Wizualizacja danych

Wykład 1: Wizualizacja danych

Mariusz Tarnopolski

Instytut Astronomii UMK

DataViz, 2024/2025

Sprawy organizacyjne

- 15 godz. konwersatorium = wykład + dyskusje + problemy (dwugodzinne bloki, co dwa tygodnie wg harmonogramu)
- Website
- Konsultacje — wg potrzeb, proszę się umawiać przez maila:

mariusz.tarnopolski@umk.pl

Sprawy organizacyjne

- 15 godz. konwersatorium = wykład + dyskusje + problemy (dwugodzinne bloki, co dwa tygodnie wg harmonogramu)
- Website
- Konsultacje — wg potrzeb, proszę się umawiać przez maila:

  mariusz.tarnopolski@umk.pl

Zaliczenie (warunki inkrementalne)

- uczestnictwo w zajęciach (maks. 1 nieob. nieuspr. $\Rightarrow$ 3.0)
- aktywność ($\Rightarrow$ 4.0)
- omówienie* jednego wykresu/diagramu/ilustracji/itp. z artykułów jakie pojawią się w czasie semestru na https://arxiv.org/ ($\Rightarrow$ 5.0):

*(1) Coś więcej niż „za mała czcionka" lub "to jest wykres kołowy". (2) W formie pisemnej elektronicznej, przesłanej mailem: w treści maila lub w załączonym pliku. Równoważność 5 min. prezentacji, czyli ok. 0.5 strony A4. Podać link do artykułu (identyfikator arXiv) oraz numer omawianego wykresu.

Cornell University

We gratefully acknowledge support from the Simons Foundation and member institutions.

arXiv

Login

Search... All fields Search

Help | Advanced Search

arXiv is a free distribution service and an open-access archive for 2,257,602 scholarly articles in the fields of physics, mathematics, computer science, quantitative biology, quantitative finance, statistics, electrical engineering and systems science, and economics. Materials on this site are not peer-reviewed by arXiv.

Subject search and browse:

Physics | Search | Form Interface | Catchup

News

Read about recent news and updates on arXiv's blog. (View the former "what's new" pages here).

Physics

- Astrophysics (**astro-ph** new, recent, search)
  includes: Astrophysics of Galaxies; Cosmology and Nongalactic Astrophysics; Earth and Planetary Astrophysics; High Energy Astrophysical Phenomena; Instrumentation and Methods for Astrophysics; Solar and Stellar Astrophysics
- Condensed Matter (**cond-mat** new, recent, search)
  includes: Disordered Systems and Neural Networks; Materials Science; Mesoscale and Nanoscale Physics; Other Condensed Matter; Quantum Gases; Soft Condensed Matter; Statistical Mechanics; Strongly Correlated Electrons; Superconductivity
- General Relativity and Quantum Cosmology (**gr-qc** new, recent, search)
- High Energy Physics - Experiment (**hep-ex** new, recent, search)
- High Energy Physics - Lattice (**hep-lat** new, recent, search)
- High Energy Physics - Phenomenology (**hep-ph** new, recent, search)
- High Energy Physics - Theory (**hep-th** new, recent, search)
- Mathematical Physics (**math-ph** new, recent, search)
- Nonlinear Sciences (**nlin** new, recent, search)
  includes: Adaptation and Self-Organizing Systems; Cellular Automata and Lattice Gases; Chaotic Dynamics; Exactly Solvable and Integrable Systems; Pattern Formation and Solitons
- Nuclear Experiment (**nucl-ex** new, recent, search)
- Nuclear Theory (**nucl-th** new, recent, search)
- Physics (**physics** new, recent, search)
  includes: Accelerator Physics; Applied Physics; Atmospheric and Oceanic Physics; Atomic and Molecular Clusters; Atomic Physics; Biological Physics; Chemical Physics; Classical Physics; Computational Physics; Data Analysis, Statistics and Probability; Fluid Dynamics; General Physics; Geophysics; History and Philosophy of Physics; Instrumentation and Detectors; Medical Physics; Optics; Physics and Society; Physics Education; Plasma Physics; Popular Physics; Space Physics
- Quantum Physics (**quant-ph** new, recent, search)

Mathematics

- Mathematics (**math** new, recent, search)
  includes (see detailed description): Algebraic Geometry; Algebraic Topology; Analysis of PDEs; Category Theory; Classical Analysis and ODEs; Combinatorics; Commutative Algebra; Complex Variables; Differential Geometry; Dynamical Systems; Functional Analysis; General Mathematics; General Topology; Geometric Topology; Group Theory; History and Overview; Information Theory; K-Theory and Homology; Logic; Mathematical Physics; Metric Geometry; Number Theory; Numerical Analysis; Operator Algebras; Optimization and Control; Probability; Quantum Algebra; Representation Theory; Rings and Algebras; Spectral Theory; Statistics Theory; Symplectic Geometry

Computer Science

- Computing Research Repository (**CoRR** new, recent, search)
  includes (see detailed description): Artificial Intelligence; Computation and Language; Computational Complexity; Computational Engineering, Finance, and Science; Computational Geometry; Computer Science and Game Theory; Computer Vision and Pattern Recognition; Computers and Society; Cryptography and Security; Data Structures and Algorithms; Databases; Digital Libraries; Discrete Mathematics; Distributed, Parallel, and Cluster Computing; Emerging Technologies; Formal Languages and Automata Theory; General Literature; Graphics; Hardware Architecture; Human-Computer Interaction; Information Retrieval; Information Theory; Logic in Computer Science; Machine Learning; Mathematical Software; Multiagent Systems; Multimedia; Networking and Internet Architecture; Neural and Evolutionary Computing; Numerical Analysis; Operating Systems; Other Computer Science; Performance; Programming Languages; Robotics; Social and Information Networks; Software Engineering; Sound; Symbolic Computation; Systems and Control

Quantitative Biology

- Quantitative Biology (**q-bio** new, recent, search)
  includes (see detailed description): Biomolecules; Cell Behavior; Genomics; Molecular Networks; Neurons and Cognition; Other Quantitative Biology; Populations and Evolution; Quantitative Methods; Subcellular Processes; Tissues and Organs

Quantitative Finance

- Quantitative Finance (**q-fin** new, recent, search)
  includes (see detailed description): Computational Finance; Economics; General Finance; Mathematical Finance; Portfolio Management; Pricing of Securities; Risk Management; Statistical Finance; Trading and Market Microstructure

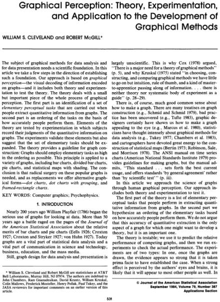
Plan ramowy kursu

1. Historia, cele i przykłady DataViz; cele kursu
2. Podstawowe rodzaje wykresów
3. Fundamentalne zasady projektowania graficznego
4. Typografia
5. Teoria koloru
6. Aspekty kognitywne (w tym złudzenia optyczne)
7. Przegląd dobrych & złych wykresów naukowych
8. Przegląd (głównie złych) wykresów z mediów
9. Wykresy specjalne
10. Wizualizacje 3D
11. Schematy blokowe (*flowcharts*)
12. Dane wielowymiarowe*: redukcja wymiaru, współrzędne równoległe
13. Tabele*

*Tylko jeśli czas pozwoli.

Literatura

podstawowa

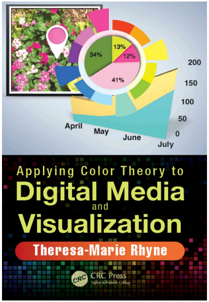

Graphical Perception: Theory, Experimentation, and Application to the Development of Graphical Methods

WILLIAM S. CLEVELAND and ROBERT McGILL*

1. INTRODUCTION

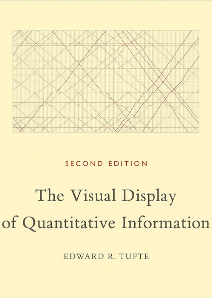
© Journal of the American Statistical Association
September 1984, Volume 79, Number 387
Applications Section

link



link

link



link


CURRENT CONCEPTS

Graphs, Tables, and Figures in Scientific Publications: The Good, the Bad, and How Not to Be the Latter

Lauren E. Franzblau, Kevin C. Chung, MD, MS

Graphs, figures, and tables can save readers time and energy, aid their understanding of an article, and reduce the word count of the main text. However, many graphics submitted to and published in scientific journals fail to meet their potential and include mistakes that jeopardize their clarity. Many formats are available for presenting data, as well as a variety of techniques for enhancing interpretability. When the appropriate format is used to depict data, it conveys the greatest amount of information in the clearest fashion, complements the text, and deepens readers' understanding. The aims of this article are to draw attention to the necessity of well-constructed graphs, tables, and figures in scientific publications, and to show how to create them. *(J Hand Surg 2012;37A:591–596. Copyright © 2012 by the American Society for Surgery of the Hand. All rights reserved.)*

Key words Graphs, tables, charts, figures, manuscripts.

IN THE HEART OF EVERY research article is its data, the information that gives rise to conclusions. However, without proper presentation, the data can be misinterpreted or, worse, ignored. To elude this trap, authors must use well-constructed graphs, figures, and tables to display data and trends, and to summarize information. Illustrations, or graphics, have 3 main purposes and advantages over text: first, they portray complex data and relationships in a way that is easier to interpret and understand[1–8]; second, they reduce reading time by summarizing and highlighting key findings or trends so that they are readily visible[5,6]; and third, they reduce the overall word count.[3,4,7]

Well-designed graphics make data easier to understand by explicitly depicting trends and key points, and they are valuable assets to any paper or presentation. These visual aids enrich readers' understanding and improve the accuracy of their interpretations of the data.[9] On the other hand, poorly fashioned graphics slow down retrieval of data and lead to incorrect interpretations of findings.[4,6,8] Thus, it is imperative that the quality of graphs, figures, and tables that are submitted to and published in scientific journals be carefully considered and for data to be presented in such a way that strengthens the overall manuscript. The aims of this article are to bring to light common mistakes in graph, table, and figure construction, and to show how to create high-quality graphics that will enhance readers' acquisition and understanding of data.

WHAT NOT TO DO

Many submitted, and even published, manuscripts contain illustrations of "suboptimal quality."[8] Table 1 shows common errors that lead to this failure to fulfill graphical and tabular potential. The most glaring oversight is selecting the wrong format to present data. Suboptimal designs omit important information about the data that should be displayed.

Many computer programs offer graphing features that are easy to use and have a plethora of design elements. However, they tend to produce overcompli-

From the Section of Plastic Surgery, Department of Surgery, University of Michigan Health System; Ann Arbor, MI.

Received for publication November 11, 2011; accepted in revised form December 27, 2011.

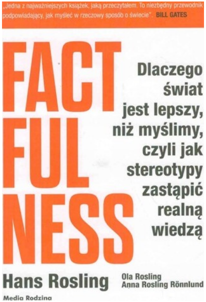
Supported in part by grants from the National Institute on Aging and the National Institute of Arthritis and Musculoskeletal and Skin Diseases, the National Institute of Arthritis and Musculoskeletal and Skin Diseases, and a Midcareer Investigator Award in Patient-Oriented Research (to K.C.C.).

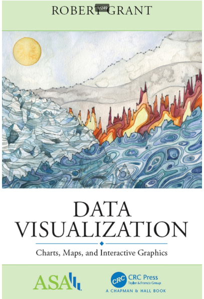
No benefits in any form have been received or will be received related directly or indirectly to the subject of this article.

Corresponding author: Kevin C. Chung, MD, MS, Section of Plastic Surgery, Department of Surgery, University of Michigan Health System, 2130 Taubman Center, SPC 5340, 1500 E. Medical Center Drive, Ann Arbor, MI 48109-0340; e-mail: kecchung@umich.edu.

Current Concepts

© 2012 ASSH • Published by Elsevier, Inc. All rights reserved. • 591





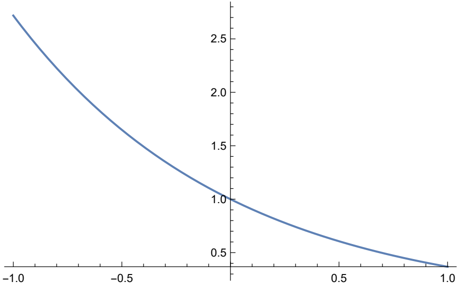
link link link

oraz dodatkowe materiały podane w trakcie zajęć.

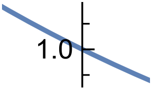
Oprogramowanie
(o tym ten kurs nie jest)

python	Matlab	Excel	Impress
matplotlib, seaborn, plotnine, bokeh, ...			
R	Mathematica	Photoshop	gimp
ggplot2, ...			
Java Script	gnuplot	Illustrator	inkscape
d3.js, ...			

Formaty plików

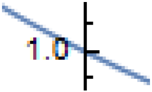


Wektorowe (.eps, .pdf, .svg):



- Nieskończona rozdzielczość

Rastrowe (.jpeg, .png, .tiff):



- Zdjęcia
- Sposób na zmniejszenie wielkości plików

Cele DataViz

1. Funkcja ilustracyjna:
   wykres jest wart więcej niż tysiąc słów, tj. opisanie przebiegu wielkości w czasie, złożona zależność jednej wielkości od drugiej, rozkład całości na składowe, itd.
2. Funkcja eksploratywna:
   walidacja; poszukiwanie wzorców, np. okresowości w czasie/przestrzeni, nadreprezentacja pewnej subpopulacji, grupowanie, itd.

1a) Kwartet Anscombe'a

Some of these points are illustrated by four fictitious data sets, each consisting of eleven (x, y) pairs, shown in the table. For the first three data sets the x-values are the same, and they are listed only once.

Data set	1-3	1	2	3		4	4
Variable	x	y	y	y		x	y
Obs. no. 1 :	10.0	8.04	9.14	7.46	:	8.0	6.58
2 :	8.0	6.95	8.14	6.77	:	8.0	5.76
3 :	13.0	7.58	8.74	12.74	:	8.0	7.71
4 :	9.0	8.81	8.77	7.11	:	8.0	8.84
5 :	11.0	8.33	9.26	7.81	:	8.0	8.47
6 :	14.0	9.96	8.10	8.84	:	8.0	7.04
7 :	6.0	7.24	6.13	6.08	:	8.0	5.25
8 :	4.0	4.26	3.10	5.39	:	19.0	12.50
9 :	12.0	10.84	9.13	8.15	:	8.0	5.56
10 :	7.0	4.82	7.26	6.42	:	8.0	7.91
11 :	5.0	5.68	4.74	5.73	:	8.0	6.89

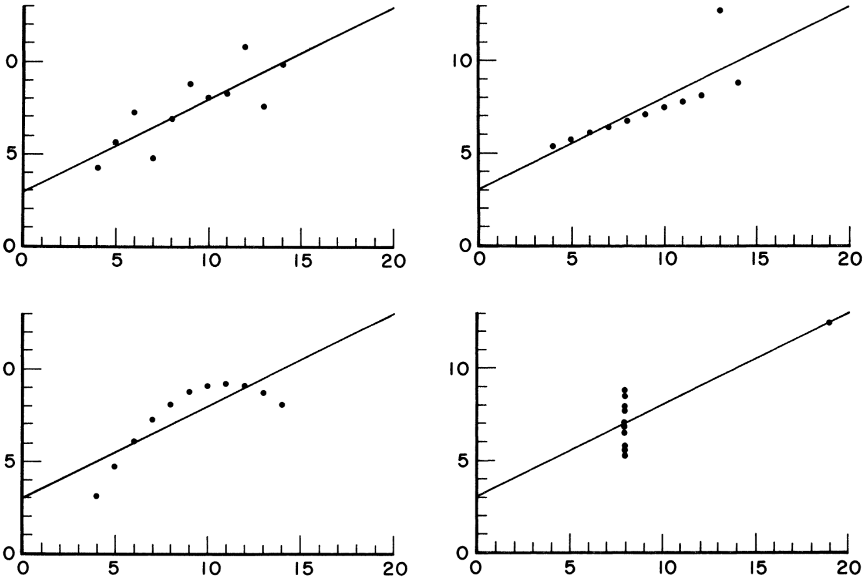
TABLE. Four data sets, each comprising 11 (x, y) pairs.

Each of the four data sets yields the same standard output from a typical regression program, namely

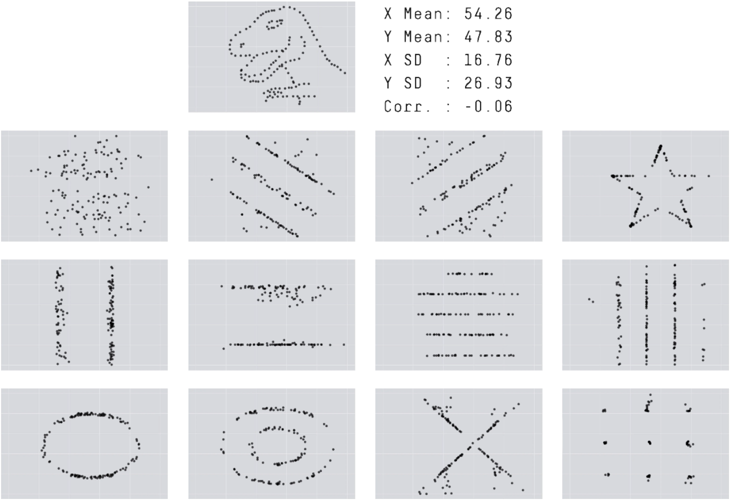
Number of observations $(n) = 11$
Mean of the x's $(\bar{x}) = 9.0$
Mean of the y's $(\bar{y}) = 7.5$
Regression coefficient (b_1) of y on $x = 0.5$
Equation of regression line: $y = 3 + 0.5\,x$
Sum of squares of $x - \bar{x} = 110.0$
Regression sum of squares $= 27.50$ (1 d.f.)
Residual sum of squares of $y = 13.75$ (9 d.f.)
Estimated standard error of $b_1 = 0.118$
Multiple $R^2 = 0.667$

https://www.jstor.org/stable/2682899

1a) Kwartet Anscombe'a

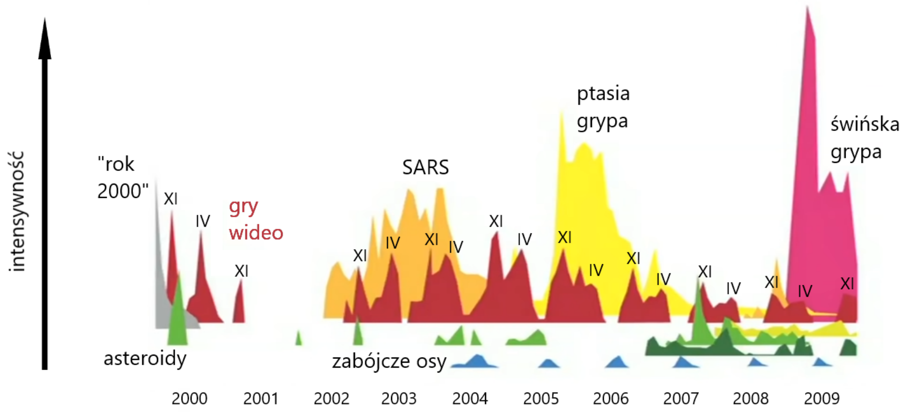


1b) DataSaurus



https://www.research.autodesk.com/publications/same-stats-different-graphs/

2) A timeline of global media panic



https://www.youtube.com/watch?v=5Zg-C8AAIGg

Historia DataViz

https://media.nature.com/original/magazine-assets/d41586-021-01429-6/d41586-021-01429-6.pdf

60 000 pne.

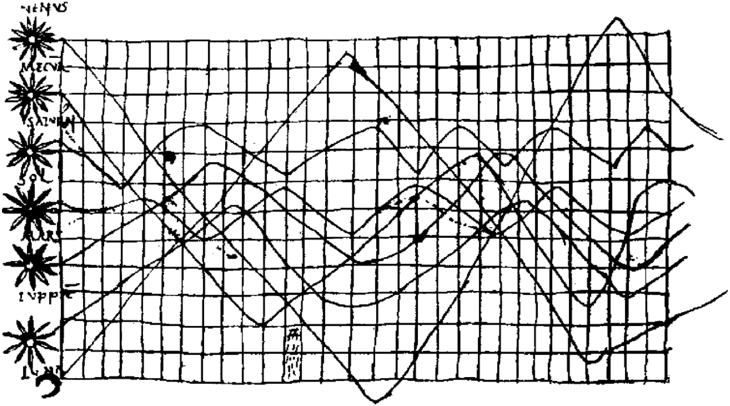
https://media.nature.com/original/magazine-assets/d41586-021-01429-6/d41586-021-01429-6.pdf

40 000 pne.

https://blog.panoply.io/history-of-data-visualization

Inklinacje ciał niebieskich

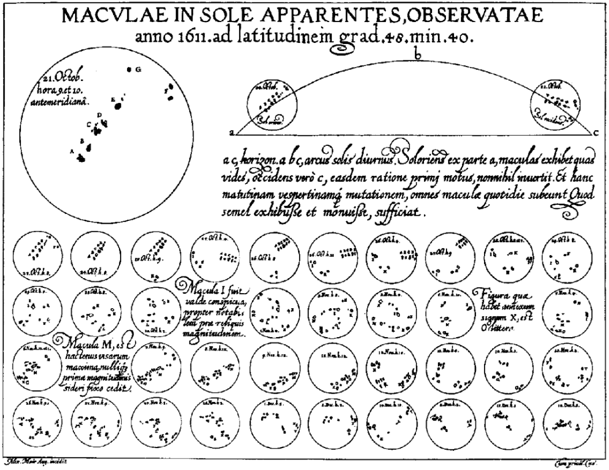
950 ne.



https://thinkinsights.net/data/data-visualization-history/

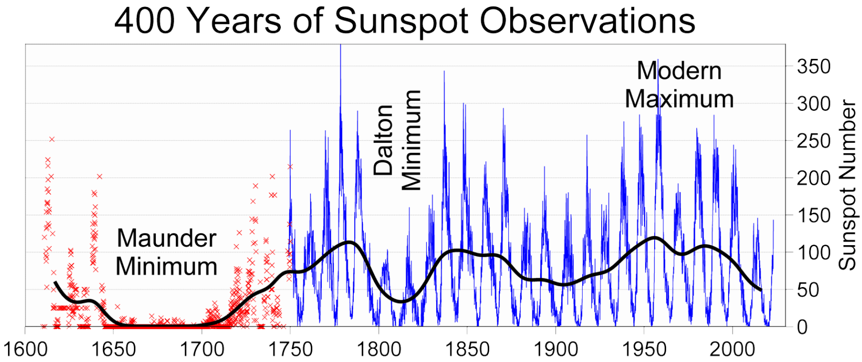
Plamy słoneczne

Scheiner, 1611



Chen, Härdle, Utwin (Eds.), „Handbook of Data Visualization”, Springer (2008)

Najdłuższy zapis astronomiczny

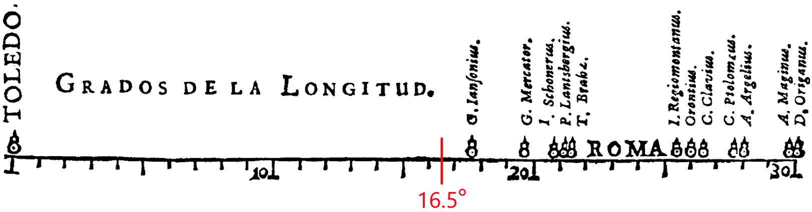


https://en.wikipedia.org/wiki/Sunspot

Odległość Toledo-Rzym

jako różnica długości geograficznych

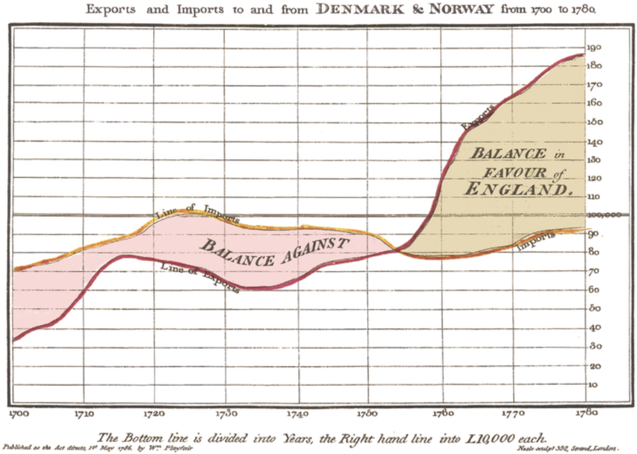
Van Langren, 1644



Równie dobrze można było przedstawić tabelę wartości, ale wykres jest bardziej obrazowy.

https://thinkinsights.net/data/data-visualization-history/

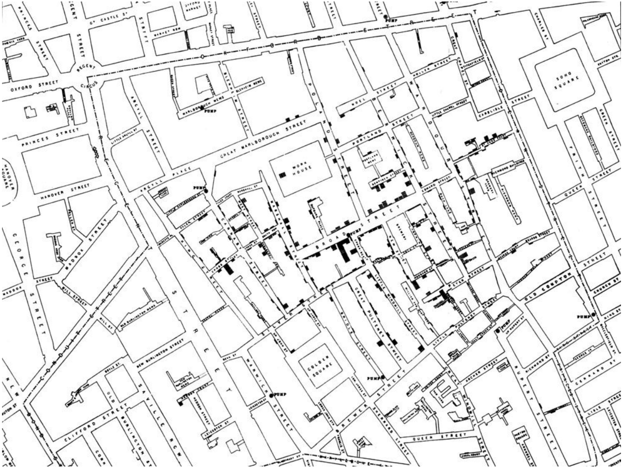
Playfair, 1821



https://thinkinsights.net/data/data-visualization-history/

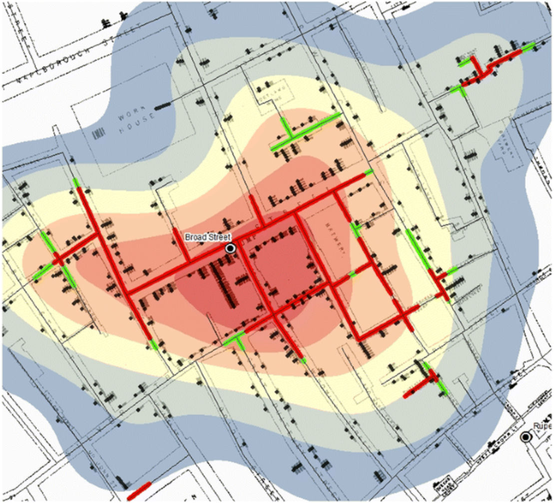
XIX-wieczne epidemie cholery

John Snow, 1854



https://ij-healthgeographics.biomedcentral.com/articles/10.1186/s12942-015-0011-y

John Snow, 1854



https://ij-healthgeographics.biomedcentral.com/articles/10.1186/s12942-015-0011-y

Naczelna zasada DataViz

Wykres ma być czytelny.

Naczelna zasada DataViz

Wykres ma być czytelny.

Cel kursu

Żeby było jak najmniej takich wykresów:

Naczelna zasada DataViz

Wykres ma być czytelny.

Cel kursu

Żeby było jak najmniej takich wykresów:

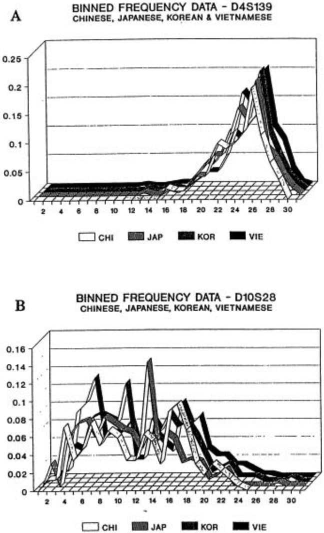
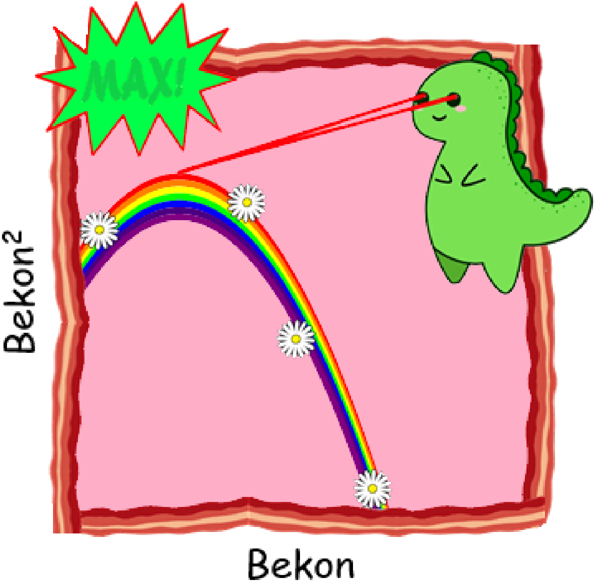


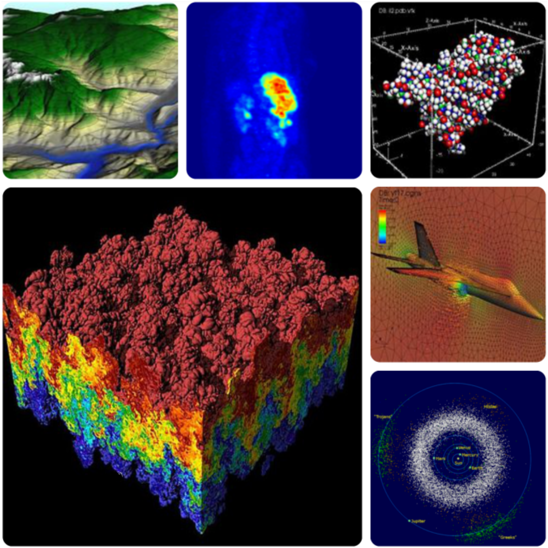
FIG. 4. *Fixed bin distribution (histogram) for two loci and four Asian subpopulations (used with permission from John Hartmann): the boundaries of the 30 bins (vertical axis) are determined by the FBI; these bins are not of equal length. Sample sizes (numbers of individuals) for Chinese, Japanese, Korean and Vietnamese are* 103, 125, 93 *and* 215 *for* D4S139 *and* 120, 137, 100 *and* 193 *for* D10S28. *The horizontal axis is the bin number; bins are not of equal length.*

www.jstor.org/stable/2246327

Walory estetyczne/artystyczne są subiektywne...

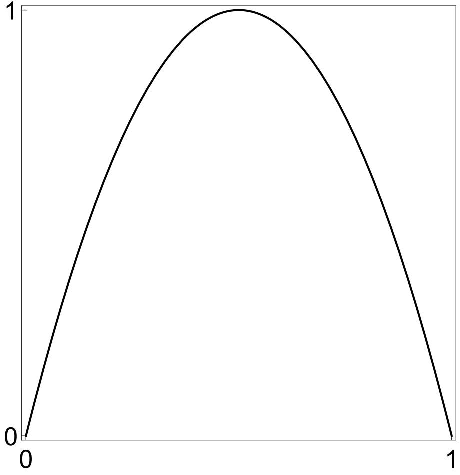


...dlatego należy dążyć do tego:



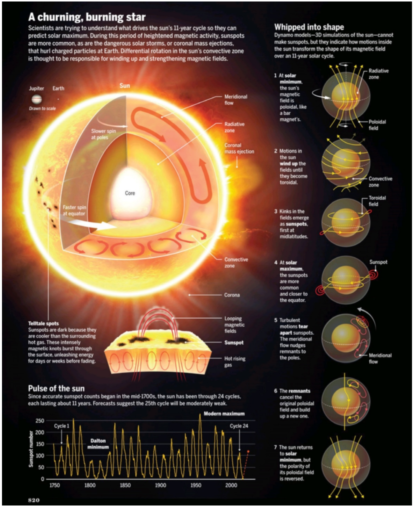
https://en.wikipedia.org/wiki/Scientific_visualization

...ale to też jest w porządku*:



*jeśli spełnia swój cel

...zaś to* wykracza poza ramy niniejszego kursu



*Tj. infografiki.
Building science graphics, Jen Christiansen