Foundations of Data Science for Biologists

Best practices: data visualization

BIO 724D

18-MAR-2024

Instructors: Greg Wray and Paul Magwene

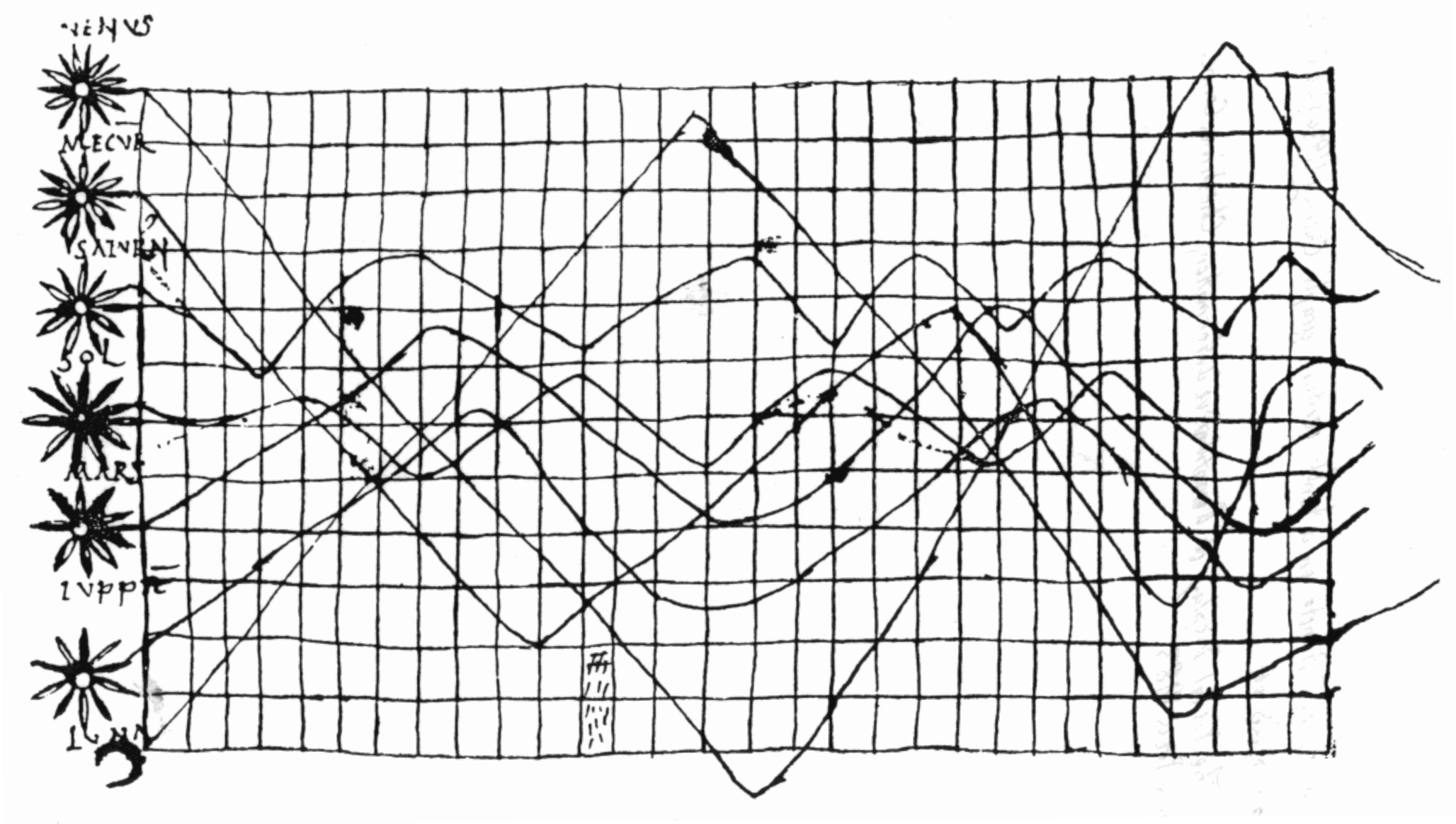
Examples of graphical excellence

The earliest known abstract graphic? (right)

Chauvet-Pont-d’Arc Cave, France; unknown artists; 32,000 - 30,000 BCE

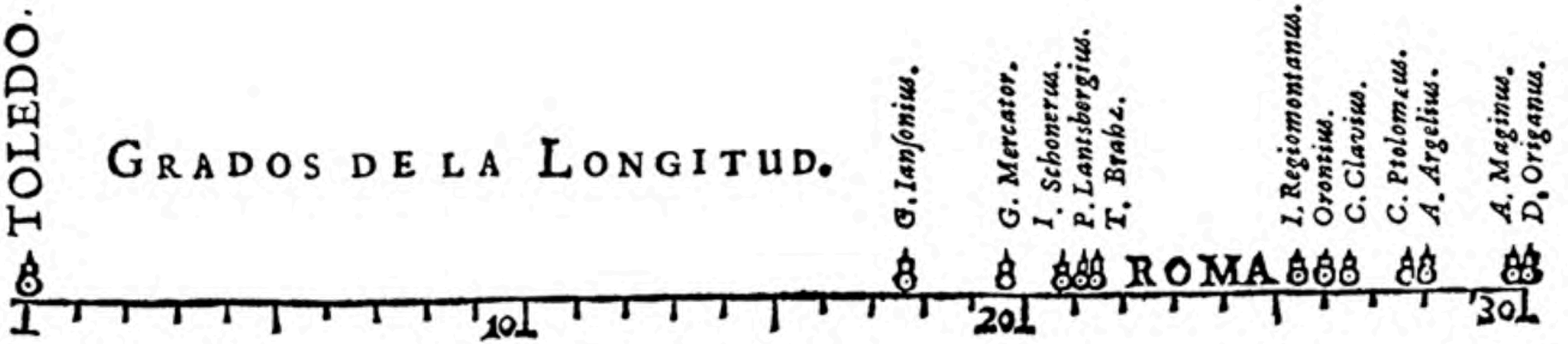
Earliest known “world map”

Babylon; unknown designer; early 6th century BCE



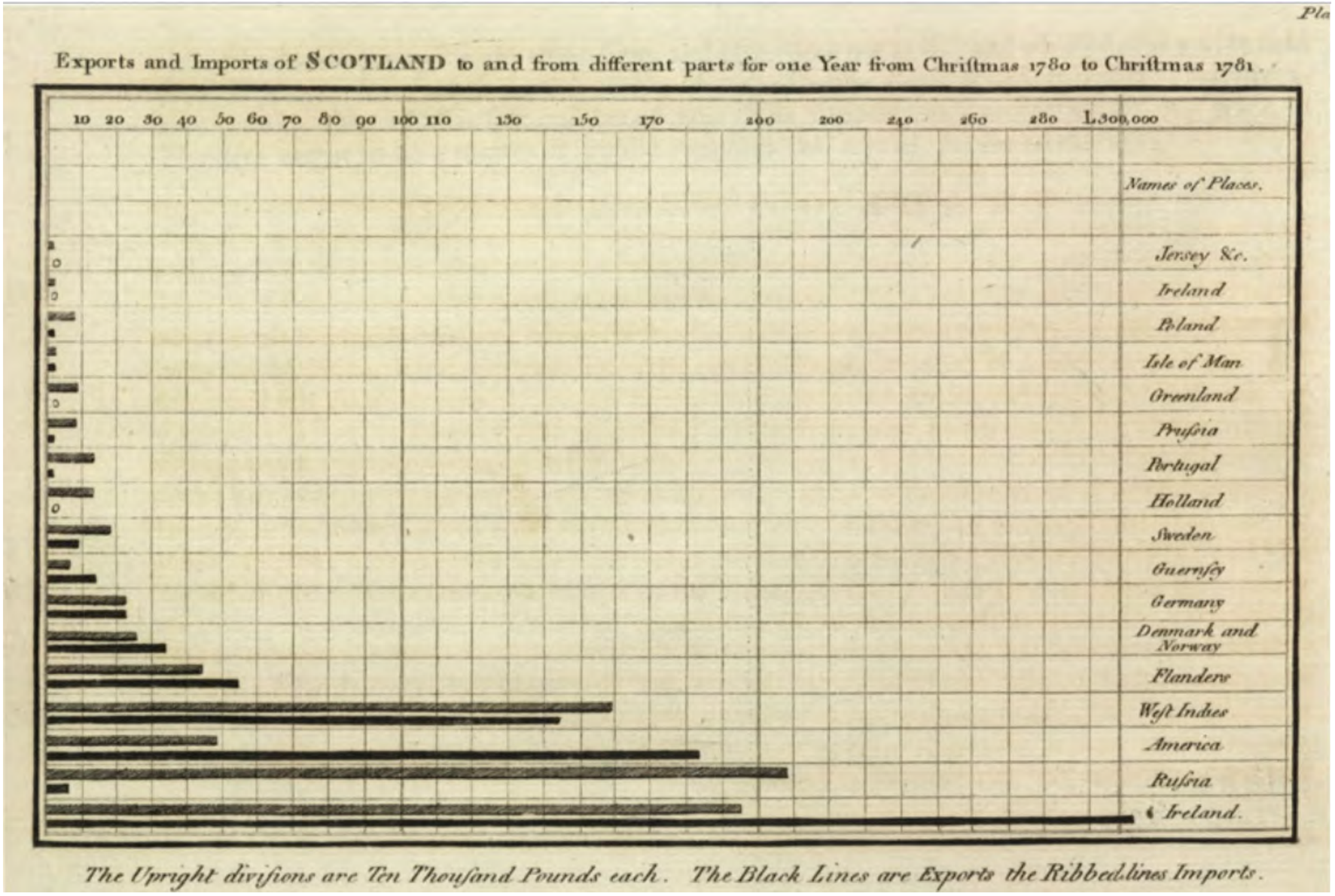
Earliest known “graph” — annotated as *De cursu per zodiacum*

Unknown designer; early 11th century



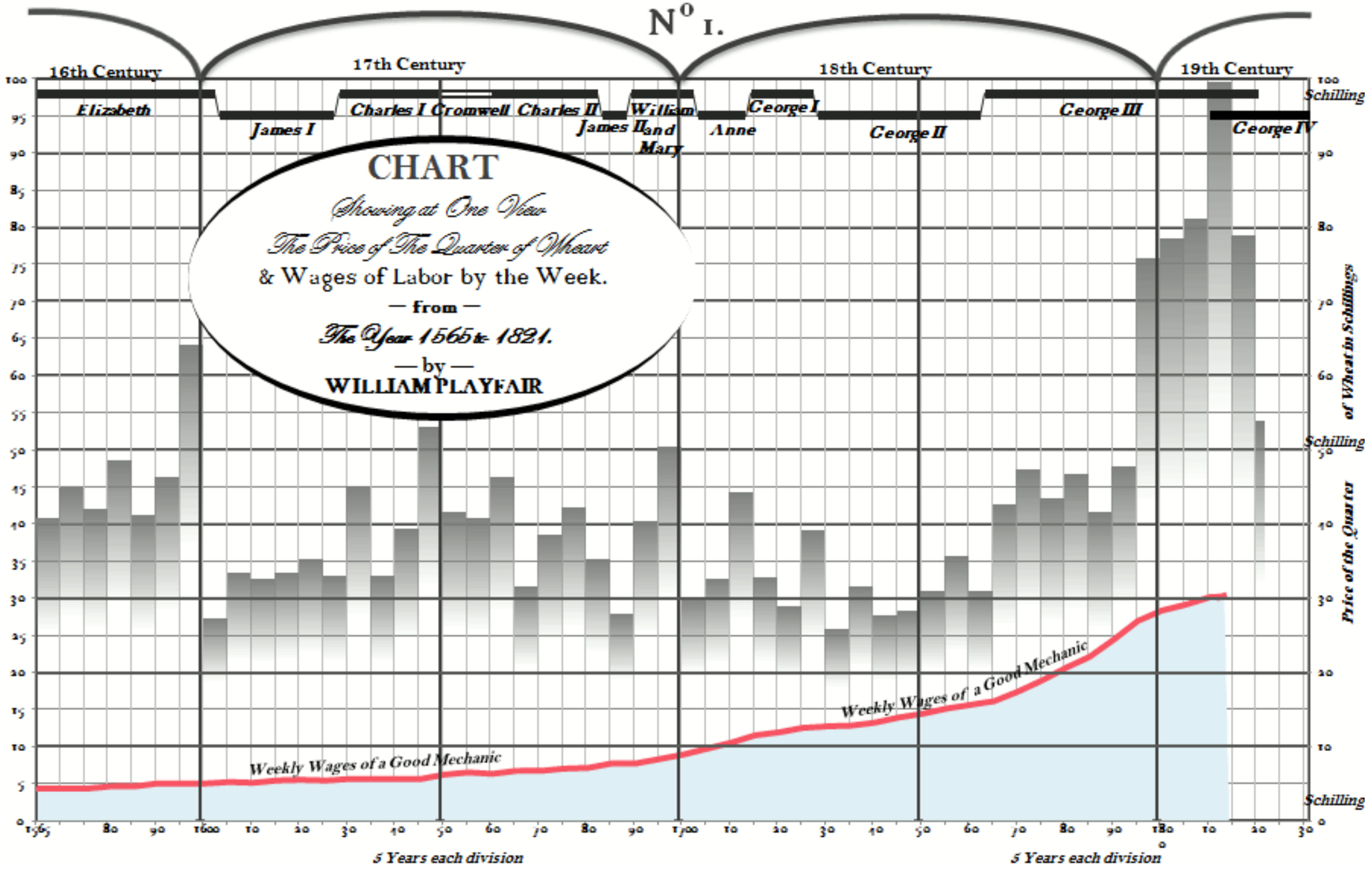
Earliest known statistical graphic

Michael Florent van Langren; 1644 (two earlier manuscript versions are known)



The first bar graph
William Playfair; 1781

(he also invented
the pie chart)

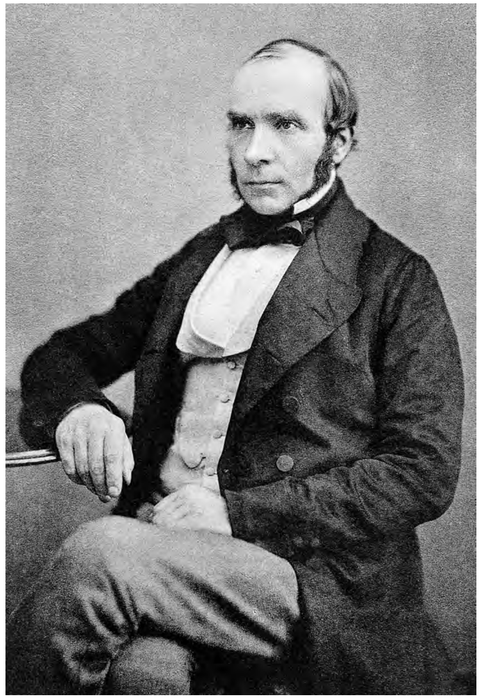


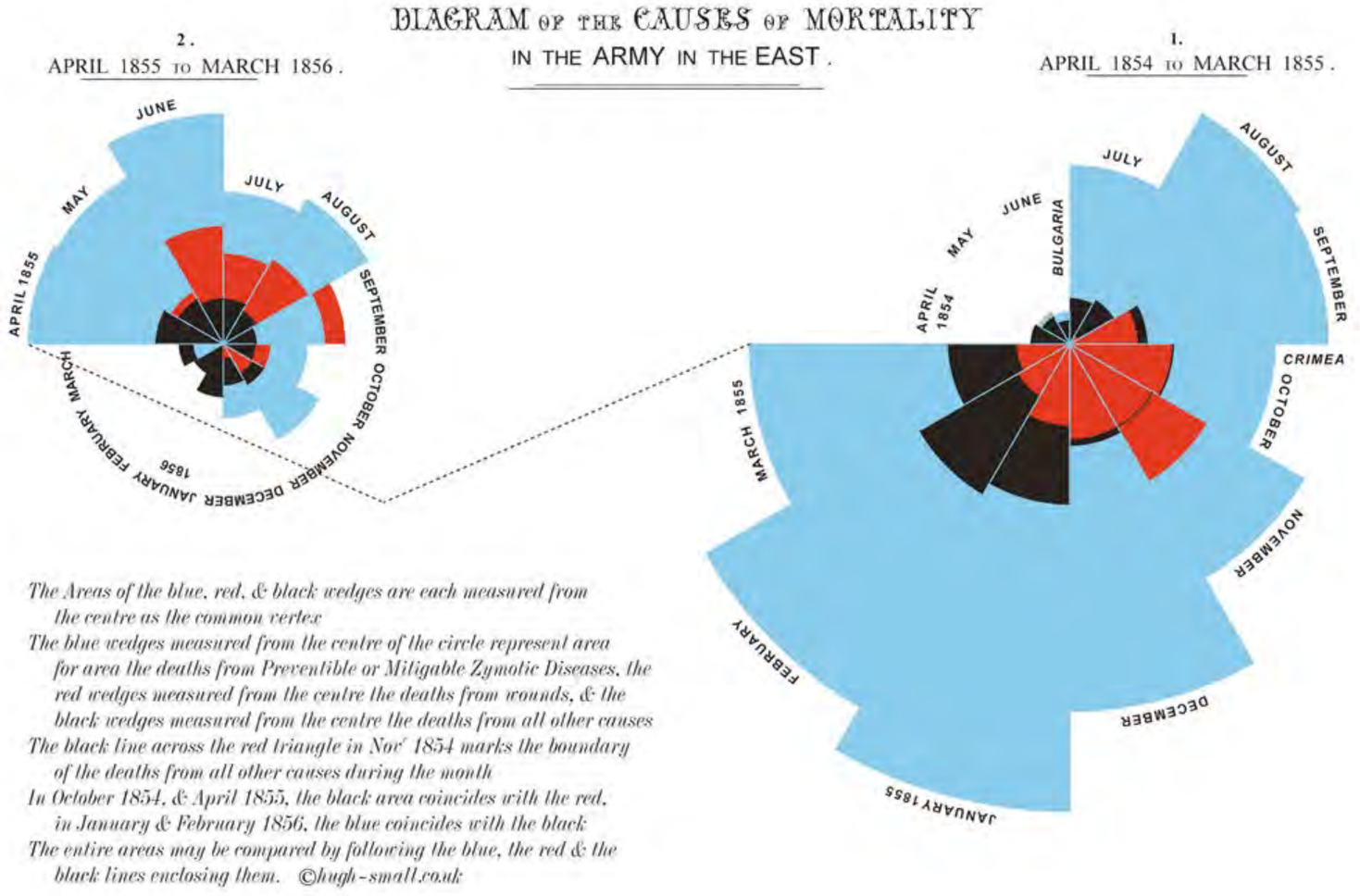
The invention of time-series graphs (also multivariate!) by Joseph Priestly and William Playfair

William Playfair; 1821

The birth of epidemiology

John Snow; 1854



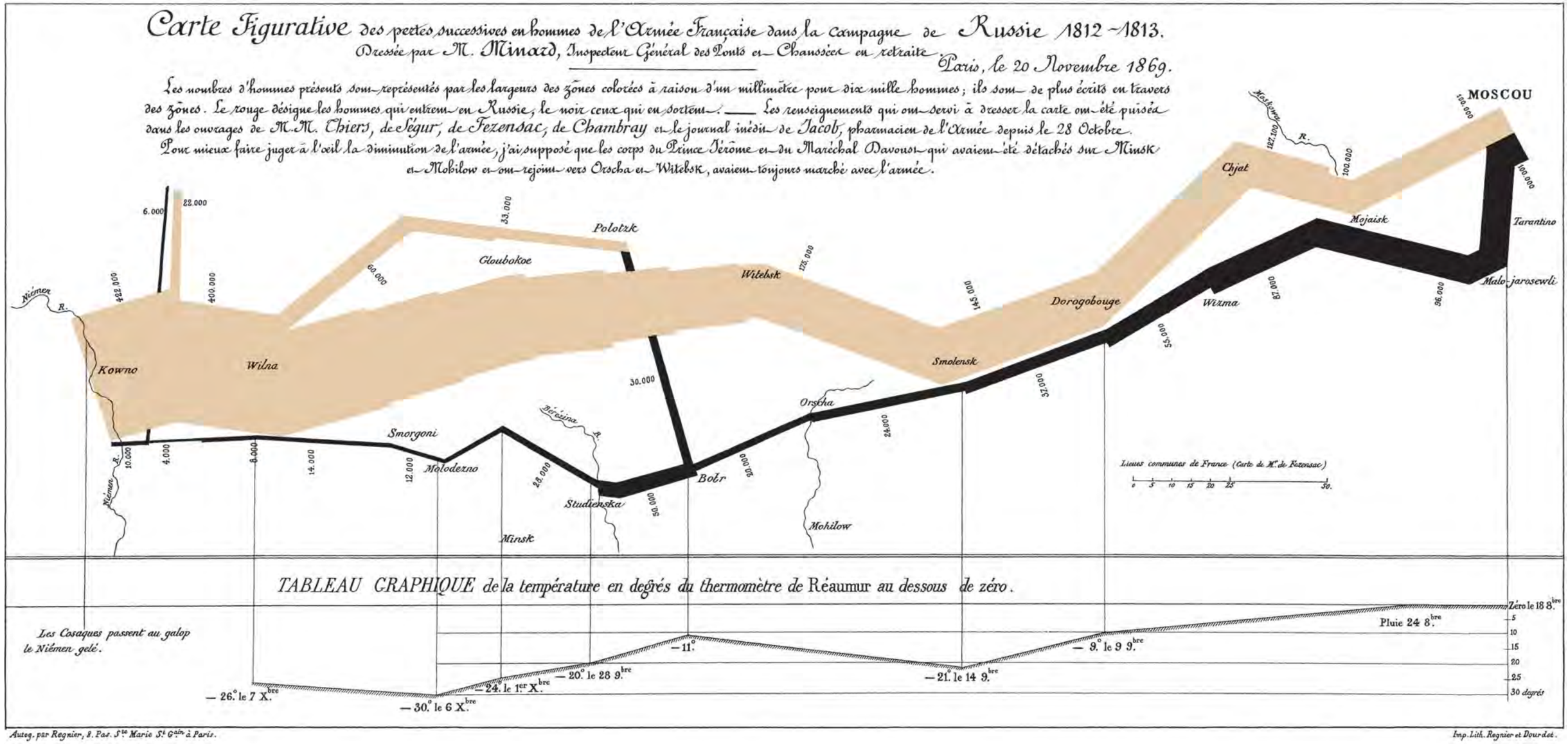


Invented the polar area diagram

Color and proportion

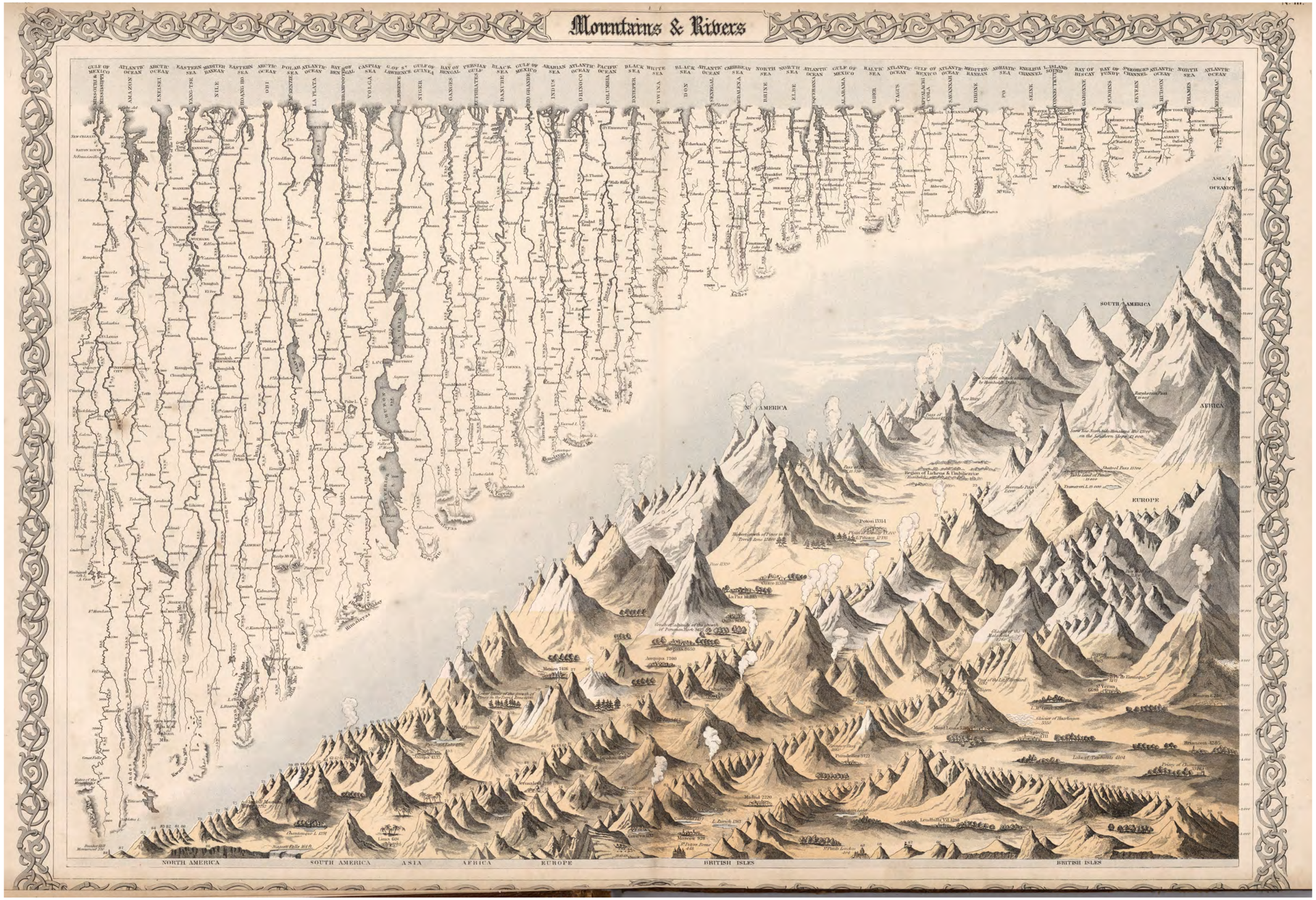
Policy reform and social justice through graphics

Florence Nightingale; 1858



Telling a story with quantitative data: the first infographic?

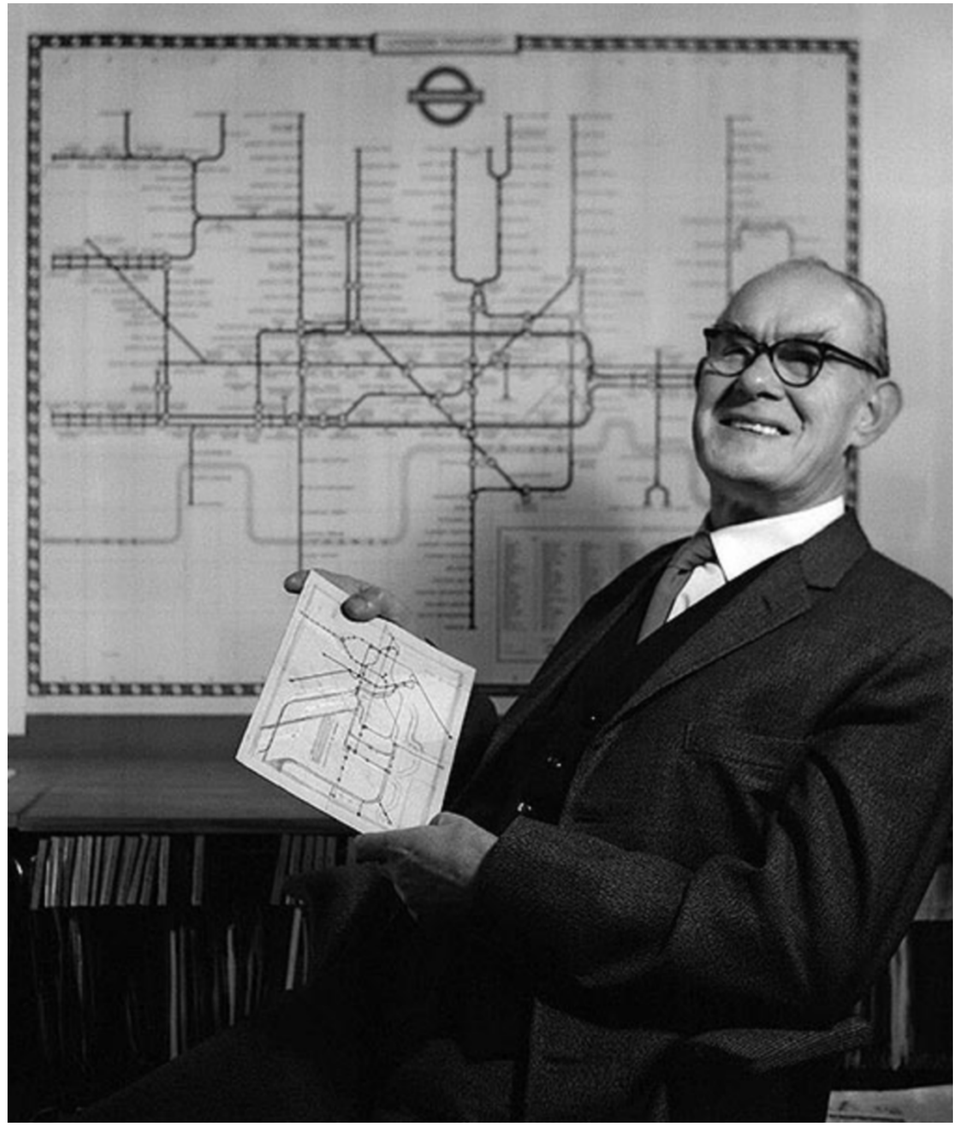
Charles Minard; 1869



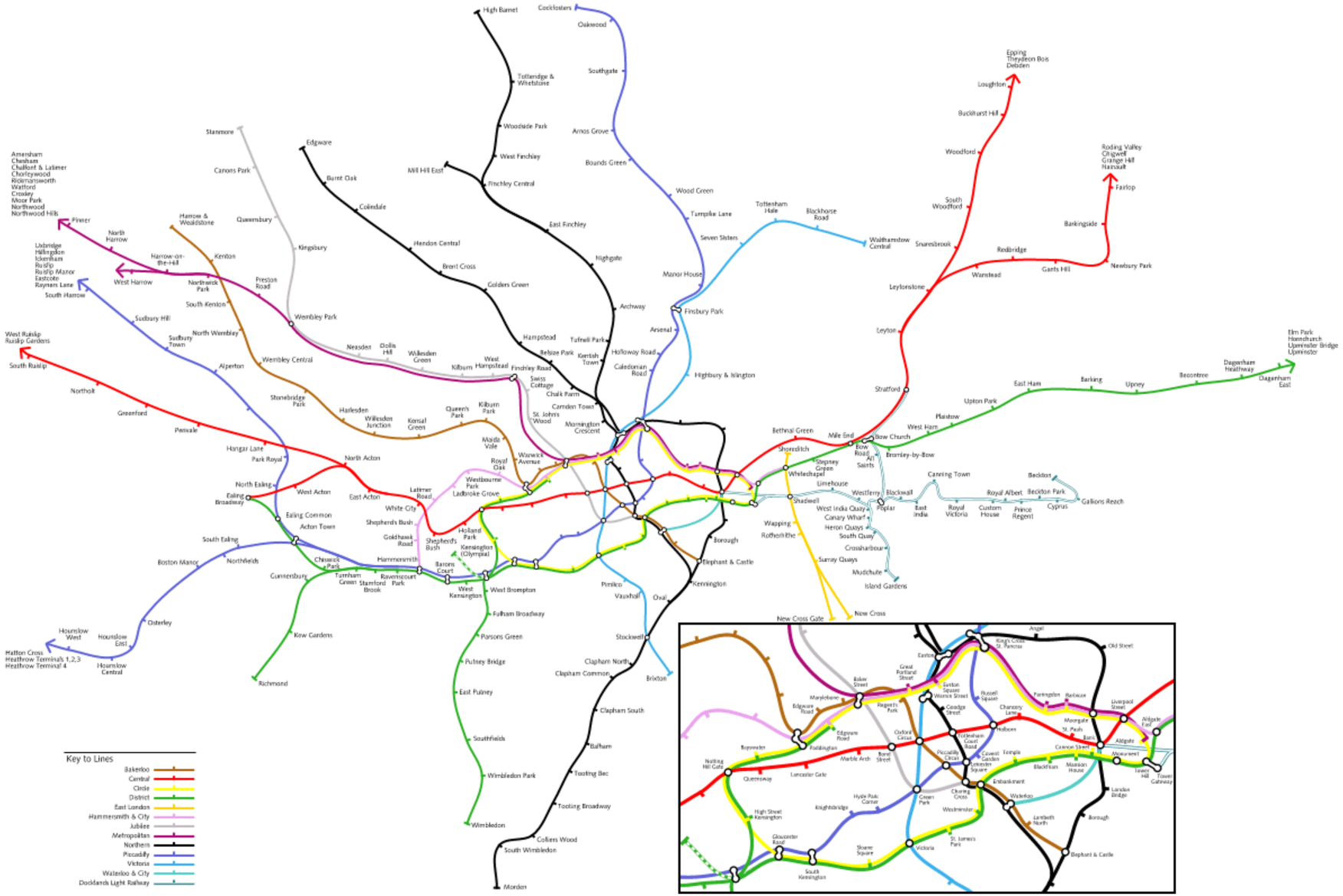
Enriching a univariate measure

Joseph Hutchins Colton; mid 1800s

Practical, intelligent simplification



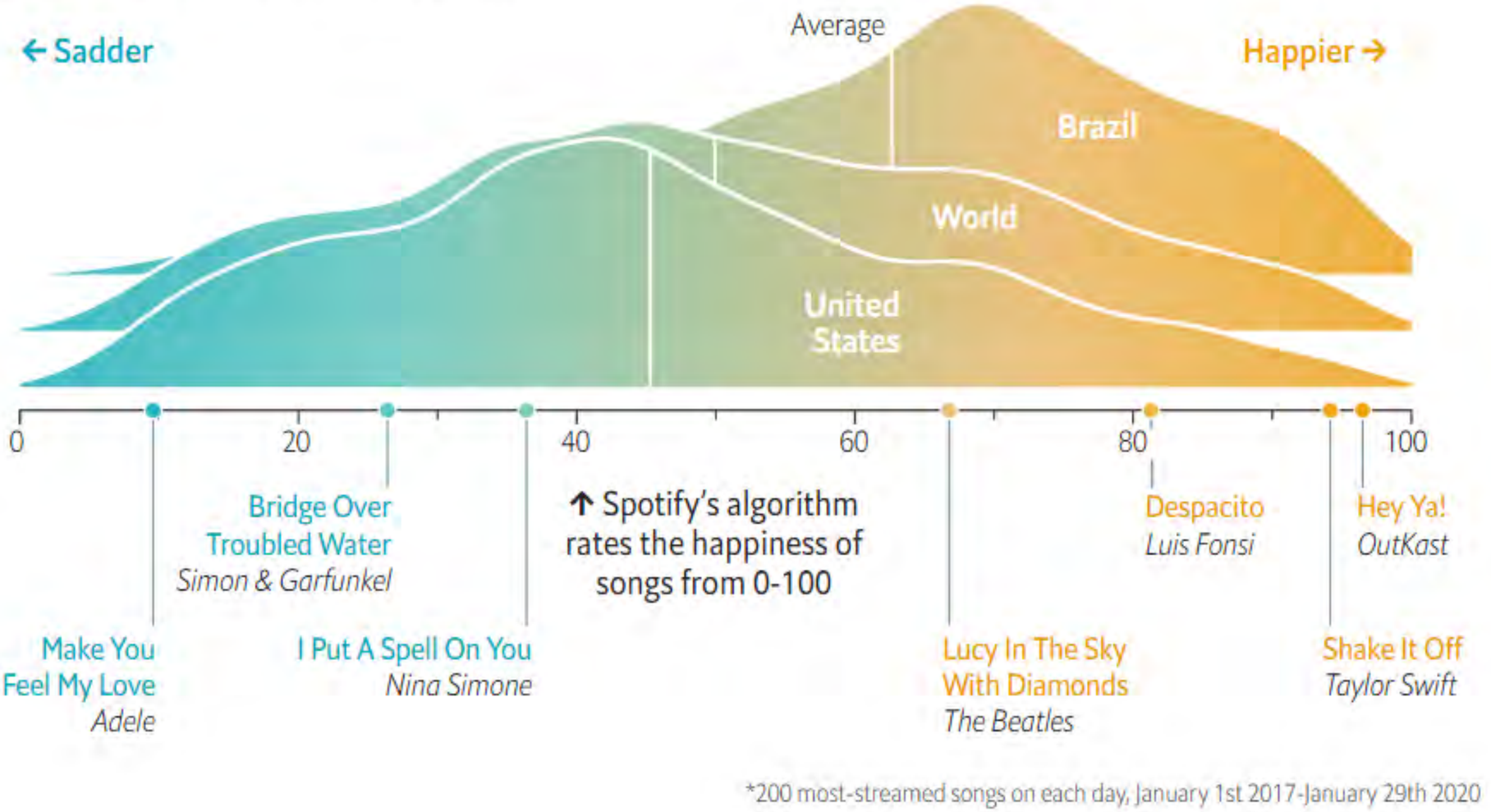
Henry “Harry” Beck; 1933



Geographically accurate London Underground map: not as easy to understand!

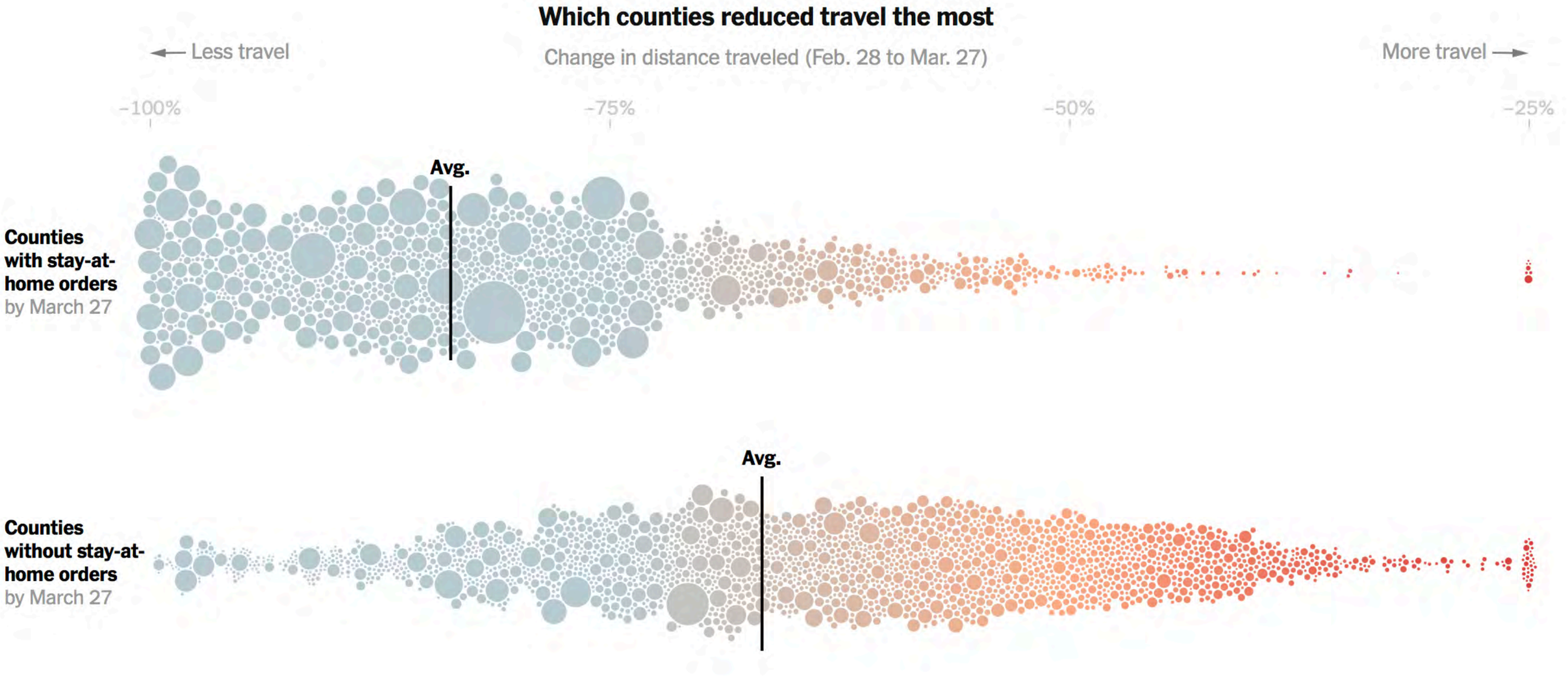
→ Some countries listen to happier music than others

Distribution of tracks streamed*, by mood

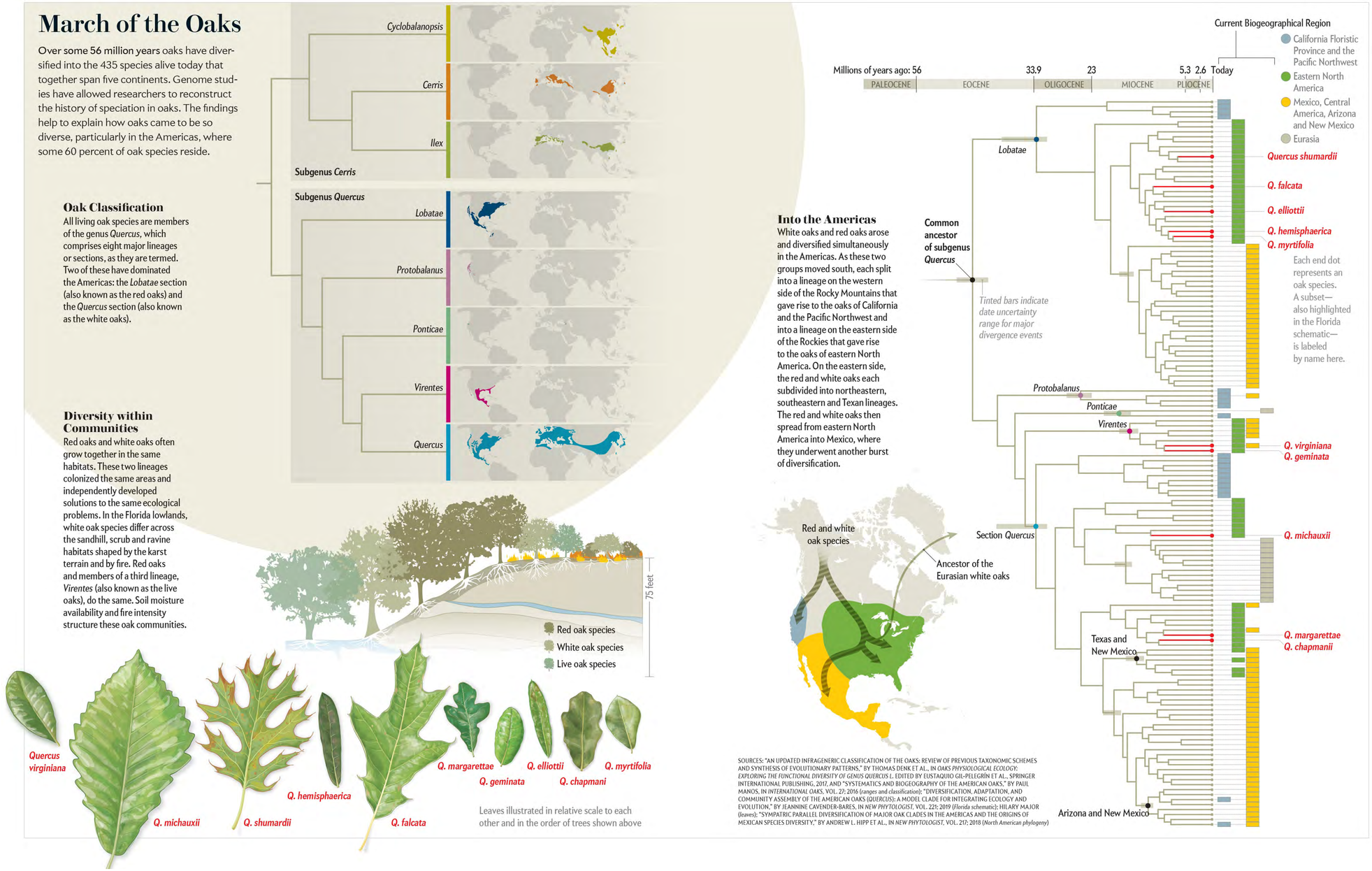


*200 most-streamed songs on each day, January 1st 2017-January 29th 2020

Economist; 2020



New York Times; 2020



Infographic based on research done at Duke

Paul Manos; Biology Dept; 2020

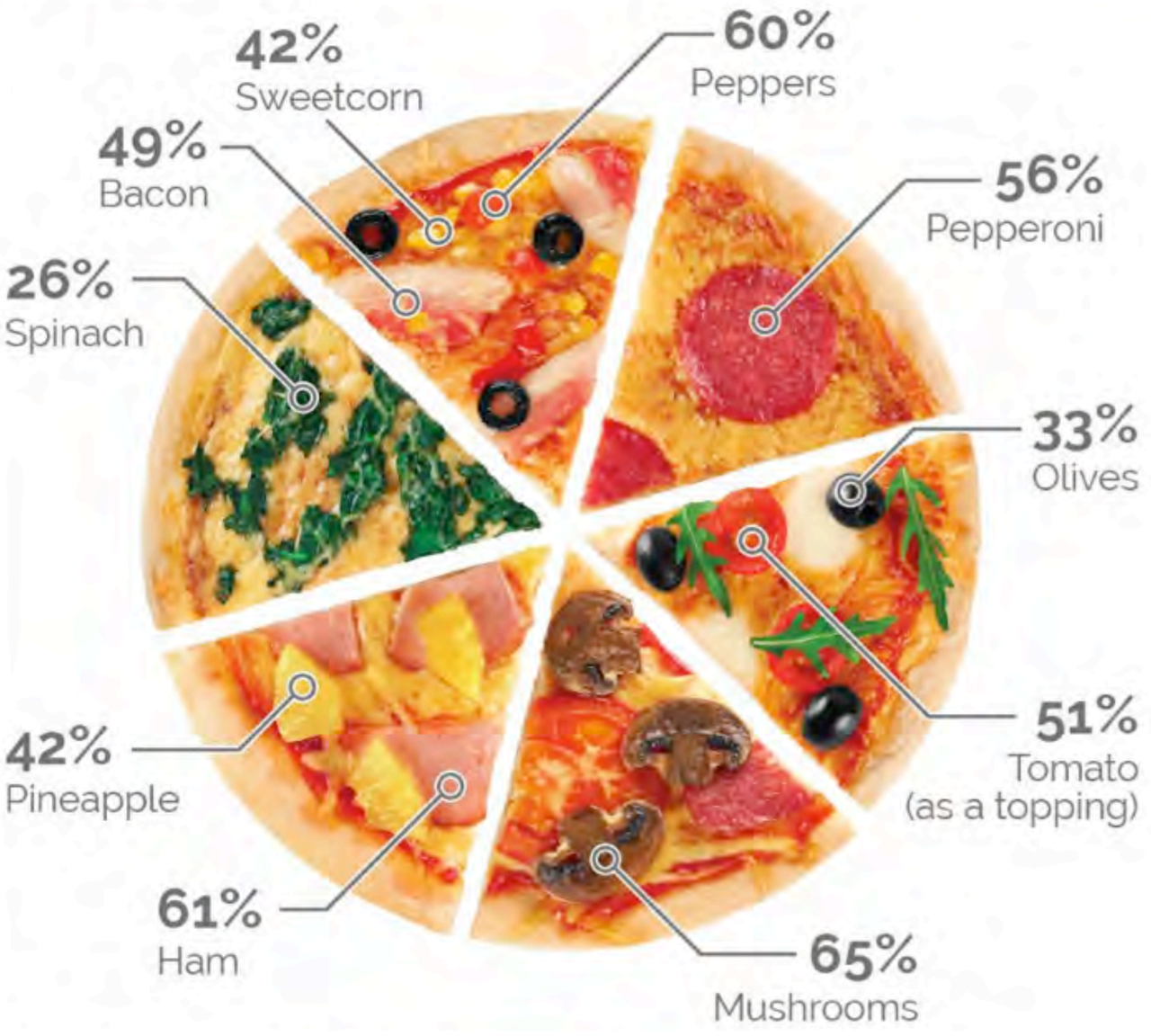
Graphical excellence and counter-examples

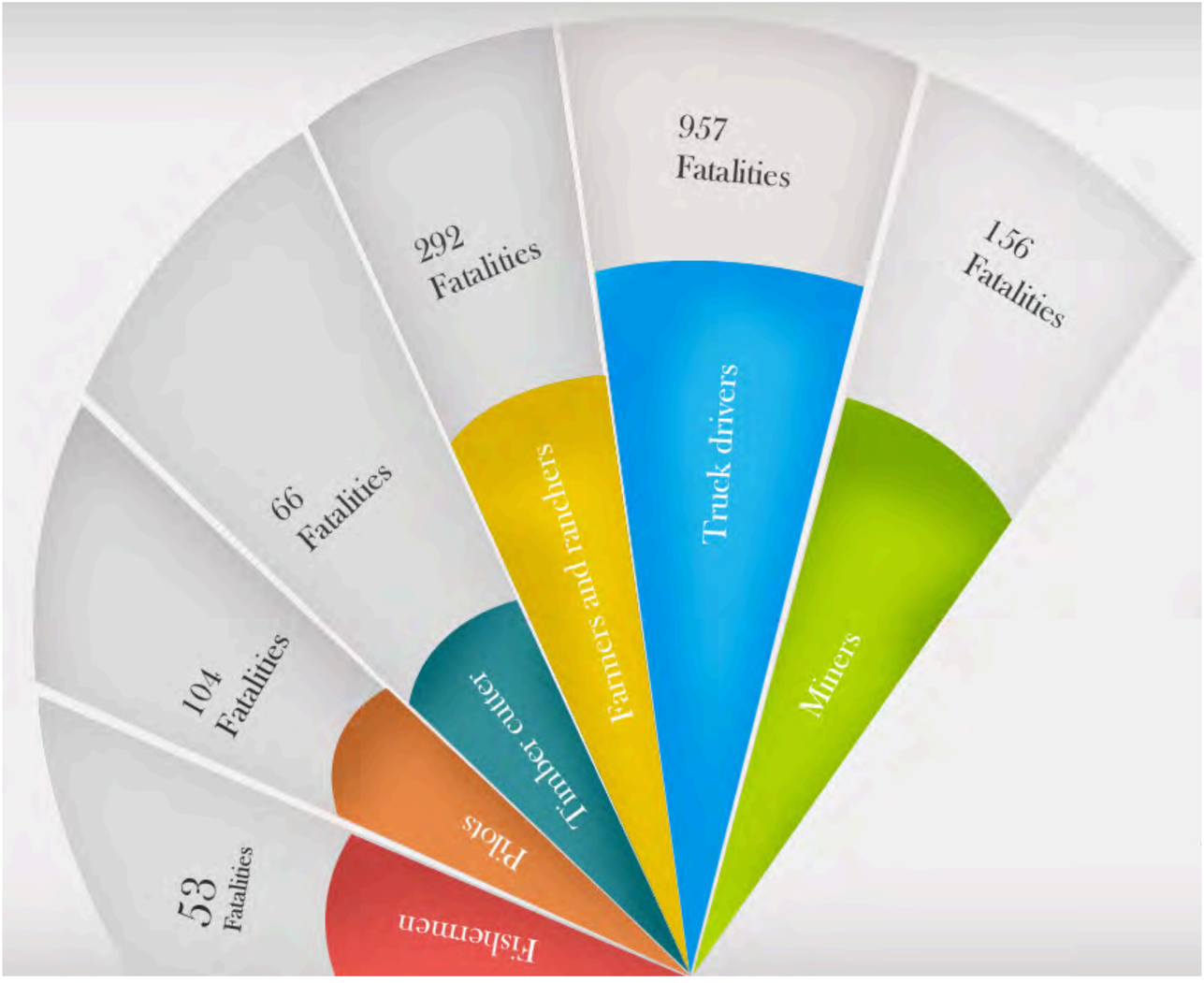
Tufte on graphical excellence

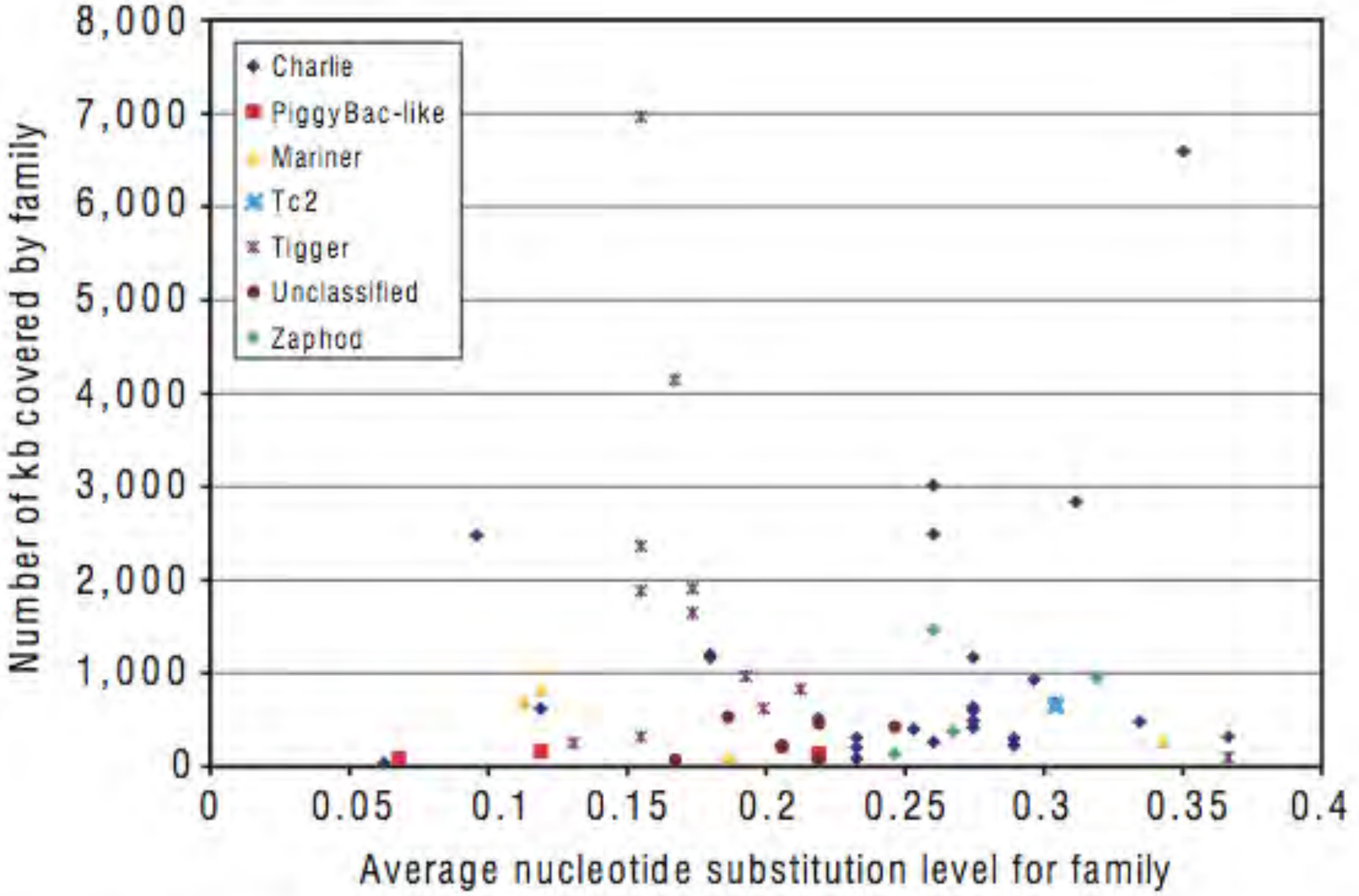
"Excellence in statistical graphics consists of complex ideas communicated with clarity, precision, and efficiency."

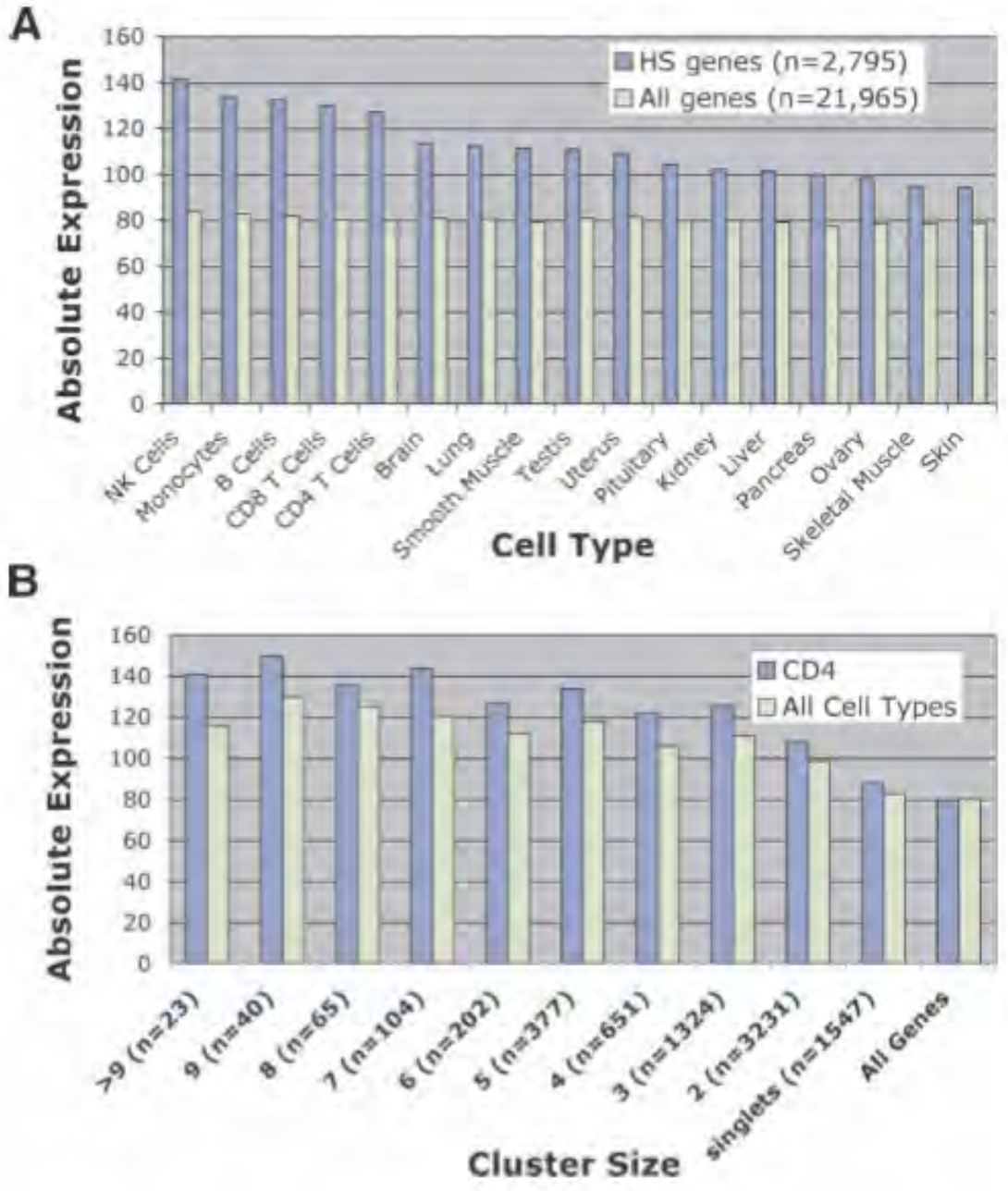
"a matter of *substance*, of *statistics*, and of *design*."

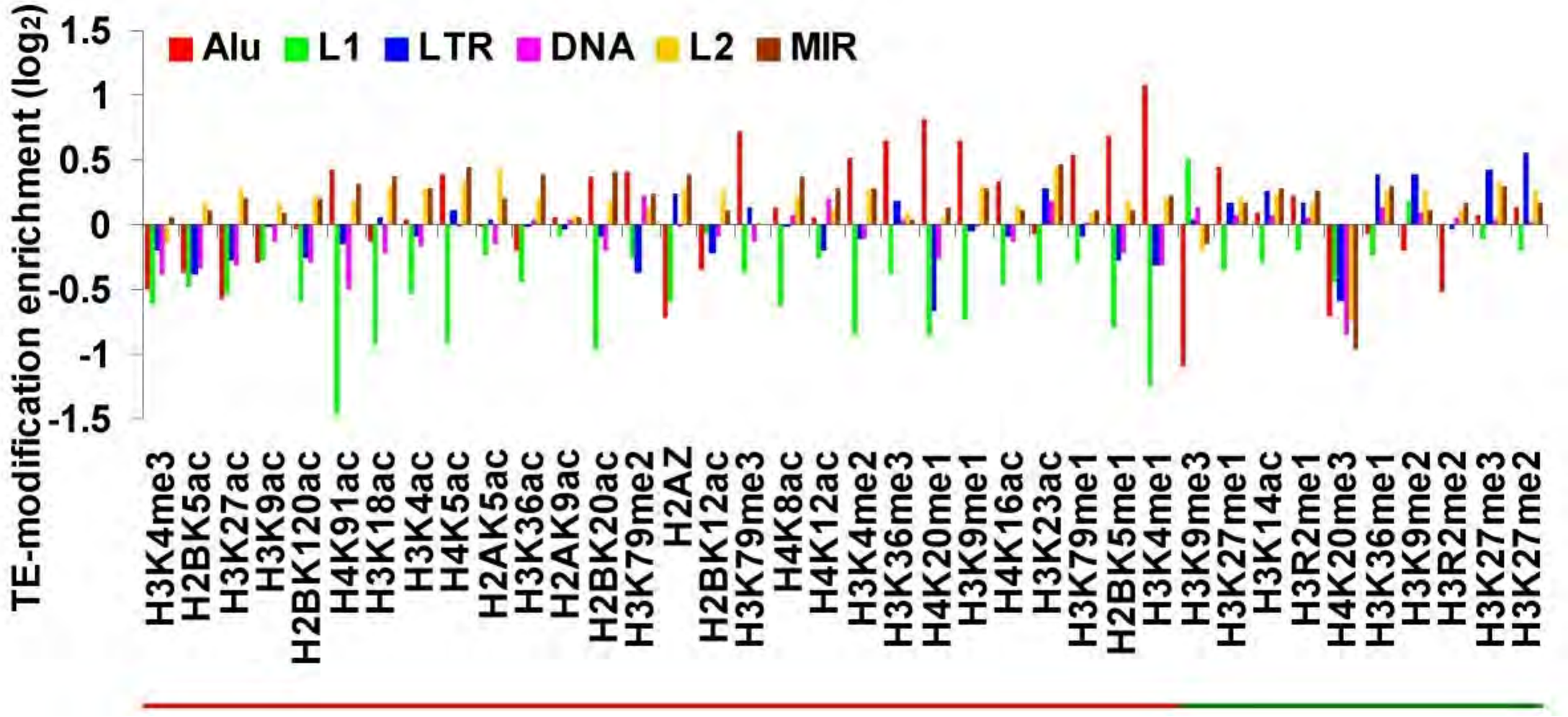
Edward Tufte, "The Visual Display of Quantitative Information"

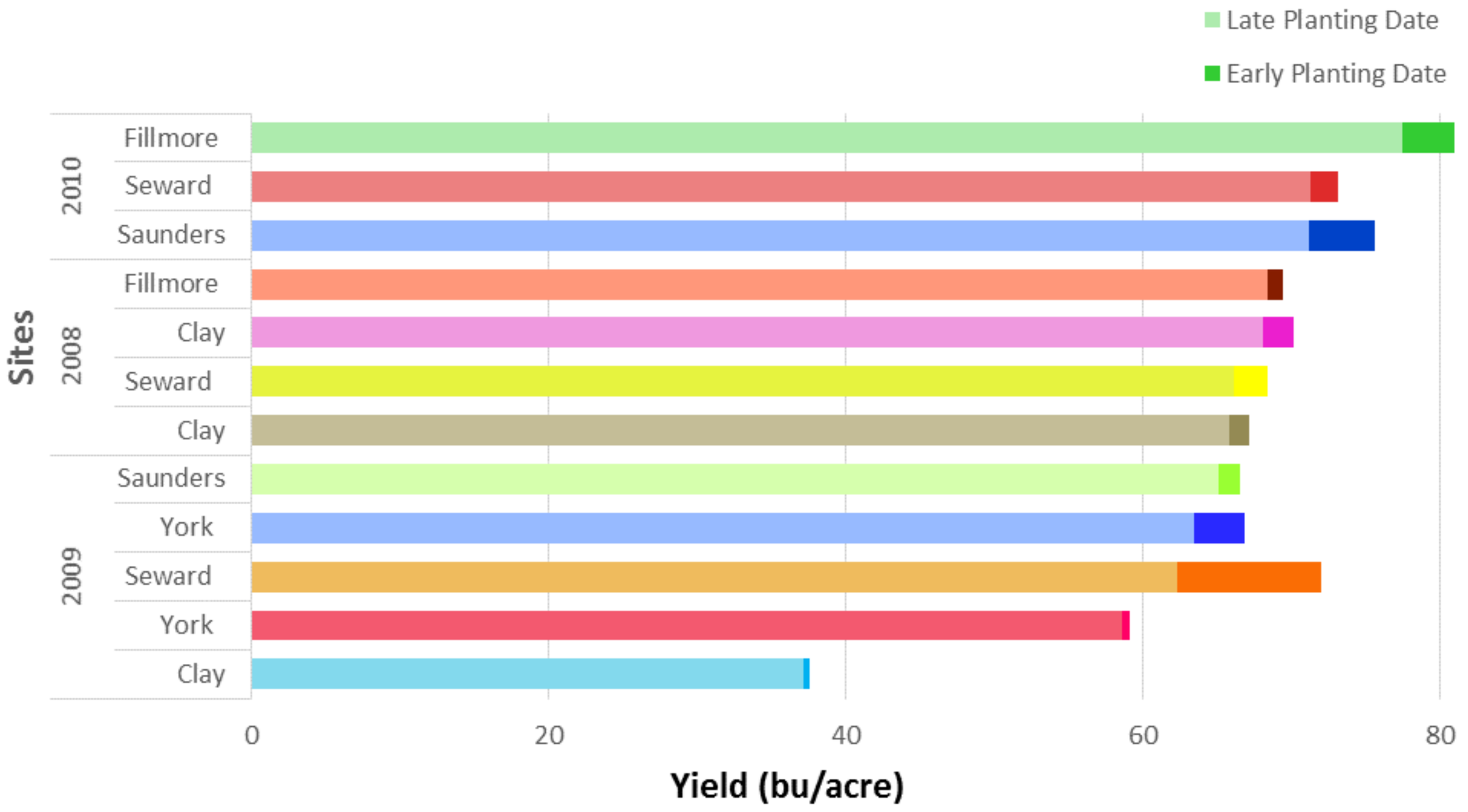

42%
Sweetcorn
60%
Peppers
49%
Bacon
56%
Pepperoni
26%
Spinach
33%
Olives
51%
Tomato
(as a topping)
42%
Pineapple
61%
Ham
65%
Mushrooms

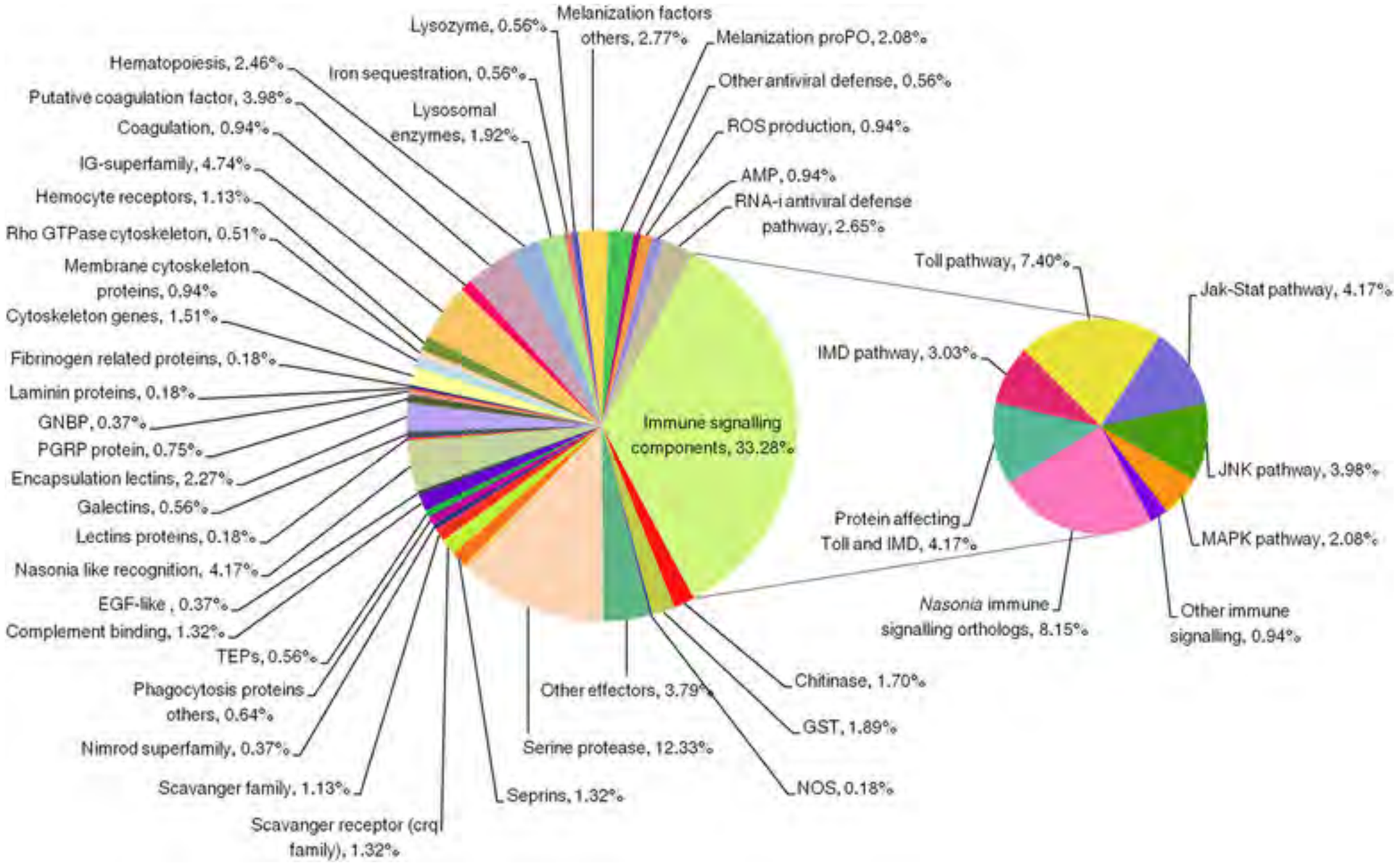

957
Fatalities
292
Fatalities
66
Fatalities
104
Fatalities
53
Fatalities
156
Fatalities
Truck drivers
Farmers and ranchers
Timber cutter
Pilots
Fishermen
Miners


8,000
7,000
6,000
5,000
4,000
3,000
2,000
1,000
0
Number of kb covered by family
Charlie
PiggyBac-like
Mariner
Tc2
Tigger
Unclassified
Zaphod
0
0.05
0.1
0.15
0.2
0.25
0.3
0.35
0.4
Average nucleotide substitution level for family


A
Absolute Expression
160
140
120
100
80
60
40
20
0
HS genes (n=2,795)
All genes (n=21,965)
NK Cells
Monocytes
B Cells
CD8 T Cells
CD4 T Cells
Brain
Lung
Smooth Muscle
Testis
Uterus
Pituitary
Kidney
Liver
Pancreas
Ovary
Skeletal Muscle
Skin
Cell Type
B
Absolute Expression
160
140
120
100
80
60
40
20
0
CD4
All Cell Types
>9 (n=23)
9 (n=40)
8 (n=65)
7 (n=104)
6 (n=202)
5 (n=377)
4 (n=651)
3 (n=1324)
2 (n=3231)
singlets (n=1547)
All Genes
Cluster Size


TE-modification enrichment (log2)
1.5
1
0.5
0
-0.5
-1
-1.5
Alu
L1
LTR
DNA
L2
MIR
H3K4me3
H2BK5ac
H3K27ac
H3K9ac
H2BK120ac
H4K91ac
H3K18ac
H3K4ac
H4K5ac
H2AK5ac
H3K36ac
H2AK9ac
H2BK20ac
H3K79me2
H2AZ
H2BK12ac
H3K79me3
H4K8ac
H4K12ac
H3K4me2
H3K36me3
H4K20me1
H3K9me1
H4K16ac
H3K23ac
H3K79me1
H2BK5me1
H3K4me1
H3K9me3
H3K27me1
H3K14ac
H3R2me1
H4K20me3
H3K36me1
H3K9me2
H3R2me2
H3K27me3
H3K27me2


Late Planting Date
Early Planting Date
Sites
2010
Fillmore
Seward
Saunders
2008
Fillmore
Clay
Seward
Clay
2009
Saunders
York
Seward
York
Clay
0
20
40
60
80
Yield (bu/acre)


Lysozyme, 0.56%
Melanization factors others, 2.77%
Melanization proPO, 2.08%
Hematopoiesis, 2.46%
Iron sequestration, 0.56%
Other antiviral defense, 0.56%
Putative coagulation factor, 3.98%
Lysosomal enzymes, 1.92%
ROS production, 0.94%
Coagulation, 0.94%
AMP, 0.94%
IG-superfamily, 4.74%
RNA-i antiviral defense pathway, 2.65%
Hemocyte receptors, 1.13%
Rho GTPase cytoskeleton, 0.51%
Membrane cytoskeleton proteins, 0.94%
Cytoskeleton genes, 1.51%
Fibrinogen related proteins, 0.18%
Laminin proteins, 0.18%
GNBP, 0.37%
PGRP protein, 0.75%
Encapsulation lectins, 2.27%
Galectins, 0.56%
Lectins proteins, 0.18%
Nasonia like recognition, 4.17%
EGF-like , 0.37%
Complement binding, 1.32%
TEPs, 0.56%
Phagocytosis proteins others, 0.64%
Nimrod superfamily, 0.37%
Scavanger family, 1.13%
Scavanger receptor (crq family), 1.32%
Seprins, 1.32%
Serine protease, 12.33%
Other effectors, 3.79%
Immune signalling components, 33.28%
NOS, 0.18%
GST, 1.89%
Chitinase, 1.70%
Toll pathway, 7.40%
Jak-Stat pathway, 4.17%
IMD pathway, 3.03%
JNK pathway, 3.98%
Protein affecting Toll and IMD, 4.17%
MAPK pathway, 2.08%
Nasonia immune signalling orthologs, 8.15%
Other immune signalling, 0.94%

Goals of graphical excellence

A graphic should convey information:

- Accurately: no distortions, distractions, or critical omissions in the data presented
- Clearly: the focus is on the data
- Efficiently: with minimal need for explanation
- Revealingly: uncovers trends and organization in data; ideally leads to insight
- Engagingly: invites the viewer to think about what results mean

And it should fit into the overall presentation:

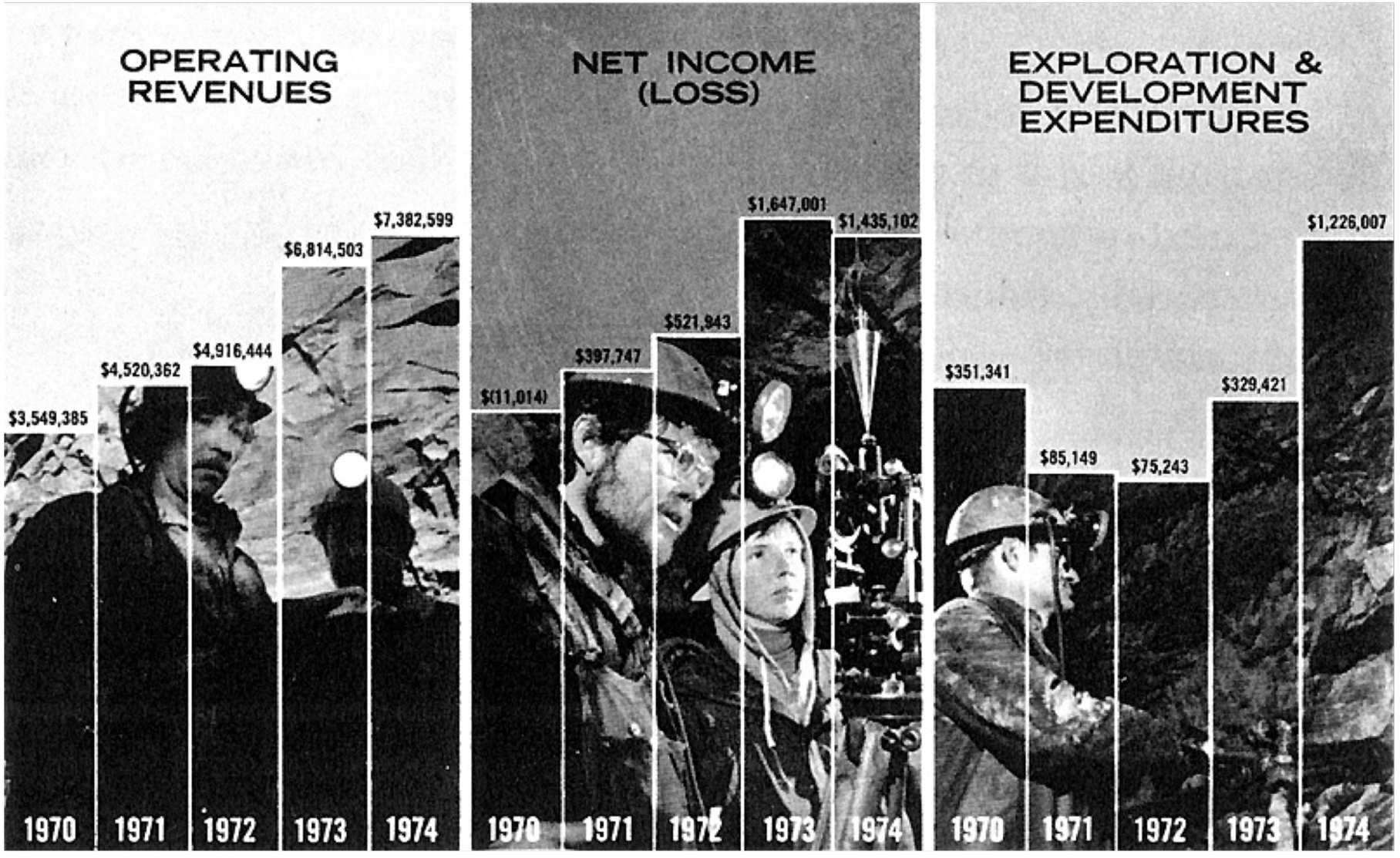
- Consistently: orientation, scale, color, font, encodings, expected location, etc.
- Relationally: connects to the overall narrative and to other visuals

Graphical integrity and counter-examples

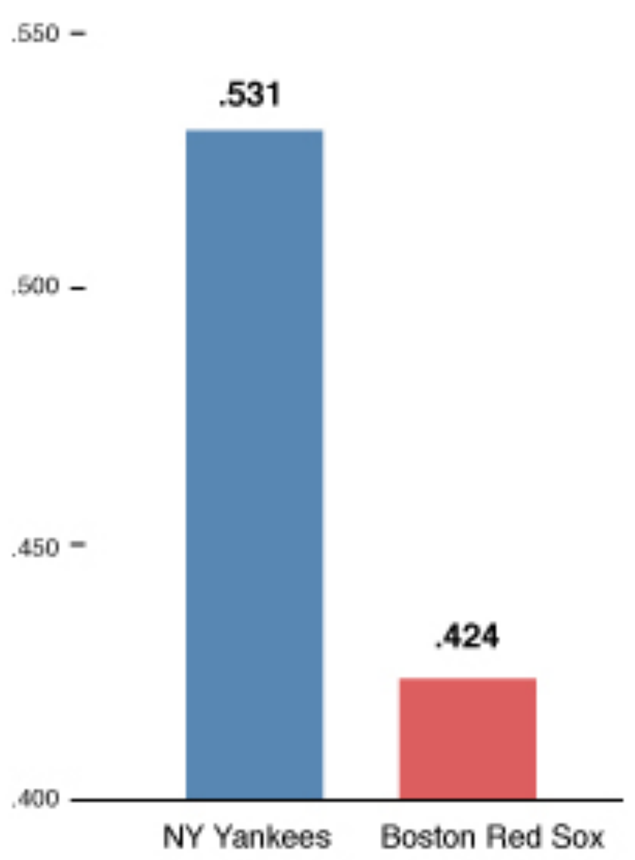
Tufte on graphical integrity

"Graphical integrity begins with telling the truth about the data"

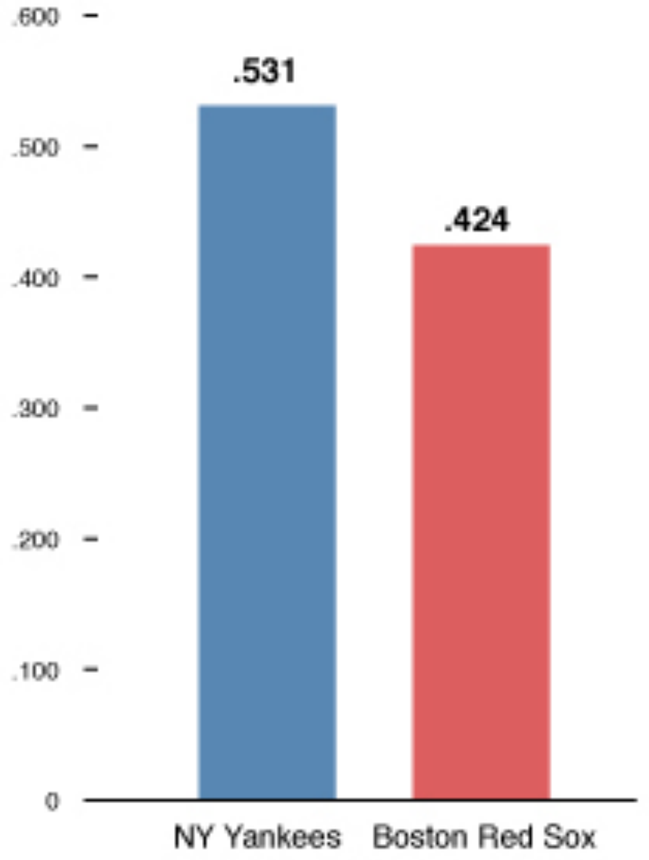
Edward Tufte, "The Visual Display of Quantitative Information"


OPERATING REVENUES
$3,549,385
$4,520,362
$4,916,444
$6,814,503
$7,382,599
1970
1971
1972
1973
1974
NET INCOME (LOSS)
$(11,014)
$397,747
$521,943
$1,647,001
$1,435,102
1970
1971
1972
1973
1974
EXPLORATION & DEVELOPMENT EXPENDITURES
$351,341
$85,149
$75,243
$329,421
$1,226,007
1970
1971
1972
1973
1974

Percentage of victories

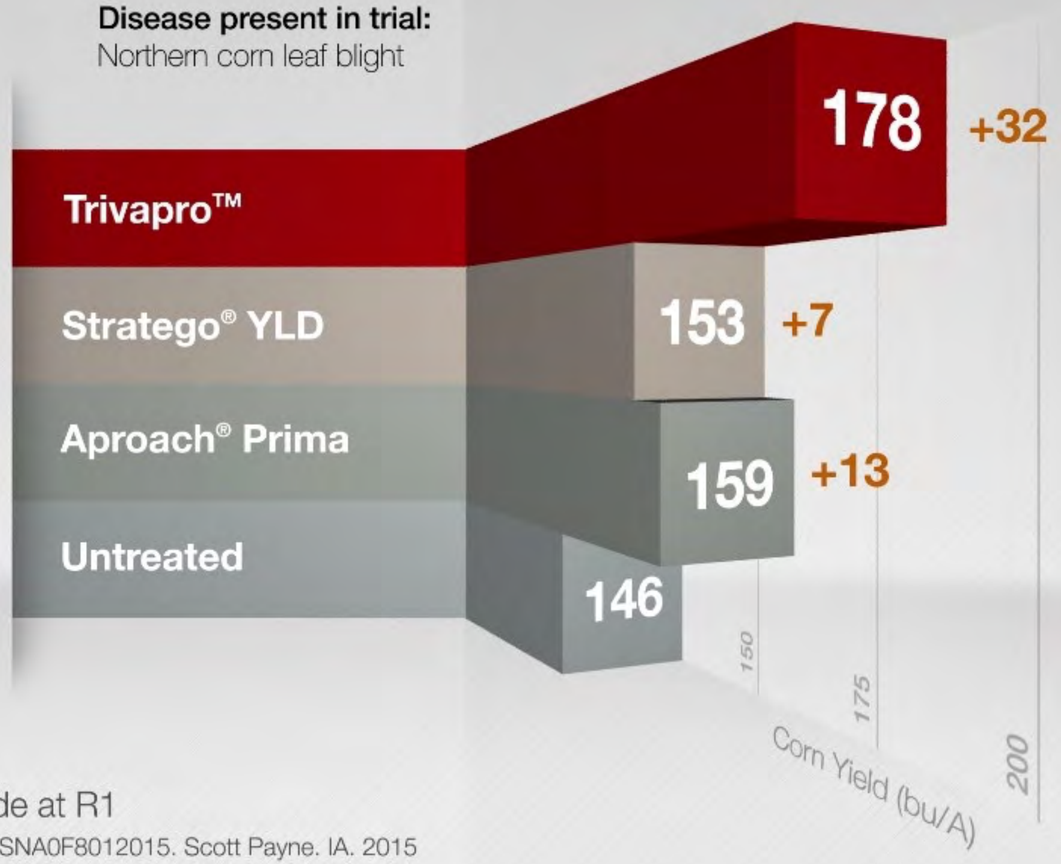


Percentage of victories



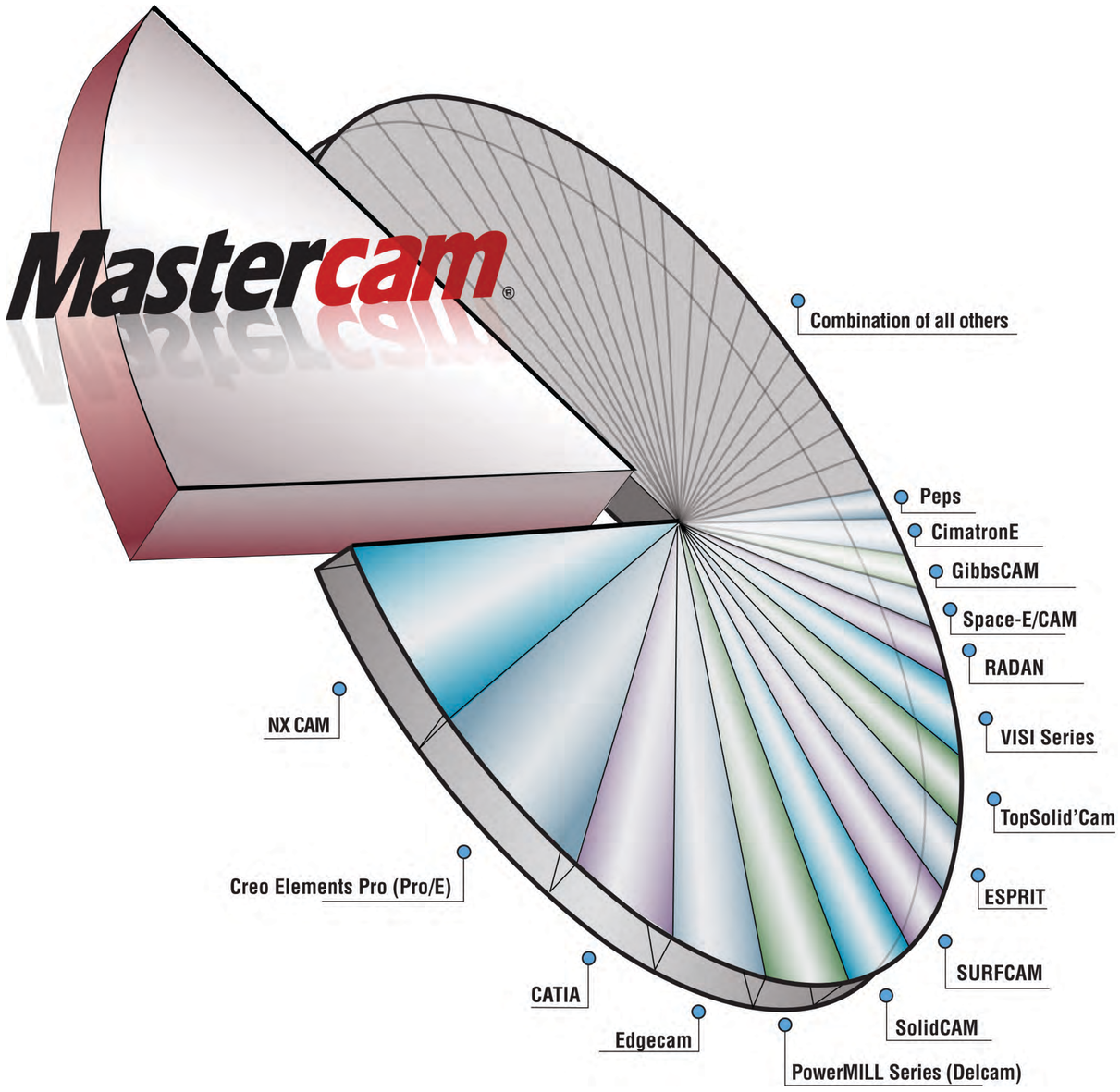
Trivapro corn yield response

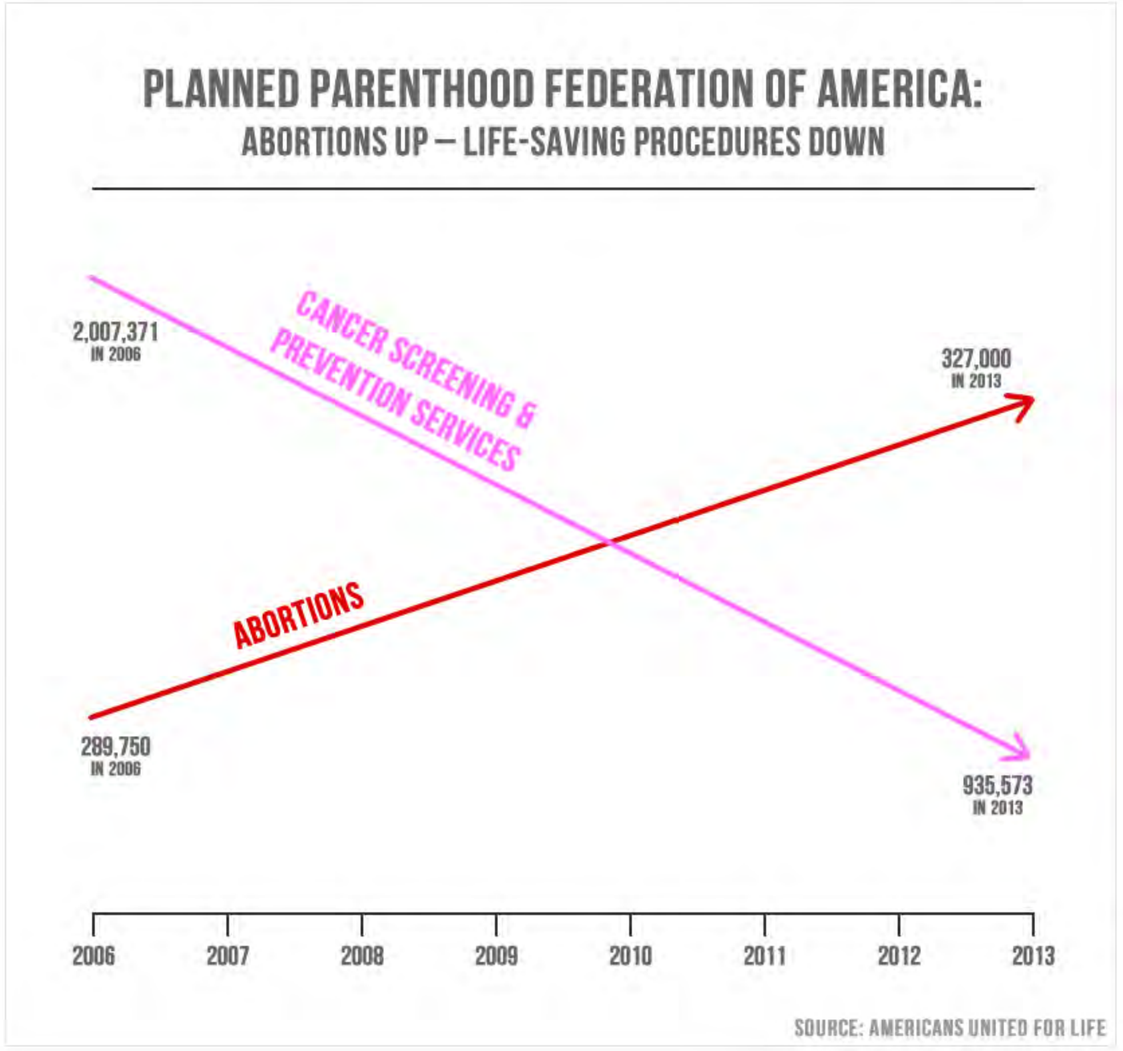
in Boone, IA

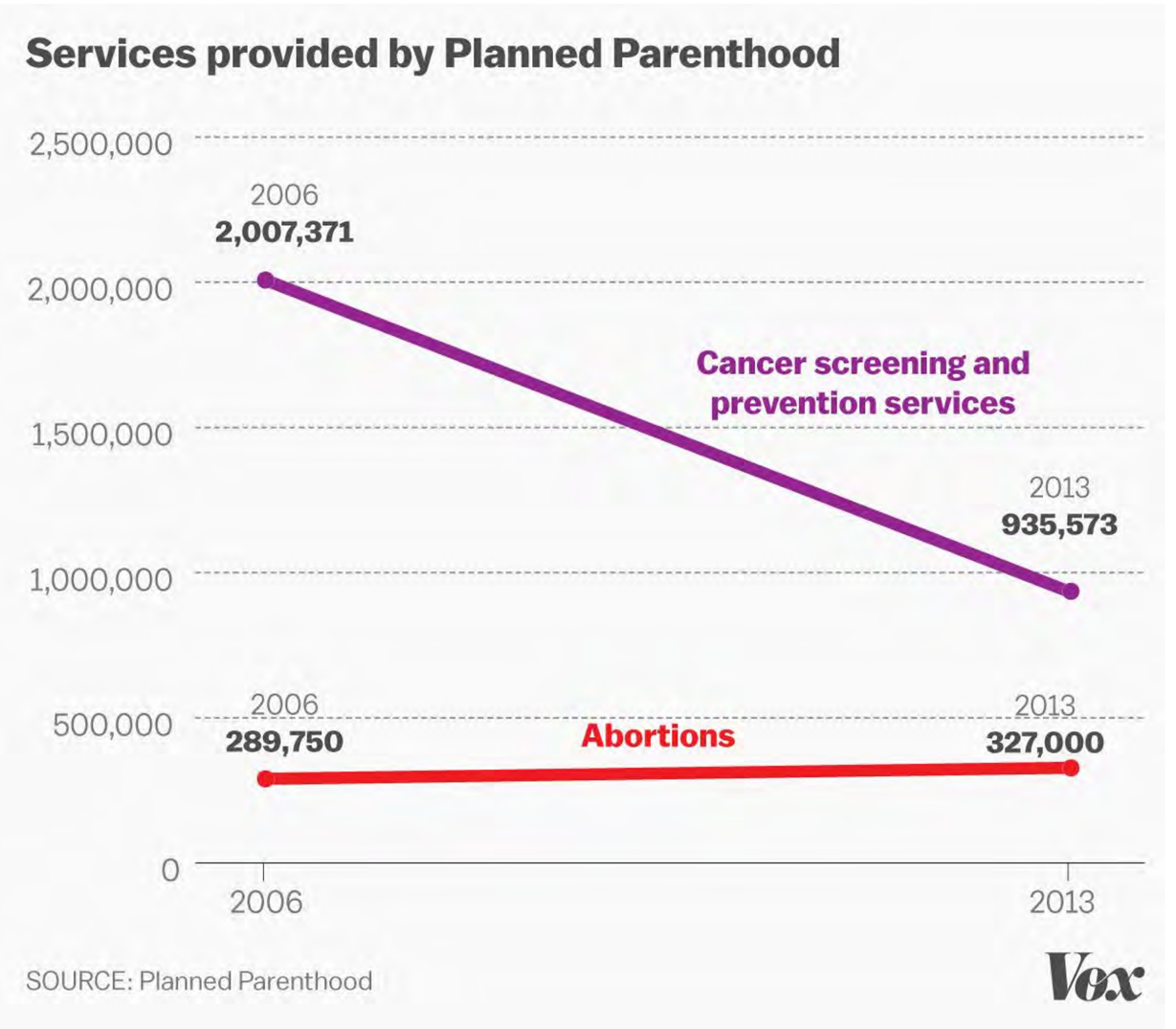


All applications made at R1

FSF001A4-2015US, Trial USNA0F8012015. Scott Payne. IA. 2015

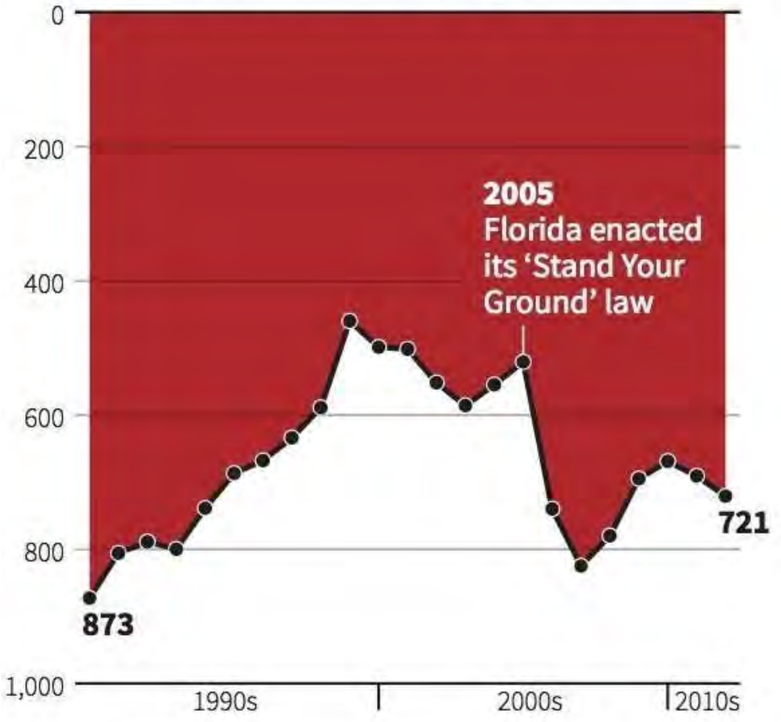

Mastercam
Combination of all others
Peps
CimatronE
GibbsCAM
Space-E/CAM
RADAN
VISI Series
TopSolid'Cam
ESPRIT
SURFCAM
SolidCAM
PowerMILL Series (Delcam)
Edgecam
CATIA
Creo Elements Pro (Pro/E)
NX CAM


PLANNED PARENTHOOD FEDERATION OF AMERICA:
ABORTIONS UP – LIFE-SAVING PROCEDURES DOWN
2,007,371
IN 2006
CANCER SCREENING &
PREVENTION SERVICES
327,000
IN 2013
ABORTIONS
289,750
IN 2006
935,573
IN 2013
2006
2007
2008
2009
2010
2011
2012
2013
SOURCE: AMERICANS UNITED FOR LIFE


Services provided by Planned Parenthood
2,500,000
2,000,000
1,500,000
1,000,000
500,000
0
2006
2,007,371
Cancer screening and
prevention services
2013
935,573
2006
289,750
Abortions
2013
327,000
2006
2013
SOURCE: Planned Parenthood
Vox

Gun deaths in Florida

Number of murders committed using firearms

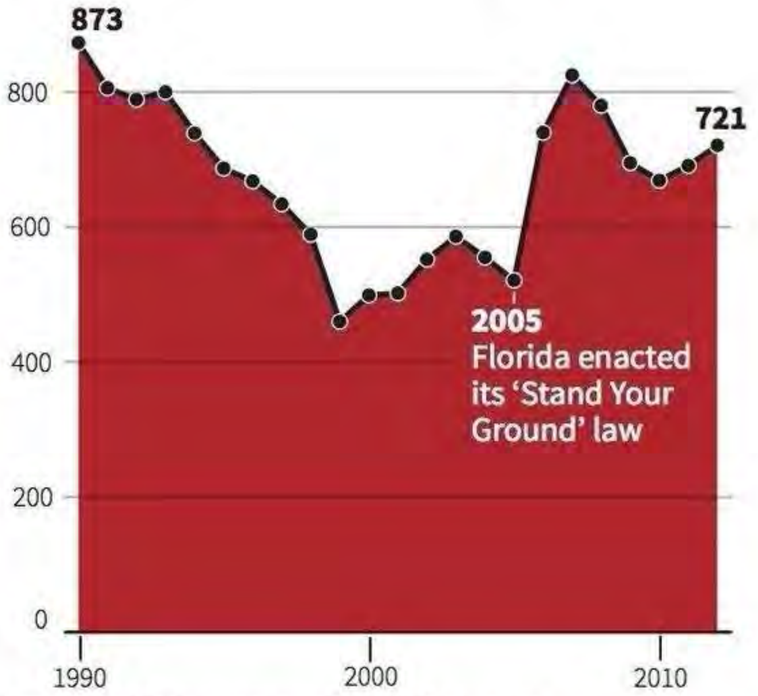


Source: Florida Department of Law Enforcement

C. Chan 16/02/2014 REUTERS

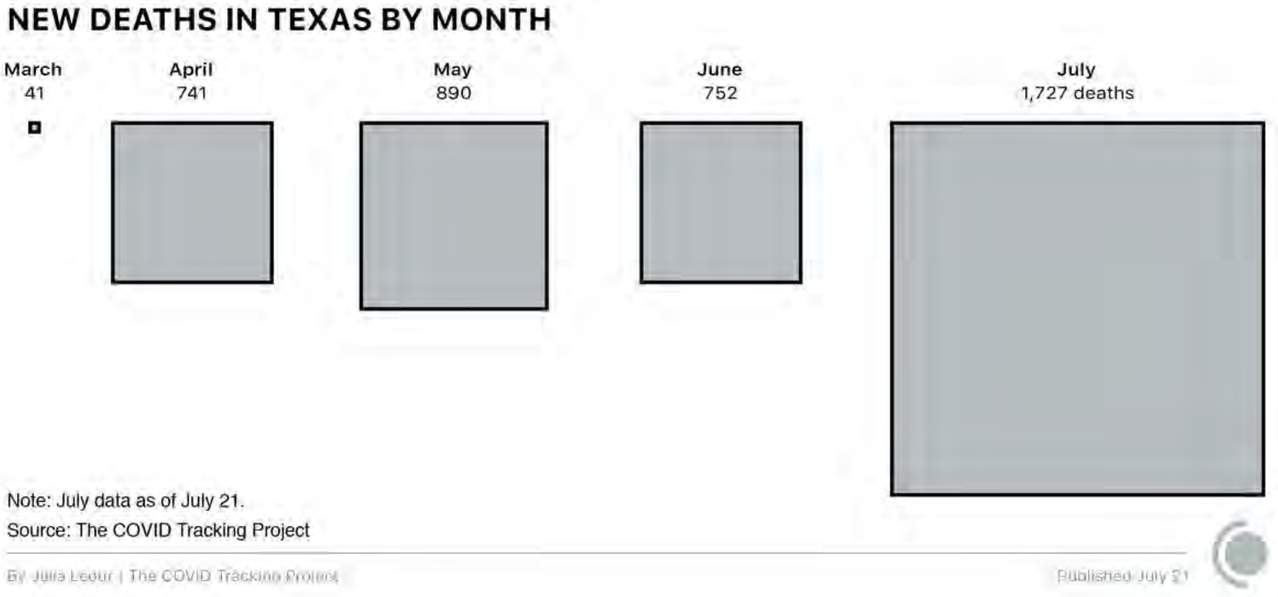
Gun deaths in Florida

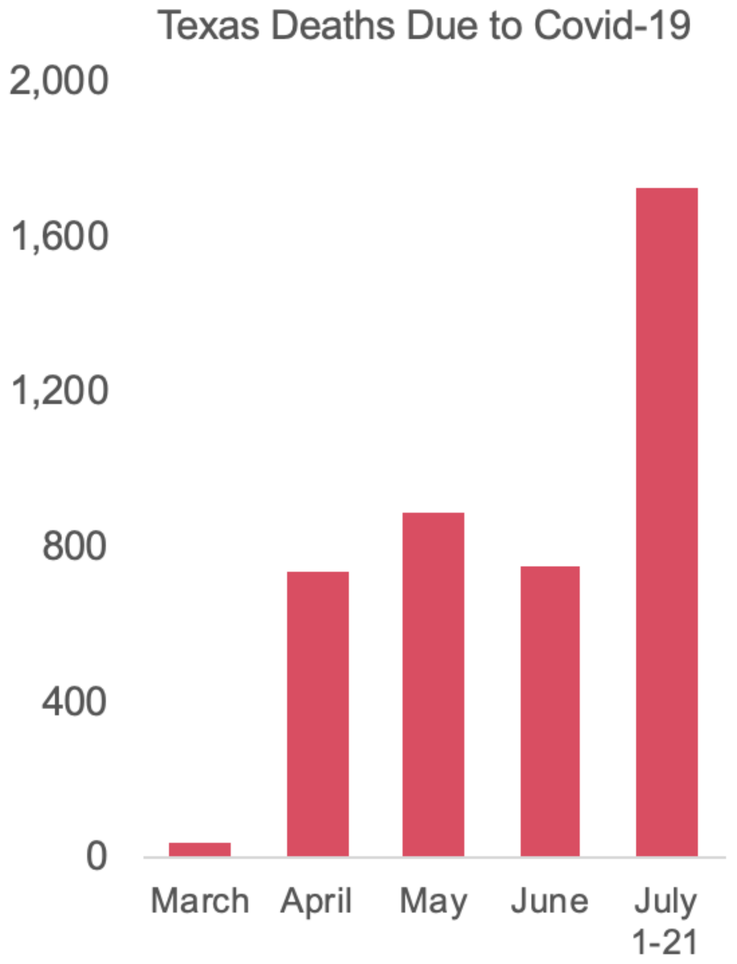
Number of murders committed using firearms



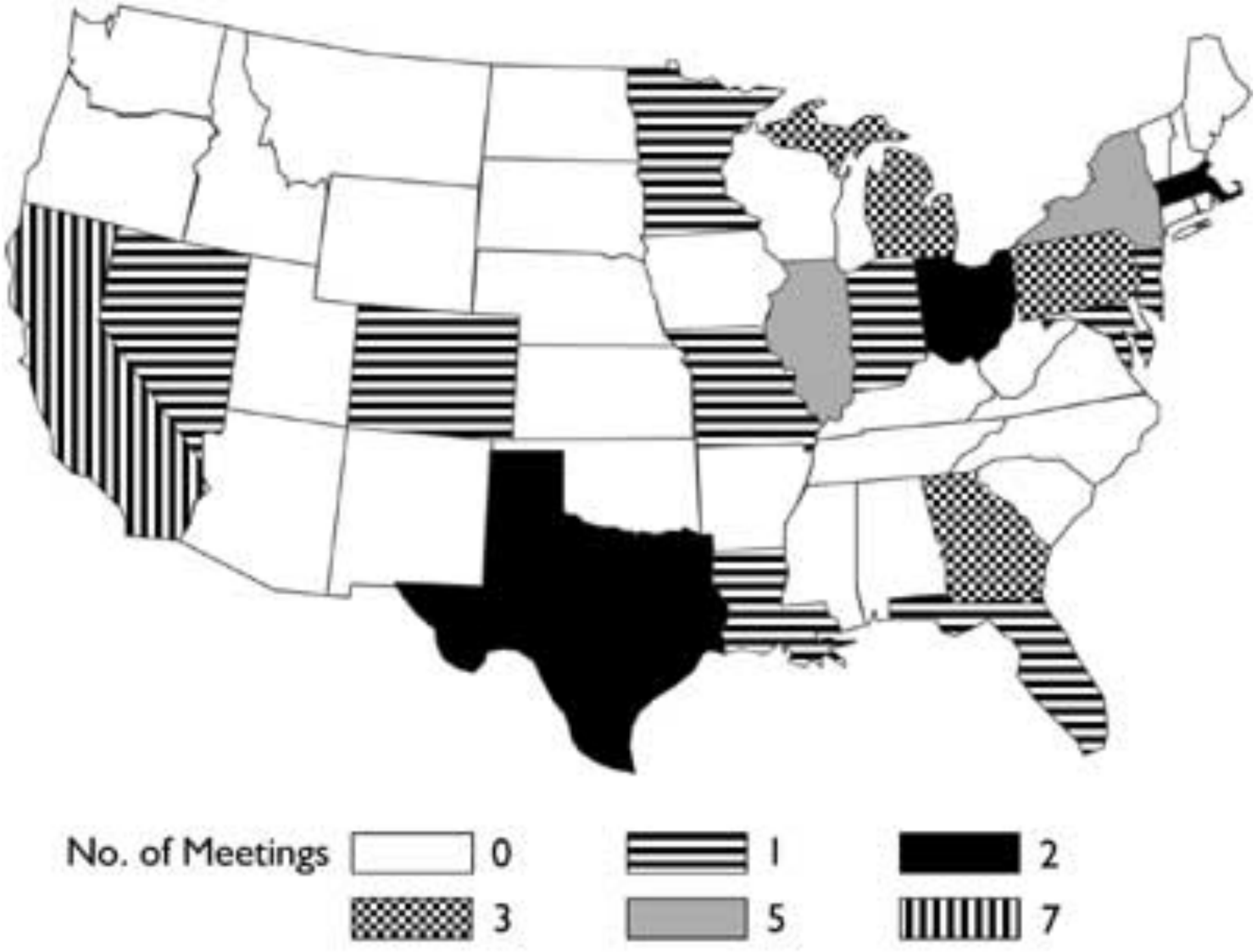
Source: Florida Department of Law Enforcement

C. Chan 16/02/2014 REUTERS





Adapted from COVID Tracking Project
Kaiser Fung / JunkCharts


No. of Meetings
0
1
2
3
5
7

Goals of graphical excellence

Visual representations should be consistent with numerical values

- Avoid non-linear representations (e.g., perspective, truncated, or log plots)
- Use consistent scales in multicomponent plots
- Avoid broken scales (often, there are better ways)

Use clear, explicit, and appropriate labeling

Use the appropriate number of variable dimensions

- Never exceed the number of data dimensions

Do not quote data out of context

- Make comparisons clear: controls, baselines, neutral models, etc.
- For time series, provide enough points to show trends

Do not selectively present data to alter conclusions

Component-based design

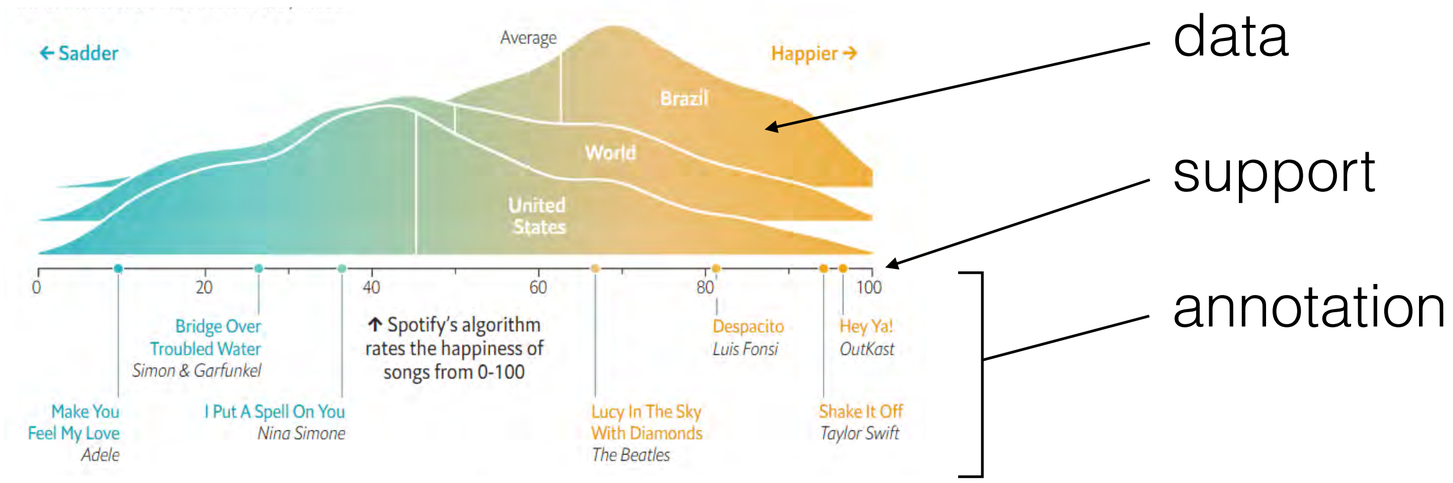
Component-based design

Quantitative graphics contain three kinds of components:

Data: geometric representations of numerical values (points, lines, bars, areas, etc.)

Data support: axes, axis labels, scales, grids, tick marks, encoding keys, etc.

Annotation: groupings, trend lines, arrows, highlights, text boxes, callouts, etc.



Data components

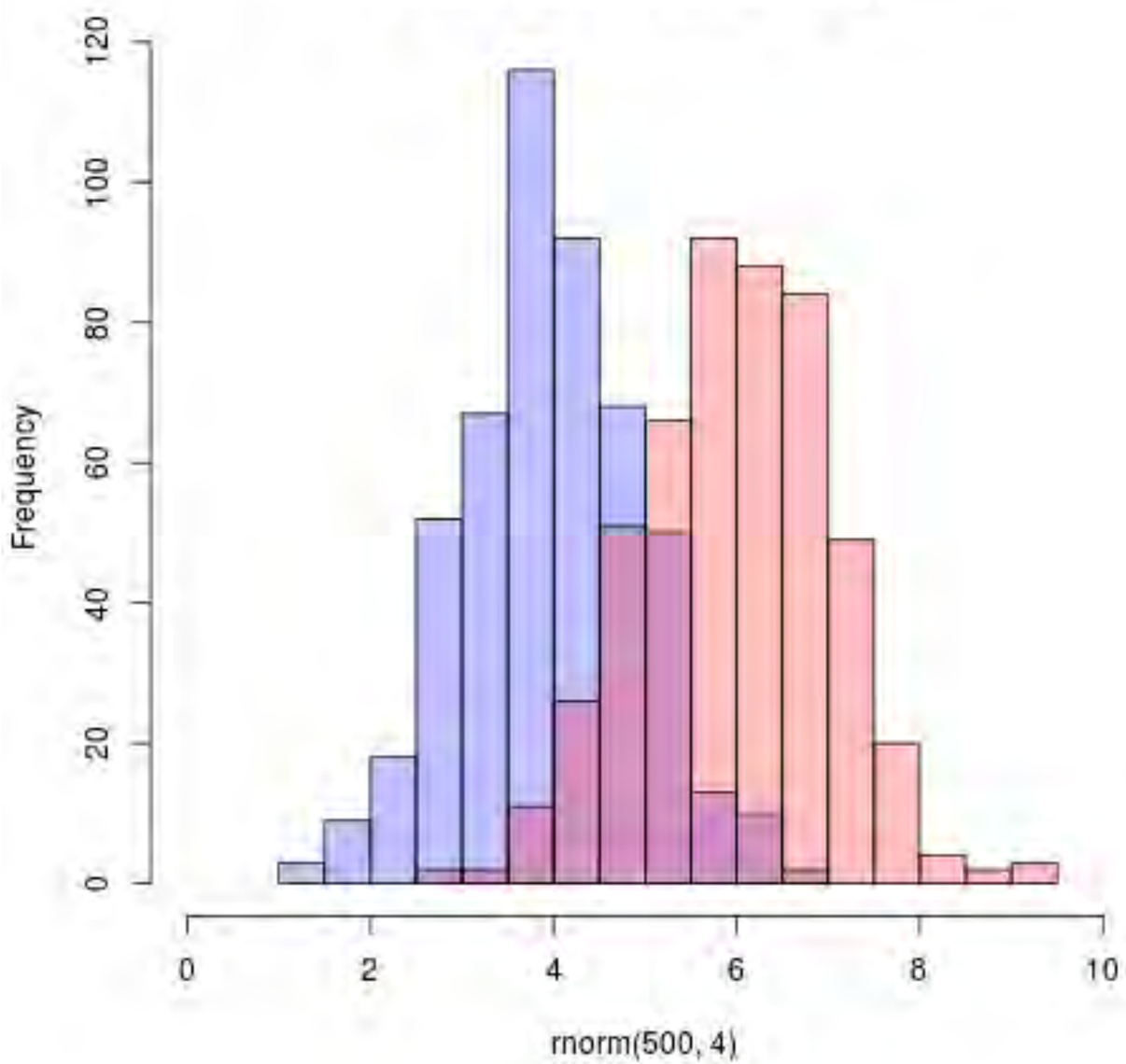
Primary objective:

- Data values and distributions should be easy to perceive
- Overall trends and differences/similarities among groups should be clear

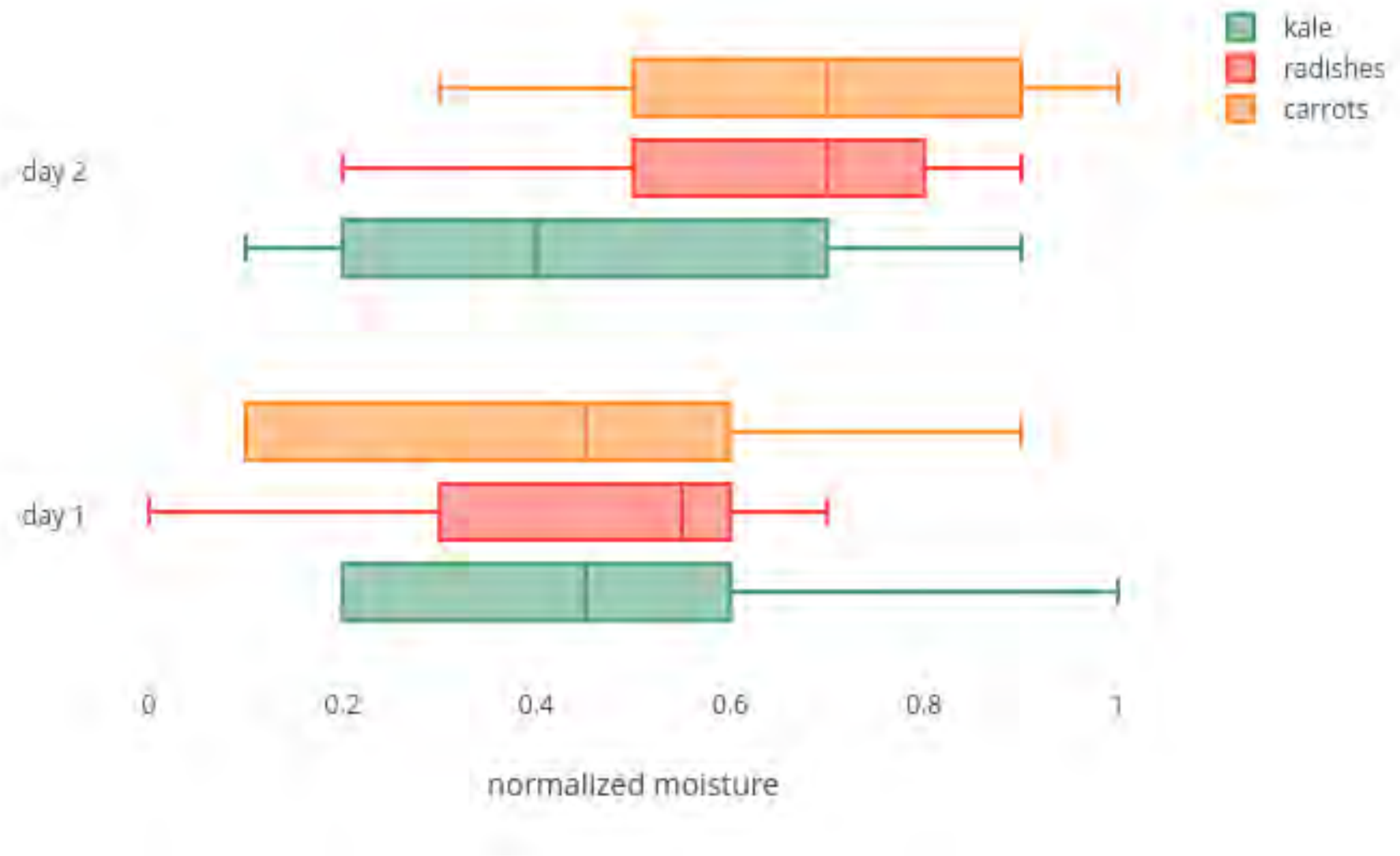
Points to consider

- Data representation: graph type, data point/line/area, size, color/shape/transparency
- How much data: all or summary or both?
- Encodings: aim to be intuitive and consistent
- Scales: linear/semi-log/log, zero-based or not, equal among all panels

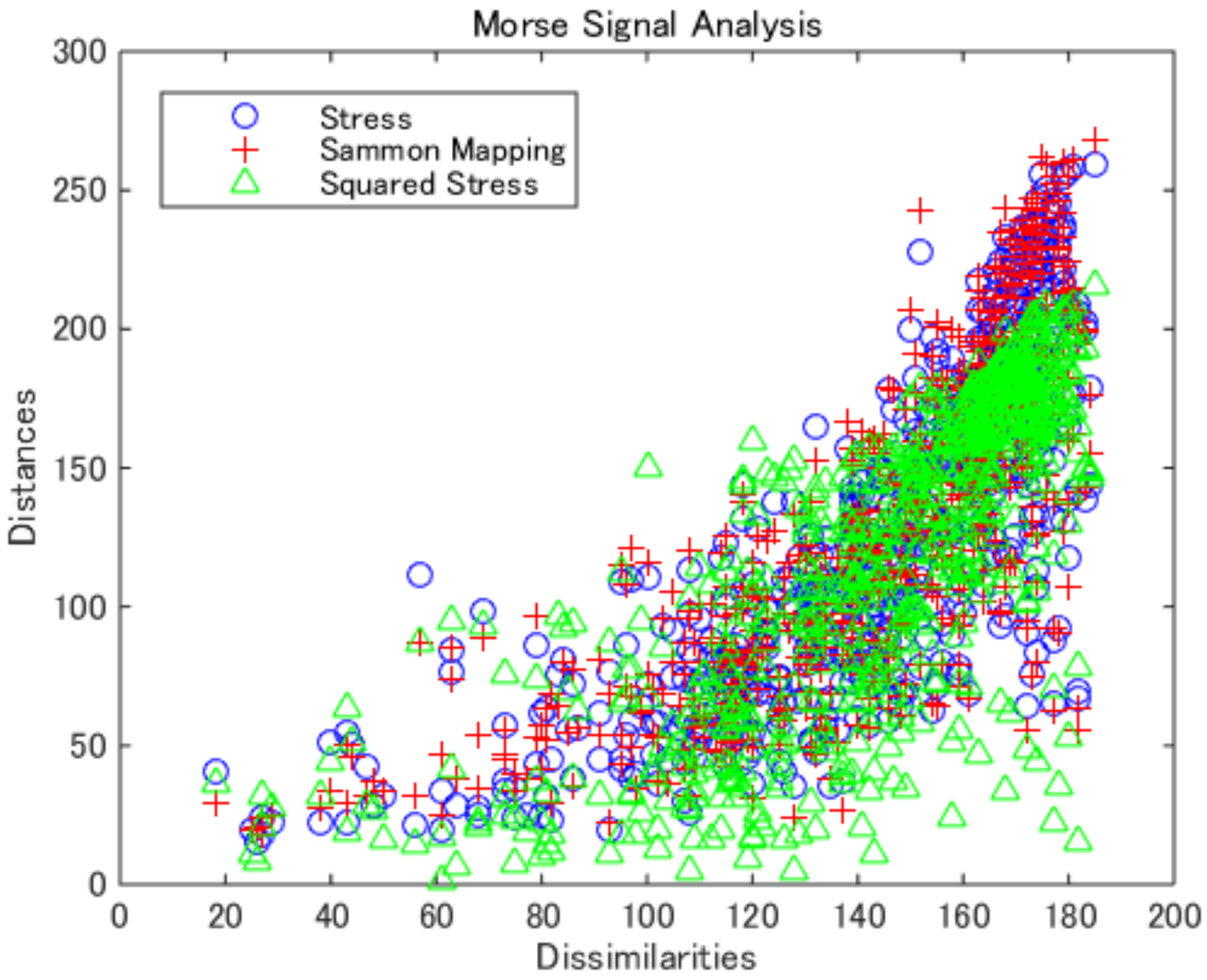
Example of intuitive encoding



Example of intuitive encoding



Example of overplotting



Support components

Primary objective: provide viewers the ability to understand the values of the data

Components to consider:

key

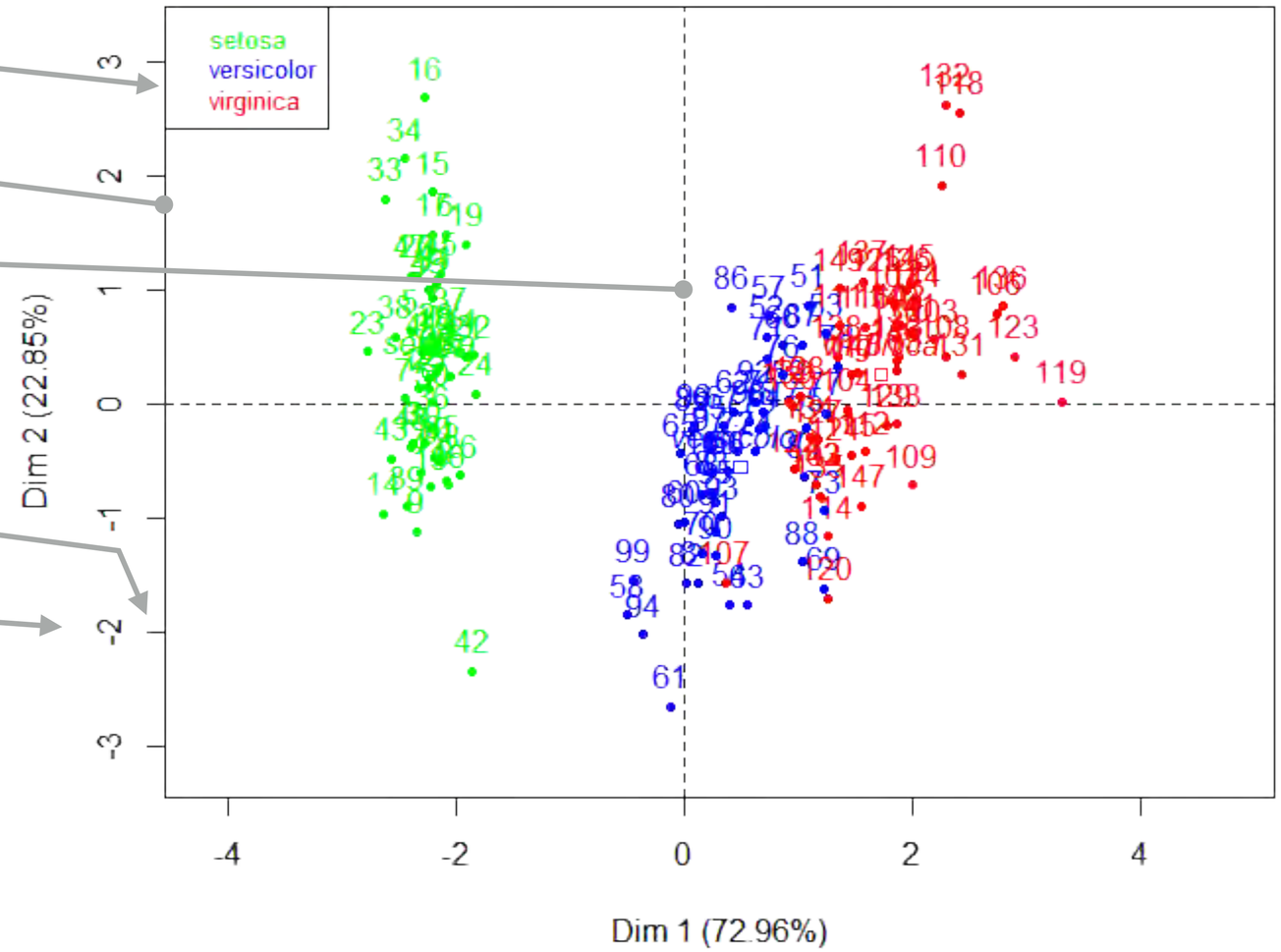
axis

grid

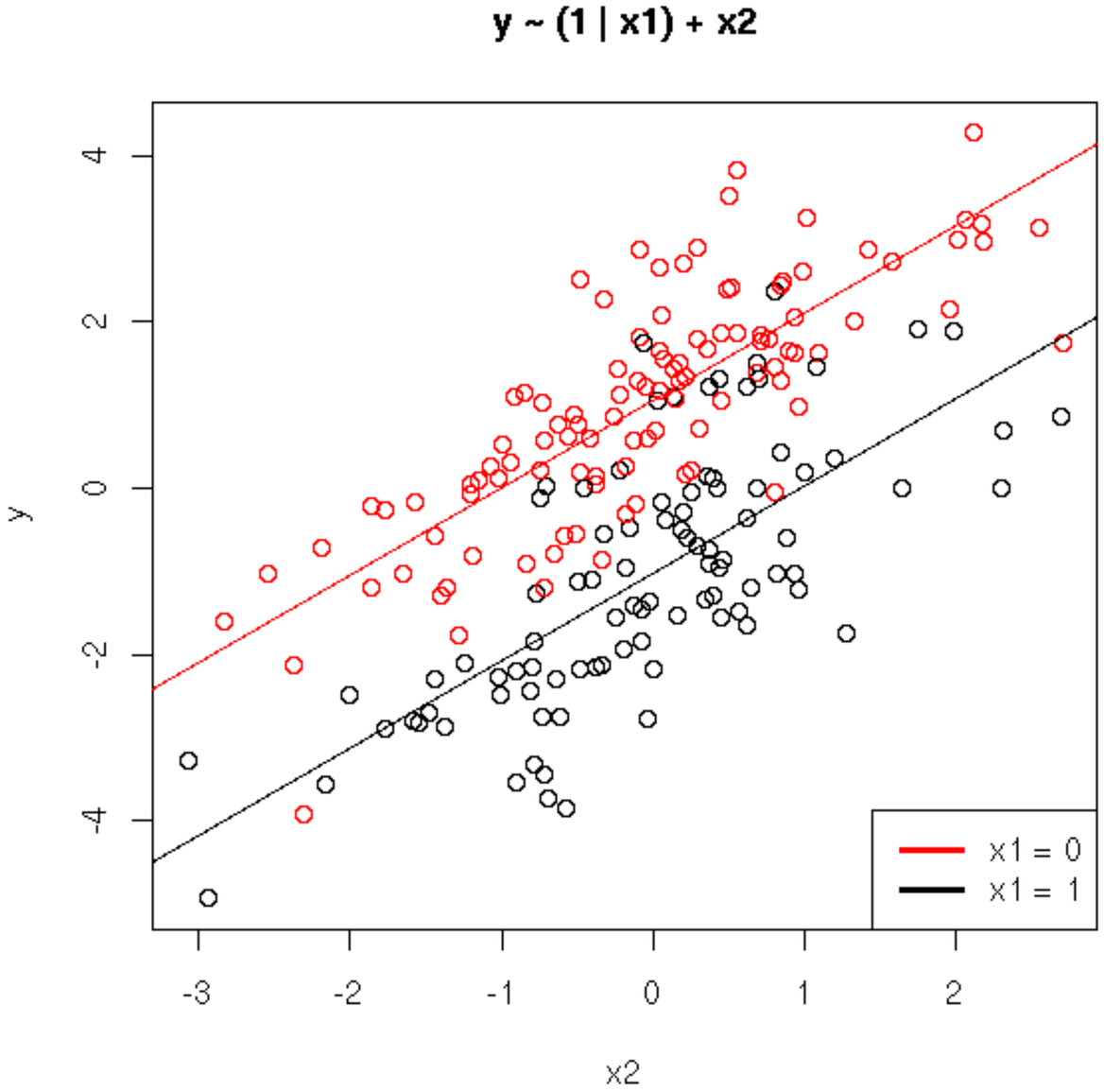
axis label

tick

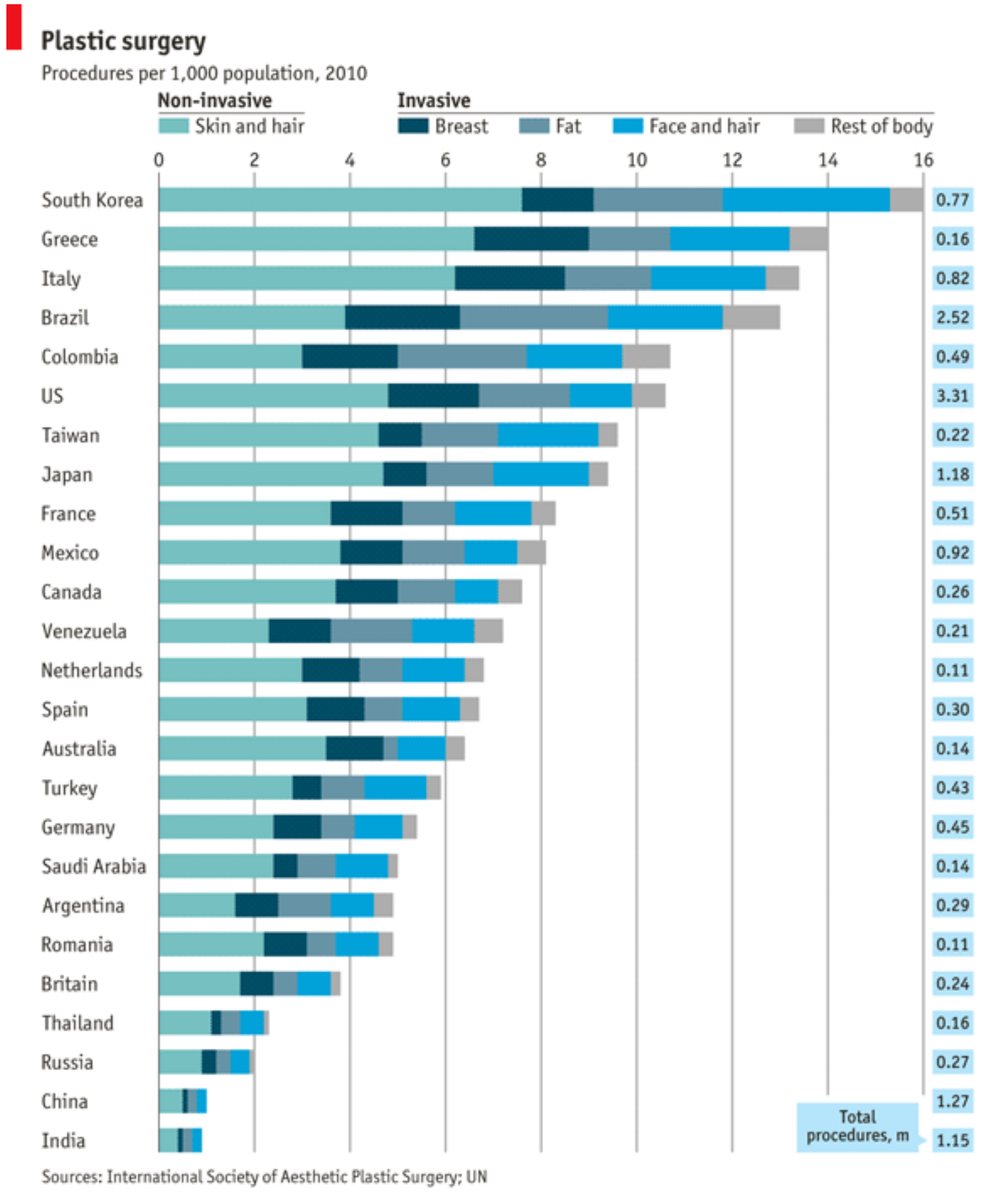
value



Example support components: trend lines



Example support components: key and alternate normalization



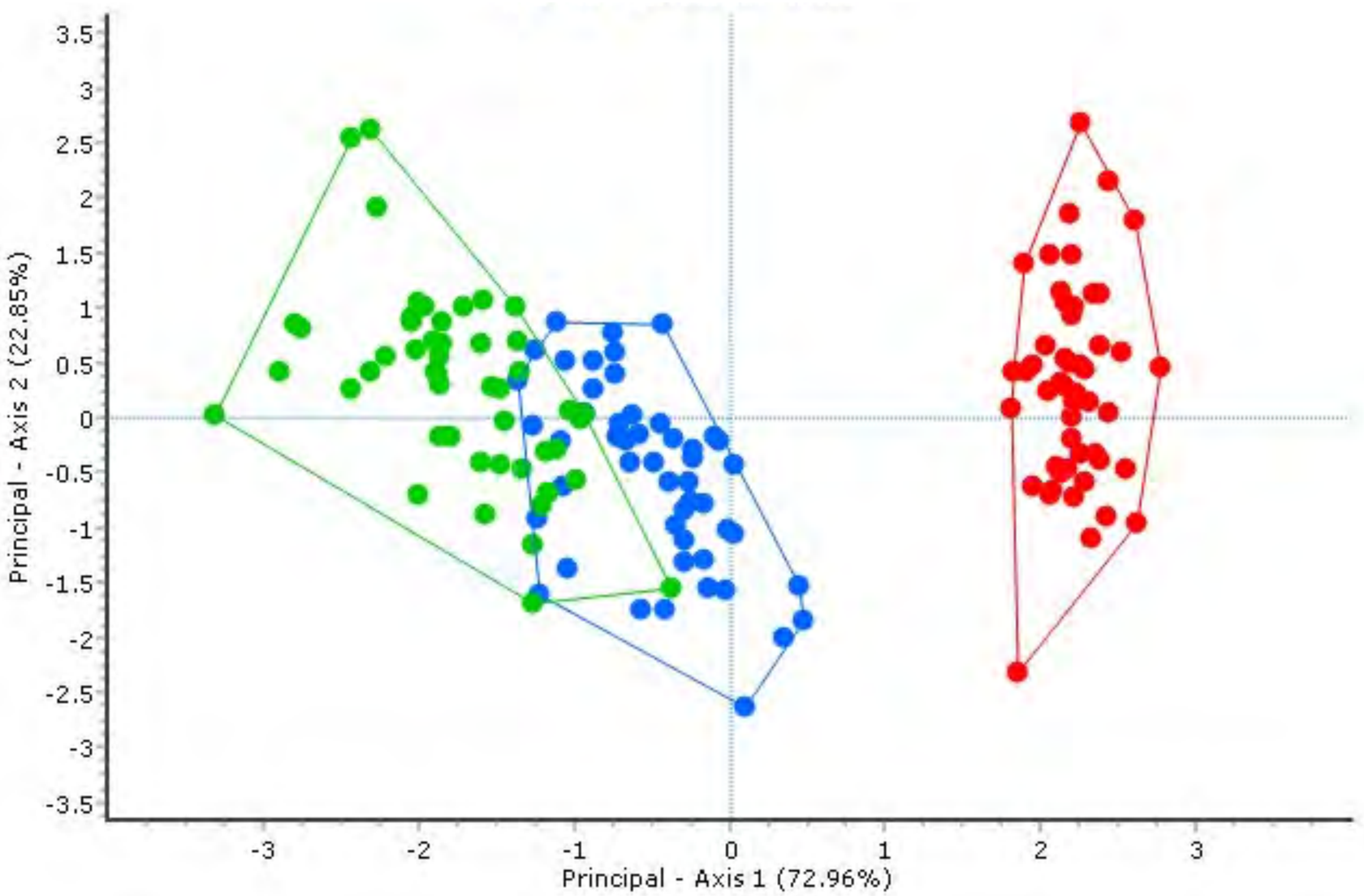
Annotation components

Primary objective: to tell a story

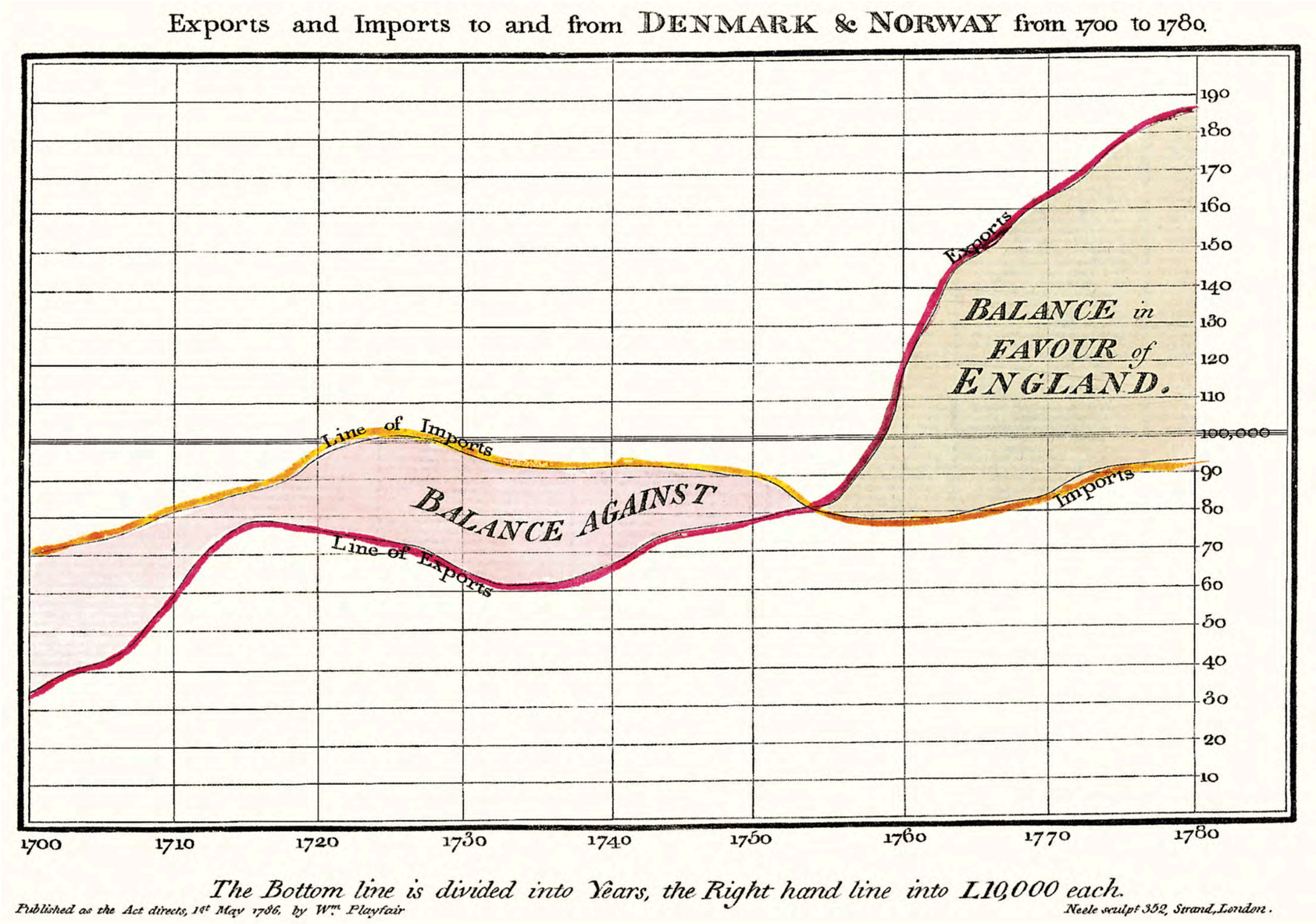
Components to consider:

- Title
- Values or labels for individual points
- Trend lines
- Groupings
- Highlighting, outlining, or transparency of data subset
- Statistical significance (asterisks, groupings, Tukey groups)
- Text labels referring to data components

Example annotation components: grouping



Example annotation components: text



Resources

Recommended resources

Books:

Tufte, E. 2001. *The Visual Display of Quantitative Information*. Graphics Press. (2nd ed.)

Wilke, C. 2019. *Fundamentals of Data Visualization*. O'Reilly Media.

Websites:

From Data to Viz: https://www.data-to-viz.com/

FriendsDontLetFriends: https://github.com/cxli233/FriendsDontLetFriends

R Graph Gallery: http://rgraphgallery.blogspot.com/

MatPlotLib Examples: https://matplotlib.org/stable/plot_types/index.html

Seaborn Gallery: https://seaborn.pydata.org/examples/index.html

Syntax for adding constraints

```
CREATE TABLE t1 (
    uid INTEGER PRIMARY KEY,
    genus VARCHAR NOT NULL UNIQUE,
    species VARCHAR NOT NULL,
    b_date DATE UNIQUE,
    age INTEGER DEFAULT = 42,
    entry_date DATE DEFAULT CURRENT_DATE,
    p1 INTEGER CHECK (c6 > 10 AND c6 != 0),
    p2 INTEGER CHECK (6 IN ('red', 'green', ‘blue’)),
);
```

Note: some versions of SQL use a different syntax for `CHECK` constraints and `CURRENT_DATE`

SQL is limited to a small set of operations

Only four kinds of operations are permitted with relational databases

Create, **R**ead, **U**pdate, **D**elete (CRUD)

`CREATE` to create a new database, table, or relation; `INSERT` to add rows to a table

`SELECT` to query a database ✔

`UPDATE` to change the information within a table

`DELETE` to remove rows from a table; `DROP` to remove a database, table, or relation

SQL has some additional keywords, but most work with the above set

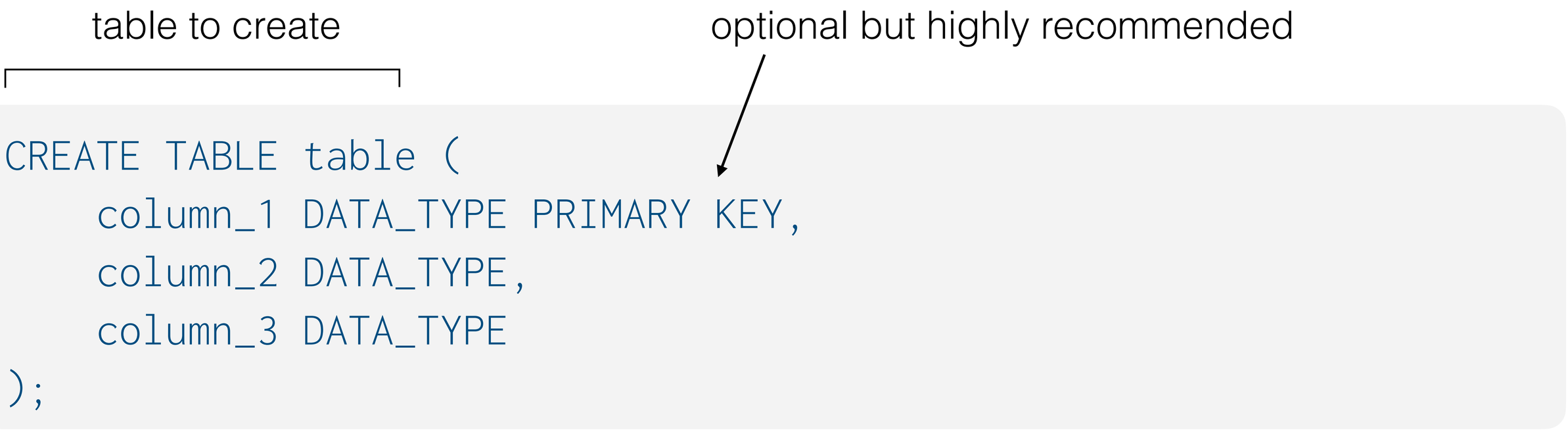
Clauses within statements: `WHERE`, `JOIN`, `GROUP BY`, `LIMIT`, etc.

Operators within statements: `=`, `>`, `IN`, `NOT`, `LIKE`, `BETWEEN`, etc.

Functions within statements: `MIN`, `MEAN`, `COUNT`, `DISTINCT`, etc.

`Pandas` and `dplyr` implement many SQL-like operations (and were inspired by it)

CREATE syntax



Data types typically include: `INTEGER`, `FLOAT`, `BOOLEAN`, `CHAR(n)`, `VARCHAR`, `DATE`, etc.

Optional qualifiers: `PRIMARY KEY`, `NOT NULL`, `UNIQUE`, etc.