What is a Dataset?
Part 1: Representing Collections

INFO-1301, Quantitative Reasoning 1

University of Colorado Boulder

January 27, 2017

Prof. Michael Paul

Overview

This lecture will…

- introduce and review some terminology for describing data collections:
    - matrices, vectors, sets;

- present concepts for describing sets.

Today's material is necessary to discuss upcoming concepts (sampling and probability)

Representation of data: **matrix**

- Each row is an **observation**

Rows

Name	Gender	Age (years)	Height (cm)	# of children
John	Male	32	179.2	2
Mary	Female	49	168.5	4
Alice	Female	25	175.0	0

Columns

- Each column is a **variable**

Cells

- Each cell is a **value**

Vectors

Another term for rows and columns: **vectors**

- Each row is a vector

John	Male	32	179.2	2

- Each column is a vector

179.2
168.5
175.0

Vectors

A vector is a *list of values*

Notation:
<179.2, 168.5, 175.0>
<John, Male, 32, 179.2, 2>

The order matters!

- Not equivalent:
    - <168.5, 179.2, 175.0>
    - <179.2, 168.5, 175.0>

Matrices

A matrix consists of…

- One vector for every variable (column)
- One vector for every observation/case (row)

Refresher: Domains

Reminder from last week:
The set of values a variable can take is called the domain of the variable

A domain is defined by a **set**

- A set is a collection of values

Examples:

- Set of genders
- Set of dog breeds
- Set of integers
- Set of real numbers

Sets

A **set** is a collection of values called **elements**

Notation:
{1, 2, 3, 4, 5}
{red, blue, green}

The order doesn't matter!

- These sets are equivalent:
  - "Integers from 1 to 5"
  - {1, 2, 3, 4, 5}
  - {5, 3, 2, 1, 4}

What about ordinal values? Ordinal values are also represented as sets. The ordering might matter when it comes time to interpret the values, but it doesn't matter for describing the set of possible values.

Sets

The elements in a set are unique (no duplicates)

- {1, 2, 3, 3, 3, 4, 5} is not a valid set

The number of different elements in a set is called the **cardinality** of the set

- Also called the **order**, but we'll avoid that in this class

- Denoted with vertical lines: **| |**

- Example: {red, green, blue}
  - Cardinality = **3**

What is the cardinality of the set of integers?
The set of real numbers?

Subsets

If every element in a set is also part of another set, then the set is called a **subset**

A = {red, green, blue, yellow} |A| = 4

B = {red, blue} |B| = 2

B is a subset of A   B $\subset$ A
A is not a subset of B   A $\not\subset$ B

Empty set

A set with no elements is called the **empty set**

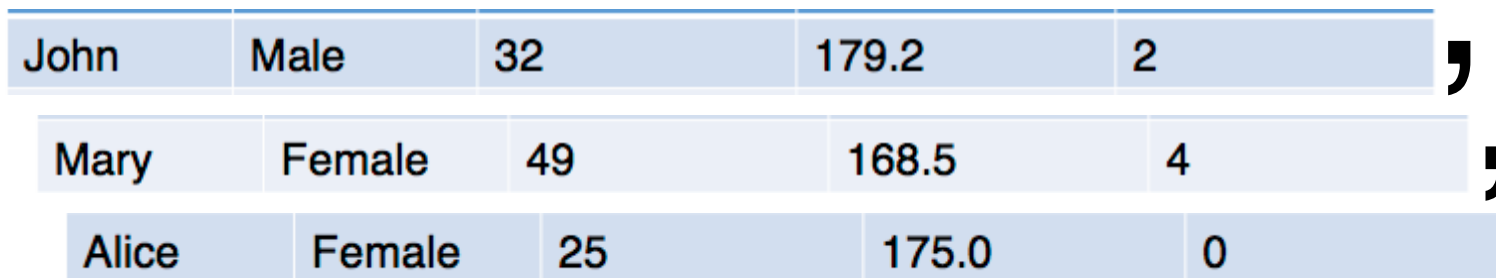- The cardinality of the empty set is 0

Notation:
{ }

Example:
Set of dinosaurs that are alive today

Where do we find sets?

- Domains of variables are sets
- Collections of observations/cases can be described as sets

Set of observations:

John	Male	32	179.2	2
Mary	Female	49	168.5	4
Alice	Female	25	175.0	0

- The elements of this set are vectors

Where do we find sets?

Matrices vs sets of observations:

- Set of people: we don't care about the order of John, Mary, and Alice

- Data matrix: we have organized the people into rows with a specified order
  - Often we don't really care about the order, but we need to decide on an order anyway so that we can refer to each row by its number (e.g., "row 15").

Confusingly, a **dataset** (or **data set**) usually refers to a data matrix, which is not a set

Visualizing sets

Sets and relationships between sets can be visualized with **Venn diagrams**
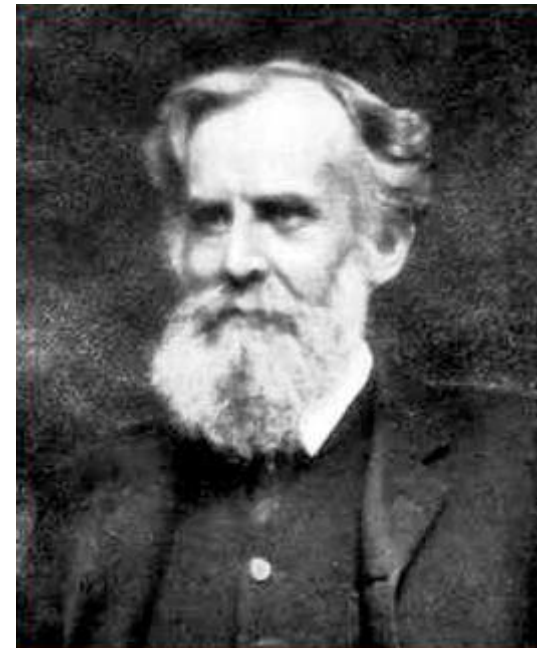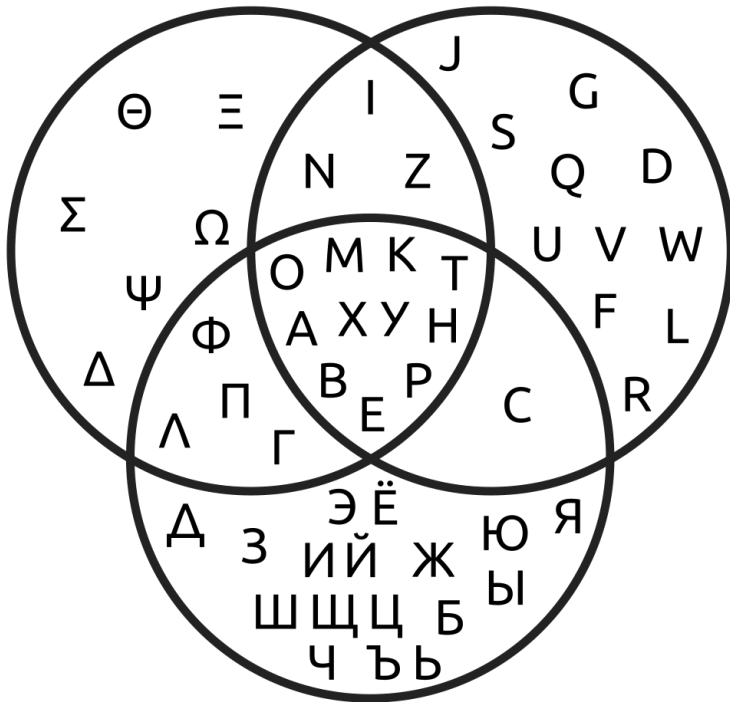
John Venn, 1834-1923

Set operations

How do we describe the relationship between sets? How do we modify sets?

In arithmetic, we have **operations** such as *addition* and *multiplication*

- What kind of operations exist for sets?
  - *Union*
  - *Intersection*
  - *Complement*
  - *Difference*