Manipulation and Machine Learning: Ethics in Data Science

DEF CON 23 Crypto & Privacy Village

Jennifer Helsby, Ph.D.
University of Chicago
@redshiftzero
jen@redshiftzero.com

GPG: 1308 98DB C324 62D4 1C7D 298E BCDF 35DB 90CC 0310

Background

- Recently: Ph.D. in astrophysics
  - Cosmologist specializing in large-scale data analysis
  - Dissertation was on statistical properties of millions of galaxies in the universe

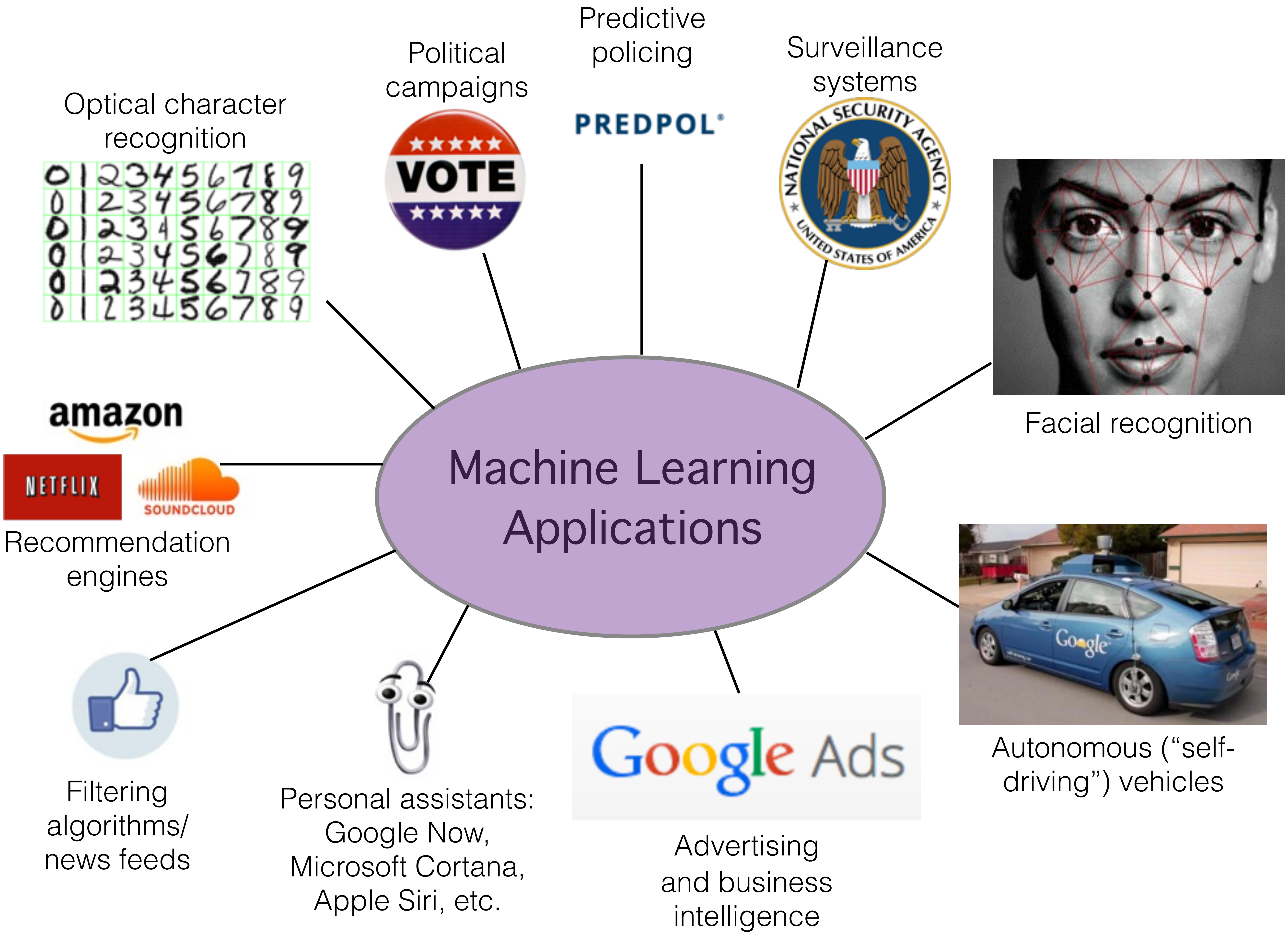
DARK ENERGY SURVEY

- Currently: Data Science for Social Good fellow at the University of Chicago
  - Machine learning/data science application to projects with positive social impact in education, public health, and international development

THE UNIVERSITY OF CHICAGO

Data Science for Social Good

My opinions are my own, not my employers



Machine Learning Applications
Optical character recognition
0 1 2 3 4 5 6 7 8 9
Political campaigns
VOTE
Predictive policing
PREDPOL
Surveillance systems
NATIONAL SECURITY AGENCY
UNITED STATES OF AMERICA
Facial recognition
amazon
NETFLIX
SOUNDCLOUD
Recommendation engines
Autonomous ("self-driving") vehicles
Google
Filtering algorithms/ news feeds
Personal assistants: Google Now, Microsoft Cortana, Apple Siri, etc.
Google Ads
Advertising and business intelligence

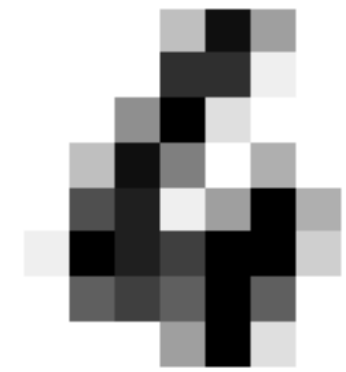
Machine Learning?

- *Machine learning* is a set of techniques for adaptive computer programming
    - learn programs from data

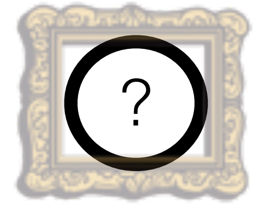
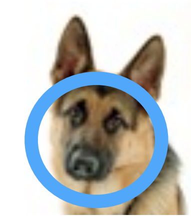
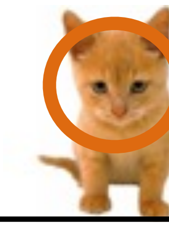
Machine Learning?

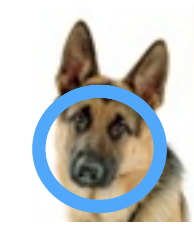
- *Machine learning* is a set of techniques for adaptive computer programming

    - learn programs from data



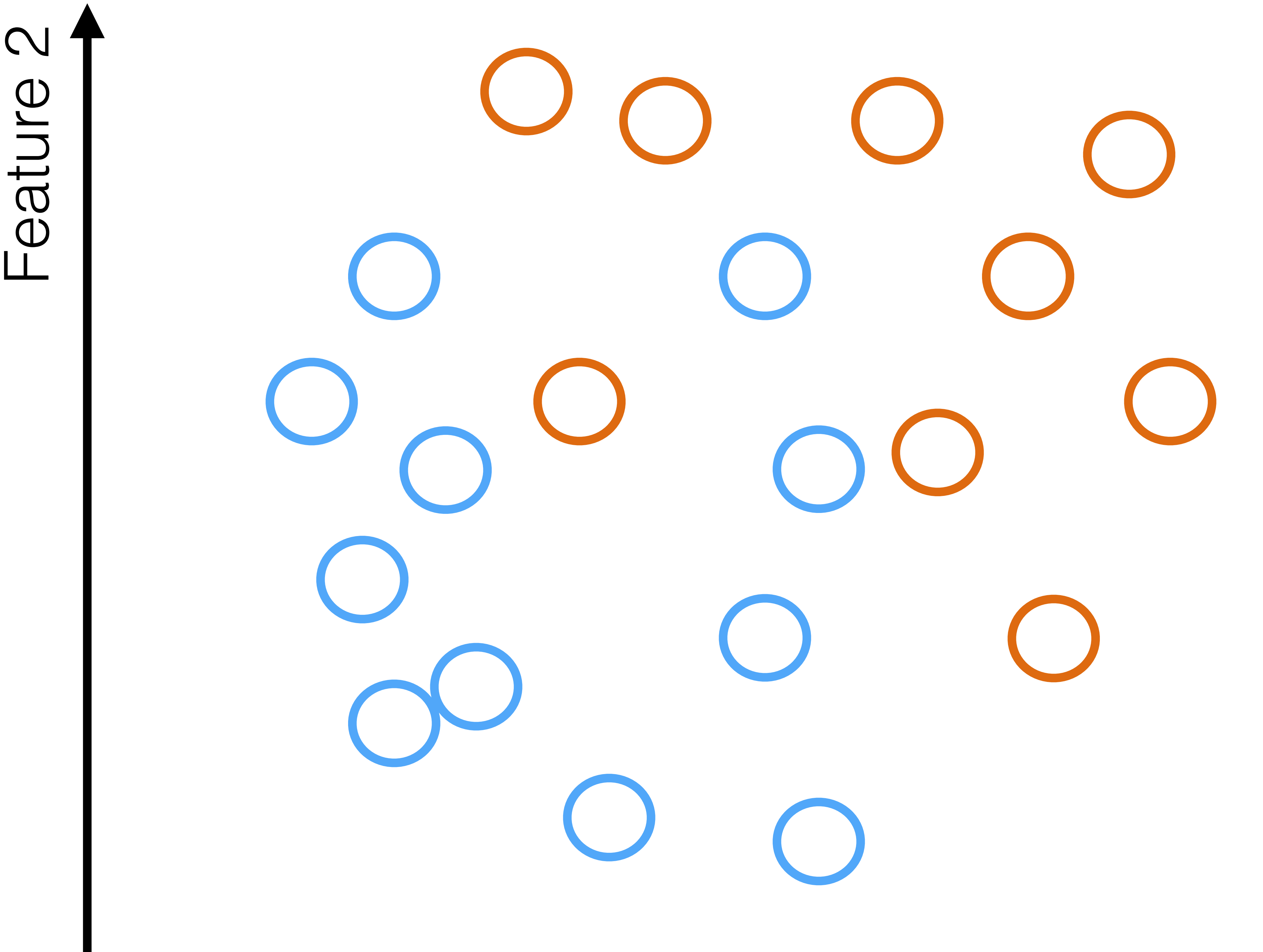
- In *supervised learning*, a computer learns some rules by *example* without being explicitly programmed

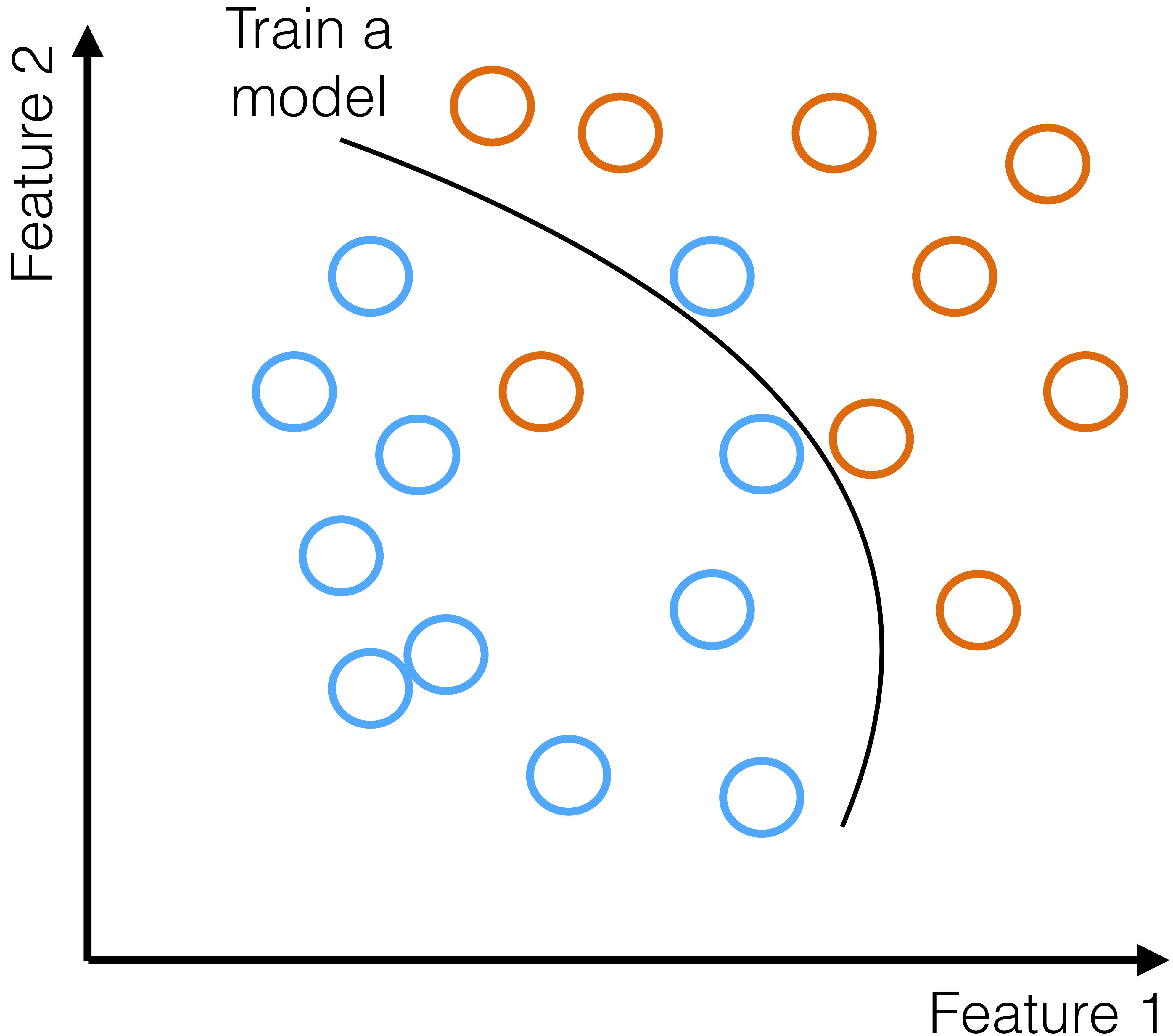
Classification problem: Classify  into  or 

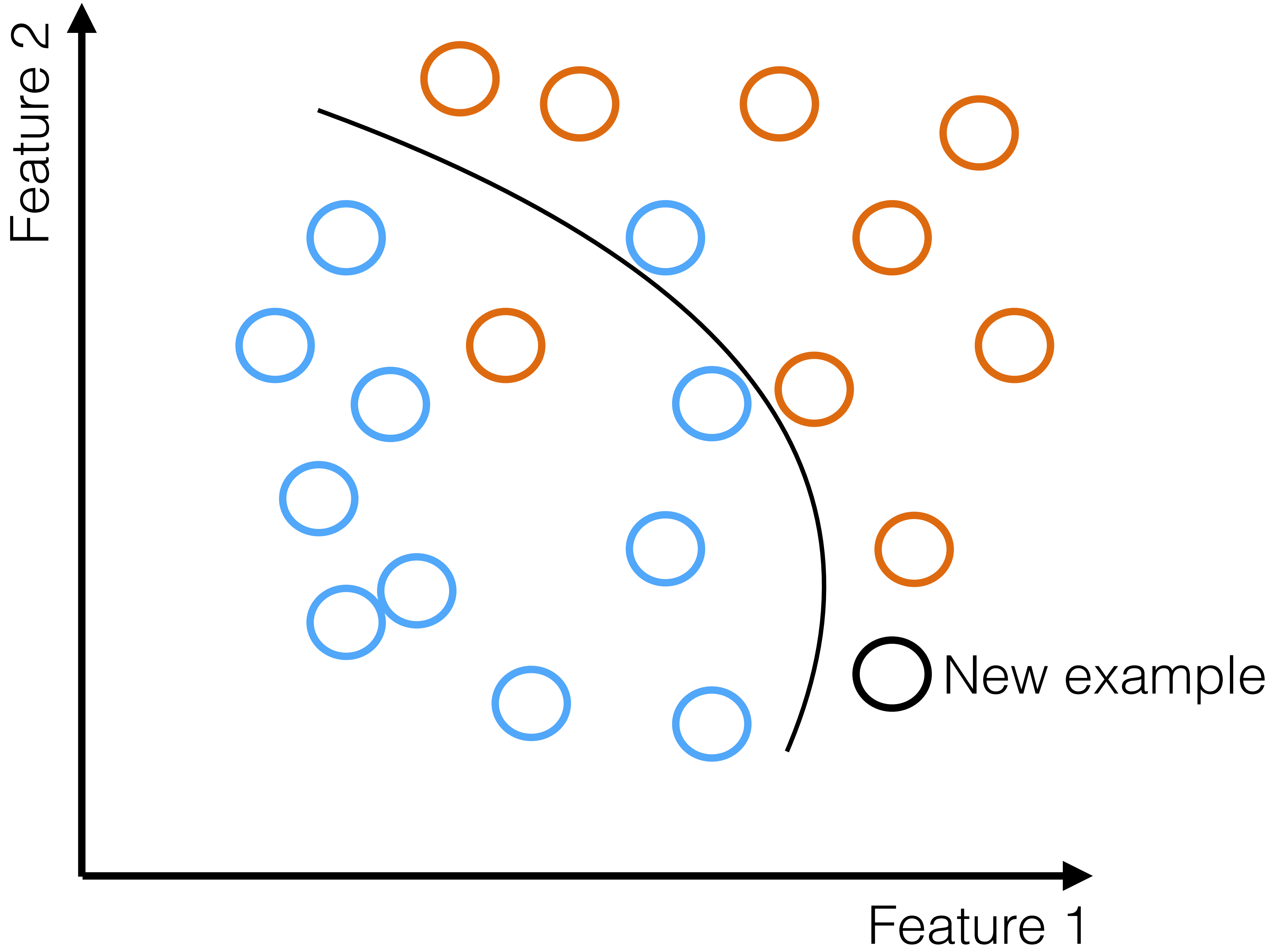
Get **examples** of past and whether they were  or 

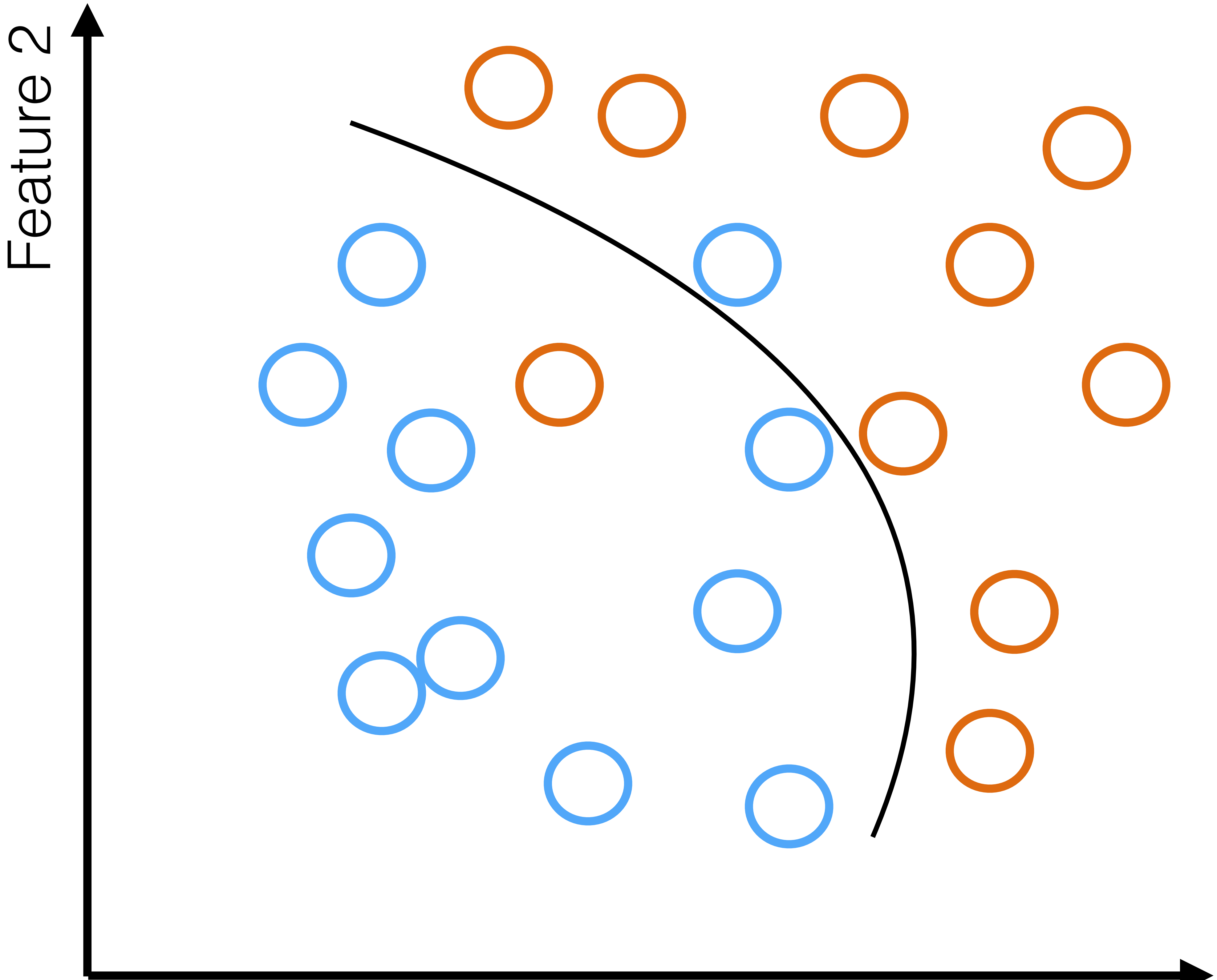
Build **features**, quantities that might be predictive of the target (cat/dog)

Use **examples** and **features** to **train** a model


Feature 2
Feature 1


Train a
model
Feature 2
Feature 1


Feature 2
New example
Feature 1


Feature 2
Feature 1

What’s the big deal?

Pitfalls

Methodological issues

Usage issues

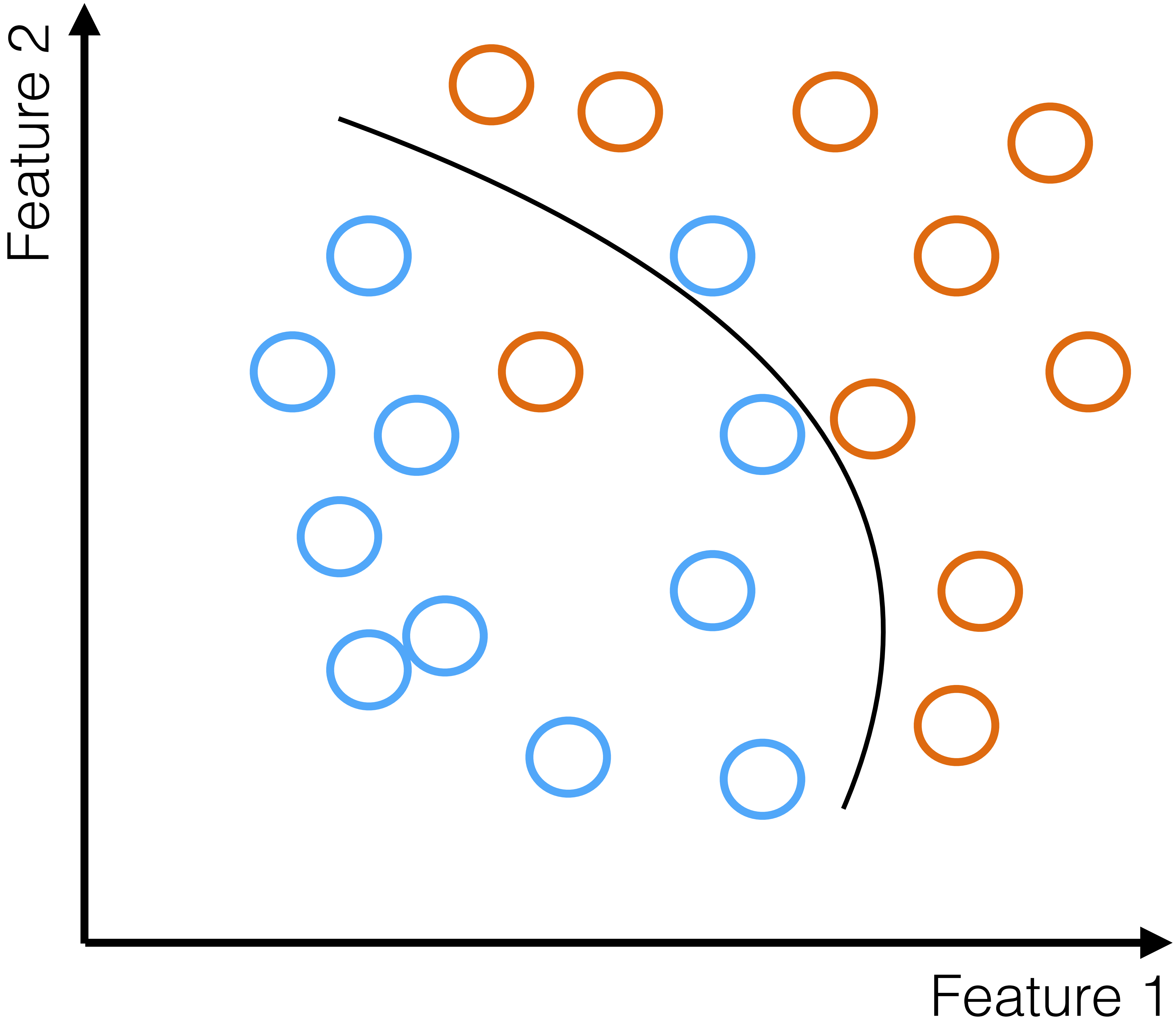
Pitfalls

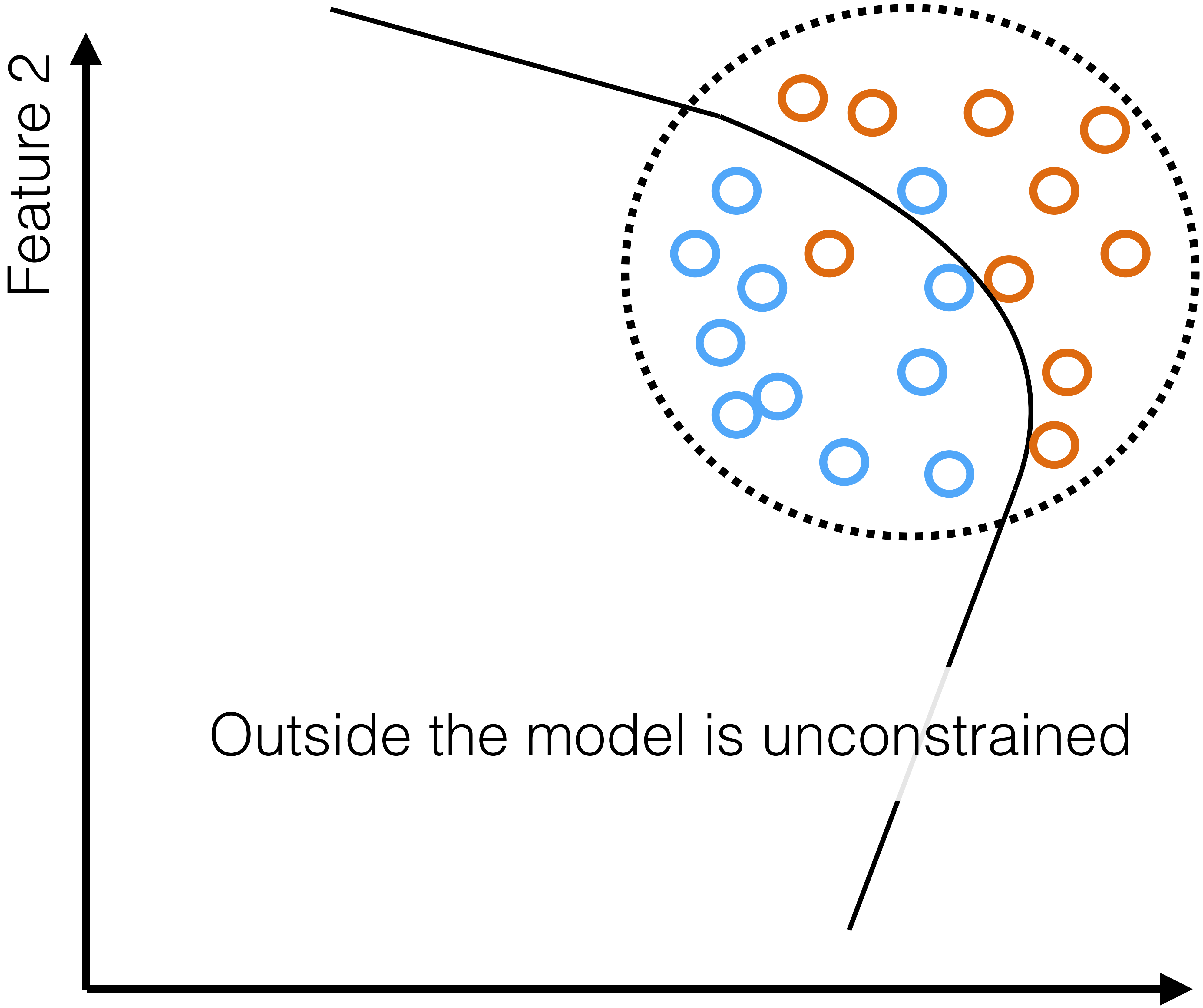
Methodological issues

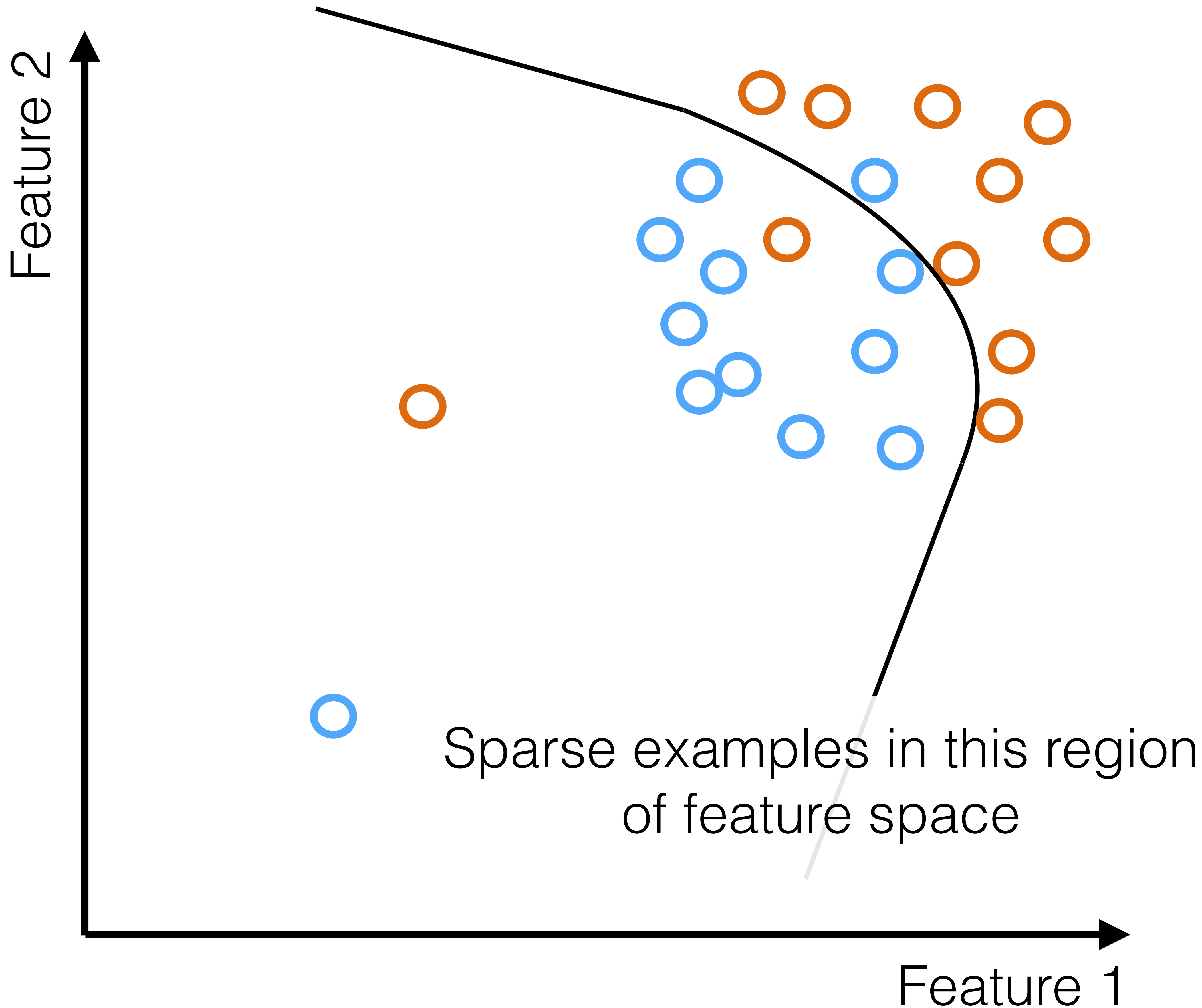
Usage issues

Representativeness

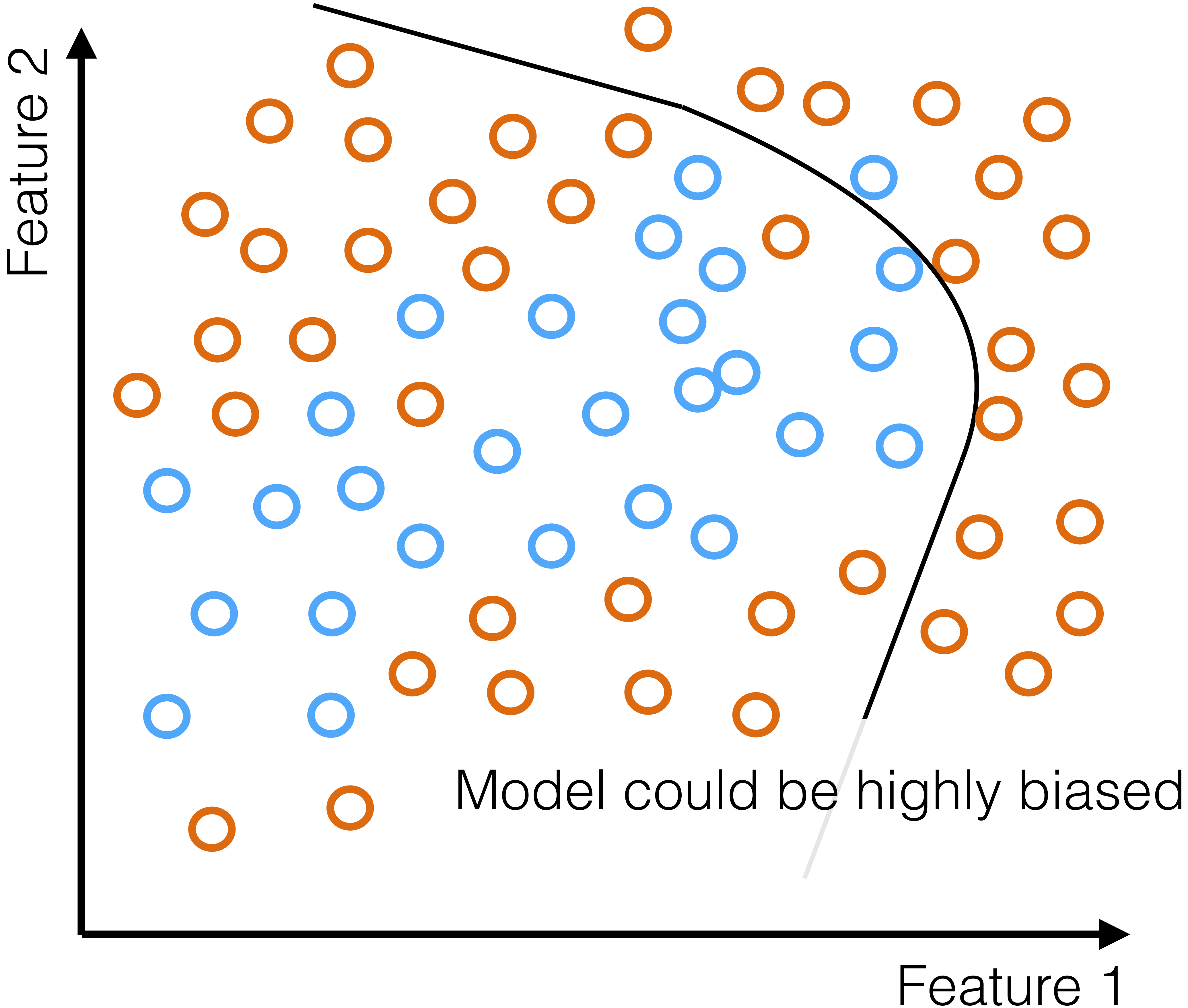
- Learning by *example:* Examples must be representative of truth
- If they are not → Model will be biased
- <u>Random sampling:</u> Probability of collecting an example is uniform
- Most sampling is *not* random
- Strong selection effects present in most training data

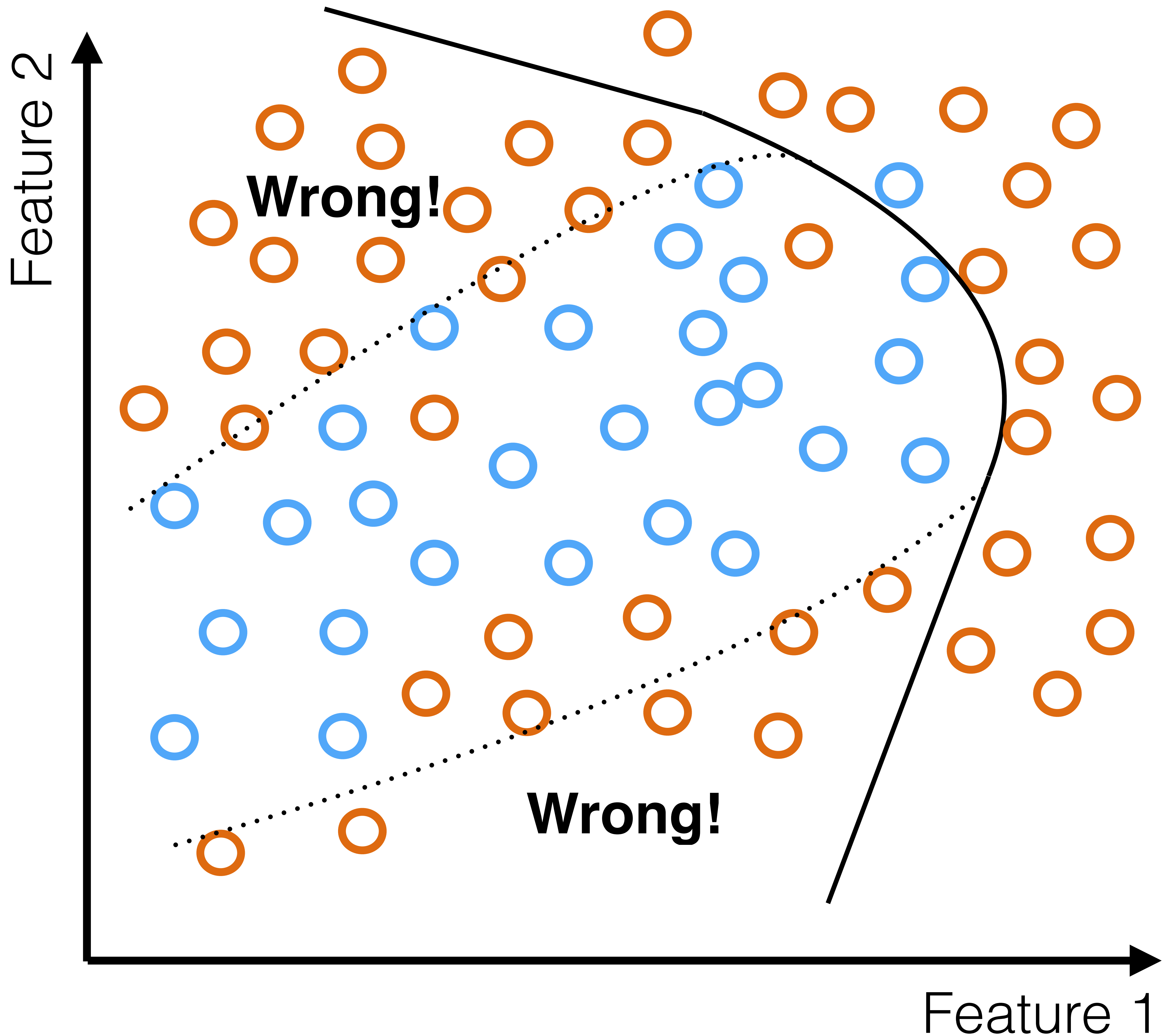

Feature 2
Feature 1


Feature 2
Outside the model is unconstrained
Feature 1

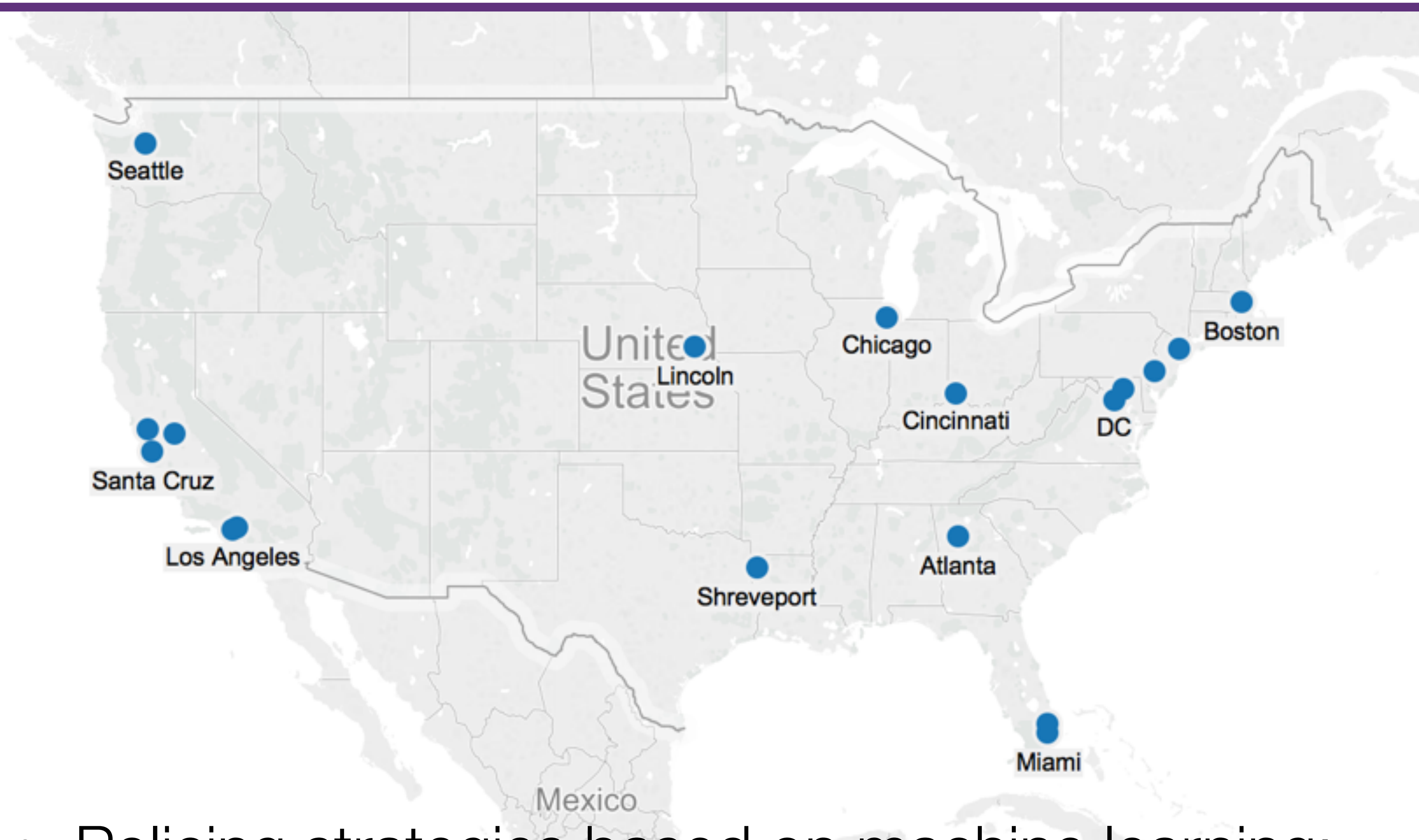


Feature 2
Sparse examples in this region
of feature space
Feature 1


Feature 2
Model could be highly biased
Feature 1


Feature 2
Wrong!
Wrong!
Feature 1

Predictive Policing



- Policing strategies based on machine learning: *proactive, preventative* or *preventative* policing

- Aim: To allocate resources more effectively

“The ‘Minority Report’ of 2002 is the reality of today”

- New York City Police Commissioner William Bratton

PREDPOL®

ABOUT HOW PREDPOL WORKS PROVEN RESULTS FOR ANALYSTS TECHNOLOGY PRESS CONTACT US BLOG

PREDICTIVE POLICING®

The Predictive Policing Company.

PredPol's cloud-based software enables law enforcement agencies to better prevent crime in their communities by generating predictions on the places and times that future crimes are most likely to occur.

Dozens of communities across the US and overseas are experiencing dramatic reductions in crime thanks in large part to PredPol software technology.

Only three pieces of data are used to make predictions – type of crime, place of crime, and time of crime. No personal data is utilized in making these predictions.

Only three pieces of data are used to make predictions – type of crime, place of crime, and time of crime.

No personal data is utilized in making these predictions.

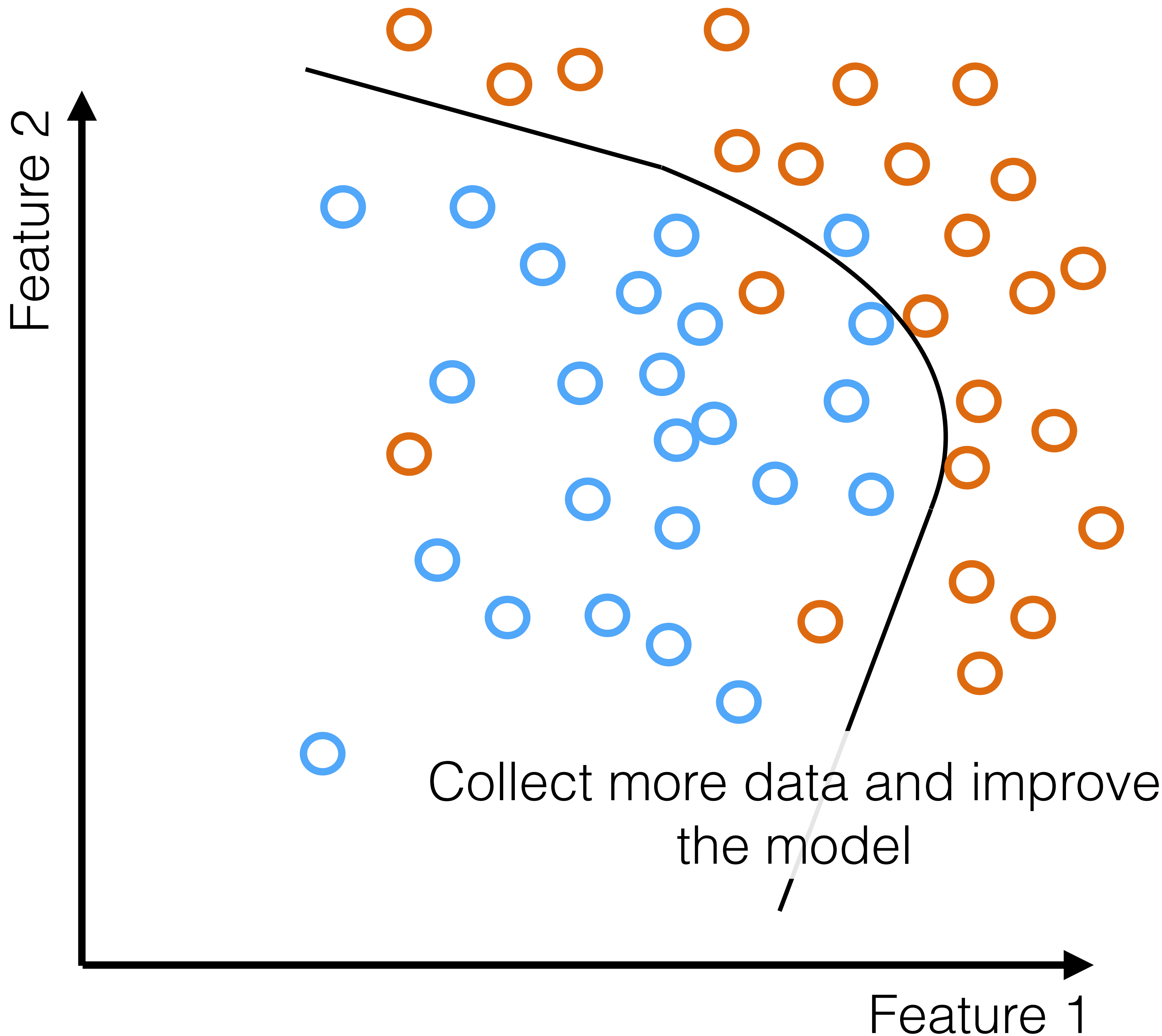
Only three pieces of data are used to make predictions – type of crime, place of crime, and time of crime.

No personal data is utilized in making these predictions.

eliminating any personal liberties and profiling concerns.

Racist Algorithms are Still Racist

- Inherent biases in input data:
    - For crimes that occur at similar rates in a population, the sampling rate (by police) is not uniform
- More responsible: Reduce impact of biased input data by exploring poorly sampled regions of feature space


Feature 2
Feature 1
Collect more data and improve the model

Pitfalls

Methodological issues:

- Selection effects in input datasets used for training
- Aggregation also provides information to a model about individuals
- Removing controversial features does not remove all discriminatory issues with the training data

Pitfalls

Methodological issues

Usage issues

Pitfalls

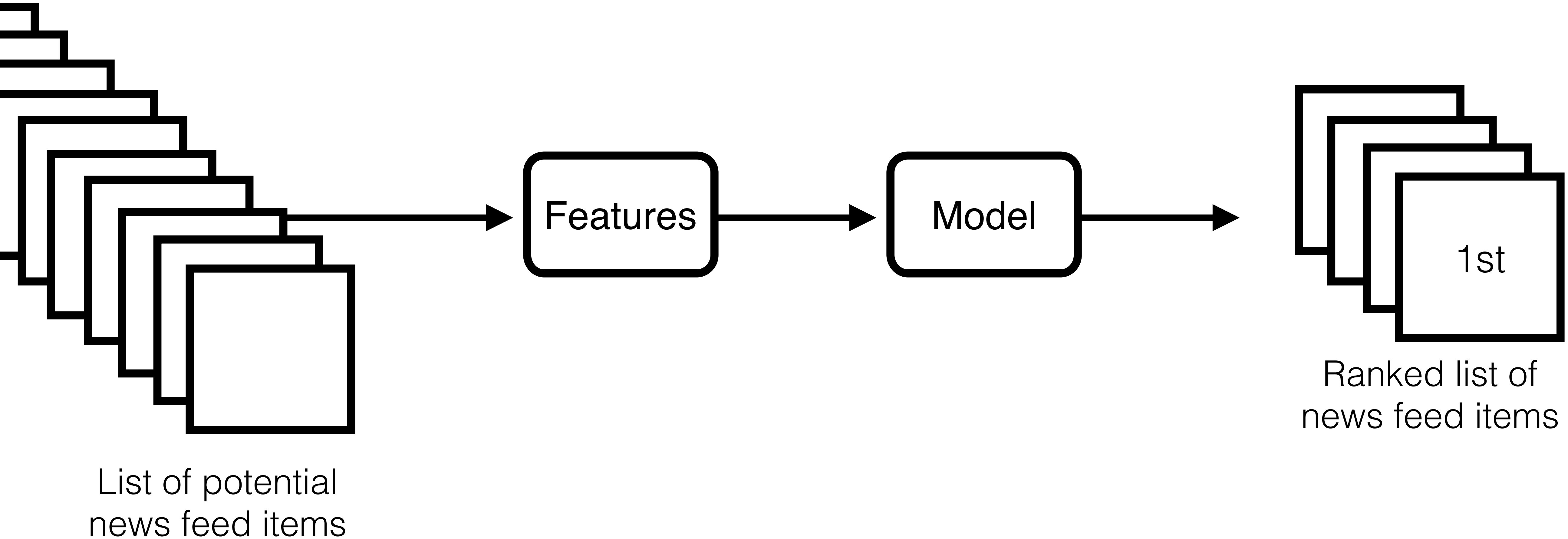
Methodological issues

Usage issues

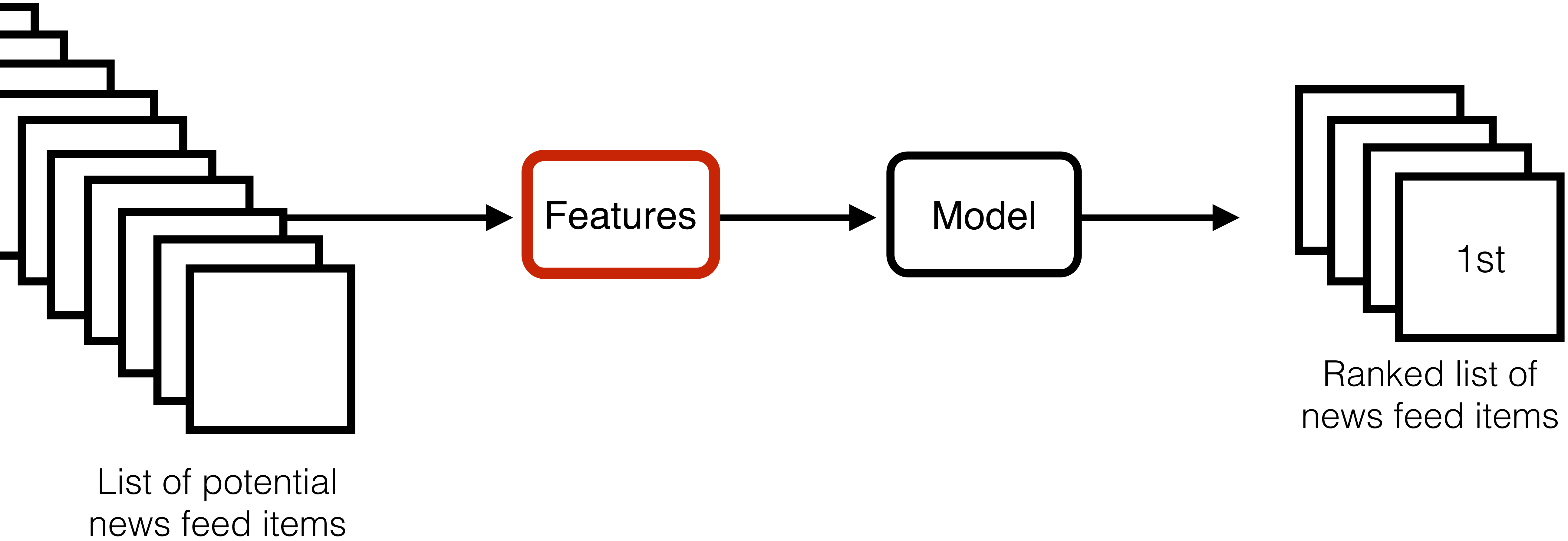
Filtering

- An avalanche of data necessitates filtering
- Many approaches:
  - Reverse chronological order (i.e., newest first)
  - Collaborative filtering: People vote on what is important
  - Select what you should see based on an algorithm

Facebook News Feed



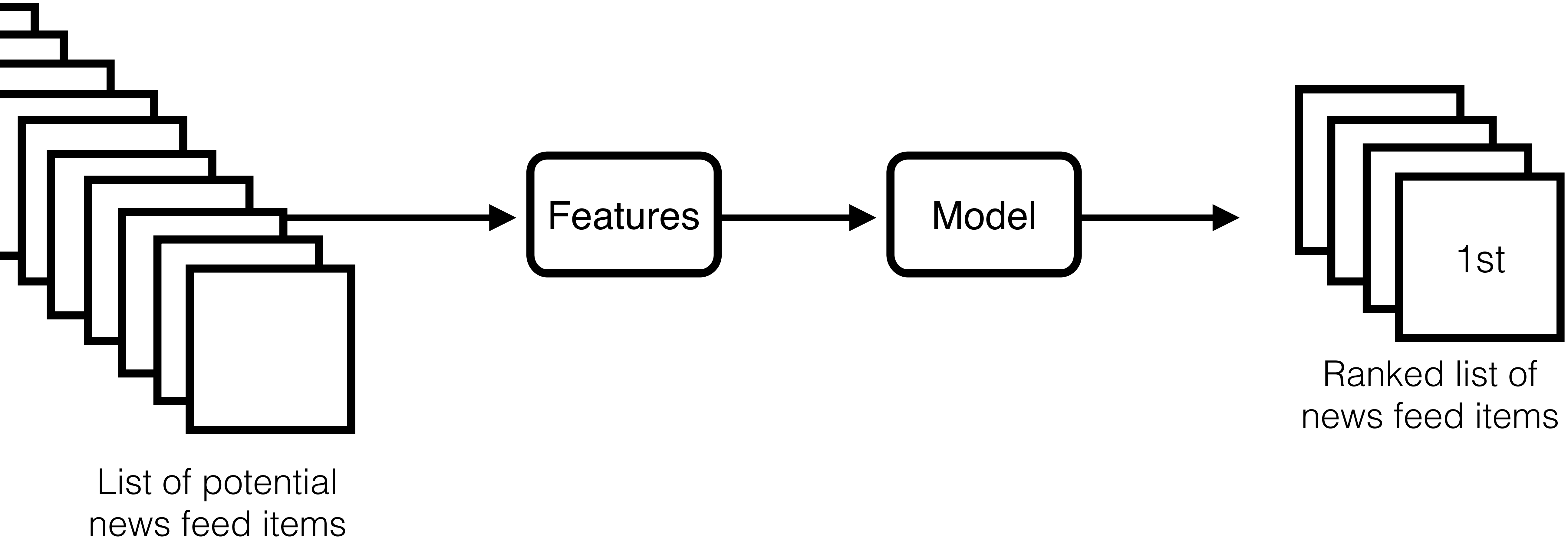
Facebook News Feed



Feature Building

- Is a **trending topic** mentioned?
- Is this an important life event? e.g. Are words like **congratulations** mentioned?
- How **old** is this news item?
- How many **likes/comments** does this item have? Likes/comments by **people I know**?
- Are the words “Like”, “Share”, “Comment” present?
- *Is **offensive content** present?*

Facebook News Feed

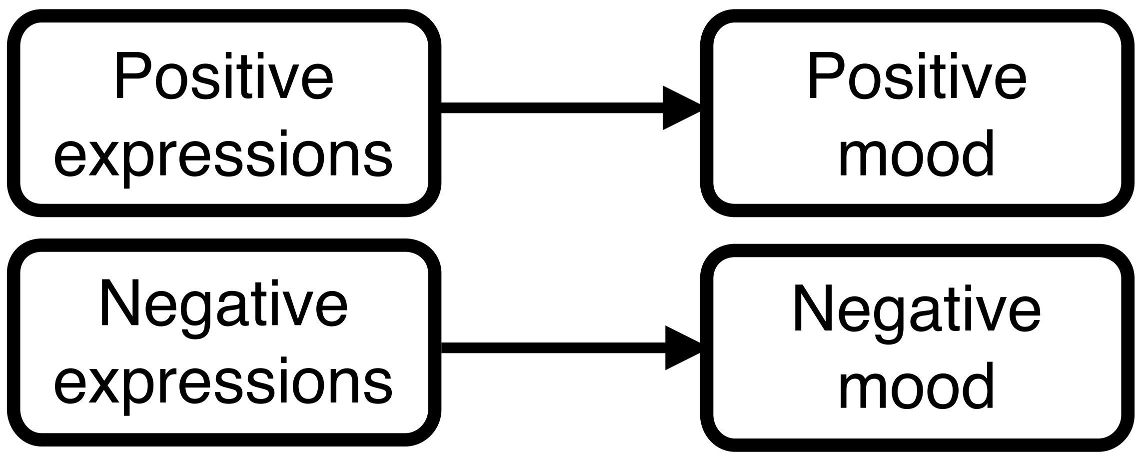


- Facebook decides what updates and news stories you get to see
- 30% of people get their news from Facebook [Pew Research]

Emotional Manipulation

Experimental evidence of massive-scale emotional contagion through social networks

Adam D. I. Kramer[a,1], Jamie E. Guillory[b,2], and Jeffrey T. Hancock[b,c]



the amount of emotional content in the News Feed. When positive expressions were reduced, people produced fewer positive posts and more negative posts; when negative expressions were reduced, the opposite pattern occurred. These results indicate that

- We know about this because *Facebook told us*

Political Manipulation

A 61-million-person experiment in social influence and political mobilization

Robert M. Bond[1], Christopher J. Fariss[1], Jason J. Jones[2], Adam D. I. Kramer[3], Cameron Marlow[3], Jaime E. Settle[1] & James H. Fowler[1,4]

Social message

- Experiment that increased turnout by 340,000 voters in the 2010 US congressional election

Behavioral Manipulation

TOP SECRET

Behavioural Science Support for JTRIG's (Joint Threat Research and Intelligence Group's) Effects and Online HUMINT Operations

Psychology-Based Influence Techniques

3.6 *Obedience* is a direct form of social influence where an individual submits to, or complies with, an authority figure. Obedience may be explained by factors such as diffusion of responsibility, perception of the authority figure being legitimate, and socialisation (including social role). Compliance can be achieved through various techniques including: Engaging the norm of reciprocity; engendering liking (e.g., via ingratiation or attractiveness); stressing the importance of social validation (e.g., via highlighting that others have also complied); instilling a sense of scarcity or secrecy; getting the "foot-in-the-door" (i.e., getting compliance to a small request/issue first); and applying the "door-in-the-face" or "low-ball" tactics (i.e., asking for compliance on a large request/issue first and having hidden aspects to a request/issue that someone has already complied with, respectively). Conversely, efforts to reduce obedience may be effectively based around educating people about the adverse consequences of compliance; encouraging them to question authority; and exposing them to examples of disobedience.

3.7 *Conformity* is an indirect form of social influence whereby an individual's beliefs, feelings and behaviours yield to those (norms) of a social group to which the

https://firstlook.org/theintercept/document/2015/06/22/behavioural-science-support-jtrig/

Pitfalls

Methodological issues

Usage issues

Pitfalls

Methodological issues:

- Selection effects in input datasets used for training
- Aggregation also provides information to a model about individuals
- Removing controversial features does not remove discriminatory issues with the training data

Usage issues:

- Proprietary data and opaque algorithms
- Unintentional impacts of increased personalization e.g. filter bubbles
- Increased efficacy of suggestion; ease of manipulation
- Need a system to deal with misclassifications

Pitfalls

Methodological issues:

- Selection effects in input datasets used for training
- Aggregation also provides information to a model about individuals
- Removing controversial features does not remove discriminatory issues with the training data

Usage issues:

- Proprietary data and opaque algorithms
- Unintentional impacts of increased personalization e.g. filter bubbles
- Increased efficacy of suggestion; ease of manipulation
- Need a system to deal with misclassifications

Detection

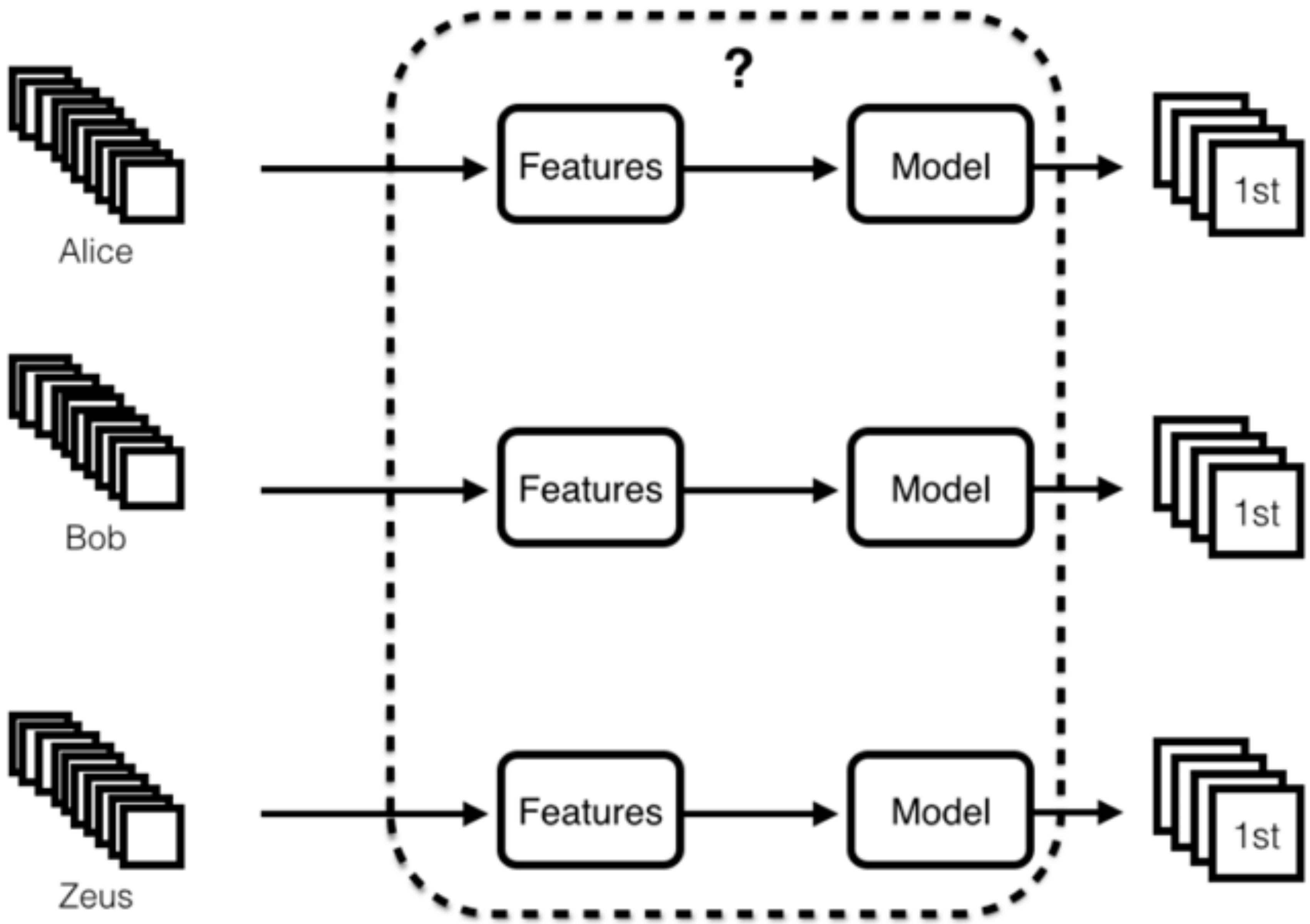
- How detectable is this type of engineering?
- Are these examples the tip of the iceberg?

How we detect this?
What can be done?

Policy

- Stronger consumer protections are needed
  - More explicit data use and privacy policies
  - Capacity to opt-out of certain types of experimentation
- Long-term: Give up less data
- Open algorithms and independent auditing: Ranking of feature importances

Black box analysis

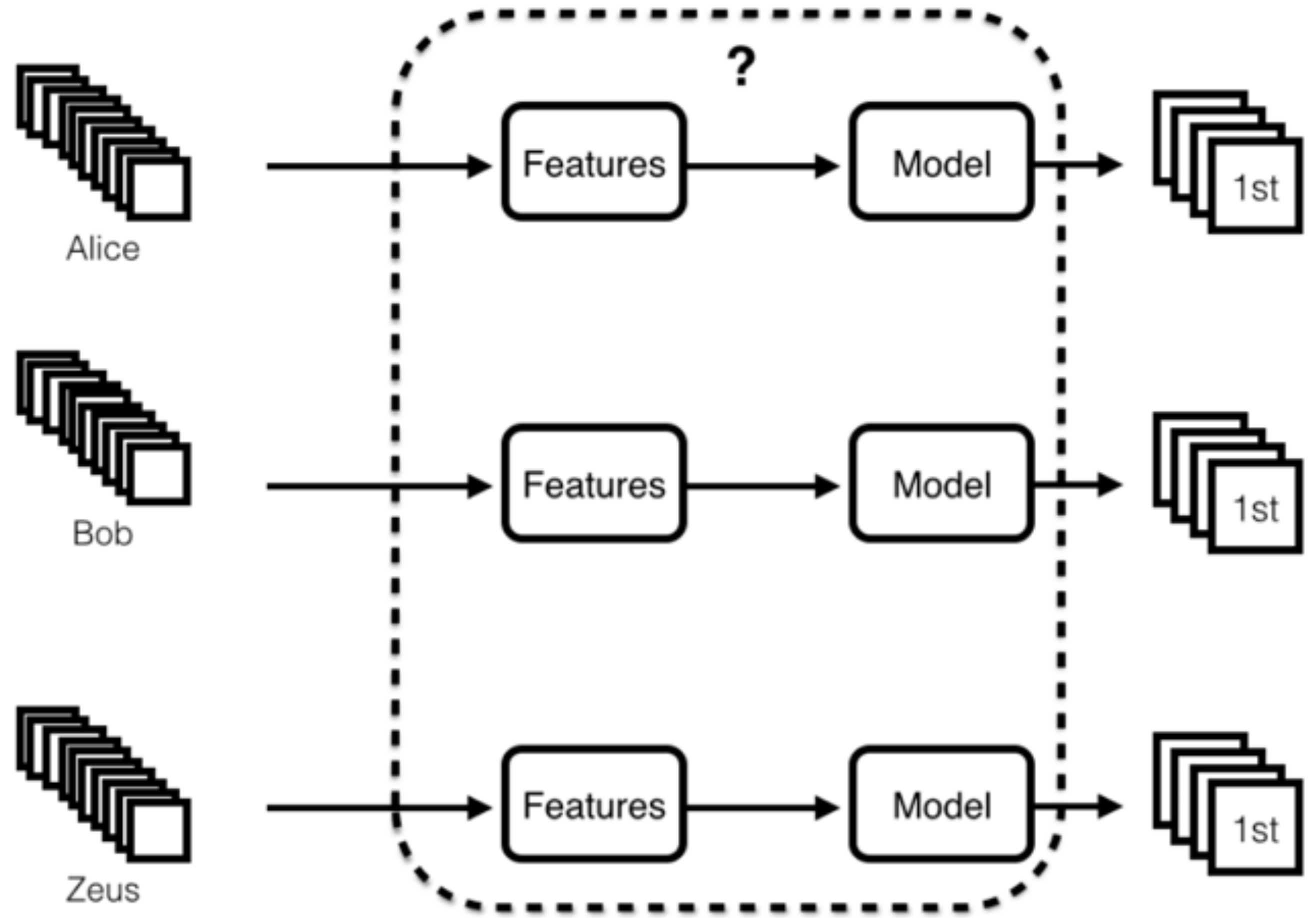


Black box analysis

Inputs:

Generate test accounts

Use real accounts



Outputs:

Compare outputs of algorithm

Why was one item shown to a given user and not another?

Black box analysis: XRay

- Nice example of how this type of analysis can be used to increase transparency [Usenix Security 2014]
- Uses test accounts on e.g. Gmail and feeds keywords and then records what ads are served

Debt/broke

Depression

http://xray.cs.columbia.edu/

Black box analysis: XRay

- Nice example of how this type of analysis can be used to increase transparency [Usenix Security 2014]
- Uses test accounts on e.g. Gmail and feeds keywords and then records what ads are served

Debt/broke	*Take a New Toyota Test Drive. Get a $50 gift card on the spot.*
Depression	*Text Coach - Get the girl you want and desire.*

http://xray.cs.columbia.edu/

Moving Forward

- To practitioners:
  - Algorithms are not impartial unless carefully designed
  - Biases in input data need to be considered
- To advocates:
  - Accountability and transparency is important for algorithms
  - We need both policy and technology to achieve this

Thanks!

twitter: @redshiftzero
email: jen@redshiftzero.com