Primer on Analysis of Experimental Data and Design of Experiments

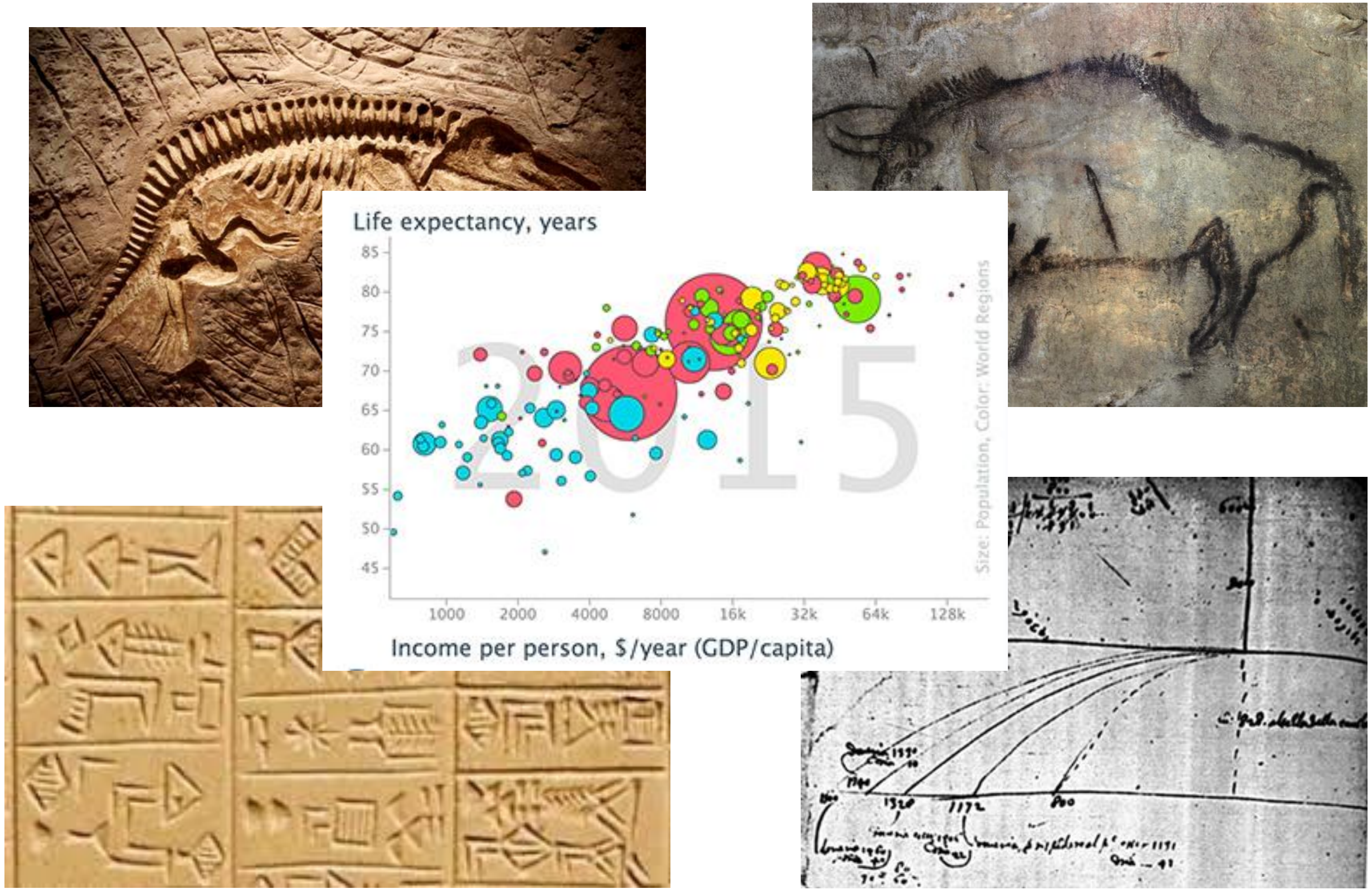
Lecture 1. Where do data come from?

Muhammad A. Alam
alam@purdue.edu

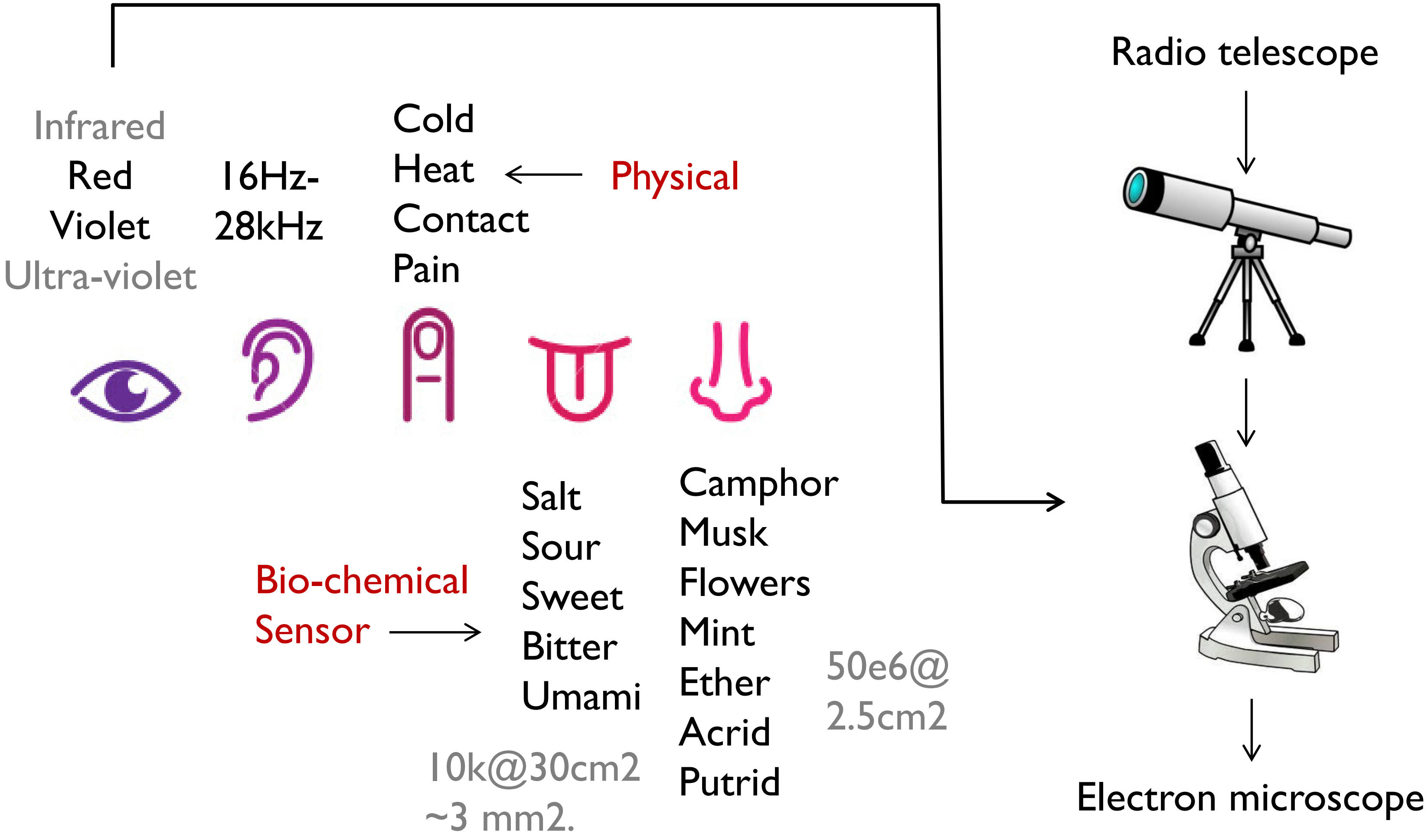
Outline

- A short history of data
- An example of small data
- Small vs. Big data
- What to expect from the class
- Conclusions

A short history of data


Life expectancy, years
85
80
75
70
65
60
55
50
45
2015
1000
2000
4000
8000
16k
32k
64k
128k
Income per person, $/year (GDP/capita)
Size: Population, Color: World Regions

Sensors and data



Infrared
Red
Violet
Ultra-violet

16Hz-
28kHz

Cold
Heat ← Physical
Contact
Pain

Bio-chemical
Sensor →

Salt
Sour
Sweet
Bitter
Umami

10k@30cm2
~3 mm2.

Camphor
Musk
Flowers
Mint
Ether
Acrid
Putrid

50e6@
2.5cm2

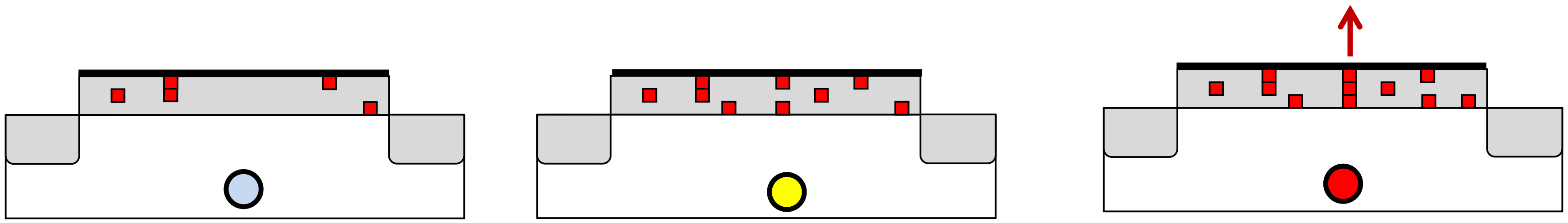
Radio telescope

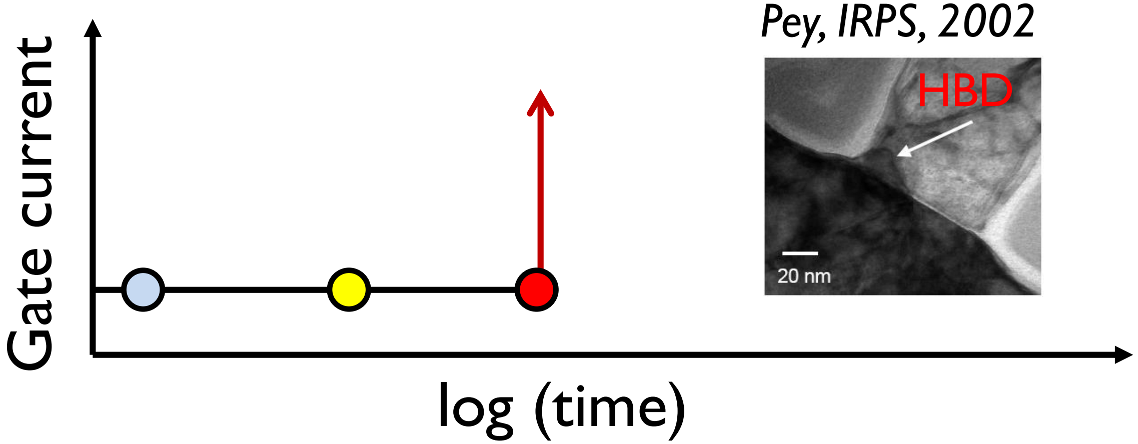
Electron microscope

Outline

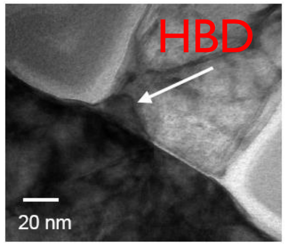
- A short history of data
- An example of small data
- Small vs. Big data
- What to expect from the class
- Conclusions

Time dependent dielectric breakdown

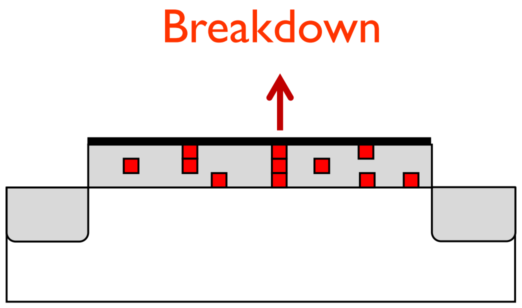


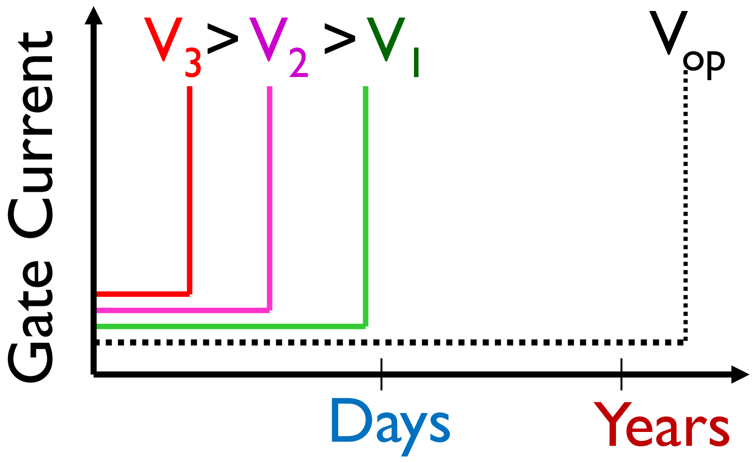


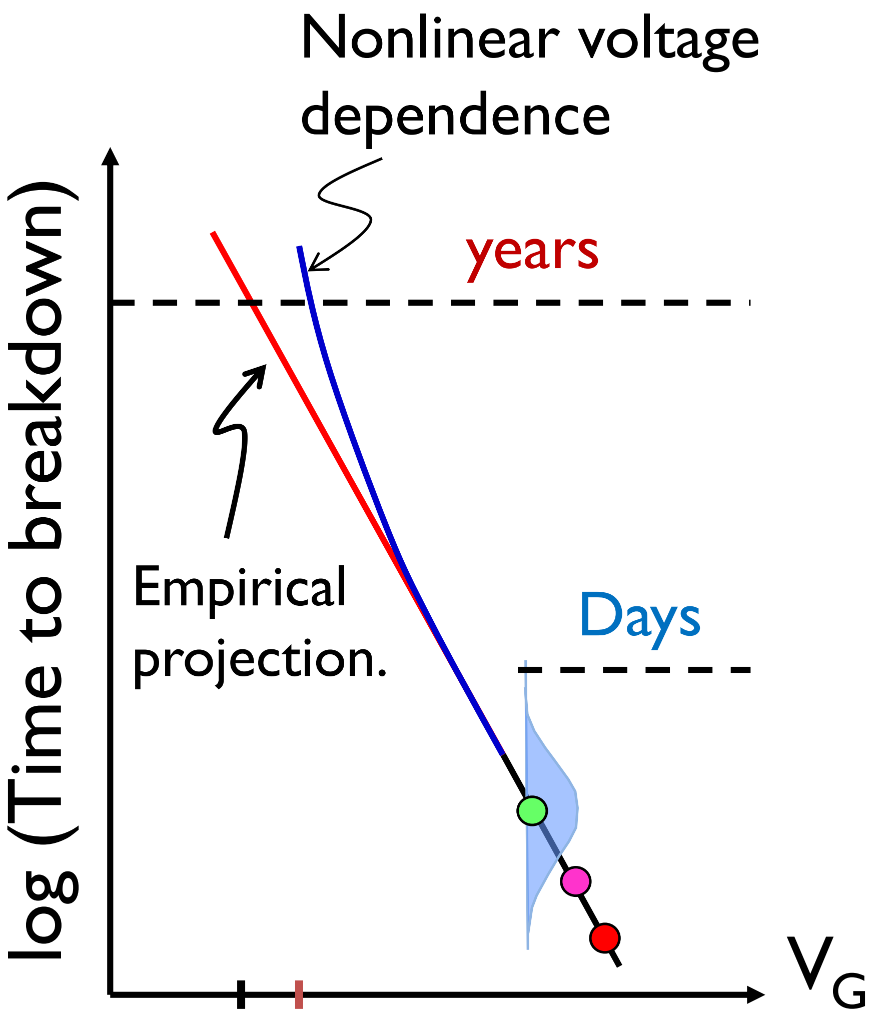
Pey, IRPS, 2002



Voltage-dependence of Dielectric Breakdown

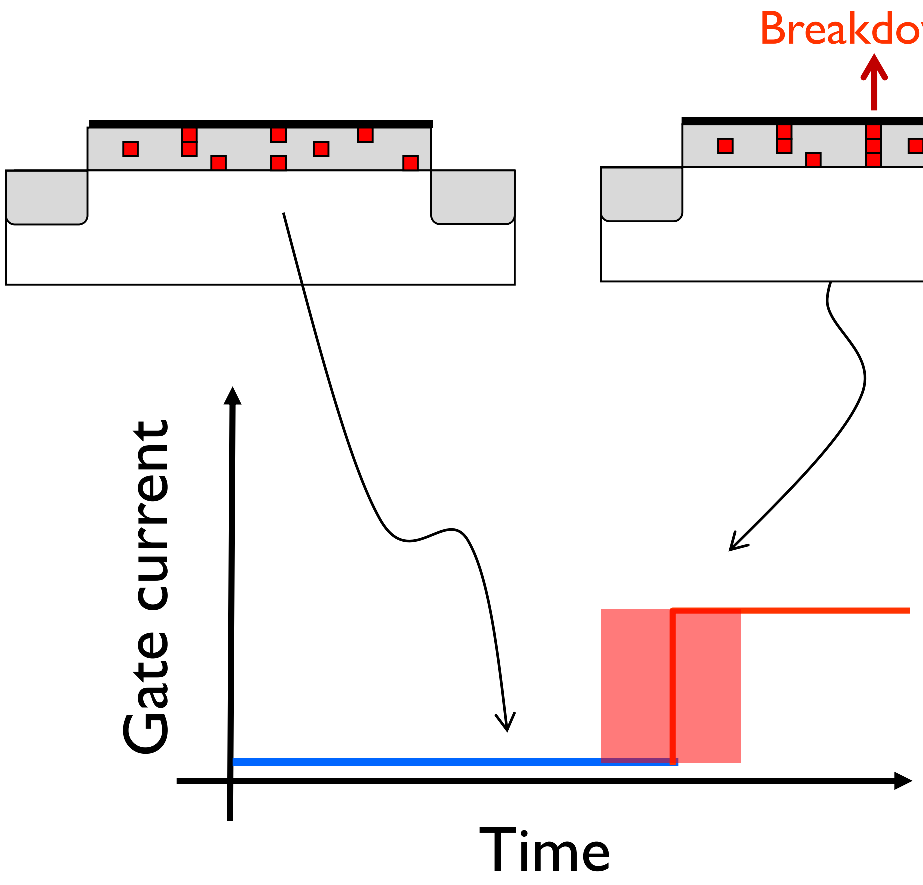




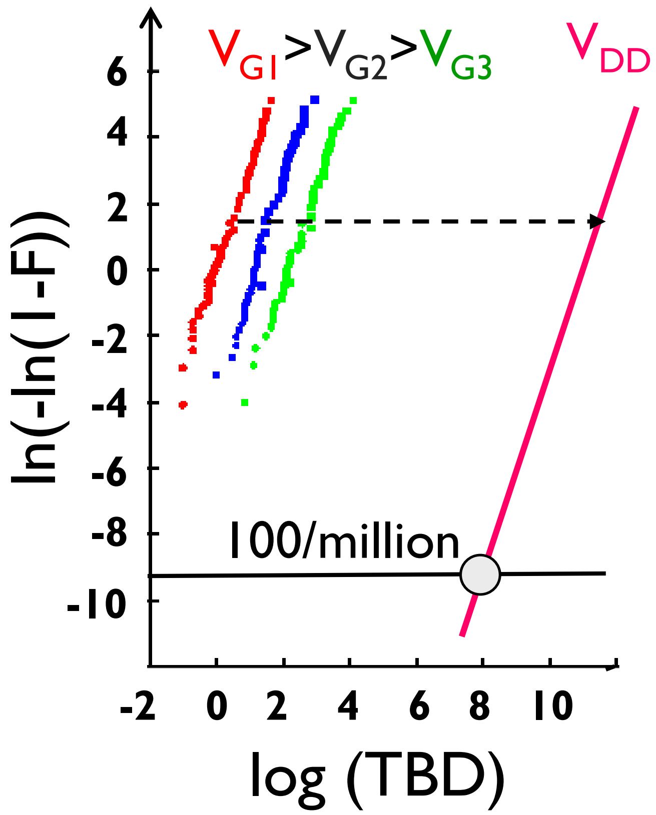


Muhammad A Alam, Purdue University

Weibull Distributed Failure times

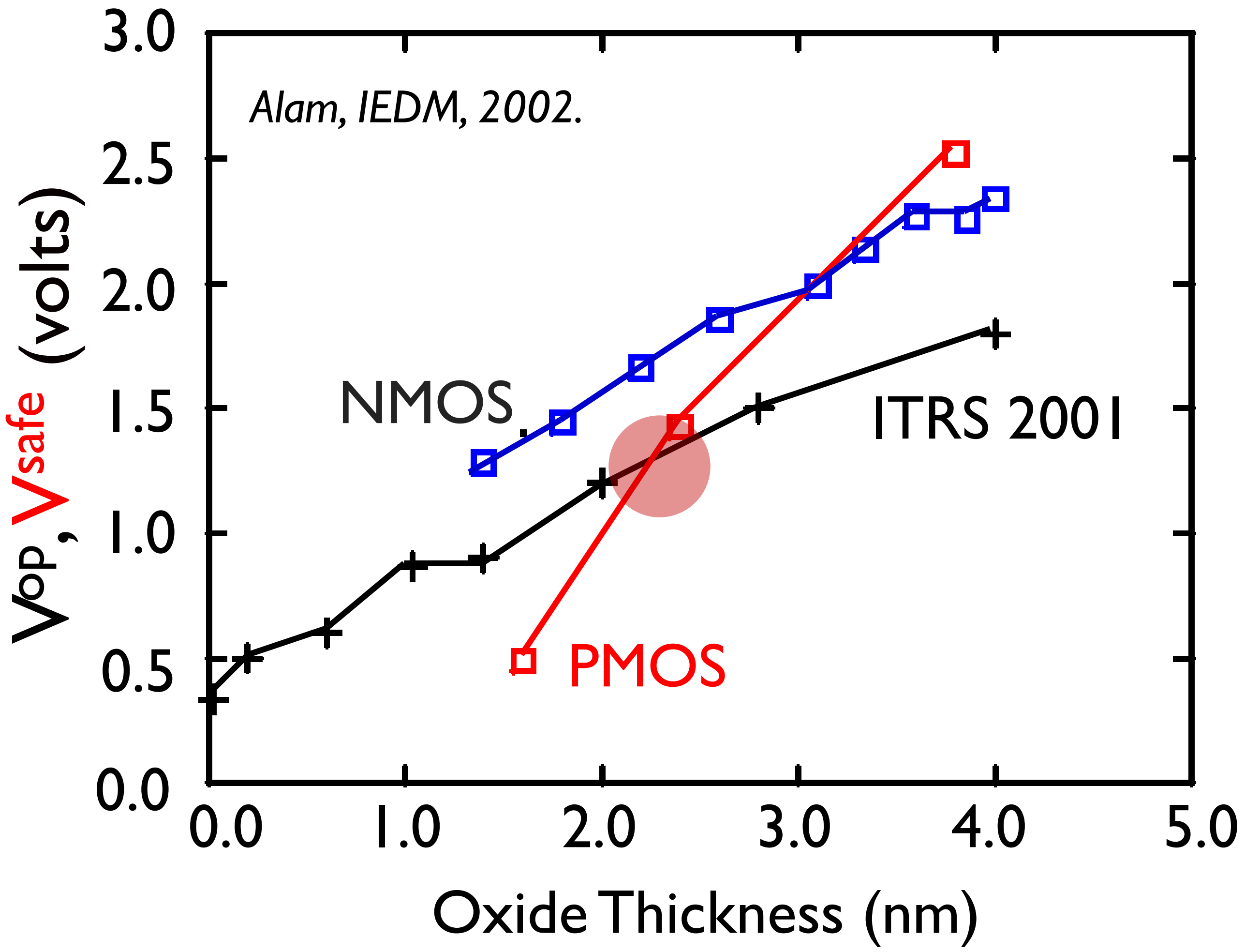


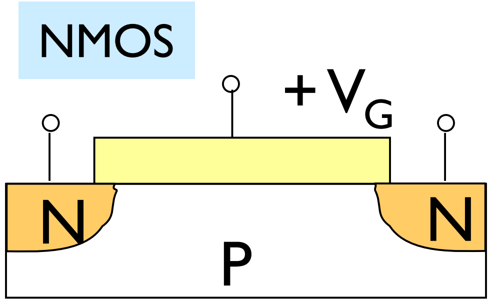
Weibull distribution

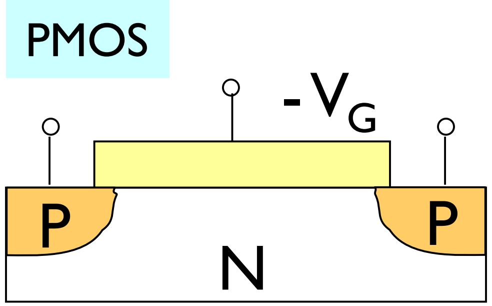


Average lifetime is not good enough ….

Predictions based on data







Issues with small data

- Small errors can have serious consequences.
- Generation of data is costly in terms of equipment, time, deadlines. Have to maximize information from small dataset.
- Often the dataset may be incomplete, the quality of the data non-uniform, and still we have to make the best decision possible.
- Often there could be competing hypothesis for a given distribution. Have to decide which one fits the data best. Based on the principles of Statistical decision theory.

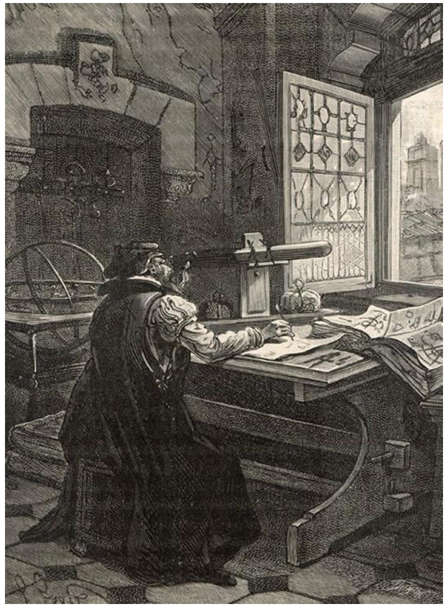
Outline

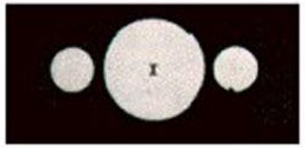
- A short history of data
- An example of small data
- Small vs. Big data
- What to expect from the class
- Conclusions

Big vs. small data

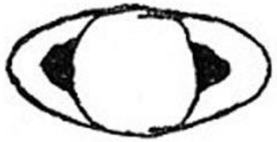
- Big data is obtained as is. One must ask intelligent questions to tease-out the answers embedded within the information. Census and insurance information are examples. Analysis is difficult, but they do represent real world conditions.
- Small data is often hypothesis driven and obtained from carefully designed experiments or survey. Data acquisition is planned and therefore expensive. The analysis is simpler, but may not represent real world conditions.

Small vs. big data

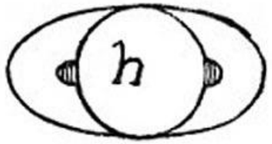




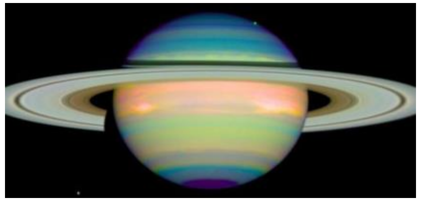
Galileo first sketch
1610



Better telescope
1616



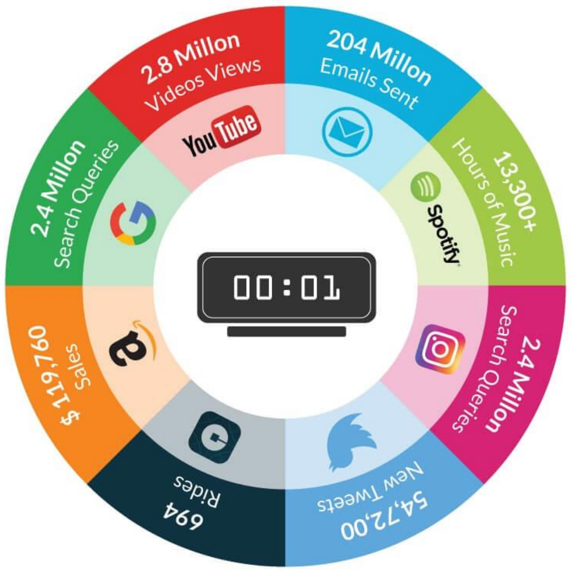
Published etch
1623



Where do data come from?

- Hundreds of petabyte of data every day.
- Social media sites
- Digital pictures
- Videos
- Purchase transaction
- GPS signals and so on.
- Scientific instrumentation
- Census data

What happens in an Internet Minute?

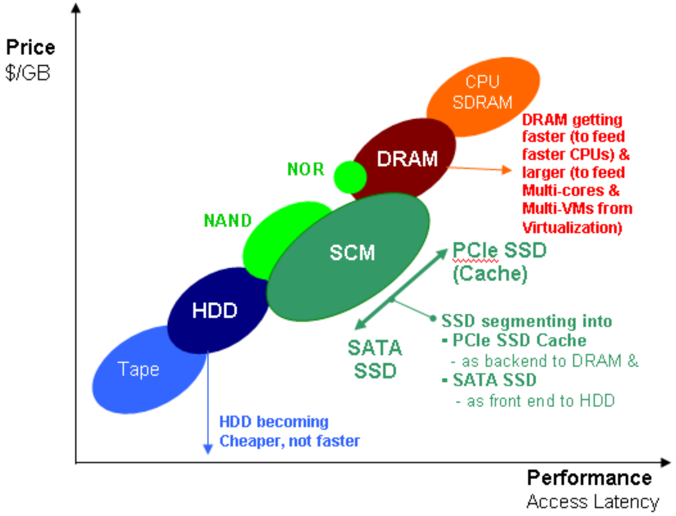


Repository of big data

- Google trends
- Federal Reserve Economic Data (FRED)
- Data.gov
- US Census Bureau
- European Union Open Data Portal
- Data.gov.uk
- The CIA World Factbook
- Healthdata.gov
- NHS Health and Social Care Information Centre
- Amazon Web Services public datasets
- Facebook Graph
- Gapminder
- Google Trends
- Google Finance
- Google Books Ngrams
- National Climatic Data Center
- DBPedia
- Topsy
- Likebutton
- New York Times
- Freebase
- Million Song Data Set
- DataScienceCentral selection of big data sets - check out the first itemized bullet list after clicking on this link
- Data sets used in our data science apprenticeship - includes both real data and simulated data - and tips to create artificial, rich, big data sets for testing models
- KDNuggets repository
- Data sets used in Kaggle competitions

……..driven by memory technology

- Cisco estimates: 1.8 ZB by 2016 and 7.2 ZB in 2021.
- If 1 MB is the size of the period at the end of sentence, 1.8 ZB is 460 km^2, eight times the size of Manhatten
- Amazon Web services, Google Cloud, IBM Cloud, Microsoft Azure.



DRAM getting faster (to feed faster CPUs) & larger (to feed Multi-cores & Multi-VMs from Virtualization)

SSD segmenting into
- PCle SSD Cache
  - as backend to DRAM &
- SATA SSD
  - as front end to HDD

HDD becoming Cheaper, not faster

Solid State Drive	
Access time	50/1000 ns
Capacity	2 terabytes
Data persistence	8-10 years
Read/Write Cycles	1000
Hard-Disk Drive	
Access time	7 millisecond
Capacity	8 terabytes
Data persistence	3-6 years
Read/write cycles	Indefinite
Magnetic Tape	
Capacity	12 terabytes
Data persistence	10-30 years
Read/write cycles	Indefinite

Outline

- A short history of data
- An example of small data
- Small vs. Big data
- What to expect from the class
- Conclusions

Outline

- Course Introduction
- Collecting and Plotting Data: Robust Data Analysis
- Physical and Empirical Distribution
- Model Selection and Goodness of Fit
- Design of Experiments: Scaling of Equations
- Design of Experiments: Buckingham Pi Theorem
- Statistical Theory of Design of Experiments
- Analysis of Data: ANOVA
- Big Data Classification: Singular Value Decomposition
- Machine Learning: Part 1
- Machine Learning: Part 2
- Physics-based Machine Learning
- Course Summary, Homework and Solutions

Outline of the course

$$\overline{y} = f\,(\overline{x}\,) \qquad \overline{x} = x_1, x_2, \dots x_n \qquad \overline{y} = y_1, y_2, \dots y_m$$

Introduction

Collecting and plotting $x_1, x_2, \dots x_n$

Physical and empirical $f, F, df/dx, \dots$

Model selection between $f_1, f_2, \dots$

Scaling theory with known f, $f(\overline{x}) = f(\overline{X})$

Scaling theory with unknown f, $\overline{x} \to X$

Principle component analysis for classifying $\{y\}$.

Design of experiments to determine $\overline{y}_{\max} = f\,(\overline{x}\,)$

Machine learning … Statistical approach to learn f

Physics-based machine learning $f = f_{\text{physics}} + \Delta f$

Conclusions

Few other information

Who should take this course

Anyone interested in modeling, simulation, collecting and analyzing the data, even reading a newspaper, etc.

What are the pre-requisites

Freshman/sophomore level preparation in physics and mathematics.

Grading

Class quizzes, homeworks, one final exam.

What to expect ….

- A deep understanding about how to analyze the data carefully, how to fit them to analytical functions, and how to use the data to make projections.
- Overfitting of the data is a general concern. Better fitting does not imply better decisions. You will be able to recognize and exclude overfitting.
- You will learn to design the experiments and simulations systematically. And then analyze the data and understand the correlation among various inputs systematically.
- The course will introduce you to basic concepts of machine learning from a simple, intuitive perspective. It will allow you to use more powerful tools currently available.
- This is however not a course on data-science or machine learning. If you are interested, you will take online courses.
- We will take a short quiz at the end of each class to make sure that the concepts are clear.

Reference Books

- Montgomery, Douglas C., and George C. Runger. Applied statistics and probability for engineers. John Wiley & Sons, 2010.
- Kirkup, Les. *Data analysis for physical scientists: Featuring Excel®*. Cambridge University Press, 2012.
- Strang, Gilbert, et al. *Introduction to linear algebra*. Vol. 3. Wellesley, MA: Wellesley-Cambridge Press, 1993.
- Machine Learning for Absolute Beginners, Oliver Therobald, ISBN 9781546172l8, 2017.

Conclusions

- To convert data into information, we must carefully process the data, with a nuanced understanding of the implications of data processing.
- Statistical data processing techniques have dramatically over the years. A deep understanding of discrete data analysis, information-theory based curve fitting, design of experiments, machine learning, etc. will maximize the information to data ratio.

References

- Montgomery, Douglas C., and George C. Runger. Applied statistics and probability for engineers. John Wiley & Sons, 2010.
- Kirkup, Les. *Data analysis for physical scientists: Featuring Excel®*. Cambridge University Press, 2012.
- Strang, Gilbert, et al. *Introduction to linear algebra*. Vol. 3. Wellesley, MA: Wellesley-Cambridge Press, 1993.
- Machine Learning for Absolute Beginners, Oliver Therobald, ISBN 9781546172l8, 2017.
- McKillup, Steve. *Statistics explained: an introductory guide for life scientists*. Cambridge University Press, 2011.
- Memory Technologies: IEEE Spectrum
- FRED Using Economic Data Is Easy With The Fred Database https://www.forbes.com/sites/billconerly/2015/03/02/using-economic-data-is-easy-with-the-fred-database/ (Mar 2, 2015, 10:30am)
- Other Resources on big data: https://www.datasciencecentral.com/profiles/blogs/17-short-tutorials-all-data-scientists-should-read-and-practice

Review Questions

1. Galileo did an experiment involving pendulum to determine its period of oscillation. Was his experimental results small data or big data?
2. Explain why incorrectly fitting a distribution may lead to incorrect decision and prediction?
3. What is the most important philosophical change in statistics over the last 300 years?
4. Give some examples of big data vs. small data before personal computers become widely available.
5. Using google trends, explain how the research topic you are interested in has changed over the last decade.