Statistics & Science

Paving the Road to a Data-informed Future

Thomas A. Louis, PhD
Department of Biostatistics
Johns Hopkins Bloomberg School of Public Health
www.biostat.jhsph.edu/~tlouis/
tlouis@jhsph.edu

Outline

- Big Data
- Old and new worlds
- Road construction & maintenance
- Opportunities & Cautions
- Statistical Science
- Realizing the potentials of big data for science and society
- Coda

The drumbeat

- Popular media and science publications sound the drum:[1]

  > "Big Data" will drive our future, from translating genomic information into new cancer therapies to harnessing the Web for untangling complex social interactions or detecting infectious disease outbreaks

- Vast amounts of new data are being generated, and statistical analysis of this information will be increasingly central to decision-making
- *The New York Times*

February 11, 2012

The Age of Big Data

By STEVE LOHR

GOOD with numbers? Fascinated by data? The sound you hear is opportunity knocking.

[1] Davidian & Louis *Science* v336: p12

Old world, "small data"

- In-person and phone interviews
- Hard-copy data forms and questionnaires
- Basic laboratory detectors
- .....

New world, "big data"

Small data, plus

- Web forms
- Transaction records
- Social networks: phone & email records, Facebook "friends"
- Road toll records
- Border crossings

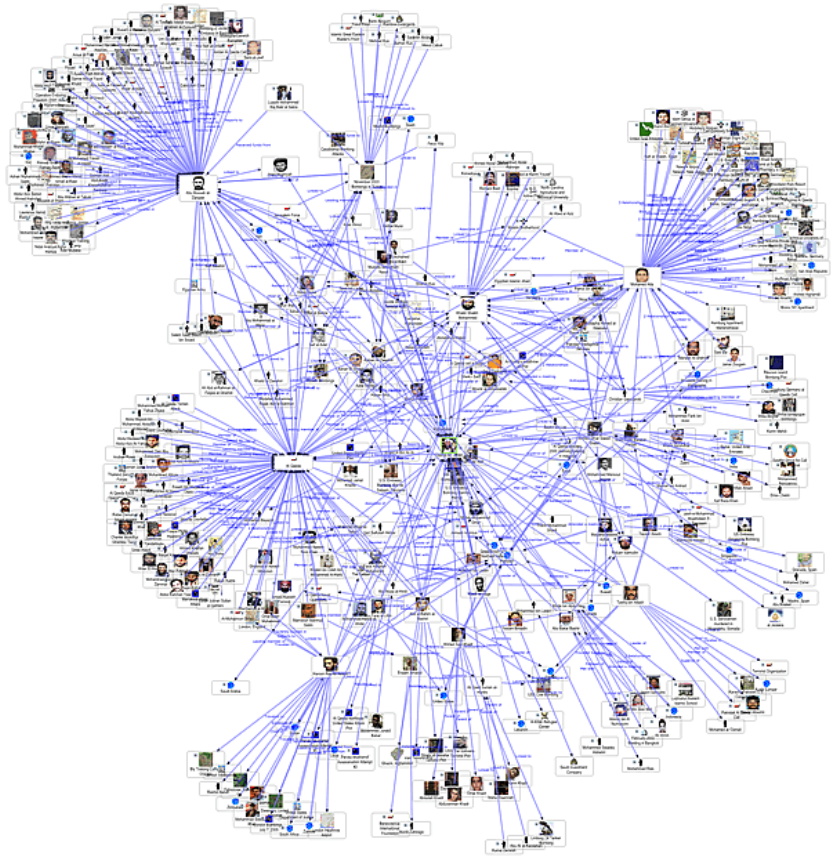
New world, "big data"

Small data, plus

- Web forms
- Transaction records
- Social networks: phone & email records, Facebook "friends"
- Road toll records
- Border crossings
- Gene chips & RFD chips
- Probes
- Cameras, movies
- Voice sampling
- Heat and other sensors
- Satellites & drones

Passive and Active; overt and covert

A social network[1]



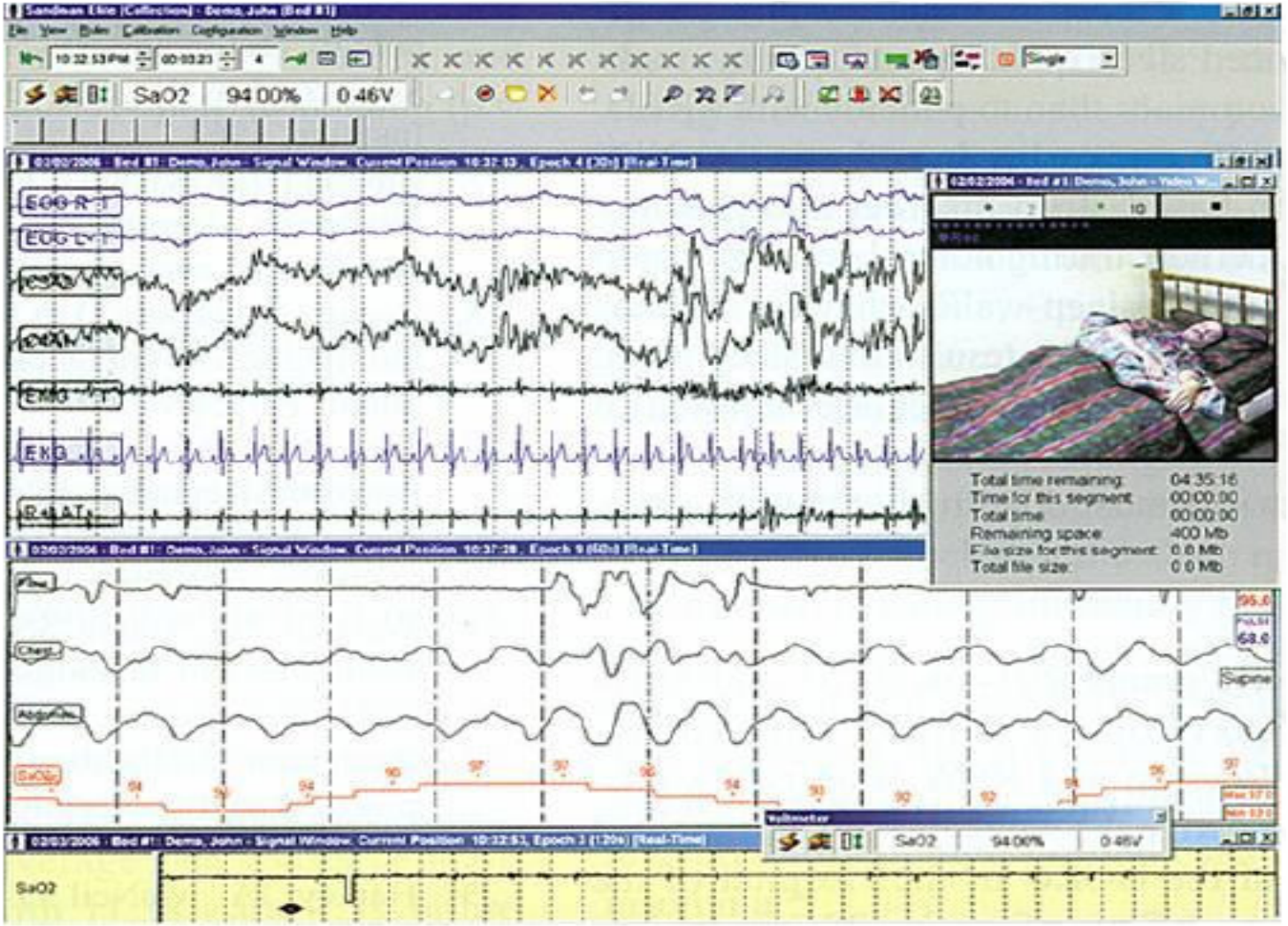
- Might also be a gene network, an electric grid, or . . .
- Graph theory in action!

[1] From the web

And more data Sources

- Facebook & Twitter & Linked-in
- Astronomical data streams
- Genomics
- E-government & administrative records
- Google and other search engines
  - Google Flu
- Amazon and other similar sites
- Acxiom Corporation, and other database marketers
- Audio and video streaming
- Brain and other images
- Brain and other waves
- Ocean temperature, wave and current monitors
- + + + +

Sleep Trajectories[1]



- Decomposition into "principal curves" that generalize principal components

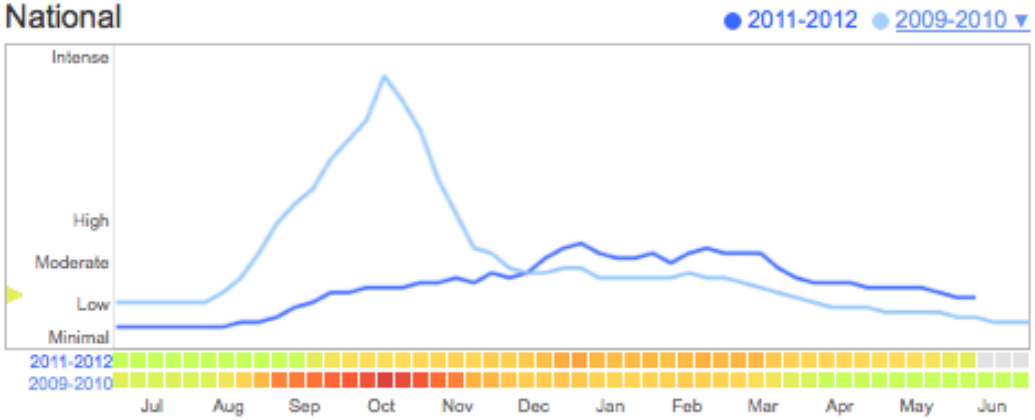
[1] From Ciprian Crainiceanu, Hopkins Biostatistics Faculty

Google Flu: can beat the CDC in predicting epidemics!

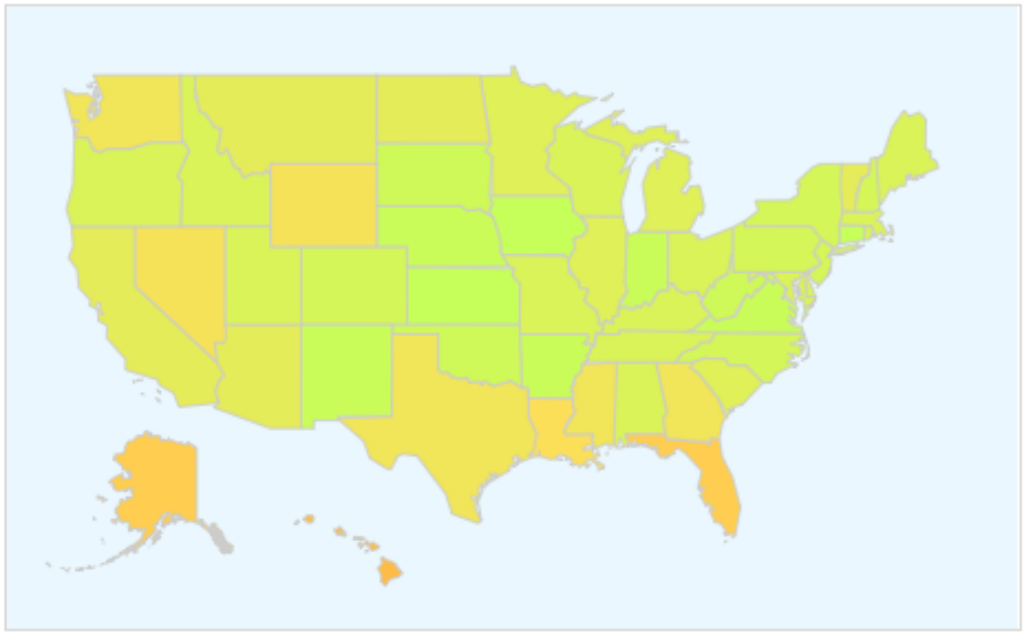
Explore flu trends - United States

We've found that certain search terms are good indicators of flu activity. Google Flu Trends uses aggregated Google search data to estimate flu activity. Learn more »

National



States | Cities (Experimental)



Road Construction & Maintenance

- Design of experiments and observational studies
- Data collection
- Storage & Husbandry
- Analysis & Reporting

Data Storage

Old World & Expensive

- Files & file cabinets
- Tape & physically BIG disks

New World & Relatively Low cost

- Storage media with LARGE capacity, that are small
- With FAST read and write
- Low in price, but with a stable price per pound!
- Inevitably, demand exceeds capacity
    - When you build a new highway, it gets crowded

Data Husbandry/Informatics

- Standards
  - Common definitions
  - Informative meta-data, annotation
- Data sharing with an infrastructure
  - Access controls
  - Disclosure protection

Data Analysis: Algorithmic, model-informed, model-based

The number and type of approaches burgeon, in part due to the very low cost of computing, and in part due to the need for innovative approaches

Golden Oldies

- t-test & confidence intervals, standard non-parametrics
- Linear & non-linear regression, contingency tables, . . .
- Fourrier Series
- Cross-validation, multiplicity control
- Jackknifing & Bootstrapping
- Bayesian methods

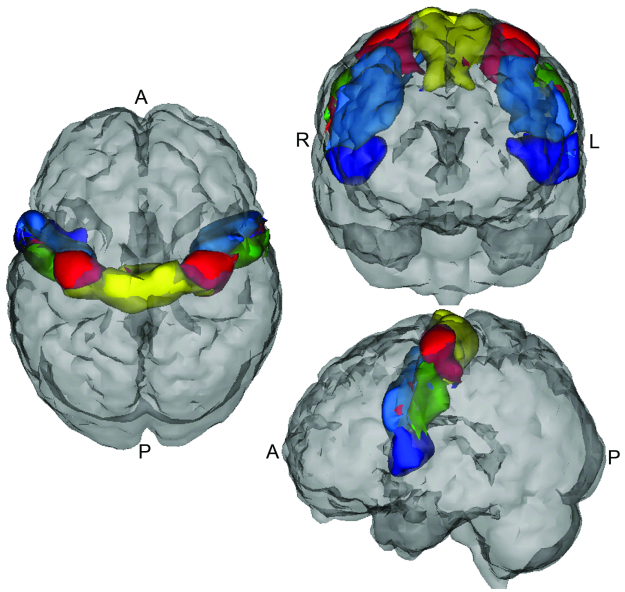
The new(ish) wave (some are "oldies on steroids")

- Causal analysis: propensity scores, principal stratification, ...
- Data mining, Knowledge Discovery, Analytics
- Functional decompositions: Splines, wavelets, movelets,
- The lasso
- Support vector machines
- Random Forests
- Slow Learning (Boosting/Bagging)
- Semi-parametric Bayes
- Genetic algorithms

ADHD Modeling via statistical decomposition[1]

Model 1

Build a statistical model for the population data based on a motor cortex parcellation and random forests for prediction. Use the model to predict ADHD diagnoses for a subject.



Prediction improved by 5 % as compared to the model using only demographic covariates.

[1] From Ani Eloyan, Hopkins Biostatistics Postdoc.
See http://www.smart-stats.org/

There are perils[1]

Big Data has its perils, to be sure. With huge data sets and fine-grained measurement, statisticians and computer scientists note, there is increased risk of "false discoveries." The trouble with seeking a meaningful needle in massive haystacks of data, says Trevor Hastie, a statistics professor at Stanford, is that "many bits of straw look like needles."

Big Data also supplies more raw material for statistical shenanigans and biased fact-finding excursions. It offers a high-tech twist on an old trick: I know the facts, now let's find 'em. That is, says Rebecca Goldin, a mathematician at George Mason University, "one of the most pernicious uses of data."

[1] *NYTimes*, February 11, 2012

Yes, care is needed

- "Big data" does not imply "big information" or "relevant information"
- **Science** requires uncovering causal relations, and while "Big Data" has produced interesting and important predictions and associations, care is needed to move from

discovery & association

to

causation

Yes, care is needed

- "Big data" does not imply "big information" or "relevant information"
- **Science** requires uncovering causal relations, and while "Big Data" has produced interesting and important predictions and associations, care is needed to move from

discovery & association

to

causation

- No amount of statistical creativity can rescue poorly designed studies and poor instrumentation
- Validity and success depend on statistical and domain-specific knowledge in design, conduct and analysis

Valid discovery depends on substantive knowledge

A very basic example

- You are analyzing two data points on wind direction (or molecule spin angles, or . . .)
- The reported values are 10° and 350°
- You report the "typical" direction

Valid discovery depends on substantive knowledge

A very basic example

- You are analyzing two data points on wind direction (or molecule spin angles, or . . .)
- The reported values are 10° and 350°
- You report the "typical" direction
  - Is it South: $180^\circ = (10^\circ + 350^\circ)/2$
  - Or, North: $0^\circ = (10^\circ + \{-10^\circ\})/2$
- What algorithm will always get it right?

Random sampling & Random Allocation

- These are mainstays of the scientific method
- Random sampling produces representativeness and impressive precision
  - Stratified sampling can produce $\pm 3\%$ in an national survey of less than 2000 individuals
- Random allocation produces statistical independence between treatment assignment and the attributes of individual "units"

Experimentation

- If you want to see what happens when you perturb a system, you must perturb it! **George Box**
- It is possible to learn from experience, but doing so is fraught with difficulties. **Lincoln Moses**
- Big Data does (do?) facilitate experimentation
  - Google conducts thousands of experiments, collecting millions of data points practically instantaneously
- But, yes, care is needed

The importance of design

- **Level 0 design**
  - To ensure that the investigative team is properly constituted and resourced
  - It's good to have a statistician involved at the very beginning
- **Level 1 and beyond**
  - For efficient use of resources
  - To address ethical considerations, for example by monitoring a clinical trial
  - To eliminate or to allow adjustment for batch effects, lab effects, technician effects . . .
  - To ensure that goals are achievable, avoiding the research equivalent of Russell Dodge's closing line in *Which way to East Millinocket*,

The importance of design

- **Level 0 design**
  - To ensure that the investigative team is properly constituted and resourced
  - It's good to have a statistician involved at the very beginning
- **Level 1 and beyond**
  - For efficient use of resources
  - To address ethical considerations, for example by monitoring a clinical trial
  - To eliminate or to allow adjustment for batch effects, lab effects, technician effects ...
  - To ensure that goals are achievable, avoiding the research equivalent of Russell Dodge's closing line in *Which way to East Millinocket*,

"Come to think of it, you can't get there from here"

Design for efficiency

- Have 8 unknown masses $M_1, \ldots, M_8$ and a balance scale
- Want to estimate their weights with precision equivalent to weighing each one 8 times
- Could weigh each one 8 times, producing

$$V(\hat{M}) = \frac{\sigma^2}{8}$$

  where σ^2 is the variance of a single weighing on the balance scale
- This approach requires $64 = 8 \times 8$ weighings
- Alternatively, consider placing some of the unknown masses on one side of the scale and some on the other, and analyzing the (Right - Left) weight differences

A design with an 8:1 advantage

Red = left side :: forestgreen = right side

Run	M_1	M_2	M_3	M_4	M_5	M_6	M_7	M_8
1	●	●	●	●	●	●	●	●
2	●	●	●	●	●	●	●	●
3	●	●	●	●	●	●	●	●
4	●	●	●	●	●	●	●	●
5	●	●	●	●	●	●	●	●
6	●	●	●	●	●	●	●	●
7	●	●	●	●	●	●	●	●
8	●	●	●	●	●	●	●	●

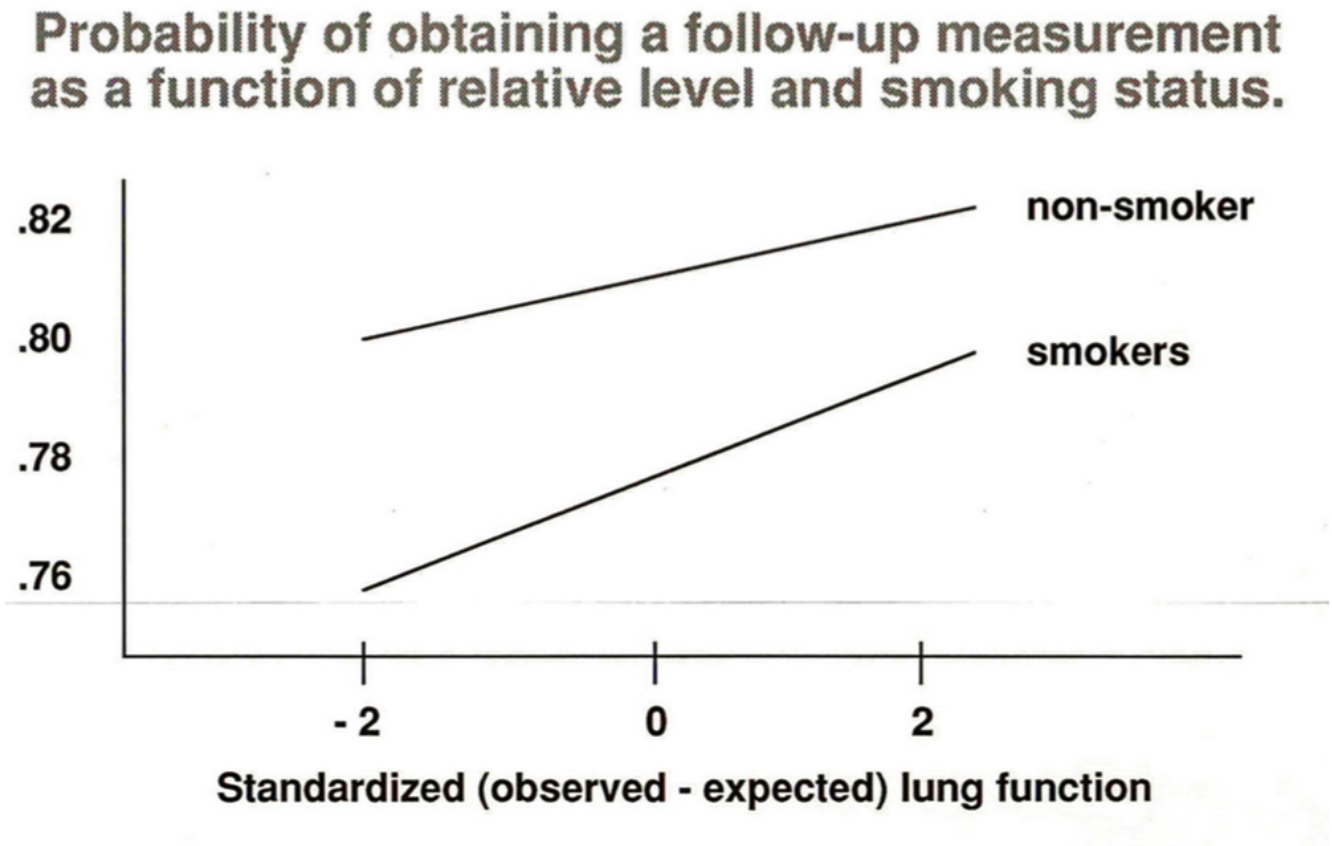
- Run 1 gives the total weight, and other runs are contrasts among the weights
- A linear model analysis produces the same $\sigma^2/8$ variance for each estimate based on 8 rather than 64 weighings

Importance of the sampling plan

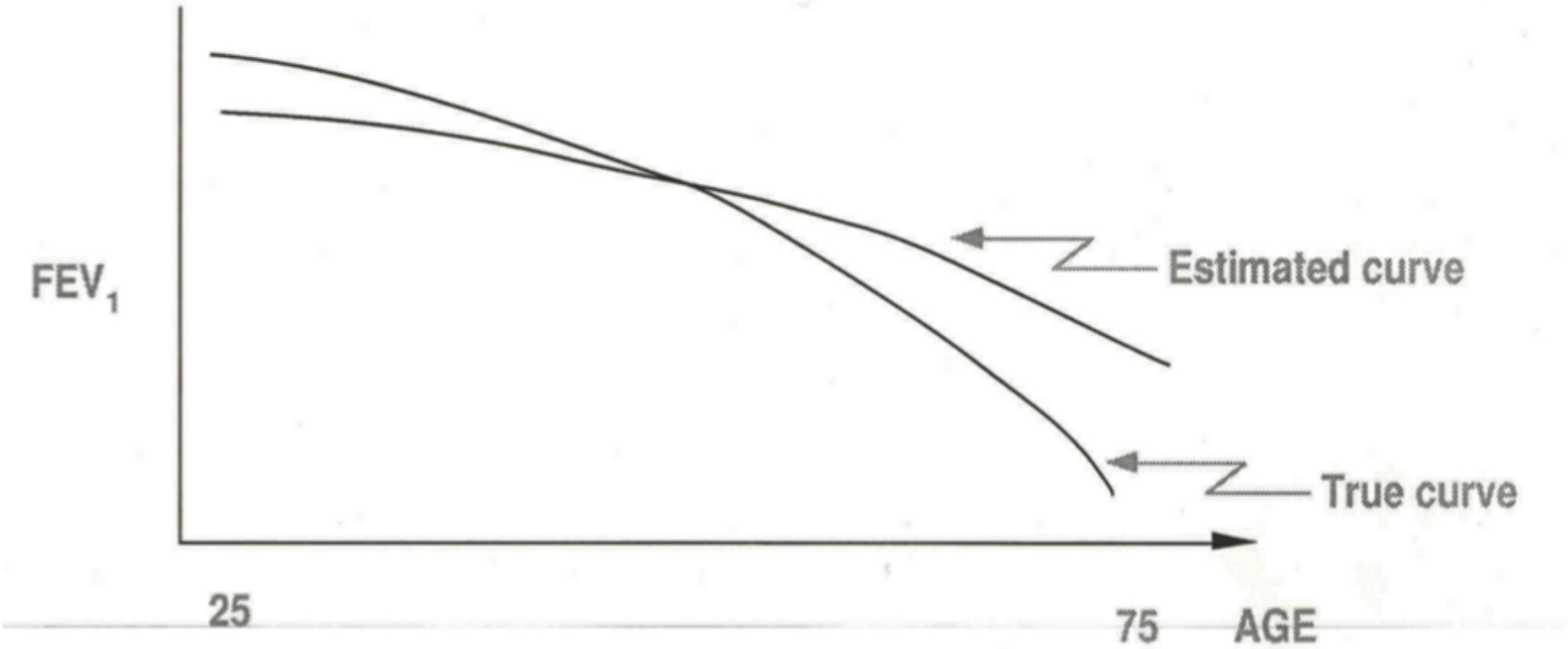
- There is always a sampling plan, but we may not know it
- Some sampling plans:
  - Random, stratified random, cluster, sno-ball
  - Haphazard, convenient, a series
  - "I have no idea"
- Selection effects and missing data affect the sampling plan

An informative missing data process

- Lung function data are not available "at random"



The healthy participant effect



Individuals are more likely to drop out if lung function decline is relatively steep or lung function relatively low.

Selection Effects: An apparent win for Clofibrate

Analyses not protected by randomization

- Association with adherence to treatment

Five-Year Mortality in Patients
(Coronary drug project*)

Adherence	Placebo	Clofibrate
<80%		24.6%
≥80%		15.0%
Total	19.4%	18.2%

*The Coronary Drug Project Research Group, NEJM. 1980;303:1038-41

However, look at the results for placebo!

Analyses not protected by randomization

- Association with adherence to treatment

Five-Year Mortality in Patients
(Coronary drug project*)

Adherence	Placebo	Clofibrate
<80%	28.2%	24.6%
≥80%	15.1%	15.0%
Total	19.4%	18.2%

- An "observational study" (covered on April 6th)

Confounding

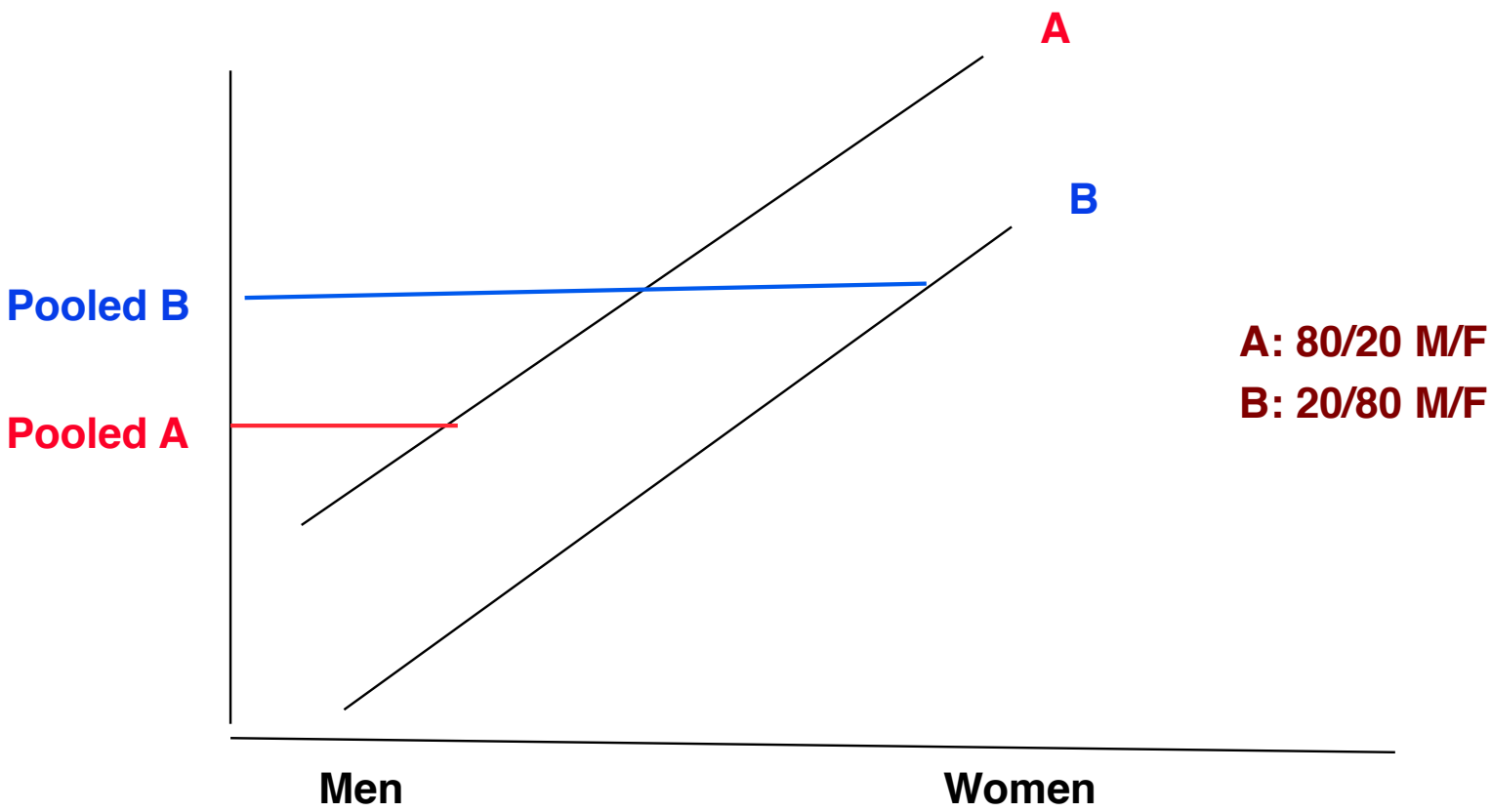
- Outcomes can be affected by experimental run/batch, lab tech/interviewer (on a specific day), clinics/community, ...
- Such associations can produce confounded relations that are not structural or causal
- For example, if cancerous tissues and healthy tissues are run on different arrays, differences gene expression can be either due to cancer/(no cancer) or to "array"
- (Stratified) randomization provides protection, and more generally a design that avoids complete confounding is needed

Simpson's paradox

"A treatment is good for men,
good for women, but bad for people!"

Simpson's Paradox

A treatment is good for men,
good for women, but bad for people



Simpson's paradox in action[1]

Victim	Defendant	Death penalty yes	Death penalty no	% yes
White	White	53	414	11.3
	Black	11	37	22.9
Black	White	0	16	0.0
	Black	4	139	2.8
	White	53	430	11.0
	Black	15	176	7.9

- For both white and black victims, black defendants received the death penalty a greater percentage of time than did white defendants
- However, pooled over victim status, white defendants received the death penalty a greater percentage of time than black defendants
  $\Longrightarrow$ Simpson's Paradox

[1] From Agresti, Categorical Data Analysis, second edition

Uncertainty: Estimate it; communicate it

- Accompany each statistical entity with comprehensive measures of uncertainty
- Make sure that they capture the influence of the full analytic process, not just the last step
- Some uncertainties are model-specific, for example a standard error or confidence interval,

Uncertainty: Estimate it; communicate it

- Accompany each statistical entity with comprehensive measures of uncertainty
- Make sure that they capture the influence of the full analytic process, not just the last step
- Some uncertainties are model-specific, for example a standard error or confidence interval,

$$17.2_{(3.1)} \text{ and } {}_{11.0}17.2_{23.4}$$

  - These formats communicate that uncertainty **is** part of the estimate; they are "connected at the hip"
- Some uncertainties result from variation induced by model choice or variable selection

Model Uncertainty in Low-dose extrapolation: Liver tumors in rats consuming 2AAF

Low dose model	"Virtually" Safe Dose (10^{-6} elevation)
Linear Multi-stage*	10^{-5}
Weibull* Gamma* Logit	10^{-2}
Probit	10^{-1}

* = model fits observed data

- Many models fit the observed data well, but produce order of magnitude differences in estimated safe dose
- Other data, scientific understanding, or a default policy are needed to deal with this uncertainty

Multiplicity & Personal Opinion/Goals

- Do you want to control the type I error or the (non)coverage of a confidence interval for,
  - This assessment of this endpoint?
  - All assessments of this endpoint?
  - A group of endpoints in a single study?
  - A group of studies?
  - Your research career?

Multiplicity & Personal Opinion/Goals

- Do you want to control the type I error or the (non)coverage of a confidence interval for,
  - This assessment of this endpoint?
  - All assessments of this endpoint?
  - A group of endpoints in a single study?
  - A group of studies?
  - Your research career?

There are procedures that will accomplish this last, but your career will likely be short!

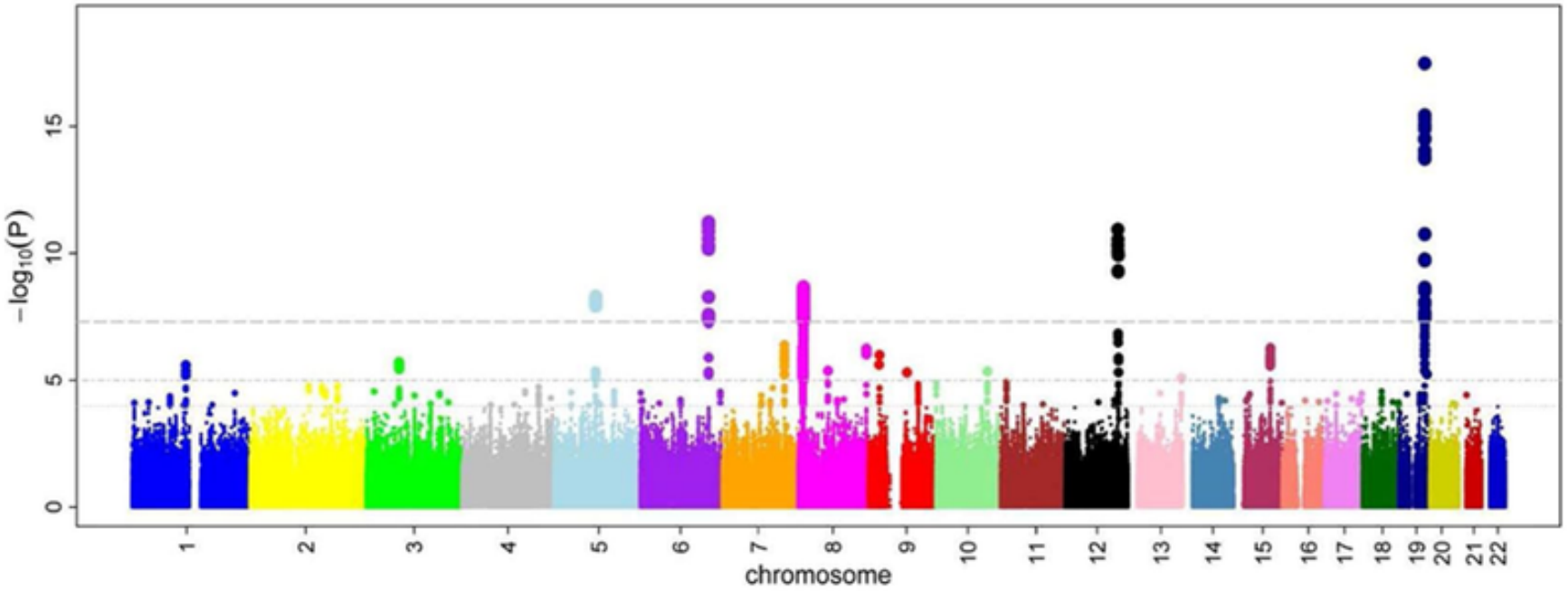
Multiplicity & Big Data

- Trevor Hastie, *NYTimes* February 11, 2012

  "...there is an increased risk of 'false discoveries.' The trouble with seeking a meaningful needle in massive haystacks of data, ..., is that 'many bits of straw look like needles."

- Old world: Family-wise error rate
  - Control the probability of even one false positive
- New World: False discovery rate
  - Control the prevalence of false positives among the positives
  - This is really old world (diagnostic testing) in a new-world context

Genomics:A Manhattan Plot[1]



- A genome-wide association study relating millions of SNPs to a phenotype
- X-axis is SNP location
- Y-axis is - log(P-value) for SNP-specific association
- Horizontal lines indicate statistical significance after adjustment for false discovery

[1] From the web

Big Data requires a culture of reproducible Research[1]

- For scientific, workload and political reasons, it is important to develop and implement a culture of reproducible research wherein there is an essentially seamless analytic system that starts with databases, feeds analyses that provide input to tables and graphs
- All assumptions, data and analyses are completely documented and if someone wants to reproduce an analysis (possibly with some changes) they can do so without disturbing the integrity of the system.
- Effective reproducibility enhances credibility and transparency, thereby benefitting science, policy and communication
- Confidentiality may impose limits on disclosure

[1] Mesirov, J. P. (2010). Computer science. accessible reproducible research. *Science* 327, 415–416

Well, just what is statistics?

- R. A. Fisher in 1948,

  "Biometry, the active pursuit of biological knowledge by quantitative methods"

- Davidian & Louis, *Science*, 2012, v336: p12
  - Statistics is the science of learning from data, of measuring, controlling, and communicating uncertainty. Thereby, it provides the navigation essential for controlling the course of scientific and societal advances.
  - The scientific and societal opportunities created by synergies between statistics and domain science, between statisticians and other scientists burgeon

And, why Statistics?

- I hope you now have at least some idea of why

Sociological imperatives

Mutual Respect & Joint Ownership

- The two-way street
  - Nothing is purely statistical, but most of science and policy has a statistical aspect; collaboration is essential
  - We all share responsibility for the statistical and domain-specific components of a study
  - Statisticians need to understand the topic under study, and collaborators need to understand the broad concepts of statistics
  - Statisticians need to invite/encourage joint ownership of statistical issues, and avoid communicating the collaboration-equivalent of,

Sociological imperatives

Mutual Respect & Joint Ownership

- The two-way street
  - Nothing is purely statistical, but most of science and policy has a statistical aspect; collaboration is essential
  - We all share responsibility for the statistical and domain-specific components of a study
  - Statisticians need to understand the topic under study, and collaborators need to understand the broad concepts of statistics
  - Statisticians need to invite/encourage joint ownership of statistical issues, and avoid communicating the collaboration-equivalent of,

"Only historians are allowed to reminisce

Educational Imperative

Mutual understanding

- Education of scientists, policy-makers, and the general public in the broad concepts of statistics and statistical thinking is essential for effective collaboration, and for informed evaluation of quantitative findings
- An excellent start is provided by,

  Cohn V, Cope D, Cohn Runkle D (2011), News and Numbers, A Writer's Guide to Statistics, 3^{rd} edition. Wiley-Blackwell

Coda

- These are exciting times and Big Data have a vast potential for generating knowledge and improving pubic policy,
- However,

Coda

- These are exciting times and Big Data have a vast potential for generating knowledge and improving pubic policy,
- However,

Space-age techniques will not rescue stone-age data, and "big data" can be "stone-age"

Coda

- These are exciting times and Big Data have a vast potential for generating knowledge and improving pubic policy,
- However,

Space-age techniques will not rescue stone-age data, and "big data" can be "stone-age"

- Therefore, as Davidian & Louis note,

"Embedding statistics in science and society will pave the route to a data-informed future, and statisticians must lead this charge"

THANK YOU

Questions?