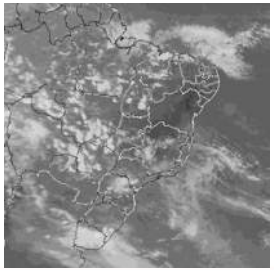


YOU'RE ALREADY A DATA SCIENTIST, NOW GO ASK FOR A RAISE

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

You're already a Data Scientist, now go ask for a raise

About this Talk

About this Talk

- What is Data Science?
- How is it defined by different parties?
- What do I need to know to be a Data Scientist?

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?

You're already a Data Scientist, now go ask for a raise

The Hype

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

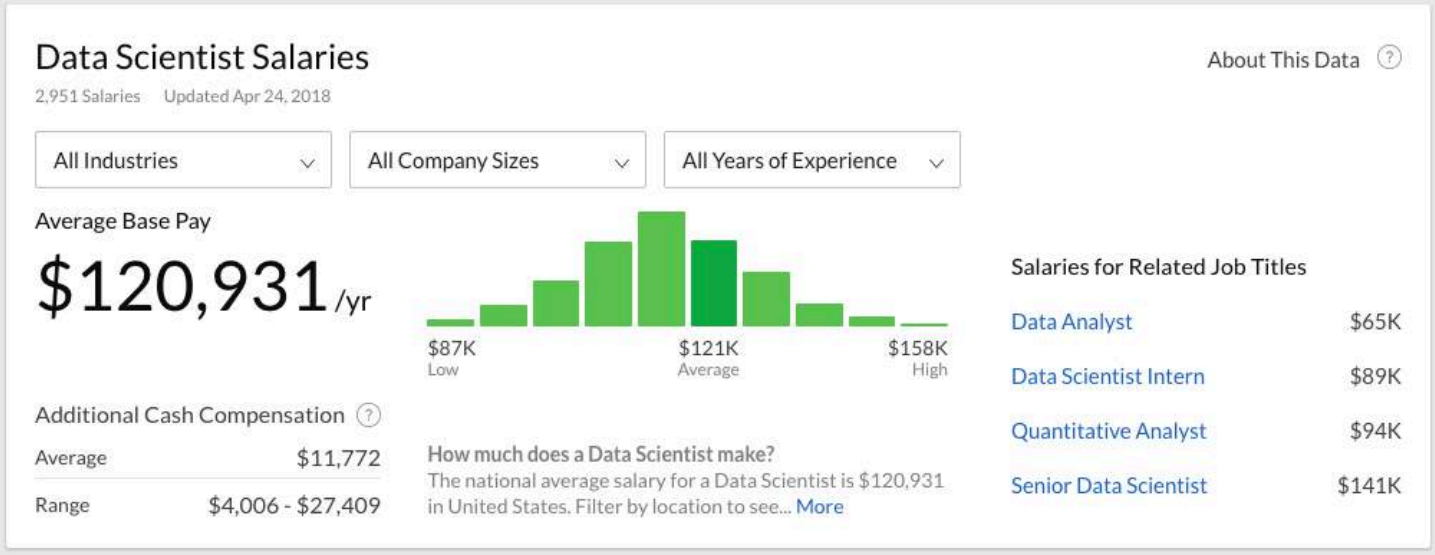
EMC Contributor Brand Contributor
EMC BRANDVOICE

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

Hype

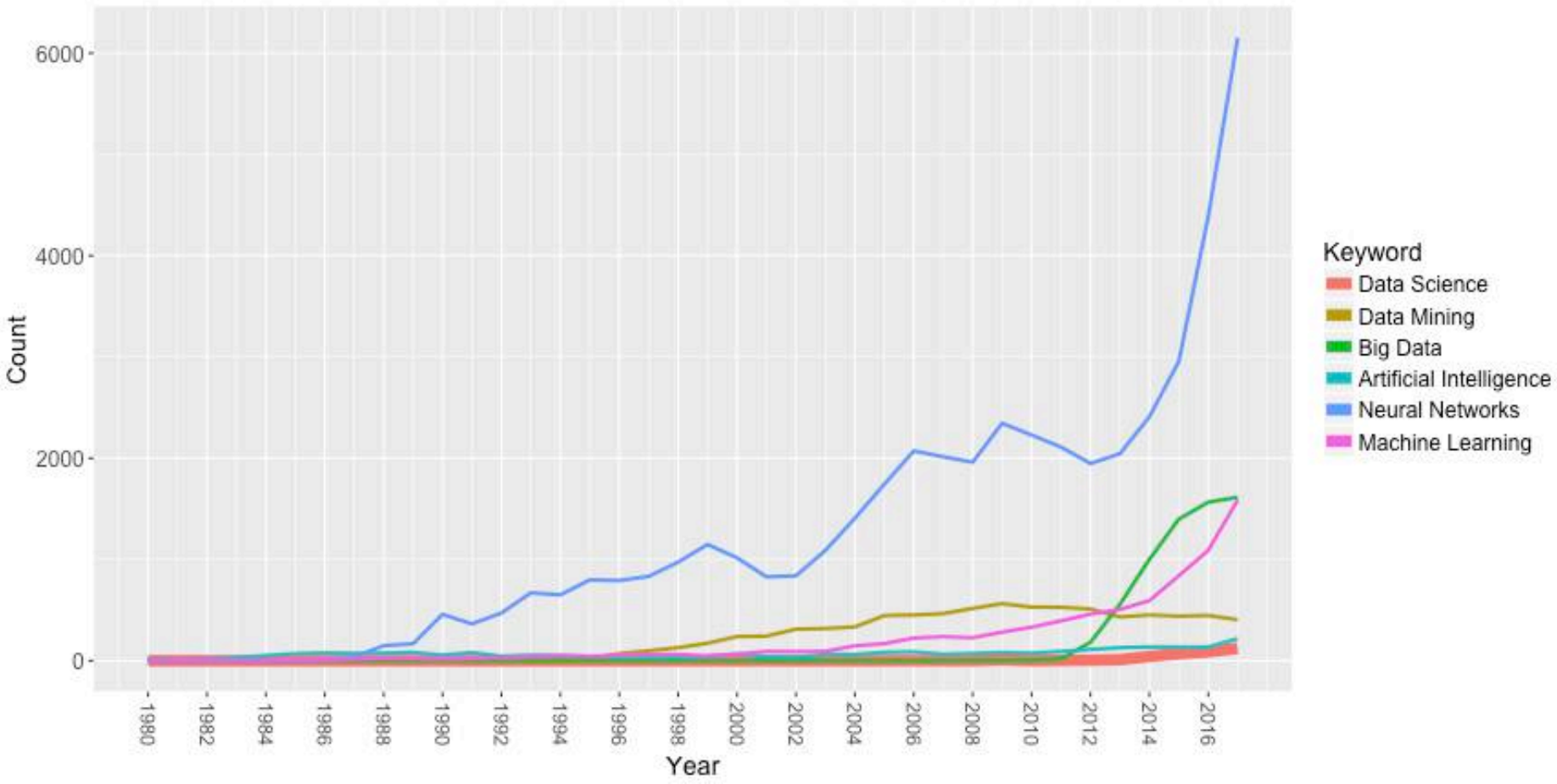
By 2018, the United States will experience a shortage of 190,000 skilled data scientists, and 1.5 million managers and analysts capable of reaping actionable insights from the big data deluge.

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

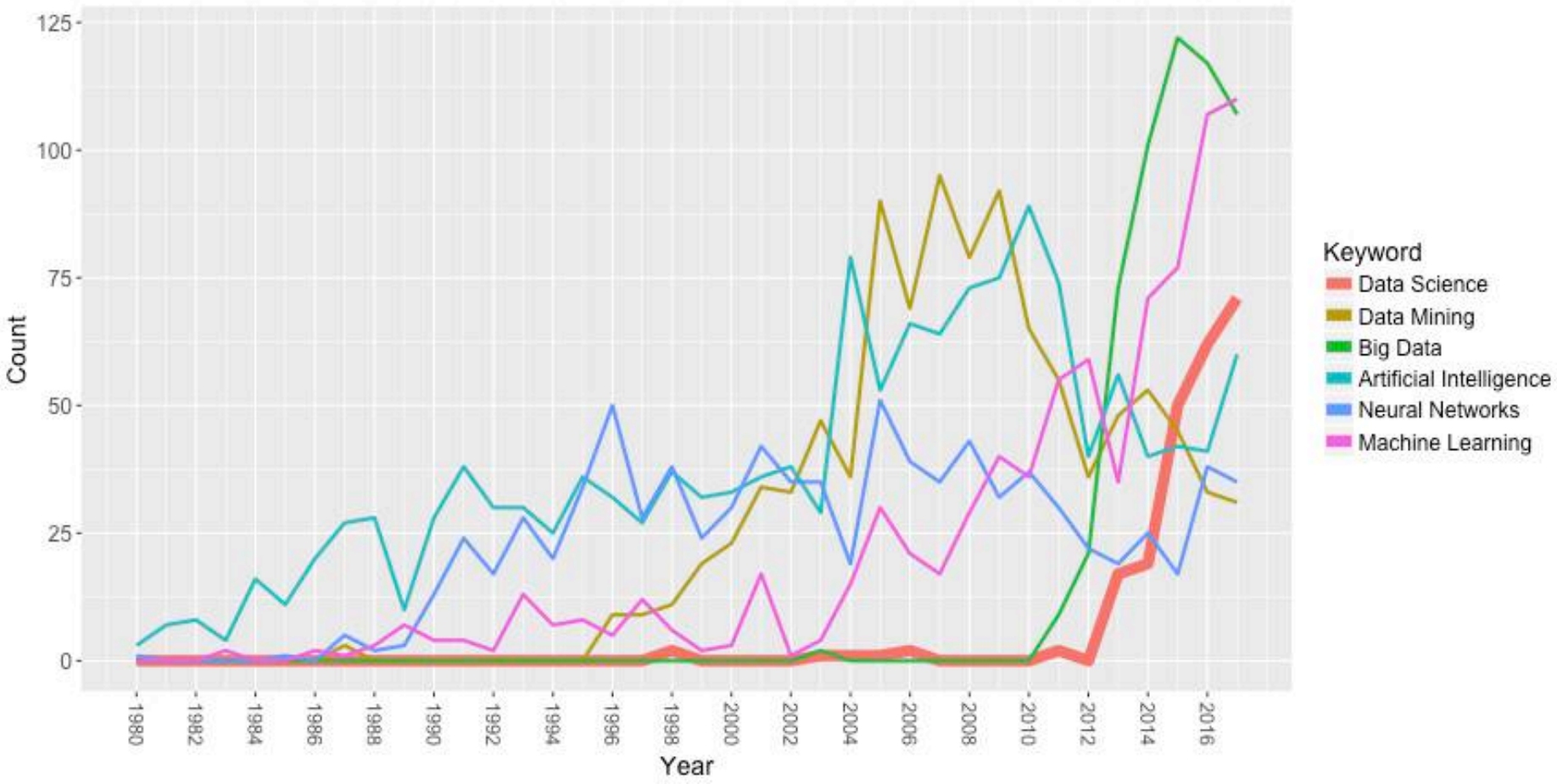


glassdoor.com

Hype


6000
4000
2000
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


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

Hype?

- Will the Real Data Scientists Please Stand Up?

https://www.kdnuggets.com/2015/05/data-science-machine-learning-scientist-definition-jargon.html

- Why most data scientists are frauds, according to a data scientist

https://thenextweb.com/syndication/2017/12/28/data-scientists-frauds-according-data-scientist/

- The problem is the definition of Data Science and the role of the Data Scientist.

You're already a Data Scientist, now go ask for a raise

What is Data Science?

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.

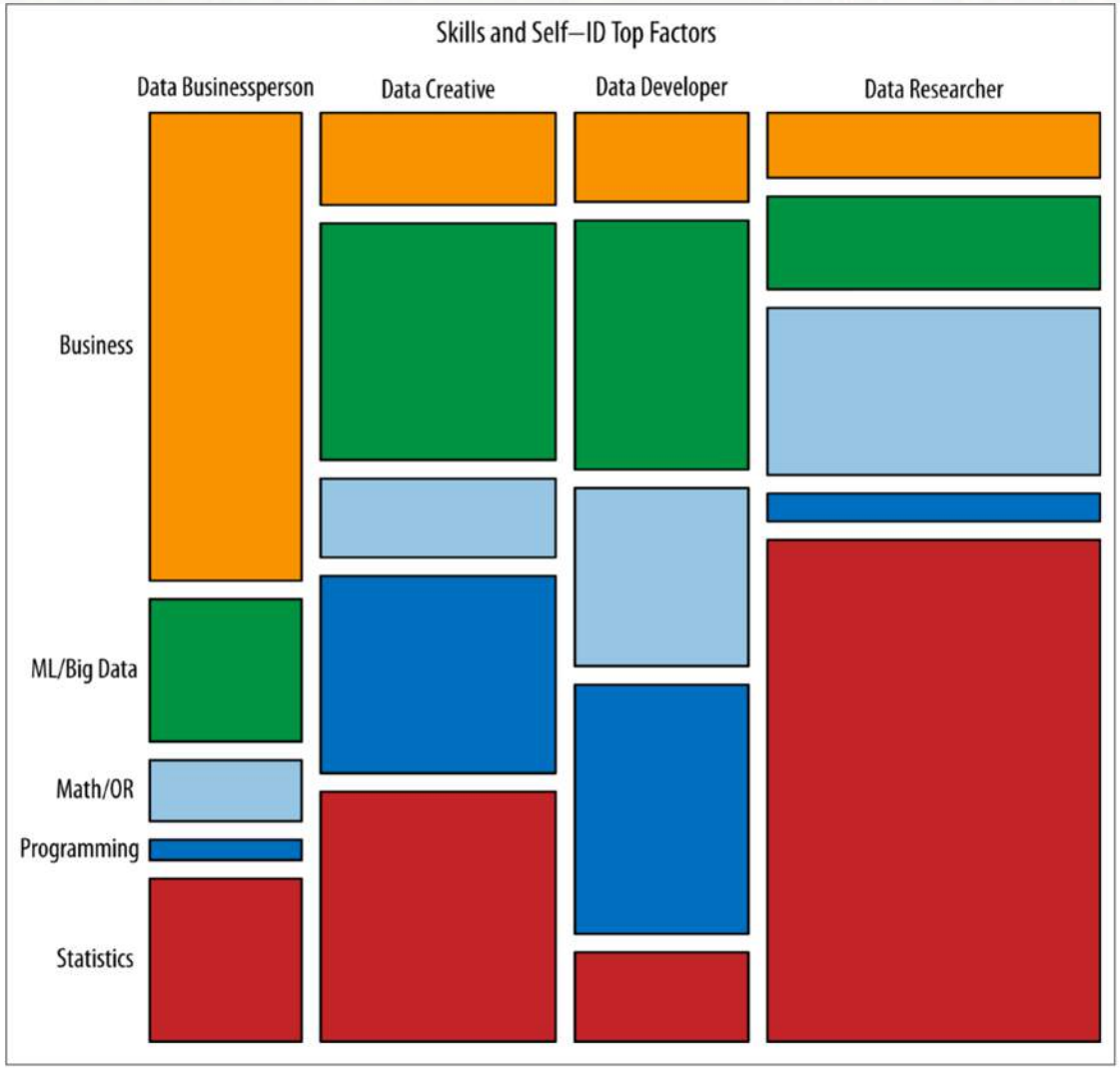
http://bigdata-madesimple.com/what-is-a-data-scientist-14-definitions-of-a-data-scientist/

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.

Harris, Harlan, 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?

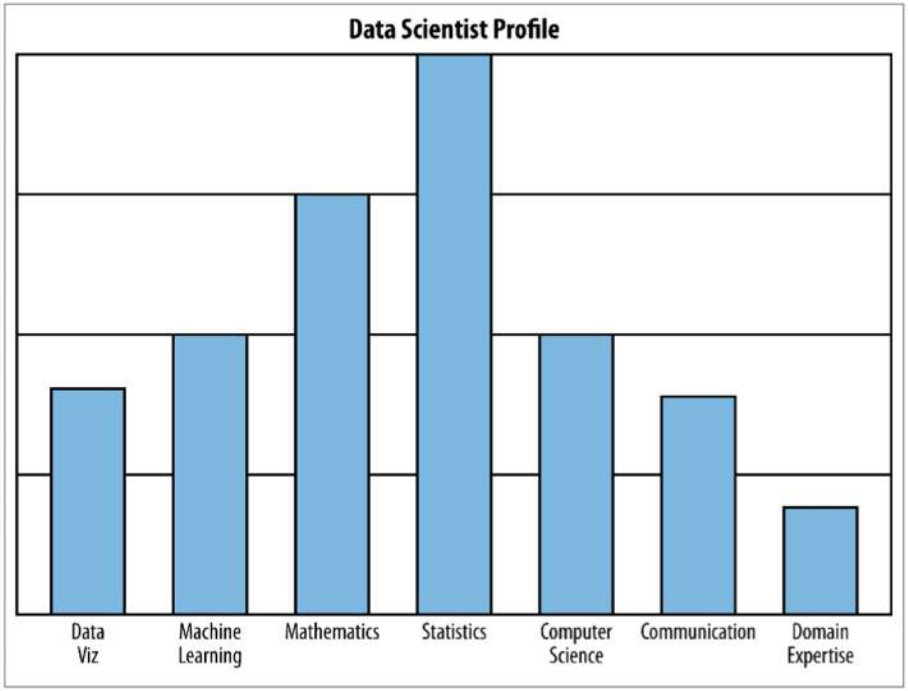


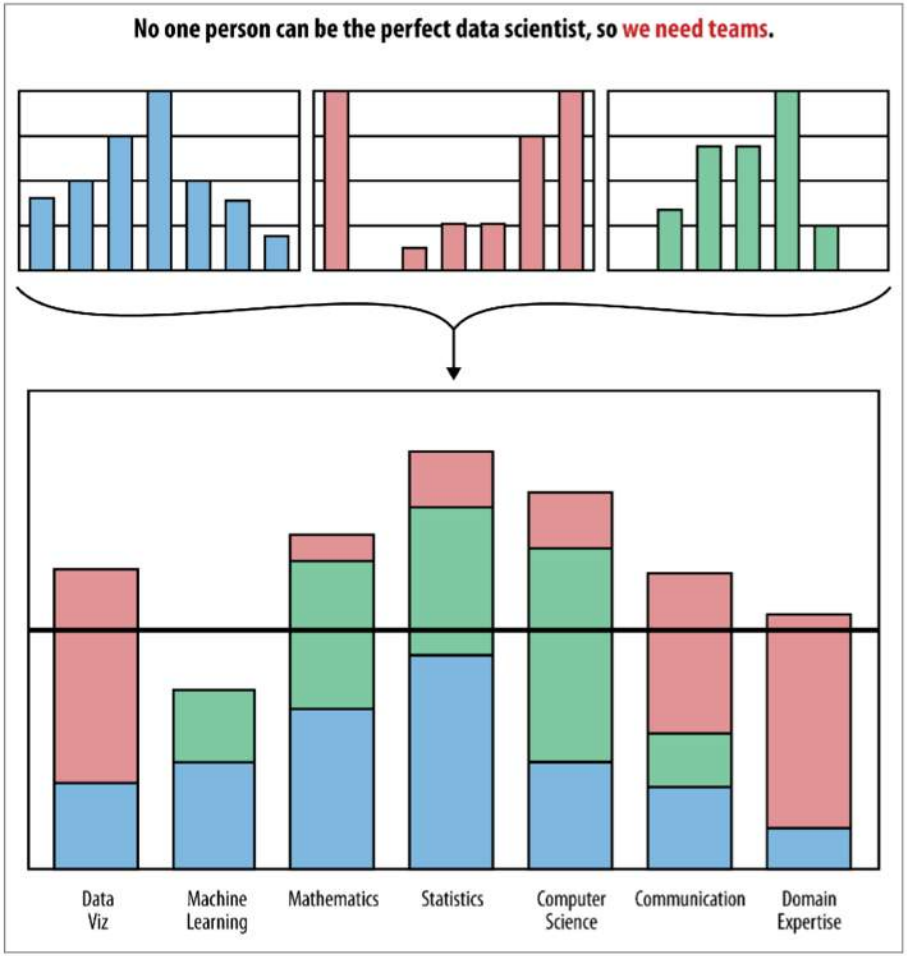
Who are the Data Scientists?

- *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

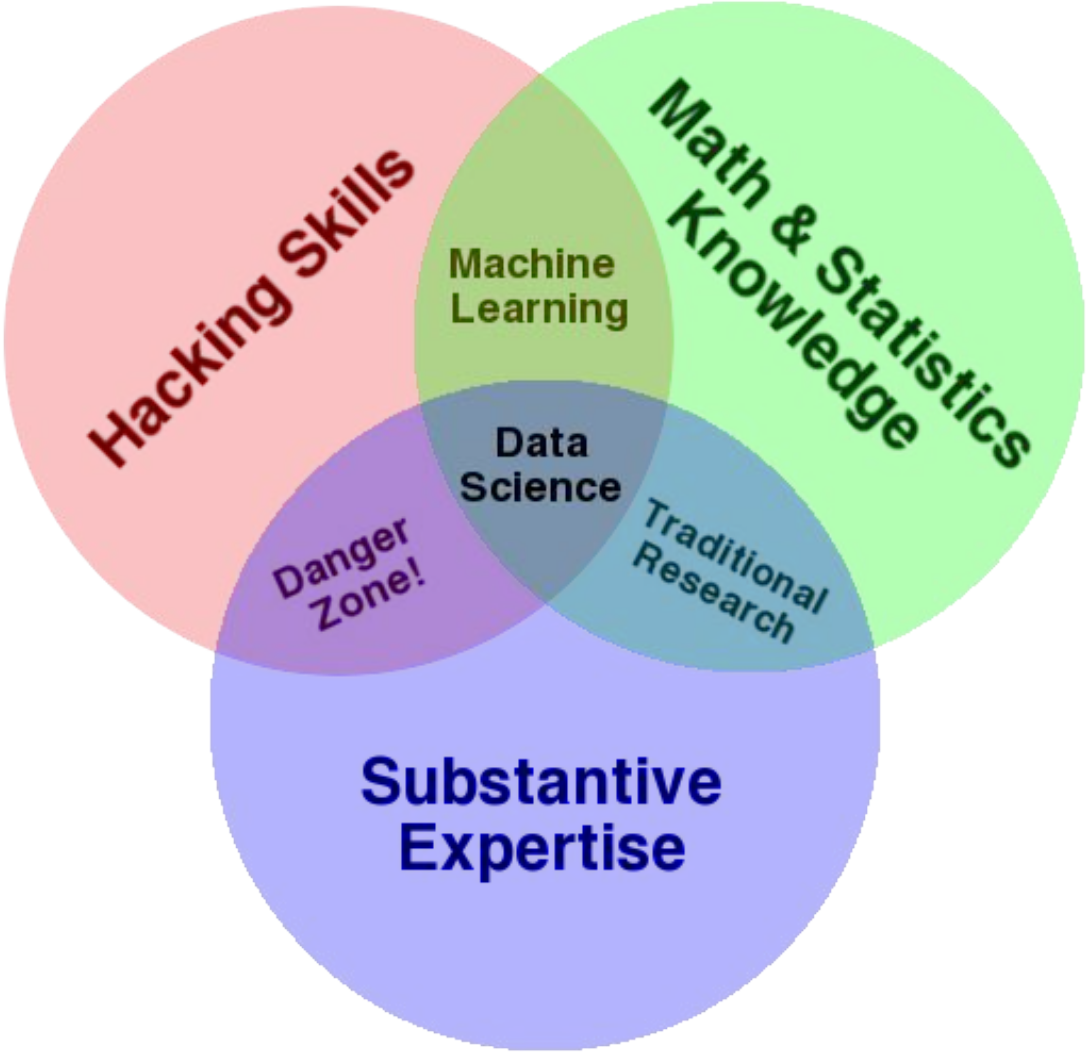
T-Shaped Data Scientist





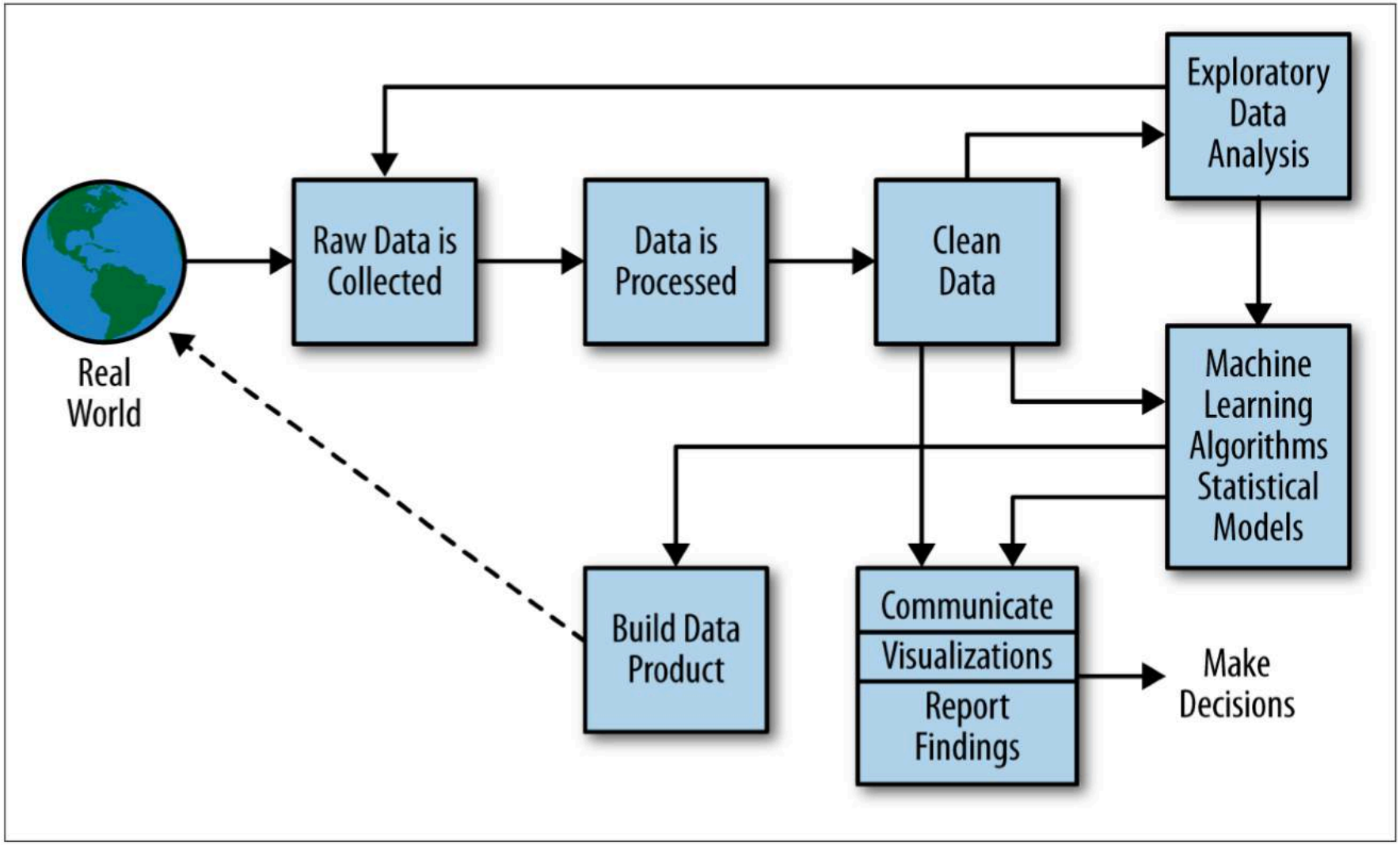
Doing Data Science, Rachel Schutt and Cathy O'Neil, OReilly, 2014

Data Science Venn Diagram



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

Data Science Process



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

For our purposes…

- *…an academic data scientist is a scientist, trained in anything from social science to biology, who works with large amounts of data, and must grapple with computational problems posed by the structure, size, messiness, and the complexity and nature of the data, while simultaneously solving a real-world problem.*

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

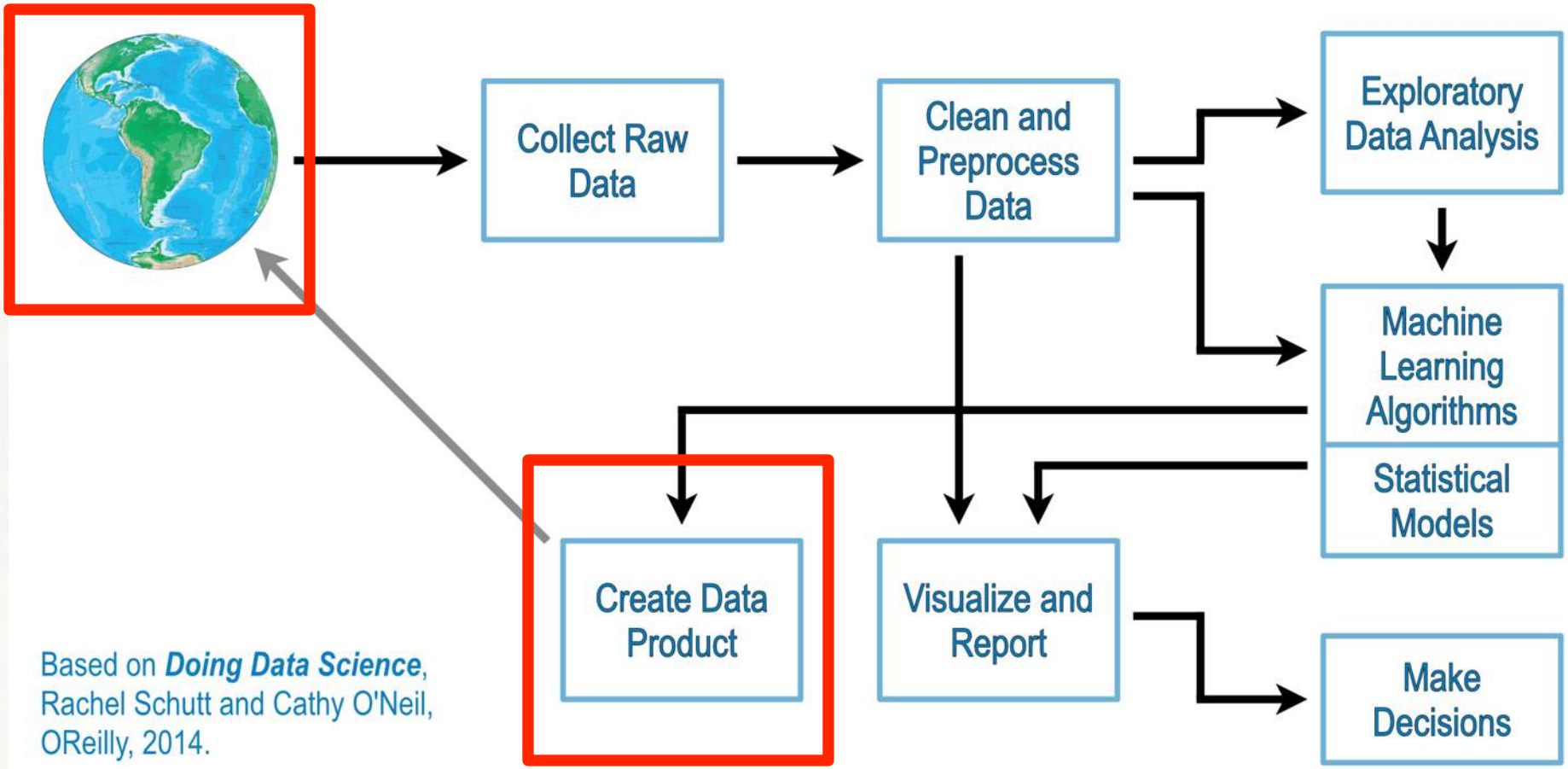
You're already a Data Scientist, now go ask for a raise

So you want to be a Data Scientist…

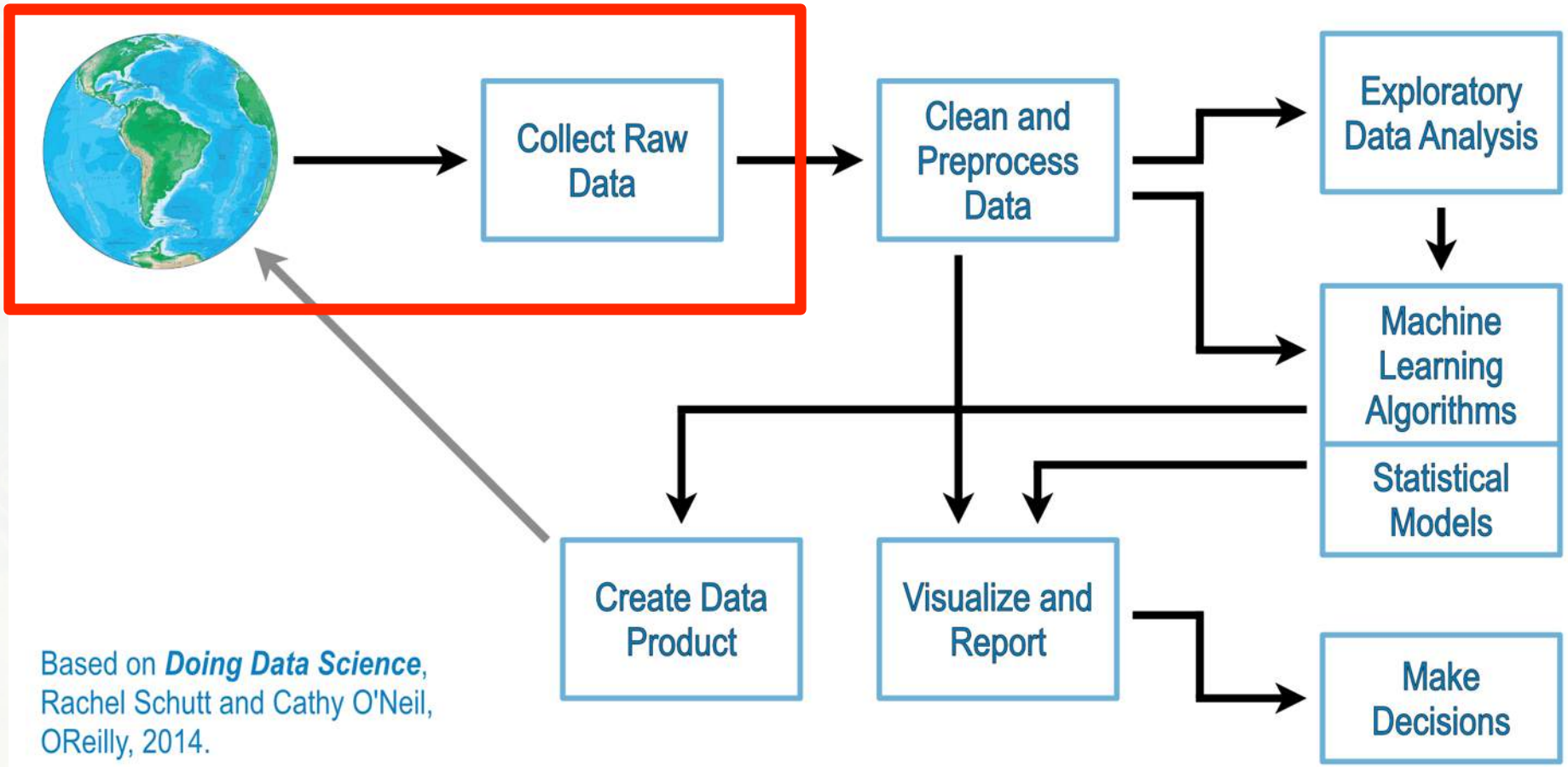
Skills

- List of useful 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: some skills are interchangeable, try to be a data science polyglot (within reasonable limits).
  - …individually impossible: *"Rockstar Programmer"*, *"Rockstar SysAdmin"*, *"Rockstar Analyst"?*
  - …not all technical: we will deal with real world problems, must talk to real world people (**scientists?**).

Skill: Understand the Problem



Skill: Find, Collect, Organize Data



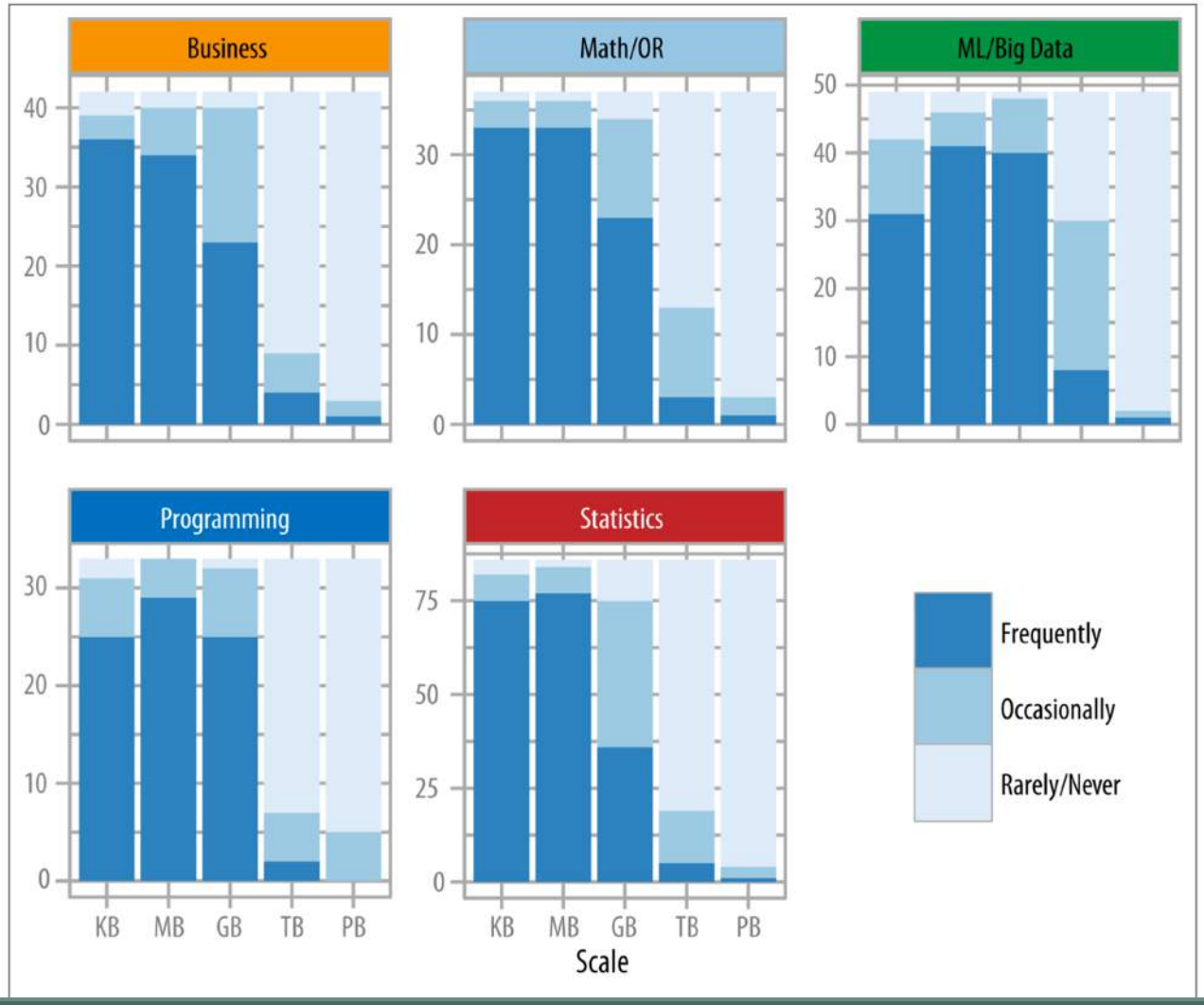
Based on ***Doing Data Science***, Rachel Schutt and Cathy O'Neil, OReilly, 2014.

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.

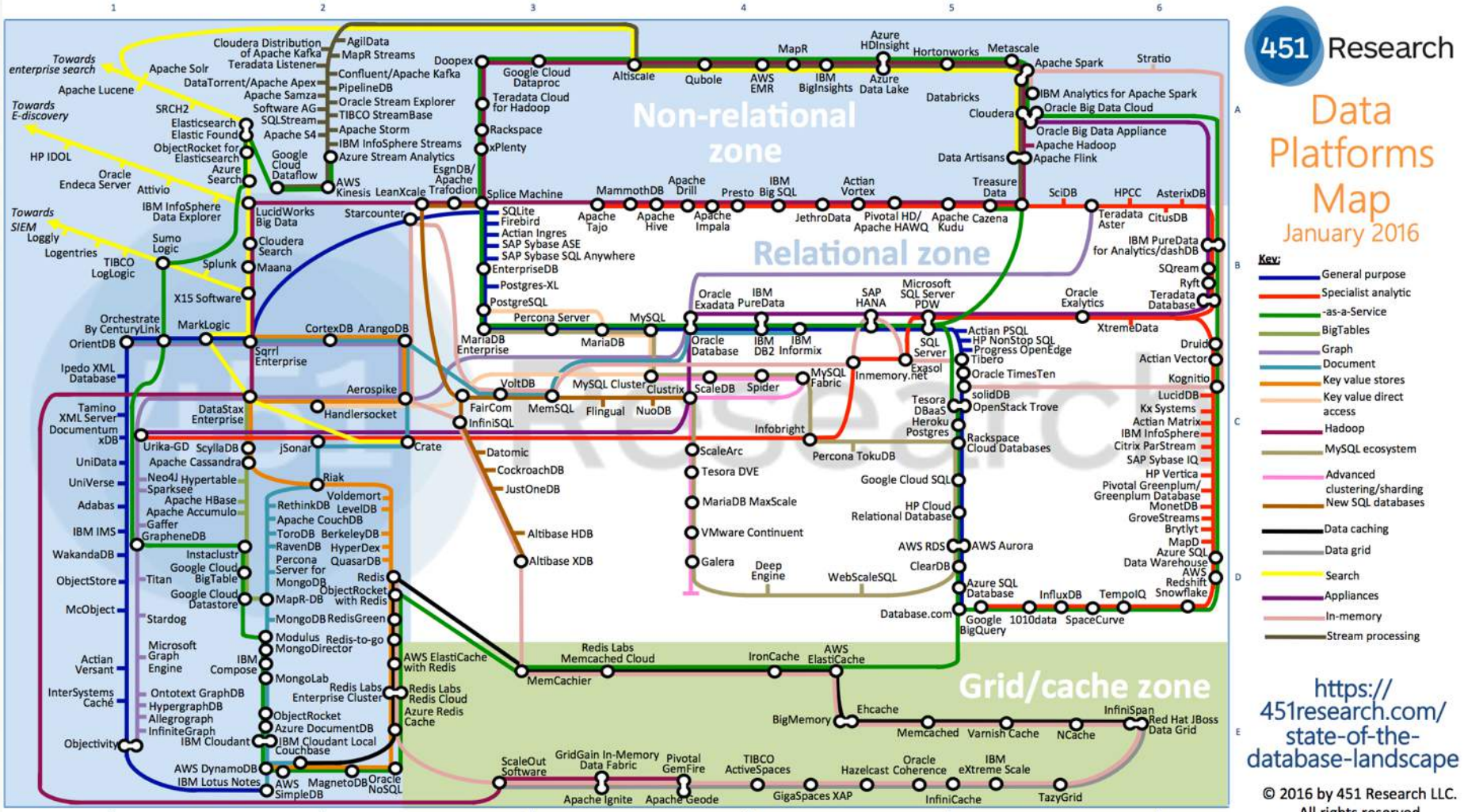
Big Data Lessons from the Climate Science Community, Seth McGinnis, 2016

Back to Skills: Big Data

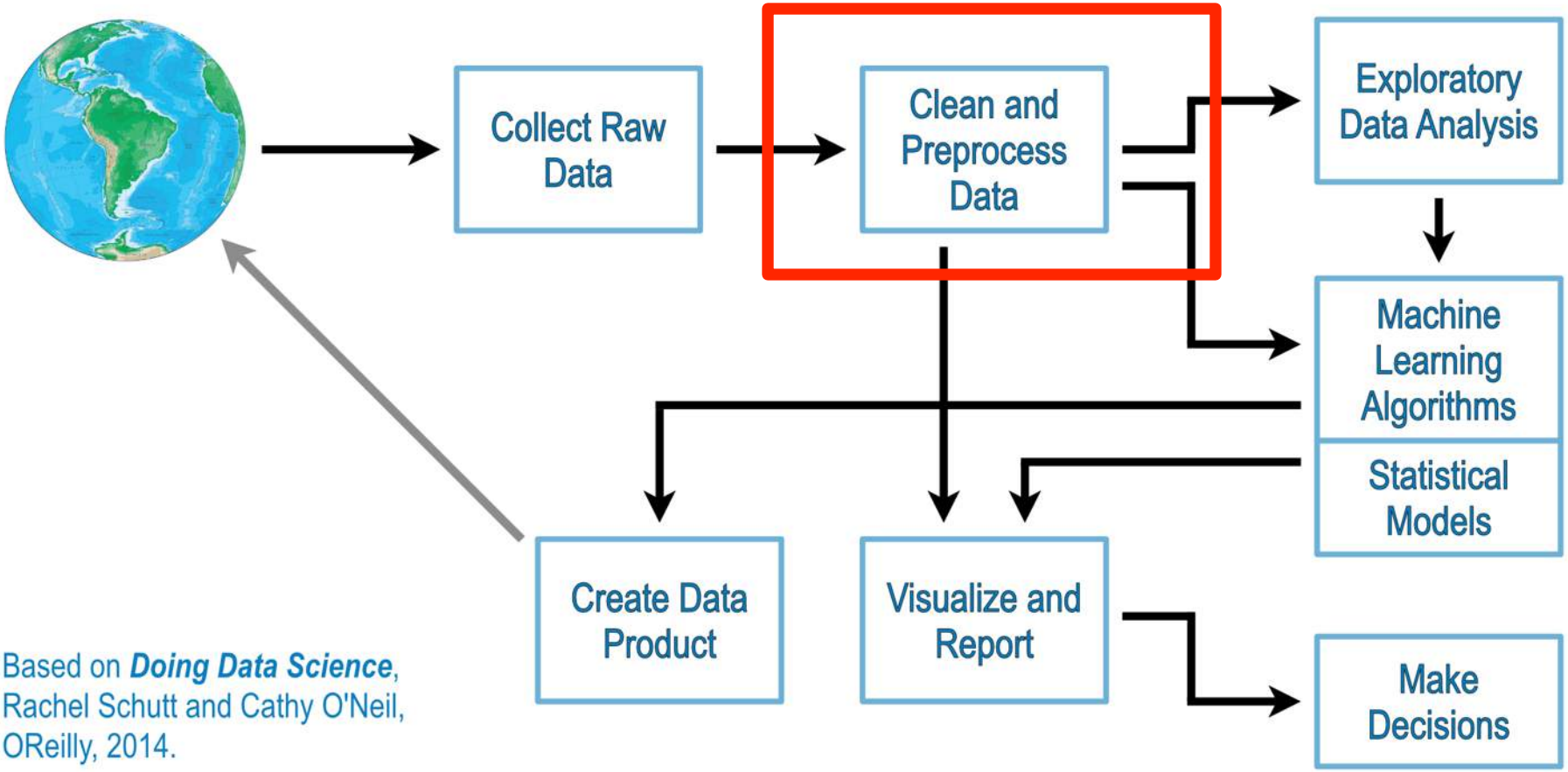


Harris, Harlan, Sean Murphy, and Marck Vaisman. Analyzing the Analyzers: An Introspective Survey of Data Scientists and Their Work. O'Reilly Media, Inc., 2013.

NoSQL


Towards enterprise search
Towards E-discovery
Towards SIEM
Apache Solr
Apache Lucene
HP IDOL
Oracle Endeca Server
Attivio
IBM InfoSphere Data Explorer
Loggly
Logentries
TIBCO LogTogic
Sumo Logic
Splunk
X15 Software
SRCH2
Elasticsearch
Elastic Found
ObjectRocket for Elasticsearch
Azure Search
Cloudera Distribution of Apache Kafka
Teradata Listener
DataTorrent/Apache Apex
Apache Samza
Software AG
SQLStream
Apache S4
Google Cloud Dataflow
AWS Kinesis
AgilData
MapR Streams
Confluent/Apache Kafka
PipelineDB
Oracle Stream Explorer
TIBCO StreamBase
Apache Storm
IBM InfoSphere Streams
Azure Stream Analytics
LeanXcale
EsgnDB/ Apache Trafodion
Doopex
Google Cloud Dataproc
Teradata Cloud for Hadoop
Rackspace
xPlenty
Splice Machine
Altiscale
Qubole
AWS EMR
MapR
IBM BigInsights
Azure HDInsight
Azure Data Lake
Hortonworks
Metascale
Databricks
Data Artisans
Apache Spark
Stratio
IBM Analytics for Apache Spark
Oracle Big Data Cloud
Cloudera
Oracle Big Data Appliance
Apache Hadoop
Apache Flink
Non-relational zone
Relational zone
Grid/cache zone
LucidWorks Big Data
Cloudera Search
Maana
Starcounter
SQLite
Firebird
Actian Ingres
SAP Sybase ASE
SAP Sybase SQL Anywhere
EnterpriseDB
Postgres-XL
PostgreSQL
Percona Server
MariaDB
MariaDB Enterprise
Apache Tajo
Apache Hive
Apache Impala
MammothDB
Apache Drill
Presto
IBM Big SQL
JethroData
Actian Vortex
Pivotal HD/ Apache HAWQ
Apache Kudu
Apache Cazena
Treasure Data
SciDB
HPCC
AsterixDB
Teradata Aster
CitusDB
IBM PureData for Analytics/dashDB
SQream
Ryft
Teradata Database
Oracle Exalytics
XtremeData
Oracle Exadata
IBM PureData
Oracle Database
IBM DB2
IBM Informix
SAP HANA
Microsoft SQL Server PDW
SQL Server
Exasol
Actian PSQL
HP NonStop SQL
Progress OpenEdge
Tibero
Oracle TimesTen
solidDB
OpenStack Trove
Tesora DBaaS
Heroku Postgres
Rackspace Cloud Databases
Inmemory.net
MySQL Fabric
ScaleDB
Spider
Clustrix
MySQL Cluster
VoltDB
FairCom
InfiniSQL
MemSQL
Flingual
NuoDB
Infobright
Percona TokuDB
Google Cloud SQL
HP Cloud Relational Database
AWS RDS
AWS Aurora
ClearDB
WebScaleSQL
Deep Engine
Azure SQL Database
Database.com
Google BigQuery
1010data
InfluxDB
SpaceCurve
TempoIQ
AWS Redshift
Snowflake
Druid
Actian Vector
Kognitio
LucidDB
Kx Systems
Actian Matrix
IBM InfoSphere
Citrix ParStream
SAP Sybase IQ
HP Vertica
Pivotal Greenplum/ Greenplum Database
MonetDB
GroveStreams
Brytlyt
MapD
Azure SQL Data Warehouse
Orchestrate
By CenturyLink
OrientDB
MarkLogic
Sqrrl Enterprise
CortexDB
ArangoDB
Aerospike
Handlersocket
DataStax Enterprise
Tamino XML Server
Documentum xDB
UniData
UniVerse
Adabas
IBM IMS
WakandaDB
ObjectStore
McObject
Actian Versant
InterSystems Caché
Objectivity
Ipedo XML Database
Urika-GD
ScyllaDB
Apache Cassandra
Neo4J
Hypertable
Sparksee
Apache HBase
Apache Accumulo
Gaffer
GrapheneDB
Instaclustr
Google Cloud BigTable
Google Cloud Datastore
Stardog
Microsoft Graph Engine
IBM Compose
Ontotext GraphDB
HypergraphDB
Allegrograph
InfiniteGraph
IBM Cloudant
jSonar
Riak
RethinkDB
LevelDB
Apache CouchDB
ToroDB
BerkeleyDB
RavenDB
HyperDex
Percona Server for MongoDB
QuasarDB
Redis
ObjectRocket with Redis
MapR-DB
MongoDB
RedisGreen
Modulus
Redis-to-go
MongoDirector
MongoLab
Redis Labs Enterprise Cluster
ObjectRocket
Azure DocumentDB
IBM Cloudant Local
Couchbase
AWS DynamoDB
IBM Lotus Notes
AWS SimpleDB
MagnetoDB
Oracle NoSQL
Crate
Datomic
CockroachDB
JustOneDB
Altibase HDB
Altibase XDB
ScaleArc
Tesora DVE
MariaDB MaxScale
VMware Continuent
Galera
Redis Labs Memcached Cloud
MemCachier
AWS ElastiCache with Redis
Redis Labs Redis Cloud
Azure Redis Cache
IronCache
AWS ElastiCache
BigMemory
Ehcache
Memcached
Varnish Cache
NCache
InfiniSpan
Red Hat JBoss Data Grid
ScaleOut Software
GridGain In-Memory Data Fabric
Apache Ignite
Pivotal GemFire
Apache Geode
TIBCO ActiveSpaces
GigaSpaces XAP
Oracle Coherence
Hazelcast
IBM eXtreme Scale
InfiniCache
TazyGrid
451 Research
Data Platforms Map
January 2016
Key:
General purpose
Specialist analytic
-as-a-Service
BigTables
Graph
Document
Key value stores
Key value direct access
Hadoop
MySQL ecosystem
Advanced clustering/sharding
New SQL databases
Data caching
Data grid
Search
Appliances
In-memory
Stream processing
https://451research.com/state-of-the-database-landscape
© 2016 by 451 Research LLC. All rights reserved

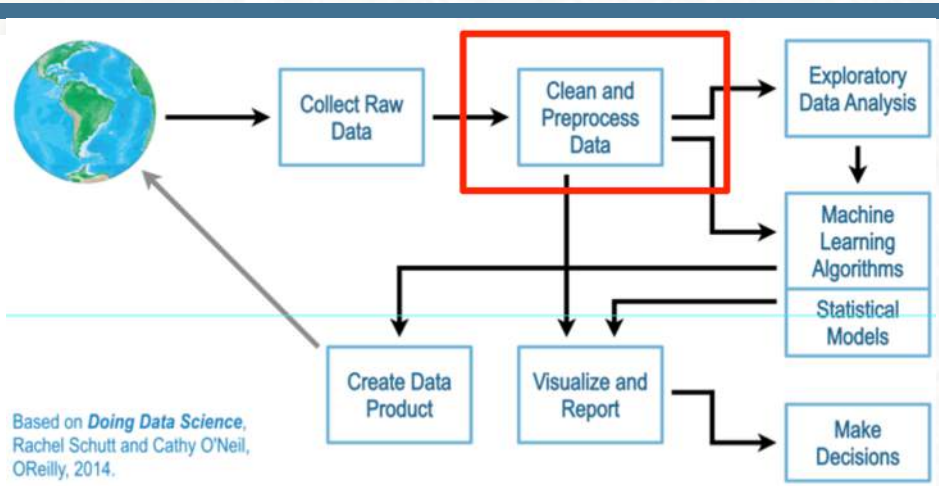
Skill: Hacking



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

Skill: Hacking

- 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

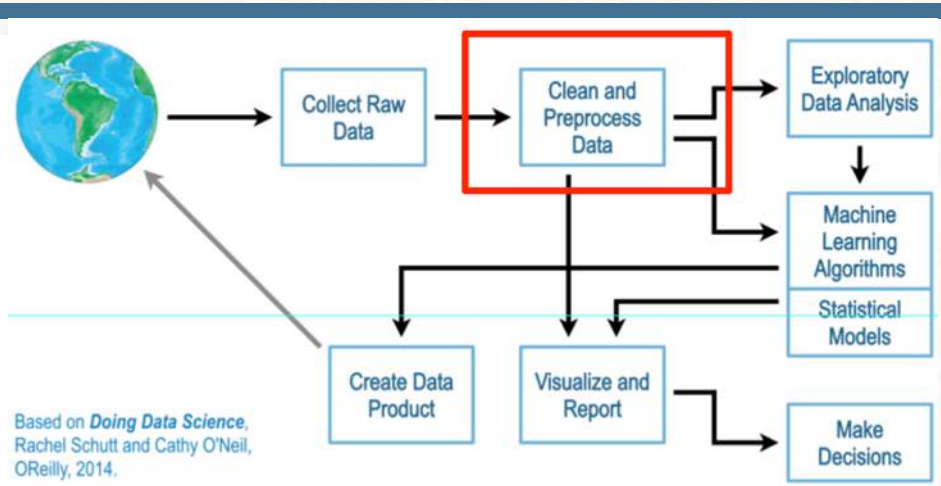


https://www.merriam-webster.com/dictionary/hacker

- 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!



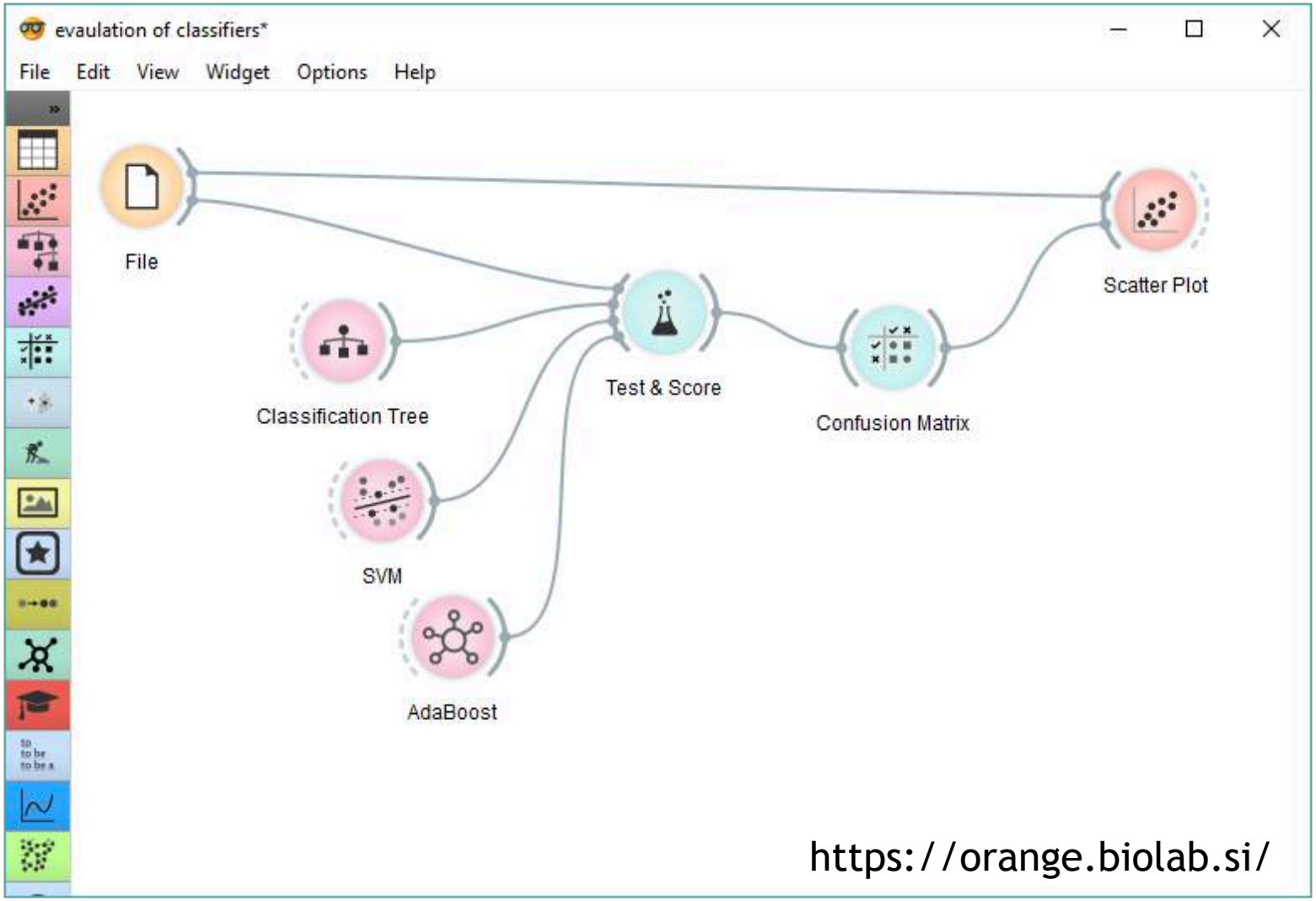
Skill: Hacking

- This:

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



Skill: Hacking Languages: Python

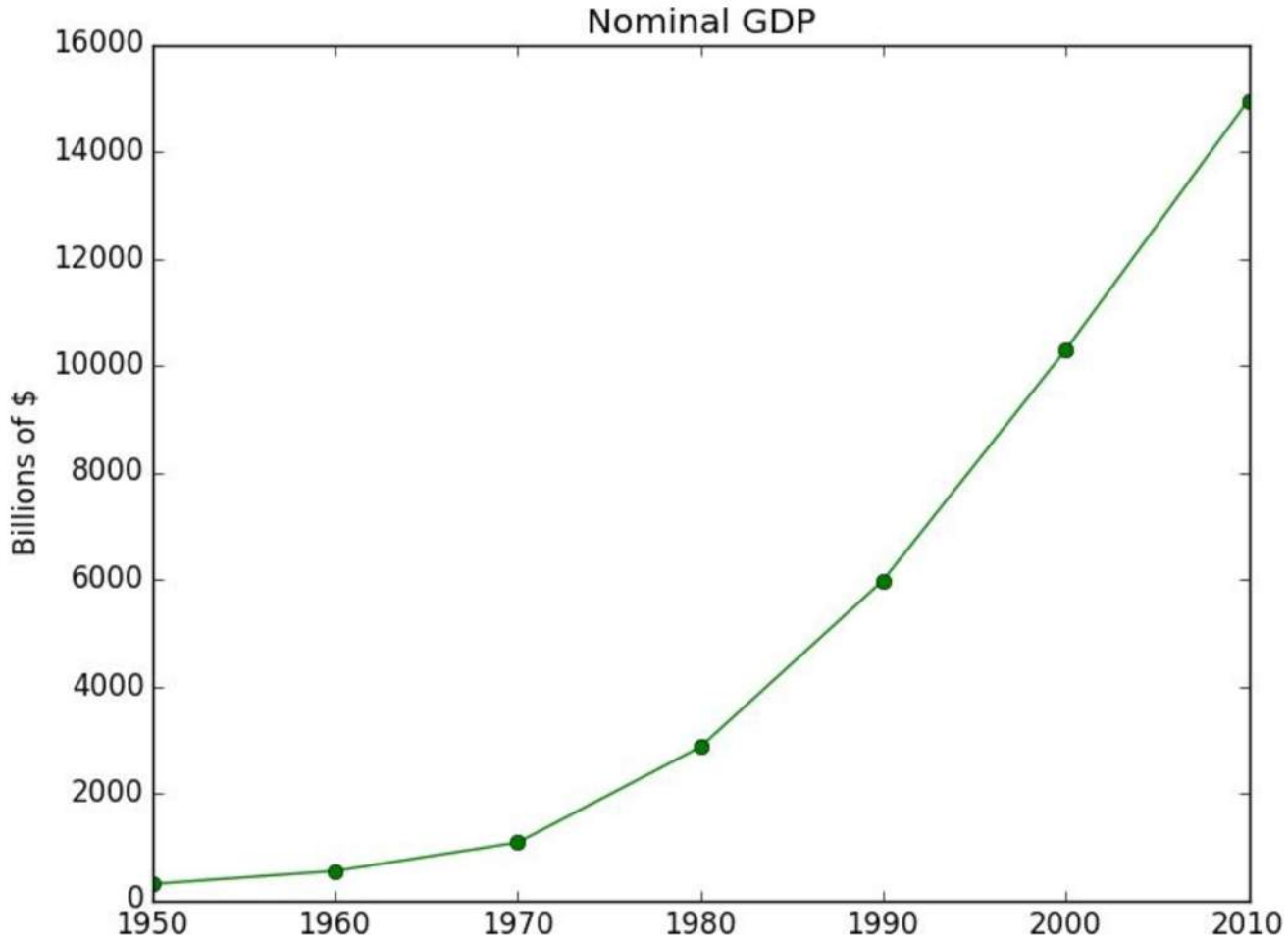
- Pros:
  - General purpose language.
  - Easy to script.
  - Lots of libraries.
- Cons:
  - Two main (sometimes incompatible) versions.
  - Many abandoned libraries.
  - *There should be one – and preferably only one – obvious way to do it.*

Skill: Hacking Languages: Python

```
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

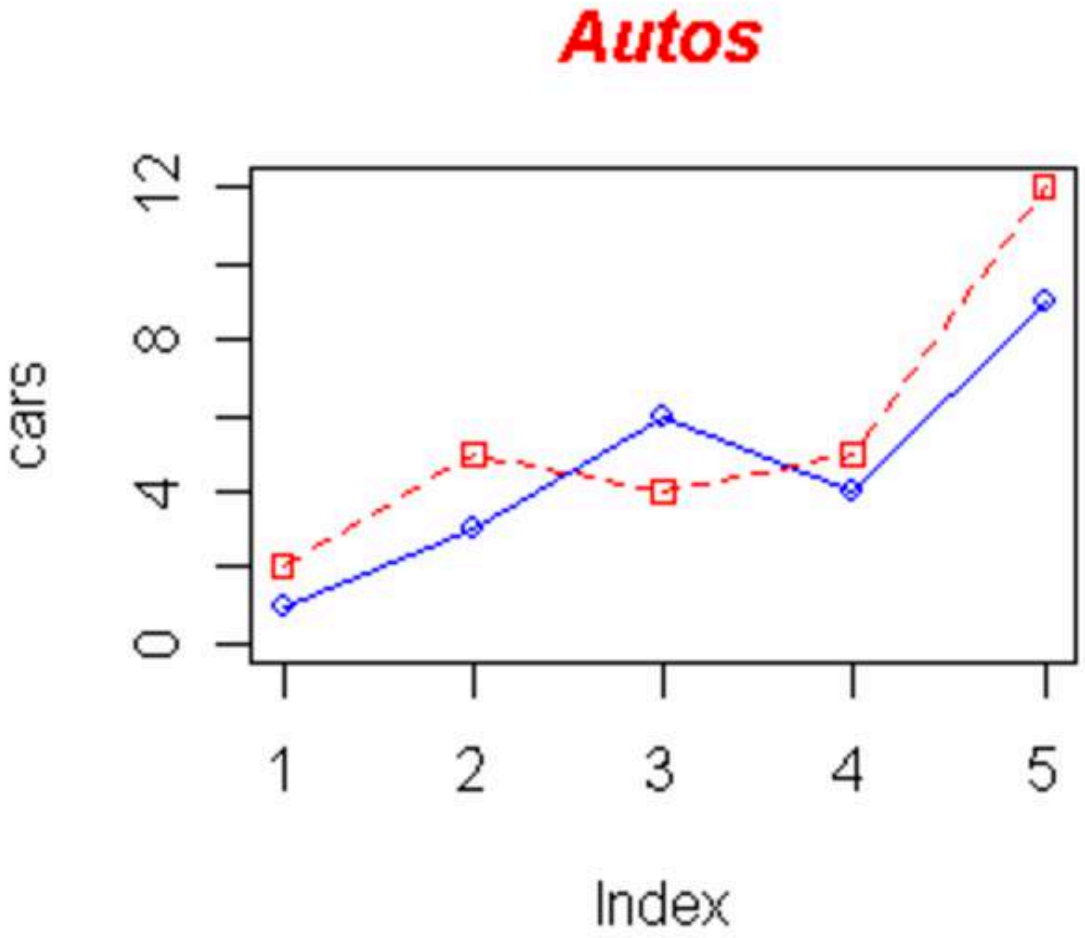
- Pros:
  - Traditionally used by scientists.
  - Strong math/statistics support.
  - Many well-organized packages: CRAN.
- Cons:
  - Steep learning curve.
  - Not everything works out of the box in every system.

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).
  - Right now, some fragmentation.
  - 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();
```

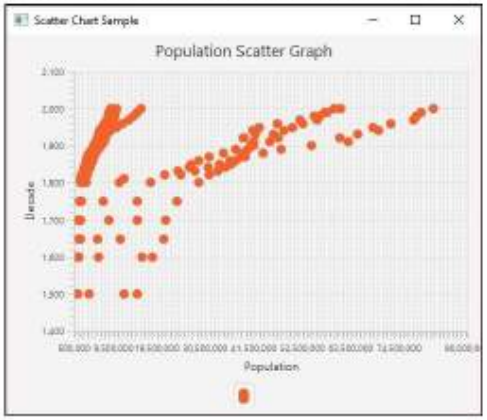
The series is populated using a `CSVReader` class instance and the file `EuropeanScatterData.csv`. This process was discussed in Chapter 3, *Data Cleaning*:

```
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

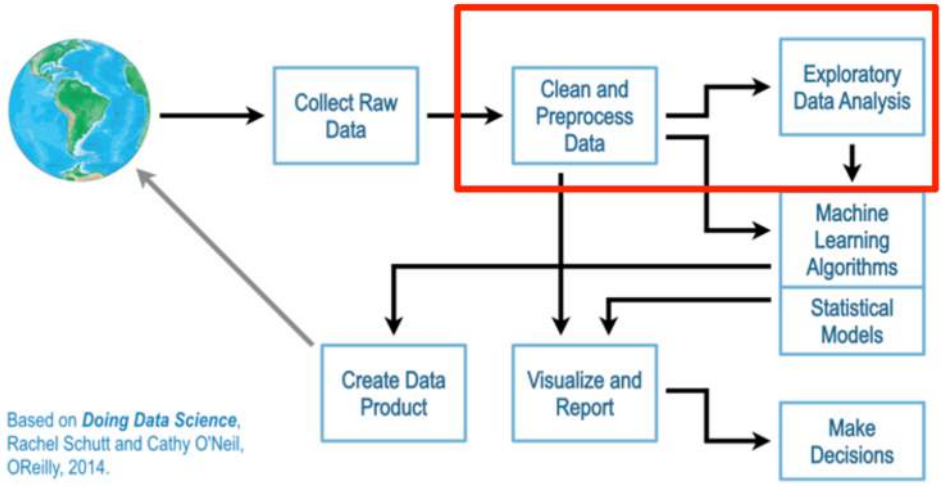
Skill: Hacking Languages: Scala

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

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

Skill: Exploratory Data Analysis

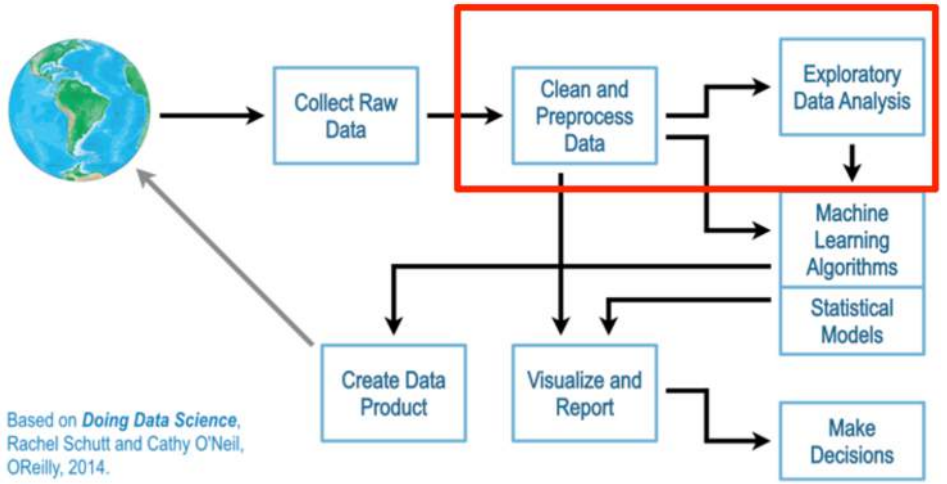
- We have the data. Now what?
  - Do we know what we want to discover?
  - We need basic skills in statistics and data modeling.



- 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

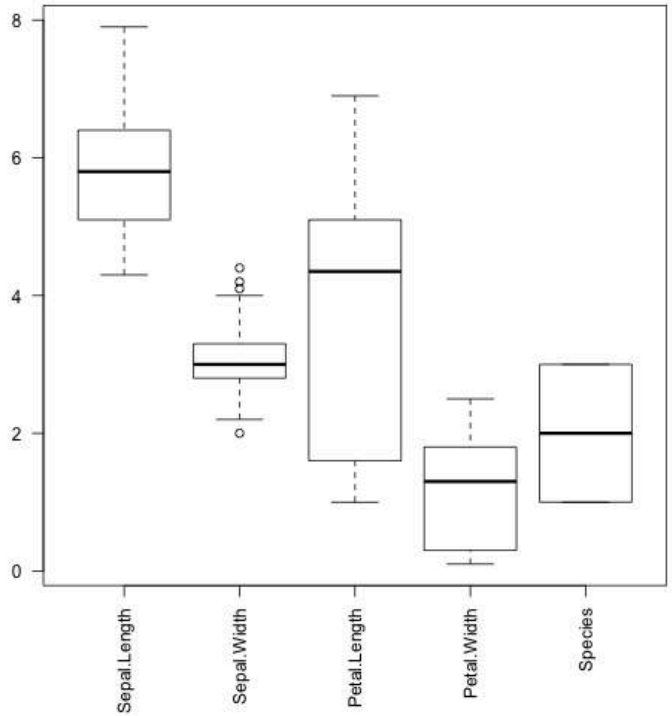
- Basic statistics – avoid complex models (for the time being).
- Basic plots that explore relations between the variables on the data.

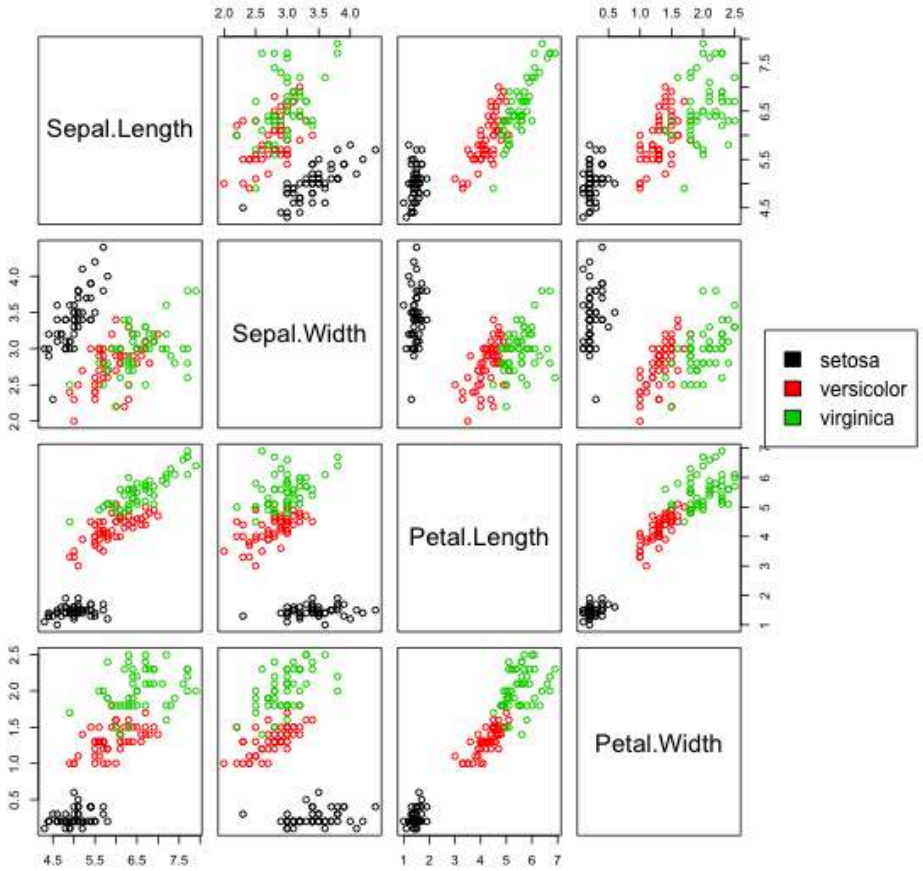


- Used to gain insight on the data and relations, may suggest which advanced analysis (e.g. machine learning) can be applied.

Skill: Exploratory Data Analysis

- Quick example: Iris Dataset.





Skill: Analysis

Collect Raw Data

Clean and Preprocess Data

Exploratory Data Analysis

Machine Learning Algorithms

Statistical Models

Create Data Product

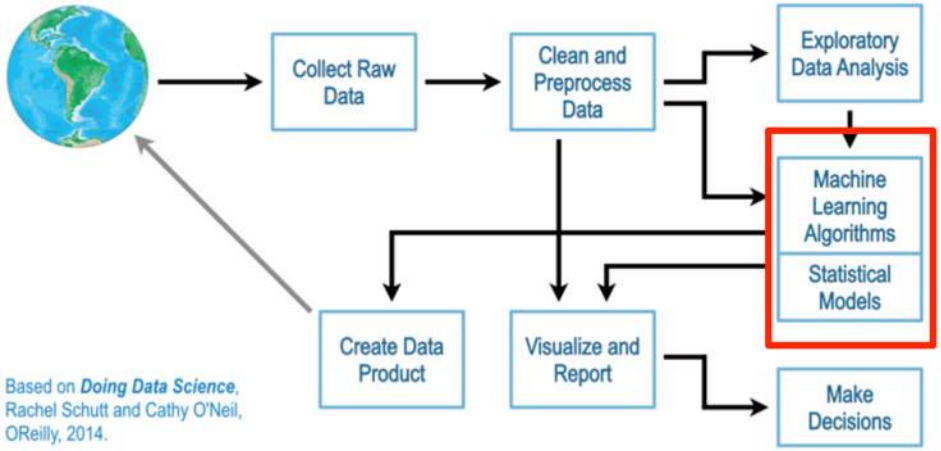
Visualize and Report

Make Decisions

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

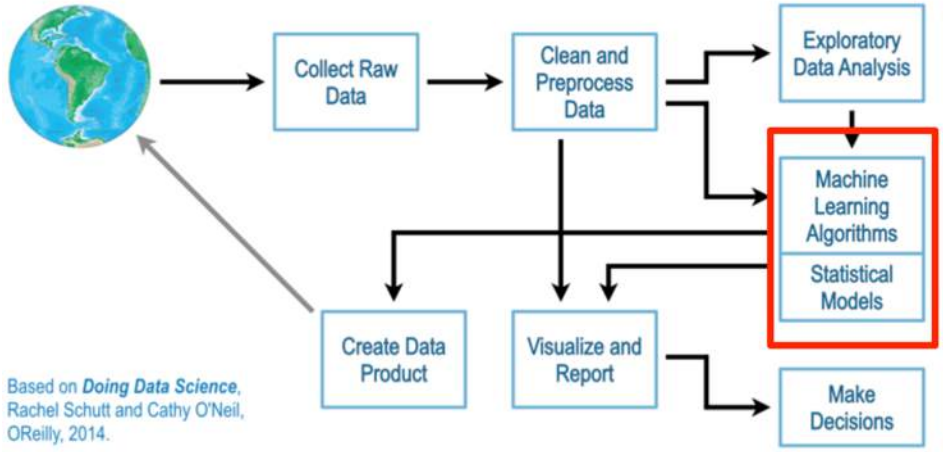
Skill: Analysis

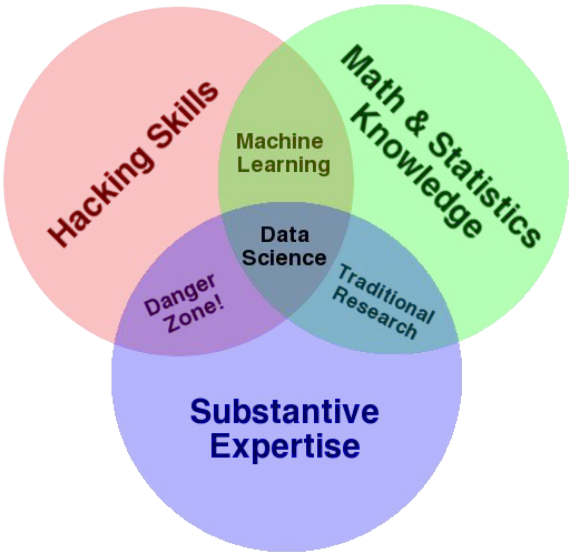
- What can I learn from my data?
- How can I describe interesting features of it?
- *Exploratory Data Analysis* can give hints on the nature of the data and which knowledge it may contain.
- *Machine Learning* and *Data Mining* can be used to **create models** that describe the data: even data we don't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



Skill: 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.





Skill: Communicate Results

Collect Raw Data

Clean and Preprocess Data

Exploratory Data Analysis

Machine Learning Algorithms

Statistical Models

Create Data Product

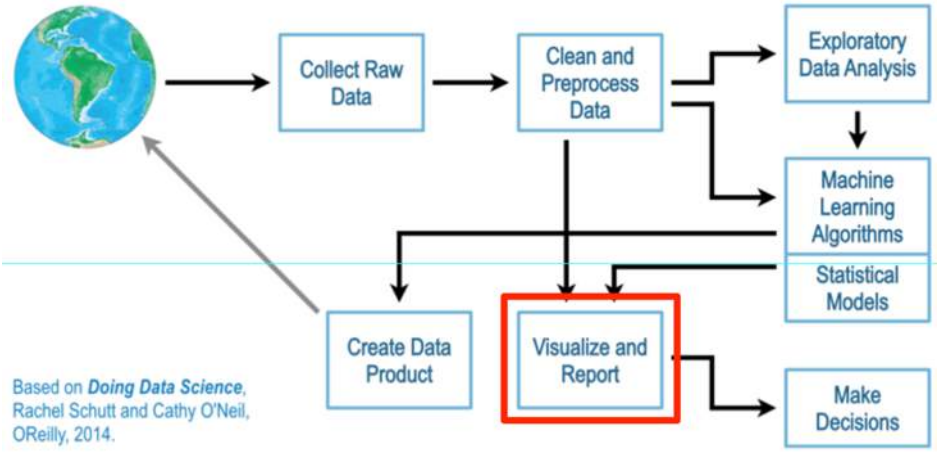
Visualize and Report

Make Decisions

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

Skill: Communicate Results

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



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))
```

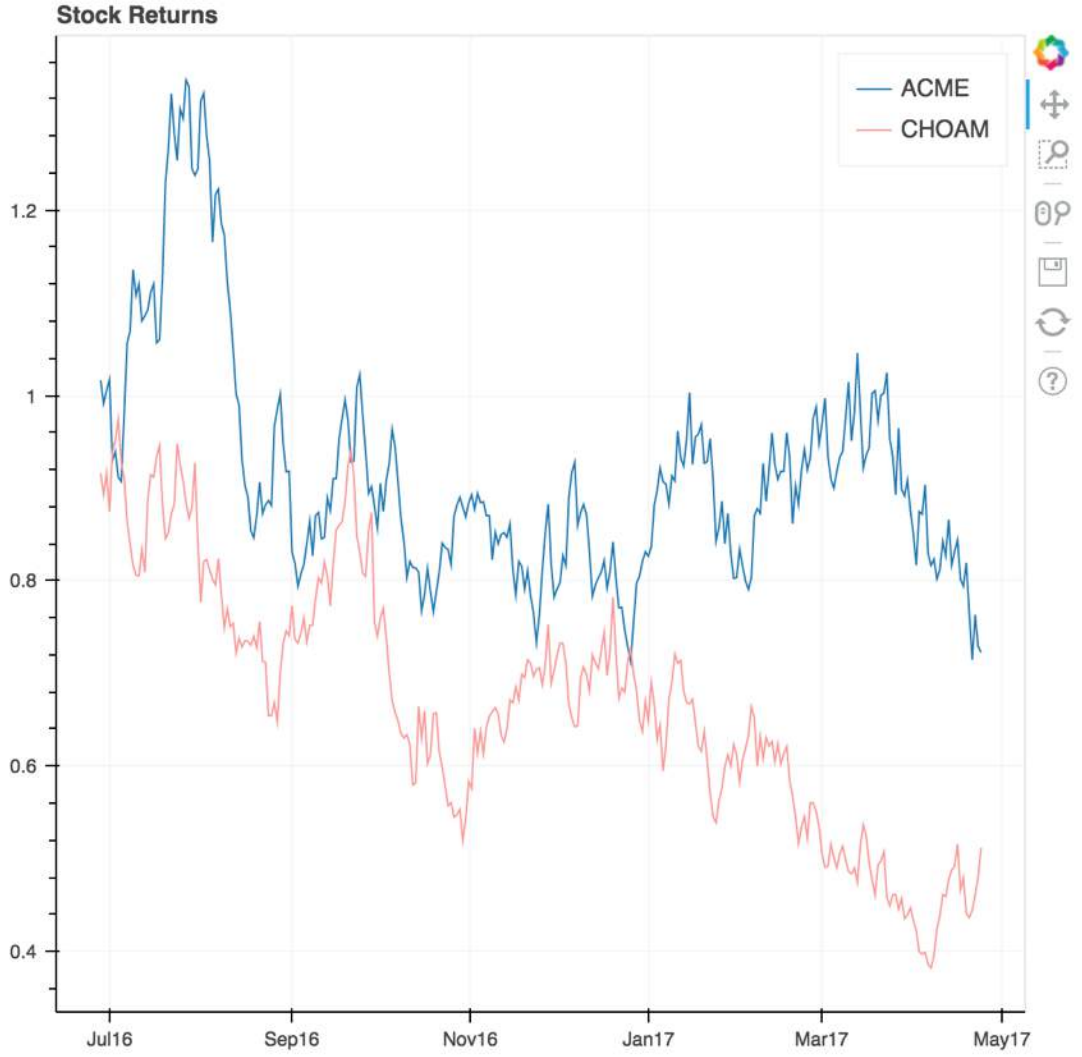
```
output_notebook()
```

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

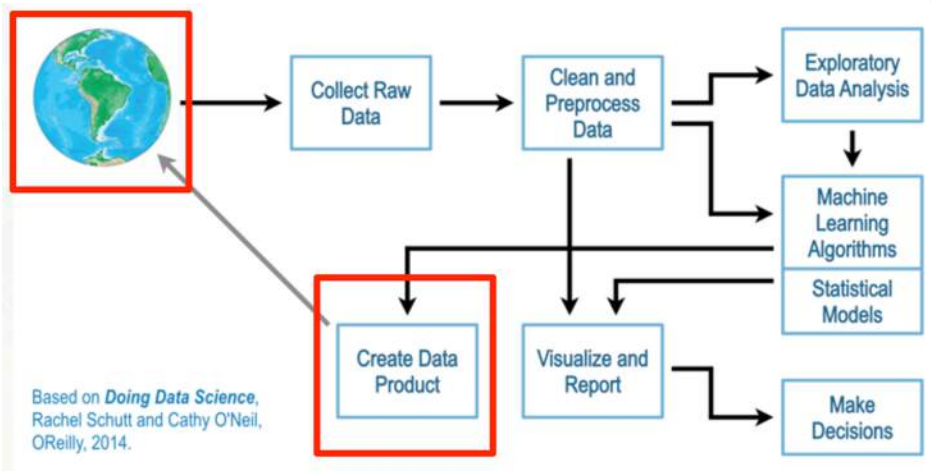
Jupyter



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!
- 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!

Let's see one (or more) examples..

http://www.lac.inpe.br/~rafael.santos/r.html

R code for the CAP394 - Introduction to Data Science course

- Data Science Example - Papers about Data Science
- Data Science Example - Books about Data Science
- Exploratory Data Analysis - Anscombes Quartet and the importance of visualization
- Data Science Example - Iris dataset
- Análise de Dados de Programas de Pós-Graduação Interdisciplinares - (in Portuguese) (veja também a Parte 2 e a Parte 3).

You're already a Data Scientist, now go ask for a raise

In conclusion…

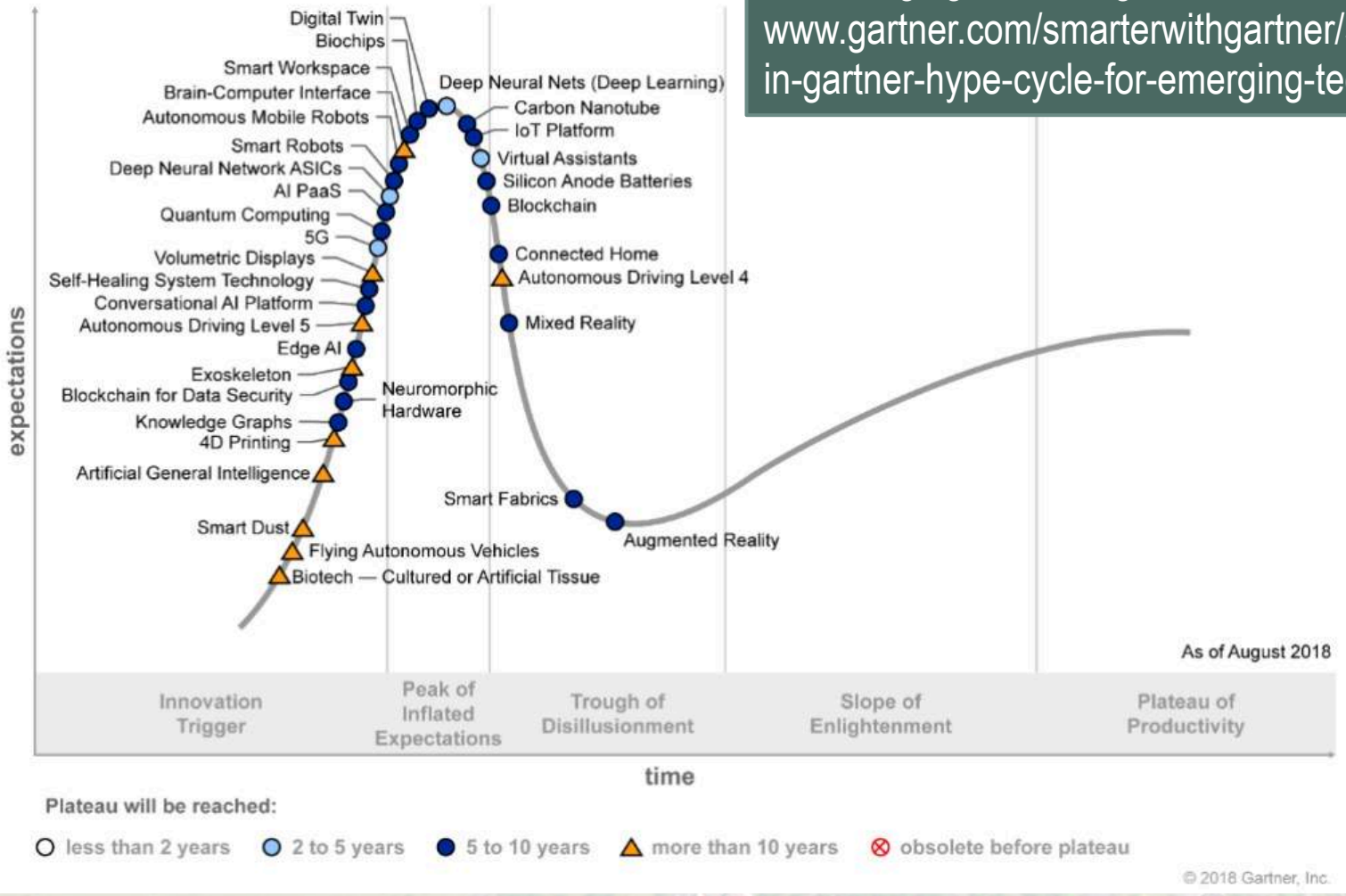
In conclusion…

- 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…

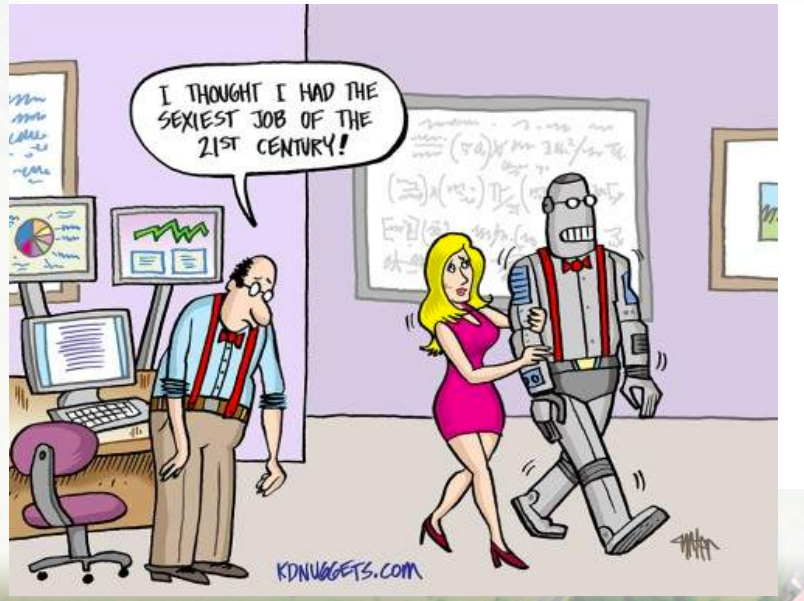
- Hype ***is*** an issue!

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/



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

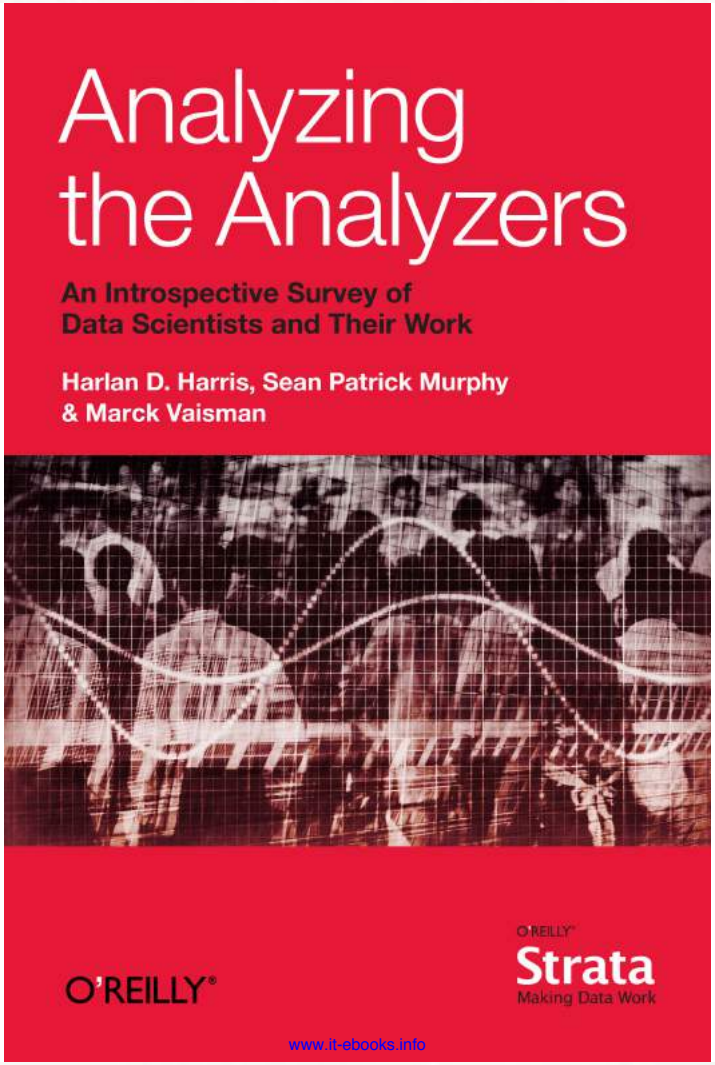
Shameless Advertising

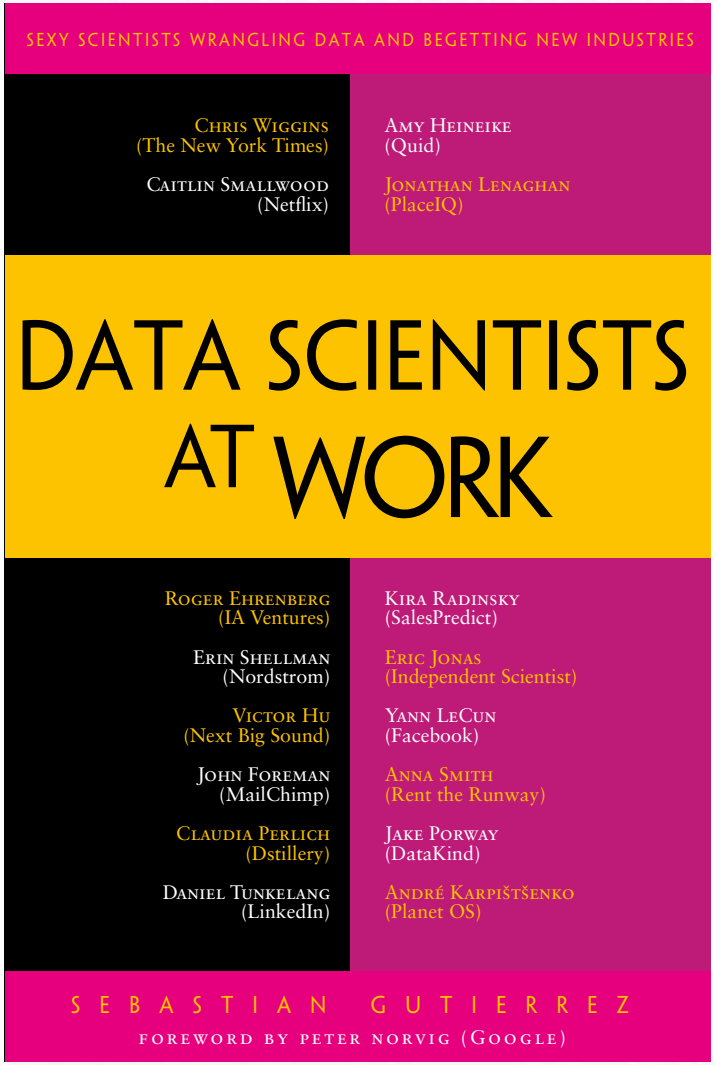
- Applied Computing Graduate Program at INPE:
  - http://www.inpe.br/pos_graduacao/cursos/cap/
- Introduction to Data Science / Data Mining
  - http://www.lac.inpe.br/~rafael.santos/index.html
- CAP's Annual Workshop:
  - http://www.inpe.br/worcap/
- LABAC's Annual Summer School:
  - http://www.inpe.br/elac2018/

rafael.santos@inpe.br

References

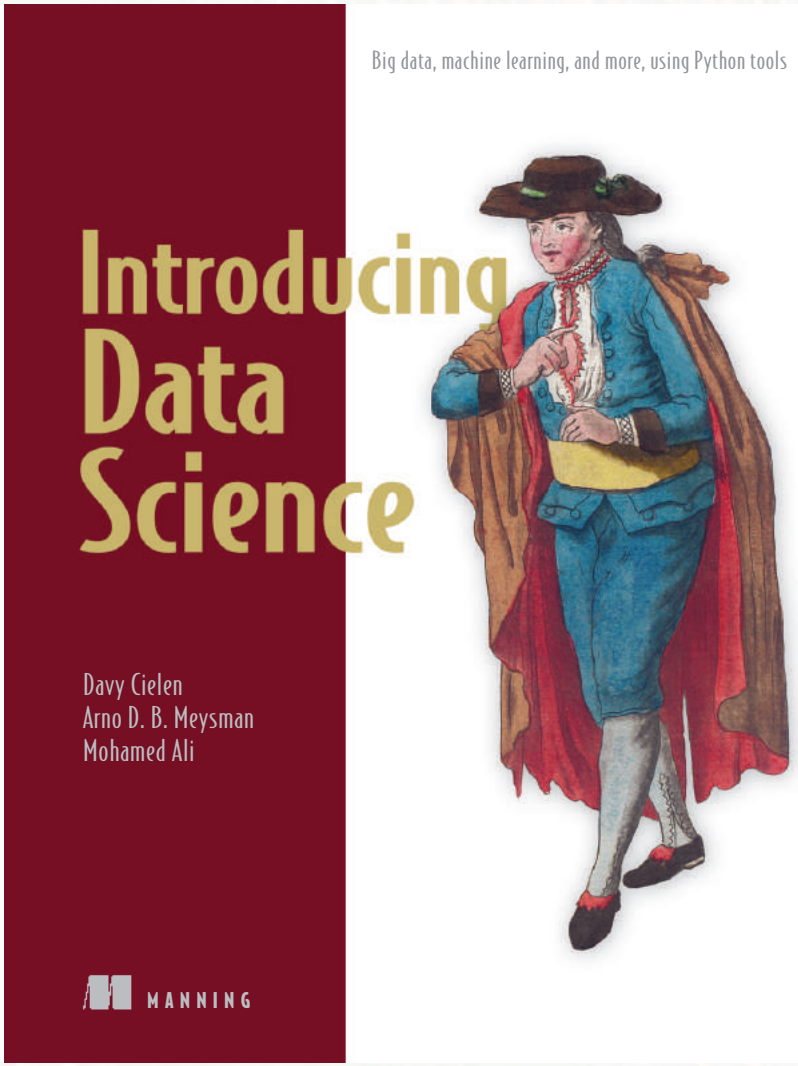
References


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

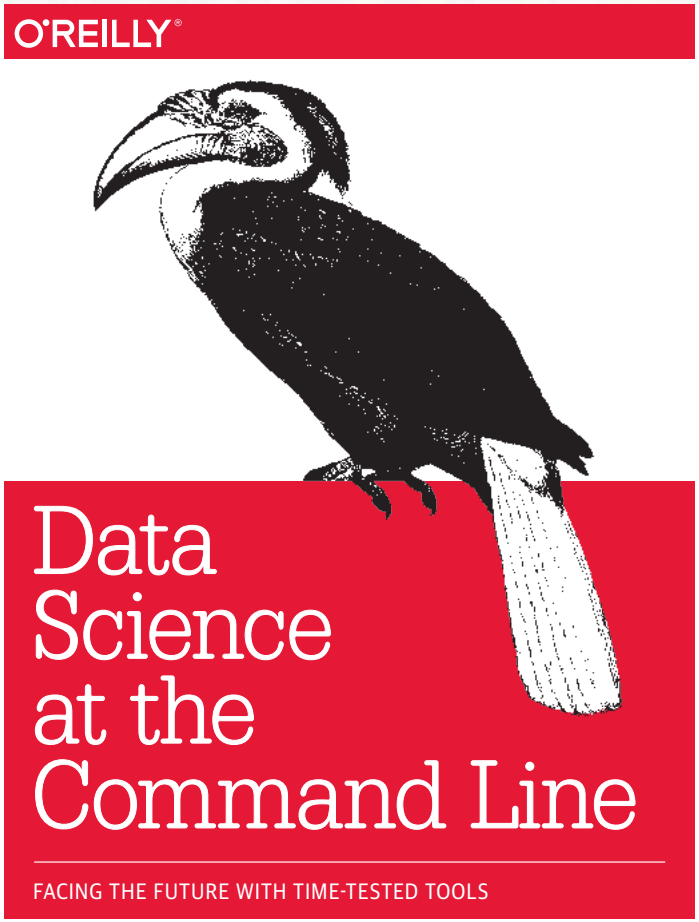

SEXY SCIENTISTS WRANGLING DATA AND BEGETTING NEW INDUSTRIES
Chris Wiggins (The New York Times)
Caitlin Smallwood (Netflix)
Amy Heineike (Quid)
Jonathan Lenaghan (PlaceIQ)
DATA SCIENTISTS AT WORK
Roger Ehrenberg (IA Ventures)
Erin Shellman (Nordstrom)
Victor Hu (Next Big Sound)
John Foreman (MailChimp)
Claudia Perlich (Dstillery)
Daniel Tunkelang (LinkedIn)
Kira Radinsky (SalesPredict)
Eric Jonas (Independent Scientist)
Yann LeCun (Facebook)
Anna Smith (Rent the Runway)
Jake Porway (DataKind)
André Karpištšenko (Planet OS)
SEBASTIAN GUTIERREZ
FOREWORD BY PETER NORVIG (GOOGLE)

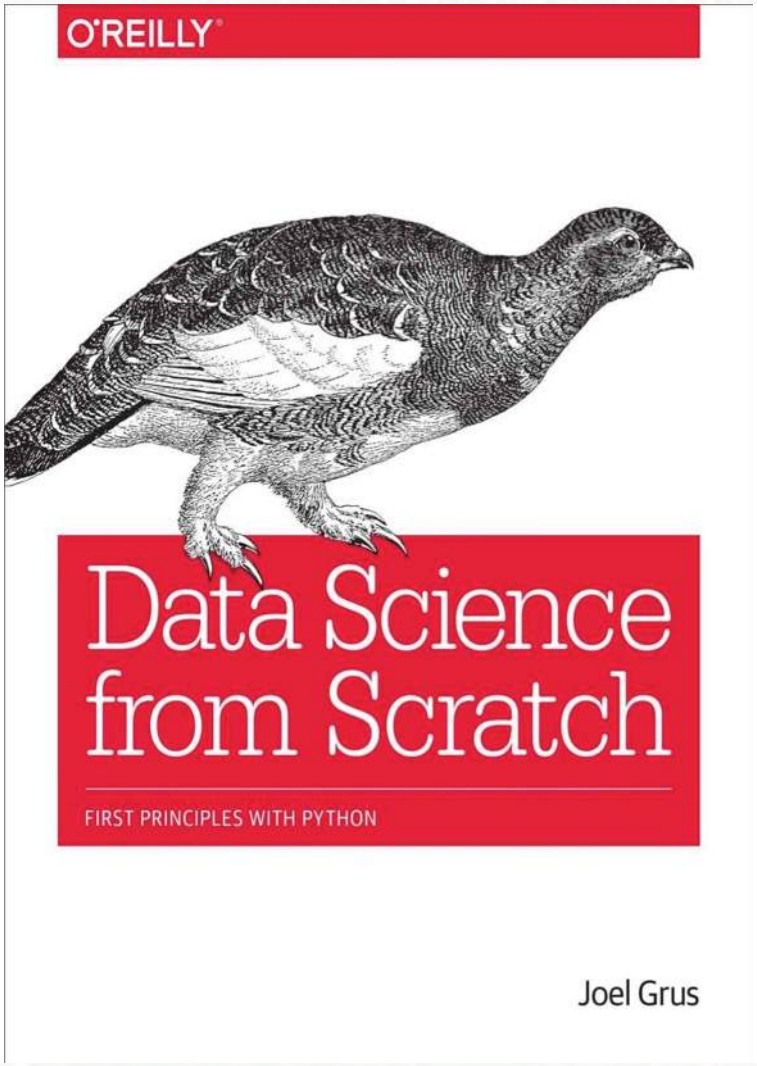
References


O'REILLY
Doing Data Science
STRAIGHT TALK FROM THE FRONTLINE
Rachel Schutt & Cathy O'Neil

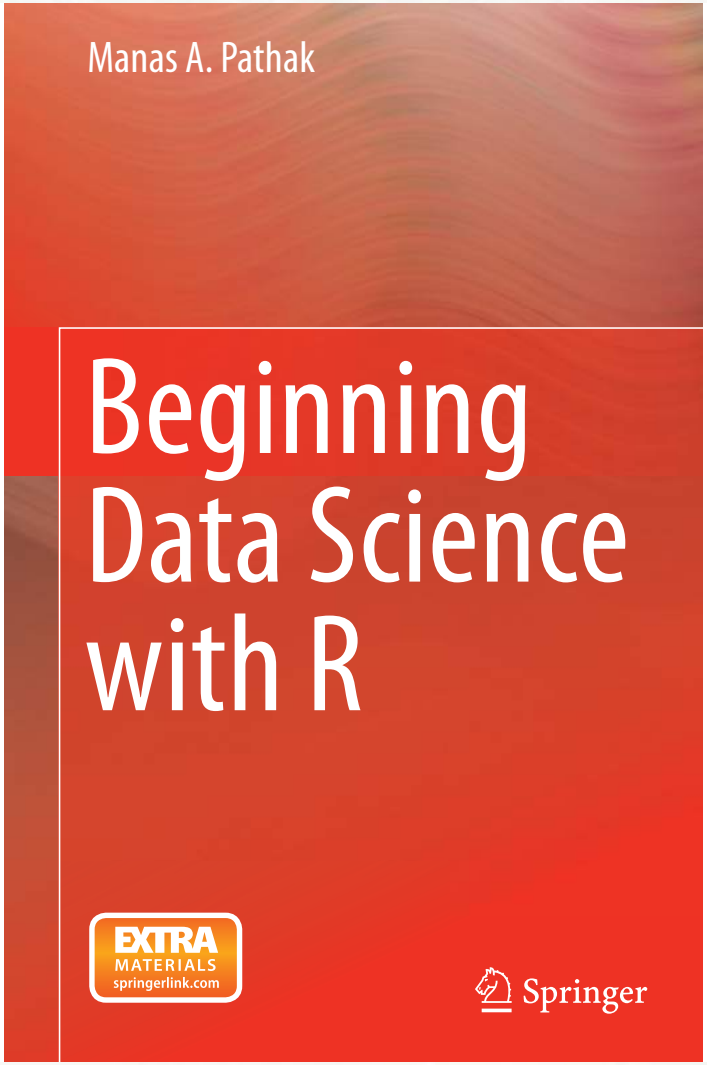

Big data, machine learning, and more, using Python tools
Introducing Data Science
Davy Cielen
Arno D. B. Meysman
Mohamed Ali
MANNING

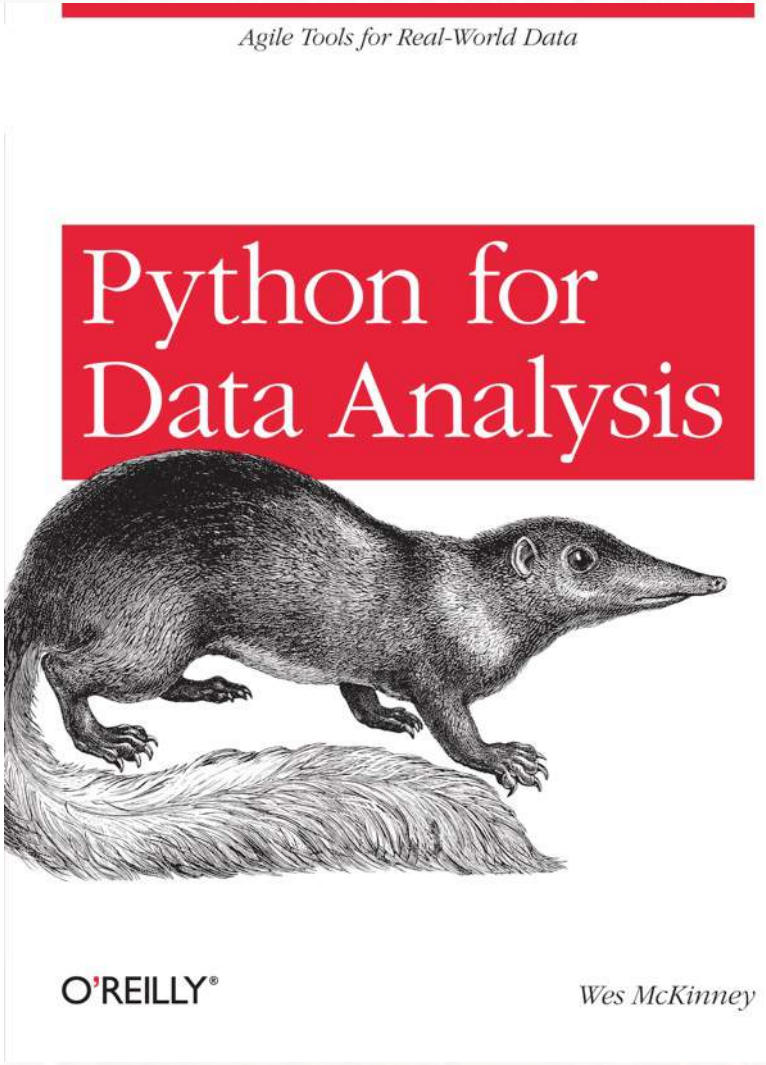
Referências

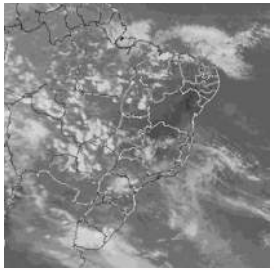




Referências







YOU'RE ALREADY A DATA SCIENTIST, NOW GO ASK FOR A RAISE

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