FACULTY/PRESENTER

FACULTY

BORIS BERNHARDT

RELATIONSHIP WITH COMMERCIAL INTERESTS

NONE

DISCLOSURES

NO CONFLICTS OF INTEREST

MITGATING POTENTIAL BIAS

NO COMMERCIAL BIAS

STATISTICAL TEACHING SESSION

Boris Bernhardt, PhD

http://mica-mni.github.io

INTRODUCTION TO TODAYS SESSION

BORIS BERNHARDT, PHD

CORE CONCEPTS
AND DATA
VISUALIZATION

COLIN JOSEPHSON, MD

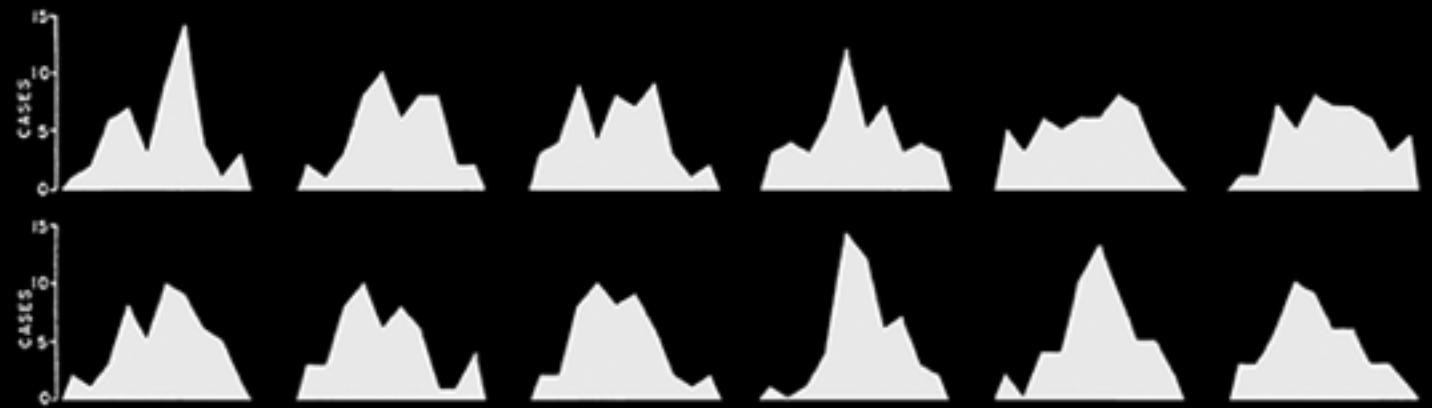
SYSTEMATIC
REVIEWS
AND
META-ANALYSIS

SEOK-JUN HONG, PhD

NEUROIMAGING
BASED
STATISTICS

JORDAN ENGBERS, PhD

MACHINE
LEARNING
TECHNIQUES



CORE CONCEPTS IN STATISTICAL ANALYSIS AND DATA VISUALIZATION

Boris Bernhardt, PhD

http://mica-mni.github.io

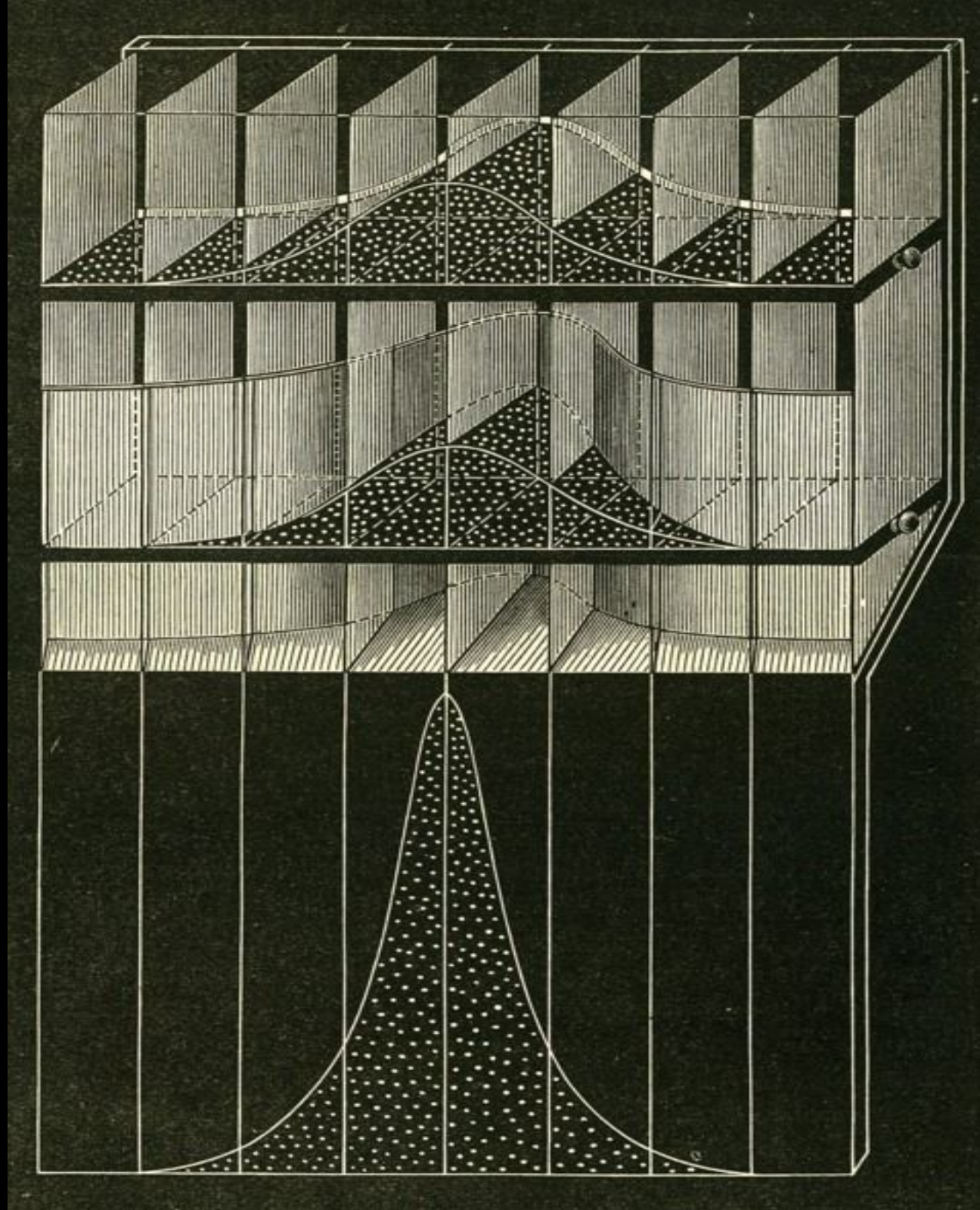
LEARNING OBJECTIVES

BASIC CONCEPTS IN DESCRIPTIVE STATISTICS

GENERALIZED LINEAR MODEL

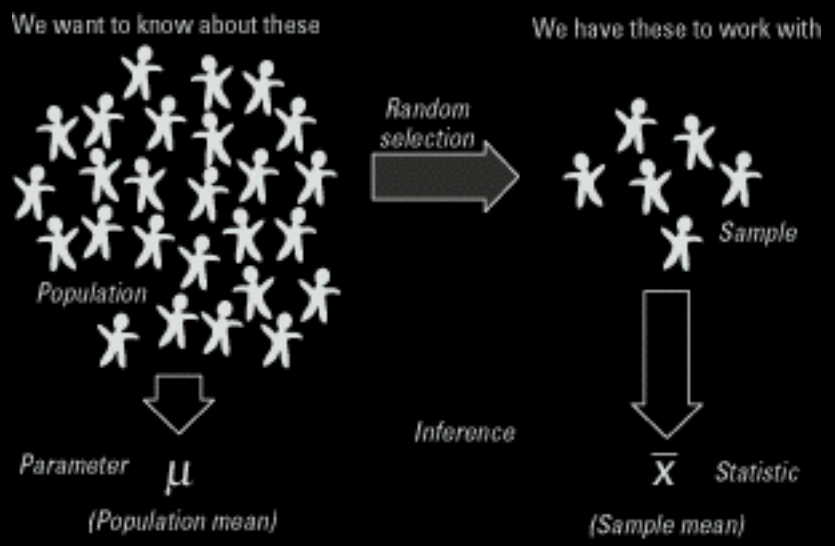
EFFECTIVE AND TRUTHFUL DATA VISUALIZATION

DESCRIBING DATA



PURPOSE OF DESCRIPTIVE STATISTICS

DESCRIBE BASIC FEATURES OF DATA IN A STUDY



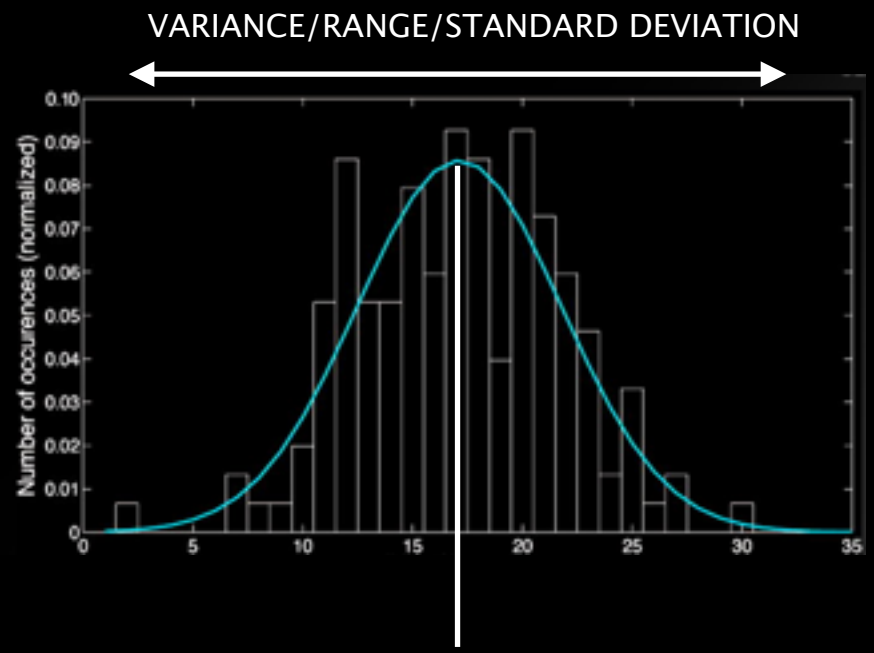
PURPOSE OF DESCRIPTIVE STATISTICS

DESCRIBE BASIC FEATURES OF DATA IN A STUDY IN A COMPACT FORM

MEASUREMENT 1
MEASUREMENT 2

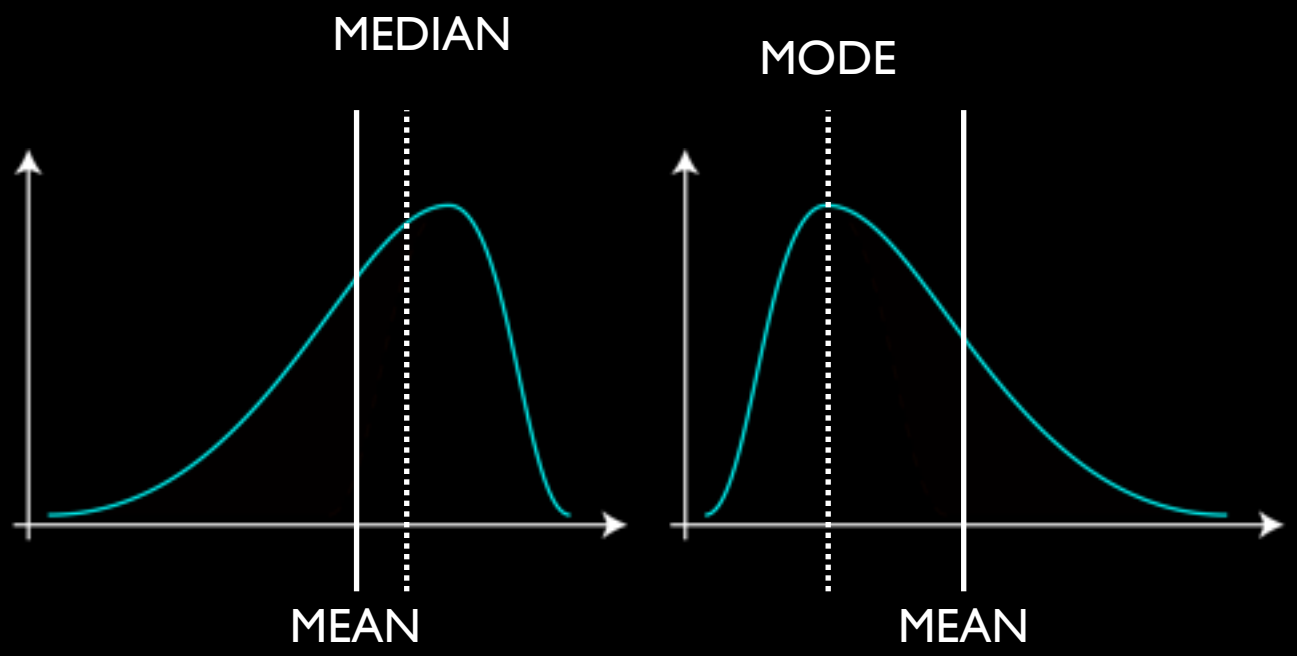
..
..
..
..
..

MEASUREMENT N



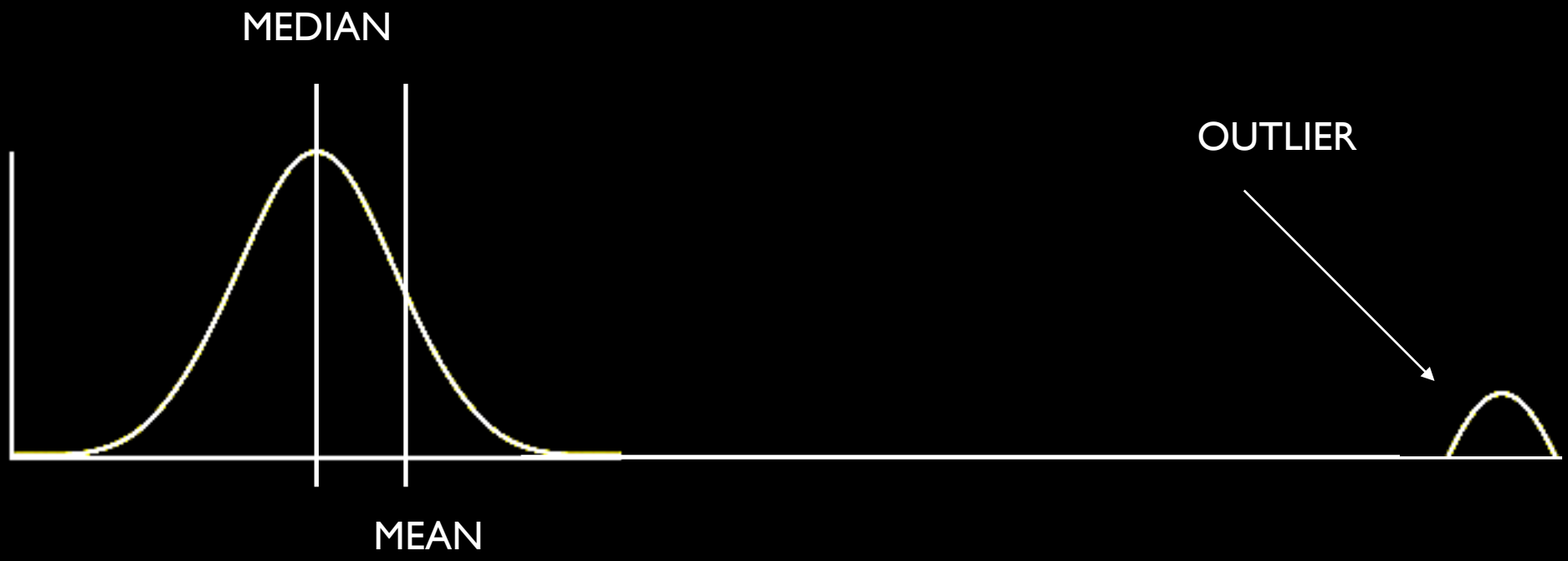
MEAN
SD

CENTRAL TENDENCY



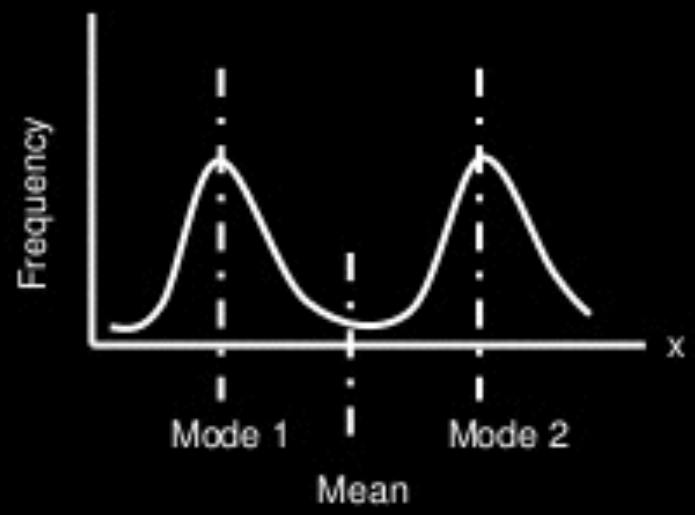
OTHER PARAMETERS SUCH AS THE MEDIAN
MAY AT TIMES PROVIDE YOU BETTER ESTIMATES OF
CENTRAL TENDENCIES

CENTRAL TENDENCY



OTHER PARAMETERS SUCH AS THE MEDIAN
MAY AT TIMES PROVIDE YOU BETTER ESTIMATES OF
CENTRAL TENDENCIES

WHERE THINGS MAY BREAK



OTHER PARAMETERS SUCH AS THE MEDIAN
MAY AT TIMES PROVIDE YOU BETTER ESTIMATES OF
CENTRAL TENDENCIES

TAKE HOME #1

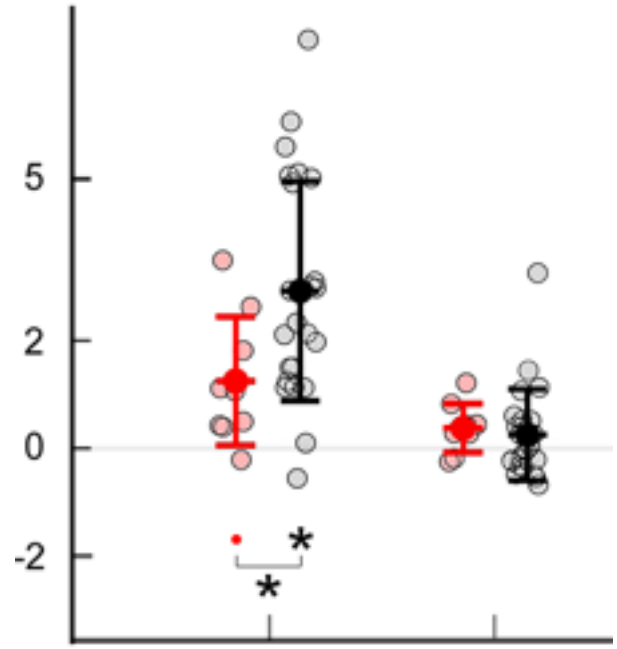
DESCRIPTIVE STATISTICS
INTENDS TO SUMMARIZE DATA

MEAN/SD
ARE POWERFUL PARAMETERS
WHEN DATA COME FROM NORMAL
DISTRIBUTION

<u>HOWEVER:</u>
THEY BECOME LESS APPROPRIATE
WHEN DATA ARE NOT NORMAL
AND ARE SENSITIVE TO OUTLIERS

<u>→VERIFY YOUR DATA</u>

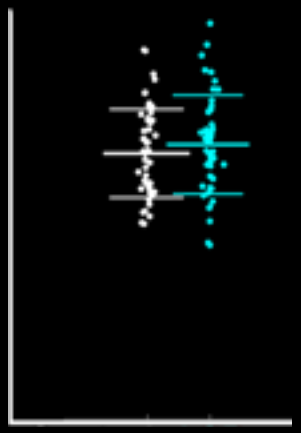
INFERENCES



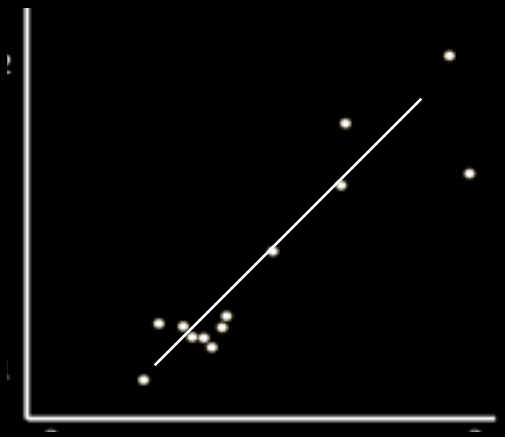
INFERENTIAL STATISTICS

MANY DIFFERENT STATISTICAL PROBLEMS

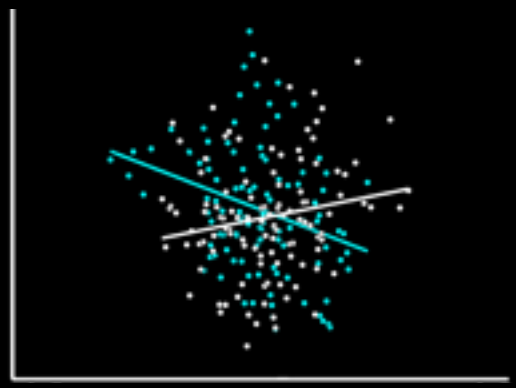
GROUP DIFFERENCE



TEST FOR SIGNIFICANCE OF CORRELATION



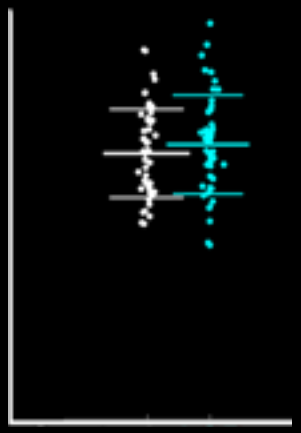
TEST FOR A DIFFERENCE IN CORRELATIONS BETWEEN TWO GROUPS



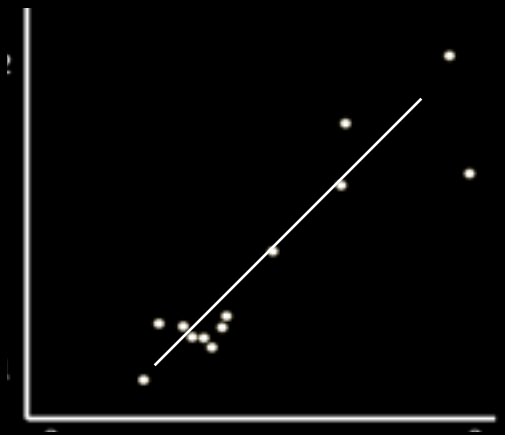
INFERENTIAL STATISTICS

CAN BE ADDRESSED WITH THE GENERALIZED LINEAR MODEL (GLM)

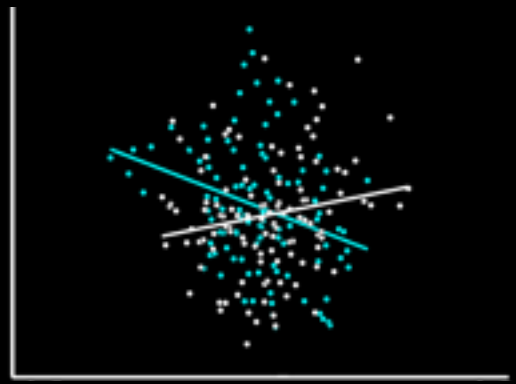
$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_2 \times X_1 + \varepsilon$$



$Y=\beta_0 + \beta_{1*}GROUP$



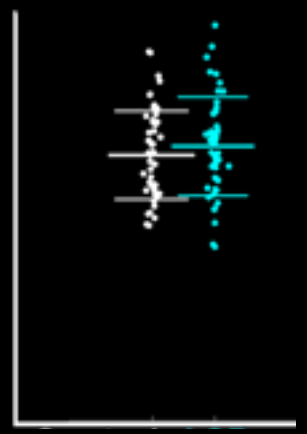
$Y=\beta_0 + \beta_{1*}AGE$



$Y=\beta_0 + \beta_{1*}AGE + \beta_{2*}GROUP + \beta_{3*}AGE \times GROUP$

THE MODEL FURTHERMORE ALLOW TO TEST FOR CONTRASTS

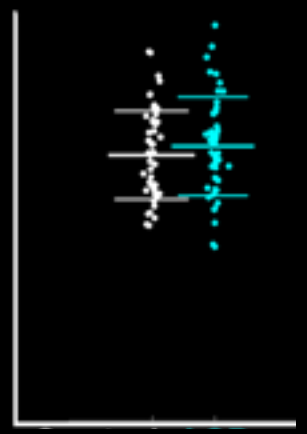
$$Y = \beta_0 + \beta_{1*}\text{GROUP} + \varepsilon$$



CONTRAST = GROUP1–GROUP2

THE MODEL FURTHERMORE ALLOW TO TEST FOR CONTRASTS
AND TO CONTROL FOR VARIABLES OF NO INTEREST

$$Y = \beta_0 + \beta_{1*}AGE + \beta_{2*}GROUP + \varepsilon$$



CONTRAST = GROUP1–GROUP2

T-STATISTIC MEANING FOR A BETWEEN-GROUP CONTRAST

ESTIMATING THE EFFECT FOR A CONTRAST PROVIDES A T-STATISTIC

H0: GROUP1=GROUP2

H1: GROUP1≠GROUP2

T IS

DIFFERENCE
IN MEANS
NORMALIZED BY
POOLED STANDARD
DEVIATION

HIGH T

DIFFERENCE
UNLIKELY TO
ARISE BY CHANCE
IF GROUP1
AND GROU2 WERE
THE SAME

T-STATISTIC MEANING FOR A BETWEEN-GROUP CONTRAST

Entry is $t(A; \nu)$ where $P\{t(\nu) \leq t(A; \nu)\} = A$

	A						
ν	.60	.70	.80	.85	.90	.95	.975
1	0.325	0.727	1.376	1.963	3.078	6.314	12.706
2	0.289	0.617	1.061	1.386	1.886	2.920	4.303
3	0.277	0.584	0.978	1.250	1.638	2.353	3.182
4	0.271	0.569	0.941	1.190	1.533	2.132	2.776
5	0.267	0.559	0.920	1.156	1.476	2.015	2.571
6	0.265	0.553	0.906	1.134	1.440	1.943	2.447
7	0.263	0.549	0.896	1.119	1.415	1.895	2.365
8	0.262	0.546	0.889	1.108	1.397	1.860	2.306
9	0.261	0.543	0.883	1.100	1.383	1.833	2.262
10	0.260	0.542	0.879	1.093	1.372	1.812	2.228
11	0.260	0.540	0.876	1.088	1.363	1.796	2.201
12	0.259	0.539	0.873	1.083	1.356	1.782	2.179
13	0.259	0.537	0.870	1.079	1.350	1.771	2.160
14	0.258	0.537	0.868	1.076	1.345	1.761	2.145
15	0.258	0.536	0.866	1.074	1.341	1.753	2.131
16	0.258	0.535	0.865	1.071	1.337	1.746	2.120
17	0.257	0.534	0.863	1.069	1.333	1.740	2.110
18	0.257	0.534	0.862	1.067	1.330	1.734	2.101
19	0.257	0.533	0.861	1.066	1.328	1.729	2.093
20	0.257	0.533	0.860	1.064	1.325	1.725	2.086
21	0.257	0.532	0.859	1.063	1.323	1.721	2.080
22	0.256	0.532	0.858	1.061	1.321	1.717	2.074
23	0.256	0.532	0.858	1.060	1.319	1.714	2.069
24	0.256	0.531	0.857	1.059	1.318	1.711	2.064
25	0.256	0.531	0.856	1.058	1.316	1.708	2.060
26	0.256	0.531	0.856	1.058	1.315	1.706	2.056
27	0.256	0.531	0.855	1.057	1.314	1.703	2.052
28	0.256	0.530	0.855	1.056	1.313	1.701	2.048
29	0.256	0.530	0.854	1.055	1.311	1.699	2.045
30	0.256	0.530	0.854	1.055	1.310	1.697	2.042
40	0.255	0.529	0.851	1.050	1.303	1.684	2.021
60	0.254	0.527	0.848	1.045	1.296	1.671	2.000
120	0.254	0.526	0.845	1.041	1.289	1.658	1.980
∞	0.253	0.524	0.842	1.036	1.282	1.645	1.960

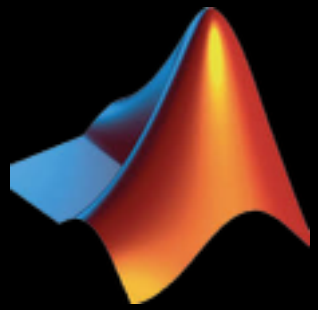
PERCENTILE

DEGREES OF FREEDOM

T-VALUES

SOFTWARE THAT SUPPORTS MODEL BASED INFERENCE

R-PROJECT.ORG



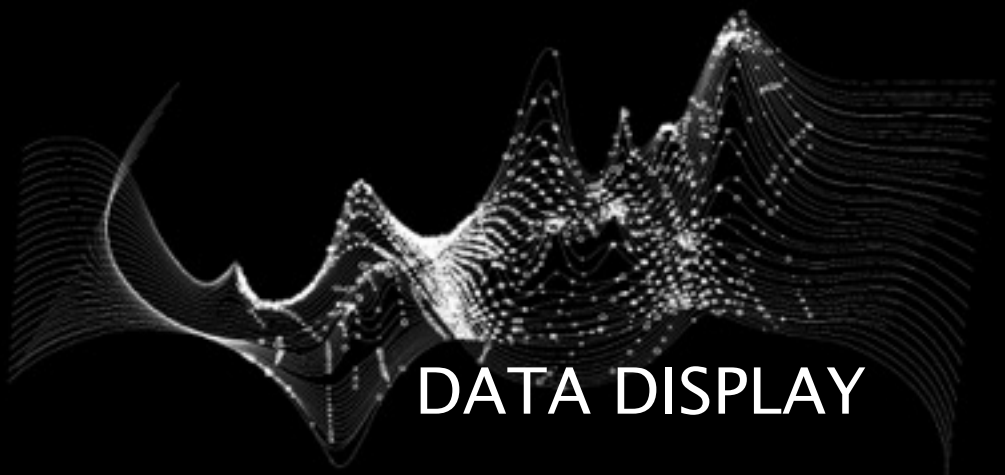
MATH.MCGILL.CA/~KEITH/SURFSTAT

TAKE HOME #2

MANY COMMON STATISTICAL TESTS ARE EXAMPLES OF GLM

GLM ALLOWS YOU TO SPECIFY VARIABLES TO CONTROL FOR

GLM WILL PROVIDE YOU EFFECTS OF CONTRASTS OF INTEREST



DATA DISPLAY

EDWARD TUFTE

STATISTICIAN

PIONEER IN FIELD OF DATA VISUALIZATION

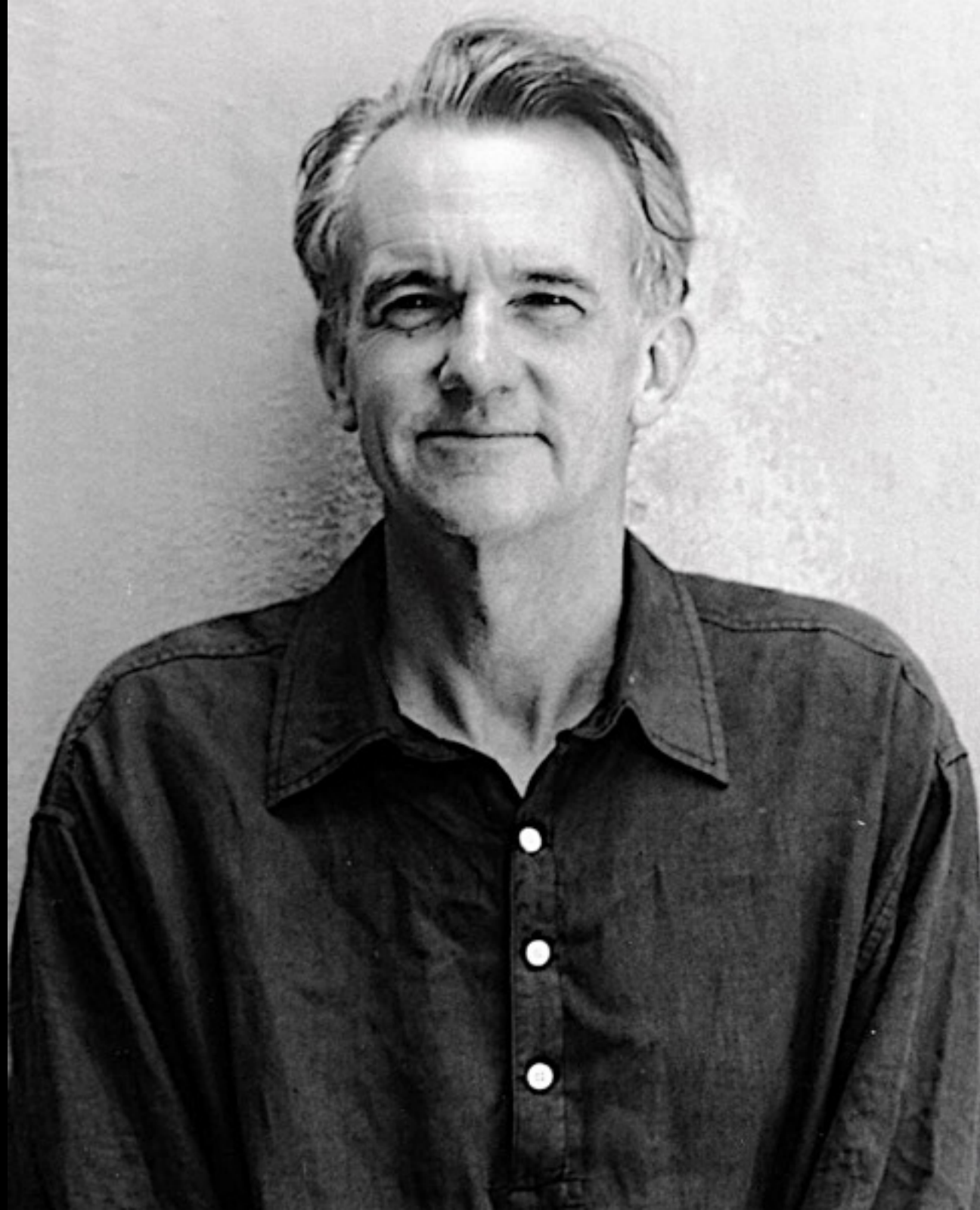
PUBLISHED SEMINAL BOOKS:

“THE VISUAL DISPLAY OF QUANTITATIVE INFORMATION”

“BEAUTIFUL EVIDENCE”

“VISUAL EXPLANATIONS”

“THE COGNITIVE STYLE OF POWERPOINT”



PRINCIPLES OF GRAPHICAL EXCELLENCE

COMPLEX IDEAS COMMUNICATED WITH
CLARITY, PRECISION, EFFICIENCY

GIVES THE READER THE GREATEST NUMBER OF IDEAS
IN SHORTEST TIME WITH LEAST INK IN SMALLEST SPACE

NEARLY ALWAYS MULTIVARIATE

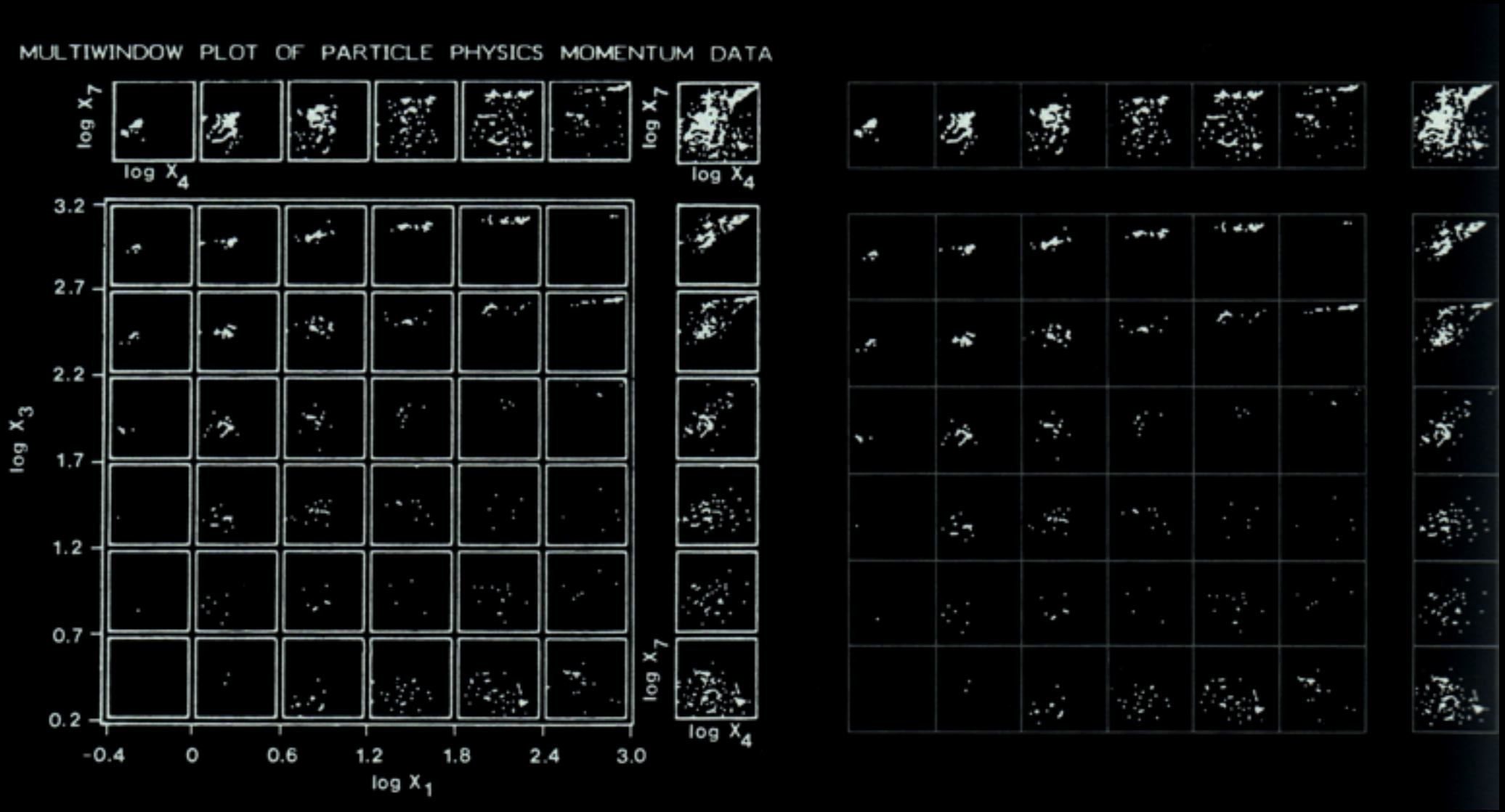
REQUIRES TELLING THE TRUTH ABOUT THE DATA

TUFTE'S DATA INK MEASURES

$$\frac{\text{INK USED TO PORTRAY DATA}}{\text{TOTAL INK}}$$

PORTRAY OF A GRAPHICS INK DEVOTED TO PORTRAY OF NON-REDUNDANT INFORMATION

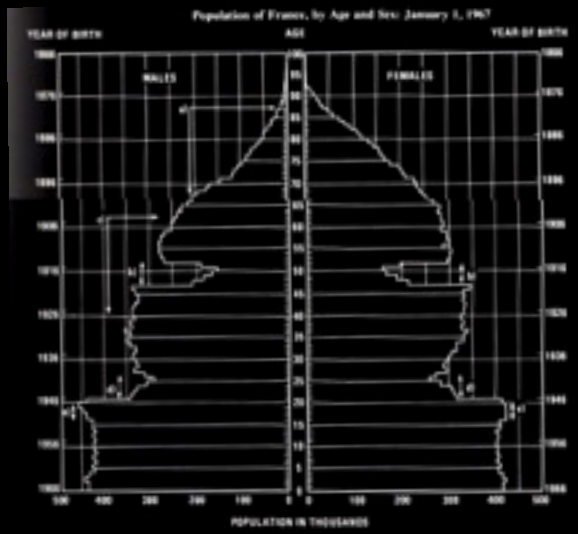
DATA-INK-MAXIMIZATION



From Tufte's book THE VISUAL DISPLAY OF QUANTITATIVE INFORMATION"

DATA-INK-MAXIMIZATION

Elimination of non-data elements and vibrations



A revision quiets the grid and gives emphasis to the data:

Based on data in Institut National de la Statistique et des Études Économiques, *Annuaire statistique de la France, 1968* (Paris, 1968), pp. 32–33; redrawn in Henry S. Shryock and Jacob S. Siegel, *The Methods and Materials of Demography* (Washington, D.C., 1973), vol. 1, 242.

From Tufte's book THE VISUAL DISPLAY OF QUANTITATIVE INFORMATION"

PRINCIPLES OF GRAPHICAL INTEGRITY

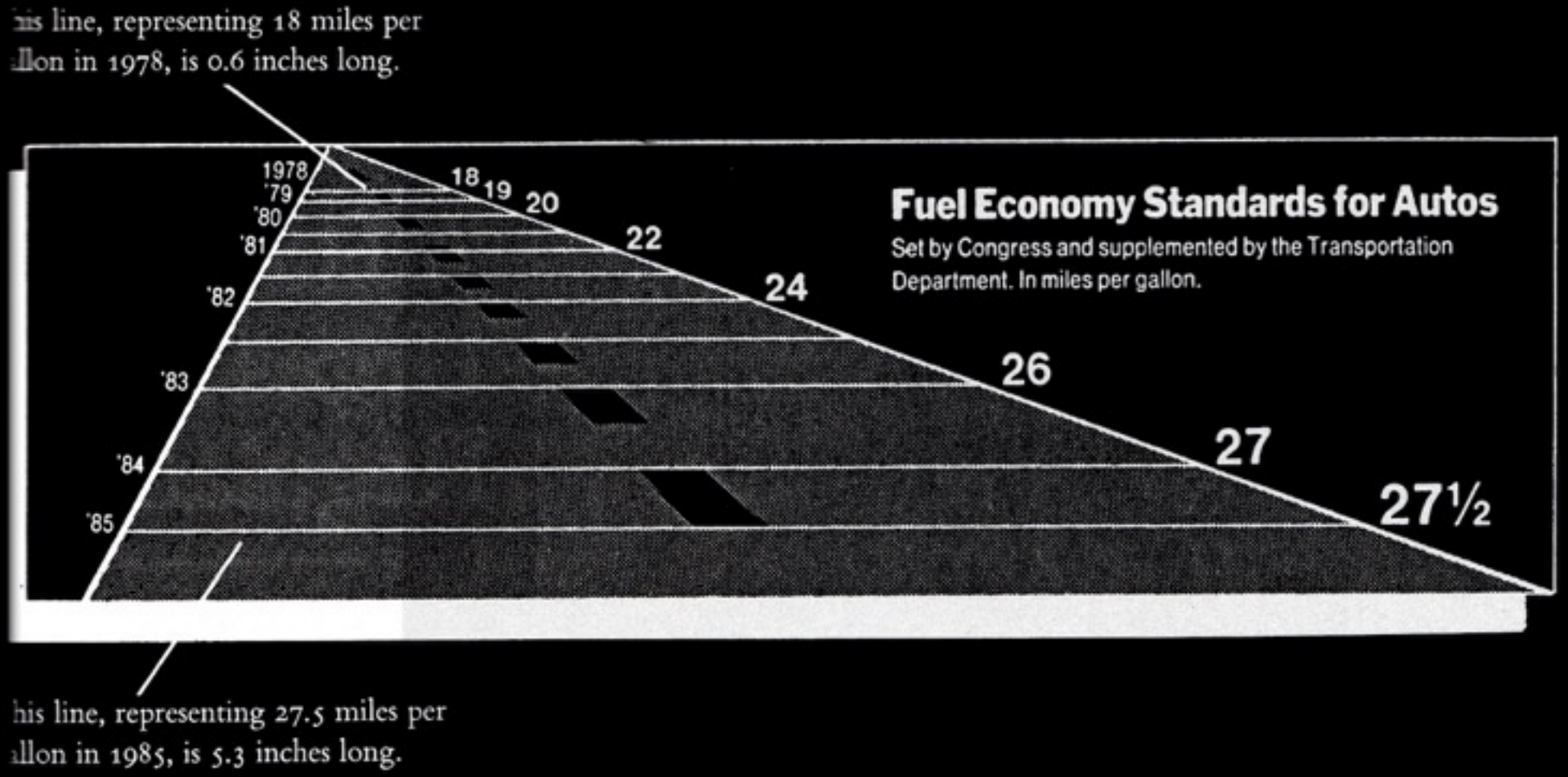
REPRESENTATION SHOULD REFLECT
THE MEASURED QUANTITIES

CLEAR LABELING SHOULD DEFEAT
AMBIGUITY AND DISTORTION

SHOW DATA VARIATION
AND NOT DESIGN VARIATION

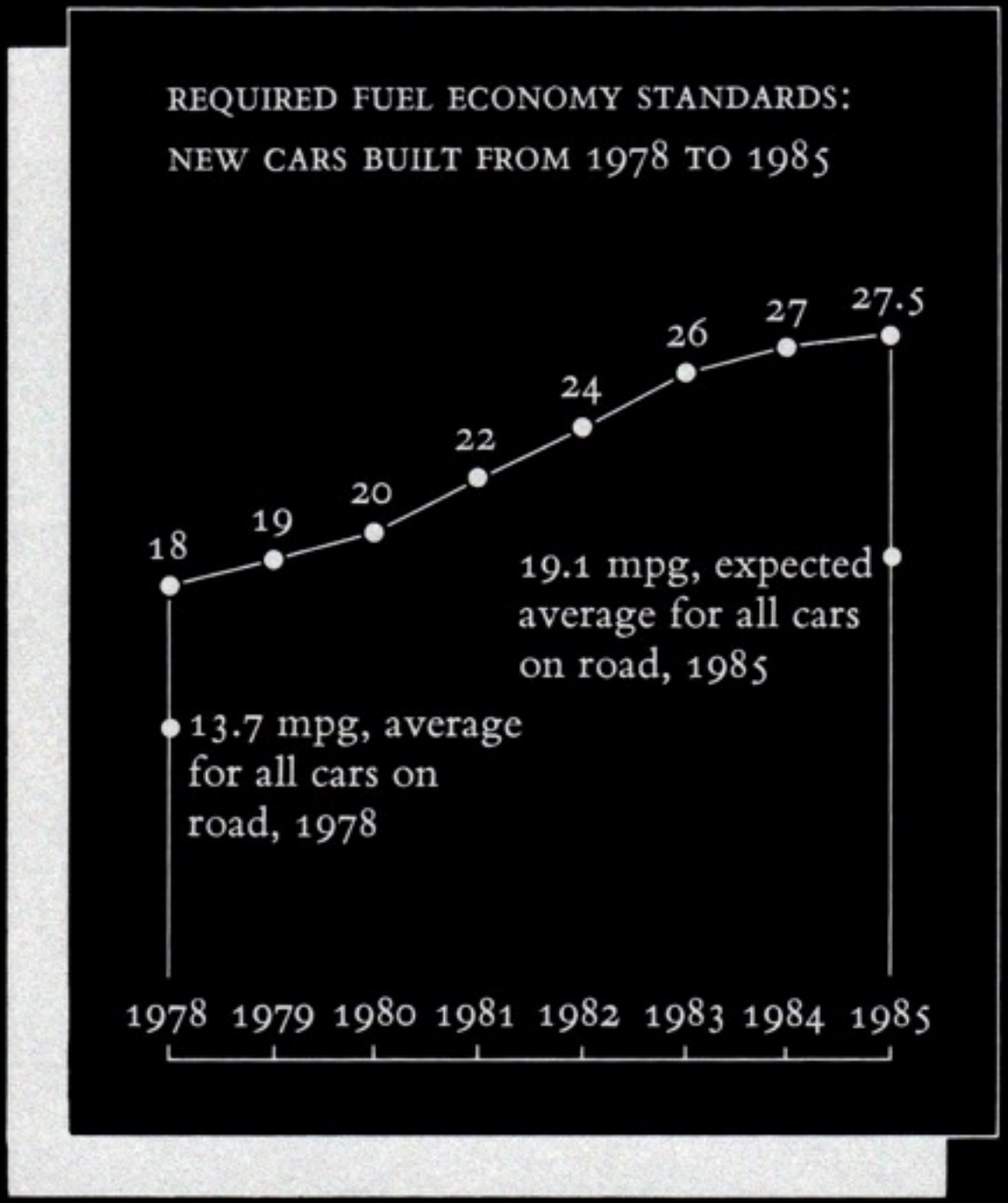
TUFTE'S LIE FACTOR

$$\frac{\text{SIZE OF EFFECT SHOWN IN GRAPHIC}}{\text{SIZE OF EFFECT IN DATA}}$$



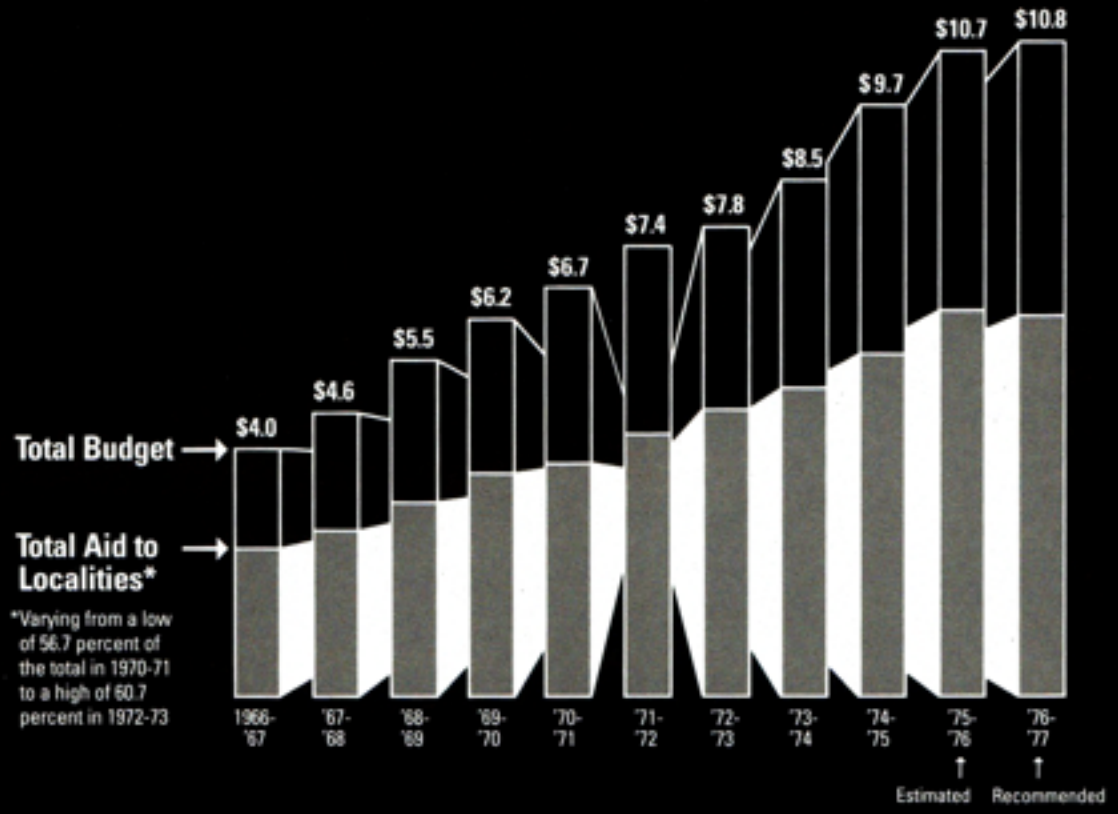
LIE FACTOR = (5.3/0.6) / (27.5/18) = 8.8/1.5=5.7

From Tufte's book THE VISUAL DISPLAY OF QUANTITATIVE INFORMATION"



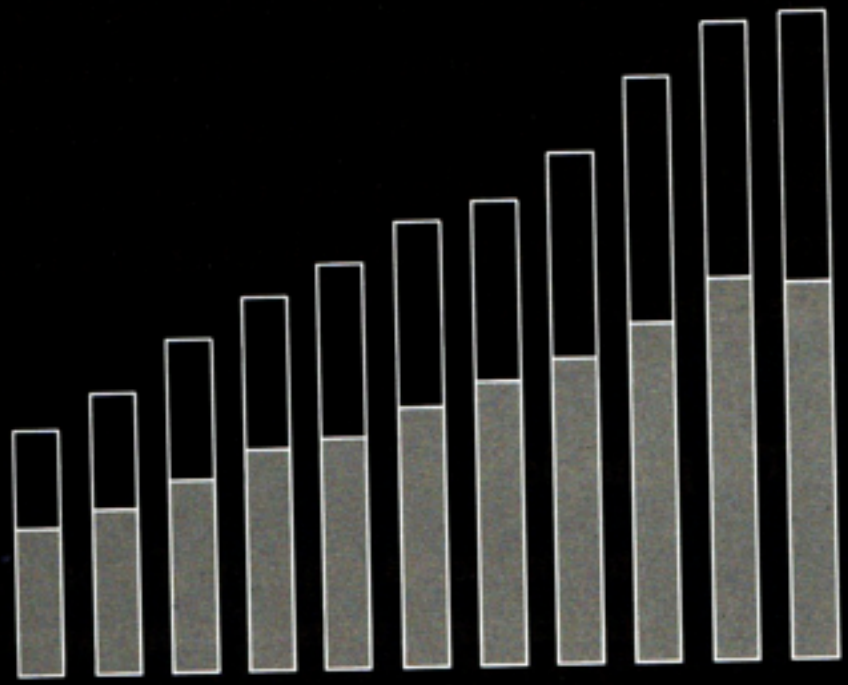
From Tufte's book THE VISUAL DISPLAY OF QUANTITATIVE INFORMATION"

EXPLODING BUDGETS?



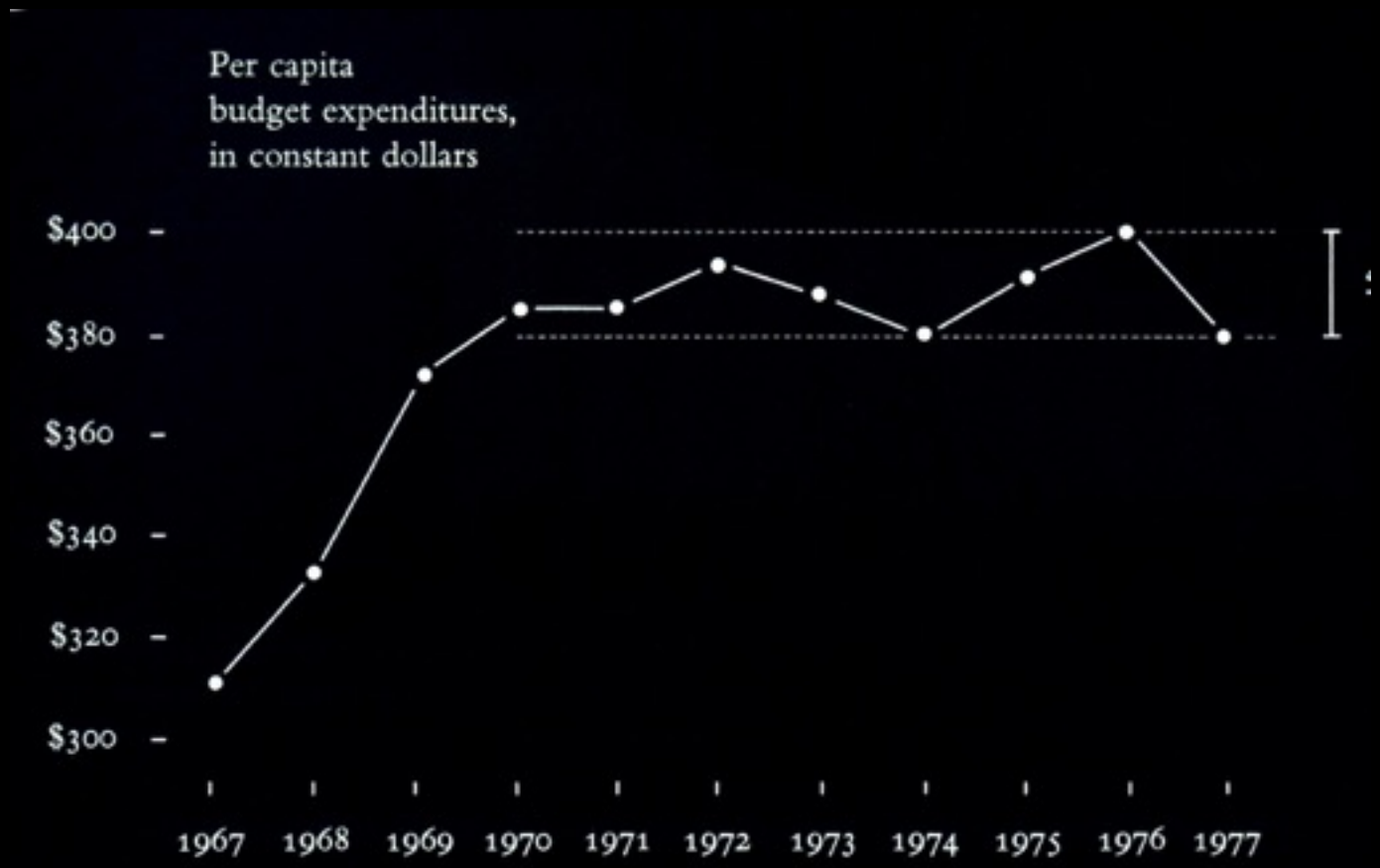
From Tufte's book THE VISUAL DISPLAY OF QUANTITATIVE INFORMATION"

REMOVING VIBRATION



From Tufte's book THE VISUAL DISPLAY OF QUANTITATIVE INFORMATION"

CONTROLLING FOR INFLATION AND POPULATION GROWTH



From Tufte's book THE VISUAL DISPLAY OF QUANTITATIVE INFORMATION"

TUFTE'S ADVICE

ABOVE ALL ELSE SHOW THE DATA

MAXIMIZE DATA-INK RATIO

REVISE AND EDIT

SUMMARY

DISPLAY DATA IN THE BEST WAY POSSIBLE AND LOOK AT IT

DETERMINE WHICH VARIABLES INTEREST YOU

EVALUATE VARIABLES YOU NEED TO CONTROL FOR