Similarities & Differences:

Traditional Statistics and DSML

Chong Ho Alex Yu, PhD
Charlene Yang, MA

TABLE OF CONTENTS

Introduction & Overview

01

Overview and Definitions

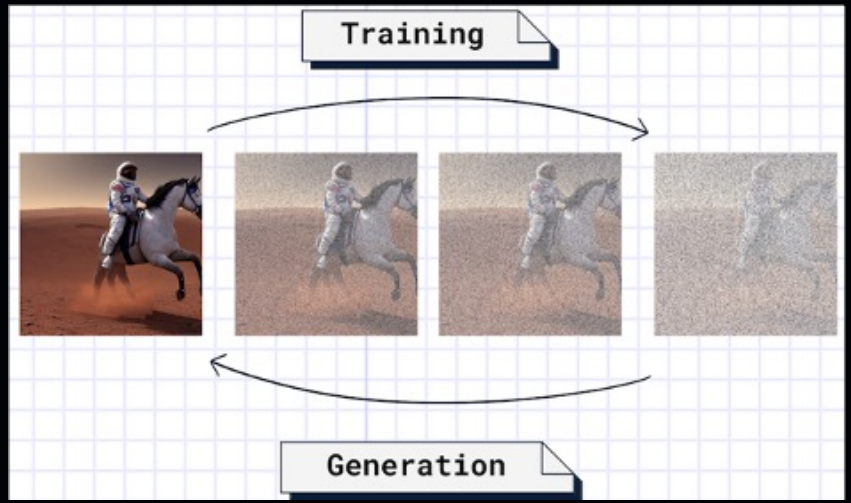
Introduction

- ➔ The roots of DSML and relationship to statistics are understudied.
- ➔ Many definitions of DS
- ➔ Coined as early as the 1970s as popularity grew
- ➔ Theory of interdisciplinary data science implemented by several institutions.

Overview of DS

Data Science and Machine Learning are distinct concepts

- ➔ Machine learning definition.
  - ◆ Used for facial recognition, image generation, etc.
  - ◆ Algorithms
- ➔ Data Mining vs. Data Science
  - ◆ Data Mining – structured data
  - ◆ Text mining – unstructured data



Images from: Analytics Insight & Hypotenuse AI

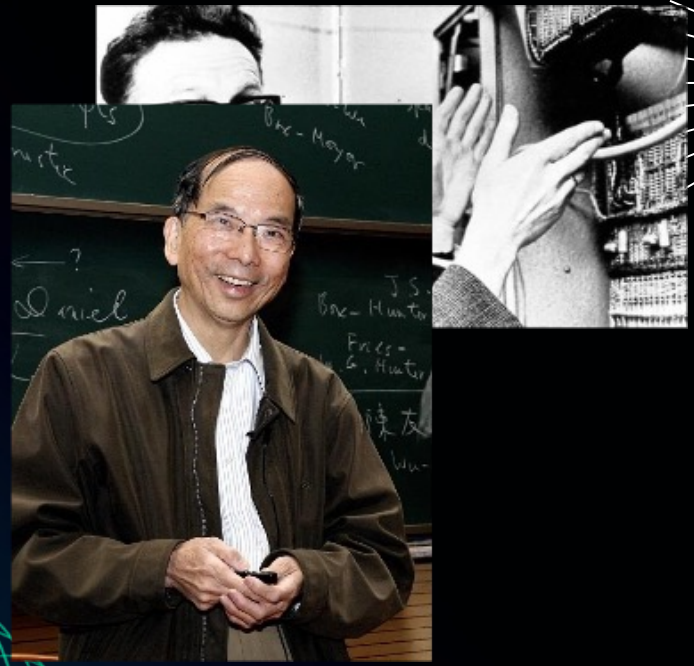
Multiple Origins

02

4 Major Sources of Contribution to DSML

Origins of Data Science

- ➔ No definitive founder of the movement of data science.
- ➔ Peter Naur's definition of DS
- ➔ Prof CF Jeff Wu's call to rename statistics as DS.
- ➔ DS traced back to at least four major sources.
  - ◆ EDA
  - ◆ Notion of Learning
  - ◆ Data Viz & Computing
  - ◆ Two-Culture Thesis

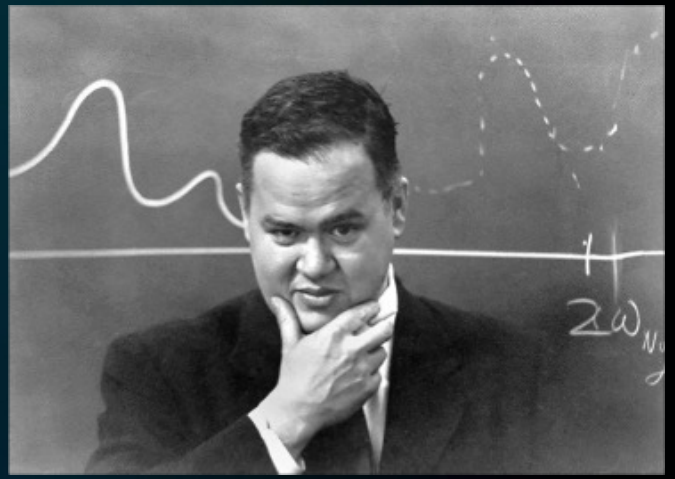


Exploratory Data Analysis

Tukey was a critical player in the world of statistics.

- Professor and founding chairman of the statistics dept at Princeton University
- His key point: Data analysis should be a new science not a branch of mathematics.
- Advocated data-driven exploration.
- Toolkits of EDA:
  - Data Visualization (revelation)
  - Data Transformation (re-expression)
- When structure cannot meet model assumptions, re-expression is needed.
- Inspired development of neural networks.

John Tukey



Notion of Learning

John Chambers

Chambers is a highly influential statistician

- ➔ Co-creator of the S programming language which grew into the R language.
- ➔ In his article, *Greater or Lesser Statistics*, he states
  - ◆ **Lesser statistics** is confined to academia.
    - Focus is on math and collaboration is rare.
  - ◆ **Greater statistics** is more inclusive.
    - Related to other disciplines, versatile, and comprehensive.
- ➔ Traditional statistics might become marginalized.

Data Viz & Advanced Computing

Cleveland was a distinguished Professor of Statistics and Computer Science at Purdue University.

- ➔ Well known for his work in data visualization.
- ➔ Argued that data viz can provide deep insight.
- ➔ Many of his graphing tools are still used today.
- ➔ In 2001, he published an article proposing a new discipline of DS which synthesizes computing methods and statistics.
- ➔ Suggested statisticians incorporate advanced computing into data analysis.

William Cleveland

Two-Culture Thesis

Leo Breiman

Breiman was a distinguished statistician from UC Berkeley

- He invented several DSML methods including
  - classification and regression tree and random forest.
- In his seminal article, Breiman outlined two directions of data analysis.
  - Traditional approach looks to infer from sample stats to true population parameters.
  - prediction-centric approach uses algorithmic models without data mechanism assumptions.

Traditional Statistics vs Data Science

03

Differences Between

Dichotomous Evidence and Decision

vs.

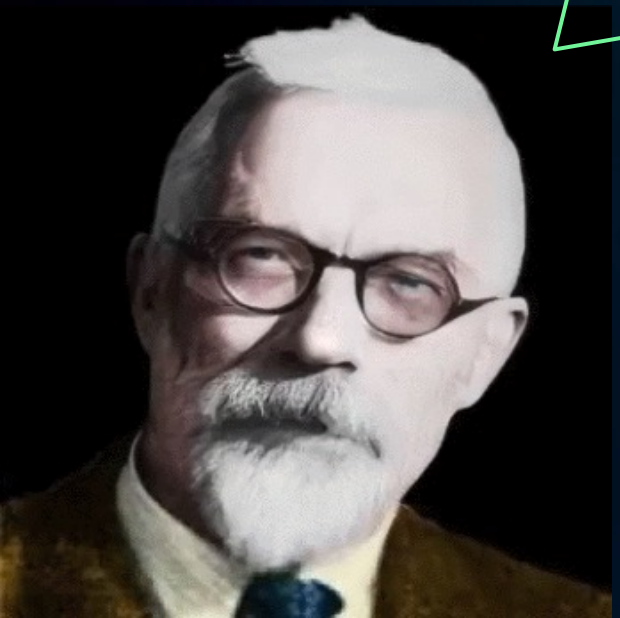
Hypothesis testing: only one fixed constant that represents the true parameter.

When alpha = .05 or .01 is adopted, not only is the final decision dichotomous, but also the evidence is binary.

Data science seeks to identify the patterns and trends in data, which is consistent with Chambers' (2001) notion of learning from the data.

Pattern Recognition and Contextual Decision

Model-Driven and Assumption-Based



Fail to reject or not to reject, that is the question!

In traditional statistics,

- Researchers begins with a pre-formulated hypothesis or model.
- Then collects data to confirm or reject.
- This approach requires parametric assumptions

R.A. Fisher's Philosophy:

- The normal distribution has two characteristics: mean and variance.
- **Mean** determines the **bias** of our estimate
- **Variance** determines its **precision**.
- **Estimation** is more precise as the variance becomes smaller and smaller.

In reality, much of data is not normally distributed.

Data-Driven and Assumption-Free

vs.

To rectify the the situation:

- ➔ Most DS methods are data-driven and assumption-free.
- ➔ Data Scientist could explore hundreds to thousands of potential predictors simultaneously.
- ➔ Most DS methods are nonparametric.

In the real world, data is noisy and complex.

Principle of DS is fully compatible with the philosophy of EDA.

Single-Modeling vs. Multiple-Modeling

The starting point in statistics is

- Usually a simple model (e.g., linear regression)
- Entire sample is usually included in modeling process.
    - Thus, prone to overfitting.
- Data are checked to find out whether the data structure meets zions.
- The model is improved by remediating the assumption violations.

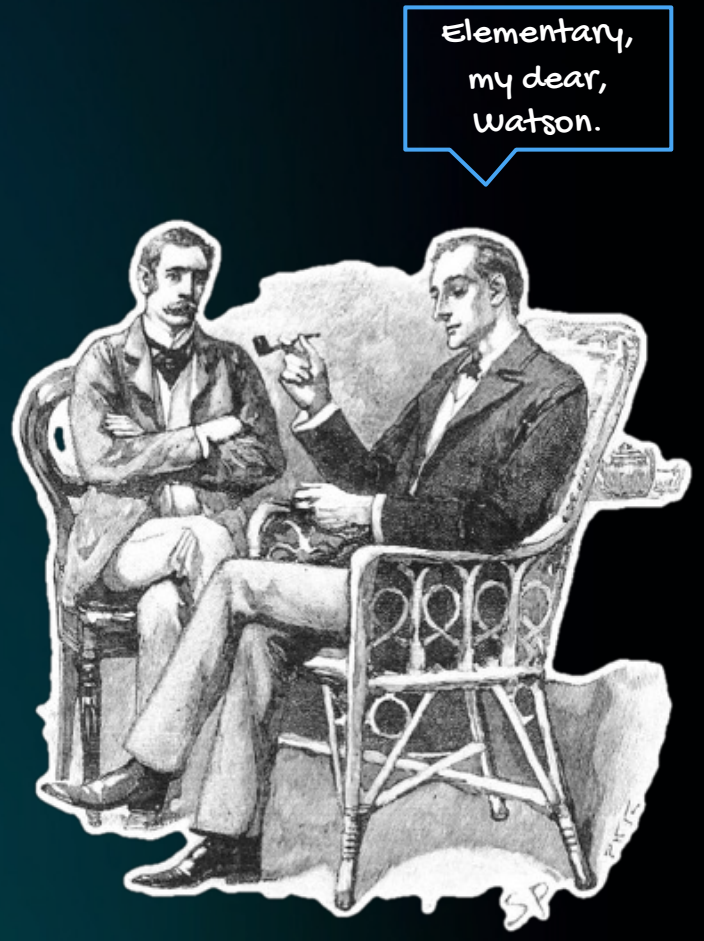
Data scientists usually run,

- Many models using a variety of modeling techniques.
- Sample is randomly partitioned into subsets for cross-validation.
- In the end, the best model is chosen on the basis of predictive accuracy, variance explained, and error rate.
- This race-to-the-top analytical approach is in accordance with inference to the best explanation.

Inference and Explanation vs. Prediction

The primary goal of traditional statistics is to,

- Infer from the sample statistics to the population parameters.
- Causal inference plays an important role in theoretical research.
- Without running a randomized experiment we cannot assert a causal explanation between variables.
- The logic: causal inferences are weakened in quasi-experiments and that non-experimental data cannot be used to infer cause and effect relationships.



Inference and Explanation vs. Prediction

Some researchers argued that many causal inferences are made without using the experimental framework.

Since the introduction of Linear Structural Equation (LISREL) in the 1970s, **Structural Equation Modeling (SEM)** has been widely applied to uncover causal structures from non-experimental data.

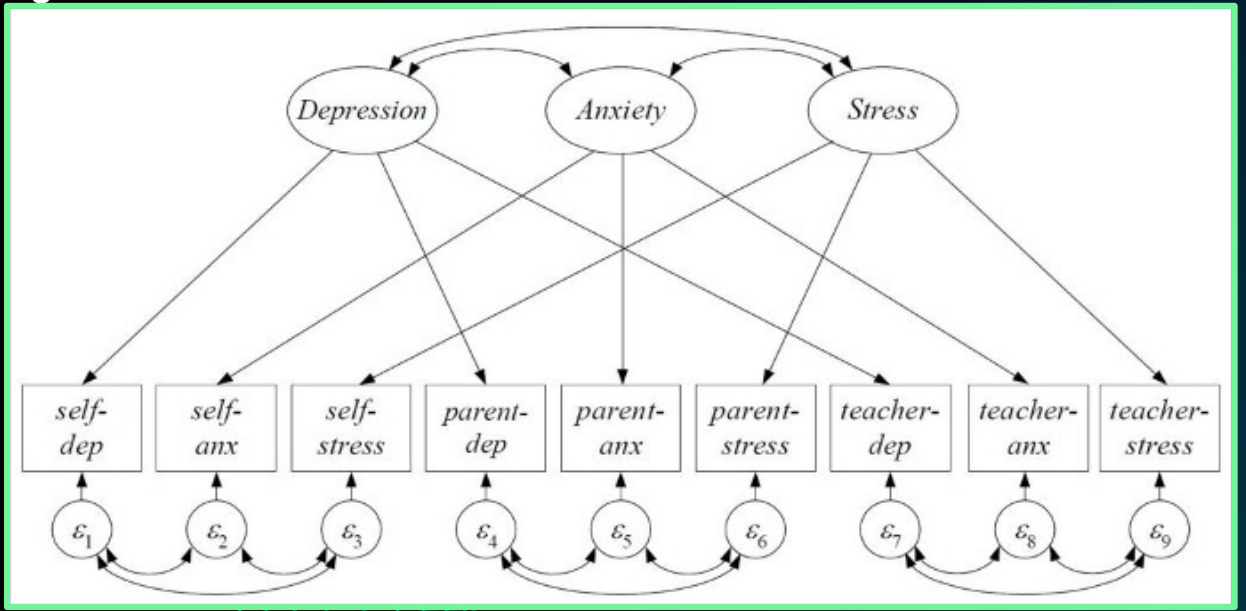


Image from: CenterStat

Inference and Explanation vs. Prediction

The preceding inference debate did not occur in the realm of data science because,

Prediction, **recommendation**, and **search optimization** for patterns and associations from big data, *rather than causal inference*, was more central to data science, especially in business analytics.

Overall Tendencies vs. Personalized Recommendation

Traditional approach to statistics emphasizes **tendencies** and **representations** derived from combining individual data points.

An example is Adolphe Quetelet's (1835) On Man: Essays on Social Physics.

Quetelet (1835) found that

- Different scientists obtained different results even though they observed the same astronomical phenomenon.
- He applied this phenomenon of astrophysics to social studies by developing the concept of the "average man."

Personalized Recommendation

Machine learning has been widely used in the development of recommendation systems that suggest appropriate content and items to users based on their preferences.

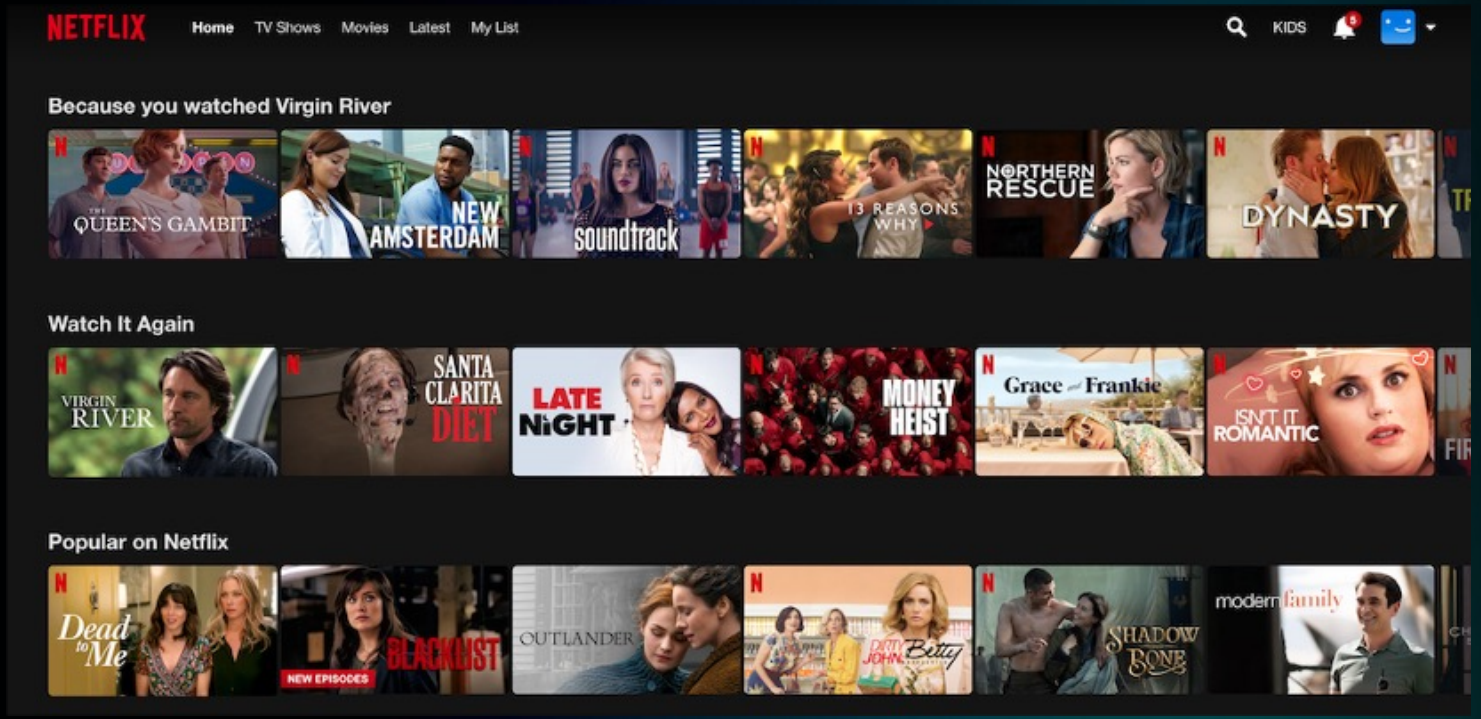


Image from: InData Labs

Personalized Recommendation

ML also plays an important role in personalized medicine.

- Utilizing algorithms for early detection of diseases.
- By using genetic profiles, biomarkers, and clinical characteristics, ML algorithms can identify patient subgroups.
- The stratification of patients enables better research design.
- Which increases successful outcomes and reduces the risks associated with non-responsive patients.

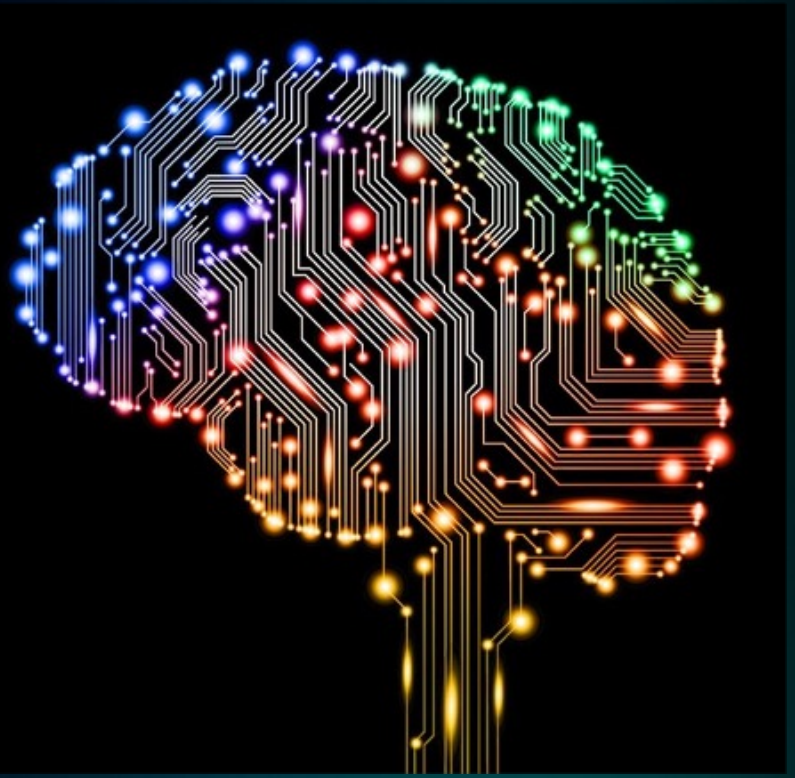


Image from: Future World Foundation

Statistics and DS: Relationships

04

Relationships Between Statistics and Data Science

Nothing can emerge from a vacuum

Certain DSML methods are based on concepts, procedures, and theorems from traditional statistics.

- LogWorth statistic used in the decision tree is derived from the p-value.
- Generalized Regression, which is a common DSML method to combat overfitting is extended from the framework of OLSR.
- Cluster Analysis and PCA are inherited from traditional multivariate statistical analysis.
- Random Forest or Bootstrap Forest is a form of resampling techniques.
- Resampling is the empirical version of the traditional sampling distribution, which is based on sampling with replacement.

Nothing can emerge from a vacuum

Not all research problems are big data problems.

- ex: Using neural networks on a data set with 50 cases and three variables is overkill.

Some simple problems can be resolved by traditional statistics

- Statistical inference works better on small data sets while ML algorithms often perform poorly on them.

In data science, predictive modeling trumps causal inferences.

- Most causal modeling techniques for observational data still rely on traditional statistics.

To develop a causal model, analysts need the right data, not necessarily big data.

Conclusion 05

Closing Thoughts

Traditional statistics and data science can supplement each other.

Can combine automated data mining techniques traditional SEM to explore alternate causal paths.

DS can help generate hypotheses and discover correlations efficiently.

For hierarchical, nested data, multi-level models (e.g. HLM) are preferable to handle the data structure properly.

Whereas traditional statistics is better suited for validating causal relationships.

CONTACT INFORMATION

Chong Ho Alex Yu, PhD, DPhil

Hawai‘i Pacific University

cayu@hpu.edu

chonghoyu@gmail.com

https://creative-wisdom.com/index.html

https://scholar.google.com/citations?user=mdGny3EAAAAJ&hl=en

Charlene J Yang, MA

Azusa Pacific University

cyang20@apu.edu

charlenejyang@gmail.com