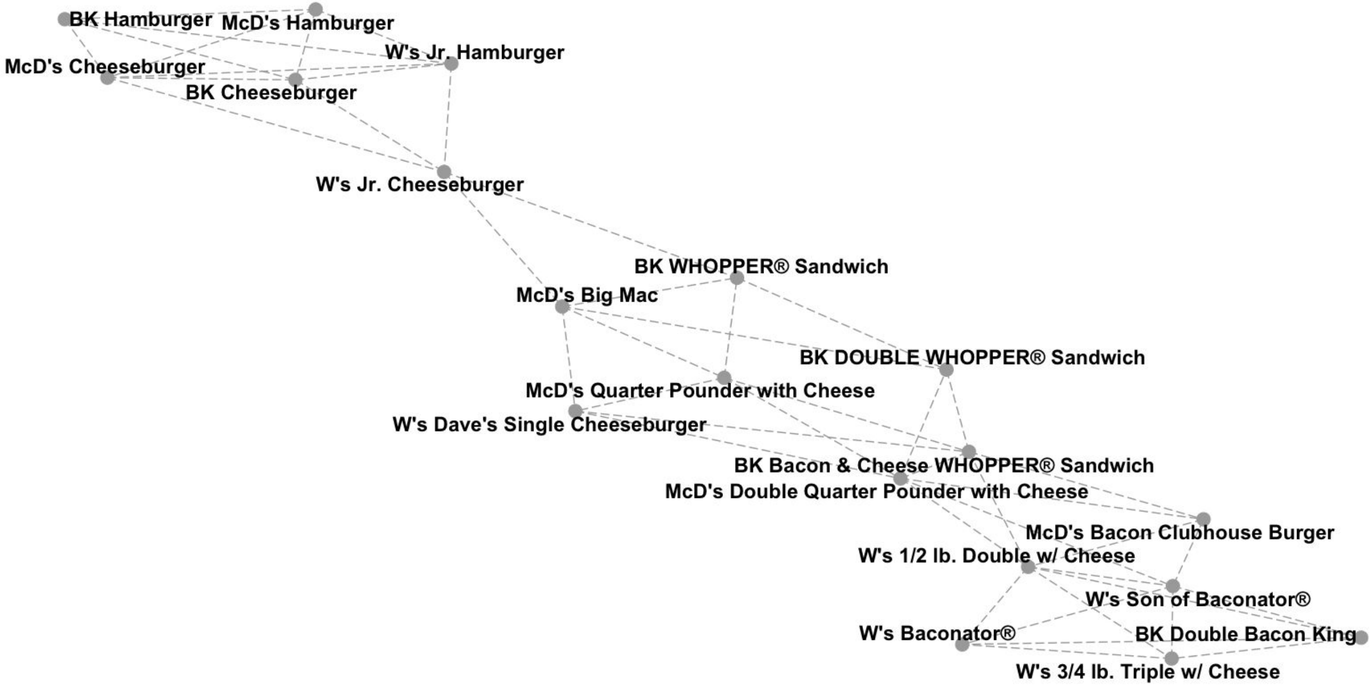
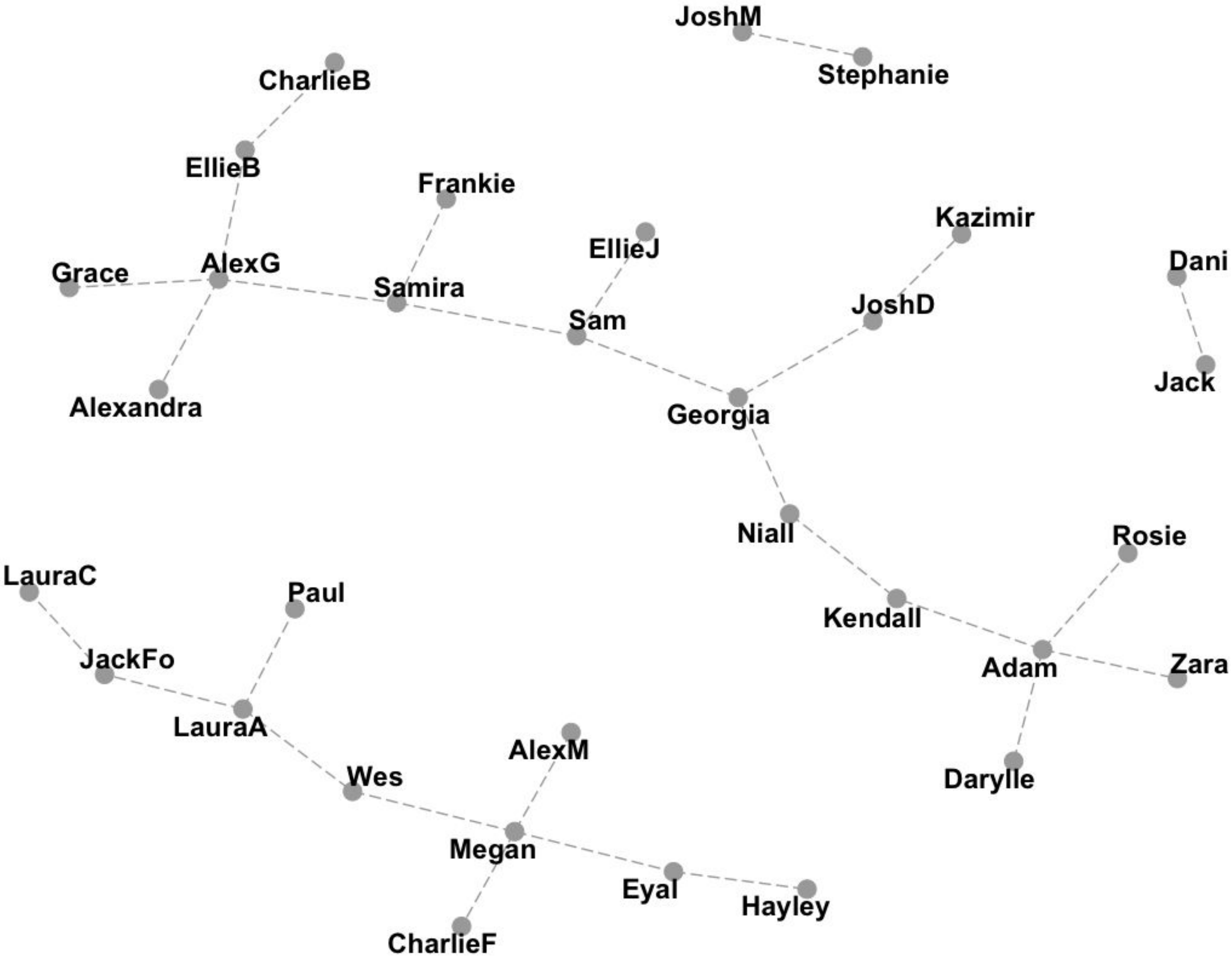
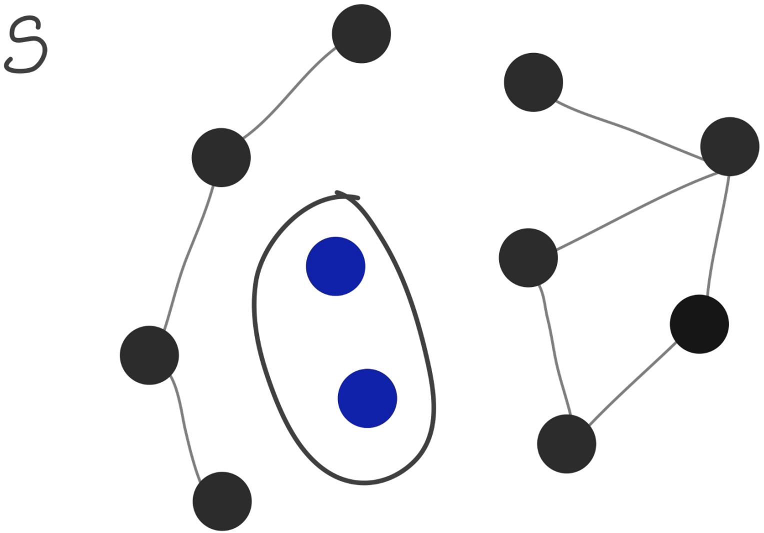
Topics In Data Science

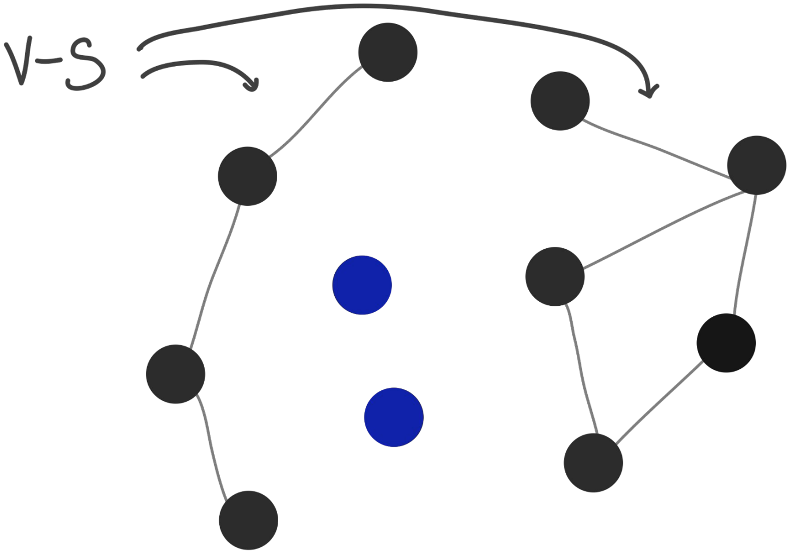
Chelsea Parlett-Pelleriti

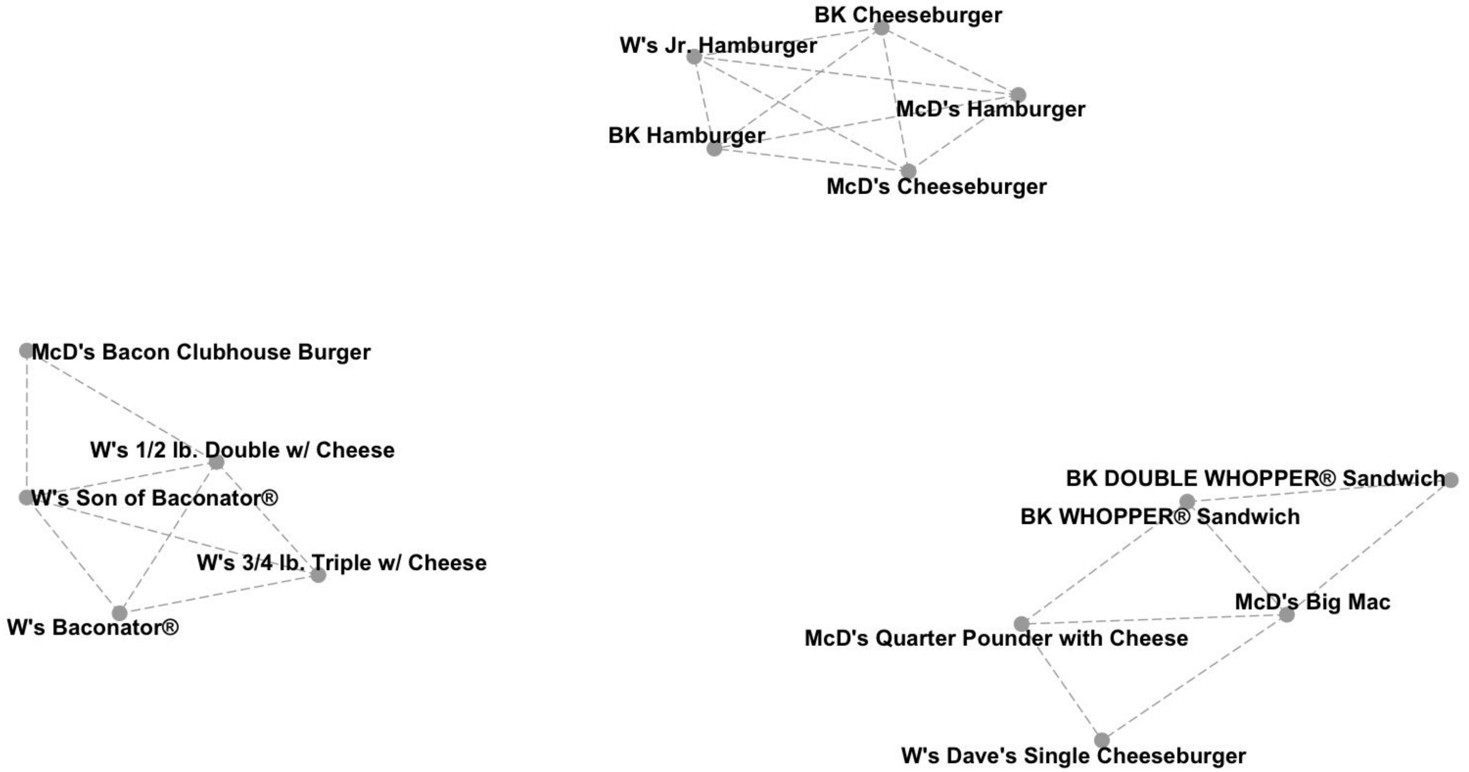
Node Based Resilience Clustering

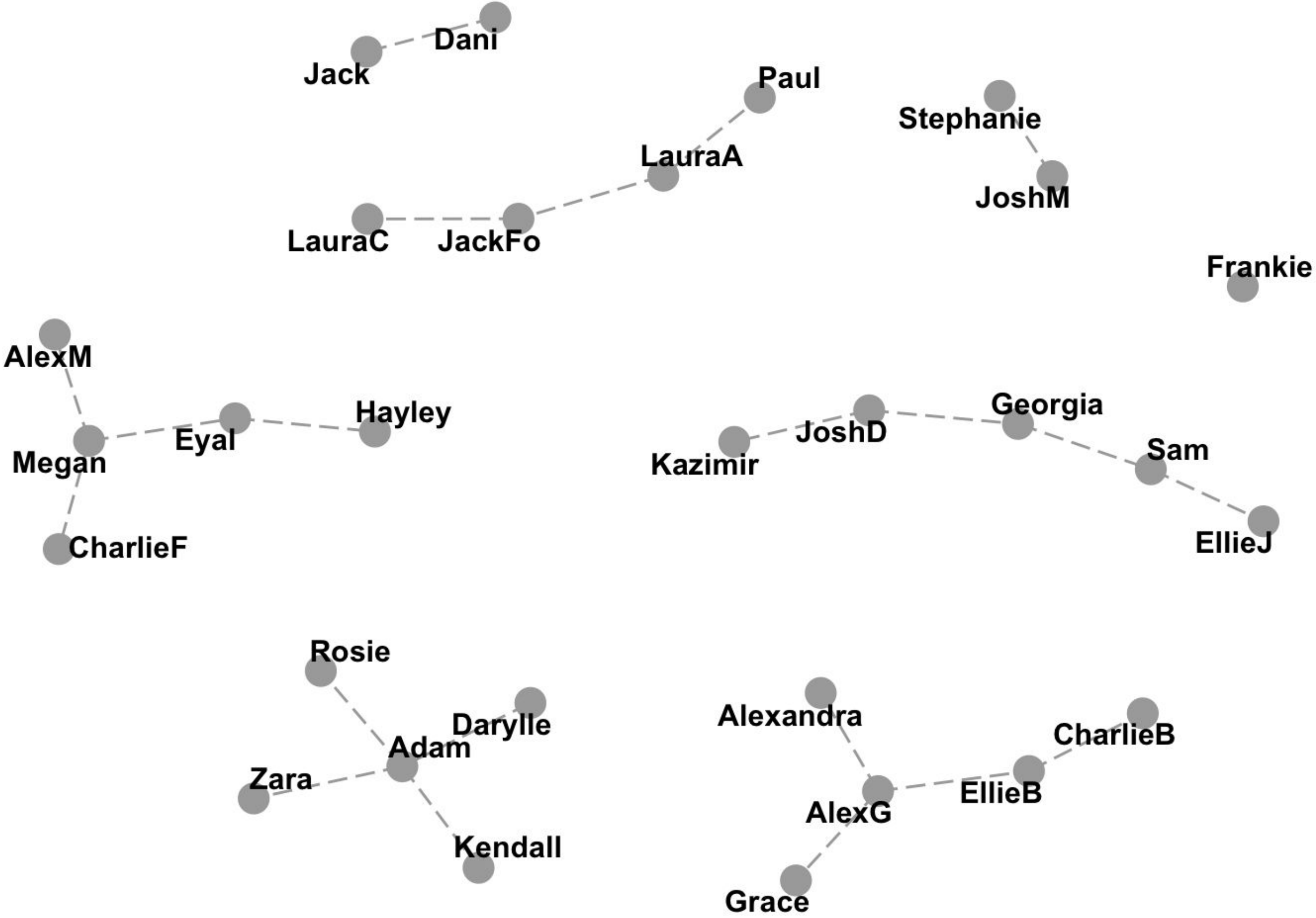

BK Hamburger
McD's Hamburger
W's Jr. Hamburger
McD's Cheeseburger
BK Cheeseburger
W's Jr. Cheeseburger
BK WHOPPER® Sandwich
McD's Big Mac
BK DOUBLE WHOPPER® Sandwich
McD's Quarter Pounder with Cheese
W's Dave's Single Cheeseburger
BK Bacon & Cheese WHOPPER® Sandwich
McD's Double Quarter Pounder with Cheese
McD's Bacon Clubhouse Burger
W's 1/2 lb. Double w/ Cheese
W's Son of Baconator®
W's Baconator®
BK Double Bacon King
W's 3/4 lb. Triple w/ Cheese


JoshM
Stephanie
CharlieB
EllieB
Frankie
EllieJ
Kazimir
Dani
Grace
AlexG
Samira
Sam
JoshD
Jack
Alexandra
Georgia
Niall
Rosie
LauraC
Paul
Kendall
JackFo
Adam
Zara
LauraA
AlexM
Darylle
Wes
Megan
Eyal
Hayley
CharlieF


S


V-S


BK Cheeseburger
W's Jr. Hamburger
McD's Hamburger
BK Hamburger
McD's Cheeseburger
McD's Bacon Clubhouse Burger
W's 1/2 lb. Double w/ Cheese
W's Son of Baconator®
W's 3/4 lb. Triple w/ Cheese
W's Baconator®
BK DOUBLE WHOPPER® Sandwich
BK WHOPPER® Sandwich
McD's Big Mac
McD's Quarter Pounder with Cheese
W's Dave's Single Cheeseburger


Dani
Jack
Paul
Stephanie
LauraA
JoshM
LauraC
JackFo
Frankie
AlexM
Hayley
Eyal
Megan
Georgia
JoshD
Sam
Kazimir
CharlieF
EllieJ
Rosie
Darylle
Adam
Zara
Kendall
Alexandra
CharlieB
EllieB
AlexG
Grace

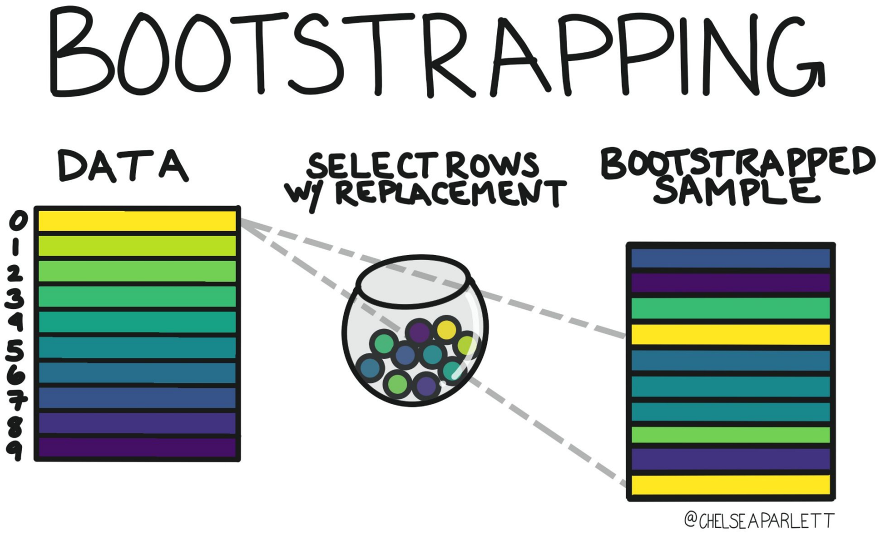
Ensemble Methods

ONE vs. MANY

Random Forest

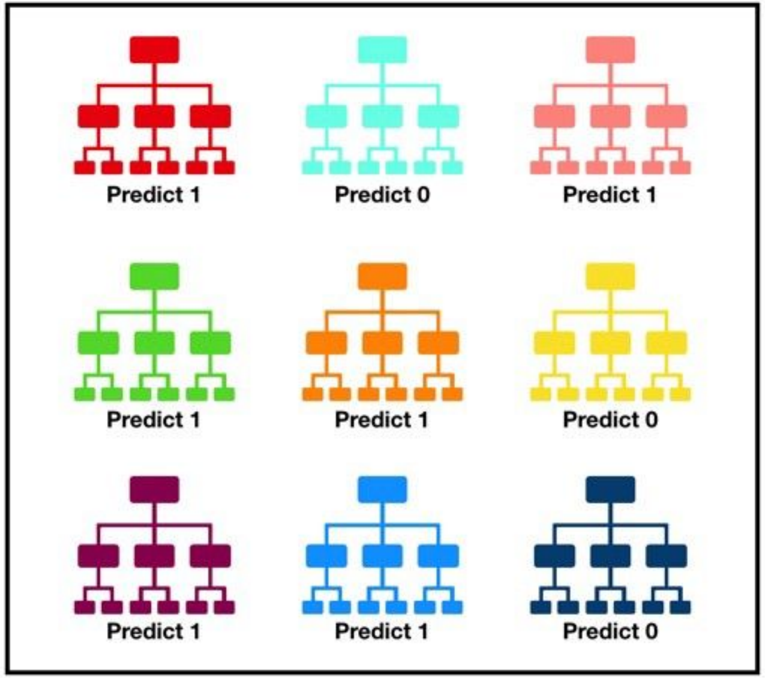
Random Forest

- **Bagging (Bootstrap Aggregating)**: Instead of using all of our training data to train each model in our sample we use **bootstrapping** to choose the samples (rows) we will include.
    - **Bootstrapping** is when you randomly sample data points *with replacement*, meaning that a data point can be included in your bootstrapped sample *more* than once, OR not at all.



Random Forest

- **Random Feature Selection**: Instead of using all the available *features/predictors* in our dataset for every model, for each model *we randomly choose a different subset of features to use* when training.
  - This helps our ensemble generalize, because it doesn't become overly reliant on one feature (since that feature might not appear in every model).



Bayesian Statistics

Bayesian Statistics

$$\underbrace{P(\theta\,|\,D)}_{posterior} \propto \underbrace{P(\theta)}_{prior}\,\underbrace{P(D\,|\,\theta)}_{likelihood}$$

Data + Expertise =

Inference

Bayesian Statistics

Data =

Inference

Bayesian Statistics

MINIMIZE:

$$\sum\Big(\overbrace{\underset{\substack{\uparrow\\\text{true}\\\text{value}}}{x_i} - \underset{\substack{\uparrow\\\text{model's}\\\text{guess}}}{\hat{x}_i}}^{\text{how off we were}}\Big)^2 + \underset{\substack{\uparrow\\\text{how HARSHLY}\\\text{we penalize}}}{\lambda}\sum \underset{\substack{\uparrow\\\text{how big the}\\\text{coefs are}}}{B_j}^{\boxed{2}}$$

MINIMIZE:

$$\sum\Big(\overbrace{\underset{\substack{\uparrow\\\text{true}\\\text{value}}}{x_i} - \underset{\substack{\uparrow\\\text{model's}\\\text{guess}}}{\hat{x}_i}}^{\text{how off we were}}\Big)^2 + \underset{\substack{\uparrow\\\text{how HARSHLY}\\\text{we penalize}}}{\lambda}\sum \underset{\substack{\uparrow\\\text{how big the}\\\text{coefs are}}}{|B_j|}$$

Bayesian Statistics

What we believed before + **Evidence from the data** =
New beliefs