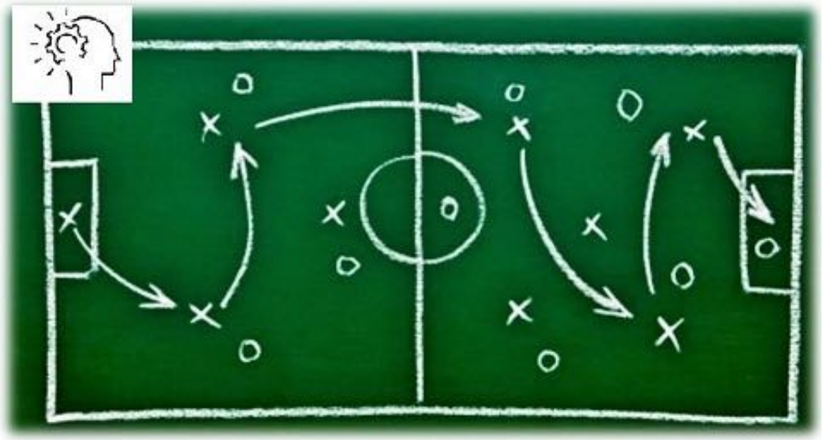
Strategic and Competitive Intelligence

Data Science Applications in R



Filippo Chiarello

filippochiarello.90@gmail.com

About me

- FILIPPO CHIARELLO

- PHD STUDENT IN MANAGEMENT ENGINEERING

- RESEARCH ON DATA SCIENCE AND TEXT MINING

- MUSICIAN

ERREQUADRO
Research over Research

Goal of this Module

Data —> Knowledge

Some Magic…

… lot of problems

- Start solving problems as soon as possibile (today)

- Learn the 20% of tools that solves the 80% problems

- Don’t solve the same problem multiple times

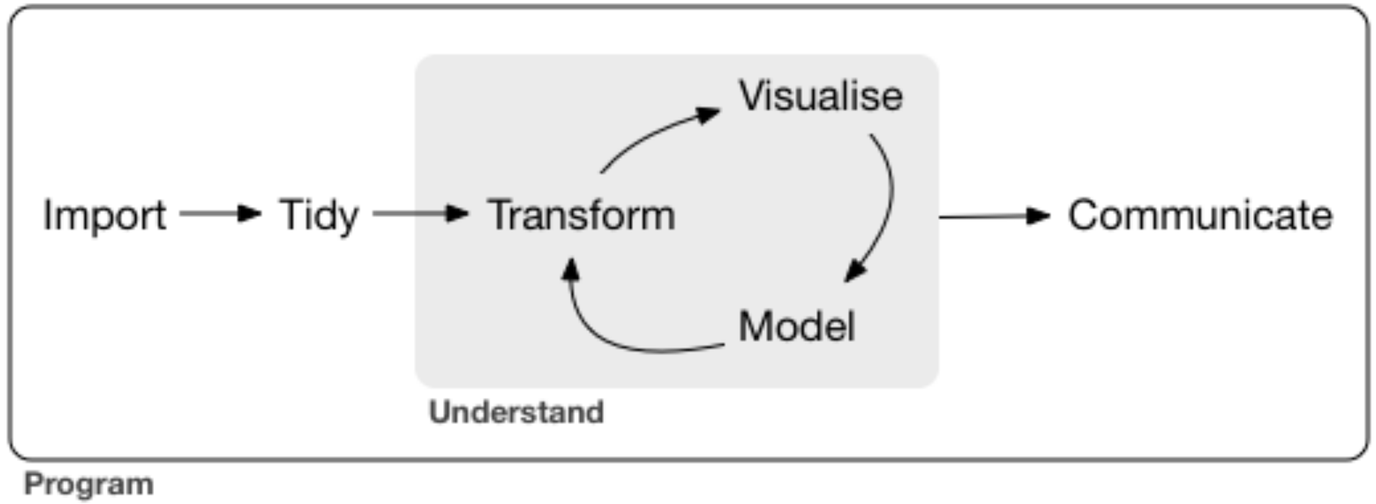
- Don’t solve problem that someone else had solved

Lessons

1. Introduction to R
2. Data Visualisation with ggplot2
3. Data Transormation with dplyr
4. Theory of Models
5. Text Mining
6. Social Networks Analysis
7. Patents Analysis
8. Papers Analysis
9. Results Comunication and Reporting with Rmarkdown

Mantras

Make Models



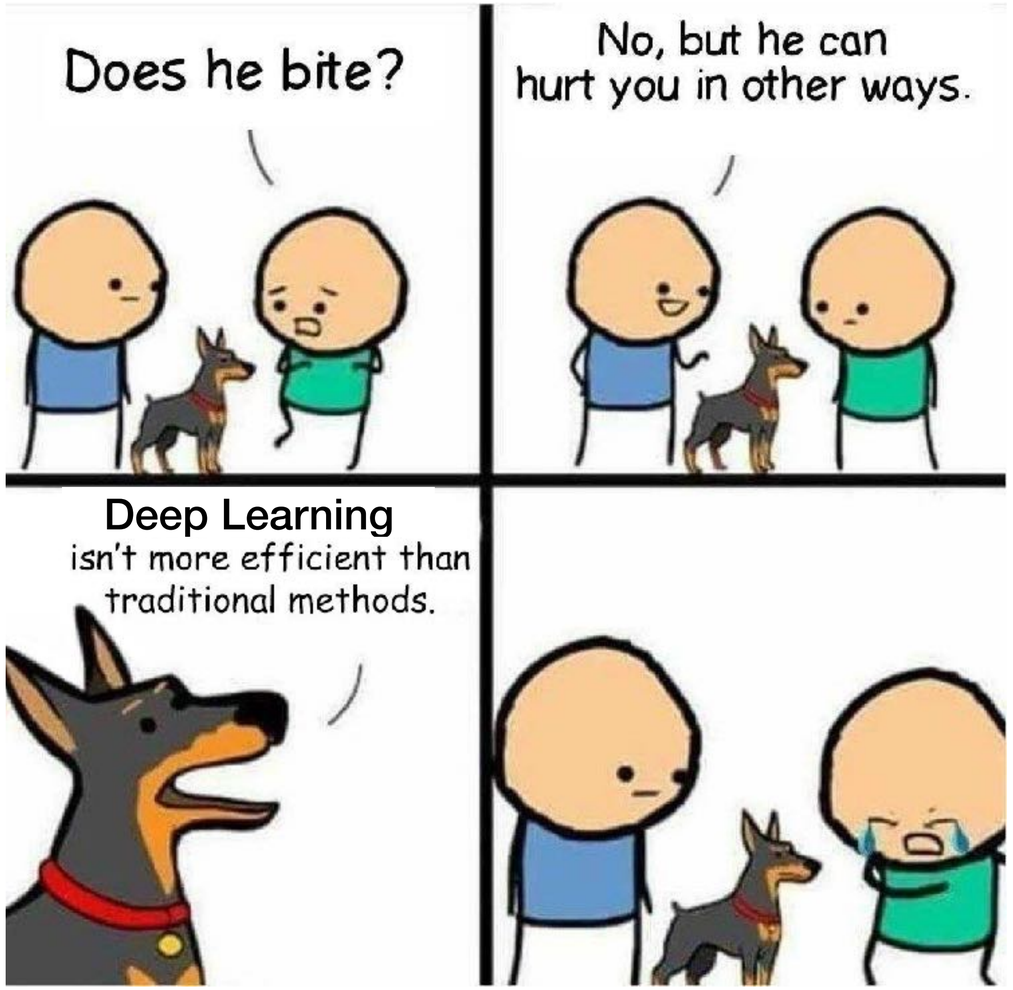
Mantras

Ask Why

Mantras

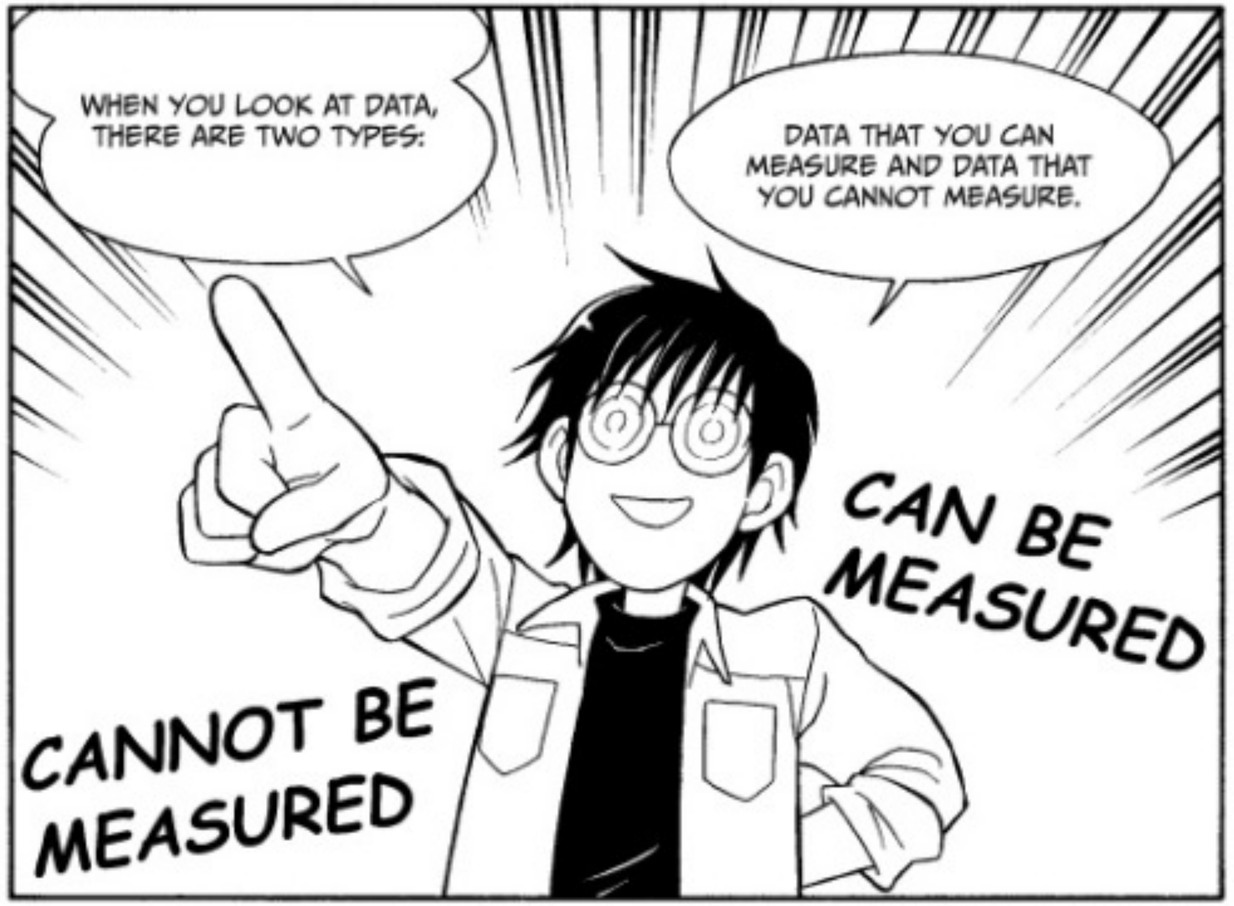
Be Sceptic

Be Agnostic



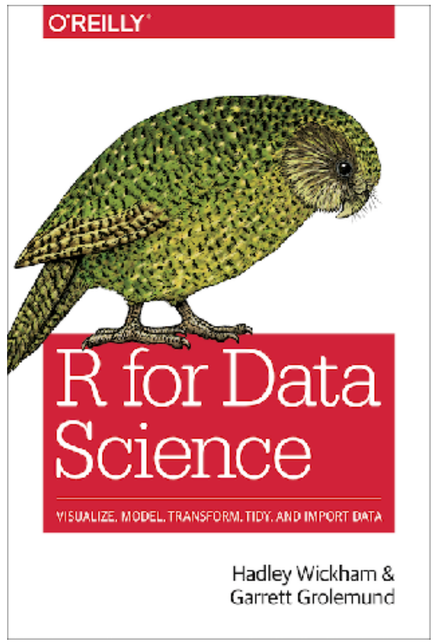
Mantras

Measure it

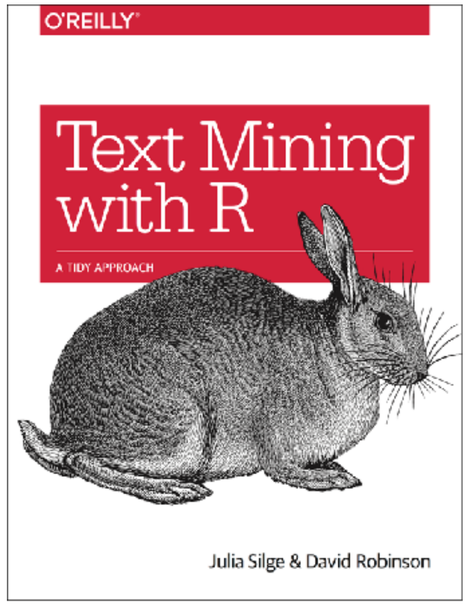


From "Manga Guide to Statistics", Shin Takahashi, 2008

Material



http://r4ds.had.co.nz/



https://www.tidytextmining.com/

Material

Home Posts Courses Material Publications Contact

Filippo Chiarello

PhD Student

University of Pisa, DESTEC

About Me

I study how to use Data Science tools for Innovation Management.

In this blog you can find:

- **Posts**: Projects i'm working on, books that i love, cool blog posts from other researchers, mind-blowing talks and so on.
- **Courses Materials**: News and materials for the my courses.
- **Pubblications**: My latest research activities.

Interests

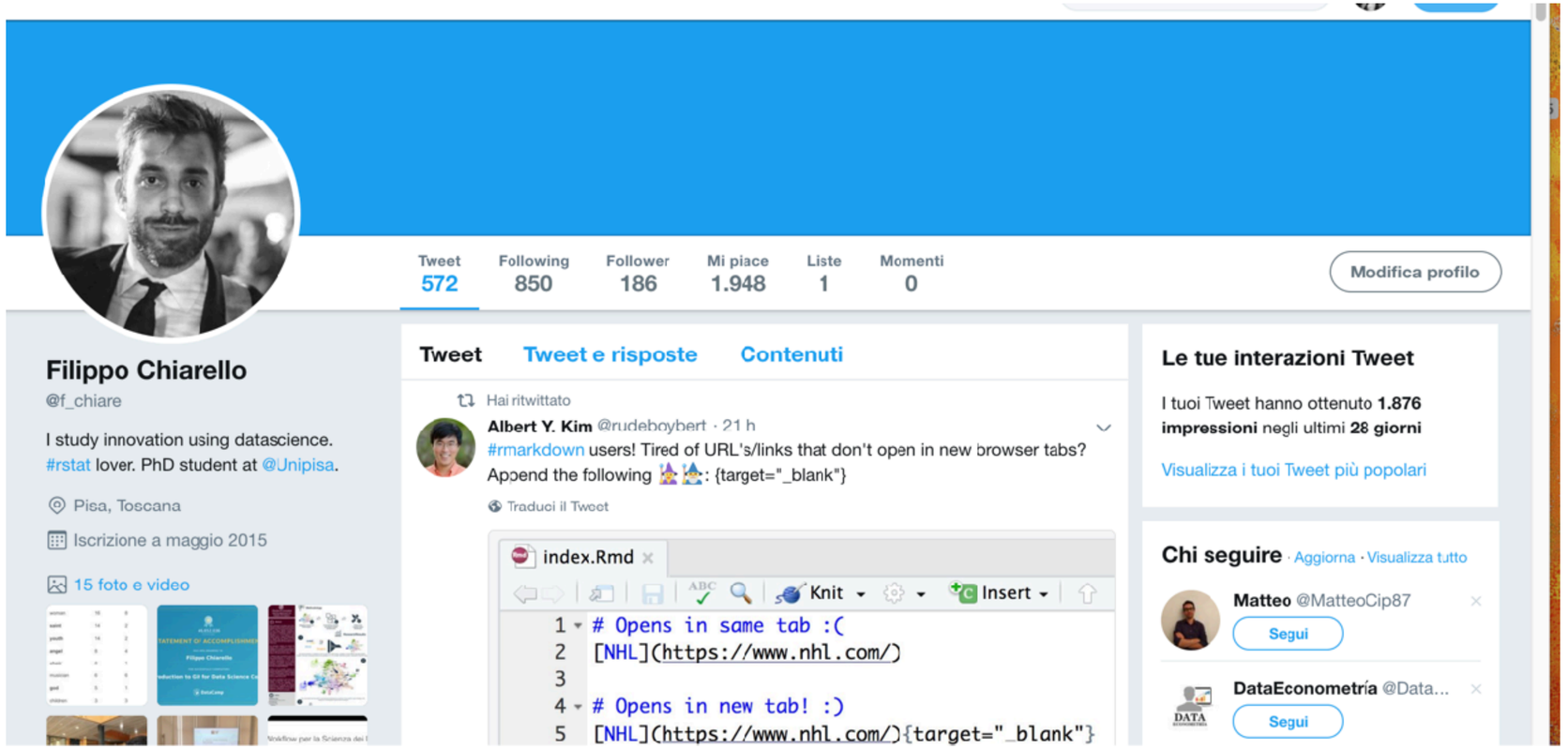
- Data Science
- R
- Engineering Design
- Natural Language Processing
- Artificial Intelligence

Education

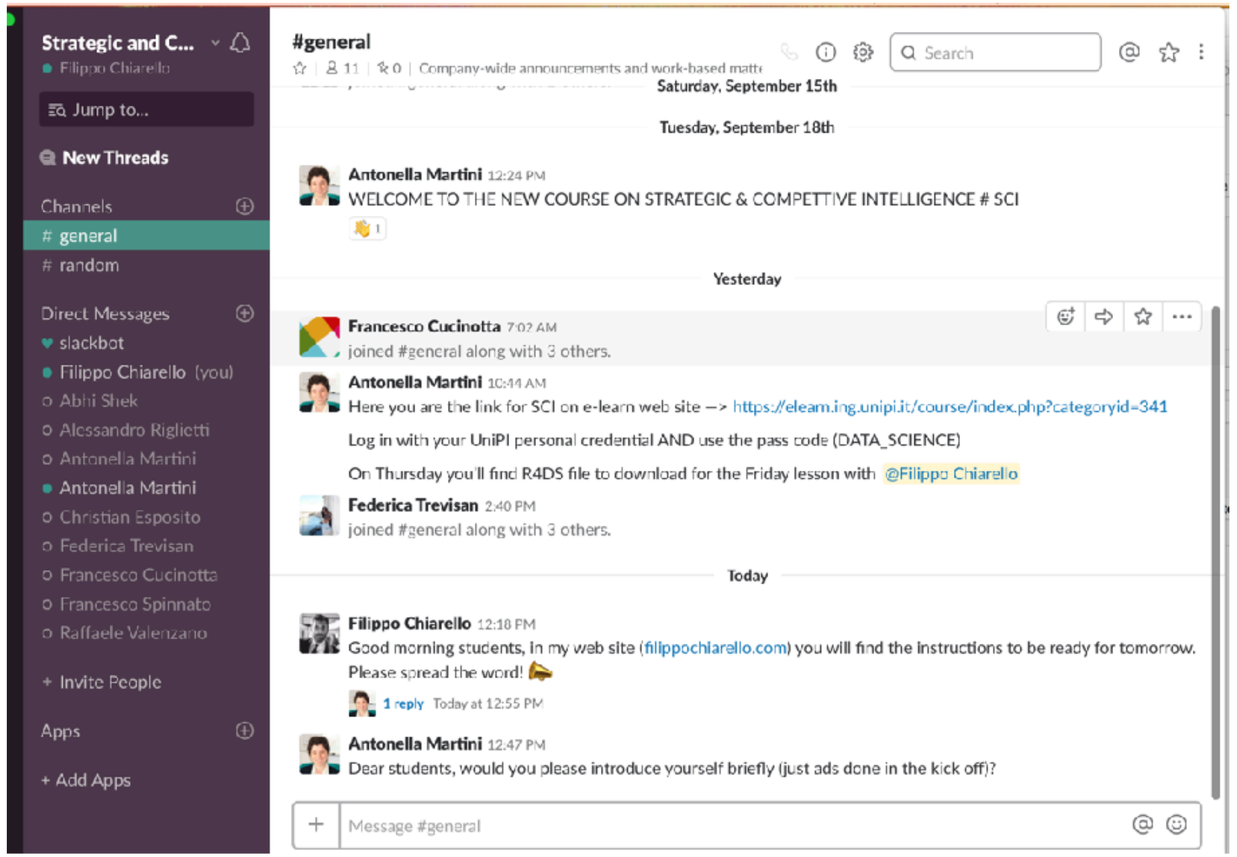
MEng in Engineering Management, 2015
University of Pisa

BSc in Engineering Management, 2013
University of Pisa

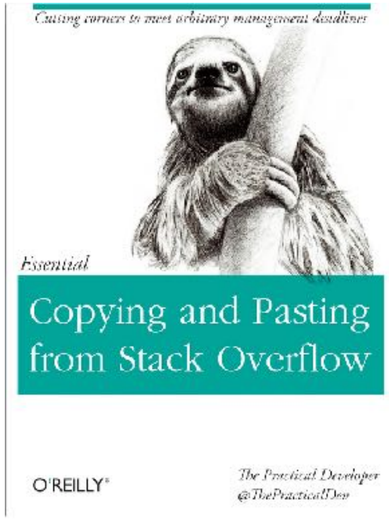
Material


Tweet 572
Following 850
Follower 186
Mi piace 1.948
Liste 1
Momenti 0
Modifica profilo
Filippo Chiarello
@f_chiare
I study innovation using datascience.
#rstat lover. PhD student at @Unipisa.
Pisa, Toscana
Iscrizione a maggio 2015
15 foto e video
Tweet
Tweet e risposte
Contenuti
Hai ritwittato
Albert Y. Kim @rudeboybert · 21 h
#rmarkdown users! Tired of URL's/links that don't open in new browser tabs?
Append the following : {target="_blank"}
Traduci il Tweet
index.rmd
Knit
Insert
1 # Opens in same tab :(
2 [NHL](https://www.nhl.com/)
3
4 # Opens in new tab! :)
5 [NHL](https://www.nhl.com/){target="_blank"}
Le tue interazioni Tweet
I tuoi Tweet hanno ottenuto 1.876 impressioni negli ultimi 28 giorni
Visualizza i tuoi Tweet più popolari
Chi seguire · Aggiorna · Visualizza tutto
Matteo @MatteoCip87
Segui
DataEconometría @Data...
Segui

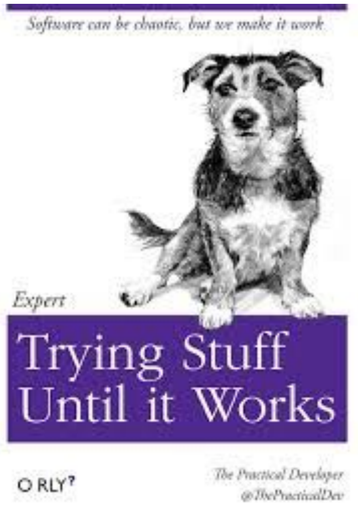
Material

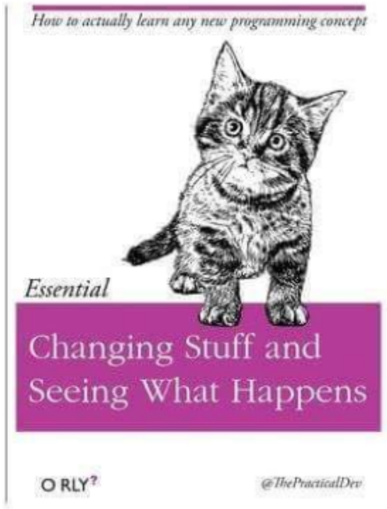

Strategic and C...
Filippo Chiarello
Jump to...
New Threads
Channels
general
random
Direct Messages
slackbot
Filippo Chiarello (you)
Abhi Shek
Alessandro Riglietti
Antonella Martini
Antonella Martini
Christian Esposito
Federica Trevisan
Francesco Cucinotta
Francesco Spinnato
Raffaele Valenzano
+ Invite People
Apps
+ Add Apps
#general
11 | 0 | Company-wide announcements and work-based matte
Search
Saturday, September 15th
Tuesday, September 18th
Antonella Martini 12:24 PM
WELCOME TO THE NEW COURSE ON STRATEGIC & COMPETTIVE INTELLIGENCE # SCI
Yesterday
Francesco Cucinotta 7:02 AM
joined #general along with 3 others.
Antonella Martini 10:44 AM
Here you are the link for SCI on e-learn web site --> https://eleam.ing.unipi.it/course/index.php?categoryid=341
Log in with your UniPI personal credential AND use the pass code (DATA_SCIENCE)
On Thursday you'll find R4DS file to download for the Friday lesson with @Filippo Chiarello
Federica Trevisan 2:40 PM
joined #general along with 3 others.
Today
Filippo Chiarello 12:18 PM
Good morning students, in my web site (filippochiarello.com) you will find the instructions to be ready for tomorrow.
Please spread the word!
1 reply Today at 12:55 PM
Antonella Martini 12:47 PM
Dear students, would you please introduce yourself briefly (just ads done in the kick off)?
Message #general

Material

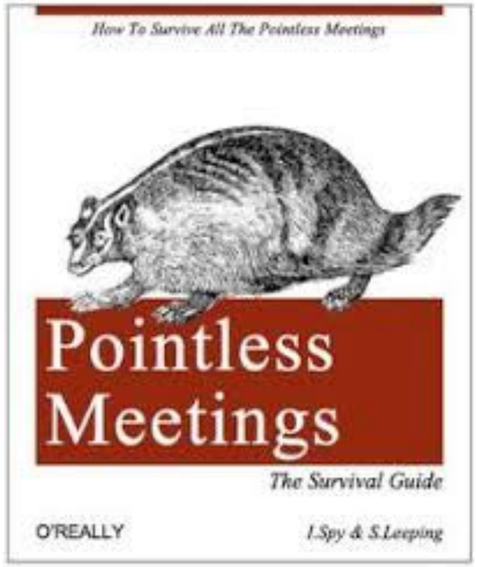

Cutting corners to meet arbitrary management deadlines
Essential
Copying and Pasting from Stack Overflow
O'REILLY
The Practical Developer
@ThePracticalDev

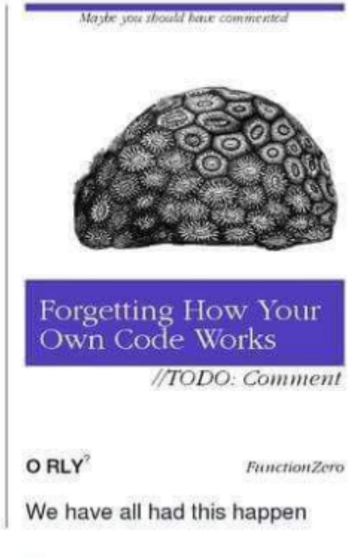

The internet will make those bad words go away
Essential
Googling the Error Message
O RLY?
The Practical Developer
@ThePracticalDev

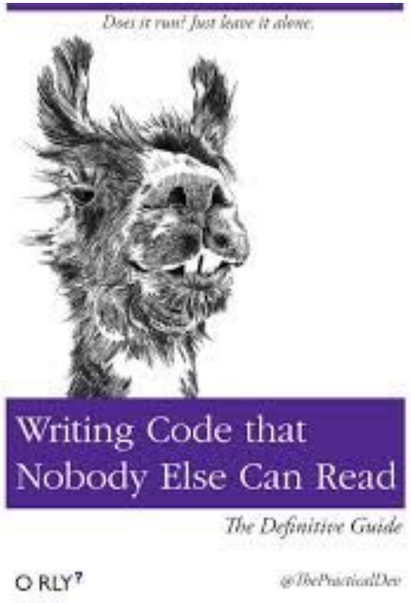

Software can be chaotic, but we make it work
Expert
Trying Stuff Until it Works
O RLY?
The Practical Developer
@ThePracticalDev

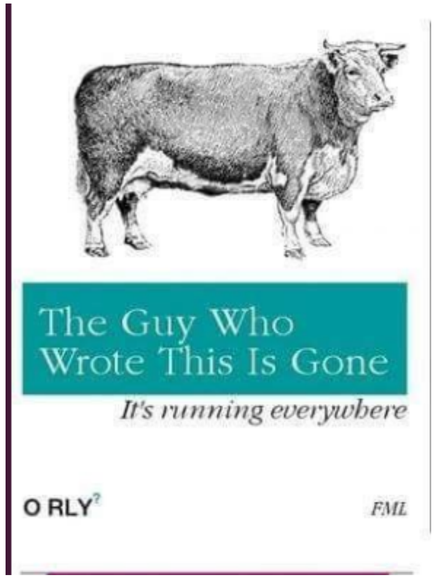

How to actually learn any new programming concept
Essential
Changing Stuff and Seeing What Happens
O RLY?
@ThePracticalDev

Material

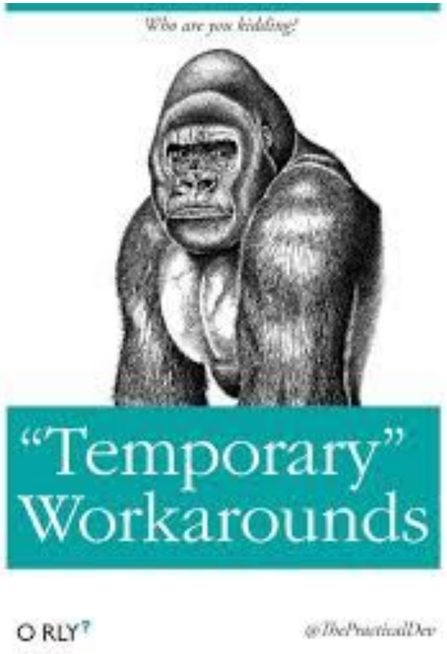

How To Survive All The Pointless Meetings
Pointless Meetings
The Survival Guide
O'REALLY
I.Spy & S.Leeping

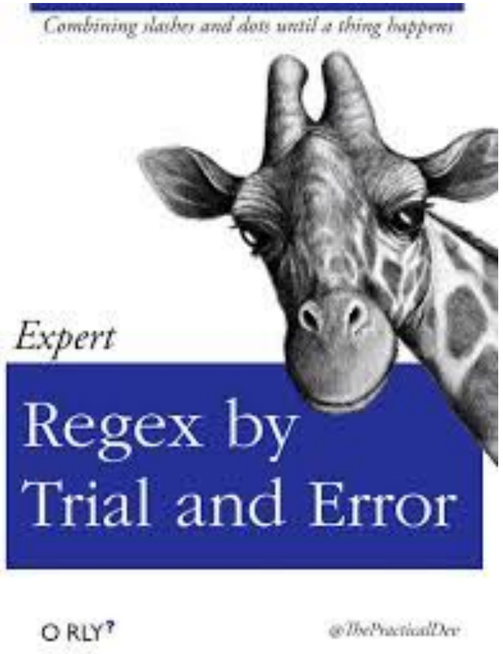

Maybe you should have commented
Forgetting How Your Own Code Works
//TODO: Comment
O RLY?
FunctionZero
We have all had this happen


Does it run? Just leave it alone.
Writing Code that Nobody Else Can Read
The Definitive Guide
O RLY?
@ThePracticalDev

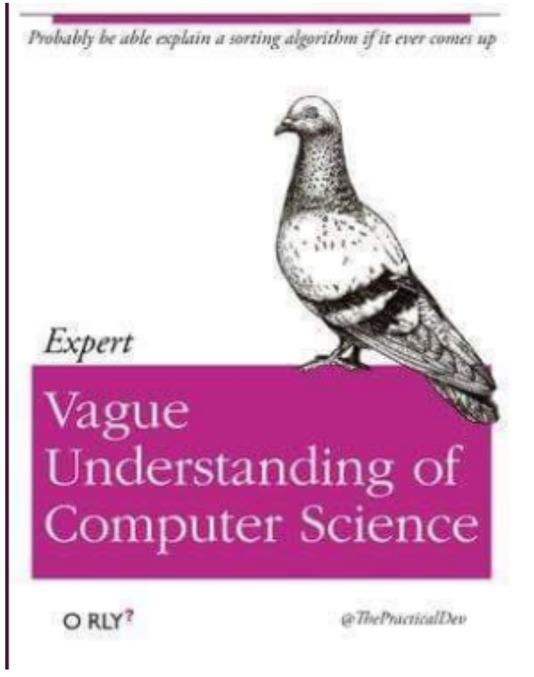

The Guy Who Wrote This Is Gone
It's running everywhere
O RLY?
FML

Material


Who are you kidding?
"Temporary" Workarounds
O RLY?
@ThePracticalDev

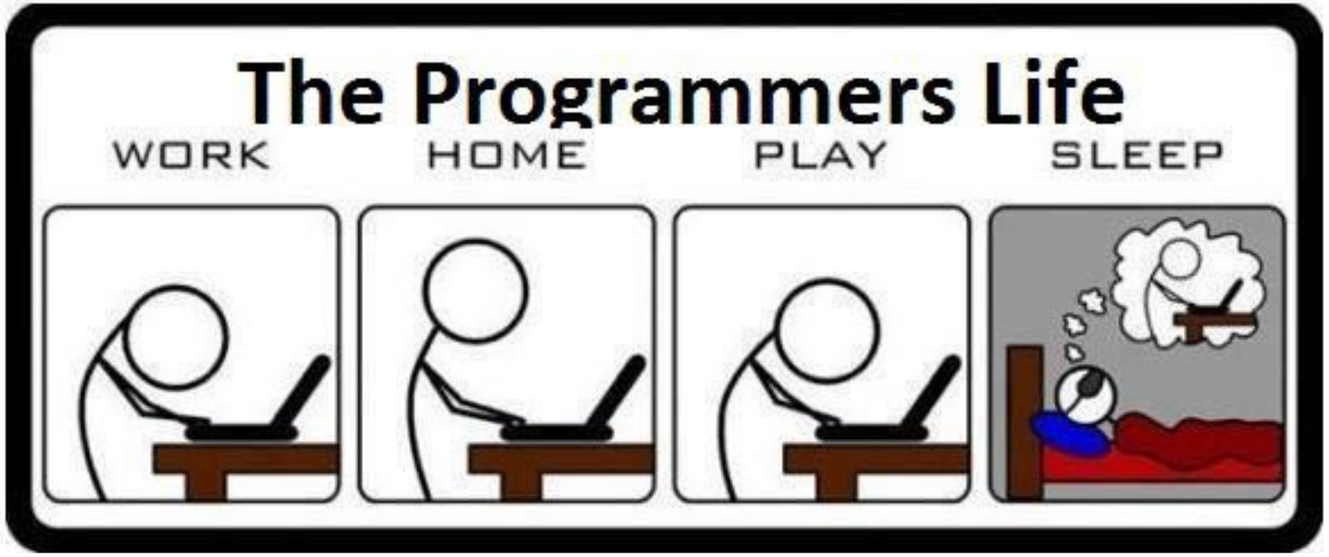

Combining slashes and dots until a thing happens
Expert
Regex by Trial and Error
O RLY?
@ThePracticalDev


git commit -m "changes"
Writing
Useless Git Commit Messages
O RLY?
@ThePracticalDev


Probably be able explain a sorting algorithm if it ever comes up
Expert
Vague Understanding of Computer Science
O RLY?
@ThePracticalDev

In other words…..

We <u>are not</u> interested in your programming skills….



In other words…..

… but in your critical thinking skills!

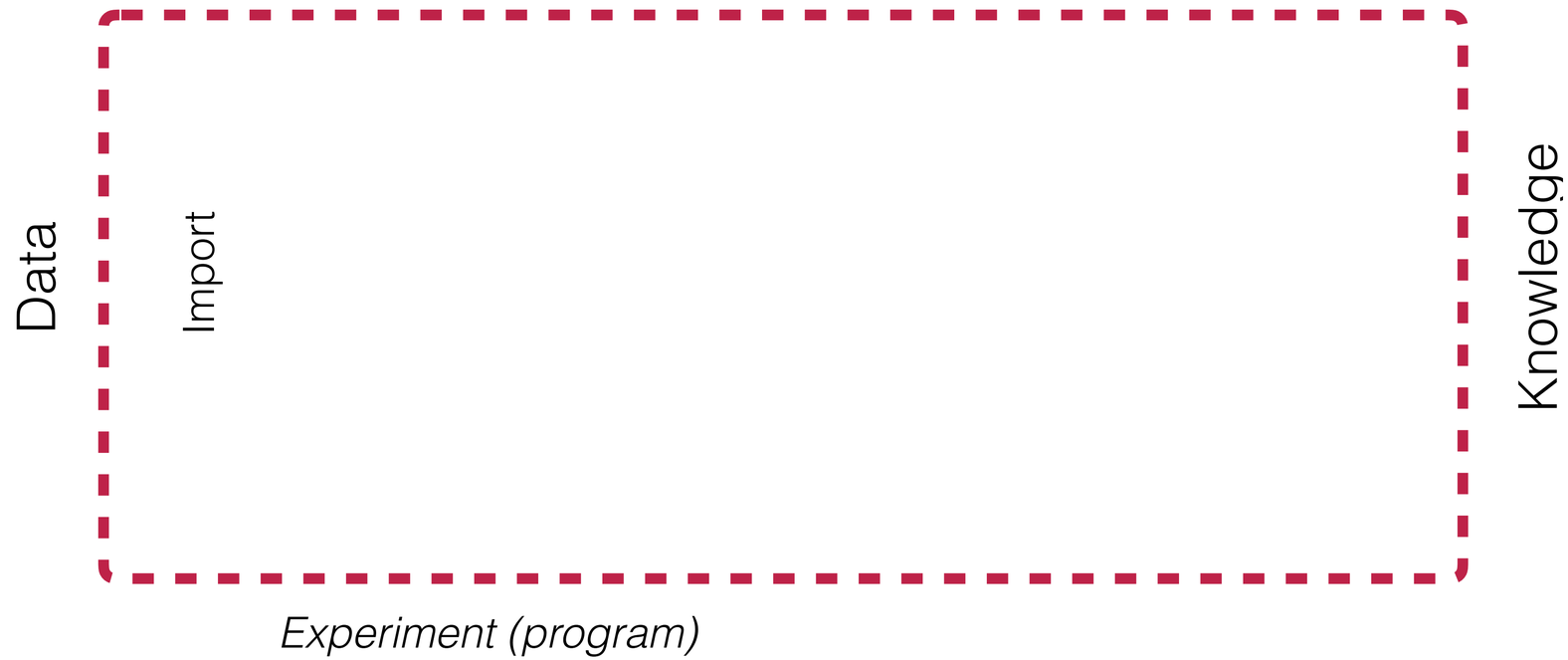
"If you don't reveal some insights soon, I'm going to be forced to slice, dice, and drill!"

Demystifying Data Science

DS != programming

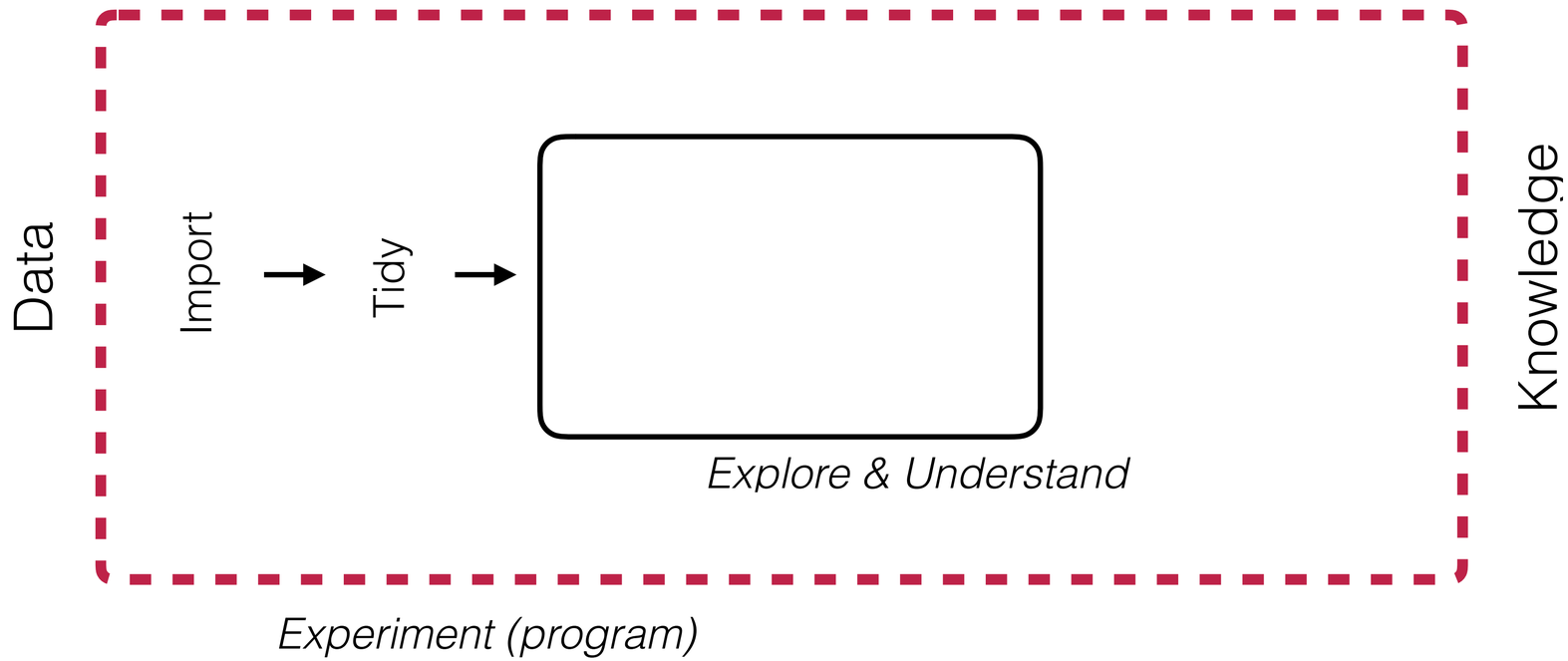
DS Workflow

DS Workflow

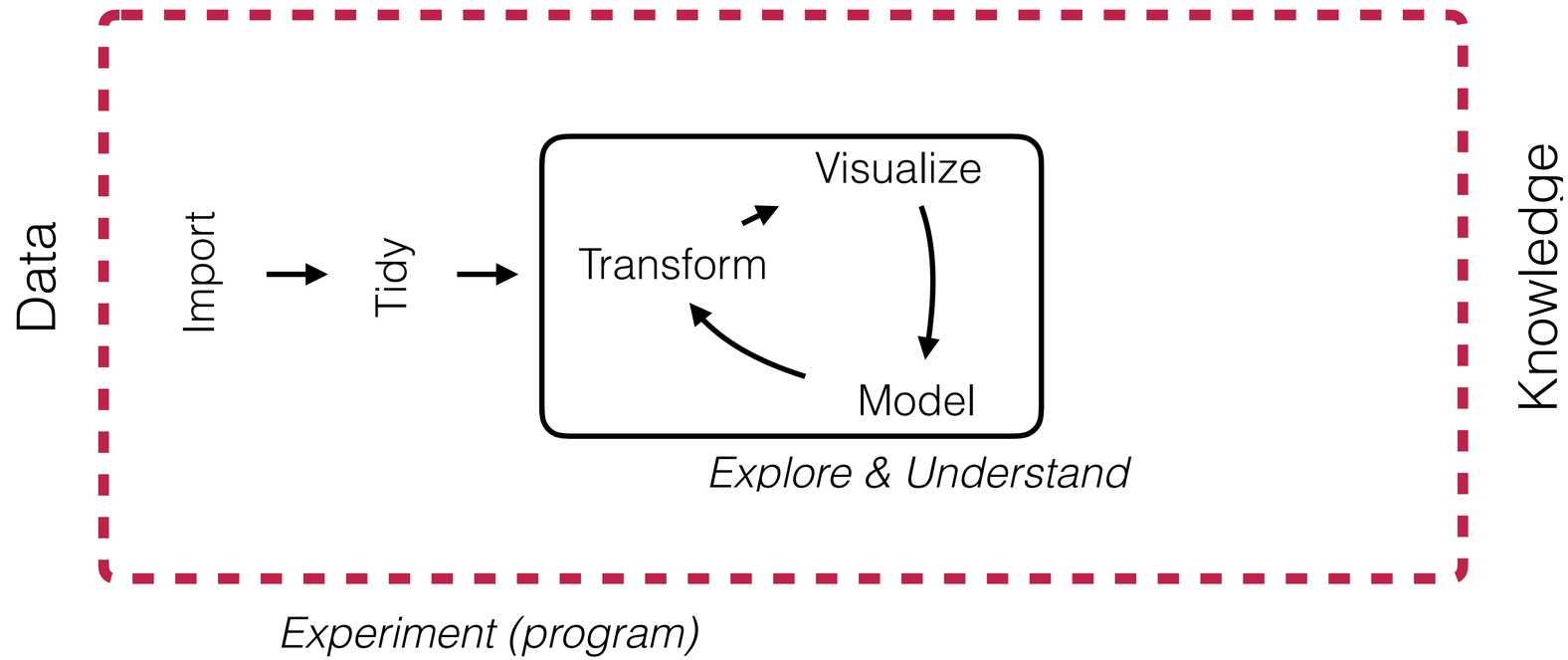


DS Workflow

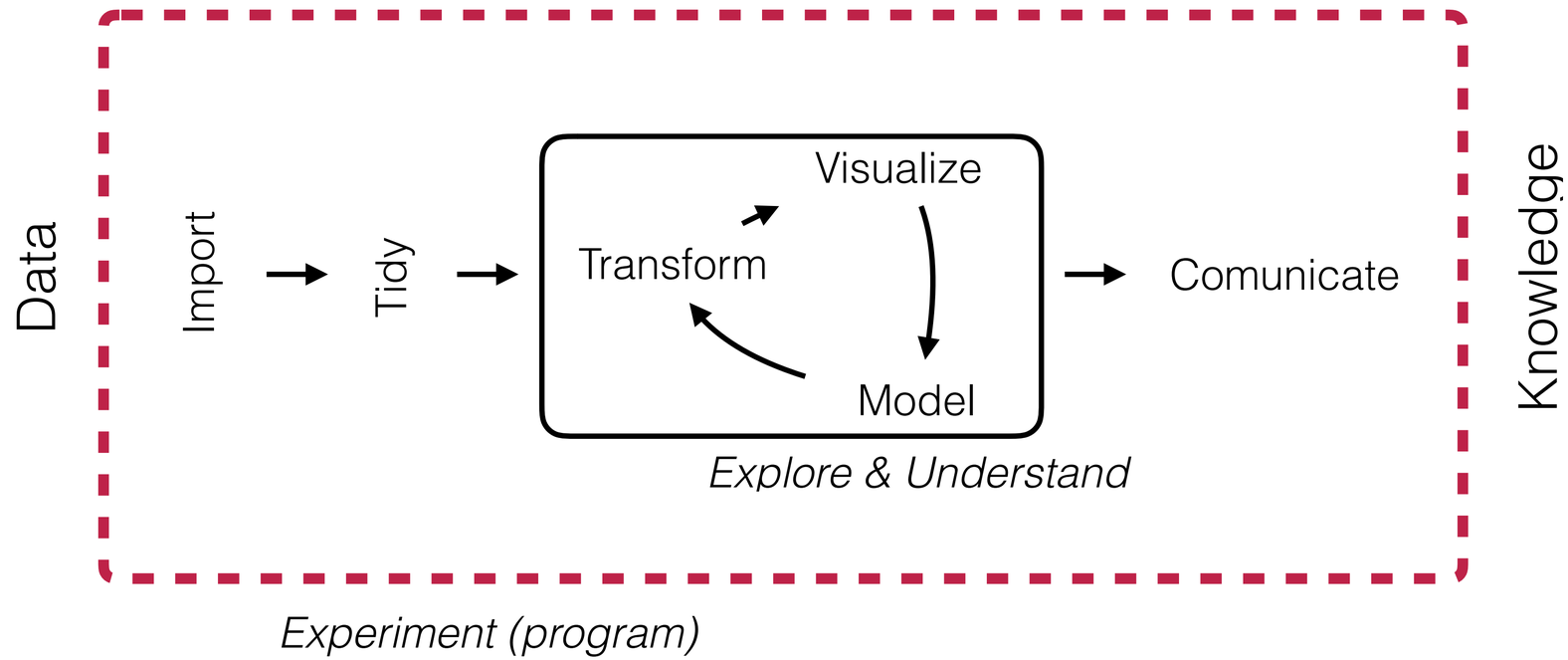
DS Workflow


Data
Import
Tidy
Explore & Understand
Knowledge
Experiment (program)

DS Workflow


Data
Import
Tidy
Visualize
Transform
Model
Explore & Understand
Knowledge
Experiment (program)

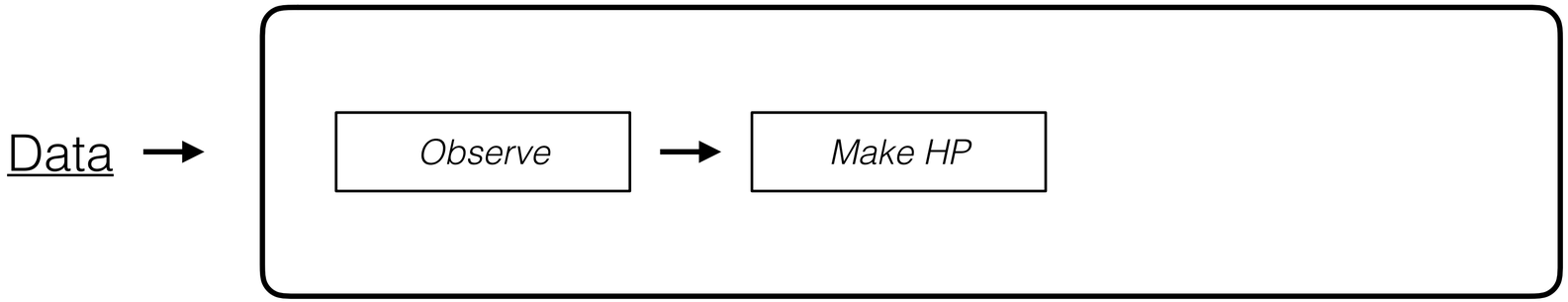
DS Workflow


Data
Import
Tidy
Visualize
Transform
Model
Explore & Understand
Comunicate
Knowledge
Experiment (program)

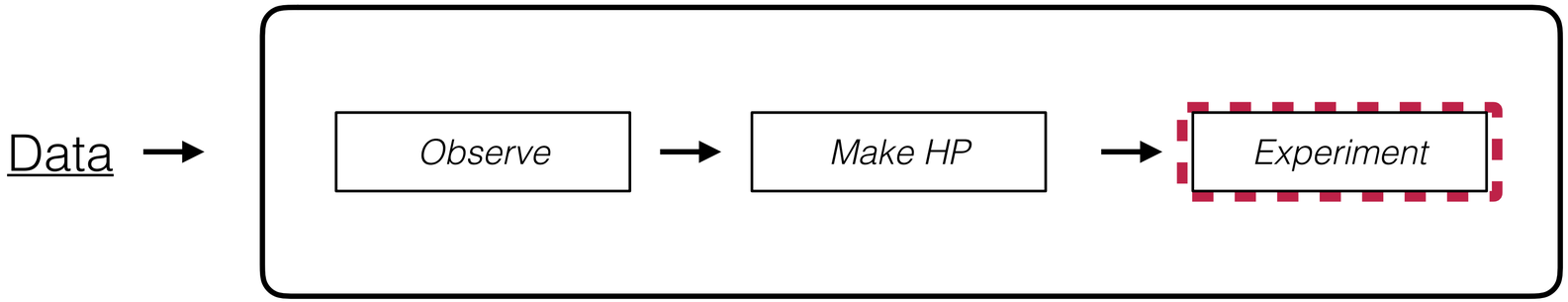
The Scientific Method

The Scientific Method

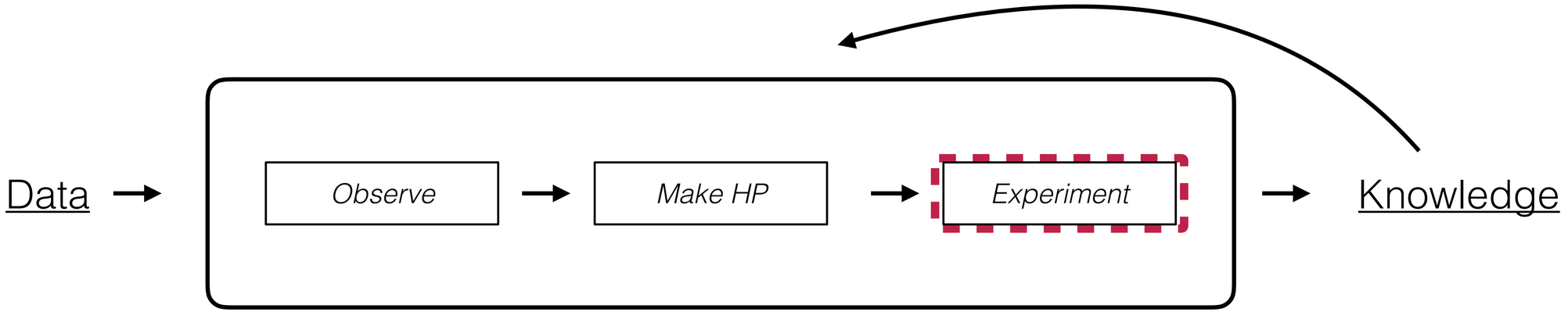
The Scientific Method



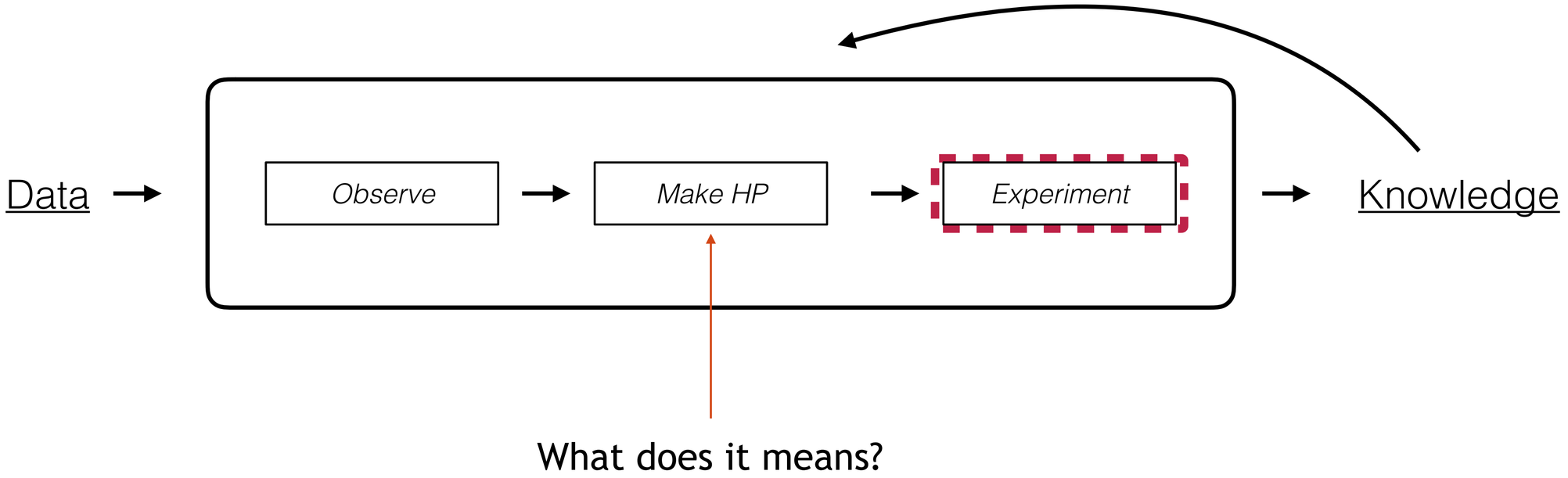
The Scientific Method



The Scientific Method



The Scientific Method



Hypothesis

A **hypothesis** is a proposed explanation for a phenomenon. For a hypothesis to be a scientific hypothesis, the scientific method requires that one can test it.

Scientists generally base scientific hypotheses on previous observations that cannot satisfactorily be explained with the available scientific theories.

WHY DO WE NEED HP TO CREATE NEW KNOWLEDGE?

Postulation

A suggestion or assumption of the existence, fact, or truth of something as a basis for reasoning, discussion, or belief.

WHY DO WE NEED POSTULATION TO CREATE NEW HYPOTHESIS?

How to Chose an Hypothesis

How to Chose an Hypothesis

- Testability (effort to accept or reject the hypothesis)
- Parsimony (discouraging excessive postulation)
- Scope (the apparent application of the hypothesis to multiple cases of phenomena)
- Fruitfulness (the prospect that a hypothesis may explain future phenomena)
- Conservatism (the degree of fit with existing recognized knowledge-systems)

Data Scientist

Skills?

Top Data Science Skills by Job Role

Business Manager

Data Skill	% who have skill
S - Communication	91%
B - Project management	86%
B - Business development	77%
T - Managing structured data	74%
B - Budgeting	71%
B - Product Design and Development	70%
M - Math	65%
S - Data Management	64%
S - Data Mining and Viz Tools	64%
B - Governance and Compliance	61%
S - Science/Scientific Method	59%
S - Statistics and statistical modeling	54%
P - Database Administration	50%
M - Optimization	48%
M - Algorithms and Simulations	46%
M - Graphical Models	46%
T - Managing unstructured data	43%
P - Front-end Programming	42%
T - Machine Learning	41%
P - Systems Administration and Design	41%
M - Bayesian Statistics	39%
P - Back-end Programming	34%
P - Cloud Management	31%
T - NLP and text mining	30%
T - Big and distributed data	29%

Developer

Data Skill	% who have skill
T - Managing structured data	91%
S - Communication	85%
S - Data Mining and Viz Tools	76%
B - Product Design	75%
M - Math	75%
S - Data Management	75%
B - Project management	74%
P - Database Administration	73%
P - Back-end Programming	70%
P - Systems Administration	65%
S - Science/Scientific Method	64%
T - Unstructured data	64%
P - Front-end Programming	63%
M - Algorithms	61%
T - Machine Learning	54%
S - Statistics and statistical modeling	52%
M - Graphical Models	51%
B - Buisness development	48%
T - Big and distributed data	48%
M - Optimization	47%
B - Budgeting	45%
B - Goverance and Compliance	44%
M - Bayesian Statistics	44%
T - NLP	43%
P - Cloud Management	42%

Creative

Data Skill	% who have skill
S - Communication	87%
T - Managing structured data	79%
B - Project management	77%
S - Data Mining and Viz Tools	77%
M - Math	75%
B - Product Design and Development	68%
S - Science/Scientific Method	68%
S - Data Management	67%
S - Statistics and statistical modeling	63%
B - Business development	58%
B - Budgeting	58%
P - Database Administration	55%
M - Graphical Models	54%
M - Algorithms and Simulations	53%
T - Machine Learning	52%
T - Managing unstructured data	52%
M - Optimization	49%
P - Front-end Programming	49%
B - Governance and Compliance	48%
P - Systems Administration and Design	46%
M - Bayesian Statistics	45%
P - Back-end Programming	45%
T - NLP and text mining	39%
T - Big and distributed data	36%
P - Cloud Management	30%

Researcher

Data Skill	% who have skill
S - Communication	90%
S - Data Mining and Viz Tools	81%
M - Math	80%
S - Science/Scientific Method	78%
S - Statistics and statistical modeling	75%
T - Managing structured data	73%
S - Data Management	69%
B - Project management	68%
T - Machine Learning	58%
M - Optimization	56%
M - Algorithms and Simulations	55%
B - Product Design and Development	54%
M - Graphical Models	51%
M - Bayesian Statistics	50%
T - Managing unstructured data	46%
P - Database Administration	46%
B - Budgeting	44%
B - Business development	43%
T - NLP and text mining	41%
P - Systems Administration and Design	40%
P - Front-end Programming	38%
P - Back-end Programming	35%
B - Governance and Compliance	34%
T - Big and distributed data	32%
P - Cloud Management	21%

Note: % who have skill reflects percent of respondents who indicated that they have, at least, an intermediate level of proficiency.
Business Manager (e.g., leader, business person, entrepreneur) N = 250; Developer (e.g., developer, engineer) N = 222; Creative (e.g., Jack of all trades, artist, hacker) N = 221; Researcher (e.g., researcher, scientist, statistician) N = 353

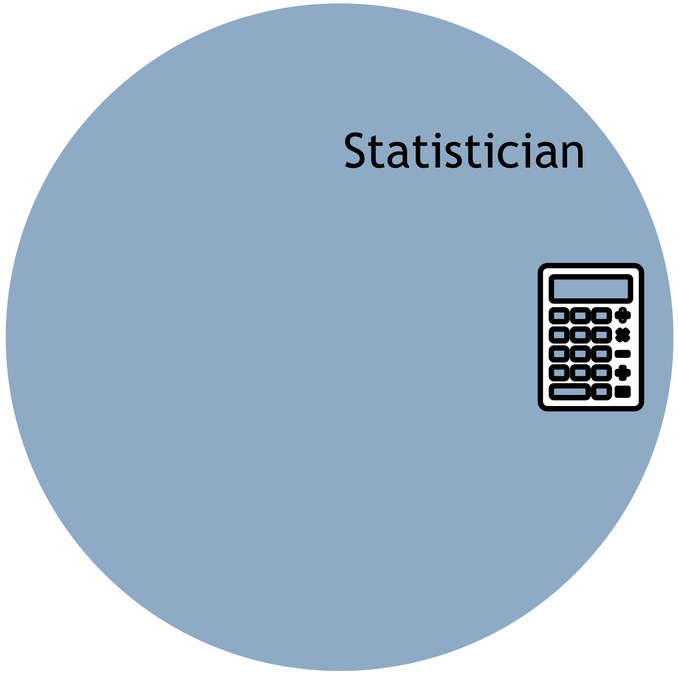
http://www.coffeewithshiva.com/tag/programming/

Data Scientist

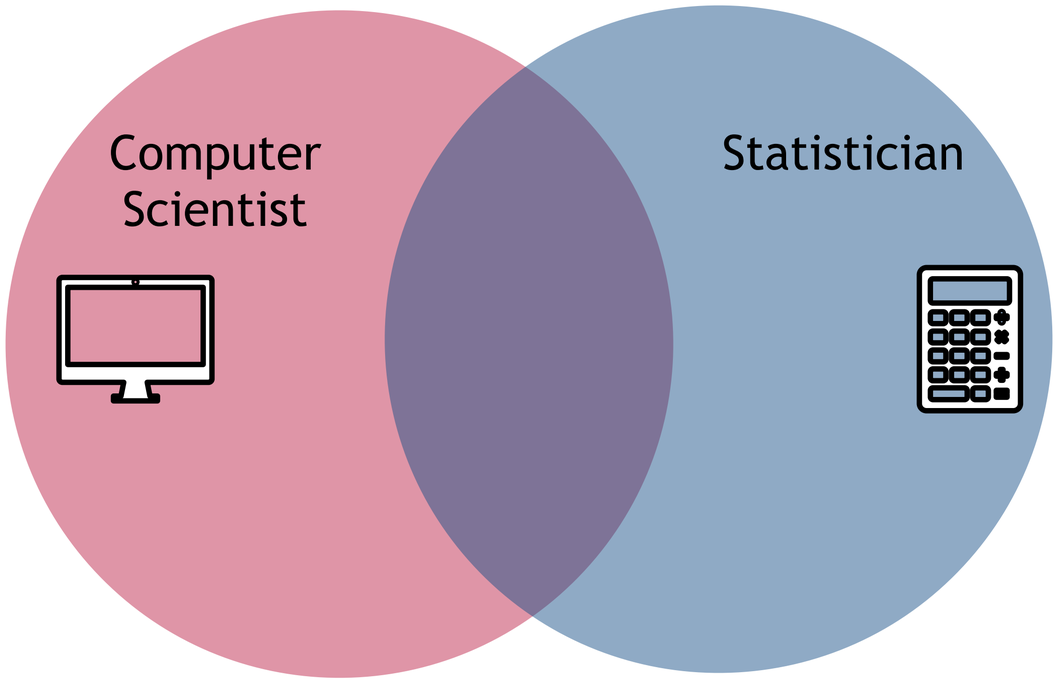
1. S – Communication (87% possess skill)
2. T – Managing Structured data (75%)
3. M – Math (71%)
4. B – Project management (71%)
5. S – Data Mining and Viz Tools (71%)
6. S – Science/Scientific Method (65%)
7. S – Data Management (65%)
8. B – Product design and development (59%)
9. S – Statistics and statistical modeling (59%)
10. B – Business development (53%)

http://www.coffeewithshiva.com/tag/programming/

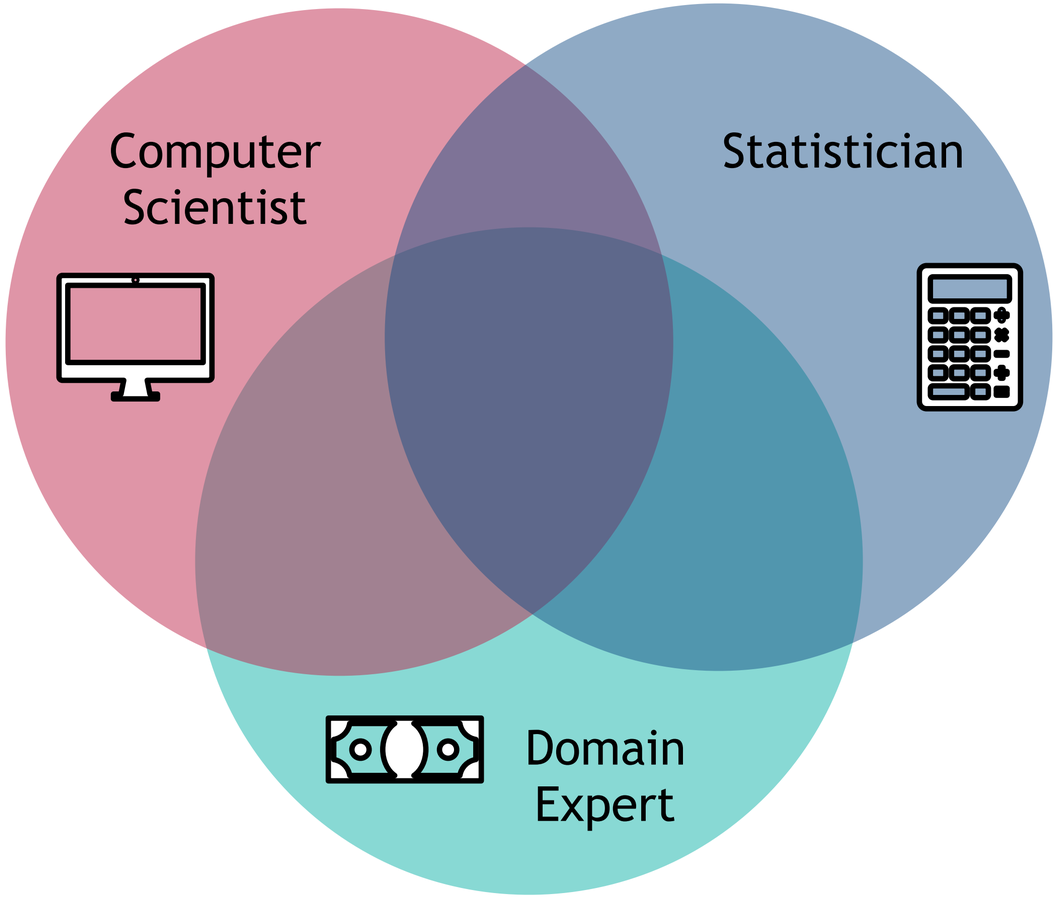
Data Scientist



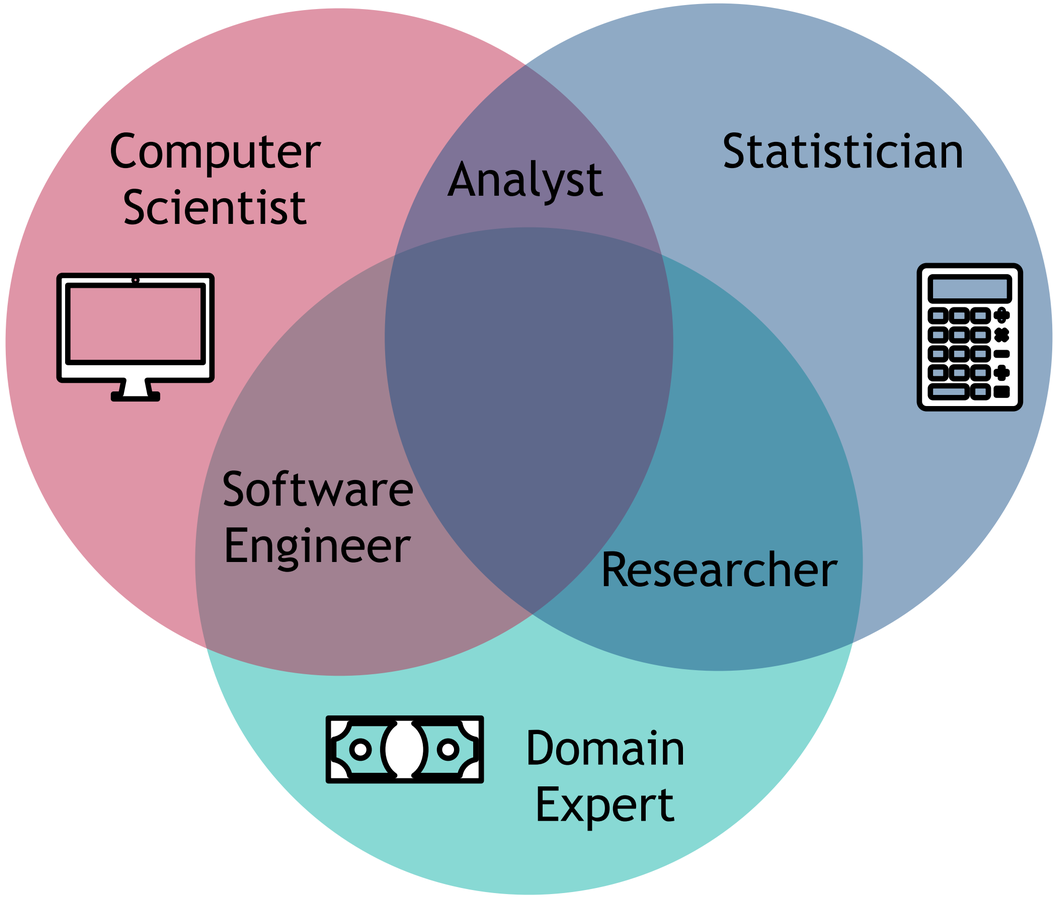
Data Scientist



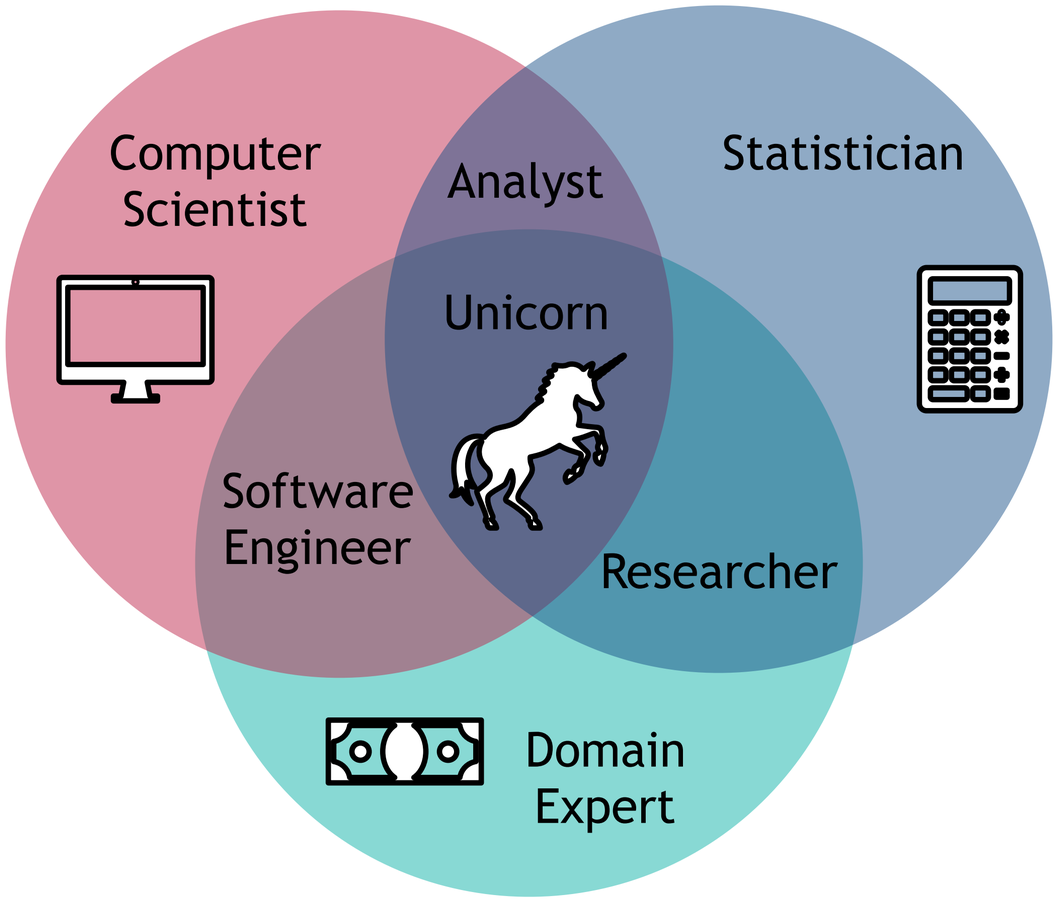
Data Scientist



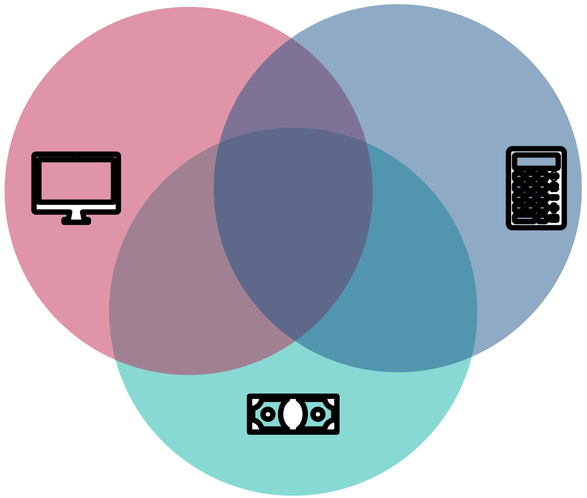
Data Scientist



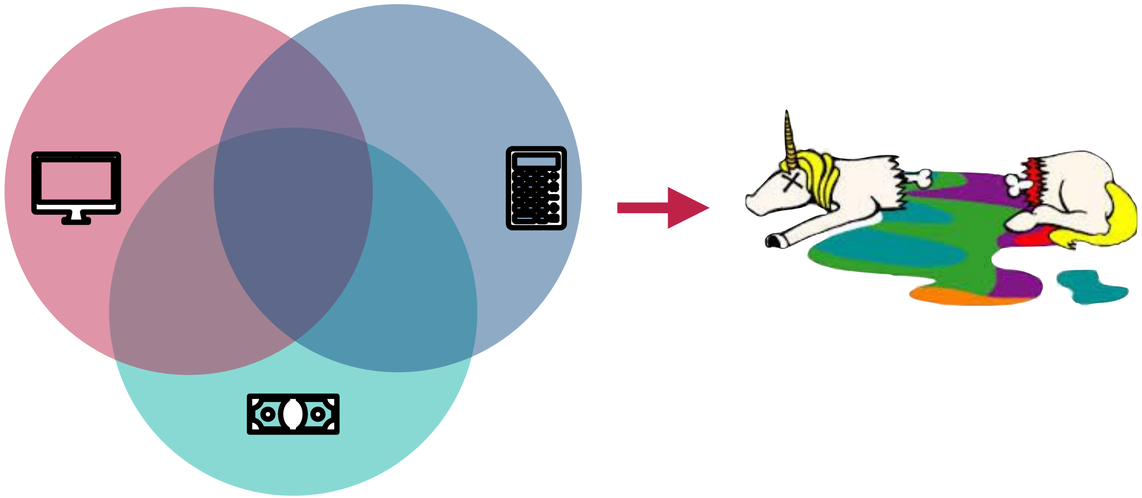
Data Scientist



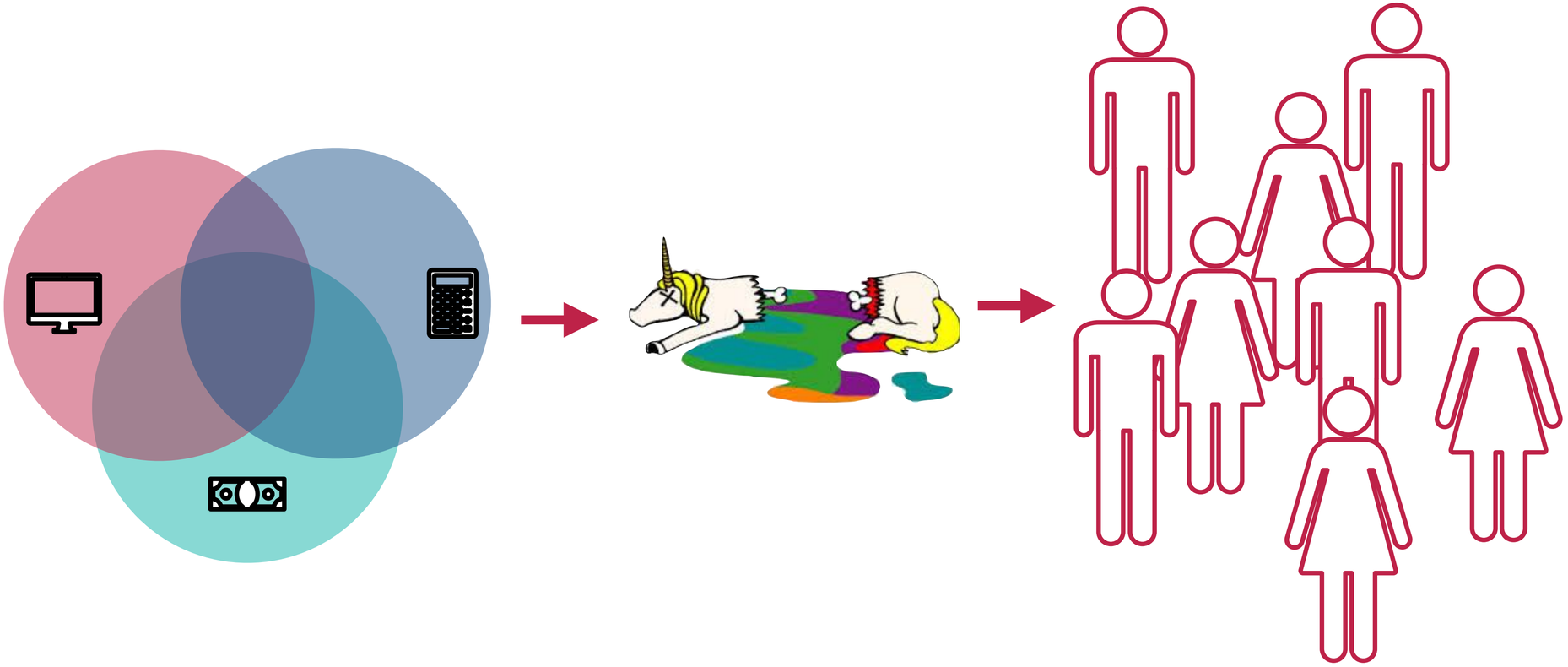
The importance of Team in Data Science



The importance of Team in Data Science

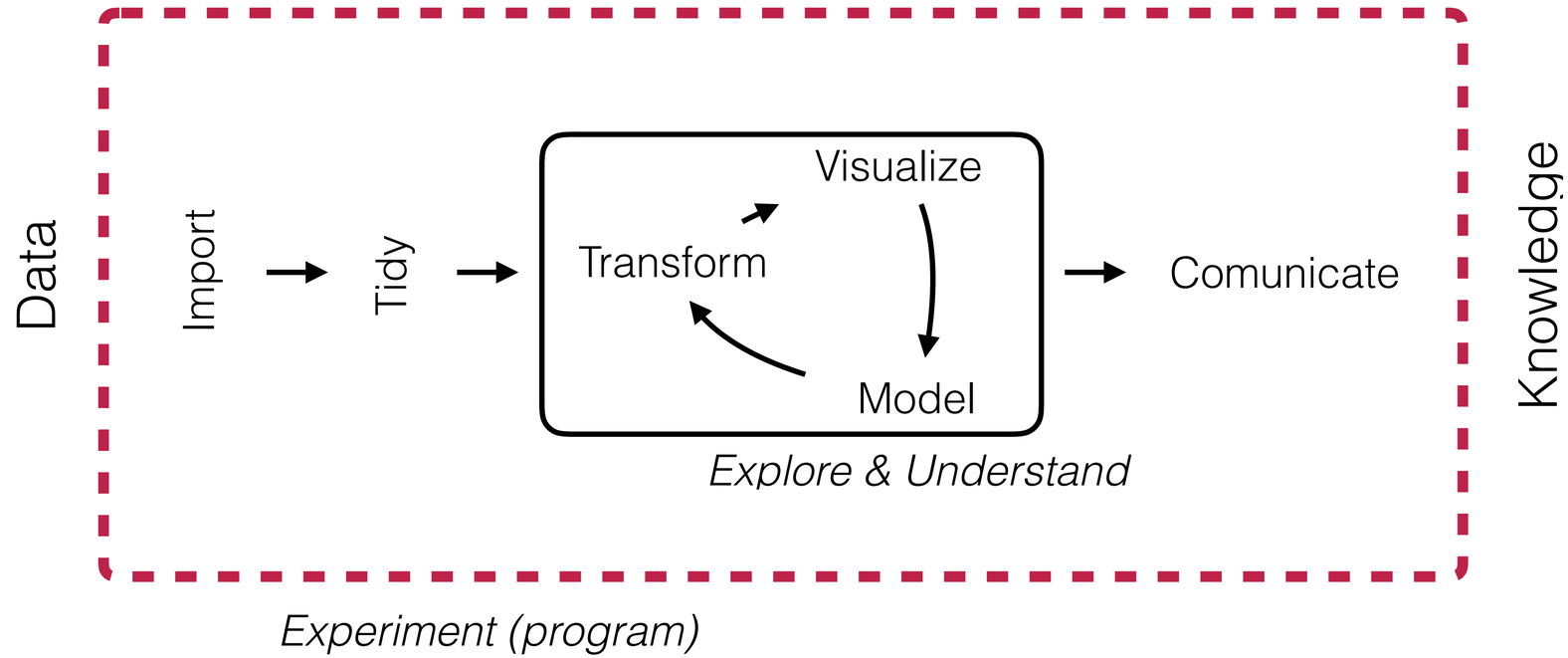


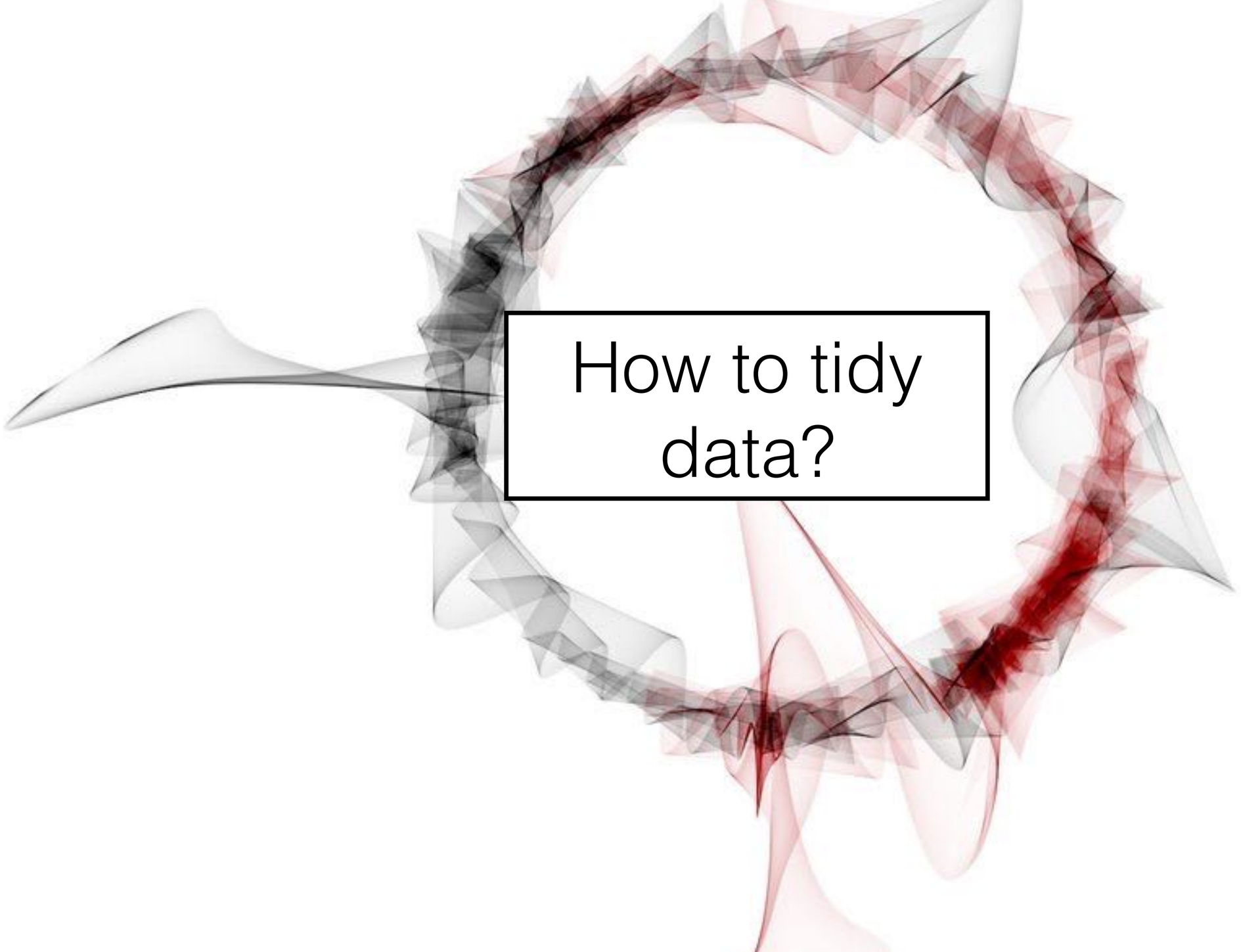
The importance of Team in Data Science



R for Data Science

DS Workflow: Which Verb you Didn't expect?





How to tidy data?

The Tydyverse

What is tidy data? Data format that:

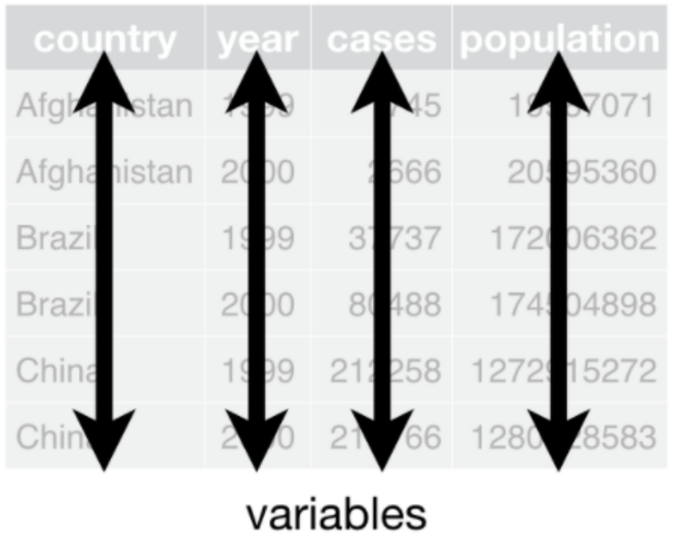
- Makes data analysis easy
- Is easy to model, visualize and transform
- Facilitate the creative process

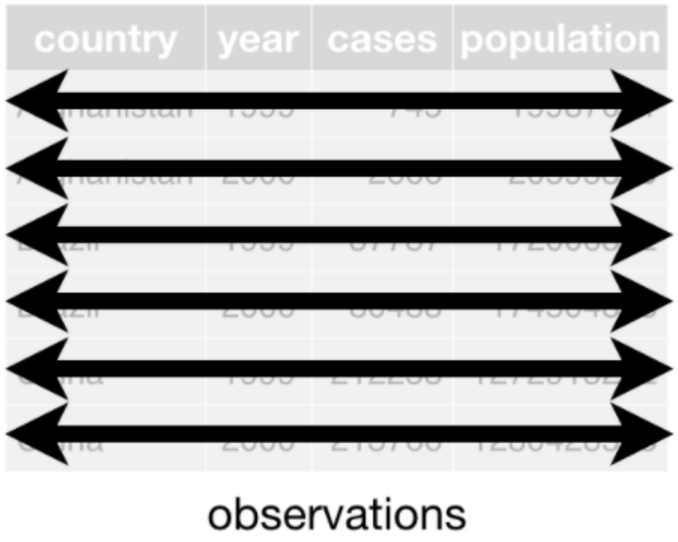
The Tydyverse

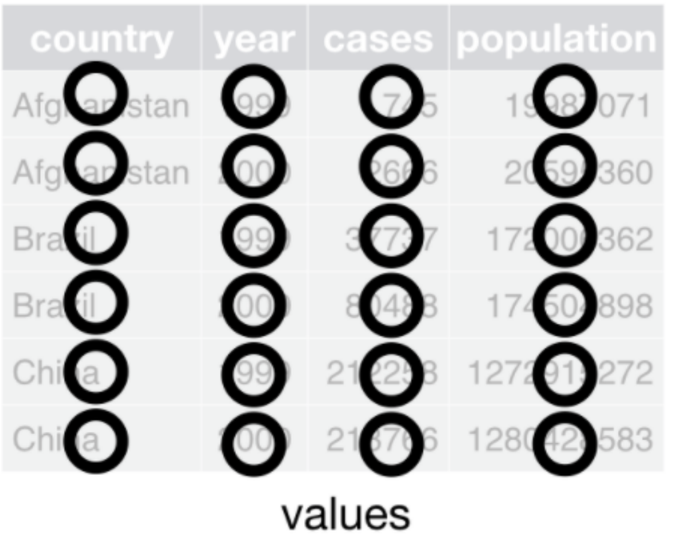
There are three interrelated rules which make a dataset tidy:

1- Each variable must have its own column.
2- Each observation must have its own row.
3- Each value must have its own cell.

The Tydyverse







Example of Messiness

	Pregnant	Not Pregnant
Male	0	5
Female	1	4

There are three variables in this data set. What are they?

Example of Messiness

Pregnant	Sex	Freq
no	female	4
no	male	5
yes	female	1
yes	male	0

- R is an **open source** programming language for statistical computing and graphics tha is supported by the R Foundation for Statistical Computing.

- Polls, surveys of **data miners**, and studies of scholarly literature databases show that R's popularity has increased substantially in recent years.

https://www.r-project.org/

Alternatives?

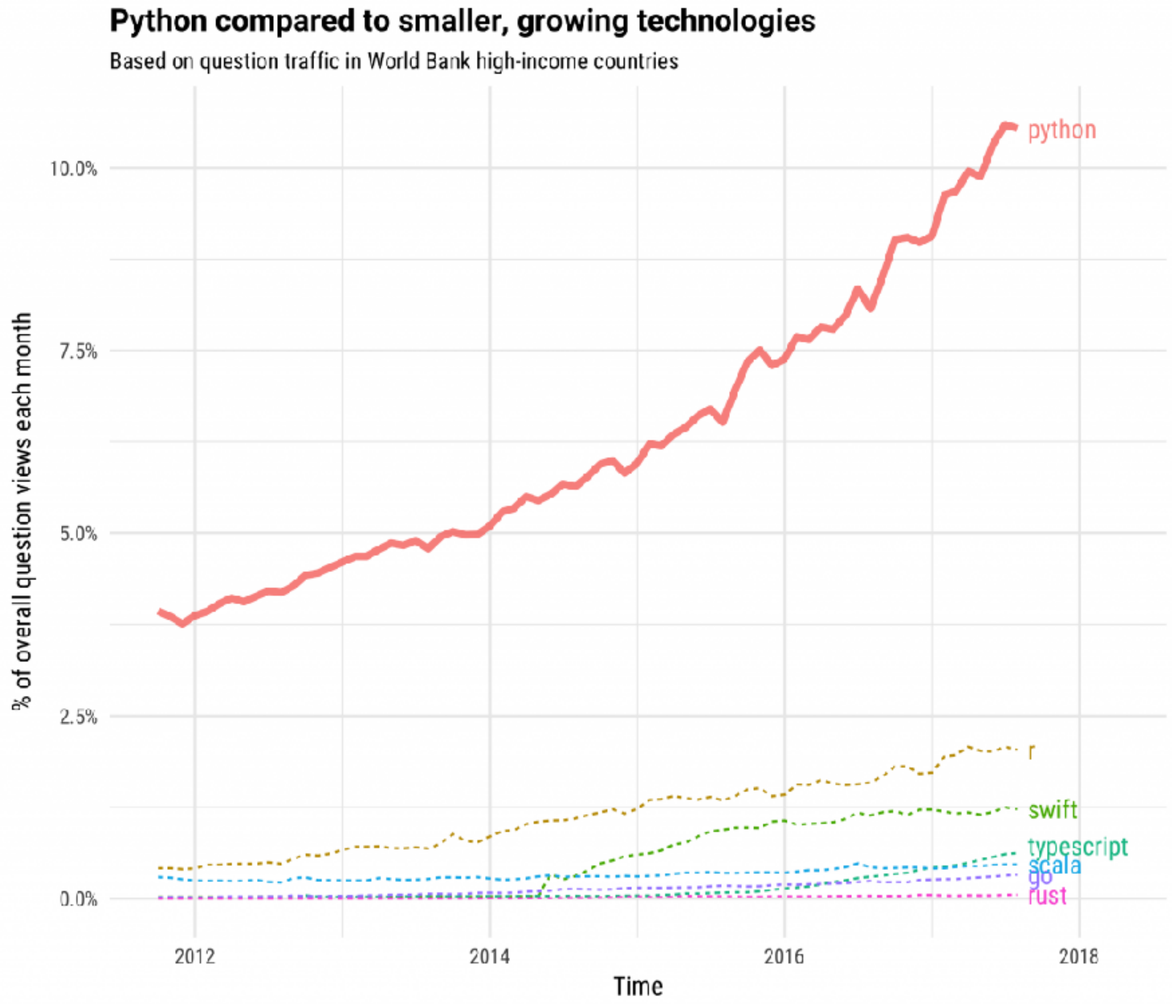
Alternatives?

VS

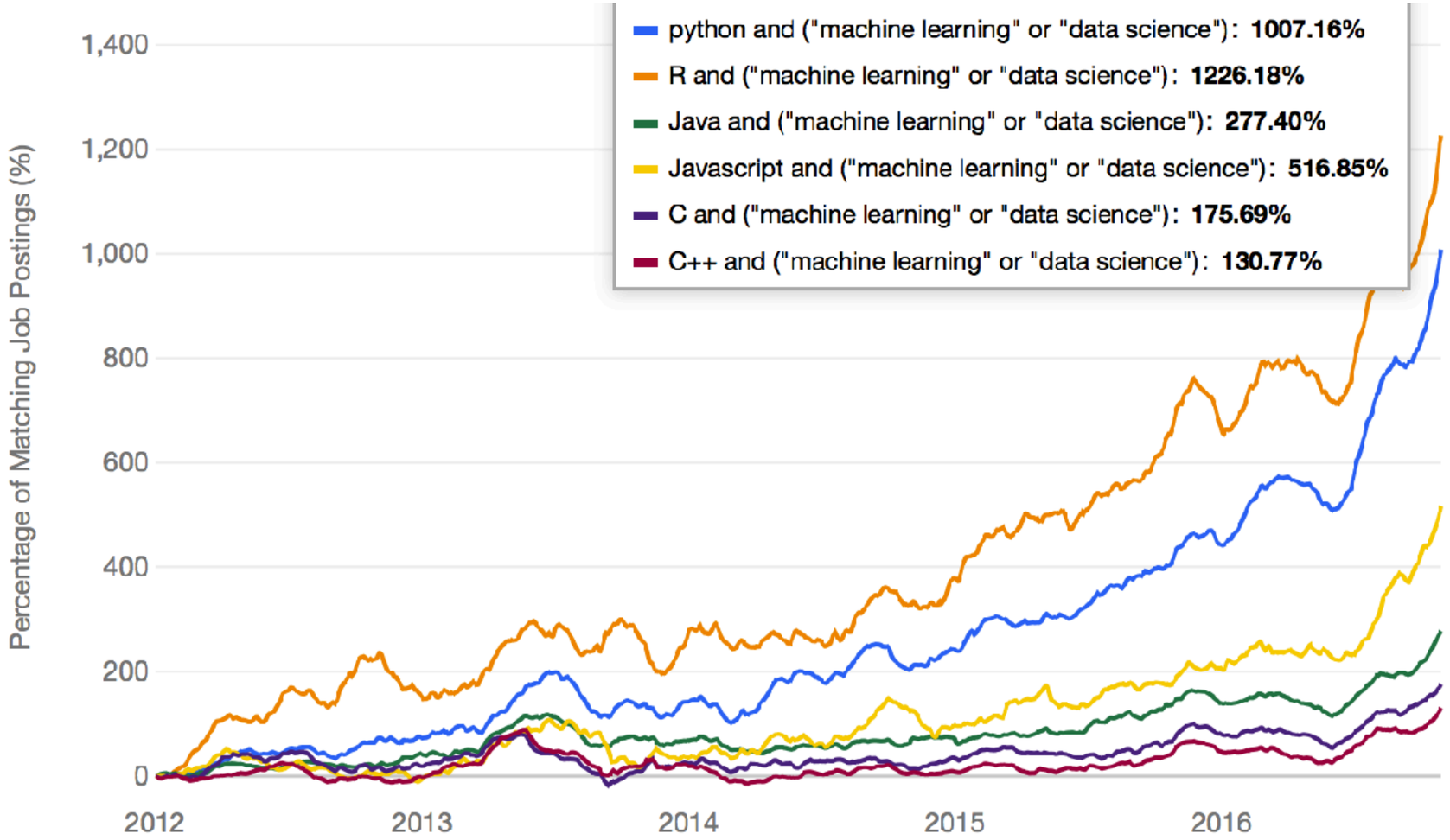
https://peerj.com/preprints/3183/

VS

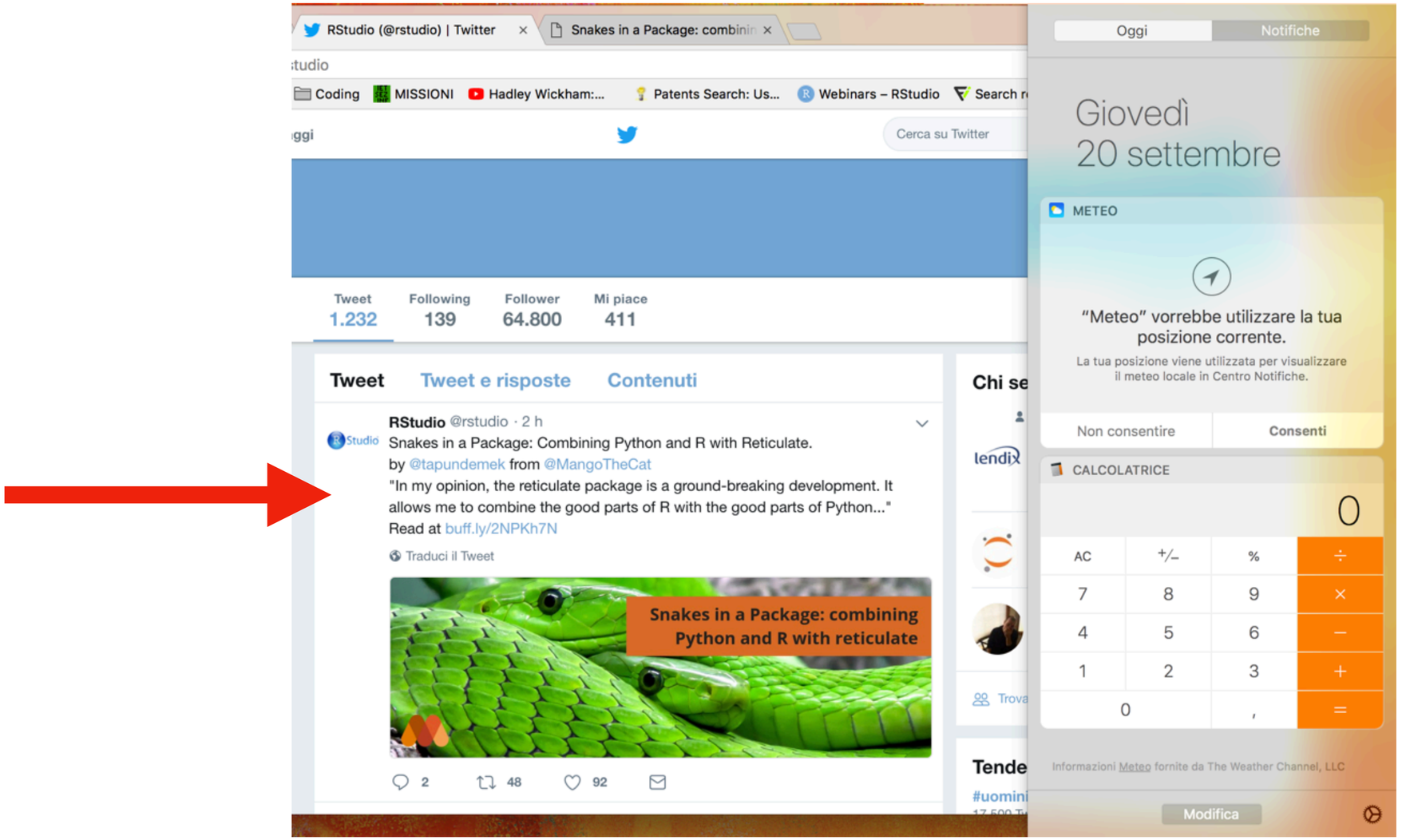
Why?


Python compared to smaller, growing technologies
Based on question traffic in World Bank high-income countries
% of overall question views each month
10.0%
7.5%
5.0%
2.5%
0.0%
2012
2014
2016
2018
Time
python
r
swift
typescript
scala
go
rust

Why?


python and ("machine learning" or "data science"): 1007.16%
R and ("machine learning" or "data science"): 1226.18%
Java and ("machine learning" or "data science"): 277.40%
Javascript and ("machine learning" or "data science"): 516.85%
C and ("machine learning" or "data science"): 175.69%
C++ and ("machine learning" or "data science"): 130.77%
Percentage of Matching Job Postings (%)
1,400
1,200
1,000
800
600
400
200
0
2012
2013
2014
2015
2016

Anyway…


RStudio (@rstudio) | Twitter
Snakes in a Package: combinin
Coding
MISSIONI
Hadley Wickham:...
Patents Search: Us...
Webinars – RStudio
Cerca su Twitter
Tweet 1.232
Following 139
Follower 64.800
Mi piace 411
Tweet
Tweet e risposte
Contenuti
RStudio @rstudio · 2 h
Snakes in a Package: Combining Python and R with Reticulate.
by @tapundemek from @MangoTheCat
"In my opinion, the reticulate package is a ground-breaking development. It allows me to combine the good parts of R with the good parts of Python..."
Read at buff.ly/2NPKh7N
Traduci il Tweet
Snakes in a Package: combining Python and R with reticulate
2
48
92
Oggi
Notifiche
Giovedì
20 settembre
METEO
"Meteo" vorrebbe utilizzare la tua posizione corrente.
La tua posizione viene utilizzata per visualizzare il meteo locale in Centro Notifiche.
Non consentire
Consenti
CALCOLATRICE
0
AC
+/–
%
÷
7
8
9
×
4
5
6
–
1
2
3
+
0
,
=
Informazioni Meteo fornite da The Weather Channel, LLC
Modifica

link

Here it Comes the King…

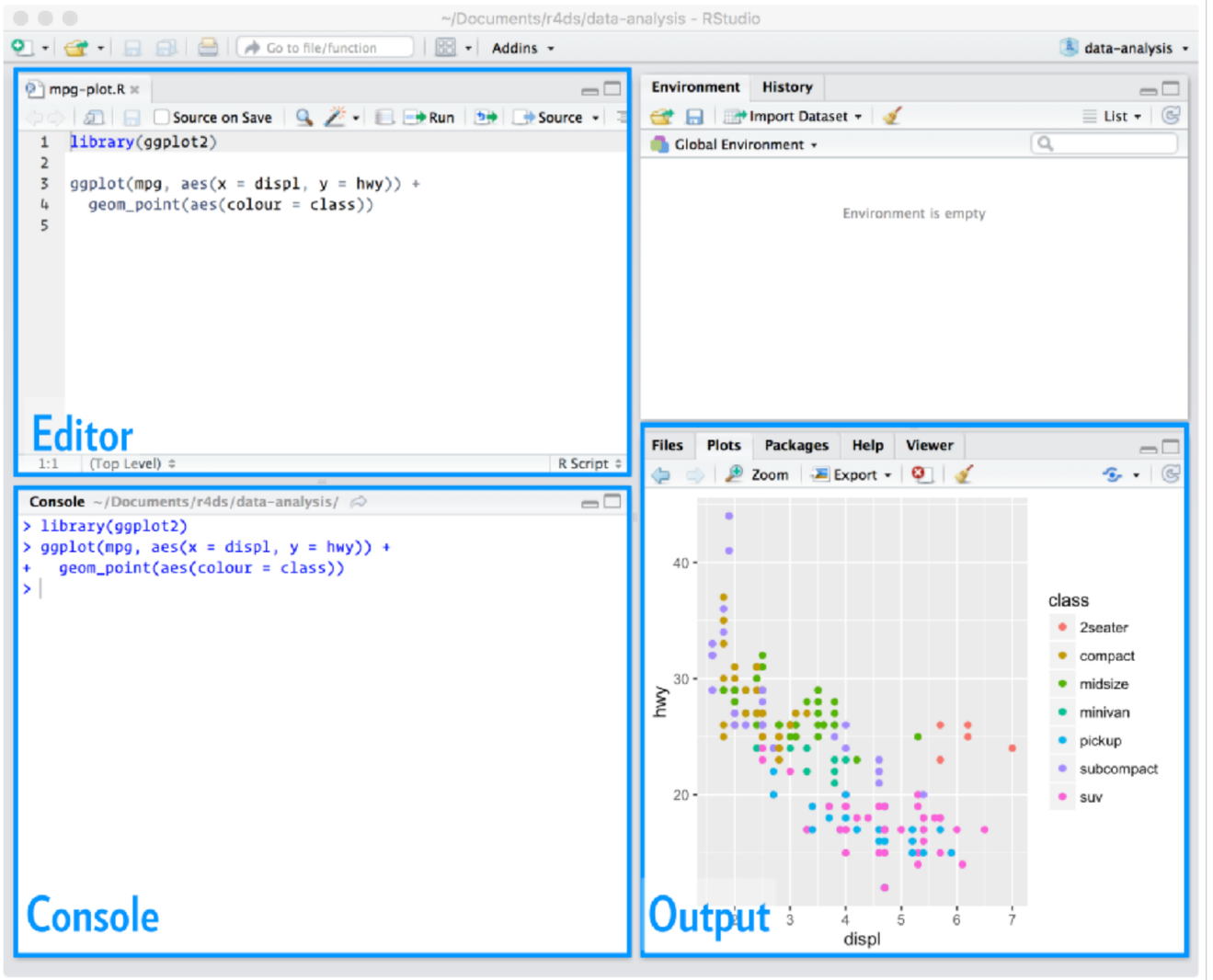
- Is a free and **open source** integrated development environment (IDE) for R.

- Available in Desktop edition and Server edition, free and commercial

- Written in the **C++** programming language

- Version 1.0 was released on 1 November **2016**

RStudio®


~/Documents/r4ds/data-analysis - RStudio
Go to file/function
Addins
data-analysis
mpg-plot.R
Source on Save
Run
Source
1 library(ggplot2)
2
3 ggplot(mpg, aes(x = displ, y = hwy)) +
4 geom_point(aes(colour = class))
5
Editor
1:1
(Top Level)
R Script
Environment
History
Import Dataset
List
Global Environment
Environment is empty
Console ~/Documents/r4ds/data-analysis/
> library(ggplot2)
> ggplot(mpg, aes(x = displ, y = hwy)) +
+ geom_point(aes(colour = class))
>
Console
Files
Plots
Packages
Help
Viewer
Zoom
Export
hwy
displ
class
2seater
compact
midsize
minivan
pickup
subcompact
suv
Output

RStudio®

	Mac	PC
Last Command	↑	↑
Last Matching Command	Command + ↑	Control + ↑
Tab Completion	Tab	Tab

RStudio®

	Mac	PC
Run Selection	Command + Enter	Control + Enter
Source	Command + Shift + S	Control + Shift + S
Source with Echo	Command + Shift + Enter	Control + Shift + Enter
Move focus to console	Control + 2	Control + 2
Move focus to source	Control + 1	Control + 1

RStudio®

	Mac	PC
<-	Option + -	Alt + -
%>%	Command + Shift + M	Control + Shift + M
Comment	Control + Shift + C	Control + Shift + C
Reflow Comment	Control + Shift + /	Control + Shift + /
Undo	Command + Z	Control + Z
Redo	Command + Shift + Z	Control + Shift + Z
Shortcuts Reference	Option + Shift + K	Alt + Shift + K

RStudio®

	Mac	PC
Multiple cursors	Control + Option + ↑	Control + Alt + ↑
	Control + Option + ↓	Control + Alt + ↓
	Control + Option + Click	Control + Alt + Click

RStudio®

	Mac	PC
Go to line	Command + Shift + Option + G	Shift + Alt + G
Find and Replace	Command + F	Control + F
Find in files	Command + Shift + F	Control + Shift + F
Close all folds	Command + Option + O	Alt + O
Open all folds	Command + Shift + Option + O	Alt + Shift + O

RStudio®

	Mac	PC
Go to line	Command + Shift + Option + G	Shift + Alt + G
Find and Replace	Command + F	Control + F
Find in files	Command + Shift + F	Control + Shift + F
Close all folds	Command + Option + O	Alt + O
Open all folds	Command + Shift + Option + O	Alt + Shift + O

An Incredible (Free) Learning Source

Lets get practical..

https://www.datacamp.com/community/tutorials/tidyverse-tutorial-r