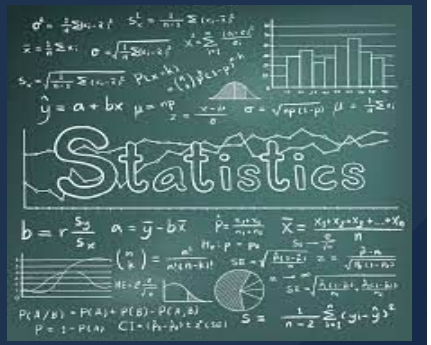
UC San Diego

50 Years of Data Science Part. 1

- Connects the discipline of DS to its 50+ years of history (John Tukey in 1960s)
- DS as an **extension** of statistics?
- Common Task Framework (e.g., Netflix Challenge)

UC San Diego

Tukey's introductory paragraphs



UC San Diego

DS as an **extension** of statistics ?

(*) Multidisciplinary investigations (25%)

(*) Models and Methods for Data (20%)

(*) Computing with Data (15%)

(*) Pedagogy (15%)

(*) Tool Evaluation (5%)

(*) Theory (20%)

UC San Diego

DS as an **extension** of statistics ?

Inference model: To [infer] how nature is associating the response variables to the input variables.

Prediction model: To be able to predict what the responses are going to be to future input variables;

UC San Diego

DS as an **extension** of statistics ?

Inference model: To [infer] how nature is associating the response variables to the input variables.

Prediction model: To be able to predict what the responses are going to be to future input variables;

Professor Breiman's paper is an important one for statisticians to read. He and Statistical Science should be applauded ... His conclusions are consistent with how statistics is often practiced in business. -Bruce Hoadley

UC San Diego

Common Task Framework and the secret sauce

- A publicly available training dataset
- A set of enrolled competitors
- A scoring referee

UC San Diego

Common Task Framework and the secret sauce

UC San Diego

Common Task Framework and the secret sauce

1. Error rates decline by a fixed percentage each year, to an asymptote depending on task and data quality.

2. Progress usually comes from many small improvements; a change of 1% can be a reason to break out the champagne.

3. Shared data plays a crucial role—and is reused in unexpected ways.