

# SBWL 1: Data Processing 1 (PI2.0)

---

Stefan Sobernig

October 6 2020

# Announcements

---

- Check out the [[SBWL Data Science club](#)] at LEARN.
- (6-months free) Access to [[DataCamp](#)]
- Beware! Assignment 1 will be published today (06.10, 18:00).

# **Data Science**

# Data Science

---

- What is Data Science?
- What problems does Data Science address?
- How do Data Scientists work?
- What tools do Data Scientists use?

# What is Data Science?

---

- "There's a joke running around on Twitter that the definition of a data scientist is 'a data analyst who lives in California'— [Malcolm Chisholm?, [@nivertech?](#)]
- "Data Scientist (n.): Person who is better at statistics than any software engineer and better at software engineering than any statistician." [Josh Wills]
- "A data scientist is that **\*\*unique blend of skills\*\*** that can both unlock the insights of data and tell a fantastic story via the data," — [DJ Patil]
- "Data scientists are involved with gathering data, massaging it into a tractable form, making it tell its story, and presenting that story to others," — [Mike Loukides]
- "taking all aspects of life and turning them into data." [Kenneth Neil Cukier and Viktor Mayer-Schoenberger]
- "Data science teams need people with the skills and curiosity to ask the big questions." [DJ Patil]

# What is Data Science?

---

... bottomline: there is no single definition, but some main recurring terms:

- **about "datafication"**
- **unique blend of skills (teamwork!)**
- **gathering data**
- **massaging data**
- **telling a story about the data**

# Datafication

---

A growing area of private and social life become reflected in computerised data to be turned into "valuable" insights.

- user tracking on the Web
- self-quantification
- cyber-physical ("smart") information systems: smart vehicles, smart stores, etc.
- "smart" marketing

... plus some recurring mention of common skills...

# Data Scientists' Skills

---

	<b>Data analyst</b>	<b>Data scientist</b>
<b>Analyt. skills</b>	Analytical thinking Apply established analysis methods	Excellent in math and statistics Visualisation, new approaches
<b>Tech. skills</b>	Data modelling, databases Use of analysis tools Programming skills of advantage	Data modelling, databases Data mining Algorithm development, method abstraction
<b>Domain knowledge</b>	Detailed domain knowledge Project management Communication skills	Background domain knowledge Creativity Team work



# Data Scientists' Skills

---

*"3 sexy skills of data geeks"* (Nathan Yau, [Rise of the Data Scientist](#), 2009)

- Statistics (data analyses as known to you; see course on "Data Analytics")
- Visualization (plots, visualisation tooling like dashboards, etc.; Data Science Lab)
- **Data munging (scraping, parsing, formatting, and cleaning data) (This course)**

# What problems does Data Science address?

## Example for data journalism

- focus on politics, economics and sports
- Who will win the presidency? (USA 2016)
- Should Travelers Avoid Flying Airlines That Have Had Crashes in the Past?
- A Statistical Analysis of the Work of Bob Ross
- 2014 World Cup Predictions

The screenshot shows the FiveThirtyEight website interface. At the top, there is a navigation bar with categories: Politics, Sports, Science & Health, Economics, and Culture. The main content area features a large article titled "World Cup Anxiety Reaches Its Boiling Point For The USMNT" by Michael Caley, accompanied by a photo of soccer players. To the right, there are sections for "THE LATEST" with several article teasers, and "INTERACTIVES" which includes "NFL Predictions" (updated 2 days ago) and "MLB Predictions" (updated 4 hours ago). The NFL predictions table is as follows:

Game	Probability
Detroit over Carolina	63%
Philadelphia over Arizona	65%
Pittsburgh over Jacksonville	80%
Buffalo over Cincinnati	51%

Below the NFL predictions is a button that says "See all NFL teams and games". The MLB predictions section is partially visible at the bottom of the screenshot.

Dataset for published articles

# **Data Science as a Process**

# What does a Data Science process look like?

---

## Example of a "classic" data-driven process: ETL in dataware housing

- ETL refers to a process in database usage and especially in data warehousing that:
- *Extracts* data from homogeneous or heterogeneous data sources
- *Transforms* the data for storing it in proper format or structure for querying and analysis purpose (includes cleansing of deduplications, inconsistencies, dealing with missing data,...)
- *Loads* it into the final target (database, more specifically,

operational data store, data mart, or data warehouse)

See., e.g. *Matteo Golfarelli, Stefano Rizzi. Data Warehouse Design: Modern Principles and Methodologies. McGraw-Hill, 2009.*

# What does a Data Science process look like?

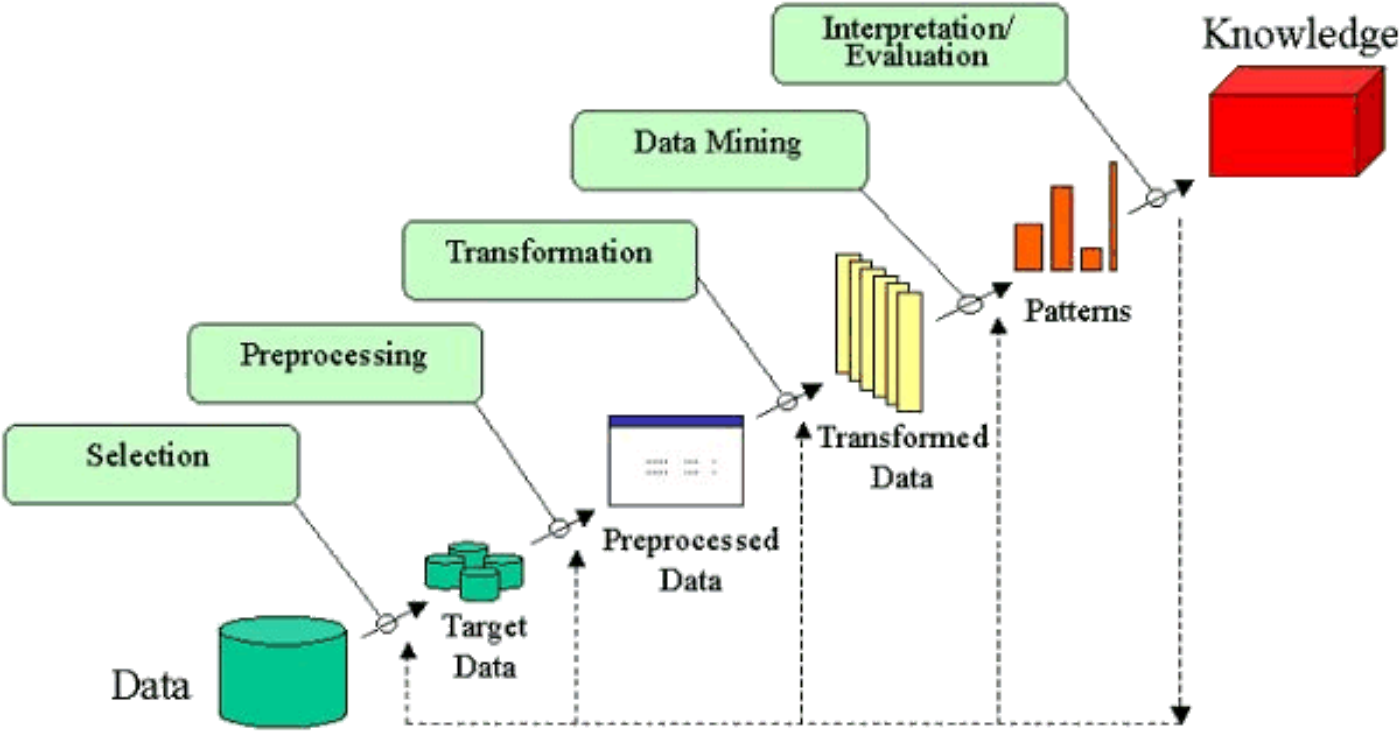
---

"Classic" views are challenged by datafication:

- The "classic view" typically assumes: **fixed, static processing pipelines** vs. iterative, dynamic pipelines in DS
- Typically assumes **complete/clean data** at the "load" stage vs. messy data in DS
- *Data cleansing* sometimes viewed as a part of a Transform step, sometimes not

# What does a Data Science Process look like?

"Knowledge Discovery in Databases (KDD)" process (often used in the course of Data Mining)



Source: [Howard Hamilton](#)

# What does a Data Science Lifecycle look like?

Towards a "Data Science workflow"

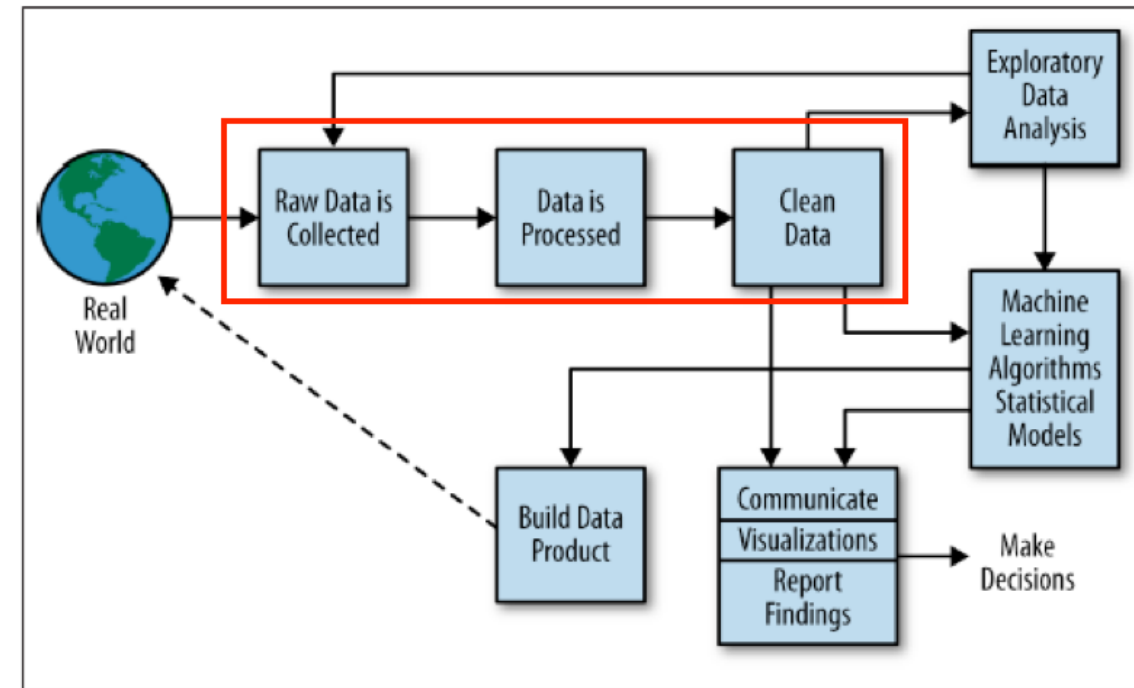
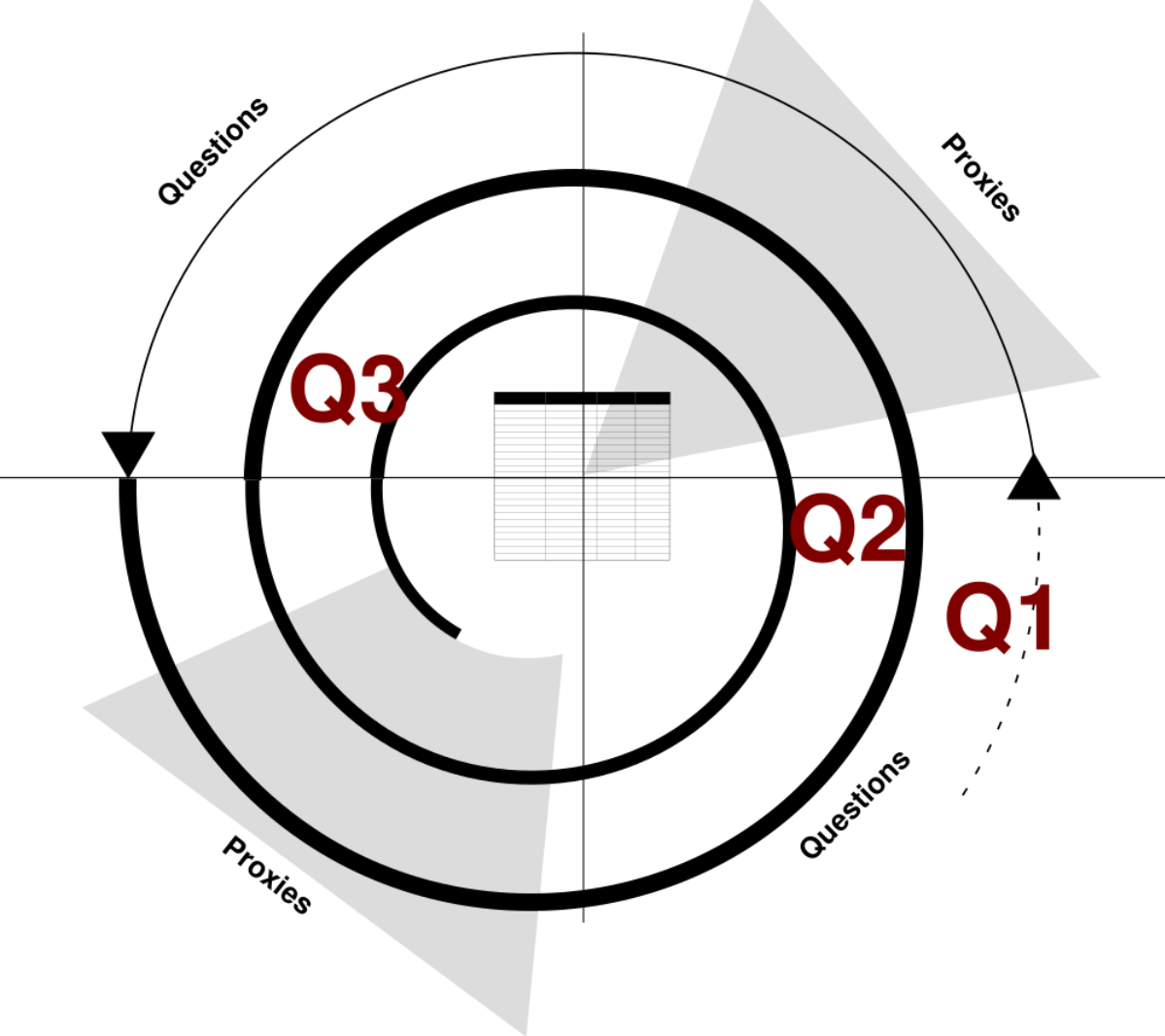


Figure 2-2. The data science process

Cathy O'Neil, Rachel Schutt. *Doing Data Science: Straight Talk from the Frontline* (O'Reilly, 2013) (Chapter 2)

# Iterative Operationalisation



*Danyel Fisher & Miriah Meyer. "Making Data Visual" (O'Reilly, 2018) (Chapter 2)\**



# Iterative Operationalisation (cont'd)

---

- Operationalisation involves searching for **proxies** (proxy tasks, proxy values) for the original question, standing-in for it at the level of the data set.
- Ex. data: a list of movies with ratings (e.g., IMDB) and a list of directors
- Q1: "Who are the best movie directors"?
- **Proxy** for best director: "Having directed many good movies"
- Q2: "What is a good movie"?
- **Proxy**: Good movie: "Having many good IMDB ratings"
- Q3: What is a "good" rating? How many ratings constitute "many" ratings?
- **Proxy**: distributions of rating scores and number of ratings per movie

# Challenges in Data Science

---

**WARNING:** At each stage, things can go wrong! Any filtering/aggregation may bias the data!

- [...] data scientists [...] **spend a lot more time trying to get data into shape than anyone cares to admit—maybe up to 90% of their time**. Finally, they don't find religion in tools, methods, or academic departments. They are versatile and interdisciplinary\*
- Yet far too much handcrafted work — what data scientists call "**data wrangling**," "**data munging**" and "**data janitor work**" — is still required. **Data scientists, according to interviews and expert estimates, spend from 50 percent to 80 percent of their time** mired in this more mundane labor of collecting and preparing unruly digital data, before it can be explored for useful nuggets.

“Data wrangling is a huge — and surprisingly so — part of the job,” said Monica Rogati, vice president for data science at Jawbone, whose sensor-filled wristband and software track activity, sleep and food consumption, and suggest dietary and health tips based on the numbers. “It’s something that is not appreciated by data civilians. At times, it feels like everything we do.”\* **New York times**

# Challenges in Data Science (cont'd)

SECTIONS 🔍 The New York Times SUBSCRIBE LOG IN

N.S.A. Suspect Is a Hoarder. But a Leaker? Investigators Aren't Sure.

Twitter's Fate: Marc Benioff of Salesforce Addresses Acquisition Talk

PAID POST: CHAUMET Napoleon Owned Jewels by This Legendary Maison

STATE OF THE MAINTENANCE: MailChimp Un-Silicon Valley Make It as

## For Big-Data Scientists, 'Janitor Work' Is Key Hurdle to Insights

By STEVE LOHR AUG. 17, 2014



Monica Rogati, Jawbone's vice president for data science, with Brian Wilt, a senior data scientist. Peter DaSilva for The New York Times

Technology revolutions come in measured, sometimes foot-dragging steps. The lab science and marketing enthusiasm tend to underestimate the bottlenecks to progress that must be overcome with hard work and practical engineering.

# The Data Science Lifecycle: your own experiences?

---

Which difficulties have you already experienced when working with data?

# The Data Science Lifecycle: your own experiences?

---

Which difficulties have you already experienced when working with data?

1. ... ever had problems loading/ importing a file someone sent to you because of an unknown file format?

# The Data Science Lifecycle: your own experiences?

---

Which difficulties have you already experienced when working with data?

1. ... ever had problems loading/ importing a file someone sent to you because of an unknown file format?
2. ... ever encountered something like this: "K❖snudl"?

# The Data Science Lifecycle: your own experiences?

---

Which difficulties have you already experienced when working with data?

1. ... ever had problems loading/ importing a file someone sent to you because of an unknown file format?
2. ... ever encountered something like this: "K❖snudl"?
3. ... ever encountered blanks in your data?

# The Data Science Lifecycle: your own experiences?

---

Which difficulties have you already experienced when working with data?

1. ... ever had problems loading/ importing a file someone sent to you because of an unknown file format?
2. ... ever encountered something like this: "K❖snudl"?
3. ... ever encountered blanks in your data?
4. ... ever saw an observation (an insight, a trend) disappear when combining from different data sets (a.k.a. "Simpson's paradox")



# The Data Science Lifecycle: your own experiences?

---

Which difficulties have you already experienced when working with data?

1. ... ever had problems loading/ importing a file someone sent to you because of an unknown file format?
2. ... ever encountered something like this: "K❖snudl"?
3. ... ever encountered blanks in your data?
4. ... ever saw an observation (an insight, a trend) disappear when combining from different data sets (a.k.a. "Simpson's paradox")
5. ... **more on that in the next lectures!**



Sign in

Home

News

Sport

Reel

Worklife

Travel

## NEWS

Home | US Election | Coronavirus | Video | World | UK | Business | **Tech** | Science | Stories | Entertainment & Arts

Tech

# Excel: Why using Microsoft's tool caused Covid-19 results to be lost

By Leo Kelion  
Technology desk editor

2 hours ago

Coronavirus pandemic



# Excursus: Simpson's paradox (1)

---

*Participants of the Whickham study*

	<b>Total</b>	<b>Deceased</b>	<b>% Deceased</b>	<b>Alive</b>	<b>% Alive</b>
<b>Female smokers</b>	369	139	37,70%	230	62,30%
<b>Female non-smokers</b>	945	443	46,90%	502	53,10%
<b>Total female</b>	1314	582	44,30%	732	55,70%

## Excursus: Simpson's paradox (2)

---

### Teilnehmerinnen nach Altersklassen zu Beginn der Studie

Alter	Raucherinnen			Nichtraucherinnen		
	Verstorben	Überlebend	Todesrate	Verstorben	Überlebend	Todesrate
18-24 J	2	53	3,60%	1	61	1,60%
25-34 J	3	121	2,40%	5	152	3,20%
35-44 J	14	95	12,80%	7	114	5,80%
45-54 J	27	103	20,80%	12	66	15,40%
55-64 J	51	64	44,30%	40	81	33,10%
65-74 J	29	7	80,60%	101	28	78,30%
ab 75 J	13	0	100,00%	64	0	100,00%

# Data Science Lifecycle: Summary

---

Again, not a single definition, but some recurring terms:

1. **find and collect all relevant data**
2. **identify issues & problems within the data**
3. **organise / transform / merge data**
4. systematically operationalise questions about the data: proxies
5. select a visualisation, a statistical technique, or a machine-learning technique as an outcome of operationalisation
6. provide interpretations and limitations of the results
7. communicate results

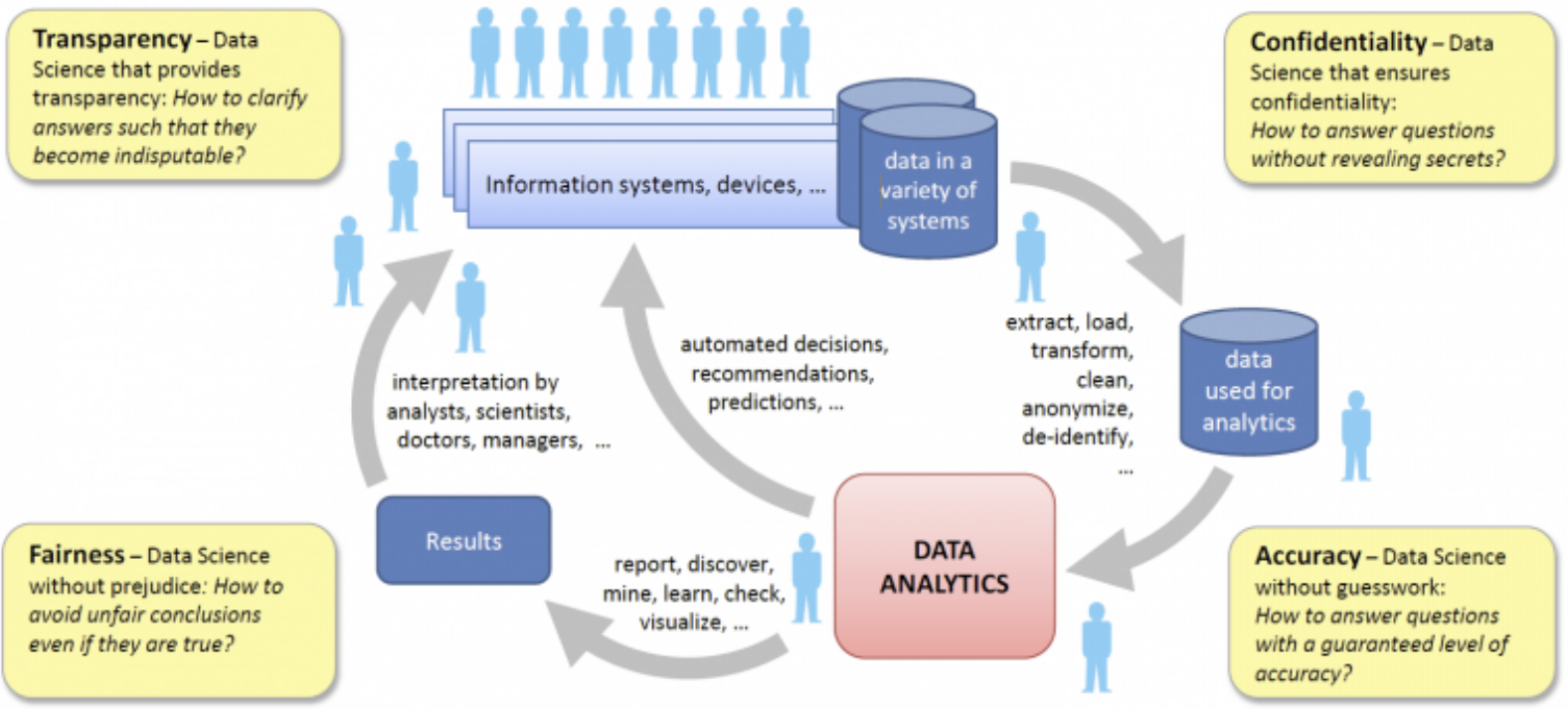
# **Data Science Ethics**

# Ethics in Data Science: FACT

---

- **Fairness** : How to avoid unfair conclusions even if they are true?
- **Accuracy** : How to answer questions with a guaranteed level of accuracy?
- **Confidentiality** : How to answer questions without revealing secrets?
- **Transparency** : How to clarify answers such that they become indisputable?

# Ethics in Data Science: FACT (cont'd)



Source <http://www.responsibledata-science.org/>



# Data Science Lifecycle: Summary

---

## NOTE:

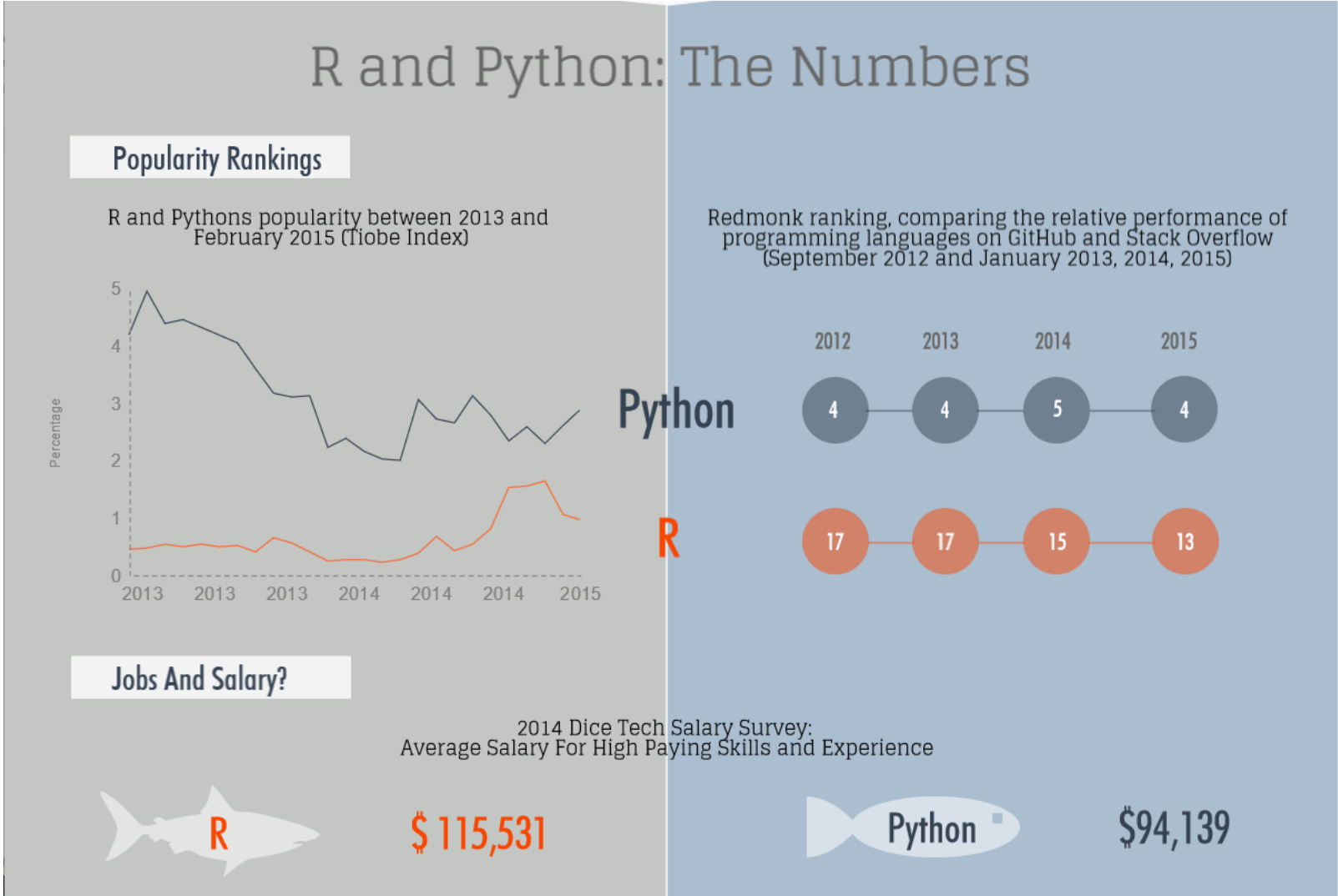
- Typically, Data Science is not a one-shot process, but an (iterative) lifecycle.
- Not ad hoc, but short-lived than building than classic processes: ETL, data mining.
- Typically, you need to revisit/ adjust your process, either for improving it or for maintenance (sources changing, source formats changing, etc.)
- Mind FACT in Data Science projects

### **Notice.**

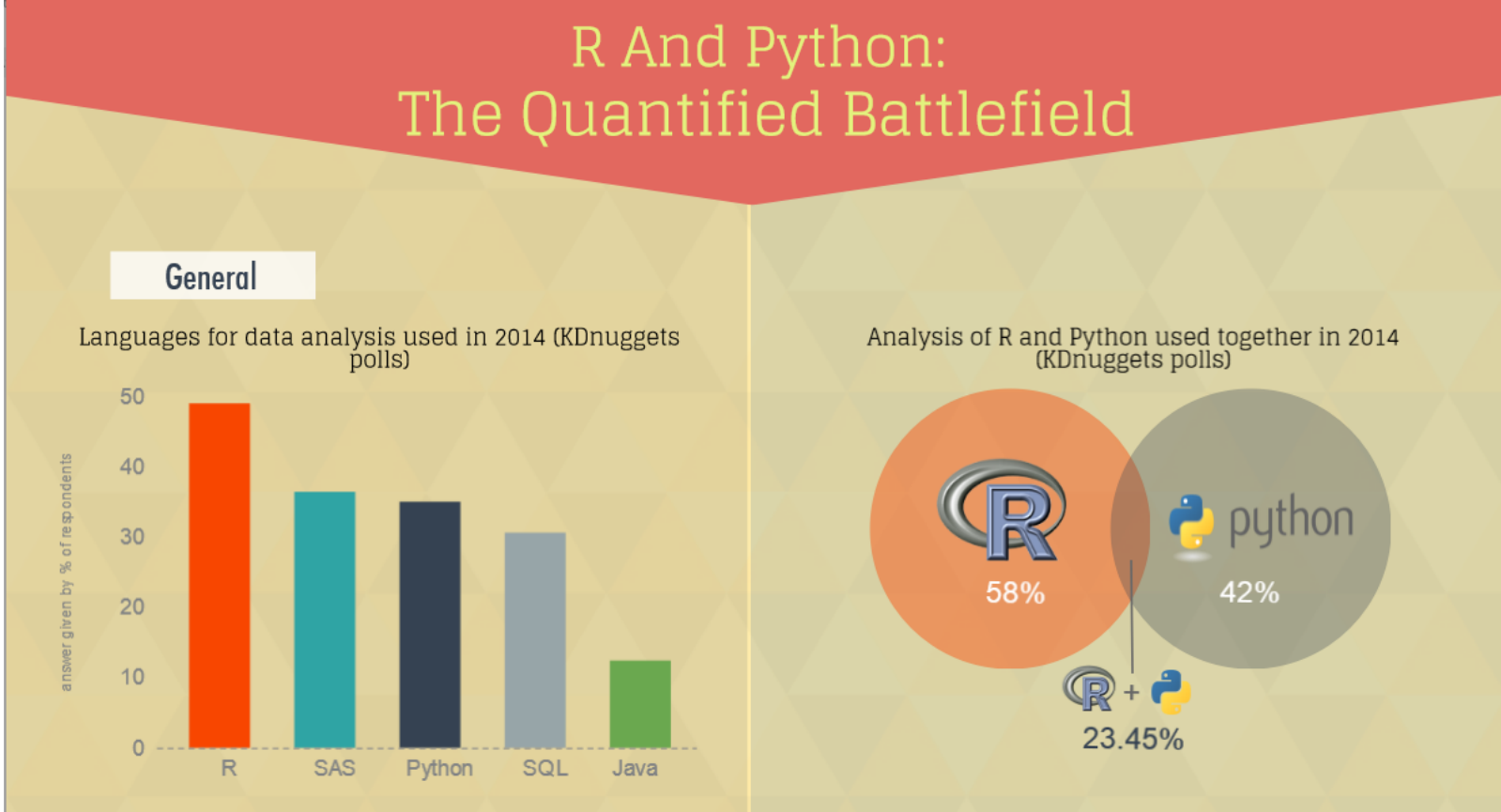
These steps may take **80% of the work** or more -> This is the focus of our course "**Data Processing I**" !!!

# **Data Science Tools**



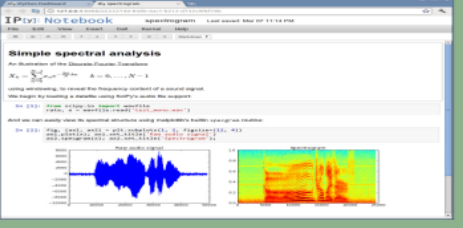
# Data Science Tools: Python and R



Source <https://www.datacamp.com/community/tutorials/r-or-python-for-data-analysis>



Source <https://www.datacamp.com/community/tutorials/r-or-python-for-data-analysis>

Graphical Capabilities	+	IPython Notebook
<p><b>A picture says more than a thousand words</b></p> <p>Visualized data can be understood more efficiently and effectively than the raw numbers alone.</p> <p><b>R + visualization = perfect match</b> </p> <p><b>ggplot2</b> To make pretty graphs, including the opportunity to use grammar of graphics to create layered, customizable plots</p> <p><b>lattice</b> To easily display multivariate relationships</p> <p><b>rCharts</b> To create, customize and publish interactive javascript visualizations from R</p> <p><b>googleVis</b> To use Google Chart tools to visualize data in R</p> <p><b>ggvis</b> To implement interactive grammar of graphics, while rendering in a web browser</p> <p>e.g.: Visualizing Facebook friends with R</p> 		<p><b>Bundle your analysis in one file</b></p> <p>The IPython Notebook makes it easier to work with Python and data.</p> <p><b>Simplify your workflow when working with data in Python</b></p> <p>It's a combination of:</p> <ul style="list-style-type: none"><li>Interactive python exploration,</li><li>prewritten programs, text, and equations for documentation in one environment</li></ul> <p><b>Share notebooks with colleagues without having them install anything.</b></p> <p>The IPython notebook drastically reduces the overhead of organizing code, output, and notes files, which allows to spend more time doing real work.</p> 

# Why Python and R

---

The Python vs R debate confines you to one programming language. You should look beyond it and embrace both tools for their respective strengths. Using more tools will only make you better as a data scientist. [[TheNextWeb](#)]

- Data Processing 1 (SBWL 1): Python
- Data Analytics (SBWL 2): R
- Data Processing (SBWL 3): Python