Towards Data Science: To Be or Not to Be a Statistician

Xinwei Deng
Department of Statistics
Virginia Tech

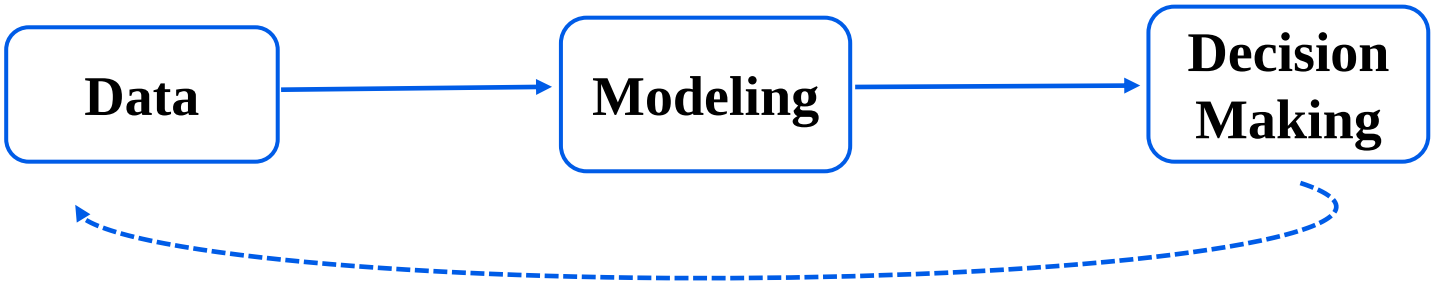
Data Points

A Random Thought
Towards Data Science

Statistician should be one of them.

But…

Data Science:
From Data to Decision Making

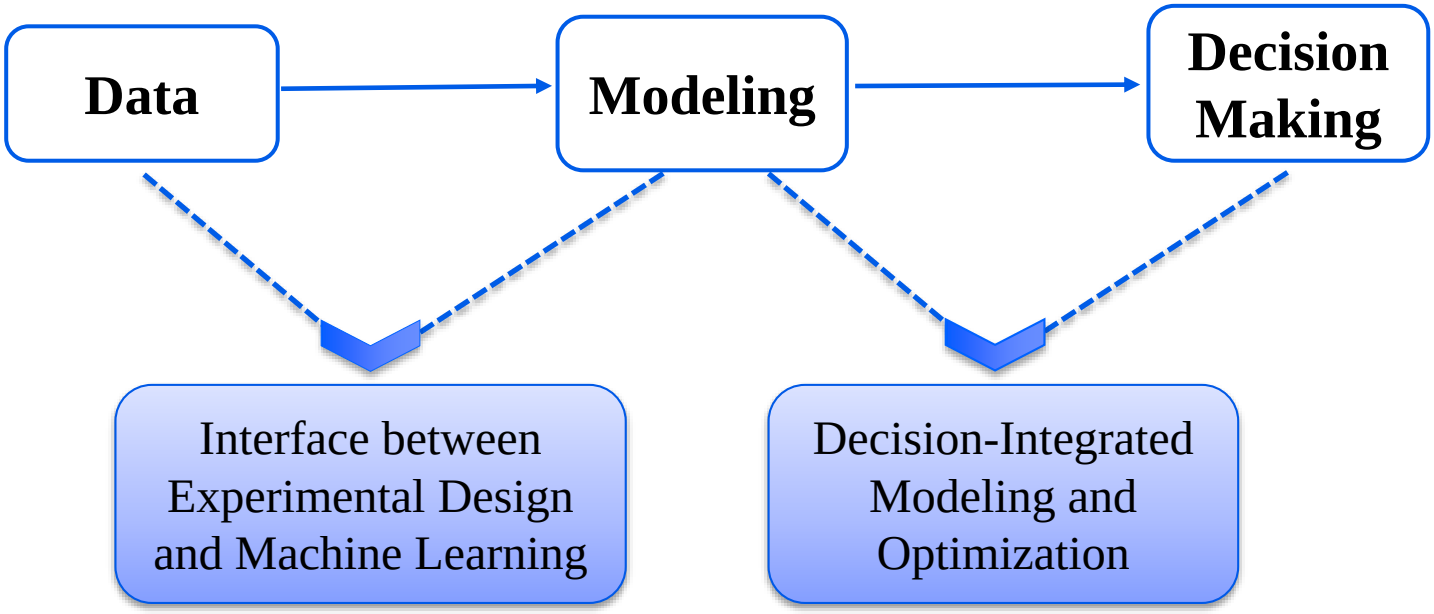


- A statistical perspective: to understand the data generation mechanism:
  - Enable model inference.
  - Enable model prediction.
  - Enable data-driven decision making (optimization).

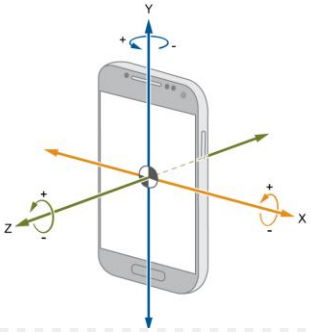
Data Science:
A Statistical Perspective

- The statistical learning is to provide a probabilistic framework for modeling and inference.
- From a statistical perspective, a challenge in Big Data mainly come from dependency:
  - Dependency among observations.
  - Dependency among features.
  - Dependency among different types of data.
  - Dependency between computation efficiency and estimation accuracy.

Towards Data Science: My Two Cents



Collaboration in Driving Risk Analytics (Mao et al., 2020)



Kinematic Data

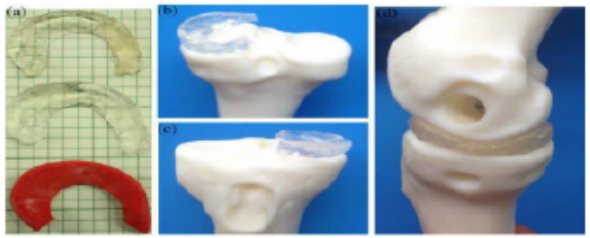
Driver Characteristics

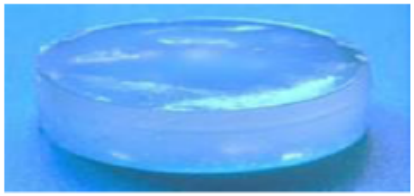
- Objective: model and assess driving risk of ride-hailing drivers.
- Big Data: millions of drivers, storage of 8 TB data daily.
- Extract informative features and develop a self-adaptive generalized additive model to predict the number of crashes.

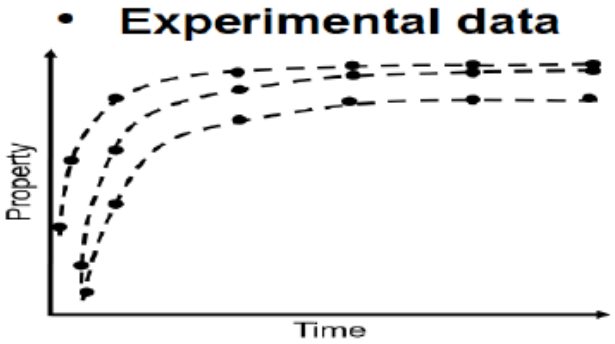
$$Y \sim \text{Poisson}(\lambda(\boldsymbol{x}) \cdot E) \text{ with } \log(\lambda(\boldsymbol{x})) = \beta_0 + f_1(x_1) + \cdots + f_p(x_p).$$

where E is the exposure measured by driving time or mileage.

Collaboration in Biofabrication (Wu et al., 2020)







Process parameter (z): Initial concentration of $CaCl_2$
Swelling time (t): 0, 0.5, 1, ..., 27 hours, 2, ..., 10 days
Swelling measurements (y): Swelling Ratio

- Propose a semi-parametric varying-coefficient model (VCM) to understand material properties in artificial meniscus.

$$y = \beta_0(t) + \beta_1(t)z + \epsilon,$$

  - $\beta_0(t)$: baseline behavior – a parametric function to incorporate knowledge of the shape.
  - $\beta_1(t)$: effect of the process parameter – a nonparametric function to handle the dynamic stability constraint.

Summary

- Data Science is truly interdisciplinary and challenging.

- Towards Data Science: Be a statistician
  - A perspective of data generation mechanism.
  - A probabilistic framework for modeling and inference.

- Towards Data Science: Not Just be a statistician
  - A perspective of system thinking.
  - A close loop of data-modeling-decision.

Thank you!

PS: To Be or Not to Be, That is Data Science.