

# 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/](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

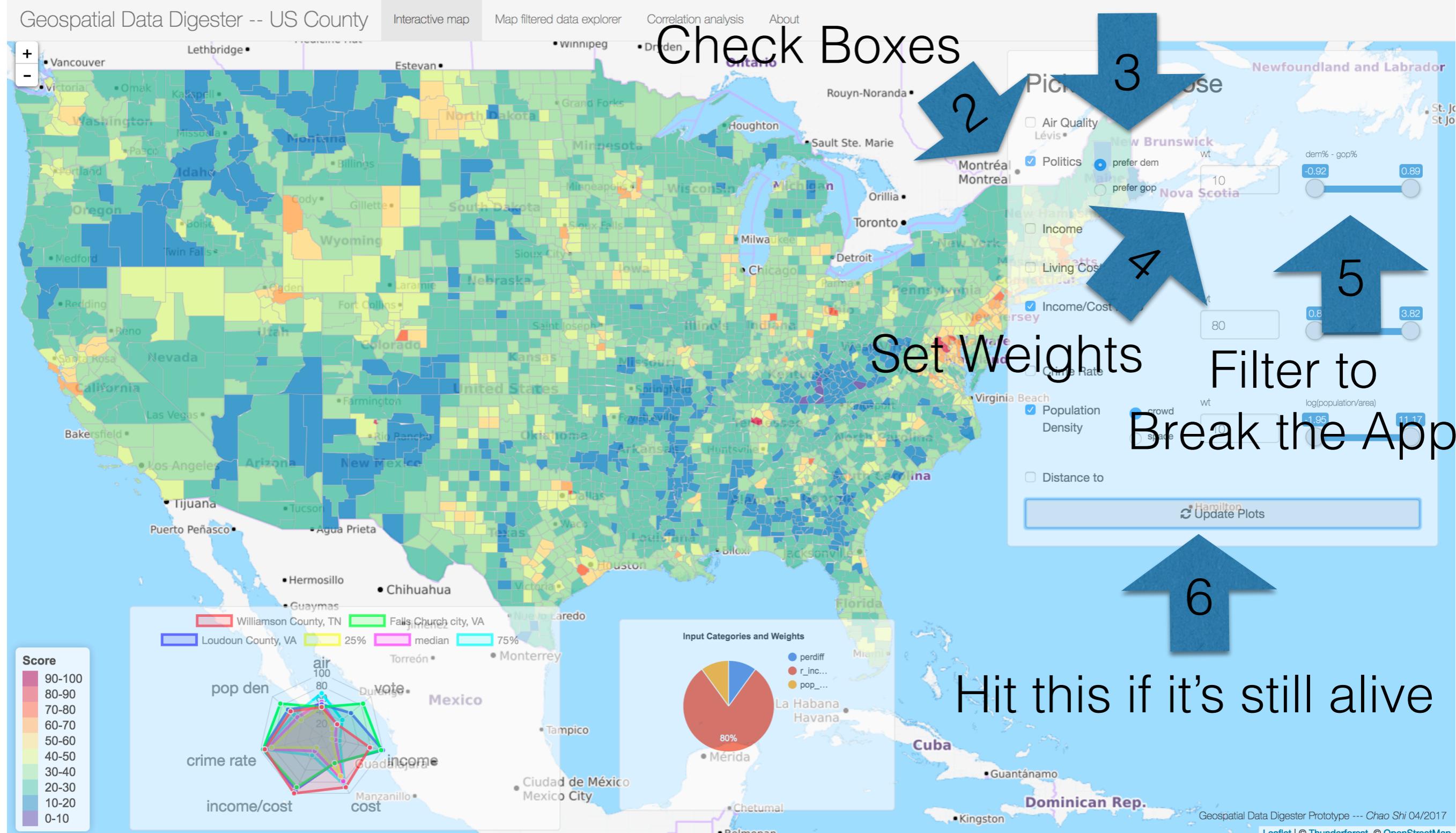
1

# Set Weights

Filter to  
wt log(population/area)  
1.95  
peak the A

6

Hit this if it's still alive



# Documentation

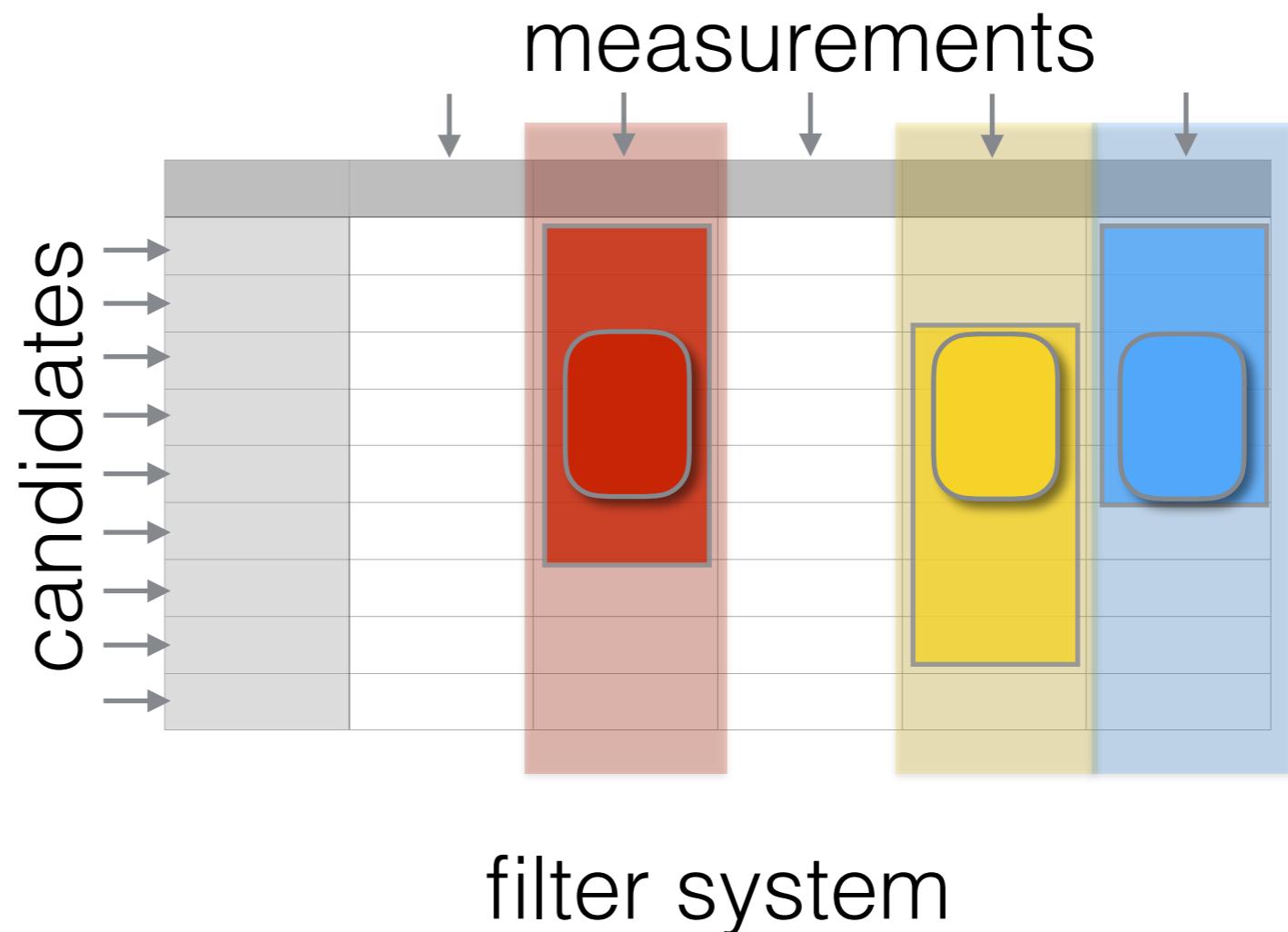
- This talk — high level intro and discussion
- Blog post — expansion of the presentation
- Readme files on Github
- Comments in source code

# 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

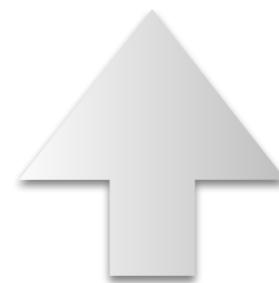
# One Decision Making Workflow

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

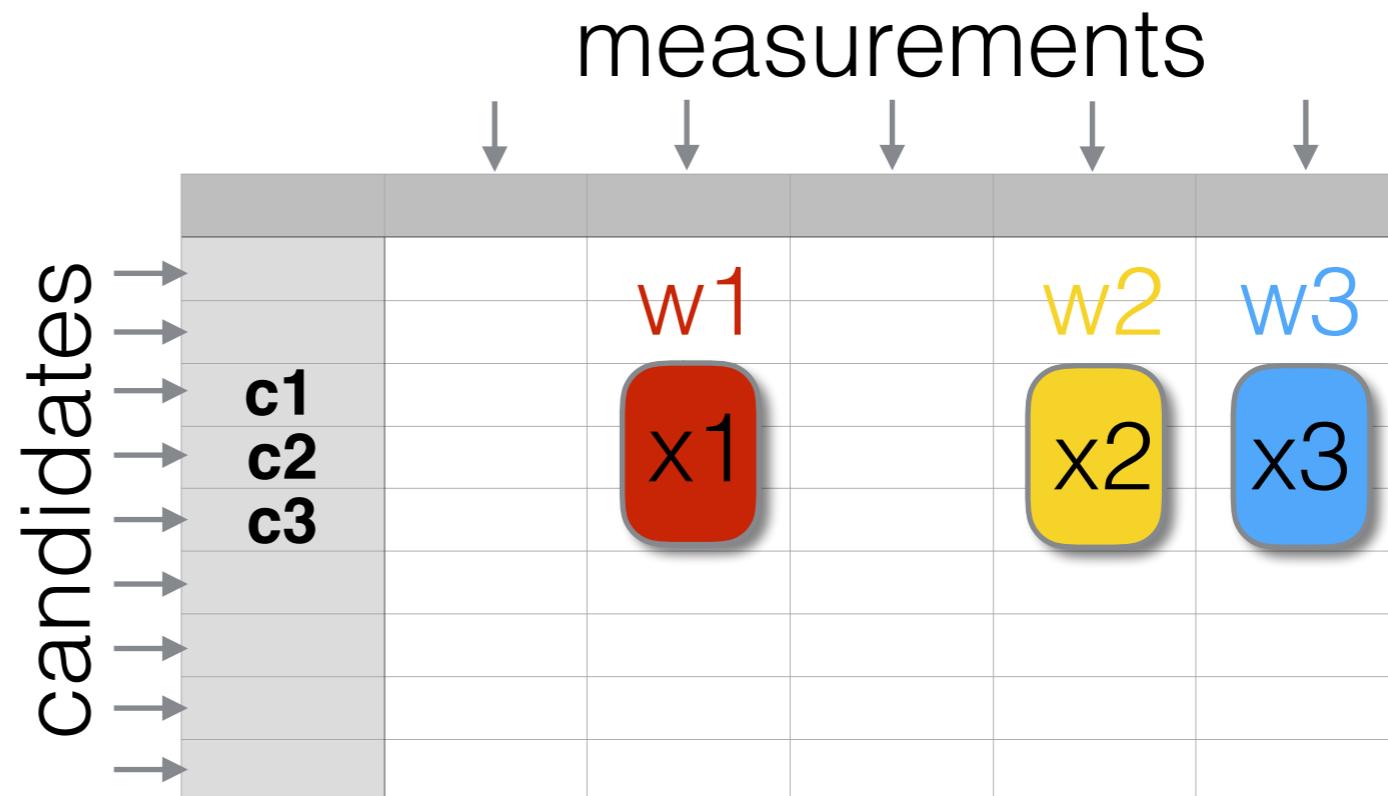
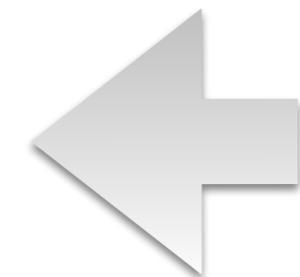


# One Decision Making Workflow

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



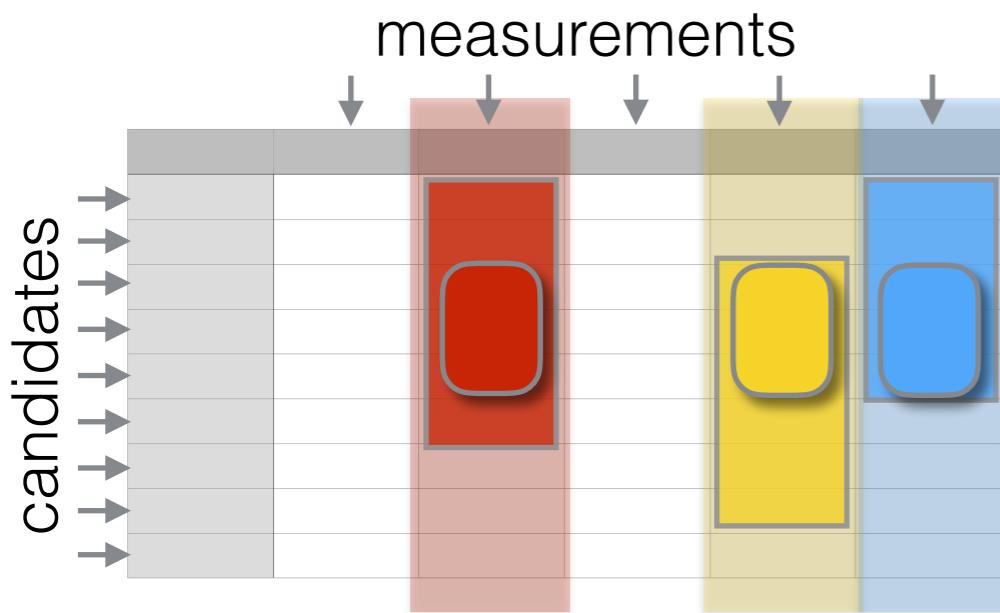
	Score
c1	s1
c2	s2
c3	s3



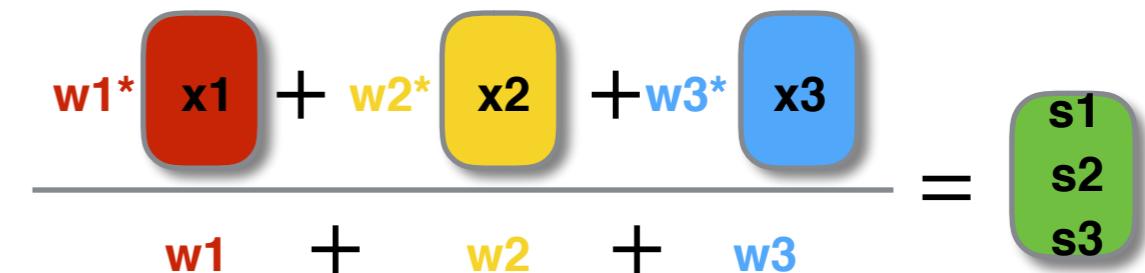
score generator

$$\frac{w_1^* x_1 + w_2^* x_2 + w_3^* x_3}{w_1 + w_2 + w_3}$$

# One Decision Making Workflow



filter system



score generator

filter system + score generator

data size reduction + dimension reduction

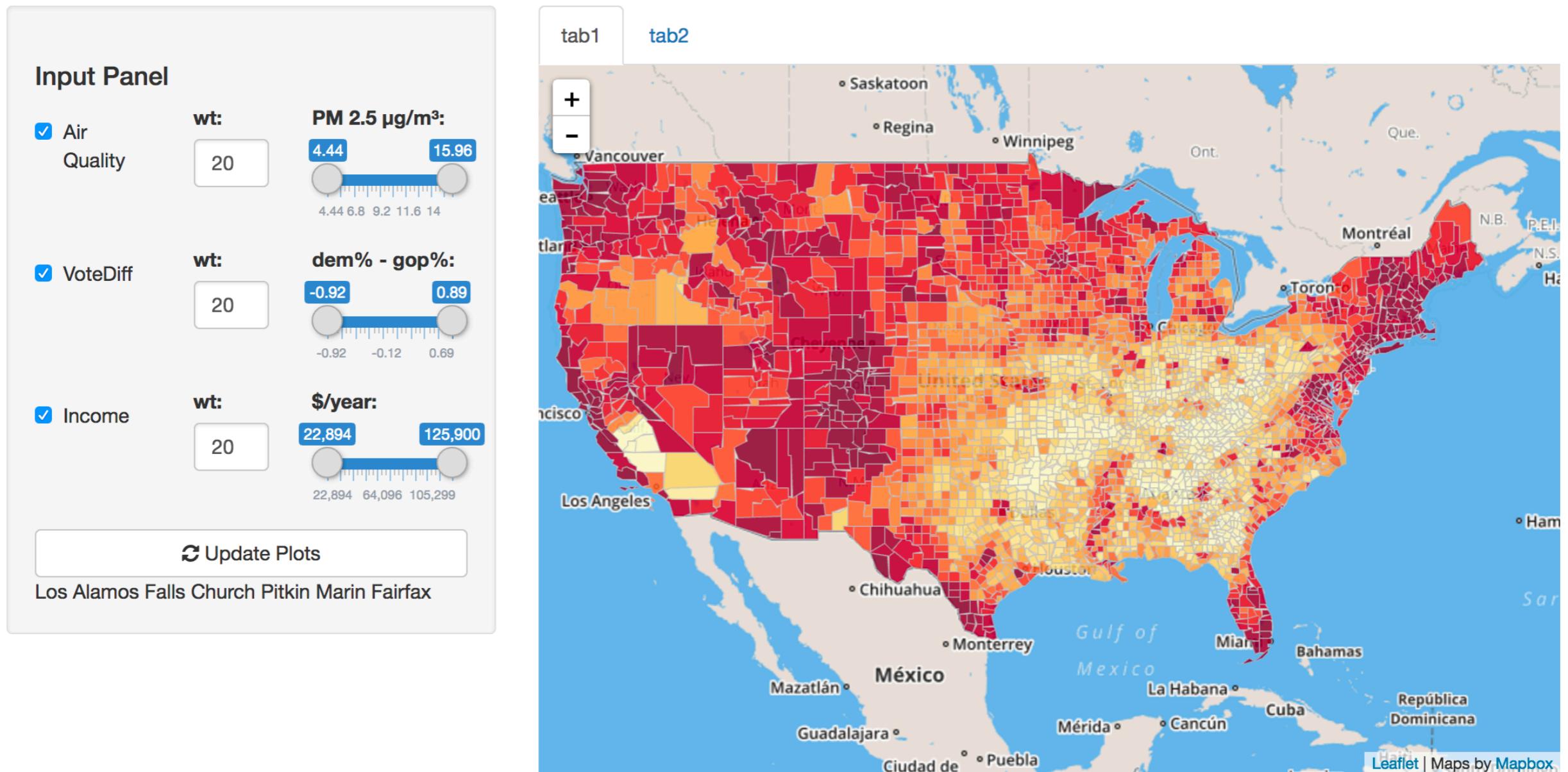
This is probably how the “top 10” lists were made. Can we do better?

# 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

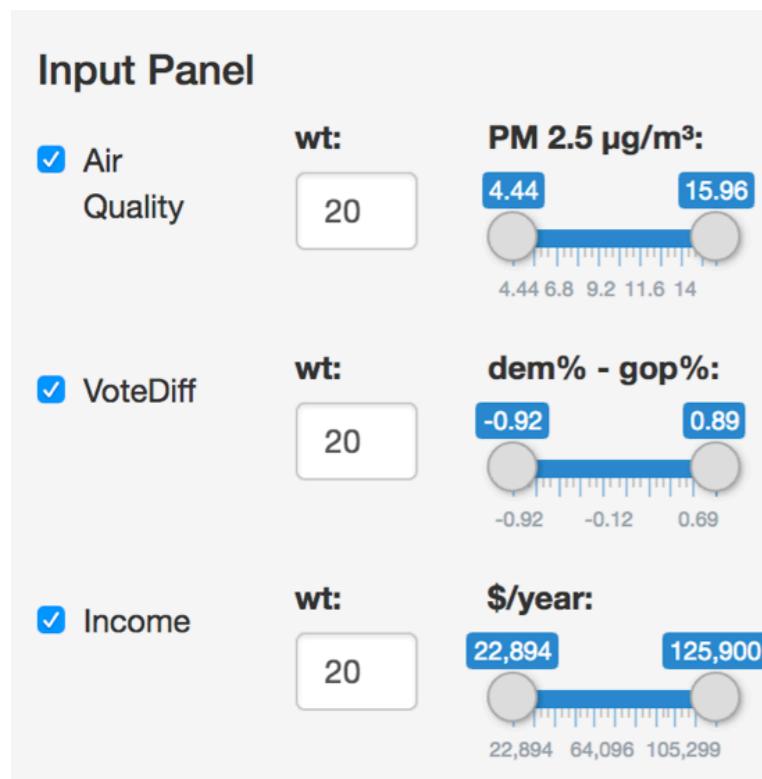
# Checkpoint — Minimum Functioning Version

## Geospatial Data Digester -- US County Ver 0.1



# 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**

$$x'' = \text{scalar} - x' \quad (\text{if needed})$$

good

[0, 100]

good

[0, 100]



[0, 100]

[0, 100]

good

**what's good?**  
need user input

# 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](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](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

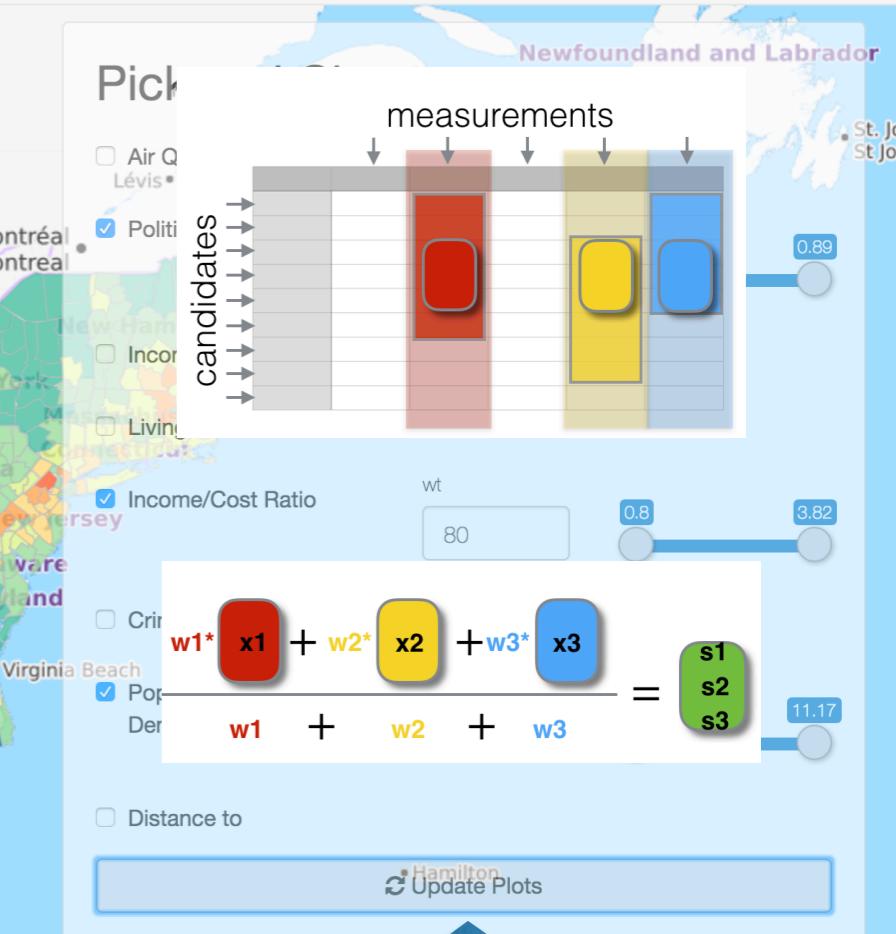
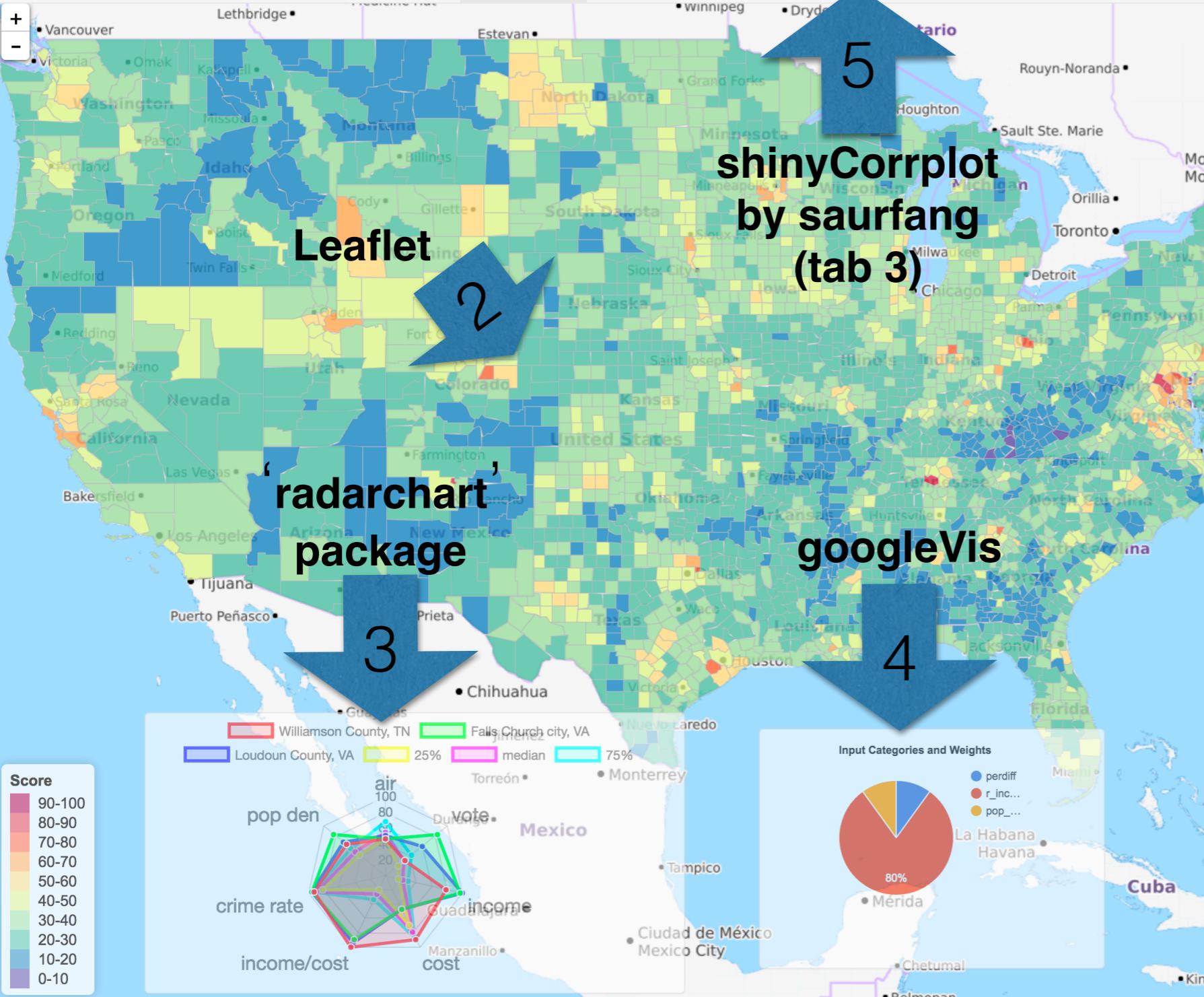
Geospatial Data Digester -- US County

Interactive map

Map filtered data explorer

Correlation analysis

About



6  
the brain

# 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

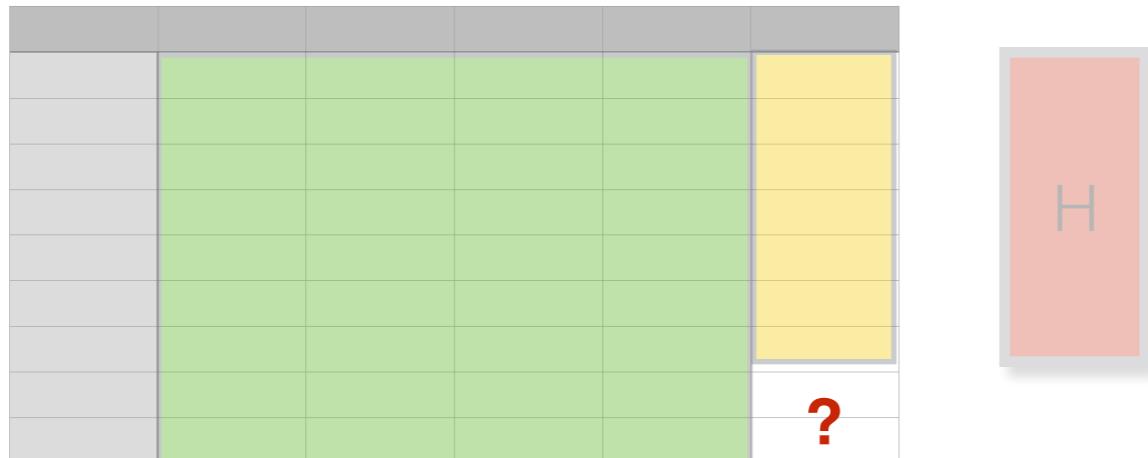
**3 offerings vs dof 2  
why including inc/cost?  
what does it mean when user selects all 3?**

**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?

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



$$\text{Happiness} = a_1 \cdot x_1 + a_2 \cdot x_2 + a_3 \cdot x_3 + \dots$$

No

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?

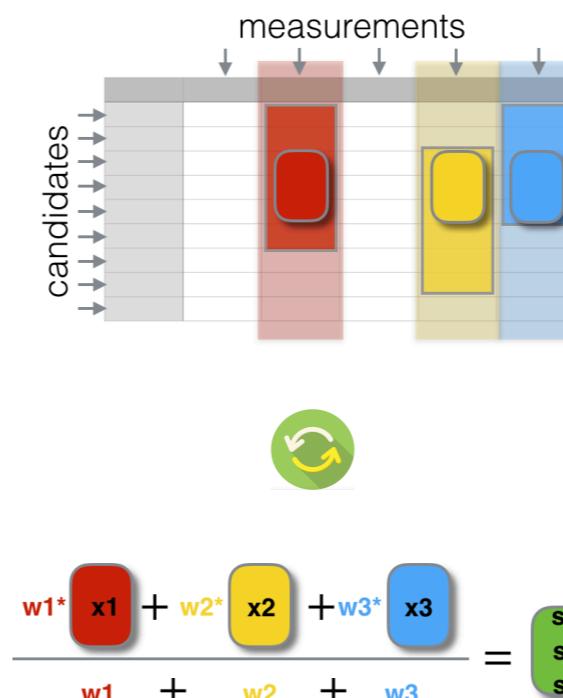
A Geospatial Data Digestion Framework  
Implemented in R with Shiny

# 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](http://datascienceriot.com)
- Correlation Matrix app ‘shinyCorrplot’ by saurfang
- Developers of all other packages I used for this project

# A few thoughts on general business opportunities with the kind of thinking demonstrated in this project

- 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 measurements to help making decisions