3251 Statistics for Data Science

Data Science Fundamentals Certificate

Thank you!

Thank you for choosing the University of Toronto School of Continuing Studies

Our Courses & Programs

The School offers more than 600 courses in 80 certificates, both classroom-based and online, covering a vast range of interests and specializations:

- Business & Professional Studies
- English Language Program
- Arts & Science
- Languages & Translation
- Creative Writing

Follow us on social

Join the conversation with us online:

- facebook.com/UofTLearnMore
- @UofTLearnMore
- linkedin.com/company/university-of-toronto-school-of-continuing-studies

Certificate in Data Science Fundamentals

- Understand the techniques and methods of predictive and Big Data analytics
- Learn how to use tools such as Python and Hadoop to tackle data analysis challenges
- Develop and use models tools to solve business problems and mine data for fresh insights

Certificate in Data Science Fundamentals (cont'd)

What You'll Learn

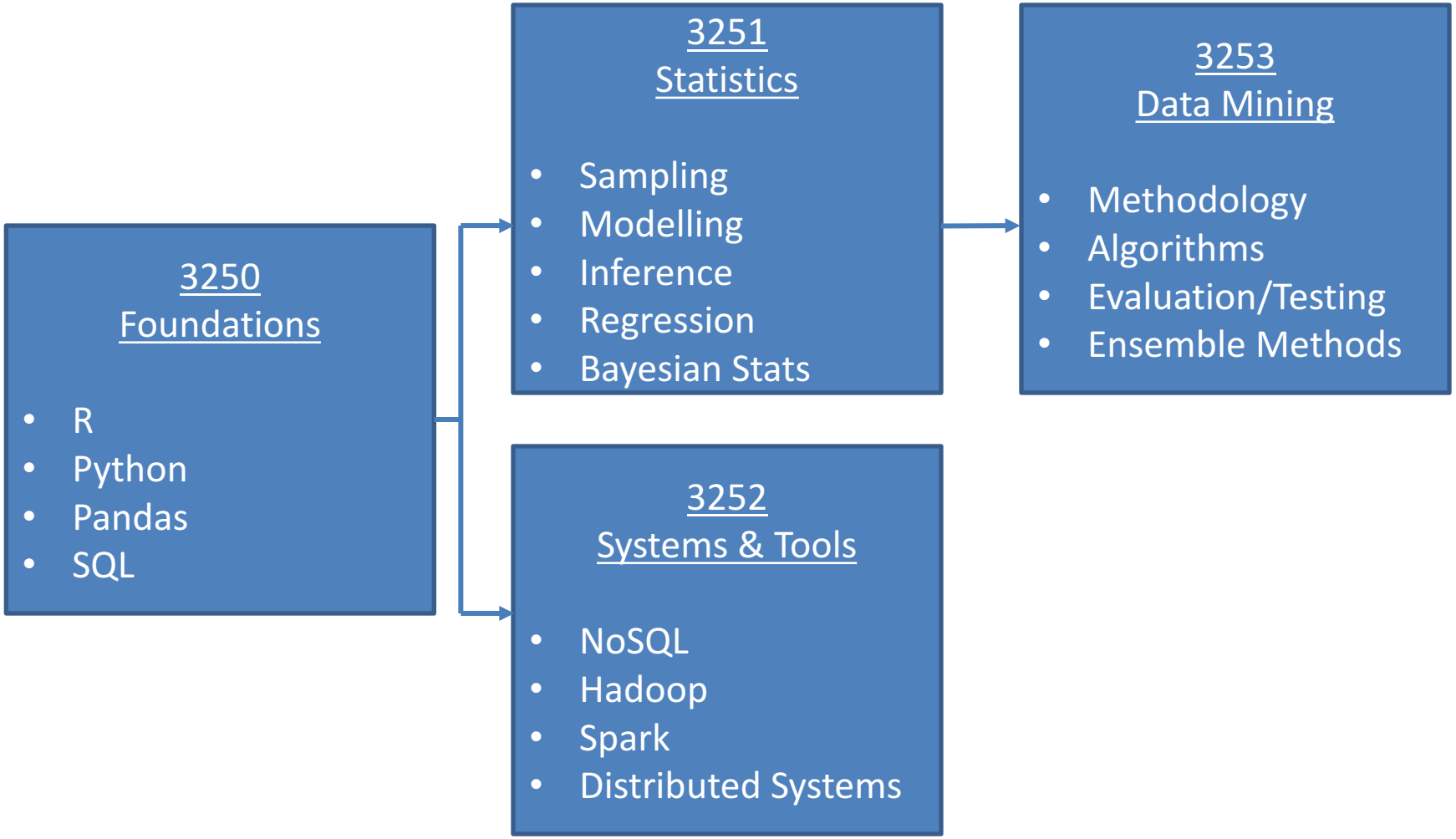
- Explore the evolution of data science and predictive analytics
- Know statistical concepts and techniques including regression, correlation and clustering
- Apply data management systems and technologies that reflect concern for security and privacy
- Adopt techniques and technologies including data mining, neural network mapping and machine learning
- Represent big data findings visually to aid decision-makers

Certificate in Data Science Fundamentals (cont'd)

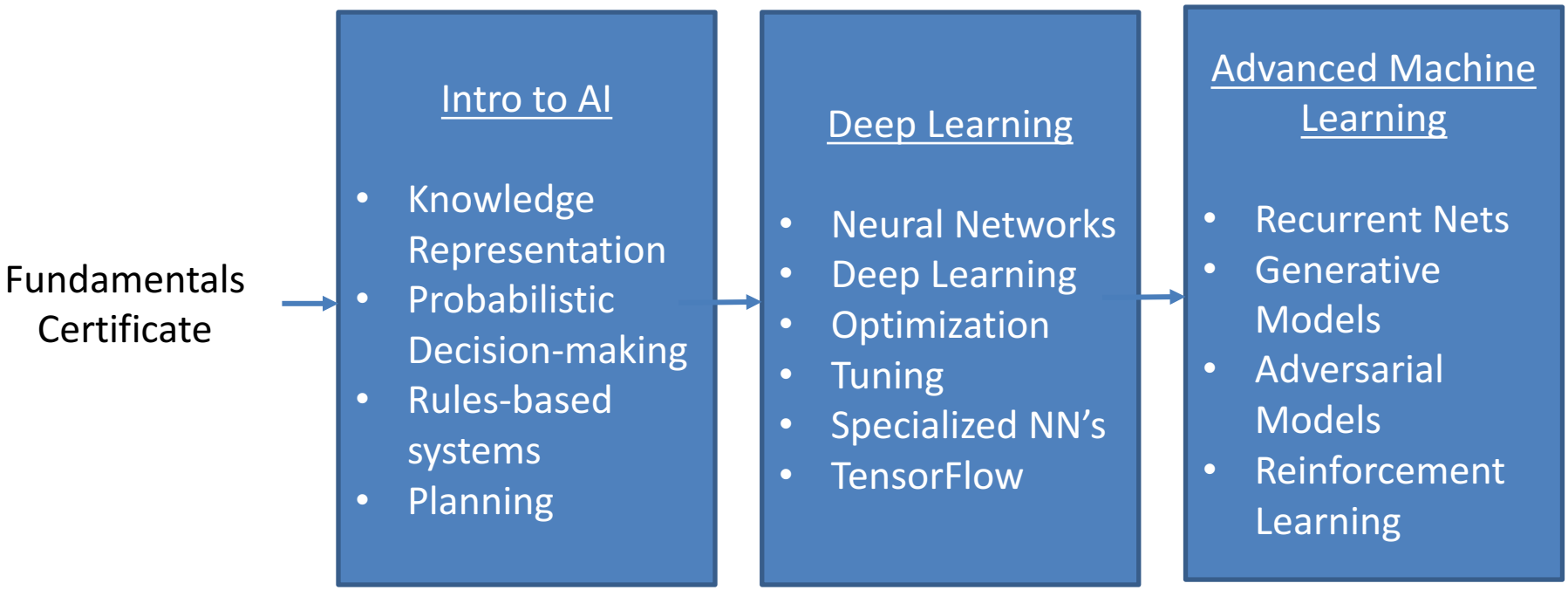
Courses

- SCS 3250 – Foundations of Data Science
- SCS 3251 – Statistics for Data Science
- SCS 3252 – Big Data Management Systems & Tools
- SCS 3253 – Analytic Techniques for Data Mining

Certificate in Data Science Fundamentals (cont'd)



Certificate in AI: Anticipated to Launch 2018



Module 1
INTRODUCTION TO STATISTICS FOR DATA SCIENCE

Course Roadmap

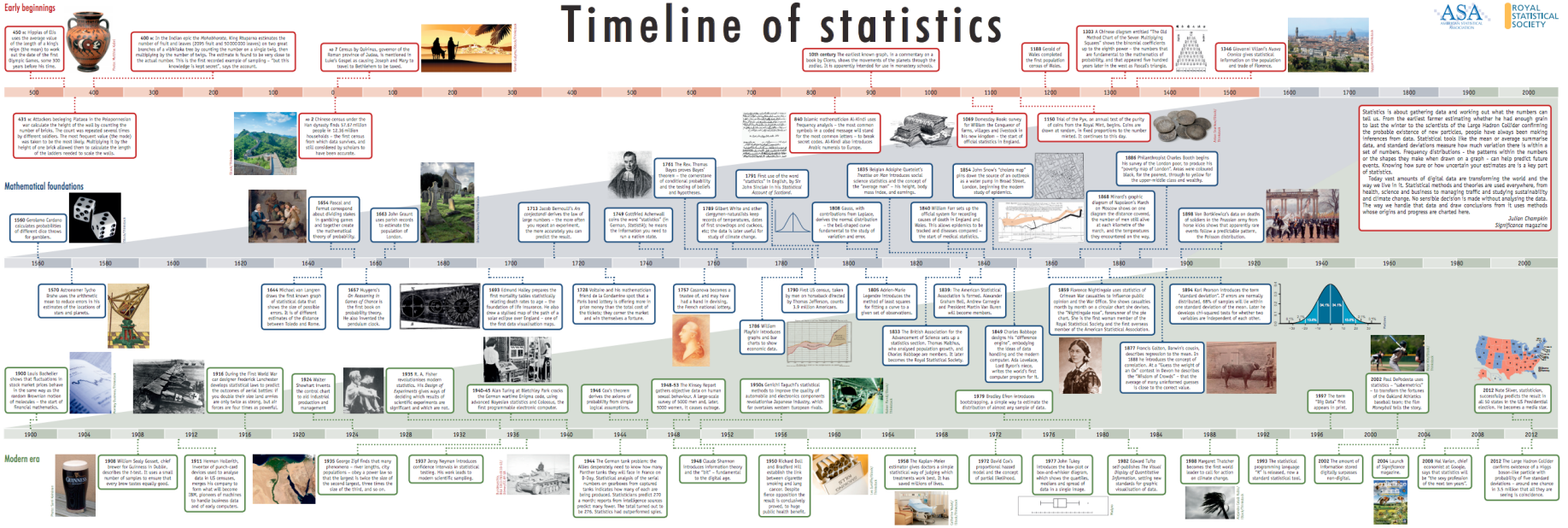
Module / Week	Title
1	Introduction to Statistics for Data Science
2	Probability
3	Distribution of Random Variables
4	Inference
5	Model Building
6	Linear Regression
7	Multiple Linear Regression
8	Logistic Regression
9	Introduction to Bayesian Inference
10	Multi-level Models
11	Markov Chain Monte Carlo
12	Presentations
13	Final Exam

Module 1: Learning Objectives

- Outline the course logistics
- Review basics of working with data
- Overview of data collection principles
- Review sampling strategies
- Discuss bias

INTRODUCTION TO STATISTICAL SCIENCE

Brief History of Statistics



Source: http://www.statslife.org.uk/images/pdf/timeline-of-statistics.pdf

Brief History of Statistics (cont'd)

- Beginning of civilization – census of the population or trade records
- 5th century BCE - Thucydides in "History of the Peloponnesian War" describes the use of mode to determine the height of the walls
- 9th century CE – earliest writing on statistics, "Manuscript on Deciphering Cryptographic Messages", written by Al-Kindi on the use of statistics and frequency analysis to decipher encrypted messages.
- 14th century, *"Nuova Cronica"*, history of Florence – use of statistical information on population, commerce and trade, education, religious facilities, and has been described as the first introduction of statistics as a positive element in history.

Brief History of Statistics (cont'd)

- 1662 – birth of Statistics, development of early statistical and census methods by John Graunt and William Petty as a framework for modern demography.
- 1713 – *Ars Conjectandi* (Latin for "The Art of Conjecturing"), book by Jacob Bernoulli on probability theory, containing the very first version of the law of large numbers. Abraham de Moivre's *The Doctrine of Chances* (1718) – first textbook on probability theory.
- 1761 – Thomas Bayes proved Bayes' theorem which describes the probability of an event, based on prior knowledge of conditions that might be related to the event.
- 1765 – Joseph Priestley invented the first timeline charts.

History of Statistics, continued

- Beginning of 18th century – central limit theorem (Pierre-Simon Laplace), and method of least squares (Carl Friedrich Gauss) and further development of the theory of errors later in 18th century
- Late 19th and early 20th centuries – emergence of modern statistics, establishment of Royal Statistical Society and American Statistical Association
- Same time – development of correlation, principal component analysis (1903), design of experiments (1935)

Examples of Application of Statistics

- **Astrostatistics** - used to process the vast amount of data produced by automated scanning of the cosmos, to characterize complex datasets, and to link astronomical data to astrophysical theory
- **Business analytics** – applies statistical methods to data sets (including Big Data) to develop new insights and understanding of business performance and opportunities.
- **Demography** – the statistical study of populations.
- **Reliability Engineering** – is the study of the ability of a system or component to perform its required functions under stated conditions for a specified period of time

Statistics Is Not Hard

John Rauser keynote: "Statistics Without the Agonizing Pain"
-- Strata + Hadoop 2014
https://youtu.be/5Dnw46eC-0o

DATA

Data - Definition

"Information, especially facts or numbers, collected to be examined and considered and used to help with making decisions"

-- Cambridge Dictionary

"Information in raw or unorganized form (such as alphabets, numbers, or symbols) that refer to, or represent, conditions, ideas, or objects. Data is limitless and present everywhere in the universe. See also *information* and *knowledge*."

-- Business Dictionary

Working with Data

- Statistics is the study of how best to collect, analyze, and draw conclusions from data
- Steps of working with data process:
  - Identify a problem
  - Collect relevant data
  - Analyze the data
  - Form a conclusion

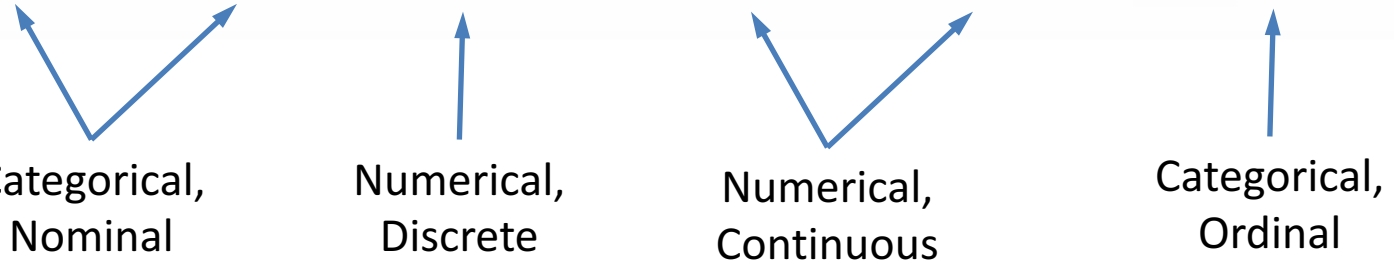
Types of Data

- Numerical (Quantitative)
  - Discrete:
    - Measured quantities
    - Results of experiments
    - Numerical values obtained by counting
  - Continuous:
    - Value obtained by measuring (e.g. height of all students)
    - All values in a given interval of numbers (e.g. federal spending)
- Categorical (Qualitative)
  - Ordinal:
    - Natural ordering (e.g. "hot", "medium", "cold")
  - Nominal:
    - Any categorical data that doesn't have an order (e.g. "blue", "red", "green)
- Other e.g. Text, Video, Binary

Data Types Example

county dataset – summarizes economics and demographic information from 3,143 counties in the United States

	name	state	pop2010	fed_spend	poverty	smoking_ban
1	Autauga	AL	54571	6.068	10.6	none
2	Baldwin	AL	182265	6.140	12.2	none
3	Barbour	AL	27457	8.752	25.0	none
⋮	⋮	⋮	⋮	⋮	⋮	⋮
3143	Weston	WY	7208	6.695	7.9	none



Relationship Between Data

- Independent variables
- Associated, or dependent:
  - Positive association
  - Negative association

DATA COLLECTION AND SAMPLING STRATEGIES

Variable Types

- **Response** variable – an observed variable
- **Explanatory** – variable or a set of variables that can influence the response variable and/or explain changes in response variable

*Is **federal spending**, on average, higher or lower in counties with high **rates of poverty**?*

Association does not imply causation

Data Collection Types

- Observational Studies:
  - Study does not directly interfere with how the data arises:
    - Surveys
    - Review medical or company records
  - Can uncover naturally occurring associations
- Experiments:
  - Help investigate a possibility of a causal connection

Observational Studies

Example: Observational study to track sunscreen use and if it is related to skin cancer.

Observed fact: the more sunscreen someone used, the more likely the person was to have skin cancer.

Question: Does this mean sunscreen **causes** skin cancer?

Confounding Variables

- **Confounding** variable (aka. confounding factor, lurking variable, confounder) – variable correlated with both the explanatory and response variables

For sunscreen example – *Sun Exposure* is a confounding variable

Observational Study or an Experiment?

- In the John Rouser video played earlier, what kind of study was used to determine if beer consumption increases human attractiveness to malaria mosquitoes: observational or experimental?

Forms of Observational Studies

- **Prospective** study – observing as events unfold
- **Retrospective** study – review data for past events

Population and Sample

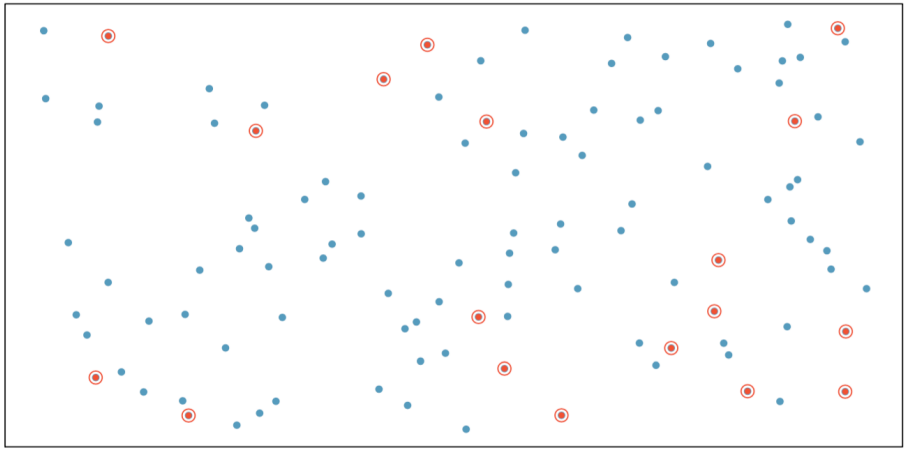
- **Population** is a collection of people, items, or events and includes all members of a defined group that we are studying or collecting information on for data driven decisions.
- **Sample** – a subset, a small fraction of a population.

Sampling Principles

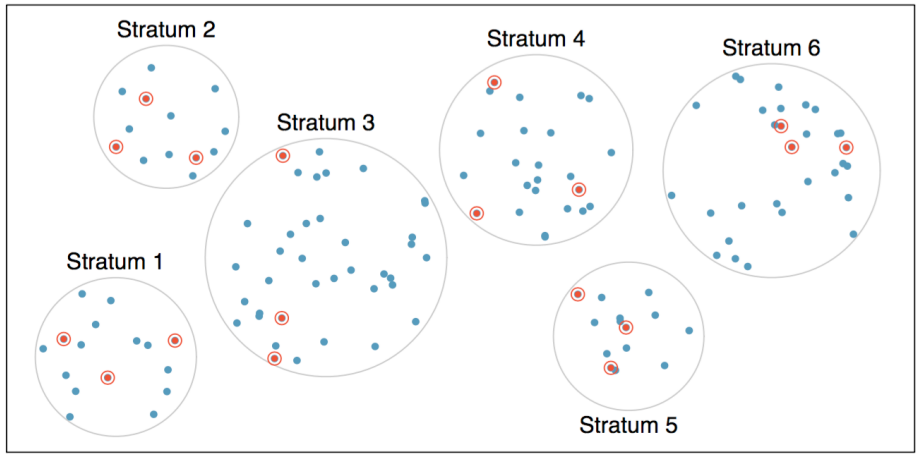
- **Representative** sample
- **Randomly** selected
- Caution if the **non-response is high** (e.g. if only 30% of the people selected respond to the survey)
- Caution if including **convenience sample** (e.g. base the analysis only on online reviews of the product)

Sampling Methods

- Simple **random** sampling – each case in the population has an equal chance of being included in the final sample



- **Stratified** sampling – random sampling from groups of similar cases



Other Types of Sampling

When there is a lot of case-to-case variability within a cluster but the clusters are very similar

- **Cluster** sample – break up the population in groups (clusters), sample fixed number of clusters and include all observations from each of those clusters in a sample.
- **Multistage** sample – similar to cluster sample, but instead of including all observations from each selected cluster, collect random sample within each selected cluster

Designing Experiments

Four Principles:

- **Controlling** the differences in the group
- **Randomization** to account for variables that the researches cannot control
- **Replication** by collecting a sufficiently large sample
- **Blocking** is used when the researches know that other variables influence the response

BIAS

Definition of Bias

Bias – Intentional or unintentional favouring of one group or outcome over other potential groups or outcomes in the population

Two main categories of biases:

- Selection bias
- Response bias

Selection Bias

The bias that results from an **unrepresentative sample** is called **selection bias:**

- **Undercoverage** - occurs when some members of the population are inadequately represented in the sample
- **Nonresponse** bias - bias that results when respondents differ in meaningful ways from nonrespondents
- **Voluntary** bias - sample members are self-selected volunteers

Response Bias

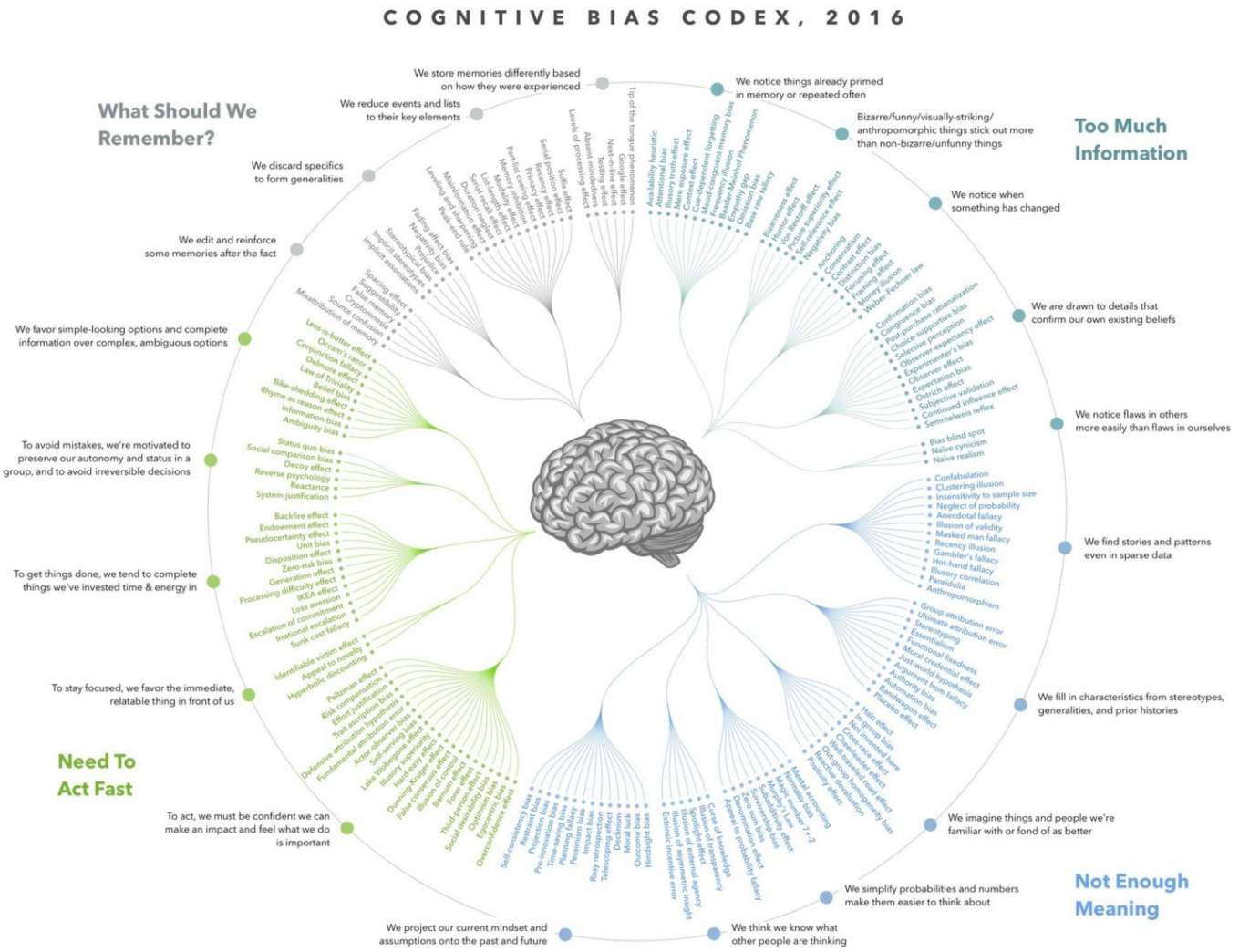
Response bias refers to the bias that results from **problems in the measurement process**:

- **Leading questions** – questions that encourage the expected answer
- **Social desirability** – responses may be biased toward what the respondents believe is socially desirable

Reducing Bias in Human Experiments

- Split participants into two groups:
  - Treatment group
  - Control group
- The study is ideally double-blind – researchers who interact with participants and participants are unaware what group the participant they are interacting with belongs to

Cognitive Bias



https://cdn-images-1.medium.com/max/2000/1*71TzKnr7bzXU_l_pU6DCNA.jpeg

Other Types of Biases

- List of cognitive biases used in the previous image - https://en.wikipedia.org/wiki/List_of_cognitive_biases
- List of types of statistical biases - https://en.wikipedia.org/wiki/Bias_(statistics)

REVIEW OF THE BASICS

Mean and Median

- **Mean** – the average of the numbers
- **Median** – "middle" of a sorted list:
  - Value in the “middle”
  - 50% of the points below; 50% above
  - Average the two middle values if there is an even number of points

Mean and Median Example

Below is a subset of data used by John Rauser in the video.

Number of mosquitos attracted to people drinking water:

21 22 15 12 21 16

Mean: $\bar{X}$ = (21+22+15+12+21+16) / 6 = 17.83

Median:

- First, - reorder:

12 15 16 21 21 22

- Find an average of 2 numbers in the middle:

(16+21) / 2 = 18.5

Variance and Standard Deviation

- Variance
  - A measure of the variability of the data
  - Roughly the average squared distance from the mean
- Standard Deviation:
  - How far away the observation is from the mean. The distance is called *deviation*
  - Standard Deviation is the square root of variance
  - 95% of the points are usually within 2 standard deviations of the mean

Variance and Standard Deviation: Example

- Same dataset as above:

21 22 15 12 21 16

Mean: $\bar{X}$ = 17.83

Variance:

$S^2 = ((21-17.83)^2 + (22-17.83)^2 + \ldots + (16-17.83)^2) / \mathbf{5} = 16.57$

Standard Deviation:

$S = \sqrt{S^2} = \sqrt{16.57} = 4.07$

Questions?

Assignment #1

- Using the dataset from the video (*mosquitos_data.csv* in the course folder), write R or Python code:
  - Create side-by-side boxplots for the number of mosquitos in each group (beer vs water)
  - Answer the question: What does the graph reveal about the data for both groups? Is there an association between beer consumption and attractiveness to mosquitos?
  - Calculate basic statistics measures for each group: the mean, median, standard deviation
  - Explain the numbers
- Write the code to implement the data simulation demonstrated in the video (this maybe a bit challenging for some, hints will be given in class)

Assignment #1 (cont'd)

- Hints
  - Use pandas, matplotlib or seaborn to do the boxplot in Python, or if using R, either R's built-in boxplot or geom_boxplot from ggplot
  - Properly label the graphs (title, legend, units if applicable)
- Upload your homework to the portal before next class
  - Upload a pdf version of your Jupyter notebook
  - Upload the Jupyter notebook
  - Don't forget your course #, name and assignment # in the name of the file, i.e. 3251-3_LastName_FirstName_A1

Resources

- Statistics does not have to be so hard: simulate! - http://blog.revolutionanalytics.com/2014/10/statistics-doesnt-have-to-be-that-hard.html
- Paper referenced in the video: http://blog.revolutionanalytics.com/2014/10/statistics-doesnt-have-to-be-that-hard.html
- Convicted on Statistics? https://understandinguncertainty.org/node/545

Next Class

- Probability
  - Frequentist Definition
  - Bayesian Definition
  - Conditional Probability
- In preparation:
  - Read OpenIntro Statistics Chap. 2