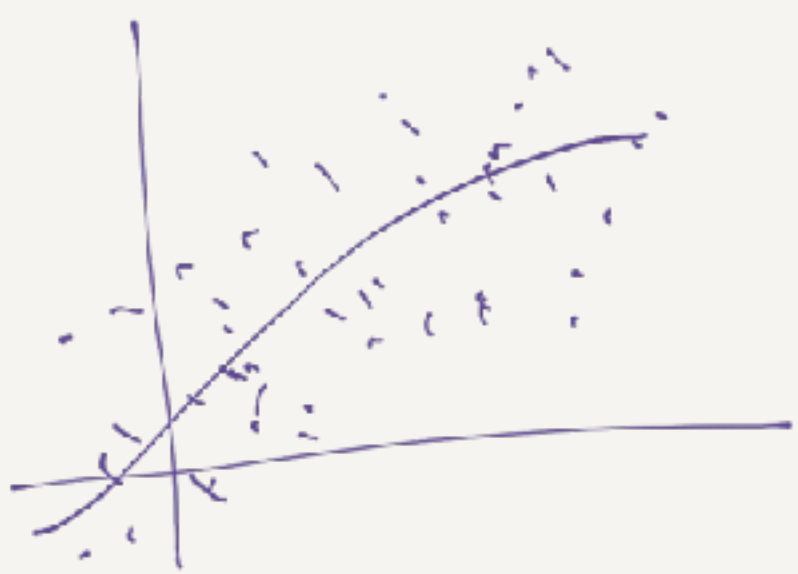
Data Science: The Good
The Bad
and the Future



Roger D. Peng

Department of Biostatistics

Johns Hopkins Bloomberg School of Public Health

@rdpeng

JOHNS HOPKINS
BLOOMBERG SCHOOL
of PUBLIC HEALTH

simplystats

simplystatistics.org

@simplystats

Not So Standard Deviations

(with Hilary Parker of Stitch Fix)

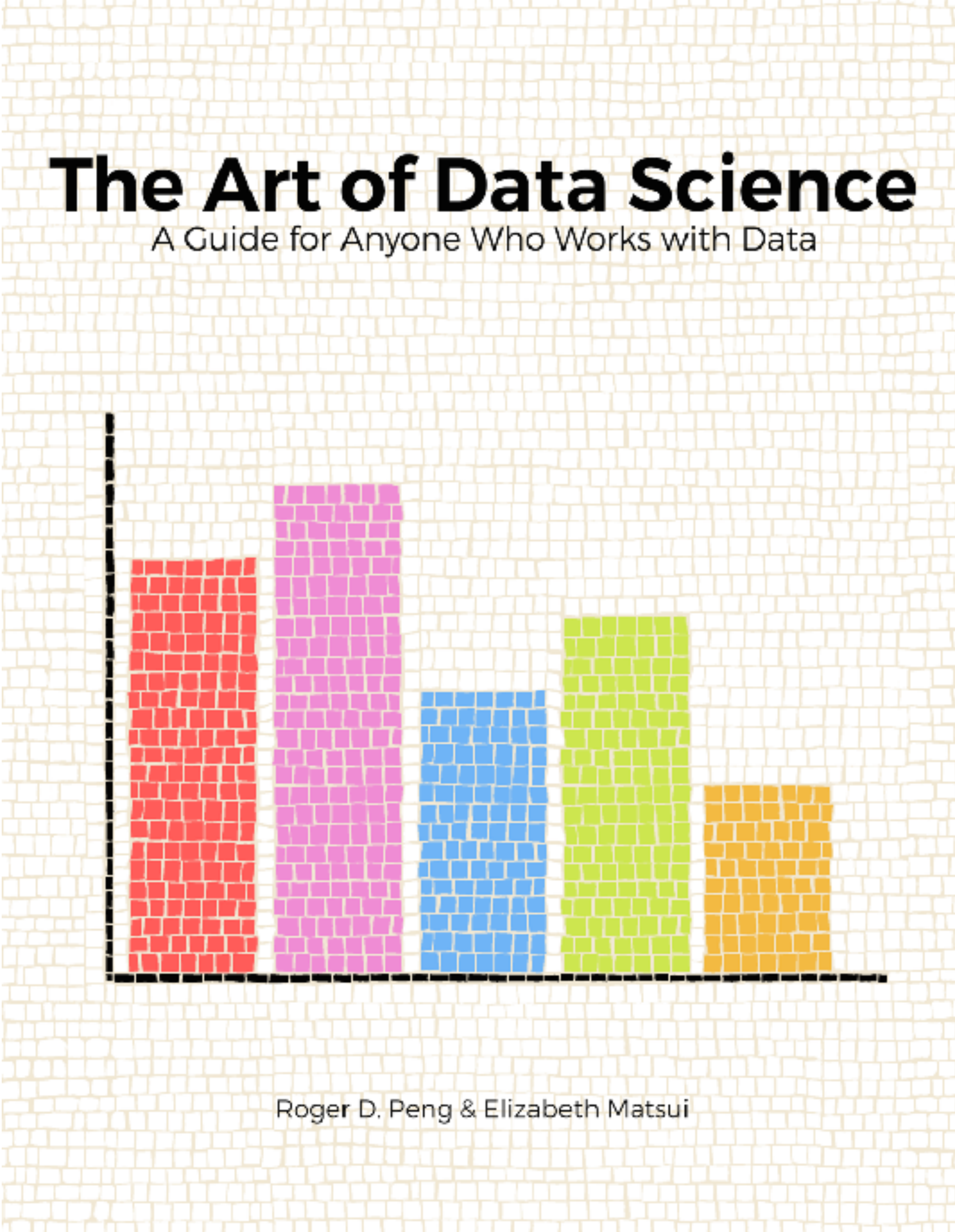
@NSSDeviations

https://soundcloud.com/nssd-podcast

Subscribe in iTunes: https://goo.gl/ZhWYbd

NSSD Episode 23

https://goo.gl/d8eszr



leanpub.com/artofdatascience

leanpub.com/conversationsondatascience

Protecting Health, Saving Lives —

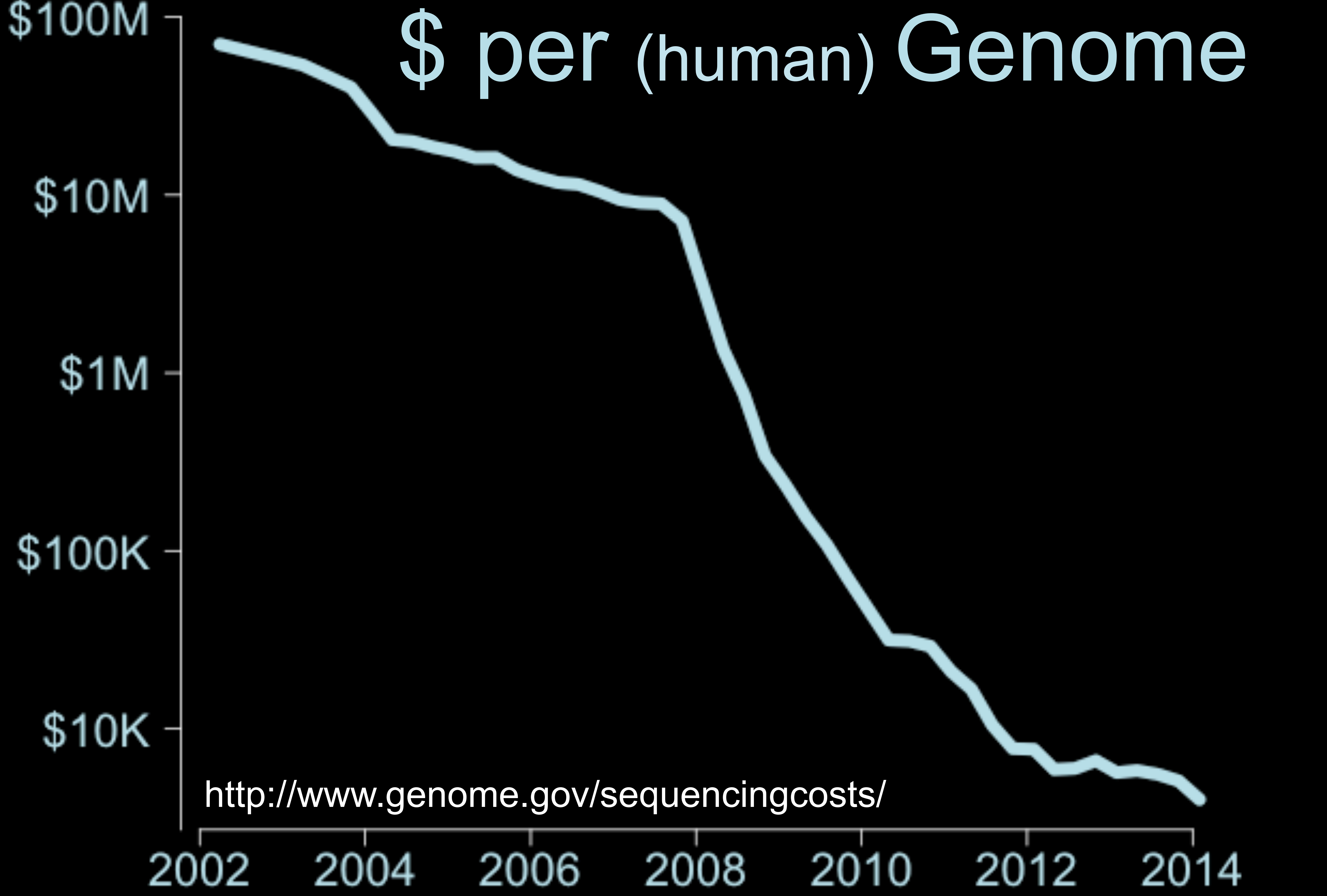
Protecting Health, Saving Lives —
Millions at a Time

Protecting Health, Saving Lives —

Millions at a Time

(of data points)

The Measurement Revolution…

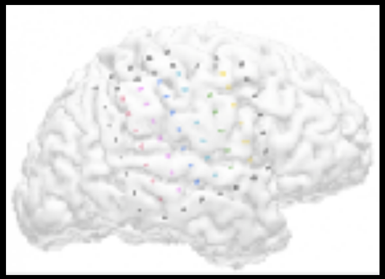


The Measurement Revolution…

Brian
Caffo

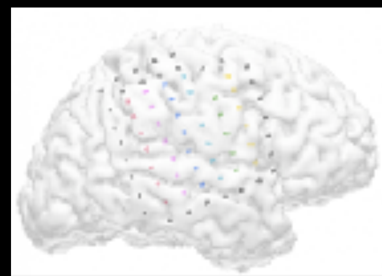
The Measurement Revolution…

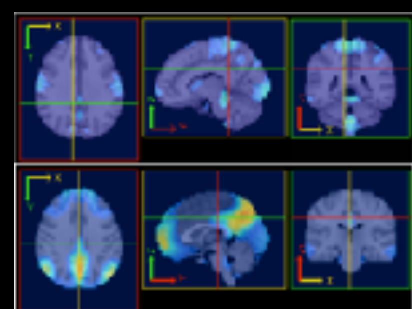
Brian
Caffo



The Measurement Revolution…

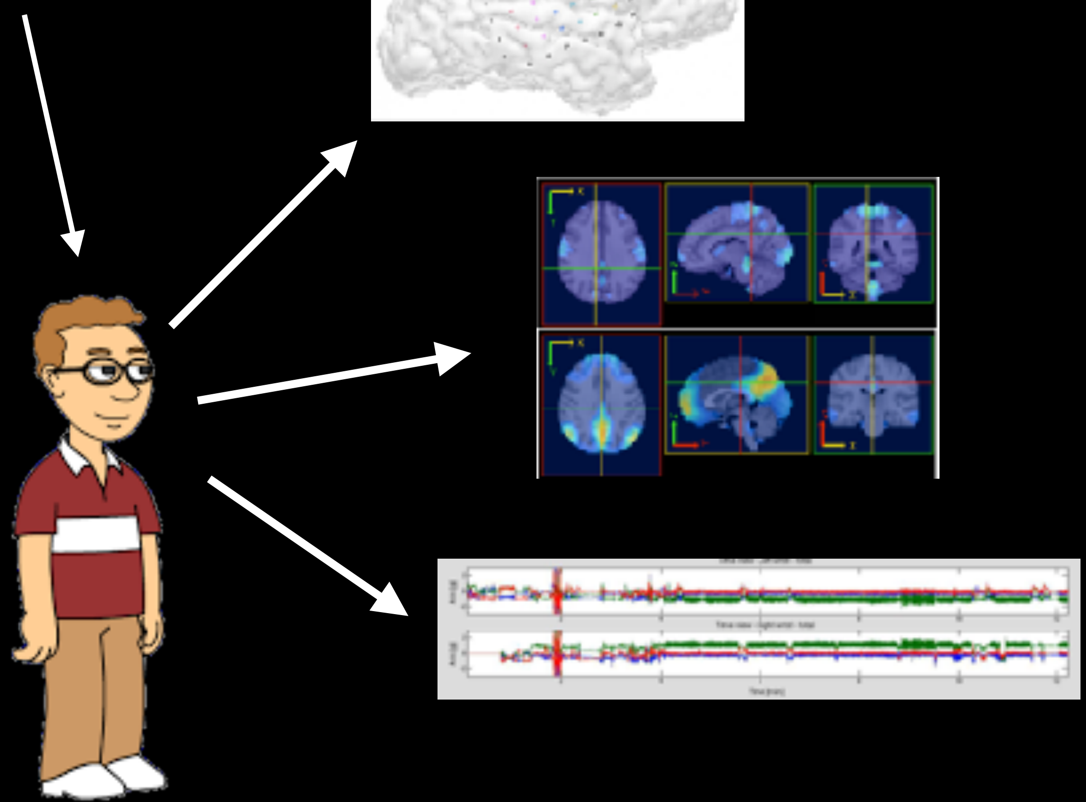
Brian
Caffo





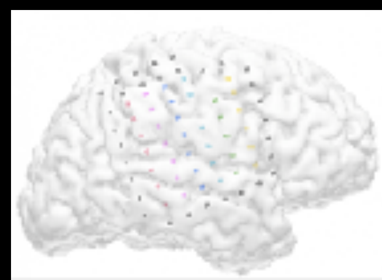
The Measurement Revolution…

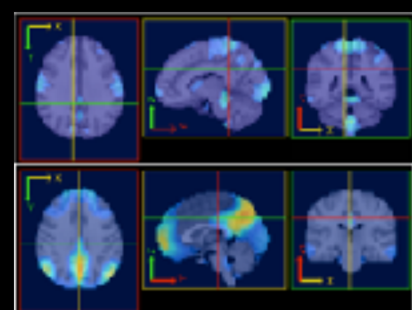
Brian Caffo

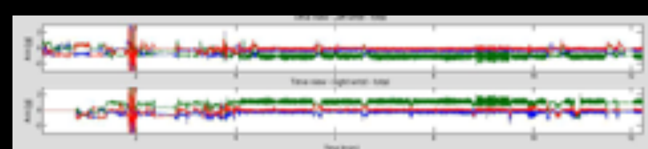


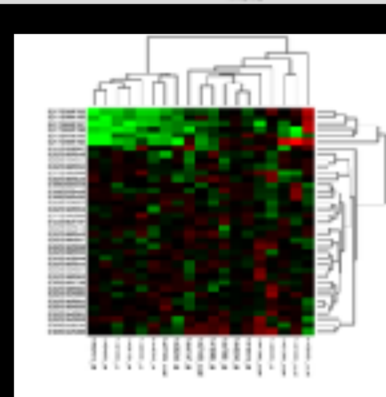
The Measurement Revolution…

Brian
Caffo



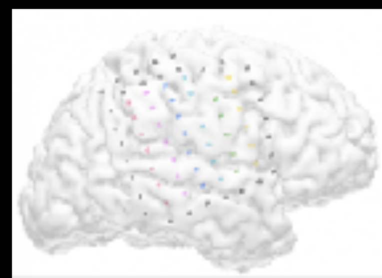


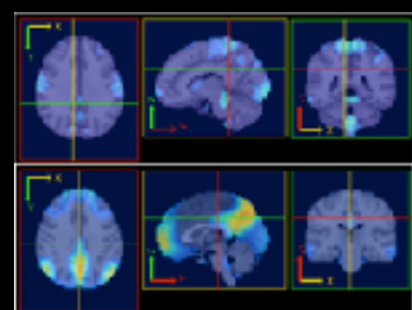


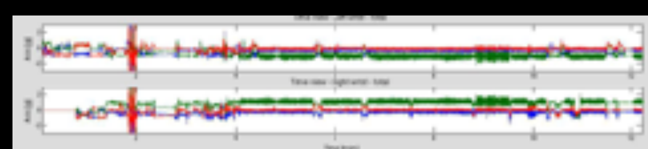


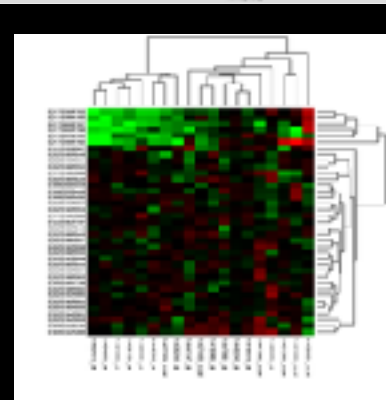
The Measurement Revolution…

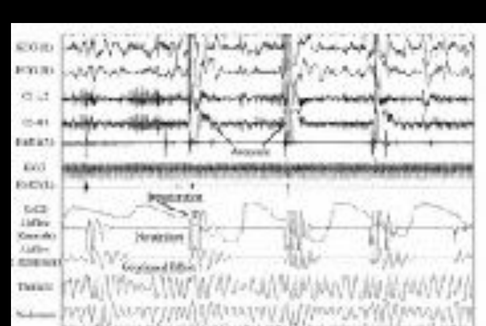
Brian
Caffo



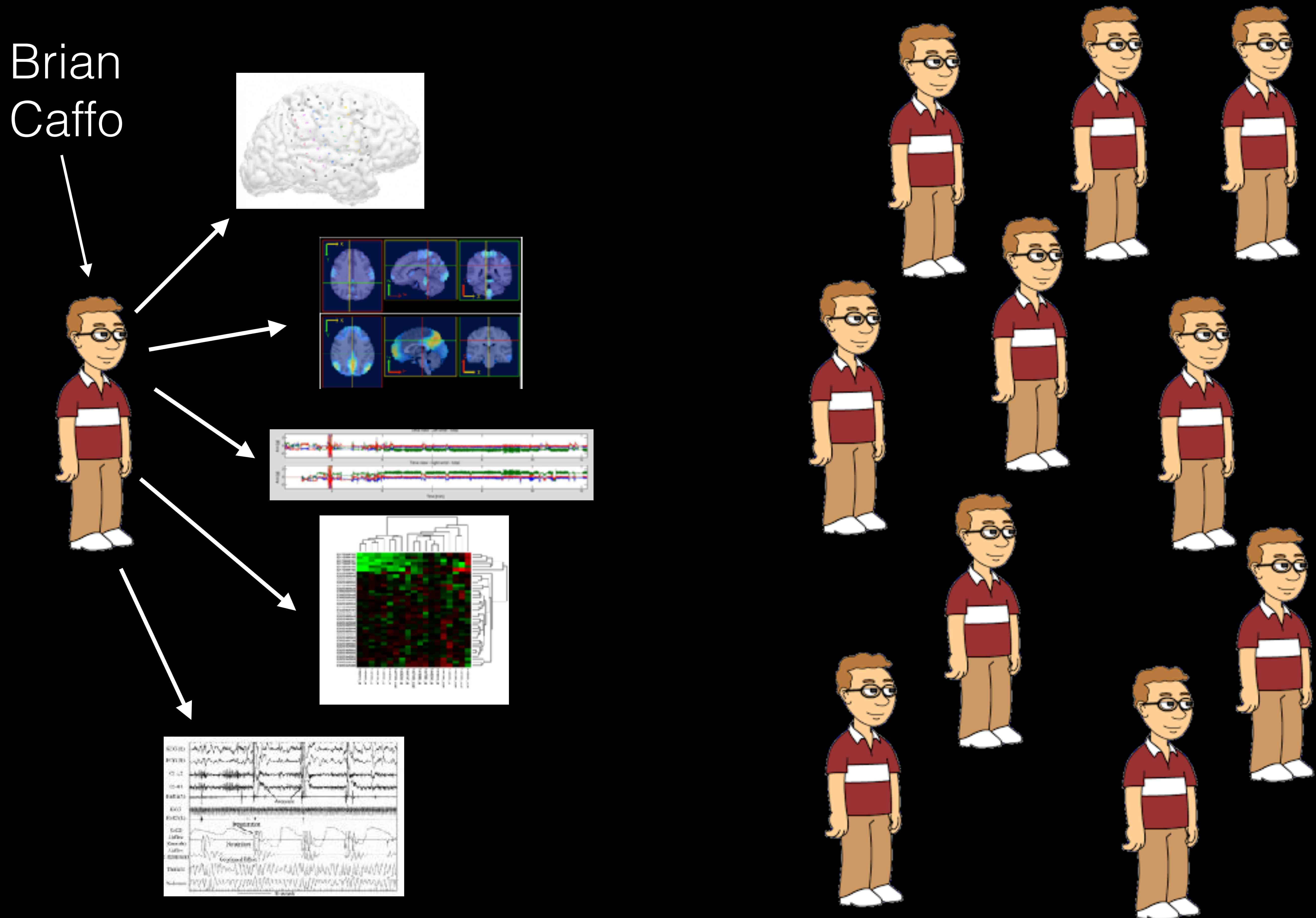








The Measurement Revolution… Multiplied!



Data is Eating the World*

*see “Software is Eating the World” by Marc Andreessen

Data is Eating the World*

(but analysis isn't yet)

*see "Software is Eating the World" by Marc Andreessen

Not Enough Geeks

06/28
2011

Critical Shortage Of "Data Geek" Talent Predicted By 2018

McKinsey&Company

New research by the McKinsey Global Institute (MGI) forecasts a 50 to 60 percent gap between the supply and demand of people with deep analytical talent. These "data geeks" have advanced training in statistics machine learning as well as the ability to analyz data sets. The study projects there will be approximately 140,000 to 190,000 unfilled positi data analytics experts in the U.S. by 2018 and a shortage of 1.5 million managers and analysts who have the ability to understand and make decisions using big data.

Not Enough Geeks

06/28
2011

Critical Shortage Of "Data Geek" Talent Predicted By 2018

McKinsey&Company

New research by the McKinsey Global Institute (MGI) forecasts a 50 to 60 percent gap between the supply and demand of people with deep analytical talent. These "data geeks" have advanced training in statistics machine learning as well as the ability to analyz data sets. The study projects there will be approximately 140,000 to 190,000 unfilled posit data analytics experts in the U.S. by 2018 and a shortage of 1.5 million managers and analysts who have the ability to understand and make decisions using big data.

Epidemic of Bad Data Analysis

Opinion: Reproducible research can still be wrong: Adopting a prevention approach

Jeffrey T. Leek[a,1] **and Roger D. Peng**[b]

[a]*Associate Professor of Biostatistics and Oncology and* [b]*Associate Professor of Biostatistics, Johns Hopkins University, Baltimore, MD*

Reproducibility—the ability to recompute results—and replicability—the chances other experimenters will achieve a consistent result—are two foundational characteristics of successful scientific research. Consistent findings from independent investigators are the primary means by which scientific evidence accumulates for or against a hypothesis. Yet, of late, there has been a crisis of confidence among researchers worried about the rate at which studies are either reproducible or replicable. To maintain the integrity of science research and the public's trust in science, the scientific community must ensure reproducibility and replicability by engaging in a more preventative approach that greatly expands data analysis education and routinely uses software tools.

We define reproducibility as the ability to been some very public failings of reproducibility across a range of disciplines from cancer genomics (3) to economics (4), and the data for many publications have not been made publicly available, raising doubts about the quality of data analyses. Popular press articles have raised questions about the reproducibility of all scientific research (5), and the US Congress has convened hearings focused on the transparency of scientific research (6). The result is that much of the scientific enterprise has been called into question, putting funding and hard won scientific truths at risk.

From a computational perspective, there are three major components to a reproducible and replicable study: (*i*) the raw data from the experiment are available, (*ii*) the statistical code and documentation to reproduce the computational tools such as knitr, iPython notebook, LONI, and Galaxy (8) have simplified the process of distributing reproducible data analyses.

Unfortunately, the mere reproducibility of computational results is insufficient to address the replication crisis because even a reproducible analysis can suffer from many problems—confounding from omitted variables, poor study design, missing data—that threaten the validity and useful interpretation of the results. Although improving the reproducibility of research may increase the rate at which flawed analyses are uncovered, as recent high-profile examples have demonstrated (4), it does not change the fact that problematic research is conducted in the first place.

The key question we want to answer when seeing the results of any scientific study is "Can I trust this data analysis?" If we think of problematic data analysis as a disease, reproducibility speeds diagnosis and treatment in

Epidemic of Bad Data Analysis

Opinion: Reproducible research can still be wrong: Adopting a prevention approach

Jeffrey T. Leek[a,1] and Roger D. Peng[b]

"The best way to prevent poor data analysis in the scientific literature is to (i) increase the number of trained data analysts in the scientific community and (ii) identify statistical software and tools that can be shown to improve reproducibility and replicability of studies."

integrity of science research and the public's trust in science, the scientific community must ensure reproducibility and replicability by engaging in a more preventative approach that greatly expands data analysis education and routinely uses software tools.

We define reproducibility as the ability to

question, putting funding and hard won scientific truths at risk.

From a computational perspective, there are three major components to a reproducible and replicable study: (*i*) the raw data from the experiment are available, (*ii*) the statistical code and documentation to reproduce the

problematic research is conducted in the first place.

The key question we want to answer when seeing the results of any scientific study is "Can I trust this data analysis?" If we think of problematic data analysis as a disease, reproducibility speeds diagnosis and treatment in

Challenges:

1. Measurement explosion!
2. Not enough analysts!
3. Reproducibility crisis!

What should
we do???

LEAD!

What is Data Science?

- Formulating a question that can be answered with data
- Assembling, cleaning, tidying data relevant to a question
- Exploring data, checking, eliminating hypotheses
- Developing a (statistical) model
- Making statistical inference
- Communicating findings

What does a data scientist do?

1. Focus the Question
2. Define the Measurement
3. Manage the process

Focus the Question

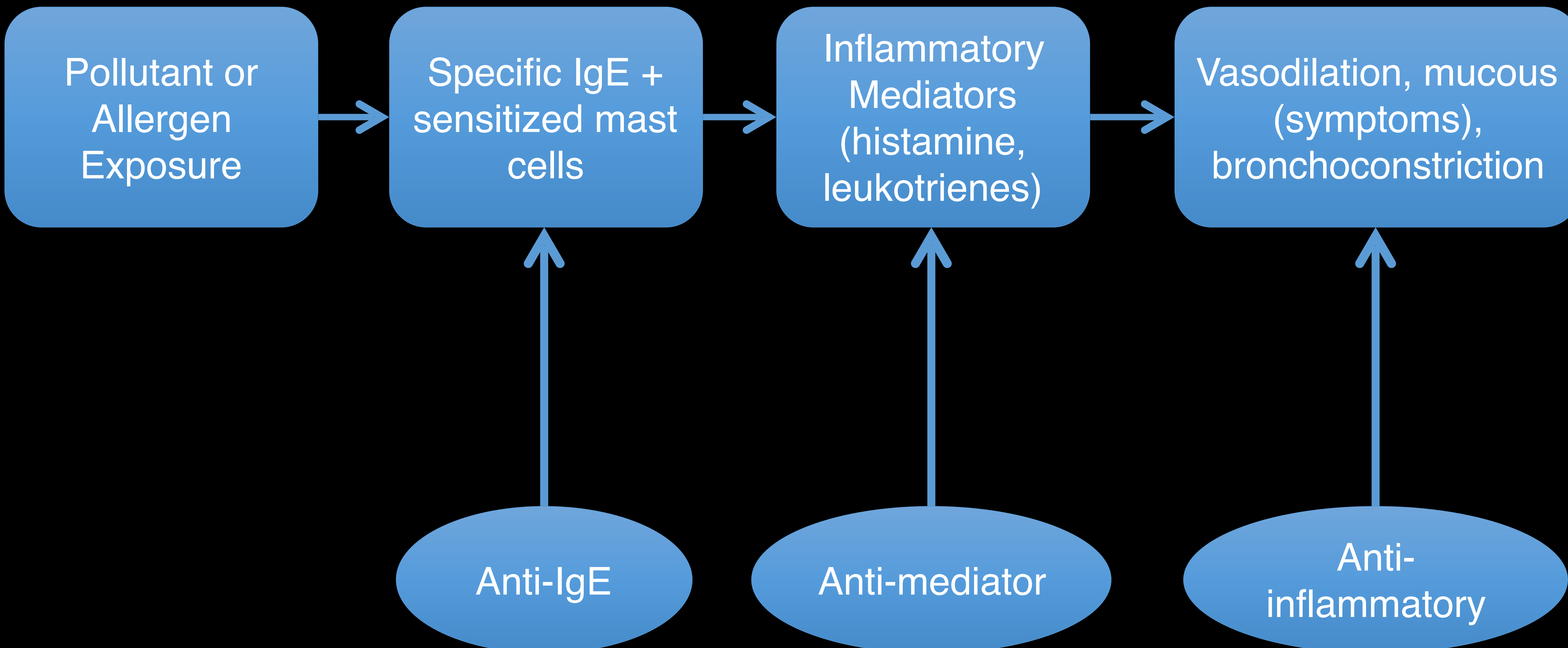
Inner-city Childhood Asthma

- Chronic inflammatory disorder of the airways
- Inflammation associated with (1) airways hyper-responsiveness; (2) airflow limitation (at least partially reversible); (3) respiratory symptoms (wheeze, cough)
- Airway inflammation can be present even in mild disease
- Racial/ethnic minorities often comprise majority of residents in inner-cities
- Asthma prevalence rates 25-28% in inner-cities

Environmental Intervention and (Allergic) Asthma Pathogenesis

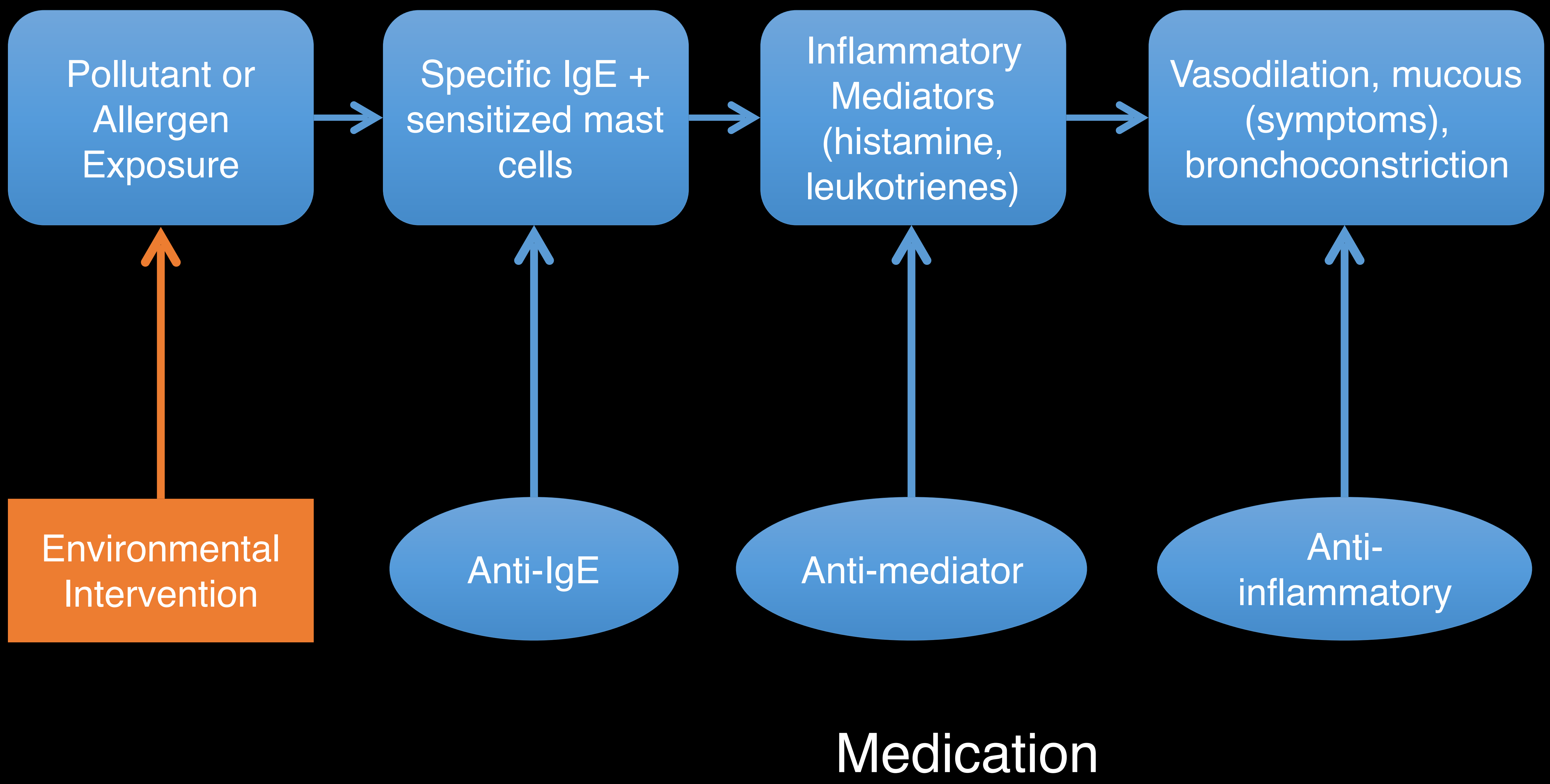
Pollutant or Allergen Exposure → Specific IgE + sensitized mast cells → Inflammatory Mediators (histamine, leukotrienes) → Vasodilation, mucous (symptoms), bronchoconstriction

Environmental Intervention and (Allergic) Asthma Pathogenesis



Medication

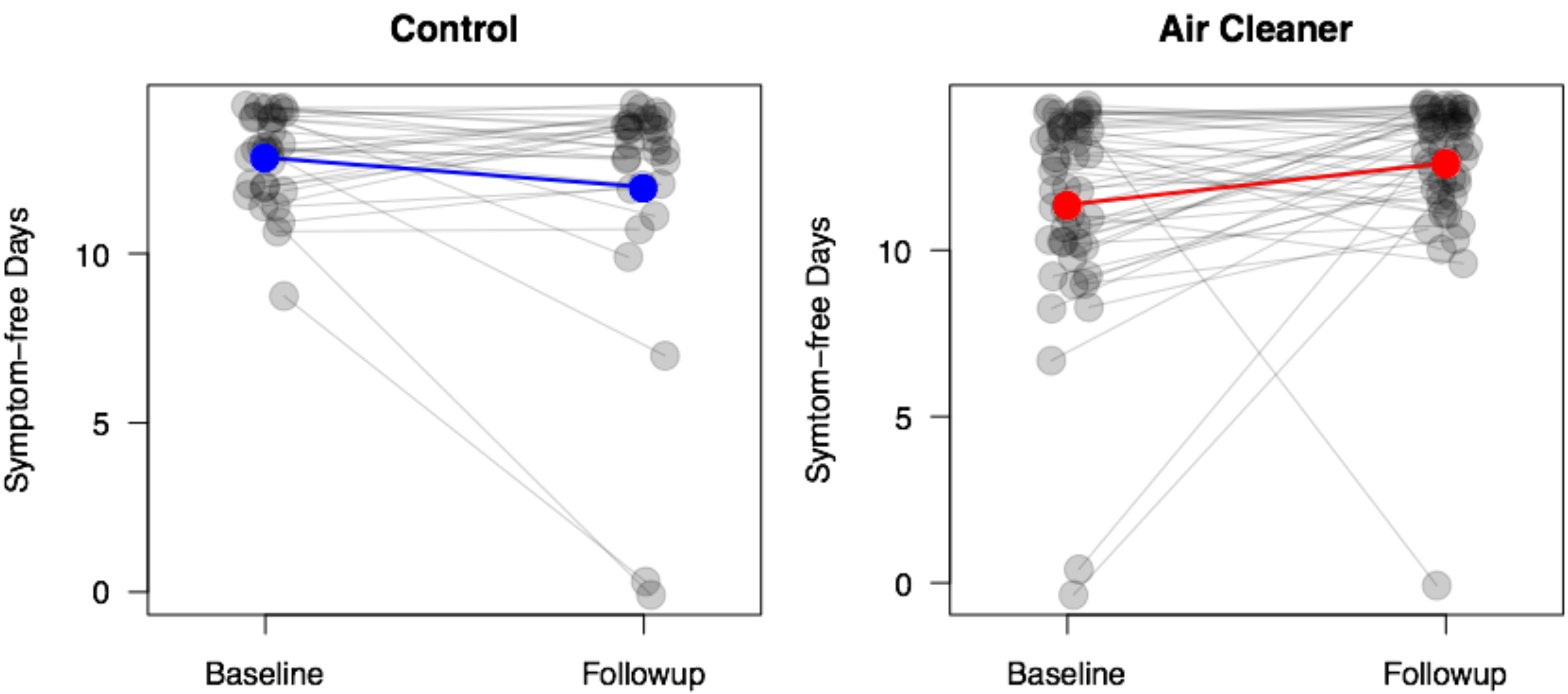
Environmental Intervention and (Allergic) Asthma Pathogenesis



PREACH Study

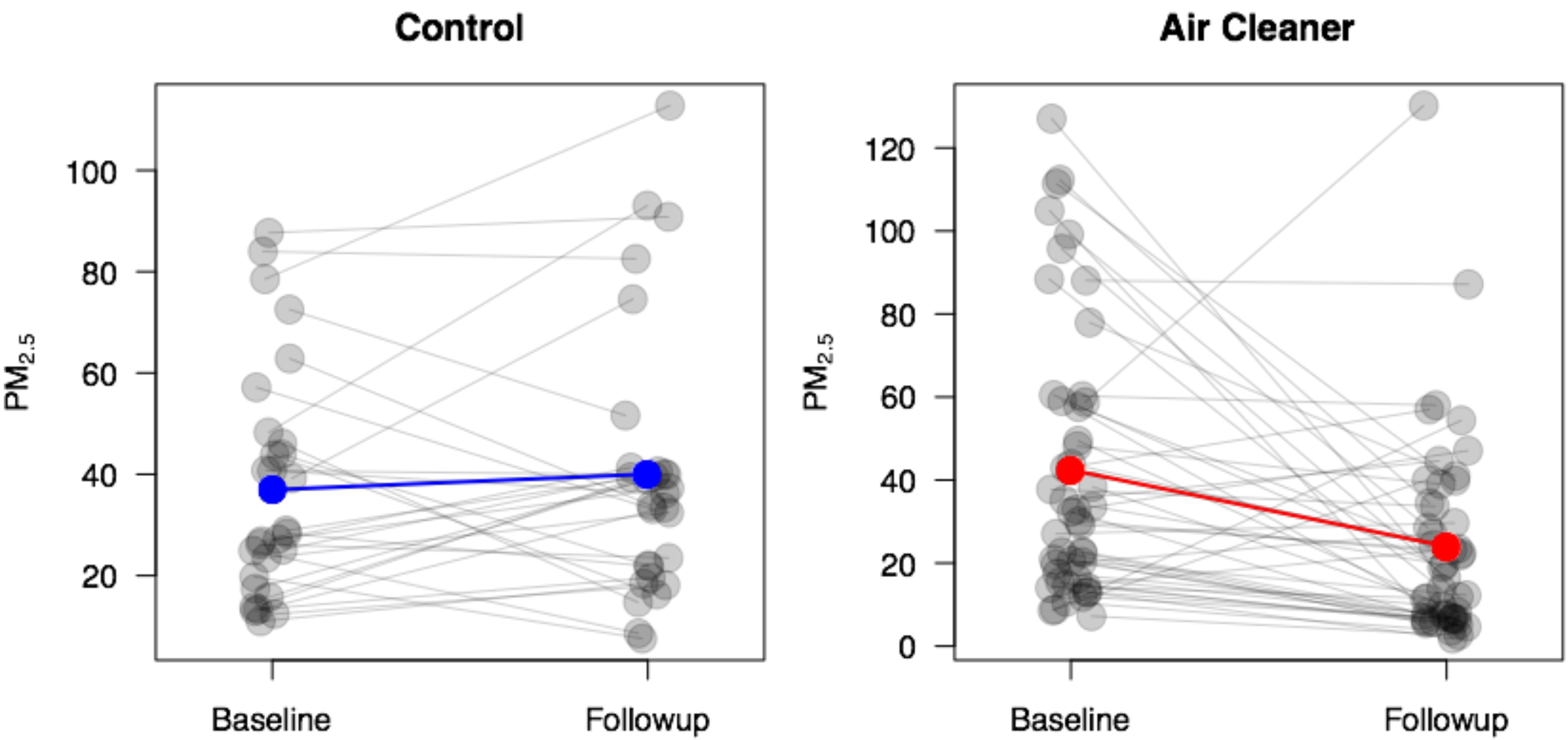
- Randomized intervention in homes in East Baltimore to lower indoor PM levels
- Two groups: **Control**, **Air Cleaner**
- Baseline and 6-month clinic and home visit
- 126 children 6-12 yrs old with asthma enrolled
- Homes had to have a smoker (> 5 cigs/day) living there at least 4 days/week
- **Goal**: Decrease PM2.5 and increase symptom-free days

PREACH Results (outcome)



Improvement of 1.9 symptom-free days

PREACH Results ($PM_{2.5}$)



Bayesian Mixture Models

$$f(y \mid S=1, D=1) = \frac{\pi_a}{\pi_a + \pi_c}\phi(y \mid \mu_{a1}, \sigma_a) + \frac{\pi_c}{\pi_a + \pi_c}\phi(y \mid \mu_{c1}, \sigma_c)$$

$$f(y \mid S=1, D=0) = \frac{\pi_n}{\pi_n + \pi_d}\phi(y \mid \mu_{n1}, \sigma_n) + \frac{\pi_d}{\pi_n + \pi_d}\phi(y \mid \mu_{d1}, \sigma_d)$$

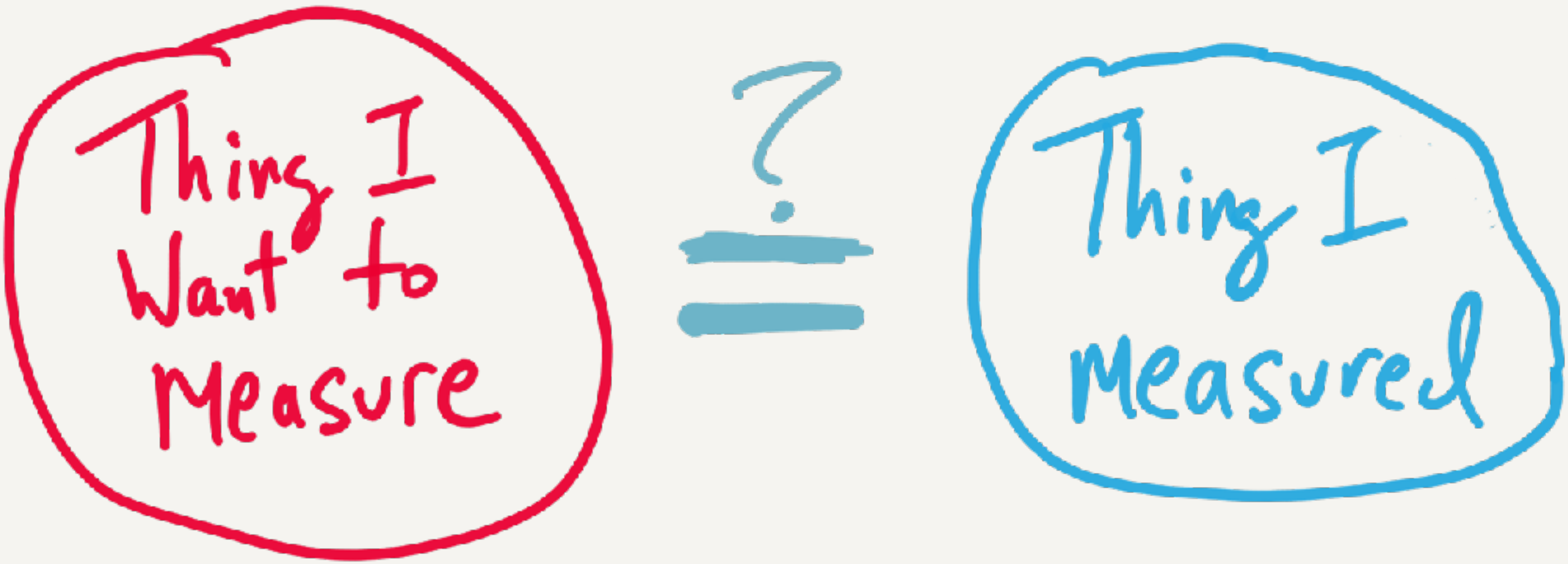
$$f(y \mid S=0, D=1) = \frac{\pi_a}{\pi_a + \pi_d}\phi(y \mid \mu_{a0}, \sigma_a) + \frac{\pi_d}{\pi_a + \pi_d}\phi(y \mid \mu_{d0}, \sigma_d)$$

$$f(y \mid S=0, D=0) = \frac{\pi_n}{\pi_n + \pi_c}\phi(y \mid \mu_{n0}, \sigma_n) + \frac{\pi_c}{\pi_n + \pi_c}\phi(y \mid \mu_{c0}, \sigma_c)$$

Comparison of Model Estimates: Change in Symptom-Free Days

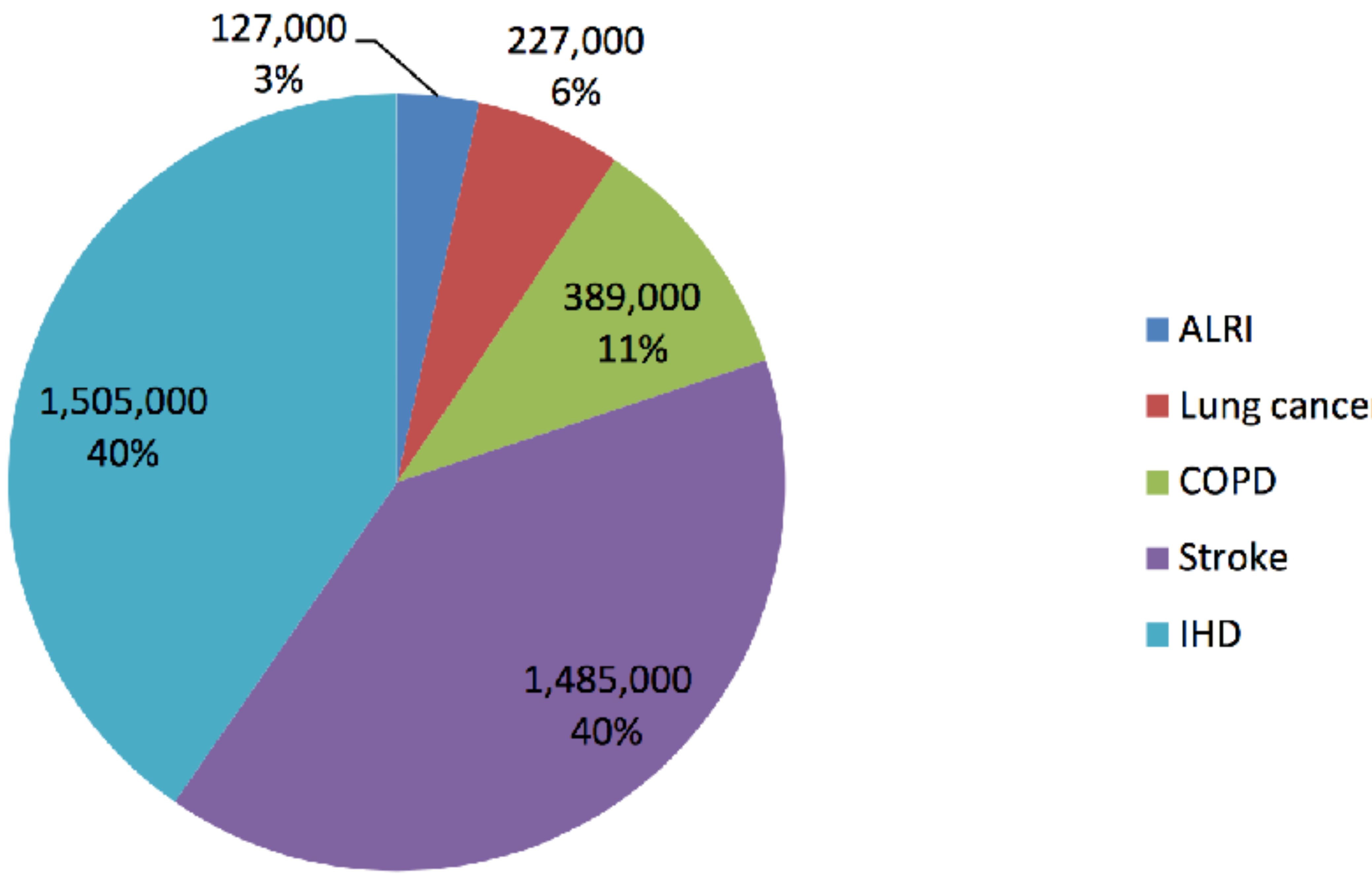
Model	Always-taker	Never-taker	Complier
1			5.2 (-0.1, 11.8)
2	-0.3 (-1.4, 0.9)		5.5 (0.4, 13.3)
3		3.0 (-2.5, 10.2)	4.1 (0.1, 10.8)
Original		1.9 (0.2, 3.6)	

Define the Measurement


Thing I Want to Measure
Thing I Measured


Air Pollution

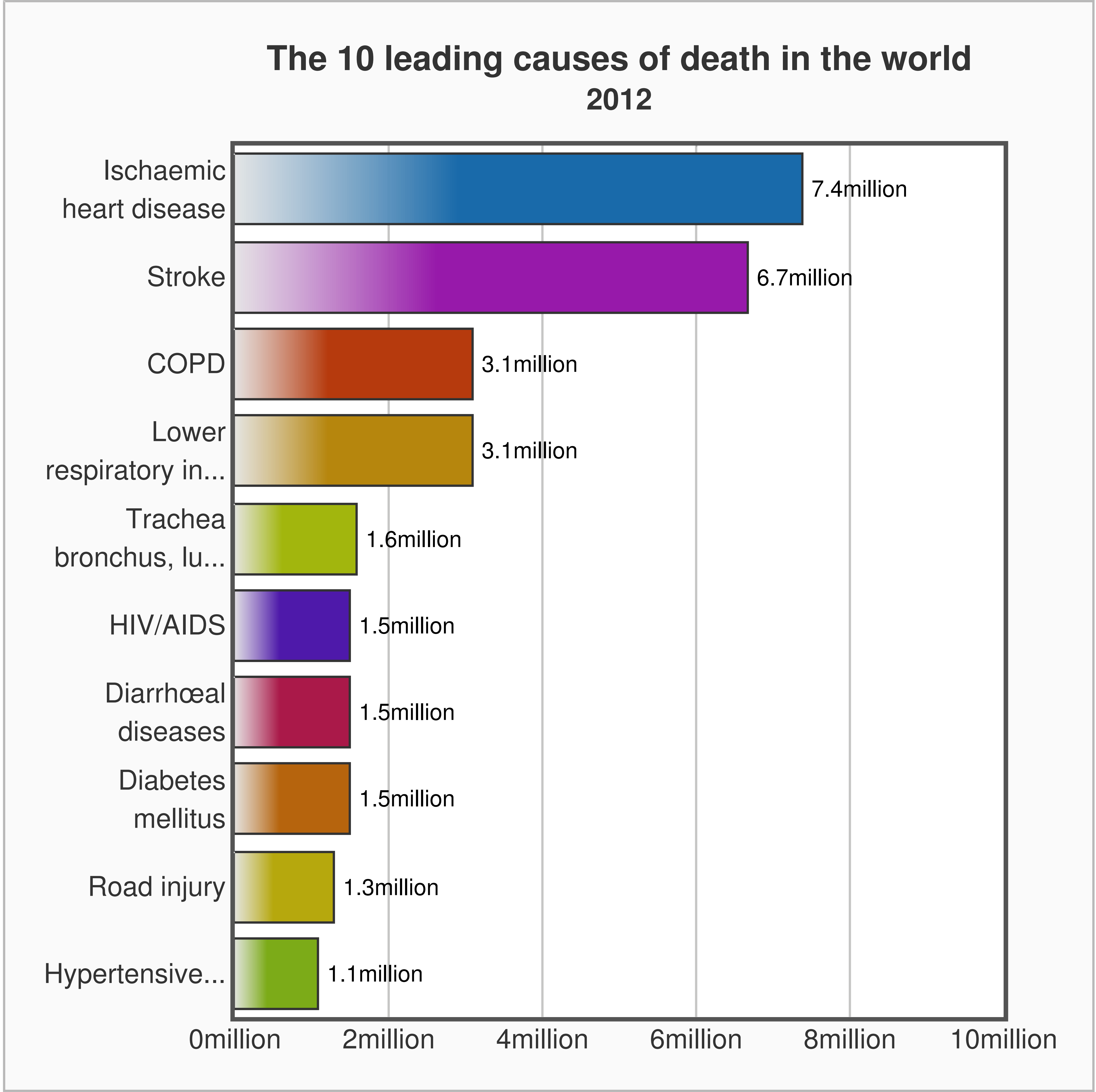
The Landscape



3.7 million deaths due to ambient air pollution (2012)

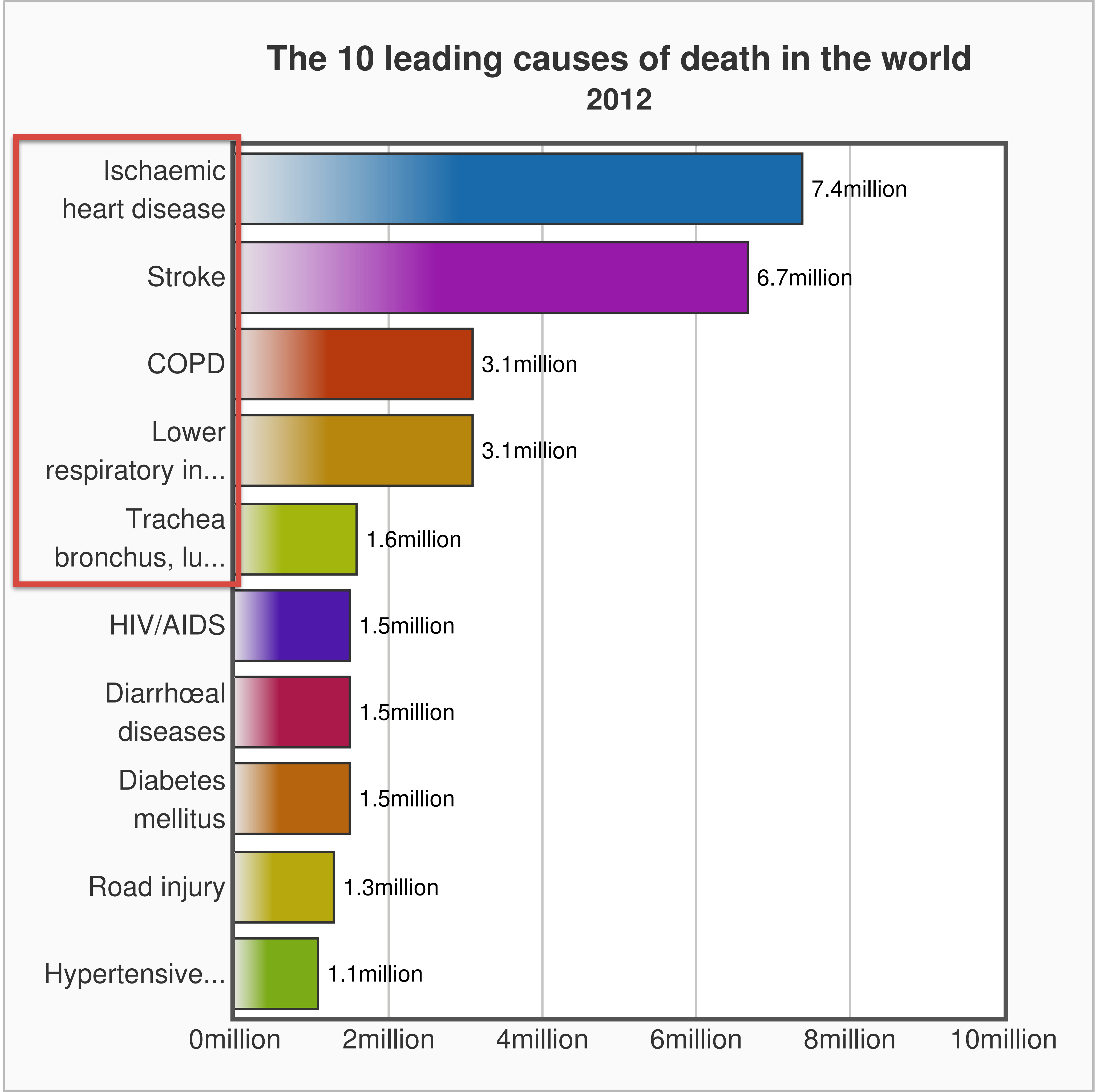
WHO Global Burden of Disease

The Landscape



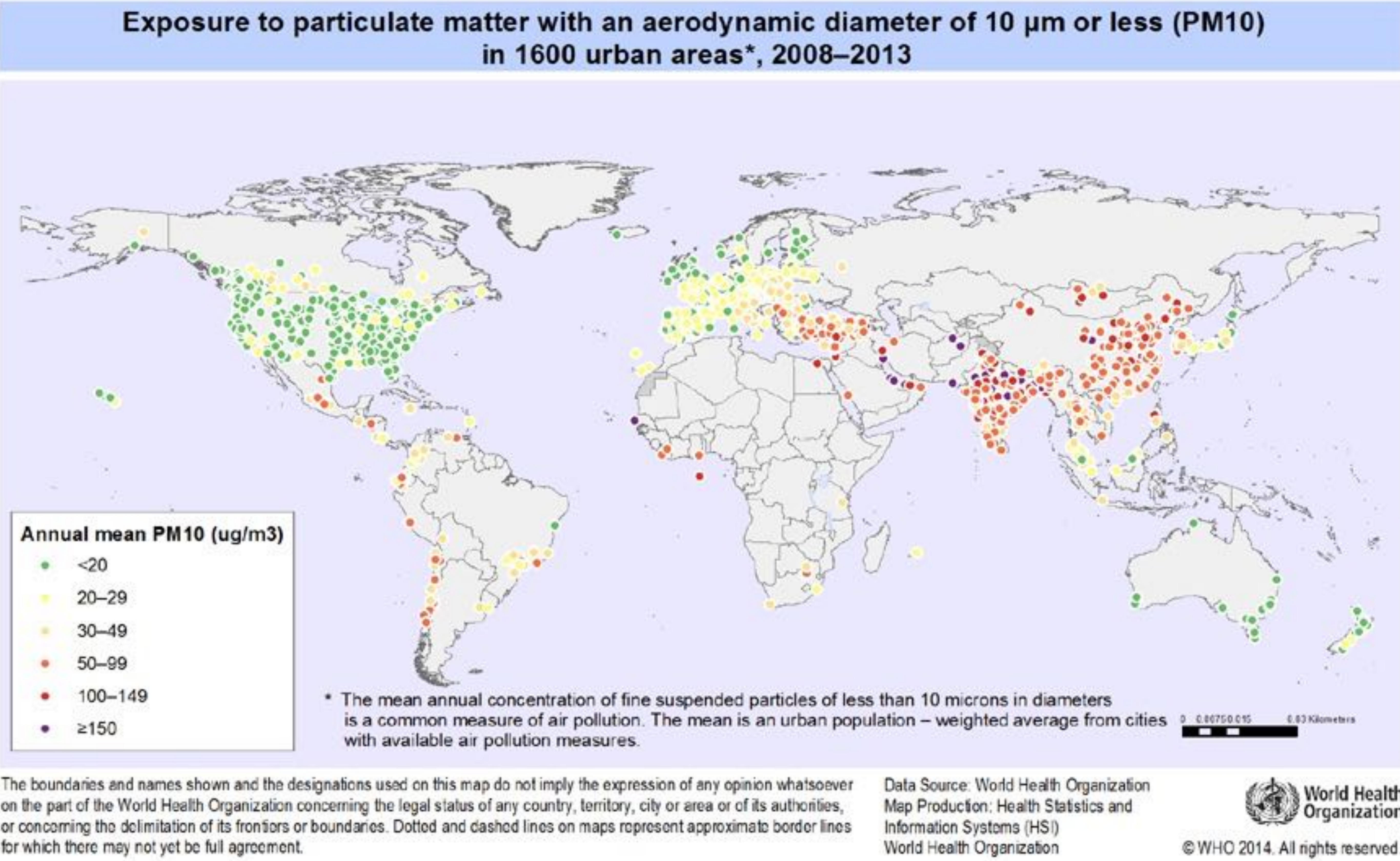
WHO Global Burden of Disease

The Landscape



WHO Global Burden of Disease

The Global Landscape



What is an air pollution study?

Beijing, August 18, 2011

What is an air pollution study?

Beijing, December 5, 2011

Beijing, August 18, 2011

What is an air pollution study?

Beijing

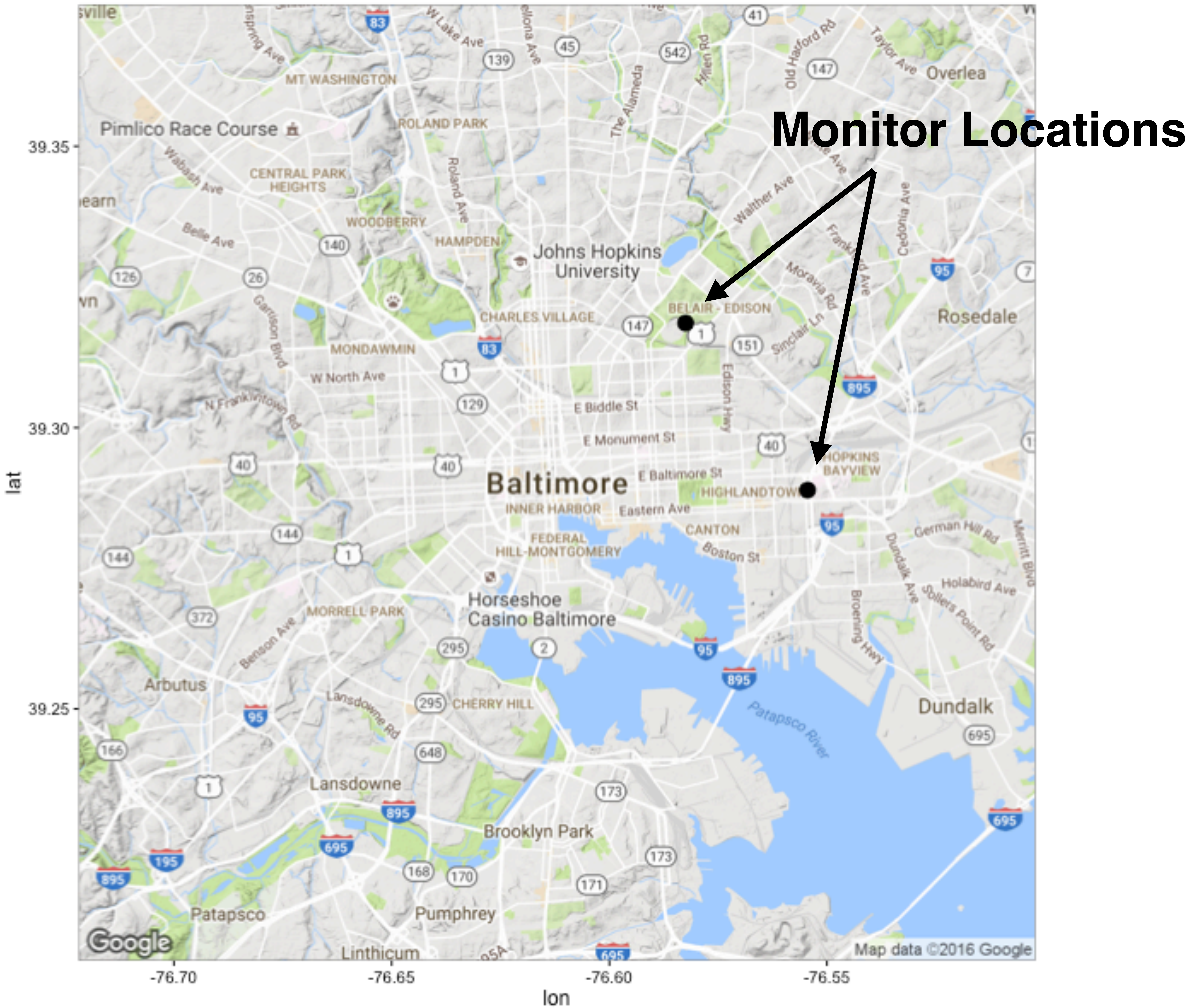
What is an air pollution study?

Shanghai

Beijing

Air Pollution Monitoring

Air Pollution Monitoring



Monitor Locations
lat
lon
39.35
39.30
39.25
-76.70
-76.65
-76.60
-76.55
Pimlico Race Course
MT WASHINGTON
ROLAND PARK
CENTRAL PARK HEIGHTS
WOODBERRY
HAMPDEN
Johns Hopkins University
CHARLES VILLAGE
BELAIR - EDISON
MONDAWMIN
W North Ave
E Biddle St
E Monument St
E Baltimore St
Baltimore
INNER HARBOR
Eastern Ave
HIGHLANDTOWN
HOPKINS BAYVIEW
CANTON
Boston St
FEDERAL HILL-MONTGOMERY
Horseshoe Casino Baltimore
MORRELL PARK
Arbutus
CHERRY HILL
Lansdowne
Brooklyn Park
Patapsco
Pumphrey
Linthicum
Patapsco River
Dundalk
Rosedale
Overlea
German Hill Rd
Holabird Ave
Sollers Point Rd
Broening Hwy
Dundalk Ave
Merritt Blvd
Edison Hwy
Sinclair Ln
Moravia Rd
Frankford Ave
Walther Ave
Wabash Ave
Belle Ave
Garrison Blvd
N Frankintown Rd
Benson Ave
Lansdowne Rd
Roland Ave
The Alameda
Hillen Rd
Old Harford Rd
Taylor Ave
W Lake Ave
Google
Map data ©2016 Google

Spatial—Temporal Model

$$w(s,t) = \mu(s,t) + z(s,t)$$

Pollutant level at location *s* and time *t*

Fixed effects

Gaussian process with correlation function $\rho(\cdot,\cdot \mid \phi, \kappa)$

County-wide block average pollutant level

$$x_t = \frac{1}{\|A\|} \int_A w(s,t)\, ds$$

Predictive Distribution for x_t

Joint distribution of monitor values and block average is Normal

$$\begin{pmatrix} w(\mathbf{v}_1, t) \\ \vdots \\ w(\mathbf{v}_m, t) \\ x_t \end{pmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \boldsymbol{\mu}_t \\ \mu_{x,t} \end{bmatrix}, \sigma^2 \begin{bmatrix} H_{11} & H_{12} \\ H_{21} & H_{22} \end{bmatrix}\right)$$

$$\begin{aligned}
[H_{11}]_{ij} &= \rho(\|\mathbf{v}_i - \mathbf{v}_j\|; \phi, \kappa) \\
[H_{12}]_i &= \frac{1}{\|A\|} \int \rho(\|\mathbf{v}_i - \mathbf{s}\|; \phi, \kappa)\, d\mathbf{s} \\
H_{21} &= H_{12}' \\
H_{22} &= \frac{1}{\|A\|^2} \iint \rho(\|\mathbf{s} - \mathbf{s}'\|; \phi, \kappa)\, d\mathbf{s}\, d\mathbf{s}'
\end{aligned}$$

Bayesian and Plug-in Approaches

- Predictive distribution of block average given monitor values

$$x_t|\mathbf{w}_t \sim \mathcal{N}(\mu_{x,t} + H_{12}' H_{11}^{-1}(\mathbf{w}_t - \boldsymbol{\mu}_t), \sigma^2(H_{22} - H_{12}' H_{11}^{-1} H_{12}))$$

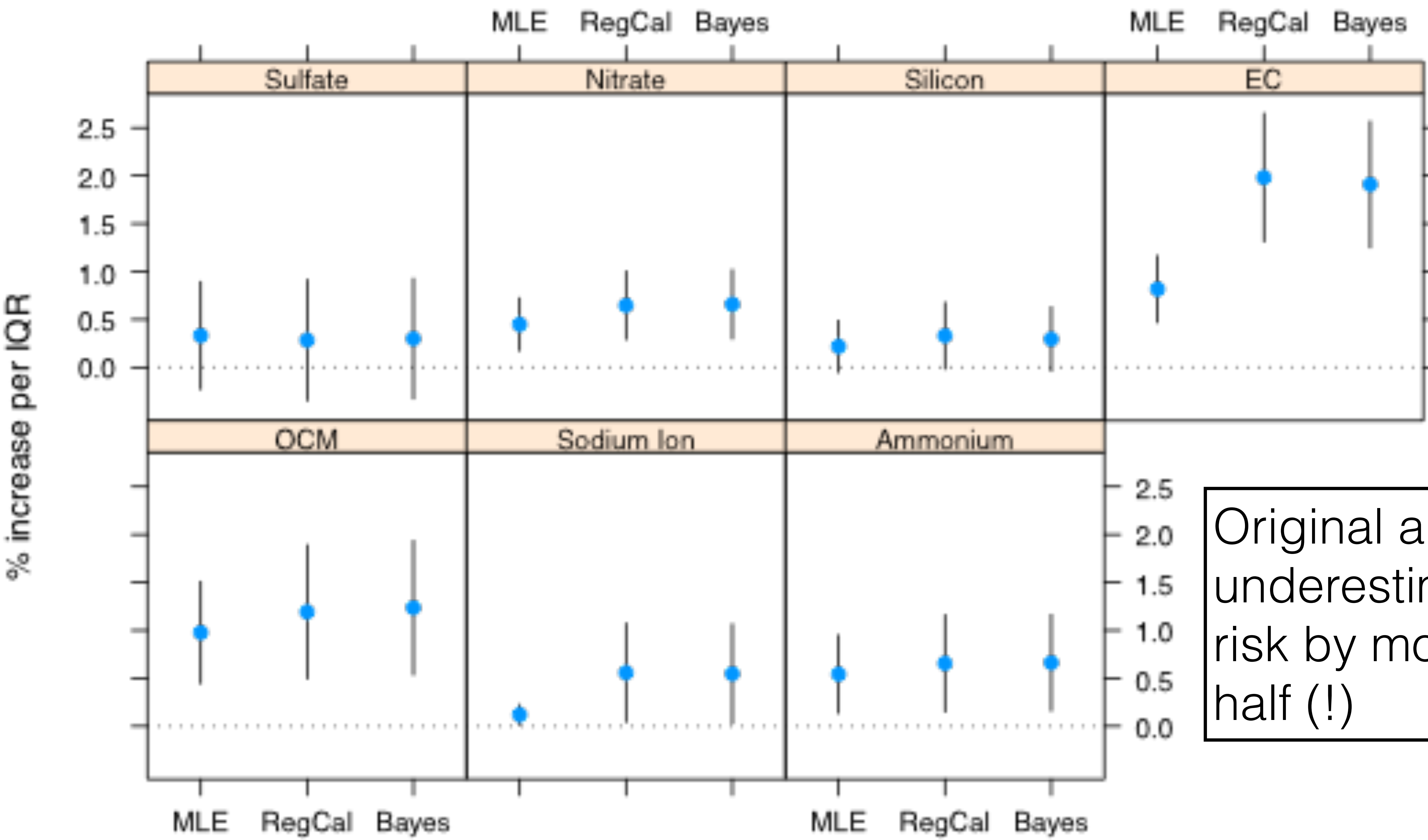
- *Bayesian Model*: We can use MCMC to sample from

$$p(\theta, \mathbf{x}|\mathbf{y}, \mathbf{w}) \propto p(\mathbf{y}|\theta, \mathbf{x}, \mathbf{w}) r(\mathbf{x}|\mathbf{w}) \pi(\theta)$$

Poisson likelihood
(time series model)

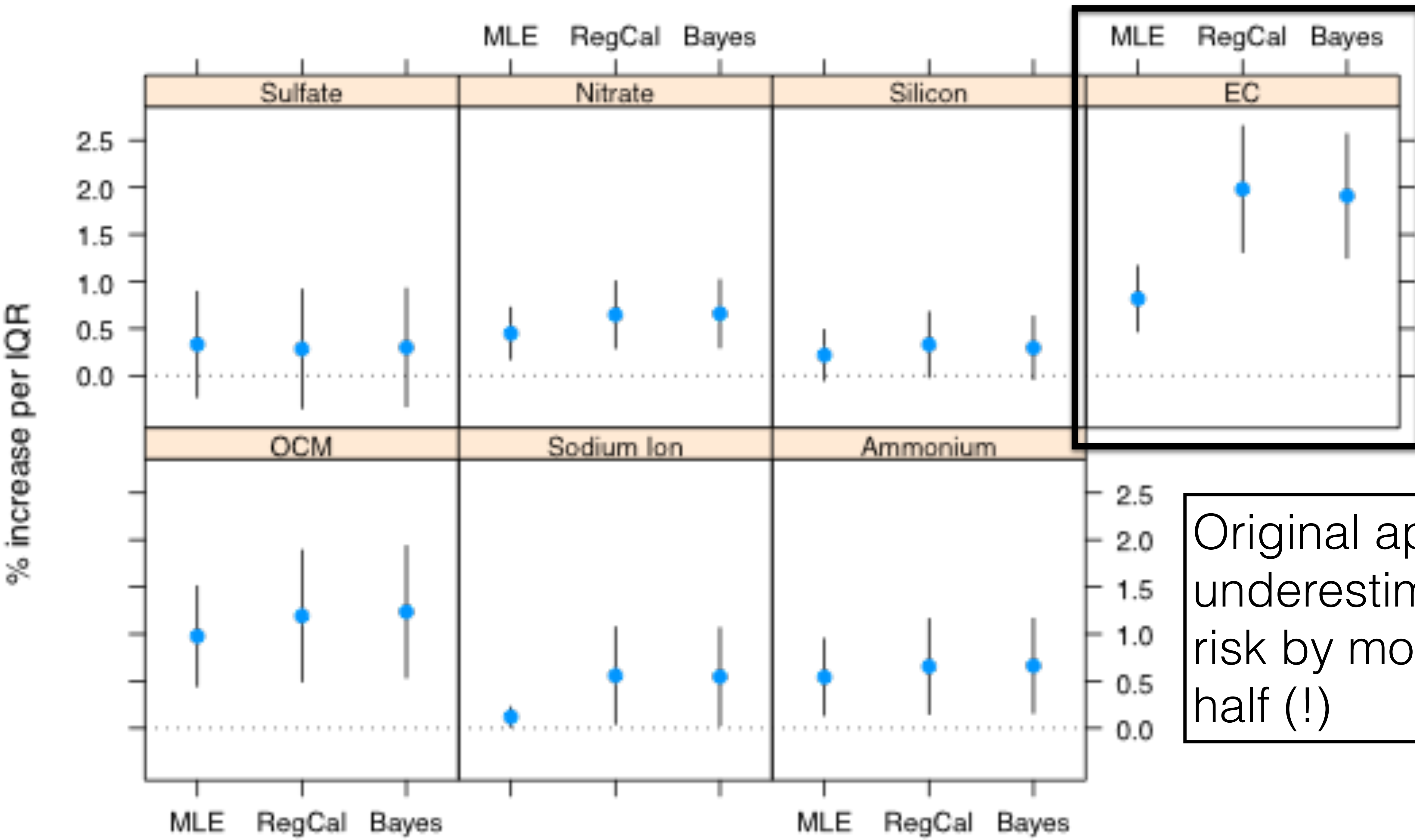
Predictive model

Combined Estimates Across 20 Counties



Original approach underestimated risk by more than half (!)

Combined Estimates Across 20 Counties

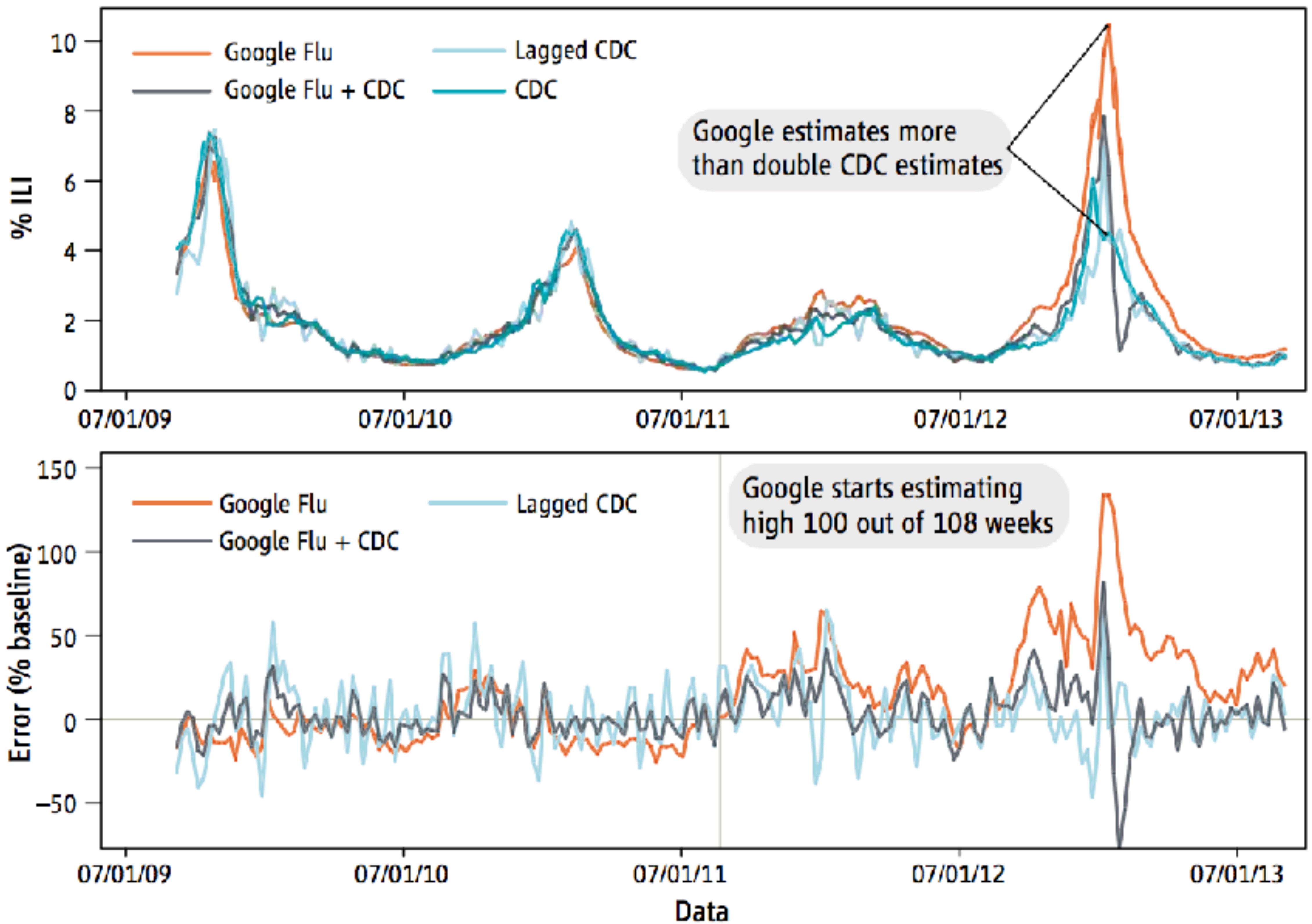


Original approach underestimated risk by more than half (!)

What could
go wrong?

¯_(ツ)_/¯

Parable of Google Flu Trends



Manage the Process

Parable of Personalized Medicine

ARTICLES

nature medicine

Genomic signatures to guide the use of chemotherapeutics

Anil Potti[1,2], Holly K Dressman[1,3], Andrea Bild[1,3], Richard F Riedel[1,2], Gina Chan[4], Robyn Sayer[4], Janiel Cragun[4], Hope Cottrill[4], Michael J Kelley[2], Rebecca Petersen[5], David Harpole[5], Jeffrey Marks[5], Andrew Berchuck[1,6], Geoffrey S Ginsburg[1,2], Phillip Febbo[1–3], Johnathan Lancaster[4] & Joseph R Nevins[1–3]

Parable of Personalized Medicine

ARTICLES

• Retracted •

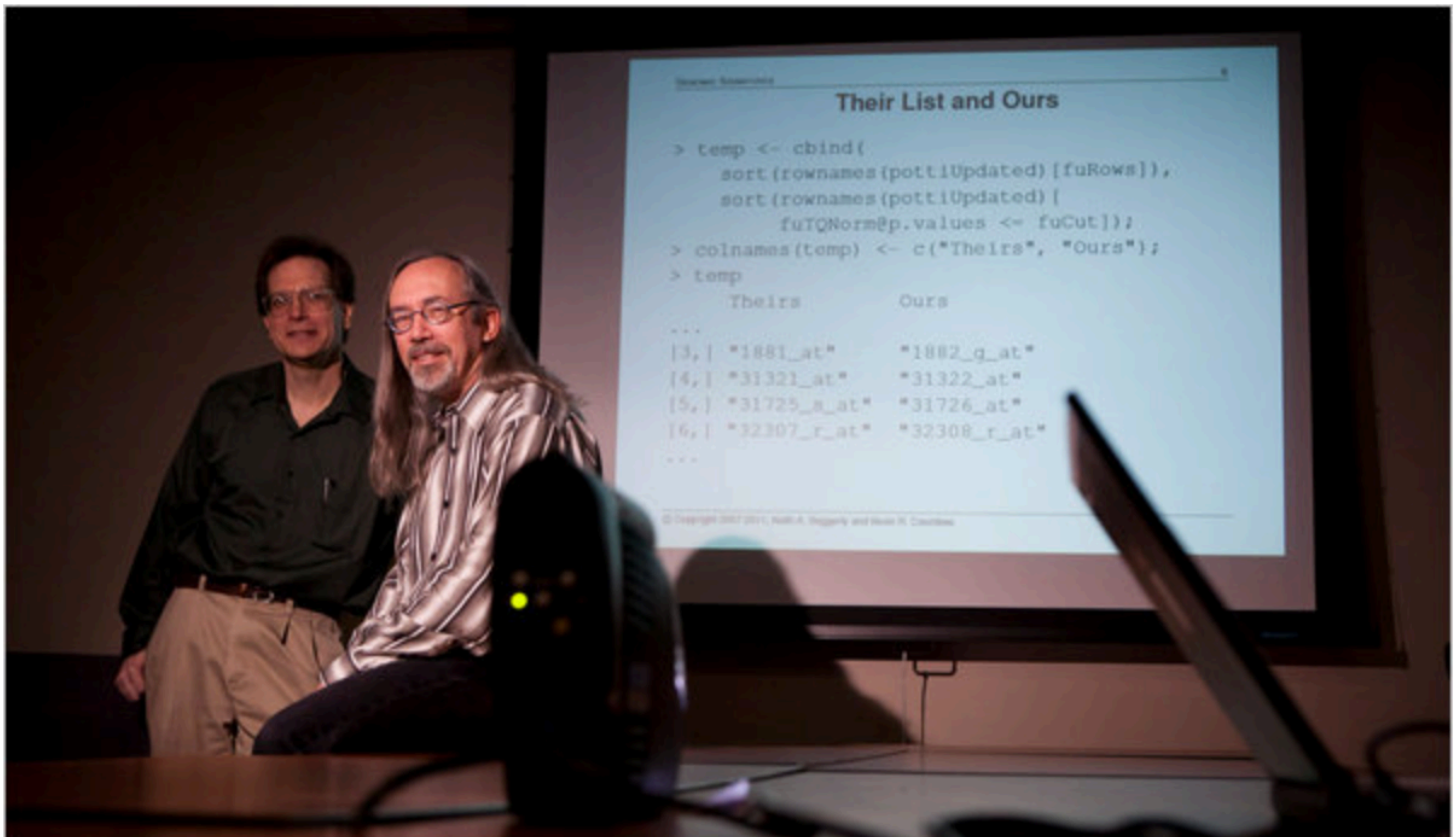
nature medicine

Genomic signatures to guide the use of chemotherapeutics

Anil Potti[1,2], Holly K Dressman[1,3], Andrea Bild[1,3], Richard F Riedel[1,2], Gina Chan[4], Robyn Sayer[4], Janiel Cragun[4], Hope Cottrill[4], Michael J Kelley[2], Rebecca Petersen[5], David Harpole[5], Jeffrey Marks[5], Andrew Berchuck[1,6], Geoffrey S Ginsburg[1,2], Phillip Febbo[1–3], Johnathan Lancaster[4] & Joseph R Nevins[1–3]

“Rock Star” Statisticians

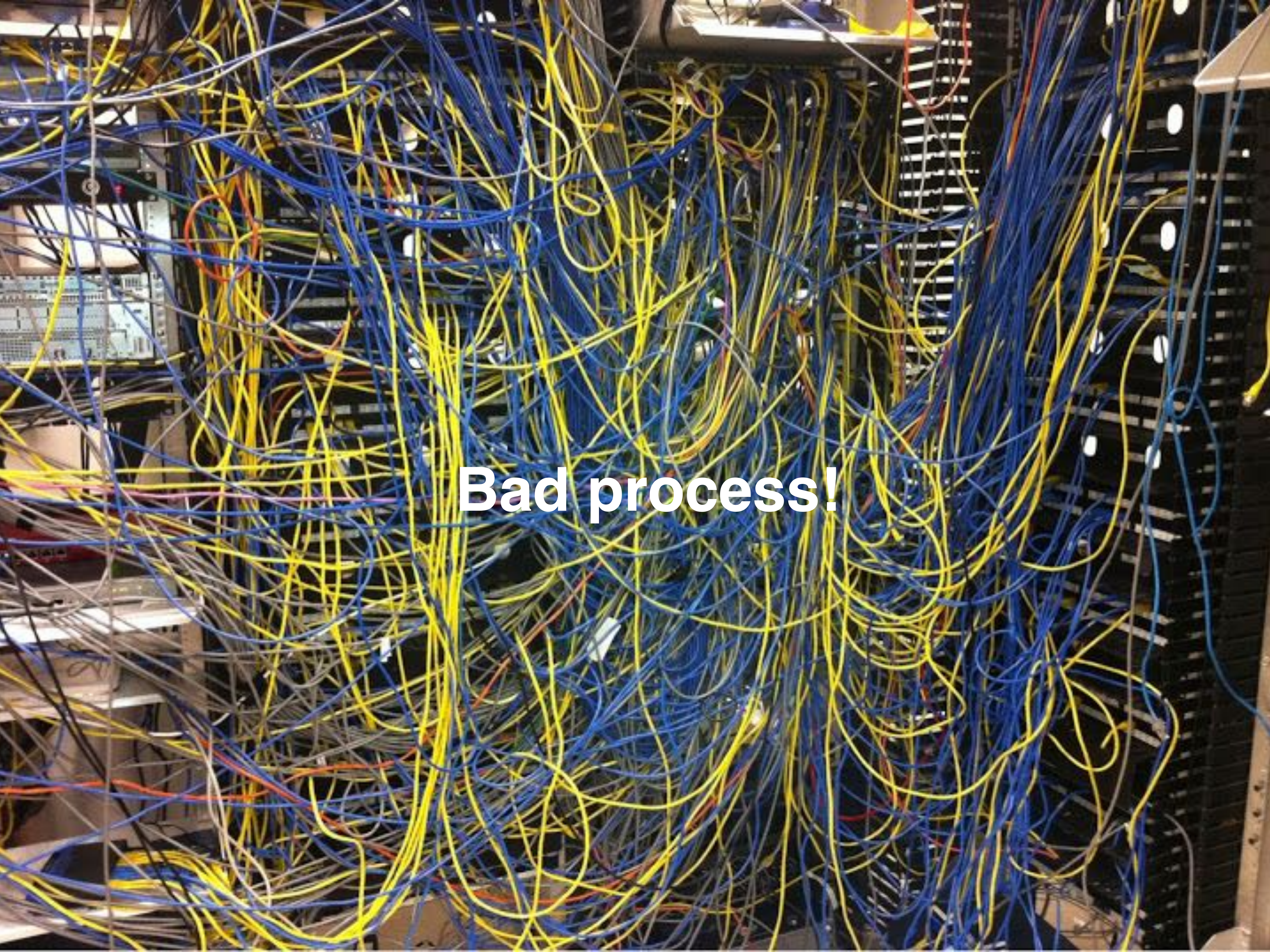
How Bright Promise in Cancer Testing Fell Apart



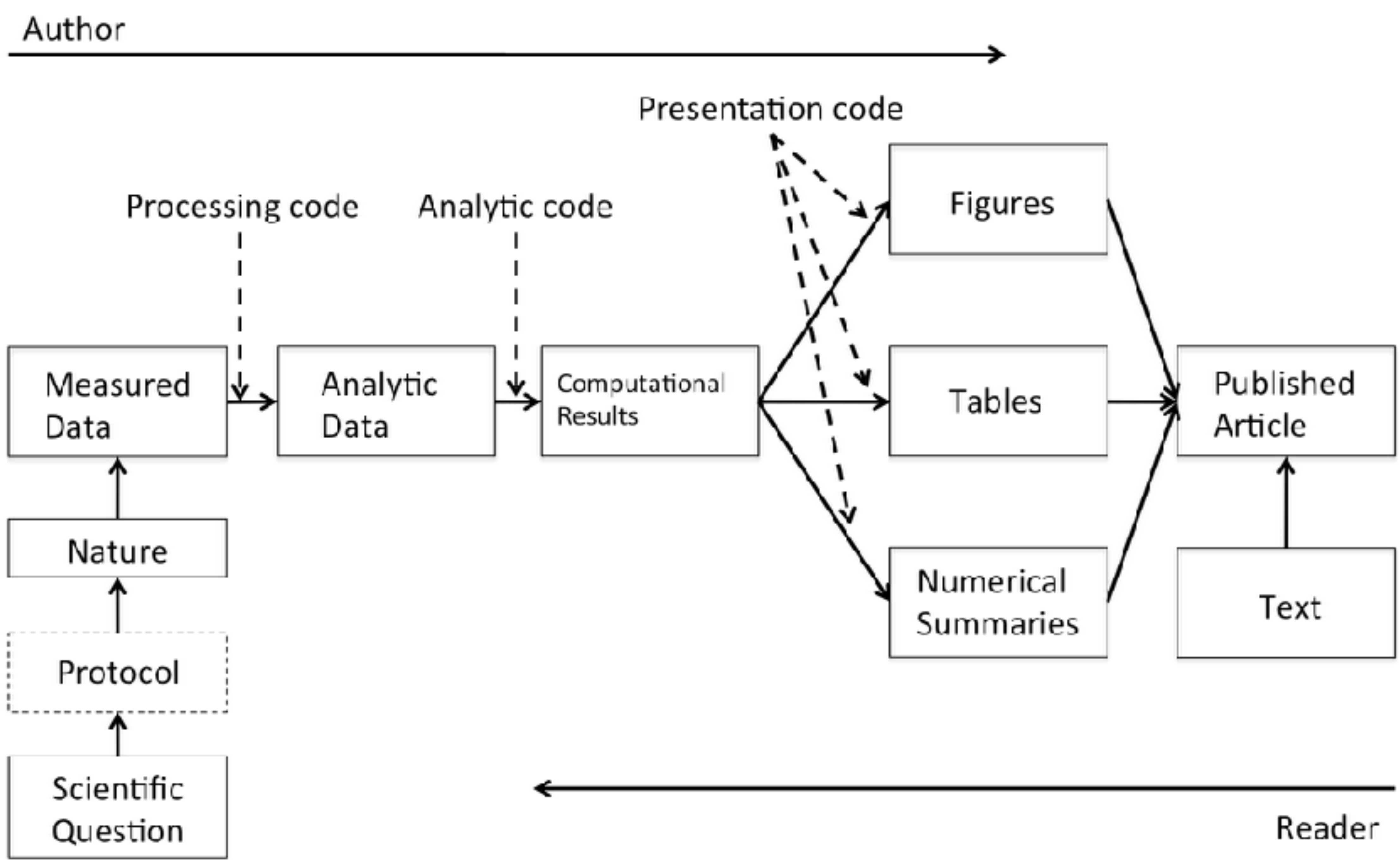
Michael Stravato for The New York Times

New York Times

Deception at Duke


Bad process!

The Data Science Process



Institute of Medicine Report

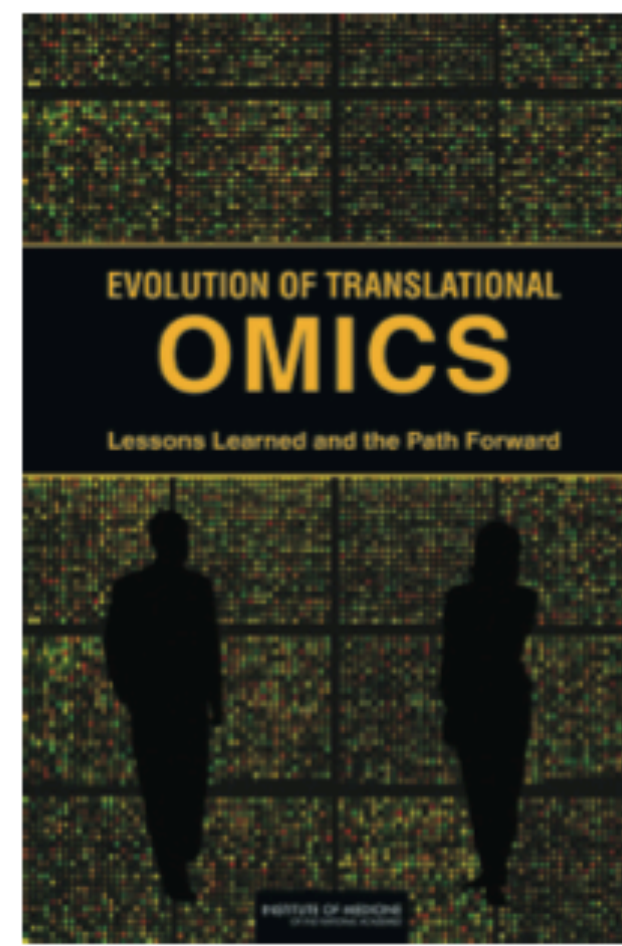
REPORT BRIEF MARCH 2012

INSTITUTE OF MEDICINE
OF THE NATIONAL ACADEMIES

Advising the nation • Improving health

For more information visit www.iom.edu/translationalomics

Evolution of Translational Omics
Lessons Learned and the Path Forward



Institute of Medicine Report

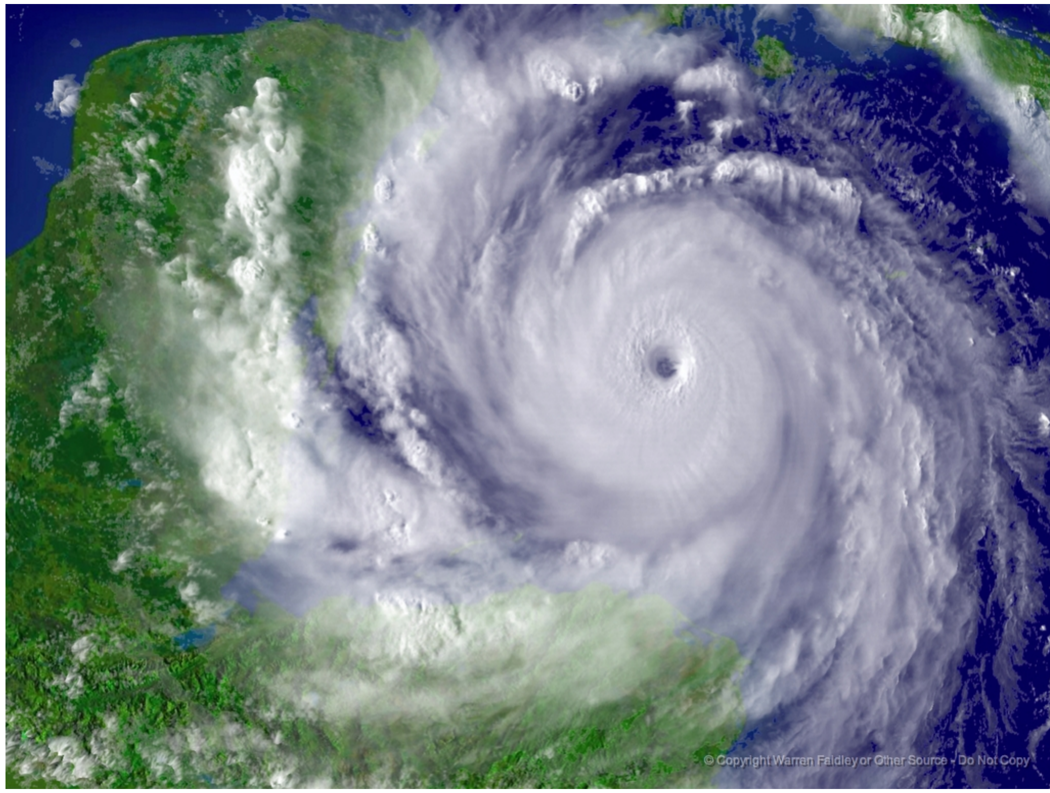
- **Data/metadata** used should be made publicly available

- The **computer code** and fully specified computational procedures used should be made available

- Ideally, the computer code that is released will **encompass all of the steps** of computational analysis, including all data preprocessing steps

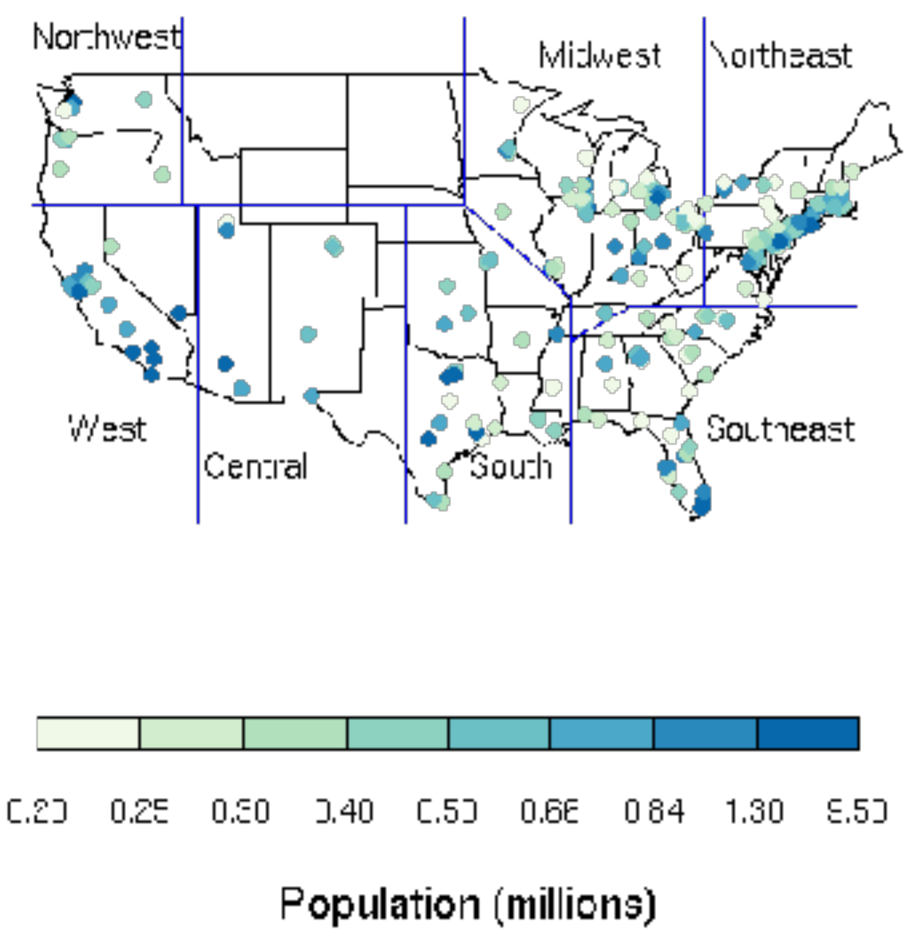
Air Pollution and Health: A Perfect Storm?

- Estimating small health effects in the presence of much stronger signals
- Results inform substantial policy decisions and affect many stakeholders
- EPA regulations can cost billions of dollars
- Complex statistical methods are needed and subjected to intense scrutiny



Medicare Cohort Air Pollution Study

Medicare Air Pollution Study (MCAPS), 1999--2002



Supplementary Materials

FREQUENTLY ASKED QUESTIONS about the MCAPS study

Press release from The Johns Hopkins Bloomberg School of Public Health

Information about the counties used in the study. More information about the counties in MCAPS is also available on our Google map

Background information particulate matter from the EPA:

- Particle Pollution in 2003
- Summary of Counties Violating the $PM_{2.5}$ Primary Standards

Materials for Reproducing Study Results

County-specific estimates

Below are tables containing county-specific estimates of the association between $PM_{2.5}$ and hospital admissions for various outcomes. The estimates here are the raw regression coefficients (`beta`) and variances (`var`) taken from the models described in the paper. **NOTE:** The HTML tables are large and may take a long time to load into your browser.

- Subset of models used for Table 1: HTML | Comma separated value (CSV) [194K]
- All models, used for Figure 2: HTML | Comma separated value (CSV) [1.8M]

Comma separated value (CSV) files are better suited for reading into statistical analysis programs.

Please note that the principal findings of the study are estimates of the **national** and **regional** effects of short-term exposure to $PM_{2.5}$. County-specific estimates are provided solely for the purpose of reproducing those findings.

iHAPSS
Internet-based Health & Air Pollution Surveillance System

NMMAPSdata R Package

Publications | Software | Data | Bweb | iHAPSS

NMMAPSdata R Package

Current version: 0.3-4

The NMMAPSdata R package contains daily mortality, air pollution, and weather data originally assembled as part of the National Mortality, Morbidity, and Air Pollution Study (NMMAPS).

There is a technical report available which contains a brief overview of the package and contains examples of multi-city time series analysis of air pollution and mortality.

- The files simple.R, seasonal.R, and tdlm.R referenced in the report contain example code and functions for reproducing NMMAPS analyses.

Database summary information

- Time frame: January 1, 1987 -- December 31, 2000
- Causes of death: Total non-accidental, CVD, respiratory, pneumonia, COPD, accidental
  - Age categories: < 65, 65--74, >= 75
- Pollutants: PM_{10}, $PM_{2.5}$, CO, O_3, SO_2, NO_2
- Weather: Temperature, dewpoint temperature, relative humidity
- Number of Cities: 108

More detailed information about the database can be found on the IHAPSS website at http://www.ihapss.jhsph.edu/.

Package requirements

- R version 1.9.0 or higher.
- bzip2 compression capability. Most people will *not* have to worry about this since R comes with bzip2 compression capability by default. However, on some Unix-like systems it is possible that the version of R was compiled without it. NMMAPSdata will give an error when the package is loaded if bzip2 capability is not present.
- Approximately 380MB of disk space to store the package.

For **Unix**, **Linux**, and **Mac OS X** users, there is a source package available.

The success of a data analysis depends on the process, not the result.

The Future

The Future is Bright!

- A tremendous infrastructure on which to build
- Cheap hardware and Moore’s Law has made powerful computing available to all with the cloud
- Advanced software has abstracted complex details of data analysis
- The Internet allows analyses to be deployed to the entire world
- Volume of data is increasing dramatically
- A never-ending supply of difficult (but interesting) questions!

The Future is Bright!

- The future will favor those trained in data science
- Problems and data are coming in too fast
- A perfect training for interdisciplinary work
- Many leadership opportunities

Johns Hopkins Biostatistics

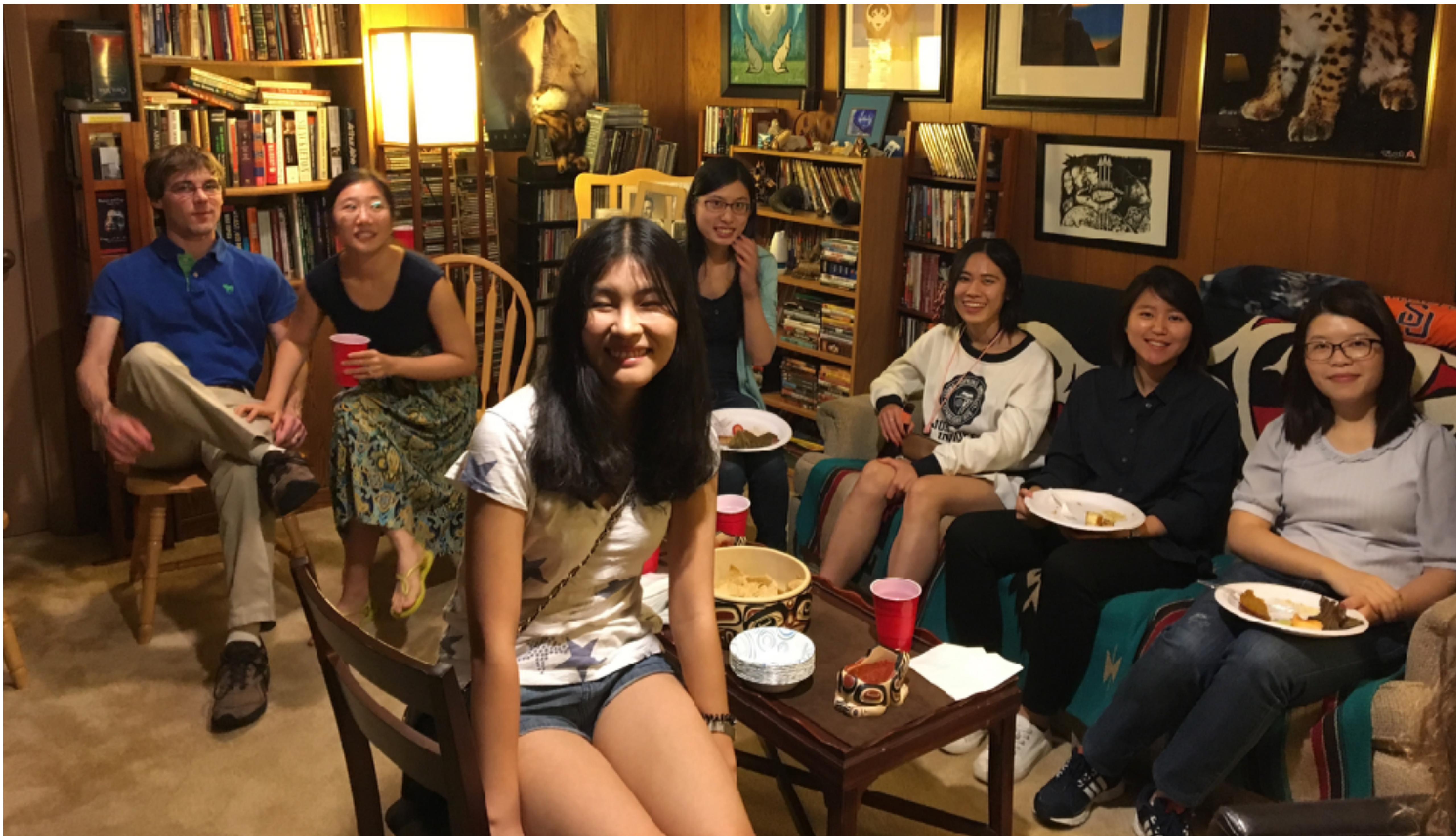
- Largest / oldest / best school of public health
- PhD program in Biostatistics (~10 per class)
- ScM program in Biostatistics (8-12 per class)
- Rigorous training with focus on science
- **Applications**: Environmental health, genomics, personalized medicine, medical imaging, wearable computing, clinical trials, infectious disease

The Johns Hopkins Data Science Lab

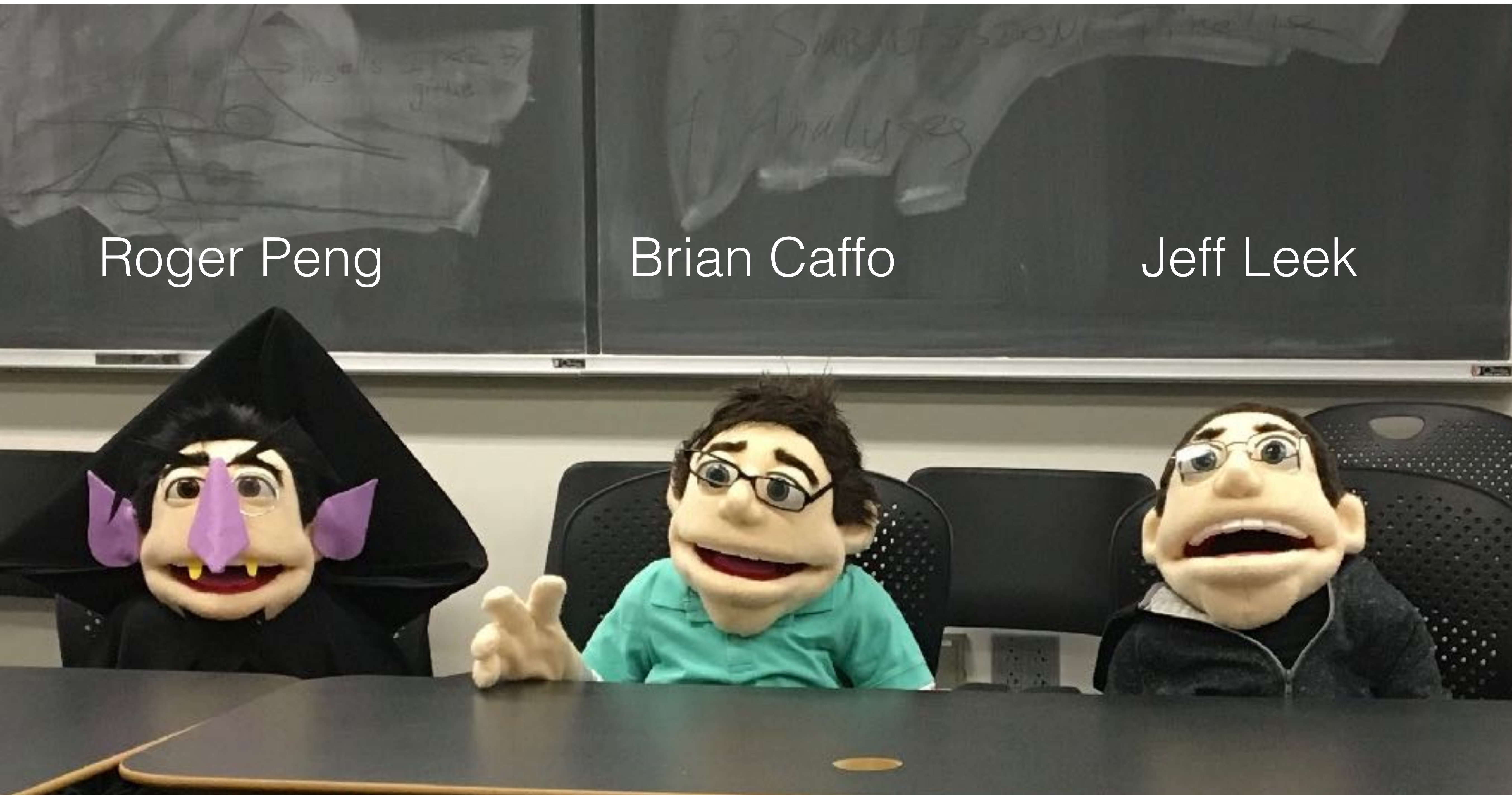
The Data Science Lab at Johns Hopkins is about all things data science. We produce courses, develop software, prototype apps, conduct research, and generally spread the word about data science. We believe that the intelligent application of data science skills can have a profound impact across all areas of our lives.

MEET THE TEAM SEE EXAMPLES

Great Students



Great Faculty



Thank You!