



# Data Science in Drug Discovery, Health & Translational Medicine

---

Joanne S. Luciano <https://www.linkedin.com/in/joanneluciano>  
Visiting Associate Professor  
Indiana University School of Informatics and Computing



DATA SCIENCE FOR DRUG DISCOVERY, HEALTH AND TRANSLATIONAL MEDICINE

# Part 1- Data Science and the Healthcare Ecosystem

## Background

- **Module 1 Data Science Concepts**
- **Module 2 Healthcare Part 1**
  - The healthcare ecosystem
  - Healthcare data
- **Module 3 Healthcare Part 2**
- **Module 4 Health Web Science**



# Module 1: Data Science Concepts

**Topics:**

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



**Learning Objectives:** Students will be able to:

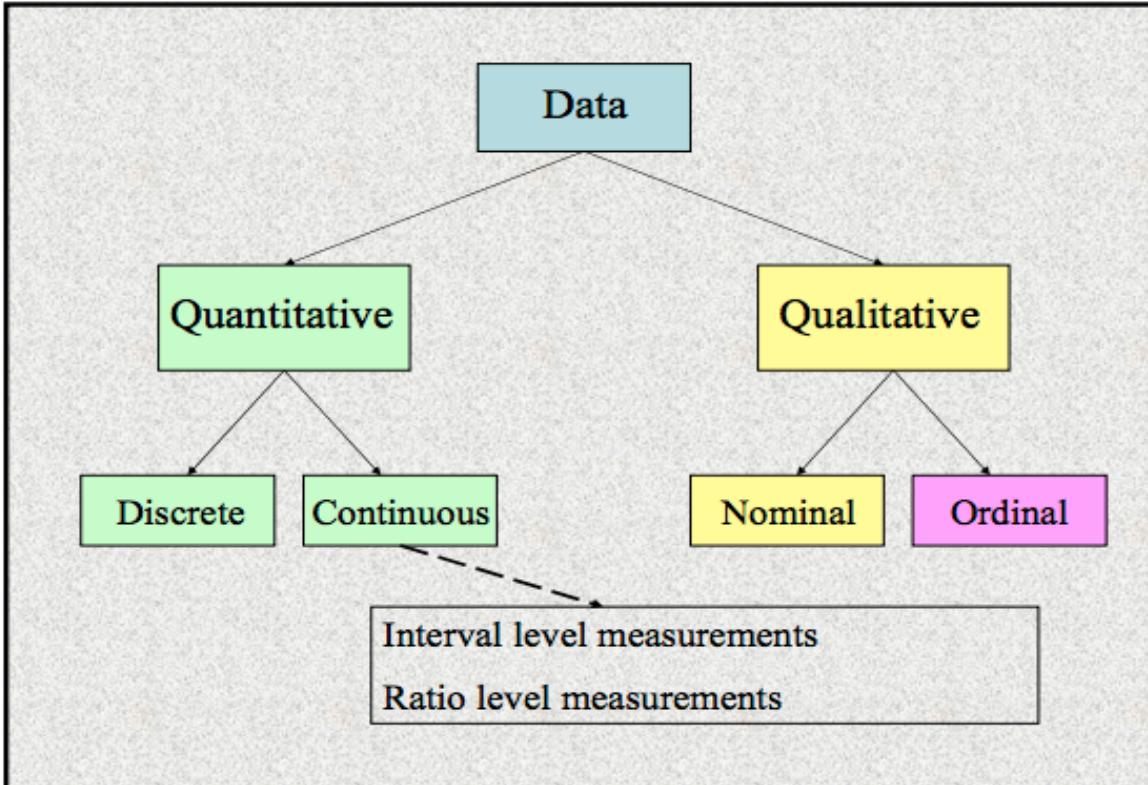
- Structure and execute a data science analysis and report the results.
- Perform exploratory data analysis, formal modeling, interpretation, and communication.
- Clean, integrate, analyze, and visually report analysis results using Tableau.

# Data - Types of data



"Data!data!data!" he cried impatiently. "I can't make bricks without clay."

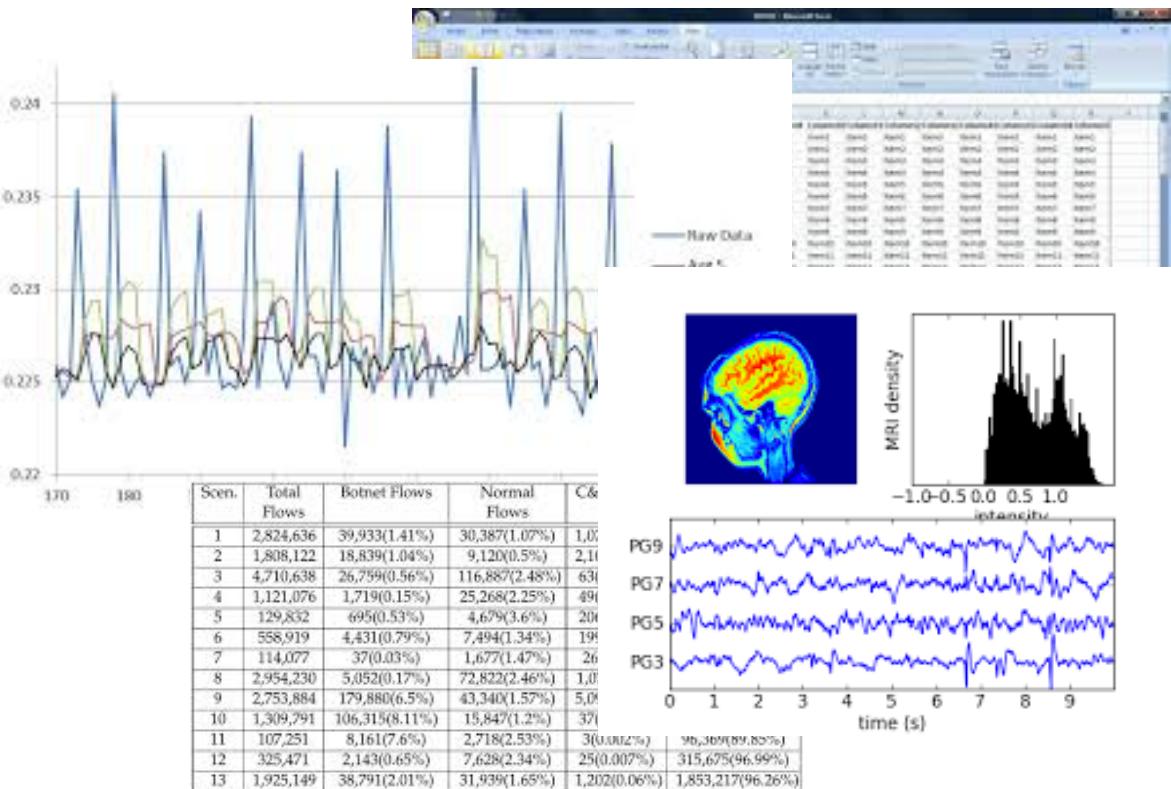
— Arthur Conan Doyle, The Adventure of the Copper Beeches



## Types of Data

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

# Data – Sources of Data



## Sources of Data

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

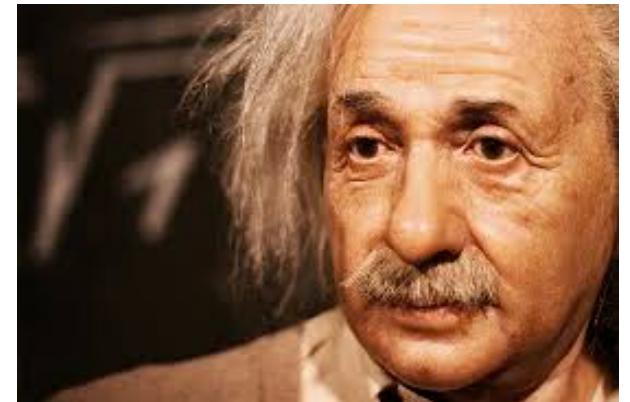
**Caveat: Context determines whether a source is primary, secondary, for tertiary.**  
[http://gethelp.library.upenn.edu/PORT/sources/primary\\_secondary\\_tertiary.html](http://gethelp.library.upenn.edu/PORT/sources/primary_secondary_tertiary.html)

# 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?

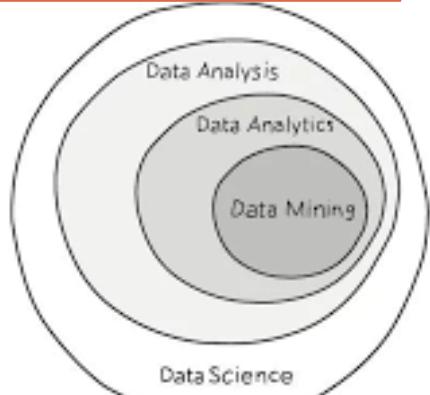
What are the standards organizations?



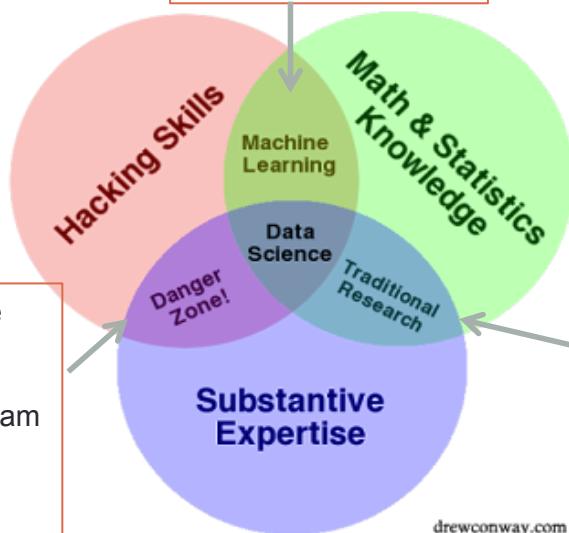
# Data Science

**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.

**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.



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



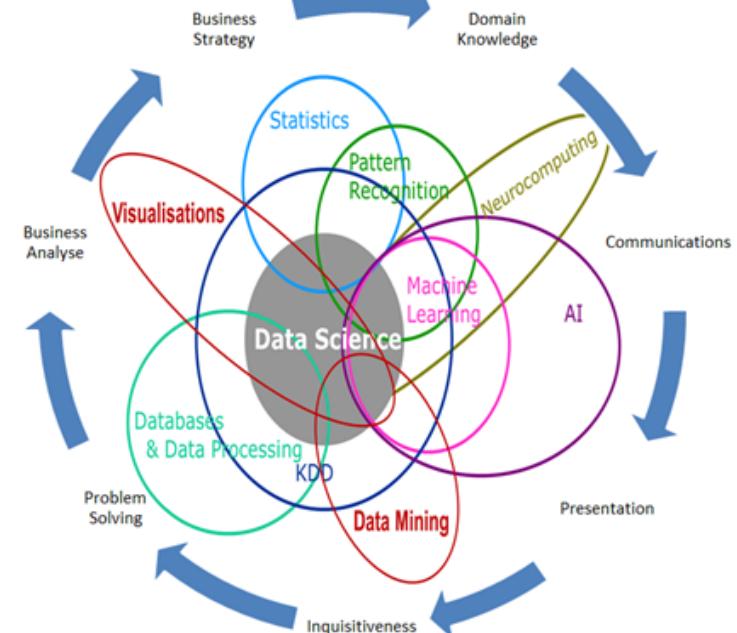
**Math & Statistics Knowledge:** Once you have acquired and cleaned the data, the next step is to actually extract insight from it. You need to apply appropriate math and statistics methods, which requires at least a baseline familiarity with these tools.



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

## Data Science Is Multidisciplinary

By Brendan Tierney, 2012



# 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 and need someone to set up the data infrastructure that the company will need moving forward. Software engineering that can provide basic insights and data-like contributions to the production code.

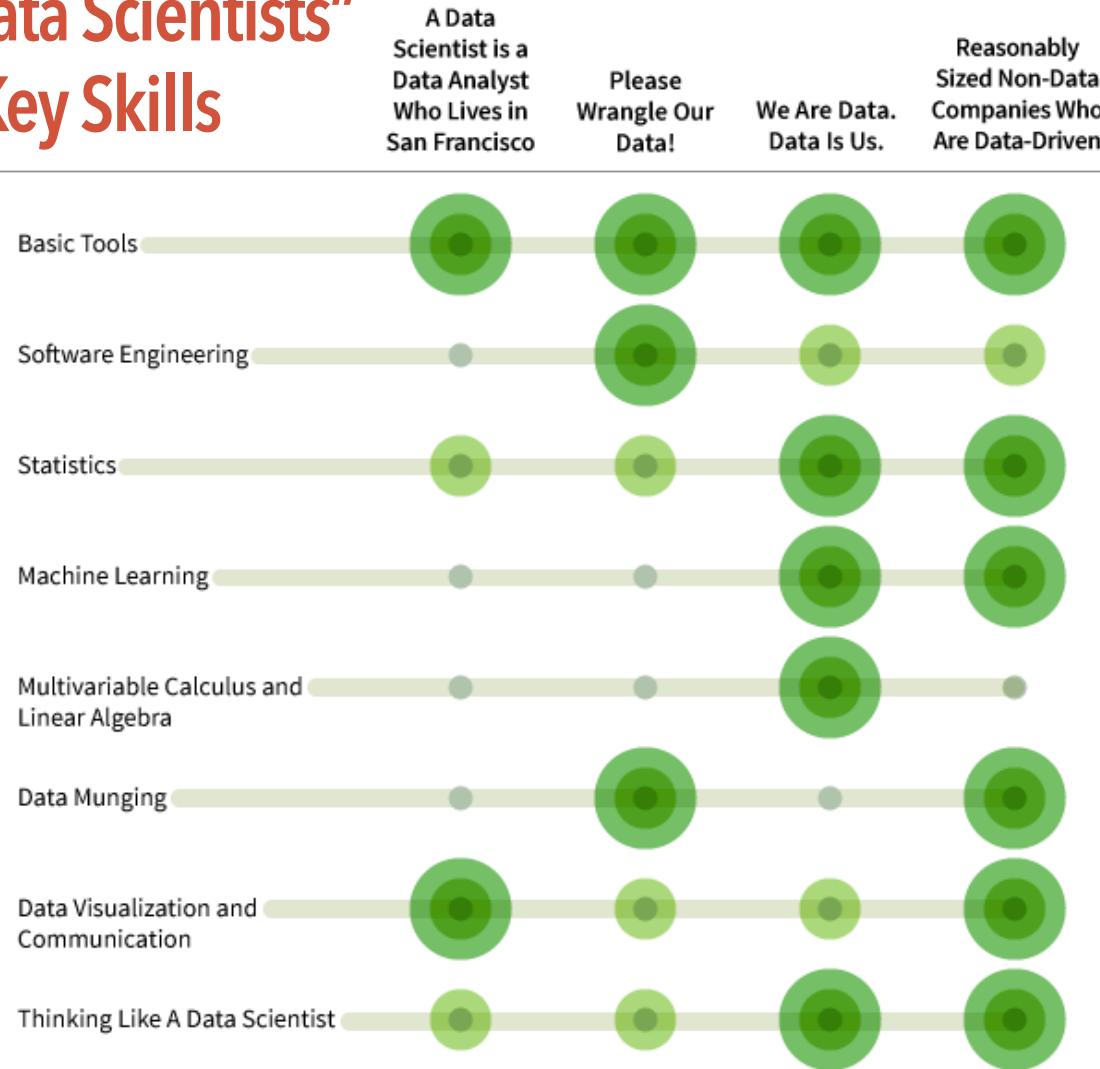
## We Are Data. Data Is Us

- 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're 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.

# "Data Scientists" 8 Key Skills



Very important



Somewhat important



Not that important

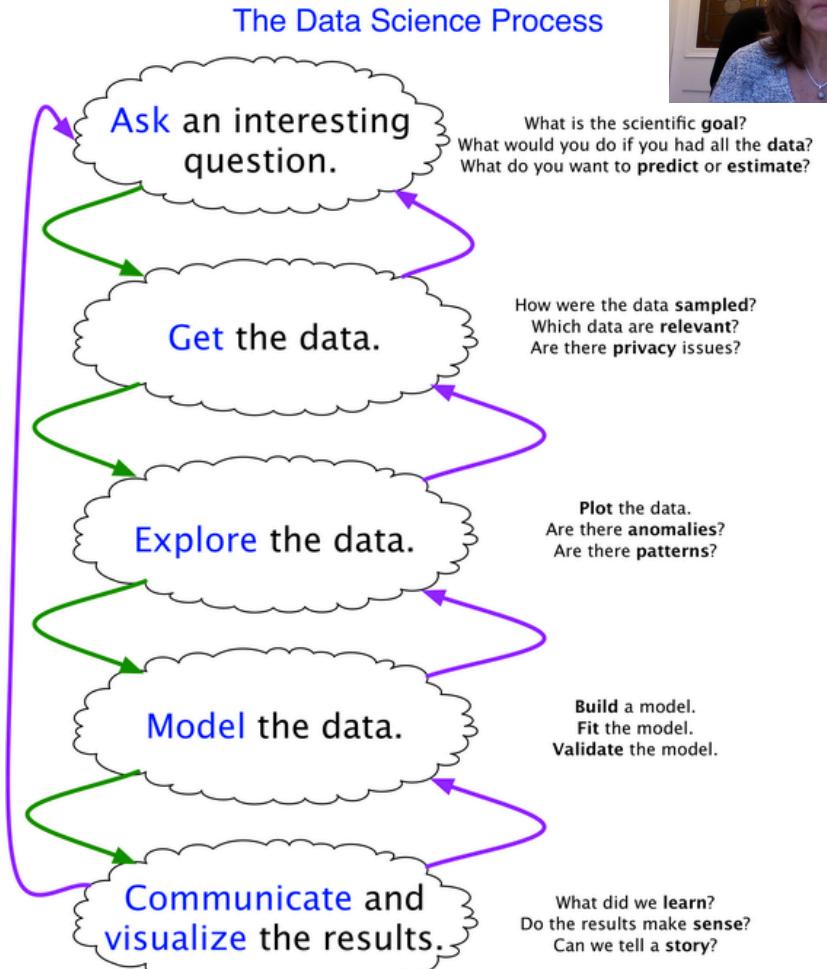
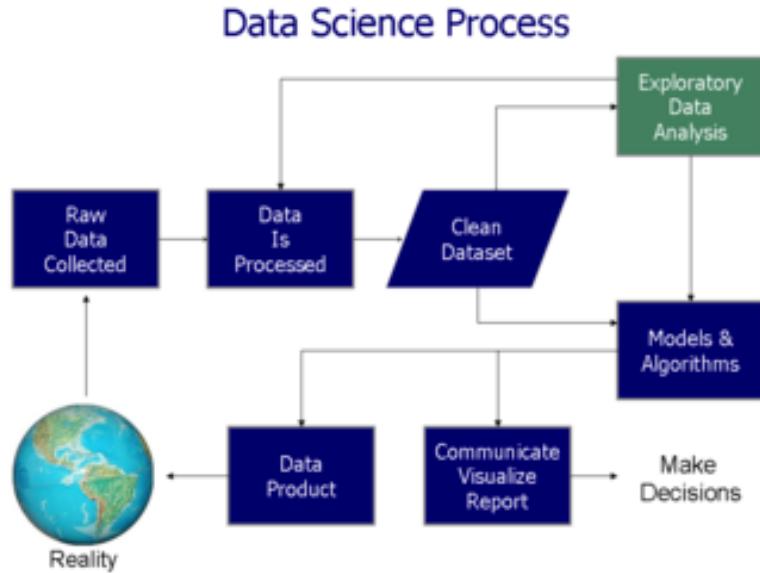


----- Candidate break point -----

----- Candidate break point

if recording is getting too long, break here -----

# Data Science Workflow



Joe Blitzstein and Hanspeter Pfister, created for the Harvard data science course <http://cs109.org/>.

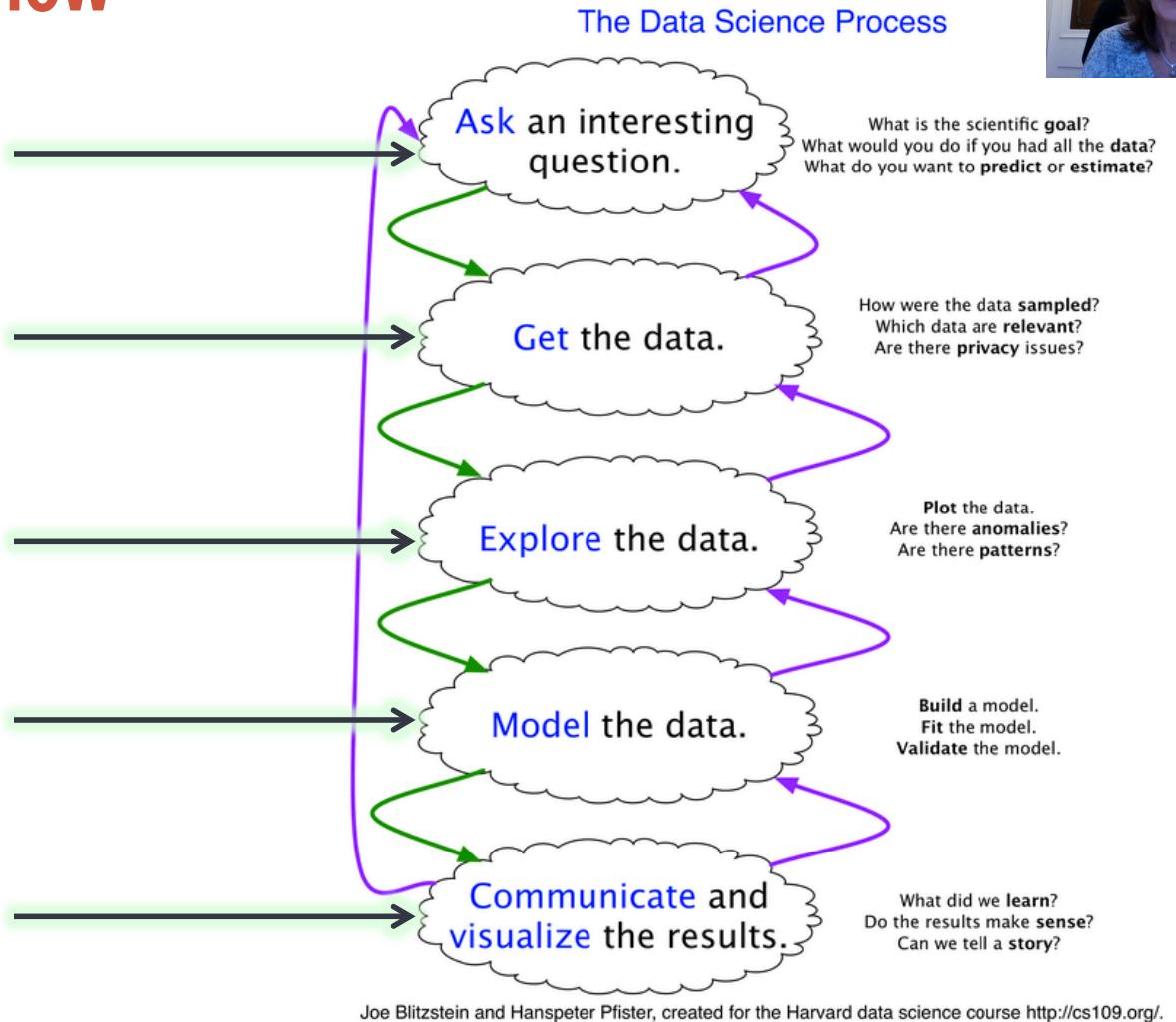
[https://en.wikipedia.org/wiki/Exploratory\\_data\\_analysis](https://en.wikipedia.org/wiki/Exploratory_data_analysis)

# Data Science Workflow



**"Workflow"**  
is a myth.

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



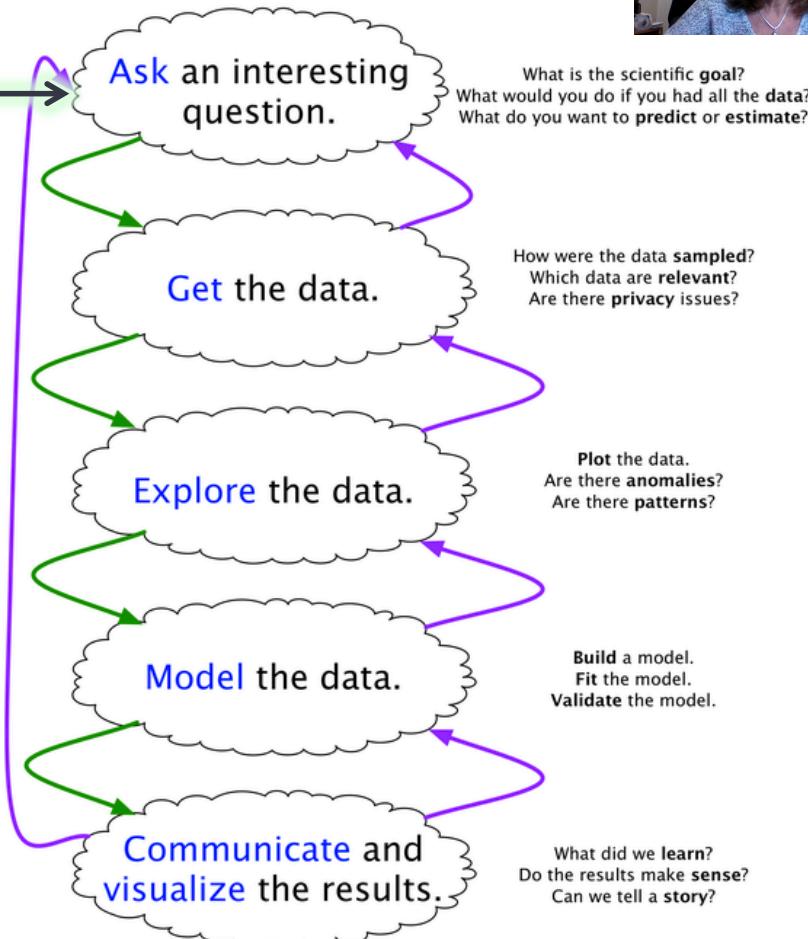
\*[https://en.wikipedia.org/wiki/Lather,\\_rinse,\\_repeat](https://en.wikipedia.org/wiki/Lather,_rinse,_repeat)

# Data Science Work

## What questions can Data Science Answer?

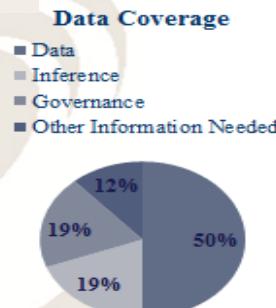
Each question is answered by a separate family of machine learning algorithms.

### The Data Science Process



### How Entrepreneur Lucy Gets the Answers to Her Questions Only half in (or derived from) the data!

Levels of Processing and Analysis	Questions
Answered by Direct SPARQL Query	13
Some Inference Required	5
Need data from external authority/governance structure	5
Need data outside Inspection & Certification and governance databases	3



# Data Science – get the data

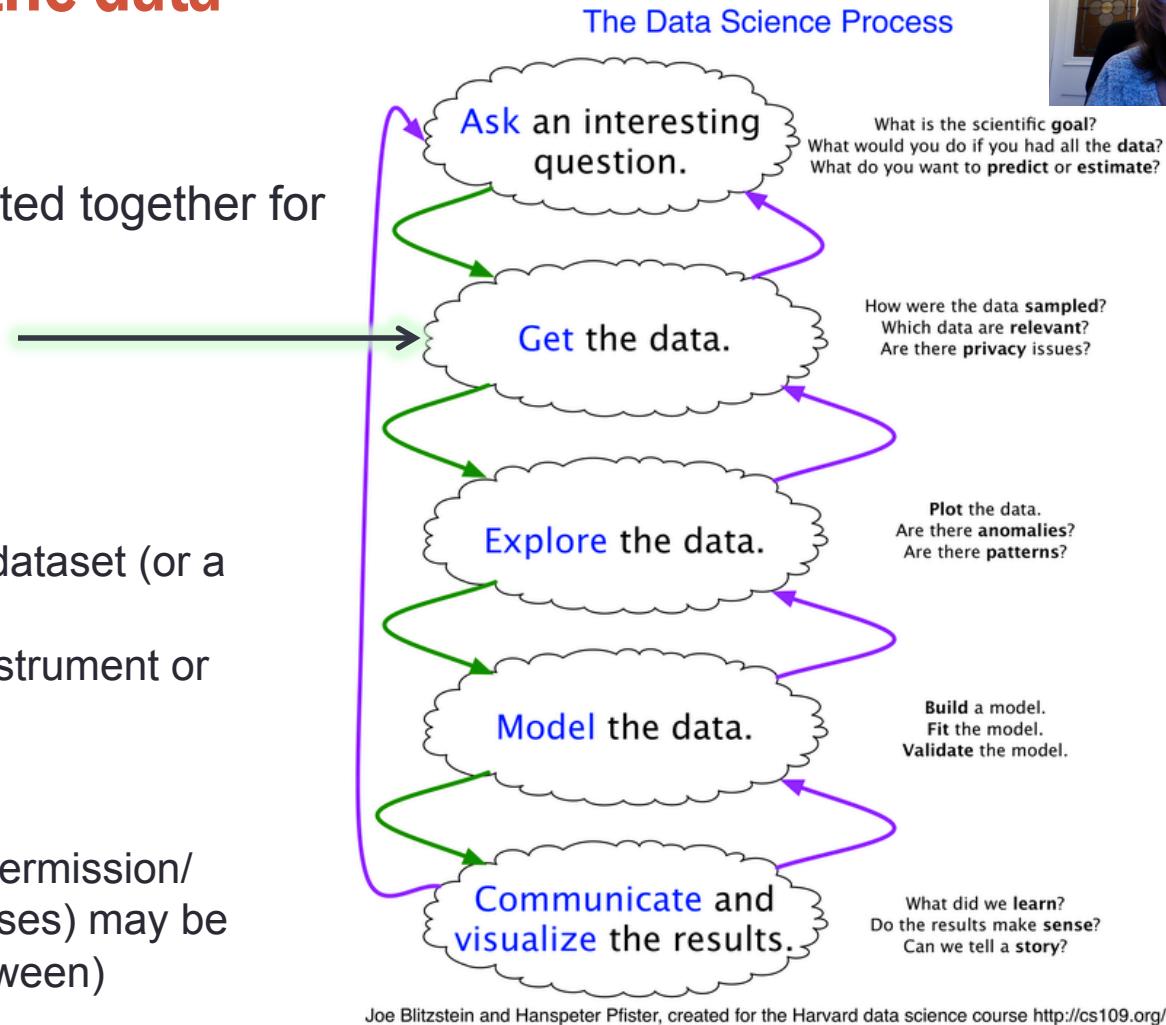


## Data

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)

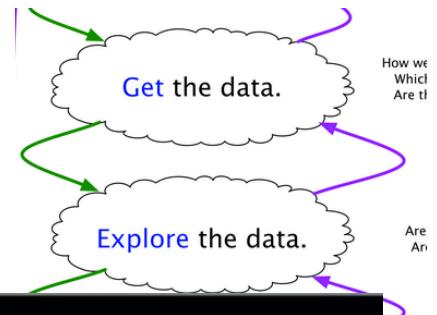


Typically as a data scientist, you would not need to collect data, someone else would have done that. The level of your contribution could be limited by your level of understanding of how the data were collected. Garbage in- Garbage out.

# Data Science Work

Get the Data-Explore the Data

Data Sharing and Management Snafu – THIS HAPPENS! ALL TOO FREQUENTLY!



Data Sharing and Management Snafu in 3 Short Acts

by Karen Hanson, Alisa Surkis & Karen Yacobucci

NYU Health Sciences Libraries

August 3, 2012 (Last Update: December 12, 2012)



# 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 associated 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.

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

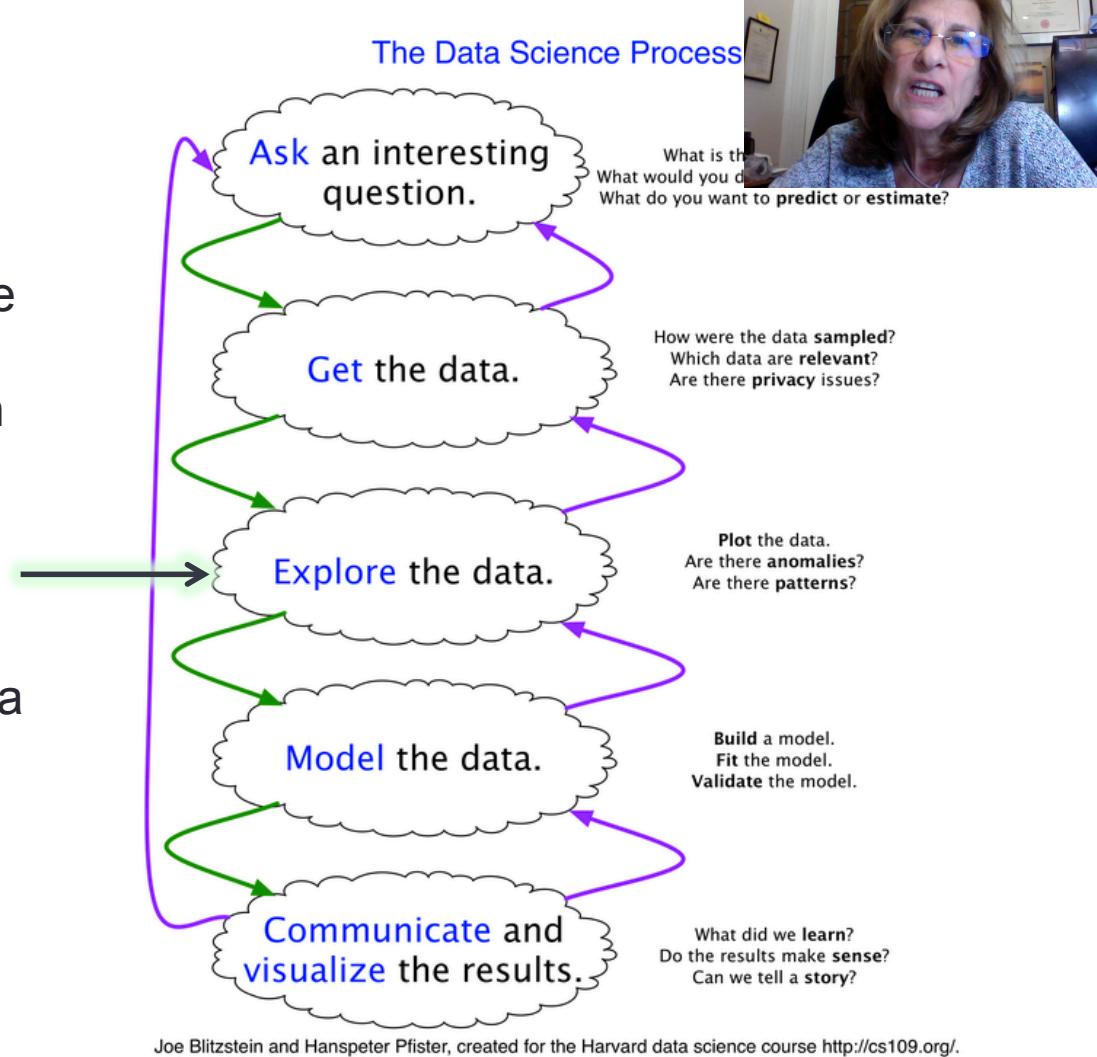


# Data Science: Explore

## 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

Graphical:



This is the tool I use for leverage.

- 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](https://en.wikipedia.org/wiki/Exploratory_data_analysis)



# Data Science Workflow

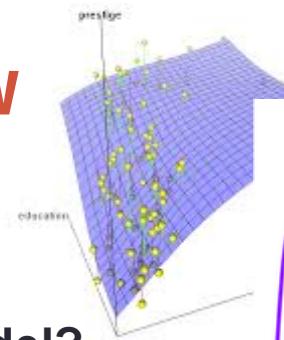
## Model the data

### Data Model or Statistical Model?

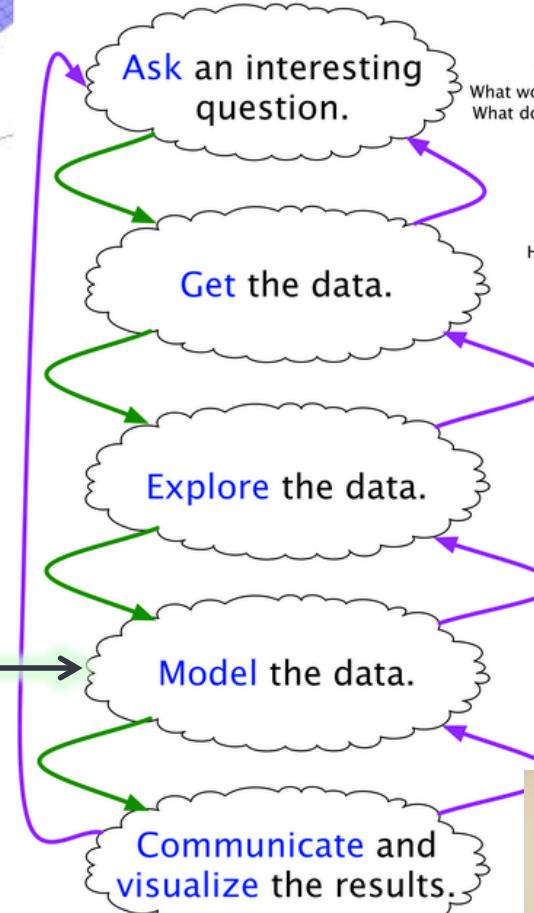
Here we mean statistical model, which Is a type of mathematical model.

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.



### The Data Science Process

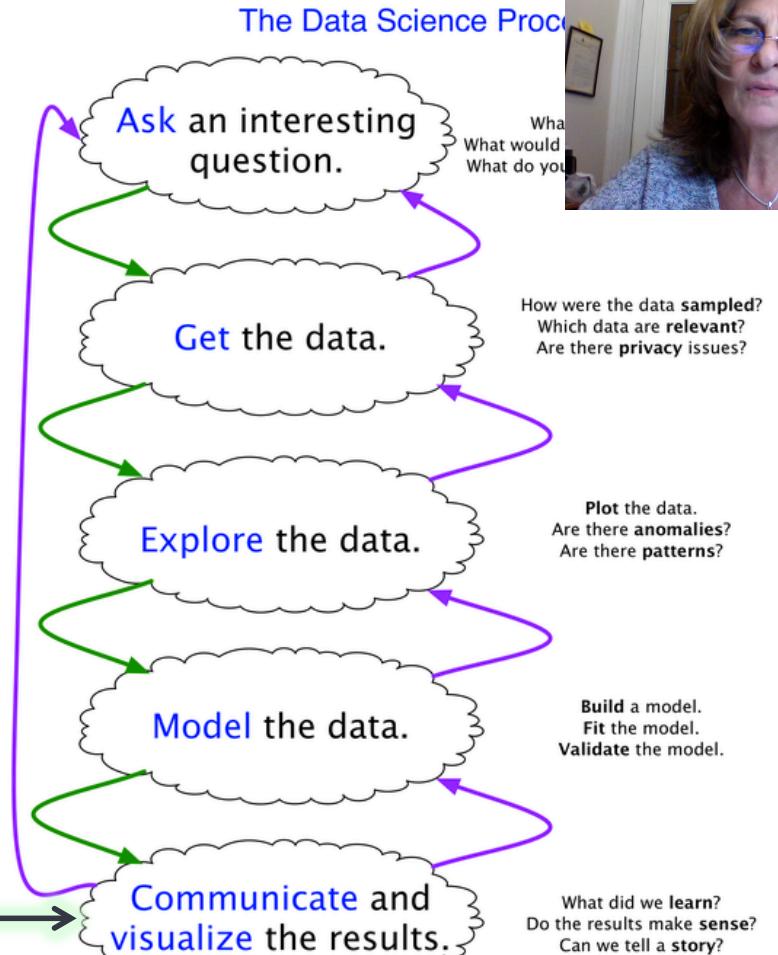
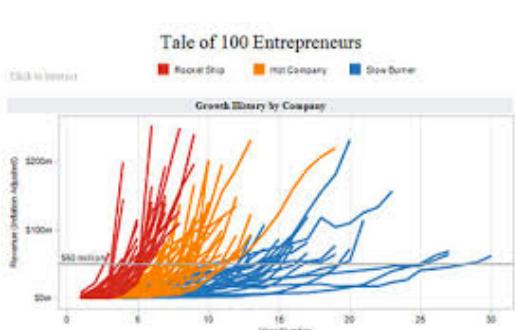
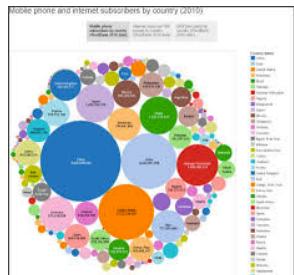


Joe Blitzstein and Hanspeter Pfister, created for the Harvard



# Data Science Workflow

## Communicate and Visualize the results.



Joe Blitzstein and Hanspeter Pfister, created for the Harvard data science course <http://cs109.org/>.

<http://www.encodedbusiness.com/blog/tableau-tips-tricks-tableau-story-telling/>

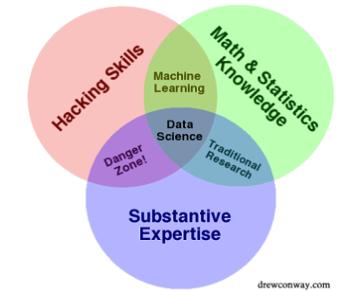
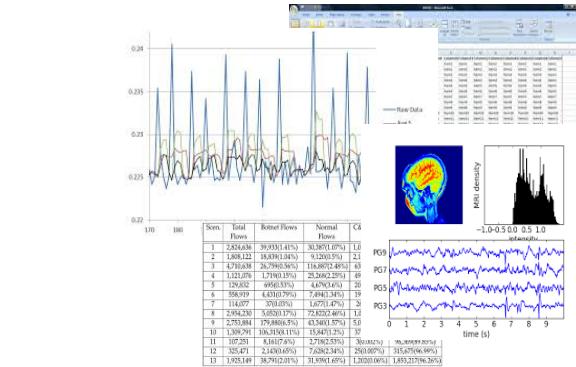
# Module 1 - Data Science Concepts Summary



**Topics:** Data  
Data Science Workflow  
Data Science Tools

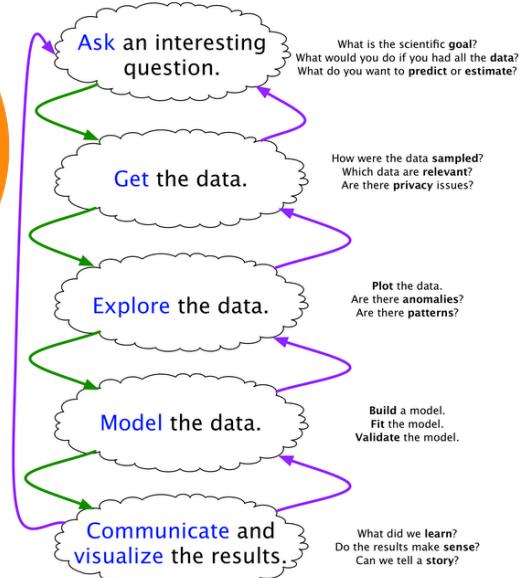
3 tools we will use in this course, R  
Python, Tableau

- 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 these 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.



drewconway.com

The Data Science Process



Joe Blitzstein and Hanspter Pfister, created for the Harvard data science course <http://cs109.org/>.

