Deconstructing Data Science

David Bamman, UC Berkeley

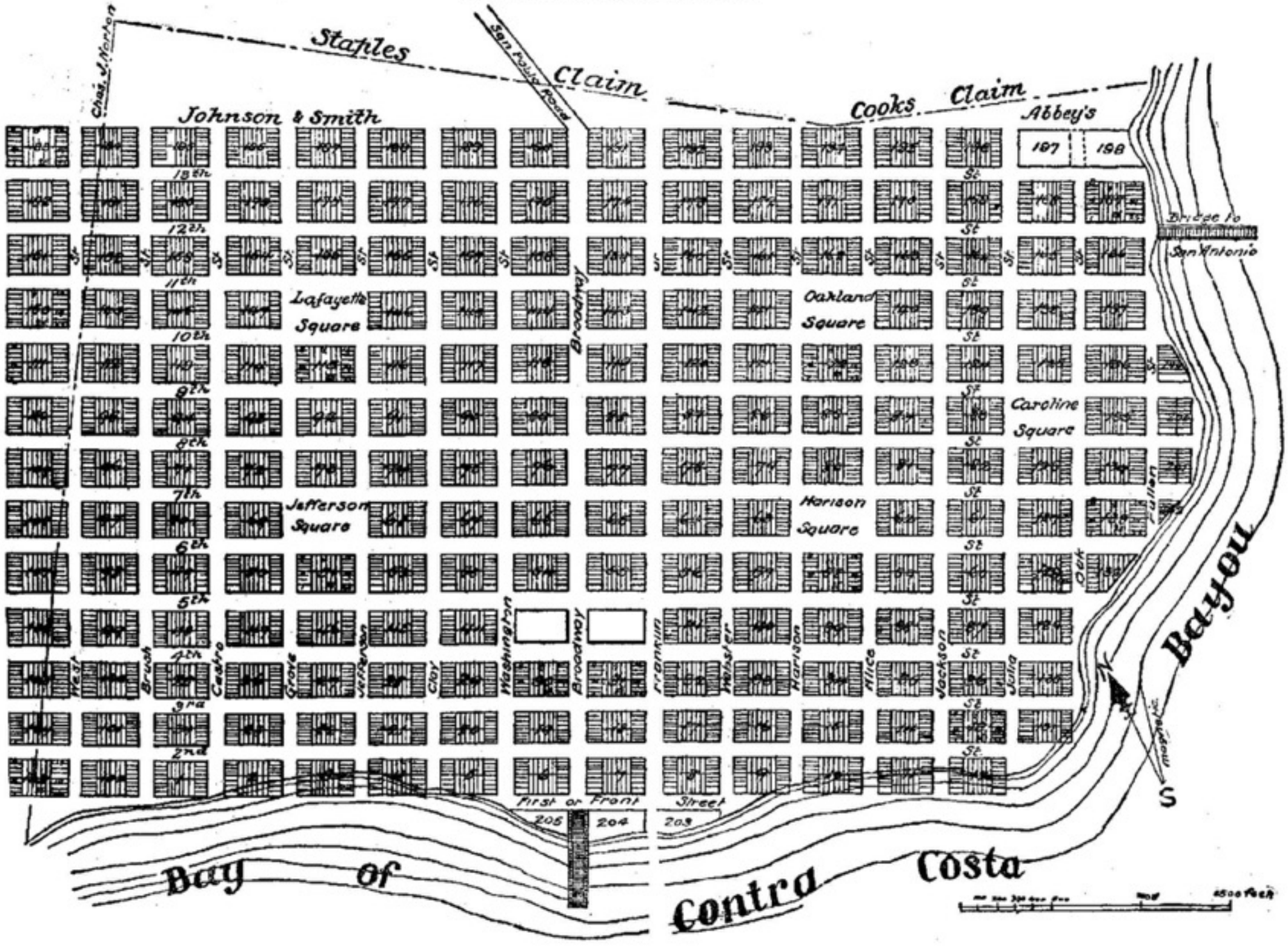
Info 290
Lecture 17: Distance models

Mar 28, 2016

A COMPLETE MAP OF OAKLAND.

RESPECTFULLY DEDICATED TO THE CITIZENS OF OAKLAND,

By J. Kellersberger, Surveyor.

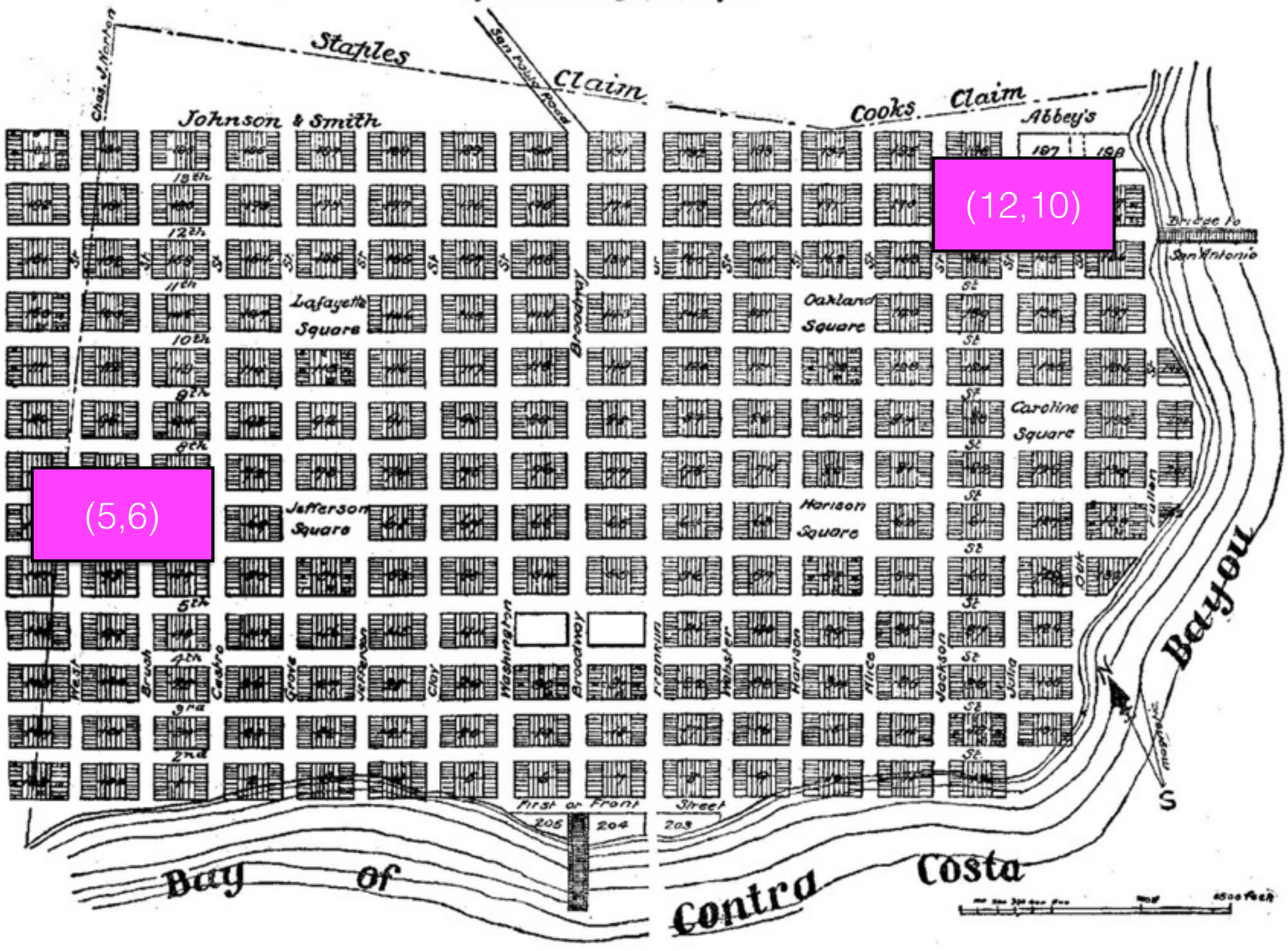


1853

A COMPLETE MAP OF OAKLAND.

RESPECTFULLY DEDICATED TO THE CITIZENS OF OAKLAND,

By J. Kellersberger, Surveyor.



1853

Feature	Jefferson Square	Oakland Square	Lafayette Square	Harrison Square
x	5	12	5	12
y	6	10	10	6

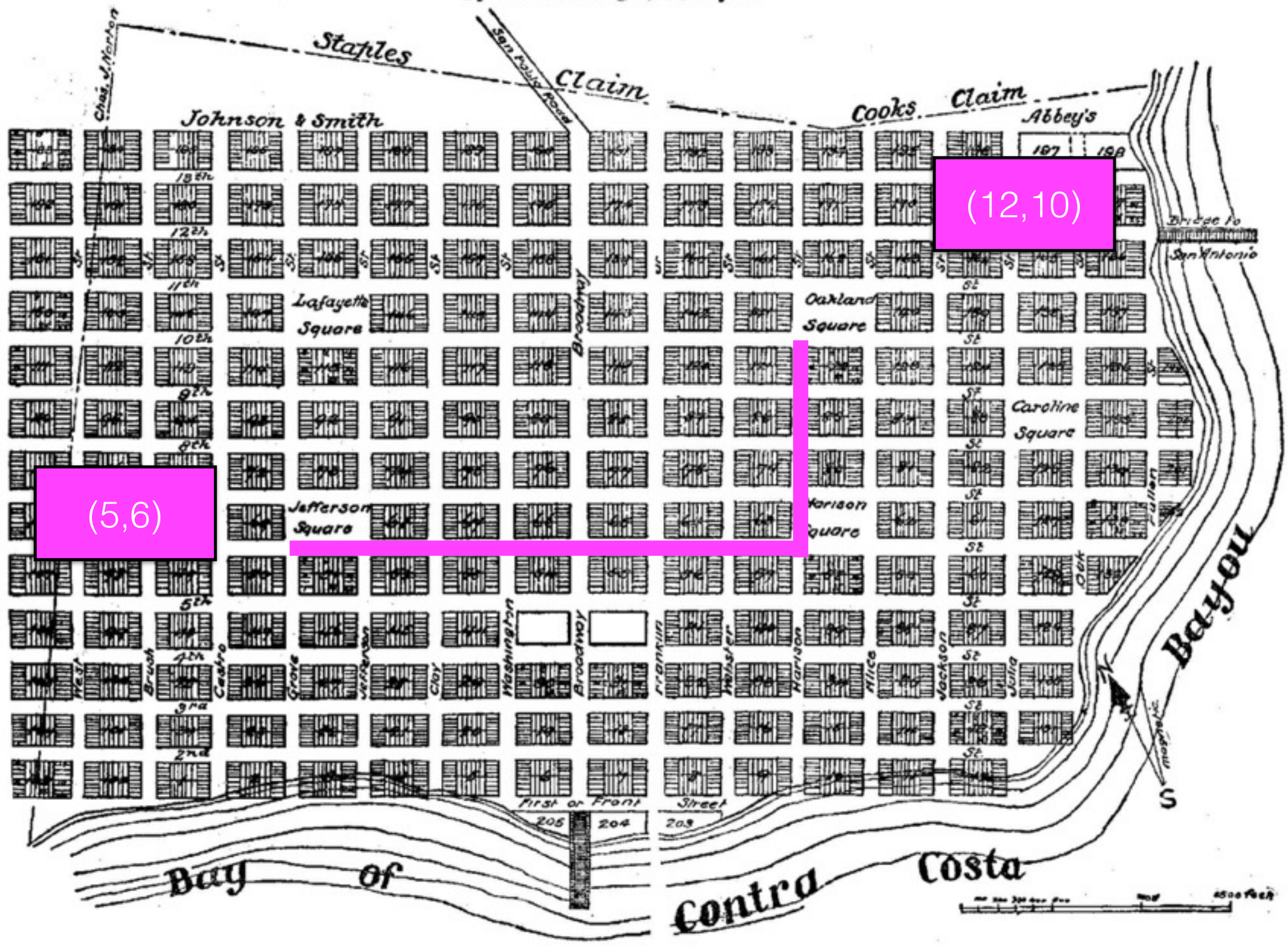
“Manhattan distance”

$$\sum_{i=1}^{F} |x_i - y_i|$$

A COMPLETE MAP OF OAKLAND.

RESPECTFULLY DEDICATED TO THE CITIZENS OF OAKLAND,

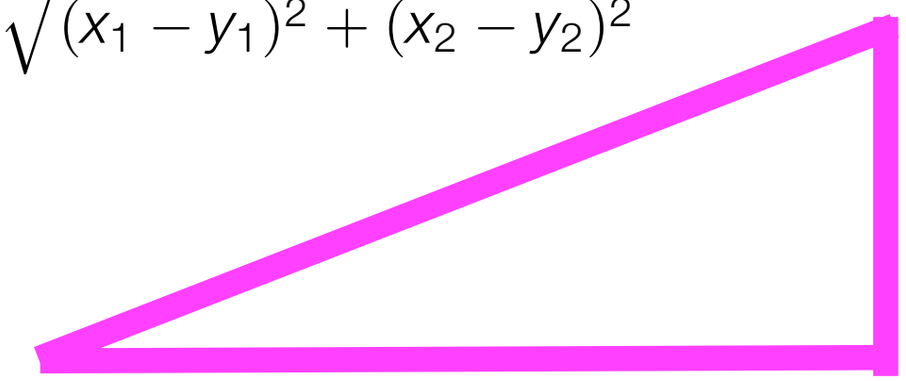
By J. Kellersberger, Surveyor.



1853

(12,10)

$$\sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2}$$



(5,6)

$|x_2 - y_2|$

$|x_1 - y_1|$

$$a^2 + b^2 = c^2$$

$$\sqrt{a^2 + b^2} = c$$

Euclidean distance

$$\sqrt{\sum_{i=1}^{F} (x_i - y_i)^2}$$

$$= \left(\sum_{i=1}^{F} (x_i - y_i)^2 \right)^{1/2}$$

1-norm (Manhattan)

$$\left(\sum_{i=1}^{F} |x_i - y_i|^1 \right)^{1/1}$$

2-norm (Euclidean)

$$\left(\sum_{i=1}^{F} |x_i - y_i|^2 \right)^{1/2}$$

p-norm

$$\left(\sum_{i=1}^{F} |x_i - y_i|^p \right)^{1/p}$$

0-norm
(Hamming)

$$\left(\sum_{i=1}^{F} |x_i - y_i|^0\right)^{1/0} = \sum_{i=1}^{F} I[x_i \neq y_i]$$

∞-norm
(Chebyshev)

$$\left(\sum_{i=1}^{F} |x_i - y_i|^\infty\right)^{1/\infty} = \max_i |x_i - y_i|$$

Metrics

$d(x, y) \geq 0$ distances are not negative

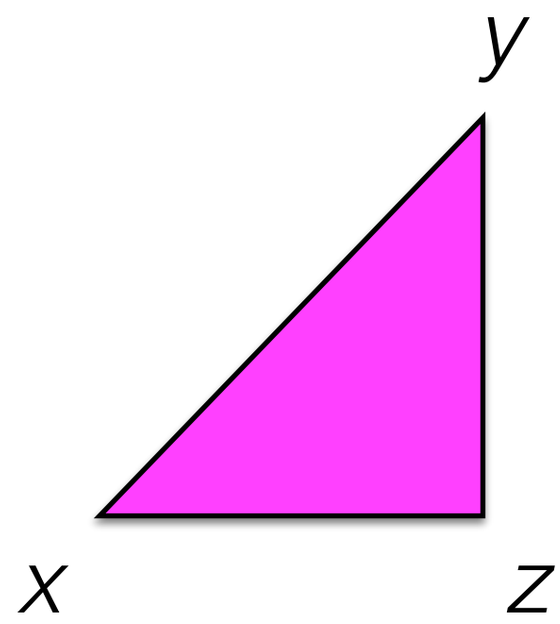
$d(x, y) = 0$ iff $x = y$ distances are positive, except for identity

$d(x, y) = d(y, x)$ distances are symmetric

Metrics

$d(x,y) \leq d(x,z) + d(z,y)$ triangle inequality

a detour to another point z
can't shorten the "distance"
between x and y



Feature	x1	x2	x3
follow clinton	1	0	0
follow trump	0	1	1
“benghazi”	0	0	1
negative sentiment + “benghazi”	0	1	0
“illegal immigrants”	0	1	1
“republican” in profile	0	0	0
“democrat” in profile	0	0	0
self-reported location = Berkeley	1	0	0

K-nearest neighbors

- Supervised classification/regression
- Make prediction by finding the closest k data points and
  - predicting the majority label among those k points (classification)
  - predicting their average of those k points (regression)

KNN Classification

Let $\mathcal{N}(x_i)$ be the K-nearest neighbors to x_i

$$P(Y = j \mid x) = \frac{1}{K} \sum_{x_i \in \mathcal{N}(x)} I[y_i = j]$$

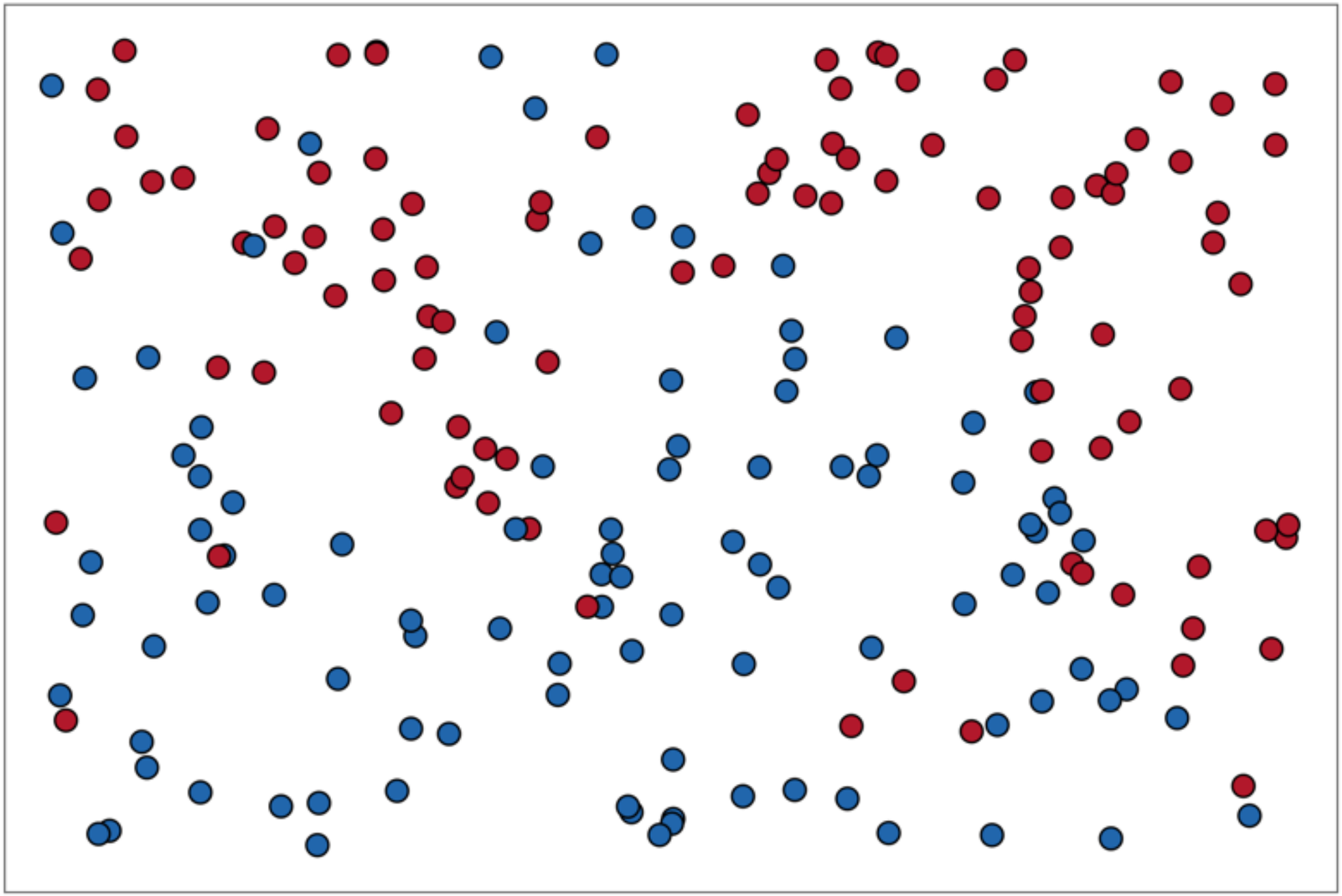
(Pick the value of Y with the highest probability)

KNN Regression

Let $\mathcal{N}(x_i)$ be the K-nearest neighbors to x_i

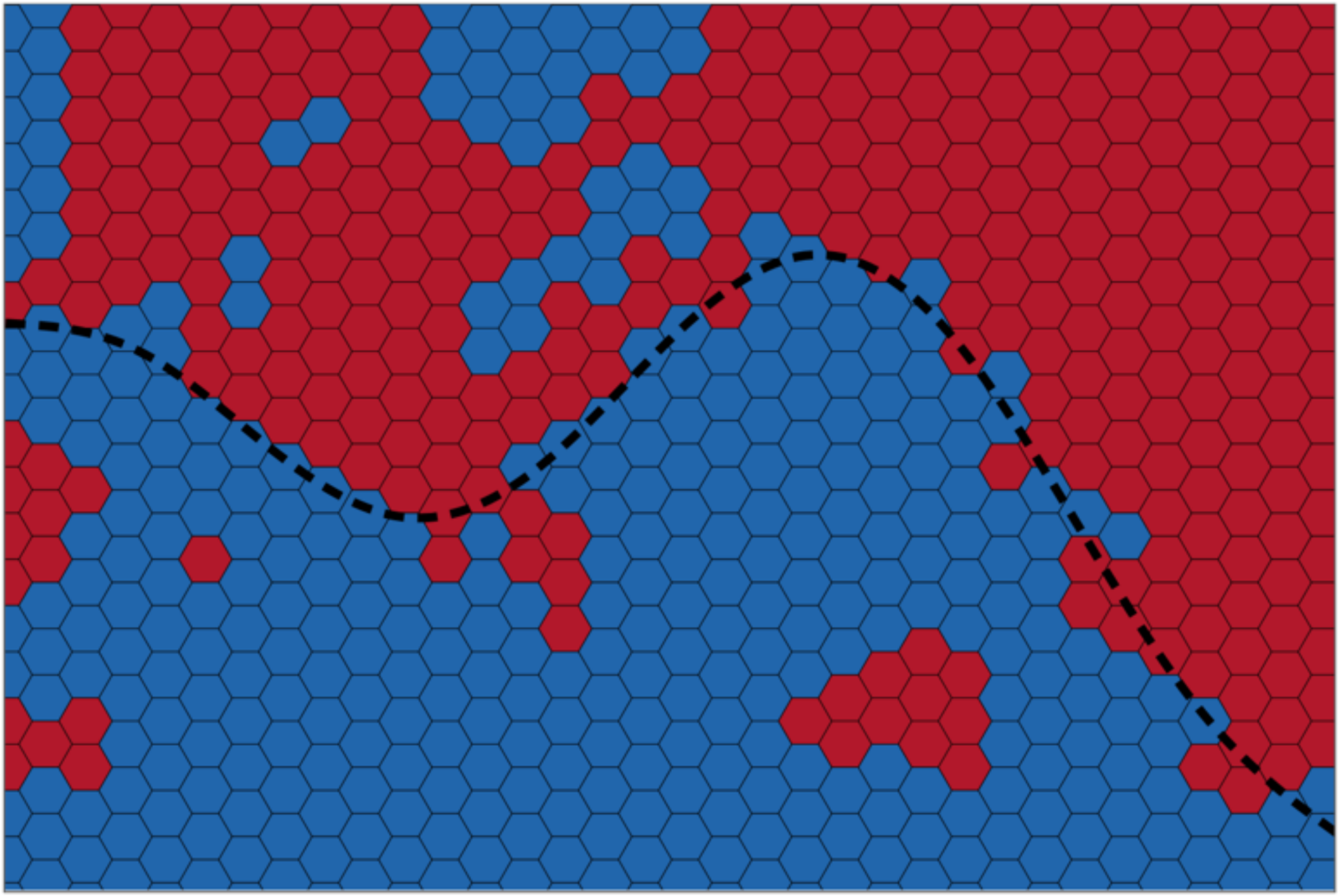
$$\hat{y}_i = \frac{1}{K} \sum_{x_j \in \mathcal{N}(x_i)} y_j$$

Data



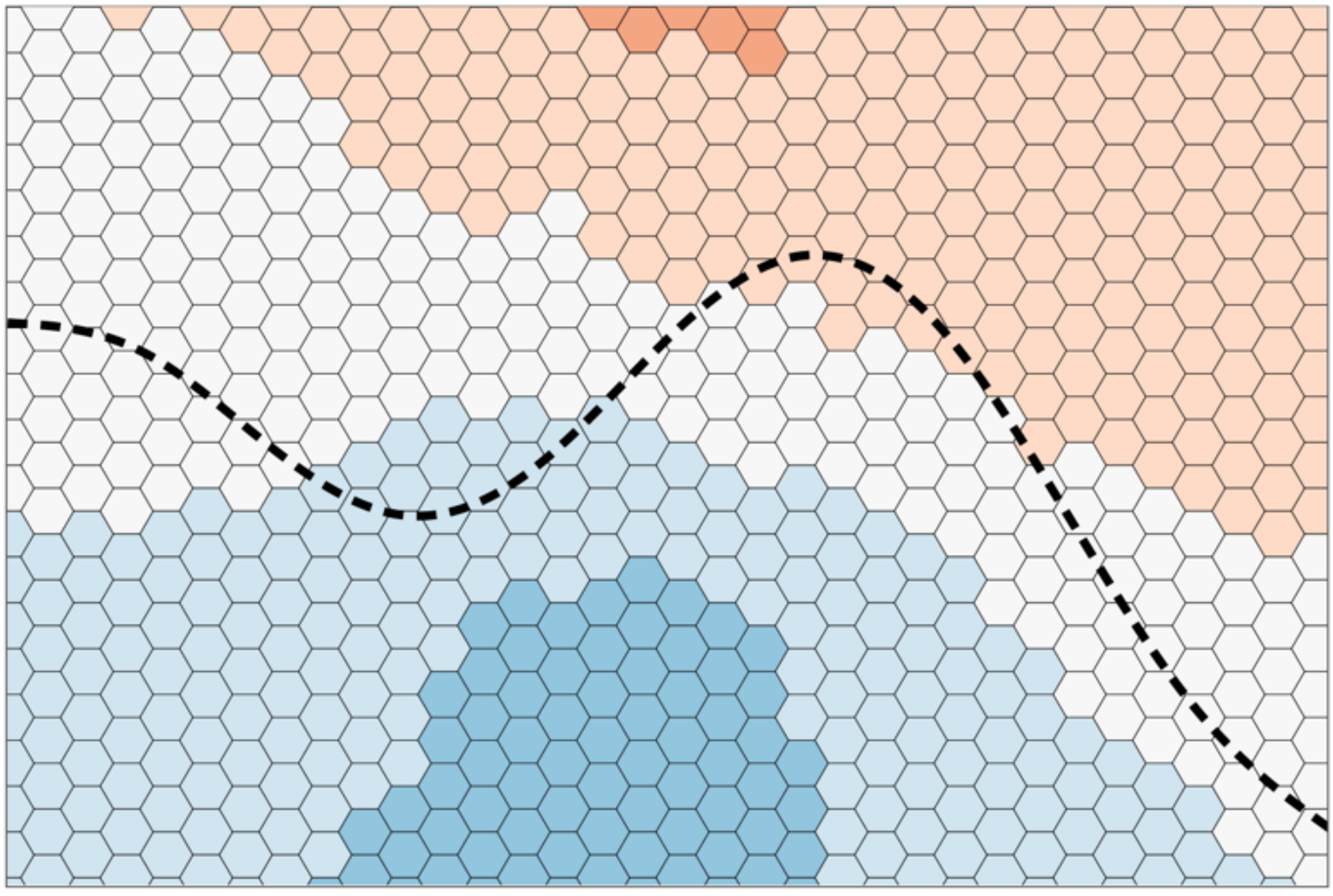
http://scott.fortmann-roe.com/docs/BiasVariance.html

K=1



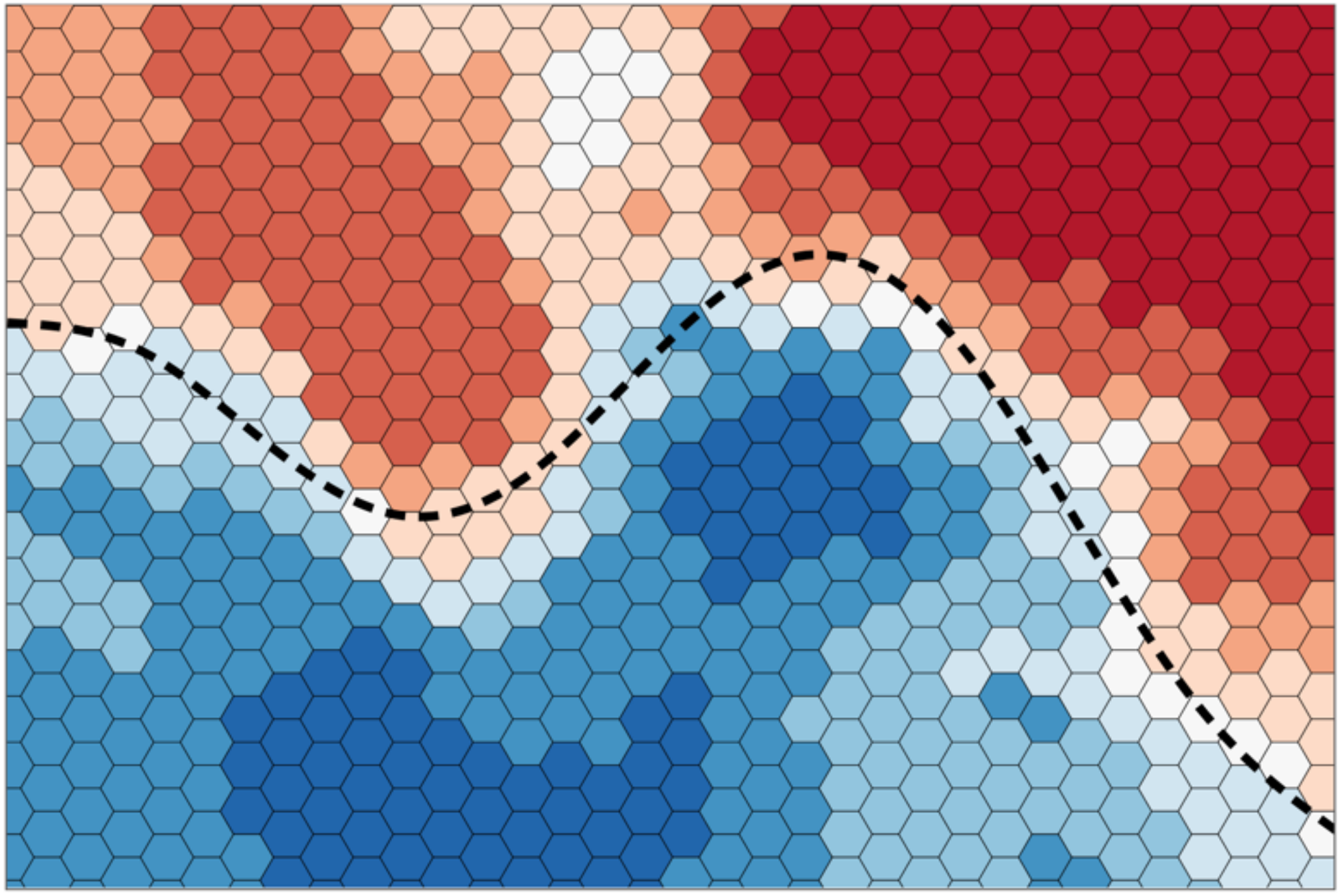
http://scott.fortmann-roe.com/docs/BiasVariance.html

K=100



http://scott.fortmann-roe.com/docs/BiasVariance.html

K=12



http://scott.fortmann-roe.com/docs/BiasVariance.html

KNN

- Properties:
    - Linear/Nonlinear?
    - Complexity of training/testing?
    - Overfitting?
    - How to choose the best K?
    - Impact of data representation

Similarity

task	method	distance
classification/regression	KNN	euclidean, etc.
classification/regression	SVM	kernel
duplicate detection		
search		

Relevance (IR)

- Similarity as an end of its own is a different paradigm from what we've been considering so far (classification, regression, clustering).

task	x	y
KNN classification/ regression	documents	genres
duplicate detection	documents	

Duplicate detection

PRESIDENT OBAMA MAKES HIS FINAL 4 PICKS; KANSAS AS CHAMPS

WASHINGTON (AP) -- President Barack Obama has made his final NCAA Tournament call in office: Rock Chalk, champions.

Obama picked Kansas, Texas A&M, North Carolina and Michigan State to all reach the Final Four in a bracket he filled out for ESPN.

AP Photo/Pablo Martinez Monsivais

President Obama Makes His Final 4 Picks
abcnews.go.com/.../president-obama-makes-final-picks
2 days ago - **His choice might be an unpopular one arou**
hasn't correctly predicted the national champion since he p

President Obama picks KU basketball as c
m.kusports.com/.../president-obama-picks-ku-basketba
2 days ago - **His choice might be an unpopular one arou**
hasn't correctly predicted the national champion since he p

WKTV.com | President Obama makes his
www.wktv.com/.../President_Obama_makes_his_Final
2 days ago - **His choice might be an unpopular one arou**
hasn't correctly predicted the national champion since he p

President Obama makes his Final 4 picks;
www.kswo.com/.../president-obama-makes-his-final-4-
His choice might be an unpopular one around Kansas,
correctly predicted the national champion since he picked N

President Obama calls for Rock Chalk Cha
www.wibw.com/.../President-Obama-calls-for-Rock-Cha
2 days ago - **His choice might be an unpopular one arou**
hasn't correctly predicted the national champion since he p

President Obama makes his Final 4 picks;
https://www.artesianews.com/.../president-obama-mak
5 days ago - **His choice might be an unpopular one arou**
hasn't correctly predicted the national champion since he p

Duplicate document detection

- What are the data points we're comparing?
- How do we represent each one?
- How do we measure "similarity"
- Evaluation?

Computational concerns

- Two sources of complexity:
  - Dimensionality of the feature space (every document in represented by a vocabulary of 1M word) [minhashing]
  - Number of documents in collection to compare (4.64 billion web pages) [locality sensitive hashing]

Feature	x1	x2	x3
the	1	1	1
and	1	1	1
obama	1	1	0
supreme	1	0	0
court	1	0	1
kansas	0	1	1
ncaa	0	1	1
four	1	1	1

Jaccard Similarity

x1	x2	x3
1	1	1
1	1	1
1	1	0
1	0	0
1	0	1
0	1	1
0	1	1
1	1	1

number of features in **both** X and Y

$$\frac{|X \cap Y|}{|X \cup Y|}$$

number of features in **either** X and Y

Text Reuse

We were many times weaker than his splendid, lacquered machine, so that I did not even attempt to outspeed him. O lente currite noctis equi! O softly run, nightmares!

Nabokov, *Lolita*

Text reuse detection

- What are the data points we're comparing?
- How do we represent each one?
- How do we measure "similarity"
- Evaluation?

Information retrieval

Google conrad heart of darkness

All Books Images Videos Shopping More

About 479,000 results (0.46 seconds)

Heart of Darkness - Wikipedia, the free encyclop
https://en.wikipedia.org/wiki/**Heart_of_Darkness** ▾ Wikipedia
Heart of Darkness (1899) is a novella by Polish-British novelist Jo
voyage up the Congo River into the Congo Free State, in the heart
Joseph Conrad - Kurtz - Disambiguation - Léon Rom

SparkNotes: Heart of Darkness
www.sparknotes.com/lit/**heart**/ ▾ SparkNotes ▾
Heart of Darkness. Joseph **Conrad** ... Buy the print **Heart of Dar**
BN.com ... Order **Heart of Darkness** and Selected Short Fiction a
Part 1 - Part 2 - Part 3 - Context

Heart of Darkness, by Joseph Conrad - Project
www.gutenberg.org/files/219/219-h/219-h.htm ▾ Project Gute
The Project Gutenberg EBook of **Heart of Darkness**, by Joseph C
for the use of anyone anywhere at no cost and with almost no rest

Heart of Darkness - Shmoop
www.shmoop.com › Literature ▾
We really can't say it better than Joseph **Conrad** himself. **Heart of**
story of a journalist who becomes manager of a station in the (Afric

Heart of Darkness at a Glance - Cliffs Notes
www.cliffsnotes.com/.../**heart-of-darkness/heart-of-darkness**
Joseph **Conrad's Heart of Darkness** retells the story of Marlow's

Information retrieval

- What are the data points we're comparing?
- How do we represent each one?
- How do we measure "similarity"
- Evaluation?

Cosine Similarity

x1	x2	x3
1	1	1
1	1	1
1	1	0
1	0	0
1	0	1
0	1	1
0	1	1
1	1	1

$$cos(x, y) = \frac{\sum_{i=1}^{F} x_i y_i}{\sqrt{\sum_{i=1}^{F} x_i^2} \sqrt{\sum_{i=1}^{F} y_i^2}}$$

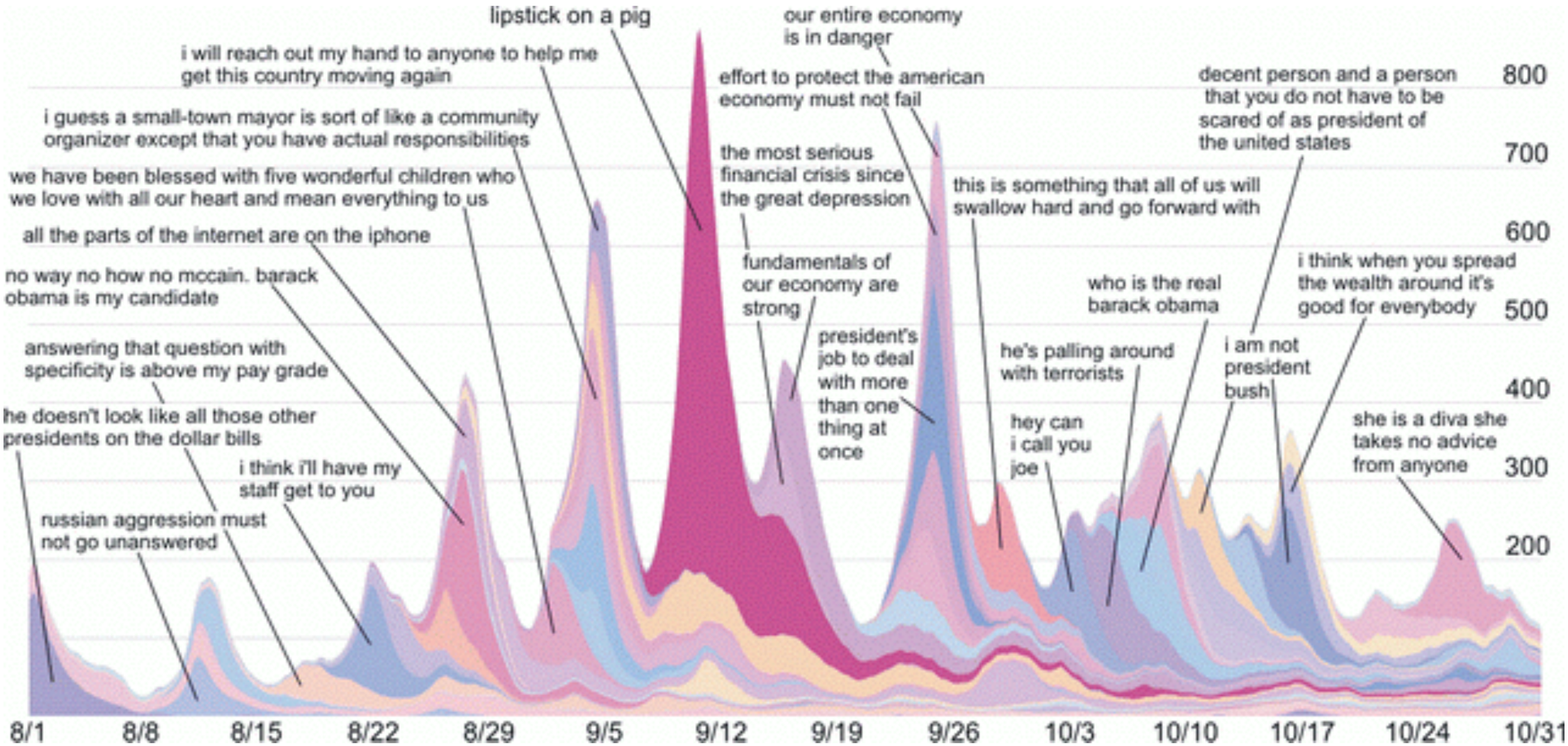
- Euclidean distance measures the magnitude of distance between two points
- Cosine similarity measures their orientation

- Often weighted by TF-IDF to discount the impact of frequent features.

Modern IR

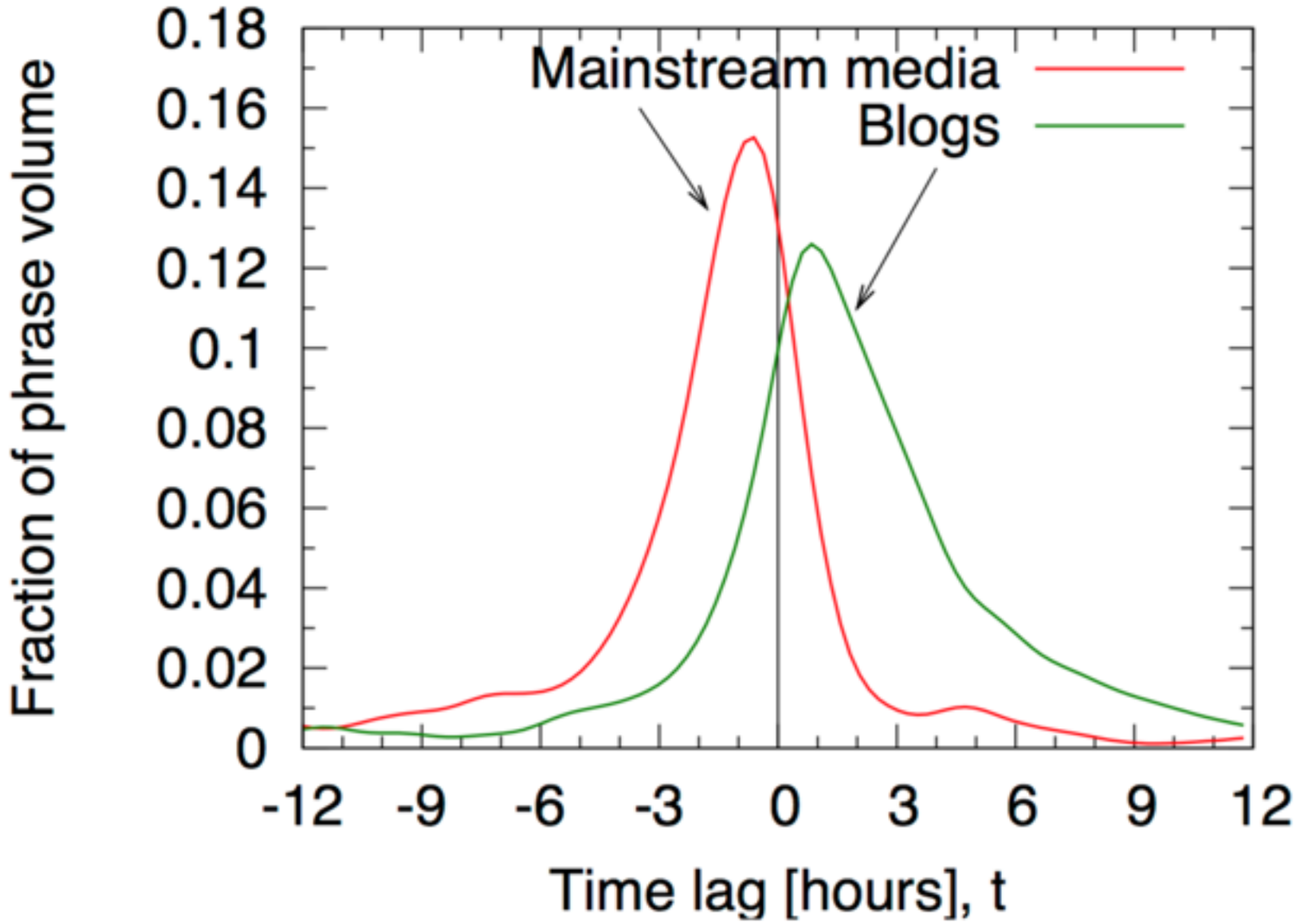
- Modern IR accounts for much more information than document similarity
    - Prominence/reliability of document (PageRank)
    - Geographic location
    - Search query history
- This can become a supervised problem to learn how to map these more elaborate features of a query/session to the search ranking. How do we represent our data?

Meme tracking



J. Leskovec et al. (2009), "Meme-tracking and the Dynamics of the News Cycle"

Meme tracking



J. Leskovec et al. (2009), "Meme-tracking and the Dynamics of the News Cycle"

Meme tracking

Rank	Lag [h]	Reported	Site
1	-26.5	42	hotair.com
2	-23	33	talkingpointsmemo.com
4	-19.5	56	politicalticker.blogs.cnn.com
5	-18	73	huffingtonpost.com
6	-17	49	digg.com
7	-16	89	breitbart.com
8	-15	31	thepoliticalcarnival.blogspot.com
9	-15	32	talkleft.com
10	-14.5	34	dailykos.com
16	-14	54	blogs.abcnews.com
30	-11	32	uk.reuters.com
34	-11	72	cnn.com
40	-10.5	78	washingtonpost.com
48	-10	53	online.wsj.com
49	-10	54	ap.org

Table 1: How quickly different media sites report a phrase.

J. Leskovec et al. (2009), "Meme-tracking and the Dynamics of the News Cycle"

http://mybinder.org/repo/dbamman/dds