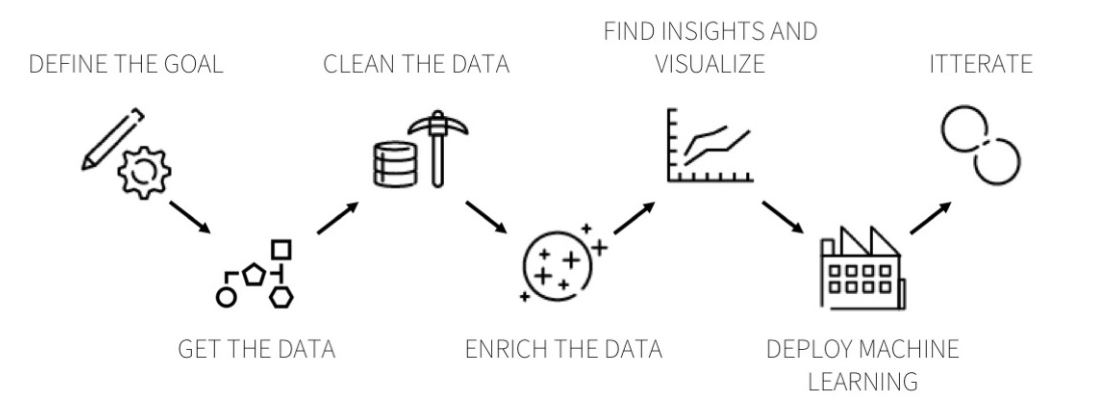
# Exploring Data with Statistics

# The Data Science Process



ds\_process\_2.png

# Agenda

* Difference between groups
* Useful Metrics
* Computational Methods
* Setting up for Hypothesis Testing

# Effect Size

One of the most common use-cases in statistics is comparing groups on certain metrics.

Examples include

* Is Version A of the homepage of our website getting us a better CTR than Version B (commonly called A/B Testing)
* Do girls score higher than boys on the SATs
* Is Drug A more effective in reducing the effects of a disease than Drug B
* Does California receive more rainfall in July than Nevada

Typically, this would mean we have 2 variables in our data - one numeric and one categorical, and we’re interested in reporting whether the groups (based on the levels of the categorical data) are significantly different from each other. The difference is called an effect and then proceed to investigate whether this effect could’ve been caused by random chance.

To explore statistics that quantify **Effect Size**, we’ll look at **the difference in height between men and women**. The question we’re trying to answer is: > *Are men, on average, taller than women?*

We start by looking at the data from recorded heights of a random sample of 1000 men and women.

male\_sample = array([180.1, 170.4, 169.1, ... 176.2, 174.7, 176.8])  
female\_sample = array([ 164.5, 160.9, 166.0, ... 162.4, 158.8, 149.3])  
  
# Visualize the two histograms  
DataFrame({'male\_sample': male\_sample, 'female\_sample': female\_sample}).plot.hist(bins=30, subplots=True);

Both samples are NumPy arrays. Now we can compute sample statistics like the mean and standard deviation.

mean1, std1 = male\_sample.mean(), male\_sample.std()  
# (178.1, 7.84)  
  
mean2, std2 = female\_sample.mean(), female\_sample.std()  
# (163.4, 7.38)

Now, there are many ways to describe the **magnitude of the difference between these distributions.**

### 1. Difference in the Means

difference\_in\_means = mean1 - mean2  
# 14.68

We see from this metric that on average, **men are 14-15 centimeters taller**.

For some applications, that would be a good way to describe the difference, but there are a few problems: - Without knowing more about the distributions (like the stddev) it’s hard to interpret whether a difference like 15 cm is a lot or not. - The magnitude of the difference depends on the units of measure, making it hard to compare across different studies.

### 2. Overlap and Misclassification Rate

Let’s try to pose this problem in a different way.

If we used height to predict gender, how often would we go wrong?

If the heights of men and women really are significantly different, they would have minimal overalp, and we should be able to find a threshold value that classifies them quite accurately. Plotting the PDFs together, we can visualise this overlap:

Next, we find the point where the PDFs cross, and count how many men are below the threshold and how many women are above it:

thresh = (std1 \* mean2 + std2 \* mean1) / (std1 + std2)  
male\_below\_thresh = sum(male\_sample < thresh)  
female\_above\_thresh = sum(female\_sample > thresh)

The “overlap” is the total area under the curves that ends up on the wrong side of the threshold, and the fraction of people who would be misclassified if you tried to use height to guess sex (50% chance) would be half that number.

overlap = male\_below\_thresh/float(len(male\_sample)) + female\_above\_thresh/float(len(female\_sample))  
misclassification\_rate = 100 \* (overlap/2.0)  
# 16.9

Thus, our simple threshold based model would have an accuracy of about 83% in predicting gender from height.

### 3. Probability of Superiority

This method uses a **simulation-based approach** to estimate the probability that

a *randomly chosen from group A would be higher than a randomly chosen from group B*

men\_taller = 0  
for i in range(5000):  
 if Series(male\_sample).sample(1).iloc[0] > Series(female\_sample).sample(1).iloc[0]:  
 men\_taller += 1  
print 100 \* men\_taller/5000.0  
# 91.2

Over 5000 runs of this simulation, we find that a randomly chosen male would be taller than a randomly chosen female about 91% of the time.

Note: *Overlap (or misclassification rate)* and *probability of superiority* have two good properties: - As probabilities, they don’t depend on units of measure, so they are comparable between studies. - They are expressed in operational terms, so a reader has a sense of what practical effect the difference makes.

# Quiz

* Why is a simple difference in means not a robust way to report the effect size?
* Name two methods which can be useful in comparing effect sizes across studies.

# Assignments

Use the cars.csv data provided to answer the following questions

* Report the difference in average mileage of Automatic vs. Manual cars.
* How well can we predict type of engine from a car’s mileage? (Hint: Find the Overlap and Misclassification Rate?)
* Find the probability that a randomly chosen Automatic Car would have a mileage higher than a randomly chosen Manual Car.