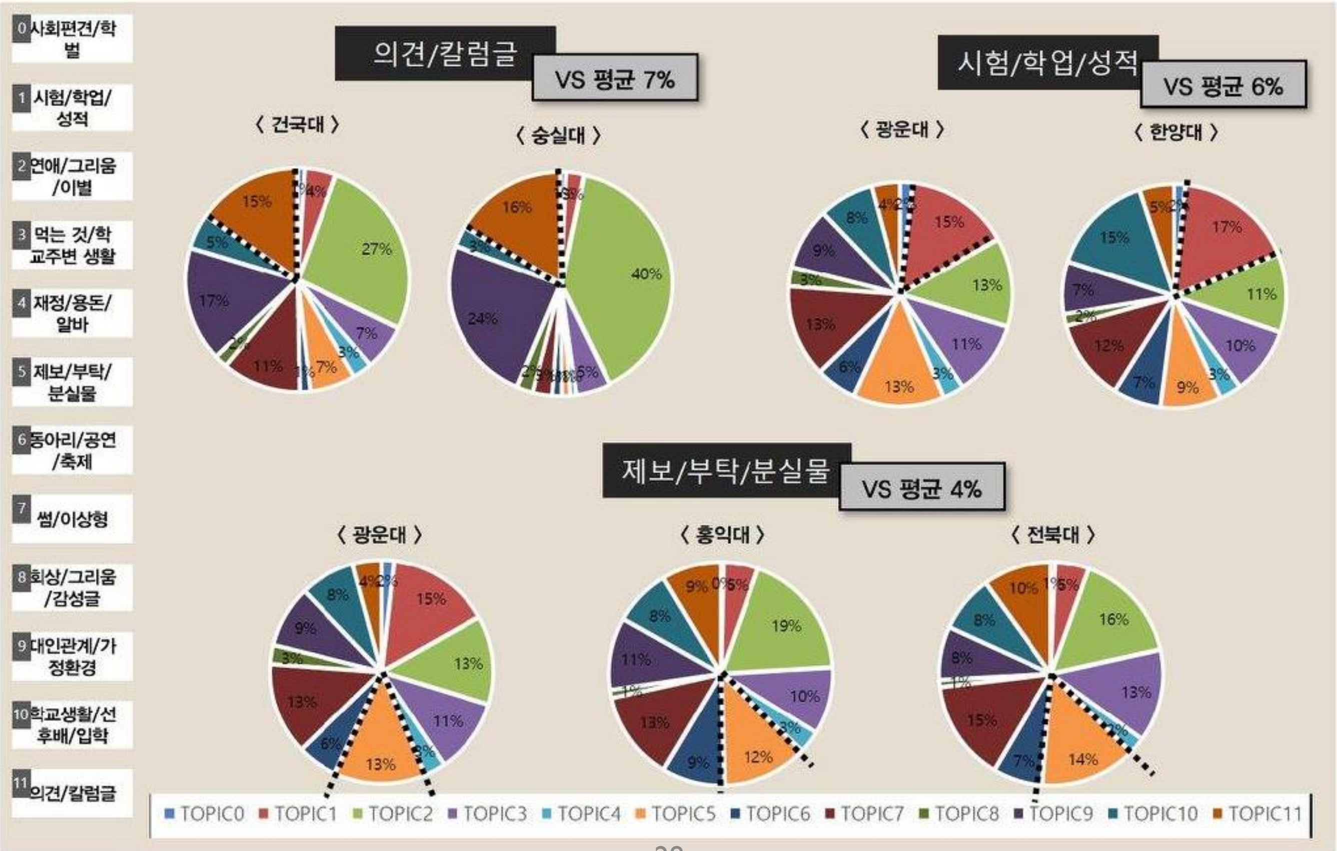
Input Data	Algorithms	Output
Facebook Posts (Bamboo forest)	Latent Dirichlet Allocation (Topic Modeling)	Topic Distribution



What are the Main Topics in Anonymous Posts of University Students?

Input Data	Algorithms	Output
Facebook Posts (Bamboo forest)	Latent Dirichlet Allocation (Topic Modeling)	Topic Distribution

[3] 대학별 특이 사항



Connecting the Dots

Connecting the Dots

You can't connect the dots looking forward; you can only connect them looking backwards.

So you have to trust that the dots will somehow connect in your future.

AGENDA

01 What is Data Analytics?

02 What Can We Do with Data Analytics?

03 Data Analytics Languages: R & Python

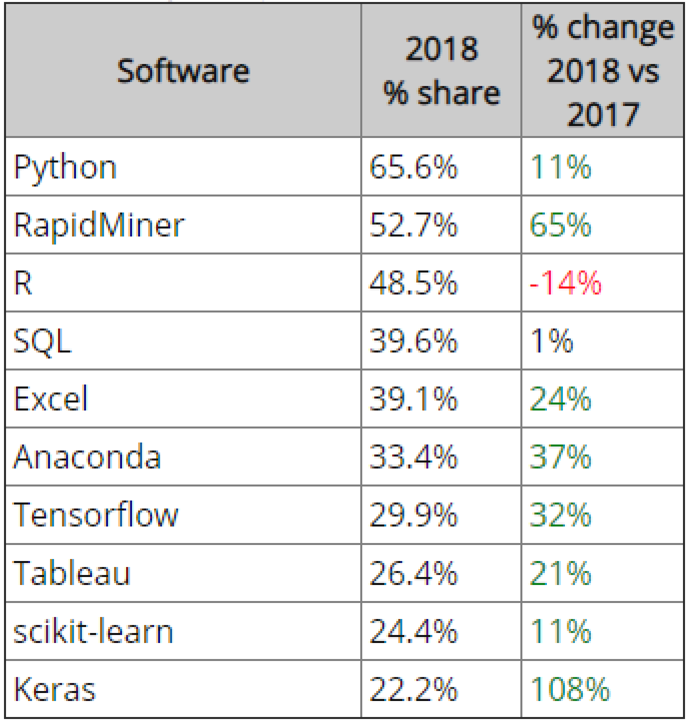
R vs. Python

- Learn Data Science, not Programming
  - ✓ R vs. Python, different brushes
  - ✓ Do not choose between R & Python, learn both
  - ✓ It strengthens your data science communication skills
  - ✓ It boosts your data science career
  - ✓ It is not that hard

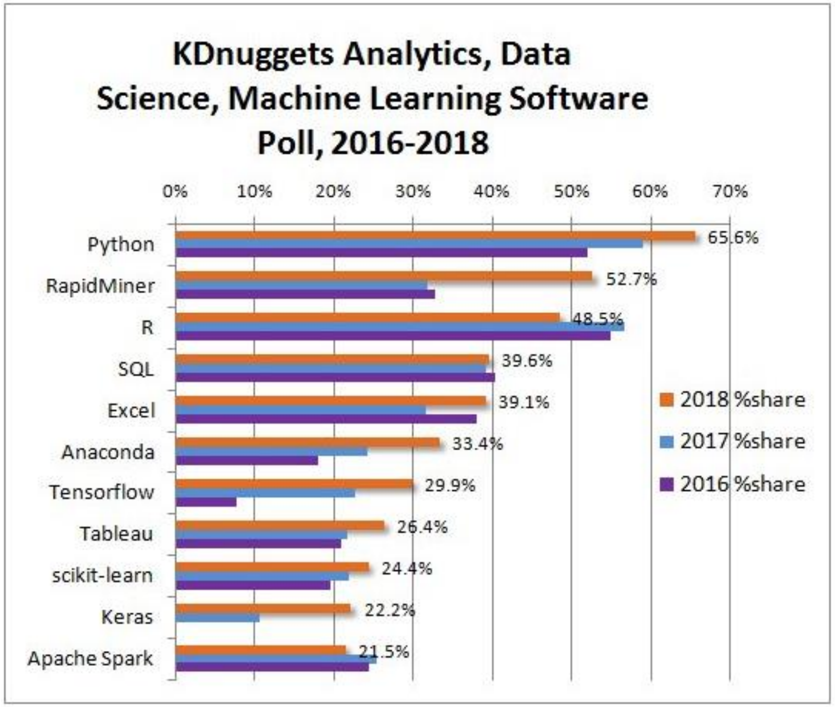
R vs. Python

- Some interesting polls (2018)

Table 1: Top Analytics/Data Science/ML Software in 2018 KDnuggets Poll

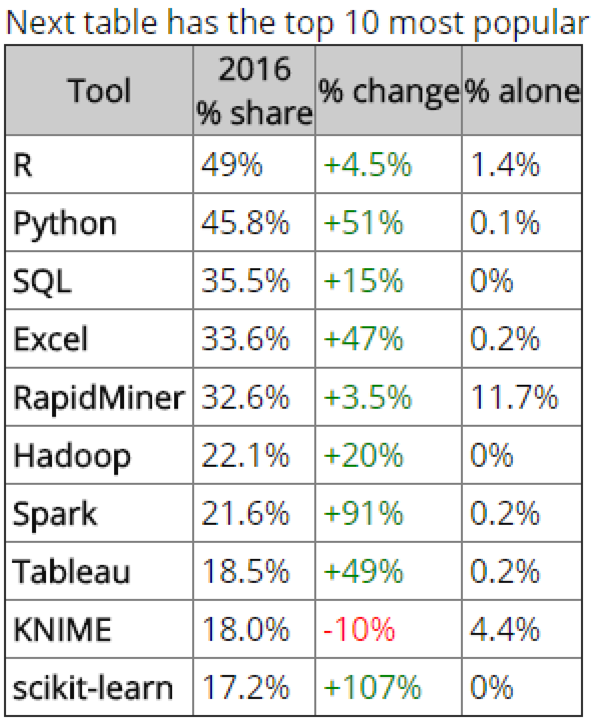


Software	2018 % share	% change 2018 vs 2017
Python	65.6%	11%
RapidMiner	52.7%	65%
R	48.5%	-14%
SQL	39.6%	1%
Excel	39.1%	24%
Anaconda	33.4%	37%
Tensorflow	29.9%	32%
Tableau	26.4%	21%
scikit-learn	24.4%	11%
Keras	22.2%	108%



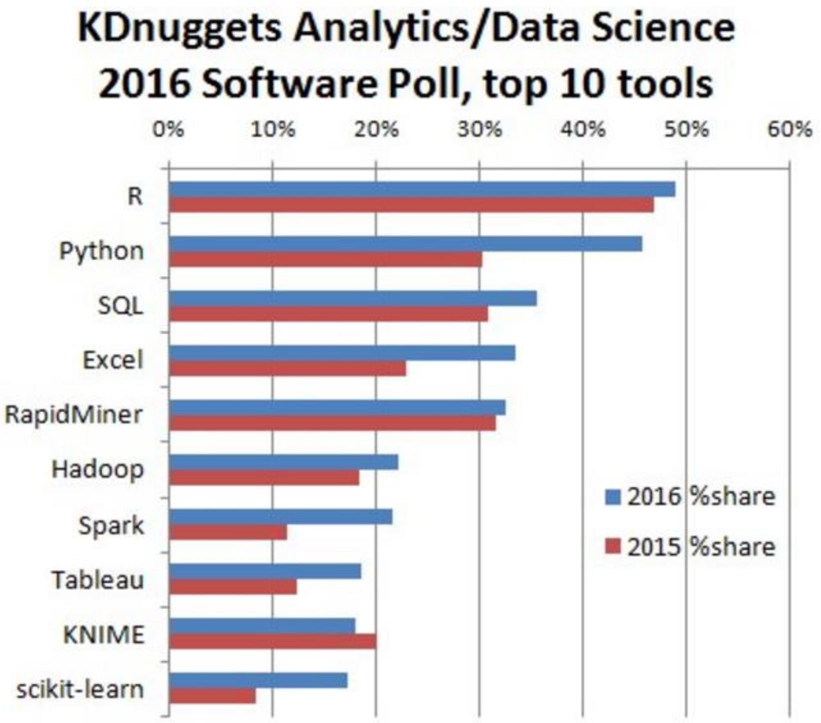
R vs. Python

- Some interesting polls (2016)



Next table has the top 10 most popular tools in 2016 poll

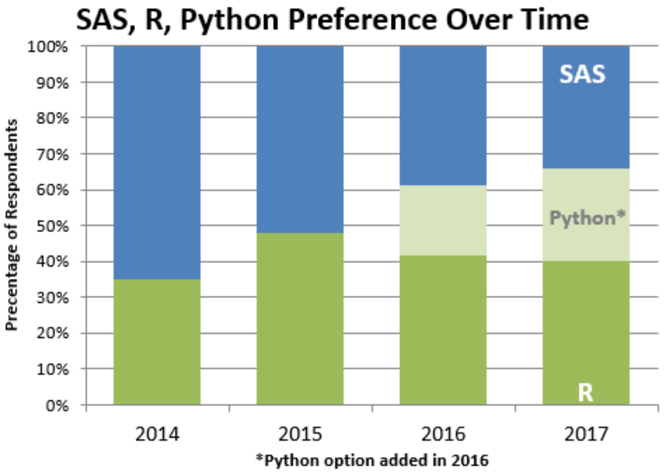
Tool	2016 % share	% change	% alone
R	49%	+4.5%	1.4%
Python	45.8%	+51%	0.1%
SQL	35.5%	+15%	0%
Excel	33.6%	+47%	0.2%
RapidMiner	32.6%	+3.5%	11.7%
Hadoop	22.1%	+20%	0%
Spark	21.6%	+91%	0.2%
Tableau	18.5%	+49%	0.2%
KNIME	18.0%	-10%	4.4%
scikit-learn	17.2%	+107%	0%

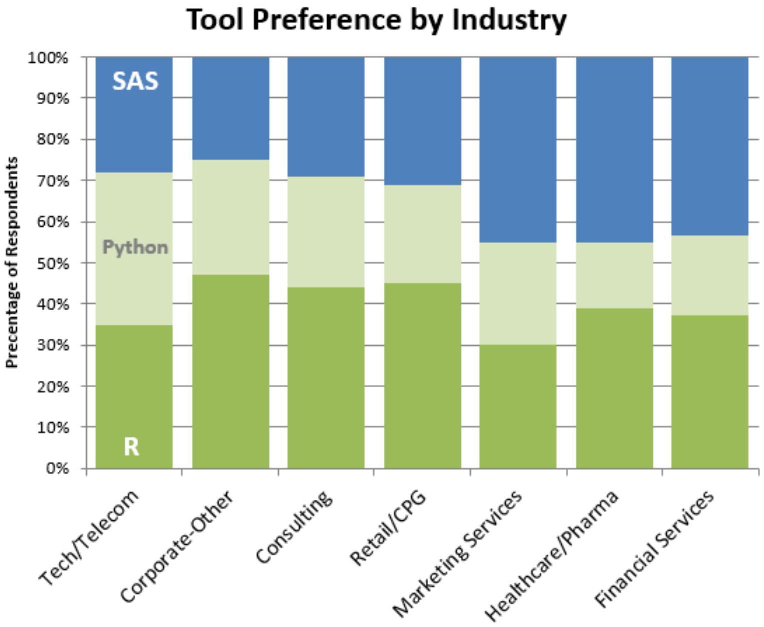


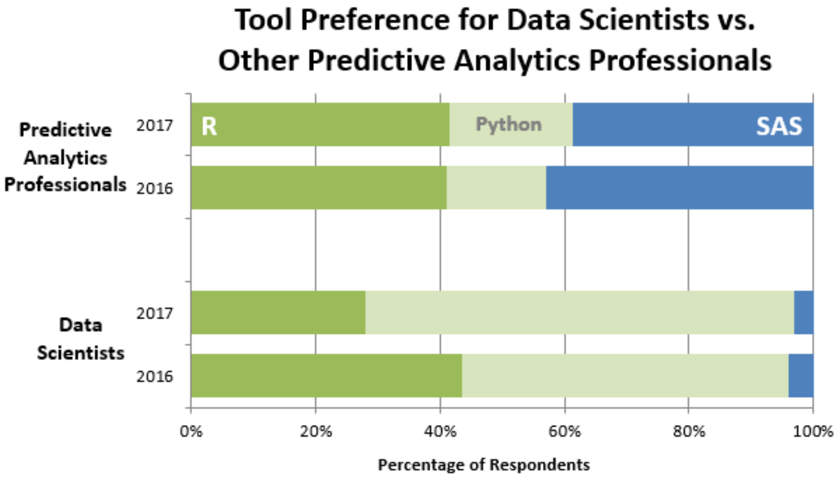
KDnuggets Analytics/Data Science 2016 Software Poll, top 10 tools

R vs. Python

- Some interesting polls

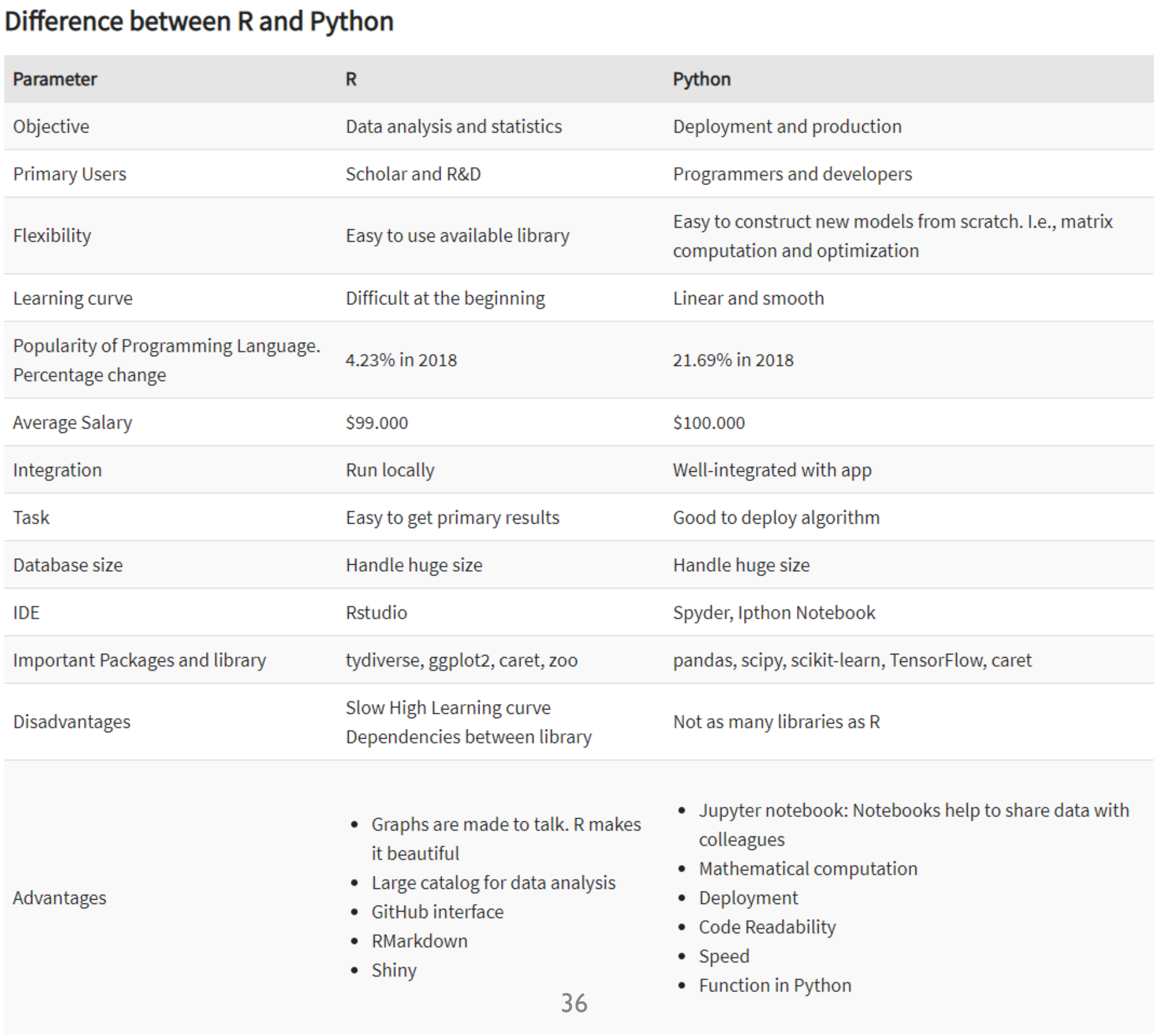






R vs. Python

- Difference between R and Python

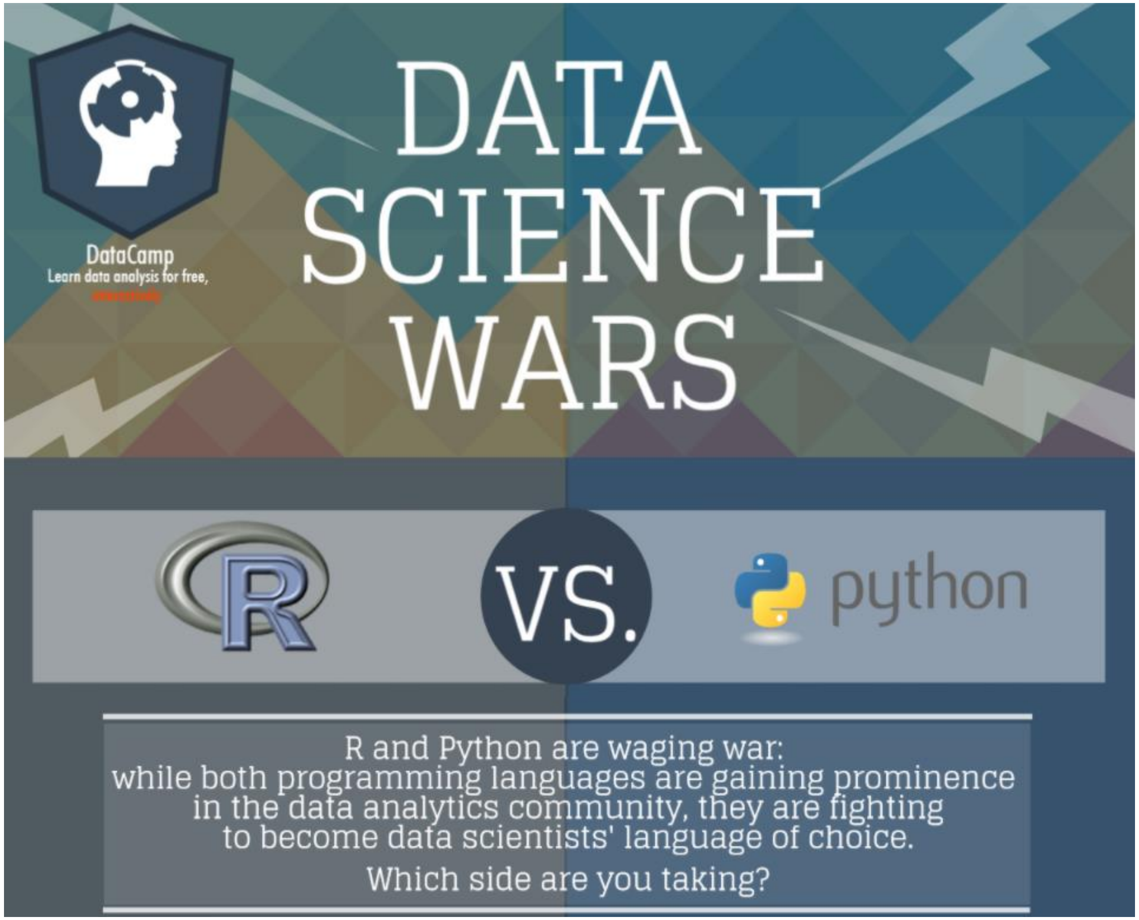


Difference between R and Python

Parameter	R	Python
Objective	Data analysis and statistics	Deployment and production
Primary Users	Scholar and R&D	Programmers and developers
Flexibility	Easy to use available library	Easy to construct new models from scratch. I.e., matrix computation and optimization
Learning curve	Difficult at the beginning	Linear and smooth
Popularity of Programming Language. Percentage change	4.23% in 2018	21.69% in 2018
Average Salary	$99.000	$100.000
Integration	Run locally	Well-integrated with app
Task	Easy to get primary results	Good to deploy algorithm
Database size	Handle huge size	Handle huge size
IDE	Rstudio	Spyder, Ipthon Notebook
Important Packages and library	tydiverse, ggplot2, caret, zoo	pandas, scipy, scikit-learn, TensorFlow, caret
Disadvantages	Slow High Learning curve Dependencies between library	Not as many libraries as R
Advantages	• Graphs are made to talk. R makes it beautiful • Large catalog for data analysis • GitHub interface • RMarkdown • Shiny	• Jupyter notebook: Notebooks help to share data with colleagues • Mathematical computation • Deployment • Code Readability • Speed • Function in Python

R vs. Python

- Data Science Wars: R vs. Python
  - ✓ https://101.datascience.community/2015/05/12/data-science-wars-r-vs-python/



R and Python are waging war:
while both programming languages are gaining prominence in the data analytics community, they are fighting to become data scientists' language of choice.

Which side are you taking?

ANY
questions?