



Lecture I: 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

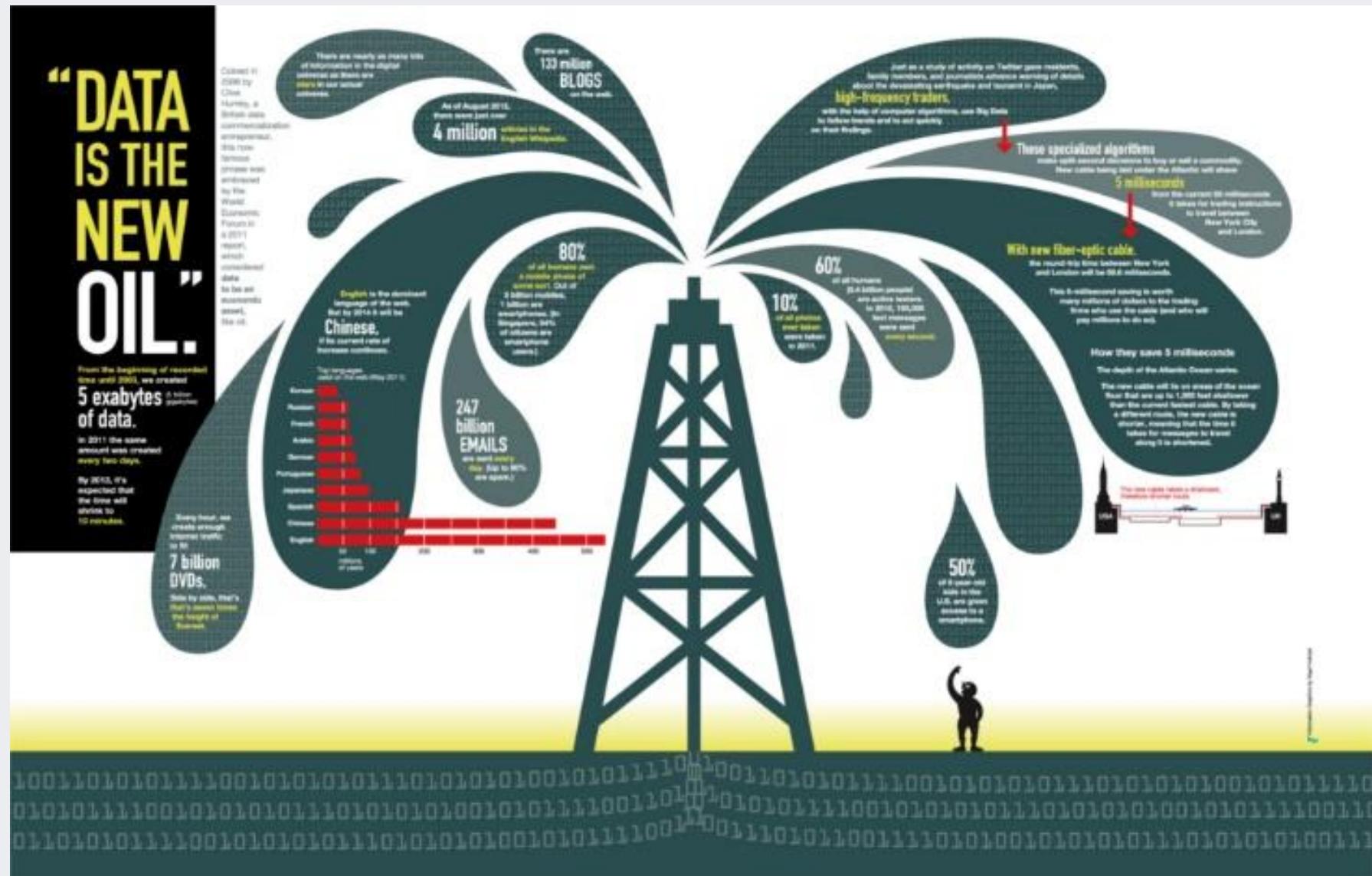
Data Analytics

미래인간
FUTURE HUMAN
AI

MBC



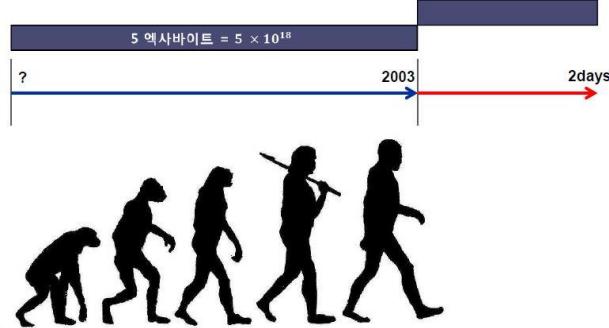
Data Analytics: The Era of “Big Data”



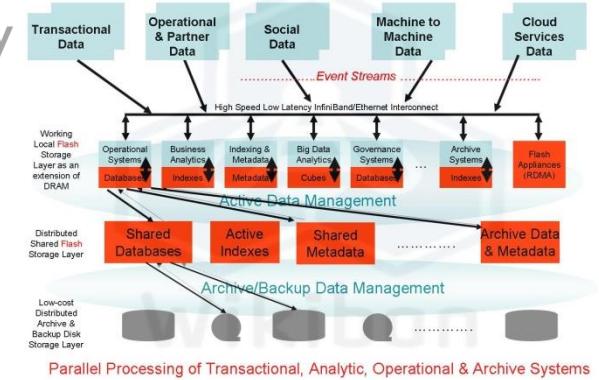
Data Analytics: The Era of “Big Data”

- 4Vs in Big Data

Volume



Velocity



Variety

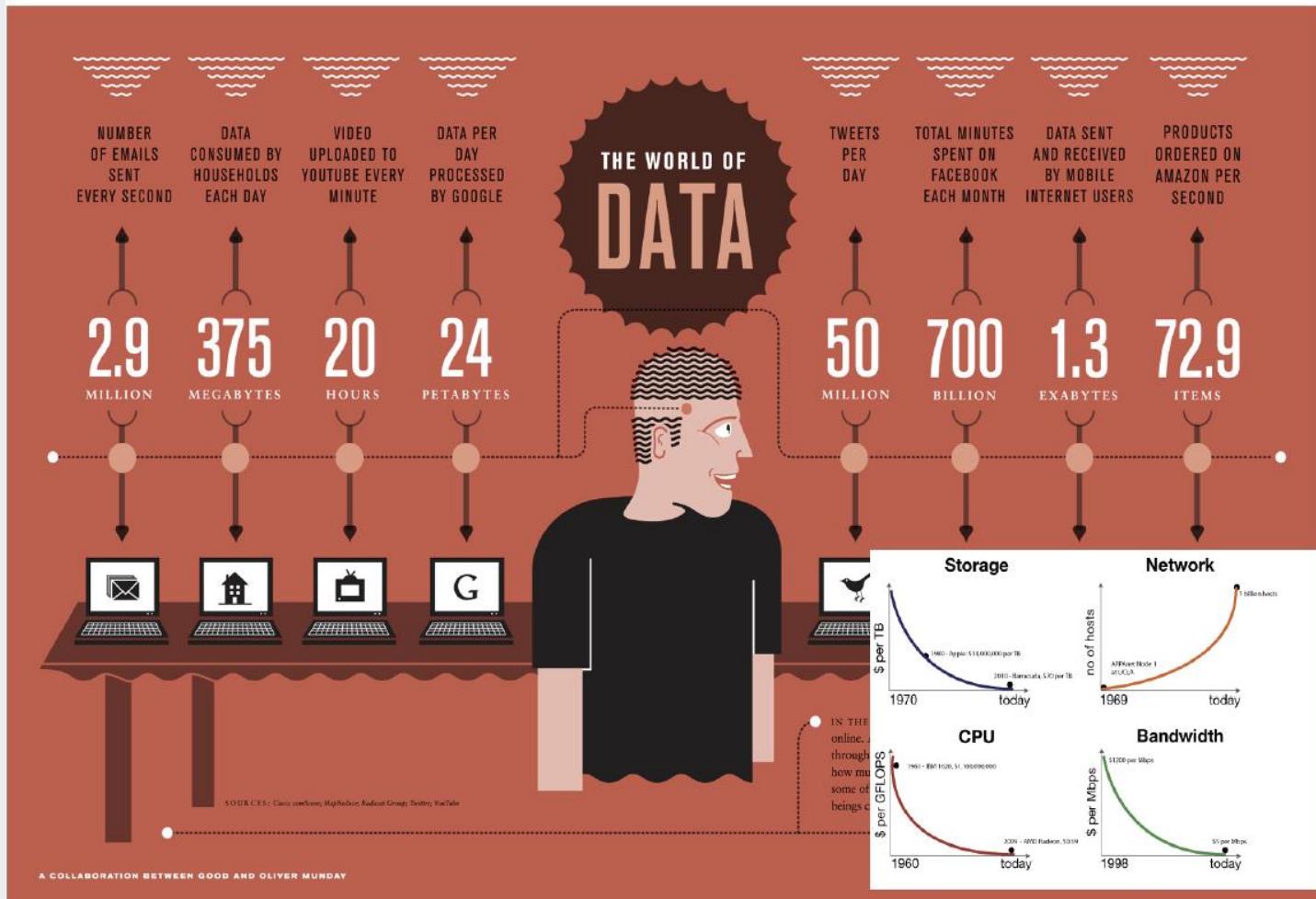


Value



Data Analytics: The Era of “Big Data”

- Volume of Big Data

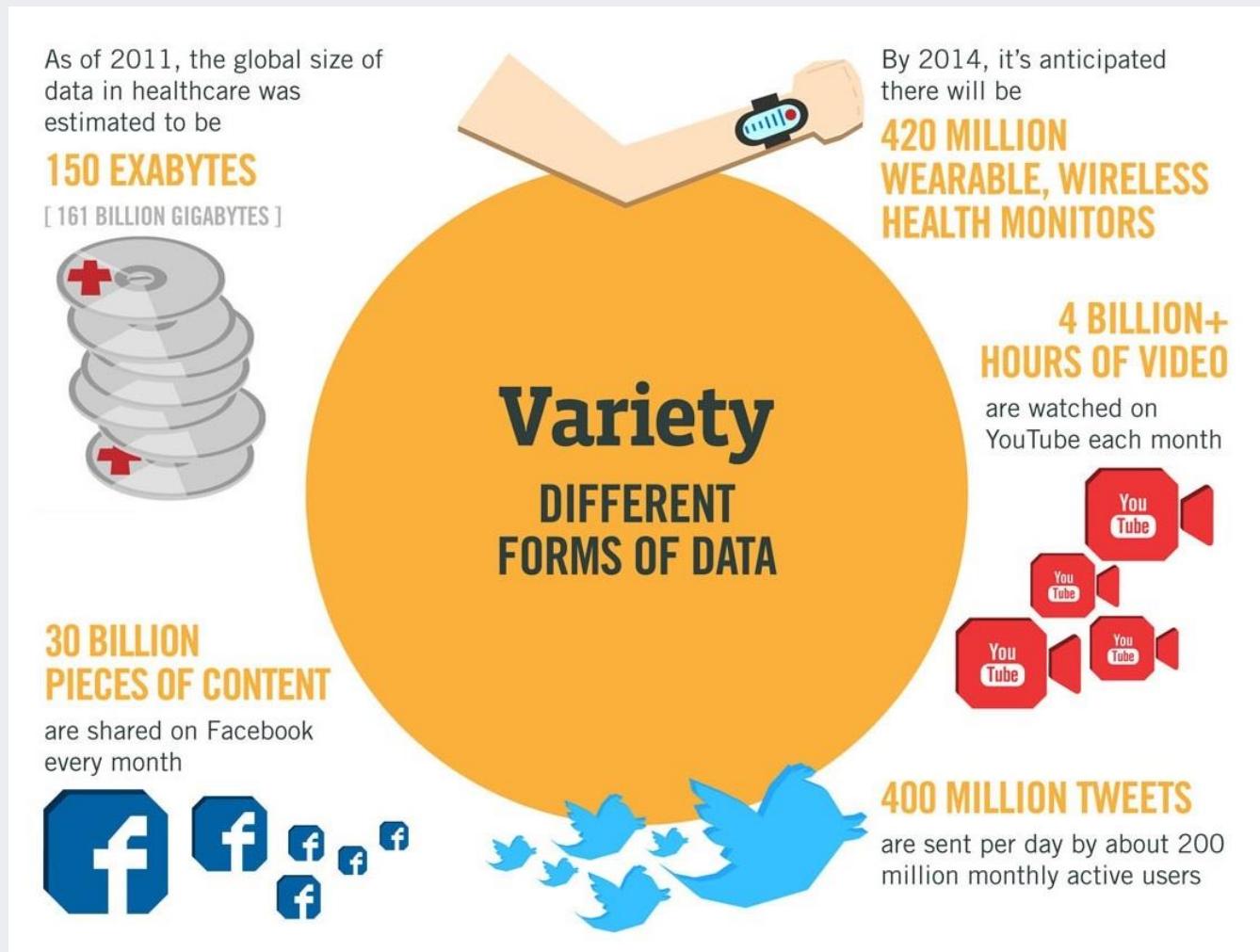


Data Analytics: The Era of “Big Data”

- Velocity of Big Data

Data Analytics: The Era of “Big Data”

- Variety in Big Data



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

Data Analytics

- Data Analytics in Industrial Engineering

Understanding Analytics

- Descriptive vs. Predictive vs. Prescriptive Analytics

Understanding analytics			
	Descriptive	Predictive	Prescriptive
What the user needs to DO	What HAS happened? <ul style="list-style-type: none">• Increase asset reliability• Reduce labor and inventory costs	What COULD happen? <ul style="list-style-type: none">• Predict infrastructure failures• Forecast facilities space demands	What SHOULD happen? <ul style="list-style-type: none">• Increase asset utilization• Optimize resource schedules
What the user needs to KNOW	<ul style="list-style-type: none">• The number and types of asset failures• Why maintenance costs are high• The value of the materials inventory	<ul style="list-style-type: none">• How to anticipate failures for specific asset types• When to consolidate underutilized facilities• How to determine costs to improve service levels	<ul style="list-style-type: none">• 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	<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

Understanding Analytics

- Business Intelligence vs. Advanced Analytics

	Business Intelligence	Advanced Analytics
Orientation	Rearview	Future
Types of questions	What happened When, who, how many	What will happen? What will happen if we change this one thing? What's next?
Methods	Reporting (KPIs, metrics) Automated Monitoring/Alerting (thresholds) Dashboards Scorecards OLAP (Cubes, Slice & Dice, Drilling) Ad hoc query	Predictive Modeling Data Mining Text Mining Multimedia Mining Descriptive Modeling Statistical / Quantitative Analysis Simulation & Optimization
Big Data	Yes	Yes
Data types	Structured, some unstructured	Structured and Unstructured
Knowledge Generation	Manual	Automatic
Users	Business Users	Data scientists, Business analysts, IT, Business Users
Business Initiatives	Reactive	Proactive

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	PROGRAMMING & DATABASE	DOMAIN KNOWLEDGE & SOFT SKILLS	COMMUNICATION & VISUALIZATION
<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	<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	<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	<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

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	PROGRAMMING & DATABASE	DOMAIN KNOWLEDGE & SOFT SKILLS	COMMUNICATION & VISUALIZATION
<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	<ul style="list-style-type: none">★ 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	<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	<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

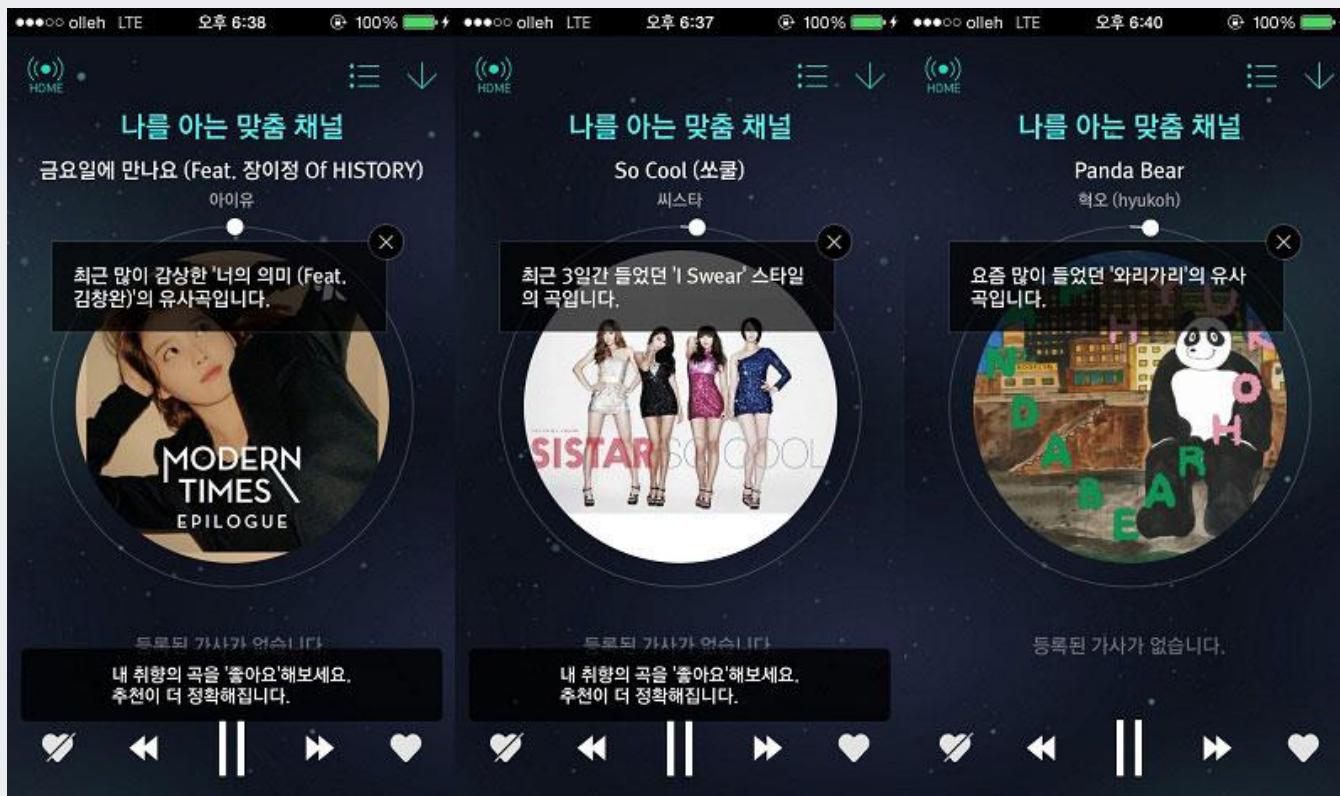
MarketingDistillery.com is a group of practitioners in the area of e-commerce marketing. Our fields of expertise include: marketing strategy and optimization: customer tracking and on-site analytics; predictive analytics; and econometrics; data warehousing and big data systems; marketing channel insights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

Marketing DISTILLERY

Data Science Tasks

- Data Mining

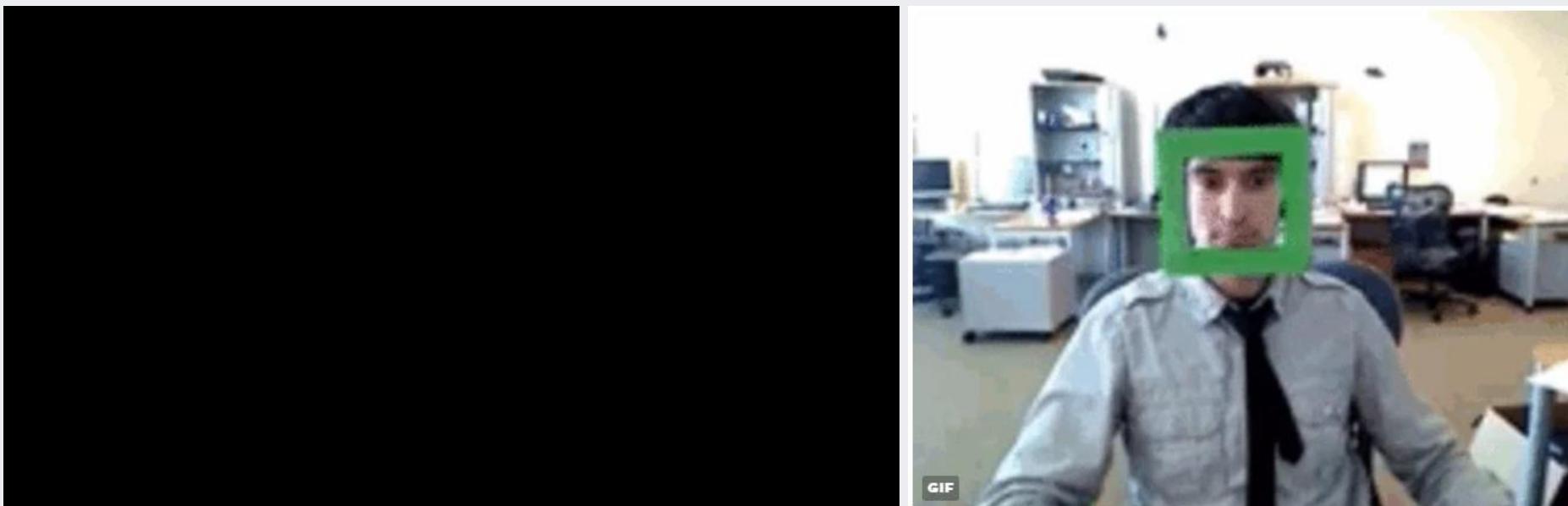
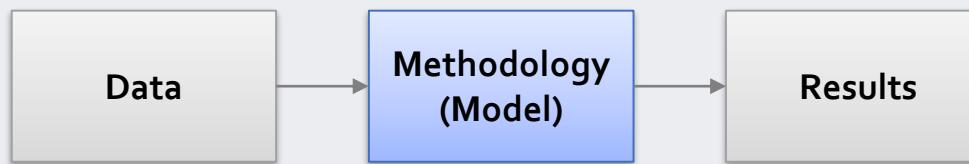
- ✓ The process of exploration and analysis, by automatic or semi-automatic means, of large quantities of data in order to discover meaningful patterns and rules. (Berry and Linoff, 1997, 2000)



Data Science Tasks

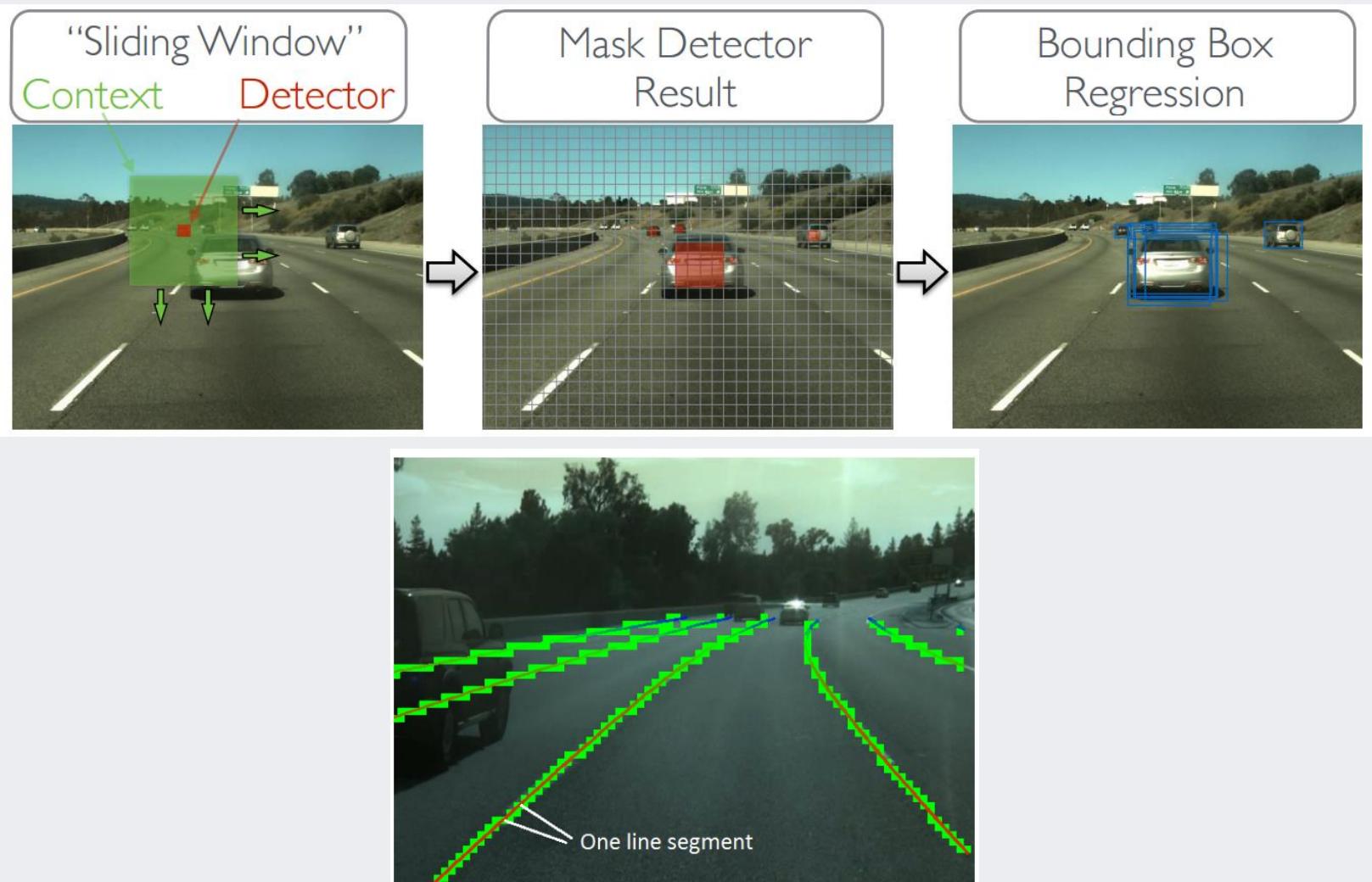
- Machine Learning

- ✓ 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 et al. (2013)



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



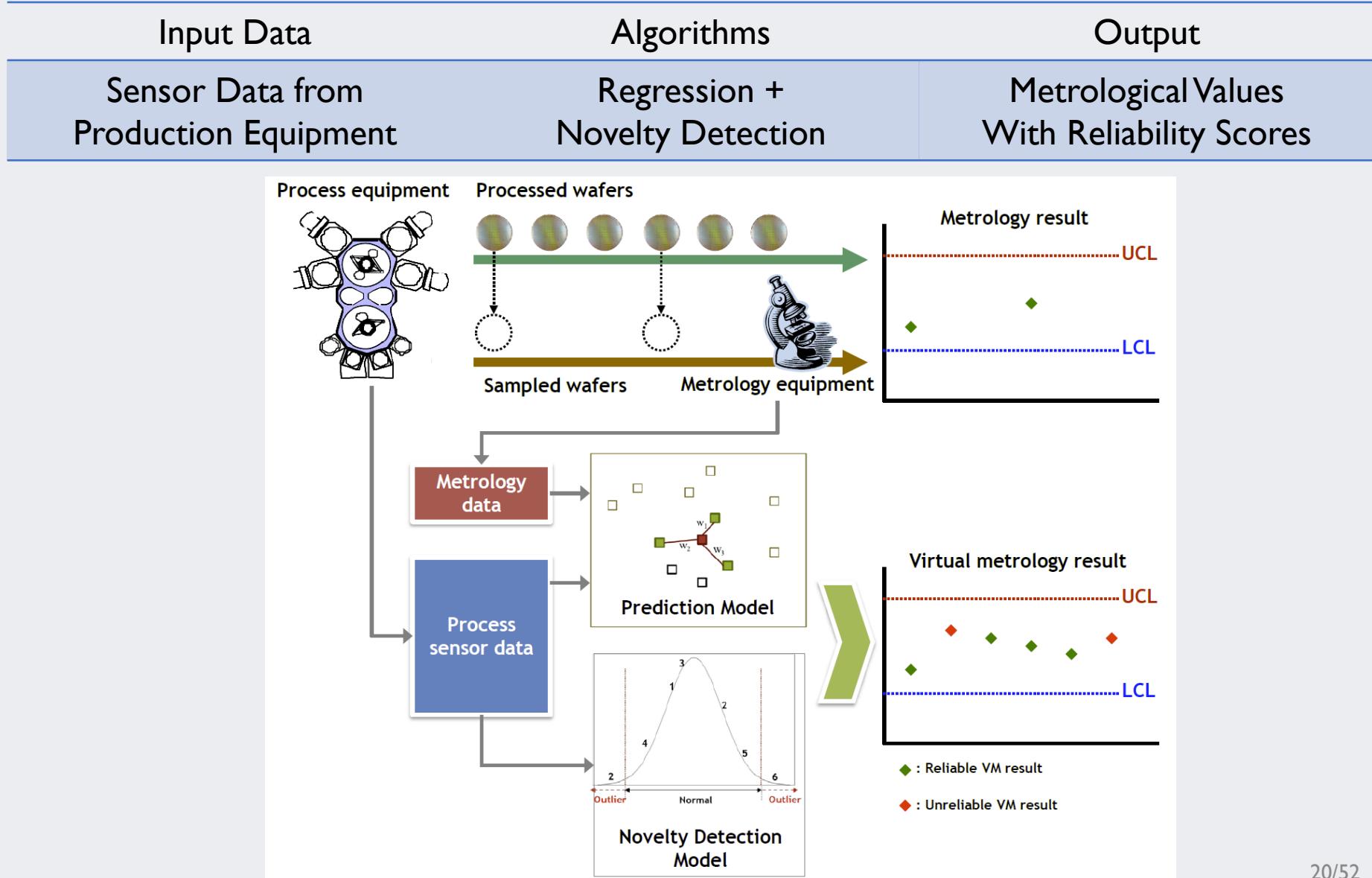
Data Science Tasks

- Artificial Intelligence (AI Speaker Amazon Echo)

AGENDA

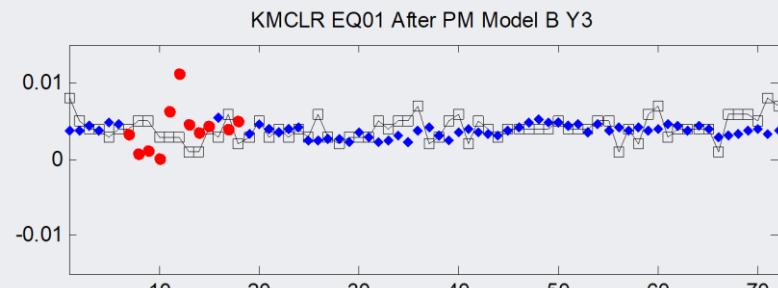
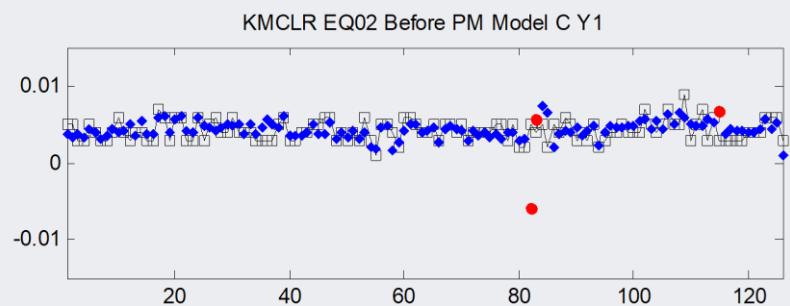
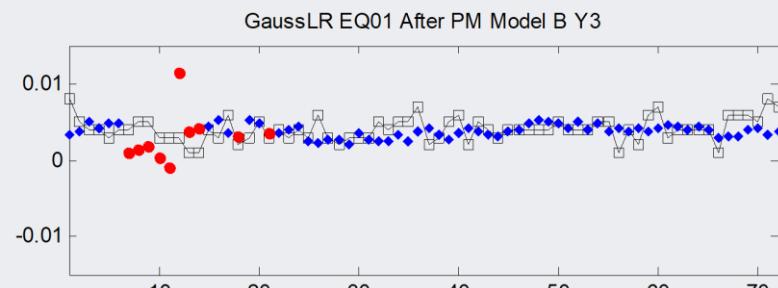
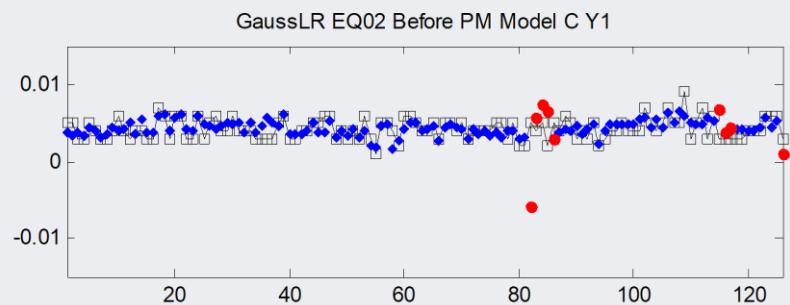
- 01 What is Data Analytics?
- 02 What Can We Do with Data Analytics?
- 03 Data Analytics Languages: R & Python

Process Monitoring & Control



Process Monitoring & Control

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



Are You a Valid User?

Input Data

Time stamps collected during typing

Through various input devices



Algorithms

Novelty Detection

In any stages

Log in

Don't have an account? [Create one.](#)

Username:

Password:

Remember me (up to 30 days)

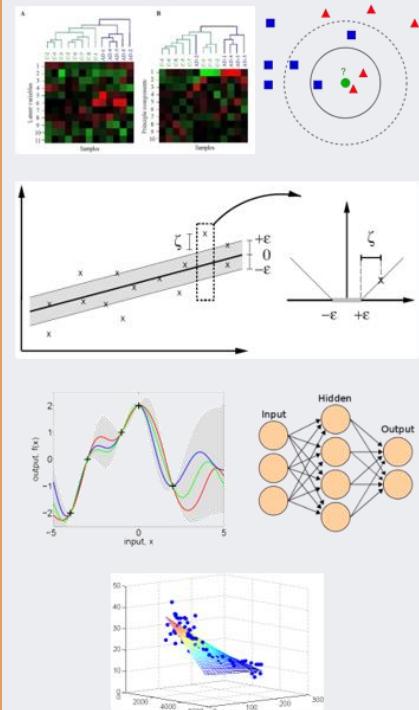
[Log in](#) [E-mail new password](#)



Output

Valid user score

Customized (real-time) authenticator



Who Looks Happy?

Input Data	Algorithms	Output
Image	Convolutional Neural Networks	Emotions Recognized
		Neutral Neutral Neutral Neutral Neutral Neutral Neutral Neutral Neutral Neutral Neutral

Favorite Artistic Style

Input Data	Algorithms	Output
Image	Convolutional Neural Networks	Image with Preferred Style
A 	B 	E 
C 	D 	F  Pinterest 1 Facebook Like 50

Understanding Emotions

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

The diagram illustrates a Convolutional Neural Network (CNN) architecture for sentiment analysis. It starts with an input sequence of words: "Padding", "this", "film", "is", "actually", "quite", "entertaining", "Padding". These words are represented as a grid of features. The CNN processes this grid through several layers. A feature map is shown with colored channels (yellow, orange, blue) representing different word embeddings or filters. These features are then processed by a fully connected layer (FC) that outputs weights w_1 , w_2 , and w_n . These weights are applied to the input features to produce a final output vector. This vector is then compared against a set of thresholds to determine the sentiment, resulting in "Positive" or "Negative".

$$w_1 \times \begin{array}{|c|c|c|}\hline & & \\ \hline & & \\ \hline & & \\ \hline \end{array} + w_2 \times \begin{array}{|c|c|c|}\hline & & \\ \hline & & \\ \hline & & \\ \hline \end{array} + \dots + w_n \times \begin{array}{|c|c|c|}\hline & & \\ \hline & & \\ \hline & & \\ \hline \end{array} = \begin{array}{|c|c|c|}\hline & & \\ \hline & & \\ \hline & & \\ \hline \end{array}$$

The equation shows the weighted sum of input features. The input features are represented as vertical vectors of colored squares (yellow, orange, blue). The weights w_1 , w_2 , and w_n are applied to these vectors. The resulting vectors are summed to produce a final output vector, which is then compared against a set of thresholds to determine the sentiment.

25/52

Understanding Emotions

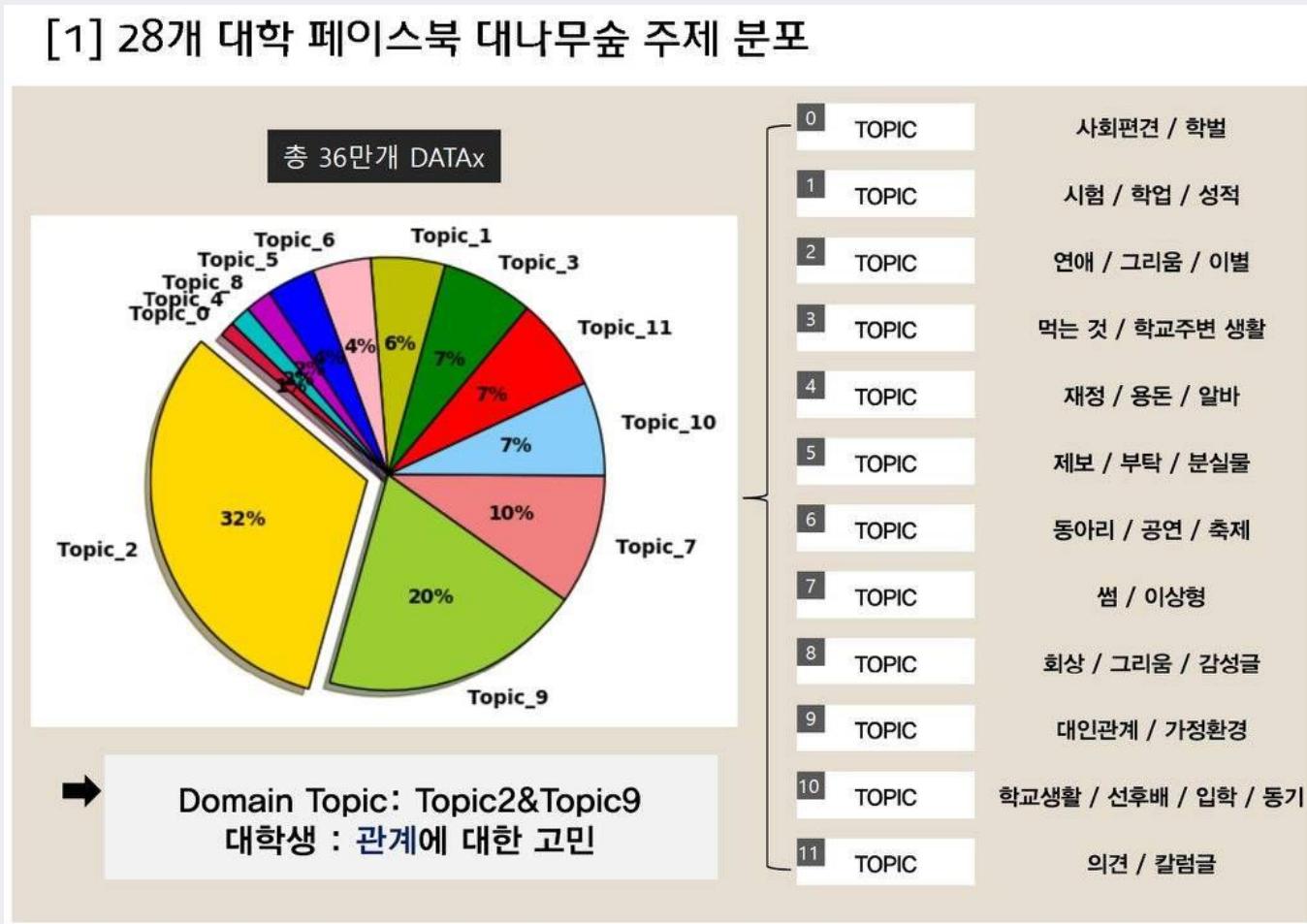
Input Data	Algorithms	Output
Movie Review Text	Convolutional Neural Networks	Sentiment (Pos/Neg) & Keyword attention
Method		Sentence
Raw text	One of the funniest most romantic and most musical movies ever; definitely worth renting/buying especially if you have a taste for older style of cinematography. 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 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	

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. <i>(3 / 10 points)</i>	
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 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 well. Giovanni Ribisi's character was quite annoying For some reason he seems to like repeating what he says. If Rain 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 well. Giovanni Ribisi's character was quite annoying For some reason he seems to like repeating what he says. If Rain Man it would been 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 been 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 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



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

Input Data

Facebook Posts
(Bamboo forest)

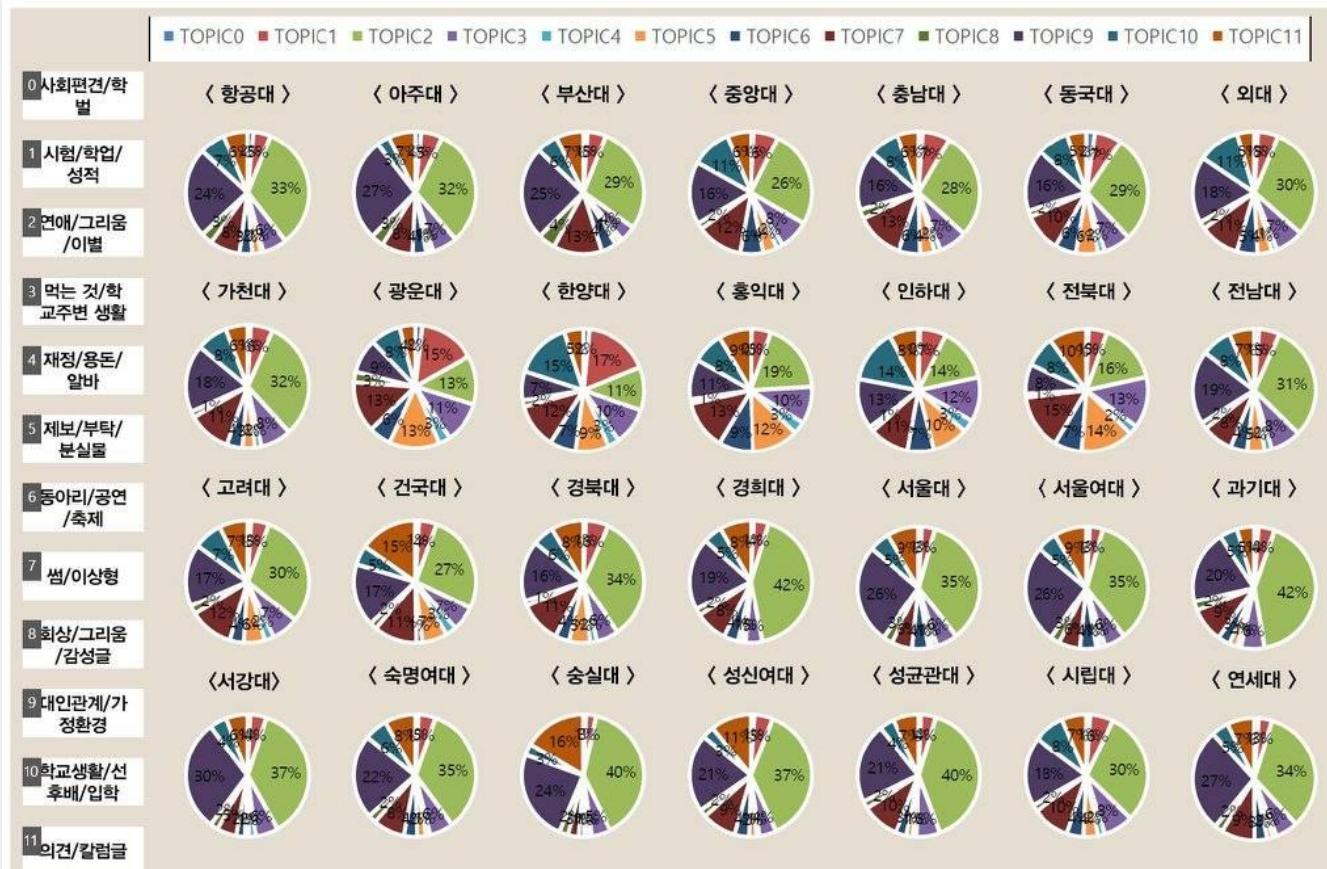
Algorithms

Latent Dirichlet Allocation
(Topic Modeling)

Output

Topic Distribution

[2] 대학별 페이스북 대나무숲 주제 분포



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

Input Data

Facebook Posts
(Bamboo forest)

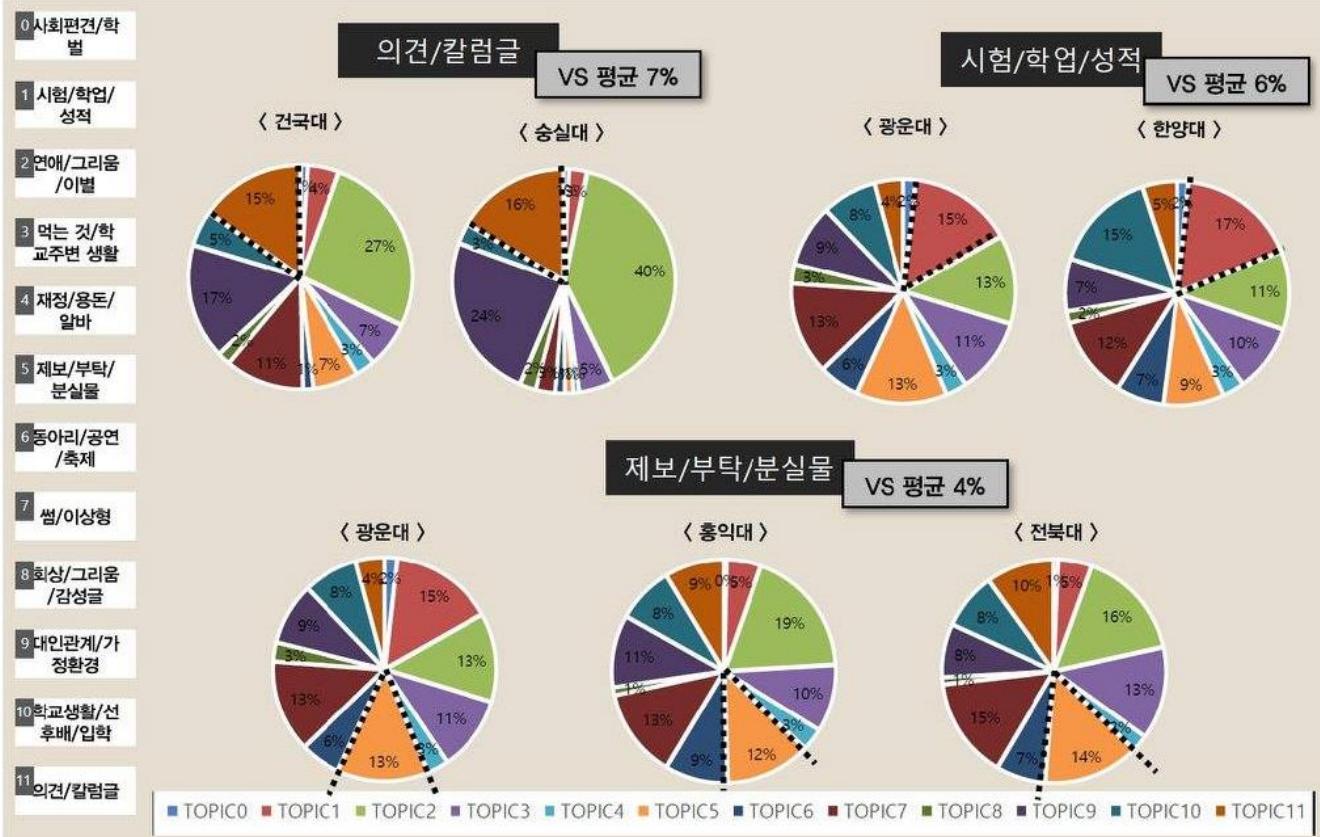
Algorithms

Latent Dirichlet Allocation
(Topic Modeling)

Output

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



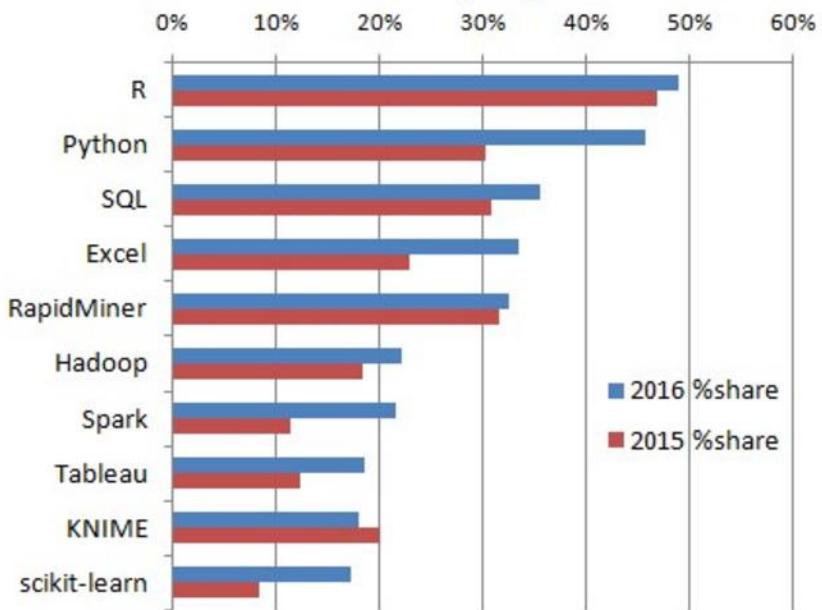
R vs. Python

- Some interesting polls

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

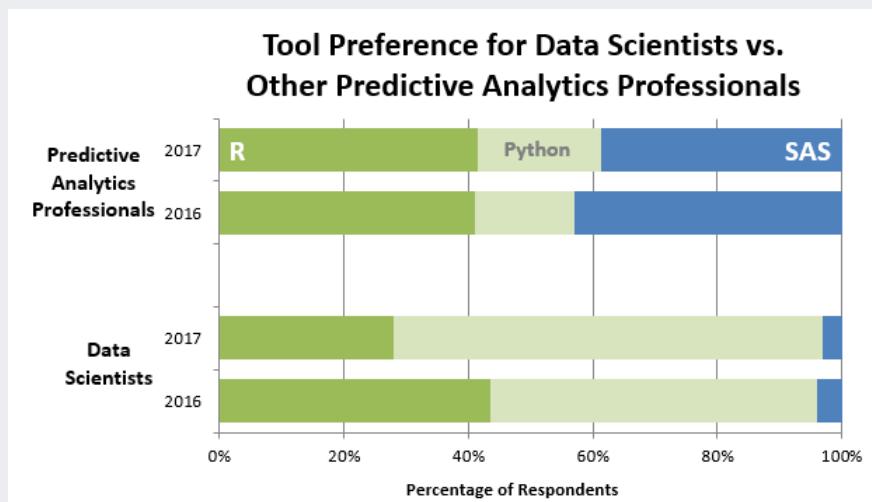
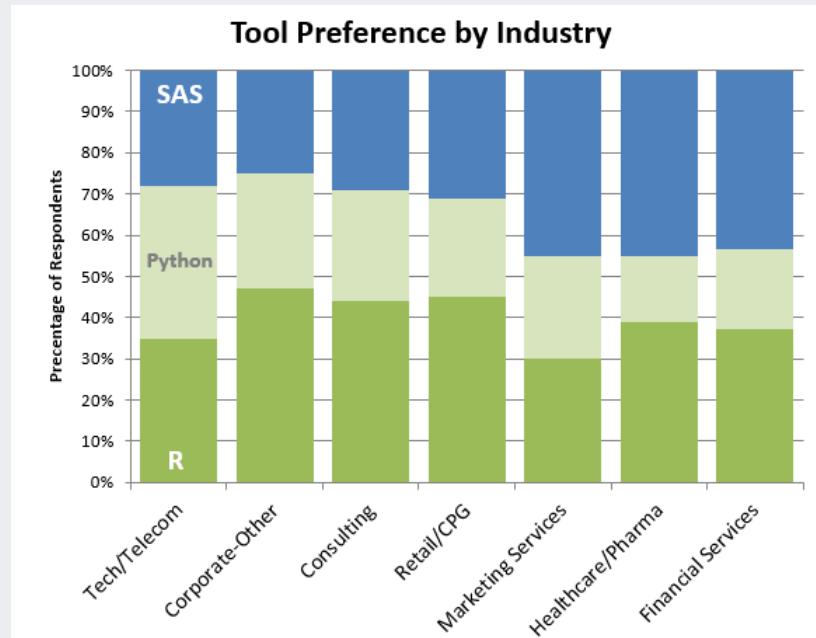
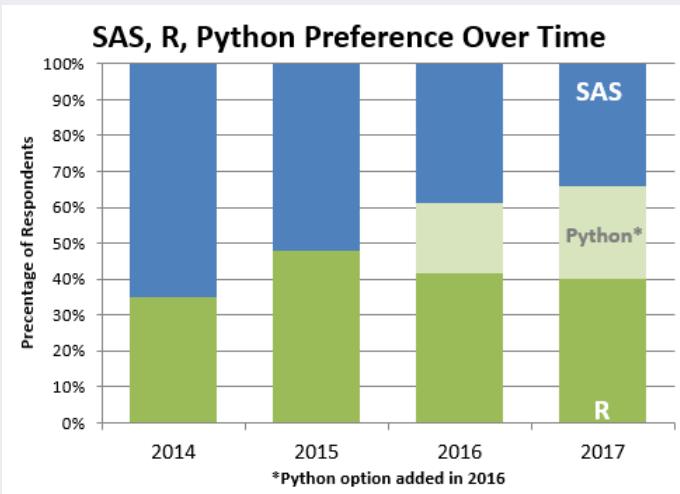
Tool	2016 % share	% change	% alone
R	49%	+4.5%	1.4%
Python	45.8%	+51%	0.1%
SQL	35.5%	+15%	0%
Excel	33.6%	+47%	0.2%
RapidMiner	32.6%	+3.5%	11.7%
Hadoop	22.1%	+20%	0%
Spark	21.6%	+91%	0.2%
Tableau	18.5%	+49%	0.2%
KNIME	18.0%	-10%	4.4%
scikit-learn	17.2%	+107%	0%

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



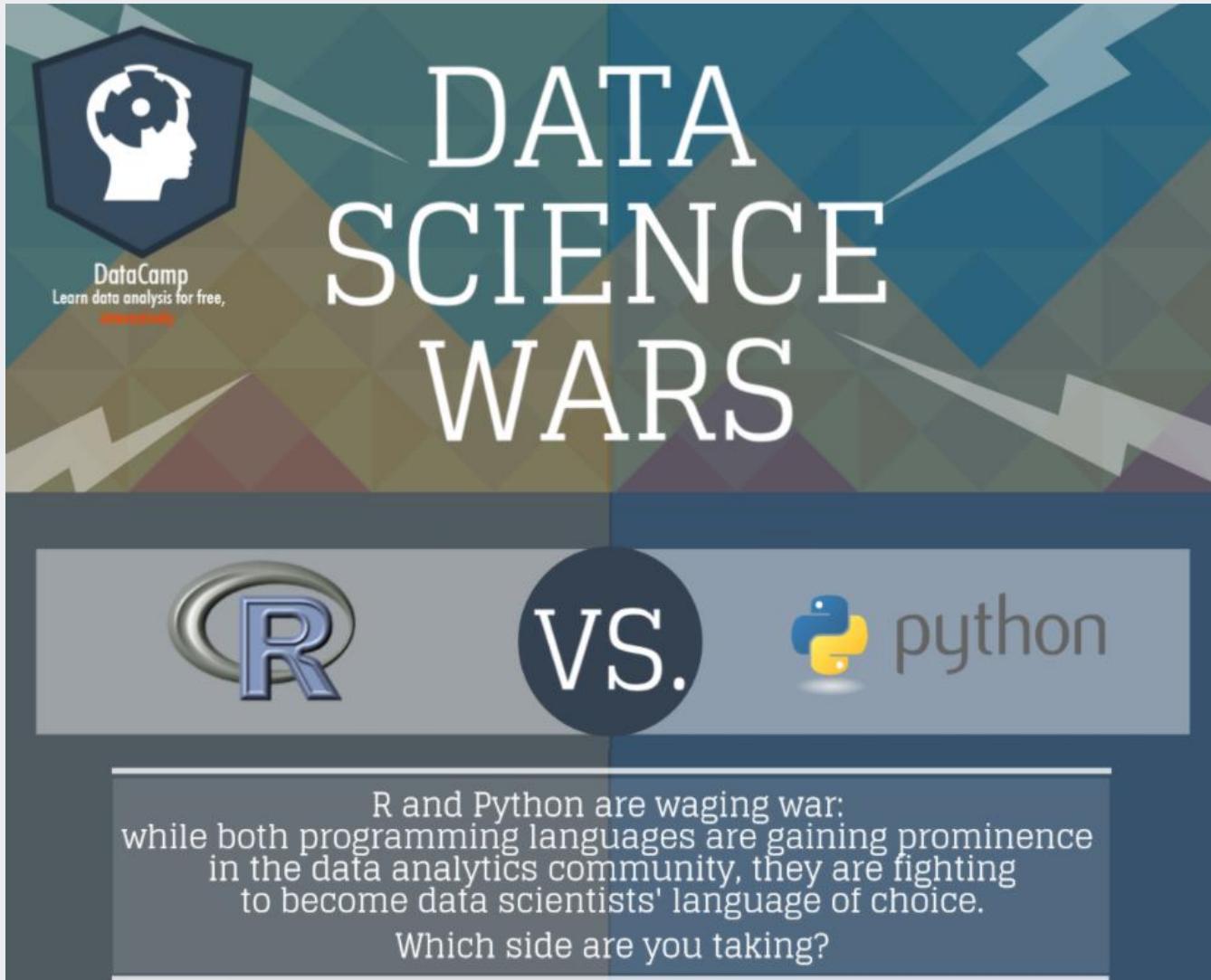
R vs. Python

- Some interesting polls



R vs. Python

- Data Science Wars: R vs. Python



R vs. Python

- Data Science Wars: R vs. Python

The infographic is titled "Introducing The Opponents" and features a large "#1" icon on the left. It compares R (left side) and Python (right side) across several categories: Current Version, History, Creators, Release Year, and Must Knows. A central lightbulb icon represents the comparison between the two languages.

Introducing The Opponents	
Current Version	3.1.3 / 2.7.9 March 2015 / February 2015 / December 2014
History	Creator
Creators	Guido Van Rossum
Release Year	1991
Must Knows	Must Knows
1. R is an implementation of S programming language (Bell Labs). 2. R's design and evolution is handled by the R-core group and R foundation. 3. R's software environment was written primarily in C, Fortran and R.	1. Python was inspired by C, Modula-3, and particularly ABC. 2. Python gets its name from the "Monty Python's Flying Circus" comedy series. 3. Python Software Foundation (PSF) takes care of Python's advances.

R vs. Python

- Data Science Wars: R vs. Python

Purpose	Used By?	Community
R focuses on better, user friendly data analysis, statistics and graphical models.	Python emphasizes productivity and code readability.	
R has been used primarily in academics and research. However, R is rapidly expanding into the enterprise market. <i>"The closer you are to statistics, research and data science, the more you might prefer R."</i>	Python is used by programmers that want to delve into data analysis or apply statistical techniques, and by developers that turn to data science. <i>"The closer you are to working in an engineering environment, the more you might prefer Python."</i>	Huge community with support coming in the form of: - Mailing lists - User-contributed documentation - Active Stackoverflow members More adoption from researchers, data scientists, statisticians, quants.

R vs. Python

- Data Science Wars: R vs. Python

Usability	
Statistical models can be written with only a few lines.	Coding and debugging is easier to do in Python, mainly because of the "nice" syntax.
There are R stylesheets but not everyone uses them.	The indentation of the code affects its meaning.
The same piece of functionality can be written in several ways in R.	Any piece of functionality is always written the same way in Python.
Flexibility	
It is easy to use complex formulas in R. All kinds of statistical tests and models are readily available and easily used.	Python is flexible for doing something novel that has never been done before. Developers can also use it for scripting a website or other applications.
Ease of Learning	
R has a steep learning curve at start. Once you know the basics, you can easily learn advanced stuff.	Python's focus on readability and simplicity makes that its learning curve is relatively low and gradual.
R is not hard for experienced programmers.	Python is considered a good language for starting programmers.
<i>Check out DataCamp's interactive exercises and tutorials.</i>	<i>Try using the book "Learn Python The Hard Way" and its accompanying site with videos and exercises.</i>

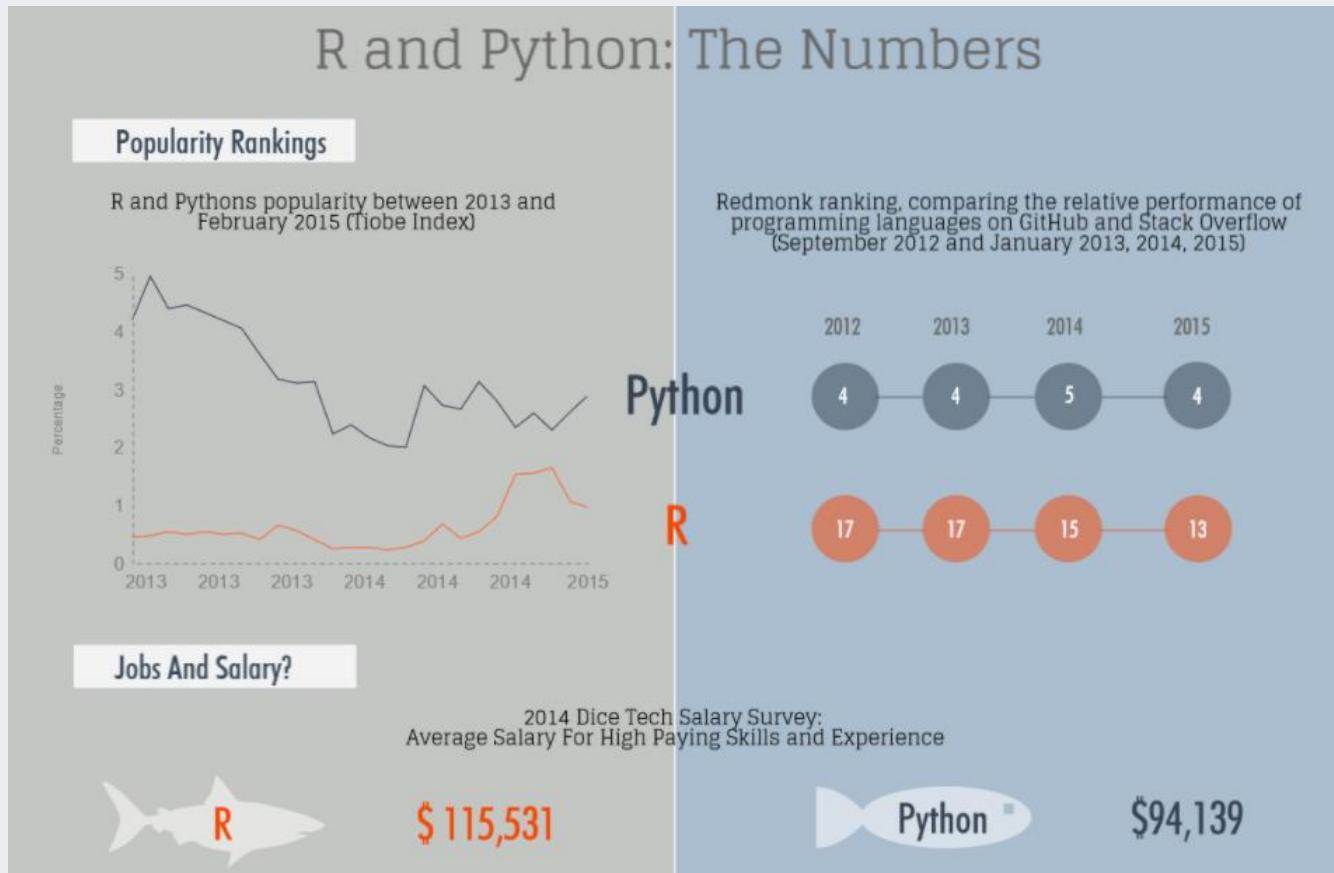
R vs. Python

- Data Science Wars: R vs. Python

Code Repositories	
CRAN stands for the Comprehensive R Archive Network: it is a huge repository of R packages to which users can easily contribute.	PyPi is the Python Package Index: it is a repository of Python software, consisting of libraries. Users can contribute to PyPi, but it is a bit complicated in practice.
Packages are collections of R functions, data, and compiled code. They can be installed in R with one line.	Watch out with dependencies and installing Python libraries!
<p><i>"I don't see Python [...] building up a huge code repository comparable to CRAN. [R has] a gigantic head start, [and] [...] statistics simply is not Python's central mission;"</i></p> <p>- Norm Matloff, professor of computer science</p>	
Miscellaneous	
Use the rPython package to run Python code from R. Pass or get data from Python, call Python functions or methods.	Use the RPy2 library to run R code from within Python. It provides a low-level interface from Python to R.

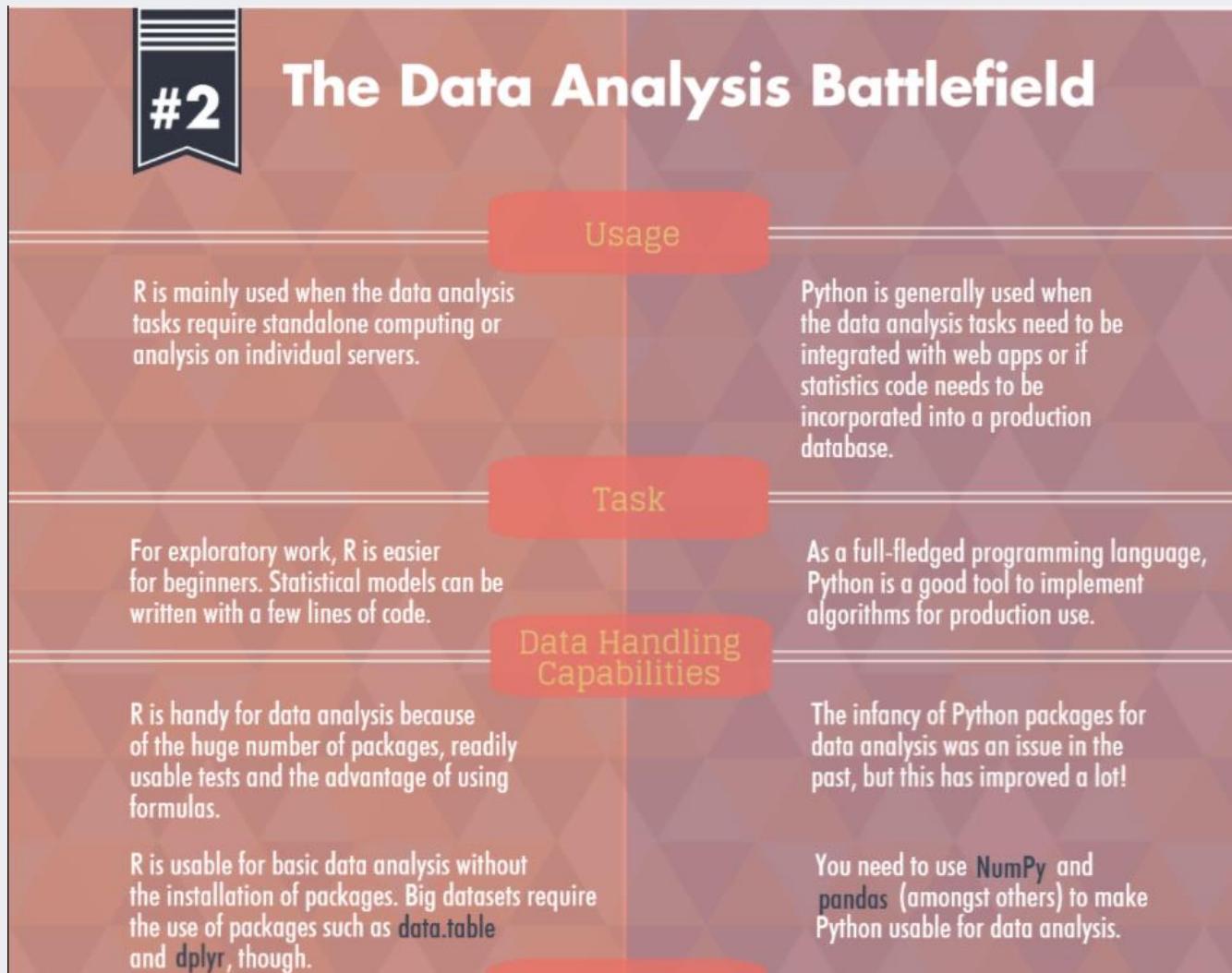
R vs. Python

- Data Science Wars: R vs. Python



R vs. Python

- Data Science Wars: R vs. Python



R vs. Python

Getting Started

IDE

R Studio

Popular Packages

- ✓ dplyr, plyr and data.table to easily manipulate data.
- ✓ stringr to manipulate strings.
- ✓ zoo to work with regular and irregular time series
- ✓ ggvis, lattice and ggplot2 to visualize data.
- ✓ caret for machine learning.

Tip: check out DataCamp's online interactive courses and tutorials!

"R is currently head-and-shoulders above Python for data analysis, but I remain convinced that Python CAN catch up, easily and quickly."
- Jan Galkowski, computational engineer

IDE

There are many Python IDEs to choose from. However, Spyder and IPython Notebook are most popular.

Tip: also look up Rodeo, the "data science IDE for Python"

Popular Libraries

- ✓ pandas to easily manipulate data.
- ✓ SciPy /NumPy for scientific computing.
- ✓ scikit-learn to use machine learning methods.
- ✓ matplotlib to make graphics.
- ✓ statsmodels to explore data, estimate statistical models, and perform statistical tests and unit tests.

Support

There's a lot of support out there for data analysis with R:

- ✓ Stackoverflow
- ✓ Rdocumentation, the R documentation aggregator
- ✓ R-help mailing list

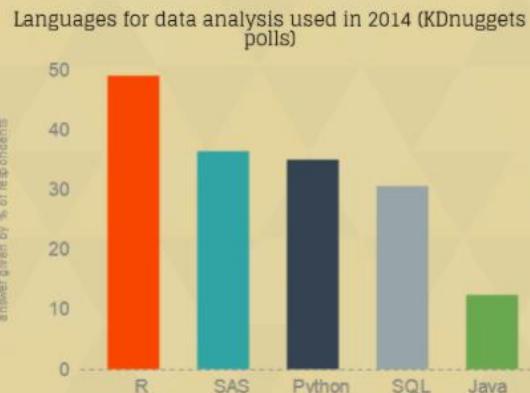
Support for data analysis issues can be found at:

✓ Stackoverflow	Questions related to Python for data analysis and pandas
✓ Mailing lists:	
pydata	Statsmodels or pandas questions
pystatsmodels	Numpy questions
numpy-discussion	General SciPy or scientific questions
sci-py user	

R vs. Python

R And Python: The Quantified Battlefield

General

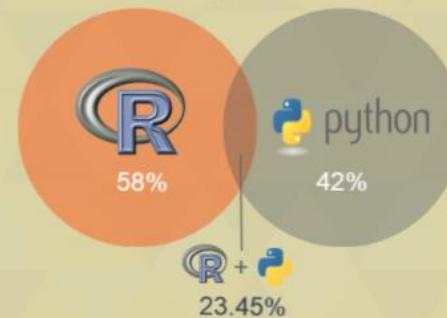


Community?

Stack Overflow Questions tagged "R" and/or "Python", "Pandas" between 2008 and April 15, 2015



Analysis of R and Python used together in 2014 (KDnuggets polls)



Twitter activities between March 12 and April 10, 2015



R vs. Python

#3

The Last Stand: Pros And Cons

Graphical Capabilities

A picture says more than a thousand words

Visualized data can be understood more efficiently and effectively than the raw numbers alone.

R + visualization
= perfect match



ggplot2 To make pretty graphs, including the opportunity to use grammar of graphics to create layered, customizable plots

lattice To easily display multivariate relationships

rCharts To create, customize and publish interactive javascript visualizations from R

googleVis To use Google Chart tools to visualize data in R

ggvis To implement interactive grammar of graphics, while rendering in a web browser

e.g.: Visualizing Facebook friends with R



IPython Notebook

Bundle your analysis in one file

The IPython Notebook makes it easier to work with Python and data.

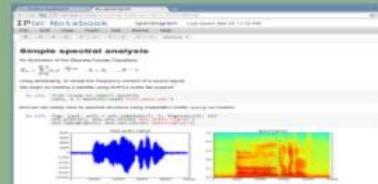
Simplify your workflow when working with data in Python

It's a combination of:

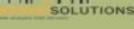
Interactive python exploration,
prewritten programs, text, and equations for documentation in one environment

Share notebooks with colleagues without having them install anything.

The IPython notebook drastically reduces the overhead of organizing code, output, and notes files, which allows to spend more time doing real work.



R vs. Python

The R Ecosystem	Python, A General Purpose Language
<h3>The R Project</h3> <p>Rich ecosystem of cutting-edge interface packages available to communicate between open-source languages.</p> <p>This allows you to string your workflow together, which is especially useful for data analysis.</p> <p>Packages are available at:</p> <ul style="list-style-type: none">Cran "Task Views" page lists a wide range of tasks for which R packages are availableBioconductor Open source software for bioinformaticsGitHub web-based Git repository hosting service <p>   </p> <p>Search through all these sources easily with Rdocumentation, the first R documentation aggregator</p> <h3>The R User Community</h3> <ul style="list-style-type: none">✓ Meetup groups<ul style="list-style-type: none">Some are sponsored by companies of the R community✓ Blogs & Social Media<ul style="list-style-type: none">R bloggers#rstats	 <h3>Readability and Learning Curve</h3> <p>Just like everyday English</p> <p>Python is easy and intuitive, and its emphasis on readability only magnifies these characteristics.</p> <p>e.g. <code>print("Hello World!")</code></p> <p>Syntactically clear and elegant code, easily interpretable and very easy to type.</p> <p>This explains why.</p> <ul style="list-style-type: none">✓ Python's learning curve is relatively flat✓ So many programmers are familiar with it <p>Also, the speed at which you can write a program is also positively impacted:</p> <p>Less time coding, more time playing</p> <h3>The Python Testing Framework</h3> <p>Guarantee your code is reusable and dependable</p> <p>A built-in, low barrier-to-entry testing framework that encourages good test coverage.</p> <h3>Python Testing Tools Taxonomy, including</h3> <ul style="list-style-type: none">UnitTest First unit test framework of the Python standard libraryNose Extends UnitTest: used in many packages such as pandasDocTest Easy generation of tests based on output from the standard Python interpreter shellPytest To write small tests, while supporting complex functional testing <p> testing-in-python (TIP) mailing list</p>

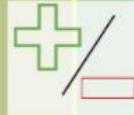
R vs. Python

- Data Science Wars: R vs. Python

R, Lingua Franca of Statistics	Python, A Multi-Purpose Language
<p>Developed by statisticians, for statisticians</p> <p>Statisticians communicate ideas and methods for statistical analysis through R code and packages.</p> <p>Statisticians, engineers and scientists without computer programming skills find it easy to use.</p> <p>Increasing industry adoption...</p> <p>R is used in finance, pharmaceuticals, media and marketing; In this last area, R's on the rise as a business analytics tool.</p> <p>"The number one value to businesses in using R is access to talent"</p> <p>Google  </p> <p>... And widespread use in academia</p> <p>R is experiencing a rapid growth, solidifying its position in third place as software used in scholarly articles, right after SAS and SPSS.</p>	 <p>Ready To Work!</p> <p>As a common, easy-to-understand language that is known by many programmers, Python also brings people with different backgrounds together.</p> <p>For example,</p> <p>Some organizations that didn't want to hire or had difficulties to hire new data scientists (re)trained their existing employees to use Python instead.</p> <p>This means that Python is a production ready language: it has the capacity to be a single tool that integrates with every part of your workflow!</p> 

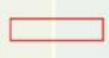
R vs. Python

- Data Science Wars: R vs. Python

R Is Slow	Python And Visualizations																
<p>R is slow, on purpose </p> <p>R was designed to make data analysis and statistics easier to do, not to make life easier for your computer.</p> <p>R has an incomplete informal definition; It is mostly defined in terms of how its implementation works.</p> <p>Beyond design and implementation, a lot of R code is slow simply because it's poorly written.</p> <p>Packages to improve R's performance:</p> <table><tbody><tr><td>pqr</td><td>A new version of the R interpreter</td></tr><tr><td>renjin, FastR</td><td>Original R rewritten in Java</td></tr><tr><td>Riposte</td><td>A fast interpreter and JIT for R</td></tr><tr><td>RevoScaleR</td><td>Commercial tool to handle big datasets</td></tr><tr><td>Foreach</td><td>Commercial tool that facilitates parallel programming</td></tr></tbody></table>	pqr	A new version of the R interpreter	renjin, FastR	Original R rewritten in Java	Riposte	A fast interpreter and JIT for R	RevoScaleR	Commercial tool to handle big datasets	Foreach	Commercial tool that facilitates parallel programming	 <p>"Visualizations are important criteria in choosing data analysis software"</p> <p>Python has some nice visualization libraries:</p> <table><tbody><tr><td>Seaborn</td><td>Library based on matplotlib</td></tr><tr><td>Bokeh</td><td>Interactive visualization library</td></tr><tr><td>Pygal</td><td>To create dynamic svg charts</td></tr></tbody></table> <p>But there are a lot of options to choose from; Maybe too many.</p> <p>Moreover, in comparison to R</p> <p>"Visualizations in Python are usually more convoluted, and the results are not nearly as pleasing to the eye or as informative."</p>	Seaborn	Library based on matplotlib	Bokeh	Interactive visualization library	Pygal	To create dynamic svg charts
pqr	A new version of the R interpreter																
renjin, FastR	Original R rewritten in Java																
Riposte	A fast interpreter and JIT for R																
RevoScaleR	Commercial tool to handle big datasets																
Foreach	Commercial tool that facilitates parallel programming																
Seaborn	Library based on matplotlib																
Bokeh	Interactive visualization library																
Pygal	To create dynamic svg charts																

R vs. Python

- Data Science Wars: R vs. Python

R's Steep Learning Curve		Python Is Immature ("It's a challenger!")						
<p>"The worst thing about R is that ... it was developed by statisticians."</p> <p>R's learning curve is nontrivial:</p> <ul style="list-style-type: none">• Even though anybody can get results using GUIs, none is comprehensive enough to totally avoid programming.• Finding packages can be time consuming <p>Using the right tools</p> <p>Good resources can help you to overcome this steep learning curve:</p> <ul style="list-style-type: none"> DataCamp's interactive exercises and tutorials Rdocumentation to search for packages		<p>Python Is Immature ("It's a challenger!")</p> <p>A more limited way to think about data analysis</p> <p>At the moment, there are no module replacements for the 100s of essential R packages</p> <p>Python's catching up, but will this make people give up R?</p> <ul style="list-style-type: none">• IPython's R extension allows you to cleanly use R in the IPython notebook.• The current landscape of conventions and resources plays a huge role:<table><tr><td>Matlab</td><td>Commonly used to publish open research code</td></tr><tr><td>Python</td><td>Used in mathematics</td></tr><tr><td>R</td><td>Used in statistics</td></tr></table> <p>Mlabwrap offers a bridge from Python to Matlab, but there are some drawbacks:</p> <ul style="list-style-type: none">- You need to work with two languages- You need a Matlab license	Matlab	Commonly used to publish open research code	Python	Used in mathematics	R	Used in statistics
Matlab	Commonly used to publish open research code							
Python	Used in mathematics							
R	Used in statistics							

R vs. Python

- Data Science Wars: R vs. Python

Shared Positive Points	
	
Open-Source	
<p>R and Python are free to download for everyone, in comparison to other statistical software such as SAS and SPSS, which are commercial tools.</p>	
Advanced Tools	
<p>Many new developments in statistics appear first in the open source packages of R and, to lesser extent, Python, before making their way to commercial platforms.</p>	
	Online Communities
	<p>While commercial softwares offer (paid) customer support, R and Python dispose of online communities that offer support to their respective users.</p>
	Paycheck
	<p>According to the O'Reilly 2013 Data Science Salary Survey, data scientists that use primarily open-source tools earned a higher median salary (US\$130,000) than those using proprietary tools (US\$90,000)</p>



ANY
questions?