

# Statistics = Data Science ?

C. F. Jeff Wu

University of Michigan, Ann Arbor

- What is “Statistics”?
- A Statistical Trilogy
- Frontier and Beyond
- A Bold Proposal

## Layman's definition and perception of statistics and statisticians

- Oh, you are doing **accounting!**
- Descriptive statistics:  
  
tables and charts (sports, economy),  
summary figures (from surveys, census, opinion pools),  
occasionally standard errors are attached.
- Lies, Damned Lies, and Statistics.

## Old definitions of “Statistics”:

*Statistik* used by the German statistician G. Achenwall in **1748**, from the Latin word “status” (state or condition)

1. ... teaches us what is the political arrangement of all the modern states of the known world (W. Hooper tr. *Bielfeld’s Elem. Universal Educ.*, **1770**)
2. ... an inquiry into the state of a country, for the purpose of ascertaining the quantum of happiness enjoyed by its inhabitants, and the means of its future improvement (Sir J. Sinclair *Statist. Acc. Scot.*, **1798**)
3. ... a form of knowledge - a mode of arranging and stating facts which belong to various sciences (*Lond. & Westm. Rev*, **1838**)
4. ... consists in the observation of phenomena which can be counted or expressed in figures (Mayo-Smith *Statist. & Sociol*, **1895**)

## Dictionary definitions of “Statistics”:

1. Science that deals with the collection, tabulation, and systematic classification of quantitative data (*Funk and Wagnalls Stand. College Dict.*, 1963).

*statistician*: one skilled in collecting and tabulating statistical data.

2. The mathematics of the collection, organization, and interpretation of numerical data (*American Heritage Dict.*, 1981).

3. Science of collecting and analysing numerical data (*Oxford Modern English Dict.*, 1996).

4. Science dealing with the collection, analysis, interpretation, and presentation of masses of numerical data (*Webster's Third New International Dict.*, 1966).

In Chinese language,

Statistics

統

collecting

計

counting

Accounting

會

計

Do statistics and statisticians deserve this public image or stereotype?

Yes and No !

The current state of statistical work can be described by a **Statistical Trilogy**:

1. Data Collection (experimental design, sample surveys)
2. Data Modeling and Analysis
3. Problem Understanding/Solving, Decision Making

## Promising Current/Future Directions:

- Large/complex data:  
neural network models,  
data mining (of massive data bases)
- Empirical - Physical Approach:  
driven by data and mechanistic knowledge,  
mechanistic:  
unknown state  $\xrightarrow{\text{deduction}}$  manifestation  
statistical:  
unknown state  $\xleftarrow{\text{induction}}$  observed data
- Representation and Exploitation of Knowledge:  
Representation of knowledge as a Bayesian prior and model (possibly in high-dimensional spaces), Computational algorithm, interaction with cognitive science

Why can neural network modeling solve some complex/tough problems?

- can model complex (i.e., nonlinearity, interaction) relationships
- use cross-validation and other statistical techniques to find parsimonious models and gain predictive power
- good at developing simple and efficient computational algorithms, develop problem-specific hardware

Think Big, Learn from Others!



- Tremendous progress has been made in image reconstruction:  
penalized maximum likelihood,  
Bayesian Gibbs sampling
- Much less is known and much needs to be done in computer vision:  
“Vision is a *process* that produces from images of the external world a *description* that is *useful* to the viewers and not cluttered with irrelevant information (Marr, 1976)”
- Computer vision:  
an infusion of psychophysics, neural physiology, statistics, engineering and artificial intelligence

Some suggestions:

- A balanced curriculum:  
more emphasis on data collection,  
scientific/mathematical basis for modeling,  
computing for large/complex systems
- Interdisciplinary training:  
requirement of a cognitive minor,  
joint teaching by statisticians and scientists
- A radical idea:  
an applied master or doctoral program with  
30% - 50% courses outside statistics

- Long tradition and deeply rooted perception of statistics  $\implies$  difficult to break this undeserving image
- It is time in the history of statistics to make a bold move
- A good role model,

Professor Harry Clyde Carver

founding editor of the Annals of Mathematical Statistics (1930 - 38),

founding member of the Institute of Mathematical Statistics (1935)

foresight, courage, unorthodox approach

A proposal:

“Statistics”       $\longrightarrow$    “Data Science”

“Statisticians”    $\longrightarrow$    “Data Scientists”

- Several good names have been taken up:  
computer science, information science,  
material science, cognitive science
- “Data Science” is likely the remaining good  
name reserved for us
- “Statistical Science” not as attractive, but  
much better than “Statistics”

## Summary

- Descriptive statistics is a small part of statistical work
- Data collection  $\implies$  data modeling/analysis  
 $\implies$  problem solving/decision making
- Statistical education:  
more balanced and science driven
- More focus on large/complex data,  
interface with other disciplines
- A joint data - knowledge approach to problem solving:  
knowledge from physical, engineering,  
cognitive, . . .