DATA FOLKZ
CATAPULT DATA LEADERS

ROC AND AUC

LOGISTIC REGRESSION METRICS

DATA FOLKZ
CATAPULT DATA LEADERS

AGENDA

- Introduction
- ROC curve
- Area under the ROC curve (AUC)
- Examples using ROC
- Concluding remarks

DATA FOLKZ
CATAPULT DATA LEADERS

ROC AND AUC

- Logistic regression gives Probability forecasts for the given data point to be in a given bucket.
- A threshold needs to be chosen to finally translate this probability to a bucket allocation
- At a given threshold, we can evaluate the classification accuracy (accuracy, sensitivity, recall, kappa etc)
- ROC curve tries to evaluate how well the regression has achieved the separation between the classes at all threshold values

DATA FOLKZ
CATAPULT DATA LEADERS

ROC CURVE DEMO

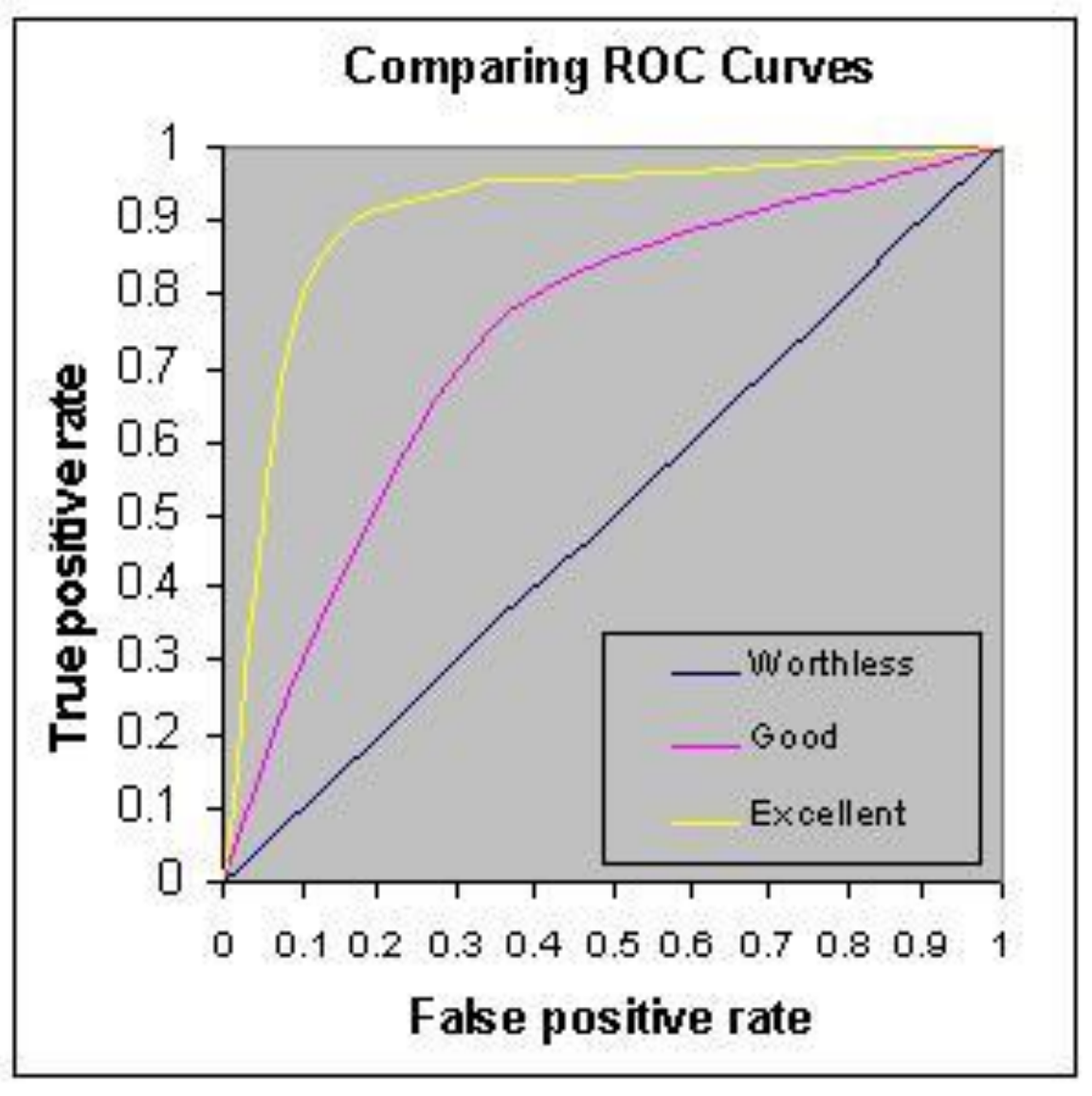
- http://www.navan.name/roc/

- https://www.youtube.com/watch?v=OAl6eAyP-yo&feature=youtu.be

DATA FOLKZ
CATAPULT DATA LEADERS

ROC CURVES AND AUC

- AUC – Measures discrimination, i.e., ability to correctly classify those with and without CHD.
- If you randomly pick one person who HAS CHD and one who DOESN'T and run the model, the one with the higher probability should be from the high risk group.
- AUC is the percentage of randomly drawn such pairs for which the classification is done correctly.



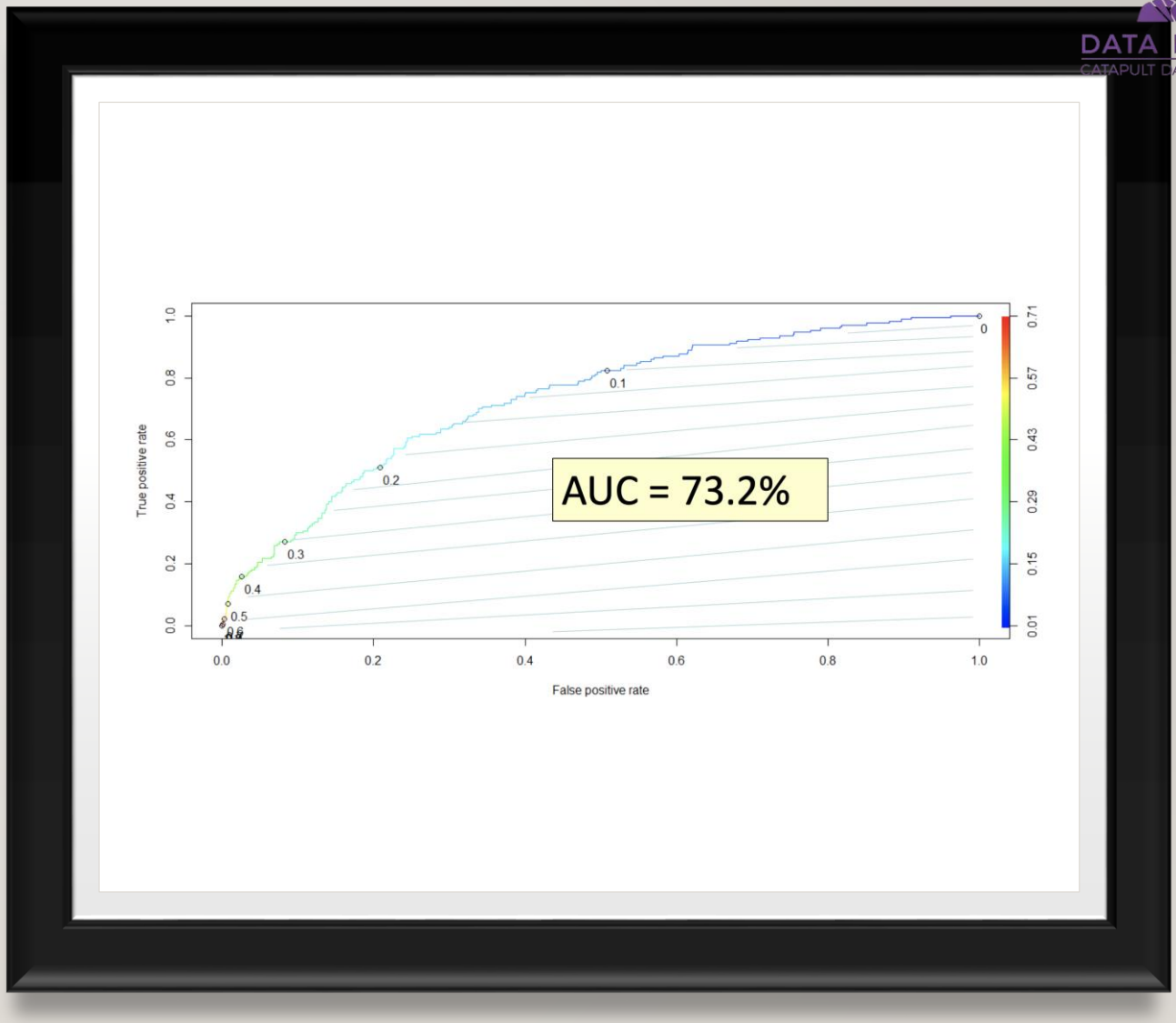
ROC CURVES

- 0.90 -1.0 = Excellent
- 0.80 – 0.90 = Good
- 0.70 – 0.80 = Fair
- 0.60 – 0.70 = Poor
- 0.50 – 0.60 = Fail

- <0.50 , coin toss is better than the prediction model

ROC CURVES AND AUC

- The model does a fair job of discrimination between high risk and low risk people.
- Useful for comparing different models.



GAINS AND LIFT CHARTS

- In some business problems, it is not good enough to just classify. For example, in direct mail or phone marketing campaigns, where it costs money to send a mail to each prospect, it is better to be able to rank the prospective buyers by their probability to buy. That way, you can order them and start calling or mailing them in their decreasing order of propensity to buy.
- Lift is a measure of the effectiveness of a predictive model calculated as the ratio between the results obtained with and without the predictive model (random selection).

DATA FOLKZ
CATAPULT DATA LEADERS

GAINS AND LIFT CHARTS

- A Lift Chart describes how well a model ranks samples in a particular class.
- The greater the area between the lift curve and the baseline (random selection), the better the model.

GAINS AND LIFT CHARTS

- A company sends mail catalogs to prospective buyers. It costs the company $1 to print and mail one catalog.
- From past data, they know the response rate is 5%, i.e., if 100,000 prospective customers are contacted, 5000 buy.
- This means that if there is no model and the company randomly contacts the prospects, they will have the following result.

No of Customers contacted	No of responses
10000	500
20000	1000
30000	1500
.	.
.	.
.	.
100000	5000

GAINS AND LIFT CHARTS

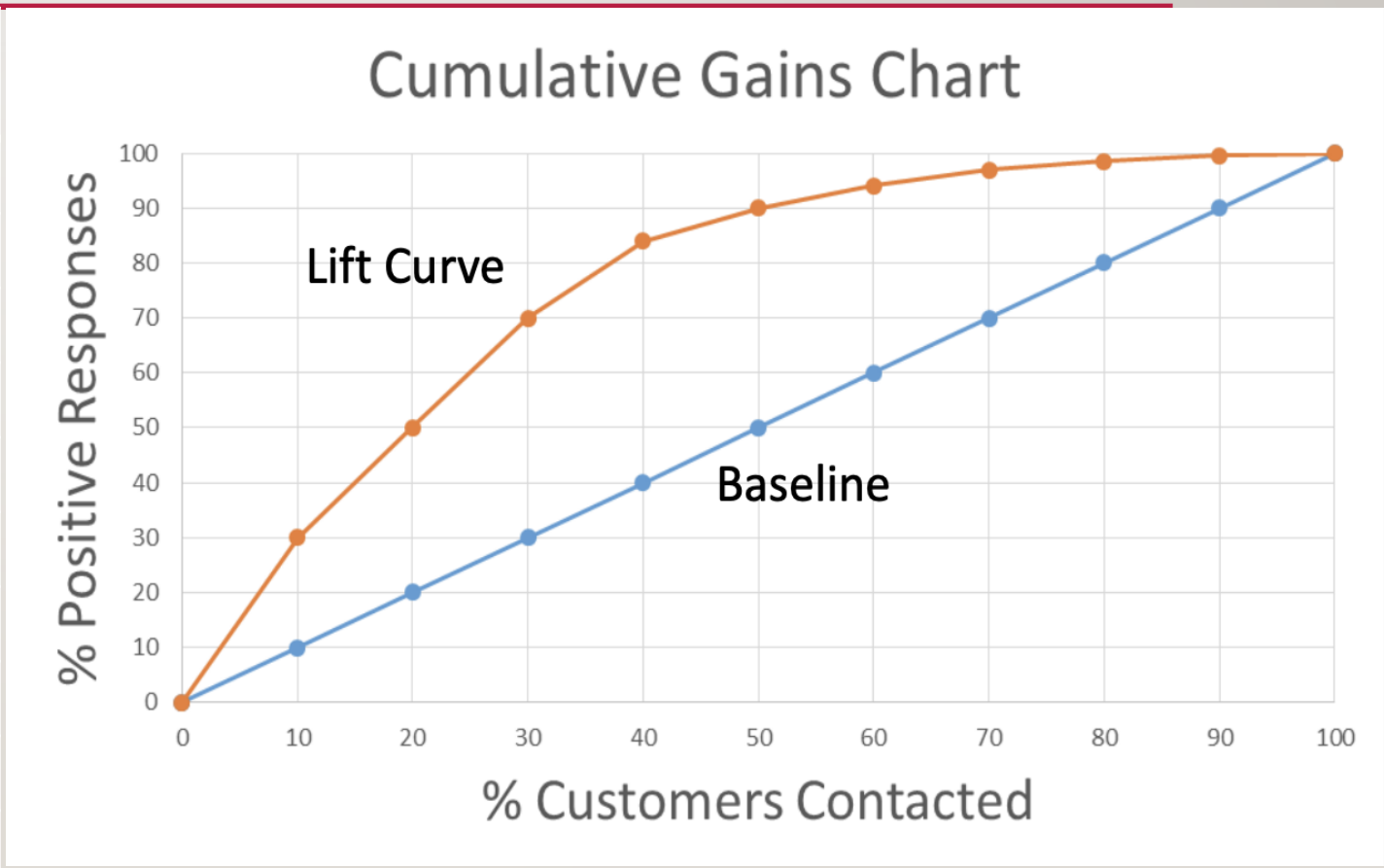
- With a predictive model, where the model assigns a probability to each customer, the customers are ordered and divided into deciles (or any other quantiles). They are then called in decreasing order of probability to buy.

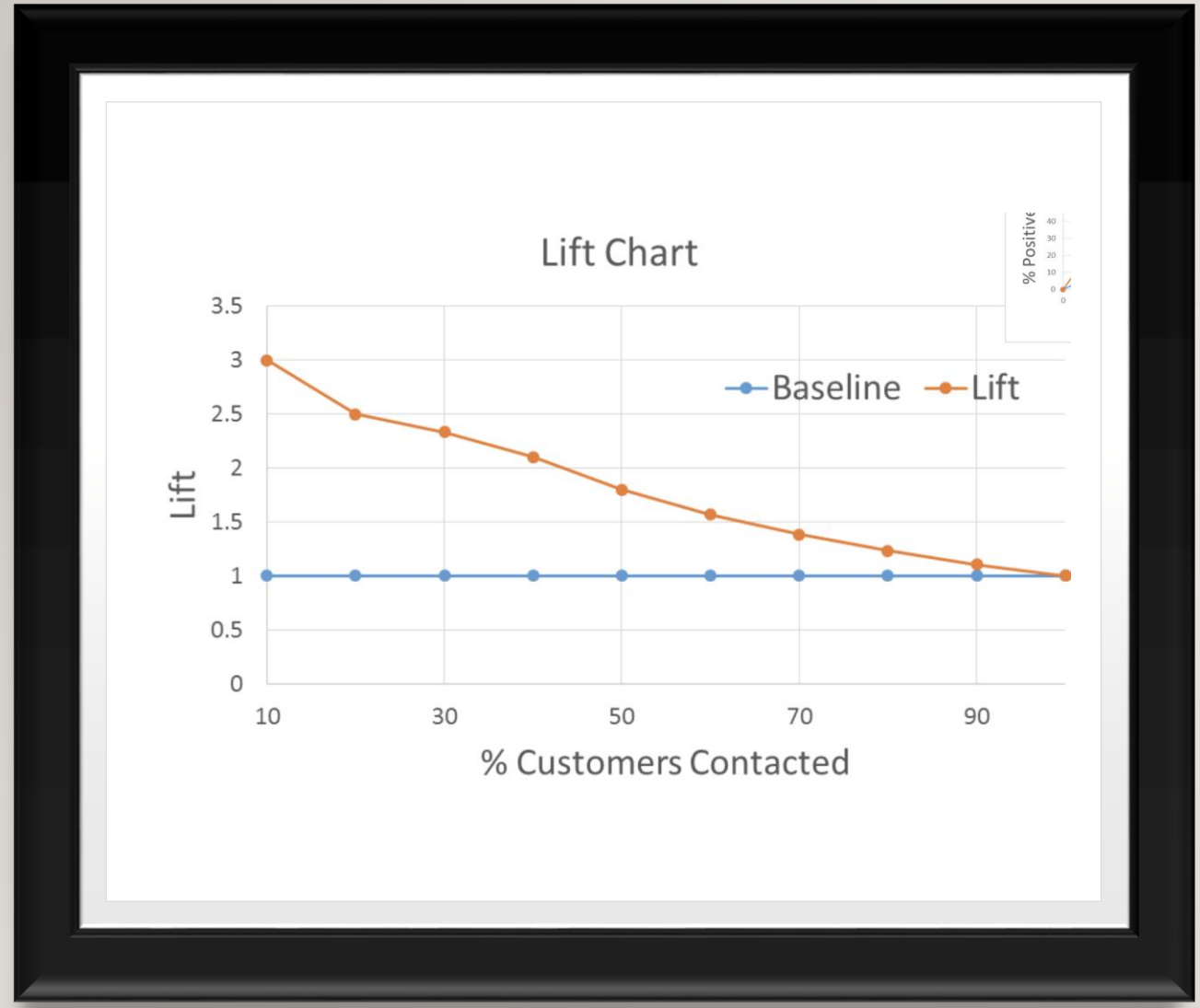
Cost ($)	Decile contacted	Cummulative responses
10000	10(top)	1500
20000	9	2500
30000	8	3500
40000	7	4200
50000	6	4500
60000	5	4700
70000	4	4850
80000	3	4925
90000	2	4975
100000	1	5000

DATA FOLKZ
CATAPULT DATA LEADERS

GAINS AND LIFT CHARTS

%called	Random	As per model
0	0	0
10	10	30
20	20	50
30	30	70
40	40	84
50	50	90
60	60	94
70	70	97
80	80	98.5
90	90	99.5
100	100	100





GAINS AND LIFT CHARTS

- Max lift of 3 at the top decile
- Model advantage diminishes as more customers are contacted, especially in lower deciles
- Useful to compare different models

DATA FOLKZ
CATAPULT DATA LEADERS

OTHER PERFORMANCE MEASURES FOR CLASSIFICATION MODELS

KAPPA METRIC

- Accuracy can often be a misleading metric, when one category occurs more often than other in the given data-set
  - – For eg: Occurrence of cancer in general population is 0.4%
  - If a prediction system blindly marks everyone as "No cancer", it will 99.6% accurate

DATA FOLKZ
CATAPULT DATA LEADERS

KAPPA METRIC

- Kappa metric quantifies how accurate the prediction algorithm is when compared to a random prediction
- $kappa = \frac{total\ accuracy - randomaccuracy}{1 - random\ accuracy}$
- $totalAccuracy = \frac{correct\ predictions}{Total}$
- randomAccuracy = $\frac{ActualFalse}{Total} * \frac{PredictedFalse}{Total} + \frac{ActualTrue}{Total} * \frac{PredictedTrue}{Total}$

DATA FOLKZ
CATAPULT DATA LEADERS

KAPPA VALUE

Kappa Value	
<0	No agreement
0-0.2	Slight
0.21 to 0.4	Fair
0.4 to 0.6	Moderate
0.6 to 0.8	Substantial
0.8 to 1	Almost perfect

KAPPA METRIC

10 - Year CHD Risk		Predicted	
		TRUE	FALSE
Actual	TRUE	30	357
	FALSE	9	2170

- Total= 30+357+9+2170=2566
- TotalAccuracy=(30+2170)/2566=0.857
- PercTrue=(30+357)/2566 = 0.15
- PercFalse=(9+2170)/2566 = 0.85
- PredTrue=(30+9)/2566=0.015
- PredFalse=(357+2170)/2566 = 0.985
- randomAccuracy= 0.15*0.015 + 0.85*0.985 = 0.84
- Kappa =0.1

DATA FOLKZ
CATAPULT DATA LEADERS

NAÏVE BAYES ALGORITHM

CLASSIFICATION PROBLEMS WITH MULTIPLE CLASSES

- Given an article – predict which section of the news paper (Current News, International, Arts, Sports,Fashion etc) it supposed to go
- Given a photo of a car number plate, identify which state it belongs to
- Audio clip of a song, identify the genre

CLASSIFICATION PROBLEMS

- All classification problems essentially equivalent to evaluating conditional probability
- P(Yi | X) i.e. Given certain evidence X, what is the probability that this if from class Yi
- Logistic Regression solves this problem by modelling the probabilistic relationship between X and Y (sigmoid function, linear in X etc)
- Such models are called Discriminative models

DATA FOLKZ
CATAPULT DATA LEADERS

NAÏVE BAYES ALGORITHM

- Naïve Bayes: Computes P(Yi | X) by using Bayes theorem and instead computes the inverse conditional probability P(X |Yi)
- A simple classifier that performs surprisingly well on a large class of problems
- This type of methods are called Generative Learning Models

EXAMPLE: US-VOTING PATTERNS

- A data frame with 435 observations on 17 variables. 168 Republicans, 267 Democrats
- Given: A congressman's voting pattern (v1 = y, v2=n), what is the probability that this person is a democrat?

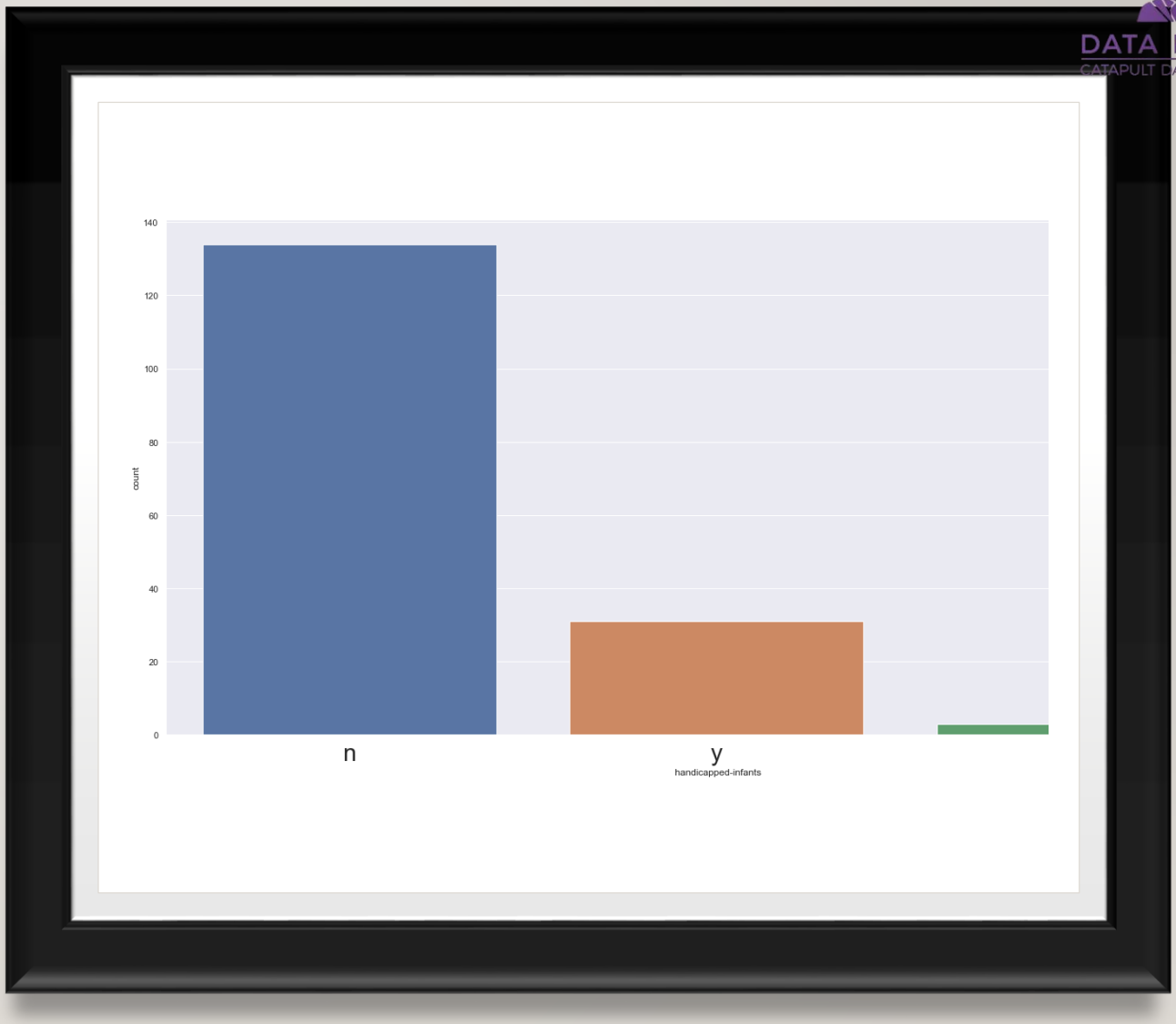
DATA FOLKZ
CATAPULT DATA LEADERS

PRIOR BELIEF - SIMPLEST SOLUTION

- The house has a majority of Democrats
  - 168 Republicans , 267 democrats

- Probability for a random person being a democrat is P(D) = 267/435 = 0.61

HANDICAPPED REPUBLICANS



HANDICAPPED DEMOCRATS