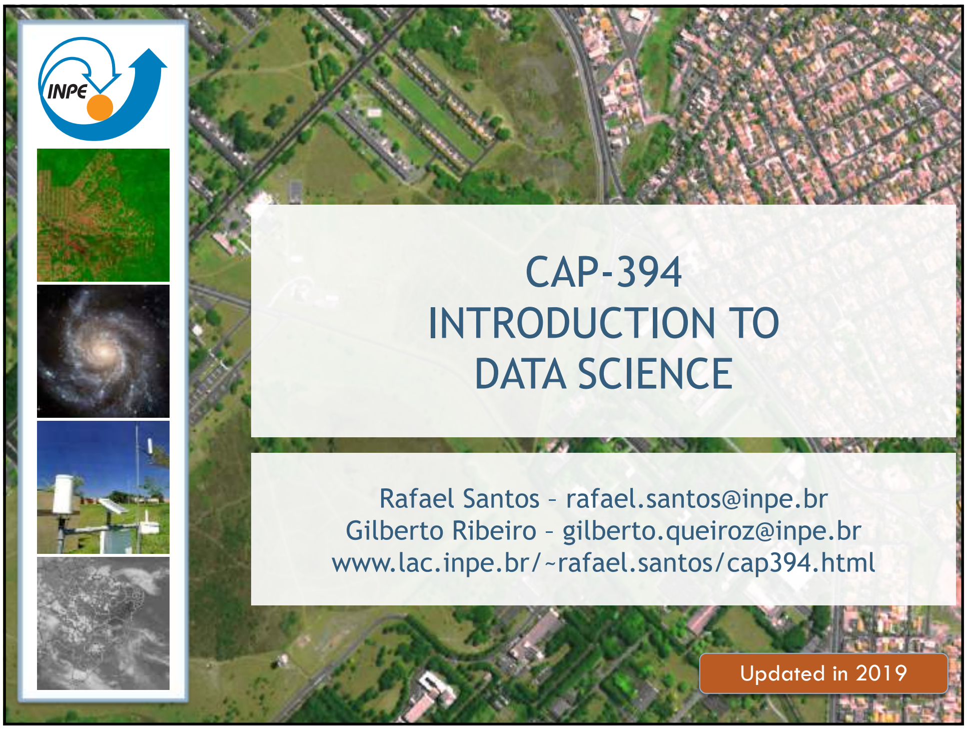
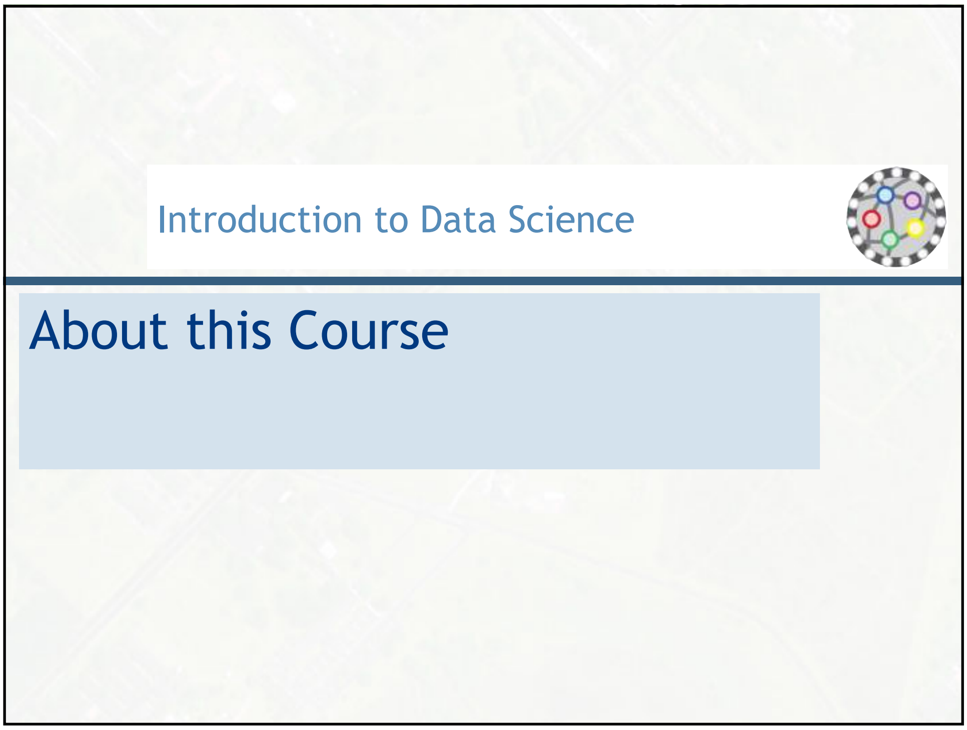


CAP-394 INTRODUCTION TO DATA SCIENCE

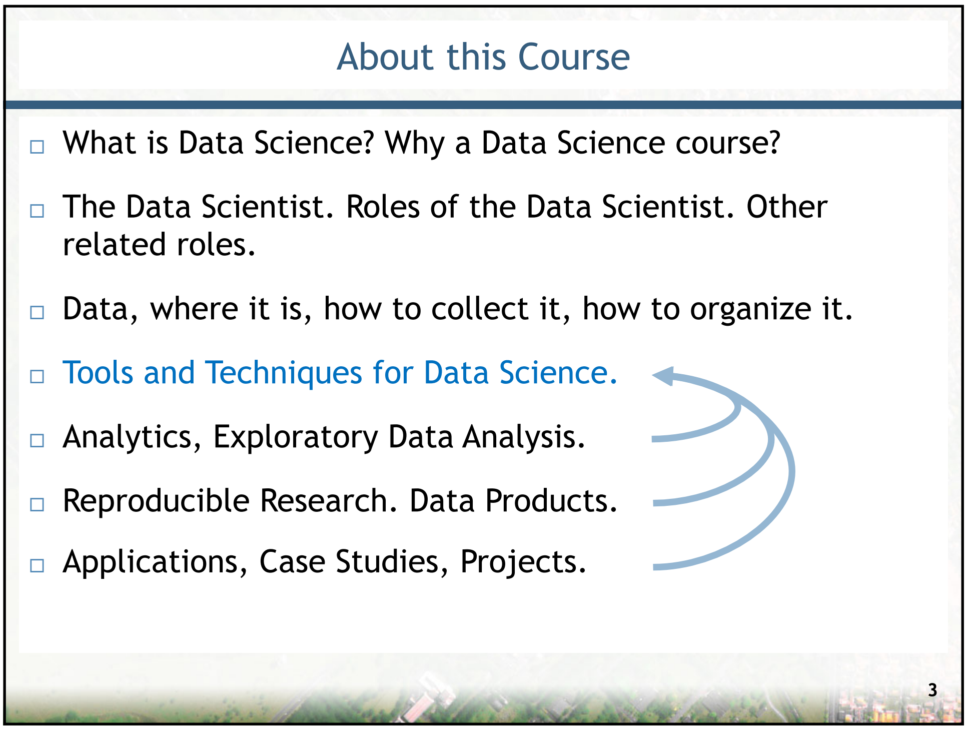
Rafael Santos - rafael.santos@inpe.br
Gilberto Ribeiro - gilberto.queiroz@inpe.br
www.lac.inpe.br/~rafael.santos/cap394.html

Updated in 2019



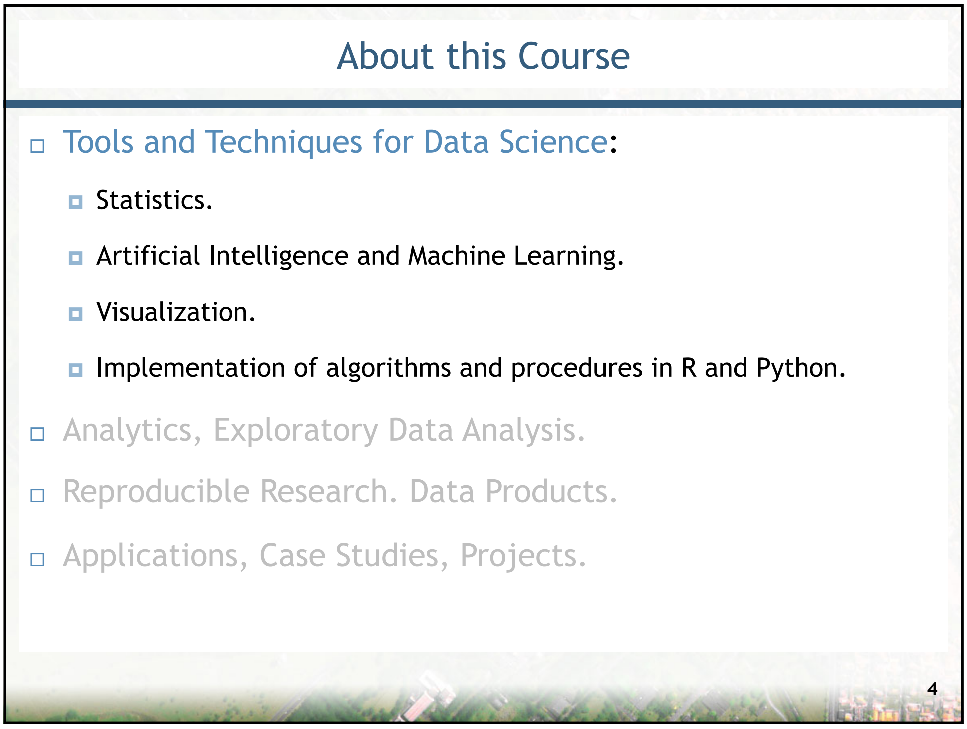
Introduction to Data Science

About this Course



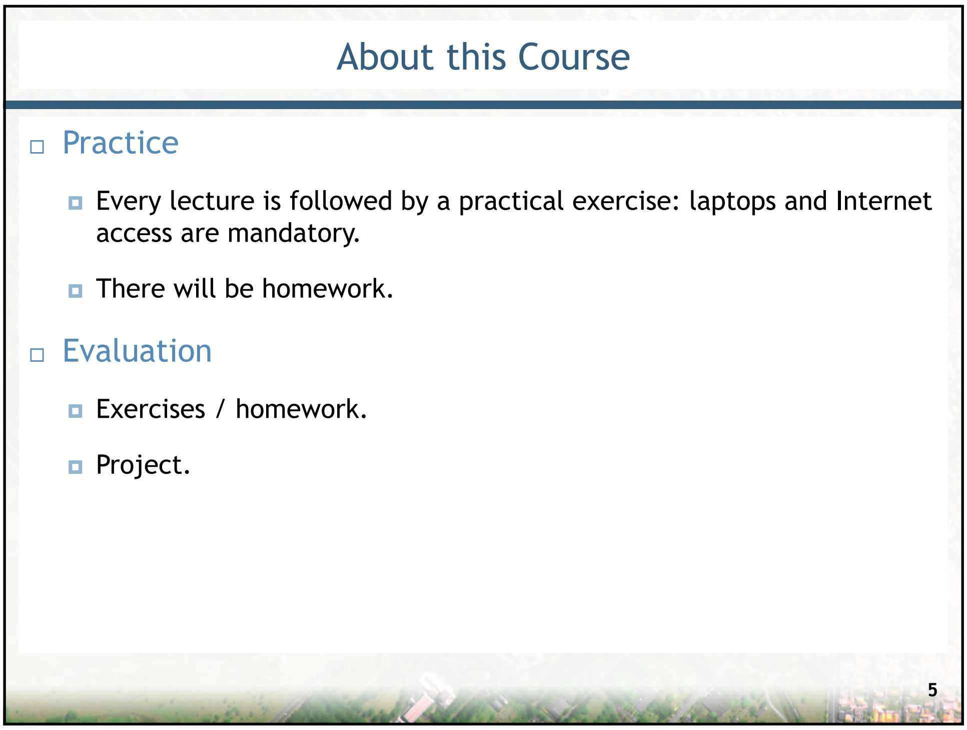
About this Course

- □ 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.
- □ Tools and Techniques for Data Science.
- □ Analytics, Exploratory Data Analysis.
- □ Reproducible Research. Data Products.
- □ Applications, Case Studies, Projects.



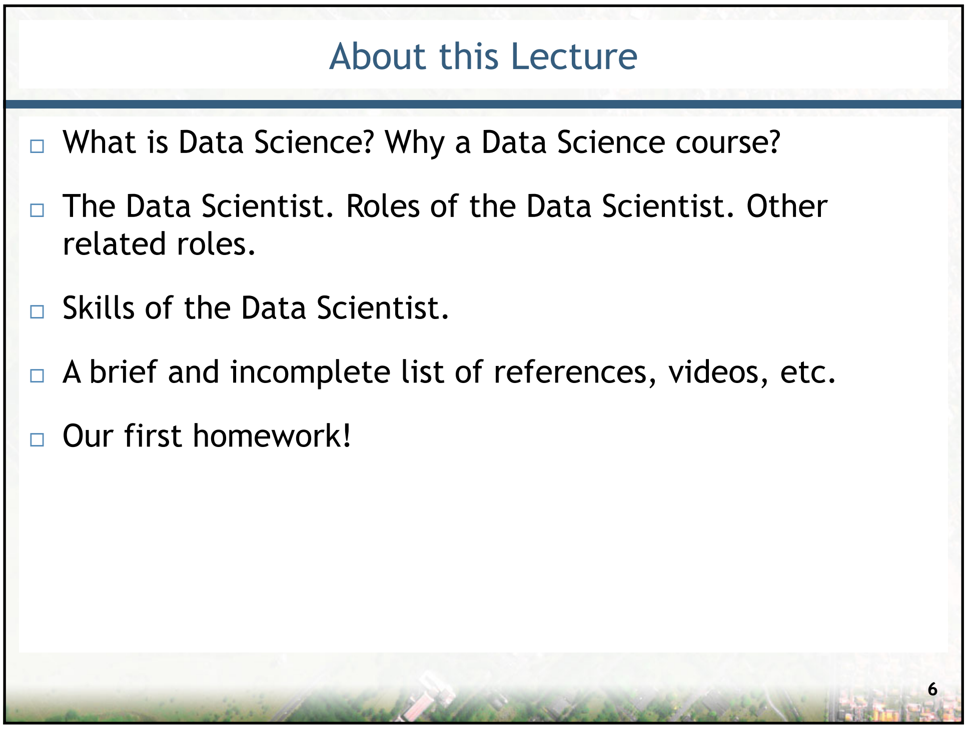
About this Course

- □ Tools and Techniques for Data Science:
  - □ Statistics.
  - □ Artificial Intelligence and Machine Learning.
  - □ Visualization.
  - □ Implementation of algorithms and procedures in R and Python.
- □ Analytics, Exploratory Data Analysis.
- □ Reproducible Research. Data Products.
- □ Applications, Case Studies, Projects.



About this Course

- Practice
  - Every lecture is followed by a practical exercise: laptops and Internet access are mandatory.
  - There will be homework.
- Evaluation
  - Exercises / homework.
  - Project.



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.
- Our first homework!

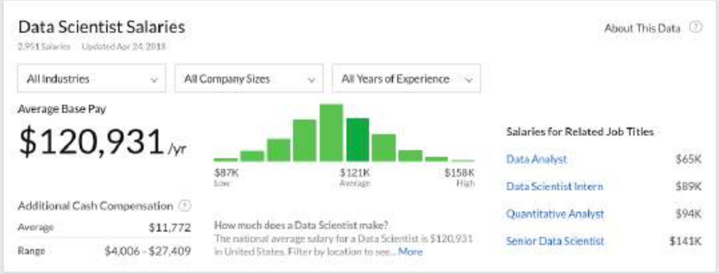
Introduction to Data Science

So you want to be a Data Scientist…

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

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

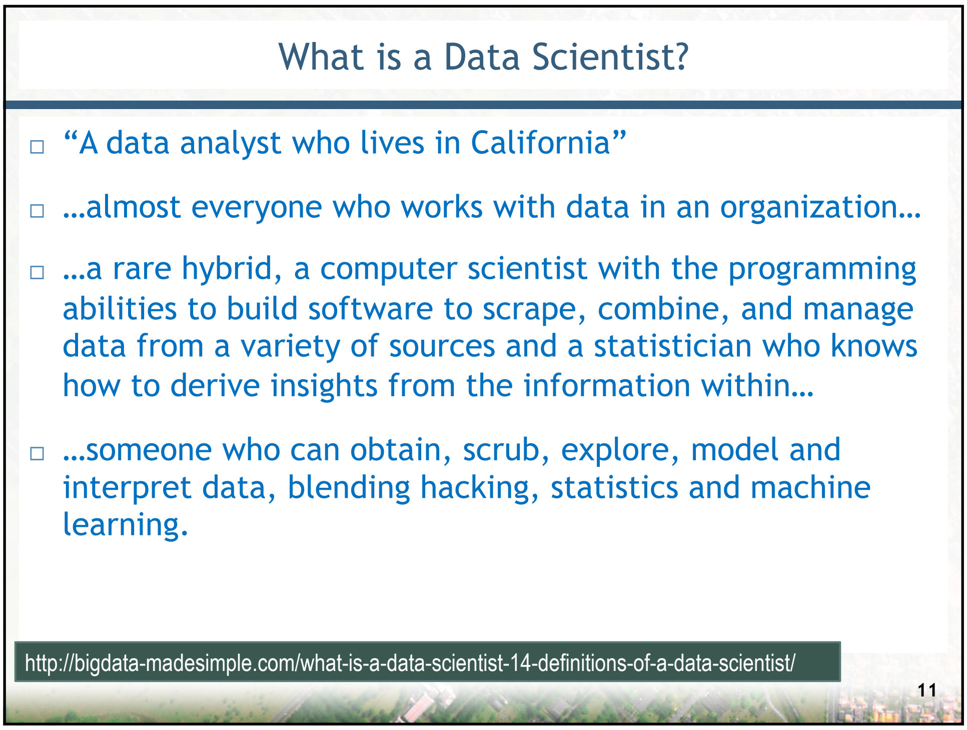
Forbes, https://www.forbes.com/sites/emc/2014/06/26/the-hottest-jobs-in-it-training-tomorrows-data-scientists/

Hype

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

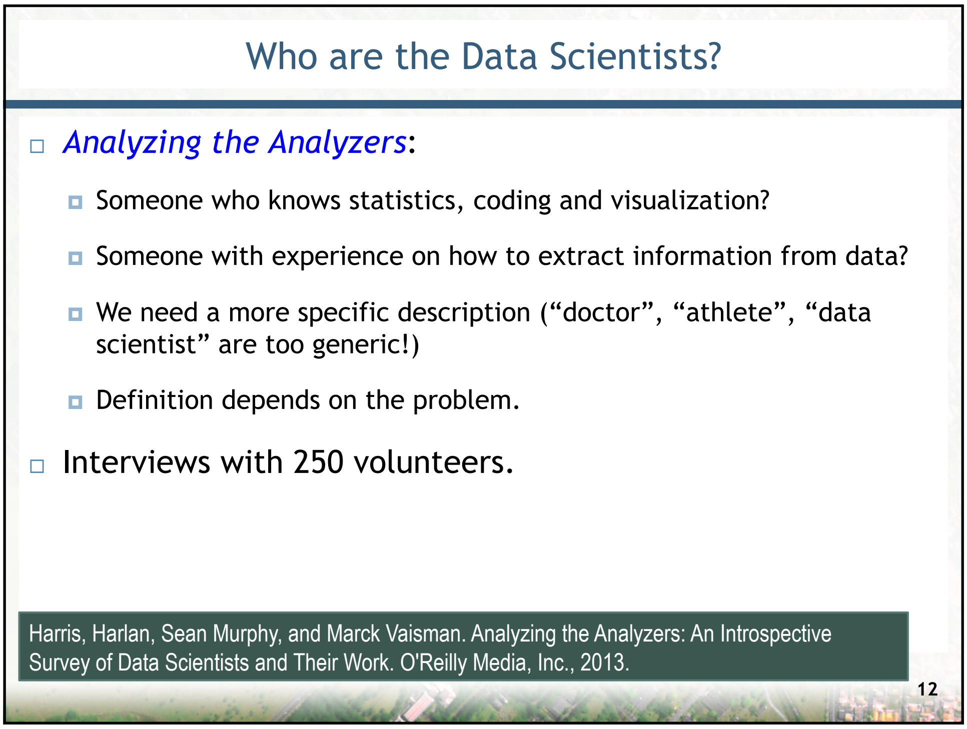


What is a Data Scientist?

- “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/

11

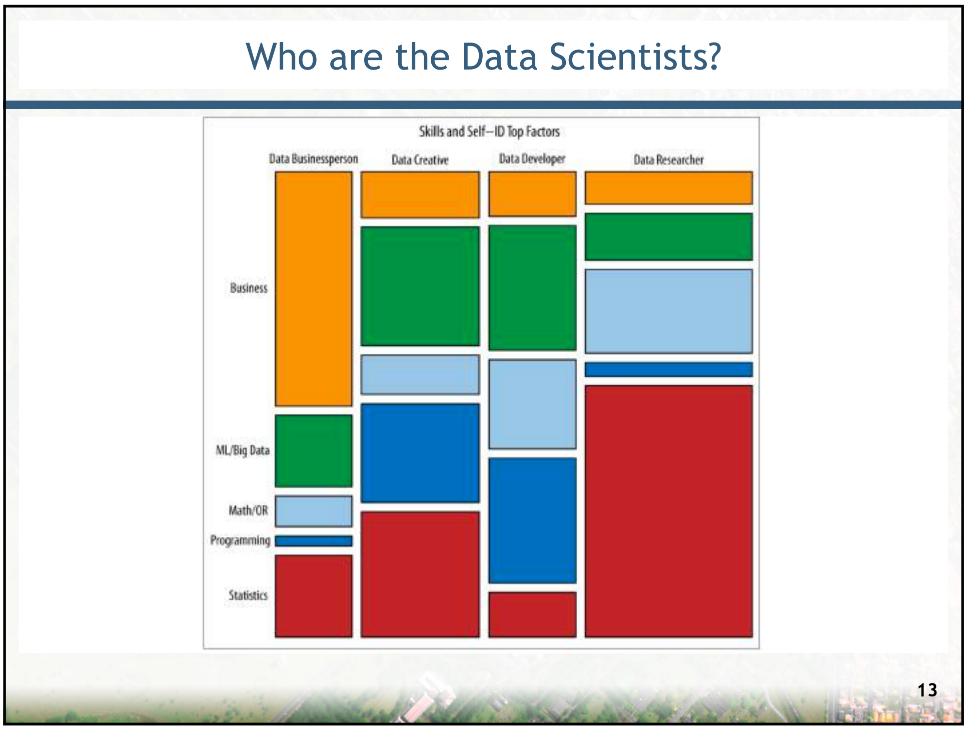


Who are the Data Scientists?

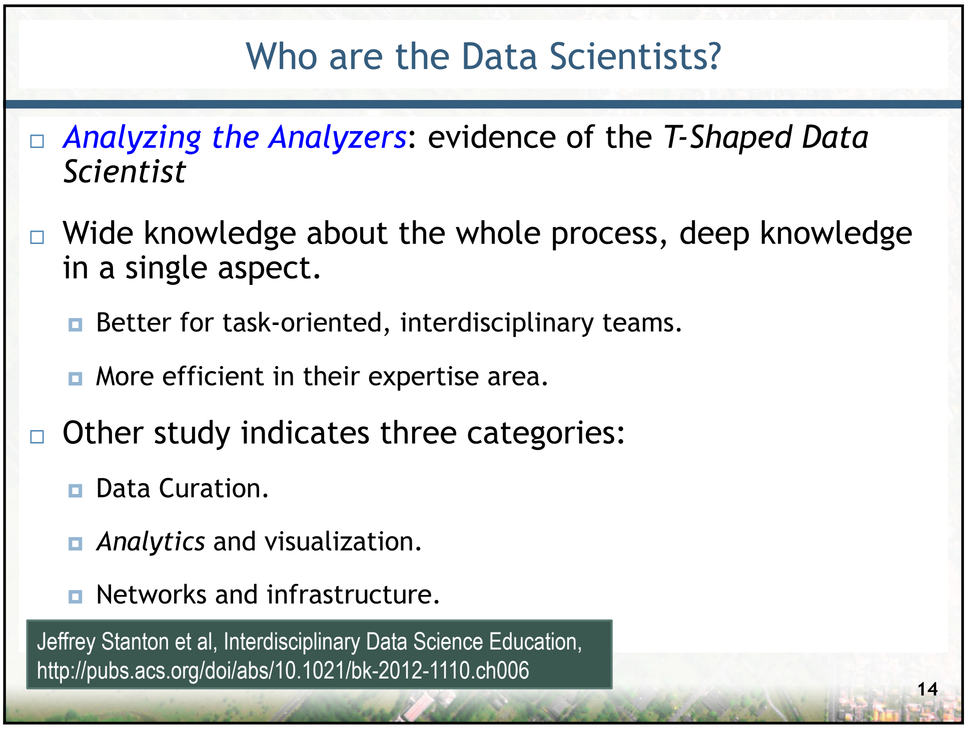
- *Analyzing the Analyzers*:
  - Someone who knows statistics, coding and visualization?
  - Someone with experience on how to extract information from data?
  - We need a more specific description (“doctor”, “athlete”, “data scientist” are too generic!)
  - Definition depends on the problem.
- Interviews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

12



Who are the Data Scientists?

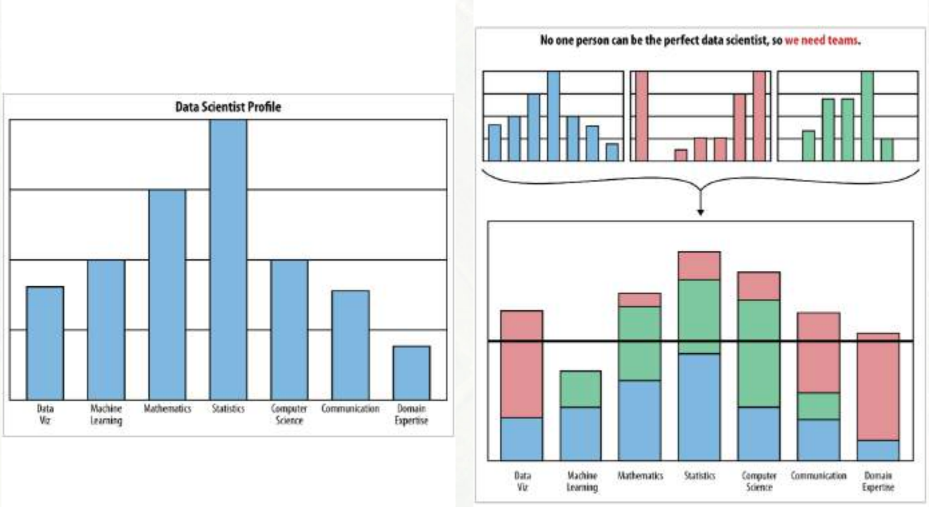


Who are the Data Scientists?

- *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

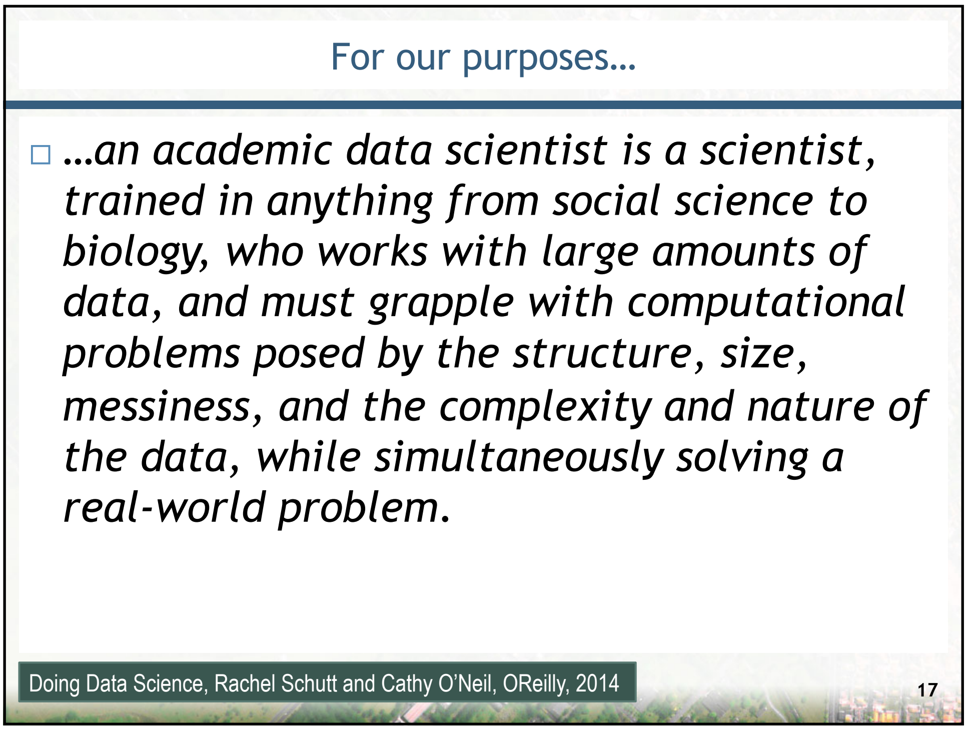
T-Shaped Data Scientist



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

So you want to be a Data Scientist…

- □ If you…
  - ☑ Have access (or can have access) to thematic data collections in different degrees of organization or tidiness, and know which kind of information can be extracted from it; and
  - ☑ Knows enough about coding in languages like R or Python, and using technologies such as SQL/NoSQL, distributed systems/web; and
  - ☑ Understand the basics of modeling, testing, algorithms, analysis, visualization; then…
- □ *You already are a Data Scientist!*

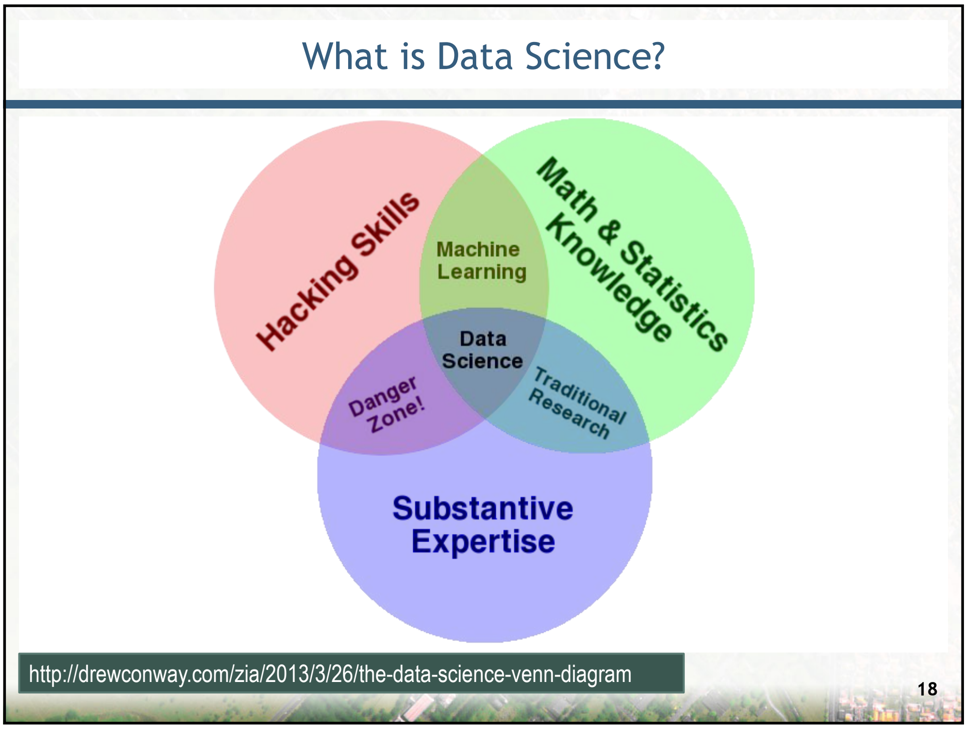


For our purposes…

□ *…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

17

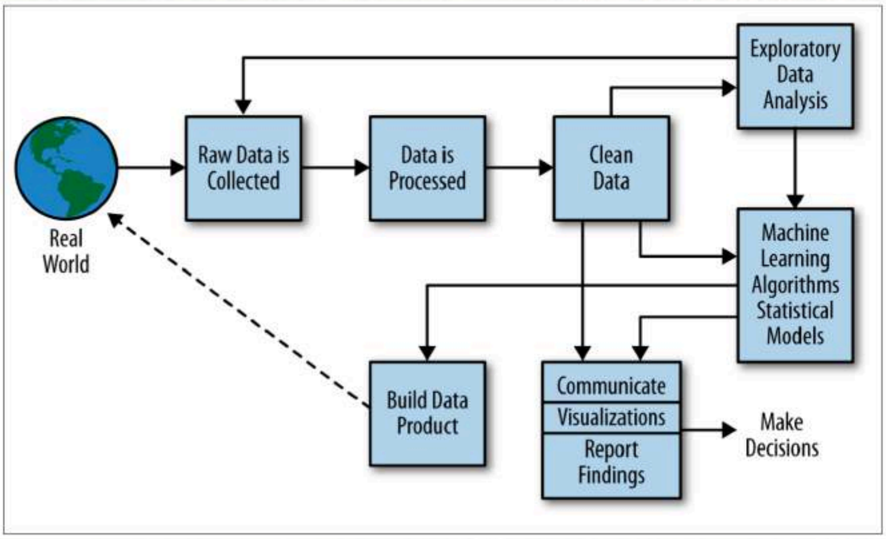


What is Data Science?

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

18

Data Science as a **Process**



Real World

Raw Data is Collected

Data is Processed

Clean Data

Exploratory Data Analysis

Machine Learning Algorithms Statistical Models

Build Data Product

Communicate Visualizations Report Findings

Make Decisions

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

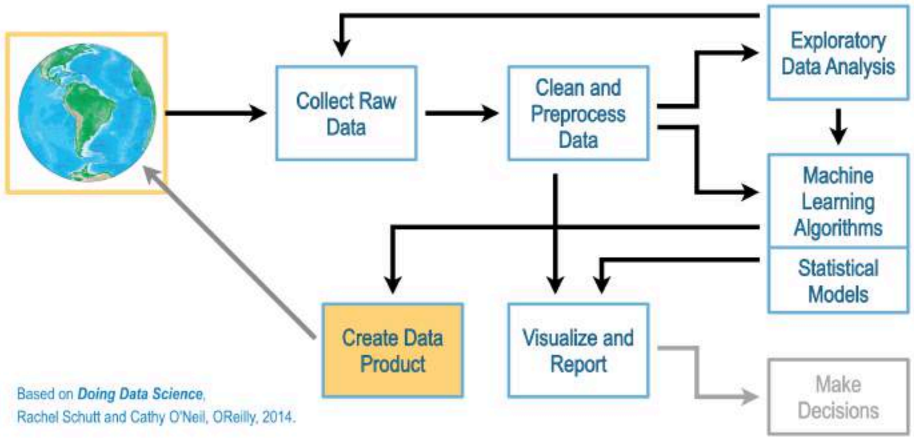
Introduction to Data Science

Skills

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.

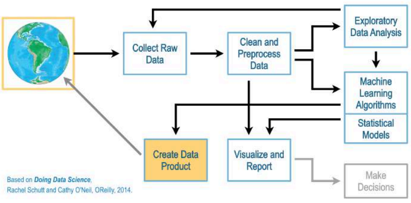
Skill: Understand the Problem



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

Skill: Understand the Problem

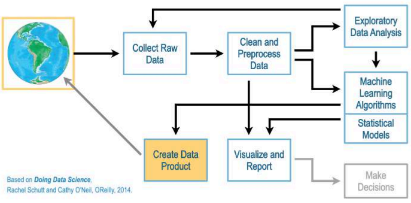
- At least enough to communicate with people that understand the problem!
  - Data Science is inherently interdisciplinary!
- What are good questions about the data?
  - …about the phenomena measured by that data?



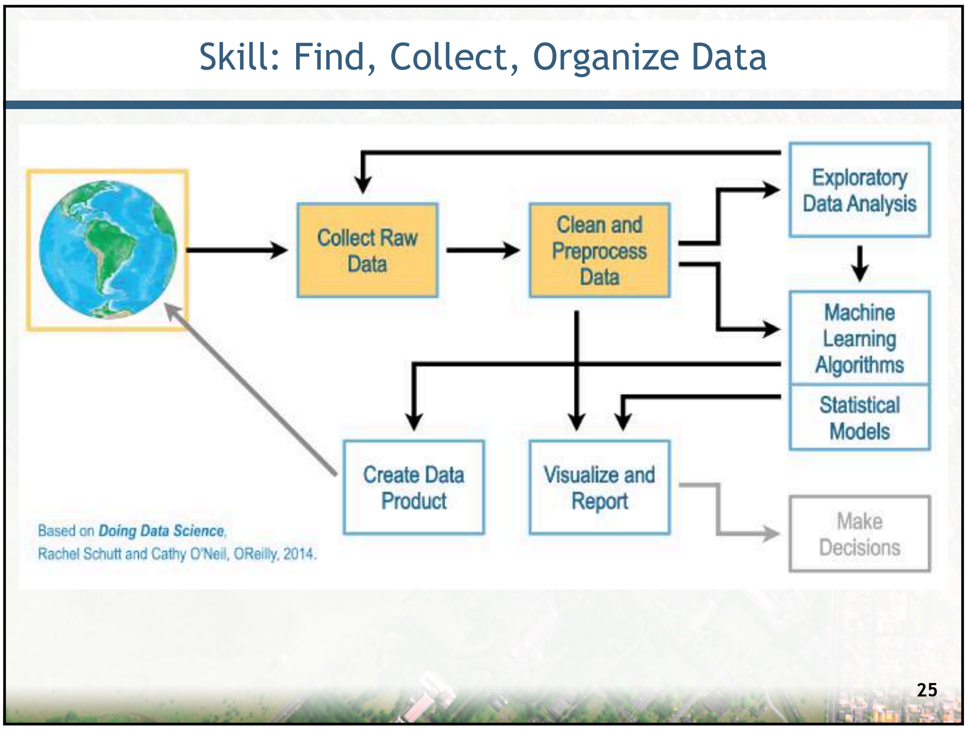
Based on *Doing Data Science*, Rachel Schutt and Cathy O'Neil, OReilly, 2014.

Skill: Understand the Problem

- 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!



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



Skill: Find, Collect, Organize Data

Collect Raw Data

Clean and Preprocess Data

Exploratory Data Analysis

Machine Learning Algorithms

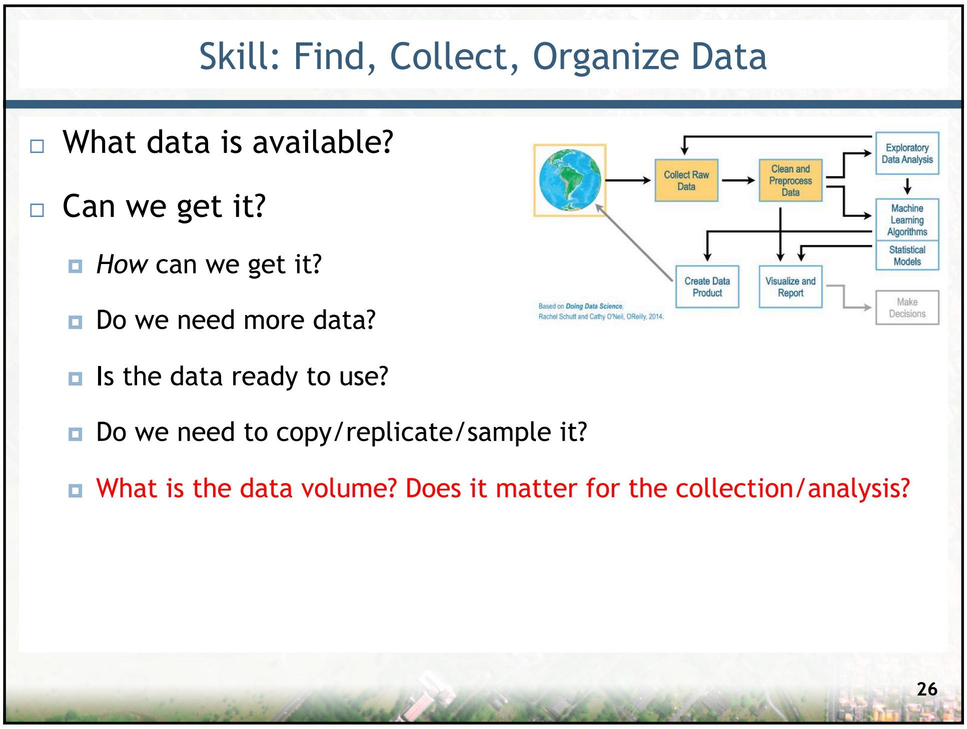
Statistical Models

Create Data Product

Visualize and Report

Make Decisions

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



Skill: 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?

Collect Raw Data

Clean and Preprocess Data

Exploratory Data Analysis

Machine Learning Algorithms

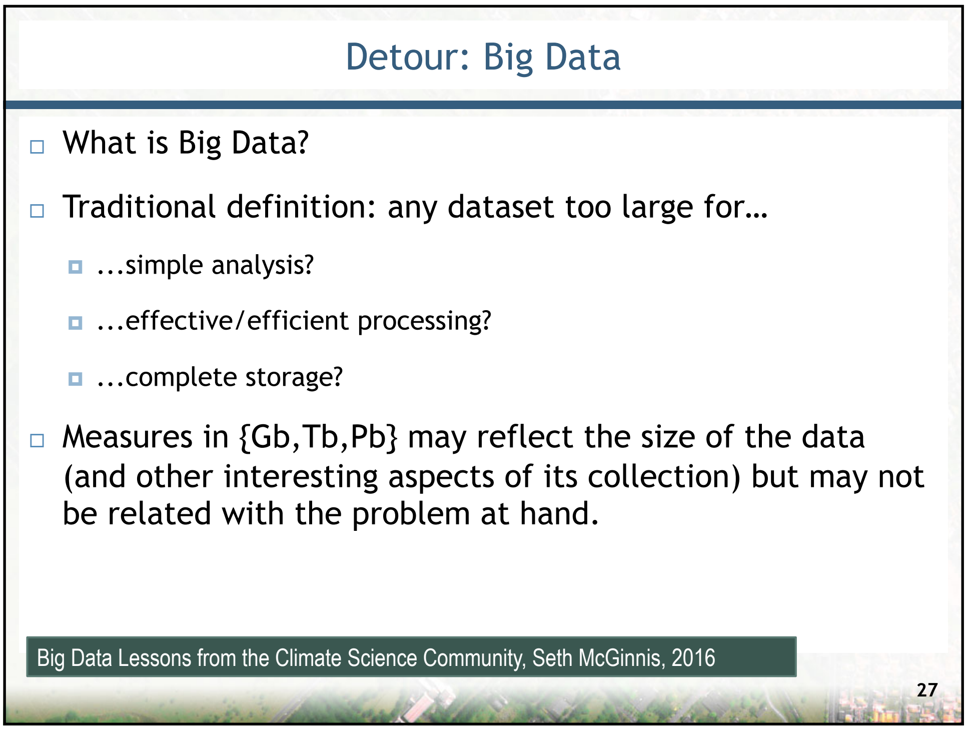
Statistical Models

Create Data Product

Visualize and Report

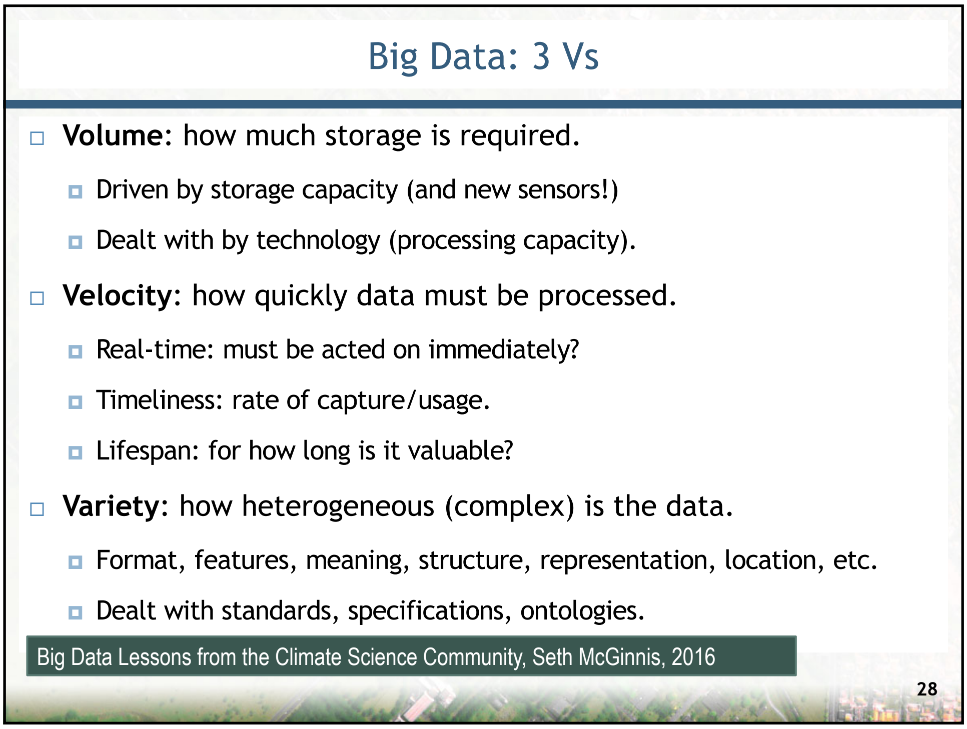
Make Decisions

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



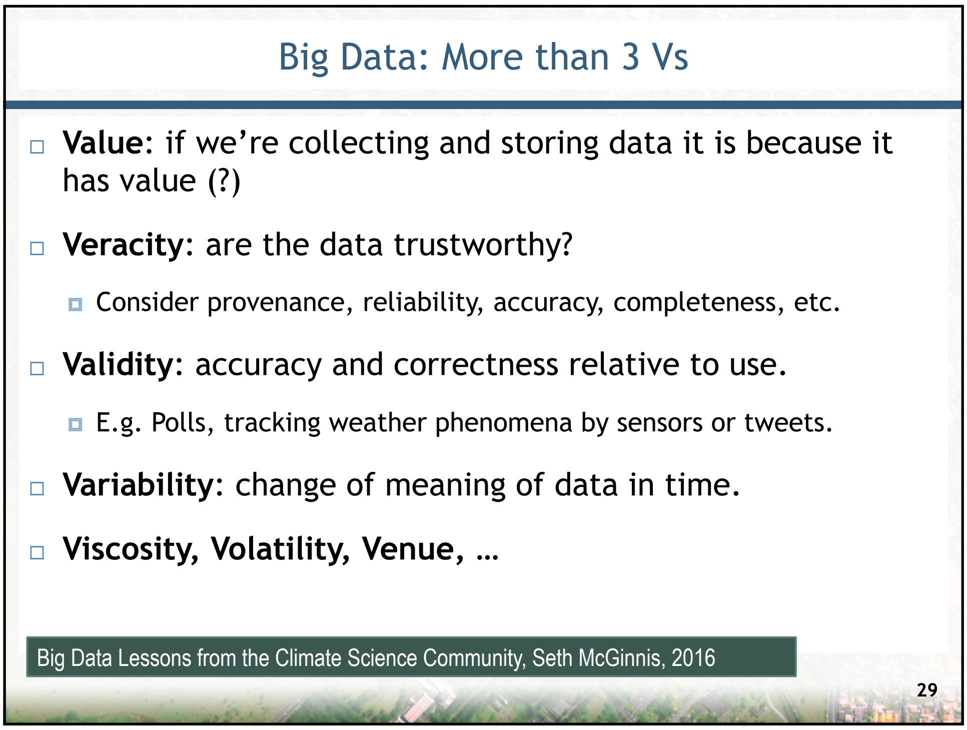
Detour: Big Data

- 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: 3 Vs

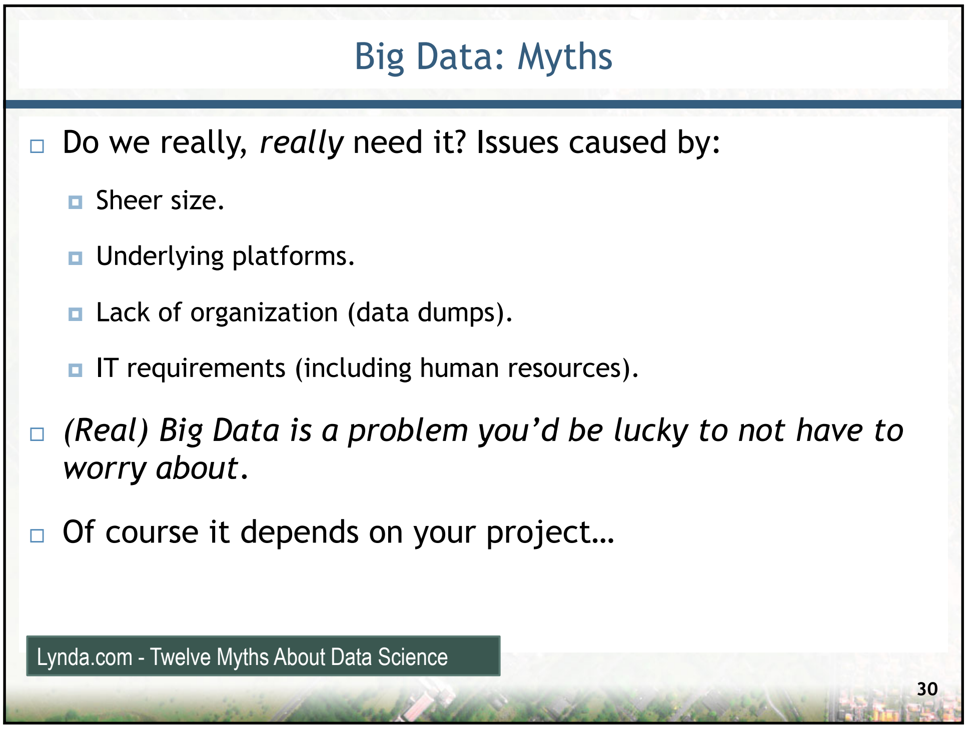
- **Volume**: how much storage is required.
  - Driven by storage capacity (and new sensors!)
  - Dealt with by technology (processing capacity).
- **Velocity**: how quickly data must be processed.
  - Real-time: must be acted on immediately?
  - Timeliness: rate of capture/usage.
  - Lifespan: for how 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: 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

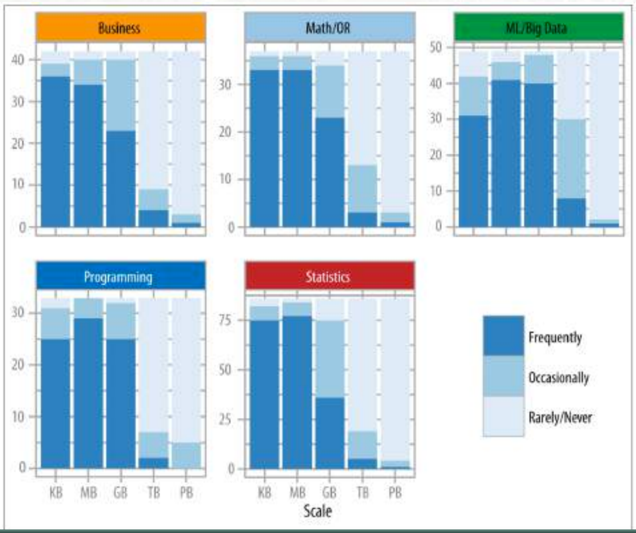


Big Data: Myths

- 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…

Lynda.com - Twelve Myths About Data Science

Back to Skills: Big Data



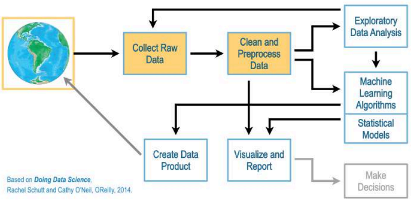
Harlan Harris, Sean Murphy, and Marck Vaisman. Analyzing the Analyzers: An Introspective Survey of Data Scientists and Their Work. O'Reilly Media, Inc., 2013.

End of Detour: Big Data

- A 3.5 petabyte pile of disks!
  - ≈1760 2Tb disks
  - ≈1100 kg

Skill: Understand and Organize Data

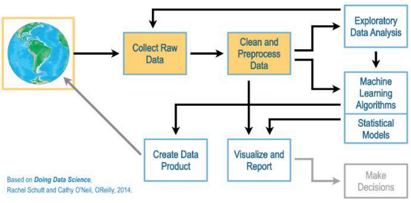
- Before Processing: How data is organized?
  - Tables, documents, images, relations, graphs, mixed?
  - Is data in a format useful for processing?
    - How can we transform it?
    - How hard it is to transform it?



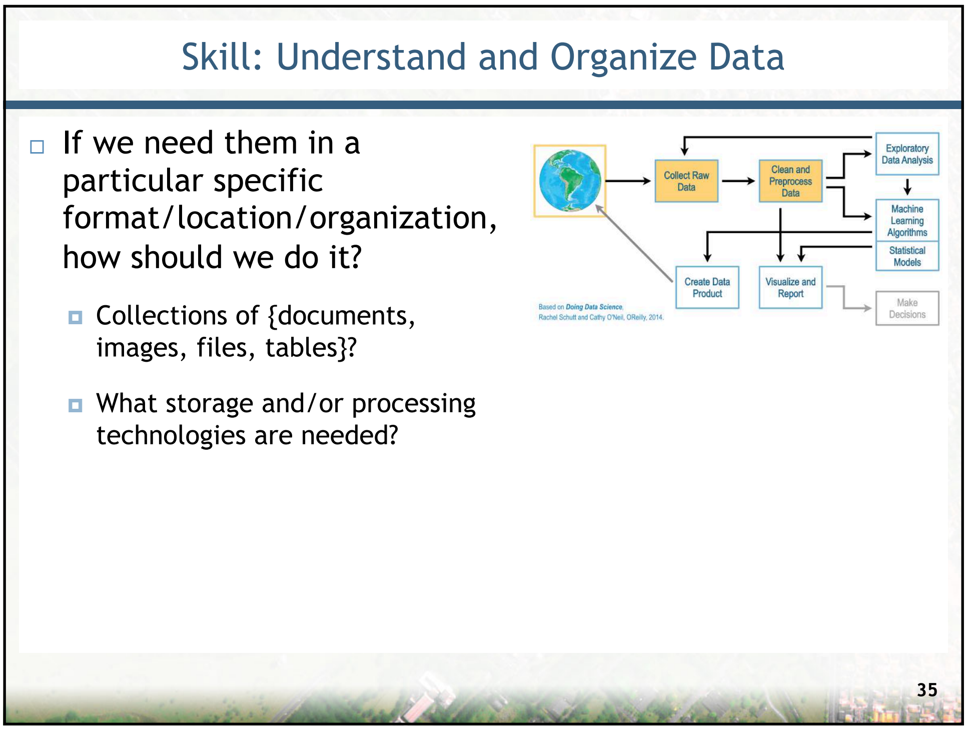
Based on *Doing Data Science*, Rachel Schutt and Cathy O'Neil, OReilly, 2014.

Skill: Understand and Organize Data

- 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?



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

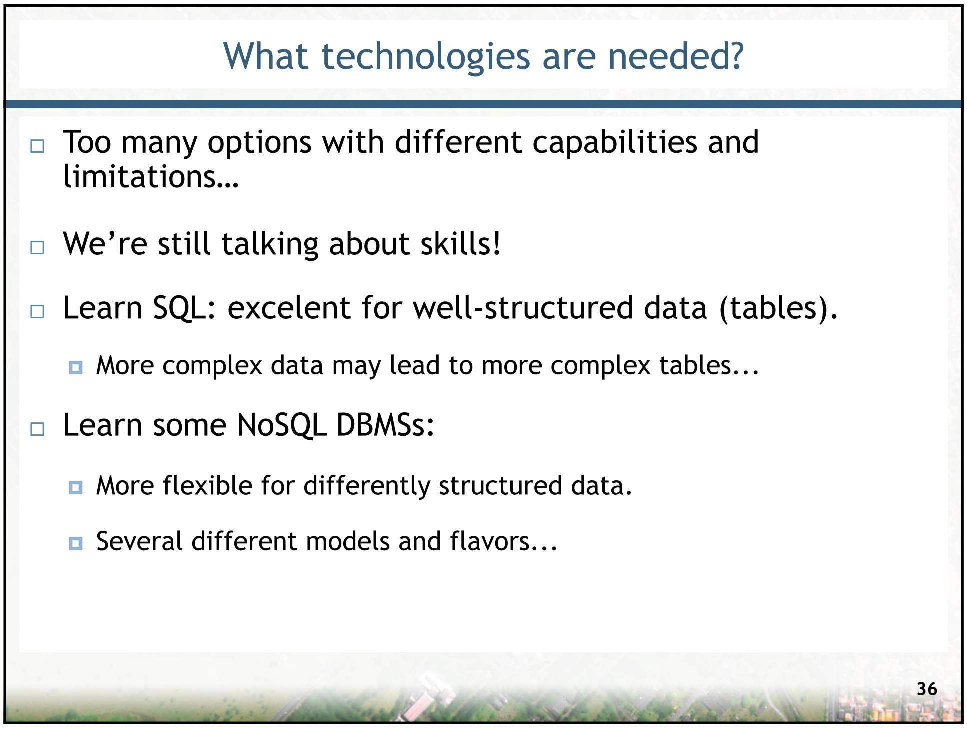


Skill: Understand and Organize Data

- 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?

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.



What technologies are needed?

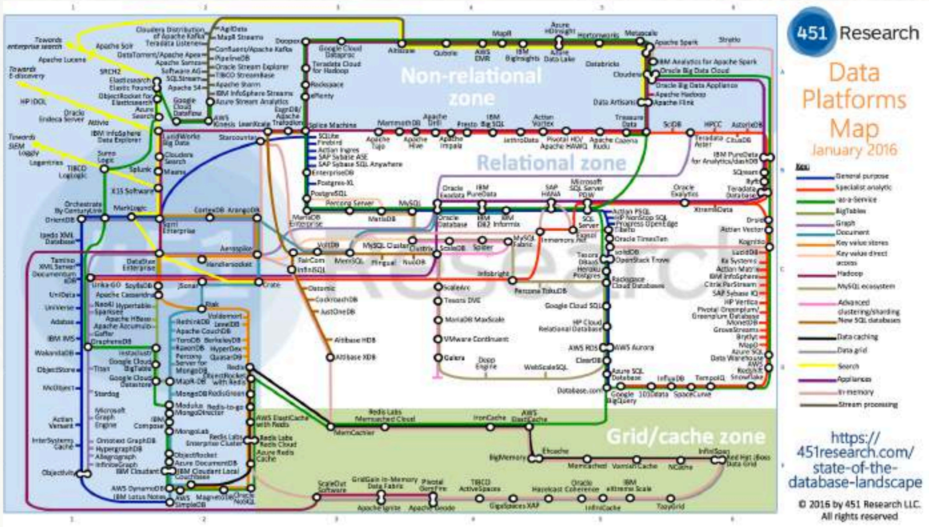
- Too many options with different capabilities and limitations…
- We’re still talking about skills!
- Learn SQL: excelent for well-structured data (tables).
  - More complex data may lead to more complex tables...
- Learn some NoSQL DBMSs:
  - More flexible for differently structured data.
  - Several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

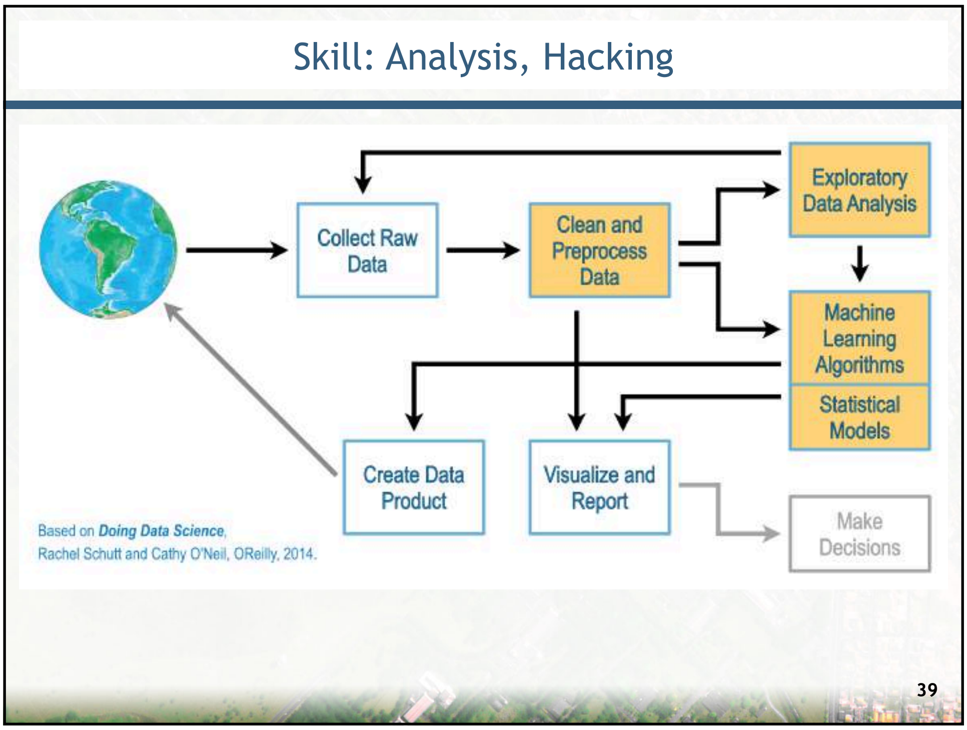
SQL/NoSQL



Data Platforms Map
January 2016

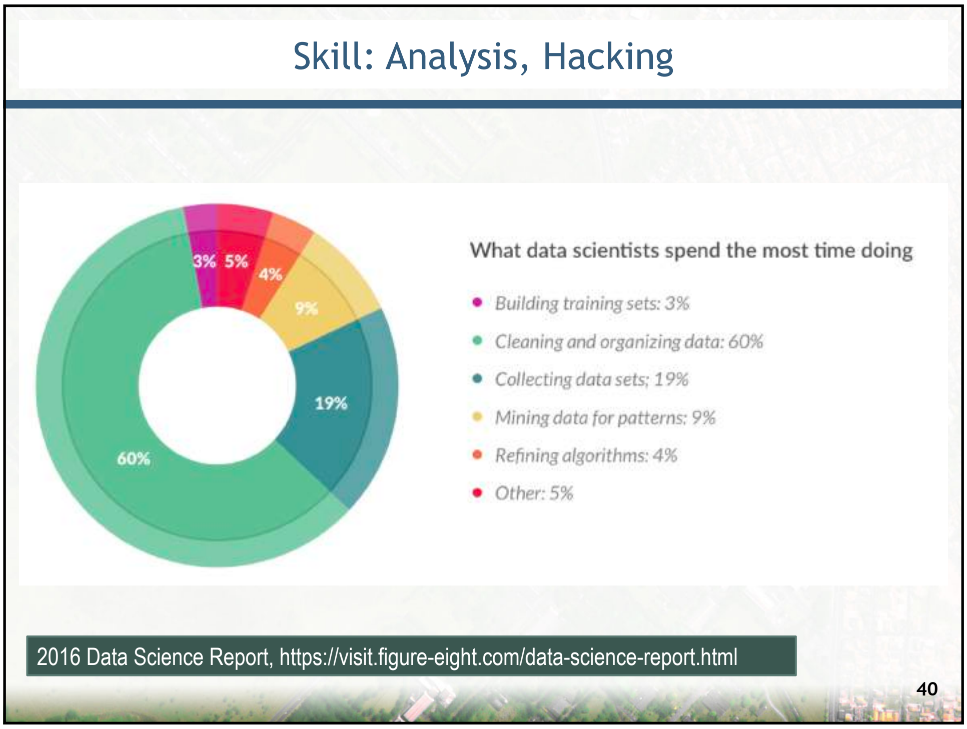
https://451research.com/state-of-the-database-landscape

© 2016 by 451 Research LLC.
All rights reserved



Skill: Analysis, Hacking

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



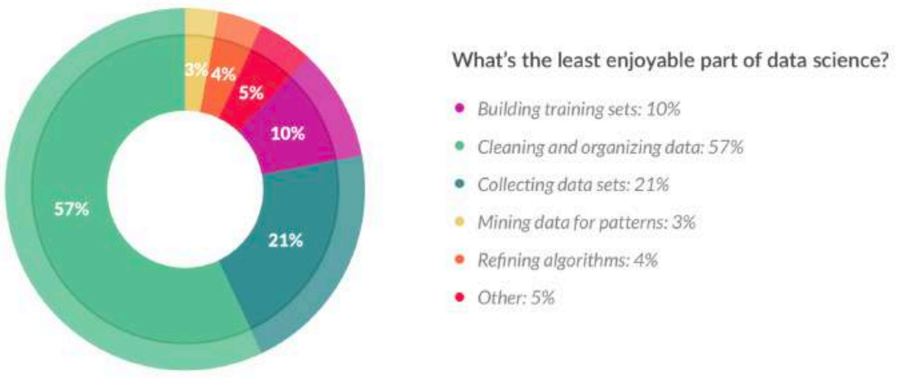
Skill: Analysis, Hacking

What data scientists spend the most time doing

- ● *Building training sets: 3%*
- ● *Cleaning and organizing data: 60%*
- ● *Collecting data sets; 19%*
- ● *Mining data for patterns: 9%*
- ● *Refining algorithms: 4%*
- ● *Other: 5%*

2016 Data Science Report, https://visit.figure-eight.com/data-science-report.html

Skill: Analysis, Hacking

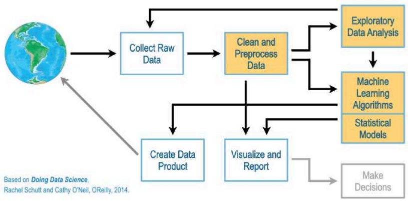


2016 Data Science Report, https://visit.figure-eight.com/data-science-report.html

Data Munging/Data Wrangling

- □ Discovering (with a domain expert)
- □ Structuring (merge, order, reshape)
- □ Cleaning (removing noise, filling gaps, normalizing)
- □ Enriching (add/change data representation)
- □ Validating (check if data “makes sense” / “looks right”)
- □ Publishing (it is a data product!)

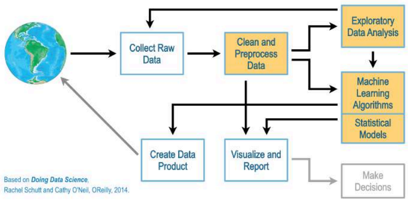
Skill: Analysis



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

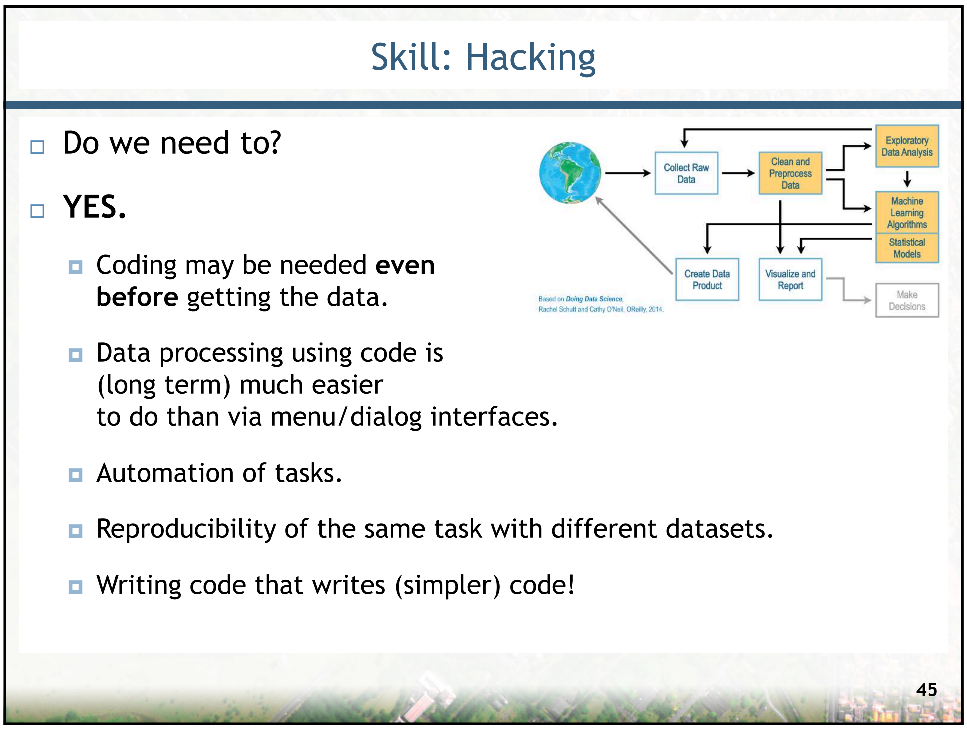
- We have the data. Now what?
  - Do we know what we want to discover?
  - We need basic skills in statistics and data modeling.
- Start exploring: Exploratory Data Analysis
  - Make different plots and charts to explore variables.
  - Get some basic statistics.
  - Evaluate the type of information we can extract from the data.

Skill: 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.

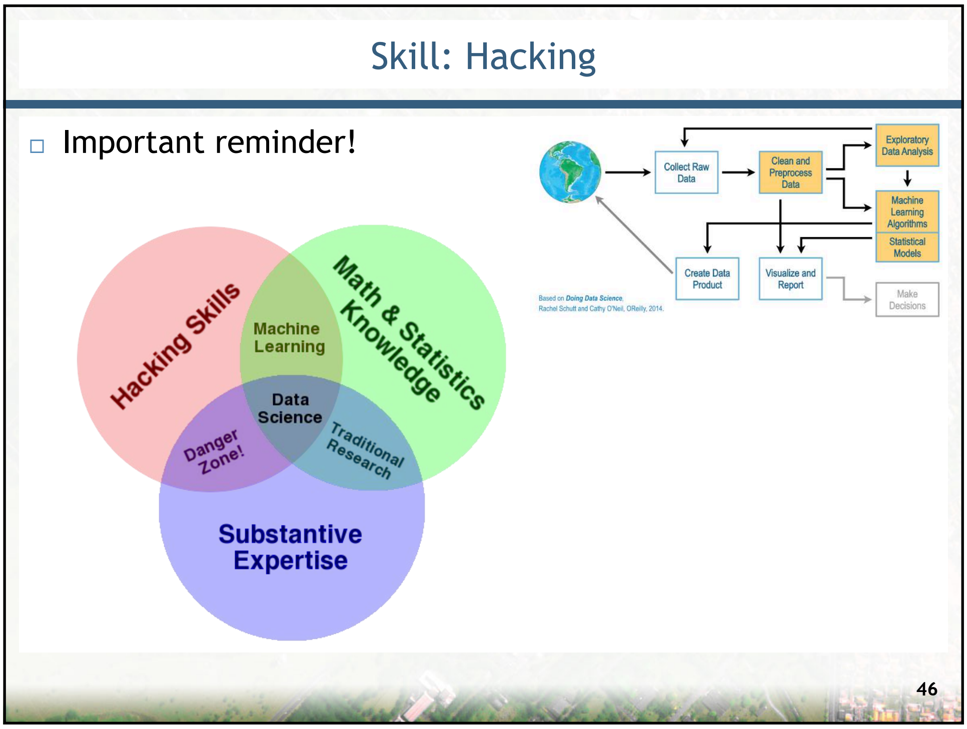


Skill: Hacking

- 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!

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, O'Reilly, 2014.



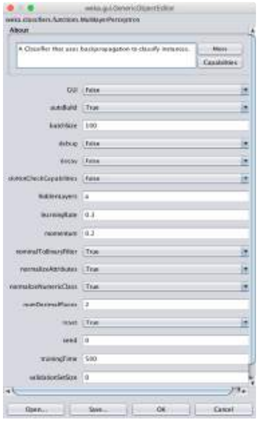
Skill: Hacking

- Important reminder!

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

Skill: Hacking

- This:



weka.classifiers.functions.MultilayerPerceptron

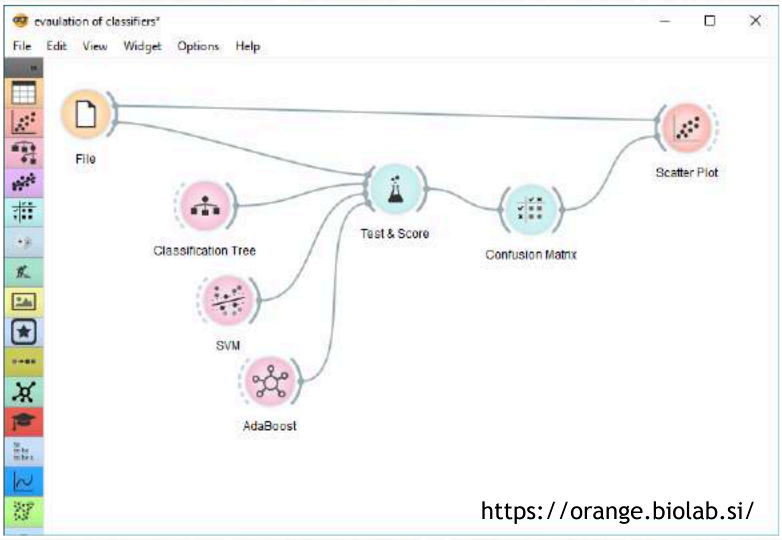
A Classifier that uses backpropagation to classify instances.

```
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



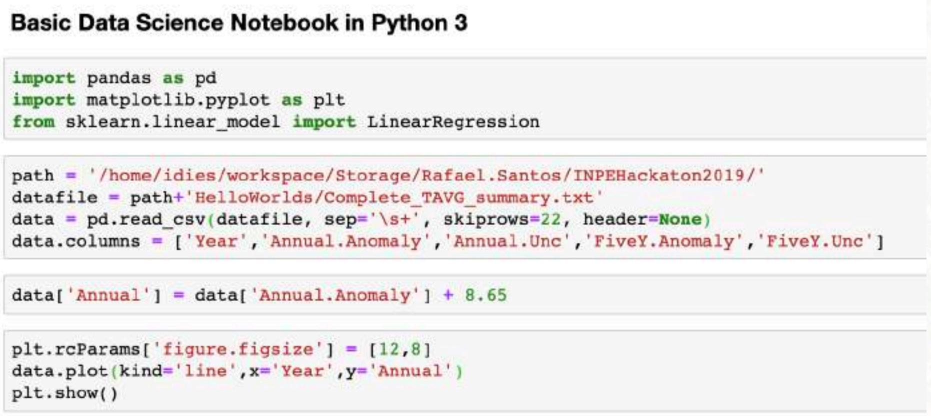
https://orange.biolab.si/

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.*

49

Skill: Hacking Languages: Python



Basic Data Science Notebook in Python 3

```
import pandas as pd
import matplotlib.pyplot as plt
from sklearn.linear_model import LinearRegression
```

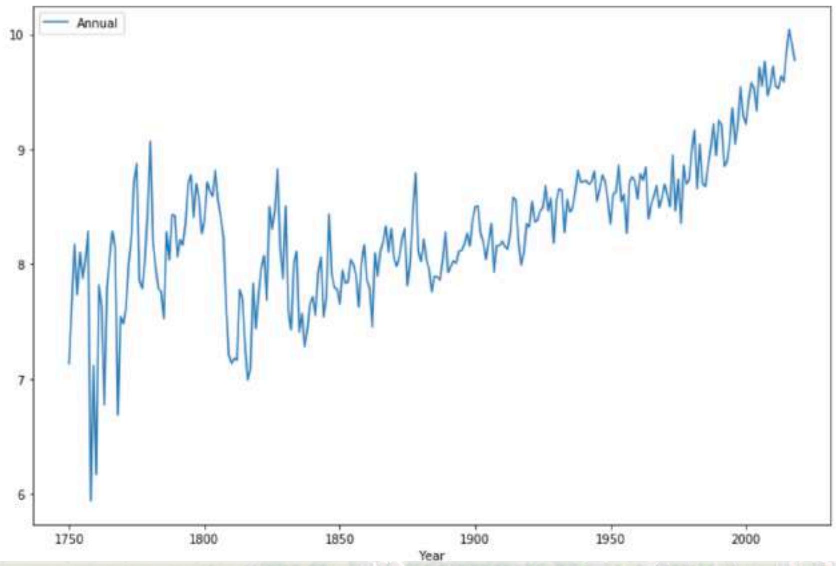
```
path = '/home/idies/workspace/Storage/Rafael.Santos/INPEHackaton2019/'
datafile = path+'HelloWorlds/Complete_TAVG_summary.txt'
data = pd.read_csv(datafile, sep='\s+', skiprows=22, header=None)
data.columns = ['Year','Annual.Anomaly','Annual.Unc','FiveY.Anomaly','FiveY.Unc']
```

```
data['Annual'] = data['Annual.Anomaly'] + 8.65
```

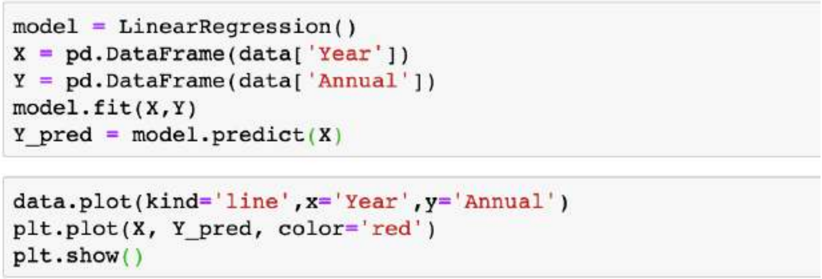
```
plt.rcParams['figure.figsize'] = [12,8]
data.plot(kind='line',x='Year',y='Annual')
plt.show()
```

50

Skill: Hacking Languages: Python



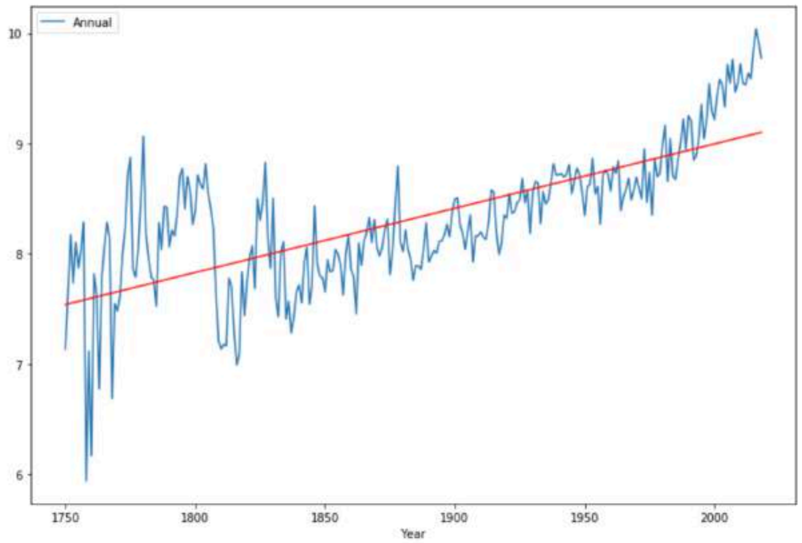
Skill: Hacking Languages: Python



```
model = LinearRegression()
X = pd.DataFrame(data['Year'])
Y = pd.DataFrame(data['Annual'])
model.fit(X,Y)
Y_pred = model.predict(X)
```

```
data.plot(kind='line',x='Year',y='Annual')
plt.plot(X, Y_pred, color='red')
plt.show()
```

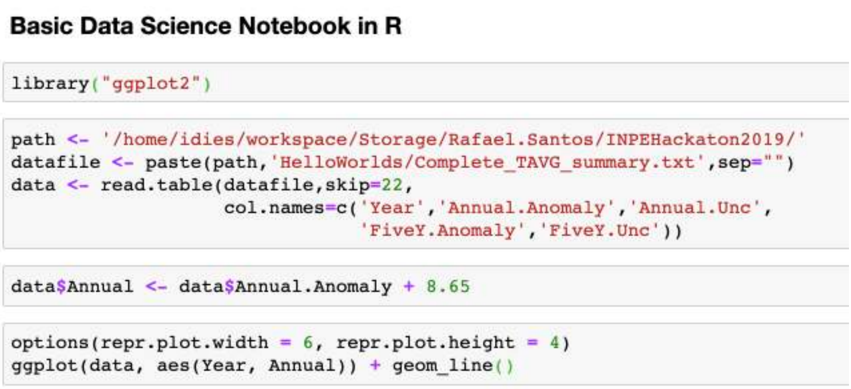
Skill: Hacking Languages: Python



Skill: Hacking Languages: R

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

Skill: Hacking Languages: R



Basic Data Science Notebook in R

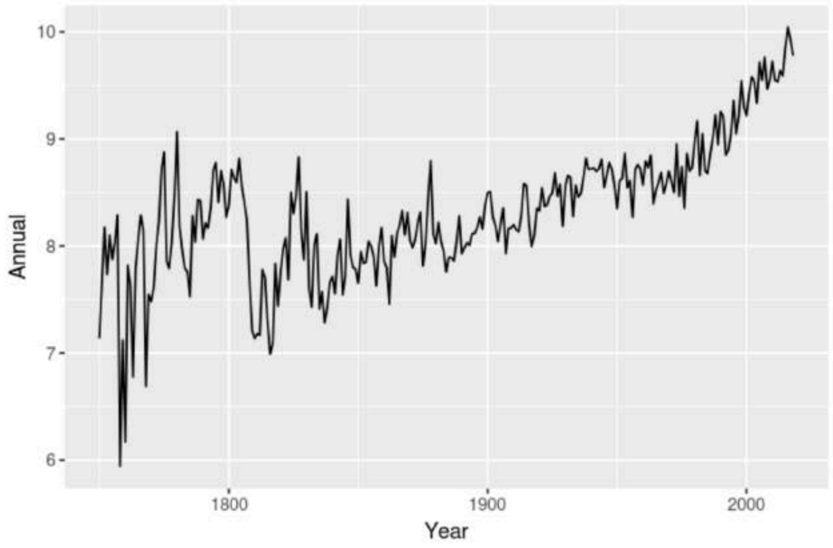
```
library("ggplot2")
```

```
path <- '/home/idies/workspace/Storage/Rafael.Santos/INPEHackaton2019/'
datafile <- paste(path,'HelloWorlds/Complete_TAVG_summary.txt',sep="")
data <- read.table(datafile,skip=22,
                   col.names=c('Year','Annual.Anomaly','Annual.Unc',
                               'FiveY.Anomaly','FiveY.Unc'))
```

```
data$Annual <- data$Annual.Anomaly + 8.65
```

```
options(repr.plot.width = 6, repr.plot.height = 4)
ggplot(data, aes(Year, Annual)) + geom_line()
```

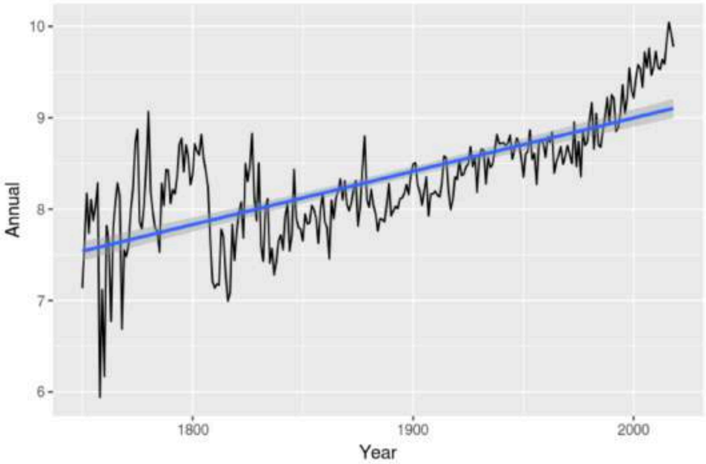
Skill: Hacking Languages: R



Skill: Hacking Languages: R

```
model <- lm(Annual ~ Year, data=data)
```

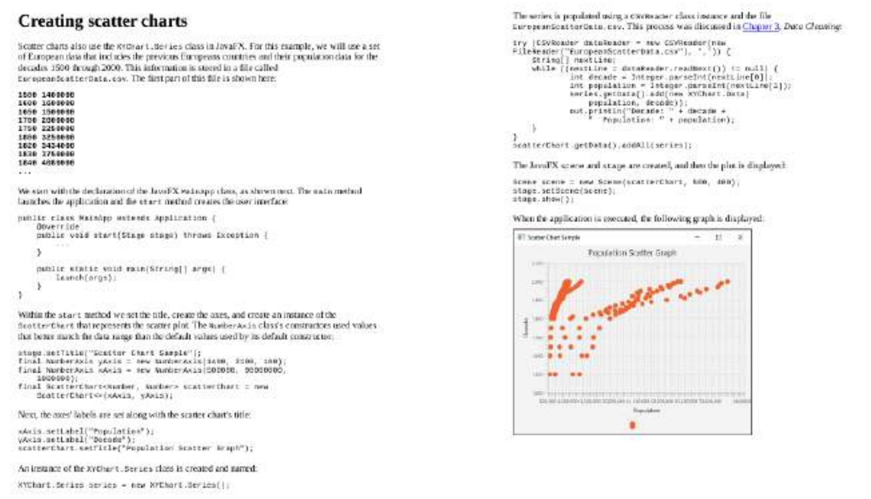
```
ggplot(data, aes(Year, Annual)) + geom_line() + geom_smooth(method='lm')
```



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 2434000
1830 3750000
1840 4000000
...
```

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();
```

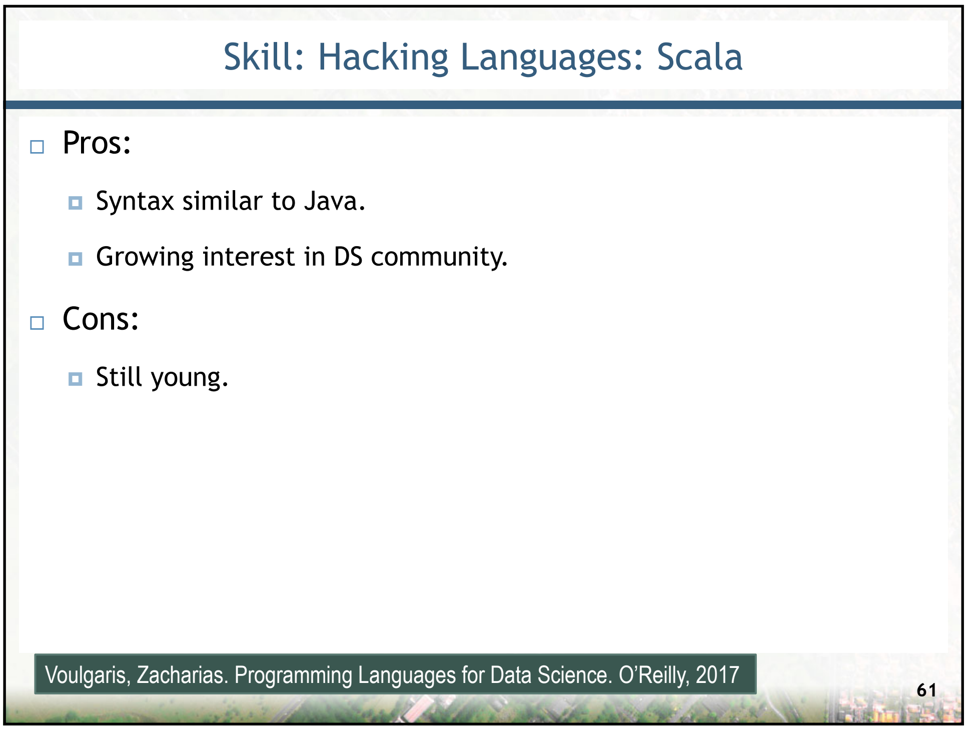
When the application is executed, the following graph is displayed:

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.

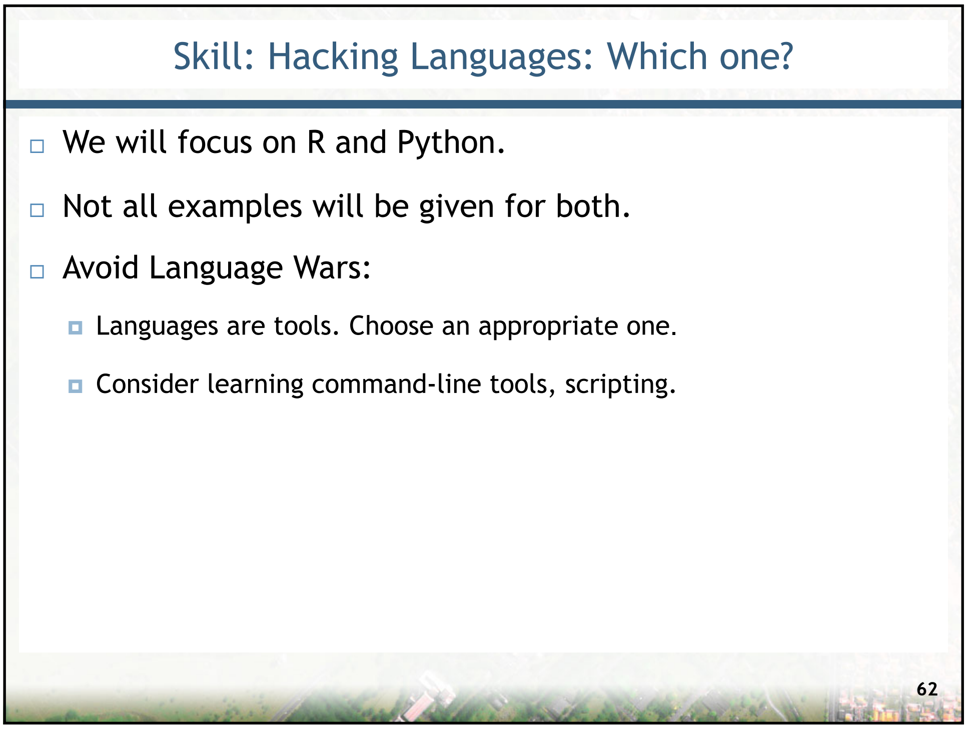
Voulgaris, Zacharias. Programming Languages for Data Science. O'Reilly, 2017



Skill: Hacking Languages: Scala

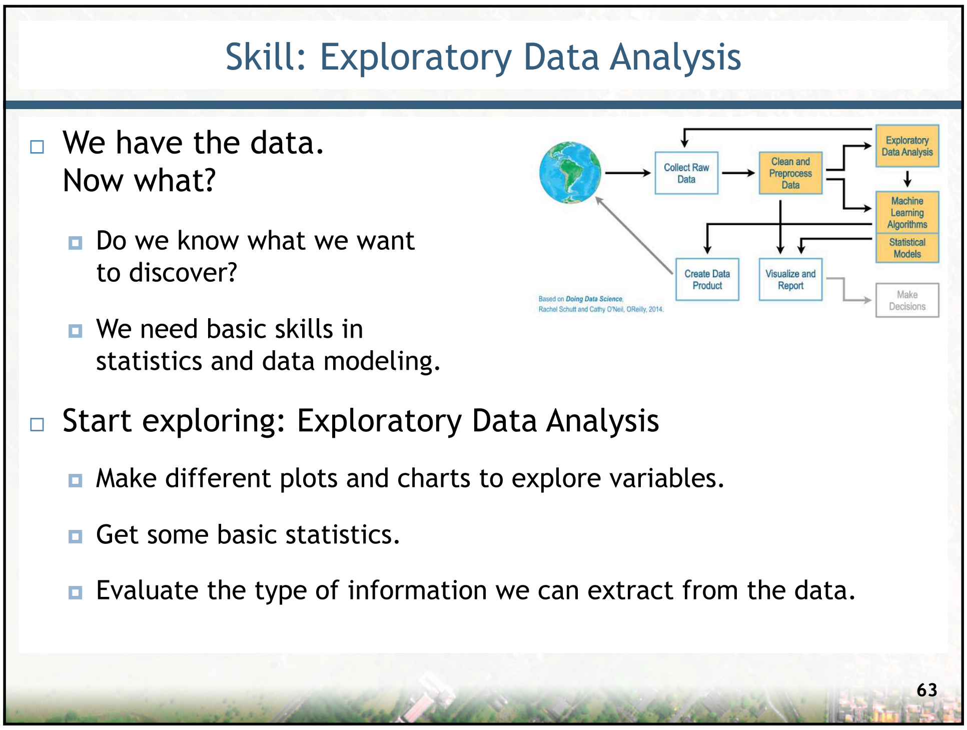
- Pros:
  - Syntax similar to Java.
  - Growing interest in DS community.
- Cons:
  - Still young.

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:
  - Languages are tools. Choose an appropriate one.
  - Consider learning command-line tools, scripting.

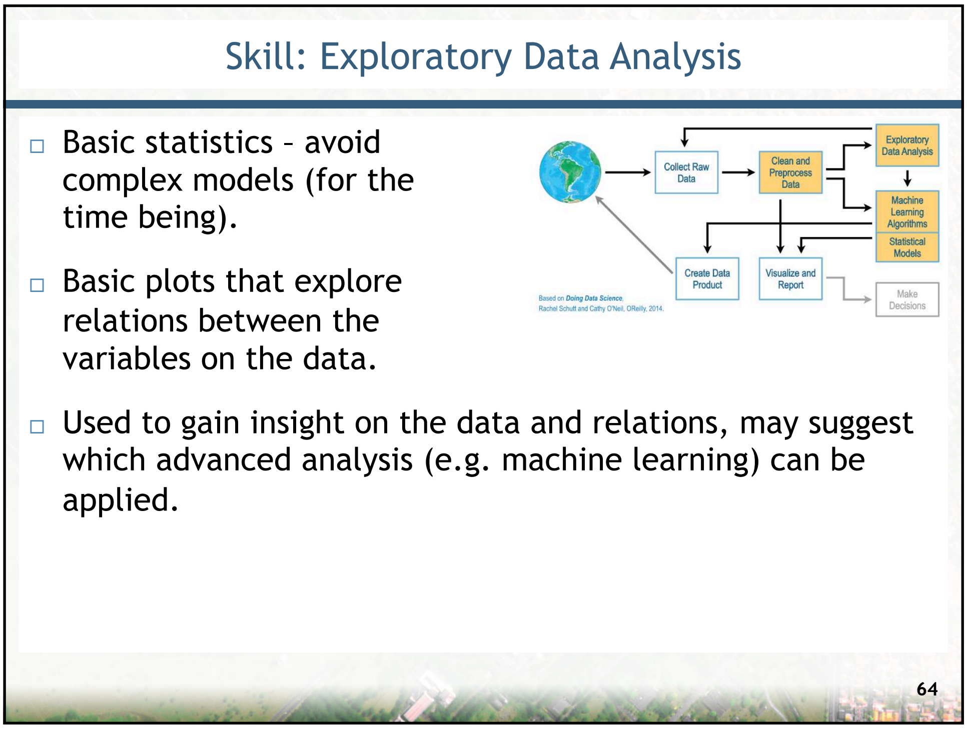


Skill: Exploratory Data Analysis

- 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.

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, O'Reilly, 2014.

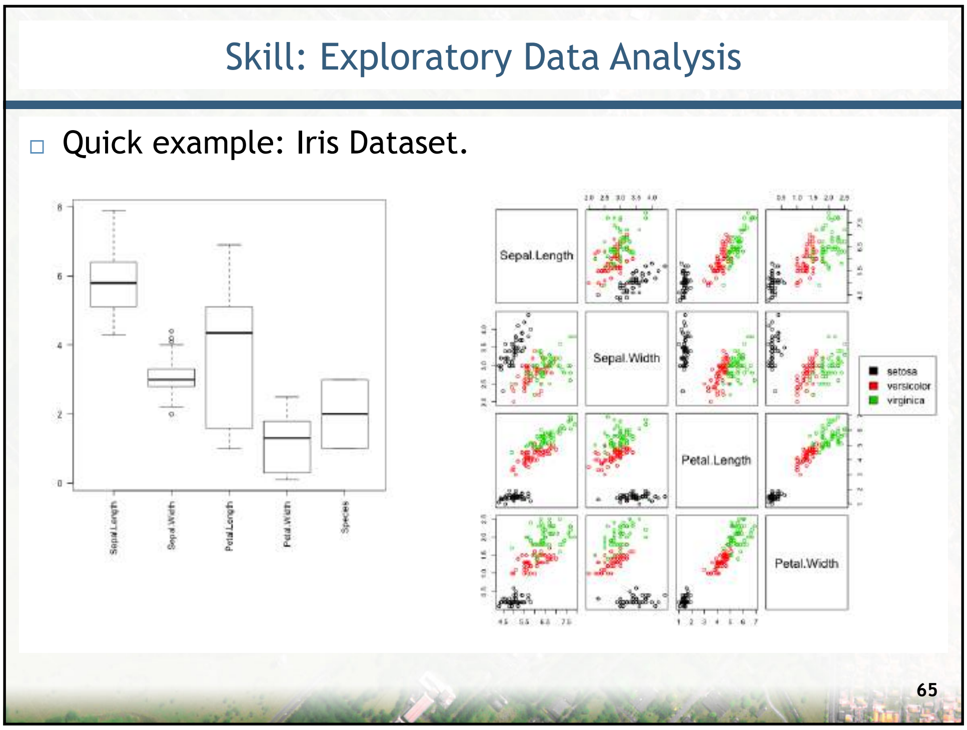


Skill: Exploratory Data Analysis

- 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.

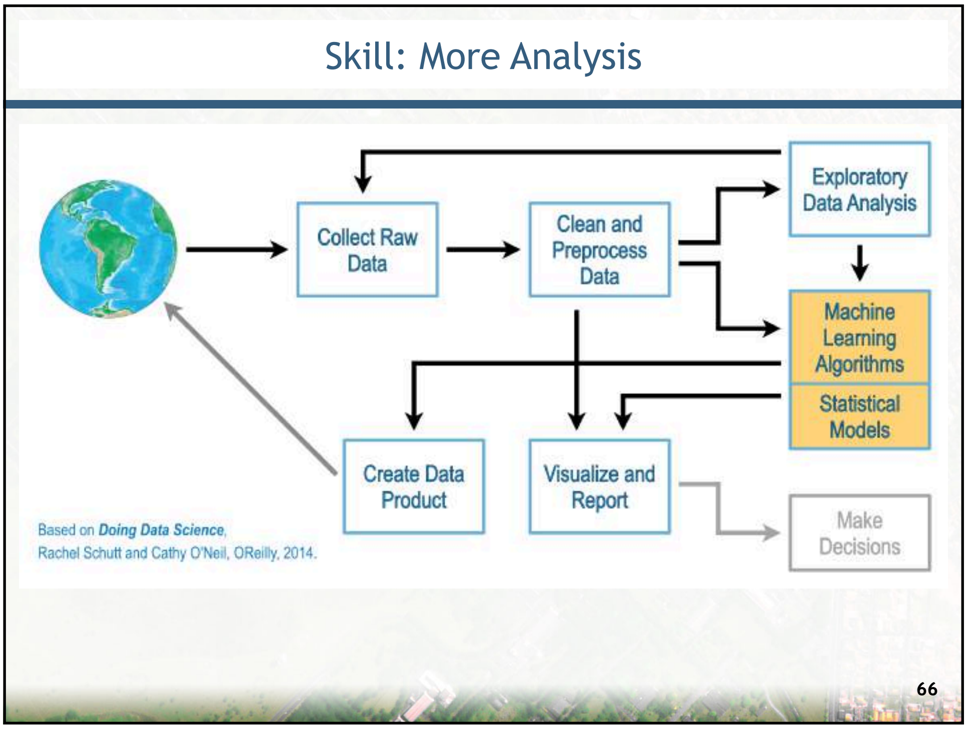
Collect Raw Data
Clean and Preprocess Data
Exploratory Data Analysis
Machine Learning Algorithms
Statistical Models
Create Data Product
Visualize and Report
Make Decisions

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



Skill: Exploratory Data Analysis

□ Quick example: Iris Dataset.



Skill: More Analysis

Collect Raw Data

Clean and Preprocess Data

Exploratory Data Analysis

Machine Learning Algorithms

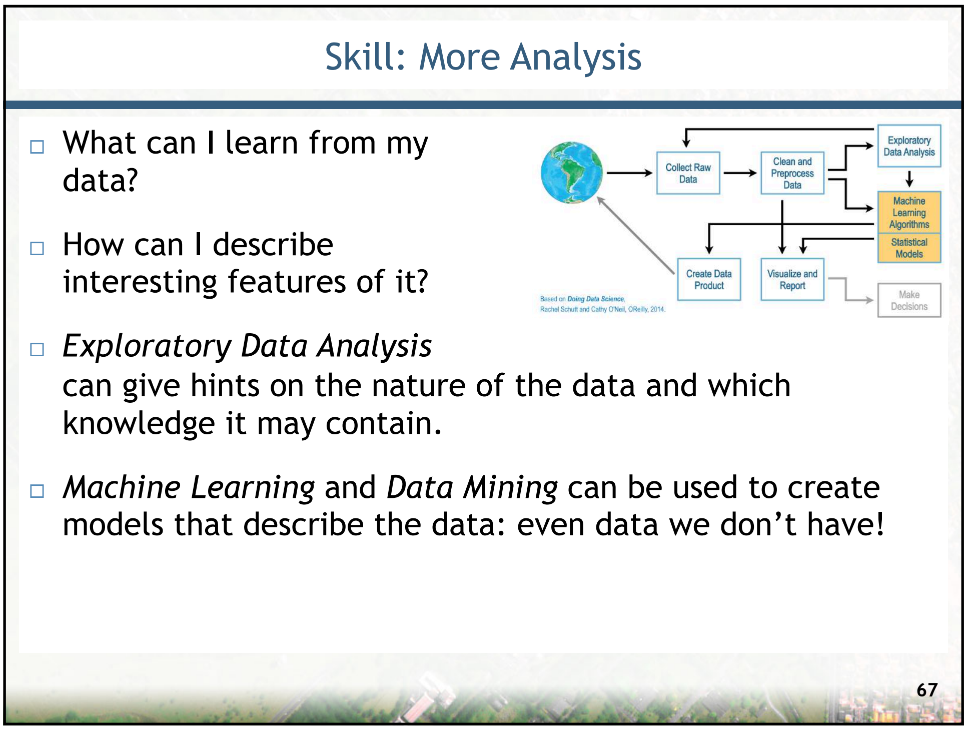
Statistical Models

Create Data Product

Visualize and Report

Make Decisions

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

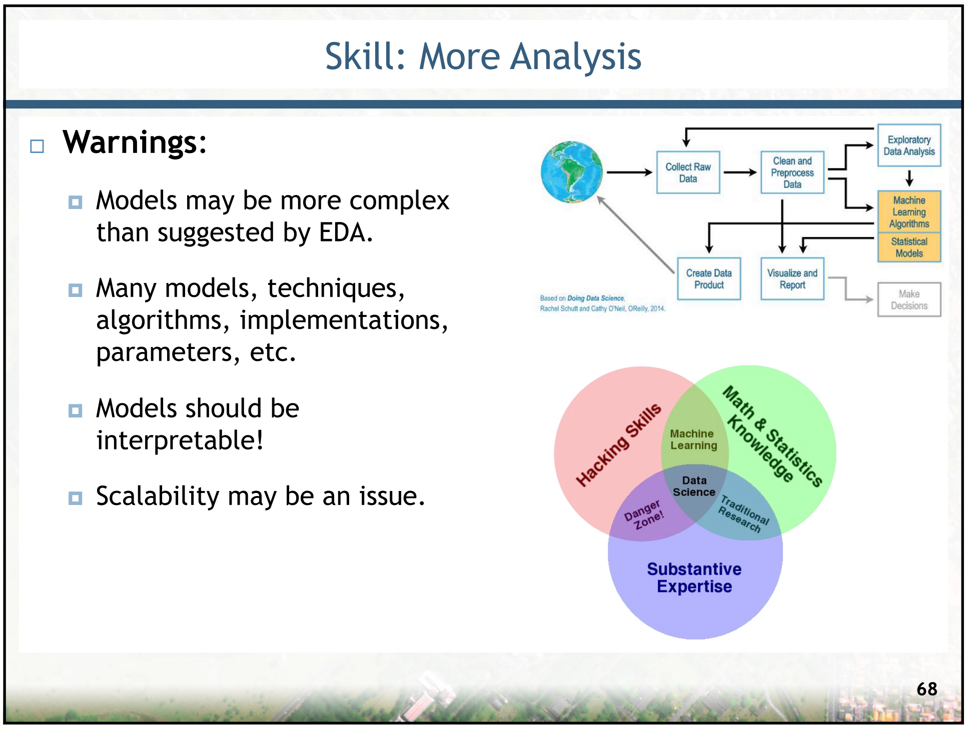


Skill: More Analysis

- 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!

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.



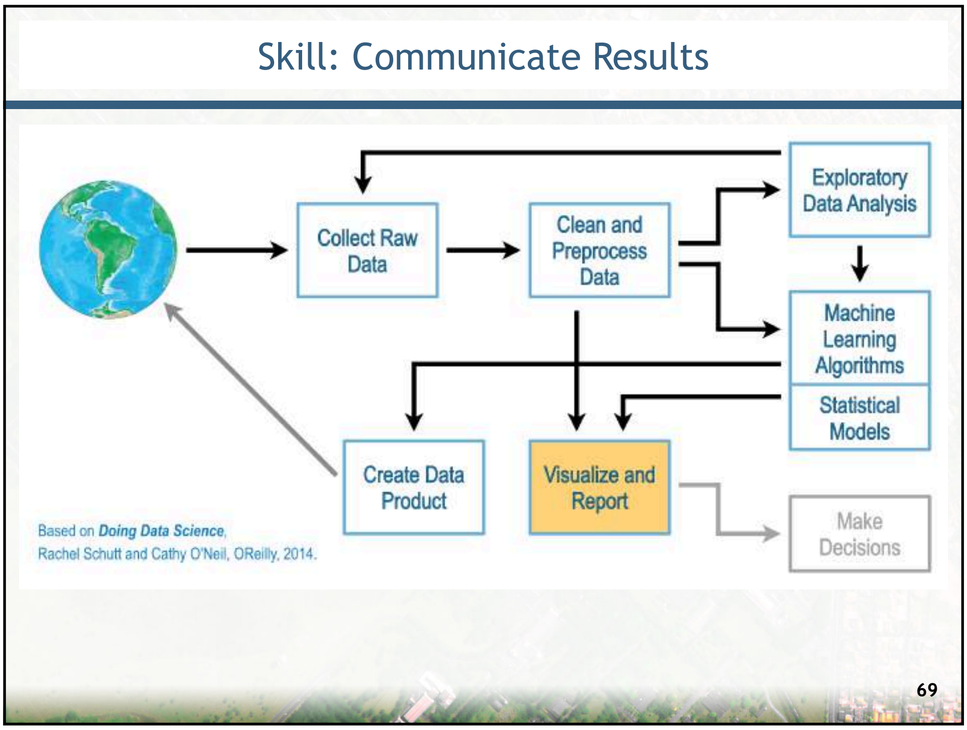
Skill: More Analysis

- **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.

Hacking Skills
Math & Statistics Knowledge
Machine Learning
Data Science
Danger Zone!
Traditional Research
Substantive Expertise



Skill: Communicate Results

Collect Raw Data

Clean and Preprocess Data

Exploratory Data Analysis

Machine Learning Algorithms

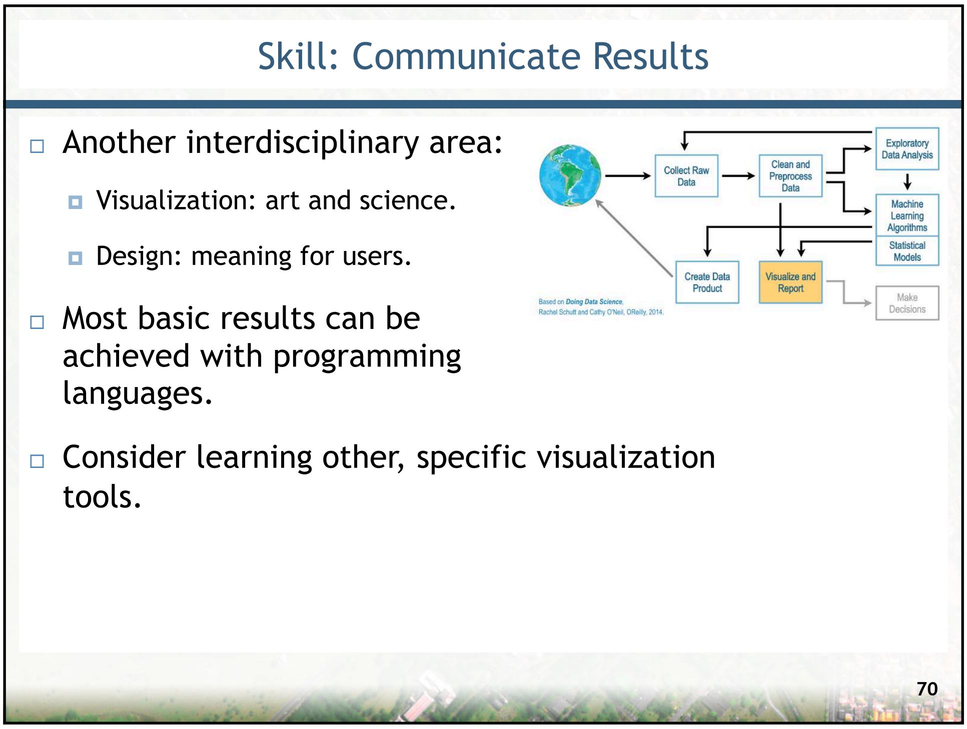
Statistical Models

Create Data Product

Visualize and Report

Make Decisions

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



Skill: Communicate Results

- 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.

Collect Raw Data

Clean and Preprocess Data

Exploratory Data Analysis

Machine Learning Algorithms

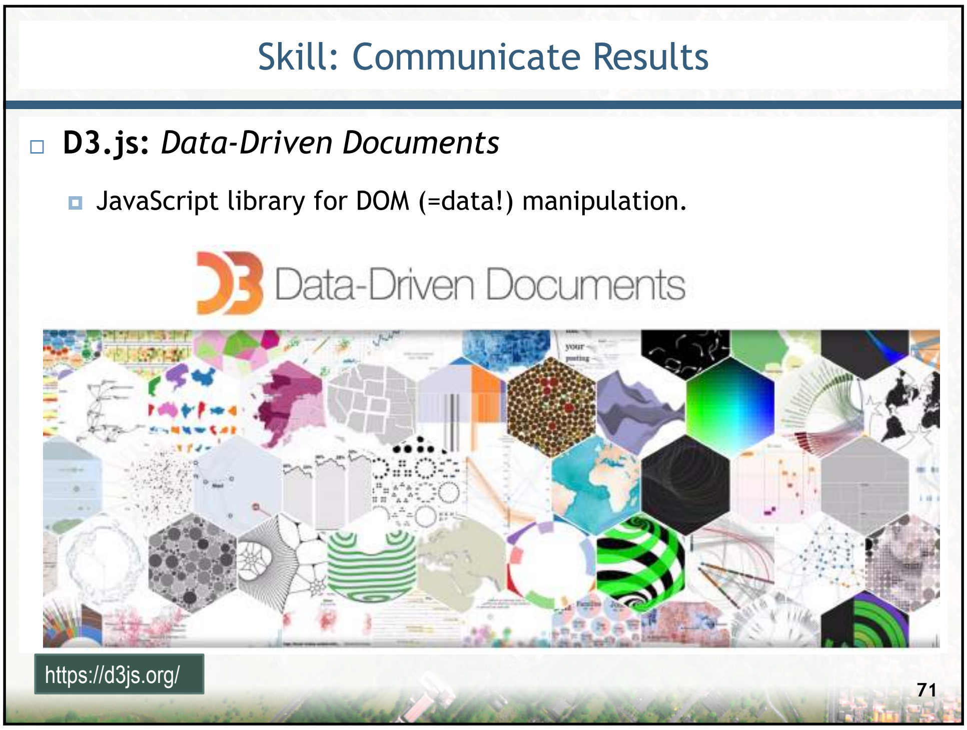
Statistical Models

Create Data Product

Visualize and Report

Make Decisions

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

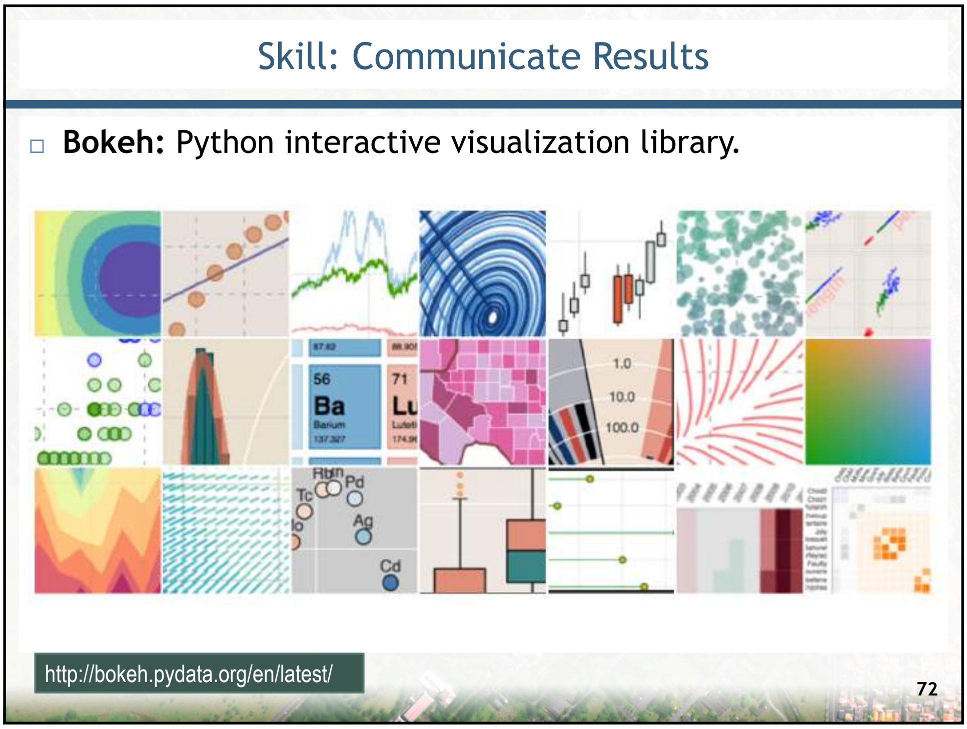


Skill: Communicate Results

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

https://d3js.org/

71

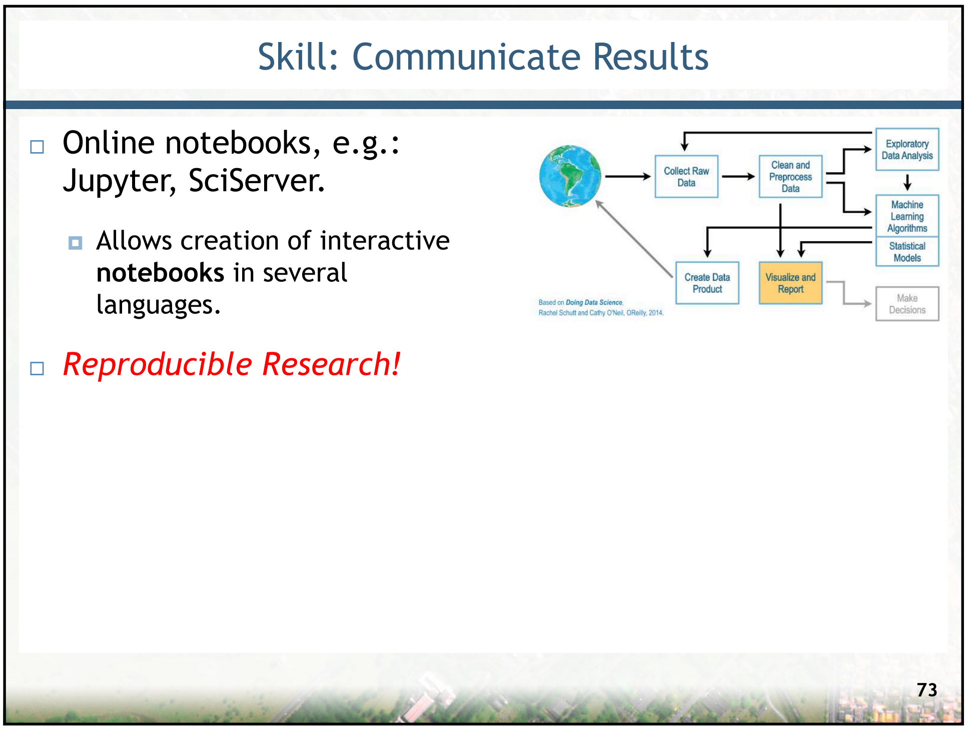


Skill: Communicate Results

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

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

72



Skill: Communicate Results

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

Collect Raw Data

Clean and Preprocess Data

Exploratory Data Analysis

Machine Learning Algorithms

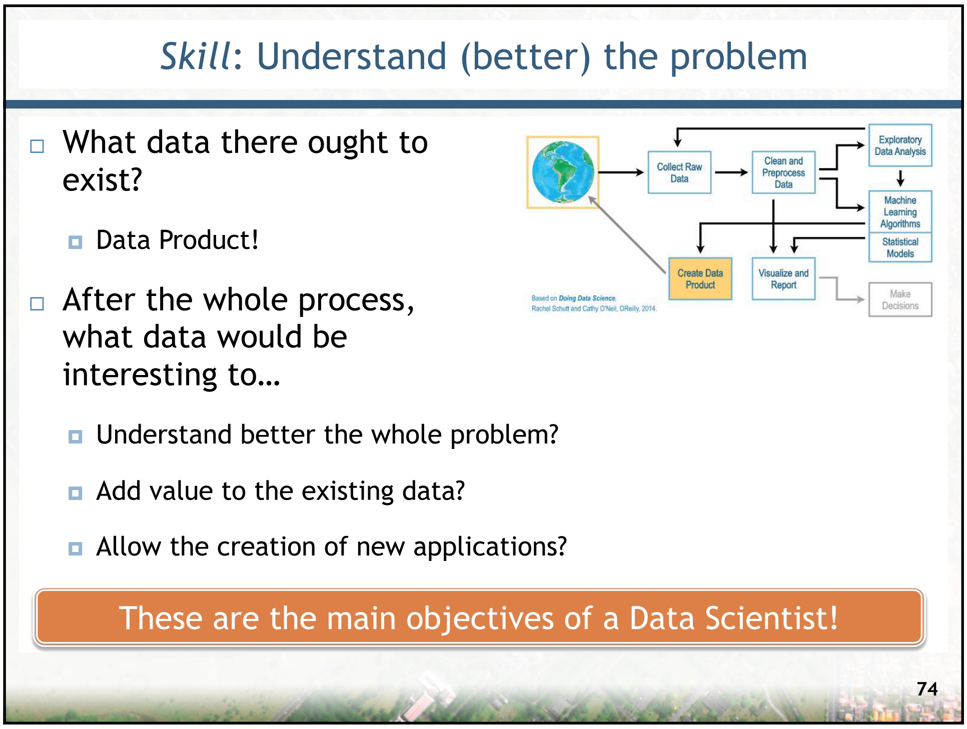
Statistical Models

Create Data Product

Visualize and Report

Make Decisions

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



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?

Collect Raw Data

Clean and Preprocess Data

Exploratory Data Analysis

Machine Learning Algorithms

Statistical Models

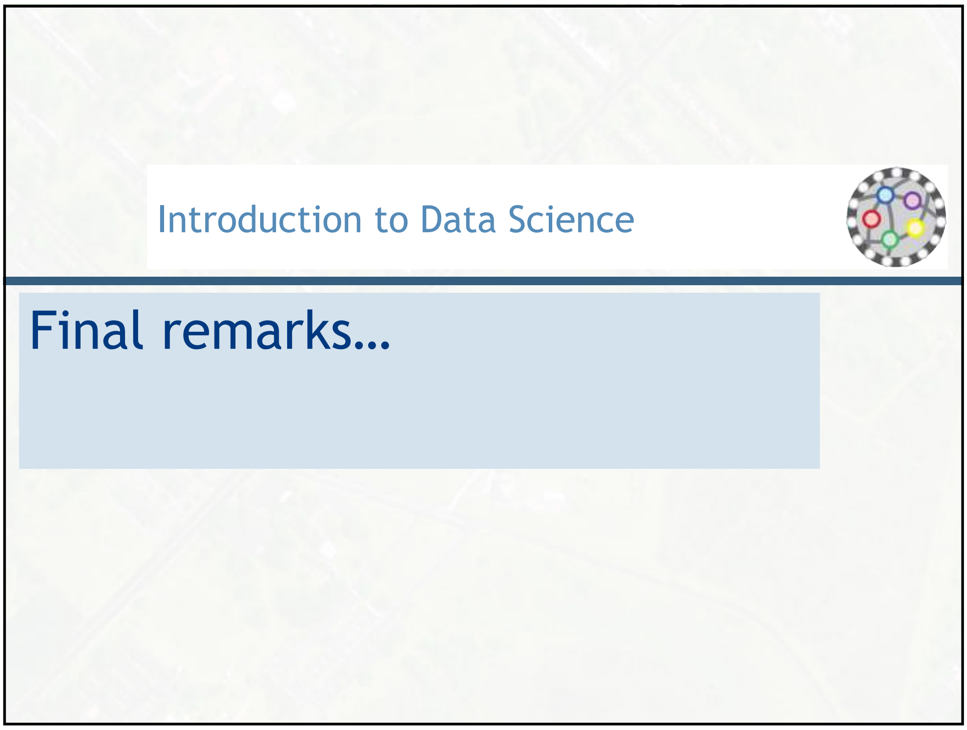
Create Data Product

Visualize and Report

Make Decisions

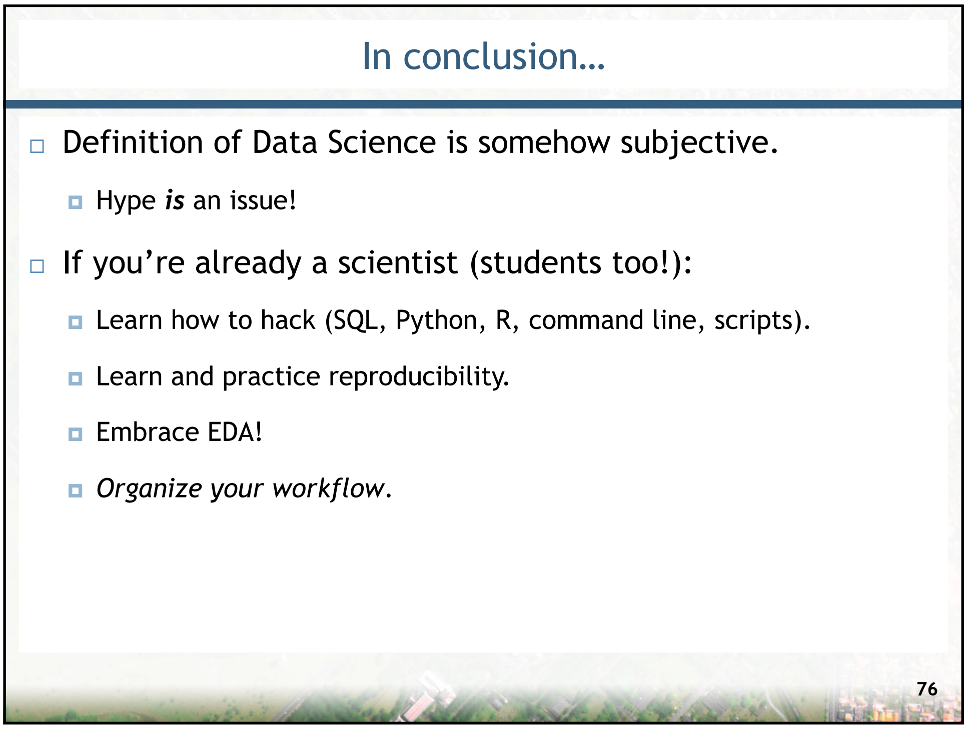
Based on *Doing Data Science*, Rachel Schutt and Cathy O'Neil, ORelly, 2014.

These are the main objectives of a Data Scientist!



Introduction to Data Science

Final remarks…



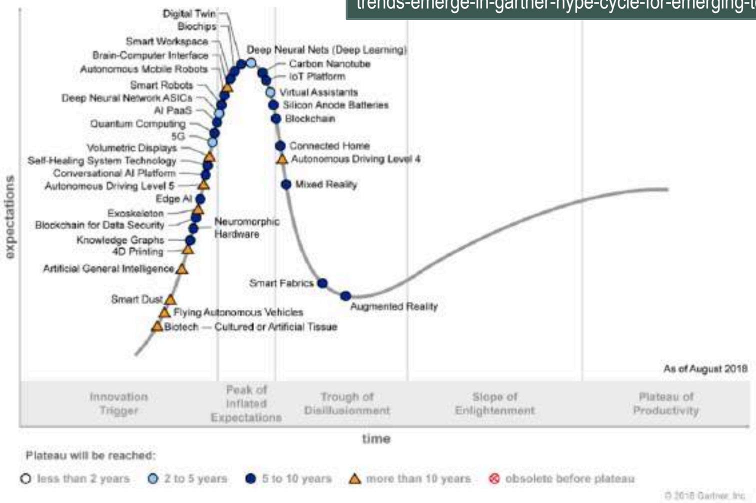
In conclusion…

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

In conclusion…

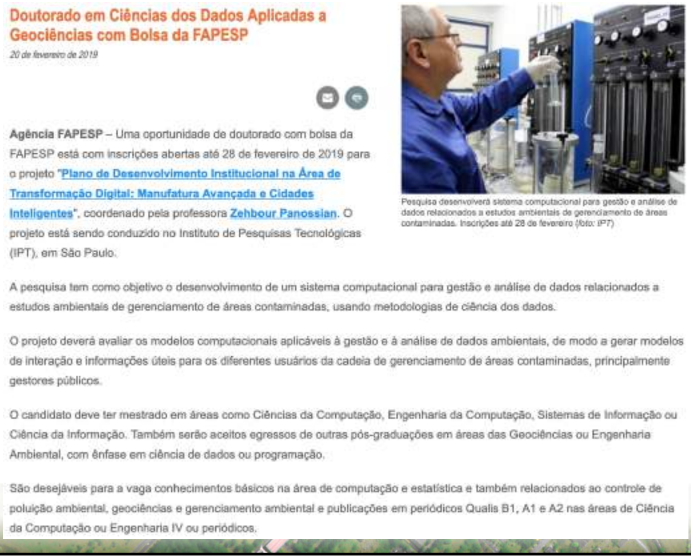
- □ Hype *is* an issue!

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.

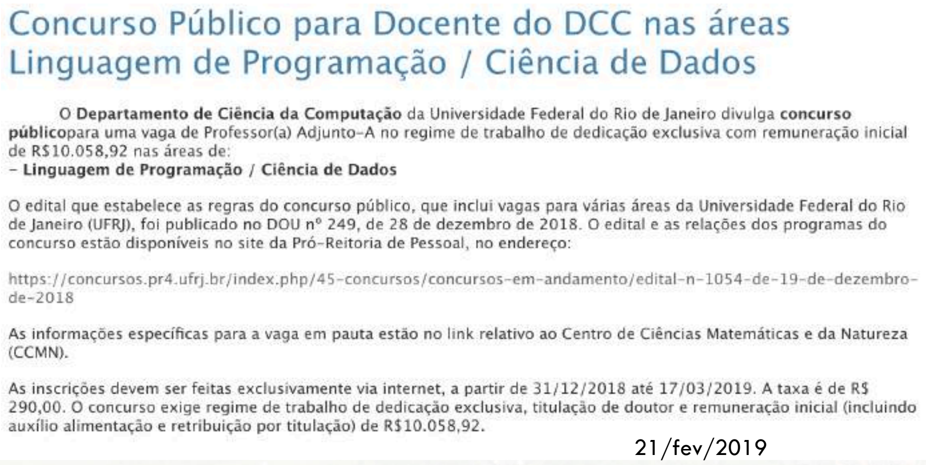
O projeto deverá avaliar os modelos computacionais aplicáveis à gestão e à análise de dados ambientais, de modo a gerar modelos de interação e informações úteis para os diferentes usuários da cadeia de gerenciamento de áreas contaminadas, principalmente gestores públicos.

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 público**para 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:

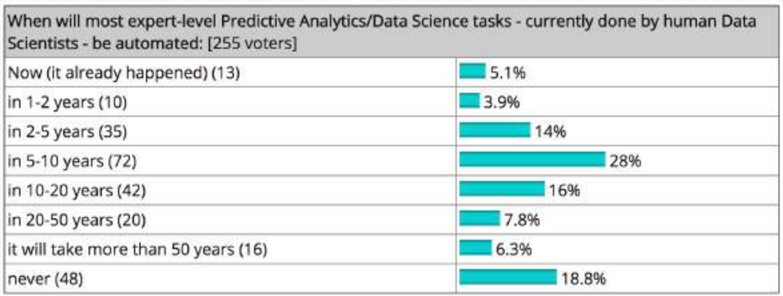
https://concursos.pr4.ufrj.br/index.php/45-concursos/concursos-em-andamento/edital-n-1054-de-19-de-dezembro-de-2018

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

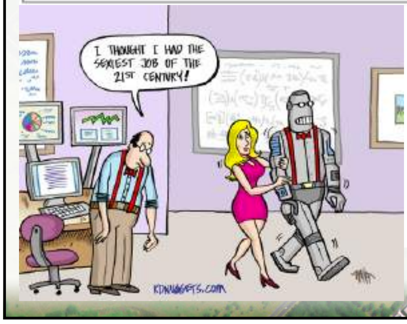
As inscrições devem ser feitas exclusivamente via internet, a partir de 31/12/2018 até 17/03/2019. A taxa é de R$ 290,00. O concurso exige regime de trabalho de dedicação exclusiva, titulação de doutor e remuneração inicial (incluindo auxílio alimentação e retribuição por titulação) de R$10.058,92.

21/fev/2019

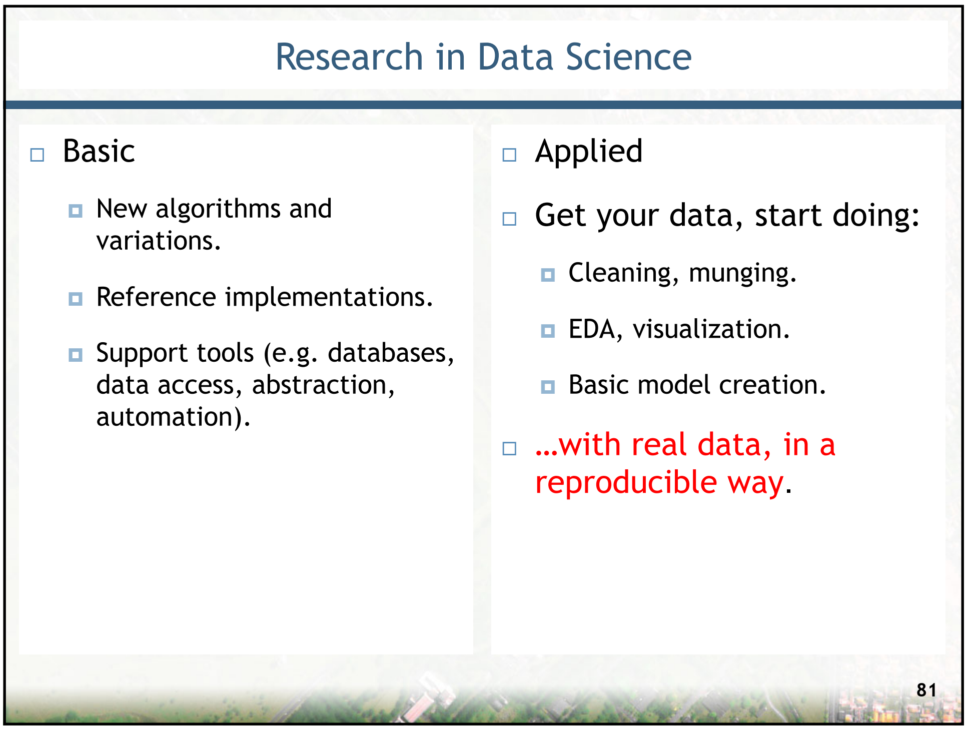
Oh No!



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%



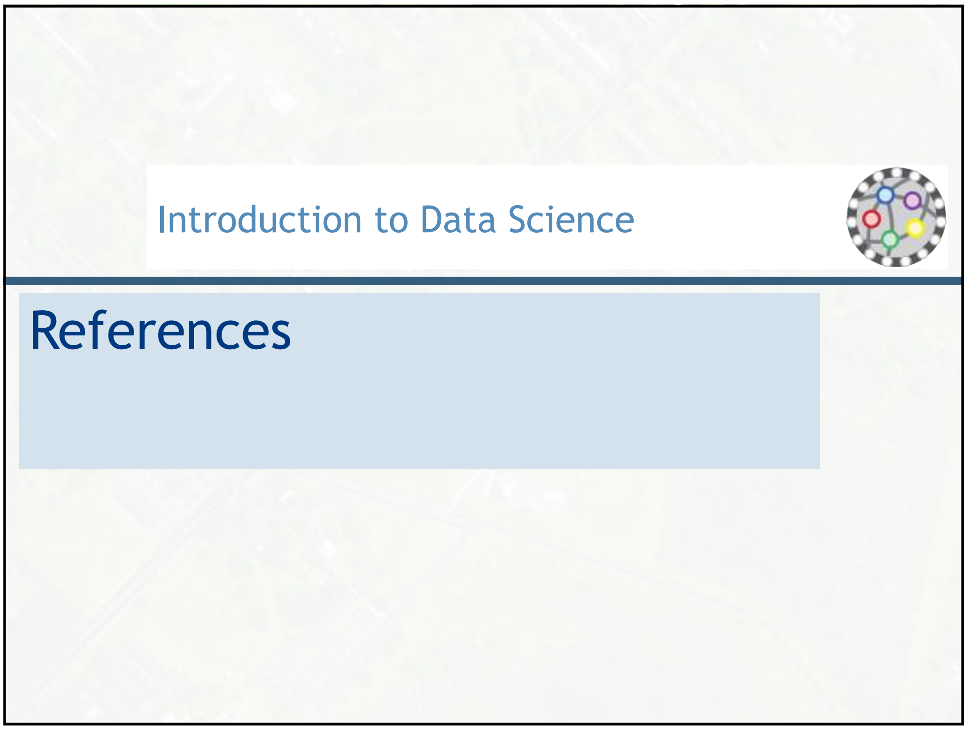
Data Scientists Automated and Unemployed by 2025?
https://www.kdnuggets.com/2015/05/data-scientists-automated-2025.html



Research in Data Science

- □ Basic
  - New algorithms and variations.
  - Reference implementations.
  - Support tools (e.g. databases, data access, abstraction, automation).

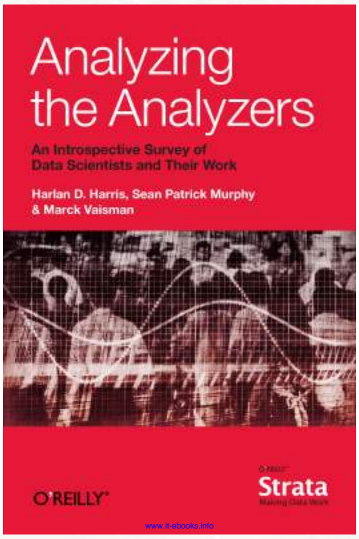
- □ Applied
- □ Get your data, start doing:
  - Cleaning, munging.
  - EDA, visualization.
  - Basic model creation.
- □ ...with real data, in a reproducible way.



Introduction to Data Science

References

References



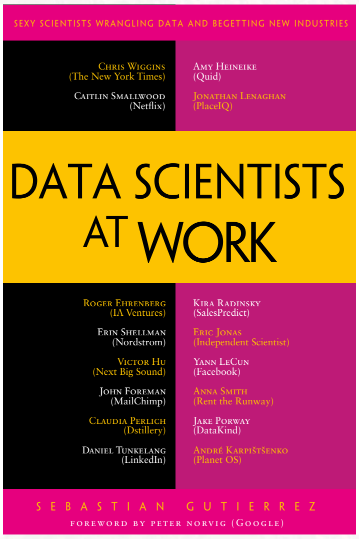
Analyzing the Analyzers

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



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)

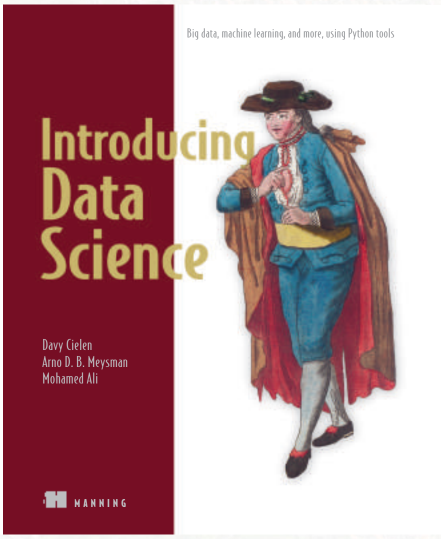
References

O'REILLY

Doing Data Science

STRAIGHT TALK FROM THE FRONTLINE

Rachel Schutt & Cathy O'Neil



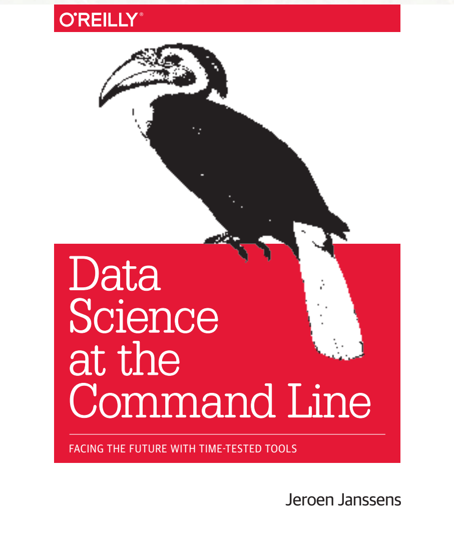
Big data, machine learning, and more, using Python tools

Introducing Data Science

Davy Cielen
Arno D. B. Meysman
Mohamed Ali

MANNING

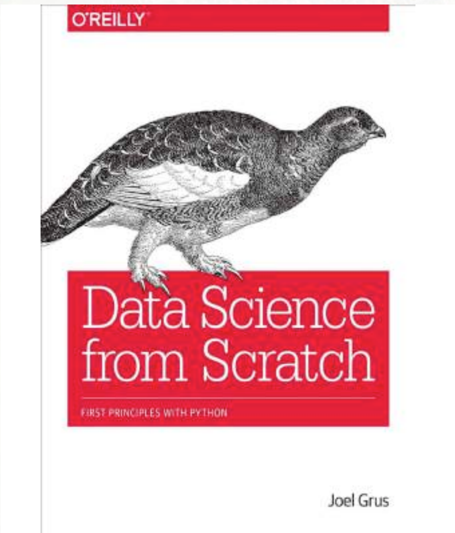
References


O'REILLY

Data Science at the Command Line

FACING THE FUTURE WITH TIME-TESTED TOOLS

Jeroen Janssens

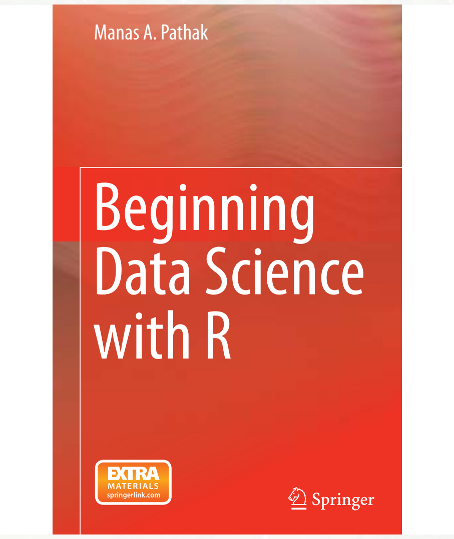

O'REILLY

Data Science from Scratch

FIRST PRINCIPLES WITH PYTHON

Joel Grus

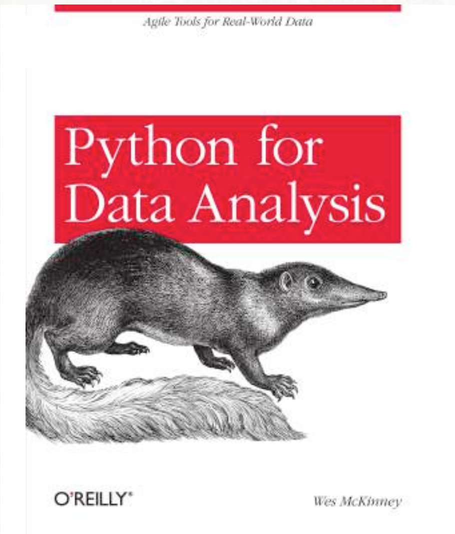
References


Manas A. Pathak

Beginning Data Science with R

EXTRA MATERIALS springerlink.com

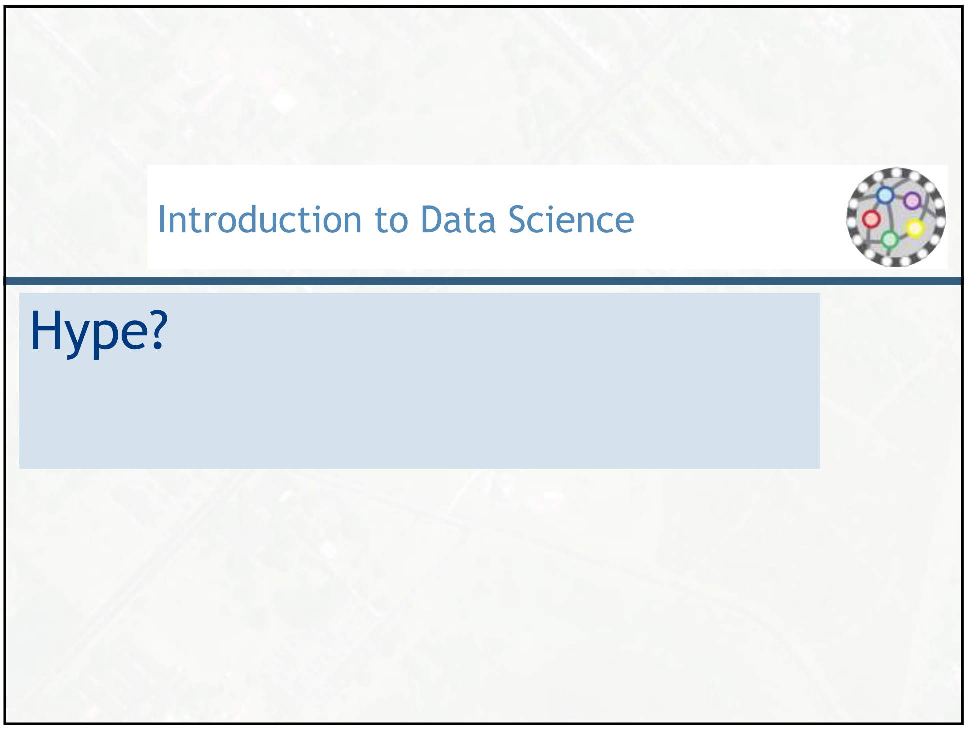
Springer


Agile Tools for Real-World Data

Python for Data Analysis

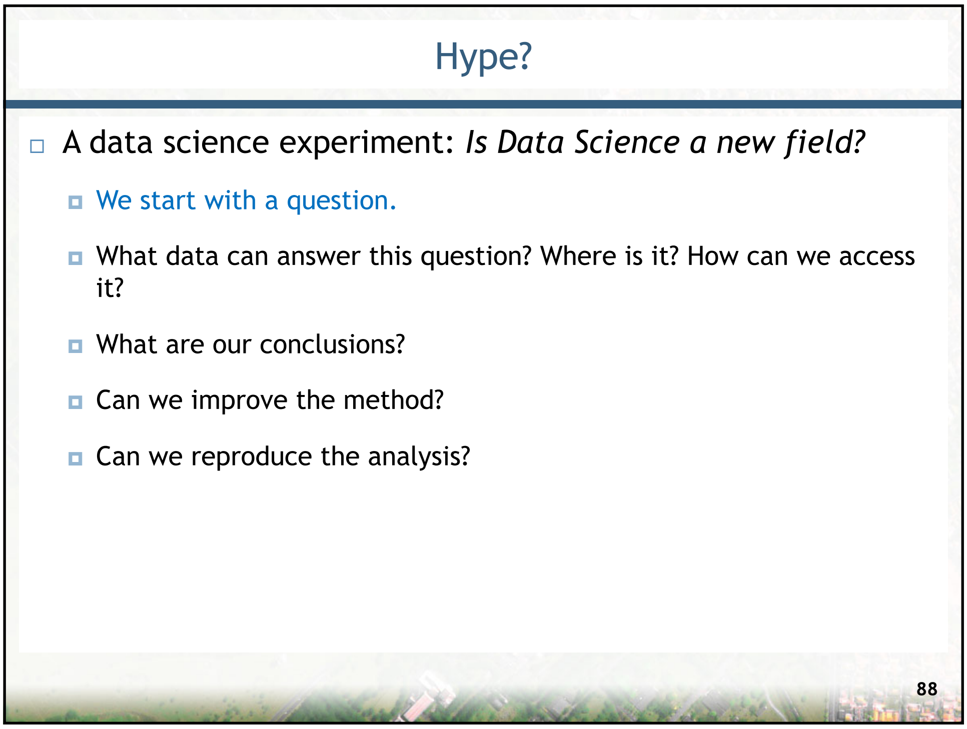
O'REILLY

Wes McKinney



Introduction to Data Science

Hype?

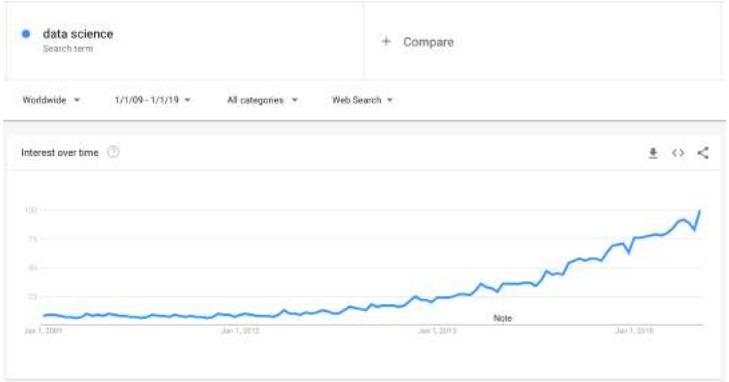


Hype?

- A data science experiment: *Is Data Science a new field?*
  - We start with a question.
  - What data can answer this question? Where is it? How can we access it?
  - What are our conclusions?
  - Can we improve the method?
  - Can we reproduce the analysis?

Is Data Science a new field?

- □ What data can answer this question? Where is it? How can we access it?
- □ Google Trends?



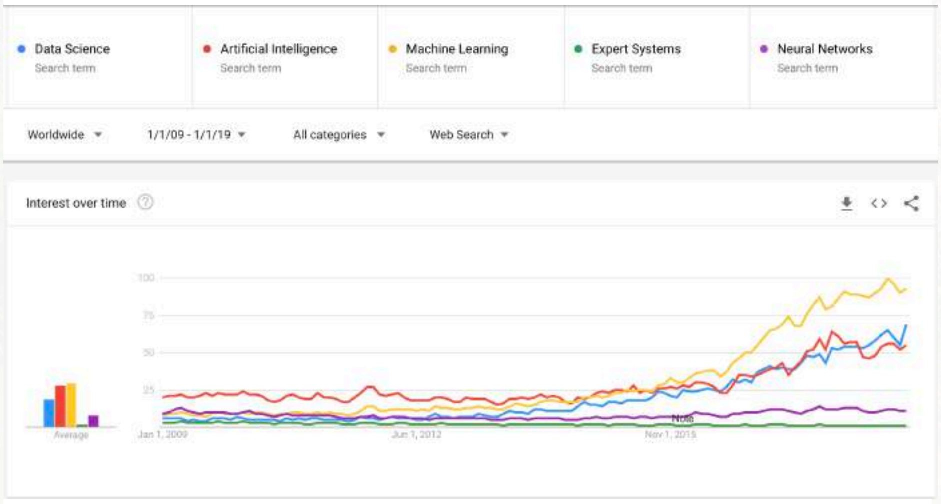
● data science
Search term

+ Compare

Worldwide ▾ 1/1/09 - 1/1/19 ▾ All categories ▾ Web Search ▾

Interest over time

Is Data Science a new field?



● Data Science
Search term

● Artificial Intelligence
Search term

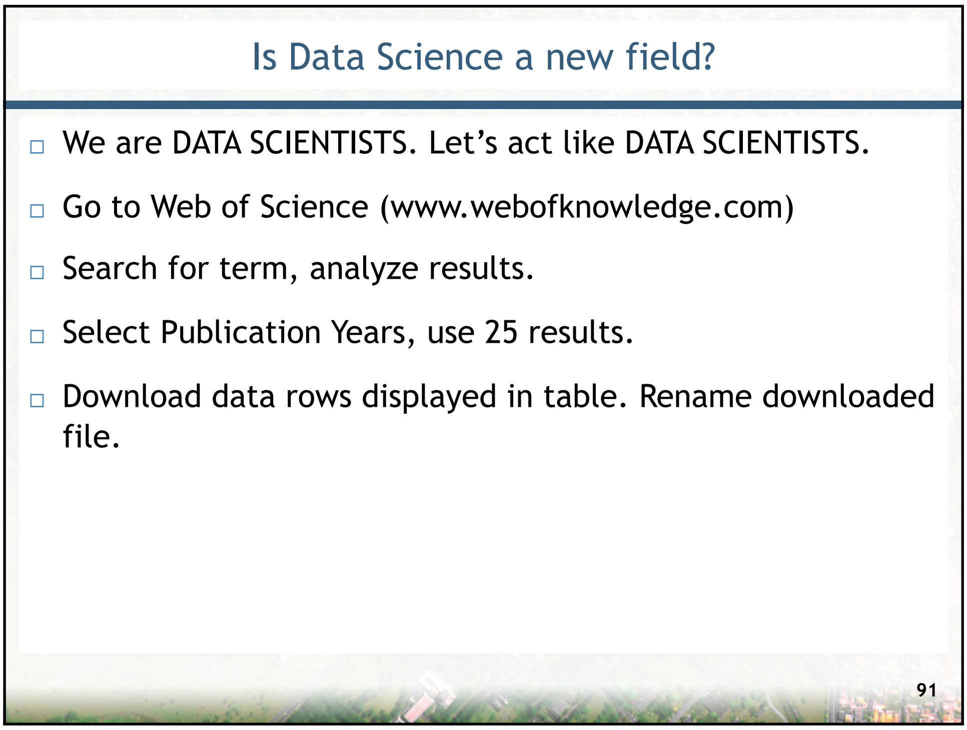
● Machine Learning
Search term

● Expert Systems
Search term

● Neural Networks
Search term

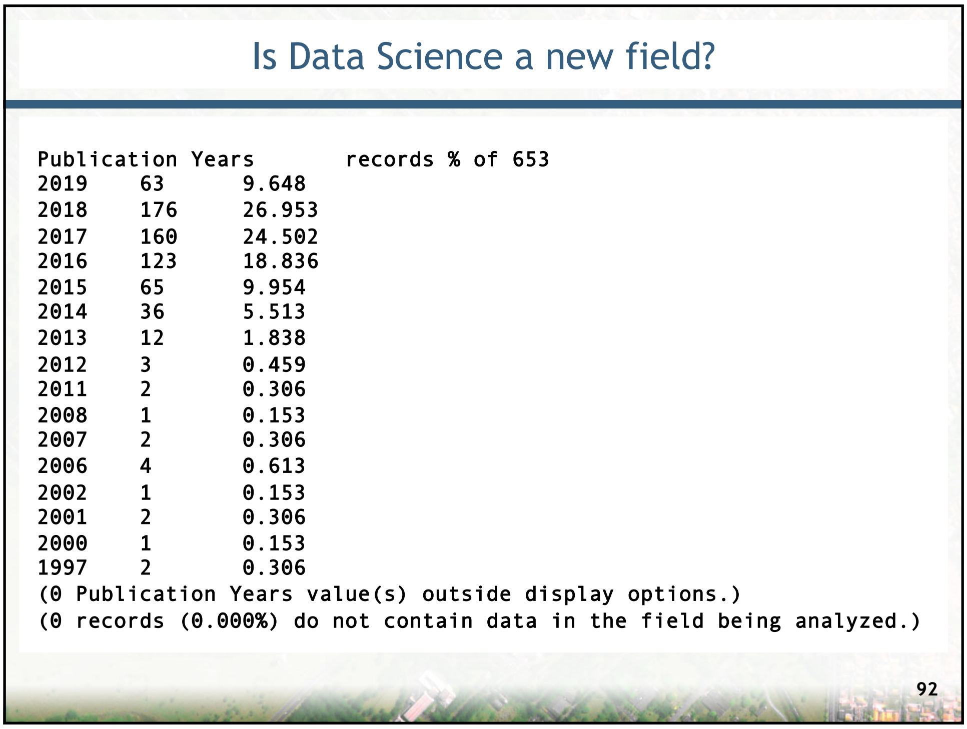
Worldwide ▾ 1/1/09 - 1/1/19 ▾ All categories ▾ Web Search ▾

Interest over time



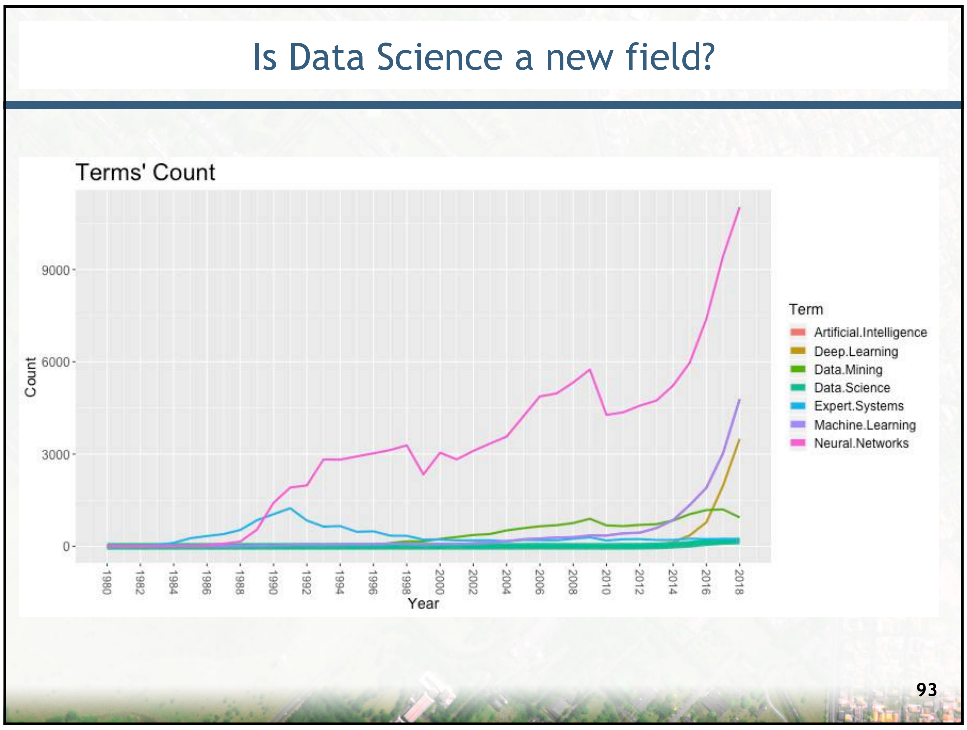
Is Data Science a new field?

- We are DATA SCIENTISTS. Let’s act like DATA SCIENTISTS.
- Go to Web of Science (www.webofknowledge.com)
- Search for term, analyze results.
- Select Publication Years, use 25 results.
- Download data rows displayed in table. Rename downloaded file.

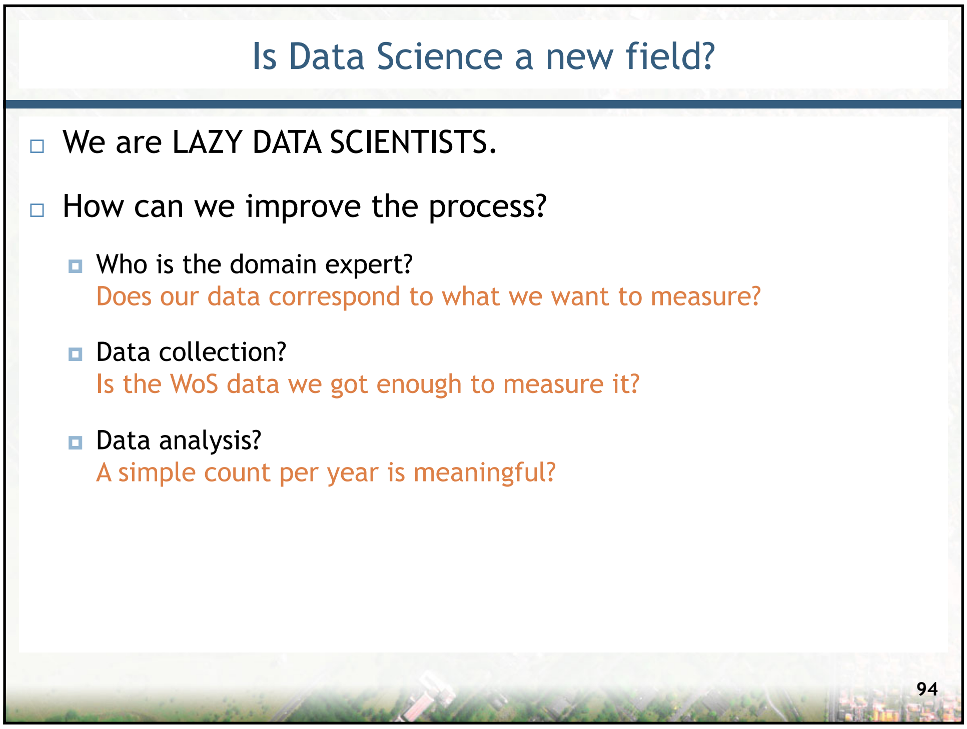


Is Data Science a new field?

```
Publication Years       records % of 653
2019    63      9.648
2018    176     26.953
2017    160     24.502
2016    123     18.836
2015    65      9.954
2014    36      5.513
2013    12      1.838
2012    3       0.459
2011    2       0.306
2008    1       0.153
2007    2       0.306
2006    4       0.613
2002    1       0.153
2001    2       0.306
2000    1       0.153
1997    2       0.306
(0 Publication Years value(s) outside display options.)
(0 records (0.000%) do not contain data in the field being analyzed.)
```

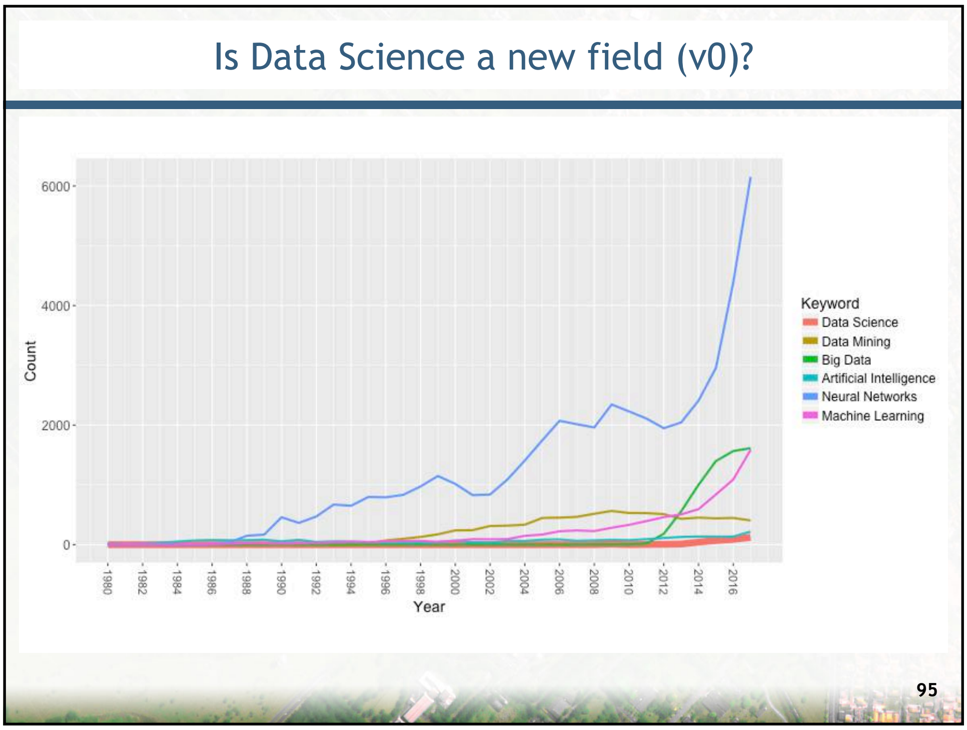


Is Data Science a new field?

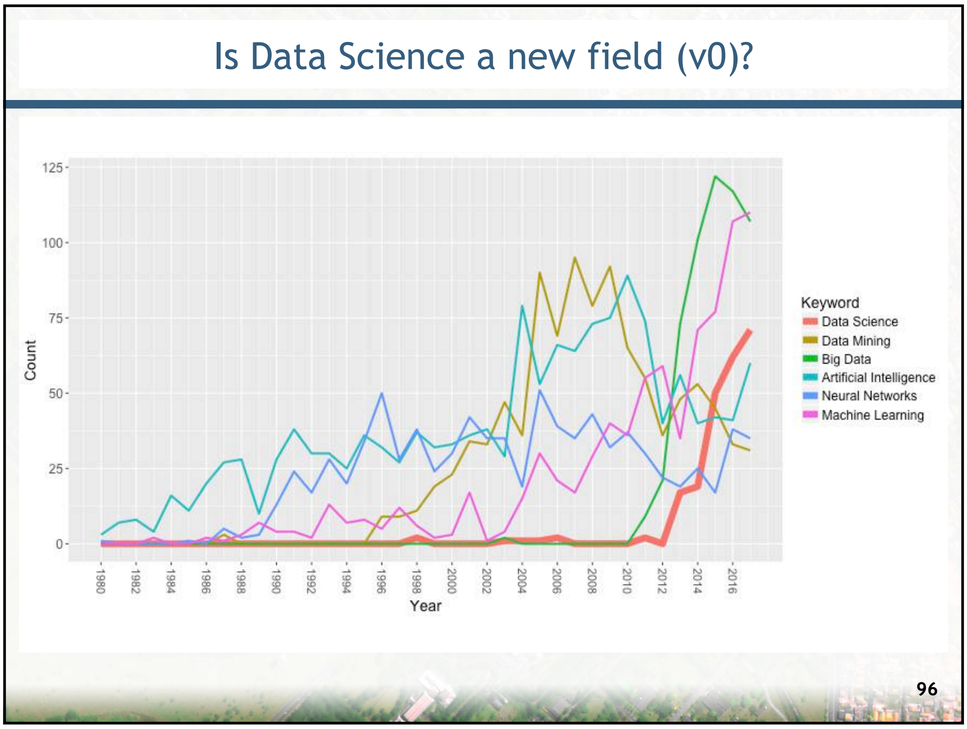


Is Data Science a new field?

- We are LAZY DATA SCIENTISTS.
- How can we improve the process?
  - Who is the domain expert?
    Does our data correspond to what we want to measure?
  - Data collection?
    Is the WoS data we got enough to measure it?
  - Data analysis?
    A simple count per year is meaningful?



Is Data Science a new field (v0)?



Is Data Science a new field (v0)?