**CS-429 Midterm Review**

1. **Review lecture slides**
2. **Review textbook: ‘’Doing Data Science”**
   1. ***Chapter 1 (what is data science)***
      1. Datafication: taking aspects of our lives and turning them into data.
      2. Explosion of data in the high-tech world.
      3. Hacking skills, math/statistics, and substantive expertise make up data science (diagram) according to Conway.
      4. Domains of data science: CS, match, statistics, machine learning, domain expertise, communication and presentation skills, data visualizations.
      5. No one person can be the perfect data scientist, so we need teams.
      6. Data scientists know how to extract meaning from and interpret data.
      7. Exploratory data analysis, combines visualization and data sense.
   2. **Chapter 2 (Statistical Inference, EDA, and the Data Science Process.)**
      1. The processes in our lives are data-generating processes.
      2. *Statistical Inference*: The overall process of going from the world to the data, and then from the data back to the world.
         1. Statistical inference is the discipline that concerns itself with the development procedures that allow us to extract meaning and information from data that has been generated by random processes.
         2. Key terms of Statistical Inference:
            1. Population – any set of objects
            2. Observation – characteristics of population (N)
            3. Sample – subset of observations (n)

Use to draw conclusions and make inference about the population.

May introduce bias. Be careful with the underlying “assumption” when you draw samples.

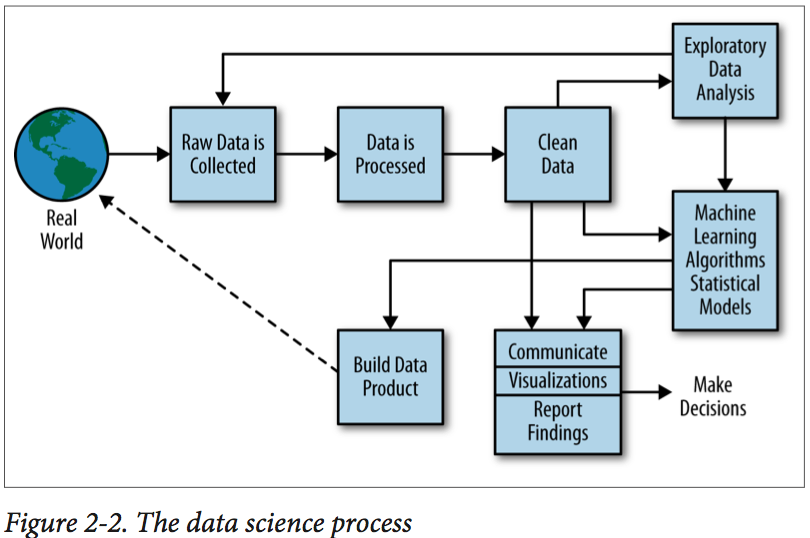
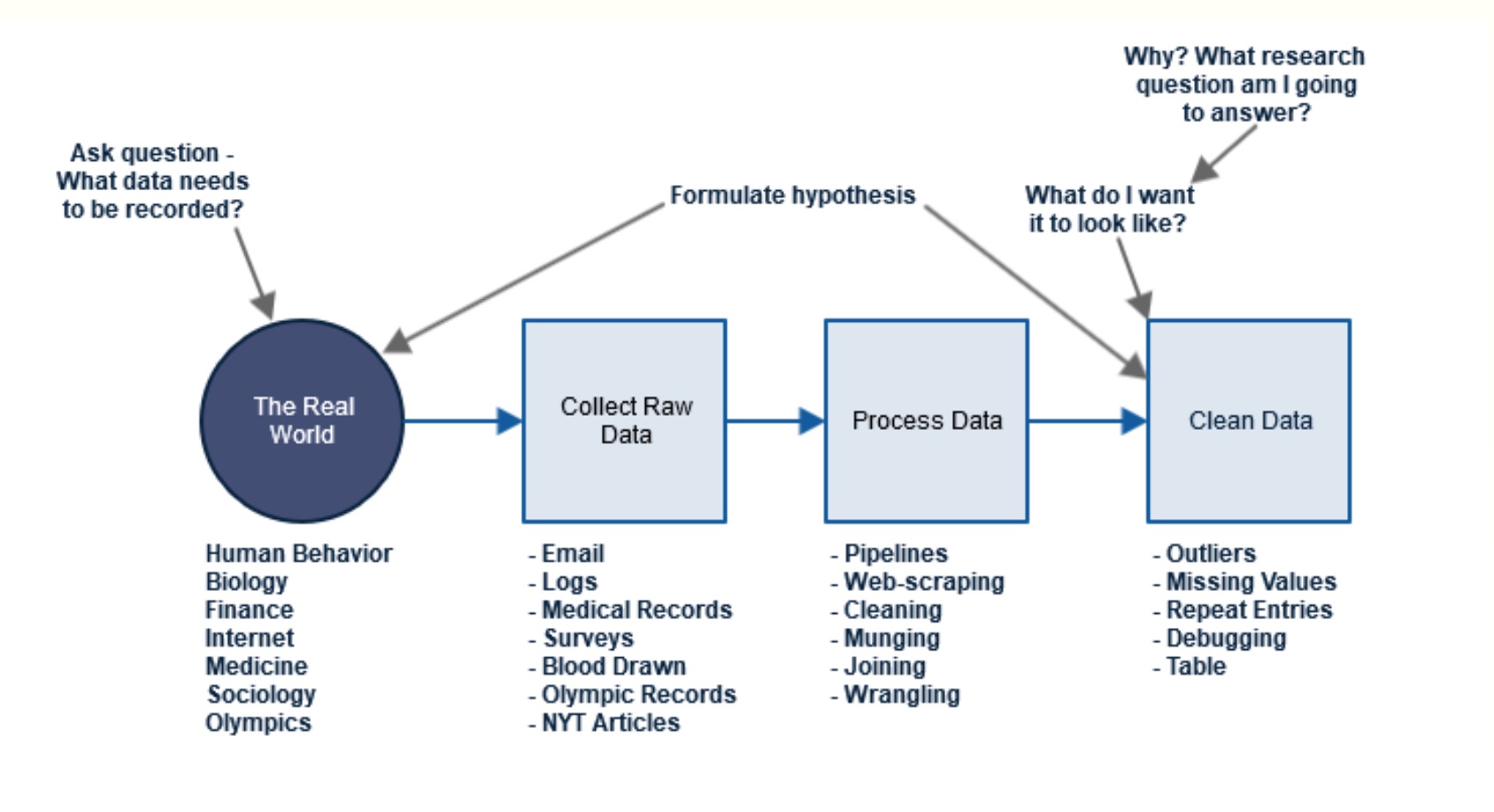
* + - 1. Big Data:
         1. Only when size becomes a challenge is it worth referring it as “Big”
         2. Big is when you can’t fit it all in one machine.
         3. The 4 Vs, Volume, Variety, Velocity and Value.
      2. Can N=All?
         1. It’s pretty much never all.

E.g. election night polls.

* + - * 1. It is wrong to believe either that data is objective or that “data speaks.” E.g. hiring men vs women example (ignoring causation).
      1. n = 1 (sample size of 1)
         1. In Big Data, for a single person, we can record tons of info about them.
      2. Modeling
         1. Our attempt to understand and represent the nature of reality through a particular lens.
         2. Artificial construction where all extraneous detail has been removed or abstracted.
         3. How do you build a model? What functional form should the data take.
         4. Need to make assumptions about the underlying structure of the reality
         5. Where to start? 🡪 EDA
      3. Exploratory Data Analysis
         1. This entails making plots and building intuition for your particular dataset.
         2. It is a method of systematically going through the data, plotting distributions of all variables (using box plots, plotting time series of data, … and generating summary statistics for all of them. EDA is a mindset
         3. EDA happens between you and the data. It isn’t about proving anything to anyone else yet!
         4. Namely to gain intuition about the data; comparisons between distributions; sanity checking (making sure the data is on the scale you expect, in the format you thought it should be.); to find out where data is missing or if there are outliers; and to summarize the data.
         5. First step towards building a model.
         6. Traditionally presented as a bunch of histograms and stem-and-leaf plots
         7. Exploratory = understanding of the problem.
         8. Basic tools: plots, graphs, and summary statistics.
         9. Helps with debugging the logging process. In the end, EDA help you make sure the product is performing as intended.
      4. Probability Distributions
         1. Foundation of statistical models.
         2. Normal Distribution

Bell-shaped curve

*u* is the mean and median and controls where the distribution is centered, and the parameter *ó* controls how spread out the distribution is.

* + - 1. Over fitting
         1. Term used to mean that you used a dataset to estimate the parameters of your model, but your model isn’t that good at capturing reality beyond your sample data.
      2. CDA vs EDA vs IDA
         1. IDA: Initial data analysis (checking assumption, handle missing value, transform variables)
         2. EDA: No Hypothesis, no model EDA encompasses IDA
         3. CDA: confirmatory data analysis concerns itself with modeling and hypothesis
      3. 
      4. 
  1. **Chapter 3: Algorithms**
     1. Types of Algorithms
        1. Data Engineering Algorithm: Data munging, preparation, processing algorithms. (sorting, MapReduce, Pregel)
        2. Optimization algorithms for parameter estimation
        3. Machine learning algorithms (predict, classify, cluster)
     2. Three basic Algorithms
        1. Linear Regression
           1. Used when you want to express the mathematical relationship between two variables or attributes.
           2. When used, one is making the assumption that there is a linear relationship between an outcome variable (response variable, dependent variable, or label, y) and a predictor (independent variable, feature, x).
           3. Loss function (measure the closeness of a fit) --Mean squared error (unbiased estimator, formula P66)
           4. How confident are you with the model estimates?

P-value: low P-value indicate high significance

R-squared: proportion of variance explained by the model.

Cross-validation: divide data into training set and test set (80%-20%), use training set to fit, compare mean squared error of test set and training set (P68 for overfitting)

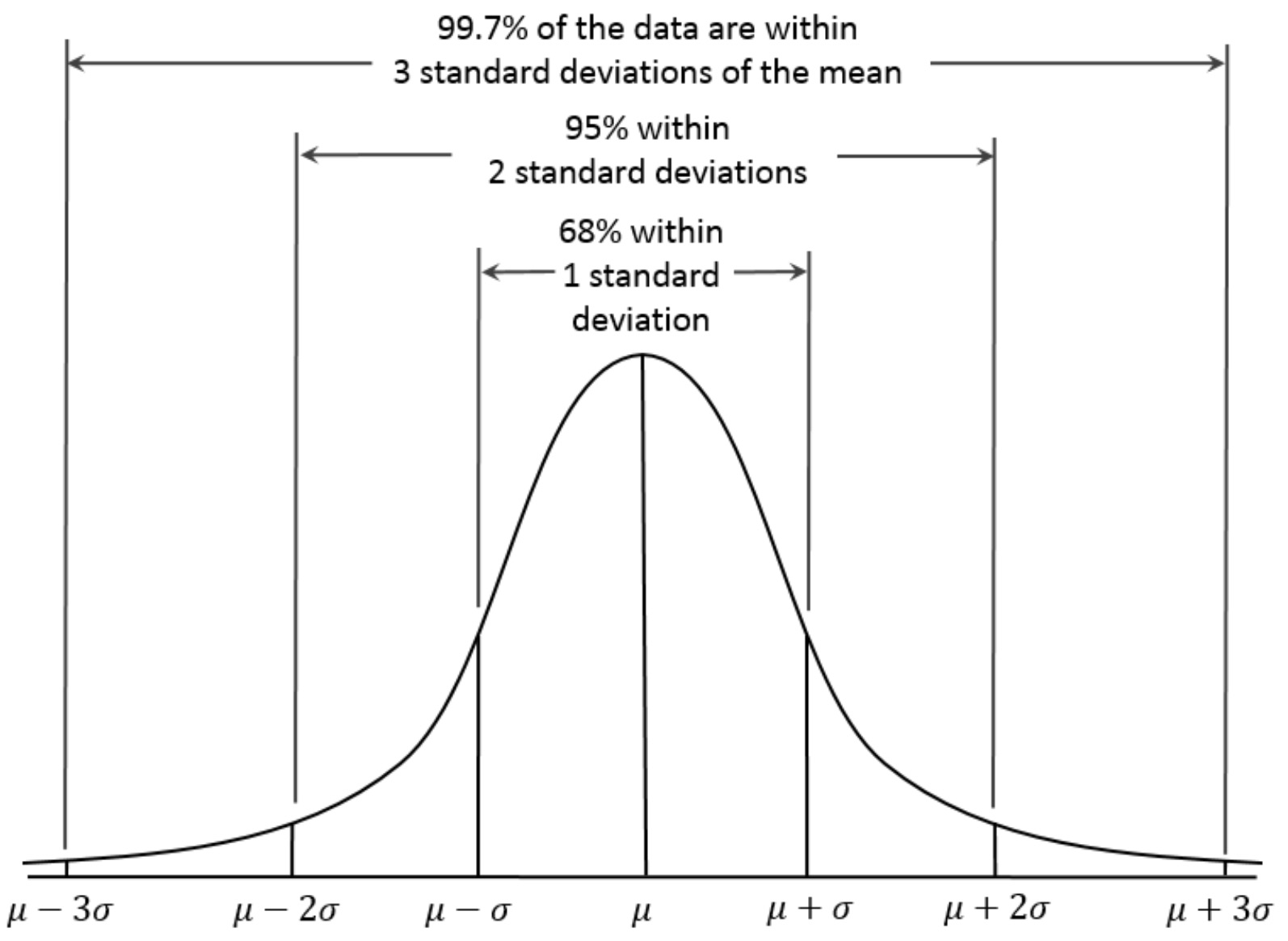
* + - 1. P-Value
         1. A number between 0 and 1

Small p-value (<=0.05) indicate strong evidence against the null hypothesis 🡪 reject the null hypothesis

Large p-value (>0.05) indicate weak evidence against the null hypothesis 🡪 fail to reject the null hypothesis

Close to the cutoff (0.05) is considered to be marginal. Always report the p-value so your readers can draw their own conclusions.

* + - 1. K-nearest Neighbors(KNN)
      2. K-Mean

1. **Review R tutorial, try out all exercises, be familiar with R basic syntax.**
   1. Basic Syntax
   2. Data Types
   3. Variables
   4. Operators
   5. Decision Making
   6. Loops
   7. Functions
   8. Strings
   9. Vectors
   10. Lists
   11. Matrices
   12. Arrays
   13. Factors
   14. Data Frames
2. **Review material posted on WISE.**
3. **Review commonly used statistical distributions, know basic facts about them**
   1. *Probability distributions describe what we think the probability of each outcome is, which is sometimes more interesting to know than simply which single outcome is most likely. They come in many shapes, but in only one size: probabilities in a distribution always add up to 1.*
   2. *Normal Distribution*
      1. Bell shaped distribution of height, IQ, etc.
      2. Completely parameterized by mean and standard deviation:
      3. Generalization of the binomial distribution
      4. 
   3. *Binomial Distribution*
      1. E.g. Heads or tails
      2. Experiments consist of n identical, independent trials which have two possible outcomes, with probabilities p and (1-p), like heads or tails.
      3. Example A biased coin comes up heads with probability 0.3 when tossed. What is the probability of achieving 0, 1, ...6 heads after six tosses?
   4. *Poisson Distribution*
      1. E.g. Over a period of time
      2. Measures the frequency of intervals between rare events.
      3. Other examples that may follow a Poisson include the number of phone calls received by a call center per hour and the number of decay events per second from a radioactive source.
   5. *Power Law*
      1. E.g. City populations, has more on an exponential curve
      2. One quantity varies as a power of another: p(x) = c^{-a} for exponent a and normalization constant c.
      3. They do not cluster around a mean like a normal distribution, instead having very large values rarely but consistently
         1. Properties of Power Laws
            1. The mean does not make sense.
            2. The standard deviation does not make sense
            3. The median better captures the bulk of the distribution
            4. The distribution is scale invariant: meaning zoomed in regions look like the whole plot
4. **Review basic statistics definitions.**
   1. Multiple Linear Regression
      1. How to choose the features?
         1. Make scatterplot of y against each of the predictors as well as between the predictors, and histograms of y|x for various values of each of the predictors to help build intuition.
      2. Adding more independent variables to a multiple regression procedure can cause OVERFITING or Multicollinearity.
      3. The ideal is for all the independent variables to be correlated with the dependent variable but NOT with each other.
5. **Understand the importance of EDA.**
6. **Understand data science as a relative new subject and data science process, data science profile.**

Questions types include answer questions, choose true or false, given R code pieces and write results.