

# Machine Learning for Data Science (CS4786)

## Lecture 24

Differential Privacy and Re-useable Holdout

# Defining Privacy

# Defining Privacy

**Dataset +**



# Defining Privacy



# Defining Privacy

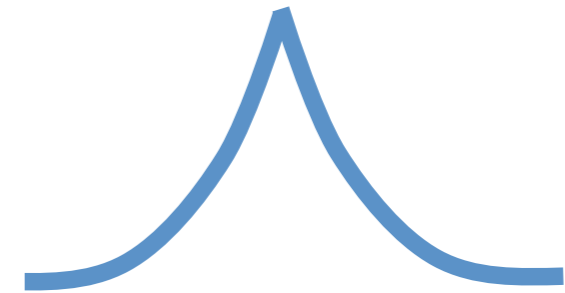


# Defining Privacy

**Dataset +**



Learning  
Algorithm



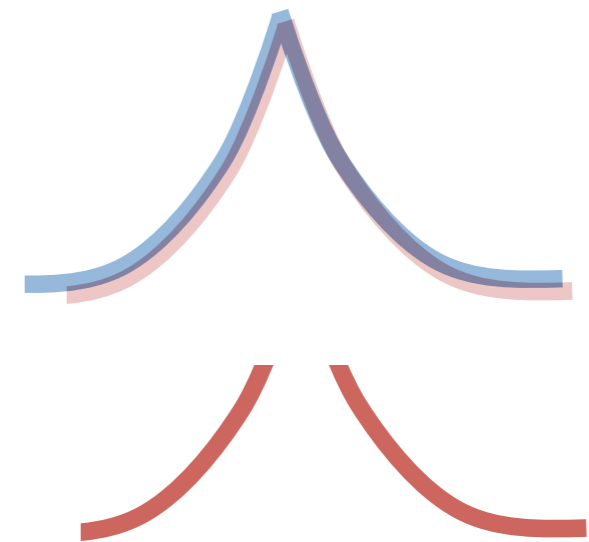
**Distribution of outcome**

**Similar**

**Dataset +**



Learning  
Algorithm



**Distribution of outcome**

# Differential Privacy

# Differential Privacy

- A deterministic algorithm cannot preserve privacy



# Differential Privacy

- A deterministic algorithm cannot preserve privacy
- Say  $S = (\text{Data}_1, \dots, \text{Data}_n)$  is the data provided to learning algorithm (be it clustering, supervised learning etc).

# Differential Privacy

- A deterministic algorithm cannot preserve privacy
- Say  $S = (\text{Data}_1, \dots, \text{Data}_n)$  is the data provided to learning algorithm (be it clustering, supervised learning etc).
- Say (randomized) learning algorithm  $A$  takes this training data and returns solution as  $A(S)$

# Differential Privacy

- A deterministic algorithm cannot preserve privacy
- Say  $S = (\text{Data}_1, \dots, \text{Data}_n)$  is the data provided to learning algorithm (be it clustering, supervised learning etc).
- Say (randomized) learning algorithm  $A$  takes this training data and returns solution as  $A(S)$
- Algorithm  $A$  is  $(\epsilon, \delta)$ - differentially private if for all samples  $S$  and  $S'$  that only differ by one data point and any set  $C$

$$P(A(S) \in C) \leq e^\epsilon P(A(S') \in C) + \delta$$

# Differential Privacy

- A deterministic algorithm cannot preserve privacy
- Say  $S = (\text{Data}_1, \dots, \text{Data}_n)$  is the data provided to learning algorithm (be it clustering, supervised learning etc).
- Say (randomized) learning algorithm  $A$  takes this training data and returns solution as  $A(S)$
- Algorithm  $A$  is  $(\epsilon, \delta)$ - differentially private if for all samples  $S$  and  $S'$  that only differ by one data point and any set  $C$

$$P(A(S) \in C) \leq e^\epsilon P(A(S') \in C) + \delta$$

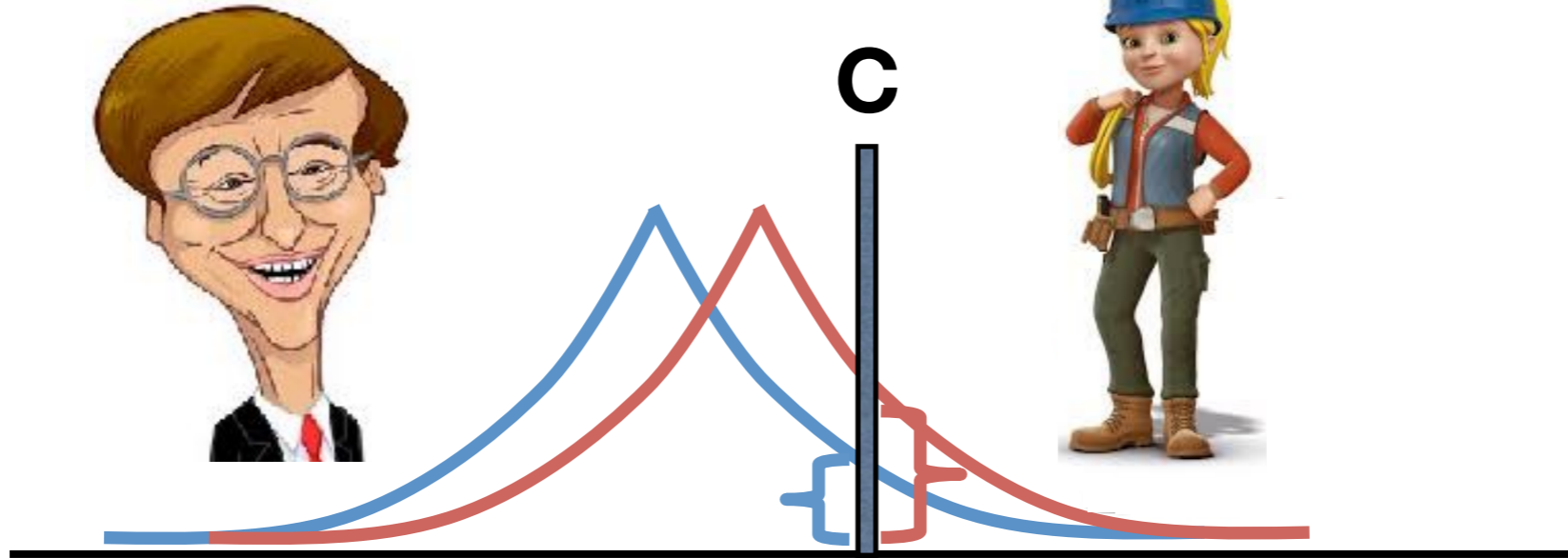
- $\delta=0$  is called pure differential privacy

# Differential Privacy

# Differential Privacy

Dataset +

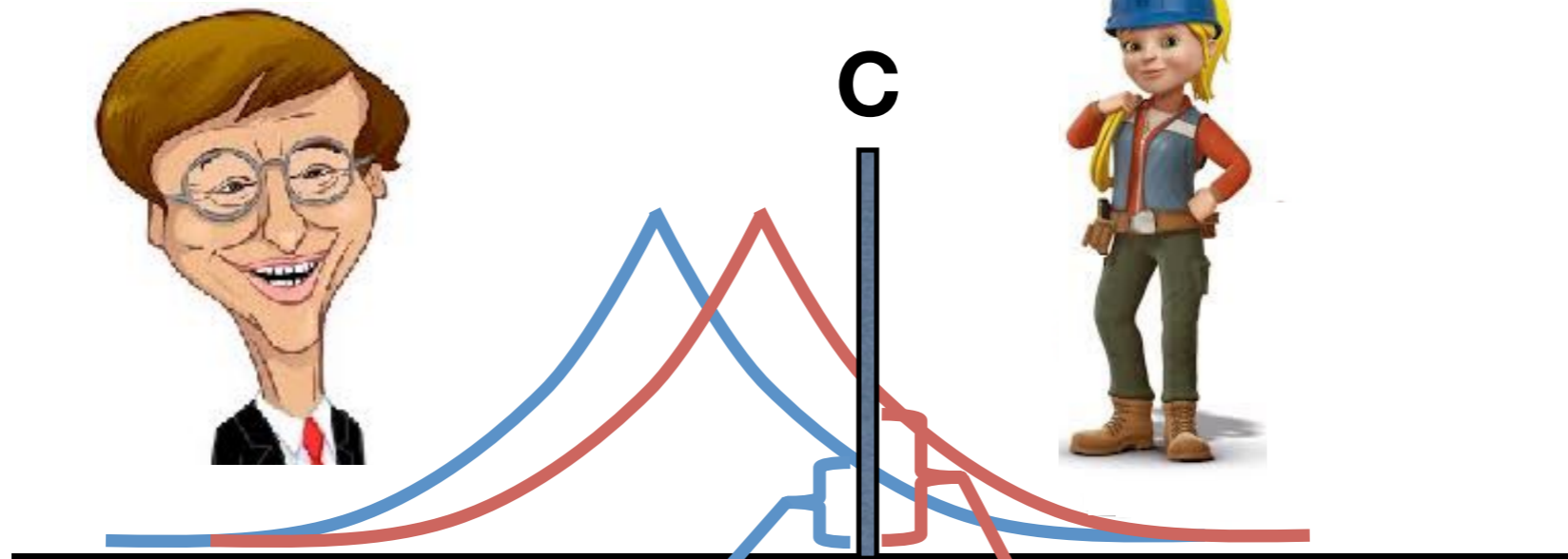
Dataset +



# Differential Privacy

Dataset +

Dataset +



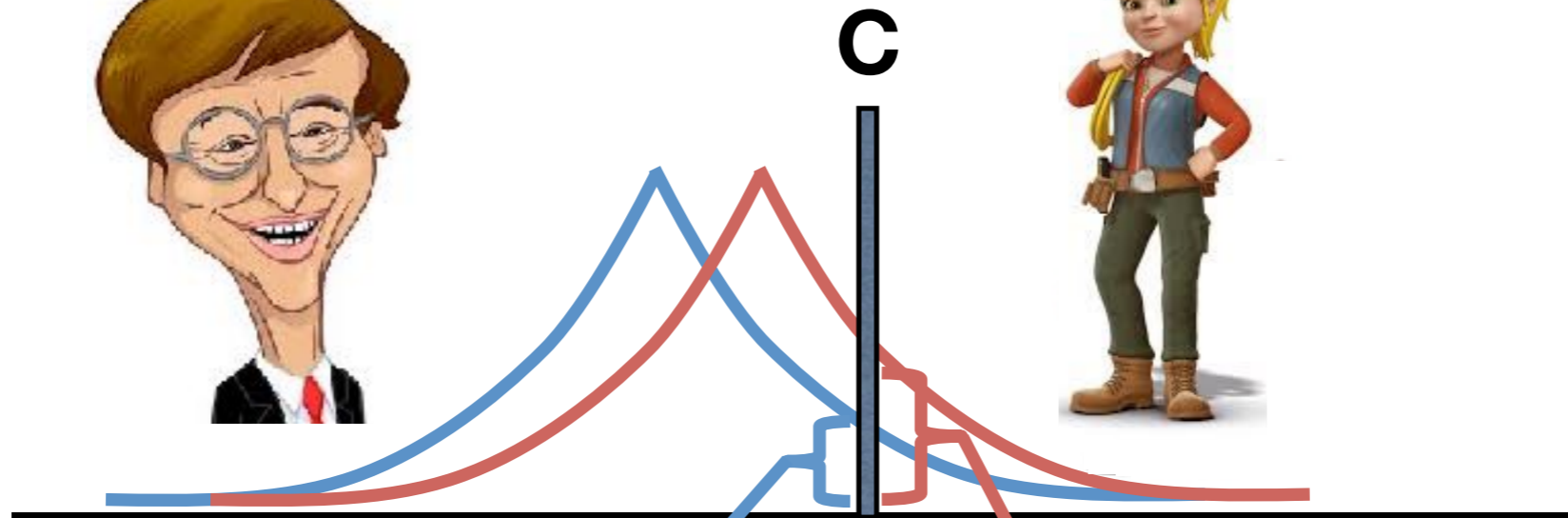
$$P(A(S) \in C) \leq e^\epsilon P(A(S') \in C)$$

# Differential Privacy

Dataset +



Dataset +



$$P(A(S) \in C) \leq e^\epsilon P(A(S') \in C)$$

$\approx$   
1 (for small  $\epsilon$ )



# Obtaining Differential Privacy

# Obtaining Differential Privacy

- Typical mechanism: Add noise to outcome or inside algorithm

# Obtaining Differential Privacy

- Typical mechanism: Add noise to outcome or inside algorithm
- More privacy we want the more noise we add

**Why it works?**

# Why it works?

- Take any arbitrary (possibly deterministic) function  $f(S)$ .

# Why it works?

- Take any arbitrary (possibly deterministic) function  $f(S)$ .
- Say  $B = \max_{S, S'} |f(S) - f(S')|$  where  $S$  and  $S'$  differ on one data point

# Why it works?

- Take any arbitrary (possibly deterministic) function  $f(S)$ .
- Say  $B = \max_{S, S'} |f(S) - f(S')|$  where  $S$  and  $S'$  differ on one data point
- $A(S) = f(S) + B \text{Laplace}(0, 1)/\epsilon$

# Why it works?

- Take any arbitrary (possibly deterministic) function  $f(S)$ .
- Say  $B = \max_{S, S'} |f(S) - f(S')|$  where  $S$  and  $S'$  differ on one data point
- $A(S) = f(S) + B \text{Laplace}(0, 1)/\epsilon$
- $A$  is  $(\epsilon, 0)$ -differentially private



# Why it works?

$$\frac{p(x; f(S))}{p(x; f(S'))}$$

# Why it works?

$$\frac{p(x; f(S))}{p(x; f(S'))} = \frac{e^{-\epsilon|f(S)-x|/B}}{e^{-\epsilon|f(S')-x|/B}}$$

# Why it works?

$$\begin{aligned}\frac{p(x; f(S))}{p(x; f(S'))} &= \frac{e^{-\epsilon|f(S)-x|/B}}{e^{-\epsilon|f(S')-x|/B}} \\ &= e^{\epsilon \frac{|f(S')-x| - |f(S)-x|}{B}}\end{aligned}$$

# Why it works?

$$\begin{aligned}\frac{p(x; f(S))}{p(x; f(S'))} &= \frac{e^{-\epsilon|f(S)-x|/B}}{e^{-\epsilon|f(S')-x|/B}} \\ &= e^{\epsilon \frac{|f(S')-x| - |f(S)-x|}{B}} \\ &\leq e^{\epsilon \frac{|f(S')-f(S)|}{B}}\end{aligned}$$

# Why it works?

$$\begin{aligned}\frac{p(x; f(S))}{p(x; f(S'))} &= \frac{e^{-\epsilon|f(S)-x|/B}}{e^{-\epsilon|f(S')-x|/B}} \\ &= e^{\epsilon \frac{|f(S')-x| - |f(S)-x|}{B}} \\ &\leq e^{\epsilon \frac{|f(S')-f(S)|}{B}} \\ &\leq e^{\epsilon}\end{aligned}$$

# The Magic Broker Scam

Based on Aaron Roth's slide

# The Magic Broker Scam

Based on Aaron Roth's slide

Day 1



Tomorrow Stock X



# The Magic Broker Scam

Based on Aaron Roth's slide

Day 1



Tomorrow Stock X



Day 2



Tomorrow Stock X





# The Magic Broker Scam

Based on Aaron Roth's slide

Day 1



Tomorrow Stock X



Day 2



Tomorrow Stock X



Day 3



Tomorrow Stock X



# The Magic Broker Scam

Based on Aaron Roth's slide

Day 1



Tomorrow Stock X



Day 2



Tomorrow Stock X



Day 3



Tomorrow Stock X



.....

Day 10

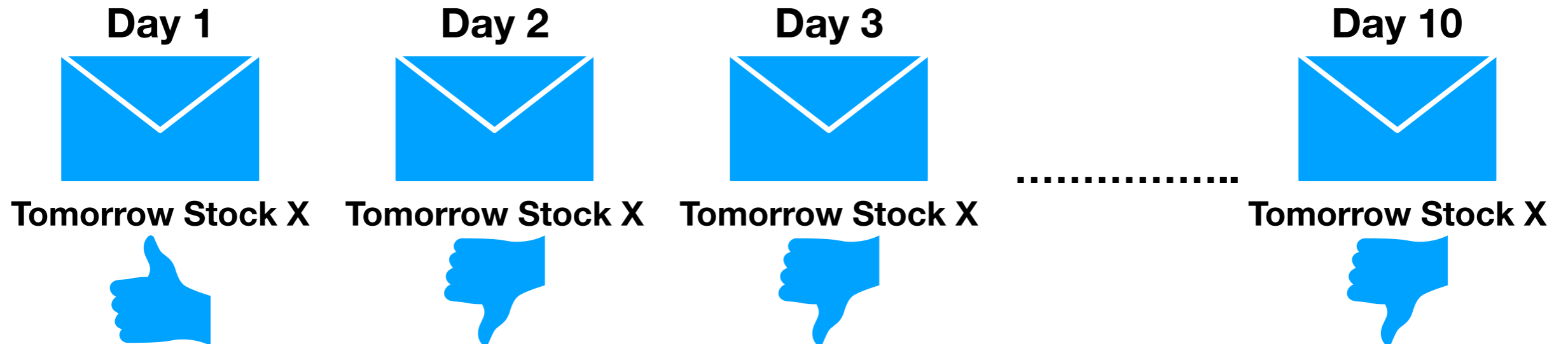


Tomorrow Stock X



# The Magic Broker Scam

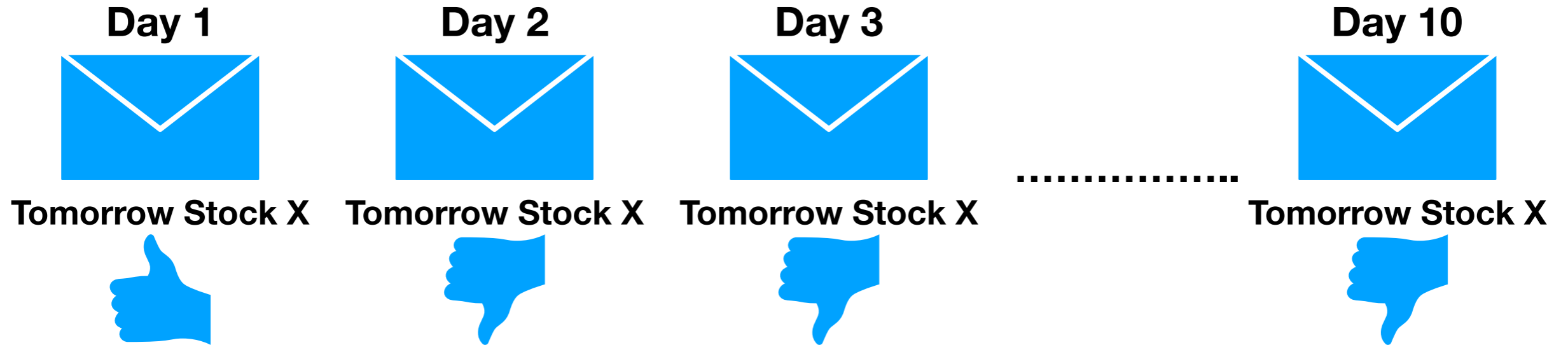
Based on Aaron Roth's slide



- Day 11: Pay me I will help you invest

# The Magic Broker Scam

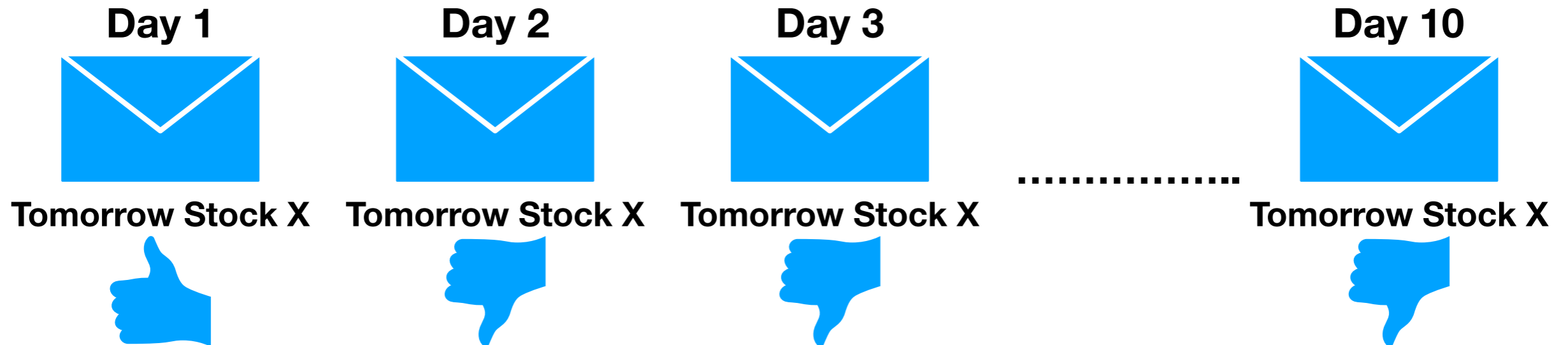
Based on Aaron Roth's slide



- Day 11: Pay me I will help you invest
- What are the chances that some one guesses randomly and gets correct?

# The Magic Broker Scam

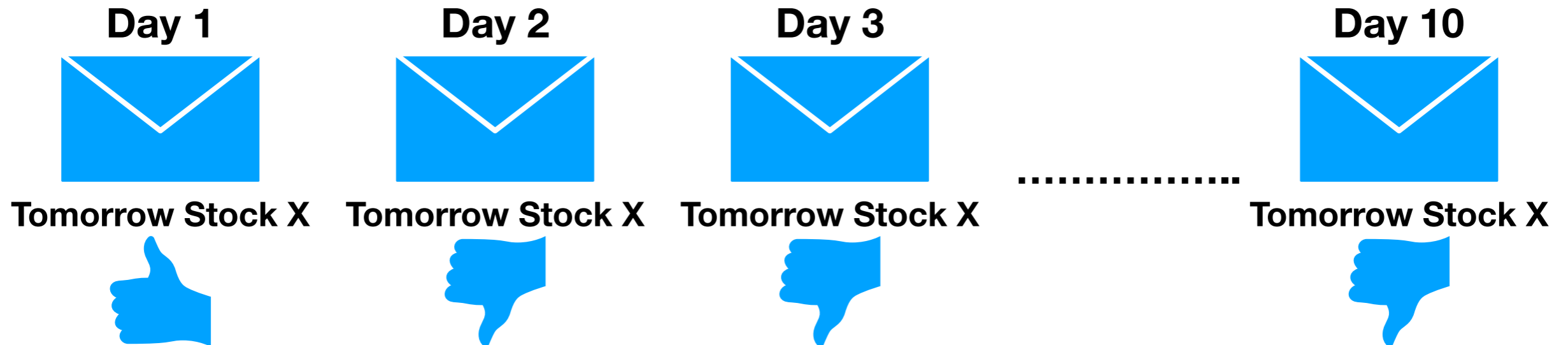
Based on Aaron Roth's slide



- Day 11: Pay me I will help you invest
- What are the chances that some one guesses randomly and gets correct?
  - 1 in 1024

# The Magic Broker Scam

Based on Aaron Roth's slide



- Day 11: Pay me I will help you invest
- What are the chances that some one guesses randomly and gets correct?
  - 1 in 1024
- But the broker will always scam people, why?

# The Magic Broker Scam

Based on Aaron Roth's slide

Day 1



Tomorrow Stock X



Day 2



Tomorrow Stock X



Day 3



Tomorrow Stock X



.....

Day 10

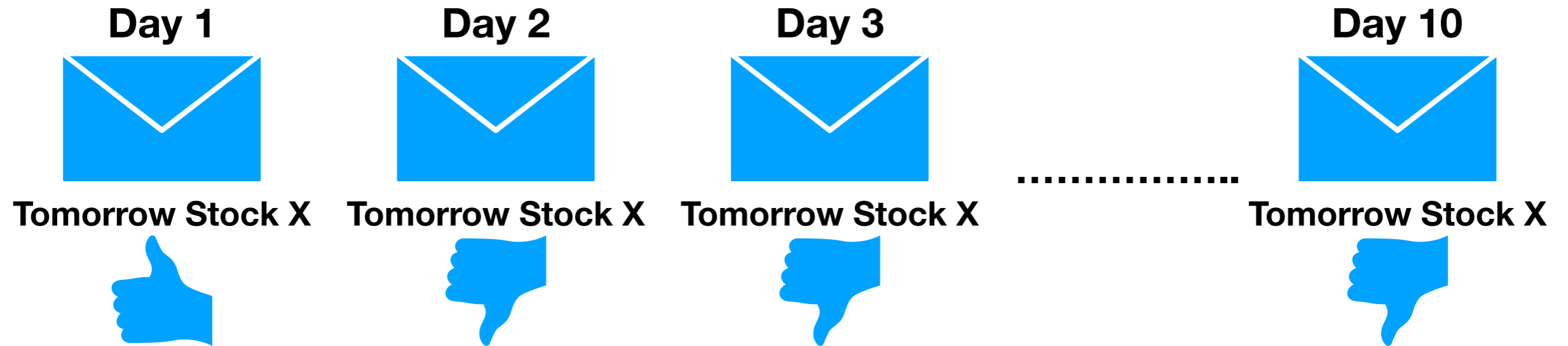


Tomorrow Stock X



# The Magic Broker Scam

Based on Aaron Roth's slide

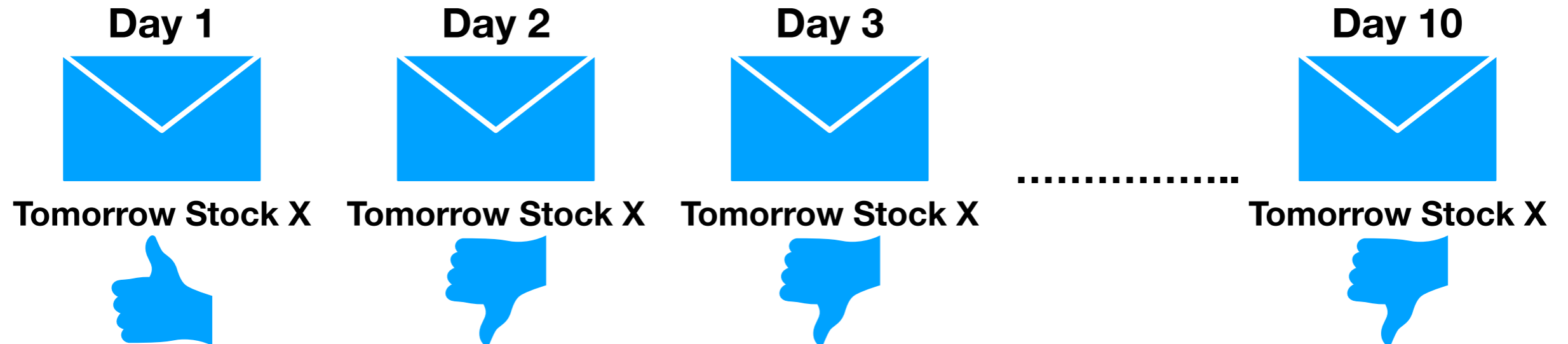


- Day 1: send email to a million people ( $1/2$  + and  $1/2$  -)



# The Magic Broker Scam

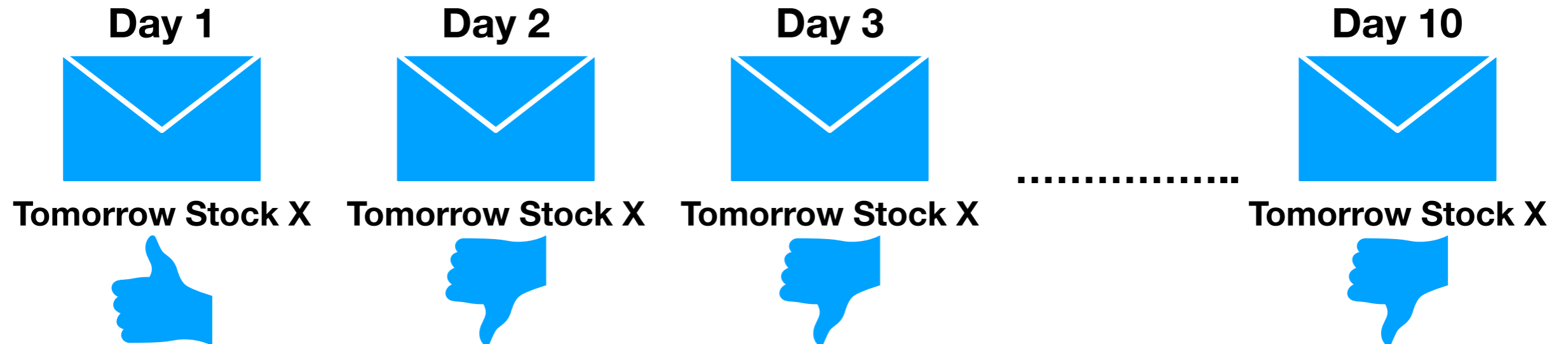
Based on Aaron Roth's slide



- Day 1: send email to a million people ( $1/2 +$  and  $1/2 -$ )
- Day 2: only send email to 500,000 people we got right ( $1/2 +$  and  $1/2 -$ )

# The Magic Broker Scam

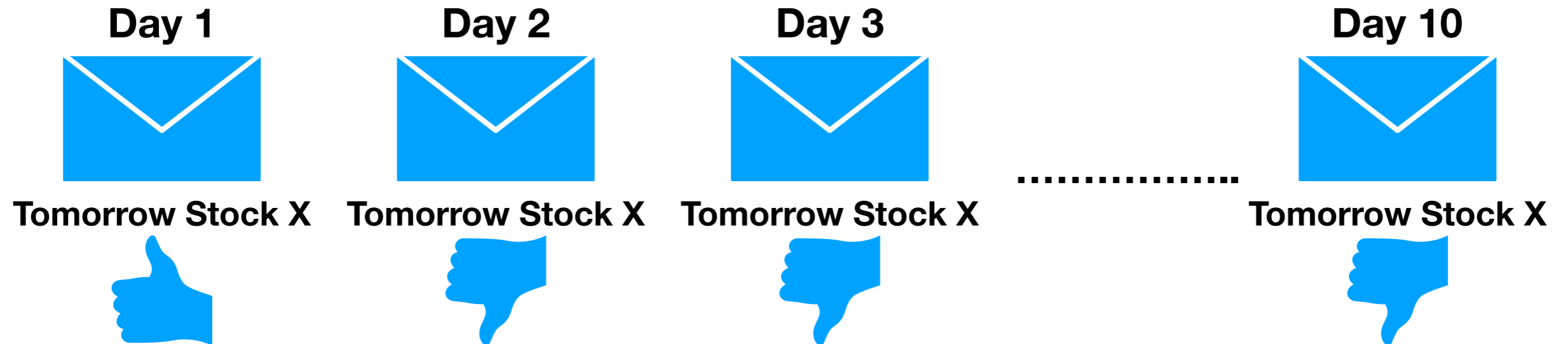
Based on Aaron Roth's slide



- Day 1: send email to a million people ( $1/2 +$  and  $1/2 -$ )
- Day 2: only send email to 500,000 people we got right ( $1/2 +$  and  $1/2 -$ )
- .....

# The Magic Broker Scam

Based on Aaron Roth's slide



- Day 1: send email to a million people ( $1/2$  + and  $1/2$  -)
- Day 2: only send email to 500,000 people we got right ( $1/2$  + and  $1/2$  -)
- .....
- Day 11: We are left with  $1000000/1024 \sim 1000$  people we can scam

# Reproducibility problem



reproducibility in biology



All

**News**

Images

Videos

Shopping

More

Settings

Tools

About 6,890 results (0.21 seconds)

## **Biological Variability Makes Reproducibility More Difficult**

**Lab Manager Magazine** - Mar 26, 2019

**Biological Variability Makes Reproducibility More Difficult** ... the question of the **reproducibility** of scientific data—an important topic that comes ...



## **Current Practices May Not Fix Reproducibility Crisis in Research**

**Laboratory Equipment** - Apr 18, 2019

... a coin toss, suggests a provocative new study published in **PLOS Biology**. ... This issue, known as the **reproducibility crisis**, has led to many ...

## **Engineering Meets Biology**

**Genetic Engineering & Biotechnology News** - Apr 1, 2019

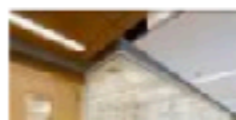
A start-to-end assembly in **biological** products, however, is far more challenging. **Biology** suffers from the **reproducibility crisis** because of the ...



## **Can flipping coins replace animal experiments?**

**Phys.Org** - Apr 9, 2019

... their paper publishing April 9 in the open-access journal **PLOS Biology**. ... while increasing the robustness and **reproducibility** of their results.



## **Core Labs Can Help Combat Issue of Research Irreproducibility ...**

**GenomeWeb** - Mar 26, 2019

... Core labs can help address the issue of research **reproducibility**. ... published in

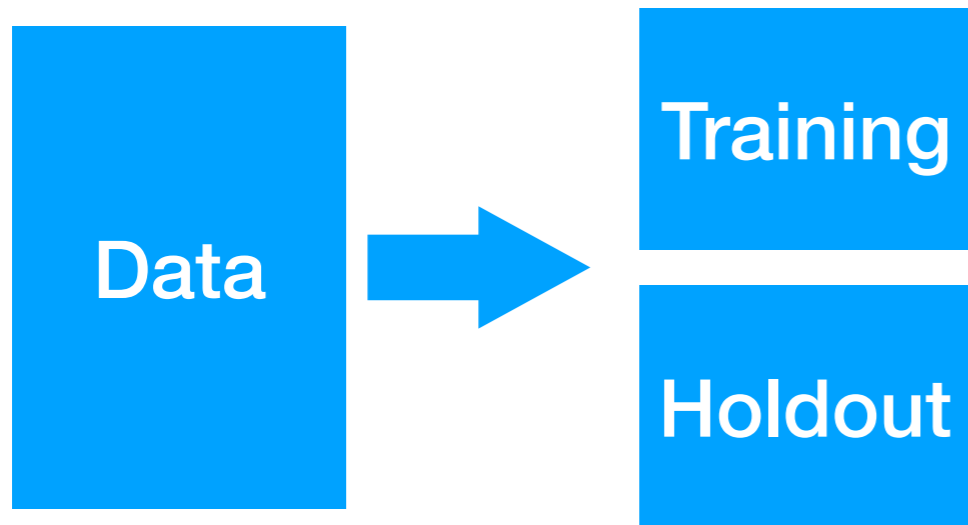
# Ideal ML Experiment

# Ideal ML Experiment

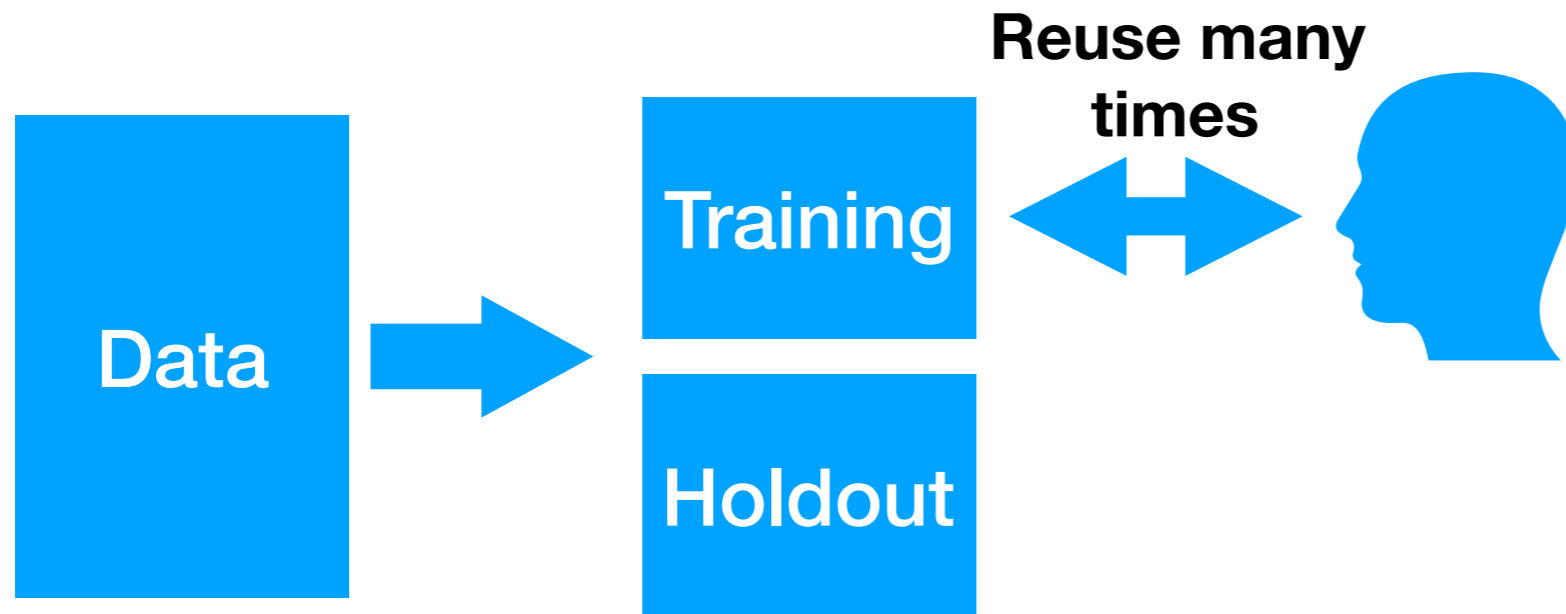


Data

# Ideal ML Experiment

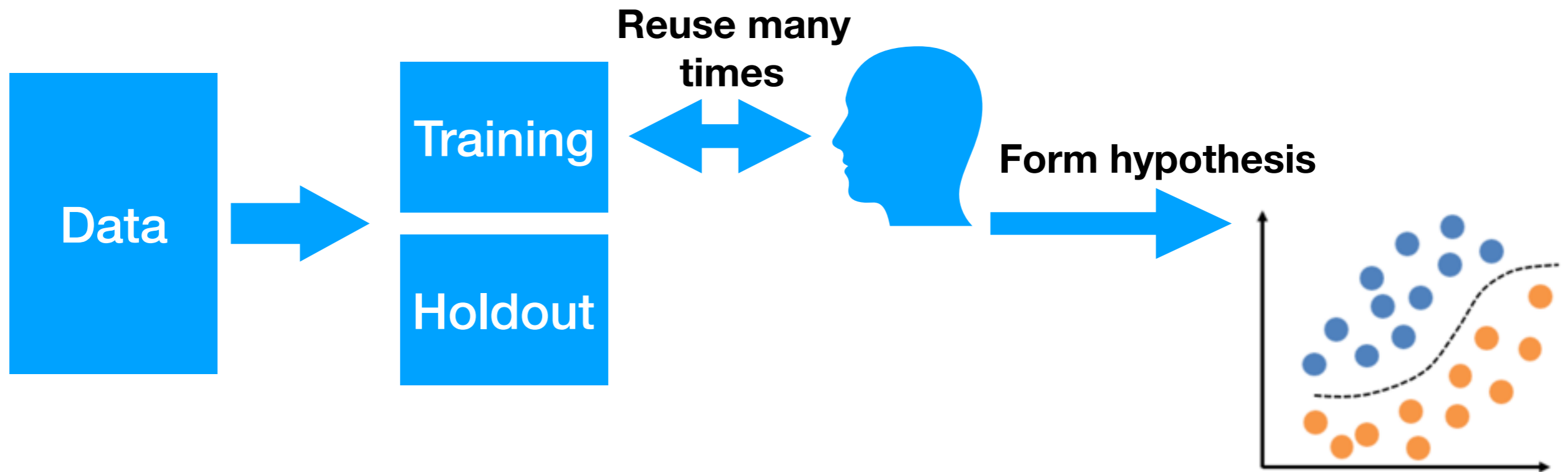


# Ideal ML Experiment

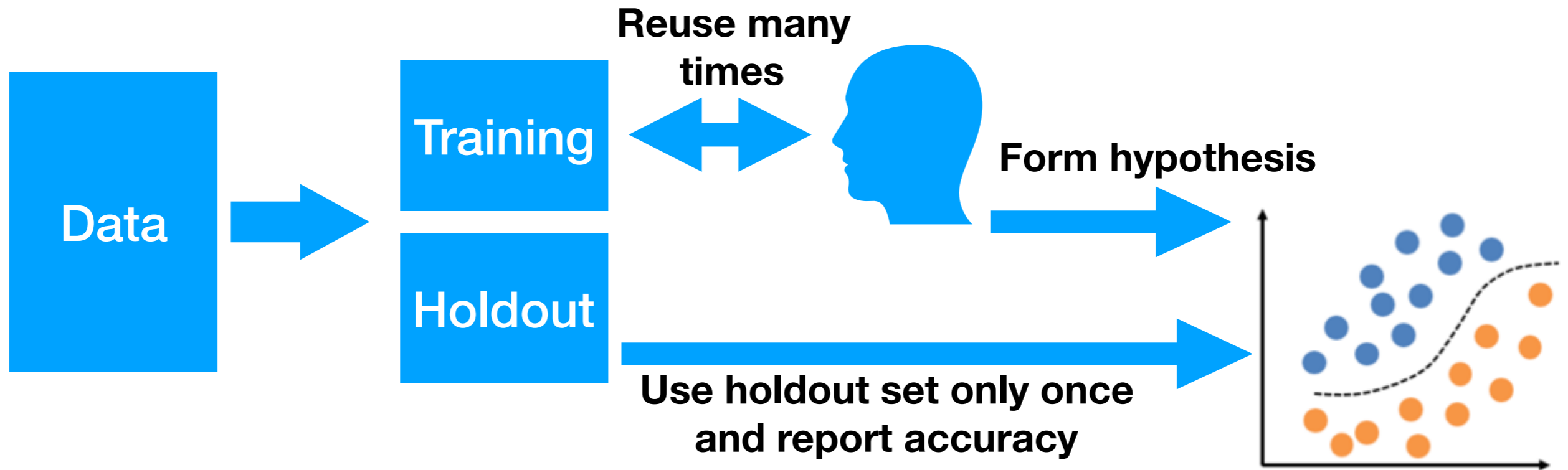




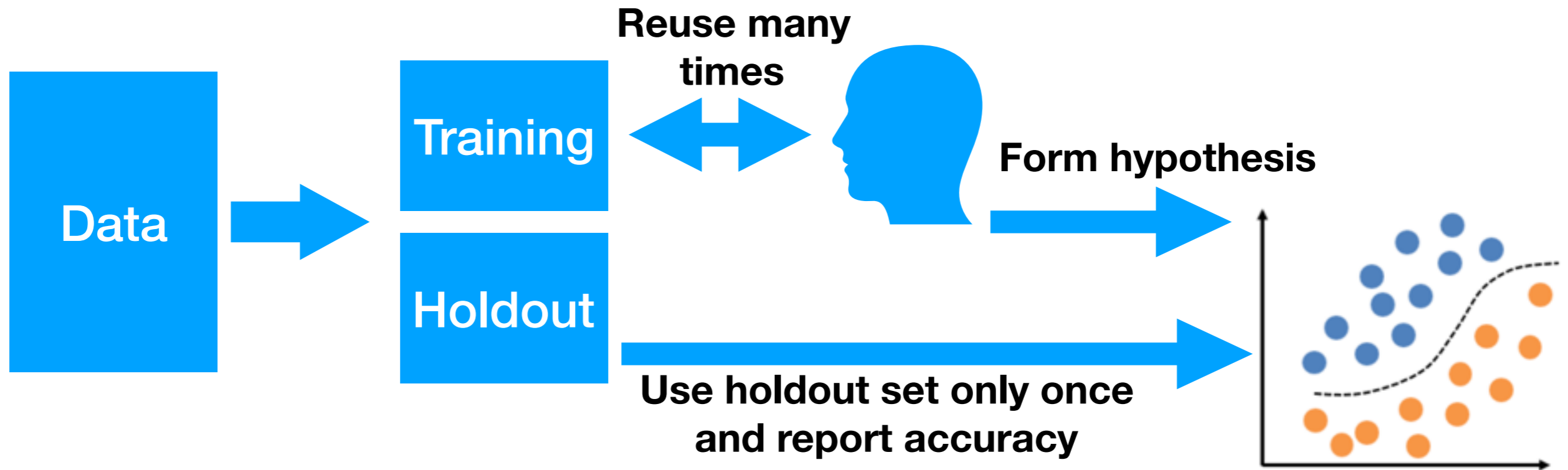
# Ideal ML Experiment



# Ideal ML Experiment

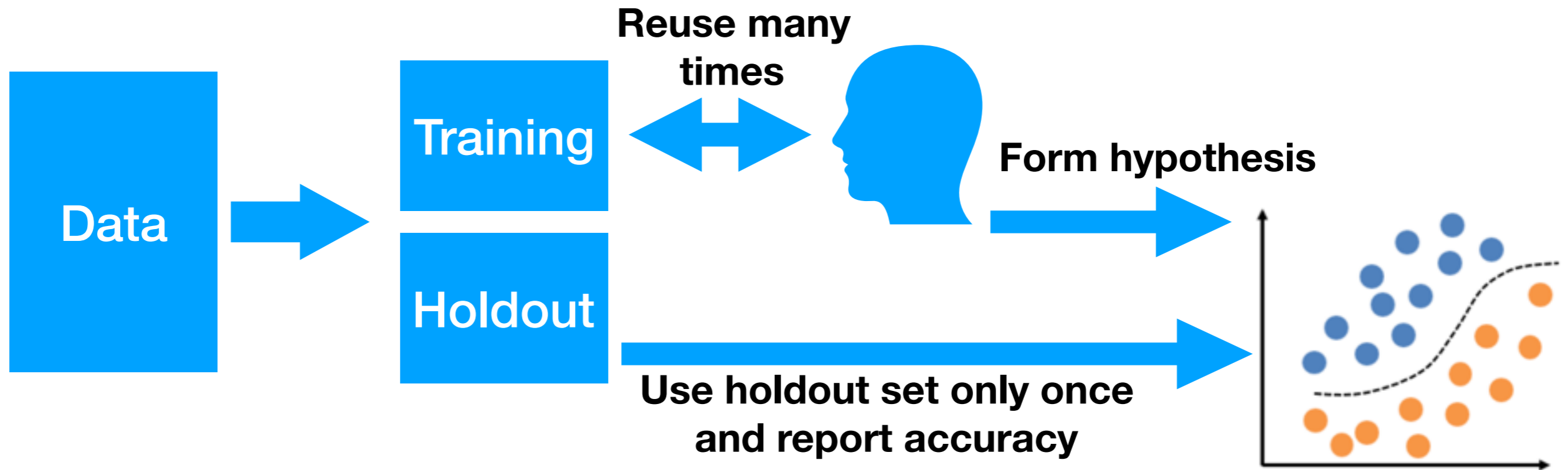


# Ideal ML Experiment



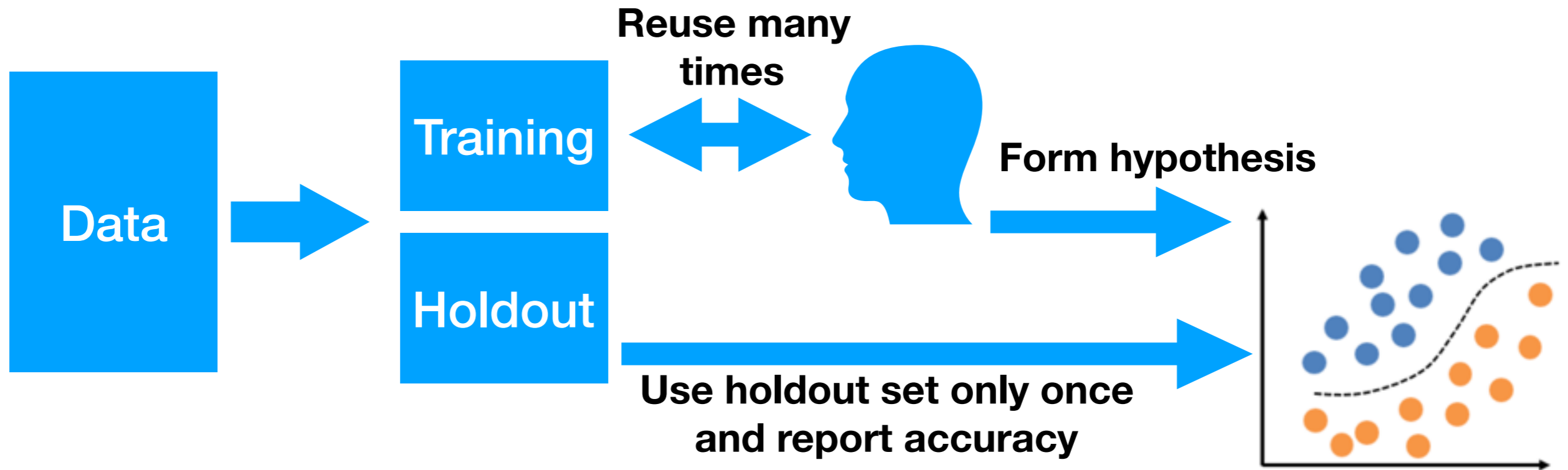
- We are not allowed to form hypothesis based on data we used to test: age old statistics

# Ideal ML Experiment



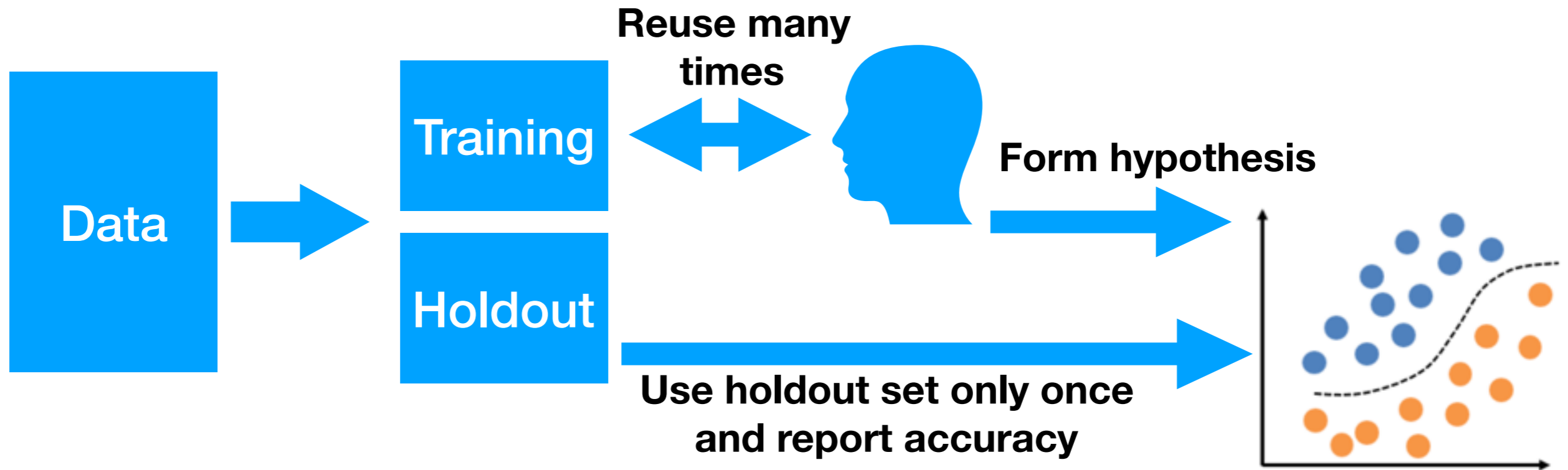
- We are not allowed to form hypothesis based on data we used to test: age old statistics
- But too tempting to form more informed opinion

# Ideal ML Experiment



- We are not allowed to form hypothesis based on data we used to test: age old statistics
- But too tempting to form more informed opinion
- We do a train/validation/test set split

# Ideal ML Experiment



- We are not allowed to form hypothesis based on data we used to test: age old statistics
- But too tempting to form more informed opinion
- We do a train/validation/test set split
- But this means we can't reuse datasets over time

# ML Today

# ML Today

- Benchmark dataset from MNIST to IMAGENET



# ML Today

- Benchmark dataset from MNIST to IMAGENET
- Competitions run with feedback from test set to competitors ;)

# ML Today

- Benchmark dataset from MNIST to IMAGENET
- Competitions run with feedback from test set to competitors ;)
- Effort to collect public dataset to share....

# ML Today

- Benchmark dataset from MNIST to IMAGENET
- Competitions run with feedback from test set to competitors ;)
- Effort to collect public dataset to share....
- All great but we need to be careful!

# ML Today

- Benchmark dataset from MNIST to IMAGENET
- Competitions run with feedback from test set to competitors ;)
- Effort to collect public dataset to share....
- All great but we need to be careful!
  - Old fix: pre-register experiment, many many examples of people fudging this....

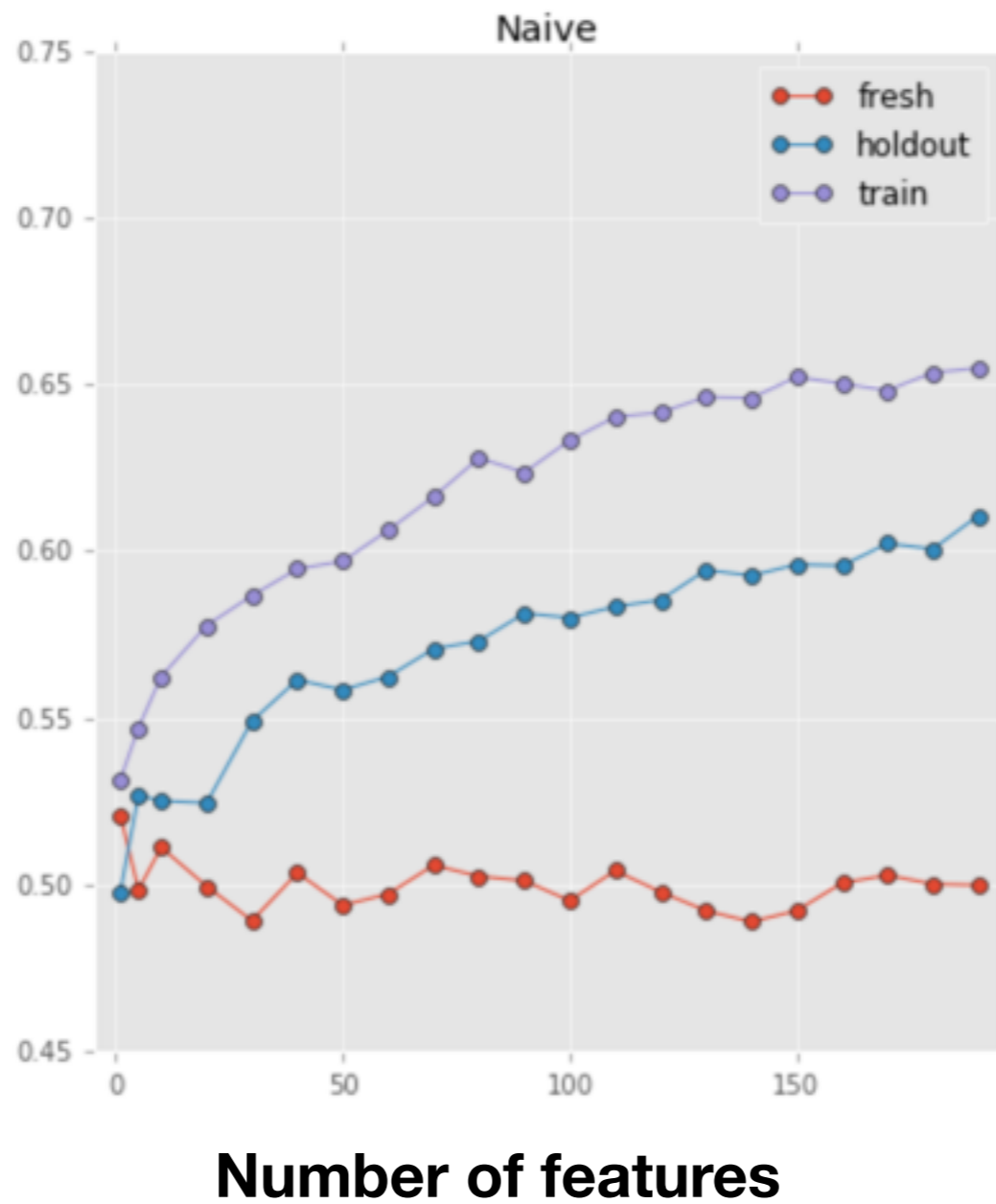
# An Example

Code at:

<https://github.com/isofer/thresholdout-experiments/blob/master/Thresholdout%20experiments.ipynb>

- Data: 20,000 data points in 10000 dimensions drawn randomly
- Labels: random +1 or -1 (no correlation with input)
- Data-scientist runs:
  - Split data into train and test of equal size
  - Select best k features on training data
  - Only use variables also good on holdout
  - Build linear predictor out of these k variables
  - Find best k = 10,20,30,40,.....

# An Example



**Can we reuse holdout set?**

**Maybe if we do it right...**



# Maybe if we do it right...

- Idea: when we report back accuracies from the dataset, we add noise so as to not to leak too much information

# Maybe if we do it right...

- Idea: when we report back accuracies from the dataset, we add noise so as to not to leak too much information
- Eg. Report back accuracies on holdout set only when training and test accuracy are significantly different

# Threshold-out Algorithm

**Input:**

Data  $S$ , holdout  $H$ , threshold  $T > 0$ , tolerance  $\sigma > 0$

Given function  $q$ :

Sample  $\eta, \eta'$  from  $N(0, \sigma^2)$

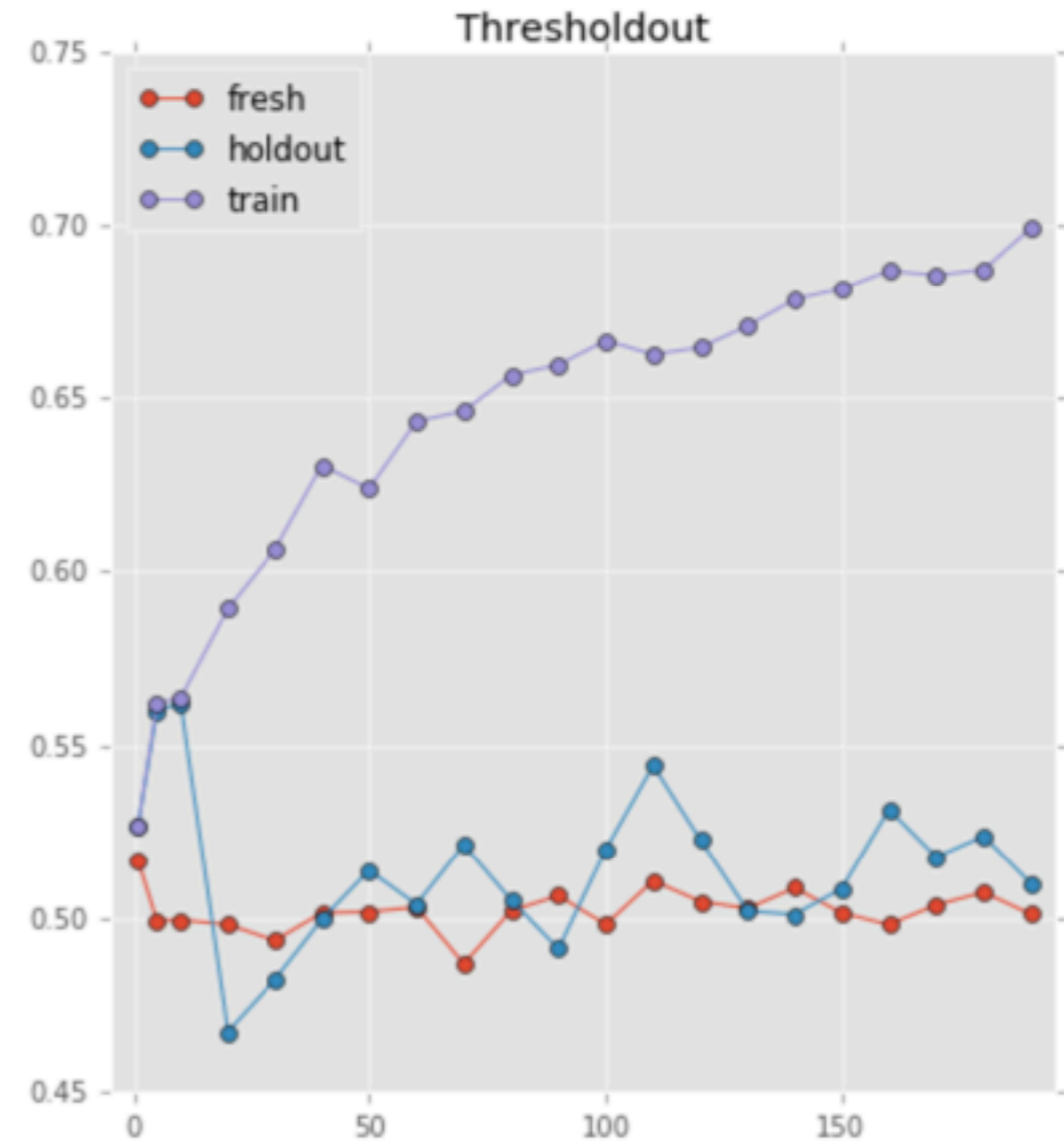
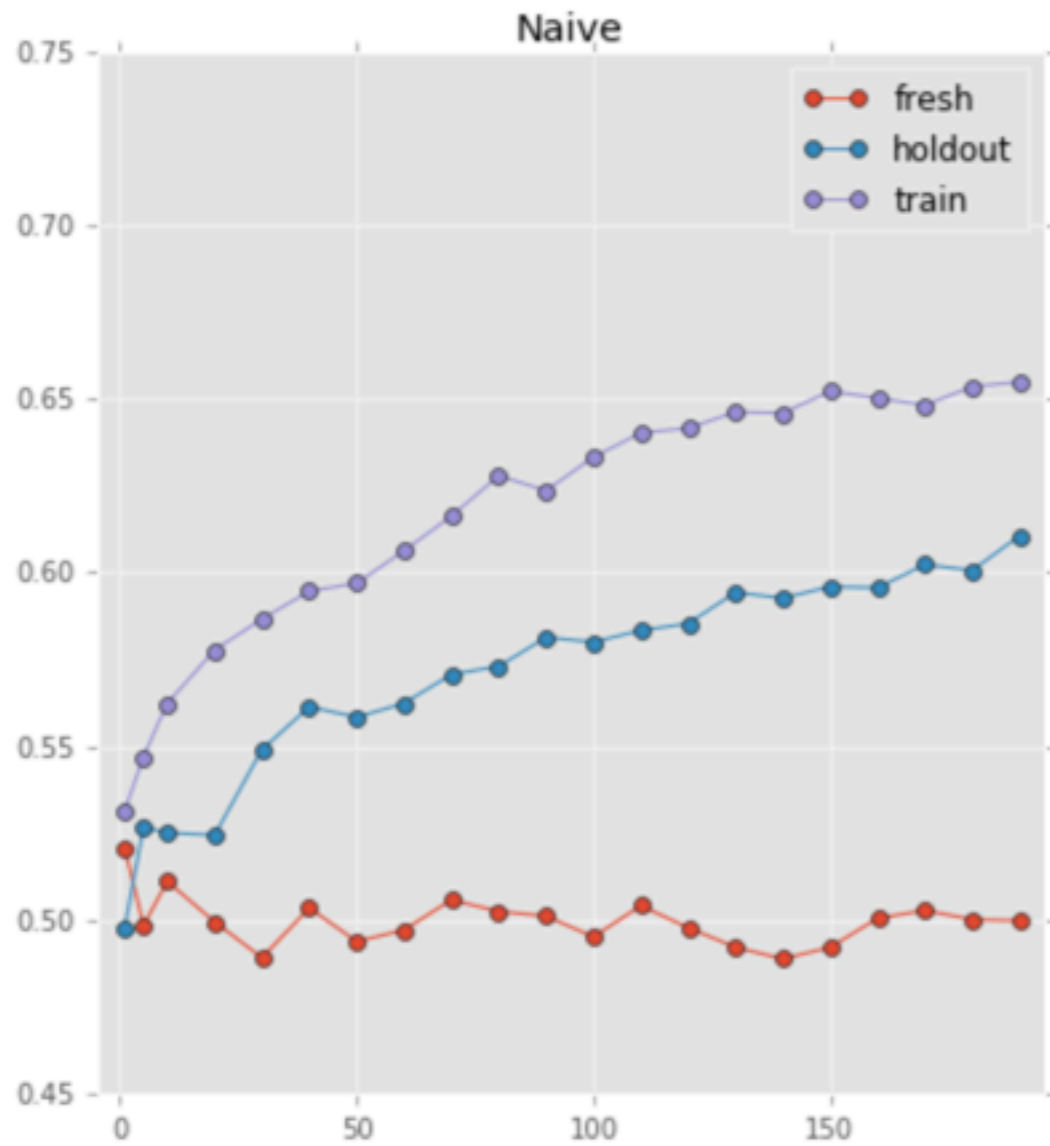
If  $|\text{avg}_H[q] - \text{avg}_S[q]| > T + \eta$ :

output  $\text{avg}_H[q] + \eta'$

Otherwise:

output  $\text{avg}_S[q]$

# The Example



# The Guarantee

- Roughly: Thresholdout guarantees high accuracy, and can handle up to *asked is order  $n^2$  over-fitting queries*

# Differential Privacy

# Differential Privacy

- A deterministic algorithm cannot preserve privacy

# Differential Privacy

- A deterministic algorithm cannot preserve privacy
- Say  $S = (\text{Data}_1, \dots, \text{Data}_n)$  is the data provided to learning algorithm (be it clustering, supervised learning etc).



# Differential Privacy

- A deterministic algorithm cannot preserve privacy
- Say  $S = (\text{Data}_1, \dots, \text{Data}_n)$  is the data provided to learning algorithm (be it clustering, supervised learning etc).
- Say (randomized) learning algorithm  $A$  takes this training data and returns solution as  $A(S)$

# Differential Privacy

- A deterministic algorithm cannot preserve privacy
- Say  $S = (\text{Data}_1, \dots, \text{Data}_n)$  is the data provided to learning algorithm (be it clustering, supervised learning etc).
- Say (randomized) learning algorithm  $A$  takes this training data and returns solution as  $A(S)$
- Algorithm  $A$  is  $(\epsilon, \delta)$ - differentially private if for all samples  $S$  and  $S'$  that only differ by one data point and any set  $C$

$$P(A(S) \in C) \leq e^\epsilon P(A(S') \in C) + \delta$$

# Differential Privacy

- A deterministic algorithm cannot preserve privacy
- Say  $S = (\text{Data}_1, \dots, \text{Data}_n)$  is the data provided to learning algorithm (be it clustering, supervised learning etc).
- Say (randomized) learning algorithm  $A$  takes this training data and returns solution as  $A(S)$
- Algorithm  $A$  is  $(\epsilon, \delta)$ - differentially private if for all samples  $S$  and  $S'$  that only differ by one data point and any set  $C$

$$P(A(S) \in C) \leq e^\epsilon P(A(S') \in C) + \delta$$

- $\delta=0$  is called pure differential privacy

# Differential Privacy

- Post-processing: If  $A$  is a differentially private algorithm, and  $f$  is any mapping on the outcome space of  $A$ , then the algorithm that maps from sample  $S$  to  $f(A(S))$  is also differentially private
- Composability:  $\text{Algo}(S)$  is given by following procedure
  - For  $i$  in 1 to  $k$ 
    - If we choose any epsilon differentially private algorithm  $A_i$  based on outcomes  $O_1, \dots, O_{\{i-1\}}$
    - Set  $O_i = A_i(S)$
  - Return  $O_1, \dots, O_n$
- The above algorithm is  $O(\epsilon \sqrt{k})$  differentially private