Computing in the Statistics Curriculum

Roger D. Peng

Johns Hopkins Bloomberg School of Public Health

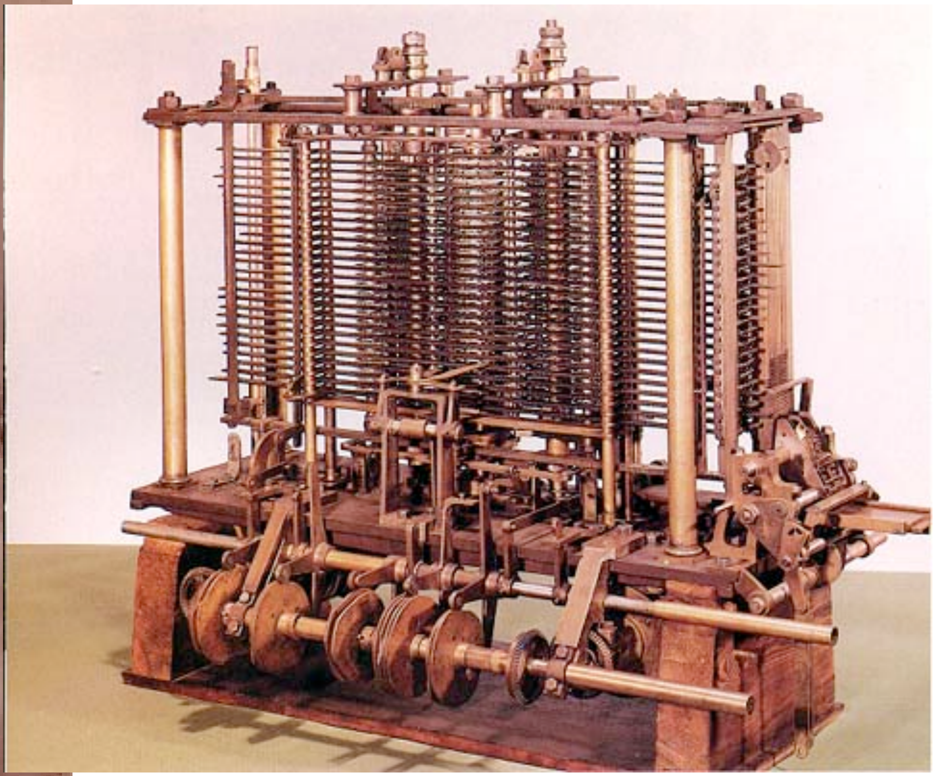
JSM 2008

Denver, CO

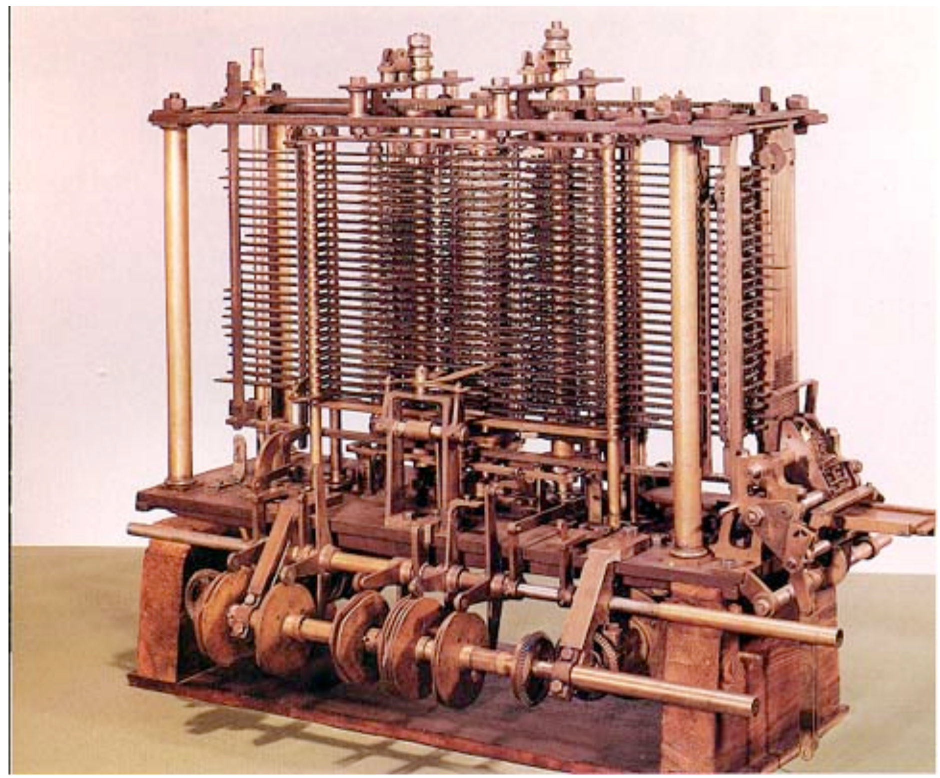
It goes against the grain of modern education to teach children to program. What fun is there in making plans, acquiring discipline in organizing thoughts, devoting attention to detail and learning to be self-critical?

Alan J. Perlis

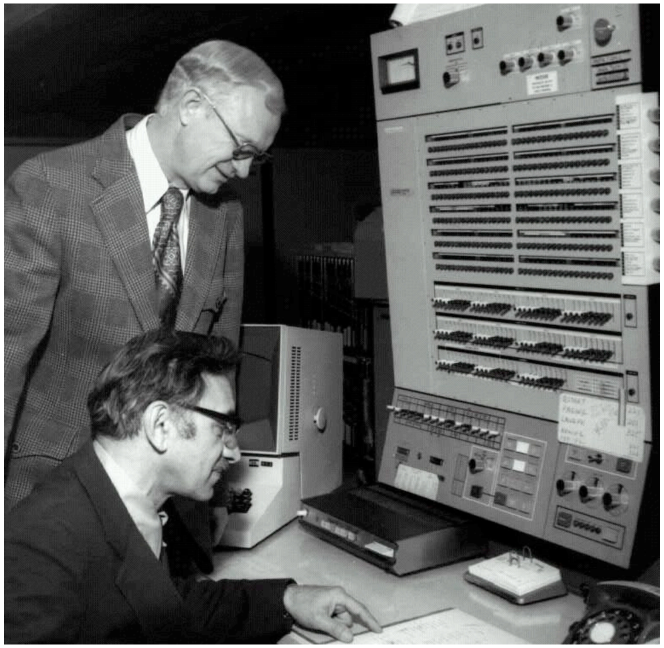
Computers have been around for a while…

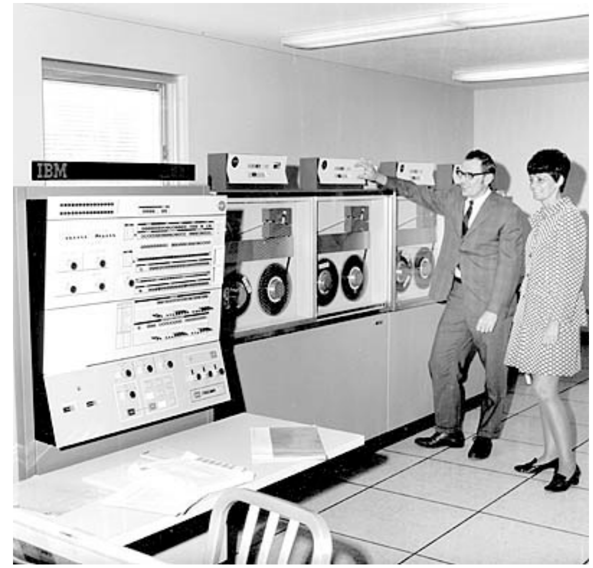


Computers have been around for a while…



Changes in Computing: Then…





…And Now


drobo


cingular
9:41 AM
SMS
Text
9
Calendar
Photos
Camera
Calculator
Stocks
Maps
73°
Weather
Notes
Clock
Settings

Statistics Curriculum: Then…

CONTENTS

TABLES

RA Fisher, *Statistical Methods for Research Workers*

SELECTED TABLES IN MATHEMATICAL STATISTICS

Sponsored by the Institute of Mathematical Statistics

...And now?

Casella & Berger

Bickel & Doksum

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]

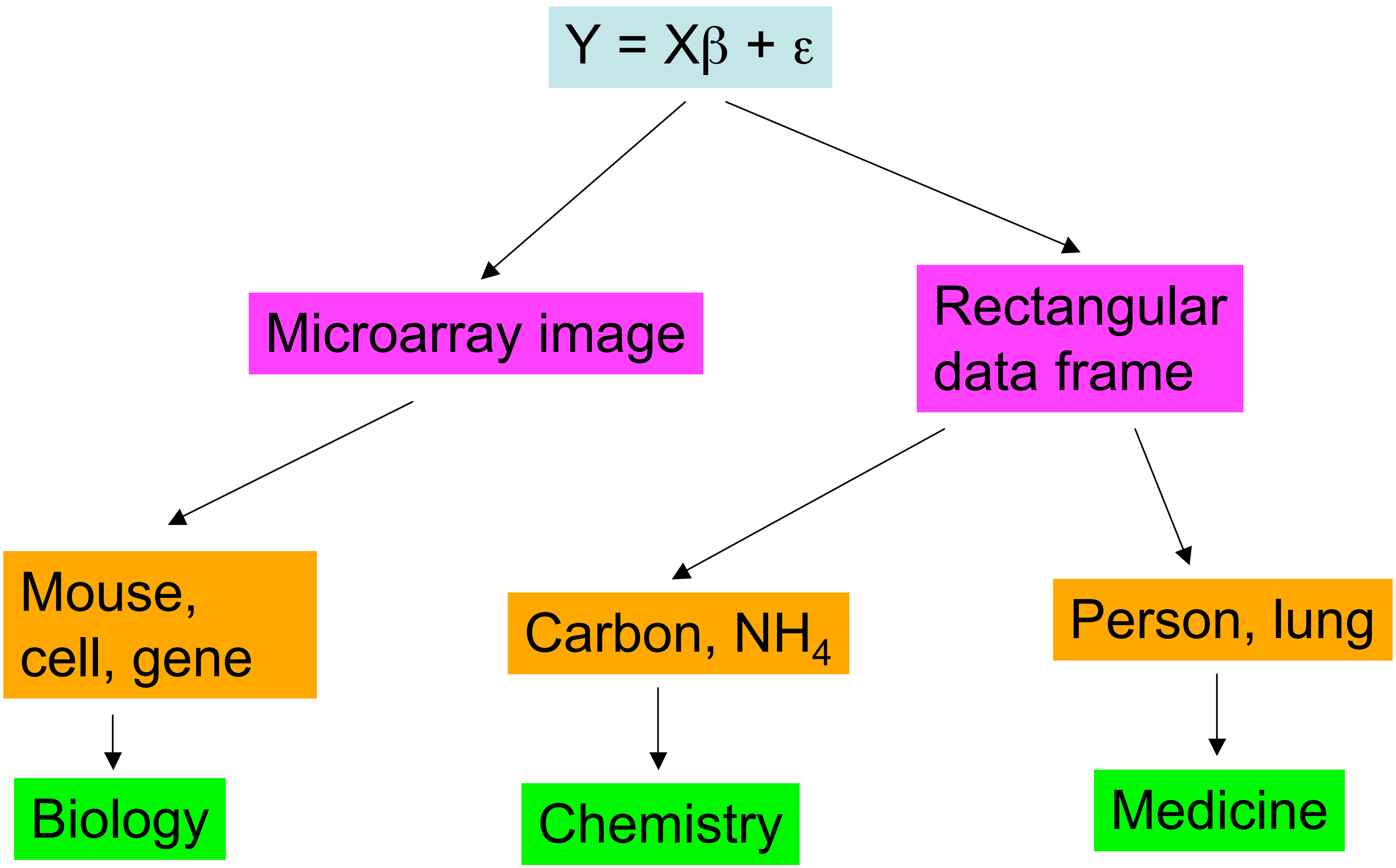
Steven A. Vokes, Hongkai Ji, Scott McCuine, Toyoaki Tenzen, Shane Giles, Sheng Zhong, William J. R. Longabaugh, Eric H. Davidson, Wing H. Wong and Andrew P. McMahon (2007)
Genomic characterization of Gli-activator targets in sonic hedgehog-mediated neural patterning.
Development, 134, 1977-1989. doi: 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?

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?

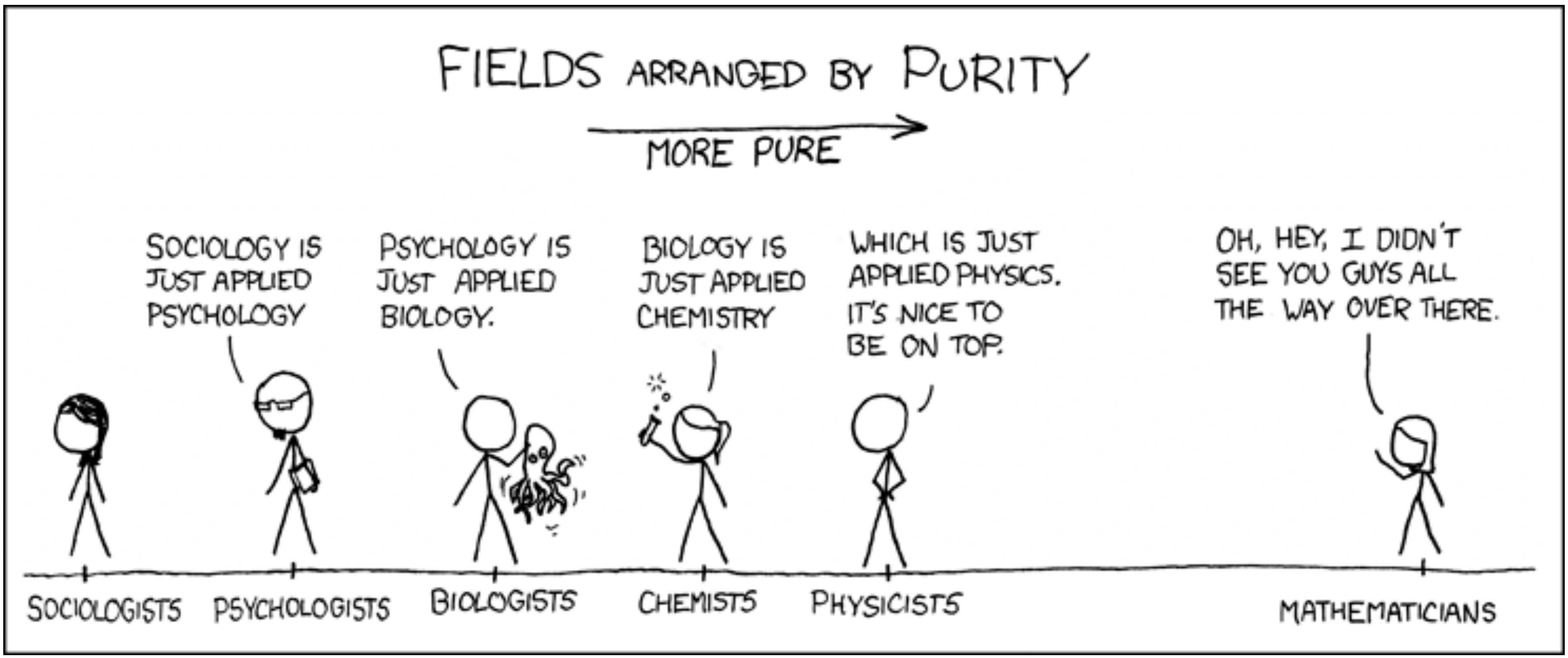
JHSPH Biostatistics

- "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.”

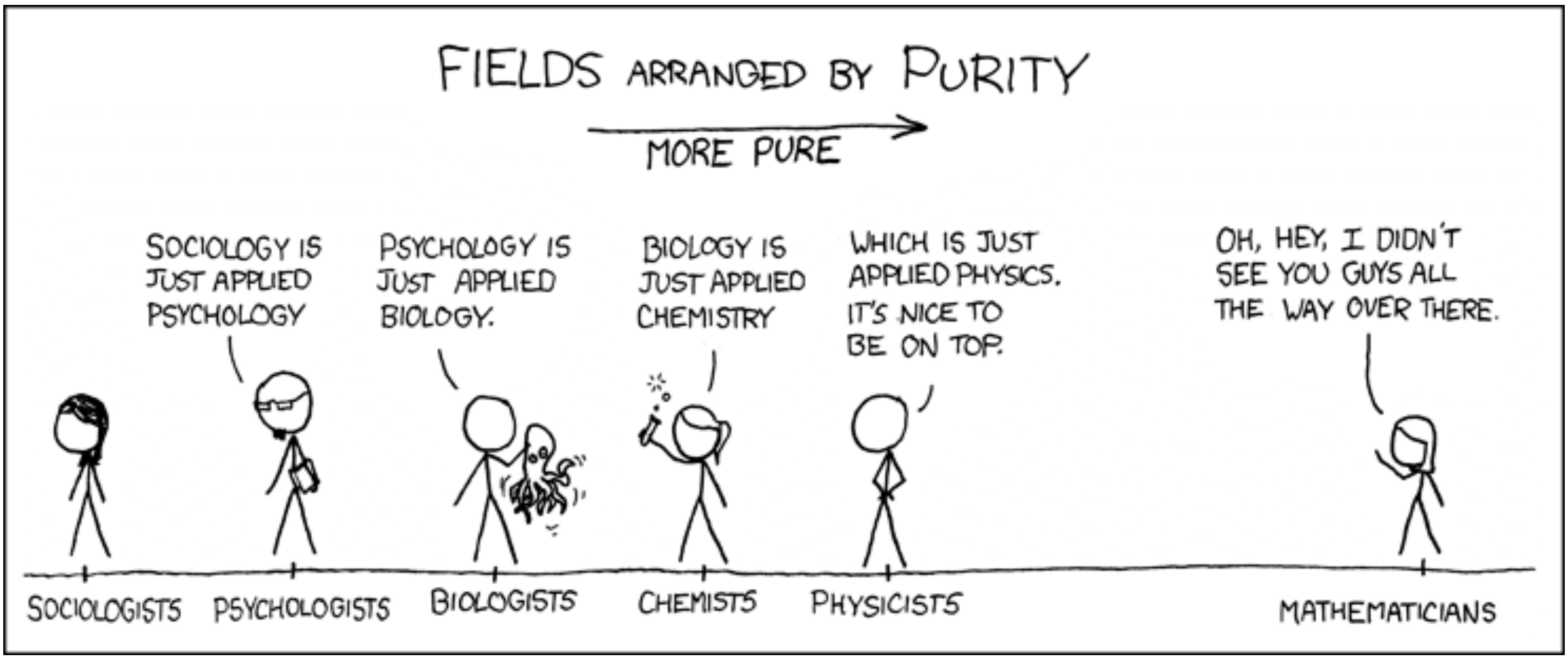
Where do statisticians belong?



xkcd.com

Where do statisticians belong?

Statisticians



xkcd.com

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)