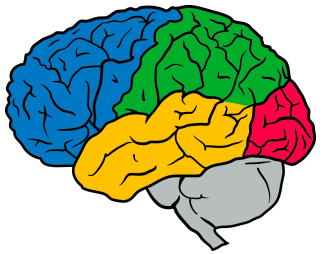
# Interpretable Machine Learning:
# The fuss, the concrete and the questions



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

Been Kim
Google Brain



with Finale Doshi-Velez, Harvard university

Tutorial, ICML 2017

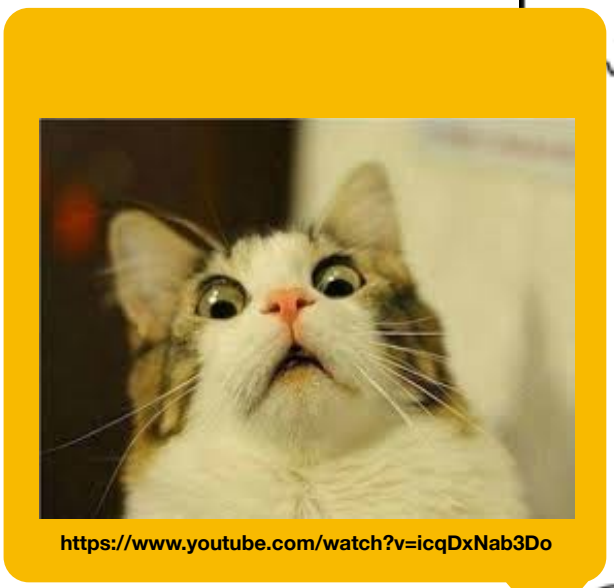
HARVARD
John A. Paulson
School of Engineering
and Applied Sciences

Contents of this tutorial is largely based on our paper

**Towards A Rigorous Science of Interpretable Machine Learning**

https://arxiv.org/abs/1702.08608


**https://xkcd.com/**



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


https://xkcd.com/

# Potentially serious consequences? Yes.

- Cost-effective Health Care (CEHC) built models to predict probability of death for patients [Cooper et al. 97]
  - HasAsthma(x) ⇒ LowerRisk for pneumonia (x)




https://xkcd.com/

Example borrowed from [Caruana et al. '15]

# Potentially serious consequences? Yes.

- Cost-effective Health Care (CEHC) b
  probability of death for patients [Co

  - HasAsthma(x) ⇒ LowerRisk for pneumonia (x)

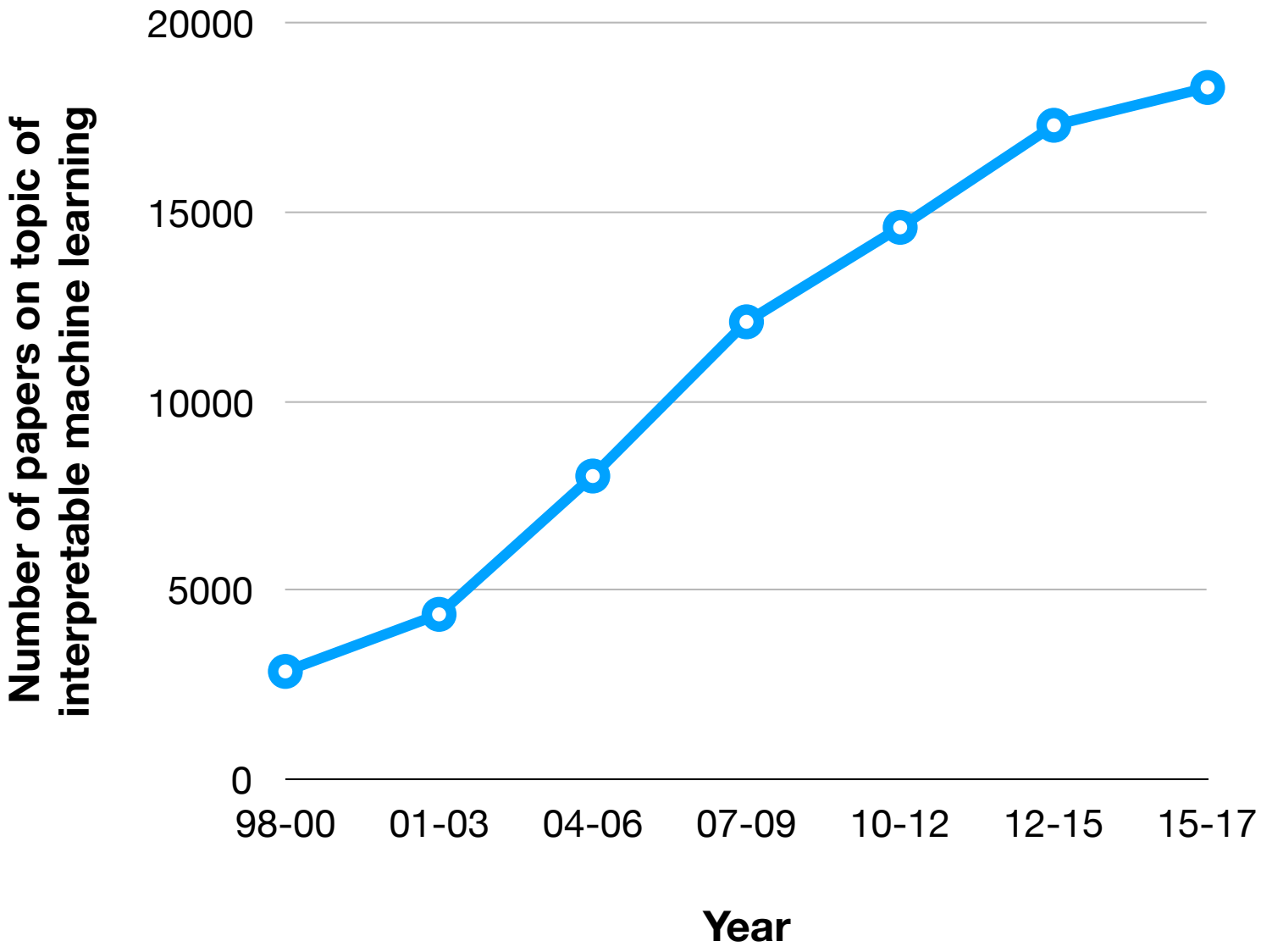
**What else did it learn?!**

Doctors think he/she is high risk → Aggressive Treatment



Example borrowed from [Caruana et al. '15]

# ML community is responding



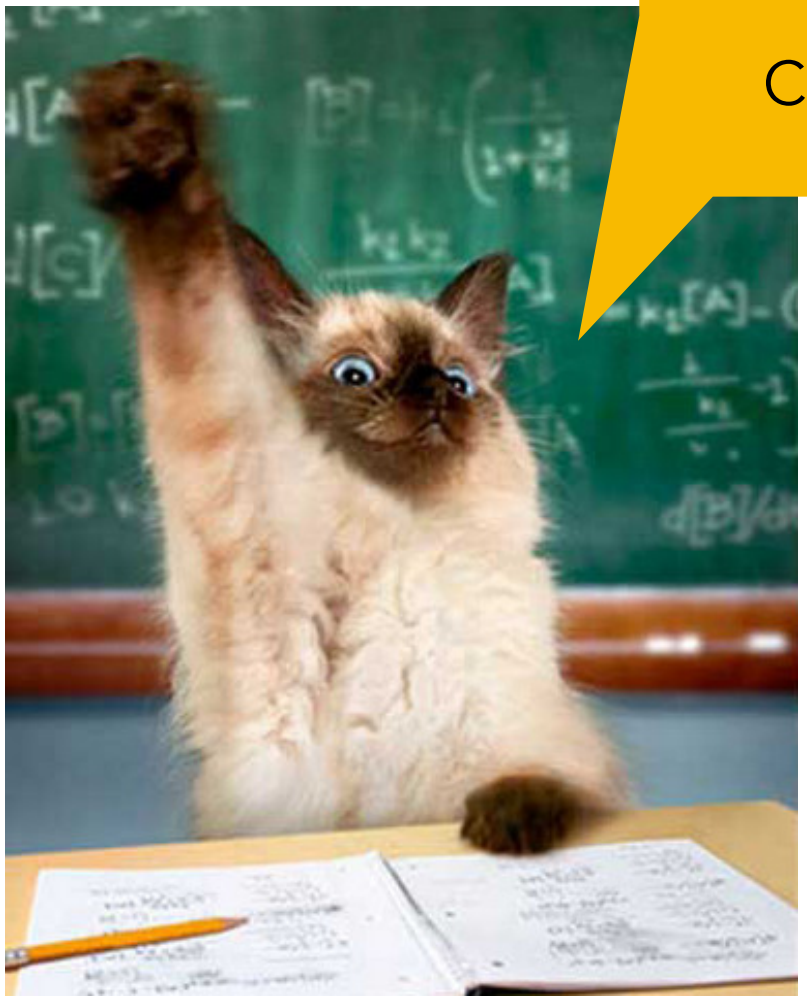
# Why now?

Widespread data collect + vast computation resources

→ ML everywhere!

I heard you can just use decision trees…

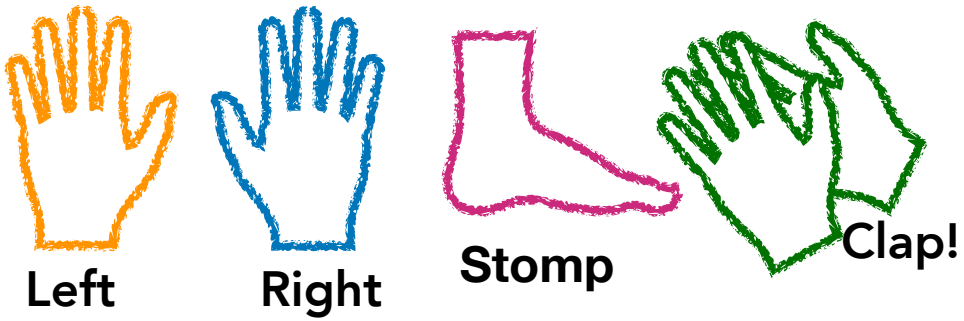
Can we go home now?



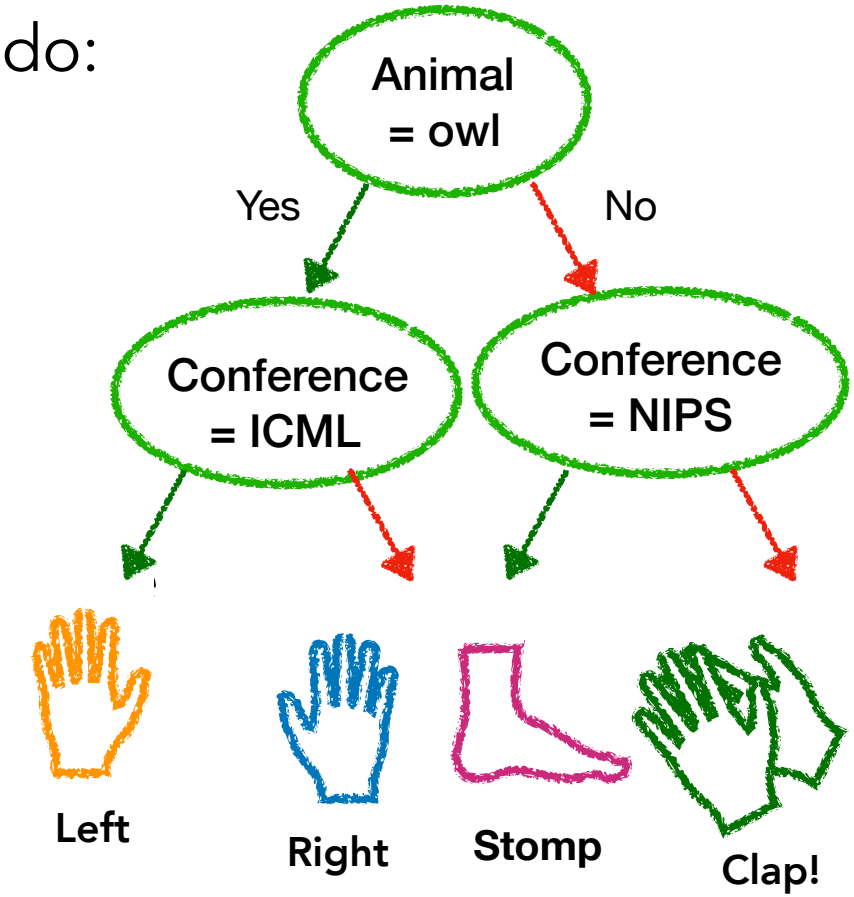
http://www.ogroup.com.au/raise-your-hand-when-you-should-and-why-you-should/

# Experiment.

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

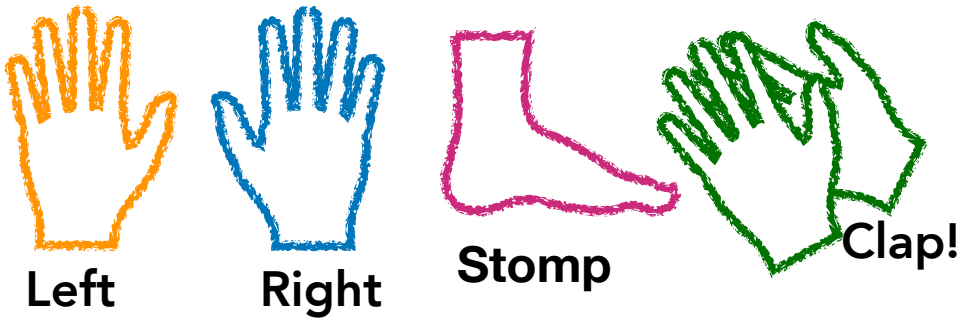



**Input = [Owl, ICML]**

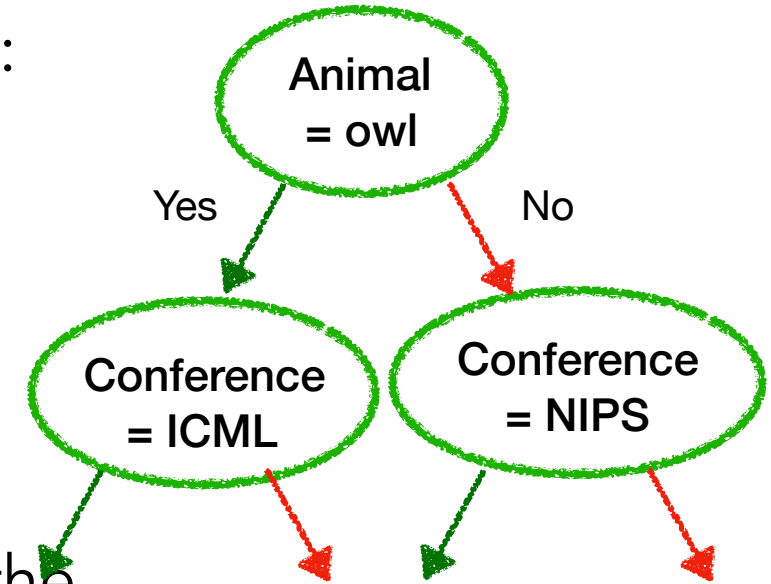


# Experiment.

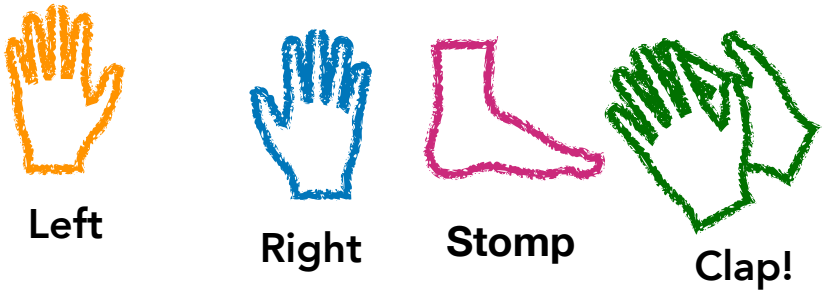
- I will show you a decision tree. Follow the right path given an input, and you do:




**Input = [Kangaroo, ICML]**

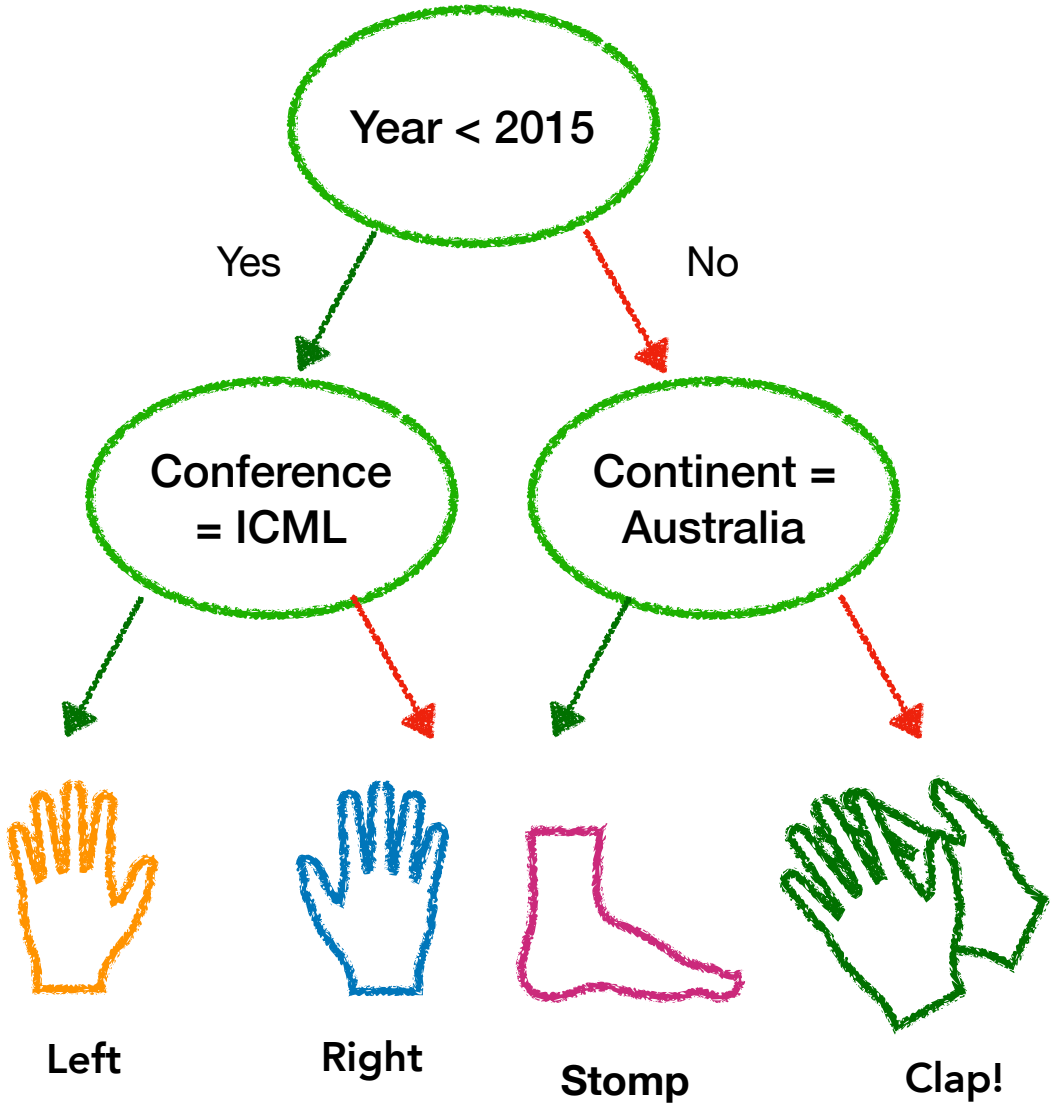



- As soon as you know the answer, do the action!



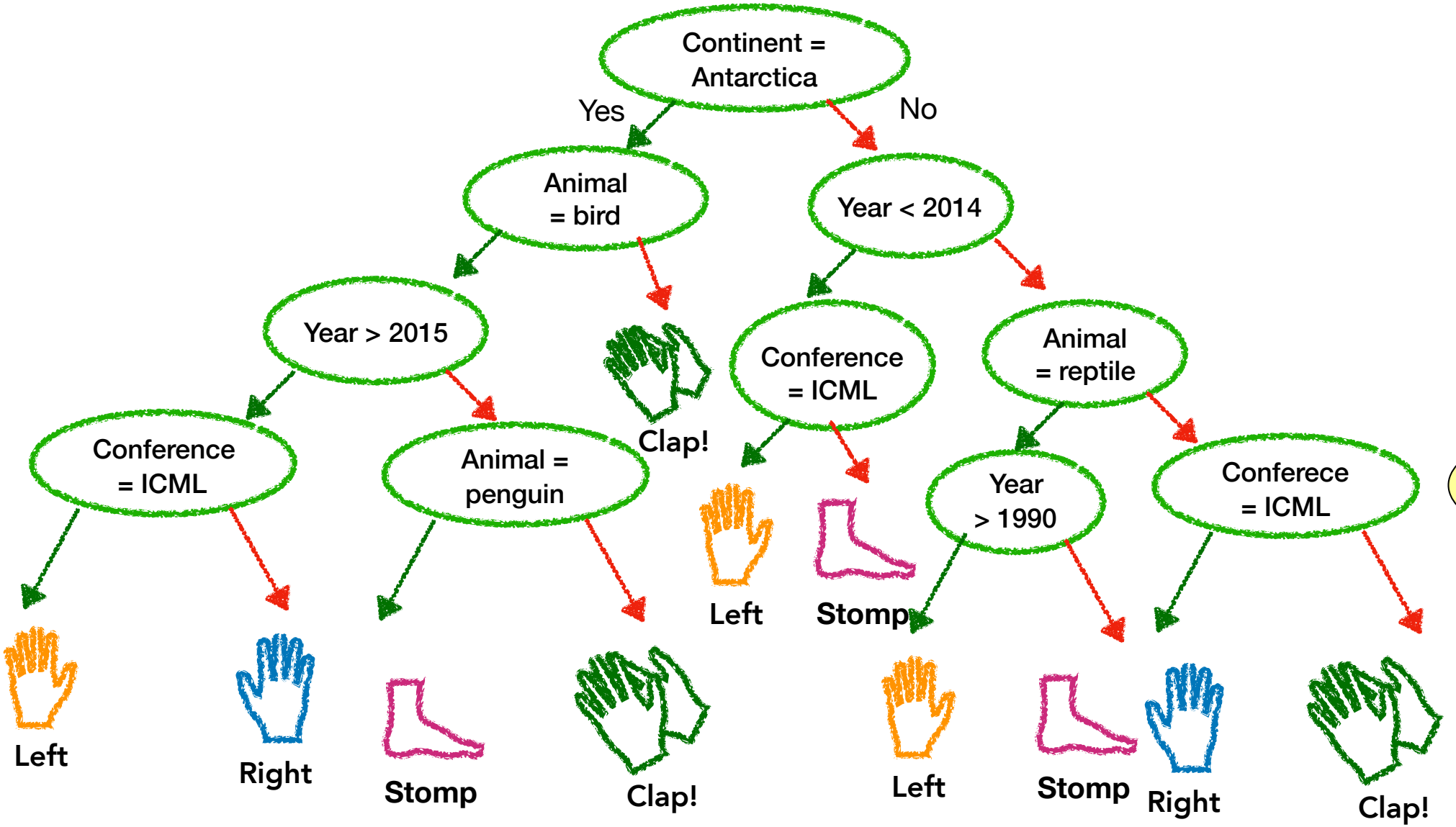
# Sample decision tree #1

Input: [ ICML, 2017, Australia, Kangaroo, Sunny ]



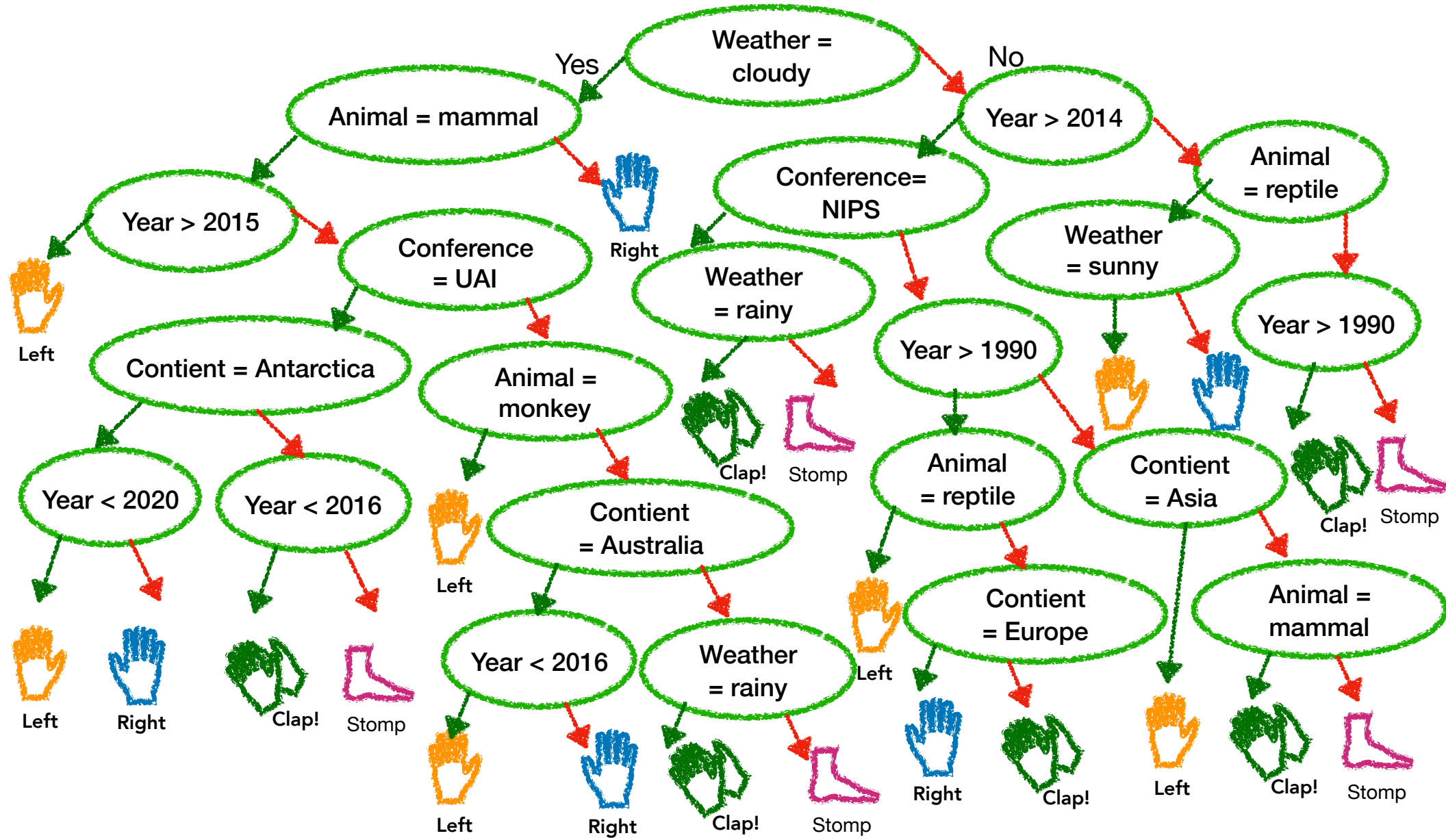
# Sample decision tree #2

Input: [ ICML, 2017, Australia, Kangaroo, Sunny ]



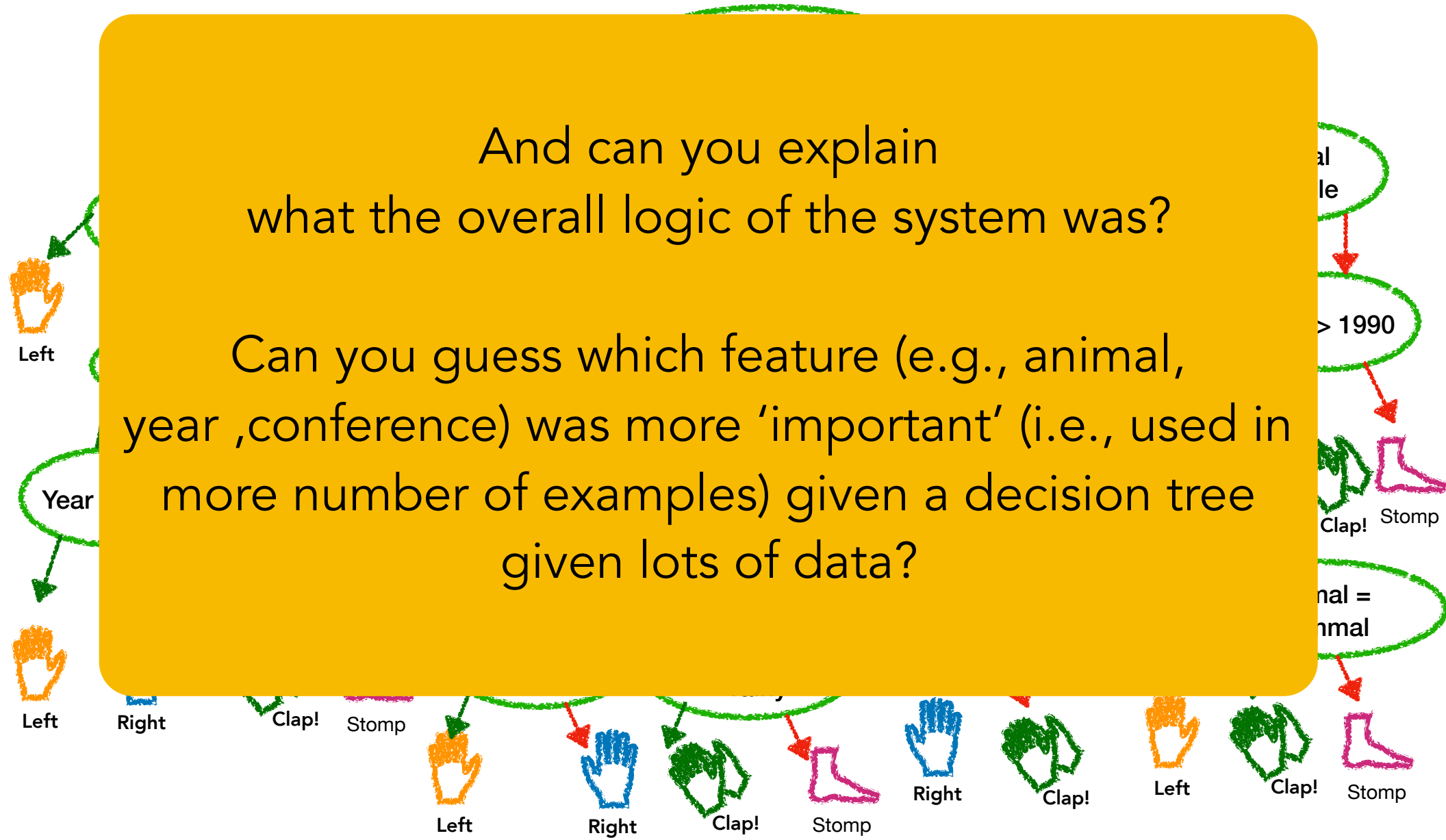
# Sample decision tree #3

## Input: [ ICML, 2017, Australia, Kangaroo, Sunny ]



# Sample decision tree #3

Input: [ ICML, 2017, Australia, Kangaroo, Sunny ]




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

Can you guess which feature (e.g., animal, year ,conference) was more ‘important’ (i.e., used in more number of examples) given a decision tree given lots of data?

# 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 ( at ICML ) | then | attend tutorial |
| Else if ( cloudy and hot ) | then | go swim |
| Else if ( snowing ) | then | go ski |
| Else if ( New Dr. Who ) | then | watch TV |
| Else if ( paper deadline ) | then | go work |
| Else if ( sick and bored ) | then | watch TV |
| 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 )
THEN go to beach
ELSE work

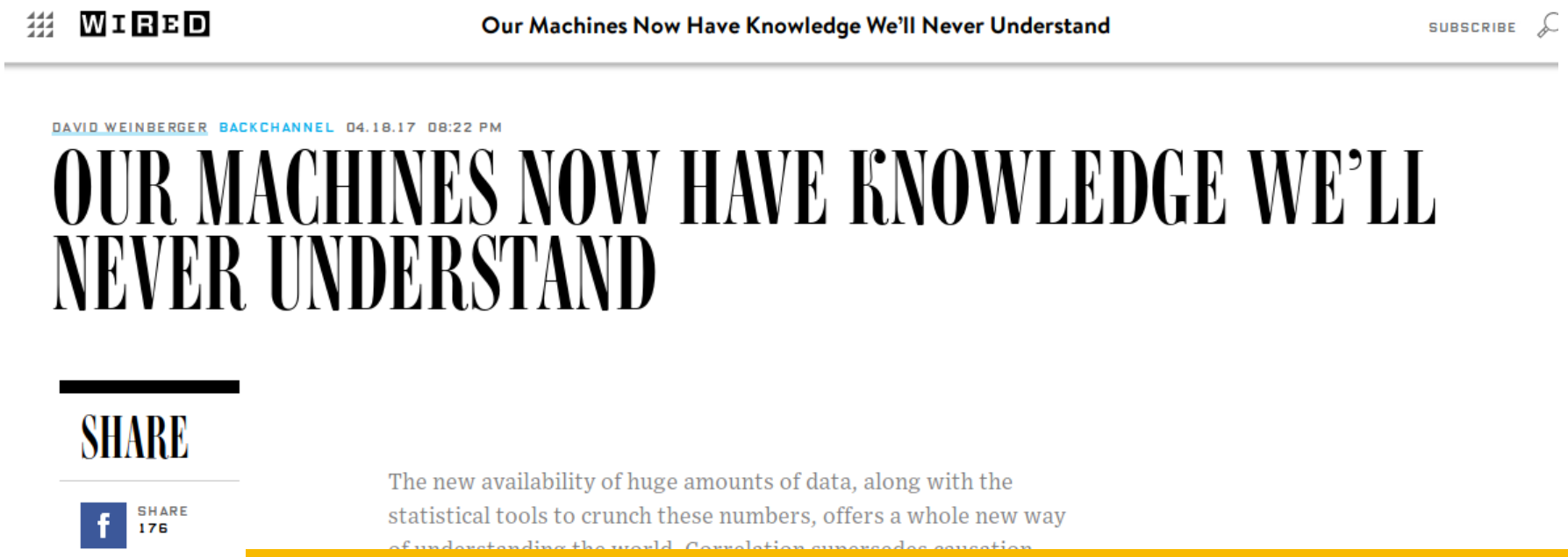
# Maybe rule sets are better?

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

# Wait… Why am I here then?

https://ameblo.jp/kamar-saya-meg/entry-12247929580.html

# Is interpretability possible at all?




Key Point:
Interpretability is NOT about understanding all bits and bytes of the model for all data points (we cannot).

It's about knowing enough for your downstream tasks.



https://www.wired.com/story/our-machines-now-have-knowledge-well-never-understand/

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

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

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

http://blog.xfree.hu/myblog.tvn?SID=&from=20&pid=&pev=2016&pho=02&pnap=&kat=1083&searchkey=&hol=&n=sarkadykati

# What is interpretability?

- Not as simple as decision rules
- Not as simple as rule lists or rule sets.
- Not about understanding every bits and bytes of the model.

Our goal:

Bring us toward more precise notion of what interpretability entails, when it is needed, and how to evaluate it.

Just the **start of a discussion**!

# Interpretability

Dictionary definition:

Interpretation is the process of giving **explanations**

# Interpretability

Dictionary definition:

Interpretation is the process of giving **explanations**

To Humans

# Interpretability

Why and when?

How can we do this?

Interpretation is the process of giving **explanations**

How can we measure 'good' explanations?

To Humans

# Agenda

1. Why and when?

2. How can we do this?

Interpretation is the process of giving **explanations**

3. How can we measure 'good' explanations?

To Humans

# Agenda

1. Why and when?

2. How can we do this?

Interpretation is the process of giving **explanations**

3. How can we measure 'good' explanations?

To Humans

# Why interpretability?

Fundamental **underspecification** in the problem

# Why interpretability?

Fundamental **underspecification** in the problem

**More data or more clever algorithm won't help.**

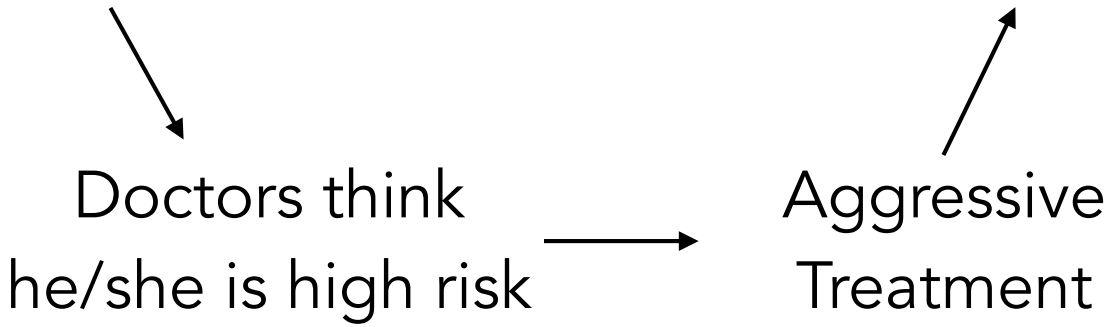
# Underspecification example 1: safety


https://www.ll.mit.edu/publications/labnotes/automation.html

**Interpretability is not safety. But it can help to expose safety issues.**

# Underspecification example 1: safety

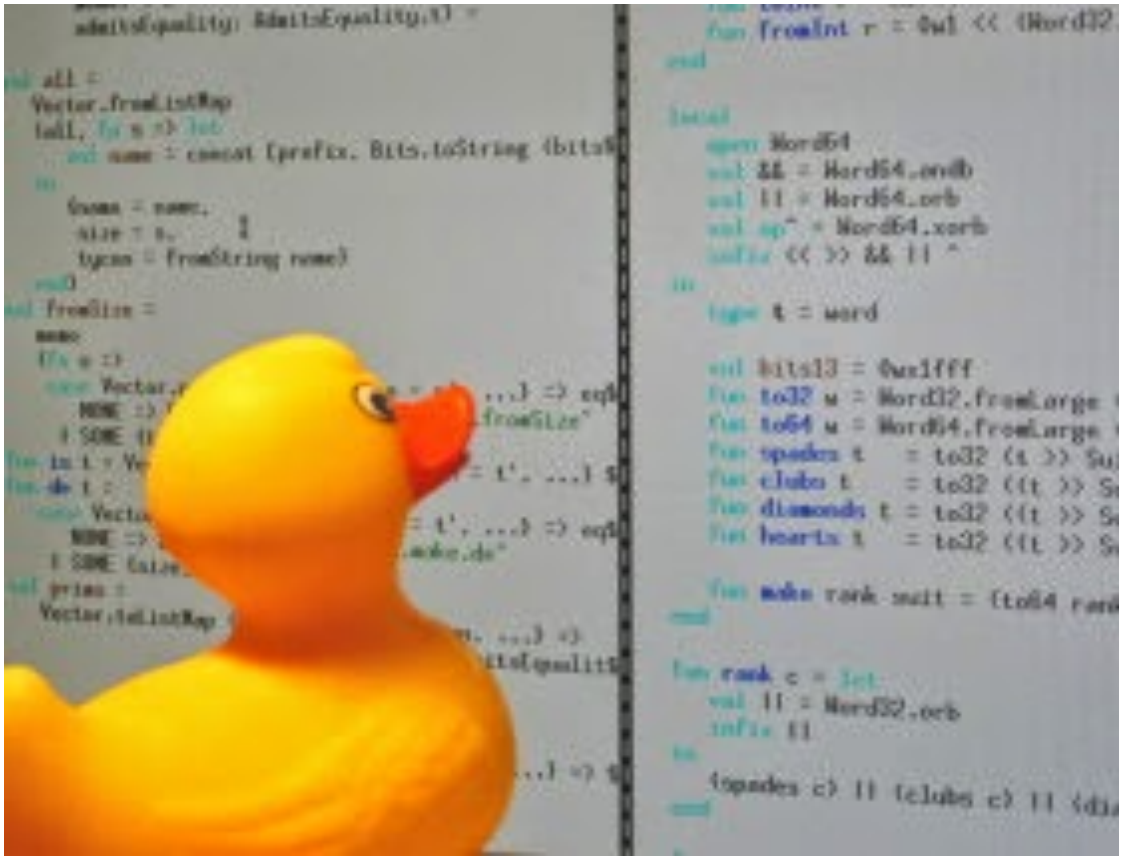
- Cost-effective Health Care (CEHC) built models to predict probability of death for patients [Cooper et al. 97]
  - HasAsthma(x) ⇒ LowerRisk for pneumonia (x)




Example borrowed from [Caruana et al. '15]

# Underspecification example 2: debugging

- We want to understand why the system doesn't work, and fix it.



https://yorktown.cbe.wwu.edu/sandvig/mis314/lectures/images/rubber-duck-debugging.jpg

# Underspecification example 3: mismatched objectives and multi-objective trade-offs

- What you optimize is not what you meant to optimize.

**All of these may address depression. Which side-effects are you willing to risk?**




http://img.medscapestatic.com/pi/features/drugdirectory/octupdate/MYN62330.jpg, , http://img.medscapestatic.com/pi/features/drugdirectory/octupdate/PLV04411.jpg , https://www.google.com/url?sa=i&rct=j&q=&esrc=s&source=images&cd=&cad=rja&uact=8&ved=0ahUKEwjKgaaZkZbVAhXDPT4KHSZ3D-MQjRwIBw&url=http%3A%2F%2Fwww.webmd.com%2Fdrugs%2F2%2Fdrug-4870-5047%2Fvenlafaxine-oral%2Fvenlafaxine-oral%2Fdetails&psig=AFQjCNHMQN9D8bhQZUFyxfHd9AoY5yxq5g&ust=1500580783703785

# Underspecification example 4: science


http://cdn.playbuzz.com/cdn/a6006912-25e4-4cb5-867d-36c333b437c2/f2519ae0-e3d9-48e9-8f0d-4e68e2c99e26.jpeg

# Underspecification example 5: legal/ethics

- We're legally required to provide an explanation and/or we don't want to discriminate against particular groups.

http://leap.utah.edu/_images/program-options/Pre-Law.jpg

# Examples of underspecification

- **Safety**: We want to make sure the system is making sound decisions.
- **Debugging**: We want to understand why a system doesn't work, so we can fix it.
- **Science**: We want to understand something new.
- **Mismatched Objectives and multi-objectives trade-offs**: The system may not be optimizing the true objective.
- **Legal/Ethics:** We're legally required to provide an explanation and/or we don't want to discriminate against particular groups.

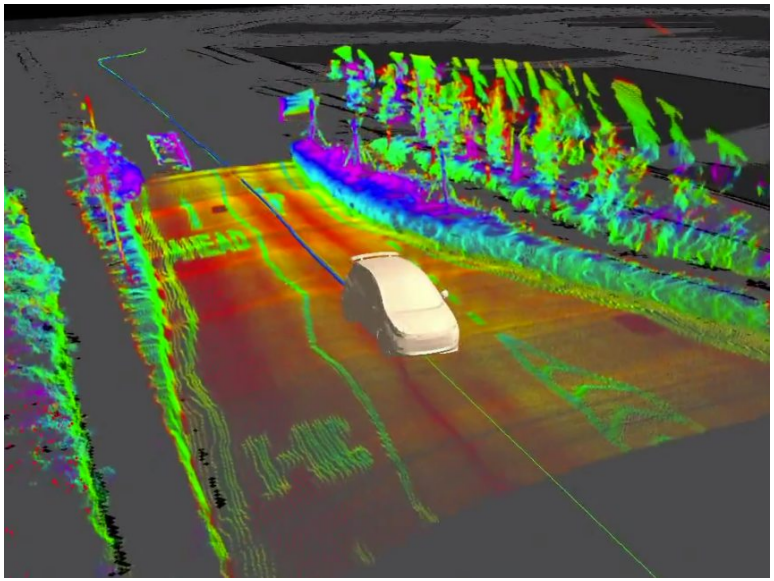
**+ Your case?**

Fundamental **underspecification** in the problem

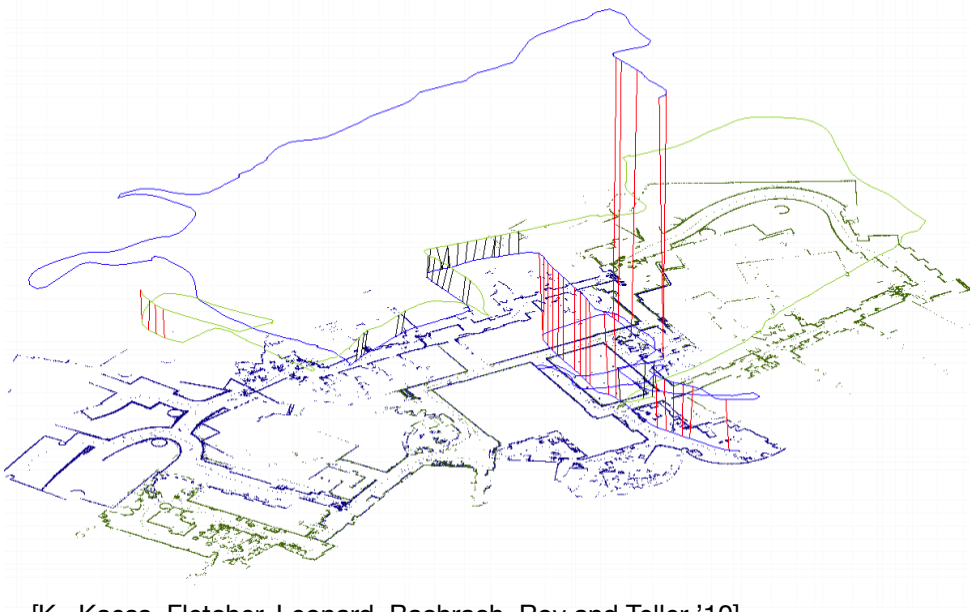
# What is NOT underspecification?


https://www.pinterest.com/dowd3128/type-o-negative/

# Underspecification is not uncertainty

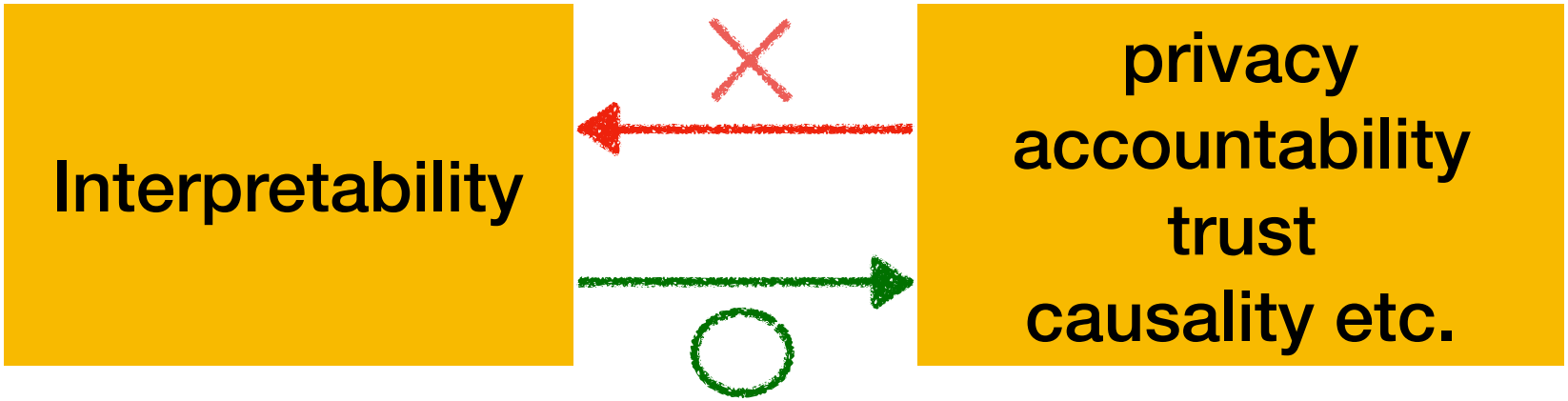


https://www.pinterest.com/pin/461126449319612657/



[K., Kaess, Fletcher, Leonard, Bachrach, Roy and Teller '10]

# Our cousins are not us




- Interpretability can help with them **when we cannot formalize these ideas**

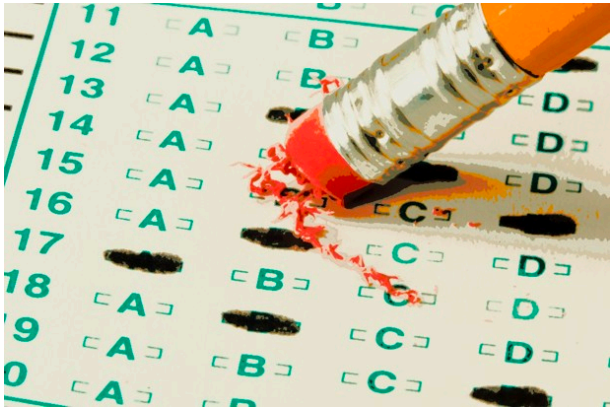
- But once formalized, you may not need interpretability.

# 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://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

# Agenda

1. Why and when?

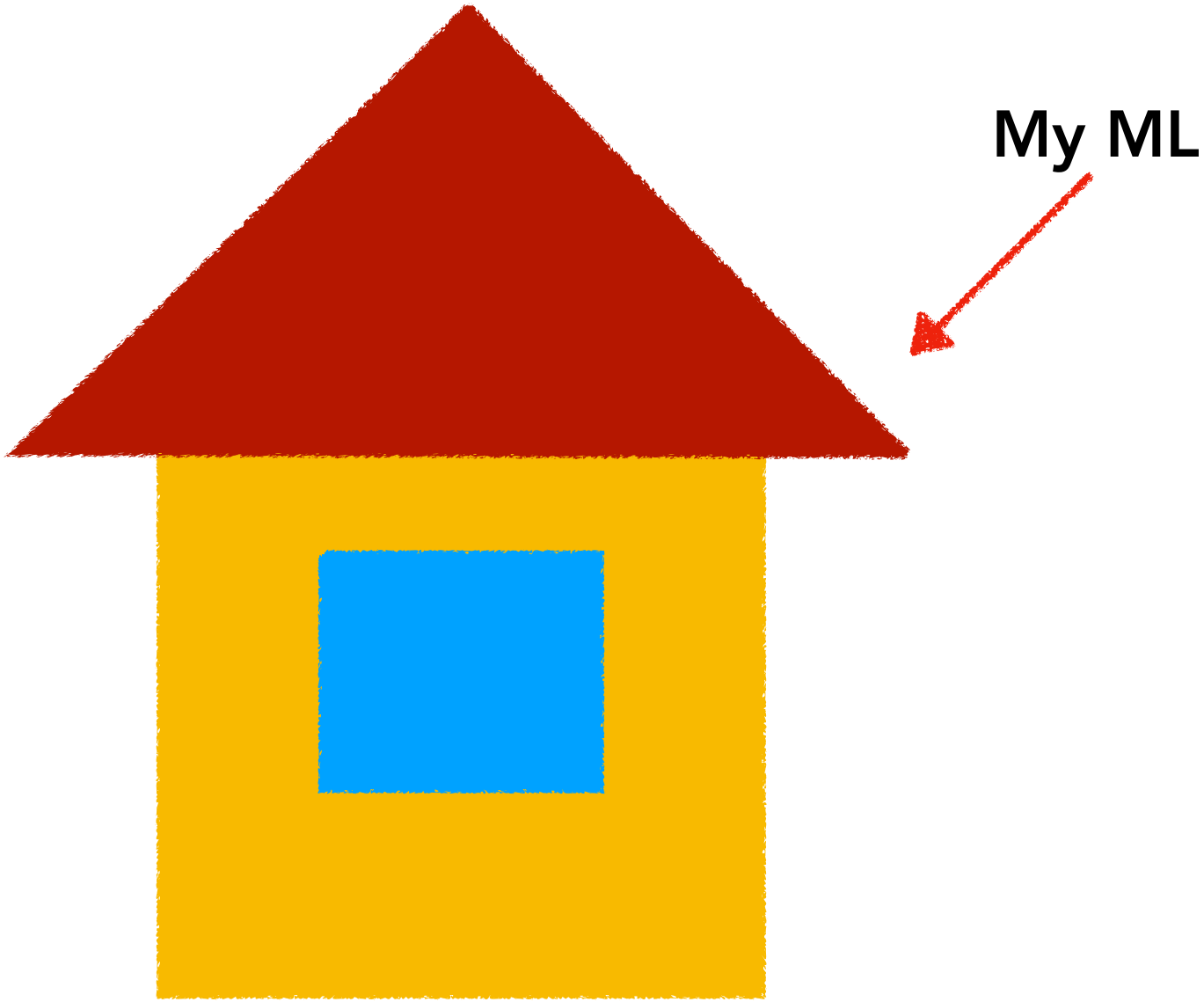
2. How can we do this?

Interpretation is the process of giving **explanations**

3. How can we measure 'good' explanations?

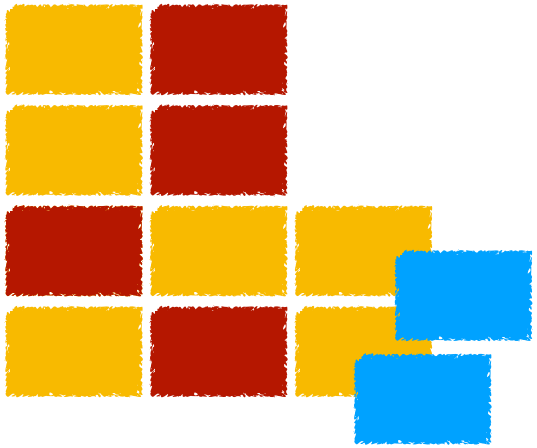
To Humans

# Types of interpretable methods

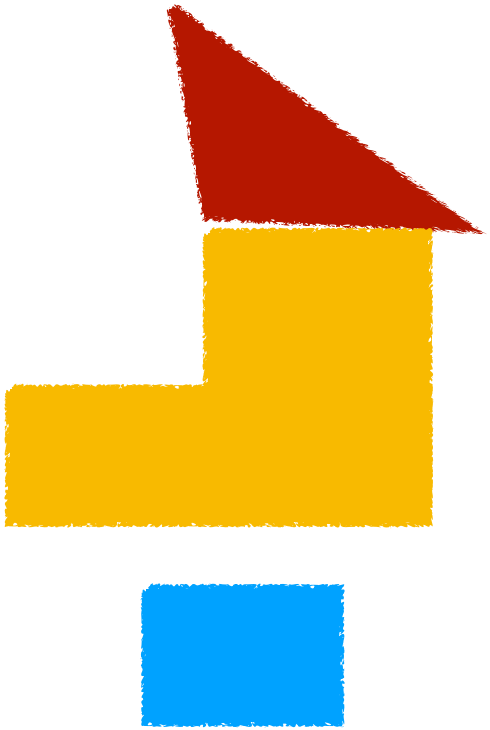


# Types of interpretable methods

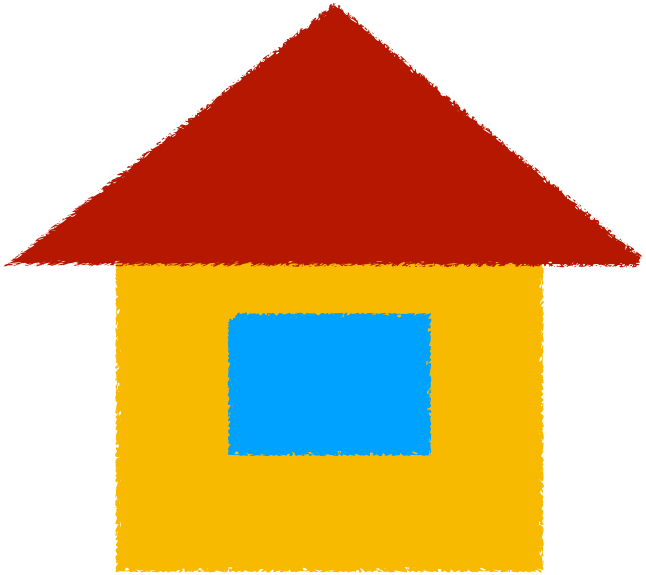
**Before building any model**



**Building a new model**



**After building a model**



Not mutually exclusive categories

Nor

Exhaustive list of literatures

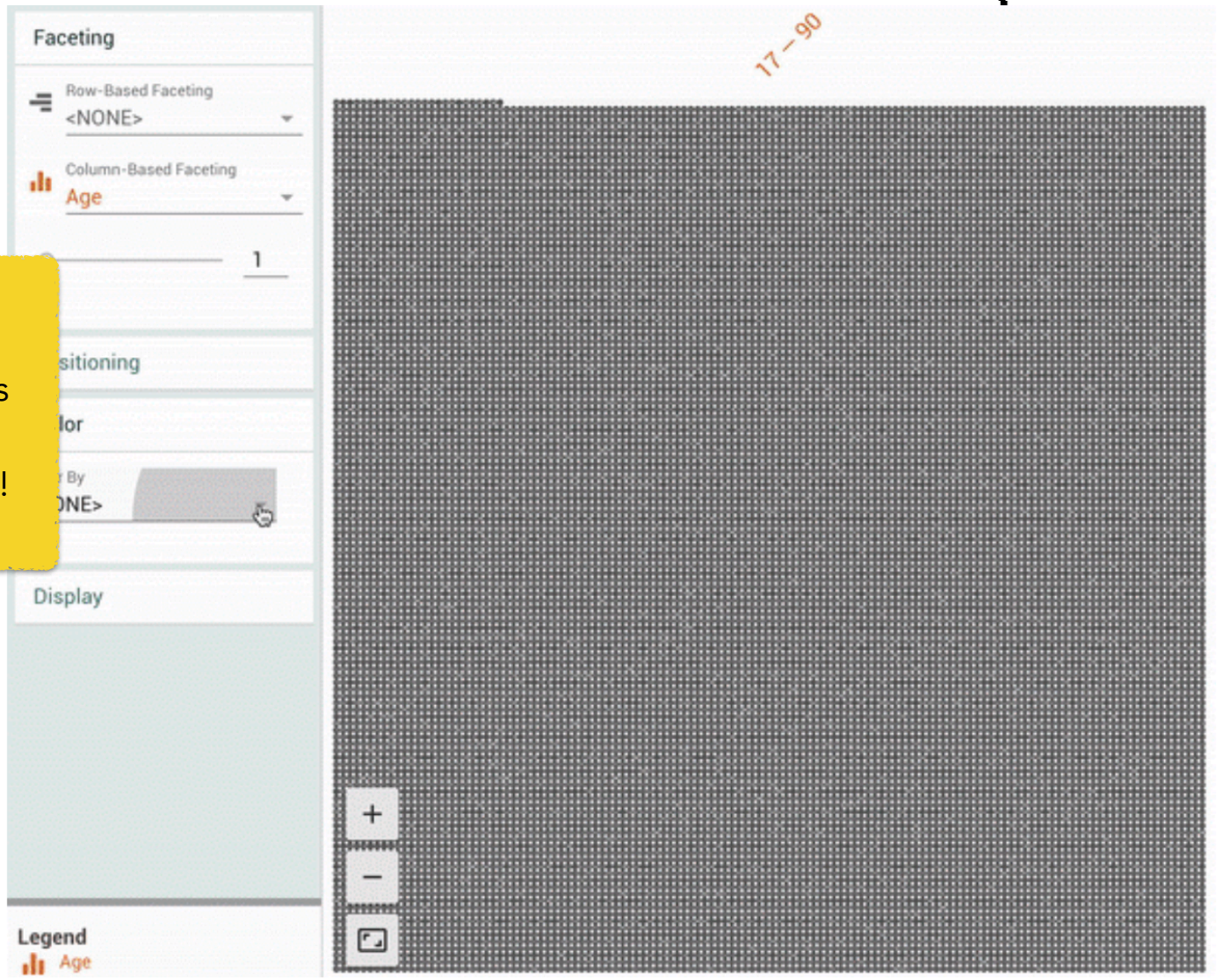
# Types of interpretable methods

**Before building any model**

- **Visualization**
- Exploratory data analysis

# Before building any model: Visualization for data exploration

Need more participations from HCI communities!

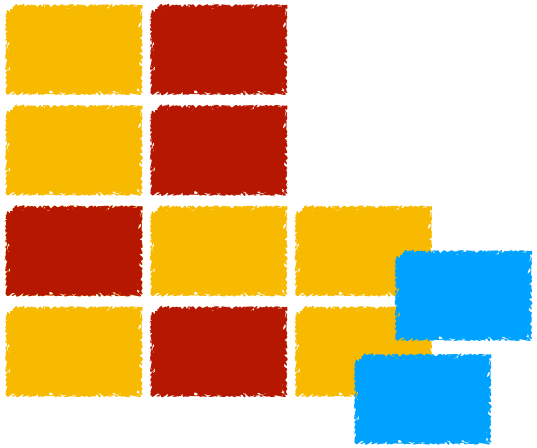



[Viégas and Wattenberg '07]
[Maaten et al. '08]
[Amershi et al. '09]
[Patel et al. '10]
[Varshney et al. '12]

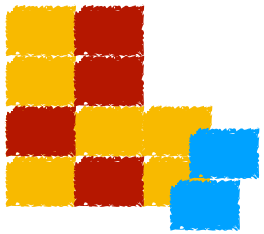
https://pair-code.github.io/facets/quickdraw.html

# Types of interpretable methods

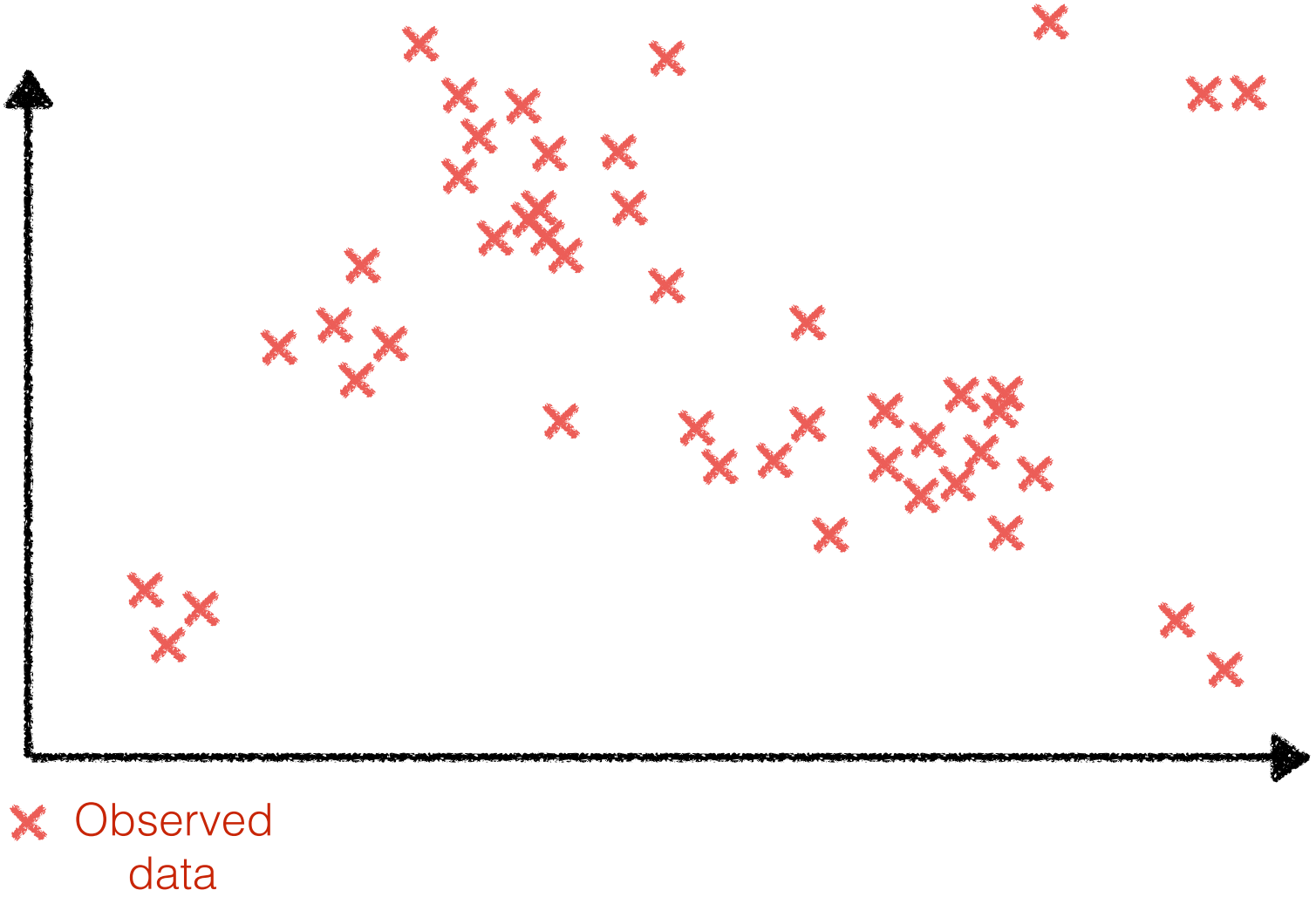
**Before building any model**

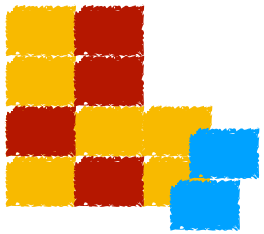


- Visualization
- **Exploratory data analysis**
  [Tukey 77]

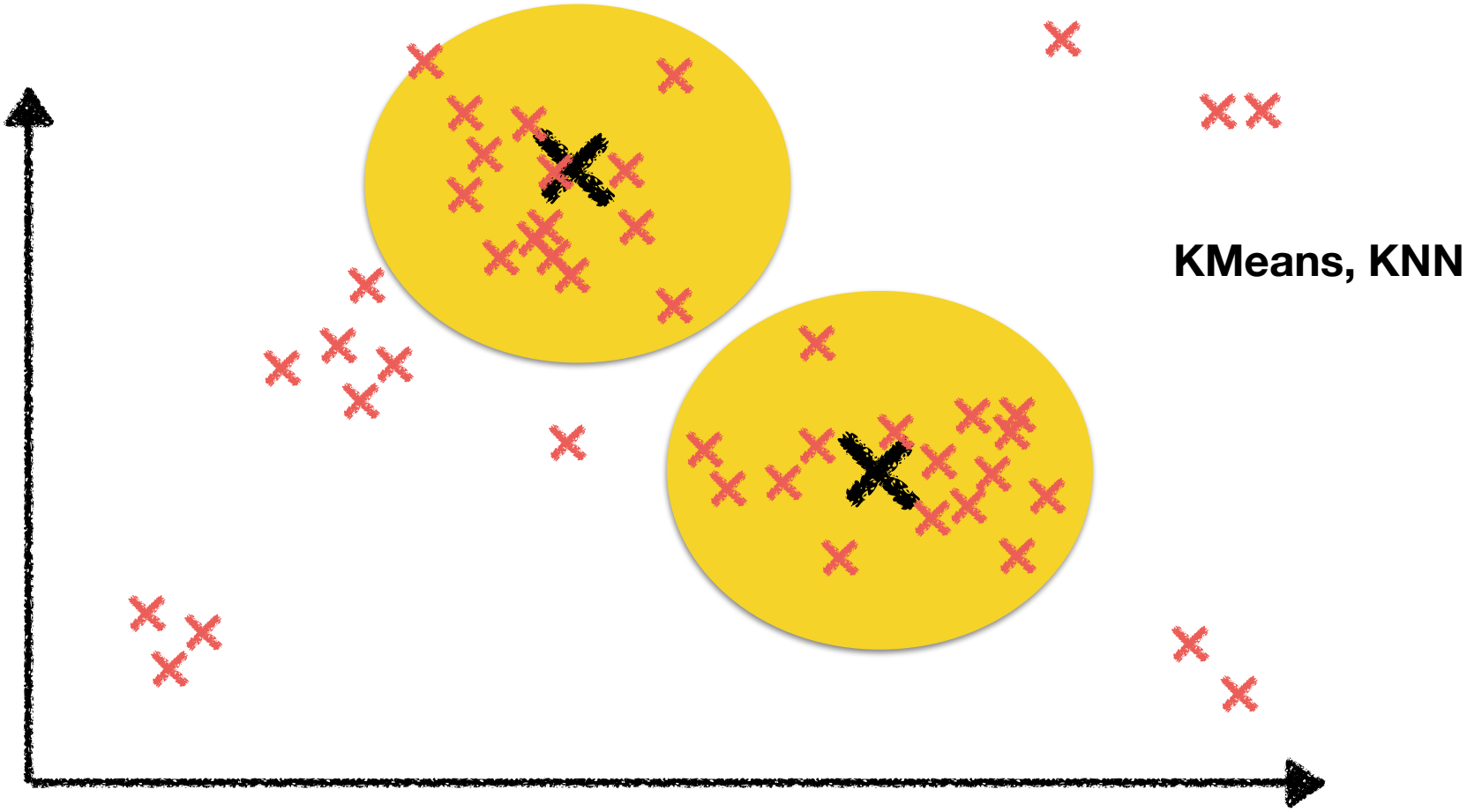


# Before building any model: Exploratory data analysis

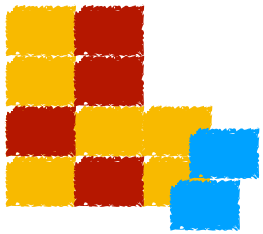




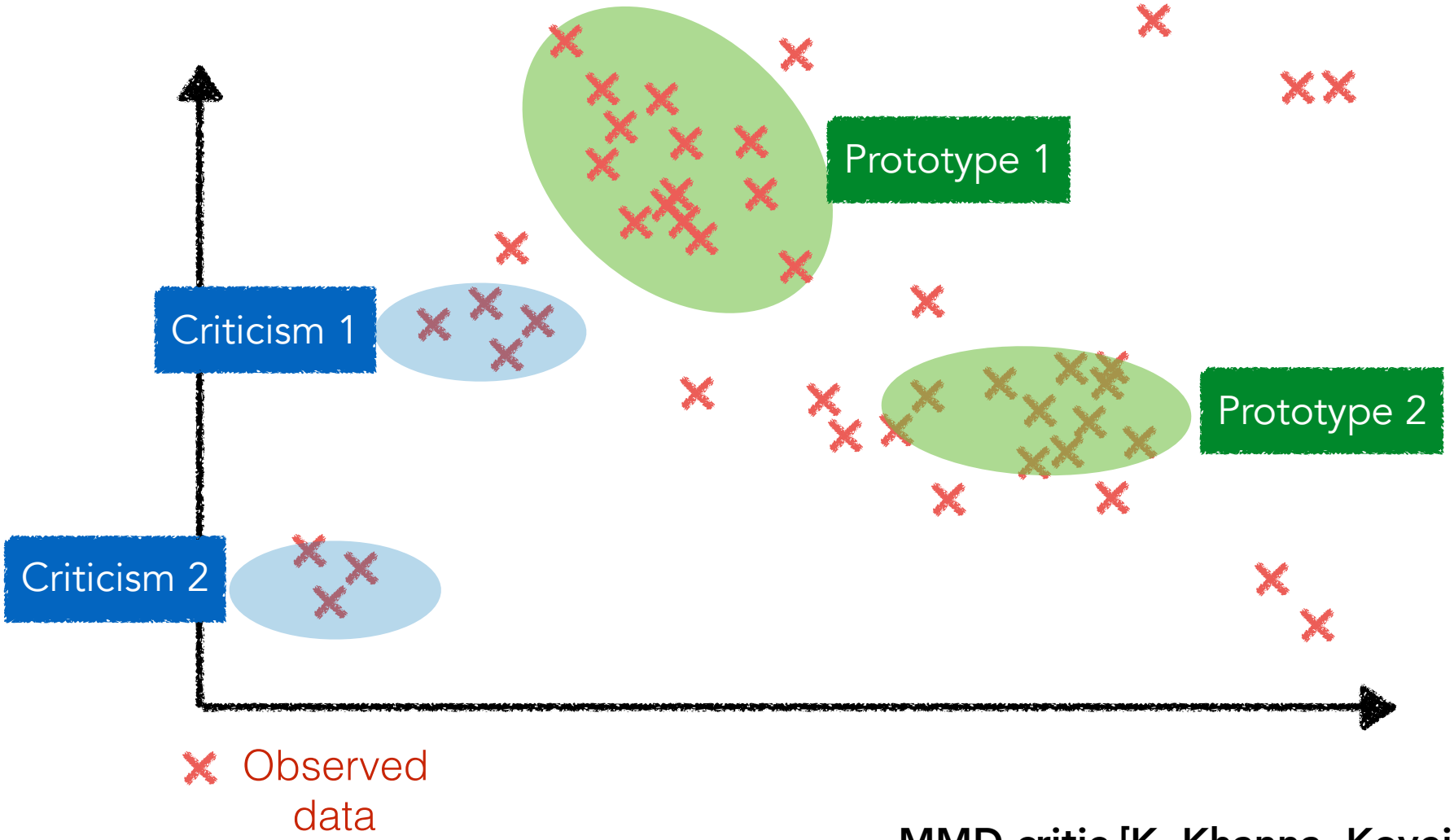
# Before building any model: Exploratory data analysis



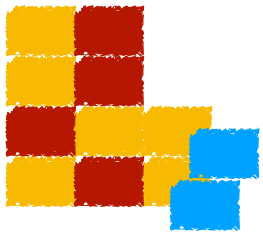

[Simon et al., '07]
[Lin and Bilmes, '11]



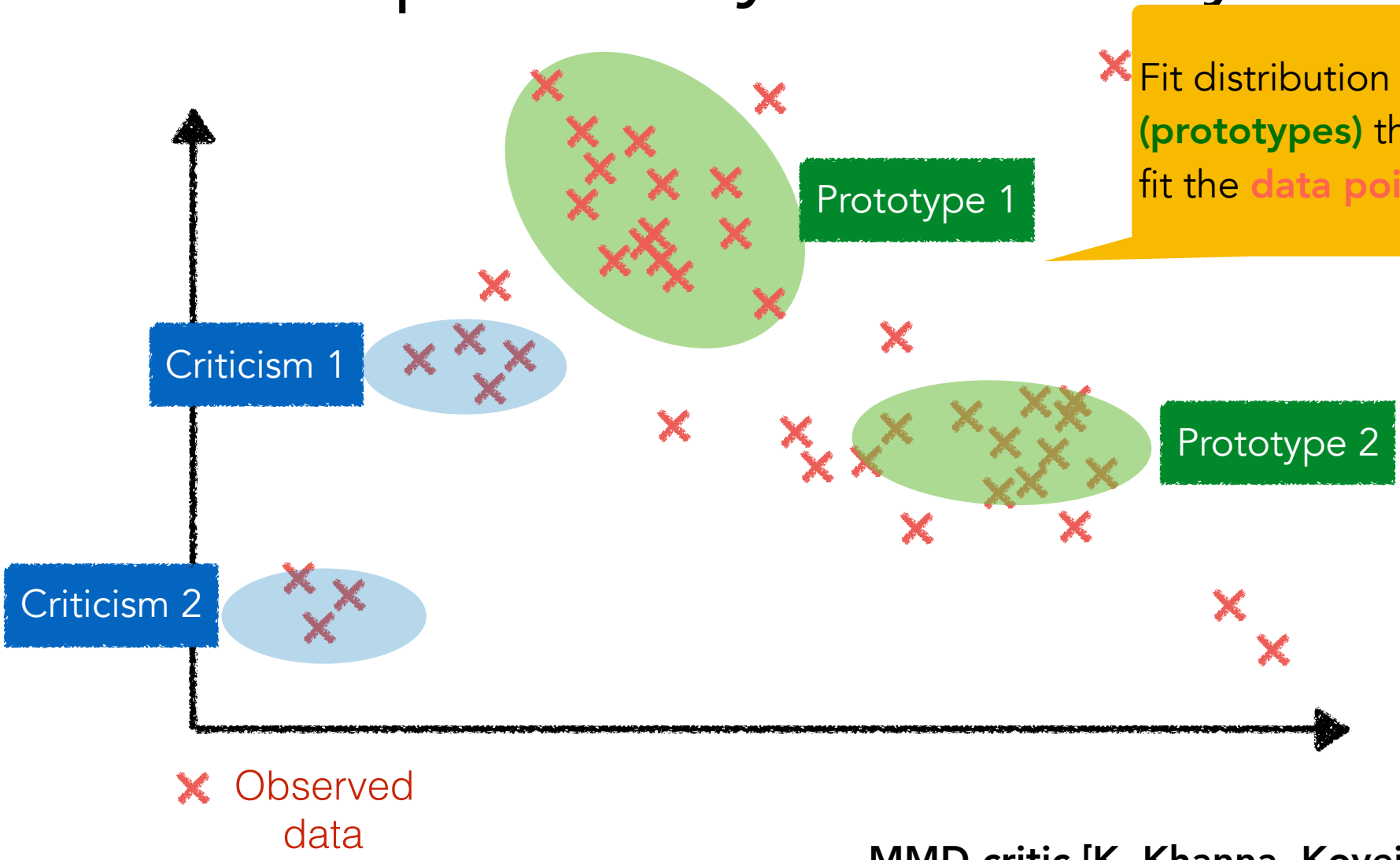
# Before building any model: Exploratory data analysis



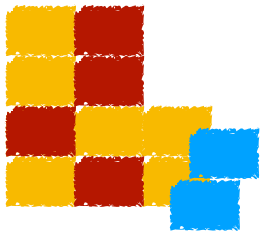

**MMD-critic [K. Khanna, Koyejo '16]**



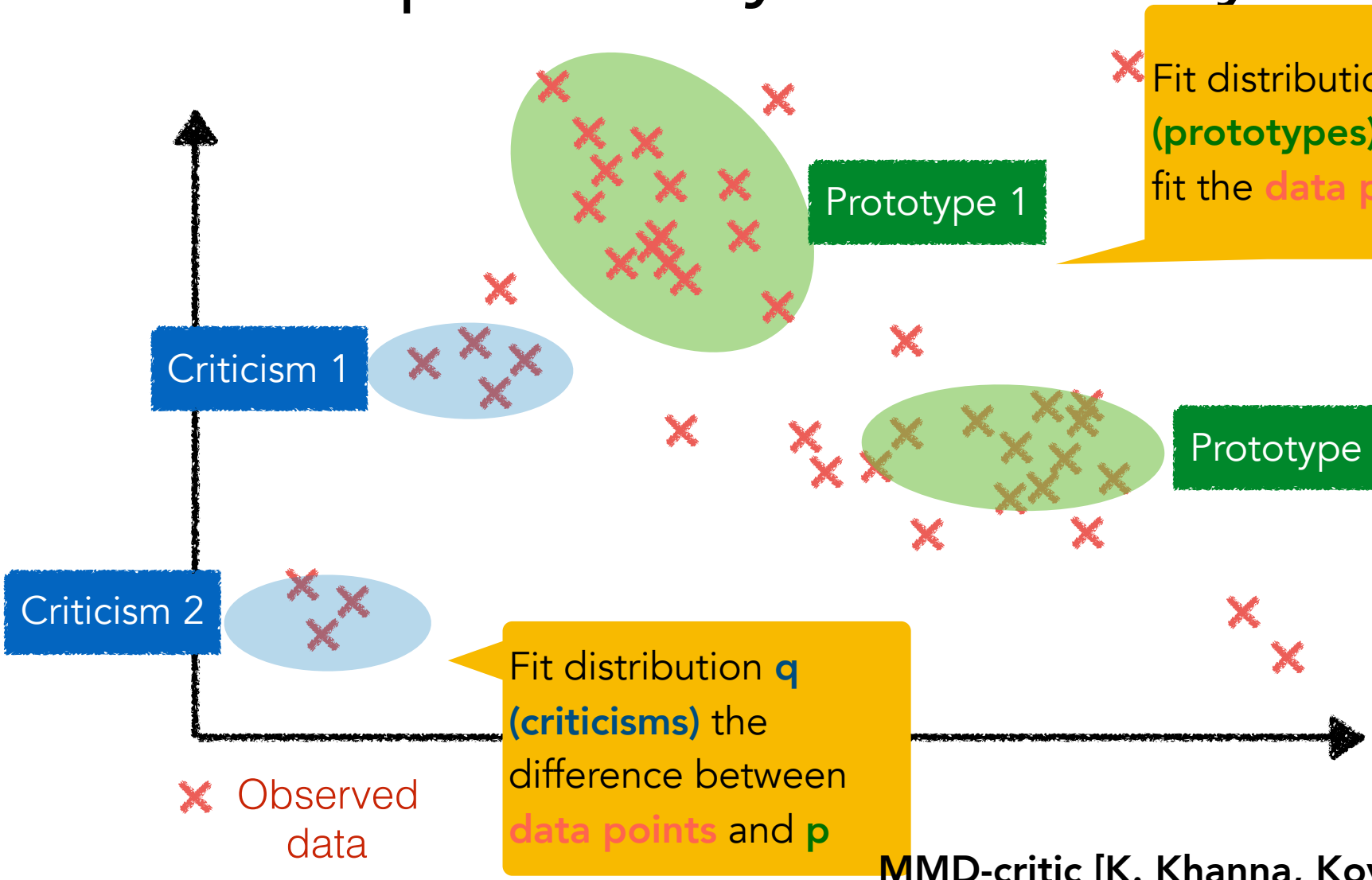
# Before building any model: Exploratory data analysis



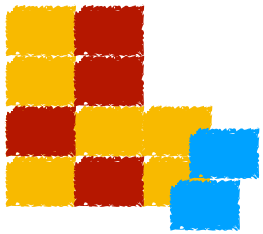

**MMD-critic [K. Khanna, Koyejo '16]**



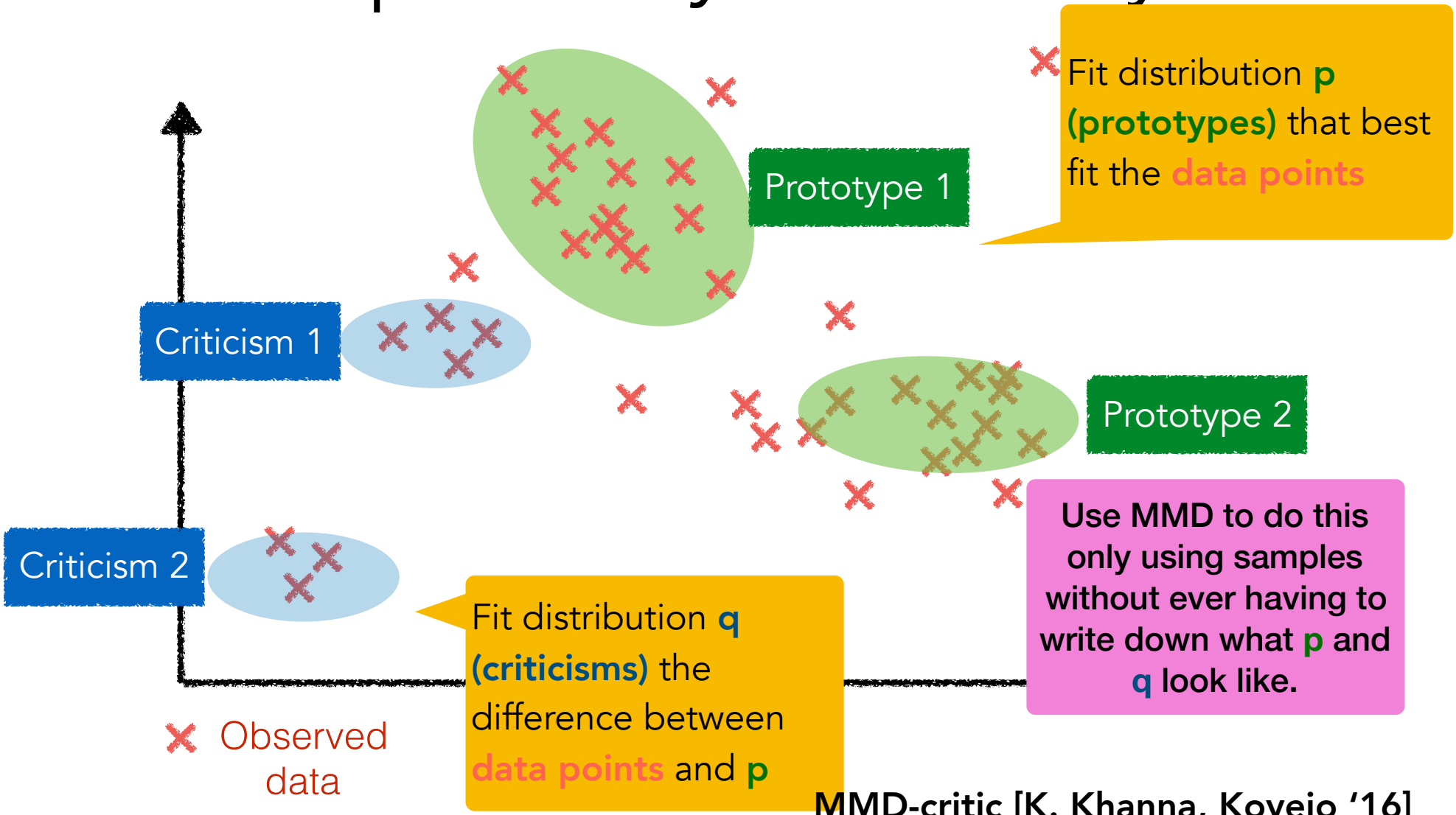
# Before building any model: Exploratory data analysis




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



# Before building any model: Exploratory data analysis




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

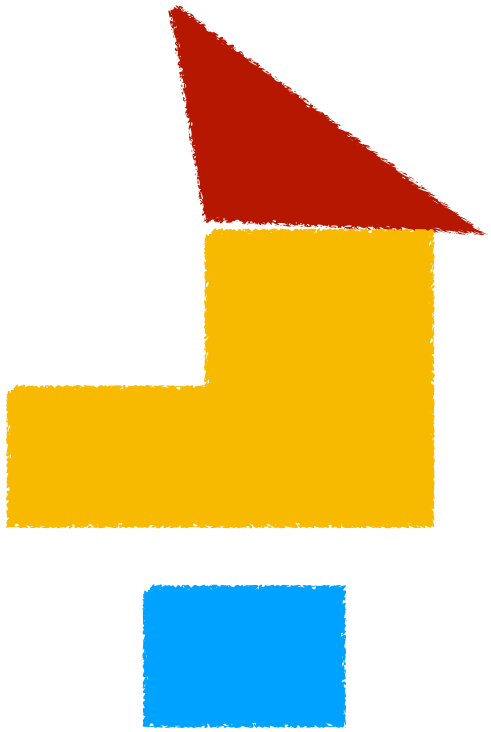
# Types of interpretable methods

**Before building any model**

- Visualization
- Exploratory data analysis

# Types of interpretable methods

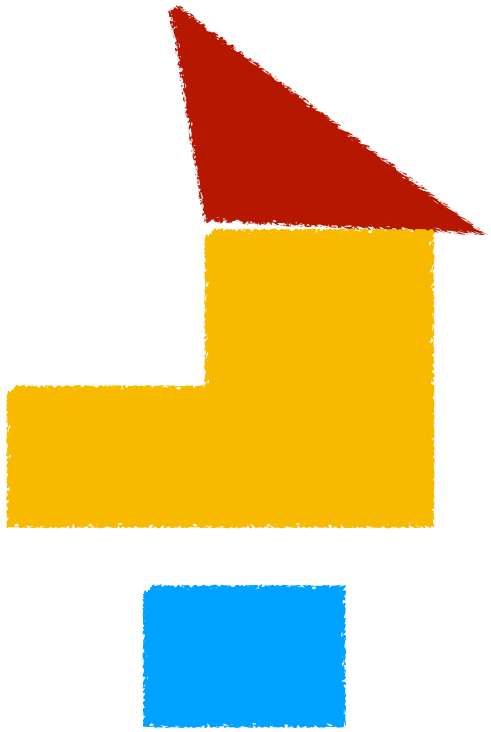
**Building a new model**



- rule-based, per-feature-based
- case-based
- sparsity
- monotonicity

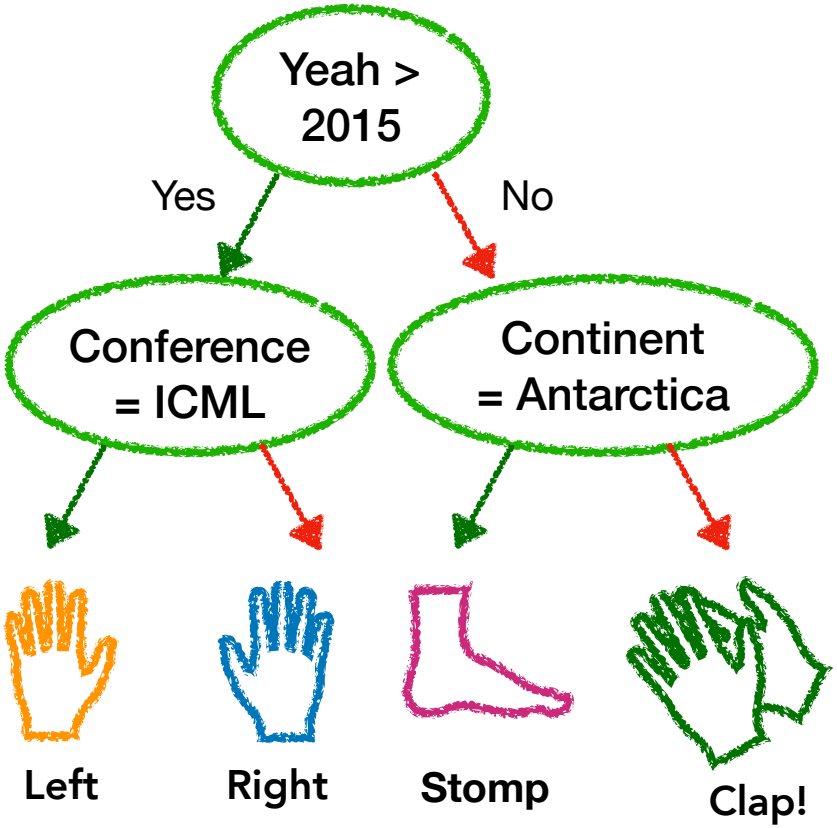
# Types of interpretable methods

**Building a new model**



- **rule-based, per-feature-based**
- case-based
- sparsity
- monotonicity

# Building a new model: Rule-based

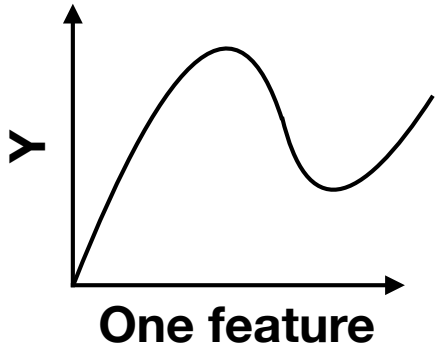



IF ( sunny and hot ) OR ( cloudy and hot )
THEN go to beach
ELSE work

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]

# Building a new model: Per-feature based




Linear model $y = \beta_0 + \beta_1 x_1 + \ldots + \beta_n x_n$

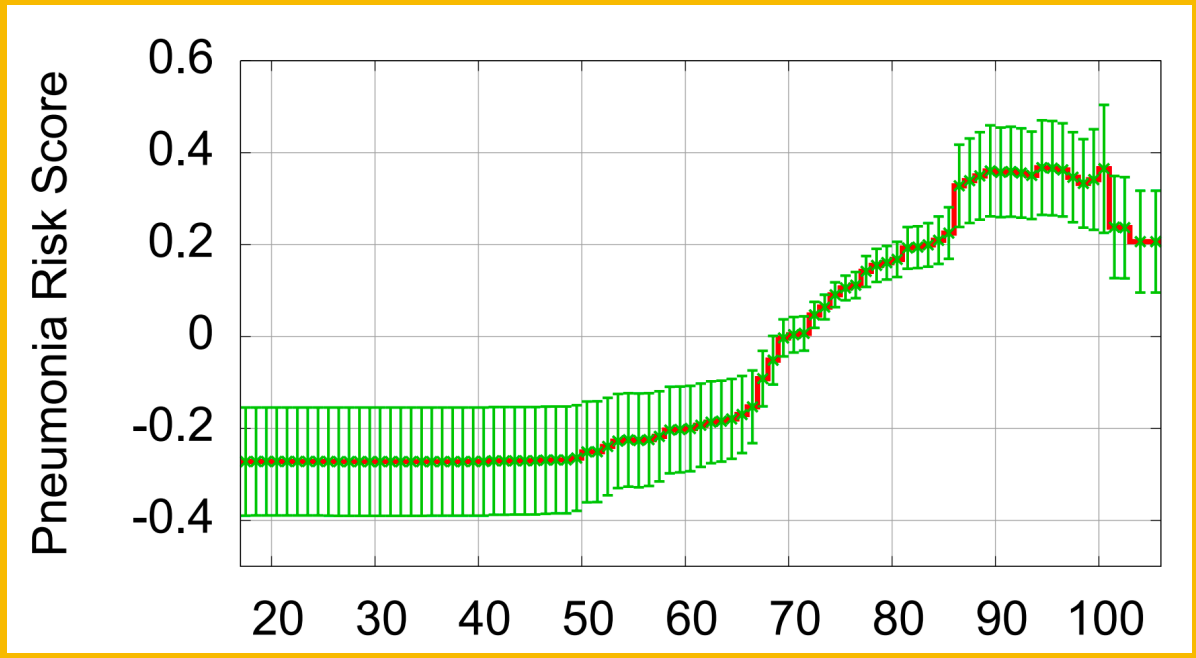
generalized linear model $g(y) = \beta_0 + \beta_1 x_1 + \ldots + \beta_n x_n$

generalized additive model $g(y) = f_1(x_1) + \ldots + f_n(x_n)$

generalized additive$^2$ model $g(y) = f_1(x_1) + \ldots + f_n(x_n) + \sum_{i \neq j} f_{ij}(x_i, x_j).$

Table from [Gehrke et al. '12]

# Bu




[Gehrke et al. '12]

Linear model

generalized linear mode

generalized additive mo

generalized additive$^2$ model

$$g(y) = f_1(x_1) + \ldots + f_n(x_n) + \sum_{i \neq j} f_{ij}(x_i, x_j).$$

Table from [Gehrke et al. '12]

# Which one is NOT 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 one is NOT 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

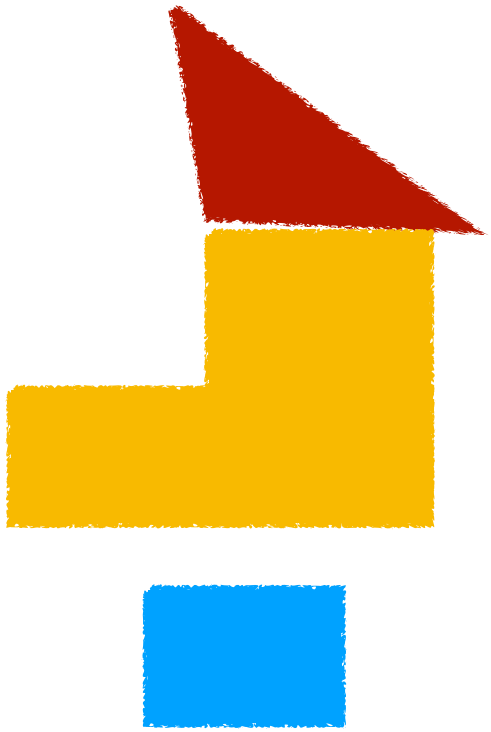
C. The data might not cluster

D. None of the above

- **Depth/Length of the tree might be too big**
- **Complexity of rules might be high**
- **May need lots of splits to fit complex function**

# Types of interpretable methods

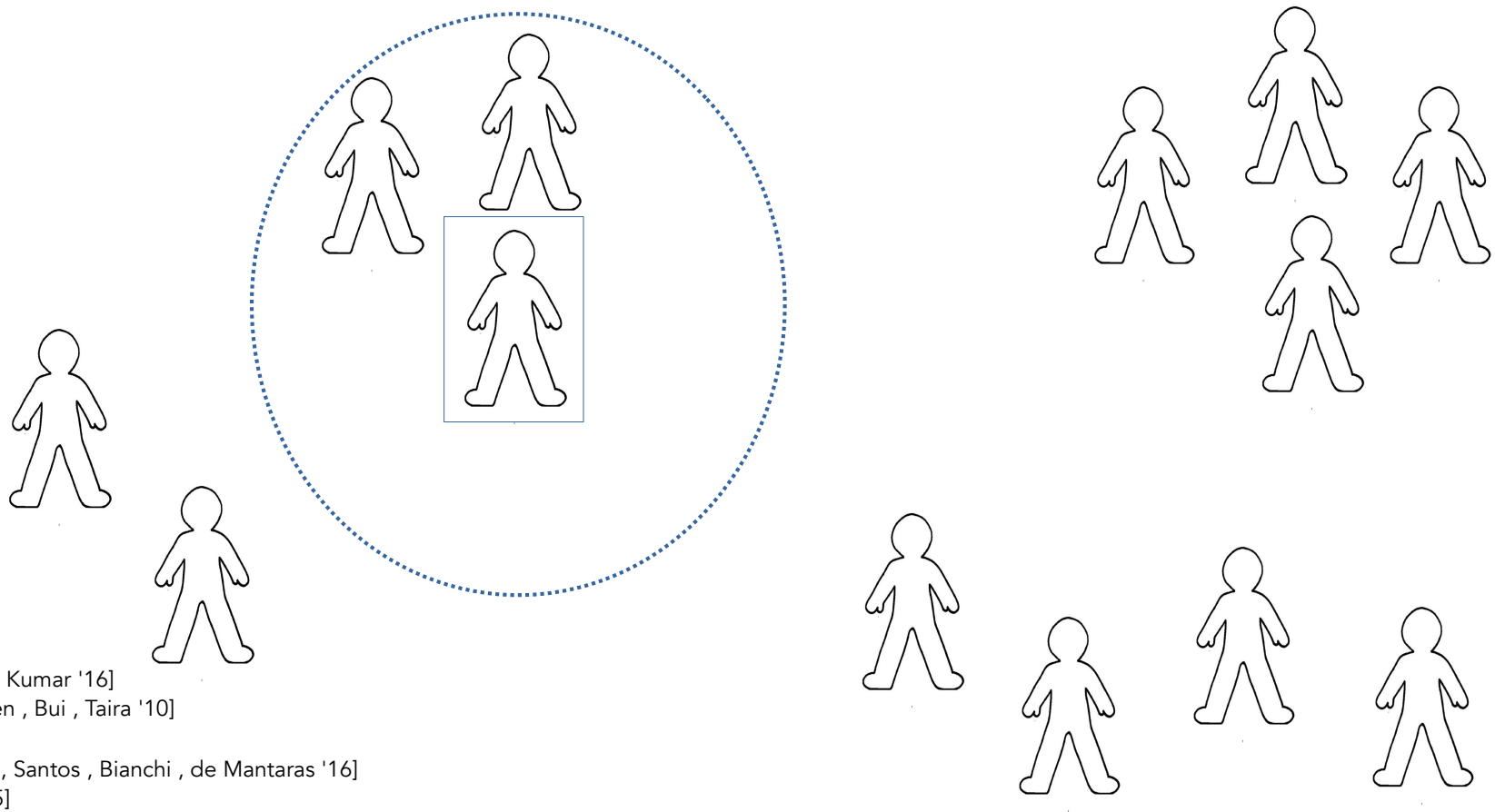
**Building a new model**



- rule-based, per-feature-based
- **case-based**
- sparsity
- monotonicity

# Building a new model: Case-based

"I recommend treatment X because it worked for other patients like you..."



[Frey, Dueck '10]
[Yen, Malioutov , Kumar '16]
[Arnold , El-Saden , Bui , Taira '10]
[Floyd , Aha '16]
[Homem, Perico , Santos , Bianchi , de Mantaras '16]
[Jalali , Leake '15]
[Reid , Tibshirani '16]

# Building a new model: Case-based

- Explain clustering results using examples (Bayesian Case Model)
- Joint inference on **prototypes**, **subspaces** and **cluster labels**

Cluster A

Cluster B

Cluster C

[K. Rudin, Shah '14]

# Building a new model: Case-based

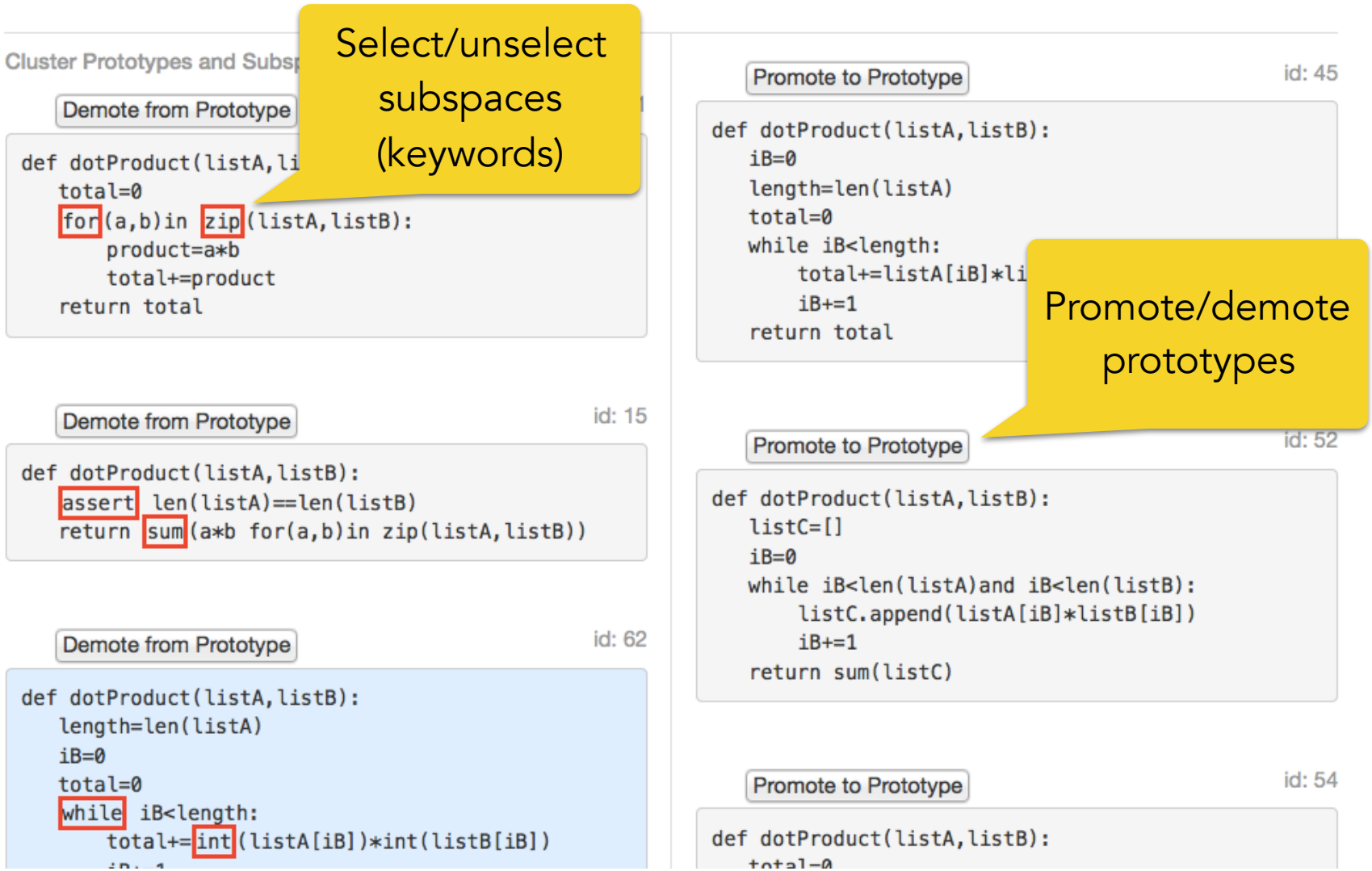
| | Cluster A | Cluster B | Cluster C |
|---|---|---|---|
| **prototypes** |  **Taco** |  **Basic crepe** |  **Chocolate berry tart** |
| **subspaces** | salsa<br>sour cream<br>avocado<br>salt, pepper, taco shell, lettuce, oil | flour<br>egg<br>water, salt, milk, butter | chocolate<br>strawberry<br>pie crust, whipping cream, kirsch, almonds |



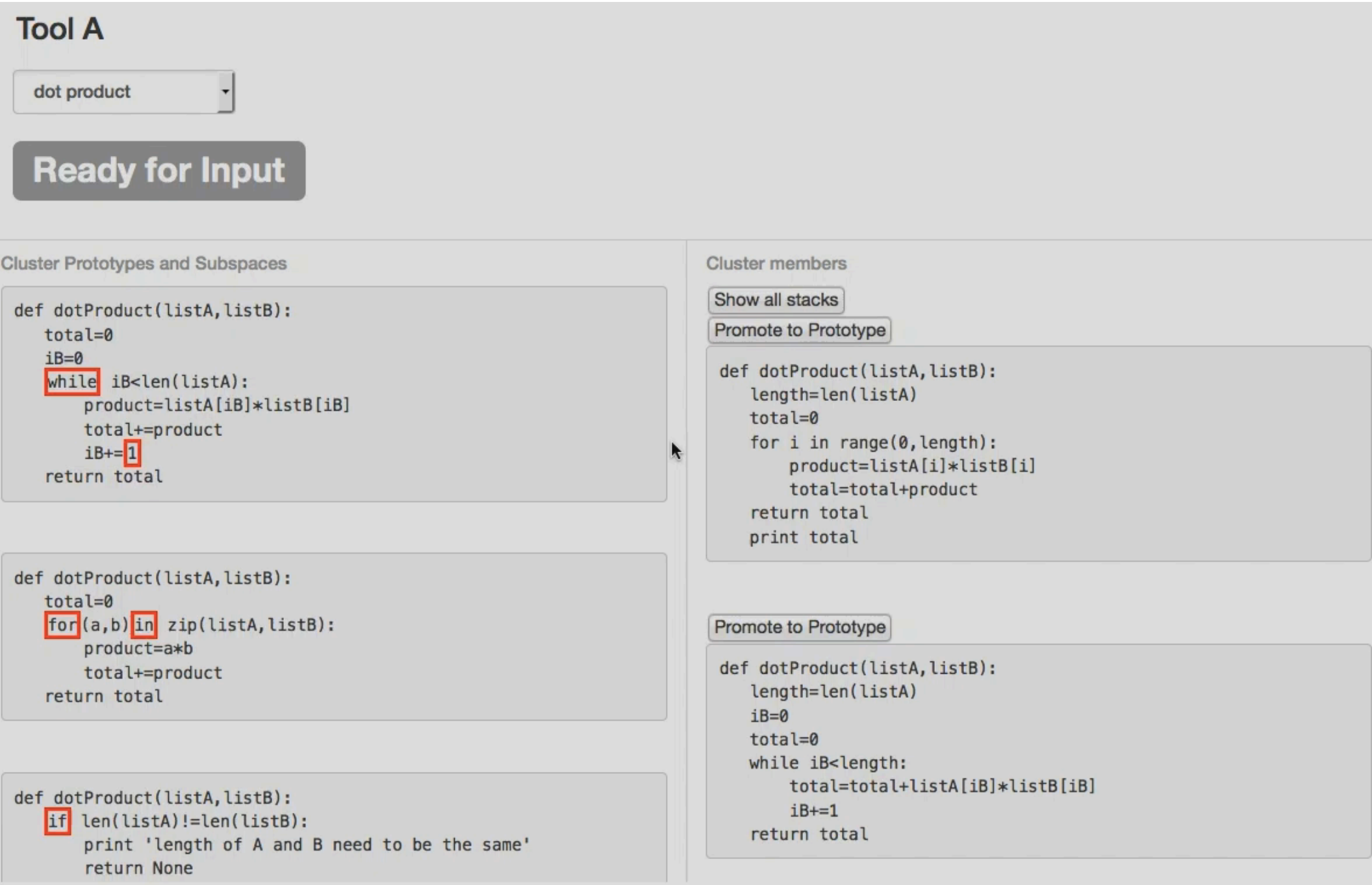
[K. Rudin, Shah '14]

# Building a new model: Case-based

## iBCM + OverCode system




[K. Glassman, Johnson, Shah '15]



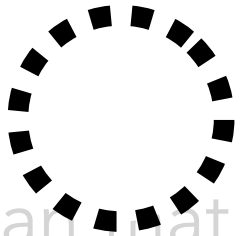

[K. Glassman, Johnson, Shah '15]



68

# Which one is NOT the limi 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

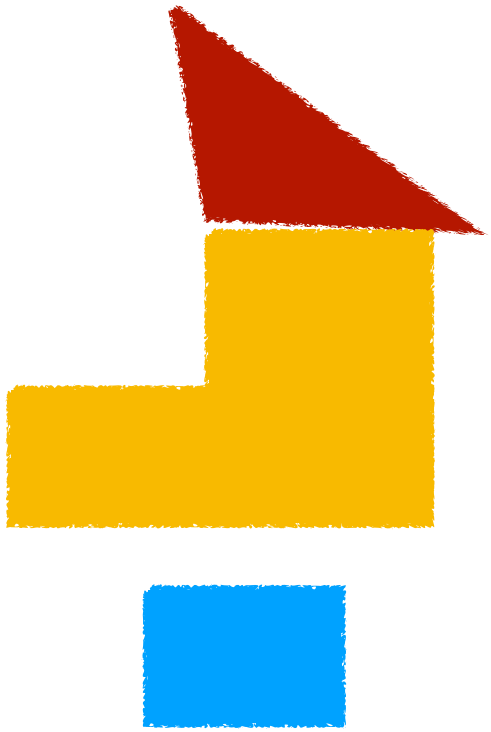
# Break-y
# 5 mins


After the break:

- Interpretability methods when you already have a model (e.g., deep learning)
- How to evaluate explanations

# Types of interpretable methods

**Building a new model**



- rule-based, per-feature-based
- case-based
- **sparsity**
- monotonicity

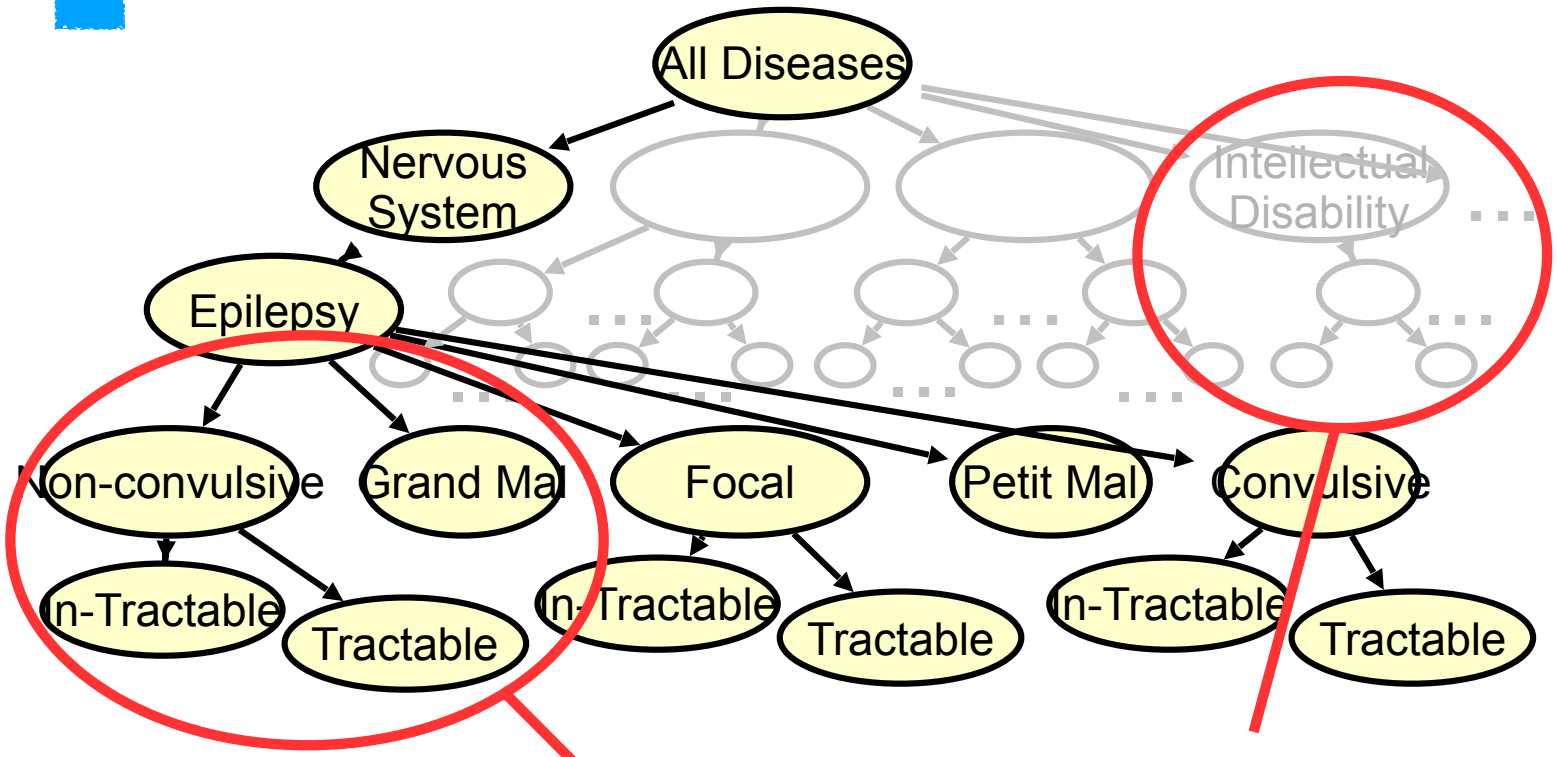
# Building a new model: Sparsity-based

$$y = a_0 + a_1 x_1 + a_{21} x_{21} + a_{1002} x_{1002}$$

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

[ Jain , Rao , Dhillon '16 ]
[ Yang , Barber , Jain , Lafferty '16 ]
[ Greenlaw , Szefer , Graham , Lesperance , Nathoo '17 ]
[ Kim , Xing '10 ]
[ Meier , Van De Geer , Buhlmann '08]
[ Bach , Jenatton , Mairal , Obozinski '12 ]

# Building a new model: Sparsity-based




Correlations across subtrees: may be a single cause manifesting in multiple aspects. Model that!

$$\Pr(\text{data}) = \text{Mult}(\ \Theta_n \ \text{patient-subtype} \ \ \Phi_k \ \text{Subtype-concept} \ \ T_c \ \text{concept-diagnosis})$$

Graph-sparse LDA [Doshi-Velez et al.'15]

# Which one is NOT the limitations of sparsity methods?

A. The model may not be able to represent what it learned in a sparse fashion.

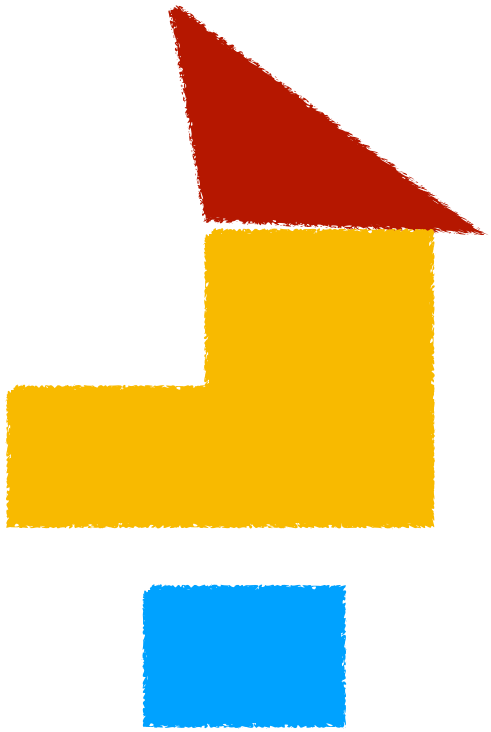
B. There might be the case that only the collections of factors make more sense

C. None of the above

"Sparsity is good, but not enough, but just because it is sparse, doesn't mean it's interpretable." [Freitas '10]

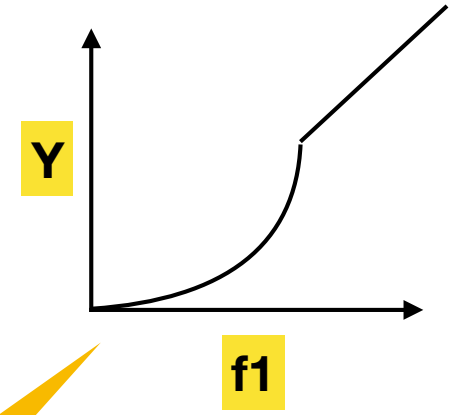
# Types of interpretable methods

**Building a new model**



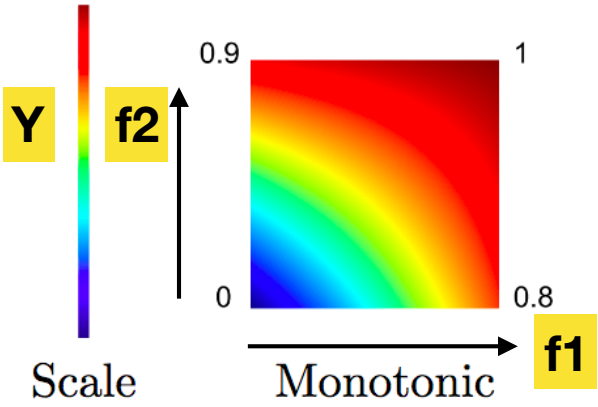
- rule-based, per-feature-based
- case-based
- sparsity
- **monotonicity**

# Building a new model: Monotonicity




Piecewise monotonic

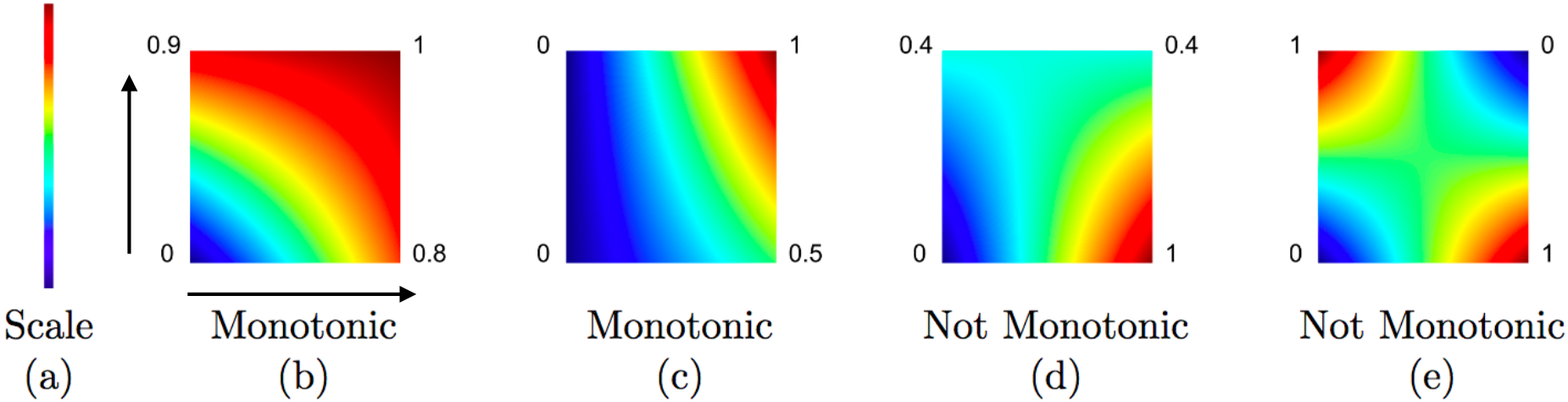
**One feature**




**Two features**


[Gupta et al. '16]

# Building a new model: Monotonicity



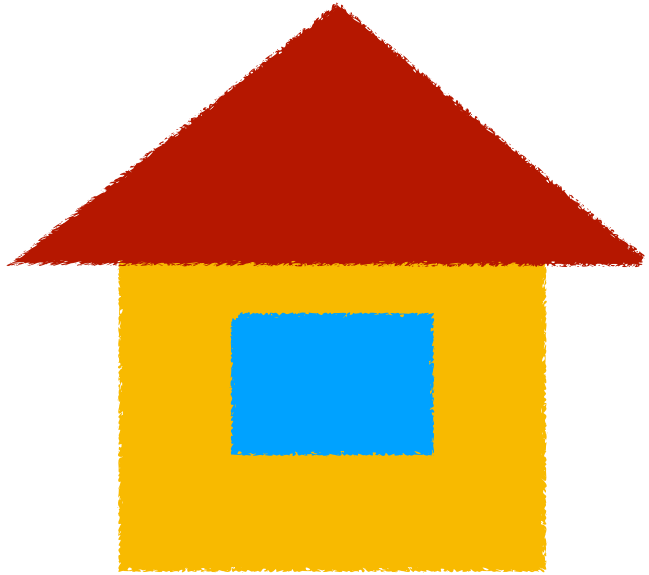

- Learn piecewise monotonic function within a user specified lattice (intervals) [Gupta et al. '16]

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

# Types of interpretable methods

- Sensitivity analysis, gradient-based methods
- mimic/surrogate models
- Investigation on hidden layers

**After building a model**



# Types of interpretable methods

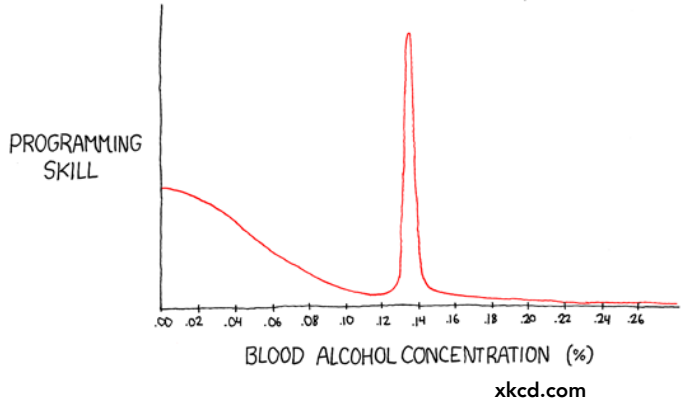
- **Sensitivity analysis, gradient-based methods**
- mimic/surrogate models
- Investigation on hidden layers

**After building a model**

# After building a model: Sensitivity analysis

What would happen to output $\hat{y}$

If we perturb the input $x \longrightarrow x + \epsilon$ ?

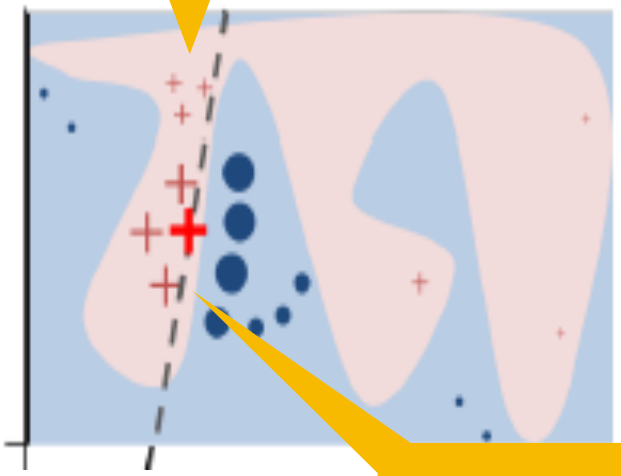



- $\epsilon$ can be group of features, data points, specific inputs
- For nonlinear functions $\hat{y} = f(x)$, higher order derivatives will get involved…

# After building a model: Sensitivity analysis

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

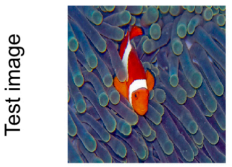
Want local explanation of the + data point



Locally fitted linear function

Influential functions
[Koh et al.'17]

To classify this image:




Model found these images most helpful

SVM

Inception

[Simonyan et al., '13]
[Li et al., '16]
[Datta et al. '16]
[Adler et al., '16]

# After building a model: Saliency/attribution Maps

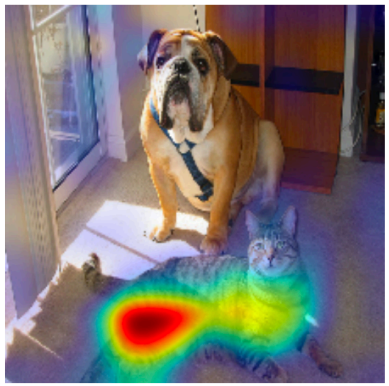
- Give me the features in the input space that mattered for the classification

$$\frac{\partial y}{\partial x_{ij}}$$

# After building a model: Saliency/attribution Maps

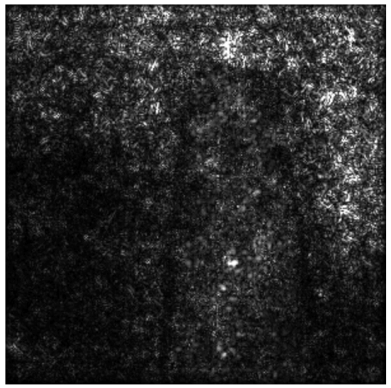
**Grad-CAM [Selvaraju et al. 16]**

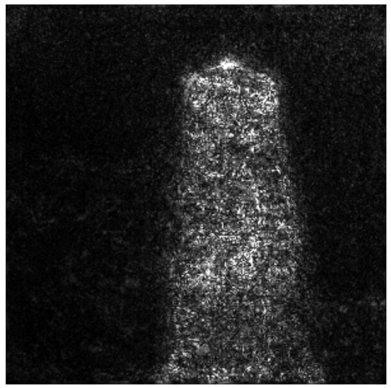
(a) Original Image



(c) Grad-CAM 'Cat'

**SmoothGrad [Smilkov et al. 17]**

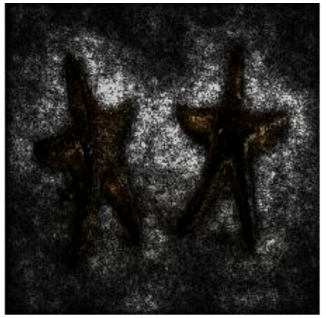



**Integrated gradients [Sundararajan et al. 17]**

Top label: starfish

Score: 0.999992



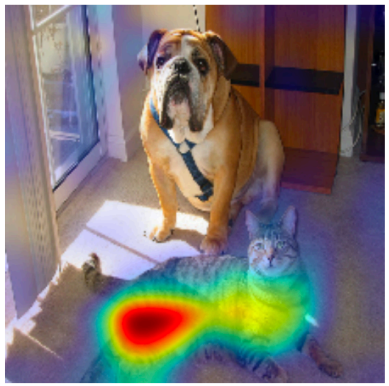


[Erhan 2009] [Springenberg, '14] [Shrikumar '17]

# After building a model: Saliency/attribution Maps

**Grad-CAM [Selvaraju et al. 16]**

(a) Original Image



(c) Grad-CAM 'Cat'

**SmoothGrad [Smilkov et al. 17]**



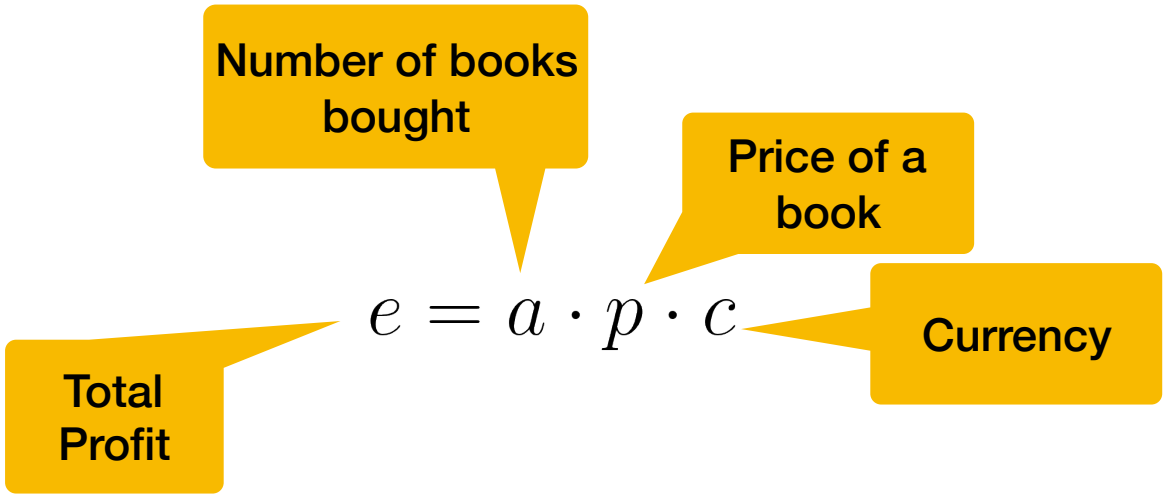
**Integrated**

Oh yeah, gradients makes sense. It's about how much the label would changes as I change the data…

[Erhan 2009] [Springenberg, '14] [Shrikumar '17]

# Pop quiz

Idea borrowed from [Sundararajan et al. 17]

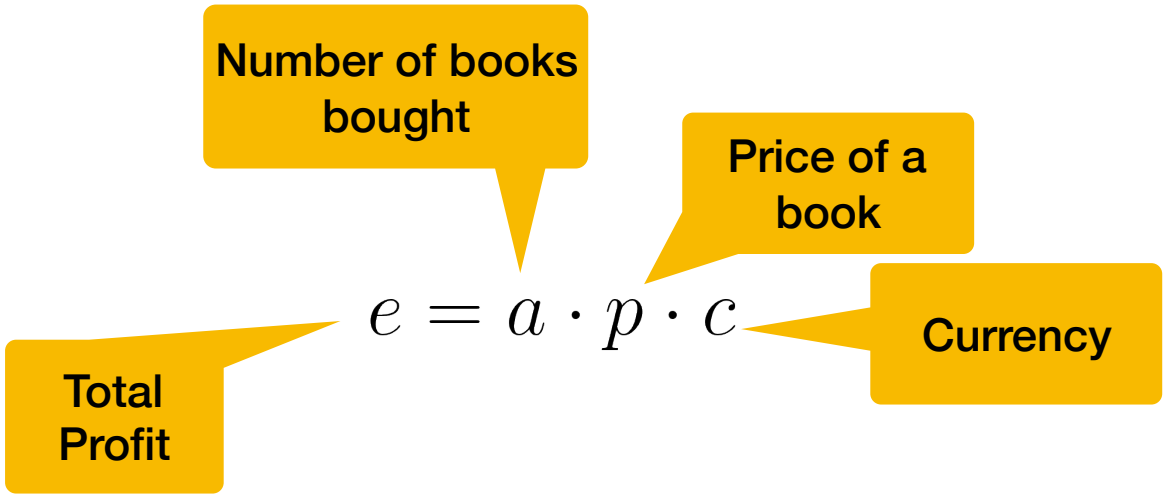



| | 2016 | 2017 |
|---|---|---|
| a | 4 | 5 |
| p | 1 | 2 |
| c | 3 | 4 |
| e | 12 | 40 |

**Increase in e: 28!**

# Pop quiz

Idea borrowed from [Sundararajan et al. 17]




| | 2016 | 2017 | Only this feature changed |
|---|---|---|---|
| a | 4 | 5 | (5-4)*1* 3 = 3 |
| p | 1 | 2 | 4*(2-1)*3 = 12 |
| c | 3 | 4 | 4*1*(4-3) = 4 |
| e | 12 | 40 | |

**Increase in e: 28!**

# Pop quiz

Idea borrowed from [Sundararajan et al. 17]

Number of books bought

Price of a book

$$e = a \cdot p \cdot c$$

Total Profit

Currency

| | 2016 | 2017 | Only this feature changed |
|---|---|---|---|
| a | 4 | 5 | (5-4)*1* 3 = 3 |
| p | 1 | 2 | 4*(2-1)*3 = 12 |
| c | 3 | 4 | 4*1*(4-3) = 4 |
| e | 12 | 40 | 19 |

What?!

Increase in e: 28!

Where is my 9?

# Which one is NOT 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

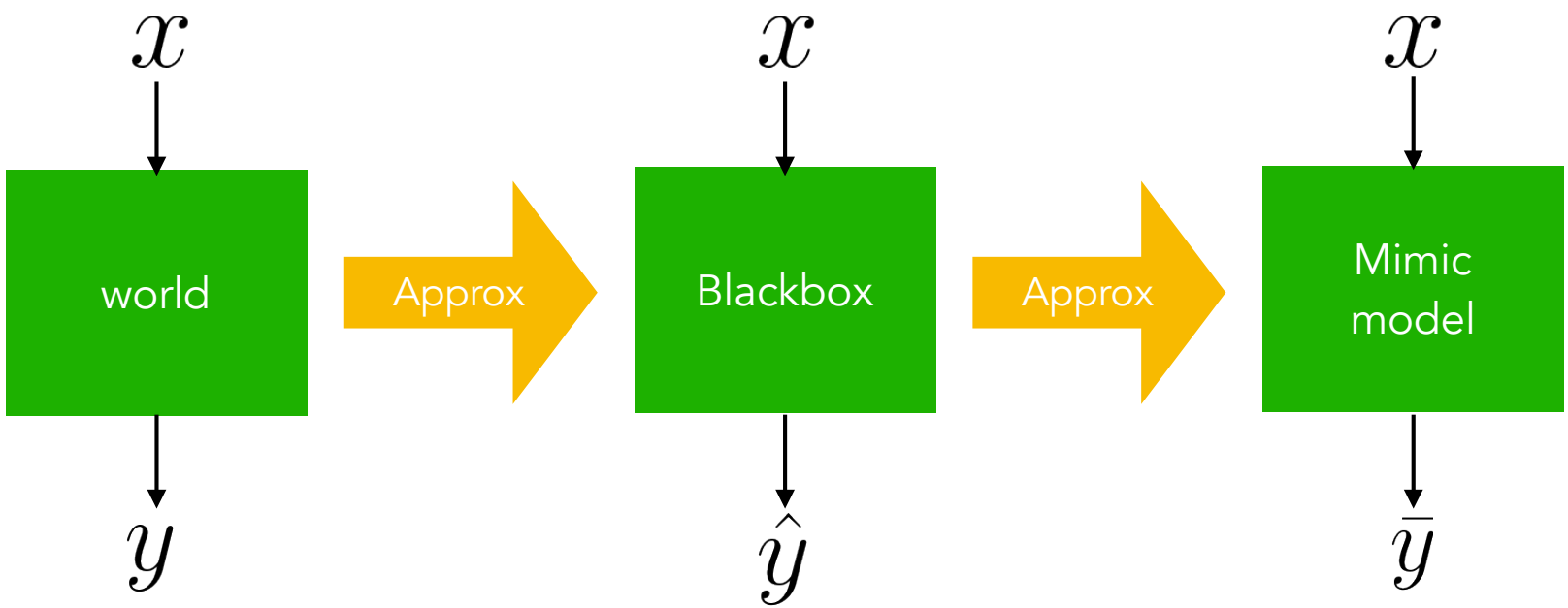
# Types of interpretable methods

- Sensitivity analysis, gradient-based methods
- **mimic models**
- Investigation on hidden layers

**After building a model**

# After building a model: Mimic models

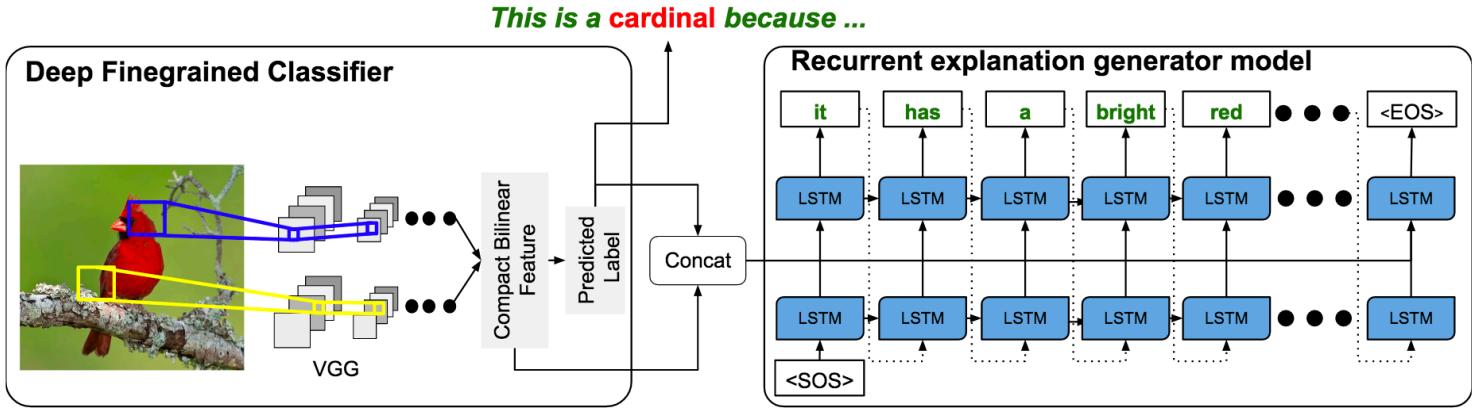
- Train a black box on $x$ and $y$ : $f(x) = \hat{y}$
- Train an interpretable model on $x$ and $\hat{y}$ : $f(x) = \bar{y}$



# After building a model: Mimic models

- Model compression or distillation [Bucila et al. ’06, Ba et al. ’14, Hinton et al. ’15]

- Visual explanations [Hendricks et al. ’16]



# Which one is NOT the limitations of mimic models?

A. You may not be able to distill - there may not be simpler model at all

B. There might be a gap between what the actual model is doing and your mimic model is doing

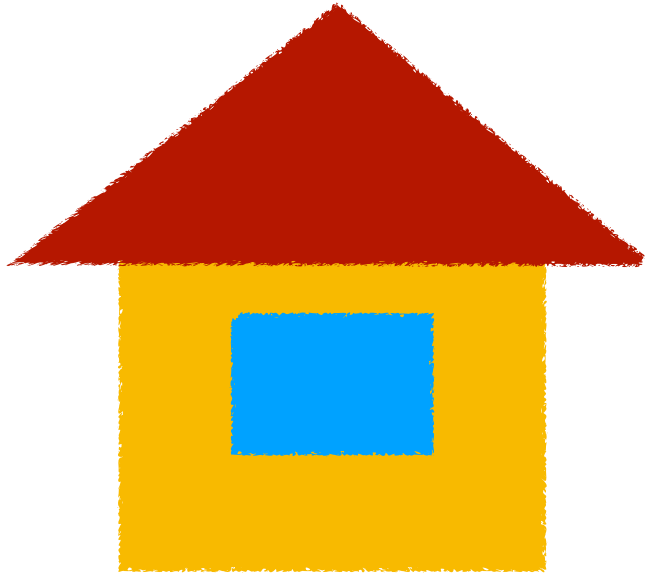
C. The simpler model may not be interpretable

D. None of the above

# Types of interpretable methods

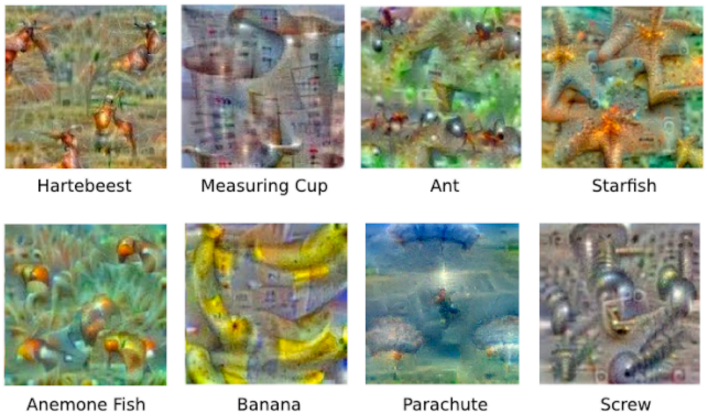
- Sensitivity analysis, gradient-based methods
- mimic/surrogate models
- **Investigation on hidden layers**

**After building a model**

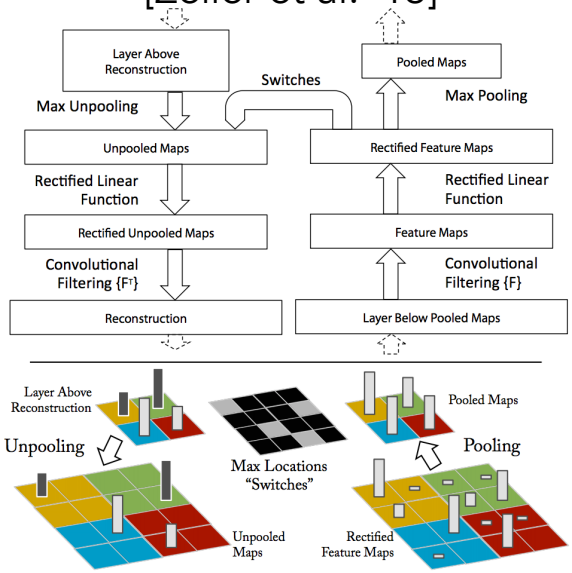


# After building a model: Investigation on hidden layers

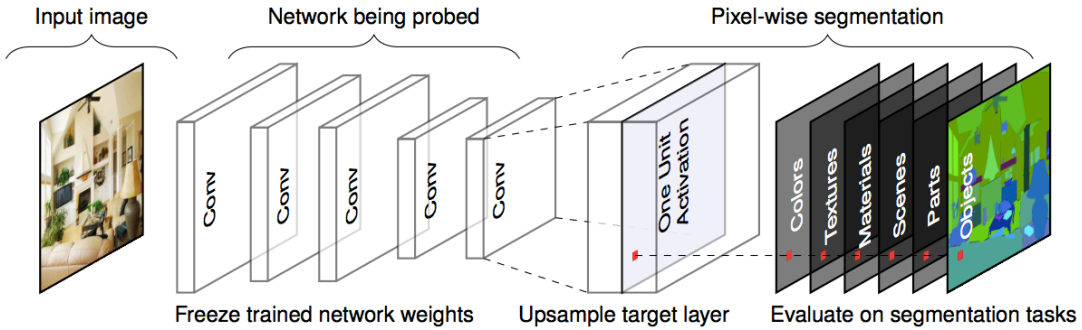
[Dosovitskiy et al. '16]




[Zeiler et al. '13]




[Bau and Zhou et al. '17]

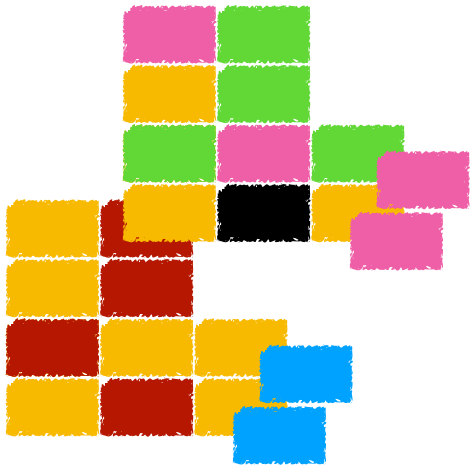


# Which one is NOT the limitations of investigation on hidden layers?

A. They may be lack of actionable insights
B. It is unclear if visualizing neuron vs. per layer vs. per subspaces is more meaningful than others
C. A golden dataset with detailed labels with human concepts are often not available
D. None of the above

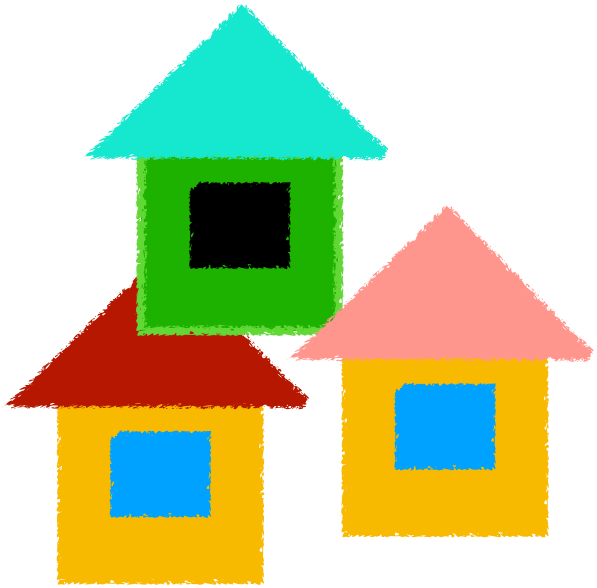
# What's the best interpretability method for me?

?



?

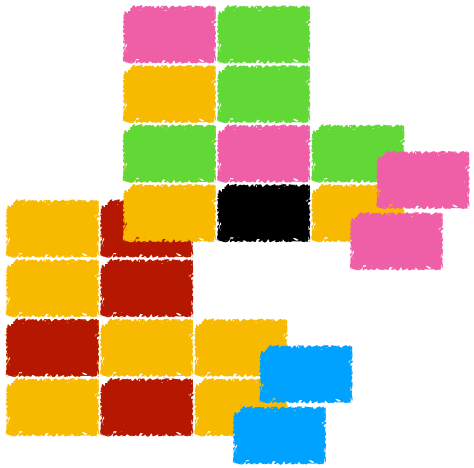
?

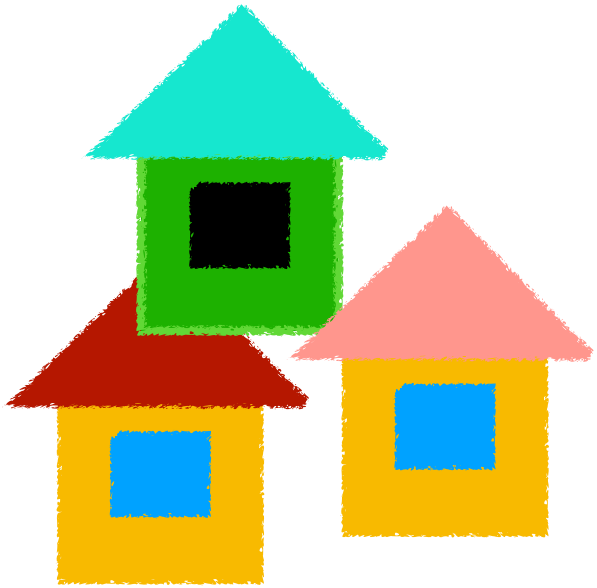


# What's the **best** interpretability method for me?

3. How can we measure 'good' explanations?

? ? ?





# Agenda

1. Why and when?

2. How can we do this?

Interpretation is the process of giving **explanations**

3. How can we measure 'good' explanations?

To Humans

# How are we measuring explanation quality now?

"You know it when you see it"

# How are we measuring explanation quality now?

"You know it when you see it"


https://www.pinterest.se/pin/365987907189893478/

# How are we measuring explanation quality now?

"You know it when you see it"

Generalized additive models (GAMs) are the gold standard for intelligibility when low-dimensional terms are considered [4, 5, 6]. Standard GAMs have the form

$$g(E[y]) = \beta_0 + \sum f_j(x_j), \qquad (1)$$

where $g$ is the link function and for each term $f_j$, $E[f_j] = 0$. Generalized linear models (GLMs), such as logistic regres-

These are great papers and I had definitely also made these claims in my work!

accurate, yet are highly interpretable. These predictive models will be in the form of sparse *decision lists*, which consist of a series of *if*...*then*... statements where the *if* statements define a partition of a set of features and the *then* statements correspond to the predicted outcome of interest. Because of this form, a decision list model naturally provides a reason for

# How are we measuring explanation quality now?

"You know it when you see it"

Generalized additive models (GAMs) are the gold standard for intelligibility when low-dimensional terms are considered [4, 5, 6]. Standard GAMs have the form

$$g(E[y]) = \beta_0 + \sum f_j(x_j), \qquad (1)$$

where $g$ is the link function and for each term $f_j$, $E[f_j] = 0$. Generalized linear models (GLMs), such as logistic regres-

accurate, yet are highly interpretable. These predictive models will be in
the form of sparse *decision lists*, which consist of a series of *if*...*then*...
statements where the *if* statements define a partition of a set of features
and the *then* statements correspond to the predicted outcome of interest.
Because of this form, a decision list model naturally provides a reason for

We want **evidence-based** so that we can compare work A to work B, and to generalize.

# How are we measuring explanation quality now?

"You know it when you see it"

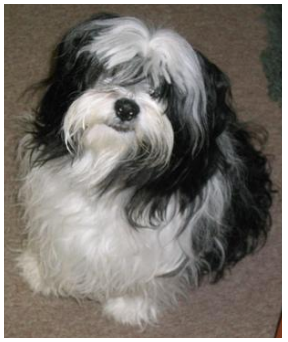
Give human a task, then measure how well they do

Generalized additive models (GAMs) are the gold standard for intelligibility when low-dimensional terms are considered [4, 5, 6]. Standard GAMs have the form

$$g(E[y]) = \beta_0 + \sum f_j(x_j), \qquad (1)$$

where $g$ is the link function and for each term $f_j$, $E[f_j] = 0$. Generalized linear models (GLMs), such as logistic regres-

accurate, yet are highly interpretable. These predictive mod
the form of sparse *decision lists*, which consist of a series o
statements where the *if* statements define a partition of a s
and the *then* statements correspond to the predicted outcom
Because of this form, a decision list model naturally provides



Q. Which group does this new data point belong to?

A. Group 1


B. Group 2

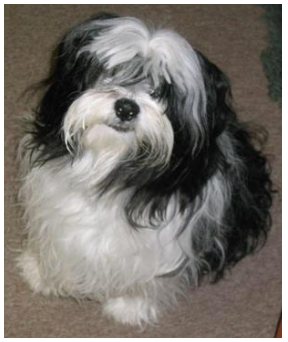

[K. 16]

# How are we measuring explanation quality now?

"You know it when you see it"

Give human a task, then measure how well they do

We want a measurement methods that can be **generalized.**



Q. Which group does this new data point belong to?

A. Group 1


B. Group 2

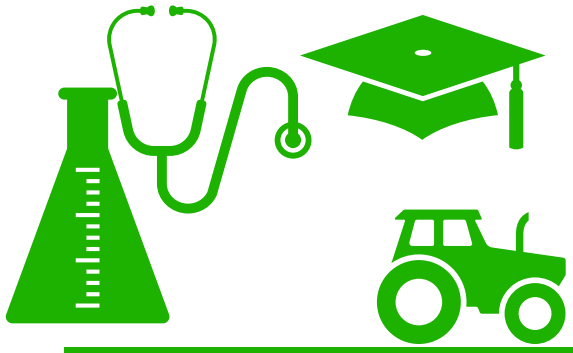

[K. 16]

# Spectrum of evaluation

**Machine Learning**

| Function-based | Application-based |
| --- | --- |
| a variety of synthetic and standard benchmarks<br>e.g, UCI datasets, imagenet | Backing up claims<br>e.g., performance on a cool medical dataset, winning Go games |

# Spectrum of evaluation



Interpretable Machine Learning

Function-based

How sparse are the features?

Does it look reasonable?

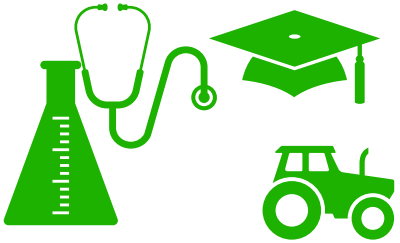
Application-based

How much did we improve patient outcomes?

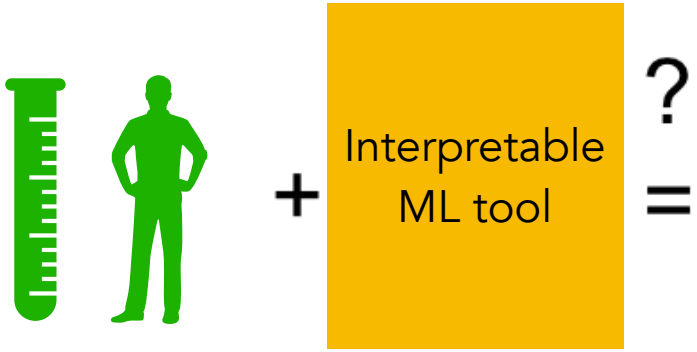
Do scientists find the explanations useful?

Quantitative Qualitative

# Evaluate: application-based



- Does providing interpretability assist with a down-stream task, such as increasing fairness, safety, scientific discovery, or productivity?

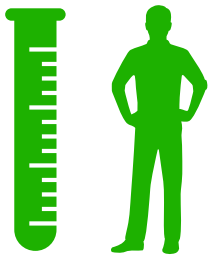



**Commonly used in HCI, visualization communities**


**https://www.bu.edu/today/2012/do-scholarly-articles-want-to-be-free/**

# Evaluate: application-based

- Does providing interpretability assist with a down-stream task, such as increasing fairness, safety, scientific discovery, or productivity?



+

Interpretable ML tool

?

=


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

# Evaluation: Function-based

- Can we use some proxy such as sparsity monotonicity or non-negativity?

# Evaluation: Function-based

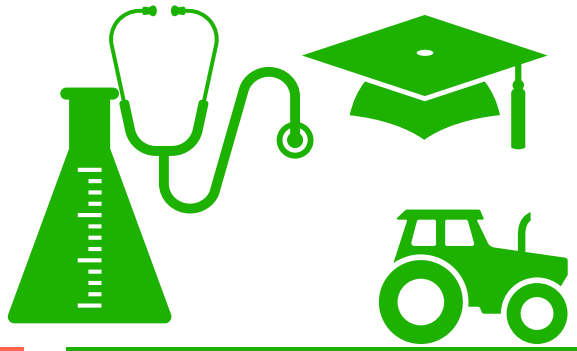
- Can we use some proxy such as sparsity monotonicity or non-negativity?

It's easy to formalize, optimize, and evaluate…

but may not solve a real need

e.g., 5 unit sparsity will save more patients than 10 unit sparsity?

# Spectrum of evaluation



Interpretable Machine Learning

Quantitative Qualitative

| Function-based | Cognition-based | Application-based |
|---|---|---|
| How sparse are the features? | | How much did we improve patient outcomes? |
| Does it look reasonable? | | Do scientists find the explanations useful? |

**Low cost** → **High cost**

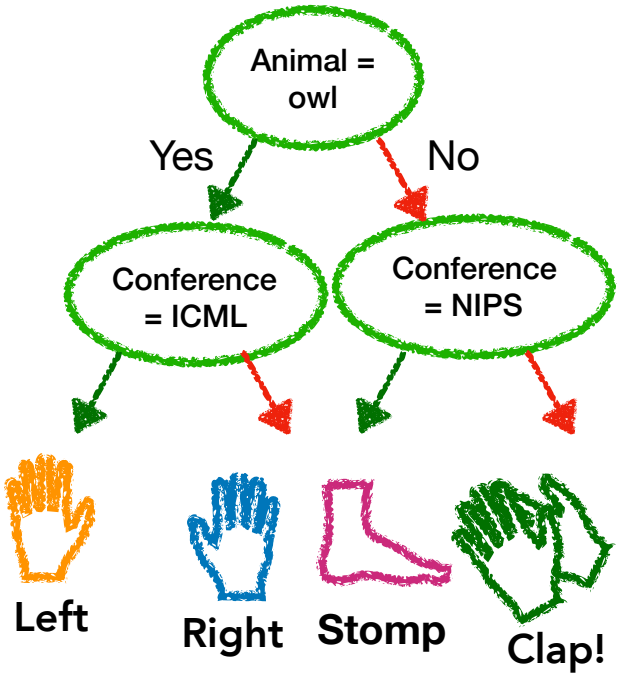
**Low validity** ← **High validity**

# Spectrum of evaluation

| Function-based | Cognition-based | Application-based | |
|---|---|---|---|
| How sparse are the features? | What factor should change to change the outcome? | How much did we improve patient outcomes? | Quantitative |
| Does it look reasonable? | What are the discriminative features? | Do scientists find the explanations useful? | Qualitative |

# Evaluations: cognition-based

- Human subject experiments on general forms



# Evaluations: cognition-based

- Human subject experiments on general forms

**What { Input / Weight / Cost } Would change { Predictions for x / Cluster of x } ?**

**e.g.,**
**Forward simulation,**
**Counterfactual reasoning**
**Identify Important features**

# Spectrum of evaluation

## Problem-related Factors

1. Global vs. Local
2. Time budget
3. Severity of underspecification

## Method-related factors

4. Cognitive chunks
5. Audience training

Cognition-based

What factor should change to change the outcome?

What are the discriminative features?

Quantitative Qualitative

# Spectrum of evaluation

Problem-related Factors

1. Global vs. Local
2. Time budget
3. Severity of underspecification

Method-related factors

4. Cognitive chunks
5. Audience training

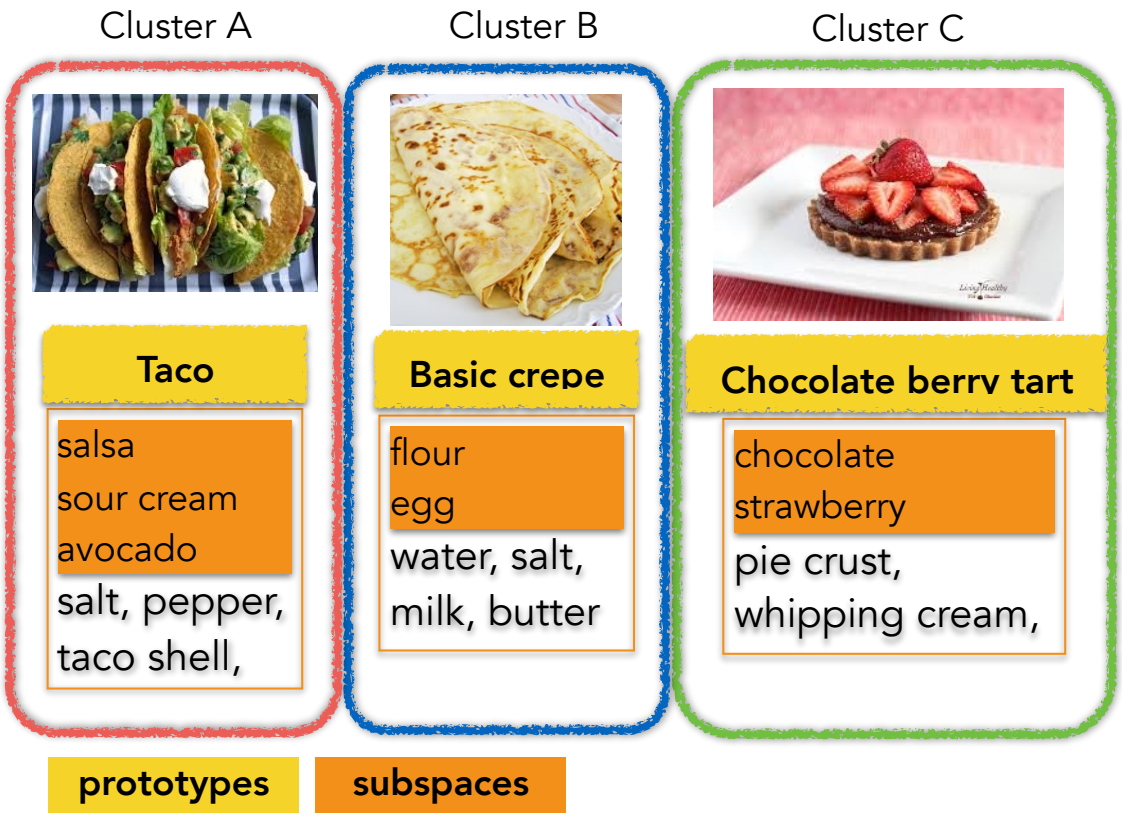
Cognition-based

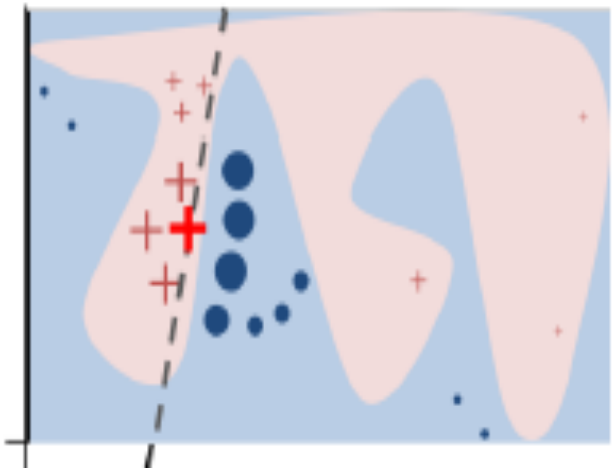
What factor should change to change the outcome?

What are the discriminative features?

e.g.,
Humans capacity as function of factors,
Set of factors that carries over well to application

# Problem-related factor: global vs. local



# Problem-related factor: time budget

https://www.greek-names.info/names-of-ancient-greek-astronomers/

http://www.idonme.com/application-medical.php

# Problem-related factor: severity of underspecification

solve f(x) + a bounded term

(stop if obstacle within 2m)

vs.

make a safe autonomous car

solve AI

# Method-related factor: Cognitive Chunks

7 plus minus 2 or what?

LMCI,AIU,PSNI

vs.

ICML,UAI,NIPS

# Method-related factor: Audience Training

http://www.ufo-blogger.com

- The expert's background will affect what cognitive chunks and relations they have available

# Spectrum of evaluation

**Recommendations**

List these factors in your work so that others can compare your work to theirs.

Find more factors.

Problem-related Factors

1. Global vs. Local
2. Time budget
3. Severity of underspecification

Method-related factors

4. Cognitive chunks
5. Audience training

Cognitive-based

What factor should change to change the outcome?

Does it look reasonable?

What are the discriminative features?

Do scientists find the explanations useful?

ualitative

# Wrap up

1. Why and when?

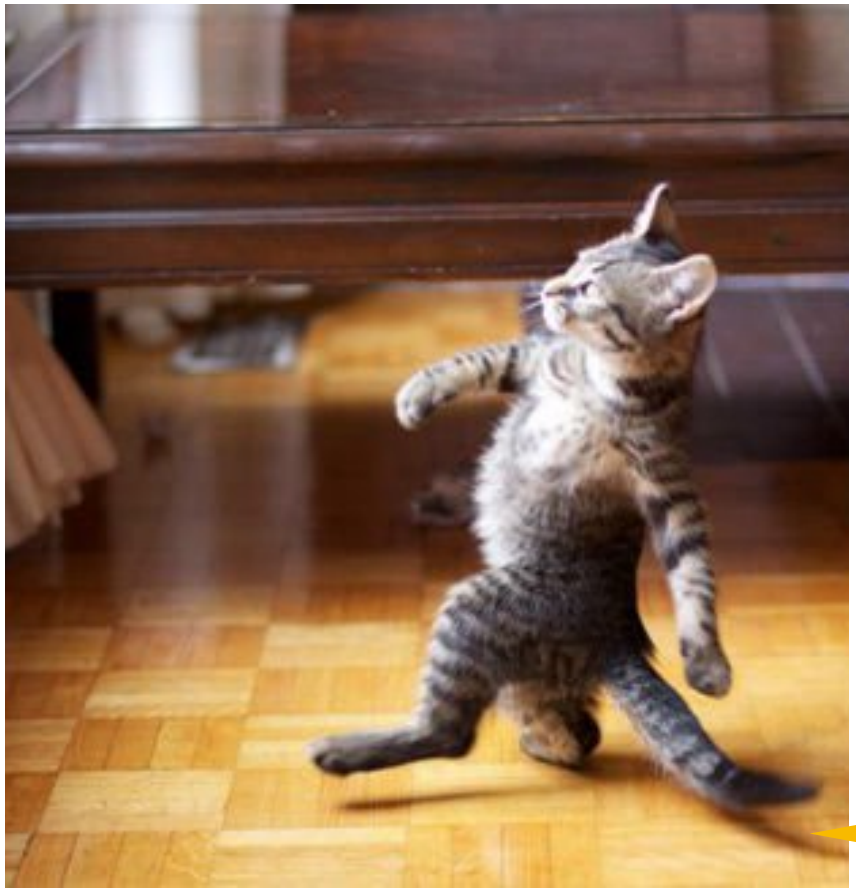
2. How can we do this?

Interpretation is the process of giving **explanations**

3. How can we measure 'good' explanations?

To Humans

# How shall we move the field forward?



https://imgflip.com

…and PAIR @ Brain, we are hiring.

**ai.google/pair**

# QnA

**Recommendations**

List these factors in your work so that others can compare your work to theirs.

Find more factors.

Problem-related Factors

1. Global vs. Local
2. Time budget
3. Severity of underspecification

Method-related factors

4. Cognitive chunks
5. Audience training

Cognitive-based

What factor should change to change the outcome?

Does it look reasonable?

What are the discriminative features?

Do scientists find the explanations useful?

ualitative