CSSS 569 · Visualizing Data and Models

PRINCIPLES FOR THE VISUAL DISPLAY OF SCIENTIFIC INFORMATION

Christopher Adolph

Department of Political Science

and

Center for Statistics and the Social Sciences

University of Washington, Seattle

CENTER *for* STATISTICS *and the* SOCIAL SCIENCES

Plan of Lecture

Turning points in the history of visual displays

Tufte's principles for information design

How data visualization differs from infovis

Scales and scaling

Making a scatterplot from scratch

Sorting in tables and table-like figures

The Invention of Visual Display

Megalaceros, a giant prehistoric deer
Source: Wikipedia/public domain

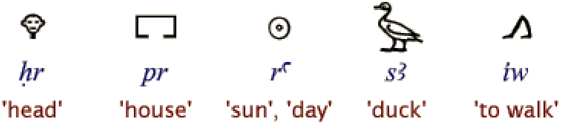
The visual representation of information dates back to the Lascaux cave paintings ($\sim$15000 BCE)

Simplified images of a more complex physical reality

The Invention of Visual Display

Pictographic writing followed – now simplified images could represent *other ideas*

These cartoons may seem primitive compared to later realistic art, but cartoons are often better communicators



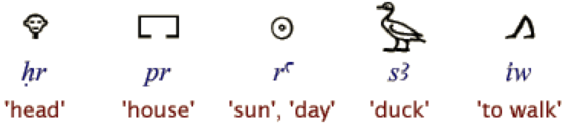
Source: XXX

The Invention of Visual Display

Pictographic writing followed – now simplified images could represent *other ideas*

These cartoons may seem primitive compared to later realistic art, but cartoons are often better communicators

Where's the hand?

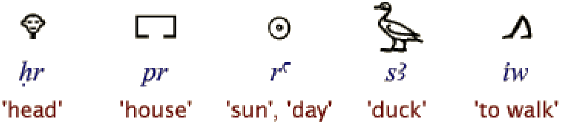


Source: XXX

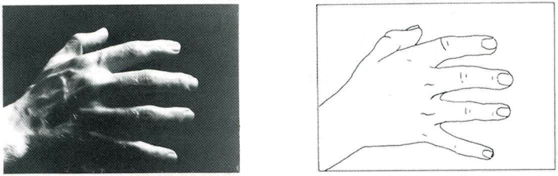
Pictographic writing followed – now simplified images could represent *other ideas*

These cartoons may seem primitive compared to later realistic art, but cartoons are often better communicators

Where's the hand?



Source: XXX



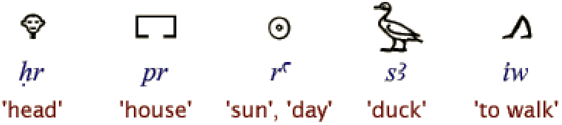
Source: Edward Tufte, *VDQI*

The Invention of Visual Display

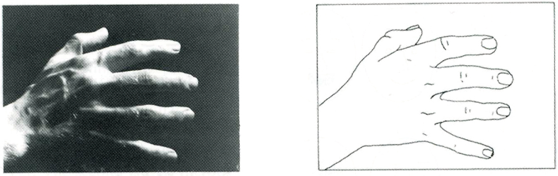
Pictographic writing followed – now simplified images could represent *other ideas*

These cartoons may seem primitive compared to later realistic art, but cartoons are often better communicators

Where's the hand?
Which hand did you see first?



Source: XXX



Source: Edward Tufte, *VDQI*

The Invention of Visual Display: Three Elements

Visual display always involves **representation** and **simplification**

As pictographic writing suggests, **abstraction** is also frequently present, but took much longer to develop

The Invention of Visual Display: Three Elements

Visual display always involves **representation** and **simplification**

As pictographic writing suggests, **abstraction** is also frequently present, but took much longer to develop

Besides writing, earliest examples are maps

One of the earliest is from c. 600 BCE and lacked a coordinate system, an innovation that would only come c. 200 CE

Source: data-art.net/resources/history_of_vis.php

Visual display always involves **representation** and **simplification**

As pictographic writing suggests, **abstraction** is also frequently present, but took much longer to develop

Besides writing, earliest examples are maps

The earliest is from c. 600 BCE and lacked a coordinate system, an innovation that would only come c. 200 CE

Source: Tufte, *VDQI*

Simplification aids visual communication

Source: Google Earth

Mapmakers have long known that the best maps simplify and even distort reality

Finding features in a photograph is hard – sharpening and repositioning features can help them pop out

To see this, try to find the pyramids from space using satellite photos

Simplification aids visual communication

Wait...Where's Giza?

Mapmakers have long known that the best maps simplify and even distort reality

Finding features in a photograph is hard – sharpening and repositioning features can help them pop out

To see this, try to find the pyramids from space using satellite photos

Simplification aids visual communication

...literally unfindable without prior knowledge

Mapmakers have long known that the best maps simplify and even distort reality

Finding features in a photograph is hard – sharpening and repositioning features can help them pop out

To see this, try to find the pyramids from space using satellite photos

Simplification aids visual communication

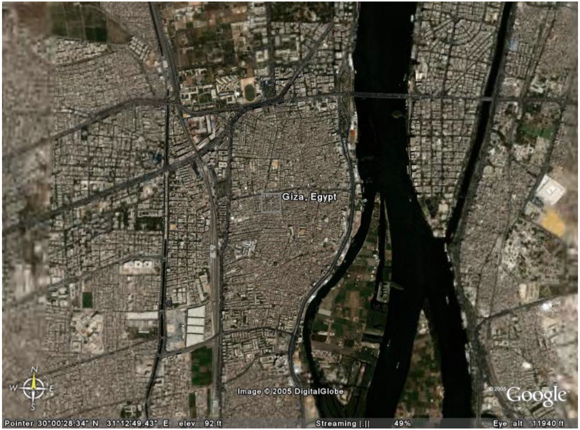
Even zoomed in, we can't spot the largest manmade objects on Earth

Mapmakers have long known that the best maps simplify and even distort reality

Finding features in a photograph is hard – sharpening and repositioning features can help them pop out

To see this, try to find the pyramids from space using satellite photos

Simplification aids visual communication



Apparently, they're not downtown...

Mapmakers have long known that the best maps simplify and even distort reality

Finding features in a photograph is hard – sharpening and repositioning features can help them pop out

To see this, try to find the pyramids from space using satellite photos

Simplification aids visual communication

No choice but to hunt around for a clue

Mapmakers have long known that the best maps simplify and even distort reality

Finding features in a photograph is hard – sharpening and repositioning features can help them pop out

To see this, try to find the pyramids from space using satellite photos

Simplification aids visual communication

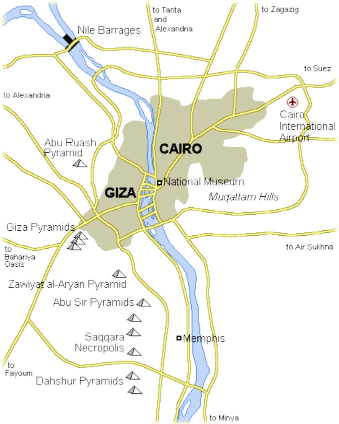
A minute of searching sandy areas later...

Mapmakers have long known that the best maps simplify and even distort reality

Finding features in a photograph is hard – sharpening and repositioning features can help them pop out

To see this, try to find the pyramids from space using satellite photos

Simplification aids visual communication



Source: `lexicorient.com/egypt/cairo_m1.htm`

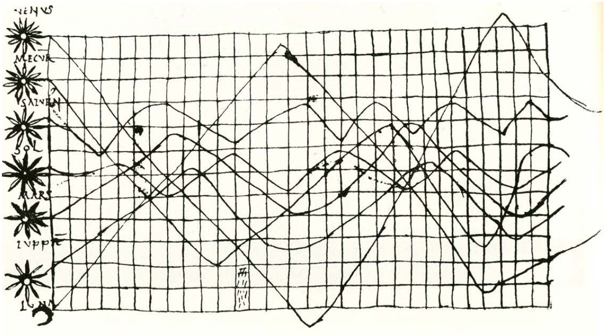
To find Waldo, use a map

The richest, most beautiful representation isn't always the most useful one

Removing lines, color, and *information* can help

Just like in statistical modeling: we need to (over-)simplify to learn

A view somewhat at odds with Tufte...

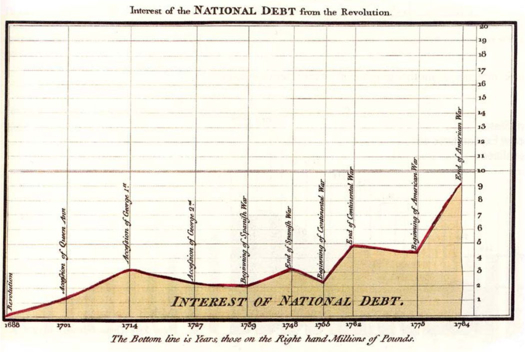


Source: Tufte, *VDQI*

Abstract visual displays are relatively new

A monk made this vague time series plot of planetary movement c. 950 CE

First known use of time as a visual dimension, but unknown until much later



Source: Tufte, *VDQI*

The unknown monk's plot didn't catch on – lost in a notebook

Time series plots weren't rediscovered until 1786 (!) by William Playfair

Playfair's time series dealt with abstract concepts like public debt and trade deficits

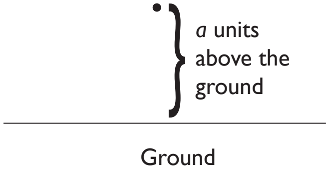
Playfair also invented the bar plot and pie chart

The ancients knew geometry backwards and forwards

Oddly, they don't appear to have discovered graphics

Before 1637 CE, visual representations = literal depictions of physical relationships

To go beyond maps, Descartes (and perhaps Fermat) had to recognize something that seems obvious in retrospect



Maps show the distance between objects and a reference...

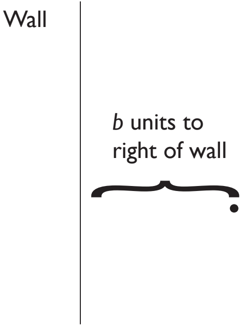
Abstract graphical representation: What took so long?

The ancients knew geometry backwards and forwards

Oddly, they don't appear to have discovered graphics

Before 1637 CE, visual representations = literal depictions of physical relationships

To go beyond maps, Descartes (and perhaps Fermat) had to recognize something that seems obvious in retrospect



...Or between one object and *multiple* references

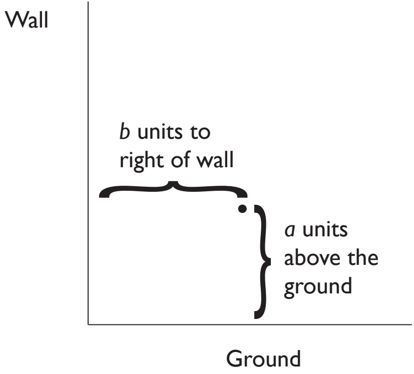
Abstract graphical representation: What took so long?

The ancients knew geometry backwards and forwards

Oddly, they don't appear to have discovered graphics

Before 1637 CE, visual representations = literal depictions of physical relationships

To go beyond maps, Descartes (and perhaps Fermat) had to recognize something that seems obvious in retrospect



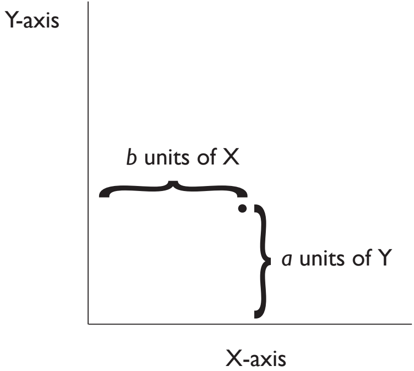
Putting these together makes a map

The ancients knew geometry backwards and forwards

Oddly, they don't appear to have discovered graphics

Before 1637 CE, visual representations = literal depictions of physical relationships

To go beyond maps, Descartes (and perhaps Fermat) had to recognize something that seems obvious in retrospect



But what if we change the names of our references to be general?

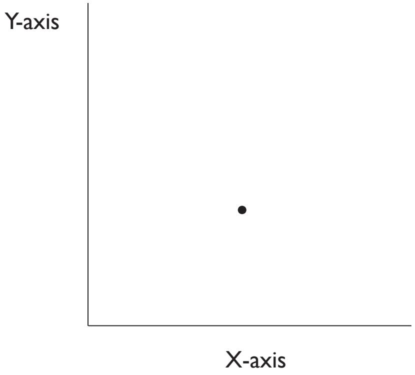
Abstract graphical representation: What took so long?

The ancients knew geometry backwards and forwards

Oddly, they don't appear to have discovered graphics

Before 1637 CE, visual representations = literal depictions of physical relationships

To go beyond maps, Descartes (and perhaps Fermat) had to recognize something that seems obvious in retrospect



We've made something revolutionary: a *relational* graphic

Abstract graphical representation: What took so long?

The axes of a Cartesian plane can measure anything

Not just space, distance, time, or motion, but any functional relationship

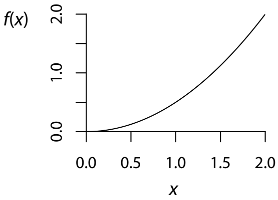
The axes of a Cartesian plane can measure anything

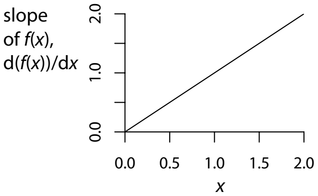
Not just space, distance, time, or motion, but any functional relationship

Can you imagine learning calculus without any visual displays of functions?

$$f(x) = \frac{1}{2}x^2 \qquad \text{slope of } f(x),\ \frac{df(x)}{x} = x$$

versus





Abstract graphical representation: What took so long?

Cartesian plane an invaluable complement to mathematical formalism, with endless scientific applications

On a Cartesian plane, any measurable concept can be plotted:
money · *attitudes* · *preferences* · *qualities* · *counts*

Abstract graphical representation: What took so long?

Cartesian plane an invaluable complement to mathematical formalism, with endless scientific applications

On a Cartesian plane, any measurable concept can be plotted:
money · *attitudes* · *preferences* · *qualities* · *counts*

Still a somewhat unintuitive concept, esp. outside the social sciences

Some natural scientists assume graphs always show *things* that physically exist

(graphs are thus cartoons of things that could be photographed, or measured in meters, if only you had a ruler the right size)

Most things we will display in this class could never be photographed or touched

But we can still learn from visual analogies

Most "types" of visual display were developed before 1900

19th century practitioners devoted enormous effort to graphics

By start of 20th century, some information designers were making very modern displays

From a 1911 conference on public health

The 20th century: A dark age of data visualization

By the mid-20th century, statistical graphics fell into disuse

Popular use of information graphics: quick and dirty, for mass media

1970s saw reemergence of statistical graphics (e.g., John W. Tukey)

Now easy to make, because of computers

But computer defaults are inelegant, clunky, misleading

Until 5 years ago, so were most media examples of visual display

Today: A new golden age of data visualization

Edward Tufte wrote *VDQI* in the early 1980s

Responding to garish, inefficient, uninformative, misleading graphics of the time

Tufte started his career as a political scientist

Clearly disappointed by the quality of graphics in social science journals

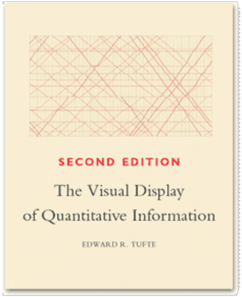
No one would have predicted *VDQI* would catch on: self-published

Enormous impact: Tufte now a "guru" of information graphics;
VDQI was one of Amazon.com's top 100 books of 20th century

Lots of new books, conferences, and jobs in infovis followed

Tufte's *VDQI* is several things:

- A beautiful, richly detailed, densely illustrated book
- A call for scientific integrity and seriousness in graphical display
- A polemic in favor of a particular aesthetics of information graphics



My personal aesthetics are similar to (& influenced by) Tufte's

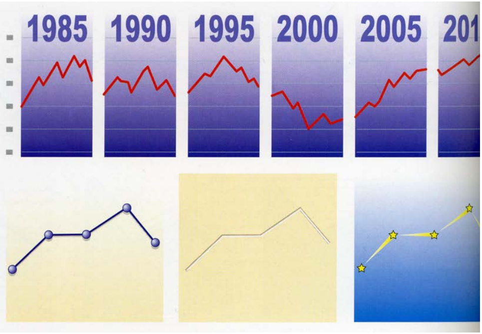
But aesthetics are not a science:
We can disagree over whether all of Tufte's ideas are "right"

Tufte's Principles

1. Show viewers substance, not method or graphic design; avoid **chartjunk**
2. Maximize data, minimize ink & space; **data-ink ratio**
3. Be honest: avoid illusions and distortions; minimize the **lie factor**
4. Show the data and facilitate comparison
5. Use **small multiples**, or repetitions of a basic design

Tufte hates

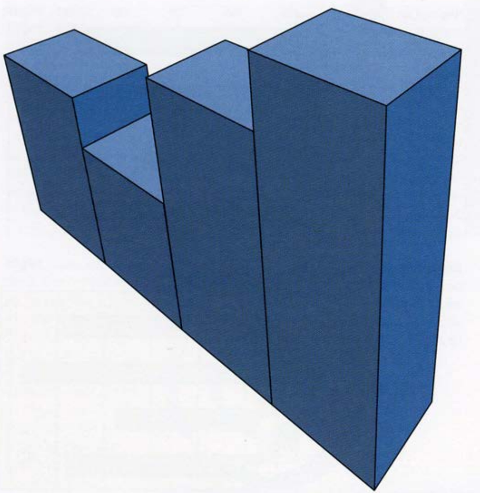
- distracting grid lines or other scaffolding
- thick lines, overlarge plots
- gratuitous use of icons, embellishment (e.g., USA Today)
- unnecessary dimensions



"If statistics is boring, then you've got the wrong numbers." Tufte

Tufte hates

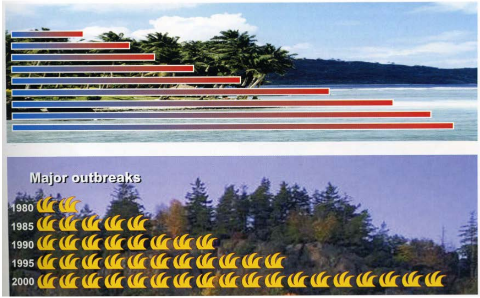
- distracting grid lines or other scaffolding
- thick lines, overlarge plots
- gratuitous use of icons, embellishment (e.g., USA Today)
- unnecessary dimensions



"If statistics is boring, then you've got the wrong numbers." Tufte

Tufte hates

- distracting grid lines or other scaffolding
- thick lines, overlarge plots
- gratuitous use of icons, embellishment (e.g., USA Today)
- unnecessary dimensions



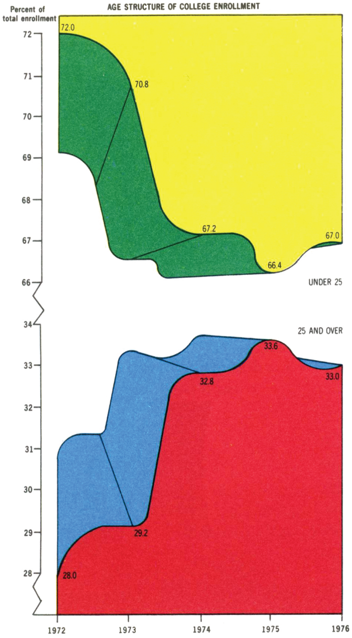
"If statistics is boring, then you've got the wrong numbers." Tufte

Tufte's Principles

2. Maximize data-to-ink

In his first book, Tufte suggested this might be the worst graphic of all time

Problems?



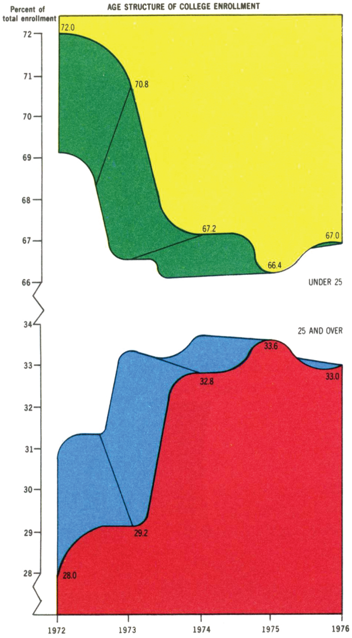
Tufte's Principles

2. Maximize data-to-ink

In his first book, Tufte suggested this might be the worst graphic of all time

Problems?

What's the scale?



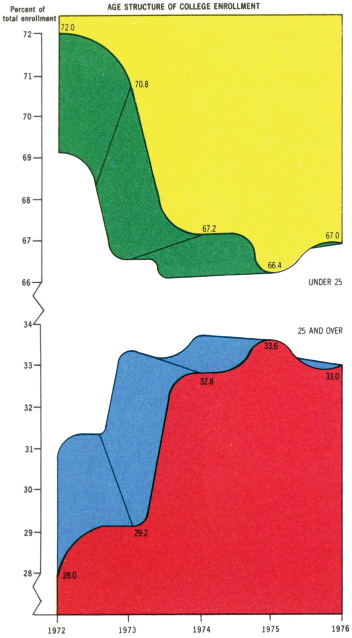
Tufte's Principles

2. Maximize data-to-ink

In his first book, Tufte suggested this might be the worst graphic of all time

Problems?

- What's the scale?
- Why the curves?



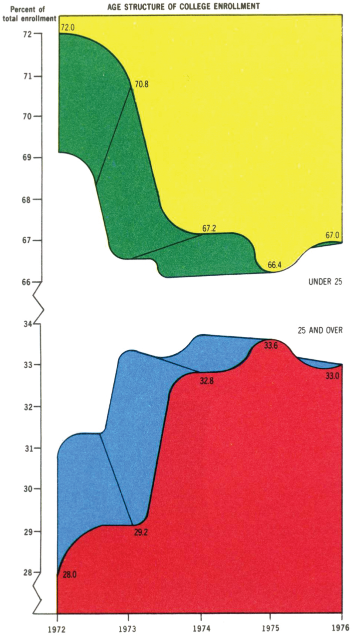
Tufte's Principles

2. Maximize data-to-ink

In his first book, Tufte suggested this might be the worst graphic of all time

Problems?

- What's the scale?
- Why the curves?
- Why are there two lines?



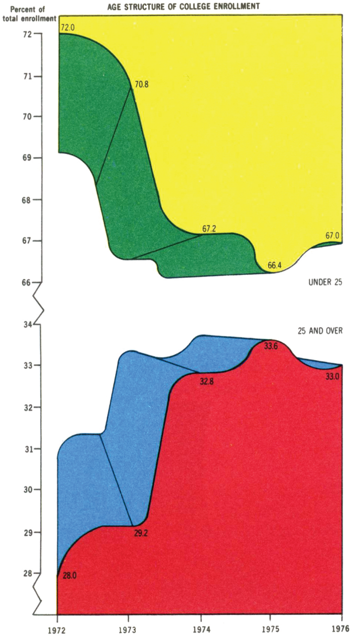
Tufte's Principles

2. Maximize data-to-ink

In his first book, Tufte suggested this might be the worst graphic of all time

Problems?

- What's the scale?
- Why the curves?
- Why are there two lines?
- How many data points?

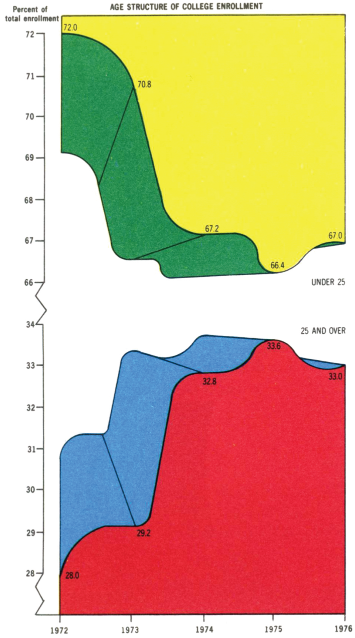


Tufte's Principles
2. Maximize data-to-ink

In his first book, Tufte suggested this might be the worst graphic of all time

Problems?

- What's the scale?
- Why the curves?
- Why are there two lines?
- How many data points?
- Why such tiny text?



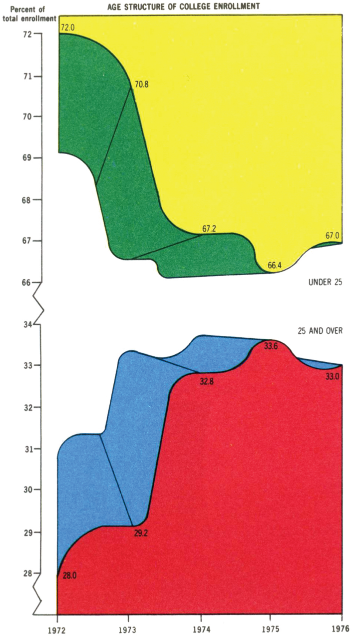
Tufte's Principles

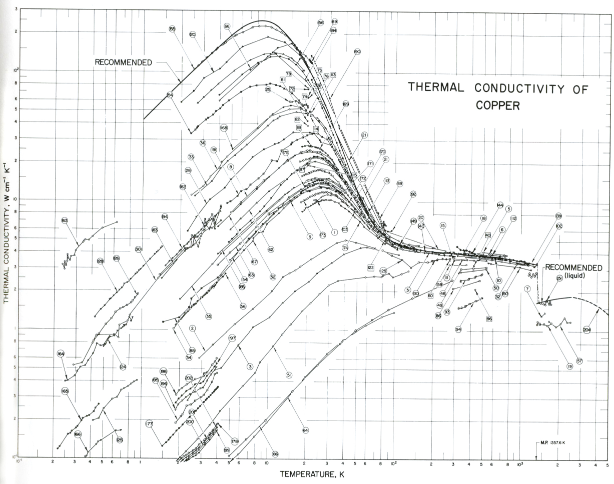
2. Maximize data-to-ink

In his first book, Tufte suggested this might be the worst graphic of all time

Problems?

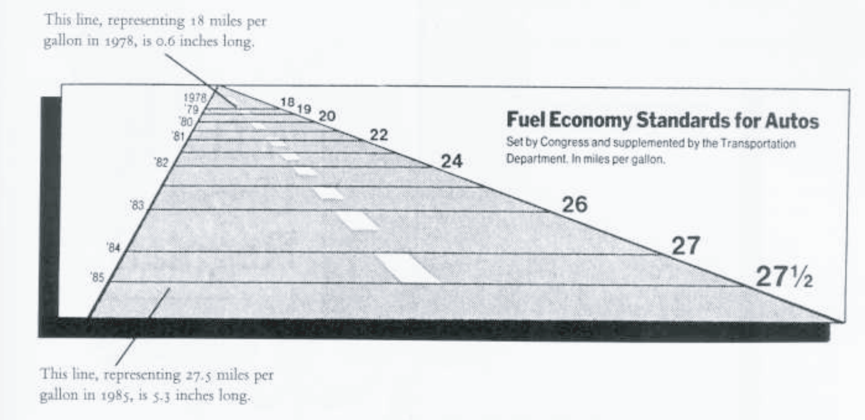
- What's the scale?
- Why the curves?
- Why are there two lines?
- How many data points?
- Why such tiny text?
- Did this need four colors?



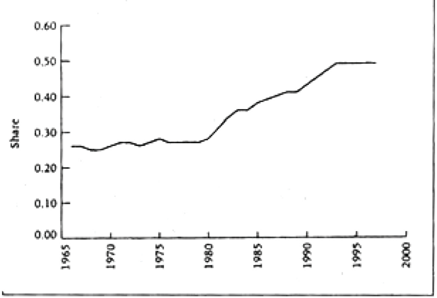


A good data/ink ratio: Literature review in a page; a model for meta-analyses

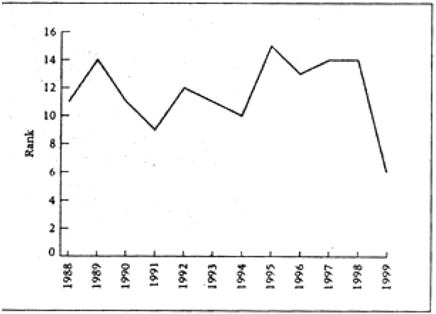
Imagine this showed estimates of the effect of class size on educational performance



Beware: Usually, distortions are more subtle than this...



BY THE NUMBERS: OVER 35 YEARS, CORNELL'S TUITION HAS TAKEN AN INCREASINGLY LARGER SHARE OF ITS MEDIAN STUDENT FAMILY INCOME.

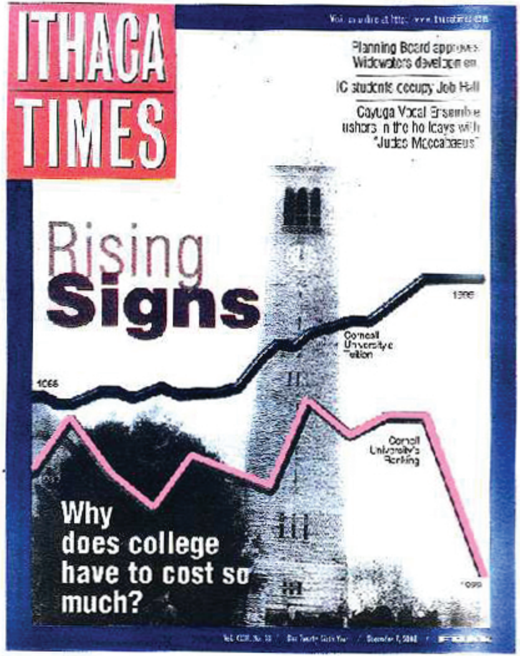


PECKING ORDER: OVER 12 YEARS, CORNELL'S RANKING IN *US NEWS & WORLD REPORT* HAS RISEN AND FALLEN ERRATICALLY.

The worst visual display I've ever seen

What is the claim?

Do you believe it?



An even more misleading combination of these plots on the cover

Scales are what allow us to make comparisons within and *across* graphics

Clearly, careful thought about scaling is essential to making good scientific visuals

Tufte's Principles

4. Facilitate comparison

Tufte presents this table of cancer survival rates

What's good about it?

What could be improved?

Source: Tufte, *Cognitive Style of Powerpoint*

Estimates of relative survival rates, by cancer site[9]

	% survival rates and standard errors							
	5 year		10 year		15 year		20 year	
Prostate	98.8	0.4	95.2	0.9	87.1	1.7	81.1	3.0
Thyroid	96.0	0.8	95.8	1.2	94.0	1.6	95.4	2.1
Testis	94.7	1.1	94.0	1.3	91.1	1.8	88.2	2.3
Melanomas	89.0	0.8	86.7	1.1	83.5	1.5	82.8	1.9
Breast	86.4	0.4	78.3	0.6	71.3	0.7	65.0	1.0
Hodgkin's disease	85.1	1.7	79.8	2.0	73.8	2.4	67.1	2.8
Corpus uteri, uterus	84.3	1.0	83.2	1.3	80.8	1.7	79.2	2.0
Urinary, bladder	82.1	1.0	76.2	1.4	70.3	1.9	67.9	2.4
Cervix, uteri	70.5	1.6	64.1	1.8	62.8	2.1	60.0	2.4
Larynx	68.8	2.1	56.7	2.5	45.8	2.8	37.8	3.1
Rectum	62.6	1.2	55.2	1.4	51.8	1.8	49.2	2.3
Kidney, renal pelvis	61.8	1.3	54.4	1.6	49.8	2.0	47.3	2.6
Colon	61.7	0.8	55.4	1.0	53.9	1.2	52.3	1.6
Non-Hodgkin's	57.8	1.0	46.3	1.2	38.3	1.4	34.3	1.7
Oral cavity, pharynx	56.7	1.3	44.2	1.4	37.5	1.6	33.0	1.8
Ovary	55.0	1.3	49.3	1.6	49.9	1.9	49.6	2.4
Leukemia	42.5	1.2	32.4	1.3	29.7	1.5	26.2	1.7
Brain, nervous system	32.0	1.4	29.2	1.5	27.6	1.6	26.1	1.9
Multiple myeloma	29.5	1.6	12.7	1.5	7.0	1.3	4.8	1.5
Stomach	23.8	1.3	19.4	1.4	19.0	1.7	14.9	1.9
Lung and bronchus	15.0	0.4	10.6	0.4	8.1	0.4	6.5	0.4
Esophagus	14.2	1.4	7.9	1.3	7.7	1.6	5.4	2.0
Liver, bile duct	7.5	1.1	5.8	1.2	6.3	1.5	7.6	2.0
Pancreas	4.0	0.5	3.0	1.5	2.7	0.6	2.7	0.8

Estimates of relative survival rates, by cancer site[9]

	% survival rates and standard errors							
	5 year		10 year		15 year		20 year	
Prostate	98.8	0.4	95.2	0.9	87.1	1.7	81.1	3.0
Thyroid	96.0	0.8	95.8	1.2	94.0	1.6	95.4	2.1
Testis	94.7	1.1	94.0	1.3	91.1	1.8	88.2	2.3
Melanomas	89.0	0.8	86.7	1.1	83.5	1.5	82.8	1.9
Breast	86.4	0.4	78.3	0.6	71.3	0.7	65.0	1.0
Hodgkin's disease	85.1	1.7	79.8	2.0	73.8	2.4	67.1	2.8
Corpus uteri, uterus	84.3	1.0	83.2	1.3	80.8	1.7	79.2	2.0
Urinary, bladder	82.1	1.0	76.2	1.4	70.3	1.9	67.9	2.4
Cervix, uteri	70.5	1.6	64.1	1.8	62.8	2.1	60.0	2.4
Larynx	68.8	2.1	56.7	2.5	45.8	2.8	37.8	3.1
Rectum	62.6	1.2	55.2	1.4	51.8	1.8	49.2	2.3
Kidney, renal pelvis	61.8	1.3	54.4	1.6	49.8	2.0	47.3	2.6
Colon	61.7	0.8	55.4	1.0	53.9	1.2	52.3	1.6
Non-Hodgkin's	57.8	1.0	46.3	1.2	38.3	1.4	34.3	1.7
Oral cavity, pharynx	56.7	1.3	44.2	1.4	37.5	1.6	33.0	1.8
Ovary	55.0	1.3	49.3	1.6	49.9	1.9	49.6	2.4
Leukemia	42.5	1.2	32.4	1.3	29.7	1.5	26.2	1.7
Brain, nervous system	32.0	1.4	29.2	1.5	27.6	1.6	26.1	1.9
Multiple myeloma	29.5	1.6	12.7	1.5	7.0	1.3	4.8	1.5
Stomach	23.8	1.3	19.4	1.4	19.0	1.7	14.9	1.9
Lung and bronchus	15.0	0.4	10.6	0.4	8.1	0.4	6.5	0.4
Esophagus	14.2	1.4	7.9	1.3	7.7	1.6	5.4	2.0
Liver, bile duct	7.5	1.1	5.8	1.2	6.3	1.5	7.6	2.0
Pancreas	4.0	0.5	3.0	1.5	2.7	0.6	2.7	0.8

There are only a few reasons to use a table instead of a graphic:

1. to convey a **handful** of numbers
2. to report precise values for **lookup**
3. to present **many different types** of quantities (i.e., dimensions) for a small number of cases

Usually graphics are more effective than tables

Estimates of relative survival rates, by cancer site[9]

	% survival rates and standard errors							
	5 year		10 year		15 year		20 year	
Prostate	98.8	0.4	95.2	0.9	87.1	1.7	81.1	3.0
Thyroid	96.0	0.8	95.8	1.2	94.0	1.6	95.4	2.1
Testis	94.7	1.1	94.0	1.3	91.1	1.8	88.2	2.3
Melanomas	89.0	0.8	86.7	1.1	83.5	1.5	82.8	1.9
Breast	86.4	0.4	78.3	0.6	71.3	0.7	65.0	1.0
Hodgkin's disease	85.1	1.7	79.8	2.0	73.8	2.4	67.1	2.8
Corpus uteri, uterus	84.3	1.0	83.2	1.3	80.8	1.7	79.2	2.0
Urinary, bladder	82.1	1.0	76.2	1.4	70.3	1.9	67.9	2.4
Cervix, uteri	70.5	1.6	64.1	1.8	62.8	2.1	60.0	2.4
Larynx	68.8	2.1	56.7	2.5	45.8	2.8	37.8	3.1
Rectum	62.6	1.2	55.2	1.4	51.8	1.8	49.2	2.3
Kidney, renal pelvis	61.8	1.3	54.4	1.6	49.8	2.0	47.3	2.6
Colon	61.7	0.8	55.4	1.0	53.9	1.2	52.3	1.6
Non-Hodgkin's	57.8	1.0	46.3	1.2	38.3	1.4	34.3	1.7
Oral cavity, pharynx	56.7	1.3	44.2	1.4	37.5	1.6	33.0	1.8
Ovary	55.0	1.3	49.3	1.6	49.9	1.9	49.6	2.4
Leukemia	42.5	1.2	32.4	1.3	29.7	1.5	26.2	1.7
Brain, nervous system	32.0	1.4	29.2	1.5	27.6	1.6	26.1	1.9
Multiple myeloma	29.5	1.6	12.7	1.5	7.0	1.3	4.8	1.5
Stomach	23.8	1.3	19.4	1.4	19.0	1.7	14.9	1.9
Lung and bronchus	15.0	0.4	10.6	0.4	8.1	0.4	6.5	0.4
Esophagus	14.2	1.4	7.9	1.3	7.7	1.6	5.4	2.0
Liver, bile duct	7.5	1.1	5.8	1.2	6.3	1.5	7.6	2.0
Pancreas	4.0	0.5	3.0	1.5	2.7	0.6	2.7	0.8

Simple ideas for effective tables

1. Minimize the use of guidelines.

Most publishers prohibit vertical lines in tables

Boxes around the whole table are chartjunk

Estimates of relative survival rates, by cancer site[9]

	% survival rates and standard errors							
	5 year		10 year		15 year		20 year	
Prostate	98.8	0.4	95.2	0.9	87.1	1.7	81.1	3.0
Thyroid	96.0	0.8	95.8	1.2	94.0	1.6	95.4	2.1
Testis	94.7	1.1	94.0	1.3	91.1	1.8	88.2	2.3
Melanomas	89.0	0.8	86.7	1.1	83.5	1.5	82.8	1.9
Breast	86.4	0.4	78.3	0.6	71.3	0.7	65.0	1.0
Hodgkin's disease	85.1	1.7	79.8	2.0	73.8	2.4	67.1	2.8
Corpus uteri, uterus	84.3	1.0	83.2	1.3	80.8	1.7	79.2	2.0
Urinary, bladder	82.1	1.0	76.2	1.4	70.3	1.9	67.9	2.4
Cervix, uteri	70.5	1.6	64.1	1.8	62.8	2.1	60.0	2.4
Larynx	68.8	2.1	56.7	2.5	45.8	2.8	37.8	3.1
Rectum	62.6	1.2	55.2	1.4	51.8	1.8	49.2	2.3
Kidney, renal pelvis	61.8	1.3	54.4	1.6	49.8	2.0	47.3	2.6
Colon	61.7	0.8	55.4	1.0	53.9	1.2	52.3	1.6
Non-Hodgkin's	57.8	1.0	46.3	1.2	38.3	1.4	34.3	1.7
Oral cavity, pharynx	56.7	1.3	44.2	1.4	37.5	1.6	33.0	1.8
Ovary	55.0	1.3	49.3	1.6	49.9	1.9	49.6	2.4
Leukemia	42.5	1.2	32.4	1.3	29.7	1.5	26.2	1.7
Brain, nervous system	32.0	1.4	29.2	1.5	27.6	1.6	26.1	1.9
Multiple myeloma	29.5	1.6	12.7	1.5	7.0	1.3	4.8	1.5
Stomach	23.8	1.3	19.4	1.4	19.0	1.7	14.9	1.9
Lung and bronchus	15.0	0.4	10.6	0.4	8.1	0.4	6.5	0.4
Esophagus	14.2	1.4	7.9	1.3	7.7	1.6	5.4	2.0
Liver, bile duct	7.5	1.1	5.8	1.2	6.3	1.5	7.6	2.0
Pancreas	4.0	0.5	3.0	1.5	2.7	0.6	2.7	0.8

Simple ideas for effective tables

2. Report only a few digits.

Don't report non-significant digits

Every extra digit distracts attention from the first, most important one

Estimates of relative survival rates, by cancer site[9]

	% survival rates and standard errors							
	5 year		10 year		15 year		20 year	
Prostate	98.8	0.4	95.2	0.9	87.1	1.7	81.1	3.0
Thyroid	96.0	0.8	95.8	1.2	94.0	1.6	95.4	2.1
Testis	94.7	1.1	94.0	1.3	91.1	1.8	88.2	2.3
Melanomas	89.0	0.8	86.7	1.1	83.5	1.5	82.8	1.9
Breast	86.4	0.4	78.3	0.6	71.3	0.7	65.0	1.0
Hodgkin's disease	85.1	1.7	79.8	2.0	73.8	2.4	67.1	2.8
Corpus uteri, uterus	84.3	1.0	83.2	1.3	80.8	1.7	79.2	2.0
Urinary, bladder	82.1	1.0	76.2	1.4	70.3	1.9	67.9	2.4
Cervix, uteri	70.5	1.6	64.1	1.8	62.8	2.1	60.0	2.4
Larynx	68.8	2.1	56.7	2.5	45.8	2.8	37.8	3.1
Rectum	62.6	1.2	55.2	1.4	51.8	1.8	49.2	2.3
Kidney, renal pelvis	61.8	1.3	54.4	1.6	49.8	2.0	47.3	2.6
Colon	61.7	0.8	55.4	1.0	53.9	1.2	52.3	1.6
Non-Hodgkin's	57.8	1.0	46.3	1.2	38.3	1.4	34.3	1.7
Oral cavity, pharynx	56.7	1.3	44.2	1.4	37.5	1.6	33.0	1.8
Ovary	55.0	1.3	49.3	1.6	49.9	1.9	49.6	2.4
Leukemia	42.5	1.2	32.4	1.3	29.7	1.5	26.2	1.7
Brain, nervous system	32.0	1.4	29.2	1.5	27.6	1.6	26.1	1.9
Multiple myeloma	29.5	1.6	12.7	1.5	7.0	1.3	4.8	1.5
Stomach	23.8	1.3	19.4	1.4	19.0	1.7	14.9	1.9
Lung and bronchus	15.0	0.4	10.6	0.4	8.1	0.4	6.5	0.4
Esophagus	14.2	1.4	7.9	1.3	7.7	1.6	5.4	2.0
Liver, bile duct	7.5	1.1	5.8	1.2	6.3	1.5	7.6	2.0
Pancreas	4.0	0.5	3.0	1.5	2.7	0.6	2.7	0.8

Simple ideas for effective tables

3. Order the table intelligently.

In a 2 dimensional table, order the rows and columns to highlight relationships

You can either

- **diagonalize** – sort based on order, or
- **cluster** – group based on similarity

More on this later...

Estimates of relative survival rates, by cancer site[9]

	% survival rates and standard errors							
	5 year		10 year		15 year		20 year	
Prostate	98.8	0.4	95.2	0.9	87.1	1.7	81.1	3.0
Thyroid	96.0	0.8	95.8	1.2	94.0	1.6	95.4	2.1
Testis	94.7	1.1	94.0	1.3	91.1	1.8	88.2	2.3
Melanomas	89.0	0.8	86.7	1.1	83.5	1.5	82.8	1.9
Breast	86.4	0.4	78.3	0.6	71.3	0.7	65.0	1.0
Hodgkin's disease	85.1	1.7	79.8	2.0	73.8	2.4	67.1	2.8
Corpus uteri, uterus	84.3	1.0	83.2	1.3	80.8	1.7	79.2	2.0
Urinary, bladder	82.1	1.0	76.2	1.4	70.3	1.9	67.9	2.4
Cervix, uteri	70.5	1.6	64.1	1.8	62.8	2.1	60.0	2.4
Larynx	68.8	2.1	56.7	2.5	45.8	2.8	37.8	3.1
Rectum	62.6	1.2	55.2	1.4	51.8	1.8	49.2	2.3
Kidney, renal pelvis	61.8	1.3	54.4	1.6	49.8	2.0	47.3	2.6
Colon	61.7	0.8	55.4	1.0	53.9	1.2	52.3	1.6
Non-Hodgkin's	57.8	1.0	46.3	1.2	38.3	1.4	34.3	1.7
Oral cavity, pharynx	56.7	1.3	44.2	1.4	37.5	1.6	33.0	1.8
Ovary	55.0	1.3	49.3	1.6	49.9	1.9	49.6	2.4
Leukemia	42.5	1.2	32.4	1.3	29.7	1.5	26.2	1.7
Brain, nervous system	32.0	1.4	29.2	1.5	27.6	1.6	26.1	1.9
Multiple myeloma	29.5	1.6	12.7	1.5	7.0	1.3	4.8	1.5
Stomach	23.8	1.3	19.4	1.4	19.0	1.7	14.9	1.9
Lung and bronchus	15.0	0.4	10.6	0.4	8.1	0.4	6.5	0.4
Esophagus	14.2	1.4	7.9	1.3	7.7	1.6	5.4	2.0
Liver, bile duct	7.5	1.1	5.8	1.2	6.3	1.5	7.6	2.0
Pancreas	4.0	0.5	3.0	1.5	2.7	0.6	2.7	0.8

Simple ideas for effective tables

3. Order the table intelligently.

In a 3+ dimensional table, nest the dimensions intelligently.

Note:

Table order applies to
1-dimensional plots,
like dot plots...

and to super tables of plots where rows or columns are categories

Estimates of relative survival rates, by cancer site[9]

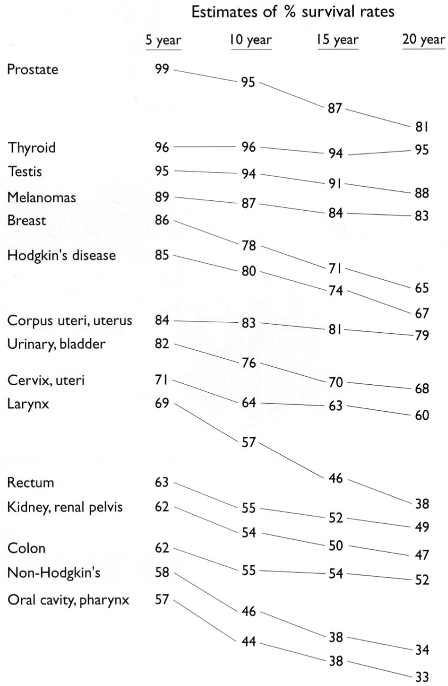
	% survival rates and standard errors							
	5 year		10 year		15 year		20 year	
Prostate	98.8	0.4	95.2	0.9	87.1	1.7	81.1	3.0
Thyroid	96.0	0.8	95.8	1.2	94.0	1.6	95.4	2.1
Testis	94.7	1.1	94.0	1.3	91.1	1.8	88.2	2.3
Melanomas	89.0	0.8	86.7	1.1	83.5	1.5	82.8	1.9
Breast	86.4	0.4	78.3	0.6	71.3	0.7	65.0	1.0
Hodgkin's disease	85.1	1.7	79.8	2.0	73.8	2.4	67.1	2.8
Corpus uteri, uterus	84.3	1.0	83.2	1.3	80.8	1.7	79.2	2.0
Urinary, bladder	82.1	1.0	76.2	1.4	70.3	1.9	67.9	2.4
Cervix, uteri	70.5	1.6	64.1	1.8	62.8	2.1	60.0	2.4
Larynx	68.8	2.1	56.7	2.5	45.8	2.8	37.8	3.1
Rectum	62.6	1.2	55.2	1.4	51.8	1.8	49.2	2.3
Kidney, renal pelvis	61.8	1.3	54.4	1.6	49.8	2.0	47.3	2.6
Colon	61.7	0.8	55.4	1.0	53.9	1.2	52.3	1.6
Non-Hodgkin's	57.8	1.0	46.3	1.2	38.3	1.4	34.3	1.7
Oral cavity, pharynx	56.7	1.3	44.2	1.4	37.5	1.6	33.0	1.8
Ovary	55.0	1.3	49.3	1.6	49.9	1.9	49.6	2.4
Leukemia	42.5	1.2	32.4	1.3	29.7	1.5	26.2	1.7
Brain, nervous system	32.0	1.4	29.2	1.5	27.6	1.6	26.1	1.9
Multiple myeloma	29.5	1.6	12.7	1.5	7.0	1.3	4.8	1.5
Stomach	23.8	1.3	19.4	1.4	19.0	1.7	14.9	1.9
Lung and bronchus	15.0	0.4	10.6	0.4	8.1	0.4	6.5	0.4
Esophagus	14.2	1.4	7.9	1.3	7.7	1.6	5.4	2.0
Liver, bile duct	7.5	1.1	5.8	1.2	6.3	1.5	7.6	2.0
Pancreas	4.0	0.5	3.0	1.5	2.7	0.6	2.7	0.8

The table at left (from Tufte) is effectively designed

It is *diagonalized*, uses *few digits*, and facilitates *lookup*

But tables always limit *comparison*

The brain is slower to grasp numerals than graphical representations of numbers

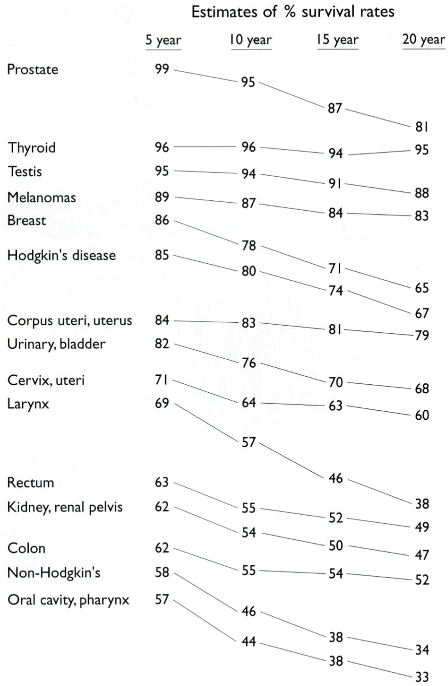


This figure (also from Tufte) may be an improvement

It keeps (almost) all the virtues of the table, but also makes comparison easier

Instead of digging information out of the table, it now hits the reader "between the eyes"

What's the scale?



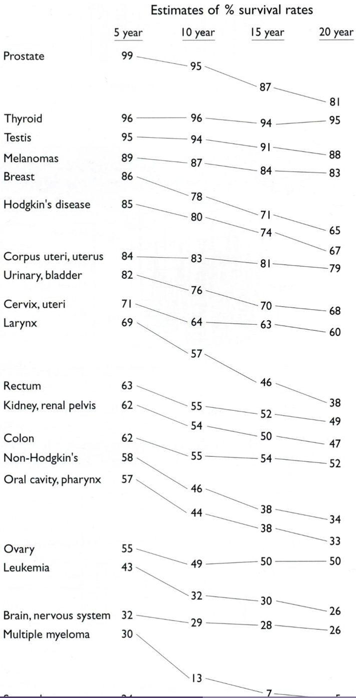
This figure (also from Tufte) may be an improvement

It keeps (almost) all the virtues of the table, but also makes comparison easier

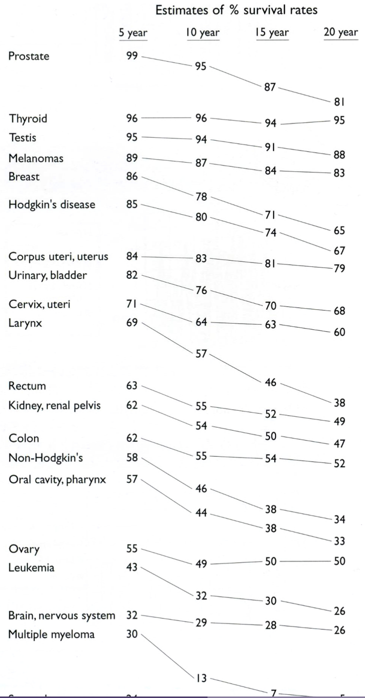
Instead of digging information out of the table, it now hits the reader "between the eyes"

What's the scale?

There isn't one!



What's missing from the figure that was in the table?



What's missing from the figure that was in the table?

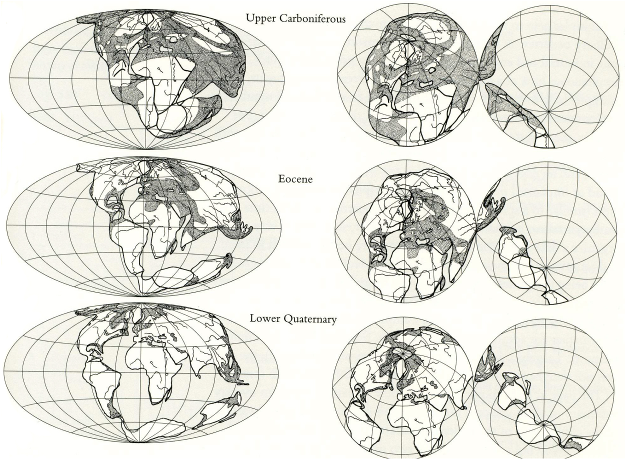
Measures of uncertainty.

The table had standard errors

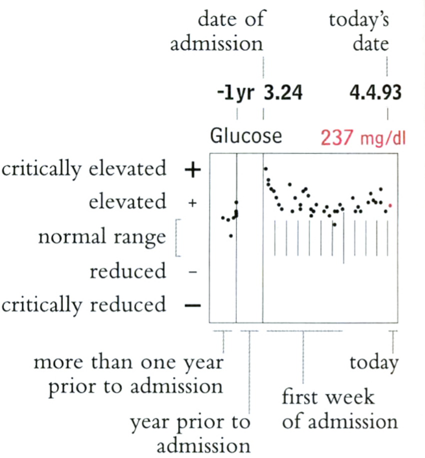
A major focus of this course is including uncertainty in plots like this one

Small Multiples: Repetition of a display concept

Tufte's most useful concept



Source: Tufte, *Visual Explanations*



Tufte proposes medical charts follow the format at left

This chart is annotated for pedagogical purposes

Lots of information; little distracting scaffolding

But the real pay off of this model plot is that it can be repeated once learned...

A complete layout using small multiples to convey hundreds of pieces of information

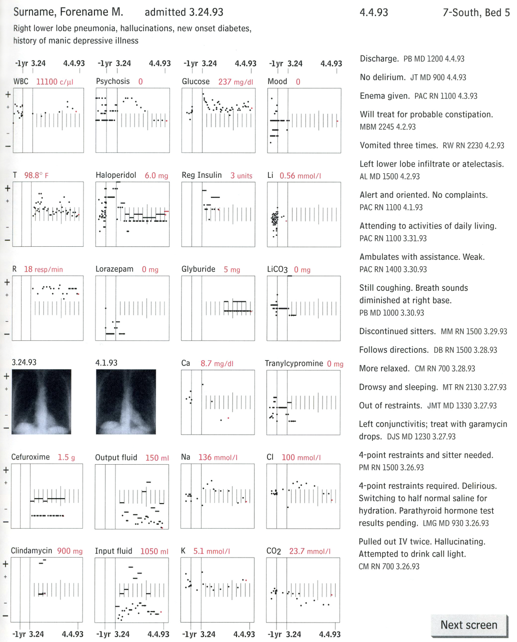
Elegant, information-rich, hard to make ...Goal: tools to make this easier

How to

Make Small Multiples

1. Start with a detailed, multifunctional plot

2. Then tile these plots to incorporate extra variables & dimensions



1. Start with a detailed, multifunctional plot

2. Then tile these plots to incorporate extra variables & dimensions

The principles of small multiples extend to virtually any VDSI

Were the inscriptions on Trajan's column cut individually, or pressed from a single mold?

Were the inscriptions on Trajan's column cut individually, or pressed from a single mold?

Small multiples – tiled & overlapped – offer an elegant solution

Source: Tufte, *Visual Explanations*

Karen Cheng (UW–Art) uses the same technique in *Designing Type* (2006, Yale Univ Press) to explain differences across fonts

Scientific Visuals and InfoVis

Big data and cheap computing created demand and opportunity for better data visuals in the media

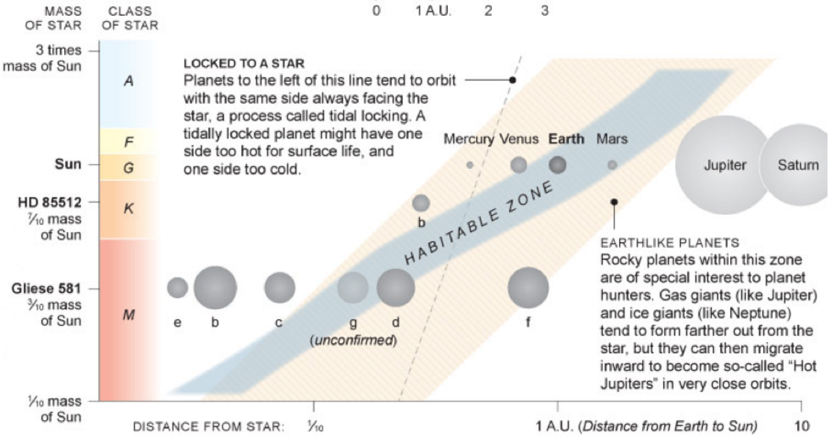
Graphic designers, computer scientists, and journalists have responded:
Information Visualization, or InfoVis

Beautiful, data-rich graphics for exploring public data

Different goals from scientific visulization of data

InfoVis: emphasis on fun, exploration, beauty, and "wow"

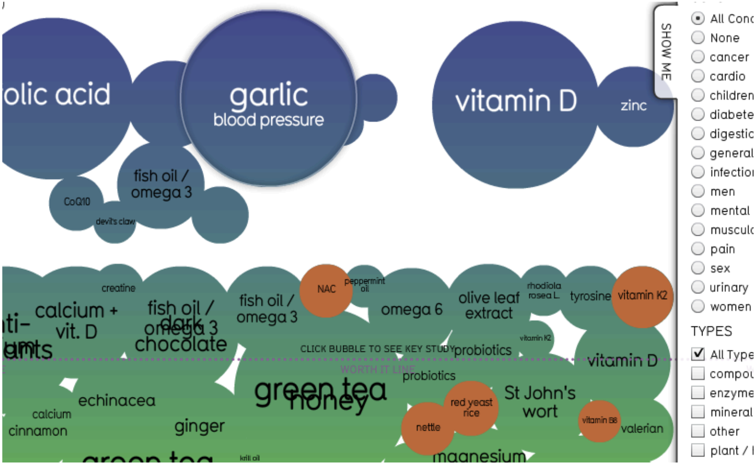
Scientific Visuals: structured comparison, precision, and inference



Source: Jonathan Corum, *New York Times*, "Habitable Zones,"

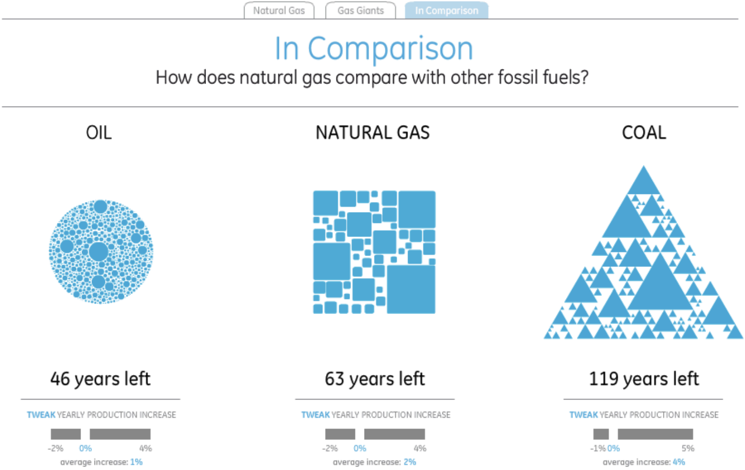
www.nytimes.com/interactive/2011/12/03/science/space/1202-planet.html

For a good reason, break the rules



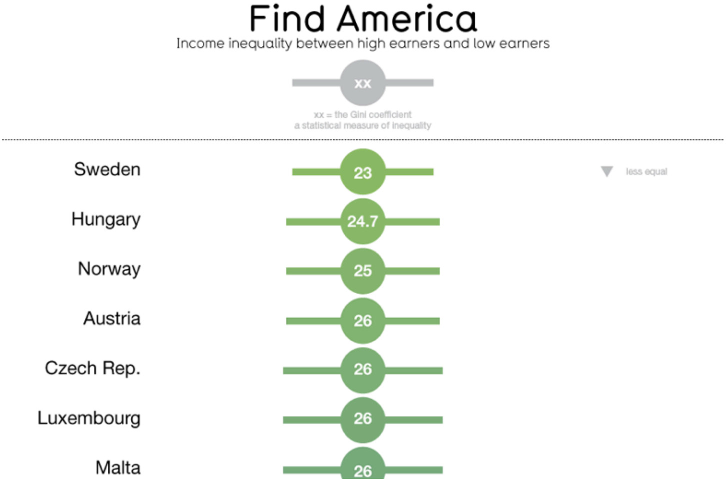
Source: David McCandless and collaborators, “Snake Oil,” informationisbeautiful.net/visualizations/snake-oil-scientific-evidence-for-nutritional-supplements-vizsweet/

But the rules exist for a reason



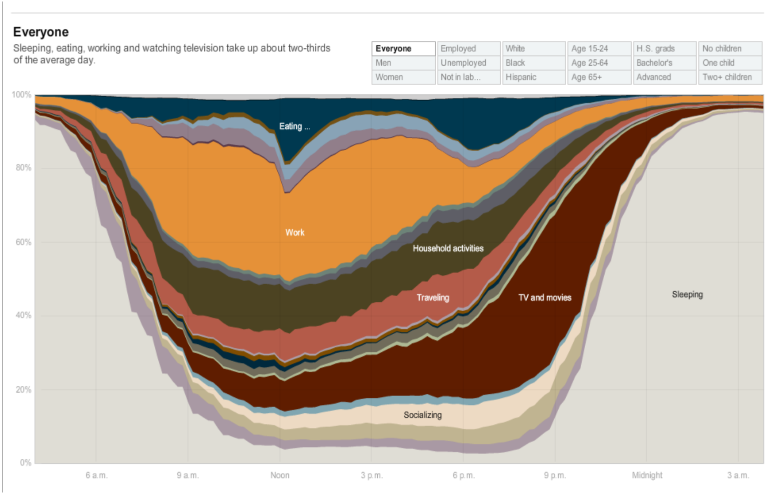
Source: David McCandless for General Electric,

`visualization.geblogs.com/visualization/gas`



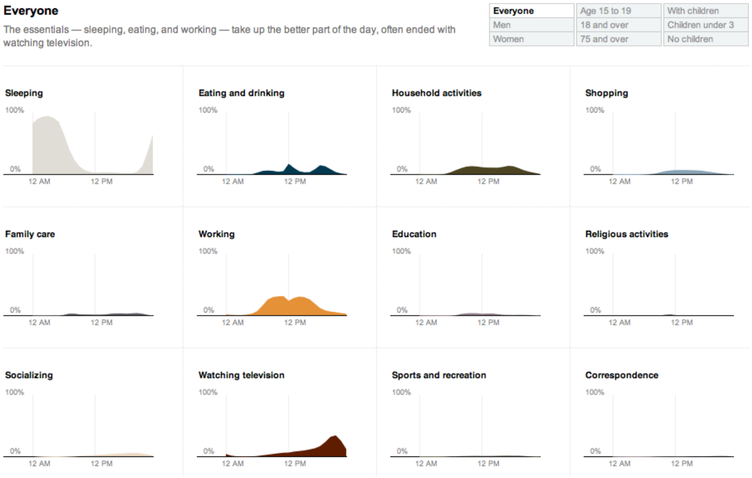
Source: David McCandless and others, www.informationisbeautiful.net/visualizations/what-are-wallst-protestors-angry-about

DataVis Approach

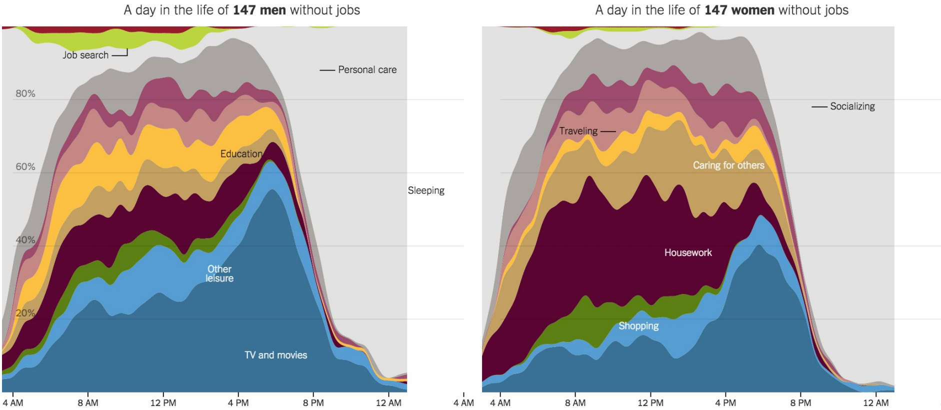


Source: Amanda Cox (CSSS grad) et al, "How different groups spend their day," *New York Times*, 8/2/2009, `http://www.nytimes.com//interactive/2009/07/31/`

VDSI approach



Source: Nathan Yau, "How Americans spend their day",
`projects.flowingdata.com/timeuse`



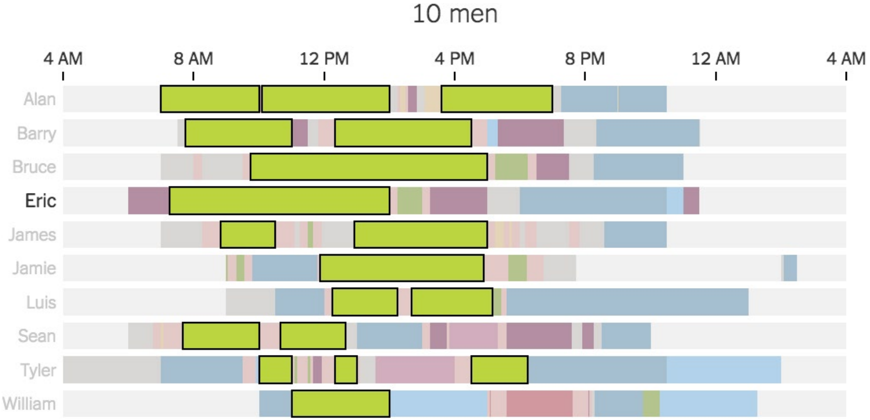
A beautiful synthesis is possible (note the example now focuses on the unemployed)

NYT's Upshot group revisited this problem in January 2015 using a blend of data visualization techniques and InfoVis polish

Josh Katz, "How nonemployed Americans spend their weekdays: Men vs. Women," *New York Times*,

`http://www.nytimes.com/interactive/2015/01/06/upshot/`

`how-nonemployed-americans-spend-their-weekdays-men-vs-women.html`



The Upshot's new approach follows Tufte's principles:

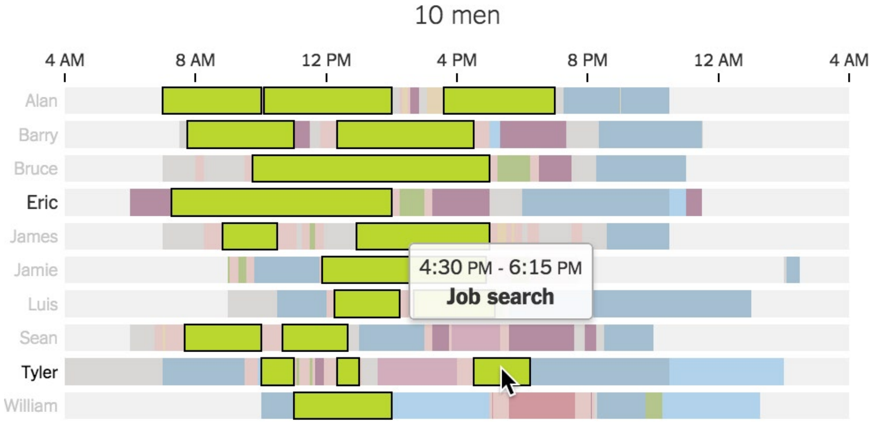
1. Show as much data as possible

here, down to lowest level of individual by time by use

Josh Katz, "How nonemployed Americans spend their weekdays: Men vs. Women," *New York Times*,

http://www.nytimes.com/interactive/2015/01/06/upshot/

how-nonemployed-americans-spend-their-weekdays-men-vs-women.html



The Upshot's approach follows Tufte's principles:

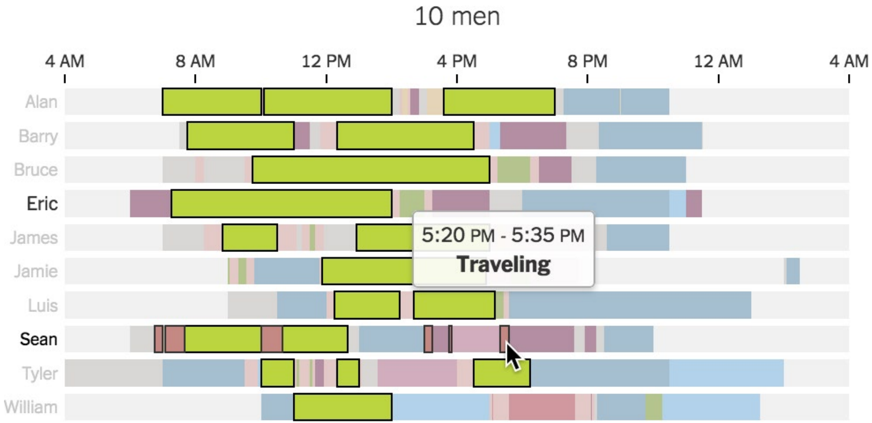
2. Establish the logic of the graphic with a detailed example

   here, through an interactive display that allows user to select each use

Josh Katz, "How nonemployed Americans spend their weekdays: Men vs. Women," *New York Times*,

`http://www.nytimes.com/interactive/2015/01/06/upshot/`

`how-nonemployed-americans-spend-their-weekdays-men-vs-women.html`



The Upshot's approach follows Tufte's principles:

2. Establish the logic of the graphic with a detailed example

   here, through an interactive display that allows user to select each use

Josh Katz, "How nonemployed Americans spend their weekdays: Men vs. Women," *New York Times*,

`http://www.nytimes.com/interactive/2015/01/06/upshot/`

`how-nonemployed-americans-spend-their-weekdays-men-vs-women.html`

TV and movies

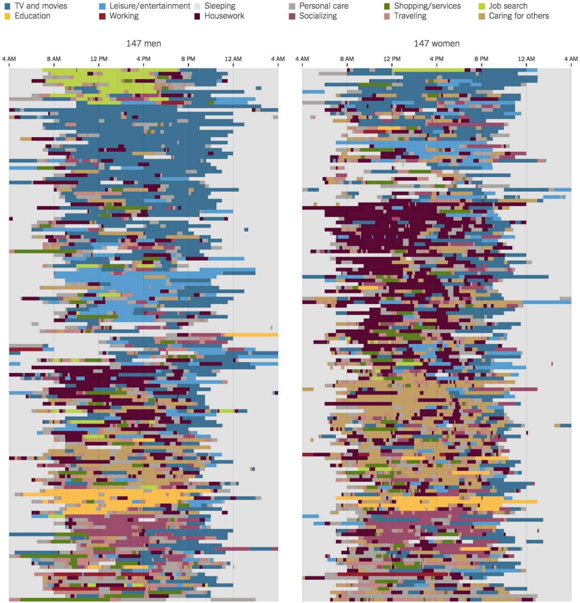
Watching television and movies is a significantly more common activity for the nonemployed than looking for work. For every one person whose main activity was job searching, there were almost six whose main activity was television and movie watching.

The gender breakdown is striking. Of the 65 people who devoted more of their daytime to watching TV and movies than any other activity, 46 are men versus 19 who are women.

With the design established, the Upshot used a series of small multiples to show differences across sexes

Here, they show all respondents whose largest time use was television

Recommendation: sort rows so similar individuals are stacked together (cluster analysis)



Finally, an ambitious graphic shows all data at once

This can only work by *grouping* individuals first by sex (columns of plots),

and then by similarity of time usage (rows of data, sorted via cluster analysis)

Turning points in the history of visual displays

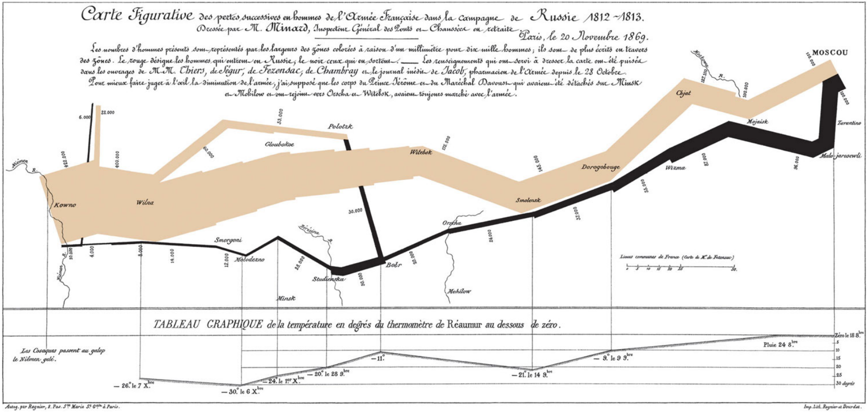
Tufte's principles for information design

How data visualization differs from infovis

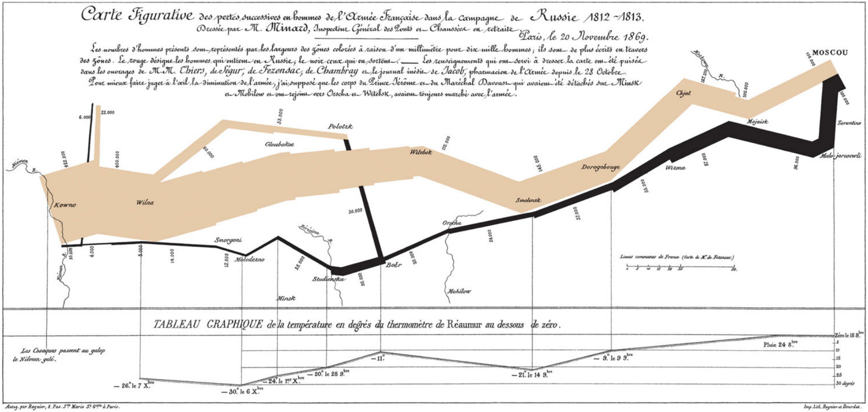
Scales and scaling

Making a scatterplot from scratch

Sorting in tables and table-like figures

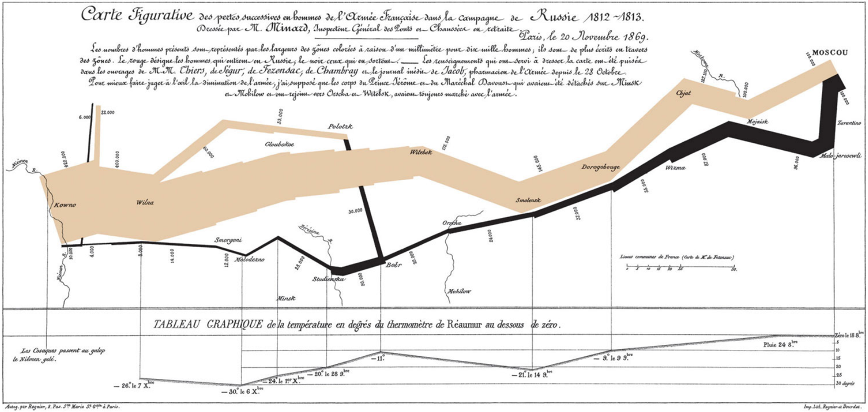


Perhaps the best plot ever: Minard's display of Napoleon's March on Moscow

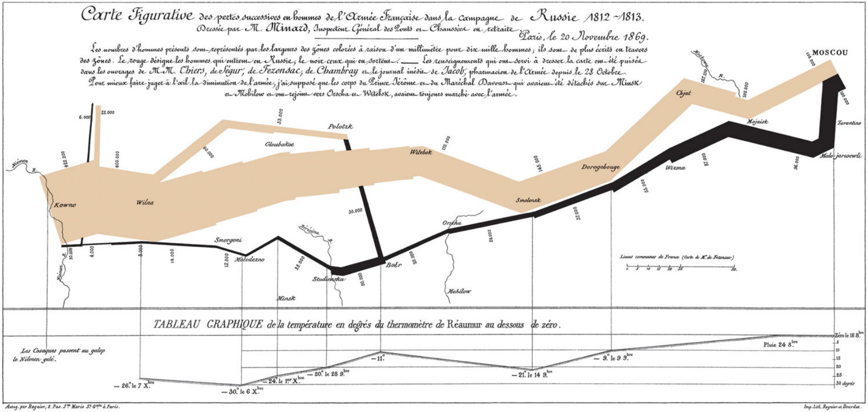


Some plots combine multiple elements into a single super-plot or **confection**

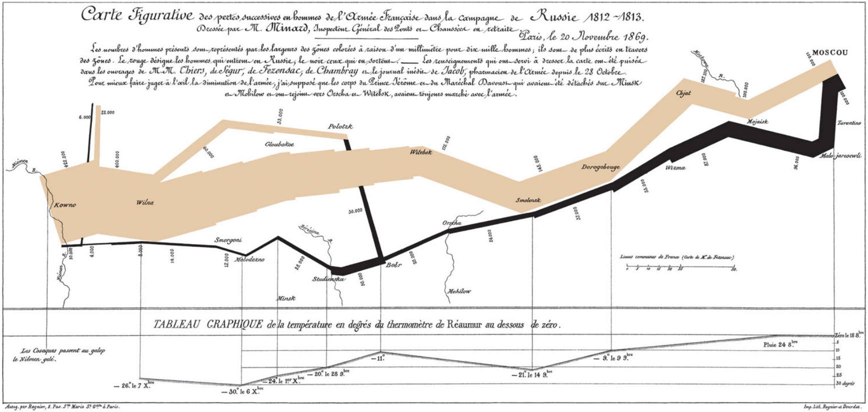
Good confections are better than the sum of their components
because they facilitate connection or comparison across elements



How many dimensions?



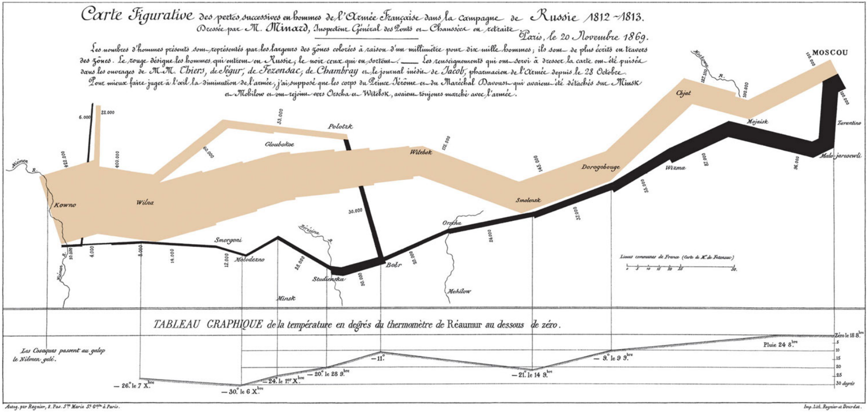
How many dimensions? 7



How many dimensions? ***7***

1. Latitude of army & features

   Y-coordinate



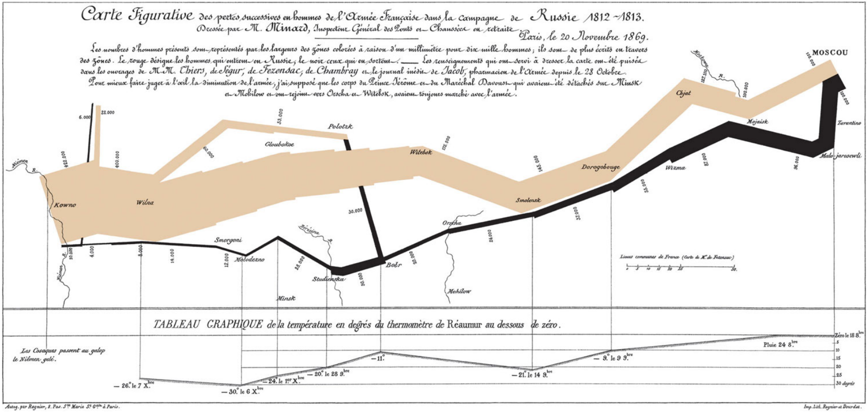
How many dimensions? ***7***

1. Latitude of army & features

   Y-coordinate

2. Longitude of army & features

   X-coordinate



How many dimensions? 7

1. Latitude of army & features

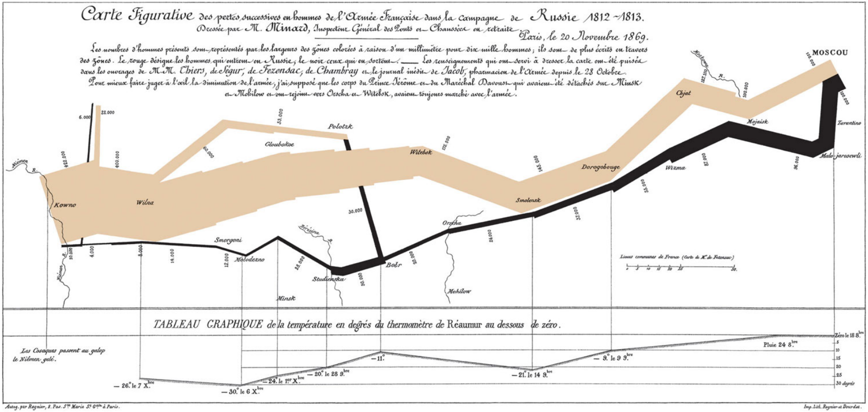
   Y-coordinate

2. Longitude of army & features

   X-coordinate

3. Size of army

   width of line, numerals



*How many dimensions? **7***

1. Latitude of army & features

   Y-coordinate

2. Longitude of army & features

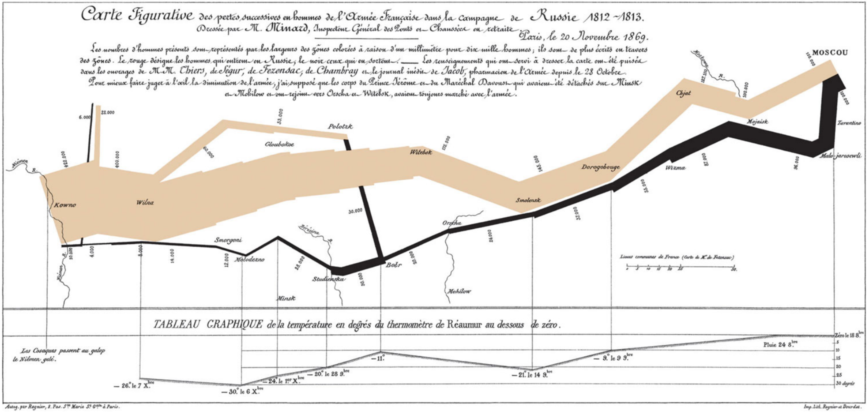
   X-coordinate

3. Size of army

   width of line, numerals

4. Advance vs. Retreat

   color of line



How many dimensions? ***7***

1. Latitude of army & features

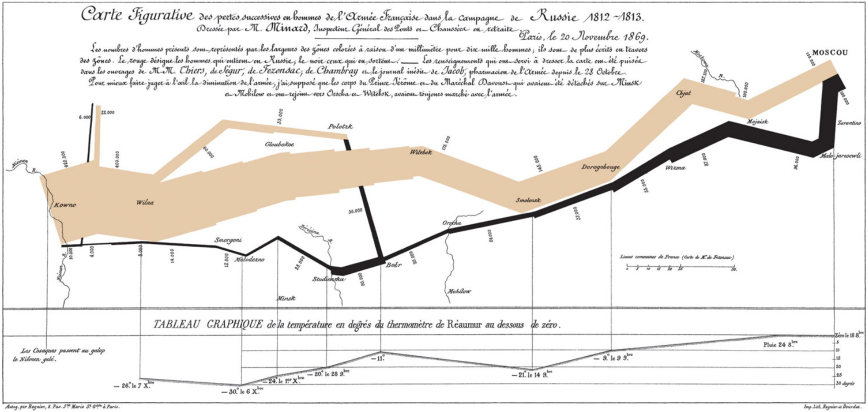
   Y-coordinate
2. Longitude of army & features

   X-coordinate
3. Size of army

   width of line, numerals
4. Advance vs. Retreat

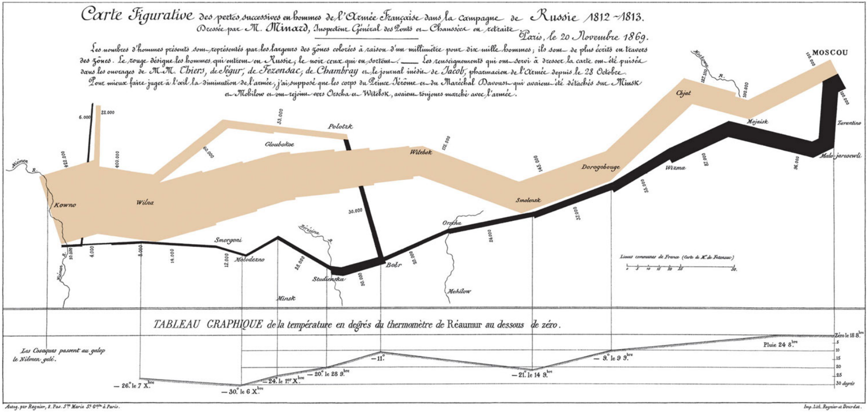
   color of line
5. Division of army

   splitting of line



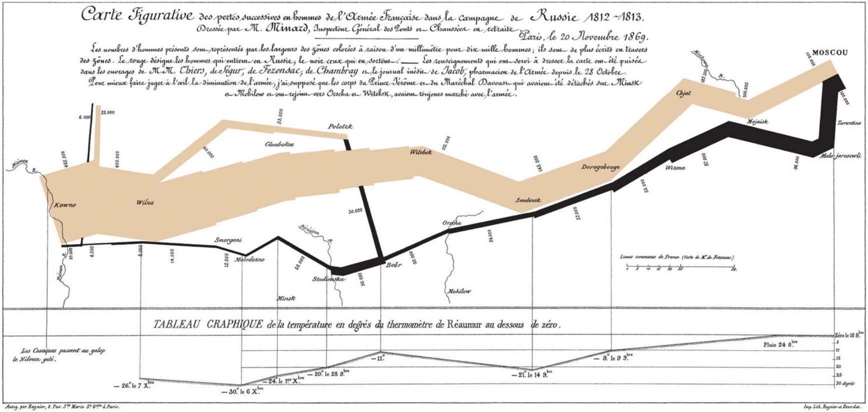
How many dimensions? ***7***

1. Latitude of army & features
   Y-coordinate
2. Longitude of army & features
   X-coordinate
3. Size of army
   width of line, numerals
4. Advance vs. Retreat
   color of line
5. Division of army
   splitting of line
6. Temperature
   linked lineplot



*How many dimensions? **7***

1. Latitude of army & features
   Y-coordinate
2. Longitude of army & features
   X-coordinate
3. Size of army
   width of line, numerals
4. Advance vs. Retreat
   color of line
5. Division of army
   splitting of line
6. Temperature
   linked lineplot
7. Time
   linked lineplot



Combines narrative & analysis: a technique mostly lost until this century

May be a spurious relationship here: time and temperature

Note the deaths at river crossings – usually, these rivers would be frozen

Did Napoleon choose too warm a winter to invade Russia?

Other approaches to VDSI?

Tufte is best known, but many other people work on VDSIs

Jacques Bertin, J. W. Tukey are foundational figures, as is...

William Cleveland (statistician), who emphasizes:

- Avoiding cognitive pitfalls
- Transforming data to highlight relationships
- Combining data and model fits: unique ability of graphics

Later weeks of 569 focus on Cleveland's contributions (Cognition, EDA)

Scaling

Let's turn from general principles to specific

Scaling is an issue for most scientific graphics

Good scaling incorporates many concerns of Tufte & Cleveland:

Graphical integrity

Highlighting relationships

Facilitating comparison

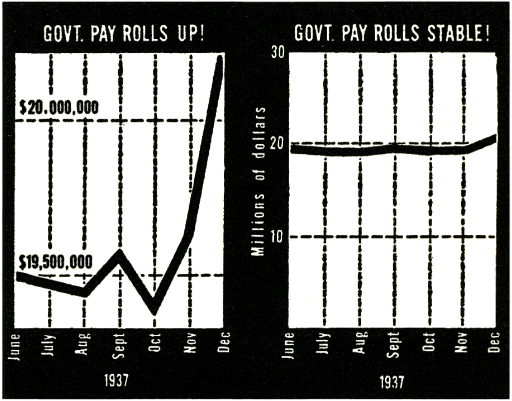
Maximixing data ink

Two parts to scaling: **anchoring** and **stretching**

Some people assert as a "rule-of-thumb" that graphical axes must always include zero

Hoff's example at the right is the likely origin of this misleading advice

(To be sure, Hoff never offers it as a rule)

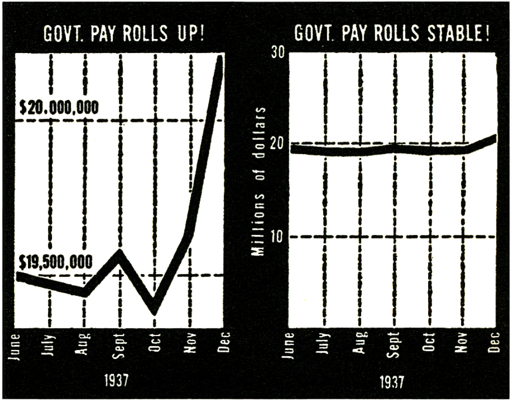


Source: Darrell Hoff, *How to Lie with Statistics*

Choose scales carefully!

- Start point
- End point
- Units (usual choices:
  linear or log)
- Parallel scales (*optional*)

You choices depend on what you want to show and *compare*, not a general rule

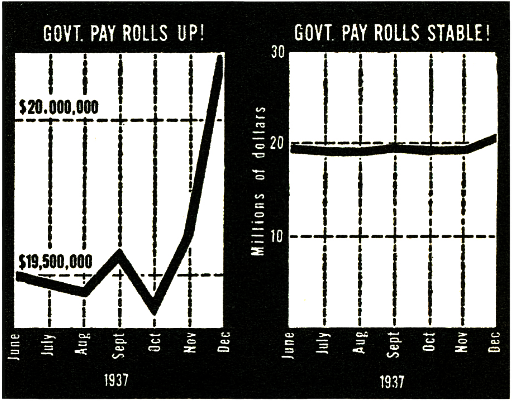


Source: Darrell Hoff, *How to Lie with Statistics*

A Myth About Scaling

Even in Hoff's example, the left plot is *better* than the right plot

Public budgets are usually very sticky, and 3% changes can be a big deal

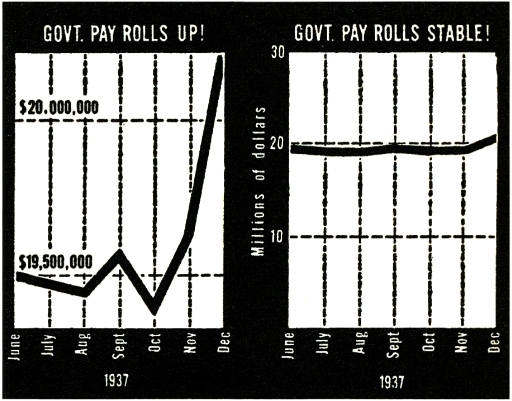


Source: Darrell Hoff, *How to Lie with Statistics*

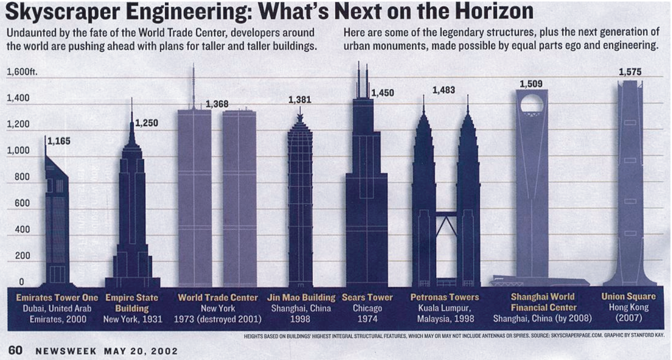
Neither plot is ideal – instead, a scale that corresponds to the “usual range” in which budgets might vary would be a better choice

But that suggests the plot is incomplete until compared with *another* set of data

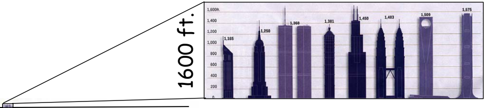
...leading back to Tufte’s recommendation to plot *small multiples*

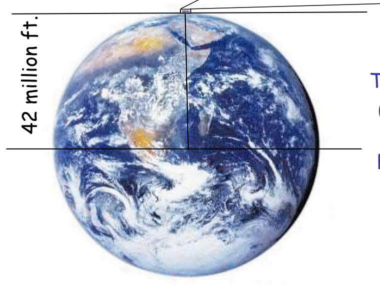


Source: Darrell Hoff, *How to Lie with Statistics*



What's "zero," exactly, in plots like this one? The ground?





Tallest buildings on Earth are a
$(1.6 \times 10^3)/(4.2 \times 10^7) \times 100 =$
0.0038%
blip added to the diameter of the planet.

Source: Alyssa Goodman, Harvard-Smithsonian Center for Astrophysics

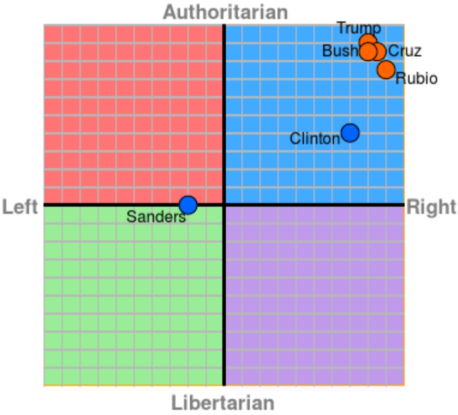
The center of the planet?

Anchoring the scale

Sometimes, there is no meaningful baseline at all

Is there a defensible "zero point" on ideology?

2016 US Presidential Primaries



Source: `www.politicalcompass.org`

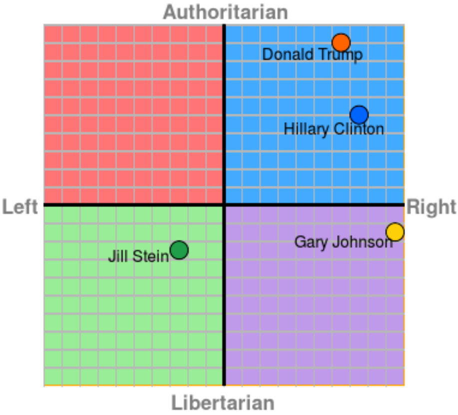
Anchoring the scale

Sometimes, there is no meaningful baseline at all

Is there a defensible "zero point" on ideology?

Can you guess Political Compass's ideological preference?

2016 US Presidential General



Source: `www.politicalcompass.org`

Stretching the scale

The most natural scaling for axes is linear

Each inch on the page $= k$ units on the scale.

Linear scales are just one option, and not helpful for skewed data

Logarthmic scaling is often better:

- doesn't hide data in the corner
- allows linear fits to log data

Remember to label on the original scale

If you do print exponents, indicate the base of the log (e, 10, etc.)

Choose carefully between linear, log-linear, and log-log plots
to facilitate useful comparison of data points

Making a scatterplot from scratch

Systems of political parties differ across democracies

In some countries (like the US), two major parties vie to win elections by turning out their base plus middle class swing voters

In other countries (most European countries) there are many parties, including large parties dominated by the working class

Some political scientists claim having more parties – including worker parties – produces more redistribution

Source: Torben Iversen & David Soskice, 2002,
"Why do some democracies redistribute more than others?" Harvard University.

Making a scatterplot from scratch

Concepts for this example

Effective number of parties:

- Number of parties varies across countries
- Electoral rules largely determine potential number parties
  - Winner take all (US) $\rightarrow \approx 2$ parties
  - Proportional representation $\rightarrow$ more parties
- To see this, discount trivial parties and use *effective number of parties*

Making a scatterplot from scratch

Concepts for this example

Effective number of parties:

- Number of parties varies across countries
- Electoral rules largely determine potential number parties
  - Winner take all (US) $\rightarrow \approx 2$ parties
  - Proportional representation $\rightarrow$ more parties
- To see this, discount trivial parties and use *effective number of parties*

Poverty reduction:

Making a scatterplot from scratch

Concepts for this example

Effective number of parties:

- Number of parties varies across countries
- Electoral rules largely determine potential number parties
  - Winner take all (US) $\rightarrow \approx 2$ parties
  - Proportional representation $\rightarrow$ more parties
- To see this, discount trivial parties and use *effective number of parties*

Poverty reduction:

- Percent lifted out of poverty by taxes and transfers
- Poverty = an income below 50% of mean income

What if we mapped the data?

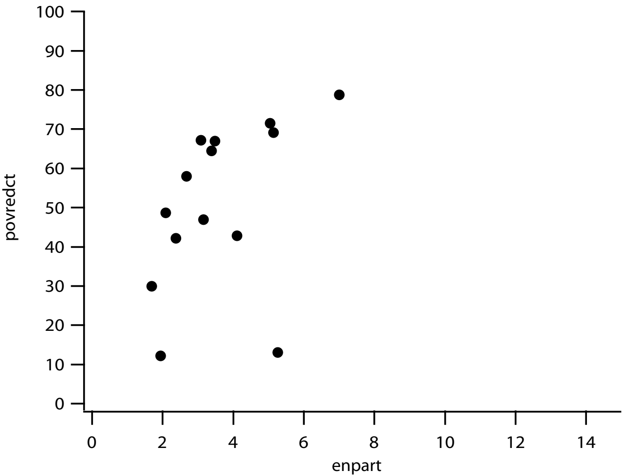
This map of poverty reduction is colorful, but how to relate it to the number of parties?

This map of poverty reduction is colorful, but how to relate it to the number of parties?

A pair of maps?

Multiple maps relate variables to geography, not to each other

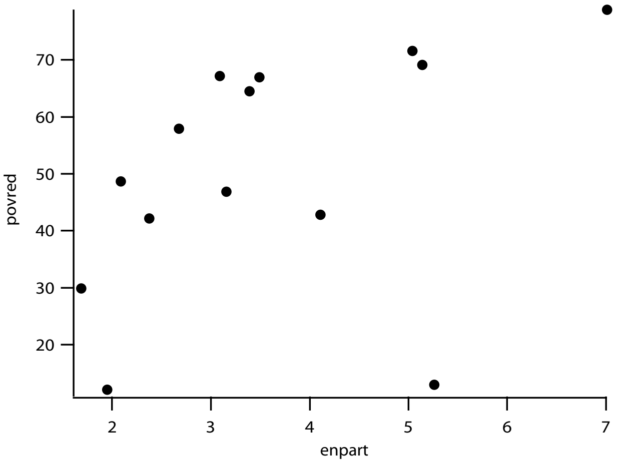
Geography is incidental to our example: we need a scatterplot to focus on variables



Initial plotting area is often oddly shaped (I've exaggerated)

This plotting area hides the relationship. Sometimes it can even exclude data!

Aside: Filled symbols are good for a little data, but open symbols are better when data overlap

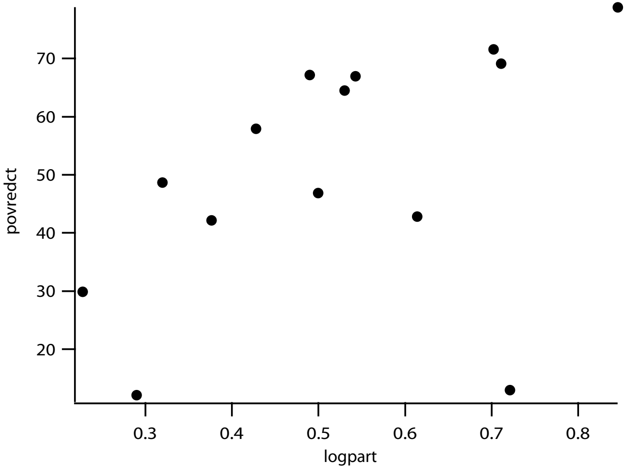


Critical *user* decision: Impose sensible, data based plot limits

Even more important with small multiples: unreadable without consistent, substantively-driven limits across plots

Don't just leave this to your package to decide

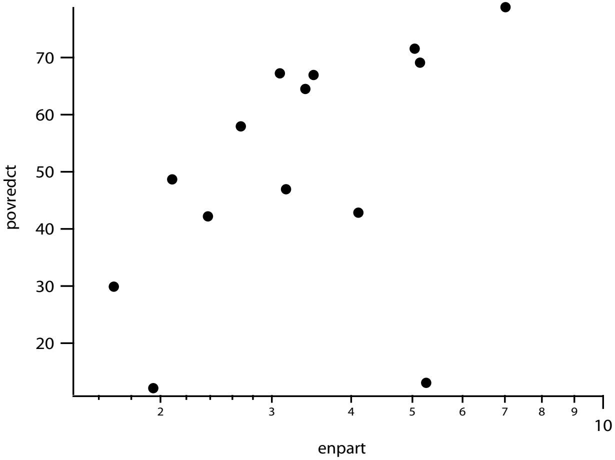
There appears to be a curvilinear relationship. We can bring that out with...



Log scaling.

But why print the exponents?

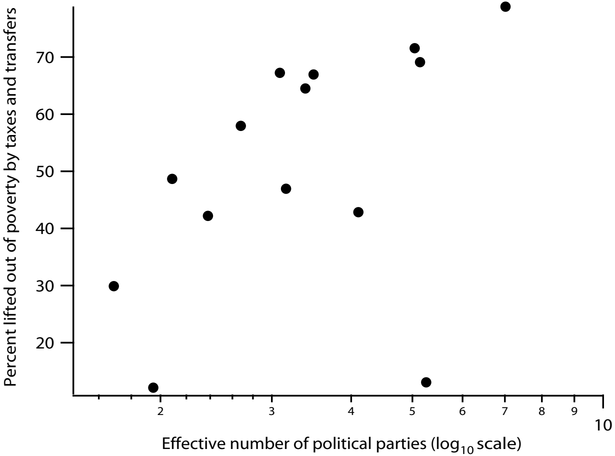
Logs aren't intuitive for many readers, but they don't need to even know we are using them in a graphic...



To make log scales easier for *everyone* to read, use a log scale but supply linear labels...

That is, plot the tick markets at the log values (exponents), but label them with the original linear scale numbers corresponding to those tick marks

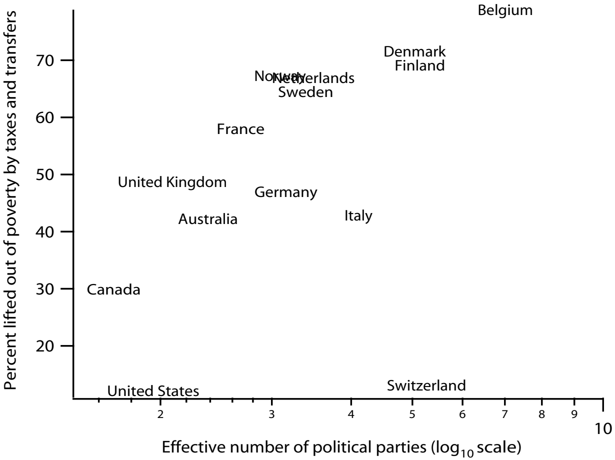
Next problem: Why use abbreviated computer labels for our variables?



Computer labels should stay in the computer

Write out informative axis titles

Next question: What are those outliers?

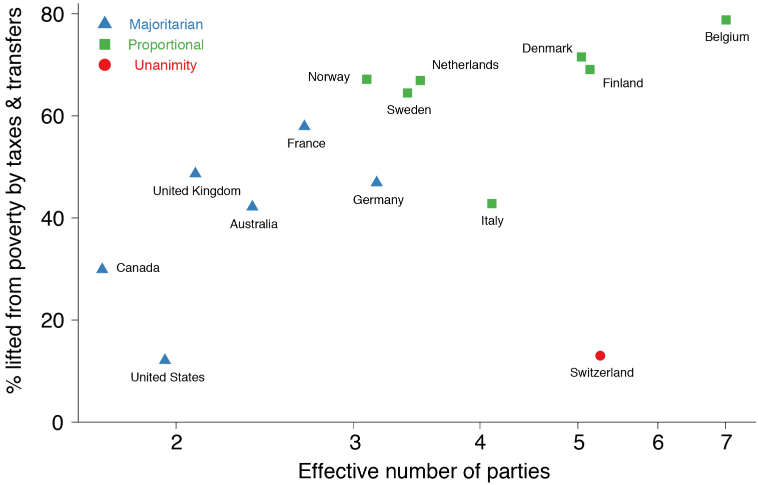


With only a little data & some big outliers, we should show the name of each case as a label

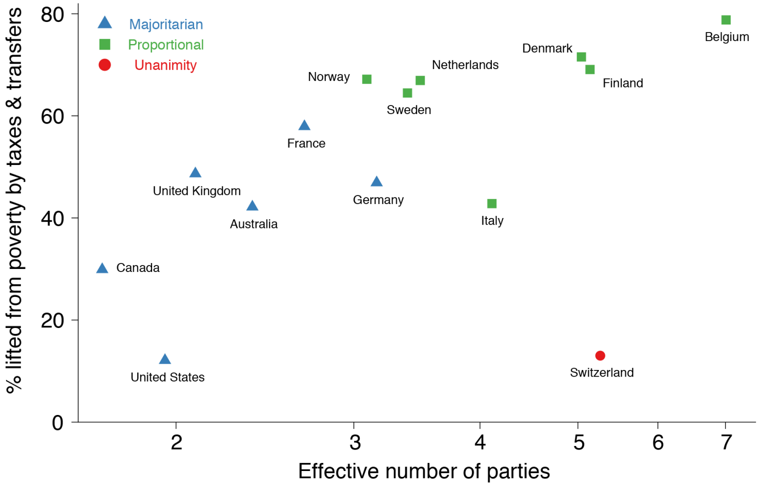
Sometimes we can just replace our plotted points with these labels

Here, let's combine the glyph (symbol) and text label for each point, so that we can use our glyphs to encode a third variable

Next, we can try to figure out what makes the US and Switzerland so different

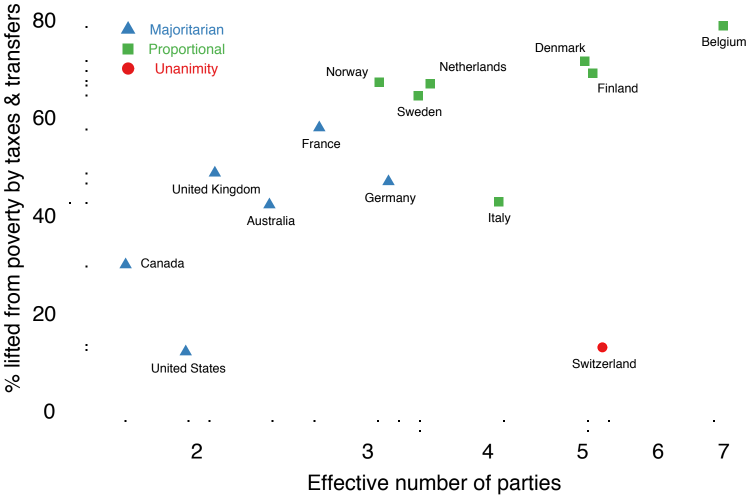


This plot and following plots are made using `scatter` (`tile` package in `R`)



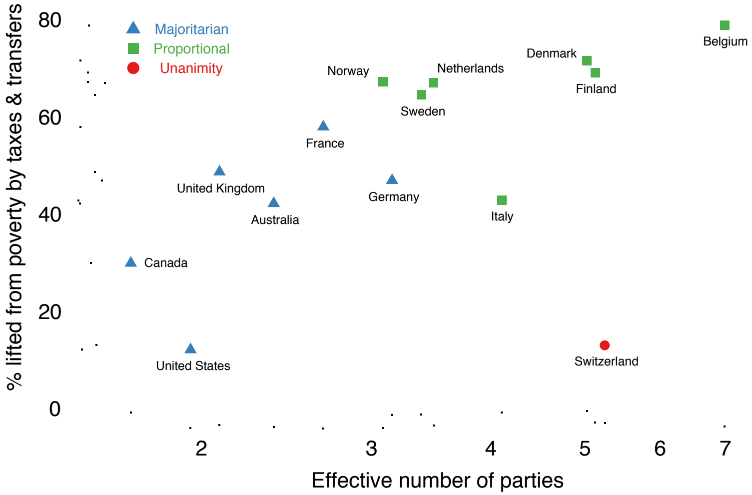
Scatterplots relate two distributions.

Why not make those marginal distributions explicit?

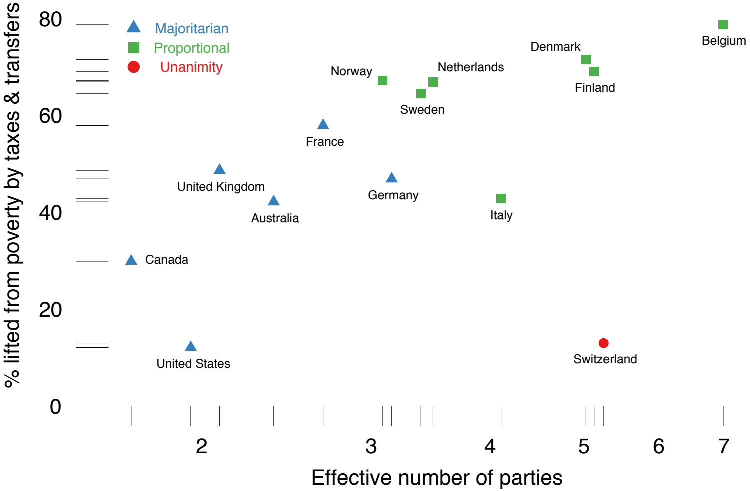


Rugs **accomplish this by replacing the axis lines with the plots**

We could choose any plotting style: from the histogram-like dots...

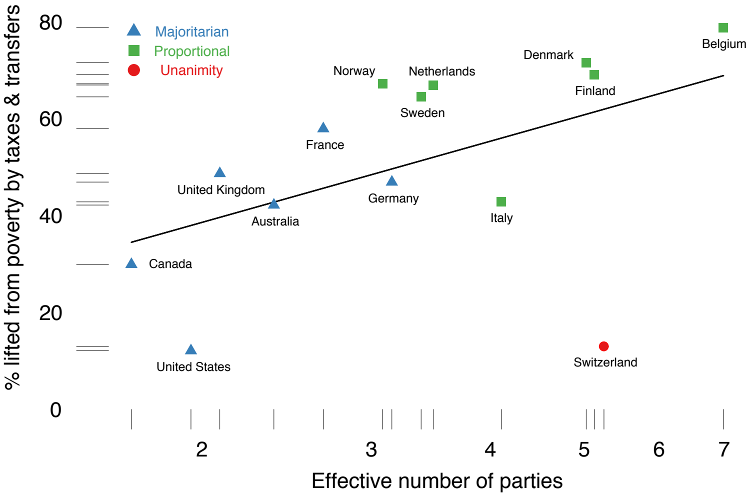


...to a strip of `jittered` data...



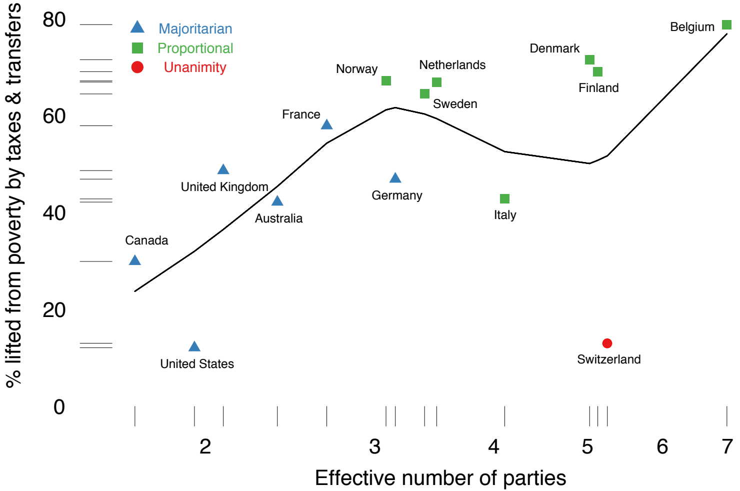
...to a set of very thin `lines` marking each observation

Because we have so few cases, thin lines work best for this example



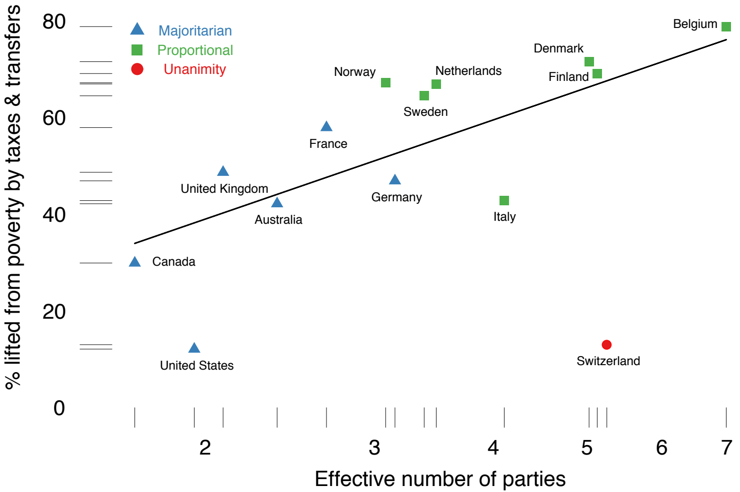
Let's add a parametric model of the data: a least squares fit line

`tile` can do this for us

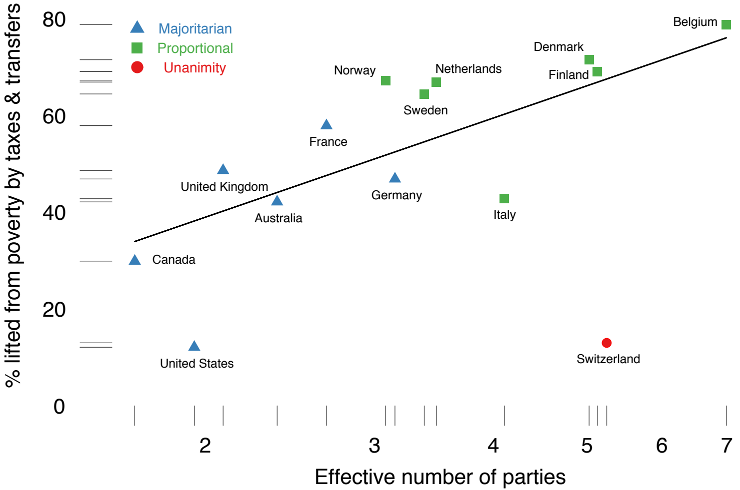


But we don't have to be parametric

A local smoother, like loess, often helps show non-linear relationships

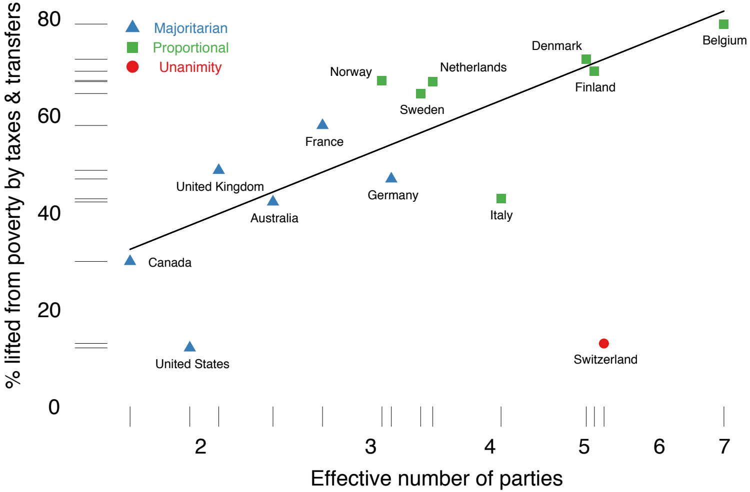


M-estimators weight observations by an influence function to minimize the influence of outliers



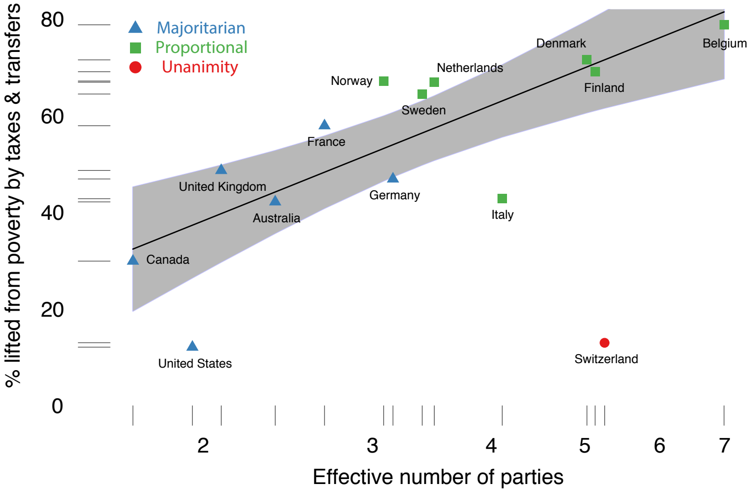
Even with an M-estimator, every outlier has some influence

Thus any one distant outlier can bias the result



A robust and resistant MM-estimator, shown above, largely avoids this problem

Only a (non-outlying) fraction of the data influence this fit. `rlm(method="MM")`



In our final plot, we add 95 percent confidence intervals for the MM-estimator

A measure of uncertainty is essential to reader confidence in the result

Steps to make a scatterplot

1. Decide on dimensions: aspect ratio, axis limits

Steps to make a scatterplot

1. Decide on dimensions: aspect ratio, axis limits
2. Add axis labels, plot titles

Steps to make a scatterplot

1. Decide on dimensions: aspect ratio, axis limits
2. Add axis labels, plot titles
3. Choose data markers: points, symbols, text

Steps to make a scatterplot

1. Decide on dimensions: aspect ratio, axis limits
2. Add axis labels, plot titles
3. Choose data markers: points, symbols, text
4. Scaling & transformation, add ticks if needed

Steps to make a scatterplot

1. Decide on dimensions: aspect ratio, axis limits
2. Add axis labels, plot titles
3. Choose data markers: points, symbols, text
4. Scaling & transformation, add ticks if needed
5. Choose a color palette

Steps to make a scatterplot

1. Decide on dimensions: aspect ratio, axis limits
2. Add axis labels, plot titles
3. Choose data markers: points, symbols, text
4. Scaling & transformation, add ticks if needed
5. Choose a color palette
6. Add annotations: labels, arrows, notes

Steps to make a scatterplot

1. Decide on dimensions: aspect ratio, axis limits
2. Add axis labels, plot titles
3. Choose data markers: points, symbols, text
4. Scaling & transformation, add ticks if needed
5. Choose a color palette
6. Add annotations: labels, arrows, notes
7. Add best-fit line(s) & confidence intervals

Steps to make a scatterplot

1. Decide on dimensions: aspect ratio, axis limits
2. Add axis labels, plot titles
3. Choose data markers: points, symbols, text
4. Scaling & transformation, add ticks if needed
5. Choose a color palette
6. Add annotations: labels, arrows, notes
7. Add best-fit line(s) & confidence intervals
8. Add extra plots (e.g., rugs) to make a confection

Steps to make a scatterplot

1. Decide on dimensions: aspect ratio, axis limits
2. Add axis labels, plot titles
3. Choose data markers: points, symbols, text
4. Scaling & transformation, add ticks if needed
5. Choose a color palette
6. Add annotations: labels, arrows, notes
7. Add best-fit line(s) & confidence intervals
8. Add extra plots (e.g., rugs) to make a confection
9. Repeat as small multiples

Steps to make a scatterplot

1. Decide on dimensions: aspect ratio, axis limits
2. Add axis labels, plot titles
3. Choose data markers: points, symbols, text
4. Scaling & transformation, add ticks if needed
5. Choose a color palette
6. Add annotations: labels, arrows, notes
7. Add best-fit line(s) & confidence intervals
8. Add extra plots (e.g., rugs) to make a confection
9. Repeat as small multiples

The `tile` package for R can help with all of the above:
sensible defaults and powerful options

Sorting in Tables and Table-like Graphs

ARMS RACE HANDOUT

Sorting in Tables and Table-like Graphs

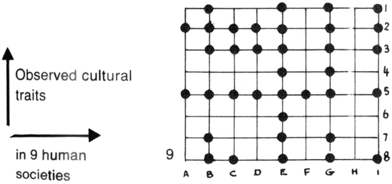
ARMS RACE HANDOUT

Just as we *scale* continuous dimensions in plots,
we *sort* categorical dimensions in tables or table-like graphs

ARMS RACE HANDOUT

Just as we *scale* continuous dimensions in plots,
we *sort* categorical dimensions in tables or table-like graphs

In *Semiologie Graphique* (1967), Jacques Bertin suggested sorting to **diagonalize**



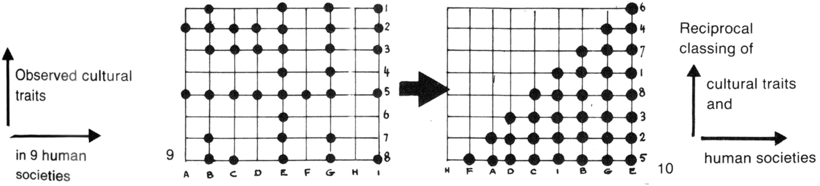
Sorting on the alphabet or a numerical ID is arbitrary, random, and pattern-hiding

Summarize the relationship in this figure

ARMS RACE HANDOUT

Just as we *scale* continuous dimensions in plots,
we *sort* categorical dimensions in tables or table-like graphs

In *Semiologie Graphique* (1967), Jacques Bertin suggested sorting to **diagonalize**



Sorting on the alphabet or a numerical ID is arbitrary, random, and pattern-hiding

Did you see the Likert-scaling of cultural traits?
Obvious when sorted to produce a "diagonal" arrangement

Sorting in Tables and Table-like Graphs

Recall the interest rate policy example

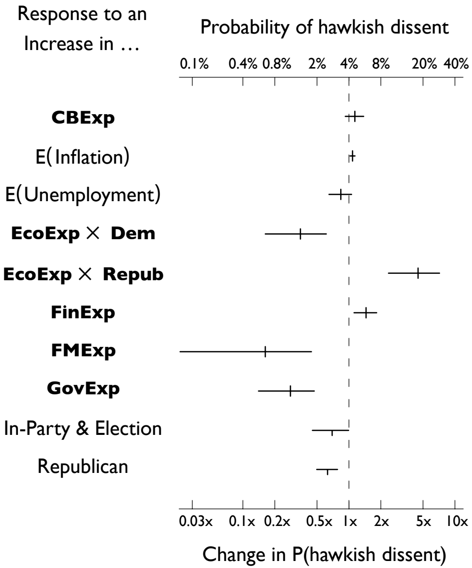
We presented correlates of conservative and liberal central banker preferences

Recall the interest rate policy example

We presented correlates of conservative and liberal central banker preferences

Imagine we had sorted the results alphabetically by covariate name

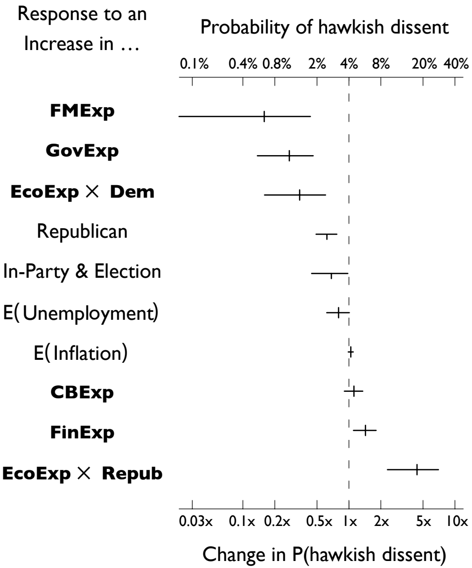
Unsorted dotplots always look the same: nothing jumps out



In the final graphic, I sorted by size of effect

Immediately tells a story, especially when I **bold** the career covariates

An arbitrary ordering in a figure is *always* a missed opportunity: seize it to highly covariation



John Gotti, a New York mob boss, was famously acquitted in a series of trials in the 1980s

Following a 1987 acquittal, the New York Times suggested the jury was swayed by a defense chart summarizing the crimes of prosecution witnesses

GOTTI IS ACQUITTED BY A FEDERAL JURY IN CONSPIRACY CASE

NEW CHARGES ARE LIKELY

Verdict is the First Setback in Recent Government Drive Against Mafia Leaders

By LEONARD BUDER

John Gotti was acquitted of Federal racketeering and conspiracy charges yesterday in the Government's first major setback in its recent assault on organized crime.

Mr. Gotti, who the Government says is the leader of the nation's most powerful Mafia family, and six co-defendants were found not guilty of charges they took part in a criminal enterprise. They were accused of carrying out illegal gambling and loan-sharking operations, armed hijackings and at least two murders over an 18-year period.

Despite yesterday's verdict, Federal investigators said the 46-year-old Mr. Gotti might face indictment on new charges as head of the Gambino crime family. "I can't comment but I won't deny it," said Thomas L. Sheer, head of the Federal Bureau of Investigation in New York, when asked if the F.B.I. was building up another case against Mr. Gotti.

'We'll Be Starting Again'

"They'll be ready to frame us again in two weeks," Mr. Gotti told a reporter before leaving the Brooklyn courthouse in a gray Cadillac that was waiting for him. "In three weeks we'll be starting again, just watch."

Until yesterday, Federal prosecutors in the Southern and Eastern Districts of New York had recorded a string of successes in major organized-crime cases.

Within the last six months, the heads of the city's four other Mafia families have been convicted after trials in Manhattan and Brooklyn. They, like Mr. Gotti and his co-defendants, had been charged under the Federal Racketeer Influenced and Corrupt Organizations Act, or RICO.

Key Witnesses Were Criminals

"Obviously they perceived there was something wrong with the evidence," said Andrew J. Maloney, the United States Attorney in Brooklyn, referring to the jury.

Many of the Government's key witnesses were criminals who testified for the prosecution under grants of immunity or in return for payments and other benefits.

The last piece of evidence requested by the jury for re-examination was a chart introduced by the defense that showed the criminal backgrounds of seven prosecution witnesses. It listed 69 crimes, including murder, drug possession and sales and kidnapping.

Mr. Gotti's lawyer, Bruce Cutler, said the jury showed "courage" because "it's not easy to say no to a Federal prosecutor." He said the jury had not been impressed with the testimony of "paid Government informants who lie, who use drugs, who kill people."

The verdict, which came on the seventh day of jury deliberations after a trial that lasted almost seven months, surprised many in the packed courtroom. Friends of the defendants cheered and applauded; the Government prosecutors, Diane F. Giacalone and John Gleeson, looked glum.

Mr. Gotti, who has been dubbed "Dapper Don" because of his expensive attire and impeccable grooming, and his co-defendants hugged and kissed each other and their lawyers.

Then they stood and applauded as the 12 members of the jury — whose identities had been kept secret to prevent possible tampering — left the room escorted by Federal marshals....

The New York Times

John Gotti

A Weakness In Gotti Case

Major U.S. Witnesses Viewed as Unreliable

By SELWYN RAAB

Many lawyers and prosecutors who followed events in the seven-month trial of John Gotti said the underlying weakness of the prosecution's case was its apparent reliance on turncoat career criminals as key witnesses against Mr. Gotti and six co-defendants.

News Analysis

A signal that the credibility of the prosecution's principal witnesses was in doubt came yesterday morning when the jury, in its final request before acquitting the defendants of all charges, reviewed an exhibit introduced by the defense.

It was a chart listing the lengthy criminal records of seven prosecution witnesses who had obtained promises of leniency and other favors from the Government in return for their testimony against Mr. Gotti....

Source: *NYT* reproduced in Tufte, *Envisioning Information*

A Well-Crafted Table Gets the Teflon Don Off

John Gotti, a New York mob boss, was famously acquitted in a series of trials in the 1980s

Following a 1987 acquittal, the New York Times suggested the jury was swayed by a defense chart summarizing the crimes of prosecution witnesses

Key Witnesses Were Criminals

"Obviously they perceived there was something wrong with the evidence," said Andrew J. Maloney, the United States Attorney in Brooklyn, referring to the jury.

Many of the Government's key witnesses were criminals who testified for the prosecution under grants of immunity or in return for payments and other benefits.

The last piece of evidence requested by the jury for re-examination was a chart introduced by the defense that showed the criminal backgrounds of seven prosecution witnesses. It listed 69 crimes, including murder, drug possession and sales and kidnapping.

Analysis

A signal that the credibility of the prosecution's principal witnesses was in doubt came yesterday morning when the jury, in its final request before acquitting the defendants of all charges, reviewed an exhibit introduced by the defense.

It was a chart listing the lengthy criminal records of seven prosecution witnesses who had obtained promises of leniency and other favors from the Government in return for their testimony against Mr. Gotti....

Source: NYT reproduced in Tufte, *Envisioning Information*

CRIMINAL ACTIVITY OF GOVERNMENT INFORMANTS

CRIME	CARDINALE	LOFARO	MALONEY	POLISI	SENATORE	FORONJY	CURRO
MURDER	X	X					
ATTEMPTED MURDER		X	X				
HEROIN POSSESSION AND SALE	X	X		X			X
COCAINE POSSESSION AND SALE	X		X	X			
MARIJUANA POSSESSION AND SALE							X
GAMBLING BUSINESS		X		X		X	
ARMED ROBBERIES	X		X	X	X		X
LOANSHARKING		X		X			
KIDNAPPING			X	X			
EXTORTION			X	X			
ASSAULT	X		X	X			X
POSSESSION OF DANGEROUS WEAPONS	X	X	X	X	X		X
PERJURY		X				X	
COUNTERFEITING					X	X	
BANK ROBBERY			X	X			
ARMED HIJACKING				X	X		
STOLEN FINANCIAL DOCUMENTS			X	X	X		
TAX EVASION				X		X	
BURGLARIES	X	X		X	X		
BRIBERY		X		X			
THEFT: AUTO, MONEY, OTHER			X	X	X	X	X
BAIL JUMPING AND ESCAPE			X	X			
INSURANCE FRAUDS					X	X	
FORGERIES				X	X		
PISTOL WHIPPING A PRIEST	X						
SEXUAL ASSAULT ON MINOR							X
RECKLESS ENDANGERMENT							X

Not a scientific chart – pure advocacy. Breaks the rules on purpose:

CRIMINAL ACTIVITY OF GOVERNMENT INFORMANTS

CRIME	CARDINALE	LOFARO	MALONEY	POLISI	SENATORE	FORONJY	CURRO
MURDER	X	X					
ATTEMPTED MURDER		X	X				
HEROIN POSSESSION AND SALE	X	X		X			X
COCAINE POSSESSION AND SALE	X		X	X			
MARIJUANA POSSESSION AND SALE							X
GAMBLING BUSINESS		X		X		X	
ARMED ROBBERIES	X		X	X	X		X
LOANSHARKING		X		X			
KIDNAPPING			X	X			
EXTORTION			X	X			
ASSAULT	X		X	X			X
POSSESSION OF DANGEROUS WEAPONS	X	X	X	X	X		X
PERJURY		X				X	
COUNTERFEITING					X	X	
BANK ROBBERY			X	X			
ARMED HIJACKING				X	X		
STOLEN FINANCIAL DOCUMENTS			X	X	X		
TAX EVASION				X		X	
BURGLARIES	X	X		X	X		
BRIBERY		X		X			
THEFT: AUTO, MONEY, OTHER			X	X	X	X	X
BAIL JUMPING AND ESCAPE			X	X			
INSURANCE FRAUDS					X	X	
FORGERIES				X	X		
PISTOL WHIPPING A PRIEST	X						
SEXUAL ASSAULT ON MINOR							X
RECKLESS ENDANGERMENT							X

Not a scientific chart – pure advocacy. Breaks the rules on purpose:

1. Unbundles categories to produce more ink

CRIMINAL ACTIVITY OF GOVERNMENT INFORMANTS

CRIME	CARDINALE	LOFARO	MALONEY	POLISI	SENATORE	FORONJY	CURRO
MURDER	X	X					
ATTEMPTED MURDER		X	X				
HEROIN POSSESSION AND SALE	X	X		X			X
COCAINE POSSESSION AND SALE	X		X	X			
MARIJUANA POSSESSION AND SALE							X
GAMBLING BUSINESS		X		X		X	
ARMED ROBBERIES	X		X	X	X		X
LOANSHARKING		X		X			
KIDNAPPING			X	X			
EXTORTION			X	X			
ASSAULT	X		X	X			X
POSSESSION OF DANGEROUS WEAPONS	X	X	X	X	X		X
PERJURY		X				X	
COUNTERFEITING					X	X	
BANK ROBBERY			X	X			
ARMED HIJACKING				X	X		
STOLEN FINANCIAL DOCUMENTS			X	X	X		
TAX EVASION				X		X	
BURGLARIES	X	X		X	X		
BRIBERY		X		X			
THEFT: AUTO, MONEY, OTHER			X	X	X	X	X
BAIL JUMPING AND ESCAPE			X	X			
INSURANCE FRAUDS					X	X	
FORGERIES				X	X		
PISTOL WHIPPING A PRIEST	X						
SEXUAL ASSAULT ON MINOR							X
RECKLESS ENDANGERMENT							X

Not a scientific chart – pure advocacy. Breaks the rules on purpose:

2. Unsorts the rows to spread the X's everywhere

CRIMINAL ACTIVITY OF GOVERNMENT INFORMANTS

CRIME	CARDINALE	LOFARO	MALONEY	POLISI	SENATORE	FORONJY	CURRO
MURDER	X	X					
ATTEMPTED MURDER		X	X				
HEROIN POSSESSION AND SALE	X	X		X			X
COCAINE POSSESSION AND SALE	X		X	X			
MARIJUANA POSSESSION AND SALE							X
GAMBLING BUSINESS		X		X		X	
ARMED ROBBERIES	X		X	X	X		X
LOANSHARKING		X		X			
KIDNAPPING			X	X			
EXTORTION			X	X			
ASSAULT	X		X	X			X
POSSESSION OF DANGEROUS WEAPONS	X	X	X	X	X		X
PERJURY		X				X	
COUNTERFEITING					X	X	
BANK ROBBERY			X	X			
ARMED HIJACKING				X	X		
STOLEN FINANCIAL DOCUMENTS			X	X	X		
TAX EVASION				X		X	
BURGLARIES	X	X		X	X		
BRIBERY		X		X			
THEFT: AUTO, MONEY, OTHER			X	X	X	X	X
BAIL JUMPING AND ESCAPE			X	X			
INSURANCE FRAUDS					X	X	
FORGERIES				X	X		
PISTOL WHIPPING A PRIEST	X						
SEXUAL ASSAULT ON MINOR							X
RECKLESS ENDANGERMENT							X

Not a scientific chart – pure advocacy. Breaks the rules on purpose:

3. Mixes legal terms with lurid descriptions to titilate & overwhelm

CRIMINAL ACTIVITY OF GOVERNMENT INFORMANTS

CRIME	CARDINALE	LOFARO	MALONEY	POLISI	SENATORE	FORONJY	CURRO
MURDER	X	X					
ATTEMPTED MURDER		X	X				
HEROIN POSSESSION AND SALE	X	X		X			X
COCAINE POSSESSION AND SALE	X		X	X			
MARIJUANA POSSESSION AND SALE							X
GAMBLING BUSINESS		X		X		X	
ARMED ROBBERIES	X		X	X	X		X
LOANSHARKING		X		X			
KIDNAPPING			X	X			
EXTORTION			X	X			
ASSAULT	X		X	X			X
POSSESSION OF DANGEROUS WEAPONS	X	X	X	X	X		X
PERJURY		X				X	
COUNTERFEITING					X	X	
BANK ROBBERY			X	X			
ARMED HIJACKING				X	X		
STOLEN FINANCIAL DOCUMENTS			X	X	X		
TAX EVASION				X		X	
BURGLARIES	X	X		X	X		
BRIBERY		X		X			
THEFT: AUTO, MONEY, OTHER			X	X	X	X	X
BAIL JUMPING AND ESCAPE			X	X			
INSURANCE FRAUDS					X	X	
FORGERIES				X	X		
PISTOL WHIPPING A PRIEST	X						
SEXUAL ASSAULT ON MINOR							X
RECKLESS ENDANGERMENT							X

Not a scientific chart – pure advocacy. Breaks the rules on purpose:

4. Centers the worst offender's column for maximum impact

Suppose you're the prosecutor. Can you turn these tricks around on Gotti?

Criminal Activity of Government Informants

	CARINALE	FORONJY	LOFARO	SENATORE	MALONEY	POLISI	CURRO	TYPE
Reckless Endangerment							3	Violent
Sexual Assult on a Minor							1	
Pistol Whipping a Priest	1							
Armed Hijacking				1		2		
Bank Robbery					1	3		
Possession of Dangerous Weapons	1		3	3	4	2	5	
Assault	2				6	1	5	
Extortion					3	4		
Kidnapping					1	2		
Armed Robberies	2			3	3	2	6	
Attempted Murder			2		2			
Murder	1		1					
Theft: Auto, Money, Other		5		4	6	2	8	Theft
Burglaries	2		3	1		2		
Marijuana Possession and Sale							5	Narcotics
Cocaine Posession and Sale	2				3	1		
Heroin Possession and Sale	1		3			1	5	
Bail Jumping and Escape		2			1	2		Judicial
Perjury		1	1					
Forgeries				3		2		Financial
Insurace Frauds		1		1				
Bribary			1			2		
Tax Evasion		1				1		
Stolen Financial Documents				1	1	2		
Counterfiting		1		2				
Loansharking			2			1		
Gambling Business		2	1			2		
Different Types of Crimes	8	7	9	9	9	17	8	
Total Number of Crimes	12	13	17	19	31	34	35	
Violations of Confidence	0	8	5	7	2	12	0	
Falsehood Index	0.00%	61.54%	29.41%	36.84%	6.45%	35.29%	0.00%	Average 24%

Violations of Confidence

Key to Cell Shades	1	2	3	4	5	6	≥ 7

Source: **Toby Braun,** `http://www.tbid.com/toybox/pg/tufte.html`

Suppose you're the prosecutor. Can you turn these tricks around on Gotti?

Criminal Activity of Government Informants

		CARINALE	FORONJY	LOFARO	SENATORE	MALONEY	POLISI	CURRO	TYPE
	Reckless Endangerment							3	Violent
	Sexual Assault on a Minor							1	
	Pistol Whipping a Priest	1							
	Armed Hijacking				1		2		
	Bank Robbery					1	3		
	Possession of Dangerous Weapons	1		3	3	4	2	5	
	Assault	2				6	1	5	
	Extortion					3	4		
	Kidnapping					1	2		
	Armed Robberies	2			3	3	2	6	
	Attempted Murder			2		2			
	Murder	1		1					
	Theft: Auto, Money, Other		5		4	6	2	8	Theft
	Burglaries	2		3	1		2		
	Marijuana Possession and Sale							5	Narcotics
	Cocaine Posession and Sale	2				3	1		
	Heroin Possession and Sale	1		3			1	5	
Violations of Confidence	Bail Jumping and Escape		2			1	2		Judicial
	Perjury		1	1					
	Forgeries				3		2		Financial
	Insurace Frauds		1		1				
	Bribary			1			2		
	Tax Evasion		1				1		
	Stolen Financial Documents				1	1	2		
	Counterfiting		1		2				
	Loansharking			2			1		
	Gambling Business		2	1			2		
	Different Types of Crimes	8	7	9	9	9	17	8	
	Total Number of Crimes	12	13	17	19	31	34	35	
	Violations of Confidence	0	8	5	7	2	12	0	
	Falsehood Index	0.00%	61.54%	29.41%	36.84%	6.45%	35.29%	0.00%	Average 24%

Key to Cell Shades						
1	2	3	4	5	6	≥ 7

Classify crimes as violent or dishonest, count them & classify witnesses into "types."
Would this work?

Criminal Activity of Government Informants

		CARINALE	FORONJY	LOFARO	SENATORE	MALONEY	POLISI	CURRO	TYPE
	Reckless Endangerment							3	Violent
	Sexual Assult on a Minor							1	
	Pistol Whipping a Priest	1							
	Armed Hijacking				1		2		
	Bank Robbery					1	3		
	Possession of Dangerous Weapons	1		3	3	4	2	5	
	Assault	2				6	1	5	
	Extortion					3	4		
	Kidnapping					1	2		
	Armed Robberies	2			3	3	2	6	
	Attempted Murder			2		2			
	Murder	1		1					
	Theft: Auto, Money, Other		5		4	6	2	8	Theft
	Burglaries	2		3	1		2		
	Marijuana Possession and Sale							5	Narcotics
	Cocaine Posession and Sale	2				3	1		
	Heroin Possession and Sale	1		3			1	5	
Violations of Confidence	Bail Jumping and Escape		2			1	2		Judicial
	Perjury		1	1					
	Forgeries				3		2		Financial
	Insurace Frauds		1		1				
	Bribary			1			2		
	Tax Evasion		1				1		
	Stolen Financial Documents				1	1	2		
	Counterfiting		1		2				
	Loansharking			2			1		
	Gambling Business		2	1			2		
	Different Types of Crimes	8	7	9	9	9	17	8	
	Total Number of Crimes	12	13	17	19	31	34	35	
	Violations of Confidence	0	8	5	7	2	12	0	
	Falsehood Index	0.00%	61.54%	29.41%	36.84%	6.45%	35.29%	0.00%	Average 24%
	Key to Cell Shades	1	2	3	4	5	6	≥ 7	

Classify crimes as violent or dishonest, count them & classify witnesses into "types."

Would this work? *No.*

Criminal Activity of Government Informants

		CARINALE	FORONJY	LOFARO	SENATORE	MALONEY	POLISI	CURRO	TYPE
	Reckless Endangerment							3	Violent
	Sexual Assault on a Minor							1	
	Pistol Whipping a Priest	1							
	Armed Hijacking				1		2		
	Bank Robbery					1	3		
	Possession of Dangerous Weapons	1		3	3	4	2	5	
	Assault	2				6	1	5	
	Extortion					3	4		
	Kidnapping					1	2		
	Armed Robberies	2			3	3	2	6	
	Attempted Murder			2		2			
	Murder	1		1					
	Theft: Auto, Money, Other		5		4	6	2	8	Theft
	Burglaries	2		3	1		2		
	Marijuana Possession and Sale							5	Narcotics
	Cocaine Posession and Sale	2				3	1		
	Heroin Possession and Sale	1		3			1	5	
Violations of Confidence	Bail Jumping and Escape		2			1	2		Judicial
	Perjury		1	1					
	Forgeries				3		2		Financial
	Insurace Frauds		1		1				
	Bribary			1			2		
	Tax Evasion		1				1		
	Stolen Financial Documents				1	1	2		
	Counterfiting		1		2				
	Loansharking			2			1		
	Gambling Business		2	1			2		
	Different Types of Crimes	8	7	9	9	9	17	8	
	Total Number of Crimes	12	13	17	19	31	34	35	
	Violations of Confidence	0	8	5	7	2	12	0	
	Falsehood Index	0.00%	61.54%	29.41%	36.84%	6.45%	35.29%	0.00%	Average 24%
	Key to Cell Shades	1	2	3	4	5	6	≥ 7	

Classify crimes as violent or dishonest, count them & classify witnesses into "types."

1. Index is meaningless: why divide by total crimes?

Criminal Activity of Government Informants

		CARINALE	FORONJY	LOFARO	SENATORE	MALONEY	POLISI	CURRO	TYPE
	Reckless Endangerment							3	Violent
	Sexual Assult on a Minor							1	
	Pistol Whipping a Priest	1							
	Armed Hijacking				1		2		
	Bank Robbery					1	3		
	Possession of Dangerous Weapons	1		3	3	4	2	5	
	Assault	2				6	1	5	
	Extortion					3	4		
	Kidnapping					1	2		
	Armed Robberies	2			3	3	2	6	
	Attempted Murder			2		2			
	Murder	1		1					
	Theft: Auto, Money, Other		5		4	6	2	8	Theft
	Burglaries	2		3	1		2		
	Marijuana Possession and Sale							5	Narcotics
	Cocaine Posession and Sale	2				3	1		
	Heroin Possession and Sale	1		3			1	5	
Violations of Confidence	Bail Jumping and Escape		2			1	2		Judicial
	Perjury		1	1					
	Forgeries				3		2		Financial
	Insurace Frauds		1		1				
	Bribary			1			2		
	Tax Evasion		1				1		
	Stolen Financial Documents				1	1	2		
	Counterfiting		1		2				
	Loansharking			2			1		
	Gambling Business		2	1			2		
	Different Types of Crimes	8	7	9	9	9	17	8	
	Total Number of Crimes	12	13	17	19	31	34	35	
	Violations of Confidence	0	8	5	7	2	12	0	
	Falsehood Index	0.00%	61.54%	29.41%	36.84%	6.45%	35.29%	0.00%	Average 24%
	Key to Cell Shades	1	2	3	4	5	6	≥ 7	

Classify crimes as violent or dishonest, count them & classify witnesses into "types."

2. Average of index is doubly meaningless

Criminal Activity of Government Informants

		CARINALE	FORONJY	LOFARO	SENATORE	MALONEY	POLISI	CURRO	TYPE
	Reckless Endangerment							3	Violent
	Sexual Assult on a Minor							1	
	Pistol Whipping a Priest	1							
	Armed Hijacking				1		2		
	Bank Robbery					1	3		
	Possession of Dangerous Weapons	1		3	3	4	2	5	
	Assault	2				6	1	5	
	Extortion					3	4		
	Kidnapping					1	2		
	Armed Robberies	2			3	3	2	6	
	Attempted Murder			2		2			
	Murder	1		1					
	Theft: Auto, Money, Other		5		4	6	2	8	Theft
	Burglaries	2		3	1		2		
	Marijuana Possession and Sale							5	Narcotics
	Cocaine Posession and Sale	2				3	1		
	Heroin Possession and Sale	1		3			1	5	
Violations of Confidence	Bail Jumping and Escape		2			1	2		Judicial
	Perjury		1	1					
	Forgeries				3		2		Financial
	Insurace Frauds		1		1				
	Bribary			1			2		
	Tax Evasion		1				1		
	Stolen Financial Documents				1	1	2		
	Counterfiting		1		2				
	Loansharking			2			1		
	Gambling Business		2	1			2		
	Different Types of Crimes	8	7	9	9	9	17	8	
	Total Number of Crimes	12	13	17	19	31	34	35	
	Violations of Confidence	0	8	5	7	2	12	0	
	Falsehood Index	0.00%	61.54%	29.41%	36.84%	6.45%	35.29%	0.00%	Average 24%
	Key to Cell Shades	1	2	3	4	5	6	≥ 7	

Classify crimes as violent or dishonest, count them & classify witnesses into "types."

3. Cell shading suggests a few crimes ≈ no crimes

Criminal Activity of Government Informants

		CARINALE	FORONJY	LOFARO	SENATORE	MALONEY	POLISI	CURRO	TYPE
	Reckless Endangerment							3	Violent
	Sexual Assult on a Minor							1	
	Pistol Whipping a Priest	1							
	Armed Hijacking				1		2		
	Bank Robbery					1	3		
	Possession of Dangerous Weapons	1		3	3	4	2	5	
	Assault	2				6	1	5	
	Extortion					3	4		
	Kidnapping					1	2		
	Armed Robberies	2			3	3	2	6	
	Attempted Murder			2		2			
	Murder	1		1					
	Theft: Auto, Money, Other		5		4	6	2	8	Theft
	Burglaries	2		3	1		2		
	Marijuana Possession and Sale							5	Narcotics
	Cocaine Posession and Sale	2				3	1		
	Heroin Possession and Sale	1		3			1	5	
Violations of Confidence	Bail Jumping and Escape		2			1	2		Judicial
	Perjury		1	1					
	Forgeries				3		2		Financial
	Insurace Frauds		1		1				
	Bribary			1			2		
	Tax Evasion		1				1		
	Stolen Financial Documents				1	1	2		
	Counterfiting		1		2				
	Loansharking			2			1		
	Gambling Business		2	1			2		
	Different Types of Crimes	8	7	9	9	9	17	8	
	Total Number of Crimes	12	13	17	19	31	34	35	
	Violations of Confidence	0	8	5	7	2	12	0	
	Falsehood Index	0.00%	61.54%	29.41%	36.84%	6.45%	35.29%	0.00%	Average 24%
	Key to Cell Shades	1	2	3	4	5	6	≥ 7	

Classify crimes as violent or dishonest, count them & classify witnesses into "types."

4. Unbundled categories now makes cells fainter

Criminal Activity of Government Informants

	CARINALE	FORONJY	LOFARO	SENATORE	MALONEY	POLISI	CURRO	TYPE
Reckless Endangerment							3	Violent
Sexual Assult on a Minor							1	
Pistol Whipping a Priest	1							
Armed Hijacking				1		2		
Bank Robbery					1	3		
Possession of Dangerous Weapons	1		3	3	4	2	5	
Assault	2				6	1	5	
Extortion					3	4		
Kidnapping					1	2		
Armed Robberies	2			3	3	2	6	
Attempted Murder			2		2			
Murder	1		1					
Theft: Auto, Money, Other		5		4	6	2	8	Theft
Burglaries	2		3	1		2		
Marijuana Possession and Sale							5	Narcotics
Cocaine Posession and Sale	2				3	1		
Heroin Possession and Sale	1		3			1	5	
Bail Jumping and Escape		2			1	2		Judicial
Perjury		1	1					
Forgeries				3		2		Financial
Insurace Frauds		1		1				
Bribary			1			2		
Tax Evasion		1				1		
Stolen Financial Documents				1	1	2		
Counterfiting		1		2				
Loansharking			2			1		
Gambling Business		2	1			2		
Different Types of Crimes	8	7	9	9	9	17	8	
Total Number of Crimes	12	13	17	19	31	34	35	
Violations of Confidence	0	8	5	7	2	12	0	
Falsehood Index	0.00%	61.54%	29.41%	36.84%	6.45%	35.29%	0.00%	Average 24%

Violations of Confidence

Key to Cell Shades	1	2	3	4	5	6	≥ 7

Classify crimes as violent or dishonest, count them & classify witnesses into "types."

"Believe the honest drug dealer, and the guy who pistol whipped a priest"

What if we wanted to approach the question of witness similarity as scientists?

1. Which crimes tend to cluster together?

2. Which criminals are similar to each other?

What if we wanted to approach the question of witness similarity as scientists?

1. Which crimes tend to cluster together?

2. Which criminals are similar to each other?

Let's order the rows and columns by *similarity* using cluster analysis

A diagram sorted in this fashion is called a **heatmap**

What if we wanted to approach the question of witness similarity as scientists?

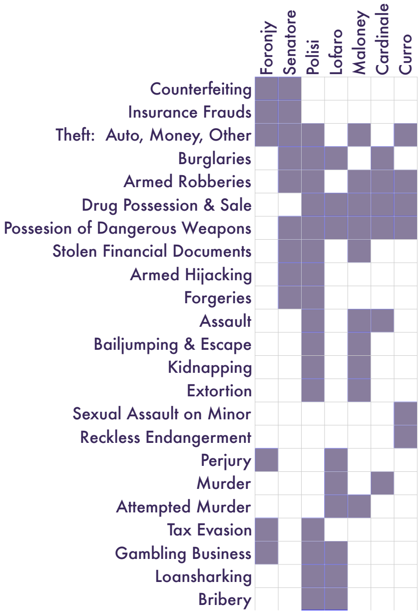
1. Which crimes tend to cluster together?

2. Which criminals are similar to each other?

Let's order the rows and columns by *similarity* using cluster analysis

A diagram sorted in this fashion is called a **heatmap**

...it also helps to combine redundant categories



My Advice for Scientific Graphics

Substantive Focus

Initial Minimalism

Develop a Style

Write Results Around Figures

Follow Through