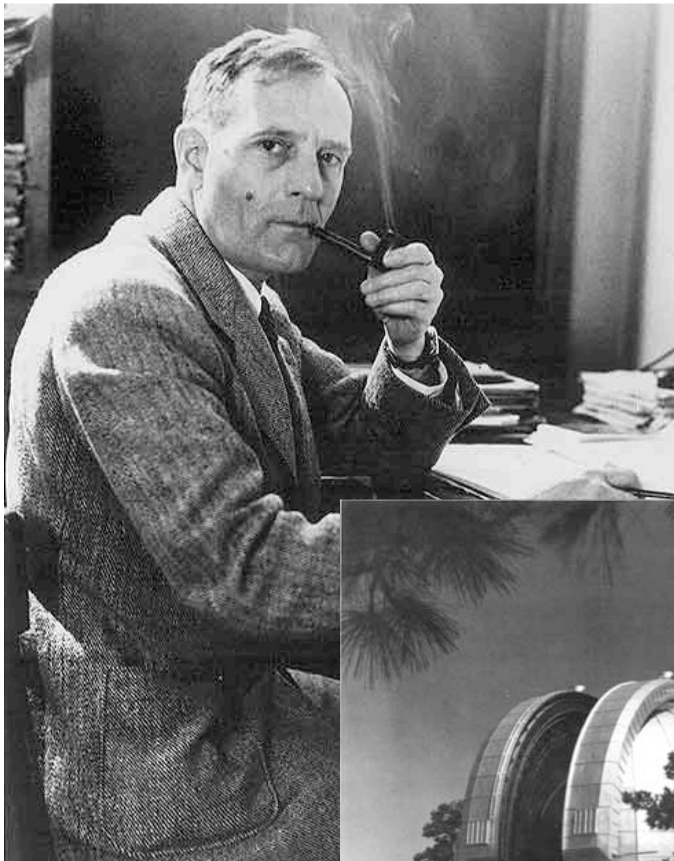
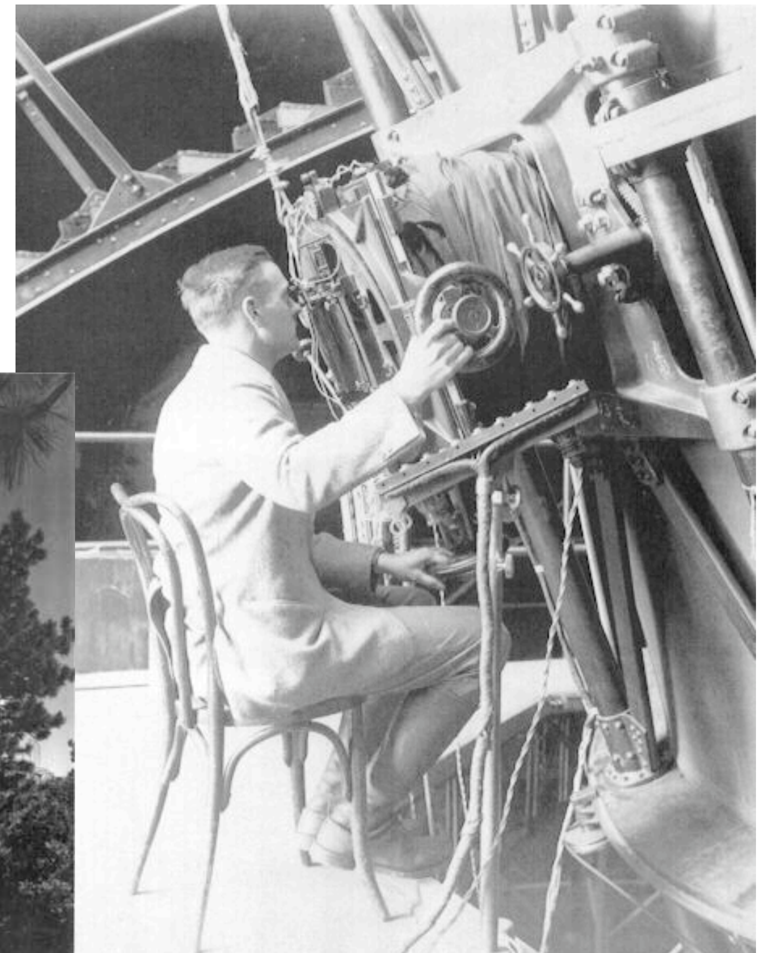
# Exploratory Data Analysis

4 March 2009

Research Methods for
Empirical Computer Science
CMPSCI 691DD

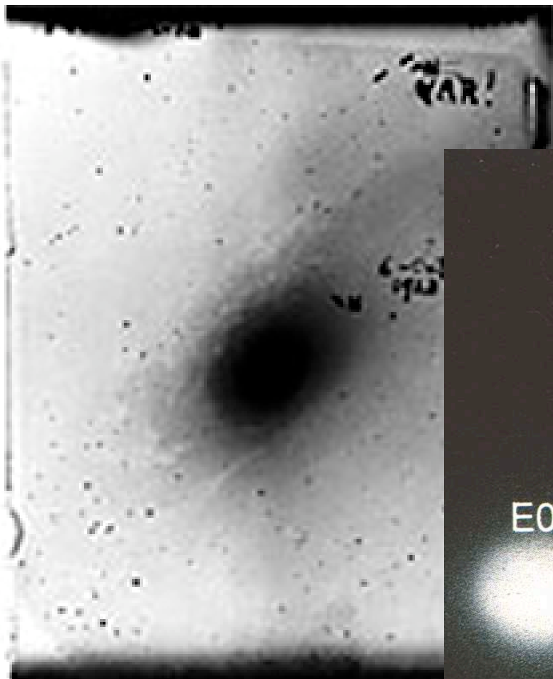
# Edwin Hubble

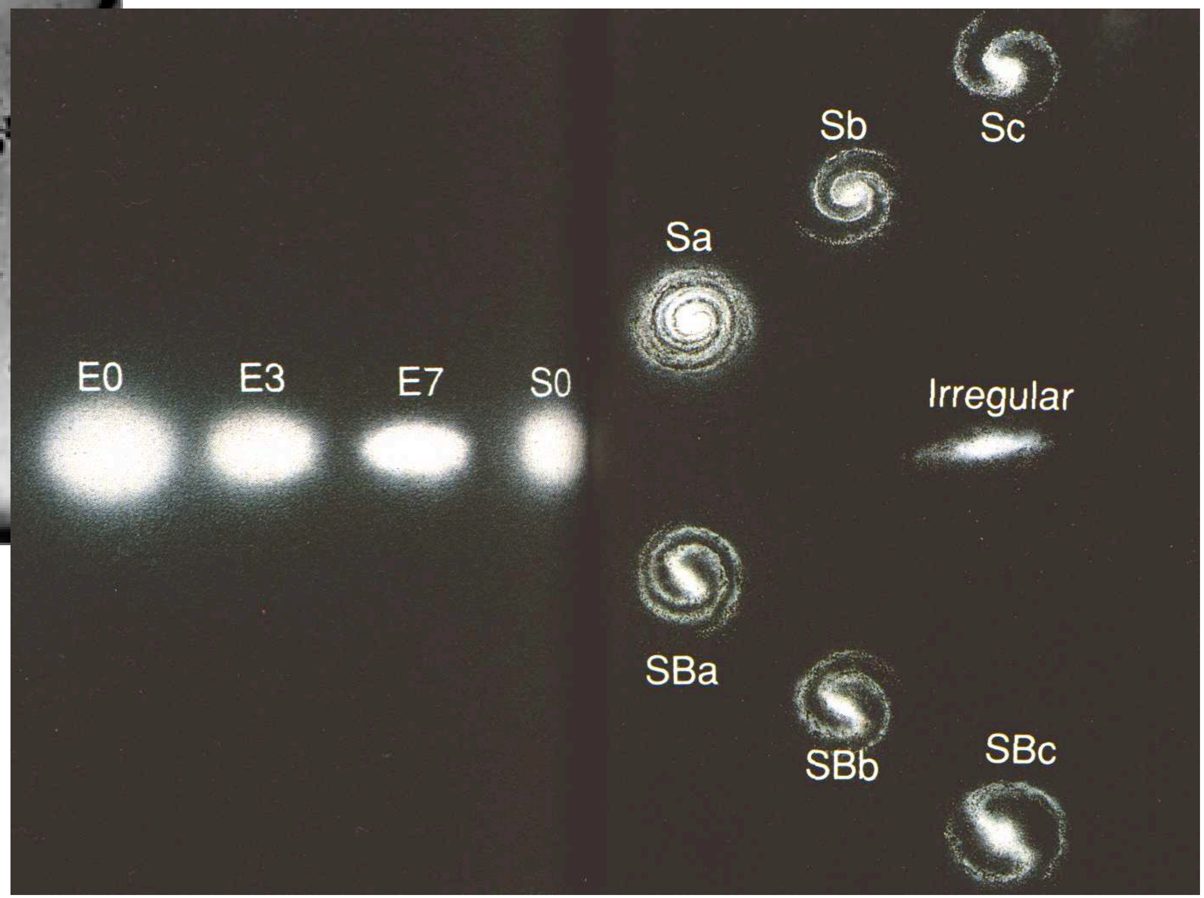




# What did Hubble see?

# What did Hubble see?



# Hubble's Law

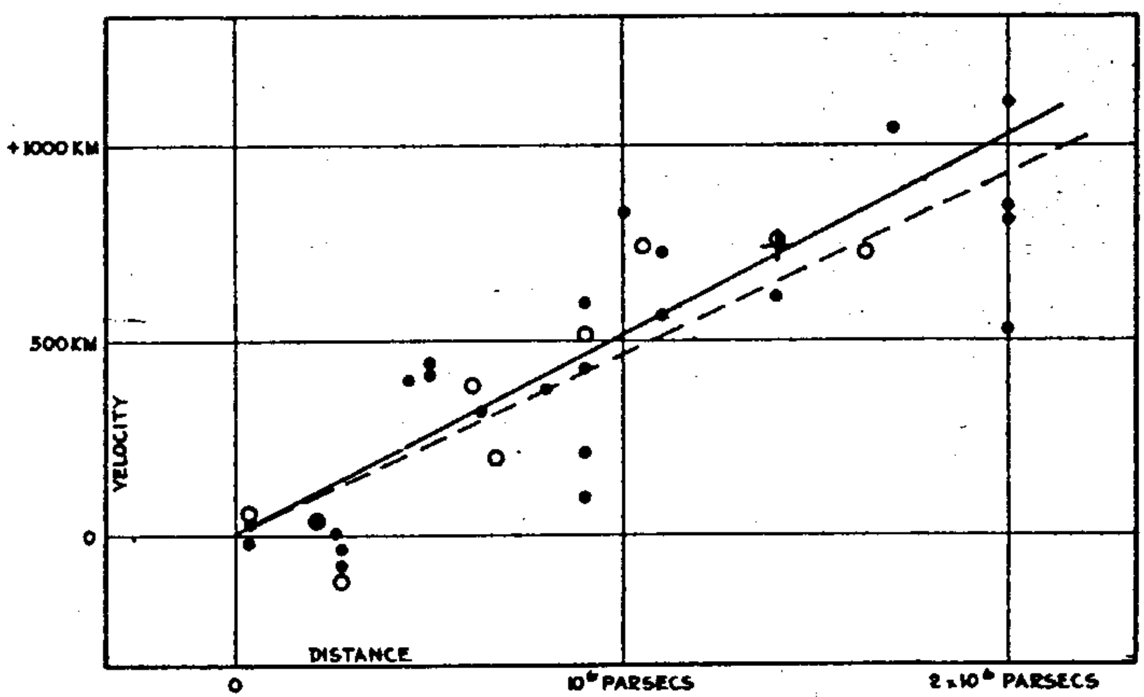



FIGURE 1

$$V = H_o r$$

*Where:*
V = recessional velocity
$H_o$= Hubble constant
r = distance (mpc)

E. Hubble (1929). A relation between distance and radial velocity among extra-galactic nebulae. *Proceedings of the National Academy of Sciences* 15(3).

# Hubble's Law

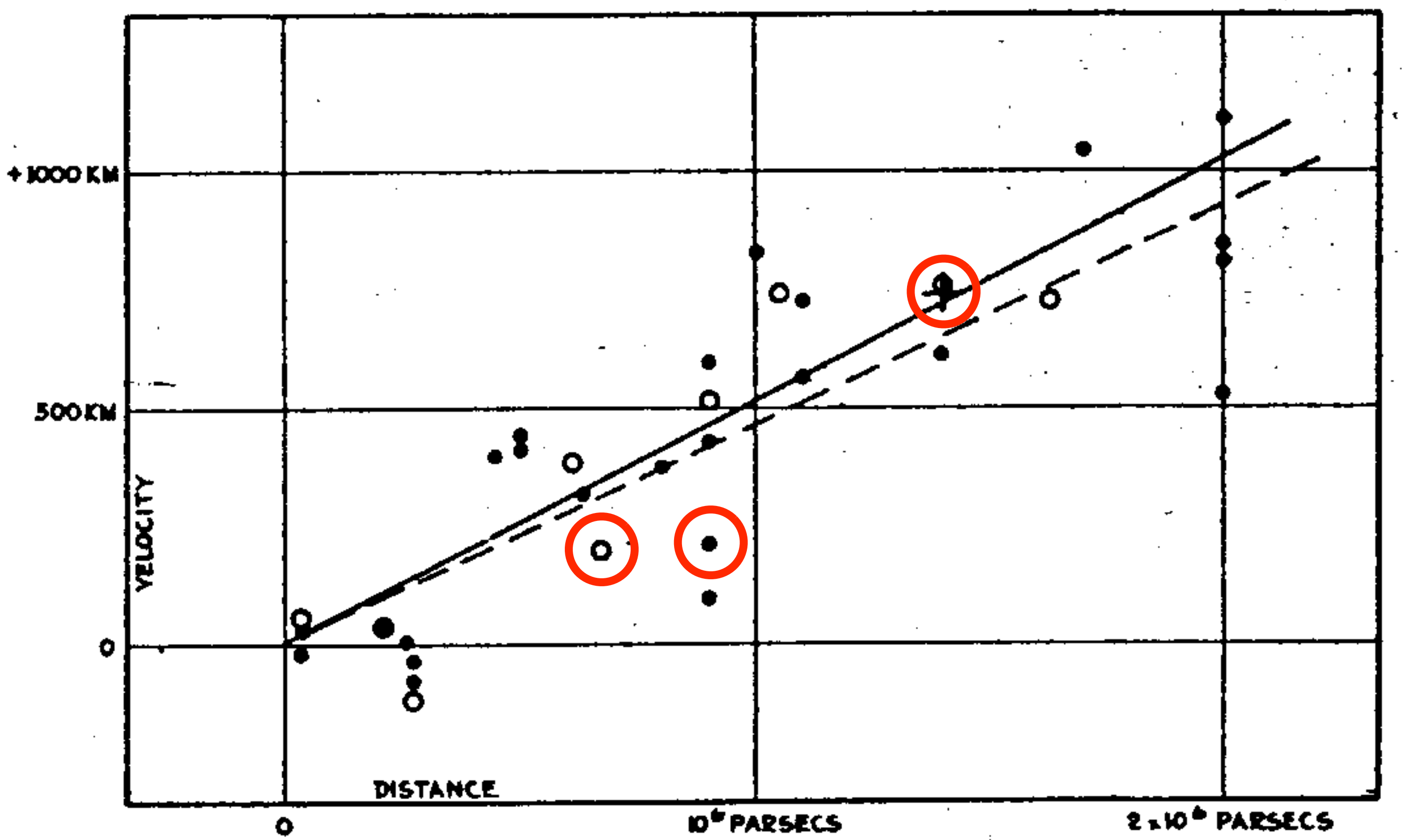
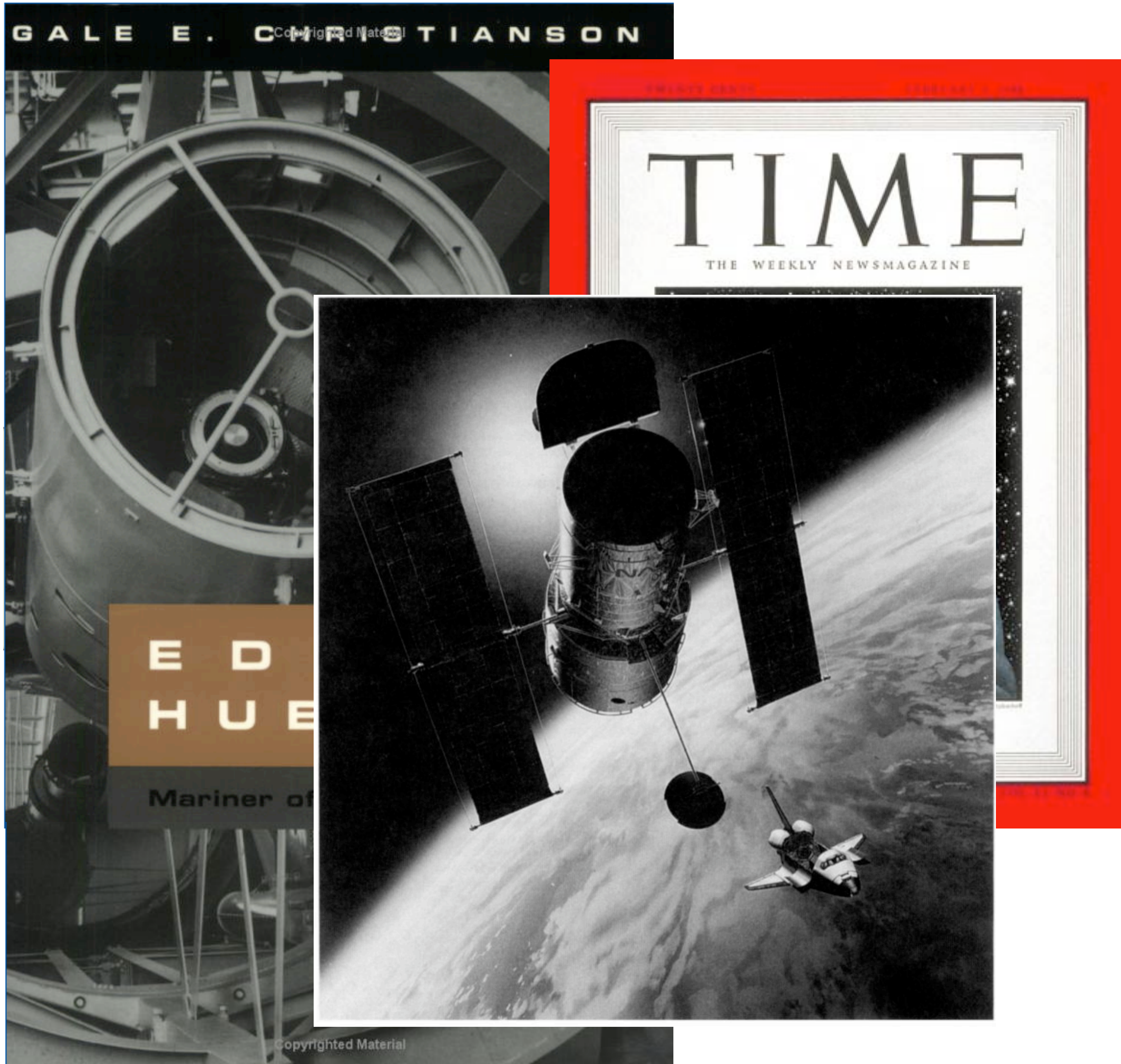



FIGURE 1


GALE E. CHRISTIANSON
Copyrighted Material
TIME
THE WEEKLY NEWSMAGAZINE
Copyrighted Material

**“The tool that is so dull that you cannot cut yourself on it is not likely to be sharp enough to be either useful or helpful.”**

– John W. Tukey

# Random variables

- The “embarrassingly dogmatic misnomer”
- They are neither *random*, nor are they *variables*
- A random variable is…
  - a function that maps from *instances* to *scales*
  - the numeric result of a non-deterministic experiment
- They can be distinguished from “fixed variables” whose value can be set or predetermined before the experiment
- They are not the individual values (e.g., 5.92), but rather the *process* of assigning value to instances or (colloquially) the set of values so assigned

# Examples

- *Recall* of an IR system, given query, corpus, and designated relevant documents
- *Size and speed* of code produced by a compiler, given source and a target processor
- *Number of database rows returned*, given an anytime query processor, query, database, and time
- *Lines of code written*, given an assignment, language, development environment, and programmer

# Notes

- The objects of study are usually the systems that enable random variables (e.g., IR systems), rather than the instances that the measures are on (e.g., queries).
- What we define as a random variable for a particular experiment can change as we discover deterministic and causal relationships in a given system

# Representation of data instances

- *i.i.d. instances* are commonly assumed
  - Independent — Knowing something about one instance tells you nothing about another
  - Identically distributed — Drawn from the same probability distribution
- Examples?
  - Queries in TREC data
  - Programs in SPEC benchmarks
  - Data sets in UCI repository
- Some alternatives
  - Time series
    (e.g., users submitting sets of slightly modified queries)
  - Relational
    (e.g., router performance embedded in a network)

# Populations and samples

- A population is a specified set of instances
  - An actual finite set of instances (e.g., the UCI data sets for machine learning research)
  - A generalization of an actual finite set (e.g., the set of all data sets that might be produced by a particular simulator in infinite time)
  - A purely hypothetical set which can be described mathematically (e.g., the set of all correct Java programs)
- Samples are finite subsets of populations

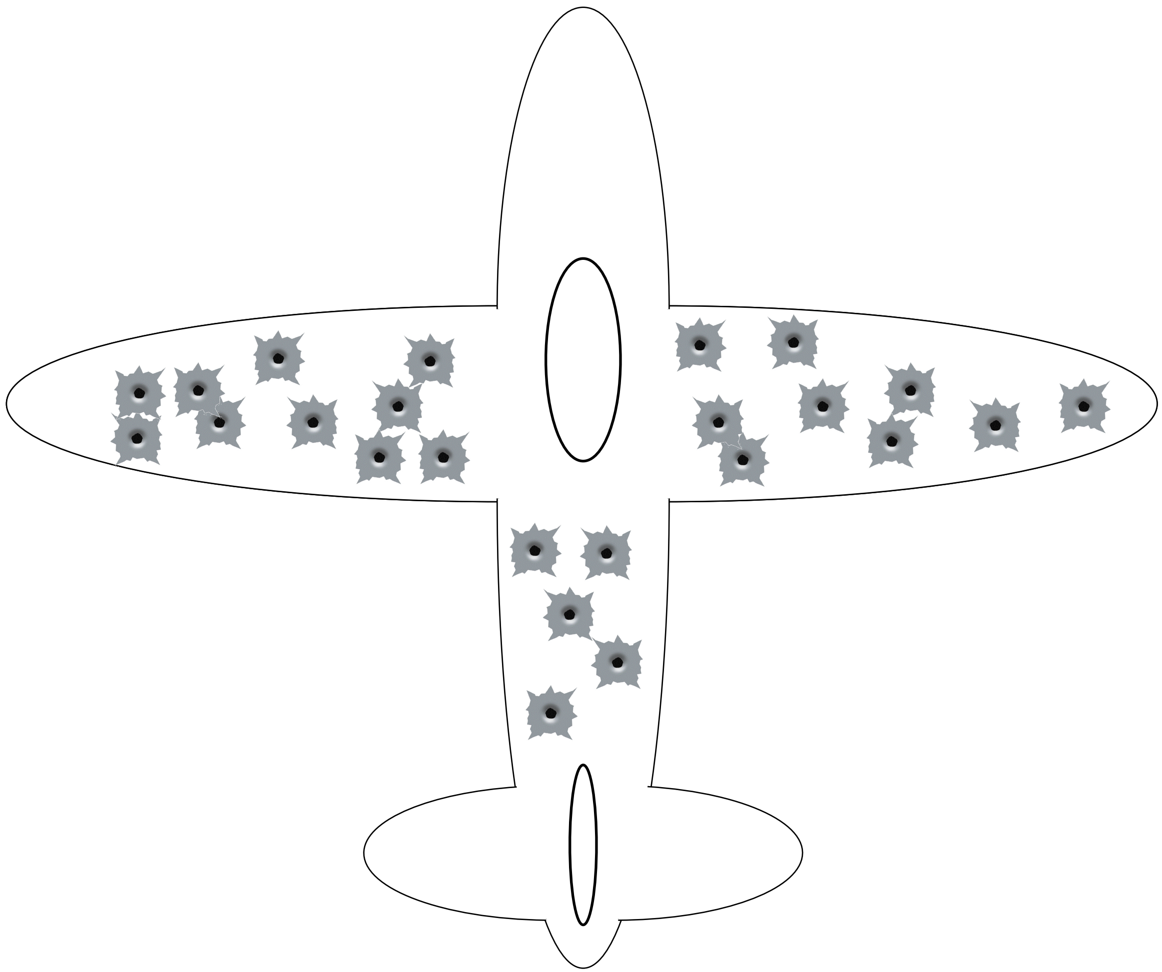
# Examples

| Populations | Actual data samples |
|---|---|
| All possible IR queries | The TREC 2005 HARD queries |
| All possible programs written in Java | The SPECjvm98 benchmarks |
| All Java programmers active in 2005 | Students taking CMPSCI 320 in Fall 2005 |
| The SPECjvm98 benchmarks | A subset of the benchmarks |

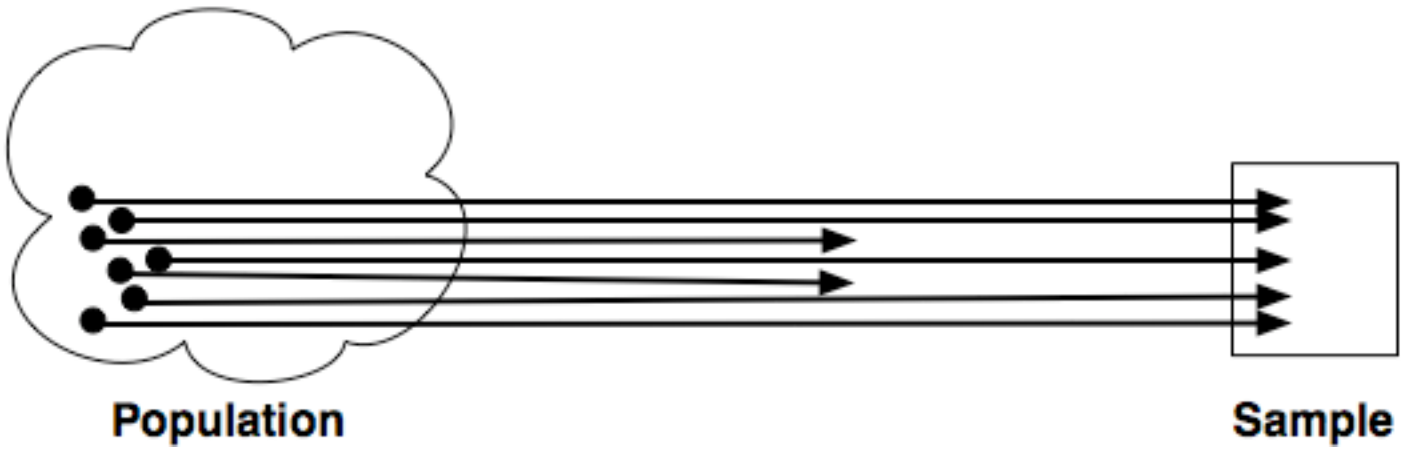
# Four stages of defining a sample

- The ***target population***
  (e.g., all computer programs)
- The ***sampling frame***
  (all programs written in Java or C++)
- The ***selected sample***
  (all programs written by CS undergraduate students in 200-level courses at UMass)
- The ***actual sample***
  (all programs actually turned in)

# Why is sampling difficult?



# Sampling problems




- The ***target population***
- The ***sampling frame***
- The ***selected sample***
- The ***actual sample***

# Random sampling in CS

- Random sampling isn't easy in CS
- ...but it's not easy in most sciences
- Answer isn't to give up, but to consider how to get closer to the ideal
  - Define the ideal population
  - Identify sources of bias in sampling and in subsequent steps of sample definition
  - Remove or mitigate as many sources of bias as possible
- Modify your confidence in your ability generalize based on your assessment of the match between your actual sample and your desired population

# Types of scales

- **Categorical, discrete, or nominal** — Values contain no ordering information (e.g., multiple-access protocols for underwater networking)
- **Ordinal** — Values indicate order, but no arithmetic operations are meaningful (e.g., "novice", "experienced", and "expert" as designations of programmers participating in an experiment)
- **Interval** — Distances between values are meaningful, but zero point is not meaningful. (e.g., degrees Fahrenheit)
- **Ratio** — Distances are meaningful and a zero point is meaningful (e.g., degrees K)

# Data transformations

- Downgrading type (e.g., interval to ordinal)
- Shifting intervals
  - Tukey's "ladder of powers": trans = original^(1-b)
  - E.g.: -2 -> original^3, 0.5 -> sqrt(original), 2 -> 1/original
- Combining several variables
  - Normalize measurements
    (e.g., Simsek & Jensen 2005, normalized to optimal)
  - Remove unwanted factors
    (e.g., remove file read times from total compile times)
  - Consider relation of two variables
    (e.g., Kirkpatrick & Selman, vertex/edge ratio)

# Exploratory data analysis

- “Exploratory data analysis (EDA)… employs a variety of techniques to…
  - maximize insight into a data set;
  - uncover underlying structure;
  - extract important variables;
  - detect outliers and anomalies;
  - test underlying assumptions;
  - develop parsimonious models; and
  - determine optimal factor settings”
- “The EDA approach is precisely that — an approach — not a set of techniques, but an attitude/philosophy about how a data analysis should be carried out.”

Source: http://www.itl.nist.gov/div898/handbook/eda/section1/eda11.htm

# Why EDA?

- Data analysis tools are typically used for
  - Hypothesis testing
  - Parameter estimation
- Graphics tools are typically used for presentation
- However, much of the quality of scientific work is determined by the quality of the hypotheses and models used by the researcher
- Can data analysis help suggest hypotheses?

# Resources

- Books
  - *Exploratory Data Analysis*, Tukey, (1977)
  - *Data Analysis and Regression*, Mosteller and Tukey (1977)
  - *Interactive Data Analysis*, Hoaglin (1977)
  - *The ABC's of EDA*, Velleman and Hoaglin (1981)
- Software
  - Data Desk (Data Description)
  - Fathom (Keypress)
  - XGobi (AT&T Research)

# Exploratory Data Analysis



For related context, please see the following paper:

Jerod Weinman, David Jensen, and David Lopatto. 2015. Teaching Computing as Science in a Research Experience. In *Proceedings of the 46th ACM Technical Symposium on Computer Science Education* (SIGCSE '15). ACM, New York, NY, USA, 24-29. http://dx.doi.org/10.1145/2676723.2677231

Other slides in this series may be found here: http://dx.doi.org/11084/10002