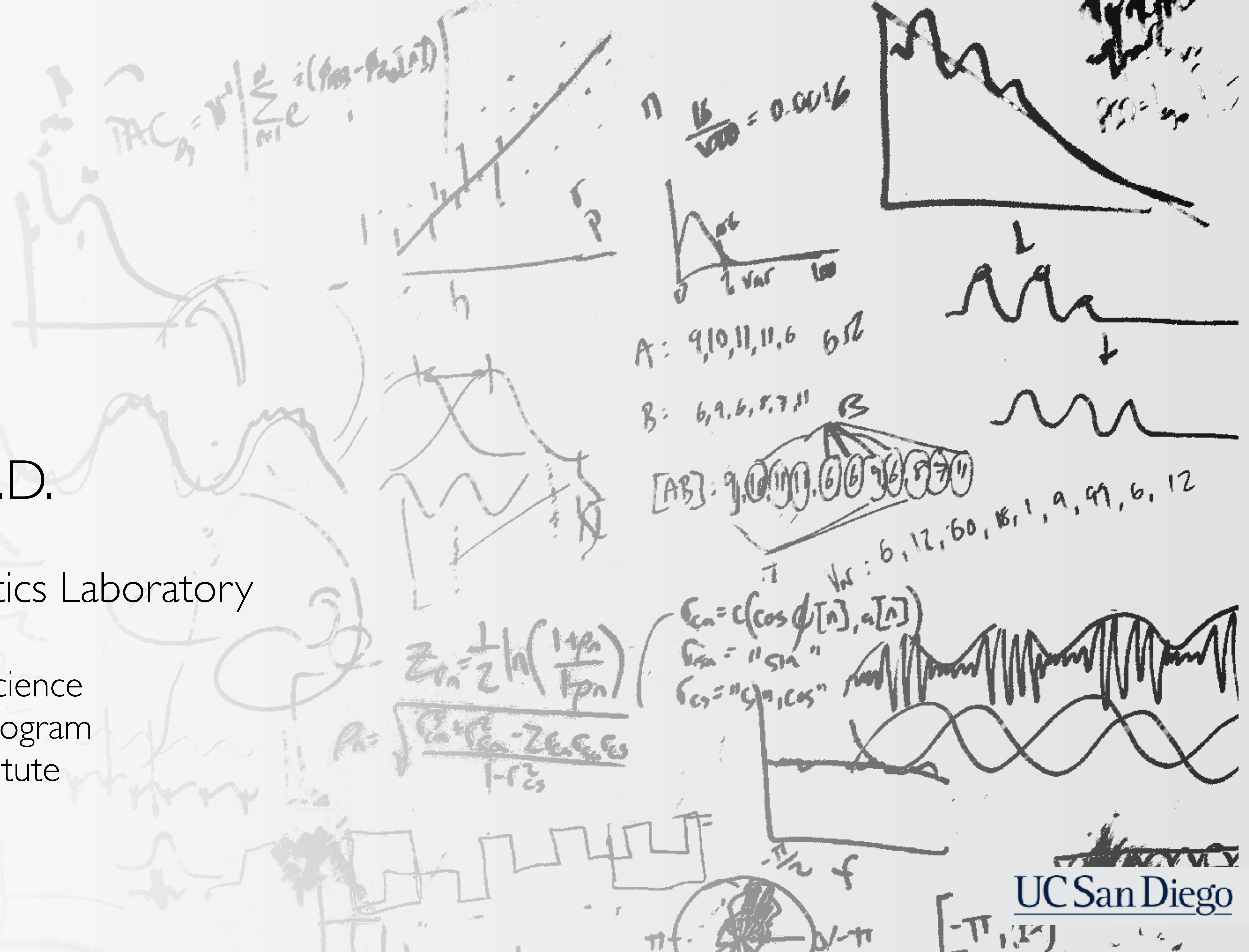


Bradley Voytek, Ph.D.

UC San Diego
Neural and Data Analytics Laboratory

Department of Cognitive Science
Neurosciences Graduate Program
Halıcıoğlu Data Science Institute

bvoytek@ucsd.edu
@bradleyvoytek

COGS 108
Data Science in Practice

More Geospatial: Practice and Stats

UC San Diego

Geocoding

Jupyter demo!

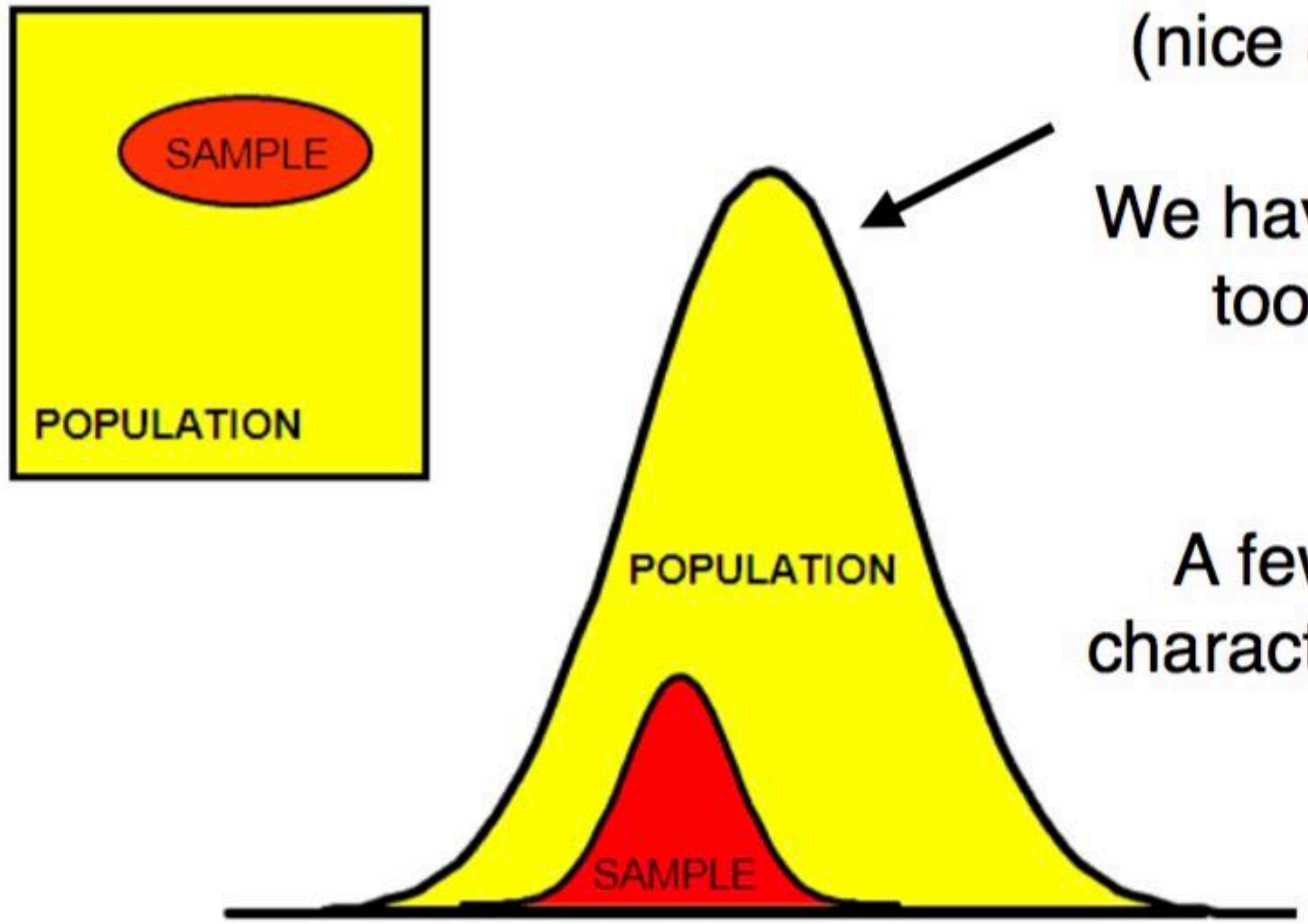
Geocoding statistics

Geo-resampling demo!

Non-parametric statistics

Central Limit reminder demo!

Non-parametric statistics - why?



Normal distribution
(nice and friendly)

We have good math tools for this.

A few parameters **fully** characterize the distribution.

Source: Richard Gao

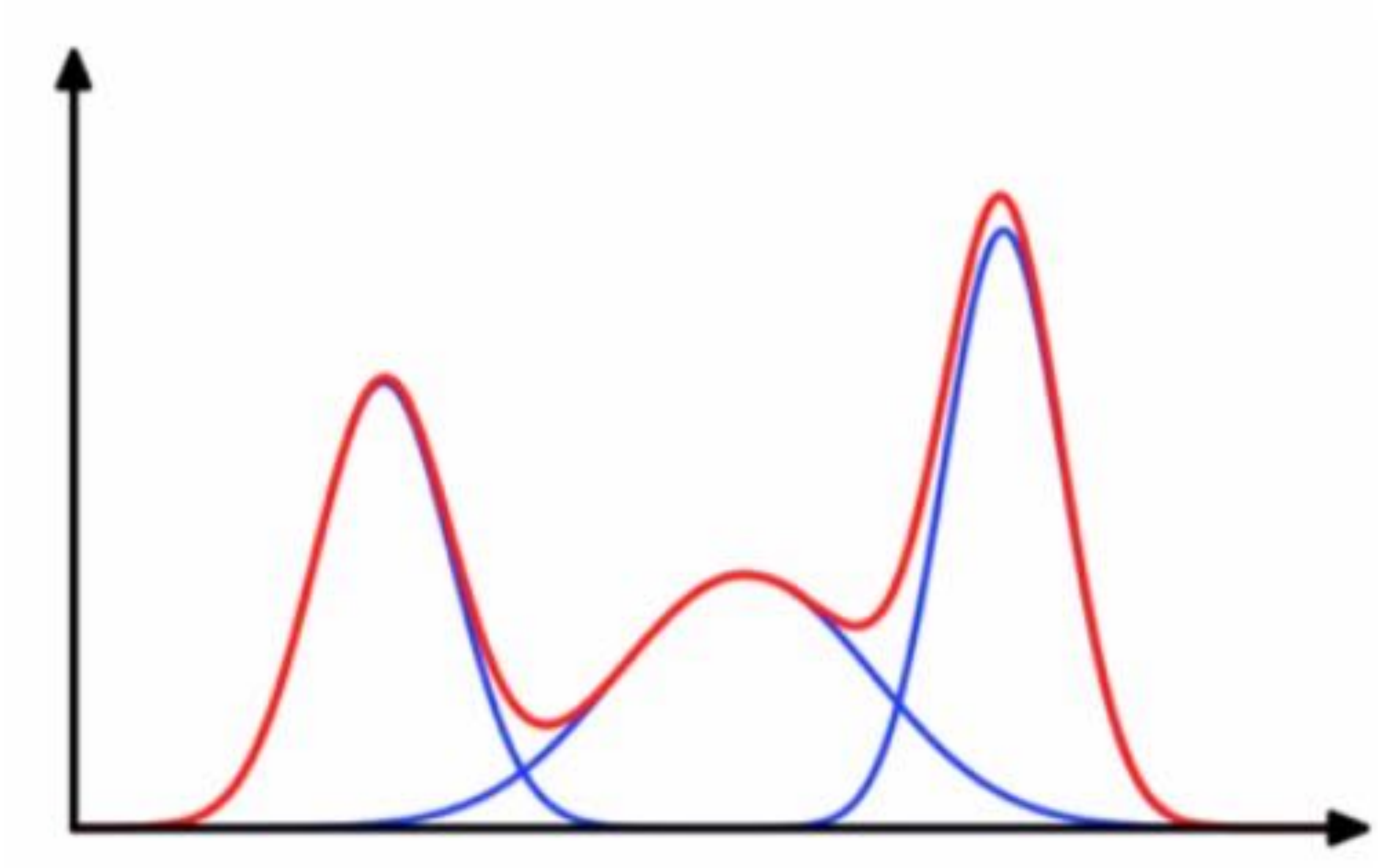
Good news and bad news

Bad news: Many of the standard techniques and methods documented in standard statistics textbooks have significant problems when we try to apply them to the analysis of the spatial distributions.

Good news: Geospatial referencing provides us with a number of new ways of looking at data and the relations among them. (e.g. distance, adjacency, interaction, and neighbor)

Non-parametric statistics - why?

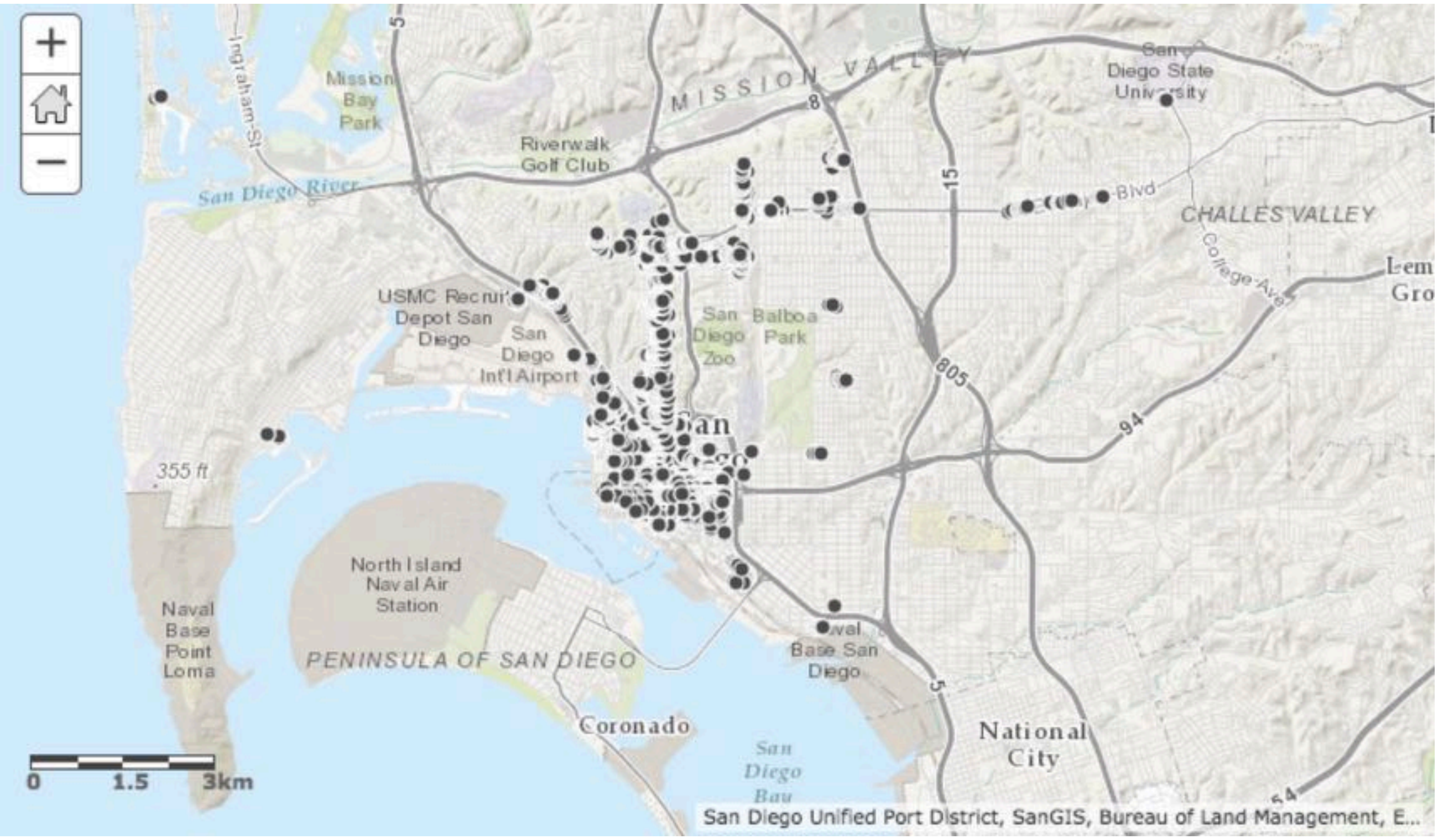
What if your population is distributed like this?



Source: Richard Gao

Non-parametric statistics - why?

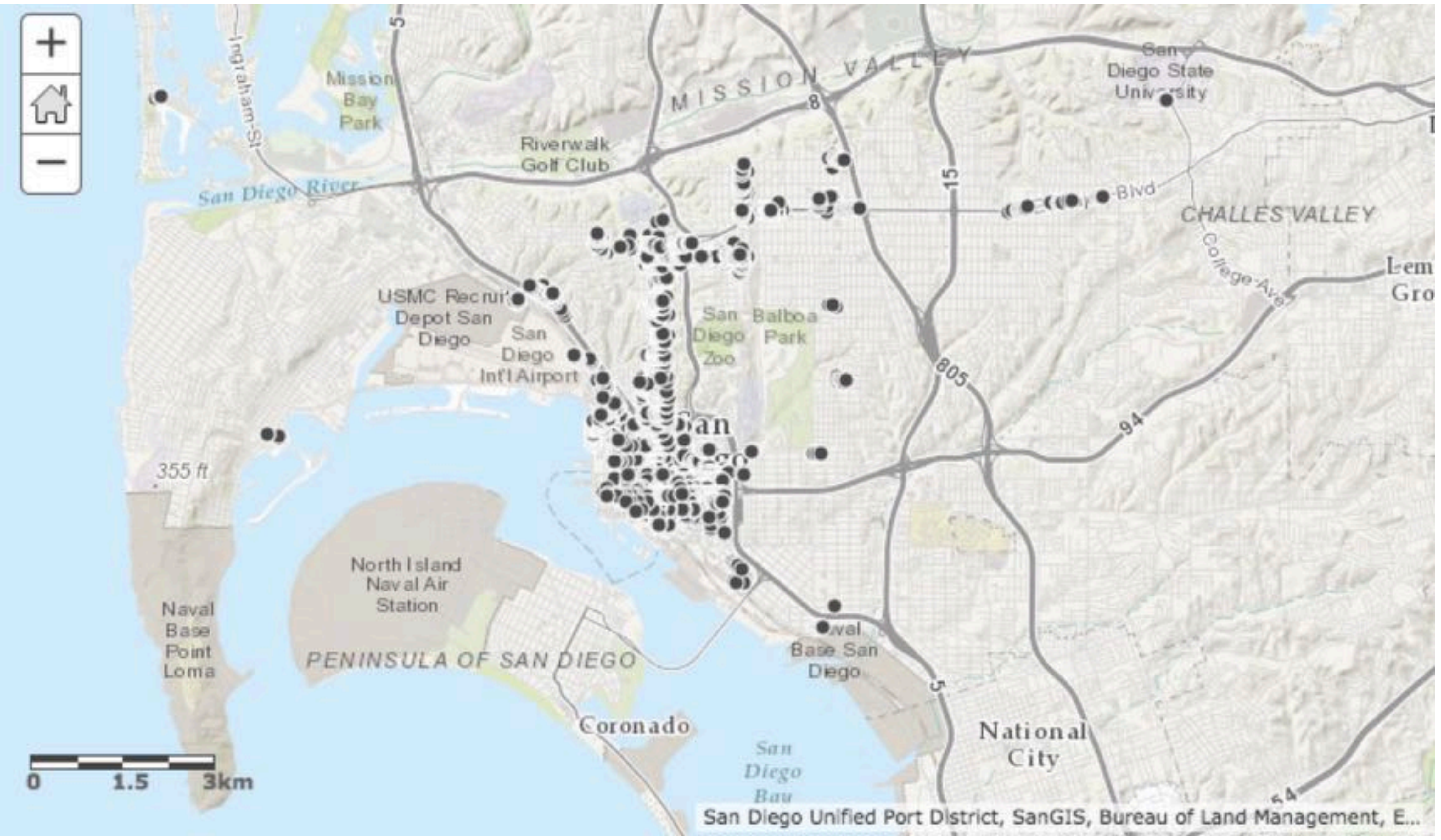
Or like this?



Source: Richard Gao

Non-parametric statistics - why?

Or like this?



Parameters (like mean and variance) cannot fully and accurately capture this distribution!

Hence, we require **non-parametric statistics**.

Source: Richard Gao

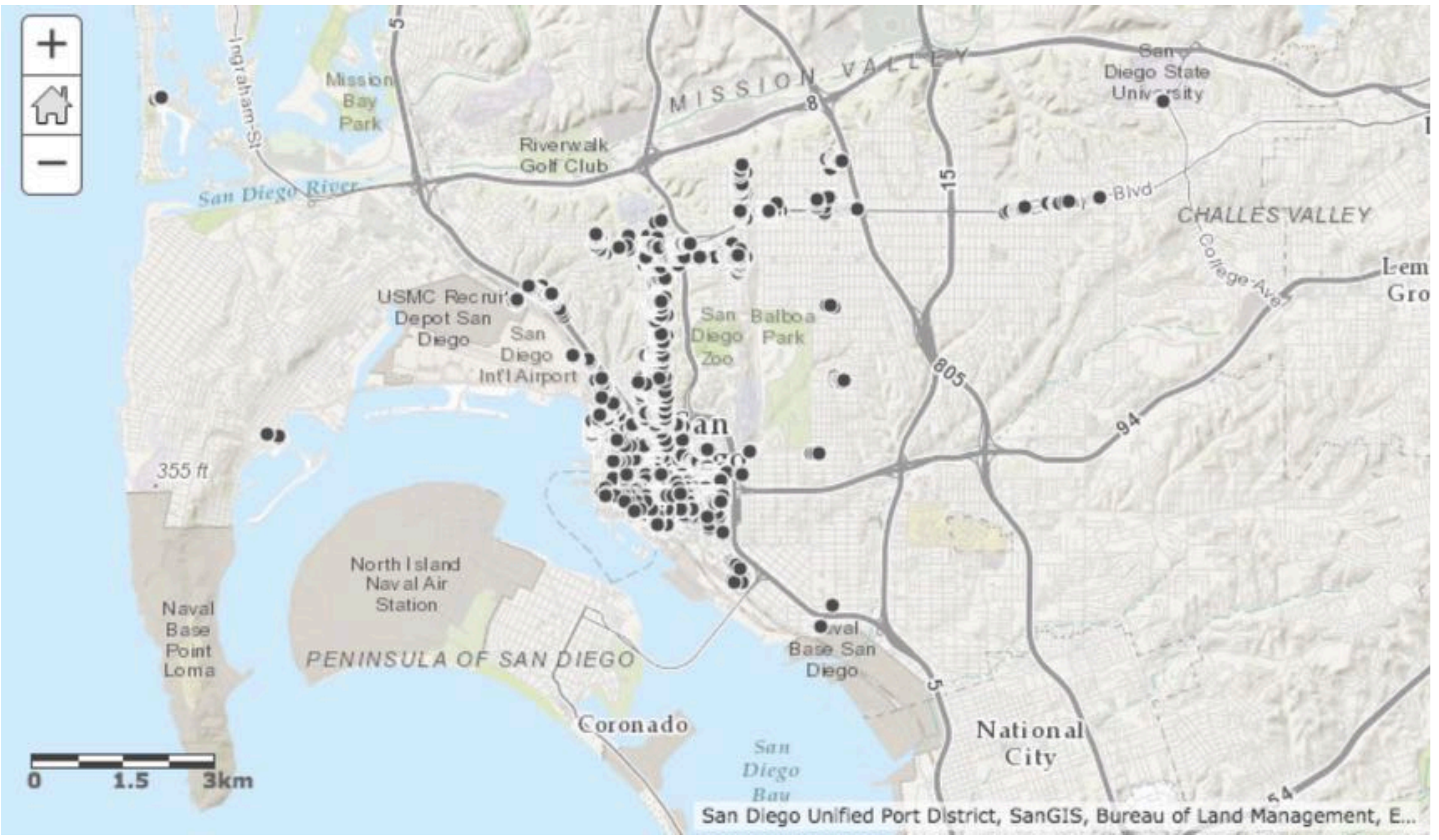
Resampling statistics

- **What & When?**
- **Kolmogorov-Smirnoff Test**
- **Rank Statistics**
- **Jackknife & Bootstrap**
- **Non-parametric prediction models**

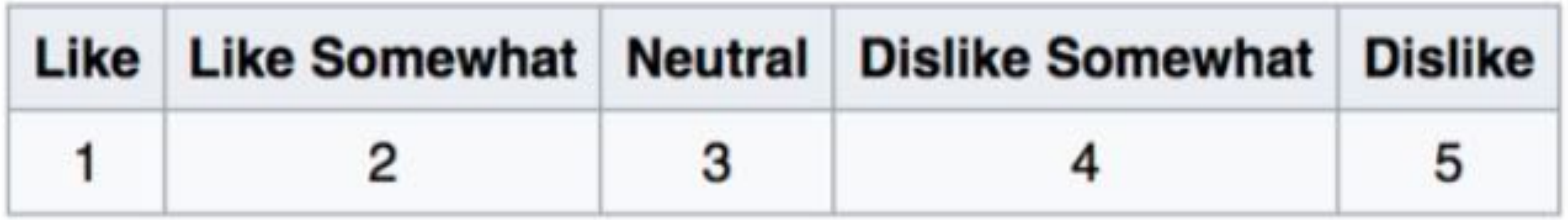
Source: Richard Gao

Non-parametric statistics - when to use?

- When underlying distributions are non-normal, skewed, or cannot be parametrized simply.



- When you have ranked (ordinal) data, e.g., preferences.



Like	Like Somewhat	Neutral	Dislike Somewhat	Dislike
1	2	3	4	5

- When you need to build an empirical "null" distribution.

Source: Richard Gao

Non-parametric statistics

- **Myth:** Non-parametric statistics does not use parameters.

- **Fact:** Non-parametric statistics does not make *assumptions about* / parametrize the underlying distribution generating the data.

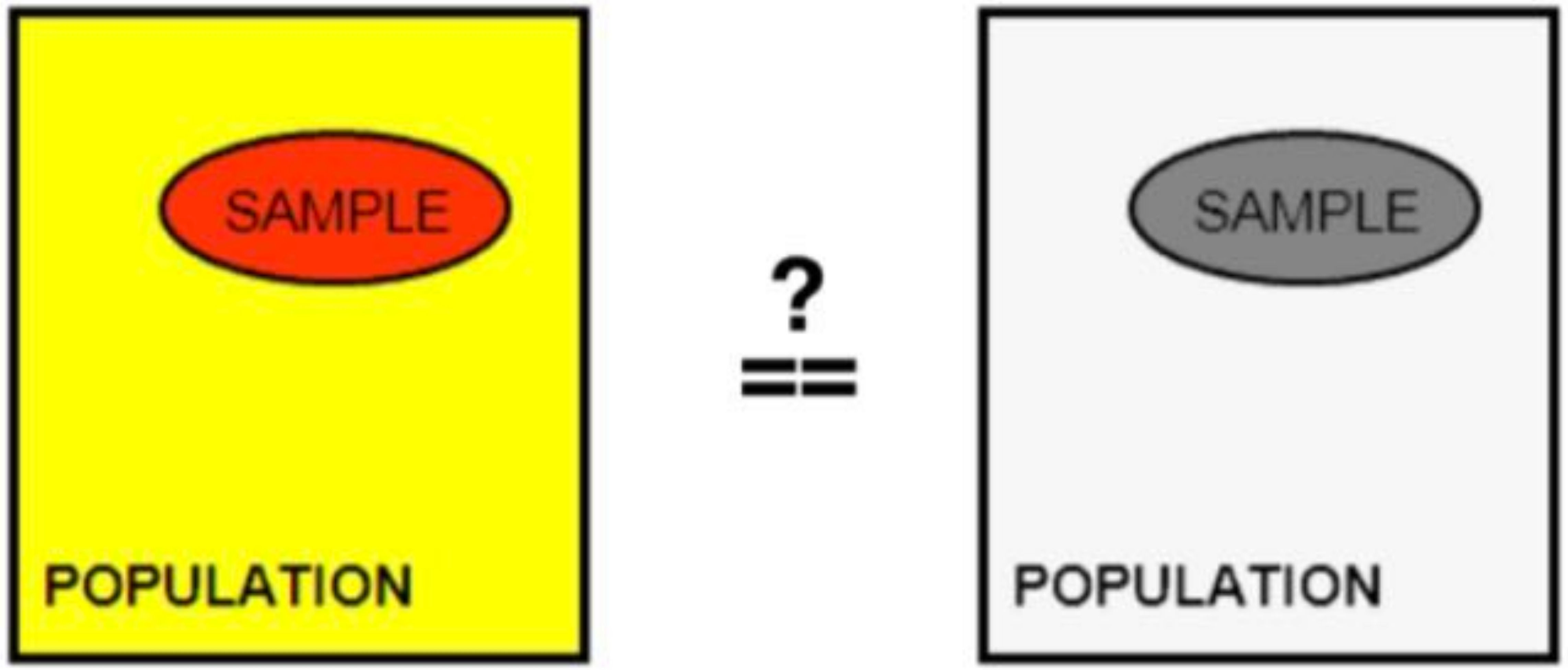
- **“Distribution-Free” statistics**
    - Meaning, it does no assume data-generating process (like heights) result in, e.g., normally-distributed data

Source: Richard Gao

Kolmogorov-Smirnov test

Question:

- Given (limited) samples from two populations, how do we quantify whether they come from the same distribution?

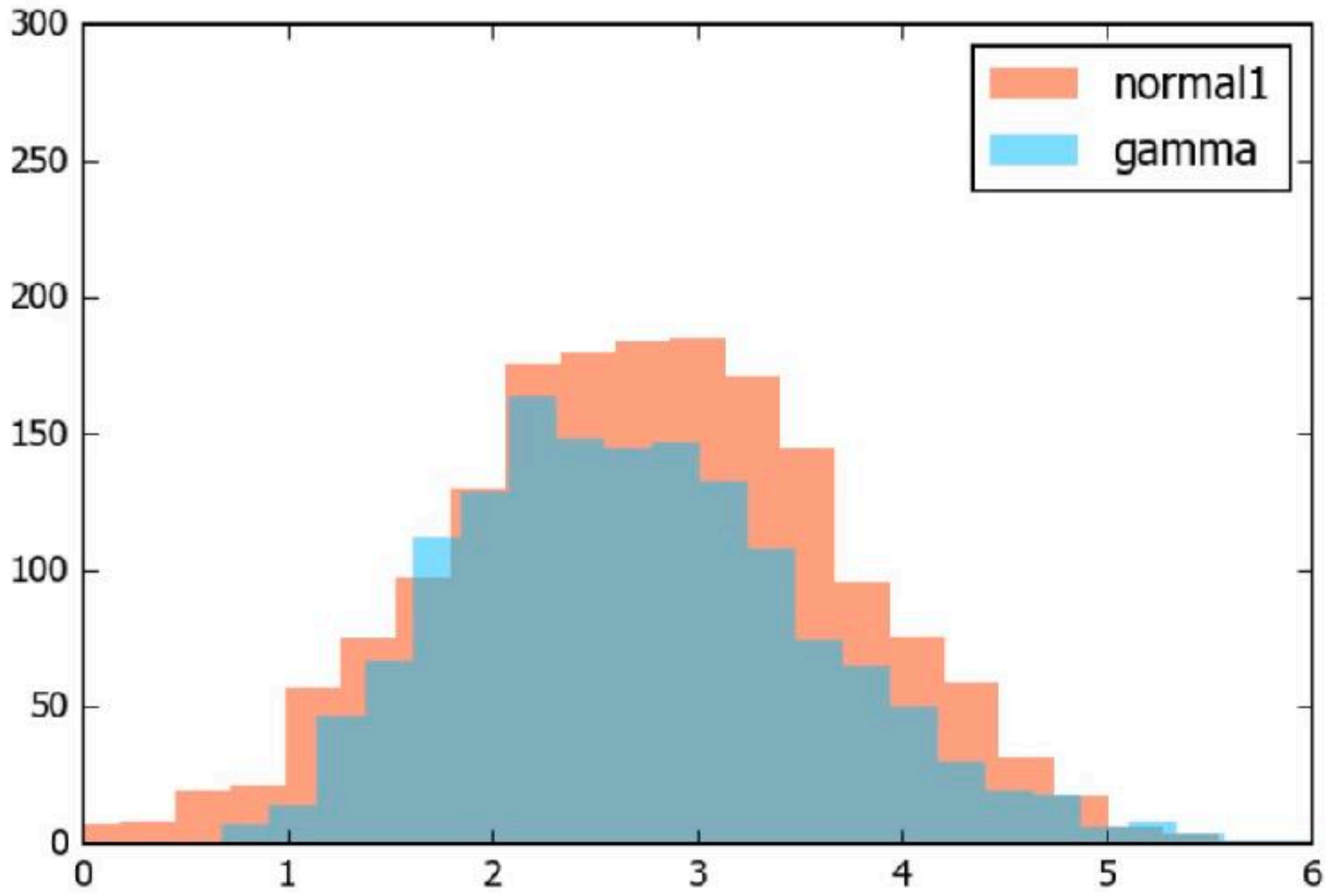


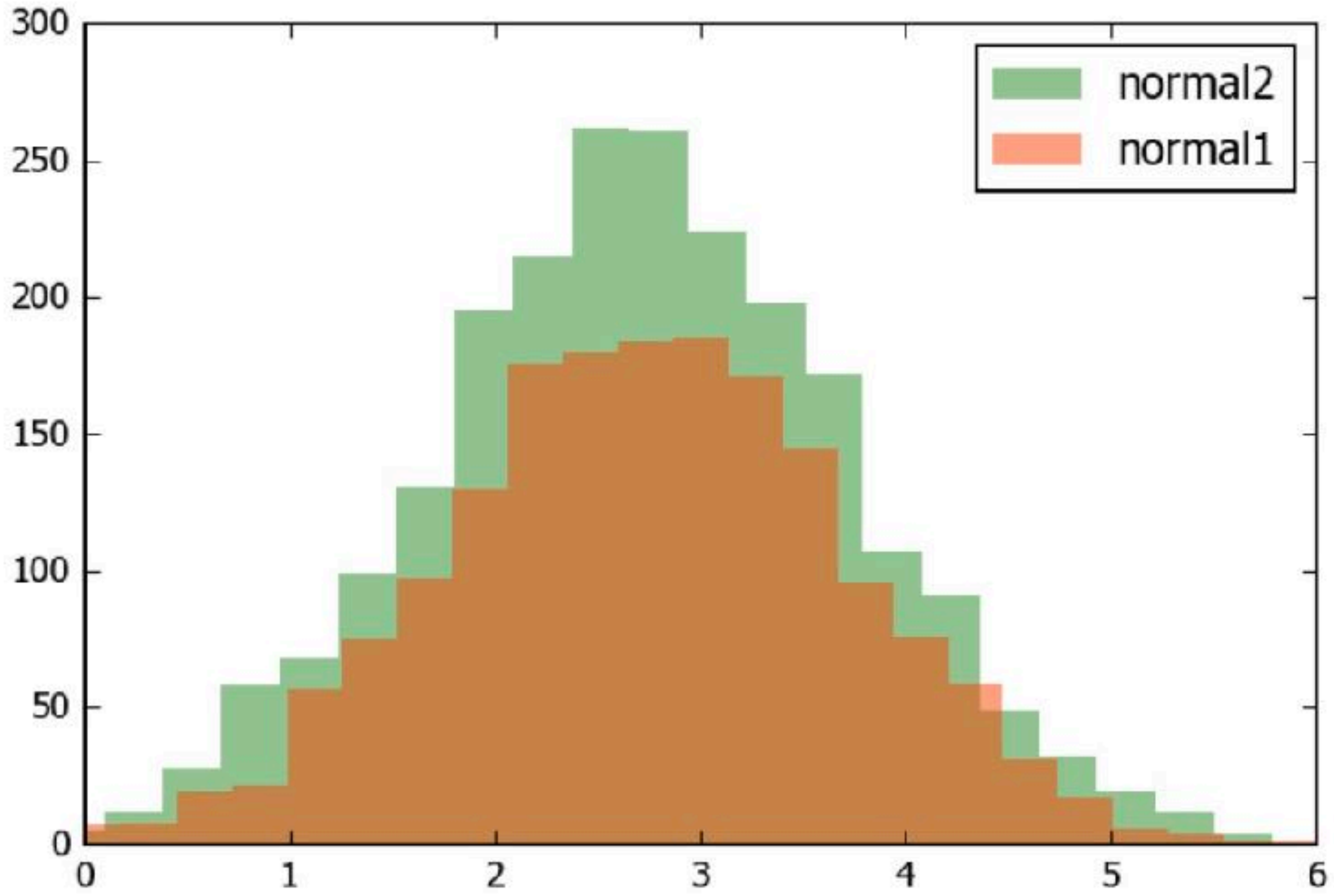
Source: Richard Gao

Kolmogorov-Smirnov test

Question:

- Given (limited) samples from two populations, how do we quantify whether they come from the same distribution?

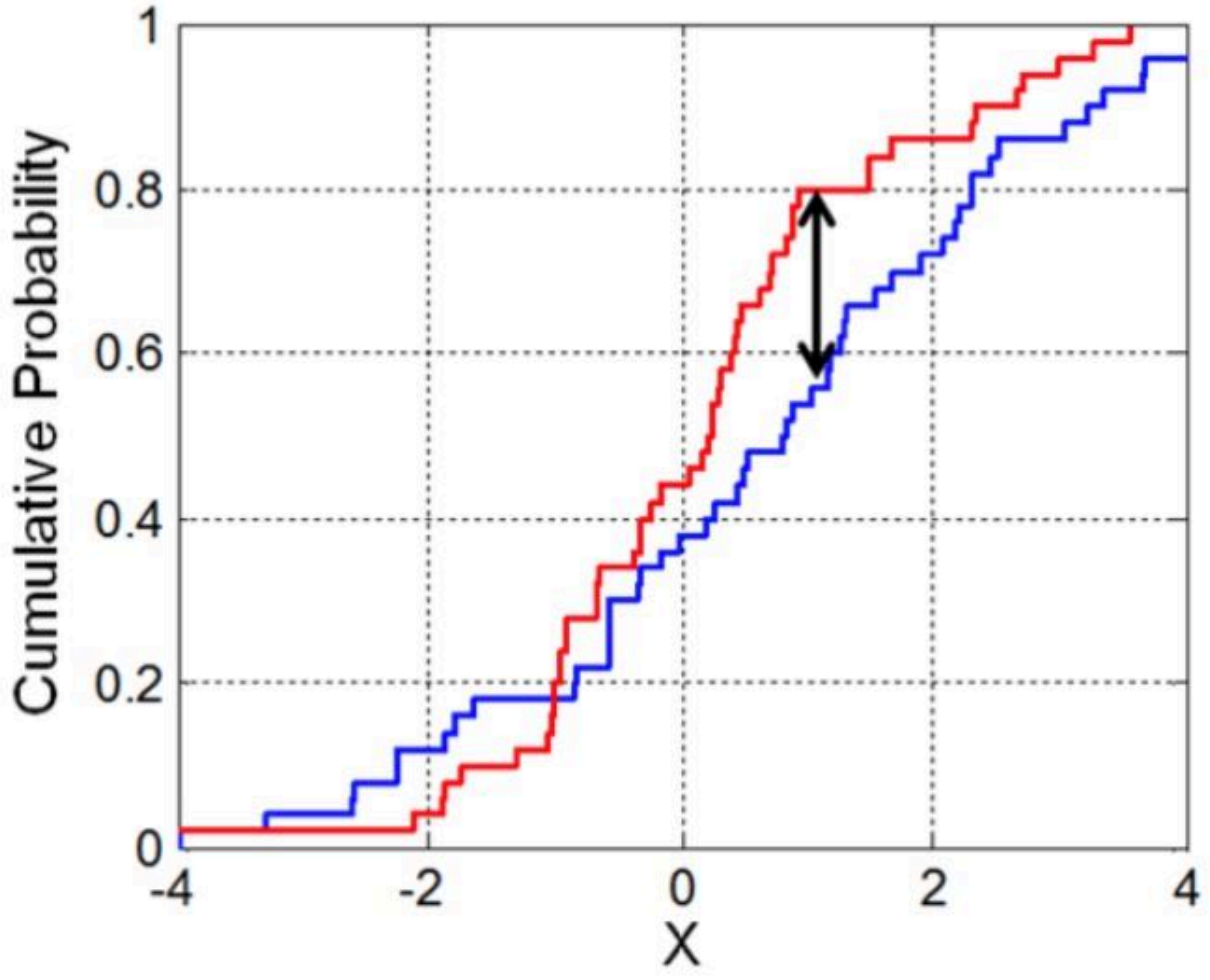




Source: Richard Gao

Kolmogorov-Smirnov test

Comparing cumulative distributions empirically

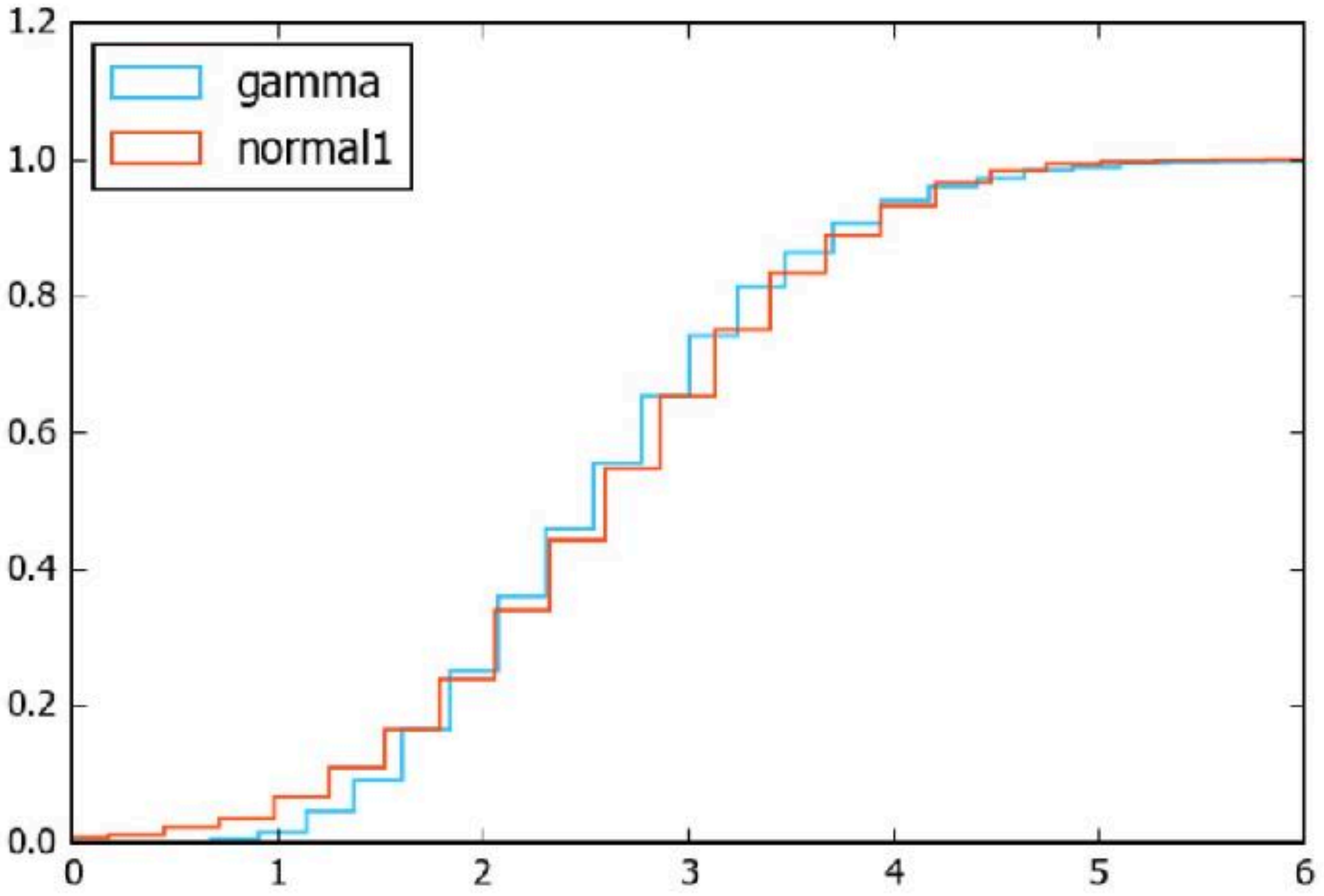


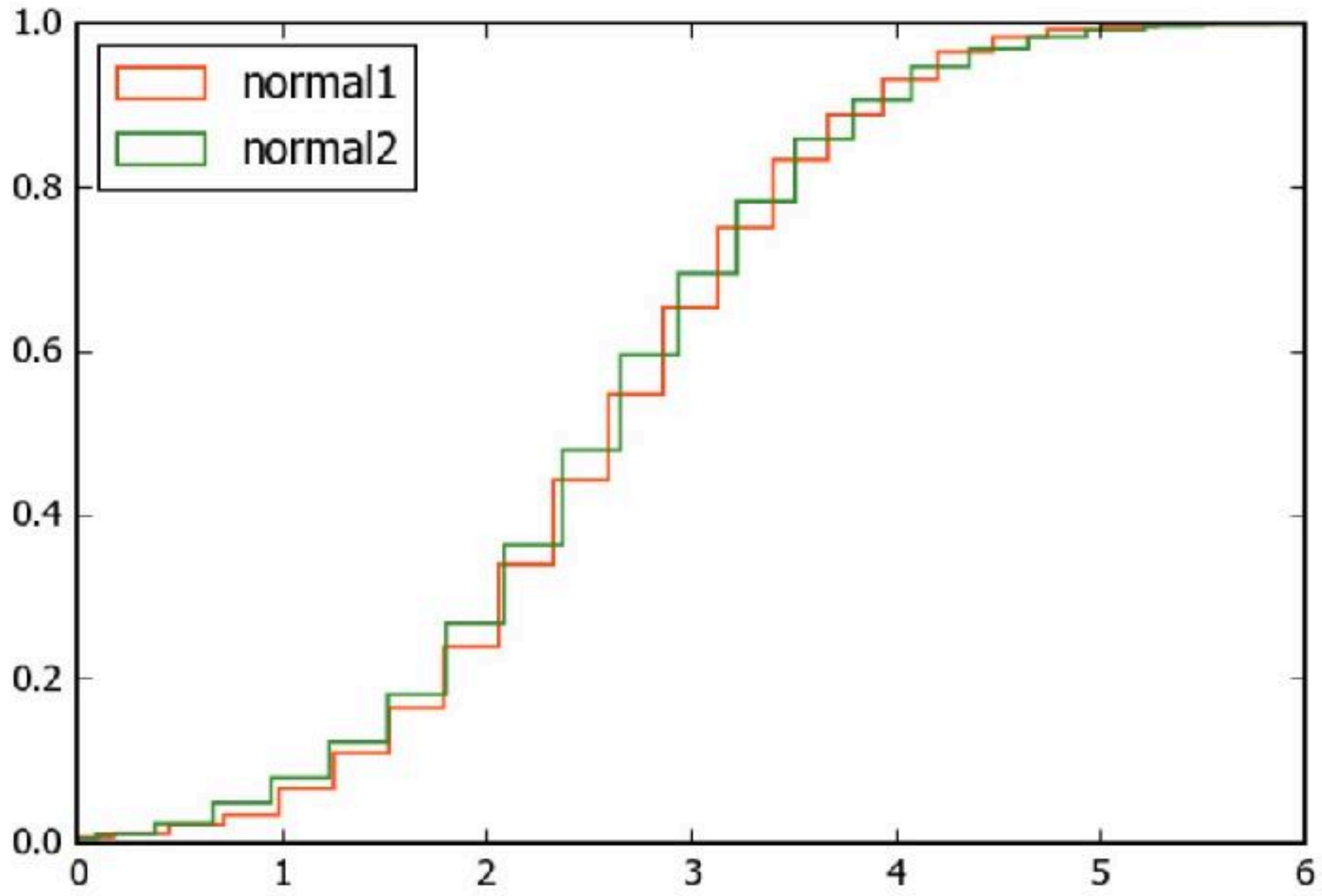
Find the maximum difference between the CDFs.

Source: Richard Gao

Kolmogorov-Smirnov test

Very sensitive!



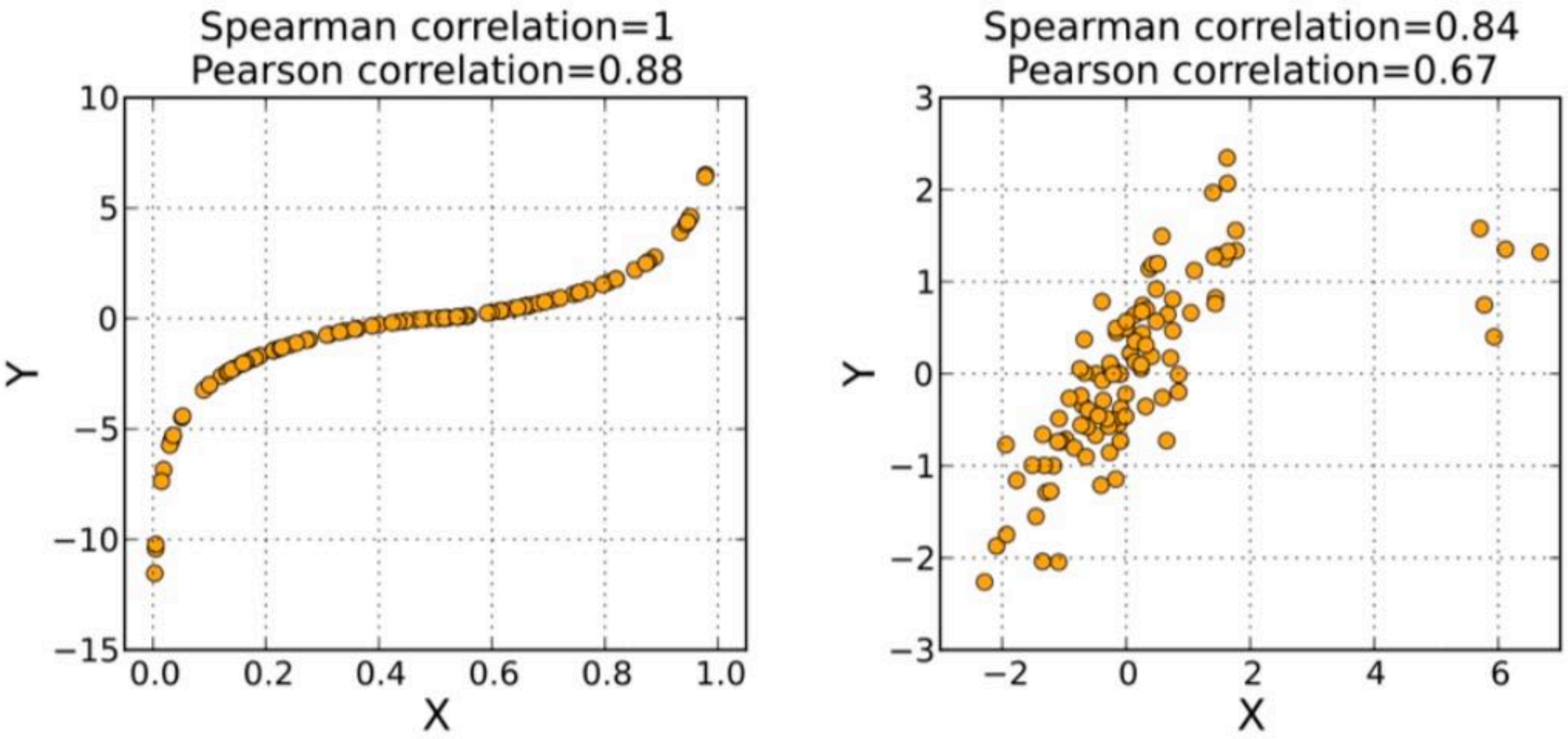


```
gamma vs. normal1: p = 0.0106803628411
normal1 vs. normal2: p = 0.550735998243
```

Source: Richard Gao

Rank correlation

Spearman Correlation - Non-linear but monotonic relationships



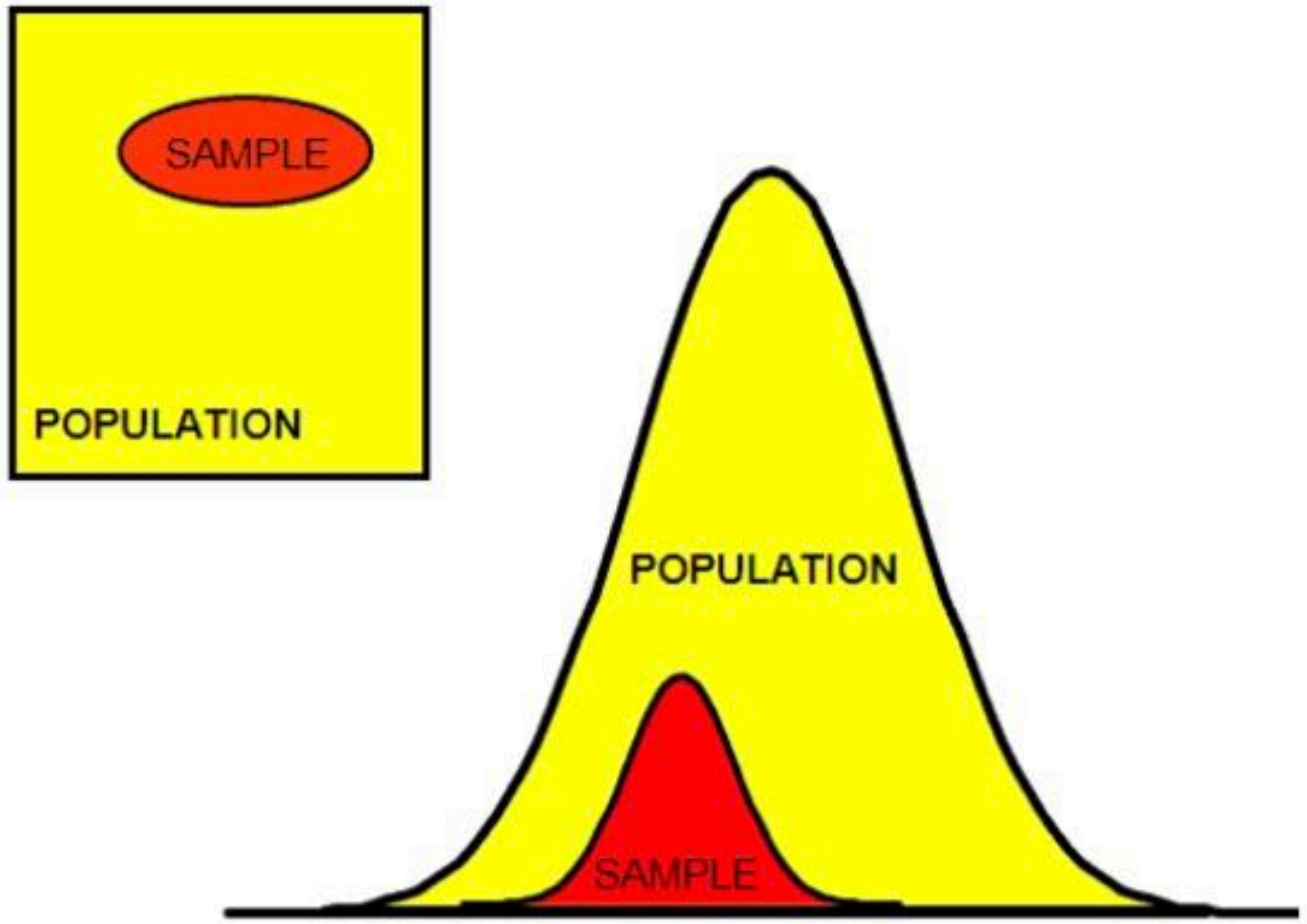
Dampens effect of outliers!

Source: Richard Gao

Bootstrapping (resampling)

Question:

- How can we build a more realistic “null distribution” for the sample estimate without knowing the population it’s drawn from?

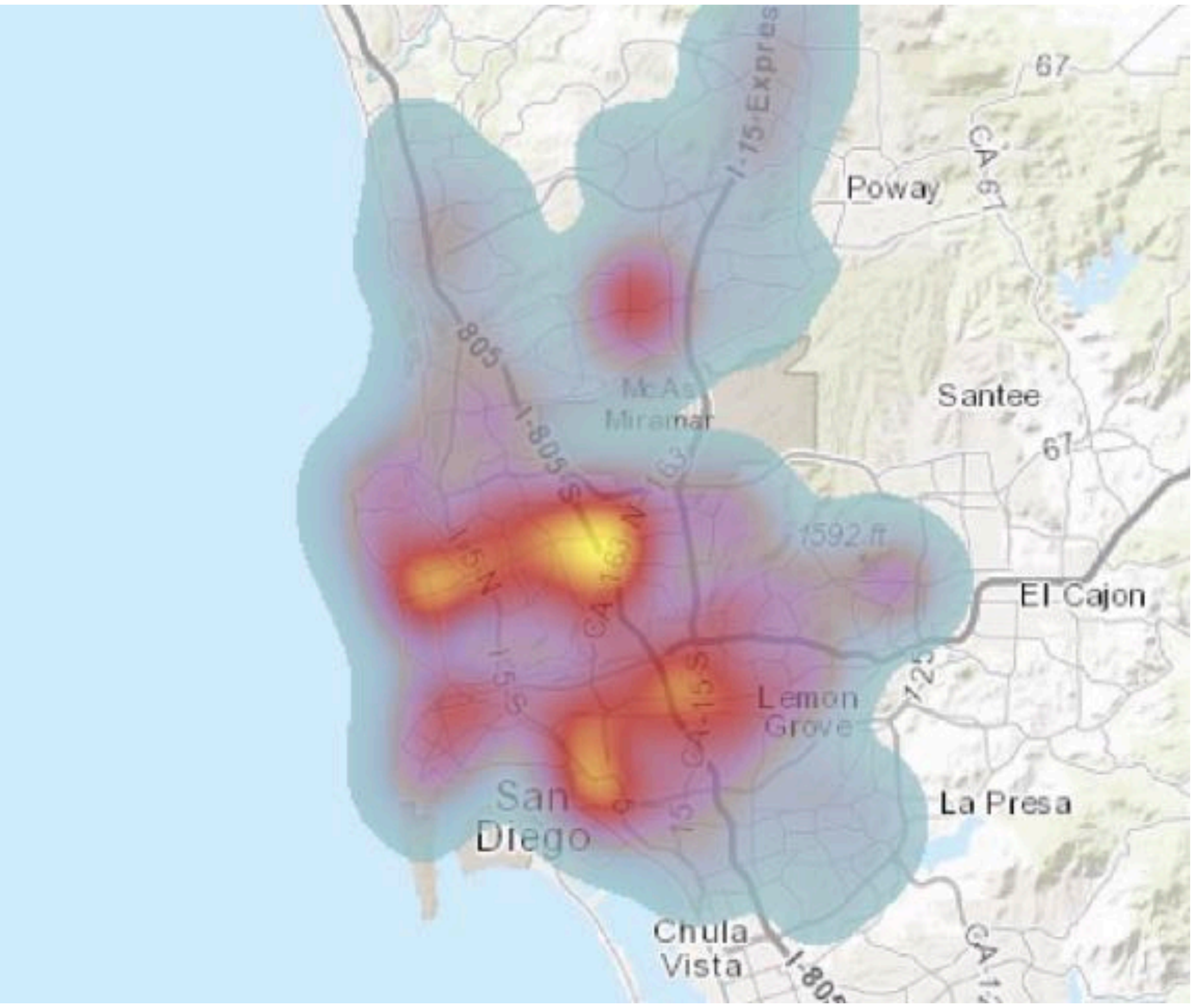


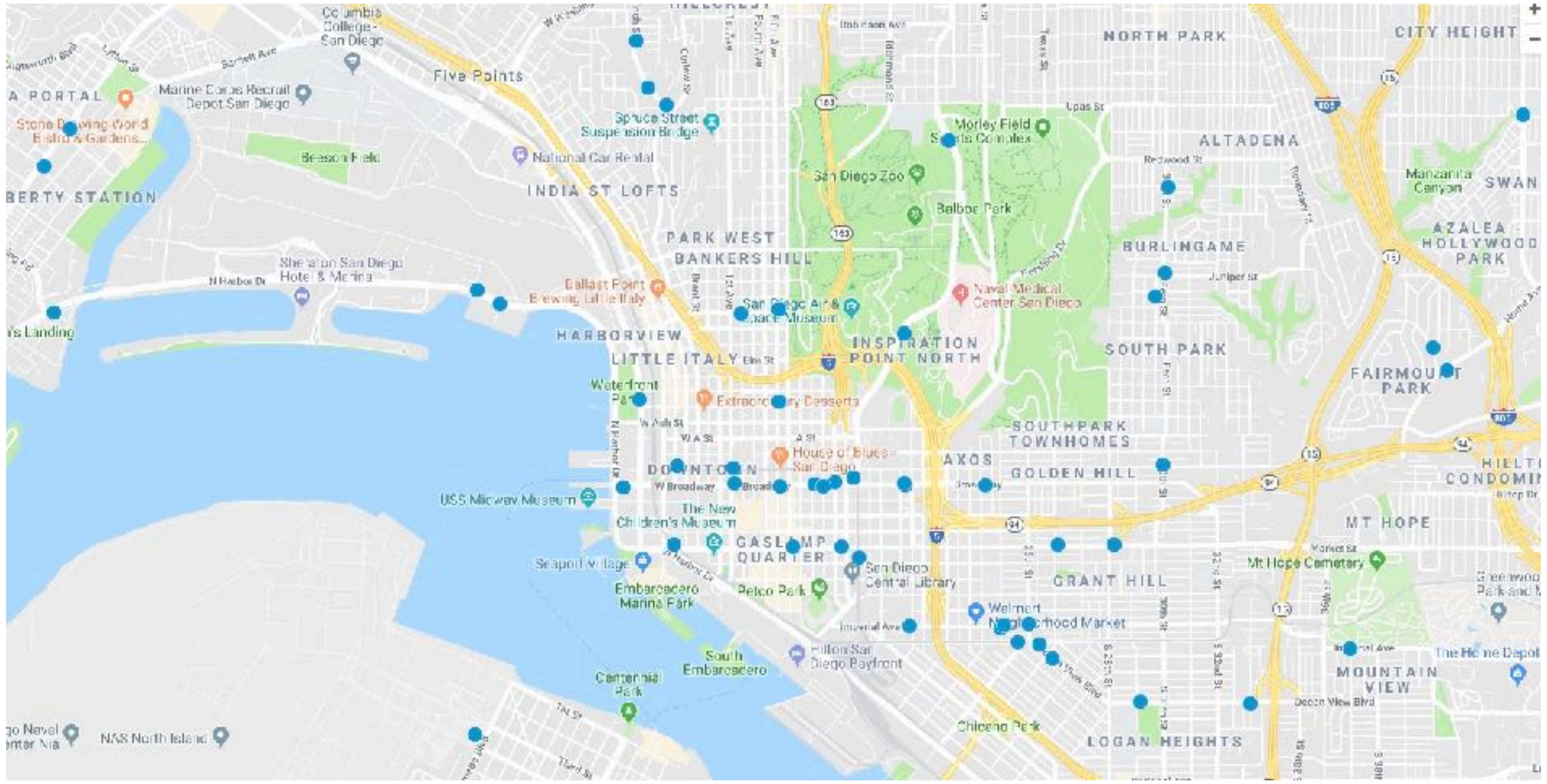
Source: Richard Gao

Bootstrapping (resampling)

Example Question:

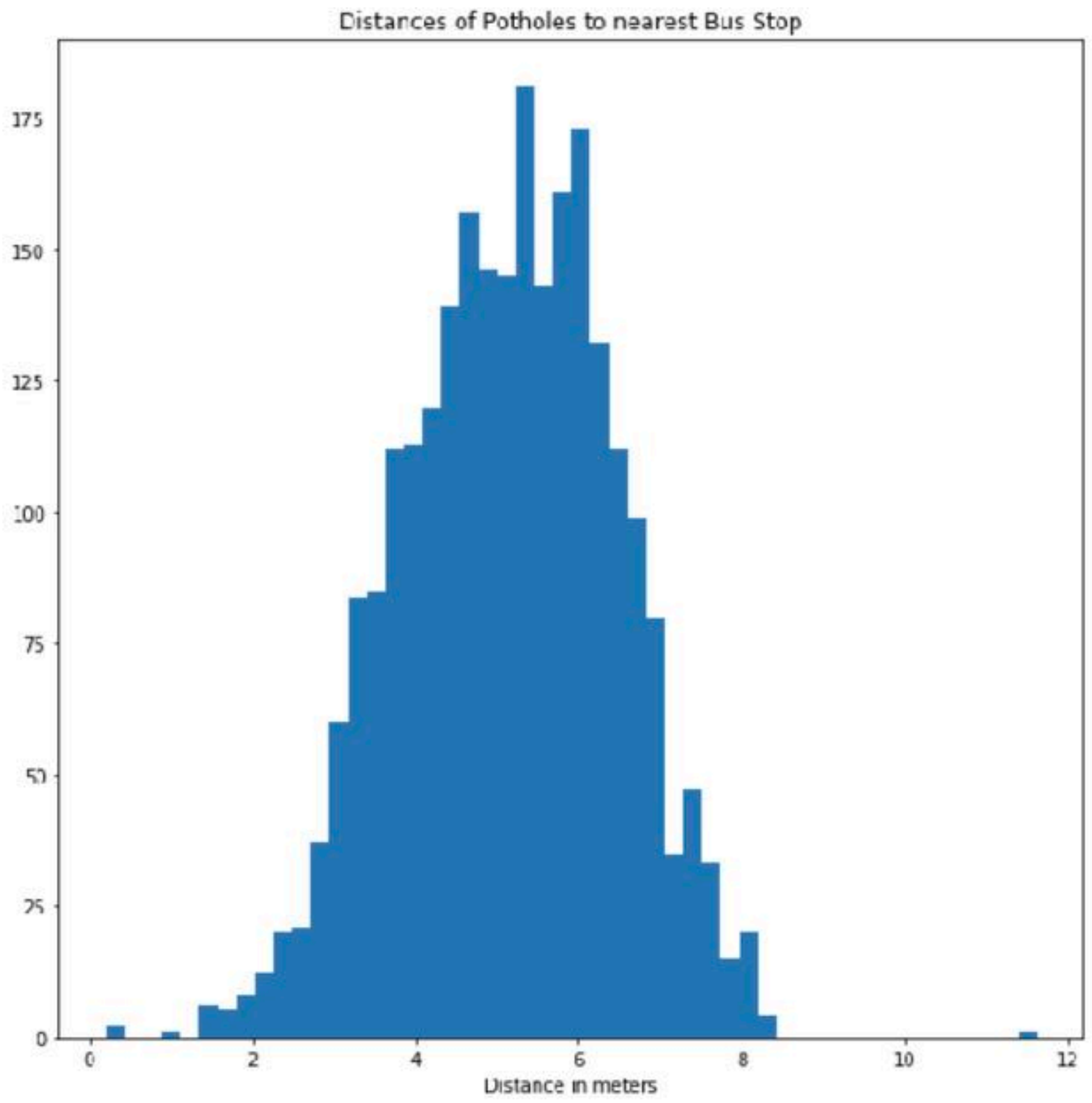
- Are San Diego's pot holes closer to bus stops than not?

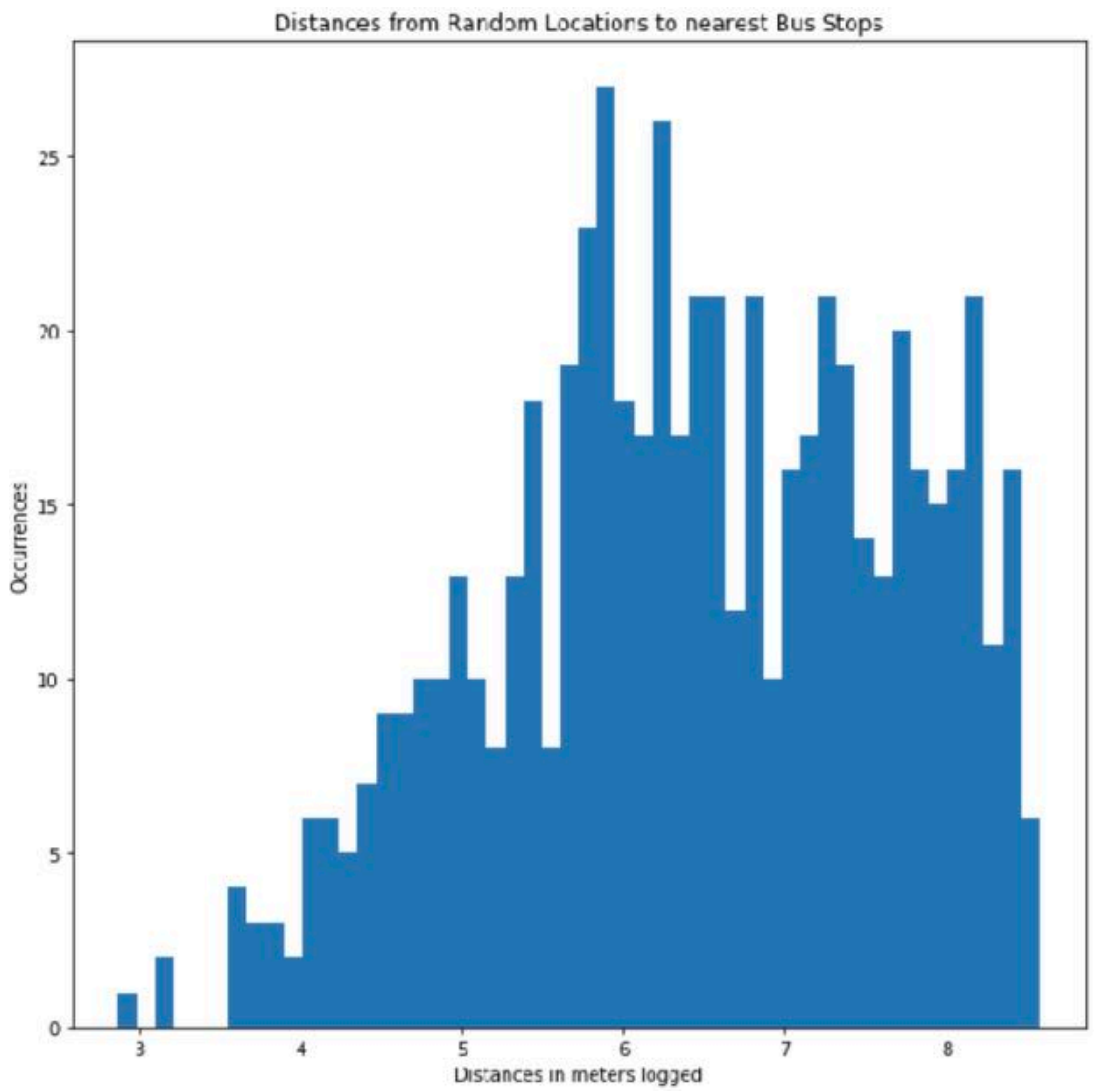




Source: Richard Gao

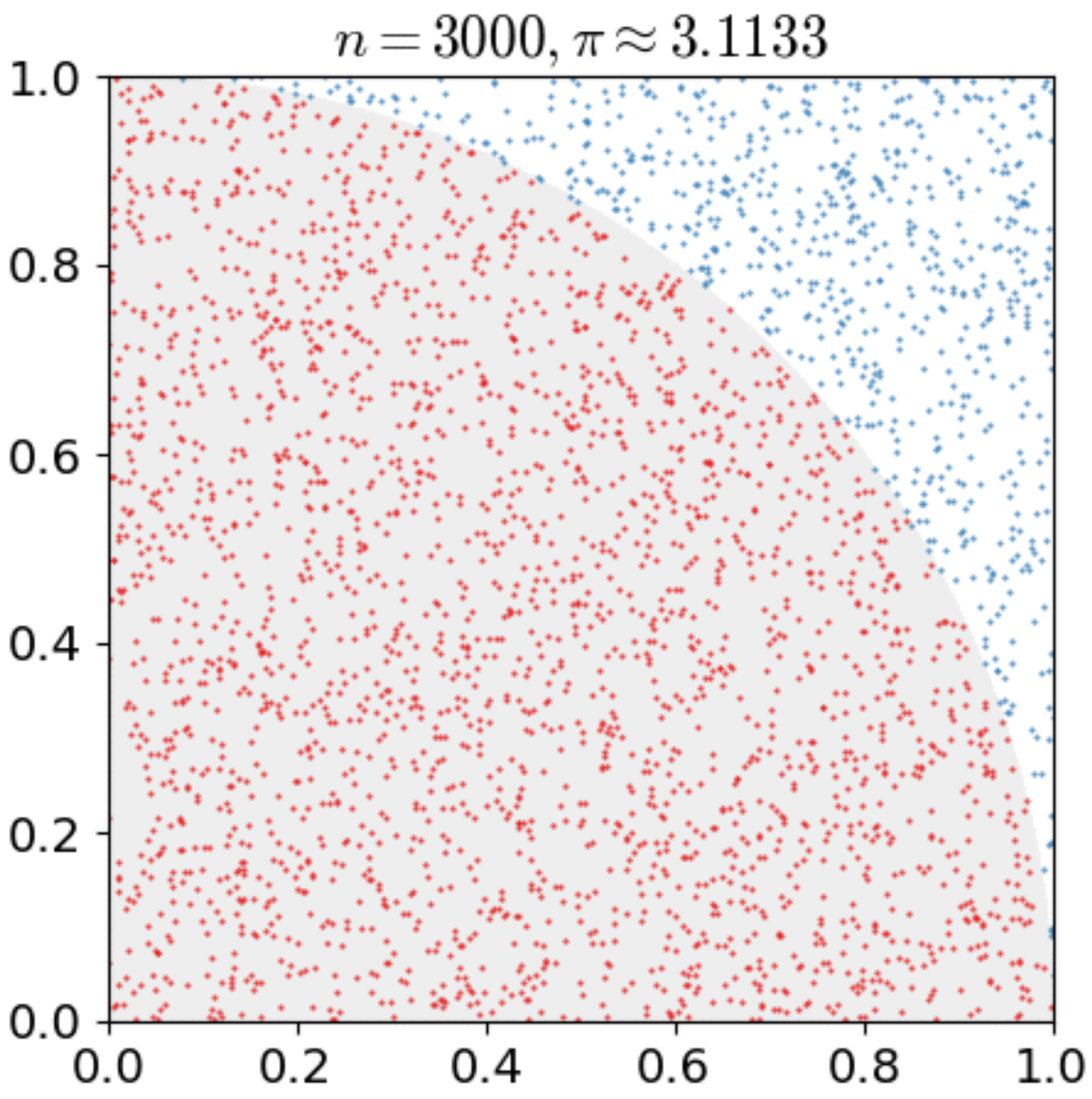
Bootstrapping (resampling)





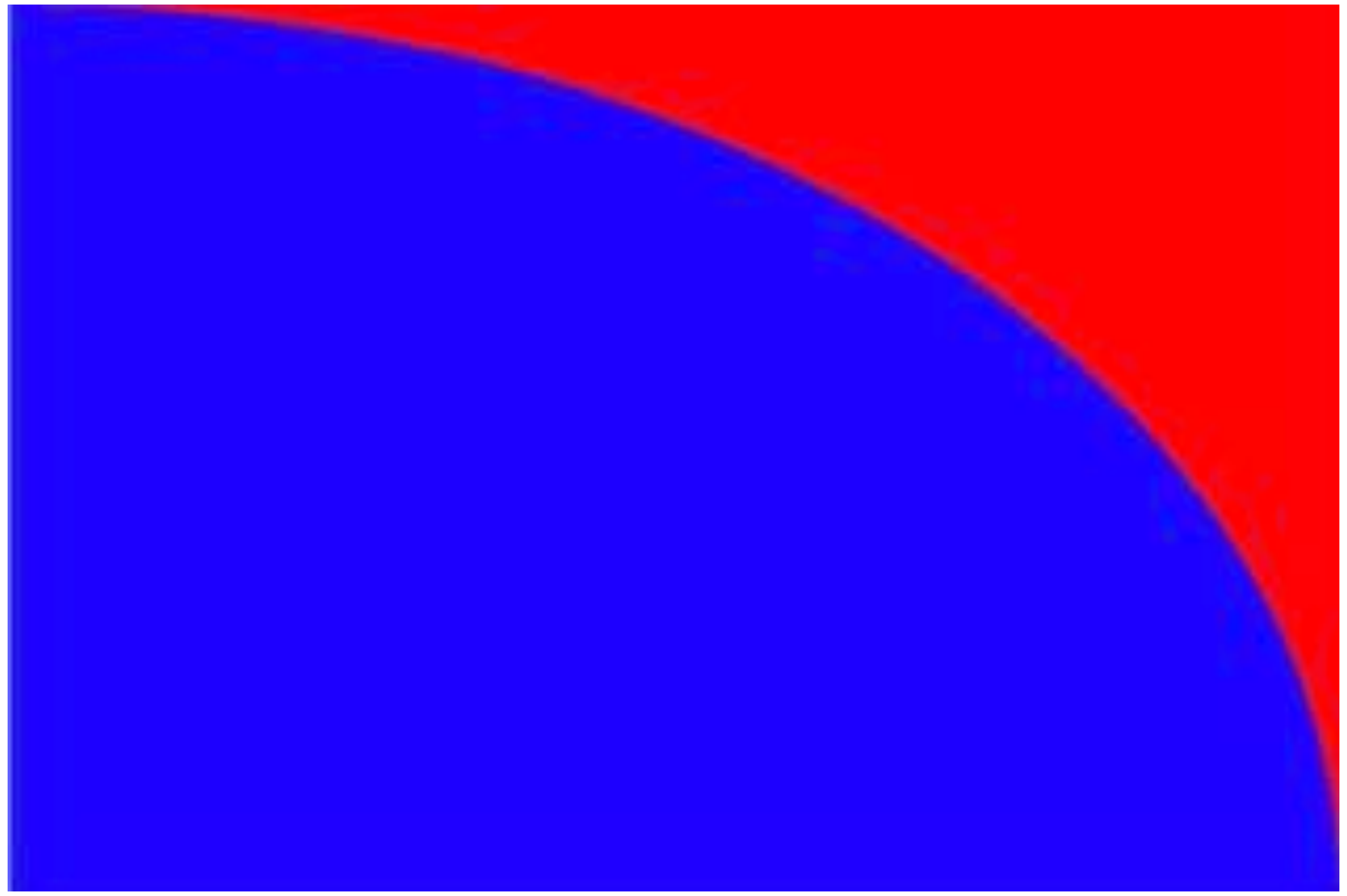
Source: Richard Gao

Monte Carlo (also resampling) - **π**



$n = 3000, \pi \approx 3.1133$

n = 10^8



```
3.1416079600000000 result
3.1415926535897931 real pi
```

Source: Wikipedia (Monte Carlo); *The Glowing Python*

Bradley Voytek, Ph.D.
UC San Diego
Neural and Data Analytics Laboratory

Department of Cognitive Science
Neurosciences Graduate Program
Halıcıoğlu Data Science Institute

bvoytek@ucsd.edu
@bradleyvoytek