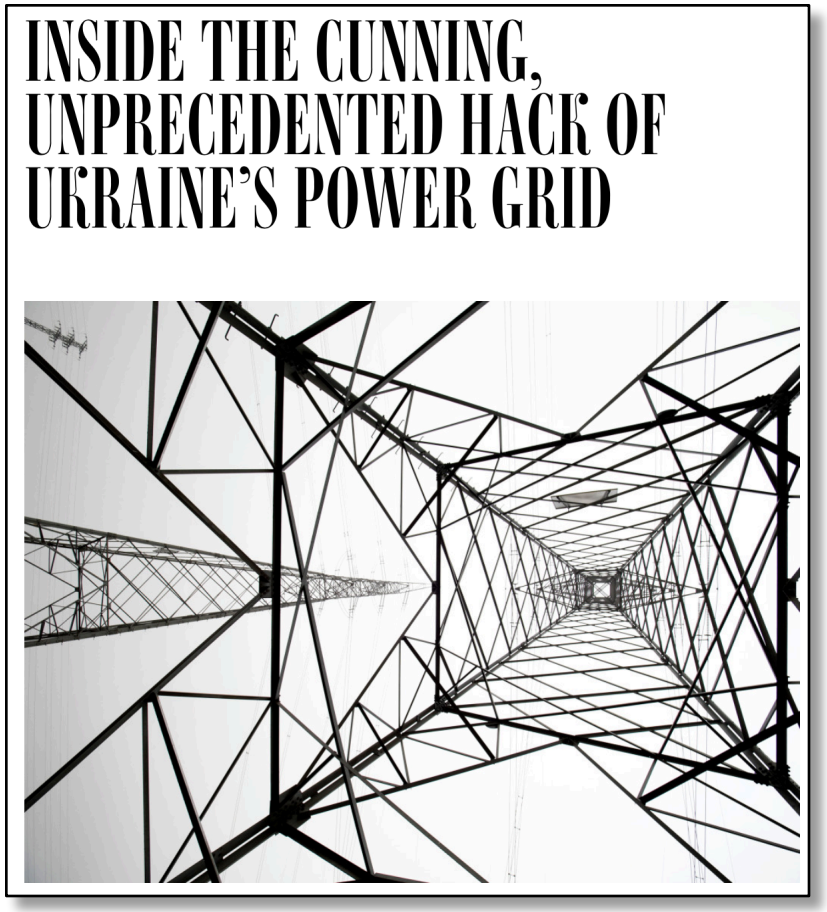
Data Summarization for Understanding the Cyber Landscape

Wingyan Chung, Ph.D.

Cyber Attack on Power Grid, 12/23/2015

"The operator grabbed his mouse and tried desperately to seize control of the cursor, but it was unresponsive. Then as the cursor moved in the direction of another breaker, the machine suddenly logged him out of the control panel. Although he tried frantically to log back in, the attackers had changed his password preventing him from gaining re-entry. All he could do was stare helplessly at his screen while the ghosts in the machine clicked open one breaker after another, eventually taking about 30 substations offline."

- The cyber attack on Ukraine's power grid caused
  - 30 substations offline
  - 230,000 residents without power
  - Two of three power distribution centers' backup supplies to be disabled



Source: http://www.wired.com/2016/03/inside-cunning-unprecedented-hack-ukraines-power-grid/

“To me what makes sophistication is logistics and planning and operations and … what’s going on during the length of it (the Ukraine cyber attack). And this was highly sophisticated.”

“Looking at the data, it looks like they would have benefited and been able to do more had they been planning and gathering intelligence longer,”

Robert M. Lee
Former cyber warfare operations officer for the US Air Force
Co-founder of Dragos Security, a critical infrastructure security company

Cybersecurity Informatics

- The interdisciplinary academic field involving information technologies, computer science, public policy, data science, and social and behavioral studies; local, state, and federal law enforcement and intelligence experts; and information technology industry consultants and practitioners.
- Focus on cybersecurity issues
  - Cybersecurity threats, exploits, vulnerabilities
- Development, application, and evaluation of informatics techniques
  - From various disciplines

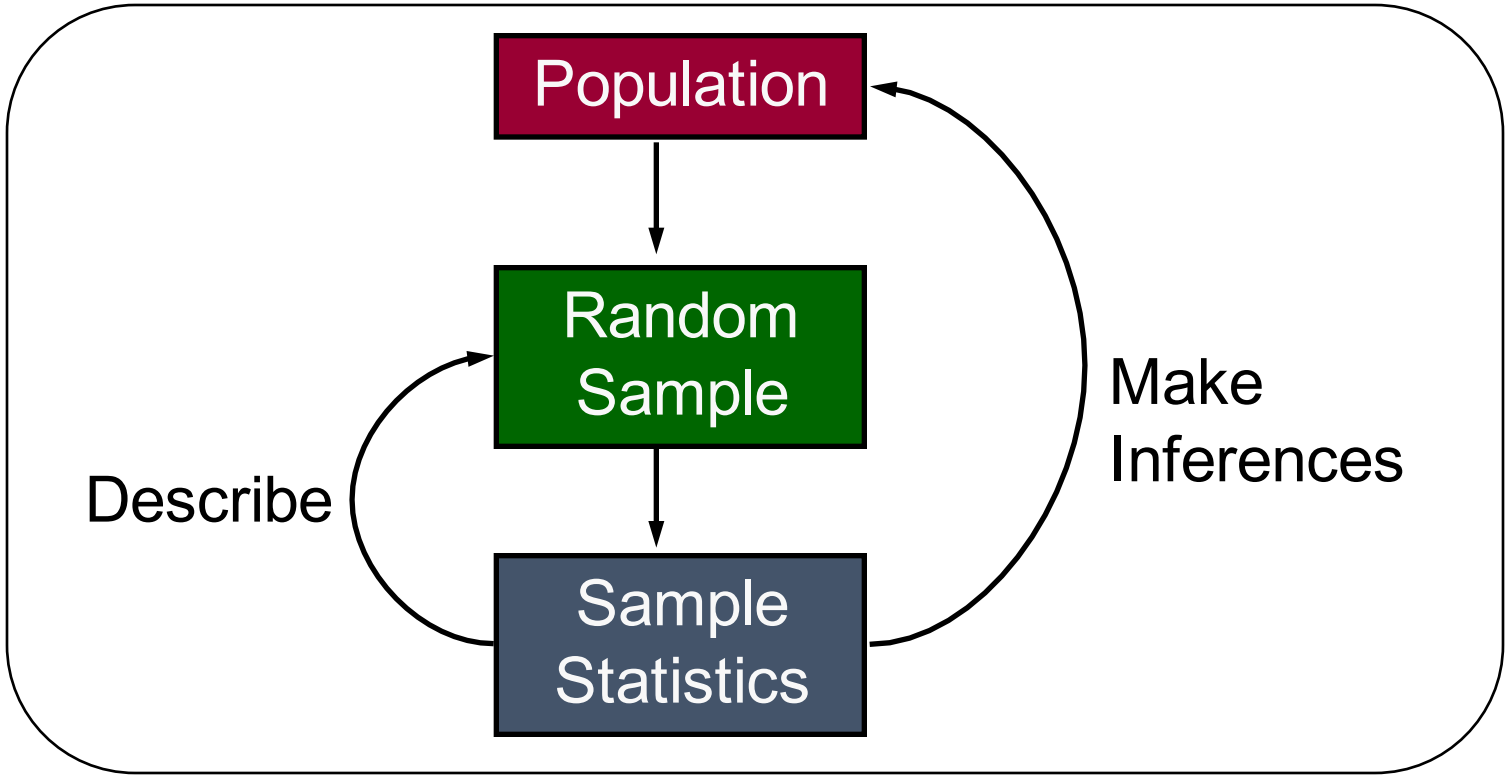
Disciplines involved in informatics

- Statistics and mathematics
- Network science and graph mining
- Machine Learning and data mining
- Computational Linguistics and NLP
- Computer science and engineering
- Social, behavioral, and economic sciences
- Human-computer interaction
- Social media analytics
- etc.

New theories,
theorems,
algorithms,
methods, constructs,
models,
mechanisms,
techniques, systems,
technologies, etc.

Describing and Summarizing Data

Process of Statistical Data Analysis

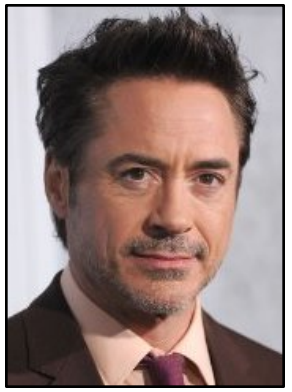


Type of Data

- Continuous Data
  - Interval Data, Ratio Data
- Categorical Data
  - Nominal Data, Ordinal Data

Cyber Terrorism Threats

- Marvo Inc. is a growing company in the entertainment and hotel industry with 3 theme parks located near Orlando, FL (US), Hong Kong (China), and London (UK).
- President and CEO, Bob Downey, has read a recent report by the US Department of Homeland Security that describes various cyber threats worldwide
- Bob examines the statistics about cyber attacks. He wants to use the analysis results to support decisions.



Bob Downey,
President and CEO

Cyber Attacks Worldwide

- Bob Downey has obtained the following data about cyber attacks (intrusion) in 2015:

Total Cyber Attacks Worldwide by Month, 2015											
Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
595	461	515	579	684	591	571	615	520	614	570	456

Assuming that the monthly data represent a sample.

- Sample size = *n* = 12
- Sample mean number of attacks per month =
- Sample SD of the number of attacks per month =
- Population variance =

$$\bar{x} = \frac{\sum x}{n} = \frac{595 + 416 + \ldots + 456}{12} = 564.25$$

$$s = \sqrt{\frac{\sum (x - \bar{x})^2}{n - 1}} = 66.13365$$

$$s^2 = \frac{\sum (x - \bar{x})^2}{n - 1} = 4373.659$$

Terminology

- Population – the set of all items or individuals of interest (e.g., all cyber attacks in U.S. in 2015)
- Sample – a subset of the population
- Statistic: a numerical quantity describing a **sample's characteristic**
  - E.g., "mean monthly number of cyber attacks in U.S."
- Parameter: a numerical measure that describes a **population's characteristic**
  - E.g., "mean weight of the U.S. population"
- Cybersecurity informatics – an interdisciplinary field devoted to the study of data, information technologies, and the related policy, social, and behavioral issues, in the context of the cybersecurity domain.

Types of Data

- Scalar
  - E.g., 461 incidents of cyber attacks happened in February

```
> feb <- 461
```

- Vector
  - E.g., The monthly numbers of cyber attacks in Quarter 1 are (595, 461, 515) respectively.
  - Shown as “numeric” in R;

```
> q1 <- c(595, 461, 515)
```

- Matrix
  - E.g., The monthly numbers of cyber attacks in four quarters of 2015 can be represented as a matrix

$$\begin{matrix} Q1 \\ Q2 \\ Q3 \\ Q4 \end{matrix} \begin{pmatrix} 595 & 461 & 515 \\ 579 & 684 & 591 \\ 571 & 615 & 520 \\ 614 & 570 & 456 \end{pmatrix}$$

Basic Commands in R Studio

- Opening a file
  - Command:`> setwd("<directory path>")`
  - Or click: `Session > Set Working Directory > Choose Directory …`
- Reading a data file into a data frame

```
> attack <- data.frame(read.table("cyberattack15.txt", sep="\t", header=TRUE))
```

  - Here, “attack” is the name of data frame; read.table() is the command to read a text file, sep=“\t” indicates field separator to be tab, header=TRUE indicates presence of header on first row
- Summarizing the data

```
> a <- attack[,2]
> m <- mean(a)
> s <- sd(a, na.rm=FALSE)
```

- Refer to m1-basics.R for details; also download data file “cyberattack15.txt”

Summary

- Cybersecurity informatics is an interdisciplinary academic field involving information technologies, computer science, public policy, data science, and social and behavioral studies.
- Many disciplines are involved in CI, including computer science, statistics, information systems, and social and political sciences, among others.
- The landscape of cybersecurity is wide and encompasses different types of incidents and all geographic regions.
- Basic computation and summarization of data can help understand a large number of cybersecurity incidents.