03 | Introduction to Machine Learning

Cynthia Rudin | MIT Sloan School of Management

Introduction to Machine Learning

- Classification
- Regression
- Statistical Learning Theory for Supervised Learning
- Clustering
- Recommender Systems

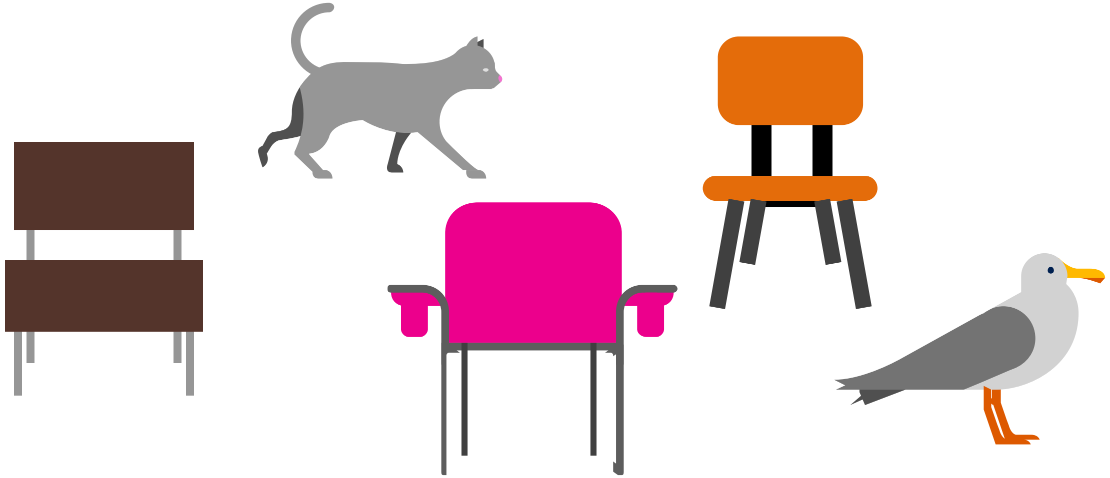
Classification

Microsoft

Cynthia Rudin | MIT Sloan School of Management

Machine Learning

- Grew out of artificial intelligence within computer science. Teaches computers by example.

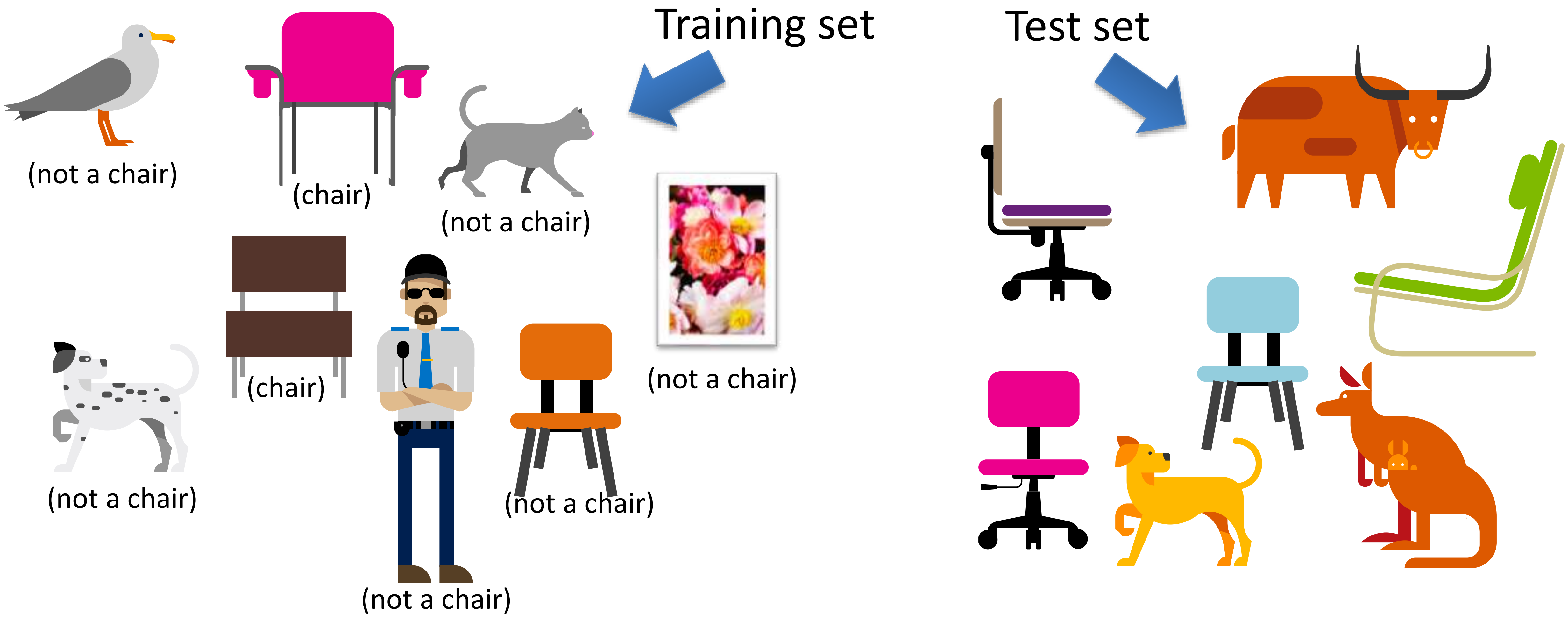


Is this a chair?

- ML is arguably part of statistics.

Classification

We have a *training set* of observations (e.g., labeled images) and a *test set* that we use only for evaluation.



Classification

- Each observation is represented by a set of numbers (features).

Each pixel gets rgb values like [1.0,0.9,0.8]

Image becomes:
[1.0,0.9,0.8,0.1,0.5,...]

(Label is -1, it's not a chair)

Classification

- Each observation is represented by a set of numbers (features).

Manhole is represented as: [5 3 120 12 1 0 …..]

Number of events last year

Number of serious events last year

Number of electrical cables

Number of pre-1930 electrical cables

Vented cover?

Inspected?

Classification

- Each observation is represented by a set of numbers (features).

Manhole is represented as: [5 3 120 12 1 0 …..]

Number of events last year
Number of serious events last year
Number of electrical cables
Number of pre-1930 electrical cables
Vented cover?
Inspected?

Training feature data is from 2014 and before

Classification

- Each observation is represented by a set of numbers (features).

Manhole is represented as: [5 3 120 12 1 0 …..]

Number of events last year
Number of serious events last year
Number of electrical cables
Number of pre-1930 electrical cables
Vented cover?
Inspected?

Training feature data is from 2014 and before
Label is 1 if it had an event in 2015

Classification

- Each observation is represented by a set of numbers (features).

Manhole is represented as: [5 3 120 12 1 0 …..]

Number of events last year
Number of serious events last year
Number of electrical cables
Number of pre-1930 electrical cables
Vented cover?
Inspected?

Testing feature data is from 2015 and before
Predict what happen in 2016

Classification

- Each observation is represented by a set of numbers (features).

Manhole is represented as:	[5	3	120	12	1	0	.....]		-1
	[0	0	89	5	1	1	.....]		1
	[1	0	20	0	0	1	.....]		-1
:			:						

Features, called X

Labels, called Y

(Predictors, Covariates, Explanatory Variables, Independent Variables)

Classification

- Formally, given training set $(x_{i,} y_i)$ for i=1…n, we want to create a classification model f that can predict label y for a new x.

Classification

- Formally, given training set $(x_{i,} y_i)$ for $i=1\ldots n$, we want to create a classification model f that can predict label y for a new x.

Manhole is represented as: [1925 15]

Year oldest cable installed

Number of events last year

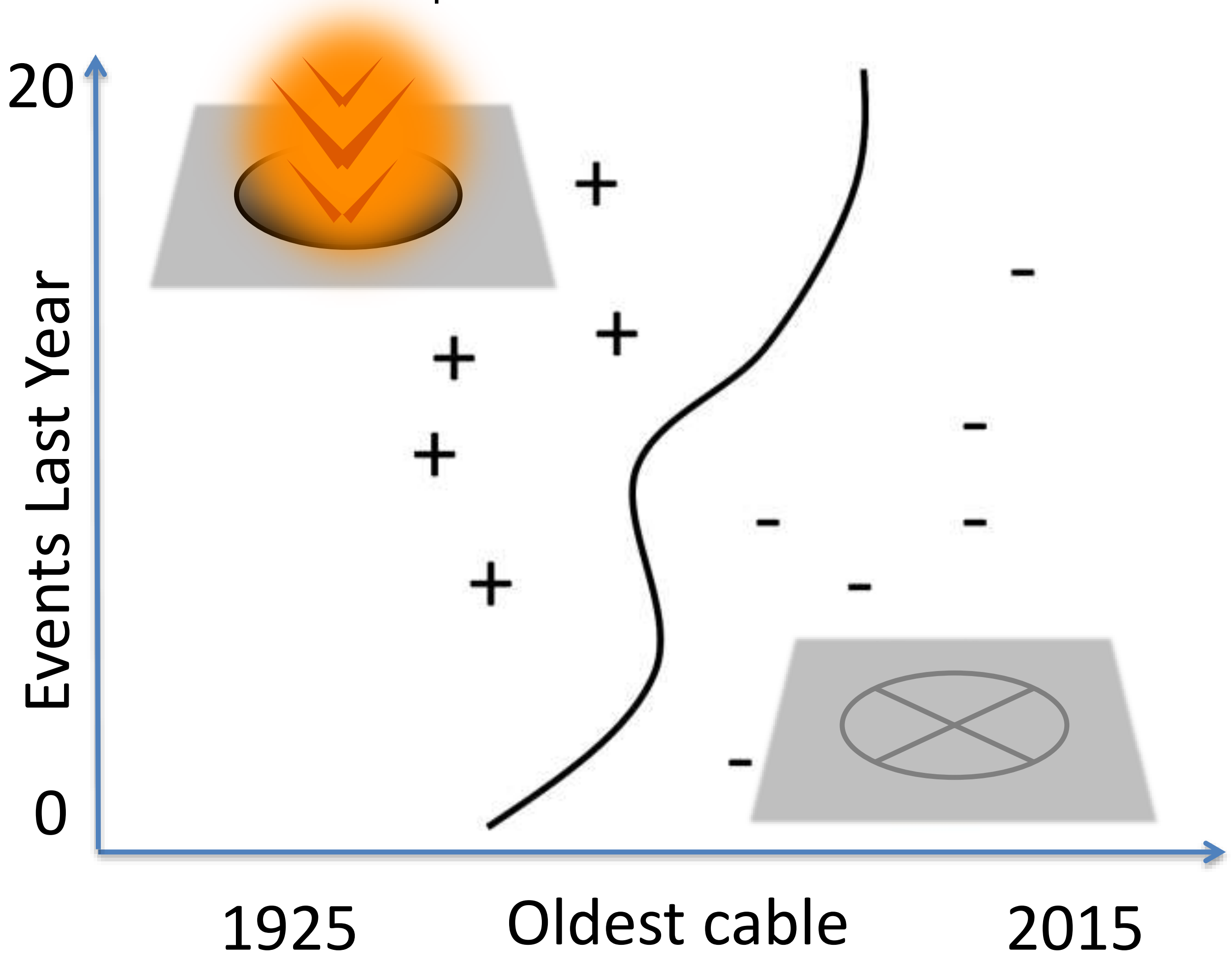
Classification

- Formally, given training set (x_i, y_i) for i=1…n, we want to create a classification model f that can predict label y for a new x.

Manhole is represented as: [1925 15]

Year oldest cable installed

Number of events last year

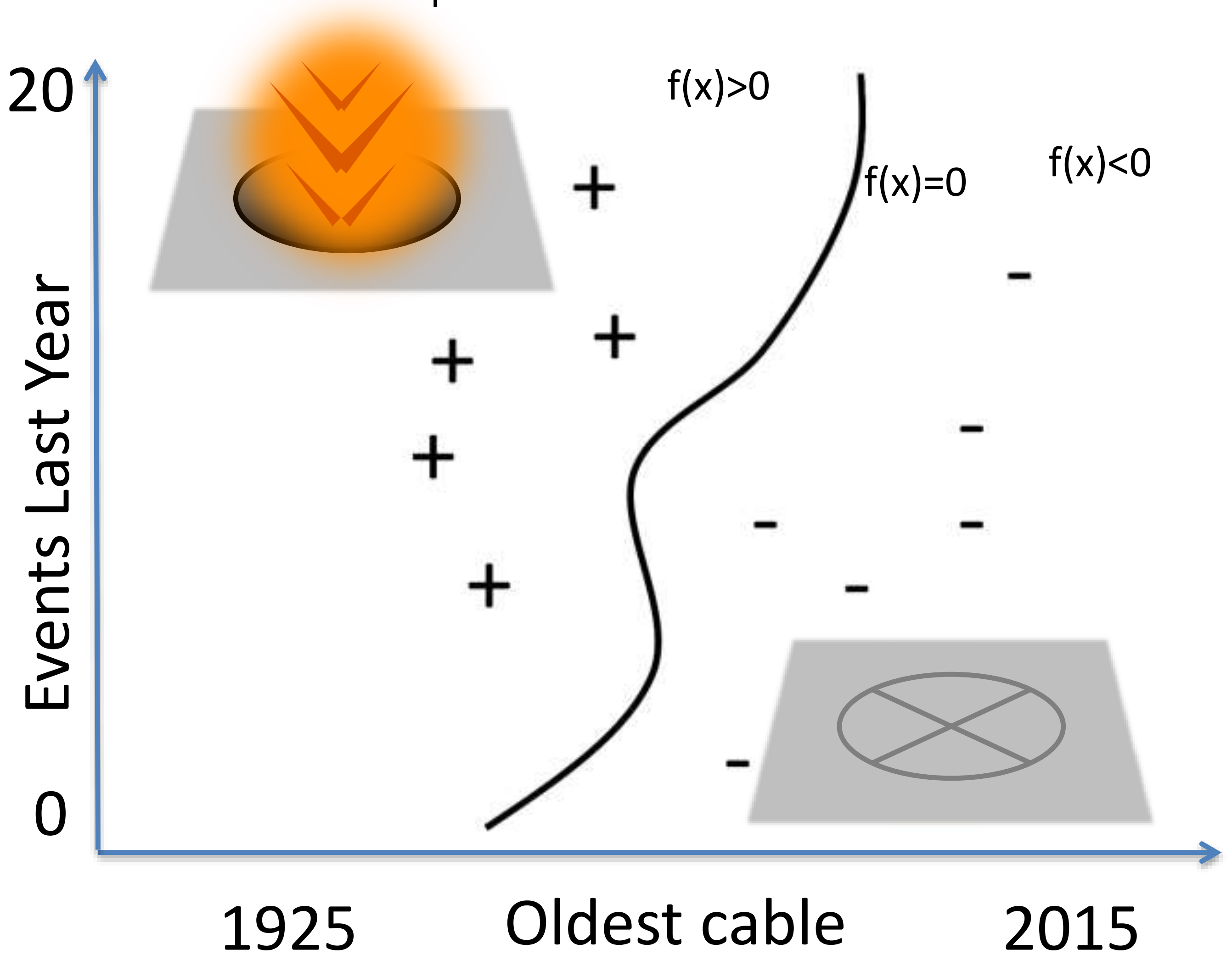


Classification

- Formally, given training set (x_i, y_i) for i=1…n, we want to create a classification model f that can predict label y for a new x.

Manhole is represented as: [1925 15]

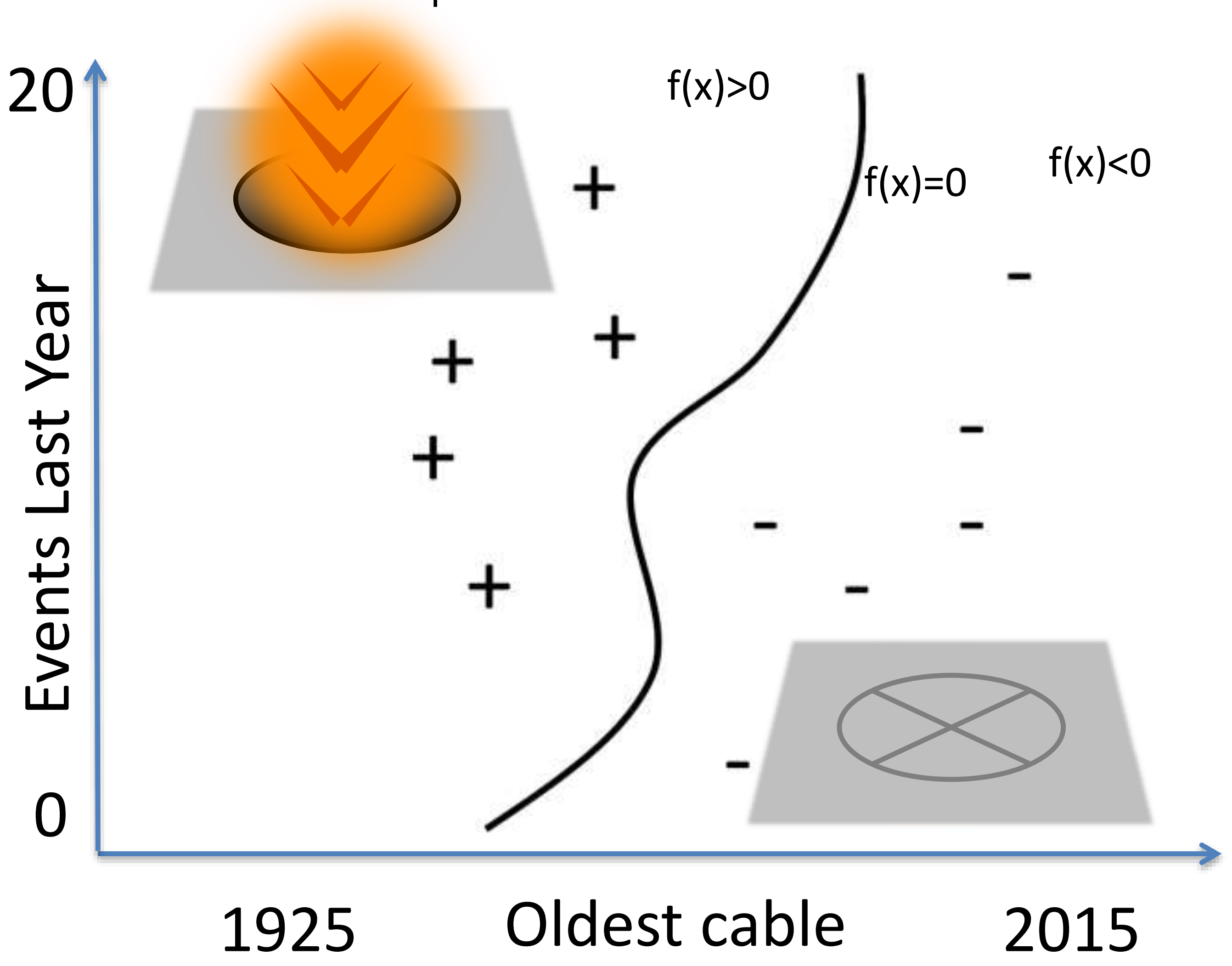
Year oldest cable installed
Number of events last year



Classification

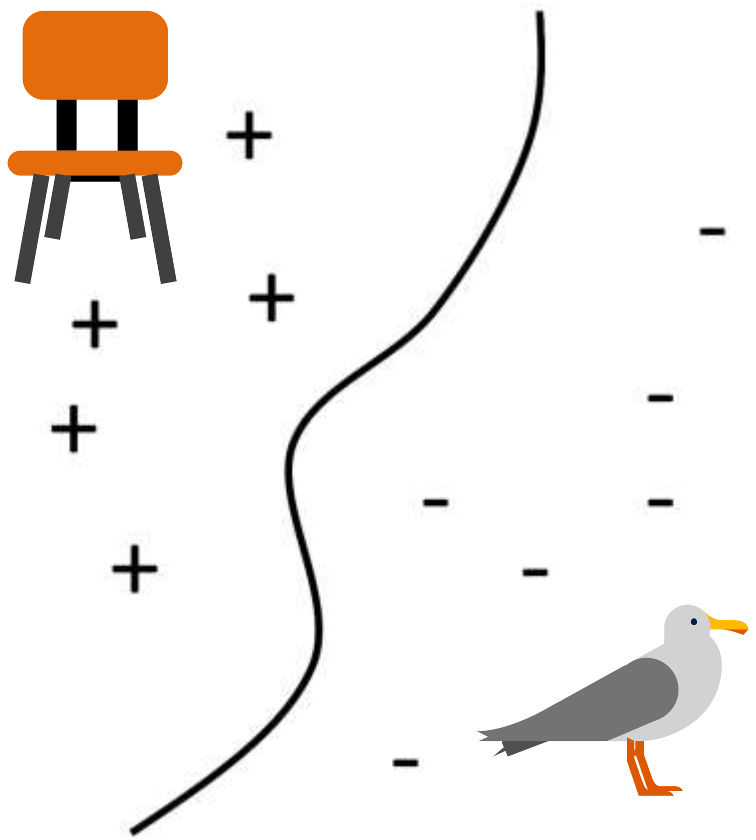
- Formally, given training set (x_i, y_i) for i=1…n, we want to create a classification model f that can predict label y for a new x.

f(x) = function(Events Last Year, Oldest Cable)



Classification

- Formally, given training set $(x_{i,}y_i)$ for i=1…n, we want to create a classification model f that can predict label y for a new x.



Classification

- Formally, given training set $(x_{i,} y_i)$ for $i=1\ldots n$, we want to create a classification model f that can predict label y for a new x.

- The machine learning algorithm will create the function f for you. It might be very complicated, but the way to use is is not complicated:

The predicted value of y for a new x is the sign of f(x).

Classification

- Yes/No questions is the most basic
- automatic handwriting recognition, speech recognition, biometrics, document classification, spam detection, predicting credit default risk, detecting credit card fraud, predicting customer churn, predicting medical outcomes (strokes, side effects, etc.)

Regression

Cynthia Rudin | MIT Sloan School of Management

Regression

- For predicting real-valued outcomes:
  - How many customers will arrive at our website next week?
  - How many tv's will we sell next year?
  - Can we predict someone's income from their click through information?

Regression

- Each observation is represented by a set of numbers.

										Income
A person is represented as:	[5	3	120	12	1	0	.....	]		84
	[0	0	89	5	1	1	.....	]		32
	[1	0	20	0	0	1	.....	]		-10
:			:							

Features, called X

Labels, called Y

Regression

- Each observation is represented by a set of numbers.

	Income
A person is represented as: [5]	84
[0]	32
[1]	-10
:	

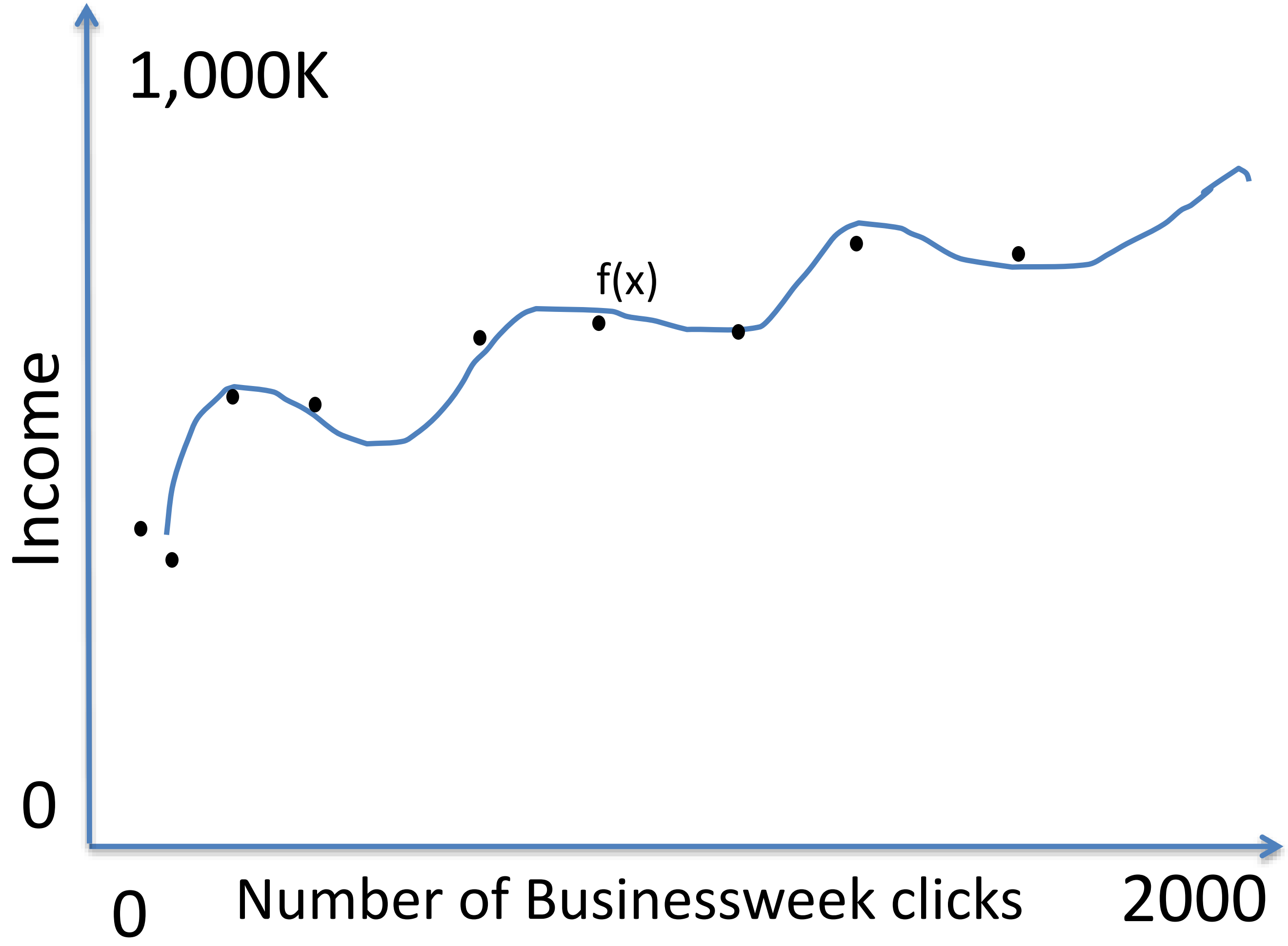
Single feature, called X

Labels, called Y

Regression

- Formally, given training set (x_i, y_i) for i=1…n, we want to create a regression model f that can predict label y for a new x.

f(x) = function(Number of Businessweek clicks)

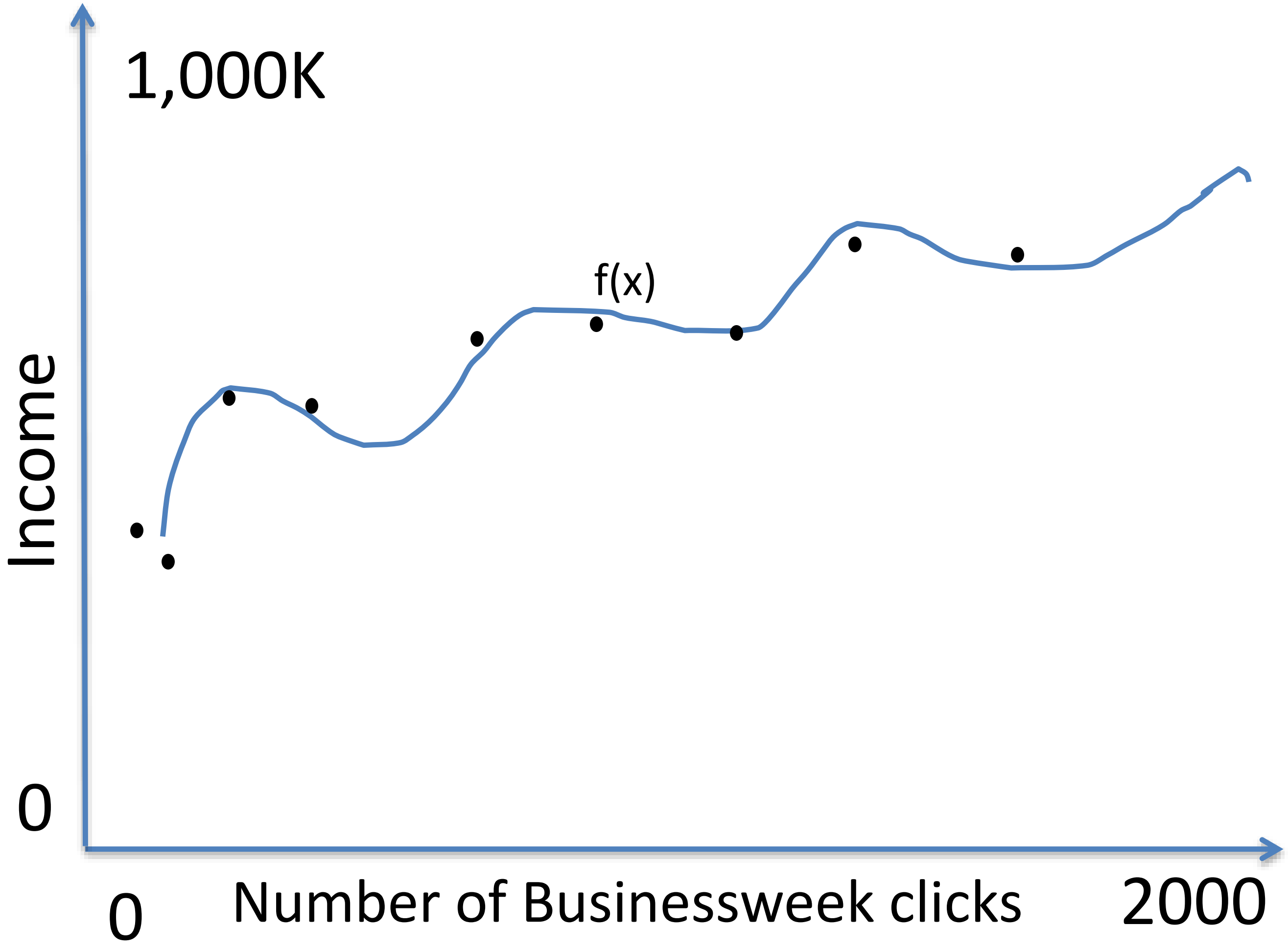


Regression

- Formally, given training set (x_i, y_i) for i=1…n, we want to create a regression model f that can predict label y for a new x.

f(x) = function(Number of Businessweek clicks)

(Overfitting?)

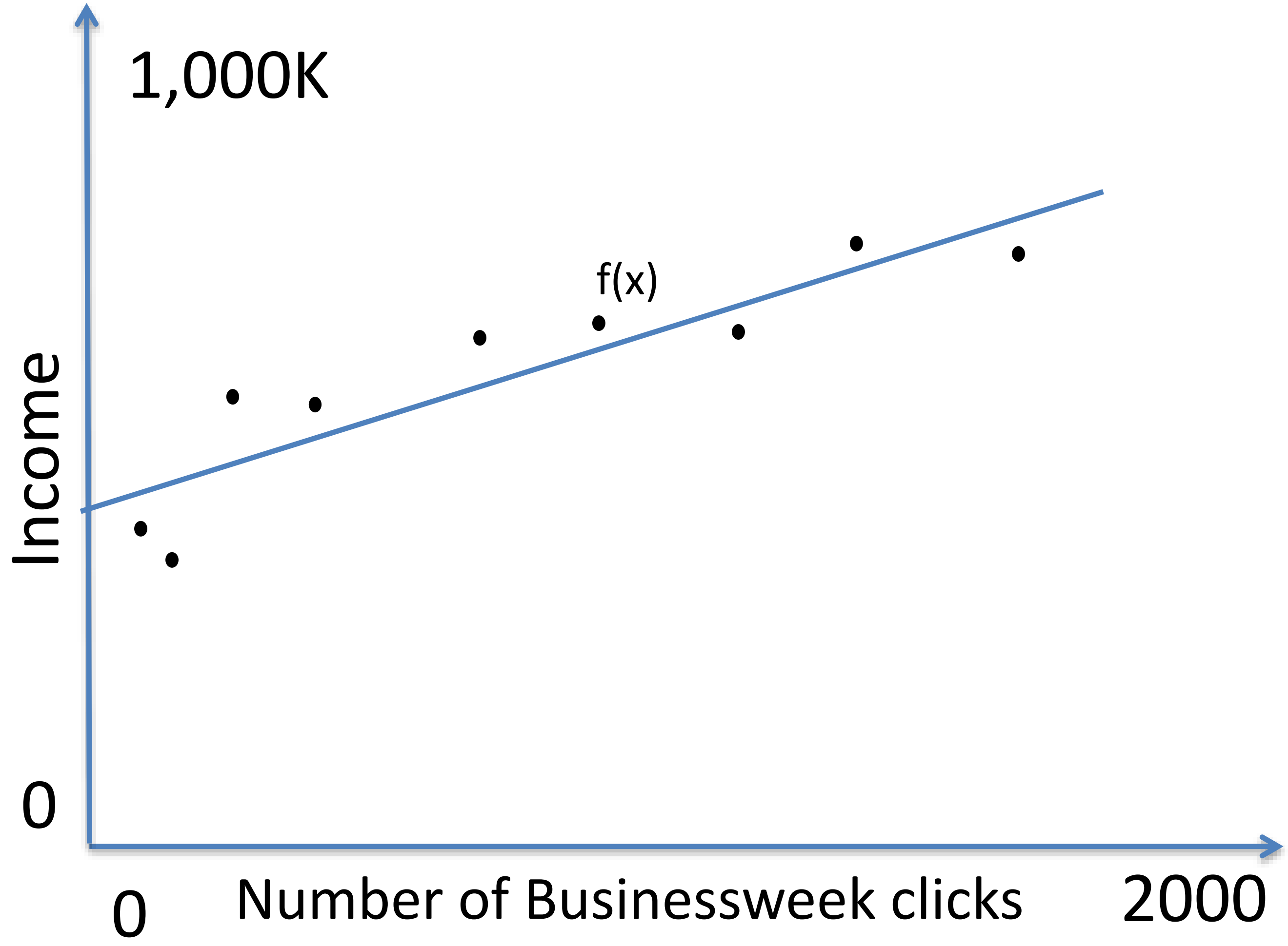


Regression

- Formally, given training set (x_i, y_i) for i=1…n, we want to create a regression model f that can predict label y for a new x.

f(x) = function(Number of Businessweek clicks)
= 5K*Number of Businessweek clicks + 100K

(Underfitting?)



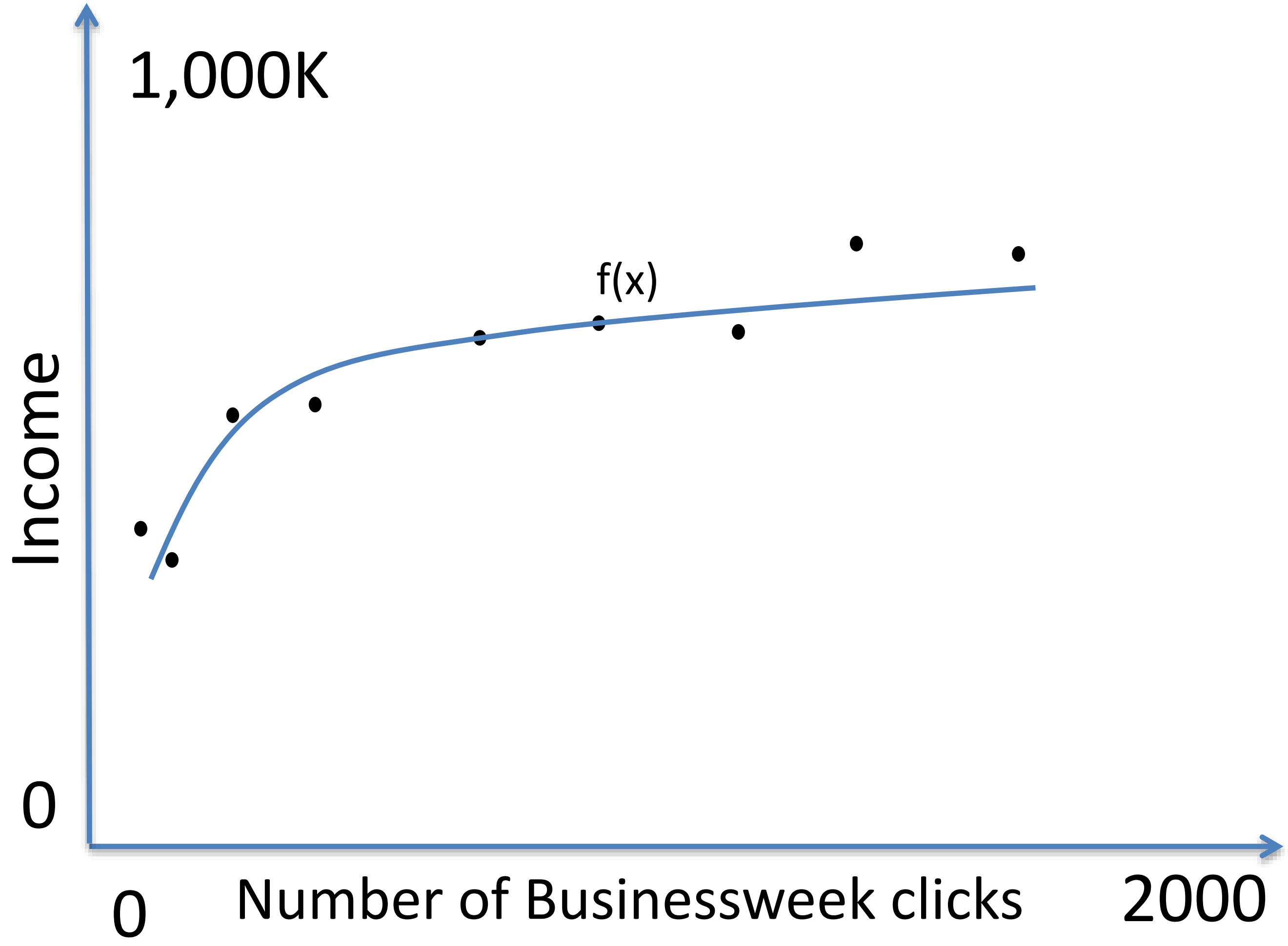
Regression

- Formally, given training set (x_i, y_i) for i=1…n, we want to create a regression model f that can predict label y for a new x.

f(x) = function(Number of Businessweek clicks)

(Just right?)

We'll talk more about this later



Regression

- Formally, given training set (x_i, y_i) for i=1…n, we want to create a regression model f that can predict label y for a new x.

Estimated income:

f(x) = function(Number of visits to upscale furniture websites, Number of Businessweek clicks, Number of distinct people emailed per day, Number of purchases of over 5K within the last month, Number of visits to airlines, etc.)

For instance,
f(x) = 3*Number of visits to upscale furniture websites
 +10*Number of Businessweek clicks
 +100*Number of distinct people emailed per day
 +2*Number of purchases of over 5K within the last month
 +10*Number of visits to airlines

But f(x) could be much more complicated

Supervised Learning

- Classification and Regression are supervised learning problems.
- “Supervised” means that the training data has ground truth labels to learn from.
- (Supervised) classification often has +1 or -1 labels.
- (Supervised) regression has numerical labels.
- There are lots of other supervised problems.
- Supervised learning algorithms are much easier to evaluate than unsupervised ones.

Statistical Learning Theory for Supervised Learning

Cynthia Rudin | MIT Sloan School of Management

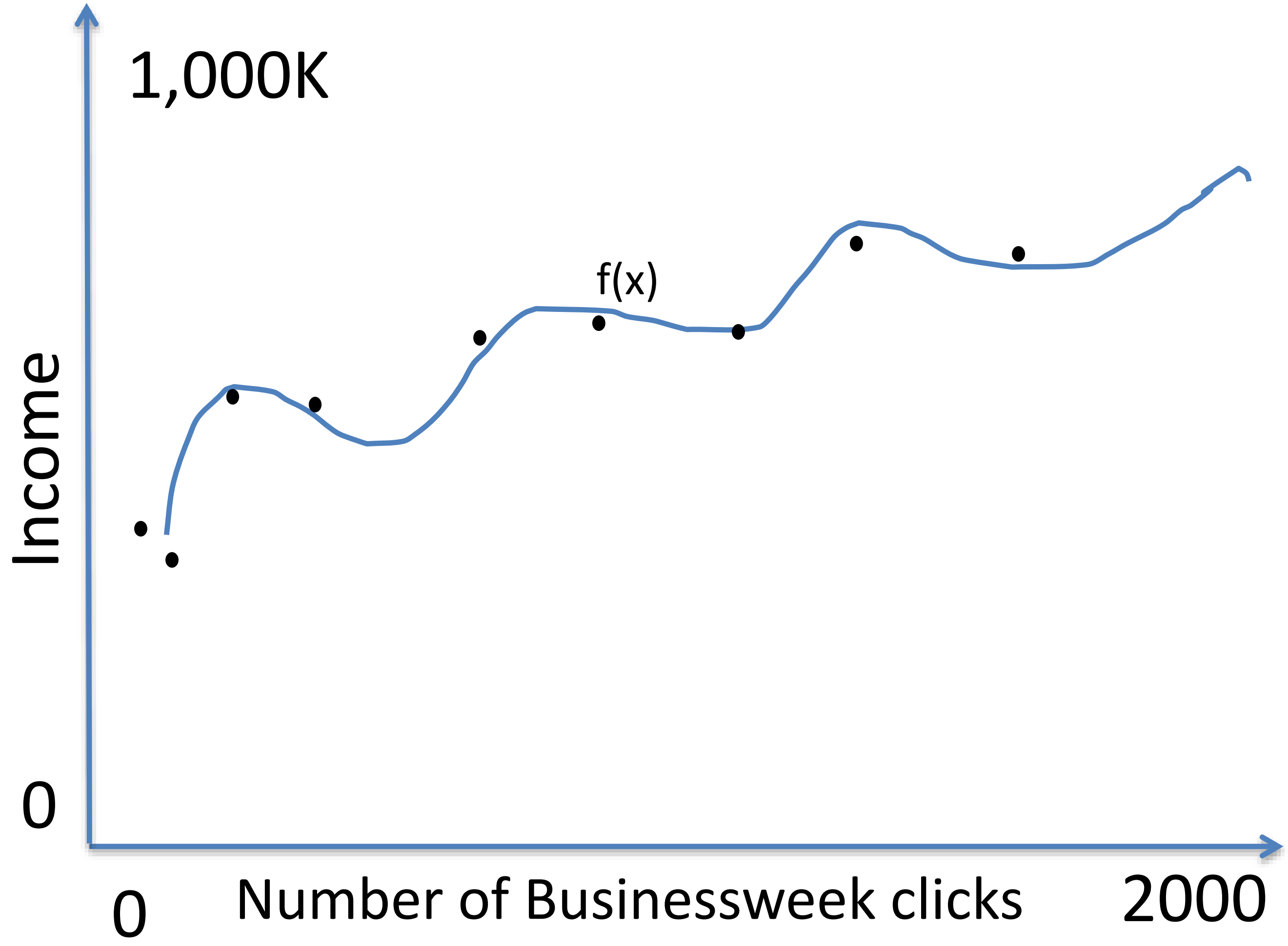
Microsoft

Statistical Learning Theory

- Occam’s Razor: The best models are simple models that fit the data well.
- William of Ockham, English frier and philosopher (1287-1347) said that among hypotheses that predict equally well, we should choose the one with the fewest assumptions.

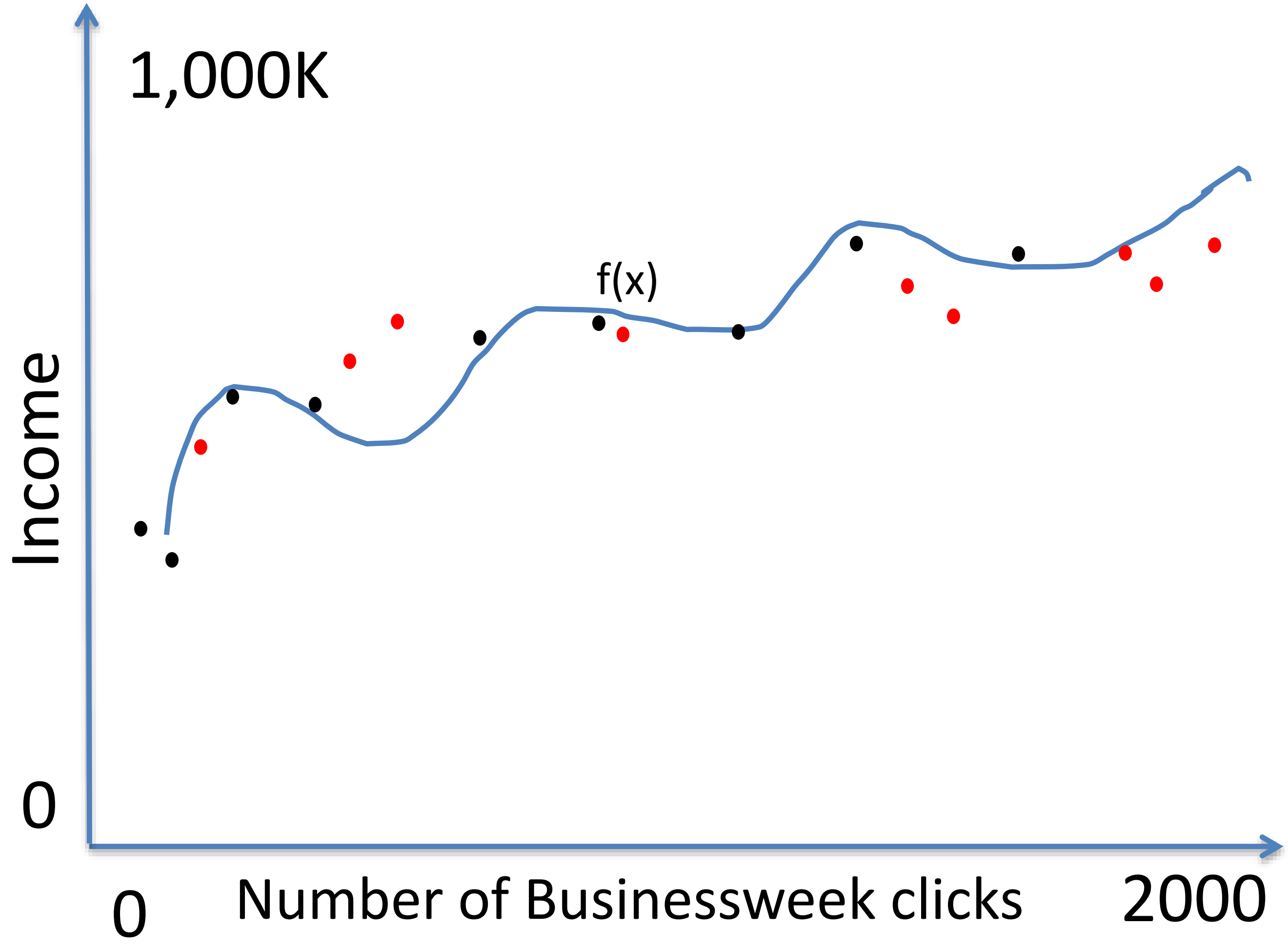
Statistical Learning Theory

- Occam's Razor: The best models are simple models that fit the data well.



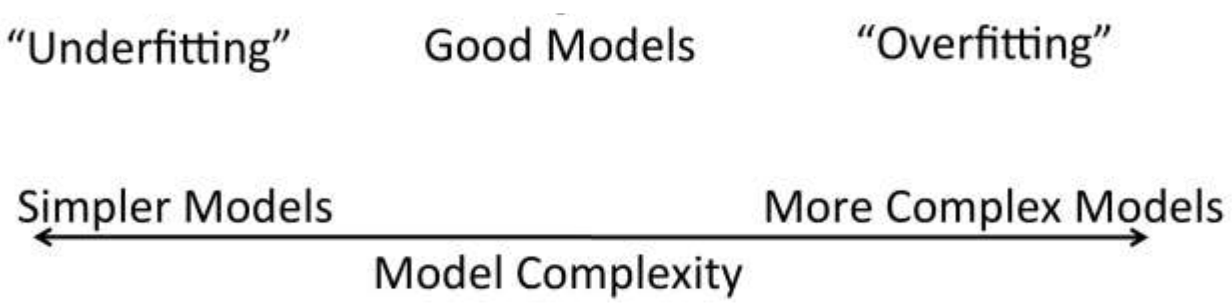
Statistical Learning Theory

- Occam's Razor: The best models are simple models that fit the data well.



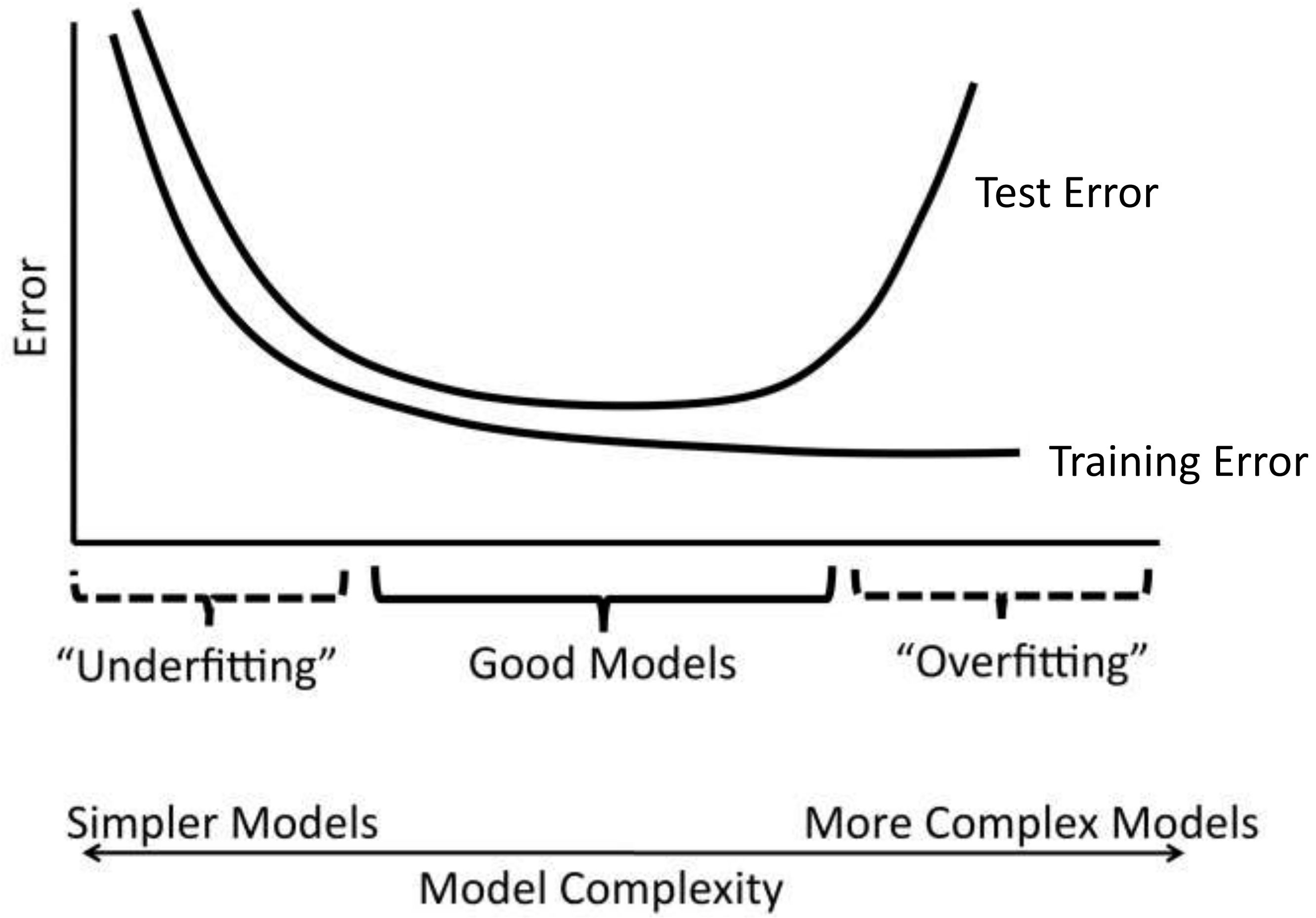
Statistical Learning Theory

- Occam's Razor: The best models are simple models that fit the data well.



Statistical Learning Theory

- Occam's Razor: The best models are simple models that fit the data well.



Statistical Learning Theory

- Occam’s Razor: The best models are simple models that fit the data well.
- We need a balance between accuracy and simplicity.

Statistical Learning Theory

- Occam's Razor: The best models are simple models that fit the data well.
- We need a balance between accuracy and simplicity.
- Most common machine learning methods choose f to minimize training error and complexity.
- Aims to thwart the "curse" of dimensionality.

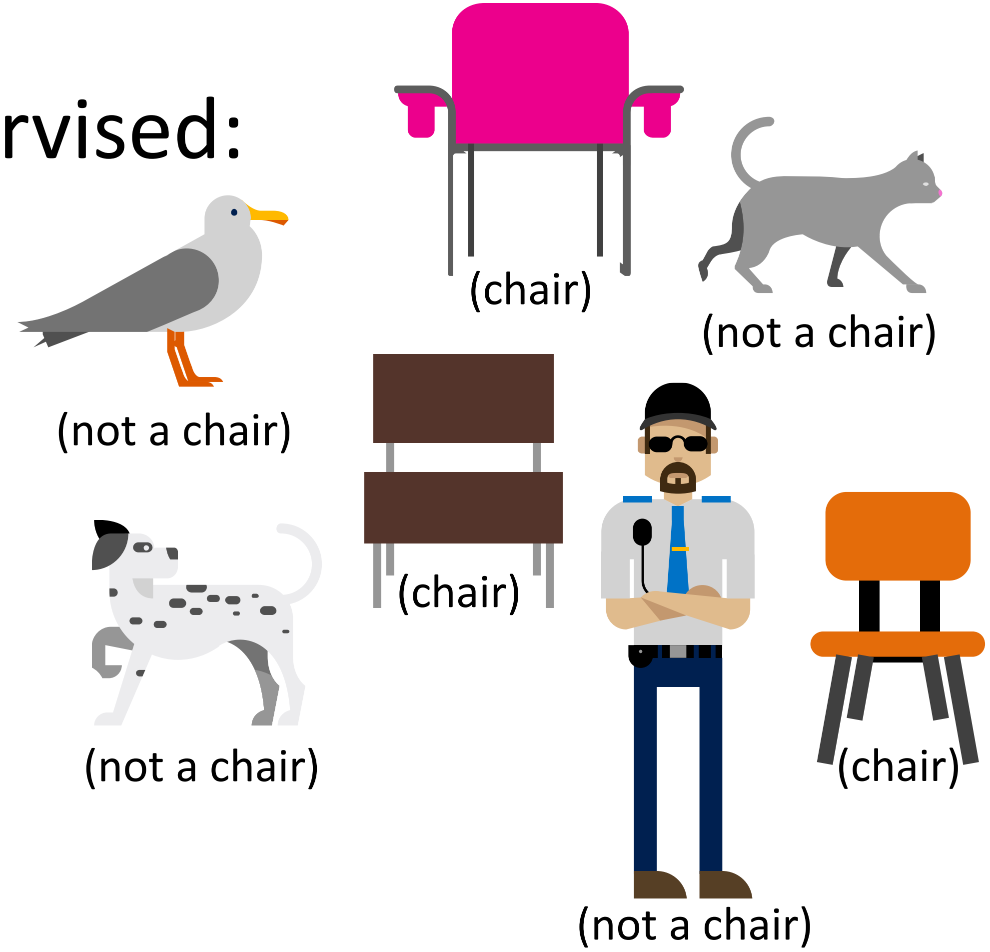
Clustering

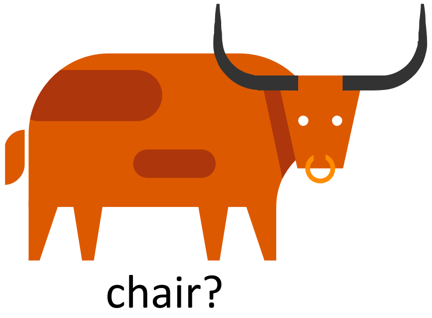
Cynthia Rudin | MIT Sloan School of Management

Unsupervised Learning

- "Unsupervised" means that the training data has no ground truth labels to learn from.

Supervised:

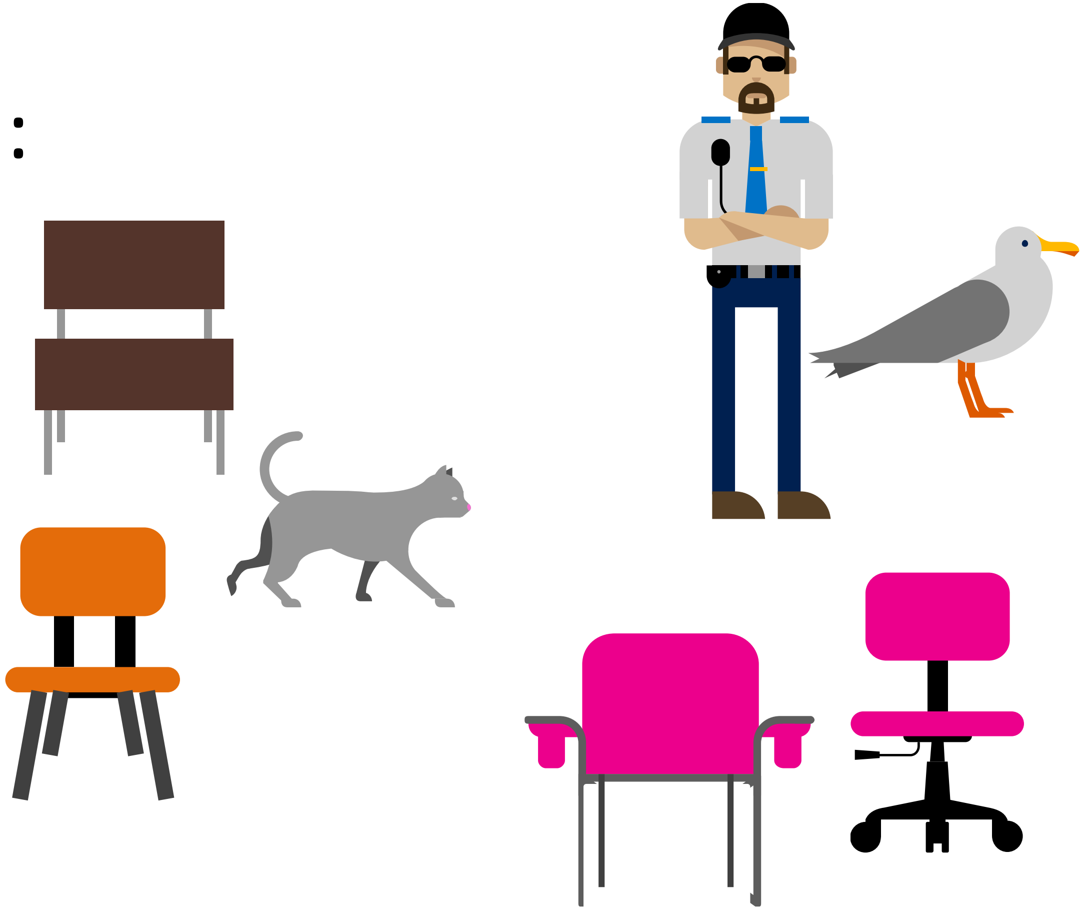




Unsupervised Learning

- "Unsupervised" means that the training data has no ground truth labels to learn from.

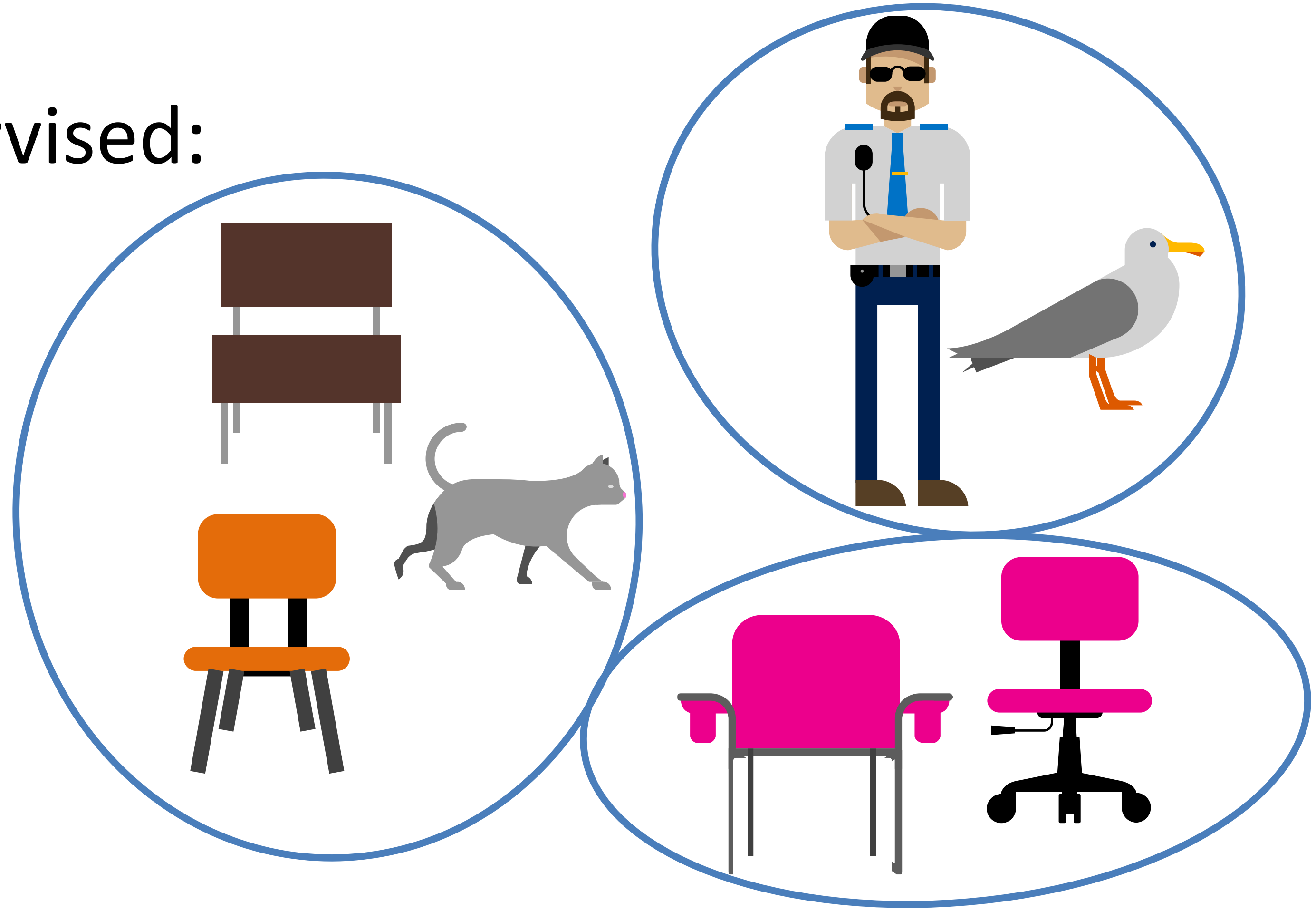
Unsupervised:



Unsupervised Learning

- "Unsupervised" means that the training data has no ground truth labels to learn from.

Unsupervised:



Unsupervised Learning

- "Unsupervised" means that the training data has no ground truth labels to learn from.
- This means they are much harder to evaluate.
- Clustering is an key unsupervised problem.

Recommender Systems and Matrix Factorization

Microsoft

Cynthia Rudin | MIT Sloan School of Management

Recommender Systems and Matrix Factorization

- The Netflix contest: Build a better recommender system from Netflix data

$$
\begin{array}{r c}
 & \begin{array}{cccc} A & B & C & D \end{array} \\
\begin{array}{r} \text{Carmen} \\ \text{Joseph} \\ \text{Leonore} \\ \text{Esmerelda} \end{array} &
\begin{pmatrix} 5 & - & 4 & 1 \\ 5 & 4 & - & - \\ 1 & 7 & - & 3 \\ 2 & 8 & 1 & - \end{pmatrix}
\end{array}
$$

- Use the crowd's votes to complete the missing entries

Coming up

Microsoft

Cynthia Rudin | MIT Sloan School of Management

Coming Up

- How to formulate and solve machine learning problems
- How to evaluate machine learning algorithms

Microsoft

©2014 Microsoft Corporation. All rights reserved. Microsoft, Windows, Office, Azure, System Center, Dynamics and other product names are or may be registered trademarks and/or trademarks in the U.S. and/or other countries. The information herein is for informational purposes only and represents the current view of Microsoft Corporation as of the date of this presentation. Because Microsoft must respond to changing market conditions, it should not be interpreted to be a commitment on the part of Microsoft, and Microsoft cannot guarantee the accuracy of any information provided after the date of this presentation. MICROSOFT MAKES NO WARRANTIES, EXPRESS, IMPLIED OR STATUTORY, AS TO THE INFORMATION IN THIS PRESENTATION.