

# Introduction to Interpretable Machine Learning

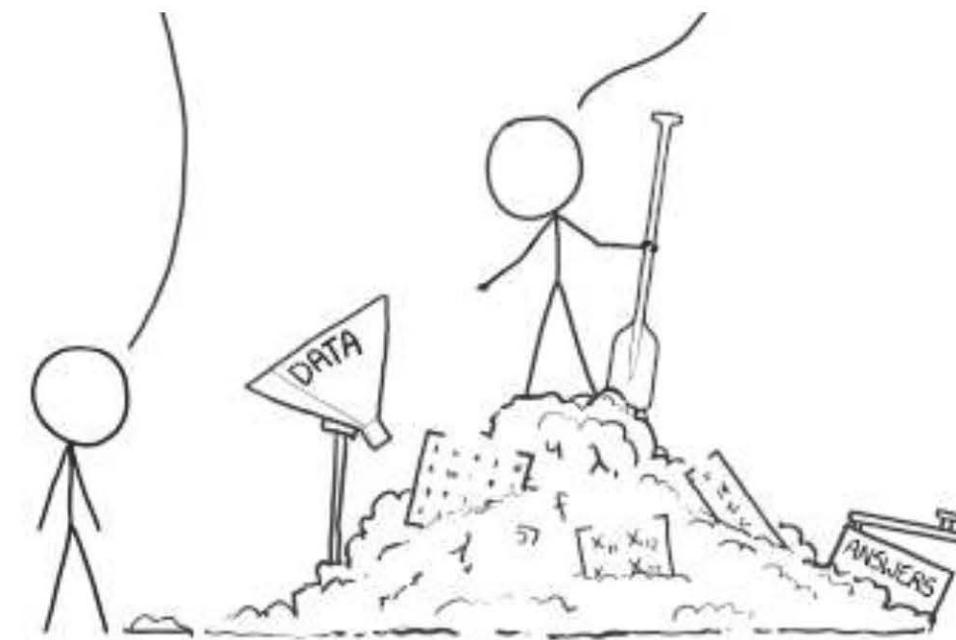


credit:<https://s-media-cache-ak0.pinimg.com>

Been Kim  
Google Brain  
Deep Learning Summer School 2018  
@Vector institute



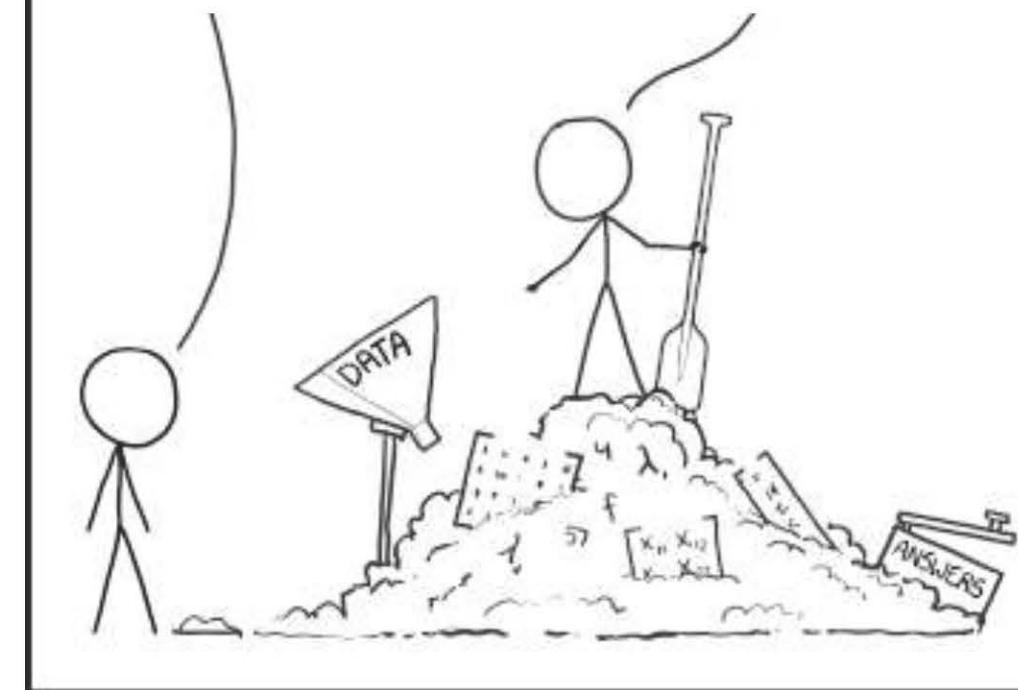
THIS IS YOUR MACHINE LEARNING SYSTEM?



<https://xkcd.com/>

THIS IS YOUR MACHINE LEARNING SYSTEM?

YUP! YOU POUR THE DATA INTO THIS BIG  
PILE OF LINEAR ALGEBRA, THEN COLLECT  
THE ANSWERS ON THE OTHER SIDE.

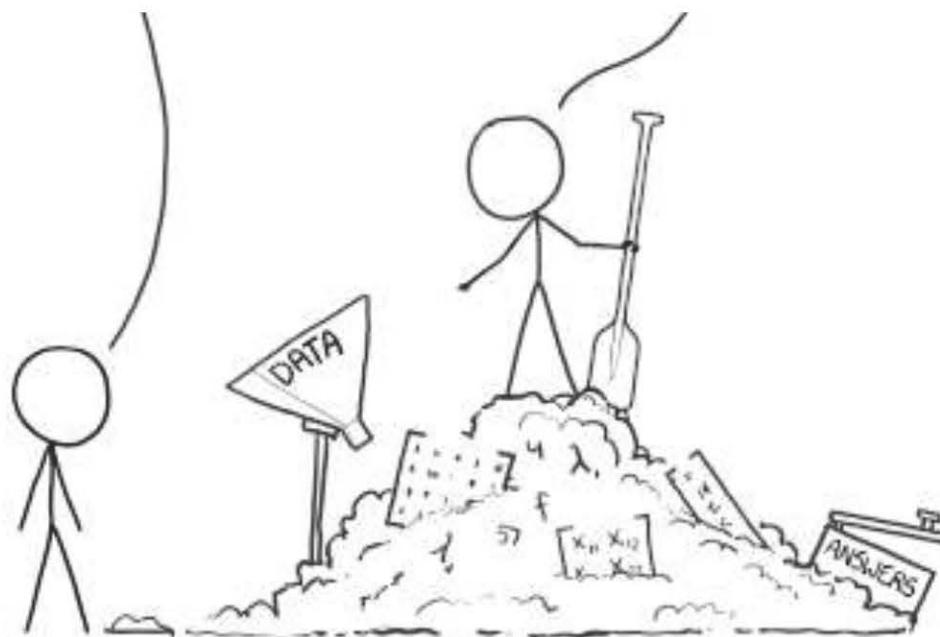


<https://xkcd.com/>

THIS IS YOUR MACHINE LEARNING SYSTEM?

| YUP! YOU POUR THE DATA INTO THIS BIG  
PILE OF LINEAR ALGEBRA, THEN COLLECT  
THE ANSWERS ON THE OTHER SIDE.

WHAT IF THE ANSWERS ARE WRONG? |

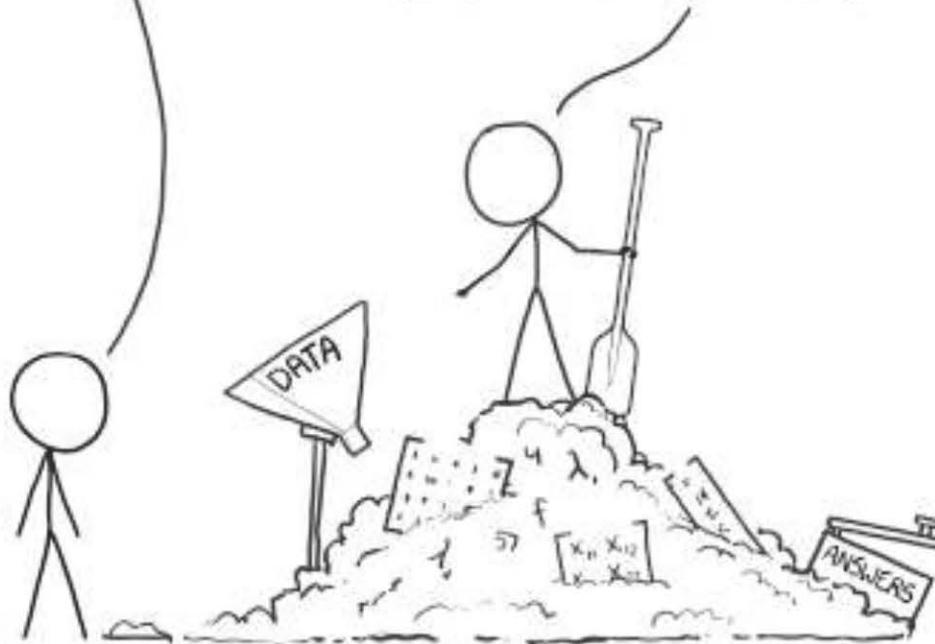


THIS IS YOUR MACHINE LEARNING SYSTEM?

YUP! YOU POUR THE DATA INTO THIS BIG  
PILE OF LINEAR ALGEBRA, THEN COLLECT  
THE ANSWERS ON THE OTHER SIDE.

WHAT IF THE ANSWERS ARE WRONG?

JUST STIR THE PILE UNTIL  
THEY START LOOKING RIGHT.



THIS IS YOUR MACHINE LEARNING SYSTEM?

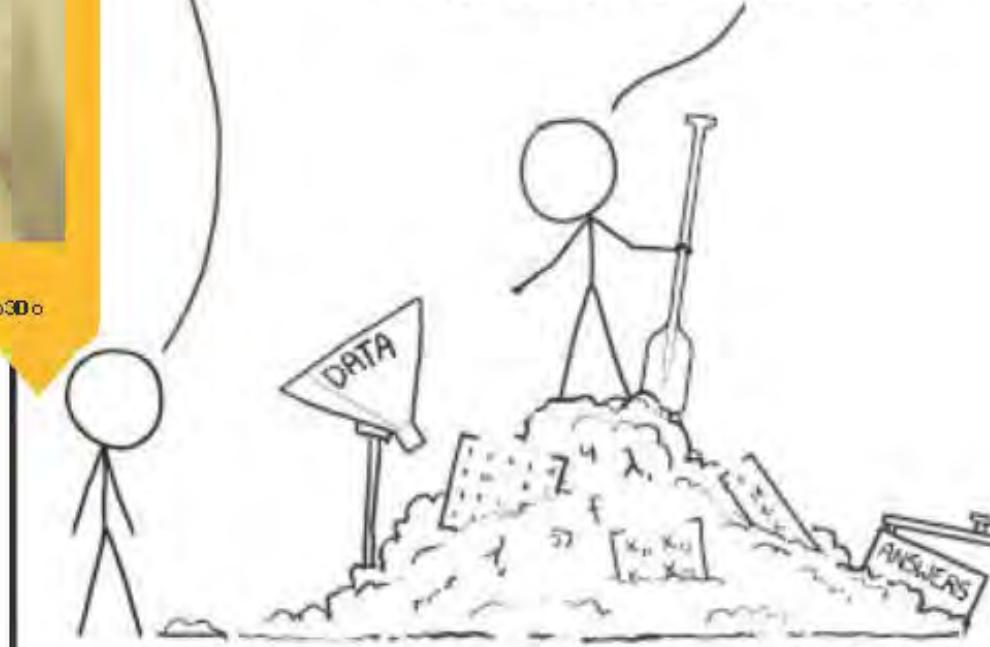
YUP! YOU POUR THE DATA INTO THIS BIG  
PILE OF LINEAR ALGEBRA, THEN COLLECT  
THE ANSWERS ON THE OTHER SIDE.

WHAT IF THE ANSWERS ARE WRONG?

JUST STIR THE PILE UNTIL  
THEY START LOOKING RIGHT.

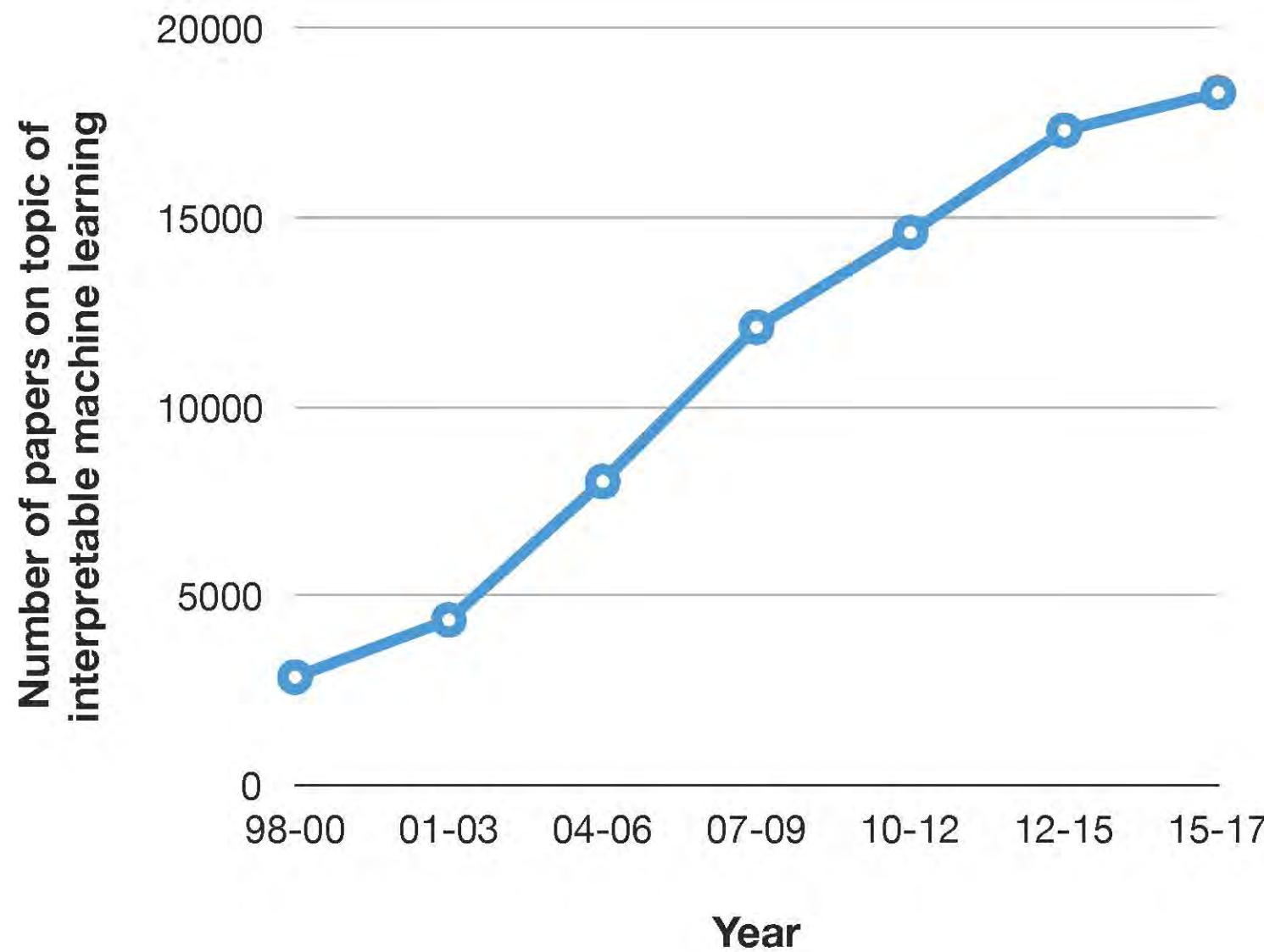


<https://www.youtube.com/watch?v=icqDxNab3Do>



<https://xkcd.com/>

# ML community is responding



This is not a new problem.  
Why now?

Complexity and prevalence!



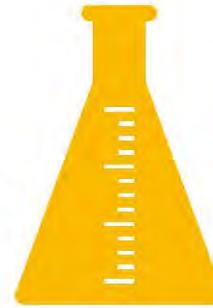
I heard you can just use  
decision trees...

Can we go home now?

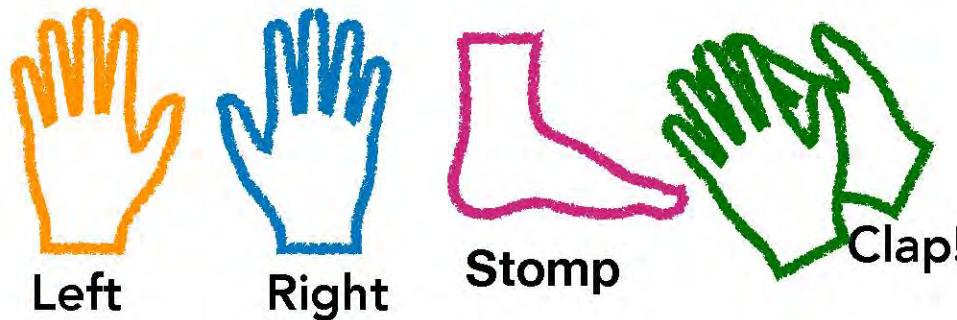


<http://www.logroup.com.au/raise-your-hand-when-you-should-and-why-you-should/>

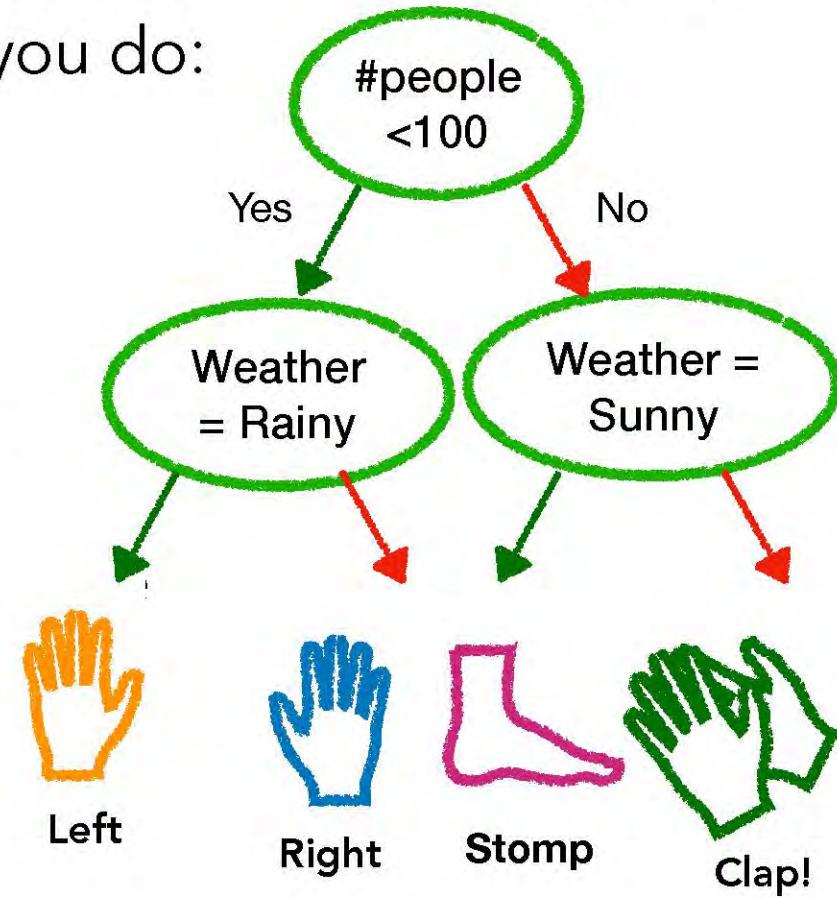
# Experiment.



- I will show you a decision tree. Follow the right path given a data point, and you do:



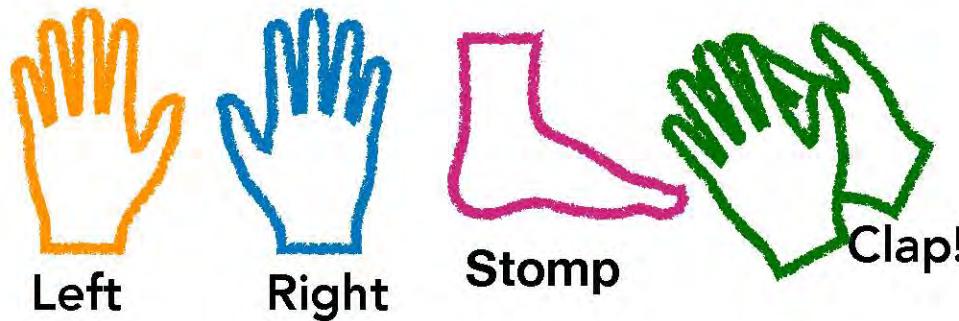
**Data = [Sunny, 200]**



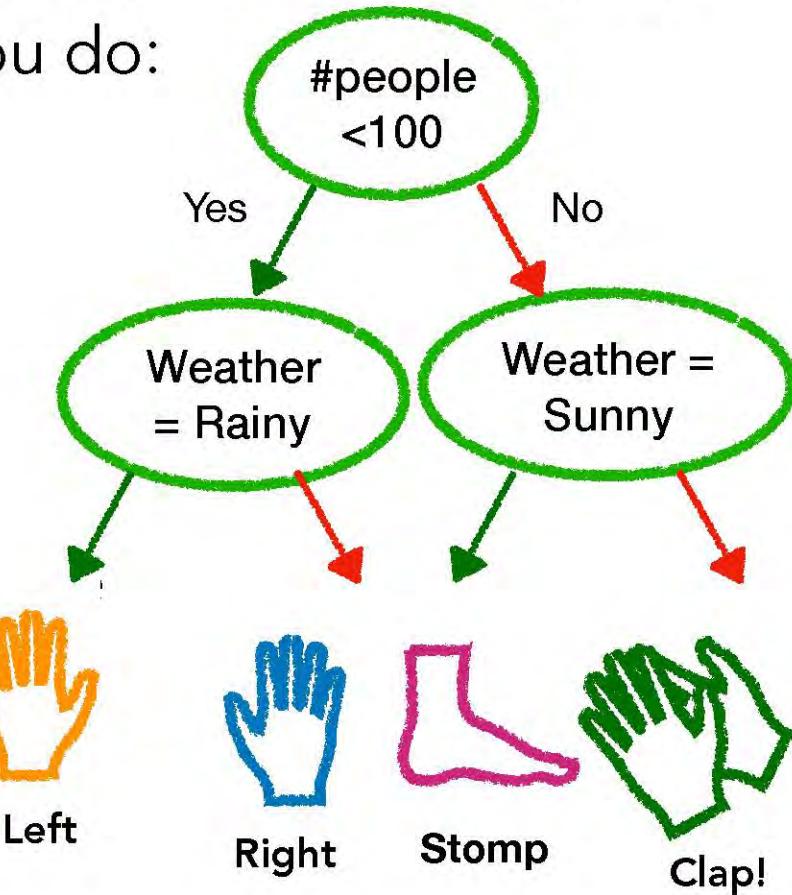
# Experiment.



- I will show you a decision tree. Follow the right path given a data point, and you do:

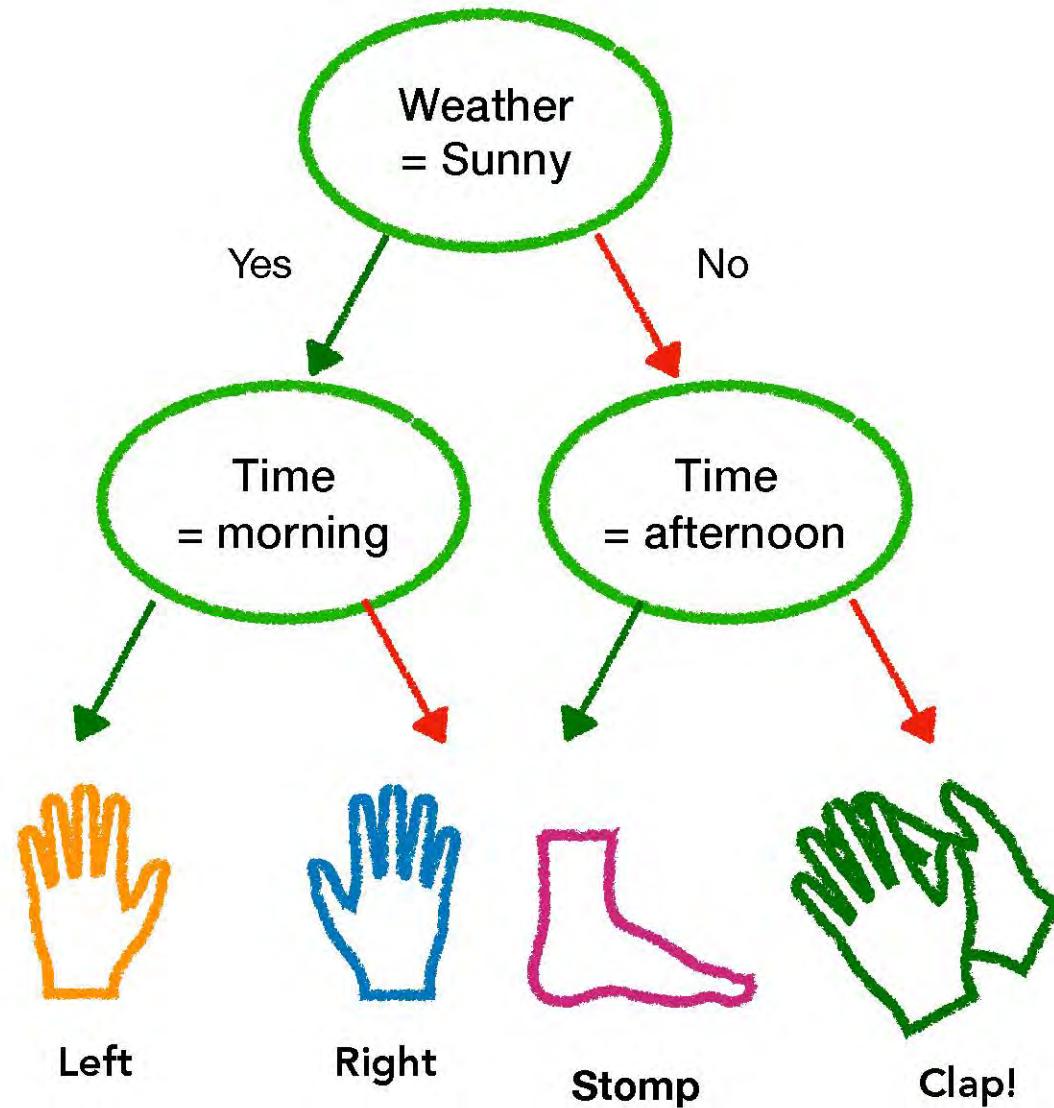


**Data = [Sunny, 200]**

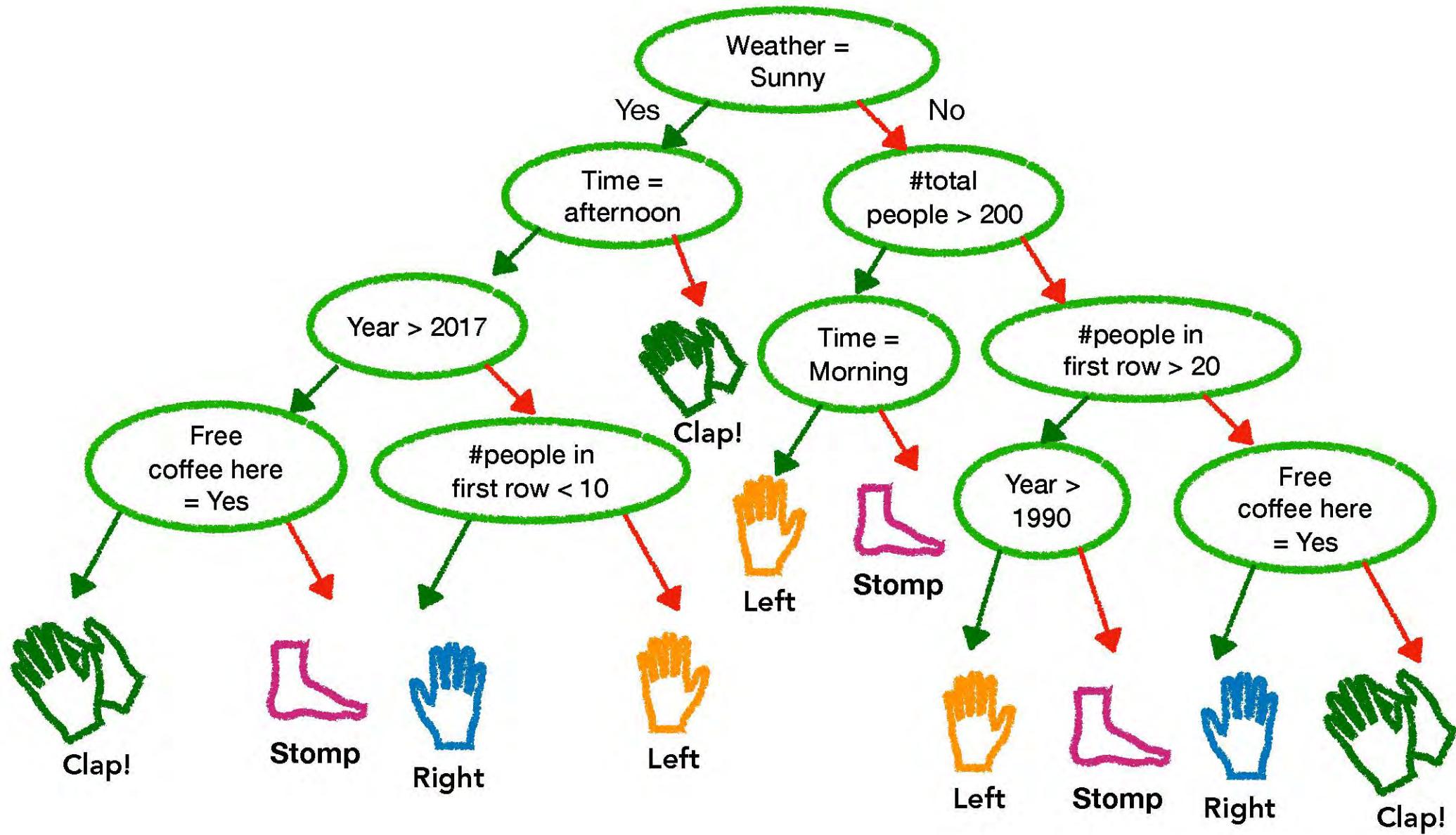


- As soon as you know the answer, do the action!

# Sample decision tree #1



# Sample decision tree #2



# Sample decision tree #3



# Sample decision tree #3

And can you explain what the overall logic of the system was?

If I give you a lot of data points, can you guess which feature was most 'important' (i.e., used in more number of examples)?



 Common misunderstanding:  
Decision trees and linear models are  
always interpretable.

# Do we need a different model?

## How about rule lists?

If ( sunny and hot )	then	go swim
Else if ( sunny and cold )	then	go ski
Else	then	go work

# Do we need a different model? How about rule lists?

If ( sunny and hot )	then	go swim
Else if ( sunny and cold )	then	go ski
Else if ( wet and weekday )	then	go work
Else if ( free coffee )	then	attend tutorial
Else if ( cloudy and hot )	then	go swim
Else if ( snowing )	then	go ski
Else if ( New Rick and Morty)	then	watch TV
Else if ( paper deadline )	then	go work
Else if ( hungry )	then	go eat
Else if ( tired )	then	watch TV
Else if ( advisor might come )	then	go work
Else if ( code running )	then	watch TV
Else	then	go work

# Maybe rule sets are better?

```
IF ( sunny and hot ) OR ( cloudy and hot ) OR  
( sunny and thirsty and bored )  
THEN go to beach  
ELSE work
```

# Maybe rule sets are better?

```
IF ( sunny and hot ) OR ( cloudy and hot ) OR  
( sunny and thirsty and bored ) OR ( bored and  
tired ) OR ( thirty and tired ) OR ( code running ) OR  
( friends away and bored ) OR ( sunny and want to  
swim ) OR ( sunny and friends visiting ) OR ( need  
exercise ) OR ( want to build castles ) OR ( sunny  
and bored ) OR ( done with deadline and hot ) OR (   
need vitamin D and sunny ) OR ( just feel like it )  
THEN go to beach  
ELSE work
```



**Are you saying decision  
trees, rule lists and rule  
sets don't work?!**

Decision trees, rule lists or rule sets may work for your case!

**The point here is that there is no one-size-fits-all method.**

# Is interpretability possible at all?

WIRED

Our Machines Now Have Knowledge We'll Never Understand

SUBSCRIBE 

DAVID WEINBERGER BACKCHANNEL 04.18.17 08:22 PM

## OUR MACHINES NOW HAVE KNOWLEDGE WE'LL NEVER UNDERSTAND

### SHARE



SHARE  
176



TWEET



COMMENT

The new availability of huge amounts of data, along with the statistical tools to crunch these numbers, offers a whole new way of understanding the world. Correlation supersedes causation, and science can advance even without coherent models, unified theories, or really any mechanistic explanation at all.

So wrote *Wired's* **Chris Anderson** in 2008. It kicked up a

# Is interpretability possible at all?

WIRED

Our Machines Now Have Knowledge We'll Never Understand

SUBSCRIBE

DAVID WEINBERGER BACKCHANNEL 04.18.17 08:22 PM

## OUR MACHINES NOW HAVE KNOWLEDGE WE'LL



Common misunderstanding:  
We need to understand every single thing  
about the model.

of understanding the world. Correlation supersedes causation,  
and science can advance even without coherent models, unified  
theories, or really any mechanistic explanation at all



TWEET

### Key Point:

Interpretability is NOT about understanding all bits and bytes  
of the model for all data points.

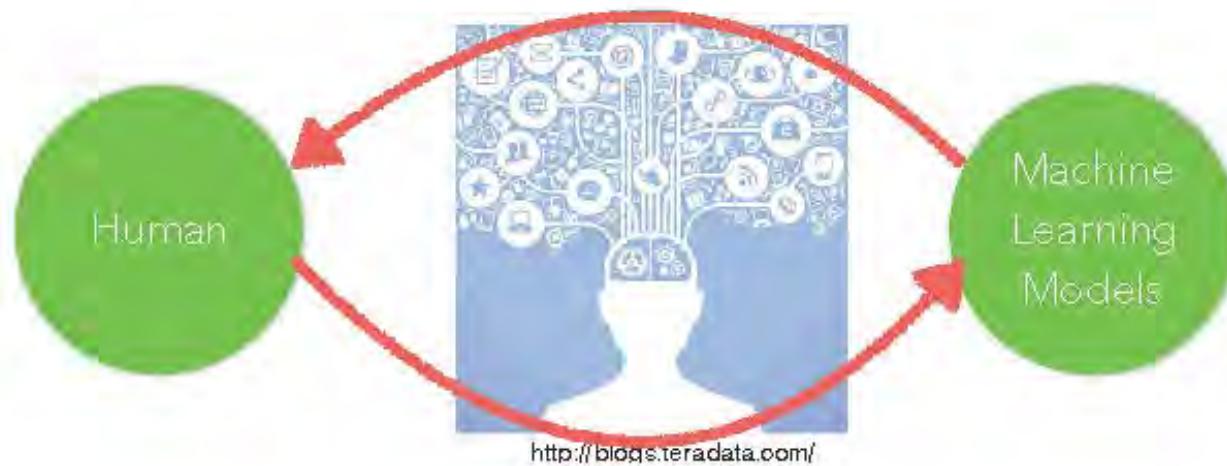
It is about knowing enough for your goals/downstream tasks.

My goal

# interpretability

To use machine learning **responsibly**  
we need to ensure that

1. our **values** are aligned
2. our **knowledge** is reflected

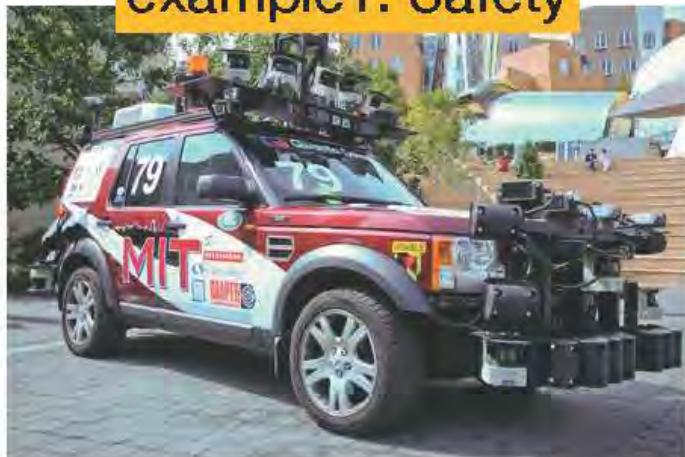


# Why interpretability?

Fundamental **underspecification** in the problem

# Why interpretability?

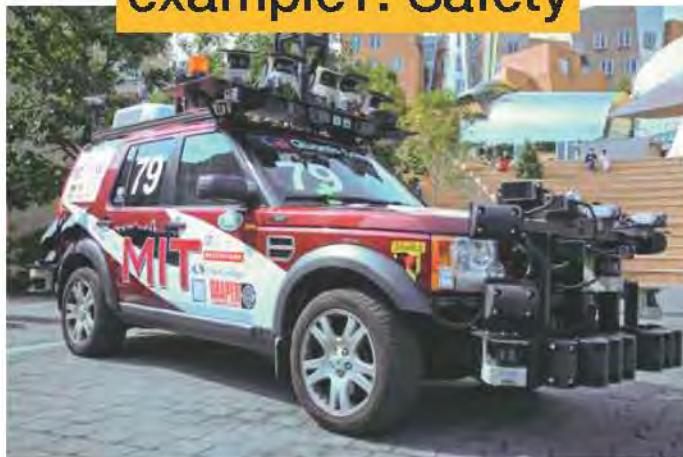
example1: Safety



Fundamental **underspecification** in the problem

# Why interpretability?

example1: Safety



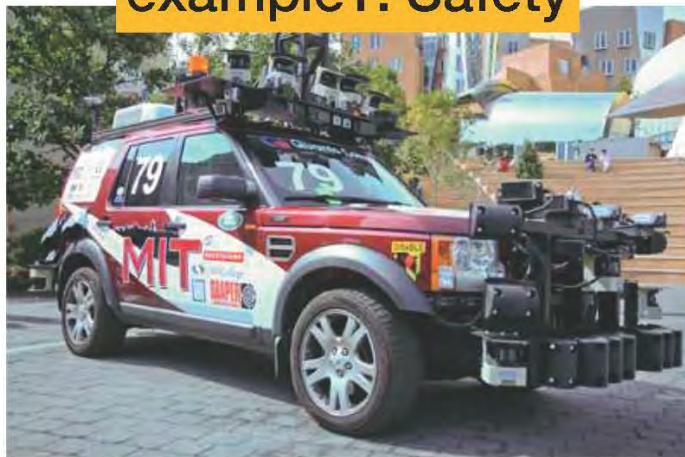
example 2: Science



Fundamental **underspecification** in the problem

# Why interpretability?

example1: Safety



example 2: Science



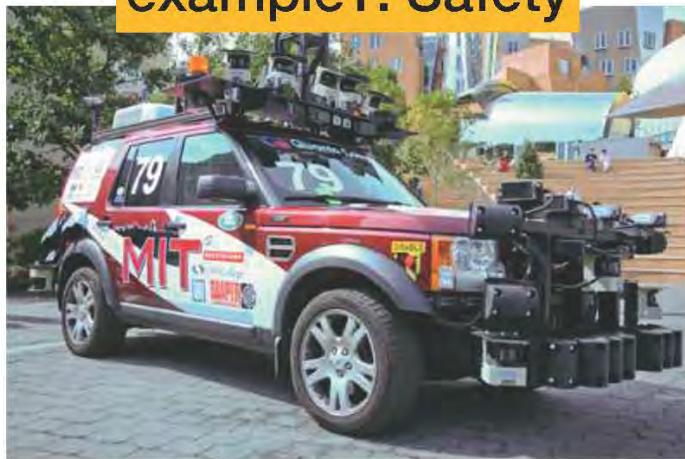
Fundamental **underspecification** in the problem

example3: mismatched objectives



# Why interpretability?

example1: Safety



example 2: Science



Fundamental **underspecification** in the problem



Common misunderstanding:

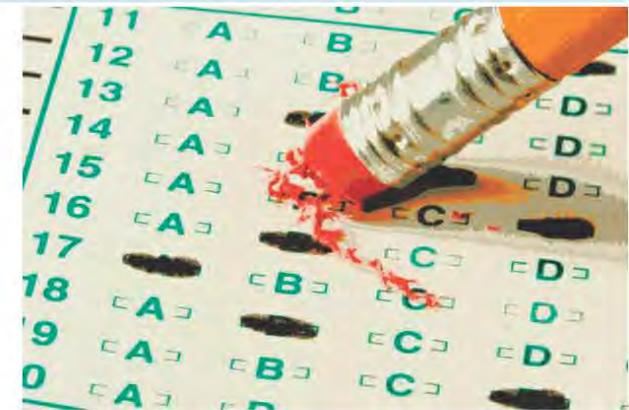
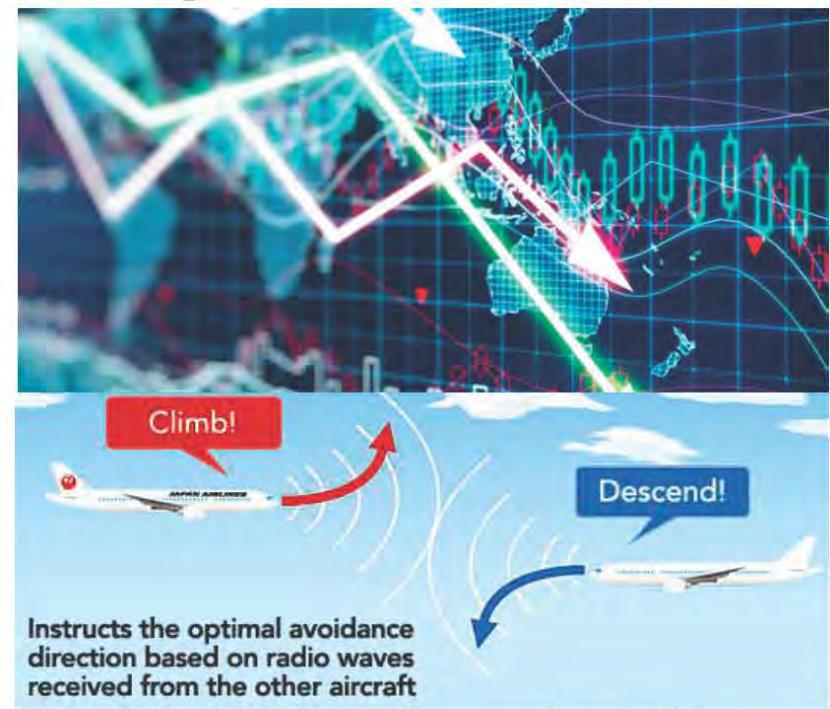
More data or more clever algorithm will solve interpretability.

# What is NOT underspecification?



# When we may **not** want interpretability

- No significant consequences or when predictions are all you need.
- Sufficiently well-studied problem
- Prevent gaming the system - mismatched objectives.



[https://cdn.theatlantic.com/assets/media/img/mt/2015/04/shutterstock\\_11926084/lead\\_large.jpg](https://cdn.theatlantic.com/assets/media/img/mt/2015/04/shutterstock_11926084/lead_large.jpg)  
[https://www.jal.com/assets/img/flight/safety/equipment/pic\\_tcas\\_001\\_en.jpg](https://www.jal.com/assets/img/flight/safety/equipment/pic_tcas_001_en.jpg)

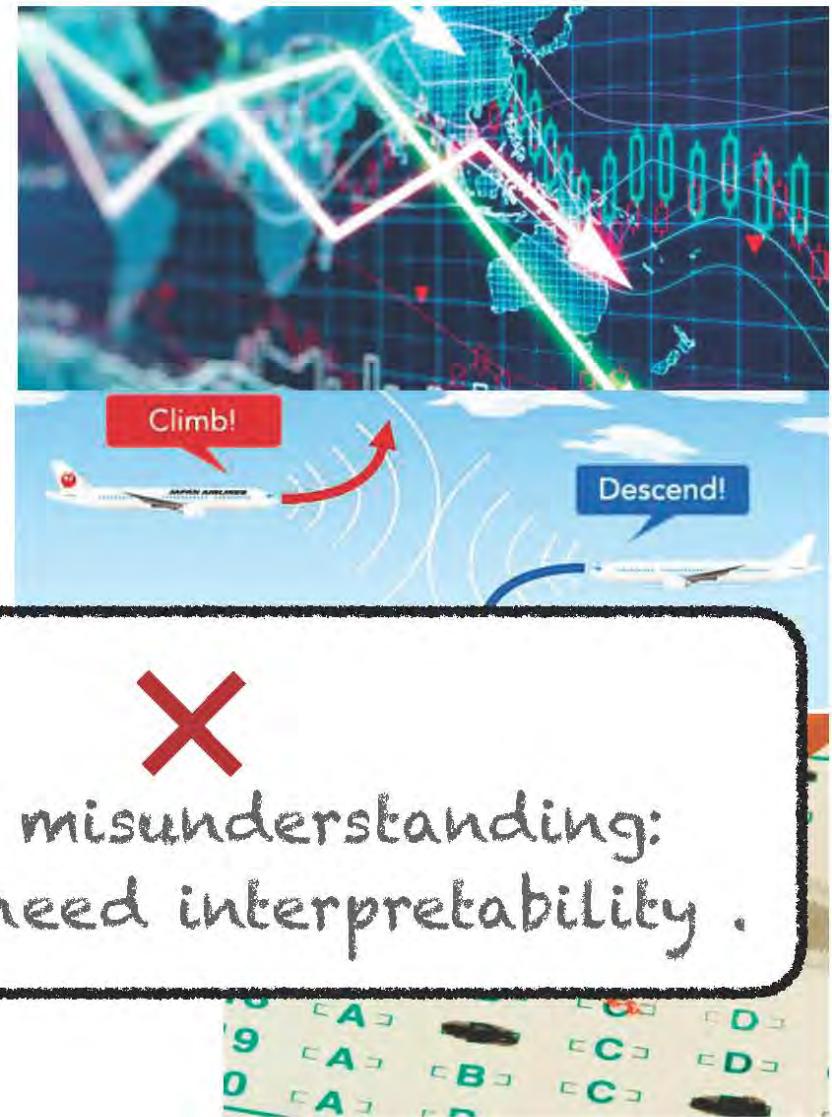
<http://www.cinemablend.com/pop/Netflix-Using-Amazon-Cloud-Explore-Artificial-Intelligence-Movie-Recommendations-62248.html>

# When we may **not** want interpretability

- No significant consequences or when predictions are all you need.

- Sufficiently well-studied problem

- Prevent gaming the mismatched objectives.



[https://cdn.theatlantic.com/assets/media/img/mt/2015/04/shutterstock\\_11926084/lead\\_large.jpg](https://cdn.theatlantic.com/assets/media/img/mt/2015/04/shutterstock_11926084/lead_large.jpg)  
[https://www.jal.com/assets/img/flight/safety/equipment/pic\\_tcas\\_001\\_en.jpg](https://www.jal.com/assets/img/flight/safety/equipment/pic_tcas_001_en.jpg)

<http://www.cinemablend.com/pop/Netflix-Using-Amazon-Cloud-Explore-Artificial-Intelligence-Movie-Recommendations-62248.html>

Interpretability

?

fairness  
accountability  
trust  
causality etc.

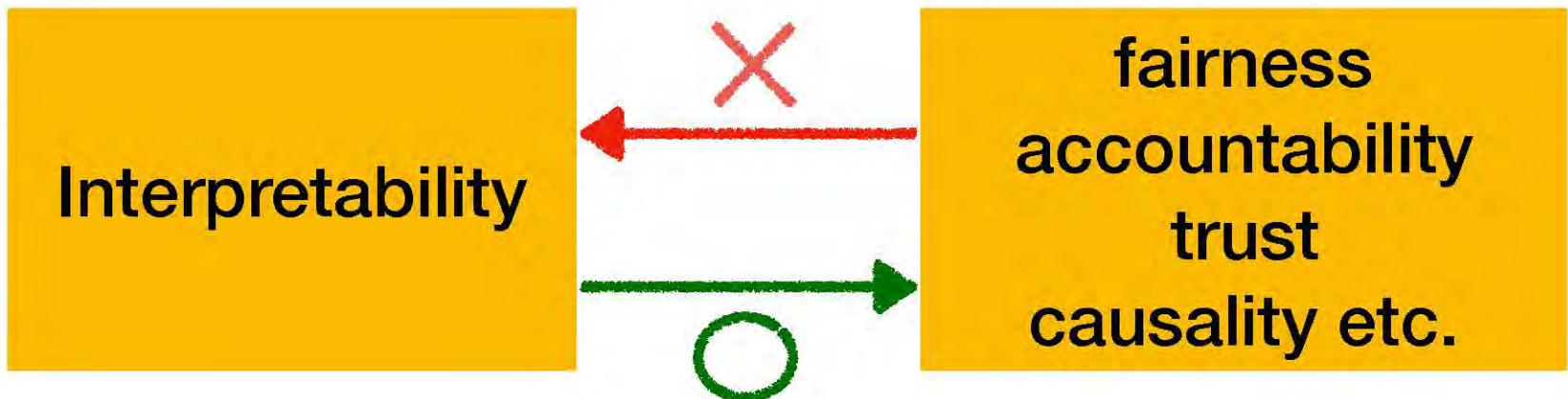
Interpretability

fairness  
accountability  
trust  
causality etc.



Common misunderstanding:  
Trust, fairness and interpretability  
are all the same thing.

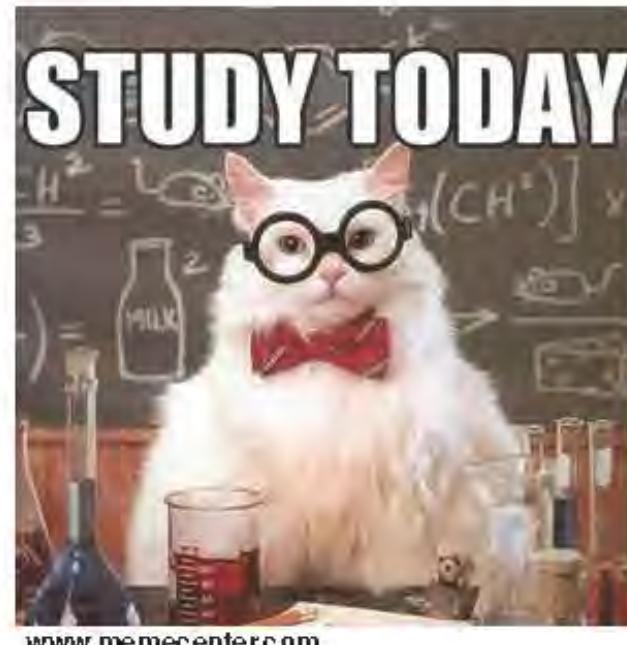
# Our cousins are not us



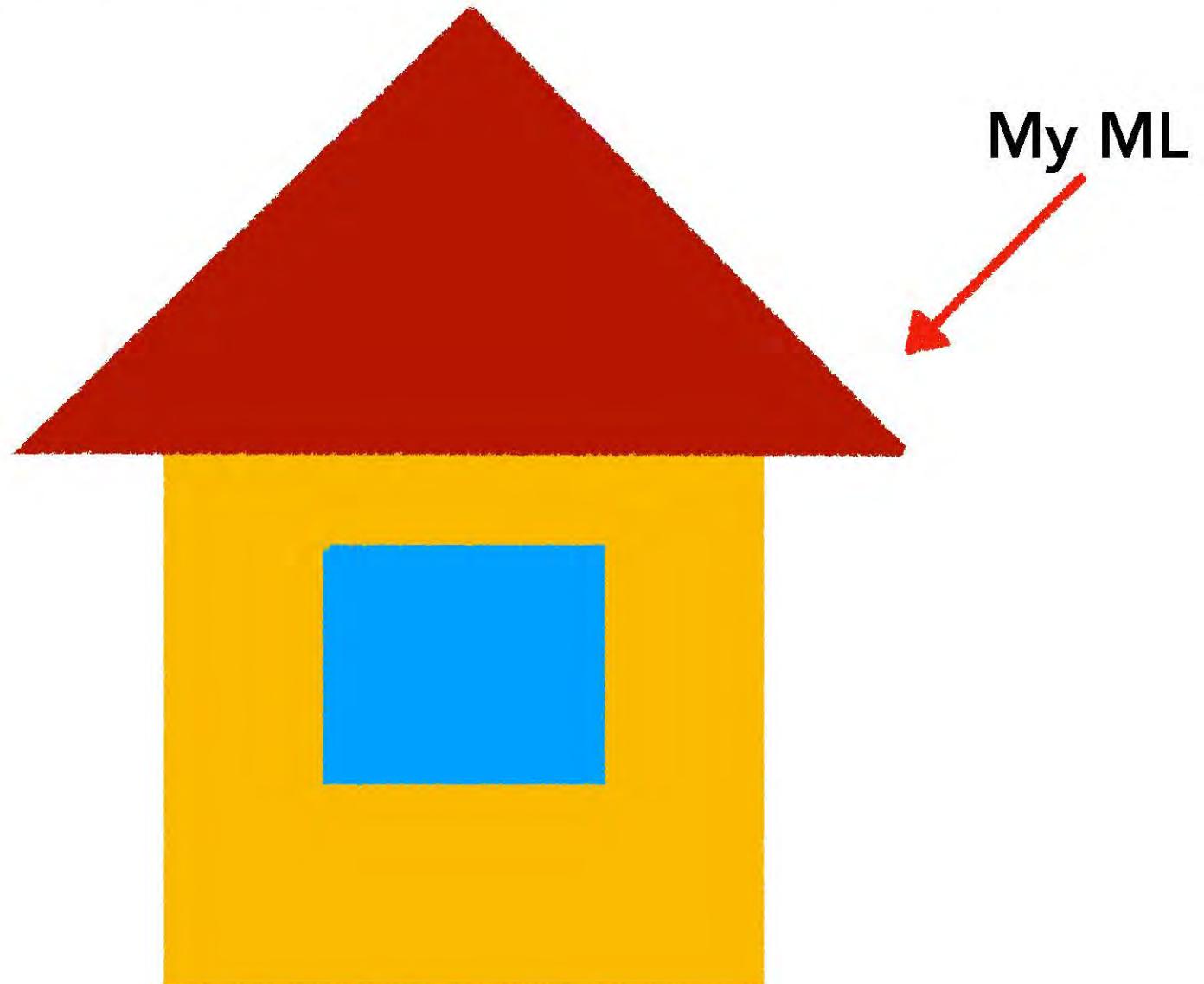
- Interpretability can help with them **when we cannot formalize these ideas**
- But once formalized, you may not need interpretability.

# Agenda

- **When** and **why** interpretability
- **Overview** of interpretability methods.
- How to **Evaluate** interpretability methods.

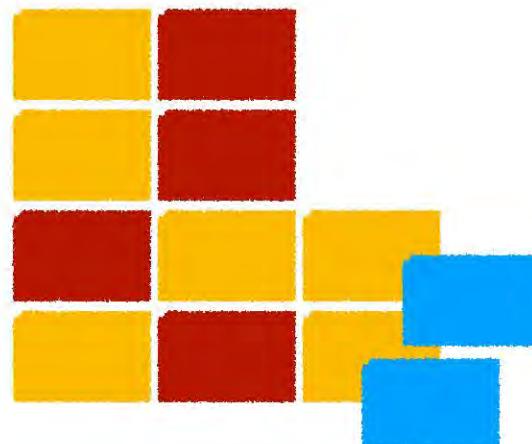


# Types of interpretable methods



# Types of interpretable methods

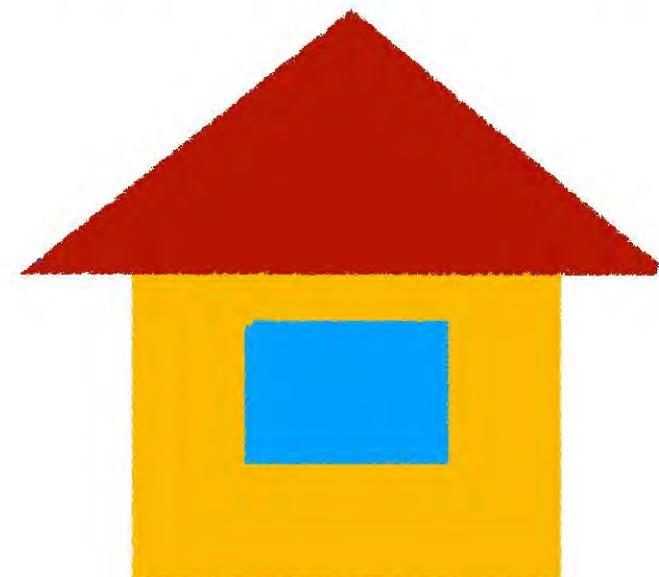
**Before building  
any model**



**Building  
a new model**

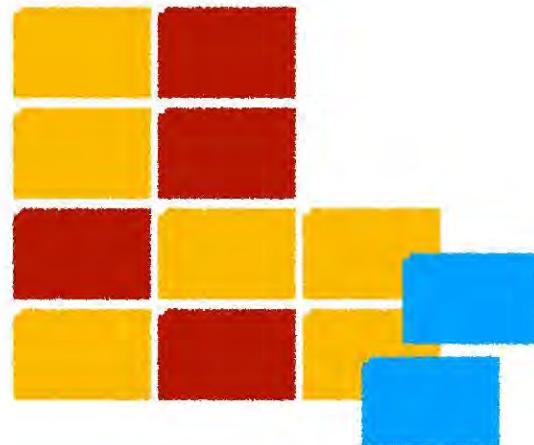


**After  
building a model**



# Types of interpretable methods

Before building  
any model



Building  
a new model

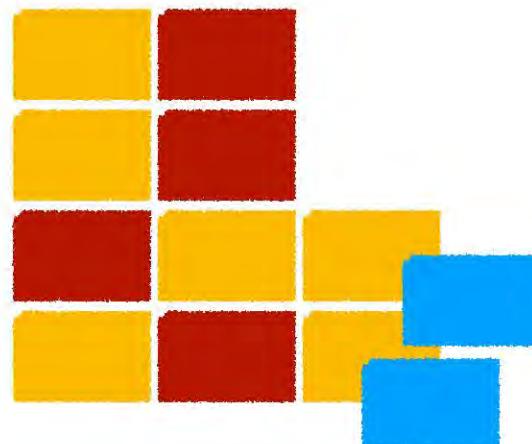
After  
building a model



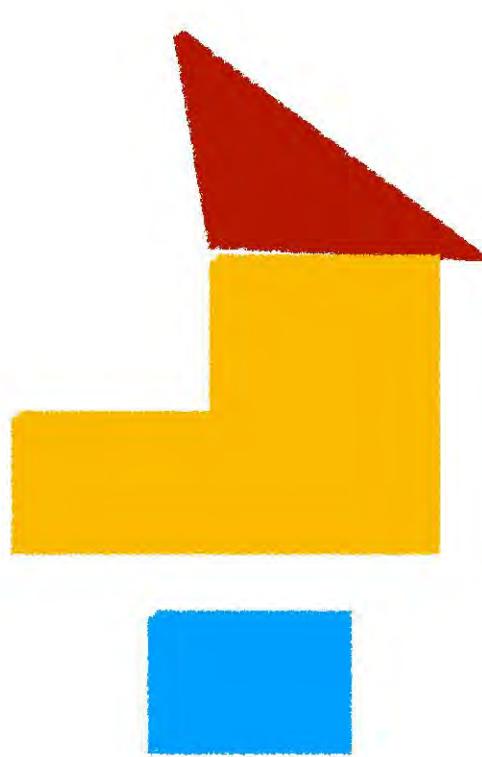
Common misunderstanding:  
Interpretability is always about  
machine learning models.

# Types of interpretable methods

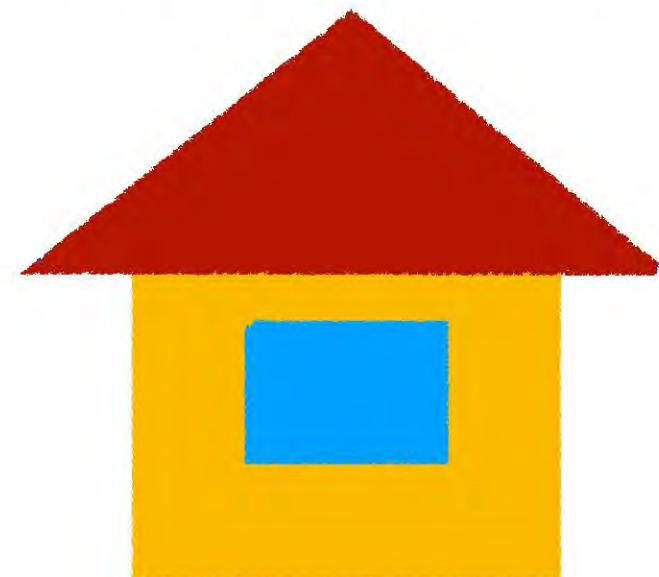
**Before building  
any model**



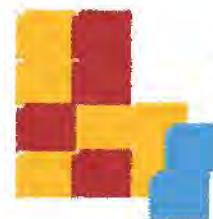
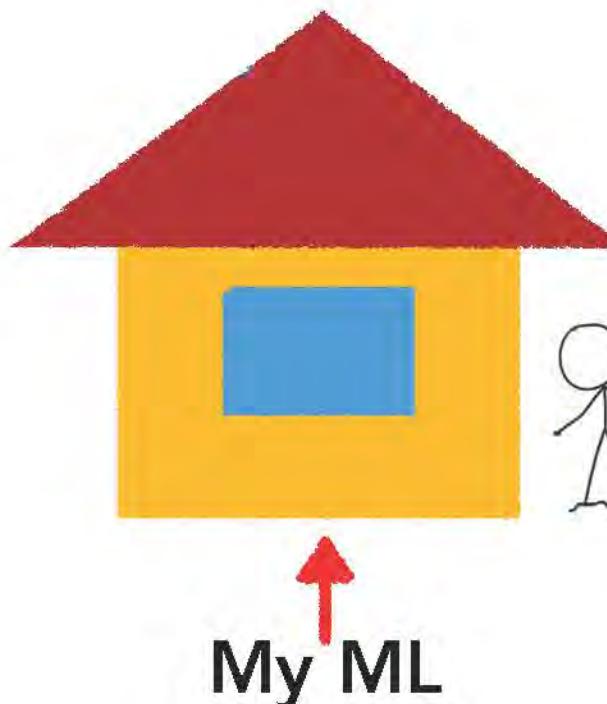
**Building  
a new model**



**After  
building a model**



# Types of interpretability methods



Before building  
any model

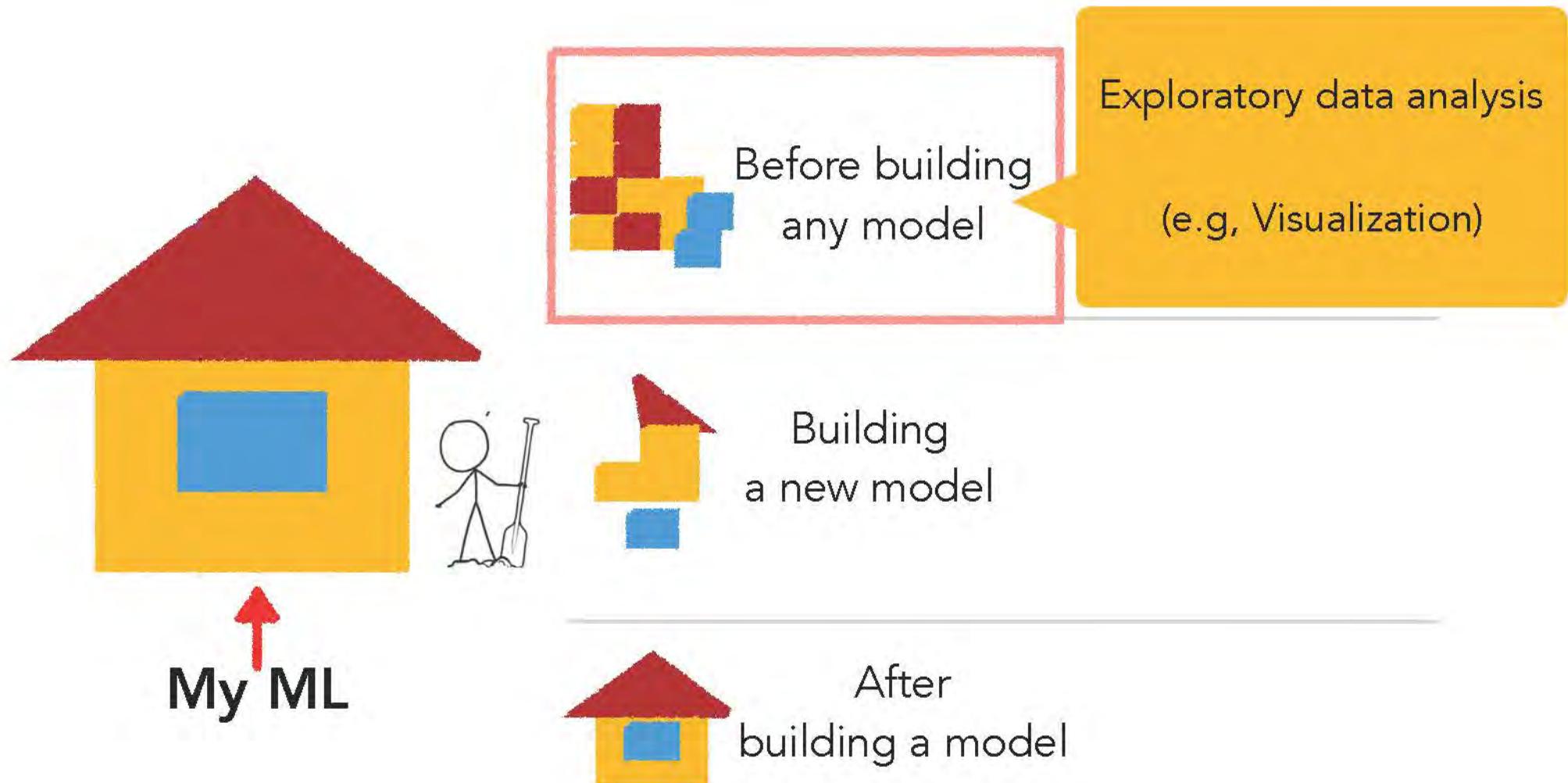


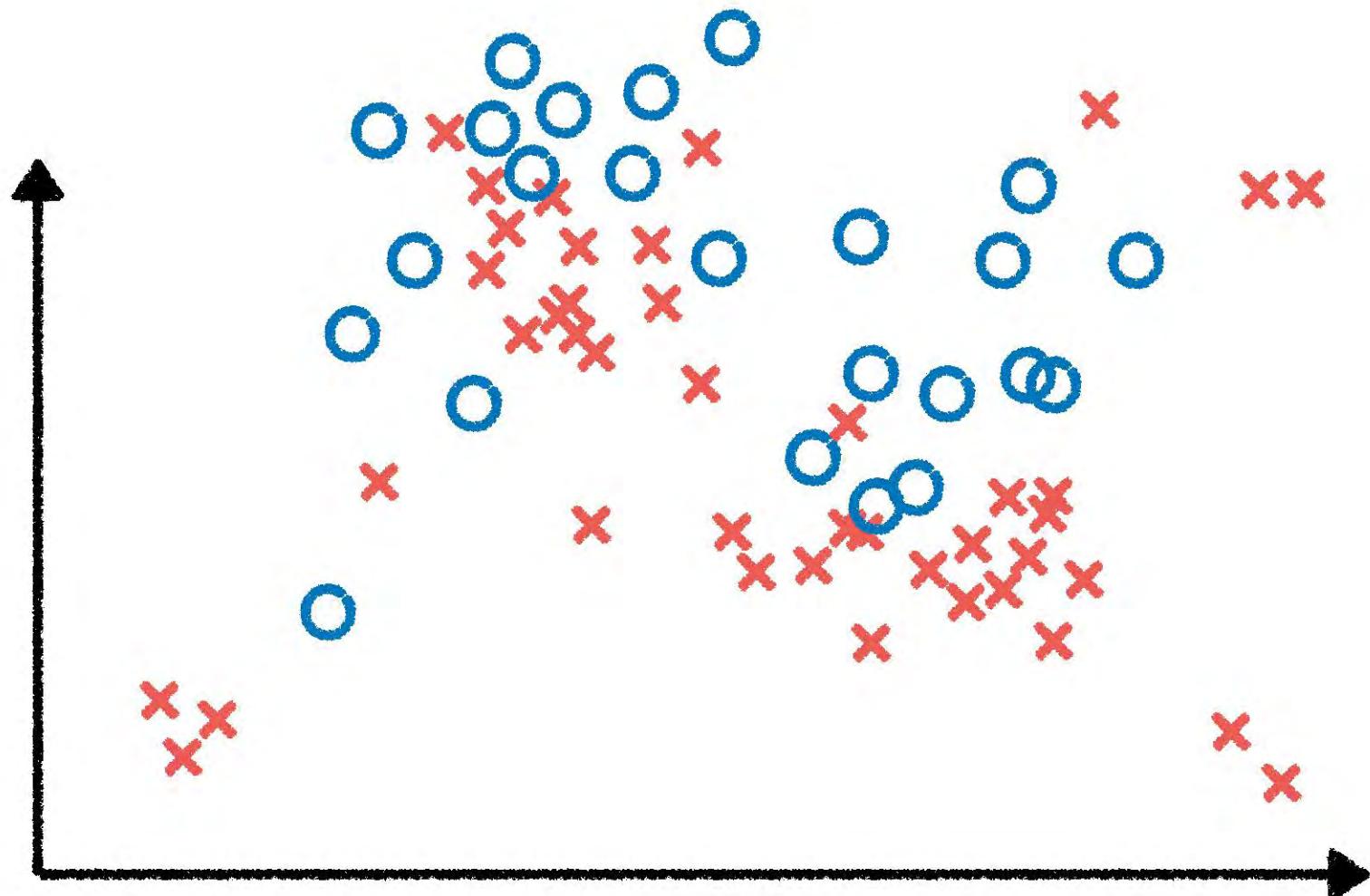
Building  
a new model



After  
building a model

# Types of interpretability methods





○ Class0

✗ Class1

## Before building any model



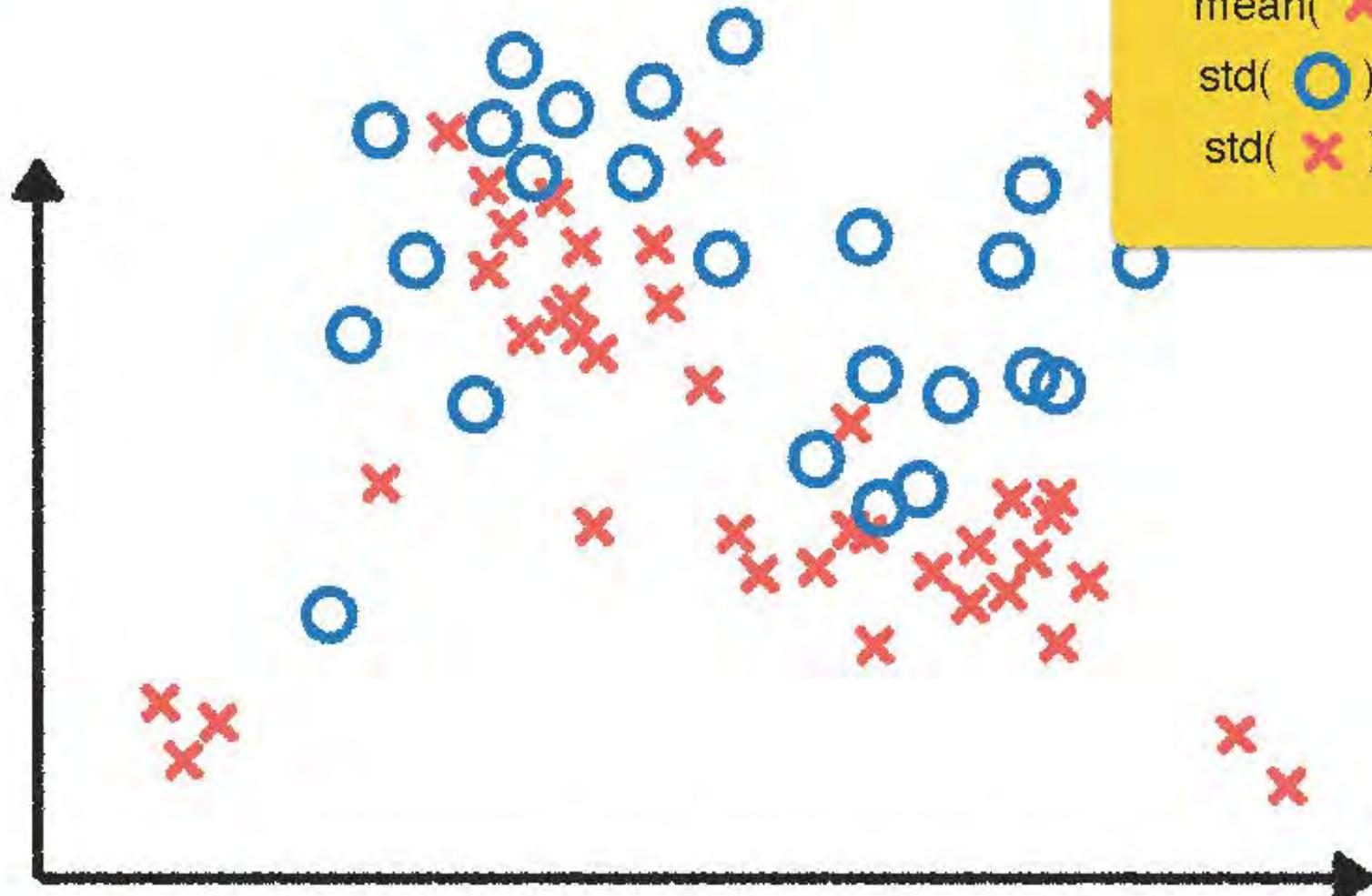
### Descriptive statistics

mean( )

mean( )

std( )

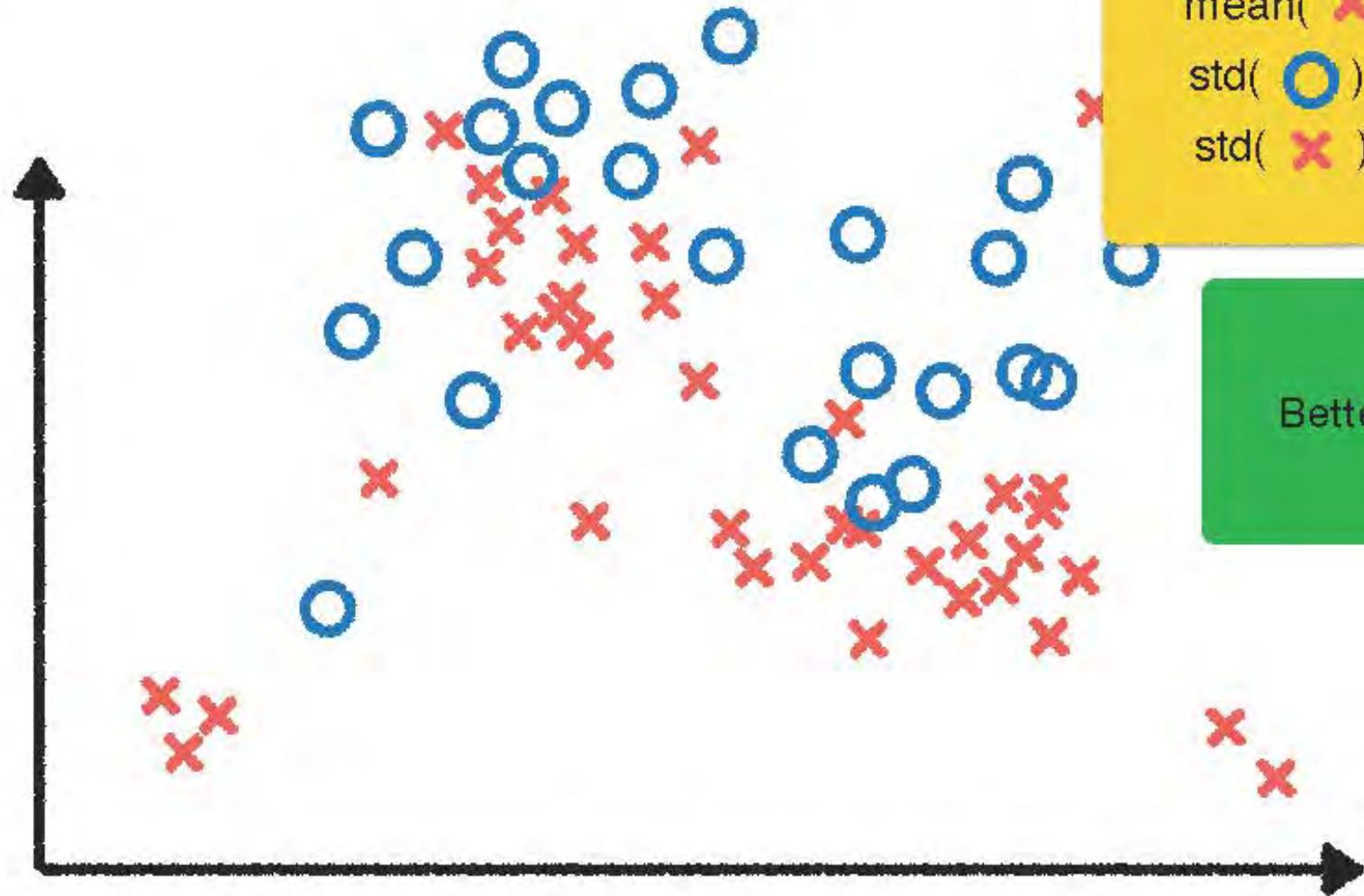
std( )



Class0

Class1

## Before building any model



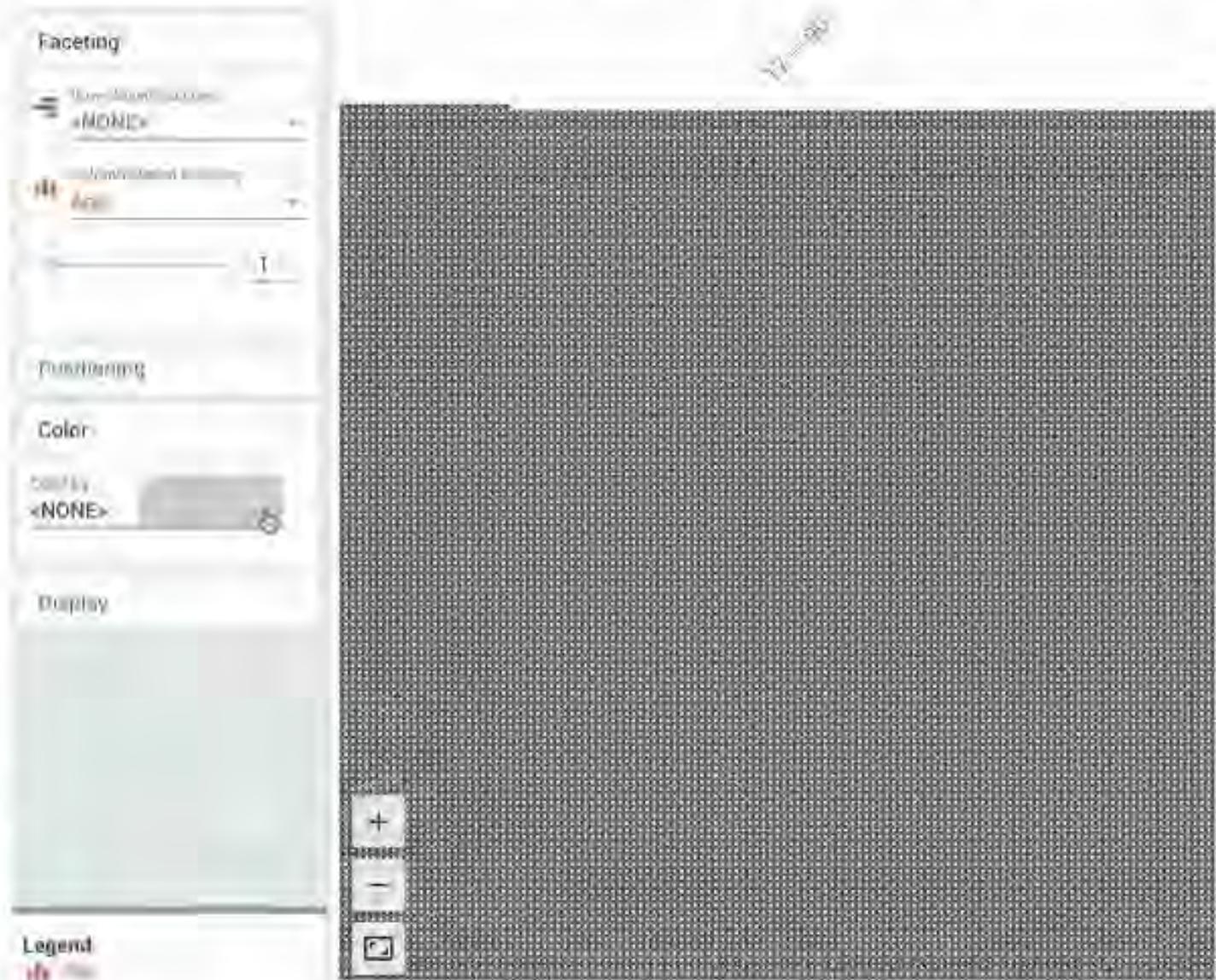
○ Class0

✖ Class1



Before building any model

# Visualization for data exploration

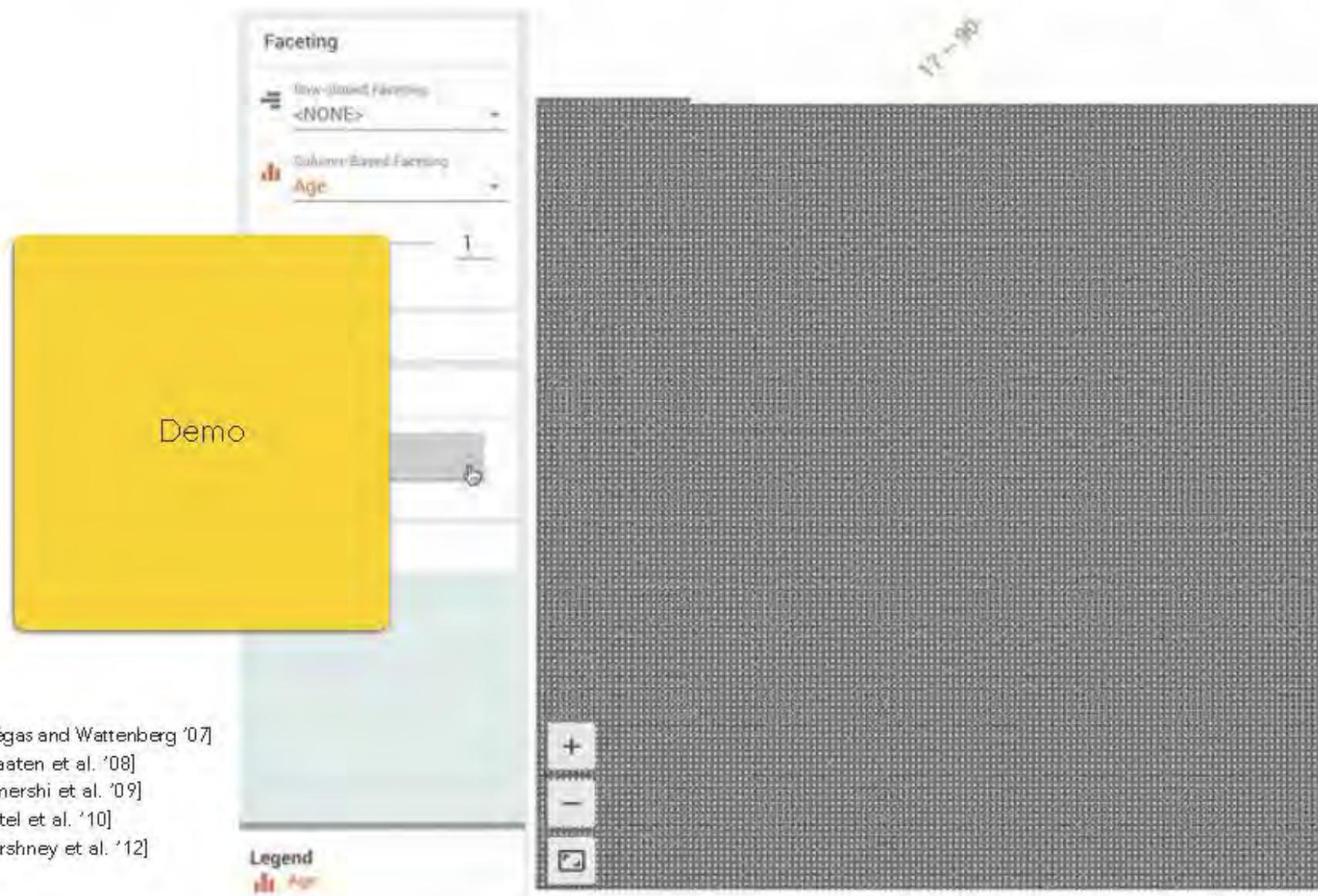


<https://pair-code.github.io/facets/quickdraw.html>



Before building any model

# Visualization for data exploration

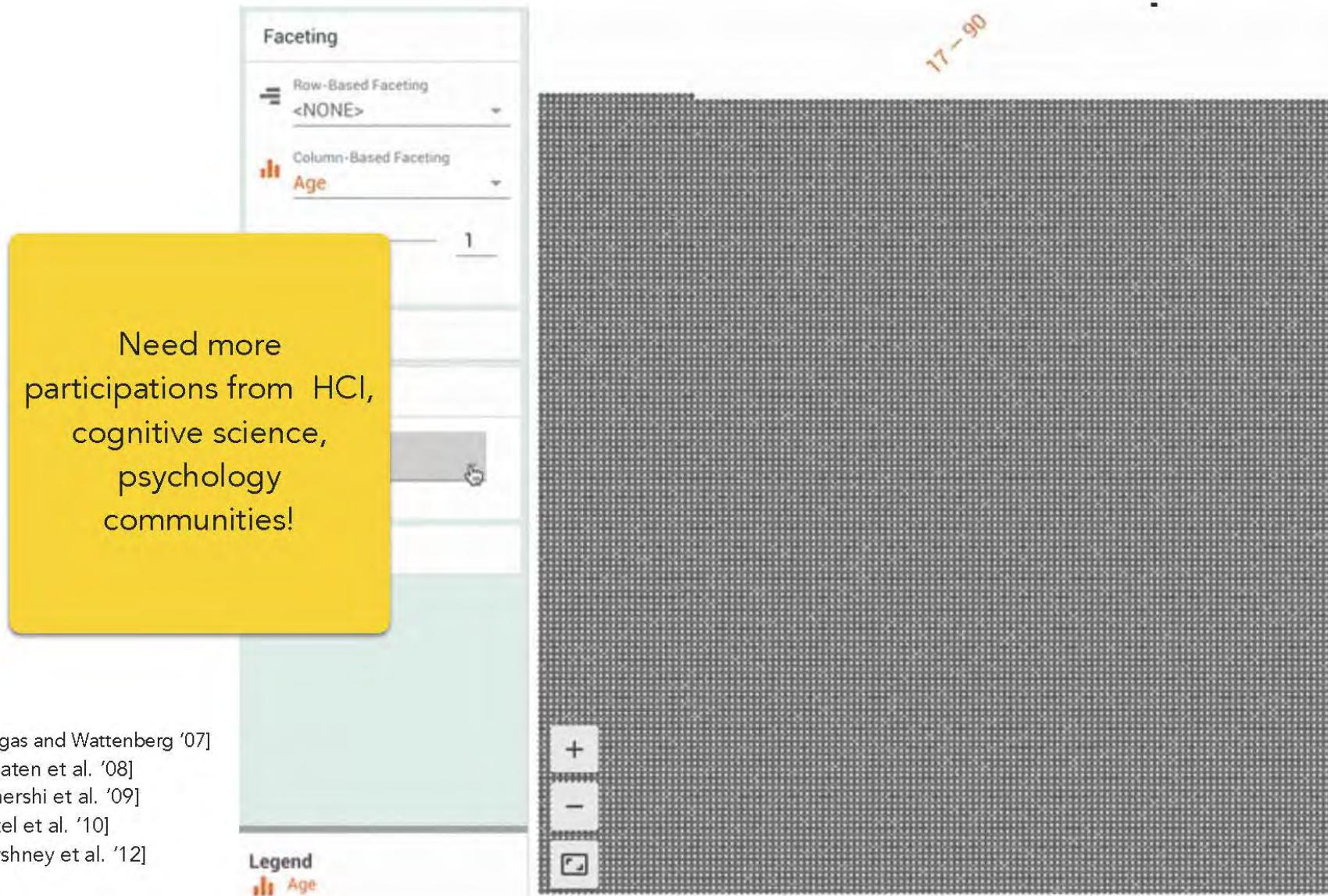


<https://pair-code.github.io/tacets/quickdraw.html>



Before building any model

# Visualization for data exploration

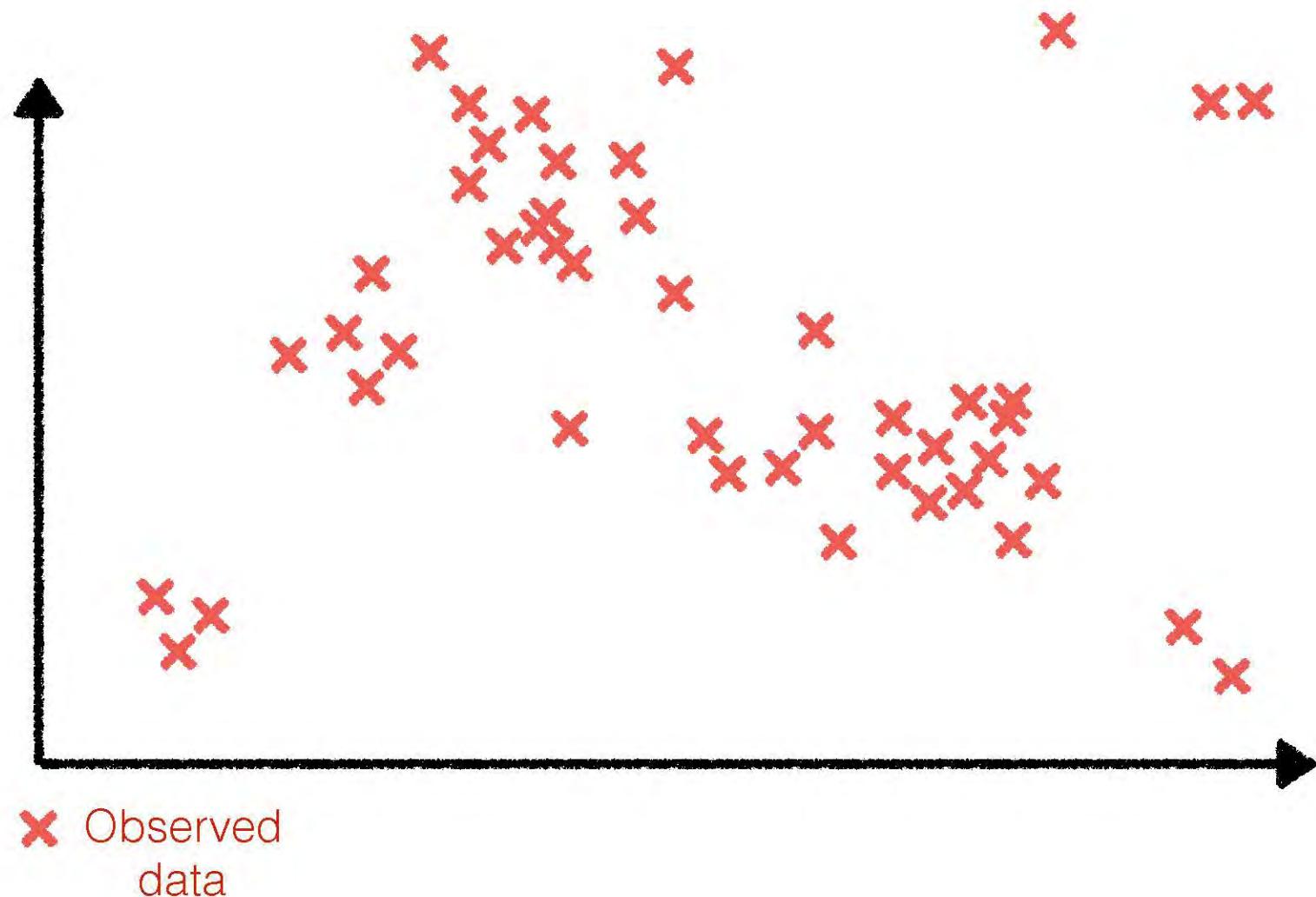


<https://pair-code.github.io/facets/quickdraw.html>



Before building any model

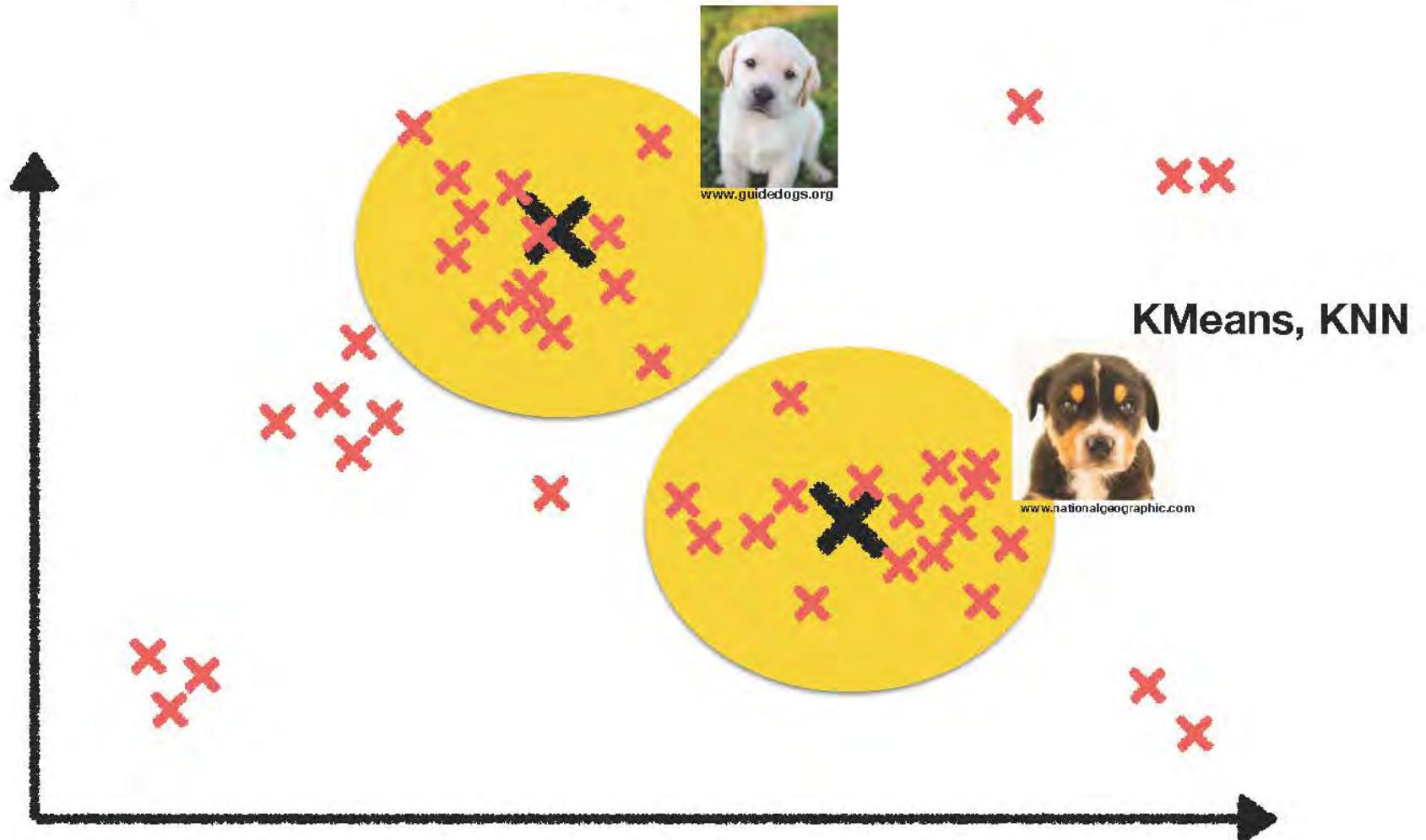
# Exploratory data analysis





Before building any model

# Exploratory data analysis



✖ Observed  
data

[Simon et al., '07]  
[Lin and Bilmes, '11]



Before building any model

# Exploratory data analysis

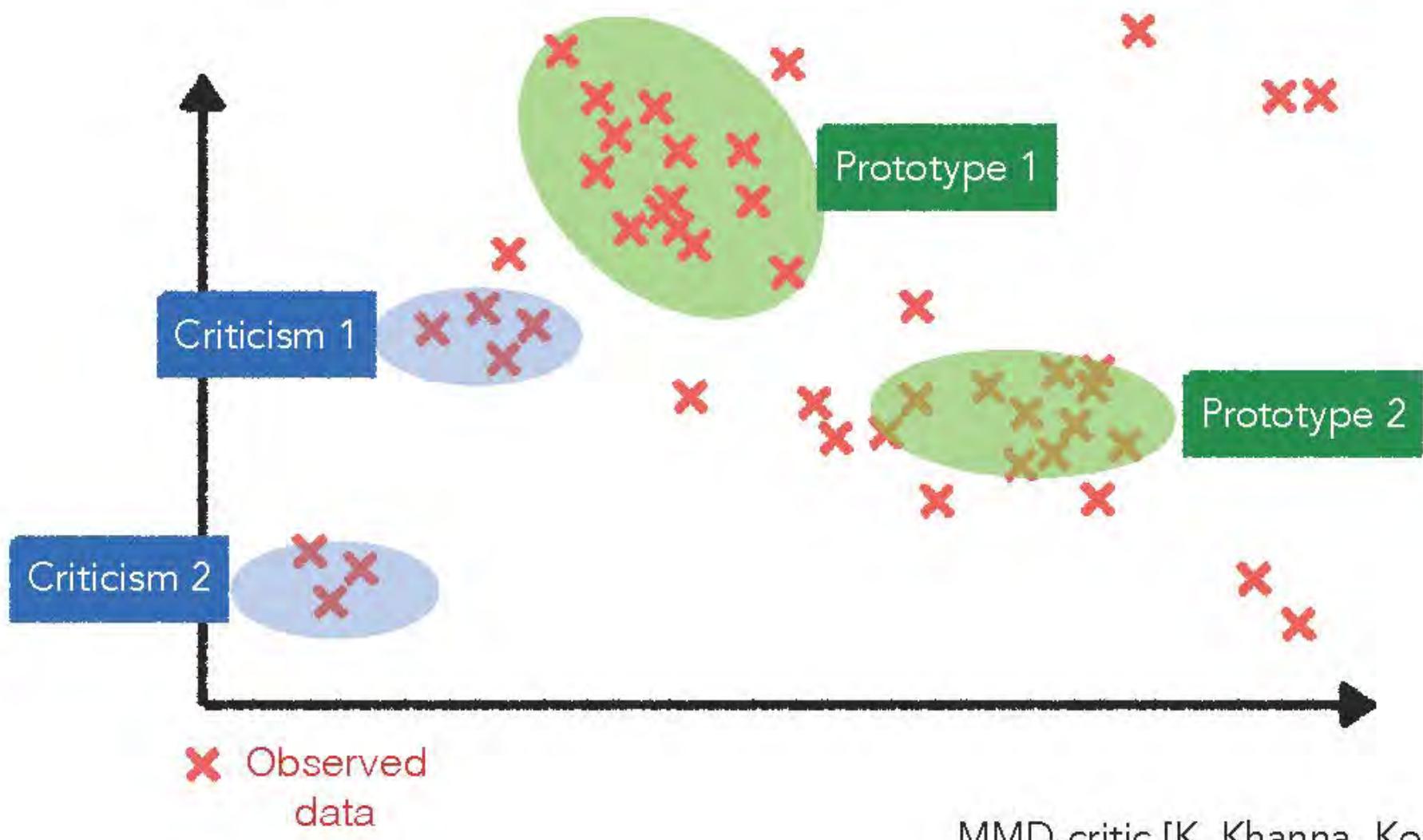


✖ Observed  
data

[Simon et al., '07]  
[Lin and Bilmes, '11]

Before building any model

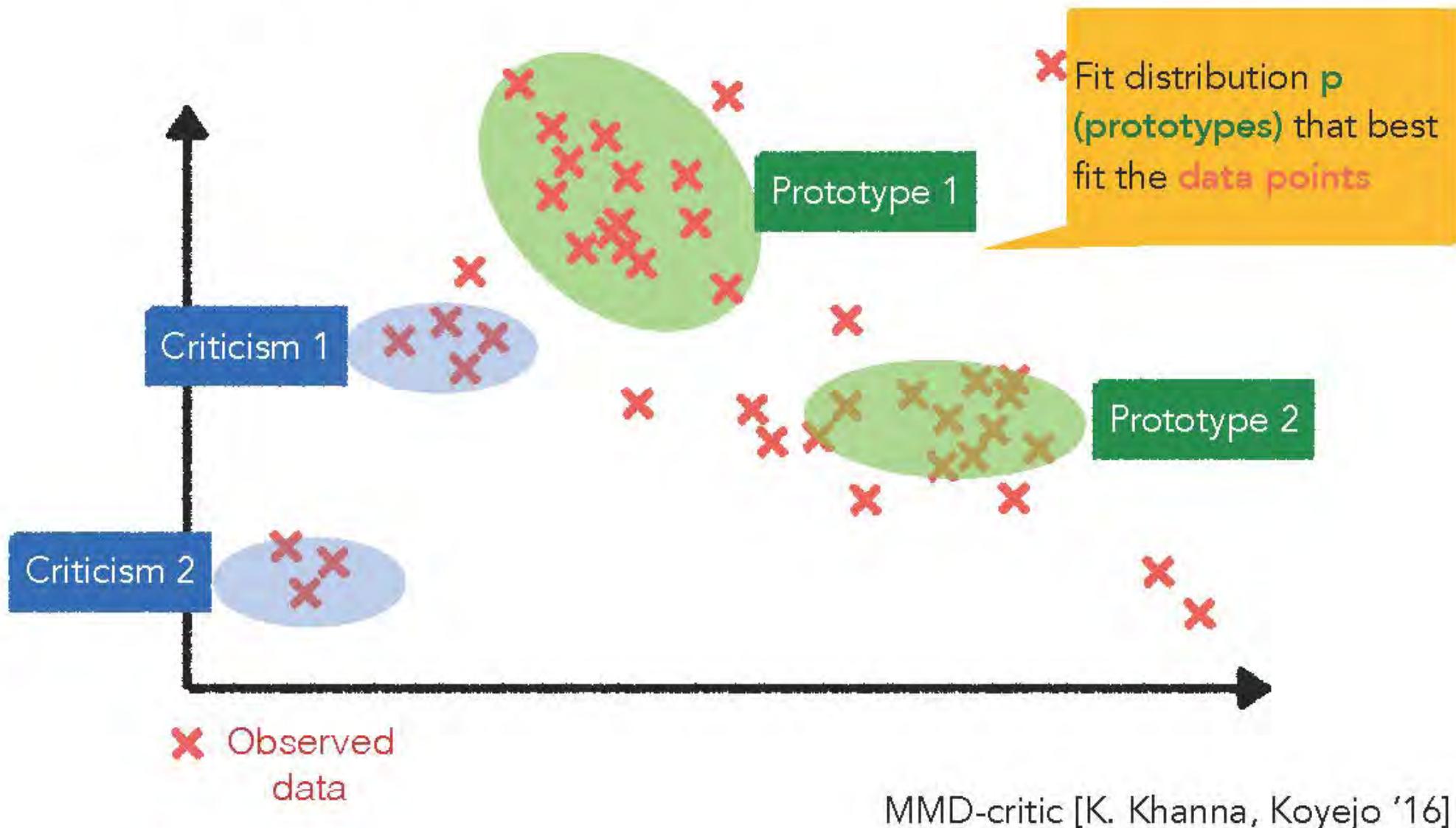
# Exploratory data analysis





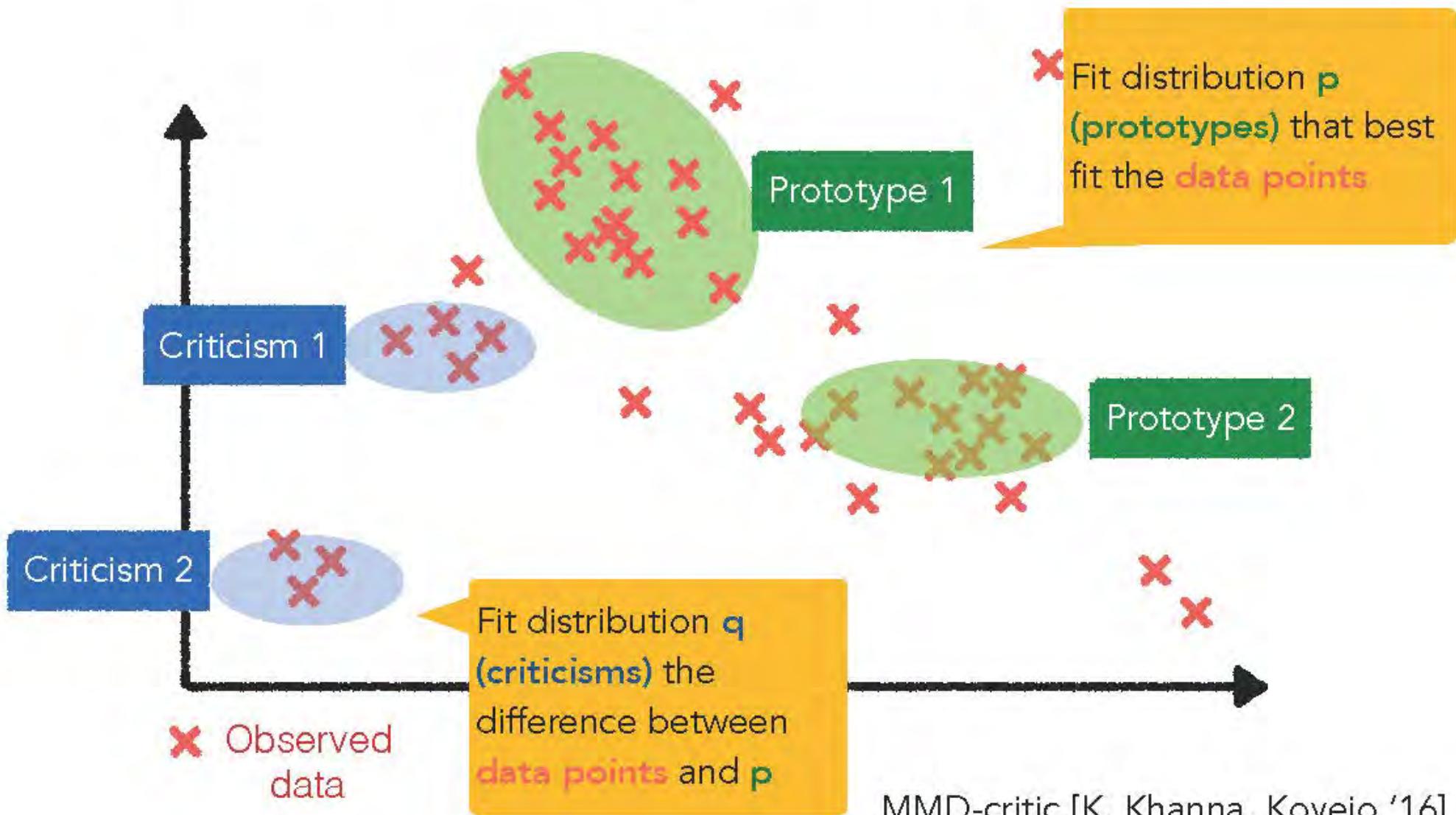
Before building any model

# Exploratory data analysis



Before building any model

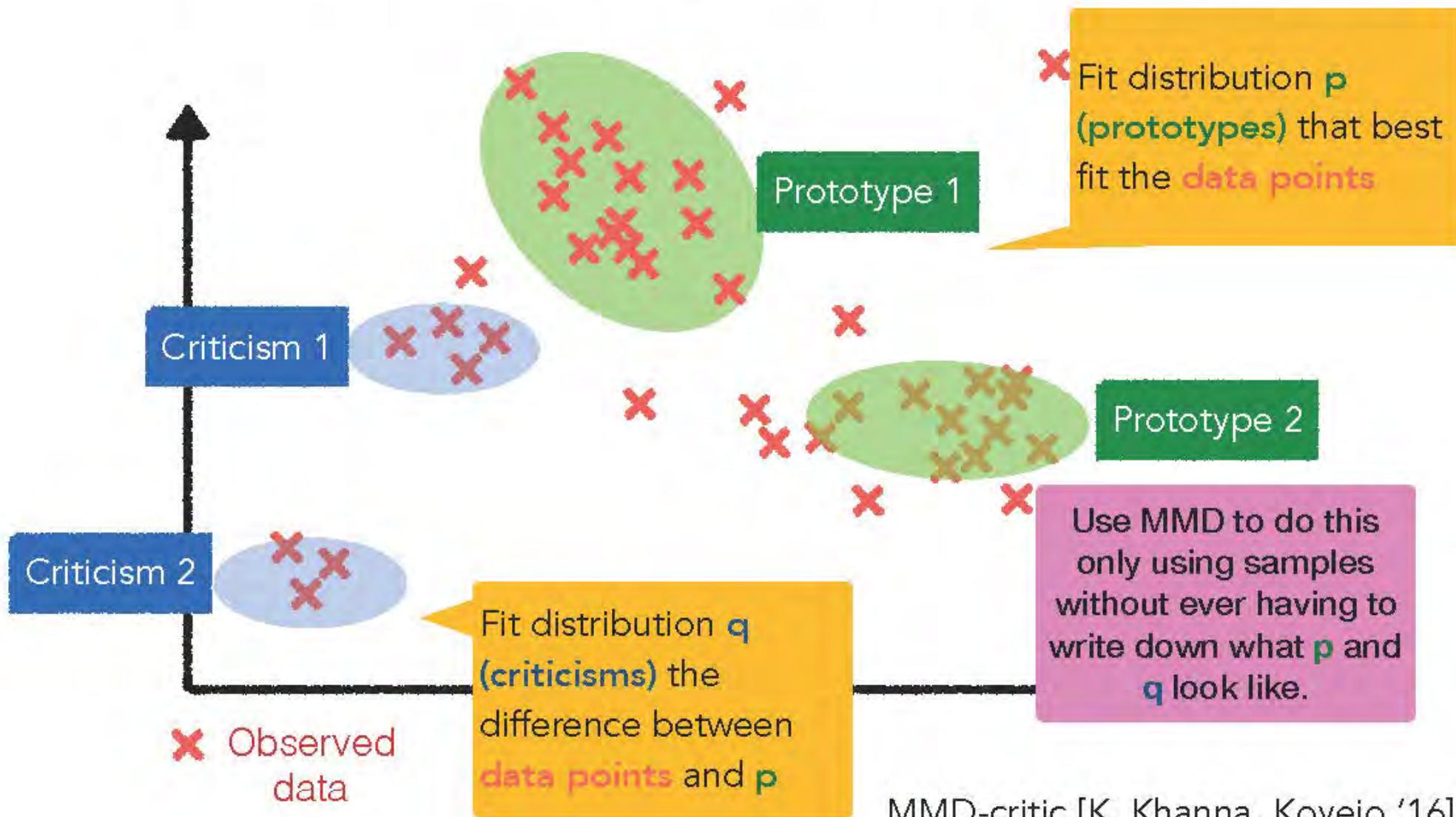
# Exploratory data analysis





Before building any model

# Exploratory data analysis



MMD-critic [K. Khanna, Koyejo '16]



Before building any model

# Exploratory data analysis

Prototypes



Criticisms



Prototypes

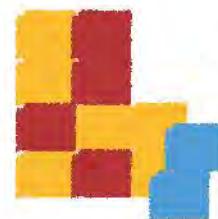


Criticisms



MMD-critic [K. Khanna, Koyejo '16]

# Types of interpretability methods



Before building  
any model

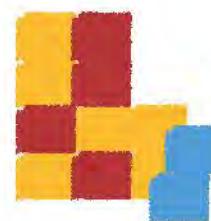


Building  
a new model



After  
building a model

# Types of interpretability methods

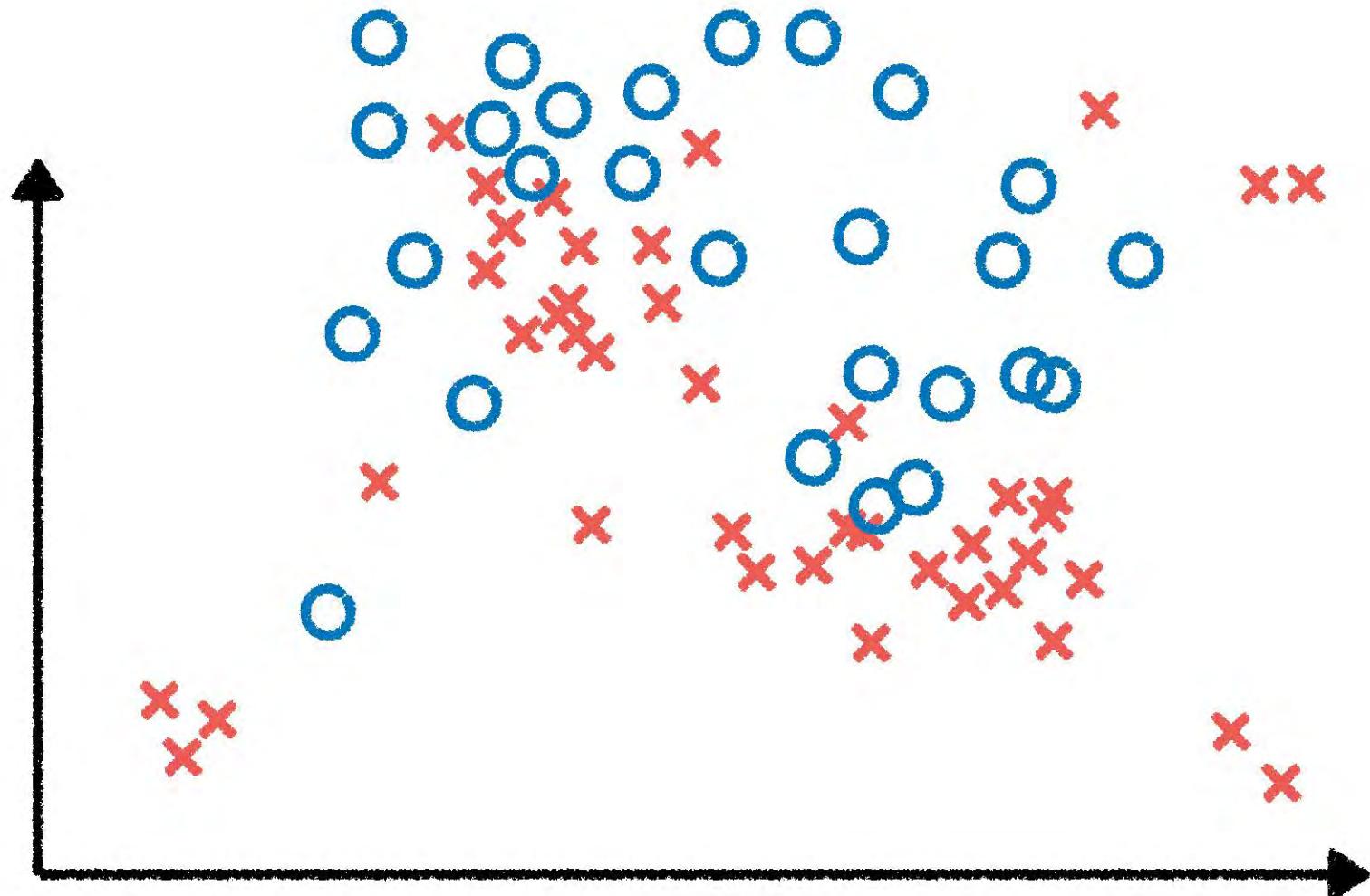


Before building  
any model



What is the medium and  
constraints we use to  
explain?

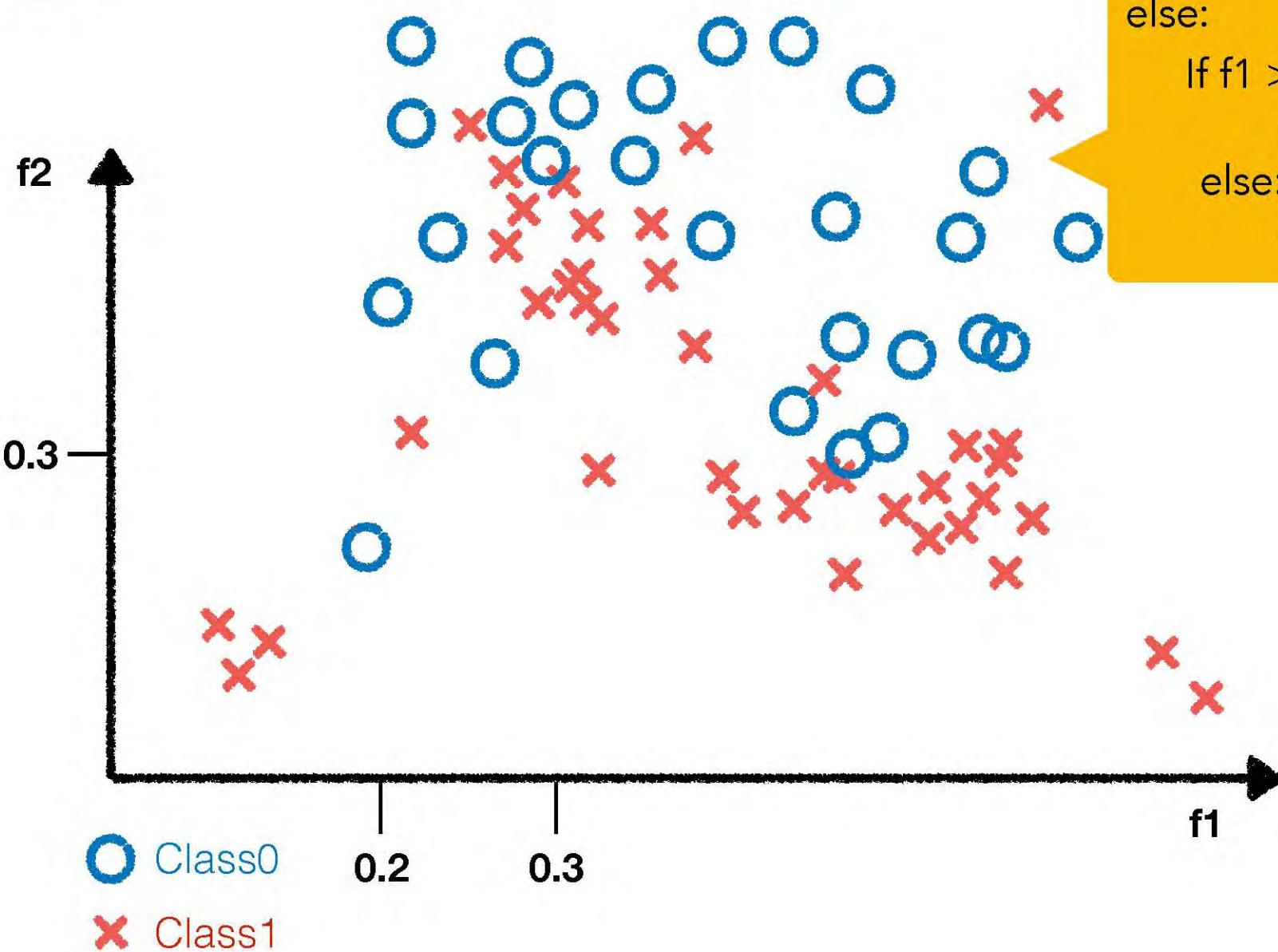
Rules, Examples, Sparsity  
and Monotonicity



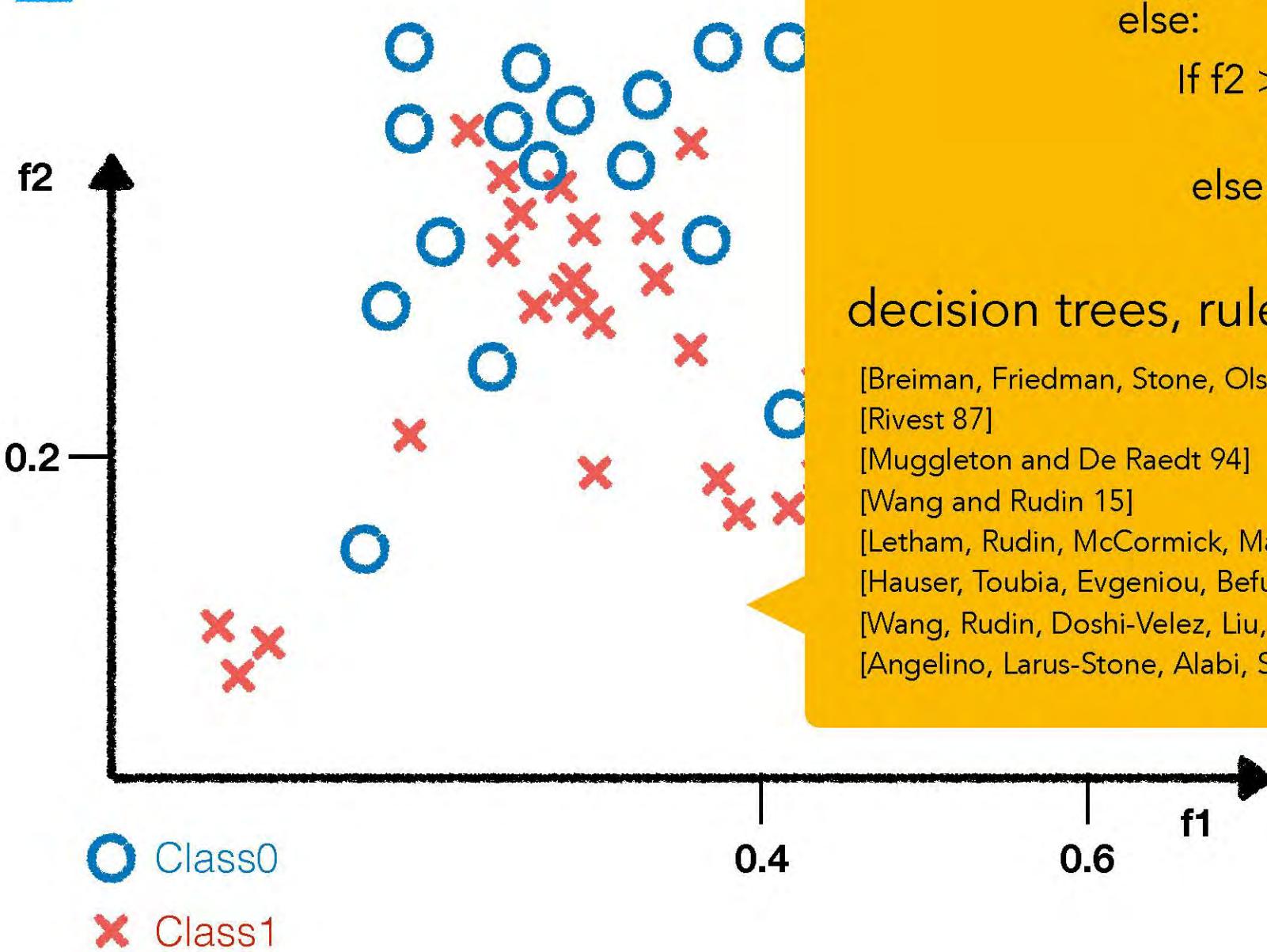
○ Class0

✗ Class1

# Building a new model



# Building a new model



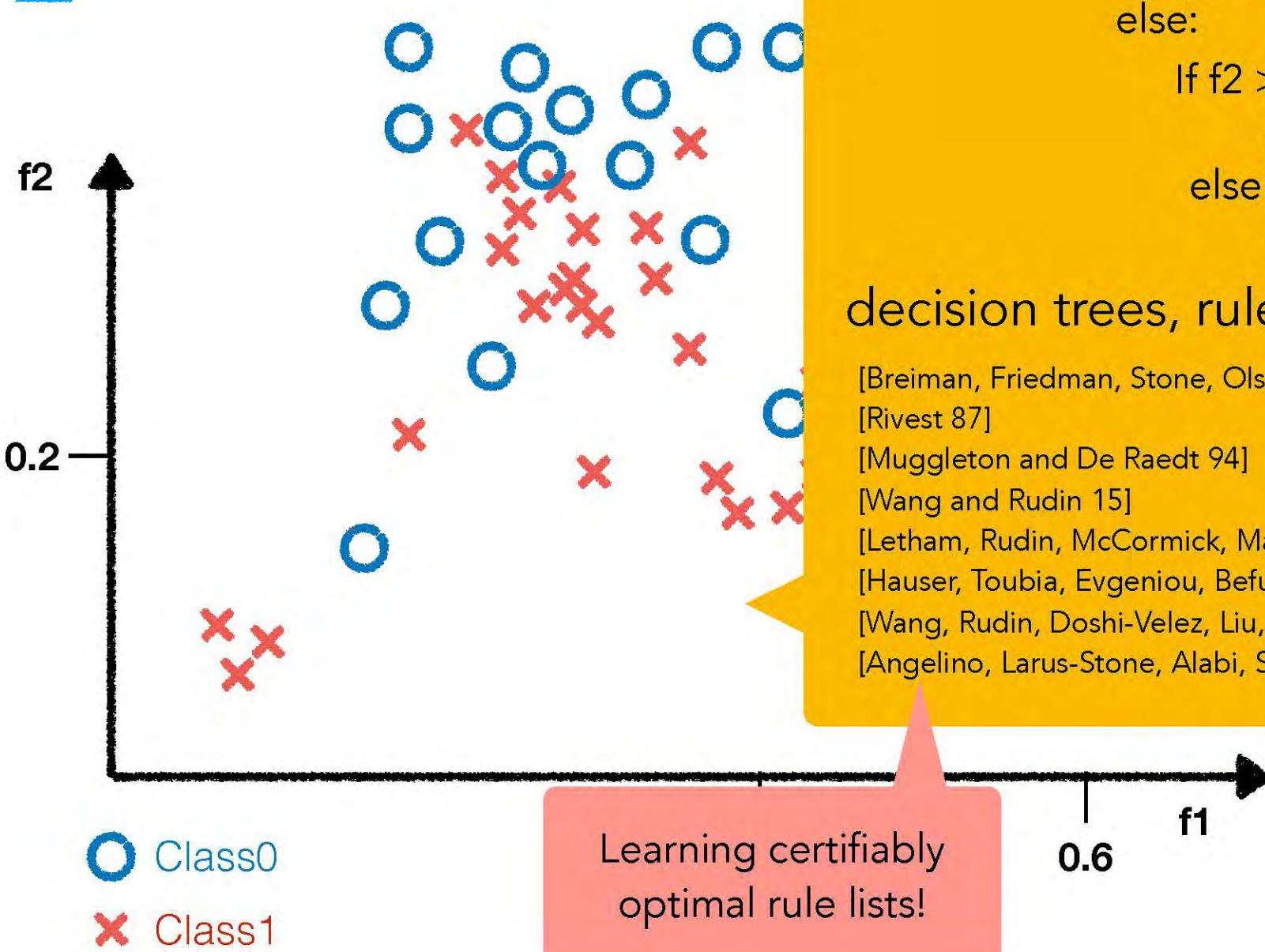
## Rule based

```
If  $f_1 < 0.1$ :  
    predict ✗  
else:  
    If  $f_2 > 0.4$  and  $f_2 < 0.6$ :  
        predict ○  
    else:  
        ...
```

## decision trees, rule lists, rule sets

- [Breiman, Friedman, Stone, Olshen 84]
- [Rivest 87]
- [Muggleton and De Raedt 94]
- [Wang and Rudin 15]
- [Letham, Rudin, McCormick, Madigan '15]
- [Hauser, Toubia, Evgeniou, Befurt, Dzyabura 10]
- [Wang, Rudin, Doshi-Velez, Liu, Klampfl, MacNeille 17]
- [Angelino, Larus-Stone, Alabi, Seltzer, Rudin '18]

# Building a new model



## Rule based

```
If  $f_1 < 0.1$ :  
    predict ✗  
else:  
    If  $f_2 > 0.4$  and  $f_2 < 0.6$ :  
        predict ○  
    else:  
        ...
```

## decision trees, rule lists, rule sets

- [Breiman, Friedman, Stone, Olshen 84]
- [Rivest 87]
- [Muggleton and De Raedt 94]
- [Wang and Rudin 15]
- [Letham, Rudin, McCormick, Madigan '15]
- [Hauser, Toubia, Evgeniou, Befurt, Dzyabura 10]
- [Wang, Rudin, Doshi-Velez, Liu, Klampfl, MacNeille 17]
- [Angelino, Larus-Stone, Alabi, Seltzer, Rudin '18]

# Which ones are the limitations of rule-based methods?

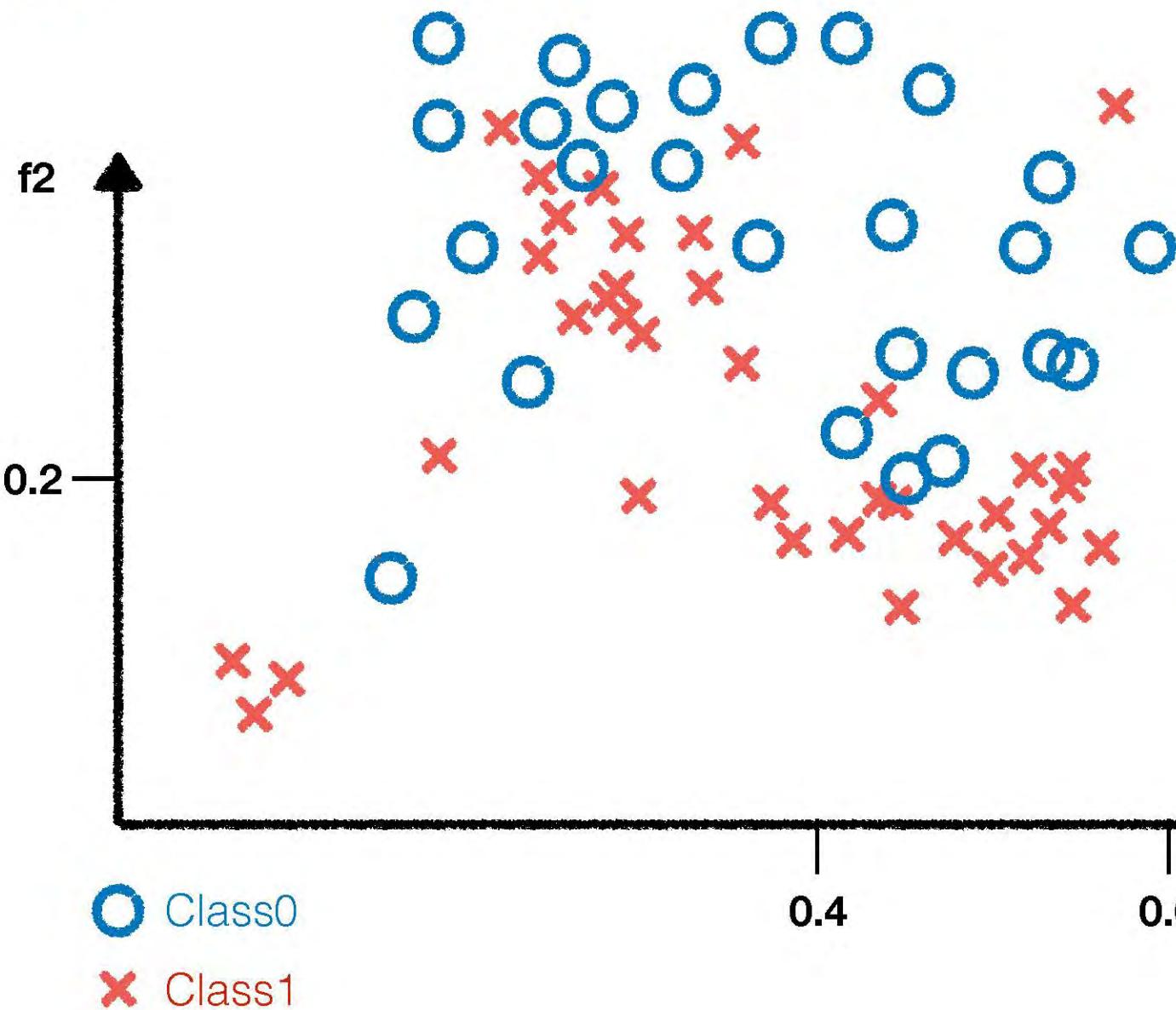
- A. It may not be as interpretable as you may think
- B. It only works if the original features are interpretable
- C. The data might not cluster
- D. None of the above

# Which ones are the limitations of rule-based methods?

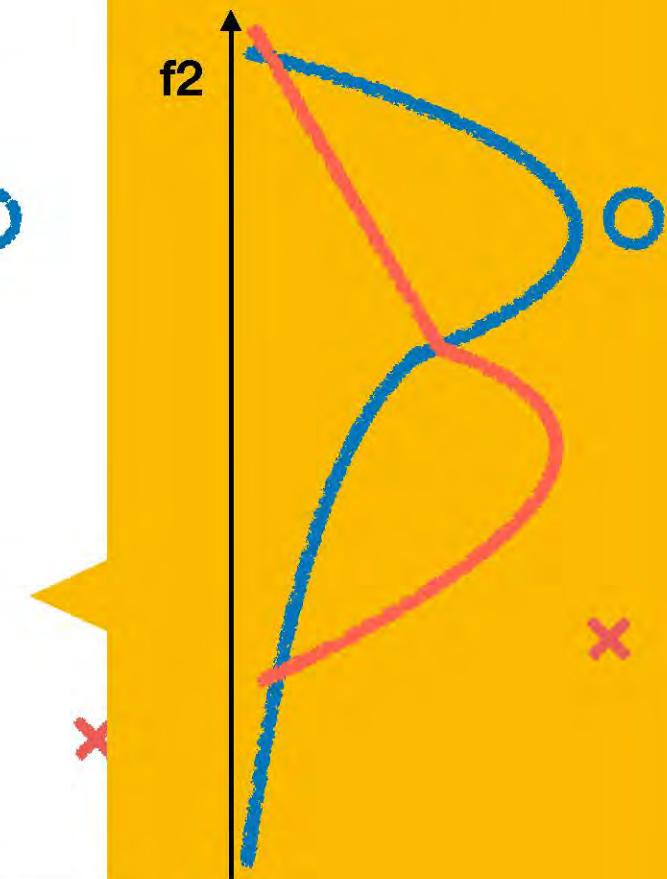
- Depth/Length of the tree might be too big
- Complexity of rules might be high
- Might not work for audio/images/embeddings

- A. It may not be as interpretable as you may think
- B. It only works if the original features are interpretable
- C. The data might not cluster
- D. None of the above

# Building a new model

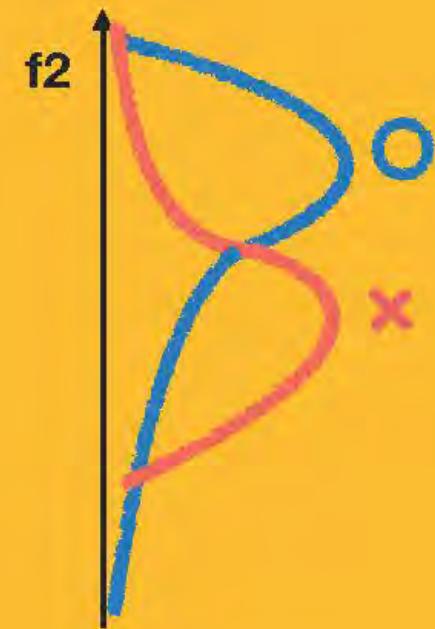
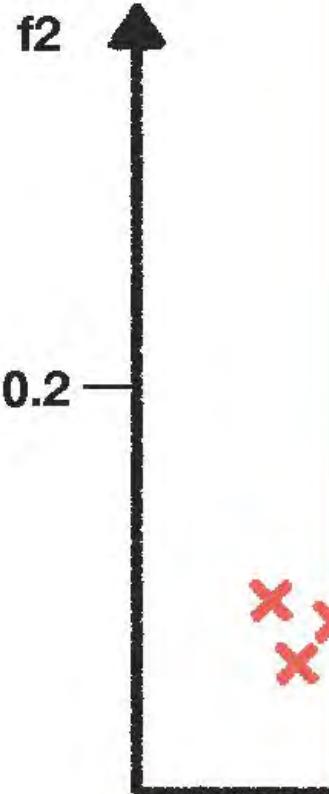


Fit a simpler function  
for each feature



# Building a new model

Fit a simpler function



Linear model

$$y = \beta_0 + \beta_1 x_1 + \dots + \beta_n x_n$$

generalized linear model

$$g(y) = \beta_0 + \beta_1 x_1 + \dots + \beta_n x_n$$

○ Class0

generalized additive model

$$g(y) = f_1(x_1) + \dots + f_n(x_n)$$

✗ Class1

Table from [Gehrke et al. '12]

# Building a new model



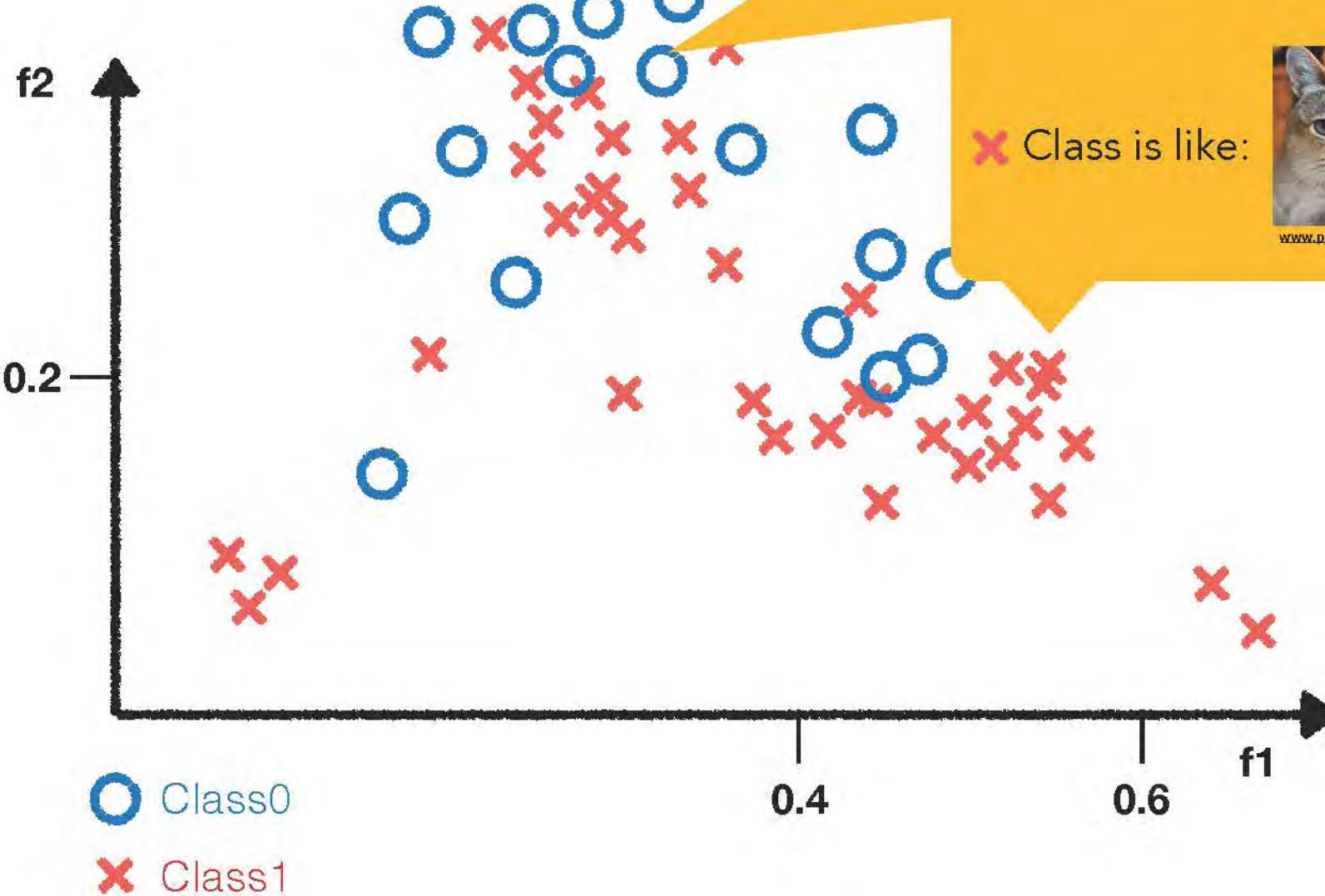
Example based



[www.guideddogs.org](http://www.guidedogs.org)

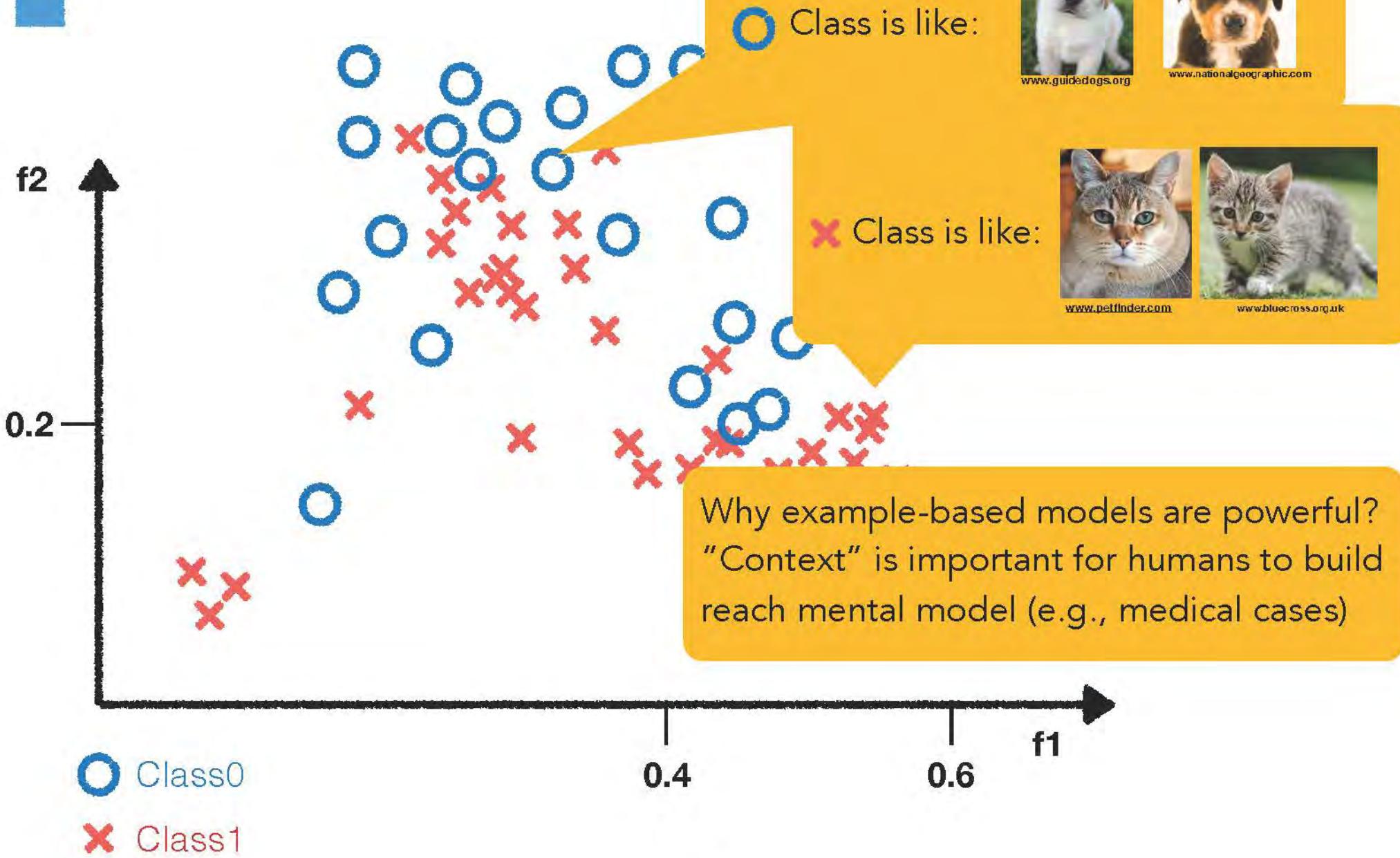


[www.nationalgeographic.com](http://www.nationalgeographic.com)



- [Frey, Dueck '10]
- [Yen, Malioutov , Kumar '16]
- [Arnold , El-Saden , Bui , Taira '10]
- [Floyd , Aha '16]
- [Homem, et al. '16]
- [Jalali , Leake '15]
- [Reid , Tibshirani '16]
- [K. Rudin, Shah '16]
- [Koh, Liang '17]

# Building a new model



# Building a new model

## Interactive Bayesian Case Model (BCM) [K. Rudin, Shah'14]

### Cluster Prototypes and Subspaces

#### Demote from Prototype

```
def dotProduct(listA,listB):
    total=0
    for(a,b) in zip(listA,listB):
        product=a*b
        total+=product
    return total
```

id: 1

#### Demote from Prototype

```
def dotProduct(listA,listB):
    assert len(listA)==len(listB)
    return sum(a*b for(a,b) in zip(listA,listB))
```

id: 15

#### Demote from Prototype

```
def dotProduct(listA,listB):
    length=len(listA)
    iB=0
    total=0
    while iB<length:
        total+=int(listA[iB])*int(listB[iB])
        iB+=1
```

id: 62

Most prototypical homework submission in cluster 3.

#### Promote to Prototype

```
def dotProduct(listA,listB):
    iB=0
    length=len(listA)
    total=0
    while iB<length:
        total+=listA[iB]*listB[iB]
        iB+=1
    return total
```

id: 45

#### Promote to Prototype

```
def dotProduct(listA,listB):
    listC=[]
    iB=0
    while iB<len(listA) and iB<len(listB):
        listC.append(listA[iB]*listB[iB])
        iB+=1
    return sum(listC)
```

id: 51

#### Promote to Prototype

```
def dotProduct(listA,listB):
    total=0
```

id: 54

Homework submissions in cluster 3

# Building a new model

## Interactive Bayesian Case Model (BCM) [K. Rudin, Shah'14]

Cluster Prototypes and Subspace

Demote from Prototype

```
def dotProduct(listA, listB):
    total=0
    for(a,b) in zip(listA, listB):
        product=a*b
        total+=product
    return total
```

Change  
important  
keywords in the  
cluster

Demote from Prototype

```
def dotProduct(listA, listB):
    assert len(listA)==len(listB)
    return sum(a*b for(a,b) in zip(listA, listB))
```

Demote from Prototype

```
def dotProduct(listA, listB):
    length=len(listA)
    iB=0
    total=0
    while iB<length:
        total+=int(listA[iB])*int(listB[iB])
        iB+=1
```

Promote to Prototype

id: 45

```
def dotProduct(listA, listB):
    iB=0
    length=len(listA)
    total=0
    while iB<length:
        total+=listA[iB]*listB[iB]
        iB+=1
    return total
```

Make this example  
prototype, and  
recluster!

Promote to Prototype

id: 52

```
def dotProduct(listA, listB):
    listC=[]
    iB=0
    while iB<len(listA) and iB<len(listB):
        listC.append(listA[iB]*listB[iB])
        iB+=1
    return sum(listC)
```

Promote to Prototype

id: 54

```
def dotProduct(listA, listB):
    +total-a
```

## Tool A

dot product

### Ready for Input

#### Cluster Prototypes and Subspaces

```
def dotProduct(listA, listB):
    total=0
    iB=0
    while iB<len(listA):
        product=listA[iB]*listB[iB]
        total+=product
        iB+=1
    return total
```

```
def dotProduct(listA, listB):
    total=0
    for(a,b) in zip(listA, listB):
        product=a*b
        total+=product
    return total
```

```
def dotProduct(listA, listB):
    if len(listA)!=len(listB):
        print 'length of A and B need to be the same'
        return None
```

#### Cluster members

Show all stacks

Promote to Prototype

```
def dotProduct(listA, listB):
    length=len(listA)
    total=0
    for i in range(0,length):
        product=listA[i]*listB[i]
        total=total+product
    return total
    print total
```

Promote to Prototype

```
def dotProduct(listA, listB):
    length=len(listA)
    iB=0
    total=0
    while iB<length:
        total=total+listA[iB]*listB[iB]
        iB+=1
    return total
```

# Which ones are the limitations of case-based models?

- A. The complexity of explanation is higher than that of data points
- B. There may not be a good representative examples
- C. Human may overgeneralize
- D. None of the above

# Which ones are the limitations of case-based models?

None of data points are representative!

- A. The complexity of explanation is higher than that of data points
- B. There may not be a good representative examples
- C. Human may overgeneralize
- D. None of the above

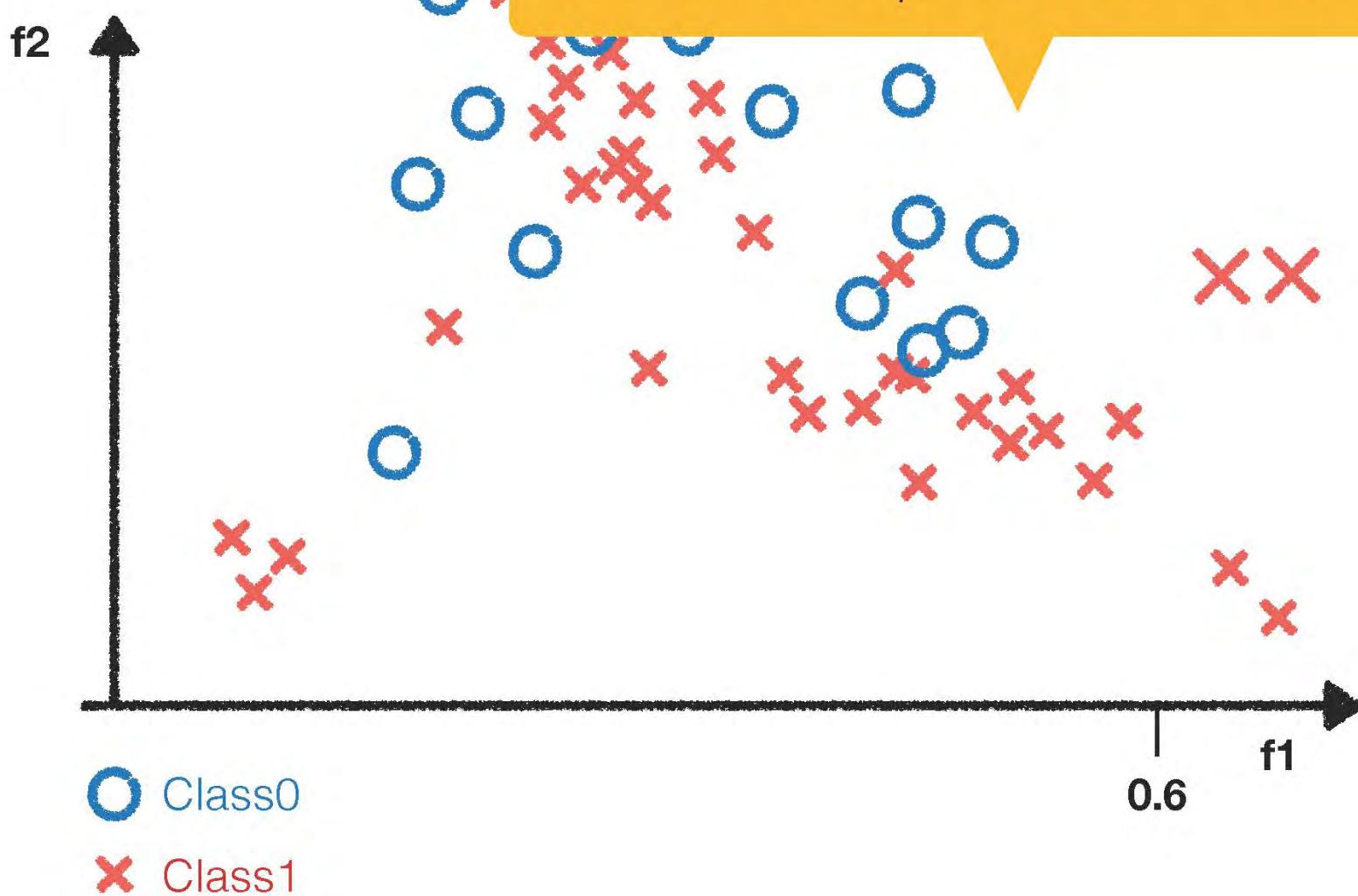


## Building a new model

Build a sparse model

$$y = a_0 + a_1x_1 + a_{21}x_{21} + a_{1002}x_{1002}$$

(all other  $a_i$ 's set to zero)



# Which ones are the limitations of sparsity methods?

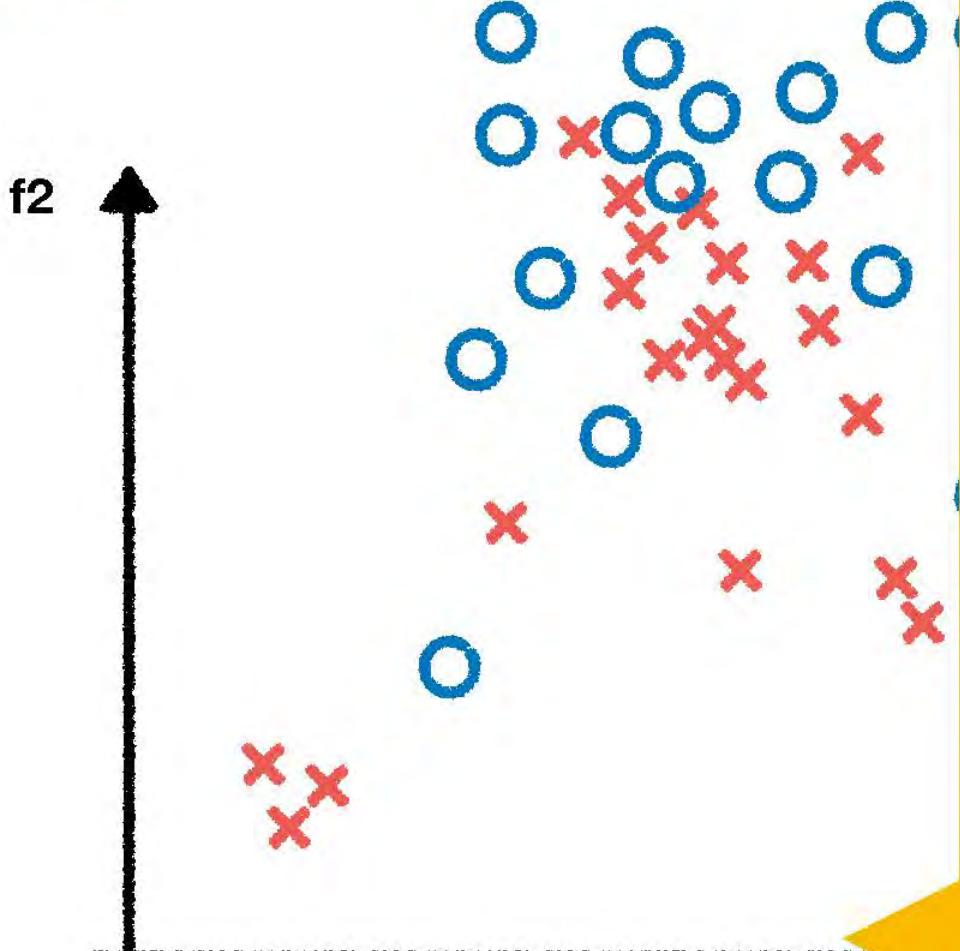
- A. The model may not be able to represent what it learned in a sparse fashion.
- B. There might be the case that only the collections of factors make more sense
- C. None of the above

# Which ones are the limitations of sparsity methods?

- A. The model may not be able to represent what it learned in a sparse fashion.
- B. There might be the case that only the collections of factors make more sense
- C. None of the above

“Sparsity is good, but not enough. Just because it is sparse, doesn’t mean it’s interpretable.” [Freitas ‘10]

# Building a new model

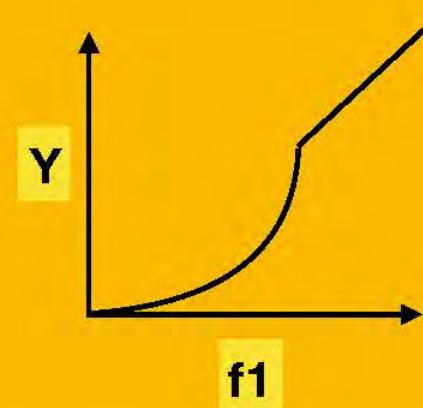


○ Class0

✗ Class1

## Monotonicity

Piecewise  
monotonic



Learn piecewise monotonic function within a user specified lattice (intervals) [Gupta et al. '16]

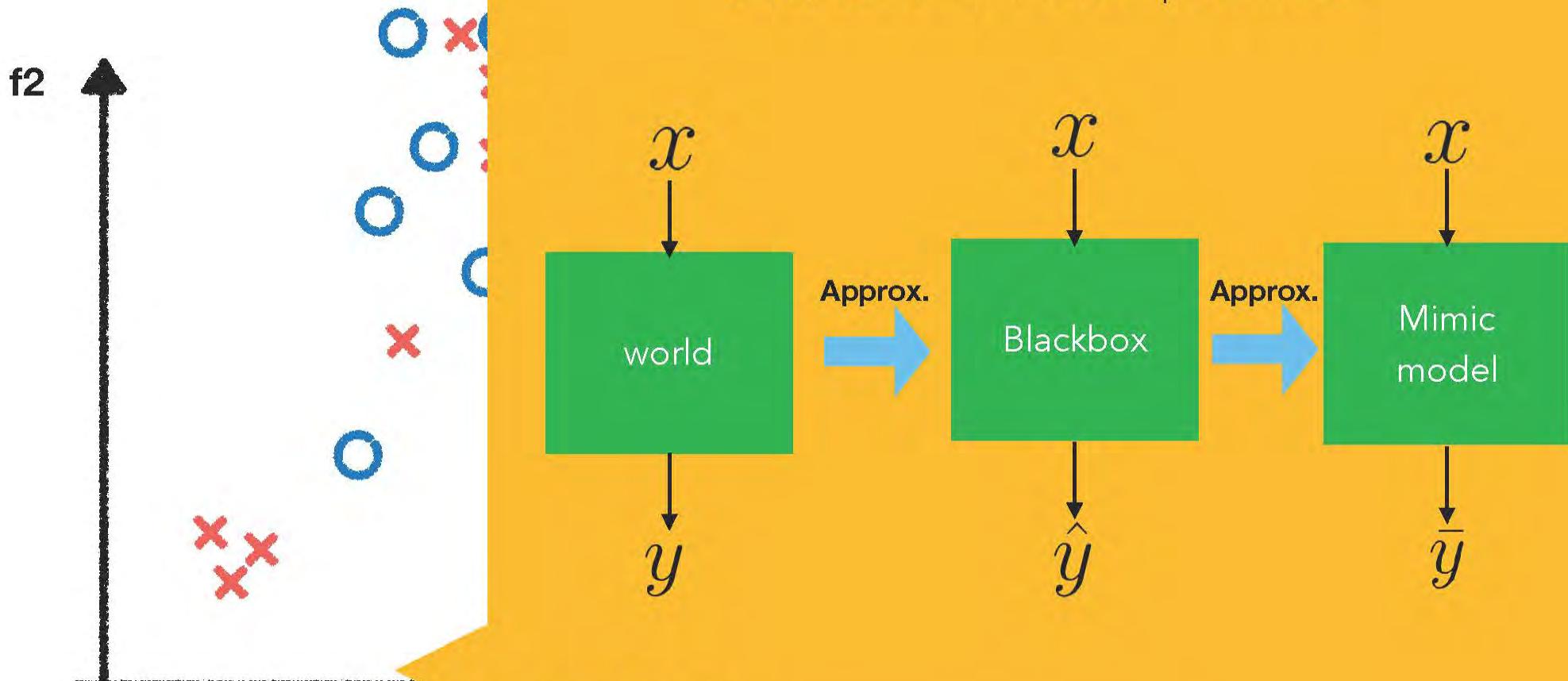
Monotonic neural networks by constraining weights  
[Neumann et al.'13, Riihimaki and Vehtari '10]

# Building a new model



## Mimic models, model distillation

Building a simpler model that walks, talks, barks like the complex model.



[Bucila et al. '06]

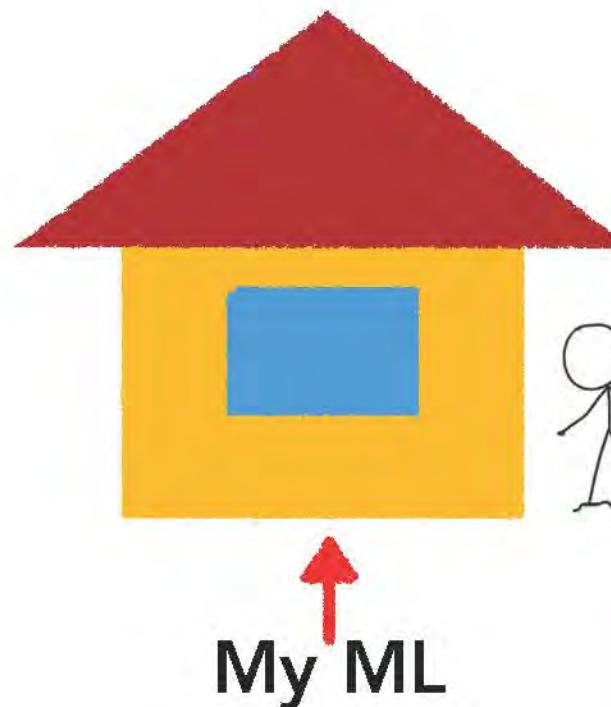
Do Deep Nets Really Need to be Deep? [Ba et al. '14]

Distilling the Knowledge in a Neural Network [Hinton et al. '15] [Frosst '17]

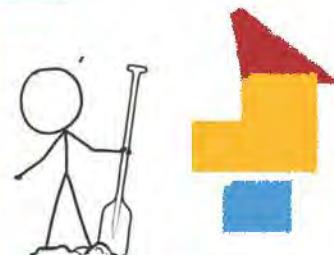
○ Class0

✗ Class1

# Types of interpretability methods



Before building  
any model



Building  
a new model

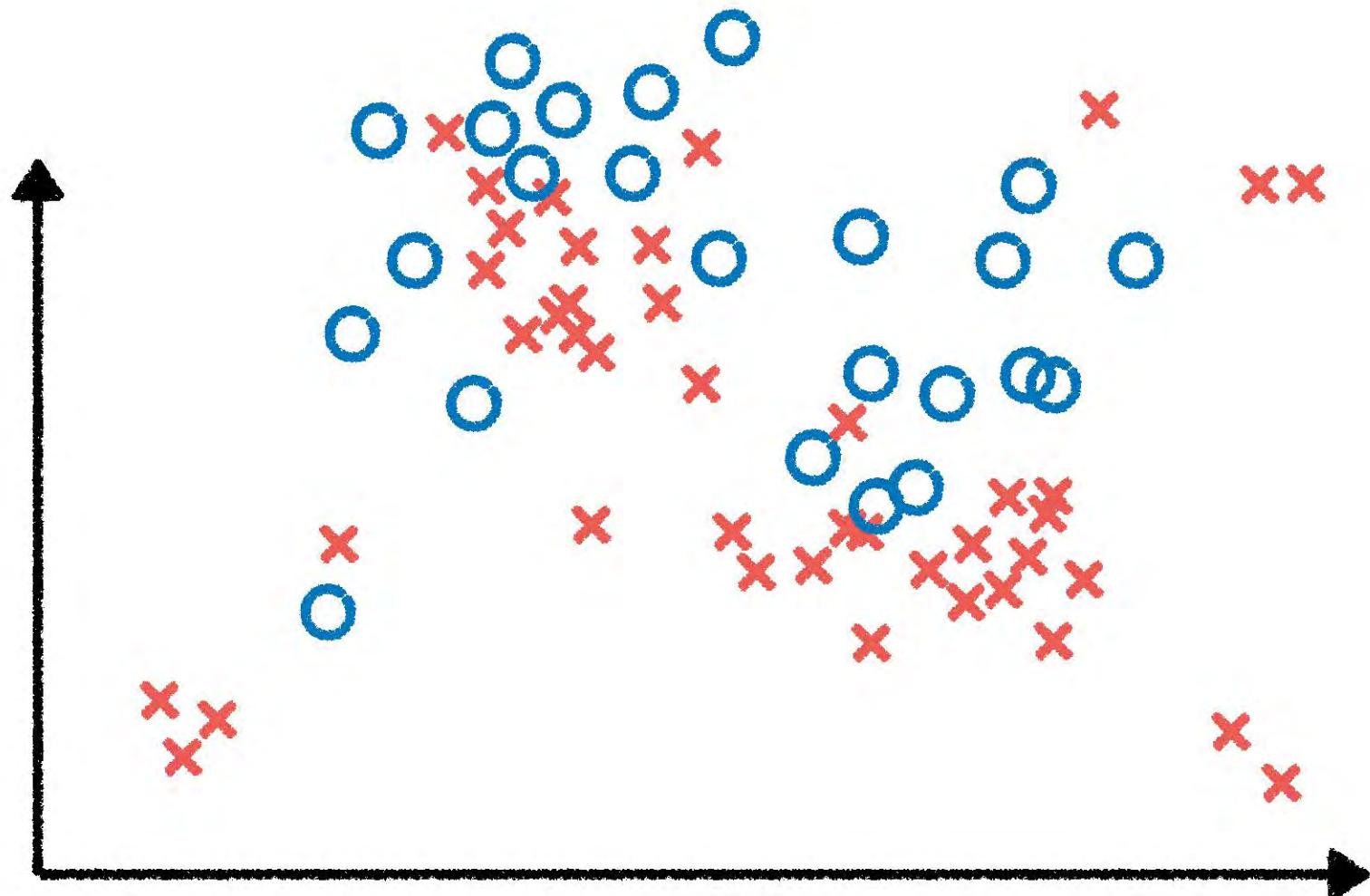


After  
building a model

Ablation test

Input-feature importance

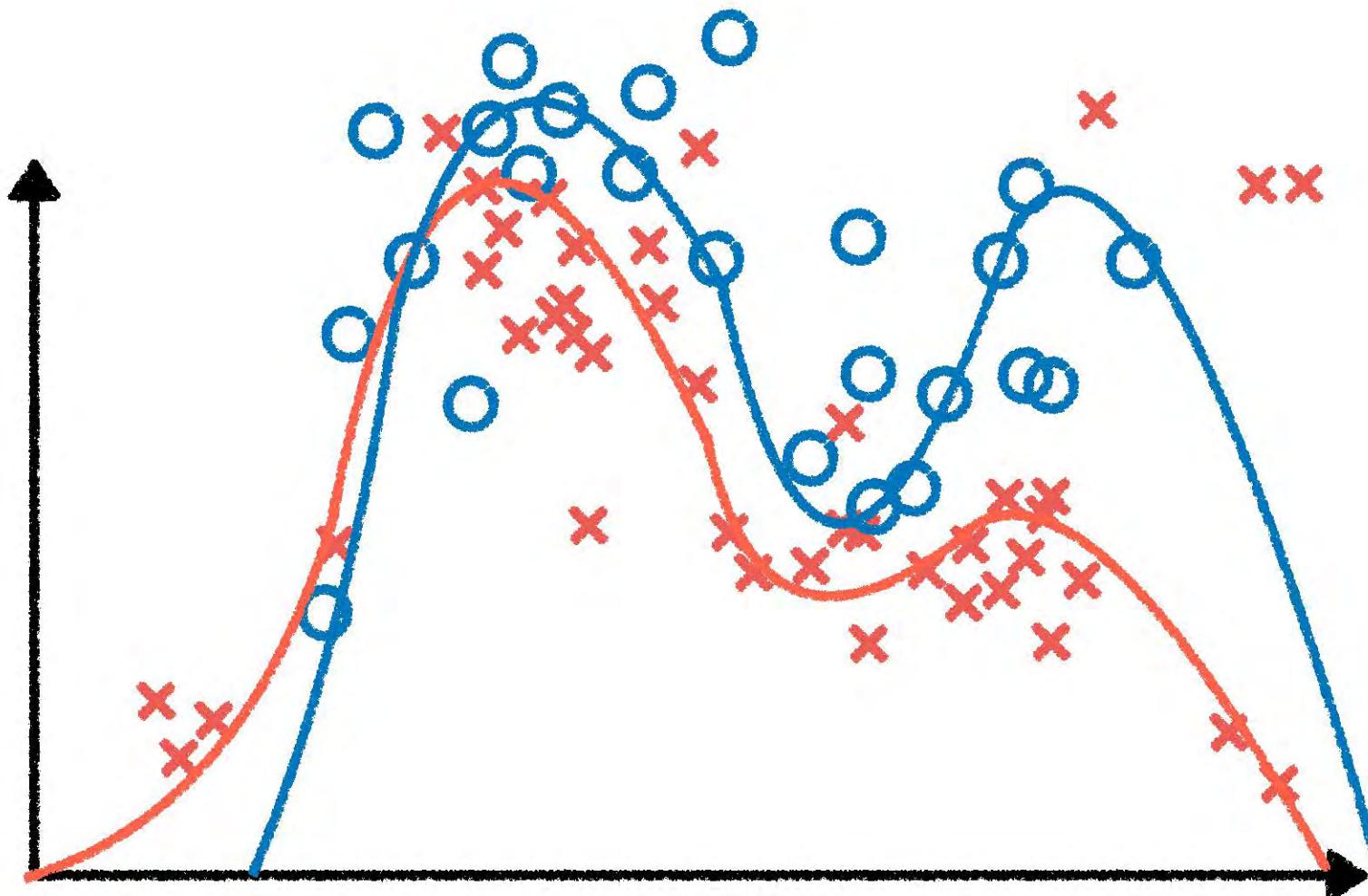
Concept importance



○ Class0

✗ Class1

# After building a model



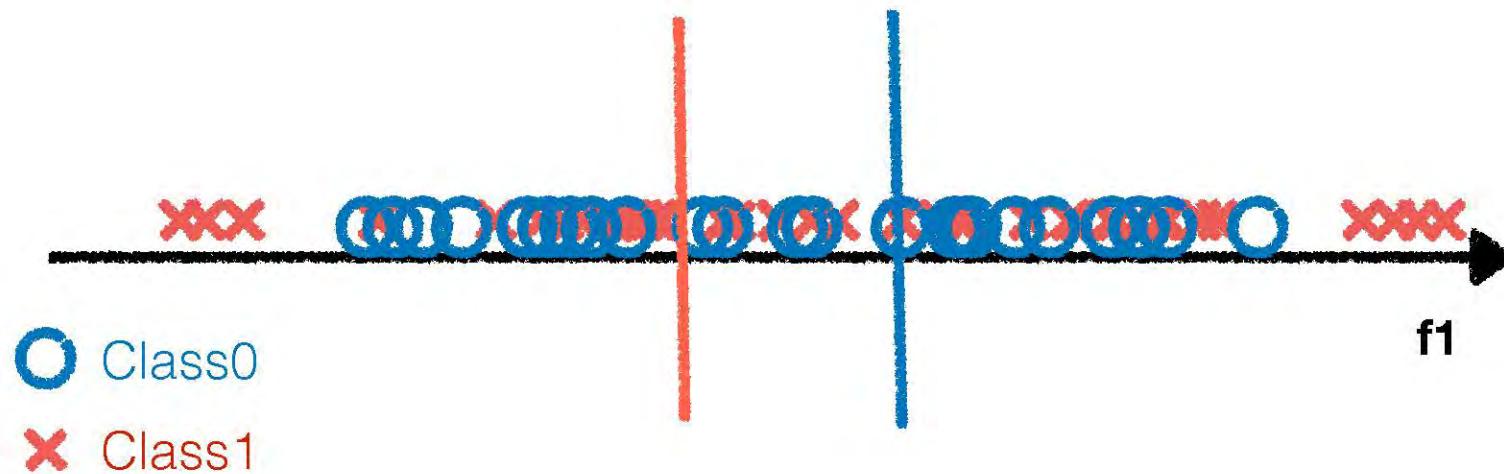
○ Class0

✗ Class1

# After building a model



1. Ablation test: train without that feature/data points and see the impact



# After building a model



1. Ablation test: train without that feature/data points and see the impact

**Smarter ablation** Influential functions [Koh et al.'17]

To classify this image:

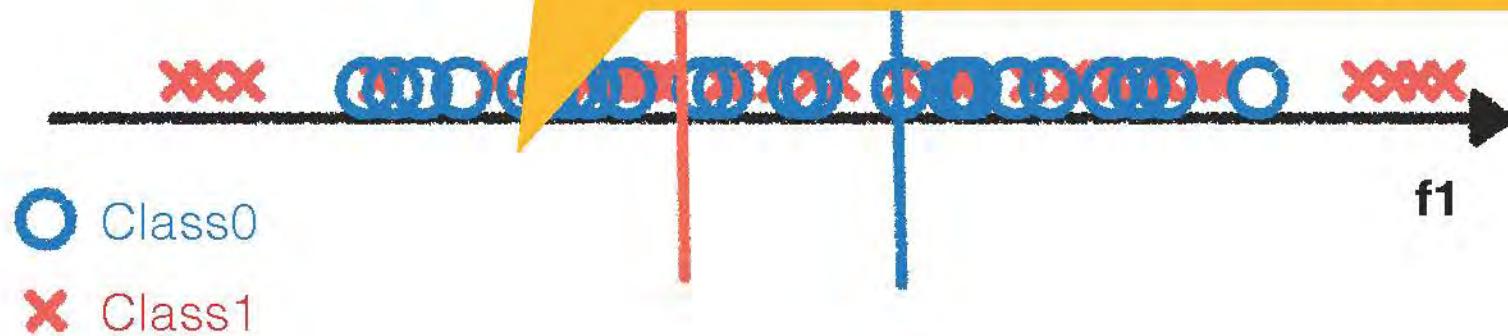


Model found these images most helpful

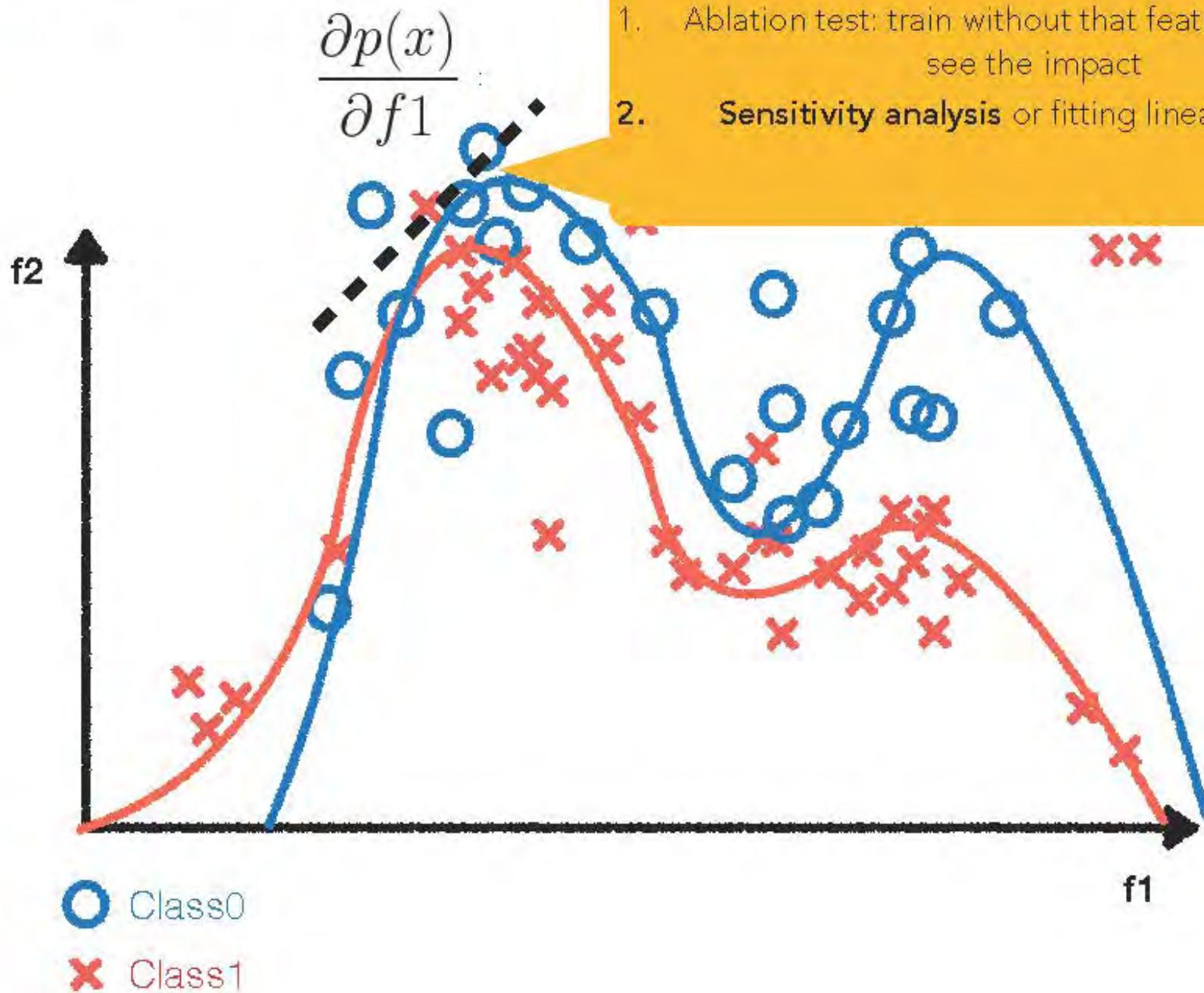
SVM



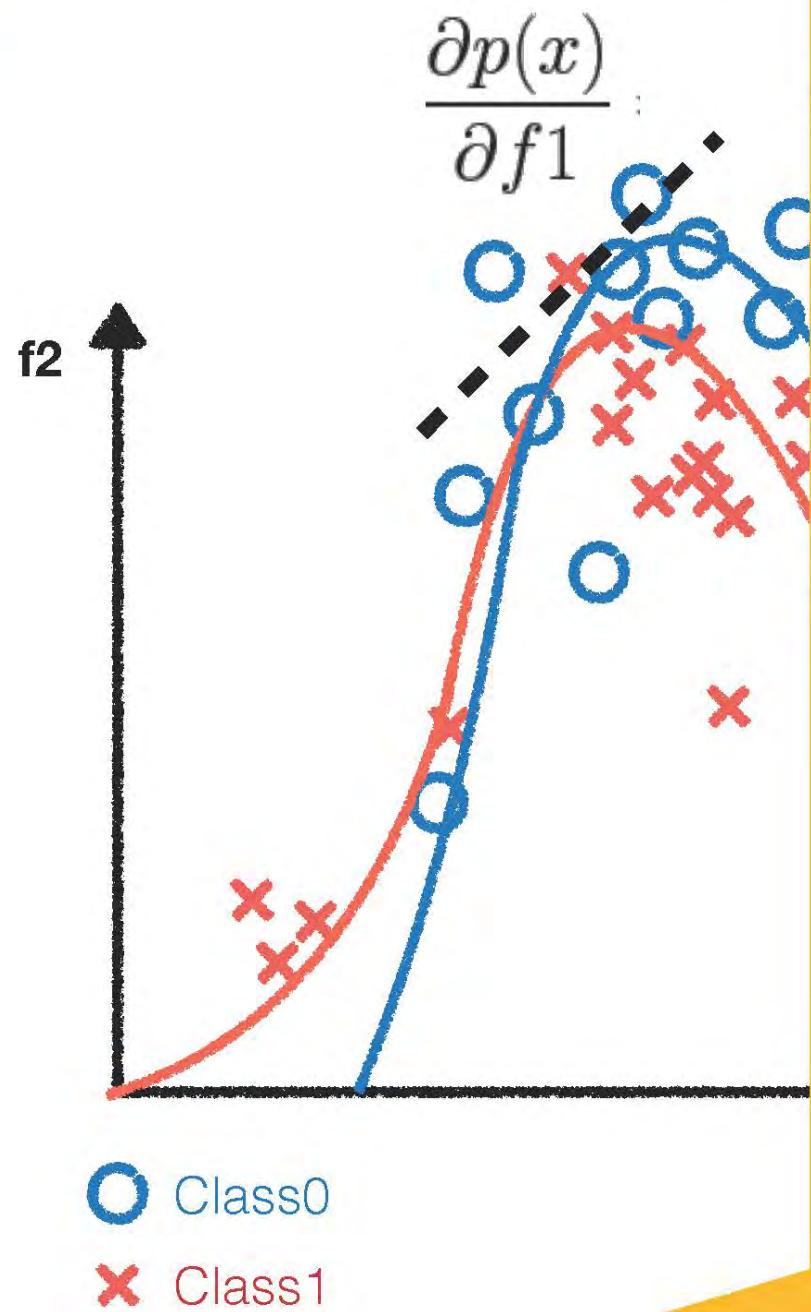
Inception



# After building a model



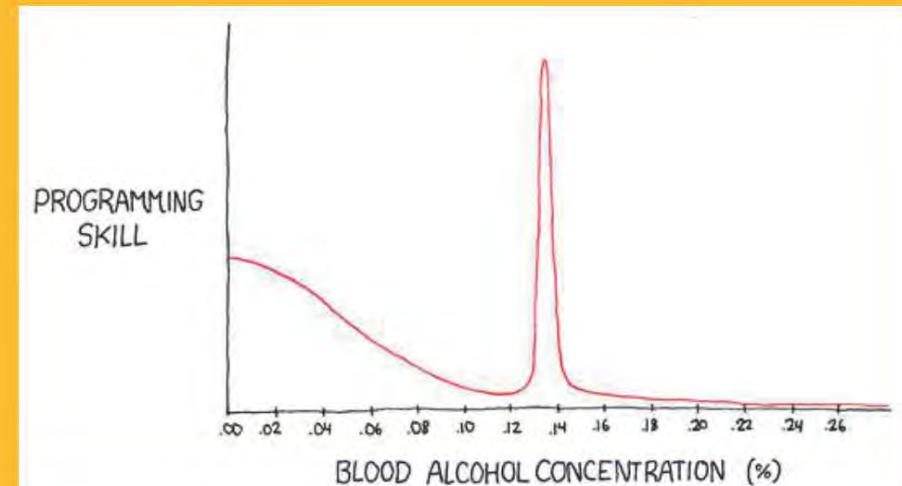
# After building a model



1. Ablation test: train without that feature/data and see the impact
2. **Sensitivity analysis** or fitting linear function

What happened to the output,  $\hat{y}$  if we perturb input  $x \rightarrow x + \epsilon$

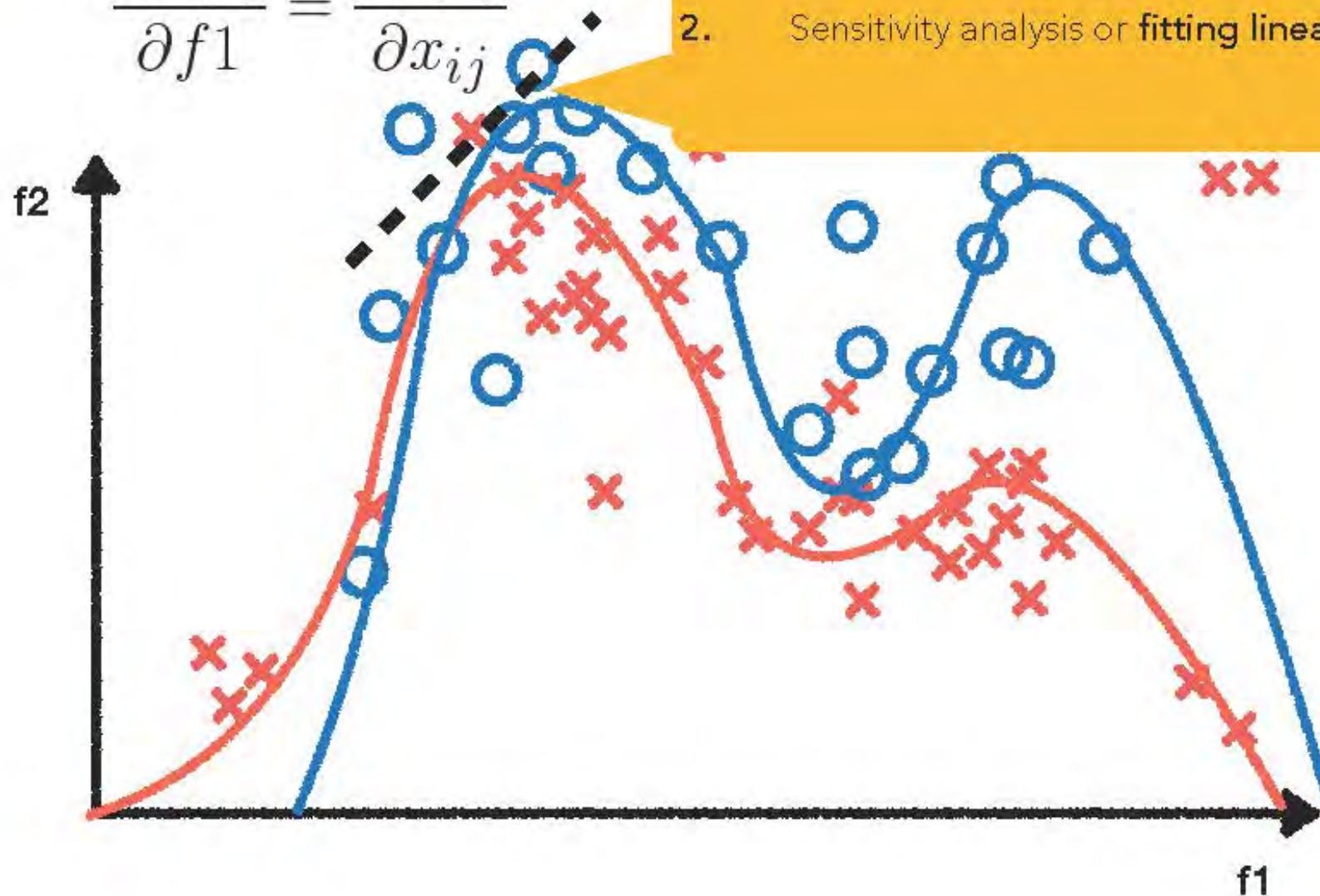
For nonlinear functions  $\hat{y} = f(x)$  higher order derivatives will get involved.



# After building a model

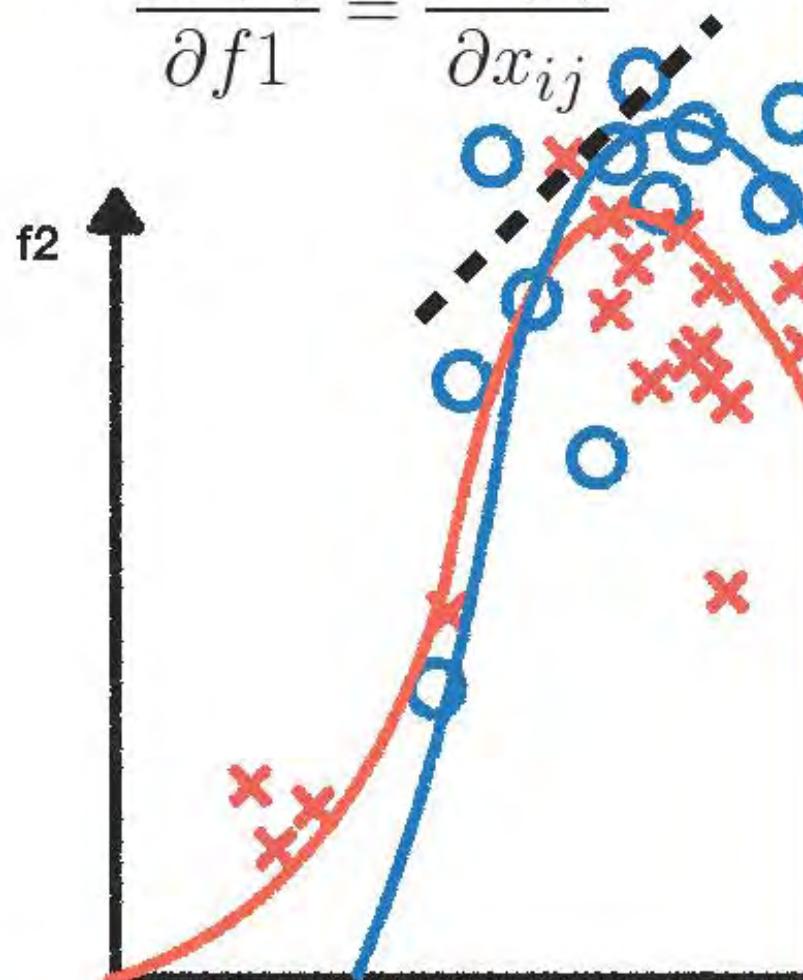


$$\frac{\partial p(x)}{\partial f_1} = \frac{\partial p(x)}{\partial x_{ij}}$$



# After building a model

$$\frac{\partial p(x)}{\partial f_1} = \frac{\partial p(x)}{\partial x_{ij}}$$



1. Ablation test: train without that feature/data and see the impact
2. Sensitivity analysis or **fitting linear function**

Sensitivity analysis on model  
[Ribeiro et al. '16]

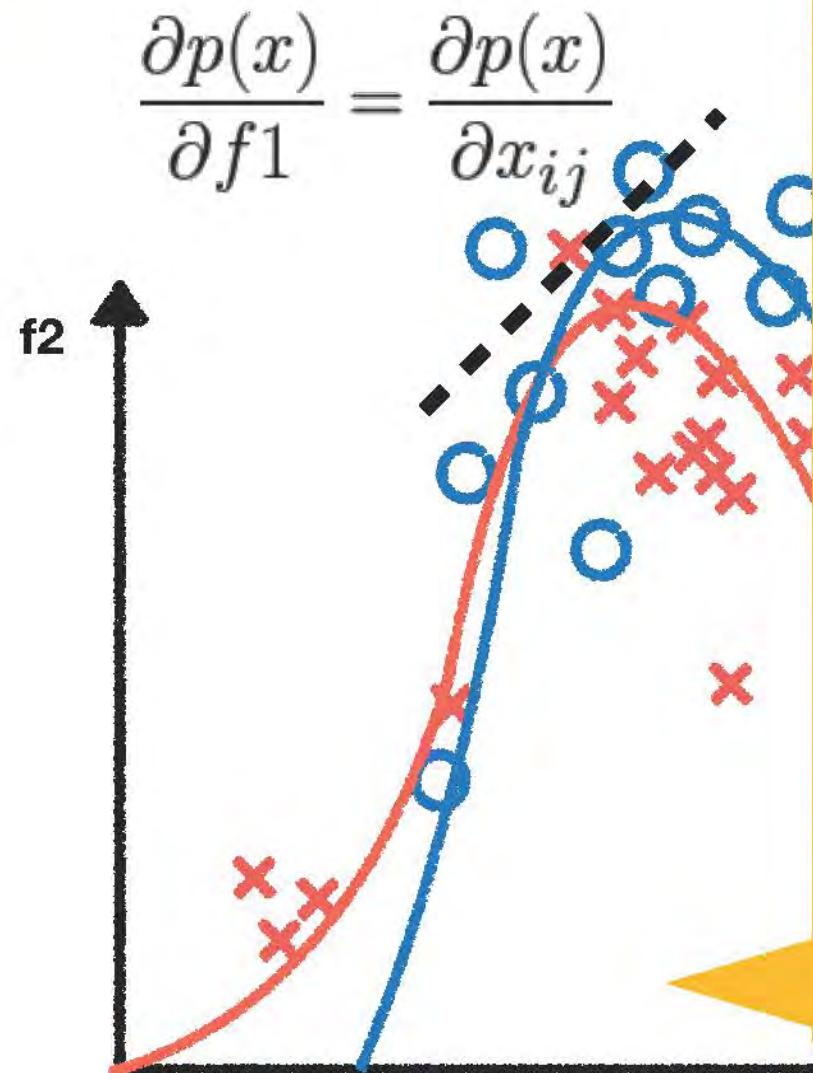
Want local explanation  
of the **+** data point



Locally fitted  
linear function

Many sensitivity analysis literature  
[Ribeiro et al. '16] [Simonyan et al., '13] [Li et al., '16]  
[Datta et al. '16] [Adler et al., '16] [Bach '15]

# After building a model



1. Ablation test: train without that feature/data and see the impact

2. **Sensitivity analysis or fitting linear function**

Integrated gradients [Sundararajan et al. 17]



SmoothGrad [Smilkov et al. 17]



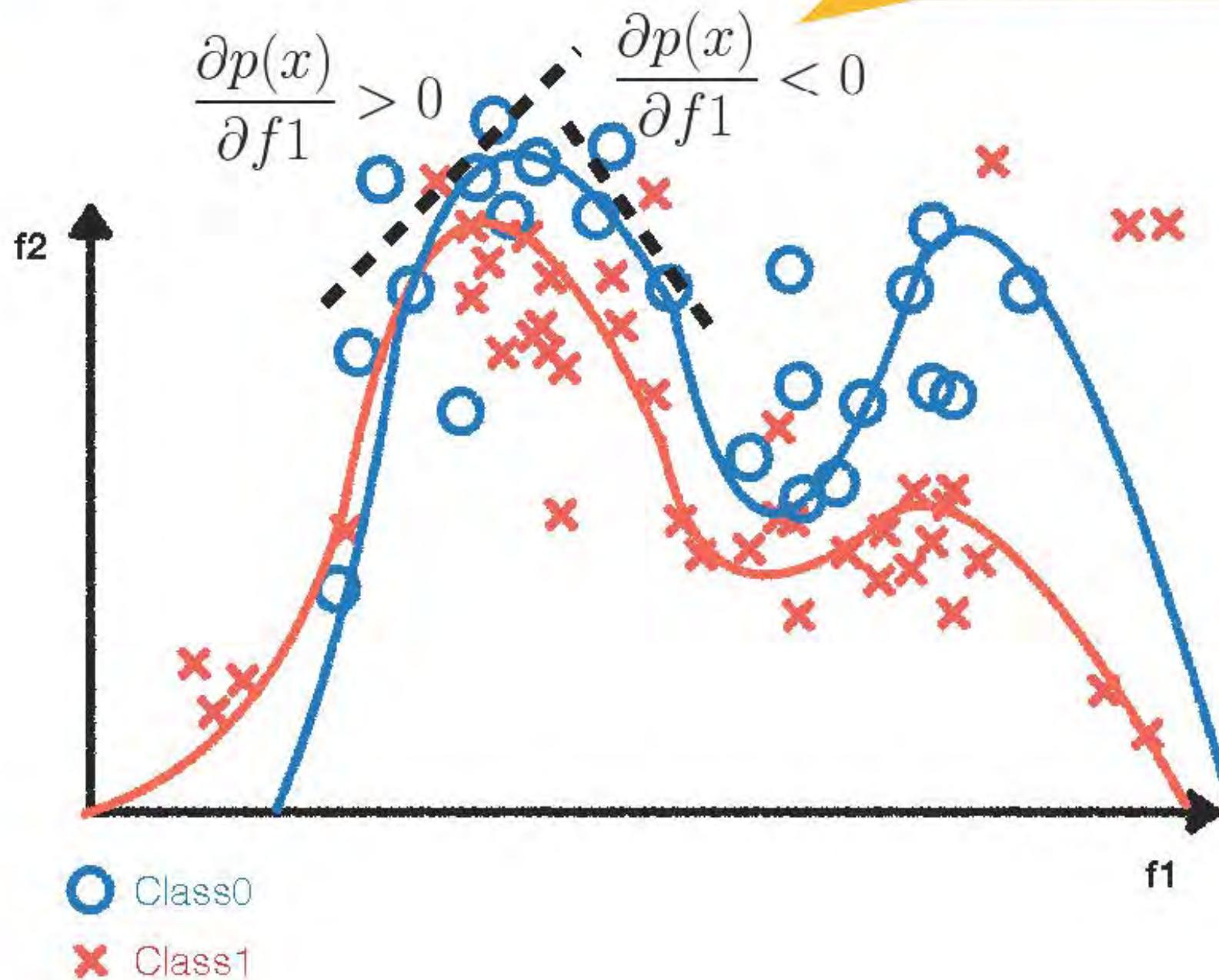
[Zeiler et al. '13] [Selvaraju et al. 16]

[Erhan 2009] [Springenberg, '14] [Shrikumar '17] and many more..

f1

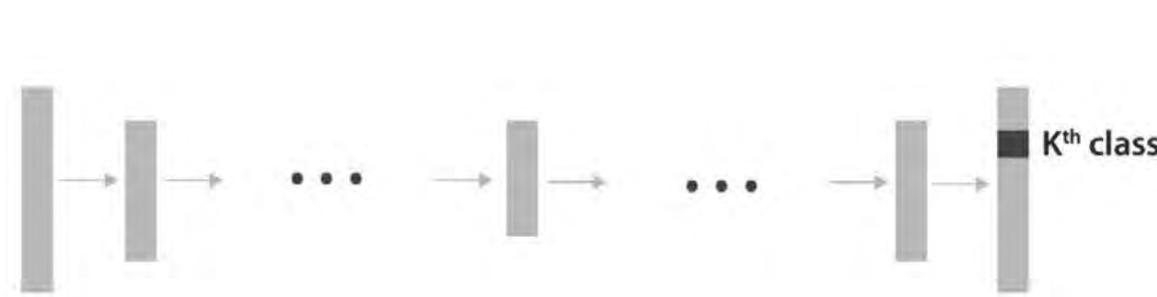
# After building a model

Local explanations



# Some confusing behaviors of saliency maps.

Original Image

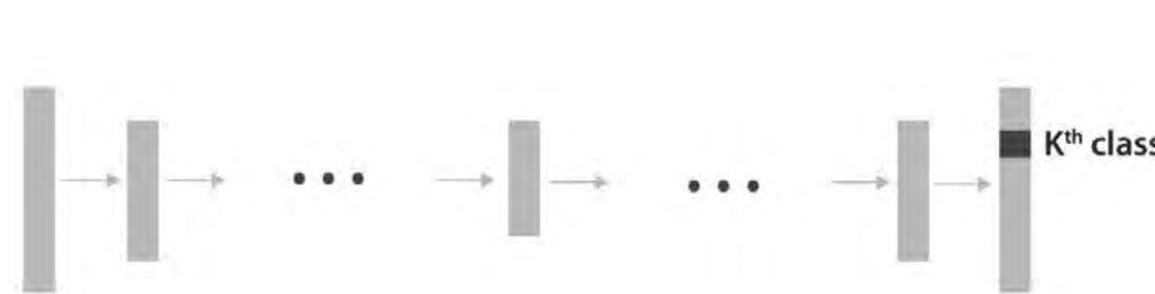


Saliency map

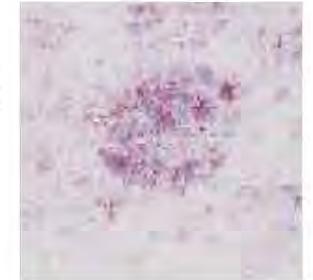


# Some confusing behaviors of saliency maps.

Original Image

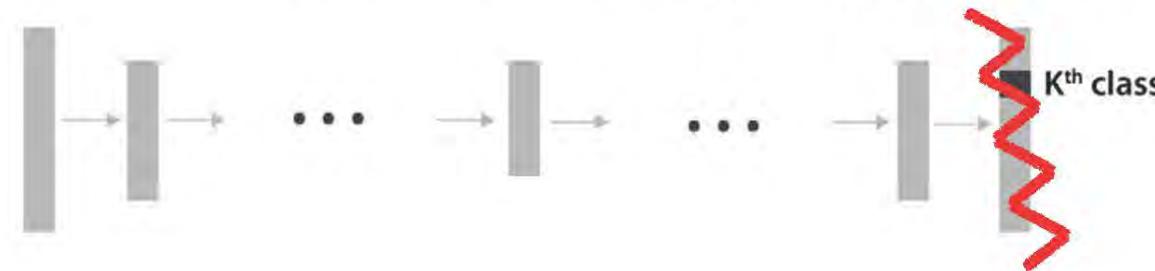


Saliency map



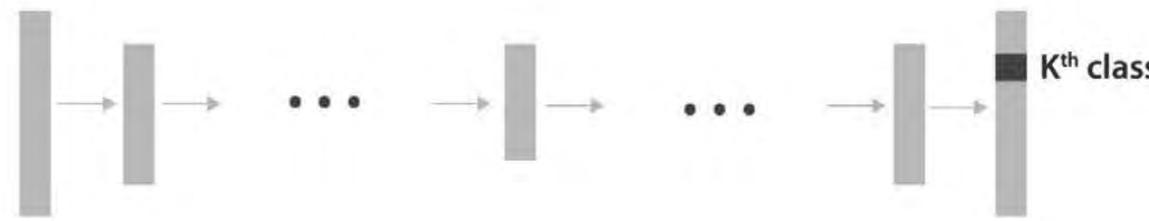
Randomized weights!

Network now makes garbage prediction.



# Some confusing behaviors of saliency maps.

Original Image



Saliency map

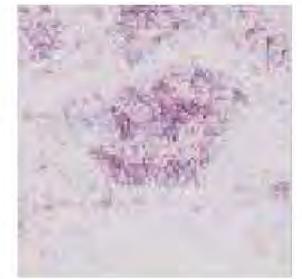
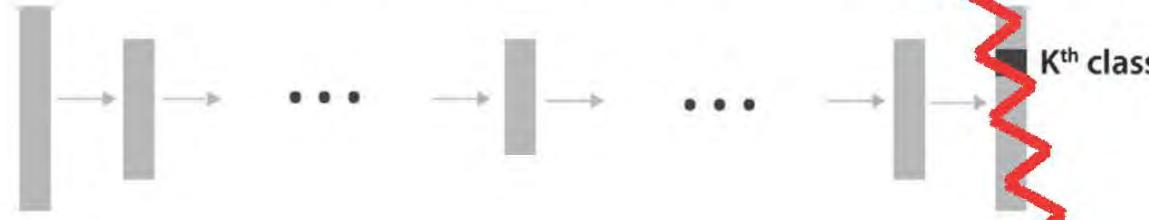


!!!!!!????!?

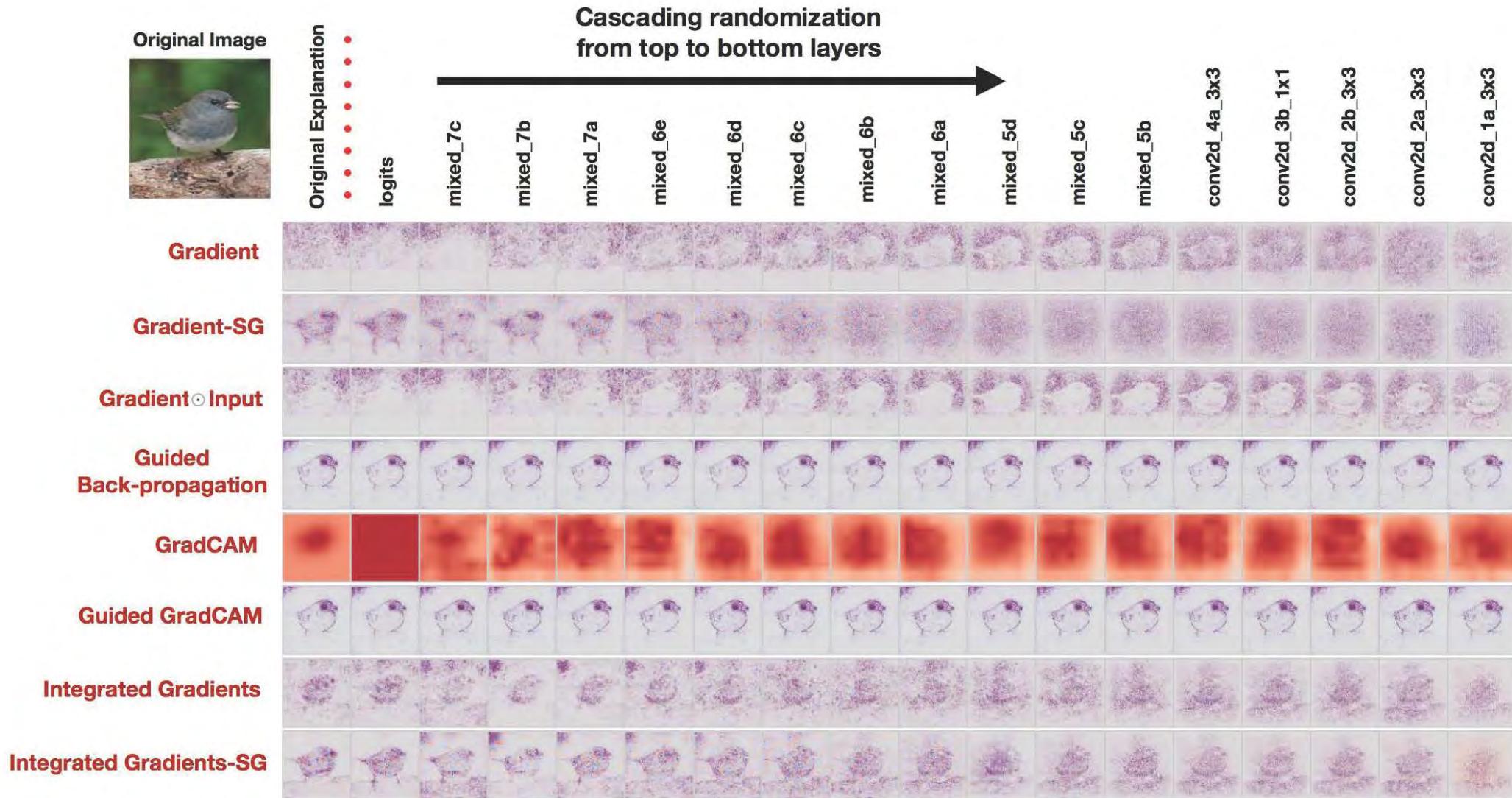
Randomized weights!

Network now makes garbage prediction.

Original Image



# Some saliency maps look similar when we randomize the network.



Which ones are the limitations of sensitivity analysis/gradient-based methods?

- A. It may not be truthful to the model
- B. The model may not allow sensitivity analysis
- C. Two local explanations may conflict
- D. The perturbed  $x$  may not be from the data distribution
- E. Interactions of sensitivity (changing two variables) is expensive

Which ones are the limitations of sensitivity analysis/gradient-based methods?

- A. It may not be truthful to the model
- B. The model may not allow sensitivity analysis
- C. Two local explanations may conflict
- D. The perturbed  $x$  may not be from the data distribution
- E. Interactions of sensitivity (changing two variables) is expensive



## After building a model



Common misunderstanding:  
An explanation IS how the model works.

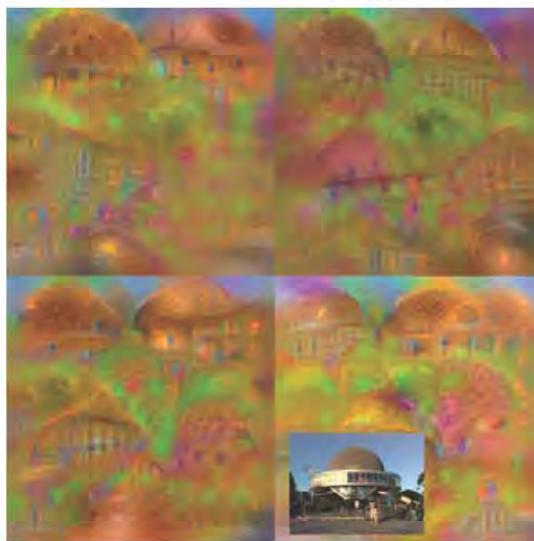
Local explanations may return  
contradictory explanations.

After building a model

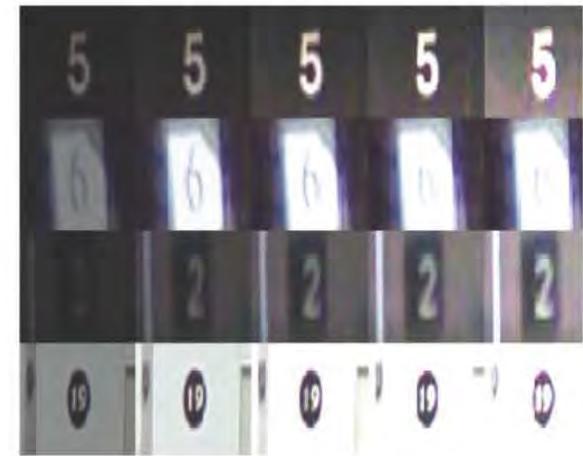


# Investigation on hidden layers

[Mahendran and Vedaldi '18]



[Adel et al. '18]

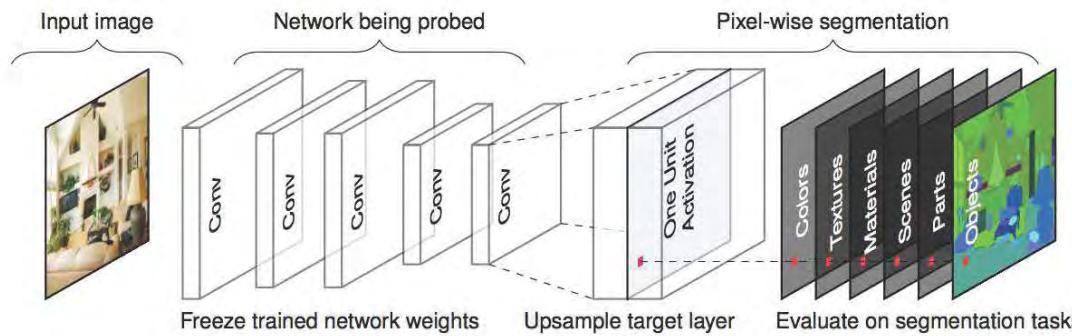


[Mordvintsev et al. '15]



[Bau and Zhou et al. '17]

[Zhou et al. '18]

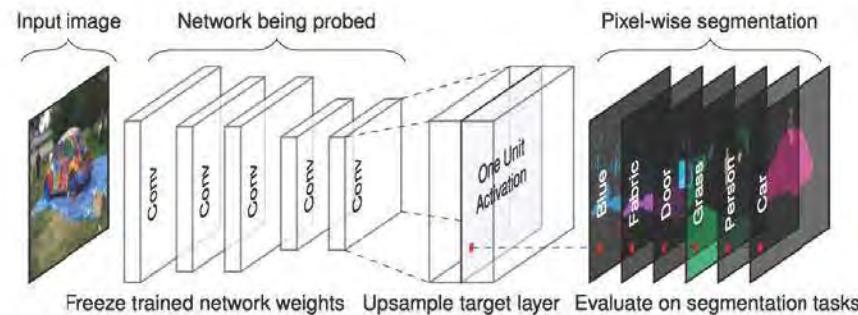


concept-based:  
NN learns concepts!

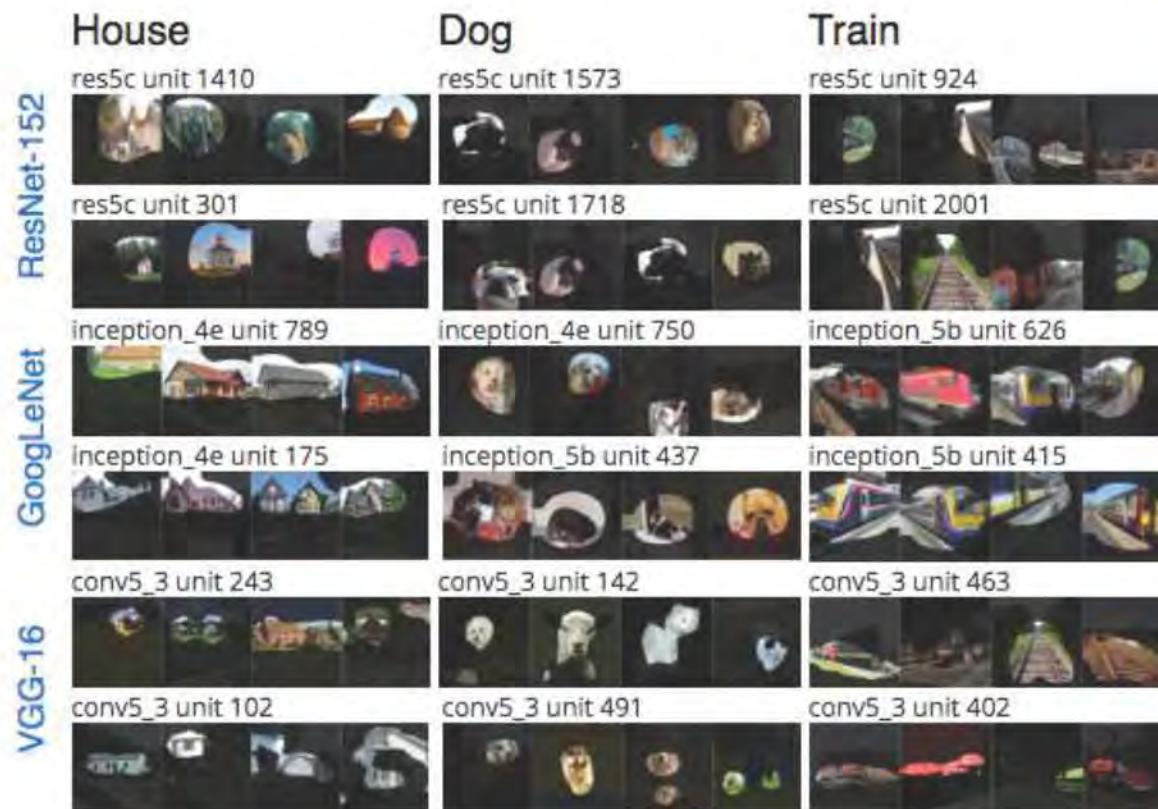


# Investigation on hidden layers

[Bau and Zhou et al. '17] [Zhou et al. '18]



concept-based:  
NN learns concepts!



After building a model

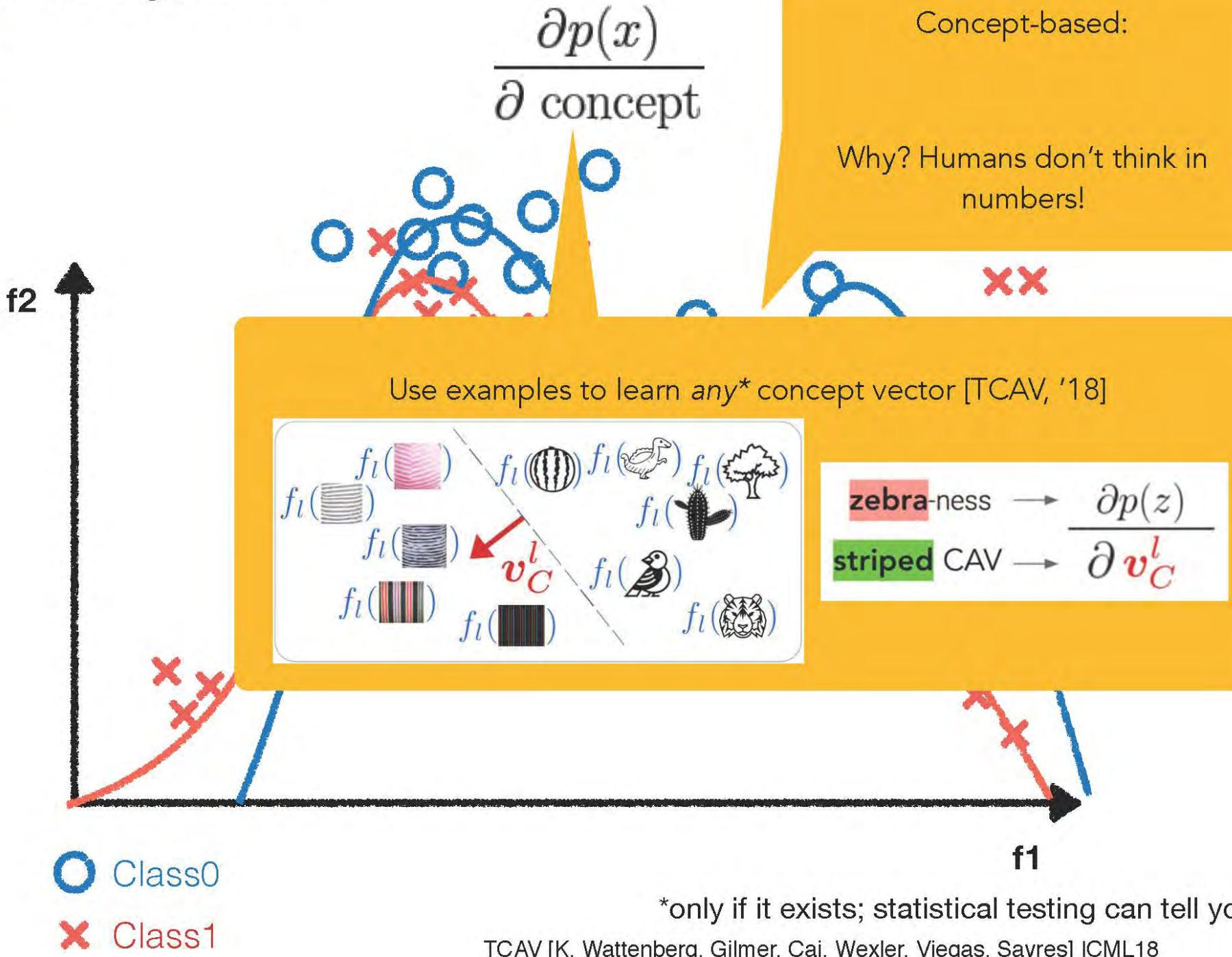
$$\frac{\partial p(x)}{\partial \text{concept}}$$



Concept-based:

Why? Humans don't think in numbers!

After building a model





# Defining concept activation vector (CAV)

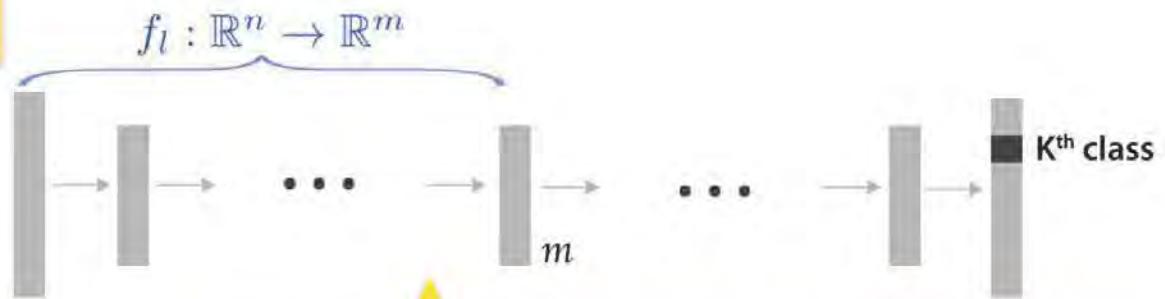
Inputs:

a



Examples of  
concepts

Random  
images

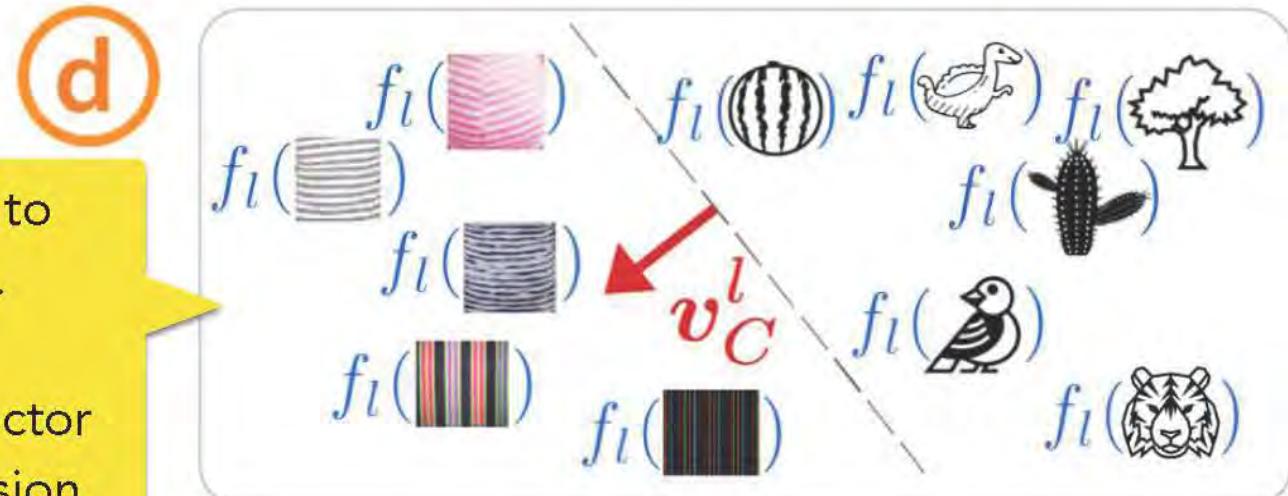
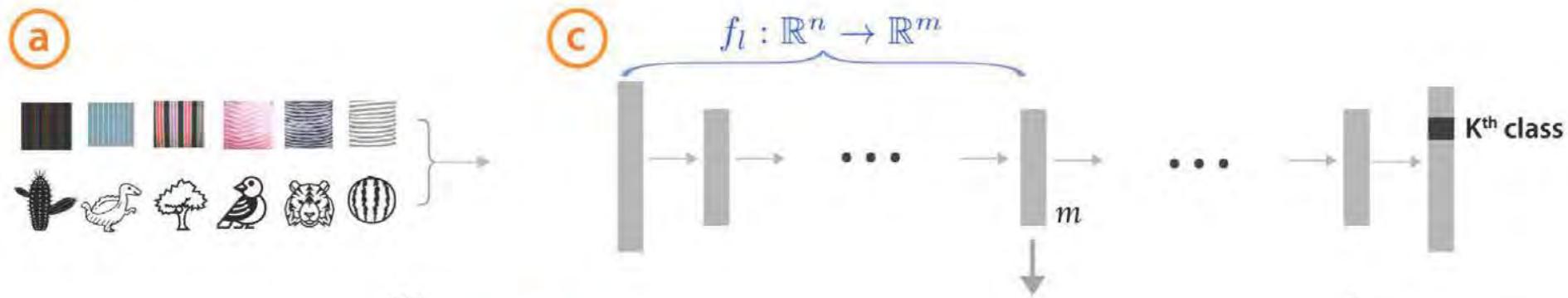


A trained network under investigation  
and  
Internal tensors



# Defining concept activation vector (CAV)

## Inputs:



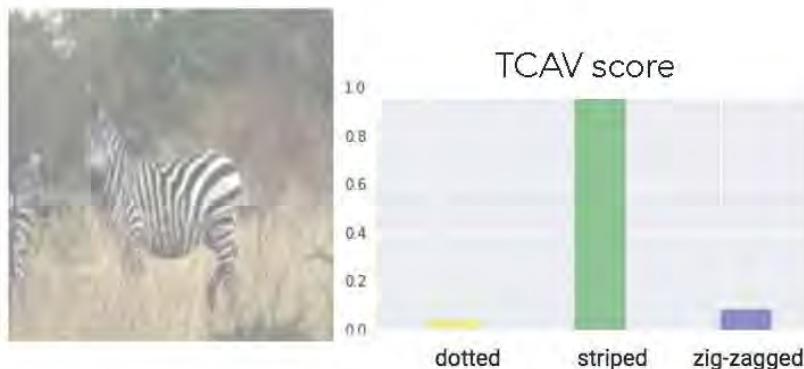
After building a model



TCAV core idea:

Derivative with CAV to get prediction sensitivity

TCAV



$$\begin{aligned}\text{zebra-ness} &\rightarrow \frac{\partial p(z)}{\partial \mathbf{v}_C^l} = S_{C,k,l}(\mathbf{x}) \\ \text{striped CAV} &\rightarrow \frac{\partial}{\partial \mathbf{v}_C^l} = S_{C,k,l}(\mathbf{x})\end{aligned}$$

Directional derivative with CAV

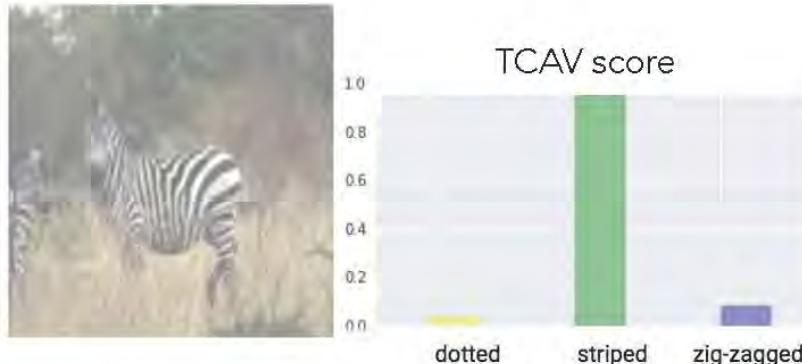
After building a model



TCAV core idea:

Derivative with CAV to get prediction sensitivity

TCAV



$$S_{C,k,l}(\text{zebra})$$

$$S_{C,k,l}(\text{horse})$$

$$S_{C,k,l}(\text{zebra})$$

$$S_{C,k,l}(\text{flamingo})$$

zebra-ness  $\rightarrow \frac{\partial p(z)}{\partial \mathbf{v}_C^l} = S_{C,k,l}(\mathbf{x})$

striped CAV  $\rightarrow \frac{\partial \mathbf{v}_C^l}{\partial \mathbf{v}_C^l} = S_{C,k,l}(\mathbf{x})$

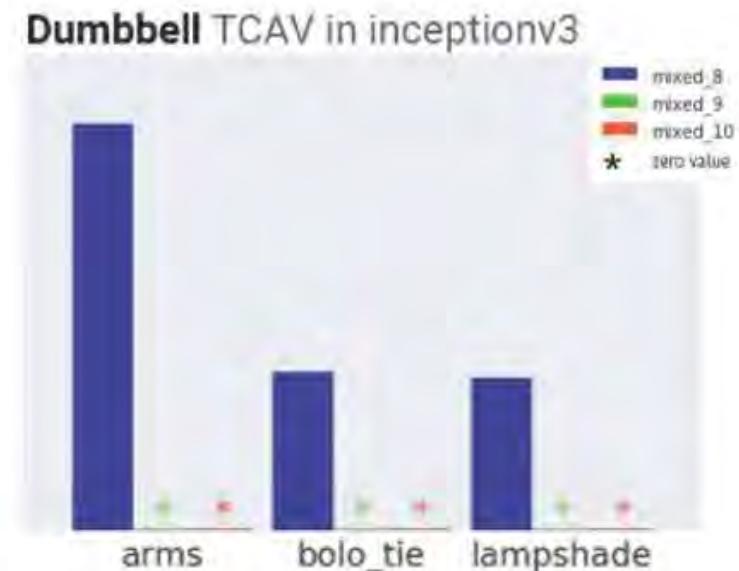
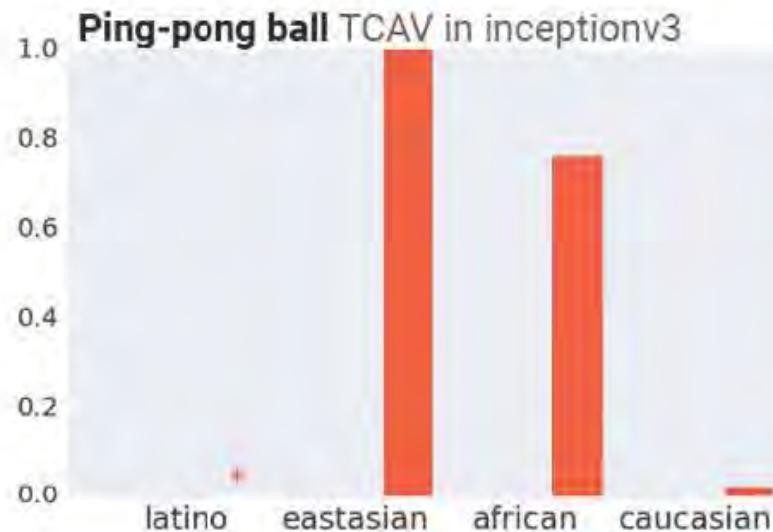
$$\text{TCAVQ}_{C,k,l} = \frac{|\{\mathbf{x} \in X_k : S_{C,k,l}(\mathbf{x}) > 0\}|}{|X_k|}$$

Directional derivative with CAV



# TCAV for widely used image prediction models

Quantitative confirmation to previously qualitative findings  
[Stock & Cisse, 2017]  
[Mordvintsev et al. '15]



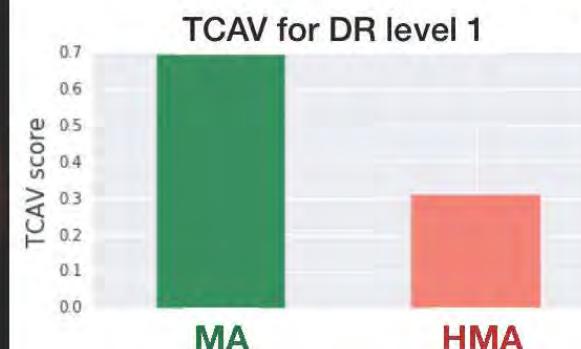
After building a model



# TCAV for Medical application: Diabetic Retinopathy

DR level 1

Med



TCAV shows the model is **inconsistent** with doctor's knowledge for classes when model is less accurate

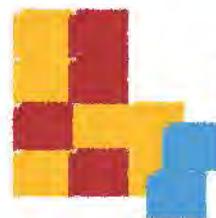
Green: domain expert's label on concepts belong to the level

Red: domain expert's label on concepts does not belong to the level

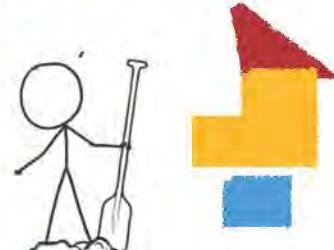
# Types of interpretability methods



My ML



Before building  
any model



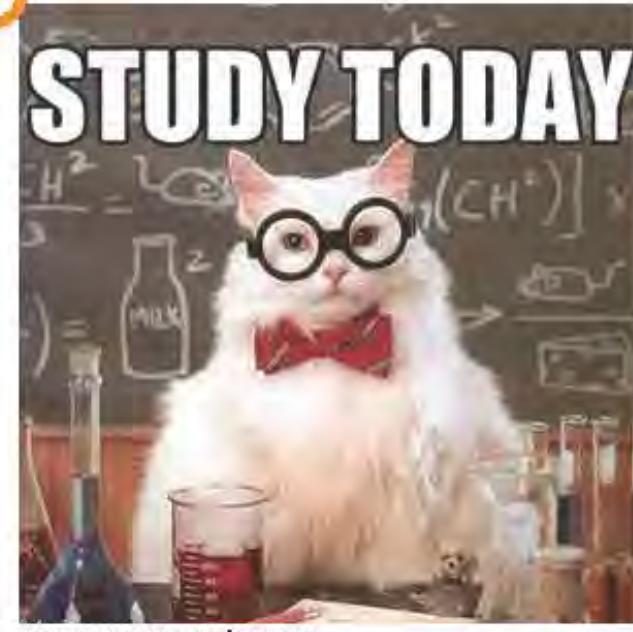
Building  
a new model



After  
building a model

# Agenda

- **When** and **why** interpretability
- **Overview** of interpretability methods.
- How to **Evaluate** interpretability methods.





# One way to evaluation interpretability...

"You know it when you see it"





# Spectrum of evaluation in machines learning



## Machine Learning

### Function-based

a variety of synthetic  
and standard  
benchmarks  
e.g., UCI datasets,  
imagenet



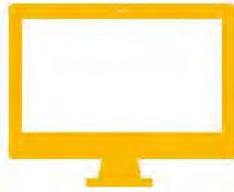
### Application-based

Backing up claims  
e.g., performance on a  
cool medical dataset,  
winning Go games



# Spectrum of evaluation in interpretable machine learning

Interpretable Machine Learning



Function-based

How sparse are  
the features?

Does it look  
reasonable?



Application-based

How much did  
we improve patient  
outcomes?

Do scientists find  
the explanations  
useful?

Quantitative Qualitative



# Spectrum of evaluation in interpretable machine learning

Interpretable Machine Learning



Function-based

How sparse are  
the features?

Does it look  
reasonable?

It's easy to formalize,  
optimize, and evaluate... but  
may not solve a real need.



Application-based

How much did  
we improve patient  
outcomes?

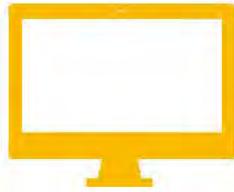
Do scientists find  
the explanations  
useful?

Quantitative Qualitative



# Spectrum of evaluation in interpretable machine learning

Interpretable Machine Learning



Function-based

How sparse are  
the features?

Does it look  
reasonable?



Application-based

How much did  
we improve patient  
outcomes?

Do scientists find  
the explanations  
useful?

Quantitative Qualitative



# Spectrum of evaluation in interpretable machine learning

Interpretable Machine Learning



Function-based

How sparse are  
the features?

Does it look  
reasonable?

It's real evaluation, but it's  
costly and hard to compare  
work A to B.



Application-based

How much did  
we improve patient  
outcomes?

Do scientists find  
the explanations  
useful?

Quantitative Qualitative



# Spectrum of evaluation in interpretable machine learning



Function-based

How sparse are  
the features?

Does it look  
reasonable?

cognition-based

What factor should  
change to change  
the outcome?

What are the  
discriminative  
features?

Application-based

How much did  
we improve patient  
outcomes?

Do scientists find  
the explanations  
useful?

Quantitative      Qualitative

Low cost

High cost

Low validity

High validity



# Spectrum of evaluation in interpretable machine learning



Factor-based

Prediction task: 1. Show explanations to humans.  
2. Ask humans what would the machine do.

Q. Which group does this new data belong to?

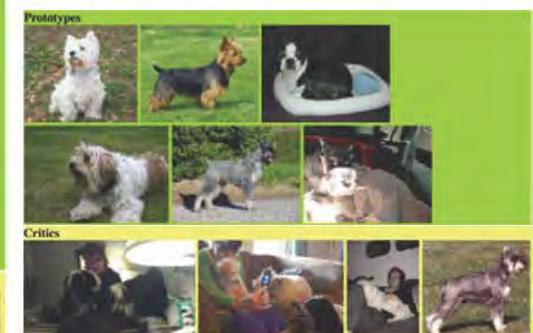


- Group A
- Group B

Group A



Group B



[K. 16]



# Spectrum of evaluation in interpretable machine learning



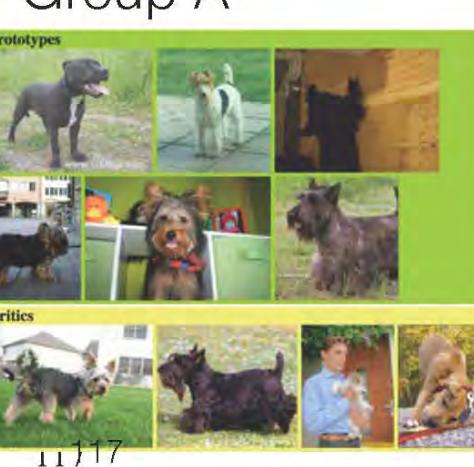
Factor-based

Validation task: 1. Show explanations to humans.  
2. Ask humans whether the machine's answer was correct.

Q. Machine thinks this image belongs to Group B. Is this correct?

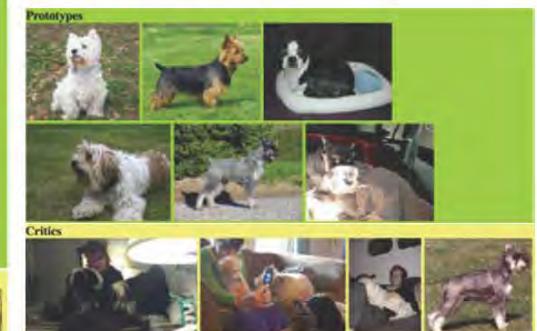


- Yes
- No



11/17

Group B



[K. 16]



# Spectrum of evaluation in interpretable machine learning

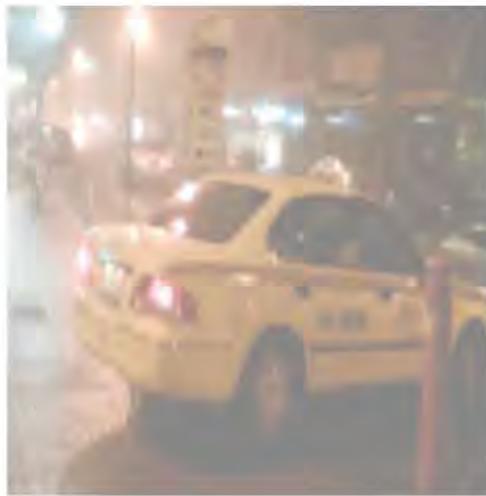


Factor-based

Formulate an experiment where you have the **ground-truth** when you can.



# An example of ground-truth experiment setup in TCAV



Goal: find out what was important for a prediction



# An example of ground-truth experiment setup in TCAV

caption  
concept



image  
concept

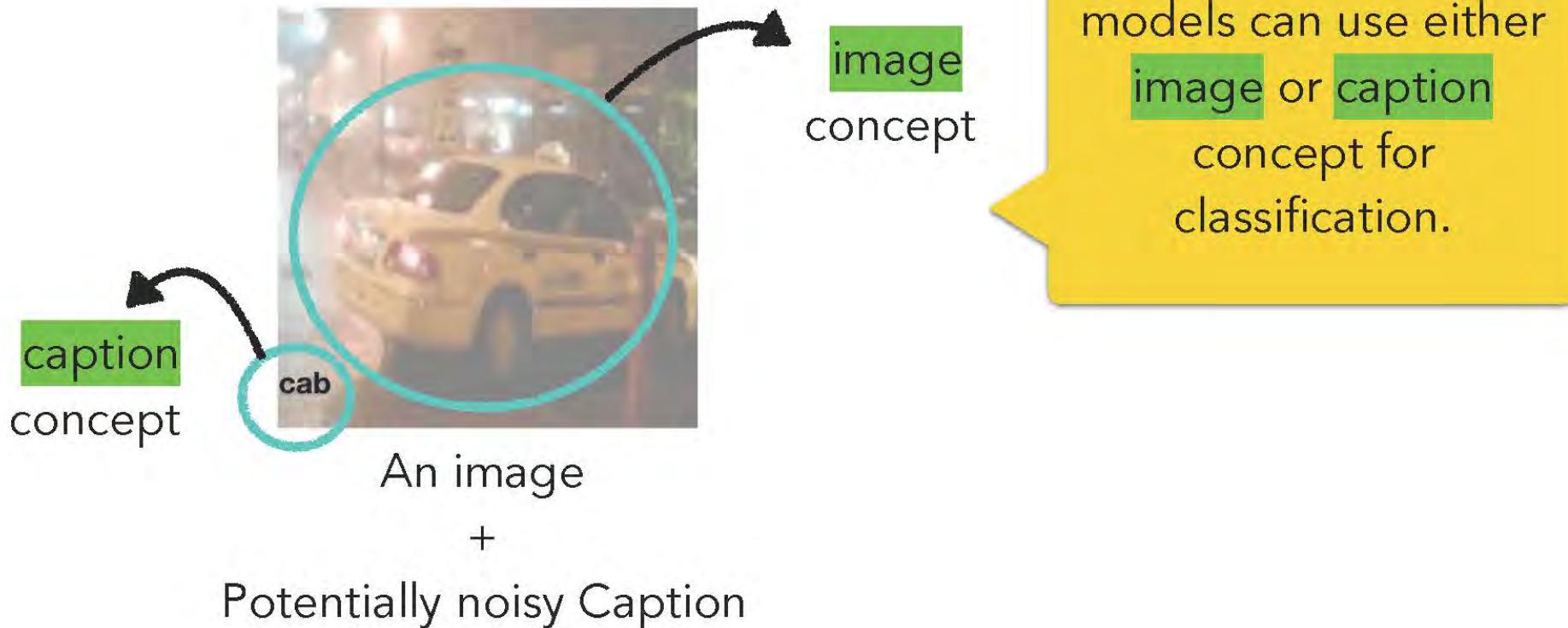
An image

+

Potentially noisy Caption



# An example of ground-truth experiment setup in TCAV





# An example of ground-truth experiment setup in TCAV



Image

+

Potentially noisy Caption



0% noisy      30% noisy      100% noisy      no captions

**Caption noise level in training set**



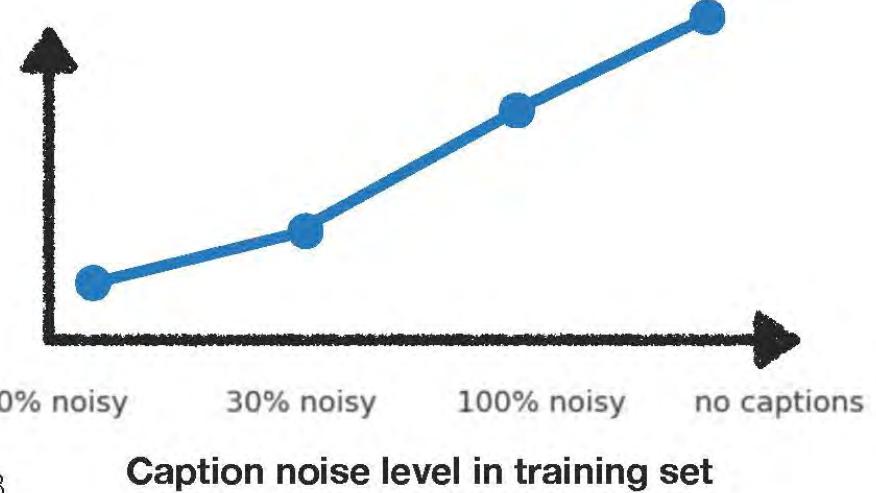
# An example of ground-truth experiment setup in TCAV



models can use either image or caption concept for classification.



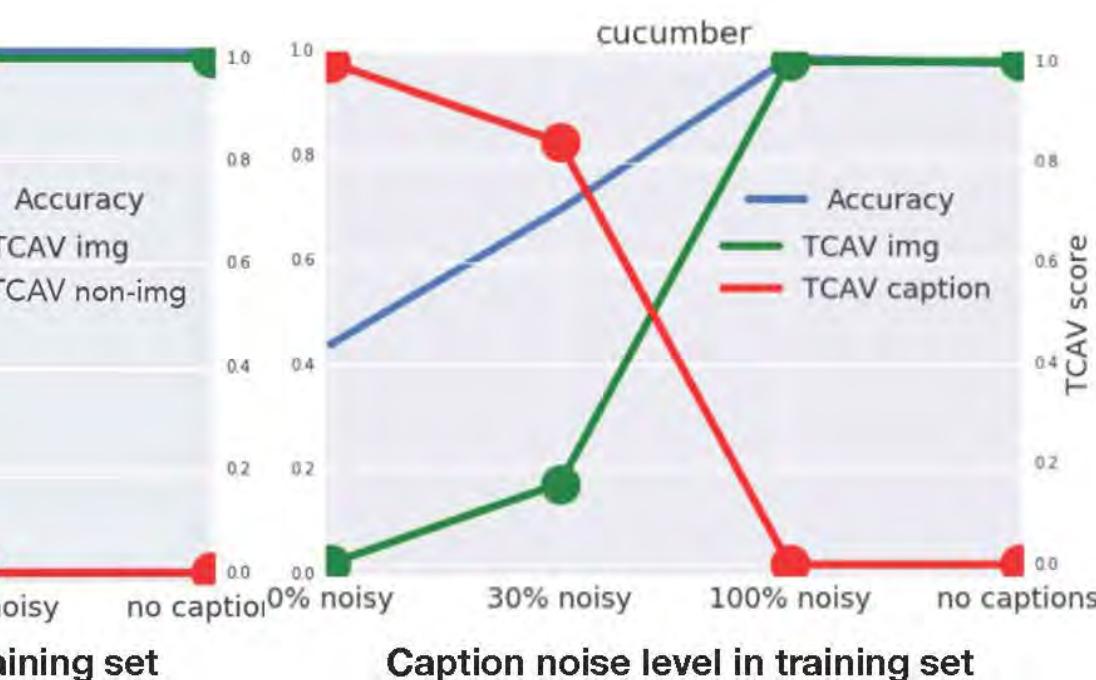
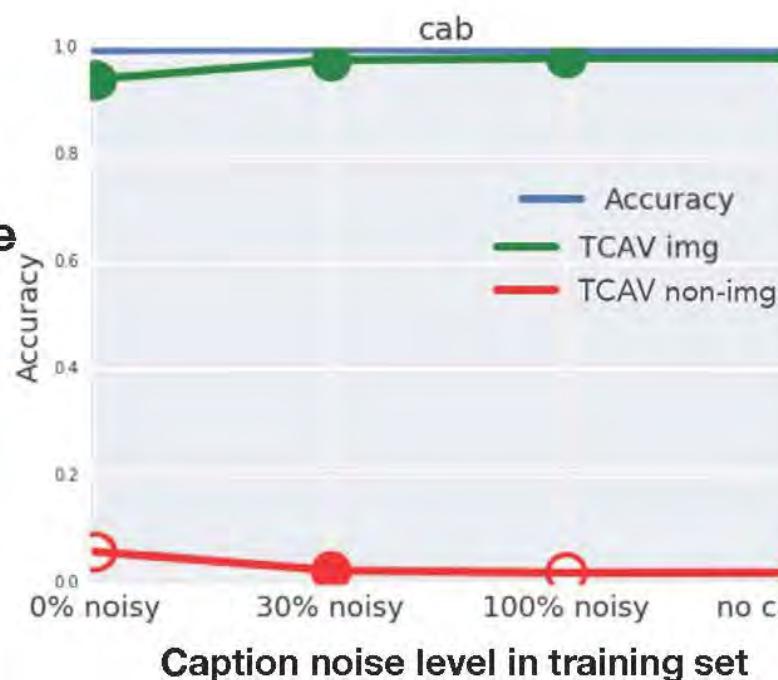
**Test accuracy with no caption image = Importance of image concept**





# An example of ground-truth experiment setup in TCAV

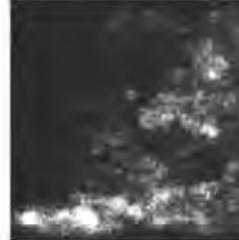
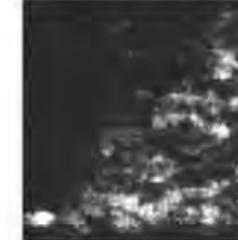
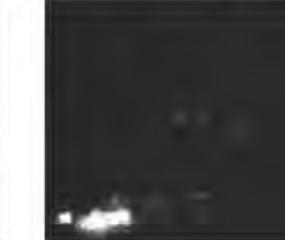
**Test accuracy with no caption image**





# An example of ground-truth experiment setup in TCAV

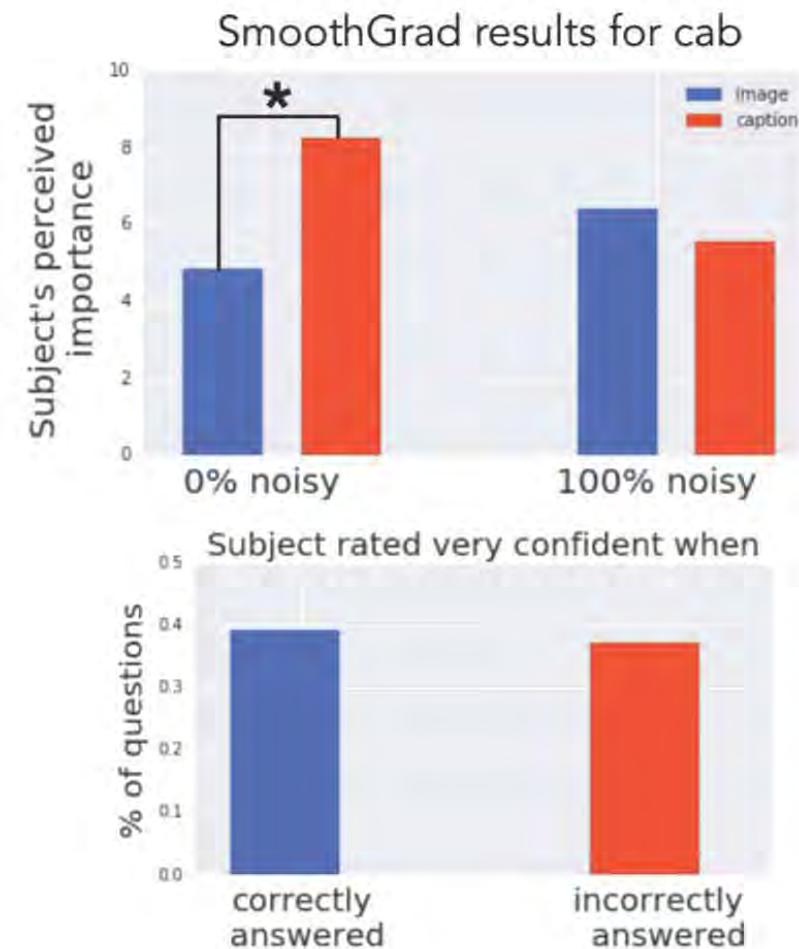
Cool, cool.  
Can saliency maps do it too?

Ground truth	Input image (with caption)	Vanilla gradient	Guided backprop	Integrated gradient	Smoothgrad
Image concept  Cab					



# Human subject experiment: Can saliency maps communicate the same information?

- Correctly communicated 52% (50% random)
- More than 50% no significant consensus among turkers
- Humans are **very** confident even when they are wrong.



50 turkers, shown 3 classes and 2 saliency maps



# Evaluating interpretability methods



Function-based



Factor-based



Application-based

- Decide which level of evaluation is needed.
- Do **human experiments** when you can.
- **Formulate an experiment** where you have the **ground-truth** when you can.

# Conclusion

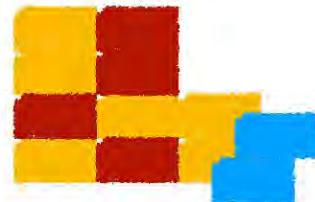
**Why and when?**

Fundamental underspecification

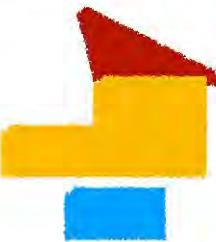
---

**How?**

Before building  
any model



Building  
a new model



After  
building a model



---

**How to evaluate?**

Human experiment and ground-truth experiment

Google's Interpretability best practices:

<https://ai.google/education/responsible-ai-practices>



<https://imgflip.com>

# Conclusion

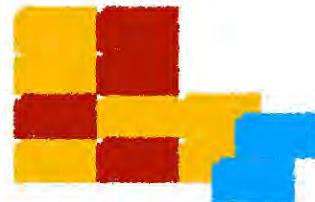
**Why and when?**

Fundamental underspecification

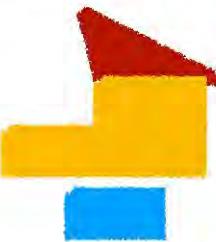
---

**How?**

Before building  
any model



Building  
a new model



After  
building a model



---

**How to evaluate?**

Human experiment and ground-truth experiment

Google's Interpretability best practices:

<https://ai.google/education/responsible-ai-practices>