
1
0:13 / 0:19
HD

Hi Friends!

- **Welcome to CSE 474/574, Introduction to Machine Learning**
- My name is Kenny, or Prof. Joseph
- Some boring facts about me:
  - I am, perhaps unsurprisingly, not a very good dancer.
  - I am sitting at a kitchen table while making these slides
  - I like the color orange.

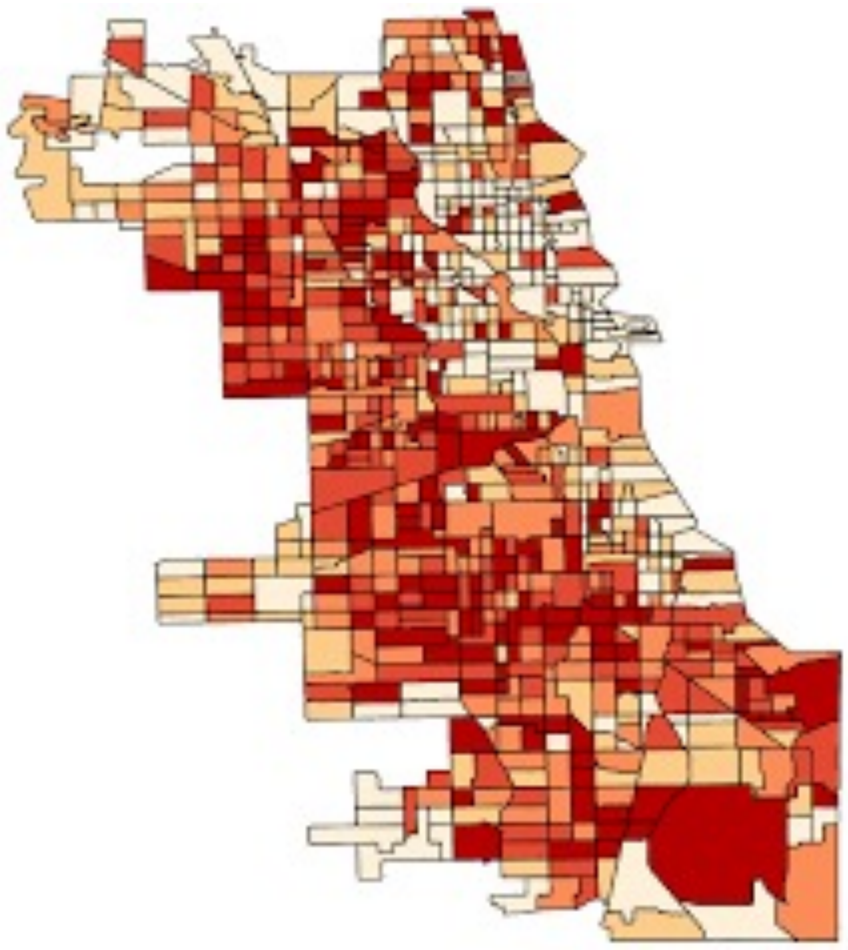
Plan today

- Intro to machine learning
- Intro to me/my teaching style
- Syllabus/Course review
- Start: A high-level example to keep in mind when we get bogged down in the details

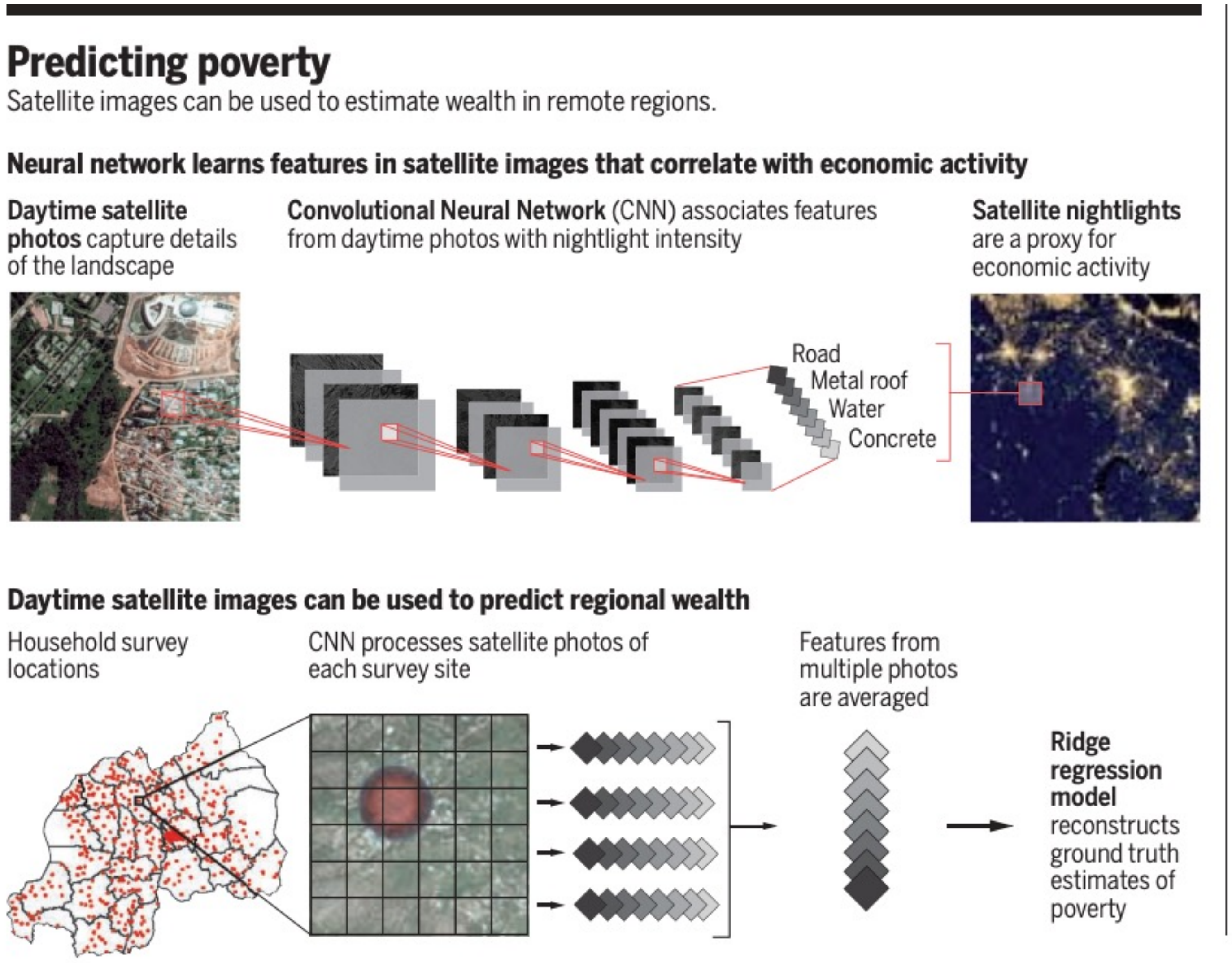
Machine Learning in the real world

- https://quickdraw.withgoogle.com/
- https://www.tiktok.com/foryou
- https://twitter.com/wowitsmrinal/status/1287175391040290816
- https://www.youtube.com/watch?v=cQ54GDm1eL0

- Any others?



Potash, E., Ghani, R., Walsh, J., Jorgensen, E., Lohff, C., Prachand, N., & Mansour, R. (2020). Validation of a Machine Learning Model to Predict Childhood Lead Poisoning. *JAMA network open*, *3*(9), e2012734-e2012734.



Jean, N., Burke, M., Xie, M., Davis, W. M., Lobell, D. B., & Ermon, S. (2016). Combining satellite imagery and machine learning to predict poverty. *Science*, *353*(6301), 790-794.

Machine learning is amazing!

Few questions here

- What is machine learning?
- (How) does it work?
- Is it always amazing?

Few questions here

- What is machine learning?
- (How) does it work?
- **Is it always amazing?**


MIT Technology Review

Subscribe

ARTIFICIAL INTELLIGENCE

Hundreds of AI tools have been built to catch covid. None of them helped.

Some have been used in hospitals, despite not being properly tested. But the pandemic could help make medical AI better.

By Will Douglas Heaven

July 30, 2021

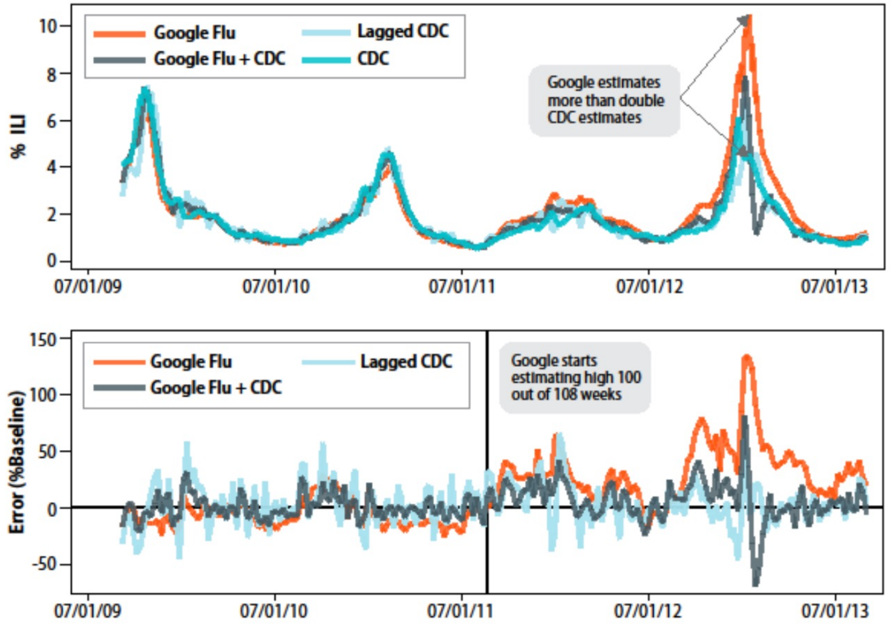


Fig. 1. GFT overestimation. GFT overestimated the prevalence of flu in the 2012–2013 season and overshot the actual level in 2011–2012 by more than 50%. From 21 August 2011 to 1 September 2013, GFT reported overly high flu prevalence 100 out of 108 weeks. (**Top**) Estimates of doctor visits for ILI. "Lagged CDC" incorporates 52-week seasonality variables with lagged CDC data. "Google Flu + CDC" combines GFT, lagged CDC estimates, lagged error of GFT estimates, and 52-week seasonality variables. (**Bottom**) Error [as a percentage of CDC baseline: (estimate by CDC)/CDC data]. Both alternative models have much less error than GFT alone. Mean absolute error (MAE) during the out-of-sample period is 0.486 for GFT, 0.311 for lagged CDC, and 0.232 for combined GFT and CDC. All of these differences are statistically significant at $P < 0.05$. See SM.

Lazer, D., Kennedy, R., King, G., & Vespignani, A. (2014). The parable of Google Flu: traps in big data analysis. *Science*, *343*(6176), 1203-1205.

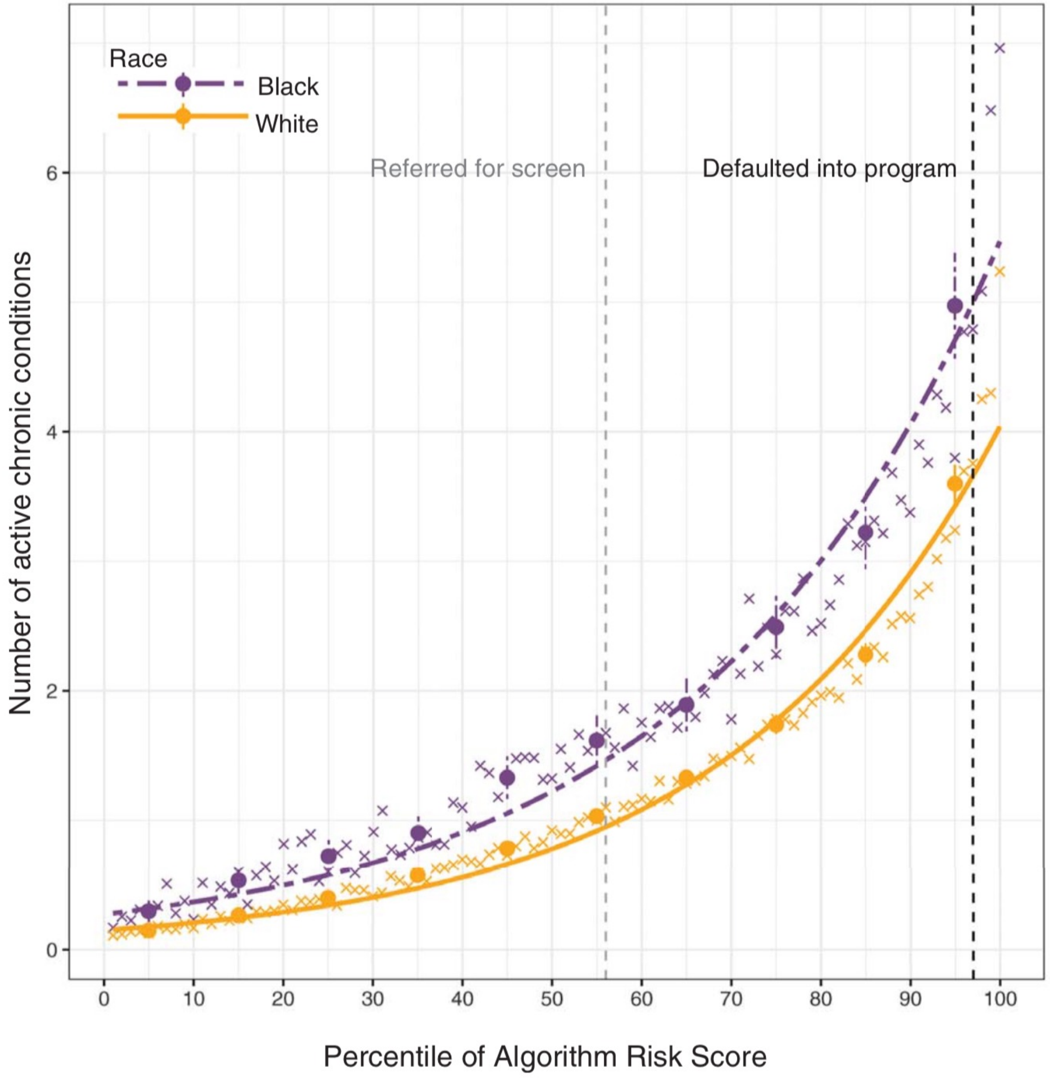


Fig. 1. Number of chronic illnesses versus algorithm-predicted risk, by race. (**A**) Mean number of chronic conditions by race, plotted against

MOTHERBOARD
TECH BY VICE

Flawed Algorithms Are Grading Millions of Students' Essays

Fooled by gibberish and highly susceptible to human bias, automated essay-scoring systems are being increasingly adopted, a Motherboard investigation has found

Research is scarce on the issue of machine scoring bias, partly due to the secrecy of the companies that create these systems. Test scoring vendors closely guard their algorithms, and states are wary of drawing attention to the fact that algorithms, not humans, are grading students' work. Only a

Meanwhile, it tended to underscore African Americans and, at various points, Arabic, Spanish, and Hindi speakers—even after attempts to reconfigure the system to fix the problem.

> **"The BABEL Generator proved you can have complete incoherence, meaning one sentence had nothing to do with another," and still receive a high mark from the algorithms.**

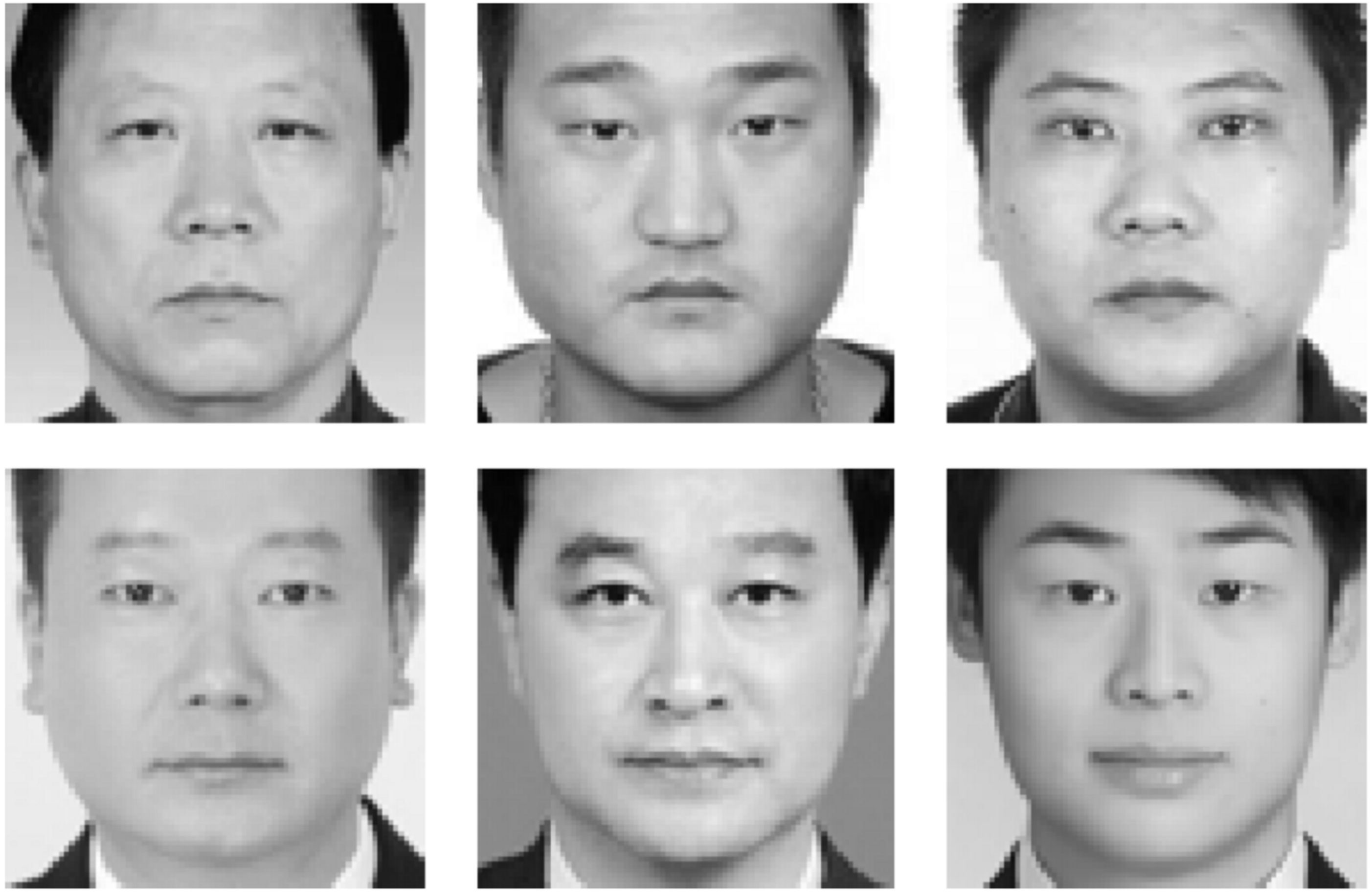


Figure 3. Wu and Zhang's "criminal" images (top) and "non-criminal" images (bottom). In the top images, the people are frowning. In the bottom, they are not. These types of superficial differences can be picked up by a deep learning system.


We're using
AI instead
of biased humans
What did you
train the AI on?
What did you
train the AI on?
imgflip.com

Is it always amazing?

- No. If you…
  - Use a bad/the wrong model
  - Ask a dumb question
  - Use crappy data
- Your work will be, **at best**, useless

Corollary: You have to know what you're doing and why you're doing it.

My aim in this class is to give you some insight into both of these.

More on my goals for us

- My goal is for you (and me!) to 1) learn and 2) have fun.
- We will learn and have the most fun if we are both…
  - Working hard, and smart, not long.
  - Engaged in lectures
  - …

Me as an instructor

- I am not scary. Wish I was, but I’m not.
- I am not out to ”get you”. Seems pointless
- ^ But,
  - I am strict. Please do not take this personally.
  - Trust-but-verify. Please do not violate this trust
  - **I have no tolerance for people making others feel not smart or not welcome.**
- I do not know you’re struggling unless you tell me.

Few questions here

- **What is machine learning?**
- (How) does it work?
- Is it always amazing?

Canonical Definition

"A computer program is said to **learn** from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E."

— Tom Mitchell, Professor at Carnegie Mellon University

Tom M Mitchell et al. "Machine learning. 1997". In: *Burr Ridge, IL: McGraw Hill* 45.37 (1997), pp. 870–877.

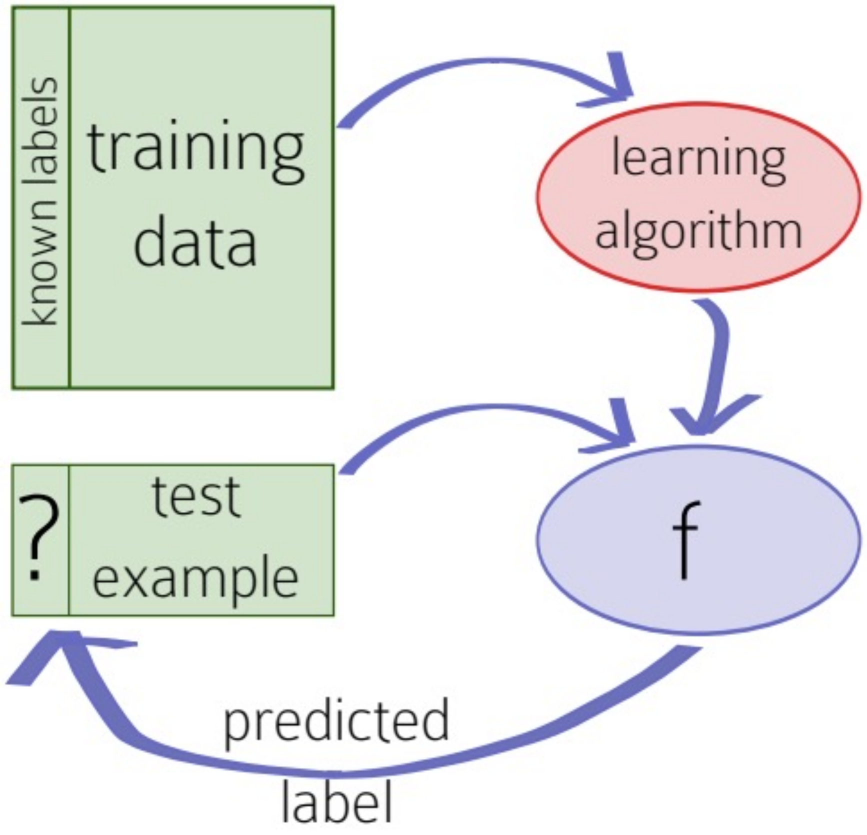
ML as a recipe creator

> *An algorithm is like a recipe. It takes "inputs" (the ingredients), performs a set of simple and (hopefully) well-defined steps, and then terminates after producing an "output" (the meal)*

A learning algorithm is a game of roulette on a 50 dimensional wheel that lands on a particular spot (a recipe) based completely on how it was trained, what examples it saw, and how long it took to search.

https://medium.com/@geomblog/when-an-algorithm-isn-t-2b9fe01b9bb5

ML as generalization of (training) data



CIML, Chapter 1

ML as generalization of (training) data

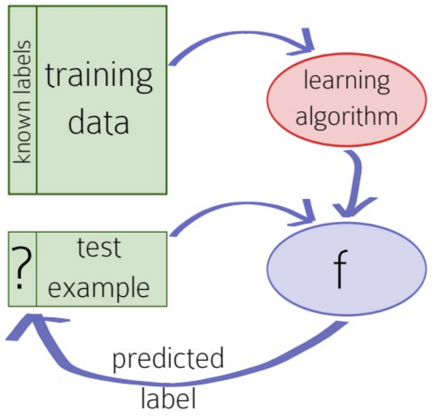
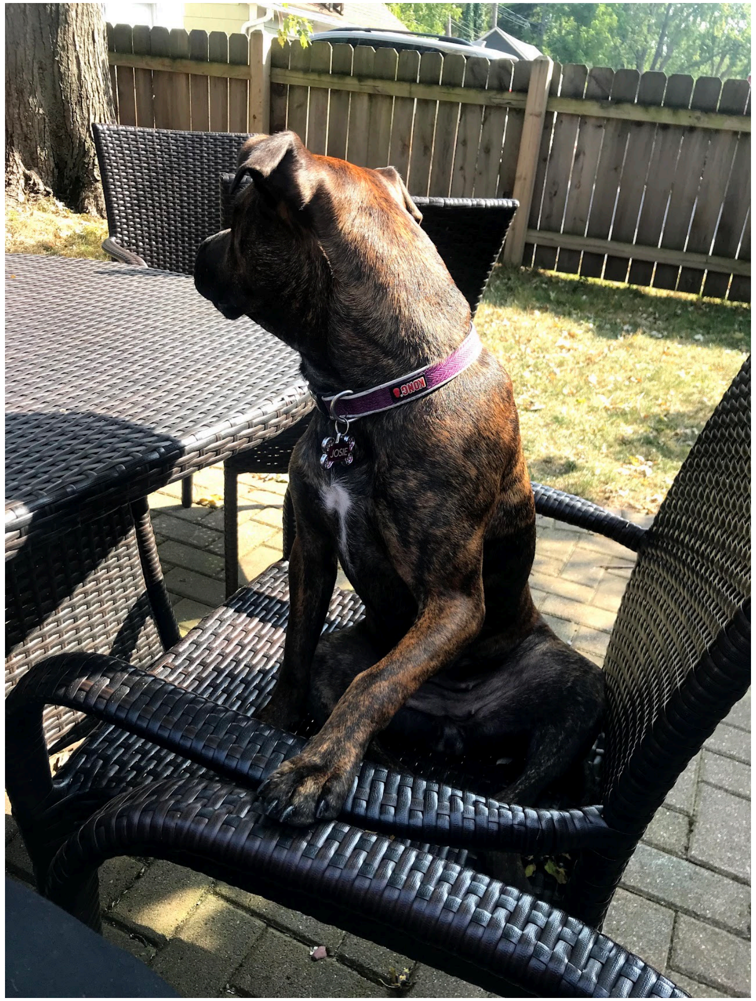


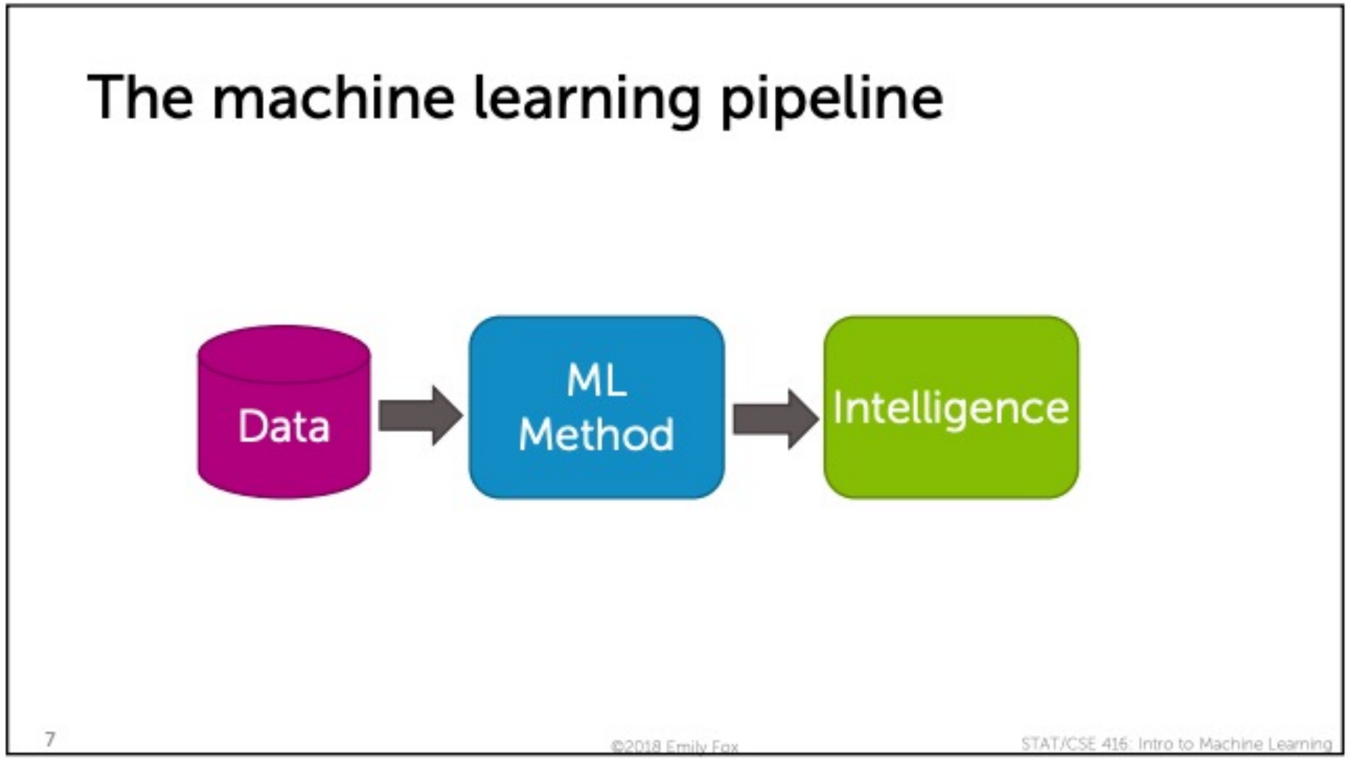
Figure 1.1: The general supervised approach to machine learning: a learning algorithm reads in training data and computes a learned function f. This function can then automatically label future text examples.

CIML, Chapter 1



Josie looks like she understand chairs. She does not. She cannot **generalize** beyond her training data.

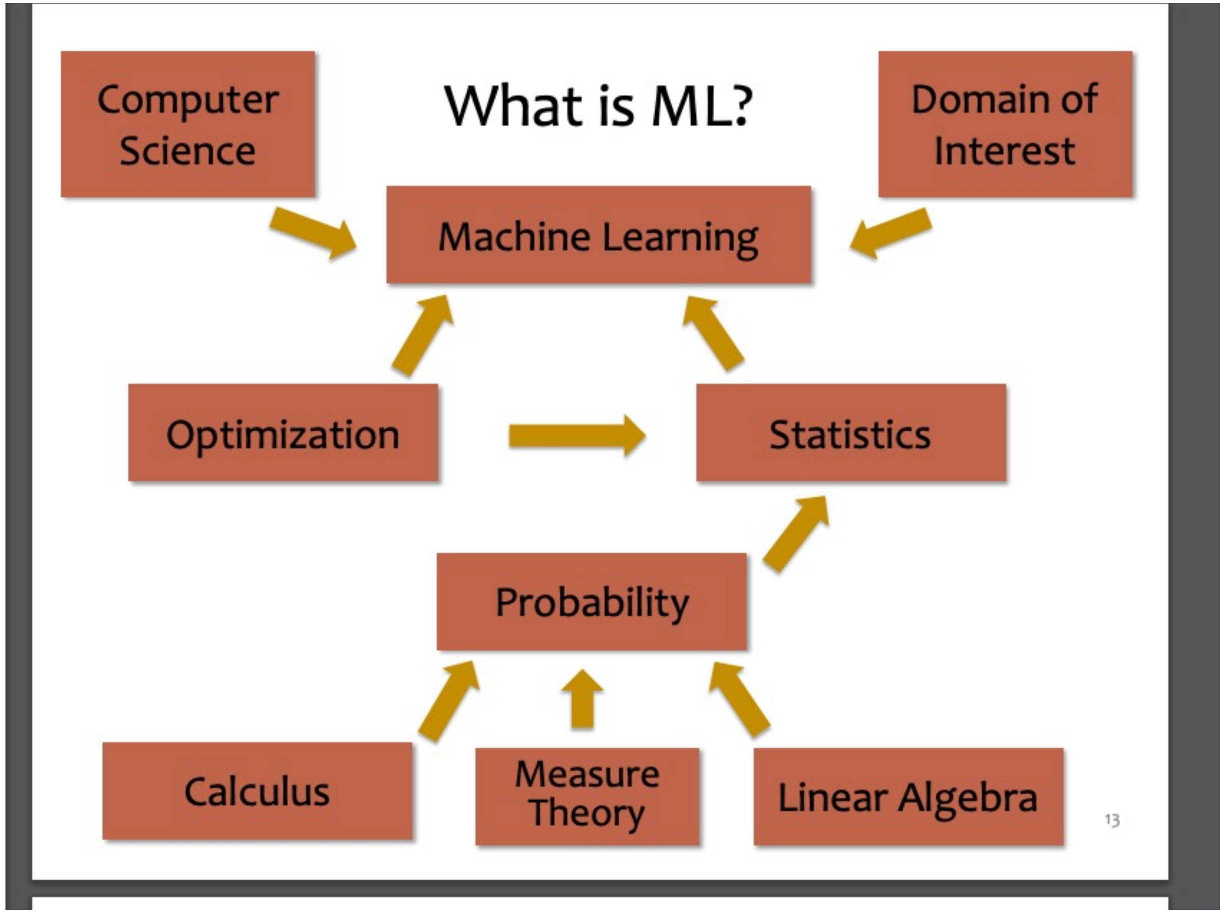
ML as the production of intelligence from data



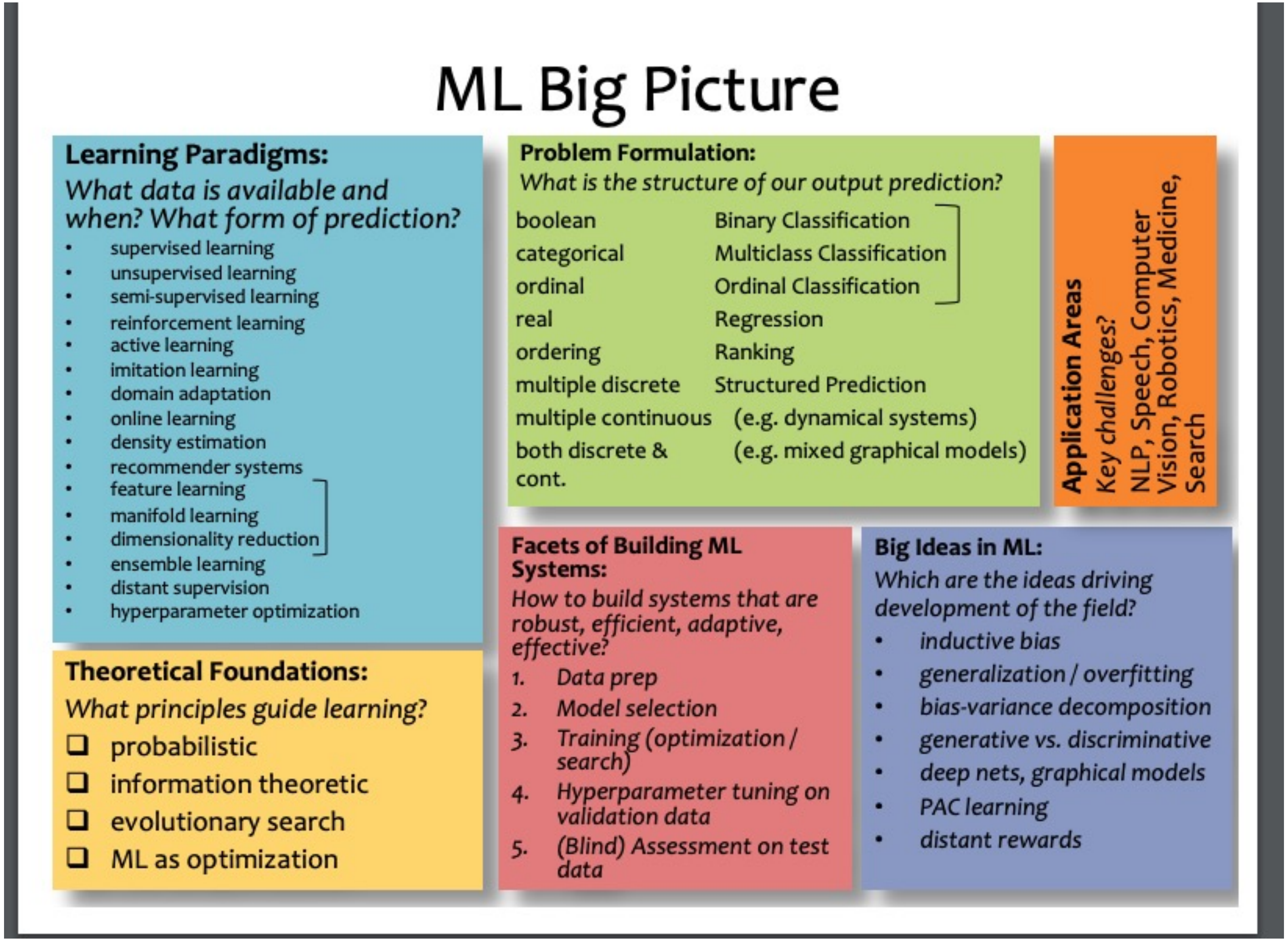
- I like this definition because it…
  - fits a number of different **learning paradigms**
  - is dead simple – we use ML to learn from data
- However, perhaps over-simplified… doesn't this match statistics too?

These images are taken from https://courses.cs.washington.edu/courses/cse416/18sp/lectures.html

ML as a composite of many things



These images are taken from http://www.cs.cmu.edu/~mgormley/courses/10601/schedule.html



ML is many things to many people.

These images are taken from
http://www.cs.cmu.edu/~mgormley/courses/10601/schedule.html

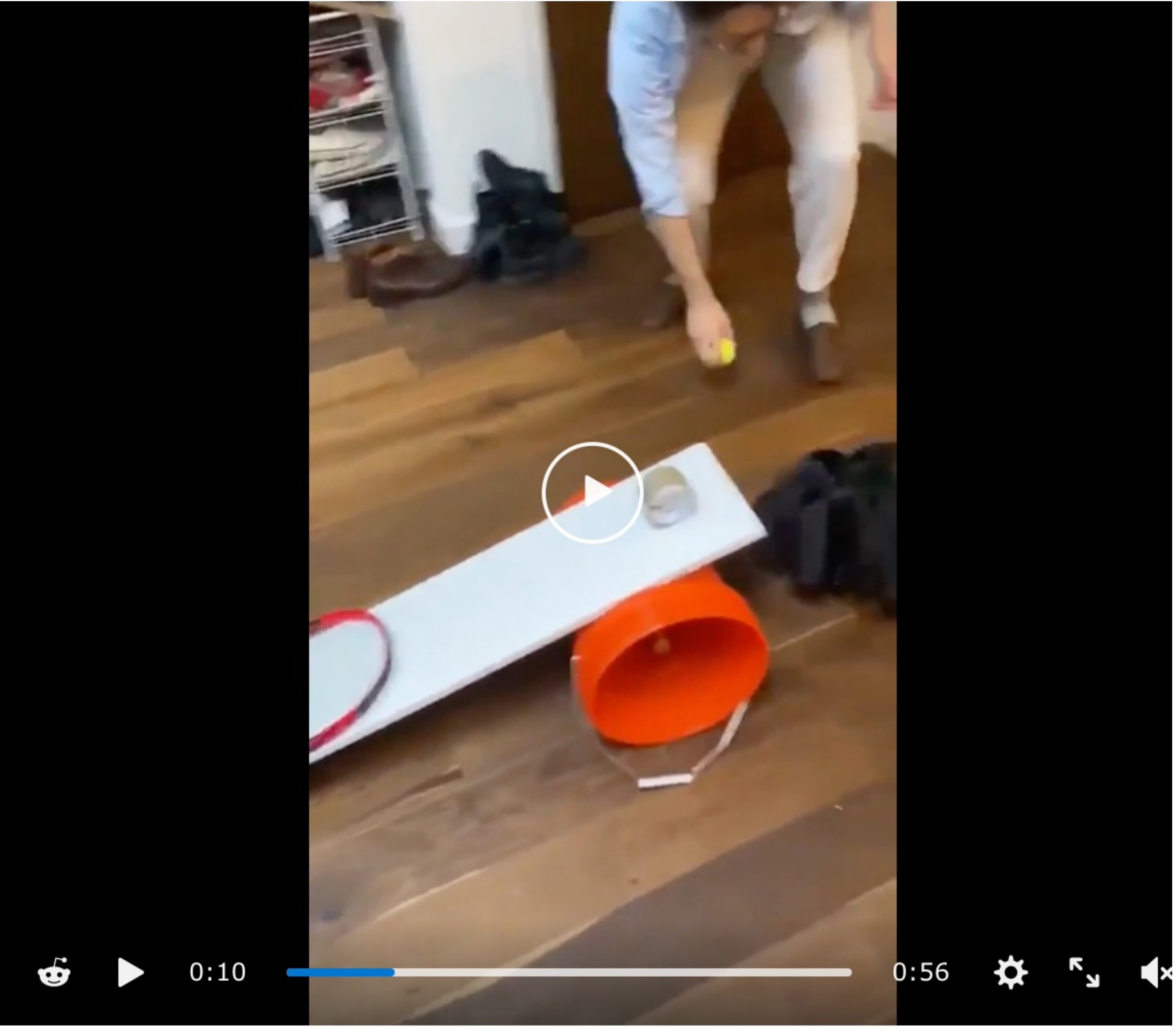
Corollary: There is more to ML than I can teach you in this class.

My aim is to give you the tools to understand and evaluate ML, even if I haven't taught you that specific math/method/paradigm/etc.

My focus will be on understanding the basics really well, and on giving you practical experience through the programming assignments

Few questions here

- What is machine learning?
- **(How) does it work?**
- Is it always amazing?



https://www.reddit.com/r/machinelearningmemes/comments/mqy9u5/machine_learning_pipelines/

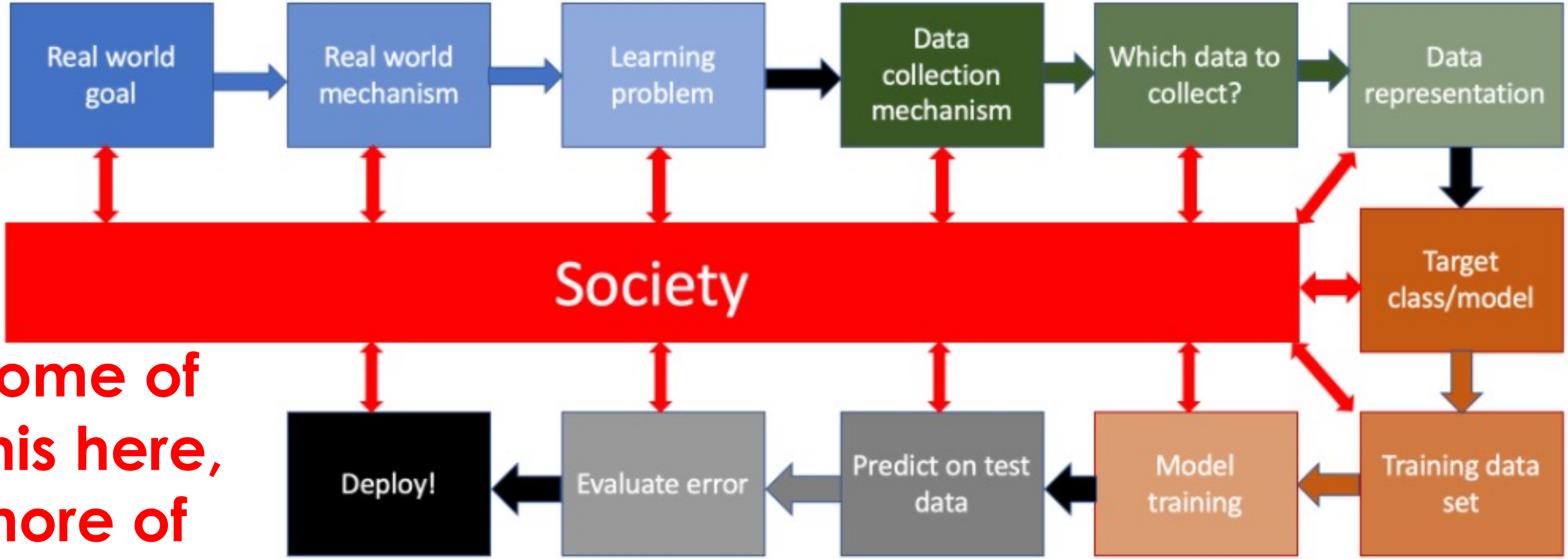
BREAK! (3 minutes)

Syllabus/Course Review

Few questions here

- What is machine learning?
- **(How) does it work?**
- Is it always amazing?

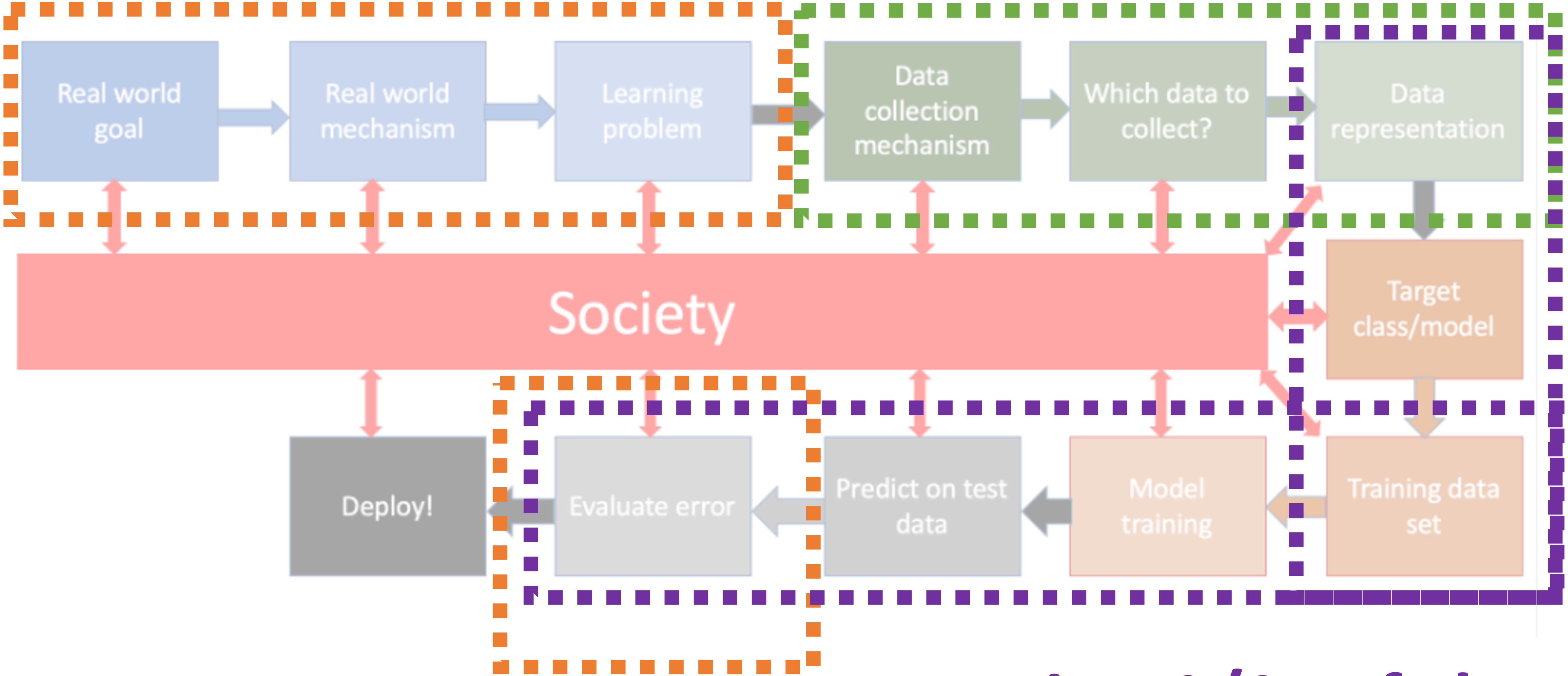
The ML Pipeline (one view)



Some of this here, more of this in 440/540

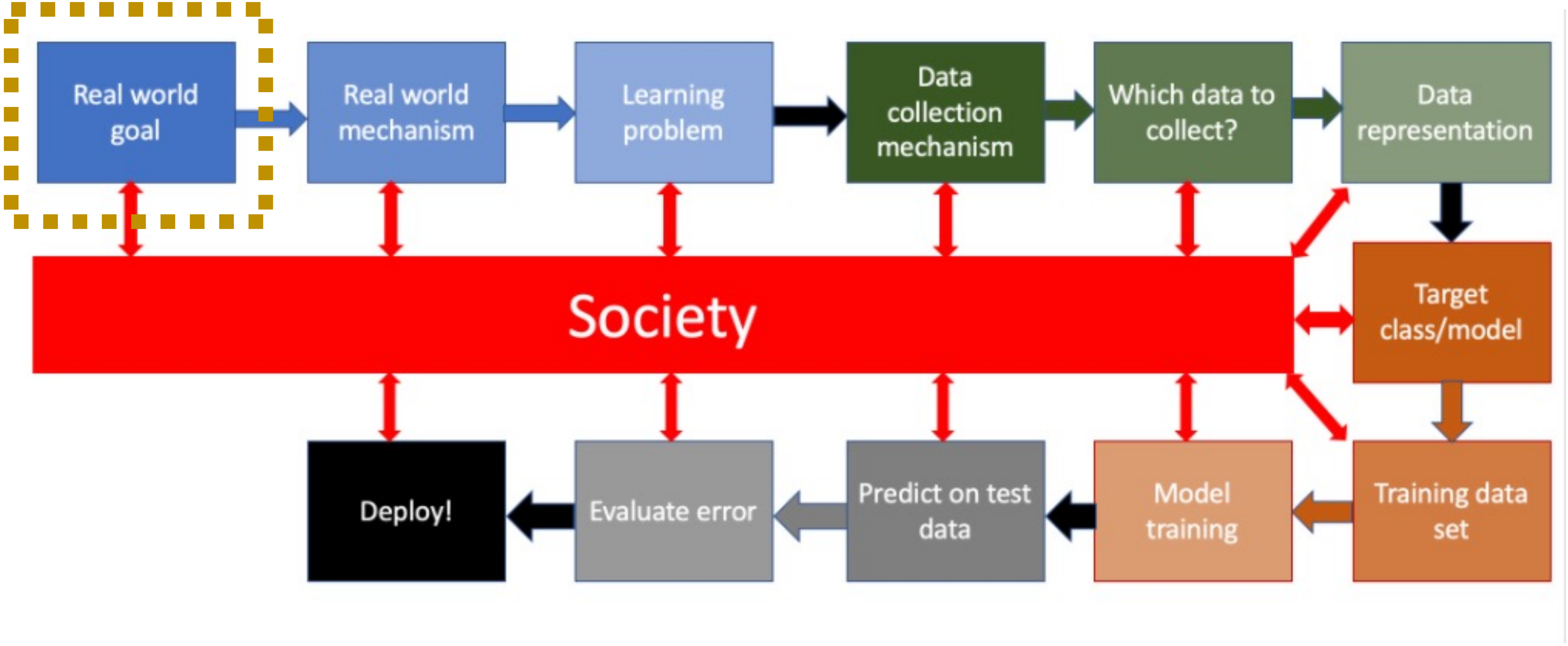
Last part of the course

PA1,2,3



First 2/3s of class

Case Study


Real world goal
Real world mechanism
Learning problem
Data collection mechanism
Which data to collect?
Data representation
Society
Target class/model
Training data set
Model training
Predict on test data
Evaluate error
Deploy!

Case Study 1 of many

Residential displacement occurs when individuals are forced, **involuntarily**, to leave their home.

We are the city of Buffalo. We want to reduce the impact of residential displacement on low income residents

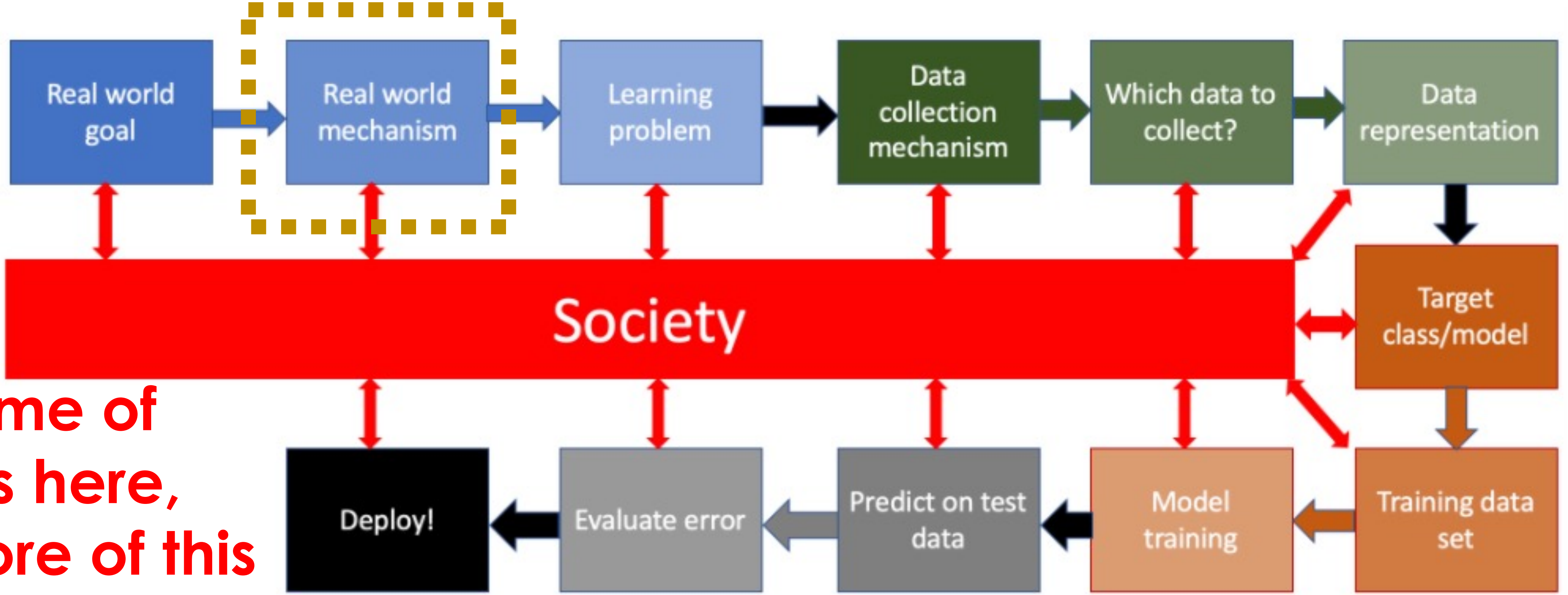
Potential ML questions - Supervised learning

- **Supervised learning** - can I predict **an outcome** from some **inputs**
- If the outcome is…
  - A category/set of discrete outcomes, this is ***classification***
    - *Given a picture of their brain, does this person have cancer, yes or no?*
  - A number, this is ***regression***
    - *How much will the price of this stock change tomorrow, given its prices over the last week?*

Why forecast residential displacement?

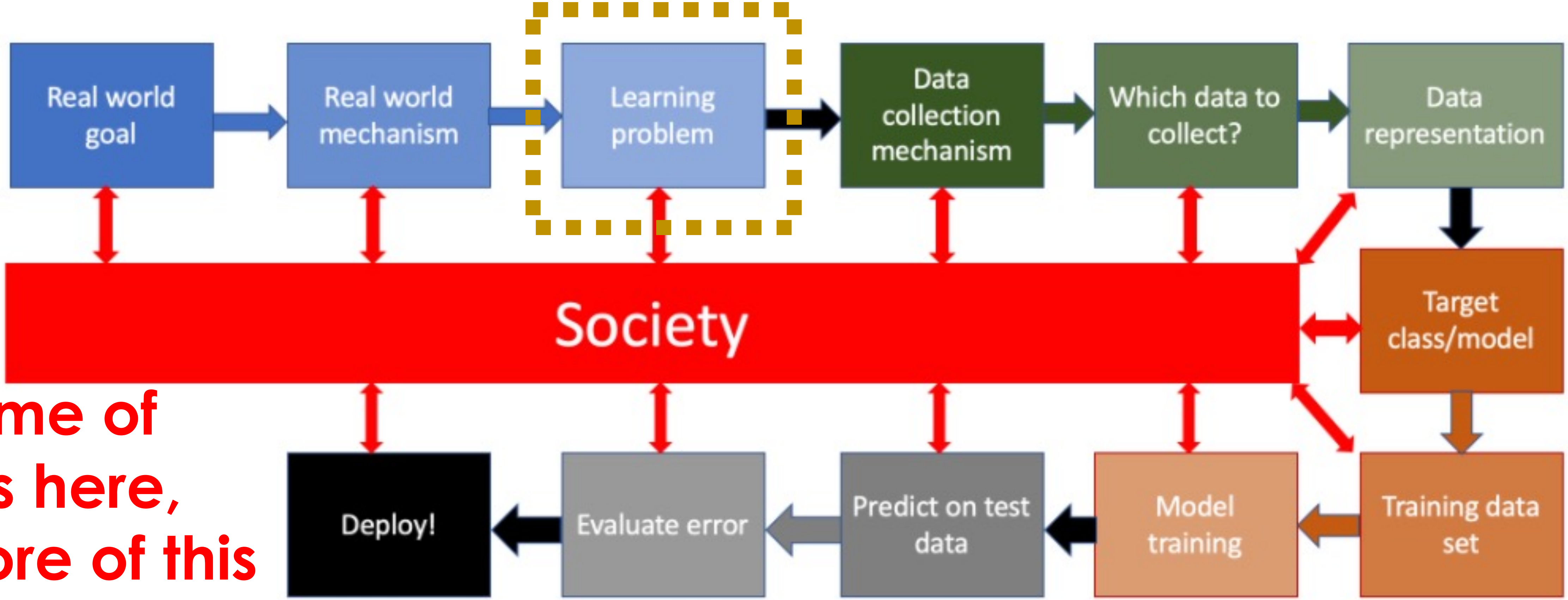
- Buffalo, New York.
  - From 2010 to 2016, one area in Buffalo (the **West Side**) saw a loss of **77%** of its Black population.

- Often due to gentrification
  - low-income regions attracts new residents and investment.
  - Rising home prices **drive out the poor, usually minorities.**



Some of this here, more of this in 440/540

We believe that if we can figure out **where** residential displacement is happening, then we can construct policy to try to mitigate it's impact

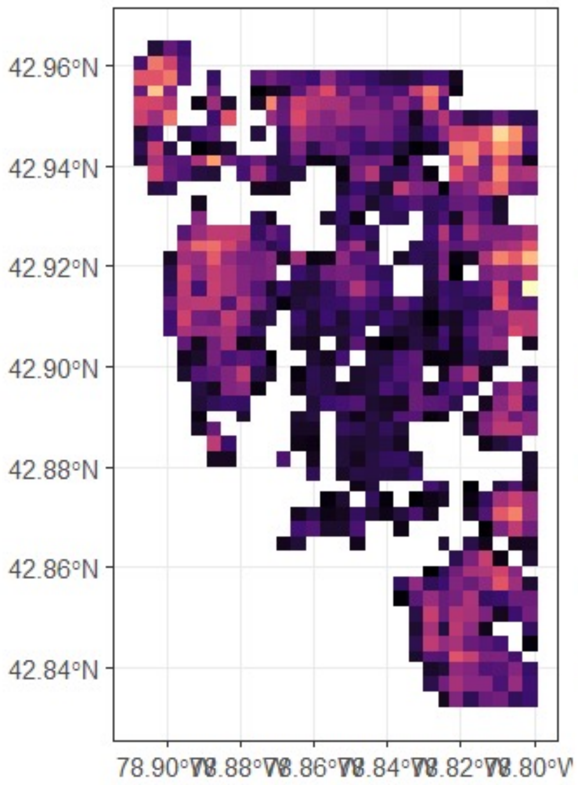


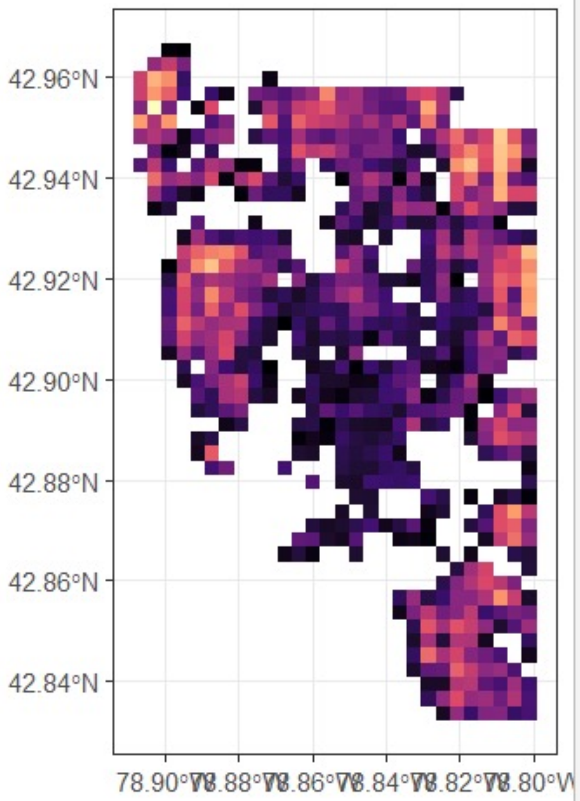
Some of this here, more of this in 440/540

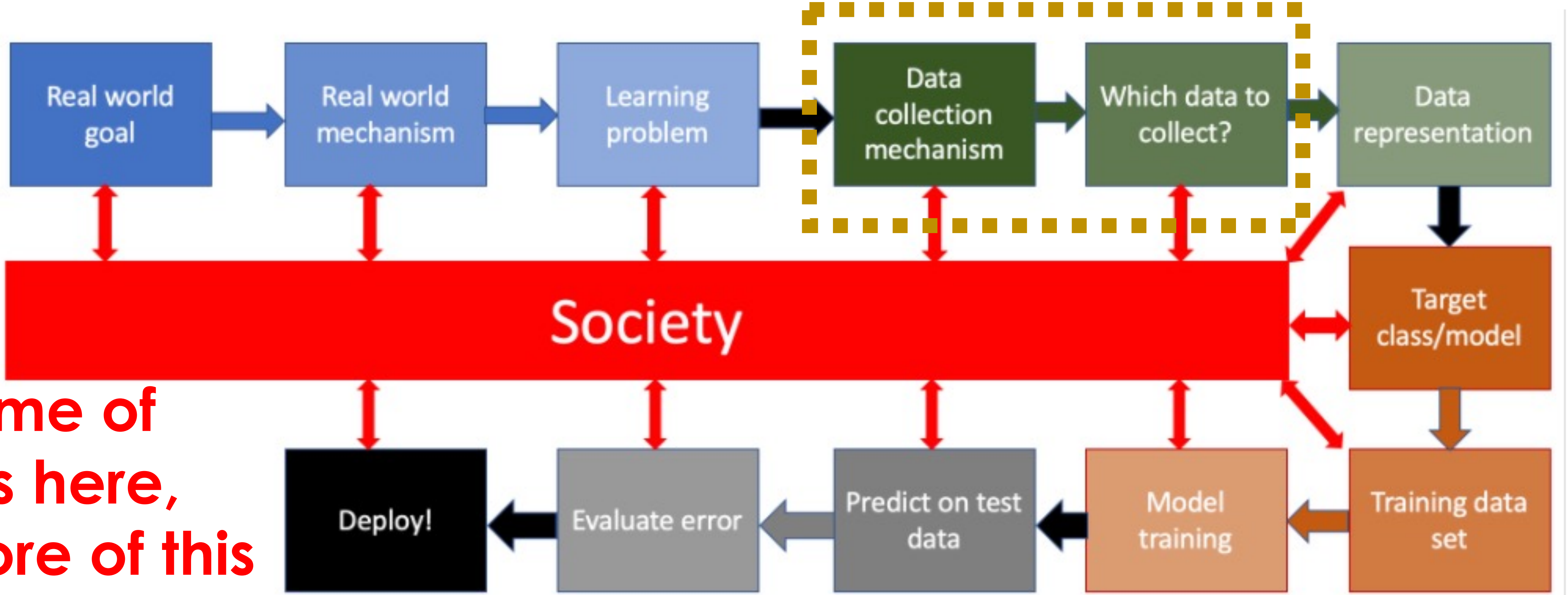
It is very hard to determine whether or not someone moved 1) **from an apartment** or 2) **involuntarily**.

So, we will try to predict **regions of the city where a lot of people have sold their homes.**

… is this specific enough? (No! Why not?)


42.96°N
42.94°N
42.92°N
42.90°N
42.88°N
42.86°N
42.84°N
78.90°W78.88°W78.86°W78.84°W78.82°W78.80°W


42.96°N
42.94°N
42.92°N
42.90°N
42.88°N
42.86°N
42.84°N
78.90°W78.88°W78.86°W78.84°W78.82°W78.80°W



Some of this here, more of this in 440/540

Zillow®

MLS

MULTIPLE LISTING SERVICE®

Real Property Parcel Search

If you're having trouble using this site, visit the page directly at h

Real Property Information

This site is a subset of the NYS Real Property System. It is composed of the most commonly referenced information.

All information is for public use.

Information as of: **July, 2021**

Erie County does not assume any liability associated with the use or misuse of this data.

Search Applications Requirements

o Microsoft Internet Explorer or Microsoft Edge (Chrome, Firefox, and Safari are not supported)

Enter ONE of the fields below to begin your search:

Owner Name {Last Name First} or

Property Address No./ Street

S-B-L

Search All | Submit Query | Clear Form

Real Property Information

Parcel Status	ACTIVE	**City\Town**	Buffalo	**Village**	
S-B-L	110.68-4-11	**Owner**	EICHEL JOHN R	**SWIS**	140200
Property Location	29 OJIBWA CIR	**Mailing Address**			
Property Class	210 1 FAMILY RES	**Line 2**			
Assessment	810000	**Line 3**			
Taxable	810000	**Street**	29 OJIBWA CIR		
Desc	269.32 E WATERFRONT CIR	**City/State**	BUFFALO NY		
Desc		**Zip**	14202		
Deed Book	11332	**Deed Page**	7562		
Frontage	49.62	**Depth**	87.75	**Acres**	0
Year Built	2016	**Square Ft**	2910		
Beds	3	**Baths**	2.5		
FirePlace	2	**School**	BUFFALO SCHOOL DIST		
Owner History	Tax Payment History				

Google maps
Click Here

Real Property Information

Parcel 140200 110.680-4-11.000					29 OJIBWA CIR		EICHEL JOHN R							
Year	Cycle/Desc	Instl	InRem	Bill No.	Principal	Adjust	Credit	Fee	Pd Intr	Paid	Balance	Interest	Total Bal	Payer Information
2022	1-County Tax	01		56772	4,072.90	0.00	0.00	0.00	0.00	0.00	4,072.90	0.00	4,072.90	n/a
2021	1-County Tax	01		56797	4,082.86	0.00	0.00	0.00	0.00	4,082.86	0.00	0.00	0.00	dovenmuehle mortgage, in
2020	1-County Tax	01		56792	3,848.38	0.00	0.00	0.00	0.00	3,848.38	0.00	0.00	0.00	dovenmuehle mortgage, in
2019	1-County Tax	01		56827	3,665.35	0.00	0.00	0.00	0.00	3,665.35	0.00	0.00	0.00	dovenmuehle mortgage, in
2018	1-County Tax	01		57101	3,567.09	0.00	0.00	0.00	0.00	3,567.09	0.00	0.00	0.00	whaley douglass g
2017	1-County Tax	01		57240	175.61	0.00	0.00	0.00	0.00	175.61	0.00	0.00	0.00	whaley douglass g
2016	1-County Tax	01		57383	161.81	0.00	0.00	0.00	0.00	161.81	0.00	0.00	0.00	1094 group llc
2015	1-County Tax	01		57499	148.28	0.00	2.19	0.00	2.19	148.28	0.00	0.00	0.00	n/a
2014	1-County Tax	01		57599	146.72	0.00	0.00	0.00	0.00	146.72	0.00	0.00	0.00	1094 group llc
2013	1-County Tax	01		57901	148.18	0.00	0.00	0.00	0.00	148.18	0.00	0.00	0.00	1094 group llc
2012	1-County Tax	01		58061	149.24	0.00	0.00	0.00	0.00	149.24	0.00	0.00	0.00	1094 group llc
2011	1-County Tax	01		57944	145.97	0.00	0.00	0.00	0.00	145.97	0.00	0.00	0.00	1094 group llc
2010	1-County Tax	01		271	148.96	0.00	0.00	0.00	0.00	148.96	0.00	0.00	0.00	1094 group llc
2009	1-County Tax	01		271	142.67	0.00	0.00	0.00	0.00	142.67	0.00	0.00	0.00	1094 group llc
	Balance as of										4,072.90	0.00	4,072.90	

Go To Tax Inquiry/Make Payment Online!

*Erie County is not responsible for clerical errors, omissions, or errors caused by taxpayers paying on the wrong property.
Notify your City or Town Assessor for any error in name, property description and change of tax bill address.

Real Property Information

Owner	1094 GROUP LLC	**Book-Page/Date**	11127-0599 * 3/26/2007 *
Owner	WHALEY DOUGLASS G/STEPHANIE	**Book-Page/Date**	11298-2183 * 6/30/2016 *
Owner	EICHEL JOHN R	**Book-Page/Date**	11332-7562 * 8/3/2018 *

Verified as of 01/20/2022

Search | Results | Document

Party | Document Type | Instrument Number | Book/Page | Map Number

Book / Page Type: ALL BOOKS

Book / Page: 11298 / 2183 [Clear] [Search]

Numbers (e.g., 12345/1234)

Limit search to maximum records [ ] Rows

Display records per page [ ] Rows per page

Help?

Search | Results | Document

Click Search Tab to initiate new search.

Print Results

* Designates From Party   Records with Status Other than V or B = Not Verified Record

Records with Flag: R = Replaced Record, C = Correction Record   Click View to See Document Details

		Name	Cross Party	Date	Type	Book	Page	Legal	Instr#
View	*	1094 GROUP LLC	WHALEY DOUGLASS G	06/30/2016	DEED	11298	2183	CTY 14 S9 C3353	20161353
View		WHALEY DOUGLASS G	1094 GROUP LLC	06/30/2016	DEED	11298	2183	CTY 14 S9 C3353	20161353
View		WHALEY STEPHANIE	1094 GROUP LLC	06/30/2016	DEED	11298	2183	CTY 14 S9 C3353	20161353

OPENDATA BUFFALO

Home About Us

2019-2020 Assessment Roll

Government

This dataset contains information pertaining to the assessed value of properties within the City of Buffalo. Accurate property inventory/information is vital to an equitable assessment and the City of Buffalo is committed to providing fair and equitable assessments in each annual Assessment Roll so that property owners pay only their fair share of property taxes.

Featured Content Using this Data

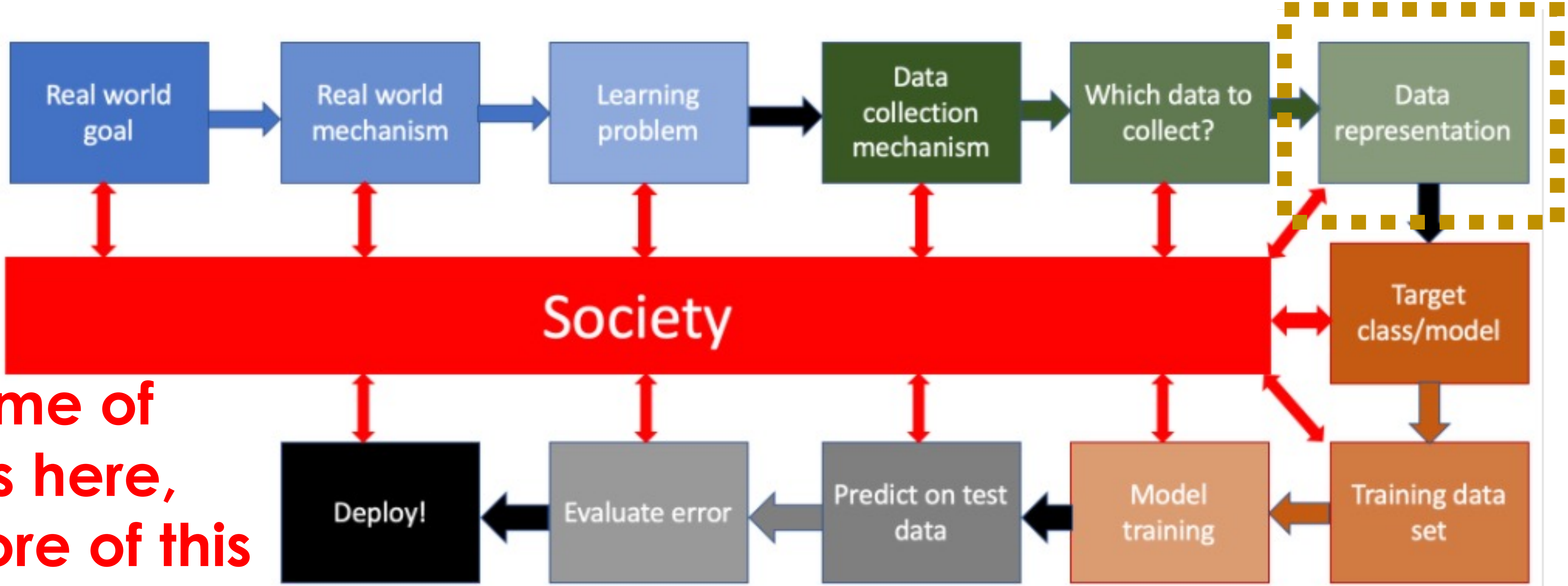
City of Buffalo Online Assessment Roll System (O...

External Content

Need to know important dates and information about your property's assessment? Click here now!

To come back to

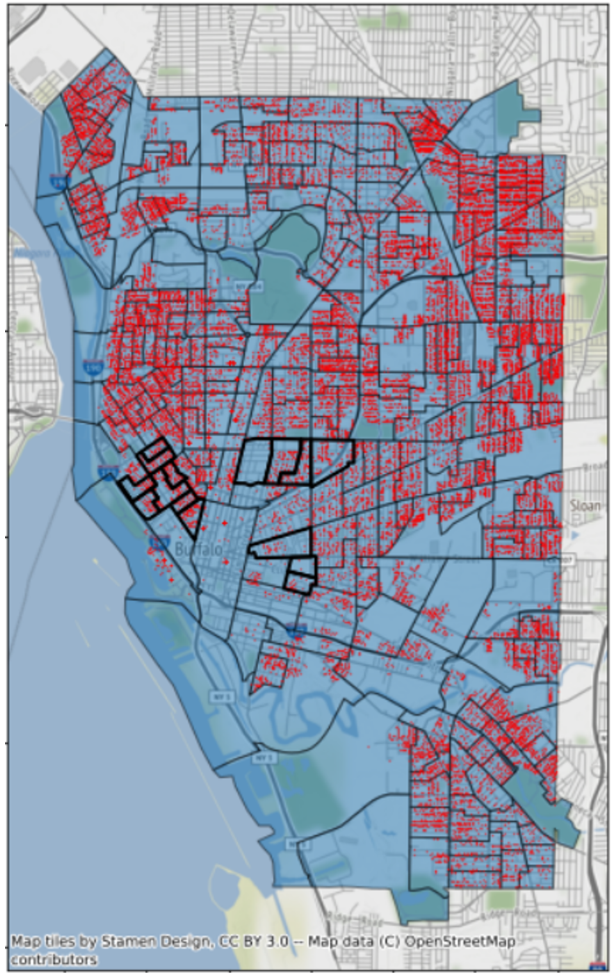
- Yikes, right?
- This semester we're gonna talk about some of this stuff
  - Should I really be using this data?
  - Does the world really need this model?
  - Is this model serving **everyone** and not just the "majority class"?



Some of this here, more of this in 440/540

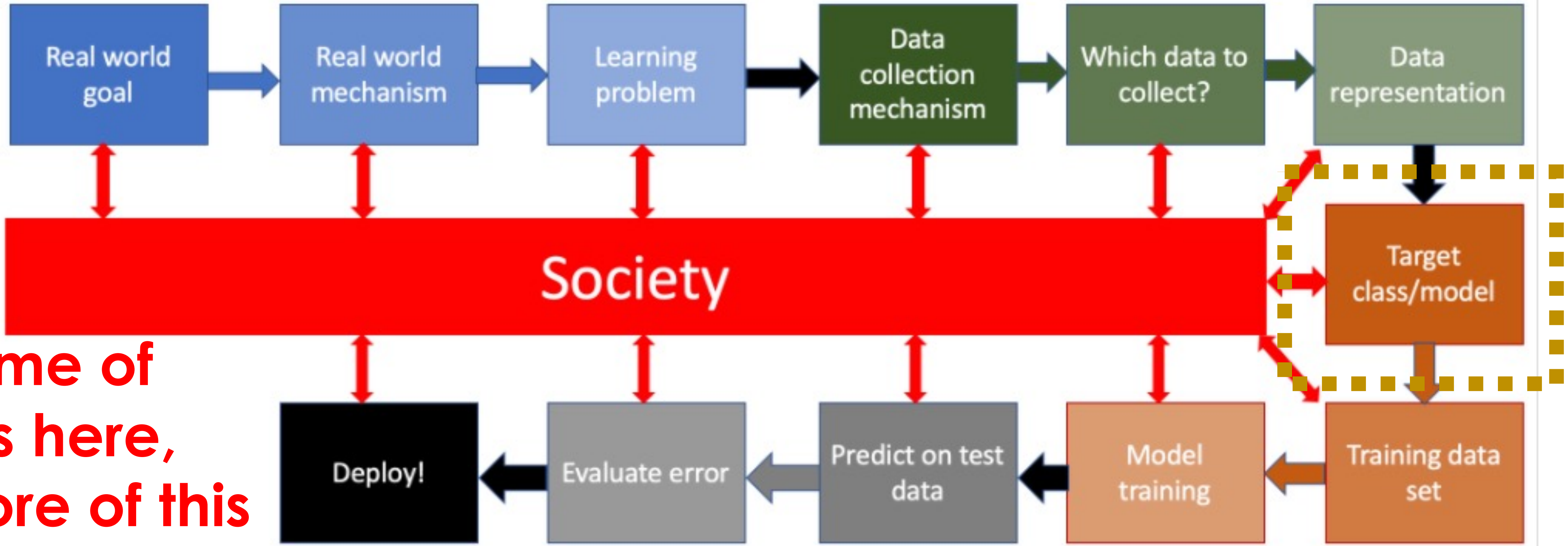
Data

- Comprises all house purchases and tax information in Buffalo from 1995 to the present
- Contains **116,438** transactions on all **51,425** homes in the city as of June, 2020



BREAK! (3 minutes)

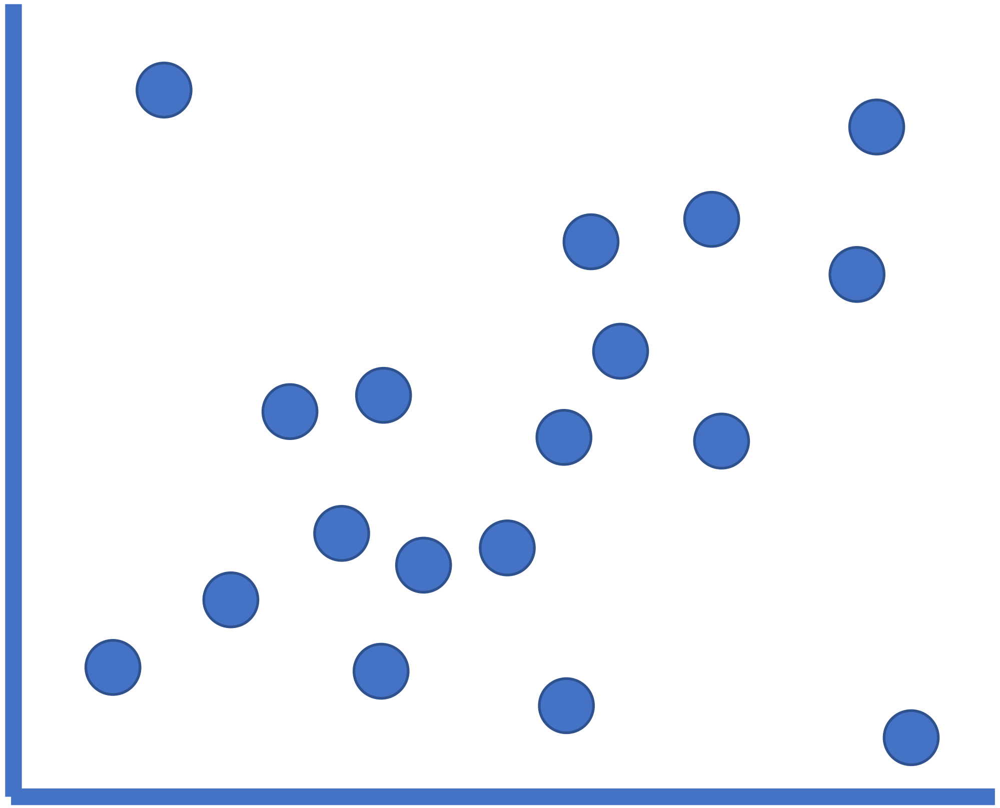
Kahoot Test



Some of this here, more of this in 440/540

Selecting a model (simplified)

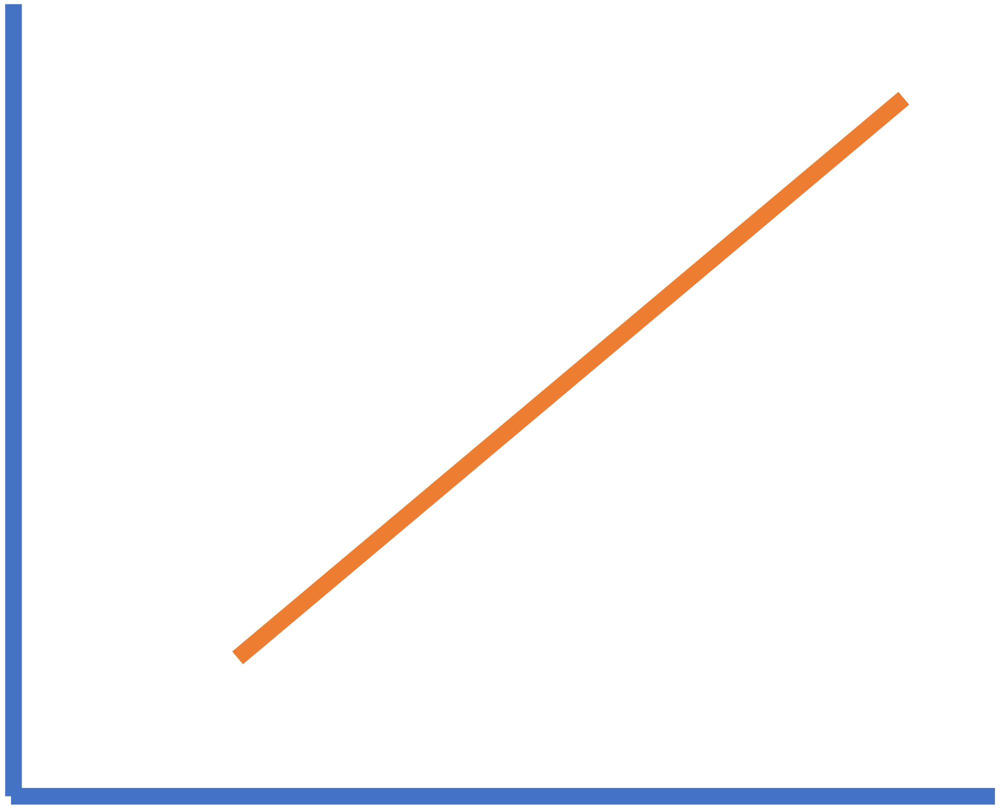
Outcome: Number of homes bought/sold this year



Feature: Number of homes bought/sold last year

Example – Supervised linear regression

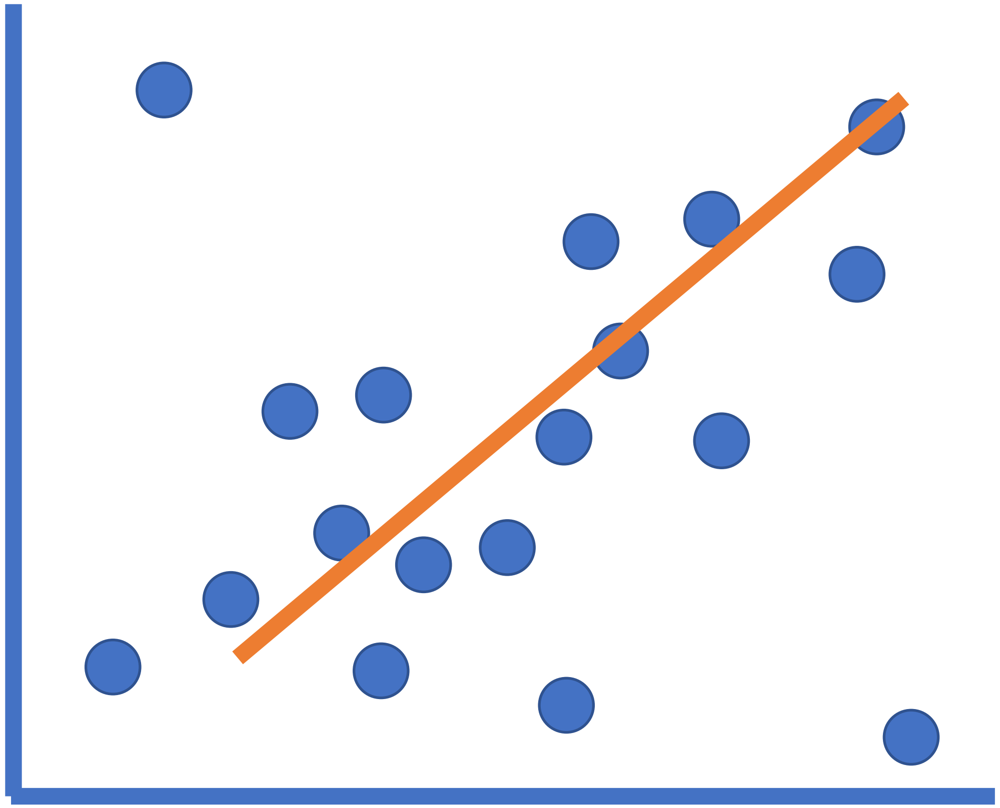
Outcome: Number of homes bought/sold this year



Feature: Number of homes bought/sold last year

Selecting a model (simplified)

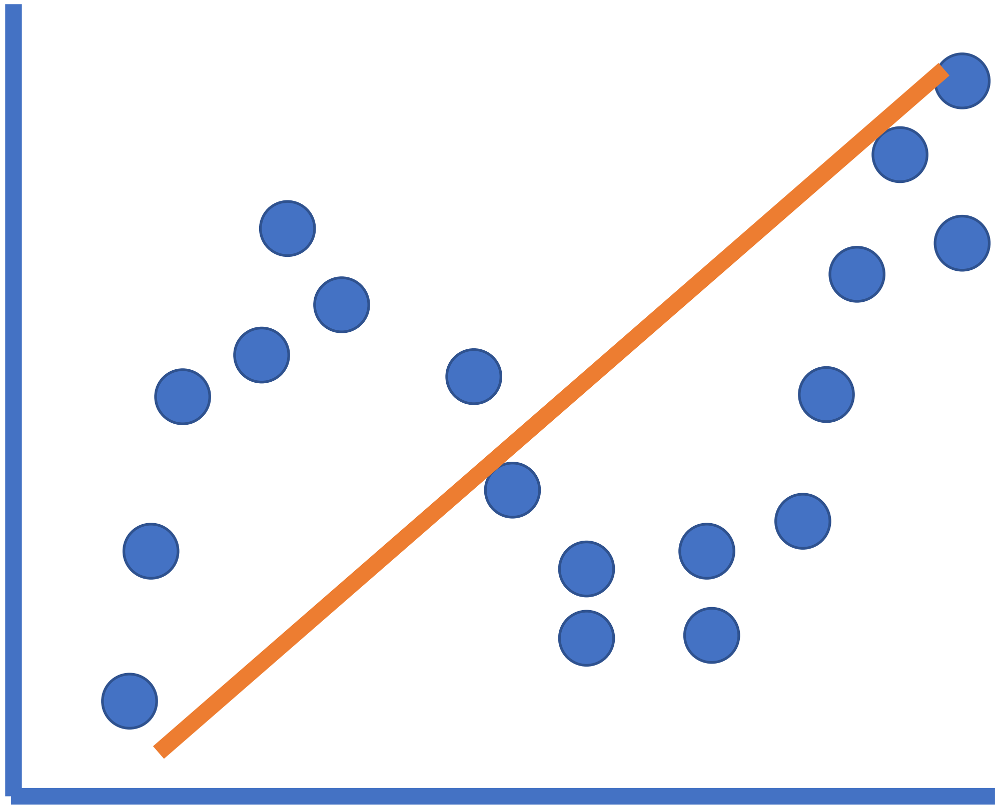
Outcome: Number of homes bought/sold this year



Feature: Number of homes bought/sold last year

Is a line the only way?

Outcome: Number of homes bought/sold this year

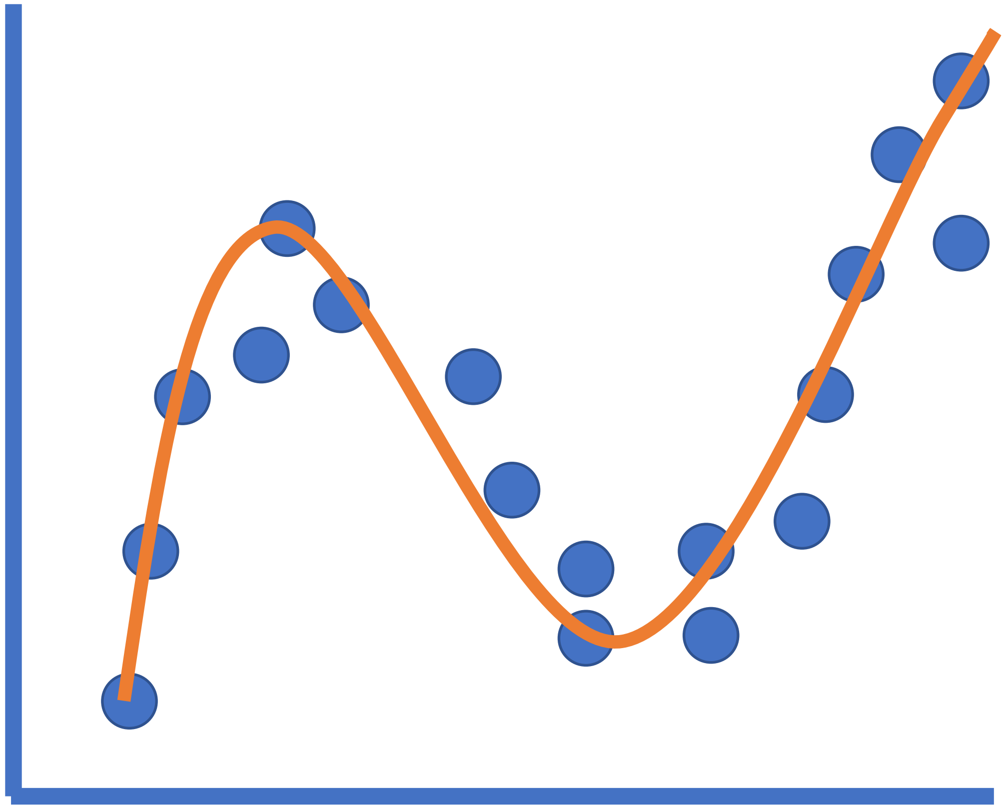


Feature: Number of homes bought/sold last year

Can we build a better model?

Is a line the only way? No!

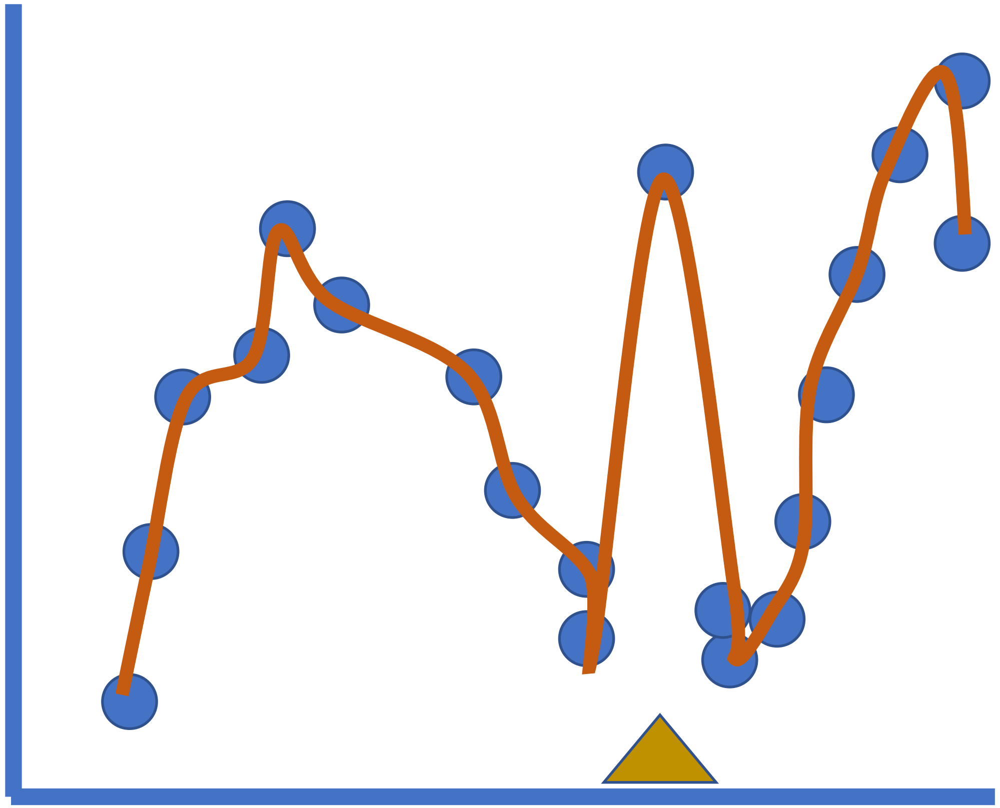
Outcome: Number of homes bought/sold this year



Feature: Number of homes bought/sold last year

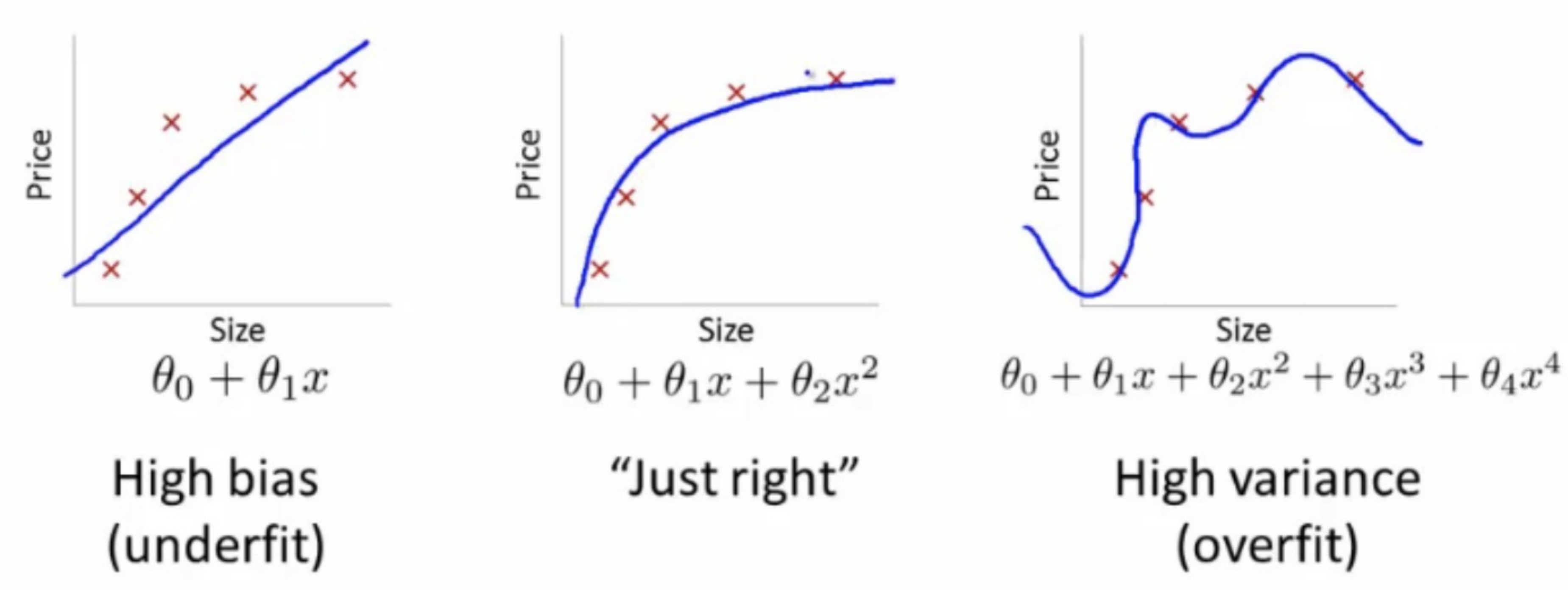
Is a line the only way? No! But careful of overfitting!

Outcome: Number of homes bought/sold this year



Feature: Number of homes bought/sold last year

Bias/Variance Tradeoff



Source: Coursera's ML course, taught by Andrew Ng

AdaBoost Classification Trees	adaboost	C
AdaBoost.M1	AdaBoost.M1	C
Adaptive Mixture Disc		
Adaptive-Network-Ba Inference System		
Adjacent Categories P for Ordinal Data		
Bagged AdaBoost		
Bagged CART		
Bagged FDA using gC		
Bagged Flexible Discr		
Bagged Logic Regres		
Bagged MARS		
Bagged MARS using		
Bagged Model		
Bayesian Additive Re		

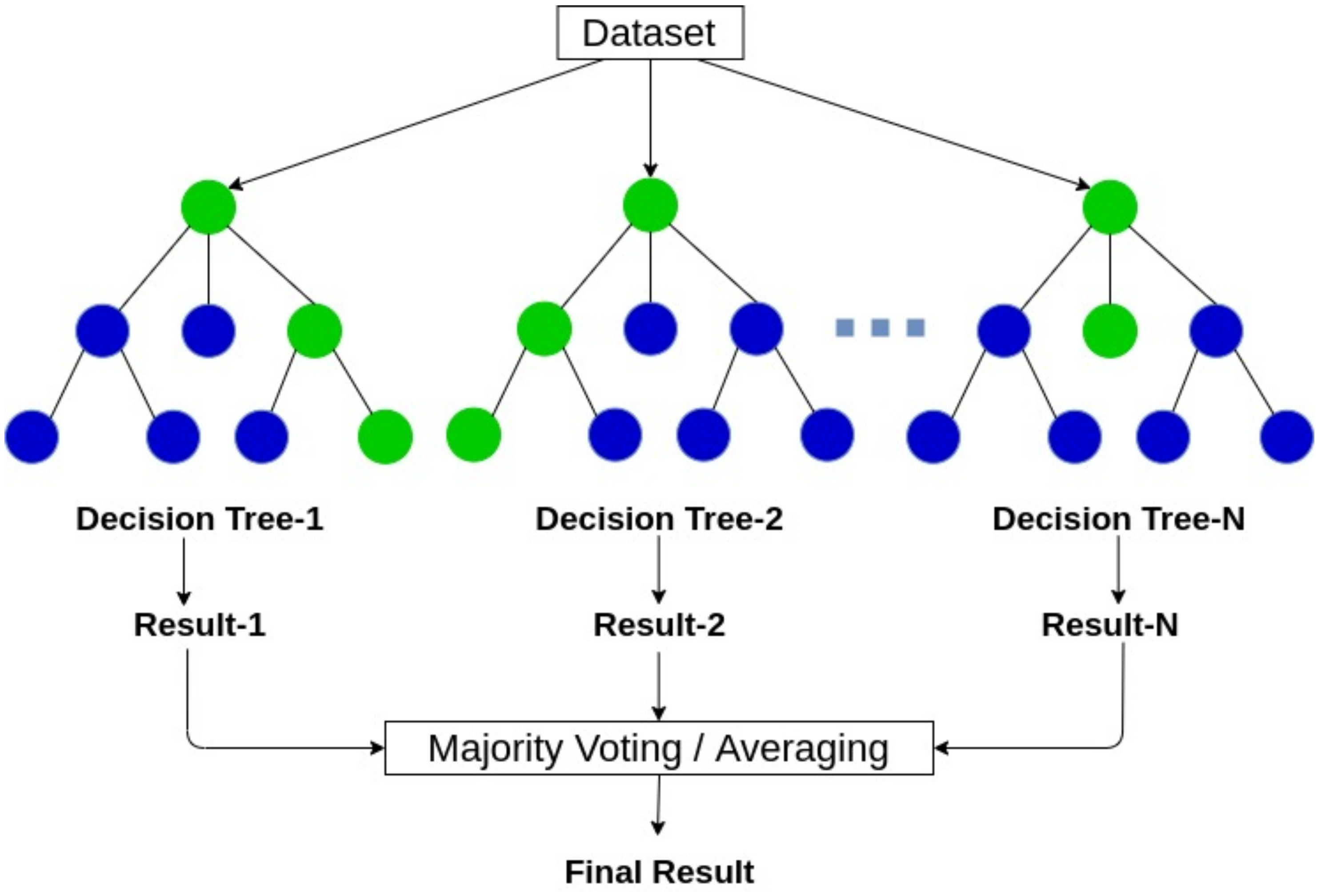
Bayesian Ridge Regression	bridge
Bayesian Ridge Regression (Model Averaged)	blassoA

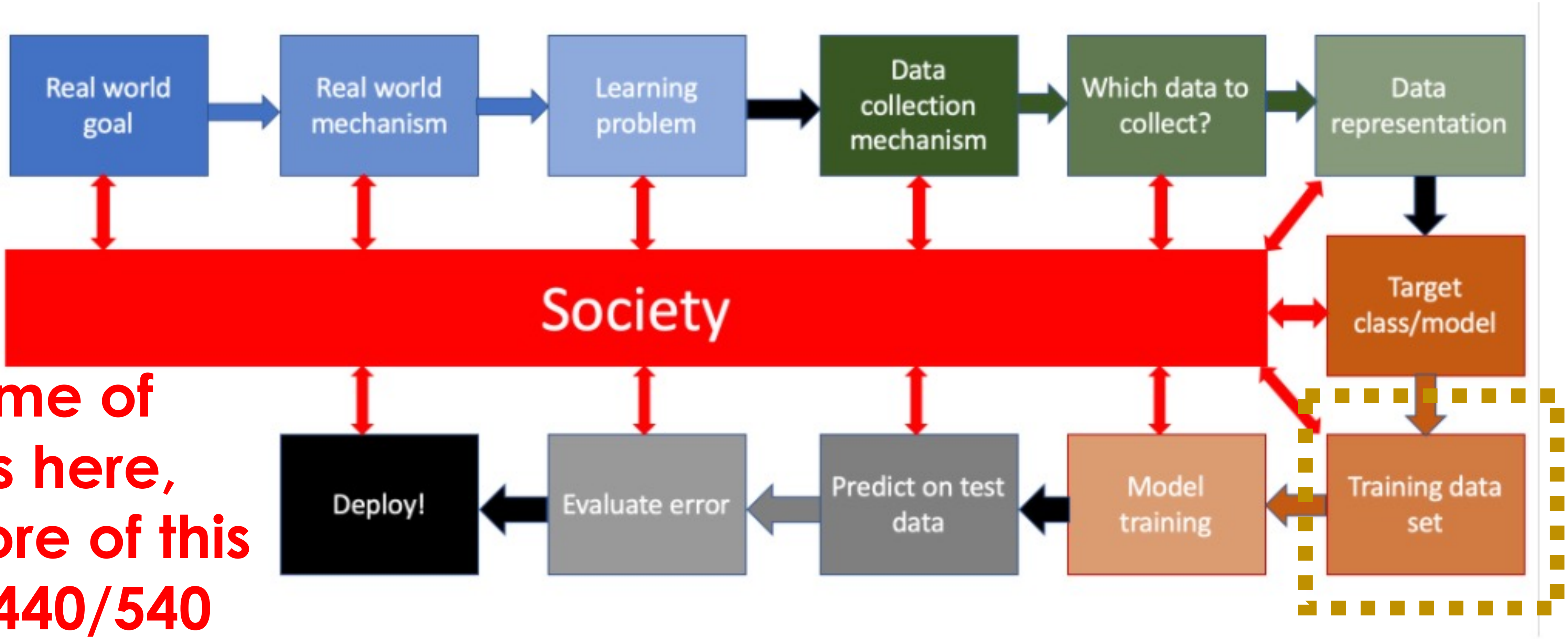
High Dimensional Discriminant Analysis	hdda	Classification	HDclassif
High-Dimensional Regularized Discriminant Analysis	hdrda	Classification	sparsediscrim
Hybrid Neural Fuzzy Inference System	HYFIS	Regression	frbs
Independent Component Regression	icr	Regression	fastICA
k-Nearest Neighbors	kknn	Classification, Regression	kknn
k-Nearest Neighbors	knn	Classification, Regression	
Knn regression via sklearn.neighbors.KNeighborsRegressor	pythonKnnReg	Regression	rPython
L2 Regularized Linear Support Vector Machines with Class Weights	svmLinearWeights2	Classification	LiblineaR
L2 Regularized Support Vector Machine (dual) with Linear Kernel	svmLinear3	Classification, Regression	LiblineaR
Learning Vector Quantization	lvq	Classification	class
Least Angle Regression	lars	Regression	lars
Least Angle Regression	lars2	Regression	lars
Least Squares Support Vector Machine	lssvmLinear	Classification	kernlab
Least Squares Support Vector Machine with Polynomial Kernel	lssvmPoly	Classification	kernlab

Extreme Learning Machine	elm	Classification, Regression	elmNN
Factor-Based Linear Discriminant Analysis	RFlda	Classification	HiDimDA
Flexible Discriminant Analysis	fda	Classification	earth, mda
Fuzzy Inference Rules by Descent Method			
Fuzzy Rules U			
Fuzzy Rules U Competitive L			
Fuzzy Rules U Competitive L			
Fuzzy Rules U Learning Algo Environment			
Fuzzy Rules v			
Fuzzy Rules v			
Fuzzy Rules w			
Gaussian Pro			ab
Gaussian Pro Kernel			ab
Gaussian Pro Function Kern			ab
Generalized A LOESS			

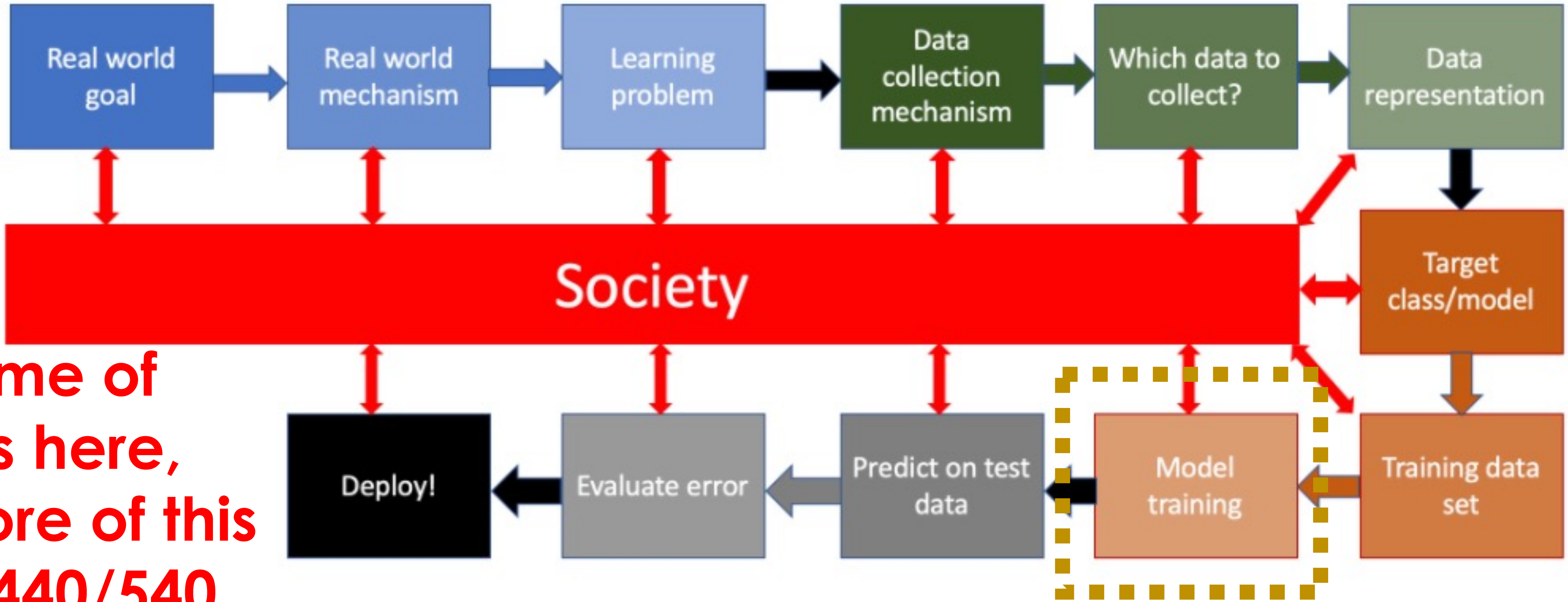
Model Averaged Neural Network	avNNet	Classification, Regression	nnet
Model Rules	M5Rules	Regression	RWeka
Model Tree	M5	Regression	RWeka
Monotone Multi-Layer Perceptron Neural Network	monmlp	Classification, Regression	monmlp
Multi-Layer Perceptron	mlp	Classification, Regression	RSNNS
Multi-Layer Perceptron	mlpWeightDecay	Classification, Regression	RSNNS
Multi-Layer Perceptron, multiple layers	mlpWeightDecayML	Classification, Regression	RSNNS
Multilayer Perceptron Network by Stochastic Gradient Descent	mlpSGD	Classification, Regression	FCNN4R, plyr
Multilayer Perceptron Network with Dropout	mlpKerasDropout	Classification, Regression	keras
Multilayer Perceptron Network with Dropout	mlpKerasDropoutCost	Classification	keras
Multilayer Perceptron Network with Weight Decay	mlpKerasDecay	Classification, Regression	keras
Multilayer Perceptron Network with Weight Decay	mlpKerasDecayCost	Classification	keras
Multi-Layer Perceptron, with multiple layers	mlpML	Classification, Regression	RSNNS
Multi-Step Adaptive MCP-Net	msaenet	Classification, Regression	msaenet

Actual Model





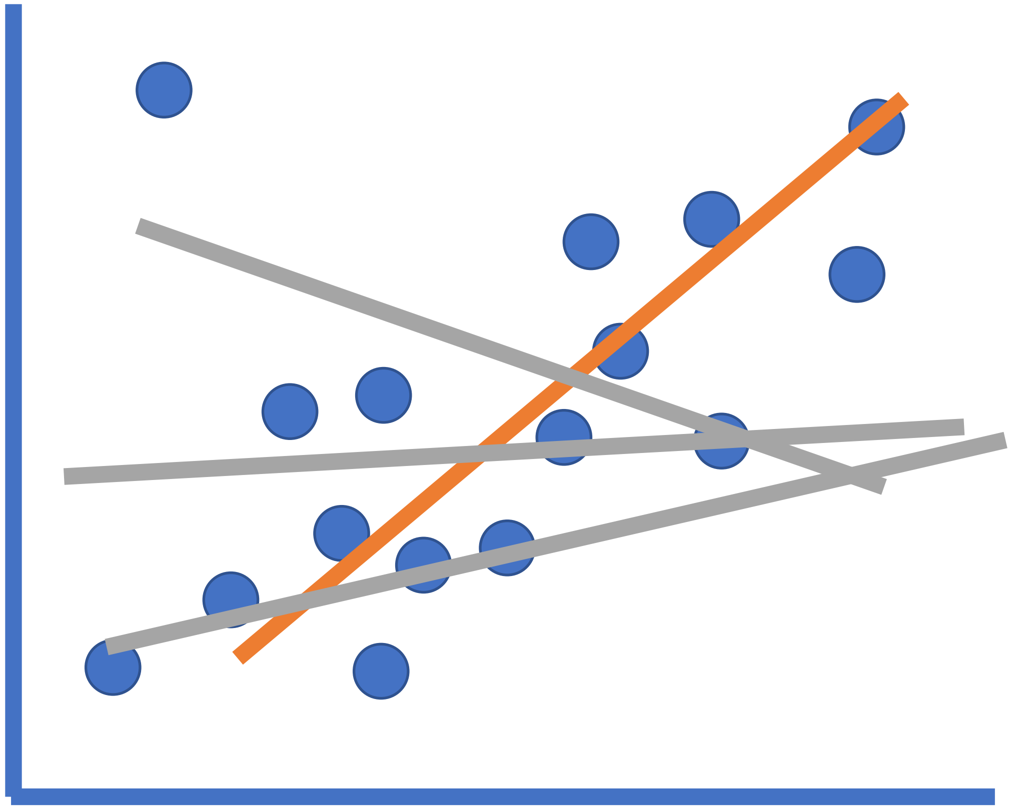
Some of this here, more of this in 440/540



Some of this here, more of this in 440/540

How do we train a model?

Outcome: Number of homes bought/sold this year

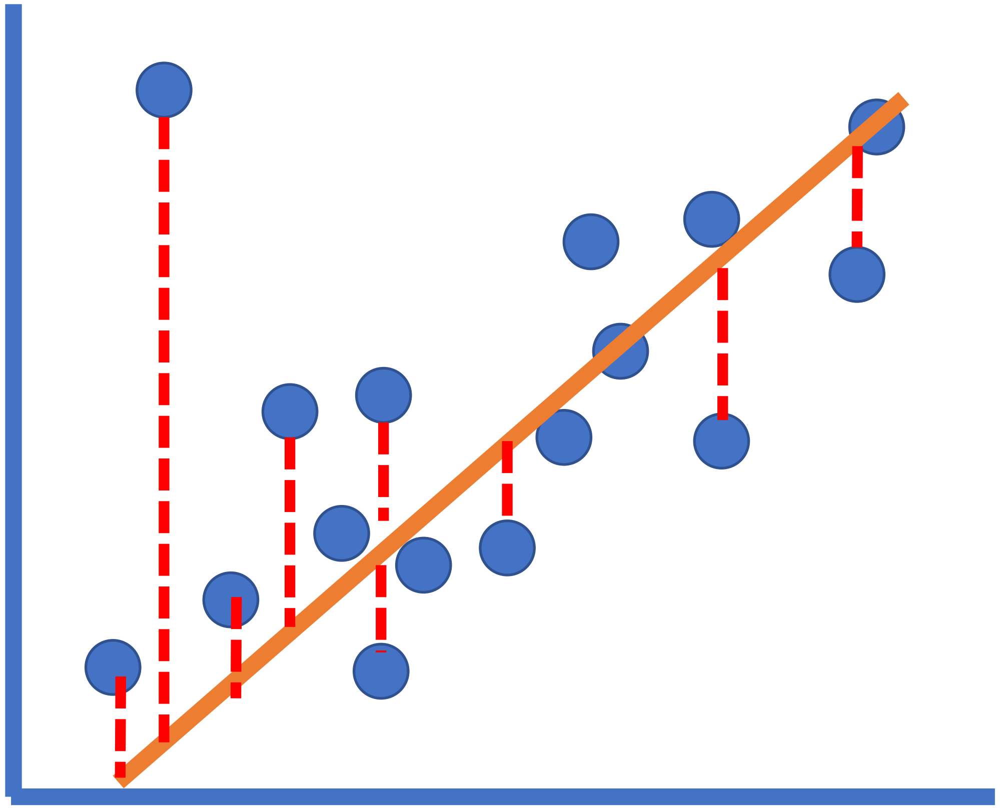


Feature: Number of homes bought/sold last year

That is, how do we find the "best" line?

How do we train a model?

Number of crimes tomorrow



Number of arrests made today

Find the one that minimizes some **objective function**

SSE – sum of squared errors

$$SS_{Total} = \sum (y_i - \bar{y})^2$$

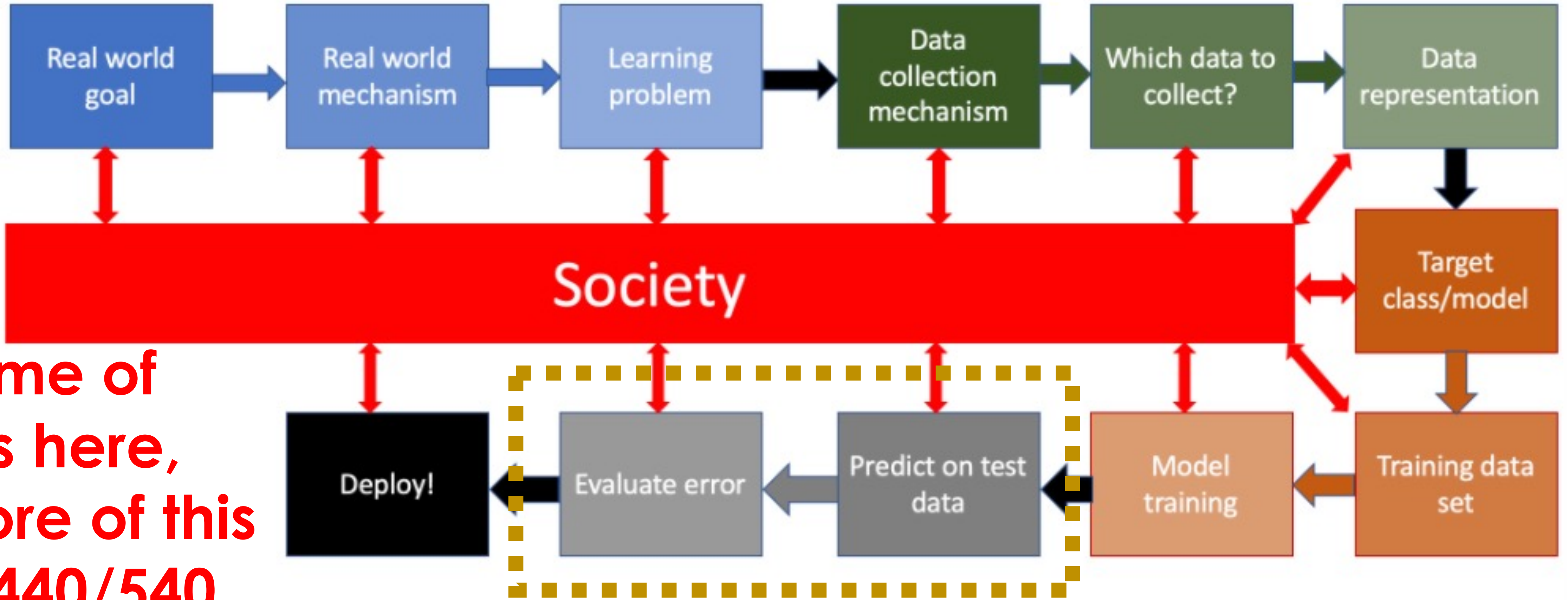
Sum Over All The Data Points

Square The Result

Sum Squared Total Error

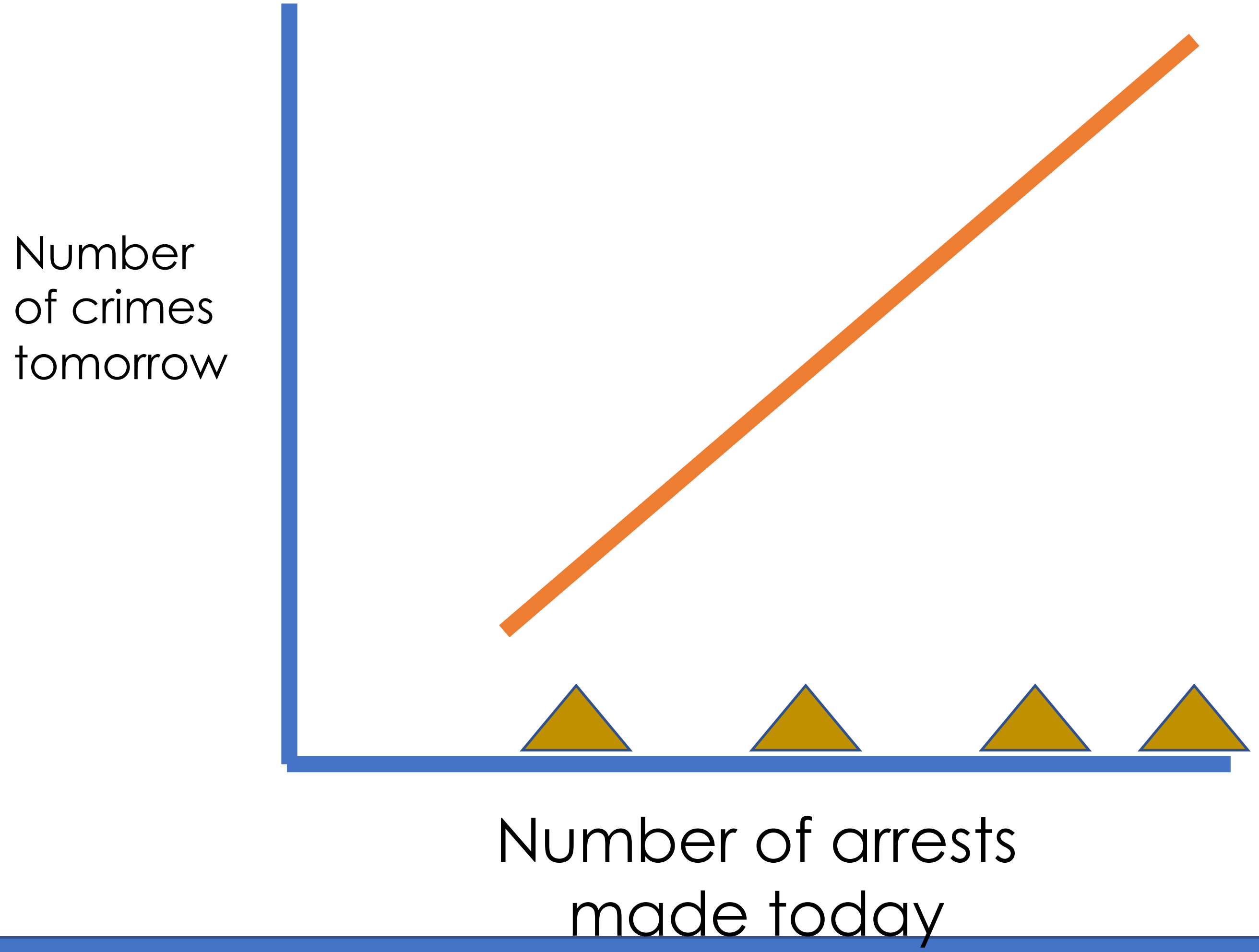
Each Data Point

Mean Value



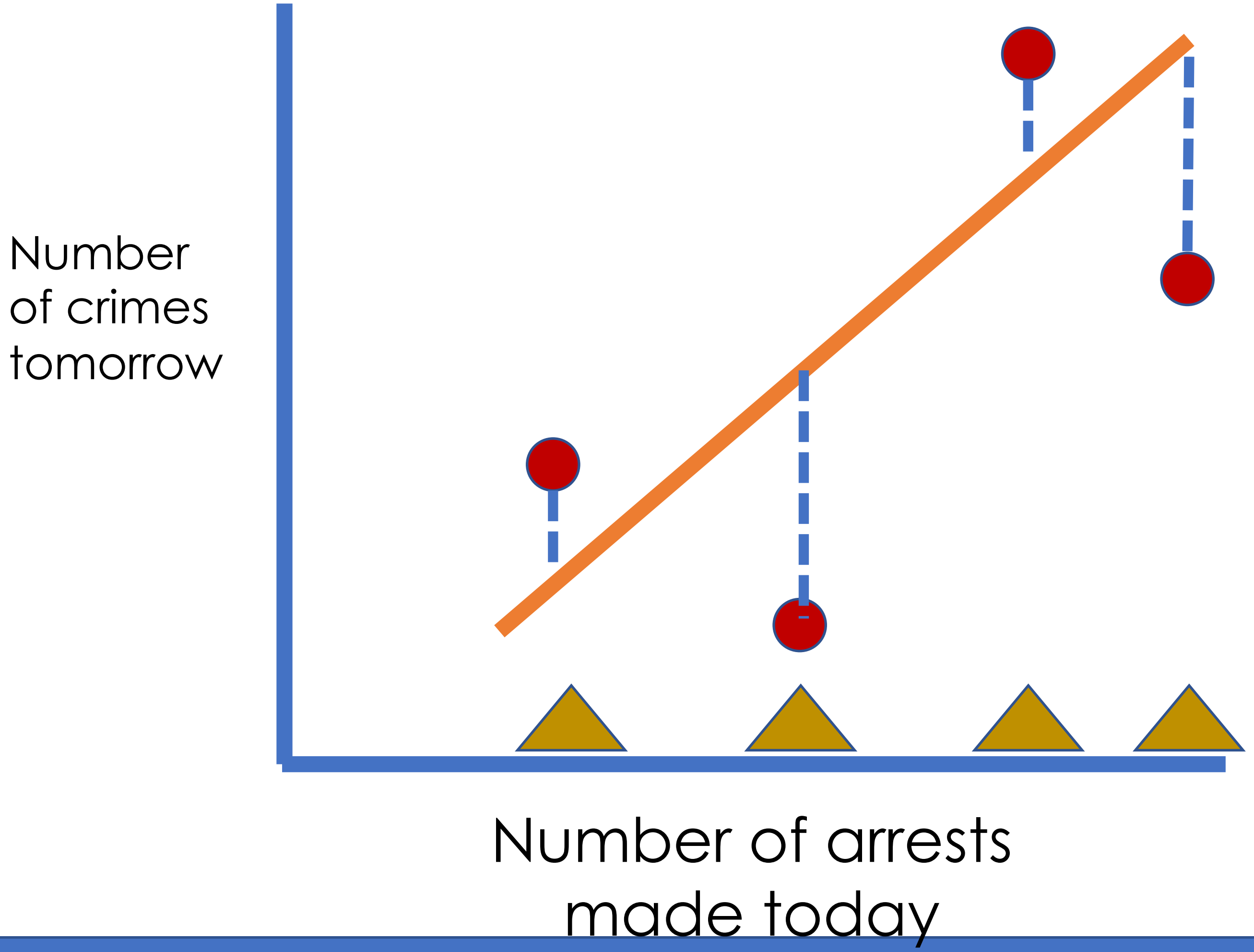
Some of this here, more of this in 440/540

Evaluating regression models



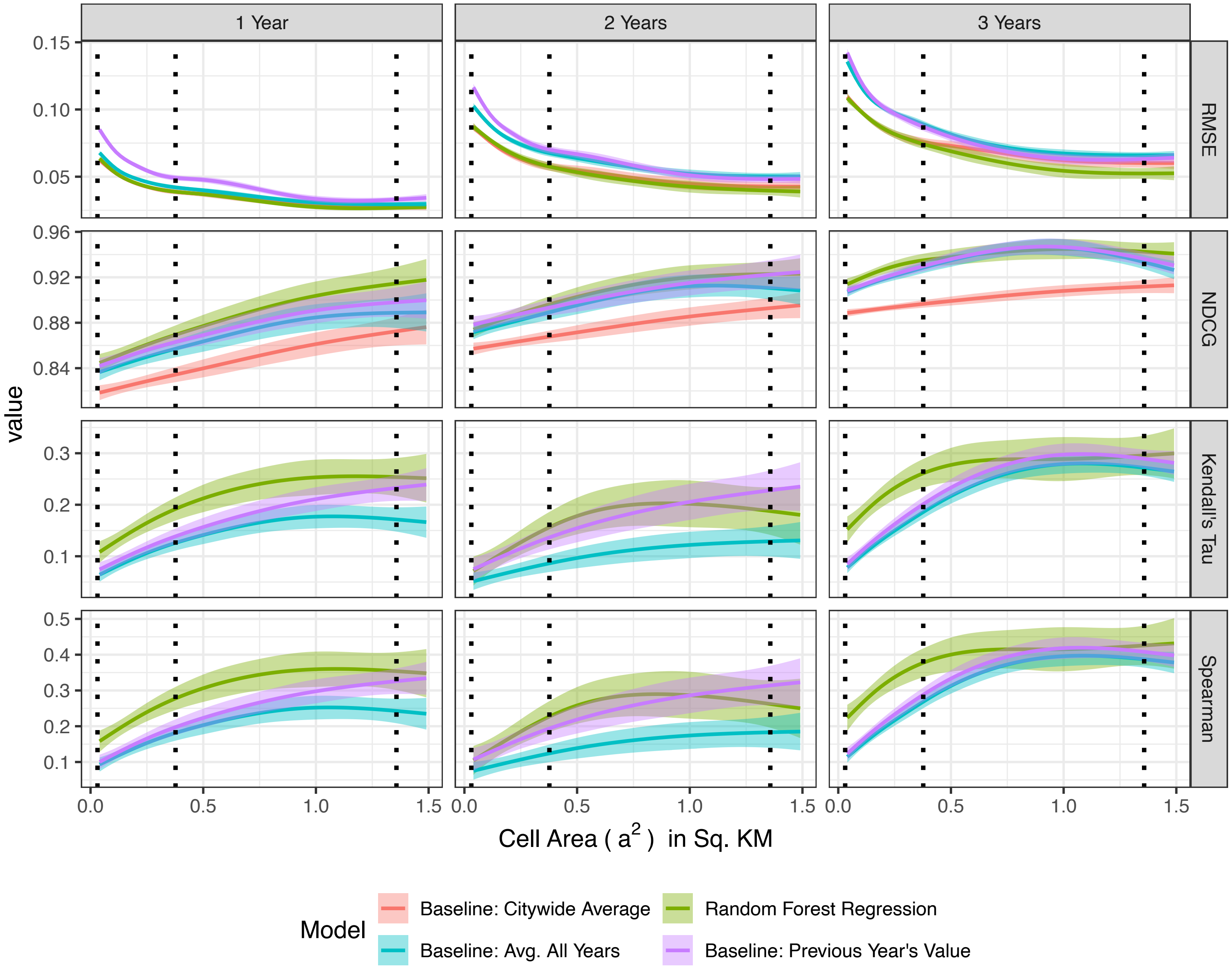
1. Make predictions for some text points

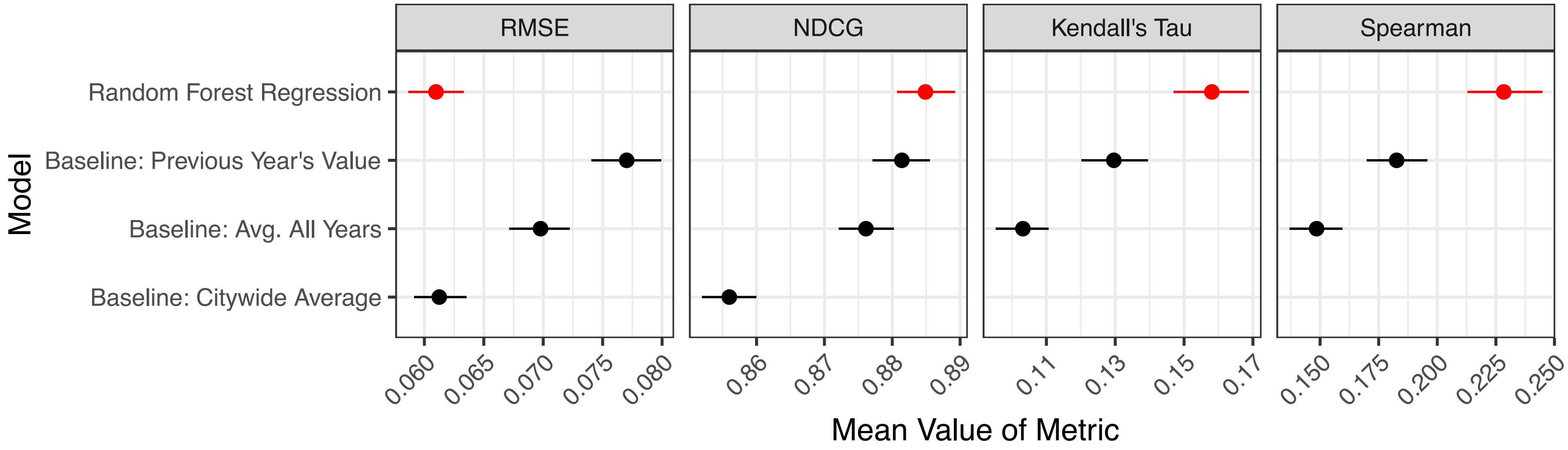
Evaluating regression models (cont.)

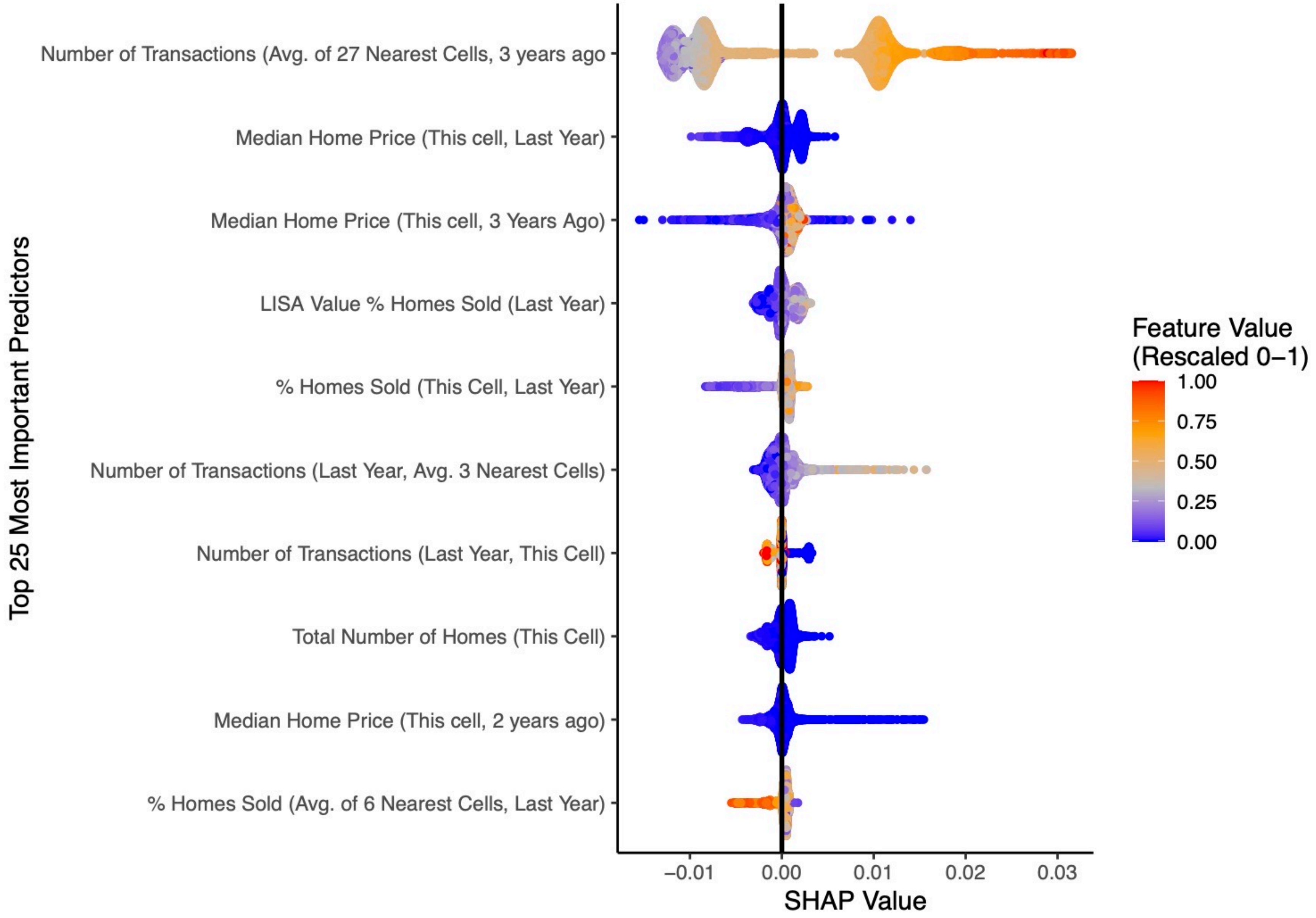


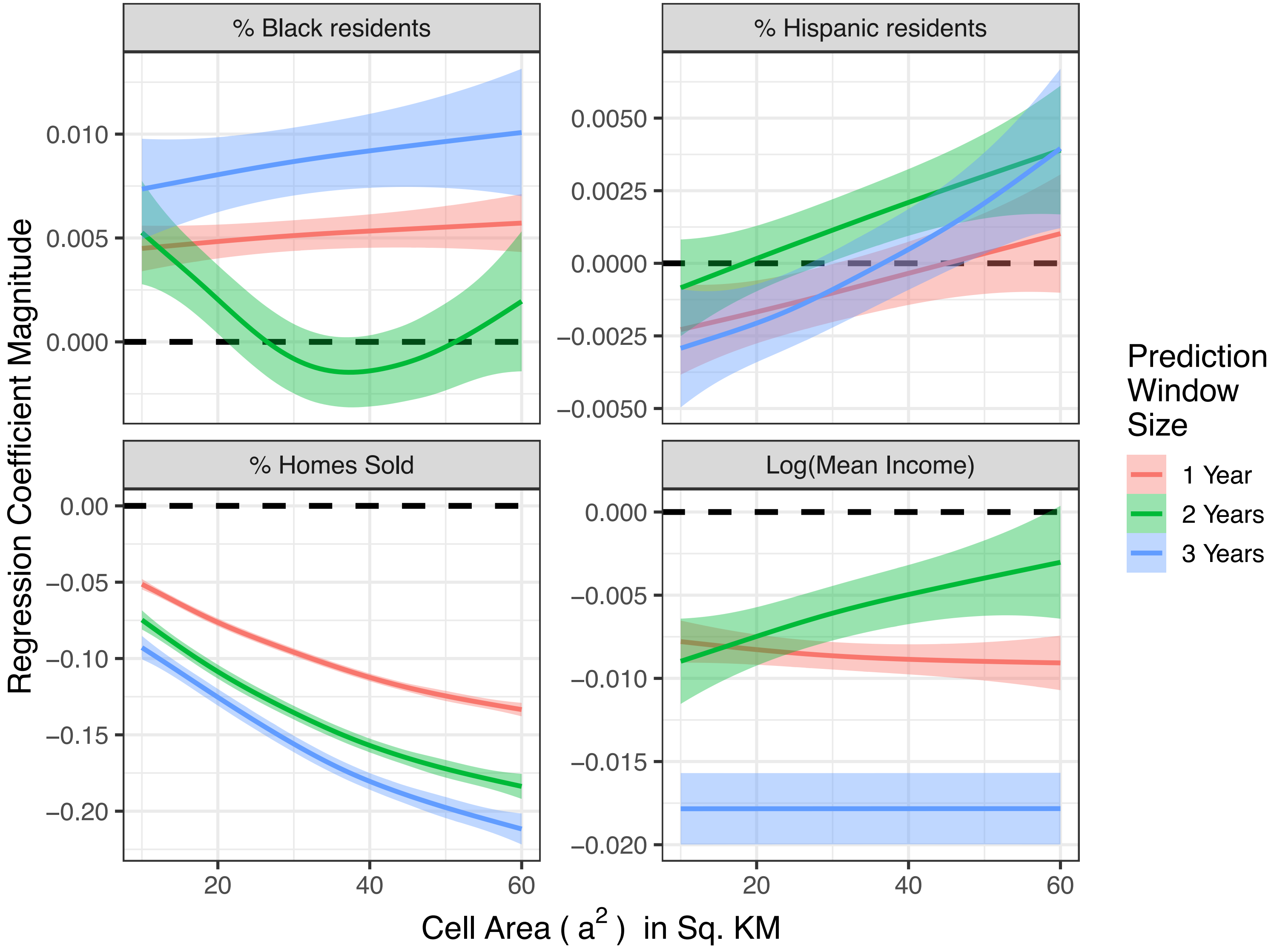
1. Make predictions for some text points
2. Get the error of those predictions
3. Make some aggregate statement about those errors

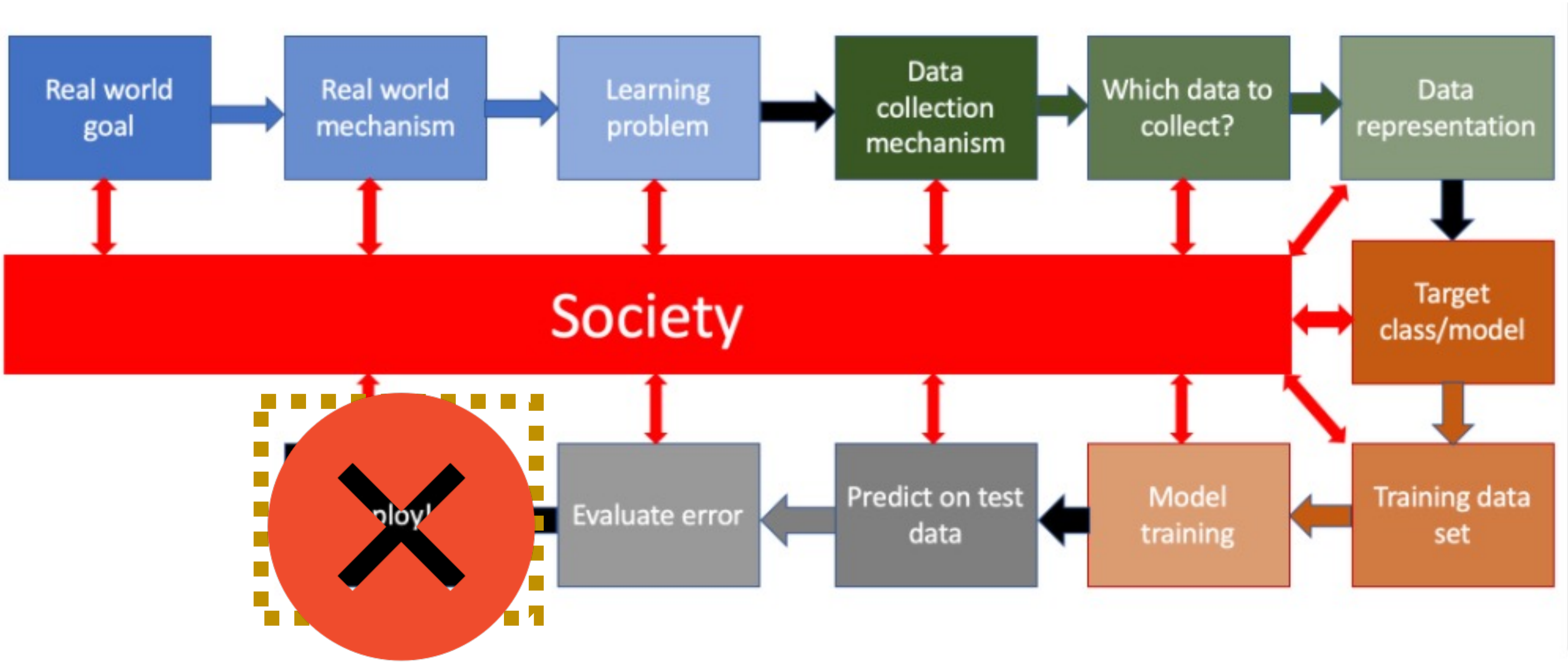
… like what?

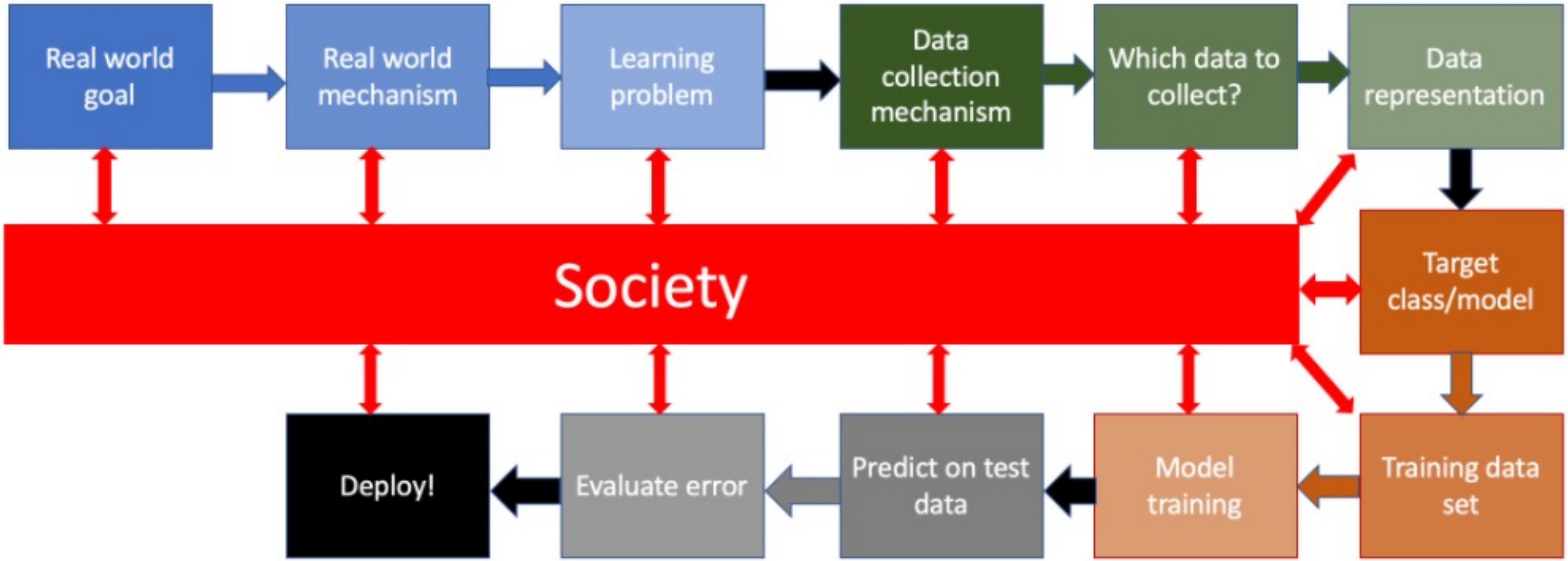

1 Year
2 Years
3 Years
RMSE
NDCG
Kendall's Tau
Spearman
value
0.15
0.10
0.05
0.96
0.92
0.88
0.84
0.3
0.2
0.1
0.5
0.4
0.3
0.2
0.1
0.0
0.5
1.0
1.5
Cell Area (a2) in Sq. KM
Model
Baseline: Citywide Average
Random Forest Regression
Baseline: Avg. All Years
Baseline: Previous Year's Value


RMSE
NDCG
Kendall's Tau
Spearman
Random Forest Regression
Baseline: Previous Year's Value
Baseline: Avg. All Years
Baseline: Citywide Average
Model
0.060 0.065 0.070 0.075 0.080
0.86 0.87 0.88 0.89
0.11 0.13 0.15 0.17
0.150 0.175 0.200 0.225 0.250
Mean Value of Metric


Number of Transactions (Avg. of 27 Nearest Cells, 3 years ago
Median Home Price (This cell, Last Year)
Median Home Price (This cell, 3 Years Ago)
LISA Value % Homes Sold (Last Year)
% Homes Sold (This Cell, Last Year)
Number of Transactions (Last Year, Avg. 3 Nearest Cells)
Number of Transactions (Last Year, This Cell)
Total Number of Homes (This Cell)
Median Home Price (This cell, 2 years ago)
% Homes Sold (Avg. of 6 Nearest Cells, Last Year)
Top 25 Most Important Predictors
−0.01
0.00
0.01
0.02
0.03
SHAP Value
Feature Value
(Rescaled 0−1)
1.00
0.75
0.50
0.25
0.00


% Black residents
% Hispanic residents
% Homes Sold
Log(Mean Income)
Regression Coefficient Magnitude
0.010
0.005
0.000
0.0050
0.0025
0.0000
−0.0025
−0.0050
0.00
−0.05
−0.10
−0.15
−0.20
0.000
−0.005
−0.010
−0.015
−0.020
20
40
60
Cell Area (a2) in Sq. KM
Prediction Window Size
1 Year
2 Years
3 Years


Real world goal
Real world mechanism
Learning problem
Data collection mechanism
Which data to collect?
Data representation
Society
Target class/model
Training data set
Model training
Predict on test data
Evaluate error
ploy!


Real world goal
Real world mechanism
Learning problem
Data collection mechanism
Which data to collect?
Data representation
Society
Target class/model
Deploy!
Evaluate error
Predict on test data
Model training
Training data set