Where's best?

A Geospatial Data Digestion Framework Implemented in R with Shiny

Chao Shi

04/25/2017

TL;DR

I built an app

I built a Shiny app, which allows users to choose parameters they care about, then score and rank US counties

https://chaoshi.shinyapps.io/SpatialDataDigester_US_County/

But hopefully at the end of the presentation you would agree that it is a bit more than that

A very brief how-to-play intro

Map Tab

1

Define "Good"

Check Boxes

2

3

4

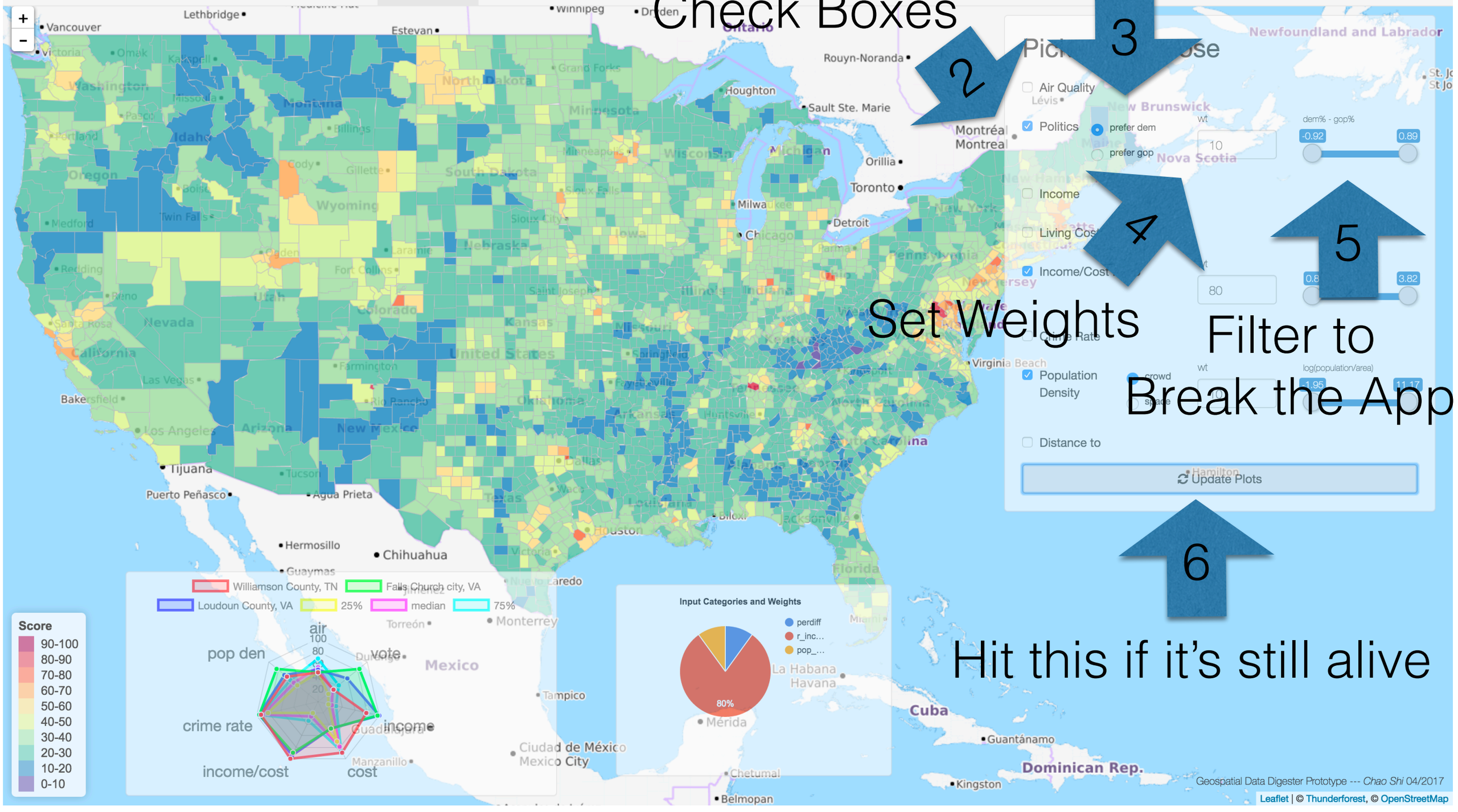
5

6

Set Weights

Filter to Break the App

Hit this if it's still alive



Documentation

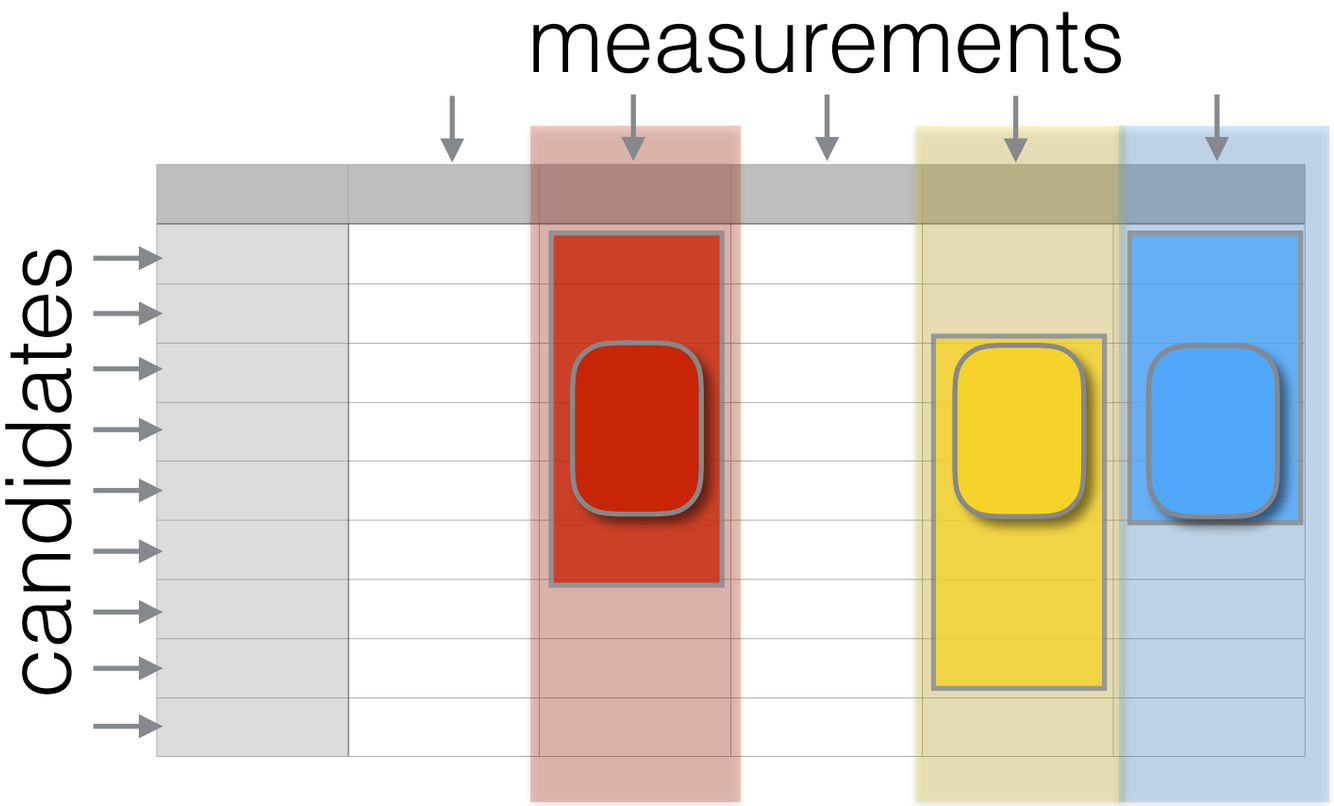
- This talk — high level intro and discussion
- Blog post — expansion of the presentation
- Readme files on Github
- Comments in source code

Motivation

- Strengths of the R + Shiny ecosystem
- Personal appreciation of the power of maps
- Interest in decision making process
- Trying to solve a *type of* problem, record logical thinking, and develop reusable code

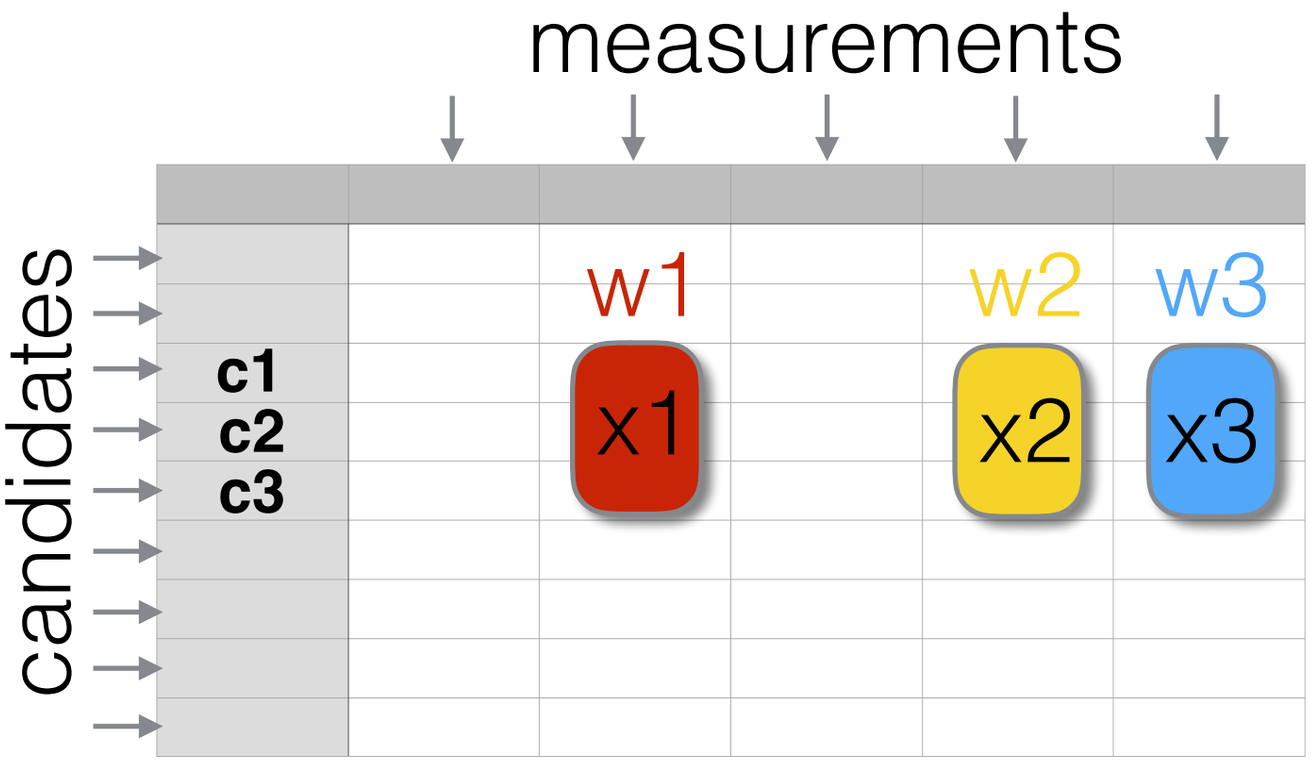
One Decision Making Workflow

- where to eat
- where to buy
- where to live



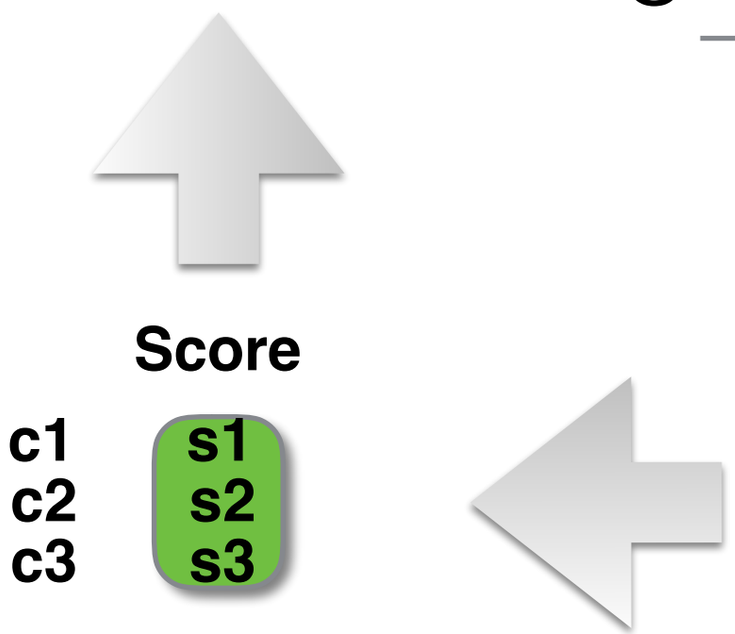
One Decision Making Workflow

- where to eat
- where to buy
- where to live

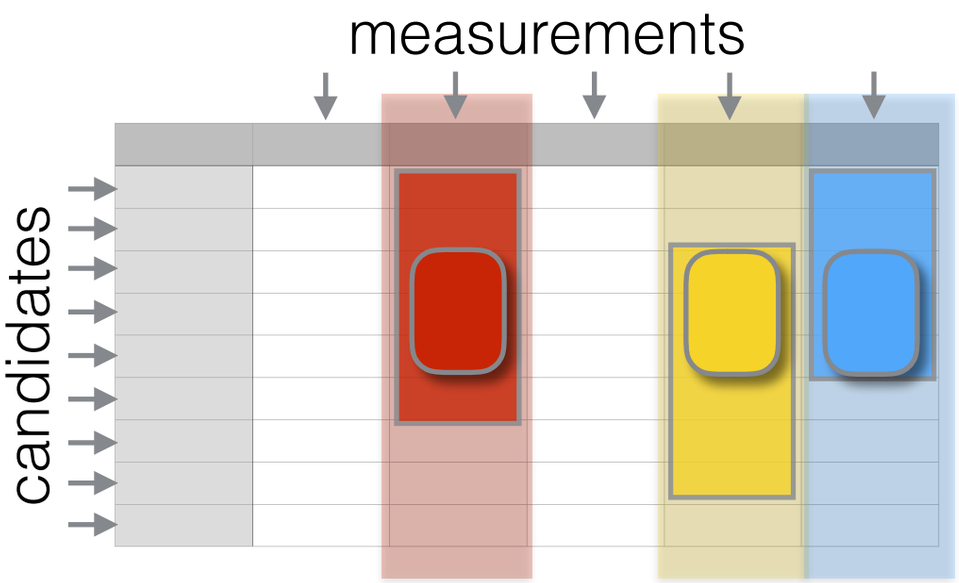


score generator

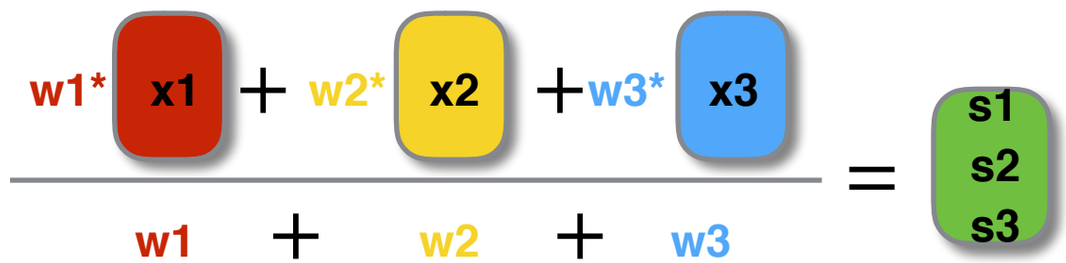
$$\frac{w1^{*}x1 + w2^{*}x2 + w3^{*}x3}{w1 + w2 + w3}$$



One Decision Making Workflow



filter system



score generator

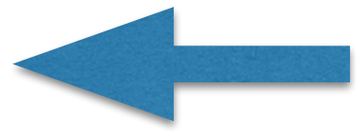
filter system + score generator

data size reduction + dimension reduction

This is probably how the “top 10” lists were made. Can we do better?

Project Design, Scope and Deliverables

- Find a spatial grid complex enough in geometry, familiar enough for the targeted audience, while small enough in variable size for one laptop to handle
- Find 3 data sets (columns) to begin with
- Develop a data filtering and ranking engine in R
- **Minimum Deliverable**: Make proof-of-concept first version



Mid-way Checkpoint

- Add data sets to enrich the complexity of the spatial decision making example
- **Hopeful Deliverable:** Beautify UI with efficiency + backend data analysis
- **Minimum Deliverable**: Document things along the way, make presentation and publish code

Checkpoint — Minimum Functioning Version

Geospatial Data Digester -- US County Ver 0.1

Input Panel

Air Quality — wt: 20 — PM 2.5 µg/m³: 4.44 – 15.96 (4.44 6.8 9.2 11.6 14)

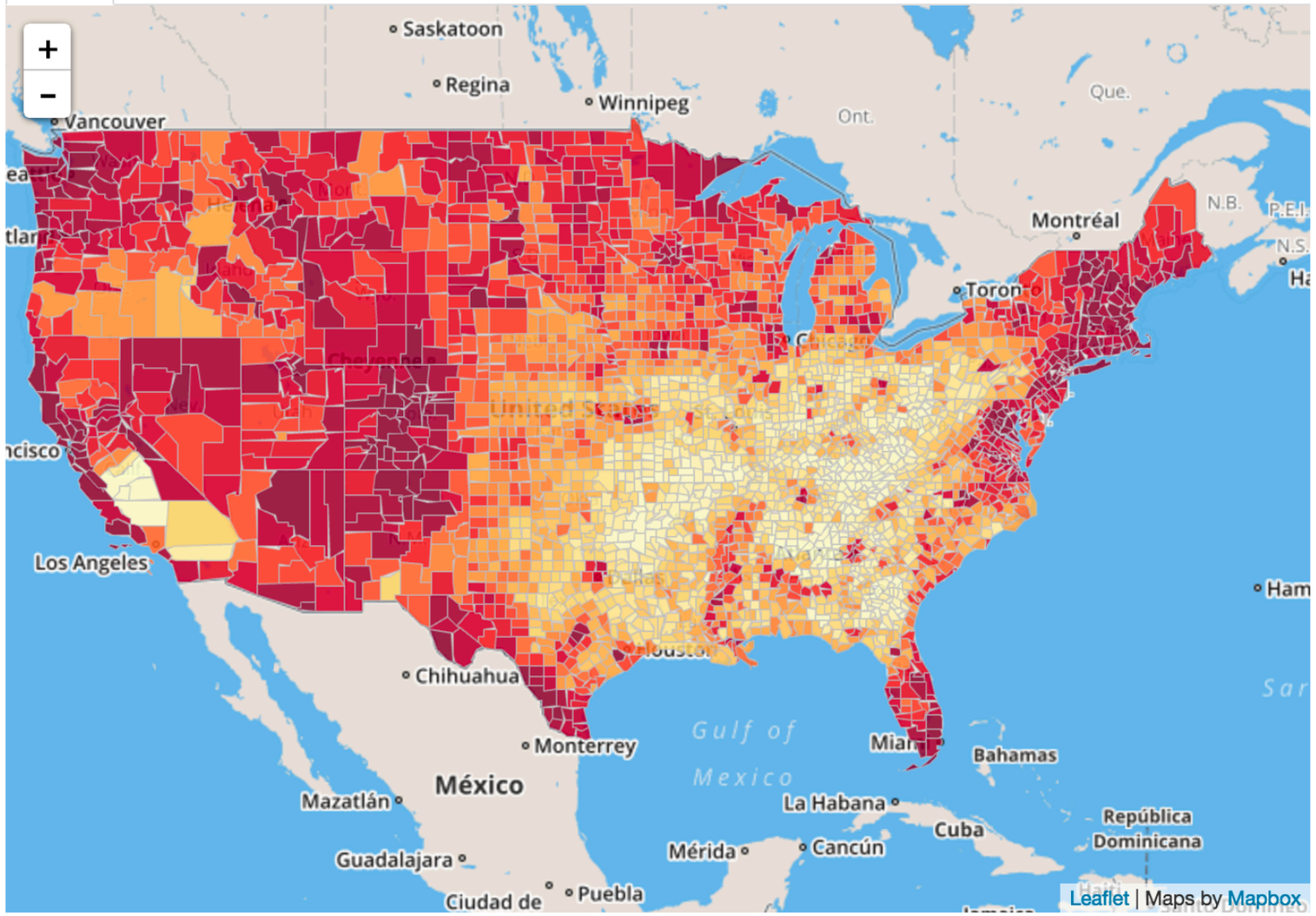
VoteDiff — wt: 20 — dem% - gop%: -0.92 – 0.89 (-0.92 -0.12 0.69)

Income — wt: 20 — $/year: 22,894 – 125,900 (22,894 64,096 105,299)

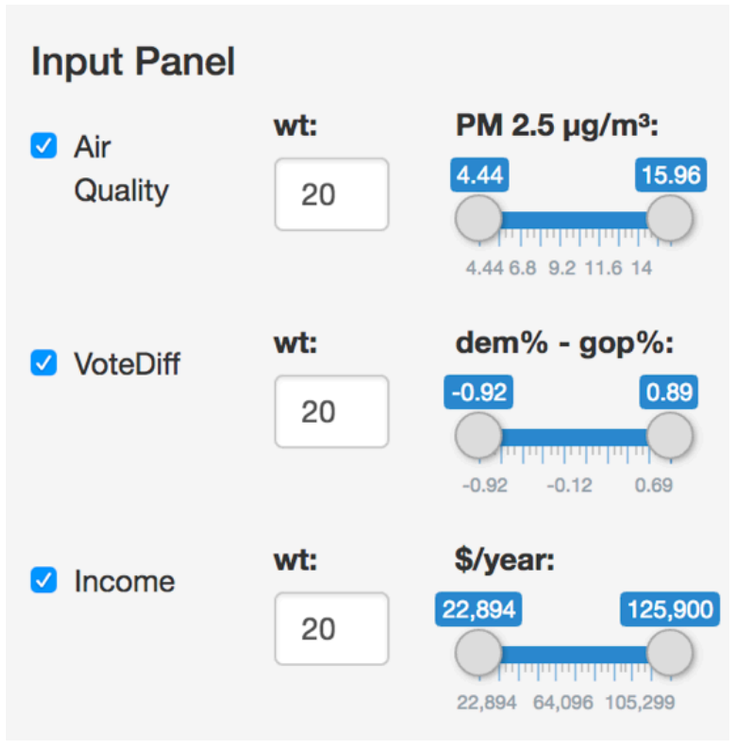
Update Plots

Los Alamos Falls Church Pitkin Marin Fairfax

tab1 tab2



The proof-of-concept version
Why these 3 data sets?



Normalization

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)} * \text{scalar}$$

Positive / "Good" Definition

x'' = scalar - x' **(if needed)**

good

[0,100]

good

[0,100]

[0,100]

what's good?
need user input

[0,100]

good

Data Source

Air Quality	https://data.cdc.gov/api/views/cjae-szjv/rows.csv?accessType=DOWNLOAD
Politics	https://github.com/tonmcg/County_Level_Election_Results_12-16
Income	https://www.census.gov/
Living Cost	https://github.com/mikeasilva/living-wage scraped from MIT living wage study
Crime Rate	https://www.kaggle.com/mikejohnsonjr/united-states-crime-rates-by-county
Population Density	https://www.census.gov/
Distance to	calculated from (lat,lng) data for each county, contained in shape files https://www.census.gov/geo/maps-data/data/cbf/cbf_counties.html
Real Estate	https://www.zillow.com/research/ not used on maps because of missing data

css style from ‘superzip’ example and lots of learning from postings by Joe Cheng

1

Geospatial Data Digester -- US County

Interactive map | Map filtered data explorer | Correlation analysis | About

Leaflet

2

‘radarchart’ package

3

shinyCorrplot by saurfang (tab 3)

5

googleVis

4

6

the brain

Score
90-100
80-90
70-80
60-70
50-60
40-50
30-40
20-30
10-20
0-10

Williamson County, TN | Falls Church city, VA | Loudoun County, VA | 25% | median | 75%

air | vote | income | cost | income/cost | crime rate | pop den

Input Categories and Weights

perdiff | r_inc… | pop_…

80%

measurements

candidates

Income/Cost Ratio

wt

80

0.8

3.82

0.89

11.17

w1* x1 + w2* x2 + w3* x3 = s1 s2 s3

w1 + w2 + w3

Distance to

Update Plots

Geospatial Data Digester Prototype --- *Chao Shi* 04/2017

Leaflet | © Thunderforest, © OpenStreetMap

Time?

```
if time_left_in_minute() <=3:
        go to correlation tab
else:
        dance
```

“Wait, I thought you were supposed to show some insights.”

Well, the way you play with the app is forming questions,
and every map you produce (hopefully) gives you some insights.

Correlation Matrix

Let’s have a second look at the input data offering

Air Quality

Politics

Income

Living Cost

Income/Cost

Crime Rate

Population Density

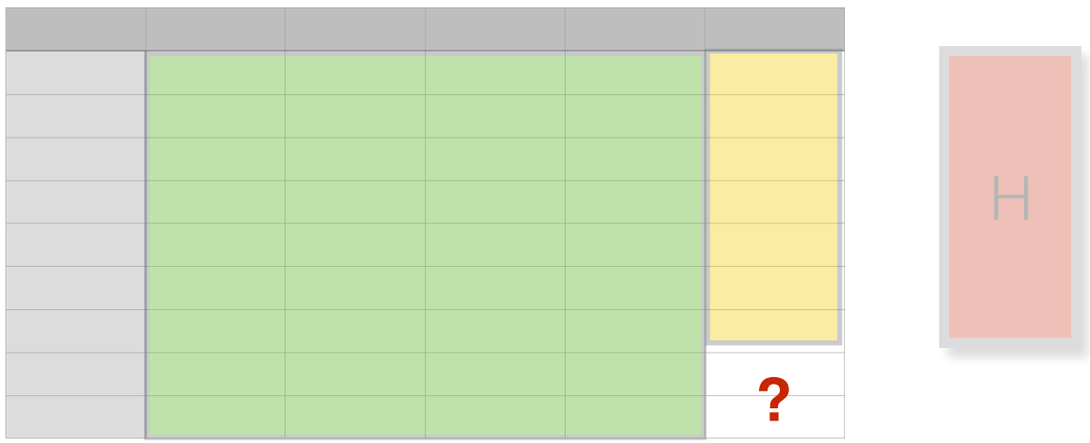
Distance to

3 offerings vs dof 2
why including inc/cost?
what does it mean when user selects all 3?

Highly correlated with several things.
What are the consequences
when filtering with this variable?

What does this ranking app do

Is this a typical ‘Kaggle’-like problem?



No

Users try to predict something outside the box, and highly subjective.

Happiness = a1*x1 + a2*x2 + a3*x3 + …

Users bring the definition of this function
Yet I pre-defined that it has to be a linear system

Hence the title…

Where’s best?

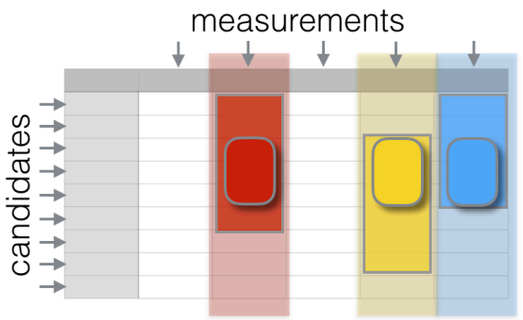
A Geospatial Data Digestion Framework
Implemented in R with Shiny

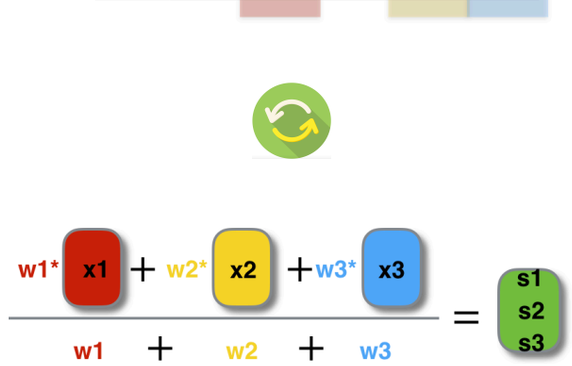
What do we have so far?

Problem Definition

Data collection
cleaning
normalization

Digestion Engine





Decision

Platform Independent

\+

Visualization on a ~~boat~~ map + Corr analysis

Where R 'Shines'

Near Future Steps

- A few features to develop before making a NYC where-to-live app
    - categorical data
    - distance calculation -> travel time calculation

The Future

- Imagine this as a service, how to make profit?
    - What kind of established companies might be interested?
    - This as a web service? What's the break even case?
- Scale up consideration
- Data with gaps — interpolation is rarely just a math problem
- Decision making based on data streams — real time digestion before data disappears, self-updating system…

Acknowledgement

- NYC Data Science: Shu Yan, Zeyu Zhang
- Inspiration from the ‘SuperZip’ example by Joe Cheng
- Leaflet mapping examples on datascienceriot.com
- Correlation Matrix app ‘shinyCorrplot’ by saurfang
- Developers of all other packages I used for this project

A few thoughts

- Too much data to handle — when facing a fast growing ocean of data, user-friendly digestion solutions will be in high demand, from corporations to consumers.

- When a majority of service providers have superb reviews, consumers may turn to other more objective measurement to help making decisions