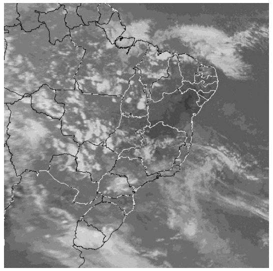


# ELAC 2019 INTRODUCTION TO DATA SCIENCE

## Day 1

Rafael Santos – rafael.santos@inpe.br
www.lac.inpe.br/~rafael.santos/talks.html

# About this Course and Lecture

# About this Course

- What is Data Science? Why a Data Science course?
- The Data Scientist. Roles of the Data Scientist. Other related roles.
- Data, where it is, how to collect it, how to organize it. Data Federation. Data Provenance.
- Tools and Techniques for Data Science.
- Analytics, Exploratory Data Analysis.
- Reproducible Research. Data Products.
- Applications, Case Studies, Projects.

# Why?

- Talks in 2015-2017, proposal of a course in our graduate program.
- Read some books, watched some videos, started an online course.
- How can I train Data Scientists?
  - Undergraduate/graduate level.
  - 4-6 hours for short courses, 45-60 hours for the graduate program.
- Am I a Data Scientist?

# About this Course

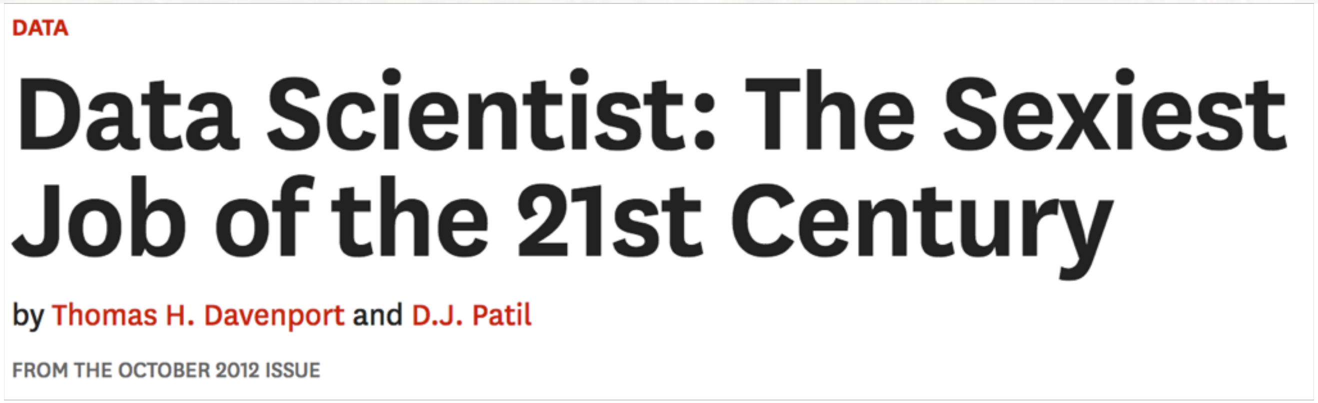
- Day 1: About Data Science
- Day 2: A bit about R
- Day 3: Tidy Data, Exploratory Data Analysis, more about R
- Day 4: A bit about Machine Learning

# About this Lecture

- What is Data Science? Why a Data Science course?
- The Data Scientist. Roles of the Data Scientist. Other related roles.
- Skills of the Data Scientist.
- A brief and incomplete list of references, videos, etc.

# What is Data Science?

# Hype



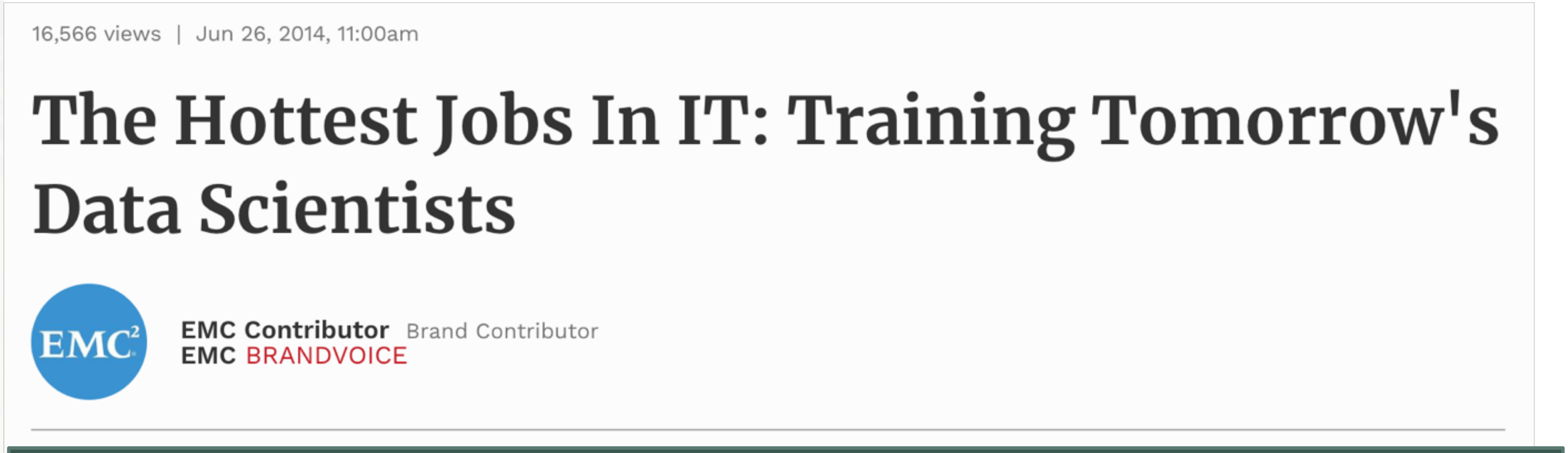
DATA

Data Scientist: The Sexiest Job of the 21st Century

by Thomas H. Davenport and D.J. Patil

FROM THE OCTOBER 2012 ISSUE

Harvard Business Review, https://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century



16,566 views | Jun 26, 2014, 11:00am

The Hottest Jobs In IT: Training Tomorrow's Data Scientists

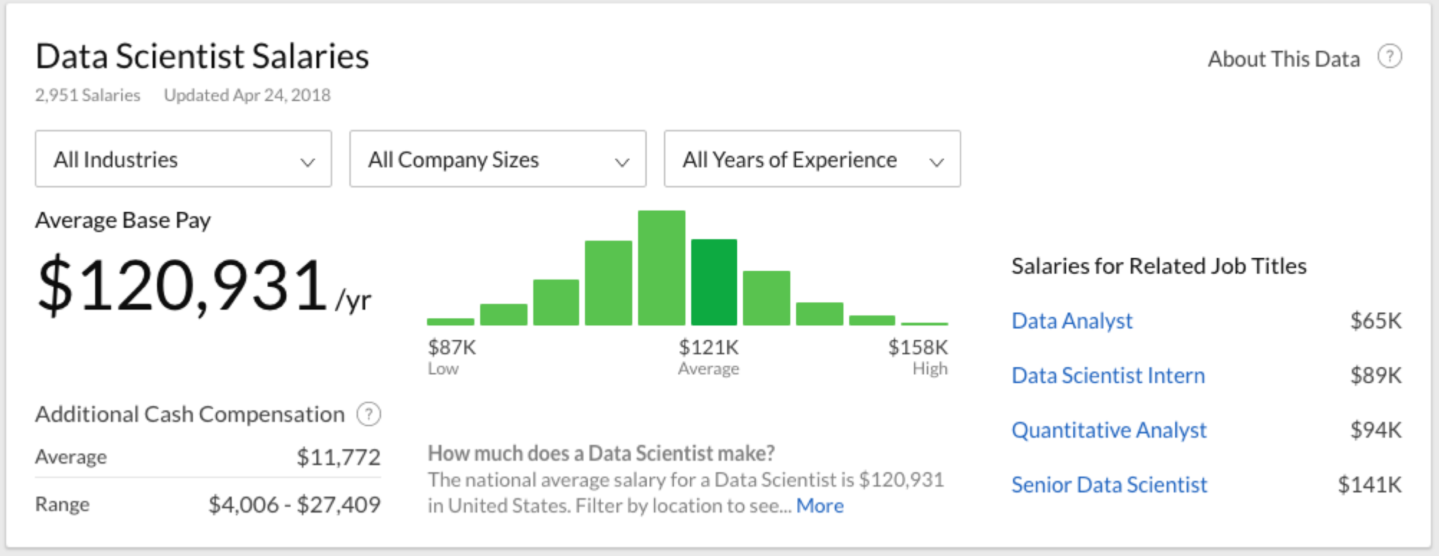
EMC Contributor Brand Contributor
EMC BRANDVOICE

Forbes, https://www.forbes.com/sites/emc/2014/06/26/the-hottest-jobs-in-it-training-tomorrows-data-scientists/

# Hype

*By 2018, the United States will experience a shortage of 190,000 skilled data scientists, and 1.5 million managers and analysts capable of reaping actionable insights from the big data deluge.*

Susan Lund et al., “Game Changers: Five Opportunities for US Growth and Renewal,” McKinsey Global Institute Report, July 2013. http://www.mckinsey.com/insights/americas/us_game_changers



Data Scientist Salaries

2,951 Salaries Updated Apr 24, 2018

About This Data

All Industries | All Company Sizes | All Years of Experience

Average Base Pay

$120,931 /yr

Additional Cash Compensation

| | |
|---|---|
| Average | $11,772 |
| Range | $4,006 - $27,409 |

How much does a Data Scientist make?

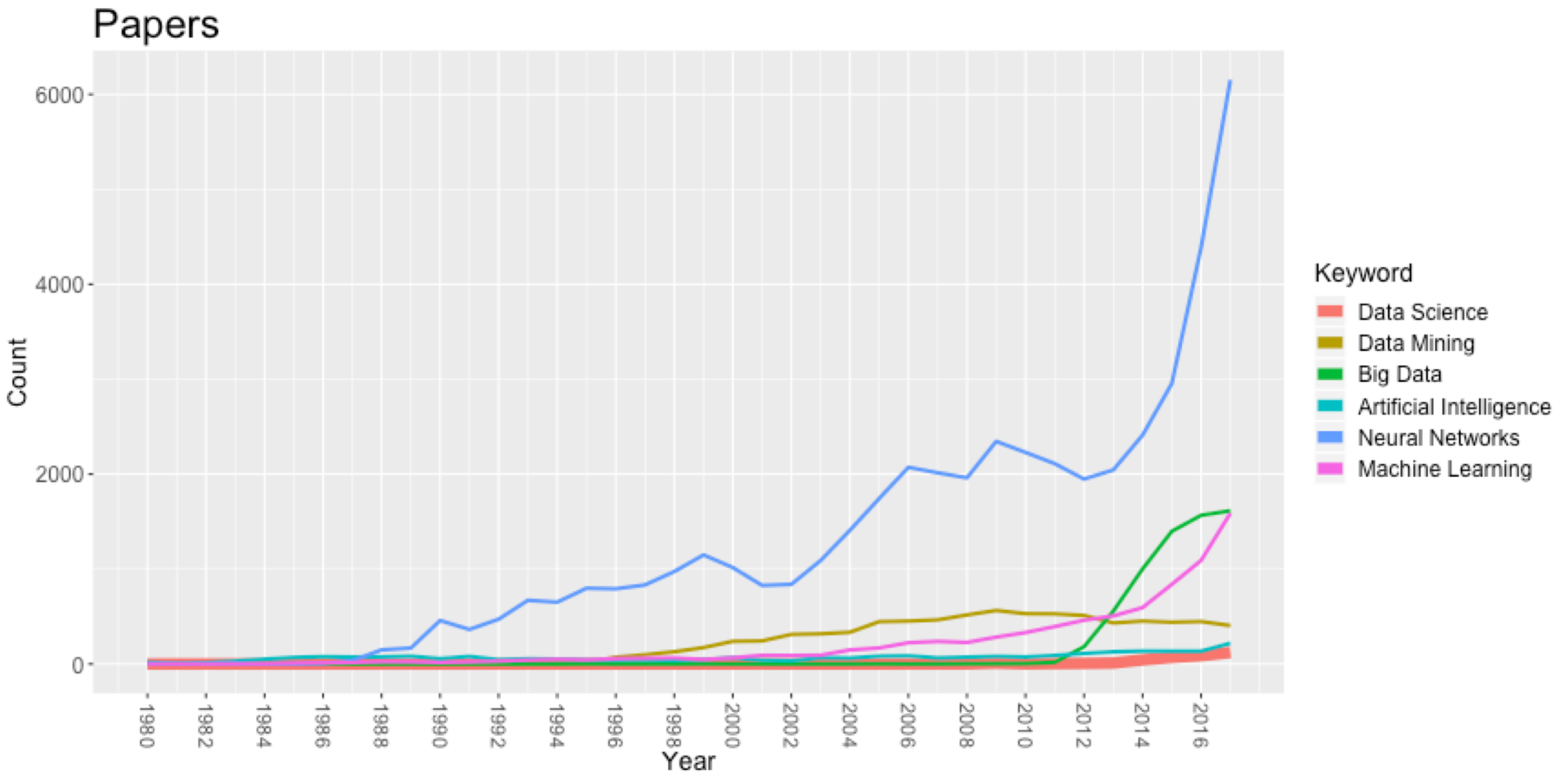
The national average salary for a Data Scientist is $120,931 in United States. Filter by location to see... More

Salaries for Related Job Titles

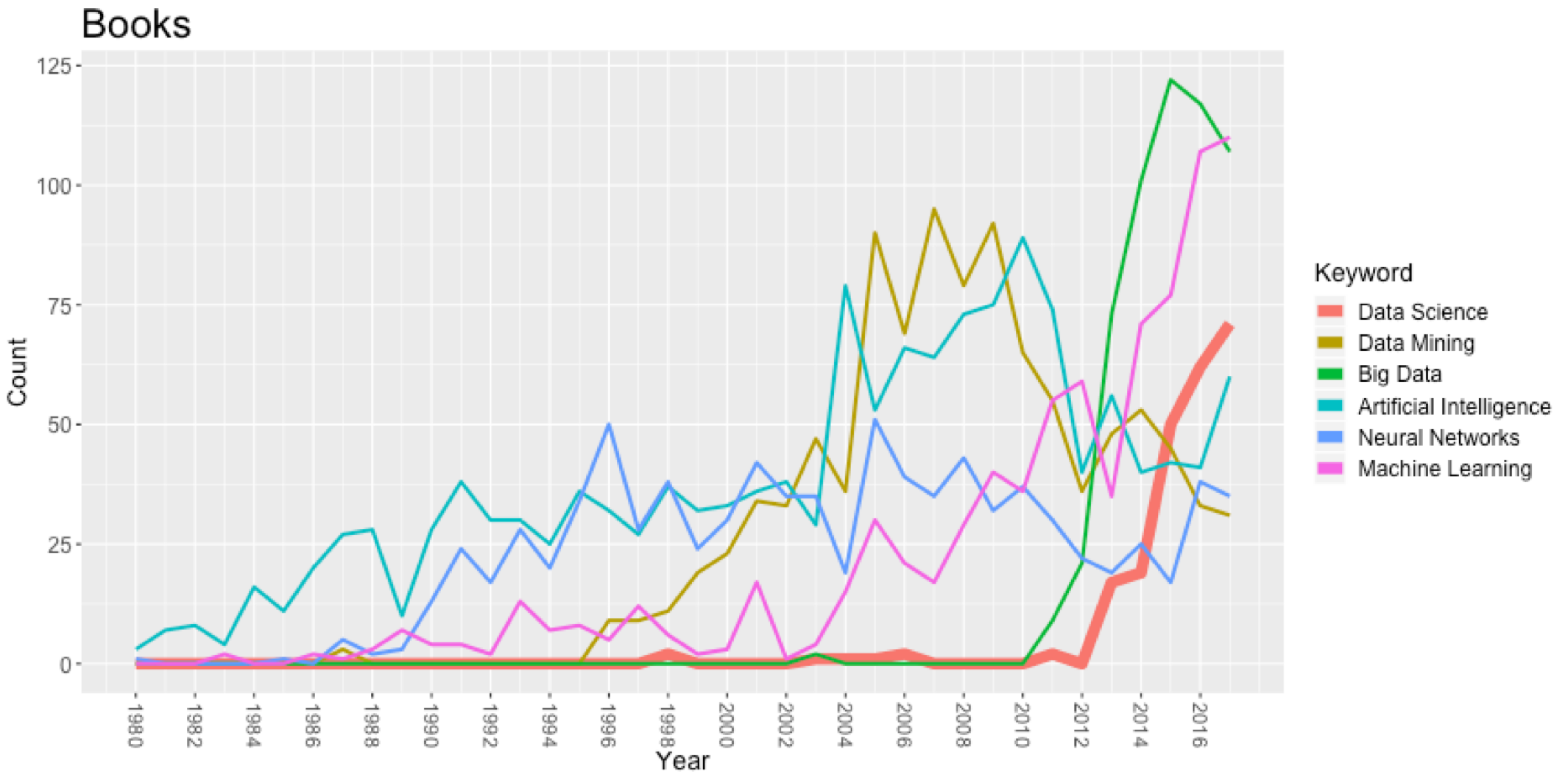
| | |
|---|---|
| Data Analyst | $65K |
| Data Scientist Intern | $89K |
| Quantitative Analyst | $94K |
| Senior Data Scientist | $141K |

glassdoor.com

# Hype


Papers
Count
6000
4000
2000
0
1980
1982
1984
1986
1988
1990
1992
1994
1996
1998
2000
2002
2004
2006
2008
2010
2012
2014
2016
Year
Keyword
Data Science
Data Mining
Big Data
Artificial Intelligence
Neural Networks
Machine Learning

# Hype



Books
125
100
75
50
25
0
Count
1980
1982
1984
1986
1988
1990
1992
1994
1996
1998
2000
2002
2004
2006
2008
2010
2012
2014
2016
Year
Keyword
Data Science
Data Mining
Big Data
Artificial Intelligence
Neural Networks
Machine Learning

# Hype

Data science and machine learning are nothing new, but several high-level trends continue to push technologies into the spotlight and generate attention and enthusiasm:

- □ Growing interest (and hype) around artificial intelligence (AI), fueled by vendor marketing combined with the understandable but erroneous conflation of AI with data science and machine learning.
- □ The data science and machine-learning talent shortage, and efforts to combat it with education, upskilling and smarter tools using more automation.
- □ Increases in computing power and availability of advanced system architectures... These advances have also fueled the hype and interest around deep learning.
- □ The explosion in popularity of open-source tools and libraries for data science and machine learning. The data science and machine-learning market is one of the most vibrant and collaborative technology market that strongly embraces open-source technologies.

Gartner, “Hype Cycle for Data Science and Machine Learning, 2017”, https://www.gartner.com/doc/reprints?id=1-4MLA3QU&ct=171220&st=sb

# What is a data scientist? 14 definitions of a data scientist!

- □ “A data analyst who lives in California”
- □ …almost everyone who works with data in an organization…
- □ …a rare hybrid, a computer scientist with the programming abilities to build software to scrape, combine, and manage data from a variety of sources and a statistician who knows how to derive insights from the information within…
- □ …someone who can obtain, scrub, explore, model and interpret data, blending hacking, statistics and machine learning.

http://bigdata-madesimple.com/what-is-a-data-scientist-14-definitions-of-a-data-scientist/

# It is about Data, but...

... *not only about data collection and management!*

- Data-based applications are common: look around!
- Use (collect, store, publish) data is not Data Science. One must add value to the data and allow new ways to use it.
  - Simple example of adding value to data: CDDB database.
  - More complex example(s): Citizen Science projects.
- Data Science allows the creation of data products.

https://www.oreilly.com/ideas/what-is-data-science

# Requires Coding, but...

*... not only programming and engineering: don't jump on the latest bandwagon!*

□ Coding for Solutions: Less emphasis on size, tools and technology, more in application of technologies to obtain answers from the data.

□ Hacking is essential (inevitable?) for Data Scientists.

http://simplystatistics.org/2013/12/12/the-key-word-in-data-science-is-not-data-it-is-science/

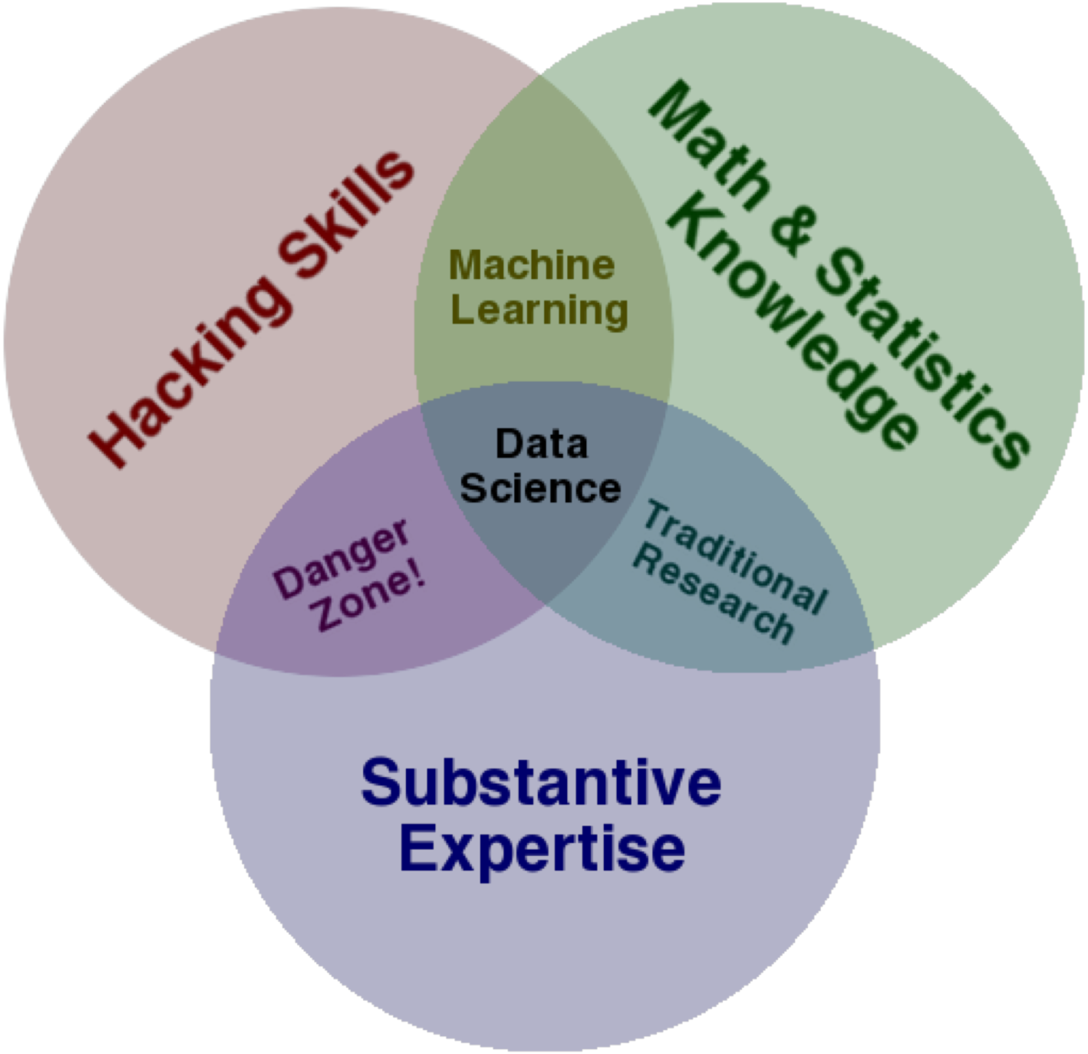
# Requires Statistics, but...

*...not only plain traditional statistics.*

- Traditional methods are used, but new ones can be aggregated.
- Data is almost never ready to process…
  - Point-and-click applications may not be enough.
  - Must understand nature and sources of the data.
  - Prototyping/Hacking (R, Python) is very useful.
- Knowledge about databases, combining/scrubbing/munging data, EDA, visualization, etc. is required.

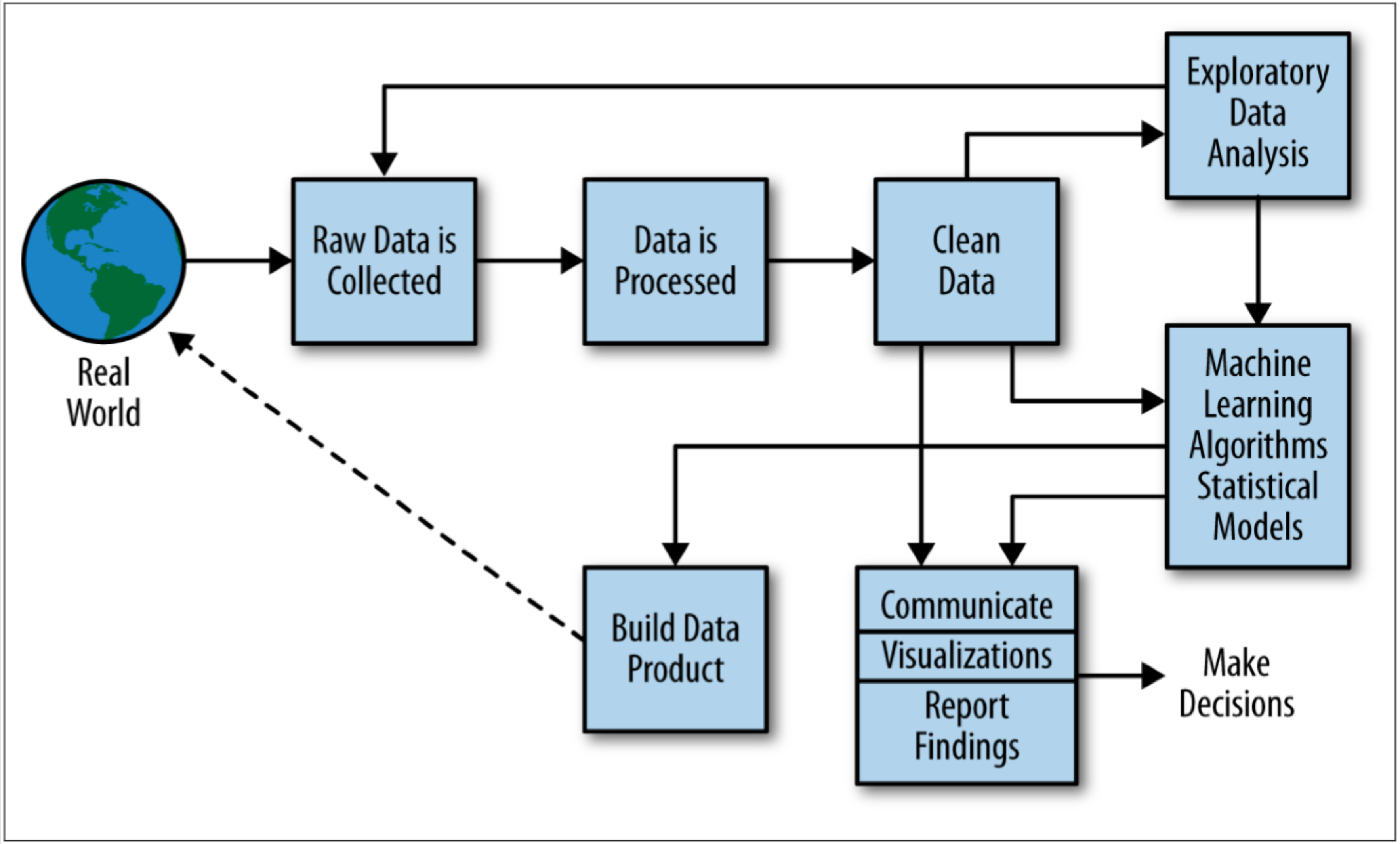
http://magazine.amstat.org/blog/2013/07/01/datascience/

# All this and much more!



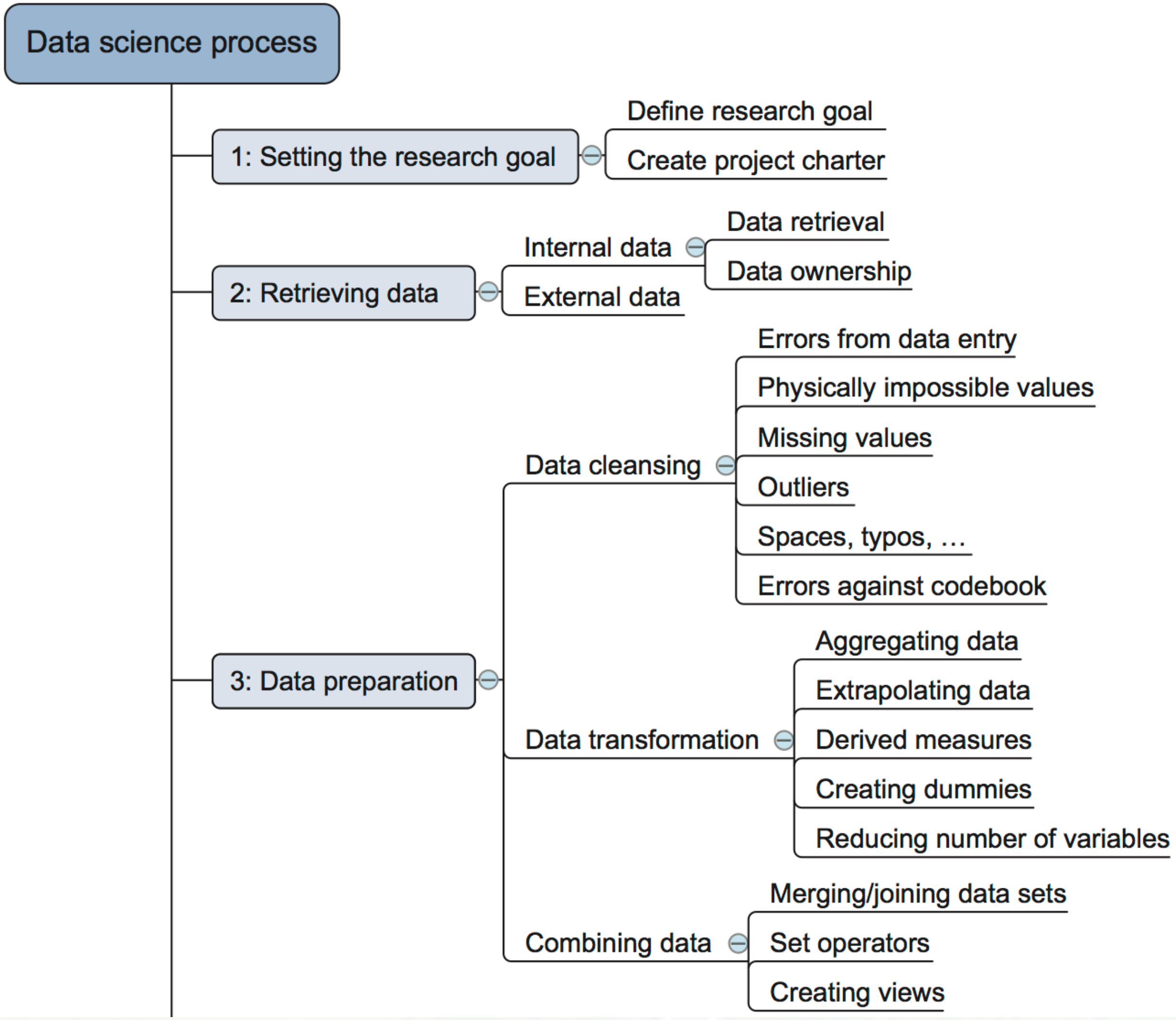
http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram

# It is a Process



Doing Data Science, Rachel Schutt and Cathy O'Neil, OReilly, 2014

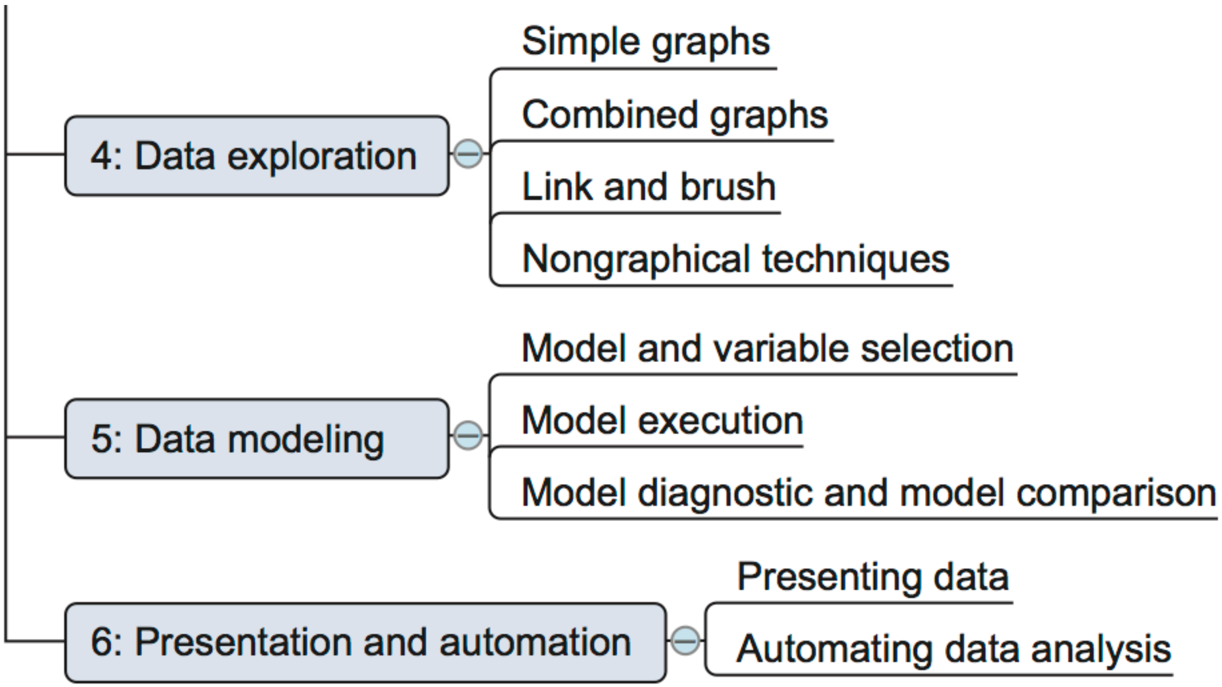
# It is a Process (2)



- Data science process
  - 1: Setting the research goal
    - Define research goal
    - Create project charter
  - 2: Retrieving data
    - Internal data
      - Data retrieval
      - Data ownership
    - External data
  - 3: Data preparation
    - Data cleansing
      - Errors from data entry
      - Physically impossible values
      - Missing values
      - Outliers
      - Spaces, typos, …
      - Errors against codebook
    - Data transformation
      - Aggregating data
      - Extrapolating data
      - Derived measures
      - Creating dummies
      - Reducing number of variables
    - Combining data
      - Merging/joining data sets
      - Set operators
      - Creating views

Introducing Data Science, Davy Cielen, Arno Meysman, Mohamed Ali, Manning, 2016

# It is a **Process** (2)



Introducing Data Science, Davy Cielen, Arno Meysman, Mohamed Ali, Manning, 2016

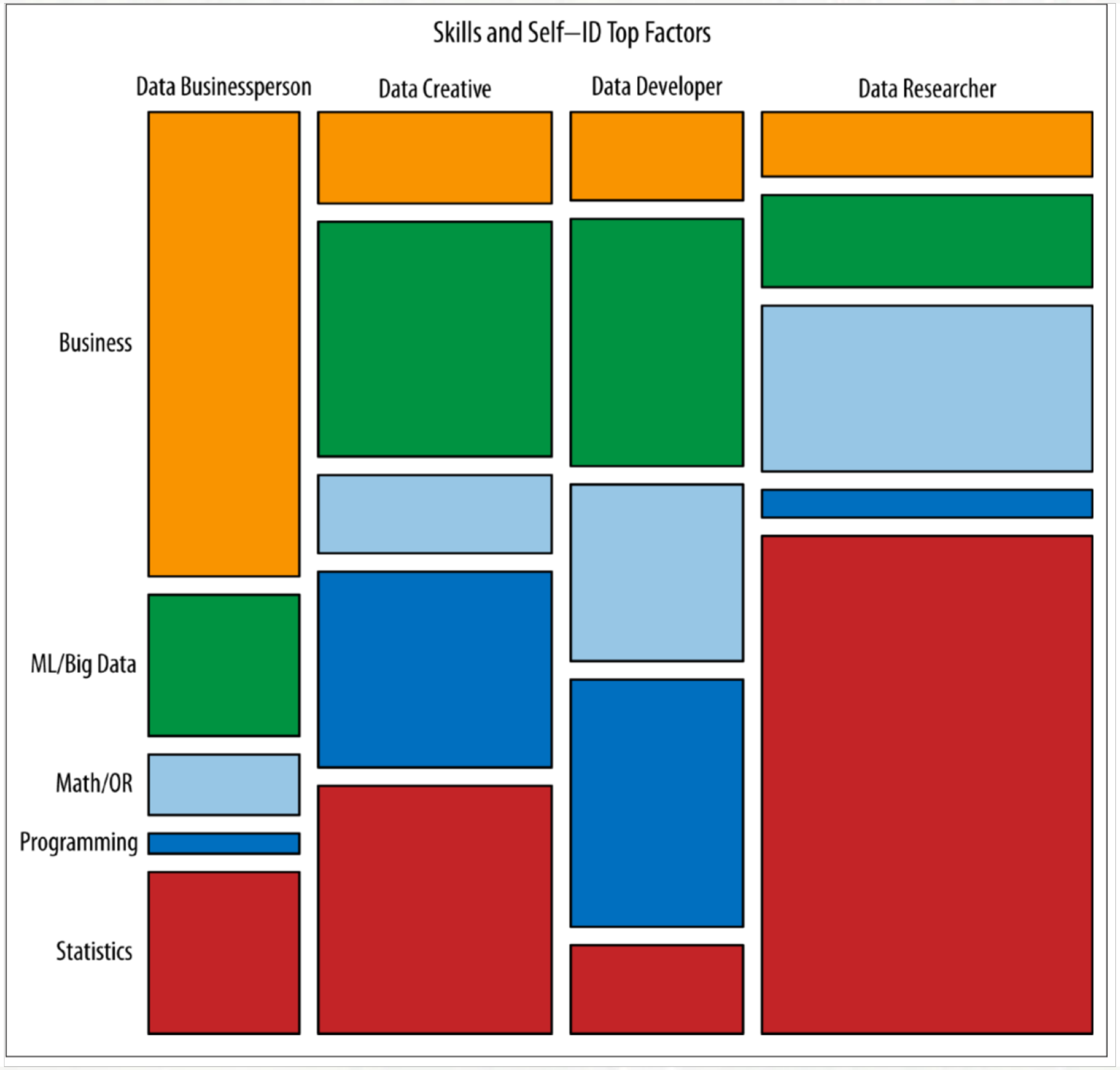
# So you want to be a Data Scientist…

# Who are the Data Scientists?

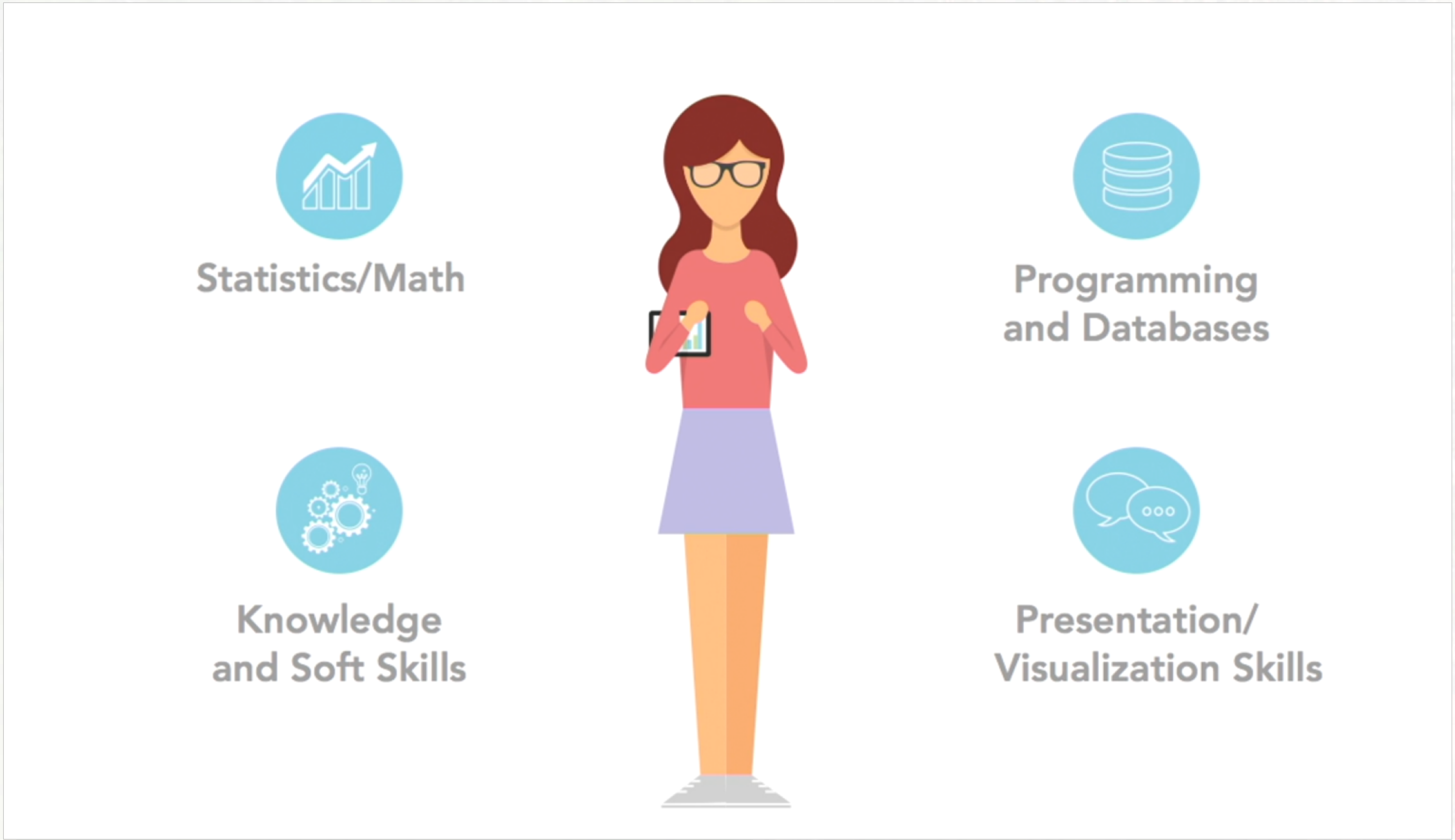
- *Analyzing the Analyzers*:
  - Someone who knows statistics, coding and visualization?
  - Someone with experience on how to extract information from data?
  - We need a more specific description ("doctor", "athlete", "data scientist" are too generic!)
  - Definition depends on the problem.
- Interviews with 250 volunteers.

Harlan Harris, Sean Murphy, and Marck Vaisman. Analyzing the Analyzers: An Introspective Survey of Data Scientists and Their Work. O'Reilly Media, Inc., 2013.

# Who are the Data Scientists?



# Who are the Data Scientists?


Statistics/Math
Programming and Databases
Knowledge and Soft Skills
Presentation/ Visualization Skills

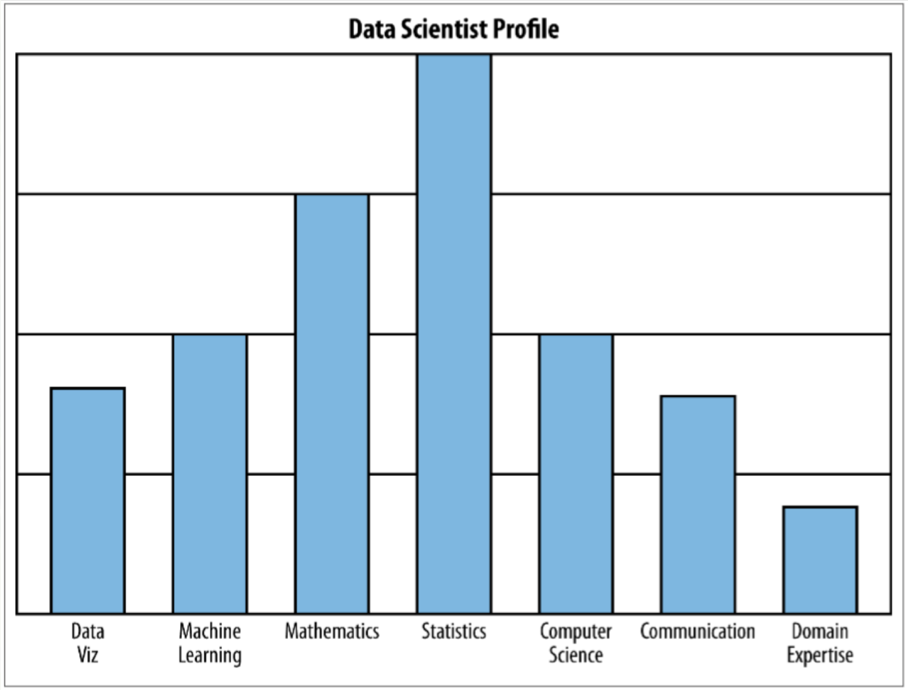
Lynda.com - Twelve Myths About Data Science

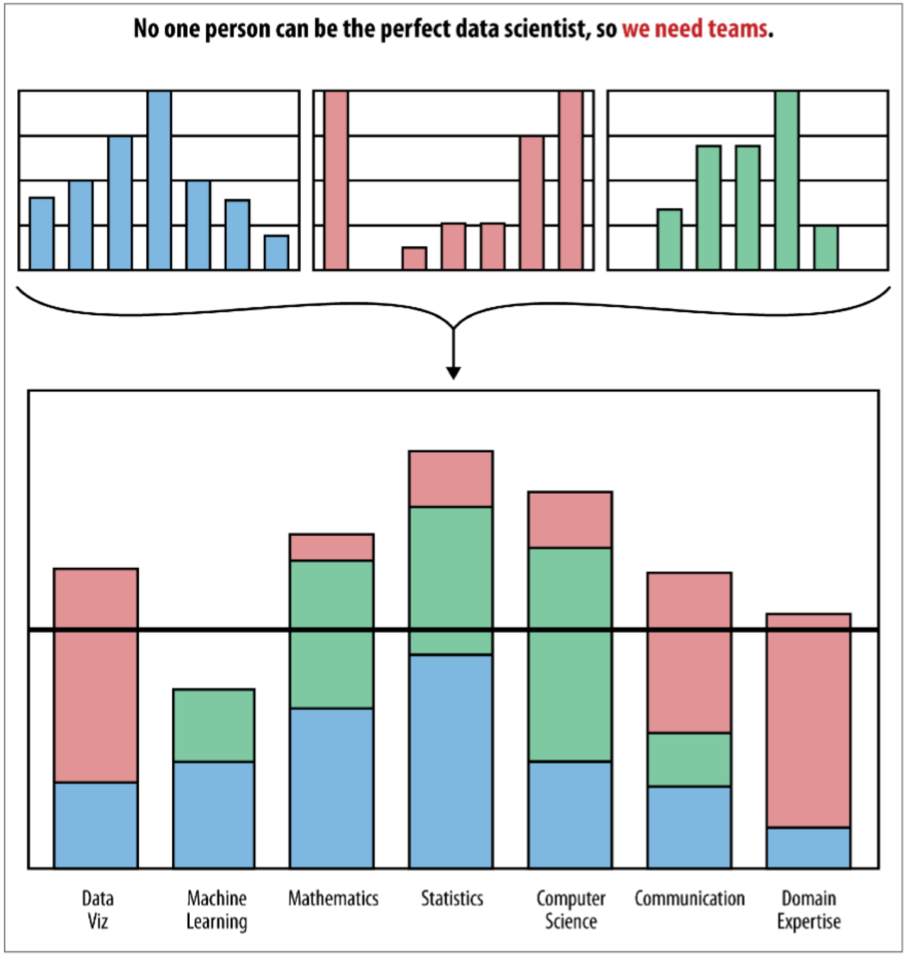
# Who are the Data Scientists?

- *Analyzing the Analyzers*: evidence of the *T-Shaped Data Scientist*
- Wide knowledge about the whole process, deep knowledge in a single aspect.
  - Better for task-oriented, interdisciplinary teams.
  - More efficient in their expertise area.
- Other study indicates three categories:
  - Data Curation.
  - *Analytics* and visualization.
  - Networks and infrastructure.

Jeffrey Stanton et al, Interdisciplinary Data Science Education, http://pubs.acs.org/doi/abs/10.1021/bk-2012-1110.ch006

# T-Shaped Data Scientist





Doing Data Science, Rachel Schutt and Cathy O'Neil, OReilly, 2014

# What is Data Science?

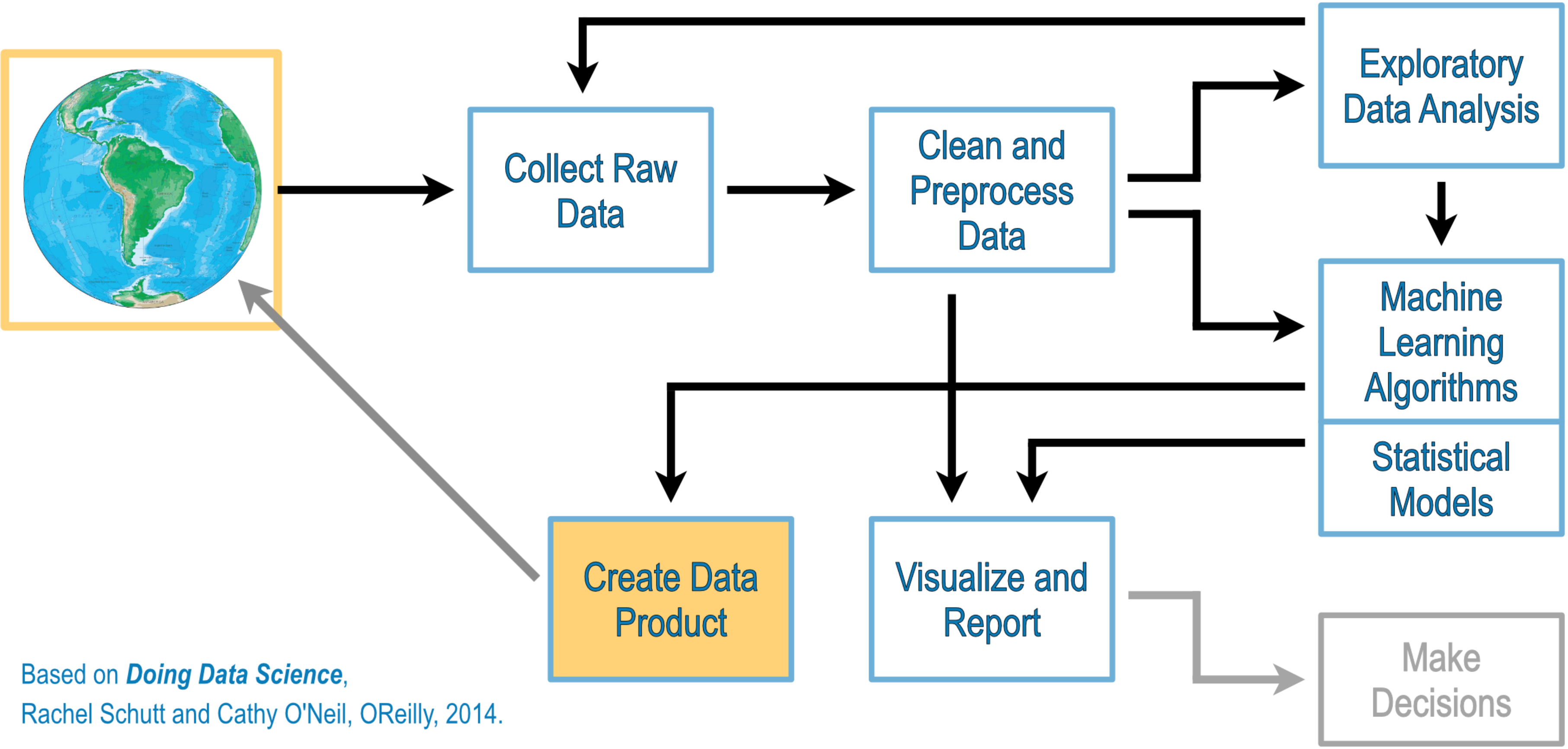
- “Data Scientist” was defined before “Data Science” was well-established.
- You’re a Data Scientist if you work with data in a scientific way.
- You’re a *good* Data Scientist if…
  - You know the data (or knows who know!)
  - You know the tools (or knows who know!)
  - You can ~~listen to~~ extract information from people.
  - Can imagine solutions and data products!

# Skills

# Skills

- List of interesting things to learn that is…
  - ...incomplete: new concepts, technologies, languages, appear all the time.
  - ...biased: everyone has some preferences. Keep a healthy, suspicious mind. Watch out for hype!
  - ...possibly redundant: a data scientist must learn how to play in different positions on several team.
  - ...individually impossible: *“Rockstar Programmer”*, *“Rockstar SysAdmin”*, *“Rockstar Analyst”?*
  - ...not all technical: we will deal with real world problems, must talk to real world people.

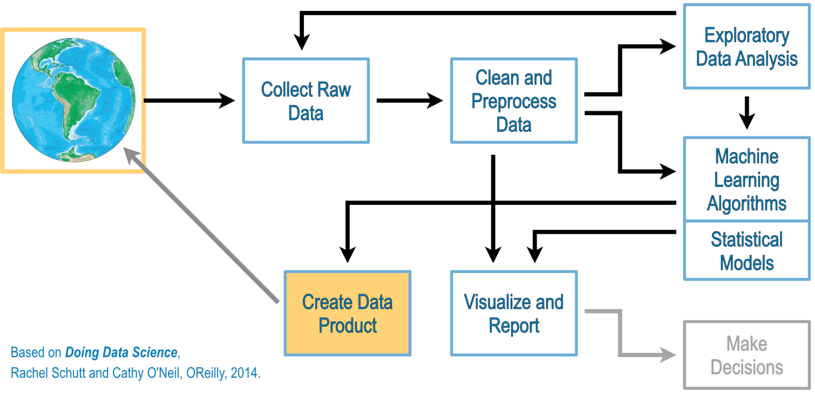
# Skill: Understand the Problem



Based on ***Doing Data Science***,
Rachel Schutt and Cathy O'Neil, OReilly, 2014.

# Skill: Understand the Problem

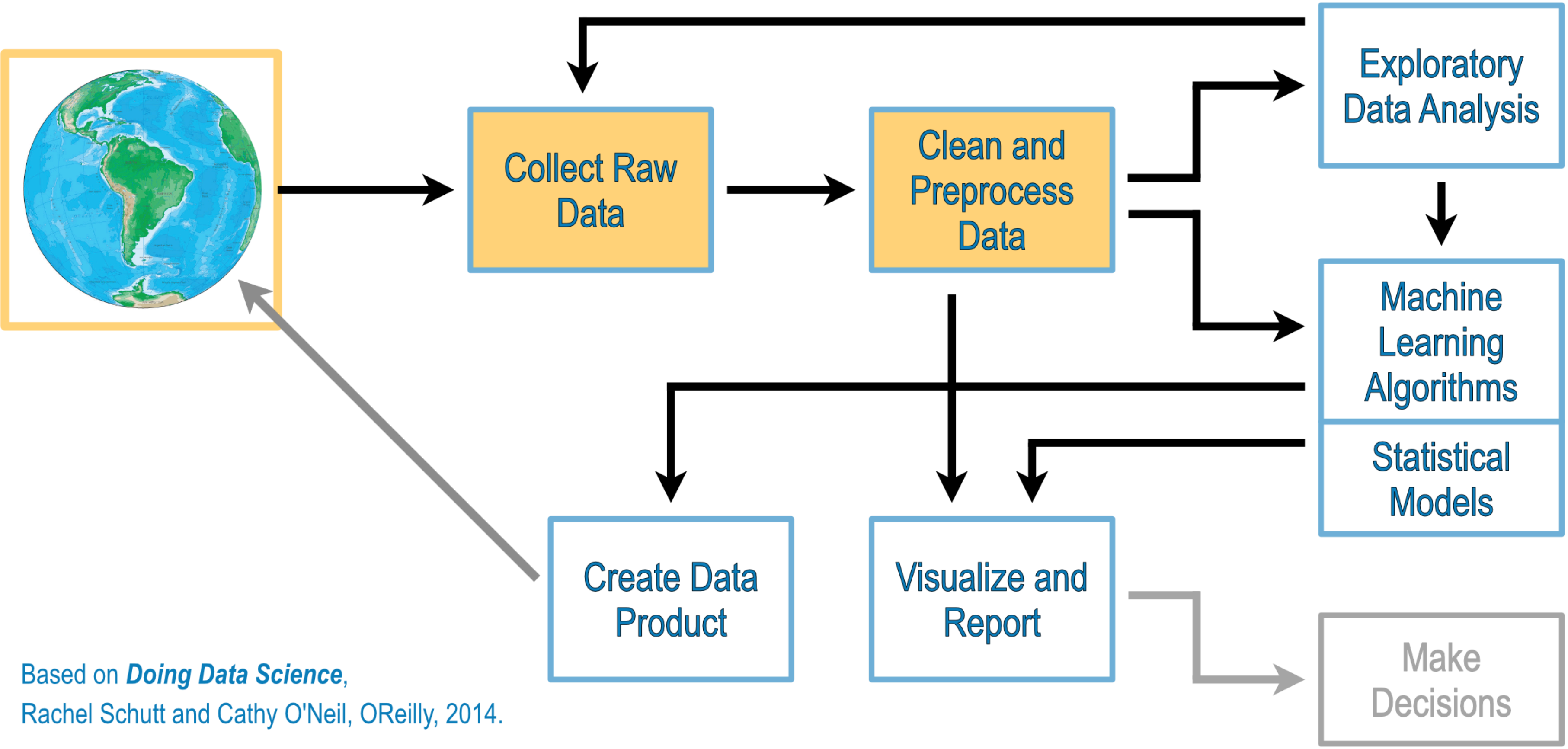
- At least enough to communicate with people that understand the problem!
  - Data Science is inherently interdisciplinary!



Based on ***Doing Data Science***, Rachel Schutt and Cathy O'Neil, OReilly, 2014.

- What data is available?
  - What data *should be* available?
  - Think about a nice Data Product to answer this!
- Don’t start without understanding the problem at hand!

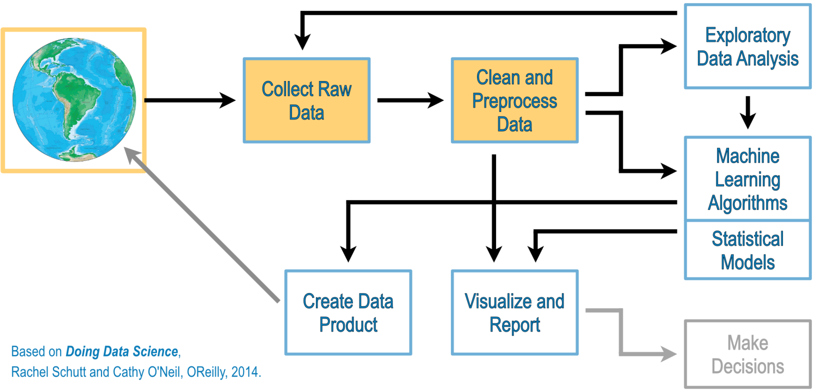
# Skill: Find, Collect, Organize Data



Based on ***Doing Data Science***,
Rachel Schutt and Cathy O'Neil, OReilly, 2014.

# Skill: Find, Collect, Organize Data

- What data is available?
- Can we get it?
  - *How* can we get it?
  - Do we need more data?
  - Is the data ready to use?
  - Do we need to copy/replicate/sample it?
  - What is the data volume? Does it matter for the collection/analysis?



Based on ***Doing Data Science***, Rachel Schutt and Cathy O'Neil, OReilly, 2014.

# Detour: Big Data

- What is Big Data?
- Traditional definition: any dataset too large for…
  - ...simple analysis?
  - ...effective/efficient processing?
  - ...complete storage?
- Measures in {Gb,Tb,Pb} may reflect the size of the data (and other interesting aspects of its collection) but may not be related with the problem at hand.

Big Data Lessons from the Climate Science Community, Seth McGinnis, 2016

# Big Data: 3 Vs

- **Volume**: how much storage is required.
  - Driven by storage capacity (and new sensors!)
  - Dealt with by technology (processing capacity).
- **Velocity**: how quickly data must be processed.
  - Real-time: must be acted on immediately?
  - Timeliness: rate of capture/usage.
  - Lifespan: for how long is it valuable?
- **Variety**: how heterogeneous (complex) is the data.
  - Format, features, meaning, structure, representation, location, etc.
  - Dealt with standards, specifications, ontologies.

Big Data Lessons from the Climate Science Community, Seth McGinnis, 2016

# Big Data: More than 3 Vs

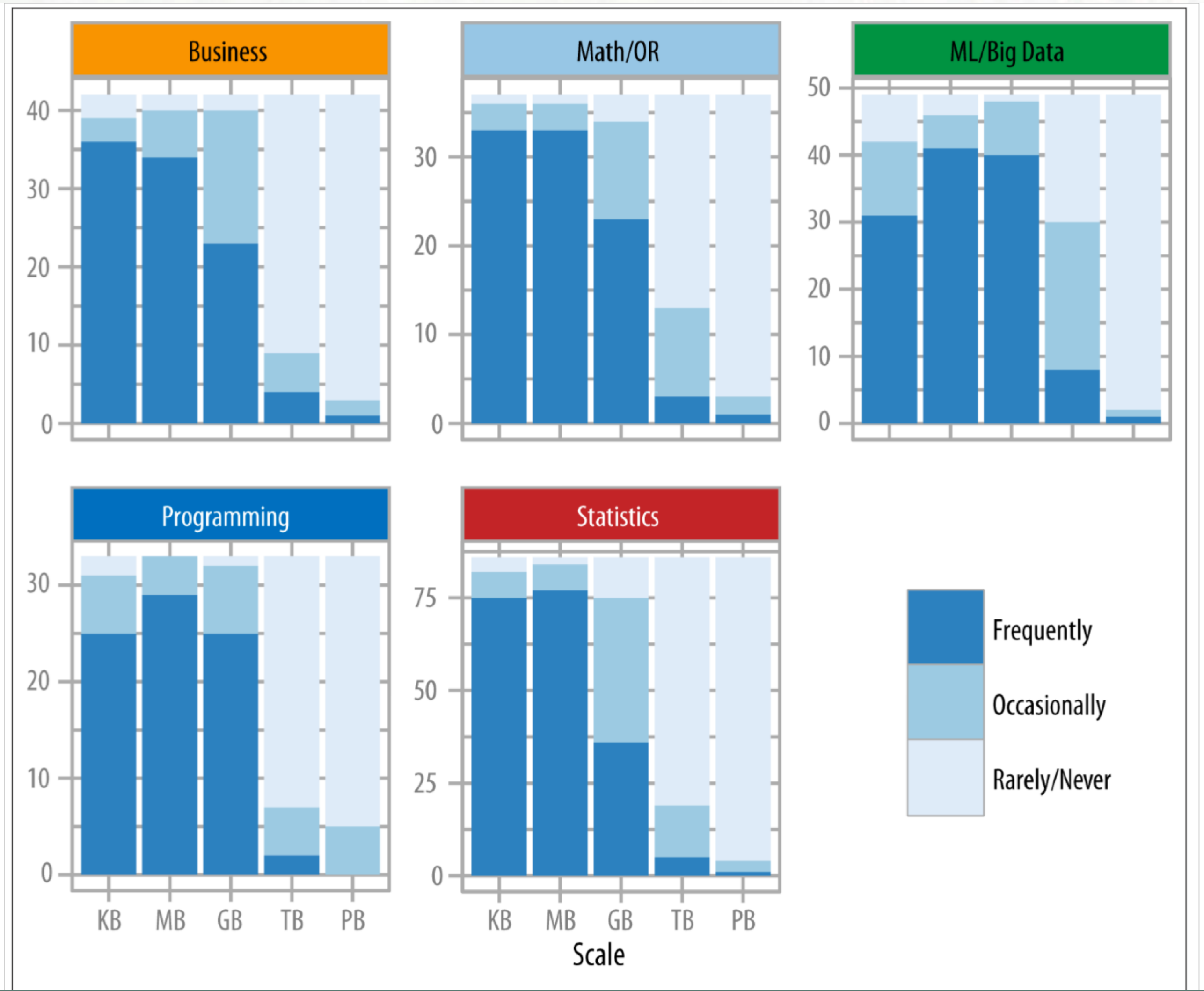
- **Value**: if we're collecting and storing data it is because it has value (?)
- **Veracity**: are the data trustworthy?
  - Consider provenance, reliability, accuracy, completeness, etc.
- **Validity**: accuracy and correctness relative to use.
  - E.g. Polls, tracking weather phenomena by sensors or tweets.
- **Variability**: change of meaning of data in time.
- **Viscosity, Volatility, Venue, …**

Big Data Lessons from the Climate Science Community, Seth McGinnis, 2016

# Big Data: Myths

- Do we really, *really* need it? Issues caused by:
  - Sheer size.
  - Underlying platforms.
  - Lack of organization (data dumps).
  - IT requirements (including human resources).
- *(Real) Big Data is a problem you'd be lucky to not have to worry about*.
- Of course it depends on your project…

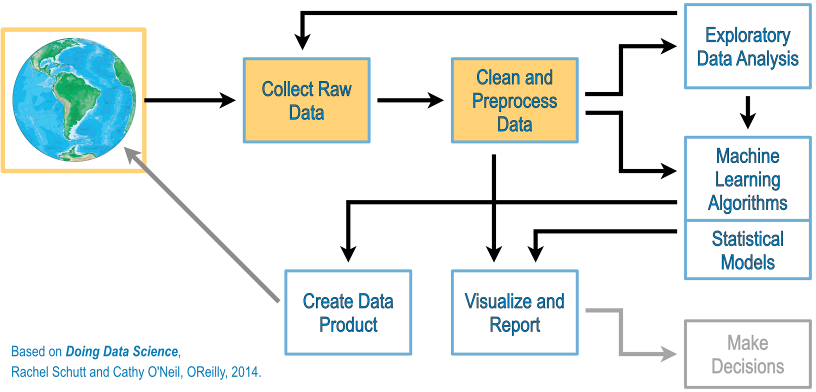
# Back to Skills: Big Data



Harlan Harris, Sean Murphy, and Marck Vaisman. Analyzing the Analyzers: An Introspective Survey of Data Scientists and Their Work. O'Reilly Media, Inc., 2013.

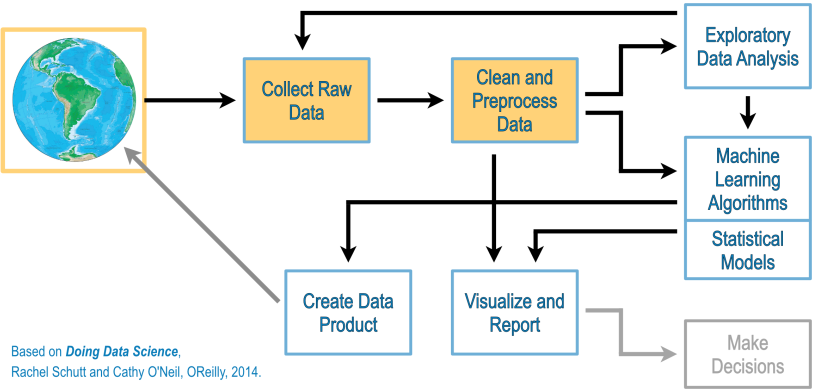
# Skill: Understand and Organize Data

- Before Processing: How data is organized?
  - Tables, documents, images, relations, graphs, mixed?
  - Is data in a format useful for processing?
    - How can we transform it?
    - How hard it is to transform it?



Based on *Doing Data Science*, Rachel Schutt and Cathy O'Neil, OReilly, 2014.

# Skill: Understand and Organize Data

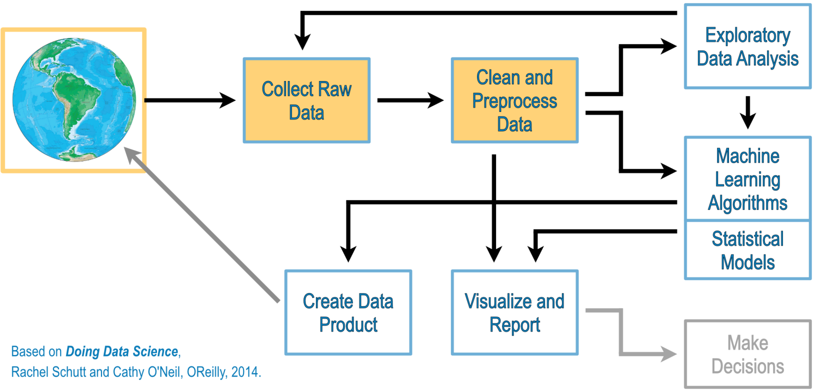


Based on *Doing Data Science*, Rachel Schutt and Cathy O'Neil, OReilly, 2014.

- Do we need this data in a specific format/location/organization?
  - Where is the data?
  - Are we going to collect it once or more?
  - Do we need provenance, metadata?
  - What does need to be stored, augmented, preprocessed?
  - Does this data (for analysis) has a different life than the original data?
    - Did we add value to the data?

# Skill: Understand and Organize Data

- If we need them in a particular specific format/location/organization, how should we do it?
  - Collections of {documents, images, files, tables}?
  - What storage and/or processing technologies are needed?



Based on *Doing Data Science*, Rachel Schutt and Cathy O'Neil, OReilly, 2014.

# What technologies are needed?

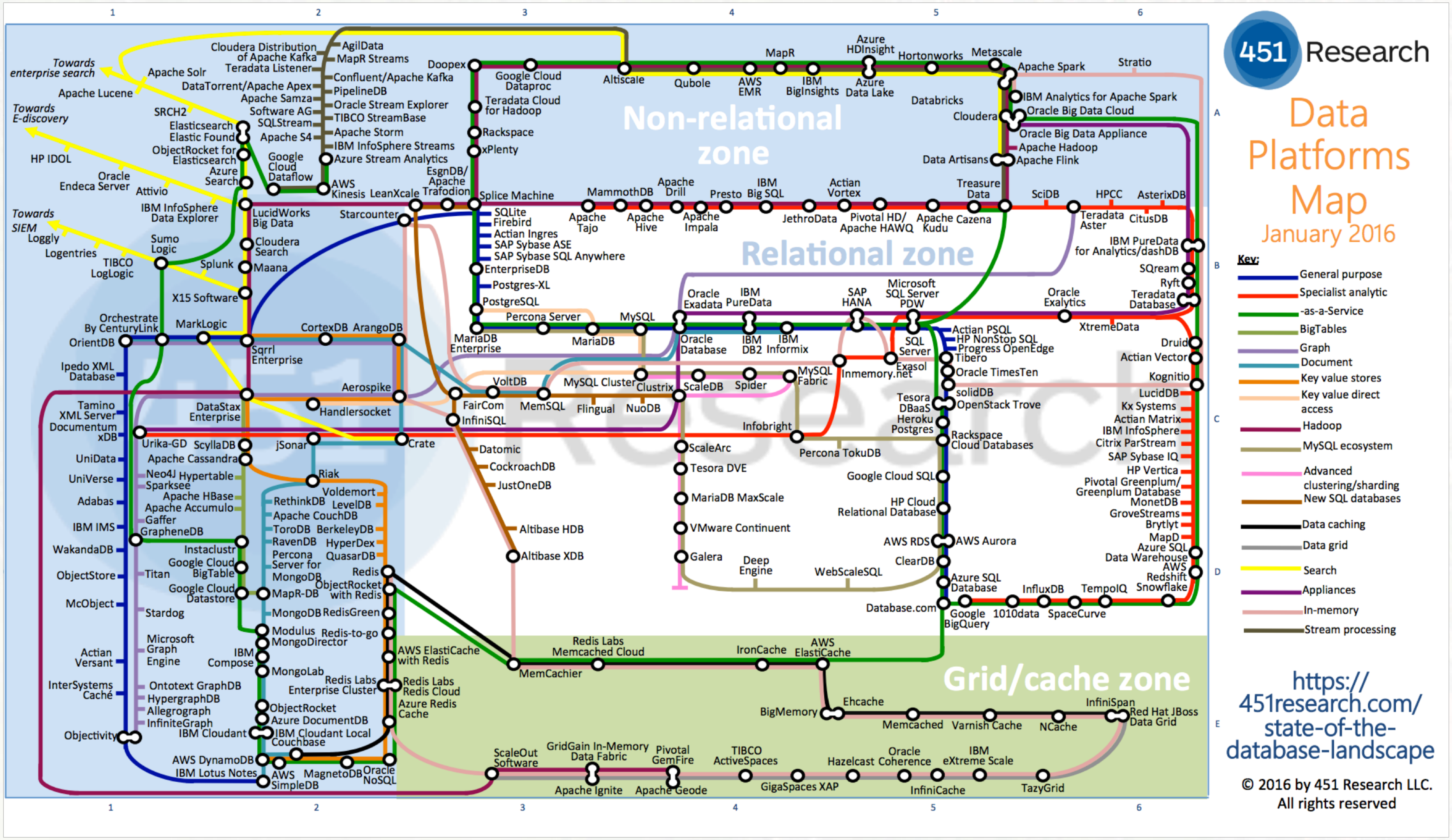
- Too many options with different capabilities and limitations…
- We're still talking about skills!
- Learn SQL: excelent for well-structured data (tables).
  - More complex data may lead to more complex tables...
- Learn some NoSQL DBMSs:
  - More flexible for differently structured data.
  - Several different models and flavors...

# NoSQL

- Key/Value: associative arrays, maps, dictionaries.
  - Redis, Riak, Memcached, etc.
- Column Based: expand Key/Value to several columns.
  - Cassandra, HBase
- Document Based: hierarchies of keys and values
  - Couchbase, CouchDB, MongoDB
- Graph Based: nodes and edges
  - Neo4J, OrientDB

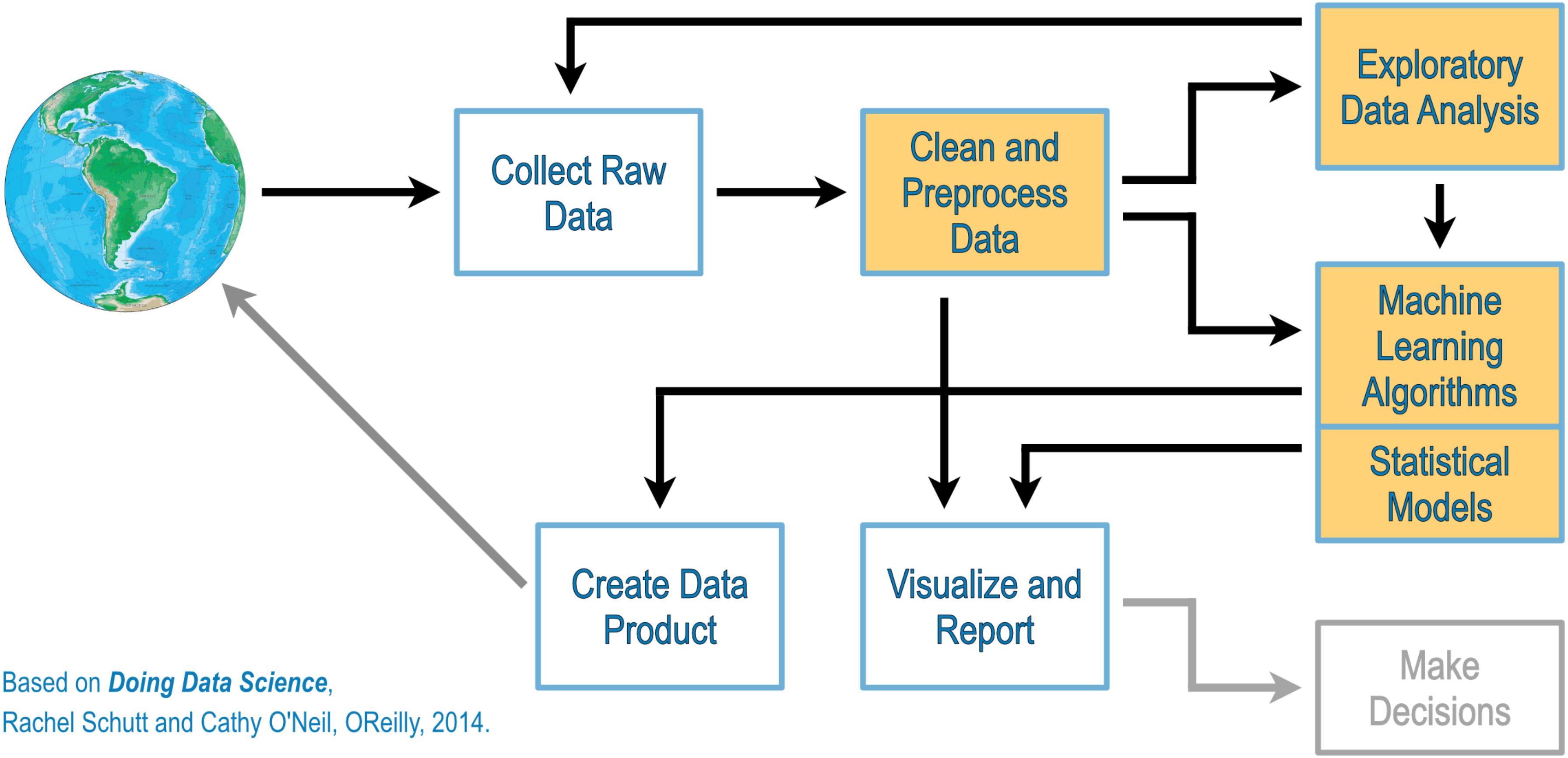
https://www.digitalocean.com/community/tutorials/a-comparison-of-nosql-database-management-systems-and-models

# NoSQL



© 2016 by 451 Research LLC.
All rights reserved

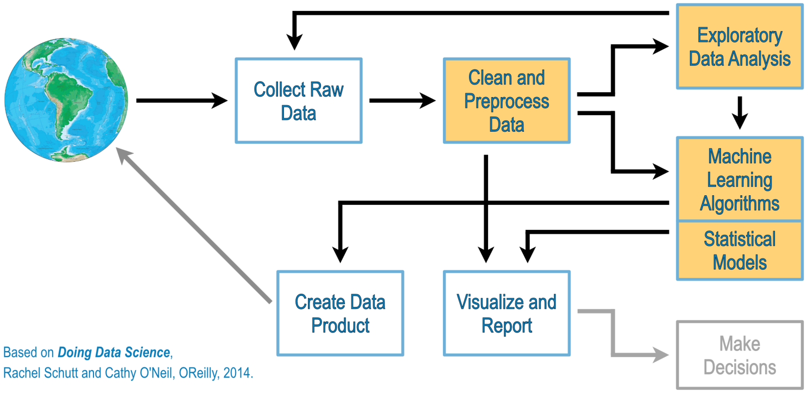
# Skill: Analysis, Hacking



Based on ***Doing Data Science***,
Rachel Schutt and Cathy O'Neil, OReilly, 2014.

# Skill: Analysis

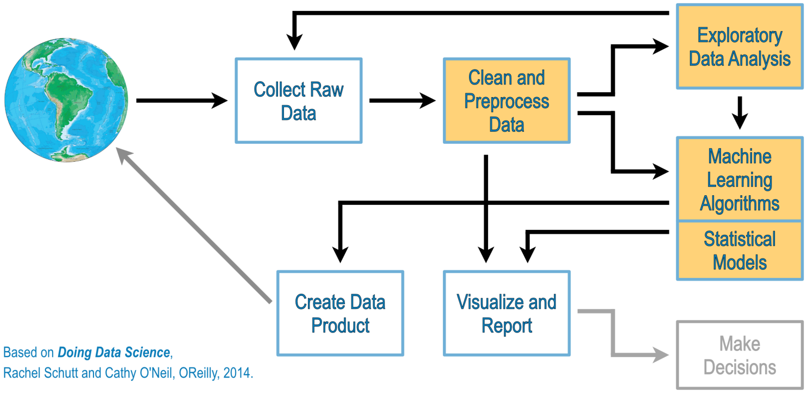
- We have the data. Now what?
  - Do we know what we want to discover?
  - We need basic skills in statistics and data modeling.



Based on ***Doing Data Science***, Rachel Schutt and Cathy O'Neil, OReilly, 2014.

- Start exploring: Exploratory Data Analysis
  - Make different plots and charts to explore variables.
  - Get some basic statistics.
  - Evaluate the type of information we can extract from the data.

# Skill: Hacking



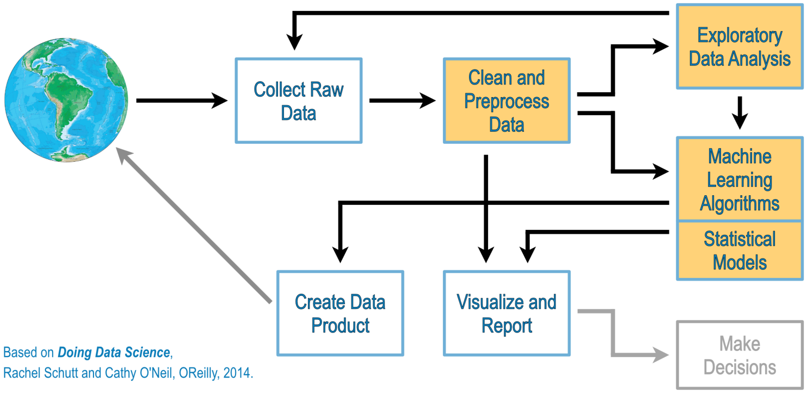
Based on *Doing Data Science*, Rachel Schutt and Cathy O'Neil, OReilly, 2014.

- □ Definition of **hacker**
  1. one that hacks
  2. a person who is inexperienced or unskilled at a particular activity – *a tennis hacker*
  3. an expert at programming and solving problems with a computer
  4. a person who illegally gains access to and sometimes tampers with information in a computer system
- □ More than expertise in Excel, not as much expertise as full applications development.

# Skill: Hacking

- Do we need to?

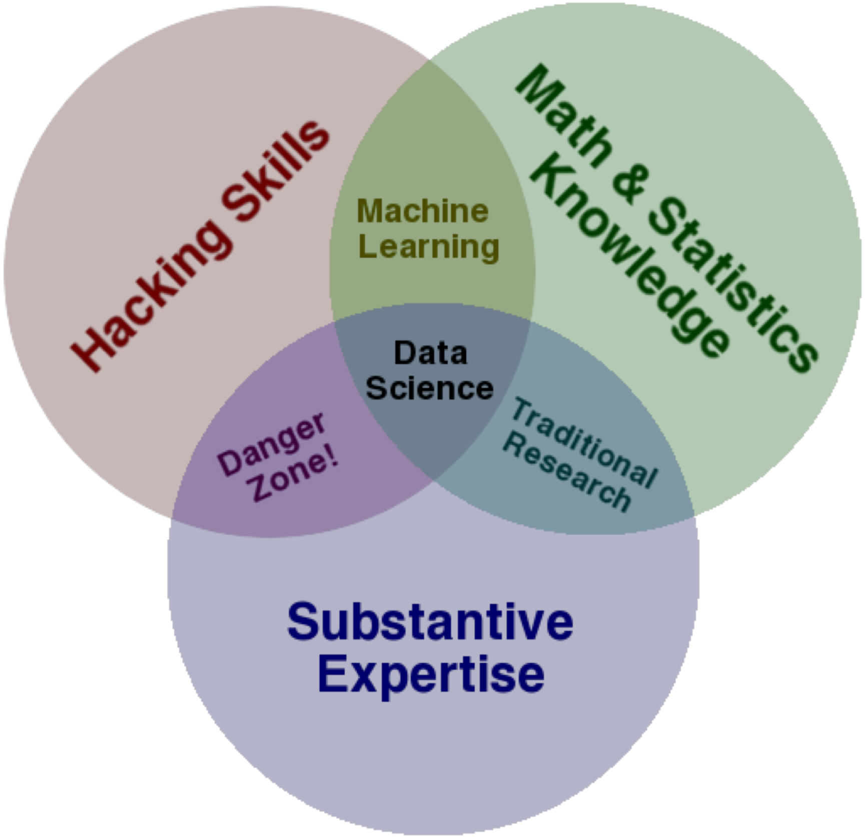
- **YES.**
  - Coding may be needed **even before** getting the data.
  - Data processing using code is (long term) much easier to do than via menu/dialog interfaces.
  - Automation of tasks.
  - Reproducibility of the same task with different datasets.
  - Writing code that writes (simpler) code!

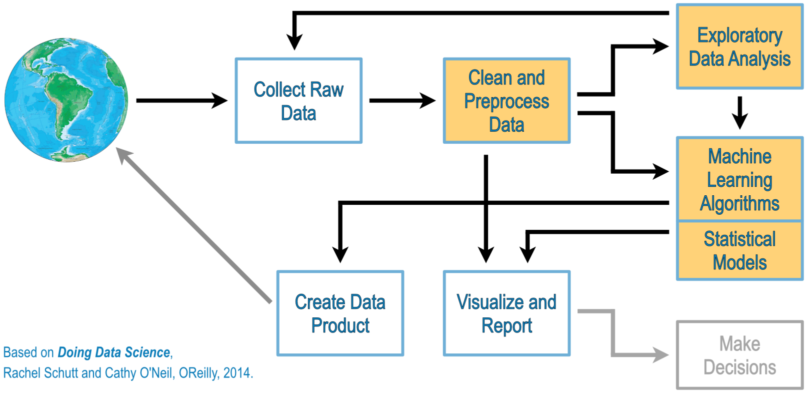


Based on ***Doing Data Science***, Rachel Schutt and Cathy O'Neil, OReilly, 2014.

# Skill: Hacking

- □ Important reminder!

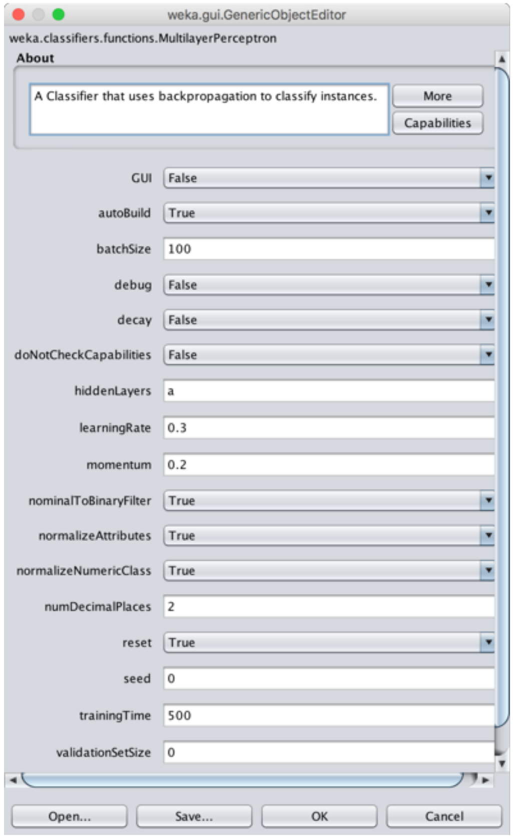




Based on *Doing Data Science*, Rachel Schutt and Cathy O'Neil, OReilly, 2014.

# Skill: Hacking

- □ This:



weka.gui.GenericObjectEditor

weka.classifiers.functions.MultilayerPerceptron

About

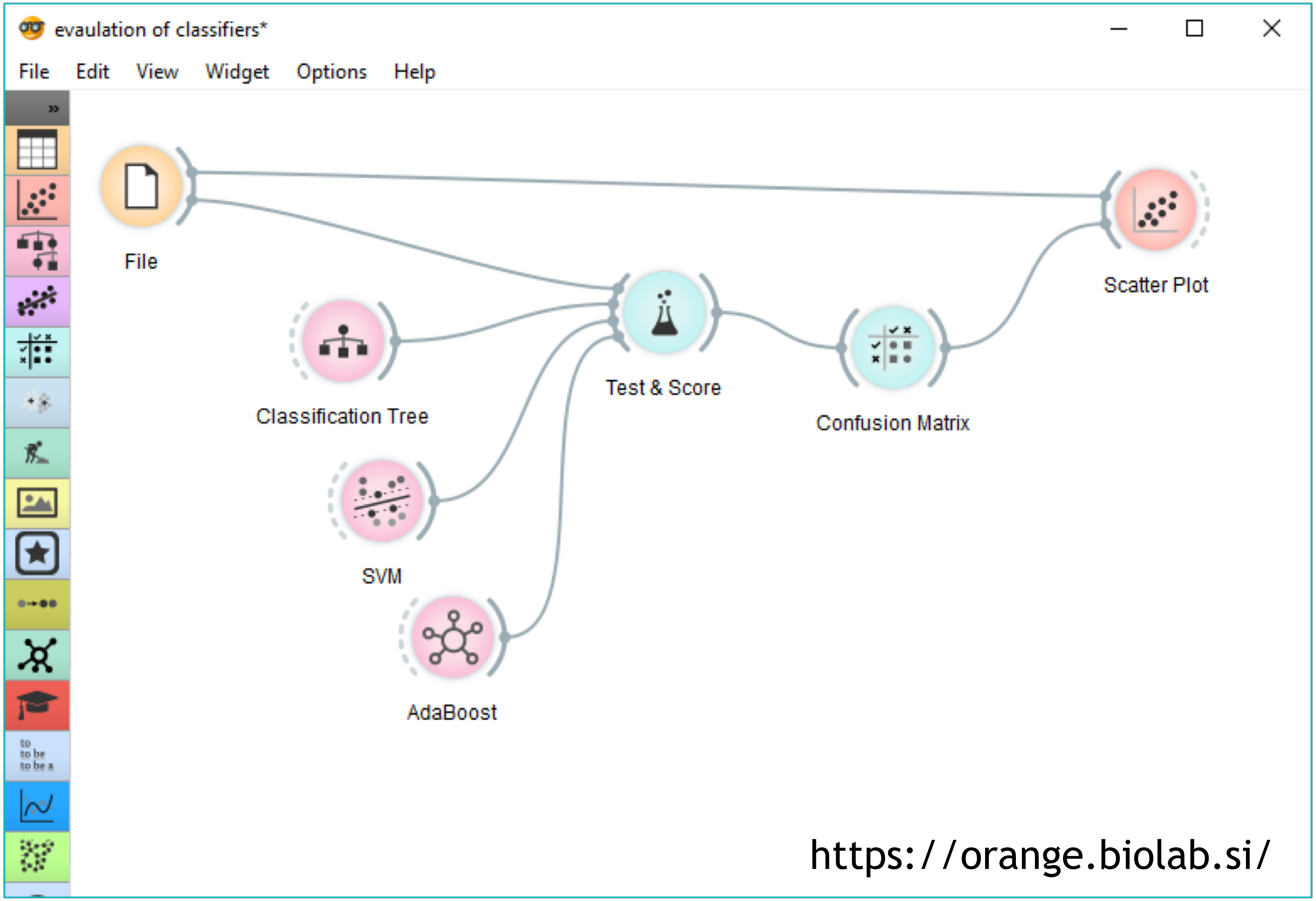
A Classifier that uses backpropagation to classify instances.

| | |
|---|---|
| GUI | False |
| autoBuild | True |
| batchSize | 100 |
| debug | False |
| decay | False |
| doNotCheckCapabilities | False |
| hiddenLayers | a |
| learningRate | 0.3 |
| momentum | 0.2 |
| nominalToBinaryFilter | True |
| normalizeAttributes | True |
| normalizeNumericClass | True |
| numDecimalPlaces | 2 |
| reset | True |
| seed | 0 |
| trainingTime | 500 |
| validationSetSize | 0 |

Parameter selection dialog for Multilayer Perceptrons Neural Network in Weka

- □ Or this:
weka.classifiers.functions.MultilayerPerceptron -L 0.3 -M 0.2 -N 500 -V 0 -S 0 -E 20 -H a

# Skill: Hacking

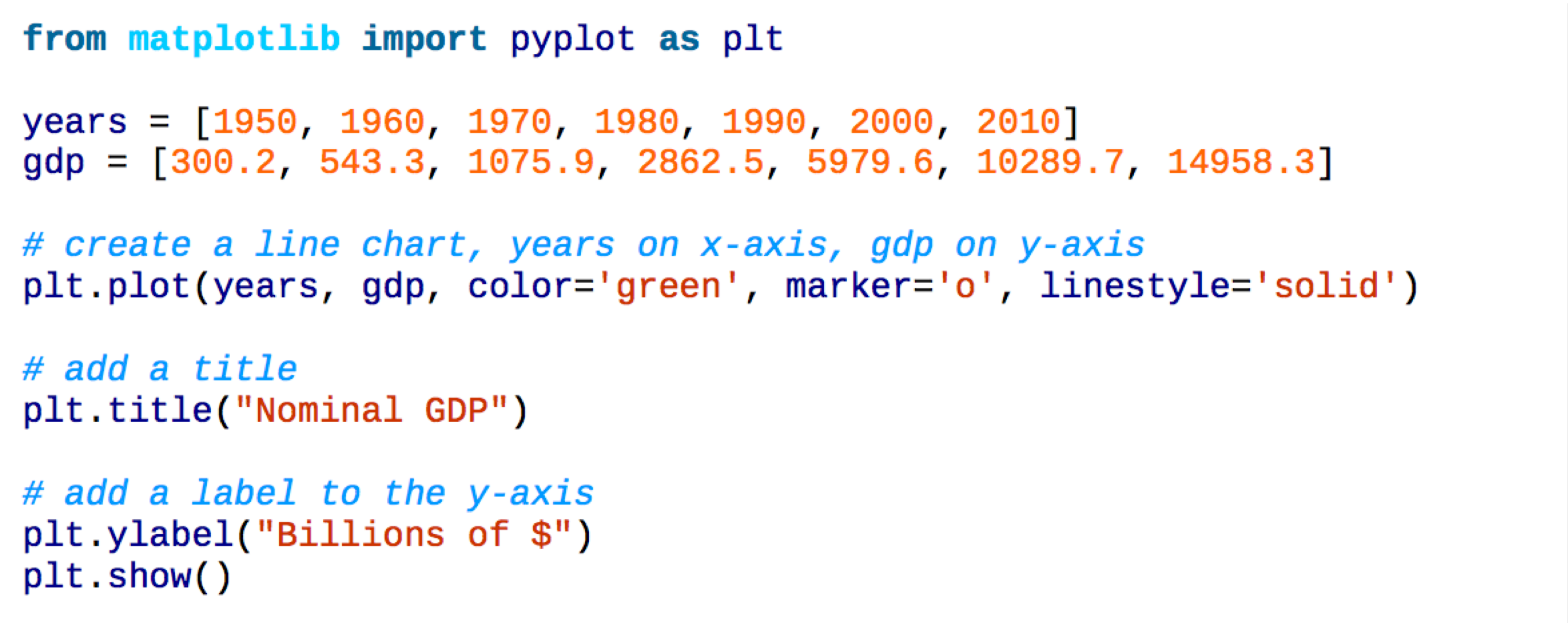


https://orange.biolab.si/

# Skill: Hacking Languages: Python

- Pros:
  - General purpose language.
  - Easy to script.
  - Lots of libraries.
- Cons:
  - Two main (sometimes incompatible) versions.
  - Many abandoned libraries.
  - *There should be one – and preferably only one – obvious way to do it.*

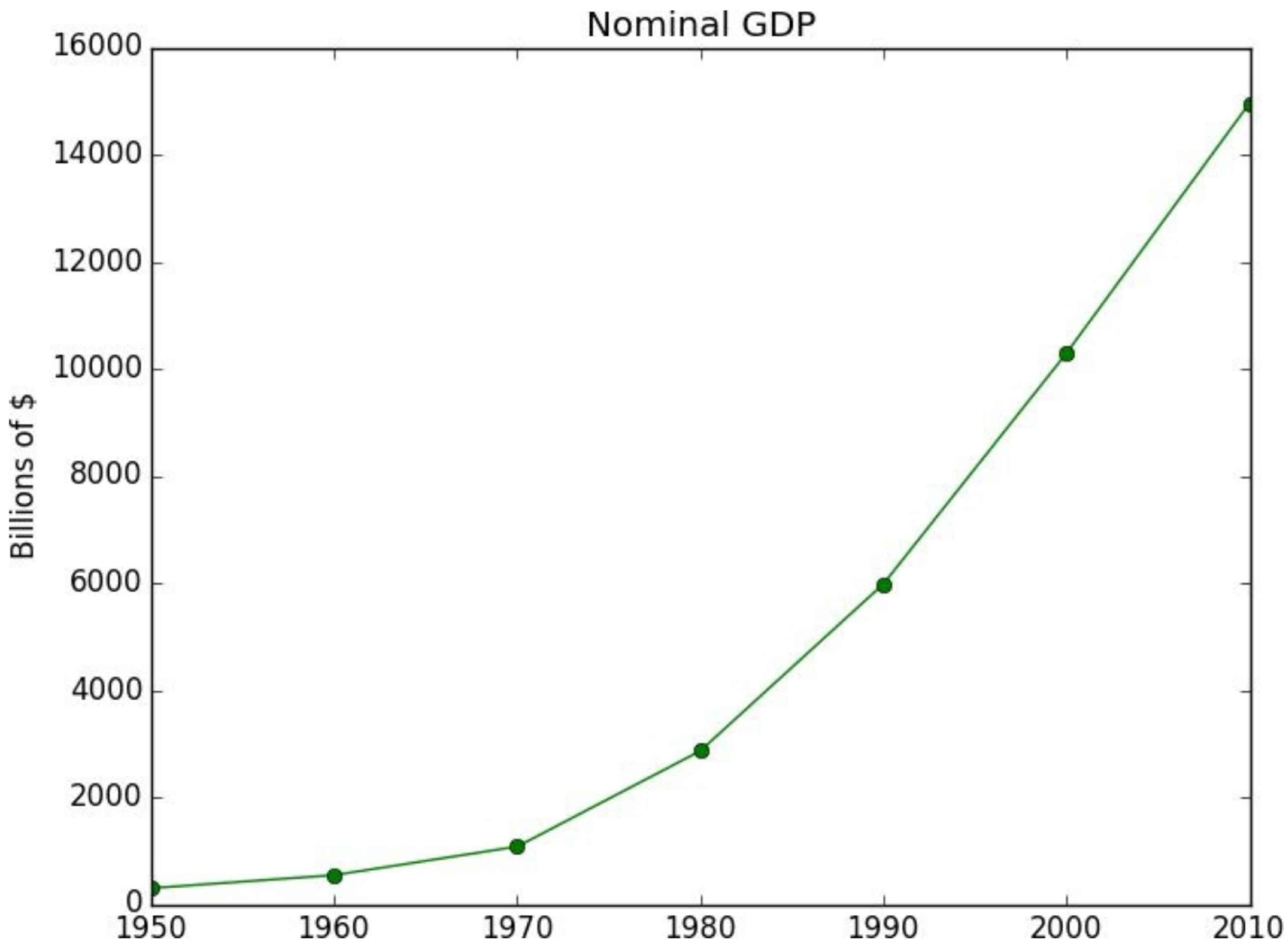
# Skill: Hacking Languages: Python



```python
from matplotlib import pyplot as plt

years = [1950, 1960, 1970, 1980, 1990, 2000, 2010]
gdp = [300.2, 543.3, 1075.9, 2862.5, 5979.6, 10289.7, 14958.3]

# create a line chart, years on x-axis, gdp on y-axis
plt.plot(years, gdp, color='green', marker='o', linestyle='solid')

# add a title
plt.title("Nominal GDP")

# add a label to the y-axis
plt.ylabel("Billions of $")
plt.show()
```

Joel Grus. Data Science from Scratch. O’Reilly, 2015

# Skill: Hacking Languages: Python

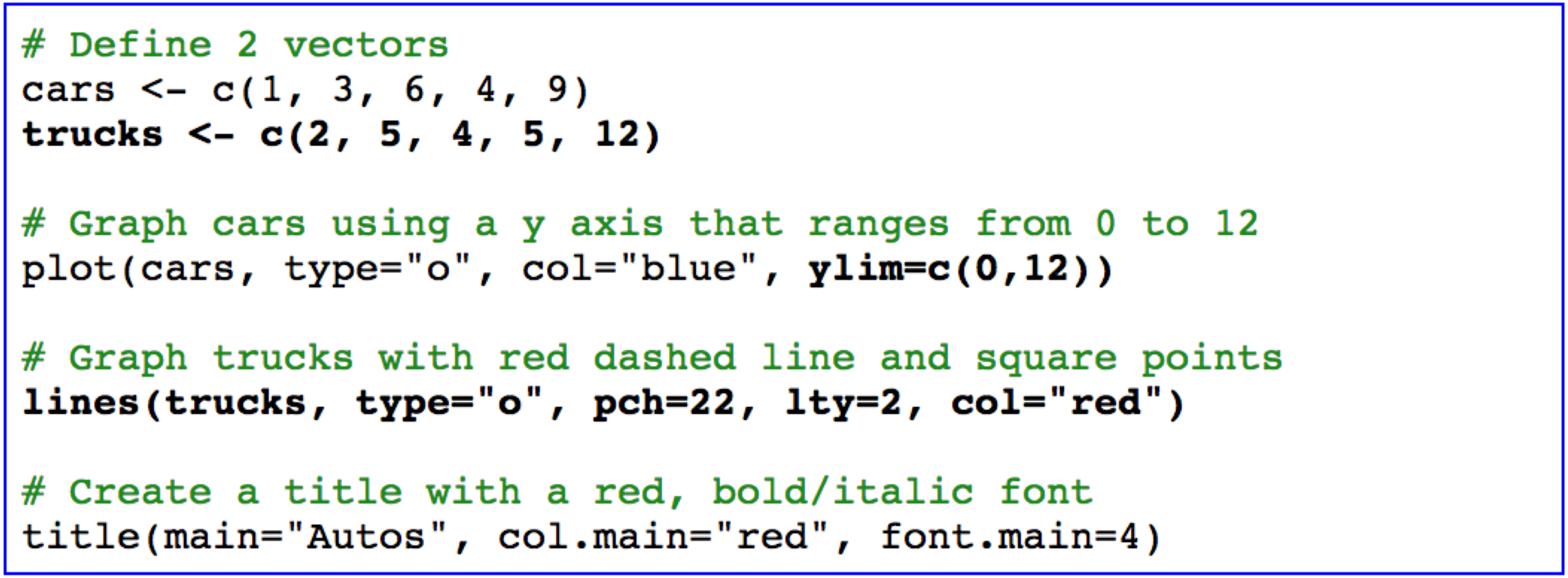


Joel Grus. Data Science from Scratch. O'Reilly, 2015

# Skill: Hacking Languages: R

- Pros:
  - Traditionally used by scientists.
  - Strong math/statistics support.
  - Many packages: CRAN.
- Cons:
  - Steep learning curve.

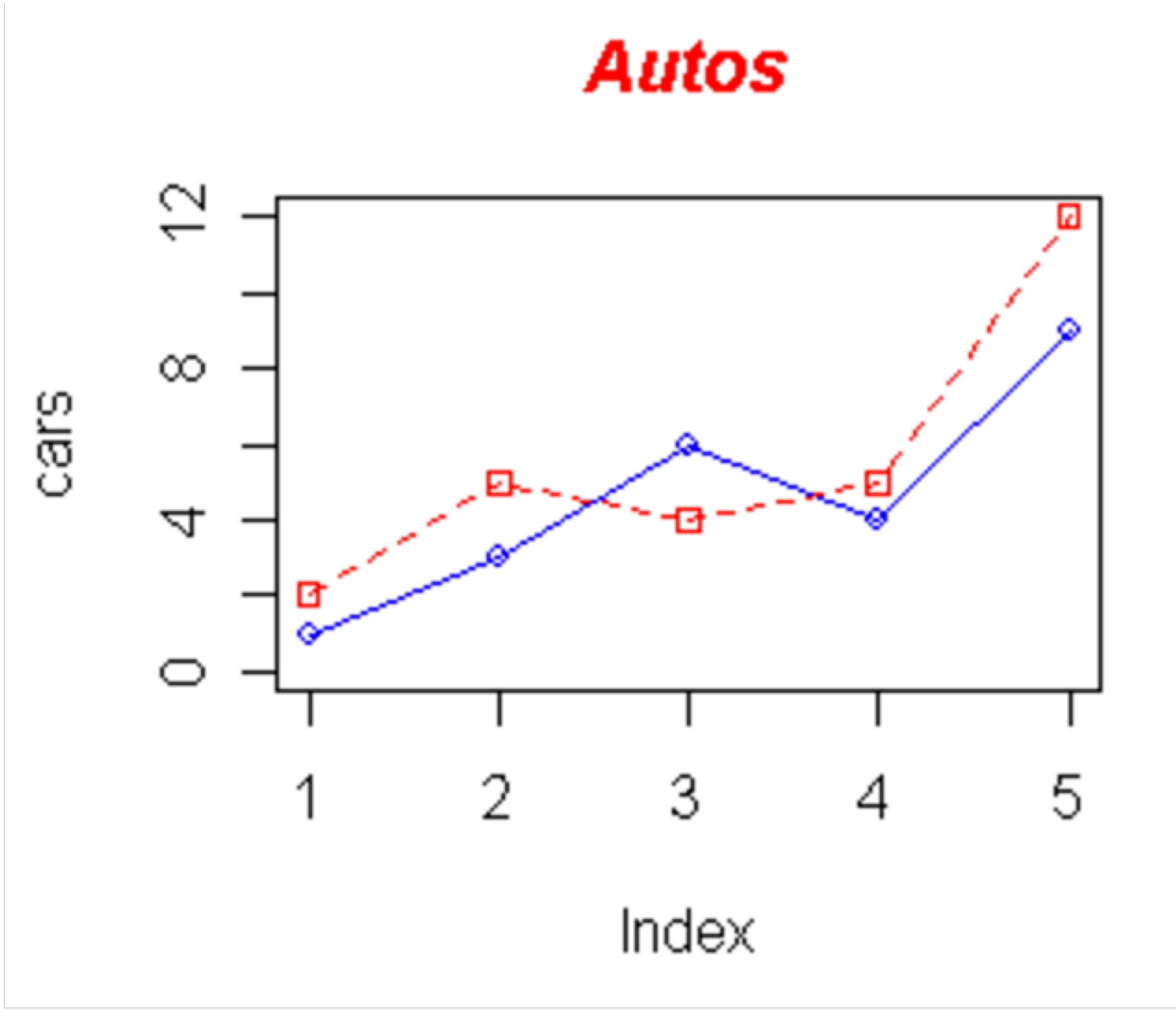
# Skill: Hacking Languages: R



```
# Define 2 vectors
cars <- c(1, 3, 6, 4, 9)
trucks <- c(2, 5, 4, 5, 12)

# Graph cars using a y axis that ranges from 0 to 12
plot(cars, type="o", col="blue", ylim=c(0,12))

# Graph trucks with red dashed line and square points
lines(trucks, type="o", pch=22, lty=2, col="red")

# Create a title with a red, bold/italic font
title(main="Autos", col.main="red", font.main=4)
```

http://www.harding.edu/fmccown/r/

# Skill: Hacking Languages: R

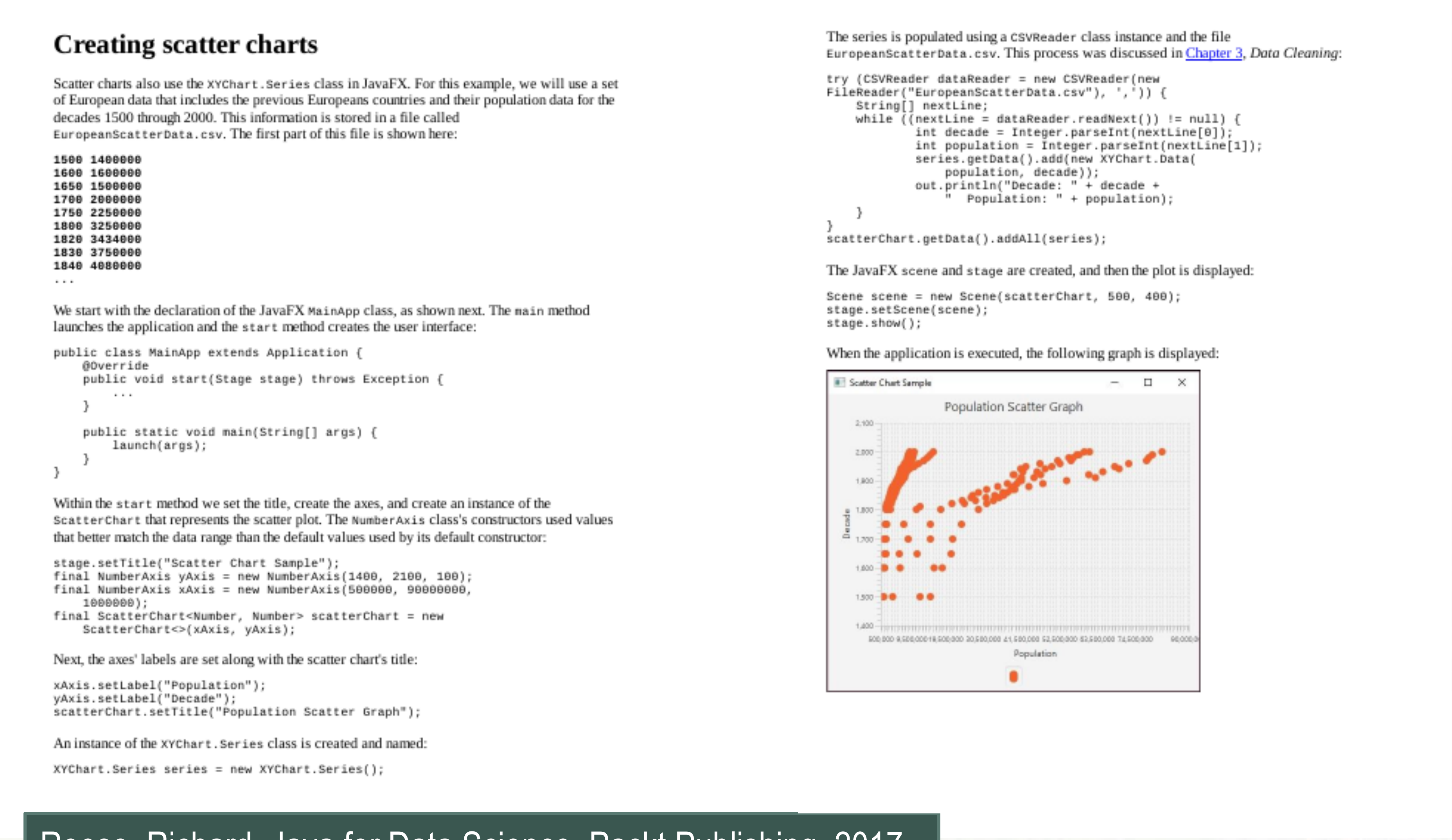


http://www.harding.edu/fmccown/r/

# Skill: Hacking Languages: Java

- Pros:
  - General purpose language.
  - Mature.
- Cons:
  - Prolix.
  - Many dependencies (for data science).
  - Not really a script language: hard to write quick hacks.

# Skill: Hacking Languages: Java



## Creating scatter charts

Scatter charts also use the `XYChart.Series` class in JavaFX. For this example, we will use a set of European data that includes the previous Europeans countries and their population data for the decades 1500 through 2000. This information is stored in a file called `EuropeanScatterData.csv`. The first part of this file is shown here:

```
1500 1400000
1600 1600000
1650 1500000
1700 2000000
1750 2250000
1800 3250000
1820 3434000
1830 3750000
1840 4080000
...
```

We start with the declaration of the JavaFX `MainApp` class, as shown next. The `main` method launches the application and the `start` method creates the user interface:

```
public class MainApp extends Application {
    @Override
    public void start(Stage stage) throws Exception {
        ...
    }

    public static void main(String[] args) {
        launch(args);
    }
}
```

Within the `start` method we set the title, create the axes, and create an instance of the `ScatterChart` that represents the scatter plot. The `NumberAxis` class's constructors used values that better match the data range than the default values used by its default constructor:

```
stage.setTitle("Scatter Chart Sample");
final NumberAxis yAxis = new NumberAxis(1400, 2100, 100);
final NumberAxis xAxis = new NumberAxis(500000, 90000000,
    1000000);
final ScatterChart<Number, Number> scatterChart = new
    ScatterChart<>(xAxis, yAxis);
```

Next, the axes' labels are set along with the scatter chart's title:

```
xAxis.setLabel("Population");
yAxis.setLabel("Decade");
scatterChart.setTitle("Population Scatter Graph");
```

An instance of the `XYChart.Series` class is created and named:

```
XYChart.Series series = new XYChart.Series();
```

The series is populated using a `CSVReader` class instance and the file `EuropeanScatterData.csv`. This process was discussed in Chapter 3, *Data Cleaning*:

```
try (CSVReader dataReader = new CSVReader(new
FileReader("EuropeanScatterData.csv"), ',')) {
    String[] nextLine;
    while ((nextLine = dataReader.readNext()) != null) {
            int decade = Integer.parseInt(nextLine[0]);
            int population = Integer.parseInt(nextLine[1]);
            series.getData().add(new XYChart.Data(
                population, decade));
            out.println("Decade: " + decade +
                "  Population: " + population);
    }
}
scatterChart.getData().addAll(series);
```

The JavaFX `scene` and `stage` are created, and then the plot is displayed:

```
Scene scene = new Scene(scatterChart, 500, 400);
stage.setScene(scene);
stage.show();
```

When the application is executed, the following graph is displayed:

Reese, Richard. Java for Data Science. Packt Publishing, 2017

# Skill: Hacking Languages: Julia

- Pros:
  - Developed for numerical computing.
  - Can easily call C code.
- Cons:
  - Still young.
  - Few DS packages and libraries.

Voulgaris, Zacharias. Programming Languages for Data Science. O'Reilly, 2017

# Skill: Hacking Languages: Scala

- Pros:
  - Syntax similar to Java.
  - Growing interest in DS community.
- Cons:
  - Still young.

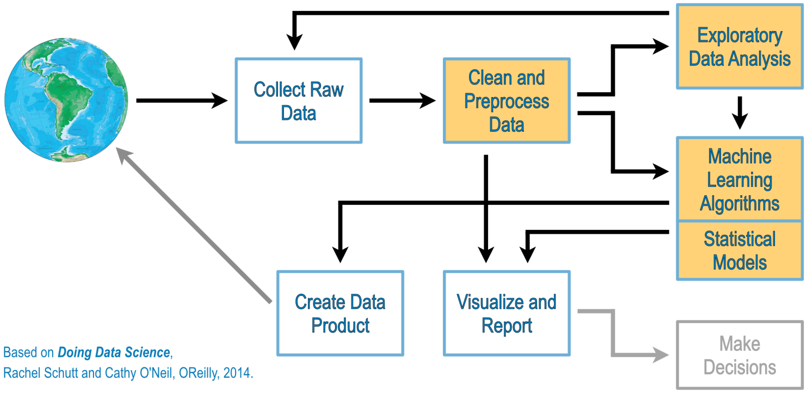
Voulgaris, Zacharias. Programming Languages for Data Science. O'Reilly, 2017

# Skill: Hacking Languages: Which one?

- We will focus on R and Python.
- Not all examples will be given for both.
- Avoid Language Wars:
  - C, C#, C++, Perl, JavaScript, Lisp, Prolog, Matlab, Octave, Cobol, Fortran…
  - Languages are tools. Choose an appropriate one.

# Skill: Exploratory Data Analysis

- We have the data.
  Now what?
  - Do we know what we want to discover?
  - We need basic skills in statistics and data modeling.



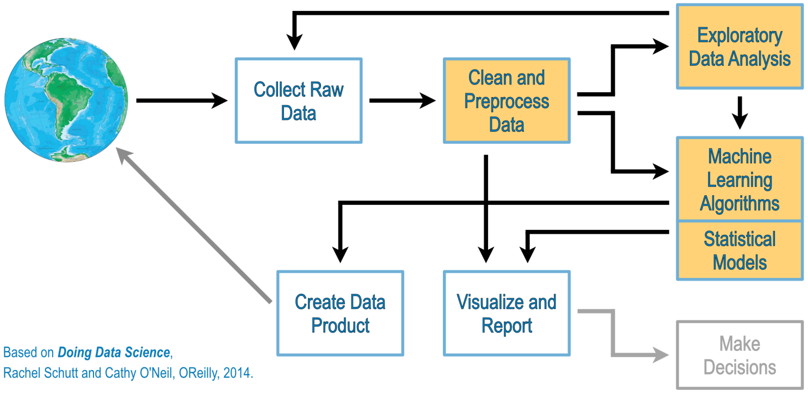
Based on ***Doing Data Science***, Rachel Schutt and Cathy O'Neil, OReilly, 2014.

- Start exploring: Exploratory Data Analysis
  - Make different plots and charts to explore variables.
  - Get some basic statistics.
  - Evaluate the type of information we can extract from the data.

# Skill: Exploratory Data Analysis

- □ Basic statistics – avoid complex models (for the time being).

- □ Basic plots that explore relations between the variables on the data.

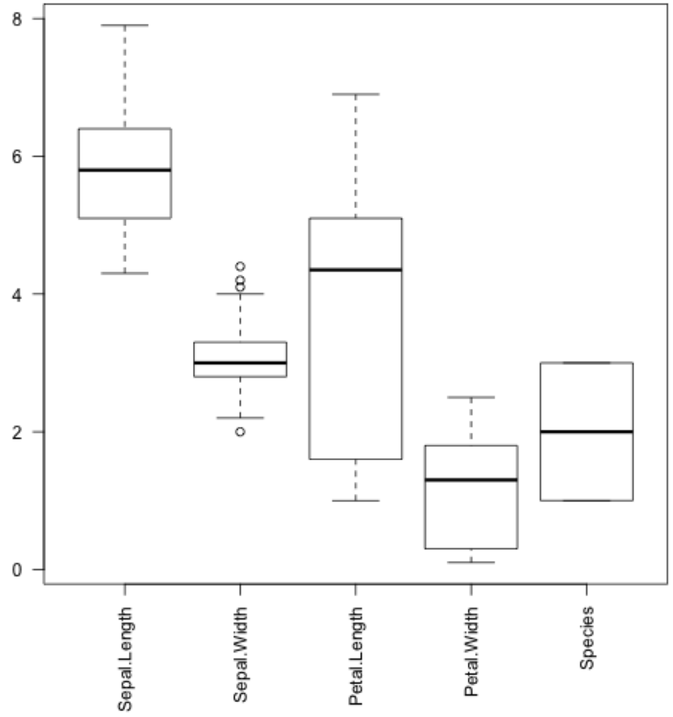


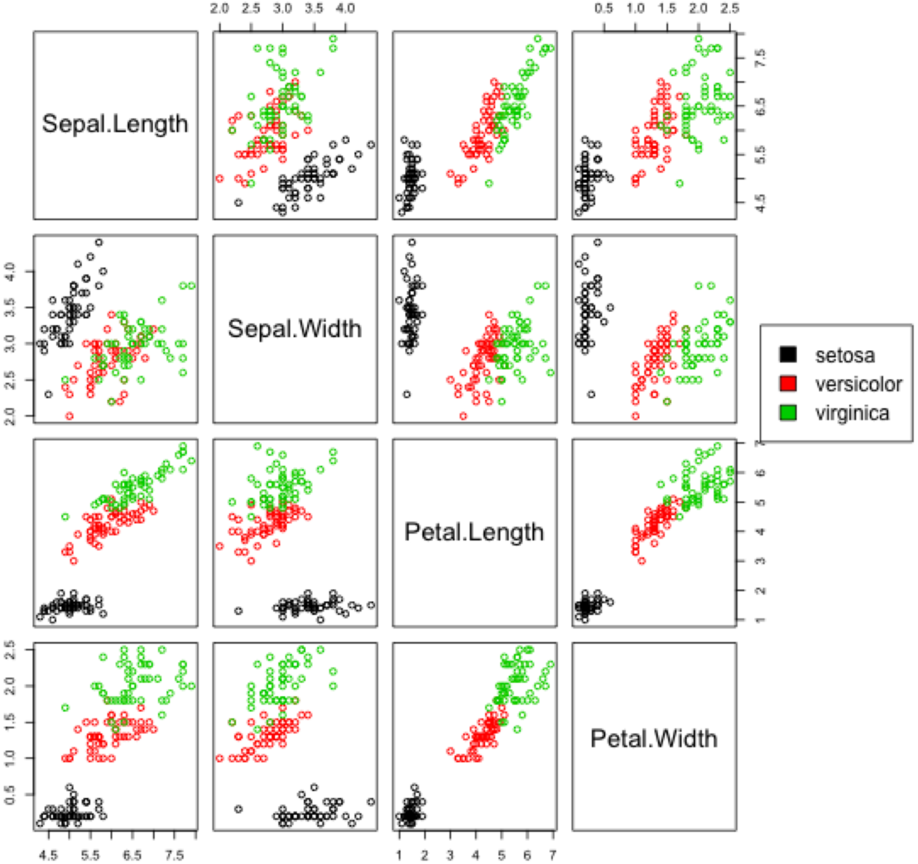
Based on ***Doing Data Science***, Rachel Schutt and Cathy O'Neil, OReilly, 2014.

- □ Used to gain insight on the data and relations, may suggest which advanced analysis (e.g. machine learning) can be applied.

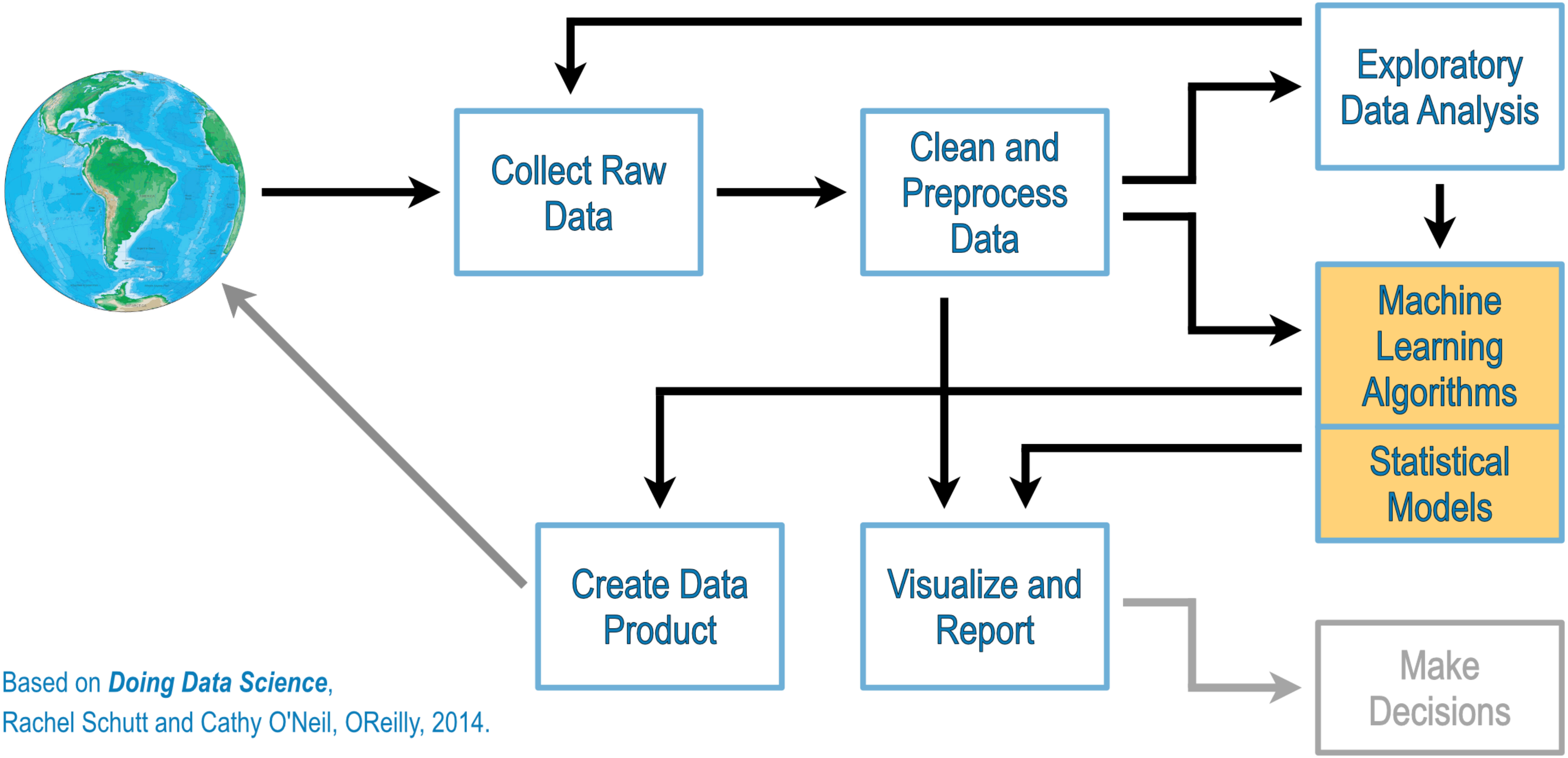
# Skill: Exploratory Data Analysis

- Quick example: Iris Dataset.





# Skill: More Analysis



Collect Raw Data

Clean and Preprocess Data

Exploratory Data Analysis

Machine Learning Algorithms

Statistical Models

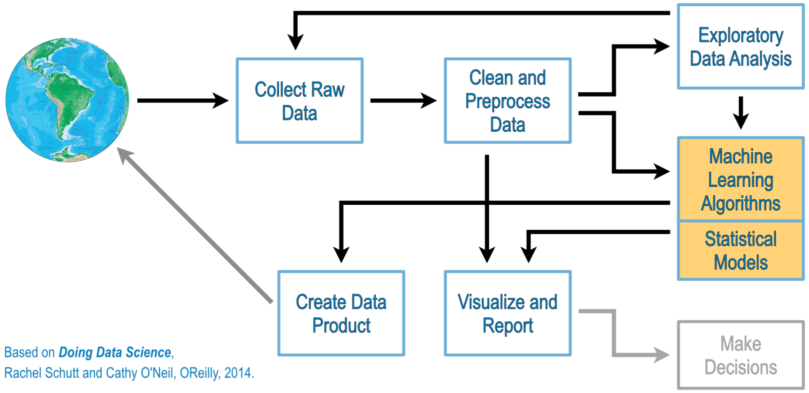
Create Data Product

Visualize and Report

Make Decisions

Based on ***Doing Data Science***,
Rachel Schutt and Cathy O'Neil, OReilly, 2014.

# Skill: More Analysis

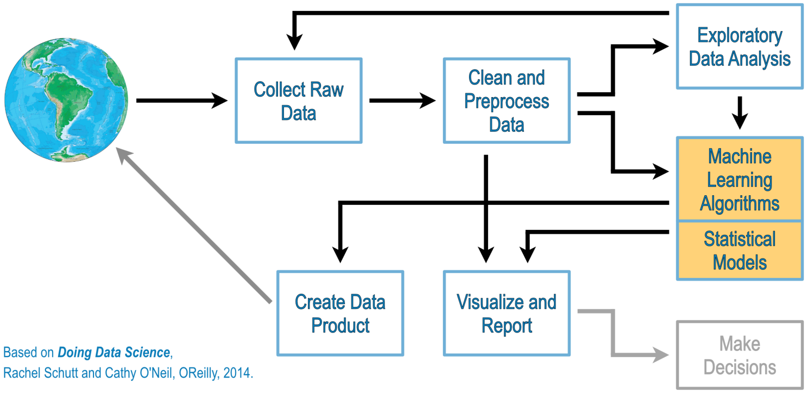


Based on ***Doing Data Science***, Rachel Schutt and Cathy O'Neil, OReilly, 2014.

- □ What can I learn from my data?
- □ How can I describe interesting features of it?
- □ *Exploratory Data Analysis* can give hints on the nature of the data and which knowledge it may contain.
- □ *Machine Learning* and *Data Mining* can be used to create models that describe the data: even data we don't have!

# Skill: More Analysis

- **Warnings:**
  - Models may be more complex than suggested by EDA.
  - Many models, techniques, algorithms, implementations, parameters, etc.
  - Models should be interpretable!
  - Scalability may be an issue.



Collect Raw Data

Clean and Preprocess Data

Exploratory Data Analysis

Machine Learning Algorithms

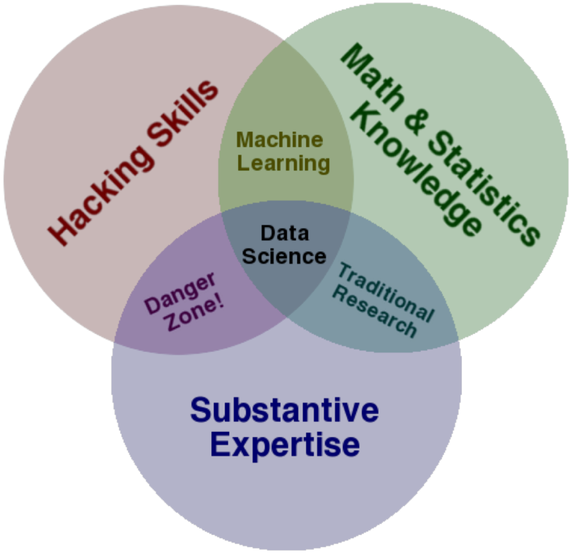
Statistical Models

Create Data Product

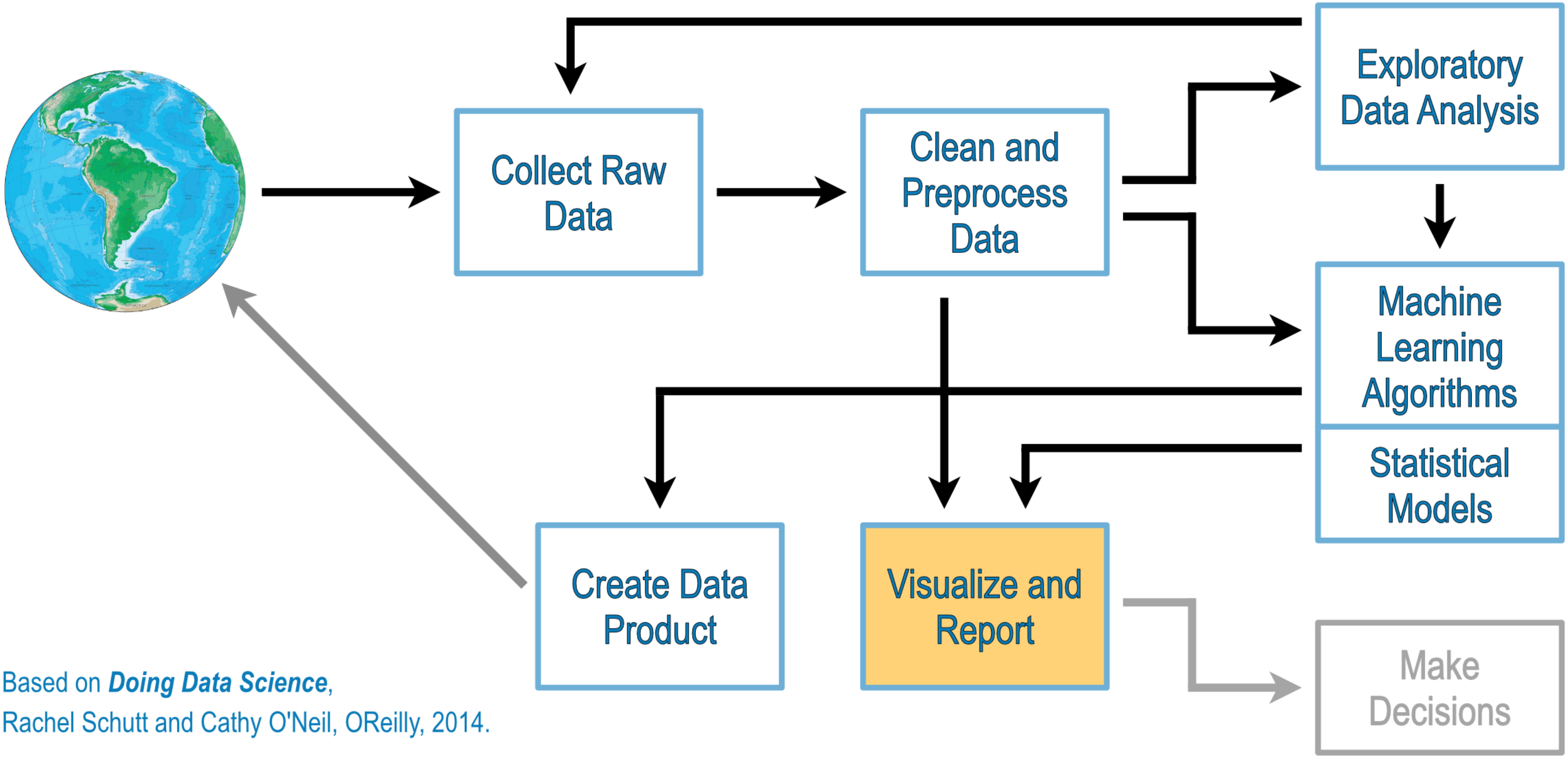
Visualize and Report

Make Decisions

Based on ***Doing Data Science***, Rachel Schutt and Cathy O'Neil, OReilly, 2014.



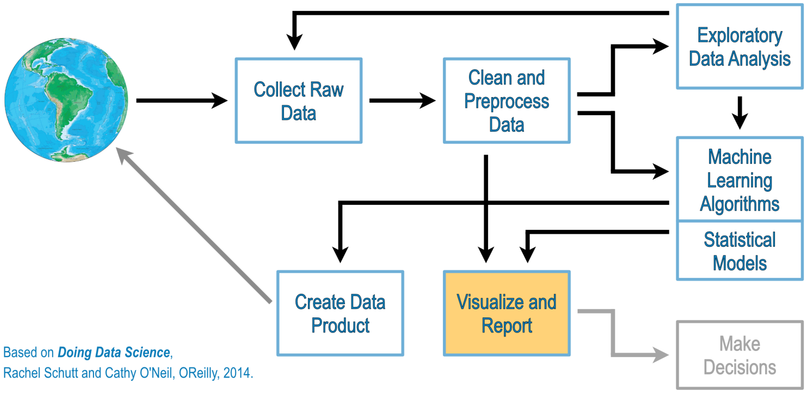
# Skill: Communicate Results



Based on ***Doing Data Science***,
Rachel Schutt and Cathy O'Neil, OReilly, 2014.

# Skill: Communicate Results

- Another interdisciplinary area:
  - Visualization: art and science.
  - Design: meaning for users.
- Most basic results can be achieved with programming languages.
- Consider learning other, specific visualization tools.

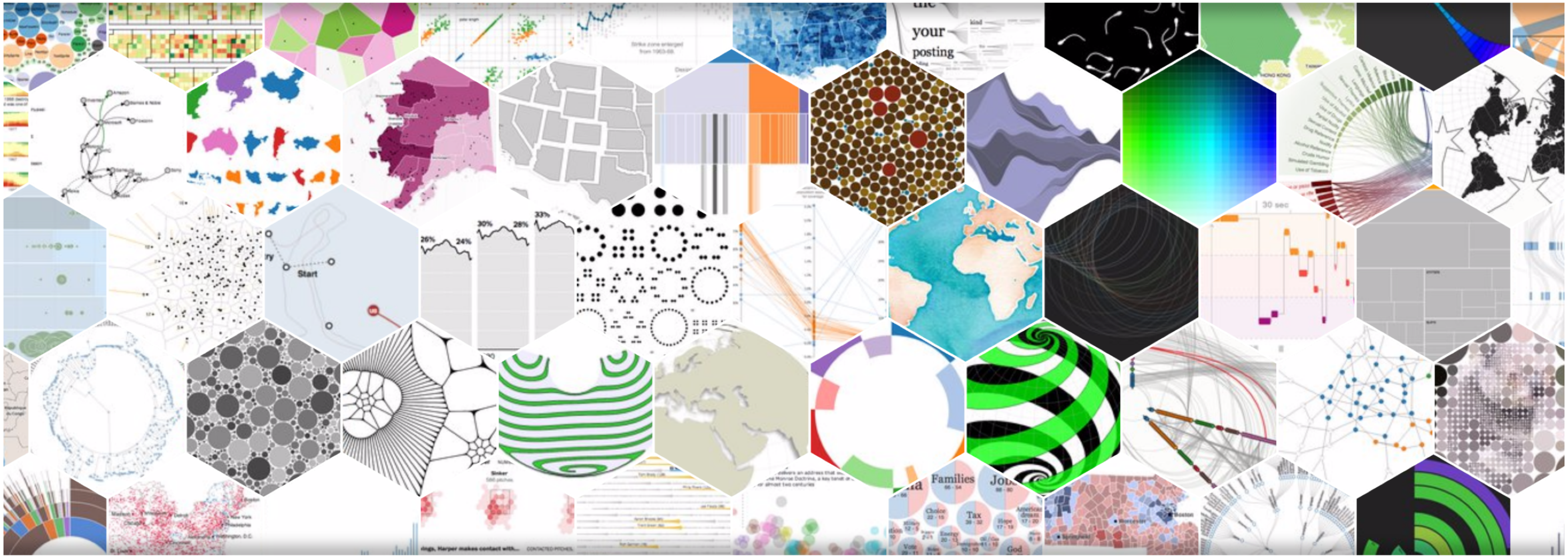


Based on *Doing Data Science*, Rachel Schutt and Cathy O'Neil, OReilly, 2014.

# Skill: Communicate Results

- **D3.js:** ***Data-Driven Documents***
  - JavaScript library for DOM (=data!) manipulation.

Data-Driven Documents



https://d3js.org/

# Skill: Communicate Results

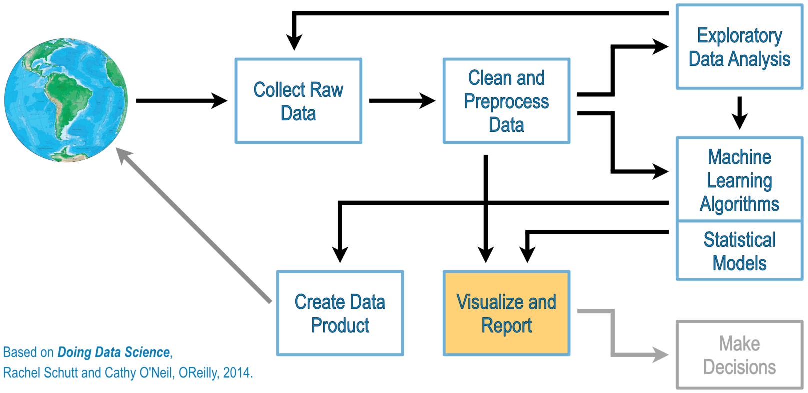
- □ **Bokeh:** Python interactive visualization library.

87.62
88.905
56
Ba
Barium
137.327
71
Lu
Luteti
174.96
1.0
10.0
100.0
Rh
Ru
Pd
Tc
Mo
Ag
Cd
length
pe

http://bokeh.pydata.org/en/latest/

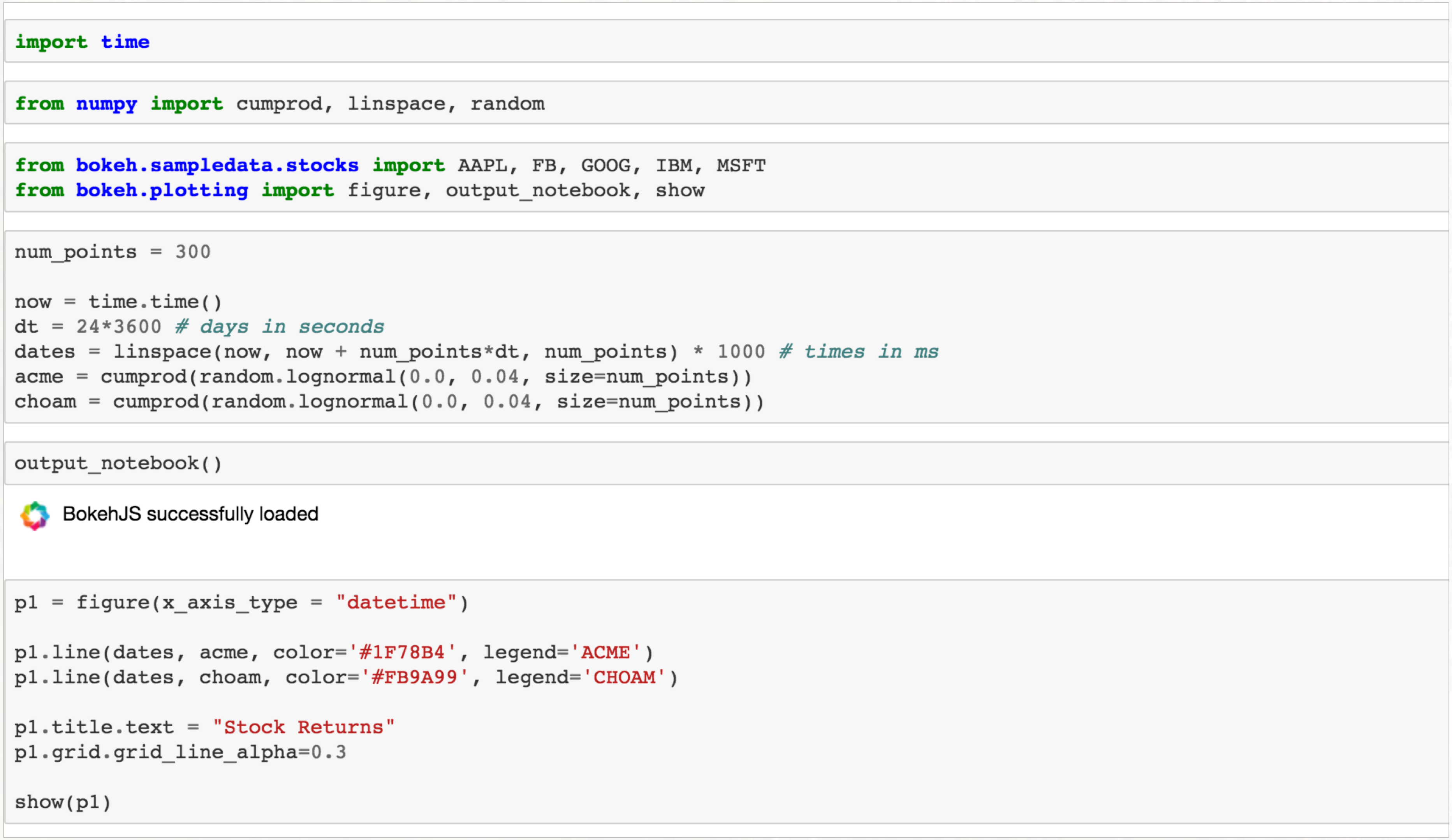
# Skill: Communicate Results

- Online notebooks, e.g.: Jupyter, SciServer.
  - Allows creation of interactive **notebooks** in several languages.
- *Reproducible Research!*



Based on ***Doing Data Science***, Rachel Schutt and Cathy O'Neil, OReilly, 2014.

# Jupyter



```
import time
```

```
from numpy import cumprod, linspace, random
```

```
from bokeh.sampledata.stocks import AAPL, FB, GOOG, IBM, MSFT
from bokeh.plotting import figure, output_notebook, show
```

```
num_points = 300

now = time.time()
dt = 24*3600 # days in seconds
dates = linspace(now, now + num_points*dt, num_points) * 1000 # times in ms
acme = cumprod(random.lognormal(0.0, 0.04, size=num_points))
choam = cumprod(random.lognormal(0.0, 0.04, size=num_points))
```

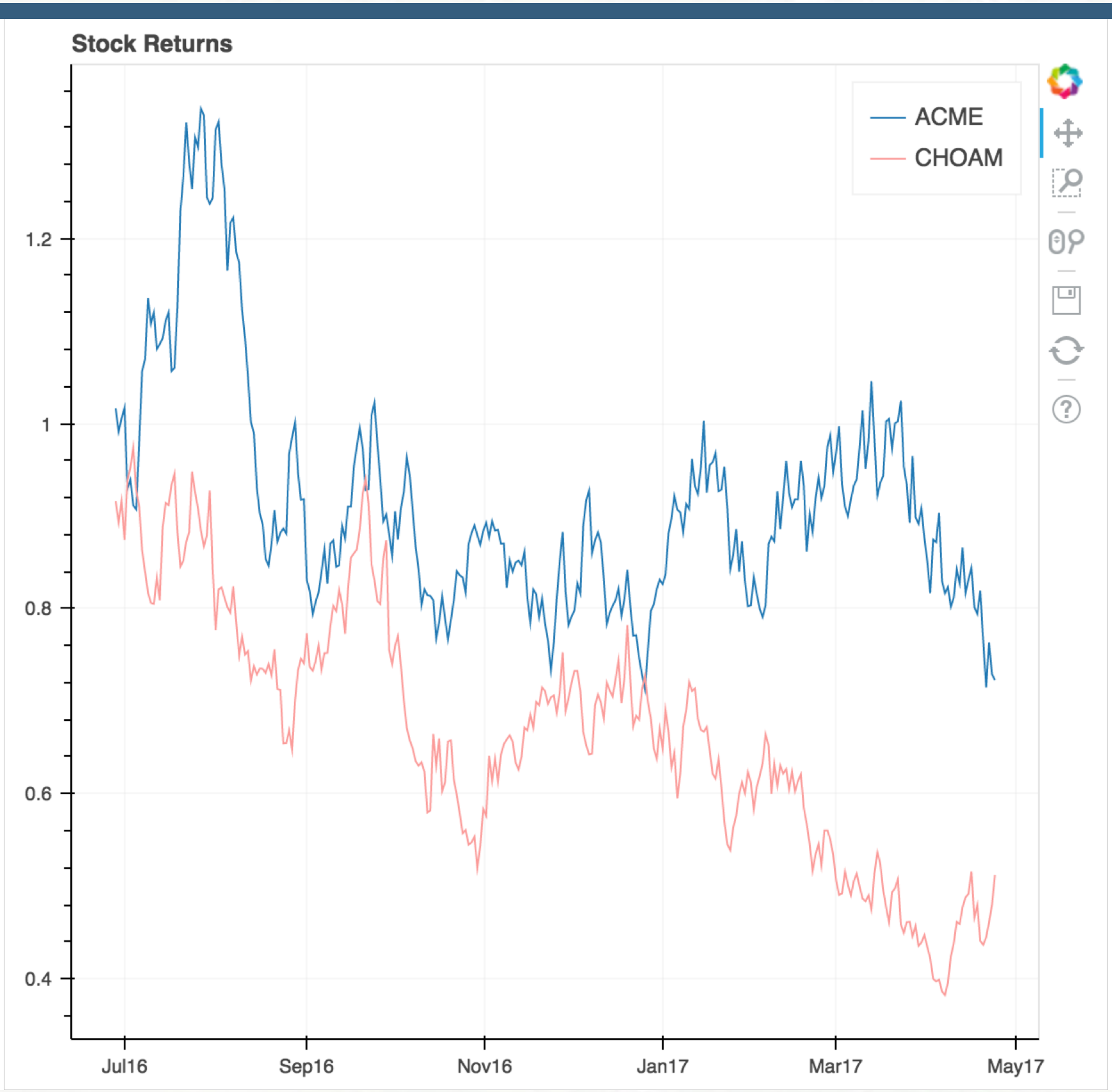
```
output_notebook()
```

```
BokehJS successfully loaded
```

```
p1 = figure(x_axis_type = "datetime")

p1.line(dates, acme, color='#1F78B4', legend='ACME')
p1.line(dates, choam, color='#FB9A99', legend='CHOAM')

p1.title.text = "Stock Returns"
p1.grid.grid_line_alpha=0.3

show(p1)
```

http://nbviewer.jupyter.org/github/bokeh/bokeh-notebooks/blob/master/index.ipynb

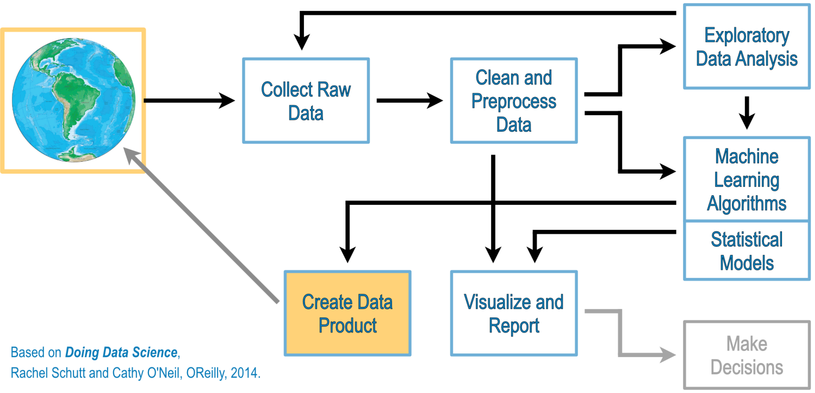
# Jupyter



http://nbviewer.jupyter.org/github/bokeh/bokeh-notebooks/blob/master/index.ipynb

# *Skill*: Understand (better) the problem

- What data there ought to exist?
  - Data Product!



Based on ***Doing Data Science***, Rachel Schutt and Cathy O'Neil, OReilly, 2014.

- After the whole process, what data would be interesting to…
  - Understand better the whole problem?
  - Add value to the existing data?
  - Allow the creation of new applications?

These are the main objectives of a Data Scientist!

# Final remarks…

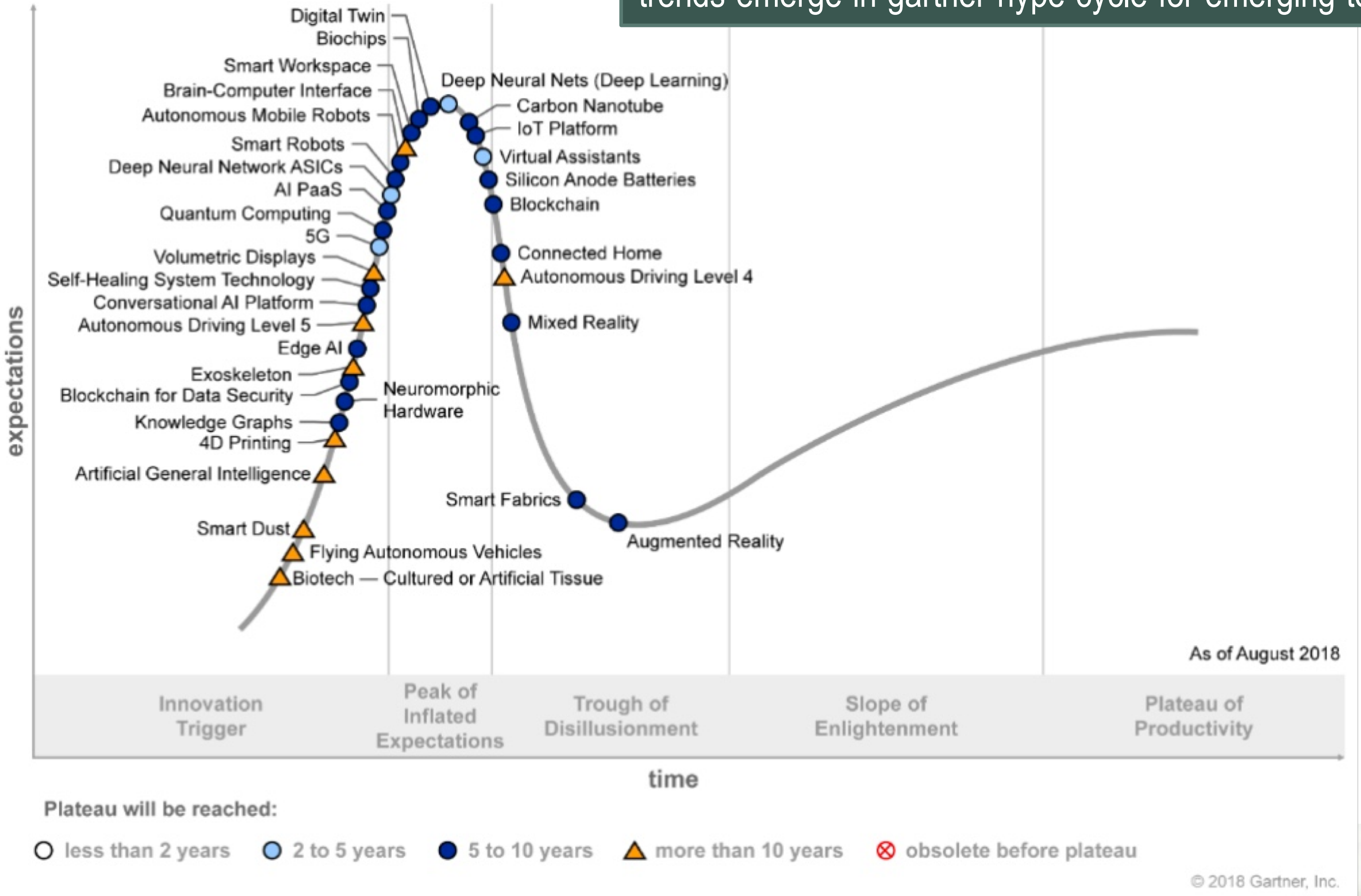
# In conclusion…

- Definition of Data Science is very subjective.
  - Hype ***is*** an issue!
- If you’re already a scientist (students too!):
  - Learn how to hack (SQL, Python, R).
  - Learn and practice reproducibility.
  - Embrace EDA!
  - *Organize your workflow.*

# In conclusion…

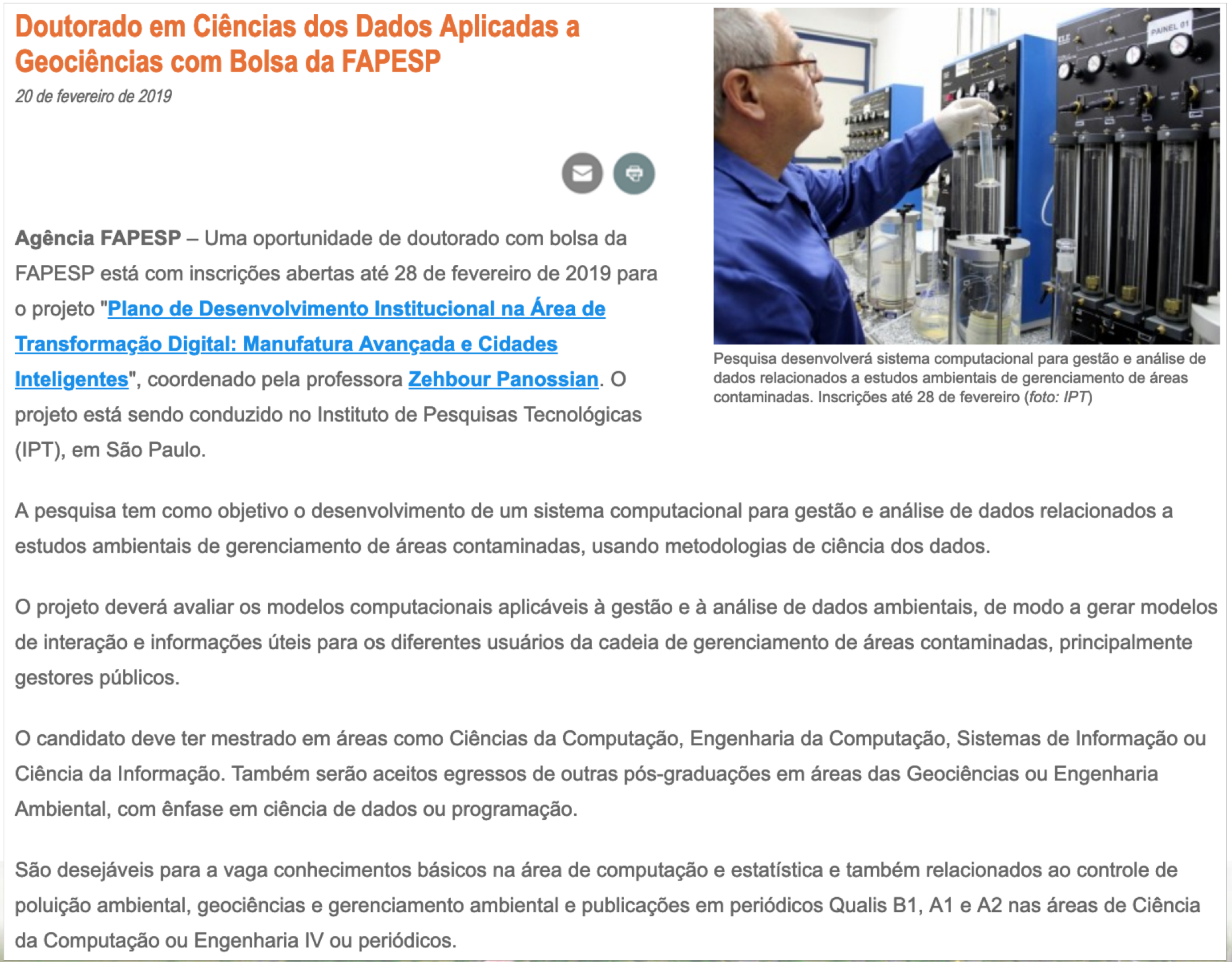
- □ Hype *is* an issue!

Gartner, “5 Trends Emerge in the Gartner Hype Cycle for Emerging Technologies, 2018”, https://www.gartner.com/smarterwithgartner/5-trends-emerge-in-gartner-hype-cycle-for-emerging-technologies-2018/



# In conclusion…

- □ But there are real opportunities!



## Doutorado em Ciências dos Dados Aplicadas a Geociências com Bolsa da FAPESP

*20 de fevereiro de 2019*

**Agência FAPESP** – Uma oportunidade de doutorado com bolsa da FAPESP está com inscrições abertas até 28 de fevereiro de 2019 para o projeto "**Plano de Desenvolvimento Institucional na Área de Transformação Digital: Manufatura Avançada e Cidades Inteligentes**", coordenado pela professora **Zehbour Panossian**. O projeto está sendo conduzido no Instituto de Pesquisas Tecnológicas (IPT), em São Paulo.

Pesquisa desenvolverá sistema computacional para gestão e análise de dados relacionados a estudos ambientais de gerenciamento de áreas contaminadas. Inscrições até 28 de fevereiro (*foto: IPT*)

A pesquisa tem como objetivo o desenvolvimento de um sistema computacional para gestão e análise de dados relacionados a estudos ambientais de gerenciamento de áreas contaminadas, usando metodologias de ciência dos dados.

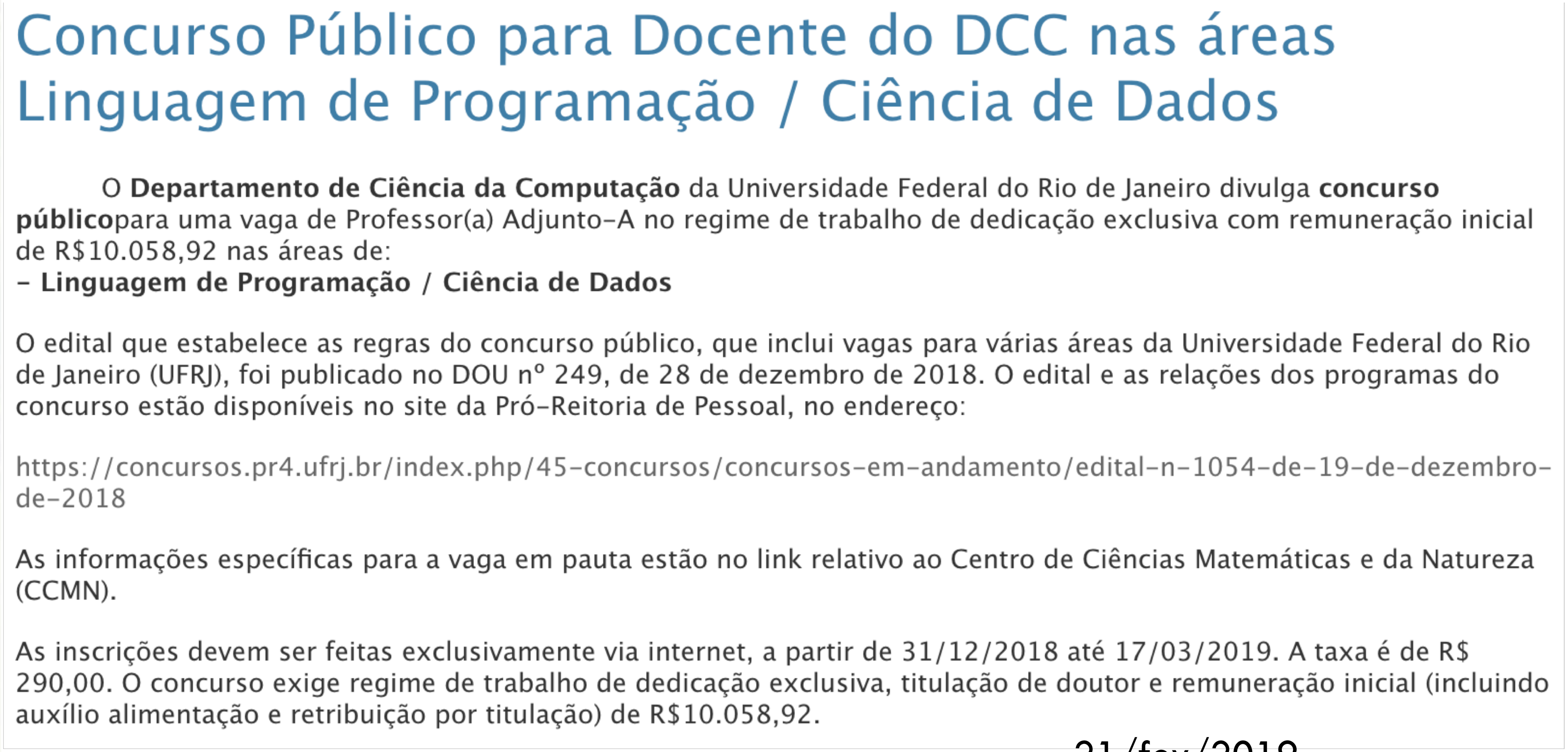
O projeto deverá avaliar os modelos computacionais aplicáveis à gestão e à análise de dados ambientais, de modo a gerar modelos de interação e informações úteis para os diferentes usuários da cadeia de gerenciamento de áreas contaminadas, principalmente gestores públicos.

O candidato deve ter mestrado em áreas como Ciências da Computação, Engenharia da Computação, Sistemas de Informação ou Ciência da Informação. Também serão aceitos egressos de outras pós-graduações em áreas das Geociências ou Engenharia Ambiental, com ênfase em ciência de dados ou programação.

São desejáveis para a vaga conhecimentos básicos na área de computação e estatística e também relacionados ao controle de poluição ambiental, geociências e gerenciamento ambiental e publicações em periódicos Qualis B1, A1 e A2 nas áreas de Ciência da Computação ou Engenharia IV ou periódicos.

# In conclusion…

- □ But there are real opportunities!



## Concurso Público para Docente do DCC nas áreas Linguagem de Programação / Ciência de Dados

O **Departamento de Ciência da Computação** da Universidade Federal do Rio de Janeiro divulga **concurso públicopara** uma vaga de Professor(a) Adjunto-A no regime de trabalho de dedicação exclusiva com remuneração inicial de R$10.058,92 nas áreas de:
**- Linguagem de Programação / Ciência de Dados**

O edital que estabelece as regras do concurso público, que inclui vagas para várias áreas da Universidade Federal do Rio de Janeiro (UFRJ), foi publicado no DOU nº 249, de 28 de dezembro de 2018. O edital e as relações dos programas do concurso estão disponíveis no site da Pró-Reitoria de Pessoal, no endereço:

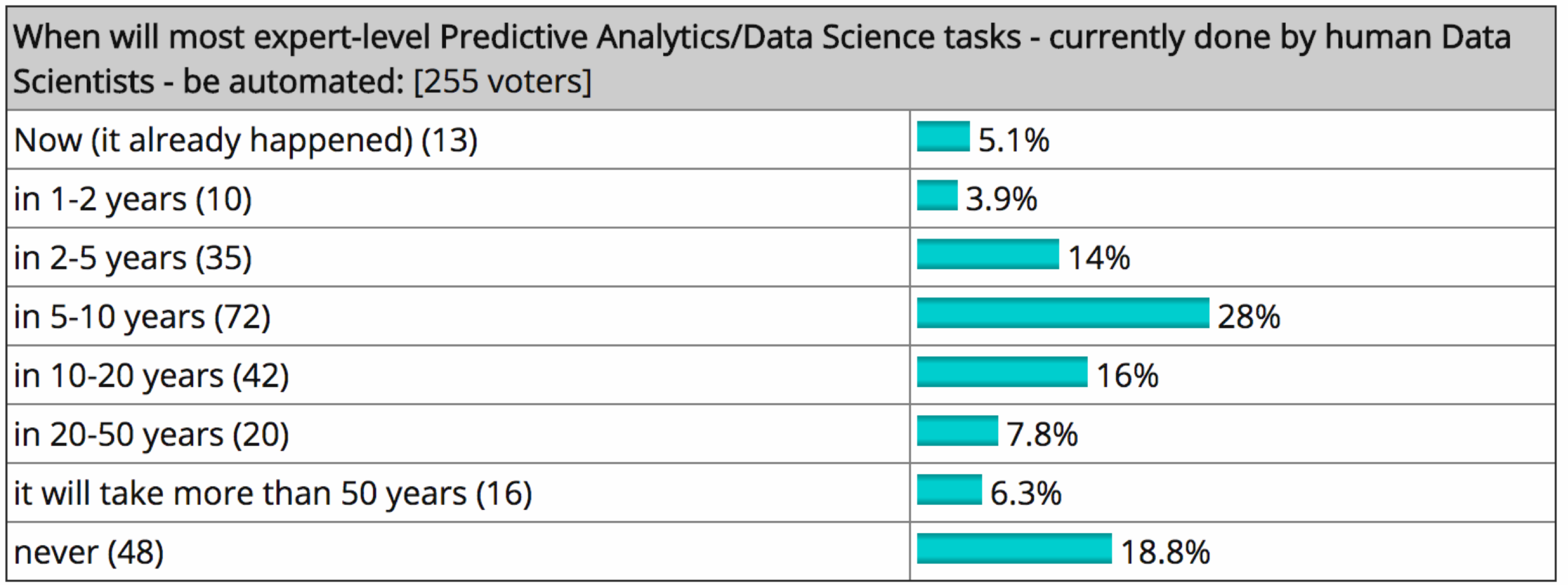
https://concursos.pr4.ufrj.br/index.php/45-concursos/concursos-em-andamento/edital-n-1054-de-19-de-dezembro-de-2018

As informações específicas para a vaga em pauta estão no link relativo ao Centro de Ciências Matemáticas e da Natureza (CCMN).

As inscrições devem ser feitas exclusivamente via internet, a partir de 31/12/2018 até 17/03/2019. A taxa é de R$ 290,00. O concurso exige regime de trabalho de dedicação exclusiva, titulação de doutor e remuneração inicial (incluindo auxílio alimentação e retribuição por titulação) de R$10.058,92.

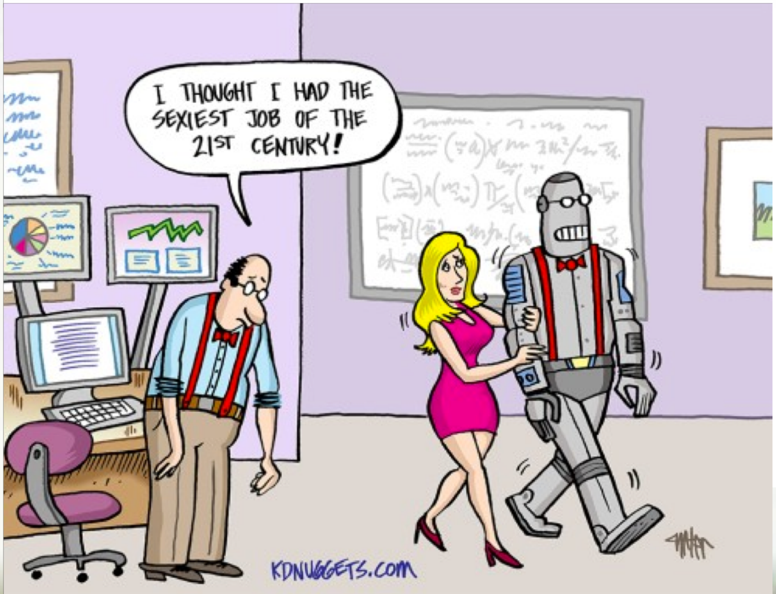
21/fev/2019

# Oh No!



When will most expert-level Predictive Analytics/Data Science tasks - currently done by human Data Scientists - be automated: [255 voters]

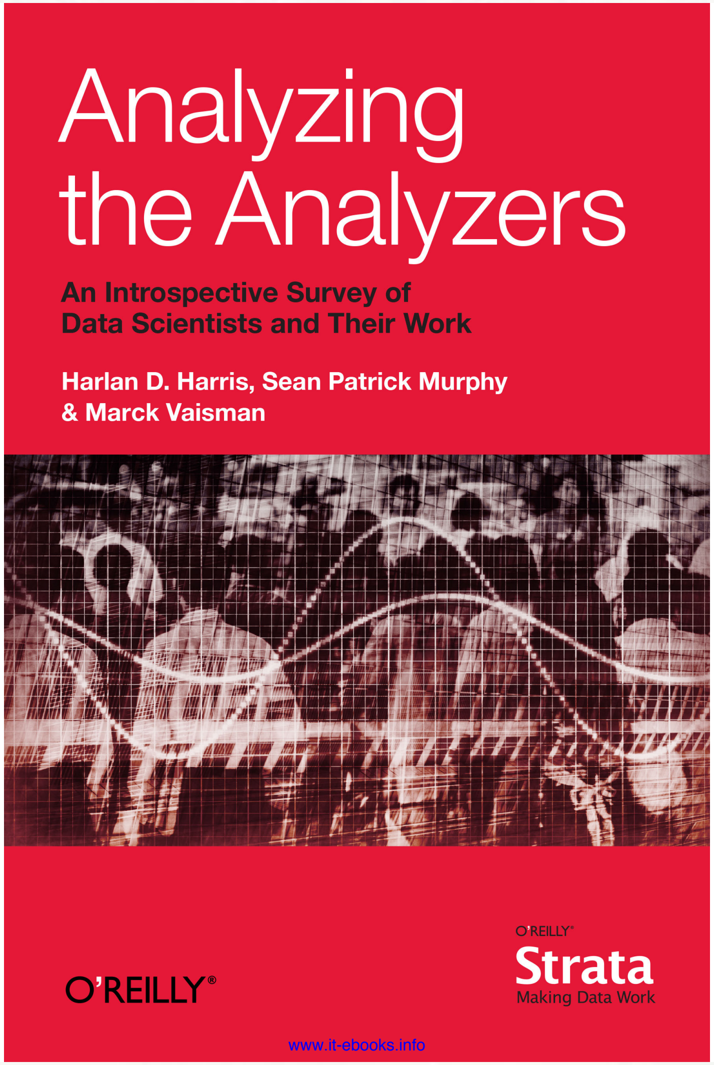
| | |
|---|---|
| Now (it already happened) (13) | 5.1% |
| in 1-2 years (10) | 3.9% |
| in 2-5 years (35) | 14% |
| in 5-10 years (72) | 28% |
| in 10-20 years (42) | 16% |
| in 20-50 years (20) | 7.8% |
| it will take more than 50 years (16) | 6.3% |
| never (48) | 18.8% |



Data Scientists Automated and Unemployed by 2025?
https://www.kdnuggets.com/2015/05/data-scientists-automated-2025.html

# References

# References



Analyzing the Analyzers

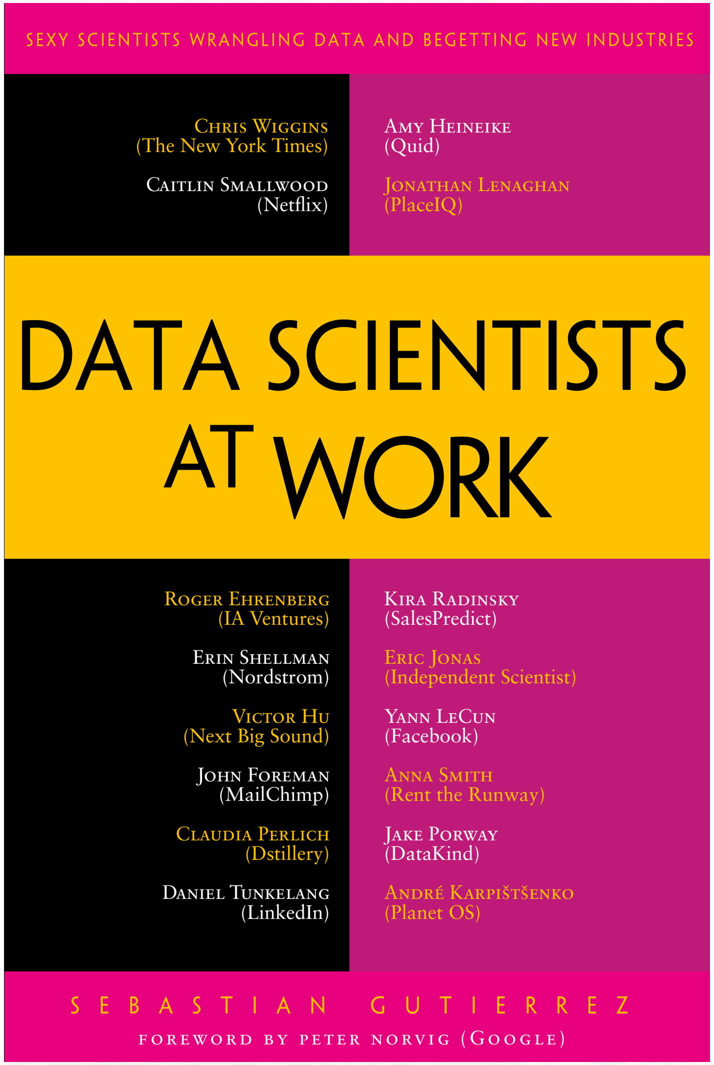
An Introspective Survey of Data Scientists and Their Work

Harlan D. Harris, Sean Patrick Murphy & Marck Vaisman

O'REILLY®

O'REILLY® Strata Making Data Work

www.it-ebooks.info



SEXY SCIENTISTS WRANGLING DATA AND BEGETTING NEW INDUSTRIES

Chris Wiggins (The New York Times)

Amy Heineike (Quid)

Caitlin Smallwood (Netflix)

Jonathan Lenaghan (PlaceIQ)

DATA SCIENTISTS AT WORK

Roger Ehrenberg (IA Ventures)

Kira Radinsky (SalesPredict)

Erin Shellman (Nordstrom)

Eric Jonas (Independent Scientist)

Victor Hu (Next Big Sound)

Yann LeCun (Facebook)

John Foreman (MailChimp)

Anna Smith (Rent the Runway)

Claudia Perlich (Dstillery)

Jake Porway (DataKind)

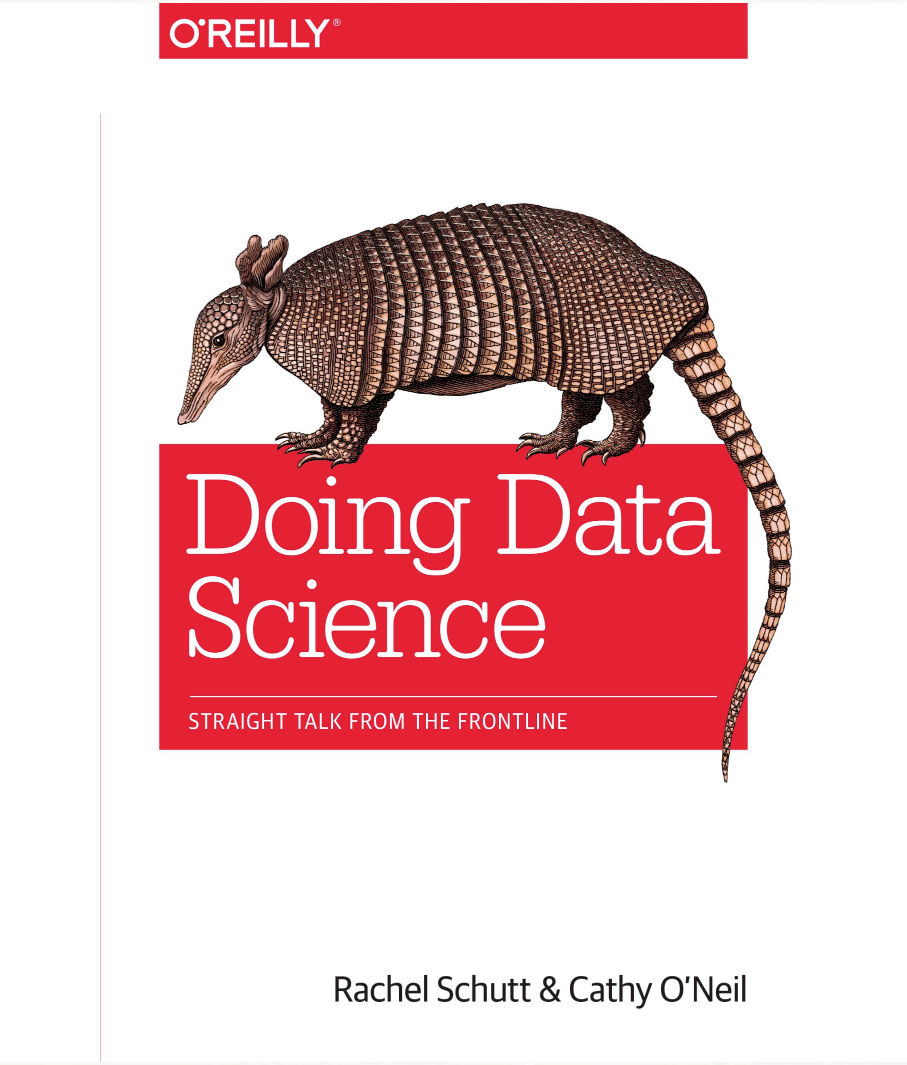
Daniel Tunkelang (LinkedIn)

André Karpištšenko (Planet OS)

SEBASTIAN GUTIERREZ

FOREWORD BY PETER NORVIG (GOOGLE)

# References

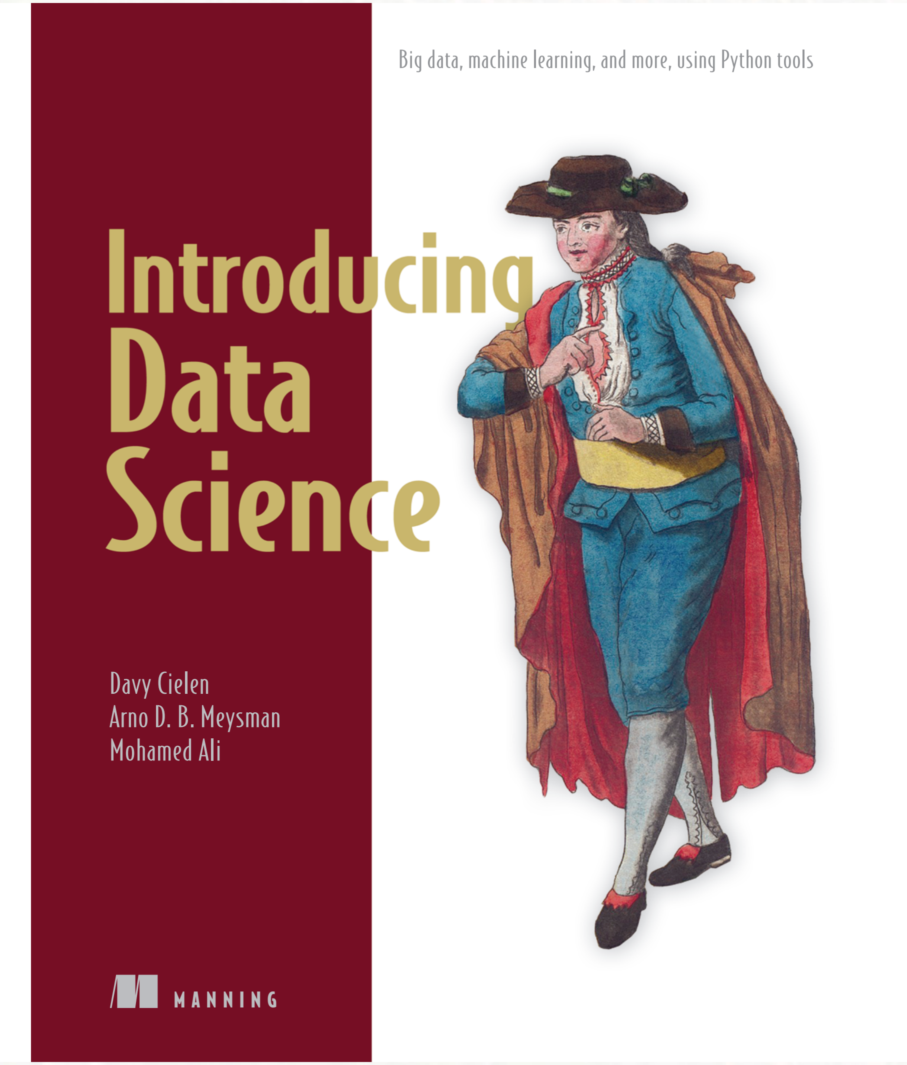


O'REILLY

Doing Data Science

STRAIGHT TALK FROM THE FRONTLINE

Rachel Schutt & Cathy O'Neil



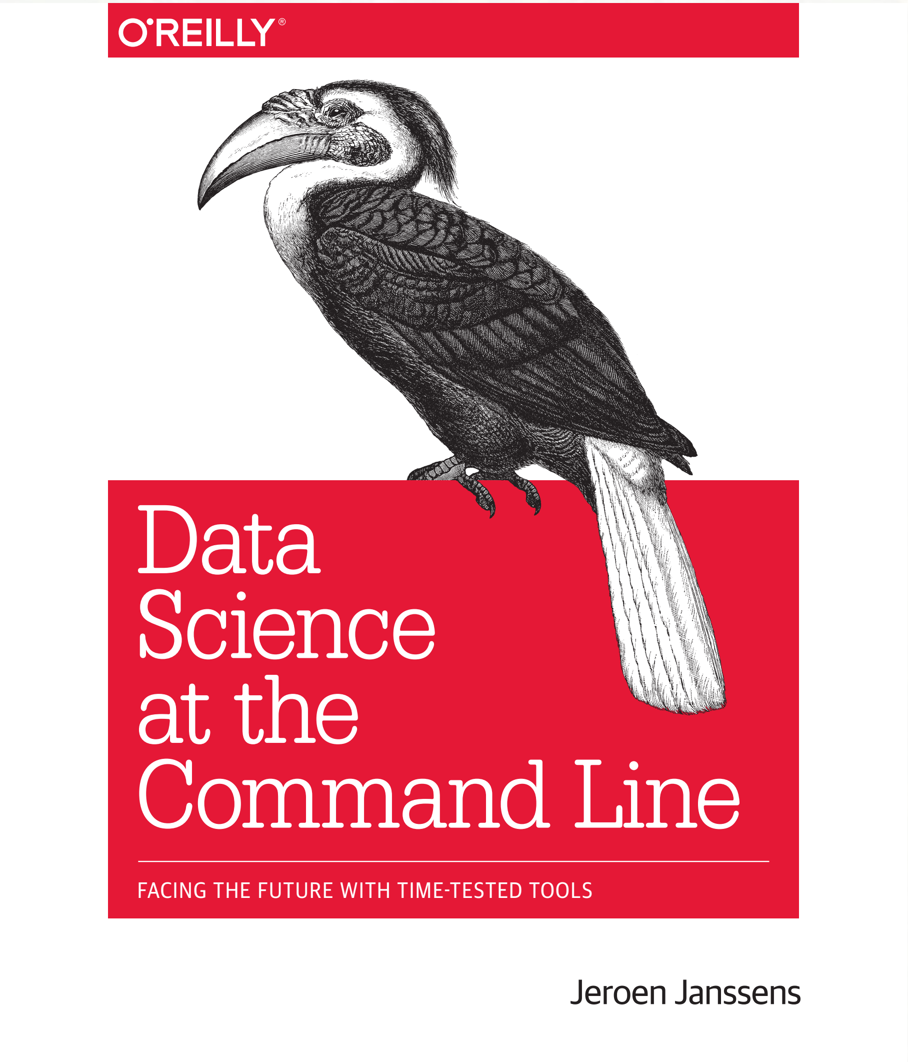
Big data, machine learning, and more, using Python tools

Introducing Data Science

Davy Cielen
Arno D. B. Meysman
Mohamed Ali

MANNING

# Referências

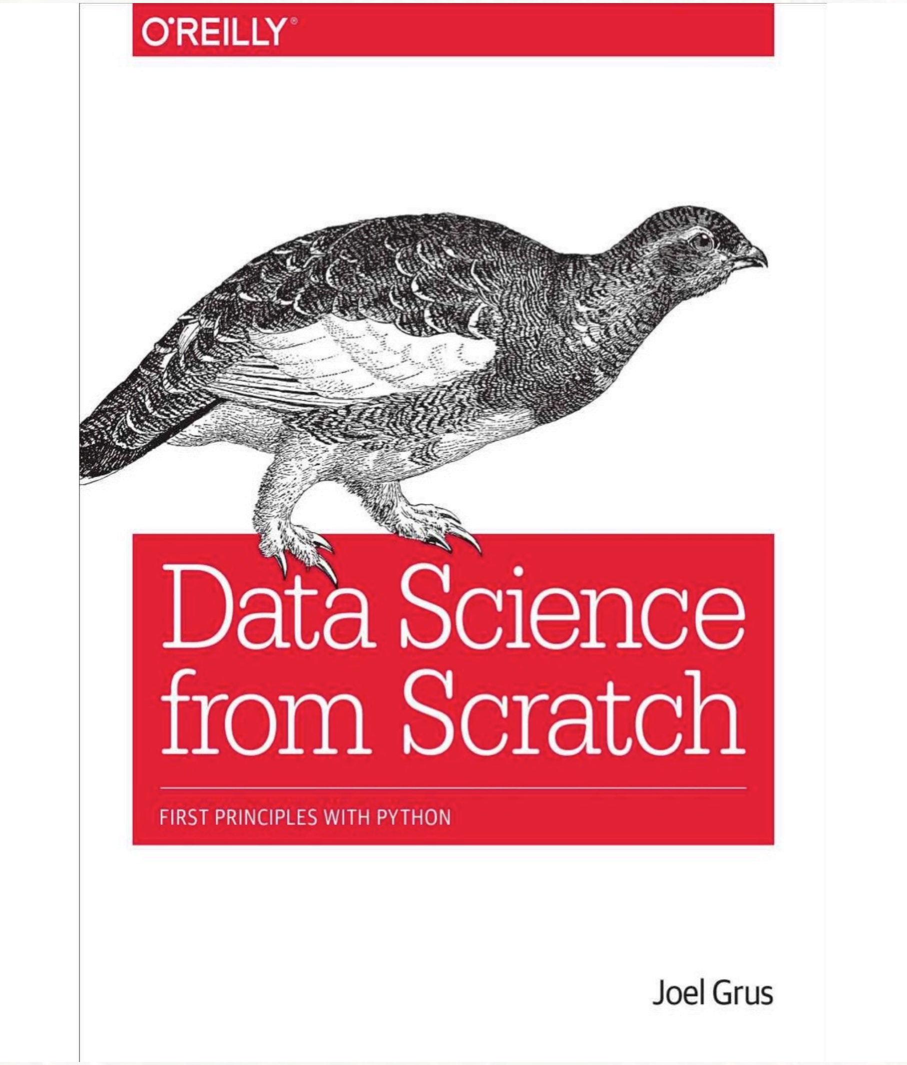


O'REILLY®

Data Science at the Command Line

FACING THE FUTURE WITH TIME-TESTED TOOLS

Jeroen Janssens



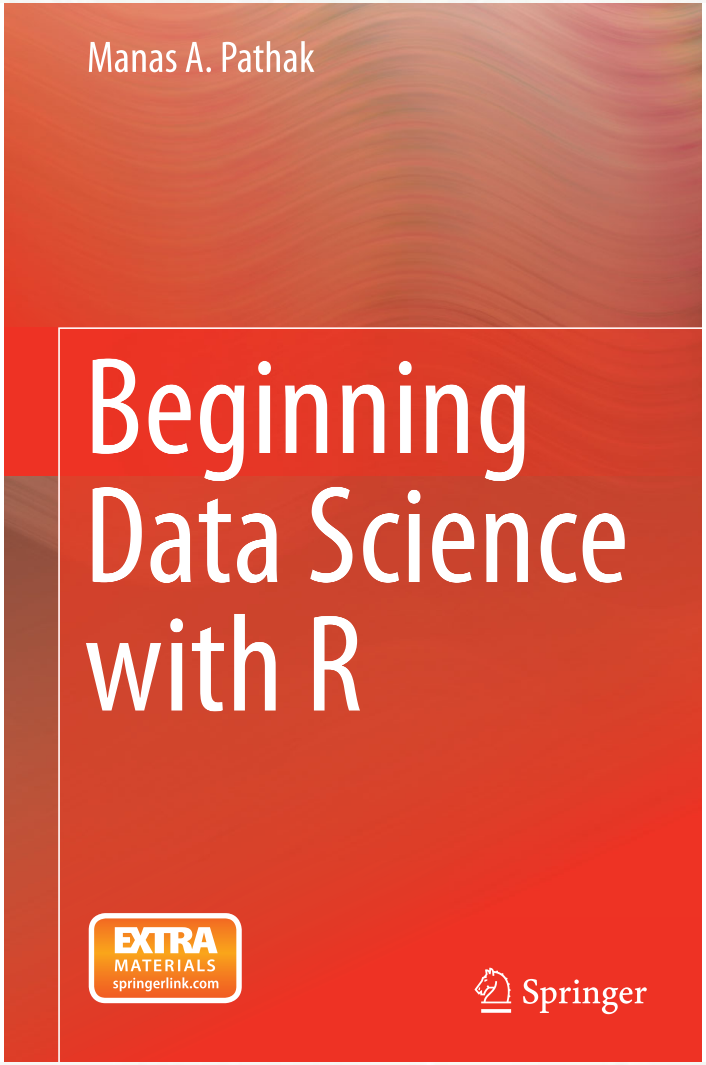
O'REILLY®

Data Science from Scratch

FIRST PRINCIPLES WITH PYTHON

Joel Grus

# Referências

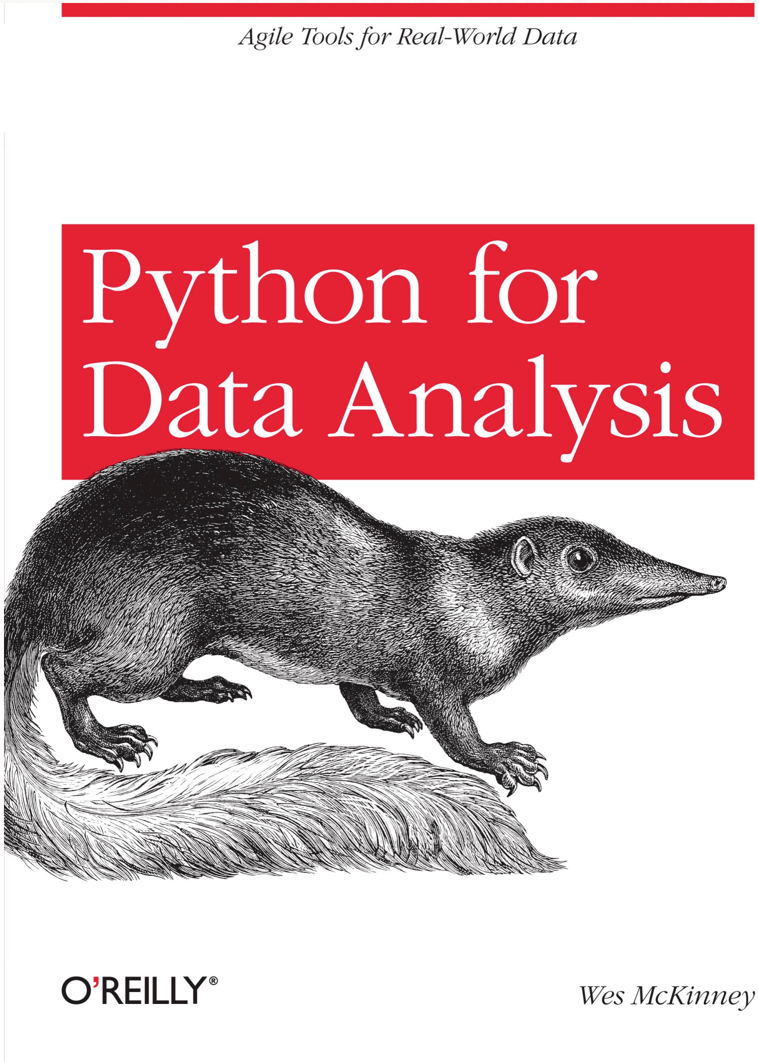


Manas A. Pathak

Beginning Data Science with R

EXTRA MATERIALS springerlink.com

Springer



*Agile Tools for Real-World Data*

Python for Data Analysis

O'REILLY®

*Wes McKinney*

# Shameless Advertising

- Applied Computing Graduate Program at INPE:
  - http://www.inpe.br/pos_graduacao/cursos/cap/
- Introduction to Data Science / Data Mining
  - http://www.lac.inpe.br/~rafael.santos/index.html
  - http://www.lac.inpe.br/~rafael.santos/r.html
- CAP's Annual Workshop:
  - http://www.inpe.br/worcap/
- LABAC's Annual Summer School:
  - http://www.inpe.br/elac2019/

**rafael.santos@inpe.br**