Computing in the undergraduate curriculum

Teaching of Statistics Seminar

May 7, 2009

Approaching Statistics 13

My focus today will be mainly on my experiences Winter Quarter 2008 with Statistics 13; it will be a report on my experiences and is, at best, **a work in progress**

I am new to teaching and really don't have a clue about what works and doesn't -- I am probably the least qualified member of the department to lead a seminar on teaching!

With those caveats...

Approaching Statistics 13

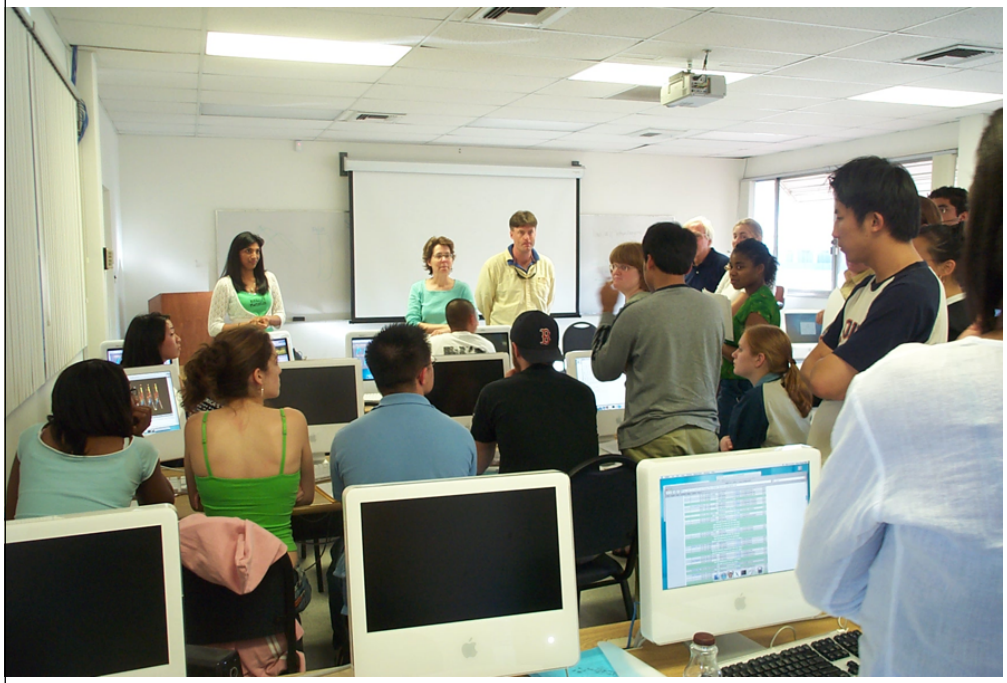
In terms of background, it had been three years since I last taught this class, and in the meantime I had been rethinking the role that **computing and data technologies should play at the undergraduate level**

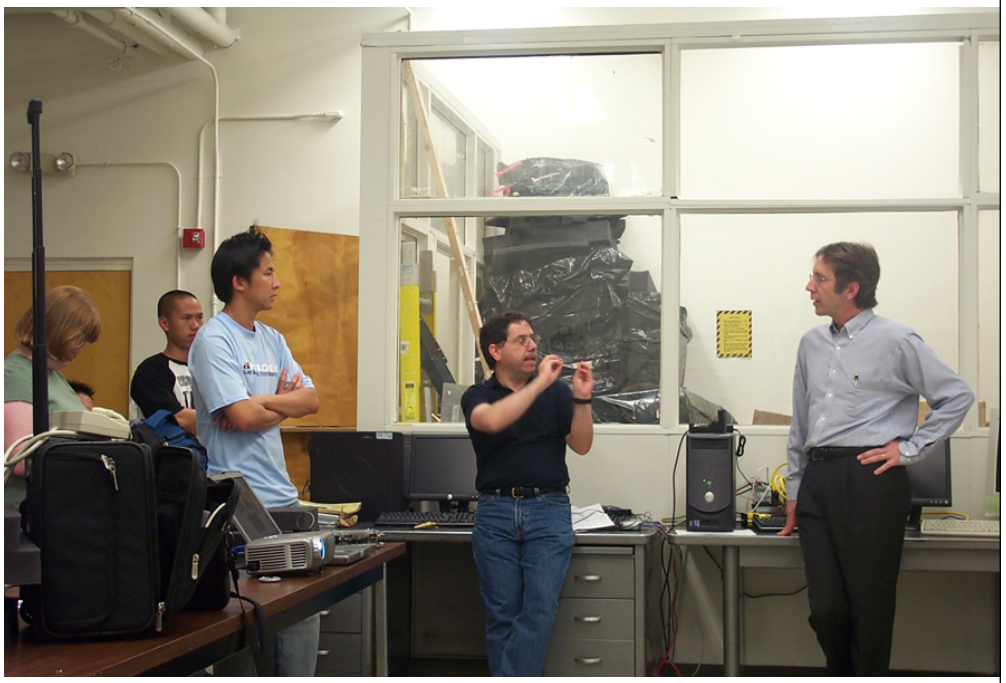
1. In 2005 and 2006 we hosted a **summer program for undergraduates** that emphasized computing and visualization

2. CENS introduced a core research focus in **participatory urban sensing**; this emphasized data collection and analysis by the public

3. Together with Co-PIs from Geography, Computer Science and Information Studies, we mapped out a field **"Data Science"** for the NSF IGERT call

4. In my graduate courses (seminars, Statistics 202a) I refined a story in which the fate of statistics **was tied to that of information technologies**



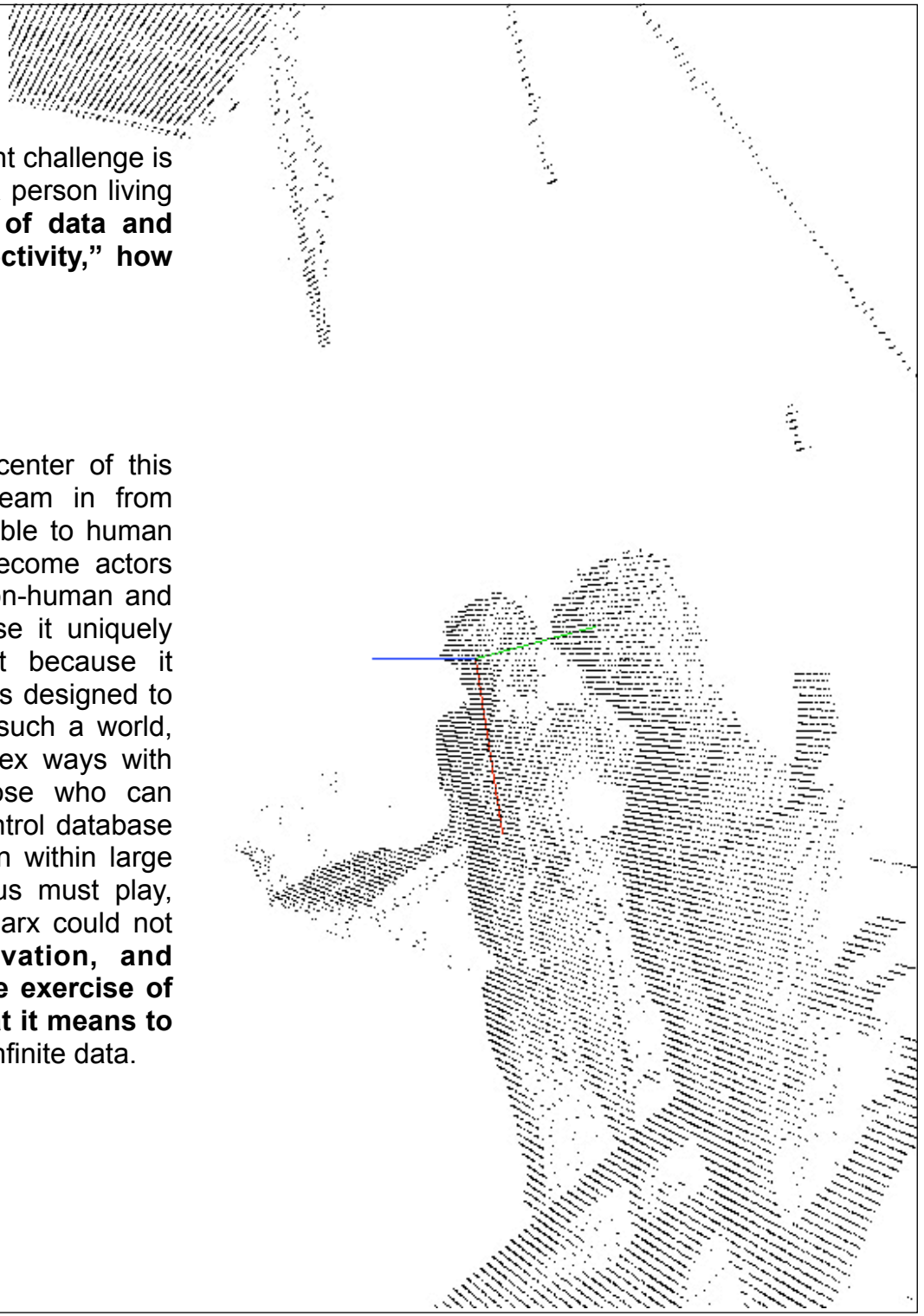


The more interesting and at the end maybe more important challenge is how to represent the personal subjective experience of a person living in a data society. **If daily interaction with volumes of data and numerous messages is part of our new "data-subjectivity," how can we represent this experience in new ways?**

Lev Manovich , The Anti-Sublime Ideal in Data Art

...The human cannot reasonably be imagined as the center of this distributed networked system. Rather, data flows stream in from everywhere, and the vast majority of them remain invisible to human perception and direct control. Humans in this world become actors among many different kinds of agents, most of them non-human and non-biological. Human intelligence is prized not because it uniquely distinguishes us from other (biological) species, but because it occupies a niche within a complex ecology of interactions designed to optimize its capabilities and minimize its limitations. In such a world, technological expertise (which correlates only in complex ways with economic class) assumes increased importance. Those who can interrogate databases, and even more so those who control database design and the standards determining how they function within large network systems, set the rules by which the rest of us must play, whether we are conscious of it or not. In a way that Marx could not have imagined, data, capital, **technological innovation, and information flows are setting new parameters for the exercise of power and control, and new landscapes defining what it means to be human.** This is the real fear, and the real promise, of infinite data.

N. Katherine Hayles
Reality Mining, RFIDs and Real Fears about Infinite Data



Approaching Statistics 13

In each case, I rediscovered my love for statistics and the important **social-political-scientific-technical** roles that our discipline plays as the acknowledged science of data

It's almost a cliche at this point to say that **the success of statistics as a discipline depends on our ability to compute**; the corollary, however, being that our practice should make more explicit our dependence on information technologies and emphasize the need to continually track new programming languages and paradigms, new database technologies, new data formats

In terms of teaching, perhaps this sums things up...

Today, software and hardware together provide far more powerful factories than most statisticians realize, factories that many of today's most able young people find exciting and worth learning about on their own. Their interest can help us greatly, if statistics starts to make much more nearly adequate use of the computer. However, **if we fail to expand our uses, their interest in computers can cost us many of our best recruits, and set us back many years.**

Today, software and hardware together provide far more powerful factories than most statisticians realize, factories that many of today's most able young people find exciting and worth learning about on their own. Their interest can help us greatly, if statistics starts to make much more nearly adequate use of the computer. However, **if we fail to expand our uses, their interest in computers can cost us many of our best recruits, and set us back many years**.

John W. Tukey, The Technical Tools of Statistics, **1964**

Approaching Statistics 13

And with all that as background, I cracked open my copy of Samuels and Witmer, the book that I had helped to select a few years back

Let’s have a look...

Contents

CONTENTS

A little retro

Fisher's **Statistical Methods for Research Workers** was first published in **1925** and the table of contents is remarkably similar to the text I had been using for the previous two years of Statistics 13

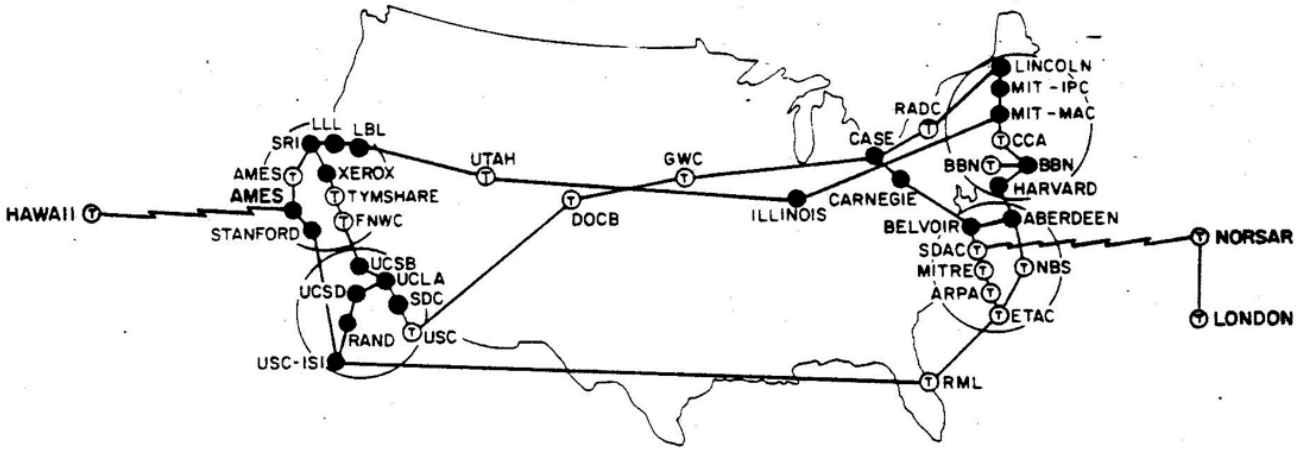
To be fair, my scan of SMRW was from the most recent edition which appeared in 1973; although that's still over 30 years old, perhaps there's nothing very different about statistics since then

Let's look back - first at the state of computing in 1973...

1972-1973

The first pocket calculator hits the market;
Texas instruments and HP will follow suit

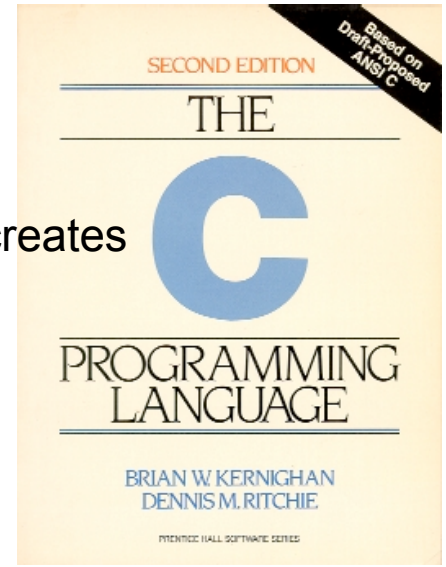
Xerox builds a personal computer (mouse, ethernet, GUI) but it's too expensive for the general public

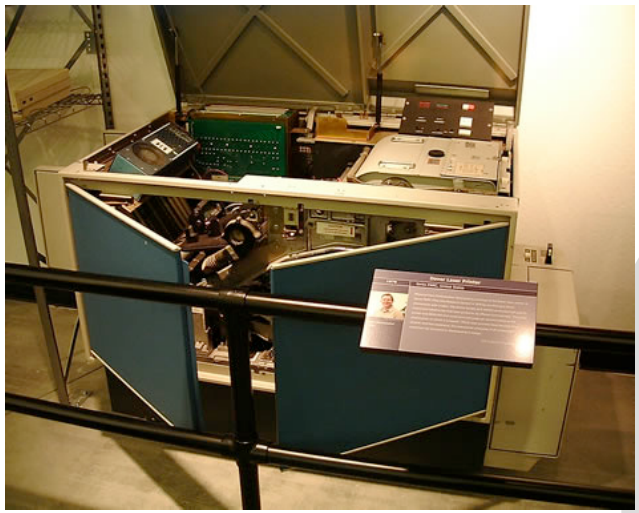


The ARPAnet connects 40 sites, the Internet consists of 25 computers

1972-1973

Dennis Richie at Bell Labs creates

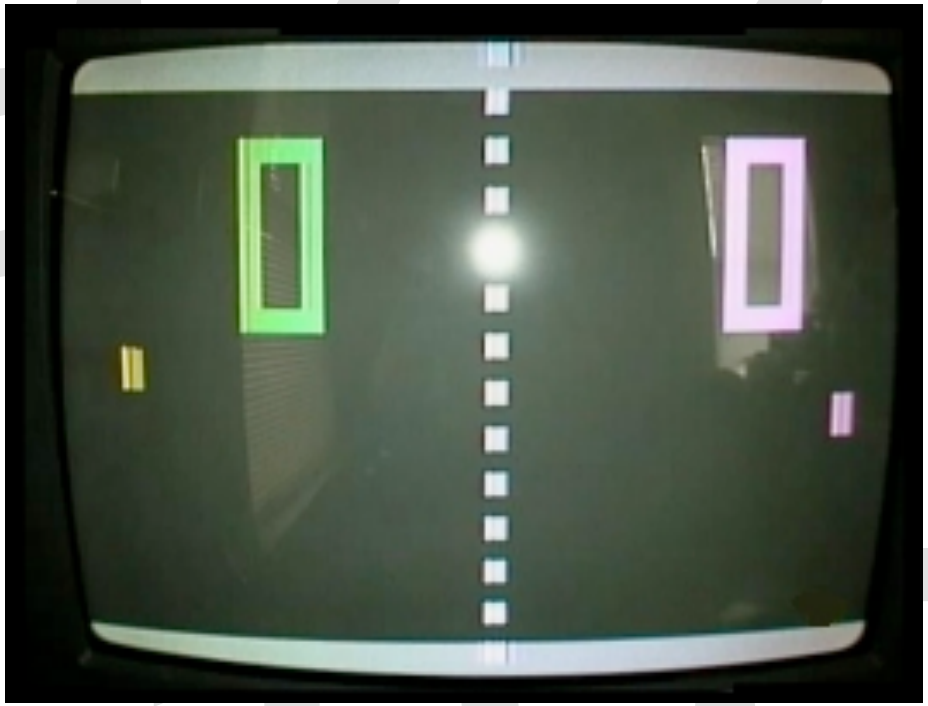




The first working
laser printer is produced

The Universal Product Code
(UPC) is developed

1973



PONG is born!

1973

A little retro

Ah, PONG

Anyway, in the early 1970s we find the beginnings of many of the computing ideas we now take for granted; what about statistics and statistical computing in particular?

Let’s have a look...

THE ANNALS
of
STATISTICS

An Official Journal of
The Institute of Mathematical Statistics

Articles

Short Communications

Continued on back cover

Vol. 1, No. 1—January 1973

"Up to that time, the Annals had published not only papers in mathematical statistics,but also had been one of the main outlets for papers in probability theory. Now the editor, Ingram Olkin, felt that the theory of probability had developed into a subject that deserved its own journal. He persuaded the IMS to create a new journal, the Annals of Probability, and at the same time to broaden the scope of the old Annals by dropping the limiting adjective "mathematical,"so that it would become more welcoming to applied work. The first of these two endeavors was wholly successful, the second less so."

AMERICAN STATISTICAL ASSOCIATION

FOUNDED 1839

806 - 15th Street, N.W. • Washington, D. C. 20005 • (code 202) 393-3253

November 17, 1971

RECEIVED
NOV 22 1971
BIOMATHEMATICS

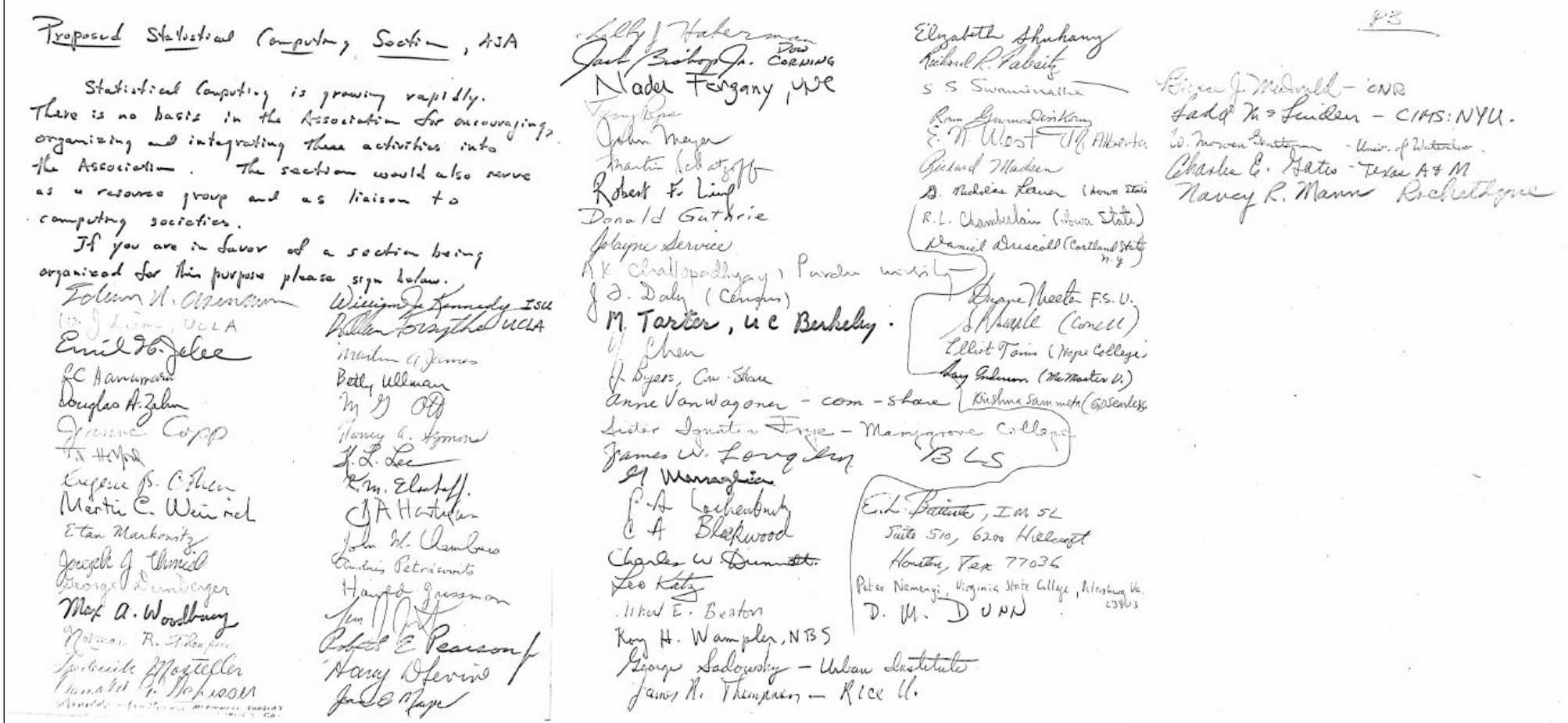
To: Paul Meier — Arthur P. Dempster
F. J. Anscombe — Wilfrid J. Dixon
Joseph M.Cameron — Michael D. Godfrey
John M. Chambers — Mervin E. Muller
Joseph F. Daly — Martin Schatzoff

From: Churchill Eisenhart, President

By vote of the membership, a new ASA Section on Statistical Computing has been established effective January 1, 1972. More than 1400 members of ASA voted for the establishment of the Section, which thus becomes the successor to the Committee on Computers in Statistics.

Will Dixon has agreed to be 1972 Chairman of the Section and Alan Forsythe will be Secretary. Art Dempster is the 1972 Program Chairman for the Section for the organization of sessions at the Annual Meetings, August 1972 in Montreal. If you have any suggestions for topics for these sessions, please write to Art Dempster at the Statistics Dept., Harvard Univ., 2 Divinity Ave., Cambridge, Mass. 02138.

On behalf of the Association, thank you for your service on the Committee and your leadership in providing the establishment of the new Section.



Proposed Statistical Computing Section, ASA

Statistical Computing is growing rapidly. There is no basis in the Association for encouraging, organizing and integrating these activities into the Association. The section would also serve as a resource group and as liaison to computing societies.

If you are in favor of a section being organized for this purpose please sign below.

Statistical Computing is growing rapidly. There is no basis in the Association for encouraging, organizing and integrating these activities into the Association. The section would also serve as a resource group and as liaison to computing societies.

Statistical computing circa 1972-1973

The ASA Section on Statistical Computing was officially founded on January 1, 1972 and sponsored an invited session at that year's JSM

In 1973, Francis and Heiberger asked the Section to form a committee to evaluate statistical packages, their report ultimately appearing in the American Statistician in 1975

Statistical Computing

This Department will carry articles of high quality on all aspects of computation in statistics.
Papers describing new algorithms, programs, or statistical packages will not contain listings of the program, although the completely documented program must be available from the author. Review of the paper will always include a running test of the program by the referee.
The description of a program or package in this Department should not be construed as an endorsement of it by the American Statistical Association or its Committees, nor is any warranty implied about the validity of the program.
The Editorial Committee will be pleased to confer with authors about the appropriateness of topics or drafts of possible articles.

Criteria and Considerations in the Evaluation of Statistical Program Packages

IVOR FRANCIS* AND RICHARD M. HEIBERGER,** AND PAUL F. VELLEMAN***

1. *Introduction*

Packages of computer programs for statistical analyses have proliferated in recent years and are now widely used, but not always widely understood. Packages can greatly assist statisticians by relieving them of tedious and error-prone computations, by making possible analyses of large data sets, and by providing a flexibility and versatility which can lead to more complete and searching analyses. In addition they are frequently used by non-statisticians to perform statistical analyses hitherto possible only with the collaboration of a statistician.

Few of these packages have ever been formally reviewed by statisticians. Consequently most users choose a package because of its availability or because of word-of-mouth recommendation. As packages have become more available, statisticians have become increasingly concerned about their impact because some contain errors, or are used for

The first task the committee set itself was to compile a set of criteria and considerations that would be useful in discussing packages. In this task the three authors of the report were greatly aided by nearly one hundred people, including both package designers and users, who had responded to an early draft sent to some five hundred people.

At the Annual meetings of the ASA at St. Louis in August 1974, the Committee presented a report on its work to the Section on Statistical Computing, and to a session with J. M. Chambers, W. J. Dixon, and H. O. Hartley as invited discussants. Copies of the full report [1], containing a list of contributors, are available from the authors.

In this article, we summarize, with some minor modifications, the section of the report on criteria and considerations. The considerations discussed in the report include those suggested by these contributors. Some appear as firm statements about which we expect little controversy. Others are opinions

Statistical computing circa 1972-1973

Francis, Heiberger and Velleman divided their criteria into **user interface** (documentation, “control language”, data structures, printed output, cost, audience and pedagogy), **statistical effectiveness** (versatility and accuracy) and **implementation** (programmer’s documentation, extensibility, portability, source language)

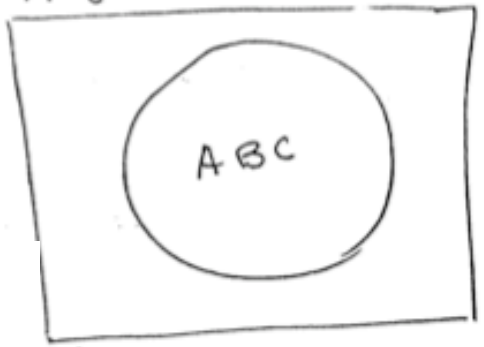
Keep in mind, however, that the early 1970s was the era of the mainframe computer, and the software market was dominated by a handful of “products” (BMDP, SPSS and SAS, for example)

And in 1976.... S

JHC

Algorithm Interface

5/5/76

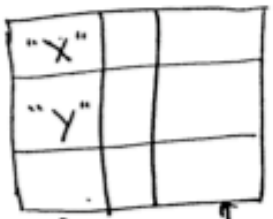


ABC: general (FORTRAN) algorithm

XABC: FORTRAN subroutine to provide interface between ABC & Language and/or utility programs

XABC (INSTR, OUTSTR)

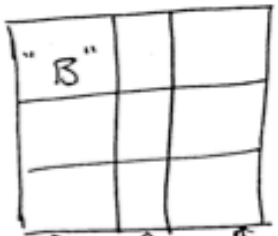
Input INSTR →



Pointers/Values

Argument Names or Blank

OUTSTR →



Pointers / Values

Types (Modes)

Result Names

Note: Names are meaningful to Algorithm, not necessarily to Language

The development of S

The image on the previous page is a scan of the first graphic produced at the May 5, 1976 meeting, and according to John Chambers:

> "The upper portion depicts the concept of an interface between a proposed user-level language and an 'algorithm,' which meant a Fortran-callable subroutine. The lower portion of the figure shows diagrammatically a hierarchical, list-like structure for data, the direct forerunner of lists with named components, structures with attributes and eventually classes with slots... ***The two portions of the sketch in fact lead to the themes of function calls and objects, in retrospect.***

By the end of 1976, Chambers and Rick Becker had produced a working version of their sketch

The program was initially referred to as "the system" locally, and attempts to come up with new names yielded lots of "unacceptable" candidates, but all of which had the letter "S" in common; and so, given the precedent of the recently-developed C language...

Approaching Statistics 13

All of this is a slight distraction, of course; but my point is that my beginning text was **completely silent** when it came to the uses, characteristics and demands of **modern data sources**

It painted statistics in an ahistorical, a-critical fashion; tables and formula were the end results of each chapter -- **statistics is not a living field**, but a set of procedures that can be applied to **a very narrow range of problems,** problems that dodged some of our really "big ideas"

Any nuance, any sense that statistics was **the product of human ingenuity,** was washed over with a kind of mathematical inevitability; there was no hint of **the debates in the field**, no sense of our **history**, and, ultimately, the students were **not prepared to question** the tools they were being offered

Approaching Statistics 13

And my students came to me with significant experience in certain kinds of data analysis, whether SW acknowledged it or not

We are all aware of the proliferation of visualization tools (popularized GIS platforms like Google Earth, for example) and analysis platforms (ManyEyes, Swivel), and even simple data collection frameworks via Twitter

Importantly, students entered the class with a profound understanding that data and data processing are important forces in their lives --
How do help them recognize that all of this is statistics?

Data and computing

The first part of the course emphasized **data and its essential character**; that a single data set can be **reformatted, reshaped, re-aggregated** to focus on different questions

We addressed the **"promiscuity" of data** and the ways in which data sets can be combined to address new questions; in future classes we will speak more directly of how data are formatted and the ease with which information can be extracted

Data and computing

Throughout, we emphasized the idea that data collection (in the health sciences, in particular) is often **a kind of social exchange** and that we need to think carefully about what happens when we turn the world into bits

In other words, we went way beyond just the taxonomy of data types presented by S&W (which has its own interesting history...)

SCIENCE

Vol. 103, No. 2684 Friday, June 7, 1946

On the Theory of Scales of Measurement

S. S. Stevens

Director, Psycho-Acoustic Laboratory, Harvard University

FOR SEVEN YEARS A COMMITTEE of the British Association for the Advancement of Science debated the problem of measurement. Appointed in 1932 to represent Section A (Mathematical and Physical Sciences) and Section J (Psychology), the committee was instructed to consider and report upon the possibility of "quantitative estimates of sensory events"—meaning simply: Is it possible to measure human sensation? Deliberation led only to disagreement, mainly about what is meant by the term measurement. An interim report in 1938 found one member complaining that his colleagues "came out by that same door as they went in," and in order to have another try at agreement, the committee begged to be continued for another year.

For its final report (1940) the committee chose a common bone for its contentions, directing its arguments at a concrete example of a sensory scale. This was the Sone scale of loudness (S. S. Stevens and H. Davis. *Hearing.* New York: Wiley, 1938), which purports to measure the subjective magnitude of an auditory sensation against a scale having the formal properties of other basic scales, such as those used to measure length and weight. Again the 19 members of the committee came out by the routes they entered, and their views ranged widely between two extremes. One member submitted "that any law purporting to express a quantitative relation between sensation intensity and stimulus intensity is not merely false but is in fact meaningless unless and until a meaning can be given to the concept of addition as applied to sensation" (Final Report, p. 245).

It is plain from this and from other statements by the committee that the real issue is the meaning of measurement. This, to be sure, is a semantic issue, but one susceptible of orderly discussion. Perhaps agreement can better be achieved if we recognize that measurement exists in a variety of forms and that scales of measurement fall into certain definite classes. These classes are determined both by the empirical operations invoked in the process of "measuring" and by the formal (mathematical) properties of the scales. Furthermore—and this is of great concern to several of the sciences—the statistical manipulations that can legitimately be applied to empirical data depend upon the type of scale against which the data are ordered.

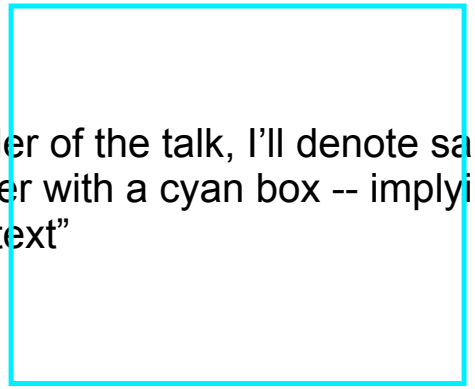
A Classification of Scales of Measurement

Paraphrasing N. R. Campbell (Final Report, p. 340), we may say that measurement, in the broadest sense, is defined as the assignment of numerals to objects or events according to rules. The fact that numerals can be assigned under different rules leads to different kinds of scales and different kinds of measurement. The problem then becomes that of making explicit (a) the various rules for the assignment of numerals, (b) the mathematical properties (or group structure) of the resulting scales, and (c) the statistical operations applicable to measurements made with each type of scale.

Scales are possible in the first place only because there is a certain isomorphism between what we can do with the aspects of objects and the properties of the numeral series. In dealing with the aspects of objects we invoke empirical operations for determining equality (classifying), for rank-ordering, and for determining when differences and when ratios between the aspects of objects are equal. The conventional series of numerals yields to analogous operations: We can identify the members of a numeral series and classify them. We know their order as given by convention. We can determine equal differences, as $8-6=4-2$, and equal ratios, as $8/4=6/3$. The isomorphism between these properties of the numeral series and certain empirical operations which we perform with objects permits the use of the series as a *model* to represent aspects of the empirical world.

The type of scale achieved depends upon the character of the basic empirical operations performed. These operations are limited ordinarily by the nature of the thing being scaled and by our choice of procedures, but, once selected, the operations determine

Samples



For the remainder of the talk, I'll denote sample slides from my Statistics 13 course last winter with a cyan box -- implying that their material might feel a little "out of context"

The registrar

The registrar maintains **a record of every class you take**; in addition to what class, it publishes a catalog of when classes meet and how many people were enrolled

On the next page, we present a few lines from a data file we will eventually consider in lab; it is was provided by the registrar (at a cost of $65) and contains the schedules for every student on campus last quarter*

In all, we have 191840 separate rows in this table, each corresponding to a different student and a single class; What can we learn from these data? And, more importantly, how?

*Note that the identification number in this table is not your student ID, or even part of it, but a random number generated to replace your real ID

id	dept	number	bldg	room	start	end	days	lev
1	MGMT	0466B			00:00:00	00:00:00	UNSCHED	G
1	MGMT	0474			00:00:00	00:00:00	UNSCHED	G
2	STATS	0013	HUMANTS	A00065	11:00:00	12:20:00	MW	U
2	SOCIOL	0101	DODD	00121	09:00:00	10:50:00	MW	U
2	SOCIOL	0101	PUB AFF	01337	14:00:00	14:50:00	M	U
2	SOCIOL	0020	BROAD	02160E	14:00:00	15:15:00	TR	U
2	SOCIOL	0020	BUNCHE	03143	13:00:00	13:50:00	R	U
2	STATS	0013	BOELTER	09413	12:00:00	12:50:00	T	U
2	STATS	0013	MS	05128	12:00:00	12:50:00	R	U
3	HIST	0107A	DODD	00078	11:00:00	12:15:00	TR	U
3	HIST	0097C	ROLFE	03120	13:00:00	15:50:00	R	U
3	ARMENIA	0106A	PUB AFF	02214	12:00:00	13:50:00	MW	U
4	PUB PLC	0219C	DODD	00175	14:00:00	15:20:00	TR	G
4	PUB PLC	0219C	PUB AFF	06362	15:45:00	16:45:00	R	G
4	PUB PLC	0290	PUB AFF	02343	18:00:00	21:50:00	T	G
4	PUB PLC	0209	PUB AFF	02343	10:00:00	11:20:00	MW	G
4	POL SCI	0222	BUNCHE	02121	13:00:00	15:50:00	M	G
5	ENGR	0183	WGYOUNG	CS00076	08:00:00	09:50:00	MW	U
5	MAT SCI	0104	BOELTER	02760	10:00:00	11:50:00	MW	U
5	ENGR	0183	BOELTER	05419	14:00:00	15:50:00	T	U

The registrar

At the end of your career here, you receive a transcript, an aggregation of all the data the registrar has on you (but printed in some reduced format); if, instead of looking at a single student, **we will look at all students in a given time period**, we get a different “view” of campus

Last time I asked you to consider this data set and what we might learn from it, specifically what it might say about “life” on campus; OK, that was vague, but many of you responded, and your ideas were great

Interestingly, not all your responses were student-centered; here’s a few questions you suggested (and I apologize for not including them all)

Questions about students

... you could find the average number of classes and units taken by a student at UCLA.

Could serve as a way to determine where the most students are at a given time and place; this would be useful for organizations, business interests, ads, etc.

Could provide information as to whether most students have a schedule with small time gaps/large time gaps; this would be useful information, as it would give more reason to add a schedule optimizer to ucla student websites (myucla).

1. How many classes/units do most students take?
2. What days of the week do the most/least students have class?
3. What day of the week do students have the most/least classes?
4. How many graduate students attend UCLA?
5. How many undergraduate students attend UCLA?
6. Do students prefer to take courses in the same subject or different?
7. Do students like to take back to back classes or have breaks?
8. Do students take more morning classes, afternoon classes, or evening classes?
9. Where do students take most/least classes?
10. Do students take 50 minutes classes or longer classes?
11. How many total classes did students take last quarter?
12. Did students take more north campus classes or south campus classes?

They could see that certain students in certain majors take on a heavier course load than other students in other majors.

Class questions

It can also determine which classes are more popular, in order to see if a particular class should be offered more than two times a year for instance.

They could see which classes are taken together concurrently commonly, and which combinations are not.

Furthermore, it can determine whether there is a trend-in which a certain class/field of study is being taken with a certain other class/field of study.

It can also compare the level/number of classes a typical undergraduate takes with that of a graduate student (last column...I'm guessing G stands for graduate and U for undergraduate?).

Facilities questions

First, it is possible to look at the rate of classroom usage - which classrooms/buildings are used most often, which would inform Facilities Management the classrooms and buildings (and nearby hallways/restrooms/stairwells) may need the most clean-up at the end of the day as well as more repairs from their constant use...

Also, it is possible for Administration to see if classrooms are being used to their full potential - that is, compare the number of students in each class against classroom capacity...

Along those lines, by analyzing the times that each building is used and the possible movement of classes into other rooms, the Administration could choose to alter the hours of the building and open later or close earlier if none of the rooms are used after a certain hour, which may yet again save electricity.

The registrar

What we see is that this one data set, this one record of students classes last quarter, can, in fact, support “studies” into a variety of questions about student life, the relationship between classes, and the functioning of campus

Before we get there, let’s try to put this data into the language we developed in the last lecture; what are our observational units (rows in the data table) and what variables have we measured? What type of data does each variable represent? Is the information coded in a useful way?

How does this data set compare to, say, the CDC BRFSS survey? Specifically, how does it differ in terms of how it was collected and what it represents?

The registrar

In its current form, the observational unit is technically is **an enrollment event** (a student signing up for a class); but for each different kind of question, **we will reshape or process the data**, changing units to students, to classes or to buildings

The registrar's data were collected for a purpose other than our analysis; you might call it a collection of convenience; the CDC data set is a survey, designed to address health questions facing the US adult population - it includes questions that help researchers track specific health trends

Unlike the CDC data set, we are given data on the **entire population of students** enrolled at UCLA in a quarter (can you find your schedule in this mess?); in what sense might we view it as a sample of something?

Because only the registrar could love enrollment events, let's consider other views of the data, other manipulated data sets...

Student view

To turn the data from a registrar-centric unit of observation (an enrollment event) into something we might care about, **we have to reshape or reformat or process the data**

Given the questions you emailed me, here is a table where **the basic unit of observation is now a student**; for each student, we record their level (graduate or undergraduate), the number of classes they've enrolled in, the number of hours spent in class, the number of days per week they have to be on campus and the time of their earliest class

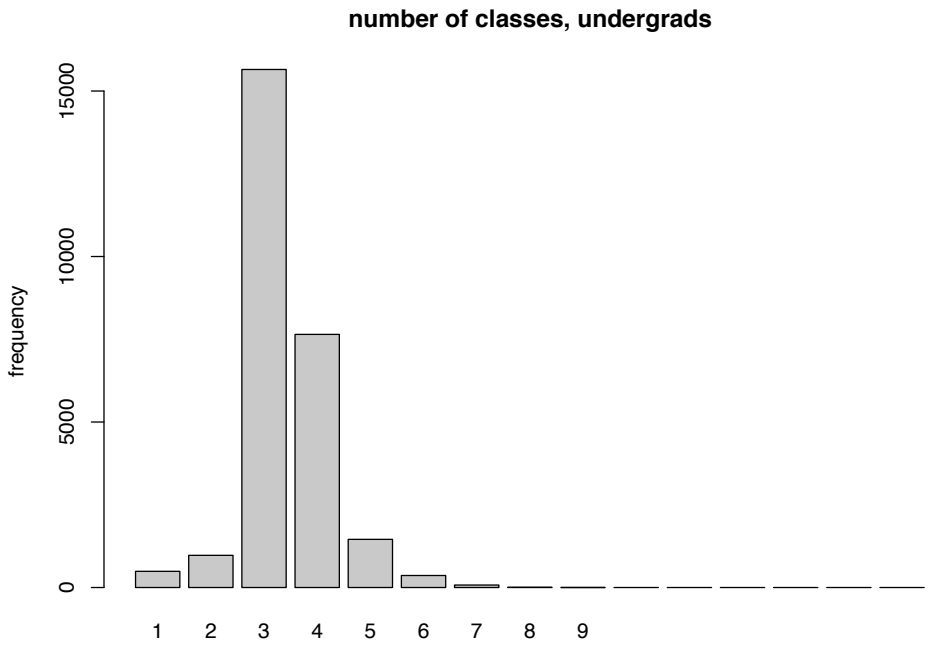
Here are a few observations and some simple plots; our data, by the way, have 1,110 professional students, 7,550 graduate students and 26,243 undergraduates

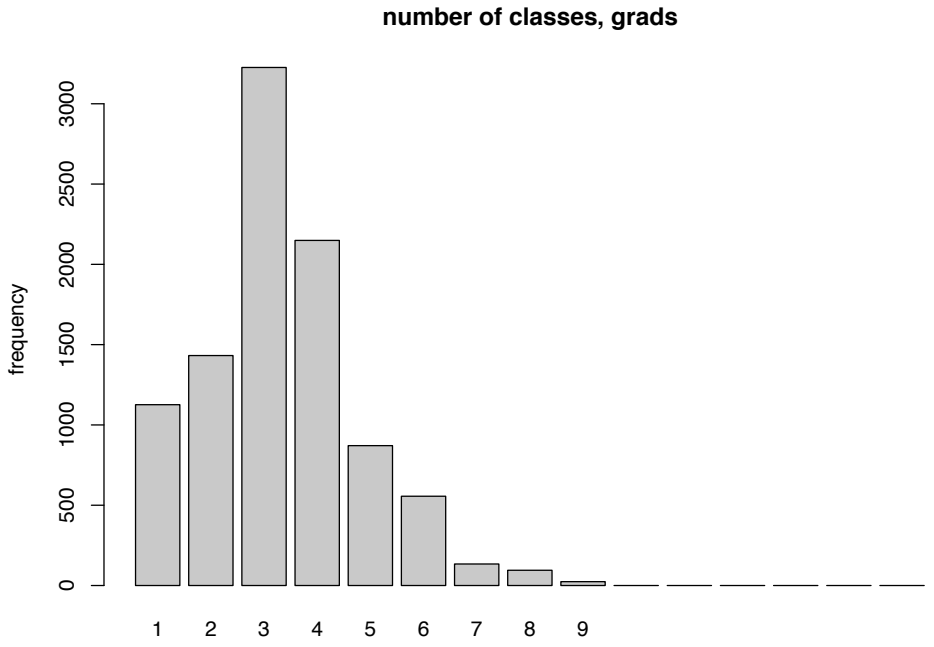
id	lev	num_classes	min_start	time_spent	days_on_campus
2	U	3	9	730	4
3	U	3	11	540	4
4	G	4	10	780	4
5	U	2	8	600	4
6	U	3	10	550	5
7	U	3	8	600	5
8	U	3	10	660	5
9	G	5	9	920	4
10	G	1	13	170	1
11	G	4	8	2100	5
12	U	4	8	820	5
13	U	4	8	1020	5
14	U	3	9	750	3
15	U	3	11	500	5
16	G	3	10	560	4
17	U	4	8	1040	5
18	U	4	9	660	5
19	U	2	9	450	4
20	G	8	8	1830	5
21	U	3	10	690	4
22	U	3	11	690	5
23	U	3	8	520	2
24	U	3	8	1470	5
25	U	3	10	520	3
26	G	4	9	850	3

Student view

Here, for example, are comparisons between the number of classes taken for graduate (G) and undergraduate (U) students

What do you see?

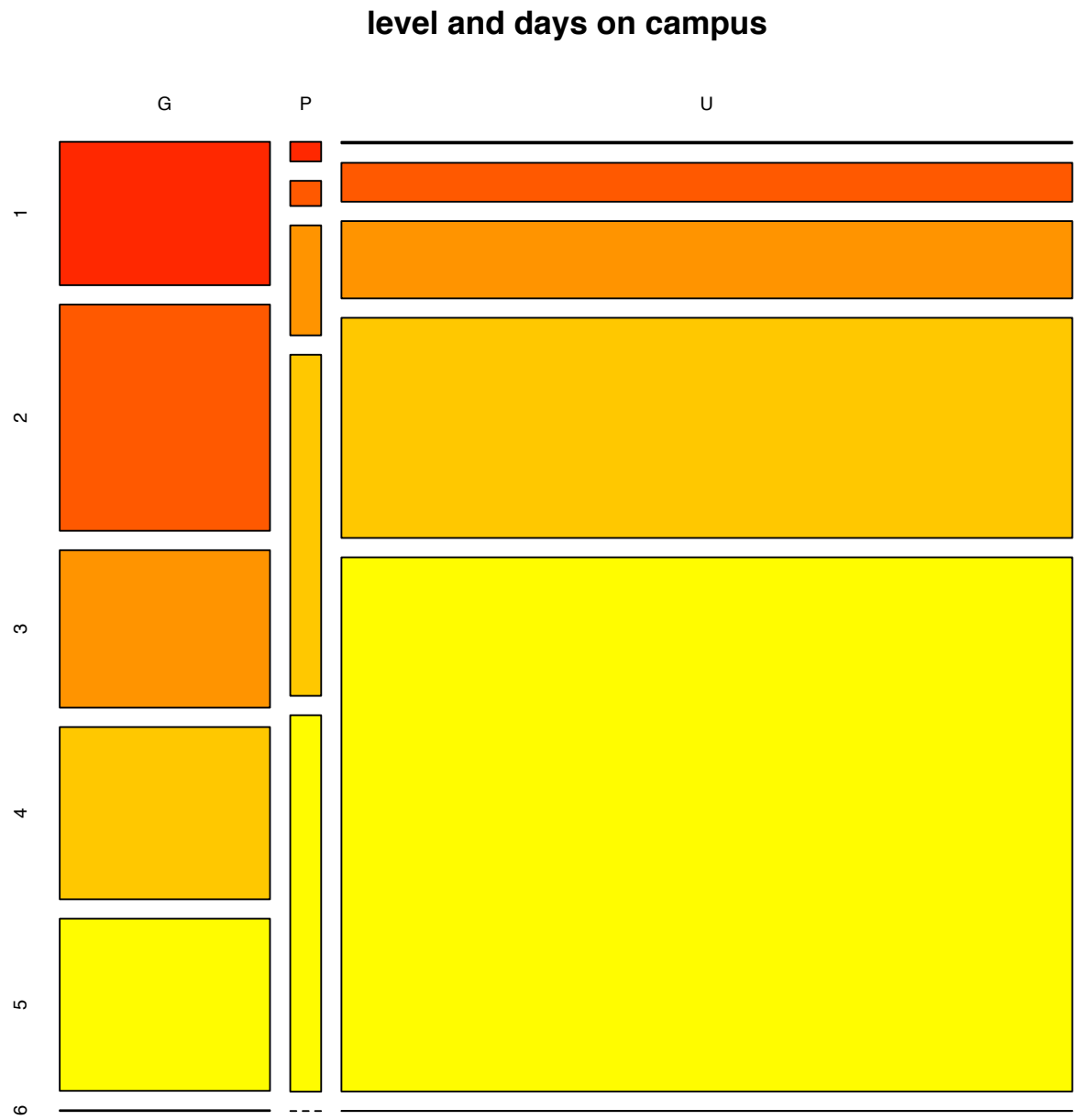




Student view

And here we examine the differences between graduate students, undergraduates, and professional students in terms of the number of days they have to be on campus

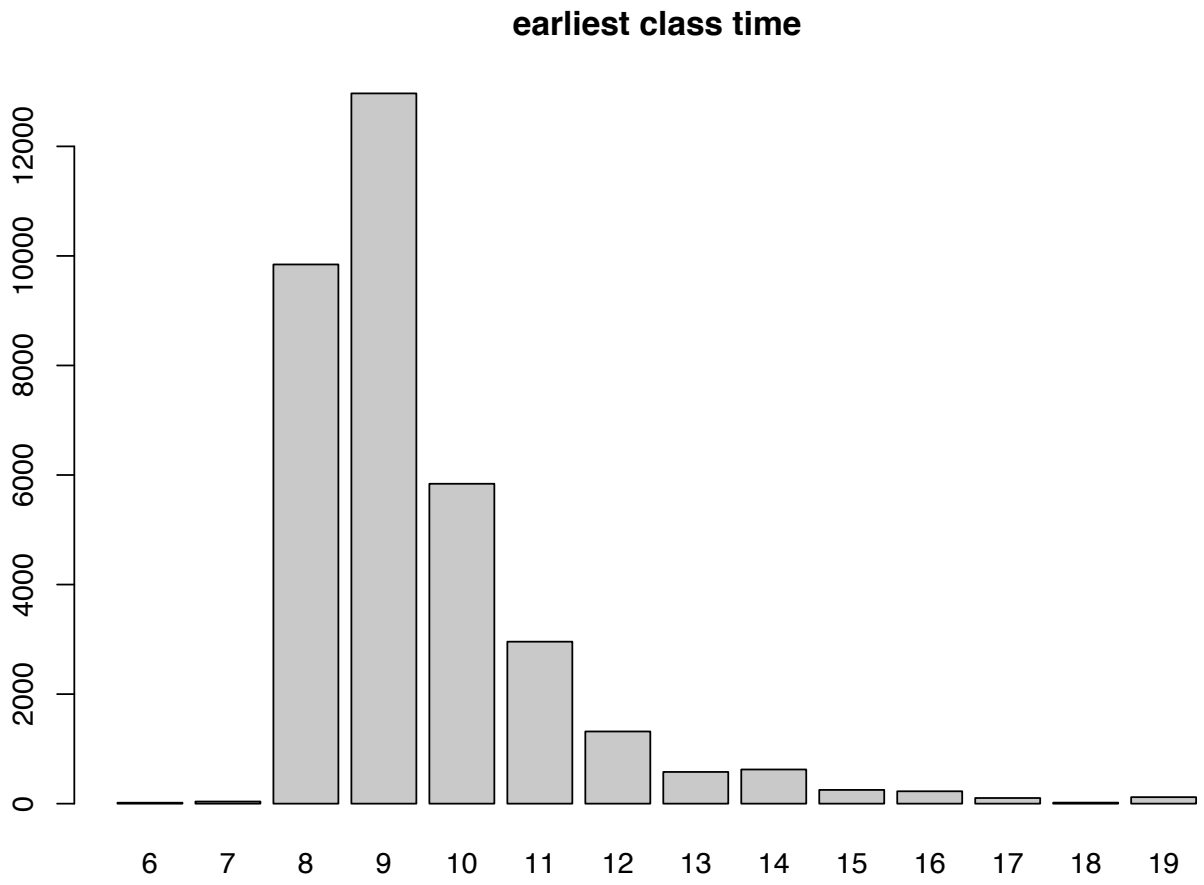
What do you see?



Student view

And finally, the earliest time we arrive on campus; on the right all the students are combined (graduate, undergraduate and professional)

What do you see? What questions might you want to ask?



	id	subj_area	cat_sort	meet_bdg	meet_rm_sort	strt_time	end_time	days_in_week	career
1	37	MGMT	0224	COLLINS	A00301	19:00:00	22:00:00	R	G
2	37	MGMT	0232D	CORNELL	D00307	19:00:00	22:00:00	W	G
3	1217	FILM TV	0434	BUNCHE	01265	19:00:00	21:50:00	T	G
4	1376	MGMT	0224	COLLINS	A00301	19:00:00	22:00:00	R	G
5	1608	MGMT	0261B	CORNELL	D00307	19:00:00	22:00:00	T	G
6	1614	MGMT	0224	COLLINS	A00301	19:00:00	22:00:00	R	G
7	1854	MGMT	0224	COLLINS	A00301	19:00:00	22:00:00	R	G
8	2397	MGMT	0267	CORNELL	D00310	19:00:00	22:00:00	R	G
9	3260	MGMT	0224	COLLINS	A00301	19:00:00	22:00:00	R	G
10	3264	MGMT	0224	COLLINS	A00301	19:00:00	22:00:00	R	G
11	3446	MGMT	0224	COLLINS	A00301	19:00:00	22:00:00	R	G
12	4867	MGMT	0224	COLLINS	A00301	19:00:00	22:00:00	R	G
13	5010	MGMT	0407	CORNELL	D00310	19:00:00	22:00:00	M	G
14	5149	MGMT	0224	COLLINS	A00301	19:00:00	22:00:00	R	G
15	5242	MGMT	0224	COLLINS	A00301	19:00:00	22:00:00	R	G
16	5445	MGMT	0224	COLLINS	A00301	19:00:00	22:00:00	R	G
17	5569	MGMT	0224	COLLINS	A00301	19:00:00	22:00:00	R	G
18	5778	MGMT	0224	COLLINS	A00301	19:00:00	22:00:00	R	G
19	6114	MGMT	0232D	CORNELL	D00307	19:00:00	22:00:00	W	G
20	6918	MGMT	0224	COLLINS	A00301	19:00:00	22:00:00	R	G
21	7211	MGMT	0224	COLLINS	A00301	19:00:00	22:00:00	R	G
22	7450	MGMT	0457	CORNELL	D00307	19:00:00	22:00:00	M	G
23	7450	MGMT	0232D	CORNELL	D00307	19:00:00	22:00:00	W	G
24	7933	MGMT	0407	CORNELL	D00310	19:00:00	22:00:00	M	G
25	8135	MGMT	0224	COLLINS	A00301	19:00:00	22:00:00	R	G
26	8529	MGMT	0224	COLLINS	A00301	19:00:00	22:00:00	R	G
27	9195	MGMT	0261B	CORNELL	D00307	19:00:00	22:00:00	T	G
28	9536	MGMT	0224	COLLINS	A00301	19:00:00	22:00:00	R	G
29	9698	MGMT	0261B	CORNELL	D00307	19:00:00	22:00:00	T	G
30	9733	FILM TV	0454B C	MELNITZ	02586B	19:00:00	21:50:00	R	G
31	10169	MGMT	0224	COLLINS	A00301	19:00:00	22:00:00	R	G
32	10245	MUSIC	0485 C	SMB	01345	19:00:00	21:50:00	W	G
33	10600	MGMT	0224	COLLINS	A00301	19:00:00	22:00:00	R	G
34	10716	MGMT	0224	COLLINS	A00301	19:00:00	22:00:00	R	G
35	10756	MGMT	0224	COLLINS	A00301	19:00:00	22:00:00	R	G
36	11090	ETHNOMU	0161Z	SMB	01846	19:00:00	21:50:00	W	G
37	11462	MGMT	0261B	CORNELL	D00307	19:00:00	22:00:00	T	G
38	11691	MGMT	0267	CORNELL	D00310	19:00:00	22:00:00	R	G
39	11734	MGMT	0407	CORNELL	D00310	19:00:00	22:00:00	M	G
40	11999	MGMT	0224	COLLINS	A00301	19:00:00	22:00:00	R	G
41	12038	FILM TV	0434	ROLFE	03106	19:00:00	21:50:00	W	G

Building view

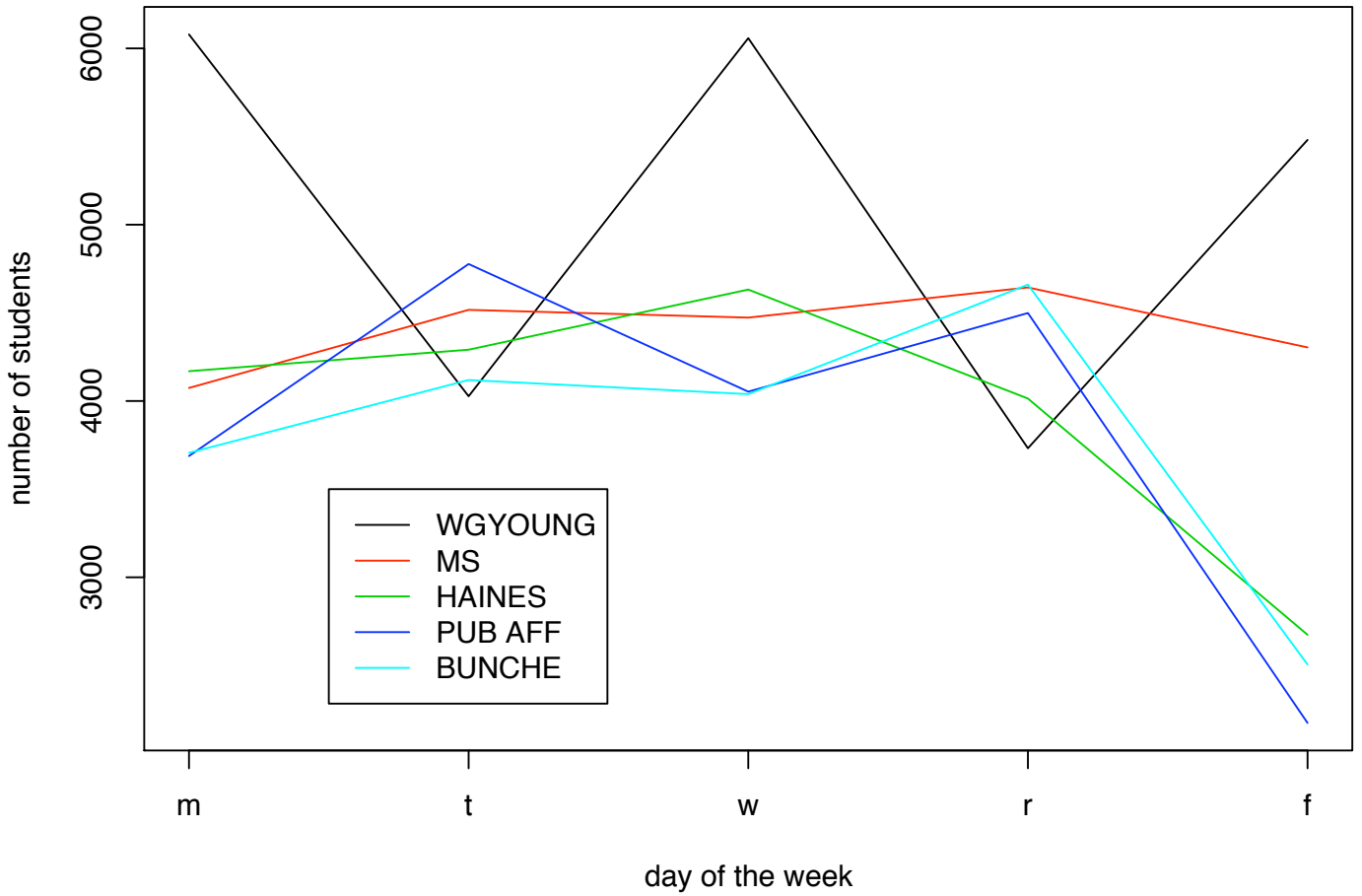
Several of you expressed interest in building usage; here is another data set, again created from the first registrar's data set, that records the number of people who enter each building each day

Note that if a person has class in MS in the morning and returns later for lab section, that counts as two in our data; think of it as the count from a turnstile posted at the entrance of MS

The rows here are sorted according to the weekly total number of students who click the turnstile for the given building; so Young sees the most students

On the next page we have a plot; what do you see?

bldg	m	t	w	r	f	s
WGYOUNG	6079	4027	6058	3732	5481	0
MS	4074	4517	4473	4644	4304	0
HAINES	4169	4291	4632	4014	2674	0
PUB AFF	3688	4777	4053	4499	2174	0
BUNCHE	3706	4119	4039	4660	2505	0
DODD	3424	3952	3569	3928	2544	0
BOELTER	3043	4003	3293	3764	3068	0
HUMANTS	3466	2750	3452	2932	2498	0
FRANZ	2430	2566	3050	2614	2566	0
LAW	1934	2589	2503	2358	720	0
MOORE	1930	2342	2096	2292	1095	156
ROLFE	1962	1911	2078	2223	1020	0
PAB	1814	1606	1849	1590	2057	0
LAKRETZ	1917	1428	2062	1512	1568	0
SMB	1870	1387	2173	1331	1448	0
KNSY PV	1987	1428	1987	1052	1659	0
ROYCE	1424	1755	1716	1914	1179	0
BROAD	1242	2001	1314	2001	621	0
FOWLER	1517	1084	1459	996	1073	0
GEOLOGY	625	919	1014	1098	688	0
PERLOFF	725	615	761	572	592	0
HLTHSCI	689	646	742	637	494	0
DE NEVE	759	905	619	905	0	0
CORNELL	690	734	701	603	0	83
GOLD	668	615	590	591	98	241
PUB HLT	467	551	452	575	486	155
MELNITZ	613	542	670	592	93	62
MACGOWN	376	629	344	860	164	32
KAUFMAN	480	483	482	524	290	0
OFF CAM	361	193	824	300	437	0


number of students
6000
5000
4000
3000
WGYOUNG
MS
HAINES
PUB AFF
BUNCHE
m
t
w
r
f
day of the week

Other views

Of course, we could reformulate these data in a variety of ways; suppose, for example, **we wanted to know how tight students' schedules are**; we could consider a transition from one class to the next as an observational unit and record items like when it started and ended and how far the student had to travel

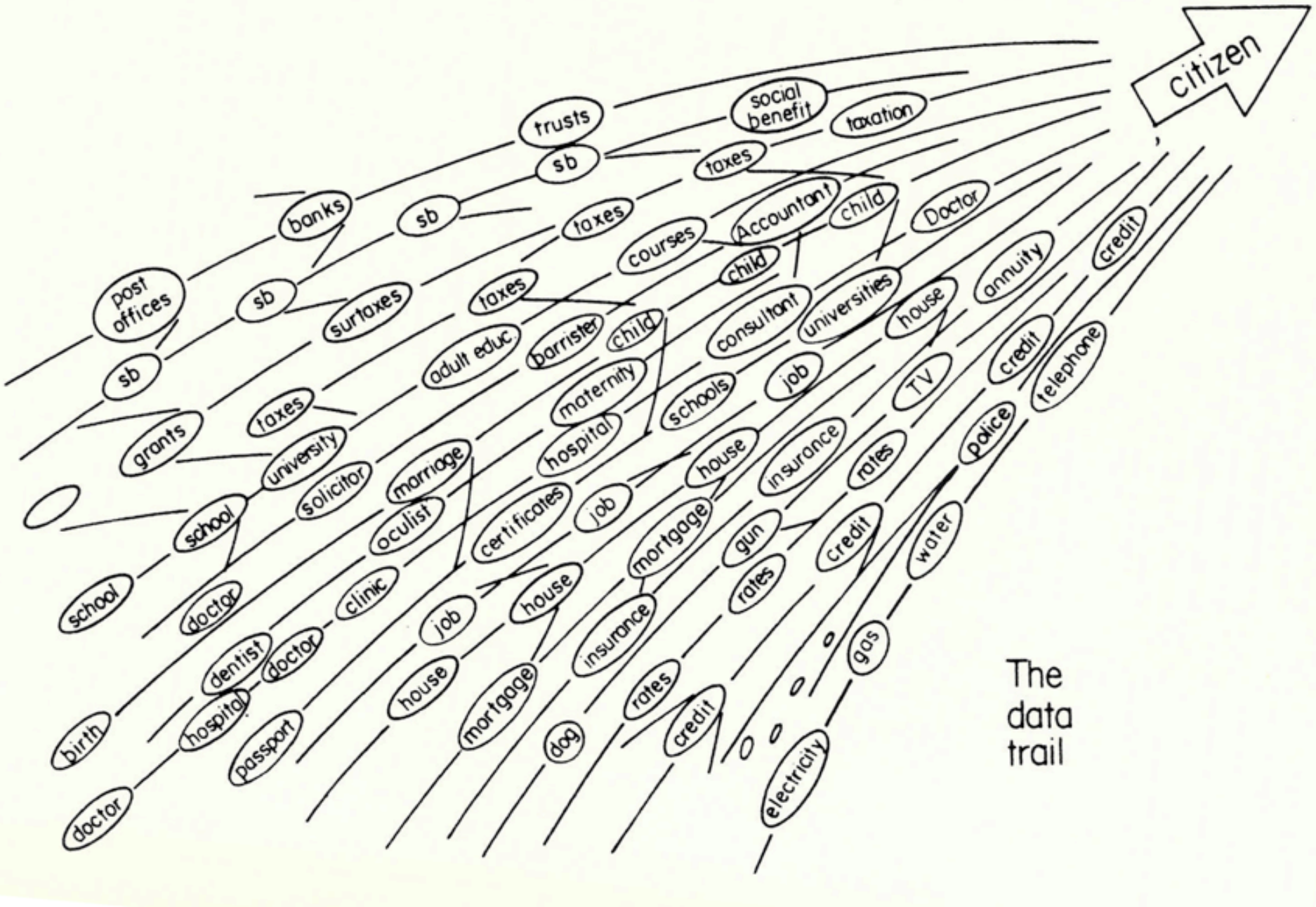
We could also consider classes as observational units, in this case asking whether MWF or TR or MW are the most popular; what variables should we include then?

Finally, several of you were interested in the popularity of classes; one way to get that would be to ask whether classes fill up or not, but perhaps another measure would be how quickly the classes fill up

Can we get that data?

Samples

And an example dealing with privacy, with data sharing, with terms of service...



The data trail

From *Platform for Change* by Stafford Beer


Facebook | Home
http://www.facebook.com/home.php?ref=home
Google
Apple
Google Maps
facebook
Home
Profile
Friends
Inbox 2
Terms of Use Update
Close
Over the past few days, we have received a lot of feedback about the new terms we posted two weeks ago. Because of this response, we have decided to return to our previous Terms of Use while we resolve the issues that people have raised. For more information, visit the Facebook Blog.
If you want to share your thoughts on what should be in the new terms, check out our group Facebook Bill of Rights and Responsibilities.
What are you doing right now?
News Feed
Status Updates
Photos
Links
Live Feed
Matthew Ericson just kissed a giraffe. about an hour ago - Comment - Like
Jane Mason at 6:14am February 18
impressive, given the ease at which you are completely grossed out.
Sarah Slobin at 6:39am February 18
Dude, I knew you were tall but that's ridiculous.
Todd Lindeman at 7:48am February 18
The average length of a giraffe's tongue is 18 to 20 inches. Just saying.
Write a comment...
Abbe Ruttenberg Serphos is two episodes behind on Heroes.
17 hours ago - Comment - Like
Katarina Holm-Didio Has misty eyes. Serena met Cinderella. Priceless.
18 hours ago - Comment - Like
Kayleigh Gillespie at 1:27pm February 17
awww! i can only imagine how happy she must have been! haha :)
Applications
Online Friends (2)

Facebook's Users Ask Who Owns Information

By BRIAN STELTER
Published: February 16, 2009

COMMENTS (92)
E-MAIL
PRINT
REPRINTS
SHARE

Reacting to an online swell of suspicion about changes to Facebook's terms of service, the company's chief executive moved to reassure users on Monday that the users, not the Web site, "own and control their information."

Related

Facebook Withdraws Changes in Data Use (February 19, 2009)

Times Topics: Facebook

Readers' Comments

Readers shared their thoughts on this article.
Read All Comments (92) »

The online exchanges reflected the uneasy and evolving balance between sharing information and retaining control over that information on the Internet. The subject arose when a consumer advocate's blog shined an unflattering light onto the pages of legal language that many users accept without reading when they use a Web site.

The pages, called terms of service, generally outline appropriate conduct and grant a license to companies to store users' data. Unknown to many users, the terms frequently give broad power to Web site operators.

This month, when Facebook updated its terms, it deleted a provision that said users could remove their content at any time, at which time the license would expire. Further, it added new language that said Facebook would retain users' content and licenses after an

Facebook Withdraws Changes in Data Use

By ALAN COWELL
Published: February 18, 2009

After a wave of protests from its users, the Facebook social networking site said on Wednesday that it would withdraw changes to its so-called terms of service concerning the data supplied by the tens of millions of people who use it.

Related

Facebook's Users Ask Who Owns Information (February 17, 2009)

Times Topics: Facebook

Readers' Comments

Share your thoughts.
Post a Comment »
Read All Comments (96) »

The about-face was made known to many users in a message posted on the Facebook home page saying : "Over the past few days, we have received a lot of feedback about the new terms we posted two weeks ago. Because of this response, we have decided to return to our previous Terms of Use while we resolve the issues that people have raised."

The posting invited users to click on a link to get more details.

Terms of service generally outline appropriate conduct and grant a license to companies to store users' data. Unknown to many users, the terms frequently give broad power to Web site operators.

Earlier this month, Facebook deleted a provision from its terms of service that said users could remove their content at any time, at which time the license would expire. It added

The original Terms of Use (put back in place as of yesterday):

"By posting User Content to any part of the Site, you automatically grant, and you represent and warrant that you have the right to grant, to the Company an irrevocable, perpetual, non-exclusive, transferable, fully paid, worldwide license (with the right to sublicense) to use, copy, publicly perform, publicly display, reformat, translate, excerpt (in whole or in part) and distribute such User Content for any purpose on or in connection with the Site or the promotion thereof, to prepare derivative works of, or incorporate into other works, such User Content, and to grant and authorize sublicenses of the foregoing. You may remove your User Content from the Site at any time. If you choose to remove your User Content, the license granted above will automatically expire, however you acknowledge that the Company may retain archived copies of your User Content."

The version from February 4th stated:

"You hereby grant Facebook an irrevocable, perpetual, non-exclusive, transferable, fully paid, worldwide license (with the right to sublicense) to (a) use, copy, publish, stream, store, retain, publicly perform or display, transmit, scan, reformat, modify, edit, frame, translate, excerpt, adapt, create derivative works and distribute (through multiple tiers), any User Content you (i) Post on or in connection with the Facebook Service or the promotion thereof subject only to your privacy settings or (ii) enable a user to Post, including by offering a Share Link on your website and (b) to use your name, likeness and image for any purpose, including commercial or advertising, each of (a) and (b) on or in connection with the Facebook Service or the promotion thereof. You represent and warrant that you have all rights and permissions to grant the foregoing licenses."

Casey Hamilton (Evansville, IN) wrote
at 8:05am

If facebook wants to use my pics, just ask.

Report

Maggie Elizabeth Hatfield (Salmen High School) wrote
at 8:05am

i do not think Face Book should be allowed to use our pictures in anyway such as ads or give them to other sites. i believe they are going to have many lawsuits with this term if they go throught with it. most people do not put their pictures up for eveyone in the world to see but just friends and family. this site is mainly about keeping in touch with others and by saying you own my pictures and information is a wrong thing to say. you should have to ask permission if we want you to do that. also you need to let people know when you are changing or adding terms like this. even if in the terms it says you dont have too but even myspace does it. i think it is copletly wrong saying you own other peoples information.

Report

Alesha Nesbeth (N.C. Central) wrote
at 8:05am

I feel like This is a way to trap users. When users either delete their accounts or something on their accounts, it's for a reason. It shouldn't be yours forever just because you feel like it. And some of the us users have been on here for years, so the terms shouldn't have changed so abrutly and i think very strongly that the new terms are ridiculous.

Report

Alicanne Trotter (Chicago, IL) wrote
at 8:05am

Why do I have the option of privacy settings, if Facebook themselves aren't going to keep it private? My photos & infomation is only for my friends not for public use! And I don't agree to share them or give up my rights to them. So I hope that Facebook will remember that.

Report

Alicia Hardy (Kitchener, ON) wrote
at 8:05am

I just reid to add all my friennnz and it didn't work! Thats myprob with facebook,, and does facebook seLL ppL the keys to my info?? Because someone doens't need to own my page to read it..... thats my question to facebook.

Homework

Examine the **Terms of Service for Google** or some other site that you contribute data to (in addition to the ToS, you might want to focus on the site's Privacy Policy -- it often contains information about data retention); then, write **a short paragraph** addressing the following questions:

1. What data do you provide to this site?
2. Who owns it?
3. How can it be shared?
4. Can you delete your data at any time?

Data and computing

In terms of computing, we made it quite clear on the first day that, while statistics is not computer science, **programming is an inevitable component** of any modern data analysis

Whether a course is based in R or Stata or SPSS, this lesson is largely the same; that (at least since the early 1970s) statistical software lets us make a good graph or two, and perform some version of the iterative analysis advocated by Tukey

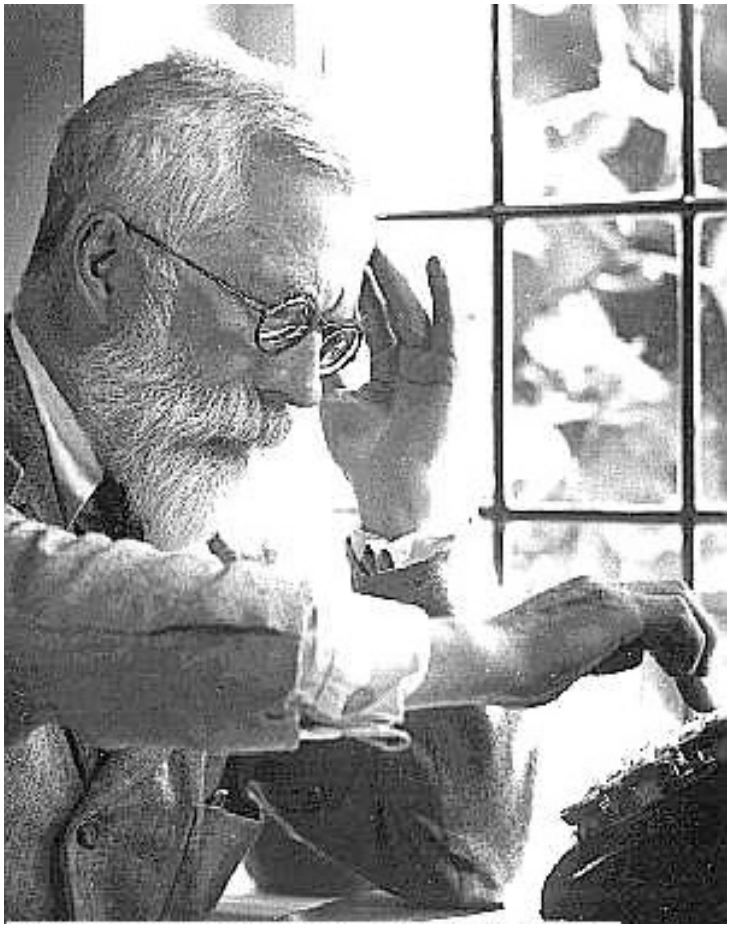
Computing

R. A. Fisher is often credited as the single most important figure in 20th century statistics; one of his books, *Statistical Methods for Research Workers* (1958) covers many of the basic techniques in in your textbook

He is responsible for creating a mathematical framework for statistics; before Fisher, statistics has been characterized as an "ingenious collection of ad hoc devices"

Fisher once commented that **"he had learned all he knew" over his hand calculator**; as we will see, these computations can reveal structures in data

Ronald Aylmer Fisher (1890-1962)



Computing

John Tukey is another pioneer of statistical theory and practice, working through the later half of the 20th century; he promoted the idea of looking at data, of exploratory analysis

In fact, he literally wrote the book on the subject; in *Exploratory Data Analysis* (1977), Tukey creates graphical tools for exploring features in data

His style is iterative, advocating many different analyses in an approach that is graphically and computationally intensive; many of the descriptive techniques presented in your textbook are due to Tukey

For Tukey, **"[w]hat the statistician often needs is enough different 'looks' to have a good chance to learn about this real world."**

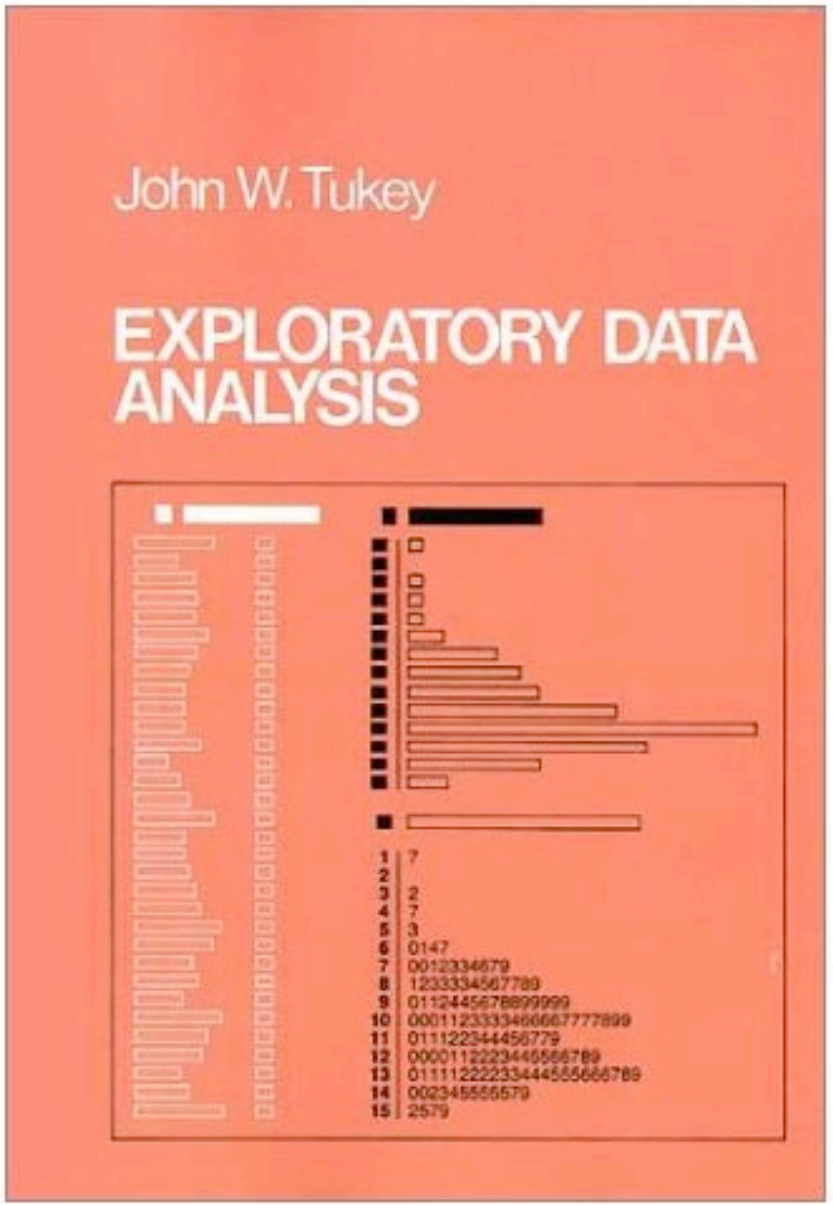
John Wilder Tukey (1915 - 2000)

Computing

To be fair, Fisher and Tukey emphasized hand calculation; even Tukey's EDA is really about manual computations, about "scratching down" numbers

Of course 30 years have passed since the first edition of this book; we've seen the rise of personal computers (first PCs and now mobile phones) and powerful networks to connect them (much more on this at the end of the lecture)

Data analysis is no longer a manual activity; and so while this is not a class in computer science, modern computing tools will play a big role in what you will do



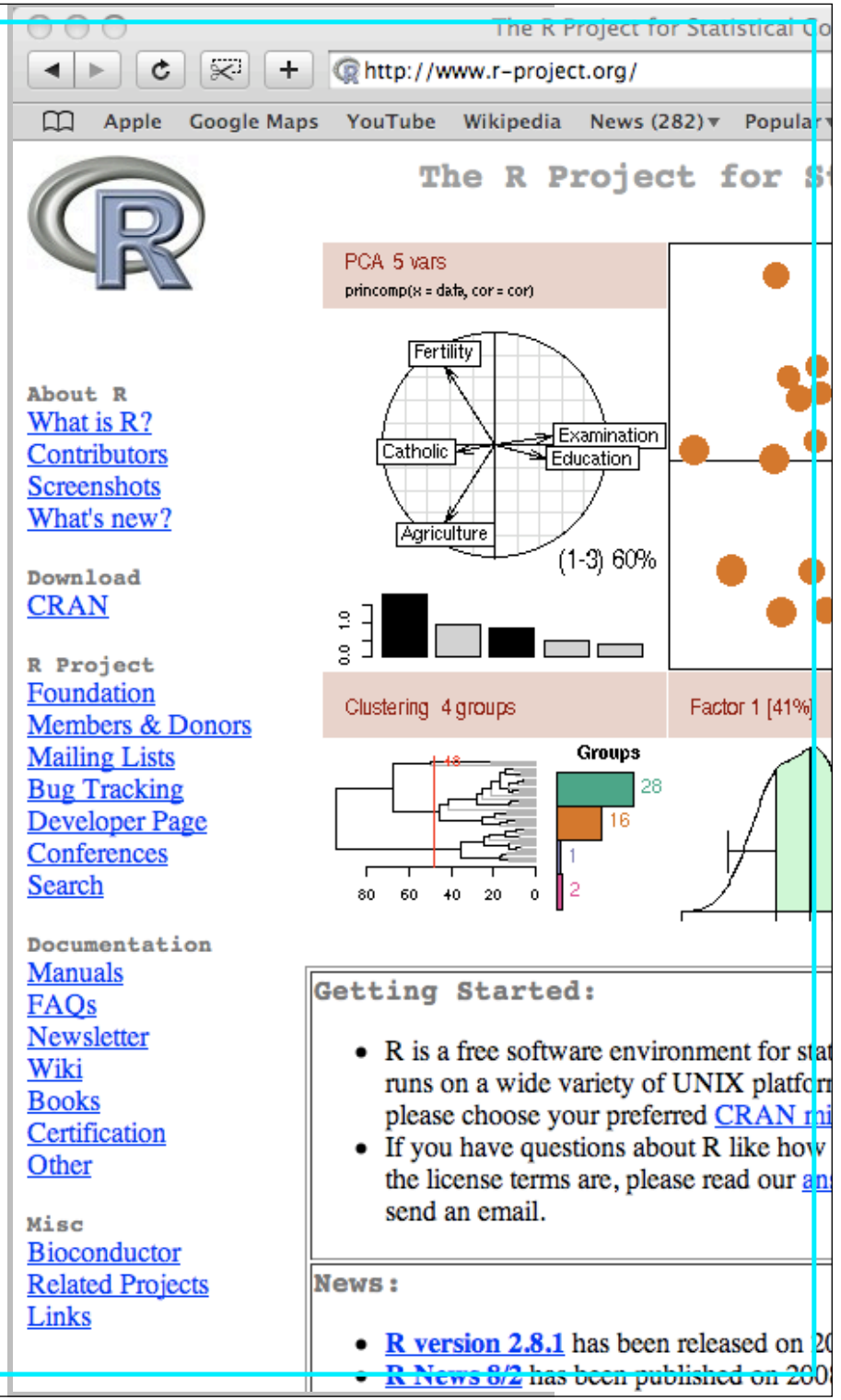
Computing: The R environment

This quarter, we will perform all of our analysis in the R environment; we've selected it because

1. **Access:** It's free and runs on any operating system you are likely to have at home

2. **Openness:** R is open source; there is a large community actively contributing to the codebase, and most practicing statisticians "share" their analyses through R programs

3. **Expressiveness:** Far from a point-and-click interface, R is programmable and extendable; with it we will consider a range of complex data types



Computing

That said, any piece of digital technology develops and matures in a particular technological and social context; it's important to remember the following chain (which holds equally for Vista, Google, Facebook or your favorite iPhone App)

Software = Model of the world = An argument

This is especially true for a piece of "technical" software like R; you will be learning how to **analyze a set of data**, with a mix of visual and numerical summaries, with a mix of simple mathematics, probability and simulation

Embedded in the commands you will type, in the small programs you will run, are a series of choices that the R designers have made; it's sensible to ask why R looks and operates the way it does (are certain analyses easier to implement than others, for example)

About lab...

This appeared in the New York Times yesterday; the main point of the article seems to be that R has a large user base (250K by one estimate) and that it's stealing significant market share from older programs like SAS

It reinforces why we chose R in the first place; it has a lot of support from the statistics community and is **gaining traction among data analysts** outside this group

According to Daryl Pregibon (a statistician at Google and my old boss) "It allows statisticians to do very intricate and complicated analyses without knowing the blood and guts of computing systems."

The New York Times

Business Computing

Data Analysts Captivated by R's Power

Stuart Isett for The New York Times

R first appeared in 1996, when the statistics professors Robert Gentleman, left, and Ross Ihaka released the code as a free software package.

By ASHLEE VANCE
Published: January 6, 2009

To some people R is just the 18th letter of the alphabet. To others, it's the rating on racy movies, a measure of an attic's insulation or what pirates in movies say.

R is also the name of a popular programming language used by a growing number of data analysts inside corporations and academia. It is becoming their lingua franca partly because data mining has entered a golden age, whether being used to set ad prices, find new drugs more quickly or fine-tune financial models. Companies as diverse as Google, Pfizer, Merck, Bank of America, the InterContinental Hotels Group and Shell use it.

Data and computing

In previous incarnations of my class, I had used R in a circa 1973 batch mode; packaged analyses and an occasional table look up when students needed to compute a P-value

Last quarter, we attempted to **push computing farther into the core of the course**, employing intuitive algorithmic descriptions of statistical methods as a means for teaching some of our most important topics

While vague, this approach involved considering the character of different algorithms...

Analysis by analogy

In previous versions of the course, I used R to explore the frequentist framework, having students simulate repeated trials and examine sampling distributions -- **R was a story generator**, but the stories were only useful in an indirect way when faced with data

This approach provides very little for students to critique, and the “narrative arc” that supports it has us **substitute real data for artificial data sets**

Data and computing

Last term we removed the indirection and chose to work with **rerandomization techniques** from the very beginning -- we gradually built up the idea that one should “analyze as you randomized”

With this approach, we avoided the (previously inevitable) detour into probability theory and the CLT; instead we start with simply stated principles and developed **the essential components of hypothesis testing on day 3**

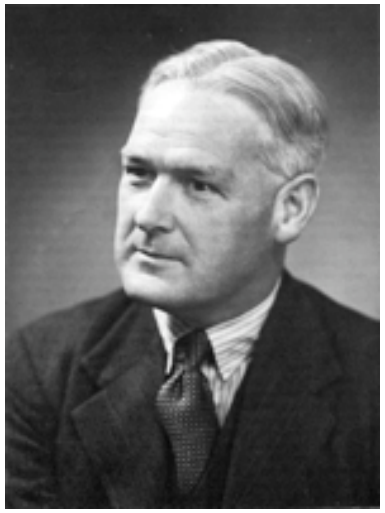
Data and computing

Rerandomization (permutation tests) are, in my mind, an example of a computational procedure that, as an algorithm, has **direct pedagogical and practical value** (during our resampling lecture students asked ME about why so many resampling distributions looked bell shaped!)

It is also extensible to a variety of contexts and not just the sample mean or the sample proportion; in fact, our discussion of testing **highlighted the choice of statistics**, on the appropriate way to judge differences between distributions

As an interesting side note, by not following the SW path, t**he sample mean was no longer an obvious choice of statistic to focus on** (and I struggled to make the mean an sensible choice); students really were free to think about other summary measures

Clinical trials



In 1948, Hill published a groundbreaking study on the effectiveness of streptomycin (an antibiotic) in treating pulmonary tuberculosis; here is how he assigned patients to the treatment and control groups

Determination of whether a patient would be treated by streptomycin and bed-rest (S case) or by bed-rest alone (C case) was made by reference to a statistical series based on ***random sampling numbers drawn up for each sex*** *at each centre by Professor Bradford Hill; the details of the series were unknown to any of the investigators or to the co-ordinator and were contained in a set of sealed envelopes, each bearing on the outside only the name of the hospital and number. After acceptance of a patient by the panel, and before admission to the streptomycin centre,* ***the appropriate numbered envelope was opened at the central office: the card inside told if the patient was to be an S or C case****, and this information was then given to the medical officer of the centre. Patients were not told before admission that they were to get special treatment; C patients did not know throughout their stay in hospital that they were control patients in a special study; they were in fact treated as they would have been in the past, the sole difference being that they had been admitted to the centre more rapidly than was normal. Usually they were not in the same wards as S patients, but the same regimen was maintained."*

Medical Research Council. Streptomycin treatment of pulmonary tuberculosis. BMJ 1948; 2: 769-782.

Fisher and randomization

"*The theory of estimation presupposes a process of random sampling. All our conclusions within that theory rest on this basis; without it our tests of significance would be worthless. ... In controlled experimentation it has been found not difficult to introduce explicit and objective randomisation in such a way that the tests of significance are demonstrably correct. In other cases we must still act in faith that Nature has done the randomisation for us....* ***We now recognise randomisation as a postulate necessary to the validity of our conclusions, and the modern experimenter is careful to make sure that this postulate is justified.***"

Fisher RA. *Development of the theory of experimental design*. Proceedings of the International Statistical Conferences 1947;3:434–39

Another aside: Fisher and Hill

There is, in fact, an interesting story that connects these two researchers; both were active in roughly the same time period and they were certainly aware of each other's work

They exchanged correspondence starting in 1929, "Dear Sir"; and then in 1931 "Dear Fisher" and "Dear Bradford Hill"; and then in 1940 "My dear Fisher" and "My dear Bradford Hill"; and then by 1952 "My dear Ron" and "My dear Tony" (Hill went by Tony, another long story)

But by 1958 they were back to "Dear Fisher" and "Dear Bradford Hill" as the two (Doll, a significant co-investigator with Hill) were on opposites sides in a dispute as to whether or not smoking caused lung cancer

From the point of view of our discussion, one of Fisher's main criticisms of the studies suggesting that smoking caused lung cancer was the fact that they were entirely observational; he wanted a "properly randomized experiment" (which of course would be difficult as you can't force people to start smoking)

Hill's tuberculosis trial

And here are Hill's original results from his 1948 paper; what do we see?

TABLE II.—*Assessment of Radiological Appearance at Six Months as Compared with Appearance on Admission*

Radiological Assessment	Streptomycin Group		Control Group	
Considerable improvement ..	28	*51%*	4	*8%*
Moderate or slight improvement	10	*18%*	13	*25%*
No material change	2	*4%*	3	*6%*
Moderate or slight deterioration	5	*9%*	12	*23%*
Considerable deterioration ..	6	*11%*	6	*11%*
Deaths	4	*7%*	14	*27%*
Total	55	*100%*	52	*100%*

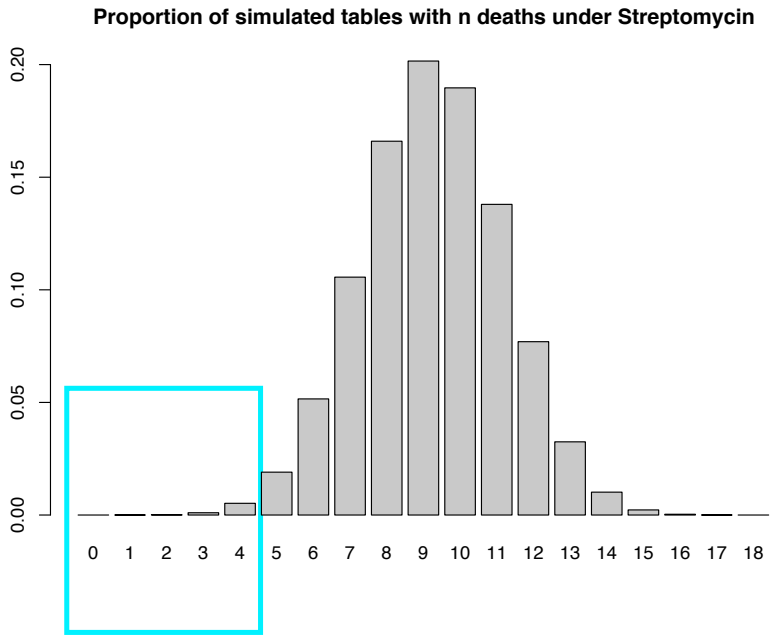
Medical Research Council. Streptomycin treatment of pulmonary tuberculosis. BMJ 1948; 2: 769-782.

Simulating random assignments

In this plot we see that a value as small or smaller than four is fairly rare; to be precise, only 0.6% of the tables have 4 or fewer deaths in the Streptomycin group

This, then, provides us with evidence that there is something more at work here than random assignment

If we believed the null hypothesis, that there was no difference between Streptomycin and bed rest, **the results HIII observed would have been extremely rare**, coming up a very small fraction of the time



Data and computing

Interestingly, this approach not only eliminated the need for our “data-free detour” but it also **re-emphasized the way in which a data set was created**; to perform resampling students immediately question whether random assignment was present in the original design (and so SW chapter 8 became our second chapter)

Moving from testing to inference, we again took a computational approach; deciding that the same sleight of hand that SW uses to move from a t-test to a confidence interval was a “thin” algorithm, **we opted instead for the bootstrap**

(I am borrowing terminology here; Kolb classifies places as being thick or thin depending on the social roles that take place there -- I will use these terms in a similar way, where we think of thin algorithms as those which cast the student in a very simple role -- typing in a call to qt() -- versus thick algorithms which allow the students to ask deeper questions about the data, about the applicability of the algorithm, to question the tool)

The bootstrap: Relative risk

The bootstrap idea can be applied quite widely; here is how it works in our A/B Travel Section trial:

1. Create two populations, the first consisting of 1211 ones and 22,684 zeroes; and a second consisting of 732 ones and 22,917 zeroes

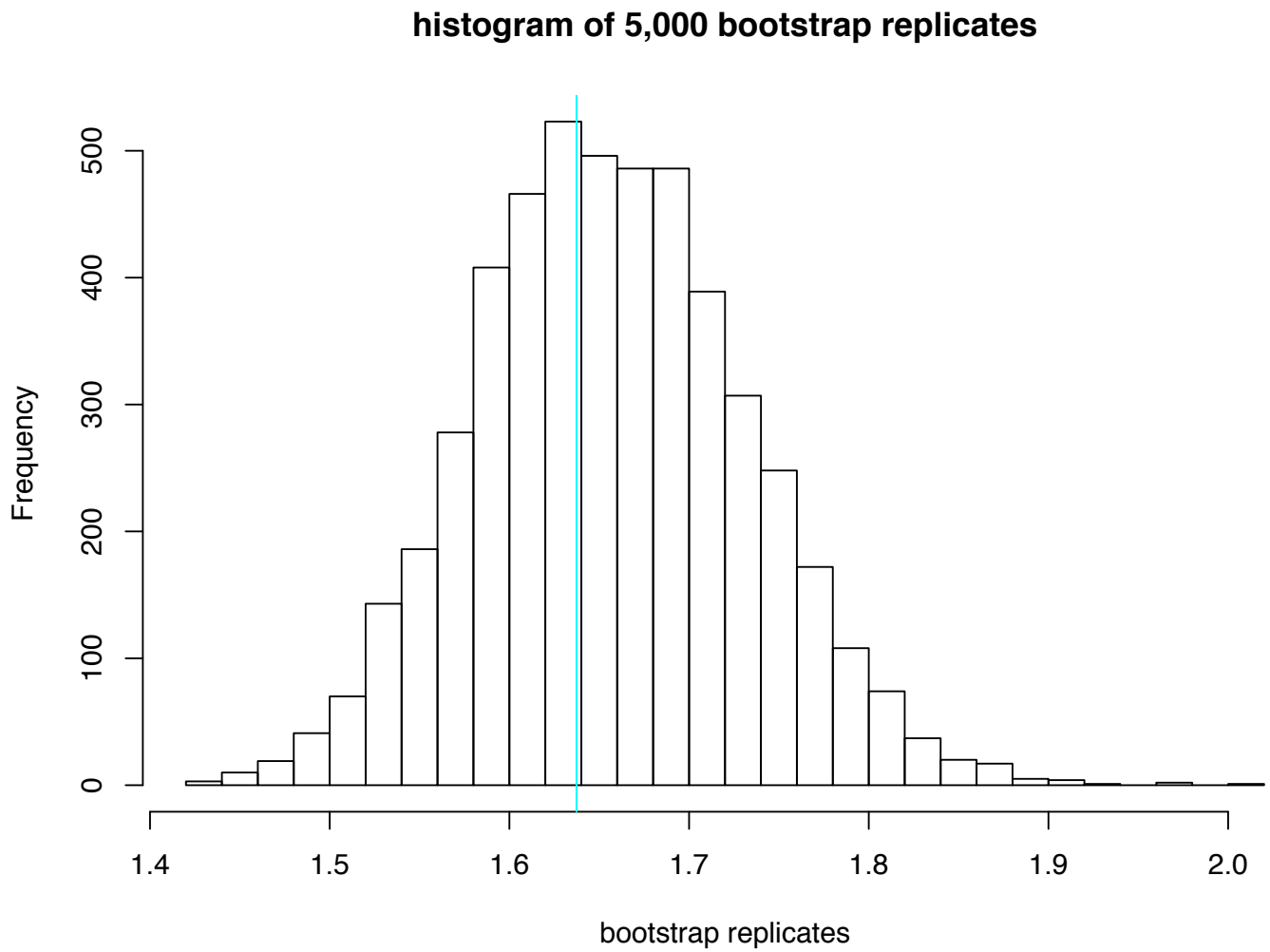
2. We then draw with replacement a sample of 1211+22,684 = 23,895 items from the first population and 732+22,917 = 23,649 items from the second -- each of these is called a bootstrap sample

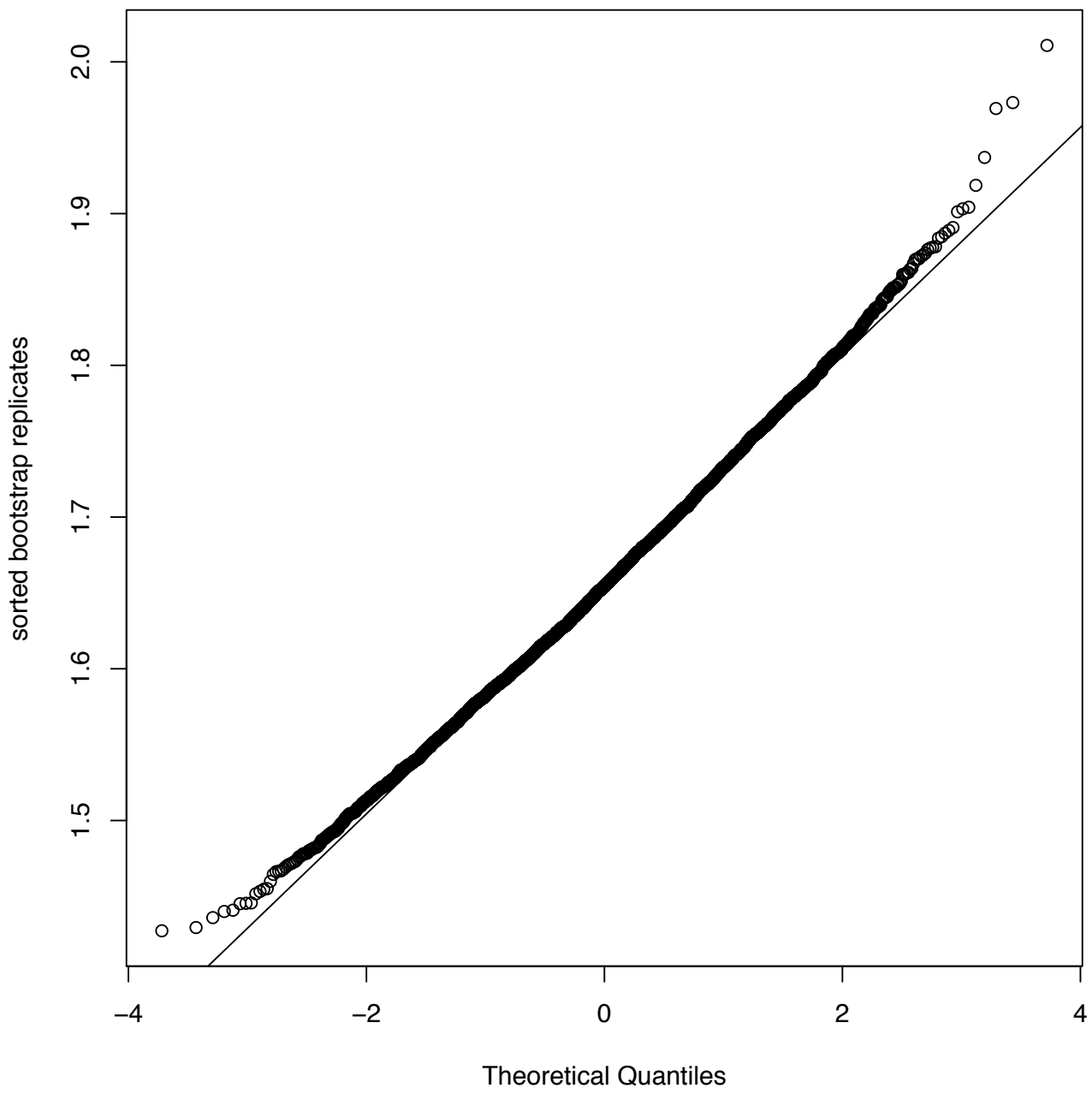
3. From these we we derive the bootstrap replicate of the relative risk

$$\widehat{r}* = \frac{\text{Proportion of ones in bootstrap sample } \#1}{\text{Proportion of ones in bootstrap sample } \#2}$$

4. Repeat steps 1-3 a large number of times, say 5,000, to obtain a set of bootstrap replicates $\widehat{r}*$

```
# some code to bootstrap the probability ratio
# 5,000 bootstrap samples

poptabs = subset(nyt$IfClicked,nyt$Variation=="Tabs")
poplist = subset(nyt$IfClicked,nyt$Variation=="List")

r = rep(0,5000)

for(i in 1:5000)
{

   btabs = sample(poptabs,replace=T)
   blist = sample(poplist,replace=T)

   bptabs = sum(btabs)/23895
   bplist = sum(btabs)/23649

   r[i] = bptabs/bplist
}

hist(r)
qqnorm(r)
```


histogram of 5,000 bootstrap replicates
Frequency
0
100
200
300
400
500
1.4
1.5
1.6
1.7
1.8
1.9
2.0
bootstrap replicates


normal probability plot of 5,000 bootstrap replicates
sorted bootstrap replicates
1.5
1.6
1.7
1.8
1.9
2.0
−4
−2
0
2
4
Theoretical Quantiles

Data and computing

While rerandomization is a fairly straightforward tool to motivate conceptually, I expected the bootstrap would be a harder “sell” -- It turns out that umbrella of “analyze as you randomized” helped make it seem natural

With the bootstrap**, students now have a rough and ready tool to assess the uncertainty in a huge number of problems**; this make it, again, a “thick” algorithm, useful pedagogically and practically

Data and computing

Rerandomization and resampling and subsample analysis all offer students an **intuition for statistical procedures**, one that is grounded in computation and one that is much broader than imagined by SW

The analytical approach is indirect, “by analogy” and seems to mimic the mathematical steps Fisher and Gosset followed when developing the techniques -- ultimately students are left with a formula, a “thin” algorithm at best

Once a computational base is established, you are really quite free to introduce a number of **“advanced” procedures** (multidimensional scaling, decision trees) that let us take a swing at more sophisticated statistical ideas

That said, we do have to circle back and hit the classics....

A small, antique example

To illustrate the use of the t-distribution, we are going to go **right back to the source**; the reason is that it will provide us with a bridge back to our computational approach and a look at a more recent set of problems

The data we will consider were originally collected by Charles Darwin; his experiment involved 15 pairs of plants (*Zea Mays*, a corn plant) **descended from the same parents, having exactly the same age and having been subjected "first to last to the same conditions"**

One individual from each pair was selected at random and cross-fertilized and the other self-fertilized; the **heights of the offspring** were then measured to the nearest eighth of an inch -- the results are on the next page

Pot	Crossed	Self-Fertilized	Difference
I	23.500	17.375	6.125
I	12.000	20.375	-8.375
I	21.000	20.000	1.000
II	22.000	20.000	2.000
II	19.124	18.375	0.749
II	21.500	18.625	2.875
III	22.125	18.625	3.500
III	20.375	15.250	5.125
III	18.250	16.500	1.750
III	21.625	18.000	3.625
III	23.250	16.250	7.000
IV	21.000	18.000	3.000
IV	22.125	12.750	9.375
IV	23.000	15.500	7.500
IV	12.000	18.000	-6.000

A small, antique example

Fisher analyzed these data by **first taking difference**s; under the assumption that plant heights in the two groups (self- and cross-fertilized) were each normally distributed, so that their difference, in turn, **had a normal distribution**

On the next page we present a normal Q-Q plot for the 15 differences; what do you notice about these values? What do you think about the normality assumption?

A small, antique example

Fisher computed the sample mean difference to be 2.62 inches (the offspring of cross-fertilized plants being taller than the self-fertilized plants by 2.62 inches) and found that the sample standard deviation is s = 4.72

With a sample size of n=15, the multiplier for the t-distribution for a 95% confidence interval is

```
> qt(0.975,df=14)
[1] 2.144787
```

This gives a 95% confidence interval of $\bar{x} \pm 2.14\, s/\sqrt{n}$ or [0.004,5.229];

We've exhibited the interval with so many significant digits to make a point -- if the confidence interval maps out a range of plausible values for the true difference in heights between cross- and self-fertilized corn plants, then what does this interval suggest?

A small, antique example

The fact that the t-distribution depends only on sample size also provides us with a way to test hypotheses; here is a simple example

In the case of differences, there is a point we consider special, zero; what we have done in the previous slide is to, essentially, **interpret the size of a difference** (the size of an "effect") **in terms of its standard error** (or an estimate of the standard error)

That is, while 2.62 inches may seem like a lot for a small plant, we are judging it's size relative to its standard error $4.72/\sqrt{15} = 1.21$; in other words, the size of the **effect of cross- versus self-fertilization is 2.62/1.21 or 2.15 standard errors from 0**

We could ask, under the null hypothesis that the population mean is zero (that there is really no difference in heights between the offspring of cross- and self-fertilized plants), how likely are we to see an effect this large or larger?

A small, antique example

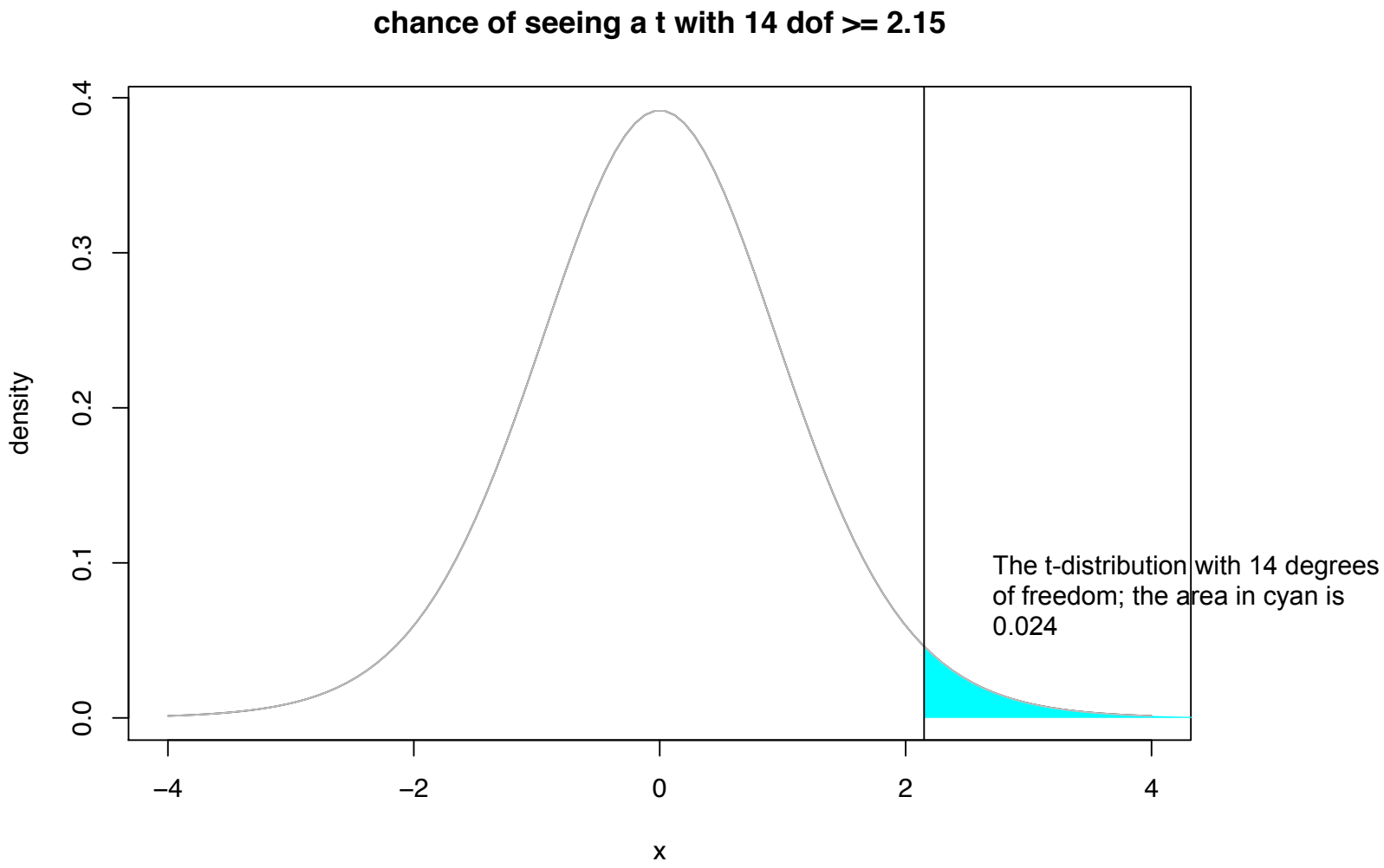
Remember that the quantity

$$\frac{\bar{x} - \mu}{s/\sqrt{n}}$$

depends only on sample size; so if we hypothesize $\mu = 0$ (in this case that the average difference in plant heights is zero), then the t-statistic

$$\frac{\bar{x}}{s/\sqrt{n}} = 2.15$$

should have a t-distribution with n-1 degrees of freedom; and so we can ask, what probability does the t-distribution assign to this event?


chance of seeing a t with 14 dof >= 2.15
density
0.0
0.1
0.2
0.3
0.4
−4
−2
0
2
4
x
The t-distribution with 14 degrees of freedom; the area in cyan is 0.024

Some comments

The last example we are using the **t-distribution as our reference distribution** for the test -- it is known as **a t-test** (or a paired t-test since we're working with differences of matched pairs)

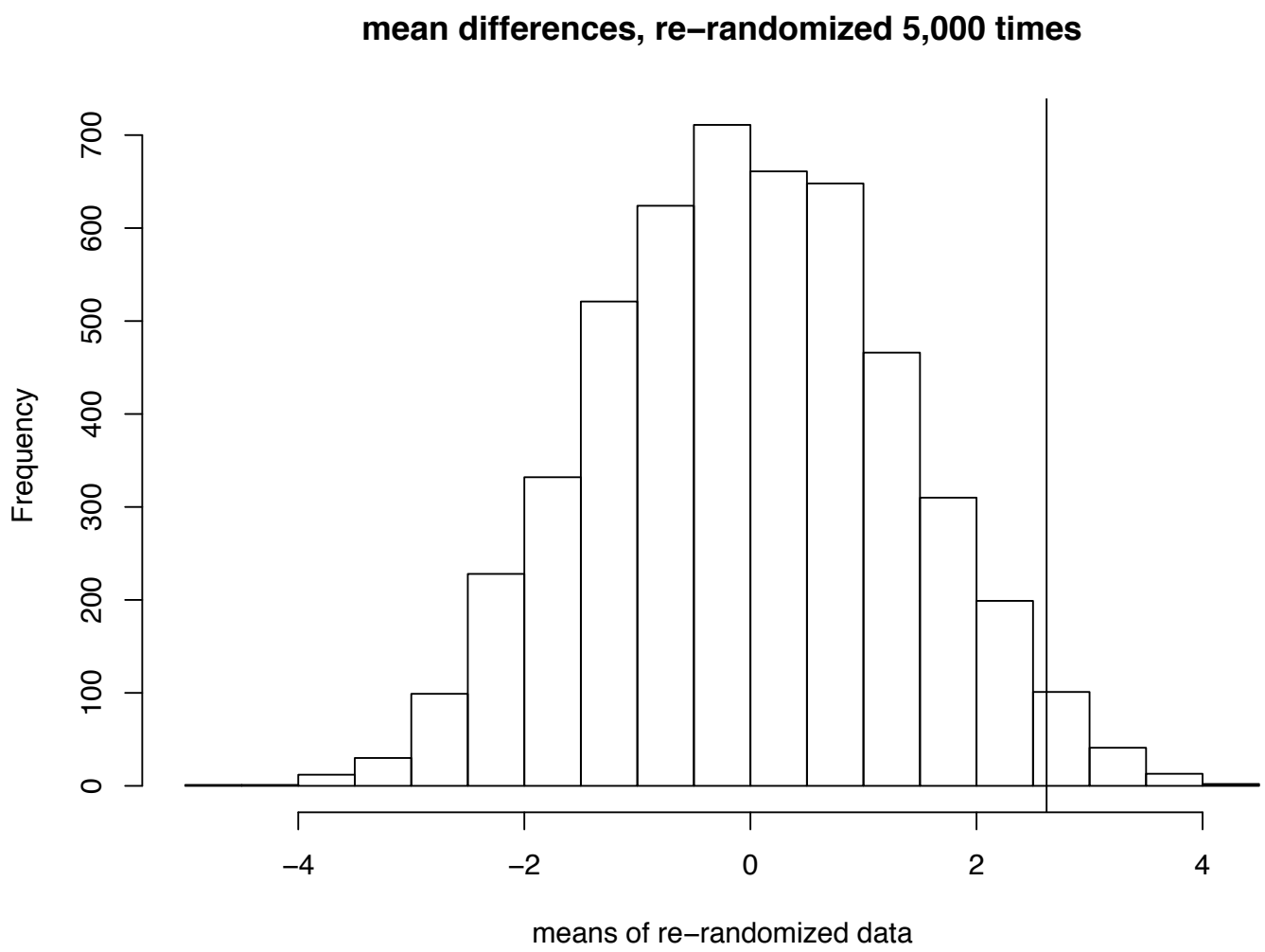
While this approach is classical, this is not how we would have approached the testing problem; instead, **we would have considered some form of re-randomization to generate a null distribution**

So... can we?

Some comments

Recall that the choice between cross- and self-fertilization was made at random; Fisher proposed re-randomizing each pair (so for each of the 15 rows in his original data table, we swap the two values on the toss of a coin) to come up with a null distribution for the sample mean

On the next slide we present the distribution of 5,000 re-randomized trials; the distribution looks fairly normal and we can (by now easily) compute a P-value...


mean differences, re−randomized 5,000 times
Frequency
0
100
200
300
400
500
600
700
−4
−2
0
2
4
means of re−randomized data

Some comments

The P-value we compute from 5,000 re-randomizations is 0.026, **in pretty good agreement with what we computed from the t-test**!

This is not surprising in that, along with the bootstrap and the other computational approaches we've been taking, **when the classical assumptions are met** (large sample sizes, or small sample sizes but normal populations) **the two approaches tend to agree**

The advantage of the computational procedure is that the same general principles ("analyze as you randomized") **apply to a variety of statistics and a range of applications**; the work we've presented on the previous 20 slides is all about one estimate, the sample mean

Epilogue

In 1912, Fisher was an undergraduate studying mathematics at Cambridge University and had just published his first paper “On an Absolute Criterion for Fitting Frequency Curves” in which he introduced the idea of the likelihood function

Noting an n versus n-1 in the formula for the standard deviation used by Gosset , Fisher was encouraged by his advisor to write to Gosset (Gosset was then 36, Fisher 22) and ultimately sent him his proof -- Gosset writes

> *This prooof, the tutor, made him send me and with some exertion I mastered it, spotted the fallacy (as I believe) and wrote him a letter showing, I hope, an intelligent interest in the matter and incidentally making a blunder. To this he replied with two foolscap pages covered with mathematics of the deepest dye in which he proved, by using n-dimensions that the formula, after all, involved n-1 and, of course, exposing my mistake. I couldn’t understand his stuff and wrote and said I was going to study it when I had time. I actually took it up to the lake with me - and lost it! Now he sends this to me [the mathematical proof of Student’s distribution]. It seemed to me that if it’s all right, perhaps you might like to put the proof in a note. It’s so nice and mathematical that it might appeal to some people.*

We will have more to say (briefly) about Gosset’s influence on the field next time...

... the sample is not sufficiently large to determine what is the law of distribution of individuals. It is usual, however, to assume a normal distribution... : since some law of distribution must be assumed, it is better to work with a curve whose area and ordinates are tabled and whose properties are well known. This assumption is accordingly made in the present paper so that its conclusions are not strictly applicable to populations known not to be normally distributed; yet it appears probable that the deviation from normality must be very extreme to lead to serious error.

- Student's 1908 paper

What I should like you to do is to find a solution for some other population than a normal one. It seems to me you might assume some sort of an equation for the frequency distribution of x which could lend itself to treatment besides the Gaussian..

- Letter to Fisher from Gosset

I have never known difficulty to arise in biological work from imperfect normality of the variation, often though I have examined data for this particular cause of difficulty; nor is there, I believe, any case to the contrary in the literature.

- Letter to Gosset from Fisher

Fisher is only talking through his hat when he talks of his experience; it isn't so very extensive and I bet he hasn't often put the matter to the test; how could he?

- Letter to E. Pearson from Gosset

The existence of these random numbers [Tippet's tables] opened out the possibility scarcely dreamed of before, of carrying out a great variety of experimental programmes, particularly answering questions in considerable depth and breadth the kind of questions about robustness of the 'normal theory' tests based on z (or t) raised by Gosset in his letter to me of 11 May 1926. This programme I started in 1927...

- From E. Pearson's memoir

History

It's probably clear, but throughout the course I attempt to **portray statistics as a real scientific discipline, one with a long history, a history of debate**

Take, for example, the uncomfortable **mash of significance and hypothesis testing** that is presented in SW....

Significance testing

As we noted last time, our use of P-values and our examination of the null distribution are in line with the methodology advocated by Fisher throughout his career; **the null hypothesis plays the role of devil's advocate, and a P-values provide evidence against the null** -- this is often called **significance testing**

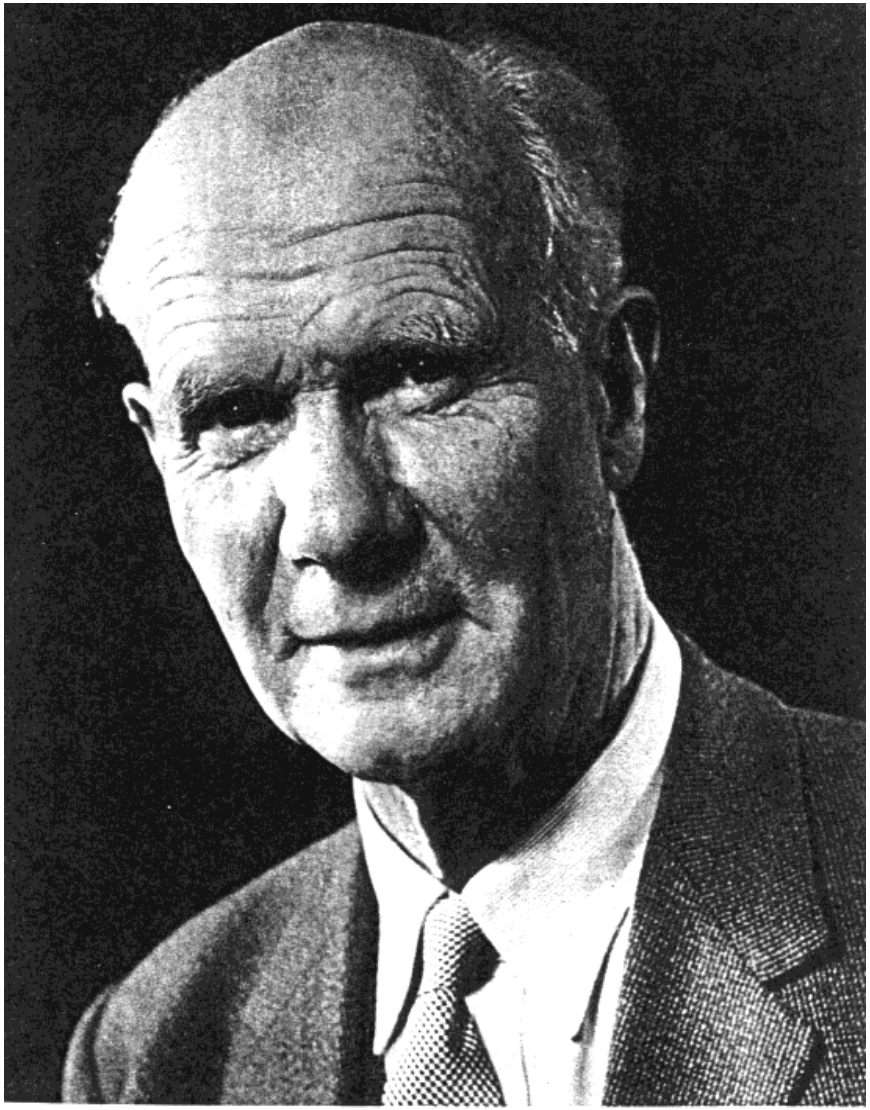
There are a few obvious questions facing practitioners, the first of which involves evaluating the evidence provided by a P-value; **is there a rule which helps you decide when you should "reject" the null hypothesis,** or, rather, decide that it's not true?

Fisher wrote: *If [the P-value] is between 0.1 and 0.9 there is certainly no reason to suspect the hypothesis tested. If it is below 0.02 it is strongly indicated that the hypothesis fails to account for the whole of the facts. We shall not often be astray if we draw a conventional line at 0.05...."* (Fisher 1950); and certainly in his own work on agricultural field trials, used **thresholds of 0.05 and 0.01** as guides to "reject" a null hypothesis

Still, Fisher believed that **the individual researcher should interpret a P-value** (a value of 0.05 might not lead to either belief or disbelief in the null, but to a decision to conduct another experiment); he wrote that the rigid use of thresholds was the *"result of applying mechanically rules laid down in advance; no thought is given to the particular case, and the tester's state of mind, or his capacity for learning, is inoperative."* (Fisher 1955, p.73-4).

J. N. Eyman

Courtesy of Berkeley Statistics Department Archives



EGON SHARPE PEARSON

"No test based upon a theory of probability, can by itself provide any valuable evidence of the truth or falsehood of a hypothesis. But we may look at the purpose of tests from another viewpoint. Without hoping to know whether each separate hypothesis is true or false, **we may search for rules to govern our behaviour** *with regard to them, in following which* **we insure that, in the long run of experience, we shall not often be wrong***."*

Neyman and Pearson, 1933

Significance v. hypothesis testing

Fisher's comment on rules "laid down in advance" was referring to the work of another very influential pair of statisticians, Jerzy Neyman and Egon Pearson

Neyman and Pearson disagreed with the subjective interpretation inherent in Fisher's approach and developed instead a procedure (which they termed hypothesis testing) based on **hard decisions about when to reject a null hypothesis;** in effect, they imposed **a threshold on the P-value called the significance level** -- they then studied what happens if you make decisions using these thresholds

Hypothesis testing, as it is covered in most introductory texts, **is a synthesis of Fisher's ideas (the P-value) together with the Neyman-Pearson framework**; we'll take some time to go over the synthesis now...

Summary

In approaching Statistics 13 this term, I took seriously the fact that students were receiving GE credit for my course -- This meant an emphasis on **history and critique** which, in turn, suggested descriptions that were intuitive and accessible and not indirect

Recall that rerandomization was introduced by Fisher who suggested that the t-test found its justification in its agreement with this computational procedure; it was also Fisher who wrote that the issues he discussed

> ... can be dissociated from all that is strictly technical in the statistician's craft, and, **when so detached**, are questions only of the right use of human reasoning powers, with which all intelligent people, who hope to be intelligible, are equally concerned, and on which the statistician, as such, speaks with no special authority.

In some sense, my approach this quarter has been less that of an "originalist" and instead treated Fisher (and statistical knowledge) as a "**living practice**"