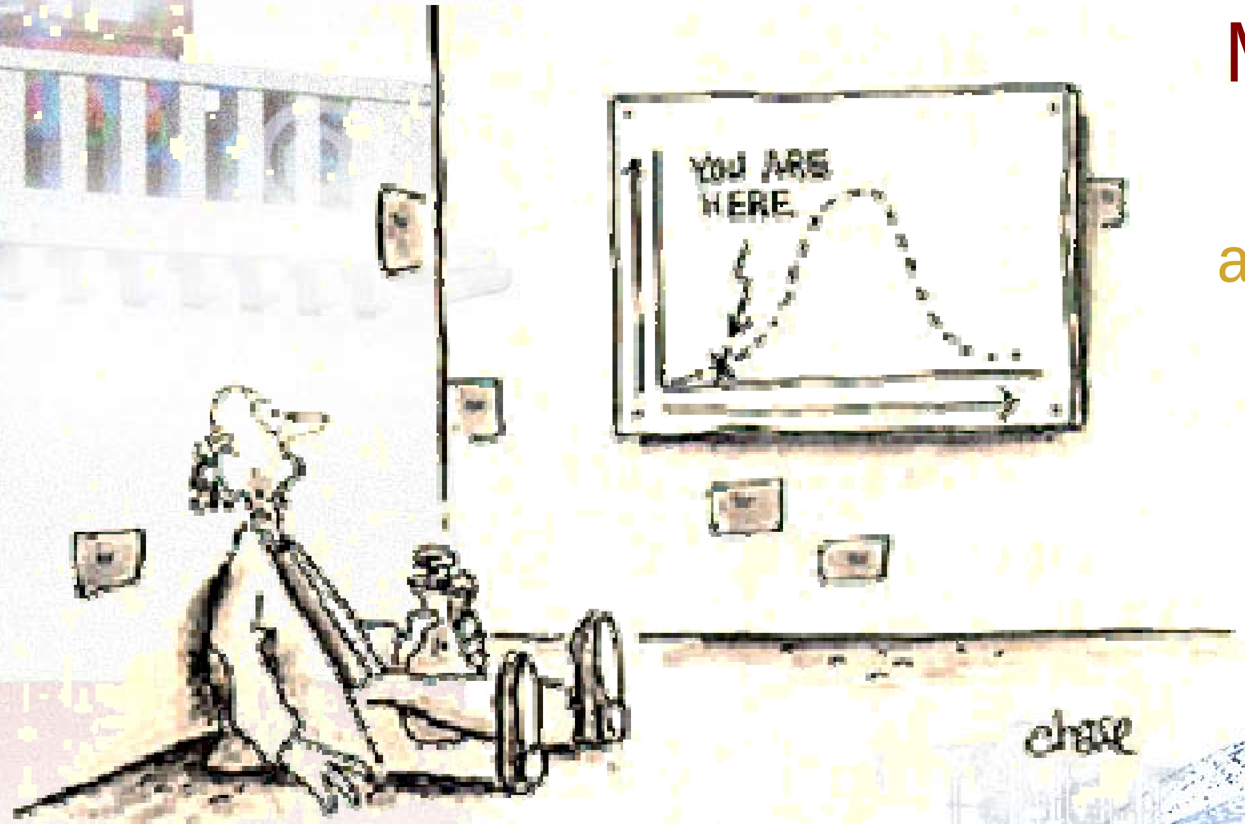
# Introduction to Informatics

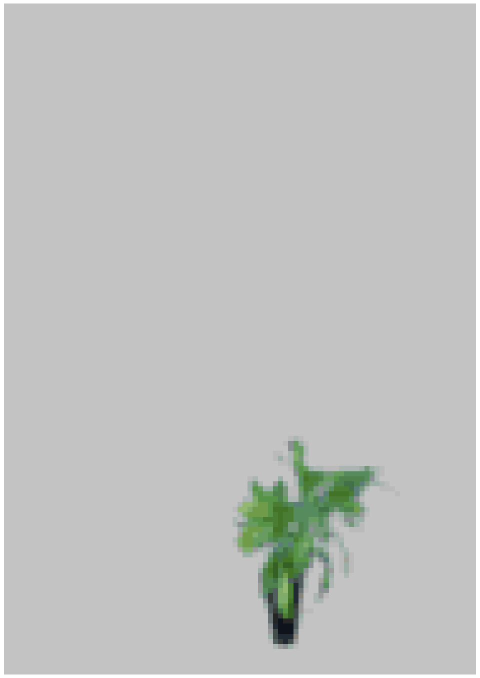
## Lecture 15: Sets and Inductive Model Building




### Measuring the World and Summarizing Data

Luis M.Rocha and Santiago Schnell

# Readings until now

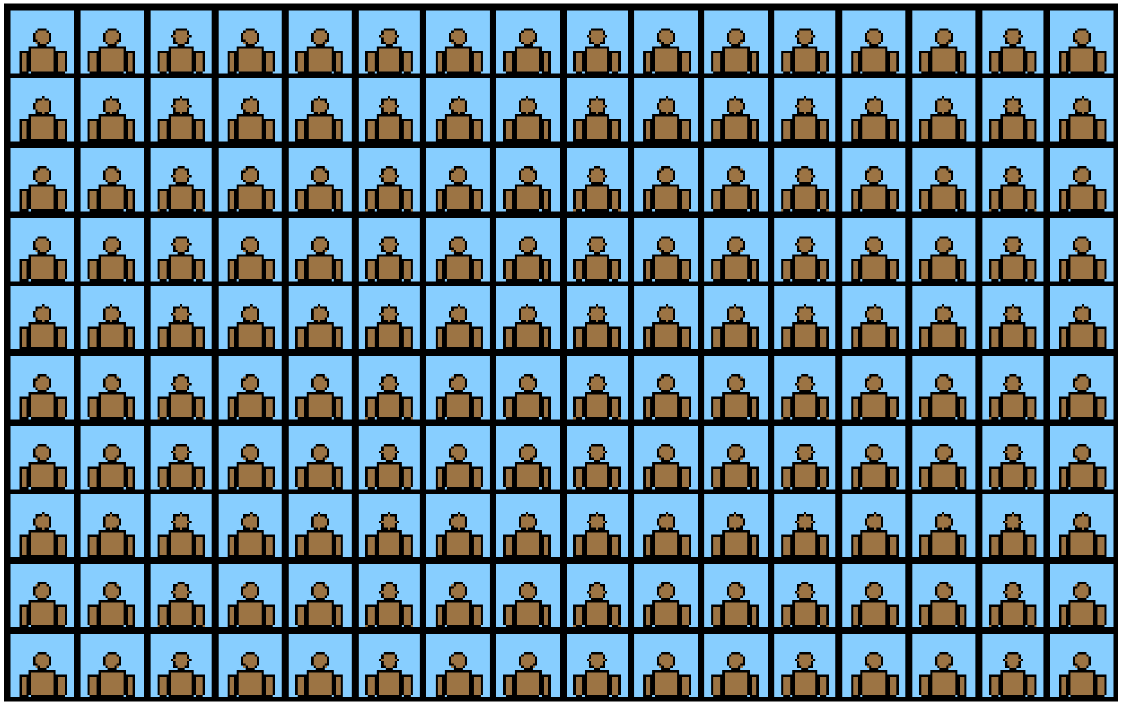


- Lecture notes
  - Posted online
    - http://informatics.indiana.edu/rocha/i101
      - *The Nature of Information*
      - *Technology*
      - *Modeling the World*
  - @ *infoport*
    - *http://infoport.blogspot.com*
  - From course package
    - Von Baeyer, H.C. [2004]. *Information: The New Language of Science*. Harvard University Press.
      - Chapters 1, 4 (pages 1-12)
    - From Andy Clark's book "*Natural-Born Cyborgs*"
      - Chapters 2 and 6 (pages 19 - 67)
    - From Irv Englander's book "*The Architecture of Computer Hardware and Systems Software*"
      - Chapter 3: Data Formats (pp. 70-86)
    - Klir, J.G., U. St. Clair, and B.Yuan [1997]. Fuzzy Set Theory: foundations and Applications. Prentice Hall
      - Chapter 2: Classical Logic (pp. 87-97)
      - Chapter 3: Classical Set Theory (pp. 98-103)

# Exam Schedule

- 11595 (T/R Class)
  - Midterm
    - March 1st (Thursday)
      - Being Graded
  - Final Exam
    - Thursday, May 3rd, 7:15-9:15 p.m.

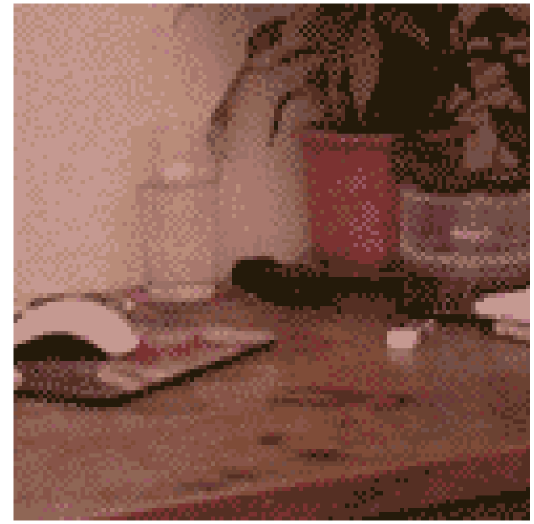
# NO LAB THIS WEEK !!!



# Assignment Situation

- Labs
  - Past
    - Lab 1: Blogs
      - Closed (Friday, January 19): Grades Posted
    - Lab 2: Basic HTML
      - Closed (Wednesday, January 31): Grades Posted
    - Lab 3: Advanced HTML: Cascading Style Sheets
      - Closed (Friday, February 2): Grades Posted
    - Lab 4: More HTML and CSS
      - Closed (Friday, February 9): Grades Posted
    - Lab 5: Introduction to Operating Systems: Unix
      - Closed (Friday, February 16): Grades Posted
    - Lab 6: More Unix and FTP
      - Closed (Friday, February 23): Grades Posted
    - Lab 7: Logic Gates
      - Closed: due Friday, March 9
  - Next: Lab 8
    - Intro to Statistical Analysis using Excel
      - March 22 & 23, Due Friday, March 30

**Get a Group NOW!**



- Assignments
  - Individual
    - First installment
      - Closed: February 9: Grades Posted
    - Second Installment
      - Past: March 2, Being Graded
    - Third installment
      - Presented on March 8th, Due on March 30th
  - Group
    - First Installment
      - Presented: March 6th, Due March 9th

# Classical Set Theory

- Propositional logic helps us make distinctions.
  - True and False, tautologies, contradictions
- Classical set theory is another form of representing the same kind of distinctions
  - Between and among groups that we perceive to share a characteristic or property.

# Definition of set

A *set* is an unordered collection of elements.

Some examples:

{1, 2, 3} is the set containing "1" and "2" and "3."

{1, 1, 2, 3, 3} = {1, 2, 3} since repetition is irrelevant.

{1, 2, 3} = {3, 2, 1} since sets are unordered.

{1, 2, 3, …} is a way we denote an infinite set (in this case, the natural numbers).

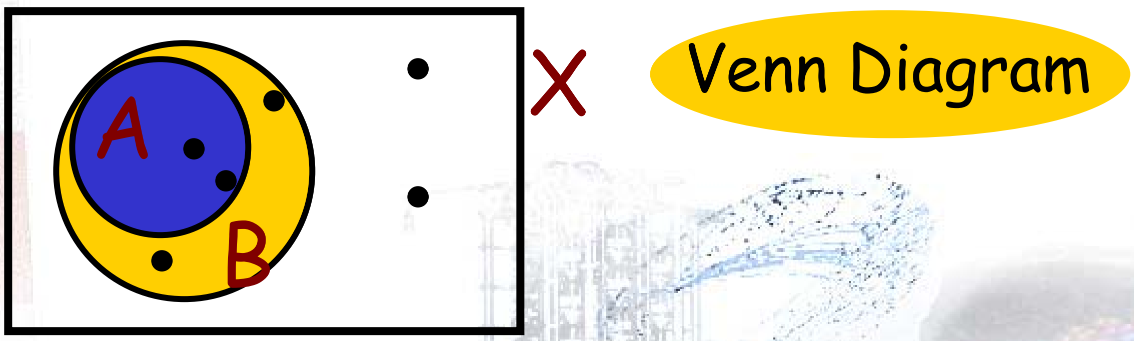
∅ = {} is the empty set, or the set containing no elements.

# Notations

x $\in$ S means "x is an *element* of set S."
x $\notin$ S means "x is *not an element* of set S."

A $\subseteq$ B means "A is a *subset* of B." (**inclusion**)

or, "B contains A."
or, "every element of A is also in B."
or, $\forall x\ ((x \in A) \Rightarrow (x \in B))$.



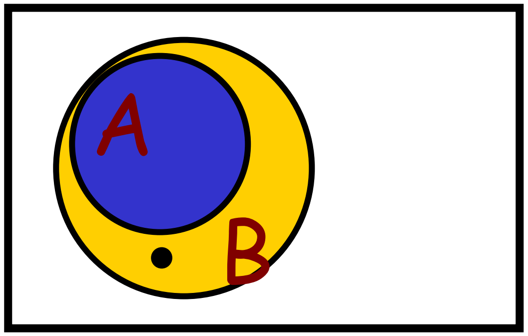

Adapted from C. Heeren

# Notations

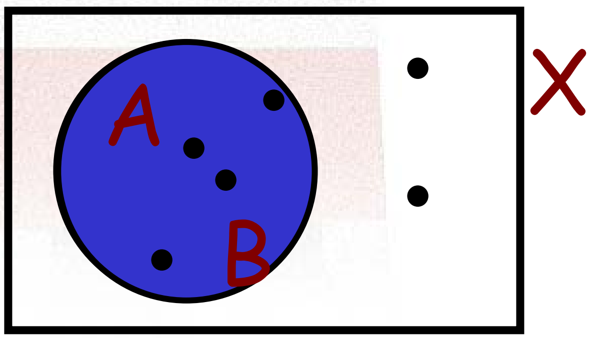
$A \subseteq B$ means "A is a *subset* of B."
$B \supseteq A$ means "B is a *superset* of A."

$A \subset B$ means "A is a *proper subset* of B."

- $A \subseteq B$, and $A \neq B$.




$A = B$ if and only if A and B have exactly the same elements.




iff, $A \subseteq B$ and $B \subseteq A$
iff, $\forall x\ ((x \in A) \Leftrightarrow (x \in B))$.

Adapted from C. Heeren

# Examples

Quick examples:

- $\{1,2,3\} \subseteq \{1,2,3,4,5\}$
- $\{1,2,3\} \subset \{1,2,3,4,5\}$

Is $\emptyset \subseteq \{1,2,3\}$ ?

Yes! $\forall x\ (x \in \emptyset) \Rightarrow (x \in \{1,2,3\})$ holds, because $(x \in \emptyset)$ is false.

IMPLICATION

| $A,p$ | $B,q$ | $A \Rightarrow B$, $p \Rightarrow q$ |
|---|---|---|
| 0 | 0 | 1 |
| 0 | 1 | 1 |
| 1 | 0 | 0 |
| 1 | 1 | 1 |

Adapted from C. Heeren

# Defining sets

- Explicitly
  - {John, Paul, George, Ringo}
- Implicitly
  - {1,2,3,…}, or {2,3,5,7,11,13,17,…}
- Set builder
  - { x | x is prime }, { x | x is odd}, {x | x ∈ Fibonacci sequence}.
    - In general { x | P(x) is true }, where P(x) is some description of the set.

Note that the symbol | reads as "such that"

# Set Operations

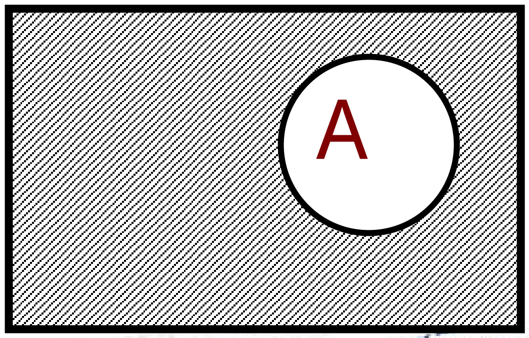
The *complement* of a set A is:

$\bar{A}$ = { x : x ∉ A}

If A = {x : x is bored}, then

$\bar{A}$ = {x : x is **not** bored} = ∅

X




$\bar{\emptyset}$ = X
and
$\bar{X}$ = ∅

NOT

| $A, p$ | $X = \bar{A}$, $\neg p$ |
|---|---|
| 0 | 1 |
| 1 | 0 |

Adapted from C. Heeren

# Set Operations

The *union* of two sets A and B is:

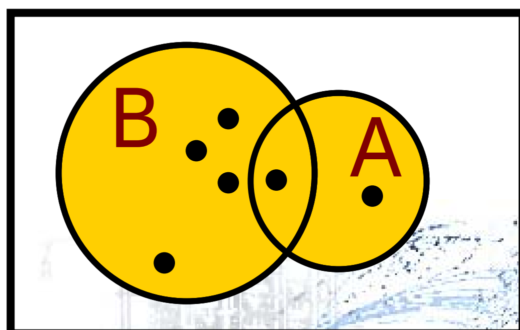
$$A \cup B = \{ x : x \in A \lor x \in B\}$$

If A = {Charlie, Lucy, Linus},
and B = {Lucy, Desi}, then

OR

A ∪ B = {Charlie, Lucy, Linus, Desi}

| *A,p* | *B,q* | *A+B, p∨q* |
|---|---|---|
| 0 | 0 | 0 |
| 0 | 1 | 1 |
| 1 | 0 | 1 |
| 1 | 1 | 1 |




Adapted from C. Heeren

# Set Operations

The *intersection* of two sets A and B is:

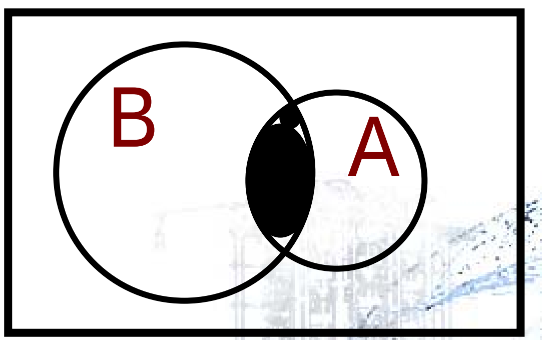
$$A \cap B = \{ x : x \in A \wedge x \in B\}$$

If A = {x : x is a US president},
and B = {x : x is deceased}, then

A ∩ B = {x : x is a deceased US president}

AND

| $A,p$ | $B,q$ | $A \bullet B,\ p \wedge q$ |
|---|---|---|
| 0 | 0 | 0 |
| 0 | 1 | 0 |
| 1 | 0 | 0 |
| 1 | 1 | 1 |




- B: Set of Funny People
- A: Set of Clowns

Adapted from C. Heeren

# Set Operations

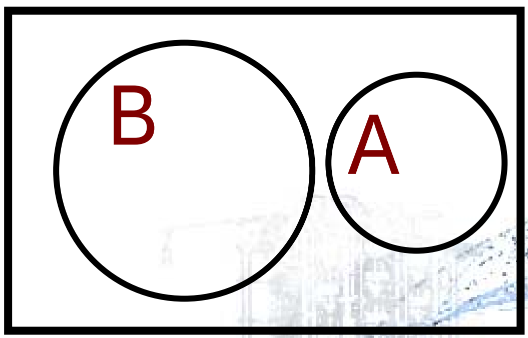
The *intersection* of two sets A and B is:

$$A \cap B = \{ x : x \in A \wedge x \in B\}$$

If $A = \{x : x \text{ is a US president}\}$, and
$B = \{x : x \text{ is in this room}\}$, then

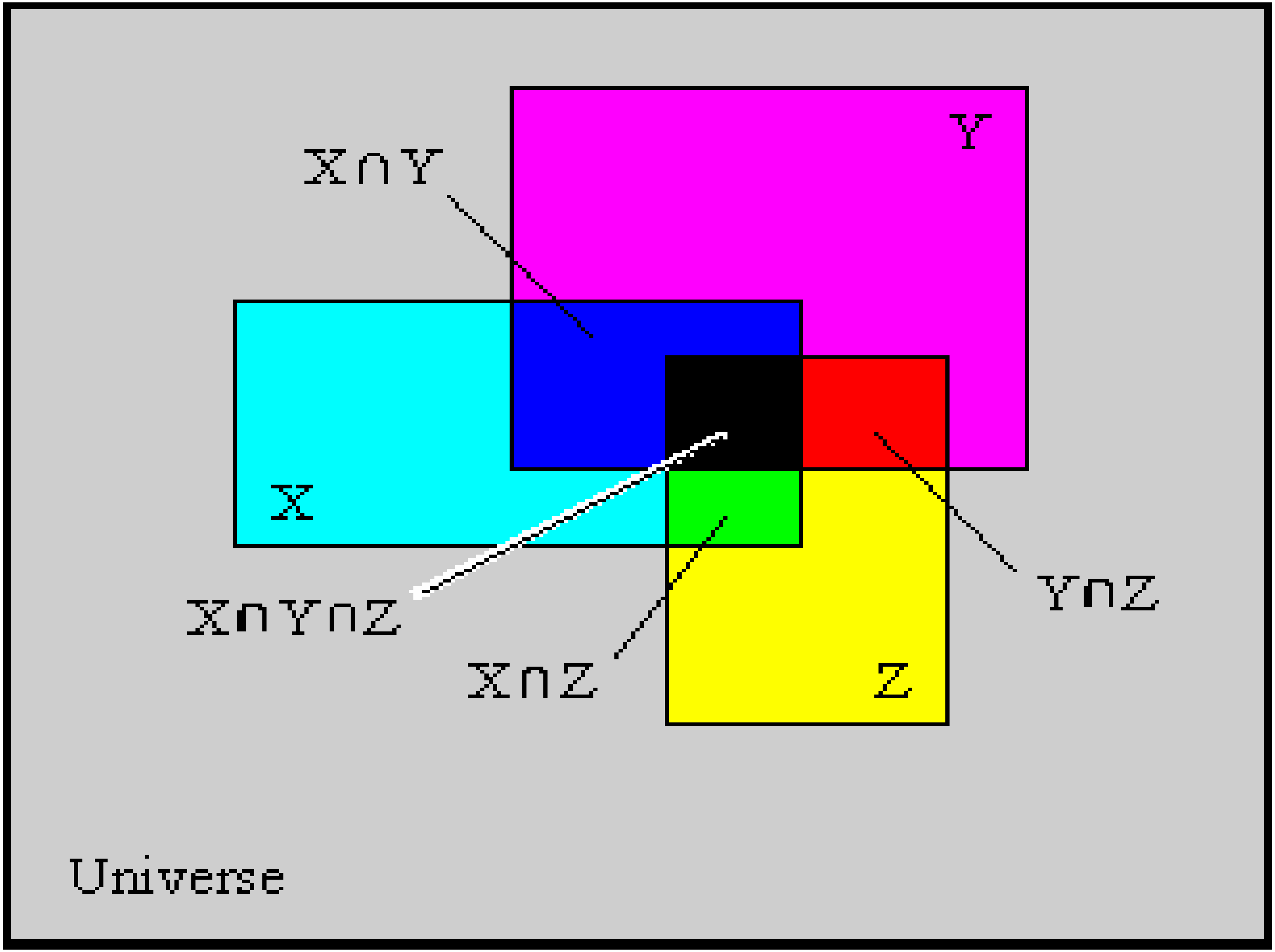
$A \cap B = \{x : x \text{ is a US president in this room}\} = \varnothing$

Sets whose intersection is empty are called *disjoint* sets

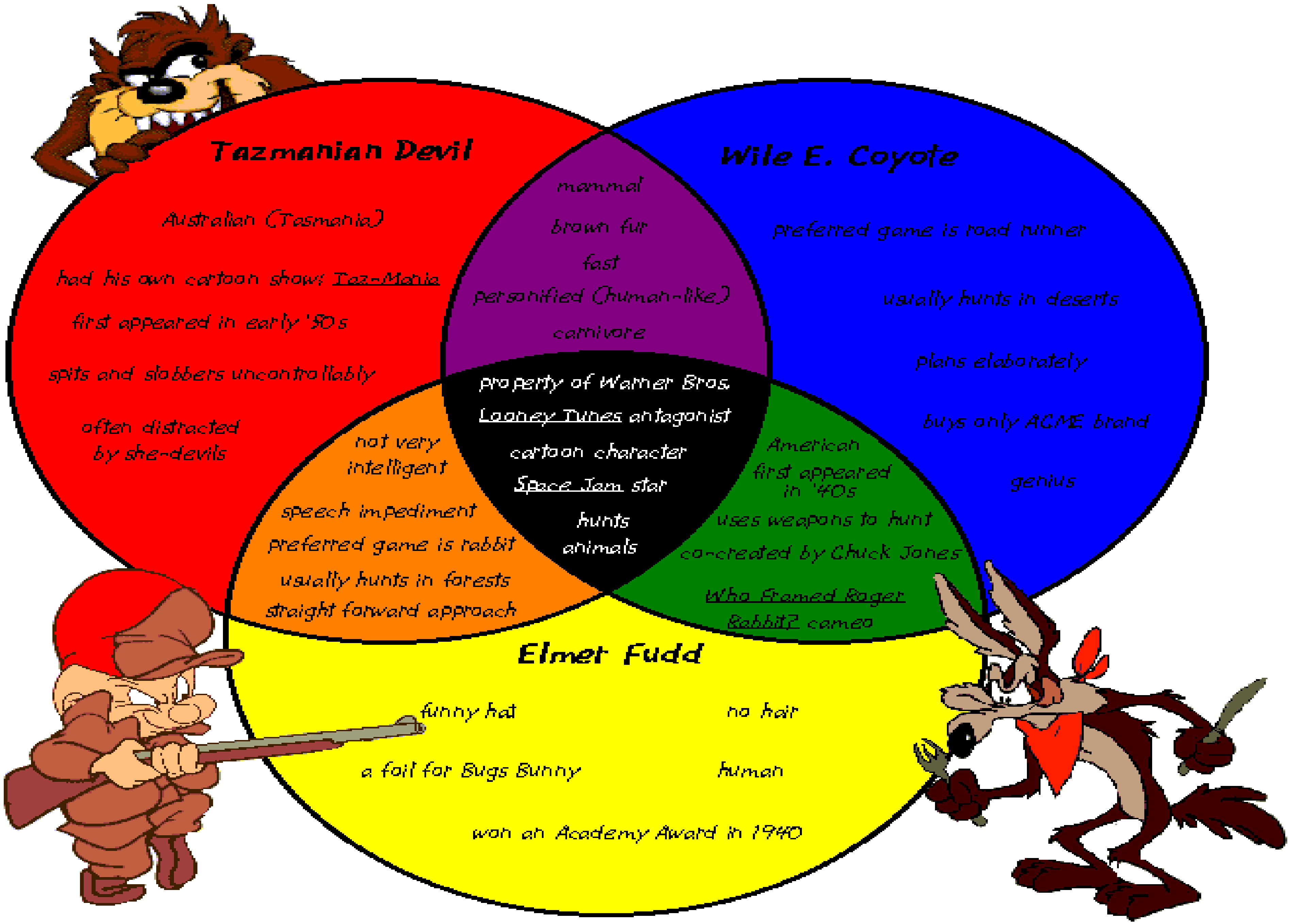



- B: Movies That Suck
- A: I101 Movies

Adapted from C. Heeren

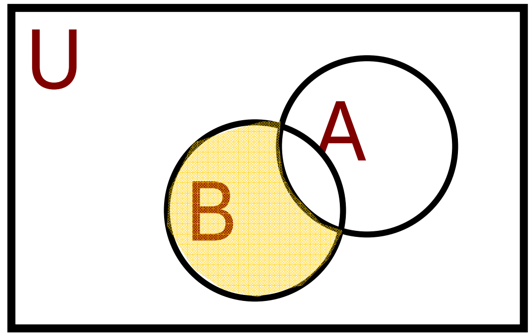

X∩Y
Y
X
X∩Y∩Z
Y∩Z
X∩Z
Z
Universe

Qualities of the Tazmanian Devil, Wile E. Coyote, and Elmer Fudd



# Set Operations

The *set difference*, B - A, is:




$$B - A = \{ x : x \in B \wedge x \notin A \}$$

$$B - A = \{ x : x \in A \wedge x \in \bar{A} \}$$

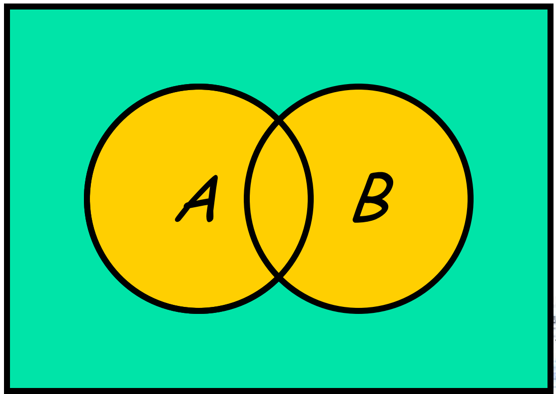
# De Morgan's Law

- *De Morgan's Law I*

  $$\overline{(A \cup B)} = \overline{A} \cap \overline{B}$$

- *De Morgan's Law II*

  $$\overline{(A \cap B)} = \overline{A} \cup \overline{B}$$




Adapted from C. Heeren

# More about Mathematics

- I201: Mathematical Foundations of Informatics
  - Steve Myers
  - An introduction to the suite of mathematical and logical tools used in information sciences.
    - finite mathematics, automata and computability theory, elementary probability, and statistics and basics of classical information theory
    - Cross listed with COGS Q250. Credit given for either INFO I201 or COGS Q250
  - Prerequisite: INFO I101, MATH M118.

# Syllabus

- *Introduction to Informatics*
- *Modeling and Problem Solving*
- *Data and Knowledge Representation*
- *Deductive Model Building*
- Inductive Model Building
- Information and Uncertainty
- Computing Models: Algorithms
- Information Technology in the Real World

# Deduction vs. Induction

- Propositional Logic is used to study ***inferences***
  - How conclusions can be reached from premises
- Deductive Inference ← **Logic**
  - If the premises are true, we have absolute ***certainty*** of the conclusion
    - February has 29 days only in leap years
    - Today is February 29th
    - This year is a leap year
- Inductive Inference ← **Uncertainty**
  - Conclusion supported by ***good evidence*** (significant number of examples/observations) but not full certainty -- ***likelihood***
    - Ran BlackBox for 1000 cycles, "dead box" observed
    - Ran BlackBox for 1000 cycles, "dead box" observed
    - Ran BlackBox for 1000 cycles, "dead box" observed
    - .................................................
    - Ran BlackBox for 1000 cycles, "dead box" observed
    - "Dead Box" always appears after 1000 cycles

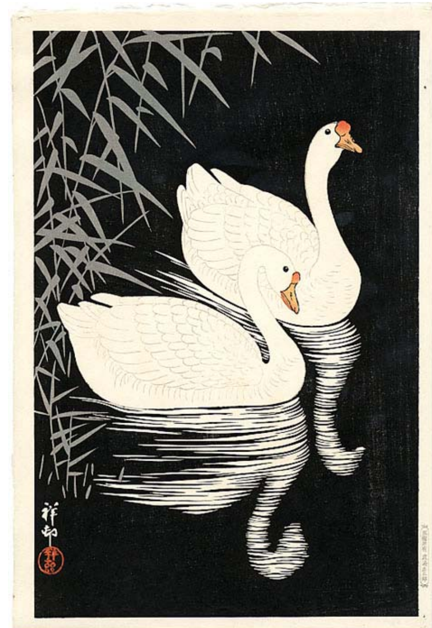
# Inductive Reasoning

- Induction
  - *The process of inferring a general law or principle from the observation of particular instances* (OED)
- A process of generalizing
  - We start from instances of an event or phenomenon, and generalize them to formulate rules
    - Apples fall from trees -> all objects are subjected to gravitational forces
  - Rules are *likely* to be true, given the premises
  - But the rules can be broken by new observations
    - Color of Kiwis

David Hume (1711-1776)

- Our everyday knowledge depends on patterns of repeated experience
  - Empiricism
    - “It is not reason which is the guide of life, but custom.”
    - “Custom, then, is the great guide of human life. ”
    - “A wise man proportions his belief to the evidence”
    - “The rules of morality are not the conclusion of our reason”

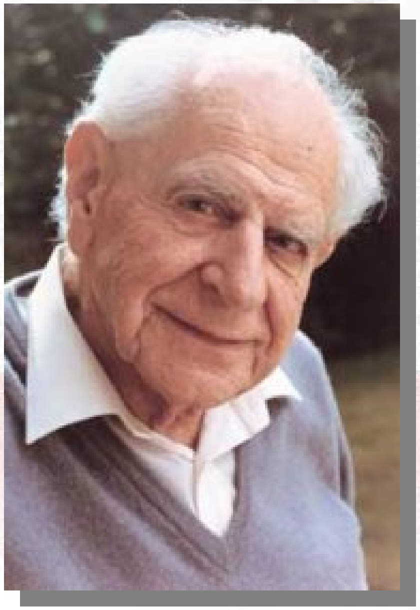
# Uncertainty in Induction



- Via Induction
  - Europeans could have thought that all Swans are White
    - by observing instance after instance
  - But black swans exist
    - From Australia

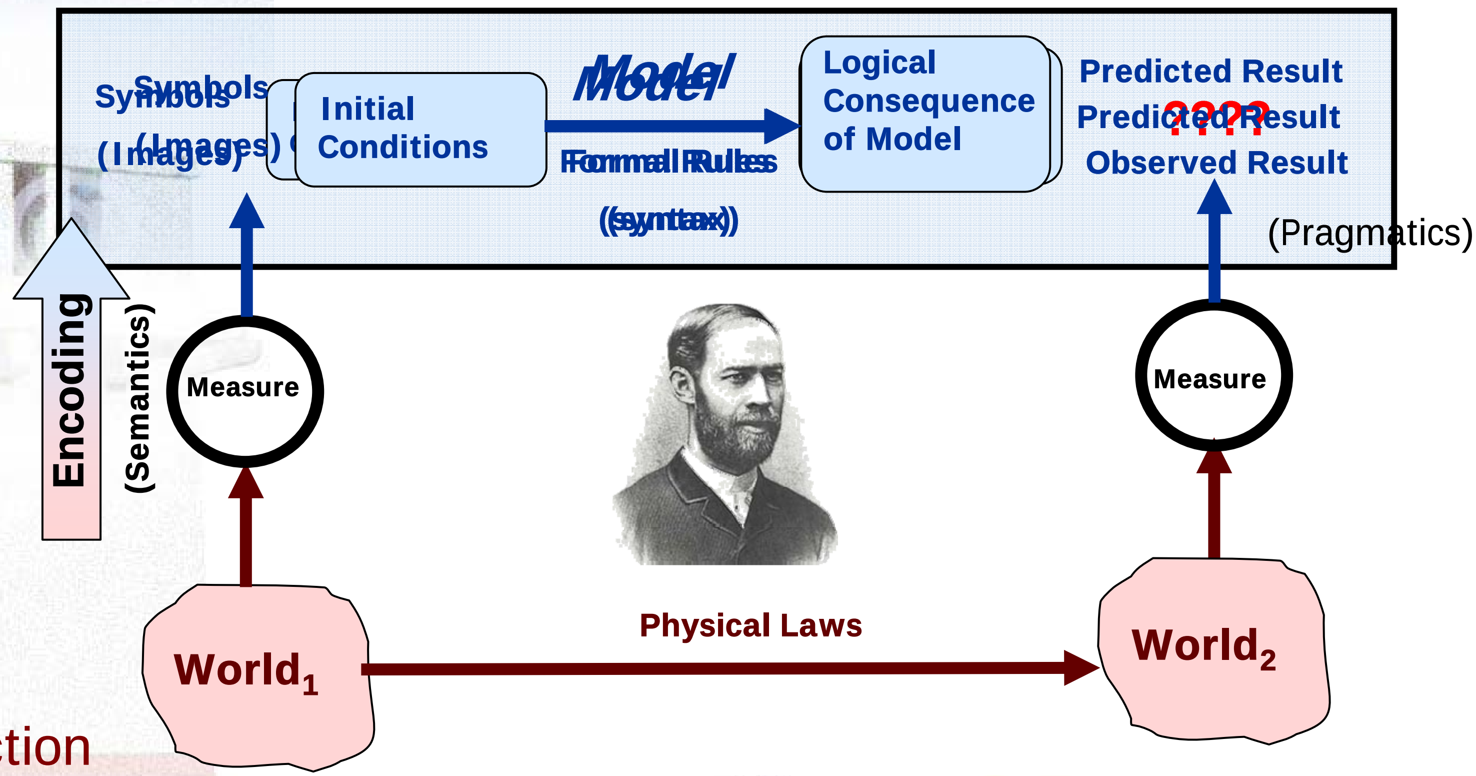
# Problems with Induction

- Karl Popper (1902-1994)
  - Used the Swan example to highlight problems with the Inductive approach
    - There could be a hypothesis that all swans are white
    - but every extra white swan that is observed does not justify the claim that all swans are white
    - Simply increases the likelihood
  - Popper warned against generalizations about the unobserved from the observed.
    - “induction is a procedure which is logically invalid and rationally unjustifiable”.
  - Proposed a deductive process of “falsification”
    - Prove the existence “in principle” of an instance that could falsify the hypothesis
    - But the hypothesis is still built by induction!
  - Also known for his views on an *Open Society*



# The Modeling Relation

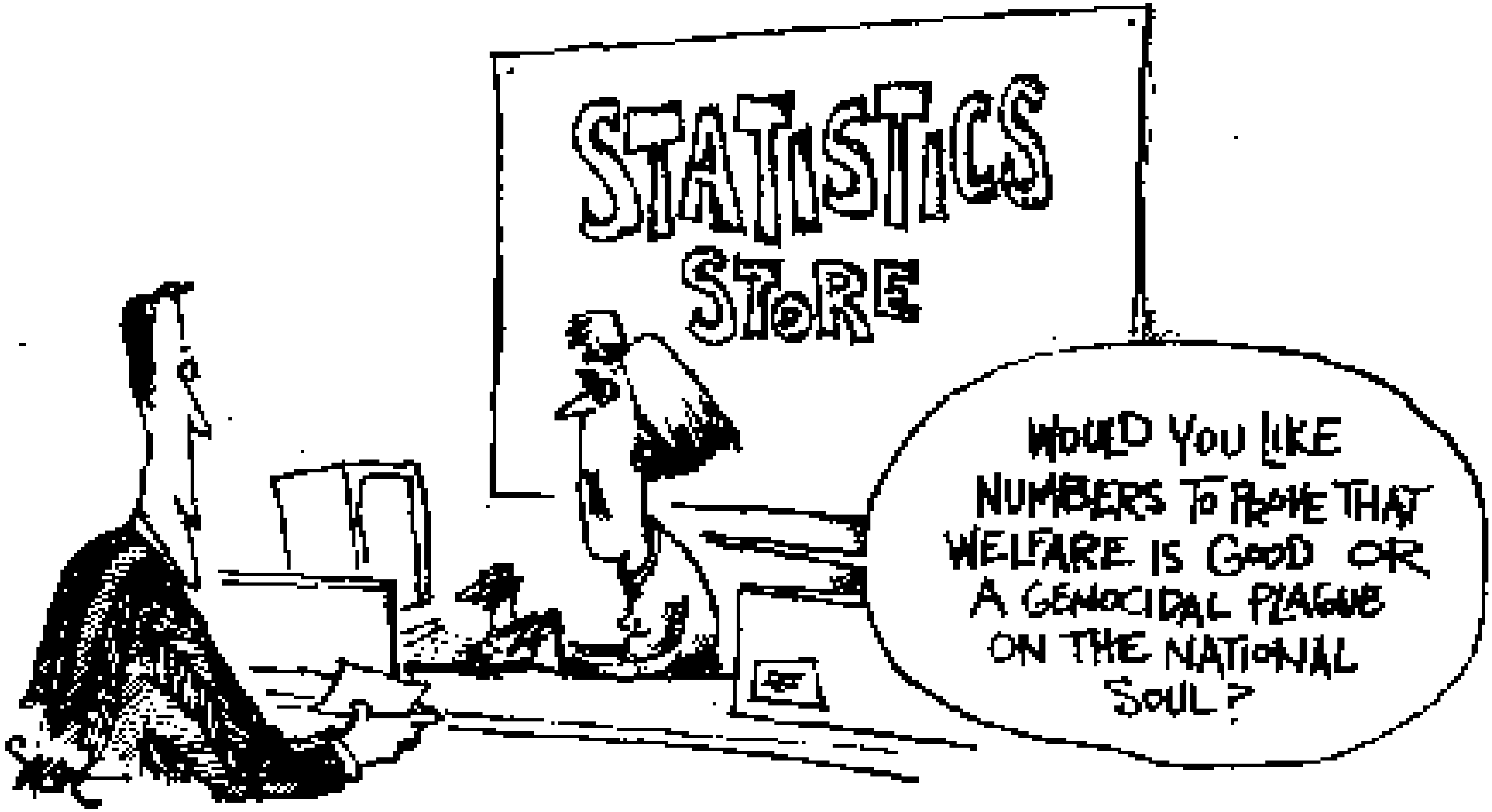
**Hertz' Modeling Paradigm**




- Induction
  - Requires attention to data collection and ***description***
- Rules from Inference
  - From Data analysis
  - Produce Conclusions

# Describing Data

- We encode our observations of the World as symbols
  - Numerical, textual, graphical
  - Data (data values)
    - The symbols without recourse to meaning
- Data Collection
  - Series of data instances
- Statistics
  - "Science of collecting, simplifying, and describing data, as well as making inferences based on the analysis of data" [Chase and Brown, "General Statistics"]
  - ***Descriptive Statistics***
    - Data collection, simplification, and characterization
  - ***Inferential Statistics***
    - Induction: Drawing Conclusions




By Signe Wilkinson, Philadelphia Daily News, Cartoonists & Writers Syndicate

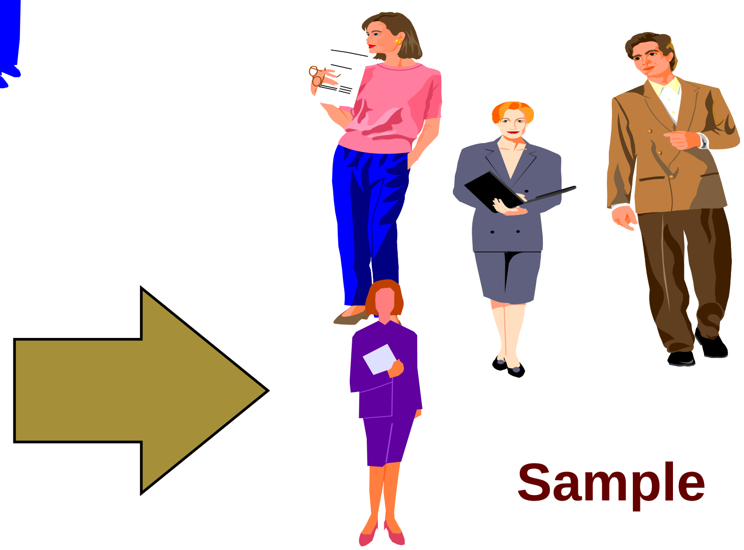
# Basic Statistics Concepts

- Population
  - The entire collection of elements we are interested in
    - Typically, elements = data values
- Sample
  - A collection of some of the elements obtained from the population
- Inferential Statistics
  - Concerned with modeling (making inferences about) a population based on the properties of a sample
- Parameter
  - A numerical property of a population
    - Average age of the US population
- Statistic
  - A numerical property of a sample
    - Average age of a selected subset of US residents
- Descriptive Statistics
  - Concerned with characterizing data from samples (their properties)
    - Organizing, describing, summarizing

# Population vs. Sample

**Population**

- *All items of interest*
- *Group of interest to investigator*



**Sample**

- *Portion of population*
- *Will be used to reach conclusions about population*

# Why Study Samples to Understand the Population?

- **Easier than studying the whole population**
- **Costs less**
- **Takes less time**
- **Sometimes testing involves risk**
- **Sometimes testing requires the destruction of the item being studied**

# Obtaining a Sample

- Random Sample
  - Sample obtained from a population such that any sample of the same size has an equal likelihood of being selected
  - Lottery Method
    - Elements are tagged, mixed up and extracted
    - http://www.dougshaw.com/sampling/

# Collecting Data

- The measurement of some quantity in a sample leads to a ***series of data values***
  - Data Array

**Raw Data: Yards Produced by 30 Carpet Looms**

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| **16.2** | **15.4** | **16.0** | **16.6** | **15.9** | **15.8** | **16.0** | **16.8** | **16.9** | **16.8** |
| **15.7** | **16.4** | **15.2** | **15.8** | **15.9** | **16.1** | **15.6** | **15.9** | **15.6** | **16.0** |
| **16.4** | **15.8** | **15.7** | **16.2** | **15.6** | **15.9** | **16.3** | **16.3** | **16.0** | **16.3** |

**Raw Data: Your favorite films**

**The Big Lebowski, Kung Fu Hustle, Team America – World Police, Kill Bill 1 + 2, Good Night, and Good Luck, Pulp Fiction,….**

# Organizing Data

**Alphabetically sorted Movies**

**12 monkeys**
**28 days later**
**a beautiful mind**
**a few good men**
**a lot like love**
**a walk to remember**
**airplane**
**ali g in da house**
**alien vs. predator**
**american history x**
**anchorman**
**anchorman**
**anchorman**
**……………………..**

1. Focus on major features
2. Data placed in ***rank*** order: ***Sorting***

   smallest to largest (or largest to smallest)
3. Data in raw form (as collected)

   24, 26, 24, 21, 27, 27, 30, 41, 32, 38
4. Data in ordered array

   21, 24, 24, 26, 27, 27, 30, 32, 38, 41

# Sorting the Data Array

- Advantages
  - Quickly notice lowest and highest values in the data
  - Easily divide data into sections
  - Easily see values that occur frequently
  - Observe variability in the data
- Disadvantage
  - Cumbersome

**Sorted Data**

| | | | | | |
|---|---|---|---|---|---|
| 15.2 | 15.7 | 15.9 | 16.0 | 16.2 | 16.4 |
| 15.4 | 15.7 | 15.9 | 16.0 | 16.3 | 16.6 |
| 15.6 | 15.8 | 15.9 | 16.0 | 16.3 | 16.8 |
| 15.6 | 15.8 | 15.9 | 16.1 | 16.3 | 16.8 |
| 15.6 | 15.8 | 16.0 | 16.2 | 16.4 | 16.9 |

**Alphabetically sorted Movies**

12 monkeys
28 days later
a beautiful mind
a few good men
a lot like love
a walk to remember
airplane
ali g in da house
alien vs. predator
american history x
anchorman
anchorman
anchorman
badboys
badboys 2
batman begins
batman begis
beer fest
beer fest
…………………..

# Summarizing Data

- Frequency
  - Number of times an item or value occurs in a collection
- Frequency Distribution
  - Given a collection of data items/values, the specification of all the distinctive values in the collection together with the number of times each of these items/values ***occurs*** in the collection
    - Table that organizes data into mutually exclusive classes
    - Shows number of observations from data set that fall into each class

[Chase and Brown, "General Statistics"]

# Frequency Distribution (values)

**Sorted Data: 30 data values (Carpet Looms)**

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 15.2 | 15.2 | 15.3 | 15.3 | 15.3 | 15.3 | 15.3 | 15.4 | 15.4 | 15.4 |
| 15.4 | 15.4 | 15.4 | 15.4 | 15.4 | 15.4 | 15.4 | 15.4 | 15.5 | 15.5 |
| 15.5 | 15.5 | 15.5 | 15.5 | 15.6 | 15.6 | 15.6 | 15.7 | 15.7 | 15.7 |

**Frequency Distribution**

| Class | Tallies | Frequency |
|---|---|---|
| 15.2 | // | 2 |
| 15.3 | ~~////~~ | 5 |
| 15.4 | ~~////~~ ~~////~~ / | 11 |
| 15.5 | ~~////~~ / | 6 |
| 15.6 | /// | 3 |
| 15.7 | /// | 3 |

# Relative Frequency Distribution (values)

| Relative Frequency Distribution | Class | Frequency (1) | Relative Freq. (1) ÷ 30 | Cumulative Relative Frequency |
|---|---|---|---|---|
| | 15.2 | 2 | 0.07 | 0.07 |
| | 15.3 | 5 | 0.16 | 0.23 |
| | 15.4 | 11 | 0.37 | 0.60 |
| | 15.5 | 6 | 0.20 | 0.80 |
| | 15.6 | 3 | 0.10 | 0.90 |
| | 15.7 | 3 | 0.10 | 1.00 |
| | | 30 | 1.00 | |

Number of values

# Freq. Distribution Film data (items)

## Raw Data: Your favorite films

**The Big Lebowski, Kung Fu Hustle, Team America – World Police, Kill Bill 1 + 2, Good Night, and Good Luck, Pulp Fiction,….**

**Sorted Movie Preferences**

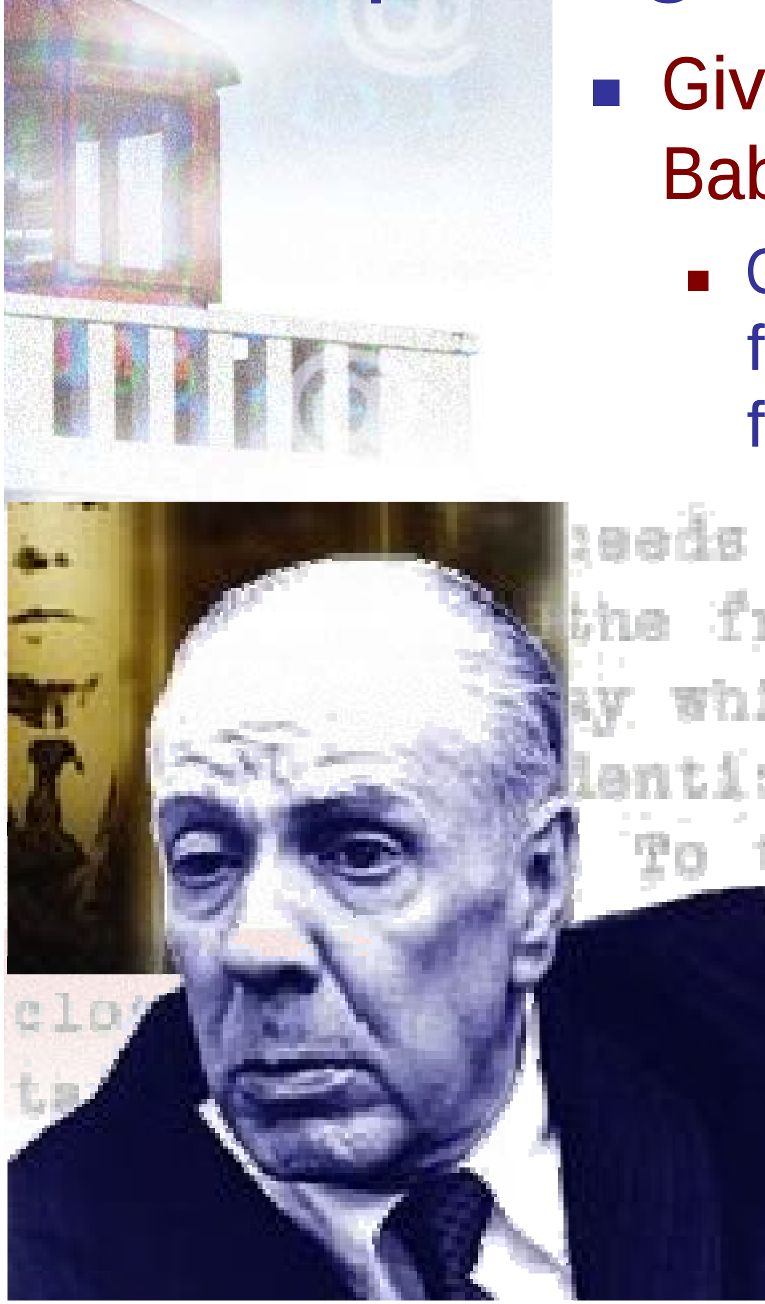
| | |
|---|---|
| dumb and dumber | 8 |
| wedding crashers | 7 |
| office space | 6 |
| the matrix | 5 |
| jackass 2 | 4 |
| old school | 4 |
| tommy boy | 4 |
| anchorman | 3 |
| mission impossible | 3 |
| scarface | 3 |
| super troopers | 3 |
| the departed | 3 |

| Class | Freq. | Rel. Freq. | % |
|---|---|---|---|
| …………………………………… | | | |
| dragon ball z | 1 | 0.004 | 0.4% |
| dream catcher | 1 | 0.004 | 0.4% |
| dumb & dumber | 8 | 0.036 | 3.6% |
| elf | 1 | 0.004 | 0.4% |
| enough | 1 | 0.004 | 0.4% |
| face off | 1 | 0.004 | 0.4% |
| Fast/furious/tokyo | 1 | 0.004 | 0.4% |
| Fear/loath/vegas | 2 | 0.009 | 0.9% |
| ………………………………………………. | | | |
| | 223 | 1.000 | 100% |

Movies

Votes (# Items)

# Group Assignment: First Installment



- Given the text of “Lottery of Babylon” by Jorge Luis Borges
  - Compute the frequency, relative frequency, and cumulative relative frequency distribution of letters
    - In the Spanish and the English Text
    - Upload to Oncourse
      - Note: in the Spanish version, lookout for ñ, á, é, í, ó, ú

# Next Class!

- Topics
  - More Inductive Reasoning Modeling
    - Measures of Central Tendency
    - Measures of Dispersion and Position
    - Probability
- Readings for Next week
  - @ *infoport*
  - From course package
    - Norman, G.R. and D.L. Streinrt [2000]. *Biostatistics: The Bare Essentials*.
      - Chapters 1-3 (pages 105-129)
- Lab 8
  - Intro to Statistical Analysis using Excel
  - NO LAB THIS WEEK!!!!