https://queue.acm.org/detail.cfm?id=1805128

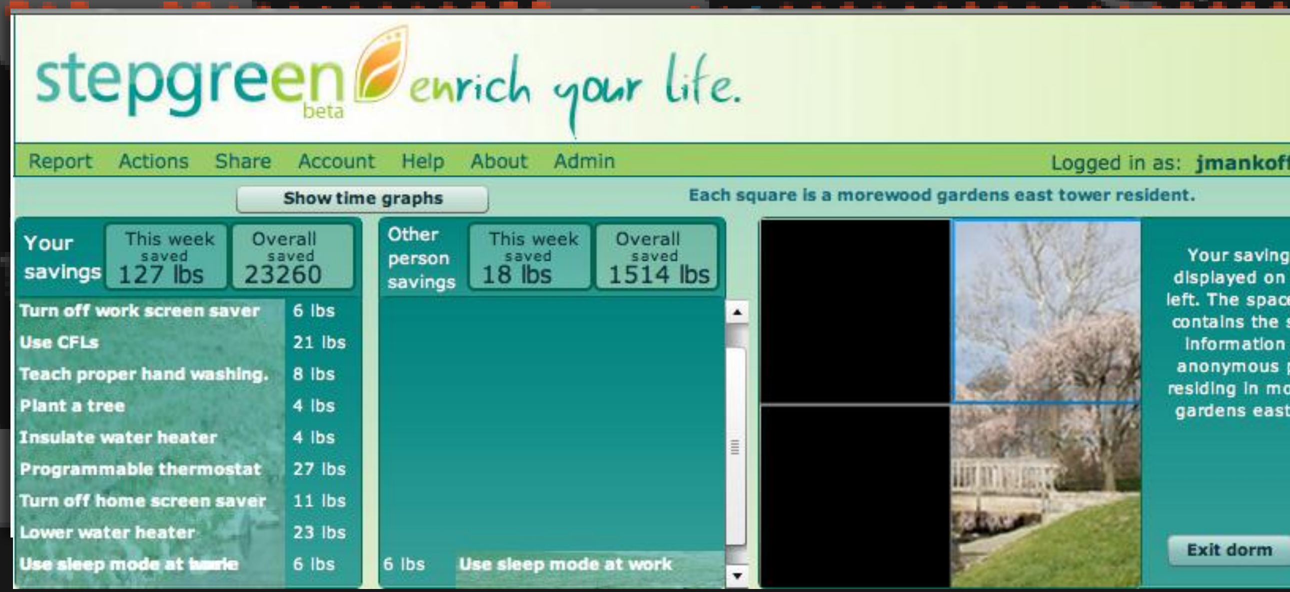
Challenges in Making the Hidden Visible

Based on material in 05-389 Interactive Data Science data.cmubi.org for more info

t/th 9-10:30, Spring

Jennifer Mankoff and many other collaborators

Information can change how we act in the world *[e.g.,* CHI'09, ICWSM'09]



Information For and About People

- As a prosthetic, changes what we can do in the world
- As a motivator, changes how we (or our machines) behave
- Shared with others, may
  - Change the balance of power [Ubicomp '09, '10]
  - Build new kinds of action and knowledge
- Beyond individuals, may support policy, politics, economics [DIS In Submission]

Making Data Actionable

Data

- Collecting and Interpreting

Information

- Sense

Knowledge

- Visualize
  Act
  Adapt

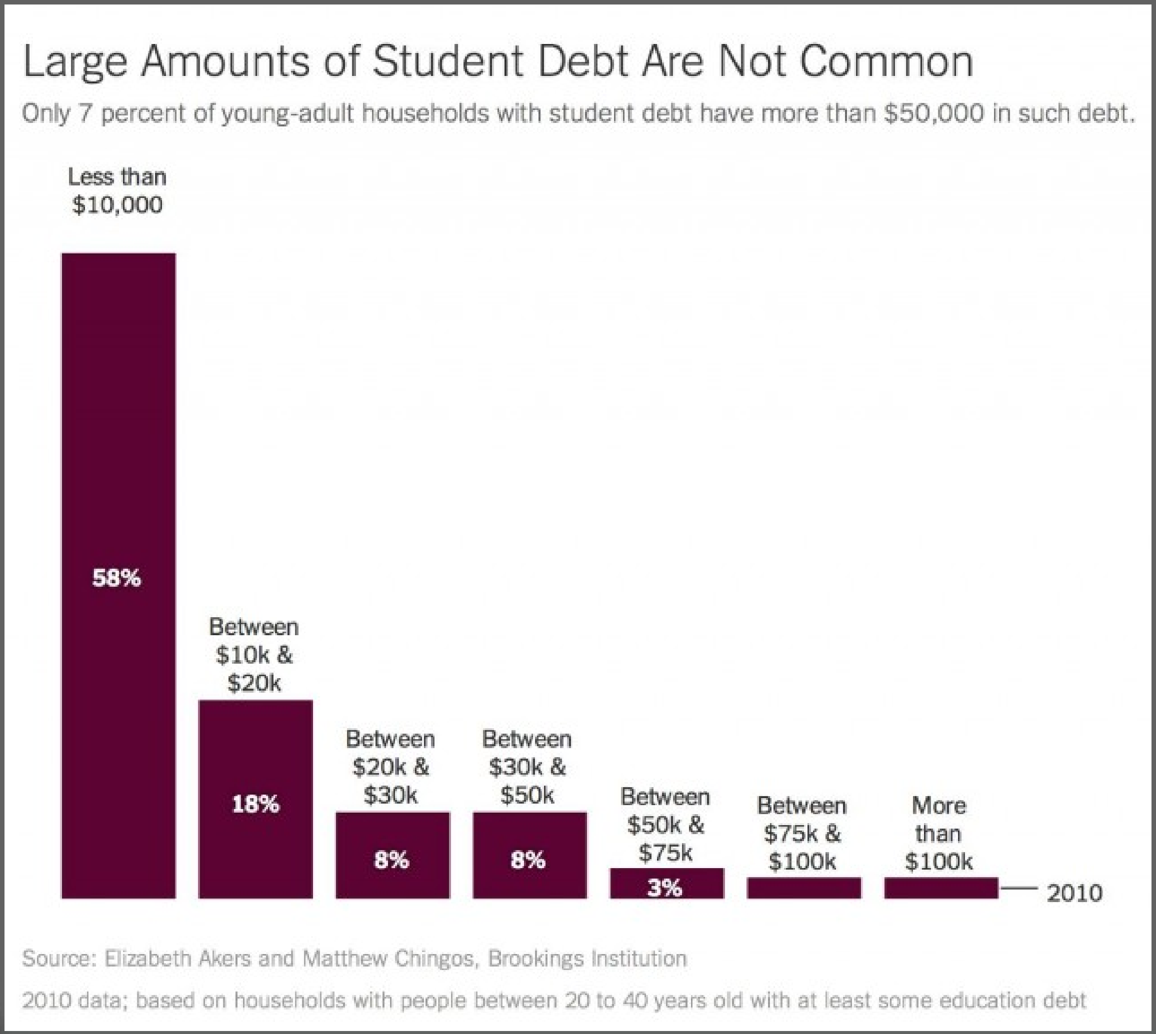
Value

Collecting the right data

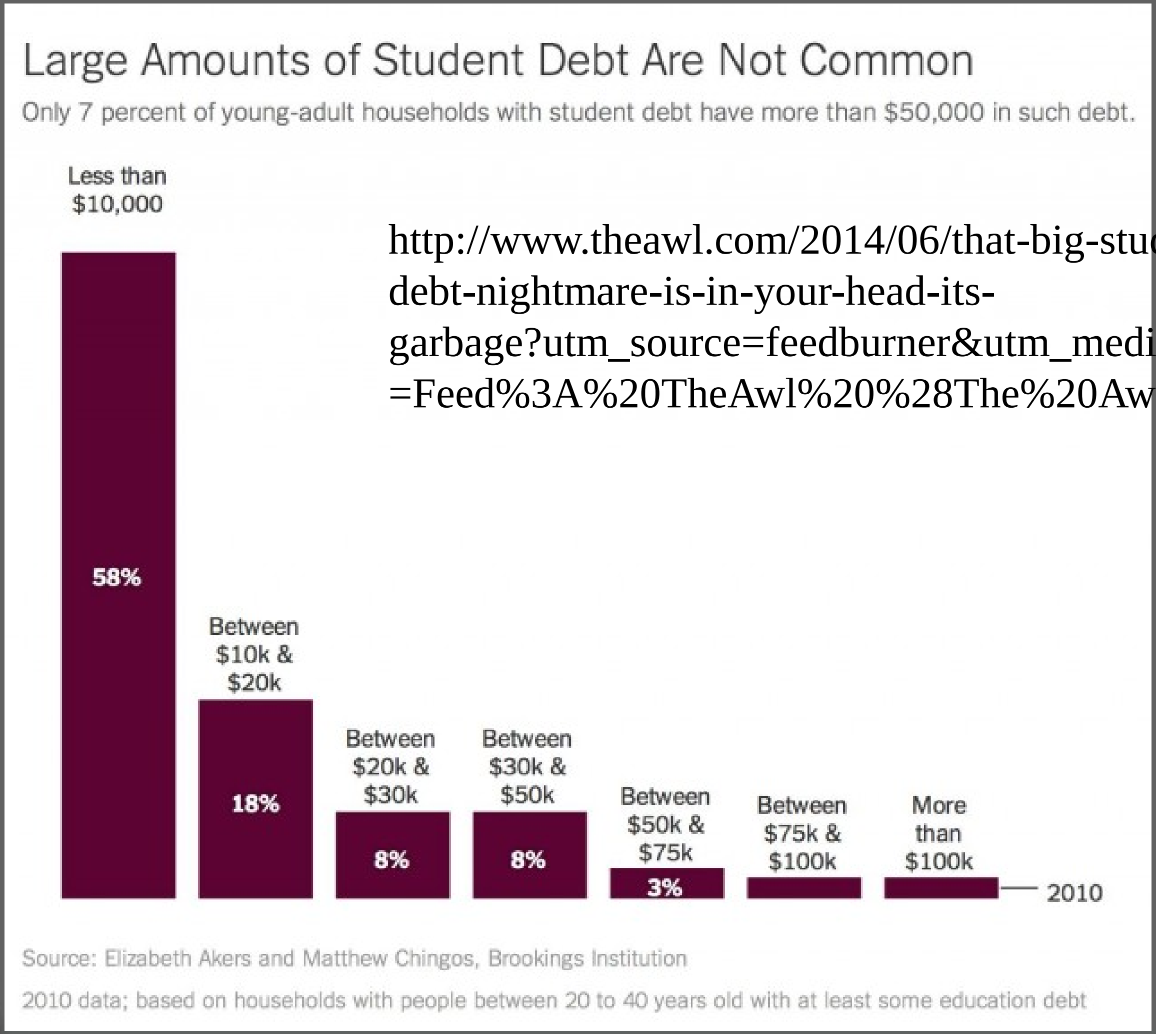
What is the problem? What data will solve the problem? How can we get that data?

- Techniques needed: Careful analysis & thought
- Tools: simulation & prototyping

Can you trust your data?



Can you trust your data?



http://www.theawl.com/2014/06/that-big-study-about-how-the-student-debt-nightmare-is-in-your-head-its-garbage?utm_source=feedburner&utm_medium=feed&utm_campaign=Feed%3A%20TheAwl%20%28The%20Awl%29

Can you trust your data?

Just because you have a lot of data, does not mean that it is *good* data

The plural of “anecdote” is not “data”.

Content borrowed from Bill Thies

Sources of Error

Sampling errors

- *Random Errors* due to the sample forming only part of the population
- *Systematic Bias* in sampling

Bias During Data Collection

- Demand Characteristics
- Illusory Superiority

Data entry / processing errors

- Data is generated accurately but errors introduced during recording or processing

What Makes a Good Sample?

Representative of the population

(Along dimensions that matter to the question being asked)

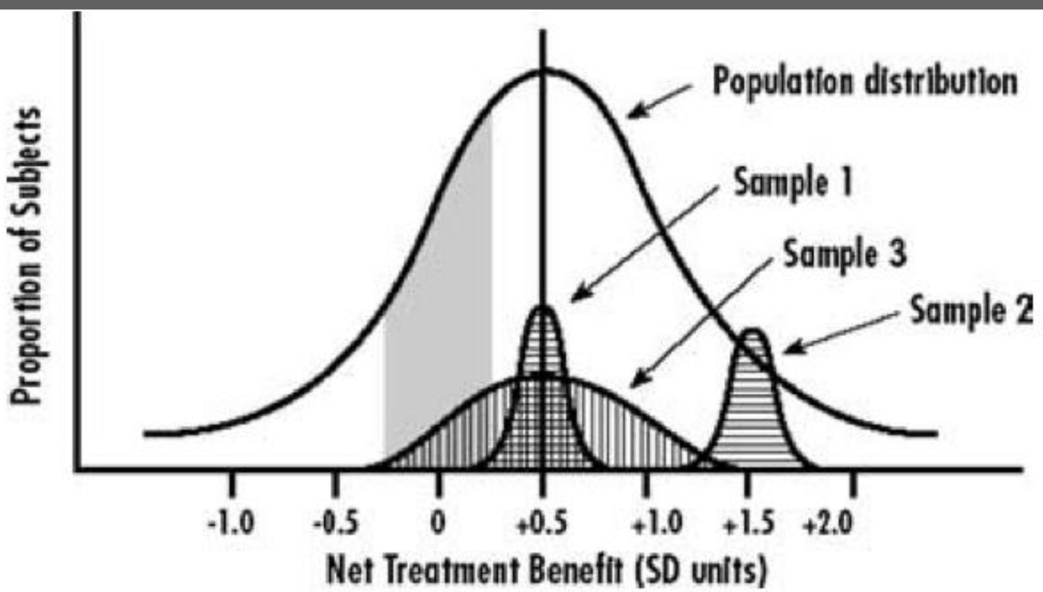


Figure 1: Kravitz *et al*, (2004) Evidence-based medicine, heterogeneity of treatment effects, and the trouble with averages. *The Milbank Quarterly* **82**(4):661-687

Sources of Error

Sampling errors

- *Random Errors* due to the sample forming only part of the population
- *Systematic Bias* in sampling

Bias During Data Collection

- Demand Characteristics
- Illusory Superiority

Data entry / processing errors

- Data is generated accurately but errors introduced during recording or processing

Example: Demand Characteristics

An experimental artifact where participants form an interpretation of the experiment's purpose and unconsciously change their behavior to fit that interpretation

Content borrowed from Bill Thies

What can we do to minimize demand characteristics in HCI?

Be aware that response bias affects all studies

Dissociate from a particular design or solution

Minimize the differences in social status between investigators and participants

Use triangulation to validate data collected

Tricks for asking sensitive questions

Measurement Errors

Badly Designed Questions

Badly Chosen Sensors

Bad Administration of Measurement Instrument

Inaccurate Measurements

See http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1323316/

Sources of Error

- Sampling errors
  - *Random Errors* due to the sample forming only part of the population
  - *Systematic Bias* in sampling
- Bias During Data Collection
  - Demand Characteristics
  - Illusory Superiority
- **Data entry / processing errors**
  - Data is generated accurately but errors introduced during recording or processing

Four C's of Data Quality

Is your data *Complete*?

Is your data *Coherent*?

Is your data *Correct?*

Is your data *aCcountable?*

Questions about Completeness

Appropriate Data: Does the data you have match the questions you want to answer?

Missing Data: Data does not exist because it was never obtained or was lost

Reporting error: The sensor (or respondent) is incorrect

Is your Data Coherent?

Does the data "add up"?

Does it make sense relative to itself?

Are the extreme values?

Examples

- Non number in a numeric field
- Month field has something other than a month
- Email has no @
- Hourly data adds up to 24 hrs per day
- *Etc.*

Is your data Correct?

Itemize aspects of your data that are easy to verify

Compare (collect twice or find alt. source)

Analyze the data collection strategy and look for sources of bias

- Within the population
- across variables (surveys with only round values; people who report everything in round numbers)

Determine how much is bad

Making Data Actionable

Data
- Collecting and Interpreting

Information
- Sensemaking

Knowledge
- Visualize
  Act
  Adapt

Value

Understanding Humans

Activities

Routines

Intent

Causality

Understanding Humans

Activities

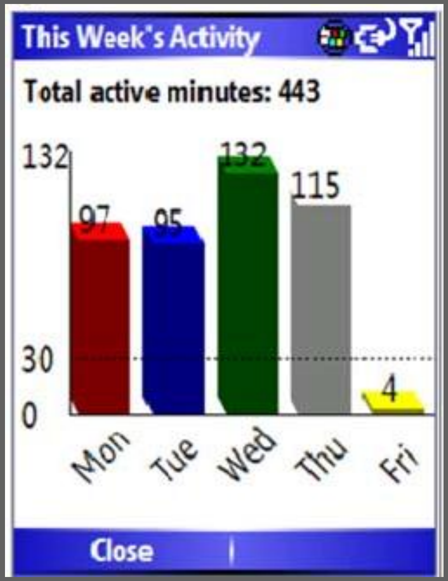
Routines

Intent

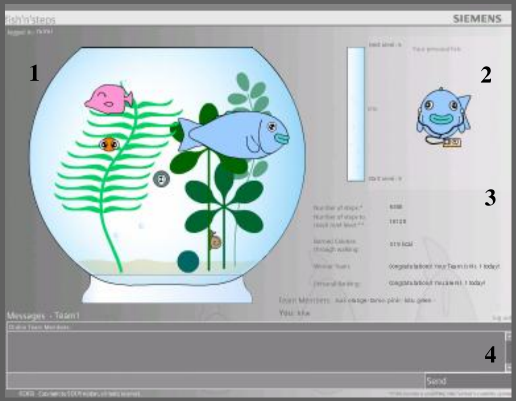
Causality

Visualizing Activities

UbiFit
Consolvo *et al.* '08



Shakra
Maitland *et al.* '06



Fish 'n Steps
Lin *et al.* '06

Stepgreen/UbiGreen
Mankoff et al. '09
Froelich et al. '09

Understanding Humans

Activities

Routines

Intent

Causality

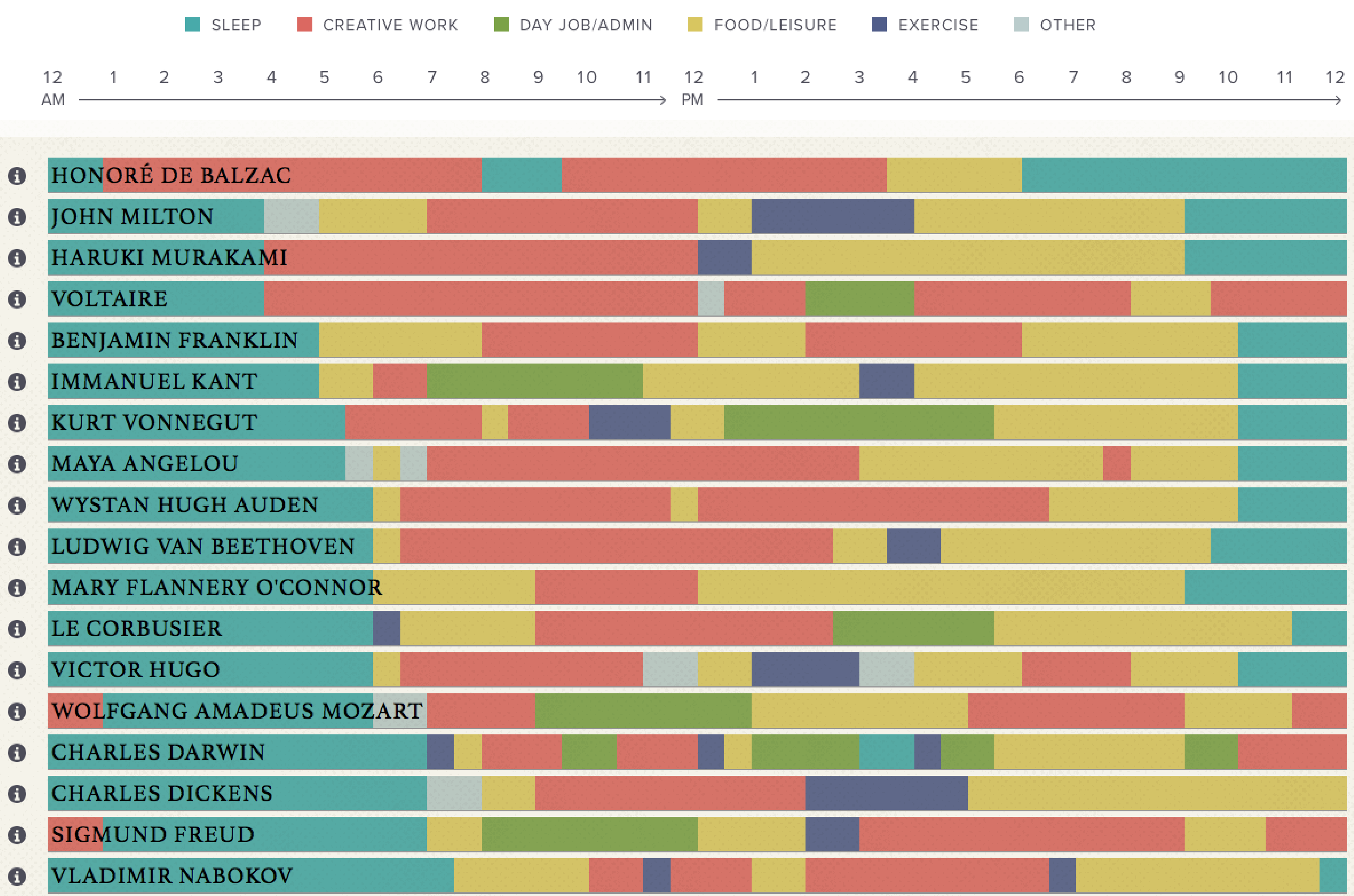
Why are routines important?

Develop routines to reduce cognitive effort

Deviations and anomalies cause stress and extra effort

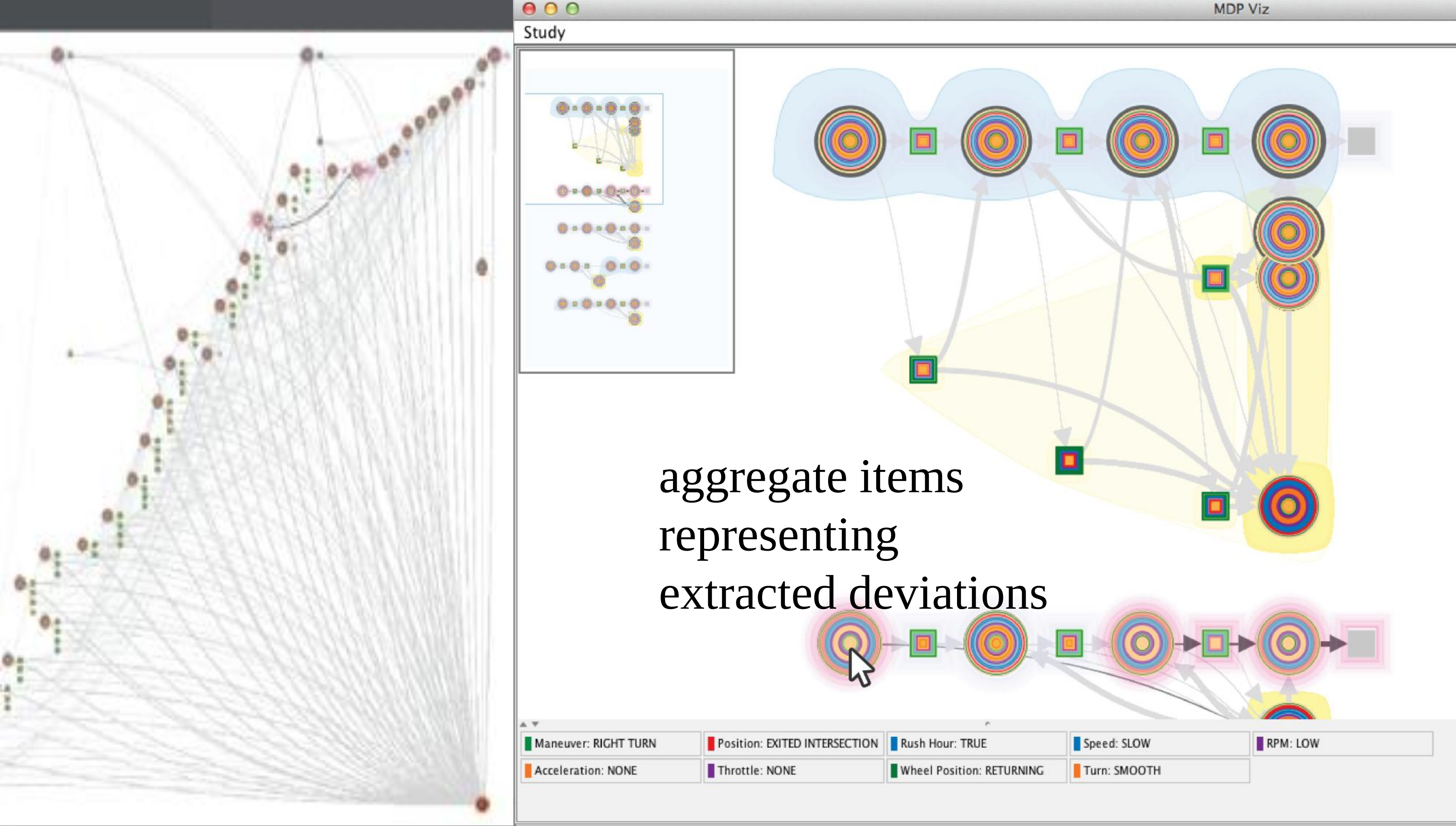
→ *at least as* actionable as inferred activities

- Longer term, more built-in
- Opportunities for change
- Barriers for change
- Leverage point for understanding human intent
- Untapped as a resource: sensing and using



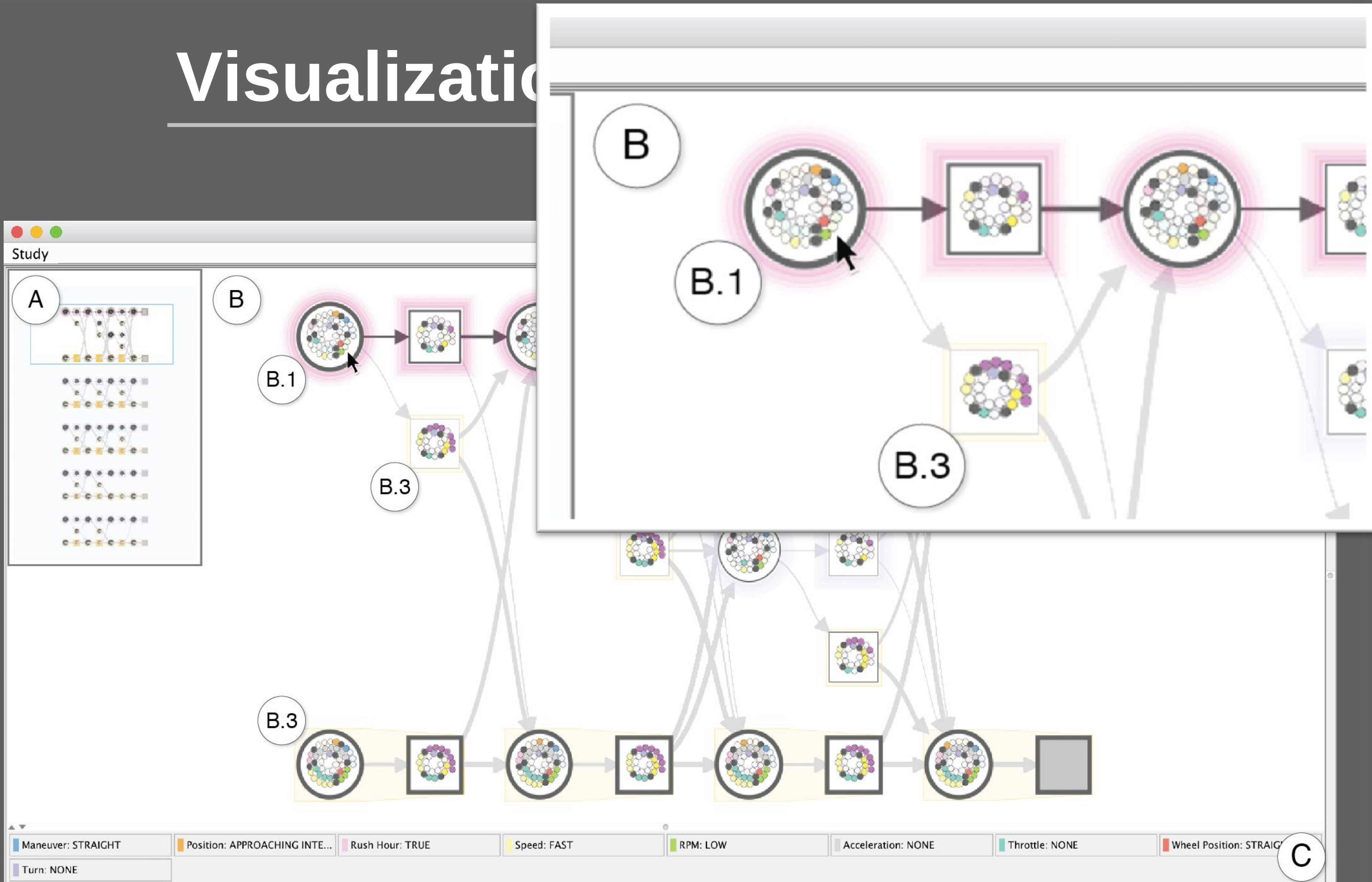
Jennifer Mankoff, 1/12

automatically extract *patterns* of human behavior that form routines and *deviations* from demonstrated behavior



aggregate items
representing
extracted deviations

Visualizatio



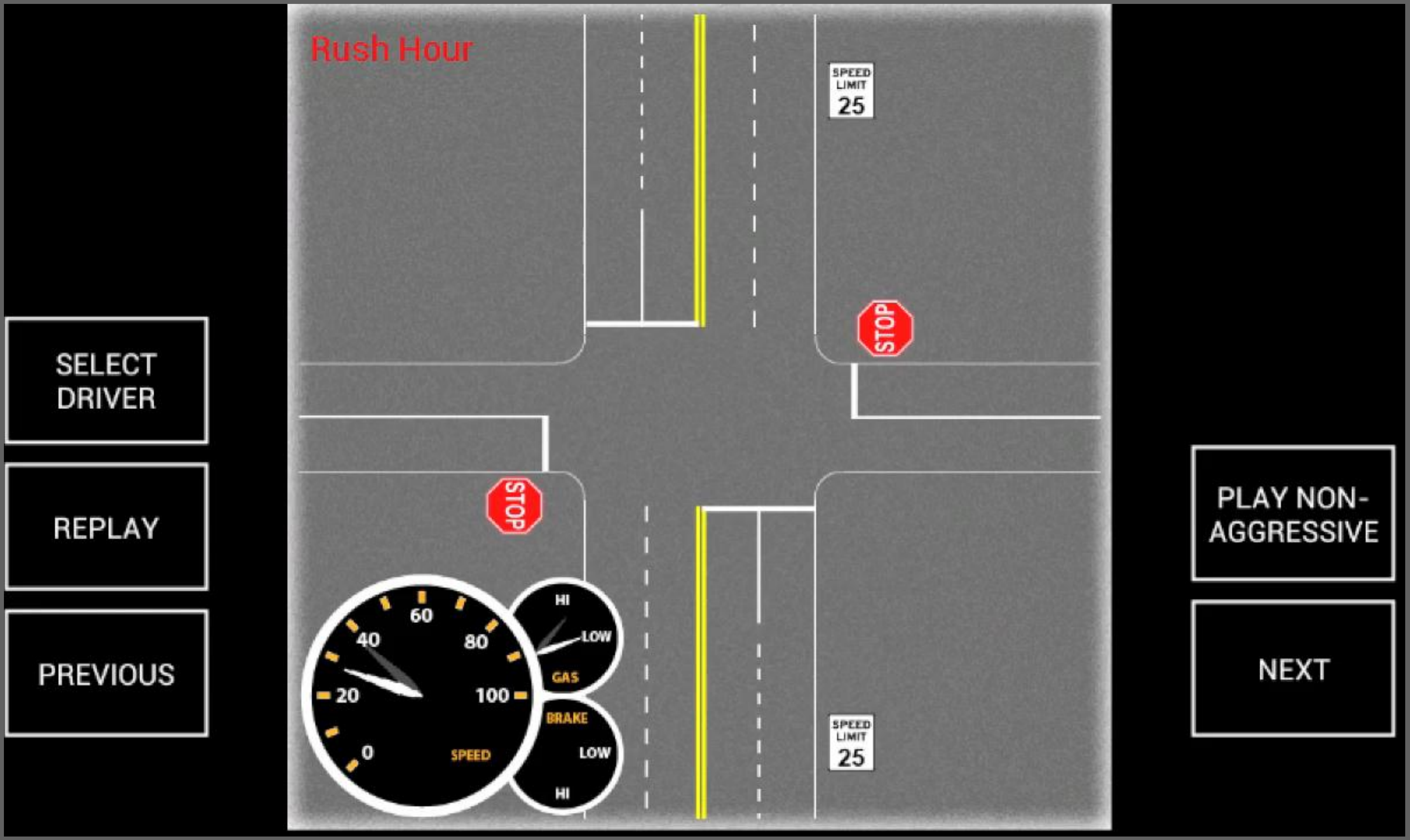
Now automated extraction (CHI 2017?)

Aggressive (Novice) Drivers
(CHI 2014, CHI 2016, CHI 2017 submission)

US: 1500 deaths/year

Cost of $40 billion from crashes

Aggressive (Novice) Drivers: Interventions



Understanding Humans

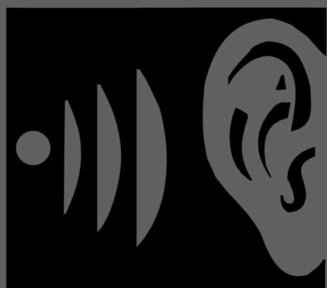
Activities

Routines

Intent

Causality

Causality



Hear noise

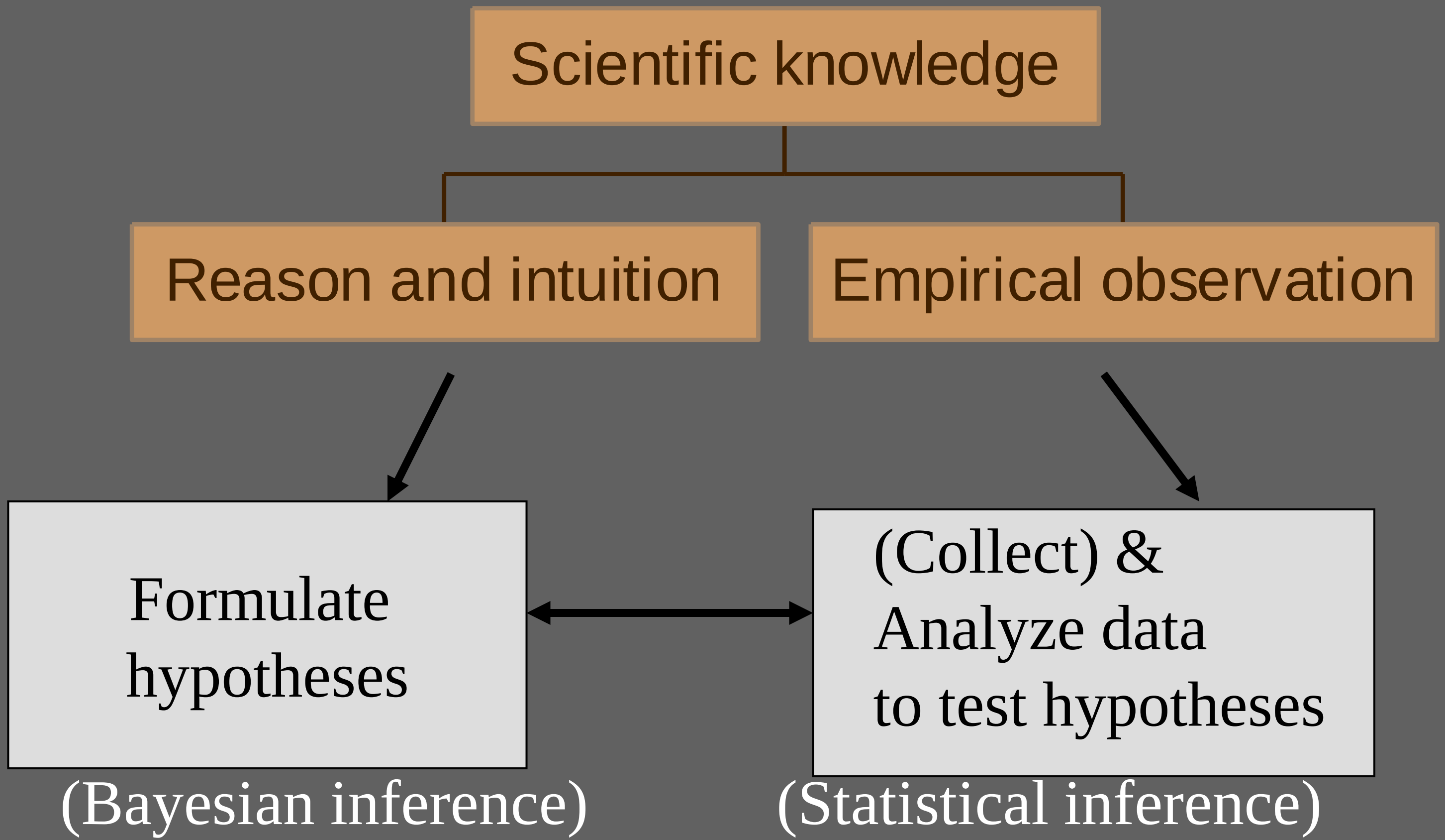
Past experience
leads to expectation

Go to window
to verify

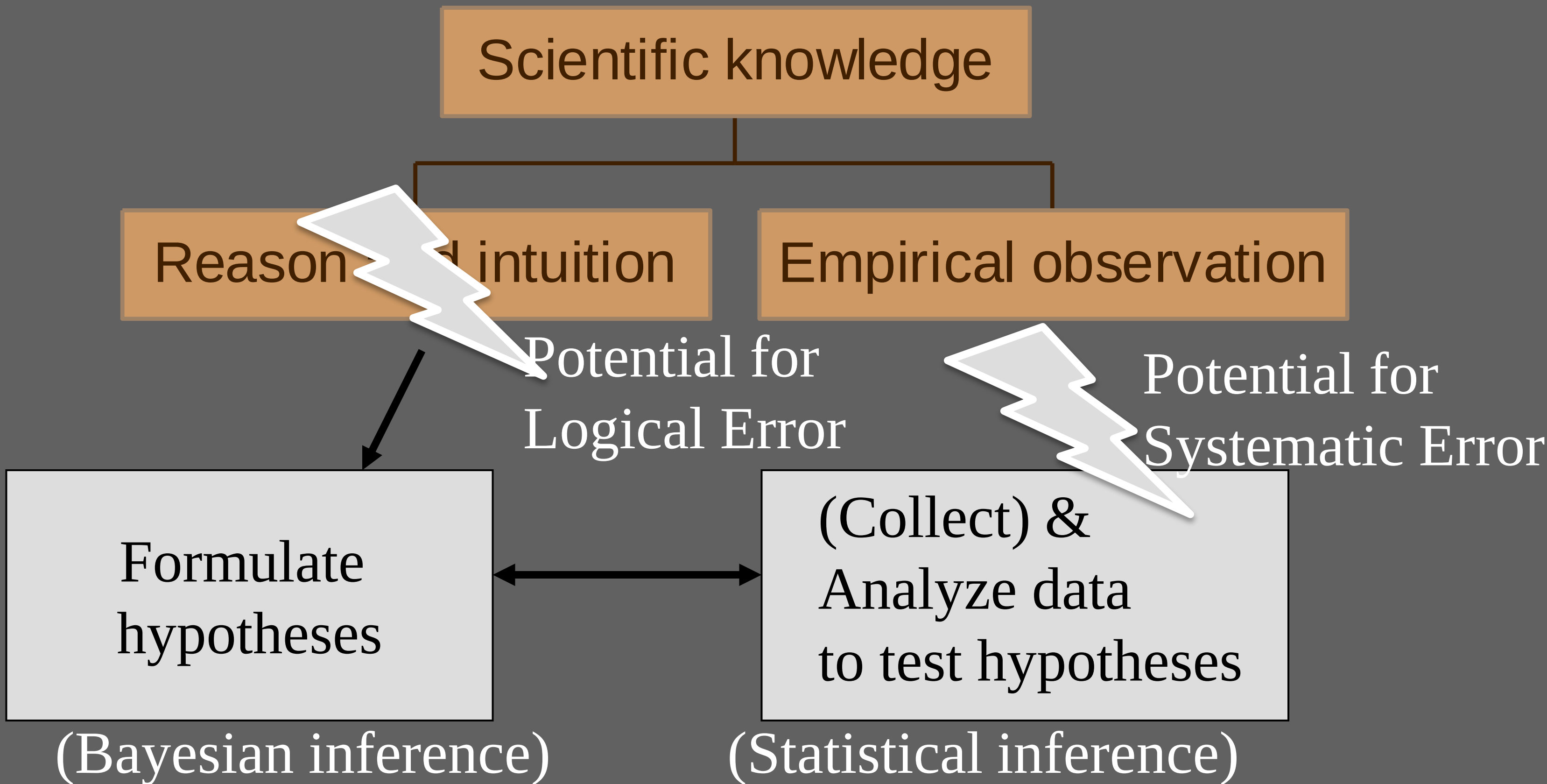
Prediction was false

Only with empirical evidence, can we make and test predictions

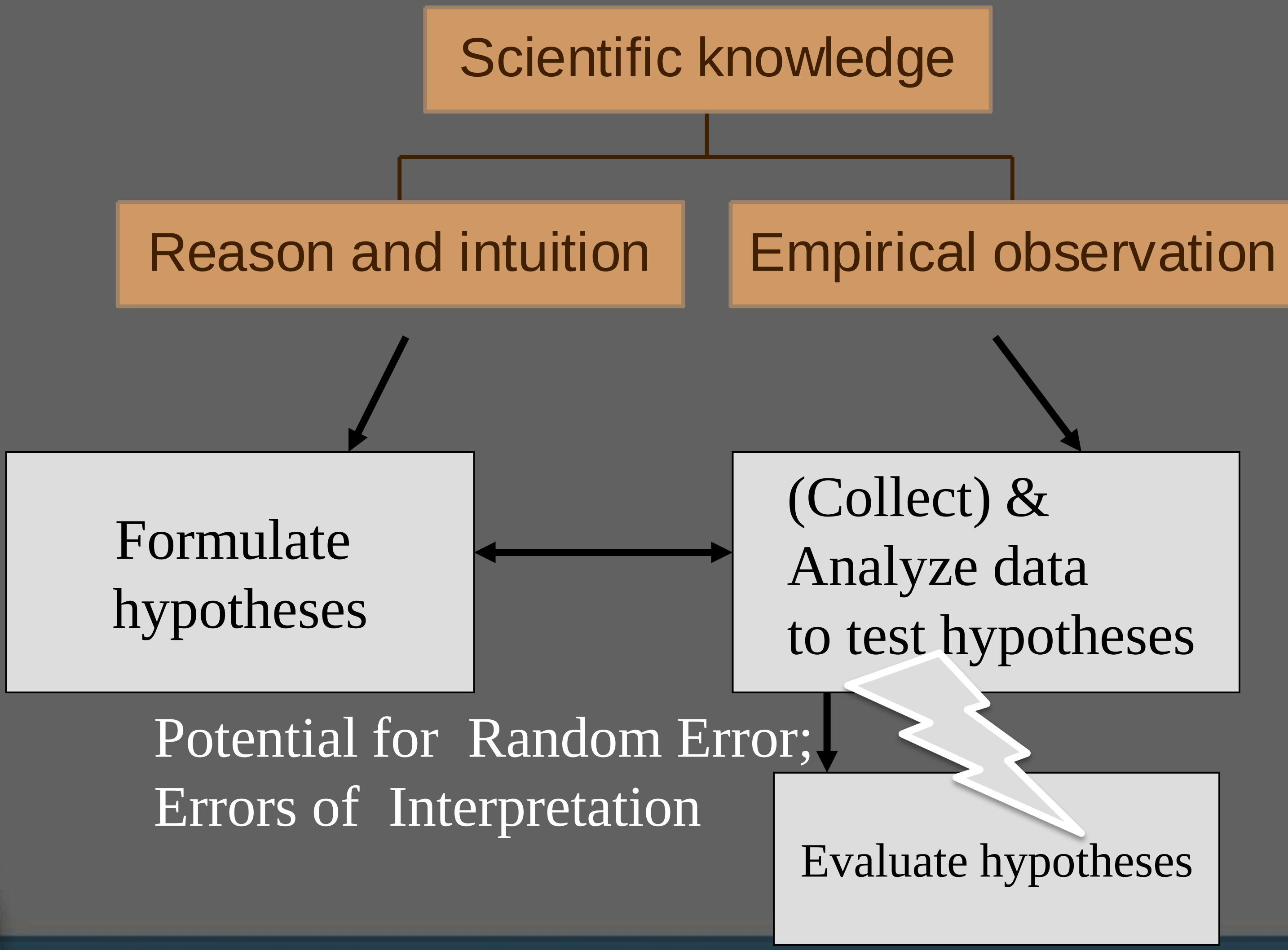
Process vs Frequency data



Process vs Frequency data



Process vs Frequency Data



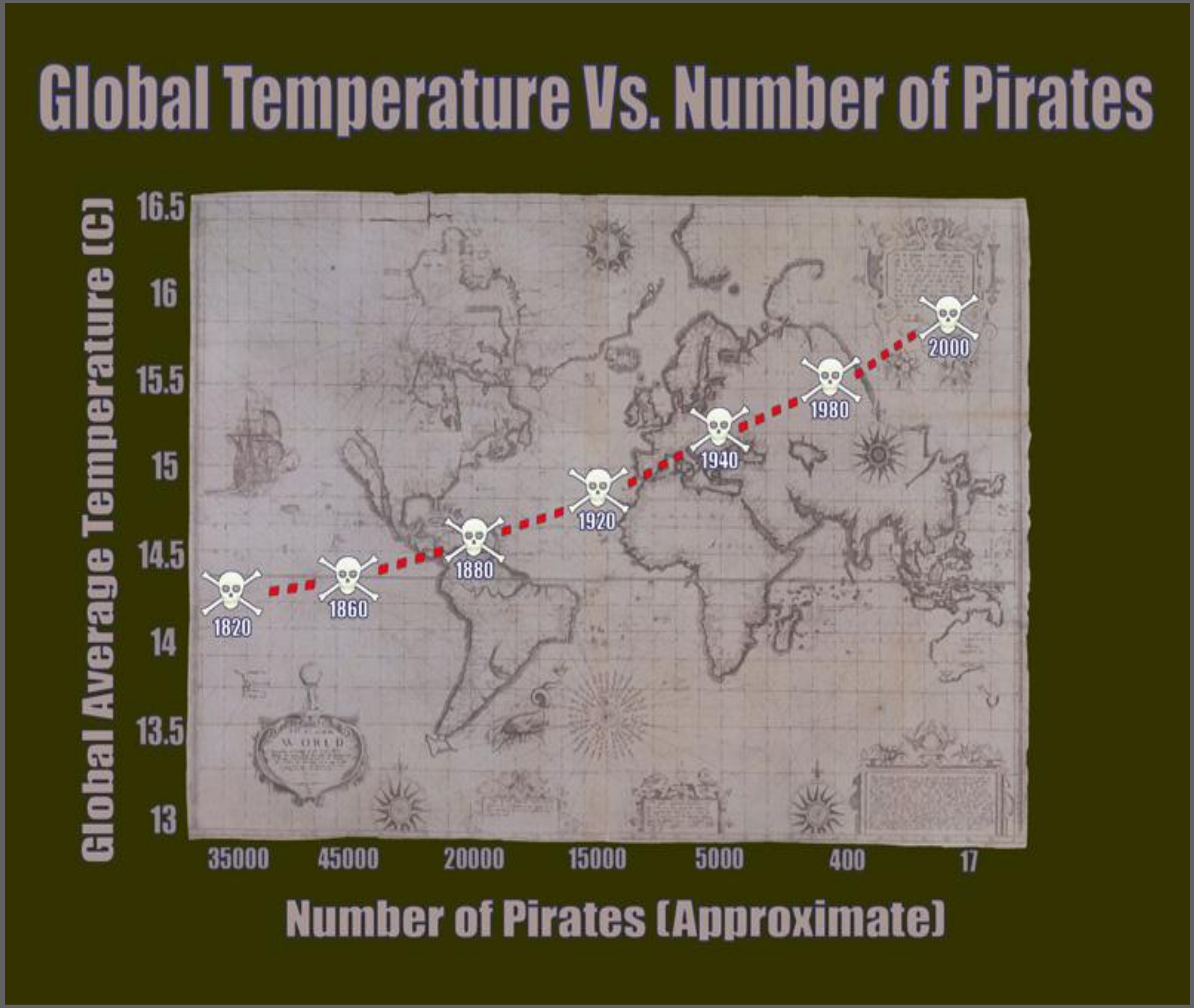
Correlation is not Causation

There is a 0.91 correlation between ice cream consumption and drowning deaths.

Does eating ice cream cause drowning?

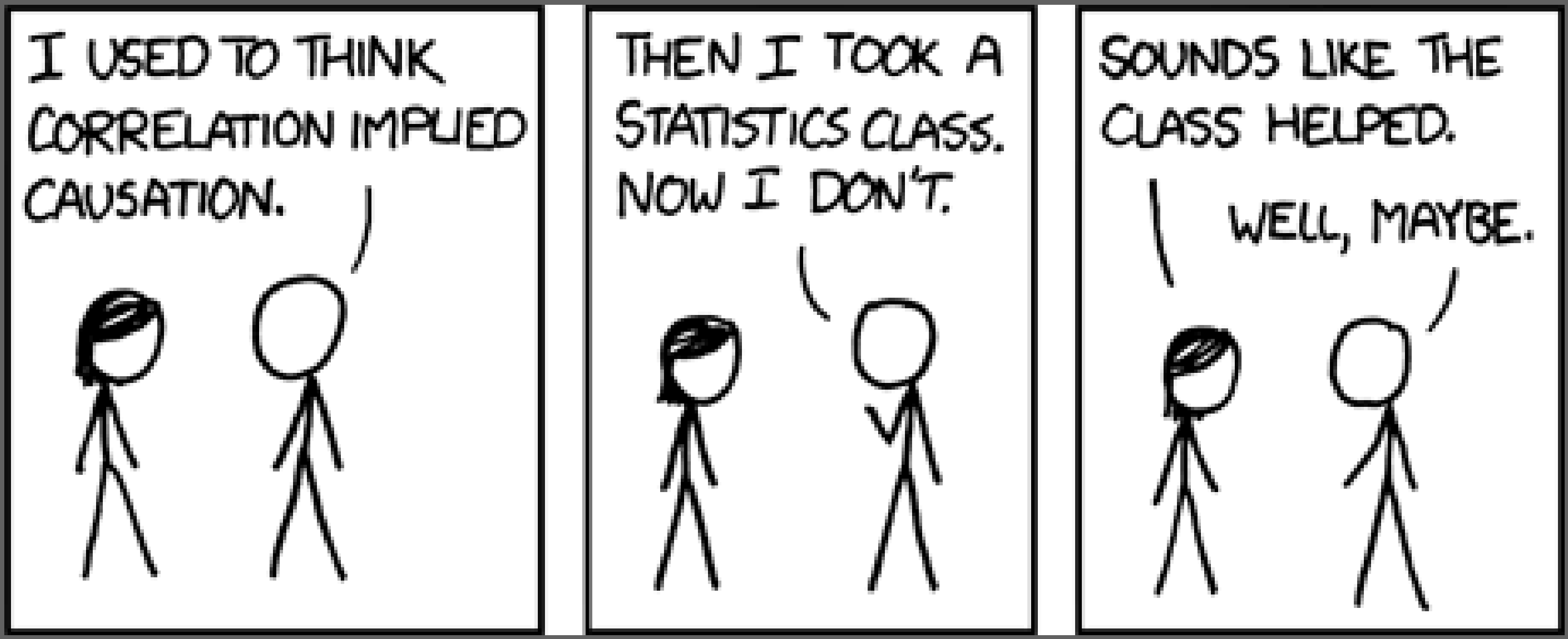
Does grief cause us to eat more ice cream?

Correlation without causation (1)



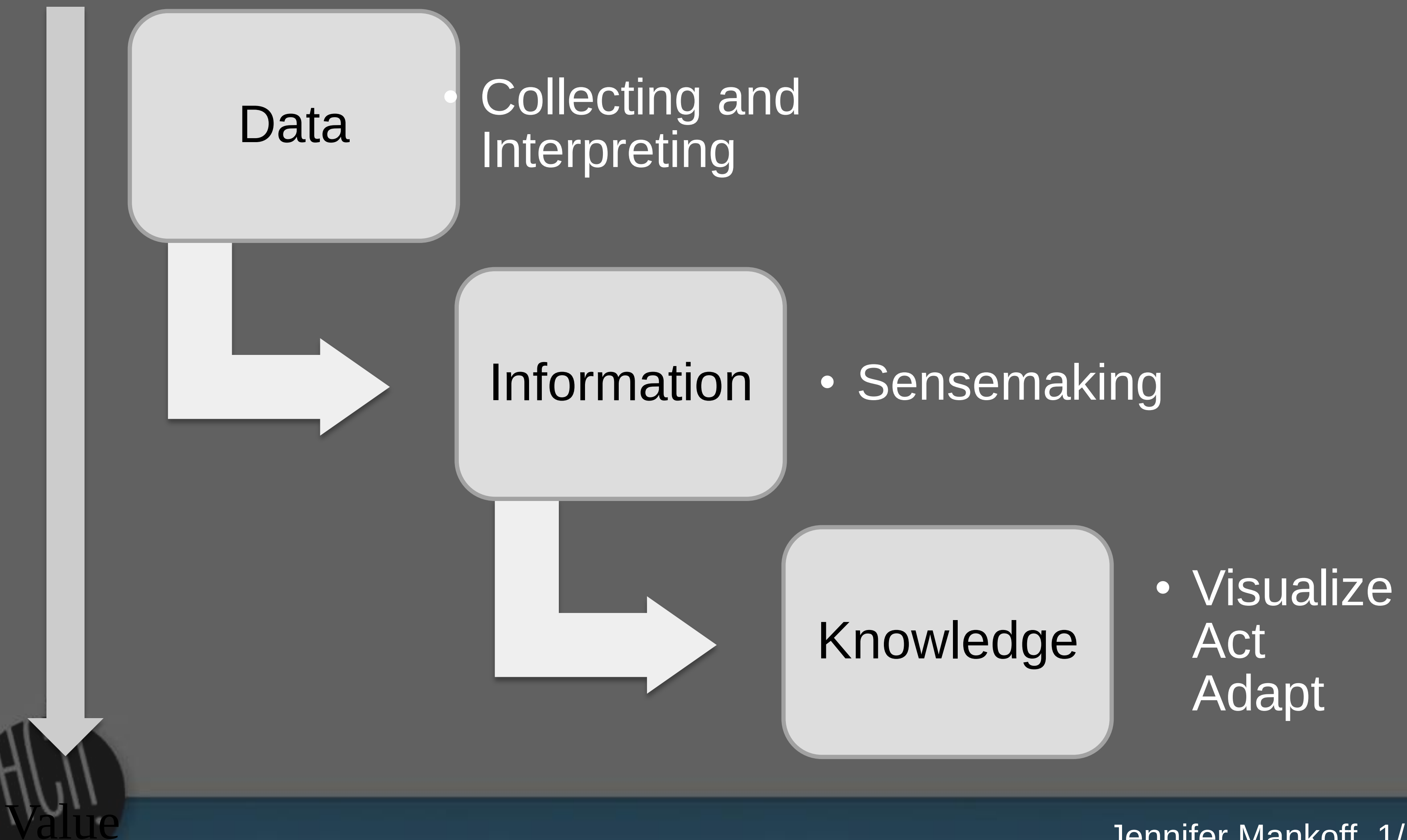
http://bama.ua.edu/~sprentic

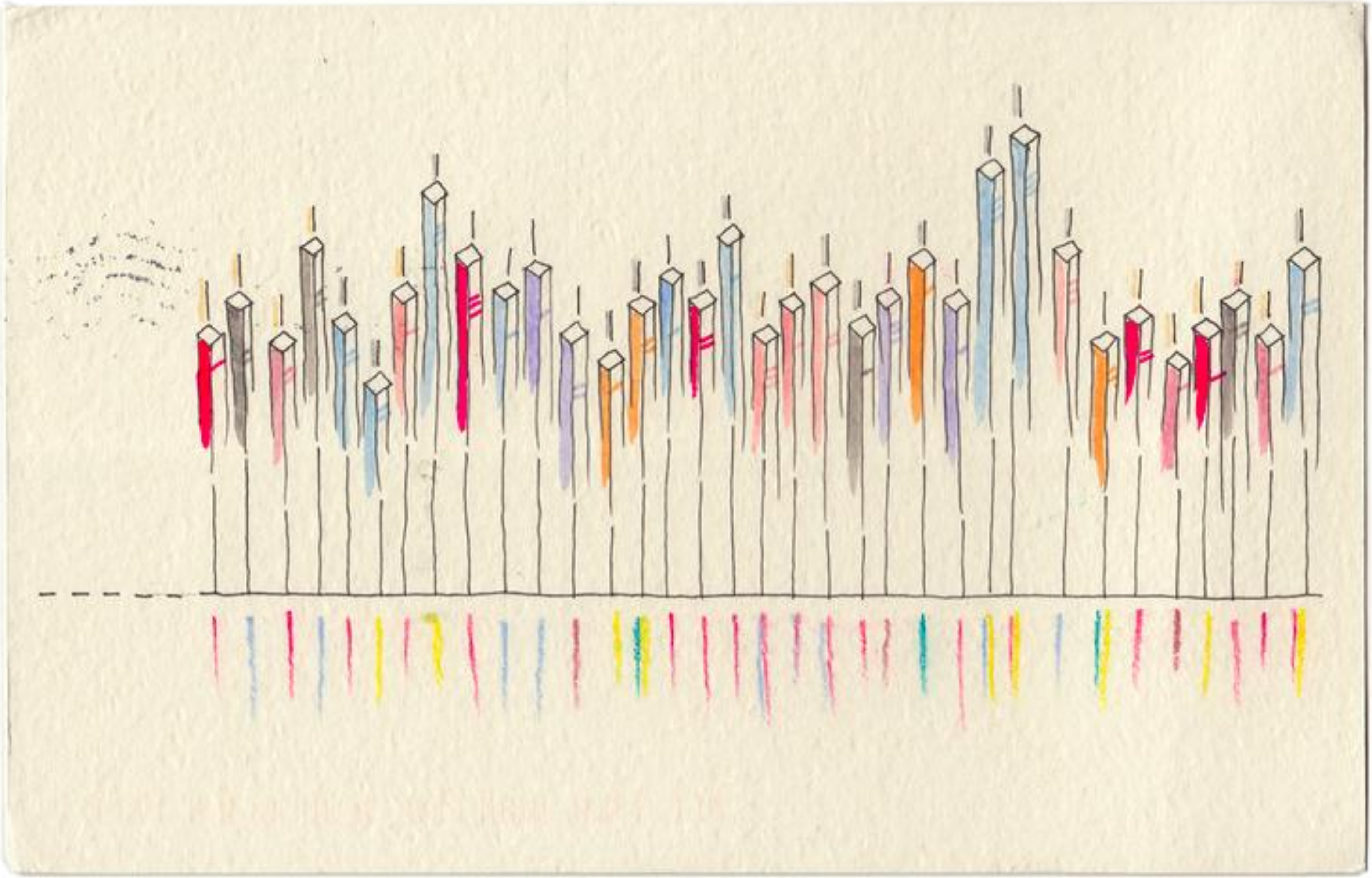
Correlation != Causality



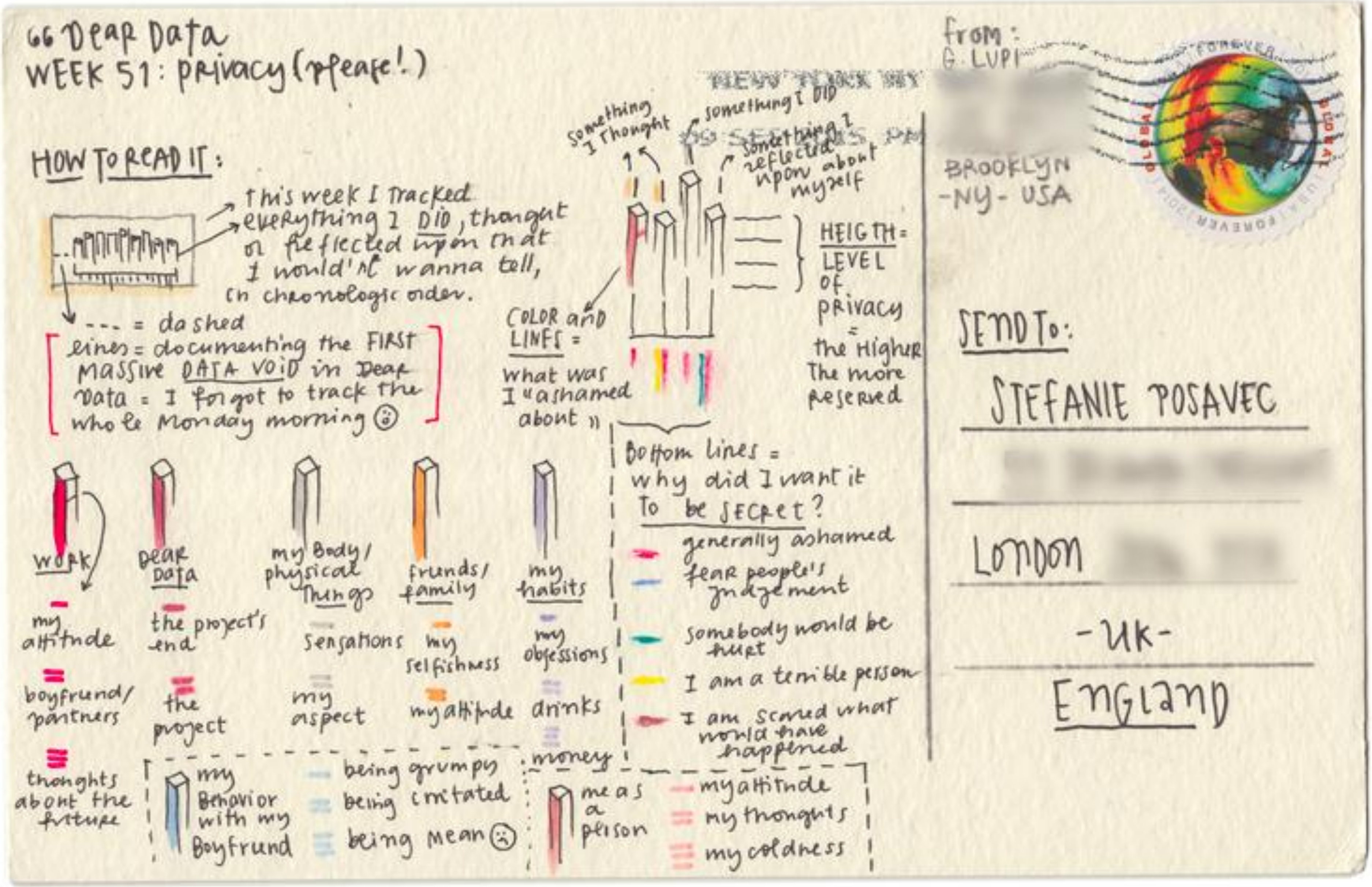
Adapted from http://www.fordham.edu/economics/Vinod/correl-regr.ppt

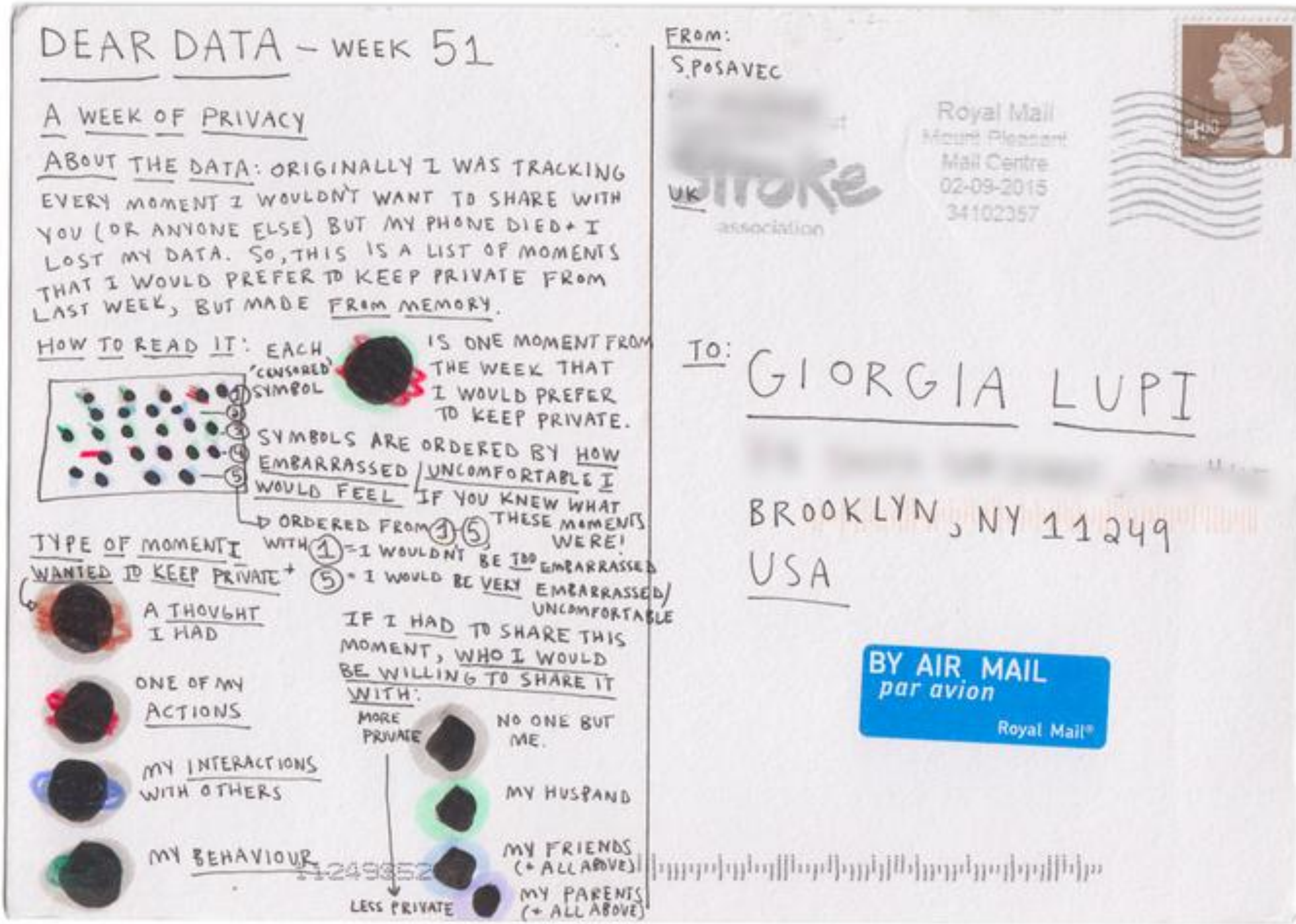
Making Data Actionable

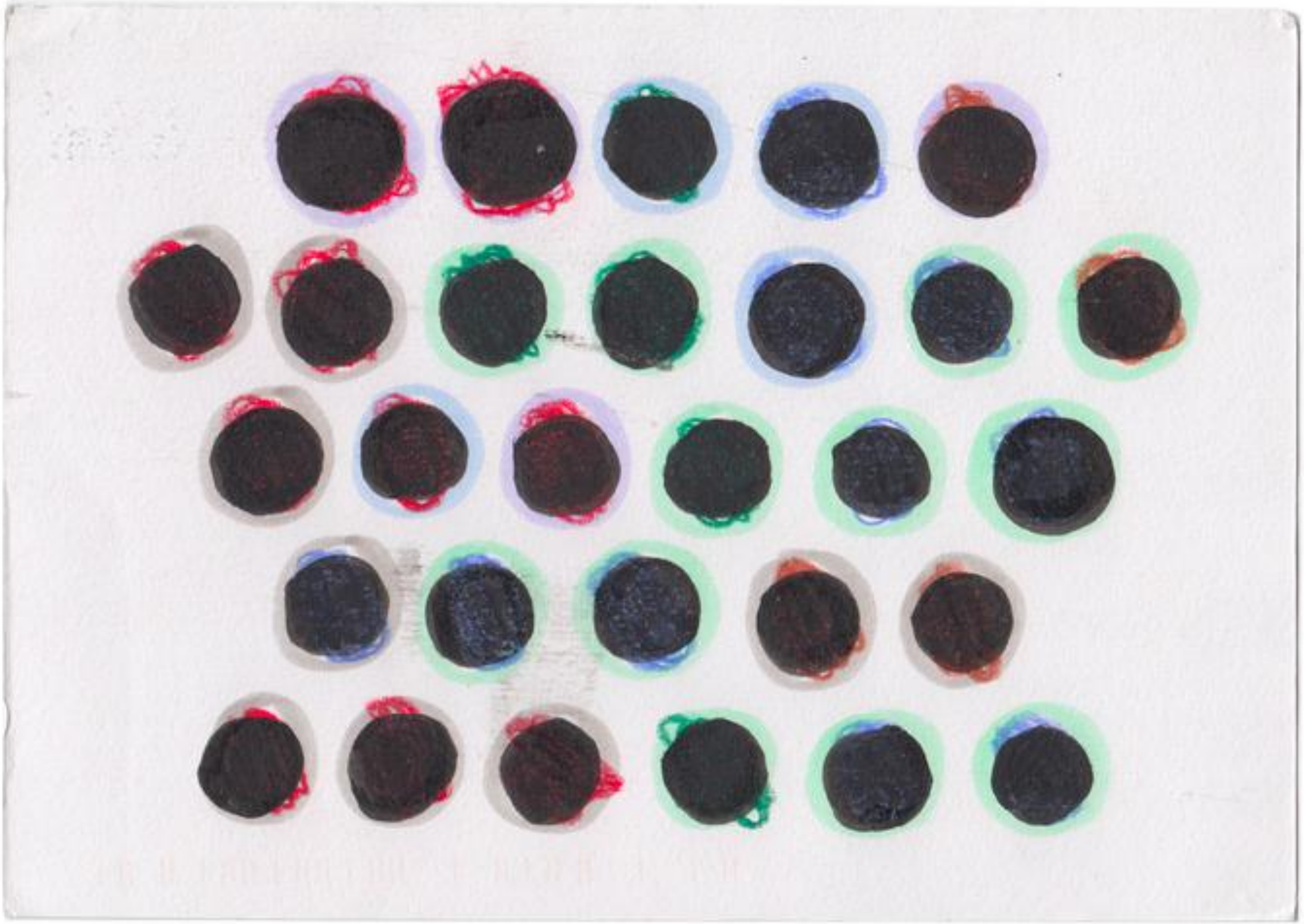




http://www.dear-data.com/all/


66 Dear Data
WEEK 51: privacy (please!)
HOW TO READ IT:
this week I tracked everything I DID, thought or reflected upon that I would'nt wanna tell, in chronologic order.
- - - = dashed lines = documenting the FIRST massive DATA VOID in Dear Data = I forgot to track the whole Monday morning
something I thought
something I DID
something I reflected upon about myself
COLOR and LINES = what was I "ashamed about"
HEIGTH = LEVEL of privacy = the higher the more reserved
Bottom lines = why did I want it to be SECRET?
generally ashamed
fear people's judgement
somebody would be hurt
I am a terrible person
I am scared what would have happened
work
Dear Data
my body/physical things
friends/family
my habits
my attitude
the project's end
sensations
my selfishness
my obsessions
boyfriend/partners
the project
my aspect
my attitude
drinks
money
thoughts about the future
my behavior with my boyfriend
being grumpy
being irritated
being mean
me as a person
my attitude
my thoughts
my coldness
from: G. LUPI
BROOKLYN -NY- USA
SEND TO:
STEFANIE POSAVEC
LONDON
-UK-
ENGLAND


DEAR DATA – WEEK 51
A WEEK OF PRIVACY
ABOUT THE DATA: ORIGINALLY I WAS TRACKING EVERY MOMENT I WOULDN'T WANT TO SHARE WITH YOU (OR ANYONE ELSE) BUT MY PHONE DIED + I LOST MY DATA. SO, THIS IS A LIST OF MOMENTS THAT I WOULD PREFER TO KEEP PRIVATE FROM LAST WEEK, BUT MADE FROM MEMORY.
HOW TO READ IT: EACH 'CENSORED' SYMBOL IS ONE MOMENT FROM THE WEEK THAT I WOULD PREFER TO KEEP PRIVATE.
1
2
3
4
5
SYMBOLS ARE ORDERED BY HOW EMBARRASSED / UNCOMFORTABLE I WOULD FEEL IF YOU KNEW WHAT THESE MOMENTS WERE!
ORDERED FROM 1-5 WITH 1 = I WOULDN'T BE TOO EMBARRASSED
5 = I WOULD BE VERY EMBARRASSED/UNCOMFORTABLE
TYPE OF MOMENT I WANTED TO KEEP PRIVATE:
A THOUGHT I HAD
ONE OF MY ACTIONS
MY INTERACTIONS WITH OTHERS
MY BEHAVIOUR
IF I HAD TO SHARE THIS MOMENT, WHO I WOULD BE WILLING TO SHARE IT WITH:
MORE PRIVATE
LESS PRIVATE
NO ONE BUT ME.
MY HUSBAND
MY FRIENDS (+ ALL ABOVE)
MY PARENTS (+ ALL ABOVE)
FROM:
S. POSAVEC
UK
Royal Mail
Mount Pleasant
Mail Centre
02-09-2015
34102357
TO:
GIORGIA LUPI
BROOKLYN, NY 11249
USA
BY AIR MAIL
par avion
Royal Mail®
11249352



What is Information Visualization?

Visualize: to form a mental image or vision of …

Visualize: to imagine or remember as if actually seeing.

American Heritage dictionary, Concise Oxford dictionary

The Power of Visualization

1. Start out going Southwest on ELLSWORTH AVE Towards BROADWAY by turning right.
2: Turn RIGHT onto BROADWAY.
3. Turn RIGHT onto QUINCY ST.
4. Turn LEFT onto CAMBRIDGE ST.
5. Turn SLIGHT RIGHT onto MASSACHUSETTS AVE.
6. Turn RIGHT onto RUSSELL ST.

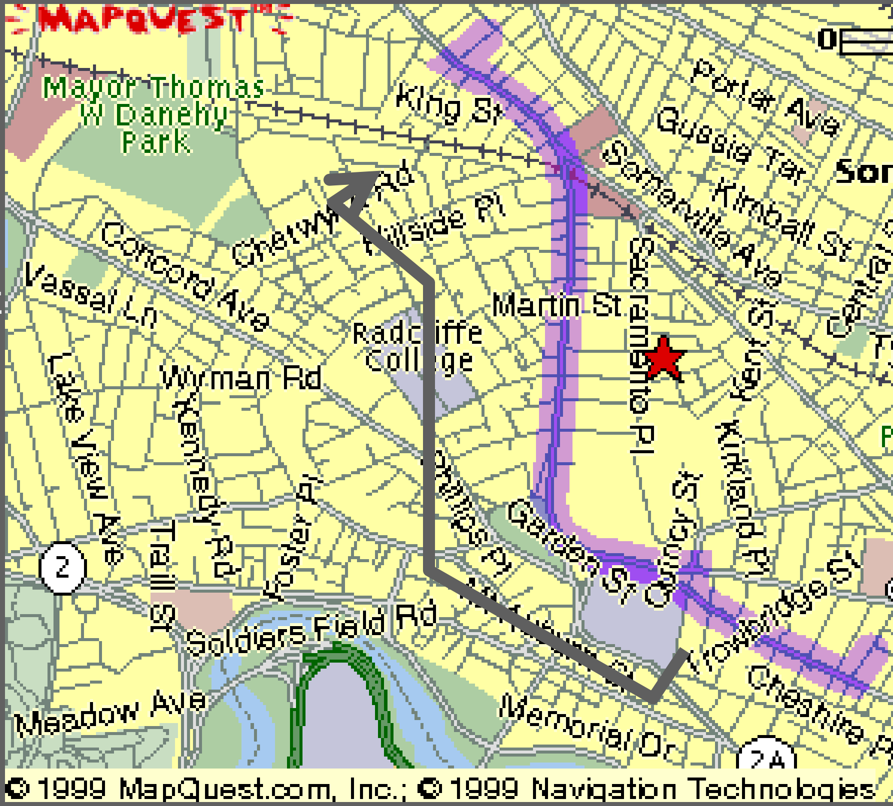
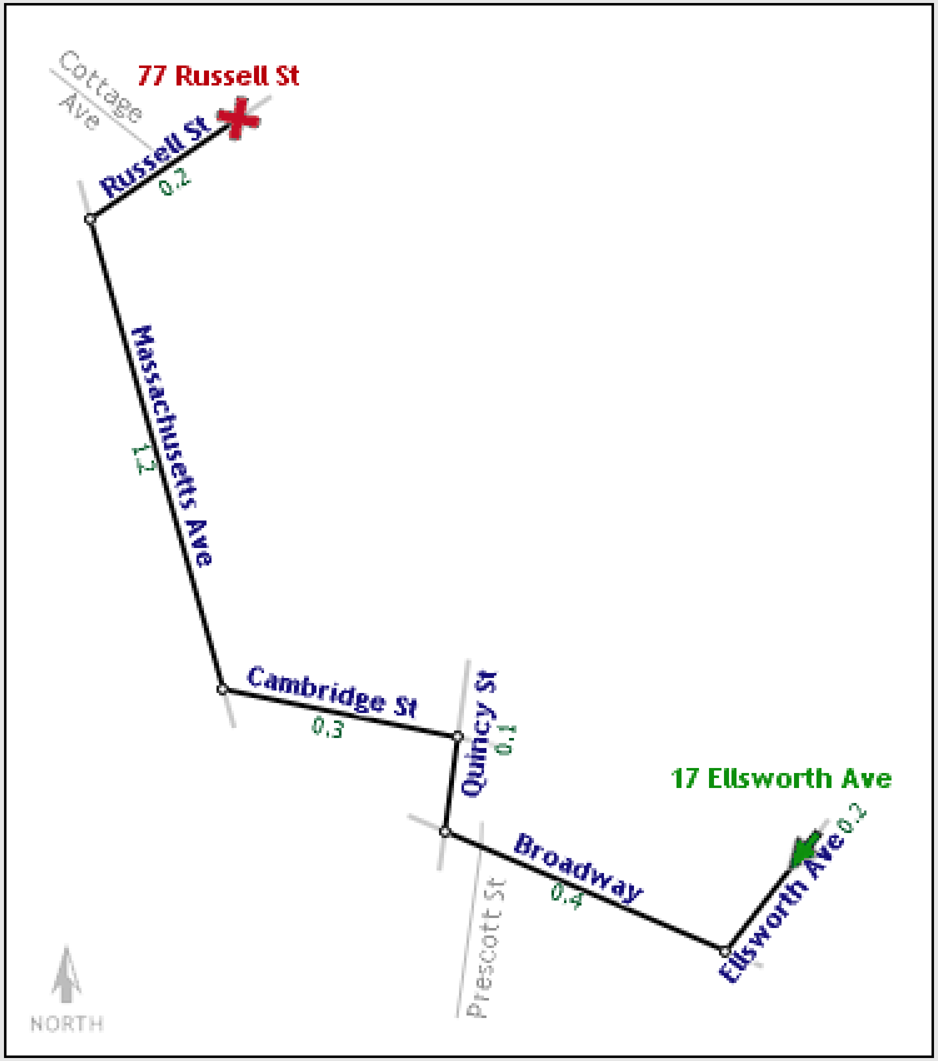


Image from mapquest.com

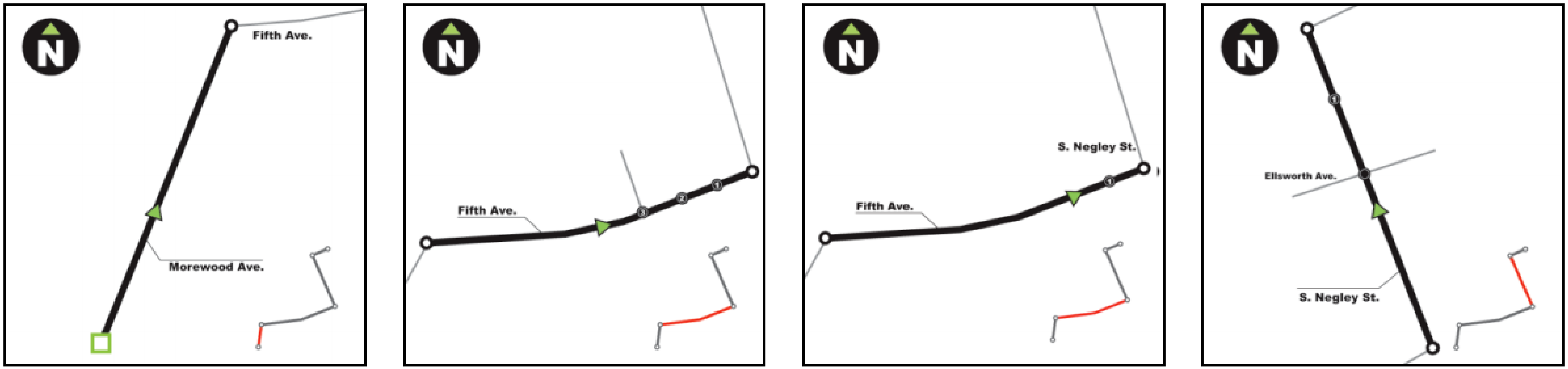
The estimated travel time is 5 minutes for 2.16 miles of travel, total of 6 steps.



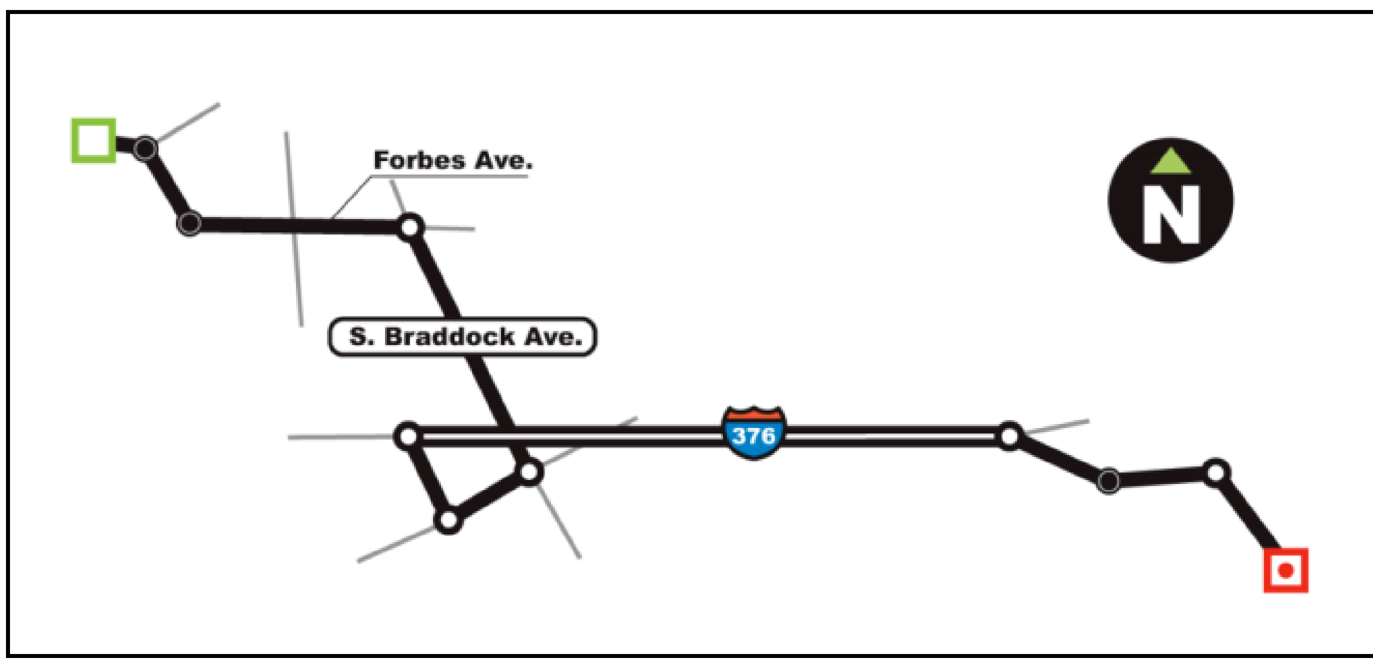
	Directions	Elapsed Distance
1	Begin at **17 Ellsworth Ave** on **Ellsworth Ave** and go Southwest for 500 feet	0.1
2	Turn right on **Broadway** and go Northwest for 0.4 miles	0.5
3	Turn right on **Quincy St** and go North for 200 feet	0.5
4	Turn left on **Cambridge St** and go West for 0.3 miles	0.8
5	Bear right on **Massachusetts Ave,Mass Ave,RT-2A** and go North for 1.2 miles	2.0
6	Turn right on **Russell St** and go Northeast for 1000 feet to **77 Russell St**	2.2

Line drive tool by Maneesh Agrawala http://graphics.stanford.ed

The Power of Design *and* interaction in Visualization



Lee, Forlizzi & H



Planning a Visualization

1. What is its goal?
2. What visual queries does it support?
3. What are some compelling, *useful* examples? [COPY COPY COPY!]
4. Could it have been done more simply?

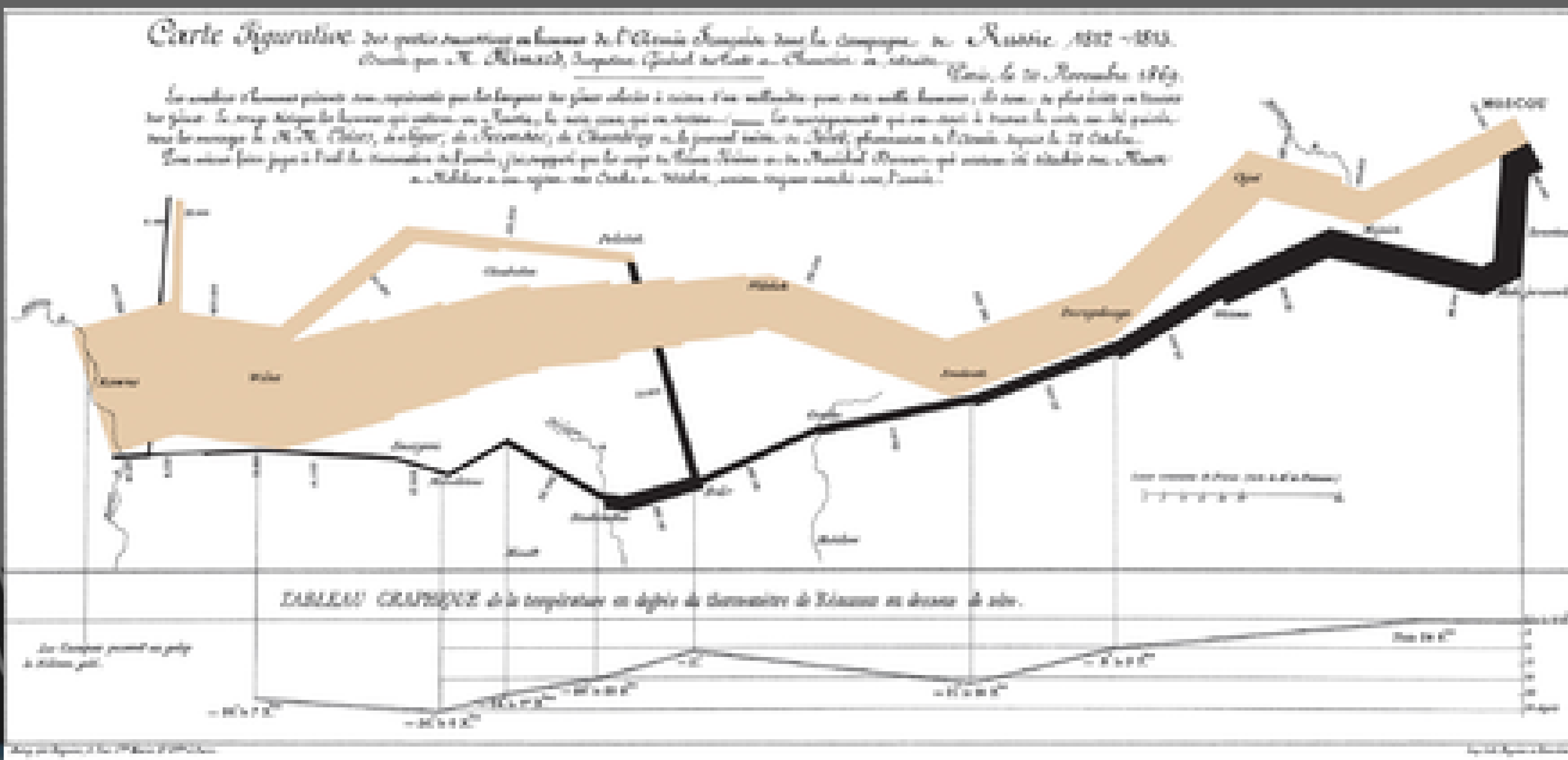
Making Queries

- Define the query[ies] you wish to support
- "The special skill of designers … [is] the talent to analyze a design in terms of its ability to support the visual queries of others…"
  - Patterns ⇔ visual system
  - Cognitive process prediction
  - A continual fresh eye

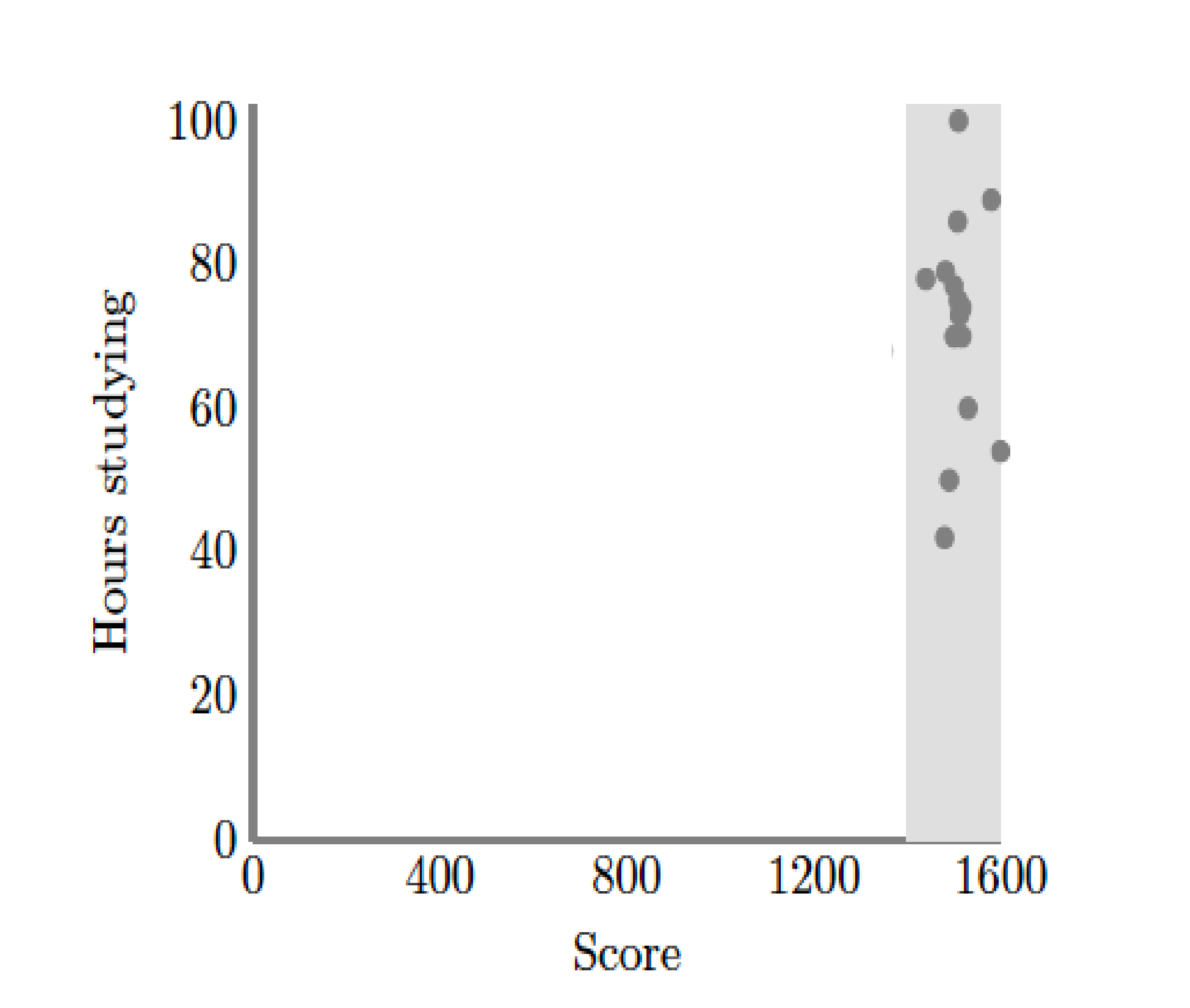
Narrative in Visualization

Documenting the question you are answering

Leading the viewer through a story

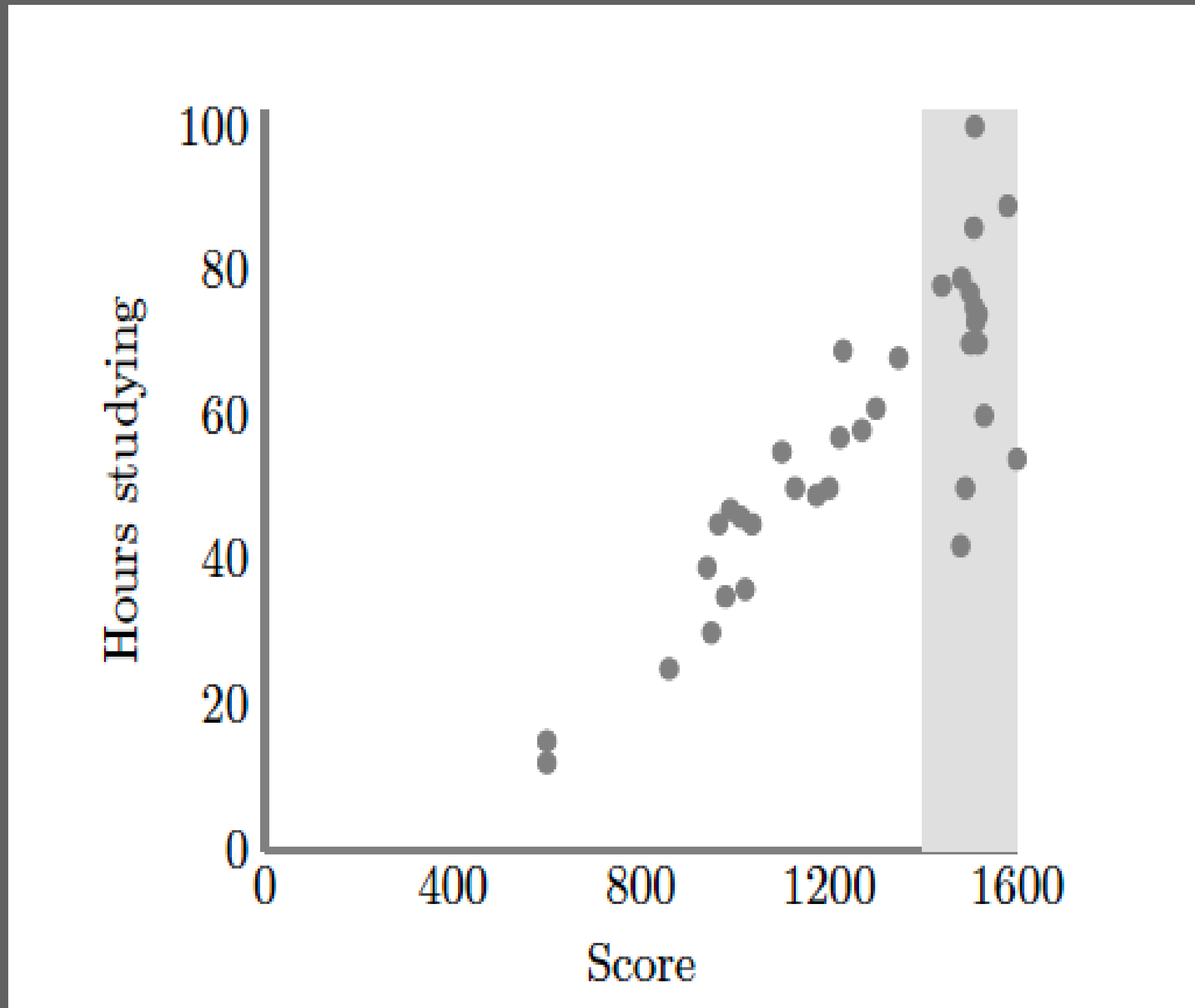


Visualizations freequently casue us to draw conclusions



S. Kleinberg (2015) Why: A Guide to Finding and Using Causes. O

Which is why visualization choices are so important…

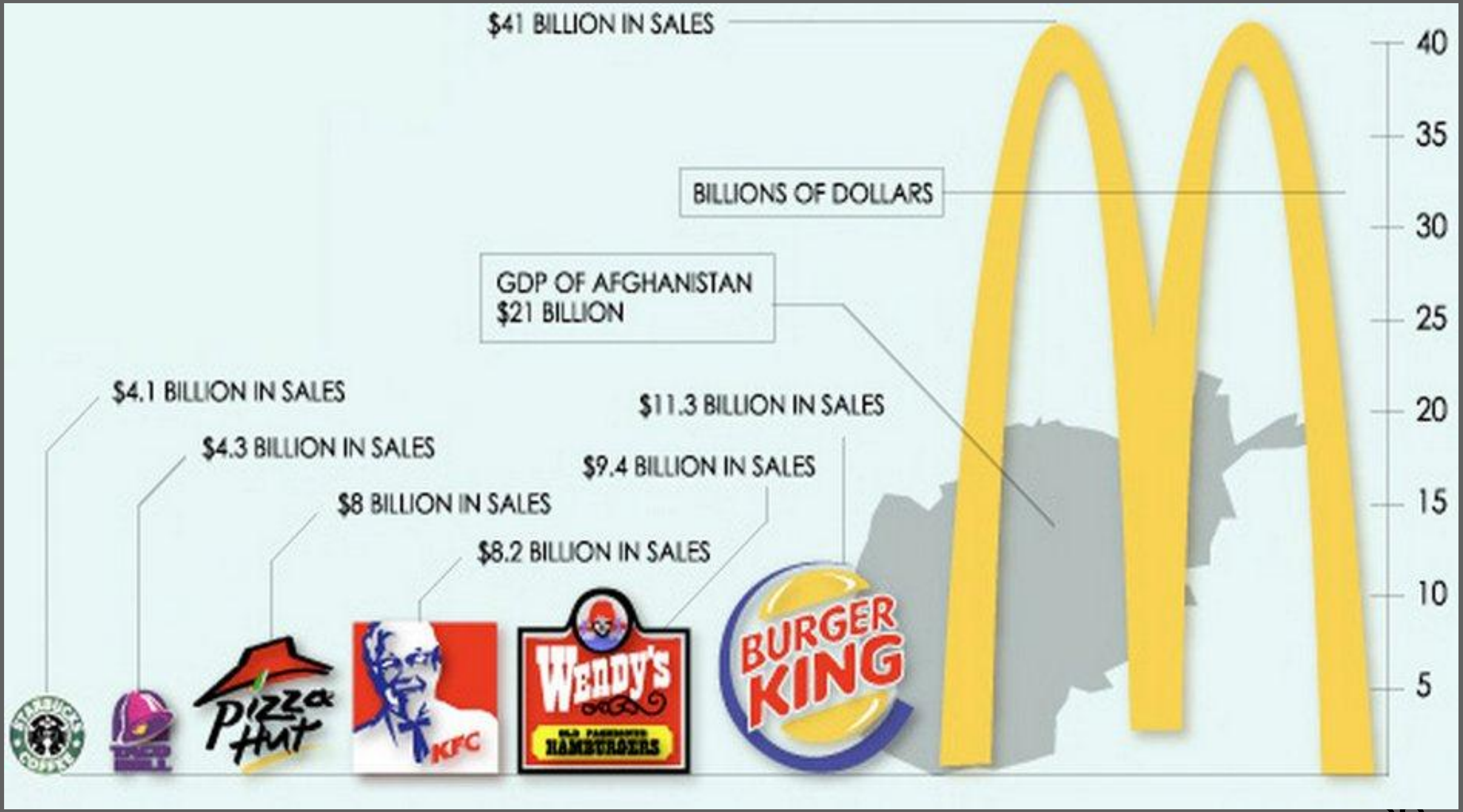


S. Kleinberg (2015) Why: A Guide to Finding and Using Causes. O

Planning a Visualization

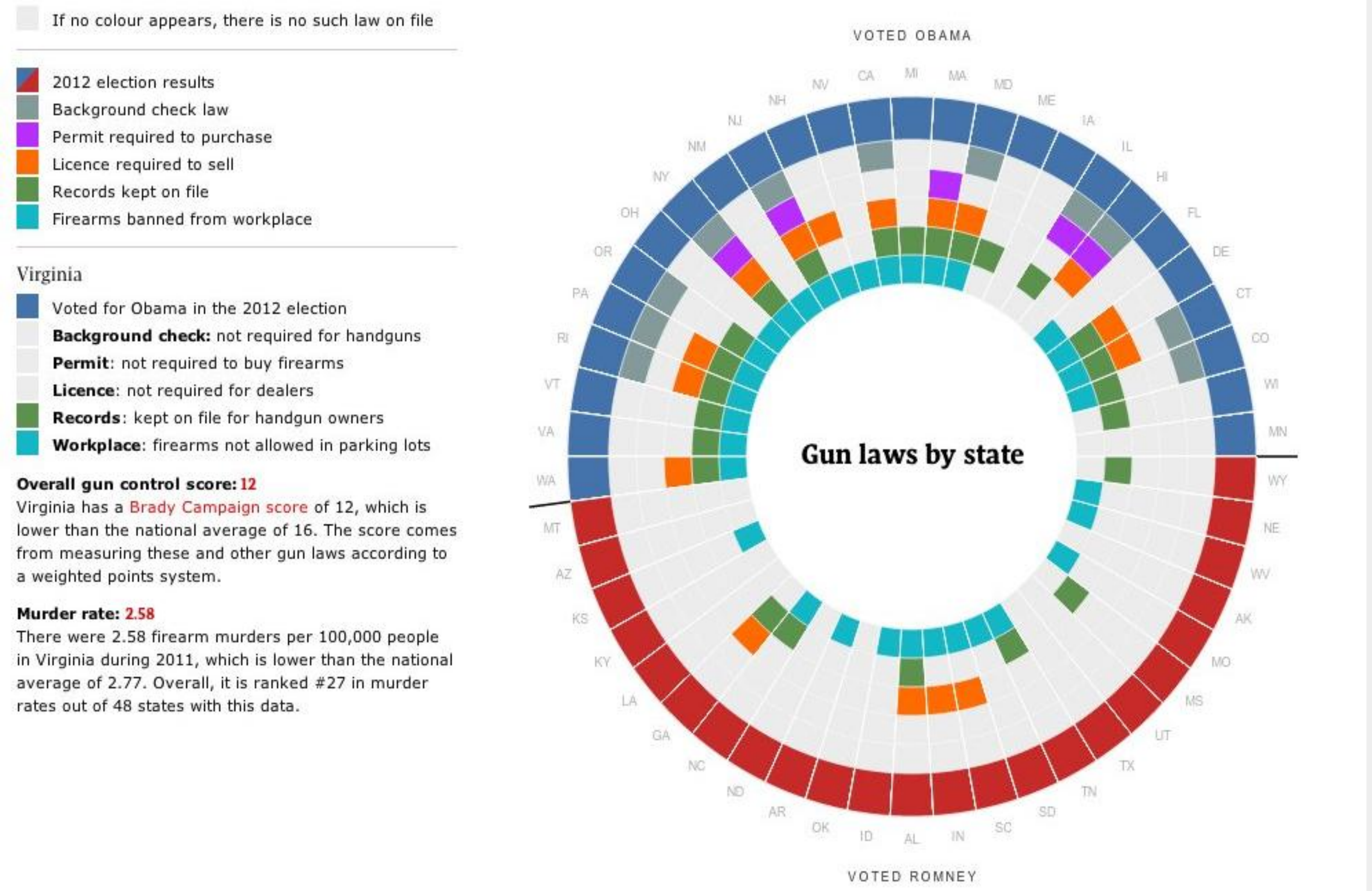
1. What is its goal?
2. What visual queries does it support?
3. What are some compelling, *useful* examples? [COPY COPY COPY!]
4. Could it have been done more simply?

Area vs Size



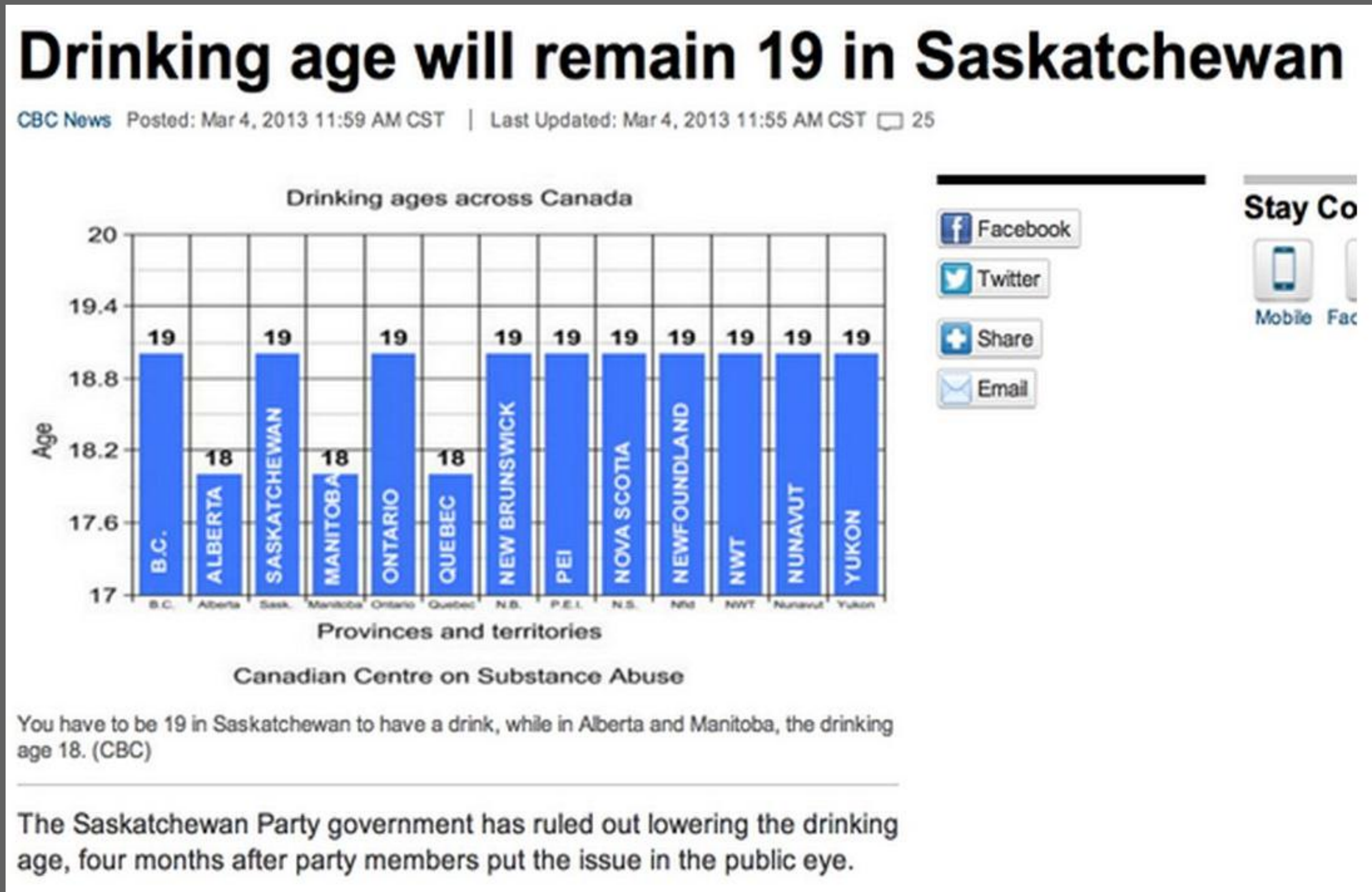
http://www.businessinsider.com/the-27-worst-charts-of-all-

Chart doesn't match data type



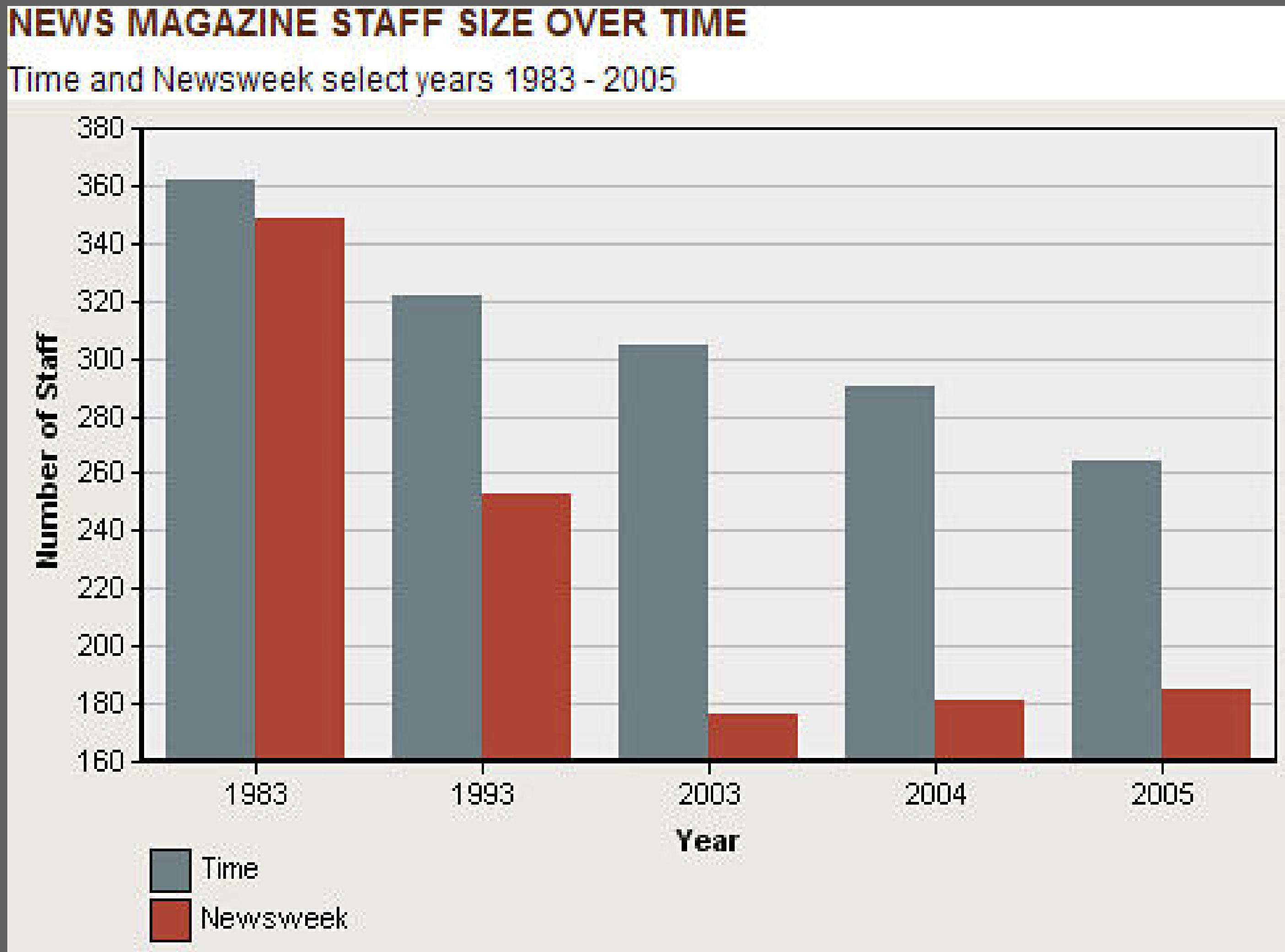
http://www.businessinsider.com/the-27-worst-charts-of-all-

Confusing Axes



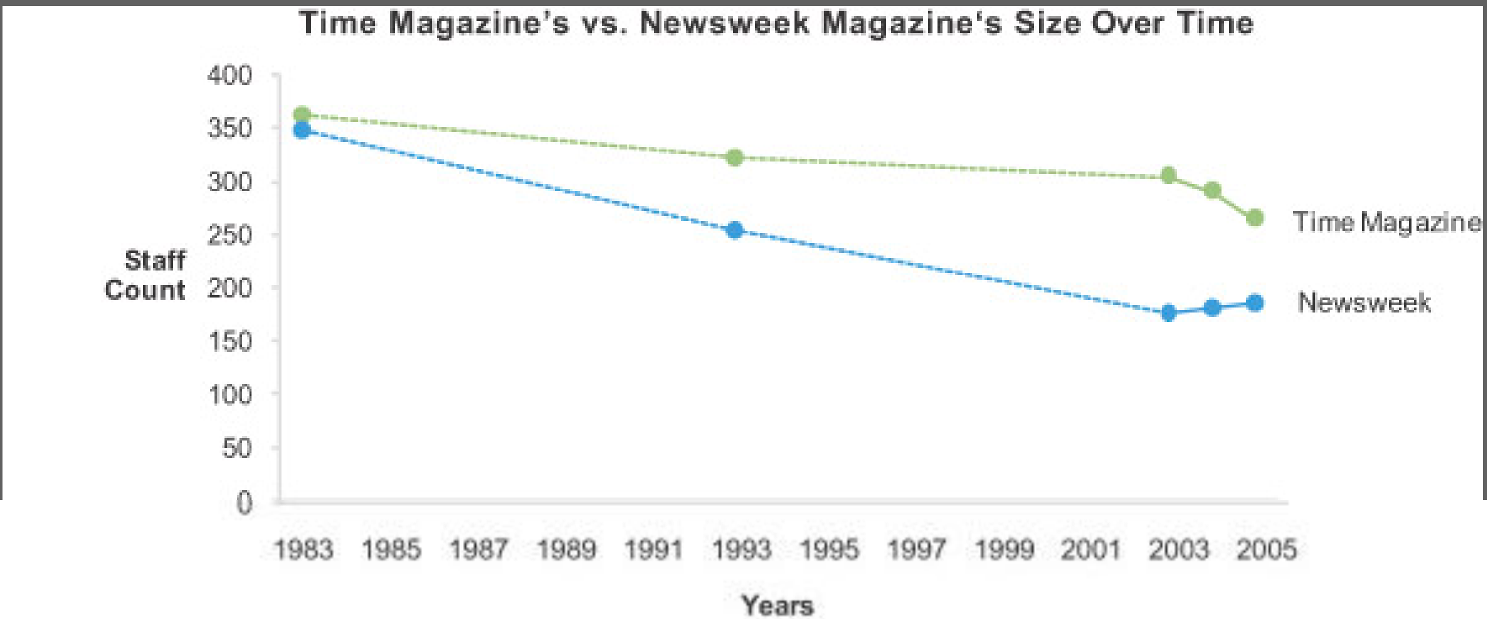
http://www.businessinsider.com/the-27-worst-charts-of-all-

Misleading Axis



https://www.perceptualedge.com/example14.php

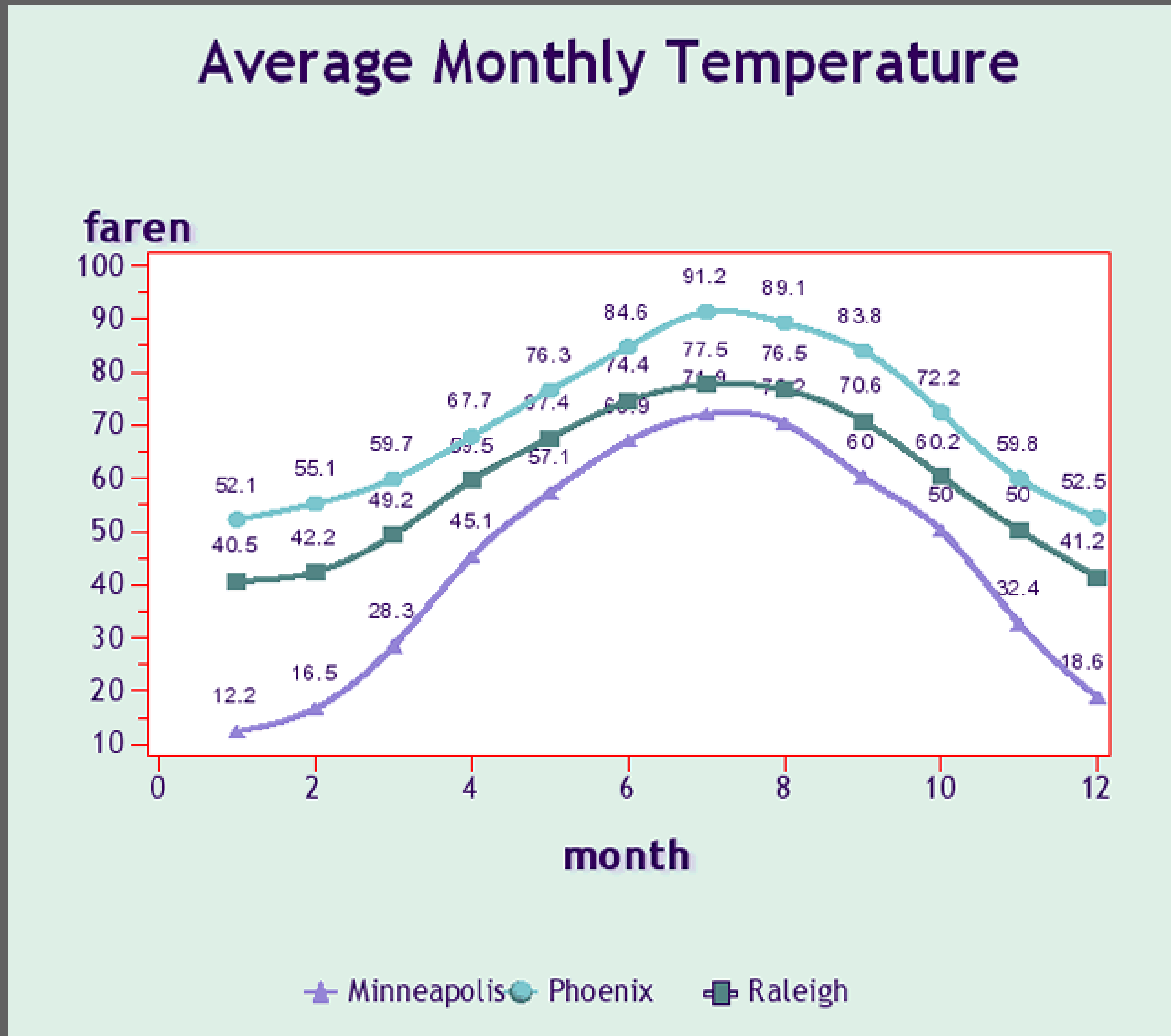
Improved



Note: A dashed line connecting two points indicates that there are years between the points for which values were not available. If the values were available, the shape of the lines might vary significantly.

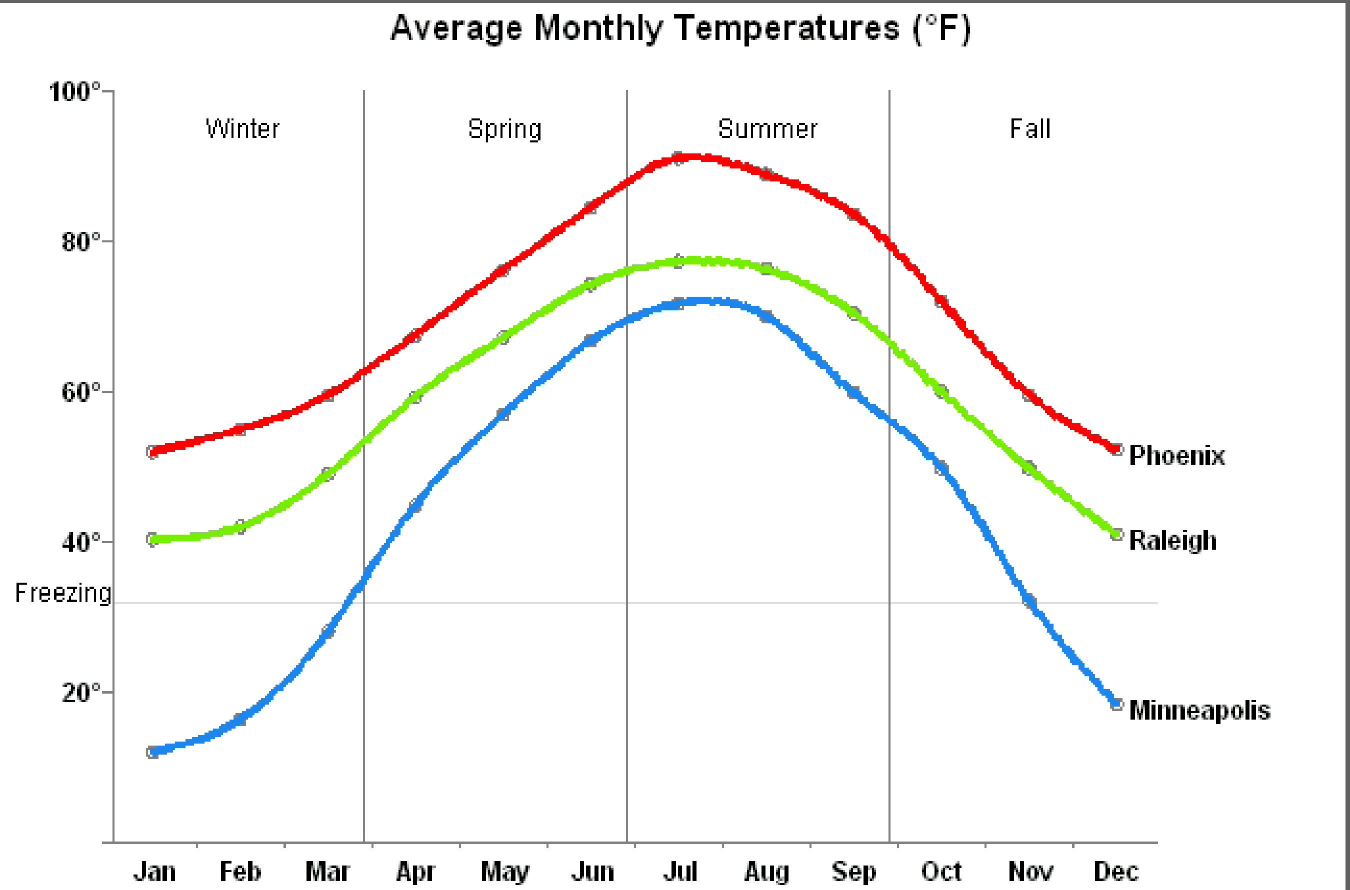
https://www.perceptualedge.com/example14.php

Overly Complex



https://www.perceptualedge.com/example2.php

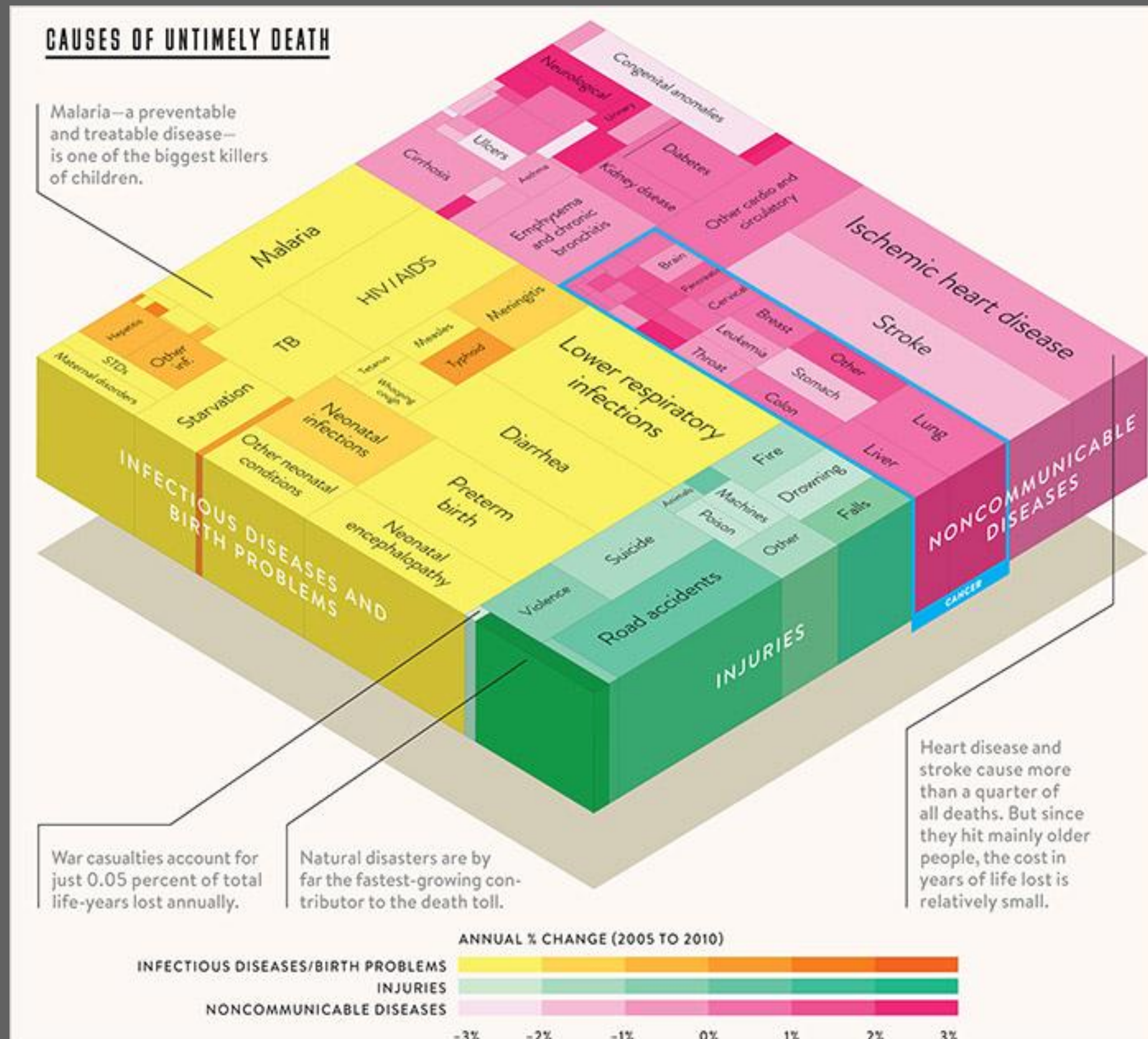
Improved



	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec	
	52.1	55.1	59.7	67.7	76.3	84.6	91.2	89.1	83.8	72.2	59.8	52.5	Phoenix
	40.5	42.2	49.2	59.5	67.4	74.4	77.5	76.5	70.6	60.2	50.0	41.2	Raleigh
	12.2	16.5	28.3	45.1	57.1	66.9	71.9	70.2	60.0	50.0	32.4	18.6	Minneapolis

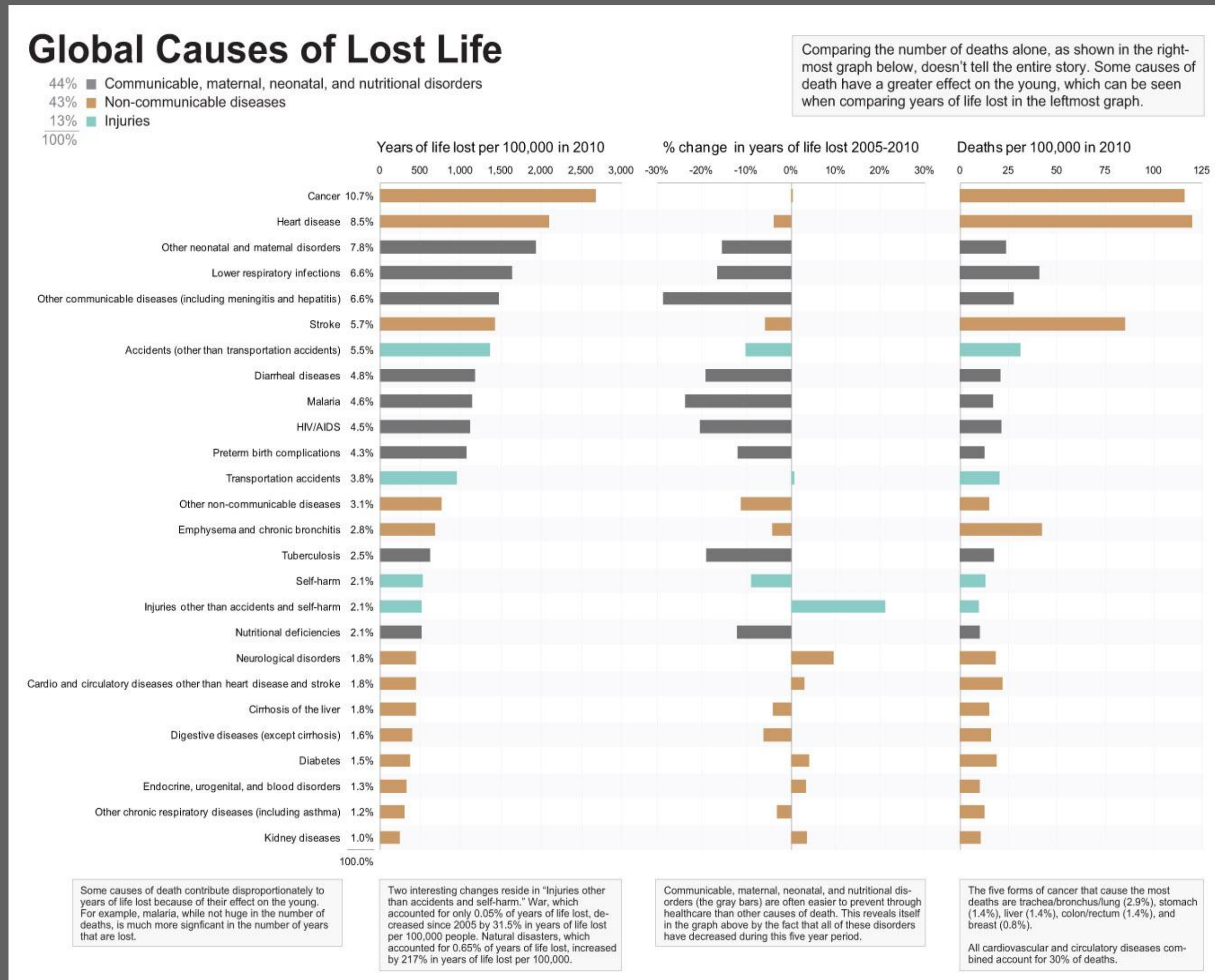
https://www.perceptualedge.com/example2.php

Hard to read



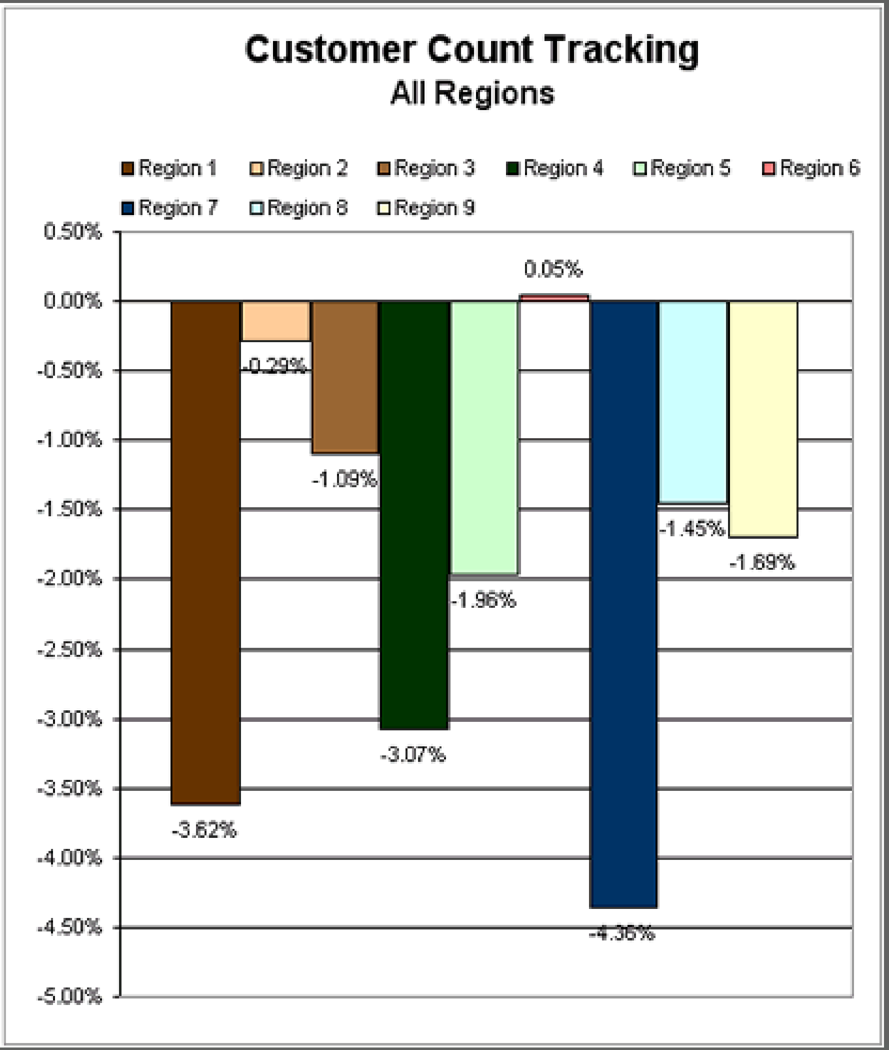
https://www.perceptualedge.com/example20.php

Improved



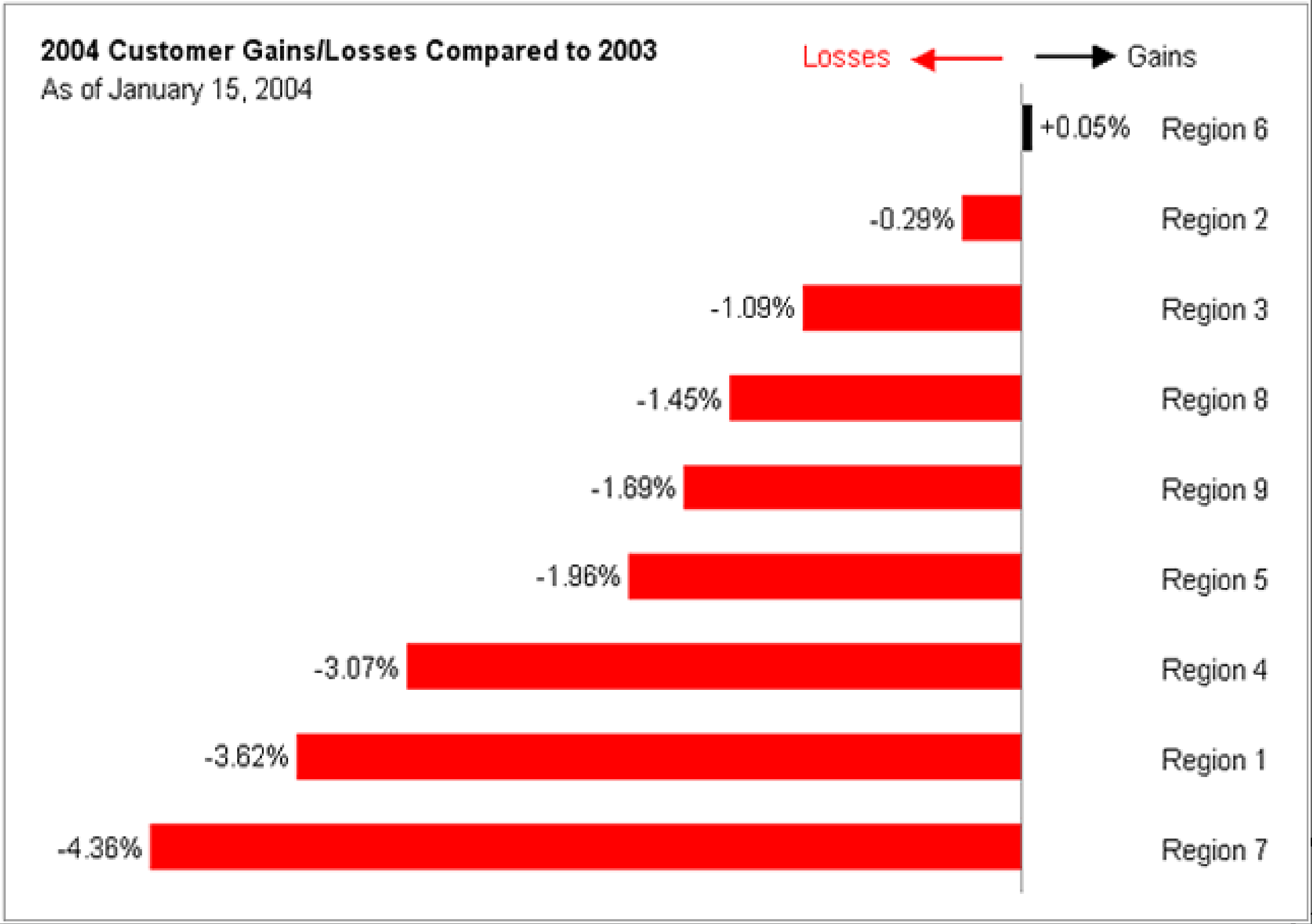
https://www.perceptualedge.com/example20.php

Comparisons Difficult



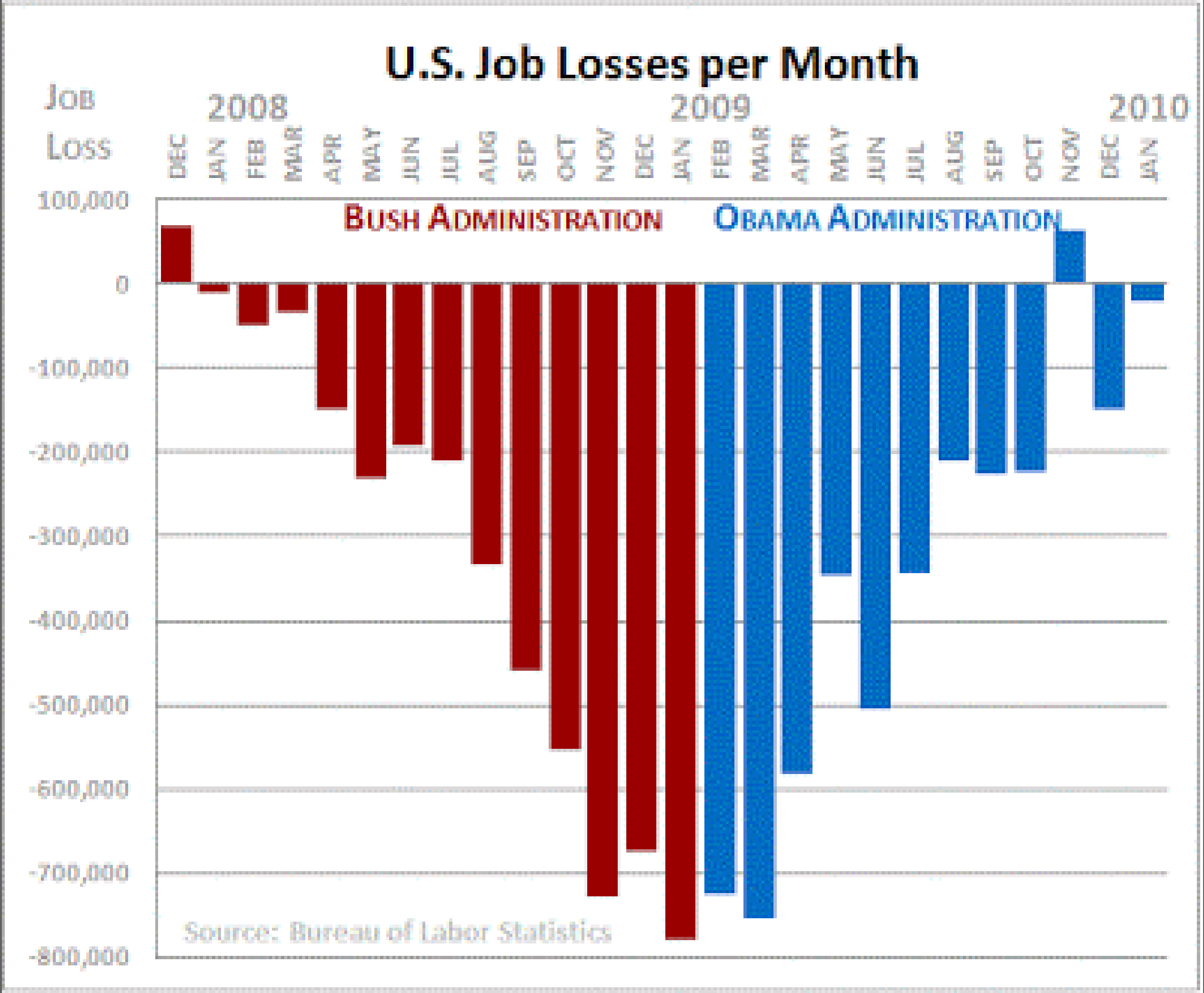
https://www.perceptualedge.com/example1.php

Improved



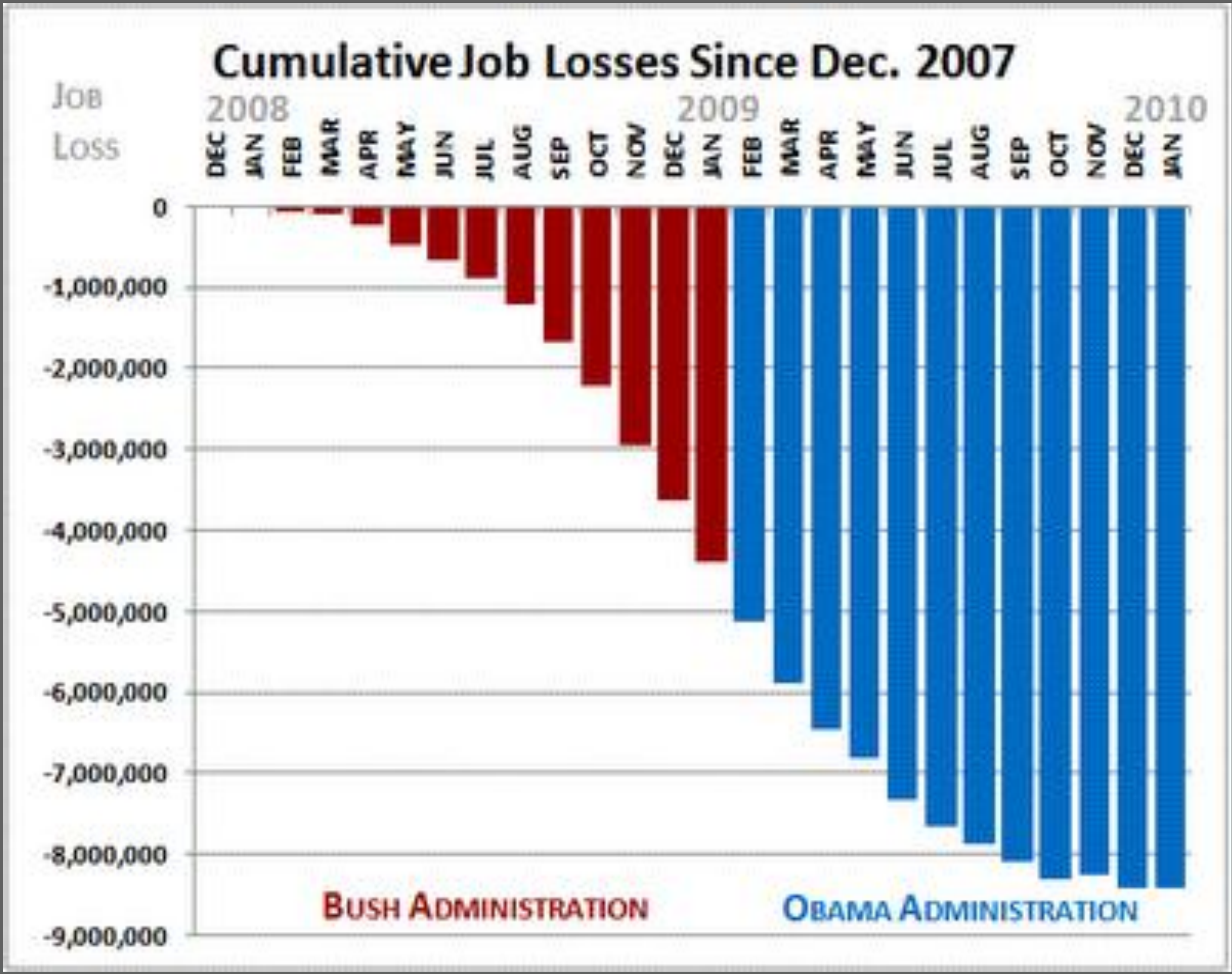
https://www.perceptualedge.com/example1.php

Highlighting wrong aspect of data?



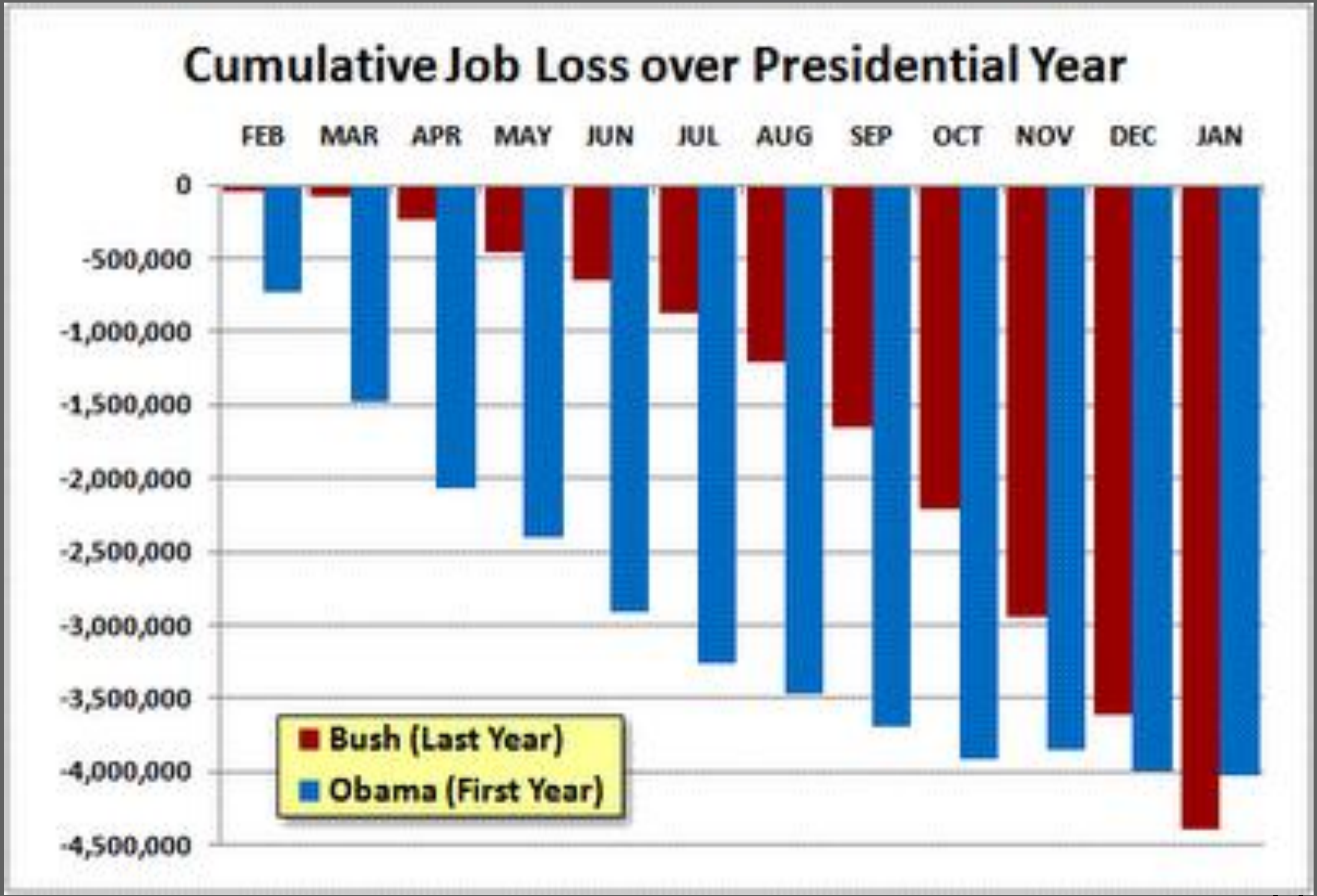
http://soquelbythecreek.blogspot.com/2010/02/what-does-

An alternative view



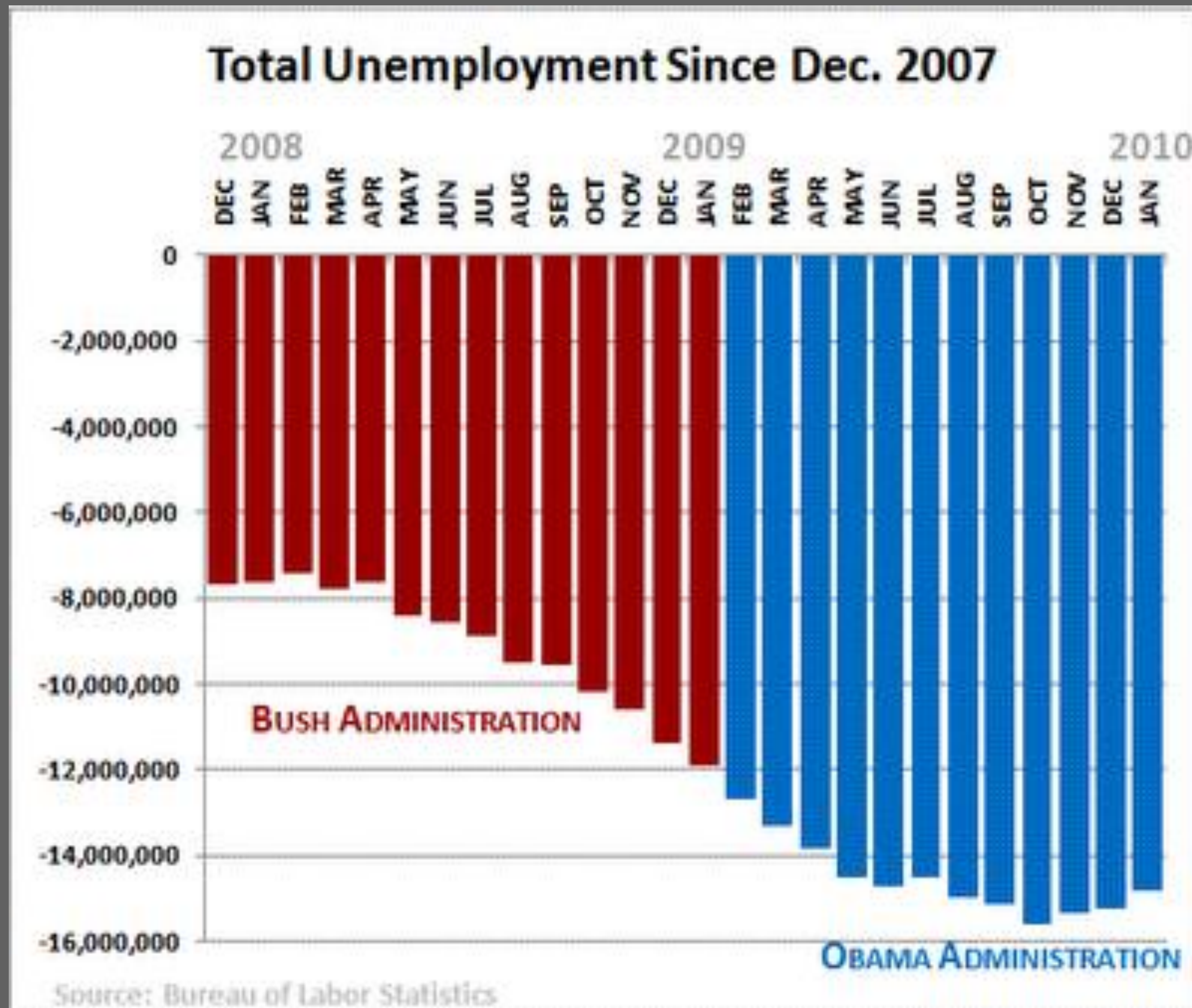
http://soquelbythecreek.blogspot.com/2010/02/what-does-

An alternative view



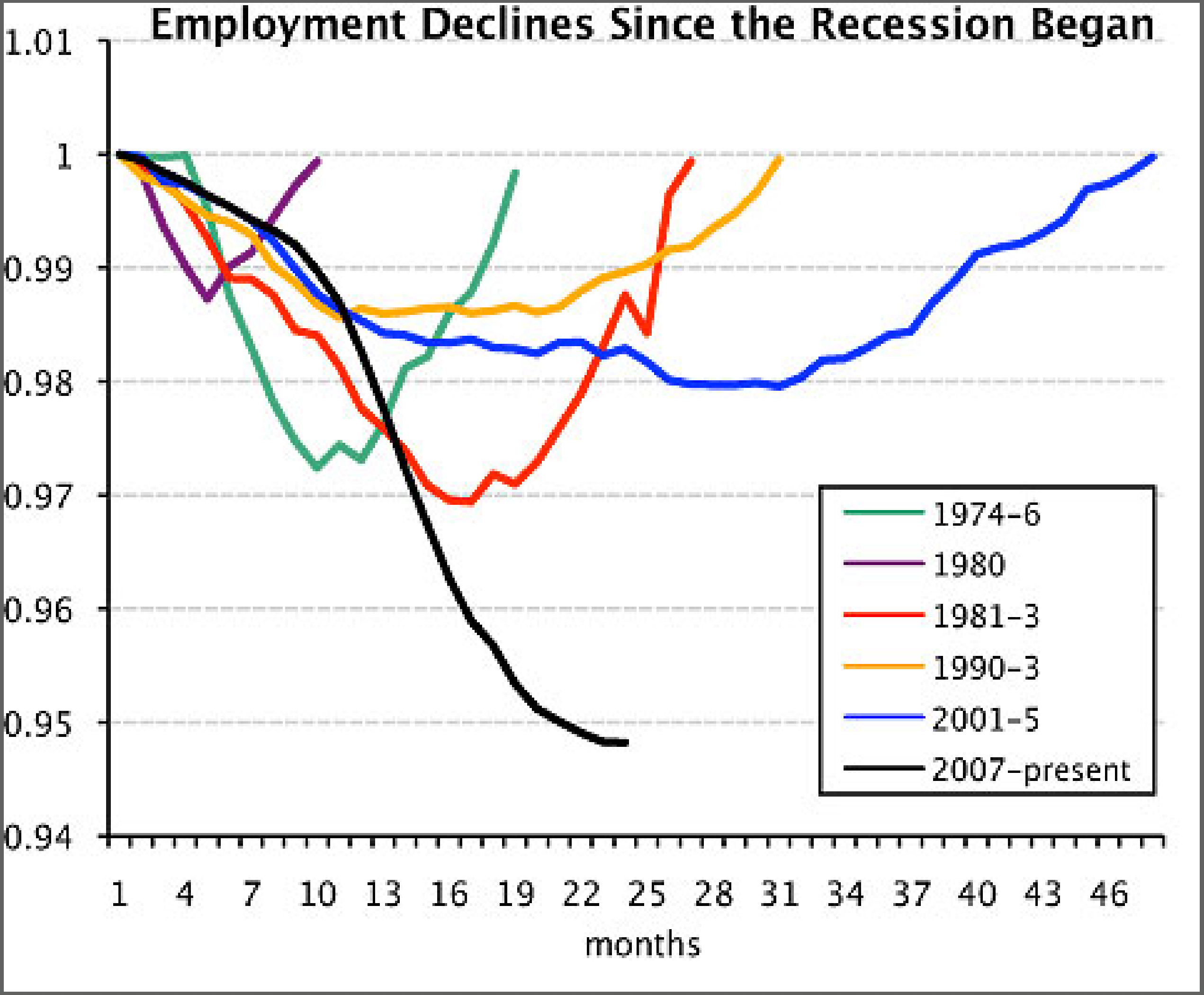
http://soquelbythecreek.blogspot.com/2010/02/what-does-

An alternative view



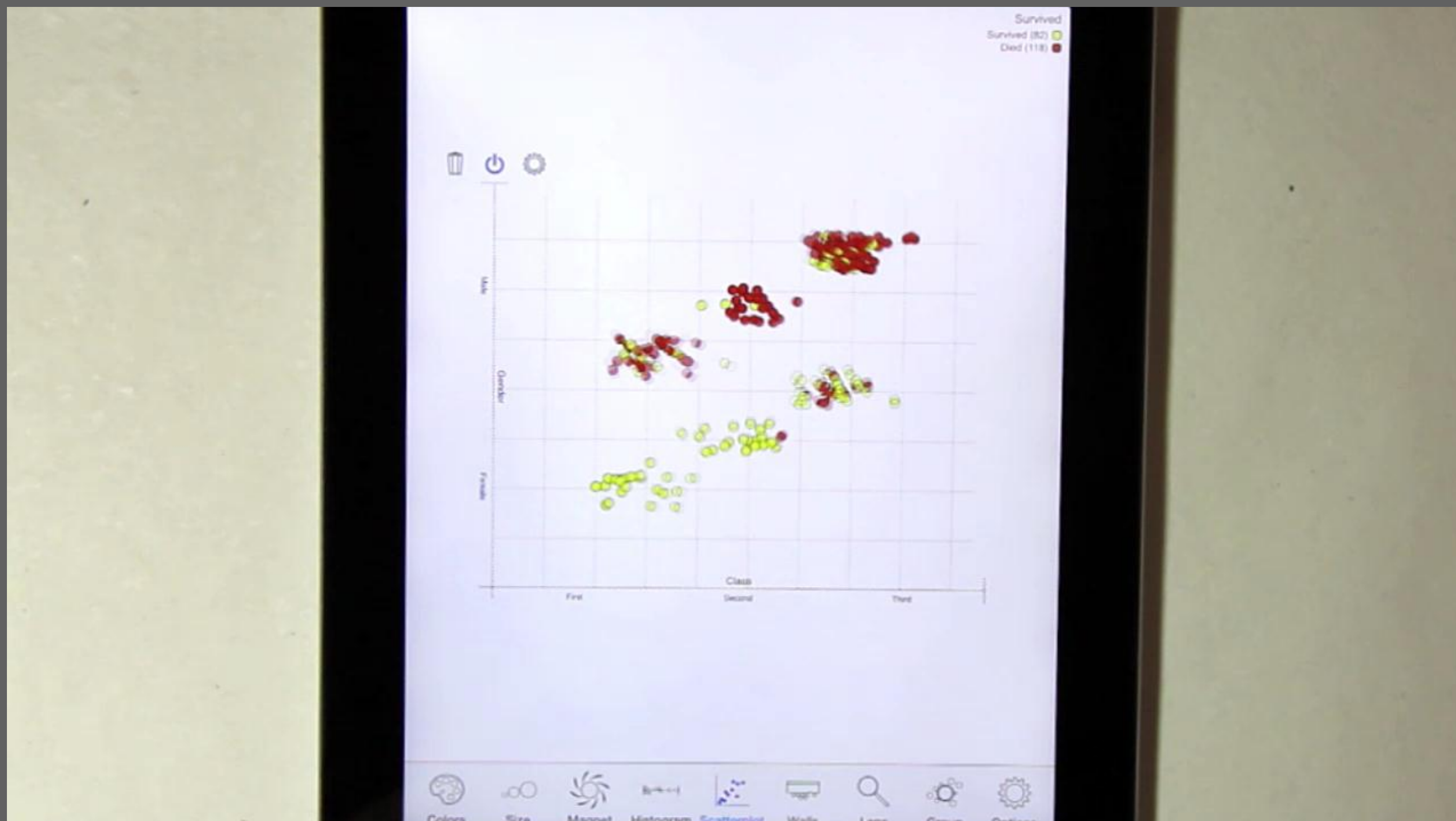
http://soquelbythecreek.blogspot.com/2010/02/what-does-

An alternative view



http://soquelbythecreek.blogspot.com/2010/02/what-does-

Interactive Visualization



Jeffrey M. Rzeszotarski and Aniket Kittur. 2014. Kinetica: naturalistic multi-touch data visualization. In Proceedings of the 32nd annual ACM conference on Human factors in

Interactive Visualization

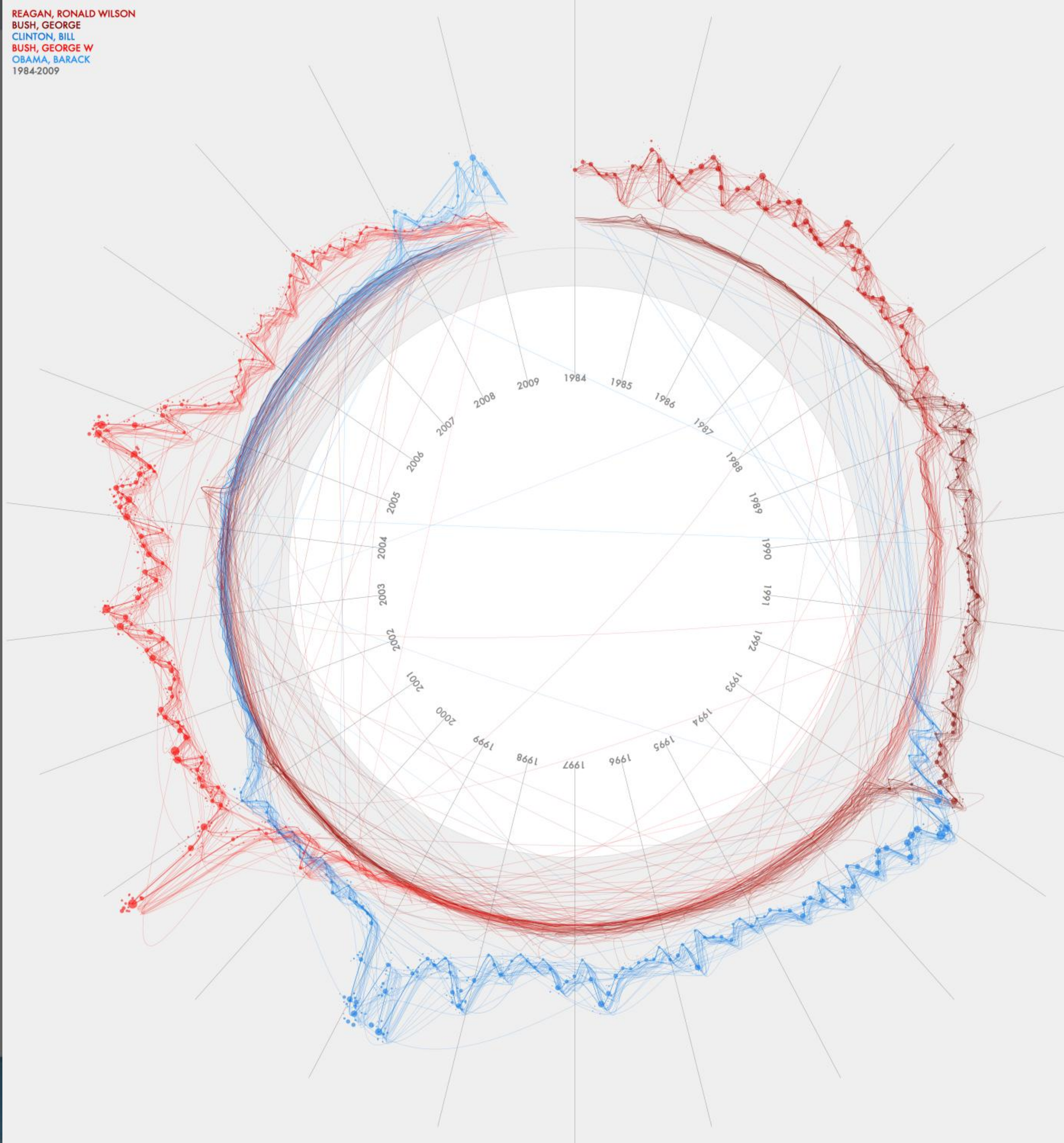
http://queue.acm.org/detail.cfm?id=2146416

Visualizing Big Data

First a tour

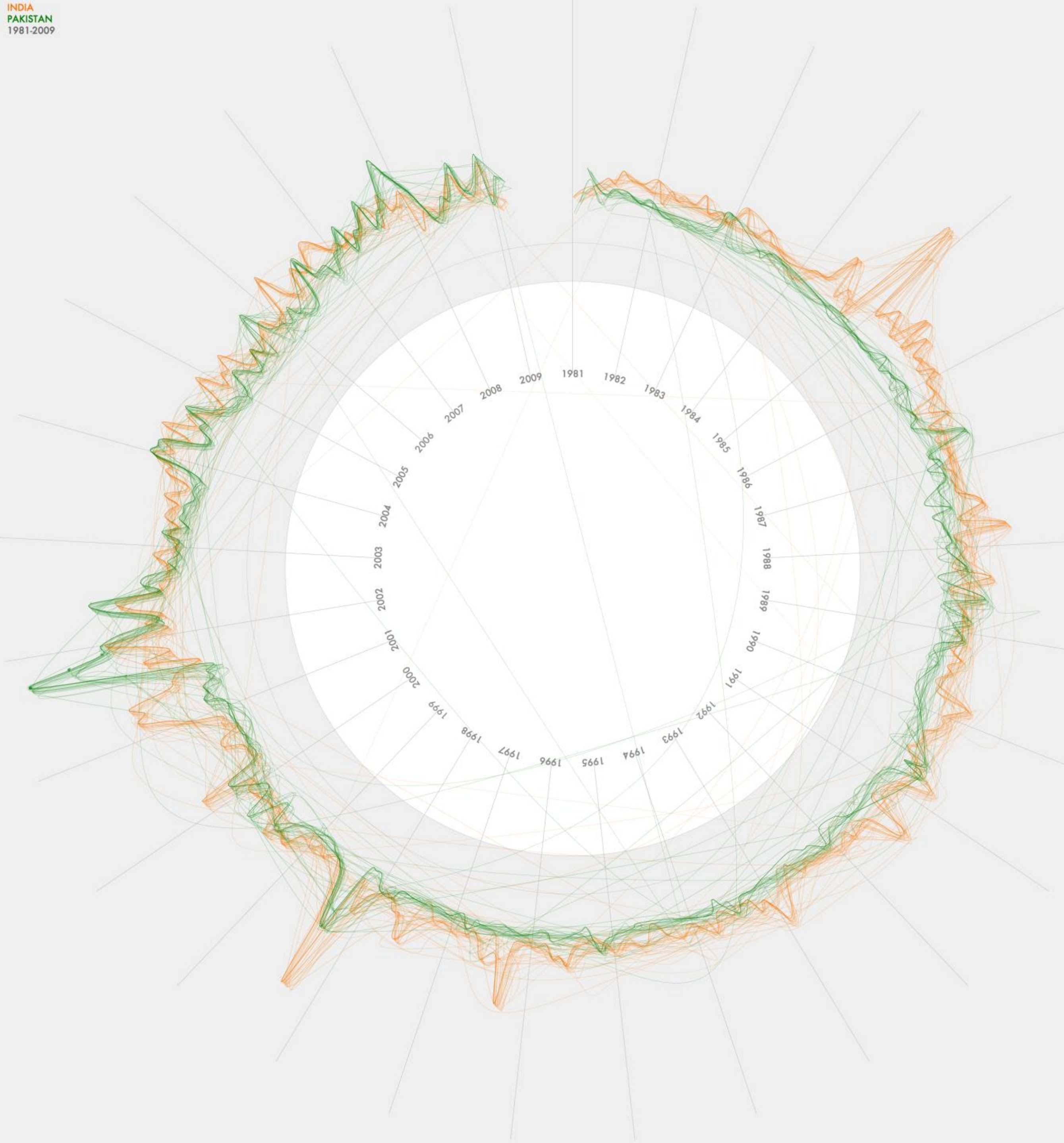
Then some techniques

http://www.smashingmagazine.com/2007/08/02/data-visualization-modern-approaches



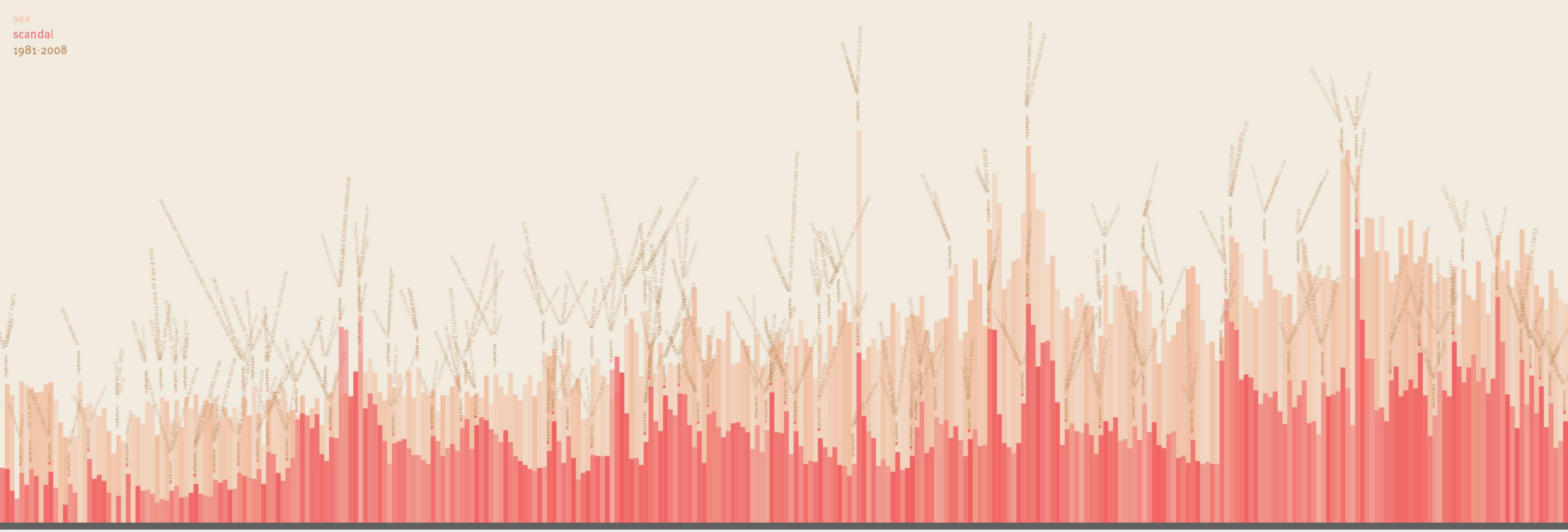
Jer Thorp
Data Artist

This graph charts the frequency of mention of the five US Presidents between 1984 and 2009. It also indicates weighting of stories - the darkest line shows front page stories while the lighter lines indicate stories buried deeper in the



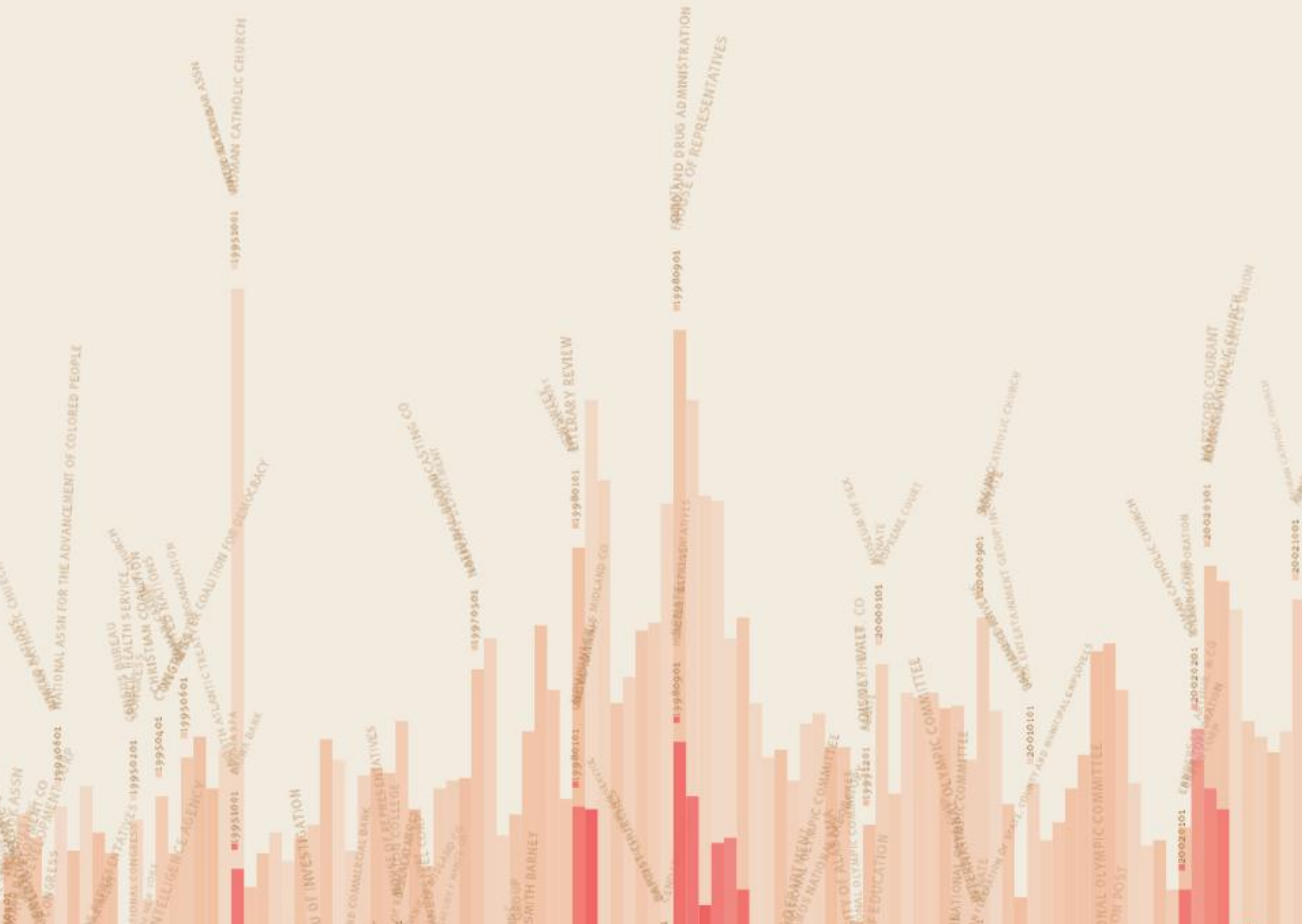
Jer Thorp
Data Artist

NYTimes Threads - India & Pakistan
This graph charts the frequency of articles mentioning India and Pakistan in the NYT between 1981 and 2009. It also indicates weighting of stories - the darkest line shows front page stories while the lighter lines indicate stories buried deeper in the paper



Jer Thorp
Data Artist

NYTimes: Sex & Scandal since 1981
This is a visualization of the frequency of occurrence of the words 'sex'

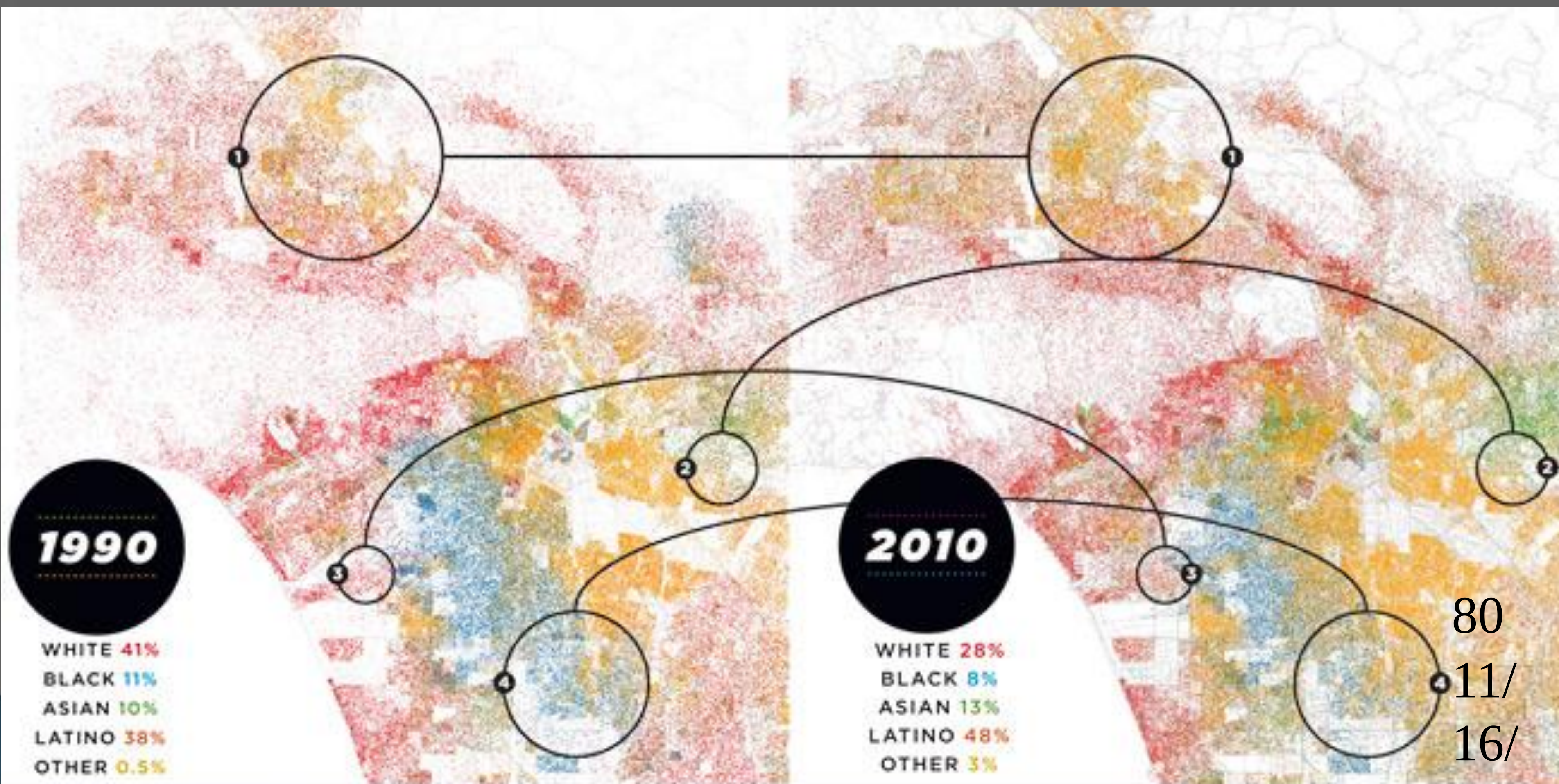


NATIONAL ASSN FOR THE ADVANCEMENT OF COLORED PEOPLE
ROMAN CATHOLIC CHURCH
19951001
FOOD AND DRUG ADMINISTRATION
HOUSE OF REPRESENTATIVES
19980901
LITERARY REVIEW
19980101
19970301
20000101
20000901
20020301
20021001
HARTFORD COURANT
SMITH BARNEY
OLYMPIC COMMITTEE

Race in LA (Eric Fi

Nice narrative @lamag.com

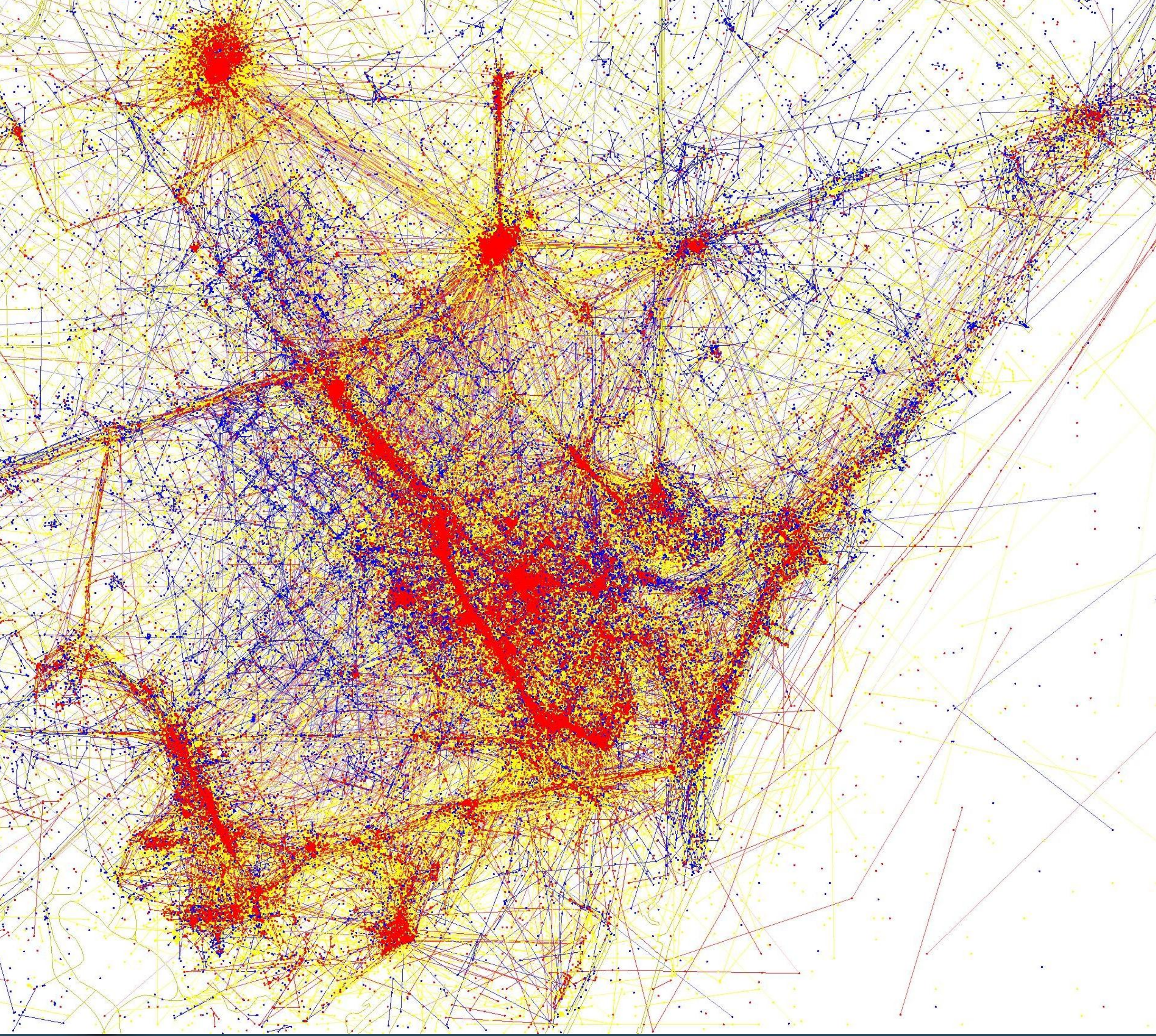
http://www.lamag.com/features-hidden/race-in-la-see-how-weve-grown/#0412infographic



World travel and communications recorded on Twitter
Green is physical movement from place to place; purple is @replies from someone in one location to someone in another; combining to white where there is both.

Reported trips to Null Island excluded; all other geotags trusted.
Endpoints of trips are real data; routes in between are fabricated.
Brightness is logarithmic.

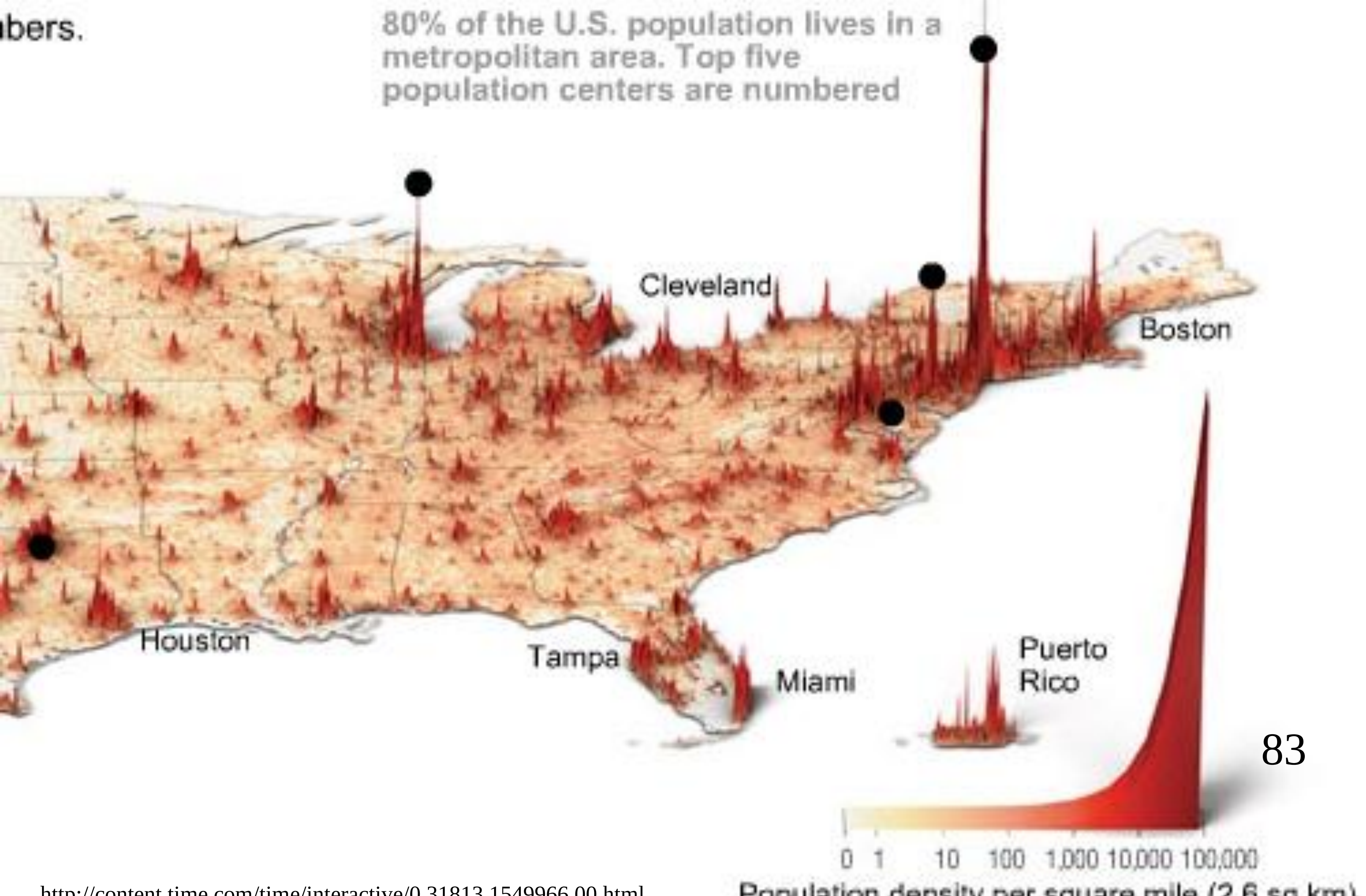
Data from the Twitter streaming API through September 1, 2011.
Continent shapes from Natural Earth. Author: Eric Fischer
https://www.flickr.com/photos/walkingsf/6635655755/in/photostream/



Tourist vs locals in Barcelona. Shows. Blue photos (Flickr) live in the city, red are tourists, yellow are unknown. Eric Fischer again

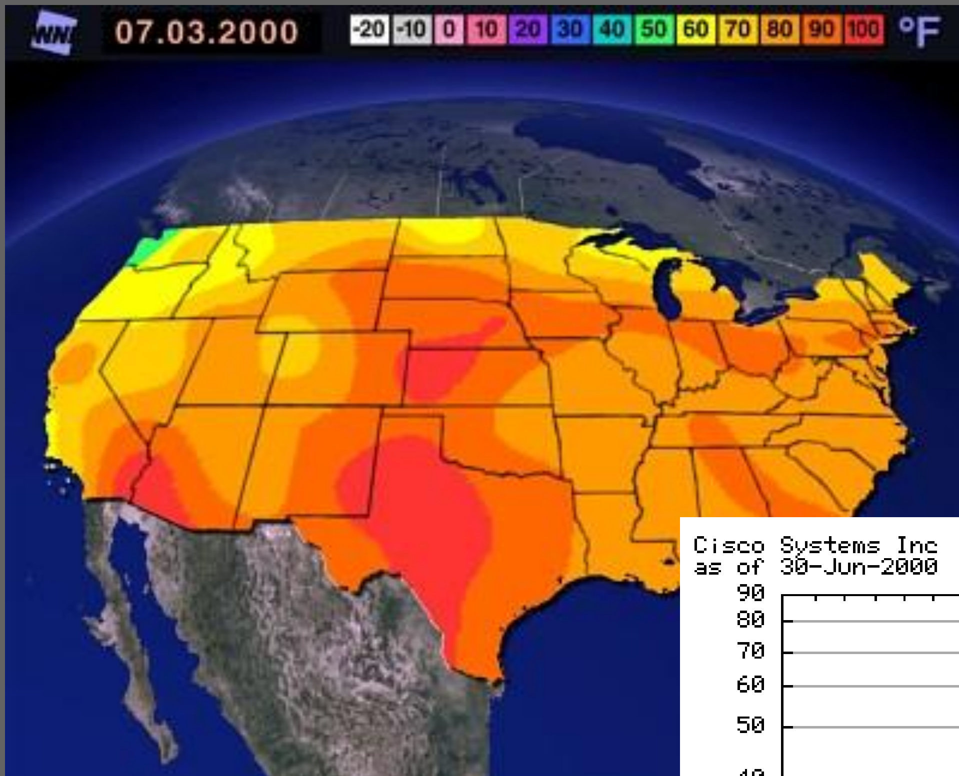
More details:

Time Magazine



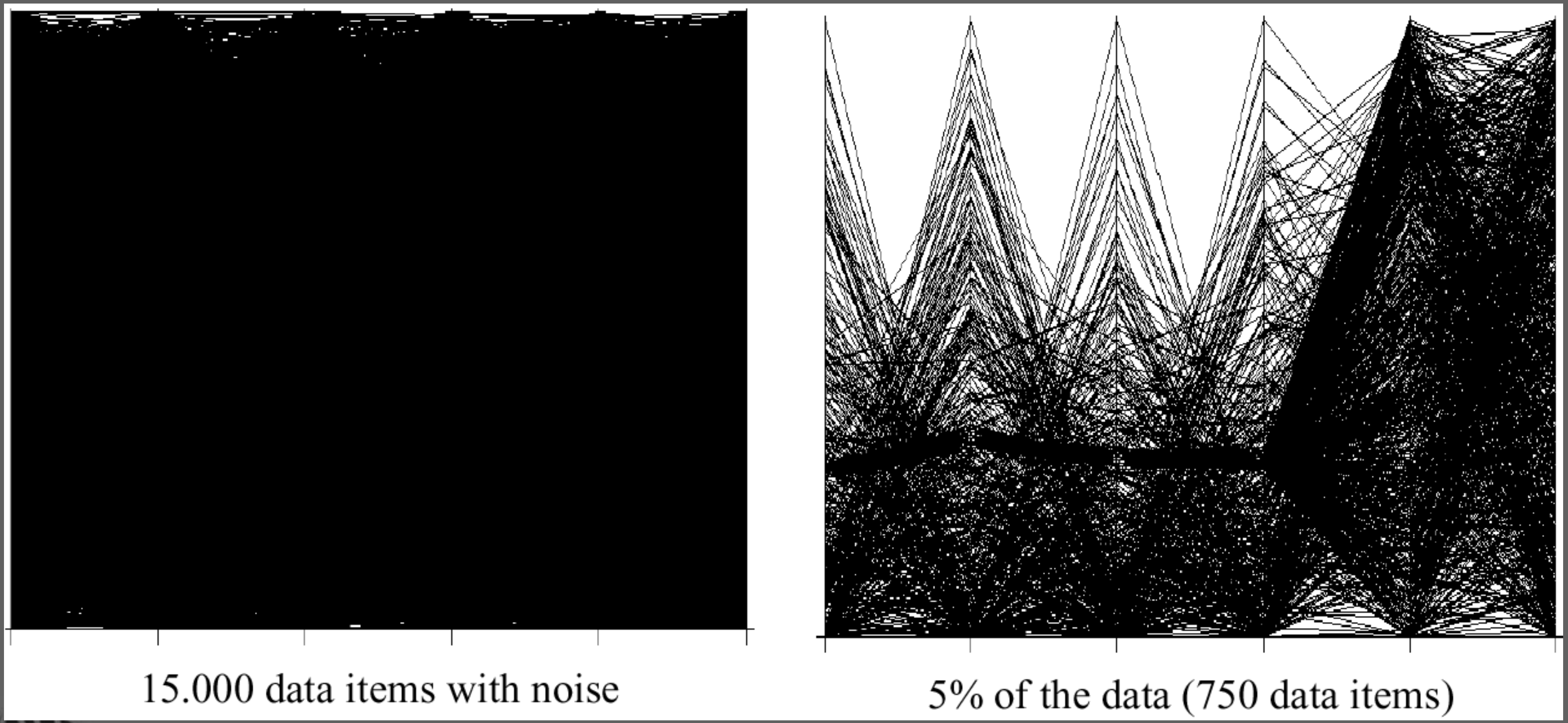
http://content.time.com/time/interactive/0,31813,1549966,00.html

Small Data?

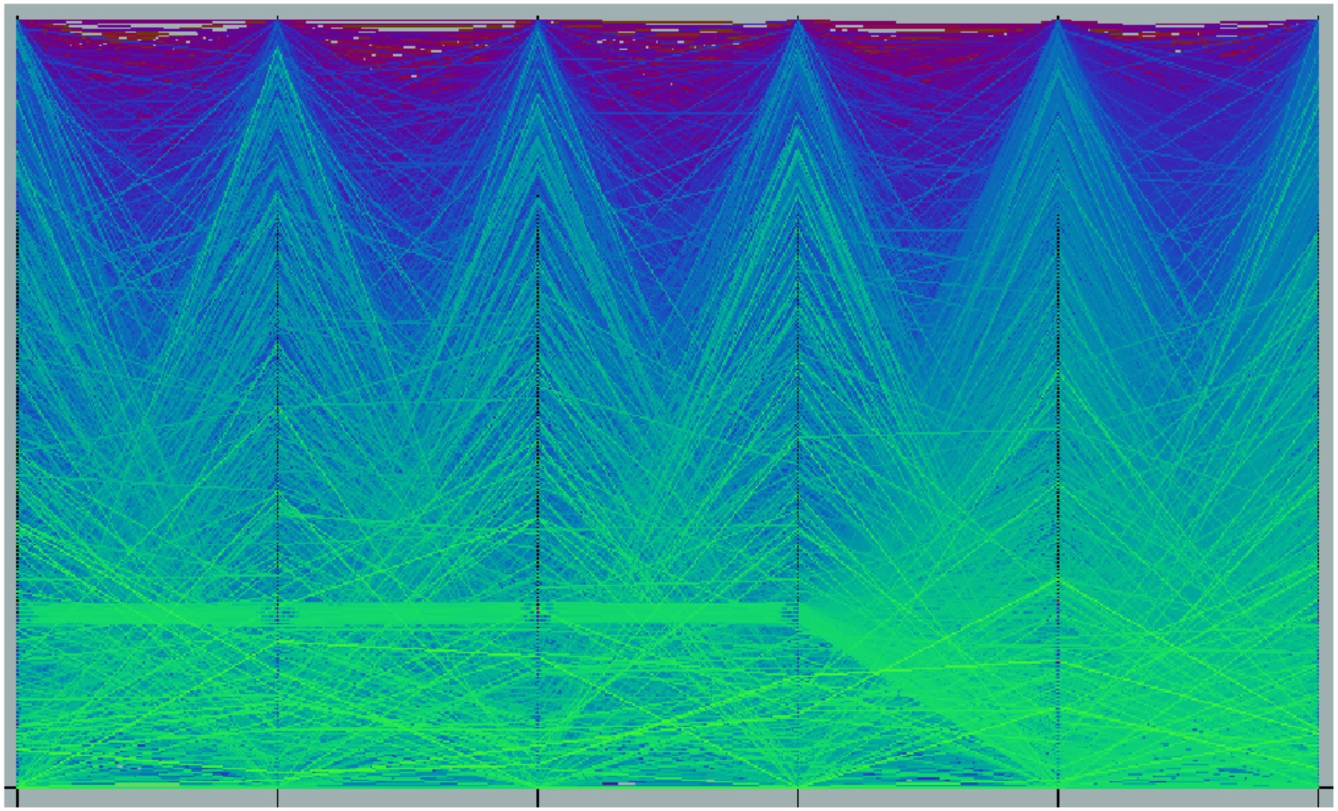


Images from yahoo.com

Visualization Techniques for Big Data

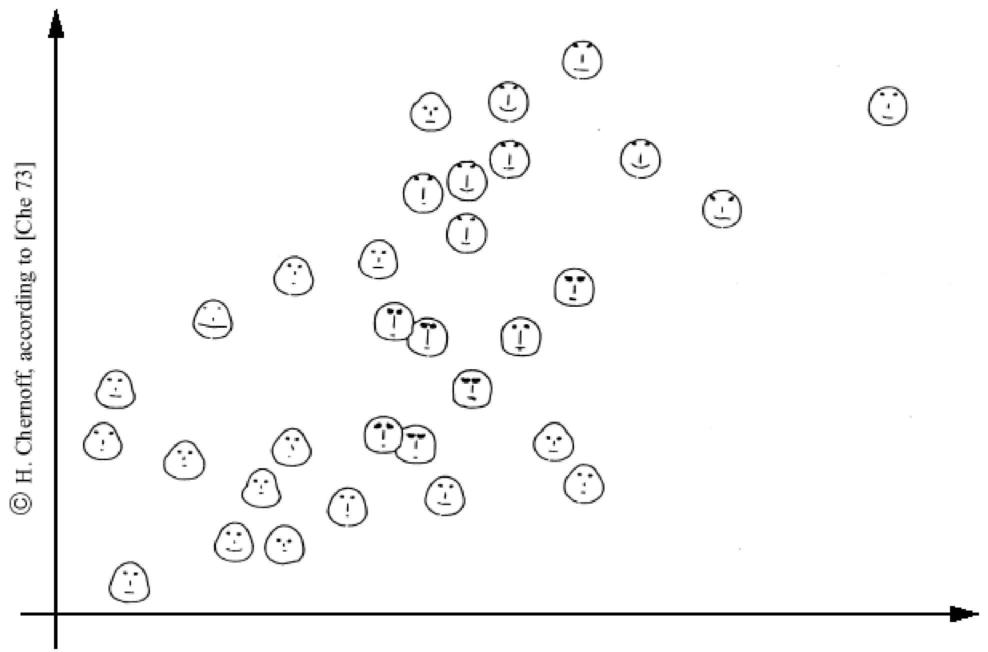


Query Dependent Coloring



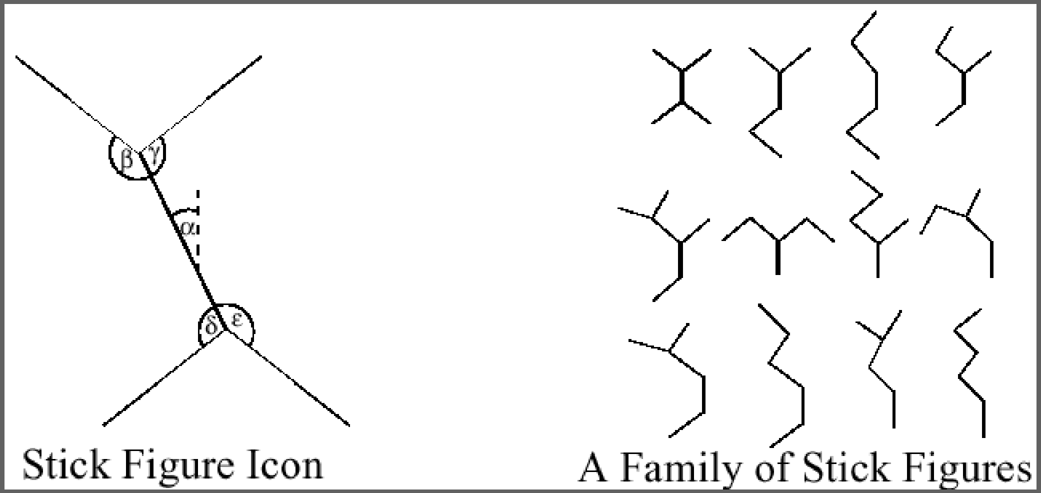
15.000 data items with a query-dependent coloring

Chernoff-faces

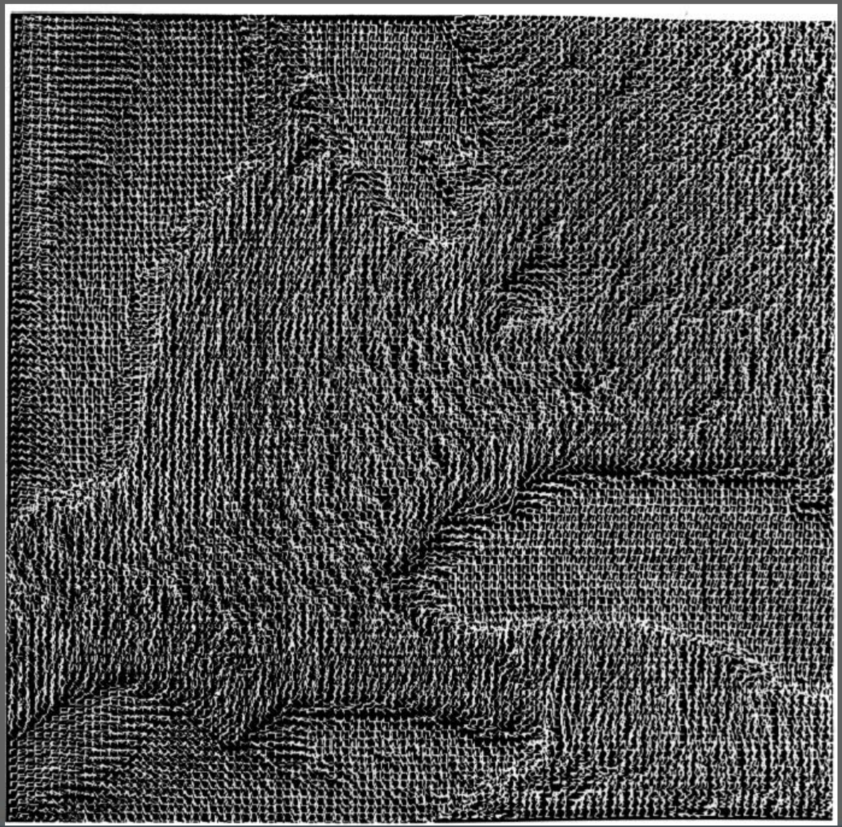


➪ visualization of the multidim. data using the properties of a face icon (shape of nose, mouth, eyes, and the shape of the face itself)

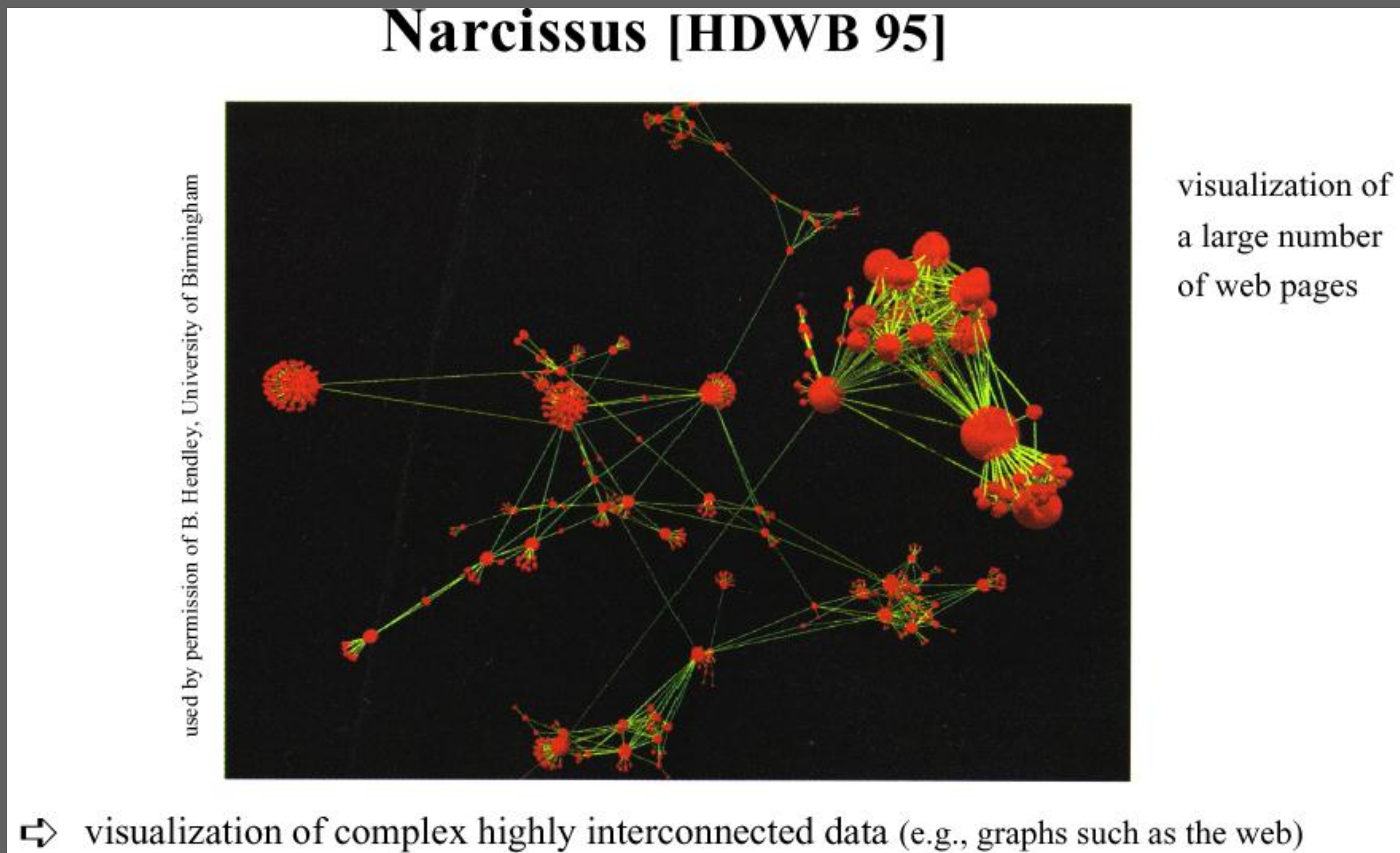
Stick Figures



Stick Figures

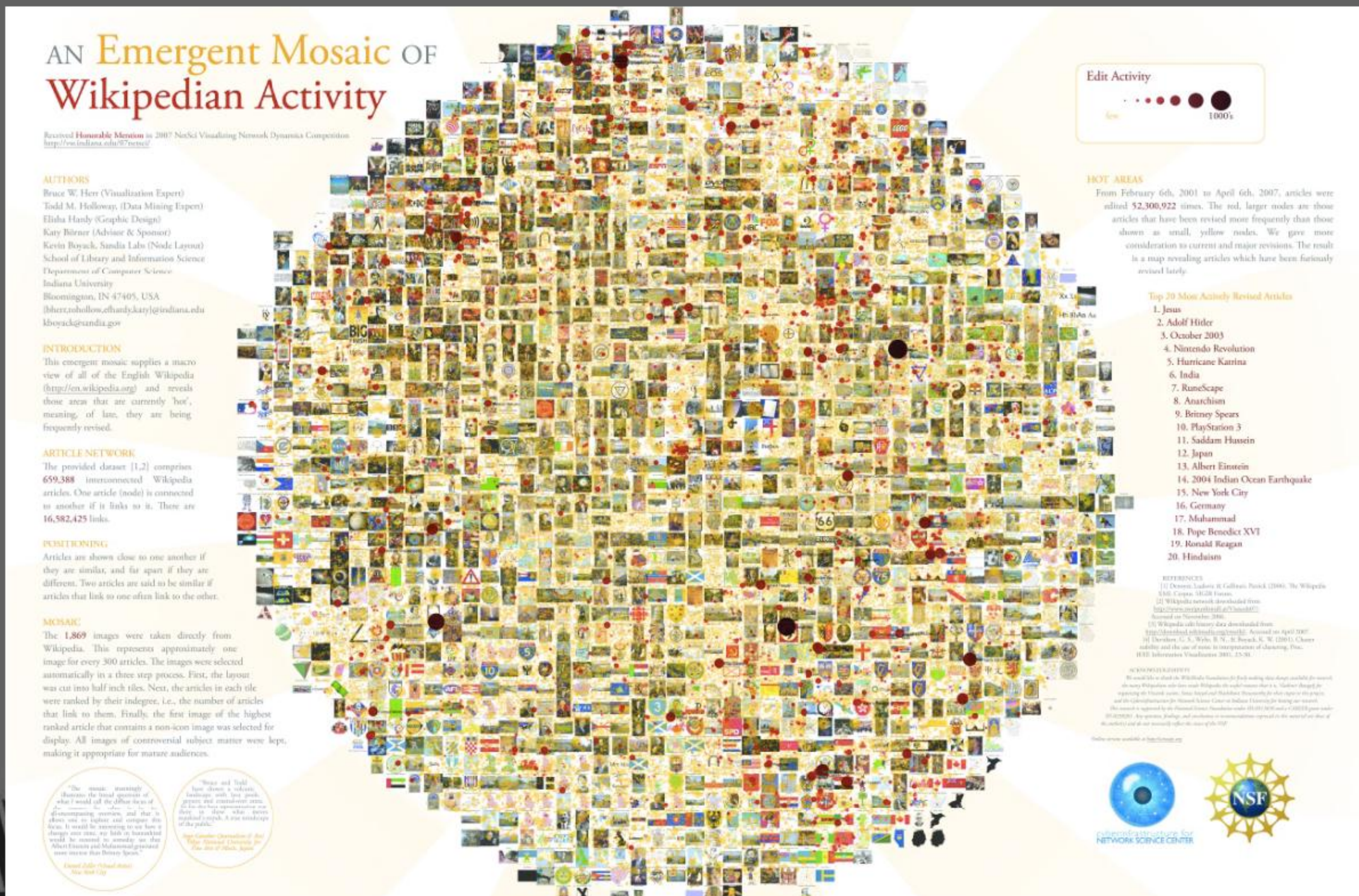


Graph-based Techniques



Zoomable UIs

http://gigapan.com/gigapans/4304



Zoomable UIs

http://gigapan.com/gigapans/4304

Distortion Techniques

Basic Idea: Distortion of the image to allow a visualization of larger amounts of data

An alternative to zoomable Uis (or complement)

Distortion Techniques

Basic Idea: Distortion of the image to allow a visualization of larger amounts of data

Simple:

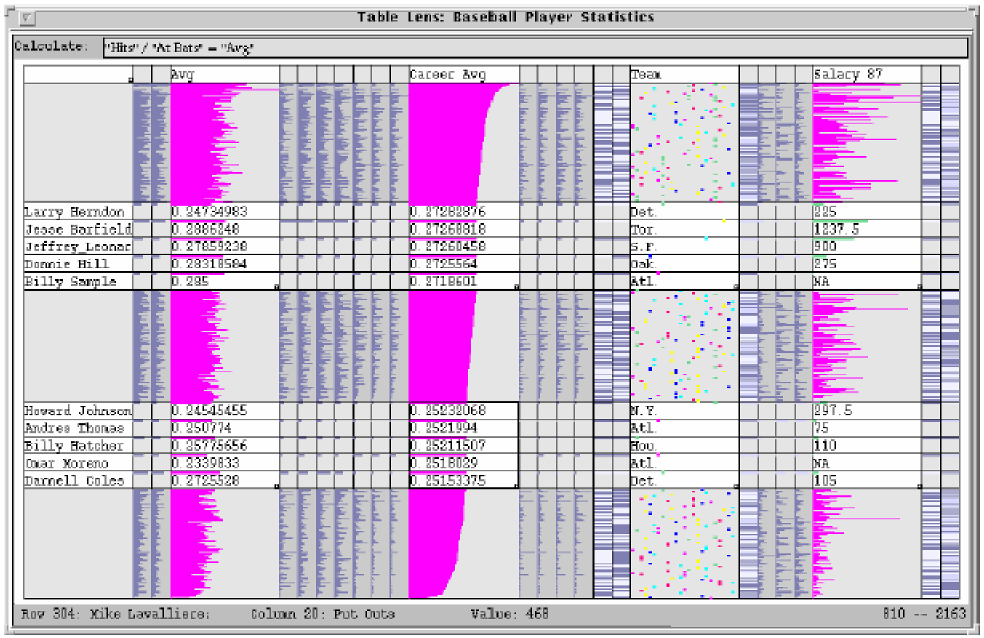
- Perspective Wall [MRC91]
- Bifocal Displays [SA 82]
- TableLens [RC94]
- Graph. Fisheye Views [Fur 86, SB94]
- Hyperbolic Repr. [LR94, LRP95]

Complex:

- Hyperbolic Repr. [LR94, LRP95]
- 3D-Hyperbolic Repr. [MB95]
- Hyperbox [AC91]

Table Lens

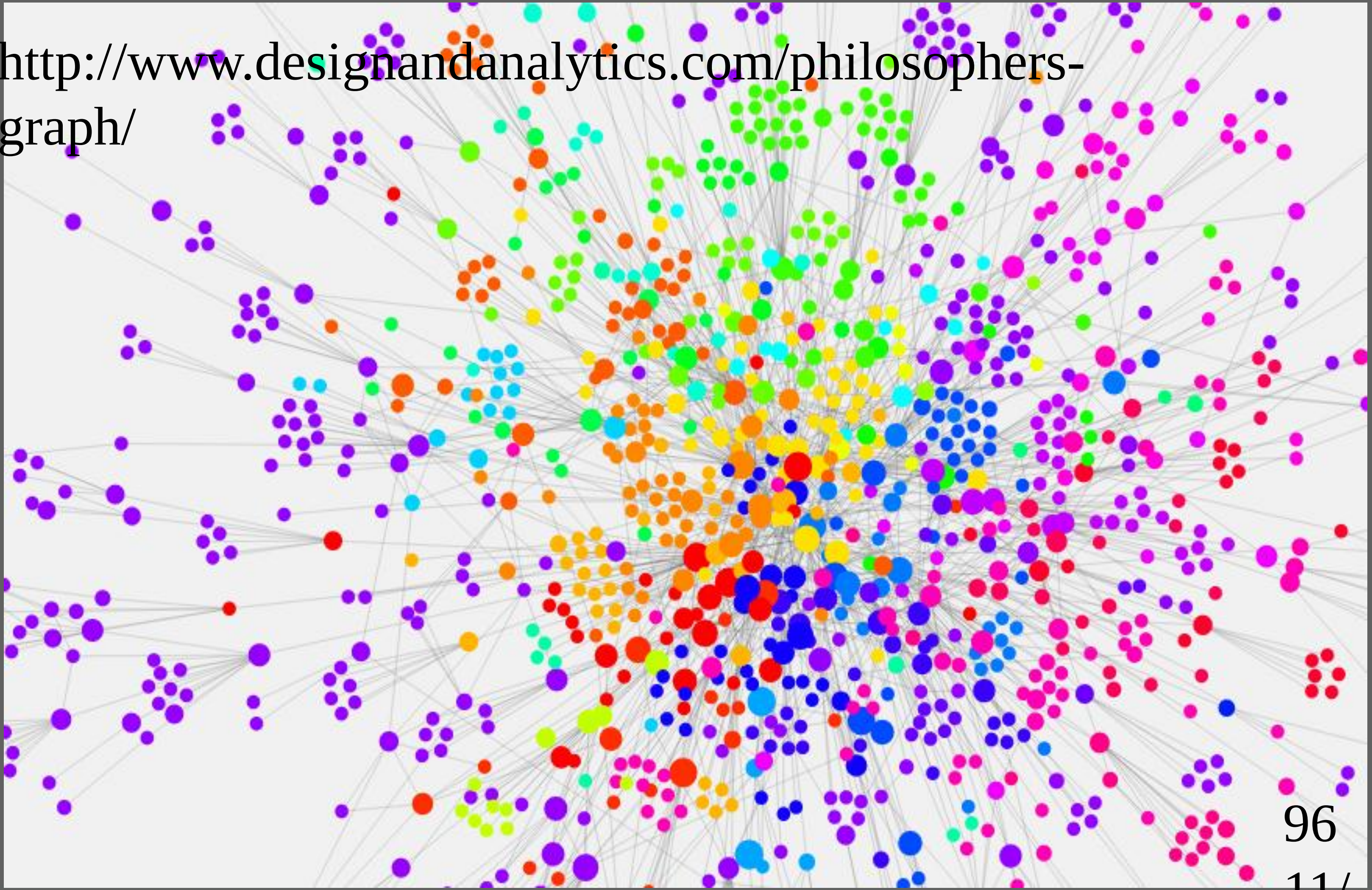
used by permission of R. Rao, Xerox PARC



visualization of a baseball database with a few rows being selected in full detail

➪ compact visualization of a table (spreadsheet / database) with the possibility of viewing portions of the table in more detail

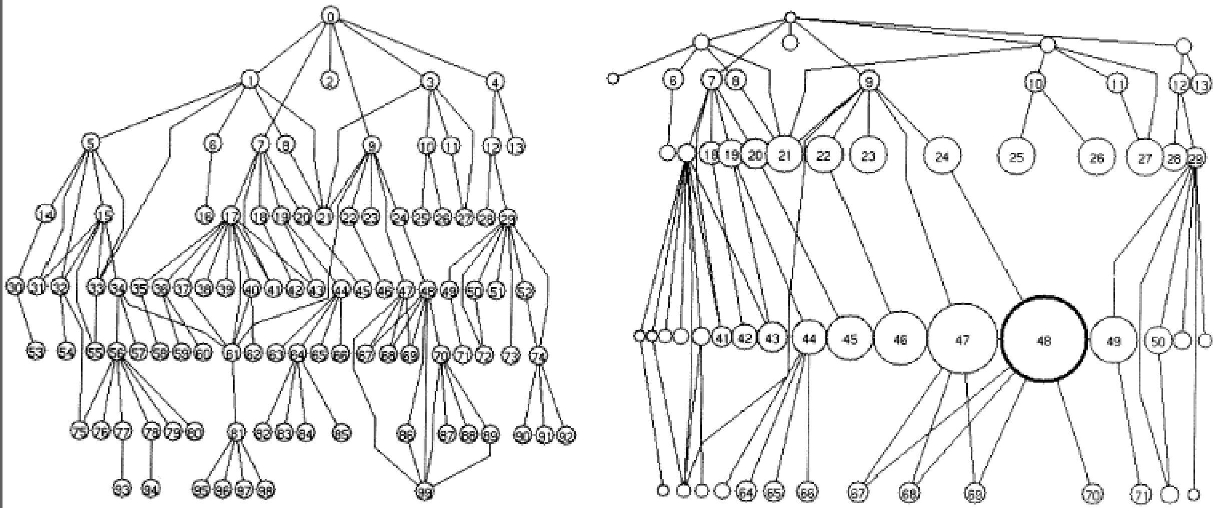
Networks of Information



http://www.designandanalytics.com/philosophers-graph/

Networks of Information – with zooming

Fisheye View

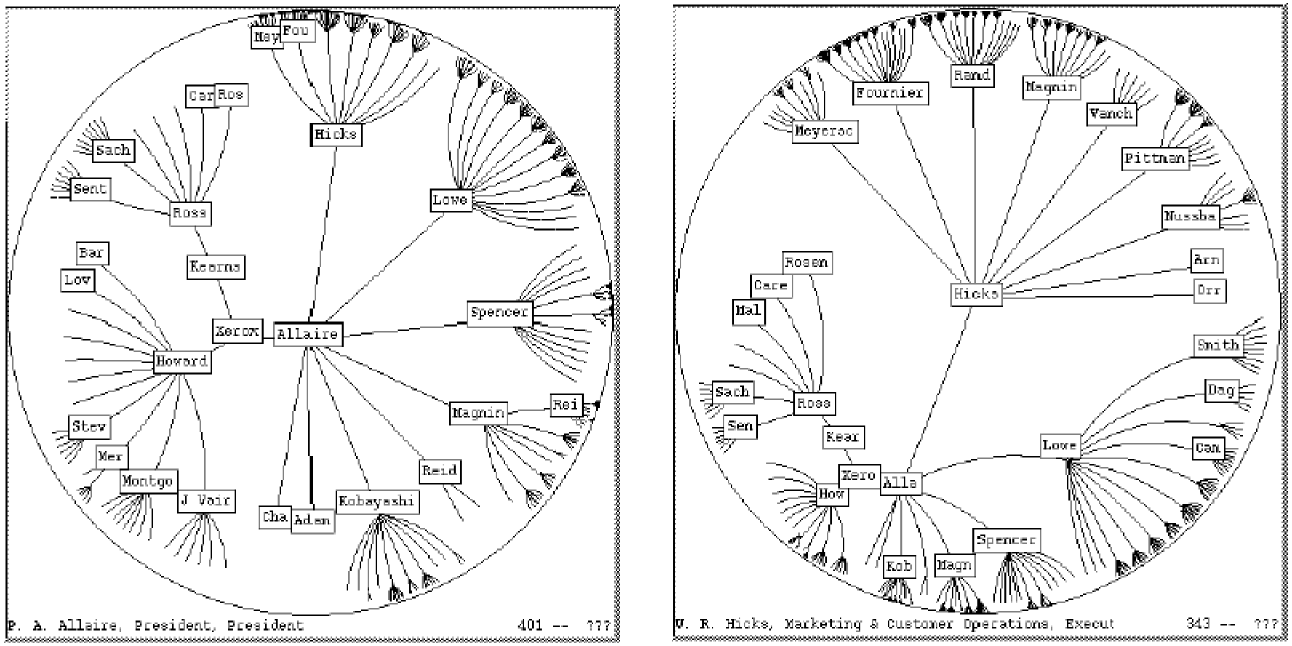


original graph

fisheye view of the graph

- graph visualization using a fisheye perspective
- shows an area of interest quite large and with detail and the other areas successively smaller and in less detail

Hyperbolic Tree



used by permission of R. Rao, Xerox PARC

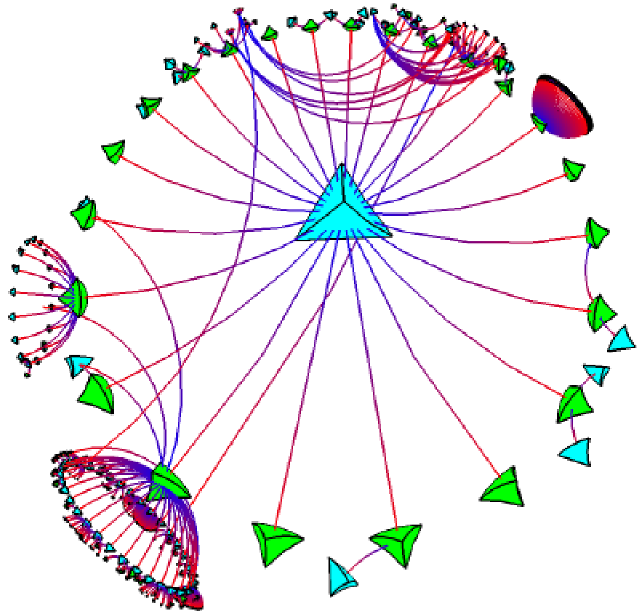
used by permission of R. Rao, Xerox PARC

visualization of a large organizational hierarchy

➪ visualization of a tree structure in hyperbolic space with different foci

3D Hyperbolic Representation

used by permission of T. Munzner, Stanford University



visualization
of a large number
of connected
web-pages

➪ visualization of a graph in 3D hyperbolic conetree-like representation

Summary

Make use of every pixel

Use color and location to break the data up and allow the viewer to easily filter

Give a sense of things and use zooming for detail

Add a dimension such as time or height for a key variable

Allow exploration through distortion, filtering, highlighting, and linking

Exploit hierarchy and connectivity

Product Tree Map of Benin (Exports) 2009 by Mcstol from

Applied Data Use: StepGreen

Capture

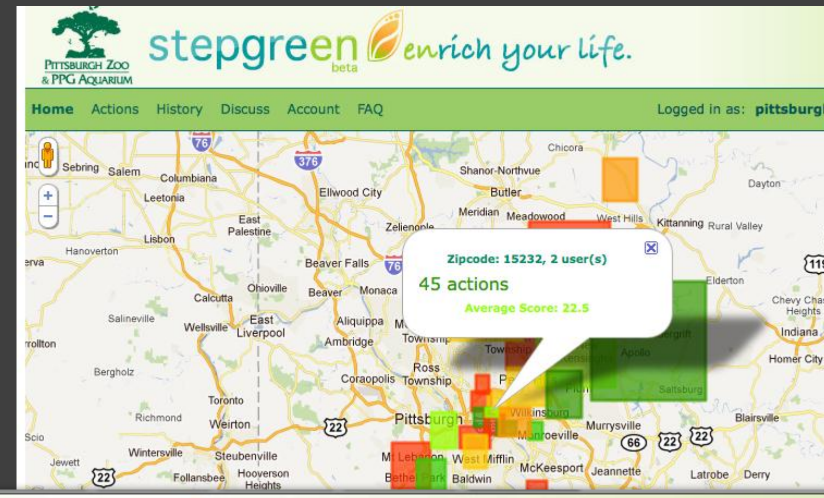
- Self reported data on behavior
- Motion & GPS
- Temperature & Energy

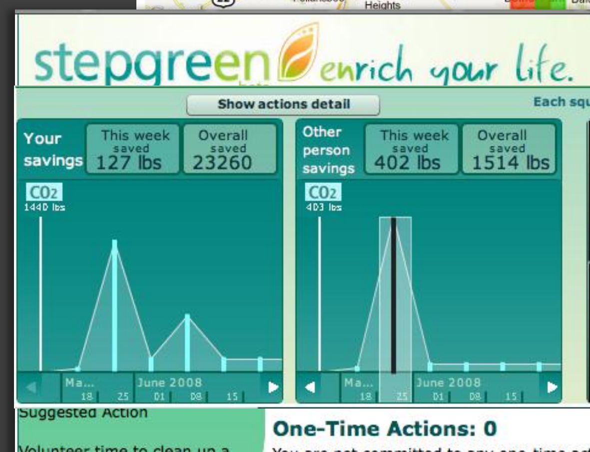
Know

- Green Actions
- Transportation choices
- Appliances in Use

Adapt and Act

- Visualize
- Expose
- …





Jenni

Applied Example: Ubigreen

Capture

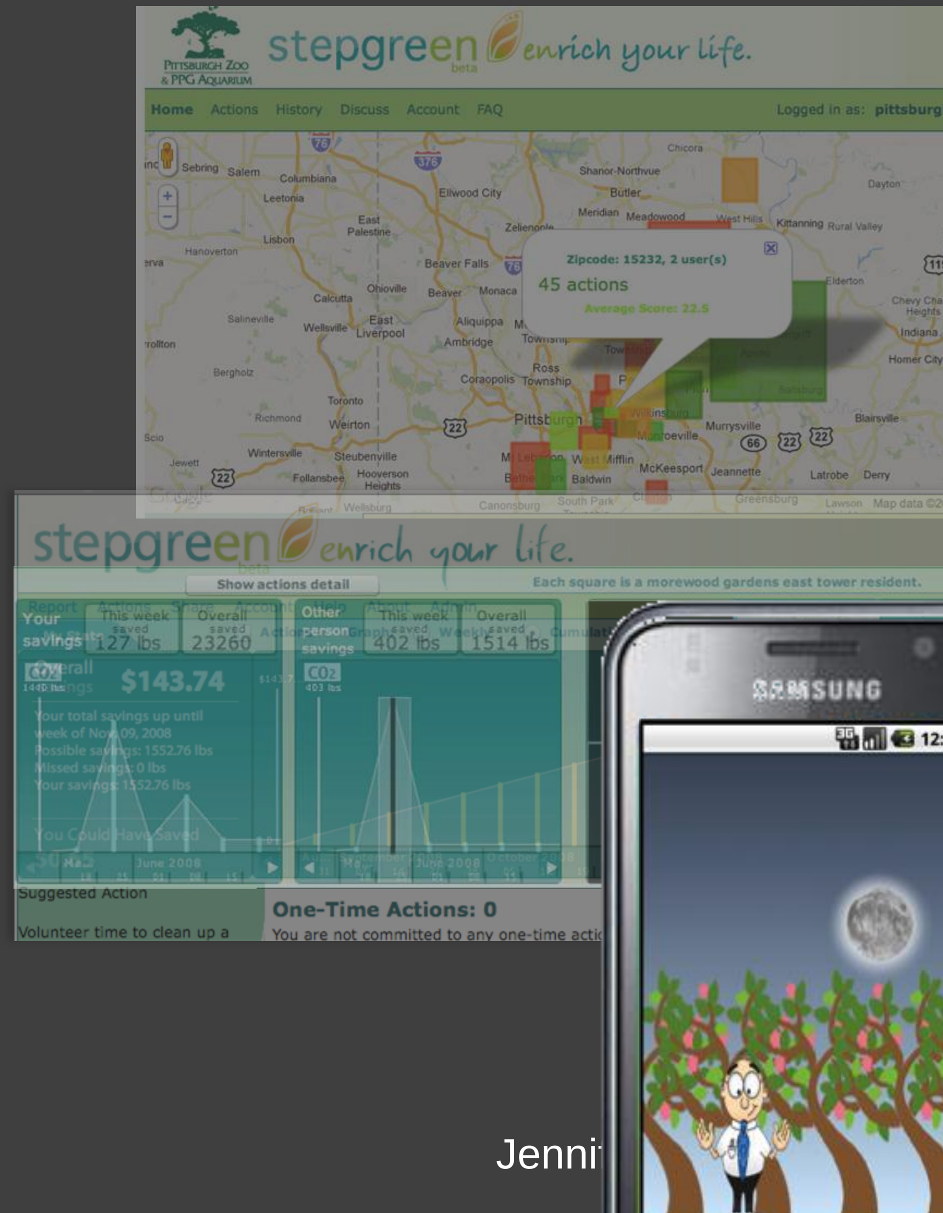
- Self reported data on behavior
- Motion & GPS
- Temperature & Energy

Know

- Green Actions
- Transportation choices
- Appliances in Use

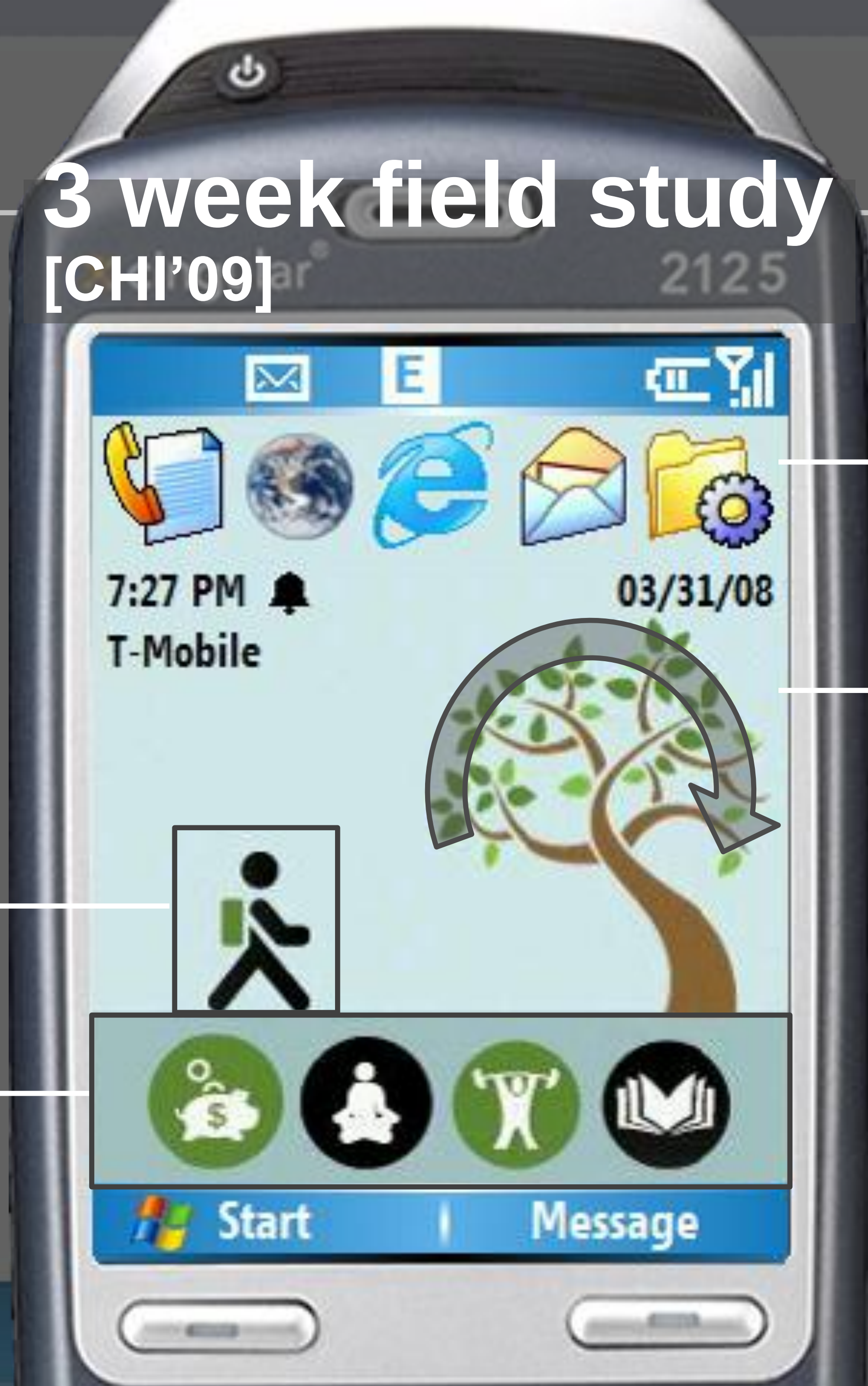
Adapt and Act

- Visualize
- Expose
- …



Jenni

3 week field study
[CHI’09]



Engagement

"It's omnipresent"

- Participant 9

"I want to have different stories every week … to maintain curiosity in the app"

- Participant 8

Real-life game

One participant complained that when a trip hadn't been automatically recorded, "I felt like I was being cheated out of my 'points'"

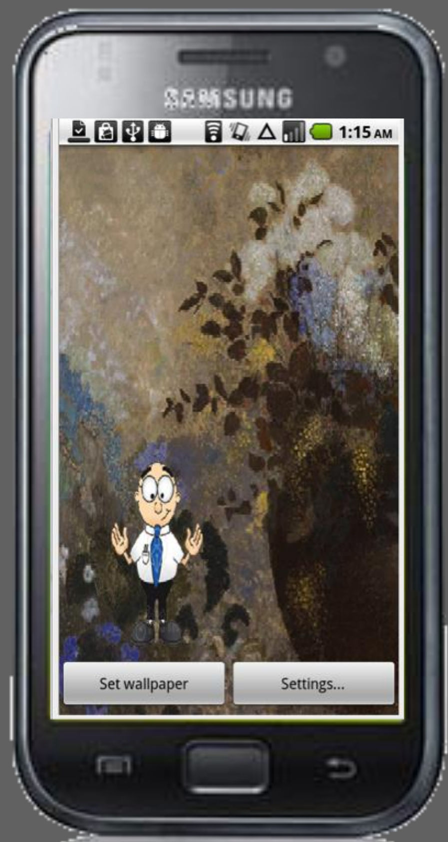
- Participant 15

Social

"Some people at work knew about the polar bear and every day they asked me about it. 'Did you get a seal today?'"

- Participant 14

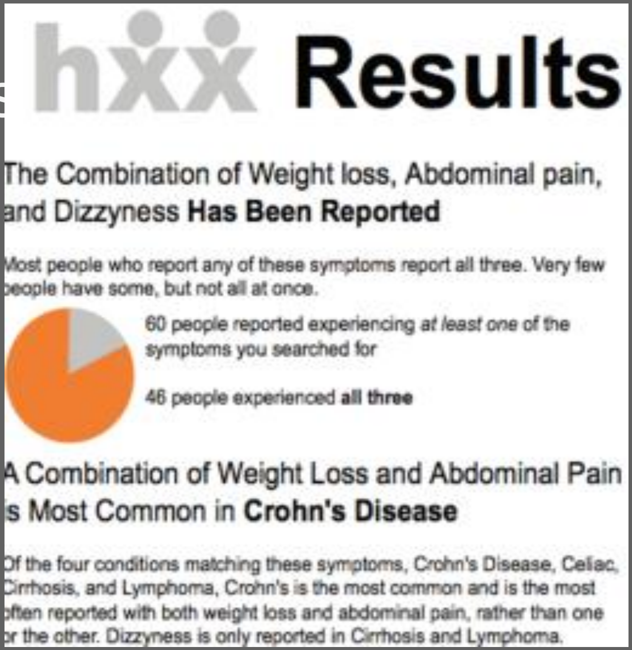
Longer Term, Real-World Deployments



Other Application Areas

Health

- Self reported data on symptoms and conditions
- Internet data: Extract argument structures & enhance search
- Forums: predict expertise, highlight time on site, *etc.*



Summary: Making Data Actionable

Data

- Collecting and Interpreting

Information

- Sensemaking

Knowledge

- Visualize
  Act
  Adapt

Value