

COMMUNICATING ENSEMBLE RESULTS

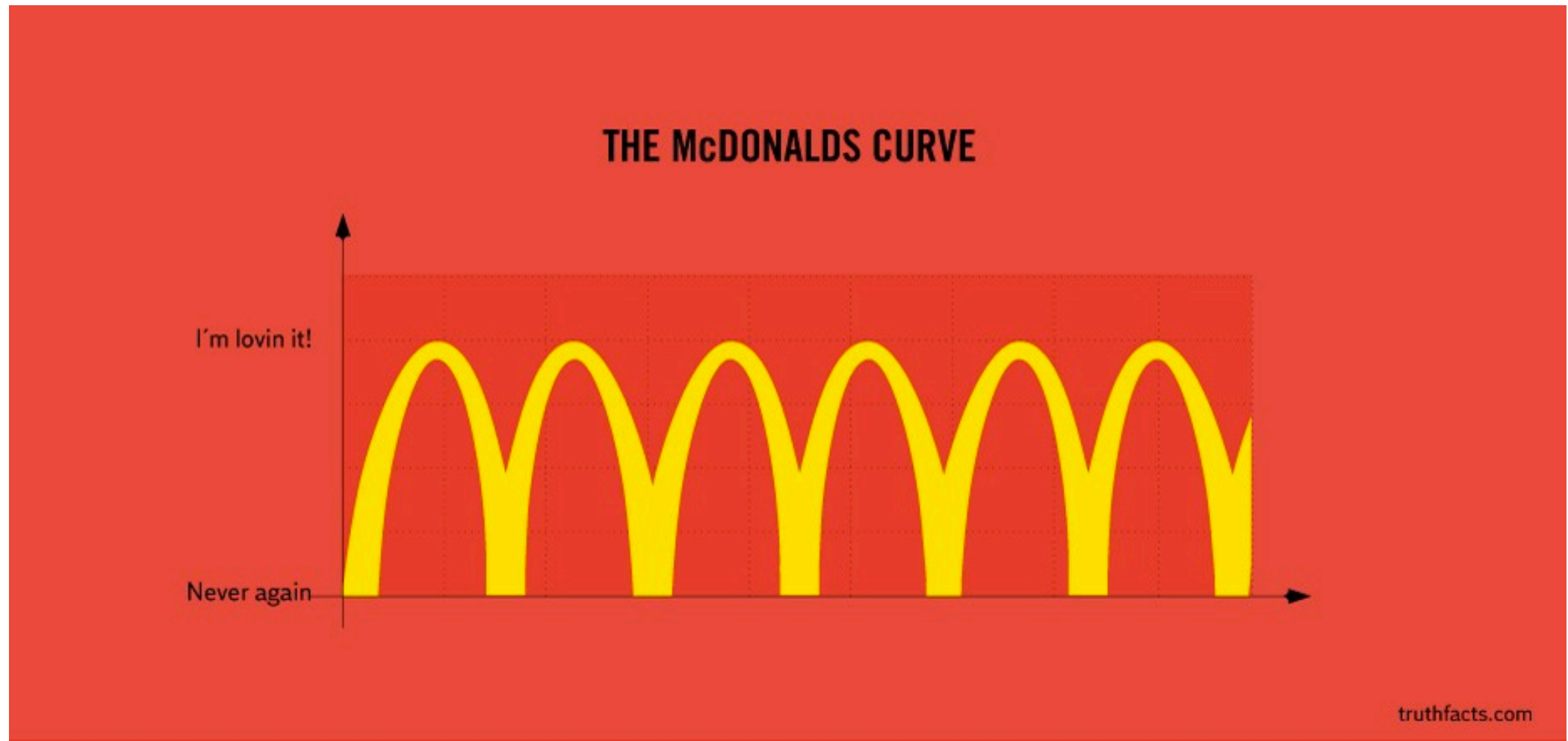
Joseph Nelson, Data Science Immersive

AGENDA

- Review of Metrics
- Cost Benefit Analysis
- Follow-up Questions
- WTFViz

COMMUNICATING RESULTS TO NON-TECHNICAL AUDIENCES

- ▶ Should you present results differently for ensemble models vs non-ensemble models?



REVIEWING MODEL QUALITY

- ▶ High level review question: what are the two types of models we've discussed?
How do they differ?



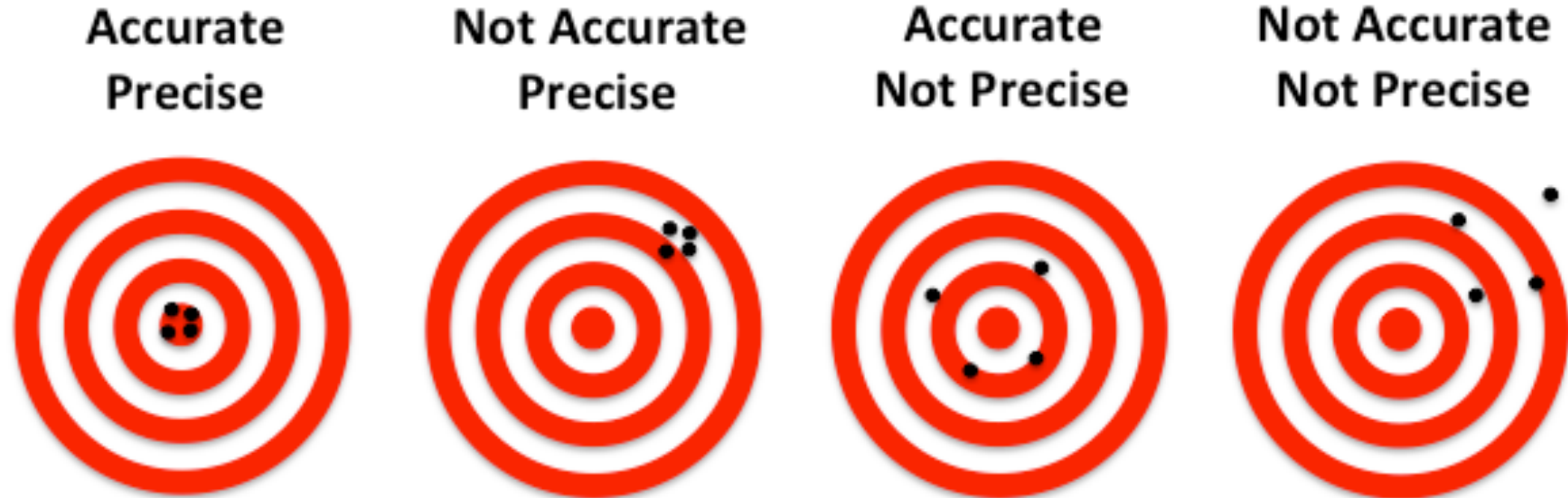
REVIEWING CLASSIFICATION MODEL QUALITY

- What are the metrics we use to assess classification models?

$$\text{accuracy}(y, \hat{y}) = \frac{1}{n_{\text{samples}}} \sum_{i=0}^{n_{\text{samples}}-1} 1(\hat{y}_i = y_i)$$

REVIEWING CLASSIFICATION MODEL QUALITY

► What are the metrics we use to assess classification models?



REVIEWING CLASSIFICATION MODEL QUALITY

- What are the metrics we use to assess classification models?
- The F-measure (F_{β} and F_1 measures) can be interpreted as a weighted harmonic mean of the precision and recall. A measure reaches its best value at 1 and its worst score at 0.

$$F_1 = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}.$$

REVIEWING CLASSIFICATION MODEL QUALITY

- What are the metrics we use to assess classification models?
- Confusion Matrix
- How many True Positives, True Negatives, False Positives, and False Negatives are there?

n=165	Predicted: NO	Predicted: YES
Actual: NO	50	10
Actual: YES	5	100

REVIEWING CLASSIFICATION MODEL QUALITY

- What are the metrics we use to assess classification models?
- Confusion Matrix
- Accuracy score?
- Misclassification Rate?
- True Positive Rate? (Recall)
- False Positive Rate?
- Specificity?
- Precision?

n=165	Predicted: NO	Predicted: YES	
Actual: NO	TN = 50	FP = 10	60
Actual: YES	FN = 5	TP = 100	105
	55	110	

REVIEWING CLASSIFICATION MODEL QUALITY

- What are the metrics we use to assess classification models?
- Confusion Matrix
- Accuracy score?
- Misclassification Rate?
- True Positive Rate? (Recall)
- False Positive Rate?
- Specificity?
- Precision?

Accuracy: Overall, how often is the classifier correct?
 $(TP+TN)/total = (100+50)/165 = 0.91$

True Positive Rate: When it's actually yes, how often does it predict yes? $TP/actual\ yes = 100/105 = 0.95$

False Positive Rate: When it's actually no, how often does it predict yes? $FP/actual\ no = 10/60 = 0.17$

Specificity: When it's actually no, how often does it predict no? $TN/actual\ no = 50/60 = 0.83$

Precision: When it predicts yes, how often is it correct? $TP/predicted\ yes = 100/110 = 0.91$

REVIEWING REGRESSION MODEL QUALITY

- What are the metrics we use to assess regression models?

REVIEWING REGRESSION MODEL QUALITY

► What are the metrics we use to assess regression models?

► Root Mean Squared Error

$$\text{RMSD} = \sqrt{\frac{\sum_{t=1}^n (\hat{y}_t - y_t)^2}{n}}.$$

► R-Squared Value

$$R^2 \equiv 1 - \frac{SS_{\text{res}}}{SS_{\text{tot}}}$$
$$SS_{\text{res}} = \sum_i (y_i - f_i)^2 = \sum_i e_i^2$$
$$SS_{\text{tot}} = \sum_i (y_i - \bar{y})^2$$

COST BENEFIT ANALYSIS

- ▶ Often, our outputs are not in terms of real financial dollars and, therefore, are difficult for our managers to make actionable decisions
- ▶ Imagine a marketing campaign where you're trying to minimize user churn (the rate at which a user signups and ultimately leaves)
- ▶ Hypothetically, say that your business makes \$10 per user per month. The cost of thinking a user signed up but realizing they ultimately did not is \$0.05.
- ▶ Given your model outputs the confusion matrix at the right, what is the cost of user retention?

TP: 0.2	FP: 0.2
FN: 0.1	TN: 0.5

COST BENEFIT ANALYSIS

‣ $P(TP) \cdot B(TP) + P(TN) \cdot B(TN) + P(FP) \cdot C(FP) + P(FN) \cdot C(FN)$

‣ In this case:

TP: 0.2	FP: 0.2
FN: 0.1	TN: 0.5

COST BENEFIT ANALYSIS

- ▶ $P(TP) \cdot B(TP) + P(TN) \cdot B(TN) + P(FP) \cdot C(FP) + P(FN) \cdot C(FN)$
- ▶ In this case:
- ▶ $(0.2 \times 10) + (0.5 \times 0) - (0.2 \times 0.05) - (0.1 \times 0)$

TP: 0.2	FP: 0.2
FN: 0.1	TN: 0.5

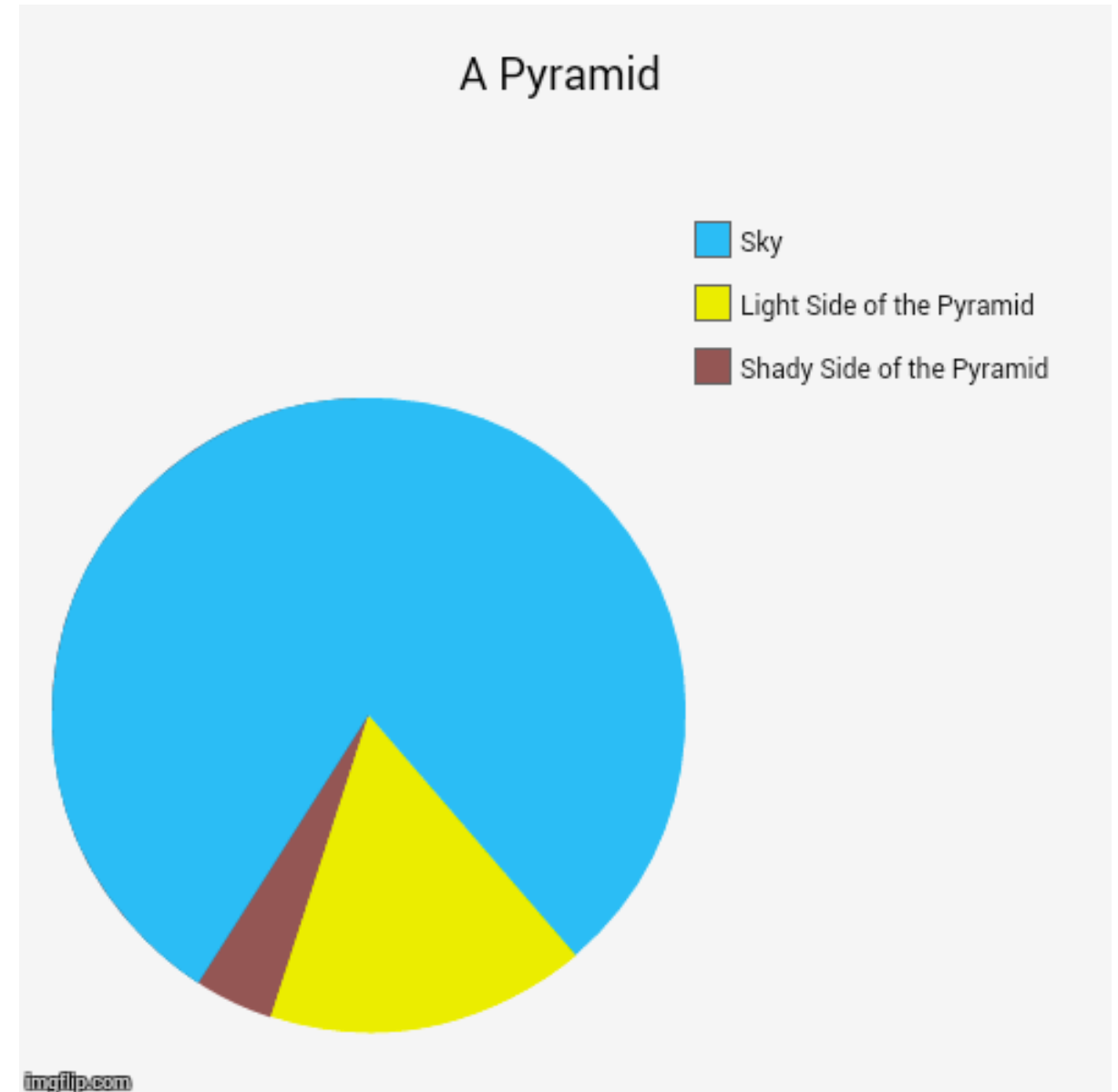
FOLLOW-UP QUESTIONS

- ▶ How would you rephrase the business problem if your model was optimizing toward precision? i.e., How might the model behave differently, and what effect would it have?
- ▶ How would you rephrase the business problem if your model was optimizing toward recall?
- ▶ What would the most ideal model look like in this case?
- ▶ Can you think of business situations where different stakeholders would take different decisions on what metric to optimize? For example, stakeholders with competing interests may decide to weigh false positives or false negatives differently. Can you think of a concrete example? > Answer: E.g. model to predict cancer: Health insurance would like to minimize false positives, patient would like to minimize false negatives.

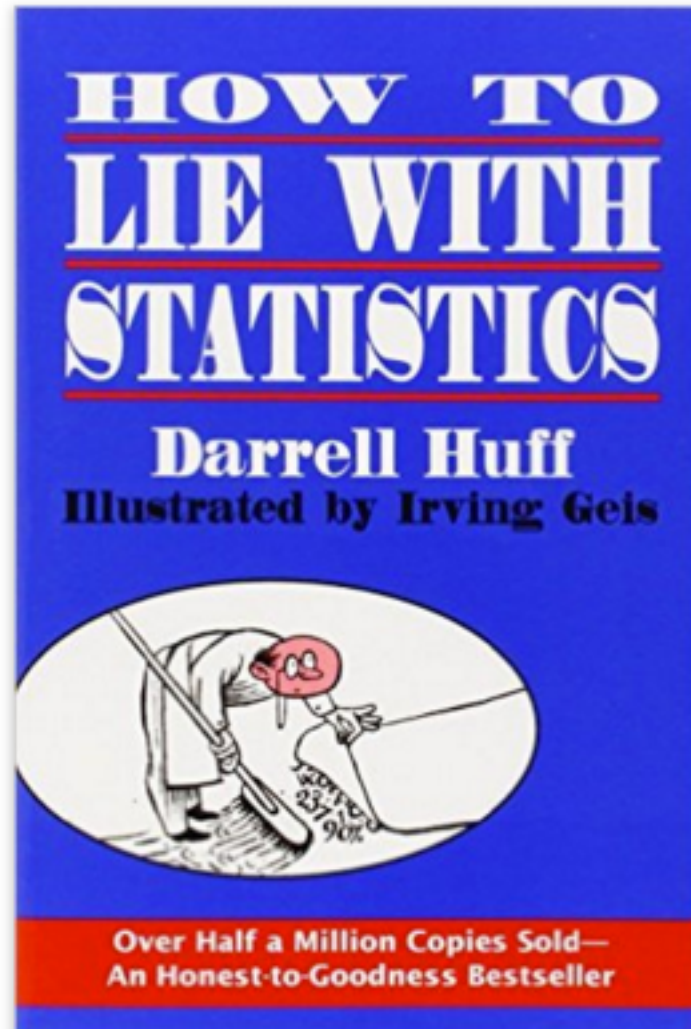
CRITERIA FOR GOOD VISUALIZATION

- ▶ Simplified
- ▶ Easy to Interpret
- ▶ Clearly Labeled

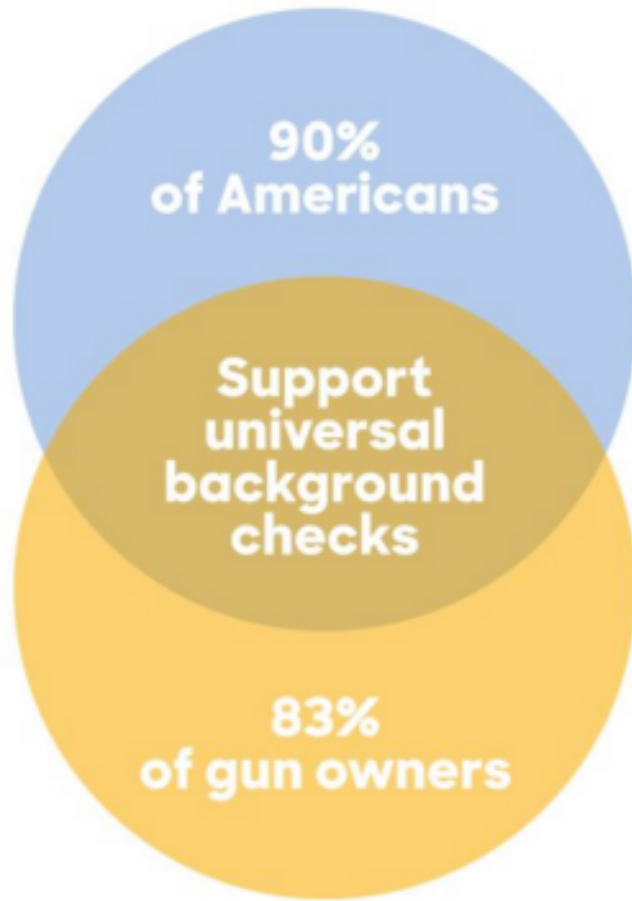
- ▶ Ask yourself:
 - ▶ Who is my target audience?
 - ▶ What do they already know, and what do they need to know?
 - ▶ How does my project affect this audience? How might they interpret (or misinterpret) the data?



CRITERIA FOR GOOD VISUALIZATION



CRITERIA FOR GOOD VISUALIZATION



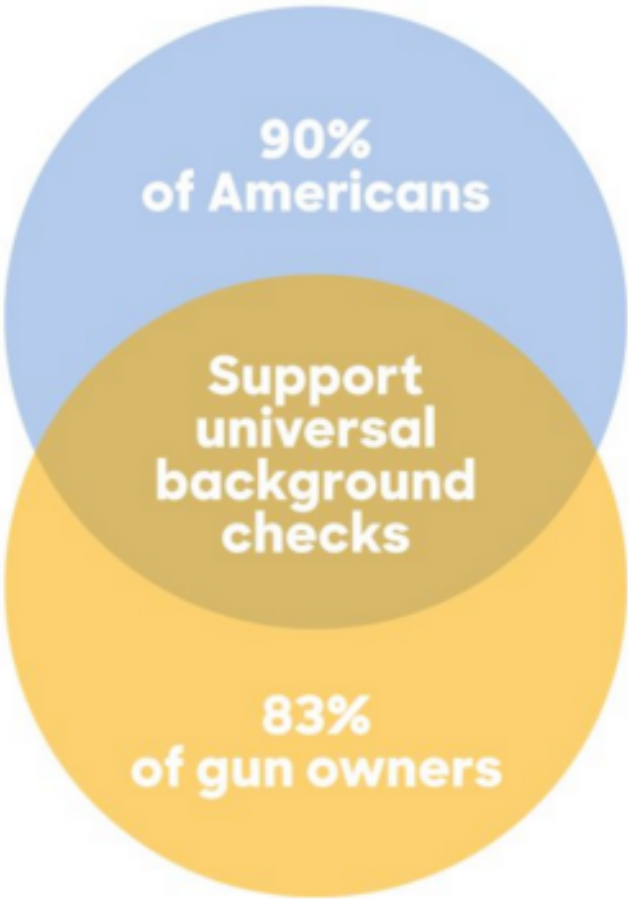
Hillary Clinton ✓
@HillaryClinton

 Follow

Dear Congress,

Let's get this done.


CRITERIA FOR GOOD VISUALIZATION



90%
of Americans

Support
universal
background
checks

83%
of gun owners



Hillary Clinton ✓
@HillaryClinton

Dear Congress,

Let's get this done.

Follow



People who know
how to make
Venn Diagrams

Hillary's graphic
design staff

In reply to Hillary Clinton



Michael Deppisch @deppisch · May 20
@hillaryclinton

3.7K

5.4K