The Role of Statistics in the Big Data Era

Feifang Hu

Department of Statistics

George Washington University

and

Institute of Statistics and Big Data

Renmin University of China

Email: feifang@gwu.edu

Dec 19, 2016, Shenzhen, China

Feifang Hu

Education

- PhD in Statistics, 1994, University of British Columbia, Vancouver, B. C., Canada. Advisor: Professor James V. Zidek.
- MSc in Statistics, 1988, Zhejiang University, Hangzhou, Zhejiang, P. R. China. Advisor: Professor Yaoting Zhang.
- BSc in Mathematics, 1985, Hangzhou Normal University, Hangzhou, Zhejiang, P. R. China

Working Experience

- Professor, 2013-present: Department of Statistics, George Washington University, USA
- Special Professor, 2012-present, Institute of Statistics and Big Data, Renmin University of China, PRC.
- Professor/Associate Professor/Assistant Professor, 2001-2013: Department of Statistics, University of Virginia, USA.
- Adjunct Professor, 2001-2013: Division of Biostatistics and Epidemiology, Department of Health Evaluation Science, University of Virginia School of Medicine. USA.

- Visiting Assistant Professor, 2001: Department of Biometrics, Cornell University, USA.
- Associate Professor/Assistant Professor, 1995-2003: Dept of Statistics & Applied Probability, National University of Singapore, Singapore.
- Post-Doctoral Fellow, 1994-1995: Department of Statistics and Actuarial Science, University of Waterloo, Canada.

UNCERTAINTY (RANDOM)

1 What is Statistics?

Statistics is a "GAME" between human and nature "THROUGH DATA".

2 Statisticians

Statisticians are experts in:

- **producing useful data;**
  Survey Sampling; Experiment Designs.

- **analyzing data to make meanful results;**
  Statistical inference and methods.

- **drawing practical conclusions.**

3 Some examples

Example 1. Measurement Error: We would like to measure the weight of a subject A by using a scale. We know that there is a error of scale. Suppose that the error follows a normal distribution with mean 0 and variance σ^2. Mathematically, we may write:

$$w_1 = A + e_1,$$

where w_A is the true weight, Y_A is the observed weight and e_1 is the measurement error.

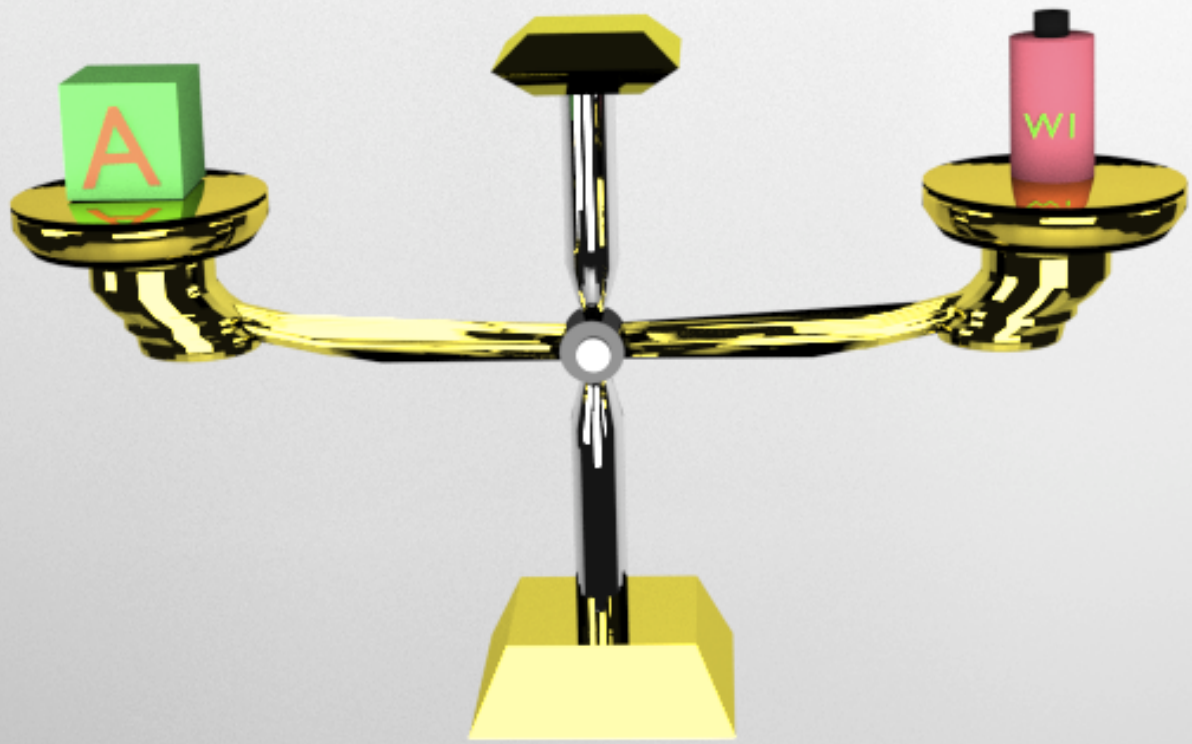


Figure 1: A scale to measure subject A

Now we would like to measure the weights of two subjects A and B by using the same scale twice. What should we do?

Method 1:

$w_1 = A + e_1$ and $w_2 = B + e_2$.

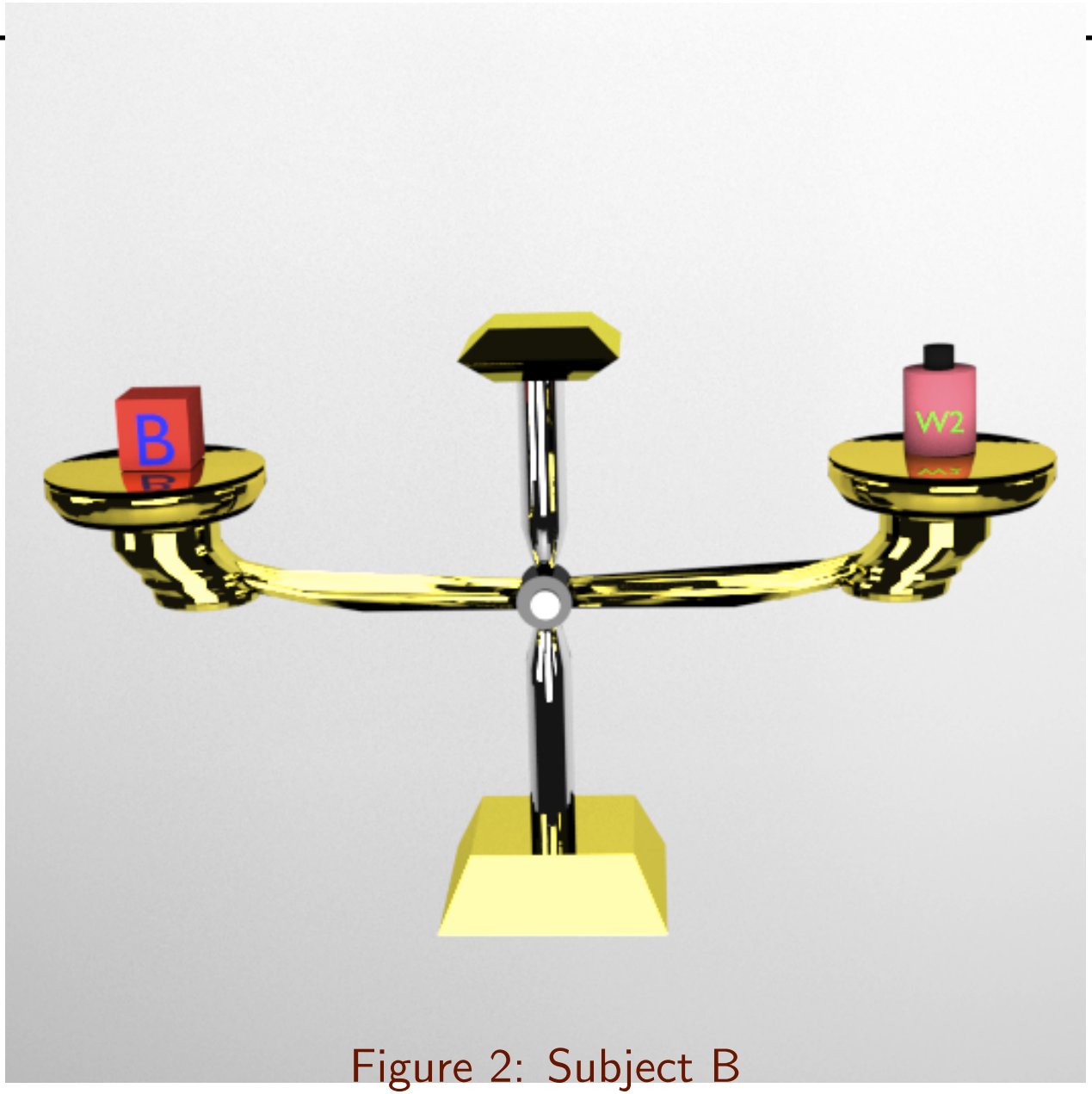


Figure 2: Subject B

Method 2:

$w_3 = A + B + e_3$ and $w_4 = A - B + e_4$.

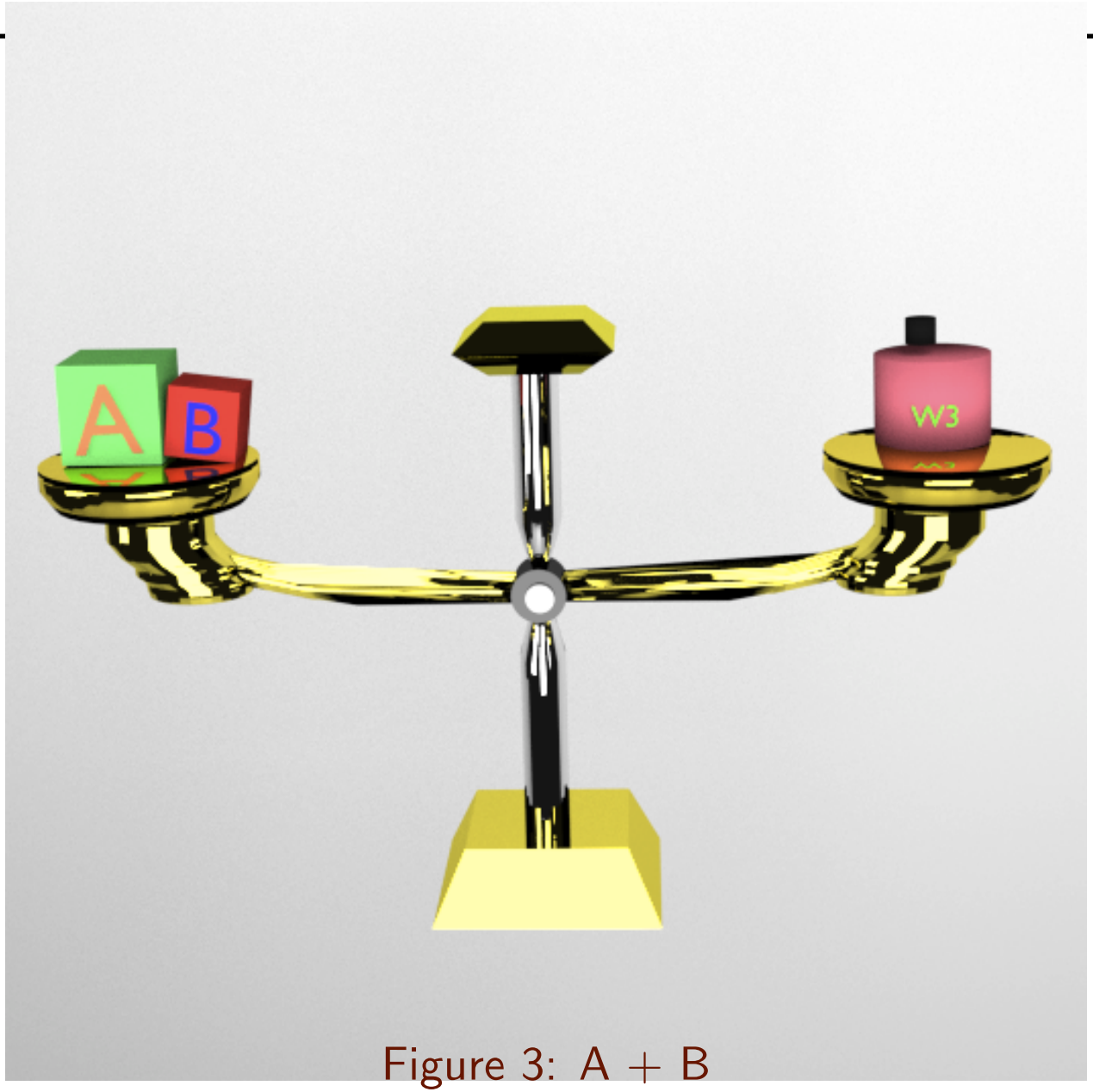


Figure 3: A + B

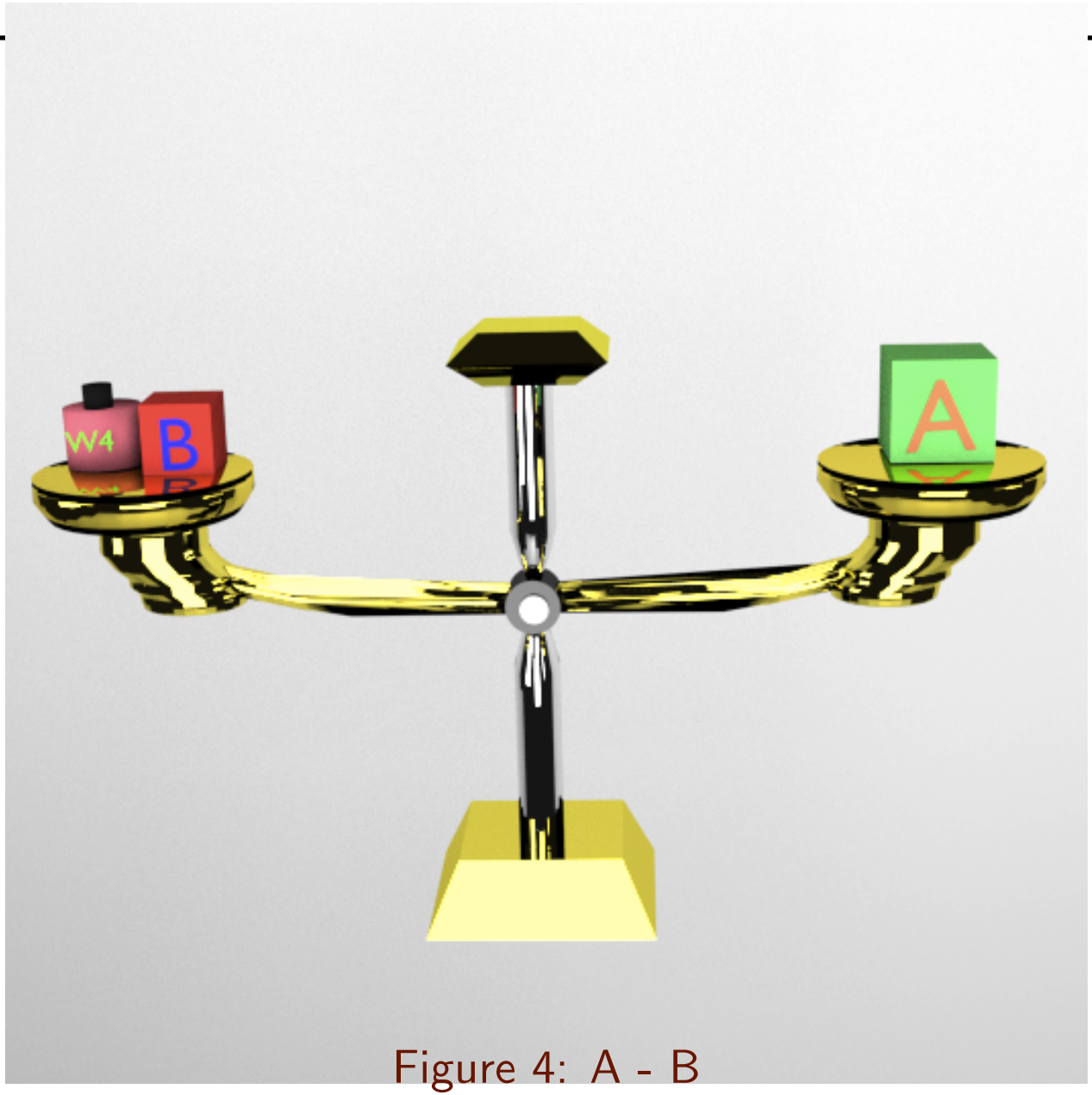


Figure 4: A - B

The measurement errors:

Method 1:
Subject A: $e_1 \sim N(0, \sigma^2)$.
Subject B: $e_2 \sim N(0, \sigma^2)$.

Method 2:
Subject A: $(e_3 + e_4)/2 \sim N(0, \sigma^2/2)$.
Subject B: $(e_3 - e_4)/2 \sim N(0, \sigma^2/2)$.

Good Design of Experiment (producing useful data):

Realistic and Efficient.

Example 2. HIV transmission. Connor et al. (1994, The New England Journal of Medicine) report a clinical trial to evaluate the drug AZT in reducing the risk of maternal-infant HIV transmission.

50-50 randomization scheme is used:

- AZT Group—239 pregnant women (**20 HIV positive infants**).
- placebo group—238 pregnant women (**60 HIV positive infants**).

Given the seriousness of the outcome of this study, it is reasonable to argue that 50-50 allocation was **unethical**. As accruing information favoring (albeit, not conclusively) the AZT treatment became available, allocation probabilities should have been **shifted from 50-50 allocation proportional to weight of evidence for AZT**. Designs which attempt to do this are called *Response-Adaptive designs (Response-Adaptive Randomization).*

If the treatment assignments had been done with the **DBCD** (Hu and Zhang, 2004, Annals of Statistics) with urn target:

- AZT Group— 360 patients
- placebo group—117 patients

then, only **60 (instead of 80)** infants would be HIV positive.

Allocation rule	AZT	Placebo	Power	HIV+
EA	239	238	0.9996	80
DBCD	360	117	0.989	60
Neyman	186	291	0.9998	89
FPower	416	61	0.90	50

4 Big Data Era

From Wikipedia:

Big data is a term for data sets that are so **large or complex** that traditional data processing applications are inadequate. Challenges include **analysis, capture, data curation, search, sharing, storage, transfer, visualization, querying and information privacy**. The term often refers simply to the use of predictive analytics or certain other advanced methods to extract value from data, and seldom to a particular size of data set. Accuracy in big data may lead to **more confident decision making, and better decisions can result in greater operational efficiency, cost reduction and reduced risk**.

Big Data affects academic and applied research in many domains, including
(i) machine translation,
(ii) speech recognition,
(iii) robotics,
(iv) search engines,
(v) digital economy,
(vi) the biological sciences,
(vii) medical informatics,
(viii) health care,
(ix) social sciences and the humanities.

It heavily influences economics, business and finance. From the business perspective, data science is an integral part of competitive intelligence, a newly emerging field that encompasses a number of activities, such as data mining and data analysis.

Some important statistical issues related to BIG DATA:

- Data structure changed:
  (i) Not i.i.d. any more; Not satisfy "our assumption".
  (ii) Usually observational data;
  (iii) Big and others

- Causality or Correlated?
  (i) Causal Inference and Big Data.
  (ii) How to combine information from many different studies.
- Need New Statistical Framework?

5 Analysis of Big Data: Big Data, Big Noise?

Example 3: Google Flu Trends Prediction.
With big data comes big noise.

Google learned this lesson the hard way with its now kaput Google Flu Trends.

(i) The online tracker, which used Internet search data to predict real-life flu outbreaks, emerged amid fanfare in 2008.
(ii) At first Googles tracker appeared to be pretty good, matching CDC datas late-breaking data somewhat closely.
(iii) But, two notable stumbles led to its ultimate downfall: an underestimate of the 2009 H1N1 swine flu outbreak and
(iv) an alarming overestimate (almost double real numbers) of the 2012-2013 flu seasons cases.
(v) Then it met a quiet death this August, 2015 after repeatedly coughing up bad estimates.

With hubris firmly in check, a team of Harvard researchers (leading by Samuel Kou) have come up with a way to tame the unruly data, combine it with other data sets, and continually calibrate it to track flu outbreaks with less error. Their new model (ARGo, auto-regression model based on google data), published in the Proceedings of the National Academy of Sciences (2015), out-performs Google Flu Trends and other models with at least double the accuracy. If the model holds up in coming flu seasons, it could reinstate some optimism in using big data to monitor disease and herald a wave of more accurate second-generation models.

Big Data has a lot of potential, Samuel Kou, a statistics professor at Harvard University. It is just a question of using the right analytics, he said.

6 From Big Data to Precision Medicine.

From Wikipedia:

Precision medicine (PM) (also called personalized medicine) is a medical model that proposes the customization of healthcare, with medical decisions, practices, and/or products being tailored to the individual patient. It usually emphasizes the **systematic use of information about an individual patient** to select or optimize that patient's preventative and therapeutic care.

From NIH:

Precision medicine is an emerging approach for disease treatment and prevention that takes into account individual variability in genes, environment, and lifestyle for each person.

Precision medicine can broadly be defined as **products and service that leverage the science of genomics and proteomics (directly and indirectly)** and capitalize on the trends toward wellness and consumerism to enable tailored approaches to prevention and care.

Over the past century, medical care has centered on standards of care based on epidemiological studies of large cohorts. However, large cohort studies do not take into account the genetic variability of individuals within a population. Precision medicine (also call **Future medicine**) seeks to provide an objective basis for consideration of such individual differences.

As stated by Margaret Hamburg (2010), commissioner, US Food and Drug administration, MD, USA,
"*However, identifying genes that seem to be linked with a disease is only the beginning of an arduous process.* **New approaches** *to the drug-development paradigm are needed such that new drugs are developed along with the tests that inform their use.* **New designs for clinical trials** *are needed so that genetics or other markers can be used to assist in patient selection.*"

Some steps to develop precision medicine:

- From BIG DATA to identify some possible predictive biomarkers and treatments.
- Identify **important predictive biomarkers** from possible predictive biomarkers and treatments with Phase II clinical studies.
- Well designed clinical studies to confirm the significance of biomarkers and treatments, identify suitable (new) treaments (drugs), then approved by FDA.
- Implement to healthcare.

The BIG data resources: fields of translational research (genomics, proteomics, and metabolomics) studies; gene data base; literatures; all related studies, family historical data, etc.

Based on data (big data), many important biomarkers or subgroups have been (and will be) identified.

Example 4: Balance many important covariates for comparison.

Advantages of covariate balance:

- Improve accuracy and efficiency of inference.
- Remove the bias and increase the power.
- Increases the interpretability of results by making the units more comparable, enhance the credibility.
- More robust against model misspecification.

p: the number of covariates.

n: sample size, i.e., the number of units.

- The phenomenon of covariate imbalance is exacerbated as p and n increase.
- Ubiquitous in the era of big data.

- Example: the probability of one particular covariate being unbalanced is $a = 5\%$. For a study with 10 covariates, the chance of at least one covariate exhibiting imbalance is $1 - (1 - a)^p = 40\%$.

Morgan and Rubin [2012] proposed rerandomization.
Qin, Li and Hu (2016) proposed CAM methods.

$$Ratio = \frac{Time_{CAM}}{Time_{Rerandomization}}.$$

CAM vs Rerandomization

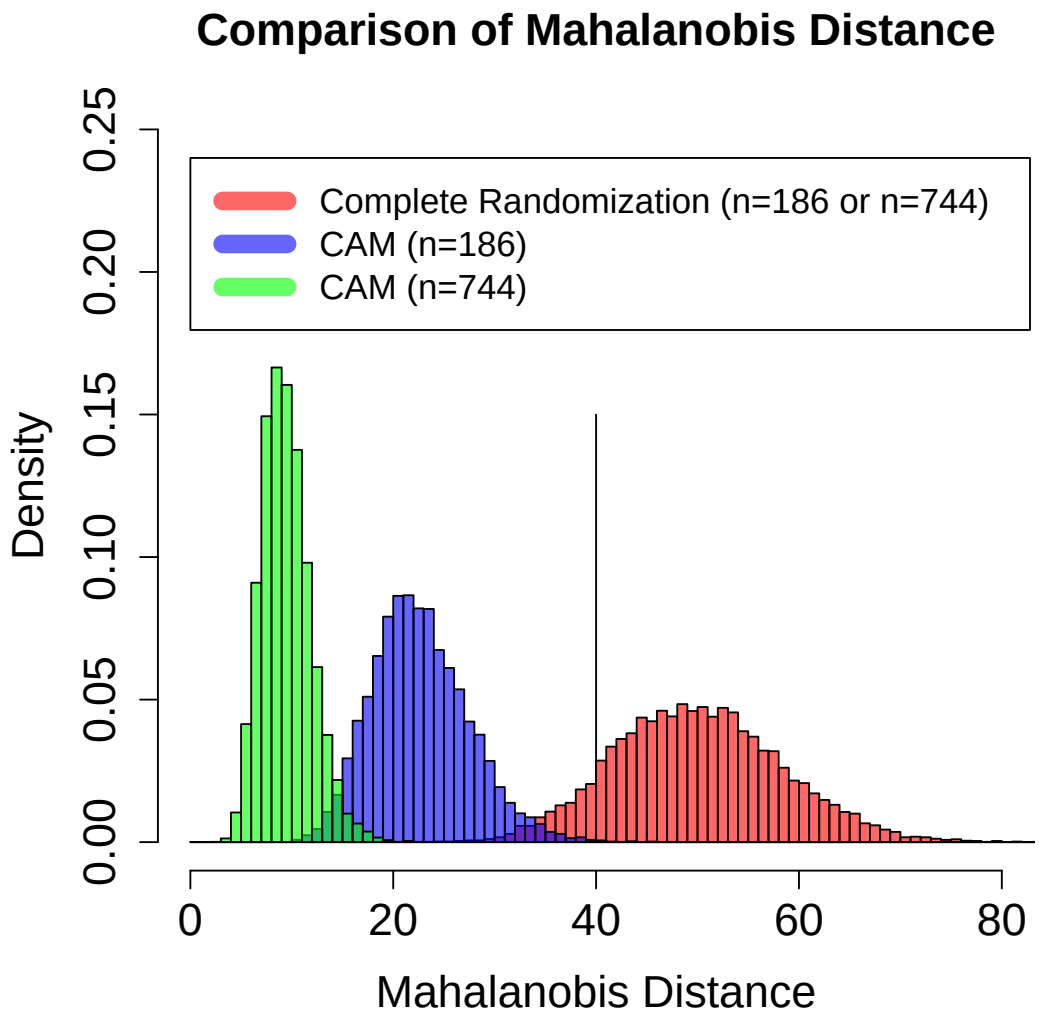
Sample size n	$p = 2$	4	6	8	10
200	0.9830	0.1084	0.0094	7.492e-04	5.686e-05
400	0.4957	0.0275	0.0012	4.884e-05	1.876e-06
600	0.3312	0.0123	0.0003	9.748e-06	2.510e-07

Real Data Example

- A real data set obtained in a clinical study of a Ceragem massage (CGM) thermal therapy bed, a medical device for the treatment of lumbar disc disease.
- Number of covariates $p = 50$.
- 30 numerical covariates: age, measurements of the patient's current conditions, including lower back pain, leg pain, leg numbness, body examination scores, and magnitudes of pain in shoulders, neck, chest, hip and so on. All are measured on 0-10 scales.
- Sample size $n = 186$.

- Replicate the data four times to have a sample size of $n = 744$.
- Original allocation $M = 57.67$, moderate covariate imbalance.
- We repeat the allocation for these patients using CAM, complete randomization and rerandomization.

.CAM vs Rerandomization



Estimation

- After the allocation, we simulate the outcome variable according to

$$y_i^{\text{sim}} = \hat{\mu}_1 T_i^{\text{sim}} + \hat{\mu}_2 (1 - T_i^{\text{sim}}) + x_i^T \hat{\boldsymbol{\beta}} + \epsilon^{\text{sim}},$$

  where $\hat{\mu}_1$, $\hat{\mu}_2$ and $\hat{\boldsymbol{\beta}}$ are obtained from fitting regression to original data. ϵ^{sim} is drawn from the residuals of that regression.

- Compare the estimation performance (PRIV and MSE) of CAM, complete randomization, and rerandomization ($M < 30$, $M < 40$).

- Optimal PRIV is 0.33 (R^2 is 0.33).

Performance Comparison

Sample Size	Method	PRIV	MSE	u_n or v_a
$n = 186$	CAM	19.7%	0.081	0.502
	Rerandomization ($M < 30$)	15.1%	0.085	0.562
	Rerandomization ($M < 40$)	12.2%	0.090	0.730
	Complete Randomization	-	0.100	-
$n = 744$	CAM	27.4%	0.018	0.205
	Rerandomization ($M < 30$)	14.6%	0.021	0.556
	Rerandomization ($M < 40$)	10.9%	0.022	0.718
	Complete Randomization	-	0.025	-

Note: The optimal PRIV is 0.33 (i.e., $R^2 = 0.33$).

7 Big Data and Data Science

"DATA Science has become a **fourth approach to scientific discovery**, in addition to experimentation, modeling, and computation" said Provost Martha Pollack, University of Michigan.

On Sept 8, 2015, University of Michigan announced a US$100 Million "Data Science Initiative" (DSI).

"Data Scientist" has become a popular occupation with Harvard Business Review dubbing it "The Sexiest Job of the 21st Century".

McKinsey Company projecting a global excess demand of 1.5 million new data scientists.

In USA, many Master programs of Data Science had been established in the past four years.

In March, 2015, Renmin University established "The Institute of Statistics and Big Data".

On Oct 8, 2015, Fudan University established "The school of Big Data" and "The institute of Big Data".

Dec 10, 2016, "The Institute of Big Data" was established CSUFL at Wuhan.

Many institutes of Big Data had been and will be established in China.

8 To become a good statistician or Data Scientist.

A statistician or data scientist requires a combination of four types of knowledge.

- Scientific: Is able to quickly learn the basic fact of the subject area.
- Statistical: Needs to have the knowledge and experience to be able to **apply appropriate procedures** for design and analysis.
- **Computational:** Has sufficient proficiency to perform the actual data using software.
- Communication: Has good verbal and written communication skills.

Education:

- BSc in Mathematics, Statistics, Computer Science, Data Science or related fields;
- MSc in Statistics, Data Science, Computer Science or related fields;
- PhD in Statistics or related fields.

Some important statistical issues:

- Data structure changed:
  (i) Not i.i.d. any more; Not satisfy "our assumption".
  (ii) Usually observational data;
  (iii) Big and others

- Causality or Correlated?
  (i) Causal Inference and Big Data.
  (ii) How to combine information from many different studies.
- Need New Statistical Framework?

Statisticians are experts in:

- **producing useful data (Big or small);**
  Survey Sampling; Experiment Designs.

- **analyzing (Big) data to make meanful results;**
  With some possible new statistical methods and **computational skills**

- **drawing practical conclusions.**

Statistics is a "GAME" between human and nature "THROUGH DATA".

Data Scientist versus Statistician

- A data scientist is a statistician who lives in certain areas (San Francisco, etc.).
- A data scientist is someone who is better in statistics than any software engineer and better at software engineering than any statistician.

Thank you!