Visualizing spatial data

CS424: Visualization & Visual Analytics

Fabio Miranda

https://fmiranda.me

Slides based on Enrico Bertini's vis course

Spatial data

flickr

twitter

Infrastructure | Environment | Social media

Spatial data

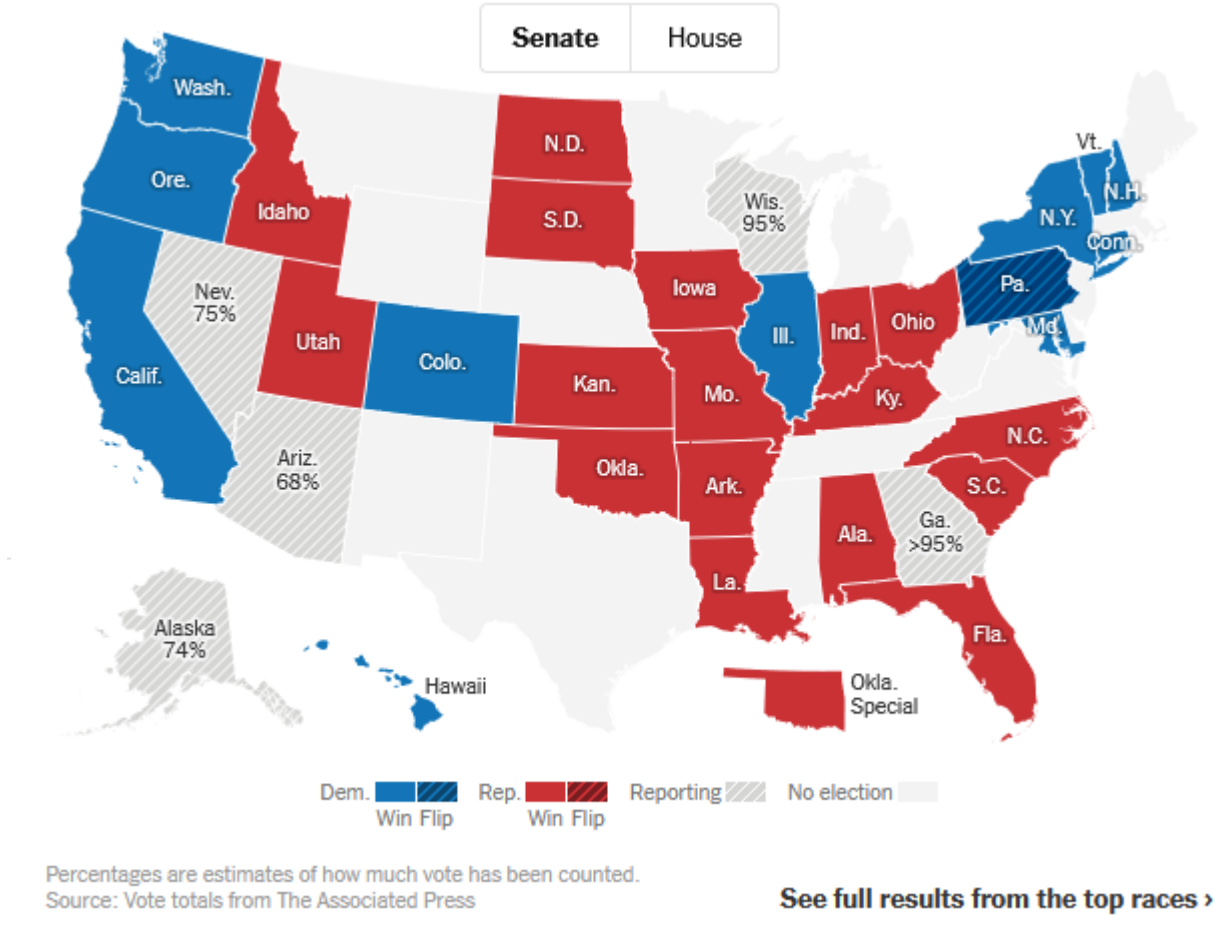
- Spatial attributes
  - 2D: (x, y)
  - 3D: (x, y, z)
- Spatial data primitives:
  - Points
  - Lines
  - Polygons
- Other attributes:
  - Time
  - …

Geographical data

- Any data containing geographical information.
- Spatial object: counties, regions, buildings, lakes.
- Geolocated objects: cars, people, animals, weather stations.

Spatial objects

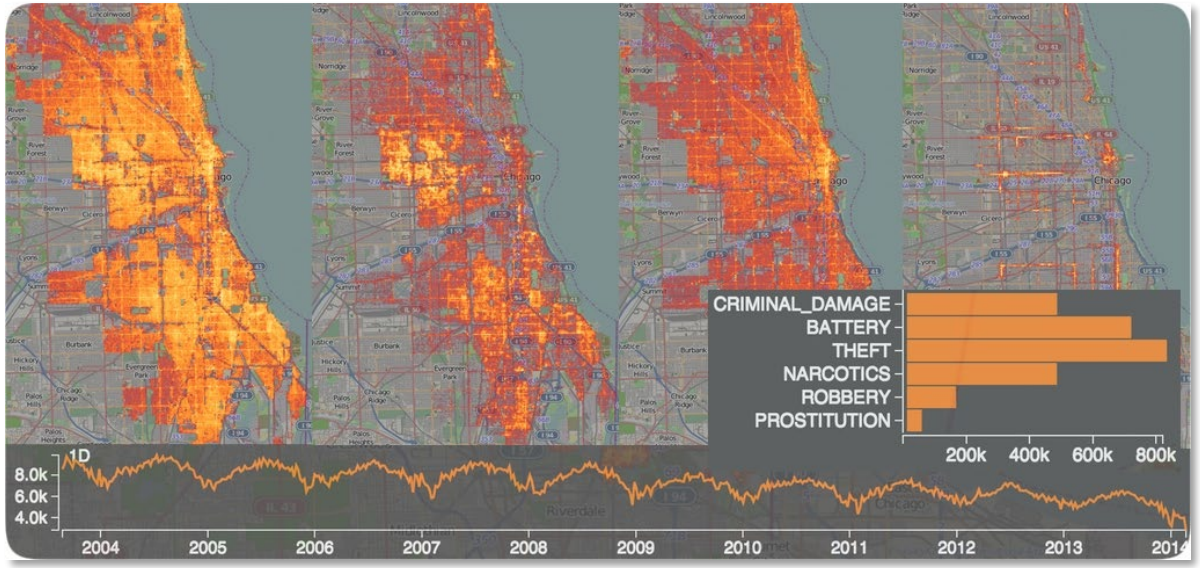
- Shape and spatial extent matters.
- Map is the main object of interest.




Search for a Bushwick address
Property Layers
All Tax Lots
Available FAR
Likely Rent Stabilized
Land Use
Vacant Land
Year Built
DOB Permits
New Buildings
A1: Major Alteration
A2, A3: Minor Alteration
Stories
Sites of Gentrification
Land Use
Multi-Family Walkup
1 & 2 Family Bldgs
Mixed Resid & Comm
Parking Facilities
Vacant Land
Commerical & Office
Industrial & Mfg
Public Facil & Instns
Open Space & Rec
N/A
Source: Map Pluto
Knollwood Park Cemetery
Leaflet | © Mapbox © OpenStreetMap, CartoDB

Geolocated objects

- Shape does not matter.
- Map used as a reference.



Types of geographical data

- Regions (e.g., county borders).
- Locations (e.g., latitude, longitude).
- Identifiers (e.g., zip code, street name, borough).

Geocoding

Geocoding

Identifier → Location

Reverse geocoding

Location → Identifier

Geocoding

```
>>> from geopy.geocoders import Nominatim
>>> geolocator = Nominatim(user_agent="specify_your_app_name_here")
>>> location = geolocator.geocode("175 5th Avenue NYC")
>>> print(location.address)
Flatiron Building, 175, 5th Avenue, Flatiron, New York, NYC, New York, ...
>>> print((location.latitude, location.longitude))
(40.7410861, -73.9896297241625)
>>> print(location.raw)
{'place_id': '9167009604', 'type': 'attraction', ...}
```

Reverse geocoding

```
>>> from geopy.geocoders import Nominatim
>>> geolocator = Nominatim(user_agent="specify_your_app_name_here")
>>> location = geolocator.reverse("52.509669, 13.376294")
>>> print(location.address)
Potsdamer Platz, Mitte, Berlin, 10117, Deutschland, European Union
>>> print((location.latitude, location.longitude))
(52.5094982, 13.3765983)
>>> print(location.raw)
{'place_id': '654513', 'osm_type': 'node', ...}
```

When & why to use a map?

- Not all geographical data should be represented using a map metaphor.
- Maps use position & space to depict geographical objects – space cannot be used to encode other types of information.

When & why to use a map?

1. When the question you want to answer is inherently spatial.
2. When the map helps find information needed.

Questions: When & why to use a map?

1. Every time a data set containing geographical information a map should be used.
2. Maps should be used when the goal is to present a spatial phenomenon.
3. Maps should be used when it is important for the reader to locate regions they are familiar with.

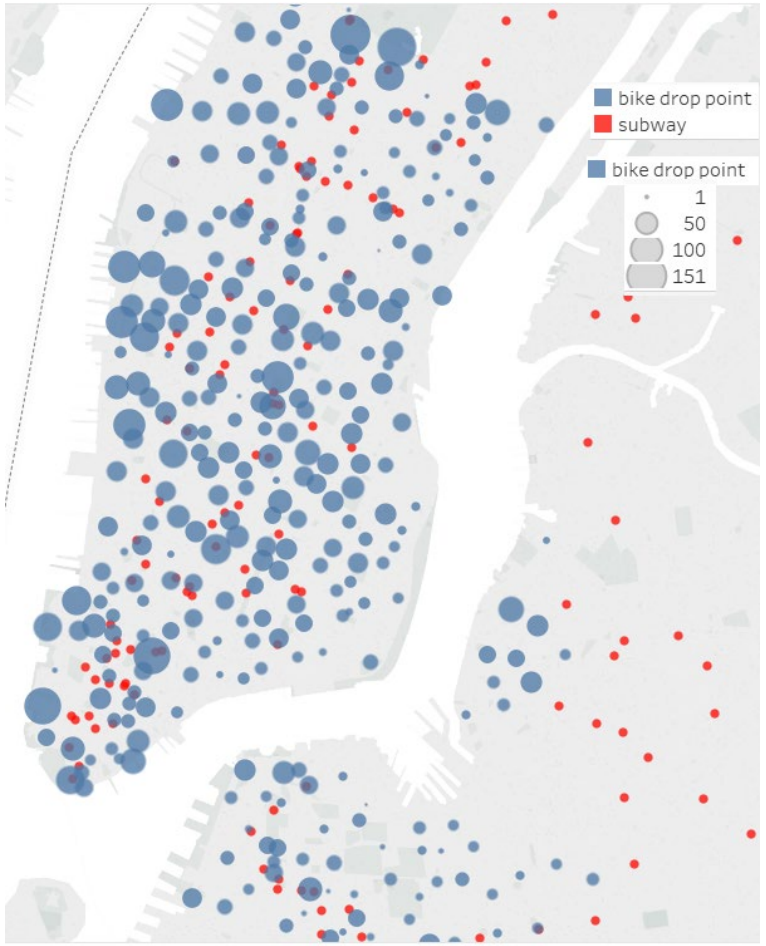
When & why to use a map?

1. Every time a data set containing geographical information a map should be used. **False**
2. Maps should be used when the goal is to present a spatial phenomenon. **True**
3. Maps should be used when it is important for the reader to locate regions they are familiar with. **True**

Examples of spatial questions

1. Questions that correlate a phenomenon to spatial locations / objects.

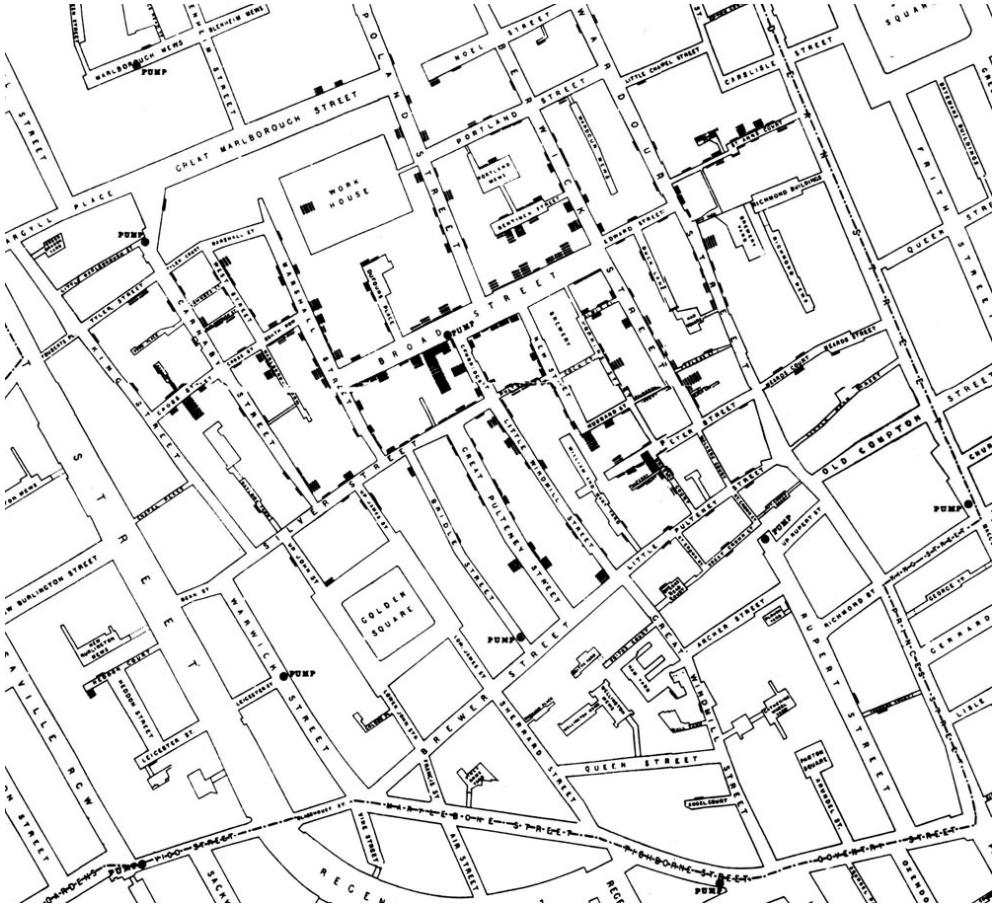
"Do city bike commuters tend to drop their bikes in proximity of subways stations?"

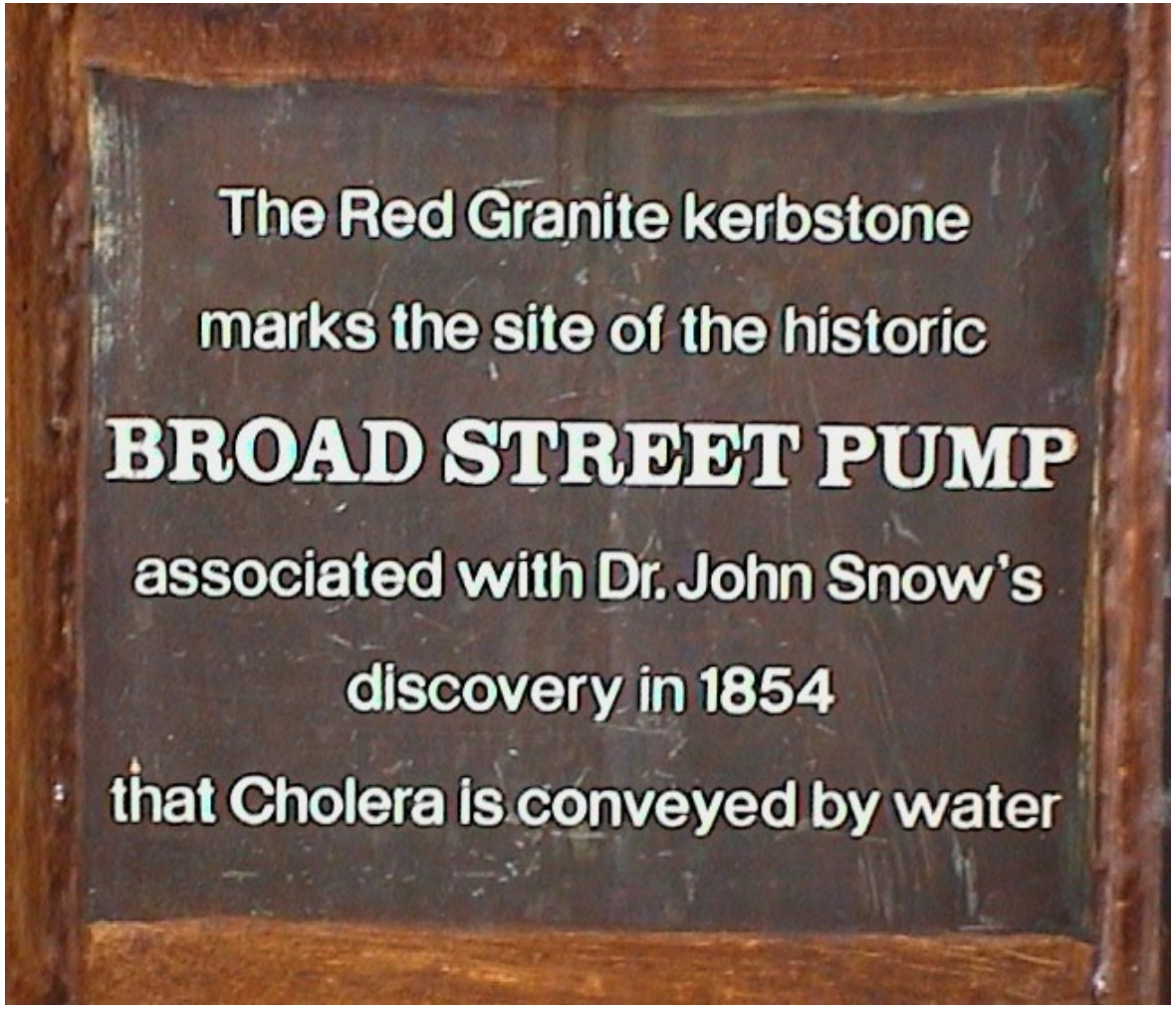


Examples of spatial questions

1. Questions that correlate a phenomenon to spatial locations / objects.

> “Do deaths caused by cholera cluster around water pumps?”



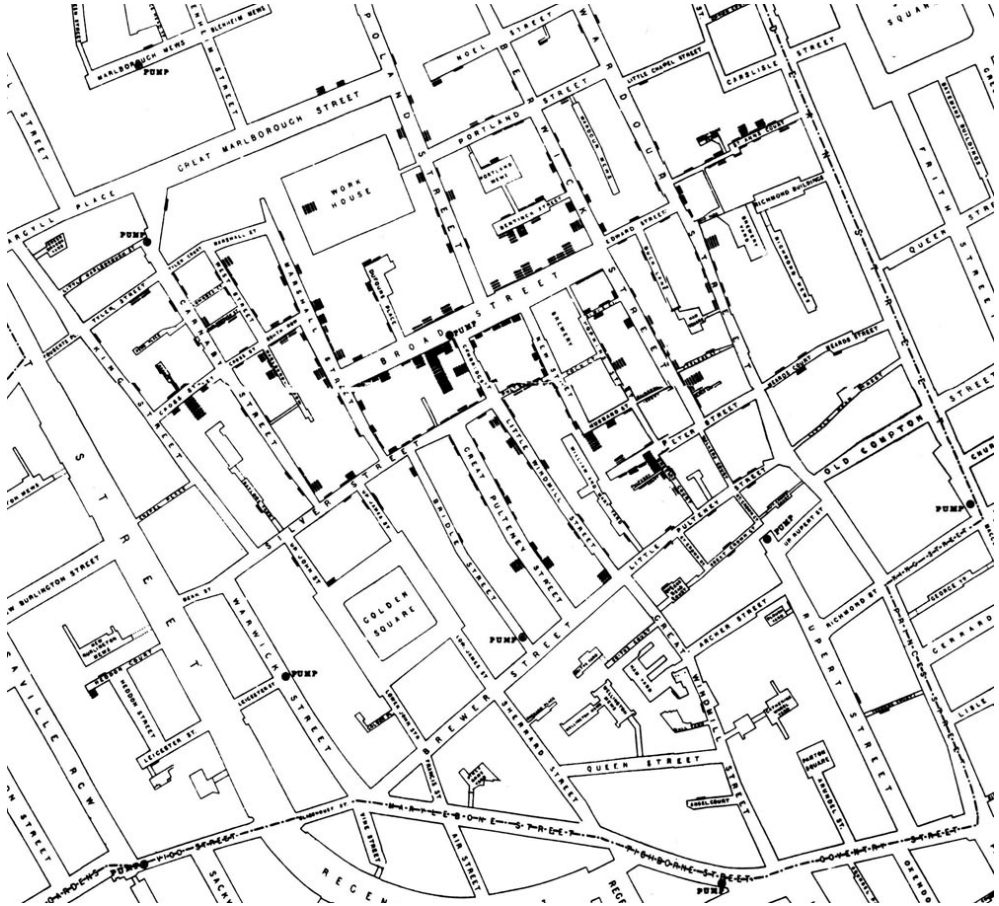

The Red Granite kerbstone
marks the site of the historic
BROAD STREET PUMP
associated with Dr. John Snow's
discovery in 1854
that Cholera is conveyed by water

Examples of spatial questions

2. Questions where phenomena for which spatial proximity or extent is relevant.

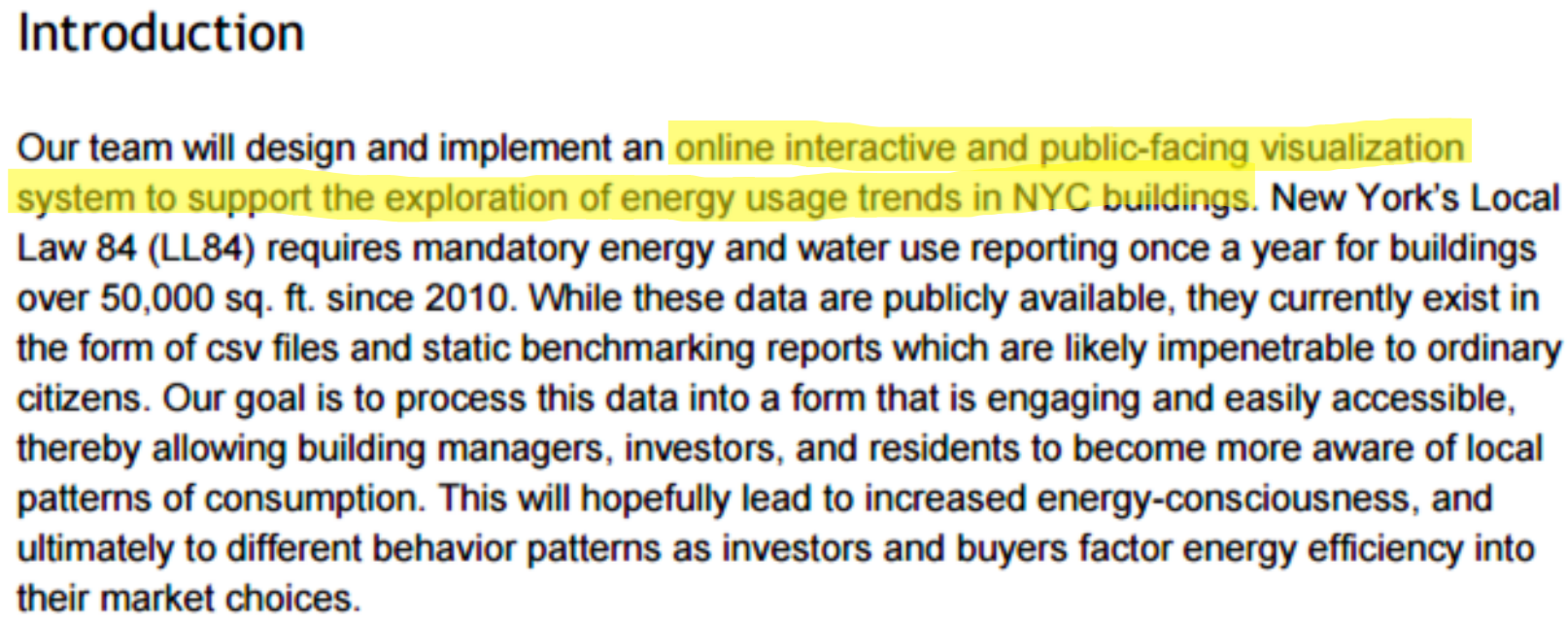
“How far has the epidemic extended from its point of origin?”

“Do neighboring regions have the same rate of people infected?”



Examples of non-spatial questions

- Maps are not always the best or only solution for geographical information.



Introduction

Our team will design and implement an online interactive and public-facing visualization system to support the exploration of energy usage trends in NYC buildings. New York's Local Law 84 (LL84) requires mandatory energy and water use reporting once a year for buildings over 50,000 sq. ft. since 2010. While these data are publicly available, they currently exist in the form of csv files and static benchmarking reports which are likely impenetrable to ordinary citizens. Our goal is to process this data into a form that is engaging and easily accessible, thereby allowing building managers, investors, and residents to become more aware of local patterns of consumption. This will hopefully lead to increased energy-consciousness, and ultimately to different behavior patterns as investors and buyers factor energy efficiency into their market choices.

Tasks

Question 1: Benchmark a Building

How does the EUI of one specific building compare to peer buildings in NYC?

Specific use-case: A building manager wants to learn about the energy usage in a building he is responsible for to detect potential for energy savings.
How does the normalized EUI of my office building (built in the 1970s) compare to peer office buildings in NYC?

Tasks

Question 1: Benchmark a Building

How does the EUI of one specific building compare to peer buildings in NYC?

Specific use-case: A building manager wants to learn about the energy usage in a building he is responsible for to detect potential for energy savings.
How does the normalized EUI of my office building (built in the 1970s) compare to peer office buildings in NYC?

Question 2: Correlation EUI and WUI

Do buildings with a high EUI also have a high WUI (water usage index)?

Specific use-case: A city employee (e.g. of the NYCHA department for public housing) needs to answer the questions which City-owned buildings to retrofit first and which measures to implement:
Which residential buildings in NYC are within the highest 10% in terms of EUI? Do they also rank in the top 10% of water usage?

Tasks

Question 1: Benchmark a Building

How does the EUI of one specific building compare to peer buildings in NYC?

Specific use-case: A building manager wants to learn about the energy usage in a building he is responsible for to detect potential for energy savings.
How does the normalized EUI of my office building (built in the 1970s) compare to peer office buildings in NYC?

Question 2: Correlation EUI and WUI

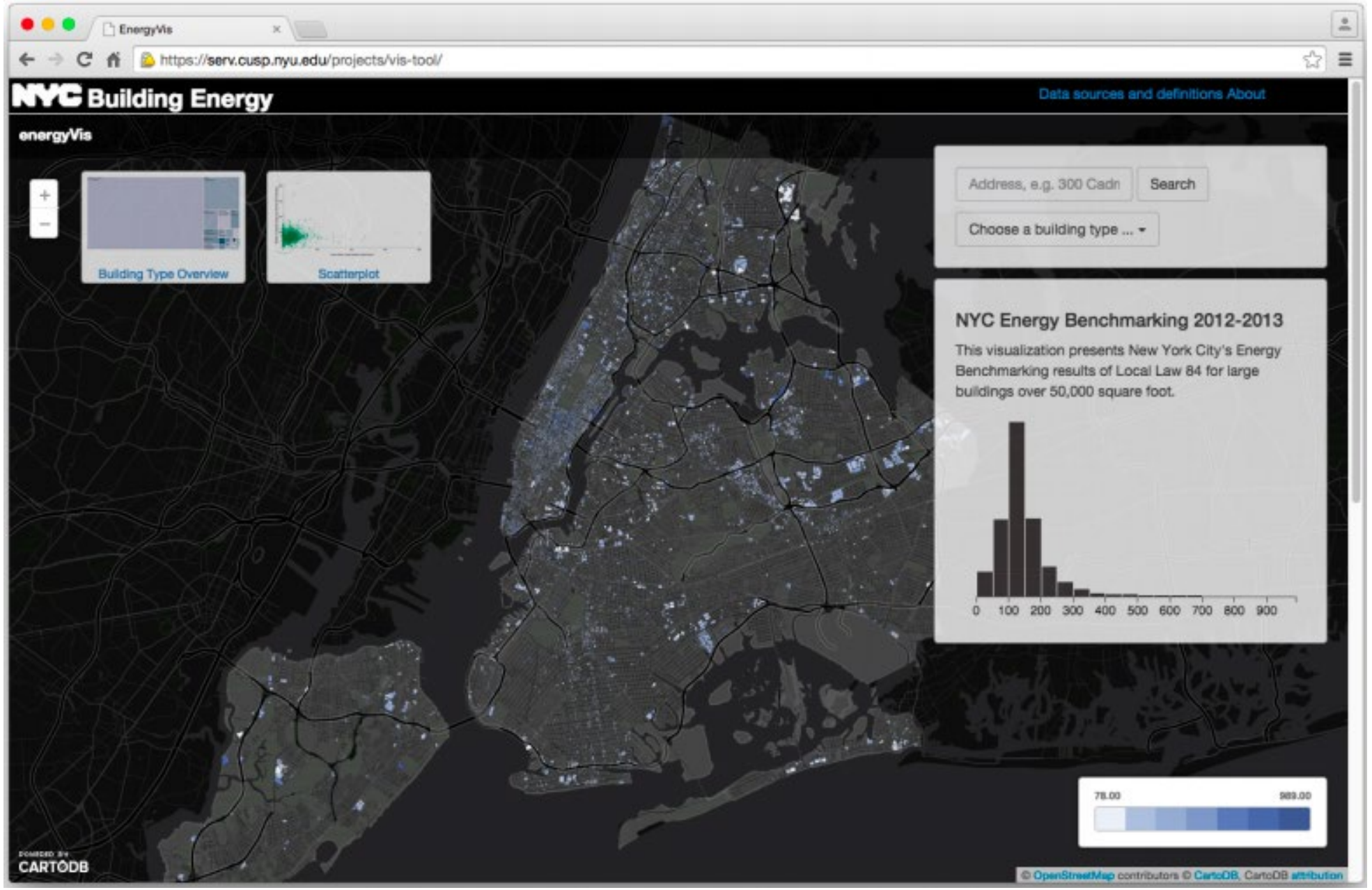
Do buildings with a high EUI also have a high WUI (water usage index)?

Specific use-case: A city employee (e.g. of the NYCHA department for public housing) needs to answer the questions which City-owned buildings to retrofit first and which measures to implement:
Which residential buildings in NYC are within the highest 10% in terms of EUI? Do they also rank in the top 10% of water usage?

Question 3: Time Analysis for a Specific Area

How does the EUI in a specific neighborhood develop over the last 4 years?

Specific use-case: A city employee of the OLTPS department wants to find out, what improvements in energy usage have been made since Local Law 84 was introduced in 2010.


EnergyVis
https://serv.cusp.nyu.edu/projects/vis-tool/
NYC Building Energy
Data sources and definitions About
energyVis
Building Type Overview
Scatterplot
Address, e.g. 300 Cadn
Search
Choose a building type ...
NYC Energy Benchmarking 2012-2013
This visualization presents New York City's Energy Benchmarking results of Local Law 84 for large buildings over 50,000 square foot.
0 100 200 300 400 500 600 700 800 900
78.00
989.00
CARTODB
© OpenStreetMap contributors © CartoDB, CartoDB attribution

However…

- Maps provide familiarity to people.
  - People know where something on a map is.
  - Maps act as an index from spatial to semantic information (and vice-versa).

FINANCIAL TIMES

myFT

HOME WORLD US COMPANIES MARKETS OPINION WORK & CAREERS LIFE & ARTS

Sign In Subscribe

Latest on Data visualisation

Data visualisation: how the FT newsroom designs maps

How to lie with maps

Dump the PowerPoints or risk losing money

How to navigate data's pink and blue problem

The Chart Doctor Data visualisation + Add to myFT

A love of maps should mean using fewer to illustrate data better

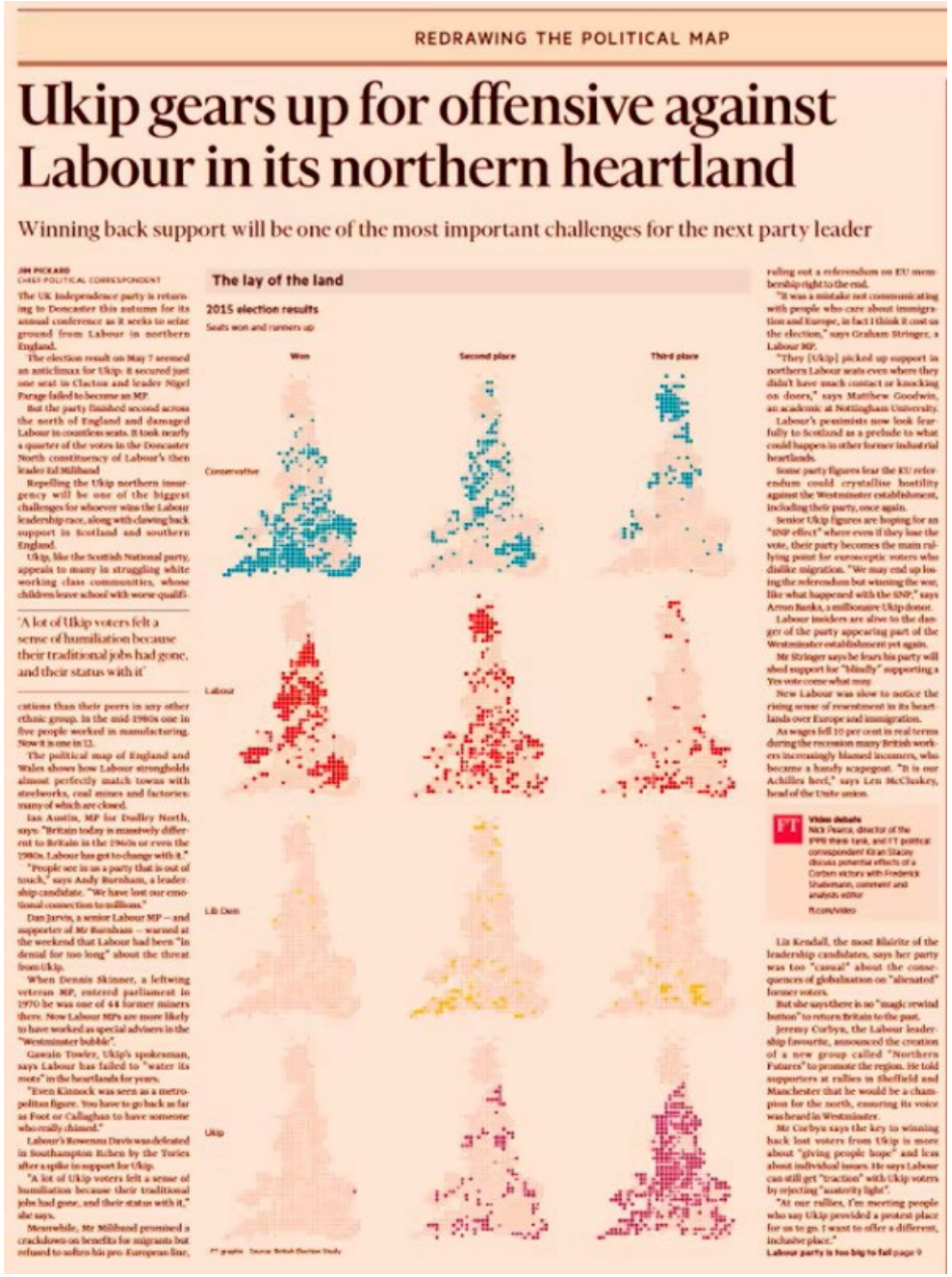
When one map is not enough, and when you should use none at all

Save

When maps are the wrong route

However, maps are not the solution to every visualisation problem with geographical data. As the FT Visual Vocabulary highlights, "spatial" is just one of many possible relationships in data and maps and should be used "only when specific locations or geographical patterns in data are more important to the reader than anything else". Ignore this advice and maps may end up leaving readers lost.

Recently, some painstaking research by Laura Noonan, the FT's investment banking correspondent, produced a data set of top banks' presence in eight cities lining up to take London's role as the EU's banking centre. Nearly 5,000 entities, each allocated to one of the eight cities.



REDRAWING THE POLITICAL MAP

Ukip gears up for offensive against Labour in its northern heartland

Winning back support will be one of the most important challenges for the next party leader

JIM PICKARD
CHIEF POLITICAL CORRESPONDENT

The UK Independence party is returning to Doncaster this autumn for its annual conference as it seeks to seize ground from Labour in northern England.

The election result on May 7 seemed an anticlimax for Ukip: it secured just one seat in Clacton and leader Nigel Farage failed to become an MP.

But the party finished second across the north of England and damaged Labour in countless seats. It took nearly a quarter of the votes in the Doncaster North constituency of Labour's then leader Ed Miliband.

Repelling the Ukip northern insurgency will be one of the biggest challenges for whoever wins the Labour leadership race, along with clawing back support in Scotland and southern England.

Ukip, like the Scottish National party, appeals to many in struggling white working class communities, whose children leave school with worse qualifi-

'A lot of Ukip voters felt a sense of humiliation because their traditional jobs had gone, and their status with it'

cations than their peers in any other ethnic group. In the mid-1980s one in five people worked in manufacturing. Now it is one in 12.

The political map of England and Wales shows how Labour strongholds almost perfectly match towns with steelworks, coal mines and factories many of which are closed.

Ian Austin, MP for Dudley North, says: "Britain today is massively different to Britain in the 1960s or even the 1980s. Labour has got to change with it."

"People see in us a party that is out of touch," says Andy Burnham, a leadership candidate. "We have lost our emotional connection to millions."

Dan Jarvis, a senior Labour MP — and supporter of Mr Burnham — warned at the weekend that Labour had been "in denial for too long" about the threat from Ukip.

When Dennis Skinner, a leftwing veteran MP, entered parliament in 1970 he was one of 44 former miners there. Now Labour MPs are more likely to have worked as special advisers in the "Westminster bubble".

Gawain Towler, Ukip's spokesman, says Labour has failed to "water its roots" in the heartlands for years.

"Even Kinnock was seen as a metropolitan figure. You have to go back as far as Foot or Callaghan to have someone who really chimed."

Labour's Rowenna Davis was defeated in Southampton Itchen by the Tories after a spike in support for Ukip.

"A lot of Ukip voters felt a sense of humiliation because their traditional jobs had gone, and their status with it," she says.

Meanwhile, Mr Miliband promised a crackdown on benefits for migrants but refused to soften his pro-European line, ruling out a referendum on EU membership right to the end.

"It was a mistake not communicating with people who care about immigration and Europe, in fact I think it cost us the election," says Graham Stringer, a Labour MP.

"They [Ukip] picked up support in northern Labour seats even where they didn't have much contact or knocking on doors," says Matthew Goodwin, an academic at Nottingham University.

Labour's pessimists now look fearfully to Scotland as a prelude to what could happen in other former industrial heartlands.

Some party figures fear the EU referendum could crystallise hostility against the Westminster establishment, including their party, once again.

Senior Ukip figures are hoping for an "SNP effect" where even if they lose the vote, their party becomes the main rallying point for eurosceptic voters who dislike migration. "We may end up losing the referendum but winning the war, like what happened with the SNP," says Arron Banks, a millionaire Ukip donor.

Labour insiders are alive to the danger of the party appearing part of the Westminster establishment yet again.

Mr Stringer says he fears his party will shed support for "blindly" supporting a Yes vote come what may.

New Labour was slow to notice the rising sense of resentment in its heartlands over Europe and immigration.

As wages fell 10 per cent in real terms during the recession many British workers increasingly blamed incomers, who became a handy scapegoat. "It is our Achilles heel," says Len McCluskey, head of the Unite union.

The lay of the land

2015 election results

Seats won and runners up

FT Video Debate

Nick Pearce, director of the IPPR think-tank, and FT political correspondent Kiran Stacey discuss potential effects of a Corbyn victory with Frederick Studemann, comment and analysis editor

ft.com/video

Liz Kendall, the most Blairite of the leadership candidates, says her party was too "casual" about the consequences of globalisation on "alienated" former voters.

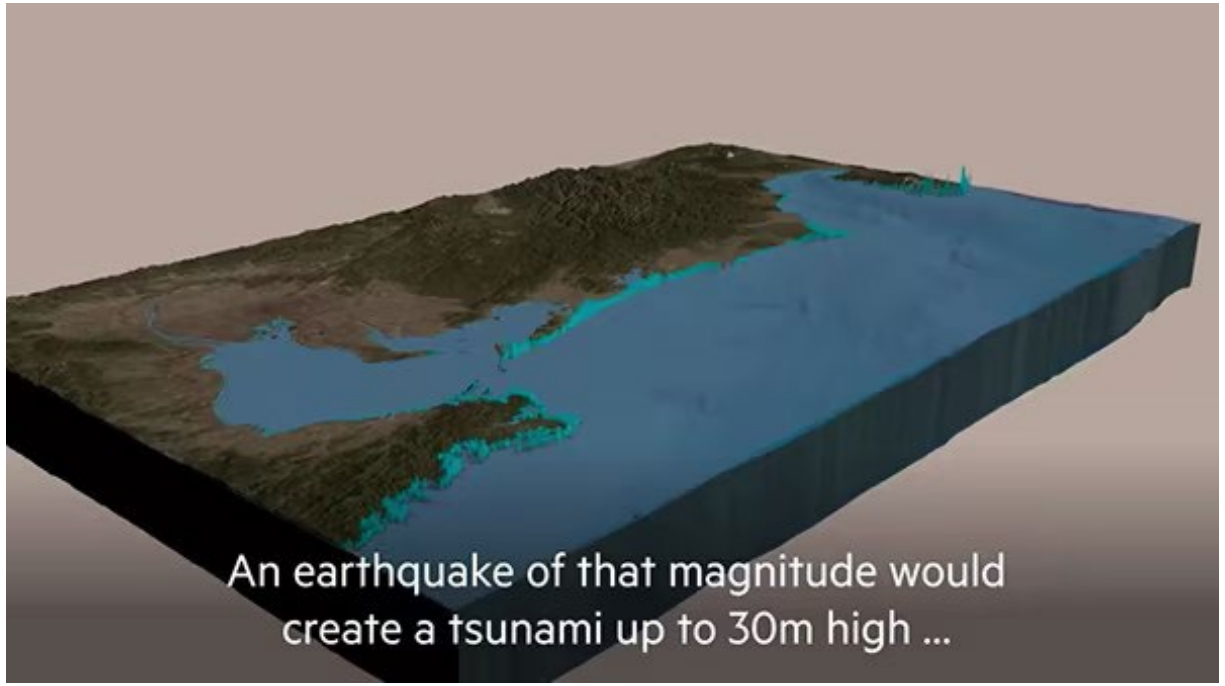
But she says there is no "magic rewind button" to return Britain to the past.

Jeremy Corbyn, the Labour leadership favourite, announced the creation of a new group called "Northern Futures" to promote the region. He told supporters at rallies in Sheffield and Manchester that he would be a champion for the north, ensuring its voice was heard in Westminster.

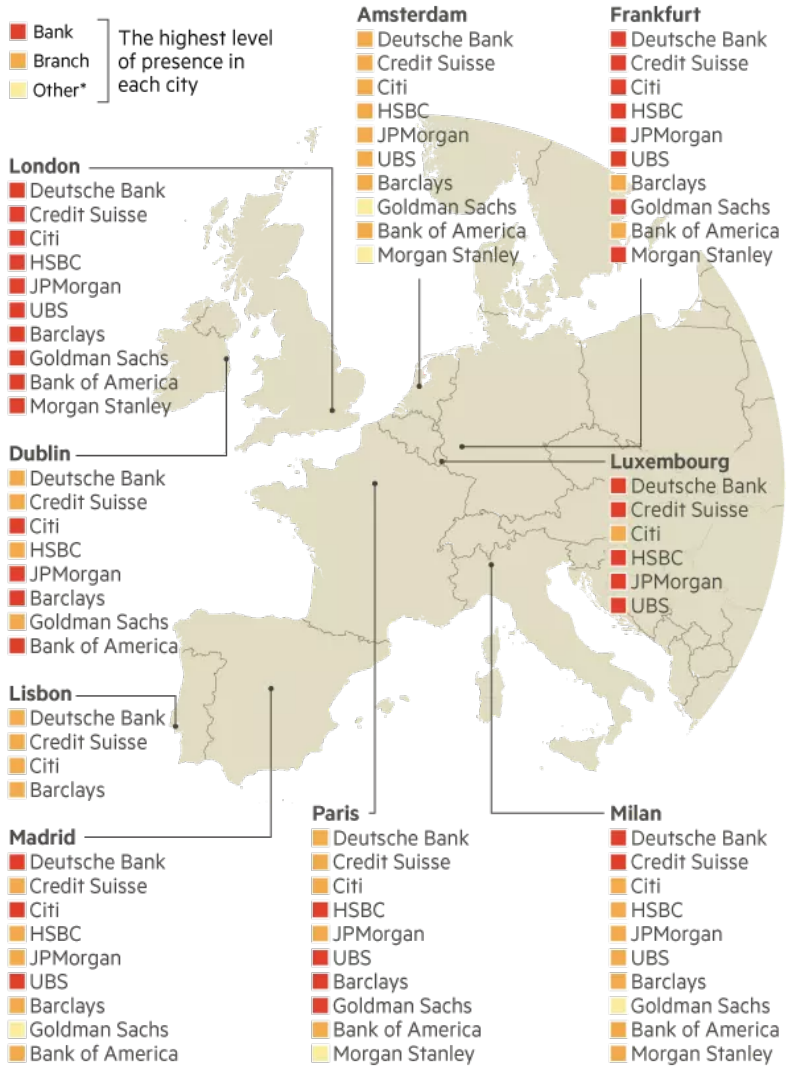
Mr Corbyn says the key to winning back lost voters from Ukip is more about "giving people hope" and less about individual issues. He says Labour can still get "traction" with Ukip voters by rejecting "austerity lite".

"At our rallies, I'm meeting people who say Ukip provided a protest place for us to go. I want to offer a different, inclusive place."

Labour party is too big to fail page 9



The Brexit banking matrix: The contenders lining up for London's crown

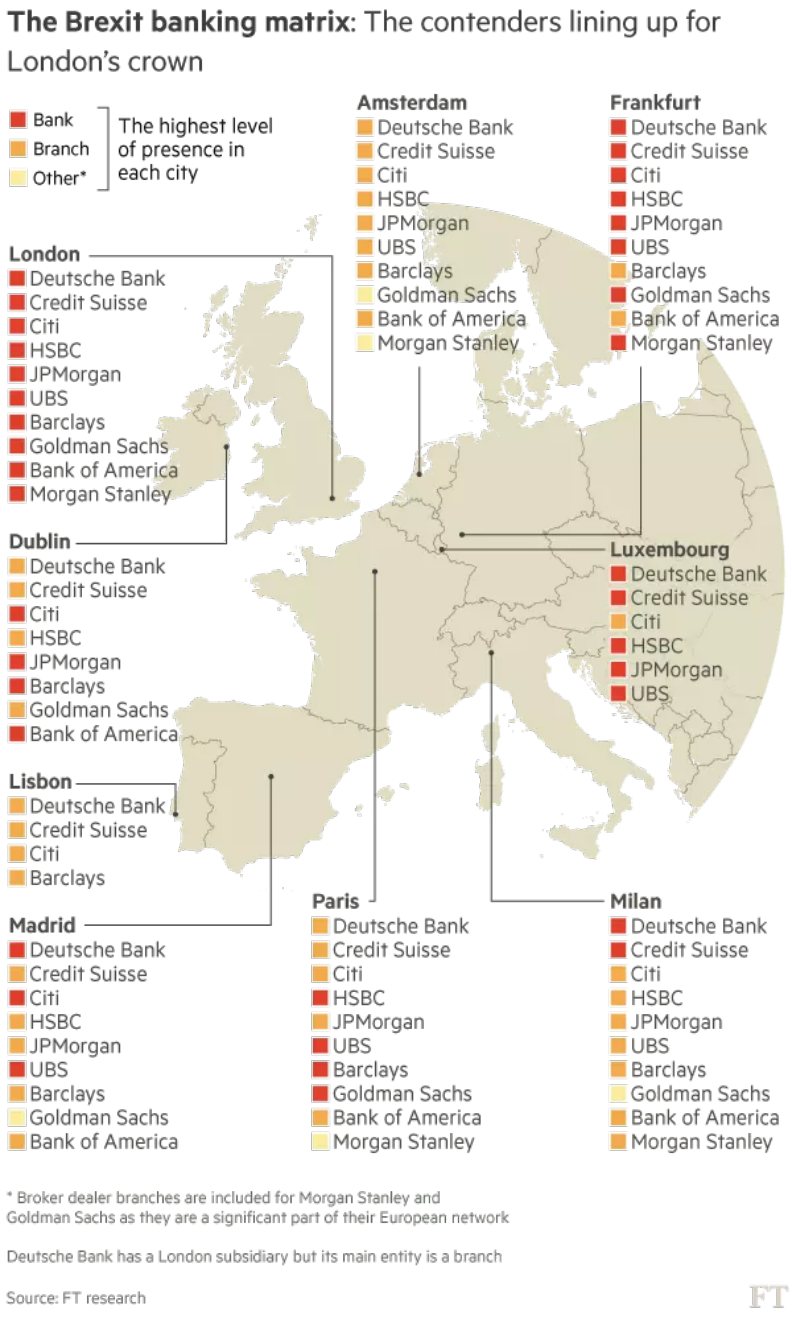


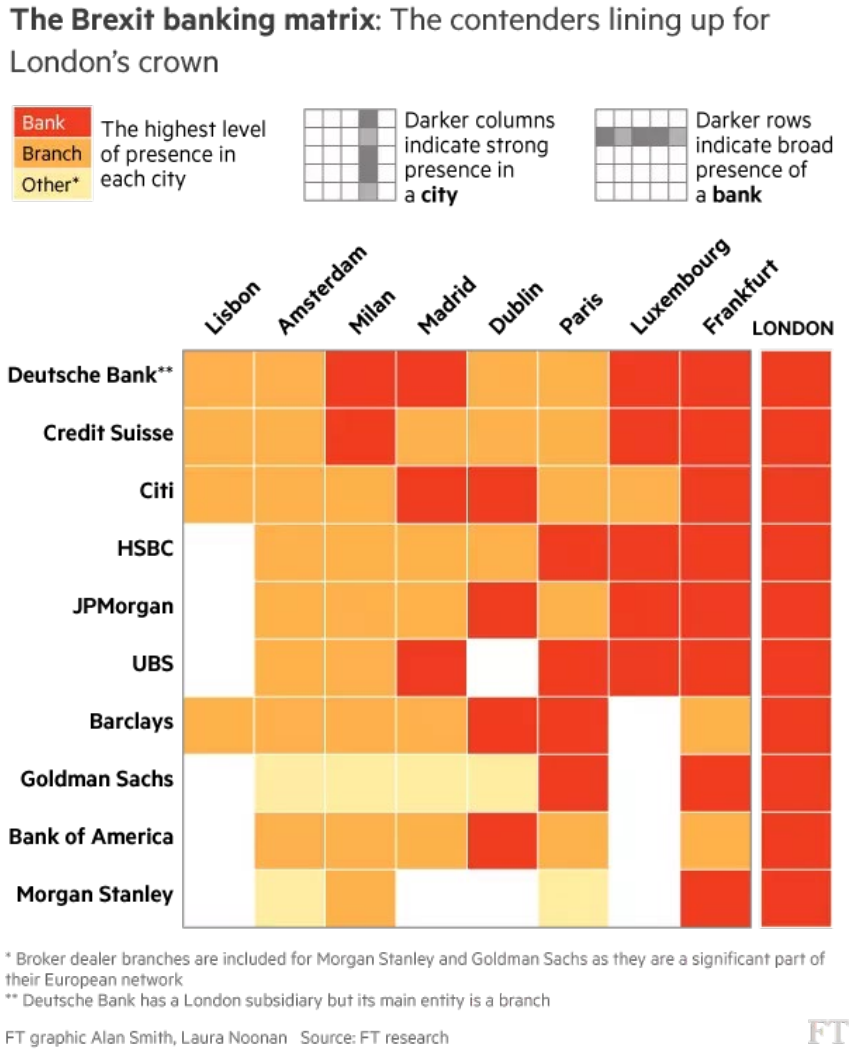
* Broker dealer branches are included for Morgan Stanley and Goldman Sachs as they are a significant part of their European network

Deutsche Bank has a London subsidiary but its main entity is a branch

Source: FT research

FT


The Brexit banking matrix: The contenders lining up for London's crown
Bank
Branch
Other*
The highest level of presence in each city
Amsterdam
Deutsche Bank
Credit Suisse
Citi
HSBC
JPMorgan
UBS
Barclays
Goldman Sachs
Bank of America
Morgan Stanley
Frankfurt
Deutsche Bank
Credit Suisse
Citi
HSBC
JPMorgan
UBS
Barclays
Goldman Sachs
Bank of America
Morgan Stanley
London
Deutsche Bank
Credit Suisse
Citi
HSBC
JPMorgan
UBS
Barclays
Goldman Sachs
Bank of America
Morgan Stanley
Dublin
Deutsche Bank
Credit Suisse
Citi
HSBC
JPMorgan
Barclays
Goldman Sachs
Bank of America
Luxembourg
Deutsche Bank
Credit Suisse
Citi
HSBC
JPMorgan
UBS
Lisbon
Deutsche Bank
Credit Suisse
Citi
Barclays
Madrid
Deutsche Bank
Credit Suisse
Citi
HSBC
JPMorgan
UBS
Barclays
Goldman Sachs
Bank of America
Paris
Deutsche Bank
Credit Suisse
Citi
HSBC
JPMorgan
UBS
Barclays
Goldman Sachs
Bank of America
Morgan Stanley
Milan
Deutsche Bank
Credit Suisse
Citi
HSBC
JPMorgan
UBS
Barclays
Goldman Sachs
Bank of America
Morgan Stanley
* Broker dealer branches are included for Morgan Stanley and Goldman Sachs as they are a significant part of their European network
Deutsche Bank has a London subsidiary but its main entity is a branch
Source: FT research
FT


The Brexit banking matrix: The contenders lining up for London's crown
Bank
Branch
Other*
The highest level of presence in each city
Darker columns indicate strong presence in a city
Darker rows indicate broad presence of a bank
Lisbon
Amsterdam
Milan
Madrid
Dublin
Paris
Luxembourg
Frankfurt
LONDON
Deutsche Bank**
Credit Suisse
Citi
HSBC
JPMorgan
UBS
Barclays
Goldman Sachs
Bank of America
Morgan Stanley
* Broker dealer branches are included for Morgan Stanley and Goldman Sachs as they are a significant part of their European network
** Deutsche Bank has a London subsidiary but its main entity is a branch
FT graphic Alan Smith, Laura Noonan Source: FT research
FT

Always ask yourself: Do I need a map here. If yes, why?

Questions: When & why to use a map?

You receive a data set containing information about where and when people pick up taxi cab in New York City. The data set contains information about latitude and longitude, time of pick up, and neighborhood. Which of the following questions would benefit the most from a map visualization?

- How many people are picked up in each neighborhood on average each day?
- Are there any specific points of interest of the city where pick-ups tend to concentrate during the weekend?
- How does the volume of pick-ups change during the course of the day?

Questions: When & why to use a map?

You receive a data set containing information about where and when people pick up taxi cab in New York City. The data set contains information about latitude and longitude, time of pick up, and neighborhood. Which of the following questions would benefit the most from a map visualization?

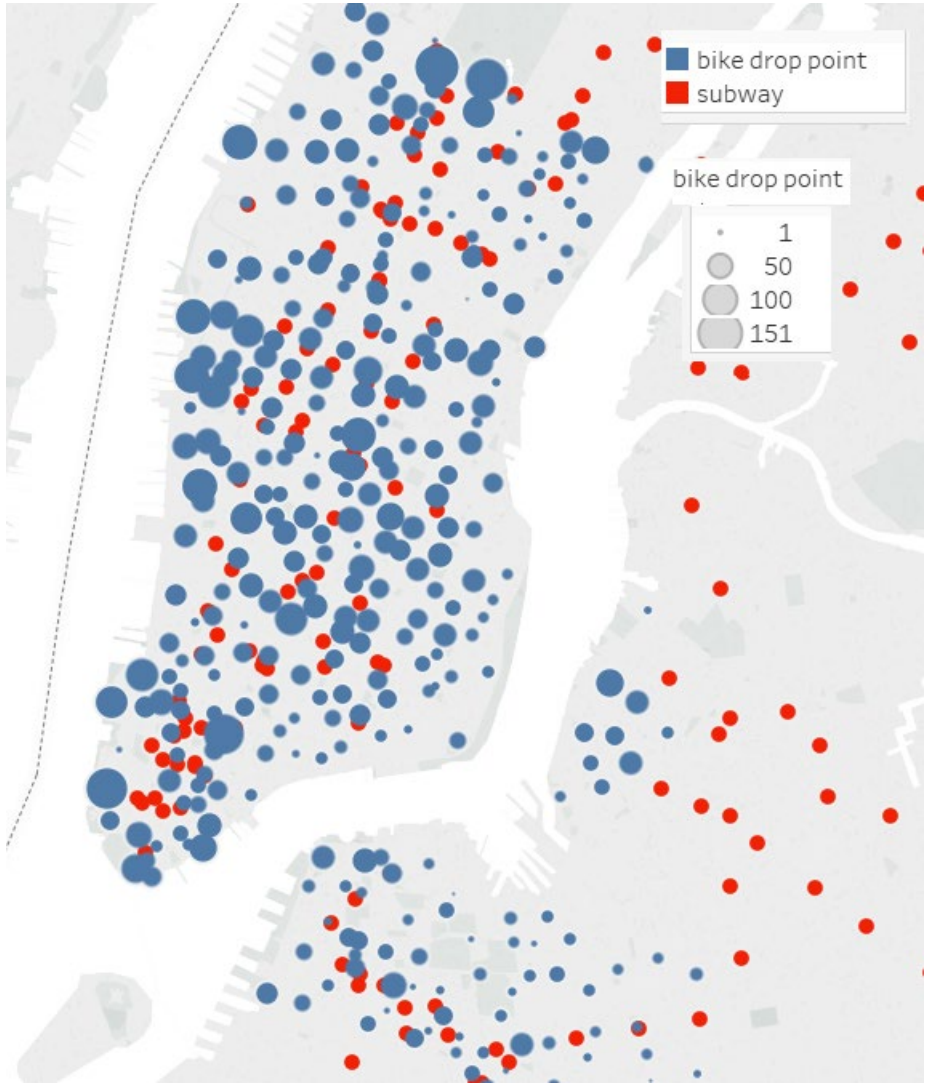
- How many people are picked up in each neighborhood on average each day? **False**
  - The question is not inherently spatial. One can just build a graph with the count of pick-ups for each neighborhood.
- Are there any specific points of interest of the city where pick-ups tend to concentrate during the weekend? **True**
  - Correct. A map is going to make it easier to figure out whether areas with high density cluster around any point of interest (better if potential point of interest are also depicted in the map).
- How does the volume of pick-ups change during the course of the day? **False**
  - The question does not require the use of any spatial information.

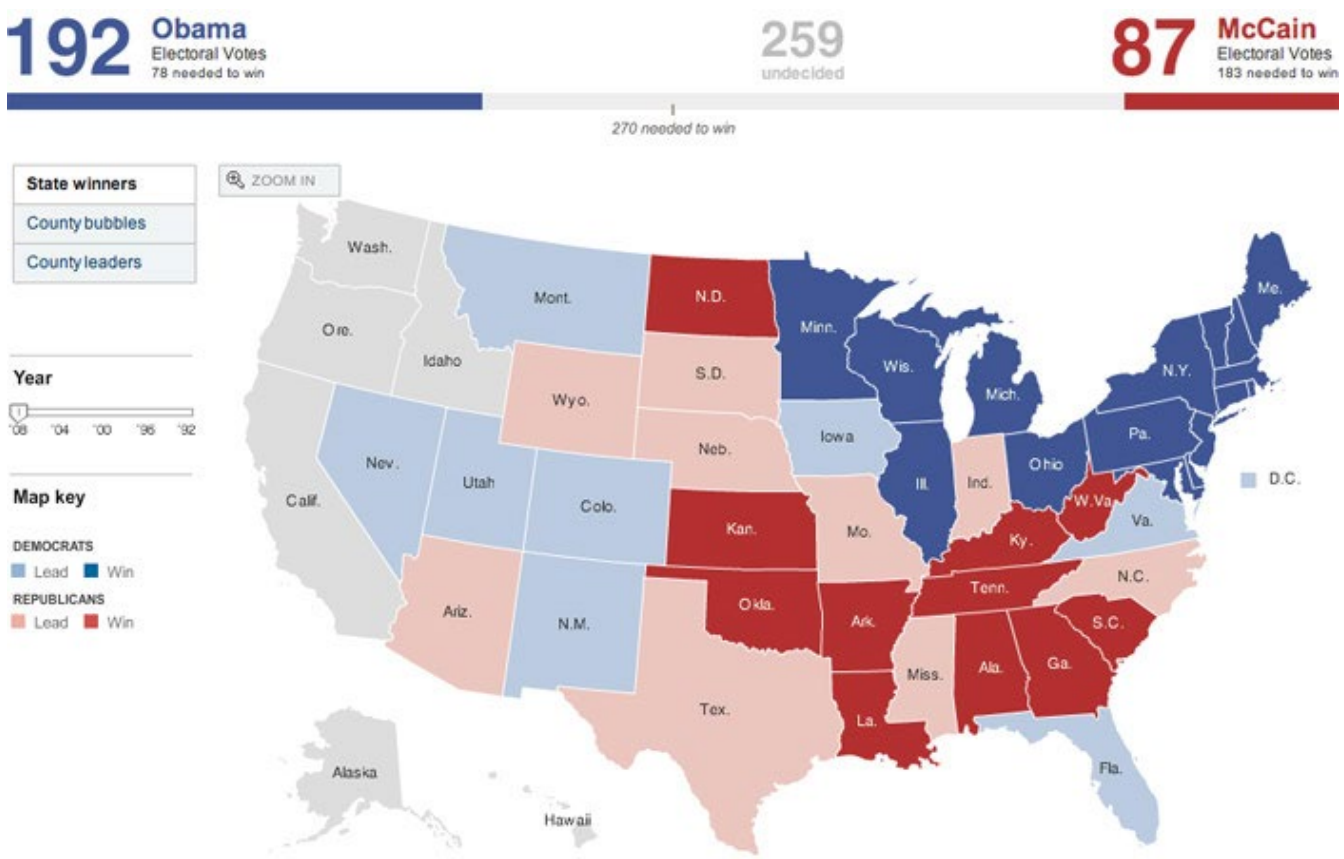
Choice of visualization

- Best choices for spatial visualization?
- Parameters for choosing best visualization:
  - Distribution of items or values.
  - Distribution is discrete or continuous.

Items vs. values

- How do the items distribute geographically?
  - People, cars, events, sensors, etc.
- How do the values distribute geographically?
  - Unemployment, number of votes, temperature, etc.


bike drop point
subway
bike drop point
1
50
100
151


192
Obama
Electoral Votes
78 needed to win
259
undecided
87
McCain
Electoral Votes
183 needed to win
270 needed to win
State winners
County bubbles
County leaders
ZOOM IN
Year
'08 '04 '00 '96 '92
Map key
DEMOCRATS
Lead Win
REPUBLICANS
Lead Win
Wash.
Ore.
Calif.
Idaho
Nev.
Mont.
Wyo.
Utah
Ariz.
Colo.
N.M.
N.D.
S.D.
Neb.
Kan.
Okla.
Tex.
Minn.
Iowa
Mo.
Ark.
La.
Wis.
Ill.
Mich.
Ind.
Ohio
Ky.
Tenn.
Miss.
Ala.
W.Va.
Va.
N.C.
S.C.
Ga.
Fla.
Pa.
N.Y.
Me.
D.C.
Alaska
Hawaii

Items → values

- Items → values (frequency)
- Items can be mapped to values through aggregations.
  - Measures of locality:
    - Mean
    - Median
    - Interquartile mean
  - Measure of spread:
    - Standard deviation
    - Range
    - Variance
    - Absolute deviation
    - Interquartile range

Discrete vs. continuous distributions

- Values associated to discrete locations (e.g., unemployment rate).
- Values associated to all locations (e.g., temperature).

Discrete → continuous

- Spatial interpolation.
- Density estimation.

Item-based vs density-based

Item-based visualization

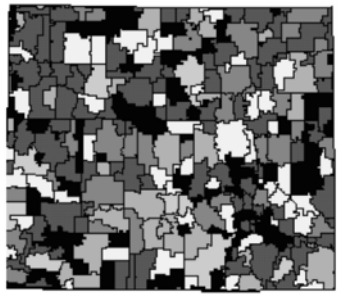


Fig. 1a. Boundaries shifted to the East

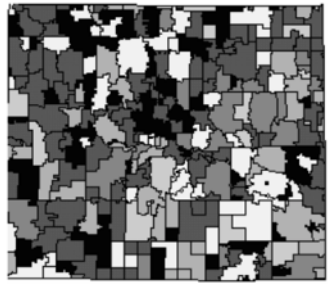


Fig. 1b. Boundaries shifted to the North

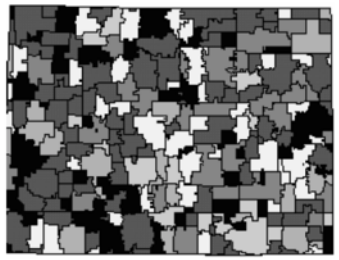


Fig. 1c. Boundaries shifted to the South

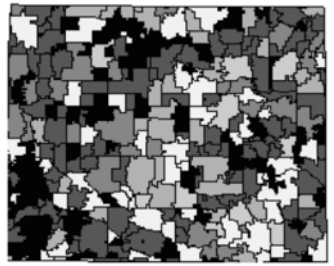
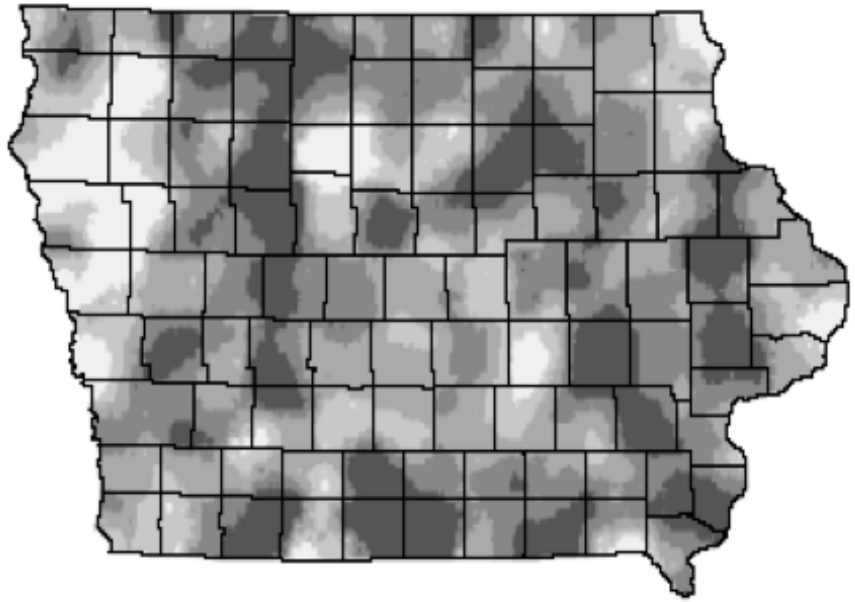


Fig. 1d. Boundaries shifted to the West

Density-based visualization



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

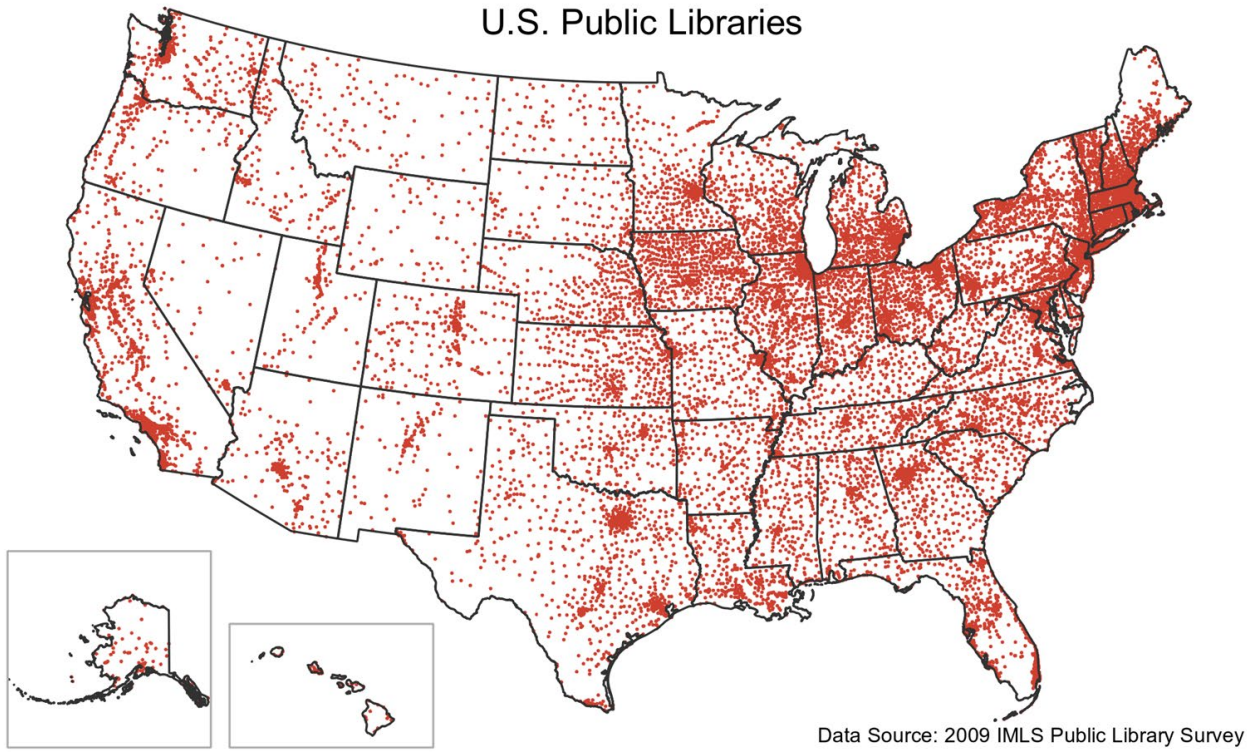
Choice of visualization

- Parameters for choosing best visualization:
  - Distribution of items or values.
  - Distribution is discrete or continuous.

1. Dot maps (distribution of items)
2. Heat maps (distribution of values, continuous)
3. Binned maps (distribution of values, continuous)
4. Choropleth maps (distribution of values, discrete)
5. Symbol maps (distribution of values, discrete)

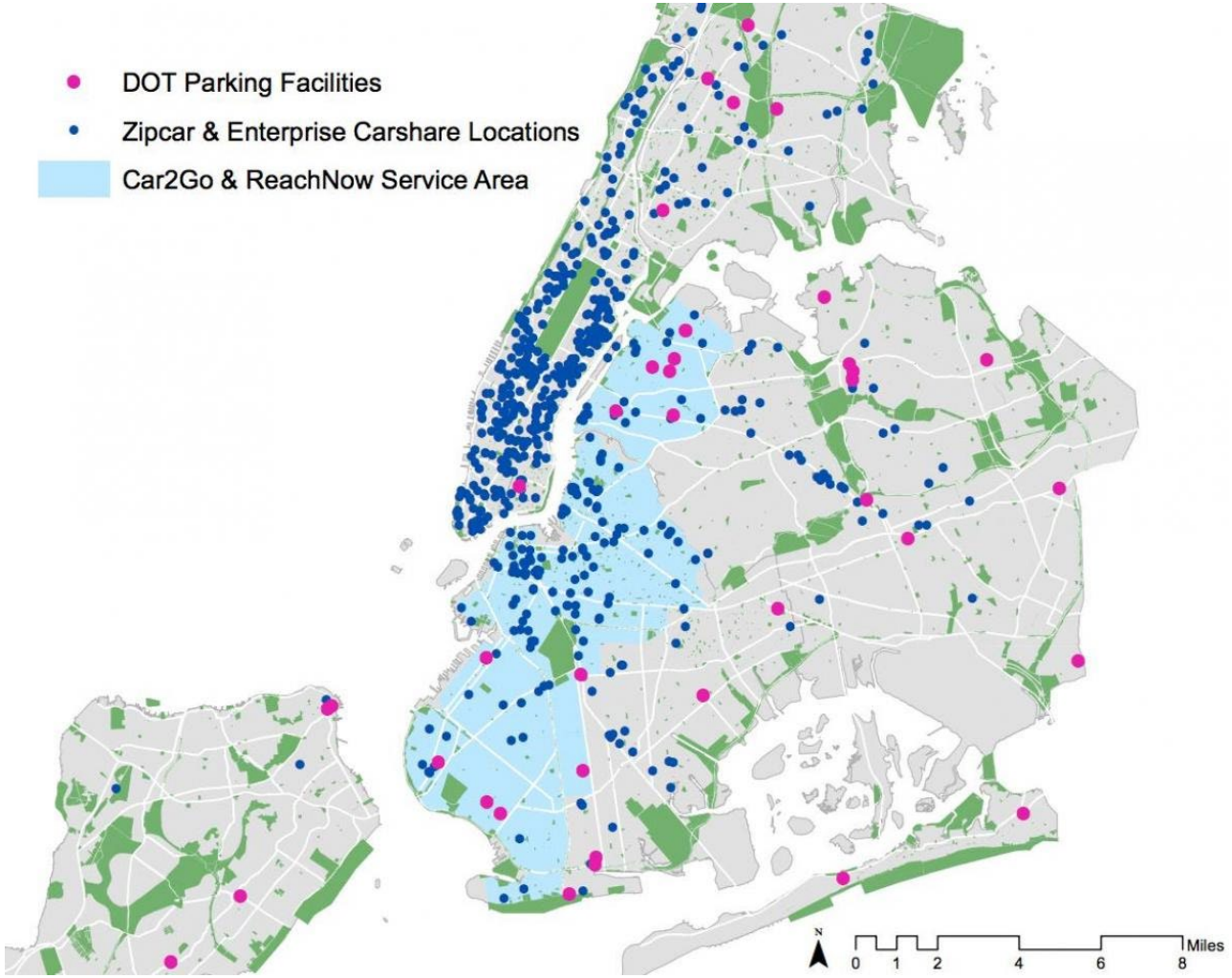
Dot maps

- Each item mapped to its location.
- Useful to depict density and distribution of events and geo-located objects.



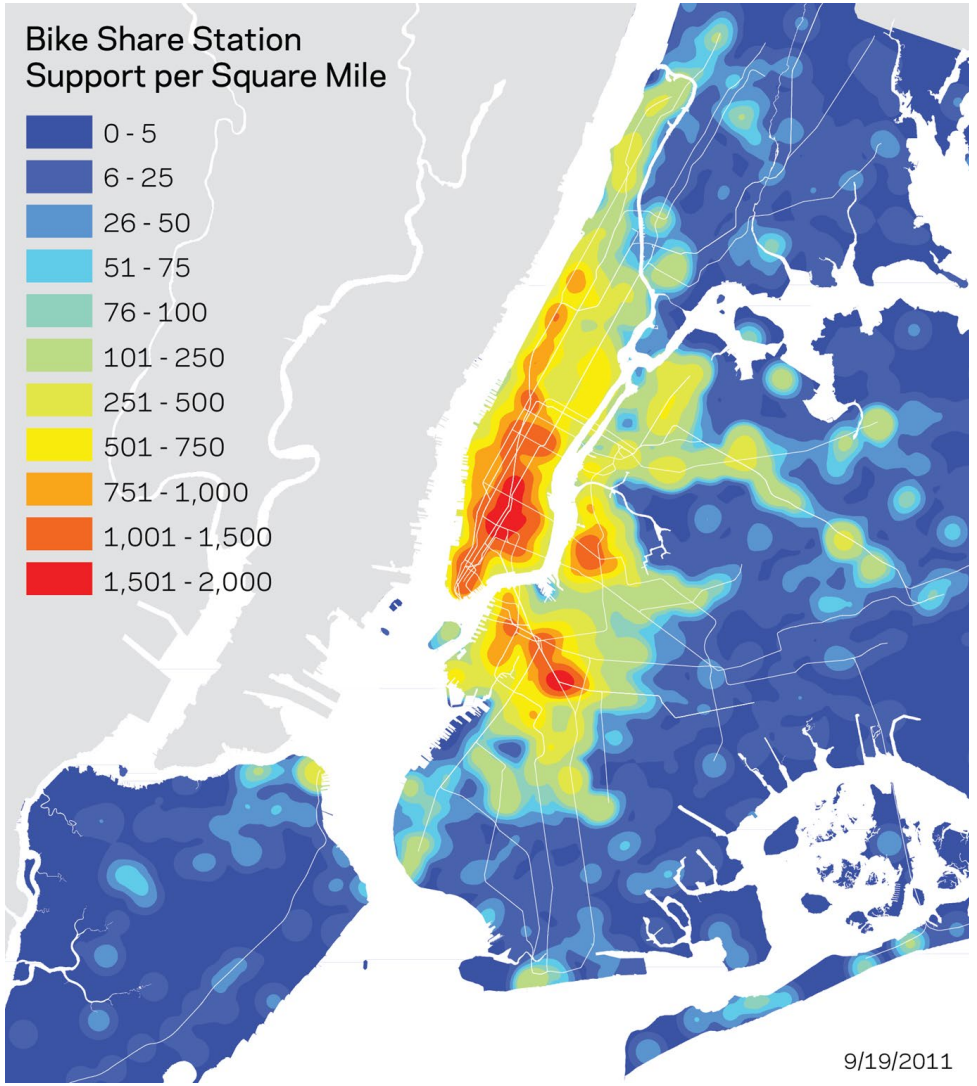
Dot maps

- It can additionally encode some categorical value with color.



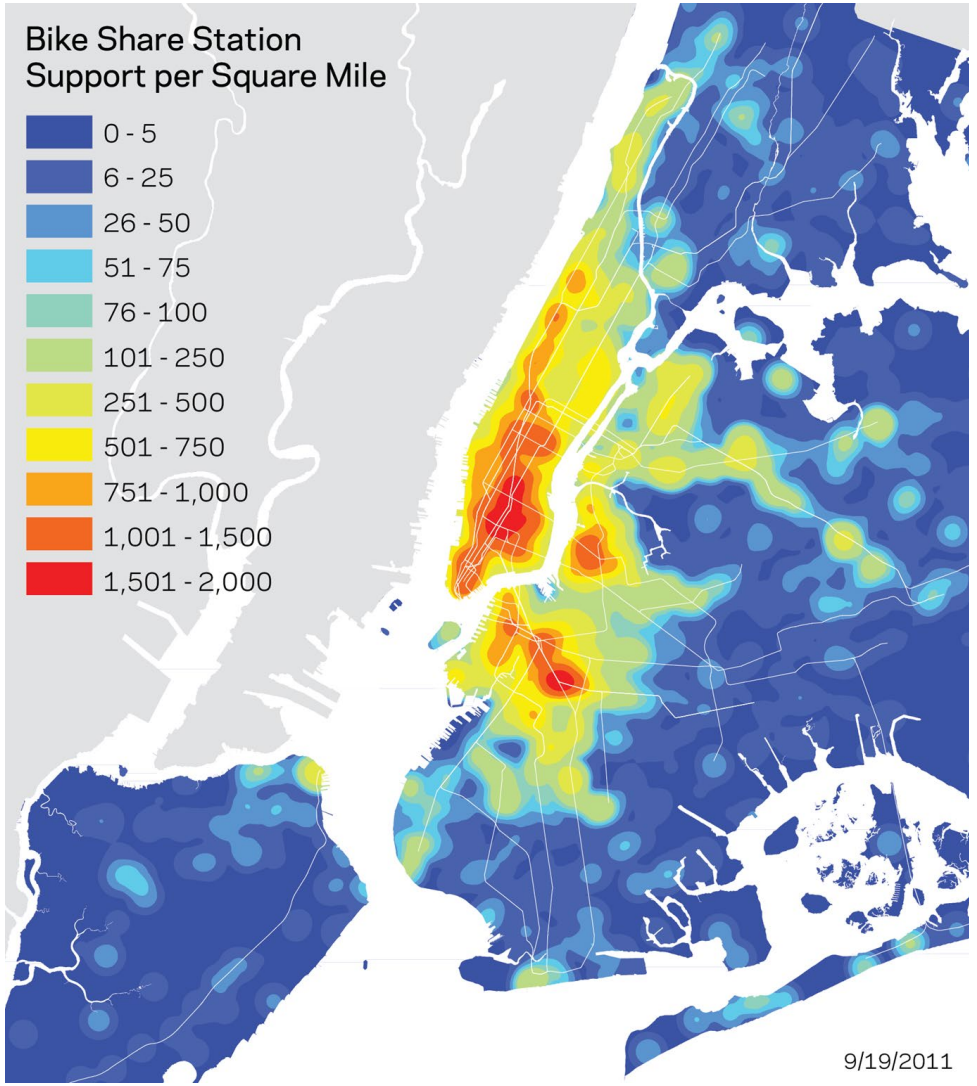
Heat maps

- Typically obtained using a density estimation method, which estimates a **continuous** density model from **discrete** data.



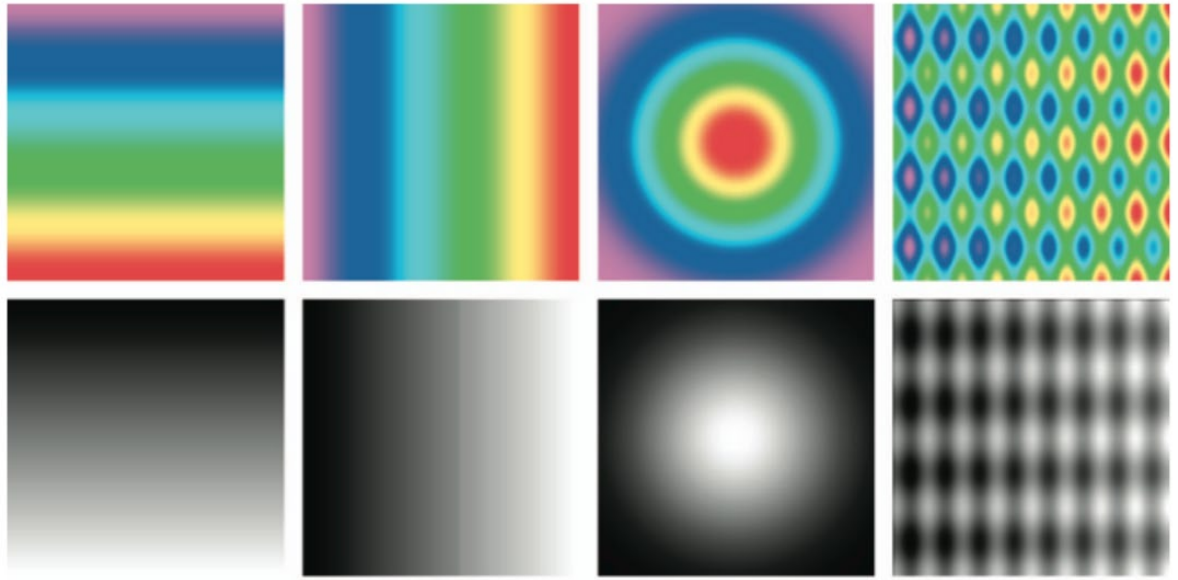
Heat maps

- Two important considerations:
  - Choice of color scale.
  - Choice of density estimation method.



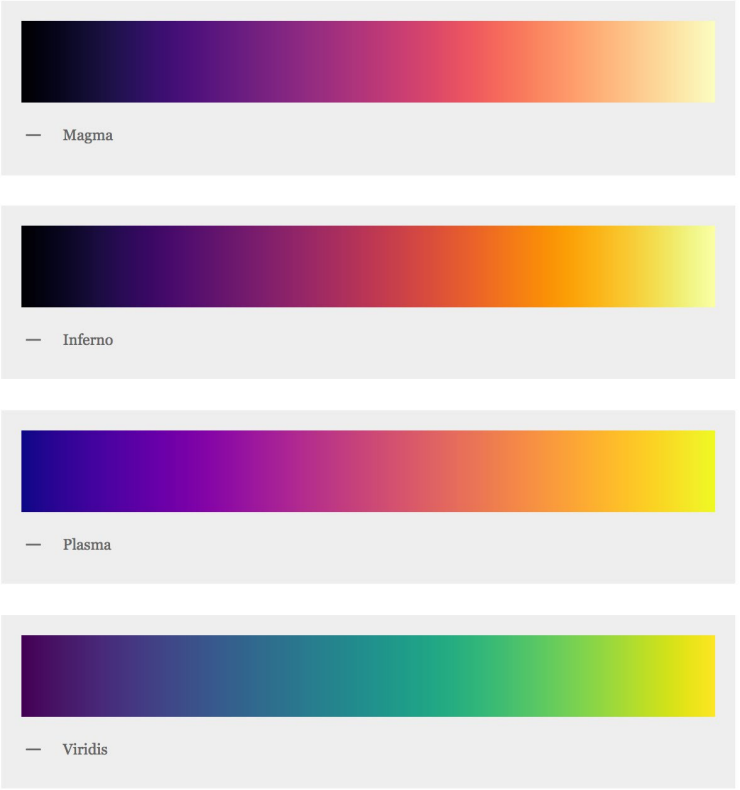
Color scales

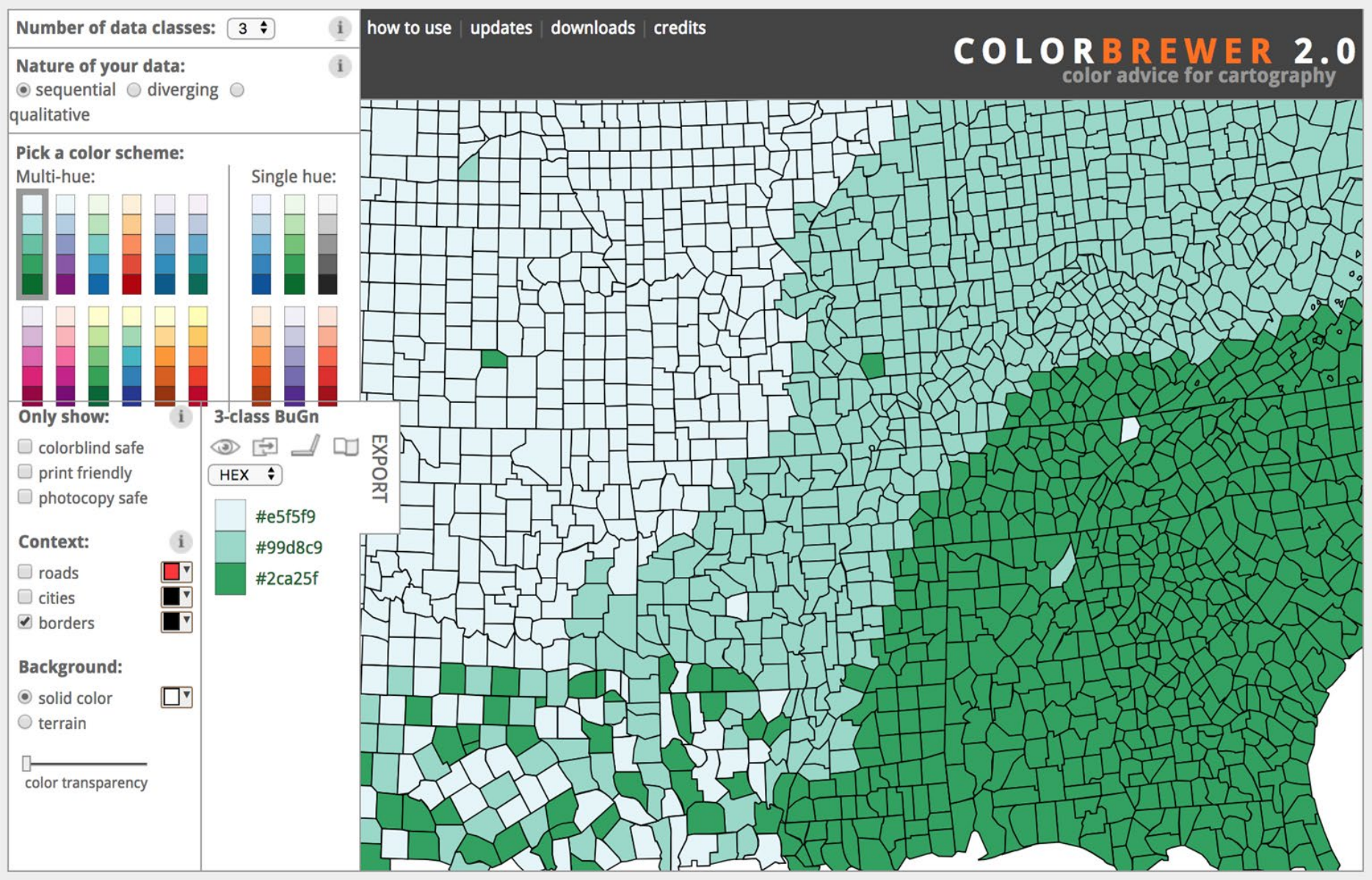
- Appropriate use and design of color scales is crucial for maps.
- Rainbow color scale should be avoided (or used with caution).



Color scales

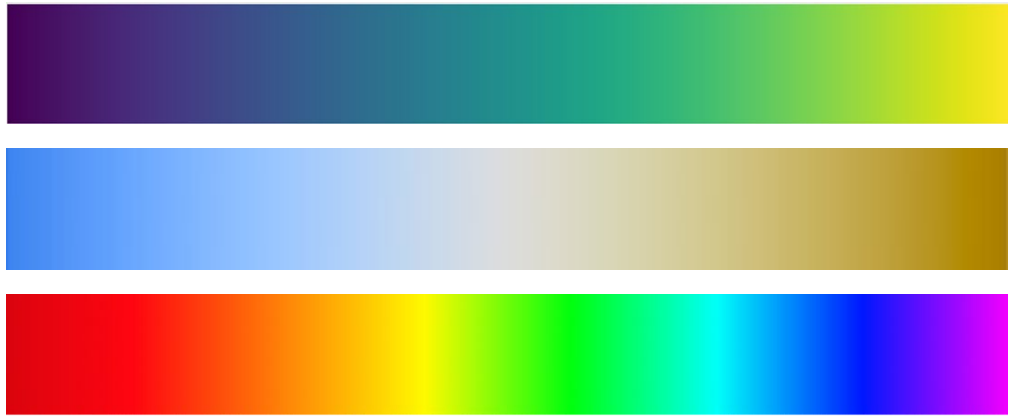
- Perceptually uniform color scales should be used instead:
- Equal steps in data values should map to equal steps in perceived color.



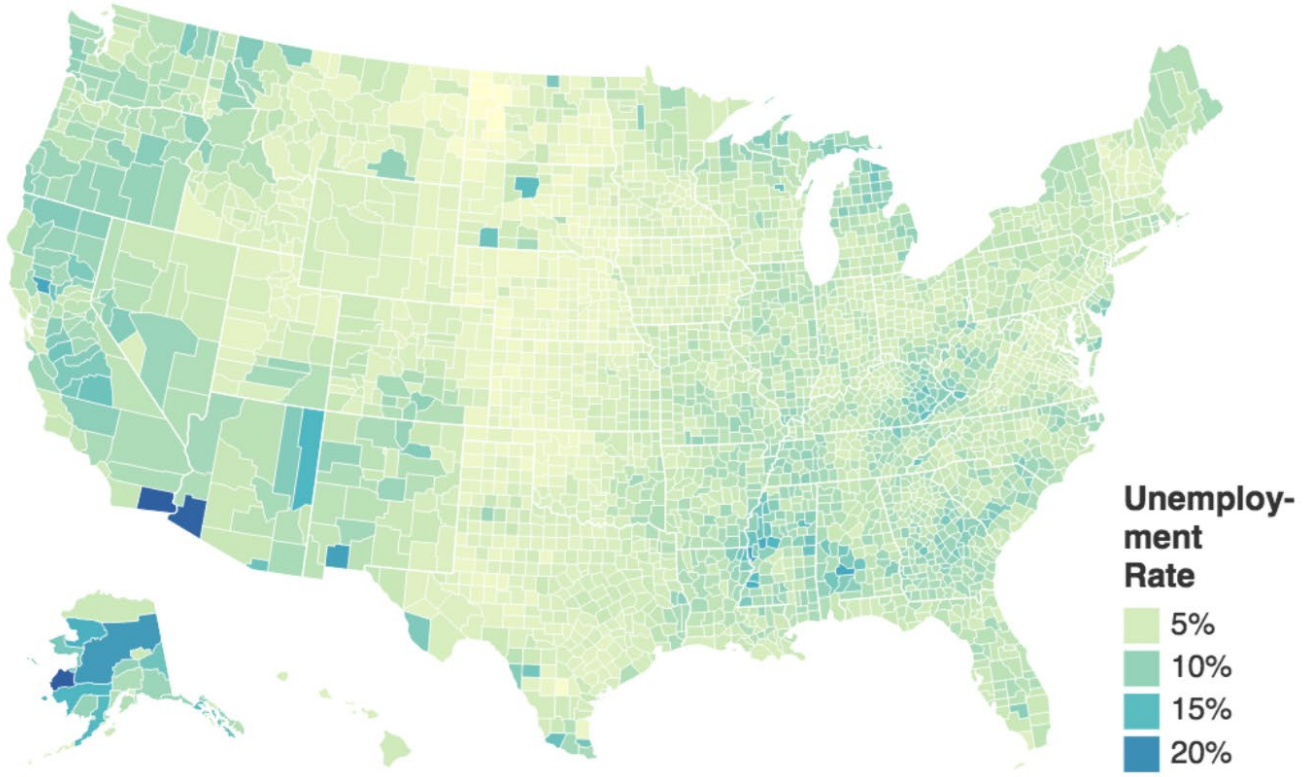

Number of data classes: 3
Nature of your data:
sequential diverging
qualitative
Pick a color scheme:
Multi-hue:
Single hue:
Only show:
colorblind safe
print friendly
photocopy safe
Context:
roads
cities
borders
Background:
solid color
terrain
color transparency
3-class BuGn
HEX
#e5f5f9
#99d8c9
#2ca25f
EXPORT
how to use | updates | downloads | credits
COLORBREWER 2.0
color advice for cartography

Color scales

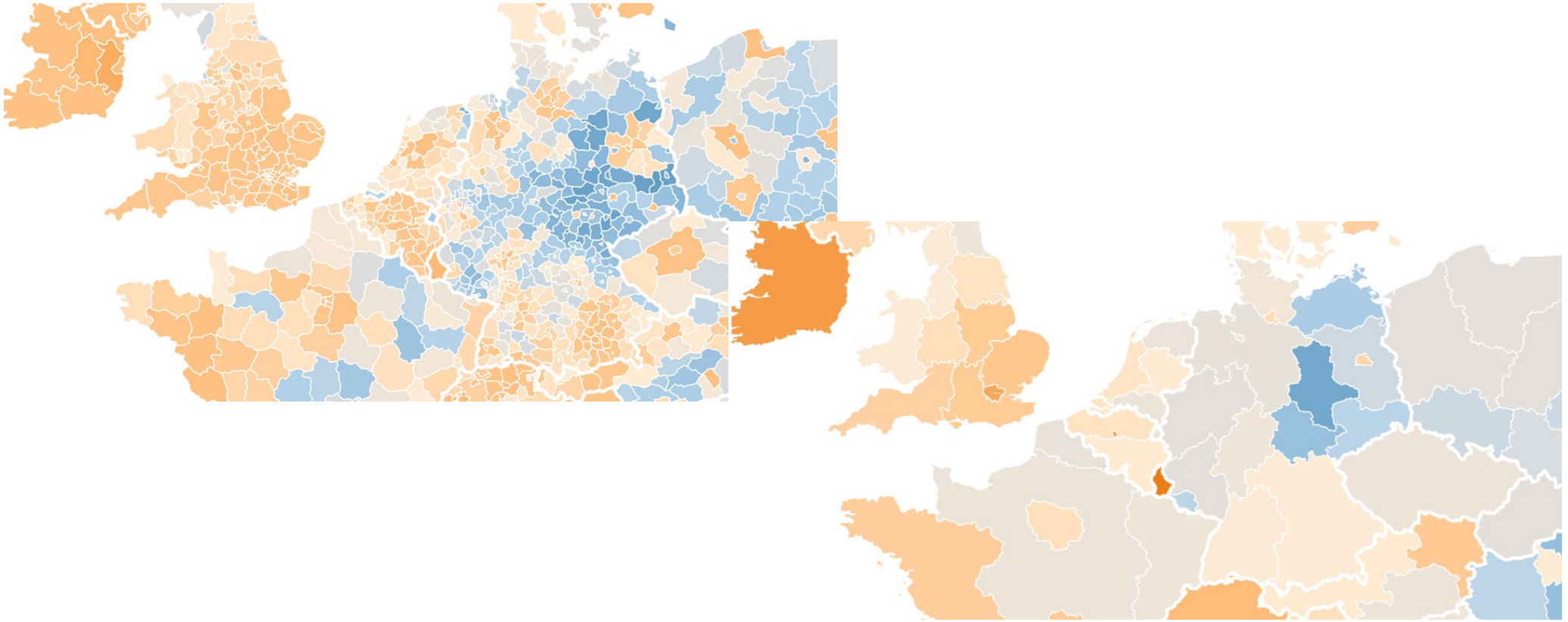
- Which of the following color scales is the most appropriate to depict how the level of noise distributes in a city using a heat map?



Choropleth maps

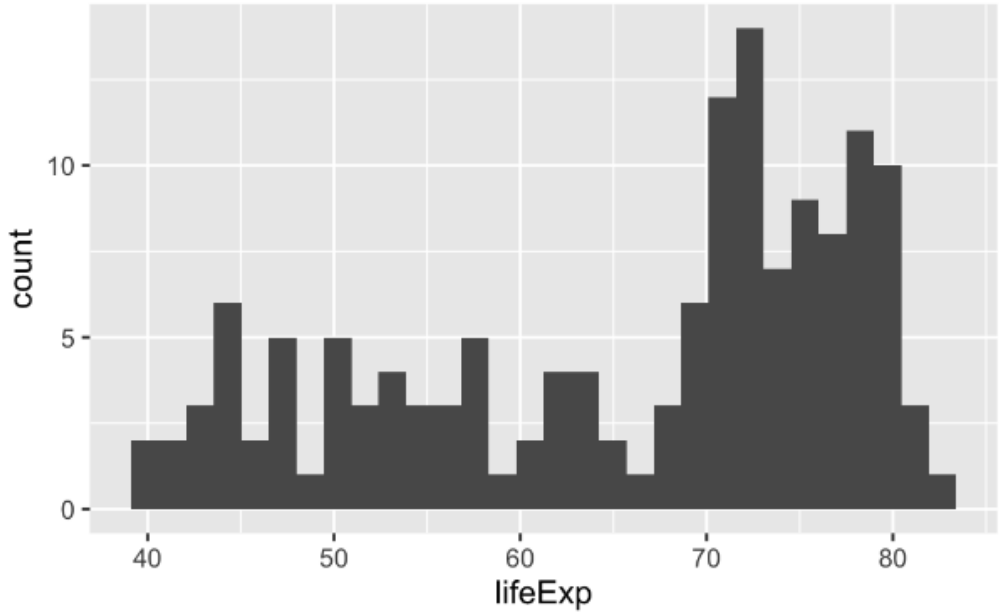


Choropleth maps: Level of detail

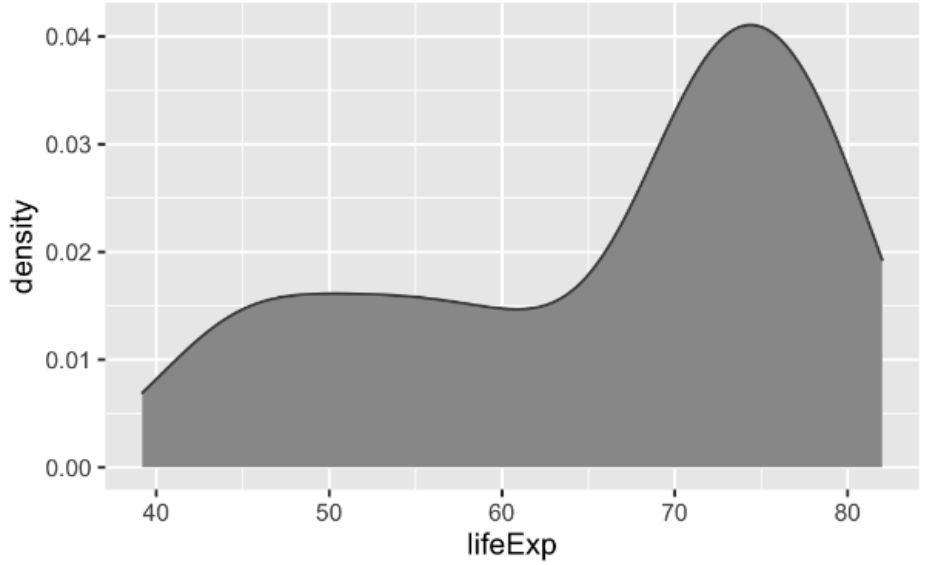


Distributions

Histogram: approximate representation of the distribution of the data.



Density plots: distribution of data over continuous interval.



Density estimation

- Estimating the **shape** of the underlying data function.
  - I.e., estimates a continuous density model from discrete data.
- 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:

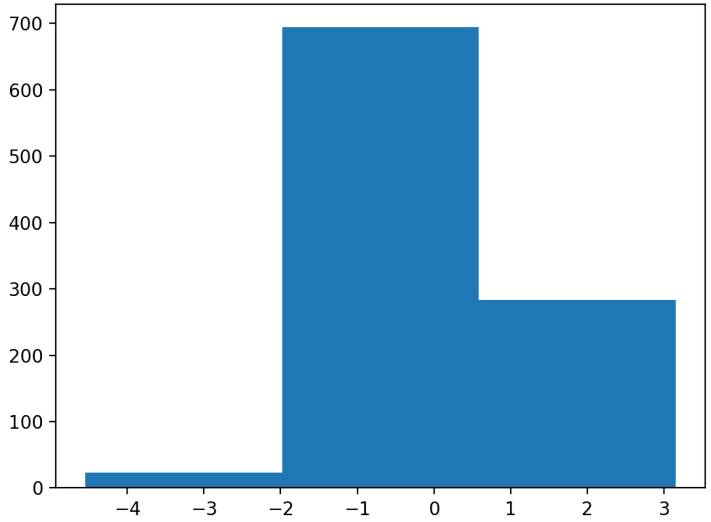
$$\hat{f}(x) = \frac{1}{n} * \frac{[\#\ observations\ in\ same\ bin\ as\ x]}{2h}$$

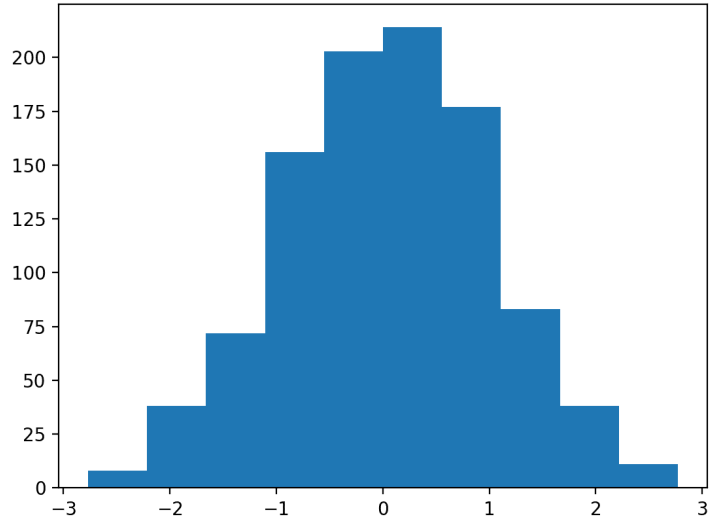
Intervals defined $[x_0 + mh, x_0 + \frac{m+1}{h}]$, origin x_0, bin width h.

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.



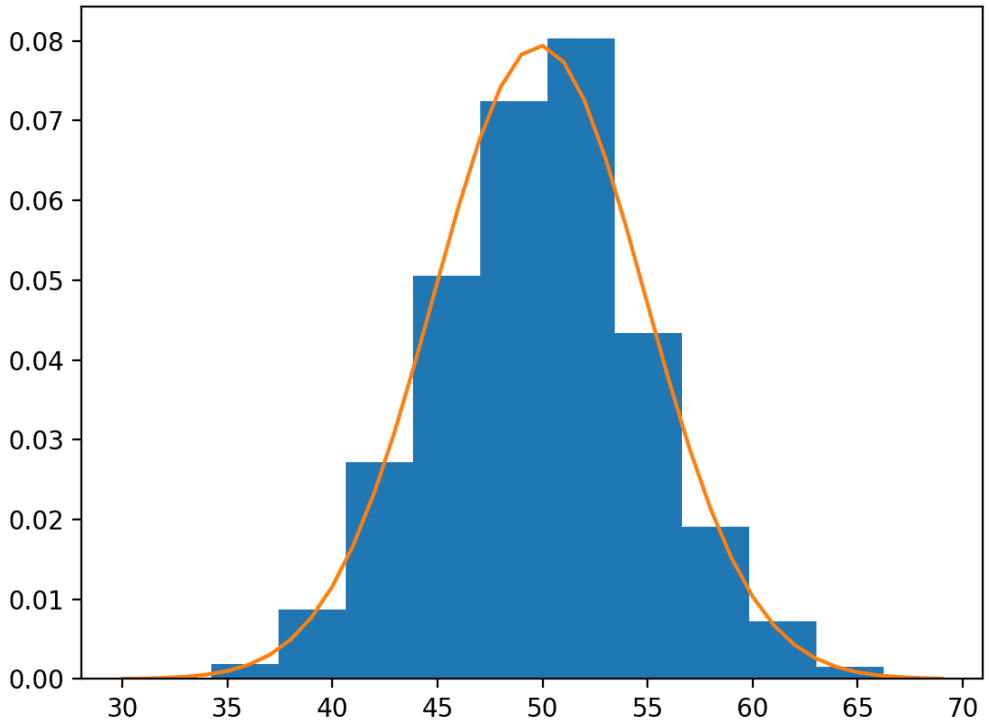


Density estimation

- 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.

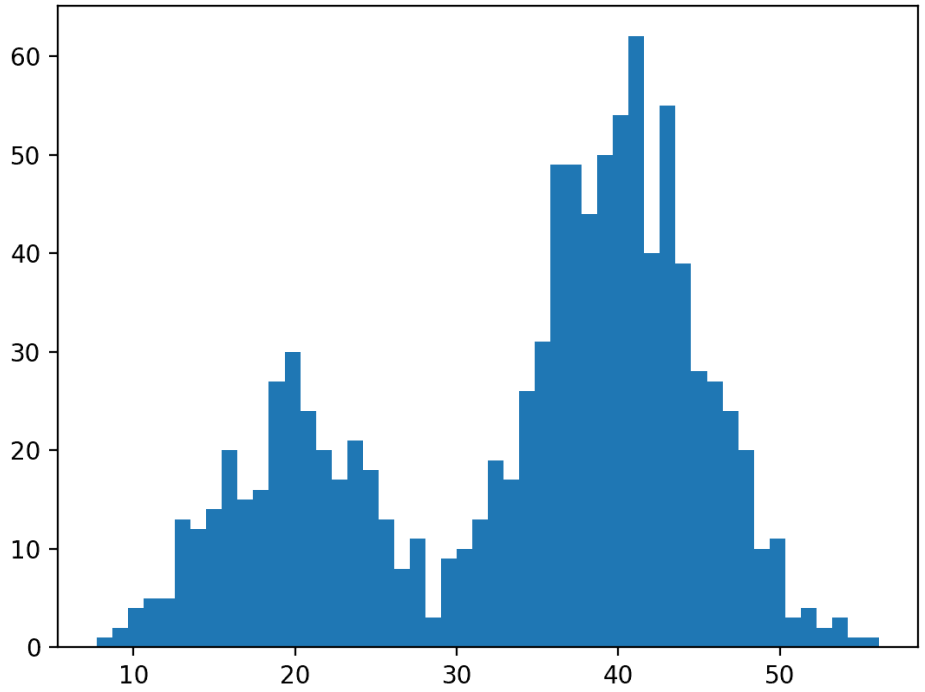
Parametric density estimation

- 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*.



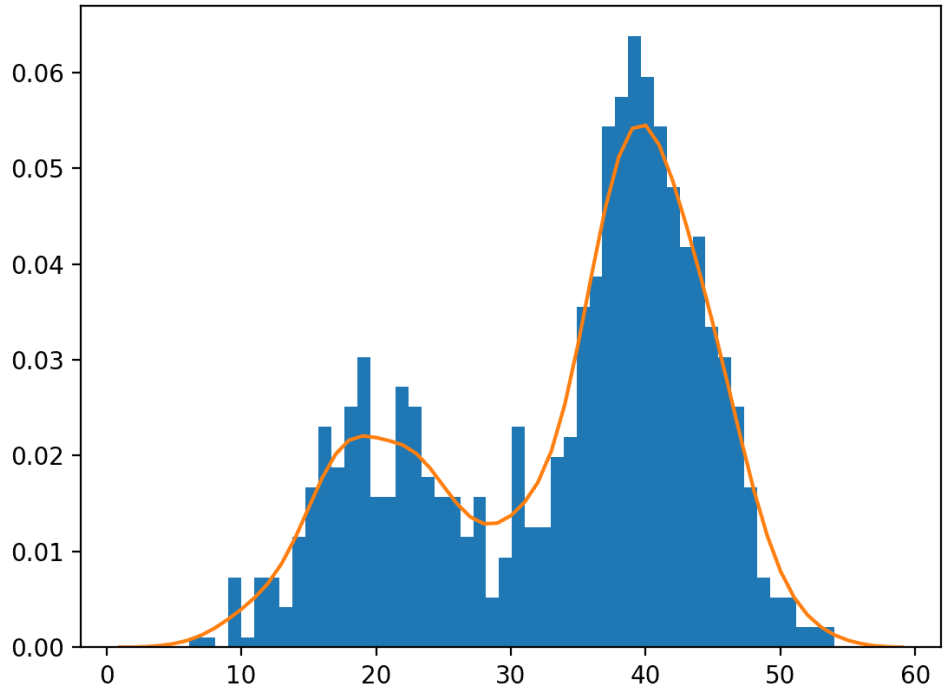
Parametric density estimation

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

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

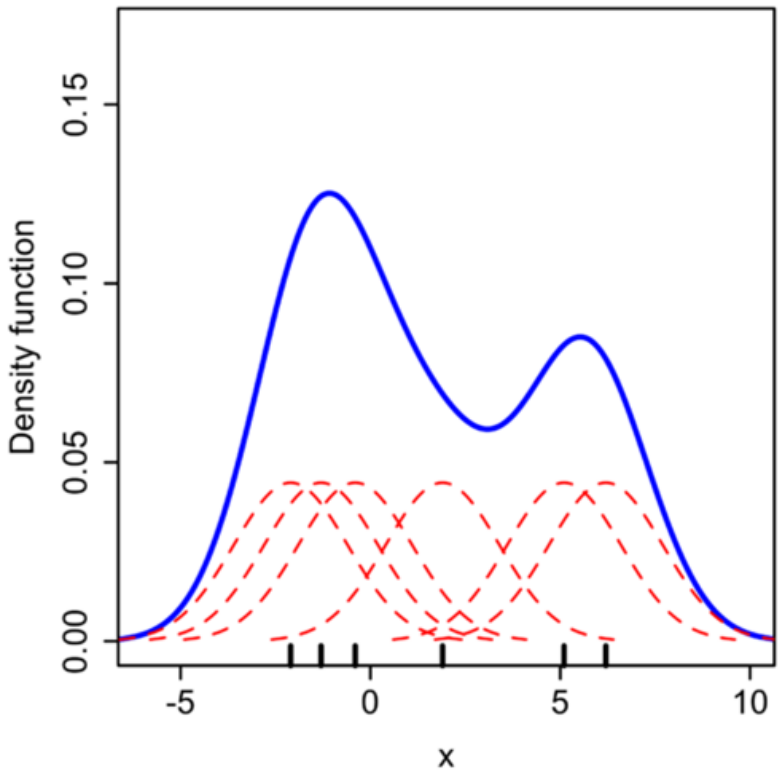
$$\hat{f}(x) = \frac{1}{nh} * \sum_{i=1}^{n} w\left(\frac{x - x_i}{h}\right)$$

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.
- Centers a function (i.e., kernel) at each data point and sums them to get a density estimation.



Kernel density estimation

- 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

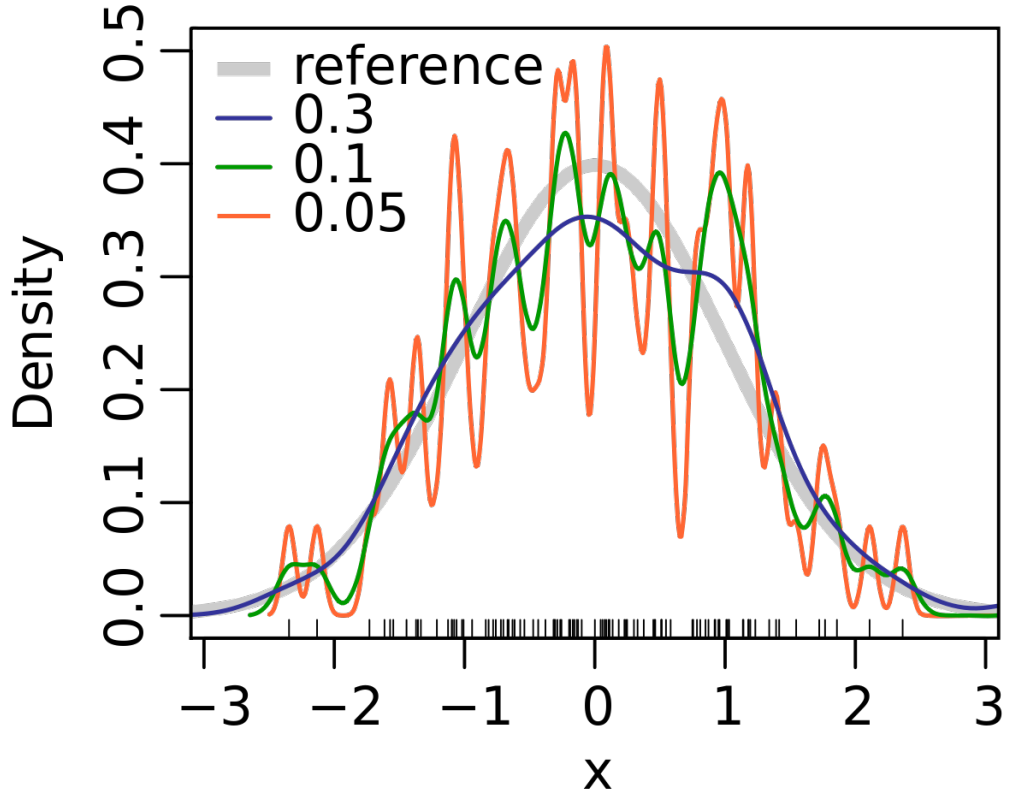
$$\hat{f}(x) = \frac{1}{nh} * \sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right)$$

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

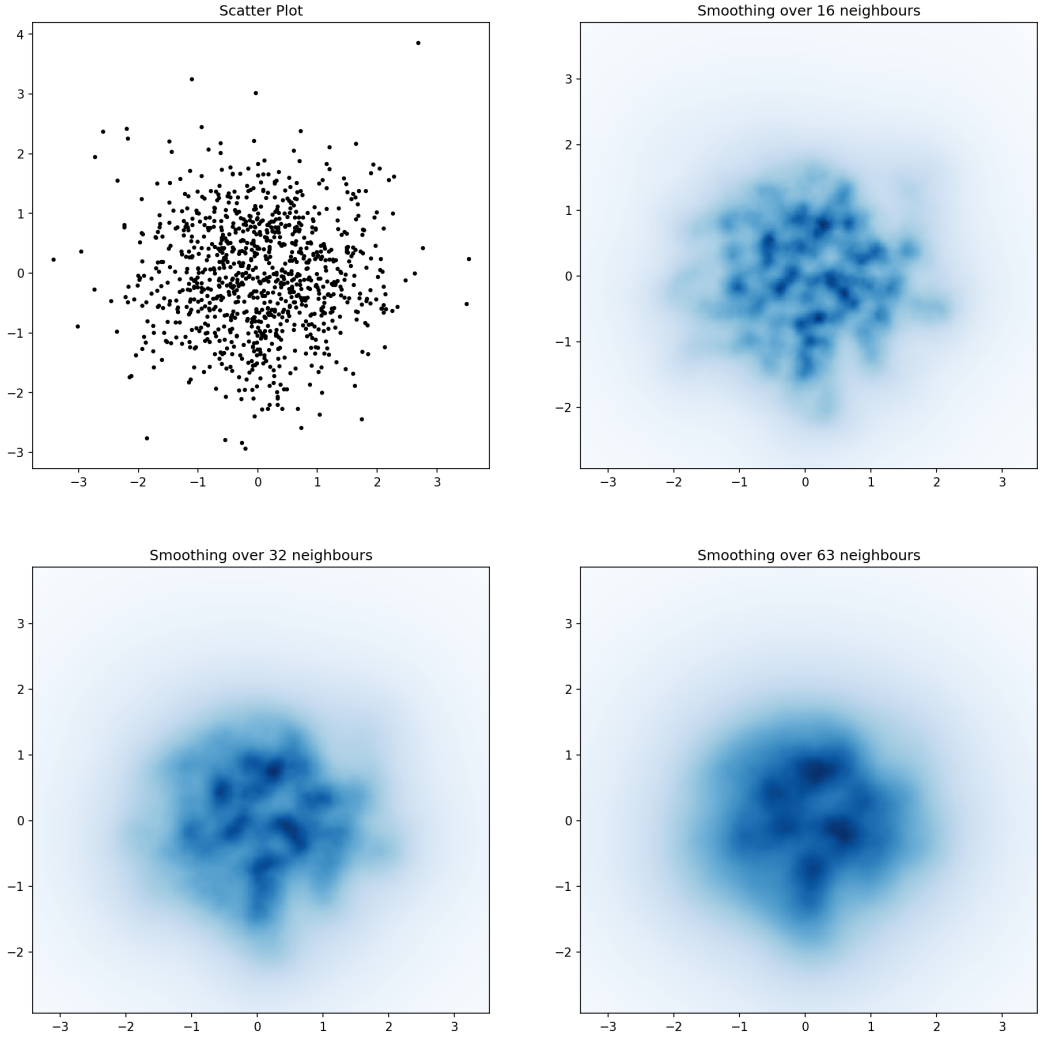
$$K(x) = \frac{1}{\sqrt{2\pi}} e^{\frac{-x^2}{2}}$$

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.

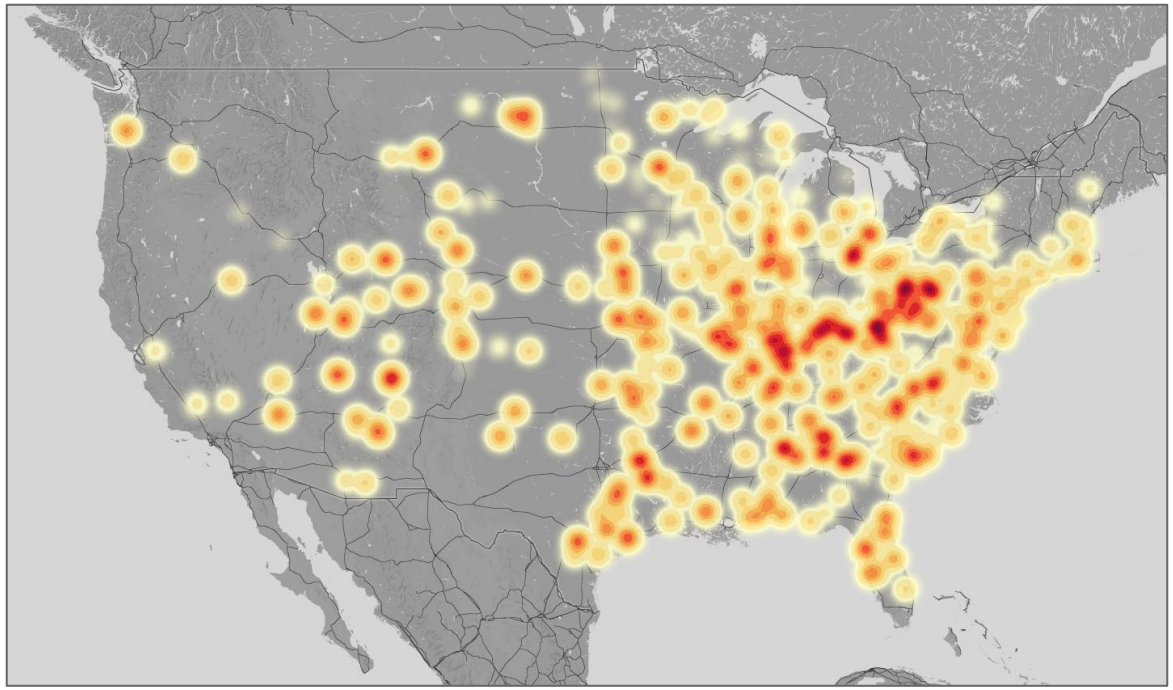


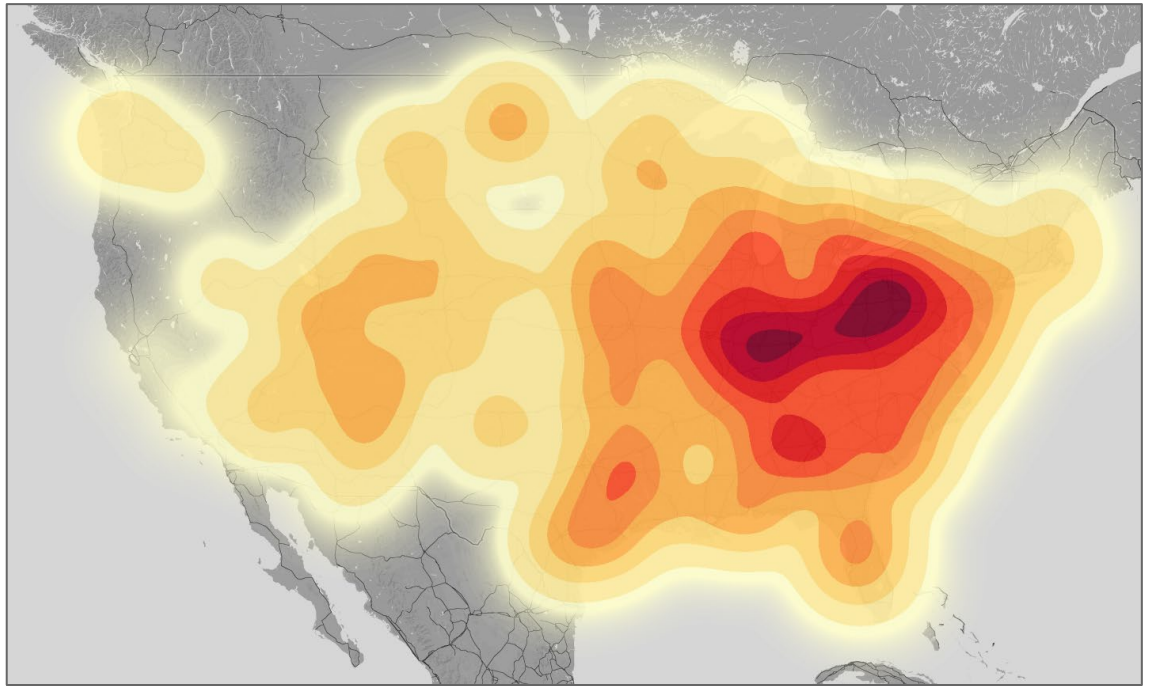
Kernel density estimation



[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

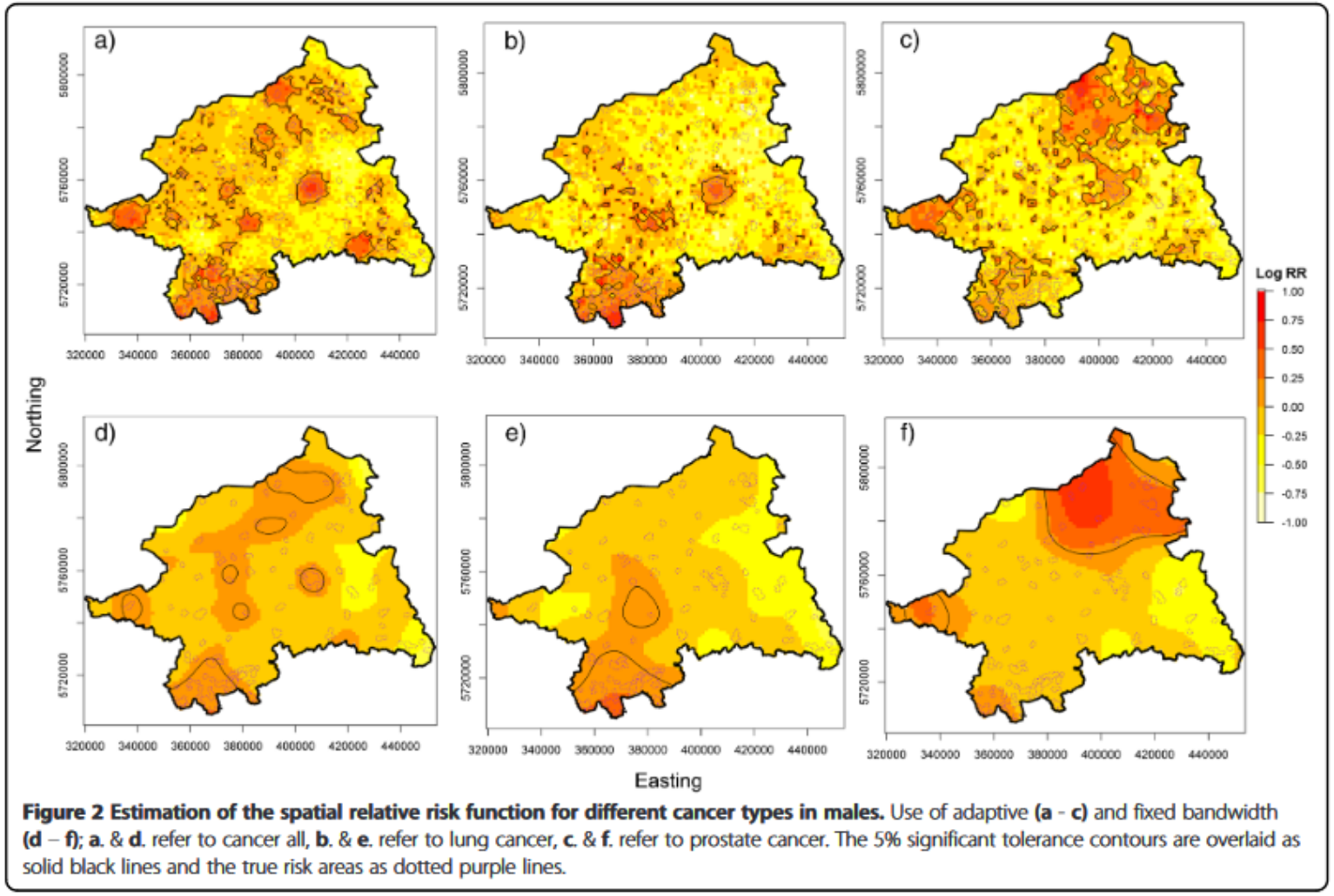


Figure 2 Estimation of the spatial relative risk function for different cancer types in males. Use of adaptive **(a - c)** and fixed bandwidth **(d – f)**; **a.** & **d.** refer to cancer all, **b.** & **e.** refer to lung cancer, **c.** & **f.** refer to prostate cancer. The 5% significant tolerance 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]