

Computing in the Statistics Curriculum

Roger D. Peng

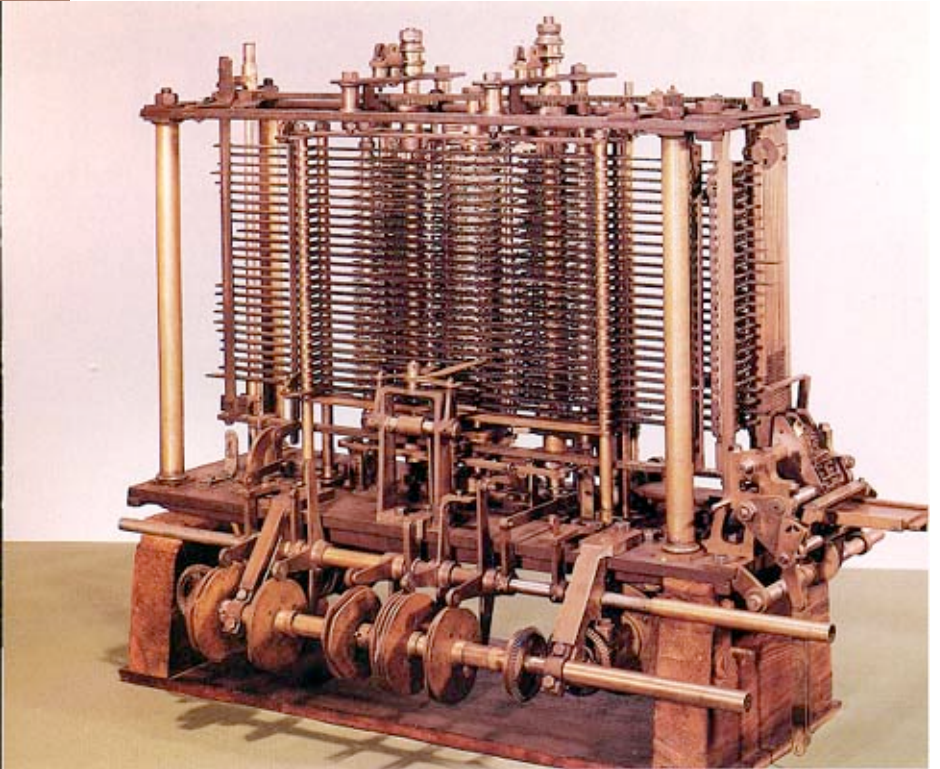
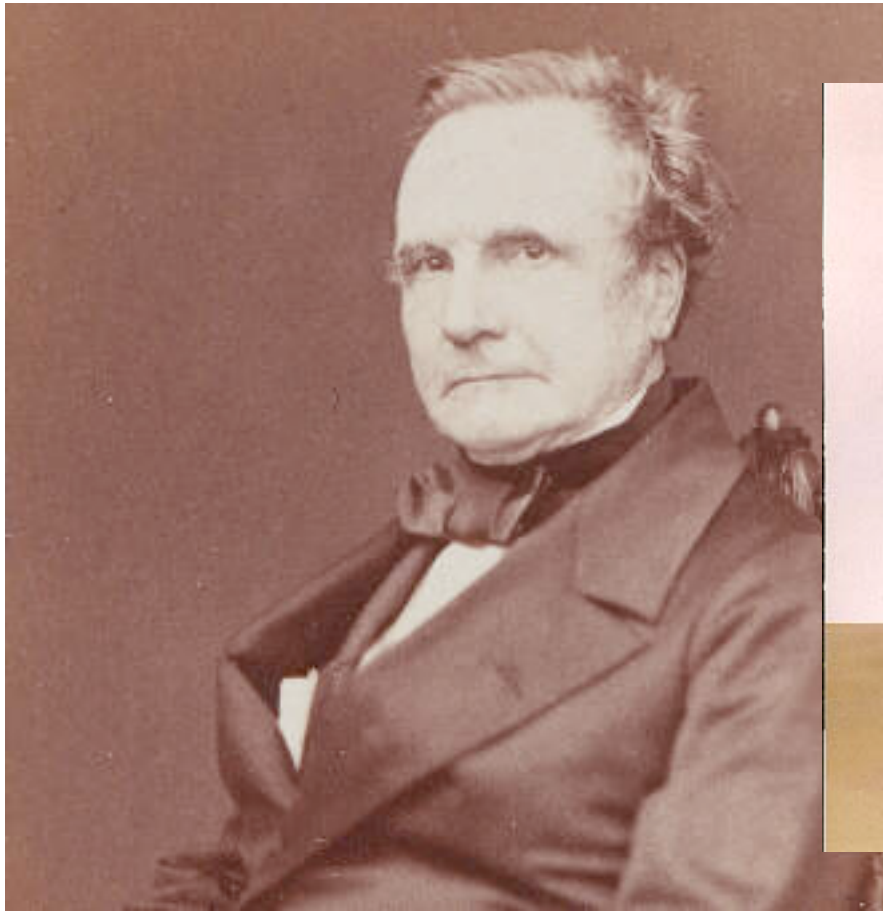
Johns Hopkins Bloomberg School of Public Health

JSM 2008
Denver, CO

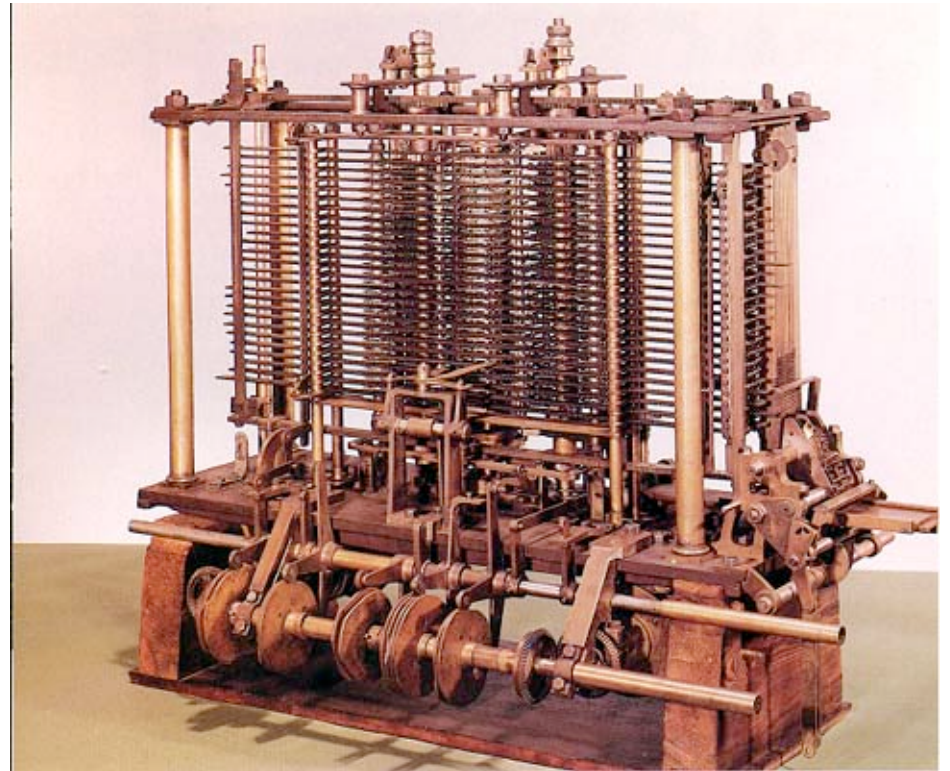
It goes against the grain of modern education to teach children to program. What fun is there in making plans, acquiring discipline in organizing thoughts, devoting attention to detail and learning to be self-critical?

Alan J. Perlis

Computers have been around for a while...



Computers have been around for a while...



Changes in Computing: Then...



...And Now



Statistics Curriculum: Then...

CONTENTS

CHAP.	PAGE
EDITORS' PREFACE	vii
PREFACE TO TENTH EDITION	ix
I. INTRODUCTORY	i
II. DIAGRAMS	24
III. DISTRIBUTIONS	41
IV. TESTS OF GOODNESS OF FIT, INDEPENDENCE AND HOMOGENEITY; WITH TABLE OF χ^2	78
V. TESTS OF SIGNIFICANCE OF MEANS, DIFFERENCES OF MEANS, AND REGRESSION COEFFICIENTS	114
VI. THE CORRELATION COEFFICIENT	175
VII. INTRAClass CORRELATIONS AND THE ANALYSIS OF VARIANCE	211
VIII. FURTHER APPLICATIONS OF THE ANALYSIS OF VARIANCE	248
IX. THE PRINCIPLES OF STATISTICAL ESTIMATION	299
SOURCES USED FOR DATA AND METHODS	336
BIBLIOGRAPHY	340
INDEX	351

TABLES

I. AND II. NORMAL DISTRIBUTION	77
III. TABLE OF χ^2	112
IV. TABLE OF t	174
V.A. CORRELATION COEFFICIENT—SIGNIFICANT VALUES	209
V.B. CORRELATION COEFFICIENT—TRANSFORMED VALUES	210
VI. TABLE OF F	242-247

RA Fisher, *Statistical
Methods for Research
Workers*

SELECTED TABLES
IN MATHEMATICAL STATISTICS

Sponsored by the Institute of Mathematical Statistics

...And now?

8 Hypothesis Testing	345	10 Decision Theory	
8.1 Introduction	345	10.1 Introduction	461
8.2 Methods of Finding Tests	346	10.2 Common Decision The	
8.2.1 Likelihood Ratio Tests	346	10.2.1 Point Estimation	
8.2.2 Invariant Tests	351	10.2.2 Hypothesis Testing	
8.2.3 Bayesian Tests	354	10.2.3 Interval Estimation	
8.2.4 Union-Intersection and Intersection-Union Tests	356	10.3 Decision Theoretic Ba	
8.3 Methods of Evaluating Tests	358	10.3.1 Bayesian Decision I	
8.3.1 Error Probabilities and the Power Function	358	10.3.2 Finding Bayes Rule	
8.3.2 Most Powerful Tests	365	10.4 Admissibility of Decis	
8.3.3 Unbiased and Invariant Tests	370	10.4.1 Comparing Decision	
8.3.4 Locally Most Powerful Tests	376	10.4.2 Finding Admissible	
8.3.5 Sizes of Union-Intersection and Intersection-Union Tests	378	10.4.3 Admissibility of the	
8.4 Other Considerations	381	10.5 Minimax Rules	487
8.4.1 Asymptotic Distribution of LRTs	381	10.6 Invariant Decision Pro	
8.4.2 Other Large-Sample Tests	383	10.7 Stein's Paradox	495
<i>Exercises</i>	385	<i>Exercises</i>	500
<i>Miscellanea</i>	400	<i>Miscellanea</i>	507
9 Interval Estimation	403	11 The Analysis of Variance	
9.1 Introduction	403	11.1 Introduction	509
9.2 Methods of Finding Interval Estimators	406	11.2 The Oneway Analysis	
9.2.1 Inverting a Test Statistic	406	11.2.1 Model and Distrib	
9.2.2 Pivotal Quantities	413	11.2.2 The Classic ANOV	
9.2.3 Guaranteeing an Interval	416	11.2.3 Inferences Regardi	
9.2.4 Bayesian Intervals	422	11.2.4 The ANOVA F Te	
9.2.5 Invariant Intervals	426	11.2.5 Simultaneous Esti	
9.3 Methods of Evaluating Interval Estimators	429	11.2.6 Partitioning Sums	
9.3.1 Size and Coverage Probability	429		
9.3.2 Test-Related Optimality	433		
9.3.3 Invariant Optimality	437		

- 3.2 The method of least squares, 94
 - A General and linear regression models
 - B Weighted least squares
- 3.3 Maximum likelihood estimates, 99
 - A One parameter families
 - B Maximum likelihood in multiparameter families
 - C Maximum likelihood and other methods
- 3.4 Comments, 107
- 3.5 Problems and complements, 108
- 3.6 References, 114

116

4 COMPARISON OF ESTIMATES—OPTIMALITY THEORY

- 4.1 Criteria of estimation, 116
- 4.2 Uniformly minimum variance unbiased estimates, 120
- 4.3 The information inequality, 126
- 4.4 Large sample theory, 132
 - A Consistency
 - B Asymptotic normality and related properties
 - C Asymptotic efficiency and optimality
- 4.5 Unbiased and maximum likelihood estimates. A comparison, 141
- 4.6 Comments, 142
- 4.7 Problems and complements, 142
- 4.8 References, 151

5 FROM ESTIMATION TO CONFIDENCE INTERVALS AND TESTING 153

- 5.1 Precision, confidence intervals, and bounds, 153
 - A The one dimensional case
 - B Confidence regions of higher dimension
 - C Other concepts of confidence regions
- 5.2 The elements of hypothesis testing, 163
 - A Introduction and the Neyman-Pearson framework
 - B The p value: the test statistic as evidence
 - C Power and sample size: indifference regions
- 5.3 Confidence procedures and hypothesis testing, 177
 - A The duality between tests and confidence regions
 - B Confidence intervals and power
 - C Applications of confidence intervals to comparisons and selections
- 5.4 Comments, 184
- 5.5 Problems and complements, 184
- 5.6 References, 191

6 OPTIMAL TESTS AND CONFIDENCE INTERVALS: LIKELIHOOD RATIO TESTS AND RELATED PROCEDURES

- 6.1 The Neyman-Pearson lemma, 192
- 6.2 Uniformly most powerful tests, 198
- 6.3 Uniformly most accurate confidence bounds, 206
- 6.4 Likelihood ratio and related procedures, 209
 - A Tests for the mean of a normal distribution—matched pair experiments
 - B Tests and confidence intervals for the difference in means of two normal populations
 - C The two-sample problem with unequal variances
- 6.5 Likelihood ratio procedures for bivariate normal distributions
 - A Testing independence, confidence intervals for ρ
 - B Tests for the bivariate mean vector
- 6.6 Large sample approximations in testing, 225
 - A Approximations to the distribution of test statistics under
 - B Consistency and local power
- 6.7 Comments, 232
- 6.8 Problems and complements, 233
- 6.9 References, 247

7 LINEAR MODELS—REGRESSION AND ANALYSIS OF VARIANCE

- 7.1 Introduction to the general linear model, 248
 - A Some examples of linear models
 - B Statement and assumptions of the general linear model
 - C What does assuming a linear model mean?
 - D Matrix formulation of the linear model
 - E Related models
- 7.2 Estimation in linear models, 260
 - A The canonical form
 - B Estimation of linear functions of the means: relations to least squares and unbiasedness theory
 - C The variance of linear least squares estimates: the Gauss-Markoff theorem
 - D Estimation of the error variance
 - E Distribution theory: confidence intervals
- 7.3 Tests in linear models, 273
 - A General theory
 - B Linear regression
 - C Analysis of variance models
- 7.4 Simultaneous confidence intervals and multiple comparisons
 - A The Tukey methods
 - B The Scheffé method

Discussing the statistics curriculum

It's personal!

How is the world different today?

- High throughput technologies for collecting vast quantities of data
- Large databases for investigating subtle associations
- Interactive computing with advanced statistical algorithms
- Sophisticated searches across models and variables to identify important risks
- Statisticians working at the interface with science

Statisticians are “part of the problem” (in a good way!)

McCall MN, **Irizarry** RA (2008) Consolidated strategy for the analysis of microarray spike-in data. *Nucleic Acids Research*. To appear.

Meluh PB, Pan X, Yuan DS, Tiffany C, Chen O, Sookhai-Mahadeo S, Wang X, Peyser BD, **Irizarry** RA, Spencer FA, Boeke JD (2008) Analysis of genetic interactions on a genome-wide scale in budding yeast: diploid-based synthetic lethality analysis by microarray. *Methods in Molecular Biology* 416:221-247.

Bjornsson HT, Albert TJ, Ladd-Acosta CM, Green RD, Rongione MA, Middle CM, **Irizarry** RA, Broman KW, Feinberg AP (2008) SNP-specific array-based allele-specific expression analysis. *Genome Research*. To appear.

Lin S, Carvalho B, Cutler D, Arking D, Chakravarti A, **Irizarry** RA (2008) Validation and Extension of an Empirical Bayes Method for SNP Calling on Affymetrix Microarrays. *Genome Biology*. To appear.

Irizarry RA, Ladd-Acosta C, Carvalho B, Wu H, Brandenburg SA, Wen B, Feinberg AP (2008) Comprehensive High-throughput Arrays for Restriction endonuclease-based Methylation (CHARM). *Genome Research*. To appear

Rodriguez-Quinones JF, **Irizarry** RA, Diaz-Blanco NL, Rivera-Molina FE, Gomez-Garzon D, Rodriguez-Medina JR (2008) Global mRNA expression analysis in myosin II deficient strains of *Saccharomyces cerevisiae* reveals an impairment of cell integrity functions. *BMC Genomics* 9(1):34

Bengtsson H, **Irizarry** R, Carvalho B, Speed TP (2008) Estimation and assessment of raw copy numbers at the single locus level. *Bioinformatics*. To appear.

Gopi Goswami, Jun S. Liu, Wing H. Wong (2007)

Evolutionary Monte Carlo Methods for Clustering.

Journal of Computational & Graphical Statistics, Vol. 16, No. 4, pp.855-876. [\[preprint\]](#)

Qing Zhou, Hiram Chipperfield, Douglas A Melton, Wing Hung Wong (2007)

A gene regulatory network in mouse embryonic stem cells.

Proc. Natl. Acad. Sci. USA, 104:16438-16443. doi_10.1073_pnas.0701014104. [\[online\]](#)

Qing Zhou and Wing Hung Wong (2007)

Coupling hidden Markov models for the discovery of cis-regulatory modules in multiple species.

Annals of Applied Statistics, 1:36-65. DOI:10.1214/07-AOAS103. [\[preprint\]](#) [\[software\]](#)

Steven A. Vokes, Hongkai Ji, Scott McCuine, Toyoaki Tenzen, Shane Giles, Sheng Zhong, William J. R. Longabaugh, Eric H. Davidson, Wing H. Wong and Andrew P. McMahon (2007)

Genomic characterization of Gli-activator targets in sonic hedgehog-mediated neural patterning.

Development, 134, 1977-1989. doi: 10.1242/dev.001966. [\[online\]](#) [\[in the news\]](#)

Karen Kapur, Yi Xing, Zhengqing Ouyang and Wing Hung Wong (2007)

Exon array assessment of gene expression.

Genome Biology, 2007, 8:R82. doi:10.1186/gb-2007-8-5-r82. [\[online\]](#)

Yi Xing, Zhengqing Ouyang, Karen Kapur, Matthew P. Scott, Wing Hung Wong (2007)

Assessing the Conservation of Mammalian Gene Expression Using High-density Exon Arrays.

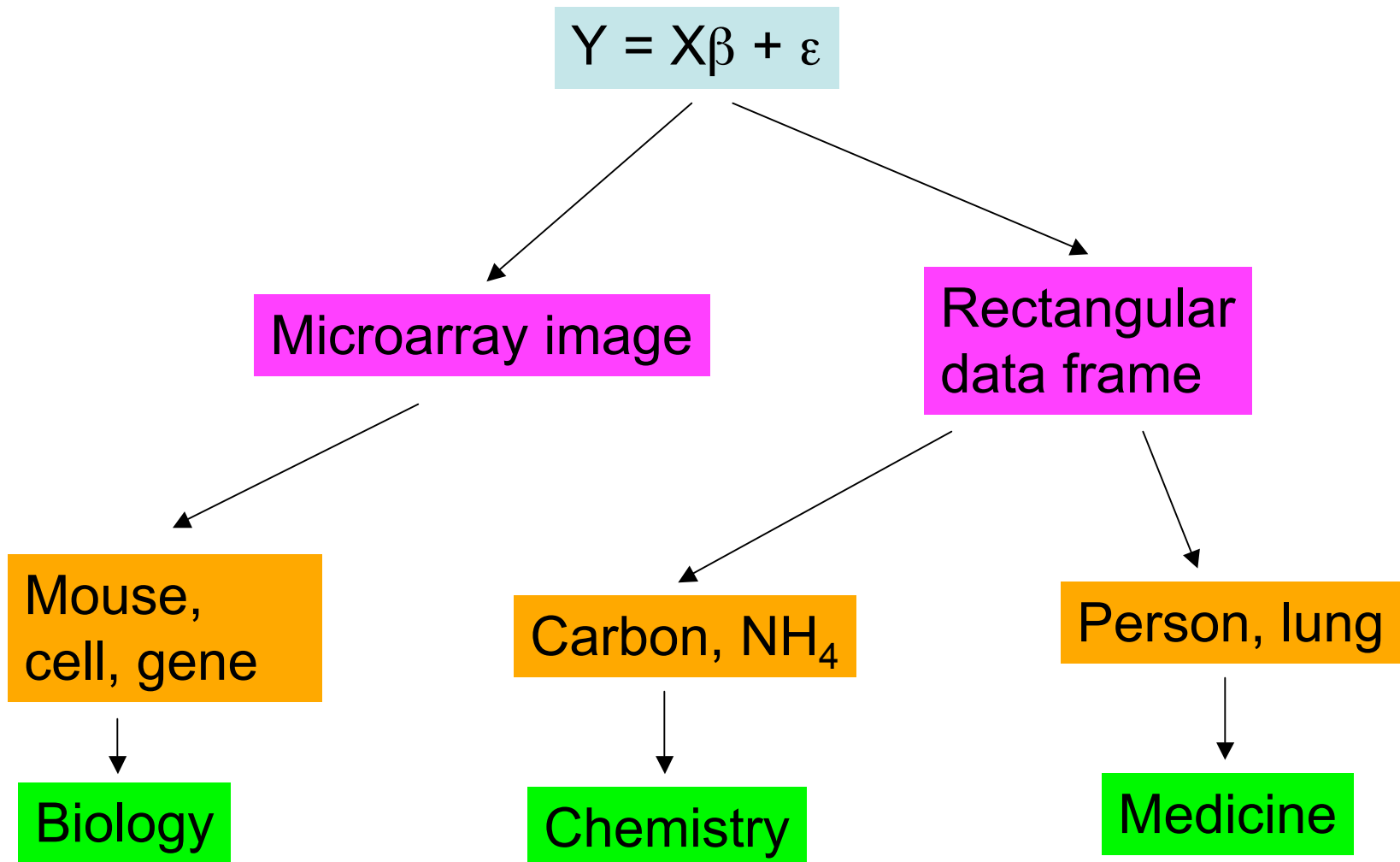
Molecular Biology and Evolution, 2007 24(6):1283-1285; doi:10.1093/molbev/msm061. [\[online\]](#) [\[Supplementary Data\]](#)

Ji-Hye Paik, Ramya Kollipara, Gerald Chu, Hongkai Ji, Yonghong Xiao, Zhihu Ding, Lili Miao, Zuzana Tothova, James W. Horner, Daniel R. Carrasco, Shan Jiang, D. Gary Gilliland, Lynda Chin, Wing H. Wong, Diego H. Castrillon, and Ronald A. DePinho (2007)

FoxOs Are Lineage-Restricted Redundant Tumor Suppressors and Regulate Endothelial Cell Homeostasis.

Cell, Vol 128, 309-323. DOI 10.1016/j.cell.2006.12.029 . [\[online\]](#)

Where do statisticians belong?



Statistician's toolbelt grows

- A facility with computational tools is becoming necessary to interact with people doing cutting edge science
 - databases
 - web services, XML
- Not everything can be crammed into a rectangular data frame
- “It's a poor workman who blames his tools (or lack thereof)”

Statistician as scientist

- Courses in computing can be used to train students to act like scientists rather than automatons
- We can collect our own data
- To interact with data, we need data technologies

“I must find out where my people are going so that I can lead them”

- Complex data are being generated in all areas and new technologies are being applied to deal with them
- Other fields are getting sophisticated
 - e.g. Majors/PhDs in bioinformatics or statistical genetics
- Should we lead or let others show us the way?

4. Time Series	54
Milk, Tea, and Coffee (Acquire and Parse)	55
Cleaning the Table (Filter and Mine)	55
A Simple Plot (Represent and Refine)	57
Labeling the Current Data Set (Refine and Interact)	59
Drawing Axis Labels (Refine)	62
Choosing a Proper Representation (Represent and Refine)	73
Using Rollovers to Highlight Points (Interact)	76
Ways to Connect Points (Refine)	77
Text Labels As Tabbed Panes (Interact)	83
Interpolation Between Data Sets (Interact)	87
End of the Series	92
5. Connections and Correlations	94
Changing Data Sources	94
Problem Statement	95
Preprocessing	96
Using the Preprocessed Data (Acquire, Parse, Filter, Mine)	111
Displaying the Results (Represent)	118
Returning to the Question (Refine)	121
Sophisticated Sorting: Using Salary As a Tiebreaker (Mine)	126
Moving to Multiple Days (Interact)	127
Smoothing Out the Interaction (Refine)	132
Deployment Considerations (Acquire, Parse, Filter)	133
6. Scatterplot Maps	145
Preprocessing	145
Loading the Data (Acquire and Parse)	155
Drawing a Scatterplot of Zip Codes (Mine and Represent)	157
Highlighting Points While Typing (Refine and Interact)	158
Show the Currently Selected Point (Refine)	162
Progressively Dimming and Brightening Points (Refine)	165
Zooming In (Interact)	167
Changing How Points Are Drawn When Zooming (Refine)	177
Deployment Issues (Acquire and Refine)	178
Next Steps	180

7. Trees, Hierarchies, and Recursion	
Using Recursion to Build a Directory Tree	
Using a Queue to Load Asynchronously (Interact)	
An Introduction to Treemaps	
Which Files Are Using the Most Space?	
Viewing Folder Contents (Interact)	
Improving the Treemap Display (Refine)	
Flying Through Files (Interact)	
Next Steps	
8. Networks and Graphs	
Simple Graph Demo	
A More Complicated Graph	
Approaching Network Problems	
Advanced Graph Example	
Mining Additional Information	
9. Acquiring Data	
Where to Find Data	
Tools for Acquiring Data from the Internet	
Locating Files for Use with Processing	
Loading Text Data	
Dealing with Files and Folders	
Listing Files in a Folder	
Asynchronous Image Downloads	
Using <code>openStream()</code> As a Bridge to Java	
Dealing with Byte Arrays	
Advanced Web Techniques	
Using a Database	
Dealing with a Large Number of Files	
10. Parsing Data	
Levels of Effort	
Tools for Gathering Clues	
Text Is Best	
Text Markup Languages	

B Fry. *Visualizing Data*

What are other fields doing?

Washington University in St. Louis

School of Medicine

- “This PhD program [in statistical genetics]...offers an interdisciplinary approach to preparing future scientists with analytical/statistical, computational, and human genetic methods for the study of human disease.”

USC Keck School of Medicine

- “The objective of the PhD program [in statistical genetics] is to produce a statistical geneticist or genetic epidemiologist with in-depth statistical and analytic skills in biostatistics, computational methods and the molecular biosciences.”

What are we doing?

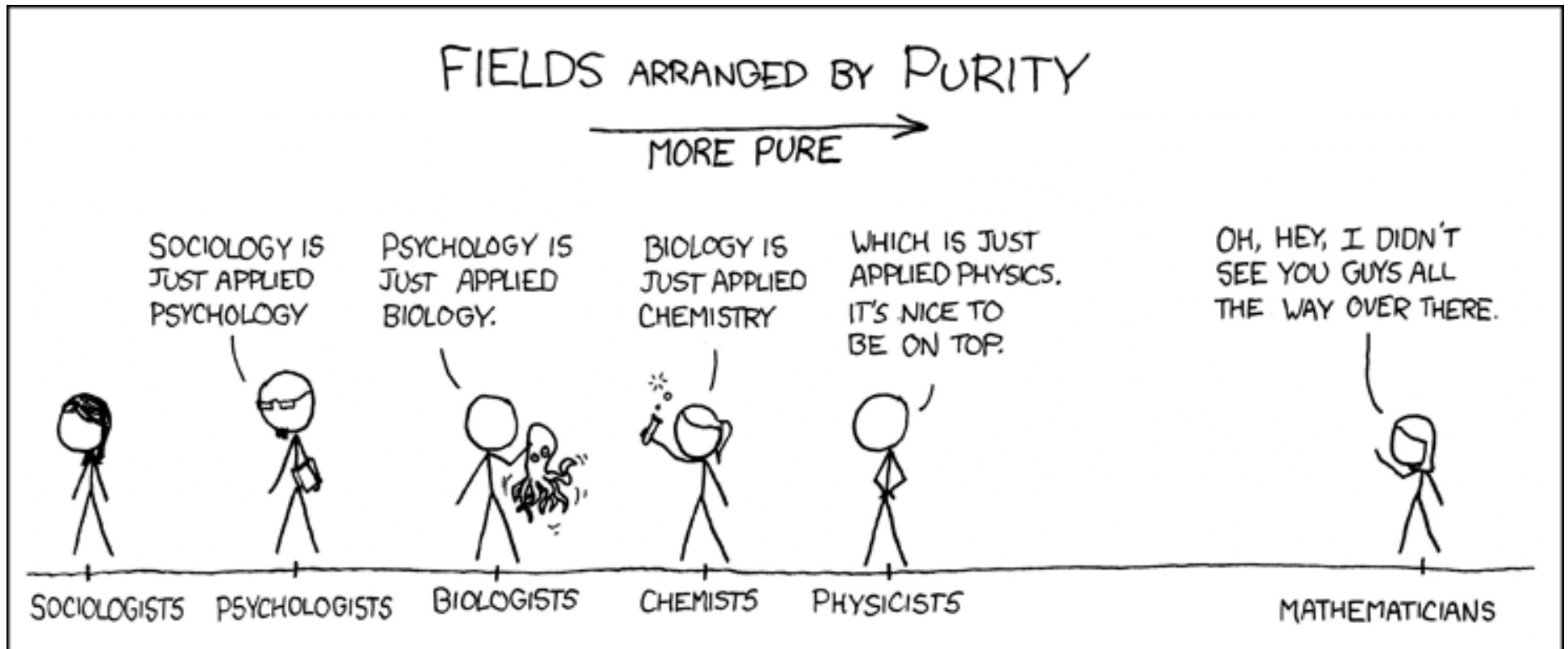
JHSPH Biostatistics

- “The PhD program of the Johns Hopkins Department of Biostatistics provides training in the theory of probability and...biostatistical methodology. The program is unique in its emphasis on...requiring its graduates to complete *rigorous training in real analysis-based probability and statistics, equivalent to what is provided in most departments of mathematical statistics.*”

UC Davis Statistics

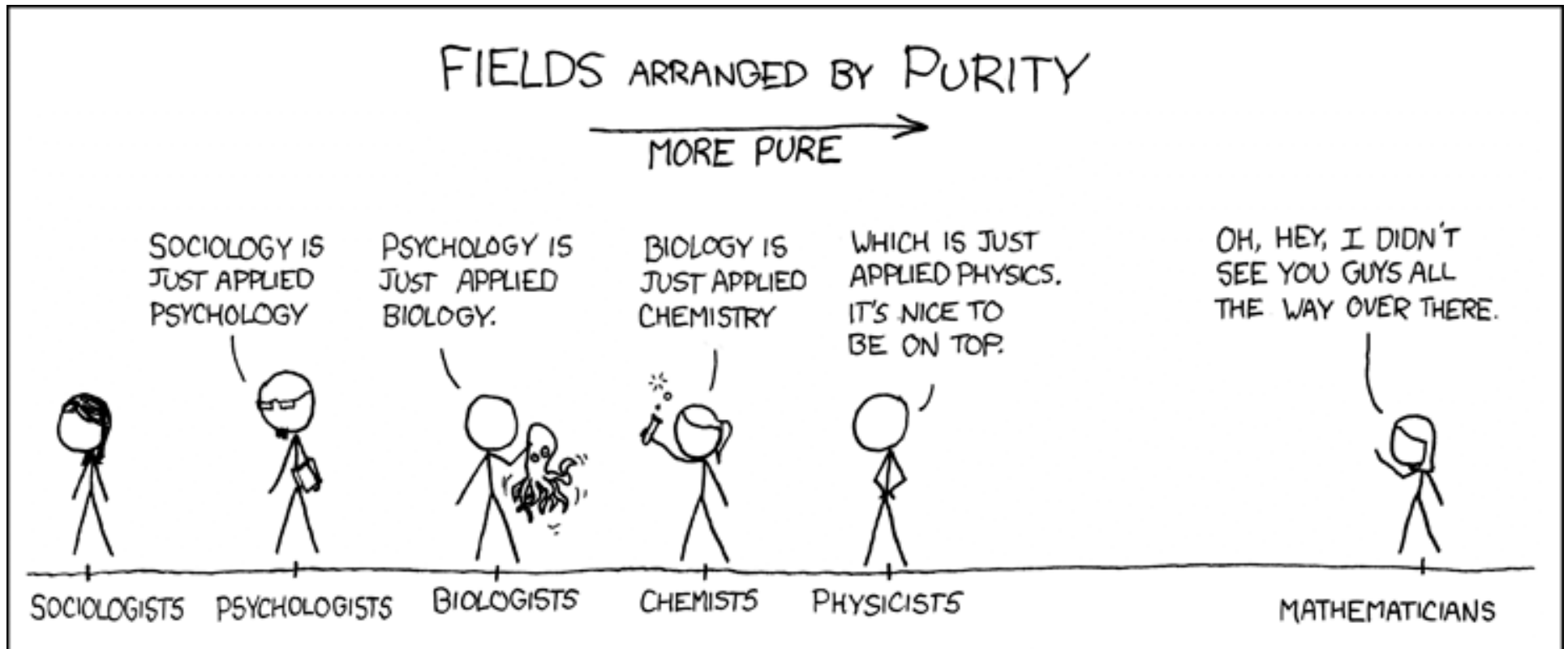
- “the core program for every graduate student in statistics includes graduate level core courses in mathematical statistics, applied statistics and multivariate analysis. Students obtain training in computational statistics and can choose from a variety of special topics courses.”

Where do statisticians belong?



Where do statisticians belong?

Statisticians



Obstacles

- Institutional: Curriculum development slow and narrow in focus (also Gibson's Law)
- Views
 - Computing can be self taught and picked up as you go
 - Computing is just a skill and should not be part of the curriculum
- Faculty training: We are not taught this; it's not natural for us like math

Obstacles (cont'd)

- It's easy to add material to the curriculum, but we can't keep students in school forever
 - What material do we subtract?
 - Is computing part of the “core” or is it “extra”?
- Resource allocation: faculty who are teaching computing to 20 students could be teaching Intro Stat to 200 students

Who can teach this?

- Statisticians with a strong computing focus appear “randomly” in the field
- Can we depend on this point process forever?
 - No: $\lambda(t)$ is going to 0.
- These people will continue to appear but there may not be a compelling reason for them to go into statistics (or be in a statistics department)

Can we depend on other departments?

- I'm not sure....
- Engage CS departments to tailor courses for us?
- Political reasons

Mathematics

Calculus I & II 110.106-107 or 110.108-109

Chemistry (for class of 2004-2006)

Introductory Chemistry I 030.101
Introductory Organic Chemistry 030.104
Introductory Chemistry Lab I & II 030.105-106
Intermediate Organic Chemistry 030.201
Intermediate Chemistry 030.204
Organic Chemistry Lab 030.225

Chemistry (for class of 2007 and later)

Introductory Chemistry I 030.101
Introductory Chemistry II 030.102
Introductory Chemistry Lab I & II 030.105-106
Introductory Organic Chemistry I 030.205
Introductory Organic Chemistry II 030.206
Introductory Organic Chemistry Lab 030.225

Biology

General Biology I & II 020.151-152
Biochemistry 020.305
Cell Biology 020.306
Biochemistry Lab 020.315
Cell Biology Lab 020.316
Genetics 020.330
Developmental Biology 020.363
Genetics Lab or 020.340
Developmental Biology Lab 020.373

Physics

General Physics 171.103-104 or 171.101-102
General Physics Lab 173.111-112

JHU BA Program
in Biology (core
courses)

We can just conduct one big observational experiment and see who wins.

*Some fields manage to absorb change, but
withstand progress.*

Alan J. Perlis (adapted)