

- Target Variable . The  $y$ , outcome, or dependent variable
- The "Unknown"
- What we are trying to predict

## Data Visualization

- 1) Explores data + understanding
- 2) Communicates to others

## Data Governance

- Data policy (security policy)
- Storing, managing & processing data in a distributed environment

Questions: **DS CAN ANSWER**

Regression - How many or how much?

Classification - Is this an observation A, or B, or C etc?

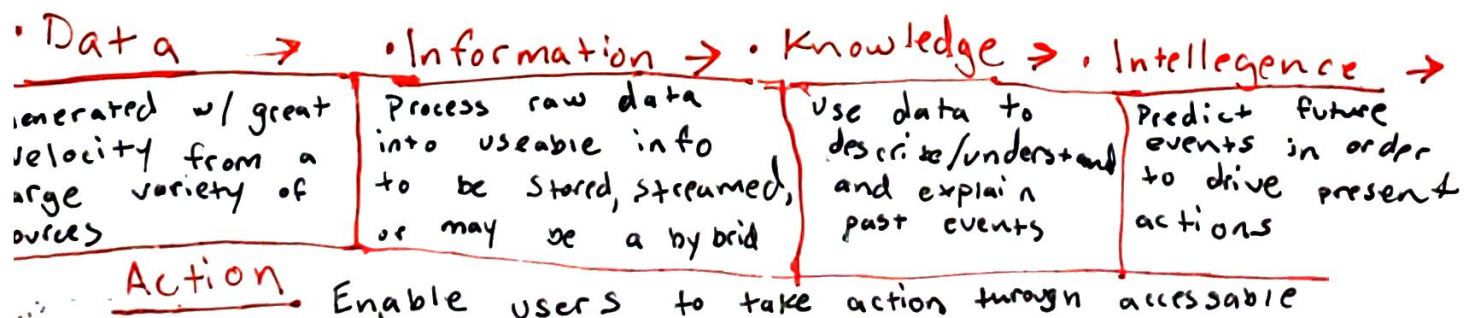
Clustering - What groupings / relationships exist in the data already?

Time Series Analysis - What's our next likely outcome?

Anomaly Detection : Is this weird?

## Data Science Pipeline

- End to end
- Left to right



## Data Planning

### Goal:

- Clearly define your goal (write it down)
- Measures of success, and plans on how to achieve that

### Deliverable:

- Documentation of your goal
- If you haven't defined success, you won't know when you have achieved it

### How you get there:

By answering questions about the final product & formulating or identifying any initial hypotheses.

## Acquisition

AKA: Data gathering  
Data Import  
Data Wrangling  
(Acquisition + Prep)

Goal: create a path from original data sources to the environment in which you will work w/ the data

Deliverable: A file, acquire.py, that contains the function(s) needed to reproduce the acquisition of data

### How to get there

: SQL: clean-up, integration, aggregation or other manipulation of data in the SQL Environment

Pylib: pandas

May use Spark and/or Hive when acquiring data from a distributed environment such as HDFS.

Examples of source types:

- RDBMS
- HDFS
- Static local flat files (CSV, txt, xls)