Visualisation and data analysis for journalism studies

Rafal Urbaniak and Nikodem Lewandowski
(University of Gdansk)
https://rfl-urbaniak.github.io/teaching/
rfl.urbaniak+teaching@gmail.com

The plan

Motivations, goals, game rules
Some history
The role of perception
Getting started with R, RStudio and ggplot2
More on what to show
Focus
Epistemic problems
Technical and mathematical problems
Statistical learning and probabilistic thinking
Statistical and analytical blunders
Basics of Bayesian thinking
Linear models
Causality and variable selection

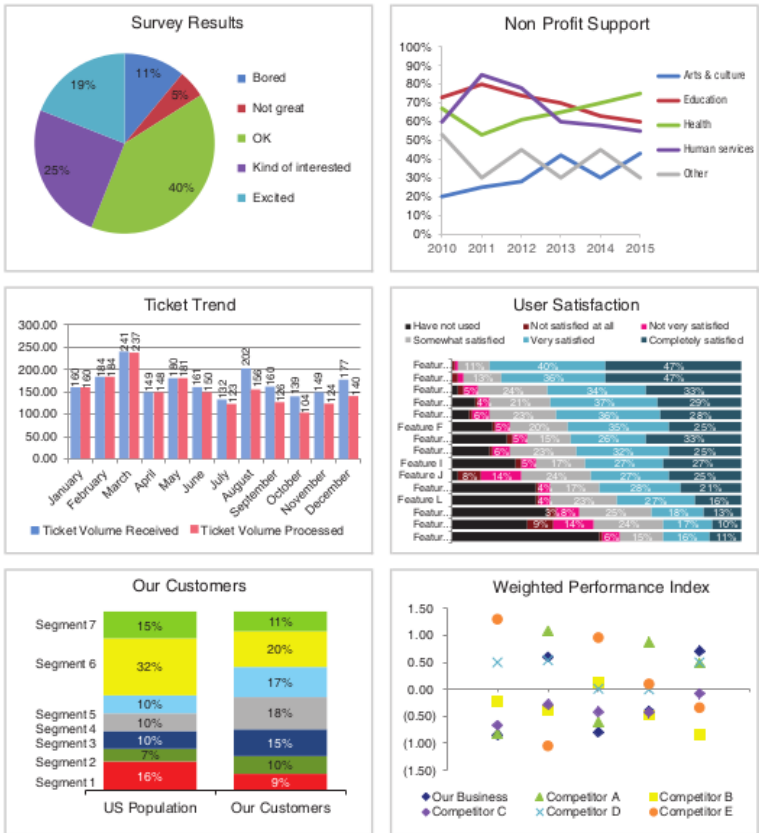
Motivations

- It's too easy to generate tables and visualisation.
- This makes communication harder!

Motivations

- It's too easy to generate tables and visualisation.
- This makes communication harder!

Bad graphs everywhere!



Cole Nussbaum [1]

Lack of background

- We learn some math at school.
- We learn some arts at school.

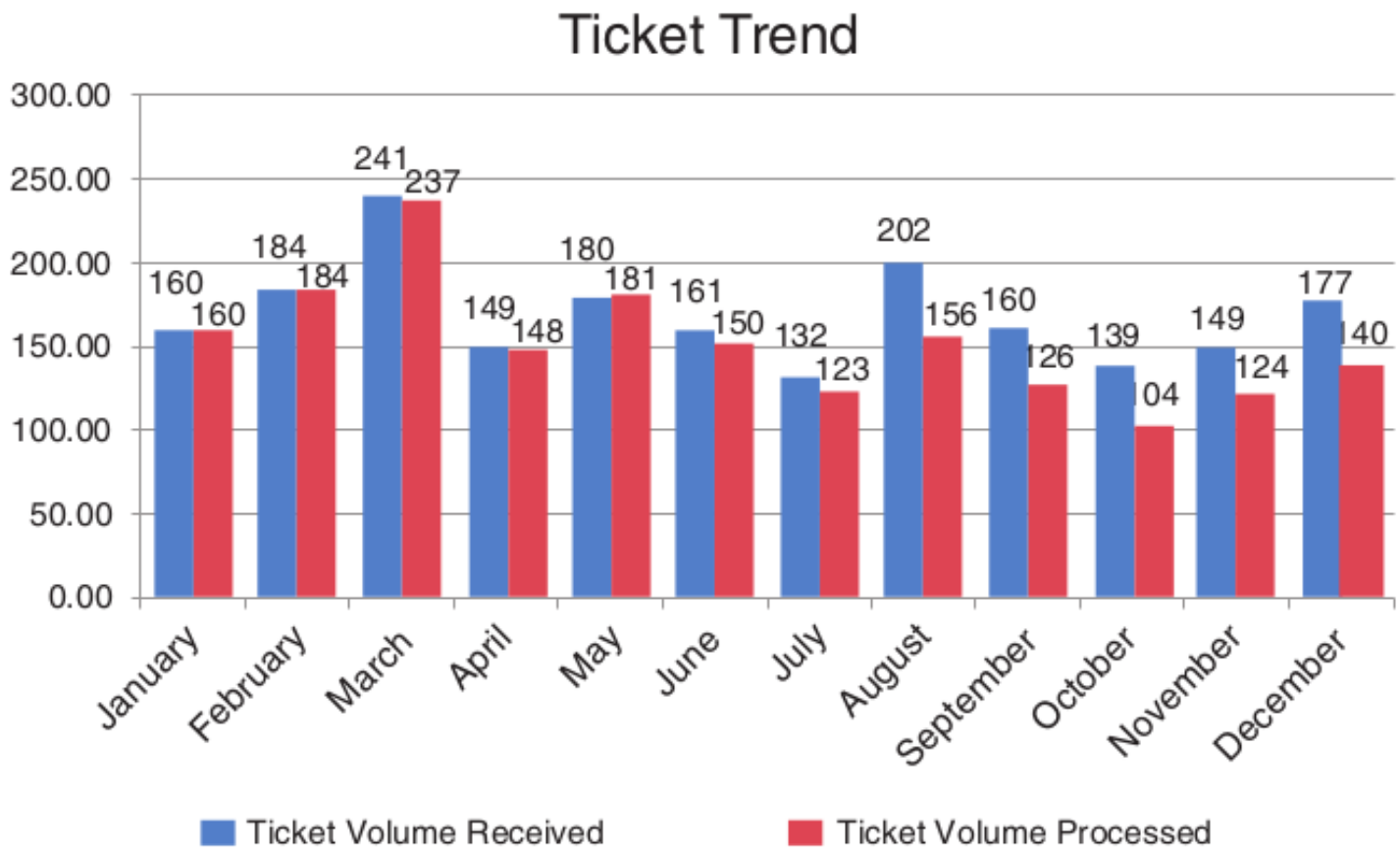
Lack of background

- We learn some math at school.
- We learn some arts at school.

Problem

We never learn to put them together, and think they're opposite.

Some examples



Cole Nussbaum [4]

Some examples

Please approve the hire of 2 FTEs

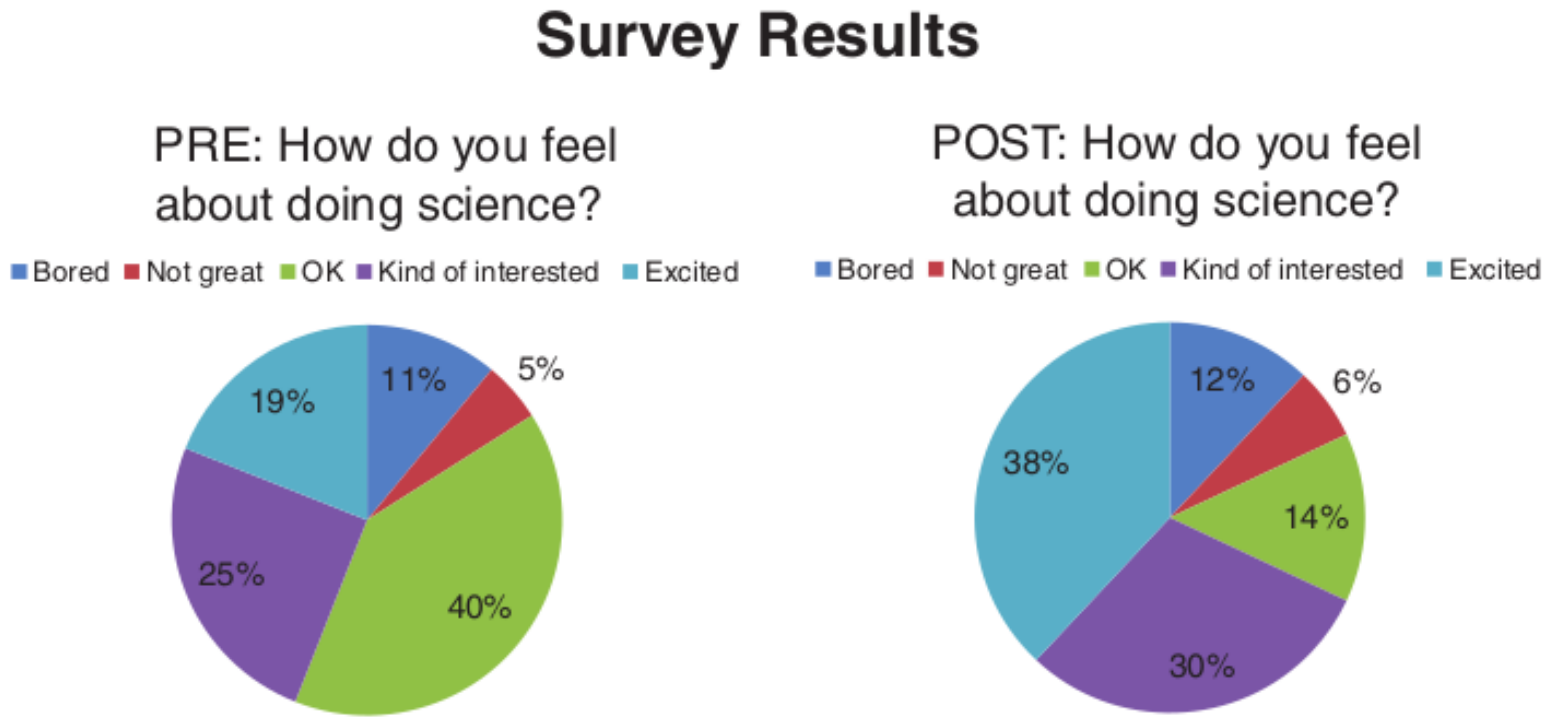
to backfill those who quit in the past year

Ticket volume over time

Data source: XYZ Dashboard, as of 12/31/2014 | A detailed analysis on tickets processed per person and time to resolve issues was undertaken to inform this request and can be provided if needed.

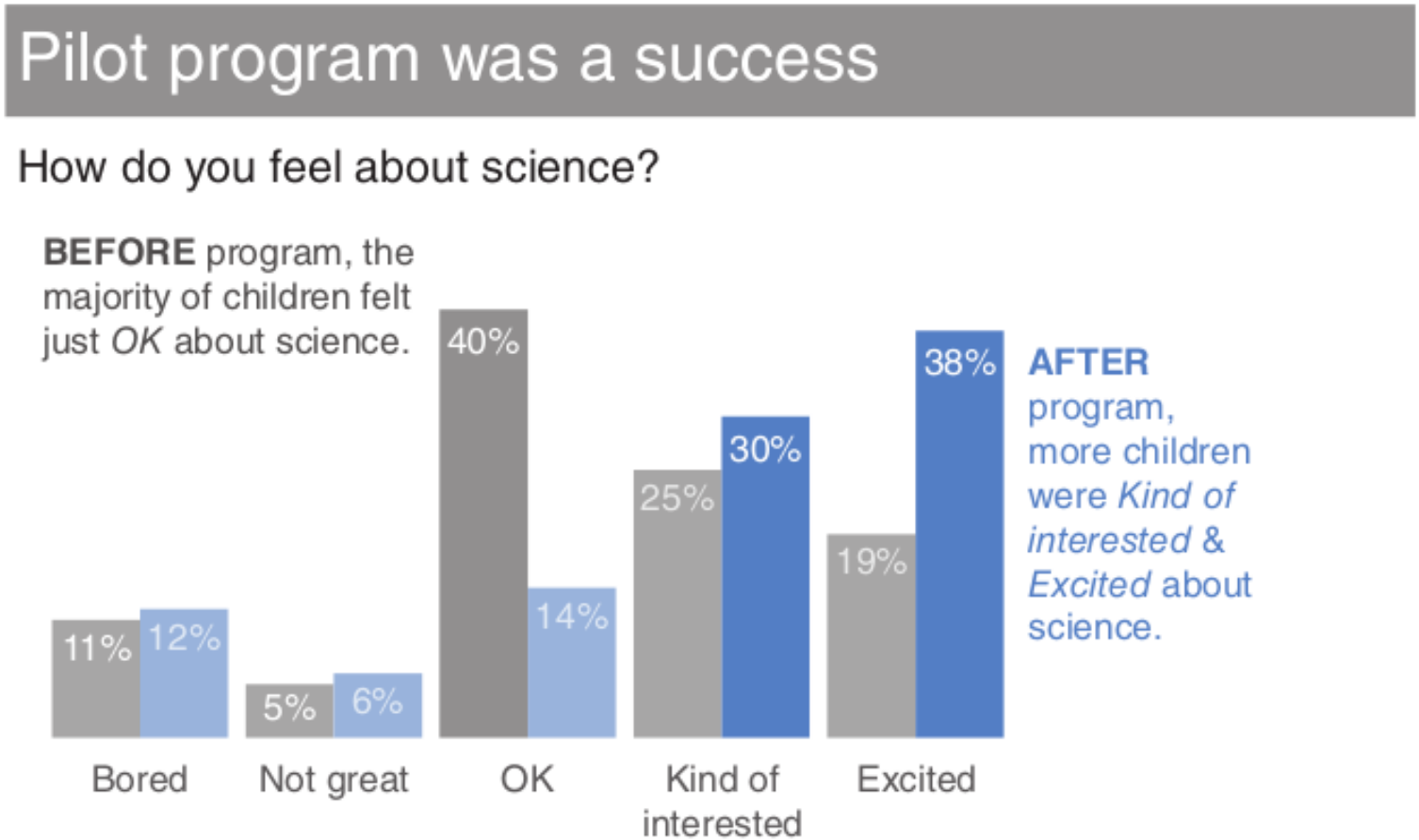
Cole Nussbaum [4]

Some examples



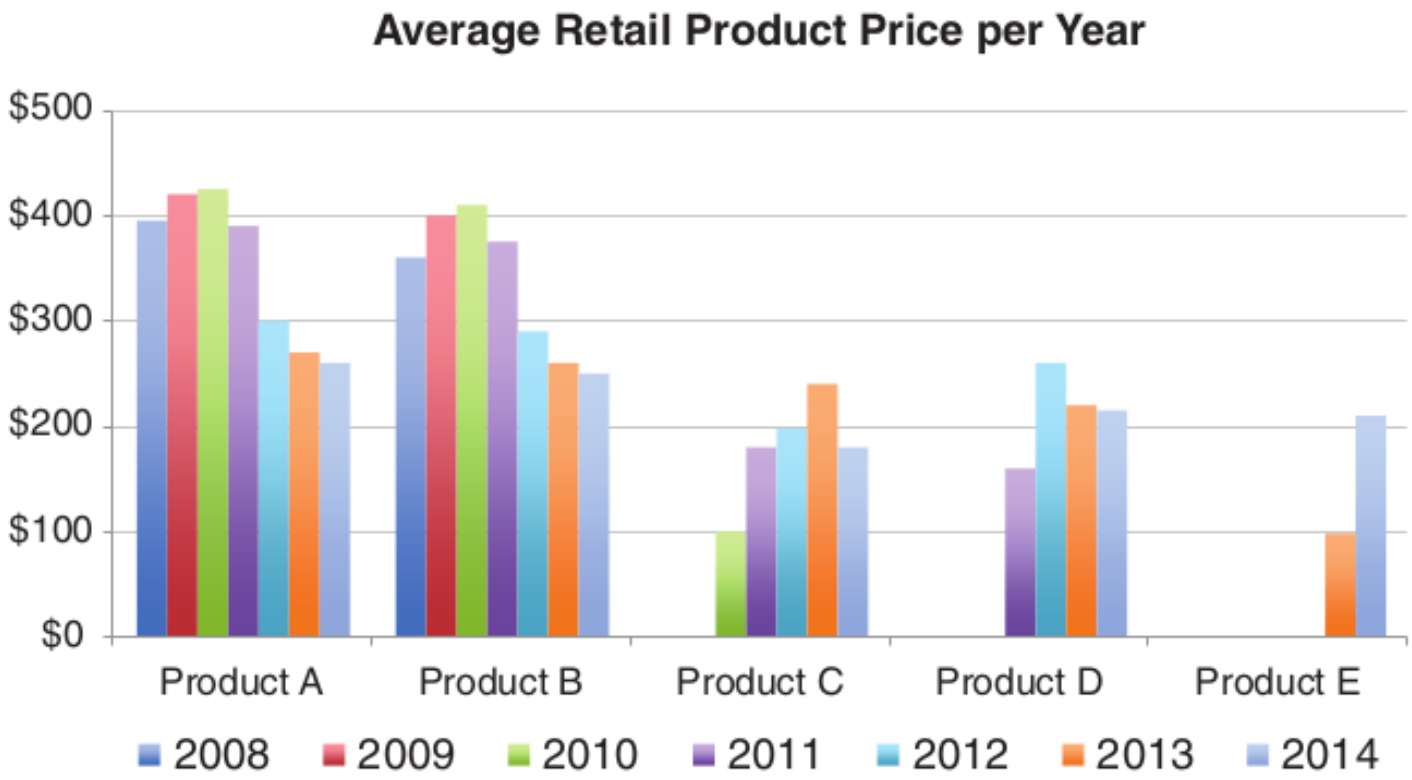
Cole Nussbaum [5]

Some examples



Cole Nussbaum [5]

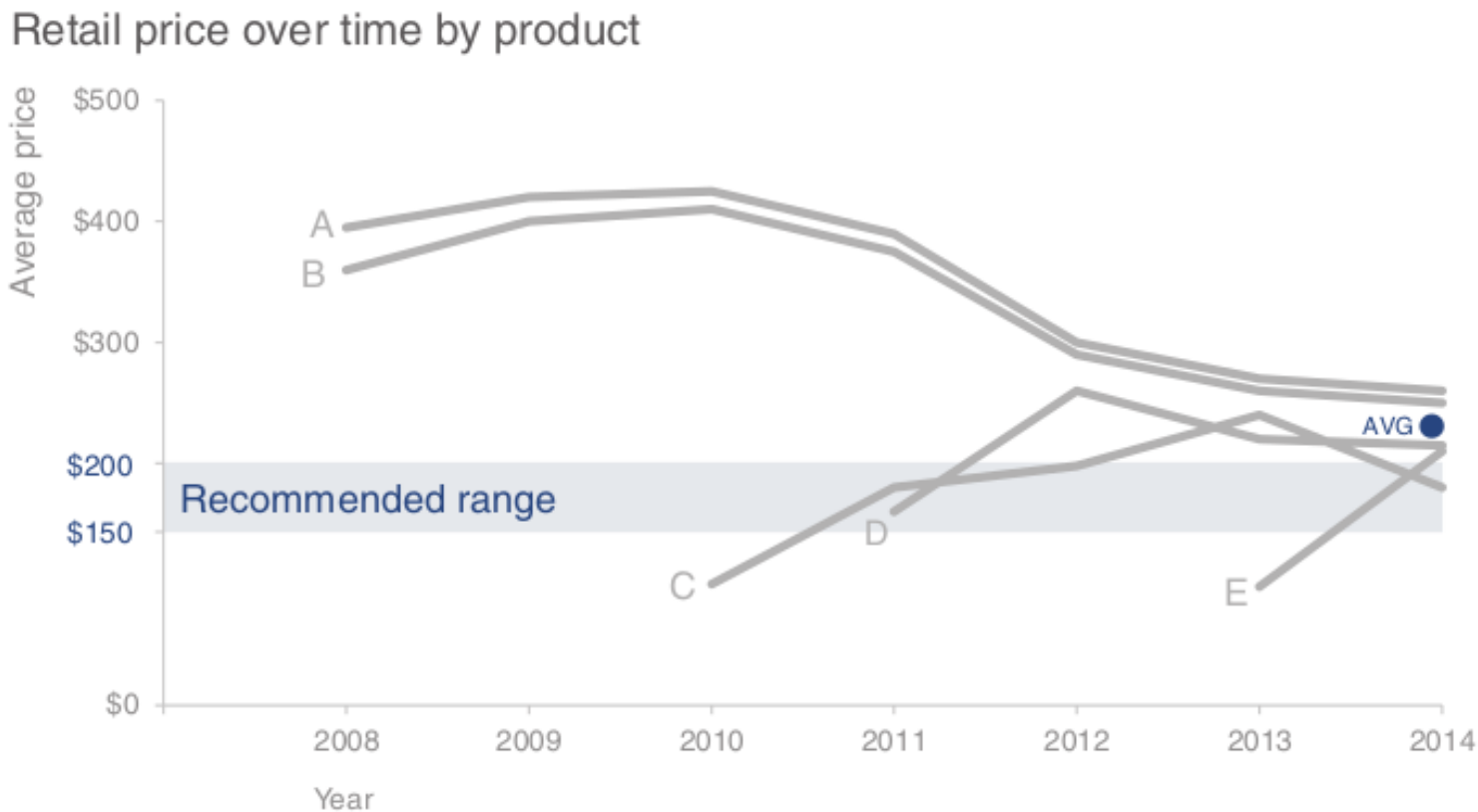
Some examples



Cole Nussbaum [6]

Some examples

To be competitive, we recommend introducing our product *below the $223 average* price point in the **$150–$200 range**



Cole Nussbaum [6]

Goal

- To understand psychological factors that guide various visualization choices
- To be able to properly analyze data yourself (at a decent level, or at least to understand some of the complexities involved)
- To be able to visualize your data insights so that they clearly convey your message
- To be able to work in R, a statistical programming language

Rules: final grade

Final test: 60 points (optional)

- multiple choice with penalty points

Project: 60 points (optional)

- two-three pages of meaningful text with at least two visualizations, bonus points for animations
- everything prepared in R markdown
- feedback loop: idea -> draft -> feedback -> revisions -> f2 -> r2

Tutorial performance: 60 points (optional)

- If you complete a free-fall exercise without much help, show us, get some points!

Final grade

As if out of 100.

Contact

Updates - only here!

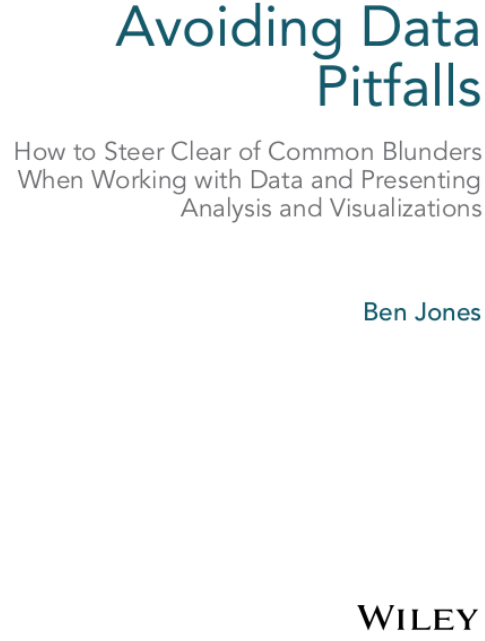
https://rfl-urbaniak.github.io/teaching/

Contact - only here!

rfl.urbaniak+teaching@gmail.com

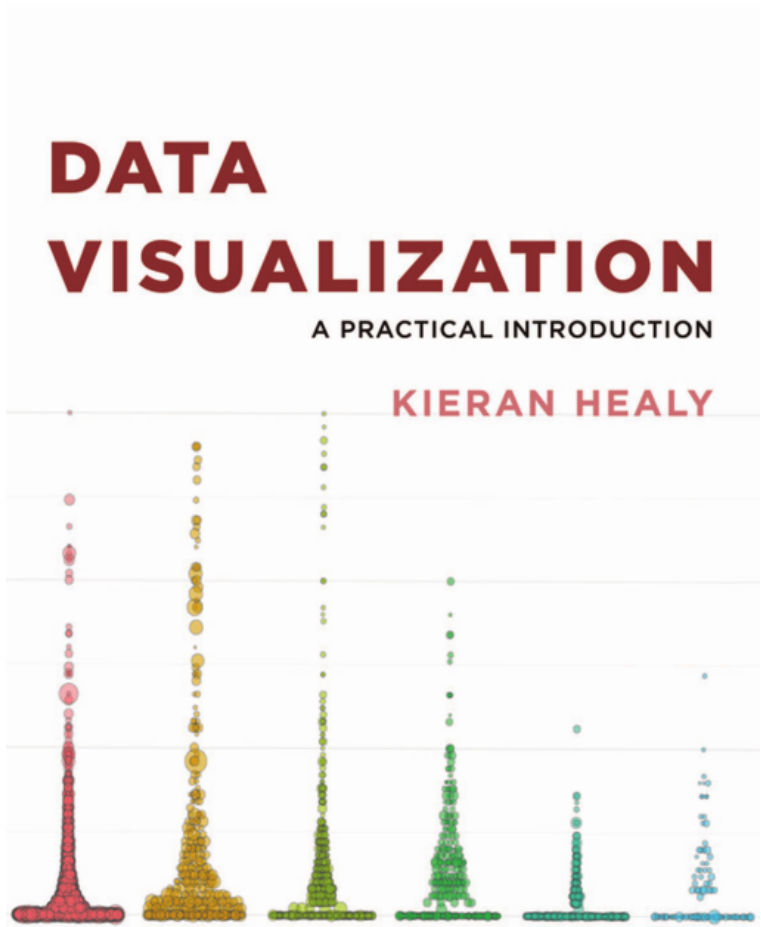
Sources


NERD
journalism
How Data and Digital Technology Transformed News Graphics
Doctoral thesis submitted by
Alberto Cairo Touriño
to the Doctoral Programme in the Information and Knowledge Society
Universitat Oberta de Catalunya
Thesis supervisor: Dr Jordi Sánchez Navarro
Thesis Supervisory Committee:
Dr Jordi Sánchez Navarro
Dr Lluis Pastor Pérez
Dr Ramón Salaverría

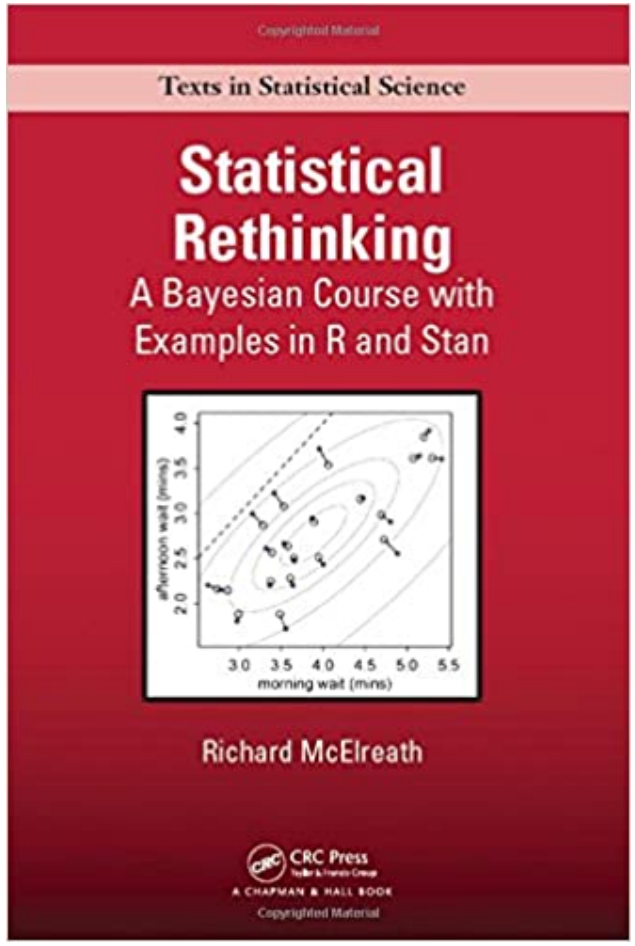

Avoiding Data Pitfalls
How to Steer Clear of Common Blunders When Working with Data and Presenting Analysis and Visualizations
Ben Jones
WILEY

Sources


cole nussbaumer knaflic
storytelling
with
data
a data
visualization
guide for
business
professionals
WILEY


DATA
VISUALIZATION
A PRACTICAL INTRODUCTION
KIERAN HEALY

Sources


Copyrighted Material
Texts in Statistical Science
Statistical Rethinking
A Bayesian Course with Examples in R and Stan
afternoon wait (mins)
morning wait (mins)
Richard McElreath
CRC Press
A CHAPMAN & HALL BOOK
Copyrighted Material

Precursors

Spanish *Relación*, 1638

Precursors

B Numb. 116.

The Daily Courant.

Saturday, September 12. 1702.

LONDON, Sept. 12.

THE Defeat of the Duke of Ormond in the Bay of Cadiz being the most considerable Enterprize that the English have undertaken abroad for this last Hundred Years; We cannot doubt but the Publick will be better pleas'd with a Description of that Island and City, which may be of some use for the clearer understanding of the Advice that is come already and that is further expected from thence; than with any thing that we could draw out of the last Foreign Prints and News-Letters, even tho' we had not already given what is most material in them.

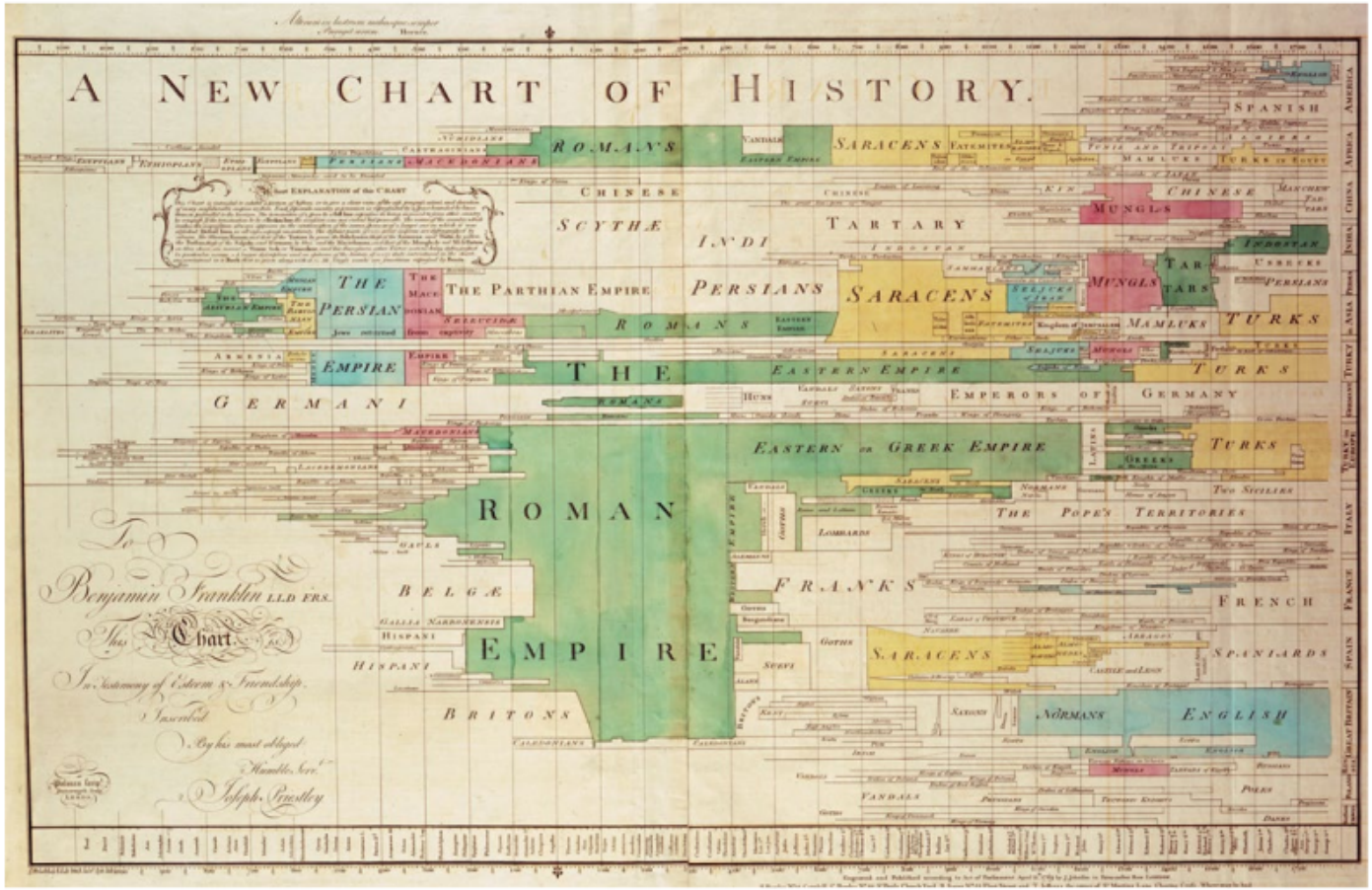
The Island of CADIZ lies between the Mouth of the River Guadalquivir, and the Streights of Gibraltar, on the Coast of Andalusia: It is joyn'd to the firm Land by a Bridge of 700 Paces long, call'd la Puente de Suaco: The Distance between Port St. Mary and that Bridge is about 12 Miles; from that Bridge to the City of Cadiz is likewise about 12 Miles: The Rocks call'd the Diamond and Las Puerco make the Entrance into the Bay pretty Dangerous: The Bay is about 12 Leagues in Circuit and 2 Broad; but the narrowest part (one Side of which is defended by a Fort call'd the Puntal, and the other by the Fort of St. Cruz,) is not above a Mile and an half over: That Part of the Island on which the Town Stands is defended towards the Sea by the Fort of St. Philip, the Point of St. Sebastian, the Walls of the Town (which towards the Port are wash'd by the Tide), and by Rocky Rocks: And there are very Strong Fortifications to secure the Passage to the Town by the narrow Neck of Land that runs from the larger Part of the Island to the City: The Fort of St. Catherine which is taken by the Duke Stands between Port St. Mary and the Bay. There are a great many Churches in Cadiz, which is well built, very rich, and full of Inhabitants.

A A. The Bay of Cadiz.
B B. The Ocean.
C. La Rotta.
D. Port St. Mary.
E. Port Real.
F. La Puente de Suaco.
G. Puntal.
H H. The Fortifications on the Neck of Land before the City.
I. The City.
K. Fort St. Philip.
L. Las Puerco.
M. Caleta.
N. Point of St. Sebastian.
O. St. Catharine.
P. The Island of St. Peter.
Q. A Scale of two Miles.

English *The Daily Courant* (invasion on Cadiz), 1702

Precursors

Joseph Priestley (1733-1804)



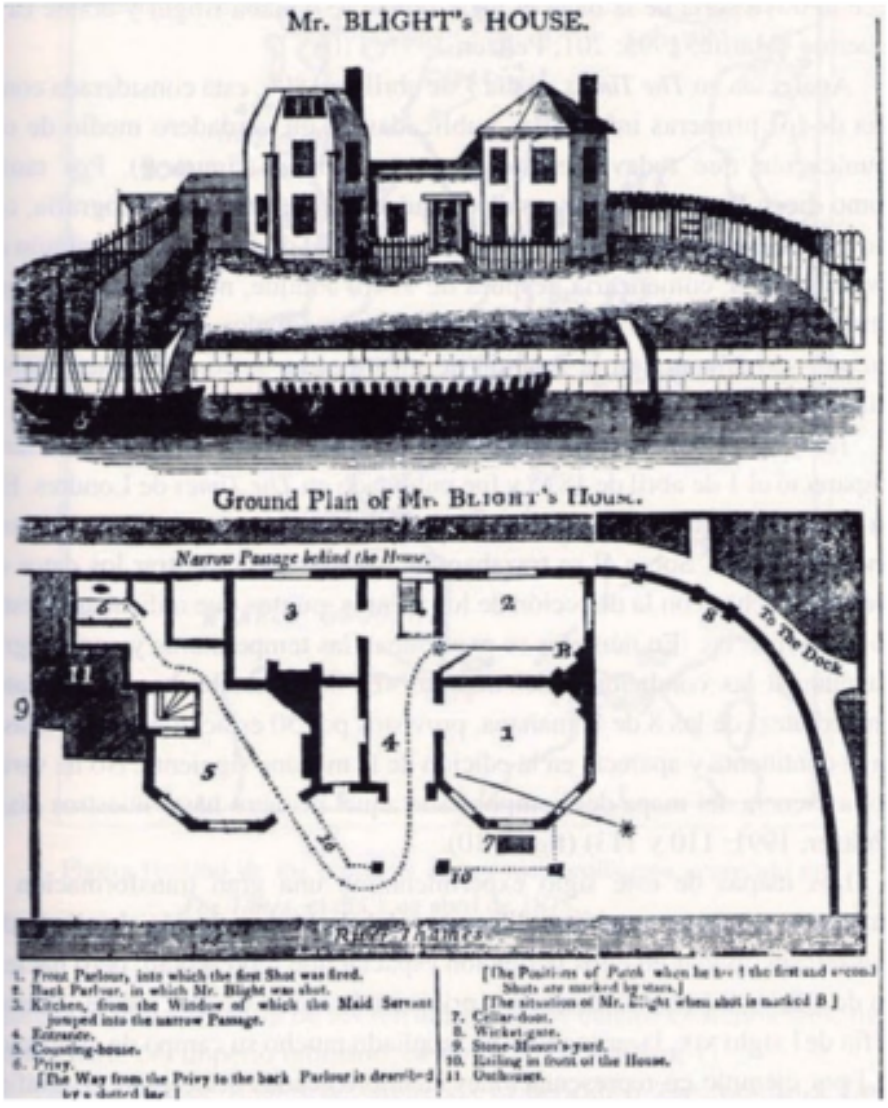
A new chart of history, 1769

Precursors

William Playfair (1759-1823)

Statistical breviary, 1801

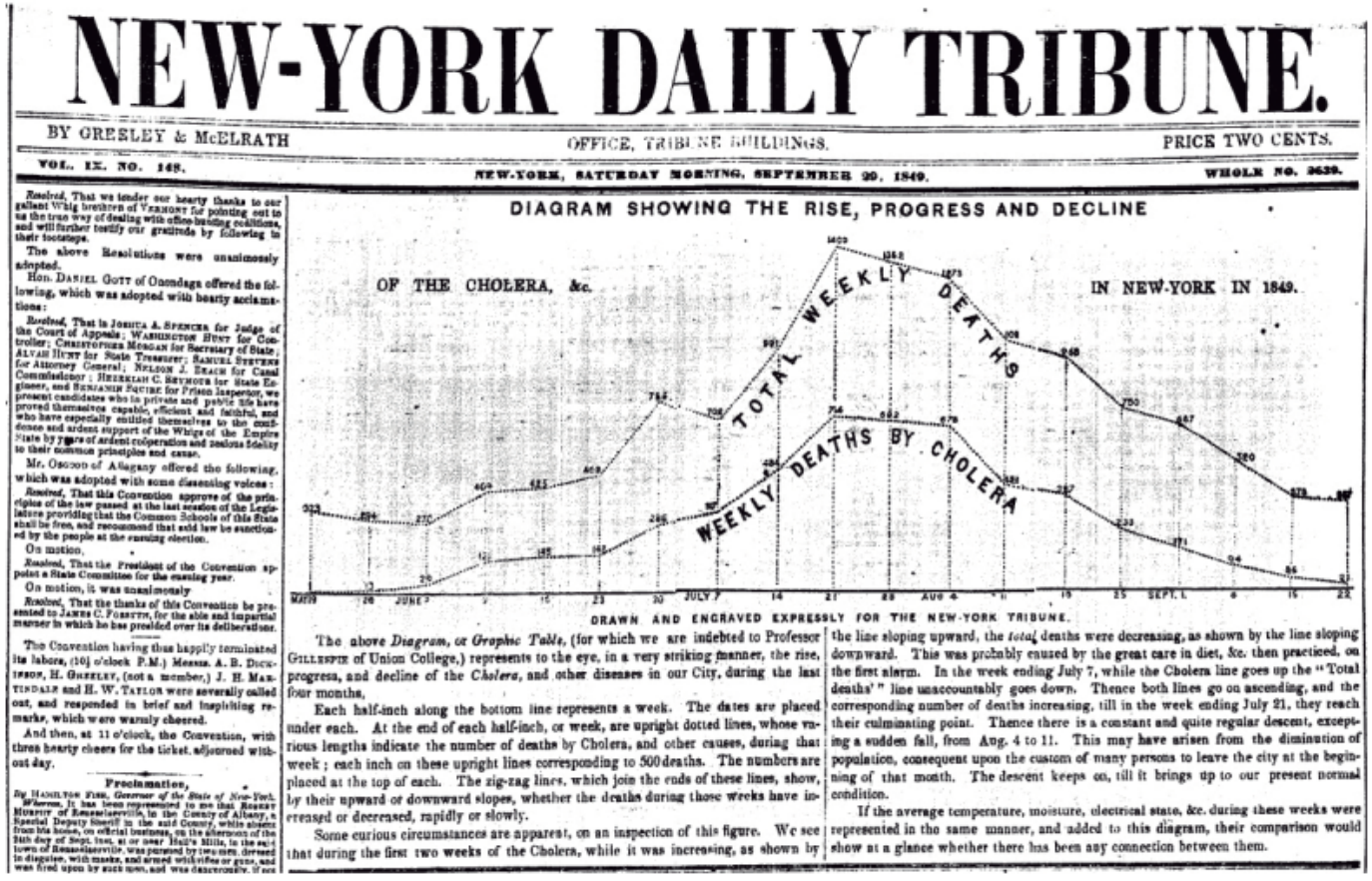
Precursors



A murder case coverage in *The Times*, 1806

Precursors

William Mitchell Gillespie (1816-1868)



NEW-YORK DAILY TRIBUNE.

BY GREELEY & McELRATH — OFFICE, TRIBUNE BUILDINGS. — PRICE TWO CENTS.

VOL. IX. NO. 148. — NEW-YORK, SATURDAY MORNING, SEPTEMBER 29, 1849. — WHOLE NO. 2639.

Resolved, That we tender our hearty thanks to our gallant Whig brethren of VERMONT for pointing out to us the true way of dealing with office-hunting coalitions, and will further testify our gratitude by following in their footsteps.

The above Resolutions were unanimously adopted.

Hon. DANIEL GOTT of Onondaga offered the following, which was adopted with hearty acclamations:

Resolved, That in JOSHUA A. SPENCER for Judge of the Court of Appeals; WASHINGTON HUNT for Controller; CHRISTOPHER MORGAN for Secretary of State; ALVAH HUNT for State Treasurer; SAMUEL STEVENS for Attorney General; NELSON J. BEACH for Canal Commissioner; HEZEKIAH C. SEYMOUR for State Engineer, and BENJAMIN FAUBE for Prison Inspector, we present candidates who in private and public life have proved themselves capable, efficient and faithful, and who have especially entitled themselves to the confidence and ardent support of the Whigs of the Empire State by years of ardent cooperation and zealous fidelity to their common principles and cause.

Mr. OSGOOD of Allegany offered the following, which was adopted with some dissenting voices:

Resolved, That this Convention approve of the principles of the law passed at the last session of the Legislature providing that the Common Schools of this State shall be free, and recommend that said law be sanctioned by the people at the ensuing election.

On motion,

Resolved, That the President of the Convention appoint a State Committee for the ensuing year.

On motion, it was unanimously

Resolved, That the thanks of this Convention be presented to JAMES C. FUSSETT, for the able and impartial manner in which he has presided over its deliberations.

The Convention having thus happily terminated its labors, (10½ o'clock P.M.) Messrs. A. B. DICKINSON, H. GREELEY, (not a member,) J. H. MARTINDALE and H. W. TAYLOR were severally called out, and responded in brief and inspiriting remarks, which were warmly cheered.

And then, at 11 o'clock, the Convention, with three hearty cheers for the ticket, adjourned without day.

Proclamation,

By HAMILTON FISH, *Governor of the State of New-York.*

Whereas, It has been represented to me that ROBERT MURPHY of Rensselaerville, in the County of Albany, a Special Deputy Sheriff in the said County, while absent from his house, on official business, on the afternoon of the 24th day of Sept. inst. at or near Hull's Mills, in the said town of Rensselaerville, was pursued by two men dressed in disguise, with masks, and armed with rifles or guns, and was fired upon by such men, and was dangerously, if not

DIAGRAM SHOWING THE RISE, PROGRESS AND DECLINE OF THE CHOLERA, &c. IN NEW-YORK IN 1849.

DRAWN AND ENGRAVED EXPRESSLY FOR THE NEW-YORK TRIBUNE.

The above *Diagram, or Graphic Table*, (for which we are indebted to Professor GILLESPIE of Union College,) represents to the eye, in a very striking manner, the rise, progress, and decline of the *Cholera*, and other diseases in our City, during the last four months.

Each half-inch along the bottom line represents a week. The dates are placed under each. At the end of each half-inch, or week, are upright dotted lines, whose various lengths indicate the number of deaths by Cholera, and other causes, during that week; each inch on these upright lines corresponding to 500 deaths. The numbers are placed at the top of each. The zig-zag lines, which join the ends of these lines, show, by their upward or downward slopes, whether the deaths during those weeks have increased or decreased, rapidly or slowly.

Some curious circumstances are apparent, on an inspection of this figure. We see that during the first two weeks of the Cholera, while it was increasing, as shown by the line sloping upward, the *total* deaths were decreasing, as shown by the line sloping downward. This was probably caused by the great care in diet, &c. then practiced, on the first alarm. In the week ending July 7, while the Cholera line goes up the "Total deaths'" line unaccountably goes down. Thence both lines go on ascending, and the corresponding number of deaths increasing, till in the week ending July 21, they reach their culminating point. Thence there is a constant and quite regular descent, excepting a sudden fall, from Aug. 4 to 11. This may have arisen from the diminution of population, consequent upon the custom of many persons to leave the city at the beginning of that month. The descent keeps on, till it brings up to our present normal condition.

If the average temperature, moisture, electrical state, &c. during these weeks were represented in the same manner, and added to this diagram, their comparison would show at a glance whether there has been any connection between them.

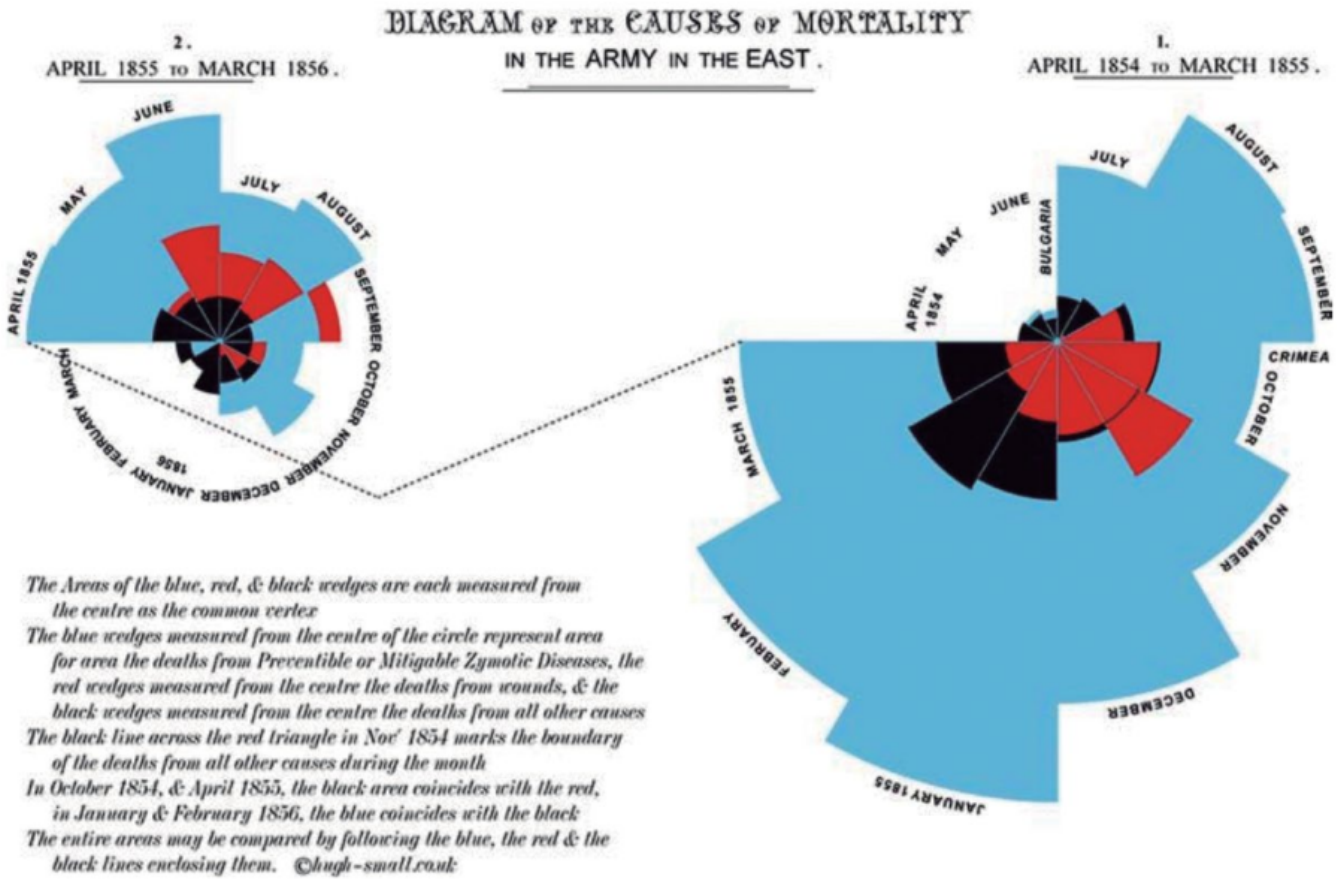
New-York Daily Tribune, 1849

XIXth century explosion

Reasons

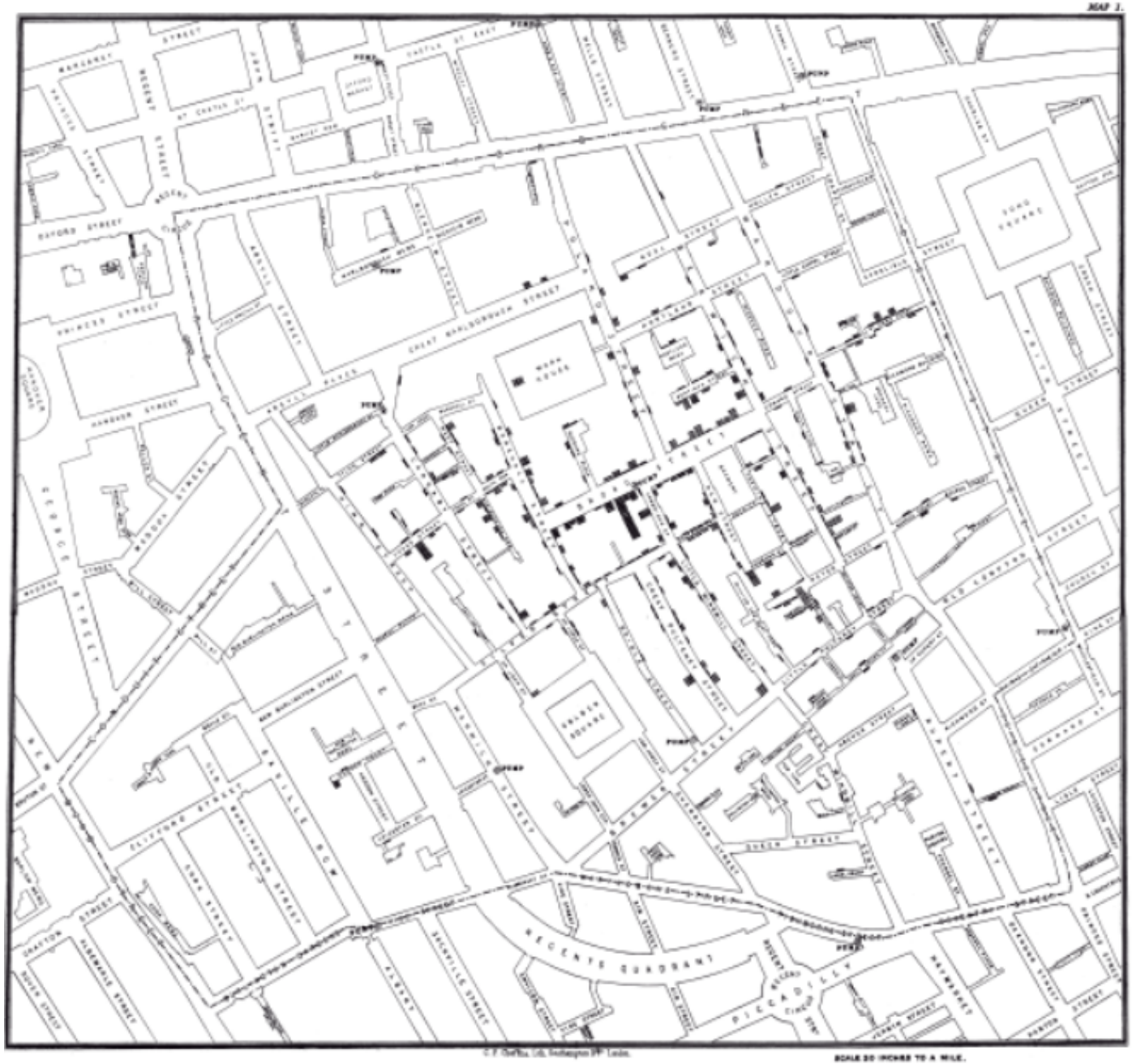
- modern nation-states with increased interest in collecting economic and demographic data
- descriptive statistical methods used before in physical sciences began to be used in social sciences (e.g. Adolphe Quelet, Francis Galton)
- dawn of new sciences, such as epidemiology

Florence Nigthingale (1820-1910) and the Crimean war

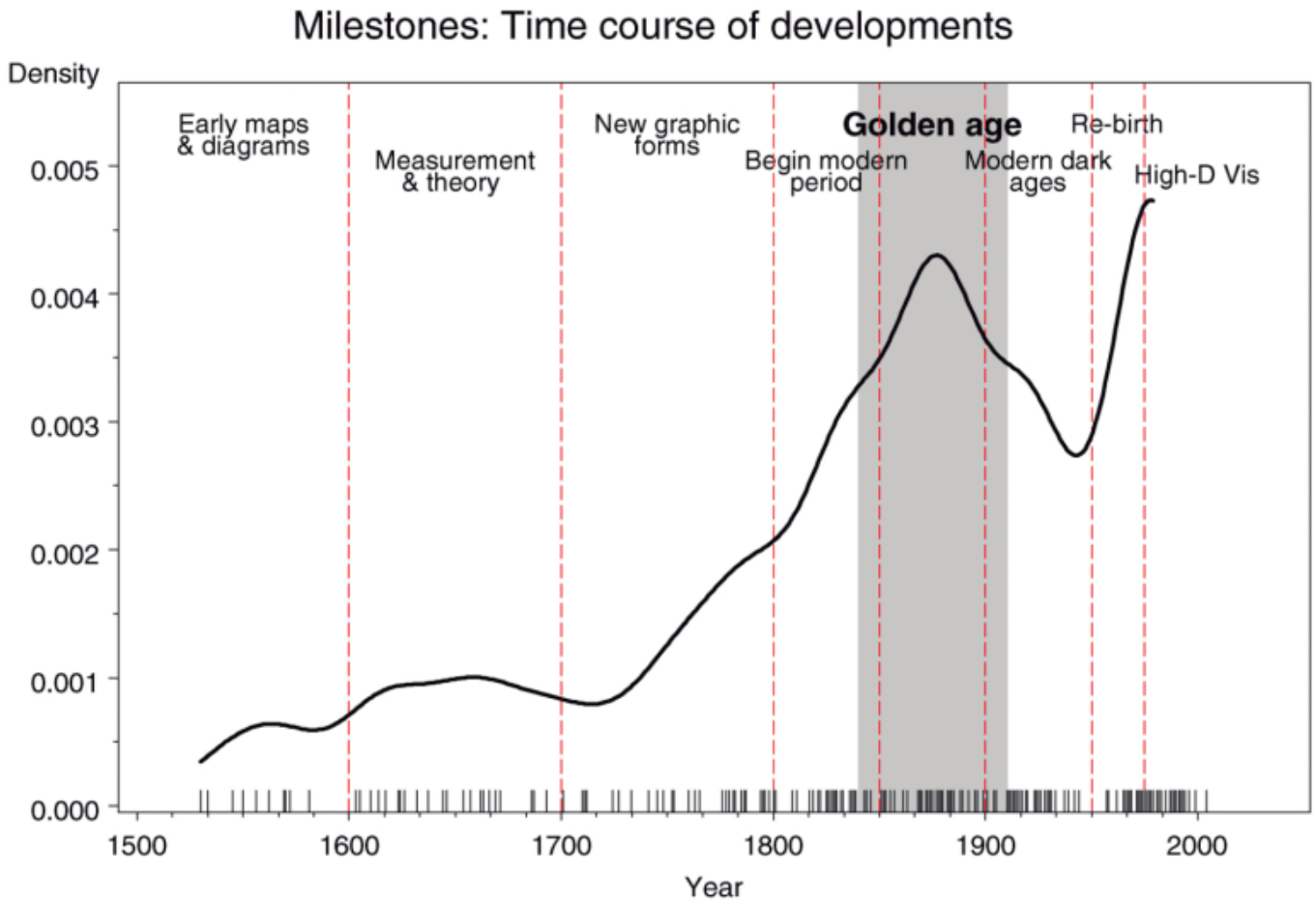


Causes of Mortality, 1856

John Snown (1813-1858) and cholera in London



Modern dark ages in statistics



Number of visualization historical landmarks per year, *Friendly 2008*

The pictorial turn in newspapers

Newspapers became a prime site where visual art and popular forces met and made their peace, and news contributed to the fullness of modernism as it arrived in the twentieth century [. . .] During the century, the newspapers in the study shifted from the abundant complexity of the Victorian era to the fixed simplicity of modernism. They adopted all the specific forms commentators identified with the modern style: fewer columns, prominent illustrations, horizontal layout, and simplified headline typography. (Barnhurst & Nerone 2001)

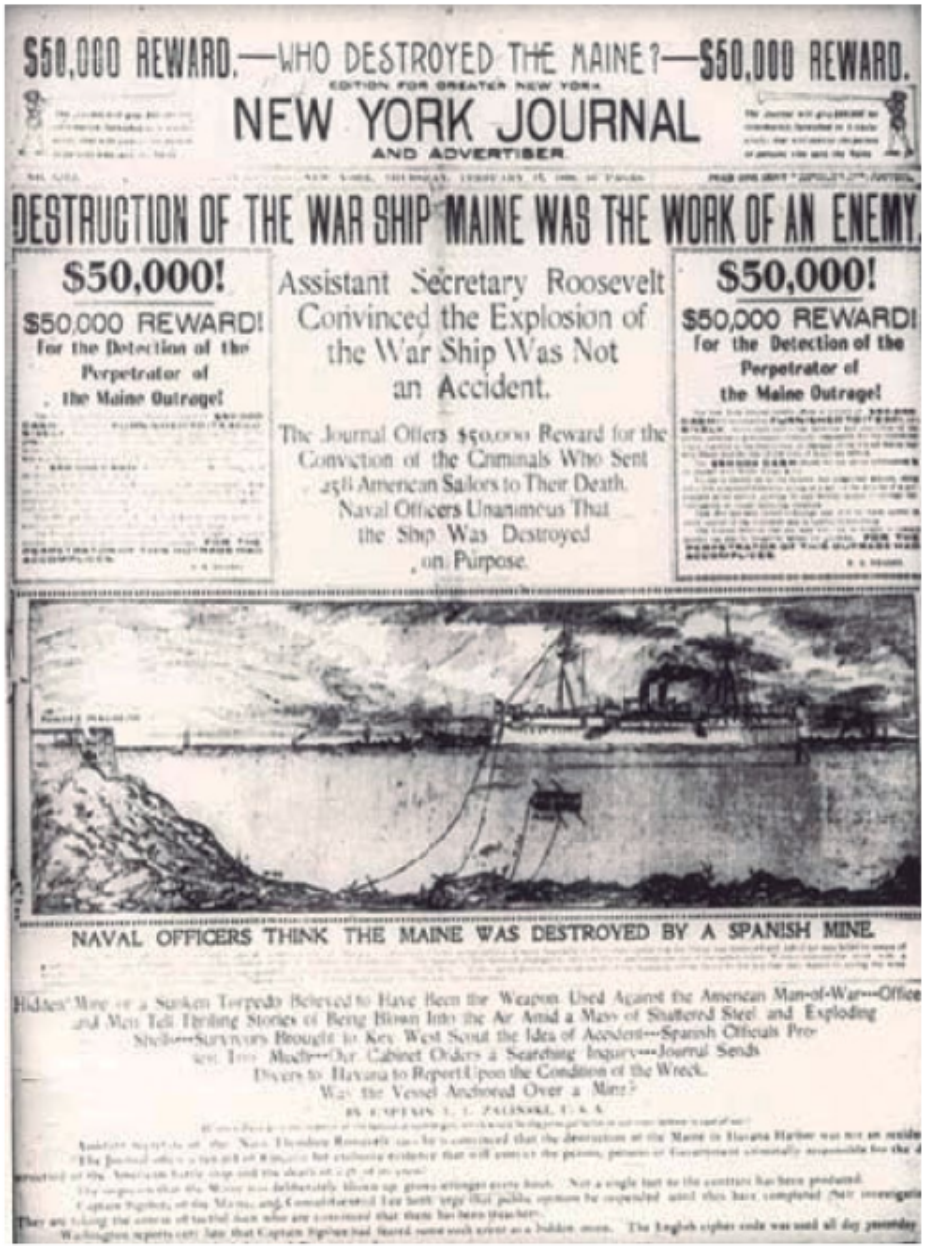
Yellow kid journalism (1895-1898)

Say what?

Sensational journalism in the circulation war between Joseph Pulitzer's *New York World* and William Randolph Hearst's *New York Journal* (Pulitzer tried to be more content-based but circulation shrank)

Yellow Kid, *New York World* and *New York Journal*

Yellow kid journalism (1895-1898)



$50,000 REWARD.—WHO DESTROYED THE MAINE?—$50,000 REWARD.

EDITION FOR GREATER NEW YORK

NEW YORK JOURNAL
AND ADVERTISER.

DESTRUCTION OF THE WAR SHIP MAINE WAS THE WORK OF AN ENEMY.

$50,000!
$50,000 REWARD!
For the Detection of the Perpetrator of the Maine Outrage!

Assistant Secretary Roosevelt Convinced the Explosion of the War Ship Was Not an Accident.

The Journal Offers $50,000 Reward for the Conviction of the Criminals Who Sent 258 American Sailors to Their Death.
Naval Officers Unanimous That the Ship Was Destroyed on Purpose.

$50,000!
$50,000 REWARD!
For the Detection of the Perpetrator of the Maine Outrage!

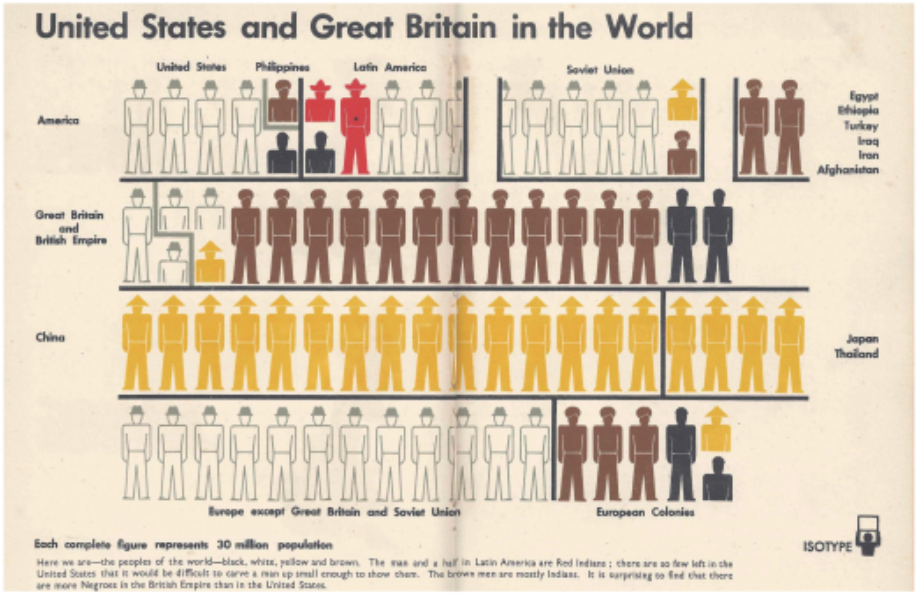
NAVAL OFFICERS THINK THE MAINE WAS DESTROYED BY A SPANISH MINE.

Hidden Mine or a Sunken Torpedo Believed to Have Been the Weapon Used Against the American Man-of-War—Officers and Men Tell Thrilling Stories of Being Blown Into the Air Amid a Mass of Shattered Steel and Exploding Shells—Survivors Brought to Key West Scout the Idea of Accident—Spanish Officials Protest Too Much—Our Cabinet Orders a Searching Inquiry—Journal Sends Divers to Havana to Report Upon the Condition of the Wreck.
Was the Vessel Anchored Over a Mine?

The sinking of Maine in the bay of Havana (notice the Spanish mine), *New York Journal*, Feb. 17, 1898

Viennese Museum for Society and the Economy (1924)

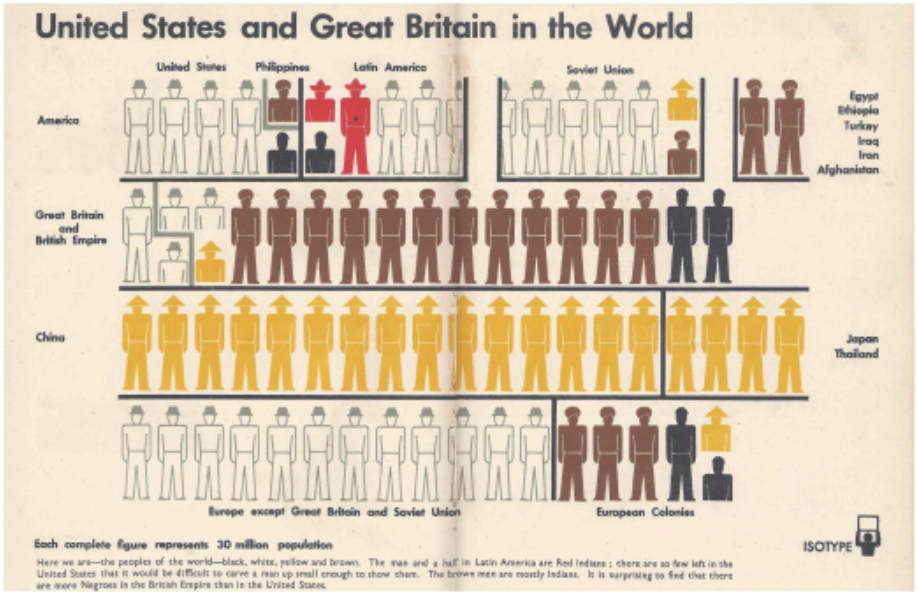
Facts for the uneducated



ISOTYPE, universal visual language by Neurath, Arntz and Reidemeister

Viennese Museum for Society and the Economy (1924)

Facts for the uneducated



ISOTYPE, universal visual language by Neurath, Arntz and Reidemeister

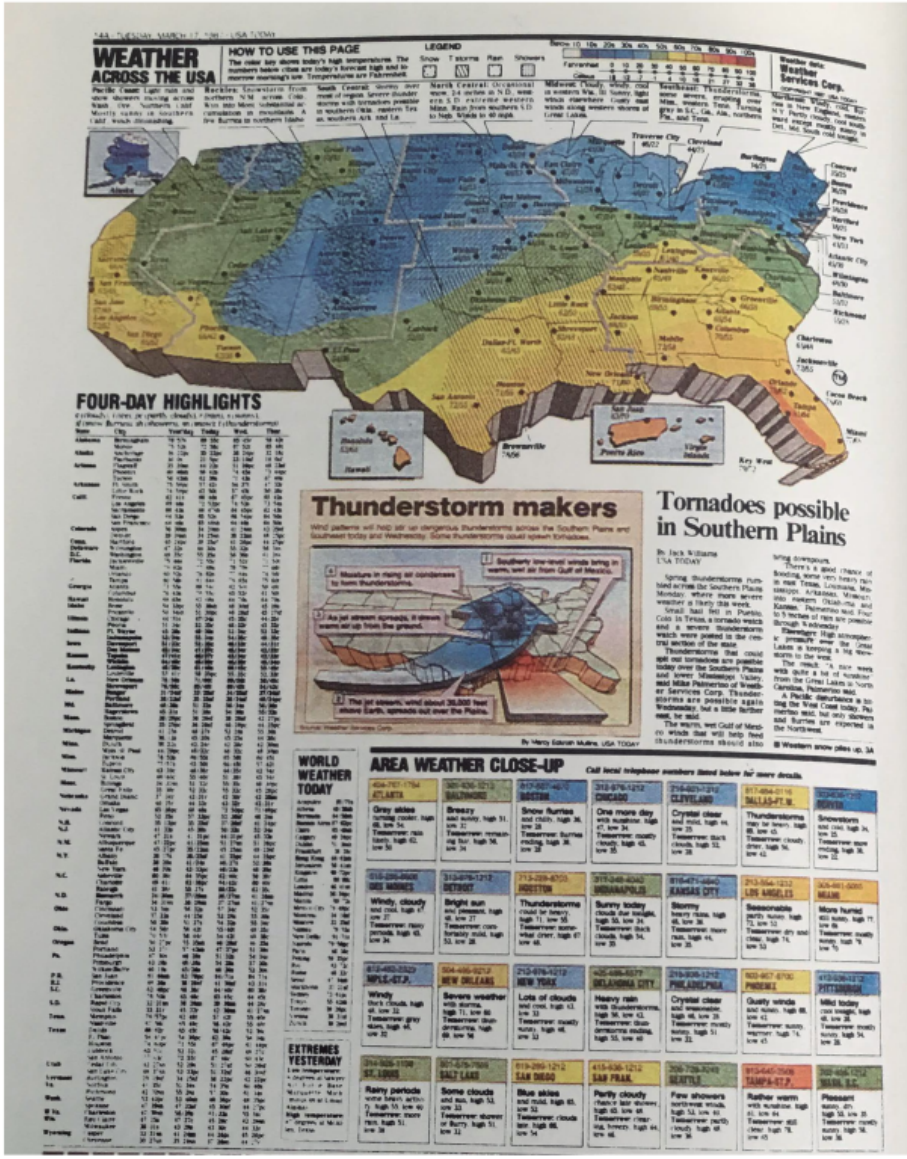
The "Bible"

Pictographs and Graphs: How to Make and Use Them, Modley & Lowenstein, 1952

ISOTYPE

A page from *Fortune*, 1929

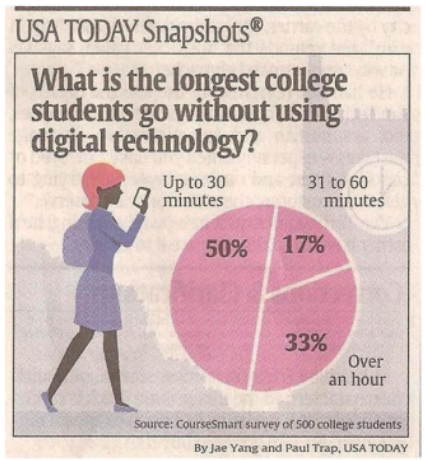
Birth of USA Today (1982)

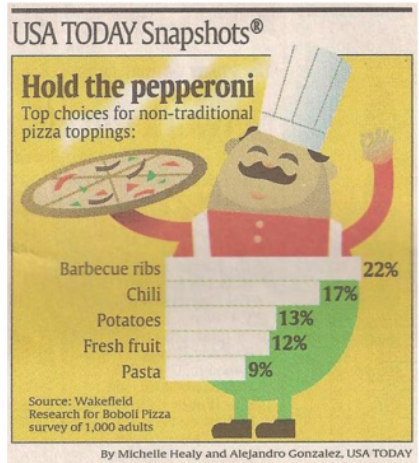


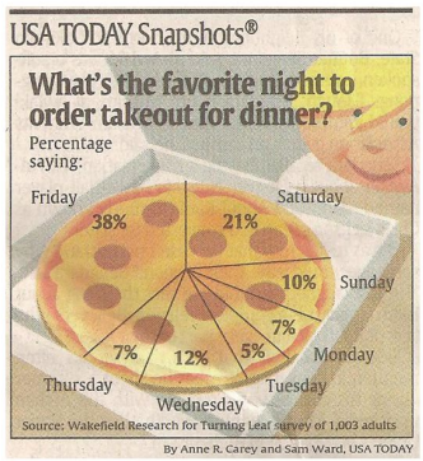
A revolutionary weather map

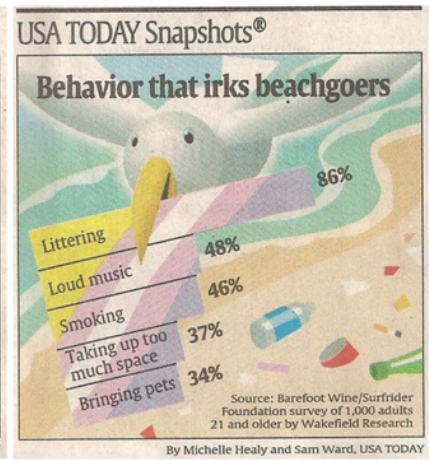
Birth of USA Today (1982)

- its success expanded the use of graphics in print publications
- tilted the stylistic balance towards the pictorial and lighthearted
- art training, no quantitative expertise
- in 1984 60% of 156 newspapers reported an increased use of news graphics, and an additional 22% said that they had just incorporated them into their pages





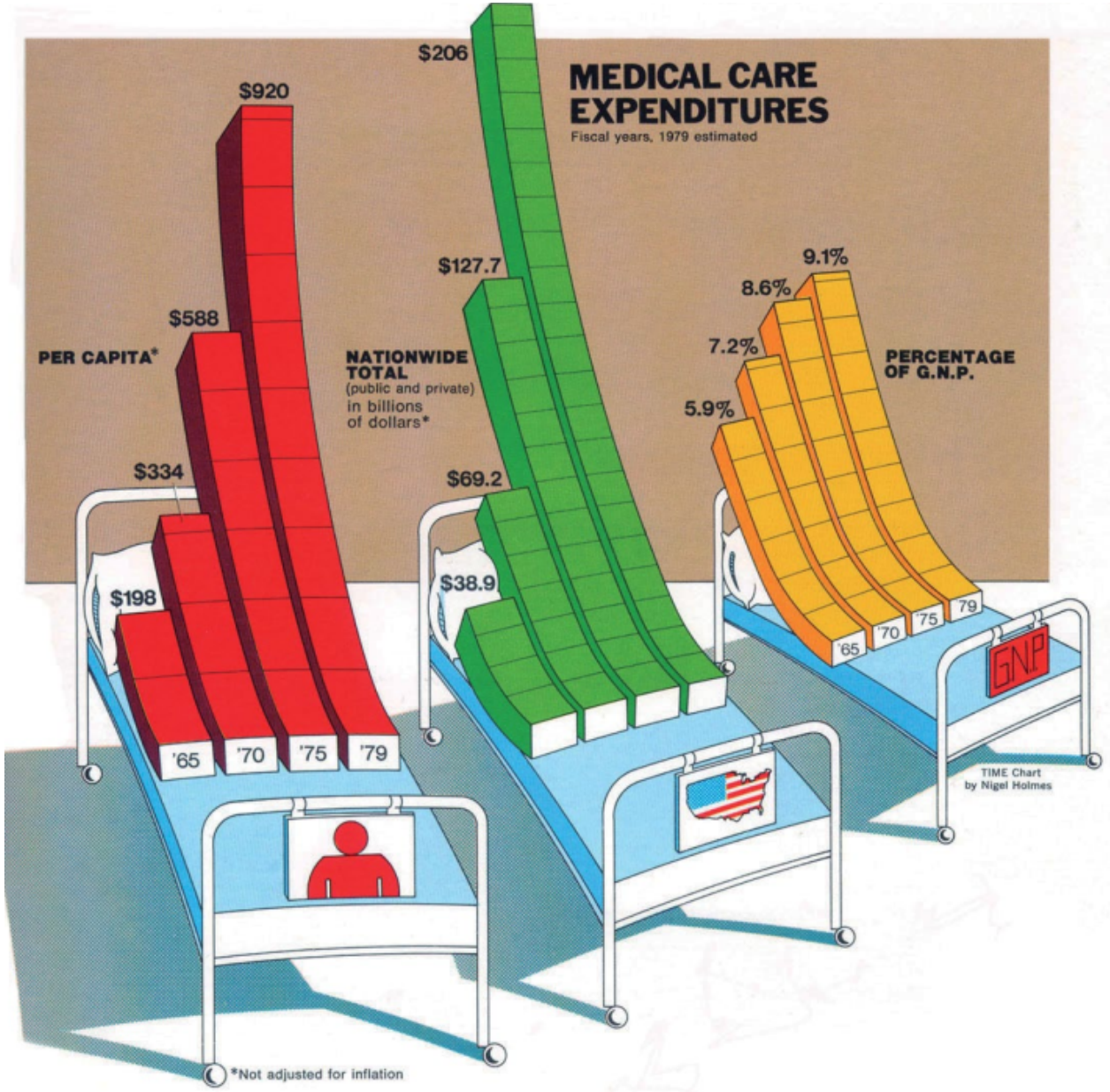




What's the problem?

Nearly all those who produce graphics for mass publication are trained exclusively in the fine arts and have had little experience with the analysis of data [. . .] Illustrators too often see their work as a exclusively artistic enterprise—the words "creative", "concept", and "style" combine regularly in all possible permutations, a Big Think jargon for the small task of constructing a time-series a few data points long. Those who get ahead are those who beautify data, never mind statistical integrity.
[Edward Tufte 1983]

Nigel Holmes

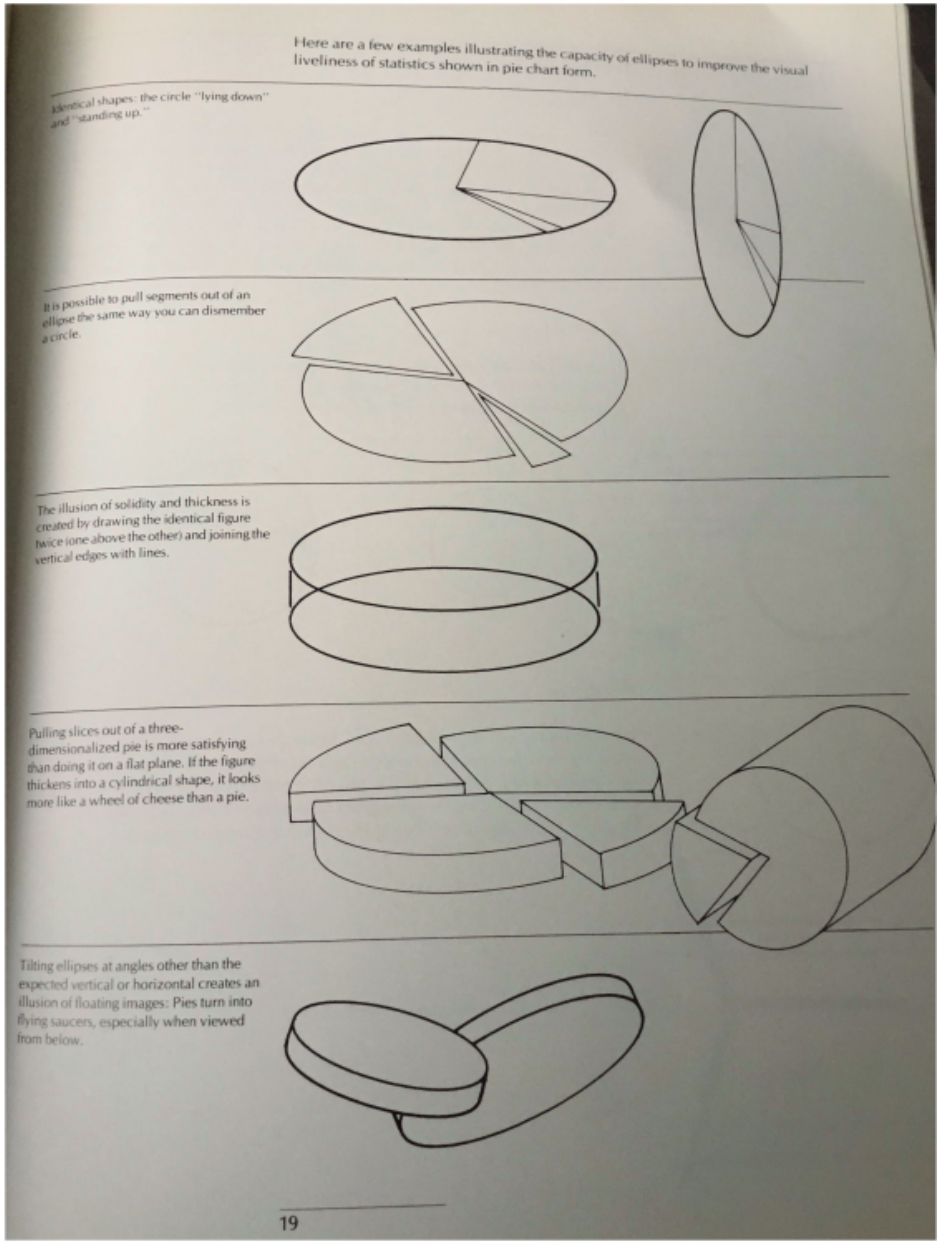


Nigel Holmes at *Time* (1978-1994)

Nigel Holmes

As long as the artist understands that the primary function is to convey statistics and respect that duty, then you can have fun (or be serious) with the image: that is, the form in which those statistics appear. Boredom is as much a threat in visual design as it is elsewhere in art and communication. The mind and eye demand stimulation and surprise.

Jan V. White



Here are a few examples illustrating the capacity of ellipses to improve the visual liveliness of statistics shown in pie chart form.

Identical shapes: the circle "lying down" and "standing up."

It is possible to pull segments out of an ellipse the same way you can dismember a circle.

The illusion of solidity and thickness is created by drawing the identical figure twice (one above the other) and joining the vertical edges with lines.

Pulling slices out of a three-dimensionalized pie is more satisfying than doing it on a flat plane. If the figure thickens into a cylindrical shape, it looks more like a wheel of cheese than a pie.

Tilting ellipses at angles other than the expected vertical or horizontal creates an illusion of floating images. Pies turn into flying saucers, especially when viewed from below.

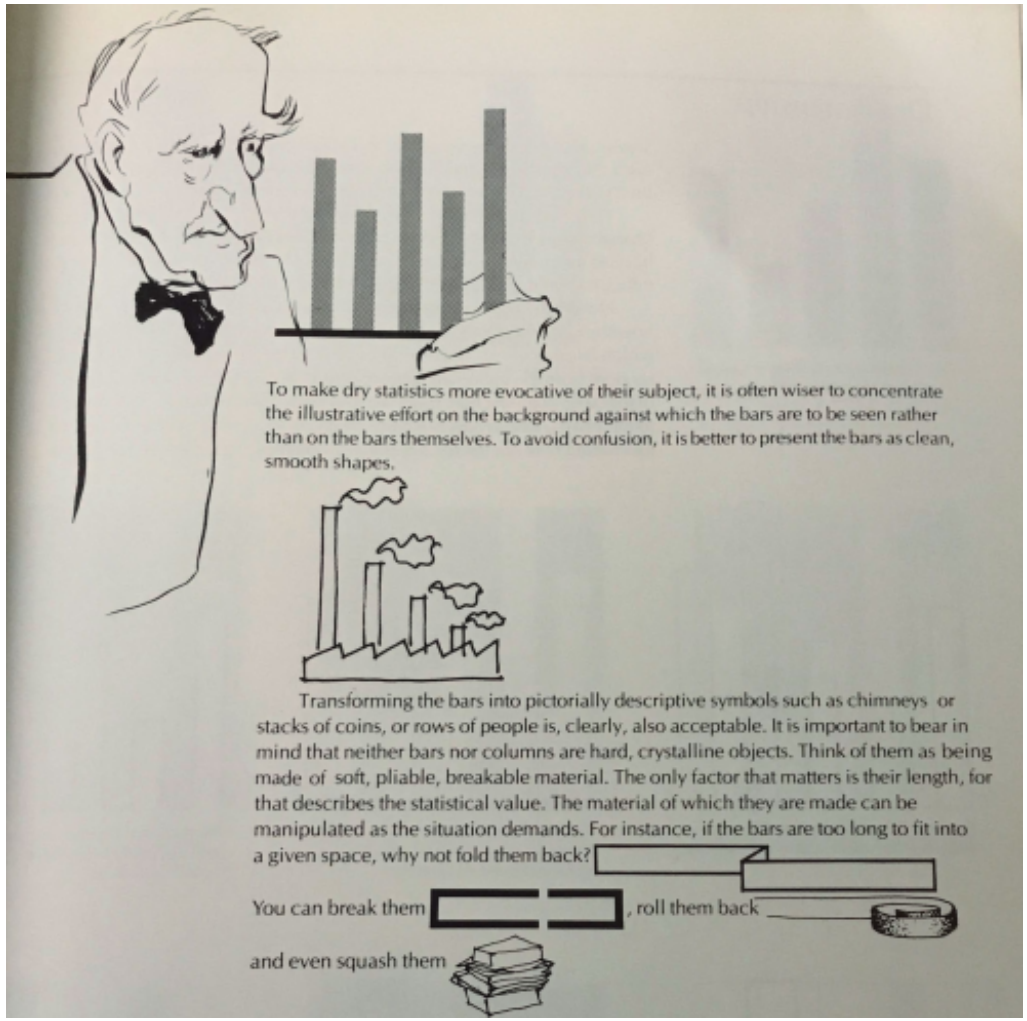
19

White's textbook on visualization, 1984

Jan V. White

To make dry statistics more evocative of their subject, it is often wiser to concentrate the illustrative effort on the background against which the bars are to be seen rather than on the bars themselves, [. . .] transforming the bars into pictorially descriptive symbols such as chimneys or stacks or coins, or rows of people is, clearly, also acceptable [. . .] The material of which they are made can be manipulated as the situation demands. For instance, if the bars are too long to fit into a given space, why not fold them back? You can break them, roll them back and even squash them.
(Jan. V. White, 1984)

Jan V. White

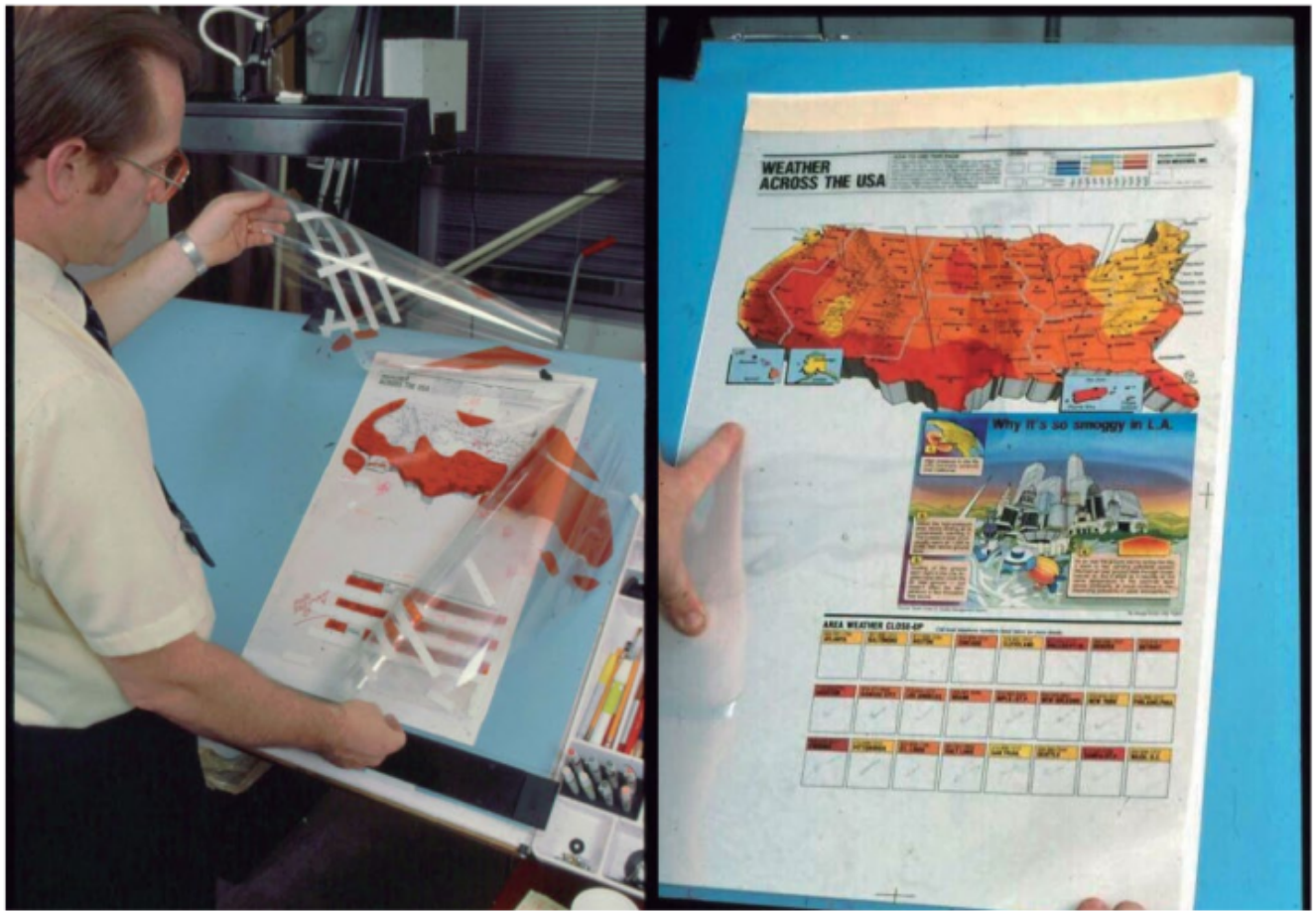


White's textbook on visualization, 1984

Computer-age graphics

George Rorick, hand-made visualisation, 11 a.m. to 6 p.m.

Computer-age graphics



George Rorick, hand-made visualisation, 11 a.m. to 6 p.m.

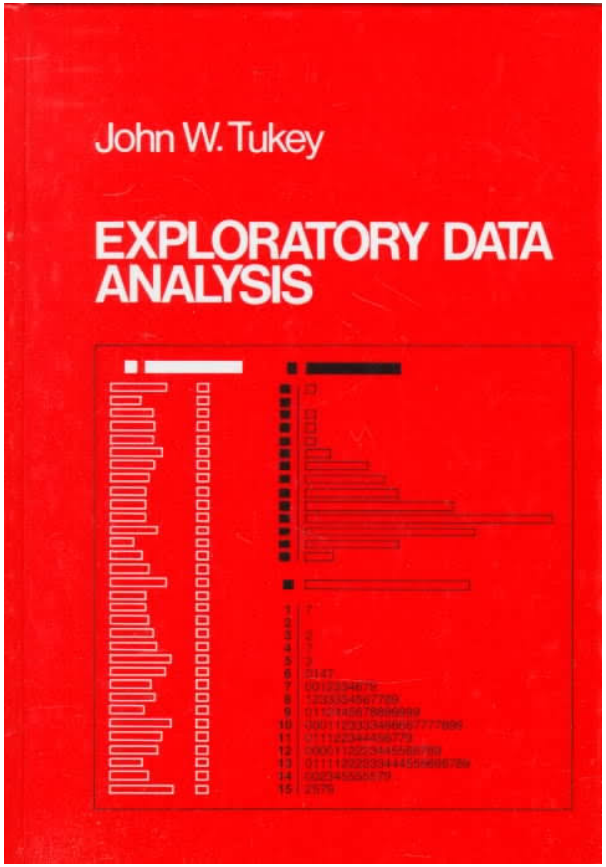
Computer-age graphics

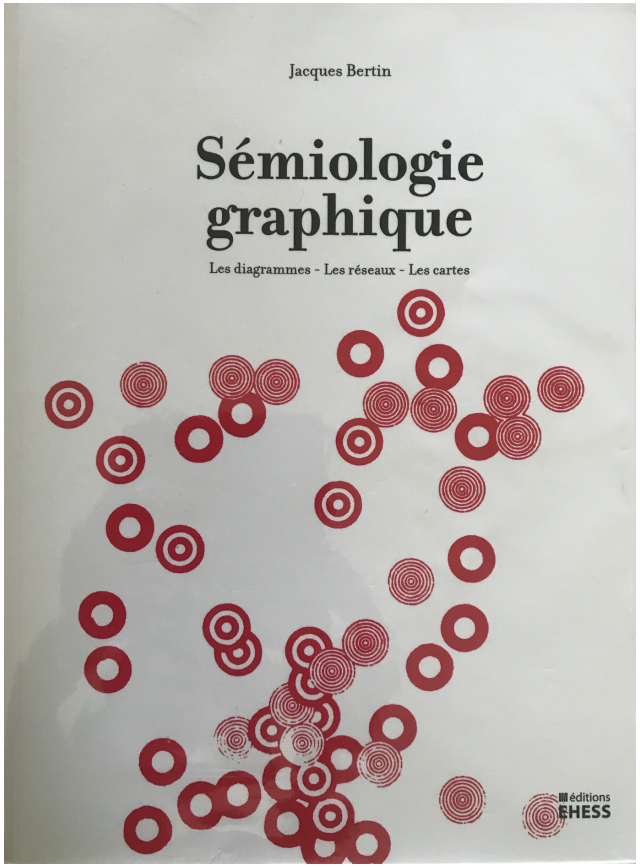
- Apple, 1984
- PostScript & Adobe Illustrator, 1987 (raster vs. vector files)
- Adobe Photoshop, 1989

We went from some very nice illustrated graphics to some very poor computer-generated graphics, but that was the limitations of the technology, and it took about at least five years, maybe more, before we started to see the computer graphics start to rise up in quality.
John Grimwade (check out his website!)

Backlash against chartoons

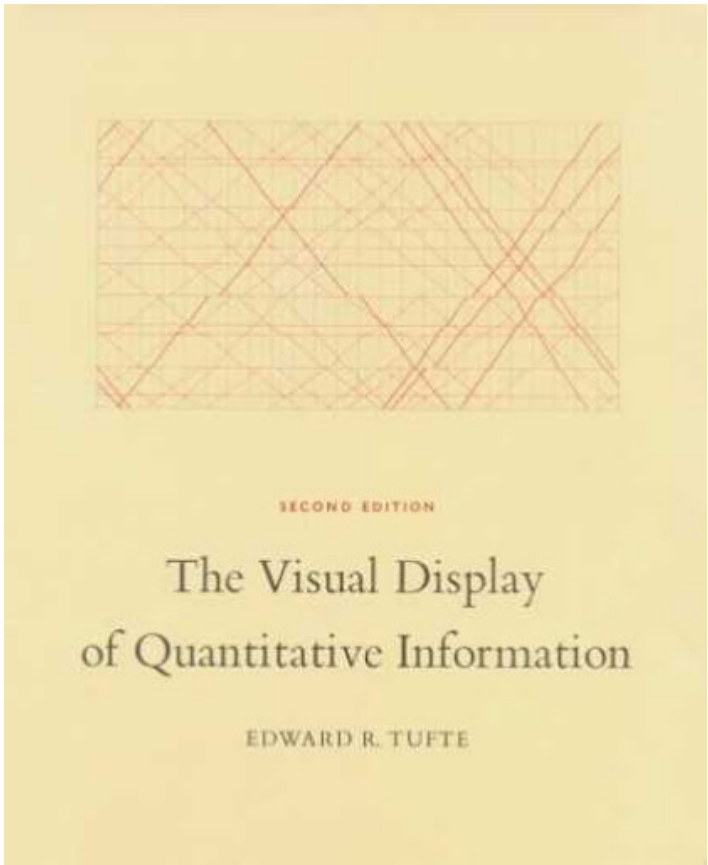
Tukey 1977, Bertin 1967

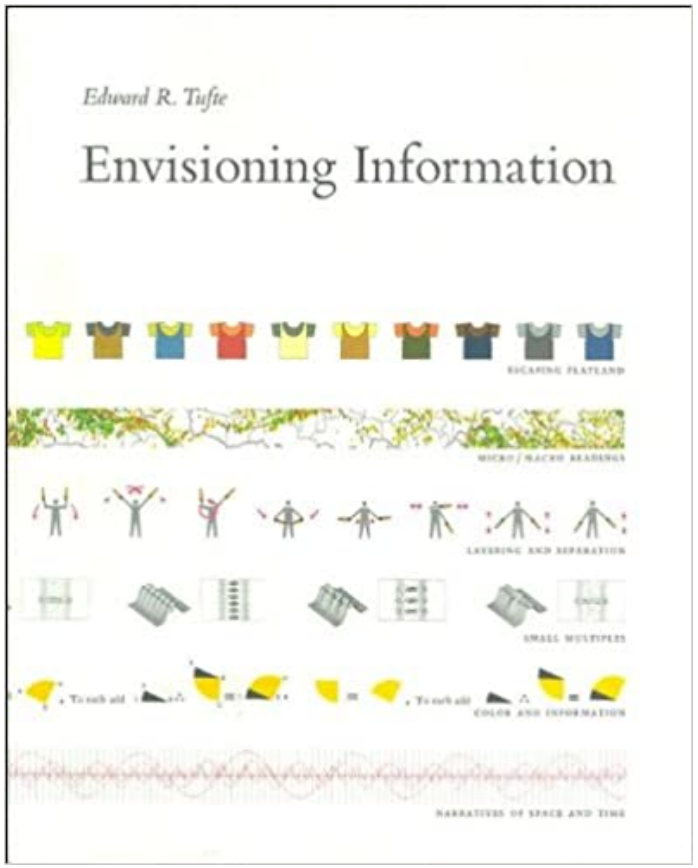




Backlash against chartoons

Tufte 1983, 1990





Backlash against chartoons

Sometimes decoration can help editorialize about the substance of the graphic. But it is wrong to distort the data measures —the ink locating values of numbers— in order to make an editorial comment or fit a decorative scheme.
(Tufte 1983: 59)

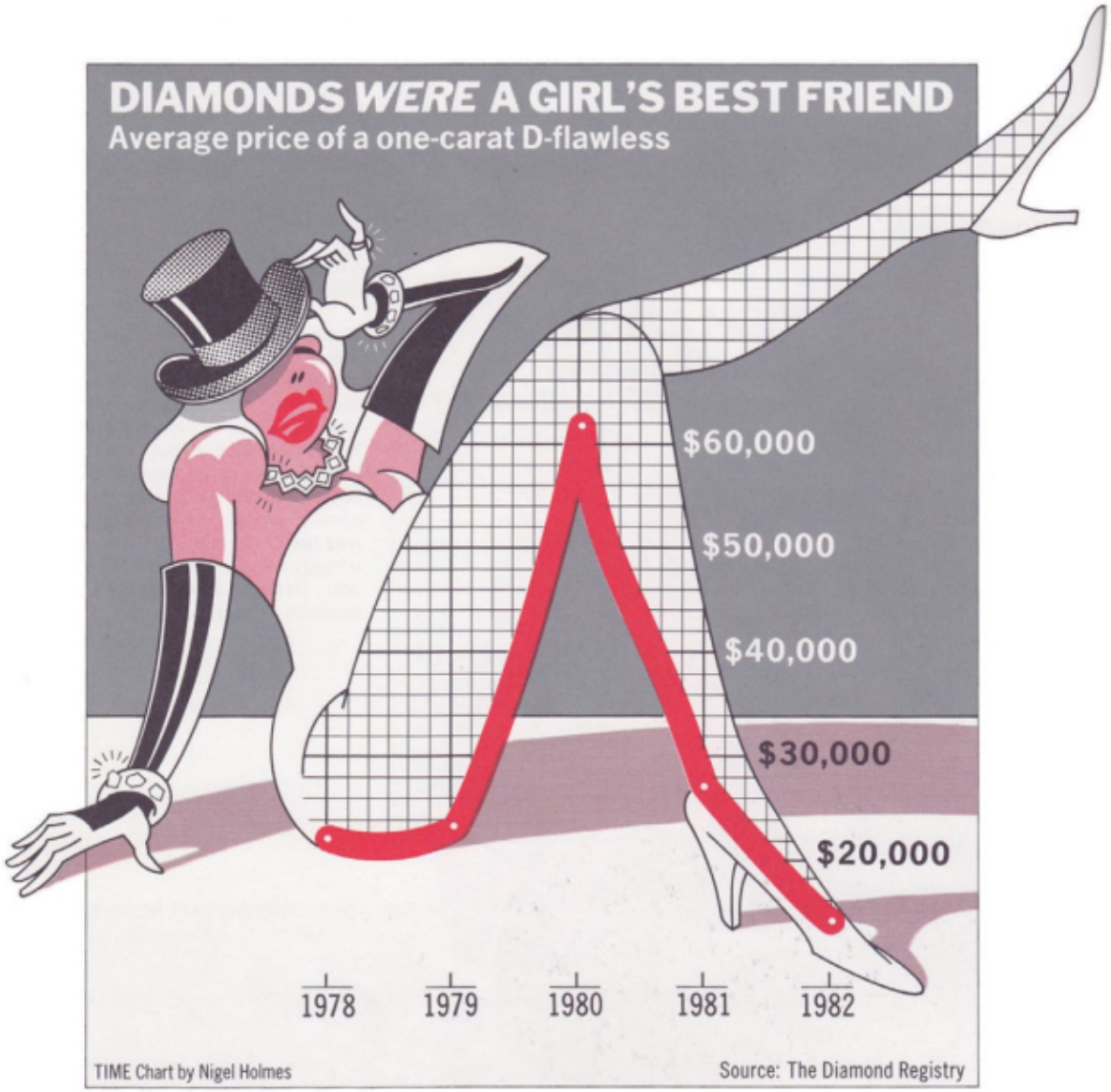
Backlash against chartoons

If you belong to the school of people who believe that charts should only present statistics in the most straightforward, plain way, with no other visual help to the reader, for example, than the bar of the bar chart, the line of the fever graph, the circle of the pie chart, or the rules of the table, then move on to another part of the book [...] Boredom is as much a threat in visual design as it is elsewhere in art and communication. The mind and eye demand stimulation and surprise [...] Even a smile will encourage a reader to look into the statistics he or she might not have thought of reading in a less embellished chart. (Holmes 1984: 72)

Backlash against chartoons

Too many data presentations [. . .] seek to attract and divert attention by means of display apparatus and ornament. Chartjunk has come to corrupt all sorts of information exhibits and computer interfaces (Tufte 1990: 33)

Backlash against chartoons



Holmes' chart in the *Times magazine*

Backlash against chartoons

Consider this unsavory exhibit at right —chockablock with cliché and stereotype, coarse humor, and a content-empty third dimension. Is it the product of a visual sensitivity in which a thigh-graph with a fishnet-stocking grid counts as Creative Concept. [...] Lurking behind chartjunk is contempt for both information and for the audience. Chartjunk promoters imagine that numbers and details are boring, dull, and tedious, requiring ornament to enliven. Cosmetic decoration, which frequently distorts the data, will never salvage an underlying lack of content. If the numbers are boring, then you've got the wrong numbers. Credibility vanishes in clouds of chartjunk; who would trust a chart that looks like a video game? (Tufte 1990: 34).

Backlash against chartoons

Graphical competence demands three quite different skills: the substantive, statistical, and artistic. Yet now [in the early 80s] most graphical work, particularly at news publications, is under the direction of but a single expertise —the artistic. Allowing artist-illustrators to control the design and content of statistical graphics is almost like allowing typographers to control the content, style, and editing of prose.
(Tufte 1983: 87).

Recent developments

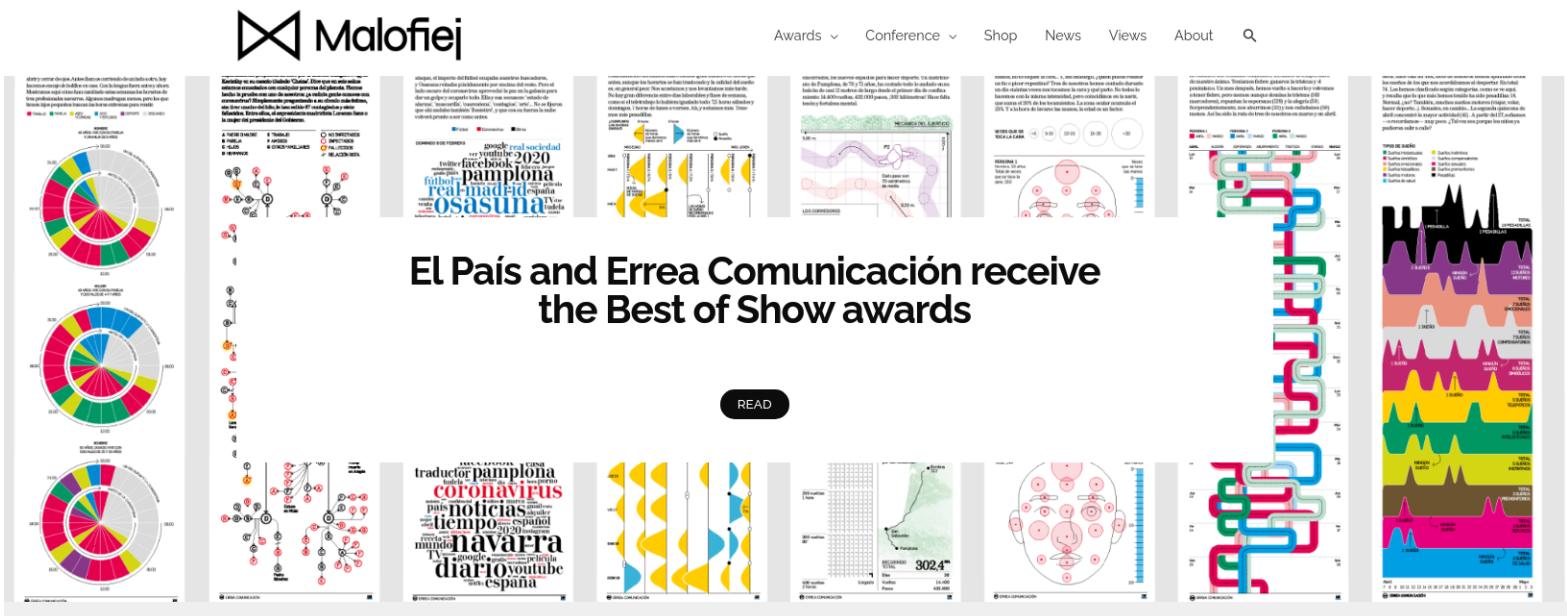
Recent developments

Geek takeover

- more information density and more data
- visualization desks more independent from arts departments
- the 90s and early 2000s: illustration-driven explanations, sometimes supplemented by small and straight-forward statistical graphs and data maps
- today, the balance has shifted to presentations that rely mainly on the visual display of data, both quantitative and qualitative
- often, no longer detached "graphics departments". Data journalists, nerd journalism!

Recent developments

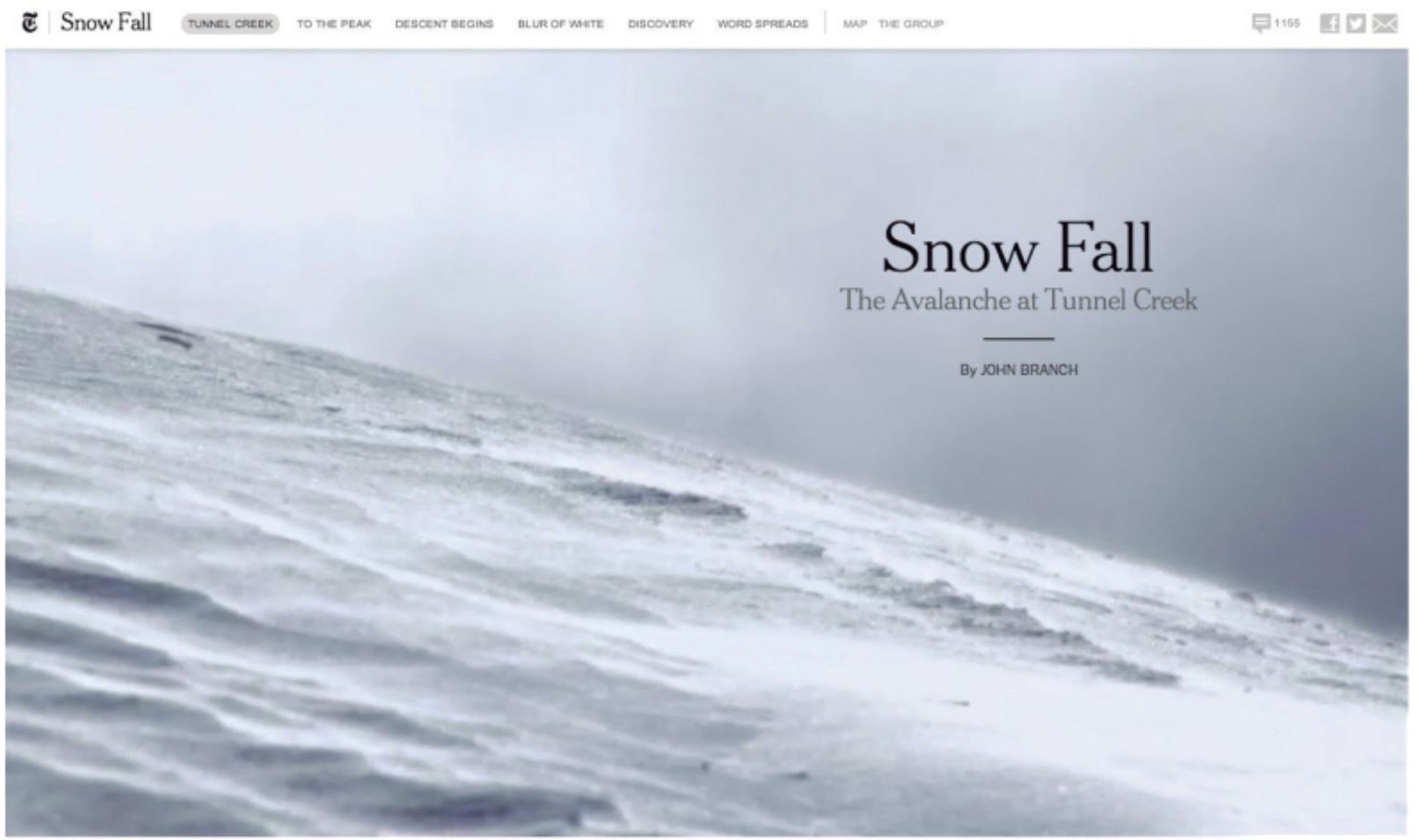
Check out Malofiej awards (1992)



Malofiej awards website

Recent developments

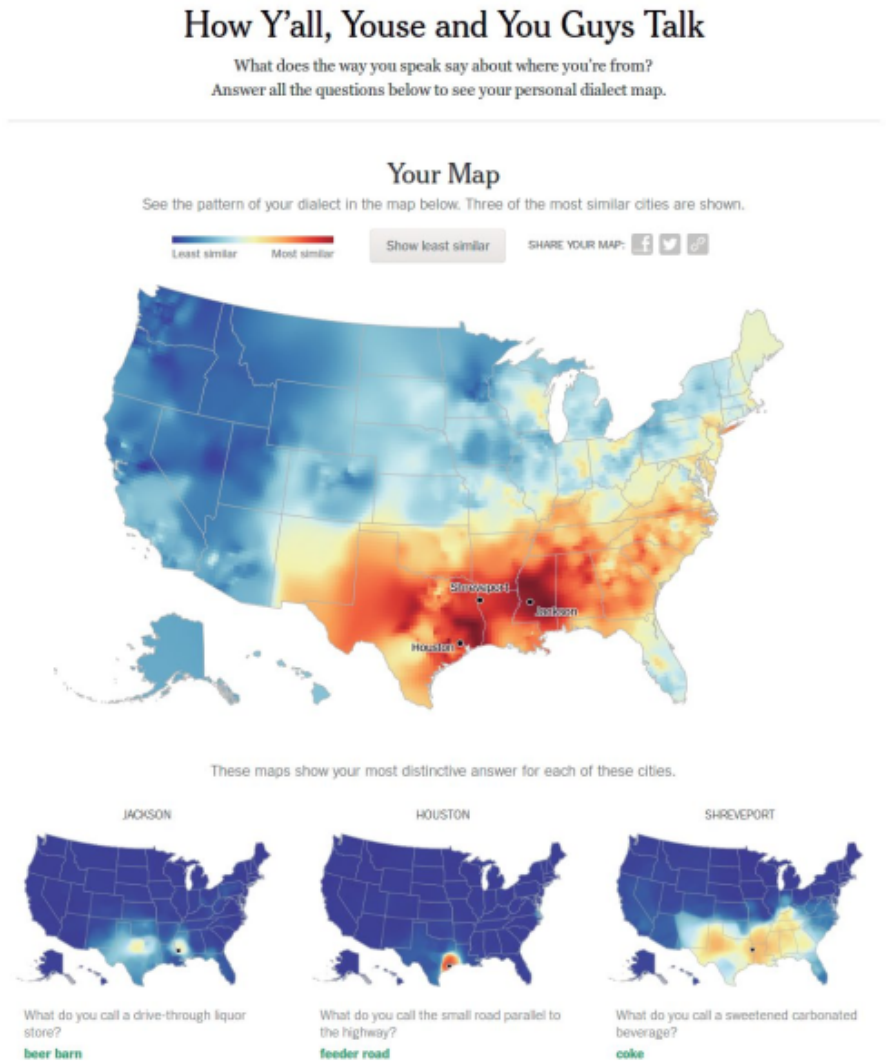
Example ("new era", 3 mln. in no time)



Snowfall at NY Times

Recent developments

Example (most popular piece in Times, 2013)



How Y'all quiz, NYT

For the tutorial

Complete the introductory instructions about github, bring a flash drive!