Lecture 1: Data Analytics

Pilsung Kang

School of Industrial Management Engineering

Korea University

AGENDA

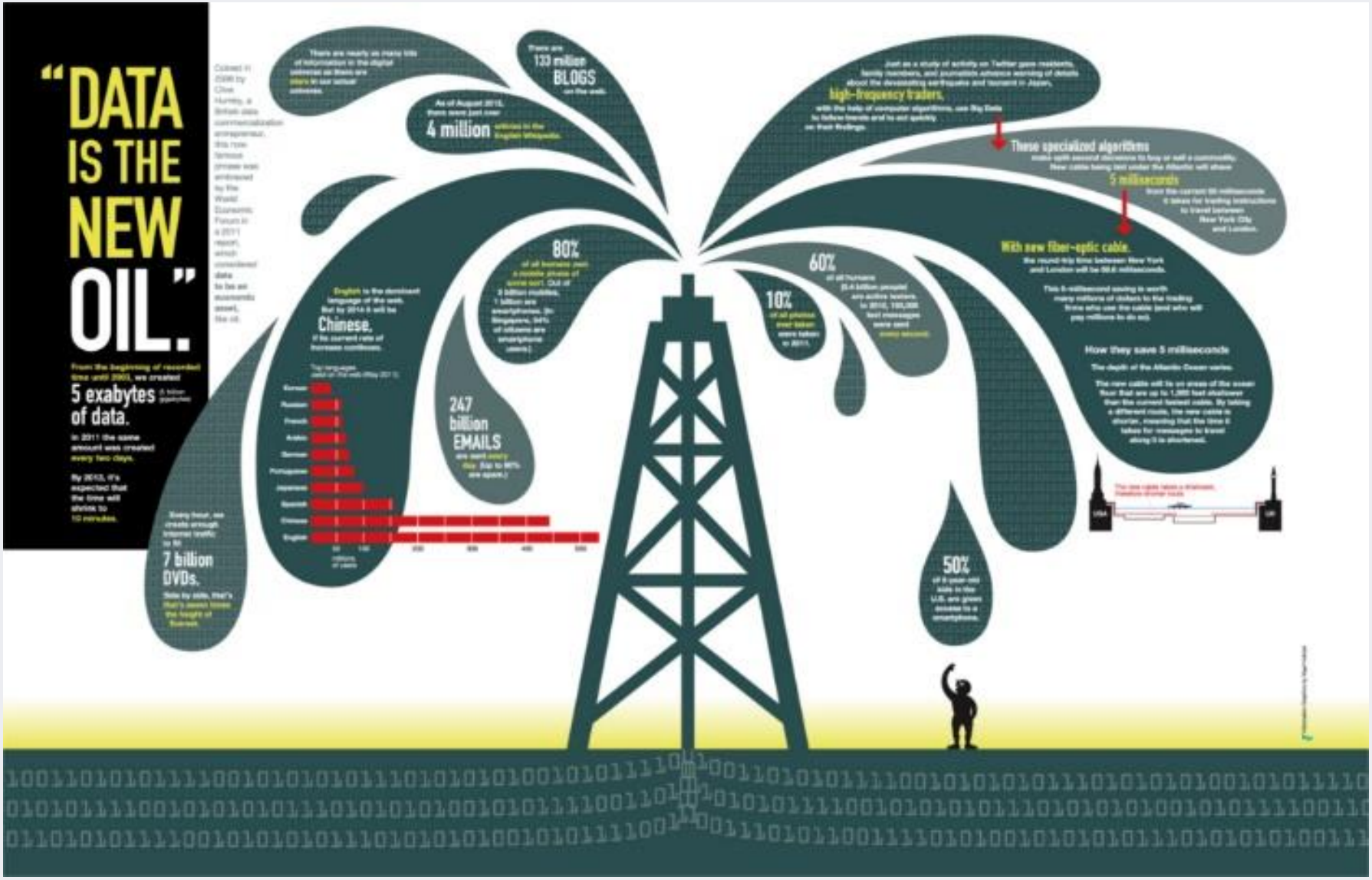
01 What is Data Analytics?

02 What Can We Do with Data Analytics?

03 Data Analytics Languages: R & Python

Data Analytics

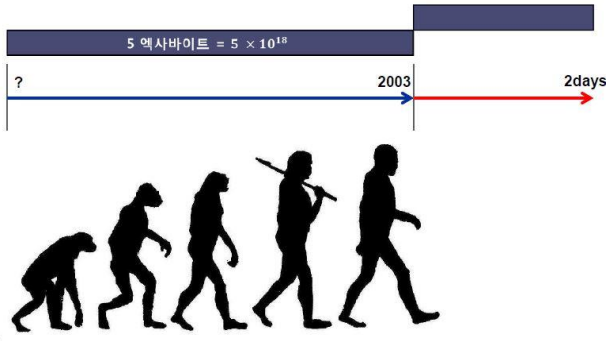
Data Analytics: The Era of "Big Data"



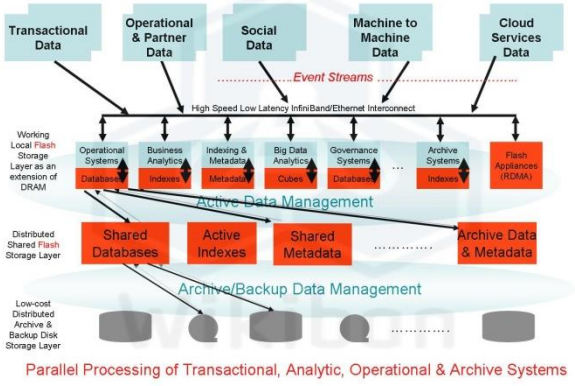
Data Analytics: The Era of "Big Data"

- 4Vs in Big Data

Volume



Velocity

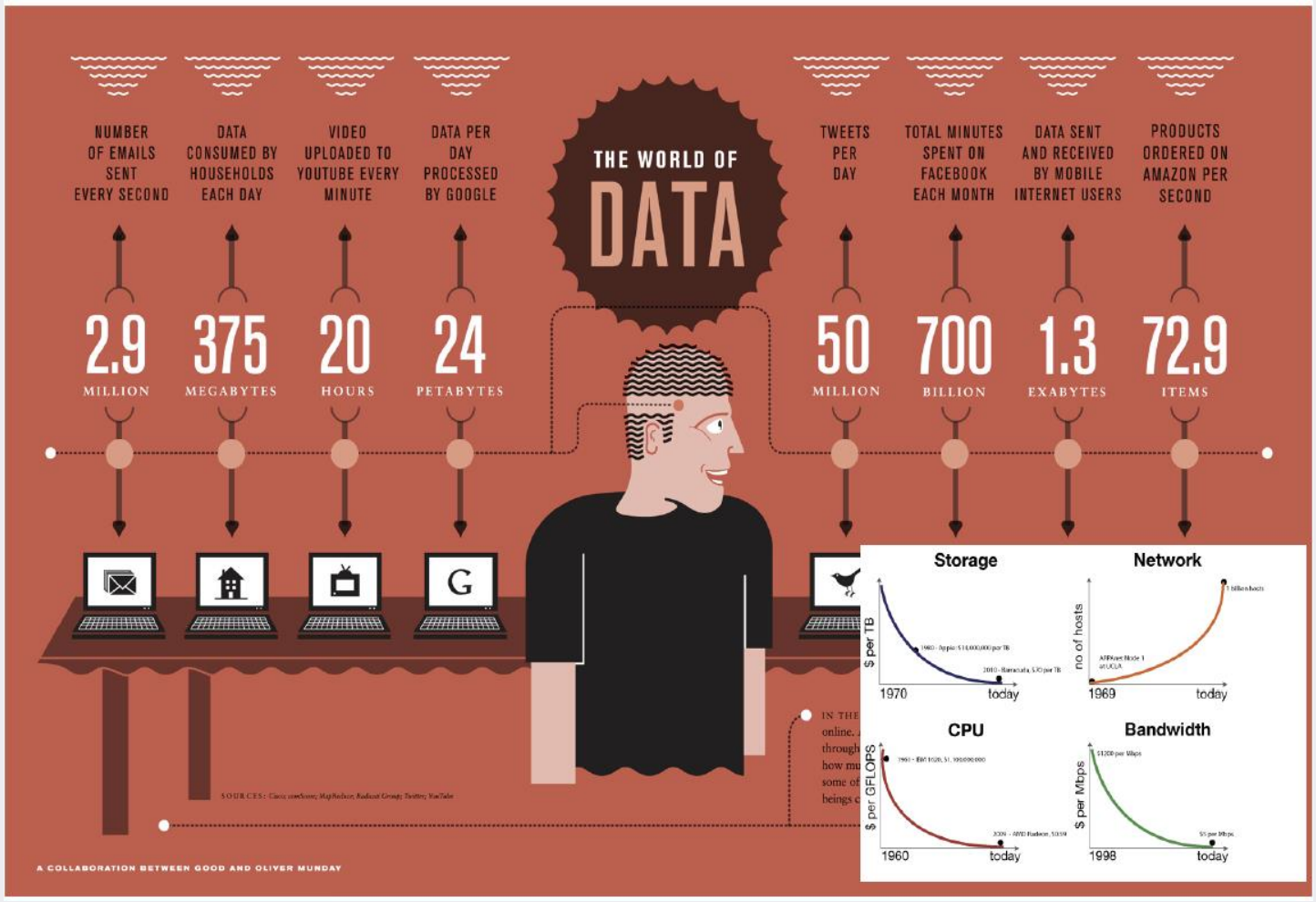


Variety

Value

Data Analytics: The Era of "Big Data"

- Volume of Big Data

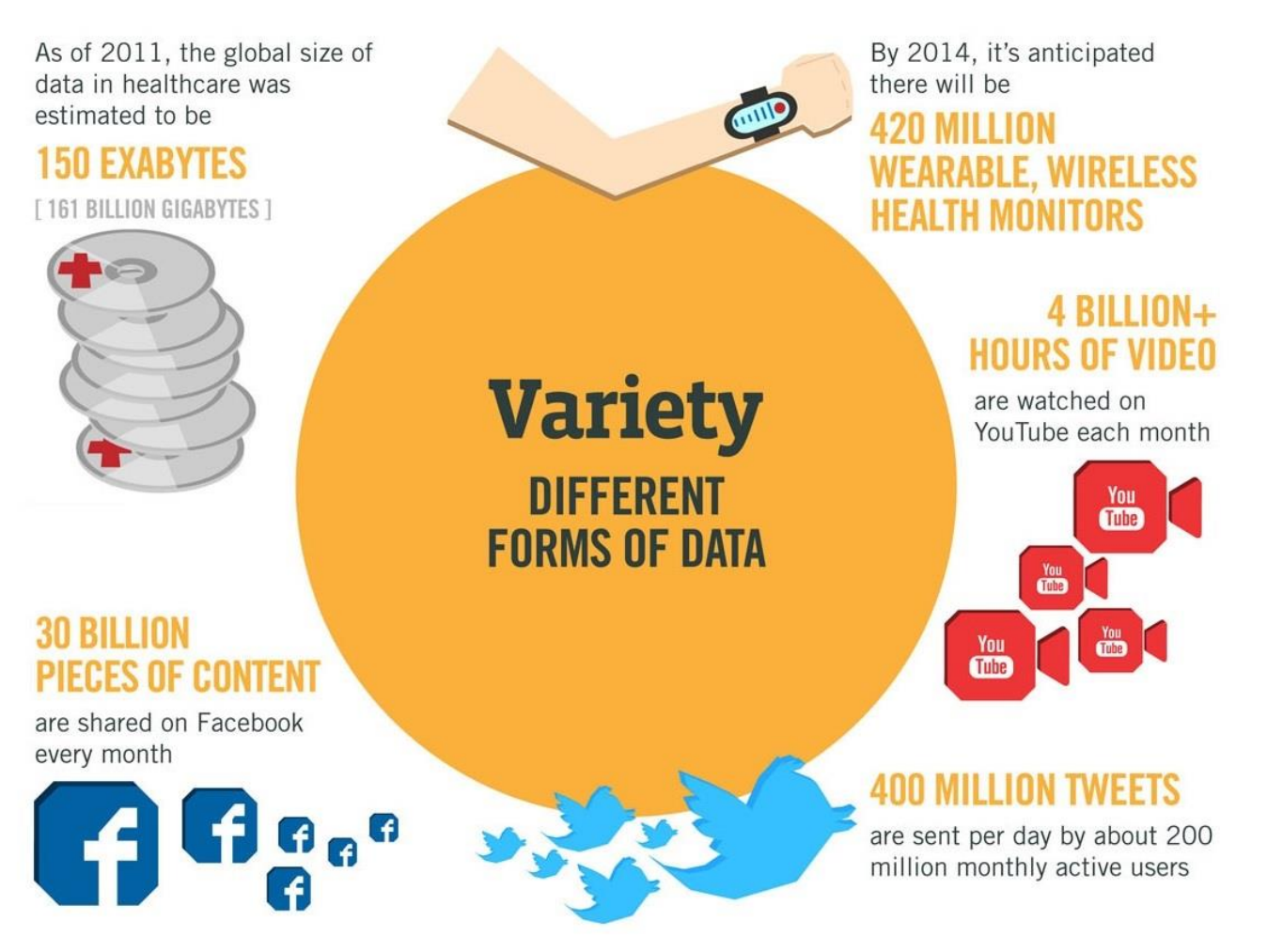


Data Analytics: The Era of "Big Data"

- 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

SOURCE: McKinsey Global Institute analysis

Data Analytics

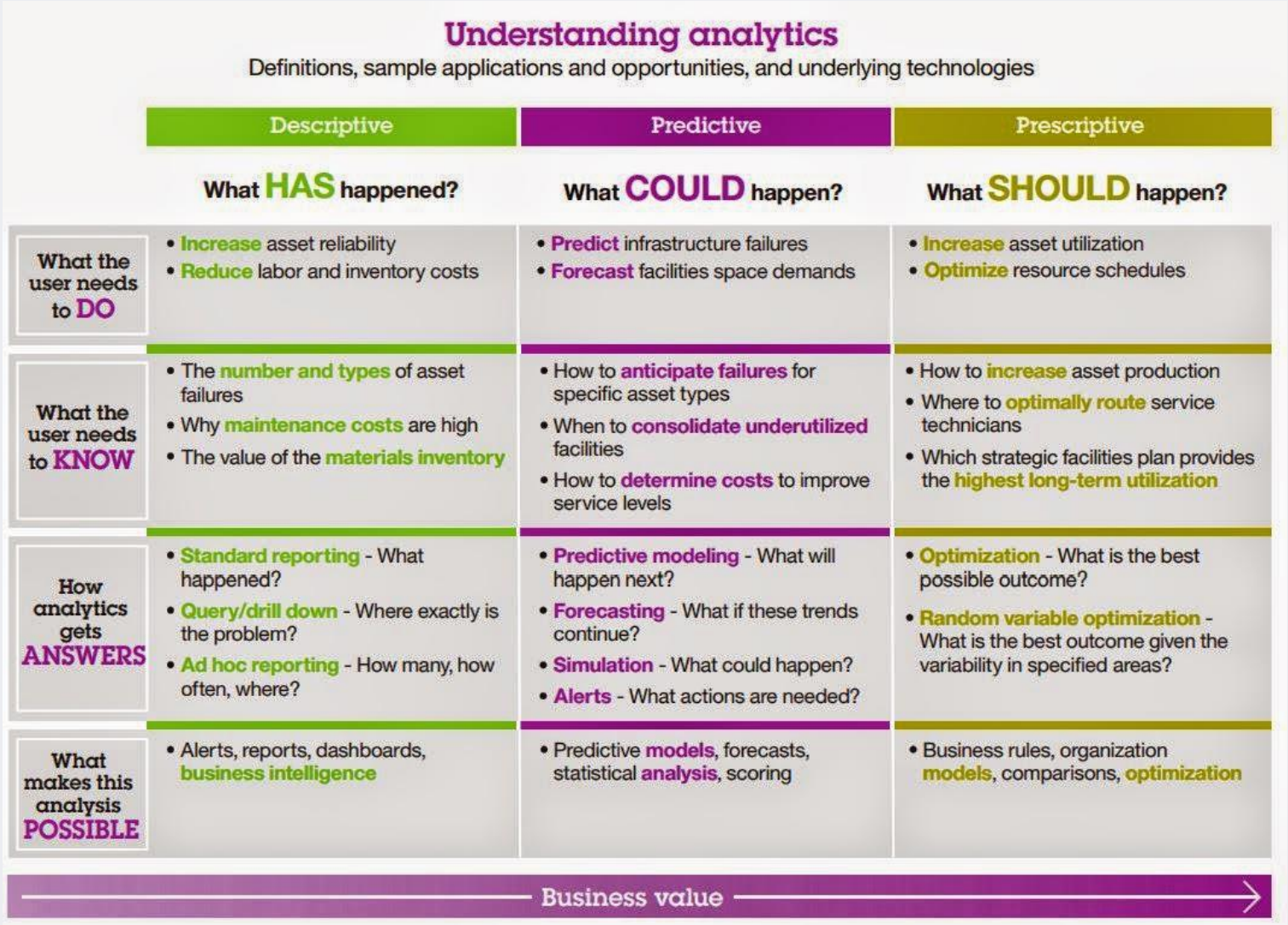
- Data Analytics in Industrial Engineering

Data Analytics

- Data Analytics by Amazon

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

Business value

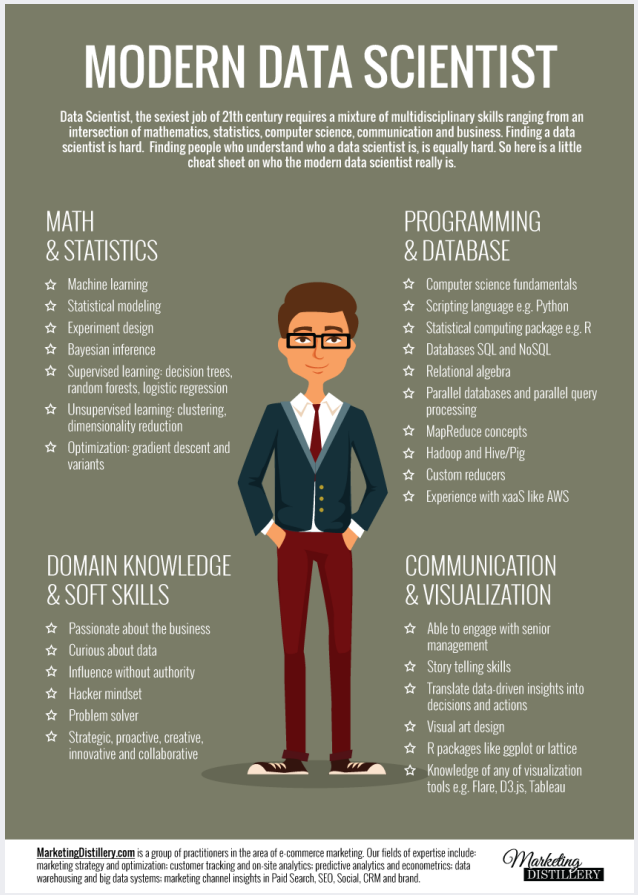
Understanding Analytics

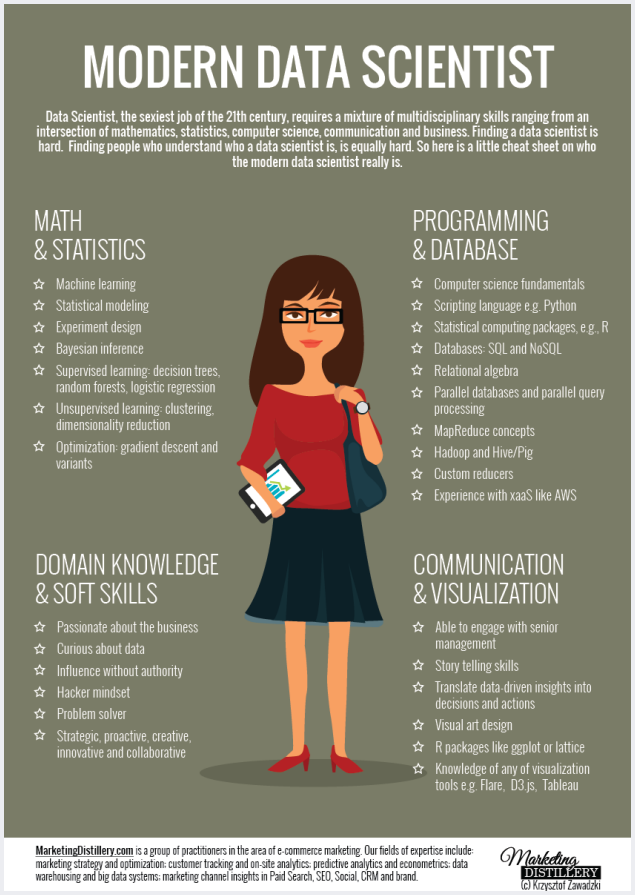
- Business Intelligence vs. Advanced Analytics

	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

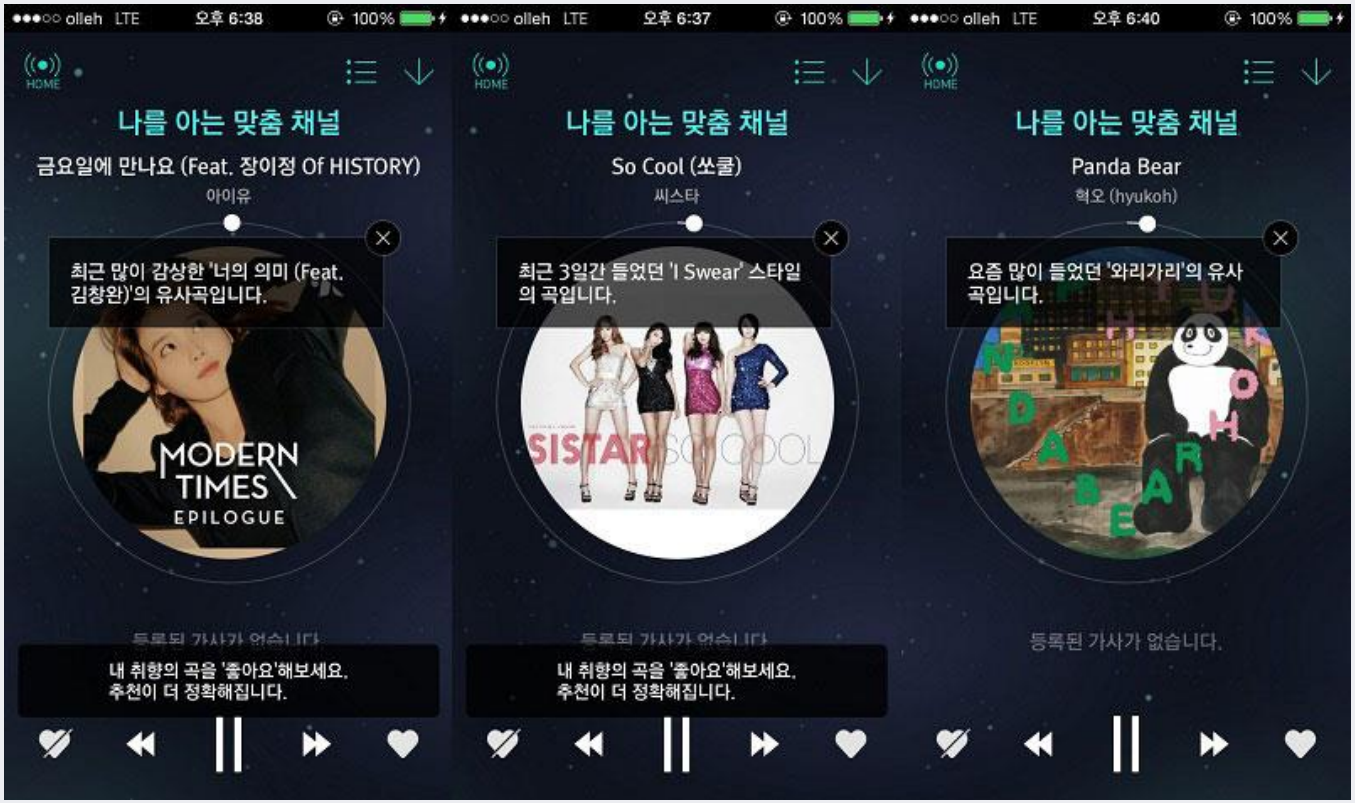
- Data Scientist: The Sexiest Job of the 21st Century
  - ✓ Harvard Business Review
  - ✓ https://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century





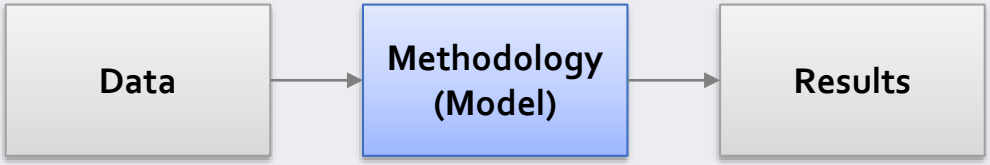
Data Science Tasks

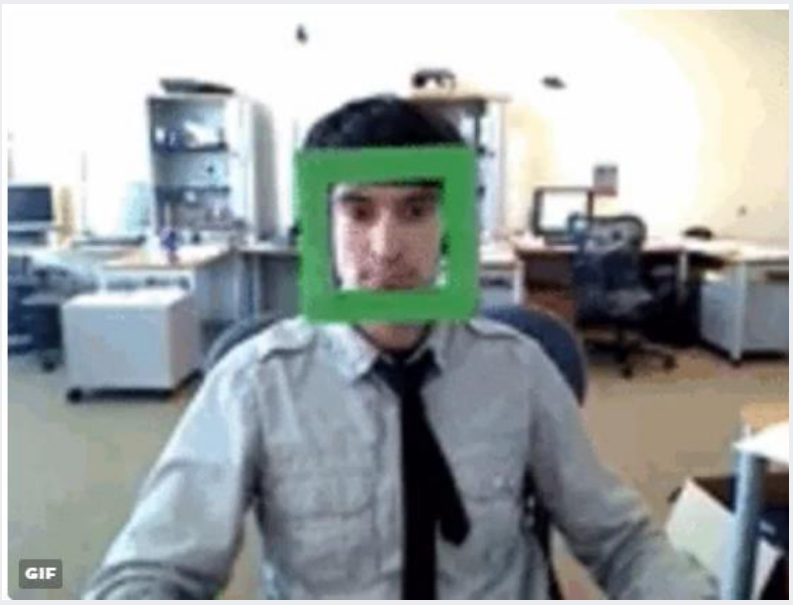
- Data Mining
  - ✓ The process of exploration and analysis, by automatic or semi-automatic means, of large quantities of data in order to discover meaningful patterns and rules. (Berry and Linoff, 1997, 2000)



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

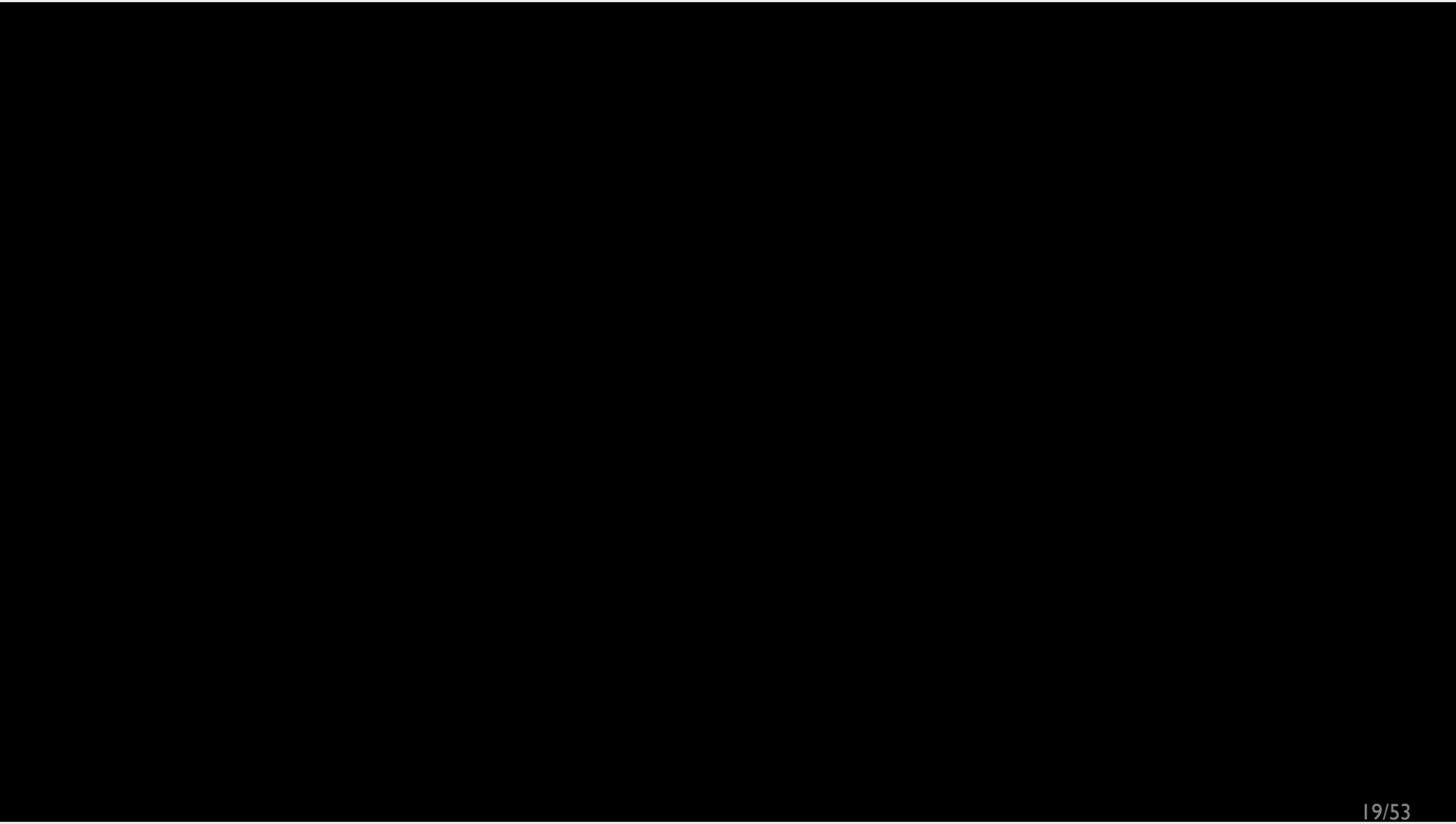
- Machine Learning Models in Self-Driving Cars

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

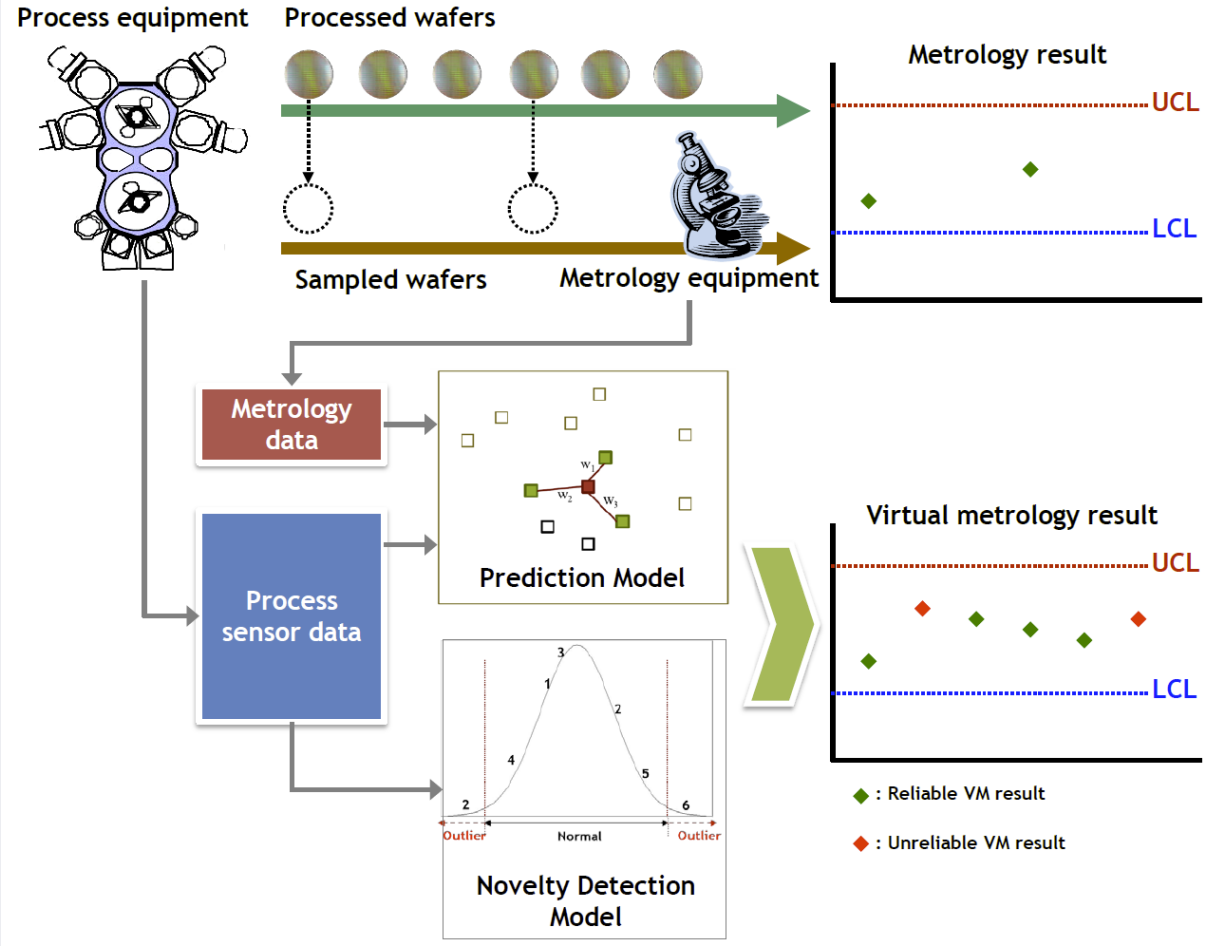
01 What is Data Analytics?

02 What Can We Do with Data Analytics?

03 Data Analytics Languages: R & Python

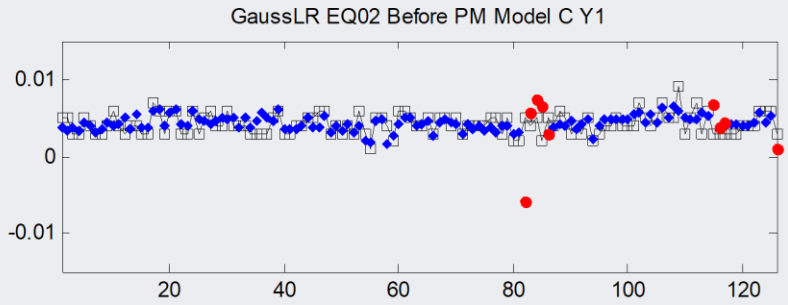
Process Monitoring & Control

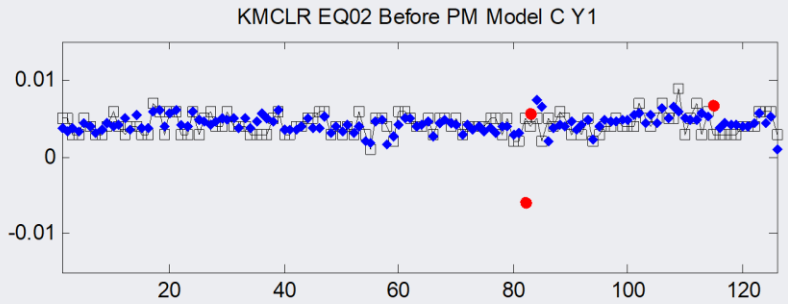
Input Data	Algorithms	Output
Sensor Data from Production Equipment	Regression + Novelty Detection	Metrological Values With Reliability Scores

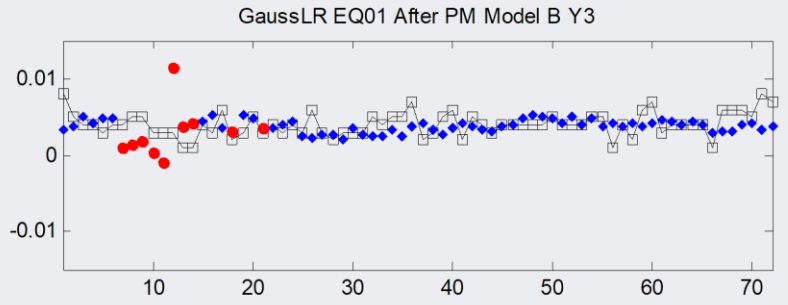


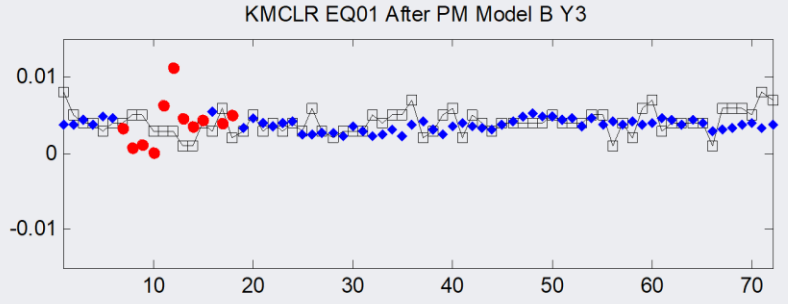
Process Monitoring & Control

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









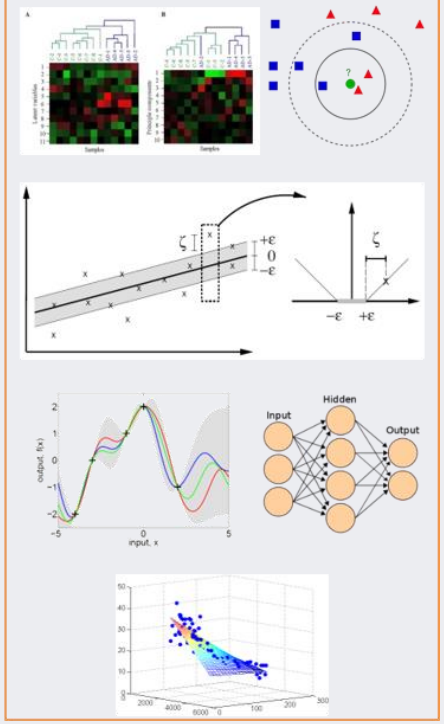
Are You a Valid User?

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

Through various input devices

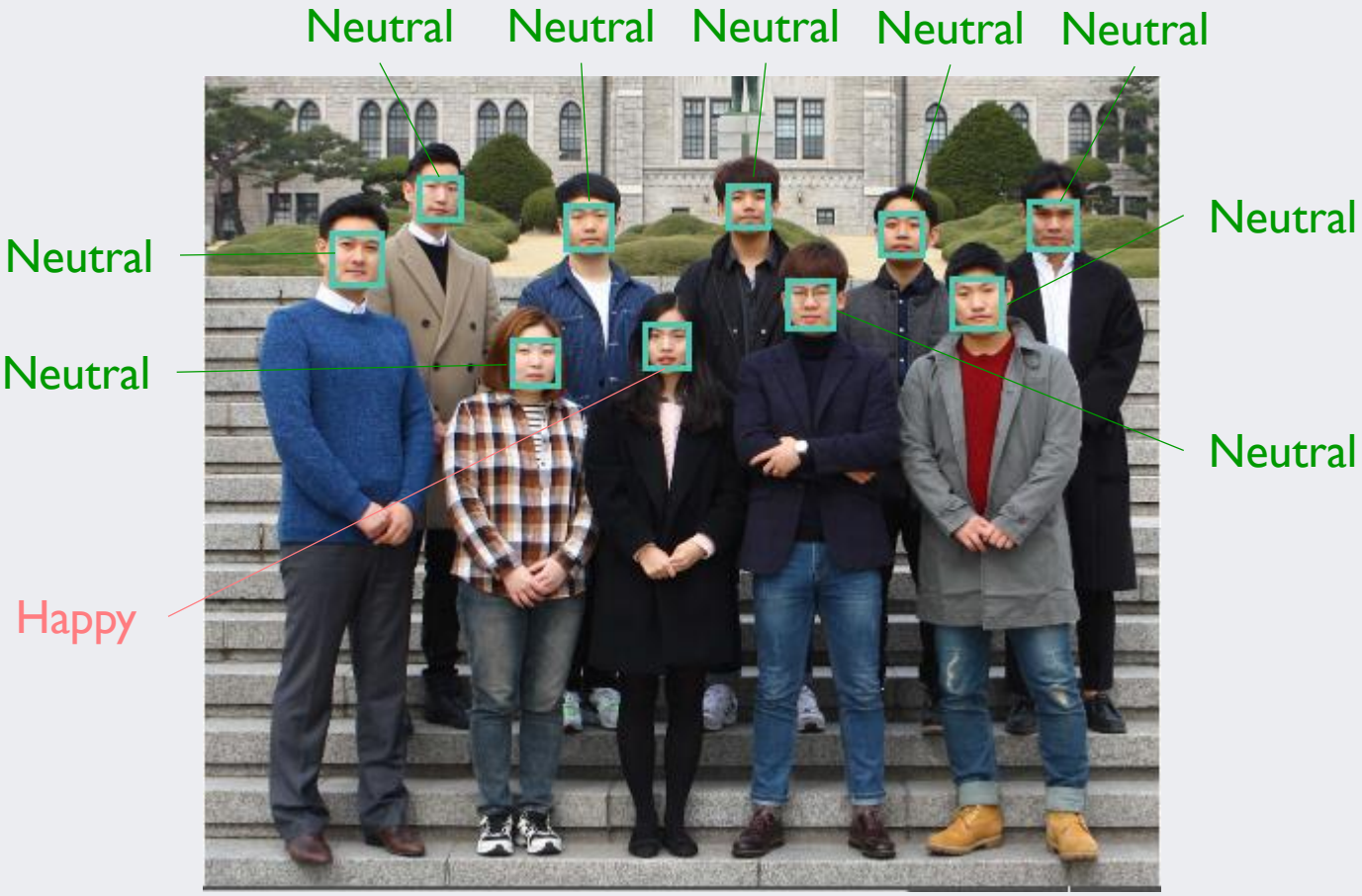
In any stages

Customized (real-time) authenticator



Who Looks Happy?

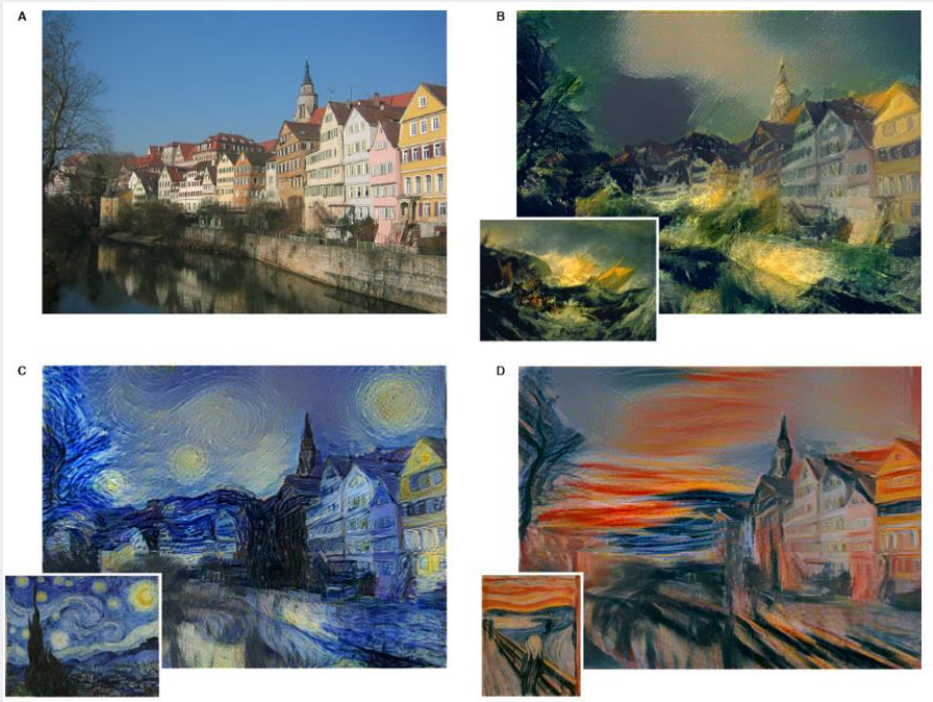
Input Data	Algorithms	Output
Image	Convolutional Neural Networks	Emotions Recognized

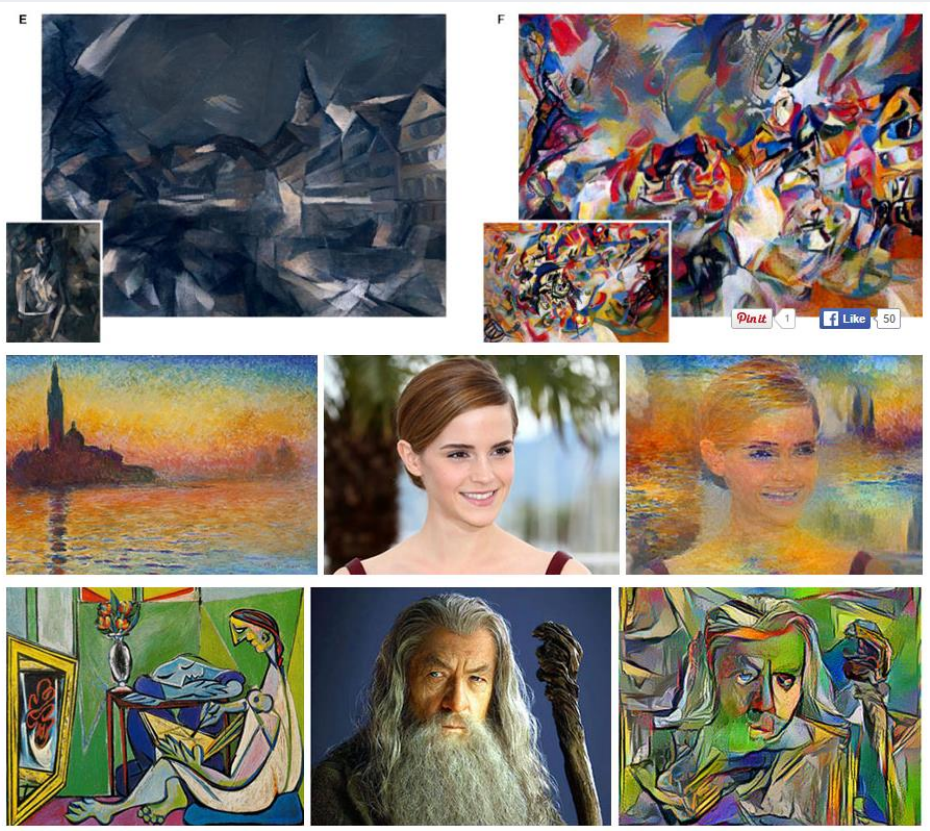


https://www.microsoft.com/cognitive-services/en-us/emotion-api

Favorite Artistic Style

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

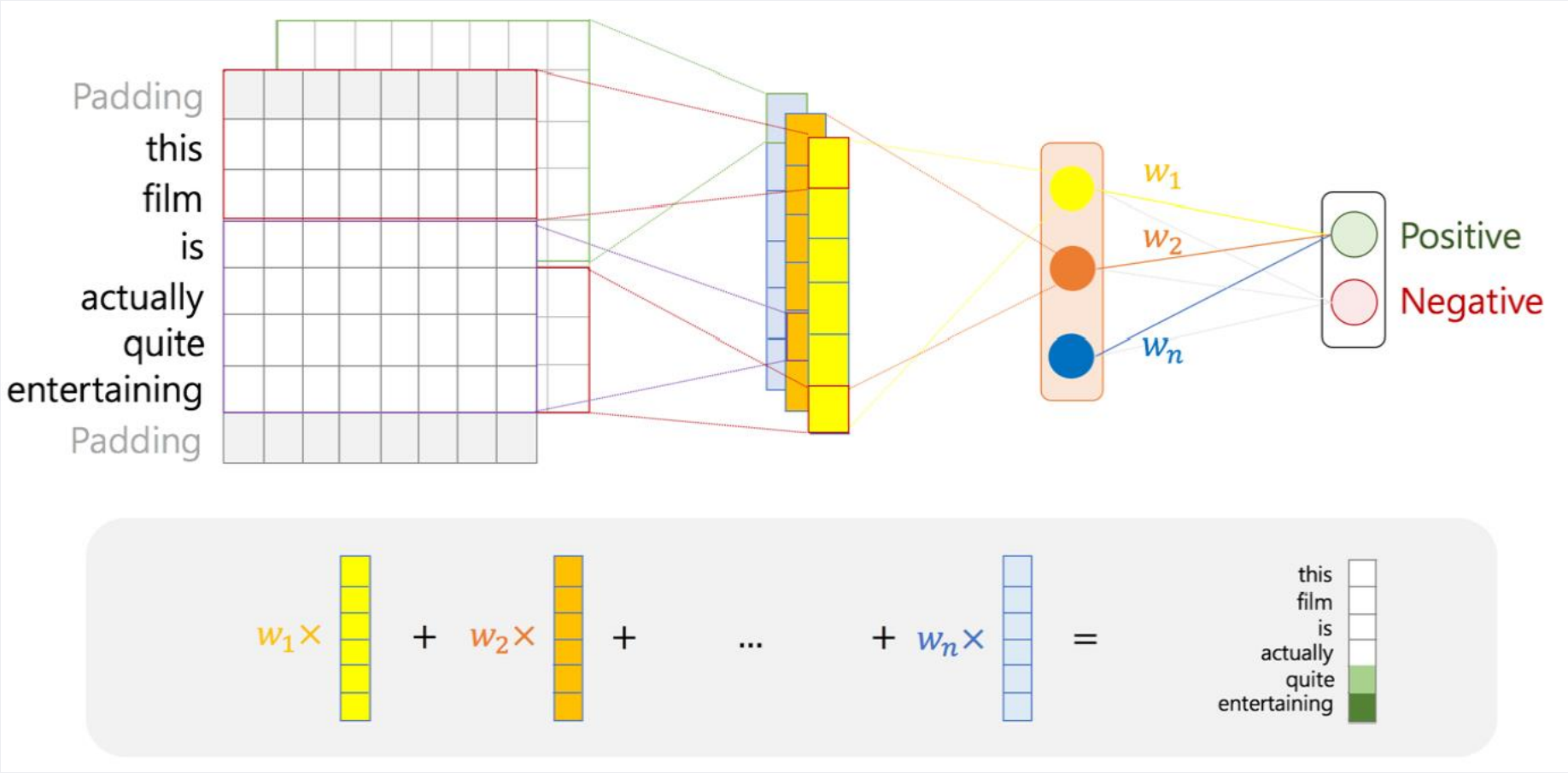




Gatys, L. A., Ecker, A. S., & Bethge, M. (2015). A neural algorithm of artistic style. arXiv preprint arXiv:1508.06576.

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 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**

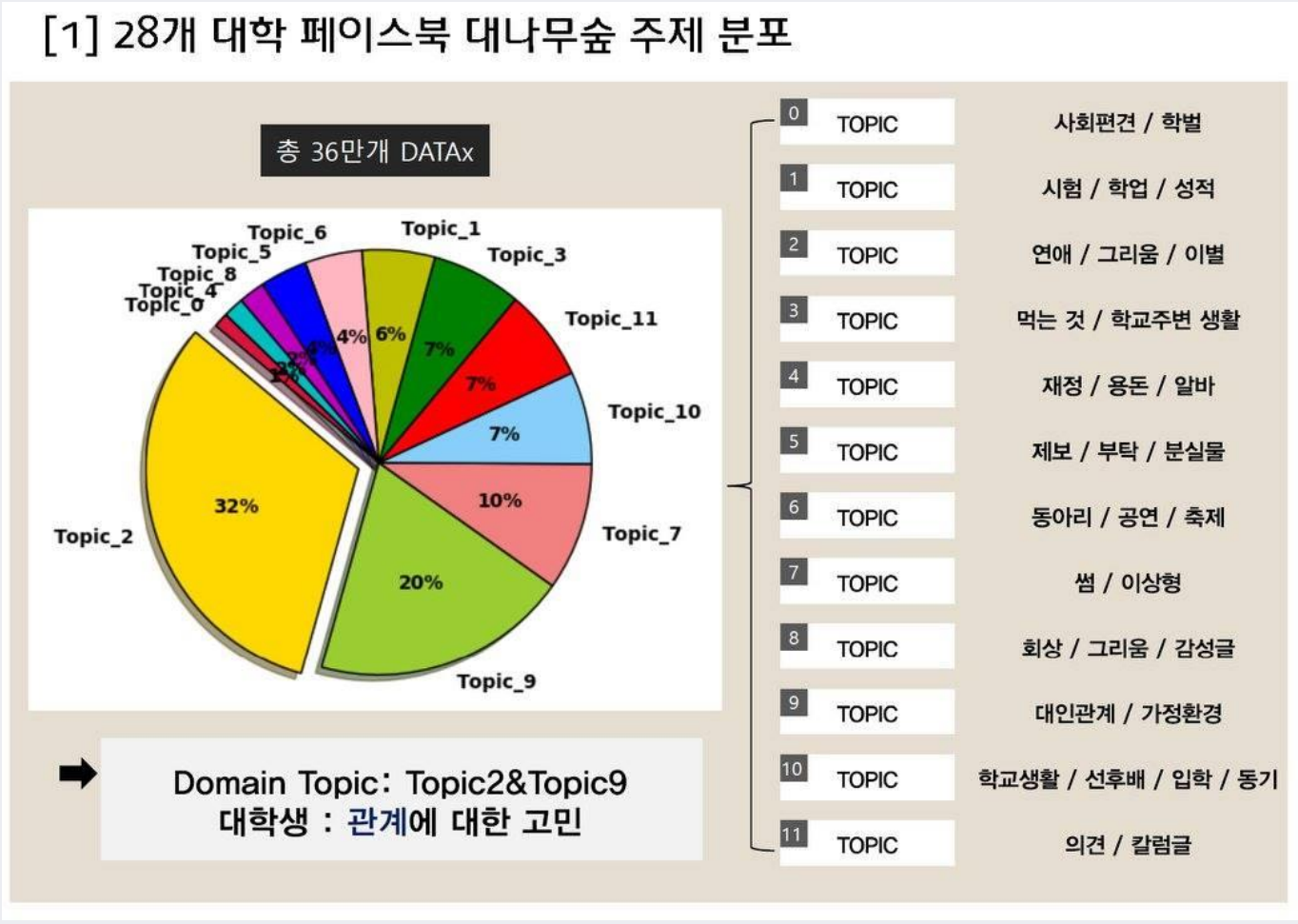
Understanding Emotions

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. *(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**

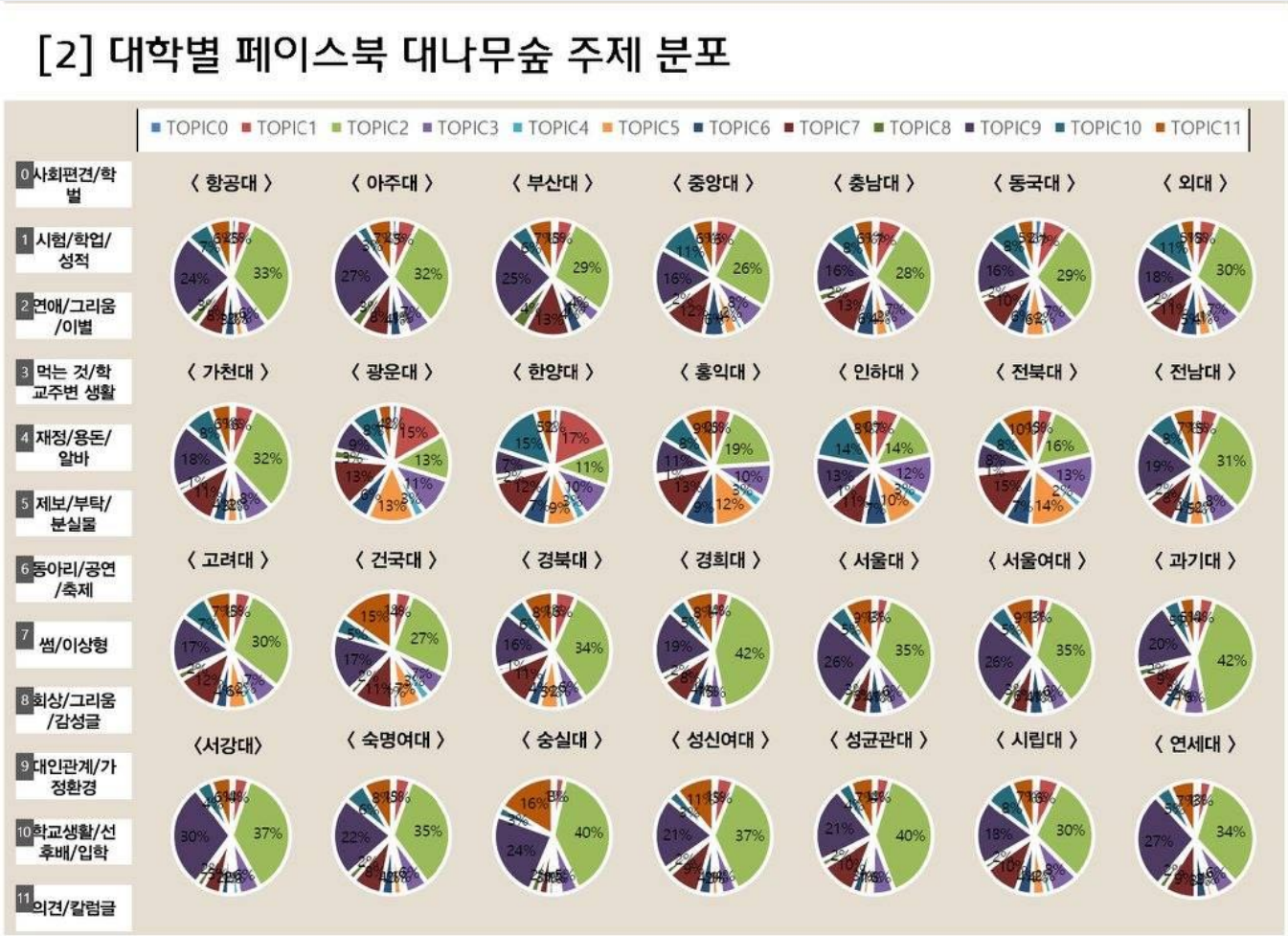
What are the Main Topics in Anonymous Posts of University Students?

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



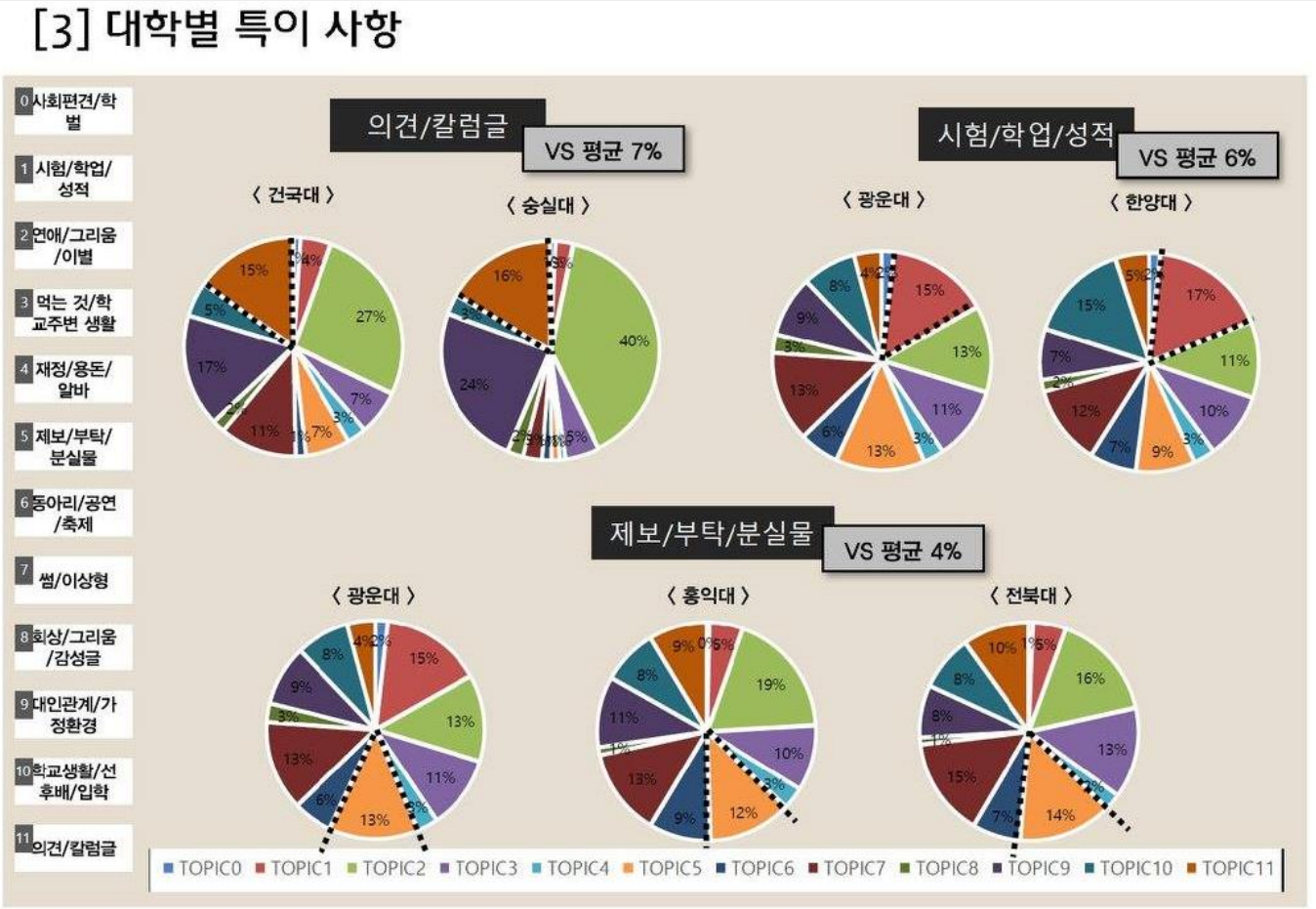
What are the Main Topics in Anonymous Posts of University Students?

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



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

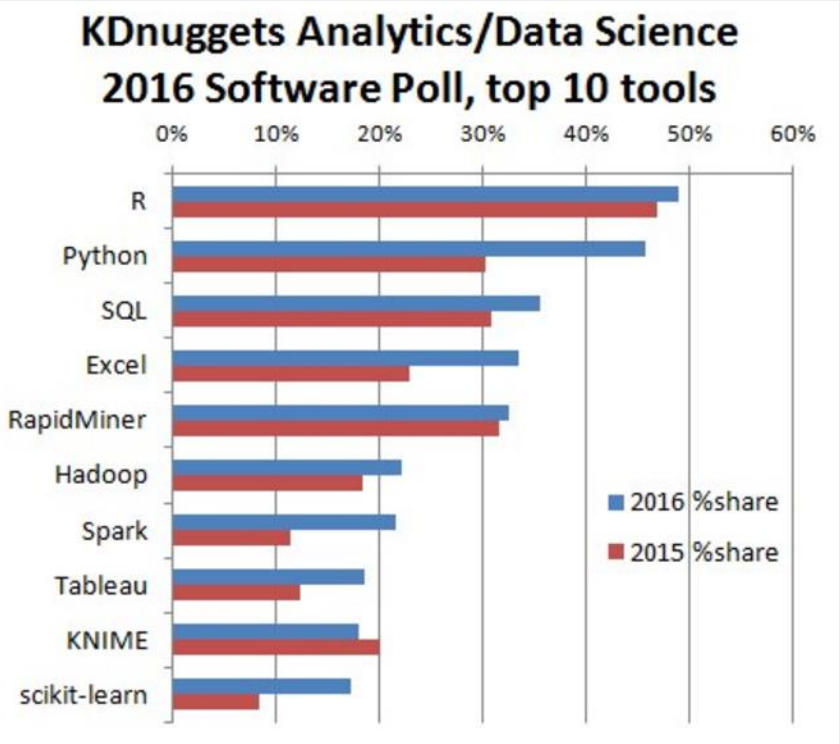
- Learn Data Science, not Programming
  - ✓ R vs. Python, different brushes
  - ✓ Do not choose between R & Python, learn both
  - ✓ It strengthens your data science communication skills
  - ✓ It boosts your data science career
  - ✓ It is not that hard

R vs. Python

- Some interesting polls

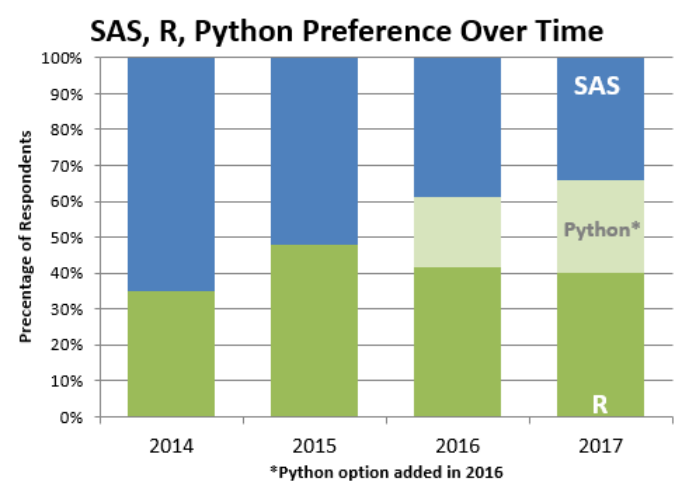
Next table has the top 10 most popular tools in 2016 poll

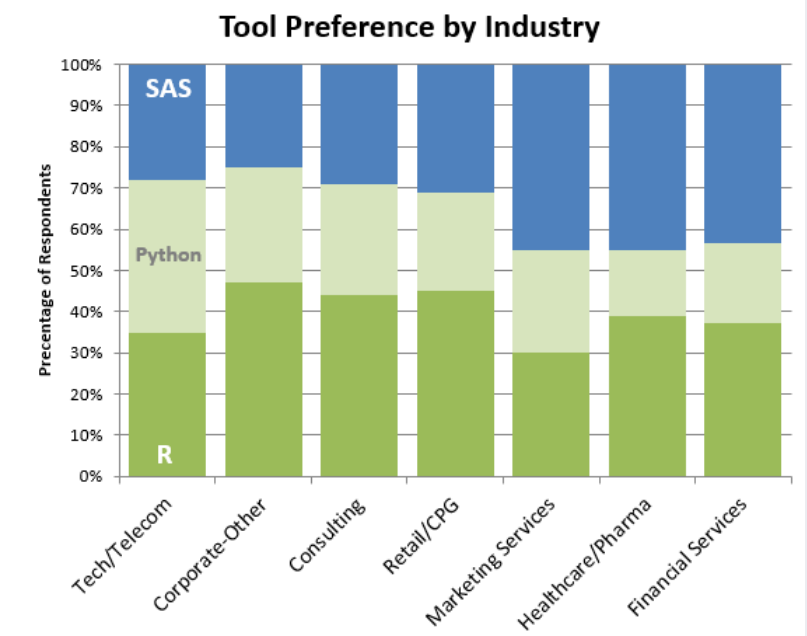
Tool	2016 % share	% change	% alone
R	49%	+4.5%	1.4%
Python	45.8%	+51%	0.1%
SQL	35.5%	+15%	0%
Excel	33.6%	+47%	0.2%
RapidMiner	32.6%	+3.5%	11.7%
Hadoop	22.1%	+20%	0%
Spark	21.6%	+91%	0.2%
Tableau	18.5%	+49%	0.2%
KNIME	18.0%	-10%	4.4%
scikit-learn	17.2%	+107%	0%

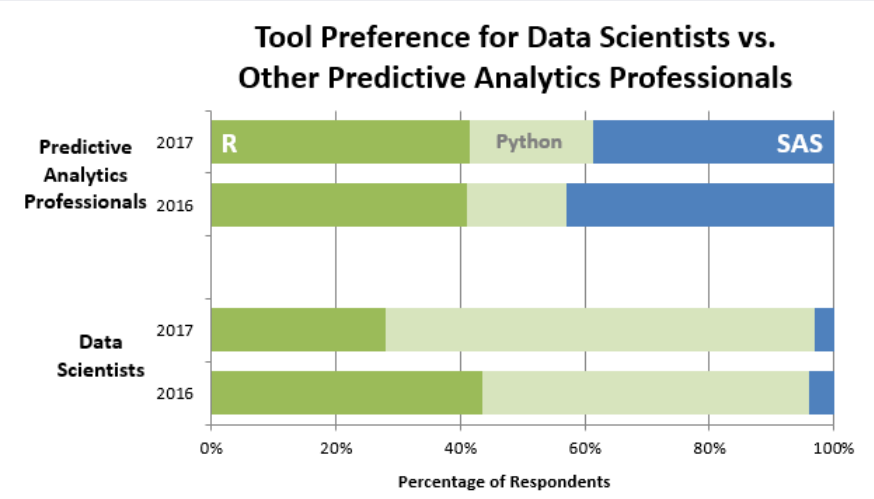


R vs. Python

- Some interesting polls

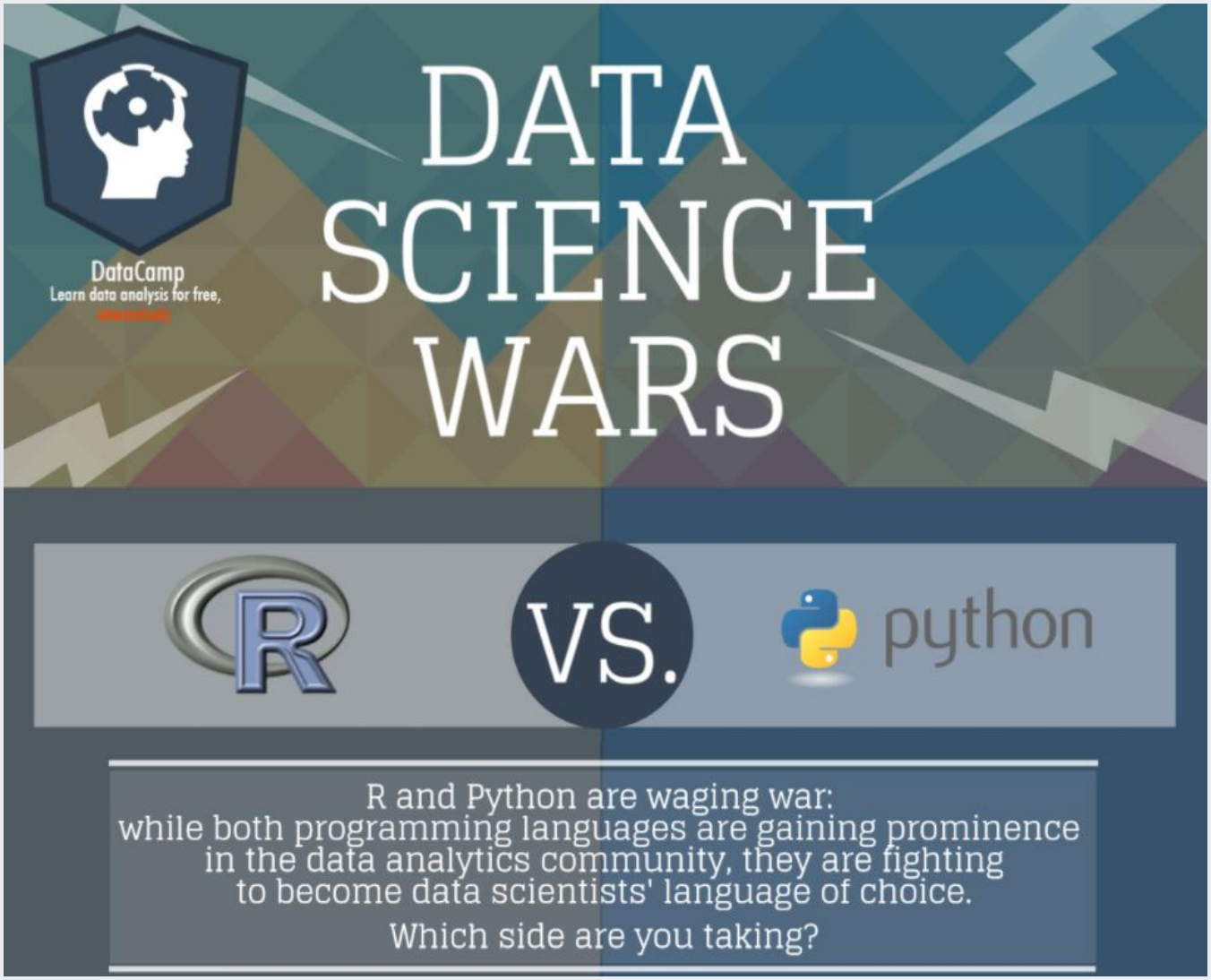






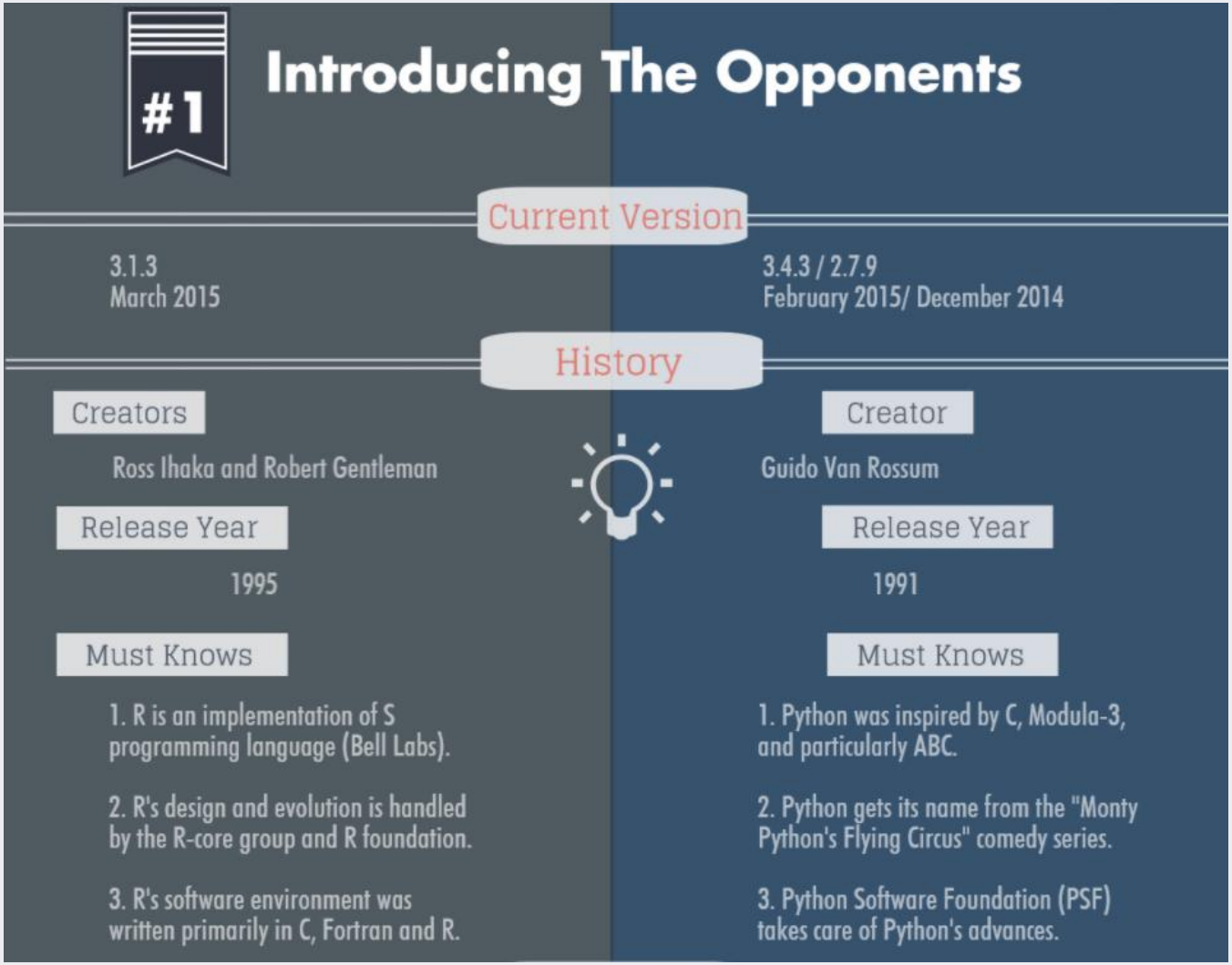
R vs. Python

- Data Science Wars: R vs. Python



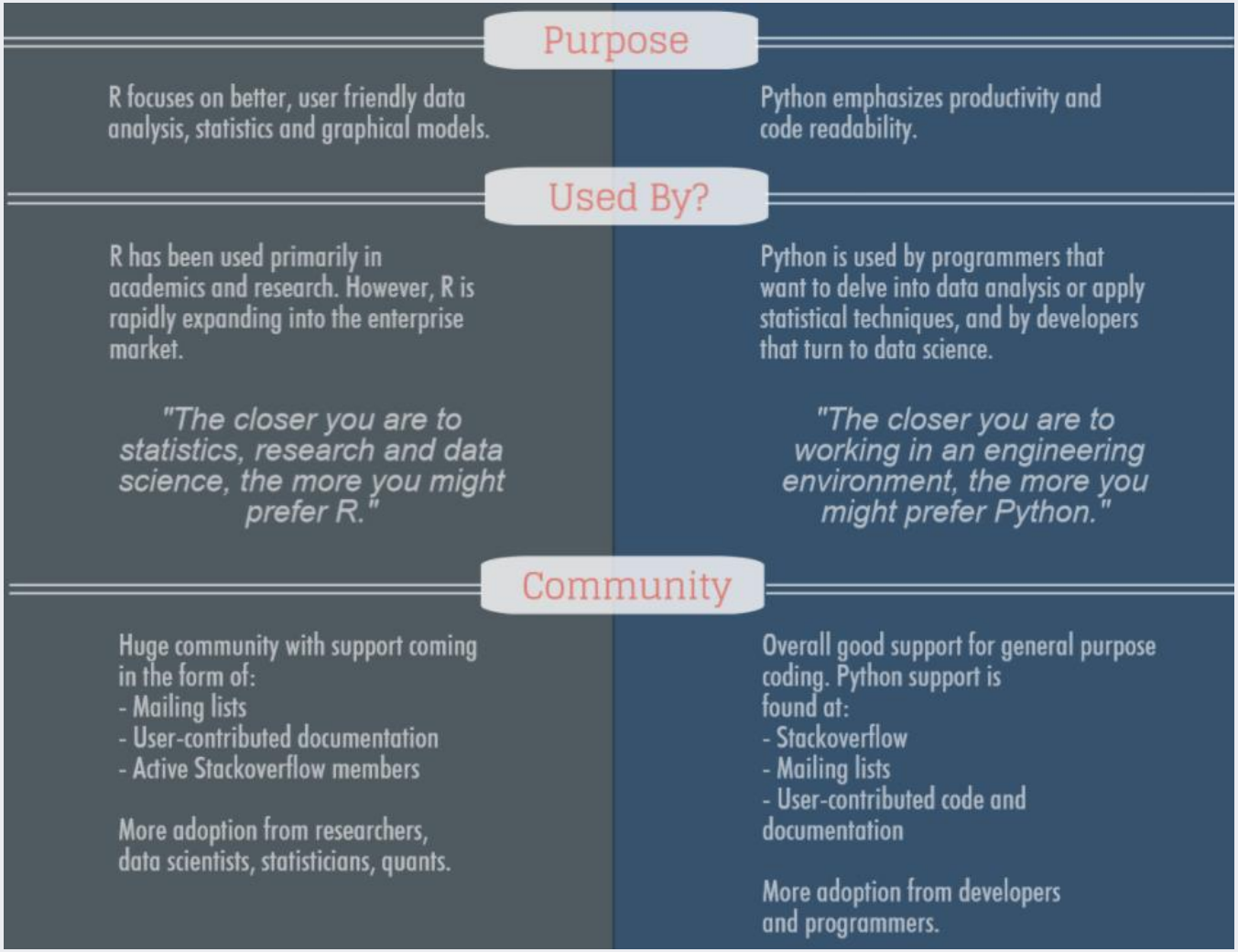
R vs. Python

- Data Science Wars: R vs. Python



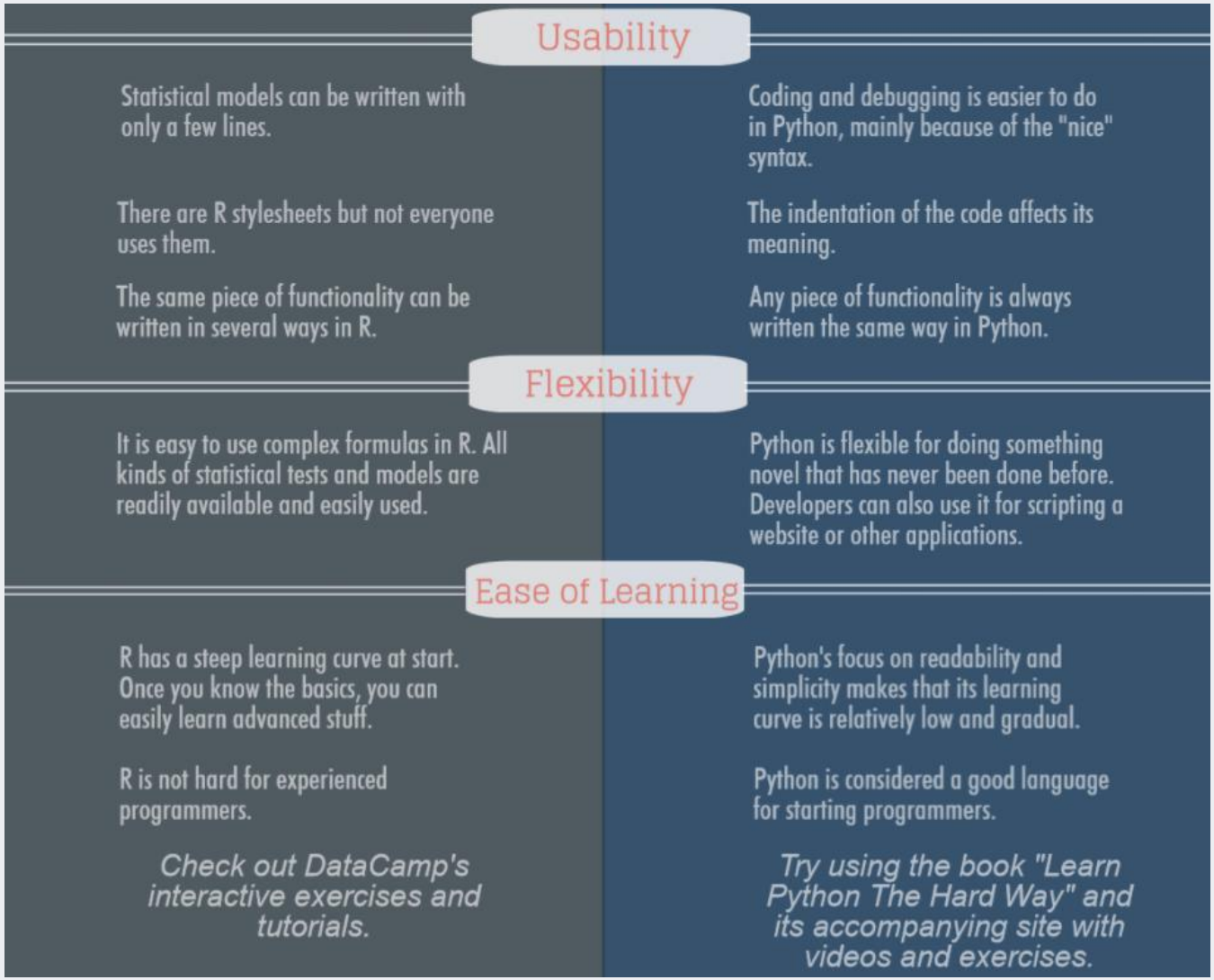
R vs. Python

- Data Science Wars: R vs. Python



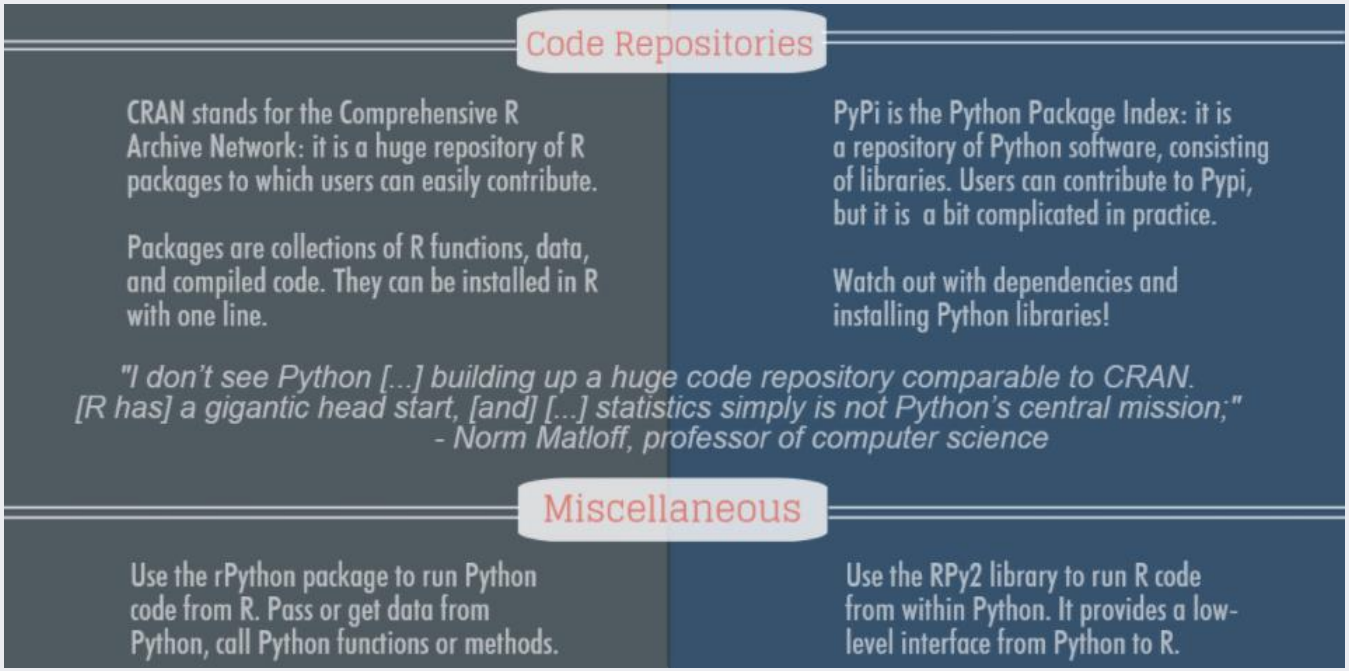
R vs. Python

- Data Science Wars: R vs. Python



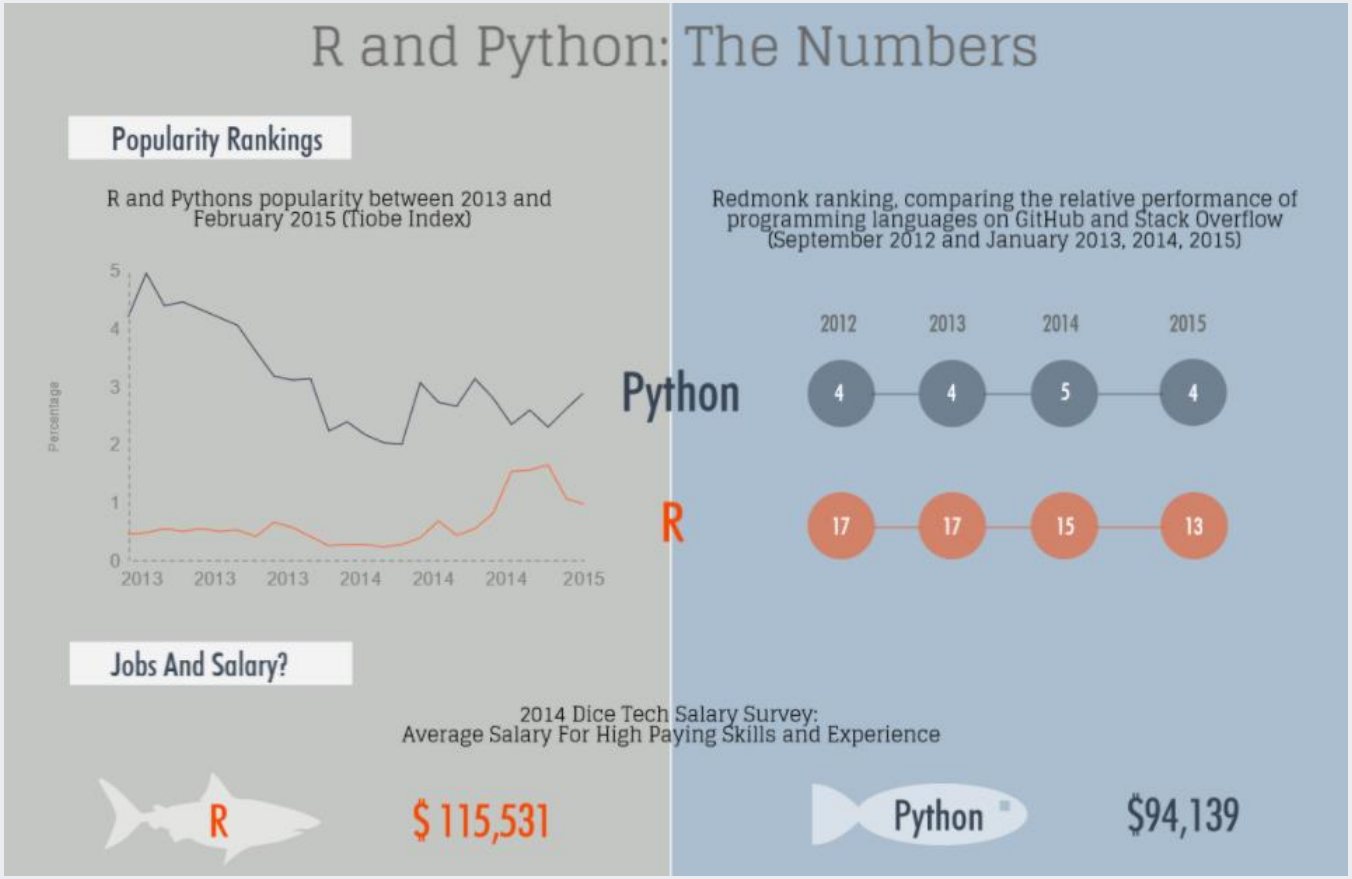
R vs. Python

- Data Science Wars: R vs. Python



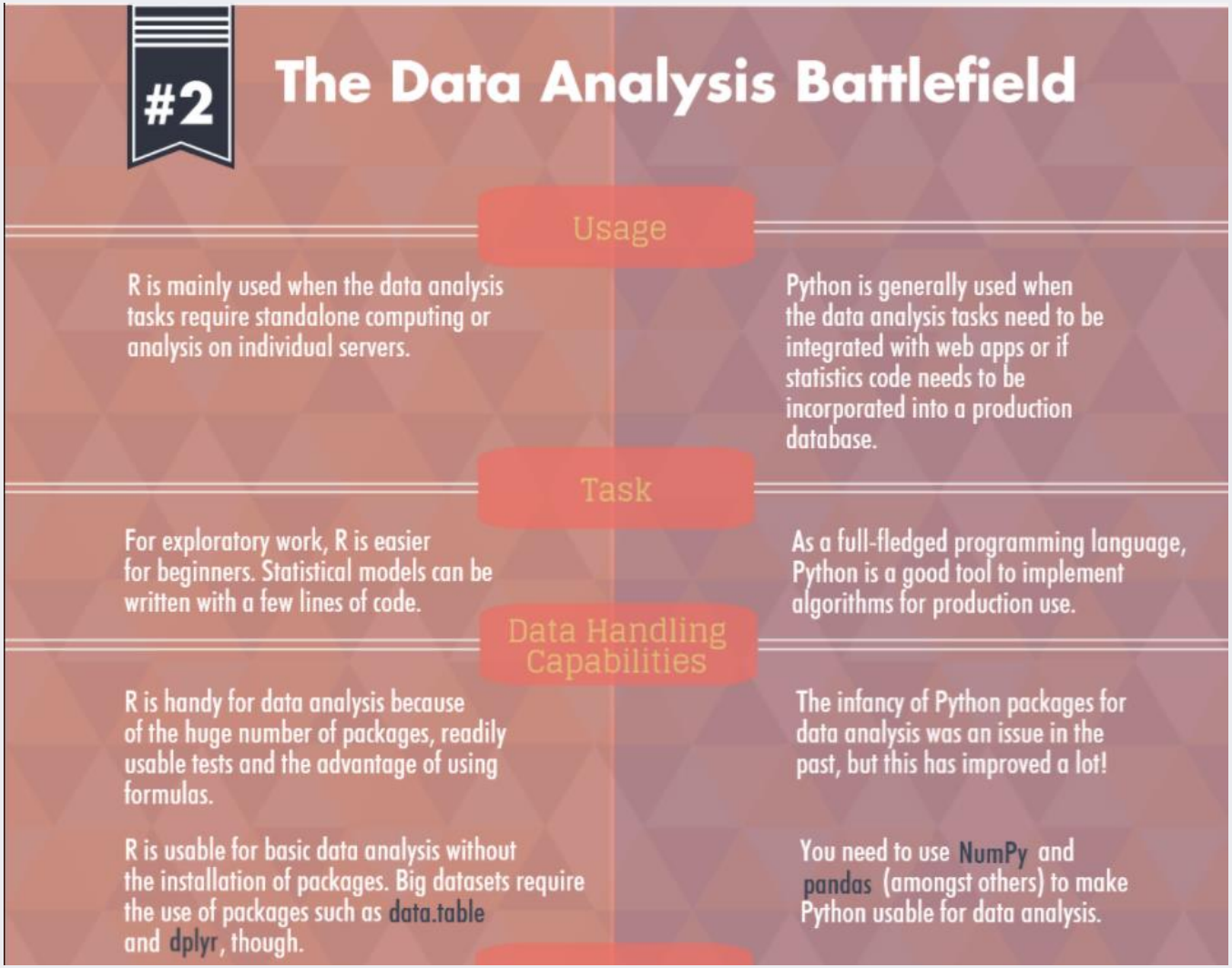
R vs. Python

- Data Science Wars: R vs. Python

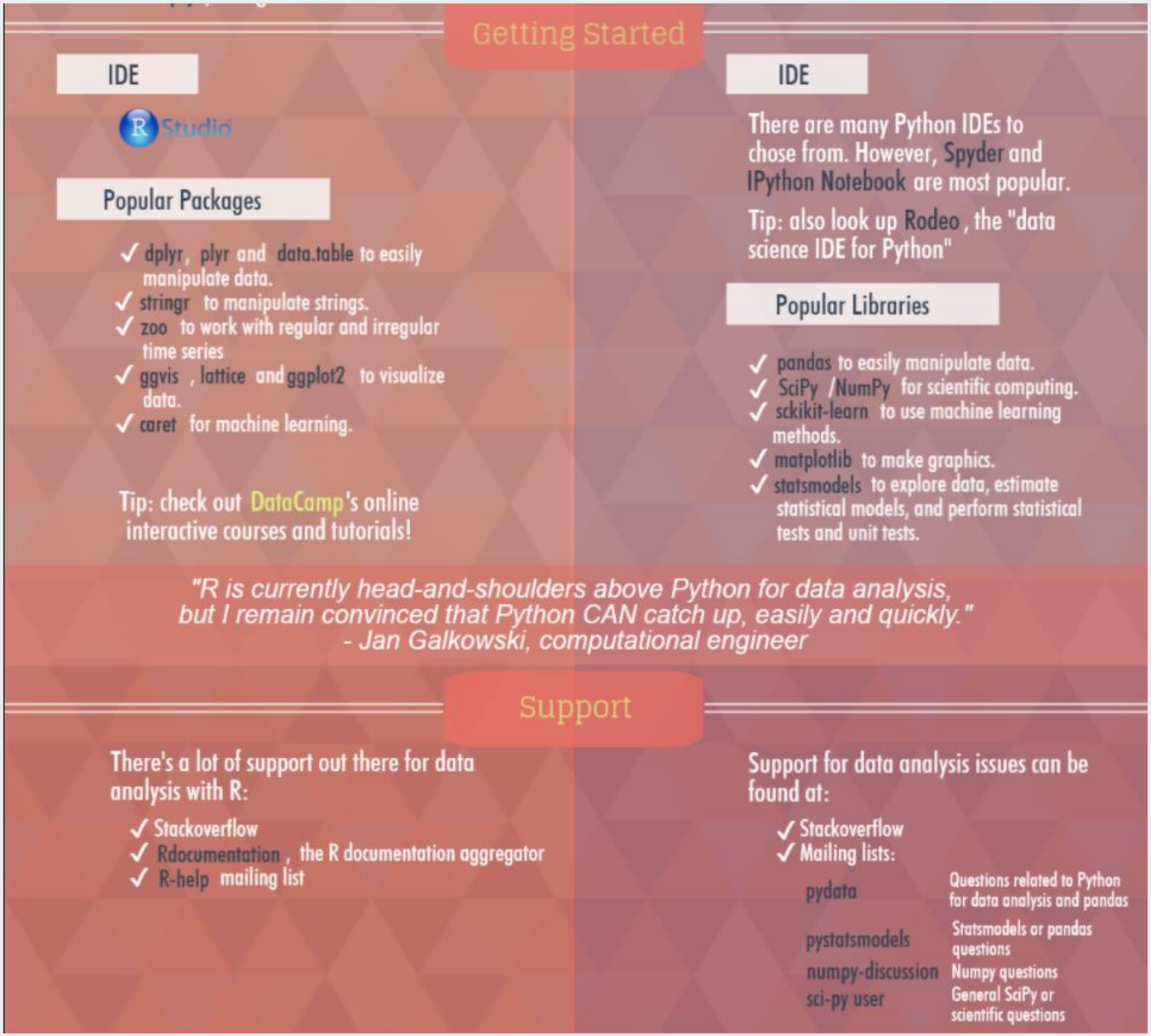


R vs. Python

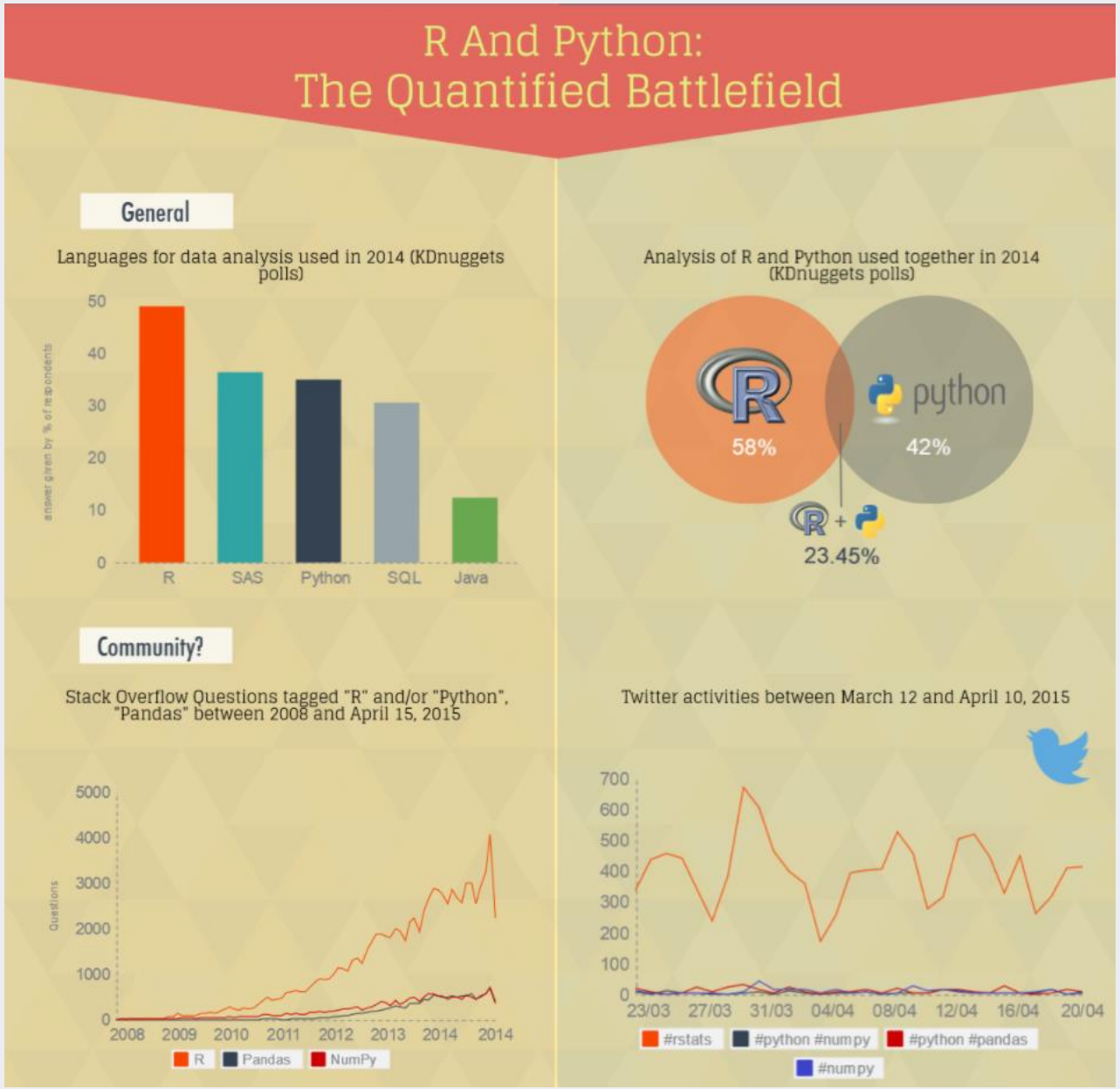
- Data Science Wars: R vs. Python



R vs. Python


Getting Started
IDE
RStudio
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!
IDE
There are many Python IDEs to chose 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.
✓ sckikit-learn to use machine learning methods.
✓ matplotlib to make graphics.
✓ statsmodels to explore data, estimate statistical models, and perform statistical tests and unit tests.
"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
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
✓ Mailing lists:
pydata — Questions related to Python for data analysis and pandas
pystatsmodels — Statsmodels or pandas questions
numpy-discussion — Numpy questions
sci-py user — General SciPy or scientific questions

R vs. Python


R And Python:
The Quantified Battlefield
General
Languages for data analysis used in 2014 (KDnuggets polls)
answer given by % of respondents
50
40
30
20
10
0
R
SAS
Python
SQL
Java
Analysis of R and Python used together in 2014 (KDnuggets polls)
58%
python
42%
23.45%
Community?
Stack Overflow Questions tagged "R" and/or "Python", "Pandas" between 2008 and April 15, 2015
Questions
5000
4000
3000
2000
1000
0
2008
2009
2010
2011
2012
2013
2014
2014
R
Pandas
NumPy
Twitter activities between March 12 and April 10, 2015
700
600
500
400
300
200
100
0
23/03
27/03
31/03
04/04
08/04
12/04
16/04
20/04
#rstats
#python #numpy
#python #pandas
#numpy

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

The R Project

Rich ecosystem of cutting-edge interface packages available to communicate between open-source languages.

This allows you to string your workflow together, which is especially useful for data analysis.

Packages are available at:

- Cran — "Task Views" page lists a wide range of tasks for which R packages are available
- Bioconductor — Open source software for bioinformatics
- GitHub — web-based Git repository hosting service

Rdocumentation

Search through all these sources easily with Rdocumentation, the first R documentation aggregator

The R User Community

✓ Meetup groups

Some are sponsored by companies of the R community

REVOLUTION ANALYTICS

✓ Blogs & Social Media

R bloggers #rstats

Python, A General Purpose Language

Readability and Learning Curve

Just like everyday English

Python is easy and intuitive, and its emphasis on readability only magnifies these characteristics.

e.g. `print("Hello World!")`

Syntactically clear and elegant code, easily interpretable and very easy to type.

This explains why.

- ✓ Python's learning curve is relatively flat
- ✓ So many programmers are familiar with it

Also, the speed at which you can write a program is also positively impacted:

Less time coding, more time playing

The Python Testing Framework

Guarantee your code is reusable and dependable

A built-in, low barrier-to-entry testing framework that encourages good test coverage.

Python Testing Tools Taxonomy, including

- UnitTest — First unit test framework of the Python standard library
- Nose — Extends UnitTest: used in many packages such as pandas
- DocTest — Easy generation of tests based on output from the standard Python interpreter shell
- Pytest — To write small tests, while supporting complex functional testing

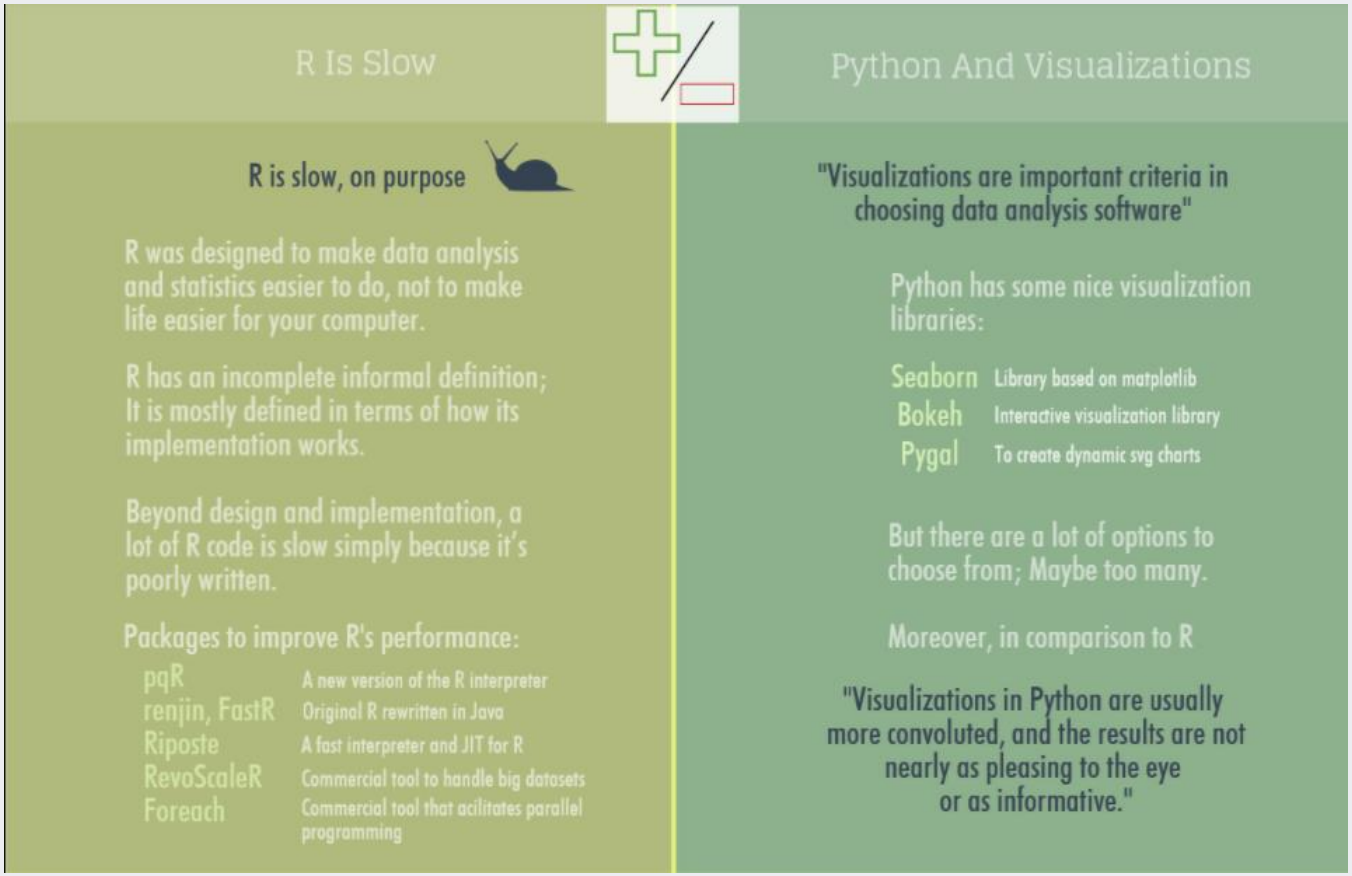
testing-in-python (TIP) mailing list

R vs. Python

- Data Science Wars: R vs. Python

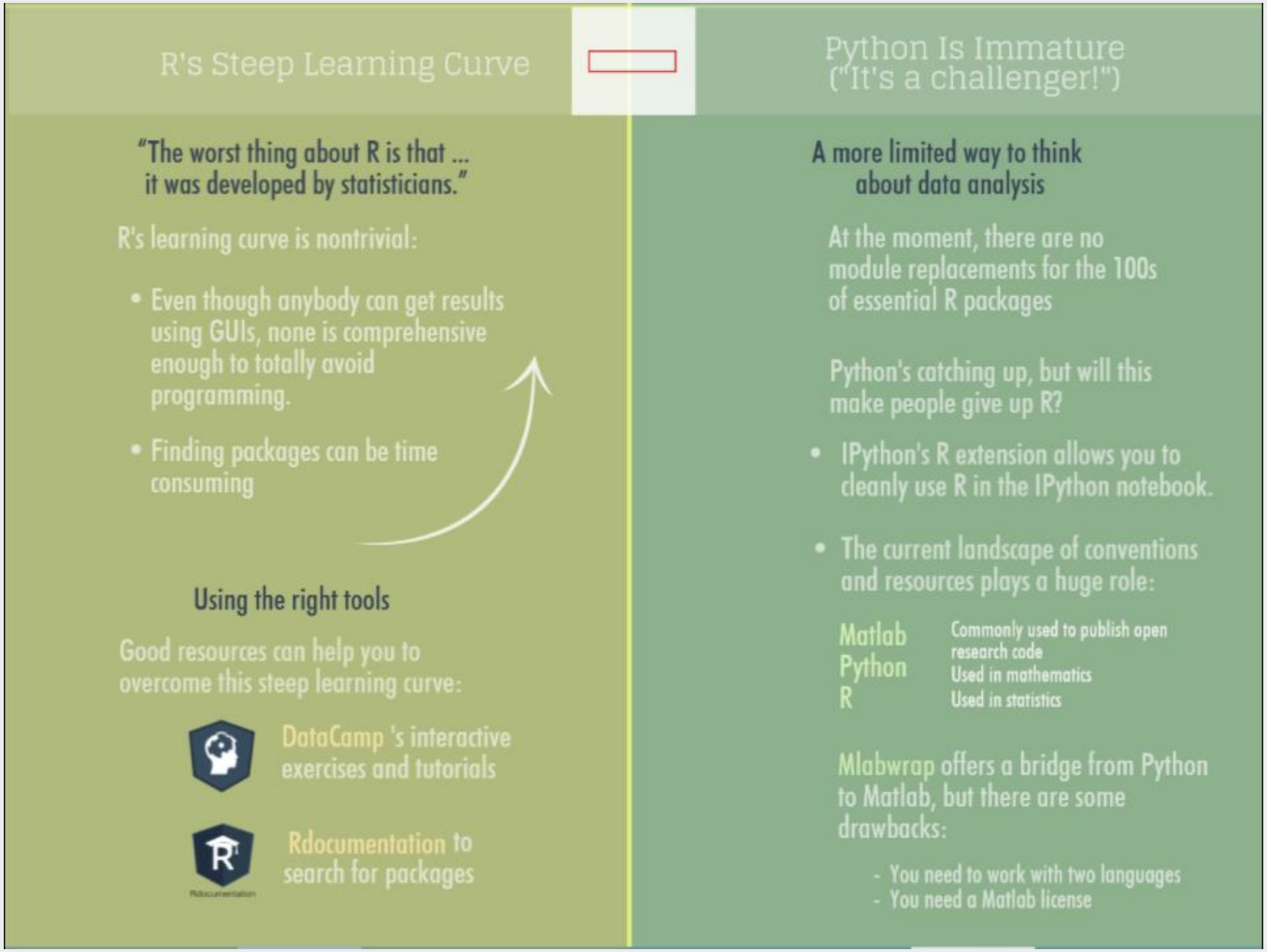
R vs. Python

- Data Science Wars: R vs. Python



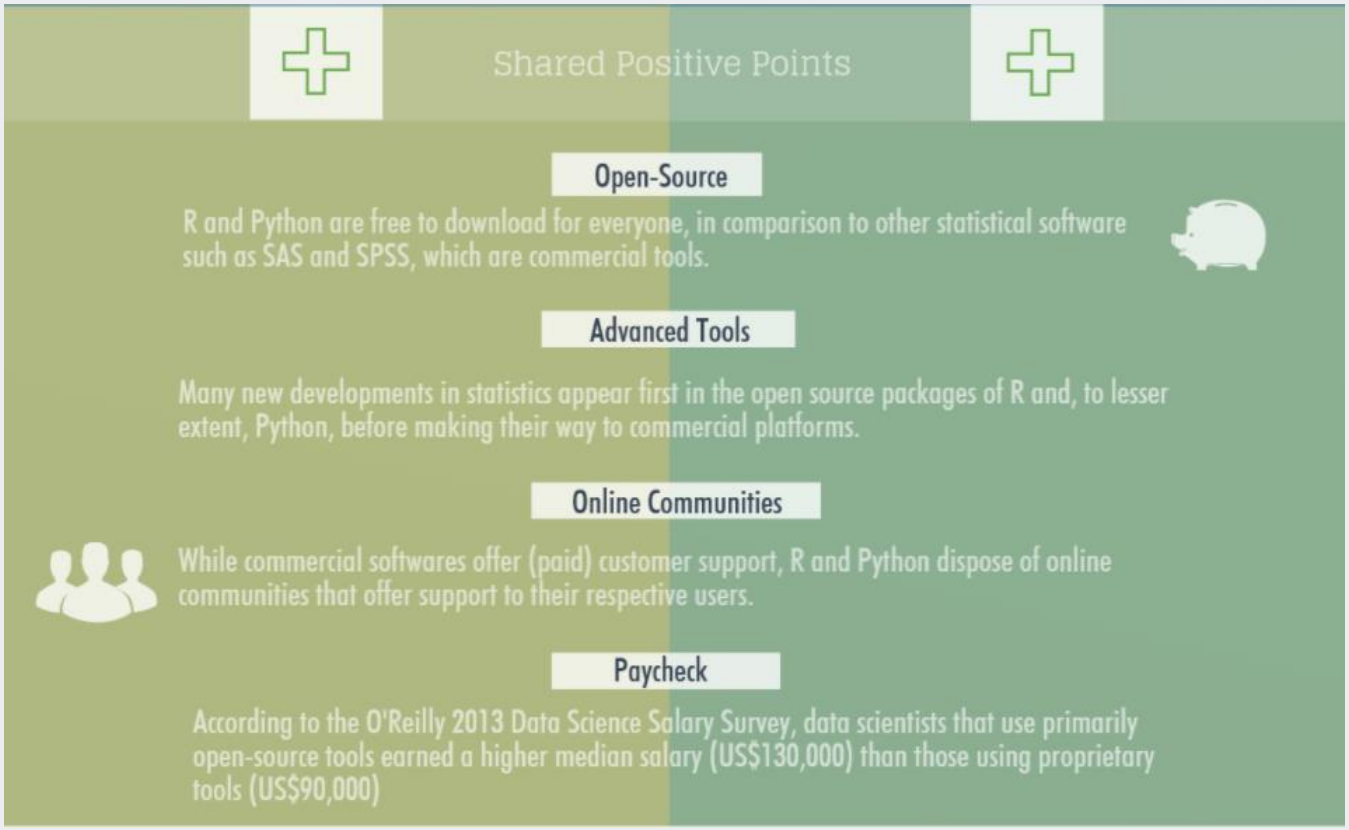
R vs. Python

- Data Science Wars: R vs. Python



R vs. Python

- Data Science Wars: R vs. Python




ANY
questions?