Lecture 6: Intro to Supervised Learning

Intro to Data Science for Public Policy
Spring 2017

Jeff Chen + Dan Hammer

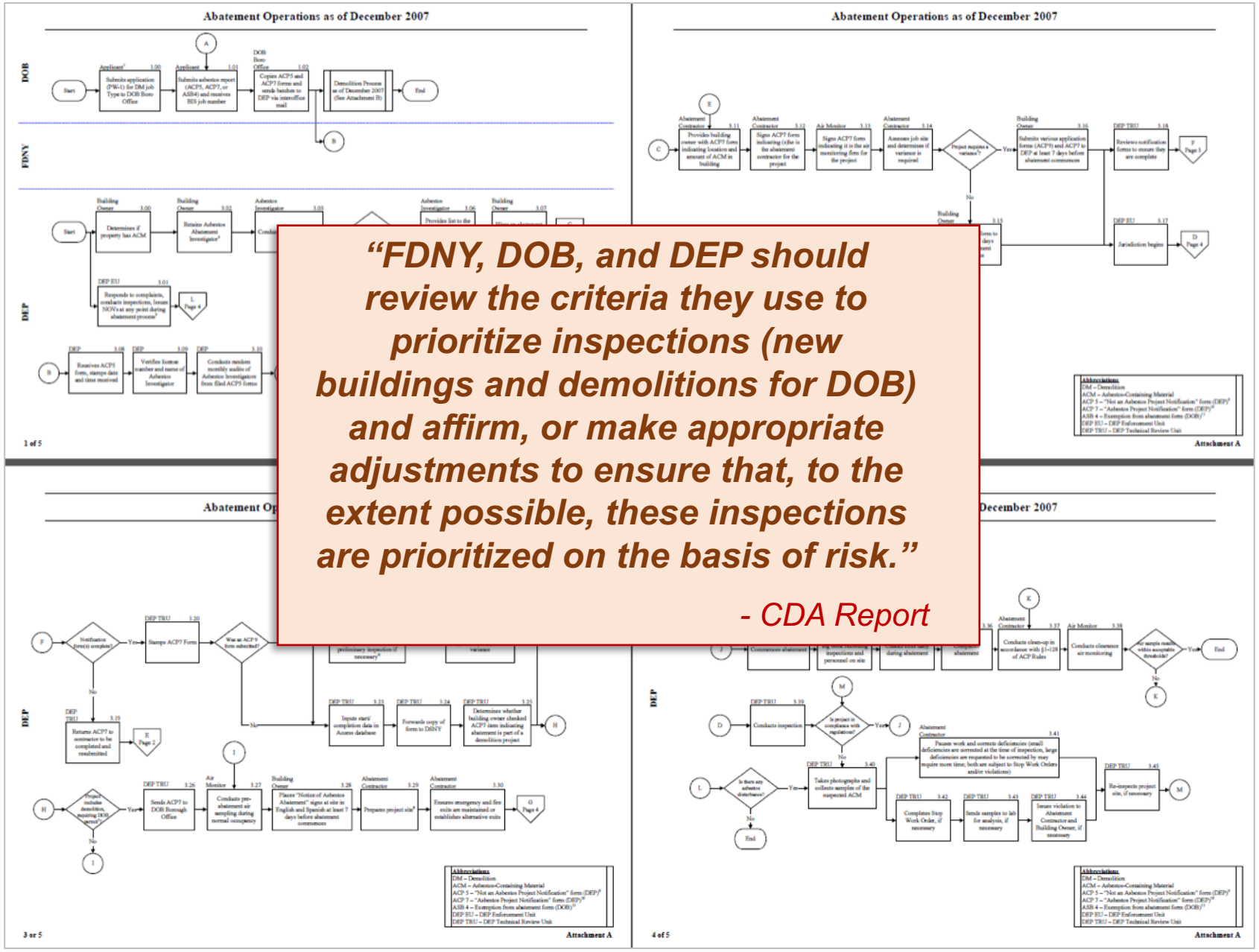
Roadmap

- Motivation
- Supervised Learning
- OLS
- <break>
- KNN
- Homework assignment

The Deutsche Bank Fire (2007)

- In 2007 the Deutsche Bank building was in the process of being remediated and demolished
- Fire engulfed the building killing two firefighters

Systemic Problems with NYC Building Policy



"FDNY, DOB, and DEP should review the criteria they use to prioritize inspections (new buildings and demolitions for DOB) and affirm, or make appropriate adjustments to ensure that, to the extent possible, these inspections are prioritized on the basis of risk."

- CDA Report

Risk Mitigation Philosophy

Fires will happen. It's just a matter of time. The Fire Department needs to get fire units to inspect the buildings of highest risk so they're ready.

Considerations for Inspection Activities

330,000 # of buildings in inspection portfolio

10% Expected proportion that will be inspected per year

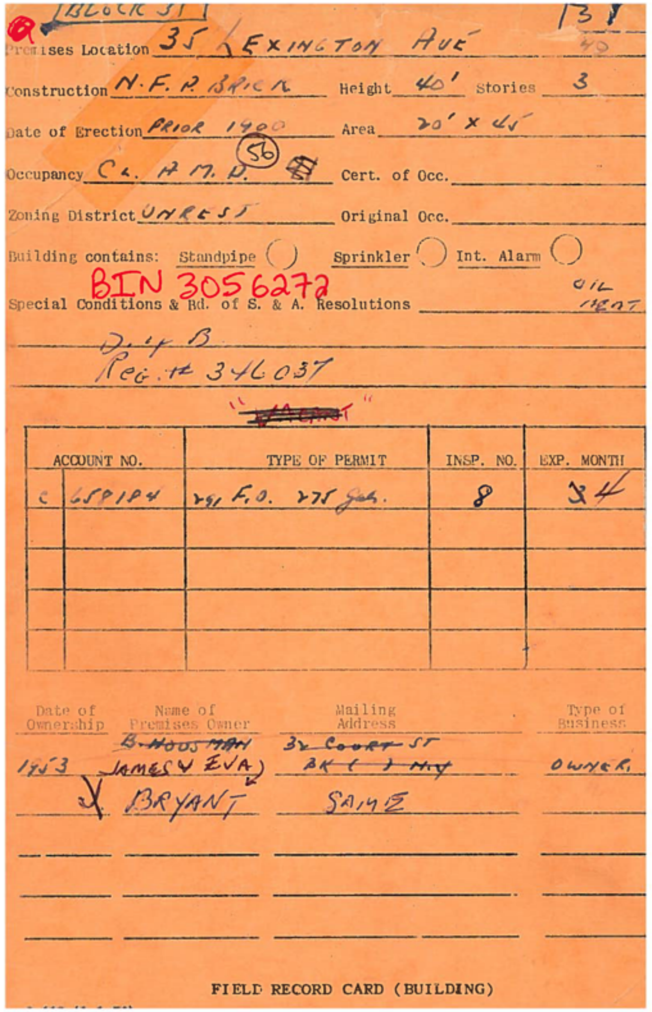
3x3 Each company performs 3 hours of building inspections, 3 times per week

493,000 # of fire incidents per year (# of times an emergency could cut into building inspection time)

105 Heat index value at which inspection operations are suspended

FDNY Tech

Pre-March 2013


BLOCK 31 138

Premises Location 35 LEXINGTON AVE

Construction N.F. R BRICK Height 40' Stories 3

Date of Erection PRIOR 1900 Area 20' x 45'

Occupancy CL. A M.D. (56) Cert. of Occ.

Zoning District UNREST Original Occ.

Building contains: Standpipe ( ) Sprinkler ( ) Int. Alarm ( )

BIN 3056272

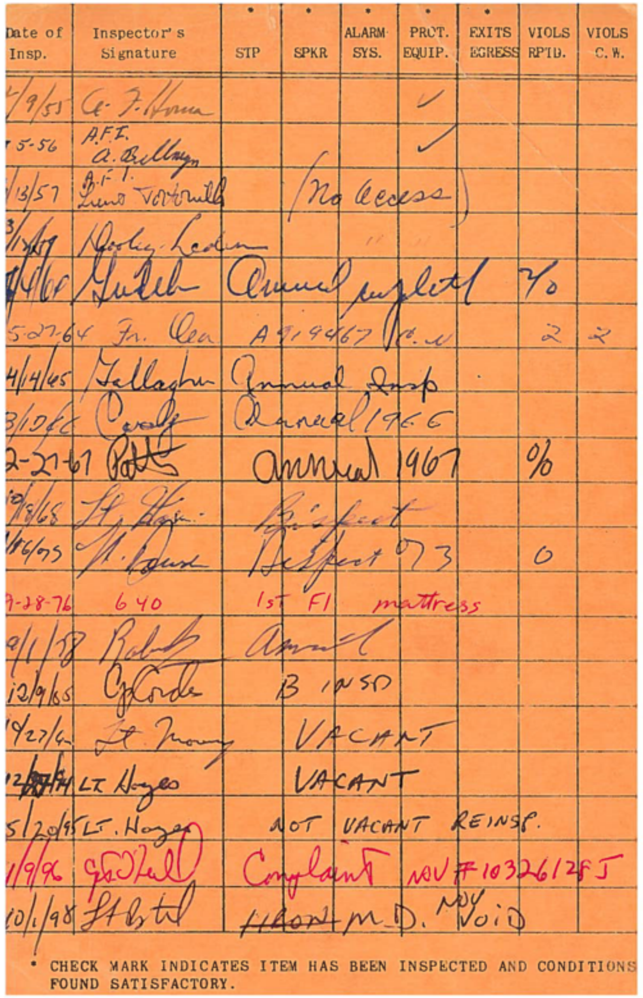
Special Conditions & Bd. of S. & A. Resolutions OIL HEAT

Reg. # 346037

ACCOUNT NO.	TYPE OF PERMIT	INSP. NO.	EXP. MONTH
C 659184	291 F.O. 275 gal.	8	34

Date of Ownership	Name of Premises Owner	Mailing Address	Type of Business
	~~B. HOUSMAN~~	~~32 COURT ST~~	
1953	JAMES & EVA	~~BK 1 N.Y.~~	OWNER
	BRYANT	SAME	

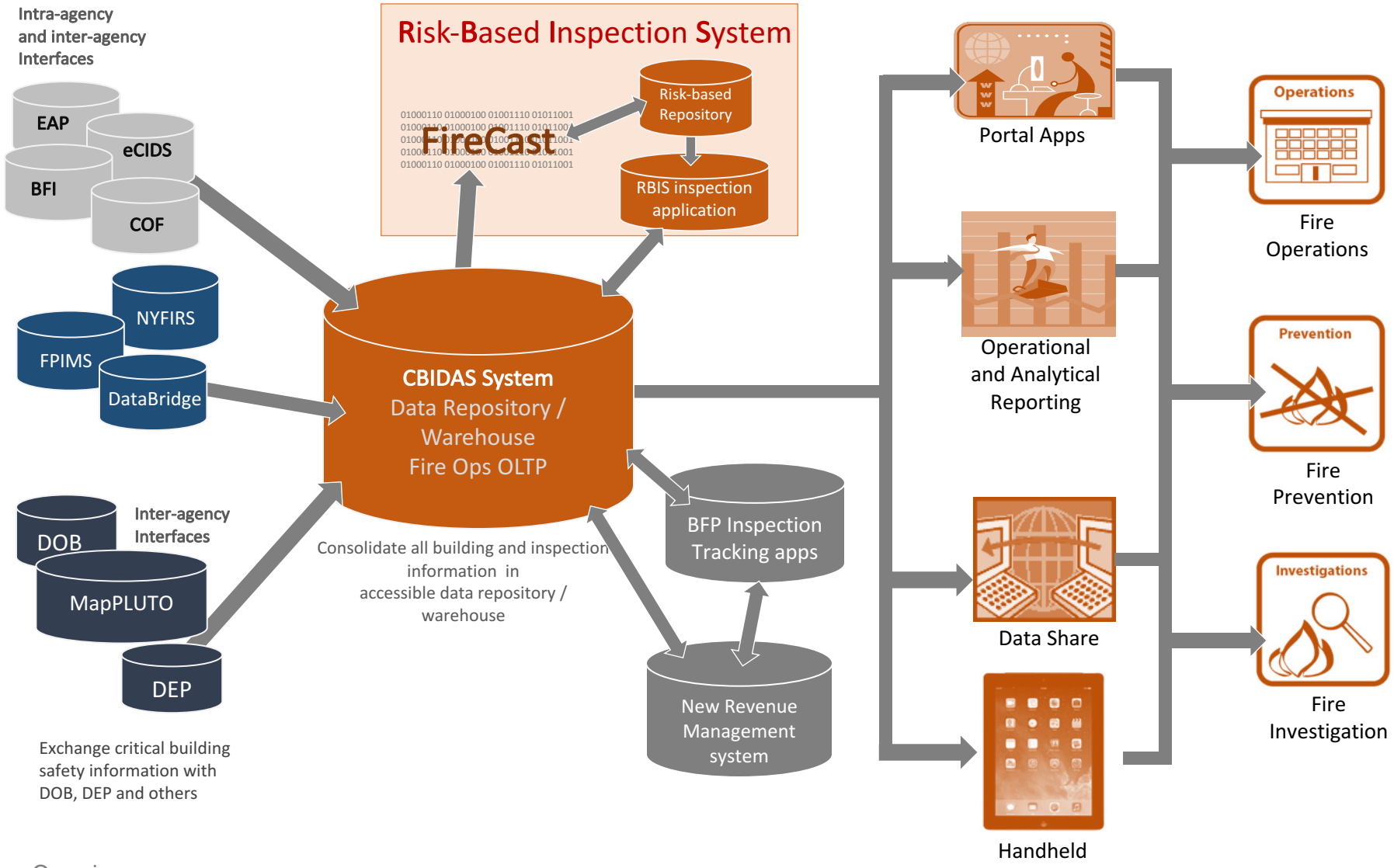
FIELD RECORD CARD (BUILDING)


Date of Insp.	Inspector's Signature	STP	SPKR	ALARM SYS.	PROT. EQUIP.	EXITS EGRESS	VIOLS RPTD.	VIOLS C.W.
1/9/55					✓			
5-56	A.F.I.				✓			
1/13/57		(No access)						
4/4/64		Annual complete					0	
5-27-64		A 919467					2	2
4/14/65		Annual Insp						
3/10/66		Annual 1966						
2-27-67		Annual 1967					0	
10/8/68								
1/6/73		1973					0	
9-28-76	640	1st Fl mattress						
9/1/78		Annual						
12/9/85		B INSP						
4/27/	Lt.	VACANT						
12/	LT Hayes	VACANT						
5/20/95	LT. Hayes	NOT VACANT REINSP.						
1/9/96		Complaint NOV # 10326128 J						
10/1/98		M.D. NOV VOID						

* CHECK MARK INDICATES ITEM HAS BEEN INSPECTED AND CONDITIONS FOUND SATISFACTORY.

FDNY Tech: Current and Future Build

Coordinated Building Inspection Data Analysis System (CBIDAS)



If a building catches fire, it's usually assumed to be random.

Bldg 1	Bldg 7	Bldg 13	Bldg 19	Bldg 25	Bldg 31
Bldg 2	Bldg 8	Bldg 14	Bldg 20	Bldg 26	Bldg 32
Bldg 3	Bldg 9	Bldg 15	Bldg 21	Bldg 27	Bldg 33
Bldg 4	Bldg 10	Bldg 16	Bldg 22	Bldg 28	Bldg 34
Bldg 5	Bldg 11	Bldg 17	Bldg 23	Bldg 29	Bldg 35
Bldg 6	Bldg 12	Bldg 18	Bldg 24	Bldg 30	Bldg 36

But there may actually be patterns in the data.

Cond A	Cond B	Cond C	Cond D	Cond E	Cond F
Bldg 1	Bldg 7	Bldg 13	Bldg 19	Bldg 25	Bldg 31
Bldg 2	Bldg 8	Bldg 14	Bldg 20	Bldg 26	Bldg 32
Bldg 3	Bldg 9	Bldg 15	Bldg 21	Bldg 27	Bldg 33
Bldg 4	Bldg 10	Bldg 16	Bldg 22	Bldg 28	Bldg 34
Bldg 5	Bldg 11	Bldg 17	Bldg 23	Bldg 29	Bldg 35
Bldg 6	Bldg 12	Bldg 18	Bldg 24	Bldg 30	Bldg 36

Pr(Fire) = *f(Input Features)*

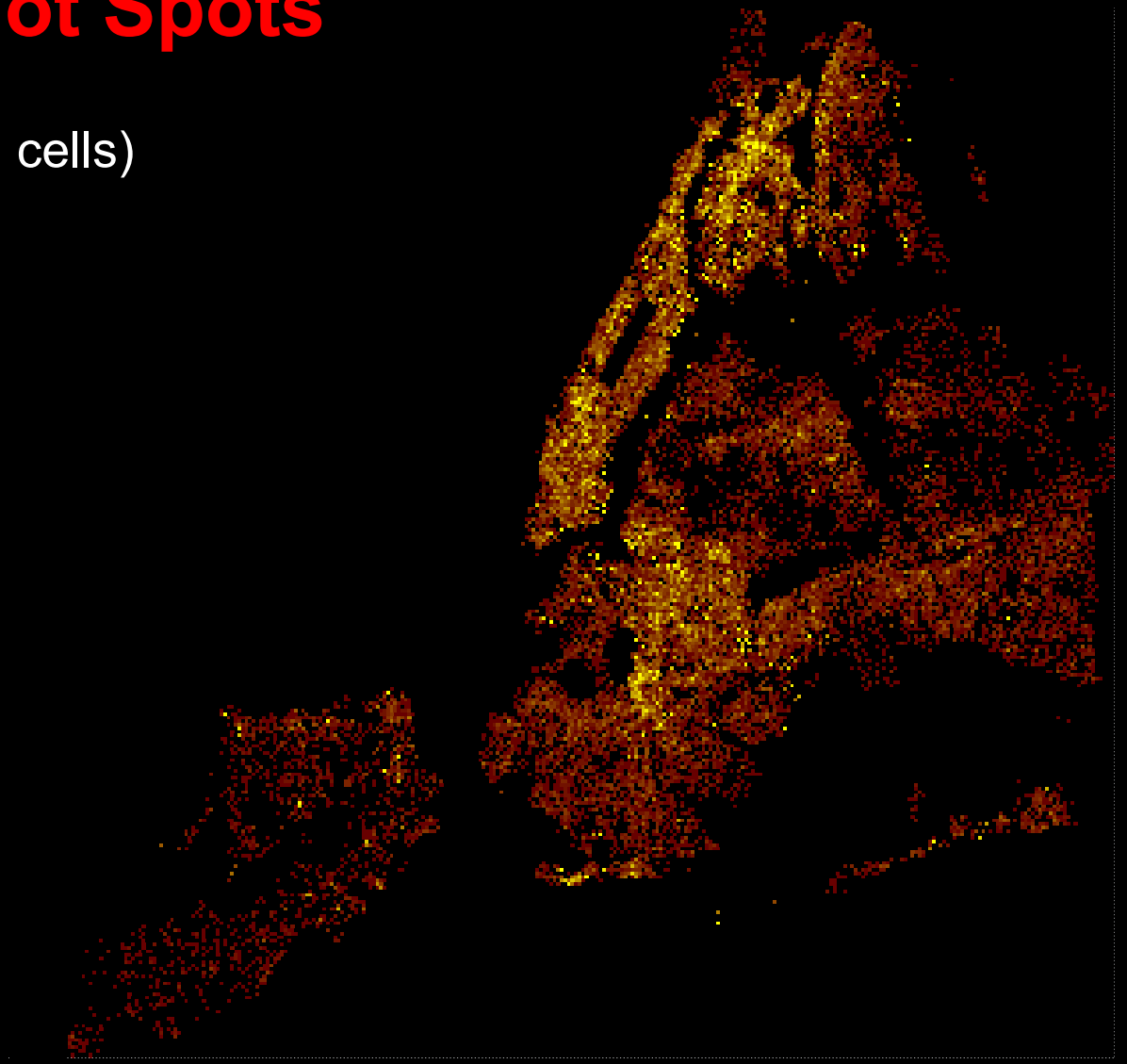
Cond A	Cond B	Cond C	Cond D	Cond E	Cond F
Bldg 1	Bldg 7	Bldg 13	Bldg 19	Bldg 25	Bldg 31
Bldg 2	Bldg 8	Bldg 14	Bldg 20	Bldg 26	Bldg 32
Bldg 3	Bldg 9	Bldg 15	Bldg 21	Bldg 27	Bldg 33
Bldg 4	Bldg 10	Bldg 16	Bldg 22	Bldg 28	Bldg 34
Bldg 5	Bldg 11	Bldg 17	Bldg 23	Bldg 29	Bldg 35
Bldg 6	Bldg 12	Bldg 18	Bldg 24	Bldg 30	Bldg 36

Known as "Independent Variable" or "Input Feature"

Known as "Dependent Variable", "Label" or "Target"

Fire + Hot Spots

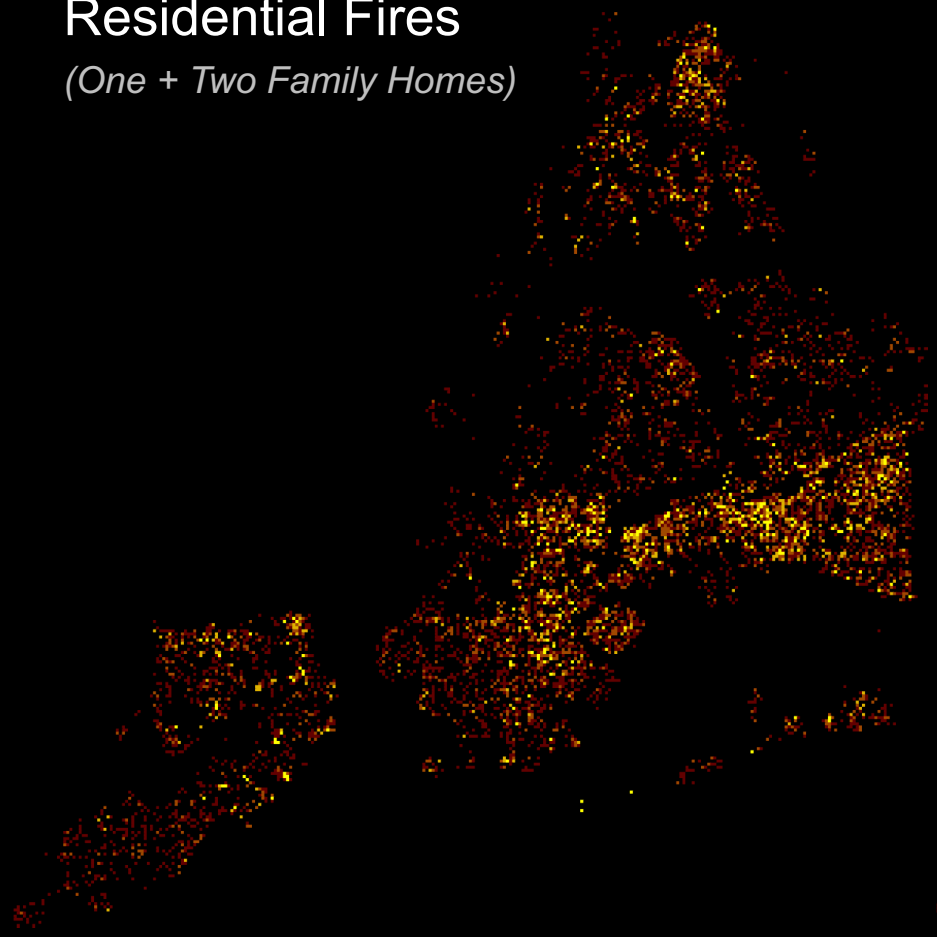
2002 to 2013
(500 foot grid cells)



Fire + Hot Spots

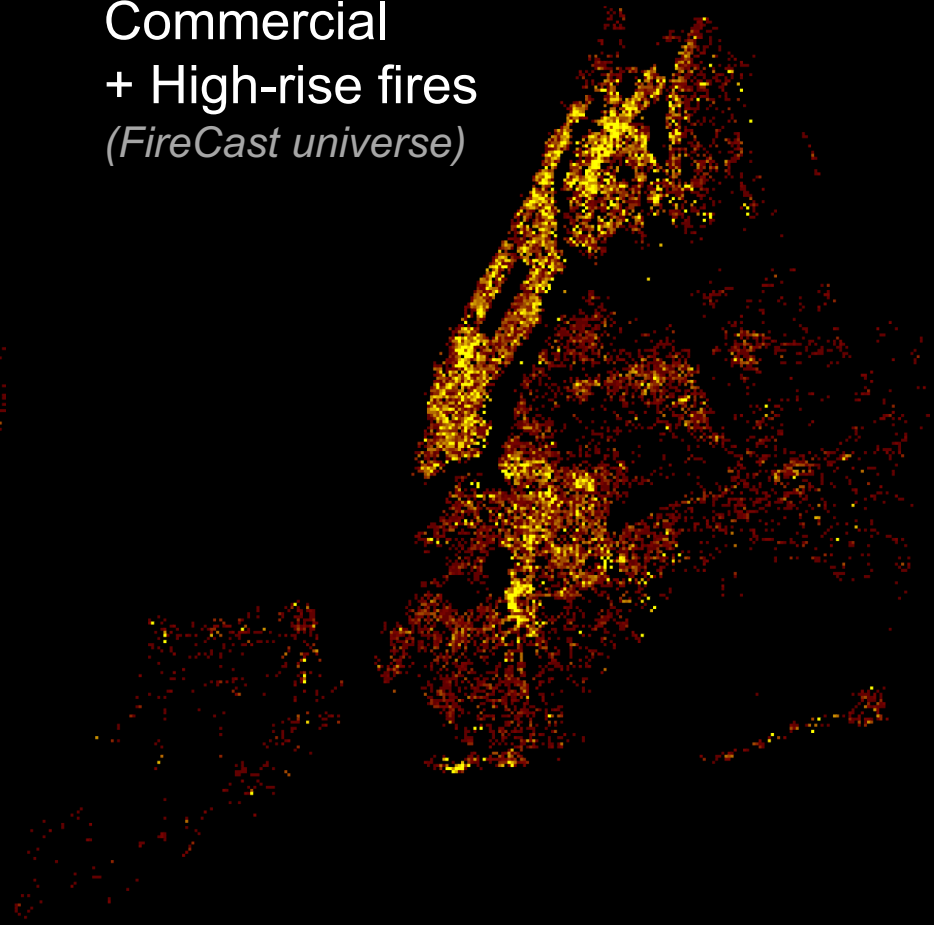
Residential Fires

(One + Two Family Homes)

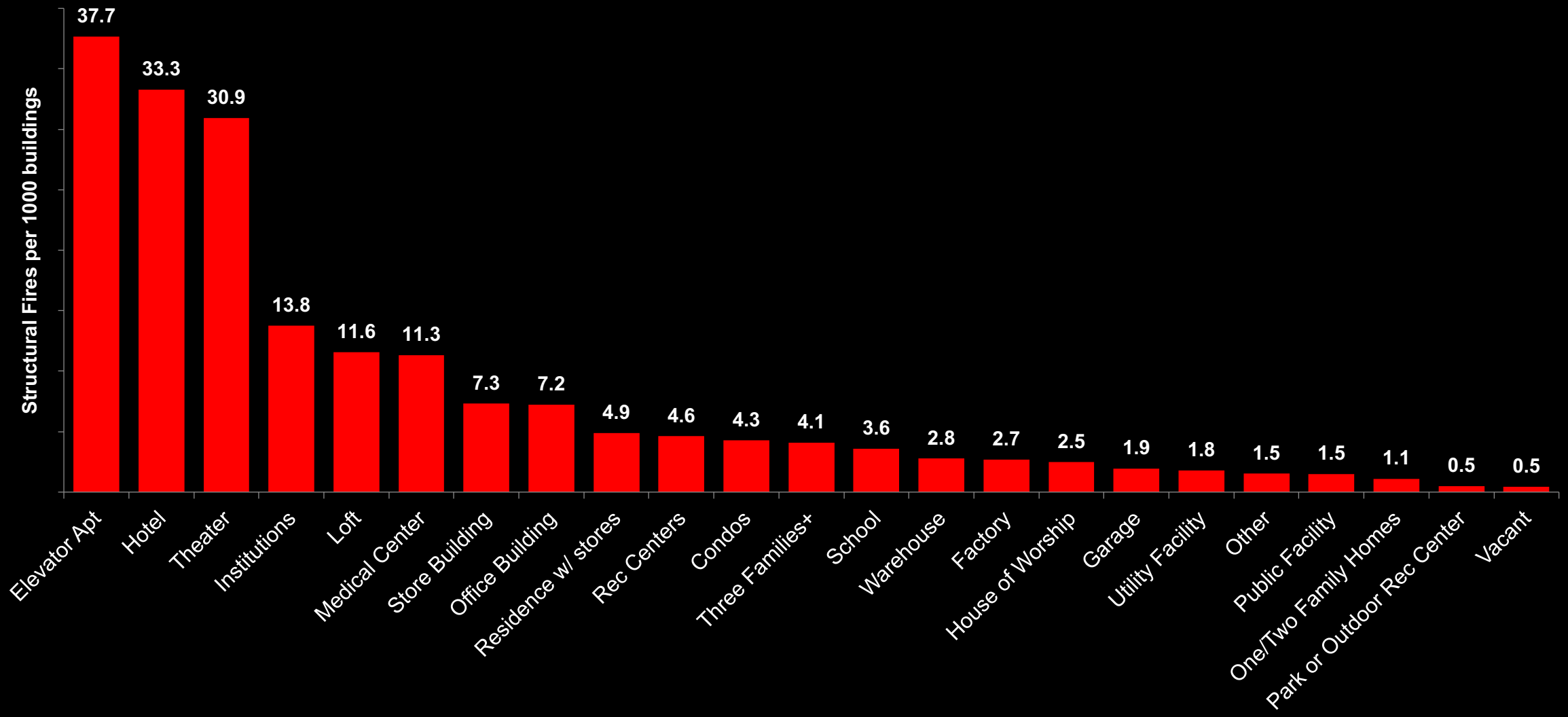


Commercial
+ High-rise fires

(FireCast universe)



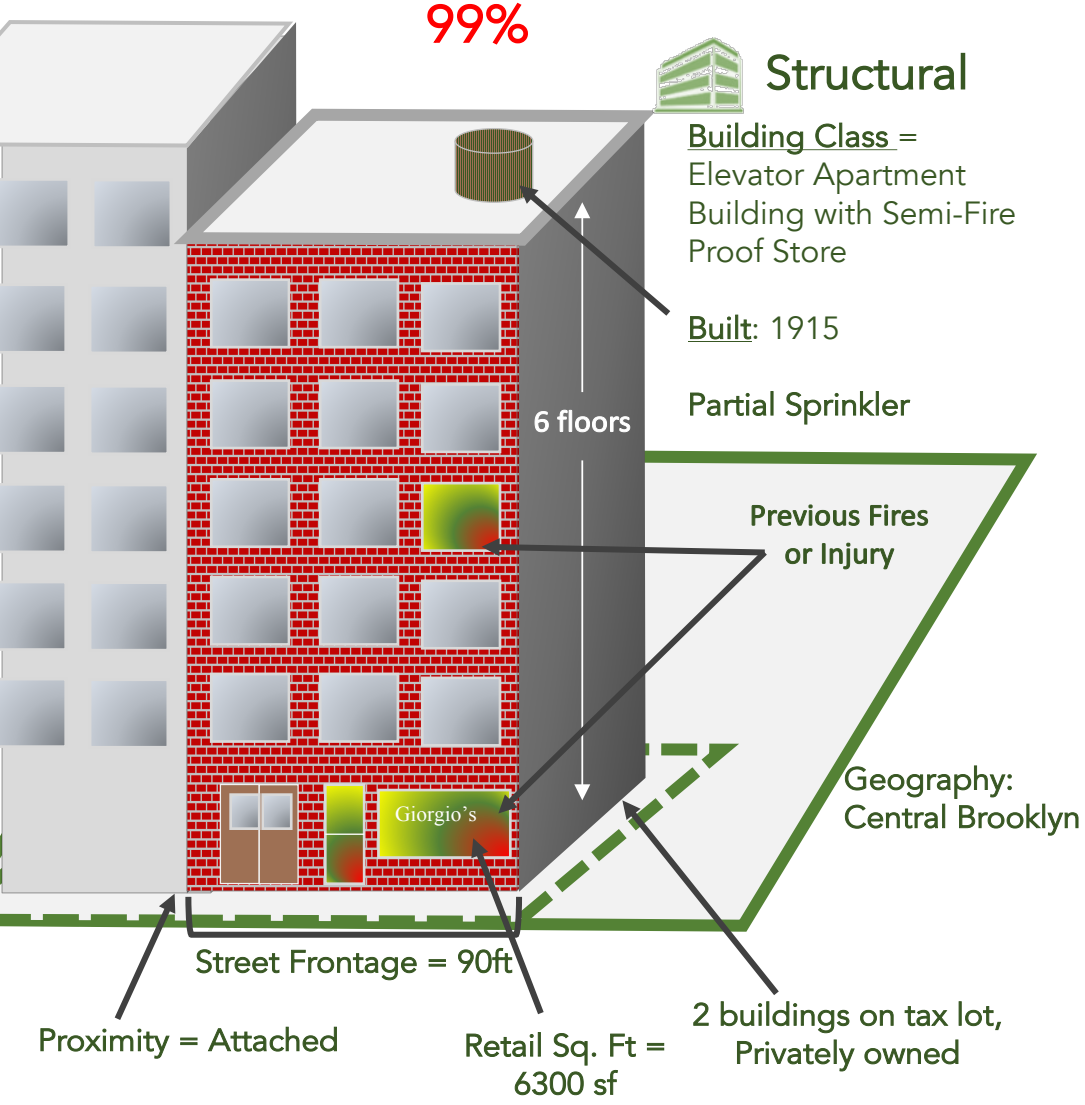
Fire Incidence + Building Use



One Model for Citywide Use

Example building with risk score

99%



FireCast 2.0

EXAMPLE

FireCast 2.0:

Consistent Risk Model

Prob(|)

*Probability of **fire ignition** given structural characteristics*

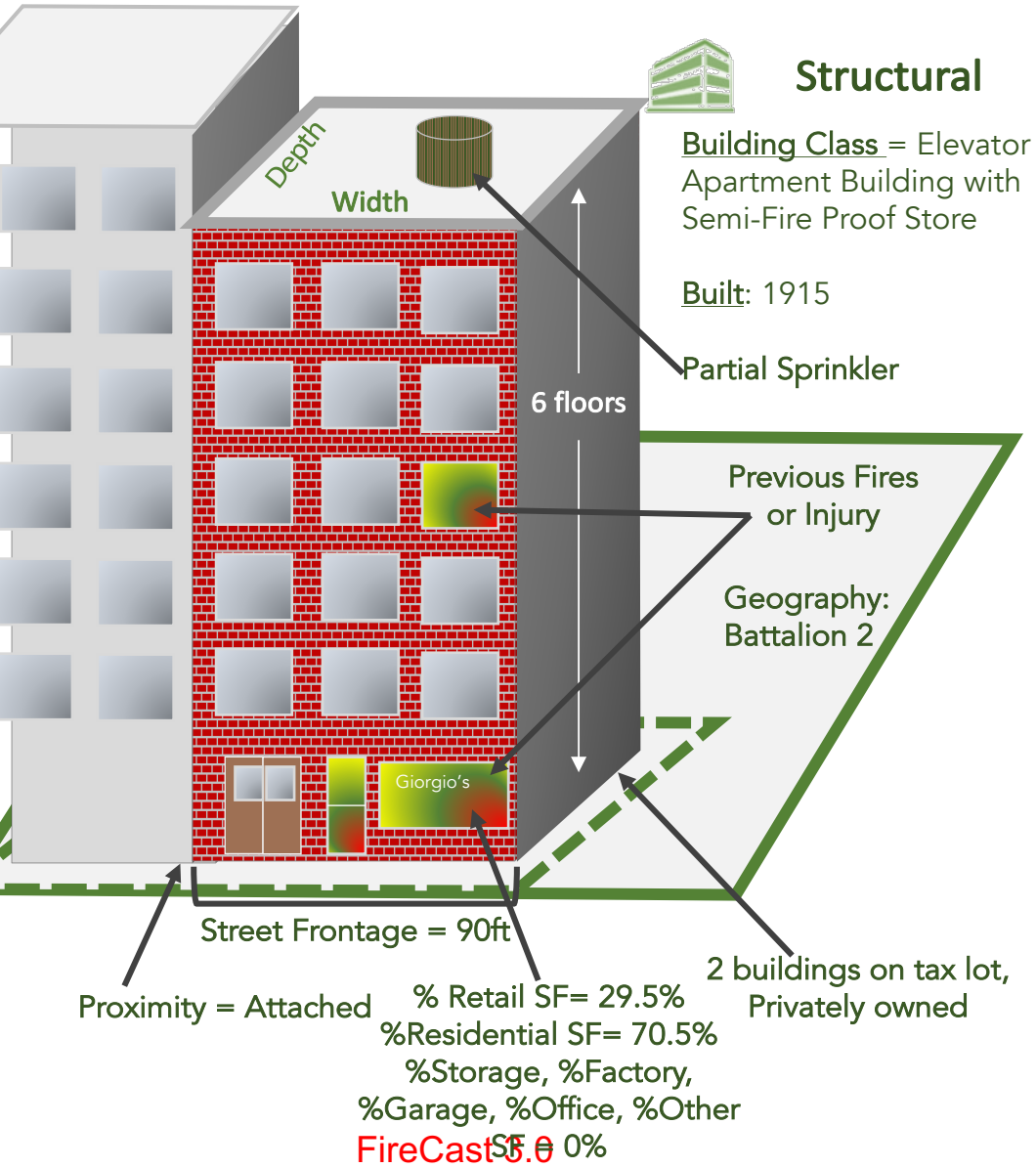
DEFINITION

IMPACT

- *Impact (Avg # violations)*
  - *First 30 days = +**19%***
  - *First 60 days = +10%*
- *Pre-Incident Rate*
  - ***16.5%** of buildings that experience a fire were inspected within 90 days before the day of incident*

One Risk Model per Incident Type for Each Battalion (49 models)

Risk score: **Depends**



FireCast 3.0

FireCast 3.0:

Machine Learning Model

Prob (|)

DEFINITION

FORMULATION

Behavioral Cues

- Excessive Noise
- Air Quality
- Sidewalk Condition
- Electrical Issues
- Rodents
- Lead
- Seasonality
- Sewer overflows
- Heating problems
- +22 other types

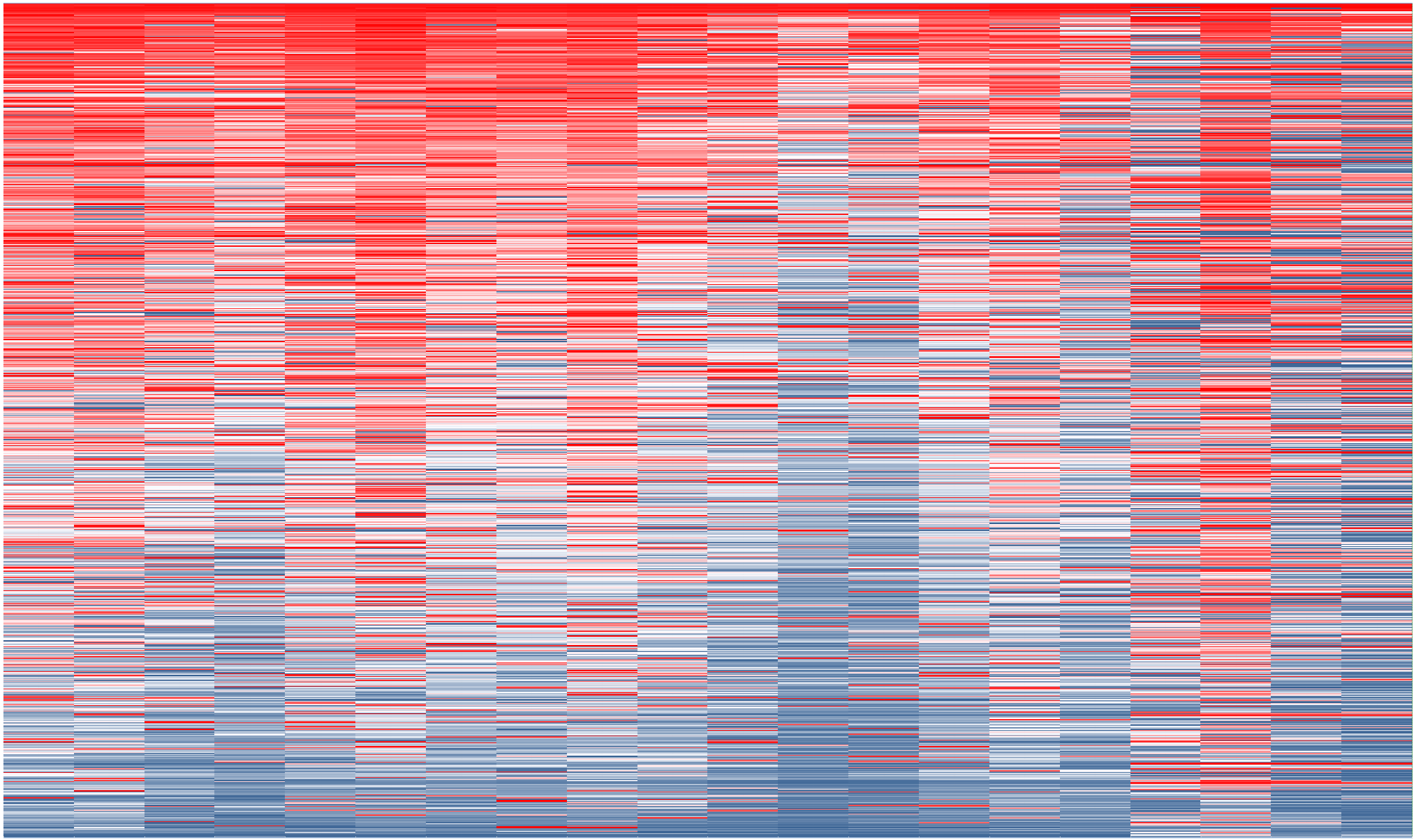
Violation Activity

- Sidewalk cleanliness
- Loose Trash
- Failure to Maintain
- Signage, Postings
- Work without permit
- Illegal alteration
- Accumulation of combustible waste
- Missing certificates of fitness
- Cert. of Occupancy not reflective
- +19 other types

Risk Sequence of FDNY Battalion 2

← Time →

← Buildings →



Roadmap

- Motivation
- Supervised Learning
- OLS
- <break>
- KNN
- Homework assignment

what is supervised learning?!

Predicting the chance that homes are with or without heat or electricity after a natural disaster.

Determining areas most likely to be hardest hit by a disaster

Mining raw data feeds to uncover critical levers to affect change in the field

Measuring the impacts of natural disasters and other extreme events.

Prioritizing or **segmenting** populations for specific action (e.g. patient triage)

Optimizing and **reallocating** resource deployment for inspections.

Identifying infrastructure that is most critical for homeland security protection.

Estimating outcomes from any policy question or risk mitigation initiative.

Monitoring for spikes in adverse activity.

what is supervised learning?!

Any empirical task where a set of **examples** / **labels** / **dependent** / **Y** is predicted using corresponding **predictors** / **independent** / **features** / **X**.

The predictions are constructed by mapping X's to Y's with the goal of minimizing prediction error.

what is supervised learning?!

It's supervised because the algorithm is learning from the examples you provide it.

Elements of Supervised Learning

1. Intended Use
2. The Data
3. Algorithms
4. Experiment Design

Elements of Supervised Learning

1. **Intended Use**
   - Prediction vs. Estimation Relationships
     - Prediction problems usually indicate that there is a greater emphasis placed on empirical accuracy
     - Estimation of relationships places emphasis on human-interpretable insight

Elements of Supervised Learning

1. **Intended Use**

- Strong prediction models can be used to direct attention and resources.
- Estimation exercises can be used to inform general strategy, but often times do not have enough prediction power to lead to action.

Elements of Supervised Learning

2. Data

- Inputs
  - Y/target/examples/labels/dependent
    - Continuous variable targets = regression
    - Discrete variable targets = classifiers
  - X/predictors/independent/features

Elements of Supervised Learning

2. Data

- Core consideration: Units must fit the decision
  - Unit of analysis
  - Data pipeline

Elements of Supervised Learning

Question

- If a fire unit inspects buildings on a schedule of every two days, how often does the data need to be updated?
  a) Real-time
  b) 36 hours
  c) 72 hours
  d) Weekly

Elements of Supervised Learning

3. Supervised Learning Algorithms

- Are a series of rules and calculations that convert raw data into insight.
- Parts
  - Input variables
  - Weights
  - Error and accuracy functions

Elements of Supervised Learning

Common error measures

- Continuous variables
  - Root Mean Squared Error
  - Mean Absolute Percentage Error
- Discrete
  - F-1 Score
  - Area Under the Curve

Elements of Supervised Learning

3. Supervised Learning Algorithms

- Linear Regression
  - Y = B0 + B1X + E
  - Gradient Descent
  - R-squared, RMSE
  - Output of regression is an estimate of Y (Y-hat)

4. Experimental design

An icebreaker example

Icebreaker #1

#	2	4		16	32	
pos	1	2	3	4	5	6

Based on the examples, what will the 3rd and 6th values likely be?

Example

Icebreaker #1

#	2	4	8	16	32	64
pos	1	2	3	4	5	6

Based on the examples, what will the 6^{th} value be?

Example

Icebreaker #2

#	**1**	**1**	**2**	**3**	**5**	
pos	1	2	3	4	5	6

Based on the examples, what will the 6^{th} value be?

Example

Icebreaker #2

#	1	1	2	3	5	8
pos	1	2	3	4	5	6

Based on the examples, what will the 6^{th} value be?

Example

Icebreaker #3

#	0.8	0.9	0.14	-.75	-.95	-.28
pos	1	2	3	4	5	6

Based on the examples, what will the 6^{th} value be?

Example

Icebreaker #3

#	0.8	0.9	0.14	-.75	-.95	-.28
pos	1	2	3	4	5	6

Based on the examples, what will the 6^{th} value be?

Example

Icebreaker #4

#	D	A	Z	Z	L	E
pos	1	2	3	4	5	6

Based on the examples, what will the 4^{th} value be?

Example

Icebreaker #4

#	D	A	Z	Z	L	E
pos	1	2	3	4	5	6

Based on the examples, what will the 4^{th} value be?

Example

Question What is the difference between #1/#2 and #3/#4?

	1	2	3	4	5	6
#1	2	4	[ ]	16	32	[ ]
#2	1	1	2	3	5	[ ]
#3	0.8	0.9	0.14	-.75	-.95	-.28
#4	D	A	Z	Z	L	E

Example

The Point

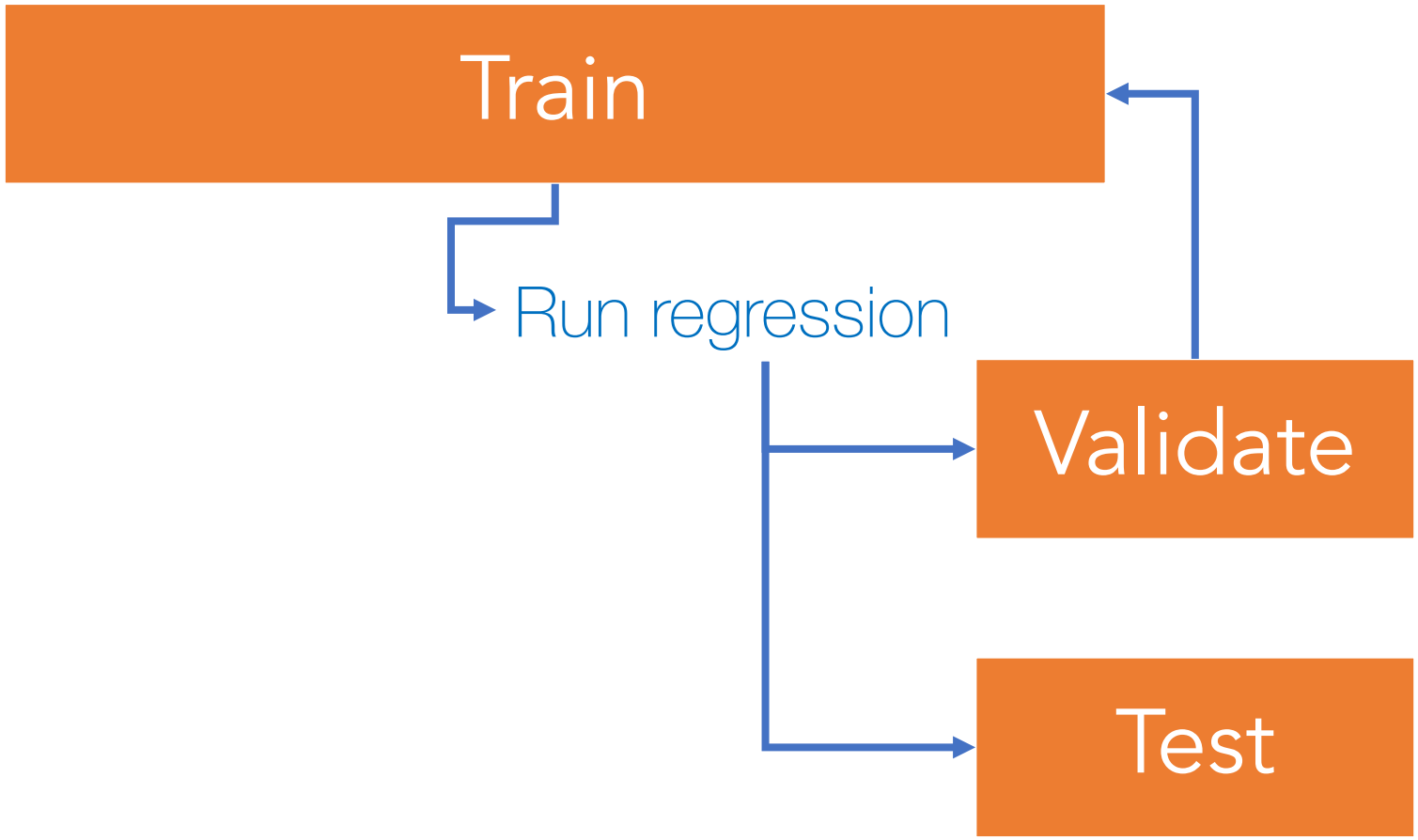
You can’t test your ability to find patterns if you know exactly what will be asked.

Experimental Design

Train/Validate/Test

- Split data into three groups (70/15/15)
- Run a model on training
- Use model weights to predict validate, iterate
- Final check is to predict test

Experimental Design: Regression



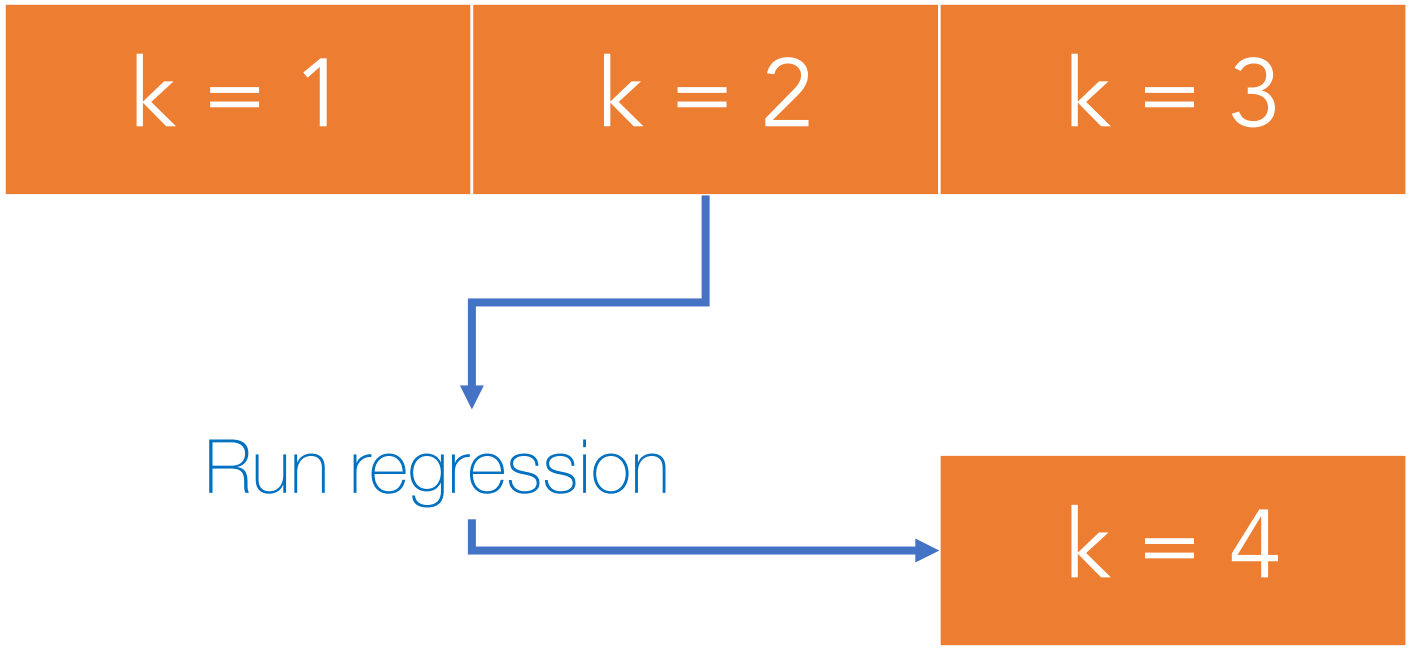
- *Run regression*
- *Predict validation sample, check error, if error is ok*
- *Then predict test set as a final check*

Experimental Design

K-folds Cross Validation

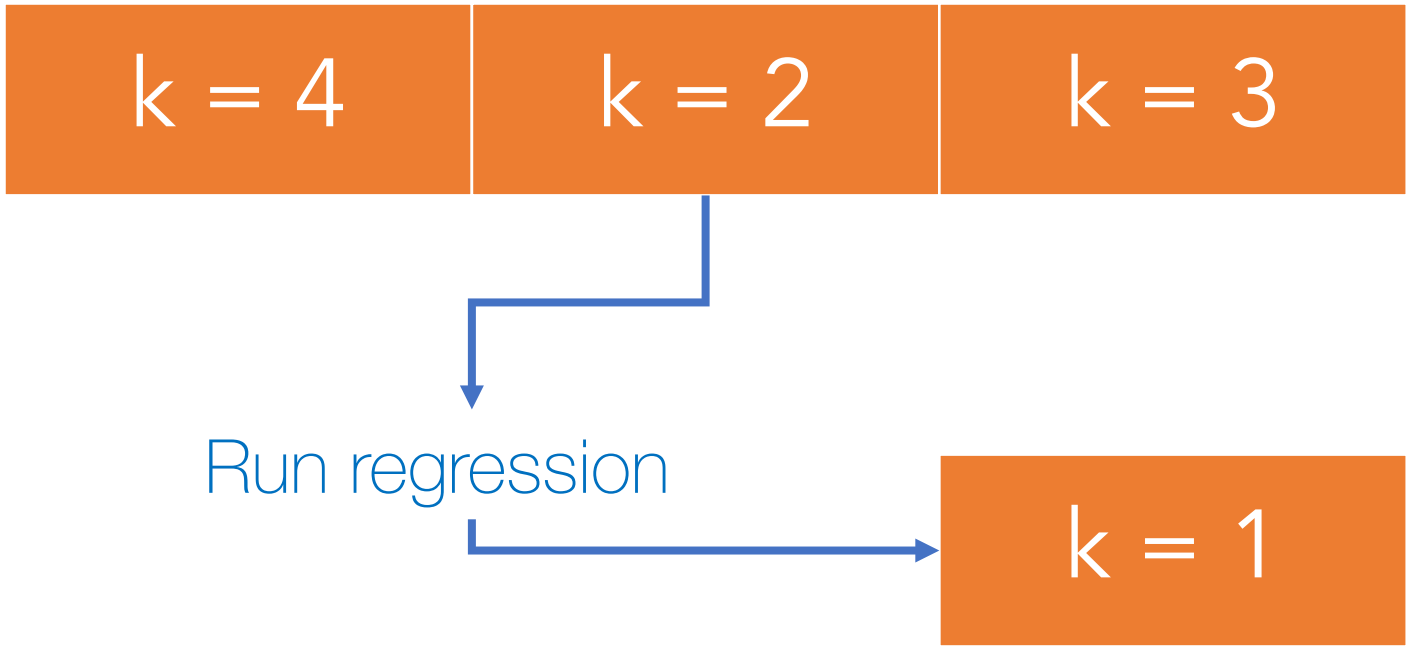
- Split data into k groups
- Run a model on k-1 groups
- Predict the kth group
- Cycle through until all k groups are predicted
- More k's means lower the variance of prediction

Experimental Design: Regression



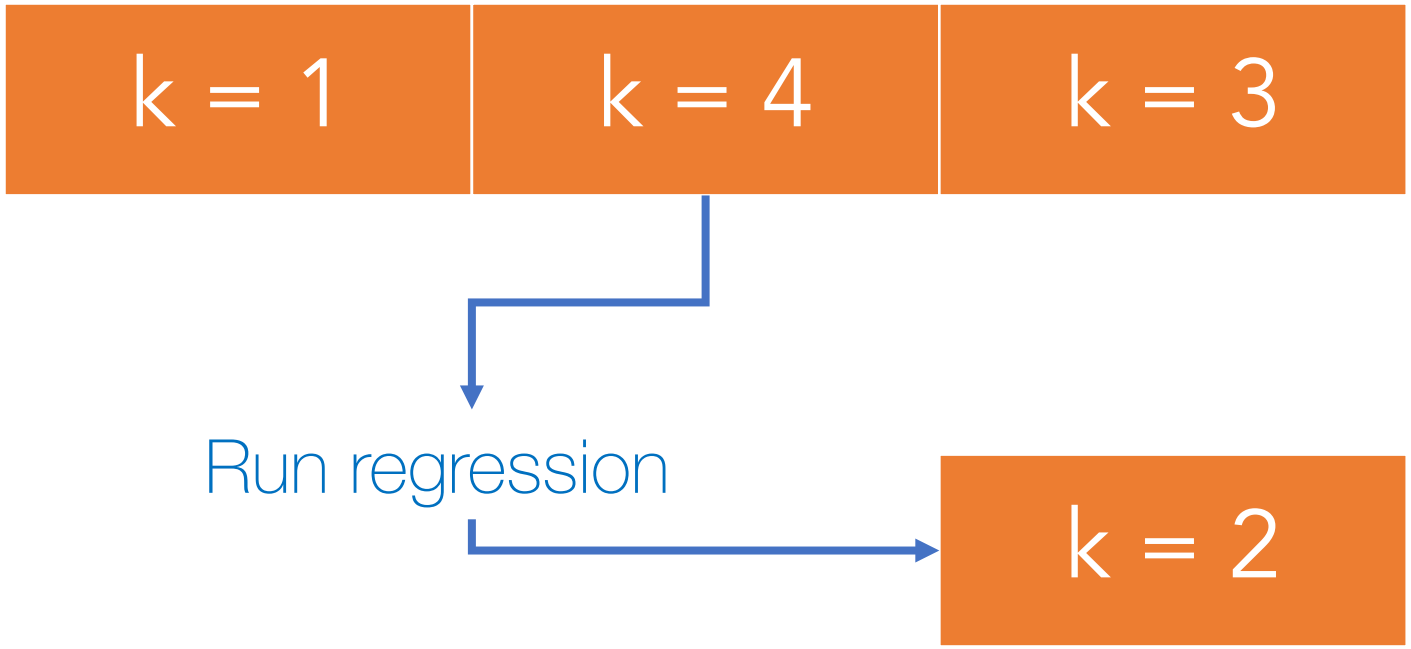
- *Run regression on k-1 samples*
- *Predict kth*
- *Iterate.*

Experimental Design: Regression



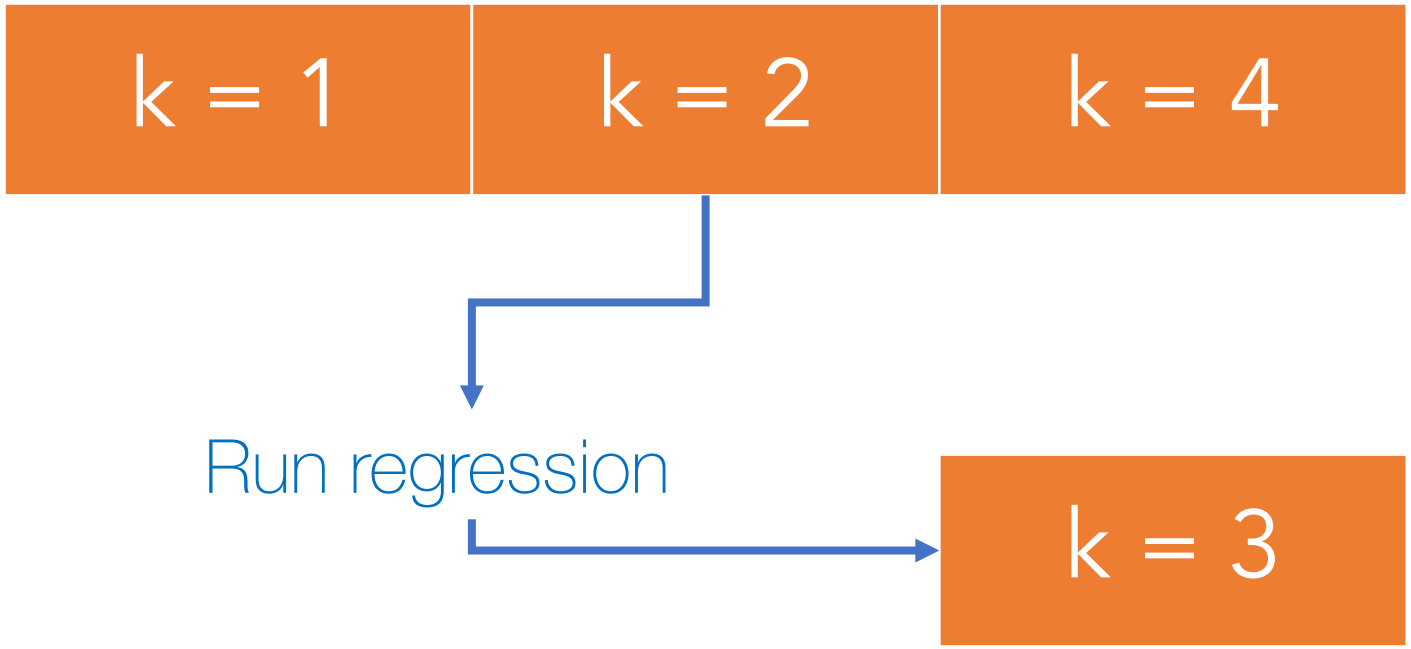
- *Run regression on k-1 samples*
- *Predict kth*
- *Iterate.*

Experimental Design: Regression



- *Run regression on k-1 samples*
- *Predict kth*
- *Iterate.*

Experimental Design: Regression

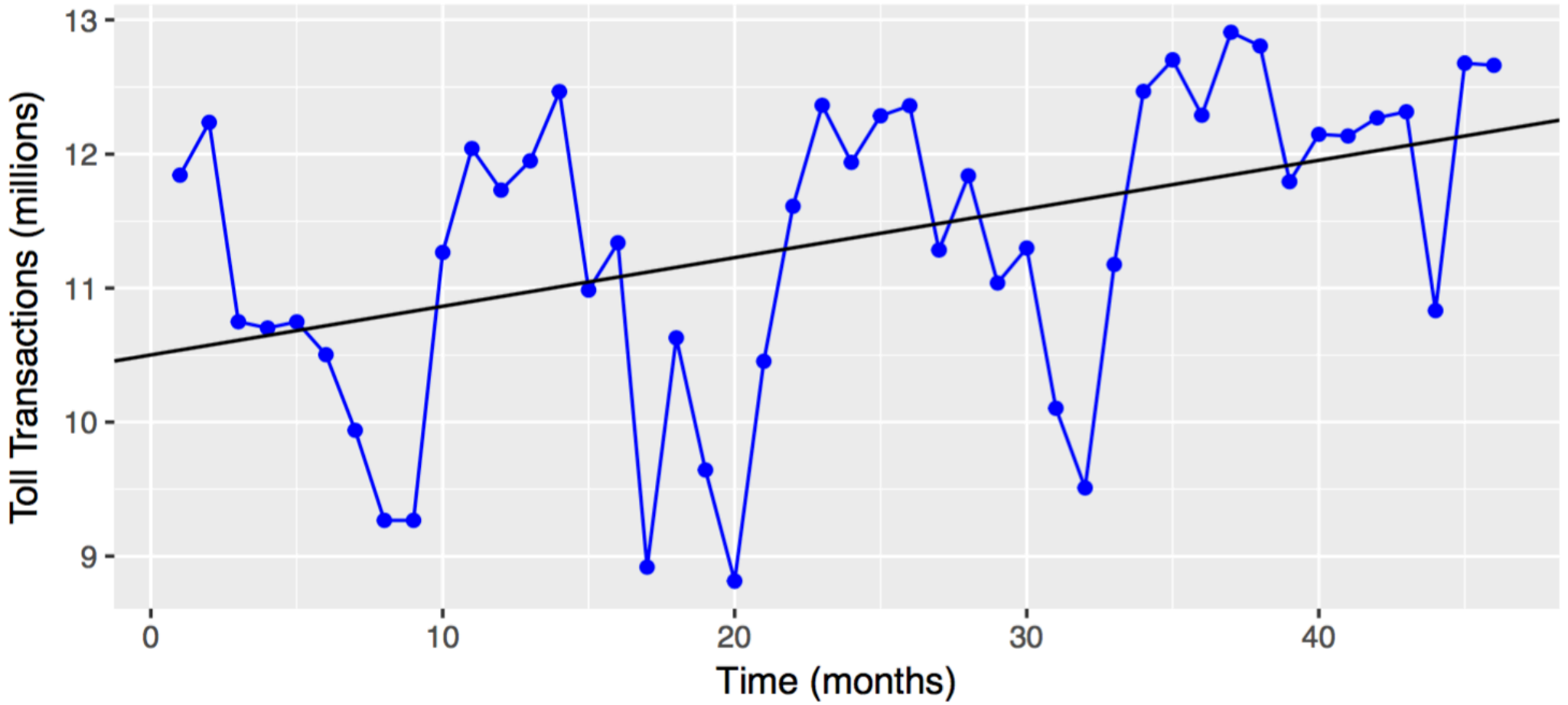


- *Run regression on k-1 samples*
- *Predict kth*
- *Iterate.*

Roadmap

- Motivation
- Supervised Learning
- OLS
- <break>
- KNN
- Homework assignment

$$\text{transactions} = 10.501 + 0.036 \times \text{months}$$



$$\text{intercept} = \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

$$\text{transactions} = 10.501 + 0.036 \times \text{months}$$

$$\text{slope} = \hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

Start

Training model

$$y_i = w_0 x_{i,0} + w_1 x_{i,1} + \ldots + w_k x_{i,k} + \epsilon_i$$

Objective: Minimize MSE or SSE via gradient descent

$$MSE = \frac{1}{n}\sum_{i}^{n}(y_i - \hat{y}_i)^2 \qquad SSE = \sum_{i=0}^{n}(y_i - \hat{y})^2$$

Evaluate Trained Model

$$R^2 = 1 - \frac{SSE}{TSS} \qquad \text{RMSE} = \sigma = \sqrt{\frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{n}}$$

$$\text{TSS} = \sigma^2(y) = \sum_{i=1}^{n}(y_i - \hat{y})^2$$

Prediction: Results in a linear equation that produces Y-hat

$$\hat{y}_i = w_0 + w_1 x_{i,1} + \ldots + w_k x_{i,k}$$

If Error is acceptable?

N

Y

Done!

Benchmark error

$$\text{MAPE} = \frac{100}{n}\sum_{i=1}^{n}\left|\frac{\hat{y}_i - y_i}{y_i}\right|$$

$$\text{RMSE} = \sigma = \sqrt{\frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{n}}$$

Coefficients

$$y_i = w_0 x_{i,0} + w_1 x_{i,1} + \ldots + w_k x_{i,k} + \epsilon_i$$

W or B are coefficients. W1 to Wk are marginal estimates.

$$\text{transactions} = 10.501 + 0.036 \times \text{months}$$

W1 = 0.036 million more transactions per month
Each W has an associated standard error

R-squared

$$R^2 = 1 - \frac{SSE}{TSS}$$

R-squared provides the proportion of variation explained by the model.

$$\text{TSS} = \sigma^2(y) = \sum_{i=1}^{n}(y_i - \hat{y})^2$$

Variance of Y

$$SSE = \sum_{i=0}^{n}(y_i - \hat{y})^2$$

Error between prediction and actual.

Error

$$\mathrm{RMSE} = \sigma = \sqrt{\frac{\sum_{i=1}^{n} (\hat{y}_i - y_i)^2}{n}}$$

Most common method of evaluating average error of predictions, but is not easily interpretted.

$$\mathrm{MAPE} = \frac{100}{n} \sum_{i=1}^{n} \left| \frac{\hat{y}_i - y_i}{y_i} \right|$$

Mean absolute Percentage Error is a normalized measure of error

OLS Assumptions

i. *Linear in parameters*
ii. *Sample is random.*
iii. *E(e|X) = 0*
    - *"Mean of errors should equal 0"*
iv. *No multi-collinearity – there needs to be variability*
v. *Var(e|X) =* $\boldsymbol{\sigma^2}$ *and Cov(*$\boldsymbol{e_t e_{t-1} \mid X) = 0}$
    - *Errors need to be homoscedastic without autocorrelation*
vi. *Errors should be normally distributed*

OLS Good/Bad

i. Good if talkable interpretation is needed
ii. Good if there are relatively few data variables
iii. Good if all relationships are linear

iv. Bad if too many variables are available
v. Bad if there are non-linear and interactions in the data

Related models

i. Non-linear least squares
ii. Robust regression – Median Absolute Deviation
iii. Quantile regression
iv. AutoRegressive Integrated Moving Average (ARIMA)

Coded example

Roadmap

- Motivation
- Supervised Learning
- OLS
- <break>
- KNN
- Homework assignment

Roadmap

- Motivation
- Supervised Learning
- OLS
- <break>
- KNN
- Homework assignment

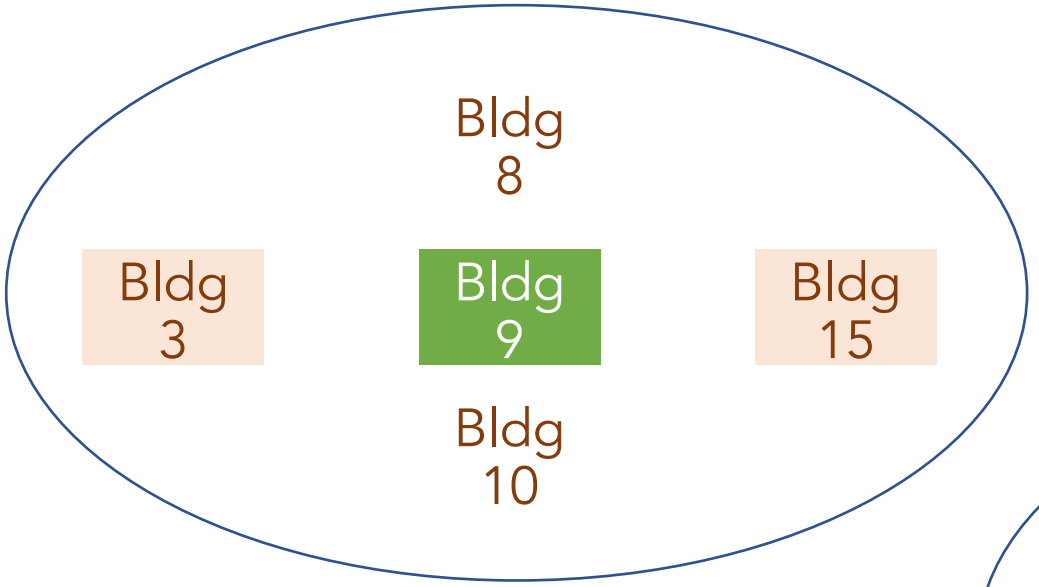
Given buildings on a grid, how would you predict whether the green buildings will catch fire?

Bldg 1	Bldg 7	Bldg 13	Bldg 19	Bldg 25	Bldg 31
Bldg 2	Bldg 8	Bldg 14	Bldg 20	Bldg 26	Bldg 32
Bldg 3	Bldg 9	Bldg 15	Bldg 21	Bldg 27	Bldg 33
Bldg 4	Bldg 10	Bldg 16	Bldg 22	Bldg 28	Bldg 34
Bldg 5	Bldg 11	Bldg 17	Bldg 23	Bldg 29	Bldg 35
Bldg 6	Bldg 12	Bldg 18	Bldg 24	Bldg 30	Bldg 36

From the visual, we can see spatial clusters of activity. Perhaps there isn't a grand unifying formula for fires…

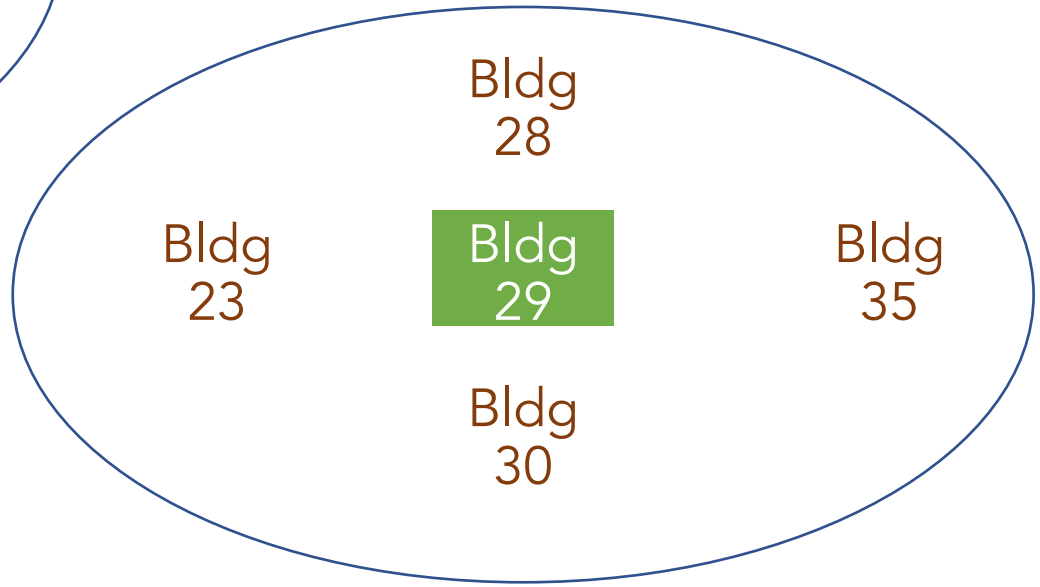
Bldg 1	Bldg 7	Bldg 13	Bldg 19	Bldg 25	Bldg 31
Bldg 2	Bldg 8	Bldg 14	Bldg 20	Bldg 26	Bldg 32
Bldg 3	Bldg 9	Bldg 15	Bldg 21	Bldg 27	Bldg 33
Bldg 4	Bldg 10	Bldg 16	Bldg 22	Bldg 28	Bldg 34
Bldg 5	Bldg 11	Bldg 17	Bldg 23	Bldg 29	Bldg 35
Bldg 6	Bldg 12	Bldg 18	Bldg 24	Bldg 30	Bldg 36

What if the 4 nearest buildings provide the most information?

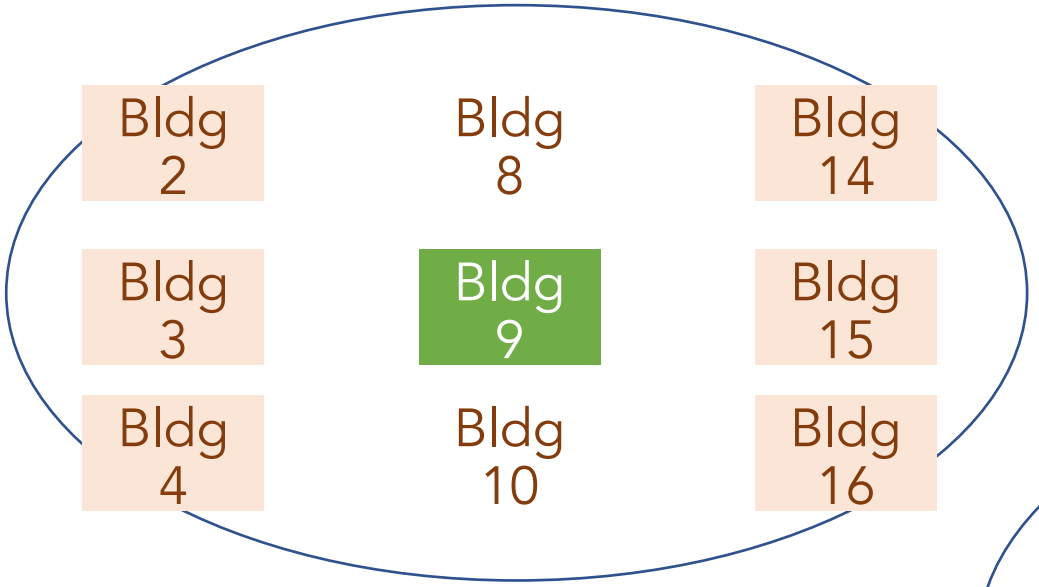


3 bldgs with fire means a 50% chance that Bldg 9 will catch fire

No buildings caught fire, meaning 0% chance that Bldg 29 will catch fire

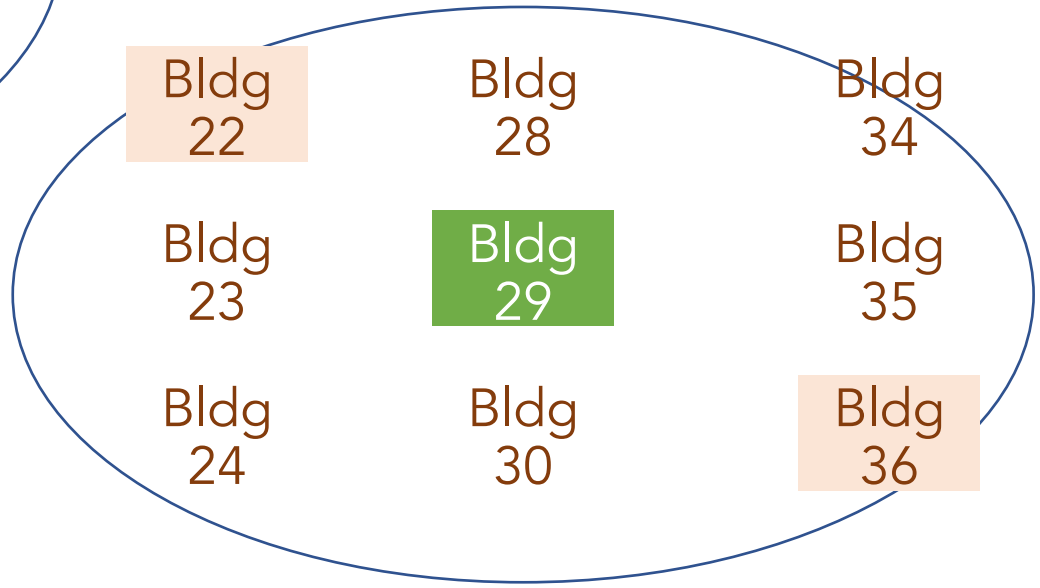


What if the 8 nearest buildings provide the most information?



6 bldgs with fire means
a 75% chance that
Bldg 9 will catch fire

2 bldgs caught fire,
meaning 25% chance that
Bldg 29 will catch fire

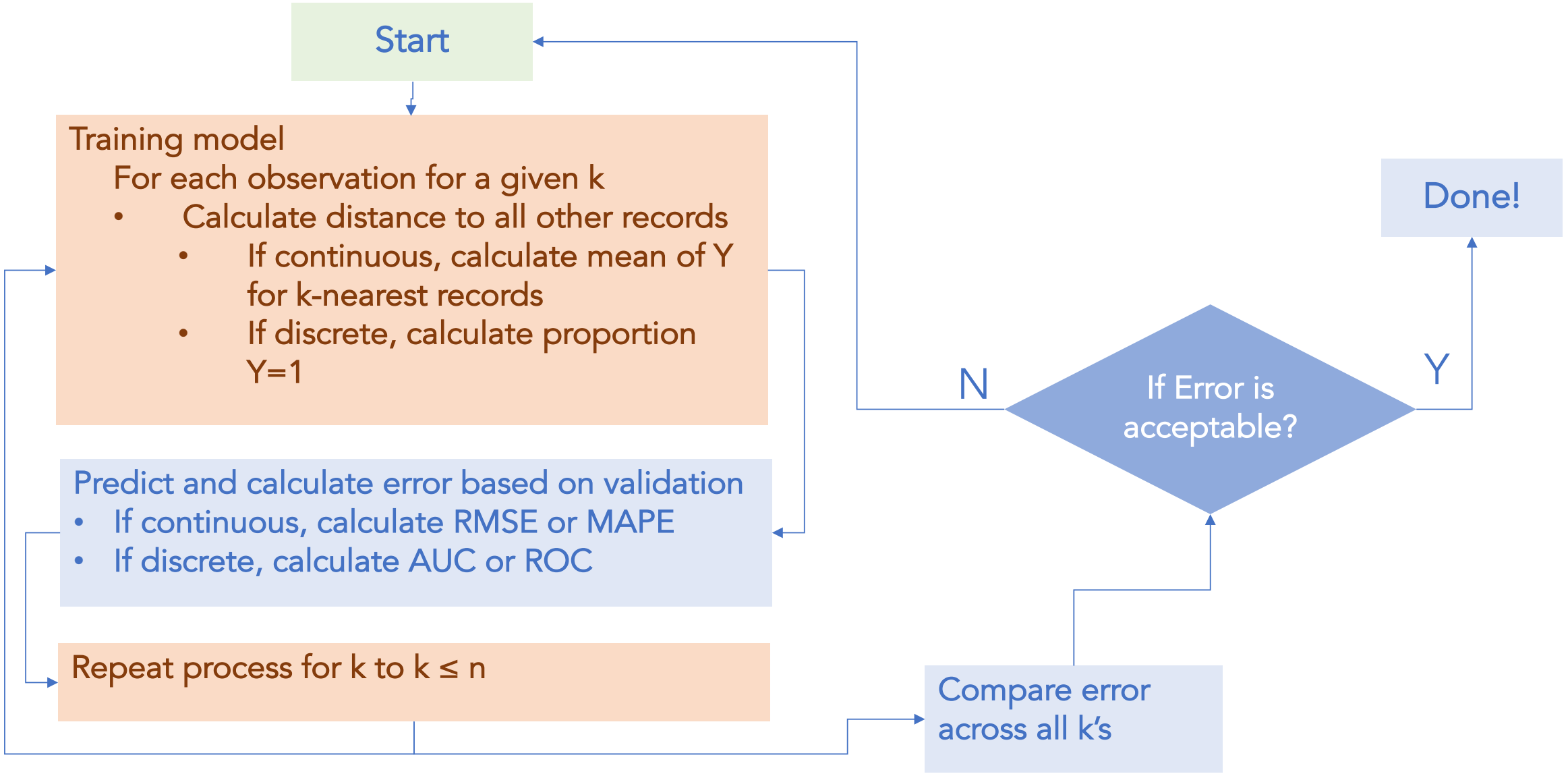


Which is better if both buildings actually caught fire?: k = 4 or k = 8?

	Correctly Predicted
k = 4	1 of 2
k = 8	2 of 2

kNNs assume that observations that are closer together are more related. Relatedness can be represented by the k-number of nearest records as measured by distance of variables.

$$\text{distance} = \sqrt{\sum (x_{ij} - x_{0j})^2}$$


Start
Training model
For each observation for a given k
• Calculate distance to all other records
• If continuous, calculate mean of Y for k-nearest records
• If discrete, calculate proportion Y=1
Predict and calculate error based on validation
• If continuous, calculate RMSE or MAPE
• If discrete, calculate AUC or ROC
Repeat process for k to k ≤ n
Compare error across all k's
If Error is acceptable?
N
Y
Done!

KNN Assumptions

- All input features need to be scaled similarly.
- All input features have equal weight.
- Uncommon assumption: While KNN is a non-parametric method, high collinearity should be kept to a minimum such that redundant variation is not disproportionately represented.

KNN Good/Bad

i. Good for low dimensional data (few variables)
ii. Good for when no theory exists
iii. Good for imputation of missing values

iv. Bad for high dimensional data as each observation needs to be processed
v. Bad if data contains both discrete and continuous
vi. Bad if you want an interpretation

Coded example

Roadmap

- Motivation
- Supervised Learning
- OLS
- <break>
- KNN
- Homework assignment