- What is Data Science
  - Exploration
  - Prediction
  - Inference
- Example: Facebook Lexicon
- **Data Makes Everything Clearer?**
- Data Makes Everything Clearer (part 1)?
  - Is there any relation between fat consumption and heart disease
    - Association equivalent to “any relation”
    - Yes - the graph points to an association
  - Does fat consumption increase heart disease?
    - Causality
    - This question is often harder to answer
- Spot Map
  - map with marking
- Comparison
  - scientists use comparison to identify association between a treatment and an outcome
    - compare outcomes of group A who go treatment to outcomes of group B who did not receive treatment
  - different results mean evidence of association
    - determining causation requires even more care
- Confounding factors
  - if treatment and control groups are similar apart from the treatment, then difference in outcomes can be ascribed to the treatment
  - if treatment and control groups have systematic differences other than the treatment, then might be difficult to identify causality
    - Such differences are often present in observational studies
  - They are called confounding factors and can lead researchers astray
- Randomize
  - If you assign individuals to treatment and control at random, then the two groups will be similar apart from the treatment
    - can account - mathematically - for variability in assignment
  - Randomized Controlled Experiment
    - may run blind experiment (placebo drug)
    - be careful with observational studies
- Comparison
  - group by some treatment and measure some outcome
    - a treatment group and a control group
  - If the outcome differs between these two groups, that’s evidence of an association (or relation)
  - if the two groups are similar in all ways but the treatment itself, a difference in the outcome is evidence of causality
  - when a group is divided randomly, it’s unlikely that there are systematic differences between sub-groups
- correlation equals causation
- **What is Data Science**
- What is Data Science
  - data science aims to derive knowledge from big data, efficiently and intelligently

    - data science encompasses the set of activities, tools, and methods that enable data-driven activities in science, business, medicine, and government
- Contrast: Databases
    - Databases/Data science
        - data value = “precious”/“cheap”
        - data volume = modest/massive
        - examples = bank records/online clicks
        - priorities = Consistency, Error recovery, Auditability / Speed, Availability, Query richness
        - structured = Strongly (schema) / Weakly or none (Text)
        - Properties = Transactions
        - Realizations Structured Query Language (SQL) / ….
    - Databases/Data Science
        - querying the past / querying the future
- Contrast: Traditional Machine Learning
    - Traditional Machine Learning / Data Science
        - Develop new (individual) models / Explore many models, build and tune hybrids
        - Prove mathematical properties of models / understand empirical properties of models
        - Improve/validate on a few, relatively clean, small datasets / develop/use tools that can handle massive datasets
        - publish a paper / take action
- **Data Science Topics**
- Data Science Topics
    - Data Acquisition
        - acquiring the data
    - Data Preparation
        - cleaning the data
    - Analysis
        - build a model and refine that model
    - Data Presentation
        - take model and present the data to people
    - Data Products
        - take complex models and turn into something that a none expert can understand
    - Observation and Experimentation
        -
- What’s Hard about Data Science
    - Overcoming assumptions
    - Making ad-hob explanations of data patterns
    - Not checking enough (validate models, data pipeline integrity, etc.)
    - Overgeneralizing
    - Communication
    - Using statistical test correctly
    - Prototype - Production transitions
    - Data pipeline complexity
- ETL (Extract Transform Load)
    - all data to a Data Warehouse
    - result of data warehouse
        - Data Products
        - Business Intelligence
        - Analytics

- Data Acquisition (Sources in Web Companies)
  - Examples from Facebook
    - Application databases
    - Web server logs
    - Event logs
    - Application Programming Interface (API) server logs
    - Ad and search server logs
    - Advertisement landing page content
    - Wikipedia
    - Images and video
- Data Acquisition and Preparation Overview
  - Extract, Transform, Load (ETL)
    - we need to extract data from the sources(s)
    - we need to load data into the sink
    - we need to transform data at the source, sink, or in a staging area
- Data Acquisition and Preparation Process Model
  - The construction of a new data preparation process is done in many phases
    - Data characterization
    - Data cleaning
    - Data integration
  - We must efficiently move data around in space and time
    - Data transfer
    - Data serialization and deserialization (for files or network)
- Data Acquisition and Preparation Workflow
  - The transformation pipeline or workflow often consists of many steps
  - If a workflow is to be used more than once, it can be scheduled
    - Scheduling can be time-based or event-based
    - Use publish-subscribe to register interest (e.g., Twitter feeds)
  - Recording the execution of a workflow is known as capturing the lineage or provenance
    - Spark's DataFrames do this for you automatically
- Impediments to Collaboration
  - The diversity of tools and programming/scripting languages makes it hard to share
  - Finding a script or computed result is often harder than just writing the program from scratch
  - view that most analysis work is "throw away"
- **Business Questions, Statistics, and Exploratory Data Analysis**
- Descriptive vs. Inferential Statistics
  - Descriptive:
    - E.g. Median - describes data but can't be generalized beyond that
    - We will talk about Exploratory Data Analysis in this lecture
  - Inferential:
    - E.g., t-test — enables inferences about population beyond our data
    - Techniques leveraged for Machine Learning and Prediction
    - Making conclusions based on data in random samples
- Applying Techniques
  - Supervised Learning: Classification and Regression
  - Unsupervised Learning: Clustering and Dimension reduction
  - Note: UL often used inside of a larger SL problem
    - e.g. auto-encoders for image recognition neural nets

- Learning Techniques
  - Supervised Learning
    - KNN (K nearest neighbors)
    - Naive Bayes
    - Logistic Regression
    - Support Vector Machines
    - Random Forests
  - Unsupervised learning
    - Clustering
    - Factor Analysis
    - Latent Dirichlet Allocation
- 5-Number Summary Statistic
  - Summary statistic provides:
    - minimum and maximum (smallest and largest observations)
    - lower quartile (Q1) and upper quartile (Q3)
    - median (middle value)
  - more robust to skewed and long-tailed distributions