Applied Data Science

Joanne S. Luciano https://www.linkedin.com/in/joanneluciano

Visiting Associate Professor
Indiana University School of Informatics and Computing

INDIANA UNIVERSITY

Part 1- Think Like a Data Scientist

- **Module 1 Talking about Data**
- **Module 2 Data Exploration**
- **Module 3 Data Visualization**
- **Module 4 Talking about Results**

Module 1: Talking about Data

Topics:

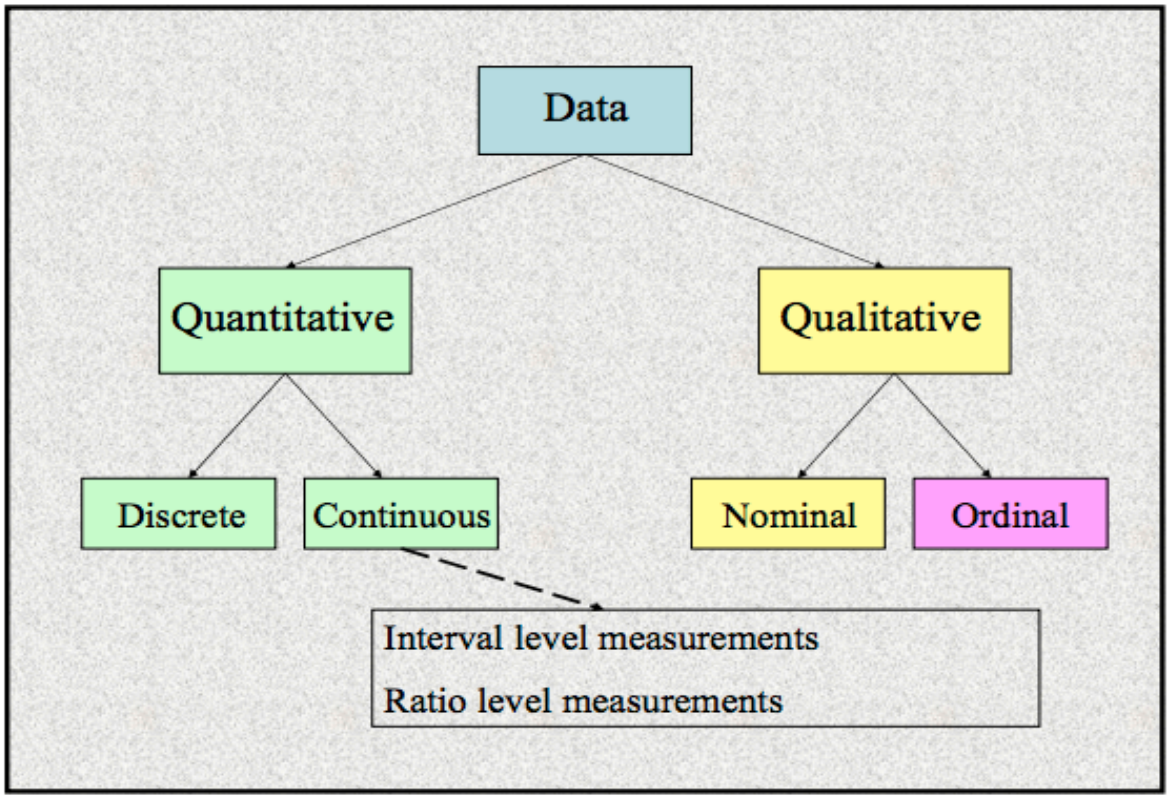
- Data – types and sources
- Data Science Workflow
- Data Science Tools

Learning Objectives: Students will be able to:

- Structure and execute a data science analysis and report the results.
- Perform exploratory data analysis, formal modeling, interpretation, and communication.
- Clean, integrate, analyze, and visually report analysis results using Tableau.

Data – Types of data

"Data!data!data!" he cried impatiently. "I can't make bricks without clay."
— Arthur Conan Doyle, The Adventure of the Copper Beeches

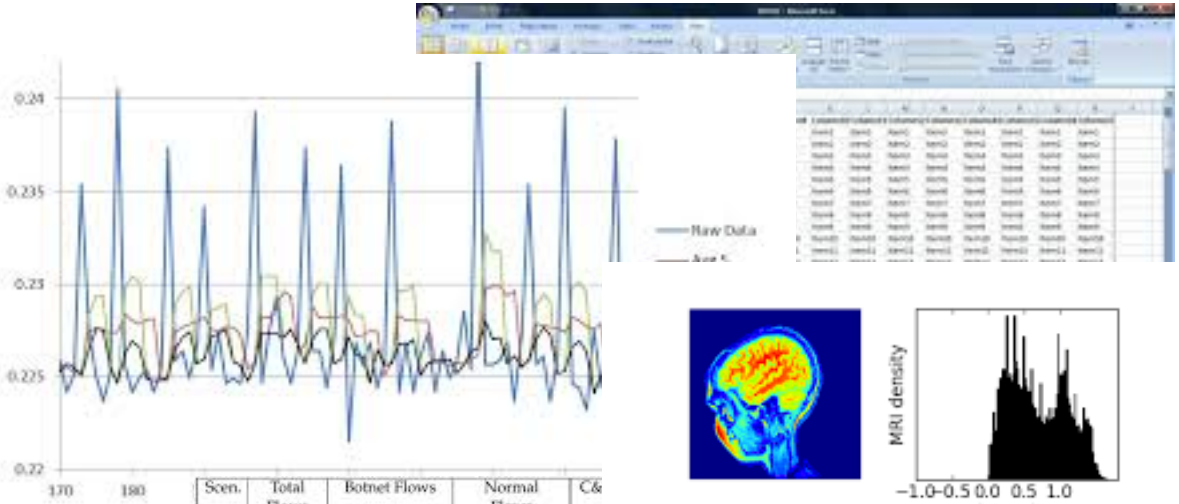


Types of Data

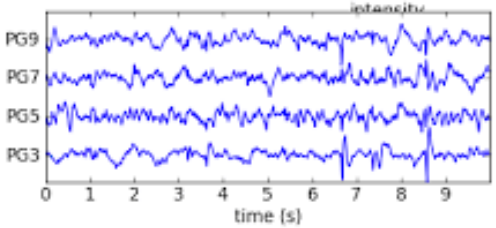
https://hsl.lib.umn.edu/biomed/help/primary-secondary-and-tertiary-sources-health-sciencesstatr

http://faculty.washington.edu/tamre/Lecture1b.pdf

Data – Sources of Data



Scen.	Total Flows	Botnet Flows	Normal Flows	C&
1	2,824,636	39,933(1.41%)	30,387(1.07%)	1,0
2	1,808,122	18,839(1.04%)	9,120(0.5%)	2,1
3	4,710,638	26,759(0.56%)	116,887(2.48%)	63
4	1,121,076	1,719(0.15%)	25,268(2.25%)	49
5	129,832	695(0.53%)	4,679(3.6%)	20
6	558,919	4,431(0.79%)	7,494(1.34%)	19
7	114,077	37(0.03%)	1,677(1.47%)	2
8	2,954,230	5,052(0.17%)	72,822(2.46%)	1,0
9	2,753,884	179,880(6.5%)	43,340(1.57%)	5,0
10	1,309,791	106,315(8.11%)	15,847(1.2%)	37
11	107,251	8,161(7.6%)	2,718(2.53%)	3(0.002%)
12	325,471	2,143(0.65%)	7,628(2.34%)	25(0.007%)
13	1,925,149	38,791(2.01%)	31,939(1.65%)	1,202(0.06%)



Sources of Data

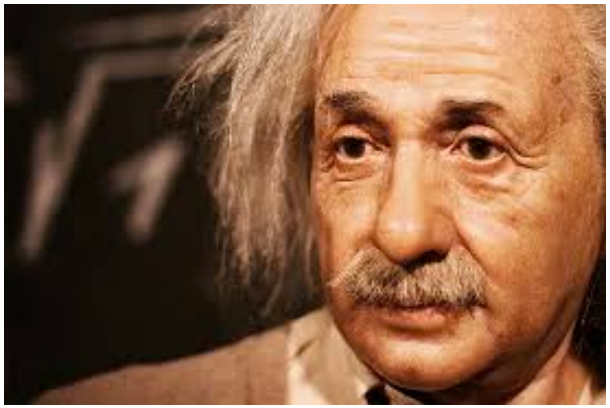
1. Primary Sources
2. Secondary Sources
3. Tertiary Sources

Caveat: Context determines whether a source is primary, secondary, or tertiary.
http://gethelp.library.upenn.edu/PORT/sources/primary_secondary_tertiary.html

Data Science

Concerned with the collection, preparation, analysis, visualization, management, and preservation of large collections of information.

Many skills are needed. Patience too!



In theory, theory and practice are the same. In practice, they are not.
-- AlbertEinstein

Data Science

Data Life Cycle

- What is data?
- Where does data come from?
- How is data created?
- How is it captured?
- How is it accessed?
- How is it stored?
- How is it maintained?
- How is it protected?
- How is it used?
- How is it re-used?
- How is it published?
- How is it archived?
- How is it destroyed?
- What are the best practices?
- What are the standards organizations?

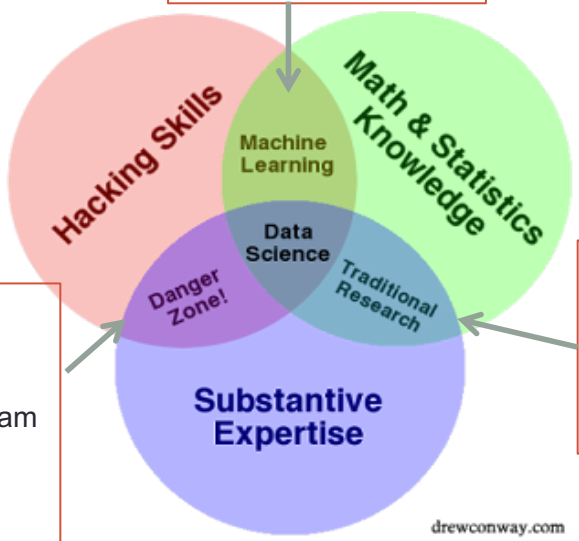
http://www.spirion.com/us/Content/Images/Solutions/lifecycle-management.png

Data Science

Machine Learning: Data plus math is machine learning, it is not data science.

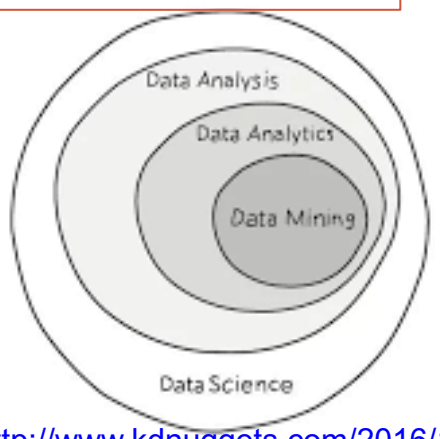
Hacking Skills: Data is a commodity traded electronically, therefore, in order to be in this market you need to be able to manipulate text files at the command-line, think algorithmically, and be interested in learning new tools.

Math & Statistics Knowledge: Once you have acquired and cleaned the data, the next step is to actually extract insight from it. You need to apply appropriate math and statistics methods, which requires at least a baseline familiarity with these tools.

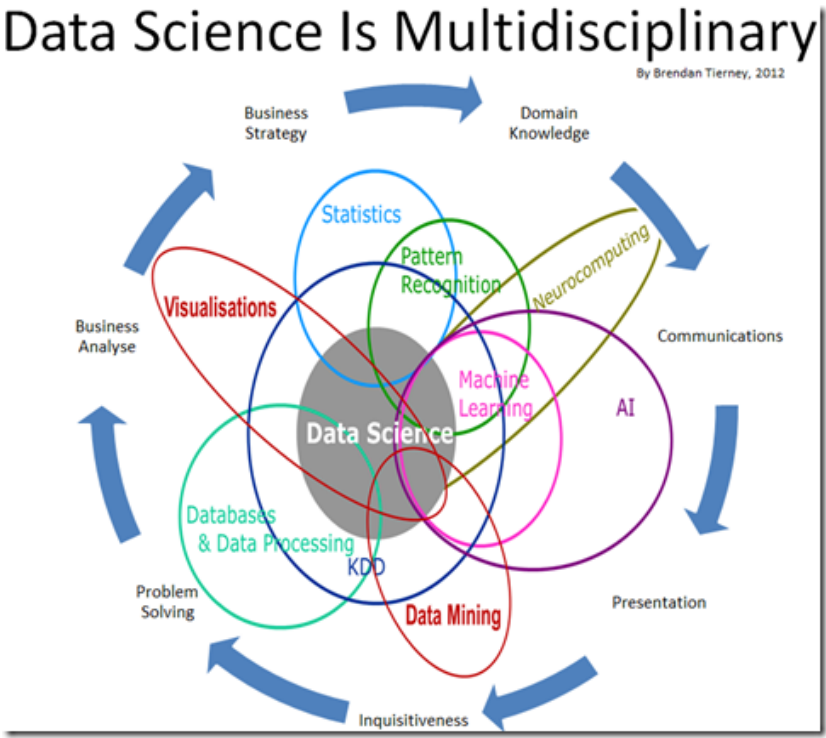


Danger Zone!: This is where people 'know enough to be dangerous,' and is the most problematic area of the diagram because this overlap of skills gives people the ability to create what appears to be a legitimate analysis without any understanding of how they got there or what they have created.

Traditional Research: Substantive expertise plus math and statistics knowledge is where the most traditional researcher falls. Doctoral level researchers spend most of their time acquiring expertise in these areas, but very little time learning about technology.



Substantive Expertise: Science is about discovery and building knowledge, which requires some motivating questions about the world and hypotheses that can be brought to data and tested with statistical methods. Questions first, then data.



http://www.kdnuggets.com/2016/10/battle-data-science-venn-diagrams.html
https://i.onthe.io/smngoz17ju2uc4v1bg.9c27179b.jpg

4 kinds of "data scientists" 8 key skills

Data Analyst

Provide analysis by pulling data out of MySQL databases, becoming a master at Excel pivot tables, and producing basic data visualizations (e.g., line and bar charts).

Data Wrangler

High traffic and an increasingly large amount of data. This requires set up of the data infrastructure that the company will need moving forward. Software engineering that can provide provide basic insights and data-like contributions to the production code.

We Are Data. Data Is Us

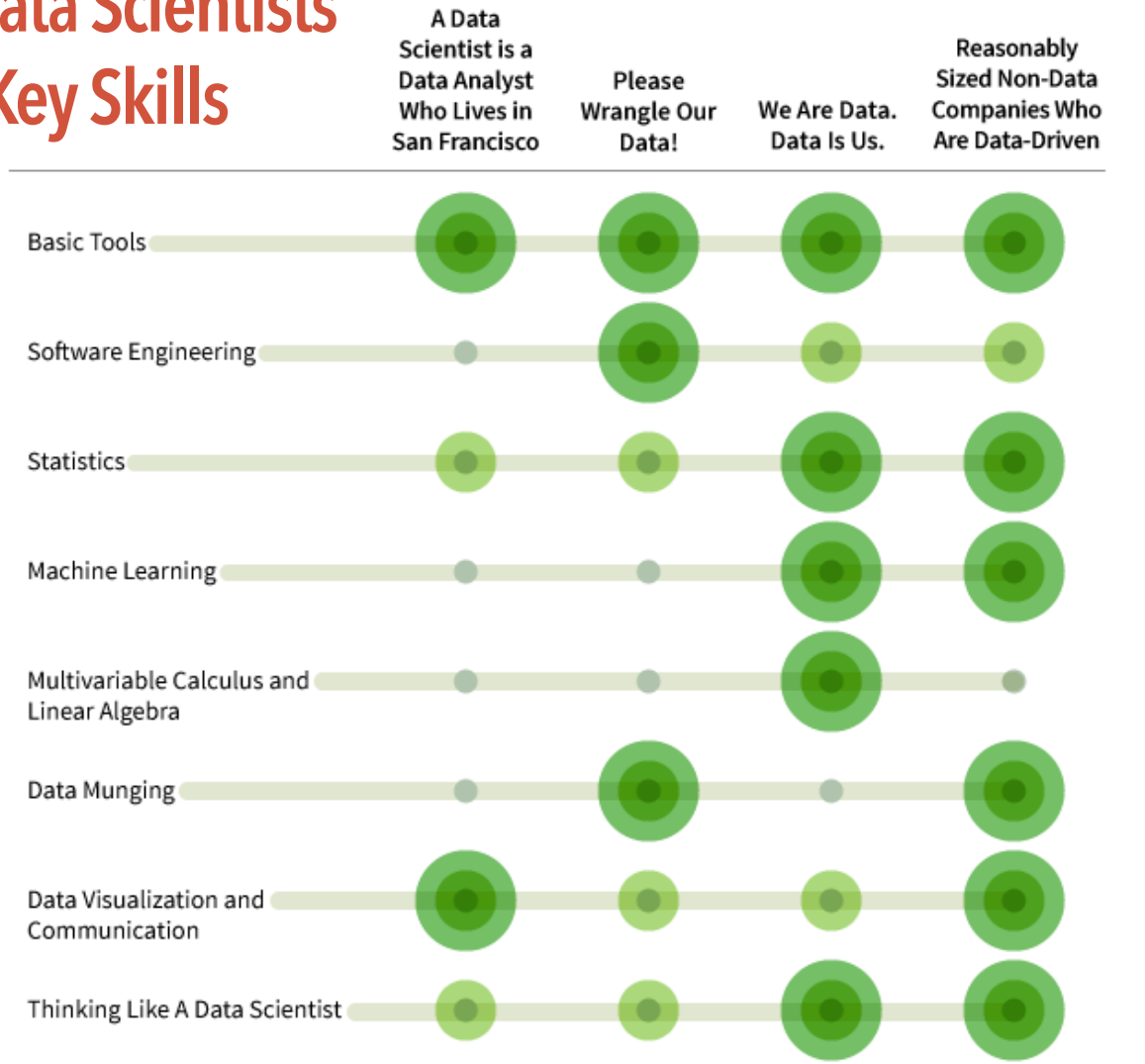
Formal mathematics, statistics, or physics background with focus on producing great data-driven products with massive amounts or a data-based service.

Reasonably Sized Non-Data Companies Who Are Data-Driven:

You would be joining an established team of other data scientists, perform analysis, touch production code, visualize data, etc. familiarity with tools designed for 'big data' (e.g., Hive or Pig) and experience with messy, 'real-life' datasets.

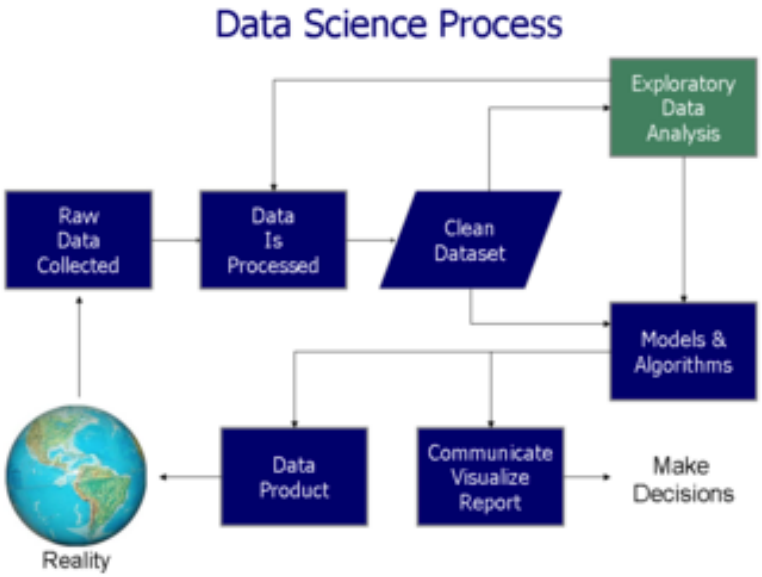
http://blog.udacity.com/2014/11/data-science-job-skills.html

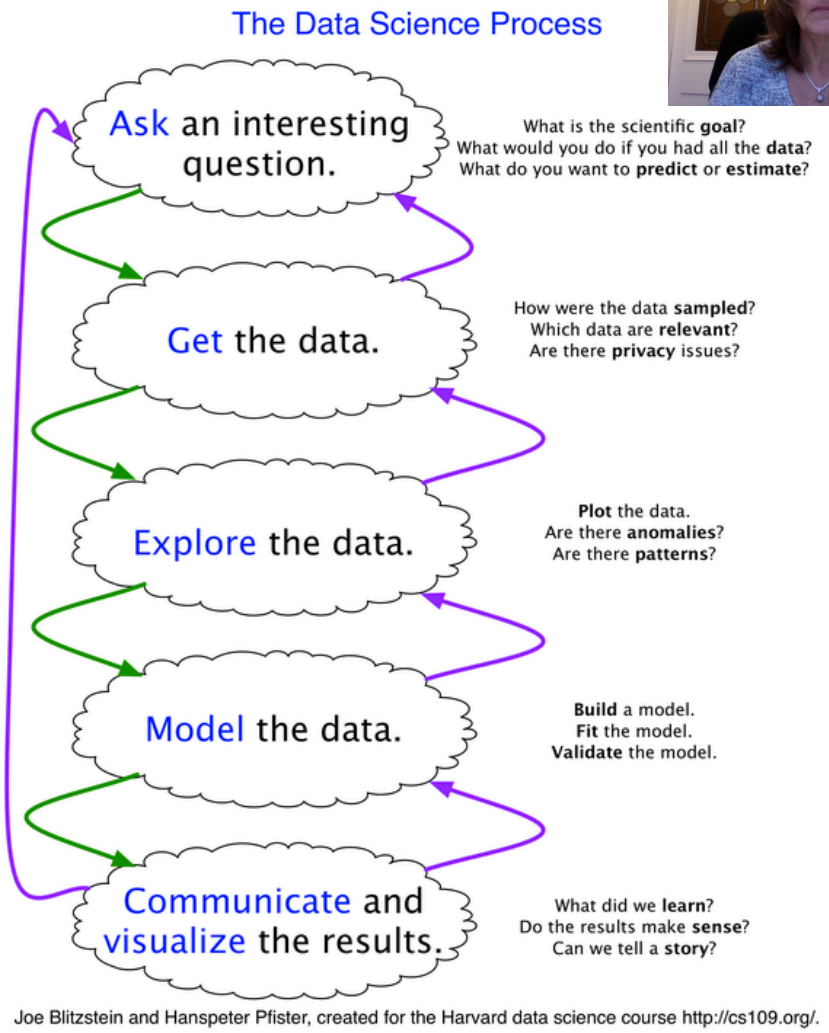
"Data Scientists"
8 Key Skills



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

Data Science Workflow



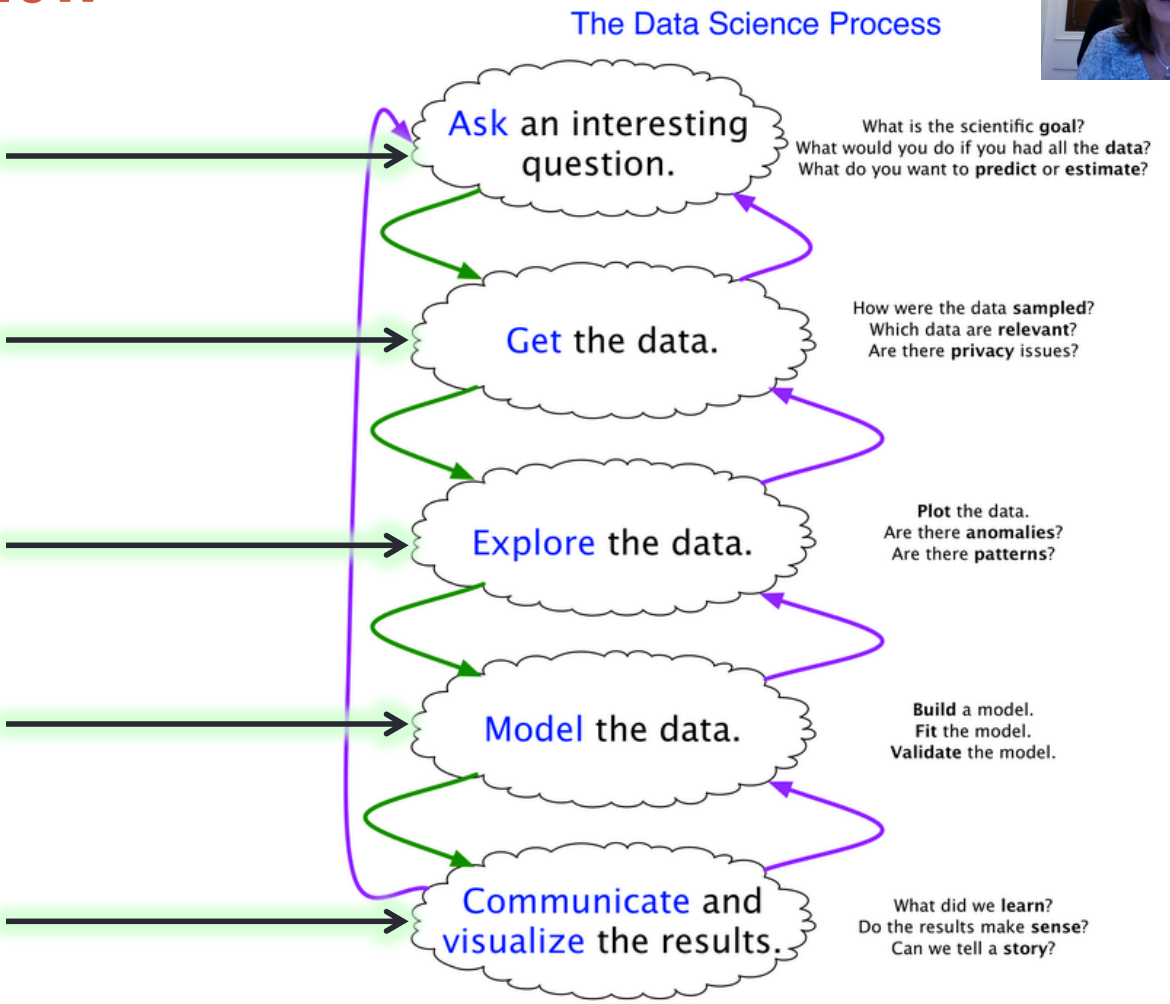


https://en.wikipedia.org/wiki/Exploratory_data_analysis

Data Science Workflow

"Workflow" is a myth.

Enter at any point, then Lather, rinse, repeat*

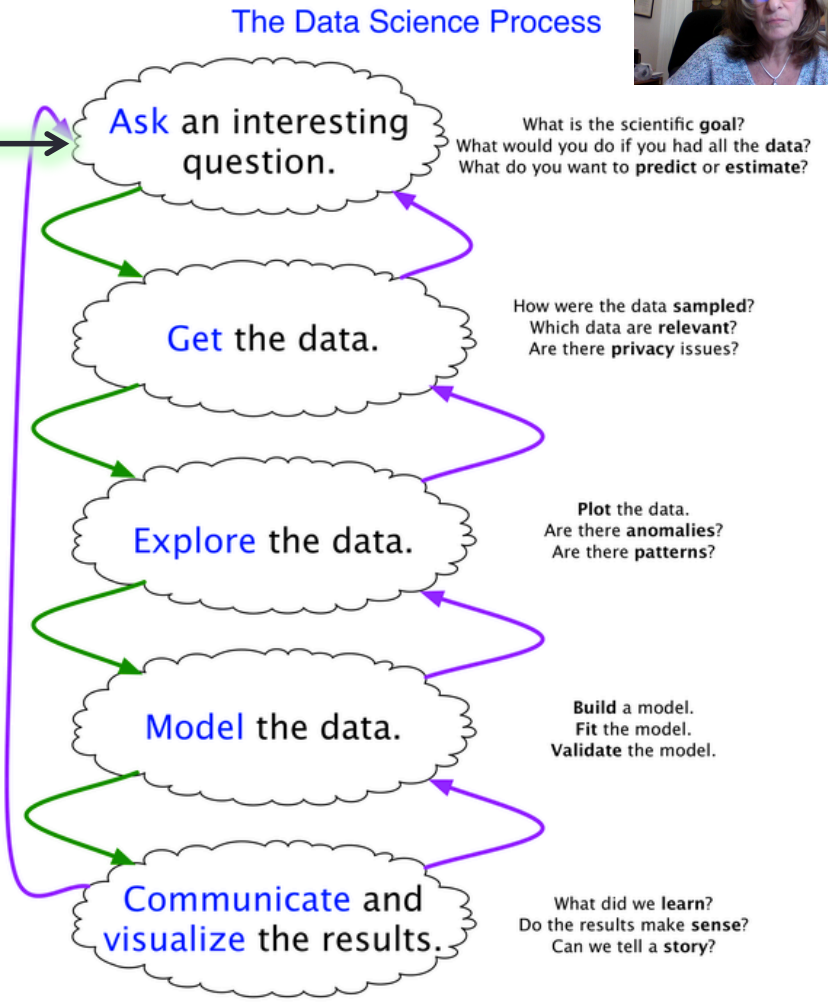


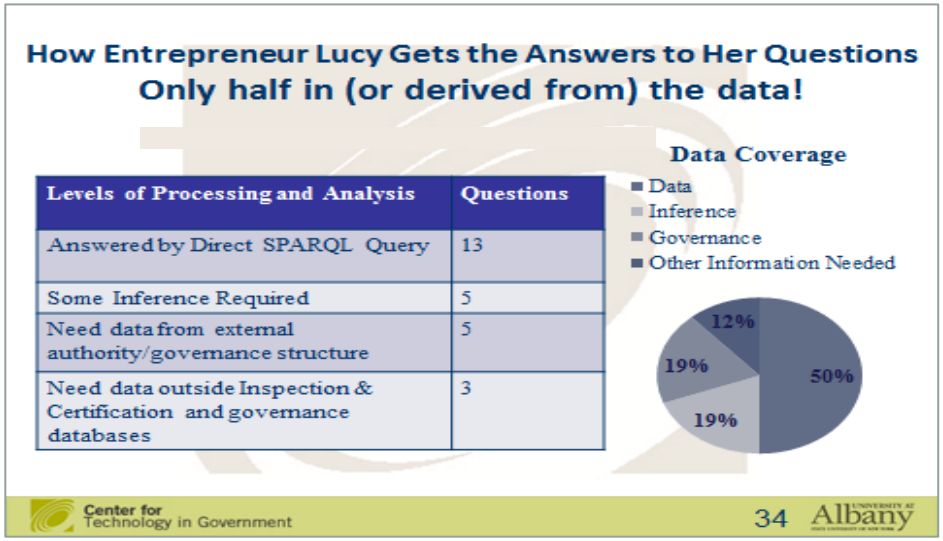
*https://en.wikipedia.org/wiki/Lather,_rinse,_repeat

Data Science Work

What questions can Data Science Answer?

Each question is answered by a separate family of machine learning algorithms.





https://docs.microsoft.com/en-us/azure/machine-learning/machine-learning-data-science-for-beginners-the-5-questions-data-science-answers

Data Science – get the data

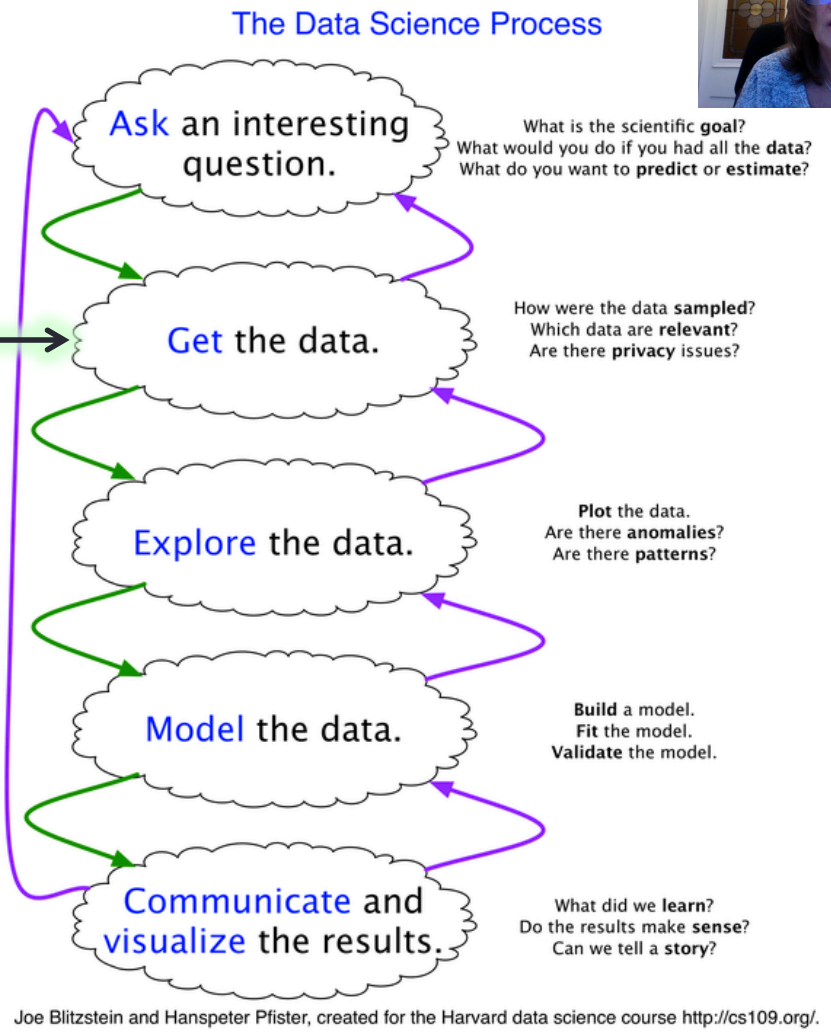
Data

facts and statistics collected together for reference or analysis.

Ways to get data
(some easy some hard)

1. Someone gives you a dataset (or a link to the dataset)
2. Data comes from an instrument or sensors

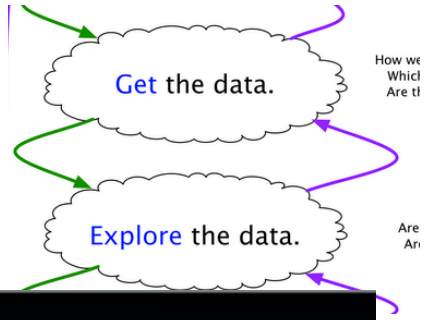
3. Data are live or static
4. May need to write for permission/access and use (purposes) may be open, restricted, in-between)

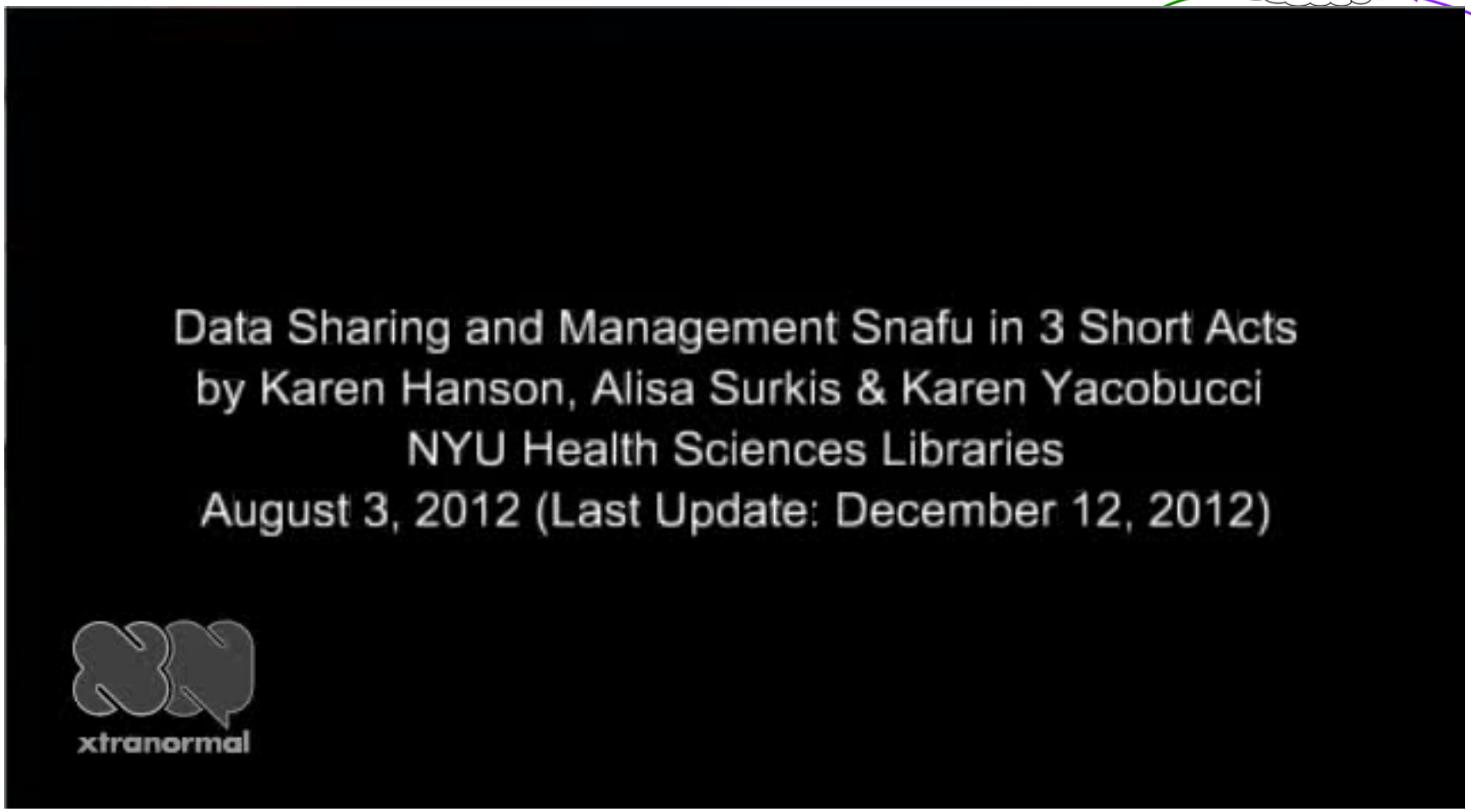


Typically as a data scientist, you would not need to collect data, someone else would have done that. The level of your contribution could be limited by your level of understanding of how the data were collected. Garbage in- Garbage out.

Data Science Work

Get the Data-Explore the Data
Data Sharing and Management Snafu –THIS HAPPENS! ALL TOO FREQUENTLY!





https://www.youtube.com/watch?v=66oNv_DJuPc

Data Principles

FAIR Data Principles

One of the grand challenges of data-intensive science is to facilitate knowledge discovery by assisting humans and machines in their discovery of, access to, integration and analysis of, task-appropriate scientific data and their asso algorithms and workflows. Here, we describe FAIR - a set of guiding principles to make data Findable, Accessible, Interoperable, and Re-usable.

To be Findable:

- F1. (meta)data are assigned a globally unique and eternally persistent identifier.
- F2. data are described with rich metadata.
- F3. (meta)data are registered or indexed in a searchable resource.
- F4. metadata specify the data identifier.

To be Accessible:

- A1 (meta)data are retrievable by their identifier using a standardized communications protocol.
- A1.1 the protocol is open, free, and universally implementable.
- A1.2 the protocol allows for an authentication and authorization procedure, where necessary.
- A2 metadata are accessible, even when the data are no longer available.

To be Interoperable:

- I1. (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.
- I2. (meta)data use vocabularies that follow FAIR principles.
- I3. (meta)data include qualified references to other (meta)data.

To be Re-usable:

- R1. meta(data) have a plurality of accurate and relevant attributes.
- R1.1. (meta)data are released with a clear and accessible data usage license.
- R1.2. (meta)data are associated with their provenance.
- R1.3. (meta)data meet domain-relevant community standards.

FORCE11
The Future of Research Communications and e-Scholarship

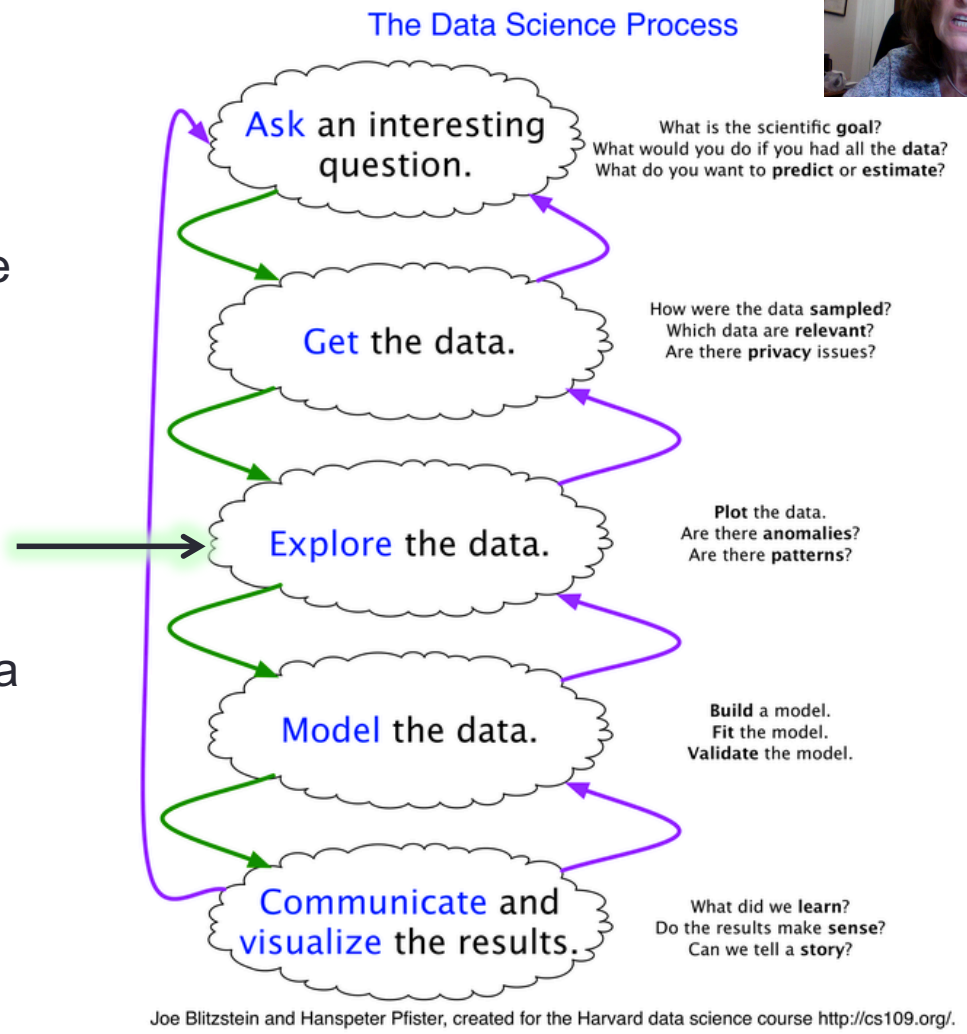
https://www.force11.org/group/fairgroup/fairprinciples

Data Science: Explore

Exploratory Data Analysis

Objectives

- Suggest hypotheses about the causes of observed phenomena
- Assess assumptions on which statistical inference will be based
- Support the selection of appropriate statistical tools and techniques
- Provide a basis for further data collection through surveys or experiments[1]



[1] Behrens-Principles and Procedures of Exploratory Data Analysis-American Psychological Association-1997

Data Science: Explore

Tools

This is the tool I use for leverage.

Graphical:

- Box plot
- Histogram
- Multi-vari chart
- Run chart
- Pareto chart
- Scatter plot
- Stem-and-leaf plot
- Parallel coordinates
- Odds ratio
- Multidimensional scaling
- Targeted projection pursuit
- Principal component analysis
- Multilinear PCA
- Projection methods such as grand tour, guided tour and manual tour
- Interactive versions of these plots

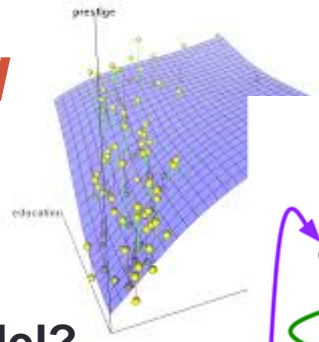
Quantitative:

- Median polish
- Trimean
- Ordination

https://en.wikipedia.org/wiki/Exploratory_data_analysis

Data Science Workflow

Model the data

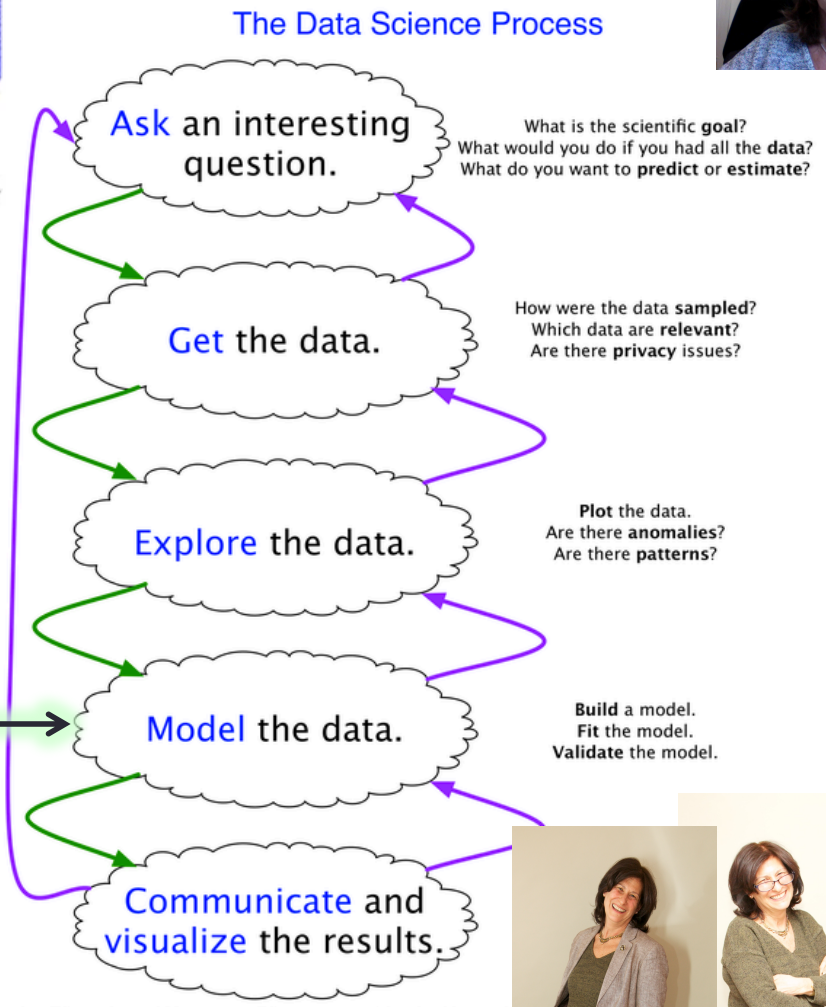


Data Model or Statistical Model?

Here we mean statistical model, which Is a type of mathematical model.

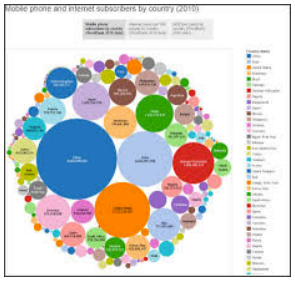
A **Data Model** organizes elements of data and how they relate to one another, e.g. a data element representing a car comprises a number of other elements which in turn represent the color, size and owner of the car.

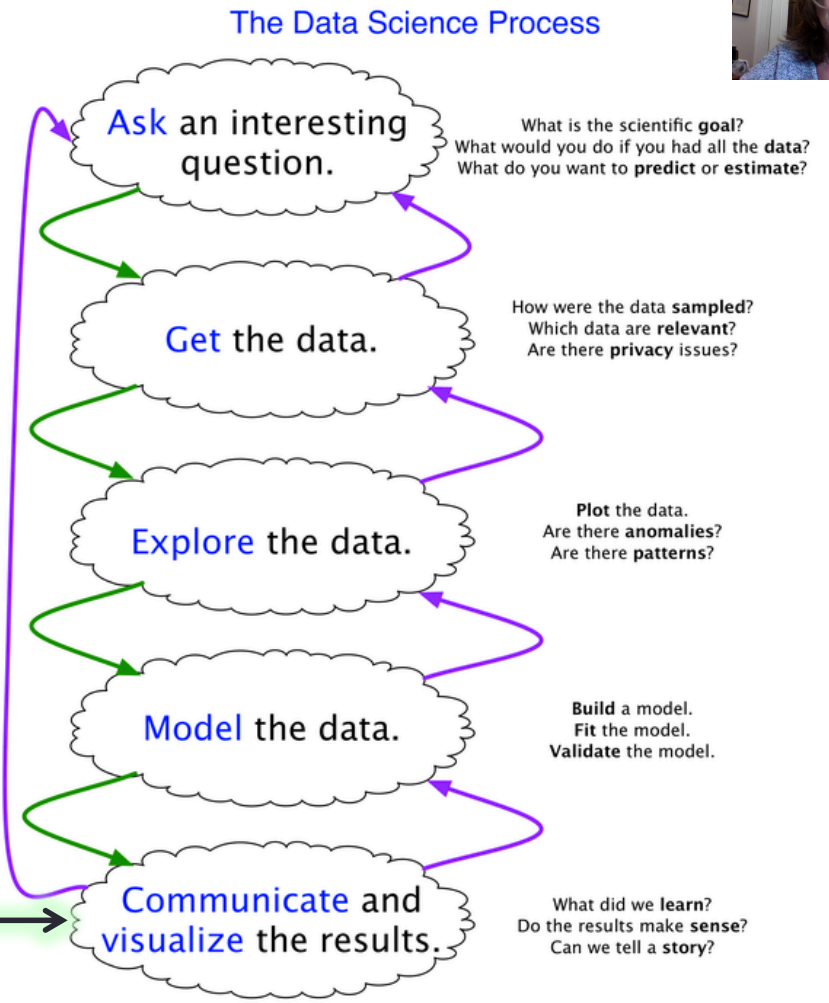
A statistical model embodies a set of assumptions concerning the generation of the observed data, and similar data from a larger population. A model represents, often in considerably idealized form, the data-generating process.



Data Science Workflow

Communicate and Visualize the results.



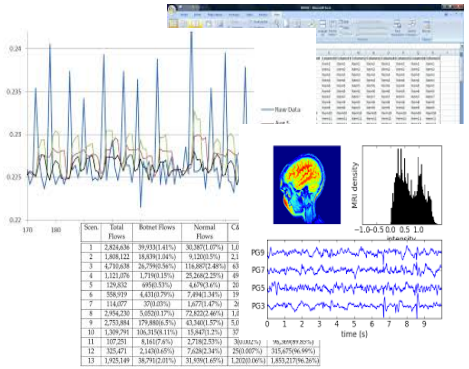


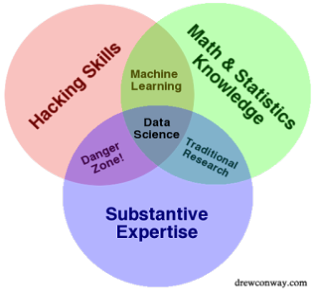
http://www.encorebusiness.com/blog/tableau-tips-tricks-tableau-story-telling/

Module 1 – Data Science Concepts Summary

Topics: Data
Data Science Workflow
Data Science Tools

3 tools we will use in this course, R Python, Tableau





- Data science is concerned with the collection, preparation, analysis, visualization, and preservation of data of all kinds and sizes.

- Data science requires a multidisciplinary skill set to address issues in all these areas.

- Data Science can be applied to many disciplines and it is important to understand the special needs of the domain and the limits of the data scientist.