

Disciplines in data science

Statistics

- 以管窺天: explore the properties of a population by a (usually small and random) sample
- Focus on scientific insights
- Model interpretability is important
- Results are usually reliable if you do it correctly

Machine learning

- Prediction, prediction, prediction!
- Build a model or an algorithm to predict future events
- What do you mean by science? model interpretability??
- Results may be irreproducible due to some careless mistakes

ML v.s. Stat

Machine learning	Statistics
network, graphs	model
weights	parameters
learning	fitting
generalization	test set performance
supervised learning	regression/classification
unsupervised learning	density estimation, clustering
large grant = \$1,000,000	large grant = \$50,000
nice place to have a meeting: Snowbird, Utah, French Alps	nice place to have a meeting: Las Vegas in August

Big data

- Statistics or machine learning when you can (nearly) access the population
 - recommendation system
 - data mining
- Unstructured data with poor quality

Data assimilation

- Putting scientific models and observations together
 - Estimate unknown parameters in a scientific model
 - Correct forecasts by observations
 - Extrapolate observations with the help of scientific (mathematical) models
- Closely related to all the above three fields