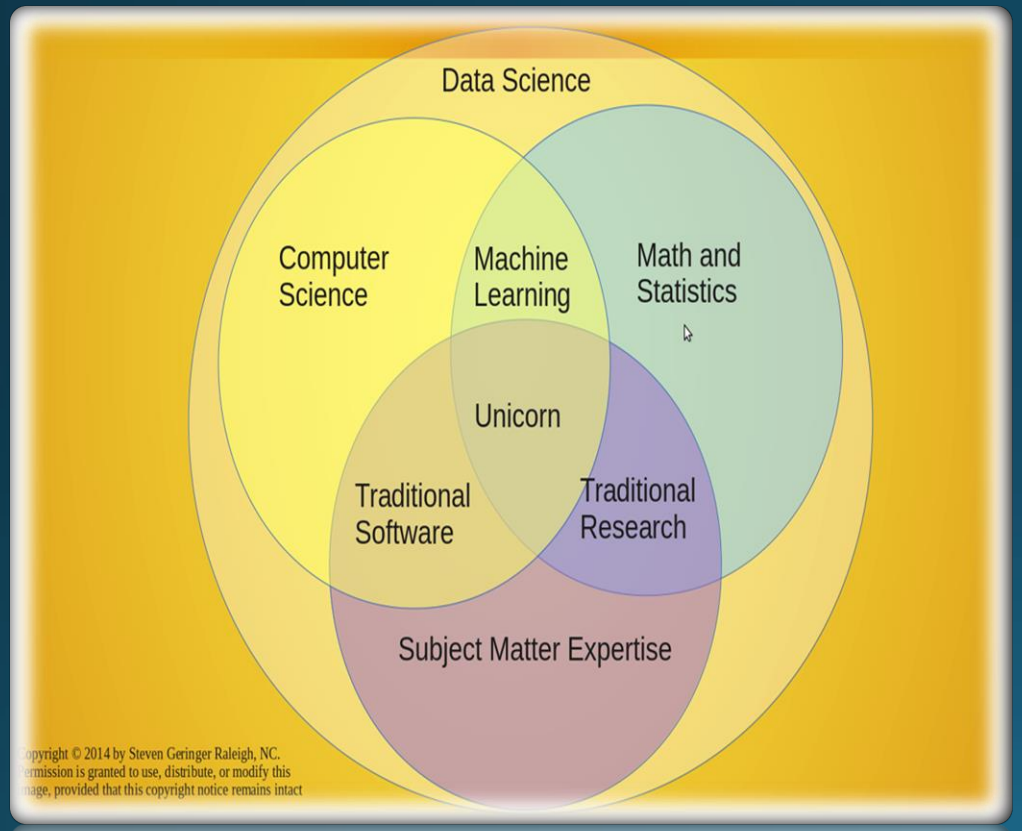
Philippe Wee

What is data science and what tools are there?

Meetup Structure

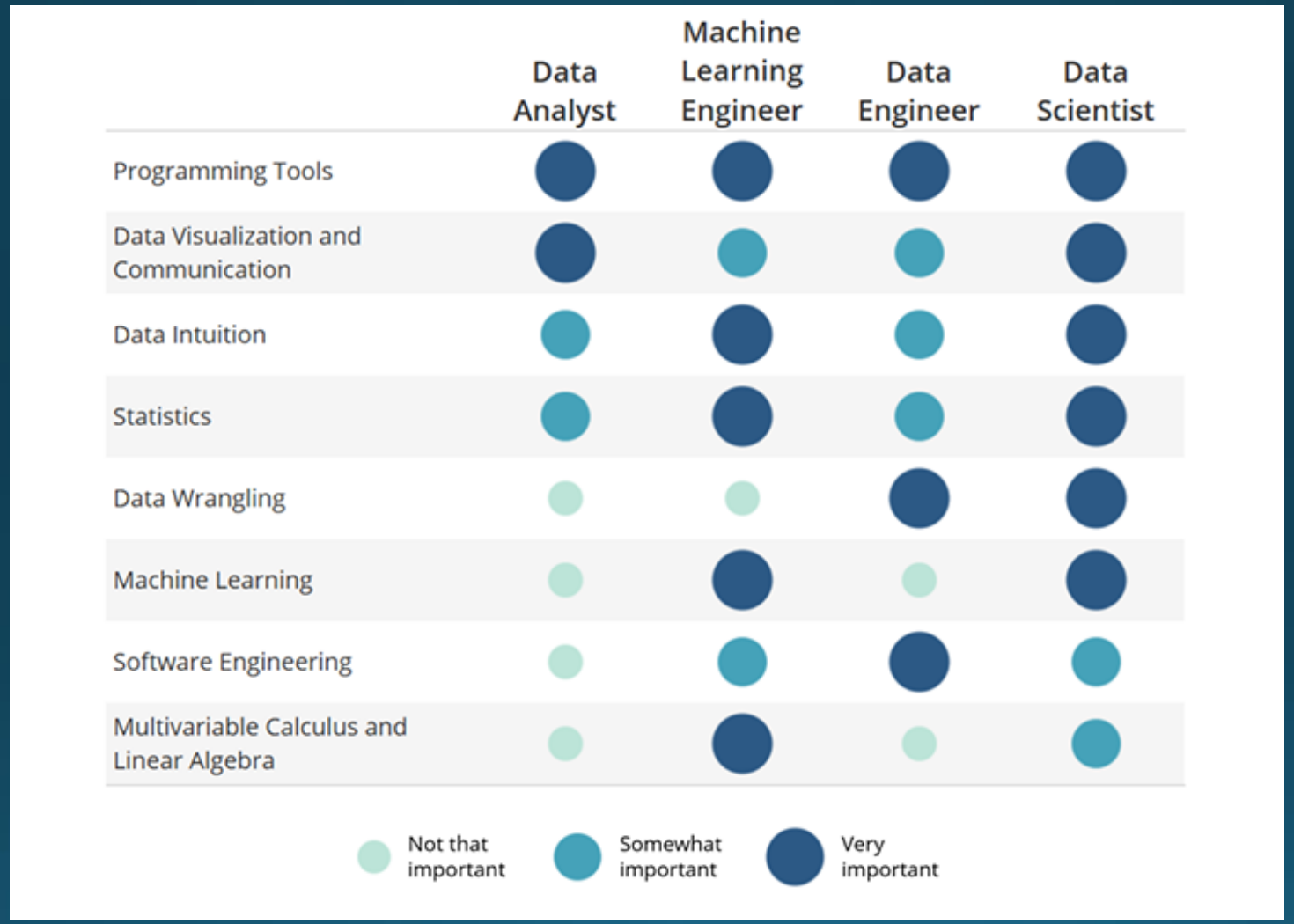
- **Regular meetups: Twice a month**

Data Science



- Interdisciplinary field
- Unification of theories, concepts, tools, and technologies that enable the extraction of valuable information from raw data that may be used for multiple purposes; such as decision making, product development, trend analysis, and forecasting.

http://www.anlytcs.com/2014/01/data-science-venn-diagram-v20.html



https://blog.udacity.com/2014/11/data-science-job-skills.html

Popular Data Science Tools

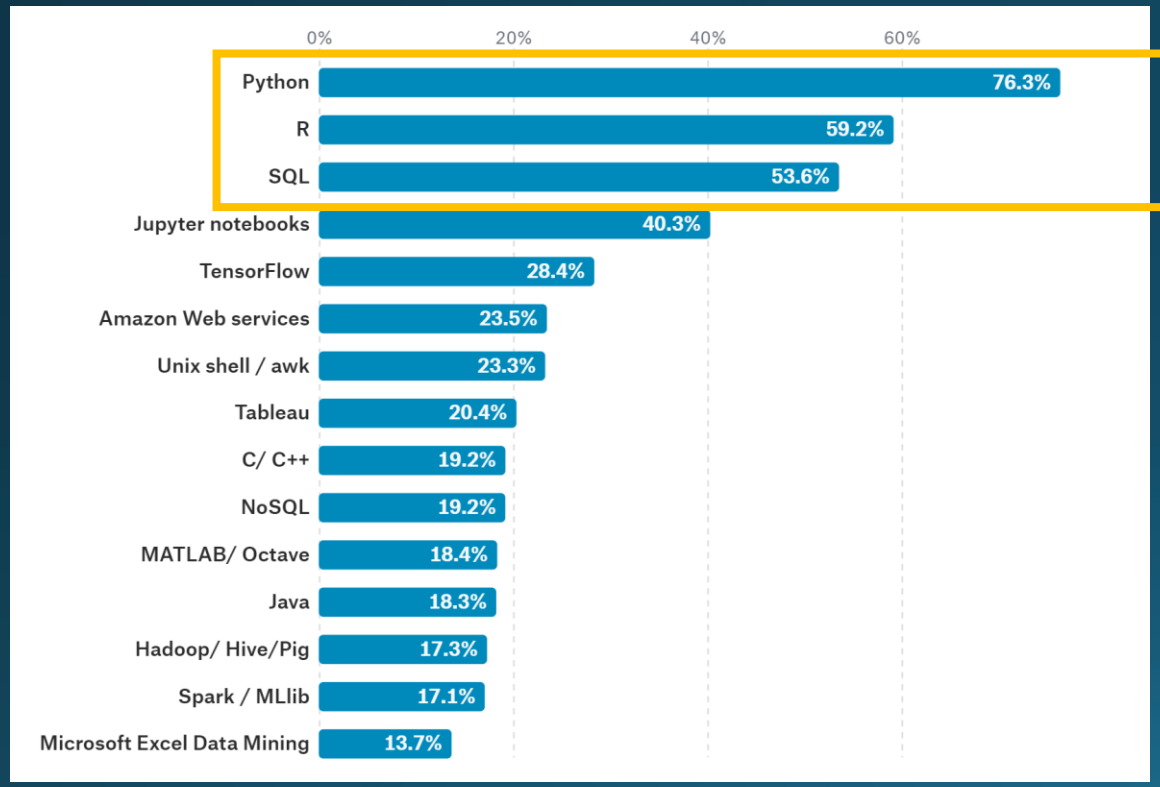


Table 1: Top Analytics/Data Science Tools in 2017 KDnuggets Poll

Tool	2017 % Usage	% change 2017 vs 2016	% alone
Python	52.6%	15%	0.2%
R language	52.1%	6.4%	3.3%
SQL language	34.9%	-1.8%	0%
RapidMiner	32.8%	0.7%	13.6%
Excel	28.1%	-16%	0.1%
Spark	22.7%	5.3%	0.2%
Anaconda	21.8%	37%	0.8%
Tensorflow	20.2%	195%	0%
scikit-learn	19.5%	13%	0%
Tableau	19.4%	5.0%	0.4%
KNIME	19.1%	6.3%	2.4%

https://www.kdnuggets.com/2017/05/poll-analytics-data-science-machine-learning-software-leaders.html

https://www.kaggle.com/surveys/2017

Some of Many Other Tools

- Programming languages: Julia, Perl, Ruby, scala
- Data mining: Rapidminer, SSAS, Knime
- Big data: Hadoop, Spark, Storm
- Libraries: Tidyverse, Epic, Numpy, Dplyr, Ggplot2, PyTorch, Tensorflow, Scikit-learn
- BI: Qlik

kaggle

- Platform for data science and machine learning
  - Hosts business competitions
  - Public datasets
  - Cloud based kernels for code
  - Tutorials

TRY THINGS OUT AND IMPLEMENT IT!

Airbnb

- Predict which country new users will book there first trip to
- Explore and analyze the datasets

Next Meetup on JUNE 14

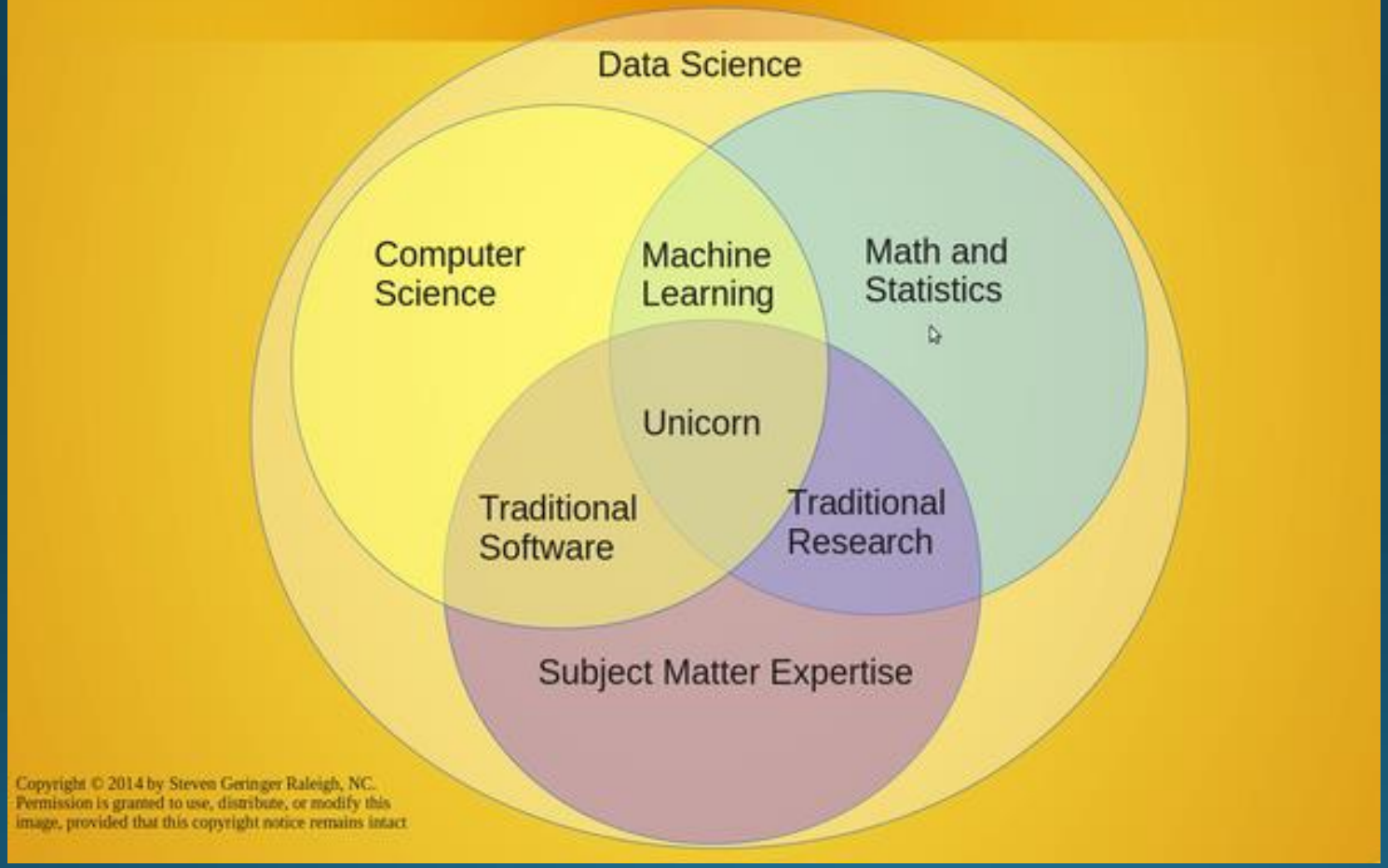
Intro to R by Olof Rännbäck-Garpinger

WORKING ON DATASETS

Join the Meetup Group to keep updated on what's going on

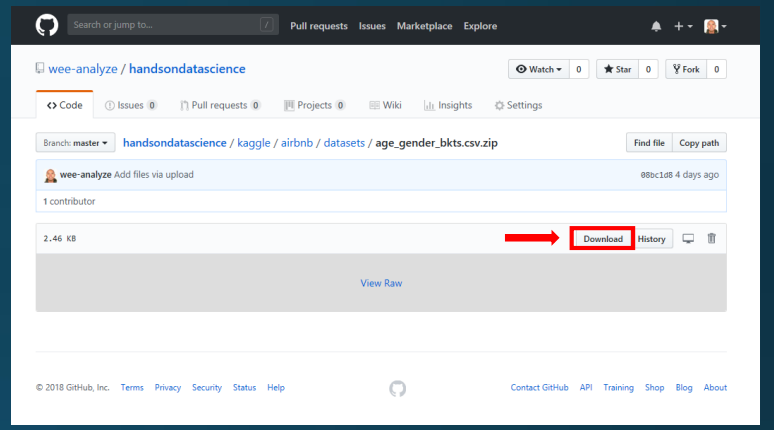
Hands On Data Science (Malmö, Sweden)

Data Science Venn Diagram v2.0



Beginner Exercises

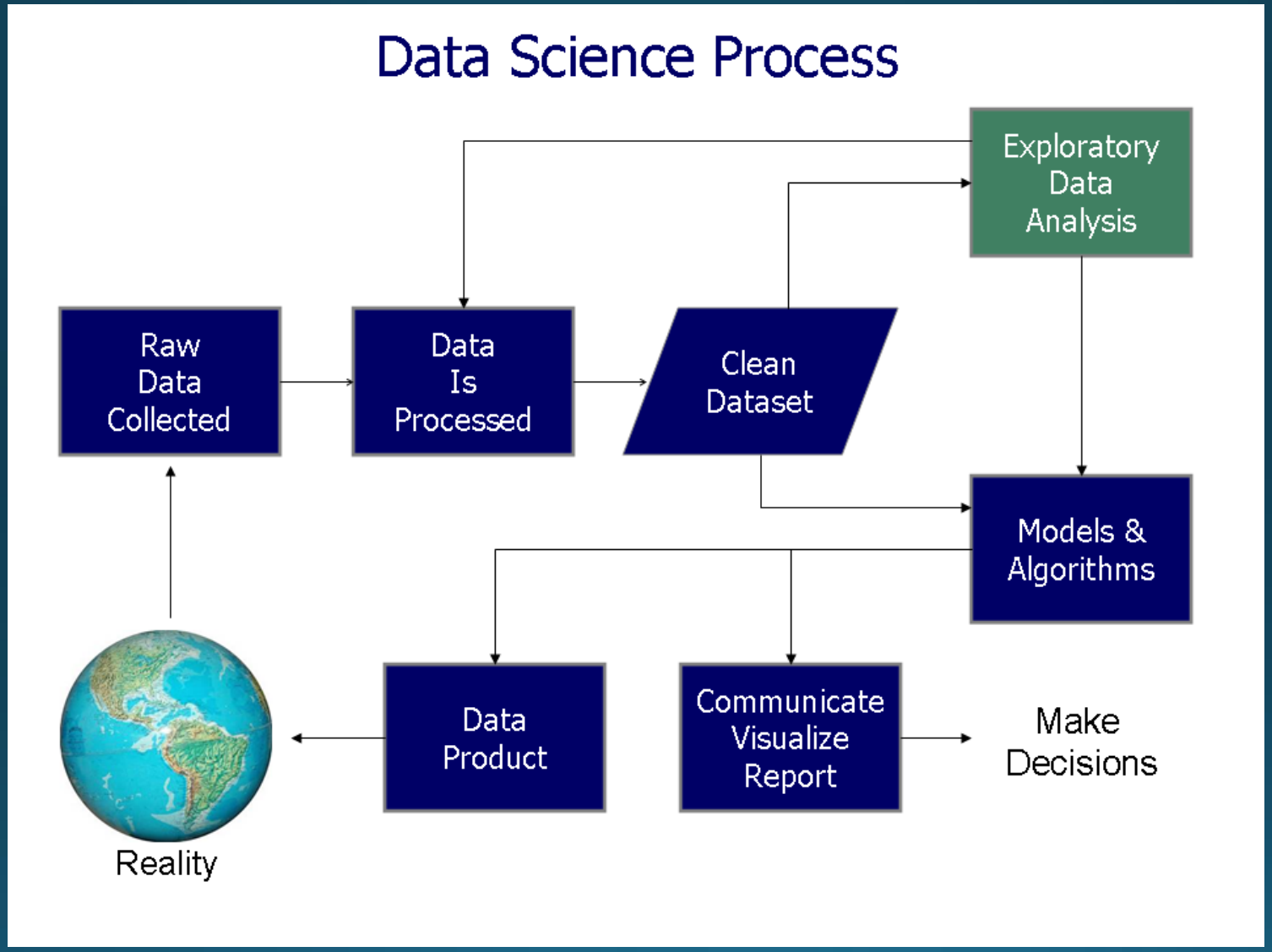
https://github.com/wee-analyze/handsondatascience/tree/master/kaggle/airbnb/datasets



- Train_user_2 set
  - Load datasets
  - Convert "account_created" column into a timestamp
  - Change column "affiliate_provider" to "marketing"
  - Return all the different categories that are in the newly made "marketing" column
  - Count how many users used Chrome as their "first_browser_type"
  - Count how many users are between the ages of 23 and 36
  - Parse "data_account_created" column into datetime object

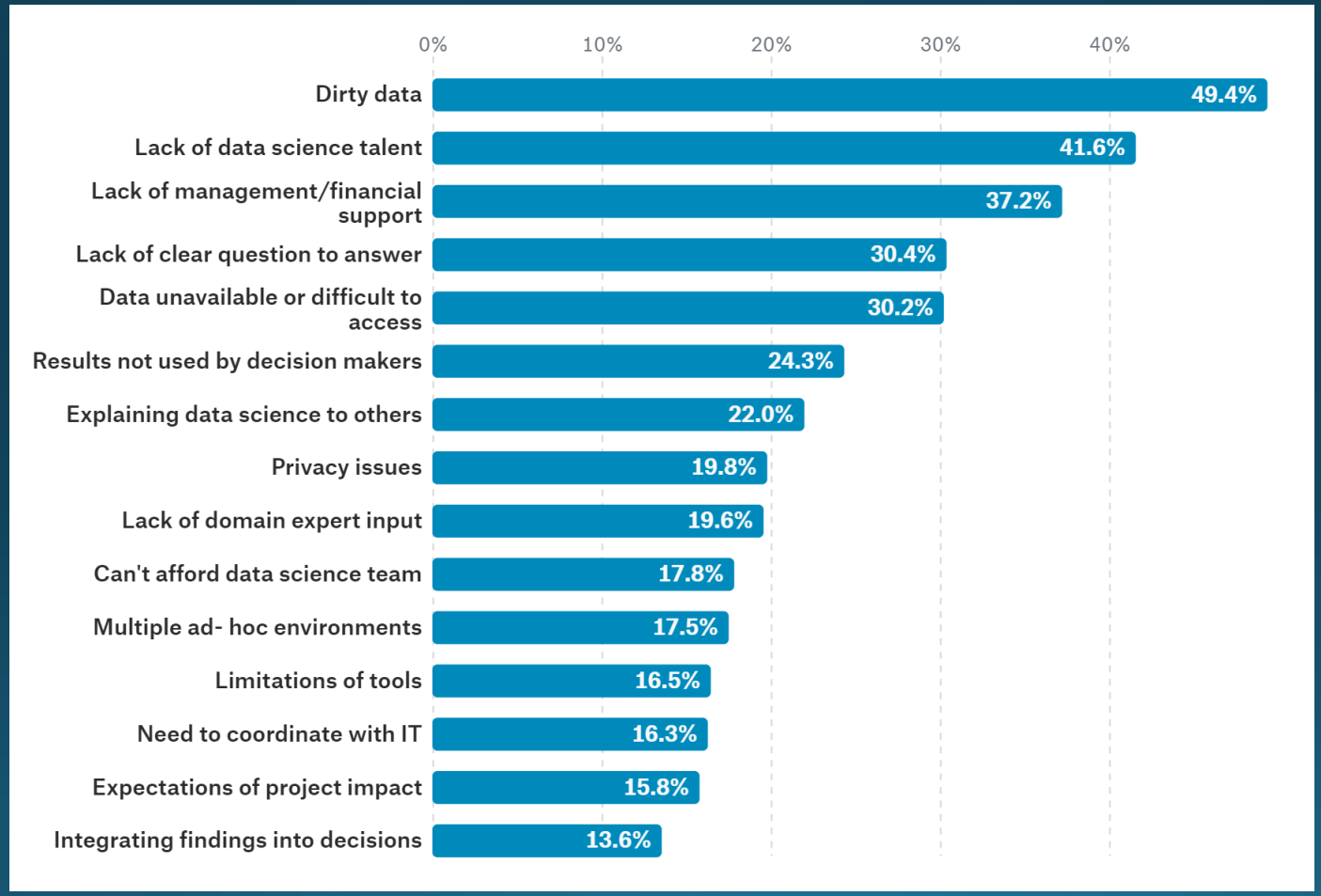
- Sessions set
  - Return all users who used "iPad Tablet"
  - how many MINUTES and SECONDS did those "iPad Tablet" users look at the Airbnb website
  - Return the date the "iPad Tablet" users created an Airbnb account

- Visualization
  - Make a histogram of the age distribution for females in the country IT
  - Make a histogram of the "country_destination" countries in the train_users_2 set



https://commons.wikimedia.org/wiki/File:Data_visualization_process_v1.png

Work Barriers



https://www.kaggle.com/surveys/2017