DePaul University
College of Computing and Digital Media

Casey Bennett, PhD

May. 14, 2019

Last Week

- HW3 due today, HW4 releases
- Project Proposal feedback
- Presentation Schedule Released
  - Online Students let me know by Sunday

Projects

- Read the instructions on the Syllabus closely
- Online students, pay special attention to special instructions in ‘For Online Students’ on D2L
- Presentations due 11/13 and 11/20, assigned *randomly* (see schedule on D2L
- Final Paper due 11/21

Projects

- **Presentation**: Each project is to be presented using PowerPoint in a modified Pecha Kucha style – 20 slides 20 seconds each, on a timer

- Effective Communication – clear succinct, “data science” is your craft

Projects

- **Final Paper:** The report will be written in the format of a paper (abstract, introduction, literature review, methodology, results, discussion, conclusions and future work).
- The literature review for the final report consists of reading and summarizing about 5 to 6 published papers on the review topic. *Proper citations in text*.
- Approximately 6-7 pages long. Single Spaced. Common IEEE conference length.

https://pollev.com/caseybennett801

or text “caseybennett801” to 37607

1) Feature Selection

- Select a subset of relevant features

2) Feature Extraction (or agglomeration)

- Smush features together

3) Feature Construction (or engineering)

- Create new features out of raw data

Feature Construction

Main idea is that we want to create more *relevant* features out of the raw data

1) Manual
   - Domain Experts

2) Automated
   - Deep Learning

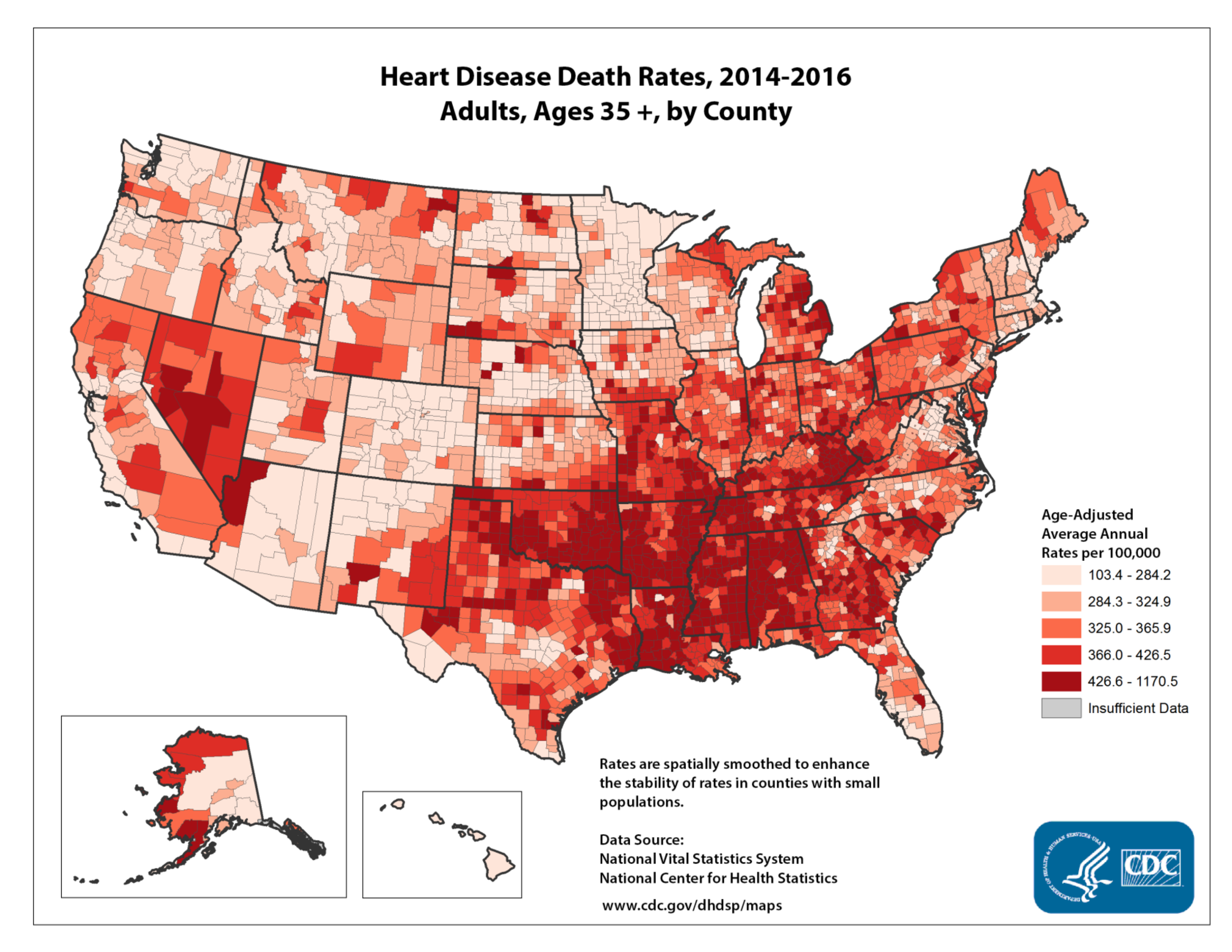
Dataset, know 2 things:

1) Where someone lives

2) If they died from heart disease

Knowing only those two pieces of info, how could I create a *more* relevant feature?

Geographic Risk Score



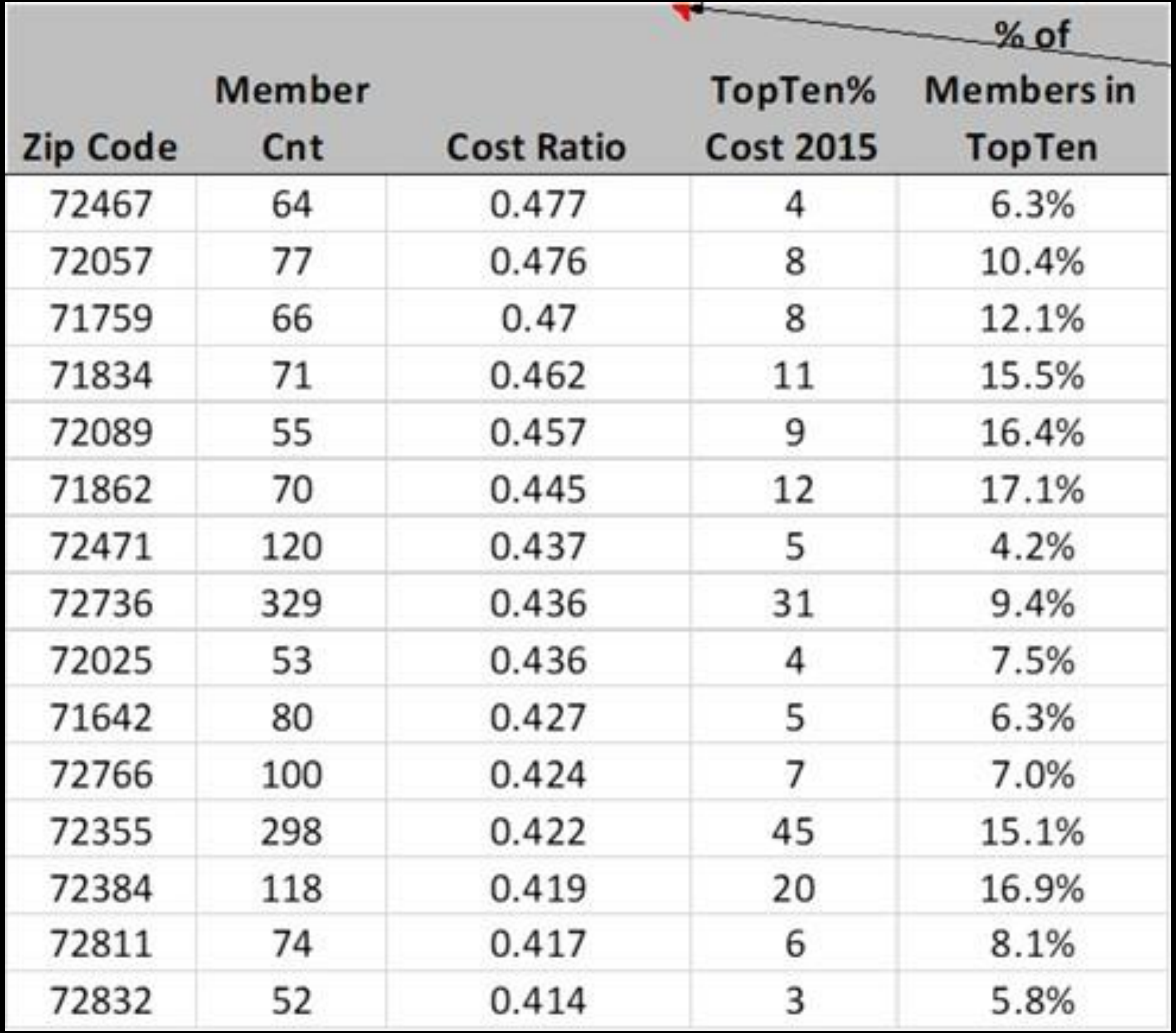
Heart Disease Death Rates, 2014-2016
Adults, Ages 35 +, by County

Rates are spatially smoothed to enhance the stability of rates in counties with small populations.

Data Source:
National Vital Statistics System
National Center for Health Statistics

www.cdc.gov/dhdsp/maps

Geographic Risk Score

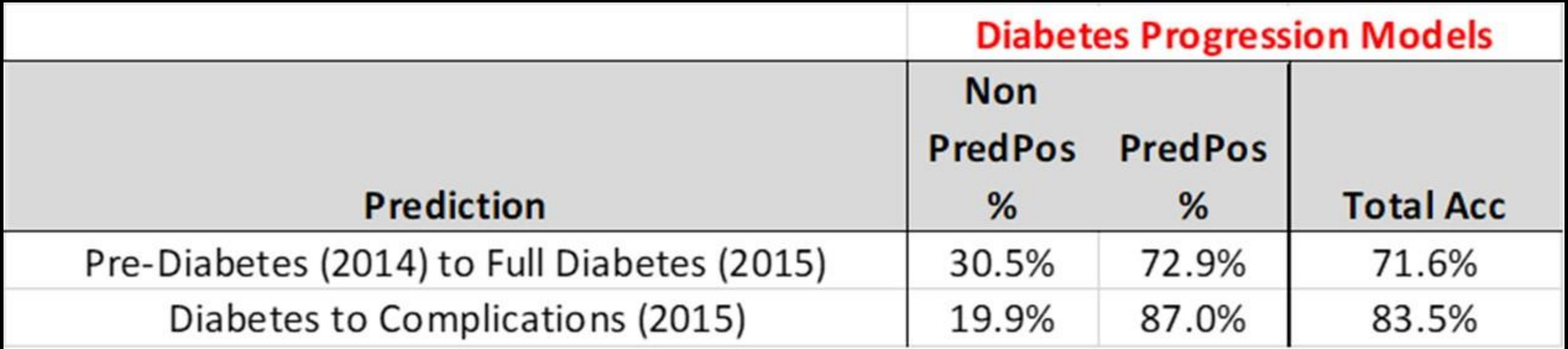


Zip Code	Member Cnt	Cost Ratio	TopTen% Cost 2015	% of Members in TopTen
72467	64	0.477	4	6.3%
72057	77	0.476	8	10.4%
71759	66	0.47	8	12.1%
71834	71	0.462	11	15.5%
72089	55	0.457	9	16.4%
71862	70	0.445	12	17.1%
72471	120	0.437	5	4.2%
72736	329	0.436	31	9.4%
72025	53	0.436	4	7.5%
71642	80	0.427	5	6.3%
72766	100	0.424	7	7.0%
72355	298	0.422	45	15.1%
72384	118	0.419	20	16.9%
72811	74	0.417	6	8.1%
72832	52	0.414	3	5.8%

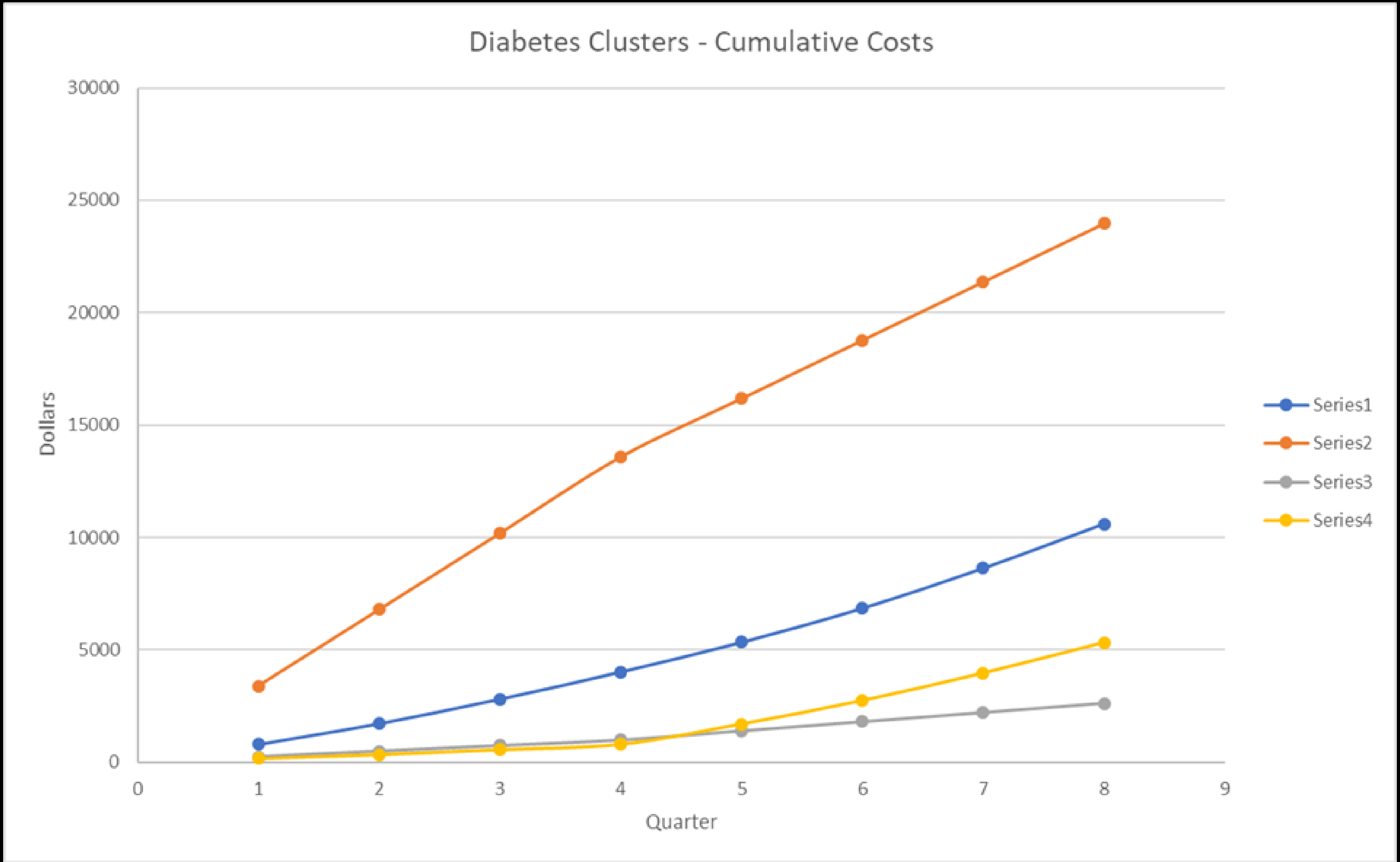
- Based on historical variations in utilization patterns
- In short, where someone lives is (not surprisingly) predictive of their future health and utilization patterns

Diabetes – Real World Example

- Evaluated a large state-wide population in the U.S. of over 300,000 unique patients spanning 3 years from 2014-2016 using random forests
- Payor claims data and social determinants of health data
- Can we detect meaningful clusters of trajectories for *diabetes progression*, in order to create cost-effective screening programs



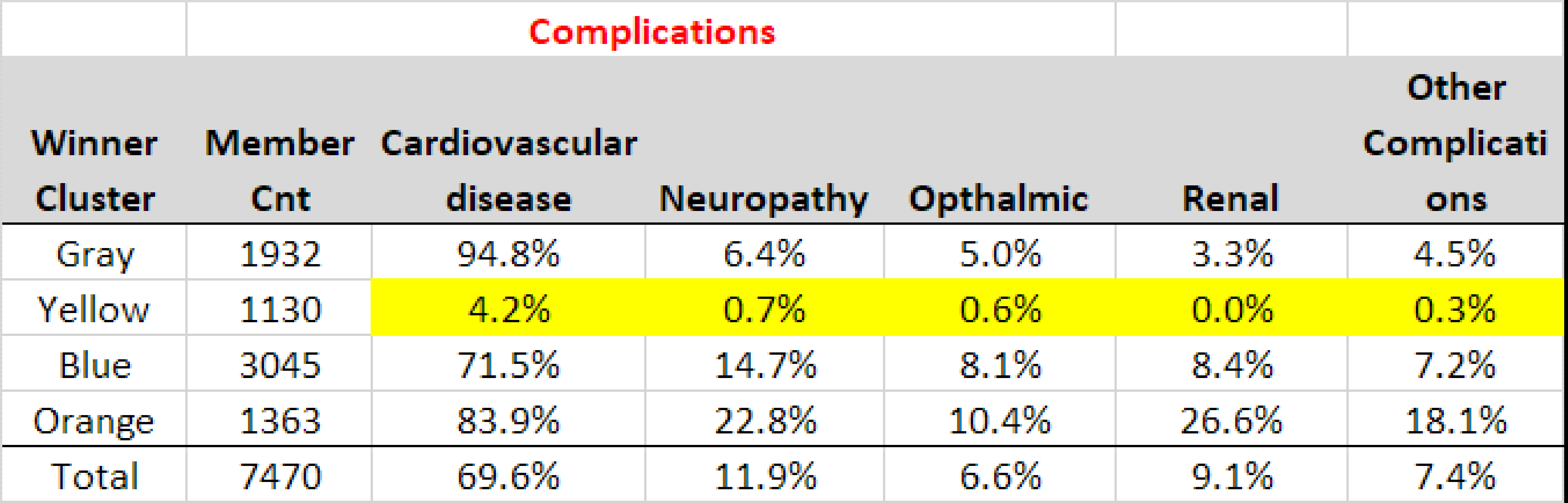
	Diabetes Progression Models		
Prediction	Non PredPos %	PredPos %	Total Acc
Pre-Diabetes (2014) to Full Diabetes (2015)	30.5%	72.9%	71.6%
Diabetes to Complications (2015)	19.9%	87.0%	83.5%



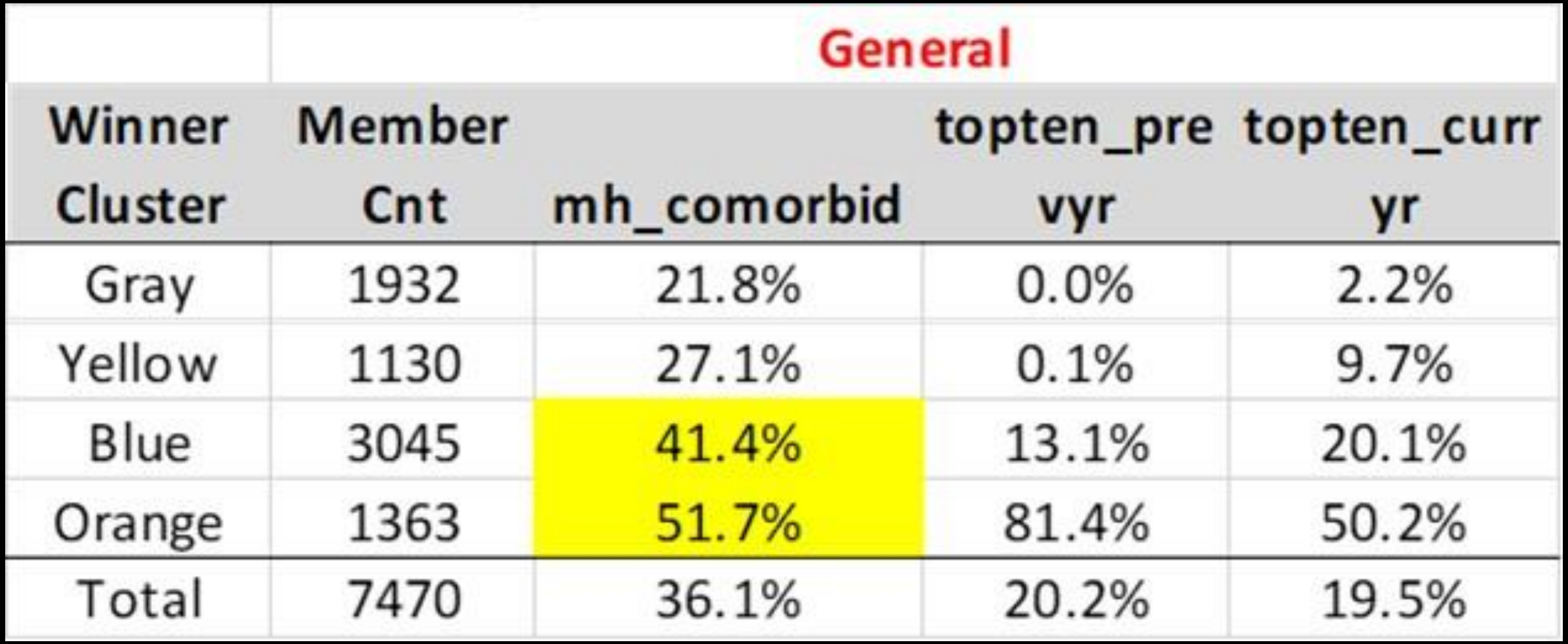
Differences in:

- Utilization Patterns
- Complications
- Medication Stage

****Orange and Blue groups were TWICE as likely to have mental health comorbidity***



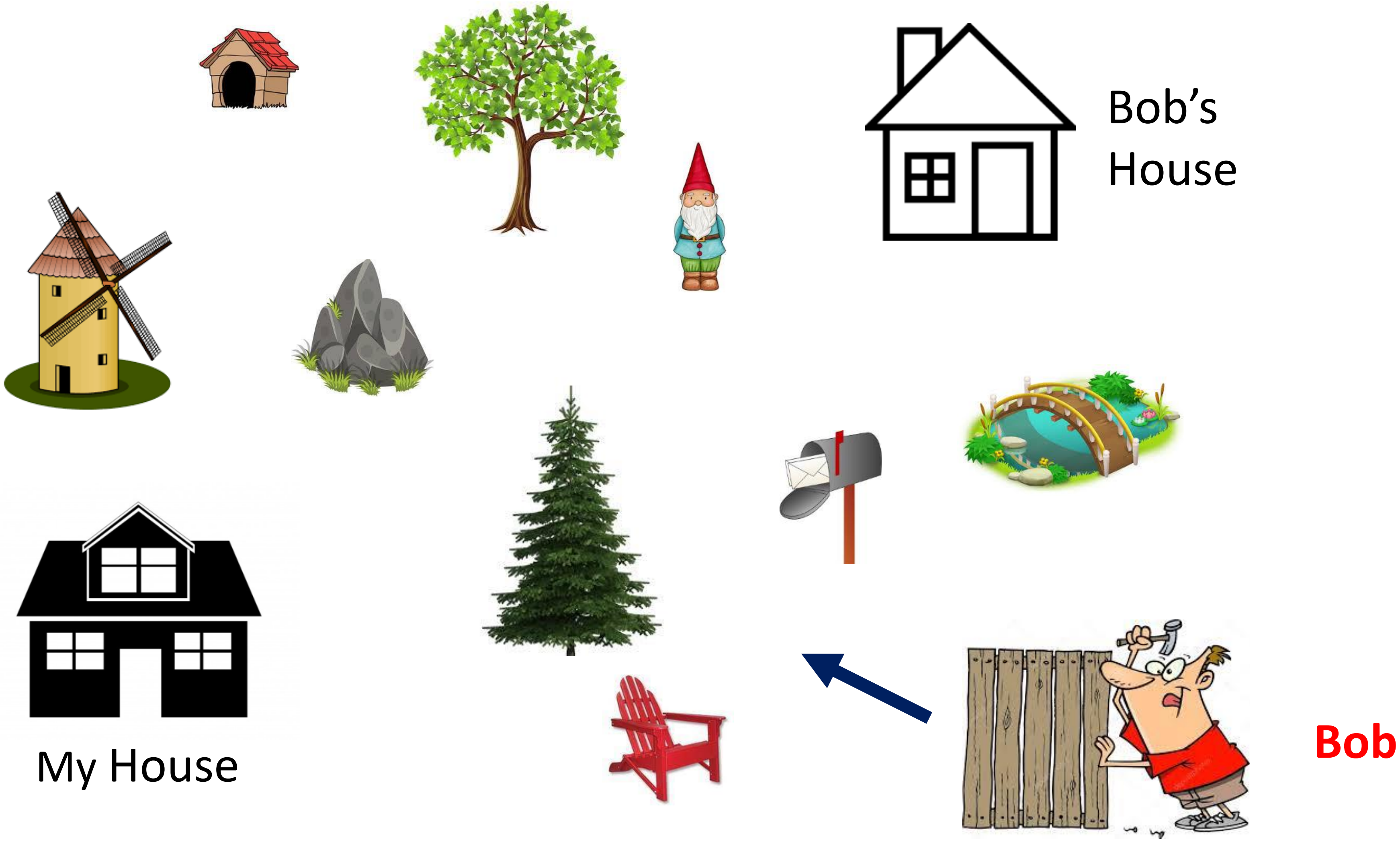
	Complications					
Winner Cluster	Member Cnt	Cardiovascular disease	Neuropathy	Opthalmic	Renal	Other Complications
Gray	1932	94.8%	6.4%	5.0%	3.3%	4.5%
Yellow	1130	4.2%	0.7%	0.6%	0.0%	0.3%
Blue	3045	71.5%	14.7%	8.1%	8.4%	7.2%
Orange	1363	83.9%	22.8%	10.4%	26.6%	18.1%
Total	7470	69.6%	11.9%	6.6%	9.1%	7.4%



	General			
Winner Cluster	Member Cnt	mh_comorbid	topten_prevyr	topten_curryr
Gray	1932	21.8%	0.0%	2.2%
Yellow	1130	27.1%	0.1%	9.7%
Blue	3045	41.4%	13.1%	20.1%
Orange	1363	51.7%	81.4%	50.2%
Total	7470	36.1%	20.2%	19.5%

SVM and Kernel Methods

BUILDING A FENCE



Bob's House

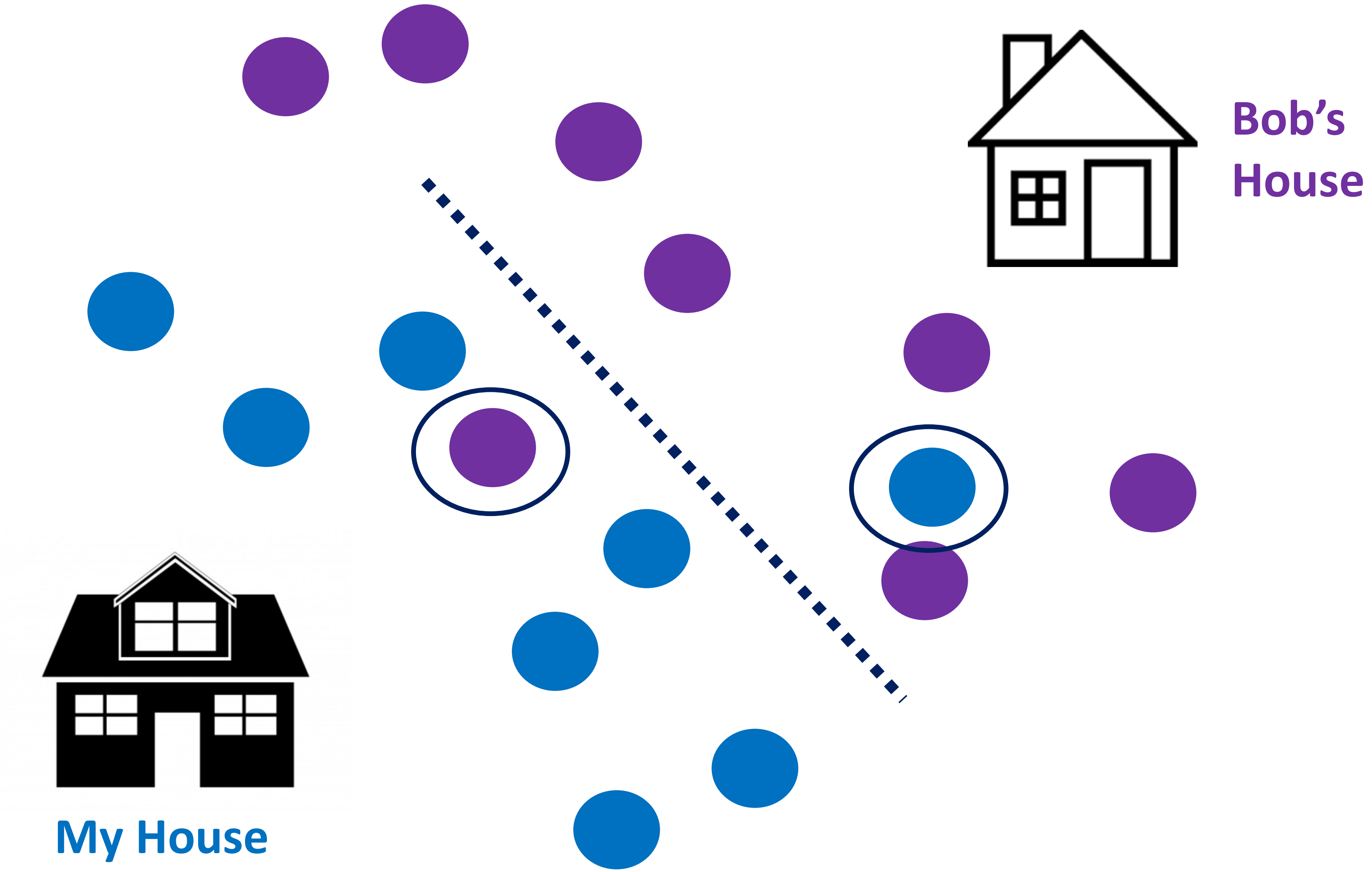
My House

Bob

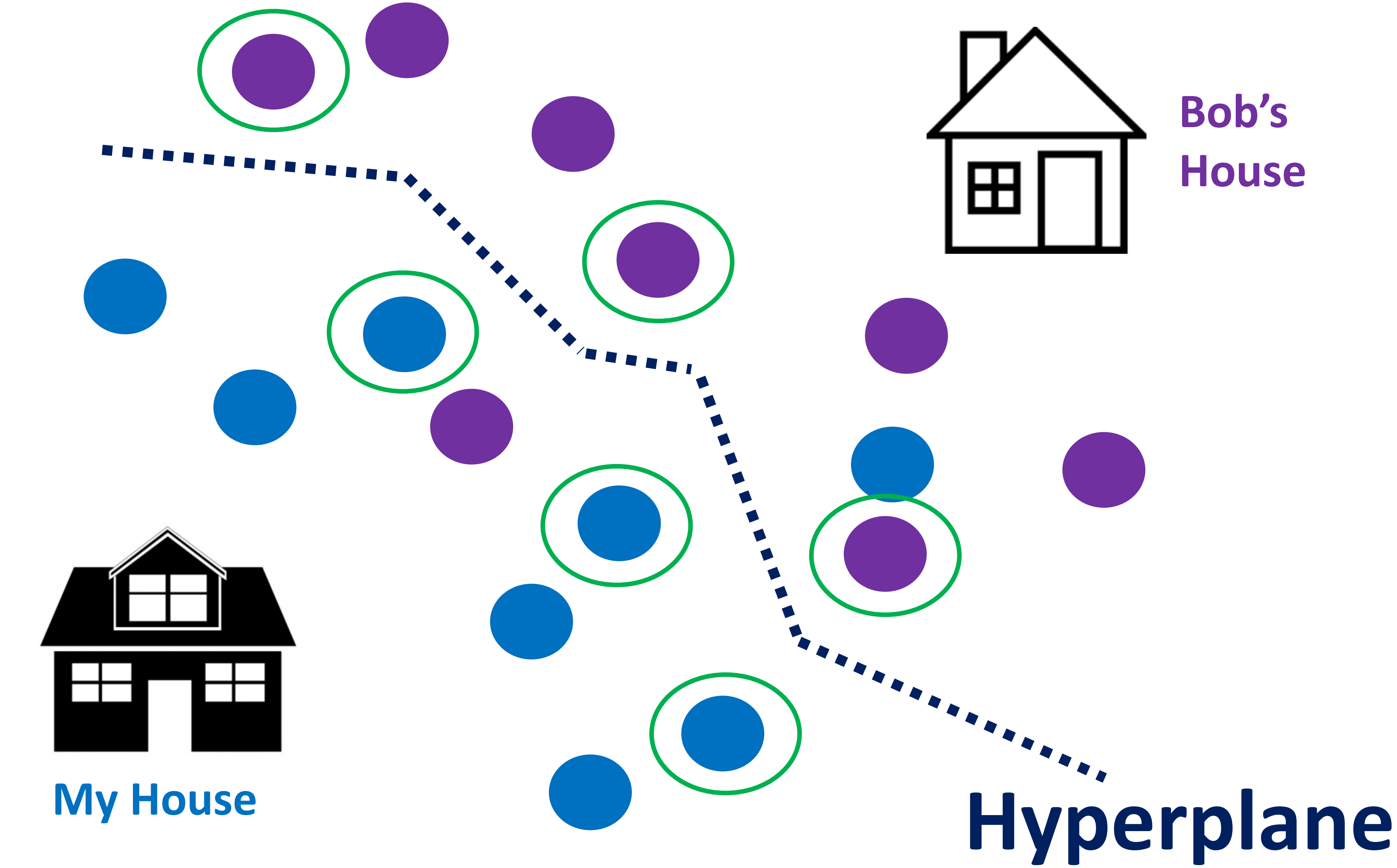
If we don’t have a property map or anything, then how could we figure out where to build the fence?

1

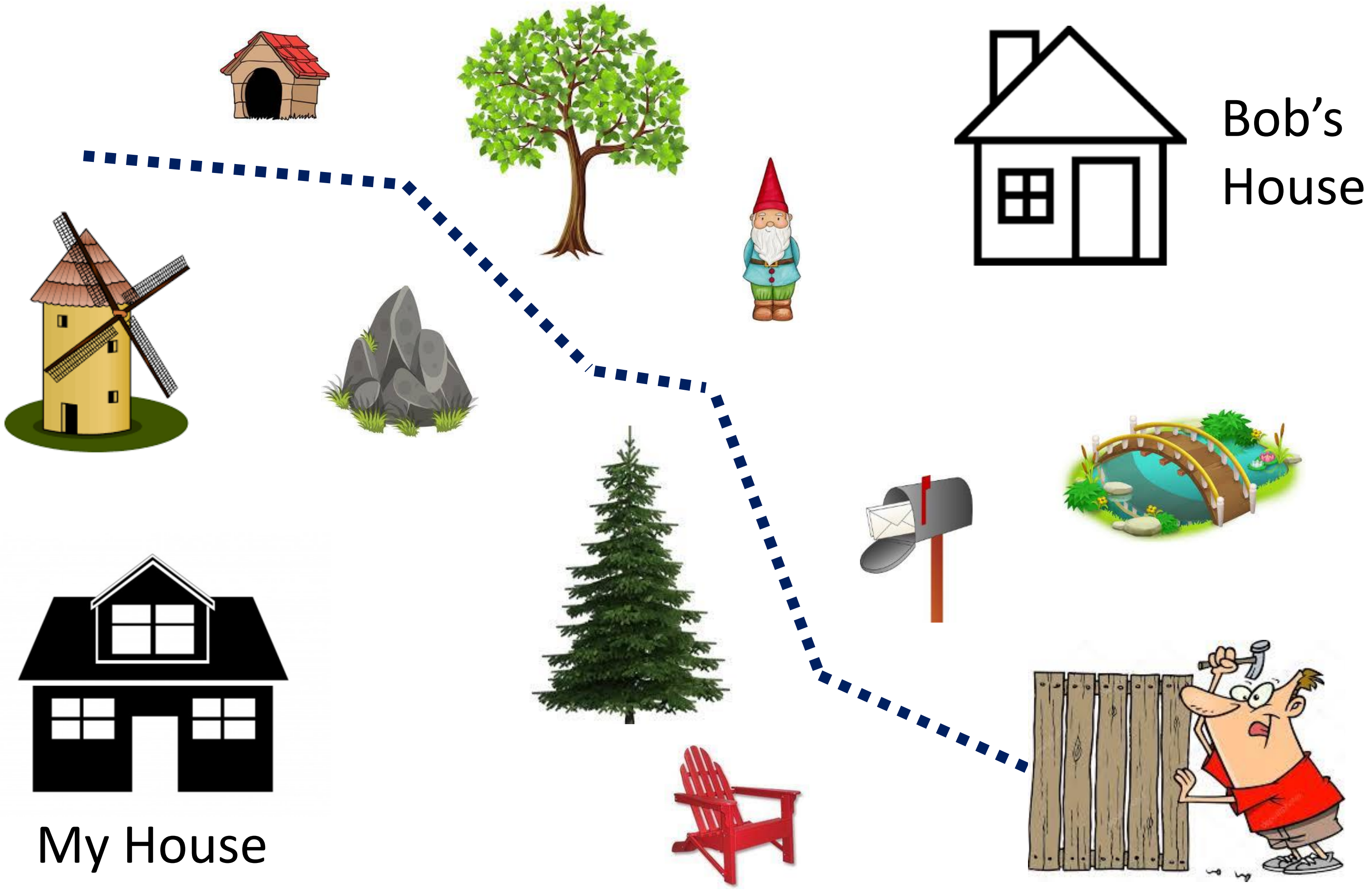
Errors



Support Vectors

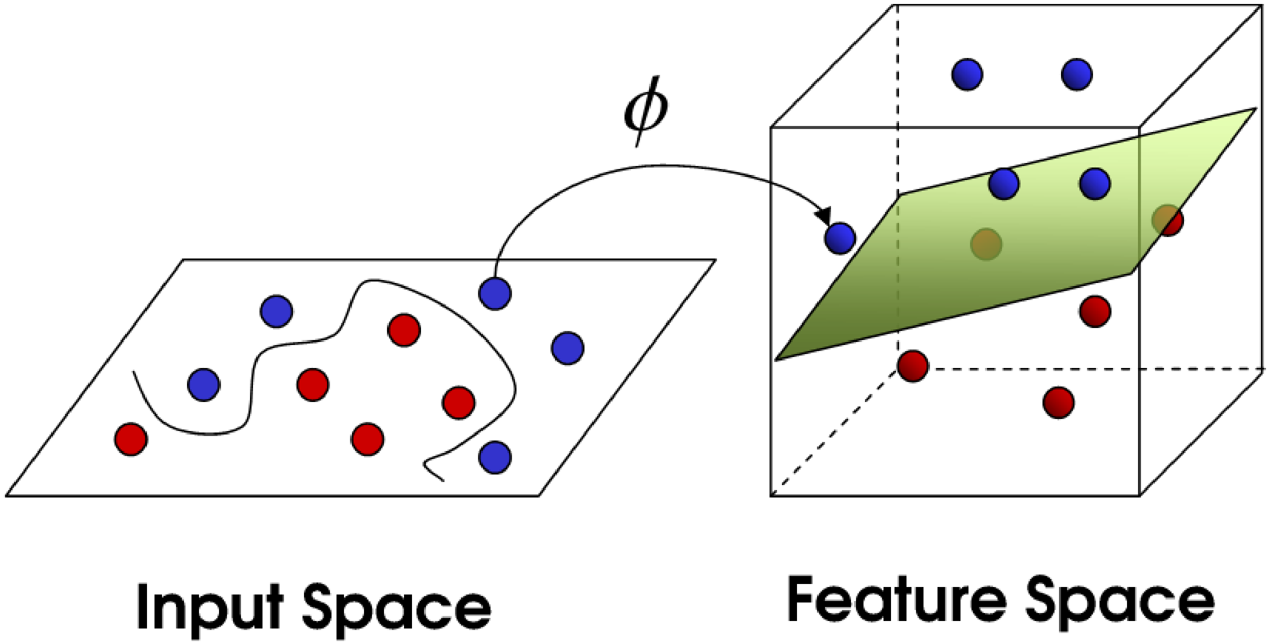


BUILDING A FENCE



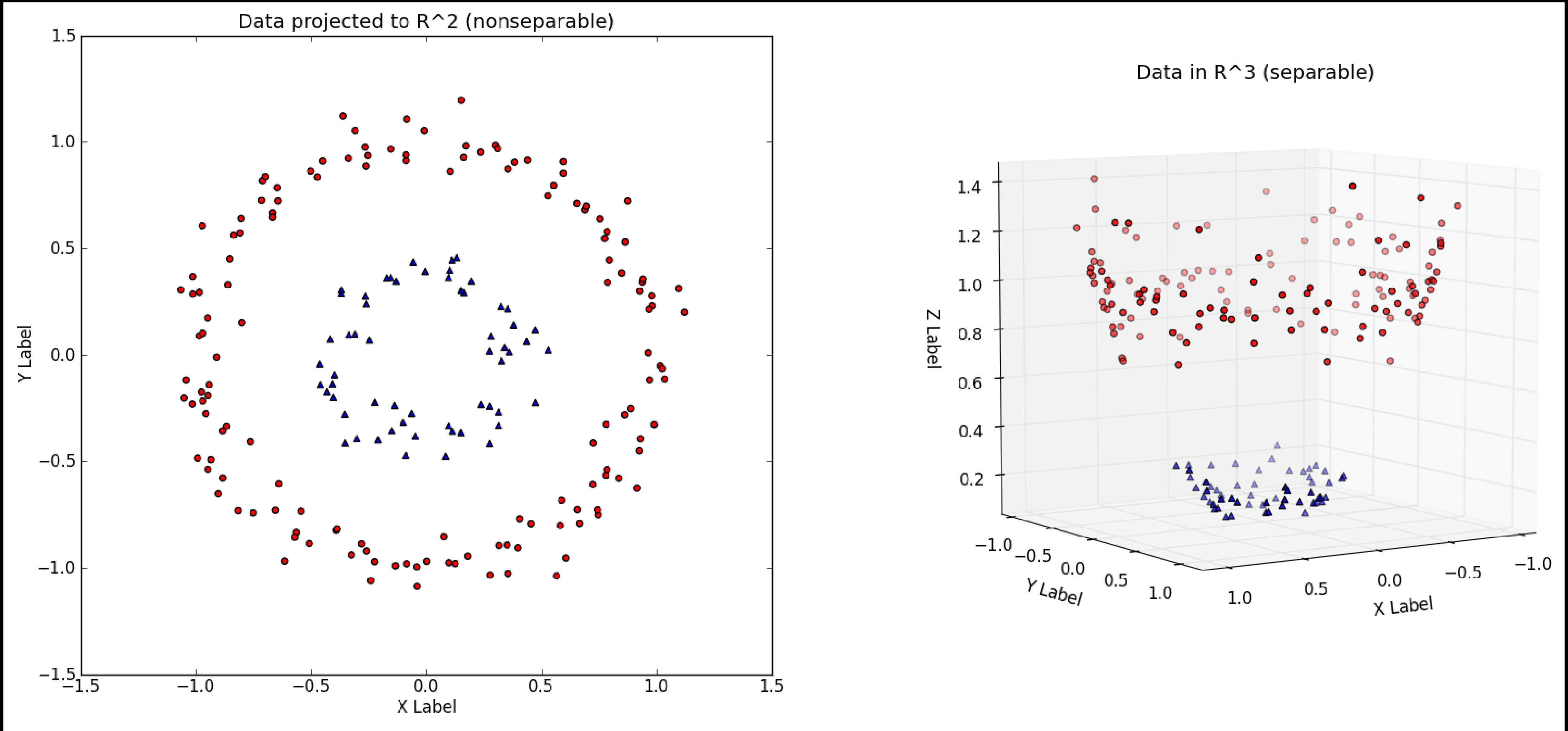
Bob’s House

My House



- Support Vector Machines (SVMs)
- Related to the general idea of discriminant analysis
- LDAs (Linear Discriminant Analysis) & QDAs (Quadratic Discriminant Analysis) are similar ideas

Kernel Trick



How can I know beforehand if my dataset is *linearly separable* in higher dimensional space?

SVM Kernel Issues

- Theoretically all data would be separable in *infinite* dimension space
- But the higher number of dimensions you map to, the greater chance of overfitting

SVM Kernels

$$K(x,y) = \langle fK(x), fK(y) \rangle$$

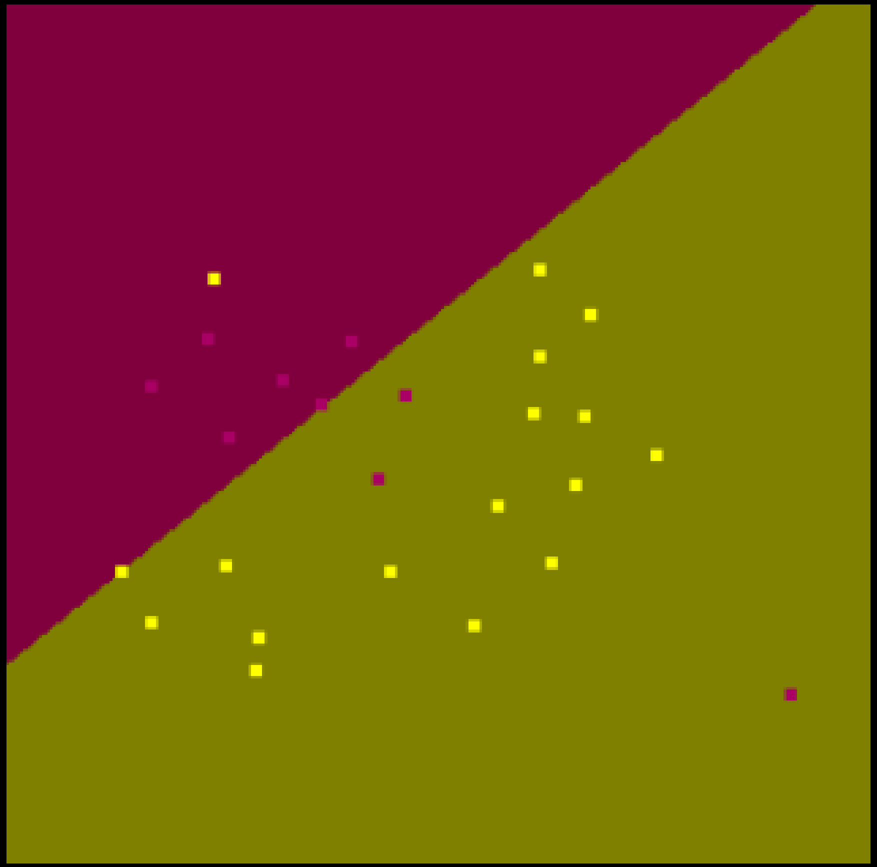
- Where fK is some kernel function
- In short, we are creating a new third dimension K(x,y) for an existing (x, y) point
- Often, we can use some dot product in place of fK

Different Types of Kernels

1) Linear
2) Polynomial
3) Sigmoid
4) RBF (radial basis function)
5) TanH (hyperbolic tangent)
6) Gaussian
7) Laplacian
8) Linear Splines

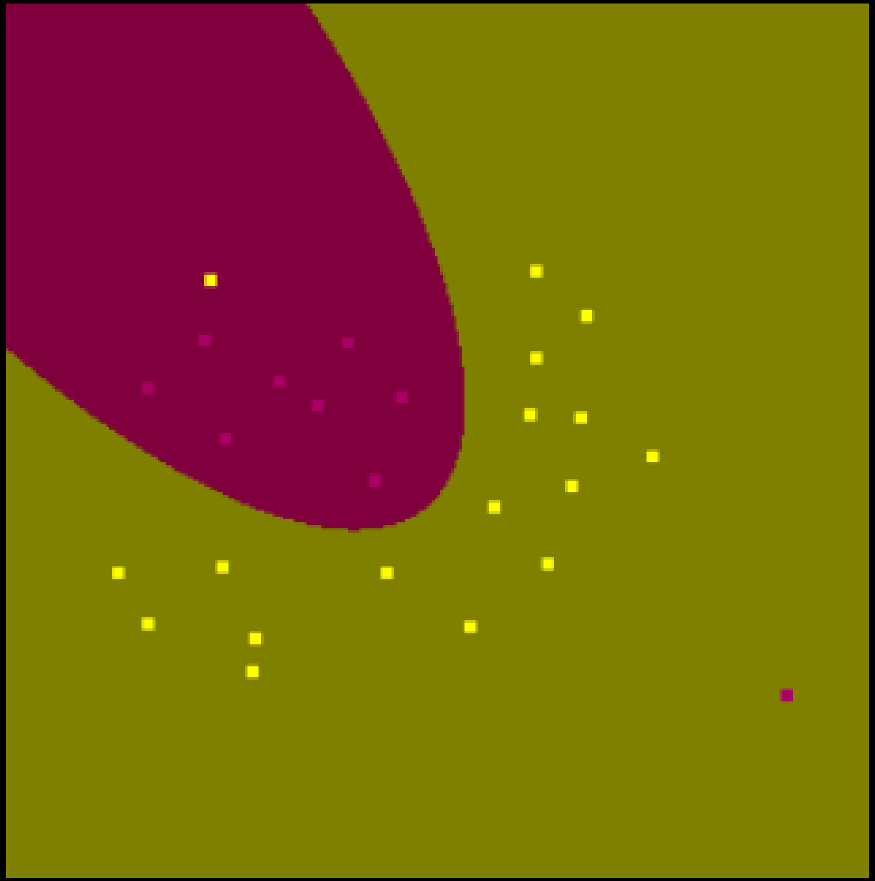
Linear Kernel

K(x,y) = plain old dot product of x and y



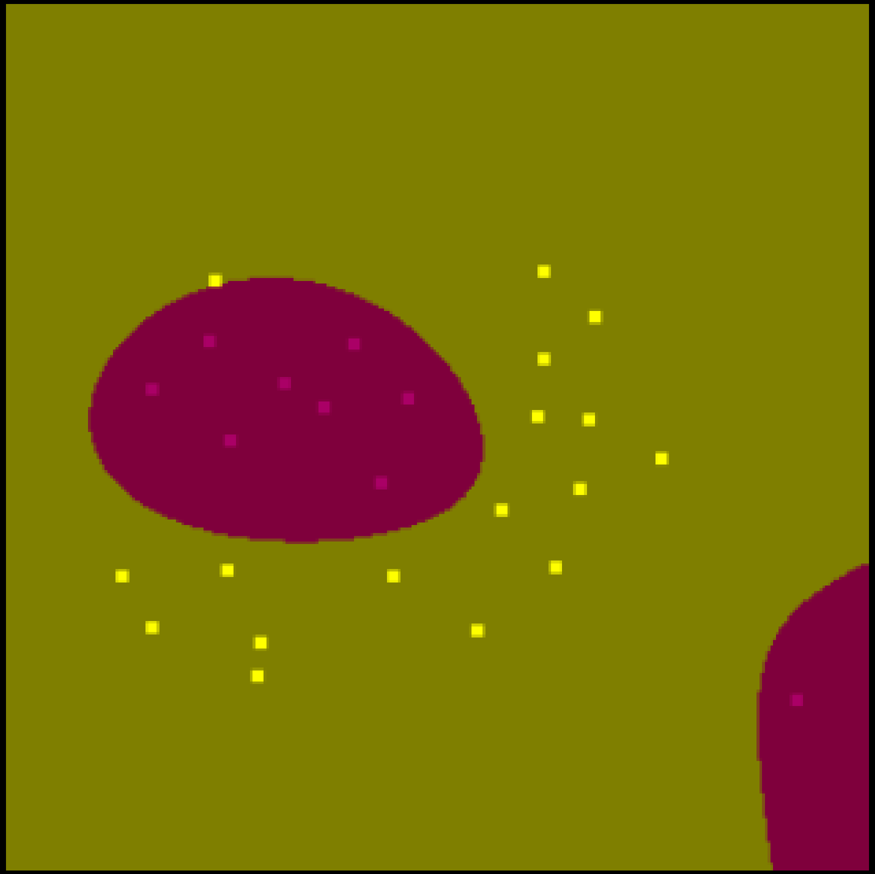
Polynomial Kernel

$$K(x,y) = (X \cdot Y + 1)^d$$



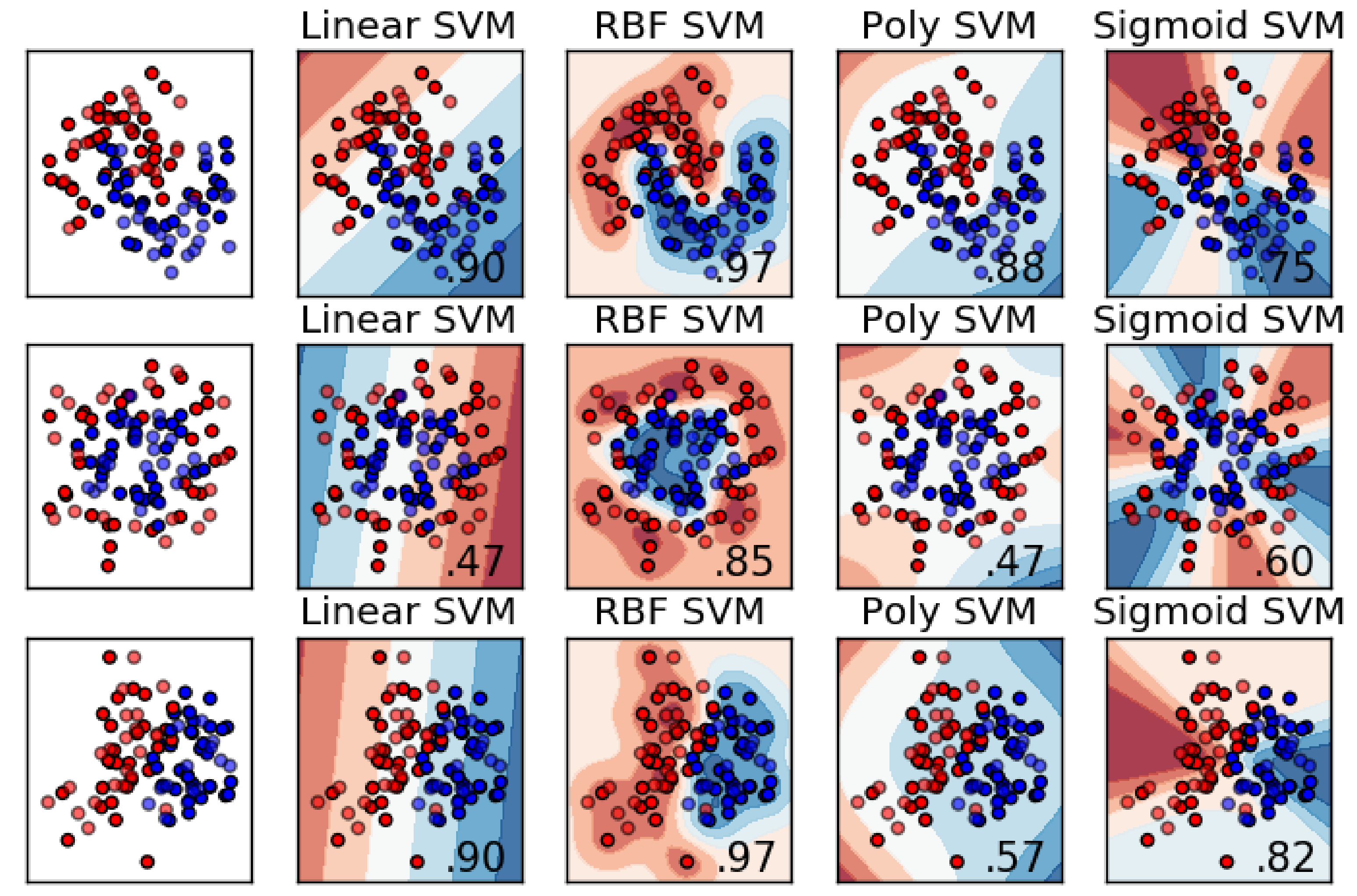
RBF Kernel

$$K(x,y) = \exp(-\gamma \cdot \|X-Y\|^2)$$



Sigmoid Kernel

$$K(x,y) = \tanh(\gamma \cdot X^{T}Y + c)$$



Code Implementation

```
#SciKit SVM
SVC(C=1.0, kernel=’rbf’, degree=3, gamma=’auto_deprecated’, coef0=0.0, shrinking=True,
      probability=False, tol=0.001, cache_size=200, class_weight=None, verbose=False, max_iter=-1,
      decision_function_shape=’ovr)
```

```
#Spark SVM
LinearSVC(labelCol="idxLabel", featuresCol="idxFeatures", maxIter=100, regParam = 0, tol = 1e-06,
                standardization = TRUE,  threshold = 0, weightCol = NULL, aggregationDepth = 2)
```

- There are also *Regressor* versions for SVM in Scikit, for when you have a continuous target variable you are trying to predict

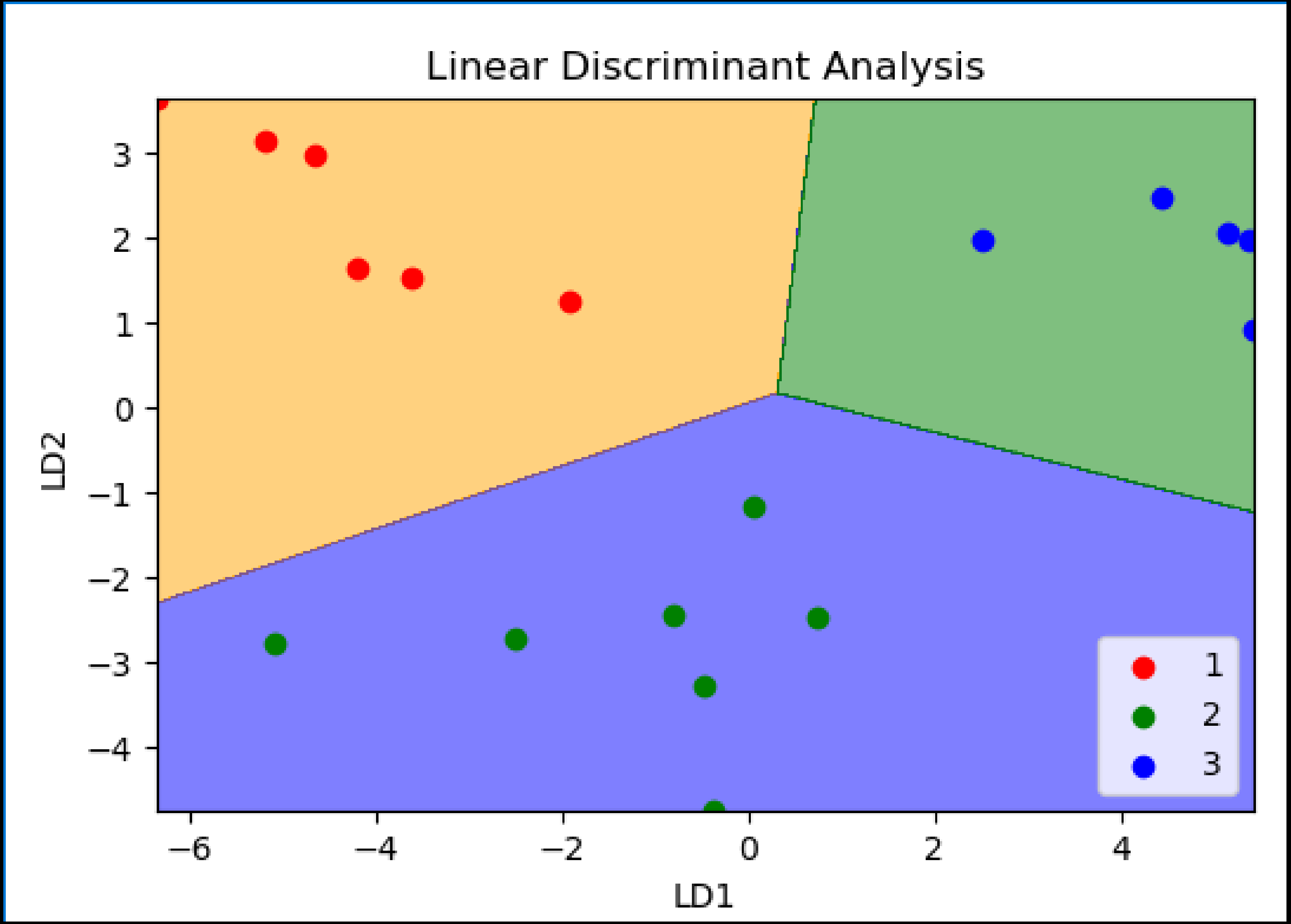
- No regressor version in Spark currently, and you can only use a linear kernel

Code Implementation

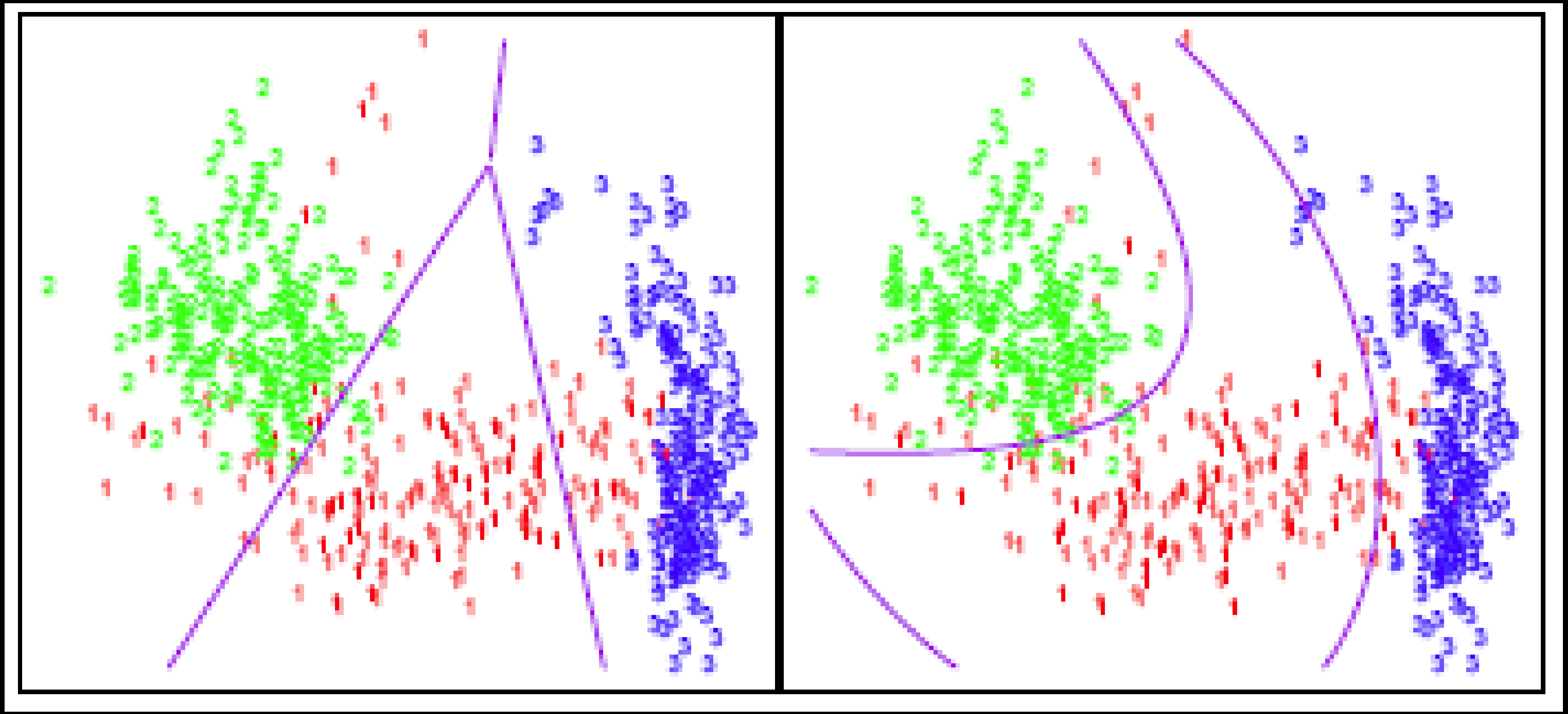
- Pay careful attention to a few parameters:
  - Number of iterations
  - Kernel
  - Gamma and degree parameters depending on kernel
  - “C” penalty parameter

Related Concepts to SVM

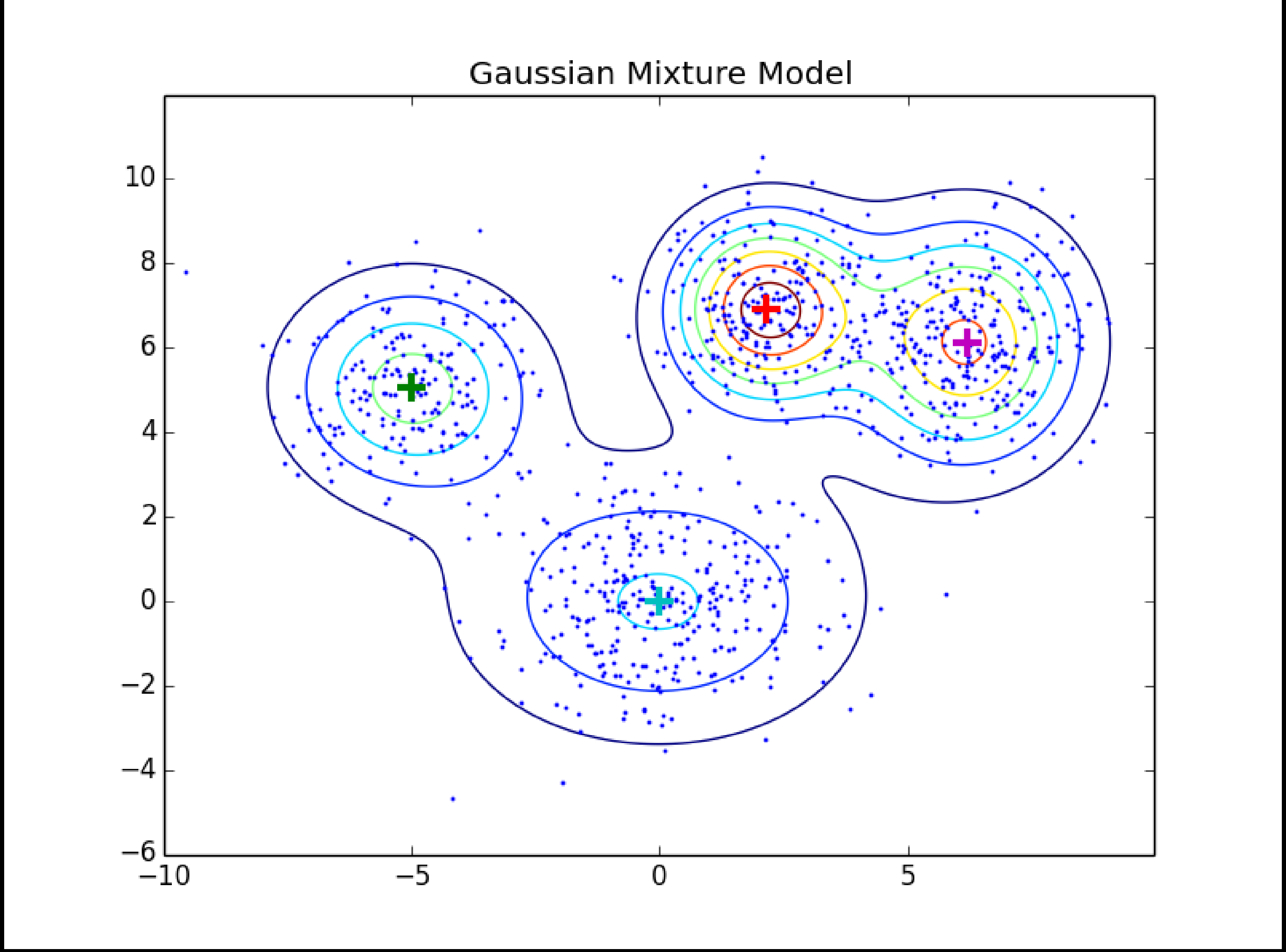
LDA



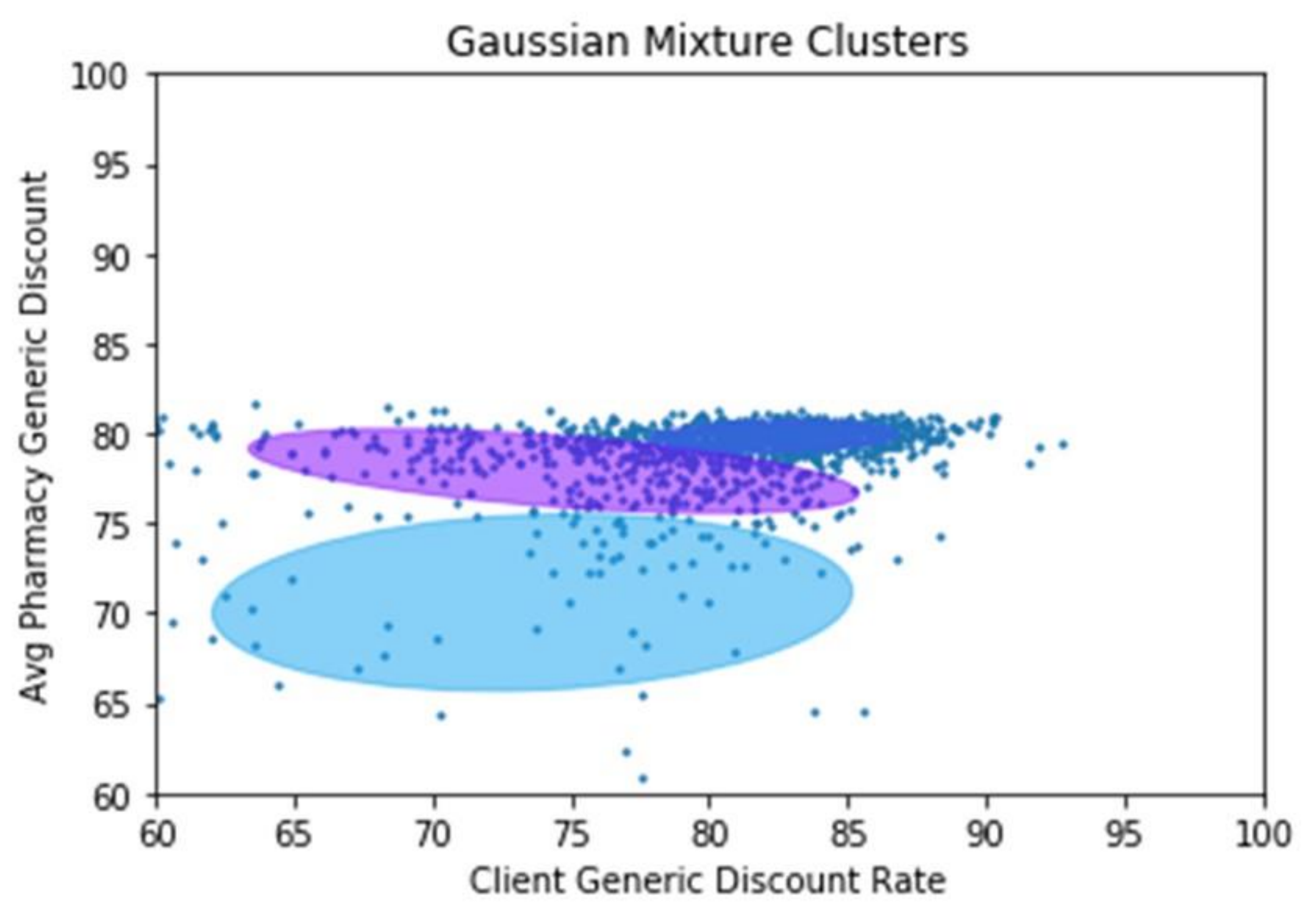
QDA - Quadratic Discriminant Analysis



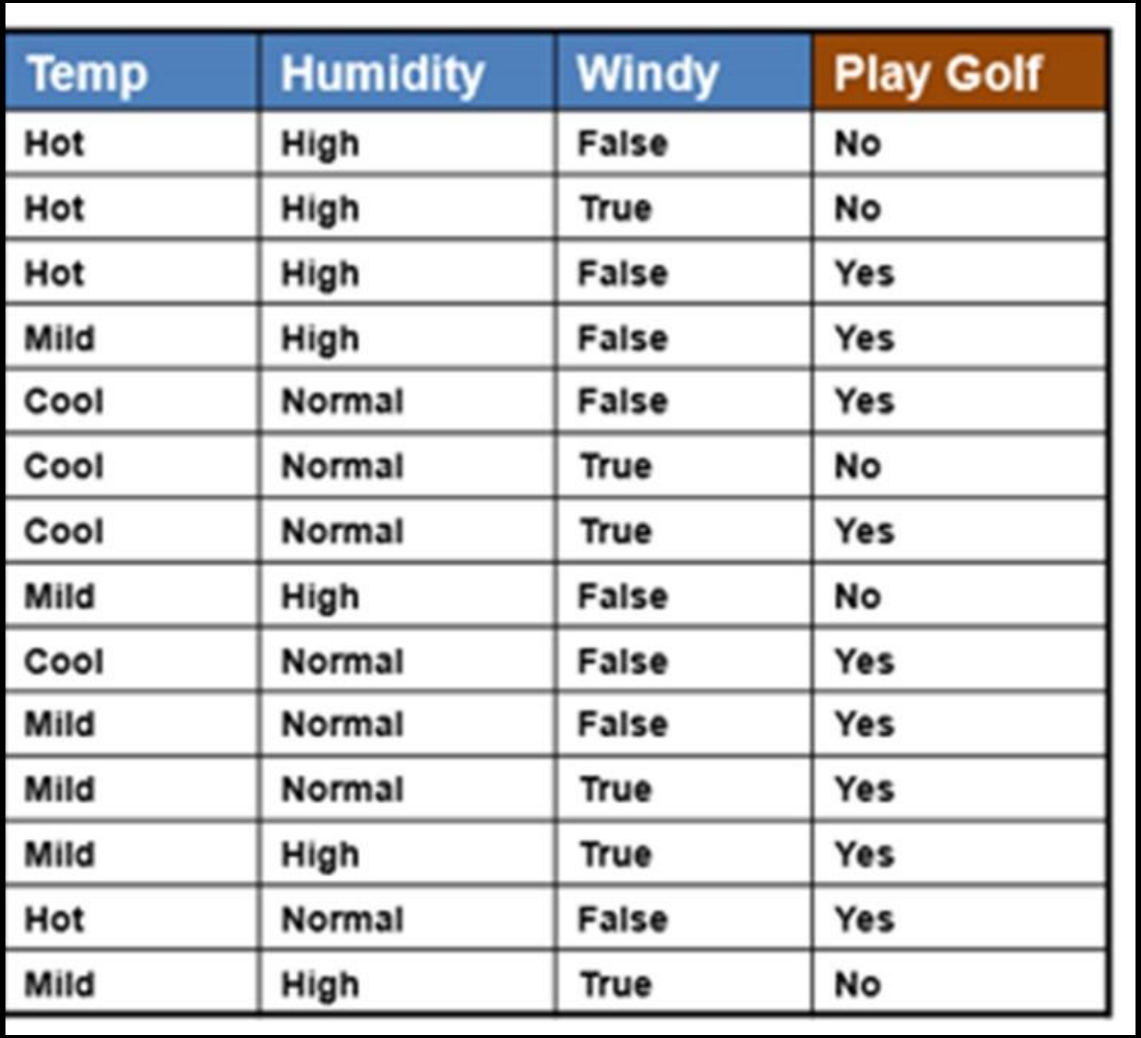
Gaussian Mixtures



Gaussian Mixtures

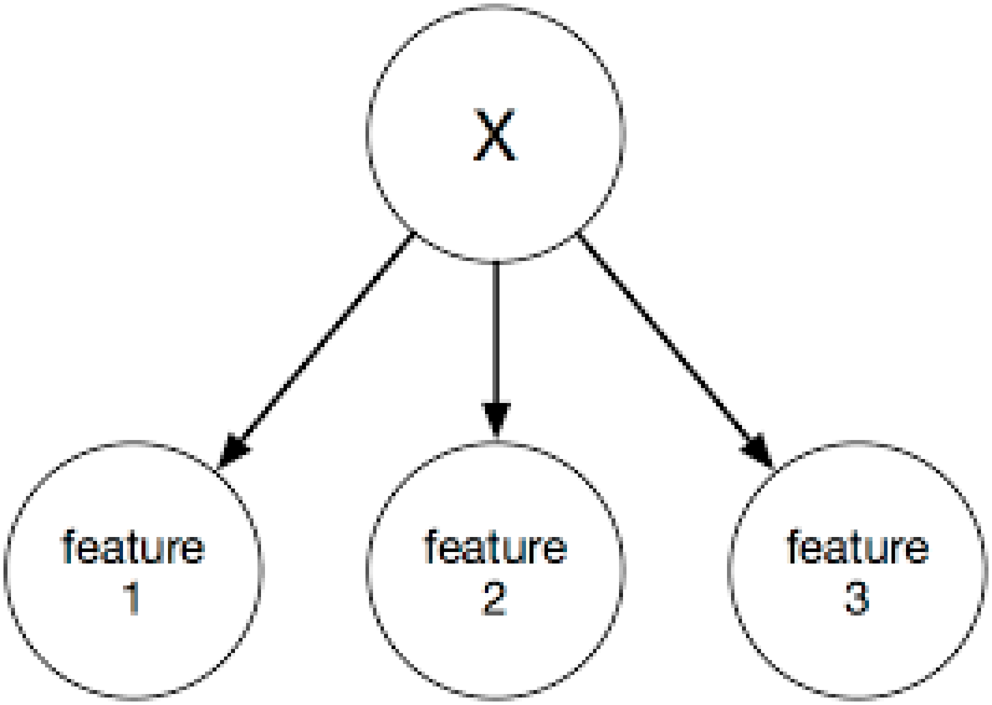


Naïve Bayes



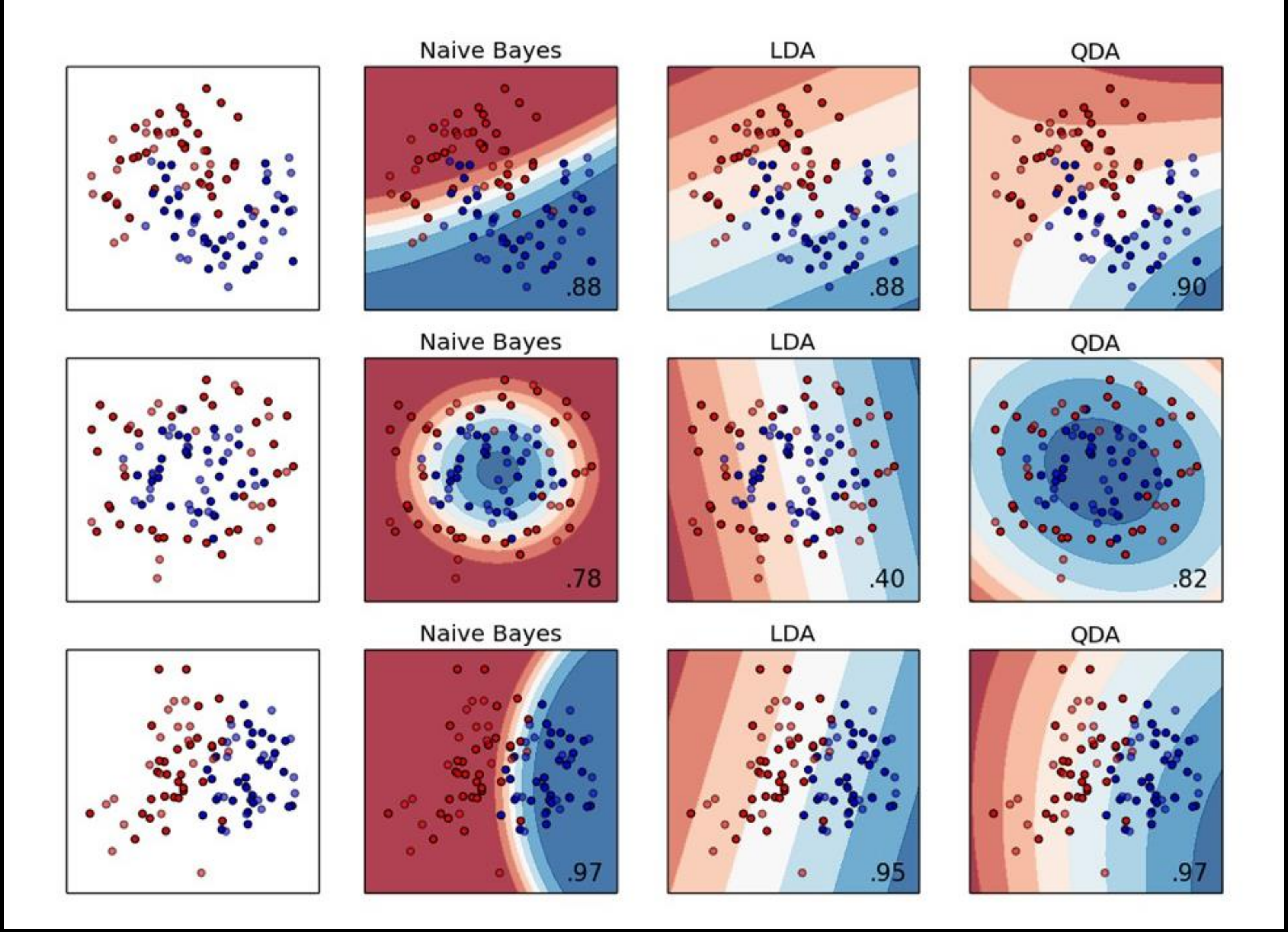
Temp	Humidity	Windy	Play Golf
Hot	High	False	No
Hot	High	True	No
Hot	High	False	Yes
Mild	High	False	Yes
Cool	Normal	False	Yes
Cool	Normal	True	No
Cool	Normal	True	Yes
Mild	High	False	No
Cool	Normal	False	Yes
Mild	Normal	False	Yes
Mild	Normal	True	Yes
Mild	High	True	Yes
Hot	Normal	False	Yes
Mild	High	True	No

Play = Yes (50% prob)



Temp = Hot

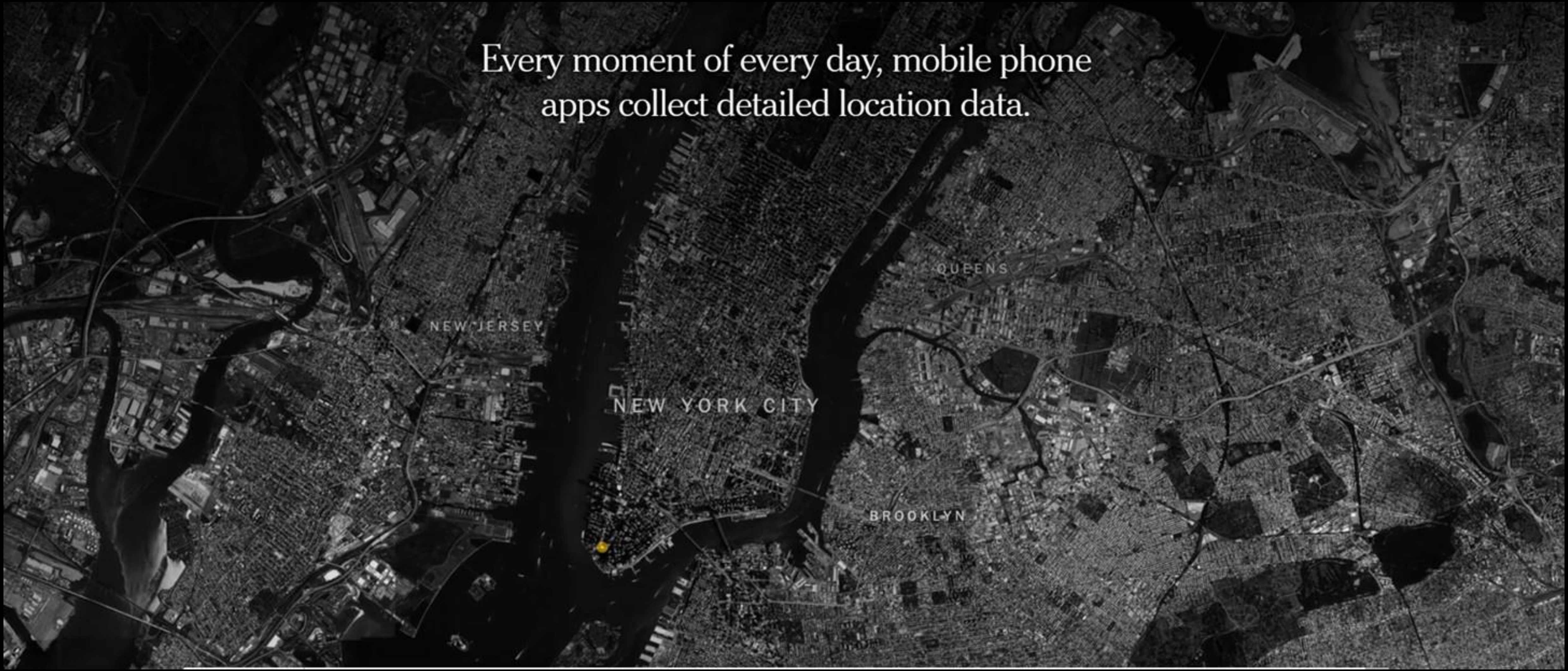
Gaussian Naïve Bayes



What should you do if you find “disturbing” patterns in the data, particularly if that pattern could make your company more money?

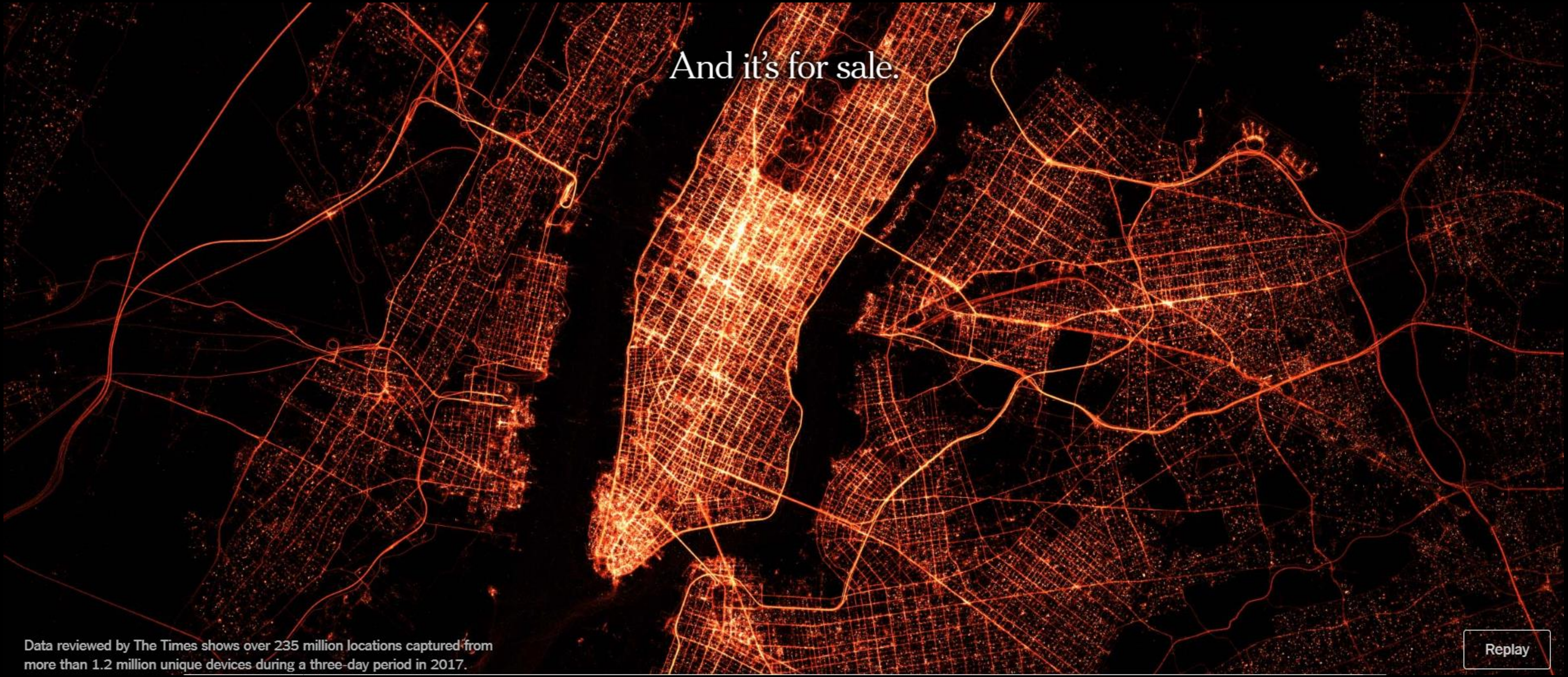
Special Topic: Ethics in Data Science

Location Tracking



Every moment of every day, mobile phone apps collect detailed location data.

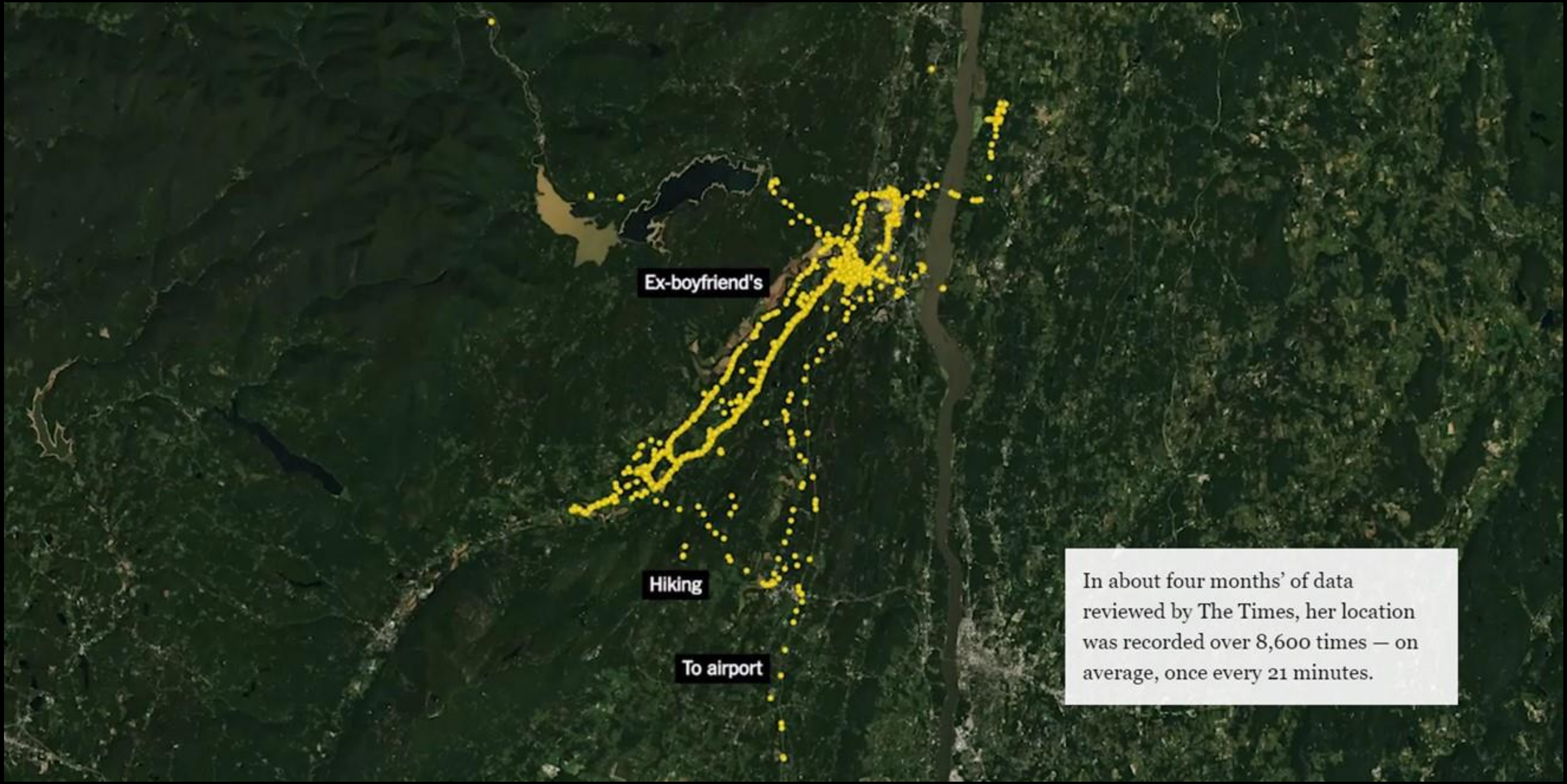
Location Tracking



And it’s for sale.

Data reviewed by The Times shows over 235 million locations captured from more than 1.2 million unique devices during a three-day period in 2017.

Location Tracking



In about four months' of data reviewed by The Times, her location was recorded over 8,600 times — on average, once every 21 minutes.

What should you do if you’re asked make predictions about people’s activities using such location data?

For next week

1) Paper Review #2 due next week
2) HW4 releases tonite after class
3) The above are your *last two* assignments, except for the final project