COSMOS 2016

The Difficulty of Science and Data Visualization

Luis F. Campos

Department of Statistics
Harvard University

July 22, 2016

Outline

1. A crisis in science? (Reproducibility)
   - p-values, huh?
   - Publication Bias
   - p-Hacking
   - Some Hope for the Future
2. Data Visualization
   - Why is visualizing data important?
   - Facebook example
   - Vernacular example
     - k-NN to predict where someone is from
   - Interactive Data Visualizations

A crisis in science? (Reproducibility)

Reading (decreasing order of required-ness):

- ▸ Nature: Statisticians issue warning over misuse of P values
- ▸ FiveThirtyEight: Science Isnt Broken
- ▸ Nature: Scientific method: Statistical errors
- ▸ Marginal Revolution: Results Free Review

p-value: what it is and what it is not

An example:

- You want to see if a new drug works
- Give it to 50 people, see that it works for 49 people.
- If we assume that the drug does not work, this is very unlikely.
- If instead it worked for only 10 people.
- Assuming that the drug does not work, 10 is fairly likely.

What it is:
p-value: If a certain hypothesis (drug doesn't work) is true, *and all other assumptions made are valid*, what is are the odds we see the results we saw.

- An implementation of the scientific process

p-value: what it is and what it is not

What it is:

p-value: If a certain hypothesis (drug doesn't work) is true, *and all other assumptions made are valid*, what is are the odds we see the results we saw.

What it is not:

- The probability of a hypothesis being true
- A replacement for science (science is hard)
- A justification for publication

▸ Nature: Statisticians issue warning over misuse of P values

Publication Bias

What is publication bias?

- Say researchers test 100 hypotheses

  e.g. test to see if a new drug/treatment works
- About 5% will be "statistically significant" even is none of them actually work
- These are usually the (only) ones we see in publication.

What's the issue?

- Because publications mean a lot to scientists, they are motivated to "hack" their studies to significance
  - p-Hacking
  - (Sometimes happens unintentionally as well)

▸ A potential solution to publication bias

A bit about p-Hacking

Science, what's the issue?

- Hard
- Expensive
- Time Consuming
- The best tool we have for reaching the truth

Why is it so hard to get a rigorous result?!

▸ Let's do an exercise

A p-Hacking Example

If you select

- Representatives and Senators to calculate "in office"
- GDP and stock-prices to measure a "better economy"

You'll find that when there are more Democrats in power, the economy is "significantly" better

If you select

- Presidents and Governors to calculate "in office"
- employment and inflation to measure a "better economy"

You'll find that when there are more Republicans in power, the economy is "significantly" better

A p-Hacking Example

- This shouldn't be interpreted as "Science is unreliable and fickle".
- Instead, "Science is harder than we give it credit for".

Hopefully going through this exercise gave you some things to look for in academic papers:

- Did they describe their data transparently?
- Did they describe their analysis in detail?
- Did they justify the choices they made? (e.g. why these variables?)

Some possible Solutions to this

Maybe we should "encourage" p-hacking?

- Of course, these decisions should be transparent
- Some call this pseudo-science
- This isn't science but that doesn't mean it's not meaningful

Maybe we should "encourage" retractions?

- Number of retractions are going up, but not fast enough
- Replication studies should be encouraged
- Should be part of the publishing/scientific process

▸ Reading Assignment 2

Some Hope for the Future

- "People want to prove something, and a negative result doesnt satisfy that craving"
- "Science is low-yield. Most experiments fail" but that's okay
- "Once an idea becomes fixed, its difficult to remove from the conventional wisdom"
- "Science is not a magic wand that turns everything it touches to truth"
- "Science isnt broken, its just more difficult than most of us realize"

Visualizing data

What are visualizations?

- Graphical representations of information.
- Tools to help us communicate and learn.
- Pretty pictures/interfaces.

Why is visualizing data important?

- Convey a large amount of information.
- Helps tell a story.
- Draws in the reader.

Facebook example

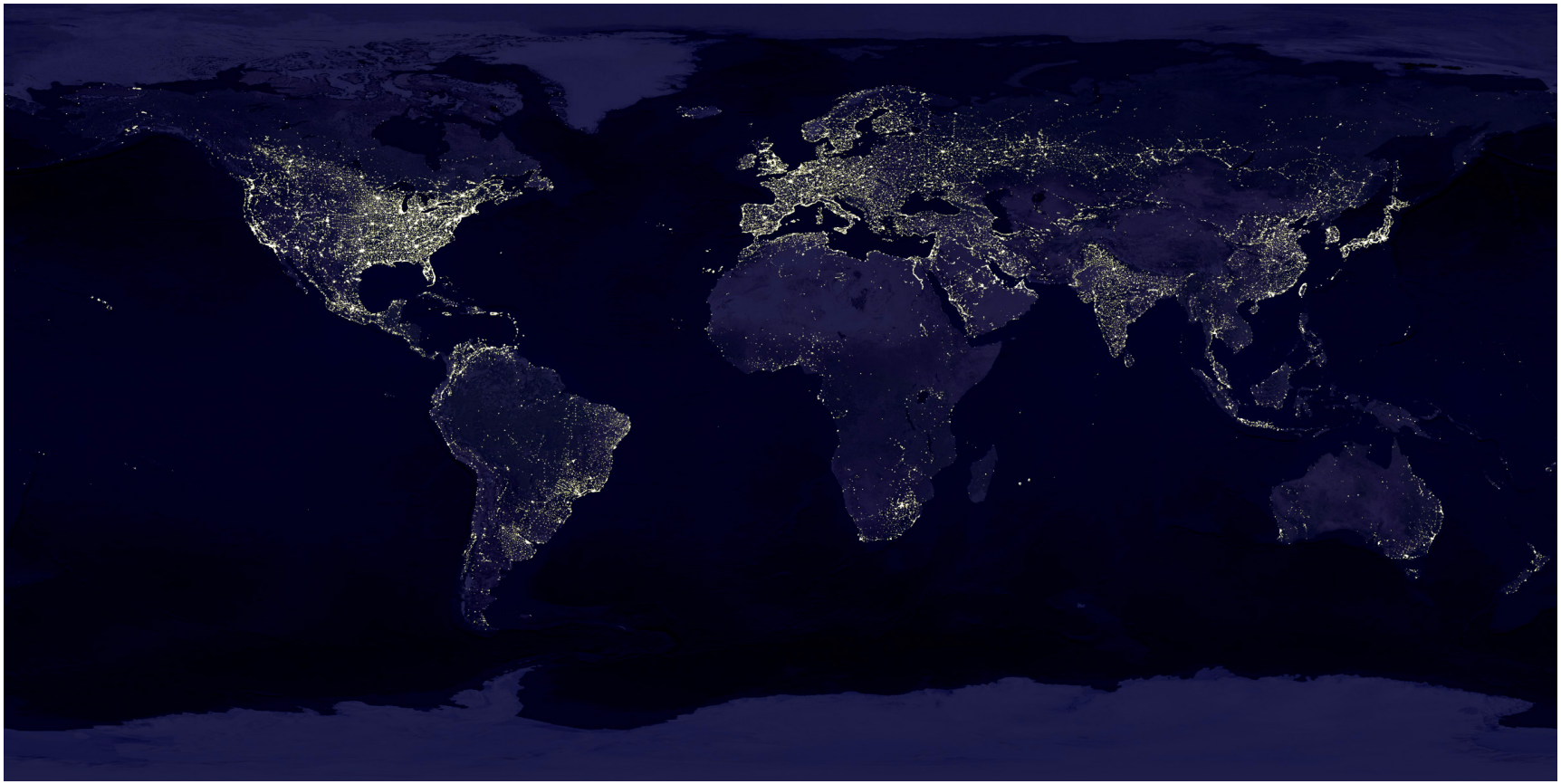
Facebook - 2010

- The line intensity from blue to white represents the number of connections between cities.

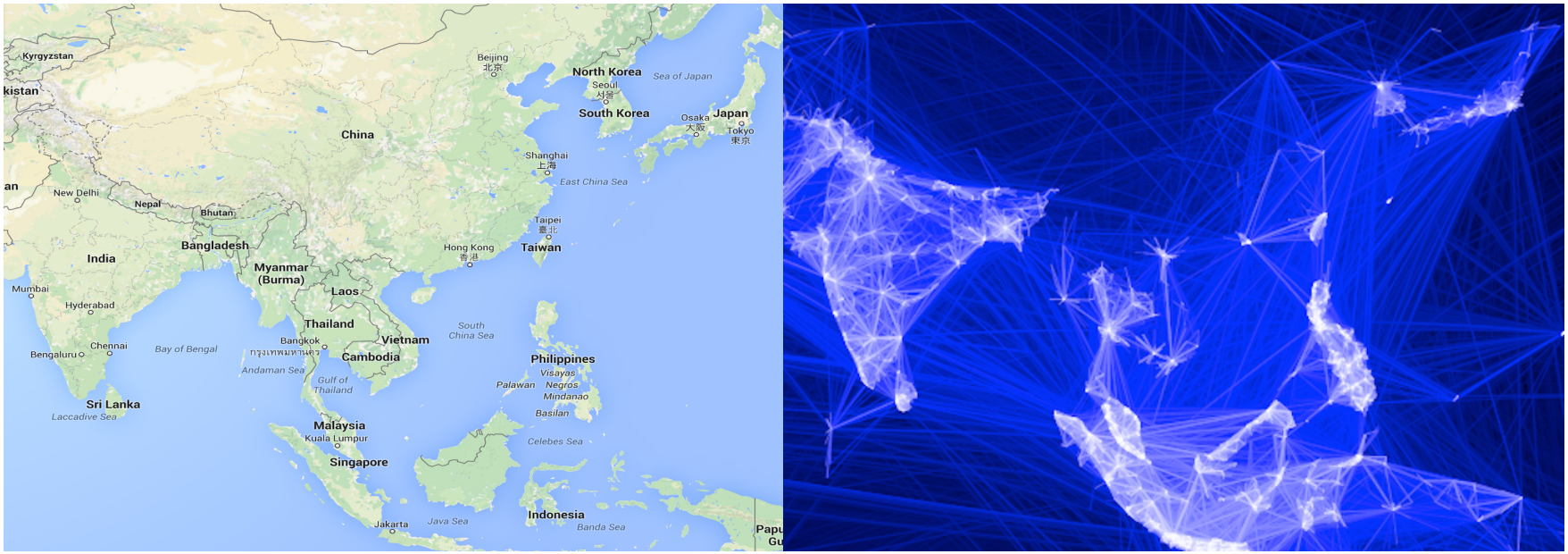
Facebook - 2013

- Most recent version I could find.

Facebook example



Facebook example



Do you notice anything interesting? Try finding the countries.

- India, Japan, Malaysia, the Philippines are all fully connected
- China is not prominent at all...

Facebook example

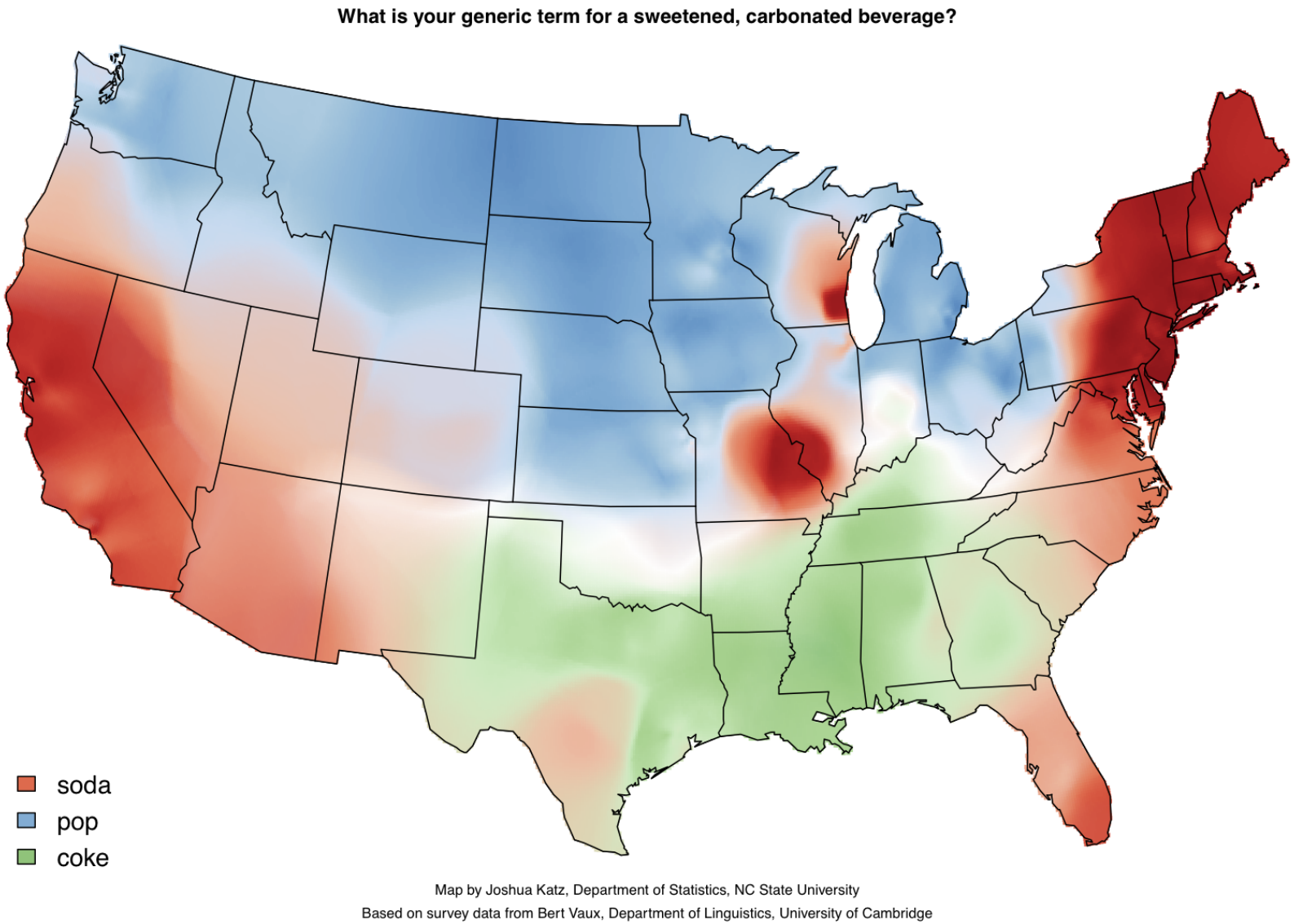
- Recall: the line intensity represents the number of connections between cities – that's all.
- We found geography.
- When data is organized in a simple and intuitive way you can explore and find other interesting things.

Vernacular across the U.S.

What do you call a sweetened carbonated beverage?

- Soda
- Pop
- Coke

Vernacular across the U.S.



Vernacular across the U.S.

We can ask other questions that might have different answers depending on where you're from

- what do you call the long sandwich that contains cold cuts, lettuce, and so on?
  - Sub, hoagie, Italian sandwich, etc.
- how do you pronounce "caramel"?
  - car-ml, carra-mel, either
- what word(s) do you use to address a group of two or more people?
  - you guys, you, y'all, you all

▸ More Questions

Vernacular across the U.S.

Vernacular across the U.S.

What are some natural questions?

- If we have how people in a particular area respond to a long list of questions, can we use this to see what cities are similar to each other?
- This might be important if you're traveling!
- It's also just fun.

Vernacular example - What does it mean to be similar?

Questions

1. what do you call a sweetened carbonated beverage?
   - soda $= 0$ pop $= 1$
2. what do you call the long sandwich that contains cold cuts, lettuce, and so on?
   - sub $= 0$ hoagie $= 1$
3. how do you pronounce "caramel"?
   - car-ml $= 0$ carra-mel $= 1$
4. what word(s) do you use to address a group of two or more people?
   - you guys $= 0$ y'all $= 1$

Vernacular example - What does it mean to be similar?

Person 1: soda, sub, carra-mel, you guys
Person 2: soda, sub, car-ml, you guys
Person 3: pop, hoagie, carra-mel, y'all

How does this translate to numbers?

1. 0, 0, 1, 0
2. 0, 0, 0, 0
3. 1, 1, 1, 1

Who is the most similar in their responses?
Who is most different in their responses?

Most Similar: $(1 <-> 2)$
Most Different: $(2 <-> 3)$

Vernacular example - What does it mean to be similar?

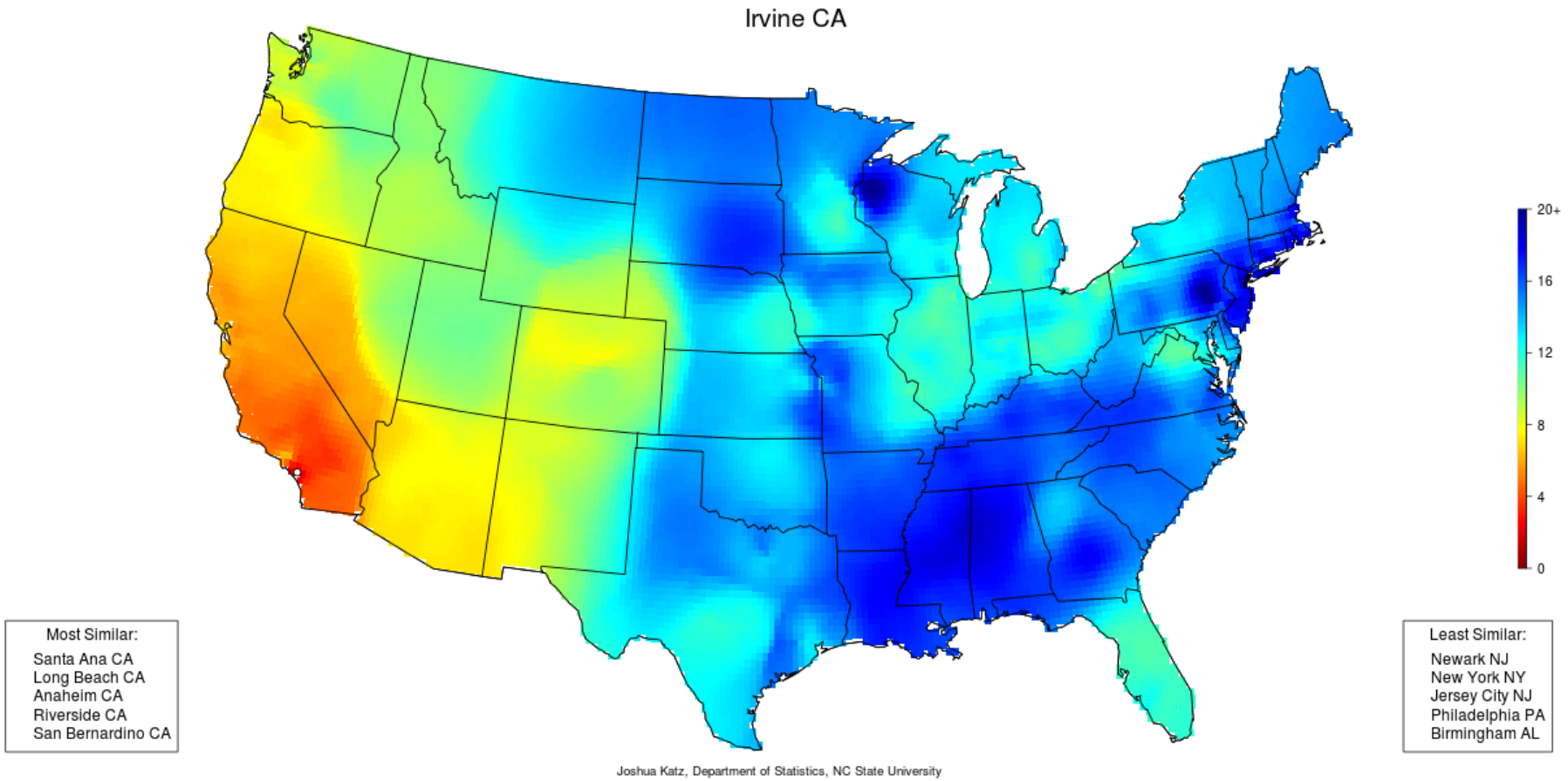
We can use the distance between these two people to assess similarity

$$d(p_1, p_2) = |0 - 0| + |0 - 0| + |1 - 0| + |0 - 0| = 1$$
$$d(p_2, p_3) = |0 - 1| + |0 - 1| + |0 - 1| + |0 - 1| = 4$$
$$d(p_1, p_3) = |0 - 1| + |0 - 1| + |1 - 1| + |0 - 1| = 3$$

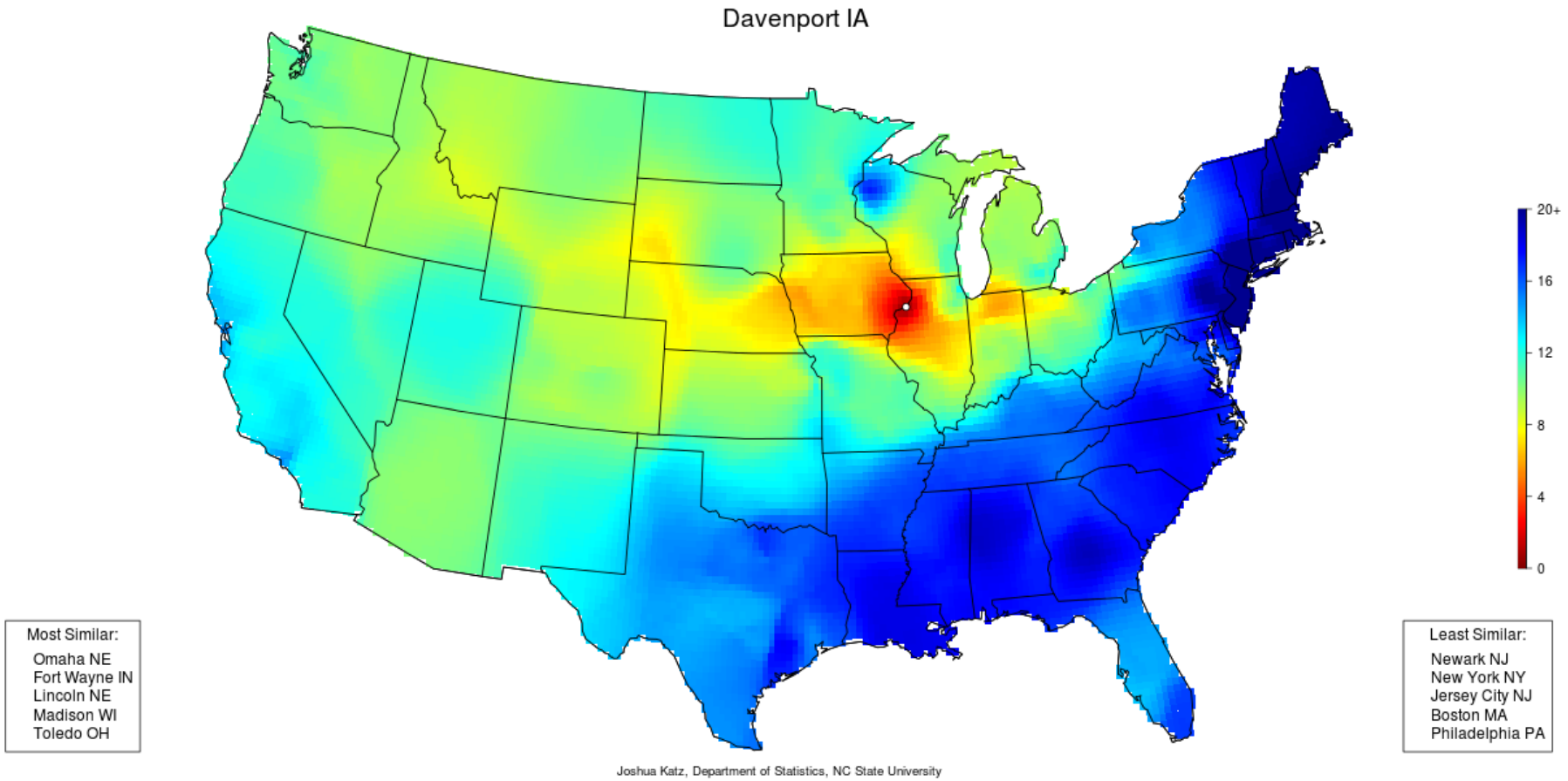
This is called the "Manhattan" distance. Can anyone guess why?

- If we do this for all the people in the survey, we can see which people are similar!
- If we want to find out what regions are similar to a particular person (p_1), we can look at all $d(p_1, p_n)$

What cities are most similar to Irvine, CA?



What cities are most similar to Davenport, IA?



Using Vernacular to predict where someone is from

Let's say you have a new person who takes the survey, how would you predict where they are from?

- calculate $d(p_{new}, p_n)$, find k people who are the most similar (have the lowest Euclidian distance $(d(p_{new}, p_n))$)
- Predict the region to be the majority of the k peoples regions

This is called **k-Nearest Neighbors** algorithm!

Where else do they use k-NN algorithm?

Amazon.com

Recommendations for You in Cell Phones & Accessories

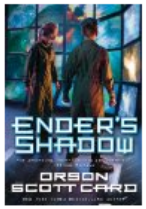
SPIGEN SGP iPhone 5 Premium Aluminum...
(103)
Why recommended?

New Trent, 10W 5V/2A Dual Port...
(439)
Why recommended?

SPIGEN SGP SGP09548 GLAS.tR Premium...
(487)
$34.99
Why recommended?

› See more recommendations

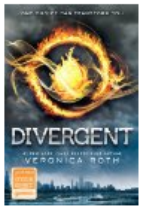
More Recommendations for You



Ender's Shadow (Ender, Book 5)
› Orson Scott Card
Paperback
(829)
~~$5.99~~ $5.39
Why recommended?

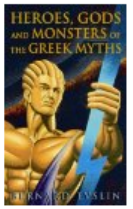
Belkin 6-by-9-inch Mouse Pad (Black)
(411)
~~$7.99~~ $3.65
Why recommended?

PowerGen Dual USB 3.1A 15w Travel...
(562)
~~$29.99~~ $14.99
Why recommended?



Divergent
› Veronica Roth
Paperback
(3,783)
~~$9.99~~ $5.49
Why recommended?

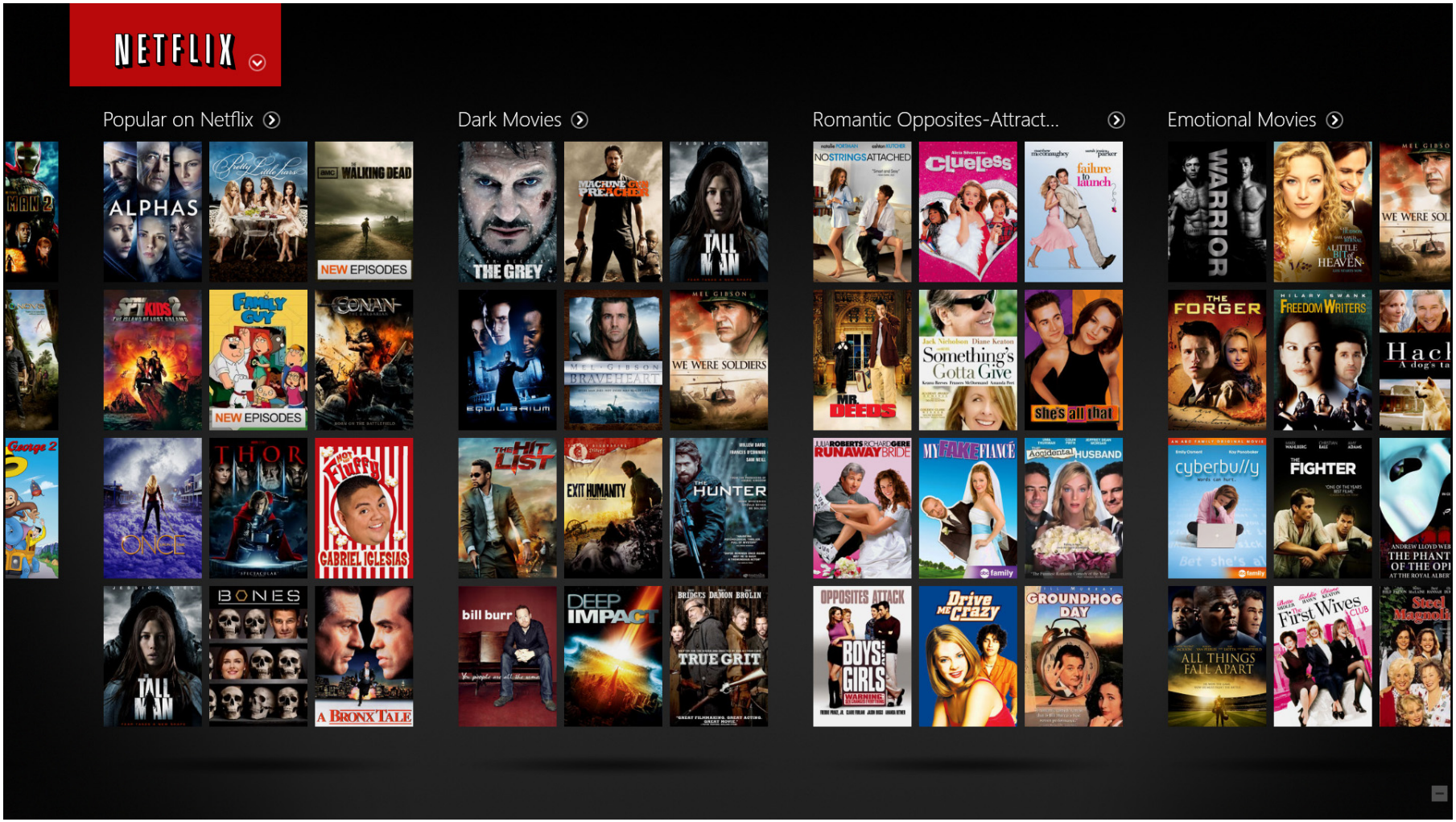
AC Power Adapter US Extension Wall...
Volex
(337)
~~$16.99~~ $7.01
Why recommended?



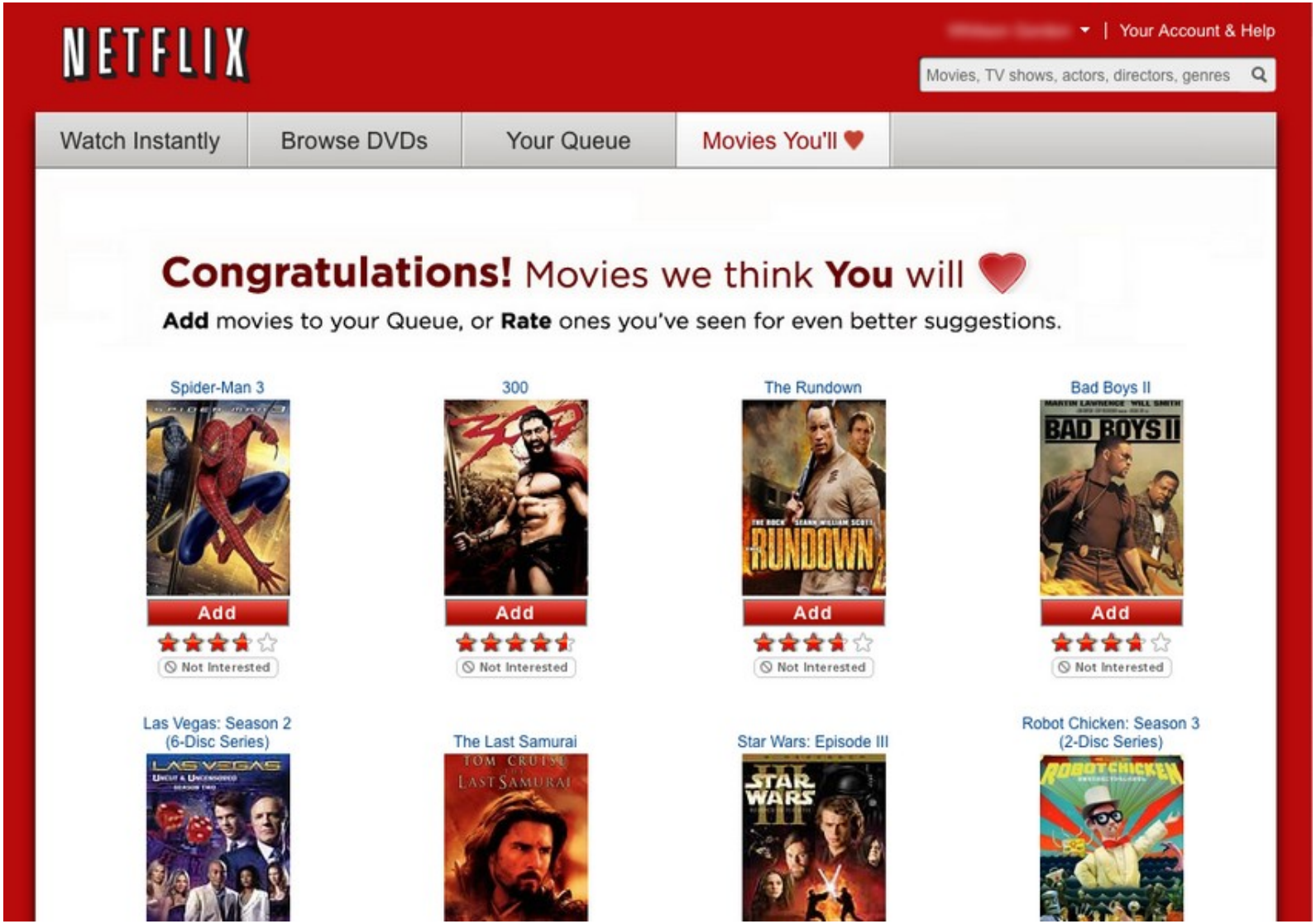
Heroes, Gods and Monsters of the...
› Bernard Evslin
Mass Market Paperback
(68)
~~$6.99~~ $6.29
Why recommended?

› See more recommendations

Where else do they use k-NN algorithm?



Where else do they use k-NN algorithm?

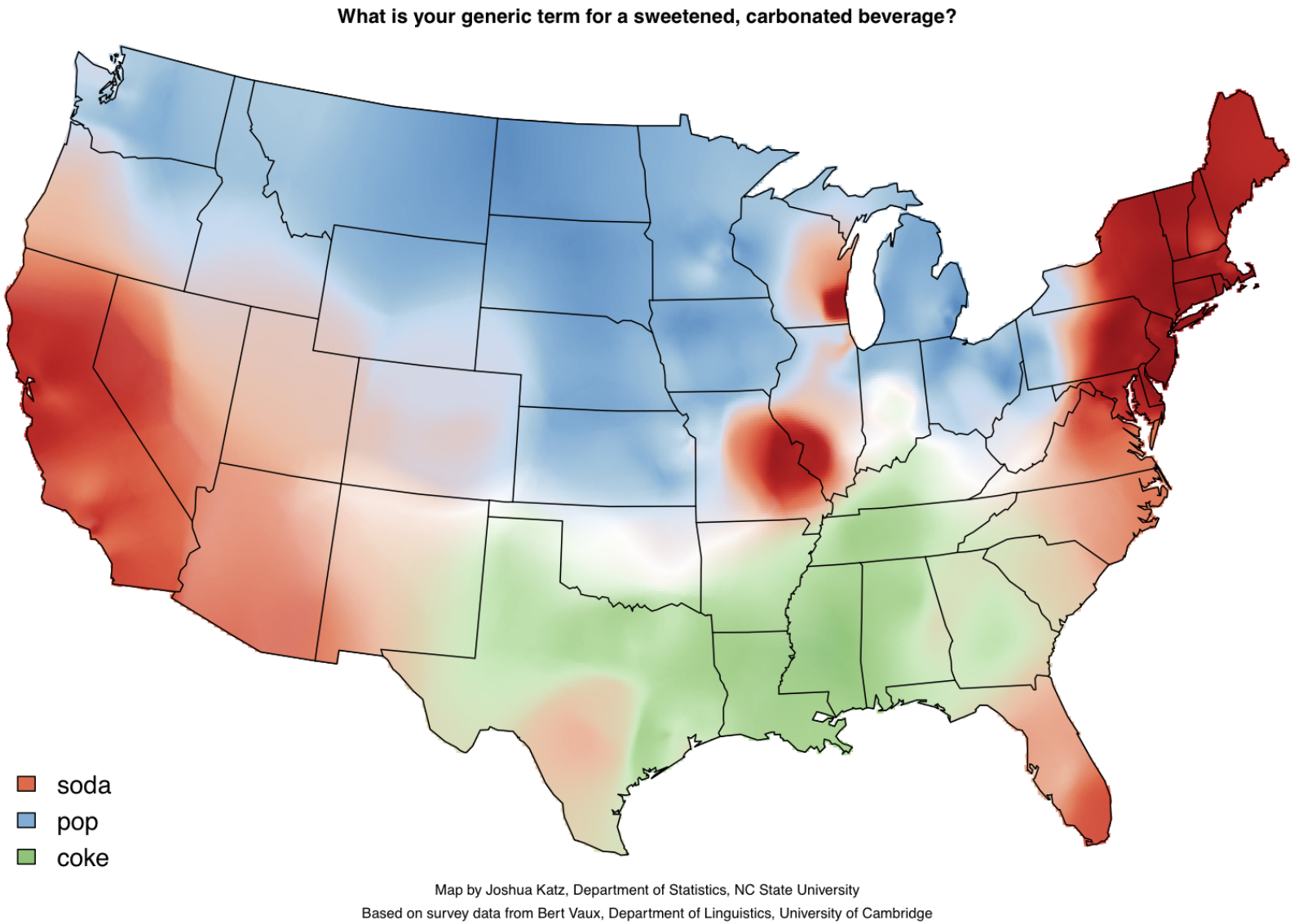


Vernacular example - What did we learn?

- Responses to questions on Vernacular give us some interesting insights into the regions of the U.S.
- We can get a measure of similarity in vernacular using people's responses to the survey
- We can use this to predict where a new person is from! (k-NN)

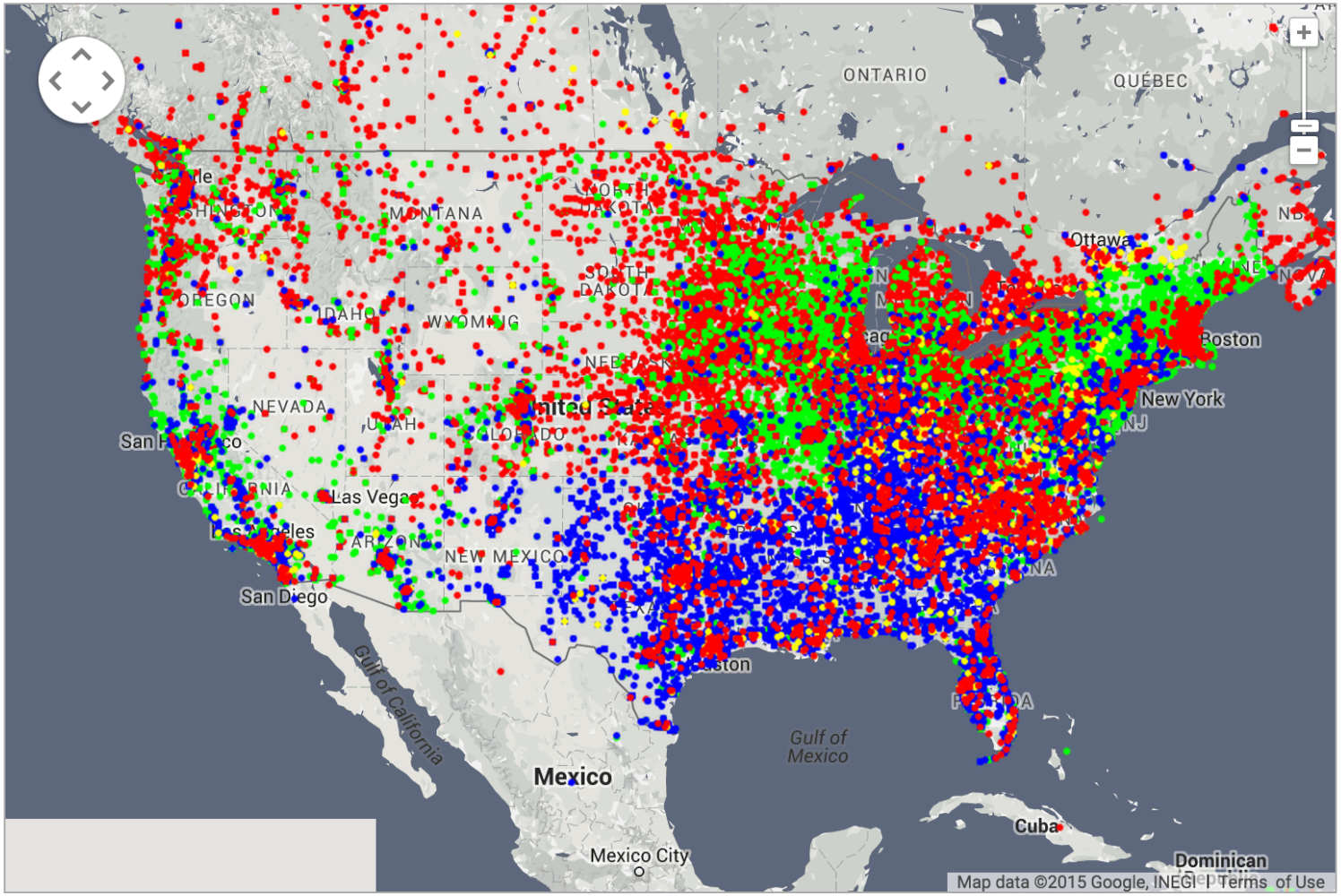
- This is one of my favorite examples because it's:
  1. Fun to learn about this stuff
  2. It shows us improvement from iteration.

Let's go back to the 'soda' example



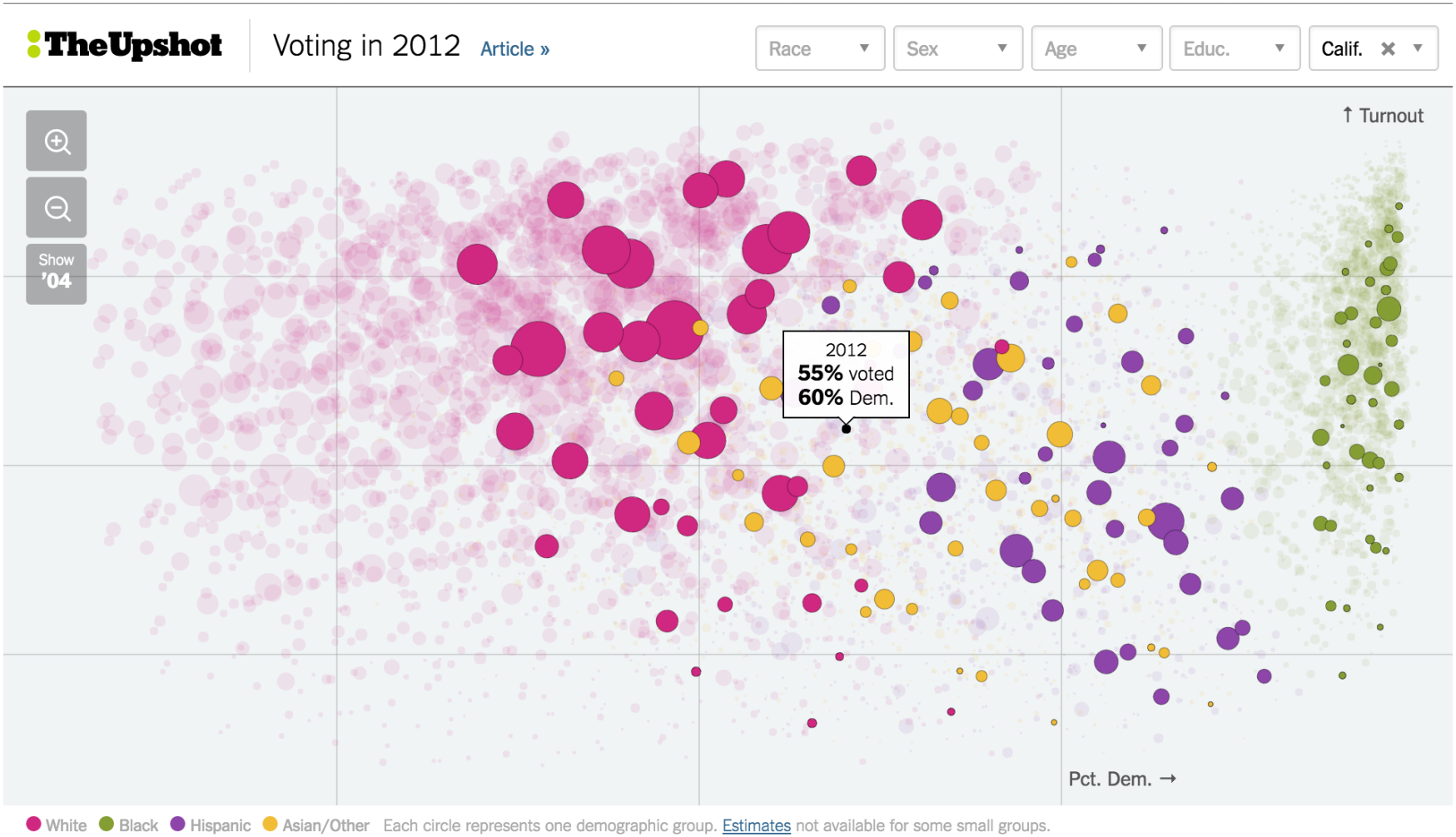
Let's go back to the 'soda' example

Do you say *pop* or *soda*?



pop (39%) soda (37%) coke (18%) soft drink (1%)

Interactive Data Visualizations



▸ The Voting Habits of Americans Like You

Conclusions: Animation and Data Visualization

We've seen a few examples of visualizations and animations that help convey complex ideas

- Facebook connection data contains tons of information
- Some basic reasoning can give us powerful tools (like k-NN)
- Animations and interactive visualizations allow people to explore and learn

Thank you!