# DATA7001
# INTRODUCTION TO DATA SCIENCE

## Module 4 Making the Data Confess Part 2

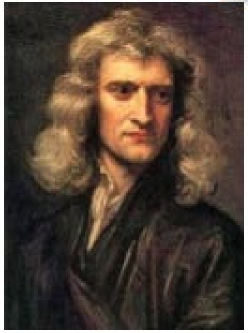
# Module Topics

- Hindsight (search and query),
  - What happened?
- **Insight (knowledge discovery)**
  - **Why is it happening?**
- **Foresight (prediction)**
  - **What will happen?**

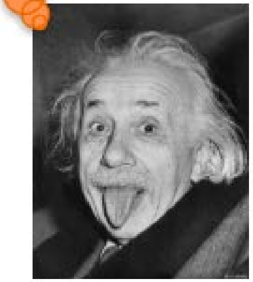
**→ What should happen (making it ‘actionable’)**

# Can We Make Data Confess?

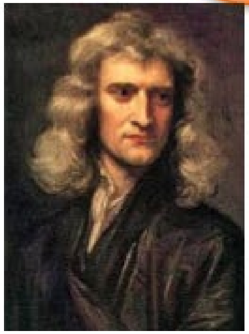
F = ma



That's not true

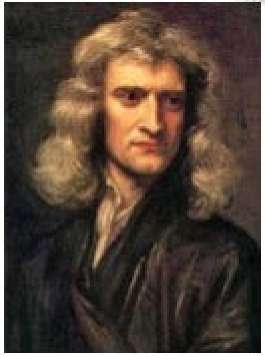


Data limits my imagination...

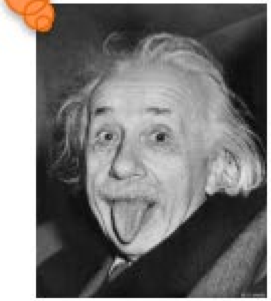


# Can We Make Data Confess?

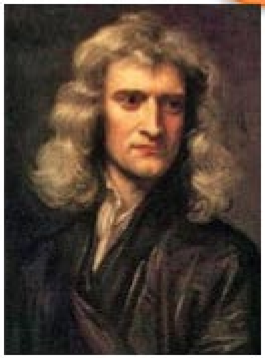
F = ma



That's not true



Data limits my imagination...

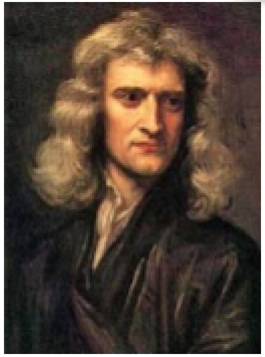


Moral of the story

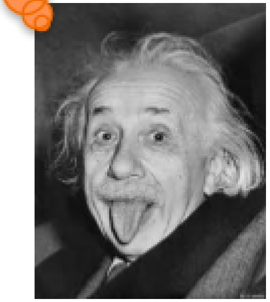
Bad news: Data never tells us that we are right, but only tells us that we are wrong.

# Can We Make Data Confess?

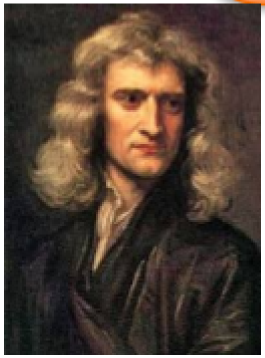
F = ma



That's not true



Data limits my imagination...



Moral of the story

Bad news: Data never tells us that we are right, but only tells us that we are wrong.

Good news: Models may be wrong, but can be useful.

# What Are Our Questions?

- We are often interested in the following questions
  - Diagnostic analysis: what caused it?
  - Predictive analysis: what will happen?
  - Prescriptive analysis: what should we do?
- Example: advertisement channels
  - We invested in advertisement via TV, radio, newspaper and observed increases in sales
  - Diagnostic analysis: which channel contributed to the increase in sales?
  - Predictive analysis: how much will the sales increase if we put in X dollars more in TV advertisement?
  - Prescriptive analysis: which channel to invest on？

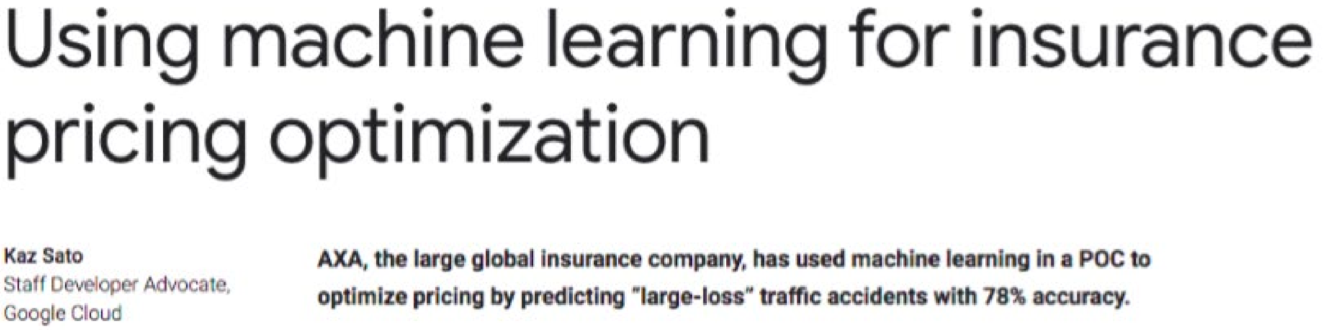
# How Do We Answer Them?

- We use a data-driven approach

  collect data -> build models using data -> use the models to answer our questions

- A model often provides a compact description for the structure of the data and relationships between variables.
- We will cover a few basic models of data
  - These are from the fields of statistics and machine learning
- All models are wrong, but some are useful.

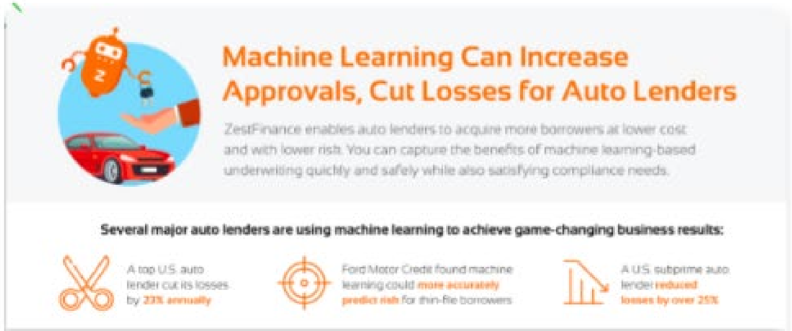
# What Are the Applications?



## Using machine learning for insurance pricing optimization

Kaz Sato
Staff Developer Advocate,
Google Cloud

**AXA, the large global insurance company, has used machine learning in a POC to optimize pricing by predicting "large-loss" traffic accidents with 78% accuracy.**



Machine Learning Can Increase Approvals, Cut Losses for Auto Lenders

ZestFinance enables auto lenders to acquire more borrowers at lower cost and with lower risk. You can capture the benefits of machine learning-based underwriting quickly and safely while also satisfying compliance needs.

**Several major auto lenders are using machine learning to achieve game-changing business results:**

A top U.S. auto lender cut its losses by 23% annually

Ford Motor Credit found machine learning could more accurately predict risk for thin-file borrowers

A U.S. subprime auto lender reduced losses by over 25%

**Financial services**

Recommended
15:45
BACK TO THE FUTURE HOVER BOARD | SKATE
Braille Skateboarding
522,458 views • 4 days ago
8:34
BOY KIDNAPPED?!
RoccoPiazzaVlogs
680,245 views • 1 week ago

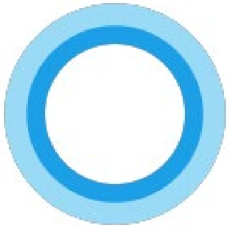
video recommendation

computer games

board games

**Entertainment**

Siri



Cortana

Alexa

Google Assistant

AliGenie

**Virtual assistants**

# Setting the Objectives

- We will NOT cover advanced applications in this course.
- We will give you basic principles and tools to dig out insights from data, and use them to make decisions.
    - Some of the techniques covered in this lecture can be directly applied to some of the applications mentioned
    - e.g. using the classifiers covered in this lecture for credit fraud detection

- We will focus on the conceptual ideas behind the tools
    - This will give you an understanding of how to apply existing software tools
- We don’t cover much of the mathematics behind the tools
    - This is the subject of a proper course on machine learning

# Task and Discussion

Give examples of (1) diagnostic analysis, (2) predictive analysis, (3) prescriptive analysis.

# What Will Be Covered

- Overview of machine learning
- Regression
- Classification
- Clustering
- Model selection

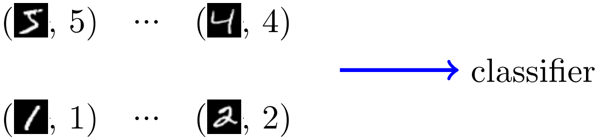
# What Will Be Covered

- Overview of machine learning
  - Machine learning approaches
  - How a learning algorithm works
  - What is the objective of learning
  - Statistical learning and prediction
- Regression
- Classification
- Clustering
- Model selection

# Machine Learning Approaches

- There are a various machine learning approaches based on
    - what data is collected, and
    - how the data is collected
- Based on the type of data being collected, there are three approaches
    - Supervised learning
    - Semi-supervised learning
    - Unsupervised learning

- Supervised Learning
  - In supervised learning, we collect a set of input-output pairs $(x_1, y_1)$, $(x_2, y_2)$...,$(x_n,y_n)$, and use it to fit a model.
  - Classification: Y is qualitative (categorical)
    - e.g. handwritten digit recognition

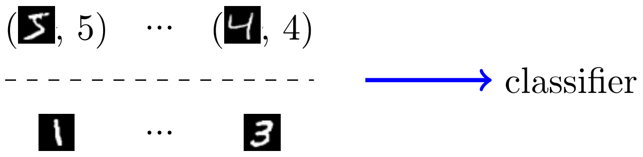


  - Regression: Y is quantitative (numerical)
    - e.g. stock price prediction

**Terminology**

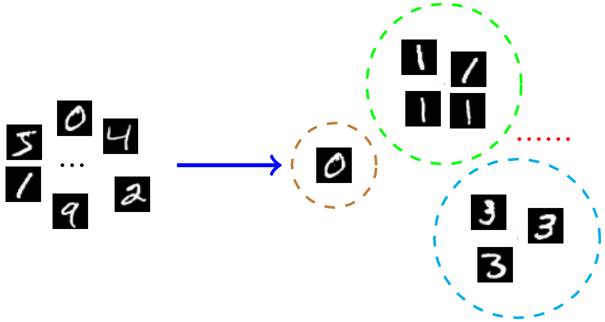
- x: input, independent variables, covariate vector, observation, predictors, explanatory variables, features.
- y: output, dependent variable, response.

- Semi-supervised learning
  - In semi-supervised learning, we collect a set of input-output pairs and a set of unlabeled inputs $x'_1$, $x'_2$, ..., $x'_m$, and use them to fit a model.
    - E.g. in handwritten digit recognition, we use both labelled and unlabeled digit images



  - Useful when it is expensive to label the inputs.

- Unsupervised Learning
    - We only collect the inputs, but not the outputs
    - Examples
        - Clustering



        - Density estimation: estimate a distribution given a sample.
        - Principal component analysis

- Many other settings of learning problems have been extensively studied
    - Online learning: examples are continuously collected
    - Reinforcement learning: learn how to act by interacting with the world
    - Active learning: we interact with a teacher to learn by asking the teacher to label example
    - …
- We do not cover these in this lecture, but to satisfy your curiosity, here are some overview articles
    - Online learning: https://en.wikipedia.org/wiki/Online_machine_learning
    - Reinforcement learning: https://en.wikipedia.org/wiki/Reinforcement_learning
    - Active learning: https://en.wikipedia.org/wiki/Active_learning_(machine_learning)

# Task and Discussion

Give examples of (1) supervised learning, and (2) unsupervised learning.

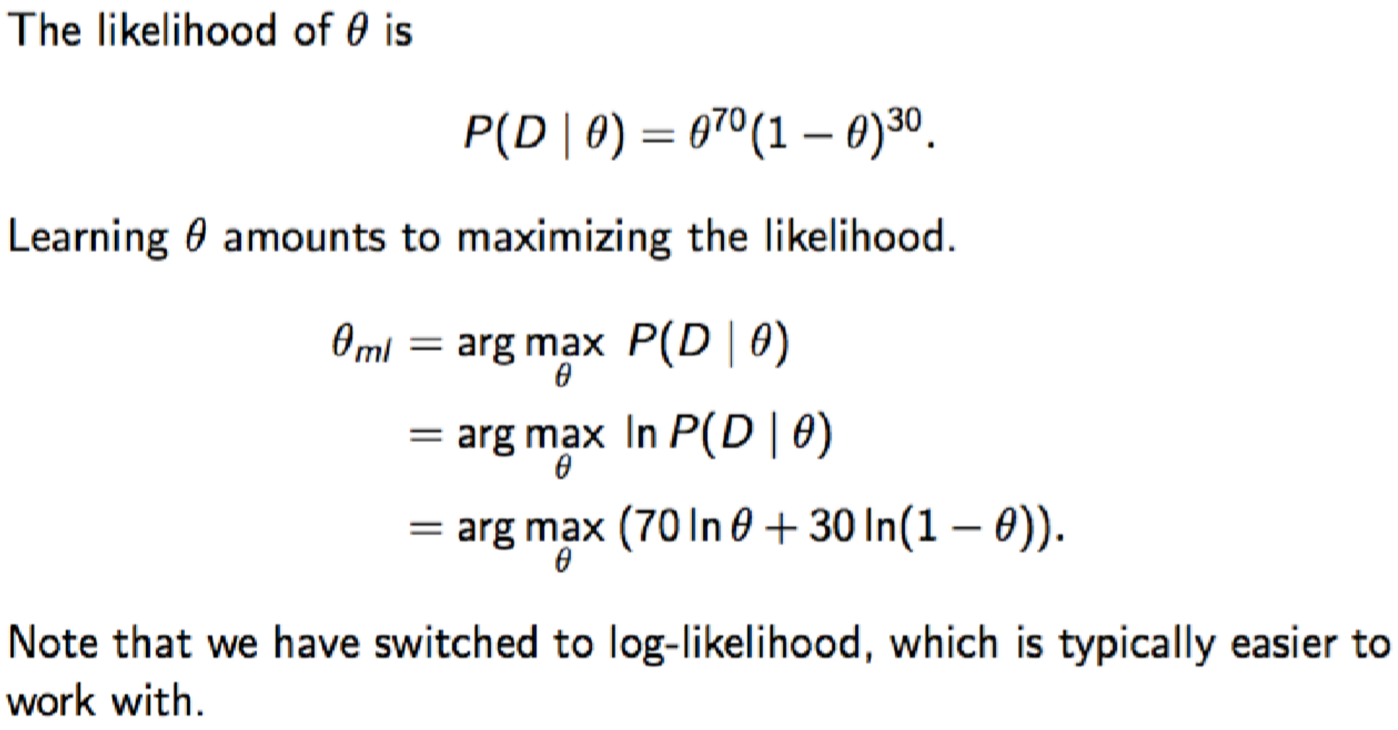
# How a Learning Algorithm Works

- We first choose a model class
- We then choose a model that best fits the training data
  - Usually the model has some real-valued parameters, and this can be cast as a numerical optimization problem
- Once we have a trained model, we can use it to make predictions on new data

# Learning a Bernoulli Distribution

- I pick a coin with the probability of heads being $\theta$, and flip it.
- You can earn a lot of money if you guess the outcome correctly.
- Assume that I've previously flip the coin 100 times for you, and you see 70 heads and 30 tails.
- What do you think $\theta$ is? What will you guess when I toss the coin?

- Maximum Likelihood estimation



The likelihood of $\theta$ is

$$P(D \mid \theta) = \theta^{70}(1-\theta)^{30}.$$

Learning $\theta$ amounts to maximizing the likelihood.

$$\begin{aligned}\theta_{ml} &= \arg\max_{\theta} \; P(D \mid \theta) \\ &= \arg\max_{\theta} \; \ln P(D \mid \theta) \\ &= \arg\max_{\theta} \, (70 \ln \theta + 30 \ln(1-\theta)).\end{aligned}$$

Note that we have switched to log-likelihood, which is typically easier to work with.

$$\theta_{ml} = \arg\max_{\theta} \left(70 \ln \theta + 30 \ln(1-\theta)\right).$$

Set derivative of log-likelihood to 0,

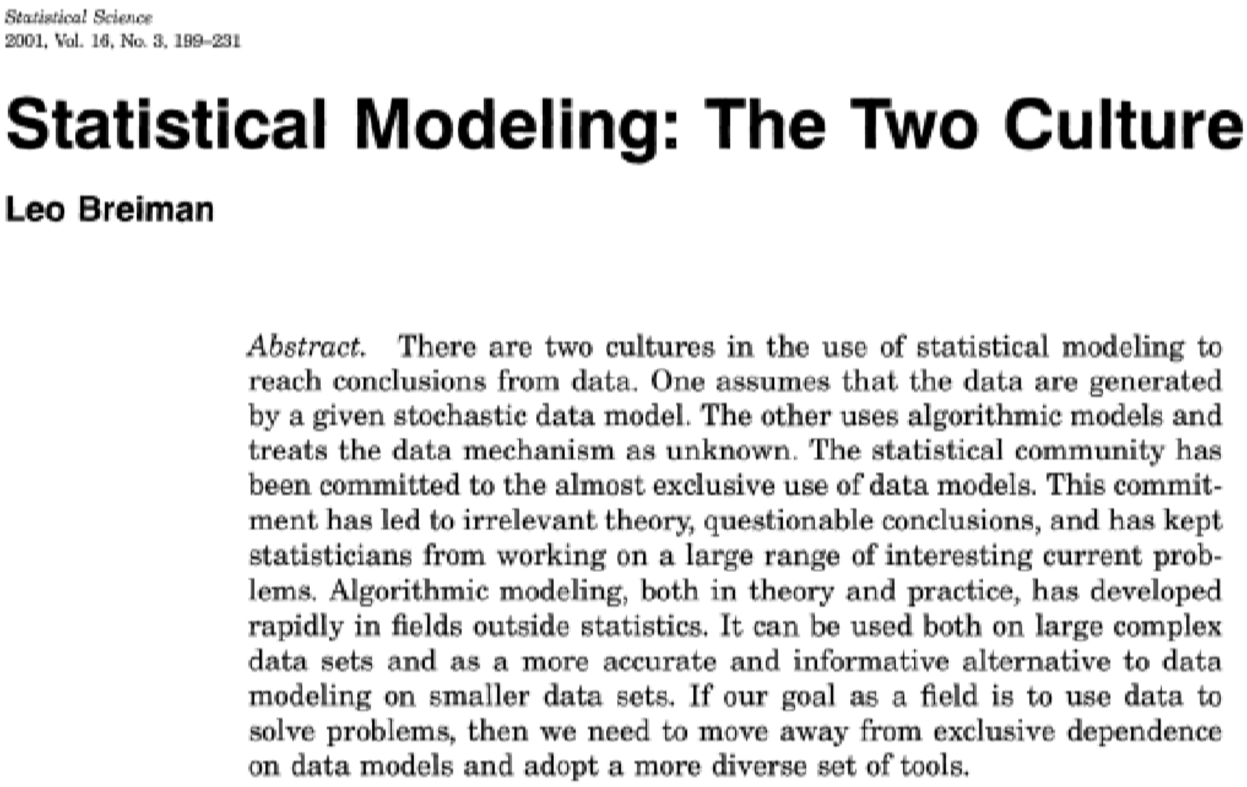
$$\frac{70}{\theta} - \frac{30}{1-\theta} = 0,$$

we have

$$\theta_{ml} = 70/(70+30).$$

- Learning is
  - Collect some data, e.g. coin toss outcomes.
  - Choose a hypothesis class, e.g. Bernoulli distribution with different head probabilities
  - Choose performance measure, e.g. log-likelihood
  - Choose an optimization procedure, e.g. set derivative to 0
  - Now you trained a model on the training data!
- Most models involve solving much complicated optimization problems
  - We will not consider how the optimization problems are solved in this lecture.

# What Is the Objective of Learning



*Statistical Science*
2001, Vol. 16, No. 3, 199–231

## Statistical Modeling: The Two Cultures

**Leo Breiman**

*Abstract.* There are two cultures in the use of statistical modeling to reach conclusions from data. One assumes that the data are generated by a given stochastic data model. The other uses algorithmic models and treats the data mechanism as unknown. The statistical community has been committed to the almost exclusive use of data models. This commitment has led to irrelevant theory, questionable conclusions, and has kept statisticians from working on a large range of interesting current problems. Algorithmic modeling, both in theory and practice, has developed rapidly in fields outside statistics. It can be used both on large complex data sets and as a more accurate and informative alternative to data modeling on smaller data sets. If our goal as a field is to use data to solve problems, then we need to move away from exclusive dependence on data models and adopt a more diverse set of tools.

- The Two Cultures paper pointed out two very different objectives of learning.
- Traditional statistics
  - We want to open nature’s black box and figure out how the data is generated
  - We make quite specific assumptions about the black box
- Predictive culture
  - It’s too hard to open nature’s black box
  - The important thing is to be able to make good predictions, possibly without knowing how data is generated

# Statistical Learning

- In statistical learning, the main assumption is that the collection of data are independently drawn from an underlying distribution
- We have a loss function that defines how a model disagrees with a single data point
    - The model may or may not represent a data generation mechanism
- The objective of learning is to find a model that has minimum expected loss
    - i.e. the model achieves good performance on new data point as well
    - This is in line with the predictive culture
- We don’t know the underlying distribution, and often choose a model that minimizes the average loss on a training set.

Learning is a.k.a. inference in statistics.

- Statistical learning for regression problems
  - Typical assumption: X and Y in the data are related by
    $Y = f(X) + \varepsilon$, where
    - f is fixed but unknown, and
    - ε is a mean zero random noise.
  - Noises often assumed to be independent, so that we can discover regularity in the data.
  - Typically, minimize mean squared error on training set

# Statistical Decision Theory

- Statistical decision theory is concerned with how to make predictions if we have learned (or are given) a model on how the data is generated
- The idea is to make a prediction that minimizes the expected loss.
- Considering predicting an output $\hat{Y}$ for a given input $X$
  - We use squared error as the loss
  - The expected squared error can be shown to decompose as follows

$$E(Y-\hat{Y})^2 = \underbrace{\left(E(Y)-\hat{Y}\right)^2}_{\text{reducible}} + \underbrace{\mathrm{Var}(Y)}_{\text{irreducible}}$$

  - Thus the optimal strategy is to predict

$$\hat{Y} = E(Y|X) = f(X)$$

# Task and Discussion

If you're betting on the outcome of rolling a biased dice, where the probabilities of 1 to 6 are 0.1, 0.1, 0.1, 0.1, 0.2, 0.4 respectively. Which outcome do you bet on?

# POLL QUESTIONS – MAKING THE DATA CONFESS - OVERVIEW