The Art of Numbers

Alyssa A. Goodman • Harvard University

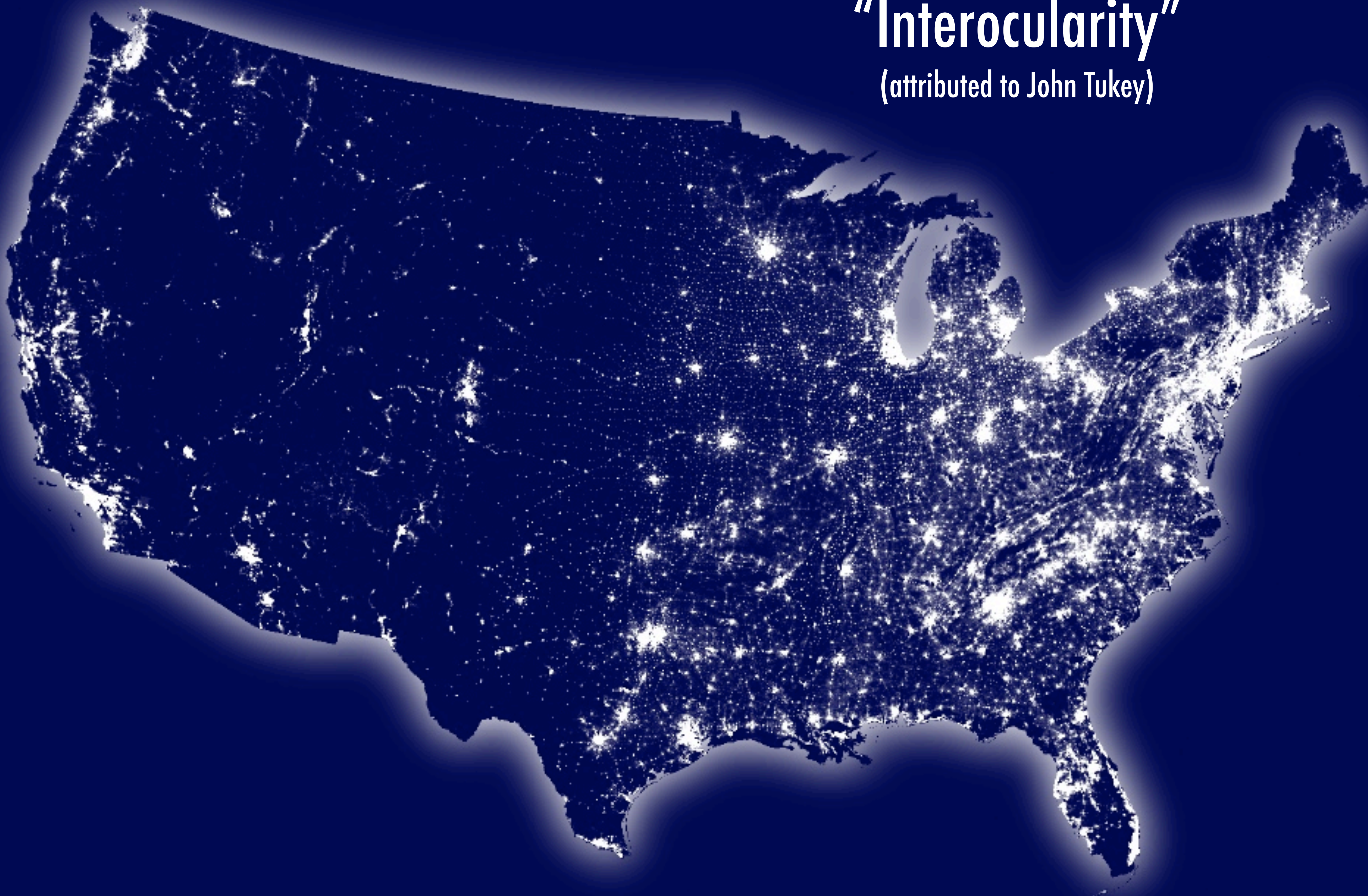
Relative Strengths

Pattern Recognition
Creativity

Calculations

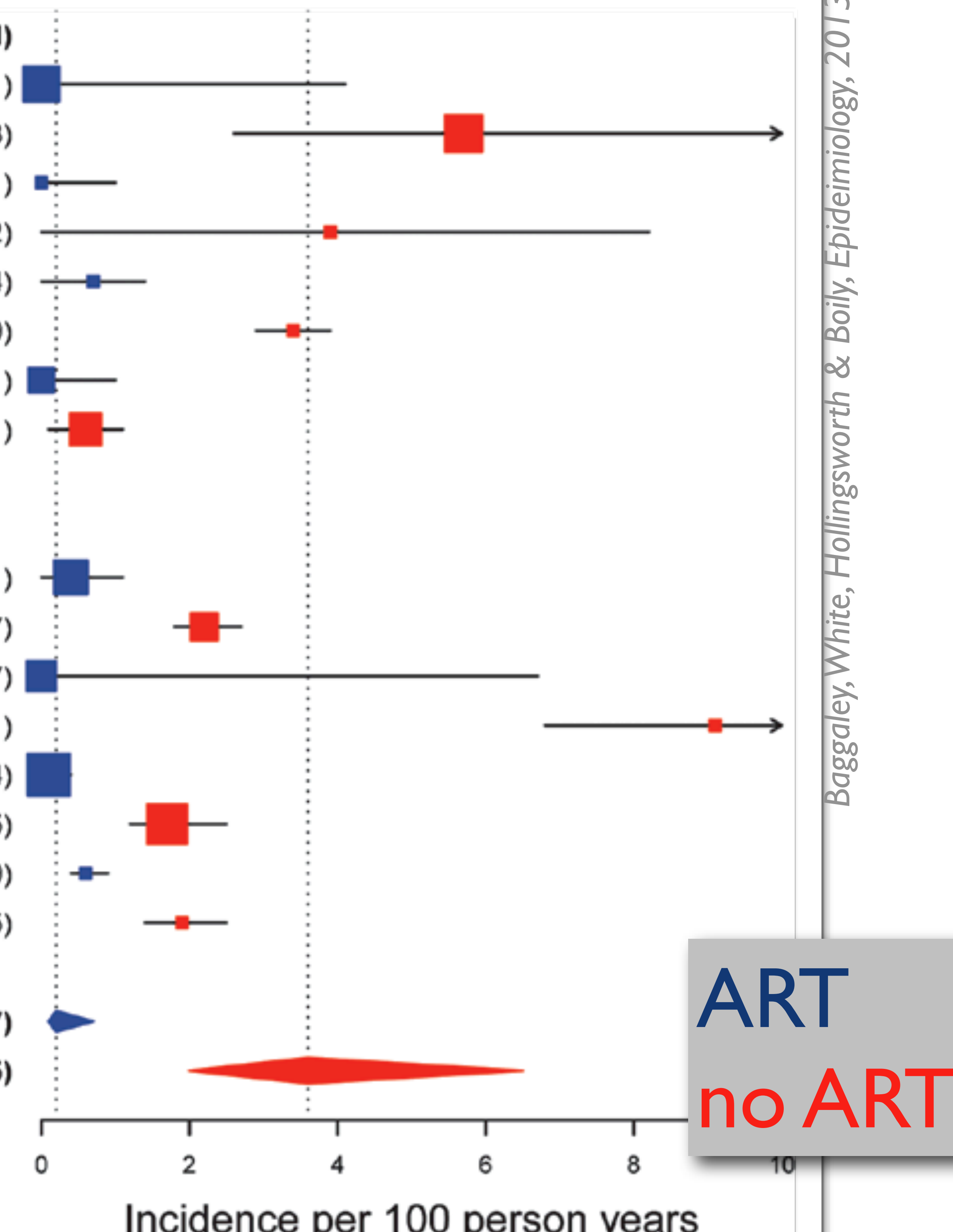
"Interocularity"

(attributed to John Tukey)



Epidemiologists' Forest Plot = Tukey's"Box Plot"

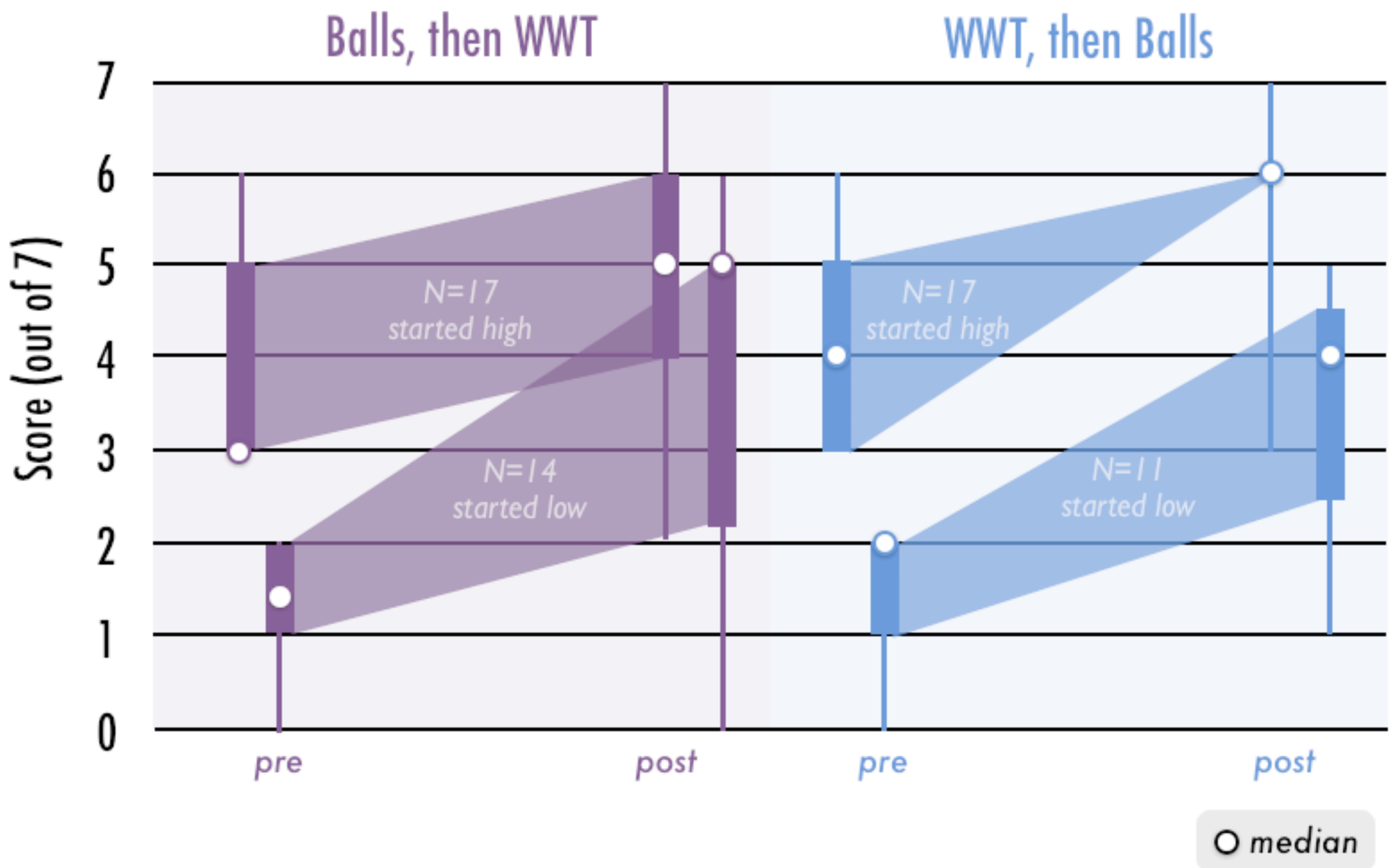
Study	x	n	py	Inc	(95%CI)
Melo et al (2008)[18]	0	41	90.4	0	(0–4.1)
	6	52	106	5.7	(2.6–11.8)
Watera et al (2009)[19]	0		19.5	0	(0–1)
	3		76.9	3.9	(0–8.2)
Sullivan et al (2009)[20]	4		575	0.7	(0–1.4)
	171		5033	3.4	(2.9–3.9)
Del Romero et al (2010)[22]	0	144	417	0	(0–1)
	5	341	863	0.6	(0.1–1.1)
Wang et al (2010)[23]	66	1369	NR	NR	
	18	558	NR	NR	
Donnell et al (2010)[24]	1	349	273	0.4	(0–1.1)
	102	3032	4558	2.2	(1.8–2.7)
Reynolds et al (2011)[25]	0	32	53.6	0	(0–6.7)
	42		459.4	9.1	(6.8–12.1)
Cohen et al (2011)[3]	1	886	1585.3	0.1	(0–0.4)
	27	877	1567.3	1.7	(1.2–2.5)
Baeten* et al (2011)[67]	31		4893	0.6	(0.4–0.9)
	47		2444	1.9	(1.4–2.5)
Summary ART	**72**	**>2821**		**0.2**	**(0.1–0.7)**
Summary No ART	**374**	**>4860**		**3.6**	**(2–6.5)**



Baggaley, White, Hollingsworth & Boily, Epidemiology, 2013

Forest plot summary of HIV-1 incidence rate estimates per heterosexual partnership for antiretroviral therapy-stratified studies, with 95% confidence intervals.

"Box Plot"

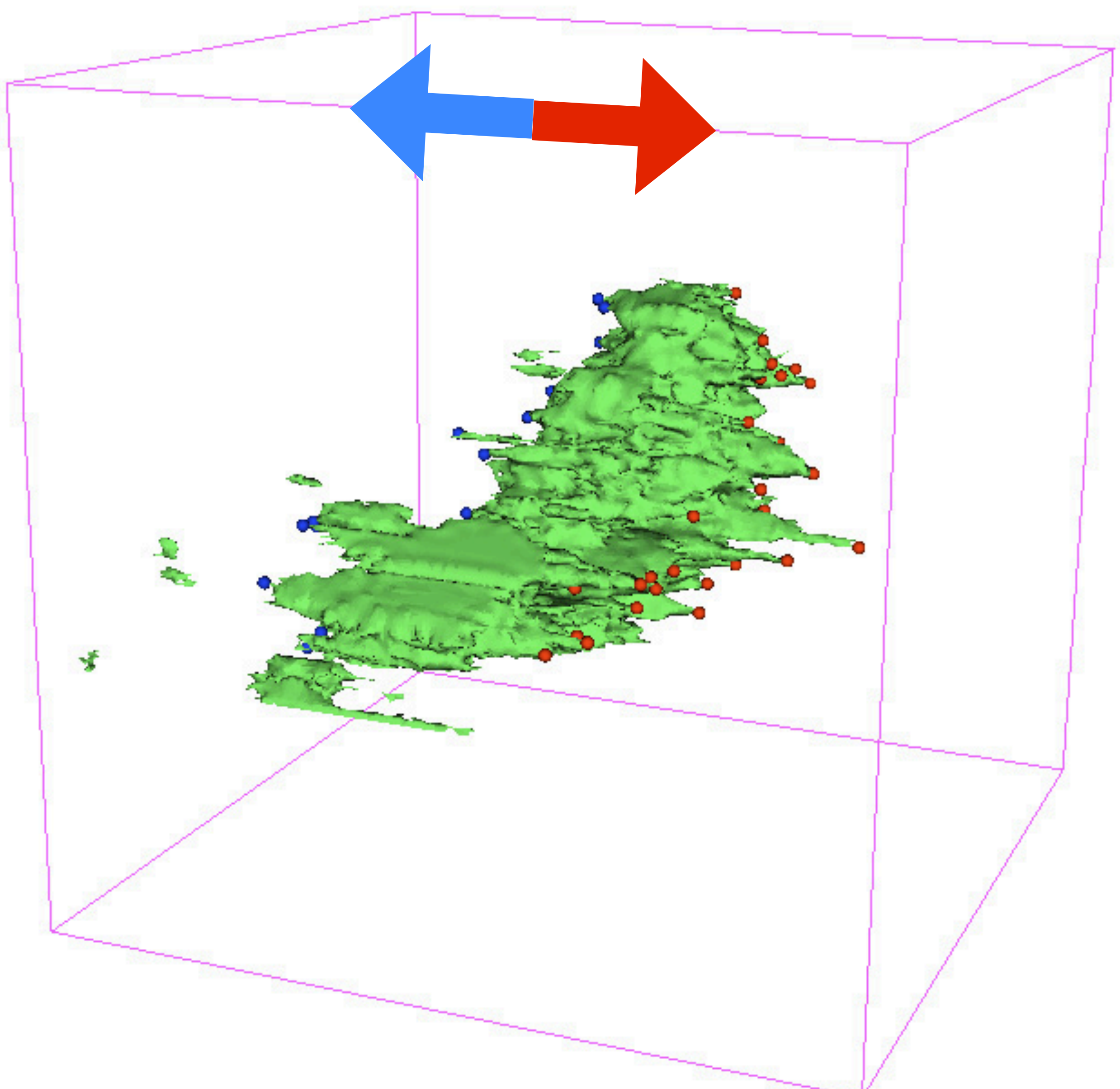


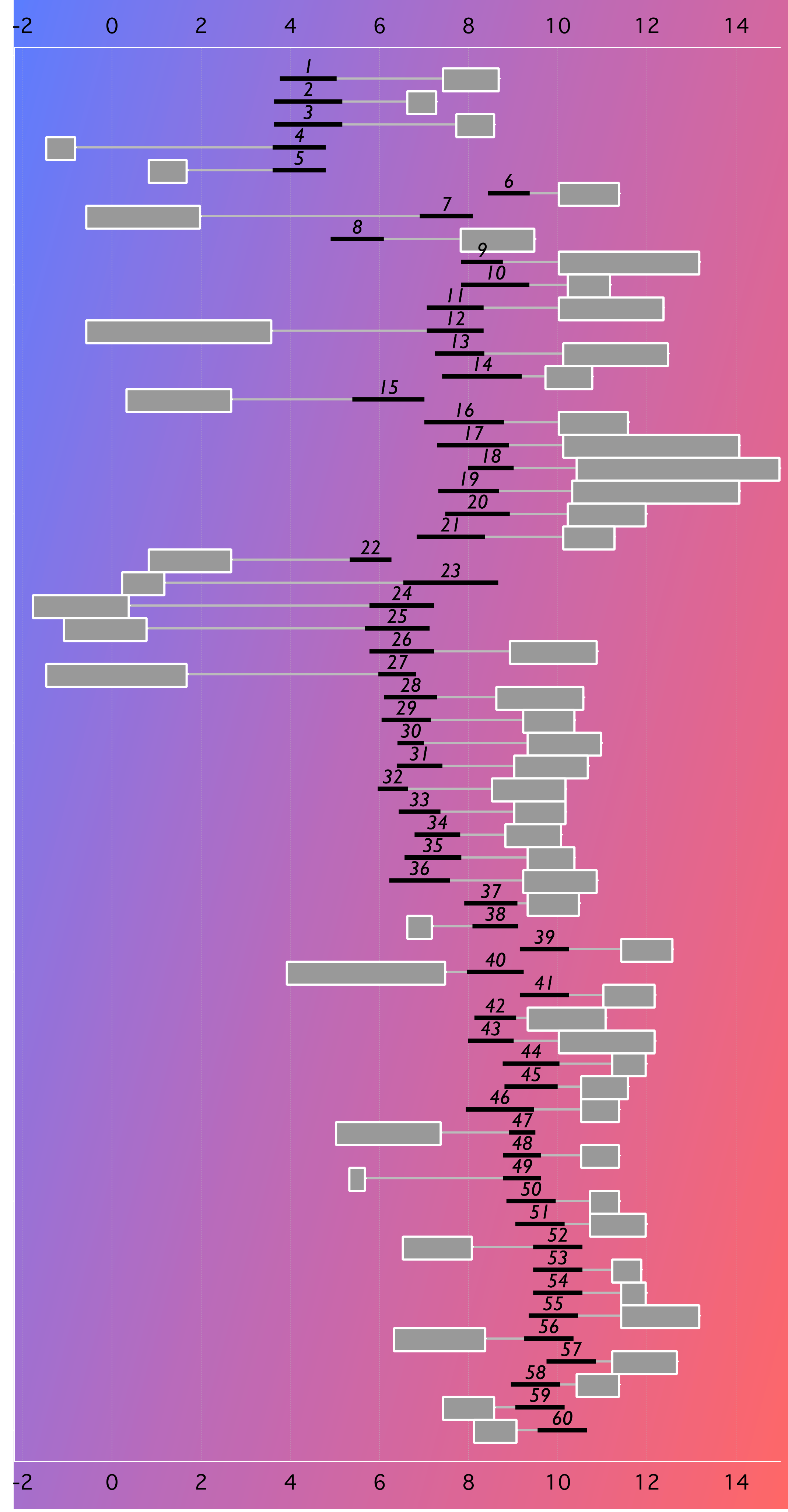
added to static display: data about time dimension and *groupings* for each of multiple trials

plot from Udomprasert et al. 2013

COMPLETE Perseus Outflow Candidates

"Box Plot?"





Arce et al. 2010; inset based on Offner et al. 2011

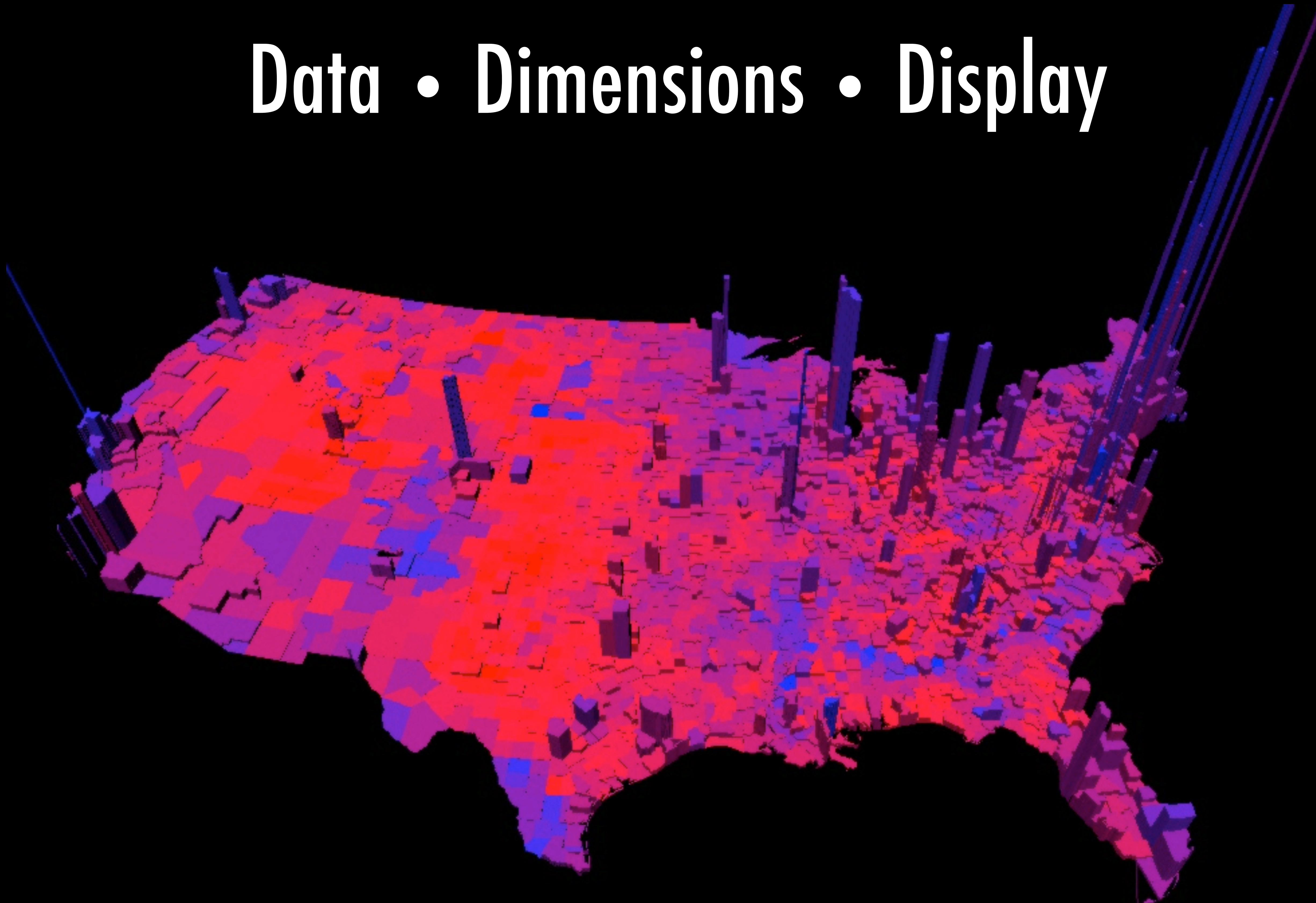
Data • Dimensions • Display

Linked Views

Data • Dimensions • Display

Linked Views

Data • Dimensions • Display

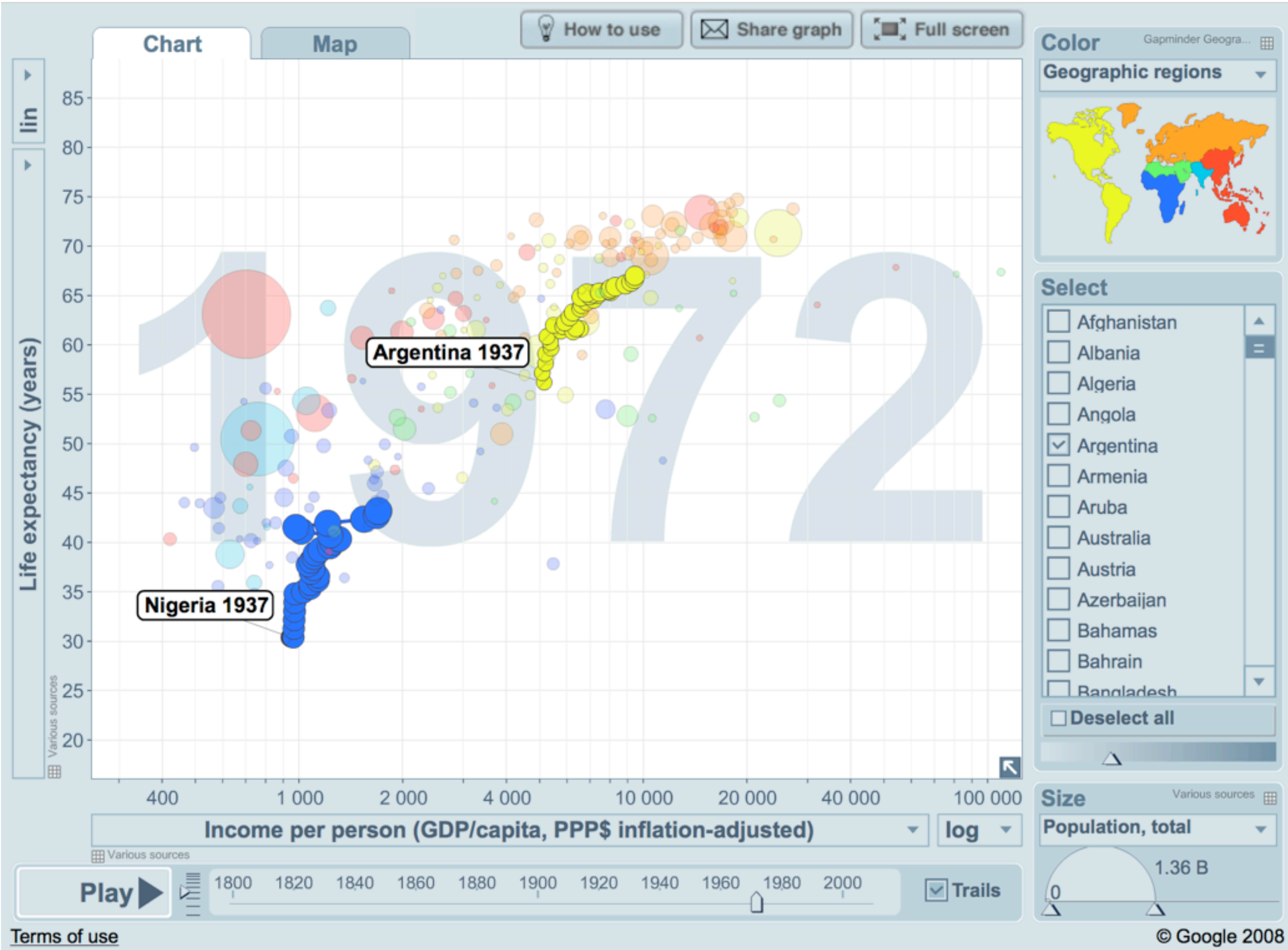


Robert J. Vanderbei

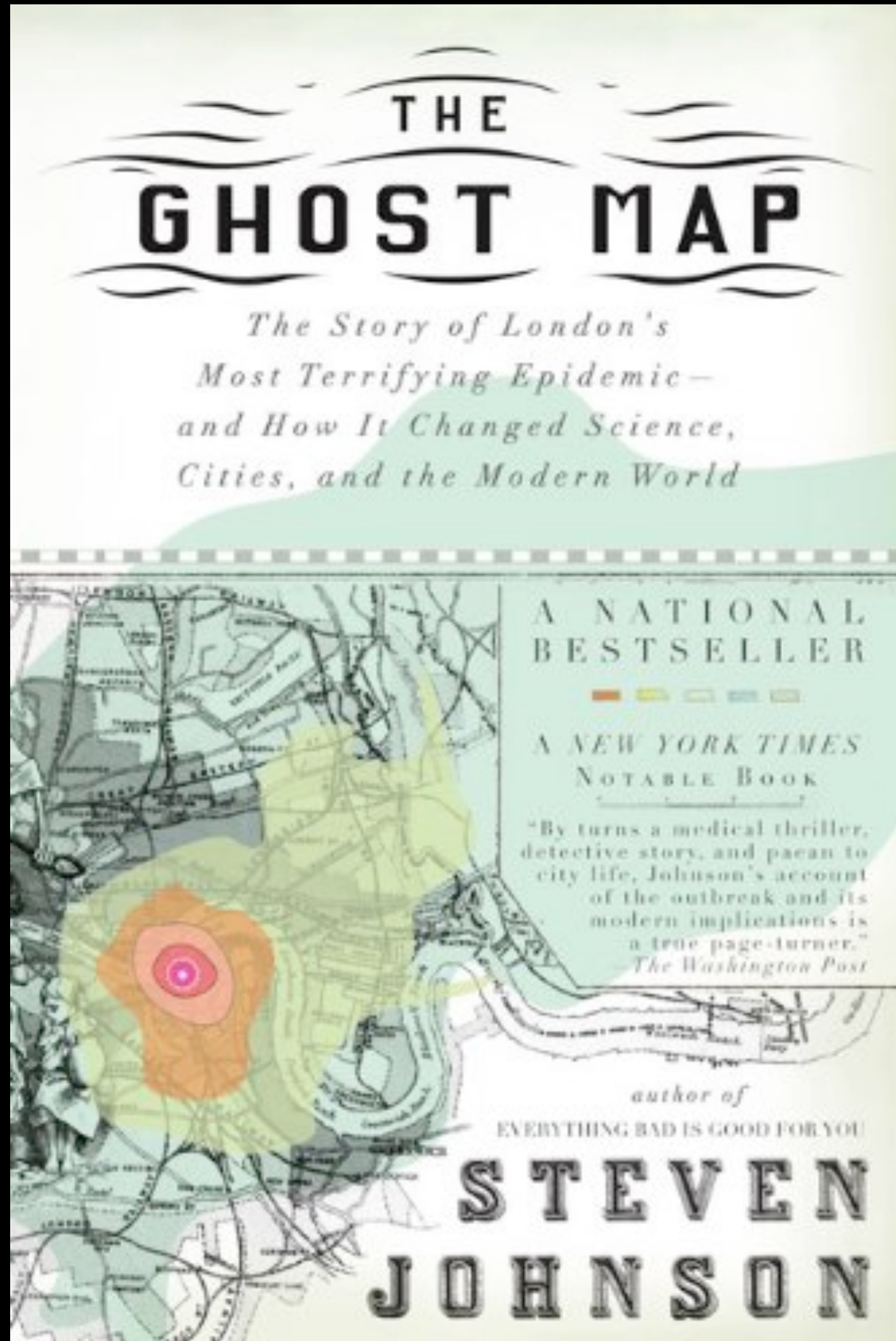
www.princeton.edu/~rvdb/JAVA/election2004/

Linked Views

Linked Views

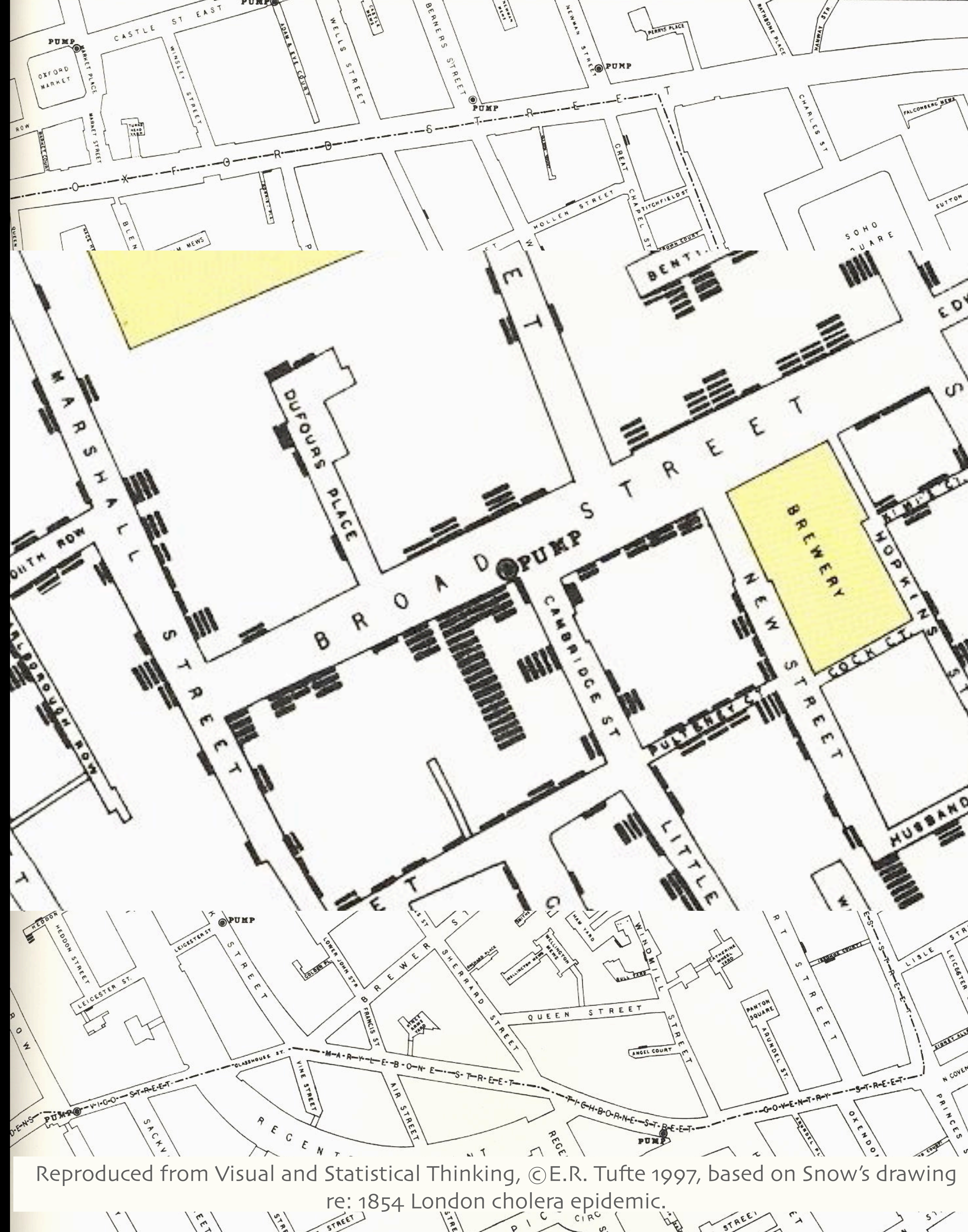


GAPMINDER for a fact-based world view


THE
GHOST MAP
The Story of London's
Most Terrifying Epidemic—
and How It Changed Science,
Cities, and the Modern World
A NATIONAL
BESTSELLER
A NEW YORK TIMES
NOTABLE BOOK
"By turns a medical thriller, detective story, and paean to city life, Johnson's account of the outbreak and its modern implications is a true page-turner."
The Washington Post
author of
EVERYTHING BAD IS GOOD FOR YOU
STEVEN
JOHNSON

Epidemiology

(in 1854)



Reproduced from Visual and Statistical Thinking, ©E.R. Tufte 1997, based on Snow's drawing re: 1854 London cholera epidemic.

Epidemiology
(in 1854)

Displaying
"high-dimensional" data

with

"multi-functioning
graphical elements"

Reproduced from Visual and Statistical Thinking, ©E.R. Tufte 1997, based on Snow's drawing re: 1854 London cholera epidemic.

Snow couldn't "interact" with the map but we should be able to,
with the right data linkages, and choice of dimensions & display.

DATA BLOG

Facts are sacred

Previous | Blog home

John Snow's cholera map of London recreated

What would John Snow's famous cholera map look like on a modern map of London, using modern mapping tools? The map changed what we know about germs and disease - and created a new way of looking at the world. With the help of mapping tool CartoDB and using the Stamen style maps, this is how it looks with larger circles representing more deaths. What do you think?

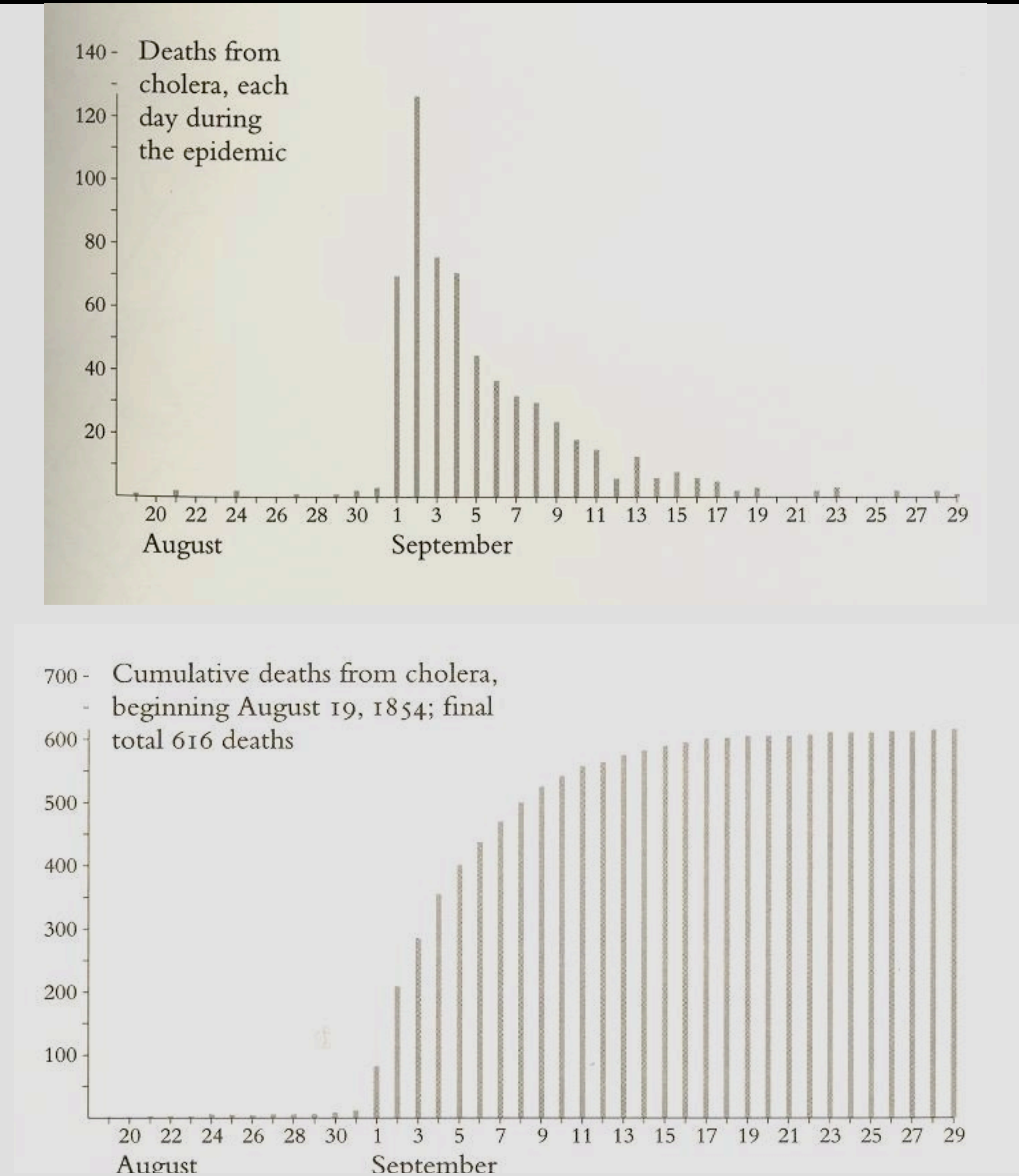
• Debate and download the data behind this map

Simon Rogers
guardian.co.uk, Friday 15 March 2013 05.29 EDT

Water pumps   Cholera deaths

What about time?

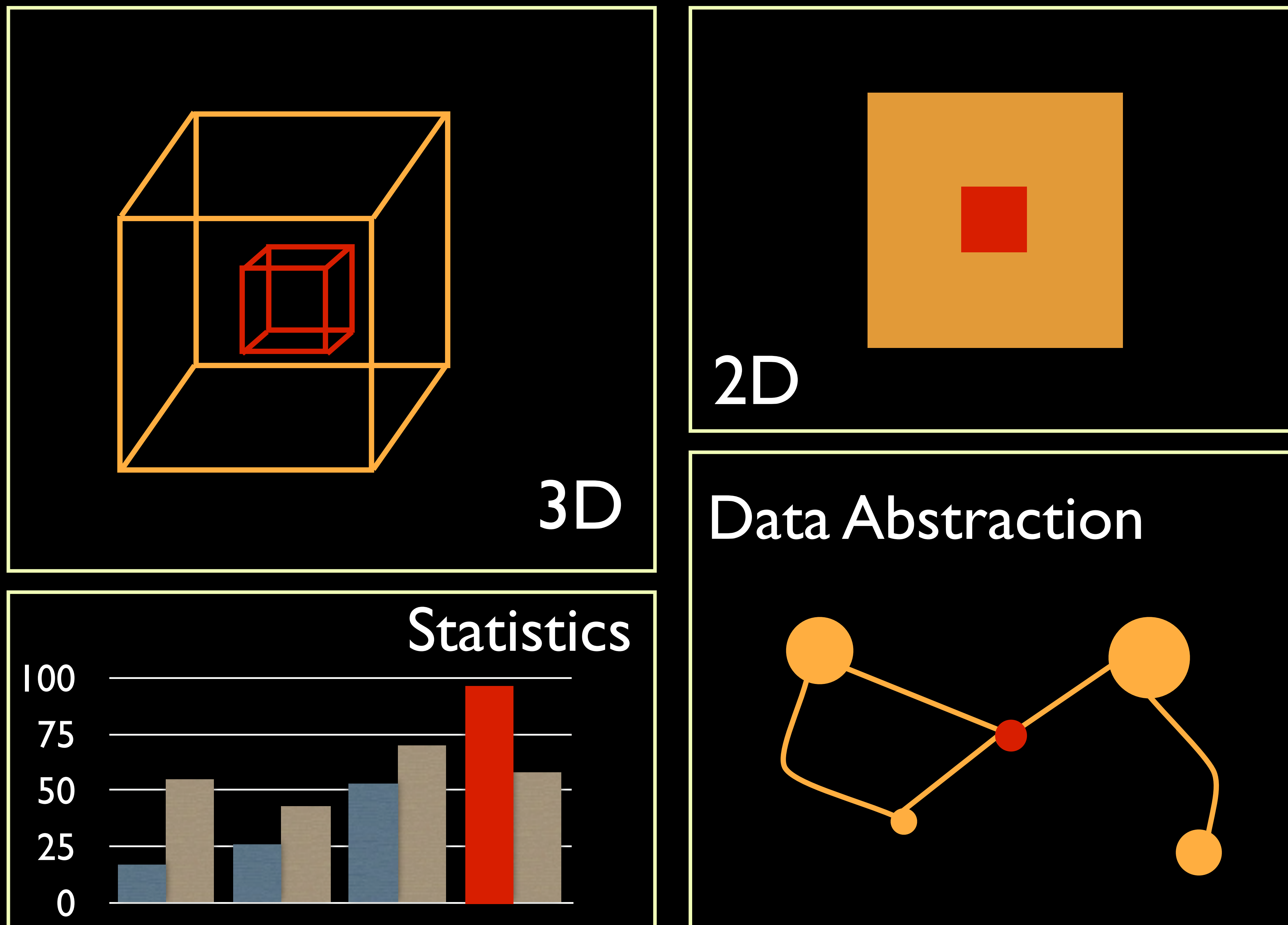
(these plots are not linked to spatial information–but they should be!)



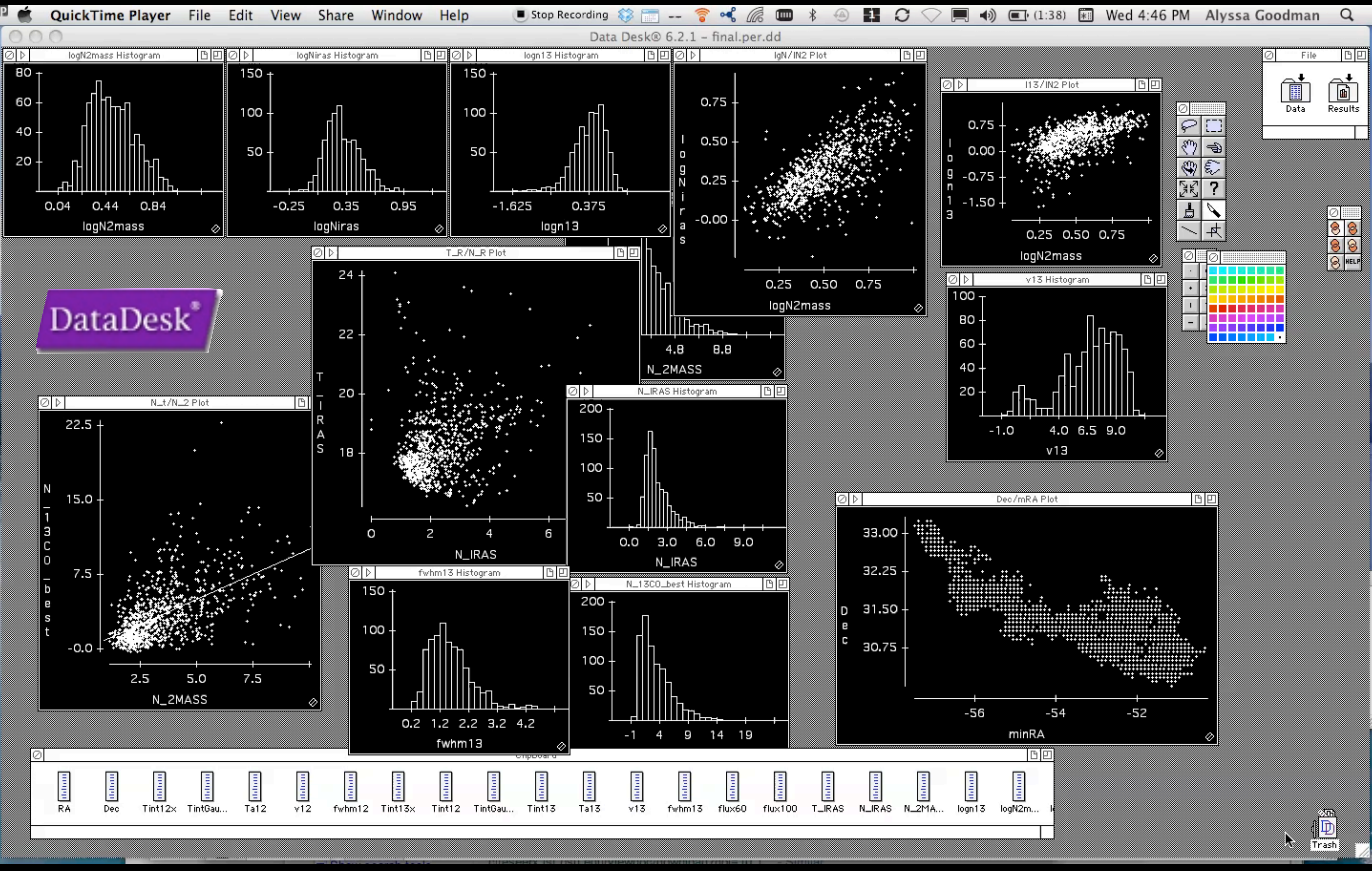
Note: Baby Lewis (index case) got sick on 28 August.

Reproduced from Visual and Statistical Thinking, ©E.R. Tufte 1997, based on Snow's data.

“Linked Views”



“Linked Views”: DataDesk (est. 1986)



JOHN TUKEY'S LEGACY

PRIM-9 → DataDesk® →

PRIM-H

XGobi → GGobi →

RGGobi

Spotfire® →

Polaris → +ableau SOFTWARE

1970 1980 1990 2000 2010

glue

multidimensional data exploration

What is glue?

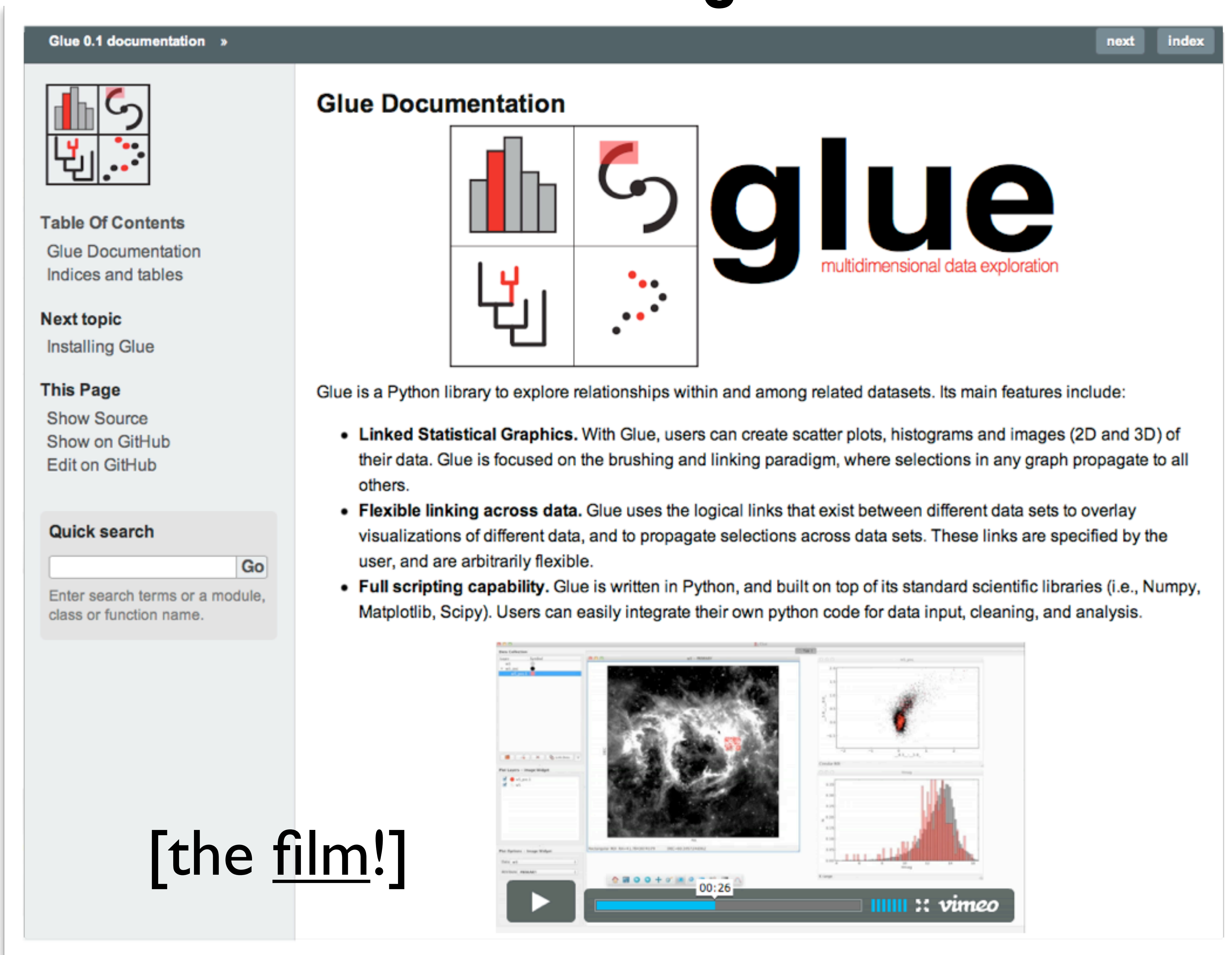

Glue 0.1 documentation »
next index

Table Of Contents
Glue Documentation
Indices and tables

Next topic
Installing Glue

This Page
Show Source
Show on GitHub
Edit on GitHub

Quick search
Go
Enter search terms or a module, class or function name.

Glue Documentation

glue
multidimensional data exploration

Glue is a Python library to explore relationships within and among related datasets. Its main features include:

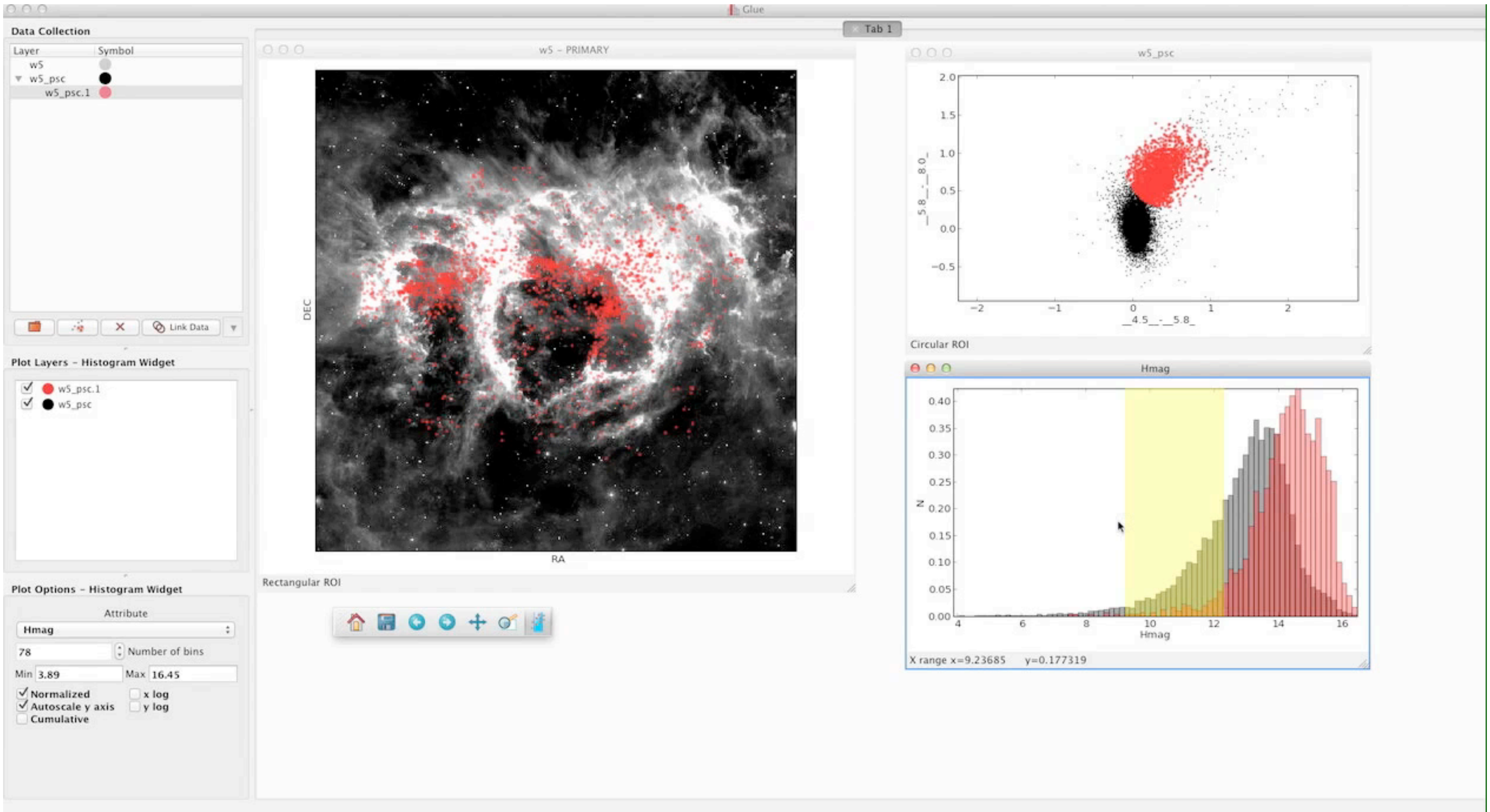
- **Linked Statistical Graphics.** With Glue, users can create scatter plots, histograms and images (2D and 3D) of their data. Glue is focused on the brushing and linking paradigm, where selections in any graph propagate to all others.
- **Flexible linking across data.** Glue uses the logical links that exist between different data sets to overlay visualizations of different data, and to propagate selections across data sets. These links are specified by the user, and are arbitrarily flexible.
- **Full scripting capability.** Glue is written in Python, and built on top of its standard scientific libraries (i.e., Numpy, Matplotlib, Scipy). Users can easily integrate their own python code for data input, cleaning, and analysis.

00:26
vimeo

[the film!]

*Glue collaboration: **Beaumont**, Borkin, Goodman, Pfister, Robitaille*

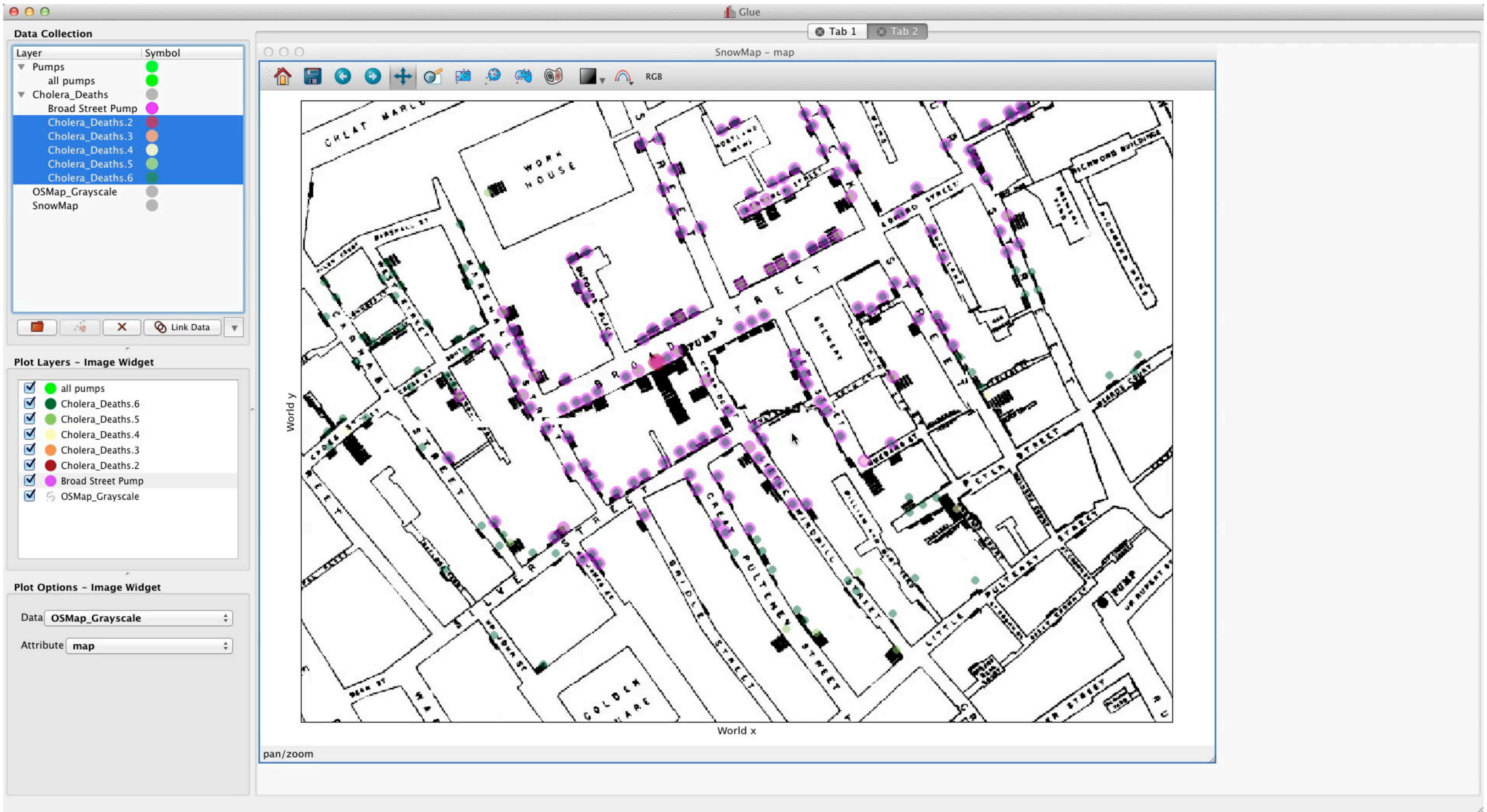
What is glue?



Glue collaboration: ***Beaumont****, Borkin, Goodman, Pfister, Robitaille*

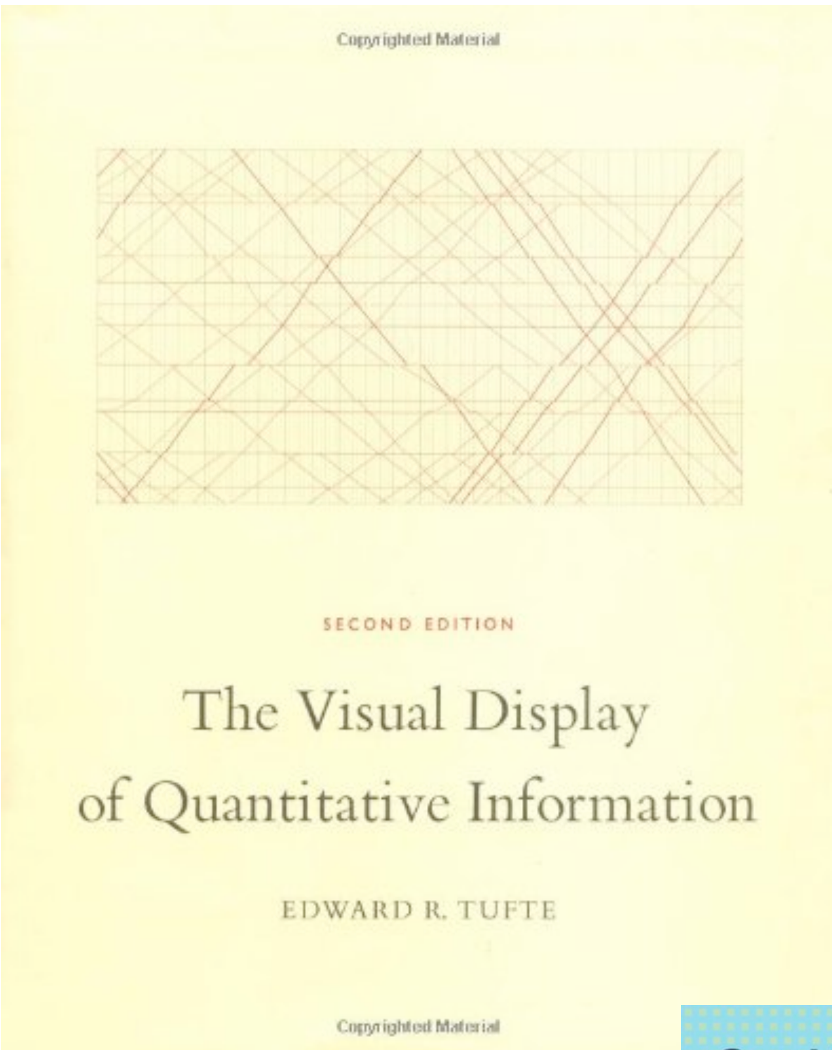
python™

Epidemiology, in Glue (used as a "linked view"GIS)



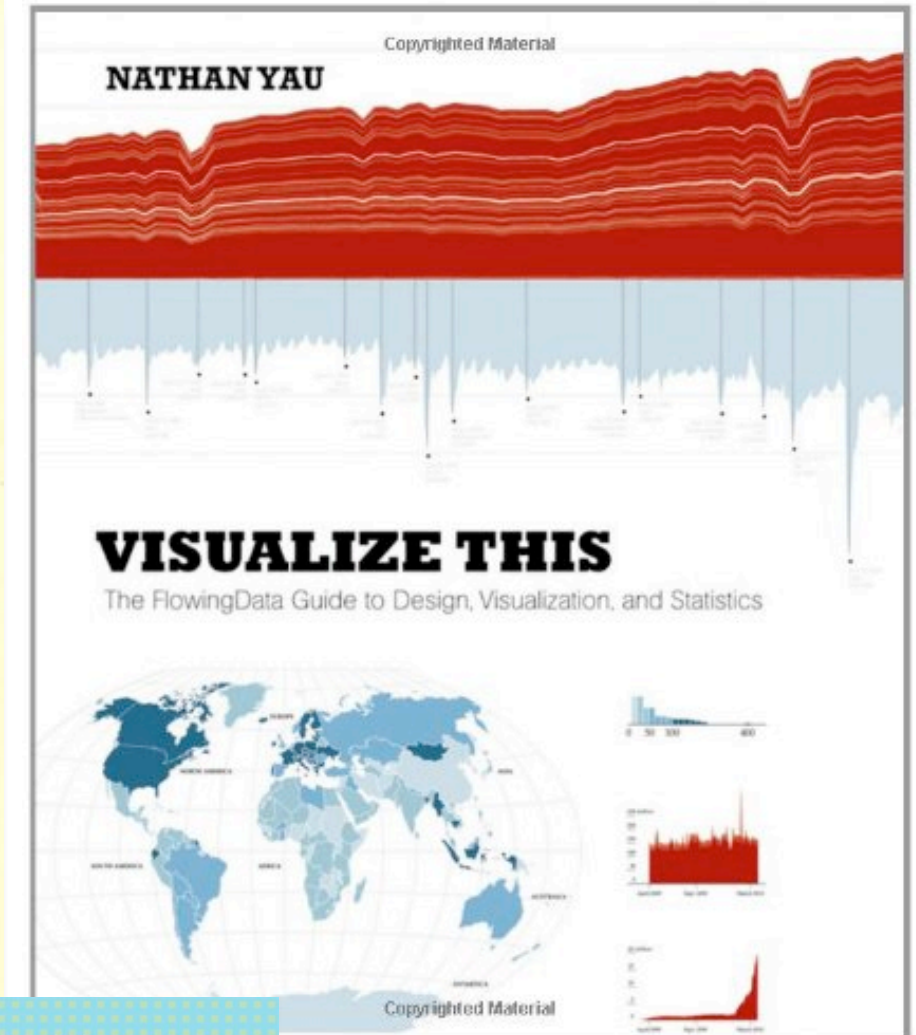
video courtesy Chris Beaumont, Glue lead

The Classic



1983

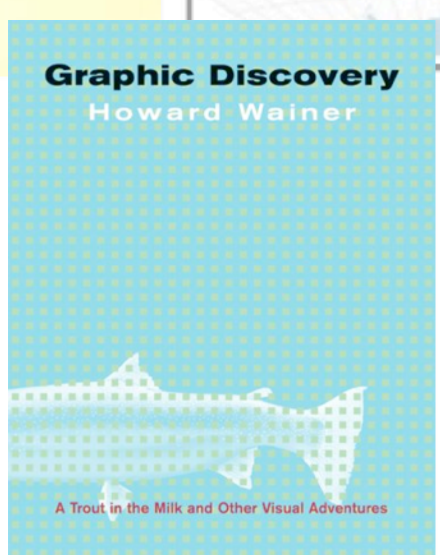
Modern "How-to"



2011

Case Studies

2012



A story about "conventions"

re: Physical Scientists & Education Researchers, excerpted from "Visual Strategies," by Frankel & DePace.

Appendix D: Likert Scale Survey results from Clarke Pilot

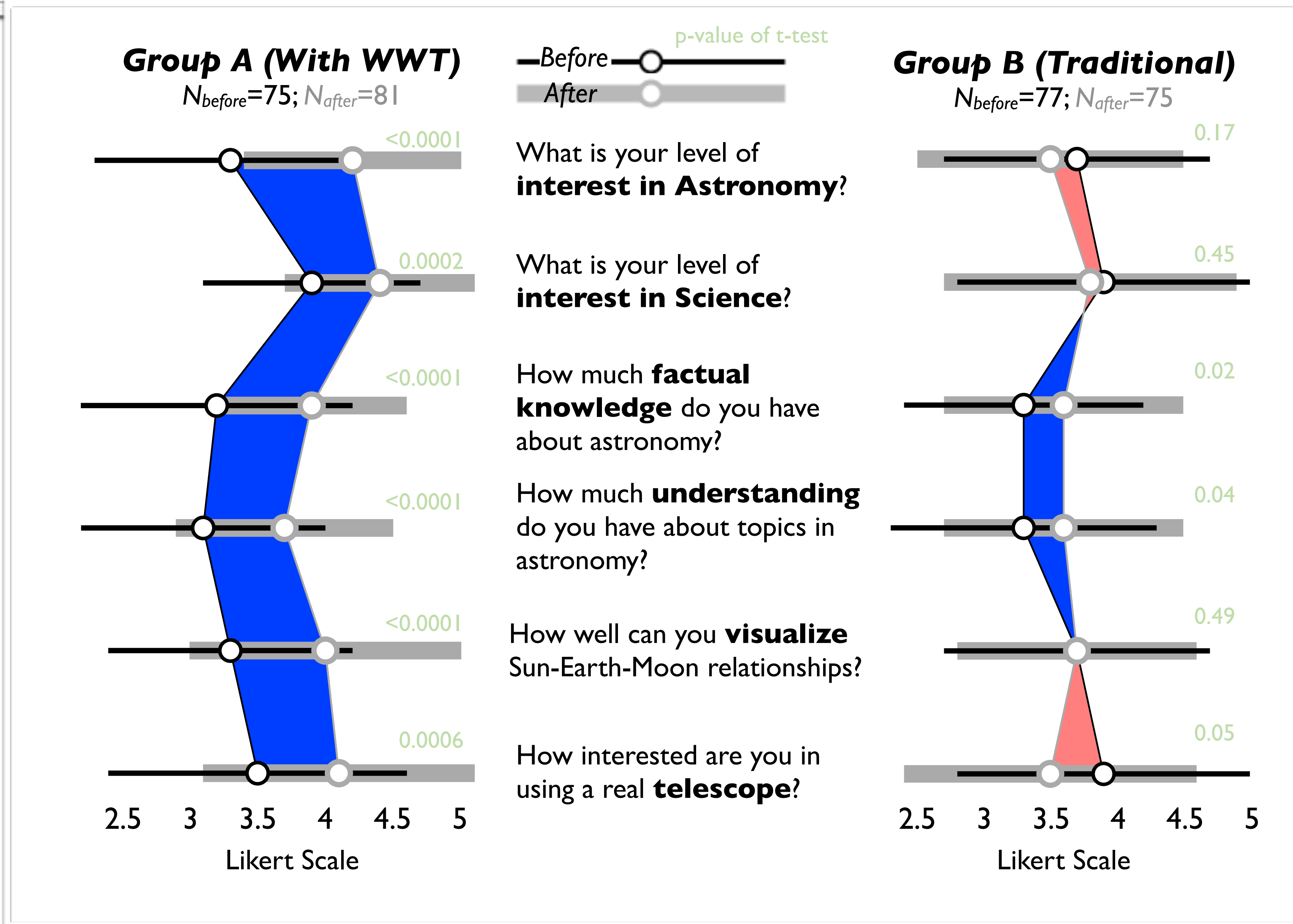
Detailed Summary of the pre-test post-test Likert Scale Surveys administered to a group of students who used WWT and a group who did not. Gains that are boldfaced are statistically significant and have a t-test p-value < 0.05.

Likert Scale Questions (1=low; 5=high)	Group A (with WWT)				**Group B (without WWT)**			
	mean (stdev)		gain	t-test p-value	mean (stdev)		gain	t-test p-value
	before N=75	**after** N=81			**before** N=77	**after** N=75		
What is your level of **interest in Astronomy**?	3.3 (1.0)	4.2 (0.8)	**0.9**	<0.0001	3.7 (1.0)	3.5 (1.0)	-0.2	0.17
What is your level of **interest in Science**?	3.9 (0.8)	4.4 (0.7)	**0.5**	0.0002	3.9 (1.1)	3.8 (1.1)	-0.1	0.45
How much **factual knowledge** do you have about astronomy?	3.2 (1.0)	3.9 (0.7)	**0.7**	<0.0001	3.3 (0.9)	3.6 (0.9)	**0.3**	0.02
How much **understanding** do you have about topics in astronomy?	3.1 (0.9)	3.7 (0.8)	**0.6**	<0.0001	3.3 (1.0)	3.6 (0.9)	**0.3**	0.04
How well can you **visualize** Sun-Earth-Moon relationships?	3.3 (0.9)	4.0 (1.0)	**0.7**	<0.0001	3.7 (1.0)	3.7 (0.9)	0	0.49
How interested are you in using a real **telescope**?	3.5 (1.1)	4.1 (1.0)	**0.6**	0.0006	3.9 (1.1)	3.5 (1.1)	**-0.4**	0.05

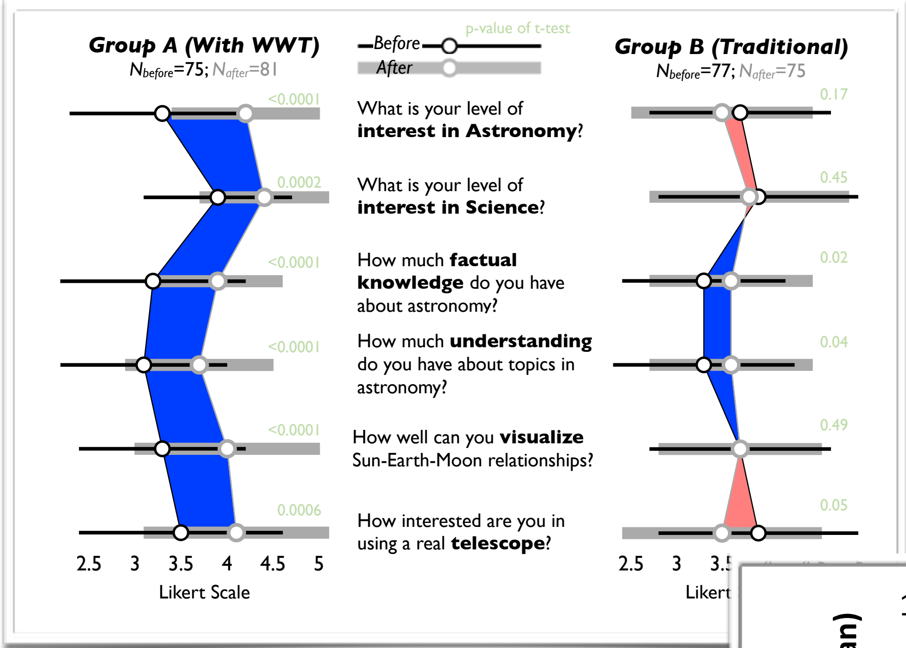
Appendix D: Likert Scale Survey results from Clarke Pilot

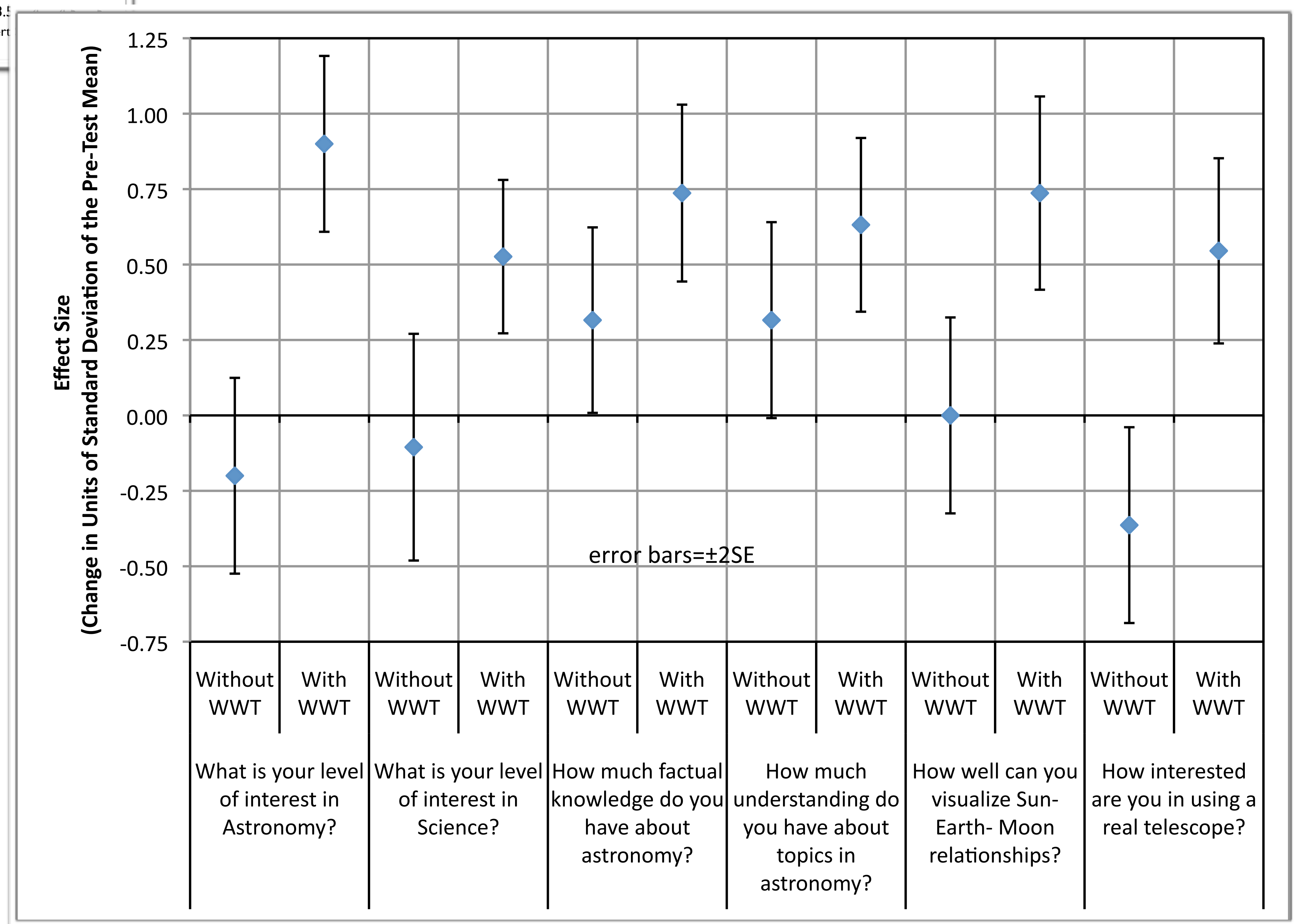
Detailed Summary of the pre-test post-test Likert Scale Surveys administered to a group of students who used WWT and a group who did not. Gains that are boldfaced are statistically significant and have a t-test p-value < 0.05.

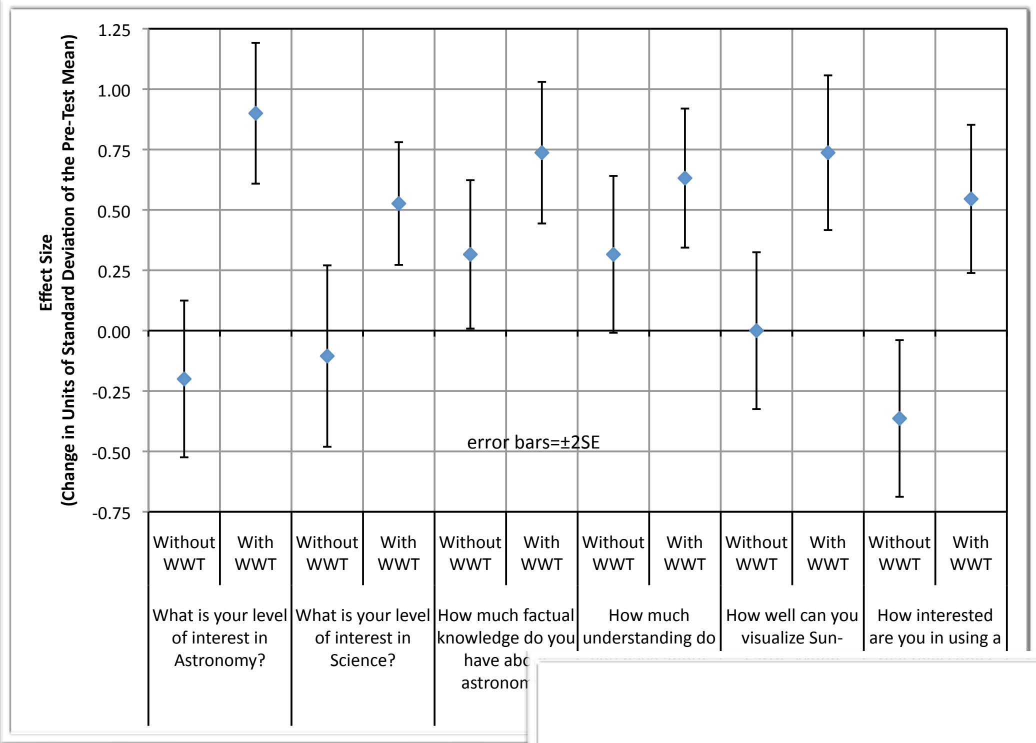
Likert Scale Questions (1=low; 5=high)	Group A (with WWT)				Group B (without WWT)			
	mean (stdev) before N=75	mean (stdev) after N=81	gain	t-test p-value	mean (stdev) before N=77	mean (stdev) after N=75	gain	t-test p-value
What is your level of **interest in Astronomy**?	3.3 (1.0)	4.2 (0.8)	**0.9**	<0.0001	3.7 (1.0)	3.5 (1.0)	-0.2	0.17
What is your level of **interest in Science**?	3.9 (0.8)	4.4 (0.7)	**0.5**	0.0002	3.9 (1.1)	3.8 (1.1)	-0.1	0.45
How much **factual knowledge** do you have about astronomy?	3.2 (1.0)	3.9 (0.7)	**0.7**	<0.0001	3.3 (0.9)	3.6 (0.9)	**0.3**	0.02
How much **understanding** do you have about topics in astronomy?	3.1 (0.9)	3.7 (0.8)	**0.6**	<0.0001	3.3 (1.0)	3.6 (0.9)	**0.3**	0.04
How well can you **visualize** Sun-Earth-Moon relationships?	3.3 (0.9)	4.0 (1.0)	**0.7**	<0.0001	3.7 (1.0)	3.7 (0.9)	0	0.49
How interested are you in using a real **telescope**?	3.5 (1.1)	4.1 (1.0)	**0.6**	0.0006	3.9 (1.1)	3.5 (1.1)	**-0.4**	0.05

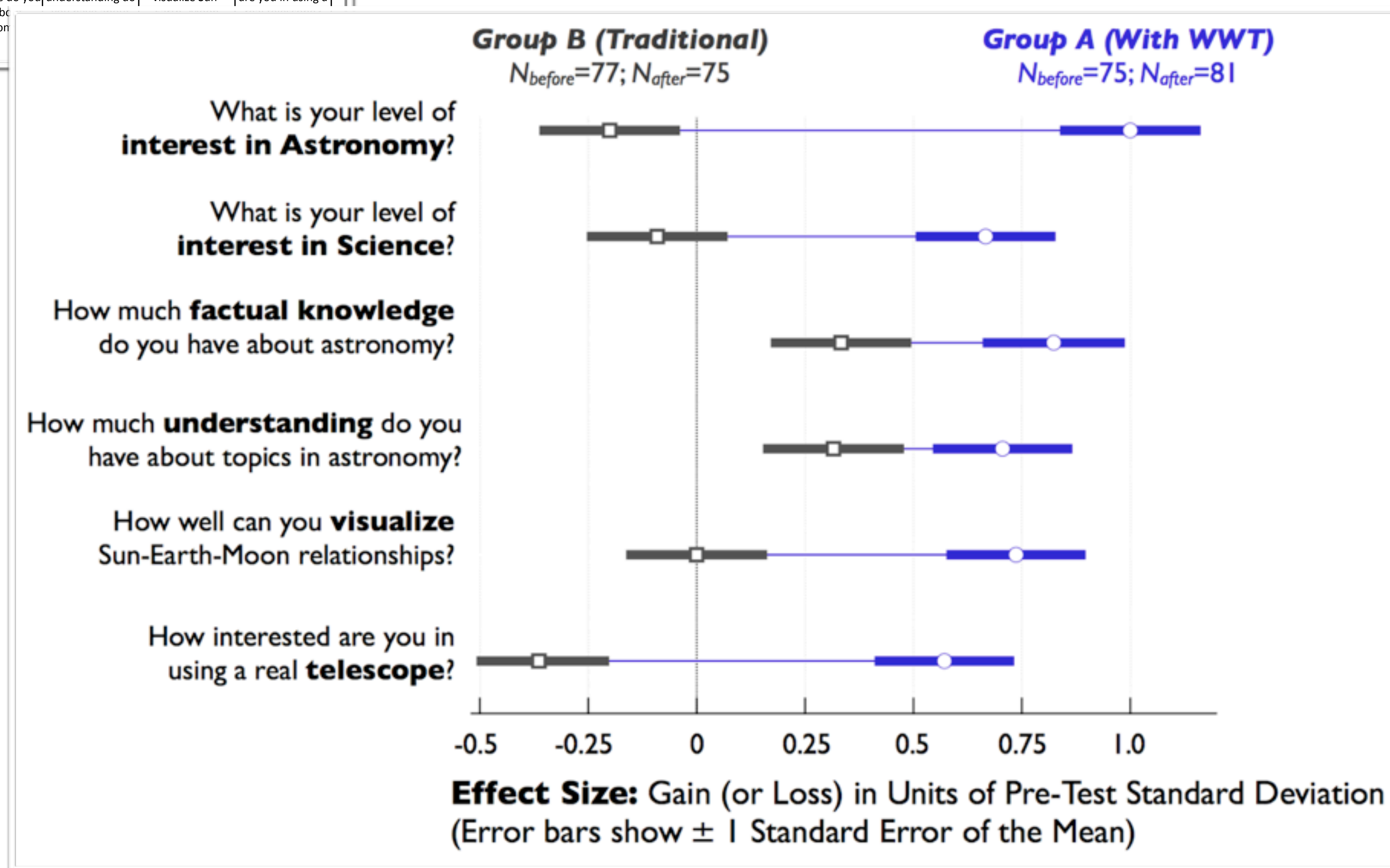


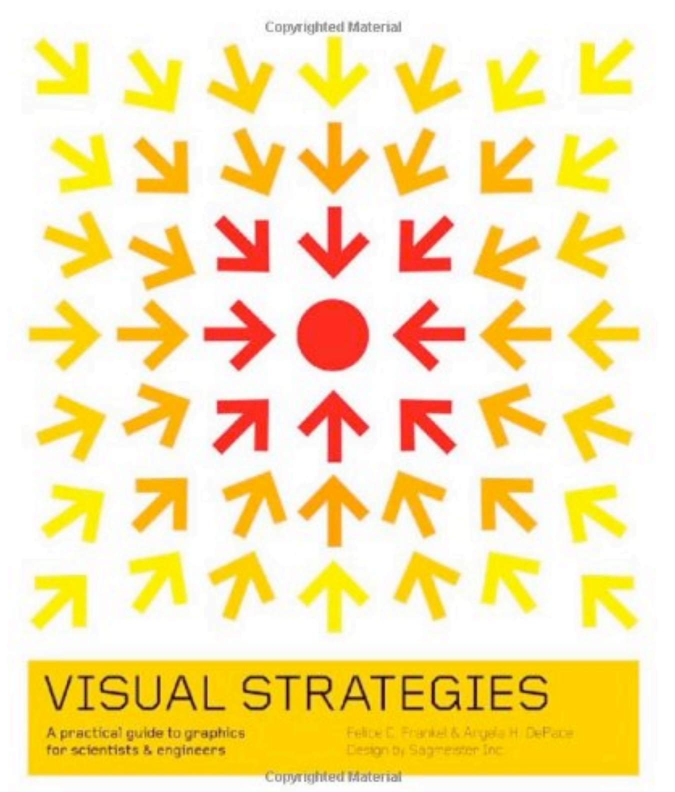
Thursday, June 20, 2013


Group A (With WWT)
N_{before}=75; N_{after}=81
Before
After
p-value of t-test
Group B (Traditional)
N_{before}=77; N_{after}=75
What is your level of **interest in Astronomy**?
What is your level of **interest in Science**?
How much **factual knowledge** do you have about astronomy?
How much **understanding** do you have about topics in astronomy?
How well can you **visualize** Sun-Earth-Moon relationships?
How interested are you in using a real **telescope**?
<0.0001
0.0002
<0.0001
<0.0001
<0.0001
0.0006
0.17
0.45
0.02
0.04
0.49
0.05
2.5 3 3.5 4 4.5 5
Likert Scale


Effect Size
(Change in Units of Standard Deviation of the Pre-Test Mean)
1.25
1.00
0.75
0.50
0.25
0.00
-0.25
-0.50
-0.75
error bars=±2SE
Without WWT
With WWT
What is your level of interest in Astronomy?
What is your level of interest in Science?
How much factual knowledge do you have about astronomy?
How much understanding do you have about topics in astronomy?
How well can you visualize Sun-Earth- Moon relationships?
How interested are you in using a real telescope?


Effect Size
(Change in Units of Standard Deviation of the Pre-Test Mean)
1.25
1.00
0.75
0.50
0.25
0.00
-0.25
-0.50
-0.75
error bars=±2SE
Without WWT
With WWT
What is your level of interest in Astronomy?
What is your level of interest in Science?
How much factual knowledge do you have abo
astronon
How much understanding do
How well can you visualize Sun-
How interested are you in using a


Group B (Traditional)
N_{before}=77; N_{after}=75
Group A (With WWT)
N_{before}=75; N_{after}=81
What is your level of **interest in Astronomy**?
What is your level of **interest in Science**?
How much **factual knowledge** do you have about astronomy?
How much **understanding** do you have about topics in astronomy?
How well can you **visualize** Sun-Earth-Moon relationships?
How interested are you in using a real **telescope**?
-0.5
-0.25
0
0.25
0.5
0.75
1.0
Effect Size: Gain (or Loss) in Units of Pre-Test Standard Deviation
(Error bars show ± 1 Standard Error of the Mean)



<u>Amazon Reviewer (“Ursiform”) says:</u>

“Some of the case studies, unfortunately, impressed me less. They often involve very intricate structures, with graphical solutions that evolved over months. Most of these will not be helpful to average (i.e., well above average) person trying to clearly explain a complex technical point. It's not that they aren't impressive. But you don't learn to paint by having someone put a Raphael in front of you.”

Small multiples (good!), confusing line styles (bad!)

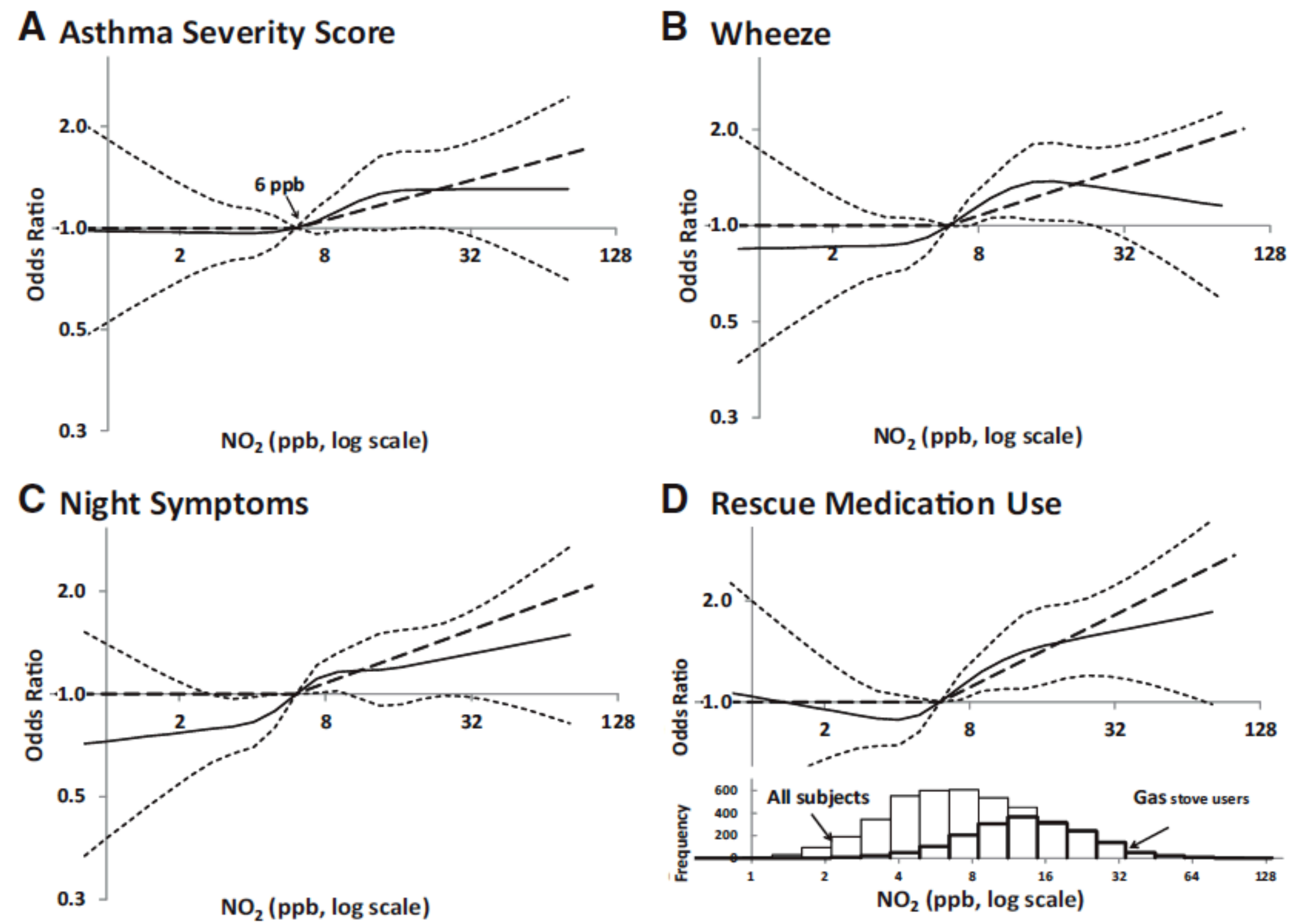
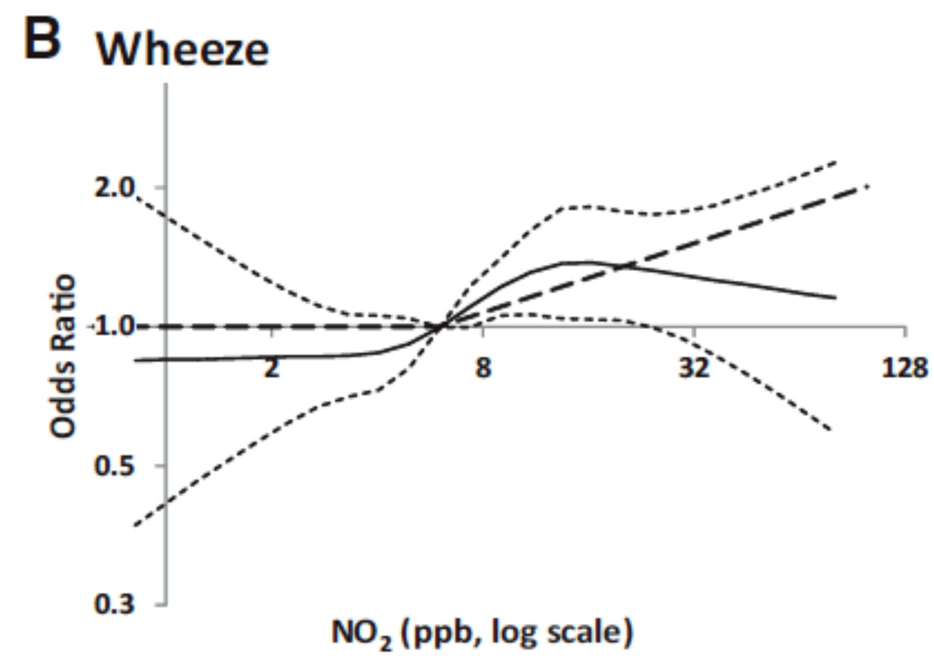
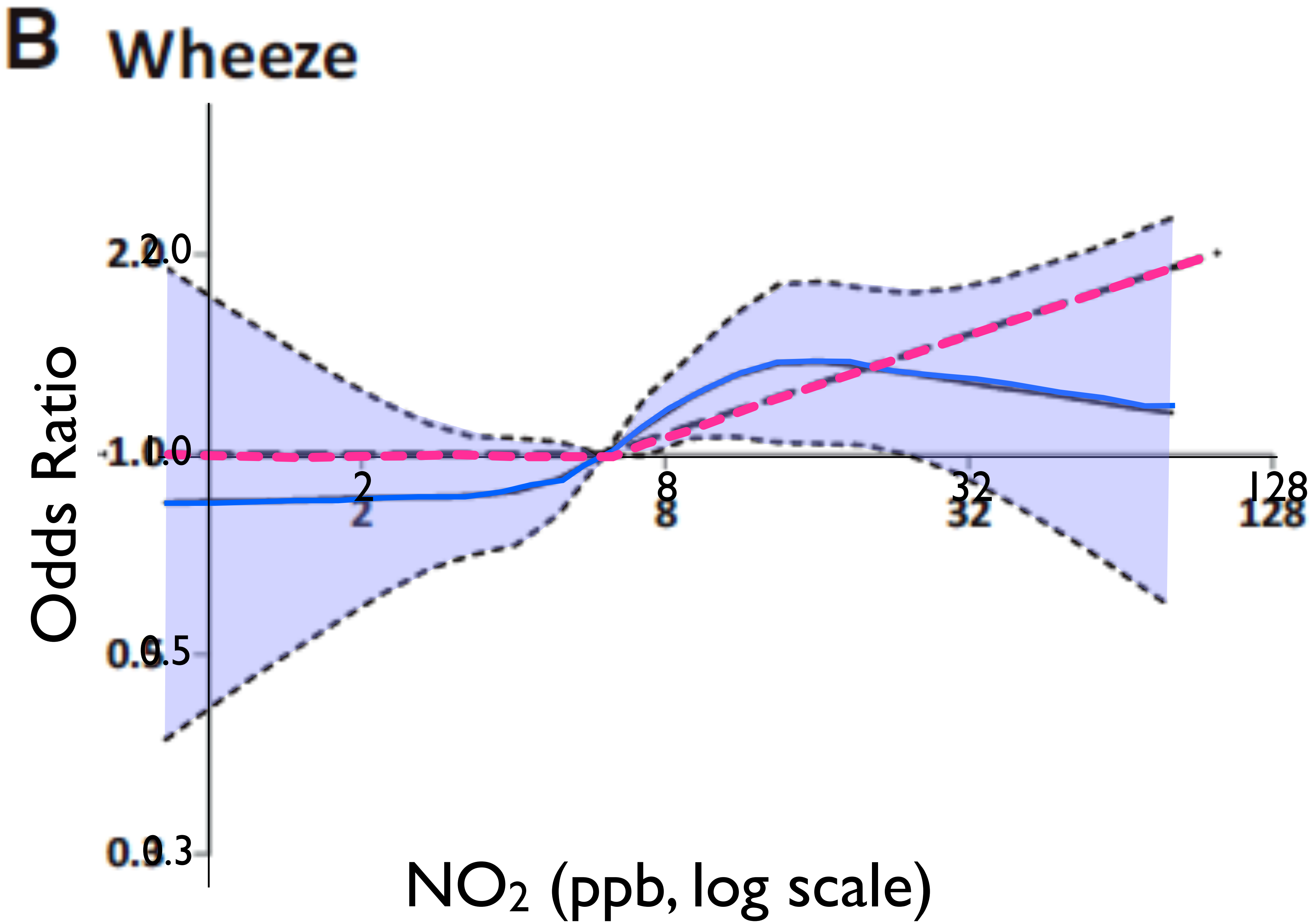
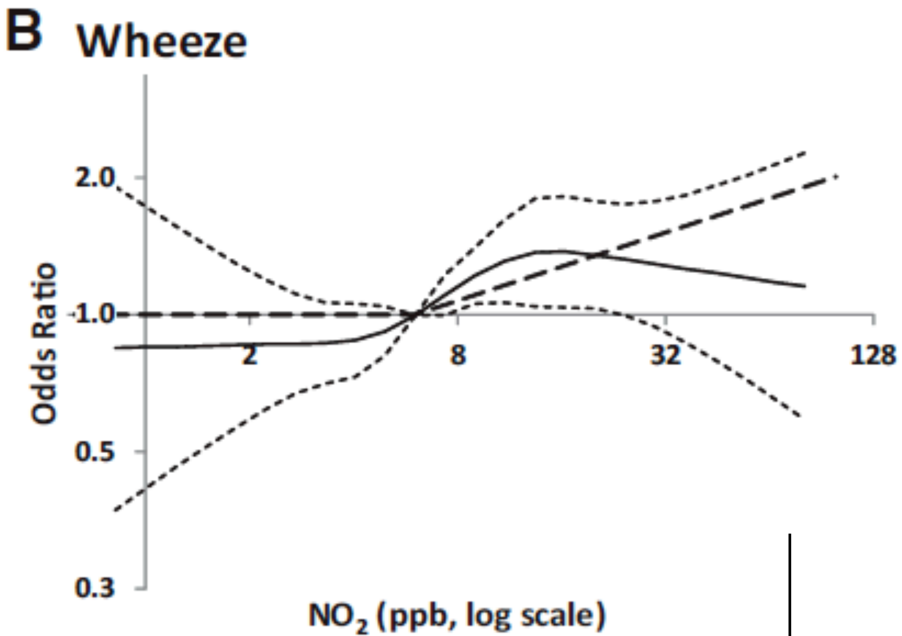


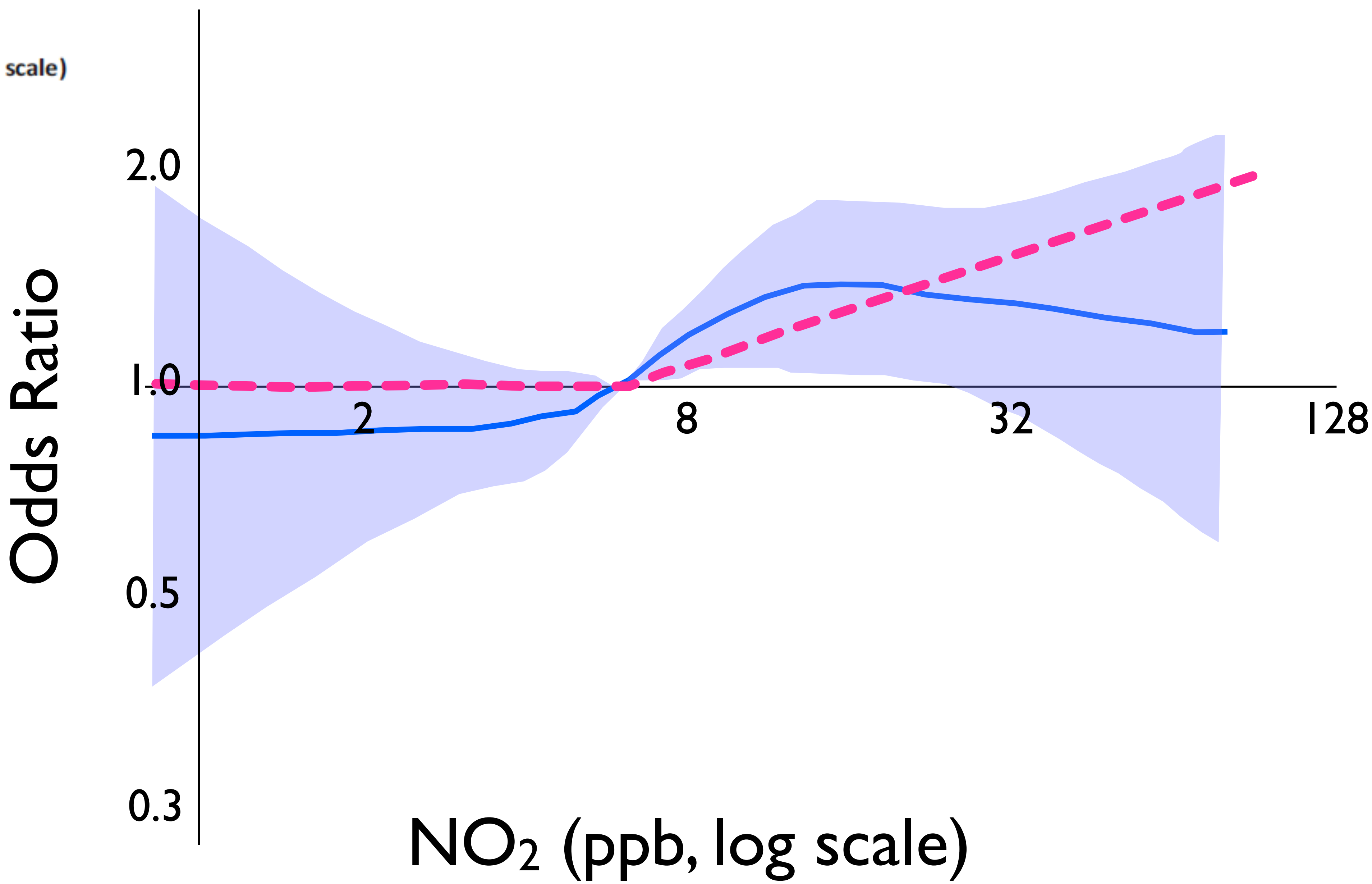
FIGURE 3. Exposure-response relationships between health outcome and NO_2 (log concentration as a continuous variable) illustrated with constrained, natural spline functions (solid lines) with 95% confidence limits (small dashed lines) and threshold function (bold dashed line) from fully adjusted, hierarchical ordered logistic regression models for (A) asthma severity score, (B) wheeze, (C) night symptoms, and (D) rescue medication use. Also shown is a histogram of NO_2 levels measured in subjects' homes (panel D) for all observations (thin border) and observations taken in homes of gas stove users (bold border).

original figure: Belanger et al. 2013, Epidemiology


B
Wheeze
Odds Ratio
2.0
1.0
0.5
0.3
2
8
32
128
NO₂ (ppb, log scale)


B
Wheeze
Odds Ratio
2.0
1.0
0.5
0.3
2
8
32
128
NO2 (ppb, log scale)


B Wheeze
2.0
1.0
0.5
0.3
Odds Ratio
2
8
32
128
NO$_2$ (ppb, log scale)


2.0
1.0
0.5
0.3
Odds Ratio
2
8
32
128
NO2 (ppb, log scale)

But, more deeply, think about...

Data • Dimensions • Display

Linked Views

The Art *of* Numbers

Alyssa A. Goodman • Harvard University

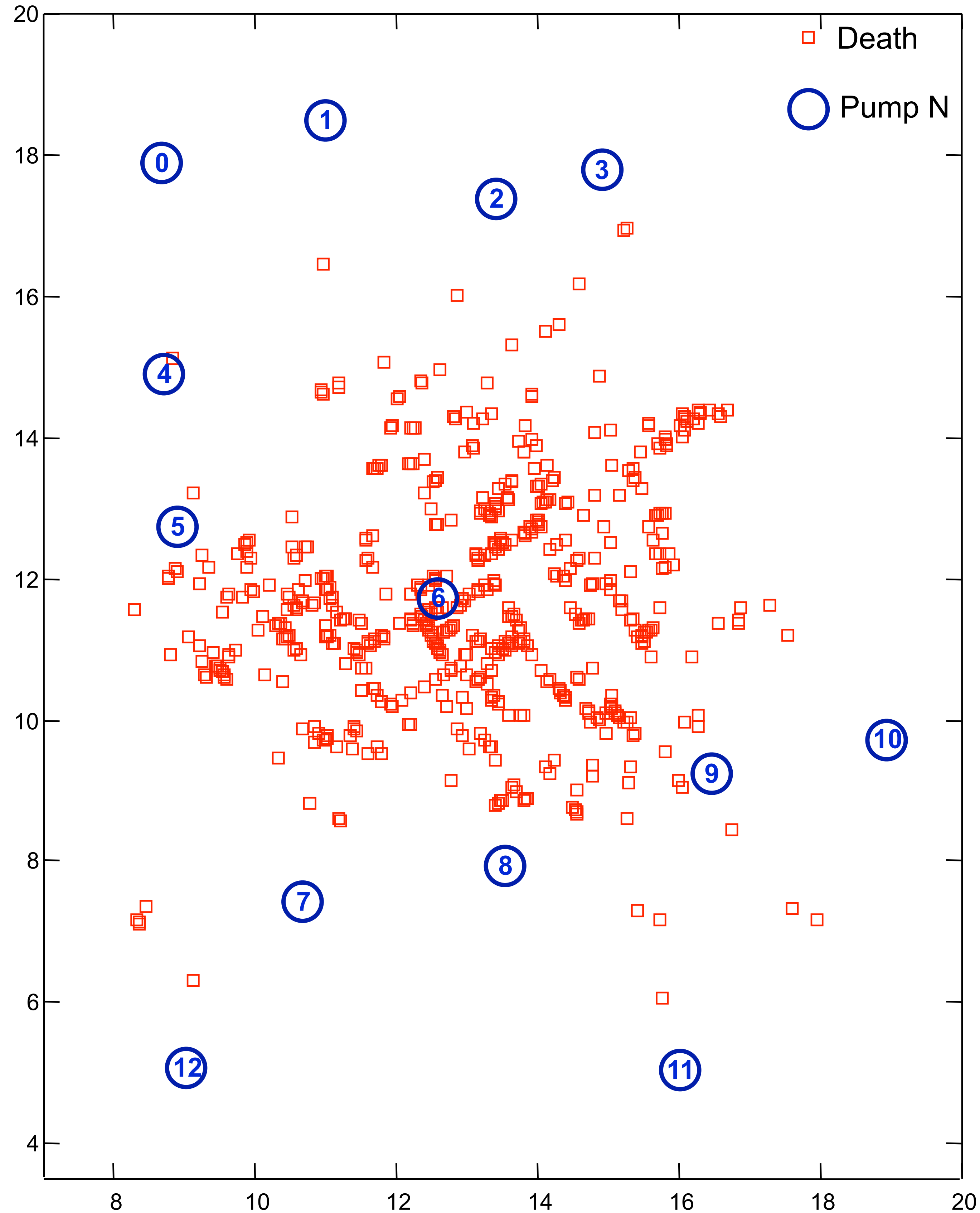
Harvard

EMR19

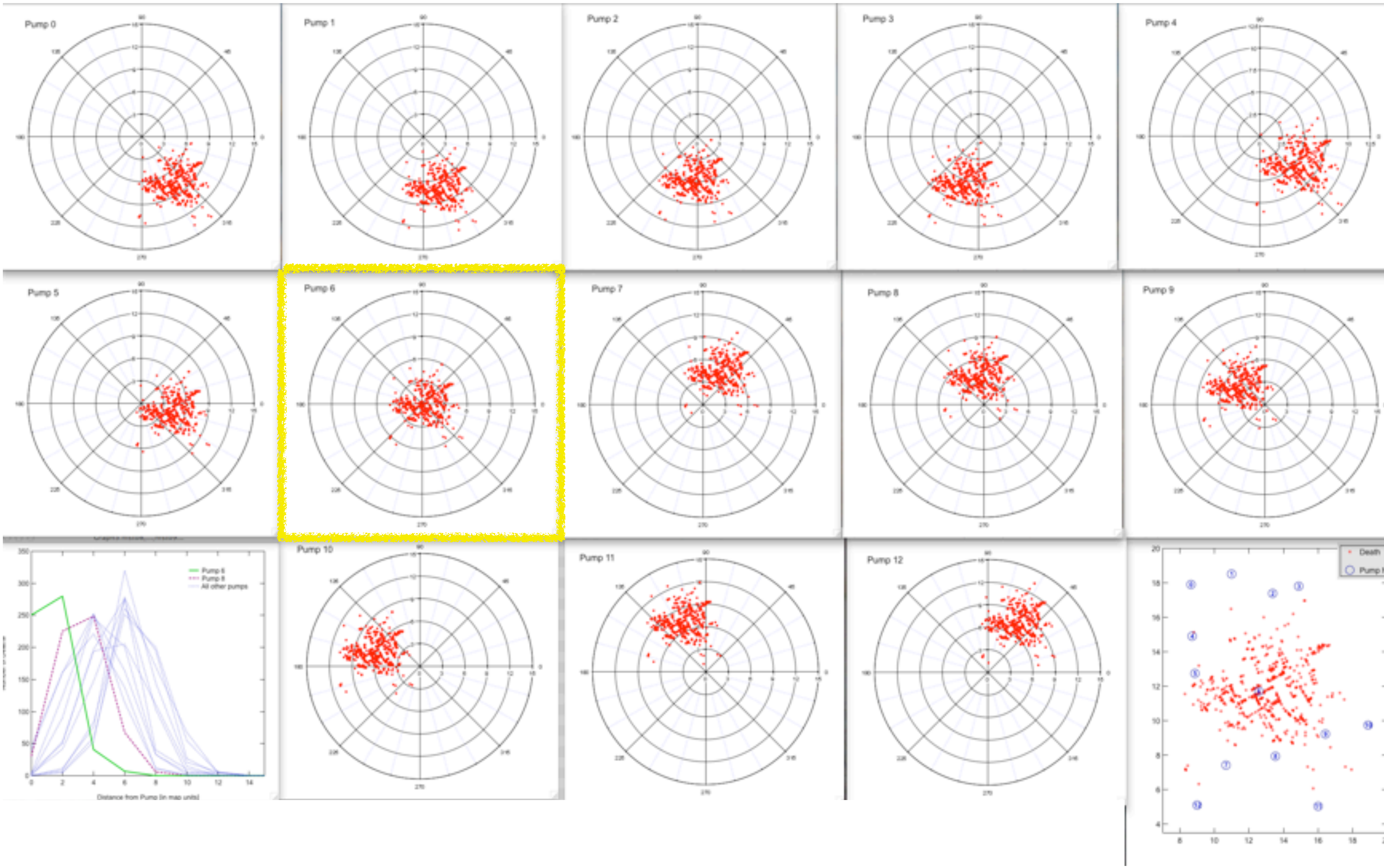
"Project Set 2"

Analysis of 1854 London Cholera Epidemic

(with "Igor")

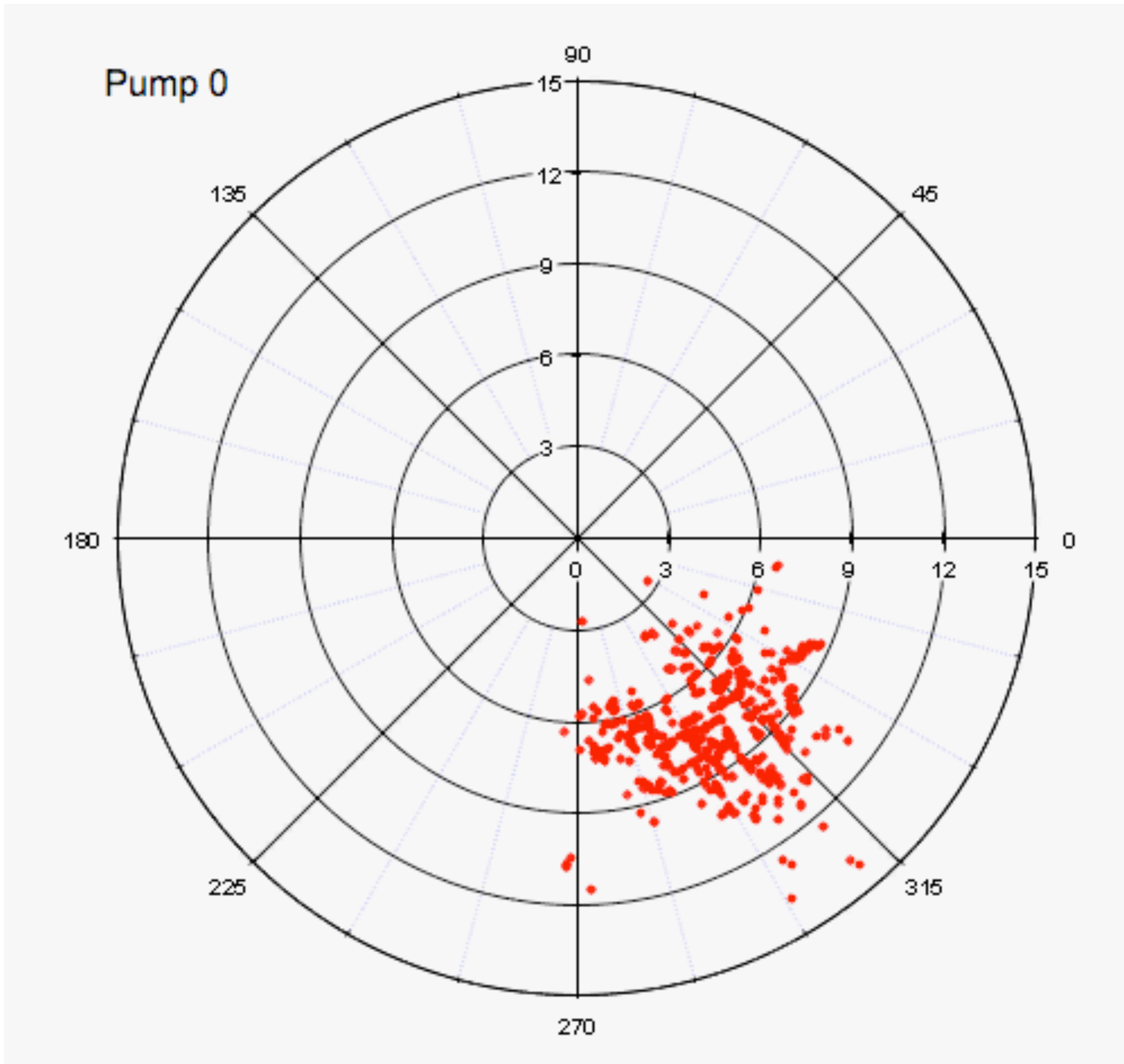


Direction?

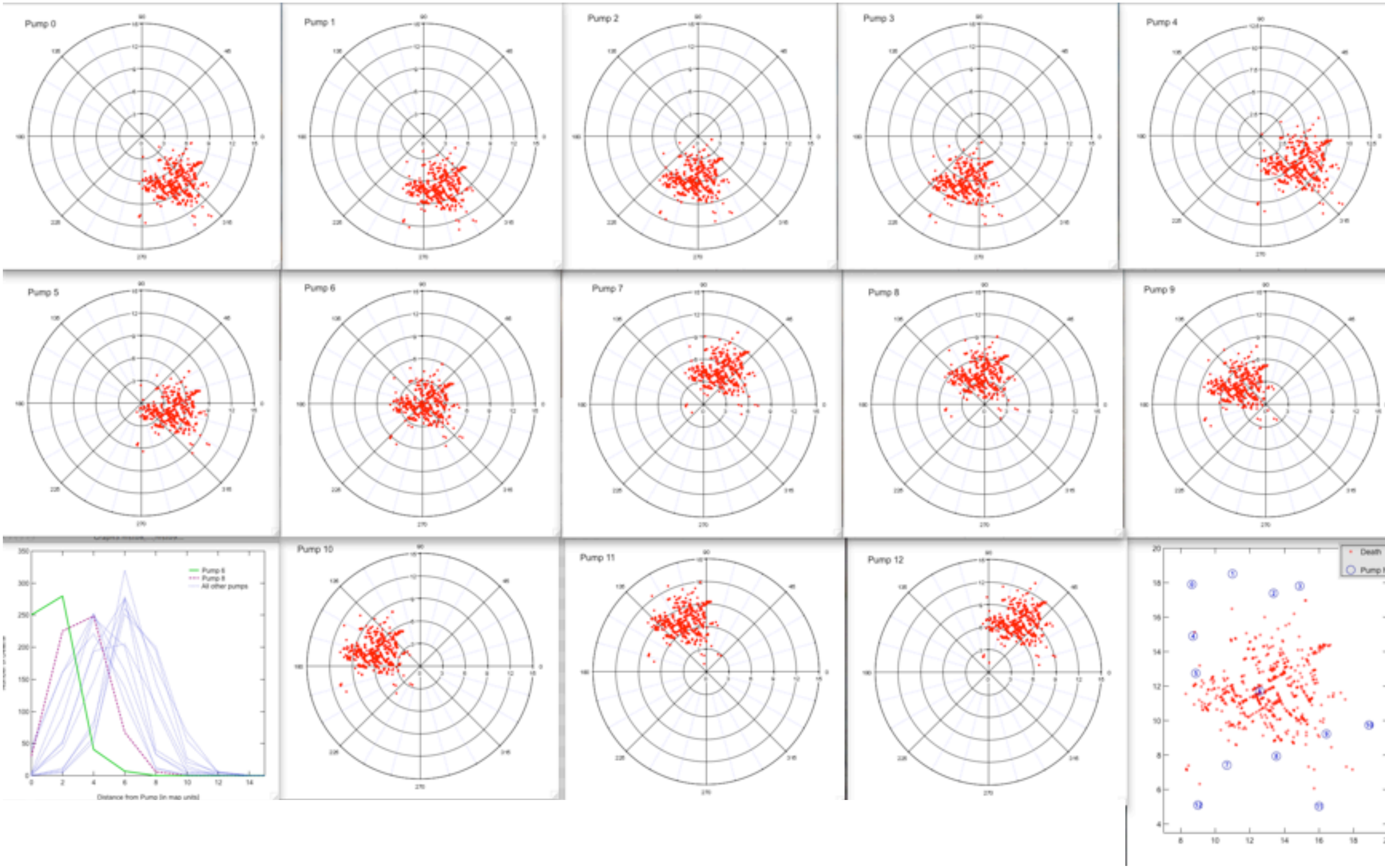


Pump 0
Pump 1
Pump 2
Pump 3
Pump 4
Pump 5
Pump 6
Pump 7
Pump 8
Pump 9
Pump 6
Pump 8
All other pumps
Distance from Pump (in map units)
Pump 10
Pump 11
Pump 12
Death
Pump N

Movie, slider, or small multiples?



Movie, slider, or small multiples?



The Art of Numbers

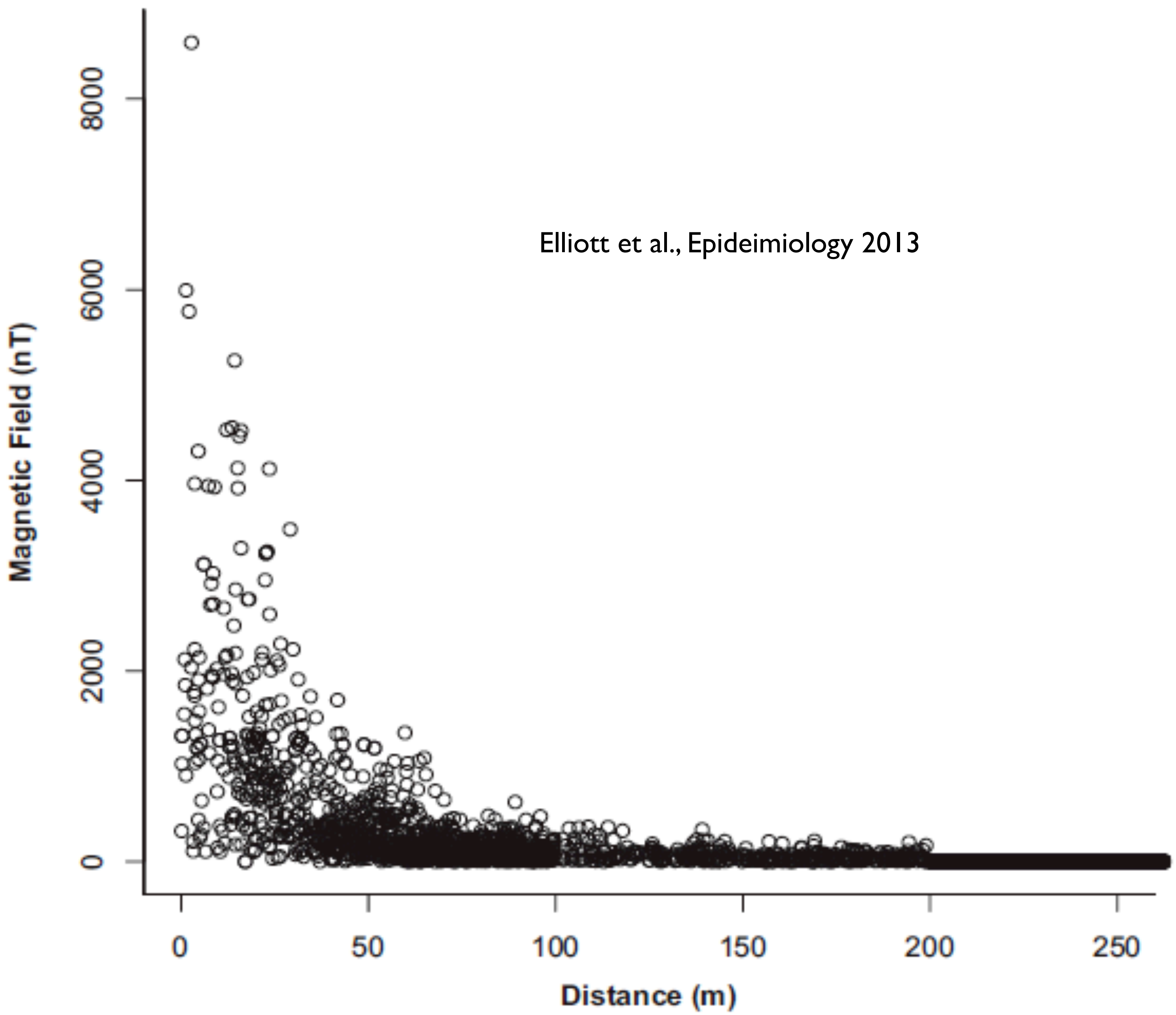


FIGURE 1. Scatter plot of distance and estimated magnetic fields from power lines.

Abstract #: 225

VISIONARY EPIDEMIOLOGY: TOWARDS BETTER GRAPHICAL PRESENTATION OF DATA. Jay Kaufman*, Allen Wilcox, (McGill University, Montreal QC CANADA)

Epidemiologists publish papers that rely on data, and graphical display of data is the most effective and compelling format for conveying this information. When graphs are properly conceived and artfully executed they appear simple and straightforward: data made into pictures. But such clarity seldom comes without effort, and a brief perusal of our journals suggests that far too few epidemiologists make this effort. Why does graphical practice in our field lag so far behind that seen in many other disciplines? Epidemiologists generally receive no formal training in this crucial area of scientific communication, and the results of this neglect can range from comic to catastrophic. For graphs to convey useful (and not misleading) information, the analyst must start with a clear understanding of the message to be conveyed, select data economically, and then give close thought to scale, proportion, choice of symbols, and labeling. Furthermore, there are no standard conventions for conveying complex epidemiologic concepts such as interaction or changes over time. The purpose of this symposium is to provide practical guidelines to epidemiologists in conveying their data and results visually, pointing out common pitfalls and suggesting criteria for assessing graphical displays.

Intro and preliminary presentation of the issues: Jay Kaufman and Allen Wilcox

Pictures at an exhibition: Sixteen visual conversations about one thing - Howard Wainer, PhD

The Visual Display of Information: The Art of Numbers - Alyssa Goodman, PhD