Introduction to Data Science

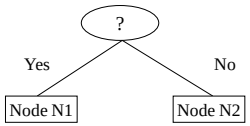
CS 5665
Utah State University
Department of Computer Science
Instructor: Prof. Kyumin Lee

Practicum

- Thursday
  - NLTK (Vahe)
  - Gephi (Aditya)
  - Processing.js (Yancy)
  - D3 (Meiling)
  - Highcharts (Jacob)

GINI for Binary Attributes

- Splits into two partitions
- Effect of Weighing partitions:
  - Larger and Purer Partitions are sought for.



	Parent
C1	6
C2	6
Gini = 0.500	

Gini(N1)
= 1 – $(5/7)^2$ – $(2/7)^2$
= 0.408

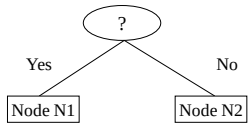
Gini(N2)
= 1 – $(1/5)^2$ – $(4/5)^2$
= 0.320

	N1	N2
C1	5	1
C2	2	4
Gini=0.371		

Gini(Children)
= 7/12 * 0.408 +
5/12 * 0.320
= 0.371

GINI for Binary Attributes

- Splits into two partitions
- Effect of Weighing partitions:
  - Larger and Purer Partitions are sought for.



	Parent
C1	6
C2	6
Gini = 0.500	

Attribute A

	N1	N2
C1	0	6
C2	6	0
Gini=0.000		

Attribute B

	N1	N2
C1	5	1
C2	1	5
Gini=0.278		

Attribute C

	N1	N2
C1	4	2
C2	3	3
Gini=0.486		

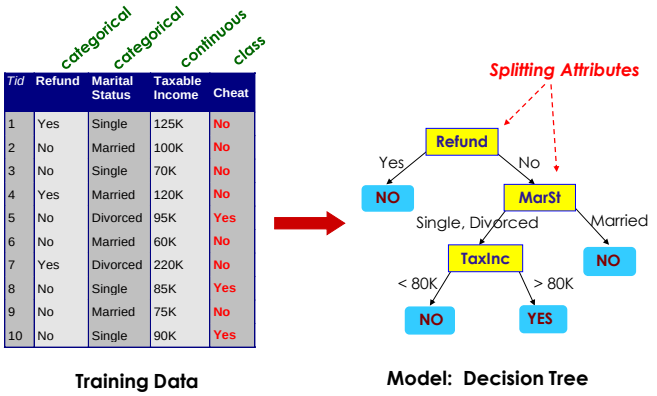
Attribute D

	N1	N2
C1	3	3
C2	3	3
Gini=0.500		

Tree Induction

- Determine how to split the records
  - Use greedy heuristics to make a series of locally optimum decision about which attribute to use for partitioning the data
  - At each step of the greedy algorithm, a test condition is applied to split the data in to subsets with a more homogenous class distribution
    - How to specify test condition for each attribute
    - How to determine the best split
- Determine when to stop splitting
  - A stopping condition is needed to terminate tree growing process. Stop expanding a node
    - if all the instances belong to the same class
    - if all the instances have similar attribute values

Example of a Decision Tree



Evaluating a Classifier

Accuracy

		PREDICTED CLASS	
		Class=Yes	Class=No
ACTUAL CLASS	Class=Yes	a (TP)	b (FN)
	Class=No	c (FP)	d (TN)

$$\text{Accuracy} = \frac{a+d}{a+b+c+d} = \frac{TP+TN}{TP+TN+FP+FN}$$

Cost Matrix

		PREDICTED CLASS	
ACTUAL CLASS	C(i|j)	**Class=Yes**	**Class=No**
	Class=Yes	C(Yes|Yes)	C(No|Yes)
	Class=No	C(Yes|No)	C(No|No)

C(i | j): Cost of misclassifying class j example as class i

Computing Cost of Classification

Cost Matrix		PREDICTED CLASS	
ACTUAL CLASS	C(i|j)	+	-
	+	-1	100
	-	1	0

Model M_1		PREDICTED CLASS	
ACTUAL CLASS		+	-
	+	150	40
	-	60	250

Accuracy = 80%
Cost = 3910

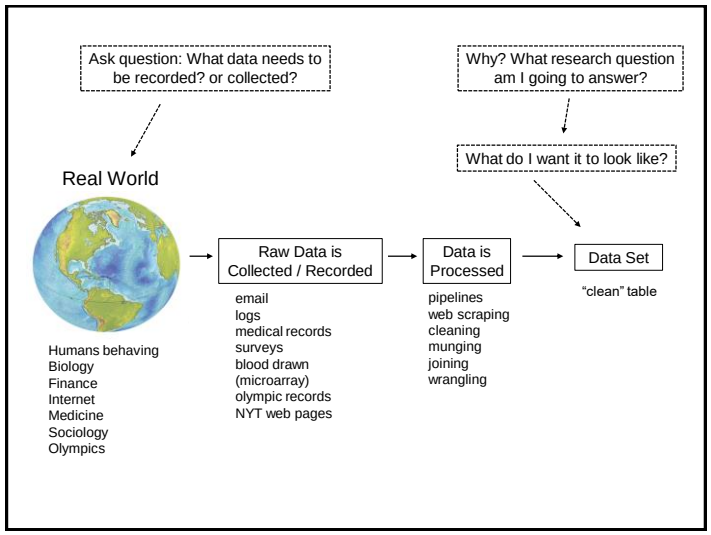
Model M_2		PREDICTED CLASS	
ACTUAL CLASS		+	-
	+	250	45
	-	5	200

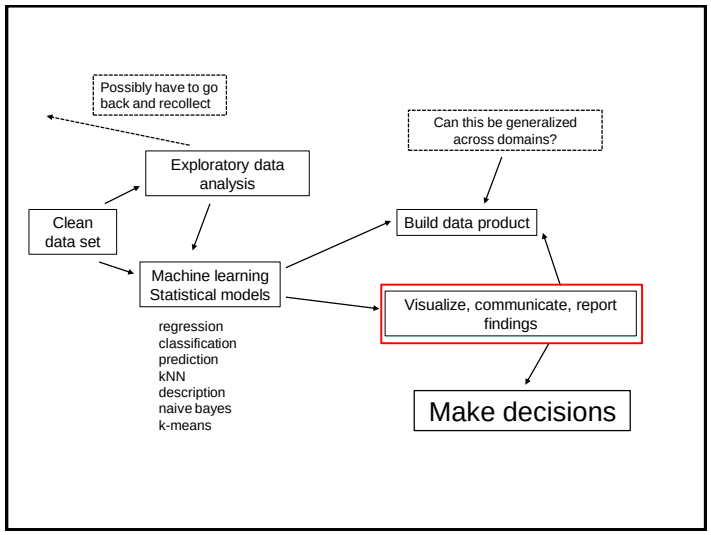
Accuracy = 90%
Cost = 4255

How to Estimate True "Accuracy" (or whatever we're measuring)

- Holdout
  - Reserve 2/3 for training and 1/3 for testing
- Cross validation
  - Participation data into k disjoint subsets
  - K-fold: training on k-1 partitions, test the remaining one
- Bootstrap
  - Sampling with replacement

Data Science: The Context





Data Visualization

What is visualization?

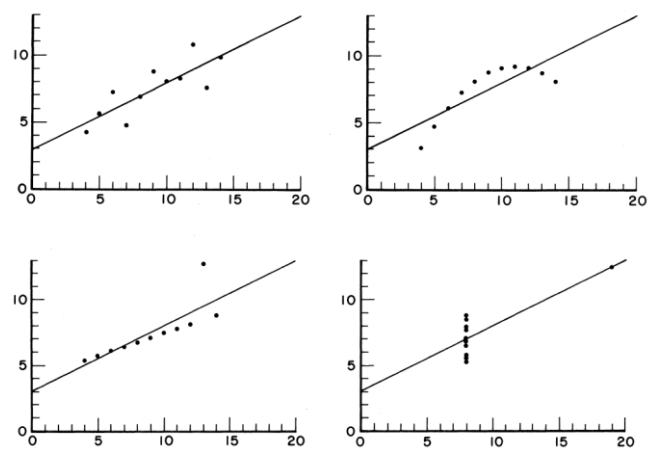
- "Transformation of the symbolic into the geometric" [McCormick et al. 1987]
- "... finding the artificial memory that best supports our natural means of perception." [Bertin 1967]
- "The use of computer-generated, interactive, visual representations of data to amplify cognition." [Card, Mackinlay, & Shneiderman 1999]

Four Datasets

I		II		III		IV	
X	Y	X	Y	X	Y	X	Y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

Anscombe "Graphs in Statistical Analysis" 1973

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



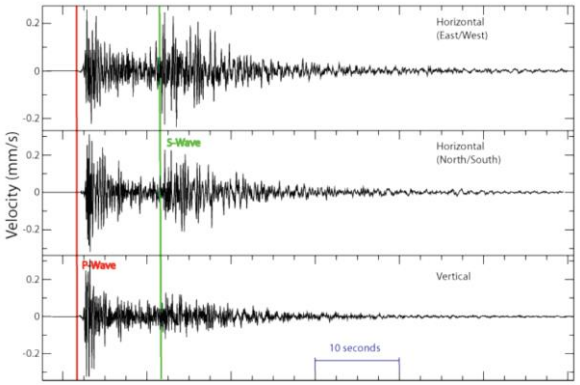
Why Do We Create Visualizations?

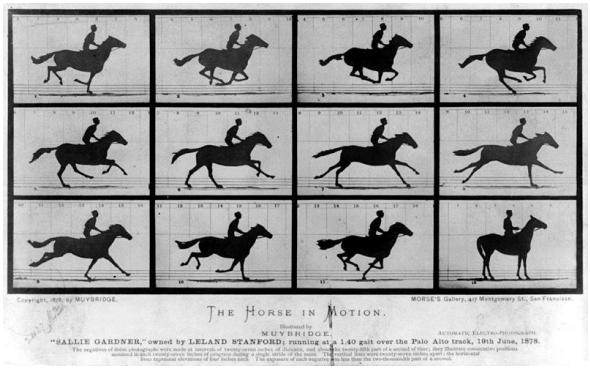
- Answer questions (or discover them)
- Make decisions
- See data in context
- Expand memory
- Support graphical calculation
- Find patterns
- Present argument or tell a story
- Inspire

Jeff Heer

The Value of Visualization

- **Record** information
  - Blueprints, photographs, seismographs, …
- **Analyze** data to support reasoning
  - Develop and assess hypotheses
  - Discover errors in data
  - Expand memory
  - Find patterns
- **Communicate** information to others
  - Share and persuade
  - Collaborate and revise





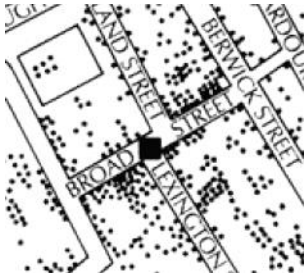
Analyze

Data in context: Cholera outbreak

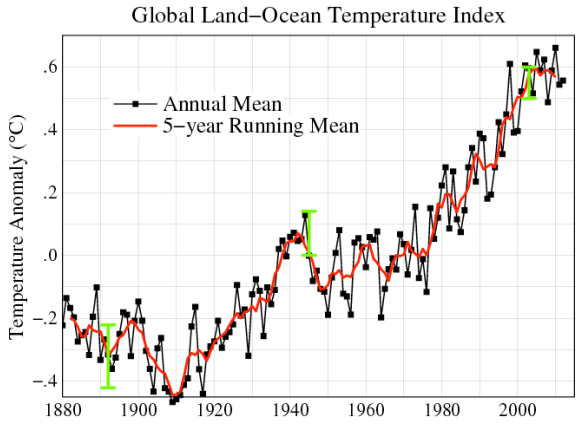
In 1854 John Snow plotted the position of each cholera case on a map. [from Tufte 83]

https://en.wikipedia.org/wiki/1854_Broad_Street_cholera_outbreak

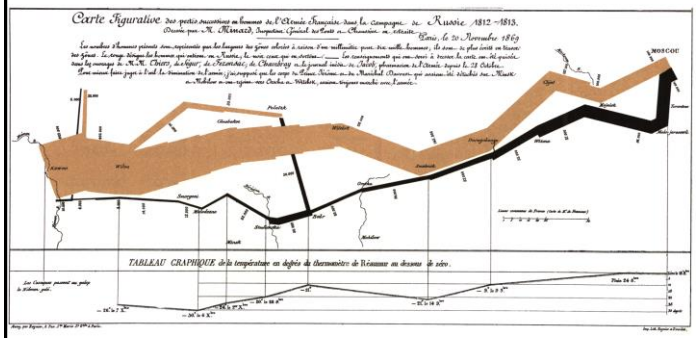
Data in context: Cholera outbreak



Used map to hypothesize that pump on Broad St. was the cause. [from Tufte 83]

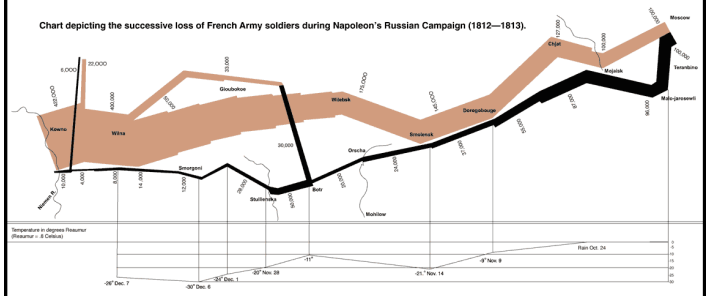


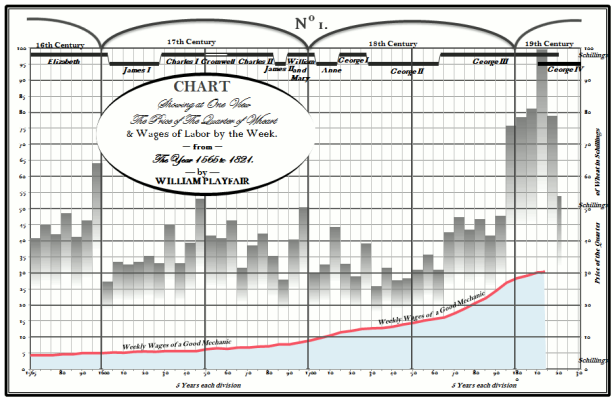
Communicate Information to Others



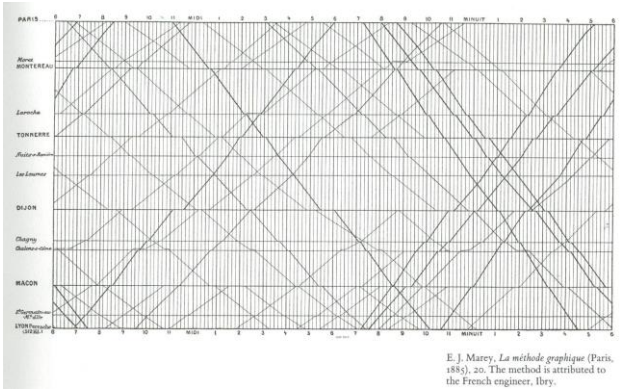
Napolean's invasion of Russia, as drawn by Charles Joseph Minard (1781-1870)

Show Space, Time and Temperature



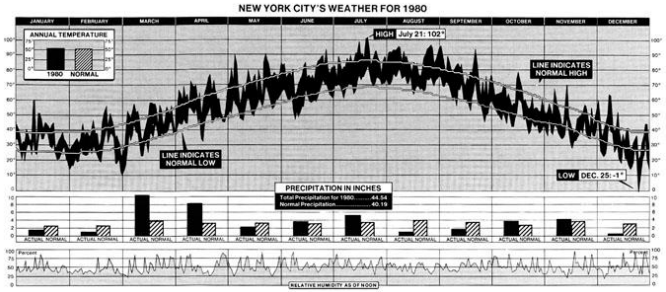


The price of wheat compared to labor wages, William Playfair (1759-1823)



E. J. Marey, *La méthode graphique* (Paris, 1885), 20. The method is attributed to the French engineer, Ibry.

French train schedule, as drawn by E.J. Marey (1830-1904)



Reasons?

- Lots of data -- compact representation
- Identify what is being represented
  - Data clarity
- Choice of presentation matters (pie chart vs. time series vs. map ...)
- Easy to compare / contrast (ANALYZE)
- Multi-data types

Tufte: Principles of Graphical Excellence

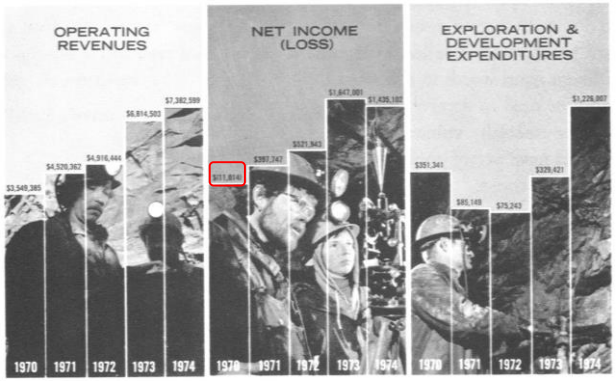
- Graphical excellence is the well-designed presentation of interesting data – a matter of *substance*, *statistics*, and *design*
- Graphical excellence consists of complex ideas communicated with *clarity*, *precision*, and *efficiency*

Tufte: Graphical Integrity

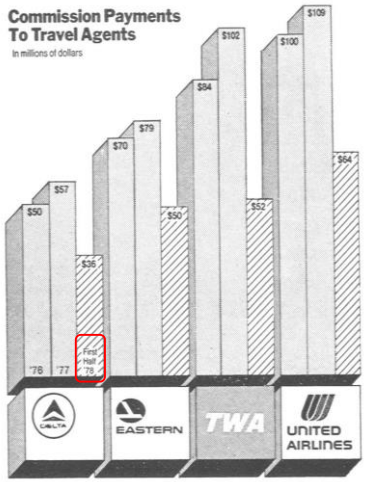
- "not lying with statistics"
- tell the truth about data

Uh oh …

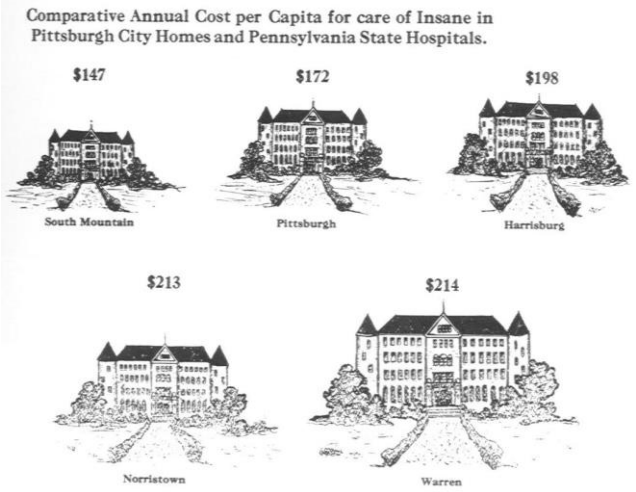
Examples of Infographics lacking integrity



Day Mines, Inc., 1974 Annual Report



New York Times, 8/8/78



Lie Factor

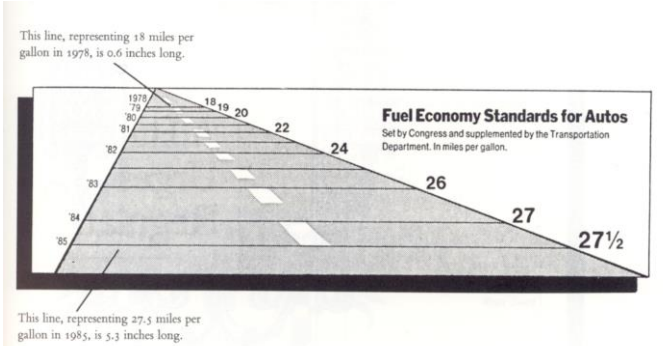
- Given perceptual difficulties – strive for uniformity (predictability) in graphics (p56)
  - 'the representation of numbers, as physically measured on the surface of the graphic itself, should be directly proportional to the numerical quantities represented.'
  - 'Clear, detailed and thorough labeling should be used to defeat graphical distortion and ambiguity. Write out explanations of the data on the graphic itself. Label important events in the data.

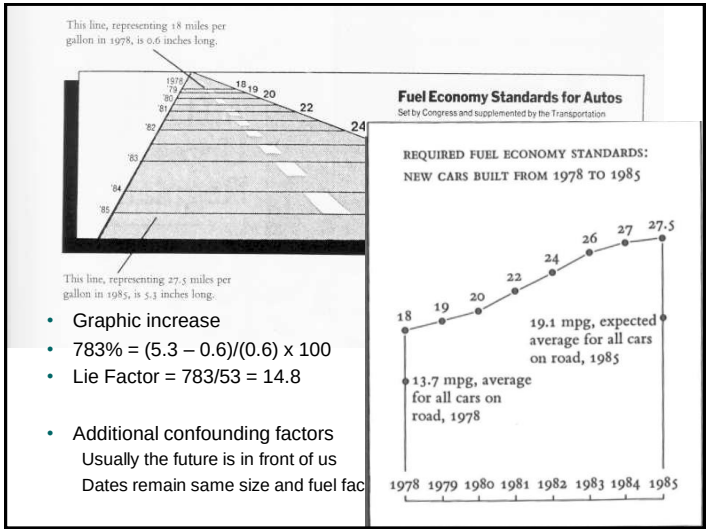
$$\text{Lie Factor} = \frac{\text{size of effect shown in graphic}}{\text{size of effect in data}}$$

- Lie factor of 1 ➔ is desirable
- Lie factor > 1.05 or < 0.95 go beyond plotting errors

Extreme example

- Fuel economy standards for automobiles
  18 miles/gallon in 1978 to 27.5 miles/gallon in 1985
  Increase of 53% = (27.5 – 18.0)/(18.0) x 100





- Graphic increase
- 783% = (5.3 – 0.6)/(0.6) x 100
- Lie Factor = 783/53 = 14.8

- Additional confounding factors
  Usually the future is in front of us
  Dates remain same size and fuel fac

Visual Area and Numerical Measure

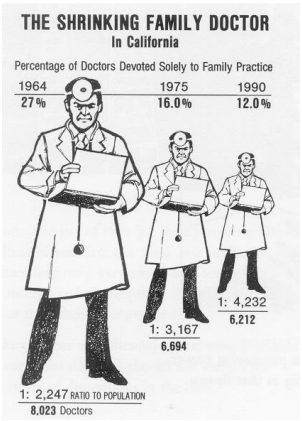
Use of area to portray 1D data can be confusing

-Area has 2 dimensions

The 'incredible' shrinking family doctor

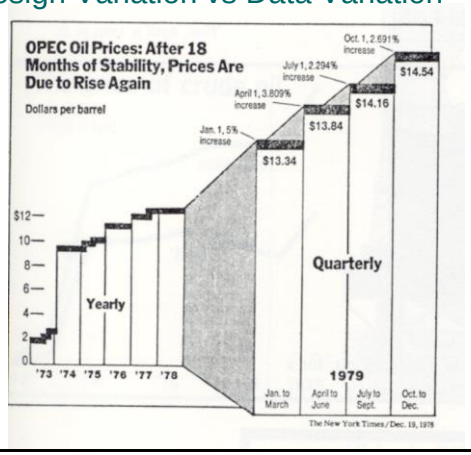
Lie factor of 2.8

Plus incorrect horizontal spacing



Los Angeles Times, August 5, 1979 p.3, (Tufte, 1983, p69)

Design Variation vs Data Variation



New York Times, Dec. 19, 1978, p.D-7 (Tufte, 1983, p61)

Design Variation vs Data Variation

- 5 different vertical scales show price

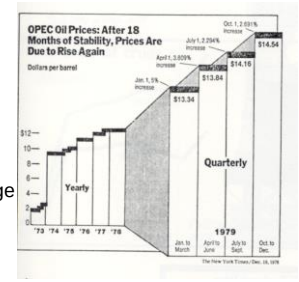
During this time	one vertical inch equals
1973 -1978	$8.00
Jan. – Mar. 1979	$4.73
Apr. – June 1979	$4.37
Jul. – Sept. 1979	$4.16
Oct. – Dec. 1979	$3.92

- 2 different horizontal scales show passage time

During this time	one horizontal inch equals
1973-1978	3.8 years
1979	0.57 years

- With both scales shifting the distortion is multiplicative

Show data variation, not design variation!



National Science Foundation, Science Indicators, 1974 (Washington D.C., 1976), p.15, (Tufte, 1983, p60)

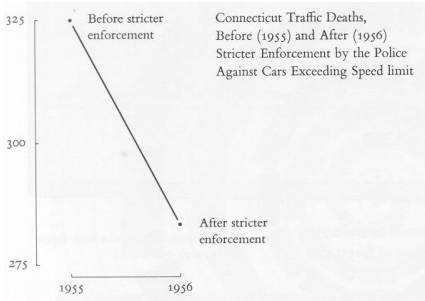
Context is Essential

Graphics must not quote data out of context

Data sparse graphics should provoke suspicion

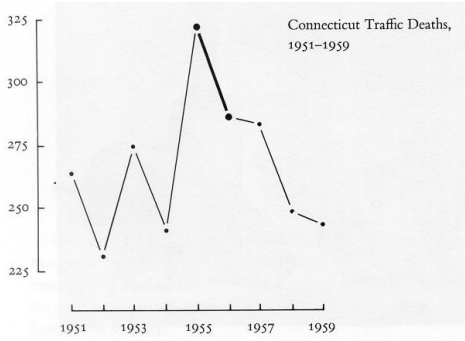
Graphics often lie by omission

Nearly all important questions are left unanswered by this graph



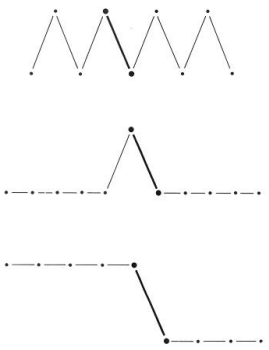
Context is Essential

A few more data points tell a more complete story



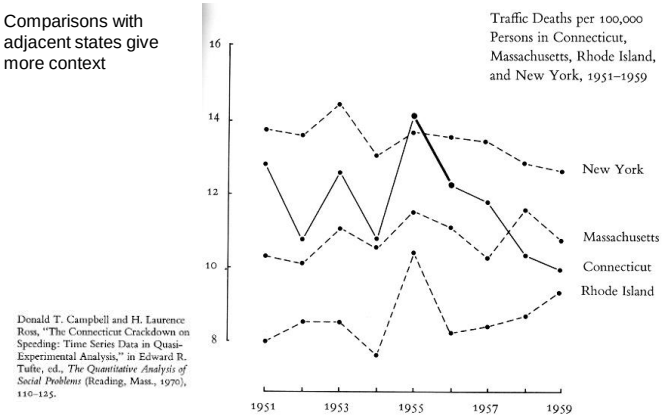
Context is Essential

Different data points would tell a different stories



Context is Essential

Comparisons with adjacent states give more context



Donald T. Campbell and H. Laurence Ross, "The Connecticut Crackdown on Speeding: Time Series Data in Quasi-Experimental Analysis," in Edward R. Tufte, ed., *The Quantitative Analysis of Social Problems* (Reading, Mass., 1970), 110–125.

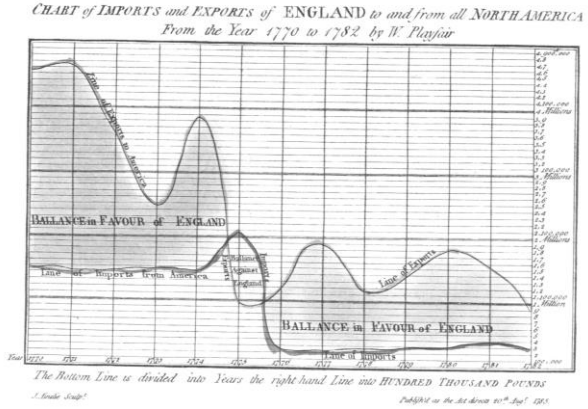
Tufte: Principles of Graphical Excellence

- Graphical excellence is the well-designed presentation of interesting data – a matter of *substance*, *statistics*, and *design*
- Graphical excellence consists of complex ideas communicated with *clarity*, *precision*, and *efficiency*

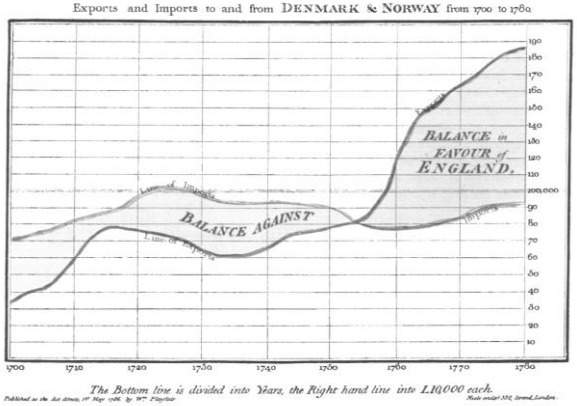
Tufte's principles for better viz?

- Above all else, show the data
- Maximize the data-ink ratio
  - Erase non-data-ink
  - Erase redundant data-ink
- Revise and edit

Above all else, show the data



Above all else, show the data



Maximize the data-ink ratio

$$\text{Data-ink ratio} = \frac{\text{data-ink}}{\text{Total ink used to print graphic}}$$

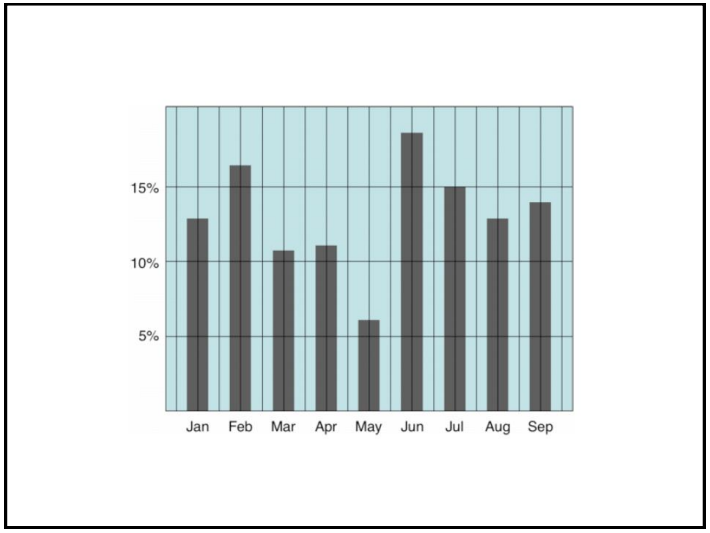
= Proportion of a graphic's ink devoted to the non-redundant display of data-information.

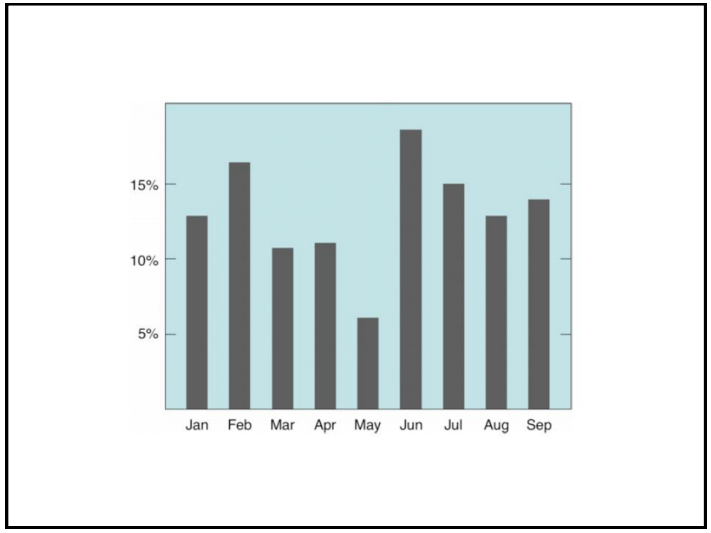
= 1.0 – proportion of graphic that can be erased without the loss of information

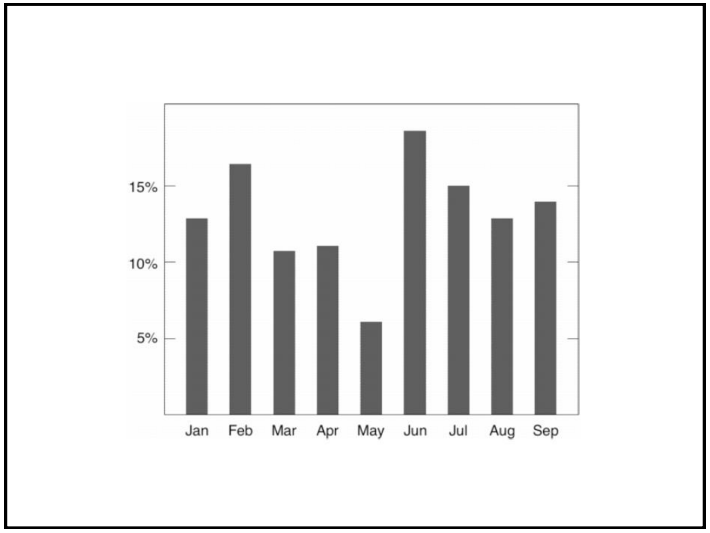
Maximize the data-ink ratio

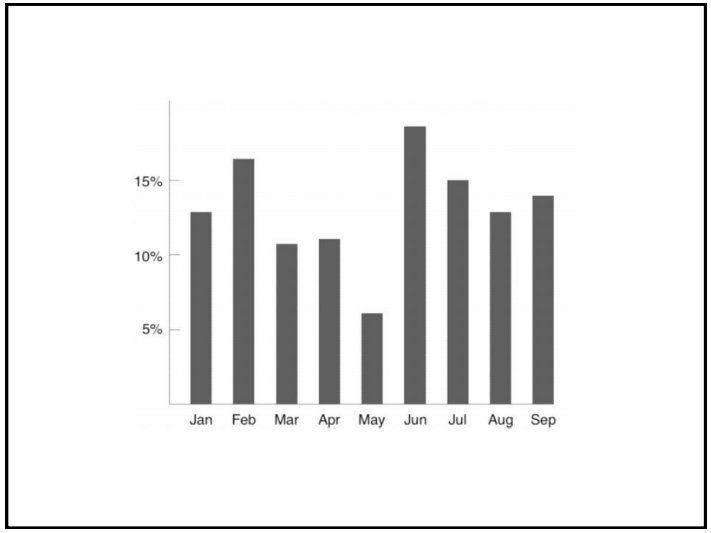
- Within reason
- In essence, you should be able to argue for every pixel

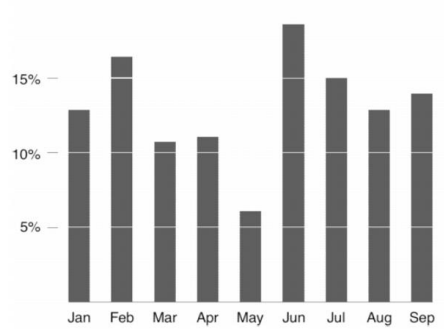
- Starting point:
  - erase non-data ink
  - erase redundant data-ink

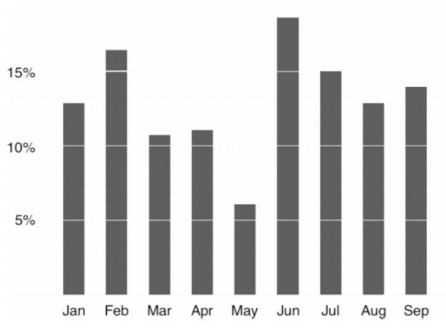

15%
10%
5%
Jan Feb Mar Apr May Jun Jul Aug Sep


15%
10%
5%
Jan Feb Mar Apr May Jun Jul Aug Sep


15%
10%
5%
Jan Feb Mar Apr May Jun Jul Aug Sep


15%
10%
5%
Jan Feb Mar Apr May Jun Jul Aug Sep





Summary

- Show data variation, not design variation
- Avoid using ink for non-data items
- Avoid redundancy
- Clear and detailed labeling should be used to defeat graphical distortion
- Revise and Edit