Uncertainty visualization

CS594: Big Data Visualization & Analytics

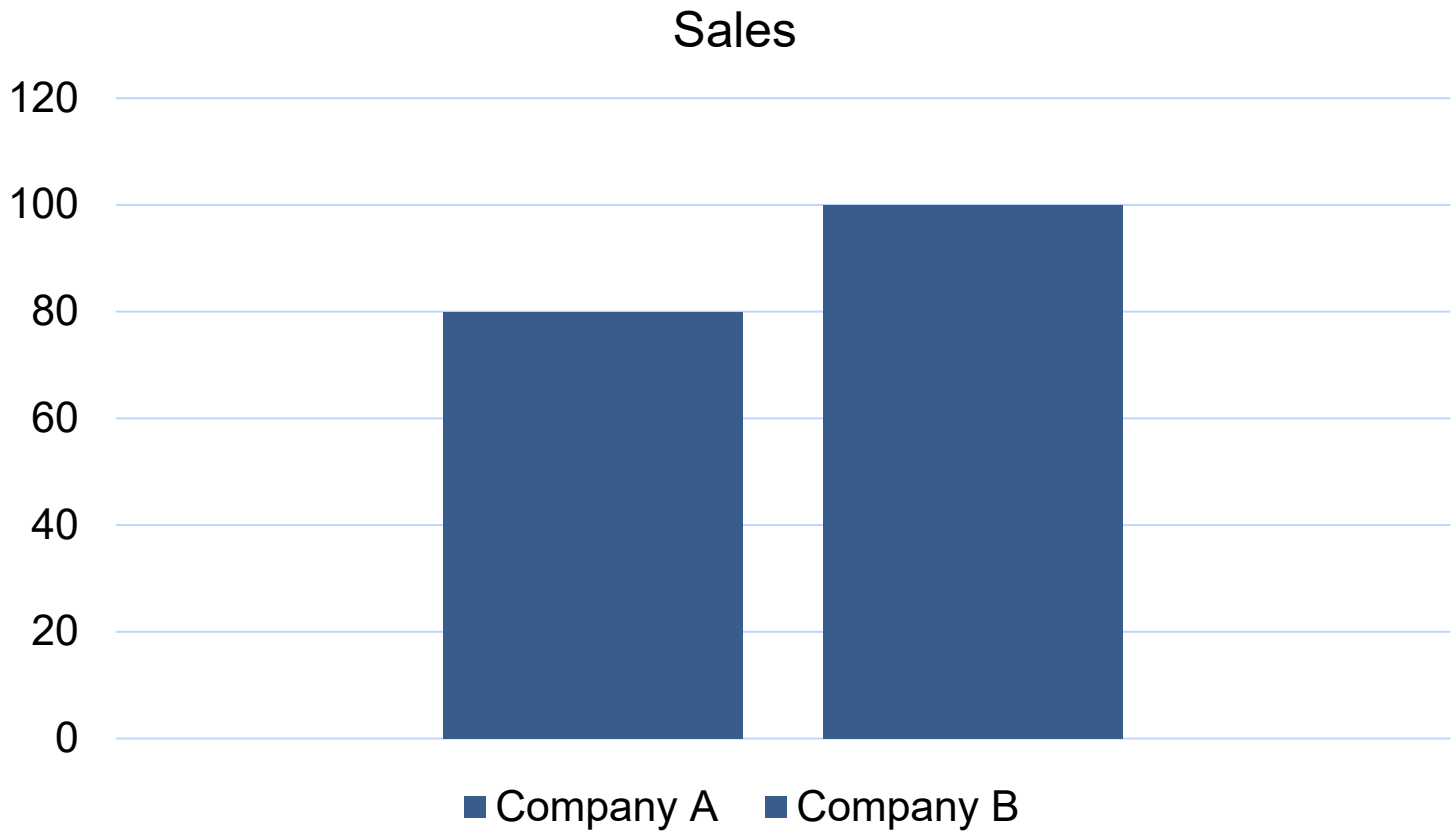
Fabio Miranda

https://fmiranda.me

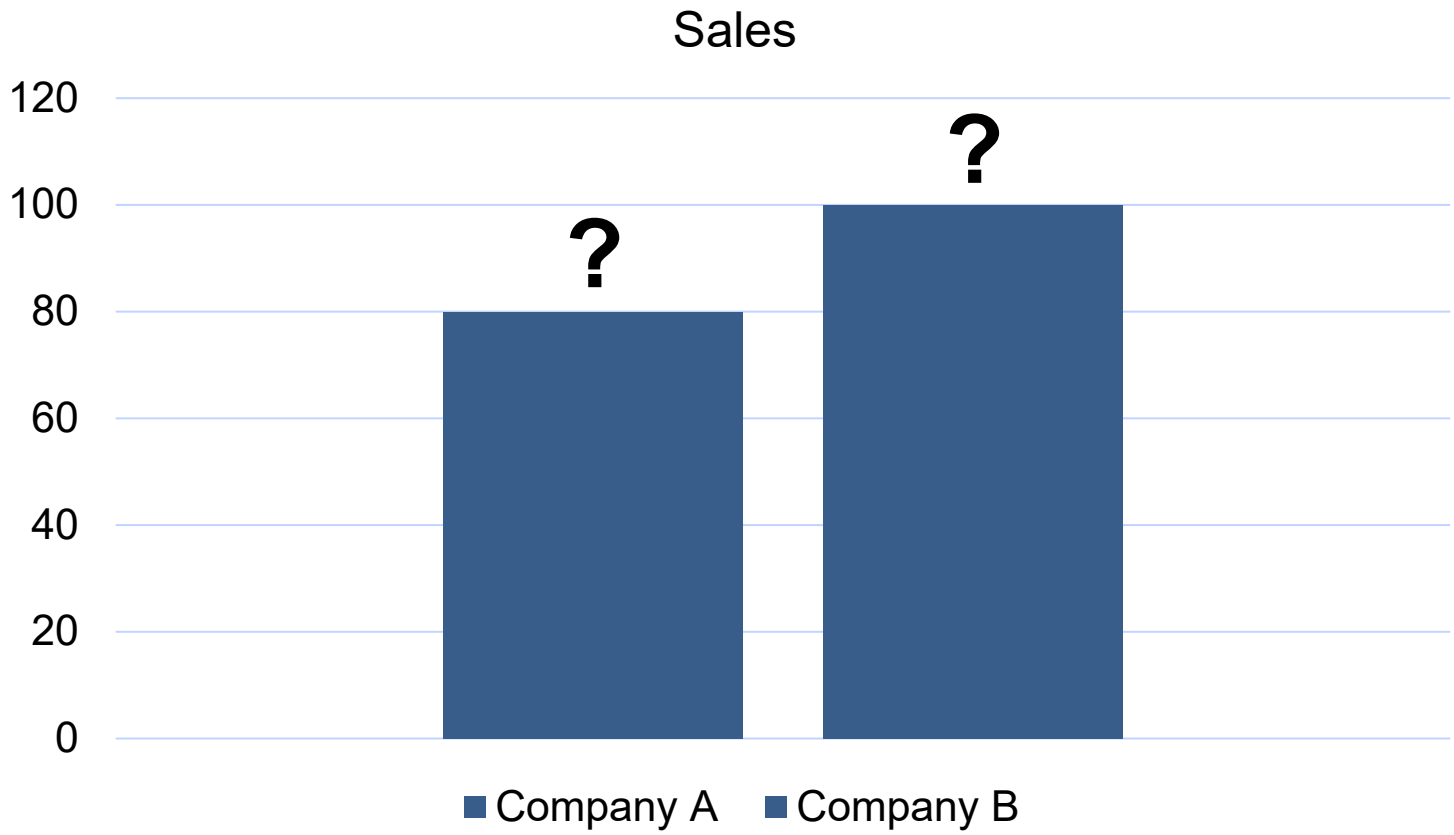
Uncertainty

- What is uncertainty?
  - Doubt
  - Risk
  - Variability
  - Error
  - Lack of knowledge
- Uncertainty is inherent to most data, and can enter the analysis pipeline during measurement, modeling and forecasting phases.
- It is important to effectively communicate uncertainty to establish transparency.

Uncertainty: a simple example

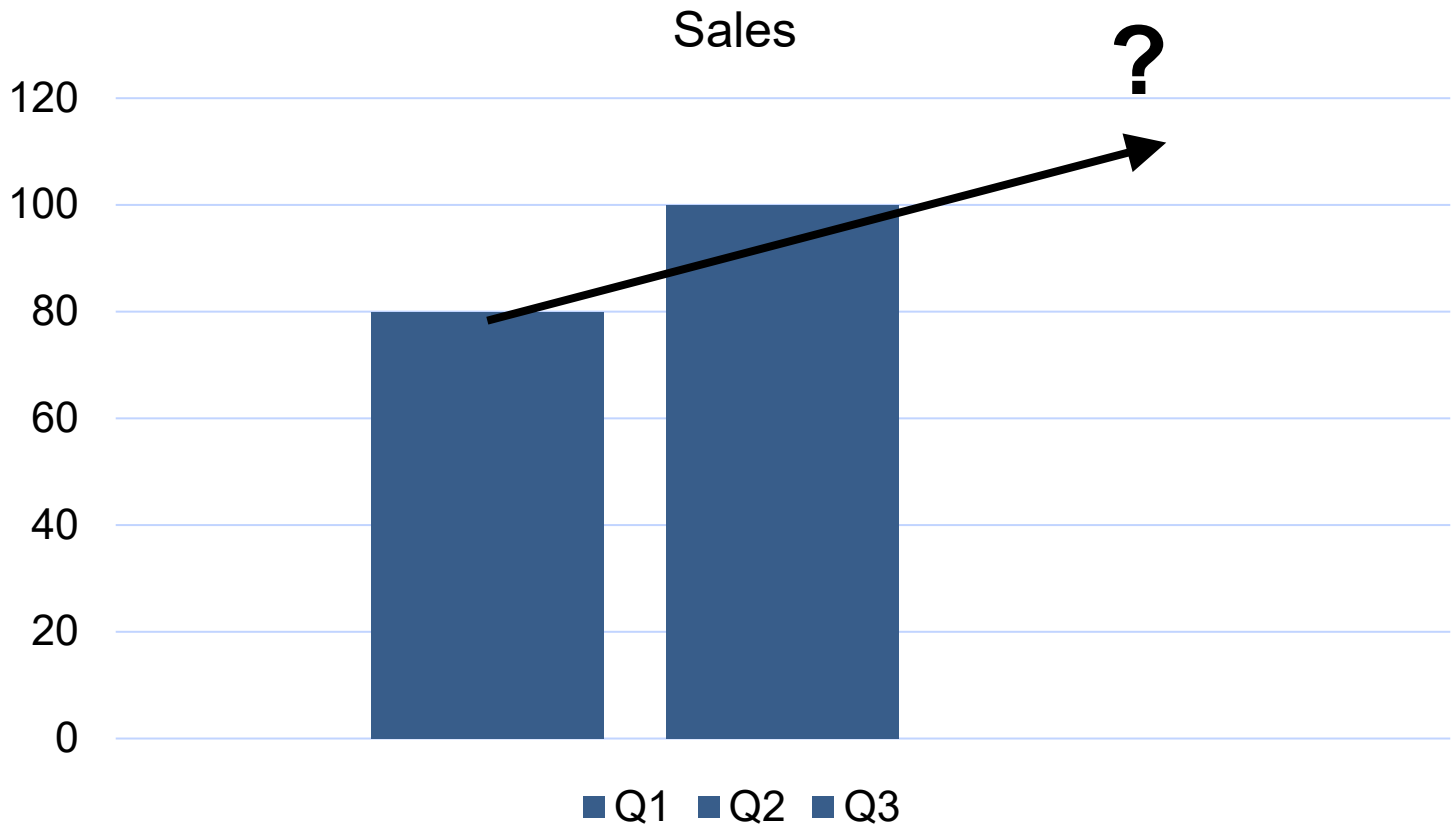


Measurement uncertainty



Forecast uncertainty

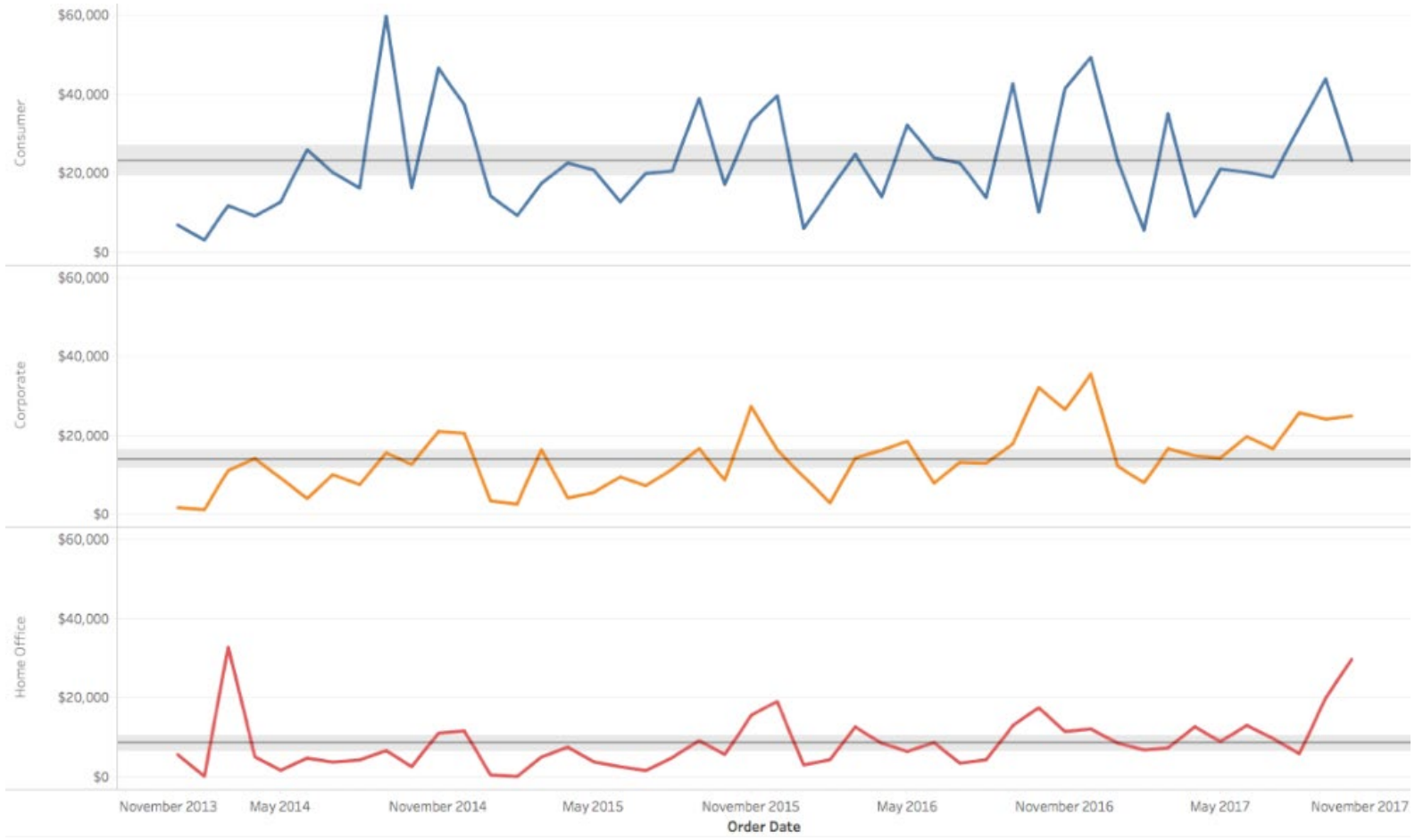
Model uncertainty



Sources of uncertainty in data

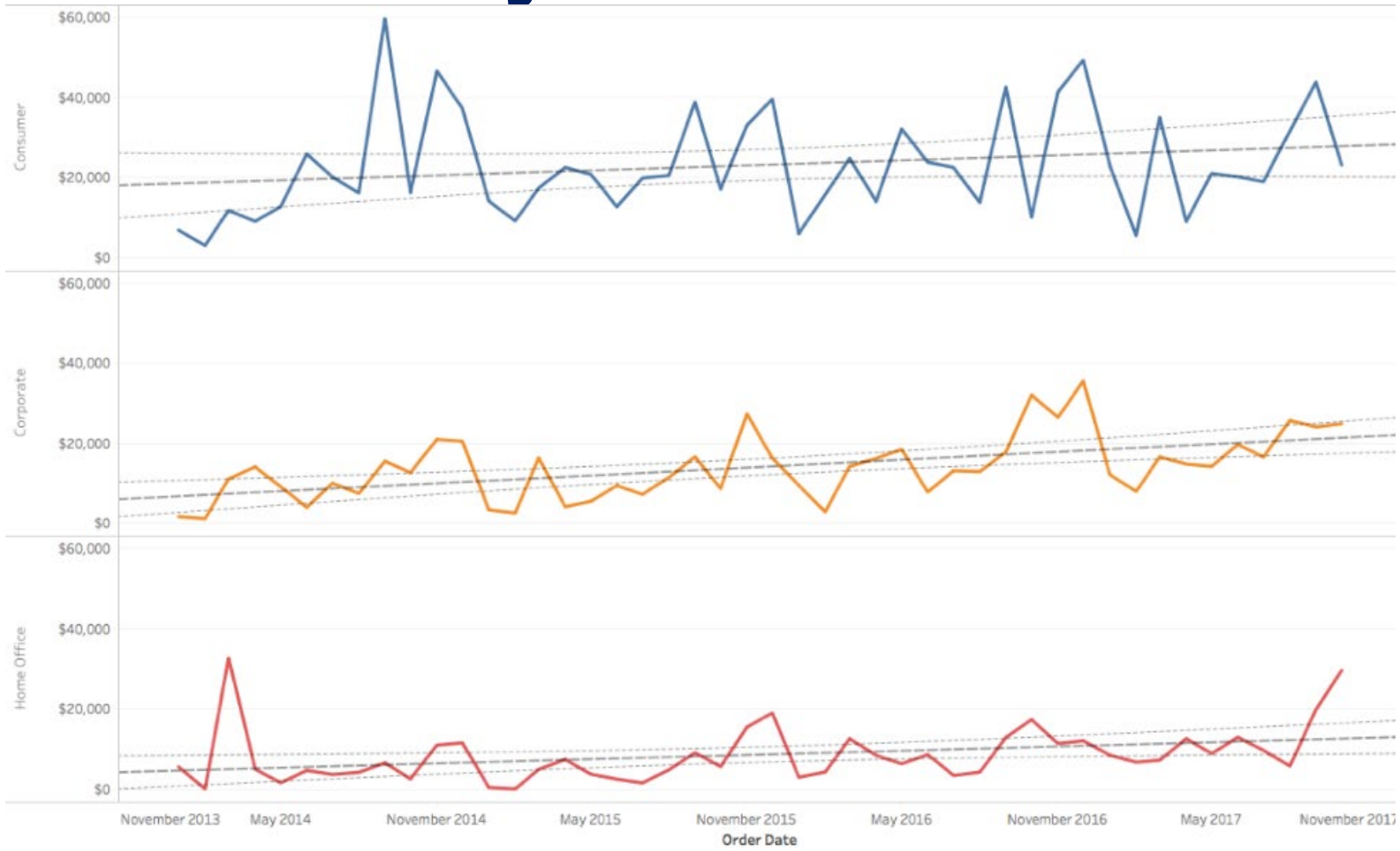
- Measurement: are we sure about the data?
- Forecast uncertainty: what will happen with the data next?
- Model uncertainty: how the data fit together?

Measurement uncertainty



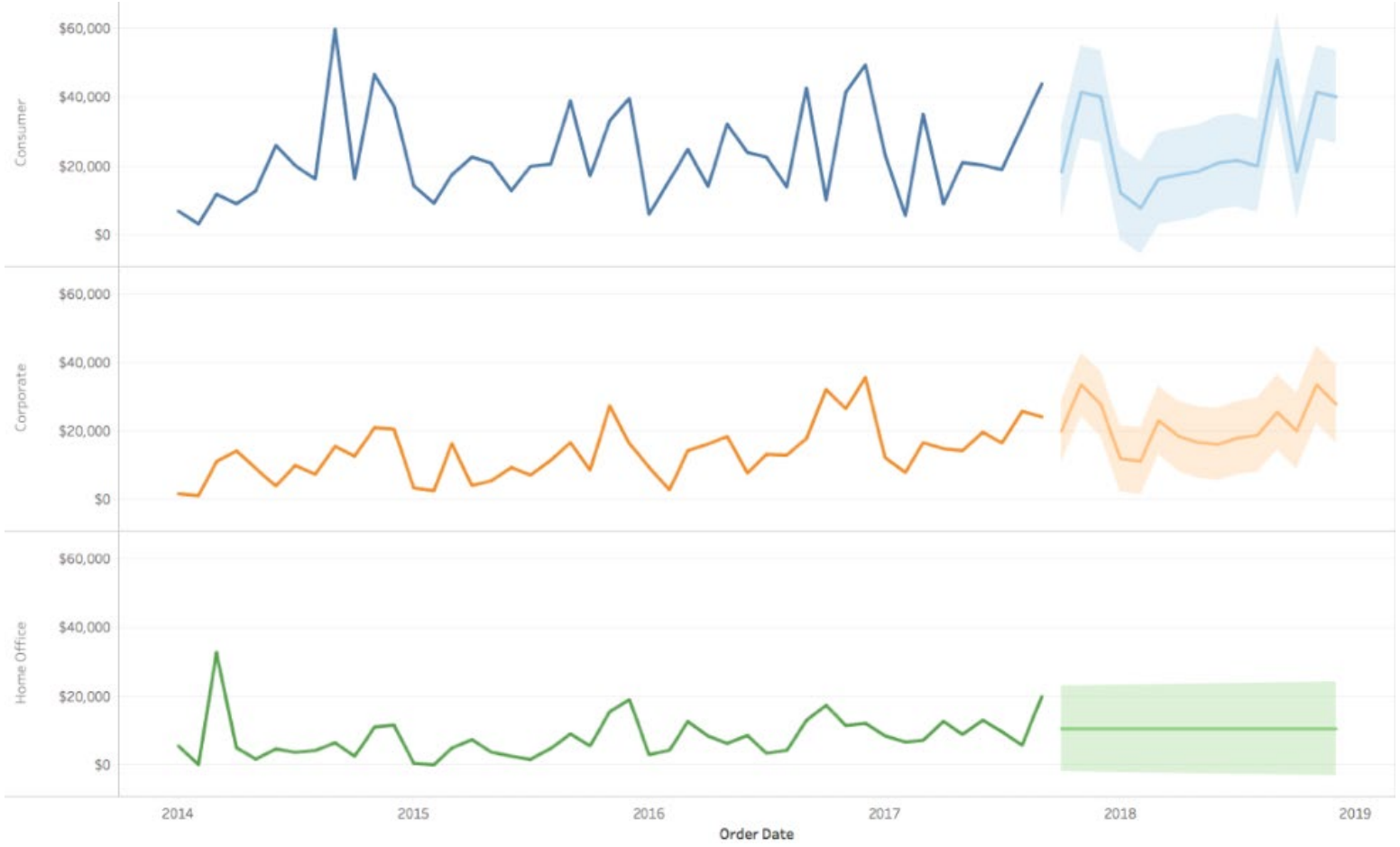
[Correll, Tableau Research]

Model uncertainty



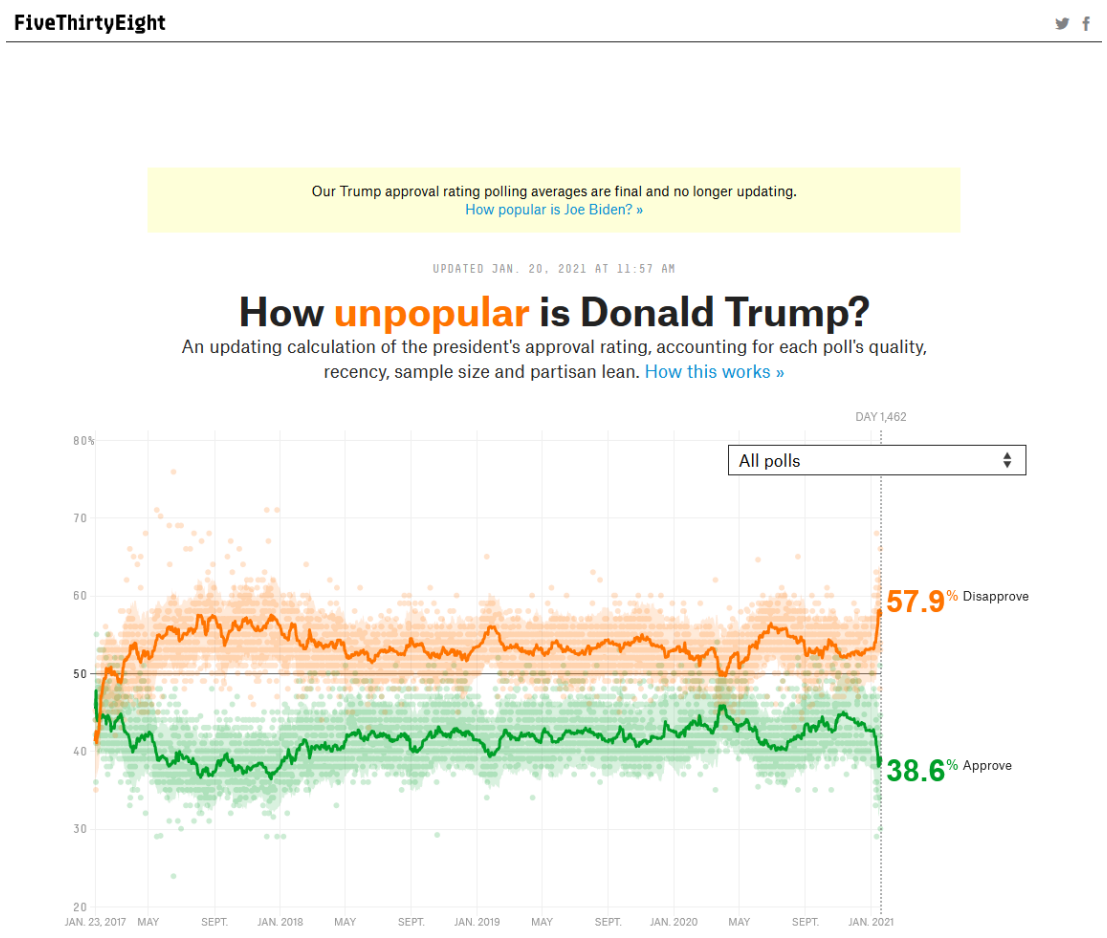
[Correll, Tableau Research]

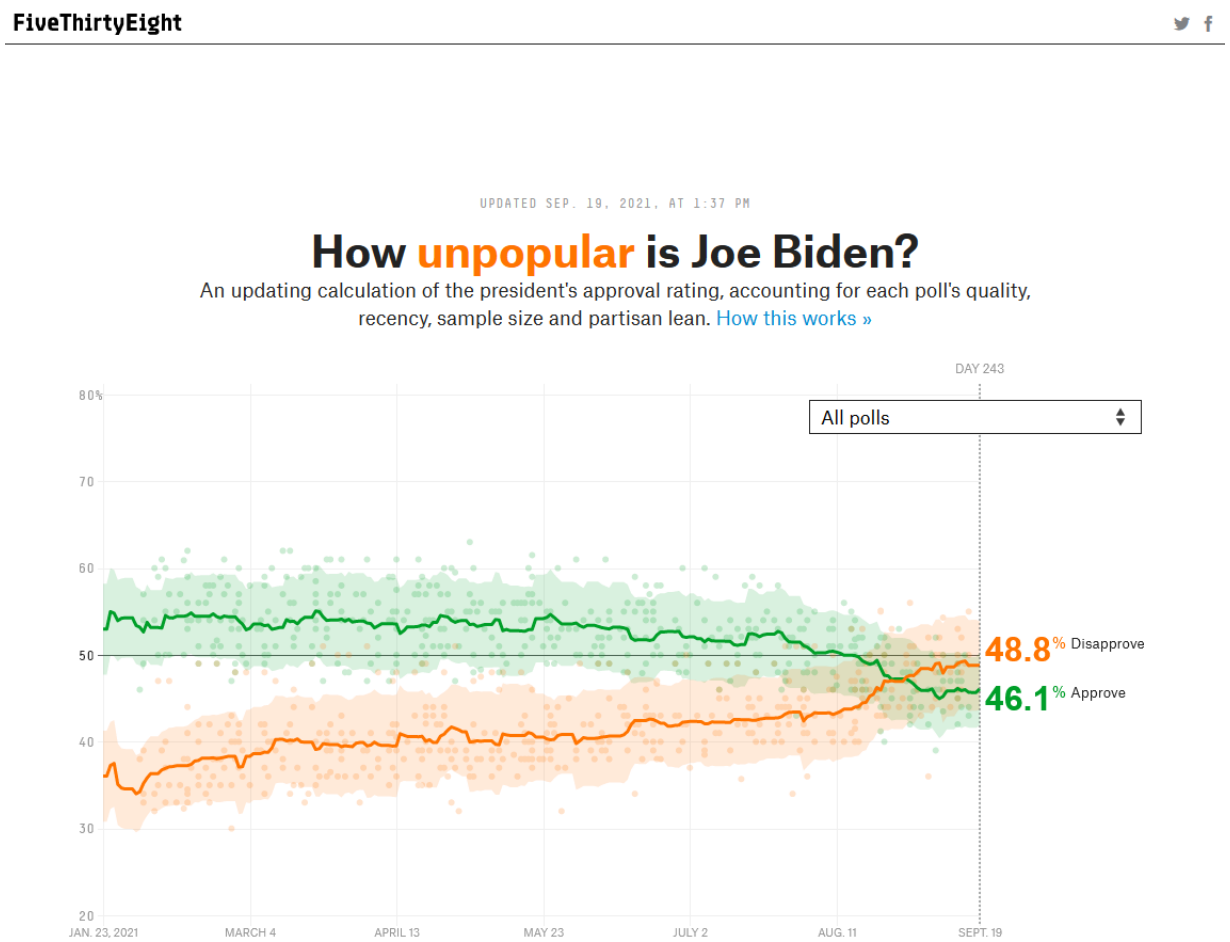
Forecast uncertainty



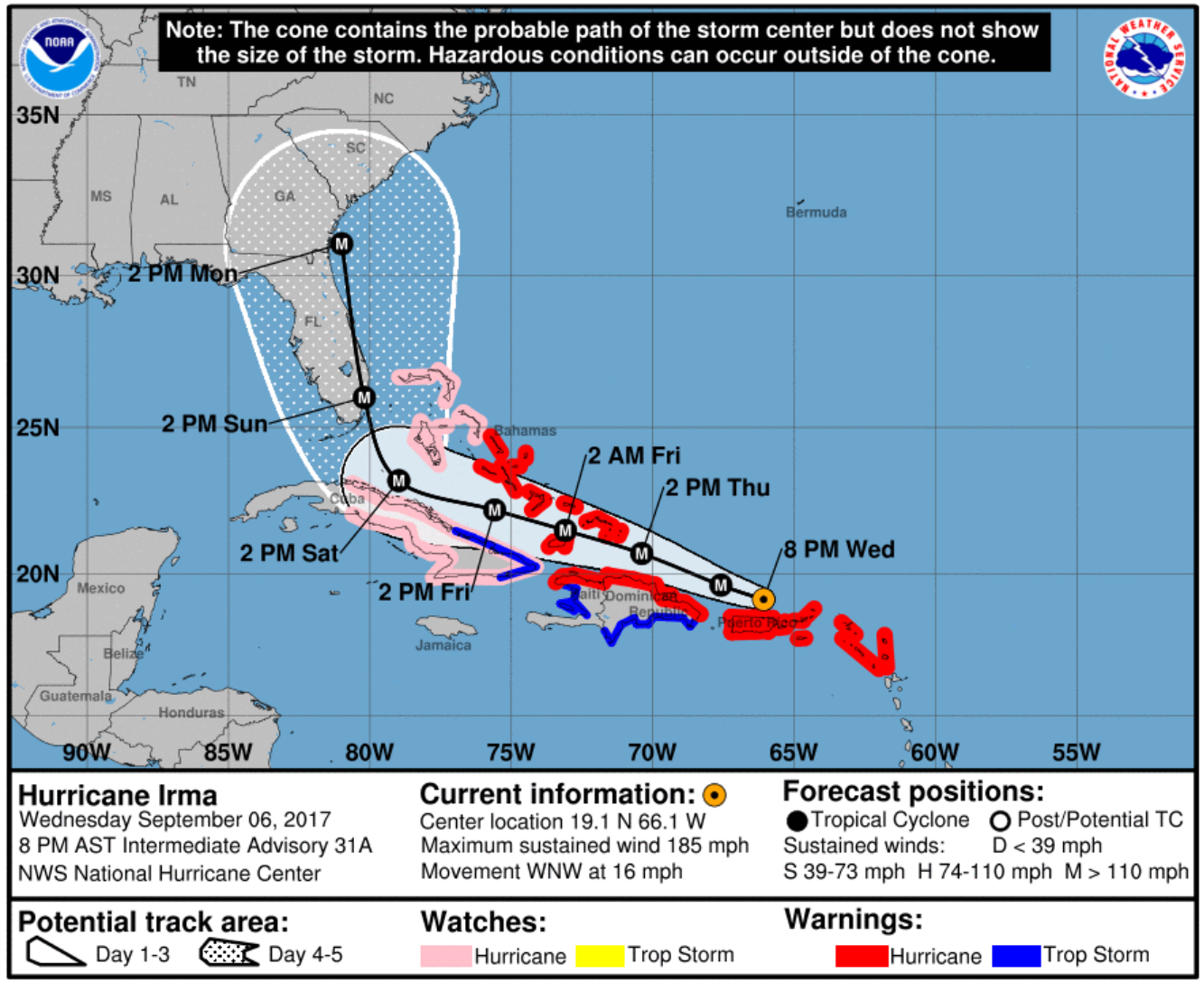
[Correll, Tableau Research]

Approval rating

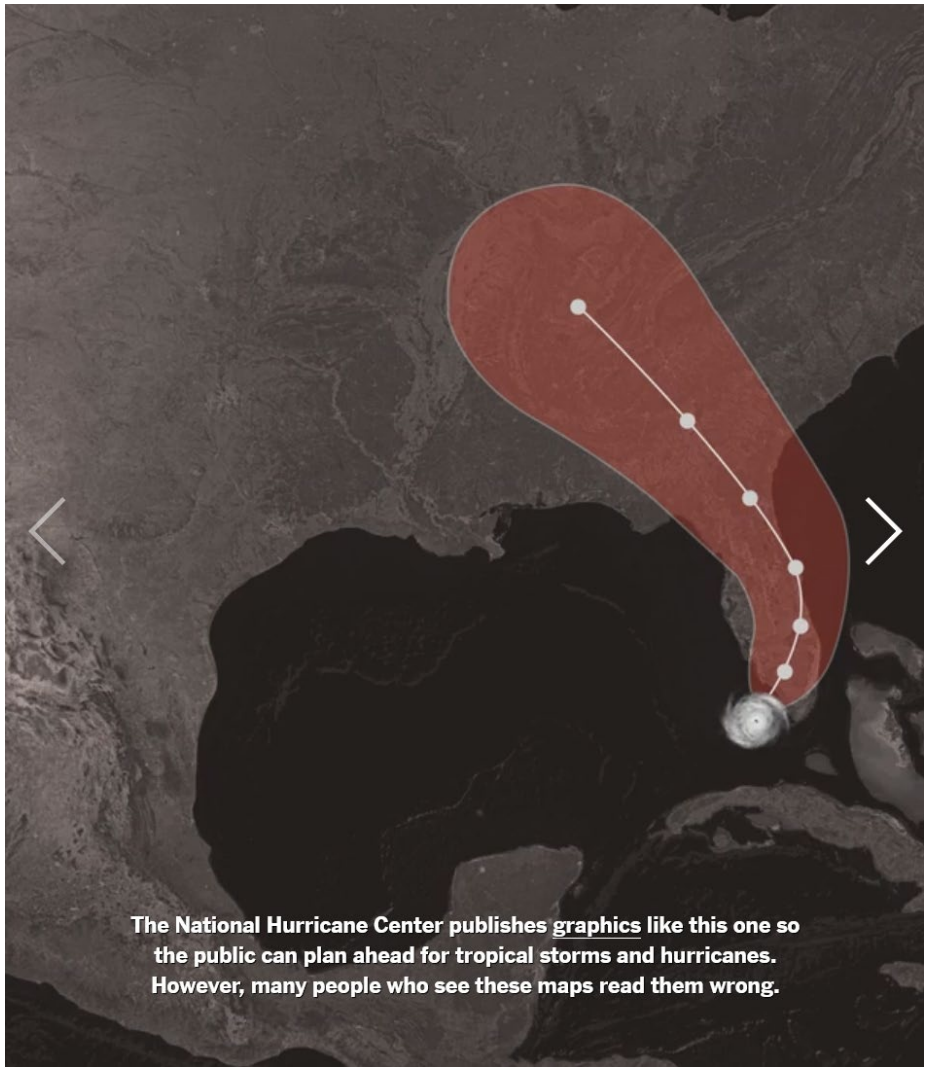


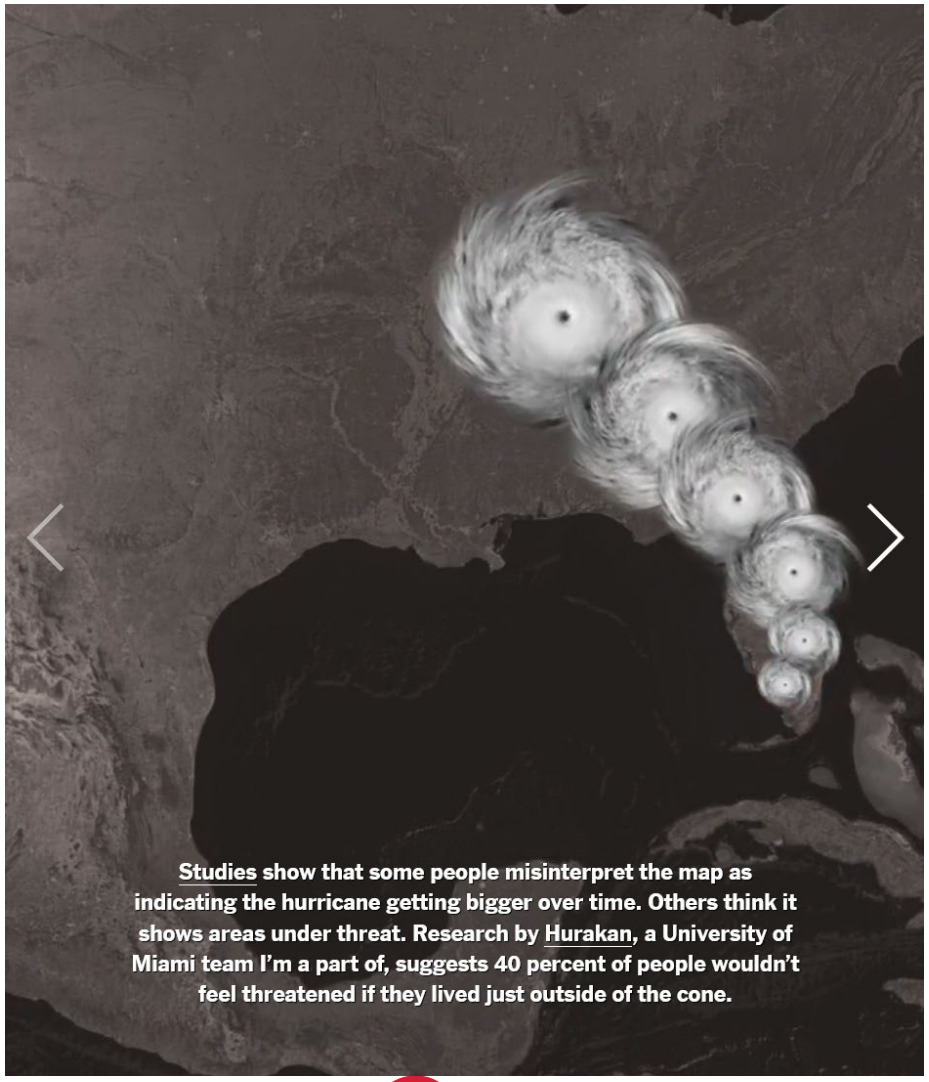


Hurricane Irma



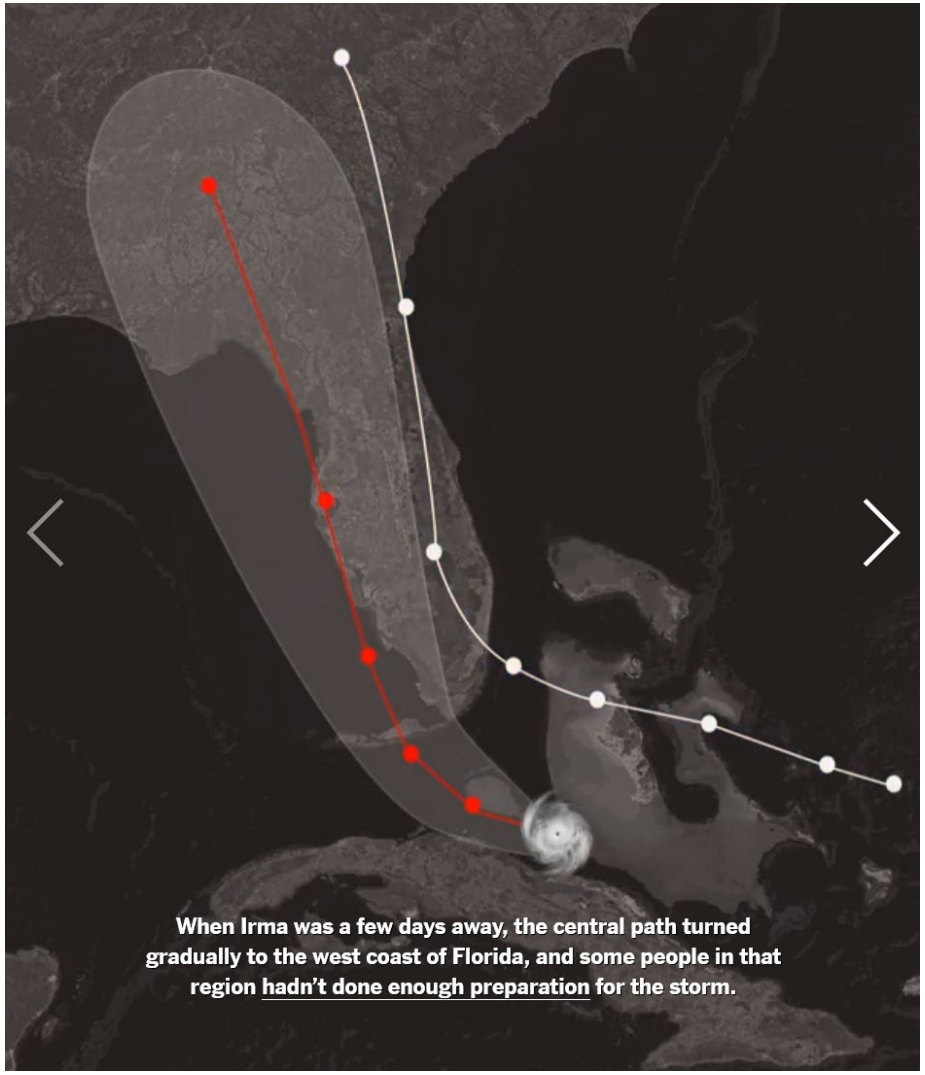
Hurricane Irma





[NY Times]

Hurricane Irma

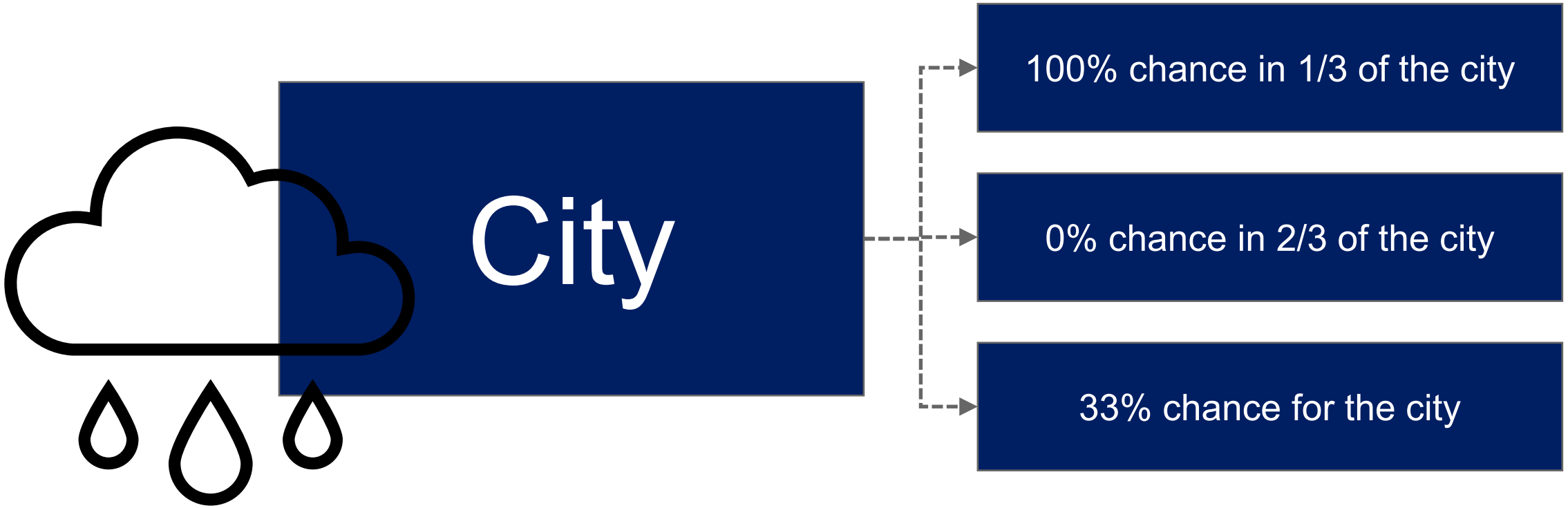


[NY Times]

Hurricane Irma

[NY Times]

Chance of rain



What does uncertainty mean?

"*Any one of a number of potentially interconnected quantitative, qualitative, or factors that affect the quality, reliability, or utility of your data or data-driven decisions. Anything that can cause you to be unsure about your data or how to use it*."

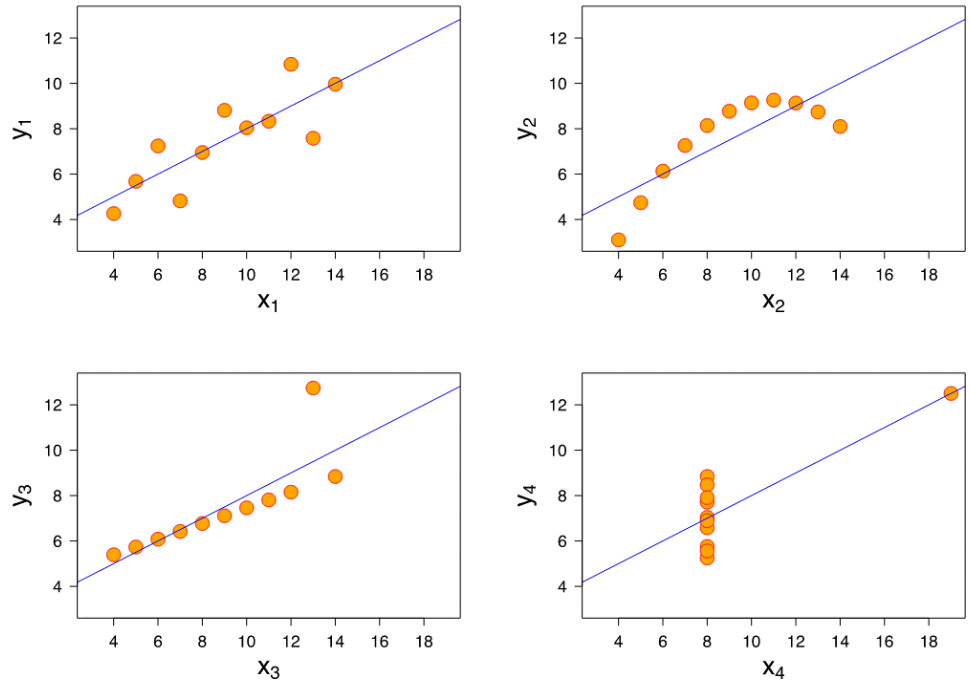
Michael Correll
Tableau Research

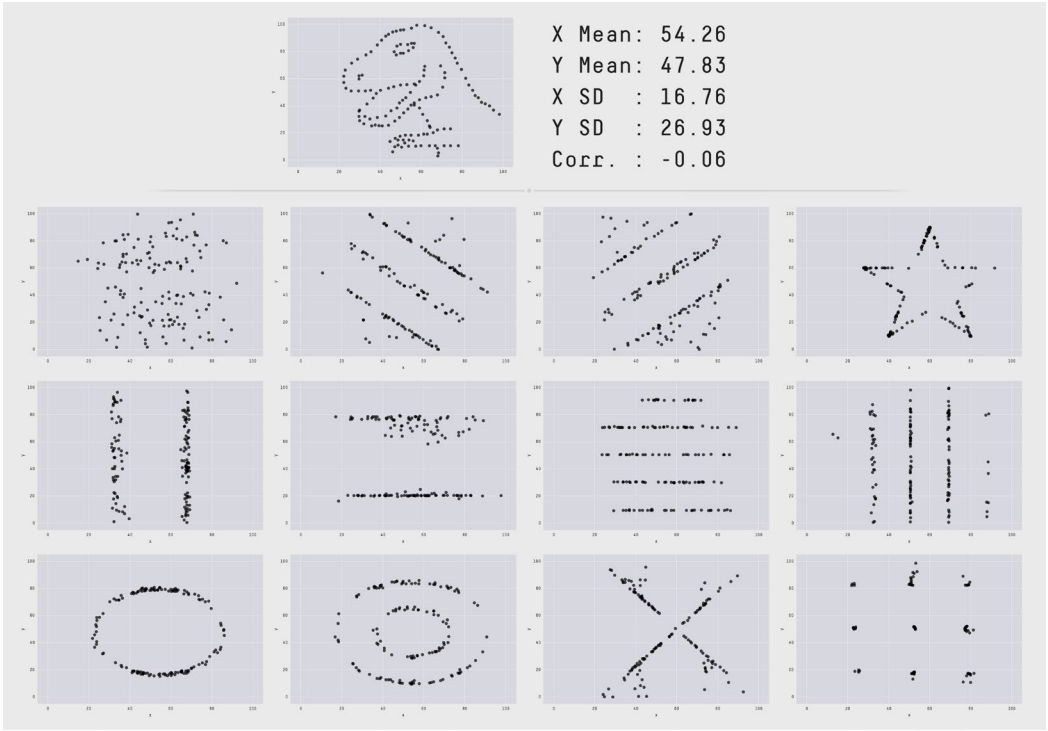
Sources of uncertainty in data

- Data ensembles: collections of simulations that explore a parameter space or realize a collection of instances.
- Model inadequacy: lack of knowledge of true phenomena that is being approximated.
- Algorithmic approximations: caused when translating models to computational settings.

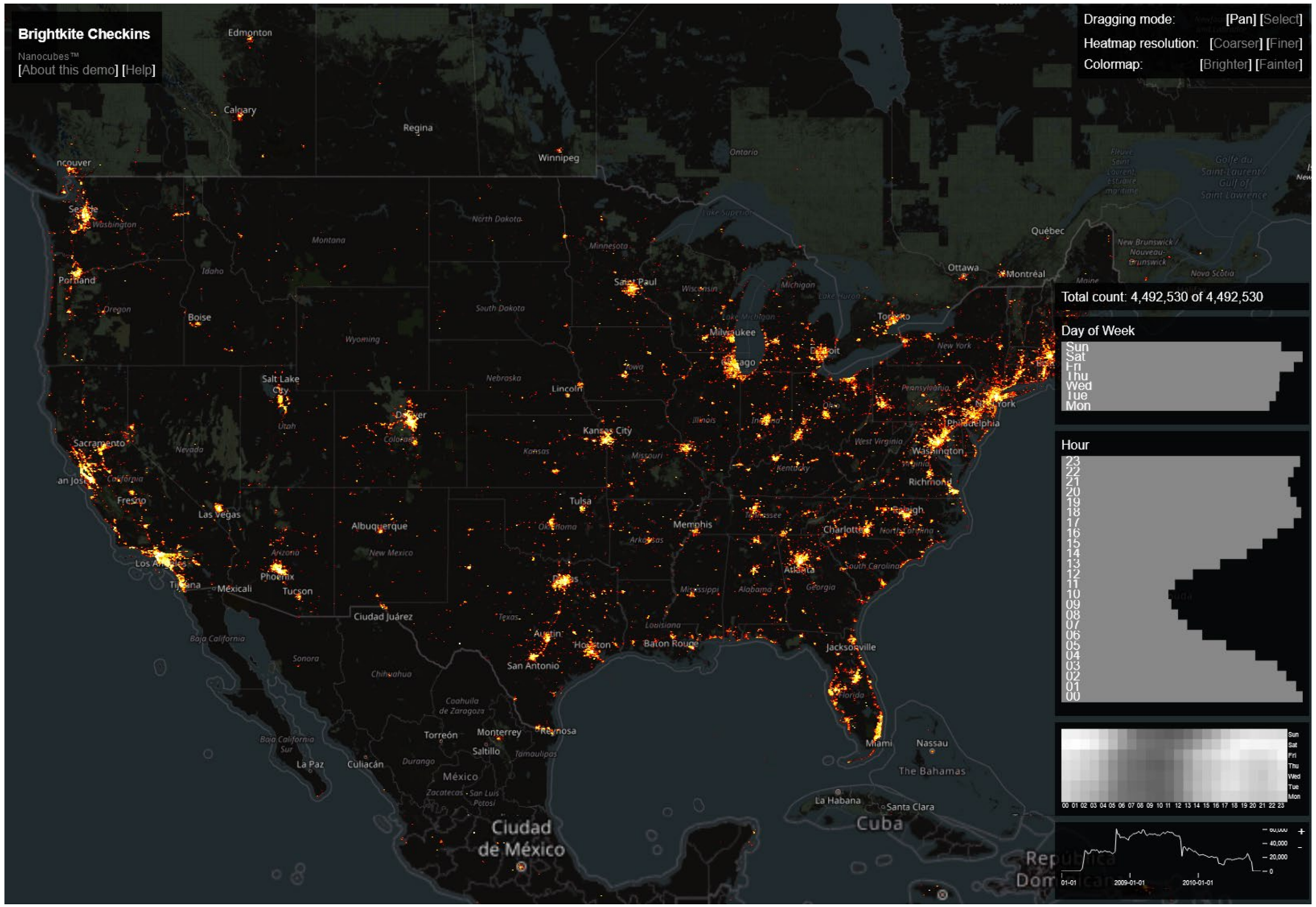
Data aggregation generates uncertainty

- Same stats, different graphs [Matejka and Fitzmaurice, 2017]





Interactive aggregations



[Lins, 2012]

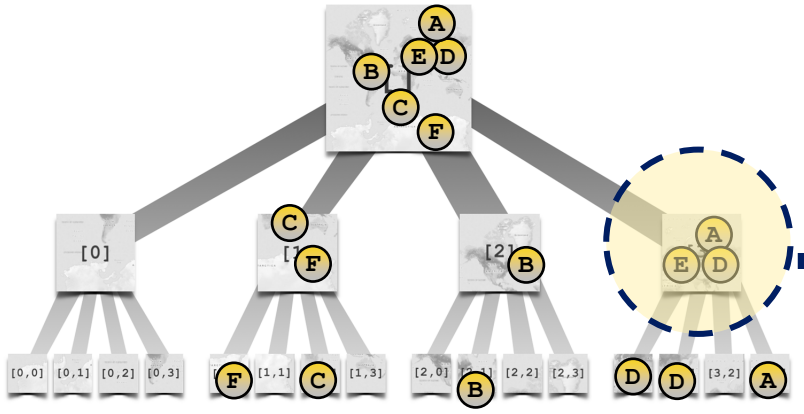
Datacube model

Following datacube model, aggregate every record along a hierarchy of bins.

The data structure is a mapping of bins to a pre-computed summary (e.g., count, timeseries).

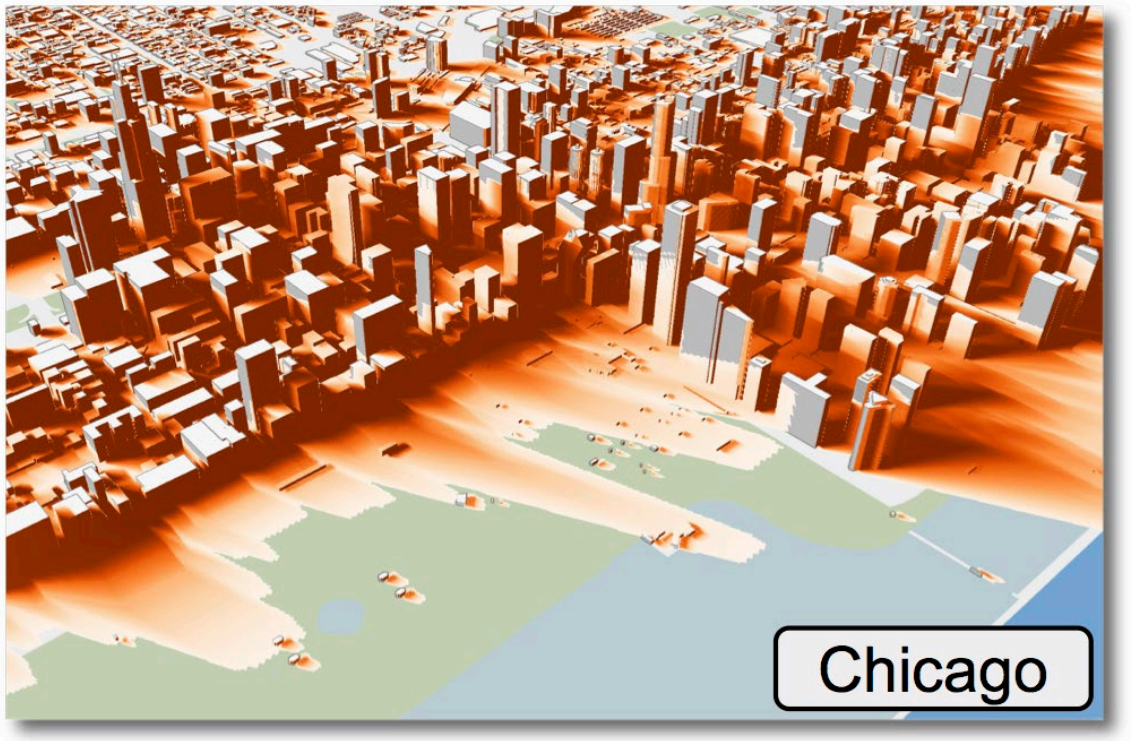
latitude	longitude	keyword
42.102908	-73.242852	#phoenix
29.617161	-81.636398	#phoenix
23.014051	75.120052	#la
26.014051	75.120052	#nyc
28.014051	74.120052	#la
23.014051	75.120052	#phoenix

	latitude	longitude
A	42.102908	-73.242852
B	29.617161	-81.636398
C	23.014051	75.120052
D	26.014051	75.120052
E	28.014051	74.120052
F	29.61161	-81.63638



K	c	p
0	10	1
1	22	2
2	15	0

Data aggregation



[Miranda et al., 2019]

Summary statistics

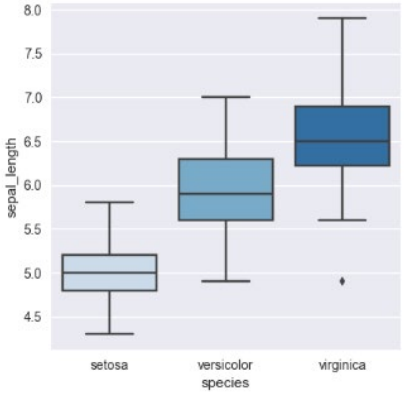
- Summarizes statistical information
- Measures of locality:
  - Mean
  - Median
  - Interquartile mean
- Measure of spread:
  - Standard deviation
  - Range
  - Variance
  - Absolute deviation
  - Interquartile range

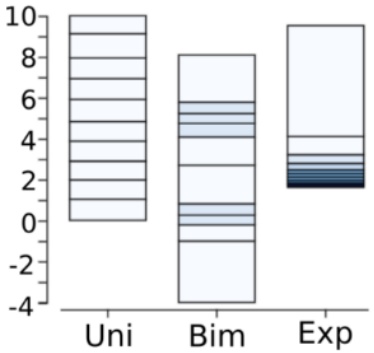
Uncertainty visualization

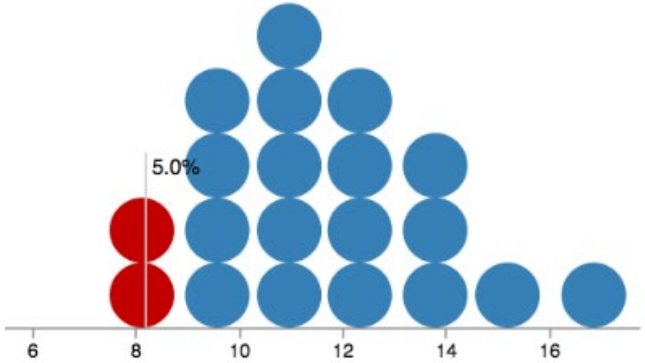
- When you have a whole range or distribution of numbers, avoid visualizing single numbers.
- Graphical annotations of distributional properties
  - Histograms and density plots
  - Intervals and ratios
  - Distributions
- Visual encodings of uncertainty
- Hybrid approaches

Possible solution: quantiles

- Robust statistics with clear interpretation







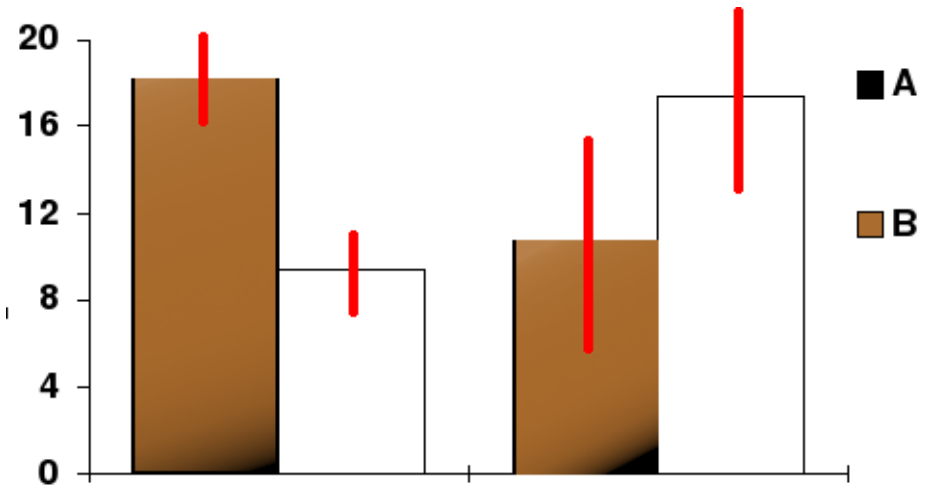
- Important: quantiles cannot be merged

$$Mean(A \cup B) = \frac{(\sum A + \sum B)}{(|A| + |B|)}$$

$Median(A \cup B)$ is not a function of $Median(A)$ and $Median(B)$

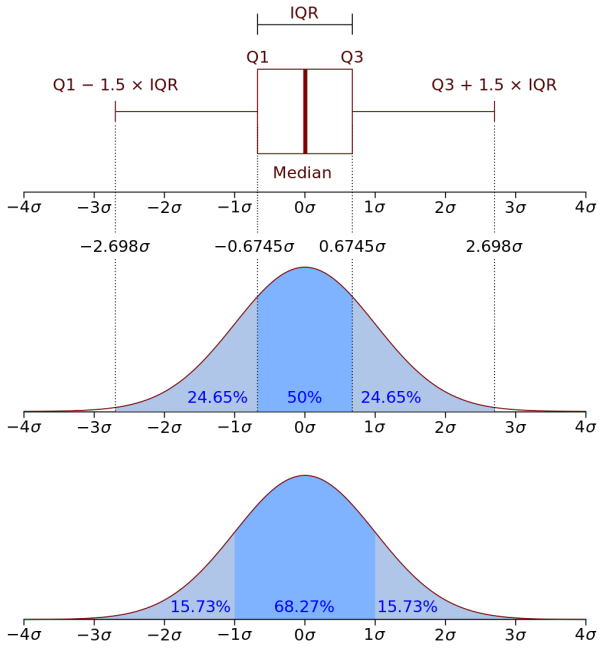
Intervals and ratios

Error bars: show variability of data, indicating error or uncertainty.



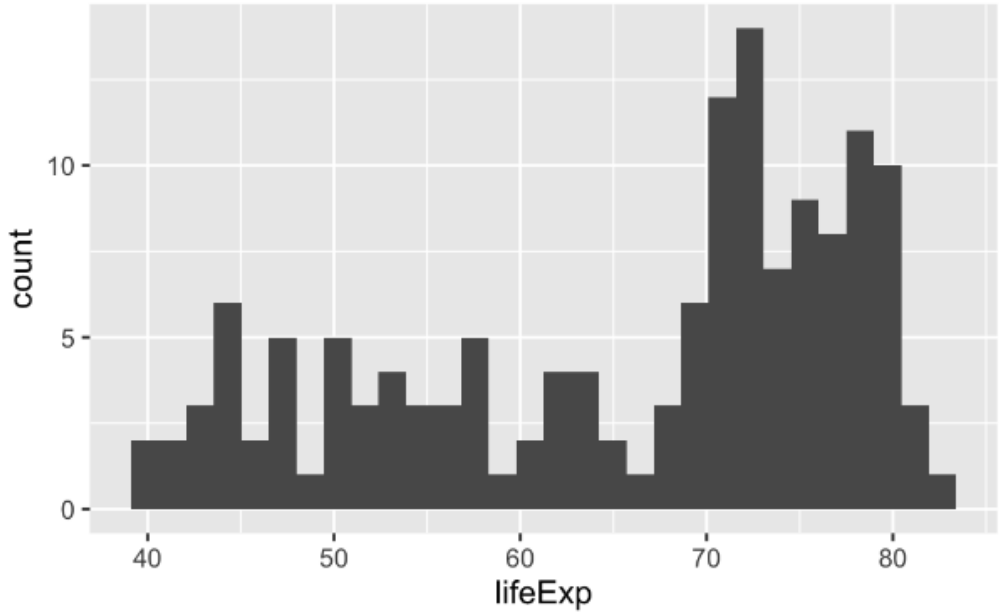
$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i + \mu)^2}$$

Error bars: show variability of data, indicating error or uncertainty.

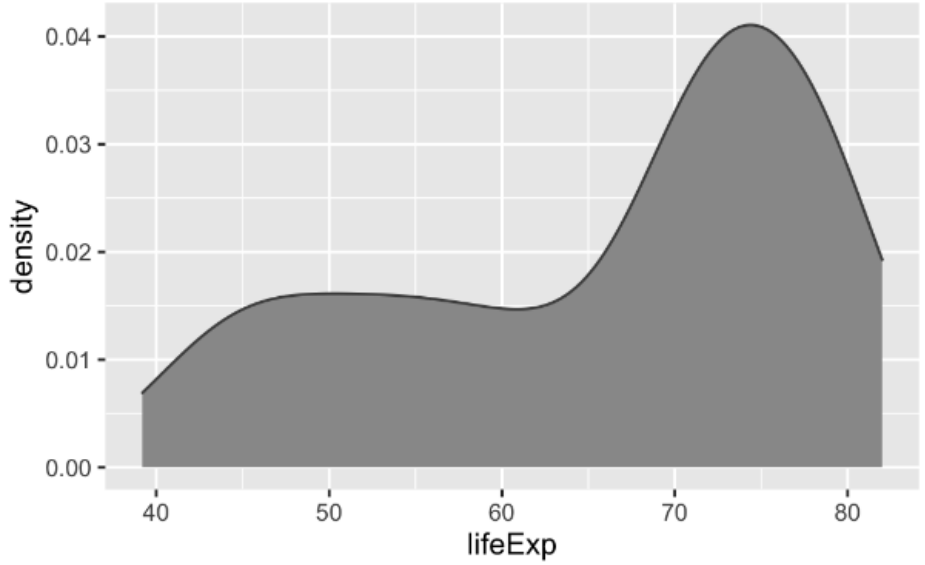


Distributions

Histogram: approximate representation of the distribution of the data.



Density plots: distribution of data over continuous interval.



Density estimation

- What is (probability) density estimation?
  - Given random variable $\mathbf{X}$, specify probability density as a function f.
  - Probability that a sample falls into an interval from a to b, calculate area under the graph of the density function:

$$P(a < X < b) = \int_a^b f(x)dx$$

  - Density estimation: estimate the unknown probability density function $\hat{f}$ from observed data points $x_1, \ldots, x_n$.

Histogram

$$\mathbf{X} = (x_1, \ldots, x_n)$$

Estimator:

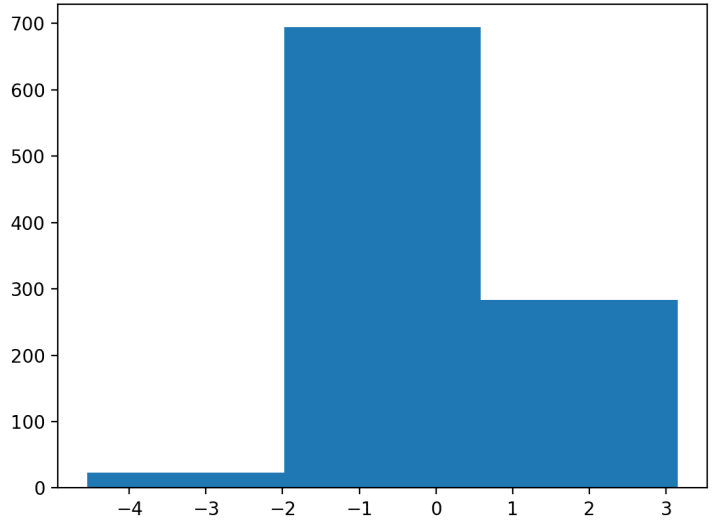
$$\hat{f}(x) = \frac{1}{n} * \frac{[\#\ observations\ in\ same\ bin\ as\ x]}{2h}$$

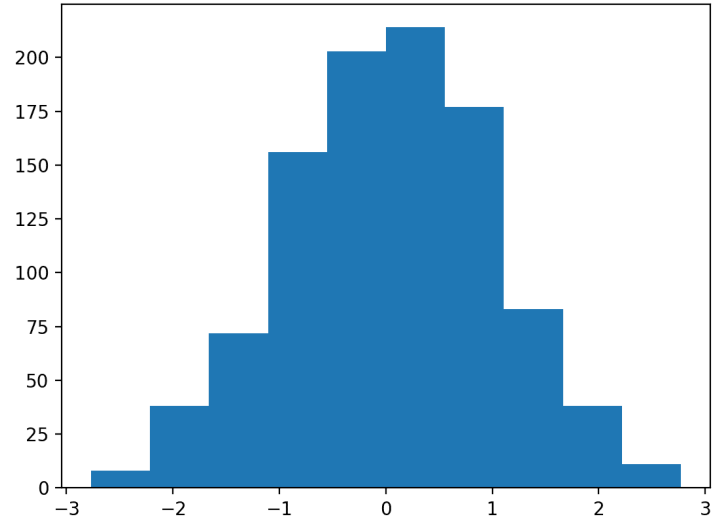
Intervals defined $[x_0 + mh, x_0 + \frac{m+1}{h}]$, origin x_0, bin width h.

Bins are not centered on data samples.

Histogram

- Fast and reliable way to visualize probability density.
- Impacted by the number of bins, i.e., depends on the width of bins.



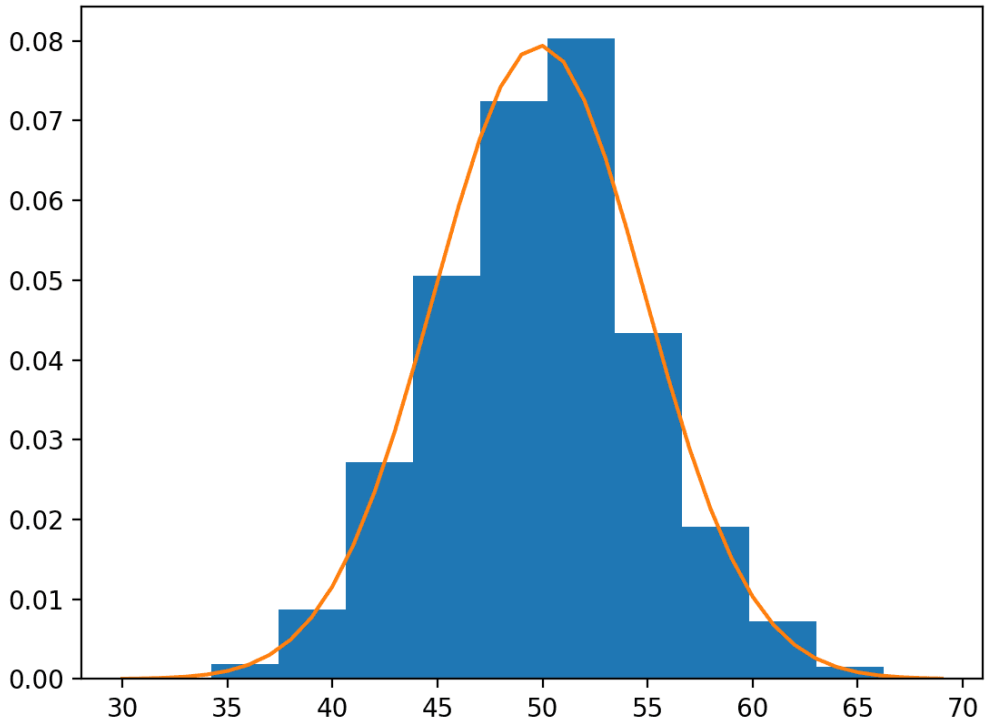


Density estimation

- Parametric density estimation: assumes that the data is from a known parametric family of distributions (e.g., normal distribution).
  - Estimate parameters from data.
- Nonparametric density estimation: less rigid assumptions about the underlying density of observed data. Data speaks for itself, no assumption that density of f comes from known parametric family.

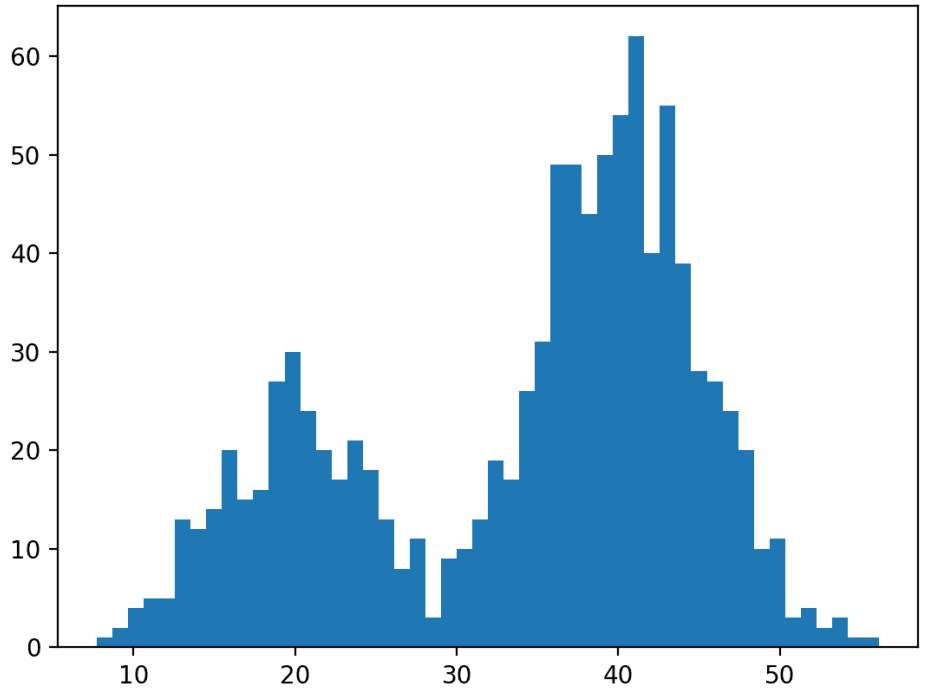
Parametric density estimation

- Estimate the density of a random variable with a known probability distribution.
- Estimate parameters from the data; for example, estimate mean and std. deviation for normal distribution.
- Summarizing relationship between observations and probabilities *through parametrization*.



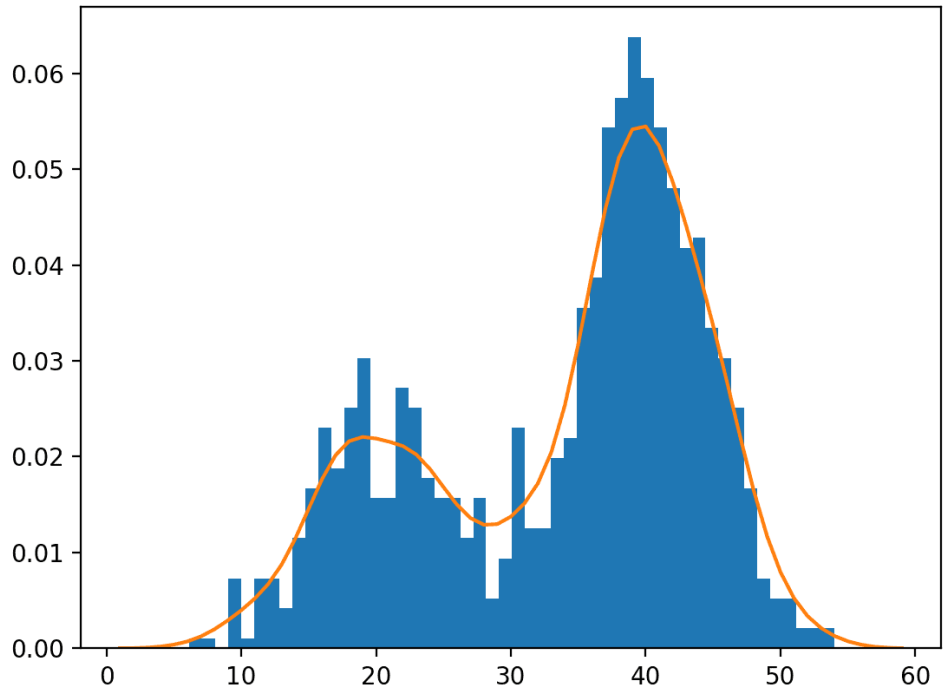
Parametric density estimation

What if data does not match common probability distribution?



Nonparametric density estimation

- Estimate the density of the random variable with no known probability distribution.
  - Two or more peaks.
- Approximate probability distribution without a pre-defined distribution.



Naïve estimator

$$\hat{f}(x) = \frac{1}{n} * \frac{[\#\ of\ x_1, \ldots, x_n\ falling\ in\ (x-h, x+h)]}{2h}$$

$$\hat{f}(x) = \frac{1}{nh} * \sum_{i=1}^{n} w\left(\frac{x - x_i}{h}\right)$$

Weight function $w(x) = \begin{cases} \frac{1}{2}, if\ |x| < 1 \\ 0, otherwise \end{cases}$

Bins are centered on every sample (in contrast to previous histogram): avoid having to choose locations of bins.

Kernel density estimation

- Nonparametric method to smooth probabilities across the range of outcomes for a random variable.
- Kernel: mathematical functions that returns probability for a given value of a random variable.

$$\hat{f}(x) = \frac{1}{nh} * \sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right)$$

- Kernel effectively smooths probabilities across the range of outcomes, such that sum equals one.

Kernel density estimation

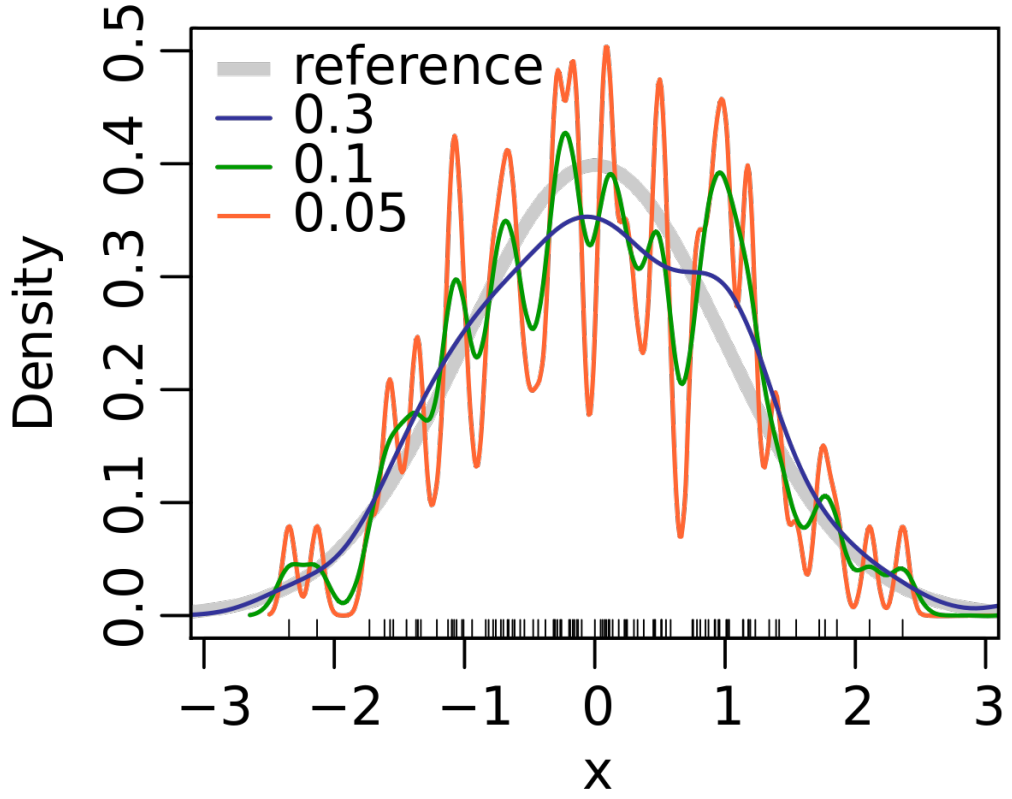
$$\hat{f}(x) = \frac{1}{nh} * \sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right)$$

Instead of a boxcar function (naïve estimator), can be a Gaussian function:

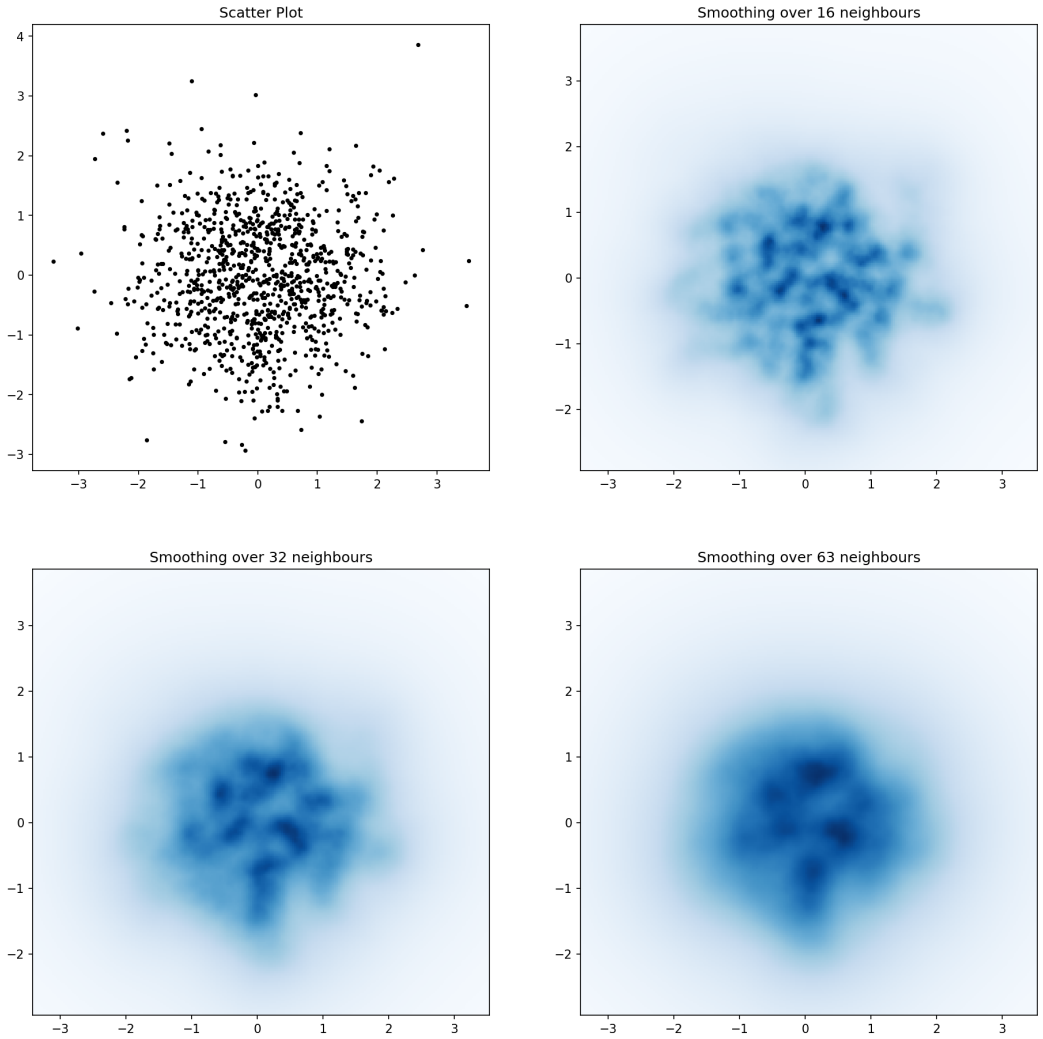
$$K(x) = \frac{1}{\sqrt{2\pi}} e^{\frac{-x^2}{2}}$$

Kernel density estimation

- Same smoothing parameter (bandwidth) for the whole domain.
  - Large bandwidth: coarse density with little details.
  - Small bandwidth: too much detail and not general enough to cover new or unseen examples.

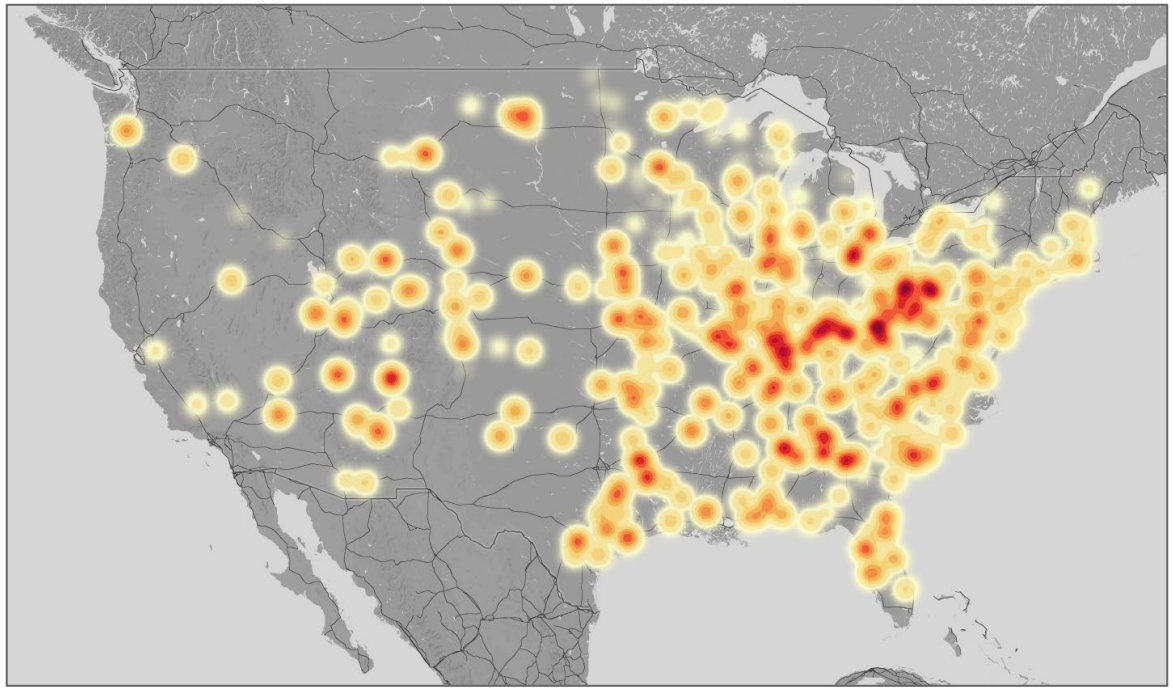


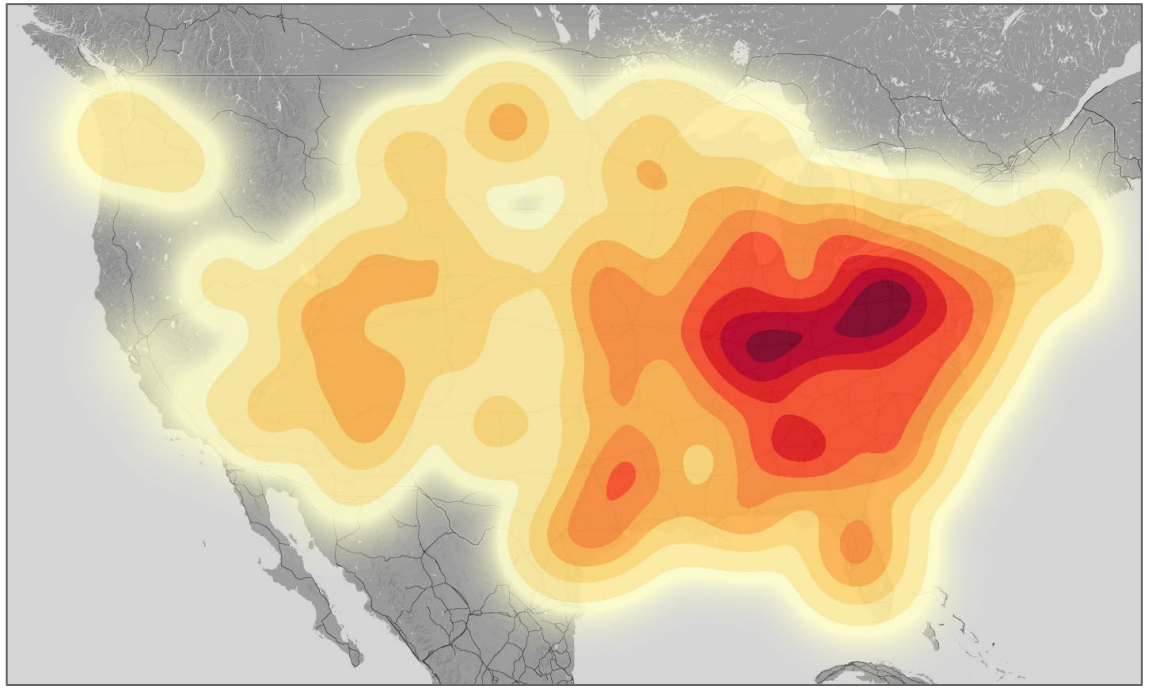
Kernel density estimation



[https://stackoverflow.com/questions/2369492/generate-a-heatmap-in-matplotlib-using-a-scatter-data-set]

Kernel density estimation





Adaptive kernel density estimation

- Different bandwidths for different x_i.

$$\hat{f}_a(x) = \frac{1}{n} * \sum_{i=1}^{n} \frac{w_i}{h_i} K\left(\frac{x - x_i}{h_i}\right)$$

$$h_i = h * \lambda_i$$

$$\lambda_i = \sqrt{G / f(x_i)}$$

$$G = \left(\prod_{i=1}^{n} \hat{f}(x_i)\right)^{\frac{1}{n}}$$

Kernel density estimation

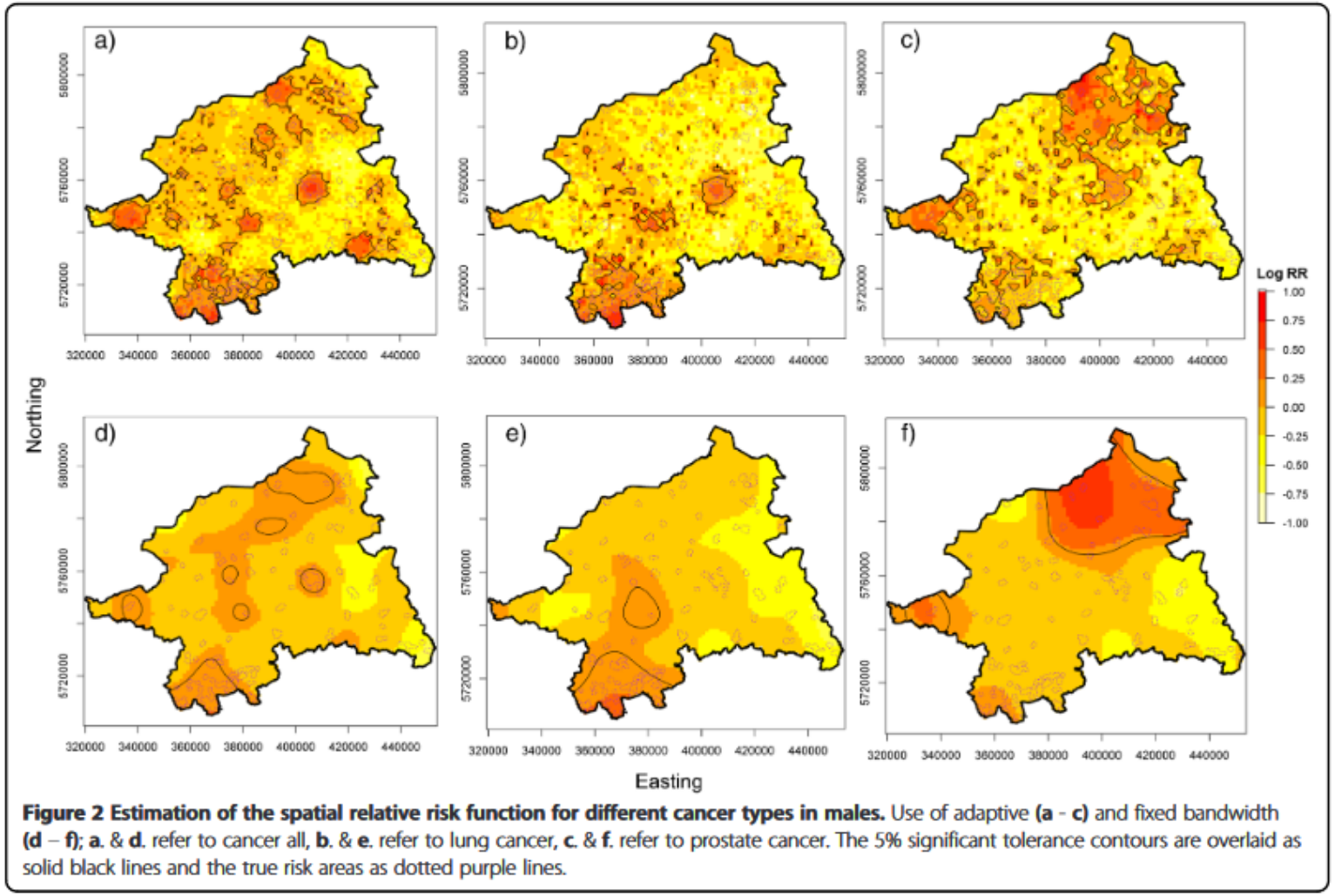
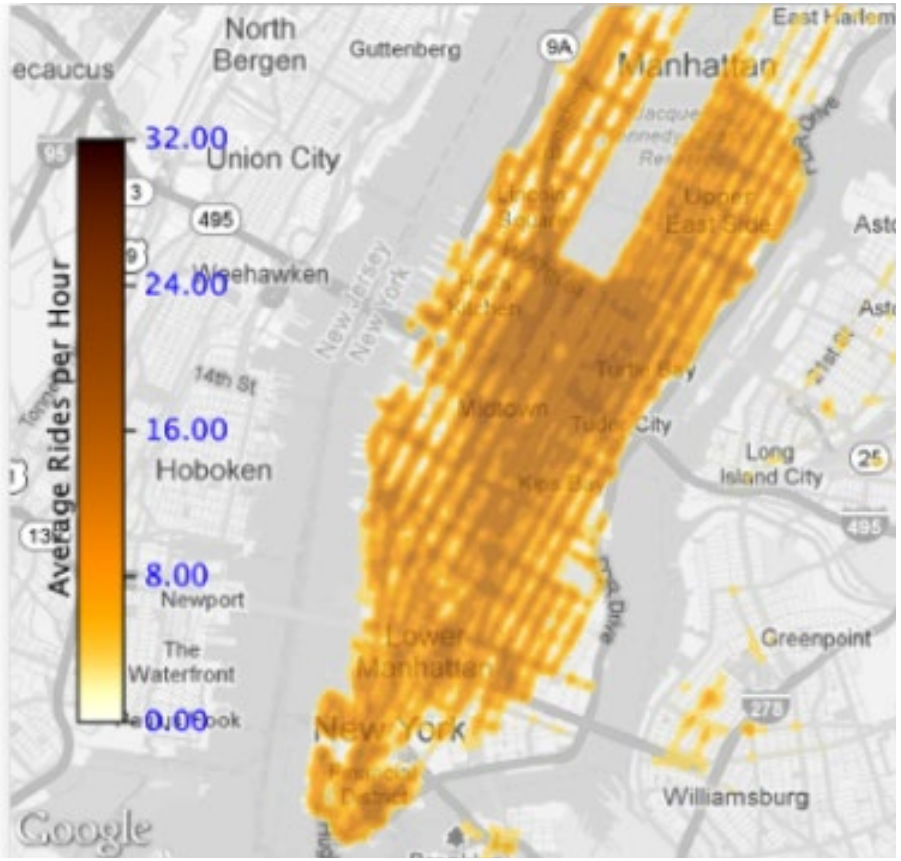


Figure 2 Estimation of the spatial relative risk function for different cancer types in males. Use of adaptive **(a - c)** and fixed bandwidth **(d – f)**; **a.** & **d.** refer to cancer all, **b.** & **e.** refer to lung cancer, **c.** & **f.** refer to prostate cancer. The 5% significant tolerance contours are overlaid as solid black lines and the true risk areas as dotted purple lines.

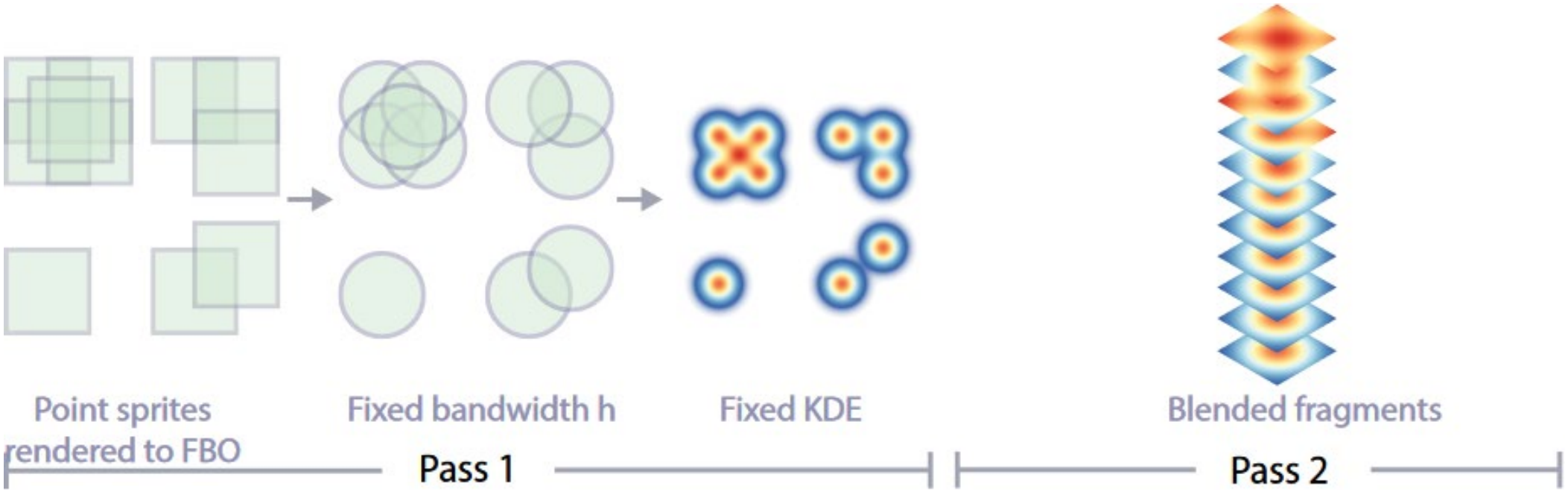
[Comparing adaptive and fixed bandwidth-based kernel density estimates in spatial cancer epidemiology]

Kernel density estimation

- TaxiVis:
  - More than 70 seconds for 100 million data points.
- How to speed up the processing if we have billions of points?
  - Each point depends only on its neighbors.



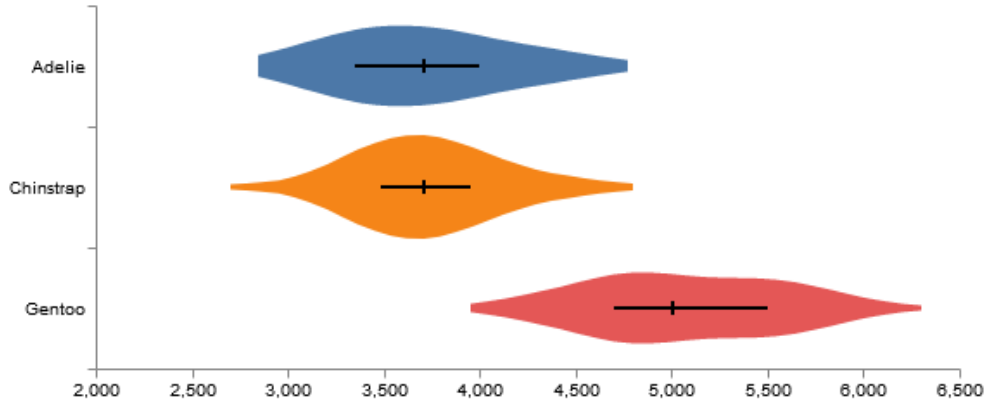
Kernel density estimation



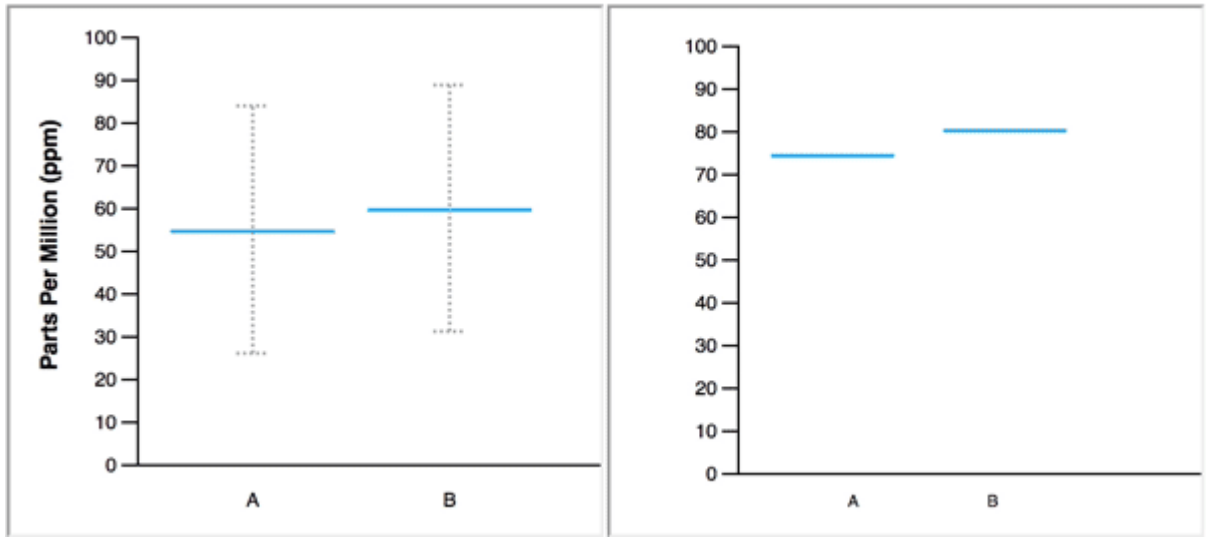
GPU pipeline

Distributions

Violin plot: similar to box plots, but show probability density at different values.



Hypothetical outcome plots: visualize a set of draws from a distribution.



Item-based vs density-based

Item-based visualization

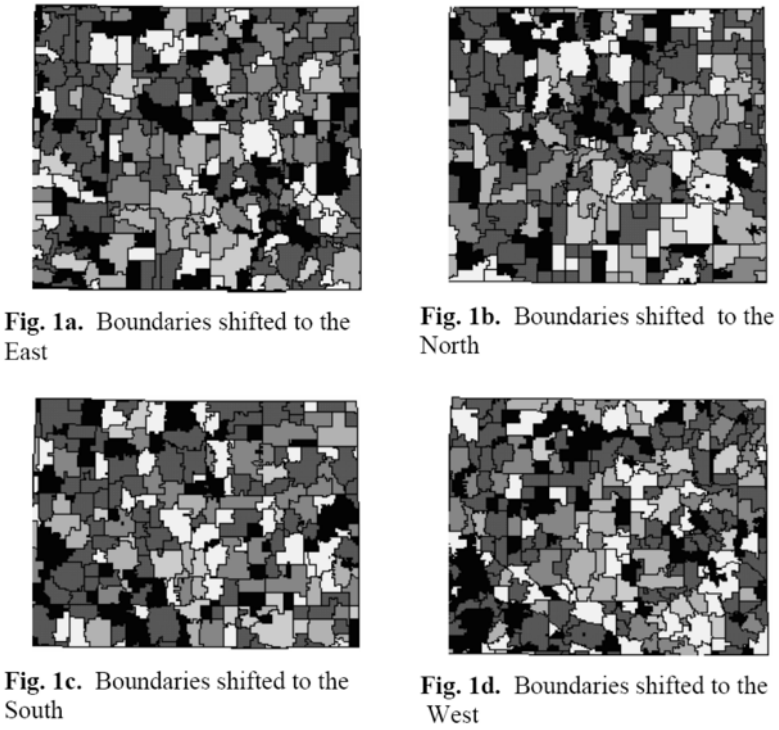


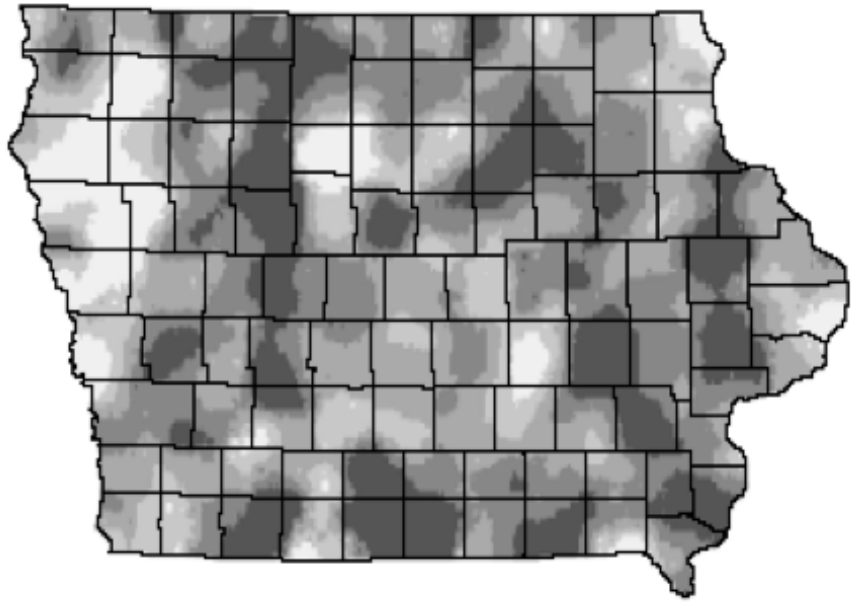
Fig. 1a. Boundaries shifted to the East

Fig. 1b. Boundaries shifted to the North

Fig. 1c. Boundaries shifted to the South

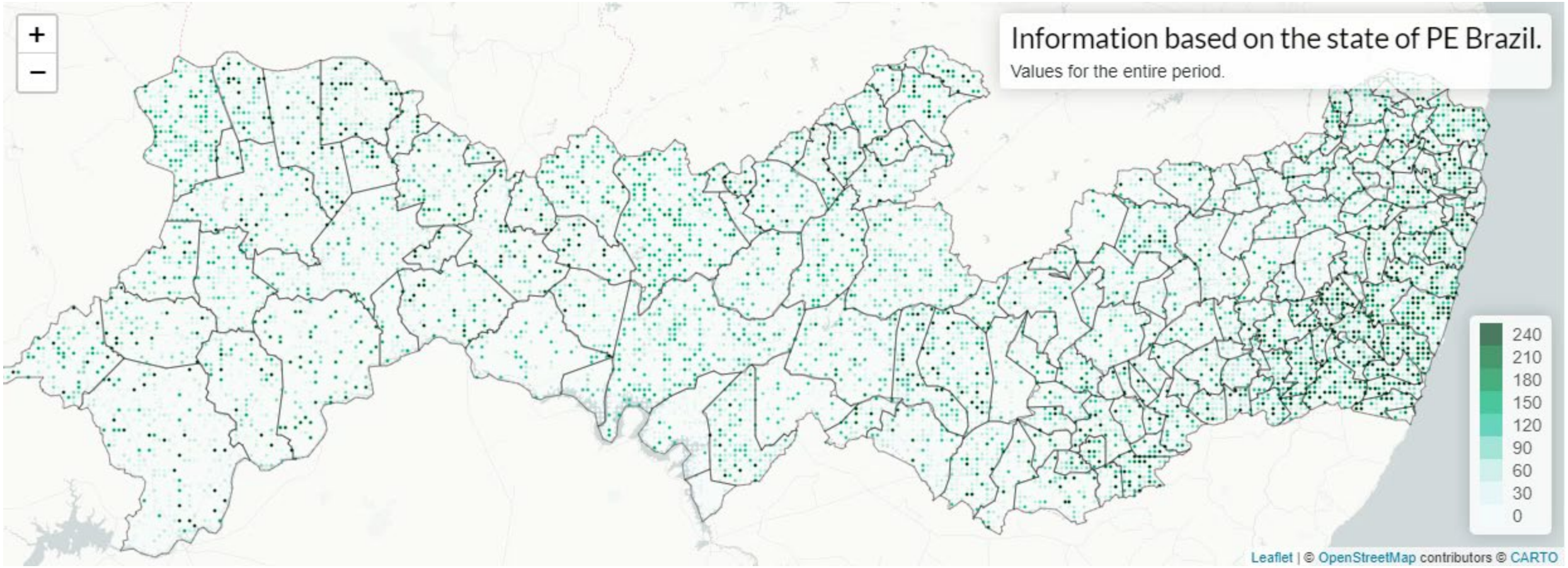
Fig. 1d. Boundaries shifted to the West

Density-based visualization

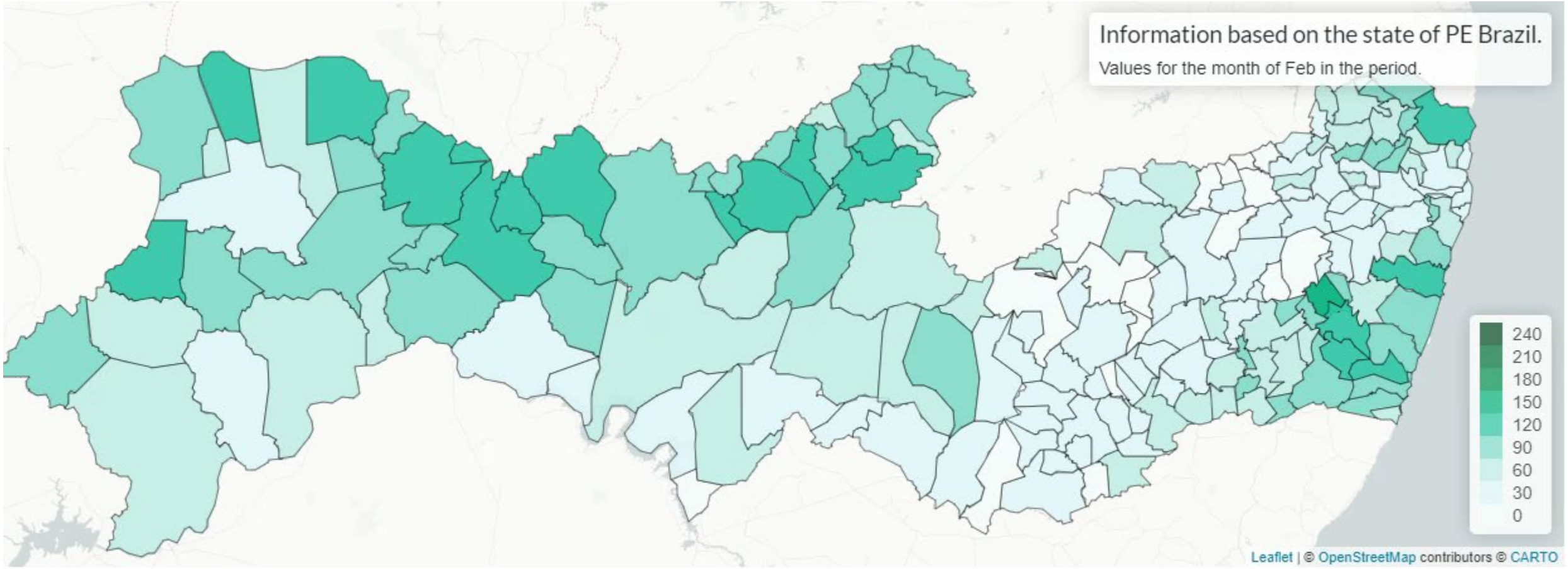


[Using Spatially Adaptive Filters to Map Late Stage Colorectal Cancer Incidence in Iowa]

Quantile dot plots

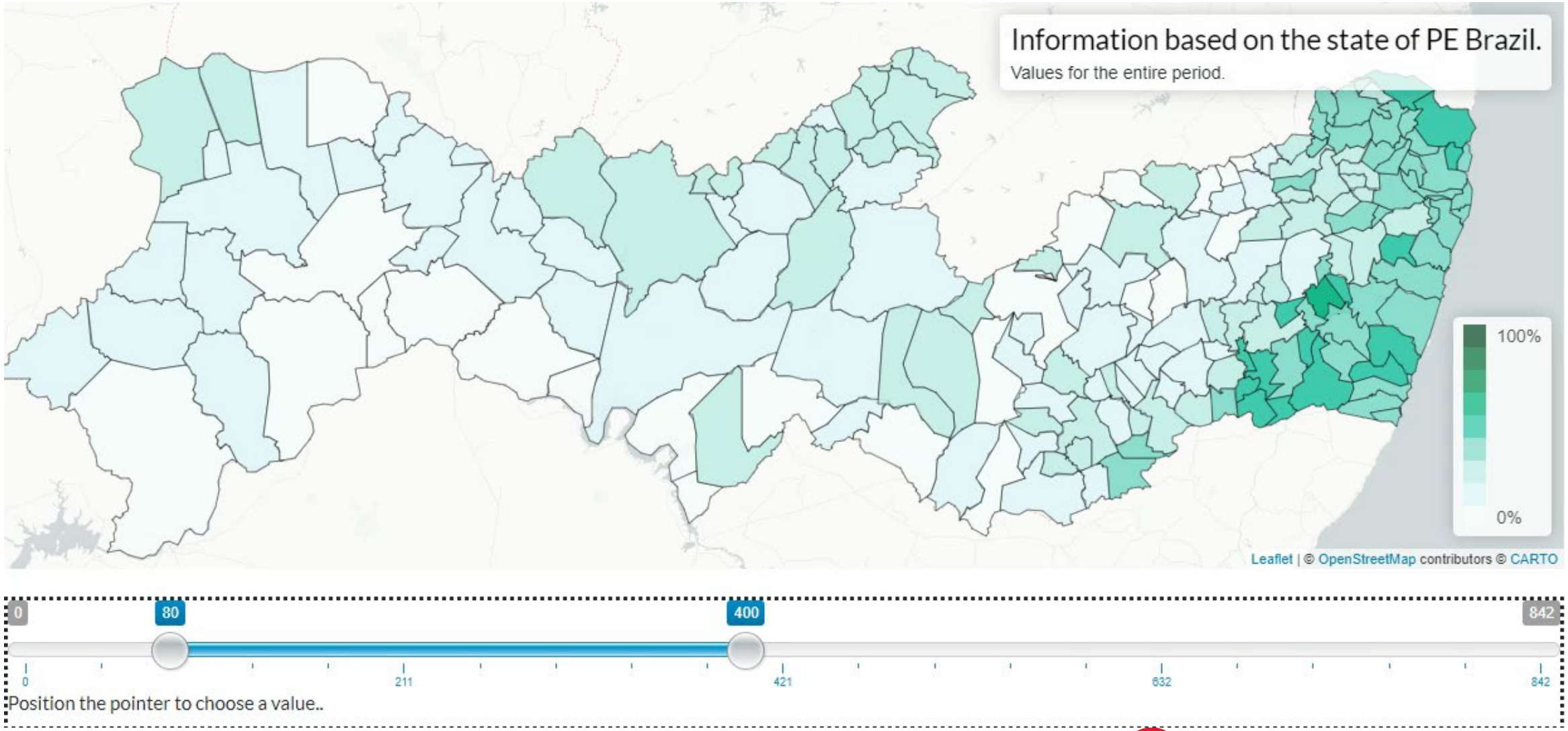


Hypothetical outcome plots

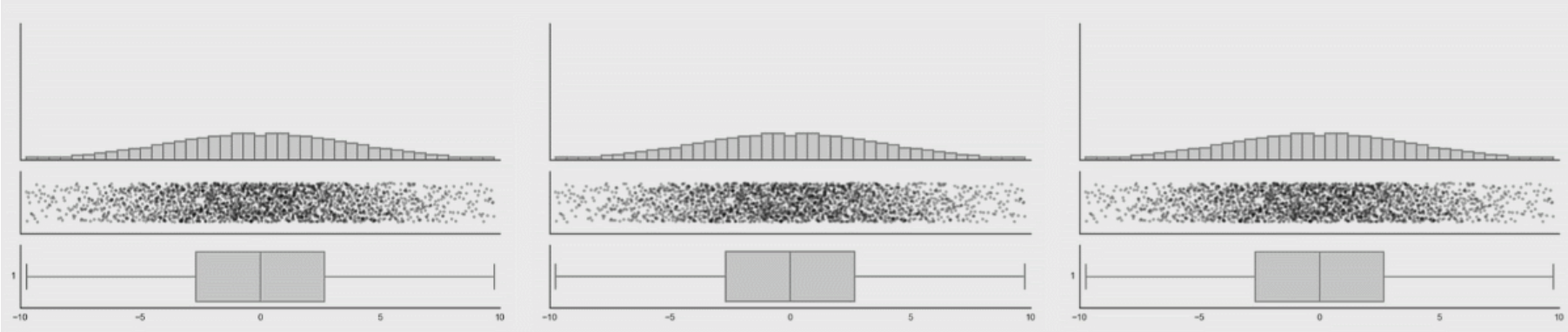


Animate Stop

Interaction maps



Data aggregation generates uncertainty

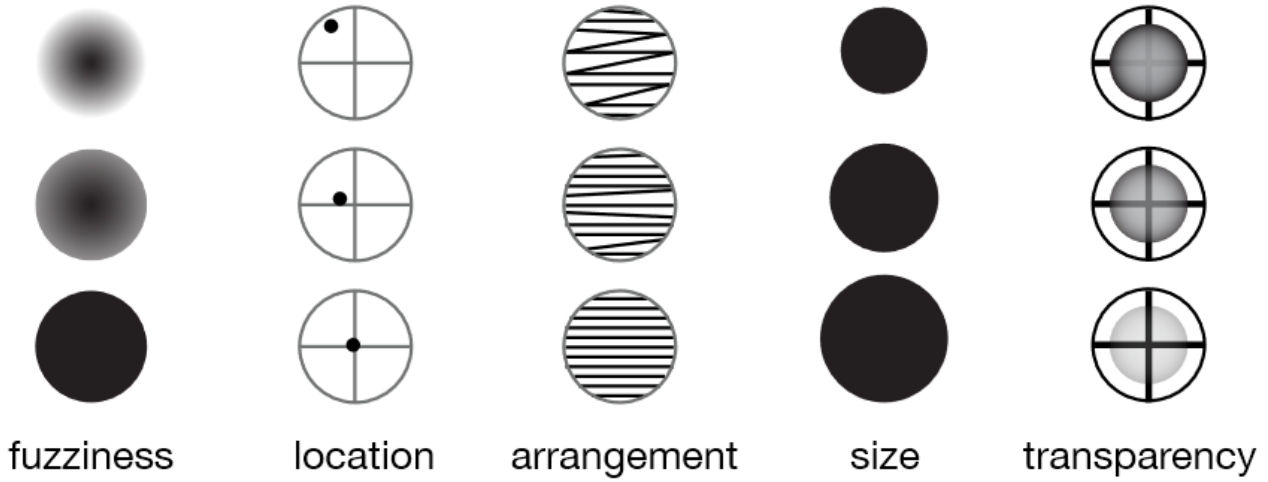


Distributions

- Histogram and density plots
- Violin plot
- Gradient plot
- Hypothetical outcome plot
- Quantile dot plot
- Ensemble plot

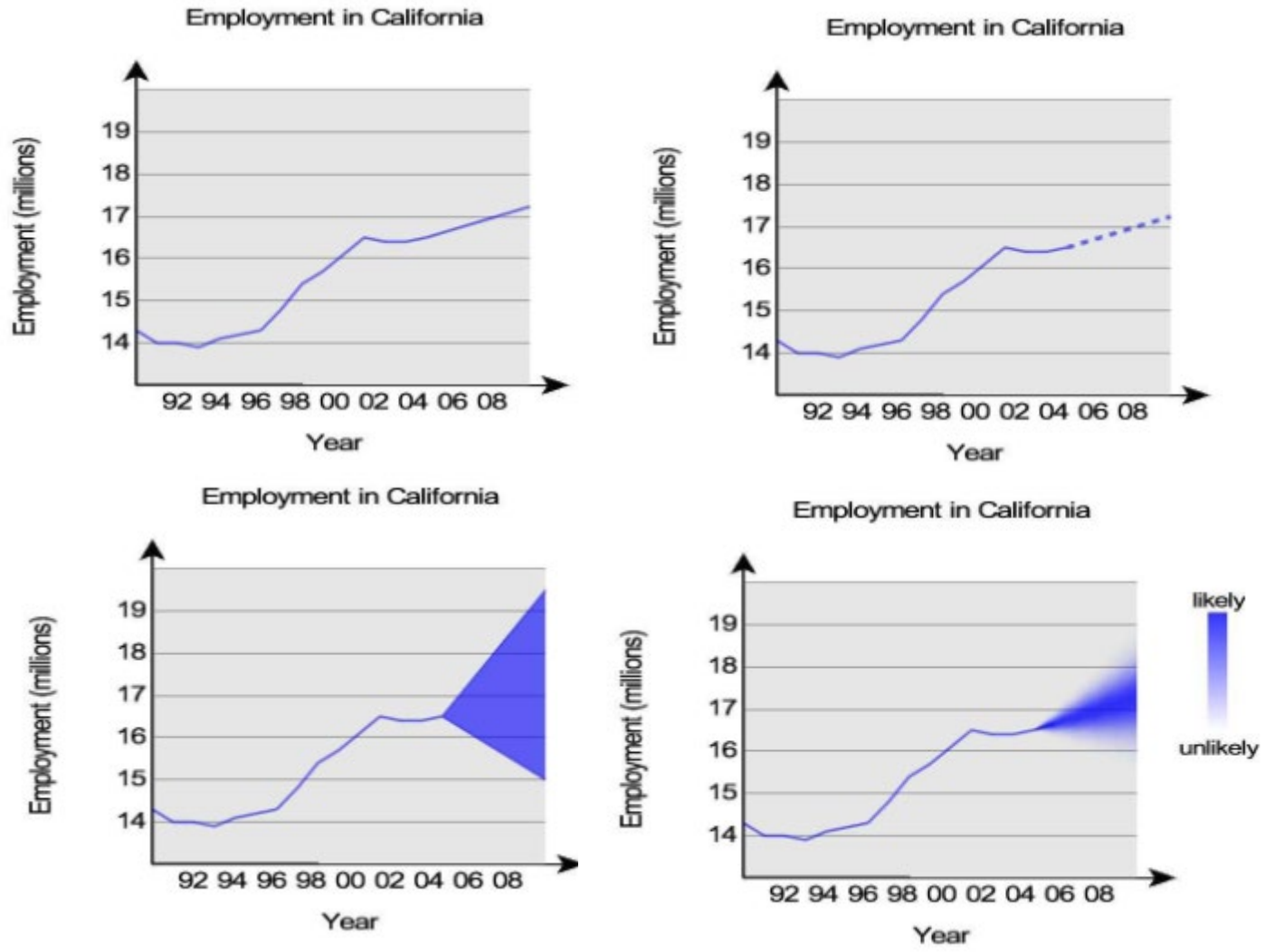
Visual encodings of uncertainty

- Additional visual channels can be used to characterize the uncertainty: hue, texture, opacity
  - Fuzziness
  - Location
  - Arrangement
  - Size transparency



[Padilla et al., 2020]

Visual encodings of uncertainty



[Streit et al., 2007]

Hybrid approaches

Contour boxplot: