

# An Introduction to Data Analysis, Design of Experiment, and Machine Learning

## *Lecture 1. Where do data come from?*

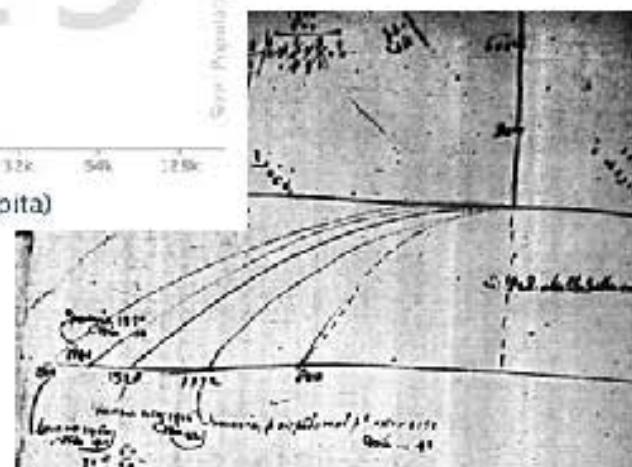
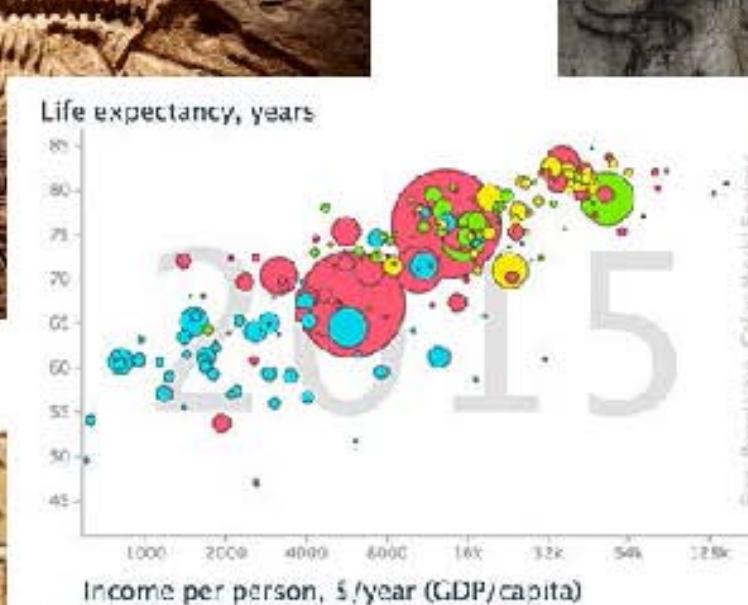
Muhammad A. Alam  
[alam@purdue.edu](mailto:alam@purdue.edu)



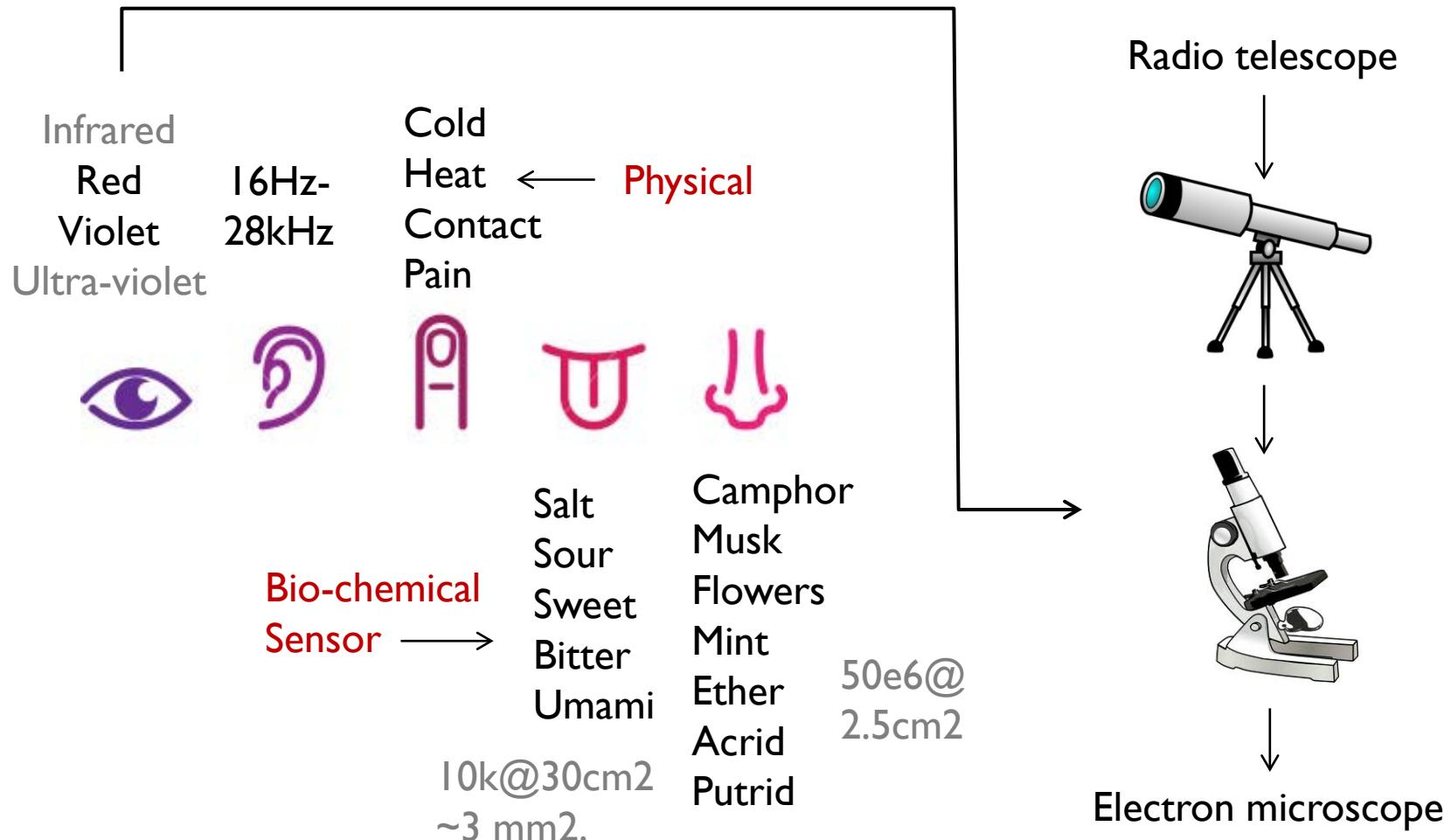
# Outline

- 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



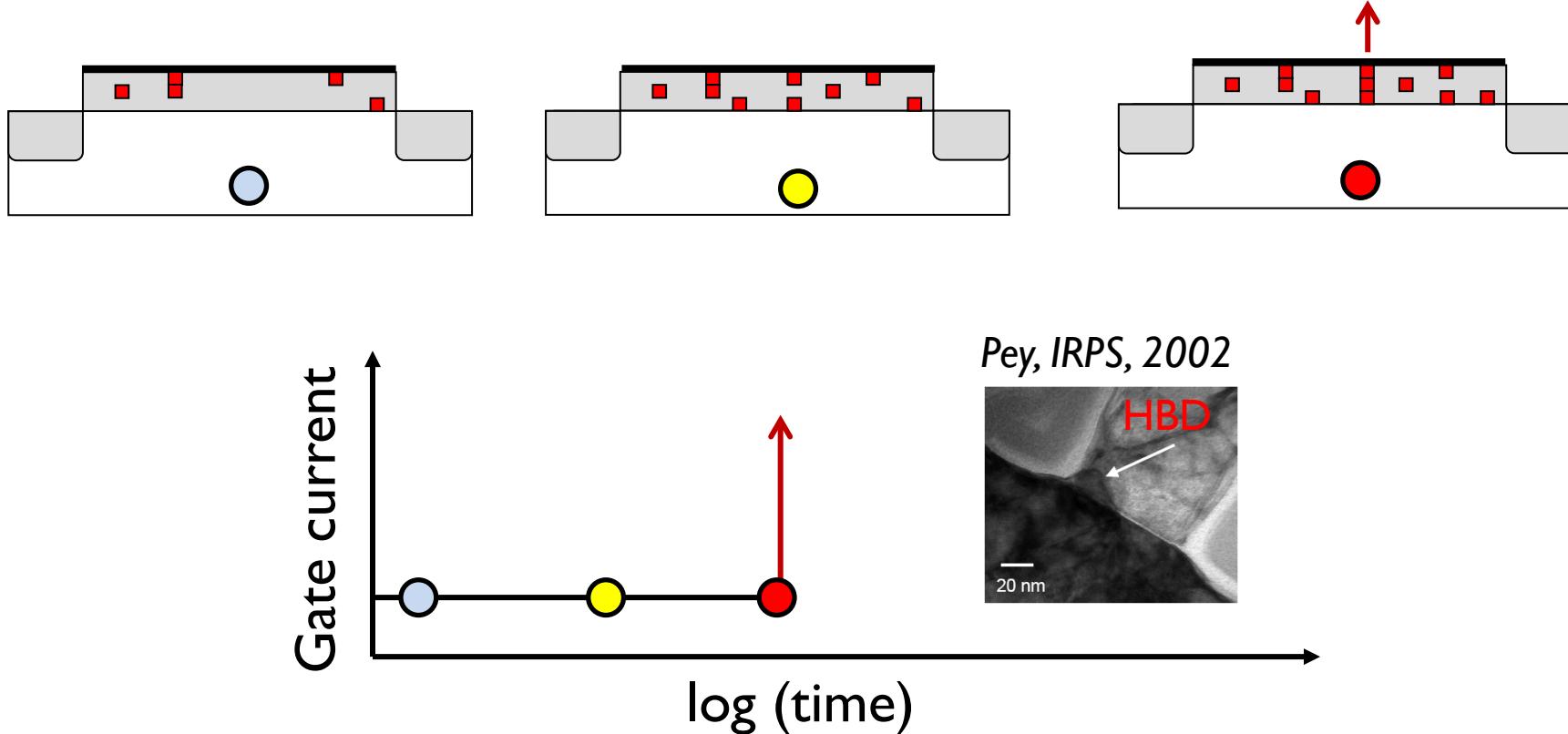
# Sensors and data



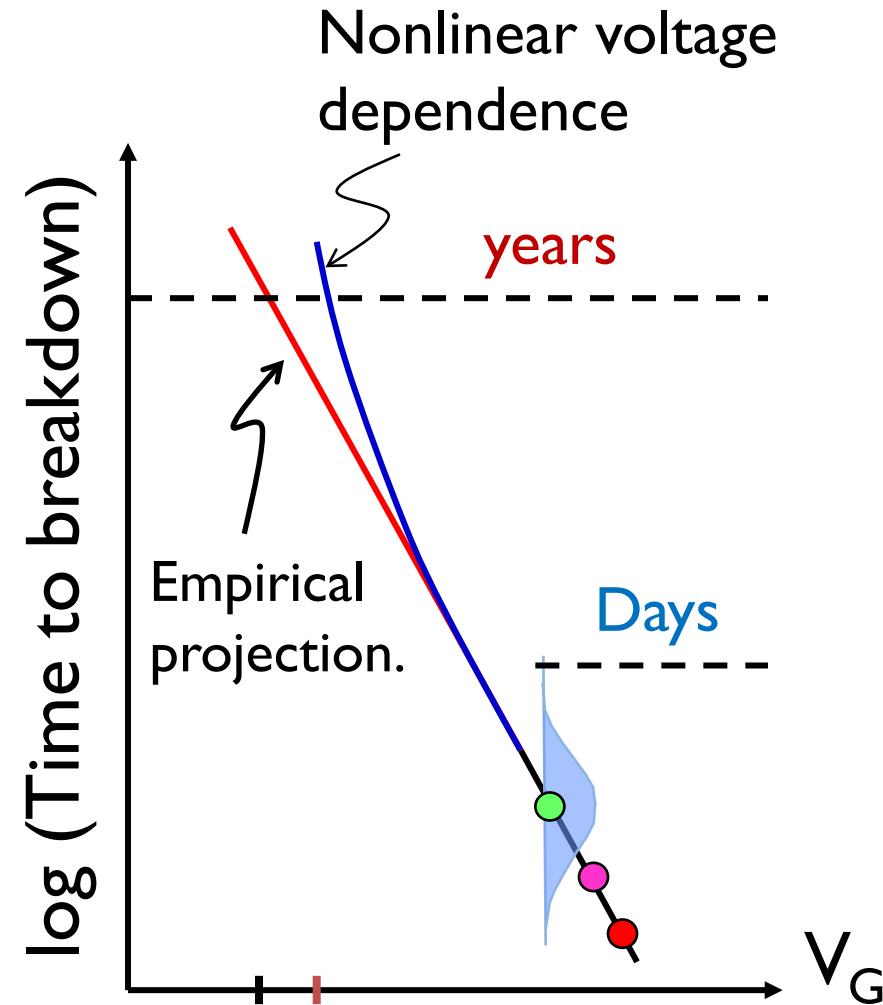
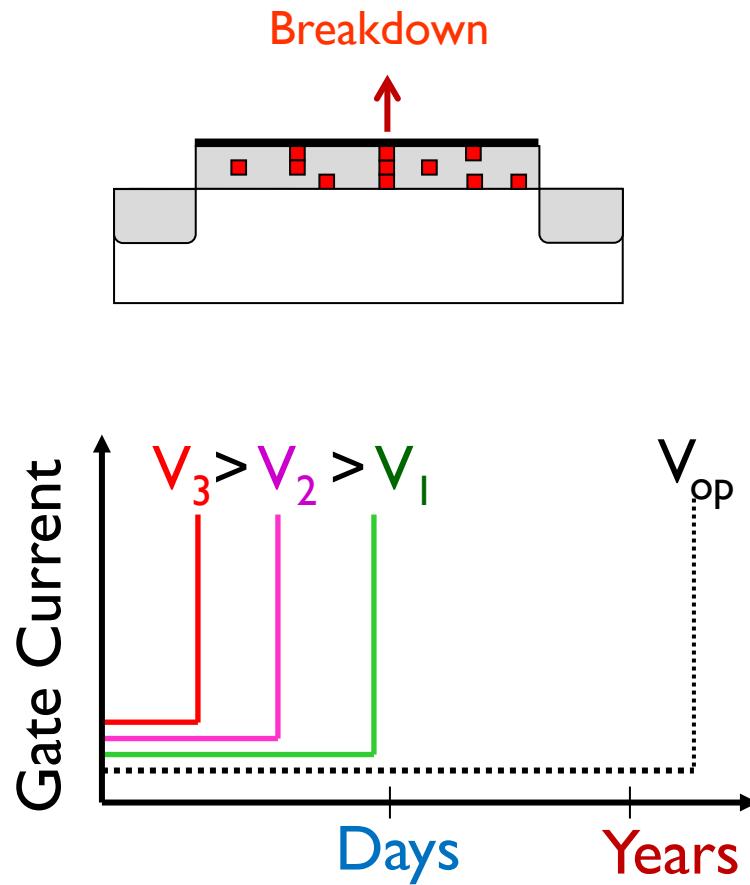
# Outline

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

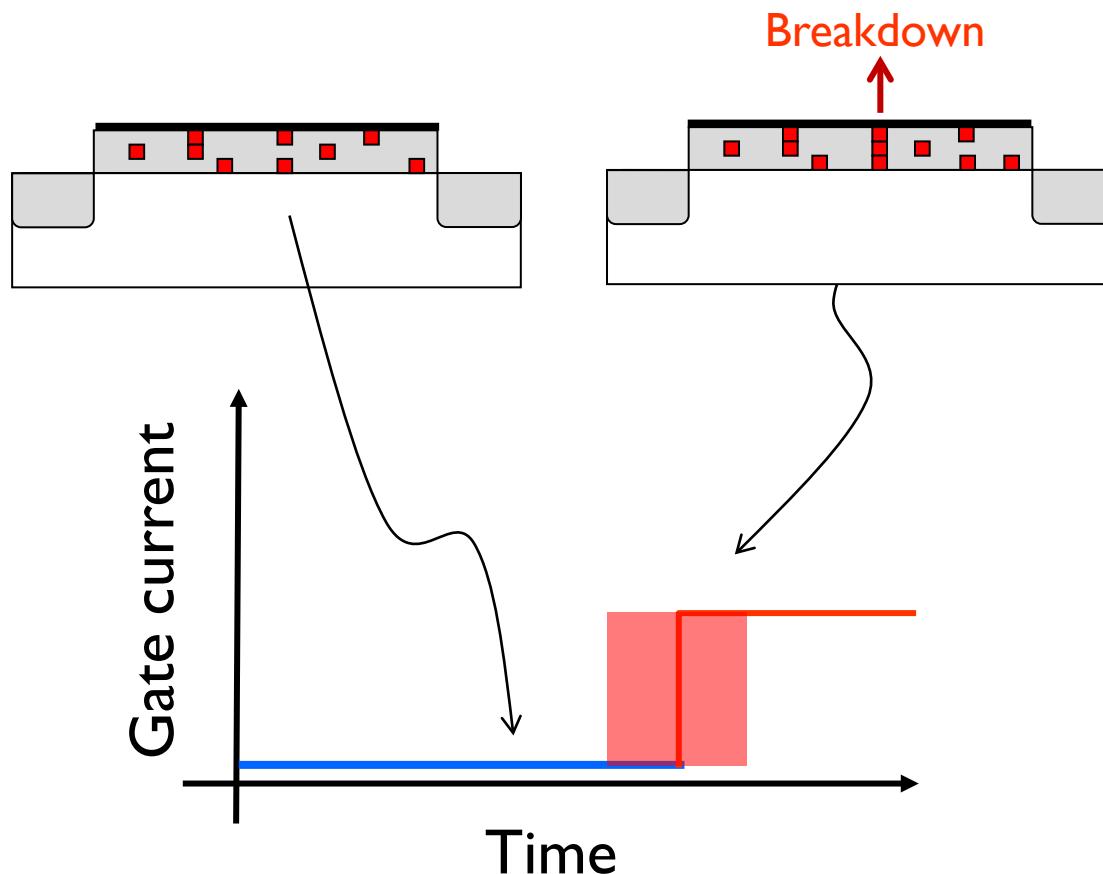
# Time dependent dielectric breakdown



# Voltage-dependence of Dielectric Breakdown

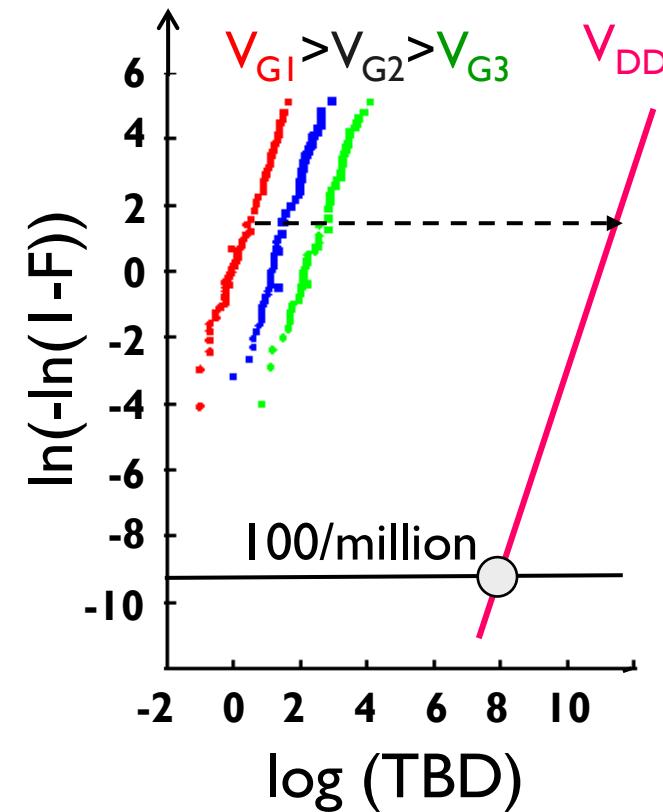


# Weibull Distributed Failure times

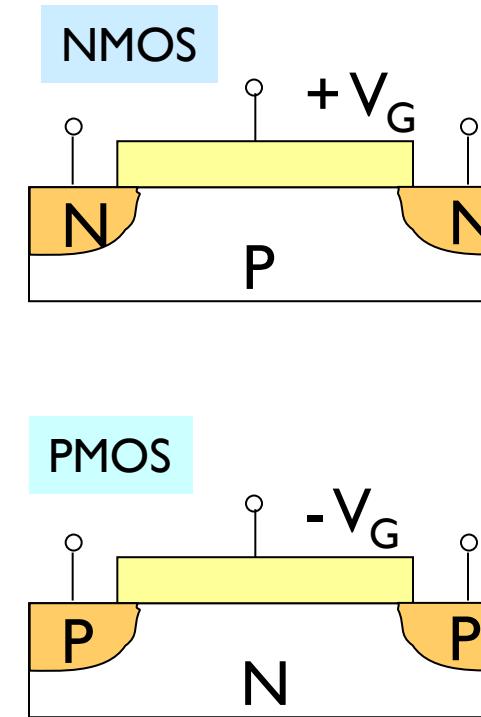
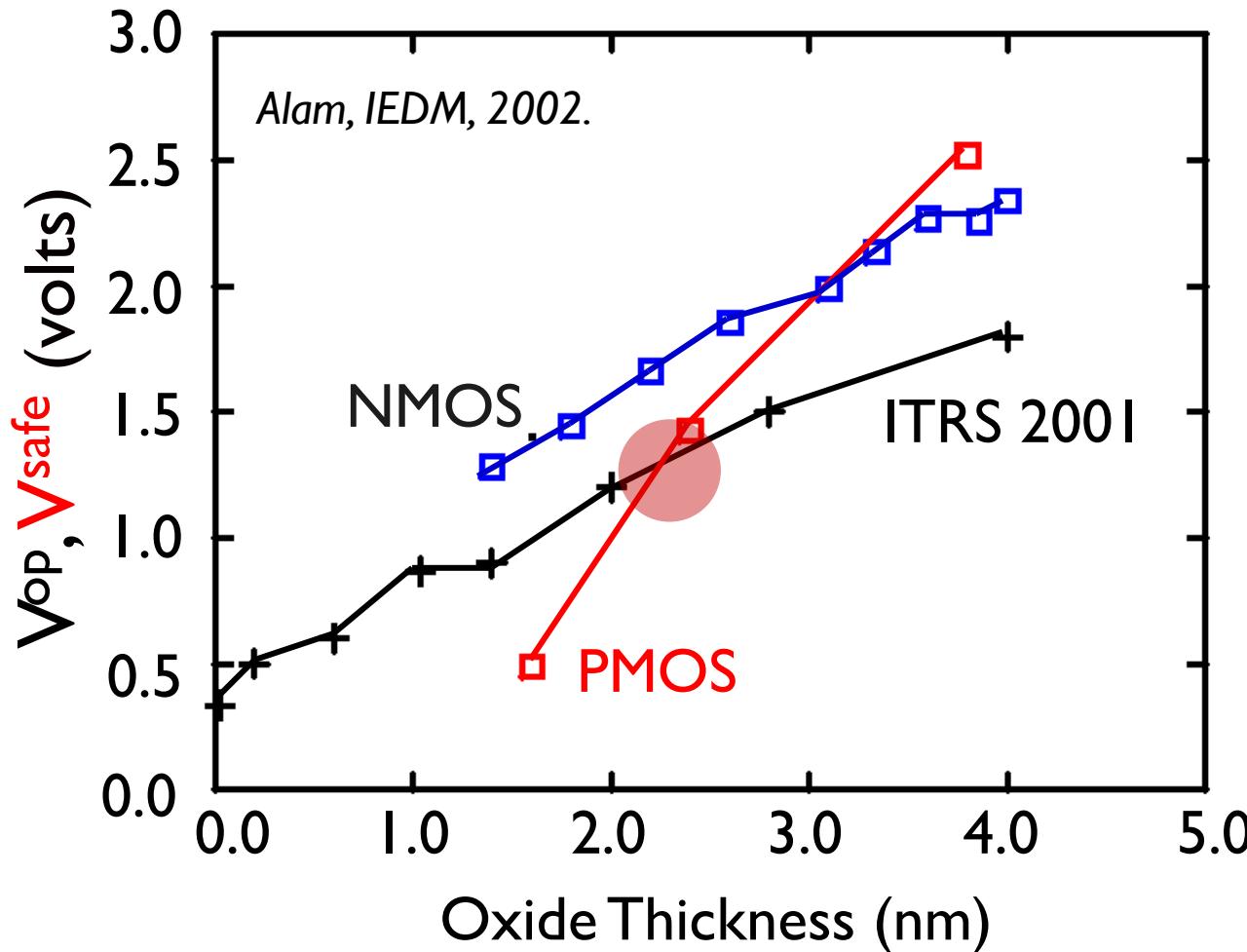


Average lifetime is not good enough ....

Weibull distribution



# 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.

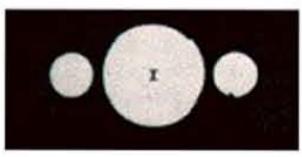
# Outline

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

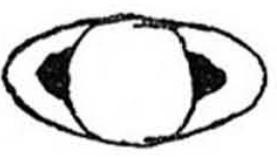
# Big vs. small data

- 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.

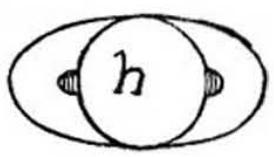
# Small vs. big data



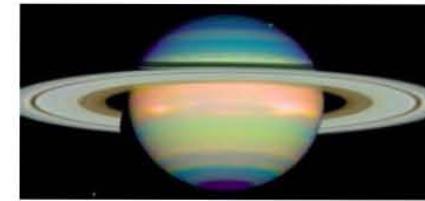
Galileo first sketch  
1610



Better telescope  
1616

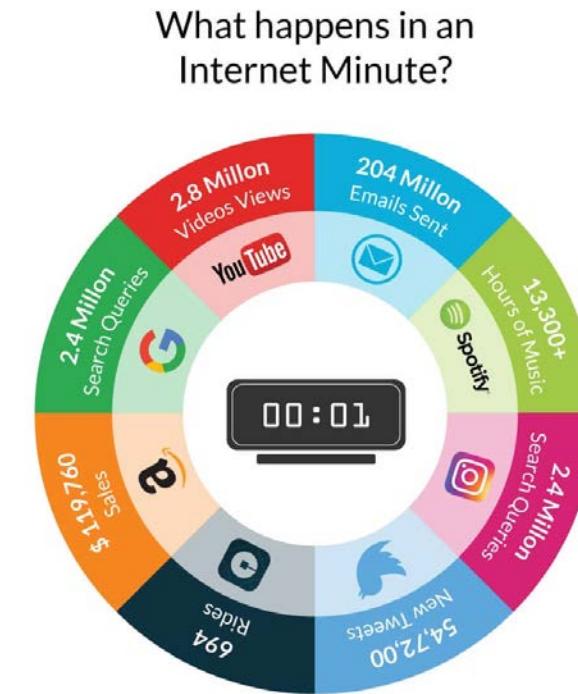


Published etch  
1623



# 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

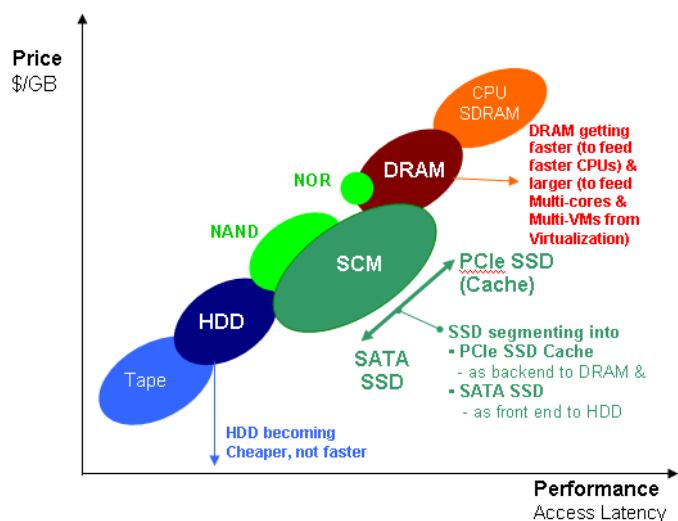


# Repository of big data

- 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](#)
- [Firebase](#)
- [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
- [KD Nuggets 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  $460 \text{ km}^2$ , eight times the size of Manhattan
- Amazon Web services, Google Cloud, IBM Cloud, Microsoft Azure.



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

# 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.

# 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 I
- Machine Learning: Part 2
- Physics-based Machine Learning
- Course Summary, Homework and Solutions

# Outline of the course

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

## 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  $f$ ,  $f(\bar{x}) = f(\bar{X})$

Scaling theory with unknown  $f$ ,  $\bar{x} \rightarrow X$

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

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

Machine learning ... Statistical approach to learn  $f$

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

## Conclusions

# 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 978154617218, 2017.

# 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.

# 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.

# An Introduction to Data Analysis, Design of Experiment, and Machine Learning

## *Lecture 2. Collecting and Plotting Data*

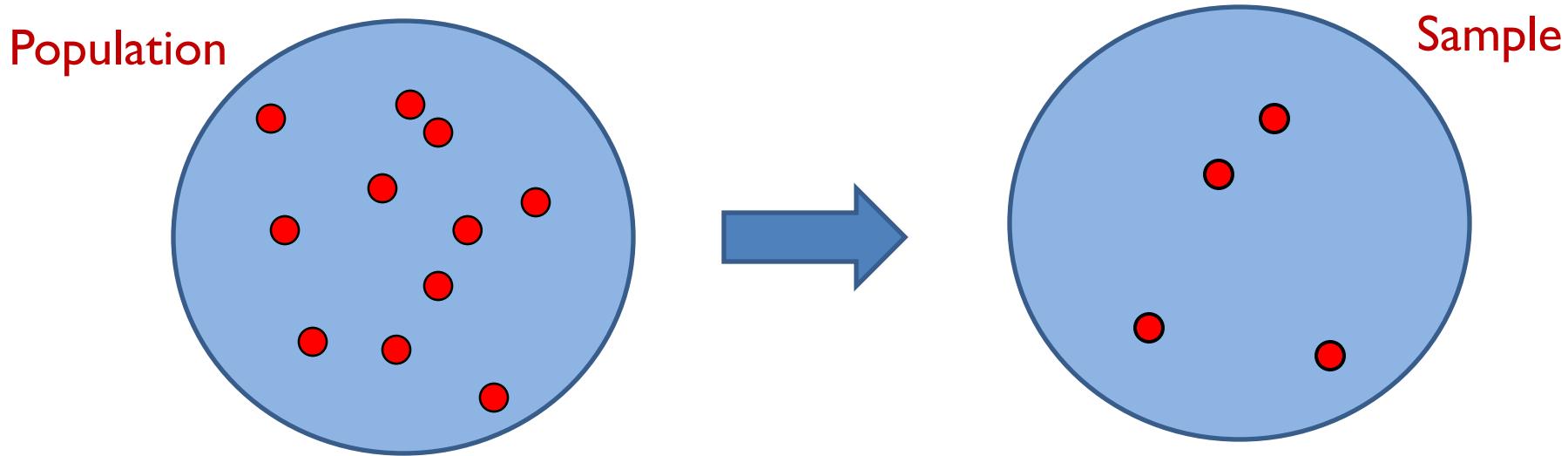
Muhammad A. Alam  
[alam@purdue.edu](mailto:alam@purdue.edu)



# Outline

- I. Review of the traditional statistical metrics
- 2. Parametric vs. Nonparametric information
- 3. Preparing data for projection: Hazen formula
- 4. Preparing data for projection: Kaplan formula
- 5. Conclusions

# Population vs. Sample Distribution



$$\langle t \rangle = \frac{\sum_{j=1,N} t_j}{N}$$

$$s^2 = \frac{\sum_{j=1,N} (t_j - \langle t \rangle)^2}{N - 1}$$

$$\sigma^2 = \frac{\sum_{j=1,N} (t_j - \langle t \rangle)^2}{N}$$

**Example Excel routines ...**

STDEV (2.1, 3.5, 4.5, 5.6) = 1.488

STDEVP= (2.1, 3.5, 4.5, 5.6) = 1.2891

# Moments of the Experimental Data (or discrete distribution)

Distribution-free statistical measure of data ....

$$\langle t \rangle = \frac{\sum_{j=1,N} t_j}{N}$$

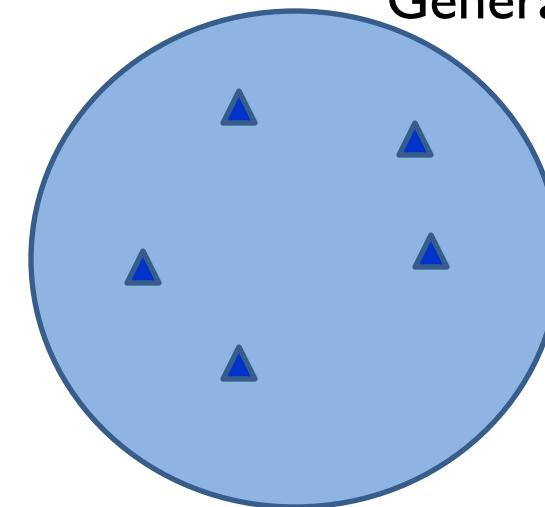
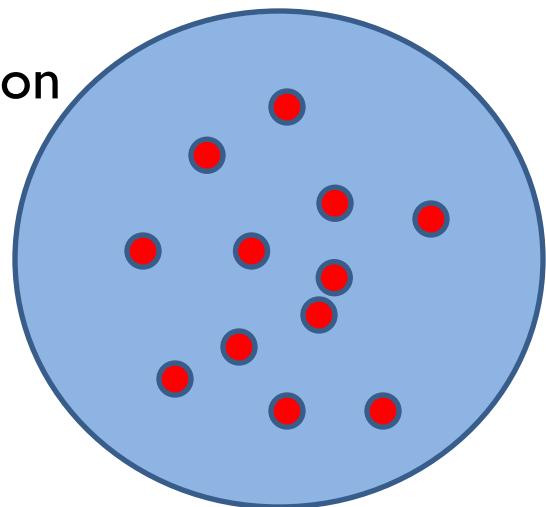
$$s^2 = \frac{\sum_{j=1,N} (t_j - \langle t \rangle)^2}{N - 1}$$

Parameter-space

$$\delta_{T_k} = \sqrt[k]{\frac{\sum_{j=1}^N (t_i - \langle t \rangle)^k}{N - k + 1}}$$

General formula

Population

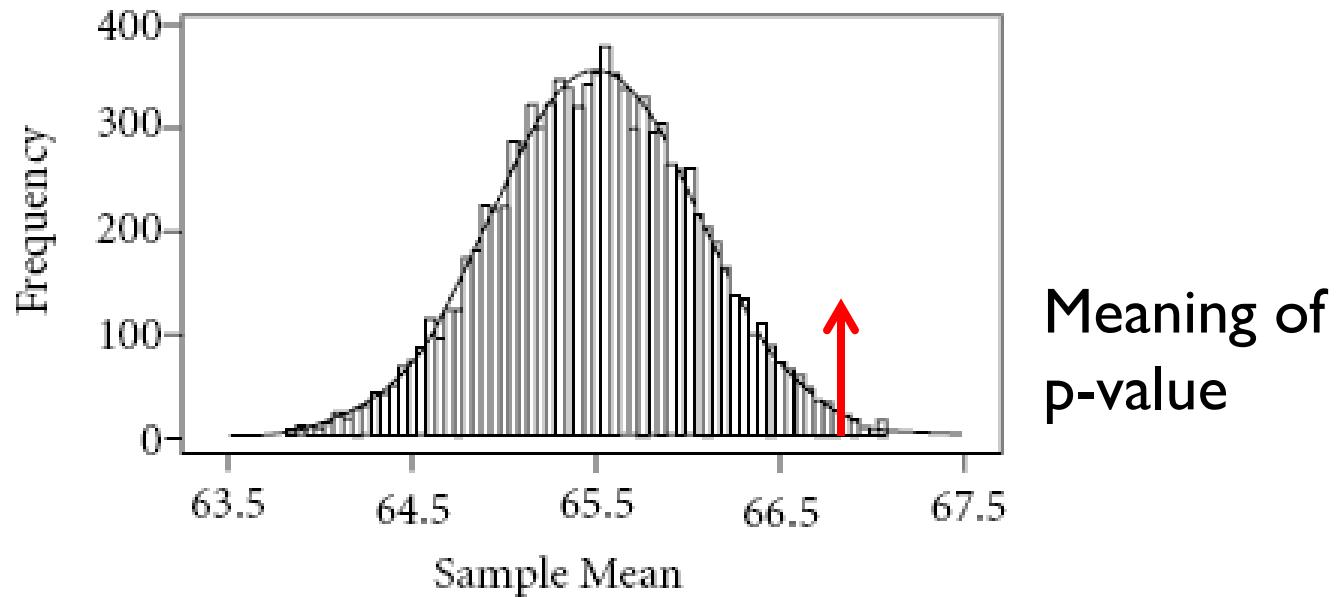


Parameter

Similar to Fourier Series, First used by Brahe for Alpha Aretis  
Good for comparison, but not appropriate for projection

# Distribution of the Sample Statistic/Moment (e.g. Mean)

Sample Size =20  
Number of samples=10k (from population)

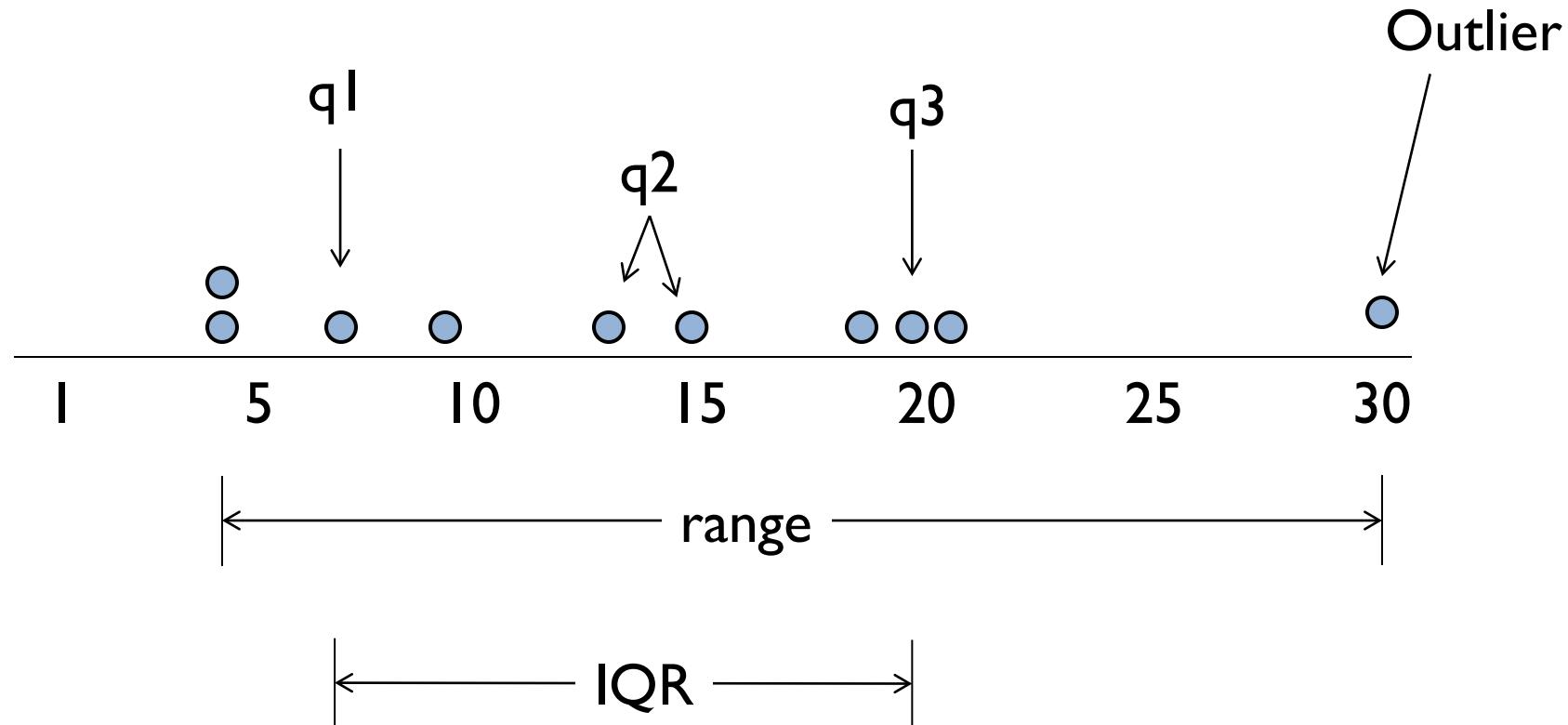


$$\begin{aligned}\mu_x &= \mu \\ \sigma_x &= \sigma / \sqrt{N}\end{aligned}$$

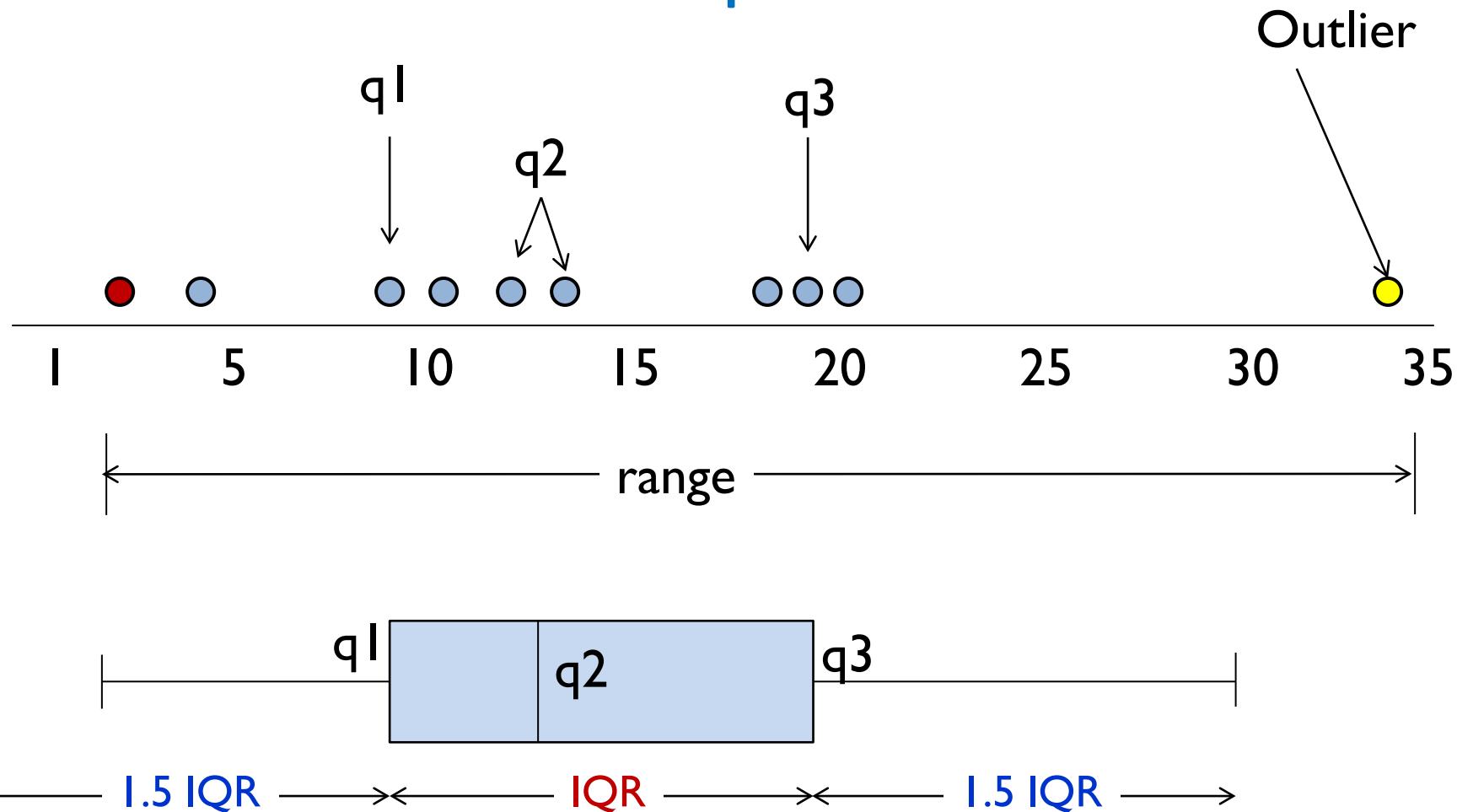
$$\begin{aligned}Z &= (X - \mu) / (\sigma / \sqrt{N}) & N > 30 \\ Z &= (X - \mu) / (s / \sqrt{N}) & N < 30\end{aligned}$$

# Problem with Sample Moments

## Quartiles and robust data description

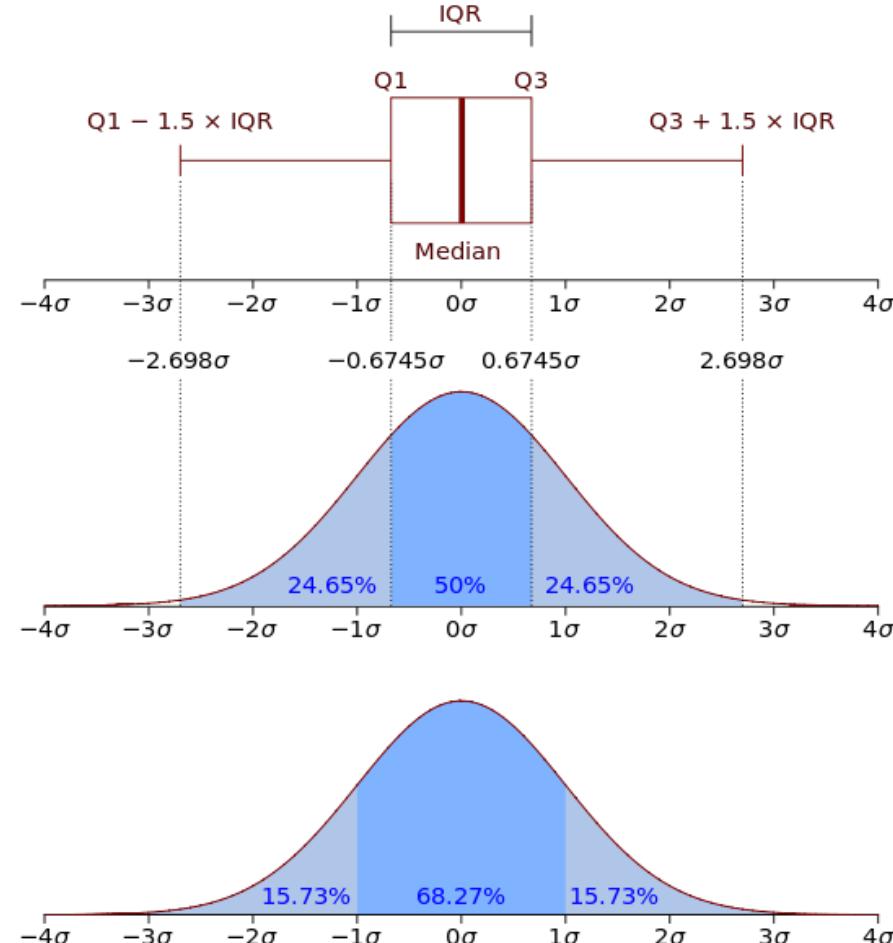


# Box plot



# Removing Outliers: $1.5 \times \text{IQR}$ Rule

Image from  
Wikipedia



Discrete  
Data

Continuous  
distribution

# Stem and leaf display: Pre-histogram

Order data

n=17

44 46 47 49 63 64 66 68 68 72 72 75 76 81 84 88 106

4 | 4679 ← Leaf

5 |

6 | 34688

7 | 2256

8 | 148

9 |

10 | 6



stem

$$L = 10 \times \log_{10} n = 10 \times \log_{10} 17 = 12.3 \sim 13$$

$$h_n = \left( \frac{\text{Range}}{L} \right) = \frac{106 - 44}{13} = 4.76$$

~ 10 rounded to 10 power

i.e. 40, 50, ... 90, 100

Should use the same approach for histogram  
Histogram should not increase precision

## Aside: Derivation of histogram size

Minimize:

$$MSE(x) = \int E[f_n(x) - f(x)]^2 dx$$

$$h_n = \left\{ \frac{6}{\int_{-\infty}^{\infty} [f'(x)]^2 dx} \right\}^{1/3} n^{-1/3}$$

$$h_n = 3.49 \times s \times n^{-(1/3)}$$

Freedman/Diaconis-1:

$$h_n = 1.66 \times s \times \left( \frac{\ln(n)}{n} \right)^{1/3}$$

Freedman/Diaconis-2:

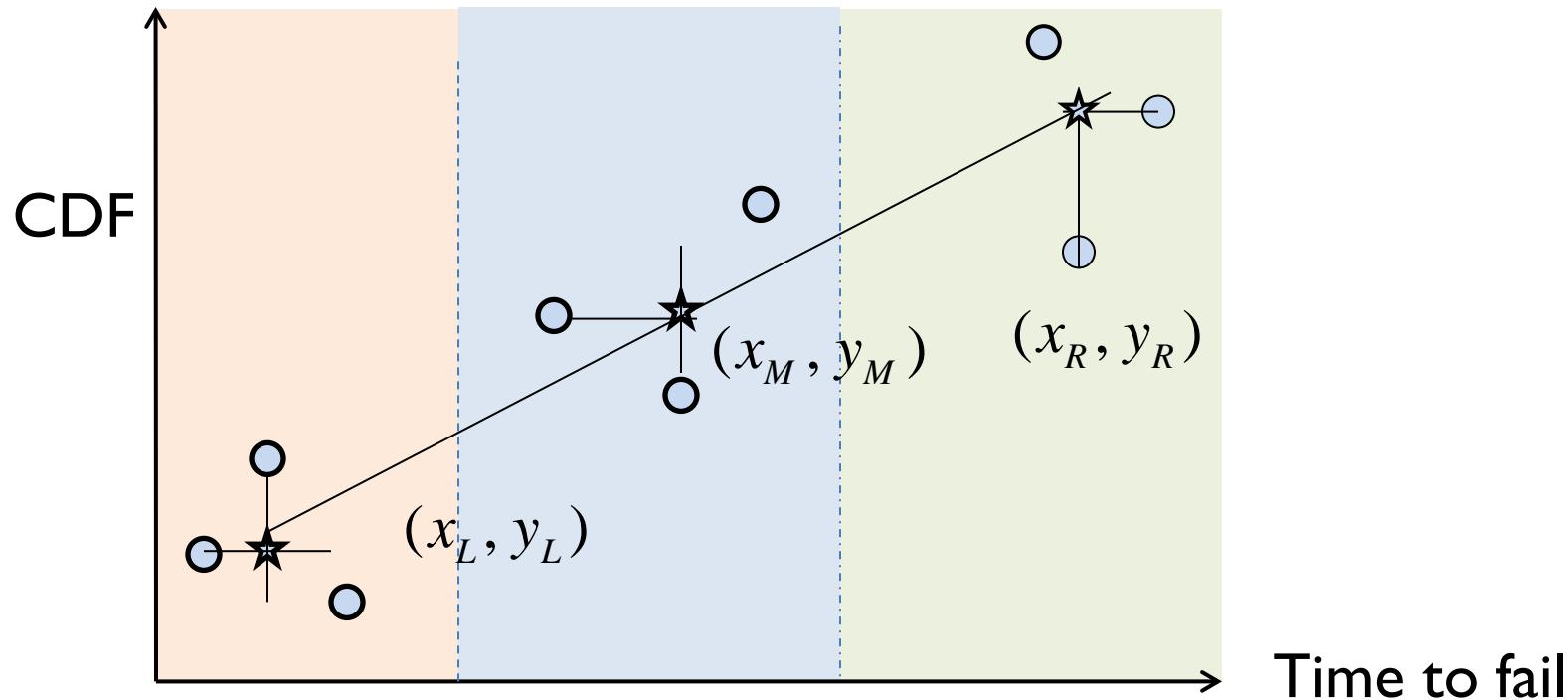
$$h_n = 2(IQR) \left( \frac{1}{n} \right)^{1/3}$$

Scott:

$$h_n = 3.49 \times s \times n^{-(1/3)}$$

Choose any of these formula, but remain consistent

# Drawing lines resistant to outliers



Divide the data into three groups, i.e.

For  $n=3k$  (  $k, k$ , and  $k$  )

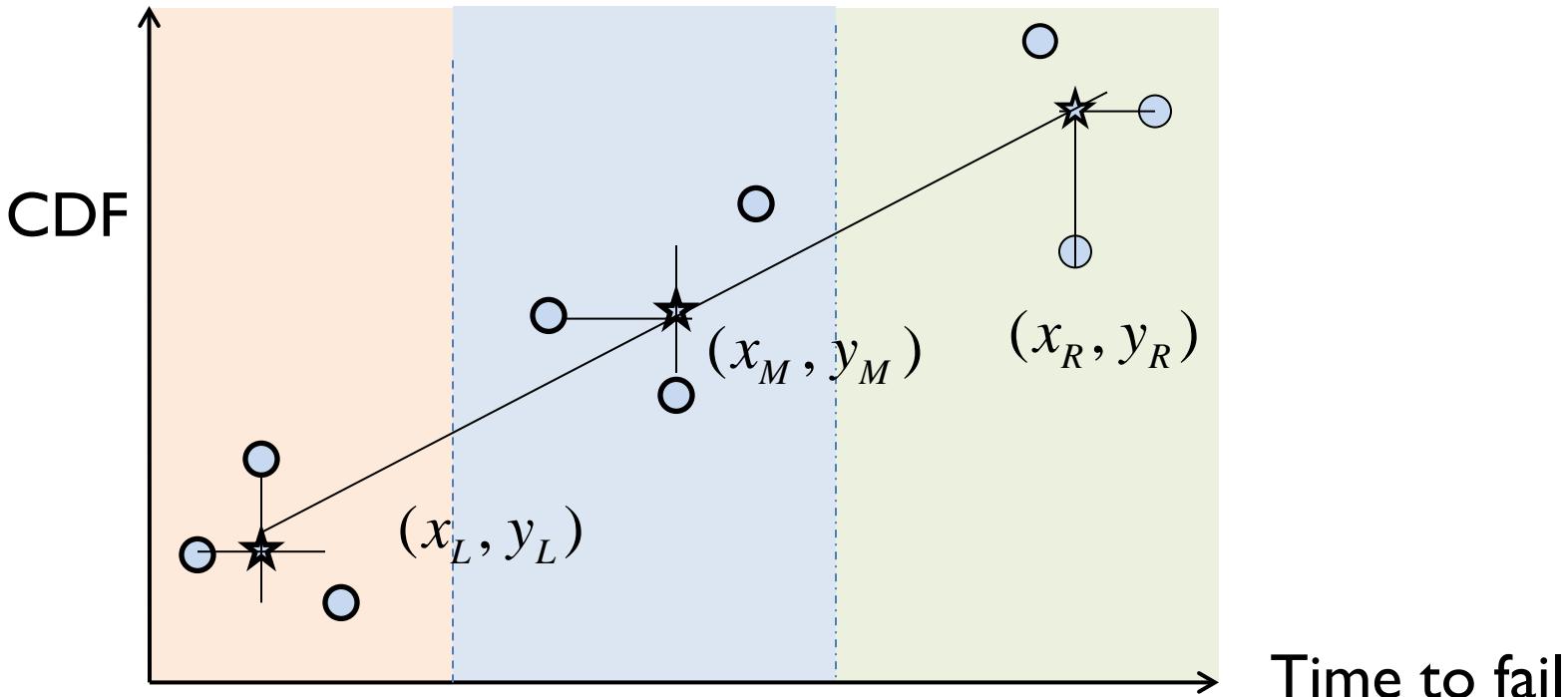
For  $n=3k+1$  (  $k, k+1, k$  )

For  $n=3k+2$  (  $k+1, k, k+1$  )

Calculate the median ( $x, y$ ) of each group.

Draw the line.

# Drawing lines resistant to outliers



$$y = b(x - x_M) + a$$

$$b_0 = (y_R - y_L) / (x_R - x_L)$$

$$3a_0 = [y_L - b_0(x_L - x_M)] + y_M + [y_R - b_0(x_R - x_M)]$$

$$r_i = y_i - [a_0 + b_0(x_i - x_0)]$$

$$a_1 = a_0 + \gamma_1 \quad b_1 = b_0 + \delta_1$$

# Problem of data plotting and numerical CDF

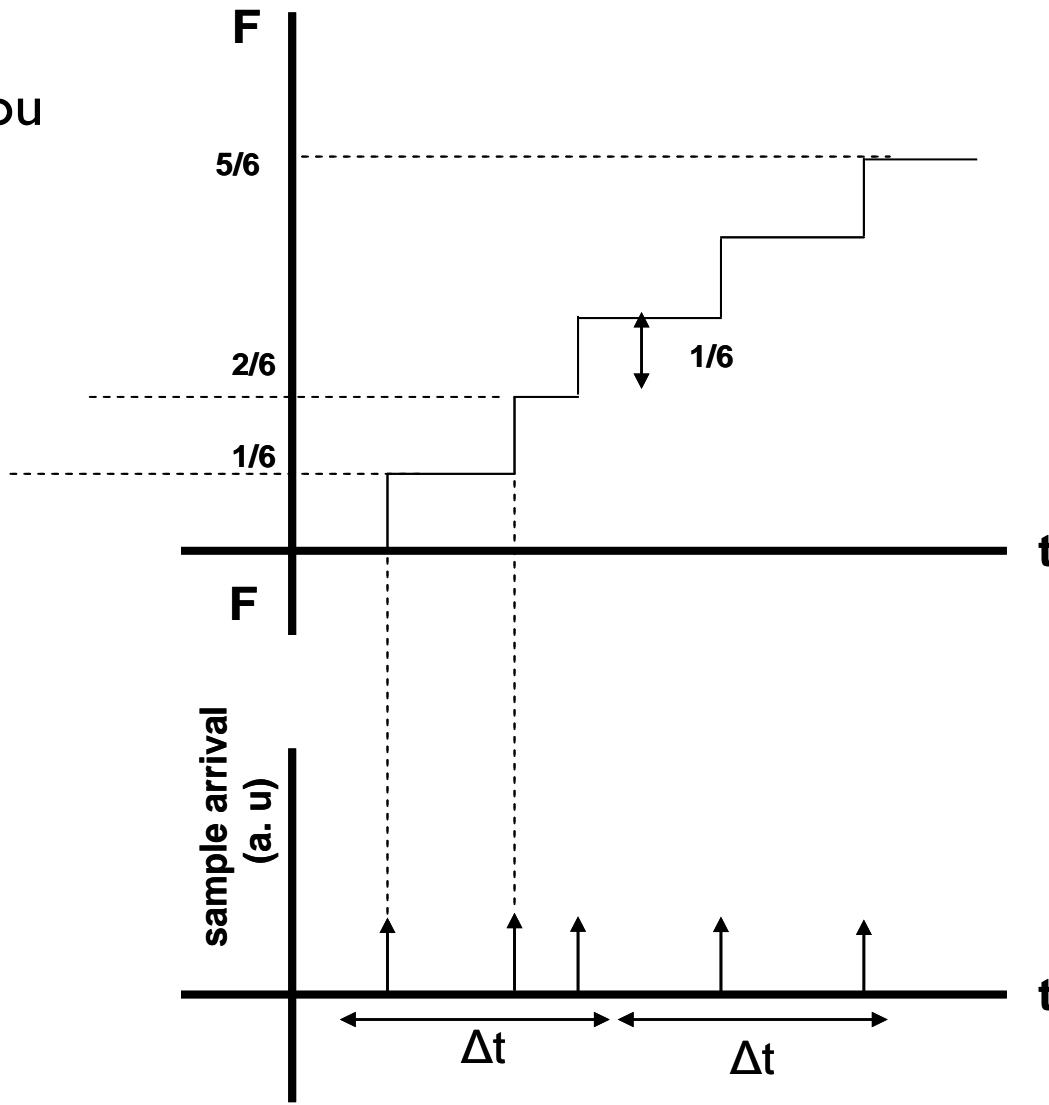
Assume you have 5 transistors and you have collected 5 breakdown times,  
 $t_1, t_2, t_3, t_4, t_5$

How do we find the CDF?

$$F_i = \frac{i}{n} \text{ or } F_i = \frac{i}{n+1}?$$

$$F_1 = \frac{1}{6} \quad F_2 = \frac{2}{6} \quad F_3 = \frac{3}{6} \quad F_4 = \frac{4}{6} \quad F_5 = \frac{5}{6}$$

$$W = \ln(-\ln(1 - F_i))$$



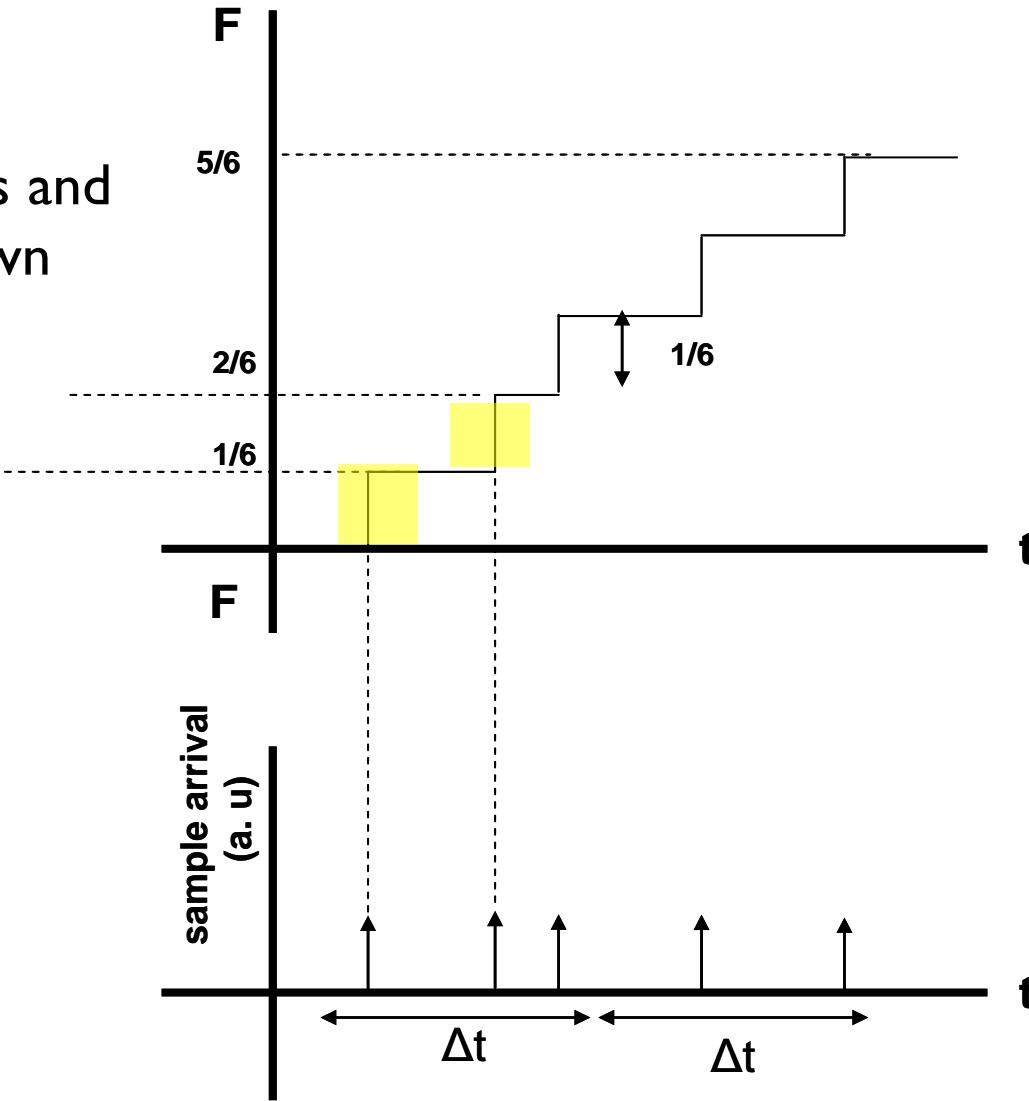
# ... there is a problem (Failure time is statistical)

Assume you have 5 transistors and you have collected 5 breakdown times,  $t_1, t_2, t_3, t_4, t_5$

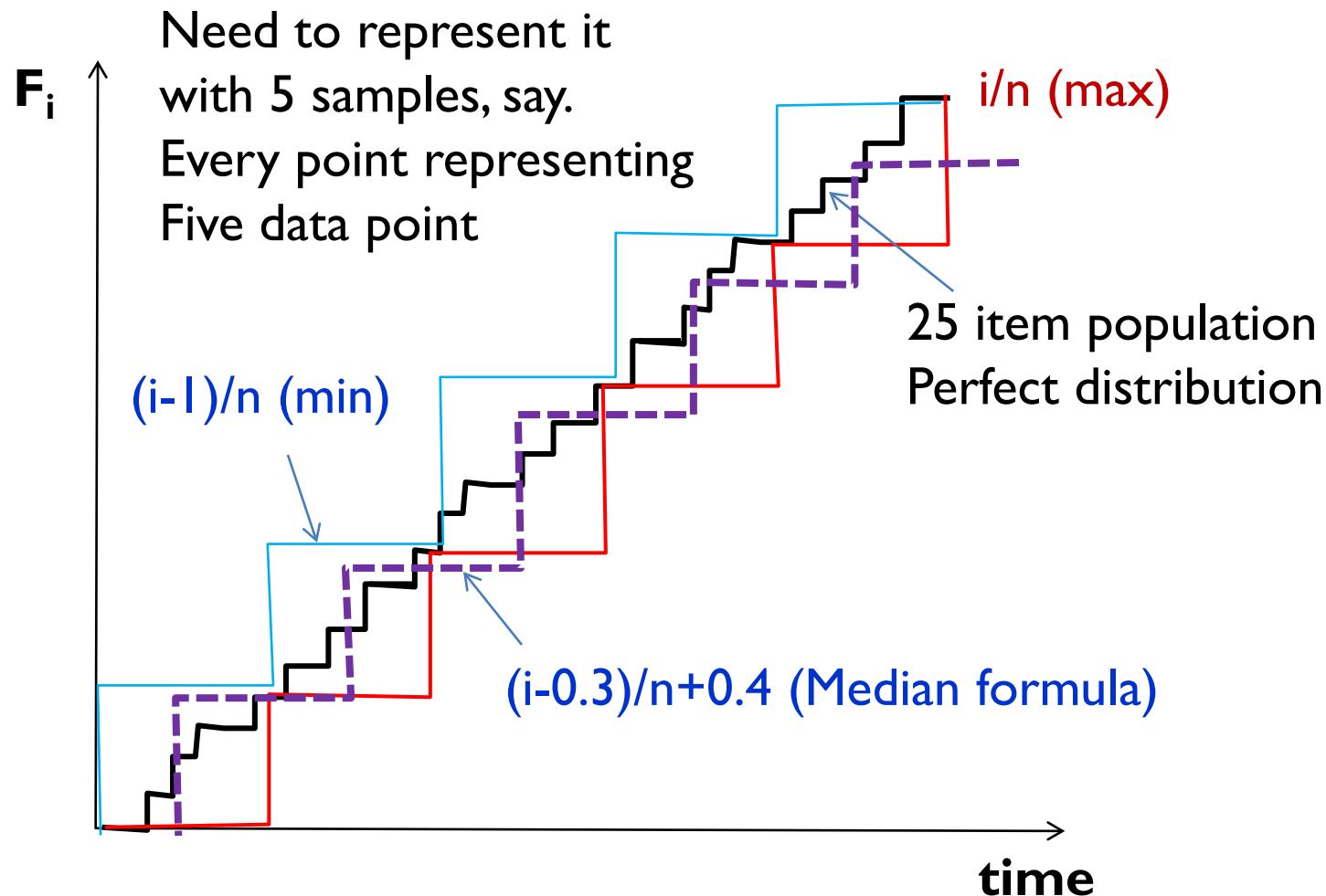
How do we find the CDF?

$$F_i = \frac{i - \alpha}{n - 2\alpha + 1}$$

$$W = \ln(-\ln(1 - F_i))$$



# Relationship among various formula

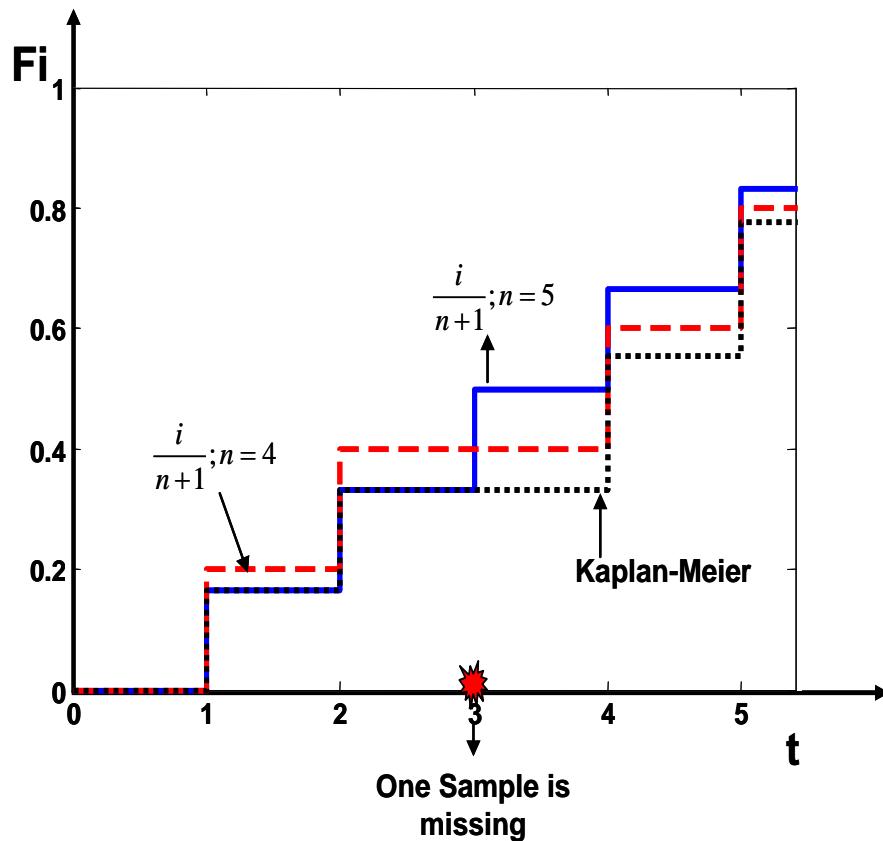
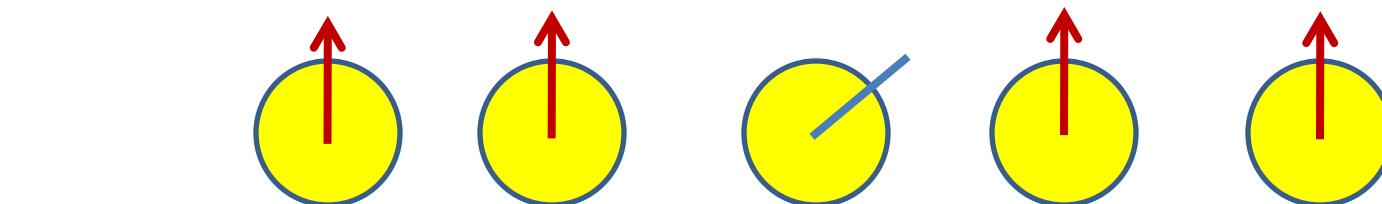


Analogous to a congressman ...

# Outline

- I. Origin of data, Field Acceleration vs. Statistical Inference
2. Nonparametric information
3. Preparing data for projection: Hazen formula
4. Preparing data for projection: Kaplan formula
5. Conclusions

# Censored data and imperfect sampling



$$F_i = \frac{i - \alpha}{n - 2\alpha + 1} \quad F_i = \frac{i}{n + 1}$$

$$F_1 = \frac{1}{6} \quad F_2 = \frac{2}{6} \quad F_3 = \frac{3}{6} \quad F_4 = \frac{4}{6} \quad F_5 = \frac{5}{6}$$

With 4 data points now, most people would do .....

$$F_1 = \frac{1}{5} \quad F_2 = \frac{2}{5} \quad F_3^* = \frac{3}{5} \quad F_4^* = \frac{4}{5}$$

... but this would be wrong!

# Hazen (approximate) formula for censored data

$$F_1 = \frac{1}{5}$$

$$F_i = \frac{i}{n+1}$$

$$F_2 = \frac{2}{5}$$

Loss of a sample  
 $N=4$

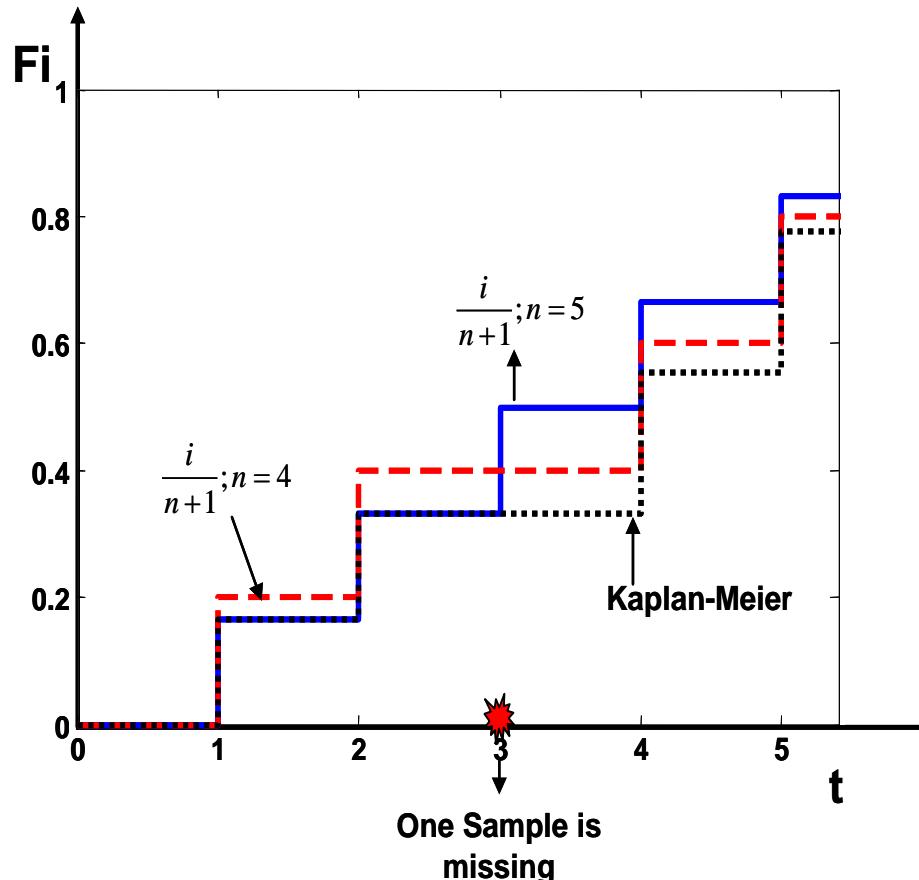
$$F_3 = \frac{2}{5}$$

but the sample  
did survive till  
 $t_2 \dots$

$$F_4 = \frac{3}{5}$$

$$F_5 = \frac{4}{5}$$

5 data-points, same as before, but with effective reduction in sample size  
Past data affected by future problems ... does not seem correct



# Kaplan-Meier (proper) Formula

$$F_i = 1 - \left( \frac{n - \alpha + 1}{n - 2\alpha + 1} \right) \prod_{i=1}^f \left( \frac{n_{si} + 1 - \alpha}{n_{si} + 2 - \alpha} \right)$$

Total number of samples      Number of surviving samples after time  $t_i$



Assume  $\alpha=0$ , so that

$$F_i = 1 - \prod_{i=1}^f \left( \frac{n_{si} + 1}{n_{si} + 2} \right)$$

# For uncensored traditional data ...

$$F_i = 1 - \prod_{i=1}^f \left( \frac{n_{si} + 1}{n_{si} + 2} \right)$$

$$F_1 = 1 - \frac{5}{6} = \frac{1}{6}$$

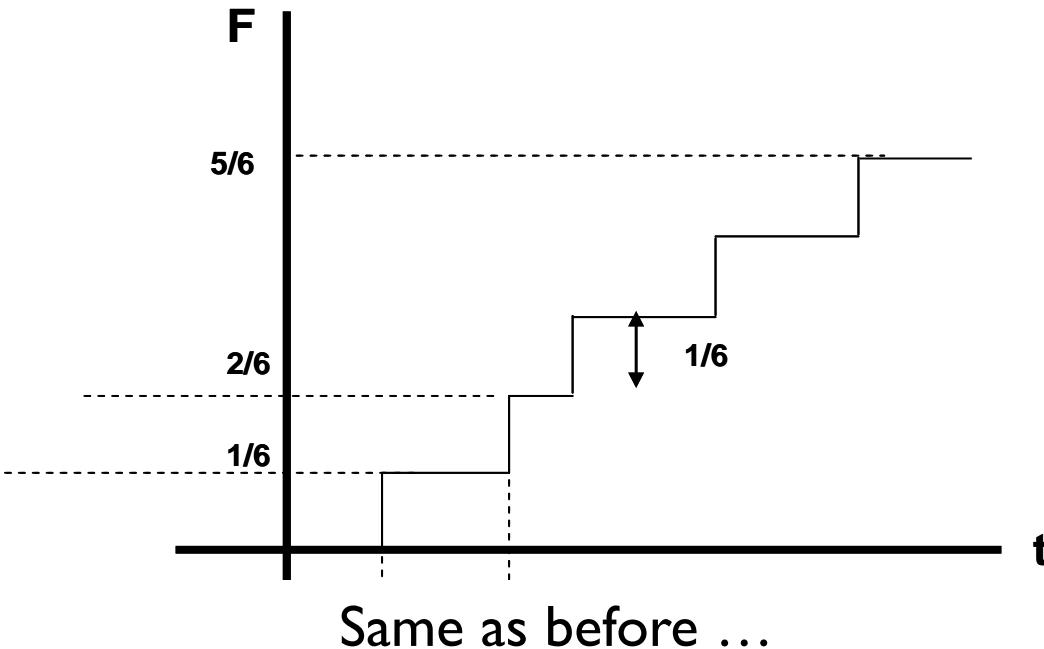
$$F_2 = 1 - \left( \frac{5}{6} \right) \cdot \left( \frac{4}{5} \right) = \frac{2}{6}$$

$$F_3 = 1 - \frac{5}{6} \cdot \frac{4}{5} \cdot \frac{3}{4} = \frac{3}{6}$$

$$F_4 = 1 - \frac{5}{6} \cdot \frac{4}{5} \cdot \frac{3}{4} \cdot \frac{2}{3} = \frac{4}{6}$$

$$F_5 = 1 - \frac{5}{6} \cdot \frac{4}{5} \cdot \frac{3}{4} \cdot \frac{2}{3} \cdot \frac{1}{2} = \frac{5}{6}$$

$n_{si}$ before $t_i$	5	4	3	2	1
$n_{si}$ after $t_i$	4	3	2	1	0

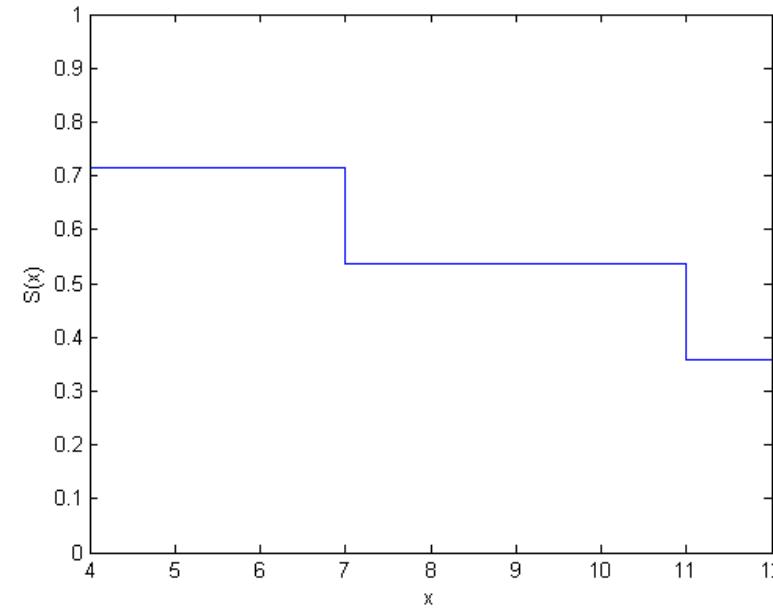


# MATLAB Routine for Censored Data

## Kaplan-Meier algorithm

```
y = [4 4 4 7 11 11 12];  
cens = [0 1 0 0 1 0 0];  
[f,x] = ecdf(y,'censoring',cens)
```

```
figure()  
ecdf(y,'censoring',cens,'function','survivor');
```



Survival function

# Conclusions

1. Treat your data with respect! They have stories to tell. A photon on your window may have the memory of a galaxy.
2. Focus on non-parametric data analysis. Simple non-parametric estimates like mean, standard deviation, median are all useful indicators that helps selecting appropriate distribution functions.
3. Non parametric plotting of distribution function is very important. Censored and uncensored data have very different plotting approaches. Outliers distort, therefore, median-based techniques is often useful.

# An Introduction to Data Analysis, Design of Experiment, and Machine Learning

## *Lecture 3. Physical and Empirical Distributions*

Muhammad A. Alam  
[alam@purdue.edu](mailto:alam@purdue.edu)



# Outline

- I. Physical Vs. empirical distribution**
- 2. Properties of classical distribution function**
- 3. Moment-based fitting of data**
- 4. Conclusions**

# Data vs. Hypothesis

Outliers identified  
(box-plot, Chauvenet)

Trend identified using median based plotting, stem-leaf histogram

CDF plotted using Kaplan-Meier formula

Non-parametric bootstrap to identify parameter uncertainty



Empirical reliability  
(Hypothesis testing)

Statistical reliability  
(Series/parallel systems)

Physical Reliability  
(Distribution function, prediction of an analytical model)

# Statistical Distribution is Physical

## Experiments



Weibull  
Log-normal  
Normal  
exponential

Fish-in-a-river  
(Inverse exp.)  
Fermi distribution  
Bose-Einstein Dist

(coin-flip)  
unknown

Flutter of  
Saturn probe  
(unknown)

If a problem can be mapped into one of the well known family,  
large number of results are available.

# Outline

1. Physical Vs. empirical distribution
2. Parametric Vs. non-parametric fits
3. Estimating various distribution functions
4. Conclusions

# Choosing distribution function

People choose functions that describe wide range of phenomena

- **Normal:** After all, everything eventually becomes normal (not really!) Distribution of last resort.
- **Log-Normal:** A variant of normal distribution that seems to describe many reliability problems phenomenologically (correlated processes, such as electromigration in interconnects, shunt distribution in solar cells)
- **Weibull:** Many physical systems are described by it.  
In the limiting case, it becomes Exponential distribution (extreme value problems such as thin oxide breakdown)

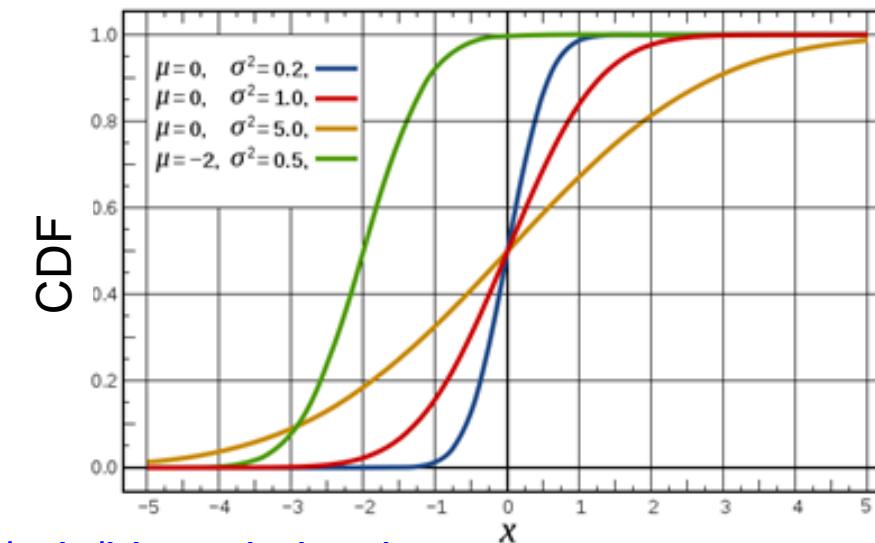
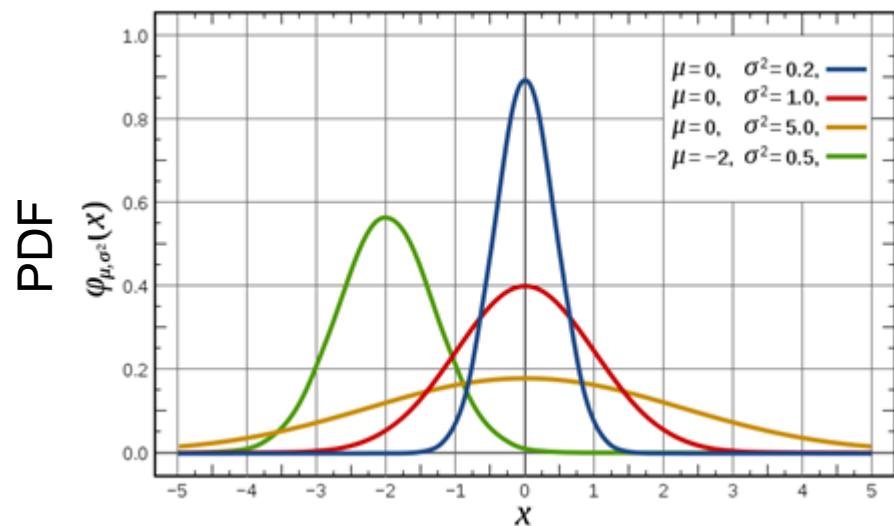
# Two parameter family: Normal distribution

$$f(t; \mu, \sigma) = \frac{1}{\sigma \sqrt{2\pi}} \cdot \exp \left[ -\frac{(t - \mu)^2}{2\sigma^2} \right]$$

$$F(t) = \Phi(\sigma^{-1}(t - \mu)) \quad \Phi(z) = [1 + \operatorname{erf}(z/\sqrt{2})]/2$$

$\mu$ =average,  $\sigma$ =standard deviation

Binomial distribution, Poisson distribution, chi-square, student-t distribution ...



[http://en.wikipedia.org/wiki/Normal\\_distribution](http://en.wikipedia.org/wiki/Normal_distribution)

## Two parameter family: log-normal distribution

$$(\text{PDF}) f(t; \mu, \sigma) = \frac{1}{t \times \sigma \sqrt{2\pi}} \cdot \exp \left[ -\frac{\{\ln(t) - \ln(\mu)\}^2}{2\sigma^2} \right]$$

$\mu$ =average,  $\sigma$ =standard deviation

$$(\text{CDF}) F(t) = \Phi \left( \sigma^{-1} \ln \frac{t}{\mu} \right) \quad \Phi(z) = [1 + erf(z/\sqrt{2})]/2$$

$$\sigma = \ln(t_2/t_1) / [\Phi^{-1}(F(t_2)) - \Phi^{-1}(F(t_1))]$$

$$= \ln(t_2 @ F = 0.5 / t_1 @ F = 0.159)$$

$$\lambda(t) = \sqrt{\frac{2}{\pi}} \frac{1}{t\sigma} \frac{\exp[-\sigma^2 \{\ln(t/\mu)\}^2/2]}{erf \left\{ \sqrt{0.5} \ln(t/\mu)/\sigma \right\}}$$

[http://en.wikipedia.org/wiki/Log-normal\\_distribution](http://en.wikipedia.org/wiki/Log-normal_distribution)

# Two parameter family: Weibull distribution

$$f(t; \alpha, \beta) = \frac{\beta}{\alpha^\beta} \cdot t^{\beta-1} \cdot e^{-(t/\alpha)^\beta} \quad (\alpha, \beta > 0)$$

$$F(t) = 1 - \exp(-(t/\alpha)^\beta)$$

$$\lambda(t) = \frac{\beta t^{\beta-1}}{\alpha^\beta}$$

$\beta$ =shape parameter,  $\alpha$ =scale parameter

$\beta=1$  .... Exponential distribution

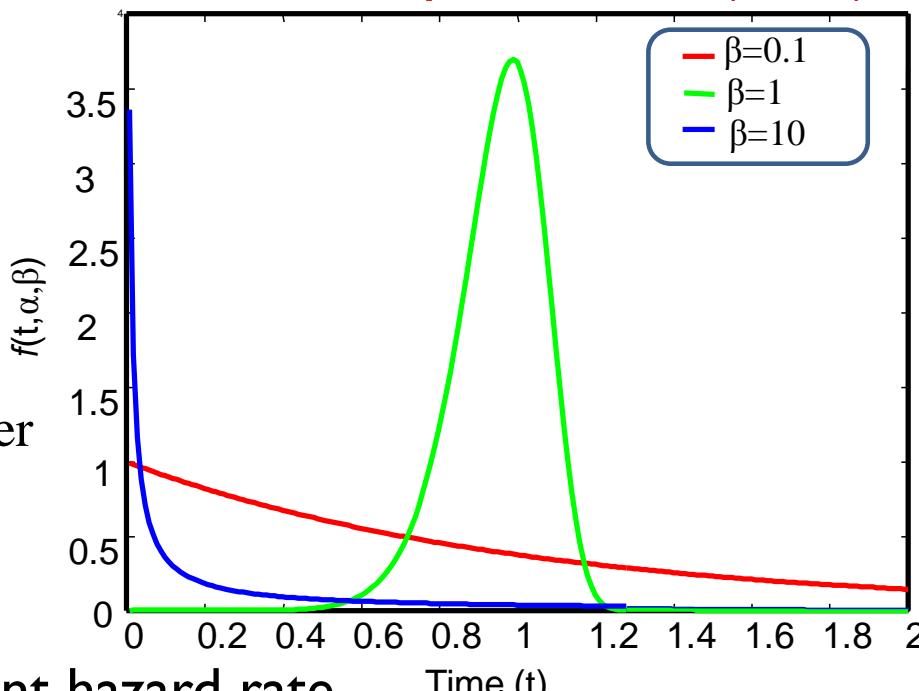
Memory-less distribution, constant hazard rate

$\beta=2$  .... Rayleigh distribution

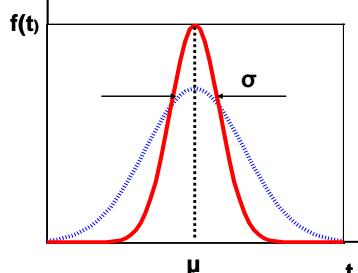
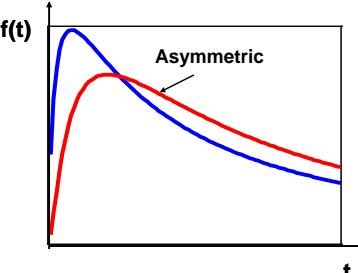
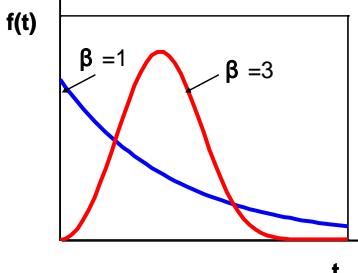
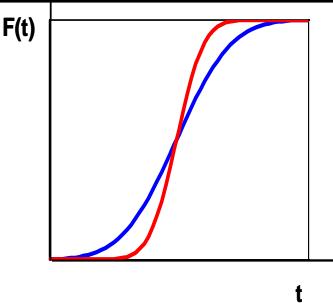
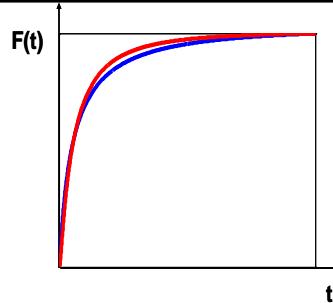
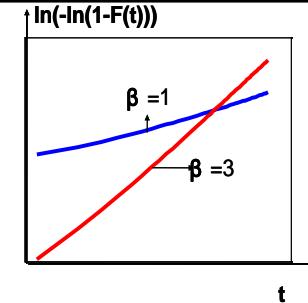
Light scattering, Corrosion in contacts, Failure rate increases with time

[http://en.wikipedia.org/wiki/Weibull\\_distribution](http://en.wikipedia.org/wiki/Weibull_distribution)

Abrahmi recrystallization (1905)



# Empirical statistical distributions

	Normal	Log Normal	Weibull
PDF	$\frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(t-\mu)^2}{2\sigma^2}}$	$\frac{1}{t\sigma\sqrt{2\pi}} e^{-\frac{(\ln t - \ln \mu)^2}{2\sigma^2}}$	$\frac{\beta}{\alpha} \cdot \left(\frac{t}{\alpha}\right)^{\beta-1} \cdot e^{-\left(\frac{t/\alpha}{\beta}\right)^\beta}$
PDF			
CDF			
Moment 1 <sup>st</sup>	$\mu$	$\mu \cdot e^{-\frac{\sigma^2}{2}}$	$\alpha \sqrt{\frac{1}{\beta}}$
Moment 2 <sup>nd</sup>	$\sigma^2$	$2\mu \cdot (e^{\sigma^2} - 1) \cdot e^{\sigma^2}$	$\sqrt{\alpha^2 \sqrt{1 + \frac{2}{\beta}} - \alpha^2 \Gamma^2 (1 + \frac{1}{\beta})}$

# Definitions of distribution functions

Name	Symbol	Expression
Prob. distribution	$f(t; \alpha, \beta, \dots)$	$f(t; \alpha, \beta, \dots)$
cumulative PDF	$F(t)$	$\int_{-\infty}^t f(t') dt'$
survival function	$R(t)$	$1 - F(t)$
hazard rate	$\lambda(t)$	$\frac{f(t)}{1 - F(t)}$
cum. hazard rate	$H(t)$	$\int_0^t \lambda(t') dt'$
average hazard	$\lambda_c(t)$	$1/t \int_0^t \lambda(t') dt'$

These functions are used in difference fields in different ways ...

# Discrete Transform: because data is discrete

$$F_i = \frac{i - \alpha}{n - 2\alpha + 1}$$

$$f_i = \frac{dF_i}{dt} = \frac{F_{i+1} - F_i}{t_{i+1} - t_i} = \frac{1}{(n - 2\alpha + 1)(t_{i+1} - t_i)}$$

$$\lambda_i = \frac{f_i}{1 - F_i} = \frac{1}{(n - i - \alpha + 1)(t_{i+1} - t_i)}$$

# Transformation among reliability functions

	$f(t)$	$F(t)$	$R(t)$	$\lambda(t)$	$H(t)$	$\lambda_c(t)$
$f(t)$	$f(t)$	$\frac{dF(t)}{dt}$	$-\frac{dR(t)}{dt}$	$\lambda(t)e^{-\int_0^t \lambda(t')dt'}$	$e^{-H(t)} \frac{dH(t)}{dt} \left( \lambda_c + t \frac{d\lambda_c(t)}{dt} \right) e^{-\lambda_c(t)t}$	
$F(t)$	$\int_0^t f(t')dt'$	$F(t)$	$1 - R(t)$	$1 - e^{-\int_0^t \lambda(t')dt'}$	$1 - e^{-H(t)}$	$1 - e^{-\lambda_c(t)t}$
$R(t)$	$\int_t^\infty f(t')dt'$	$1 - F(t)$	$R(t)$	$e^{-\int_0^t \lambda(t')dt'}$	$e^{-H(t)}$	$e^{-\lambda_c(t)t}$
$\lambda(t)$	$\frac{f(t)}{\int_t^\infty f(t')dt'}$	$-\frac{d \ln(1 - F(t))}{dt}$	$-\frac{d \ln R(t)}{dt}$	$\lambda(t)$	$\frac{dH(t)}{dt}$	$\lambda_c + t \frac{d\lambda_c(t)}{dt}$
$H(t)$	$-\ln\left(\int_t^\infty f(t')dt'\right)$	$-\ln(1 - F(t))$	$-\ln R(t)$	$\int_0^t \lambda(t')dt'$	$H(t)$	$t\lambda_c(t)$
$\lambda_c(t)$	$-\frac{1}{t} \ln\left(\int_t^\infty f(t')dt'\right)$	$\frac{-\ln(1 - F(t))}{t}$	$\frac{-\ln R(t)}{t}$	$1/t \int_0^t \lambda(t')dt'$	$\frac{H(t)}{t}$	$\lambda_c(t)$

HW: Derive a few reliability functions yourself ...

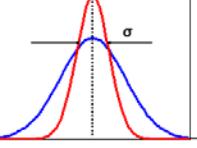
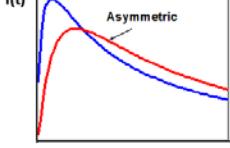
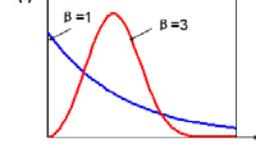
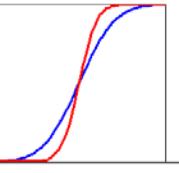
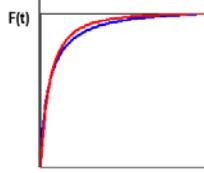
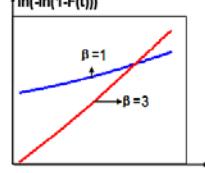
# Outline

- I. Physical Vs. empirical distribution
2. Properties of classical distribution function
3. Moment-based fitting of data
4. Conclusions

# Moment-based fitting

Of 60 oxides, 7 failed in 1000 hrs

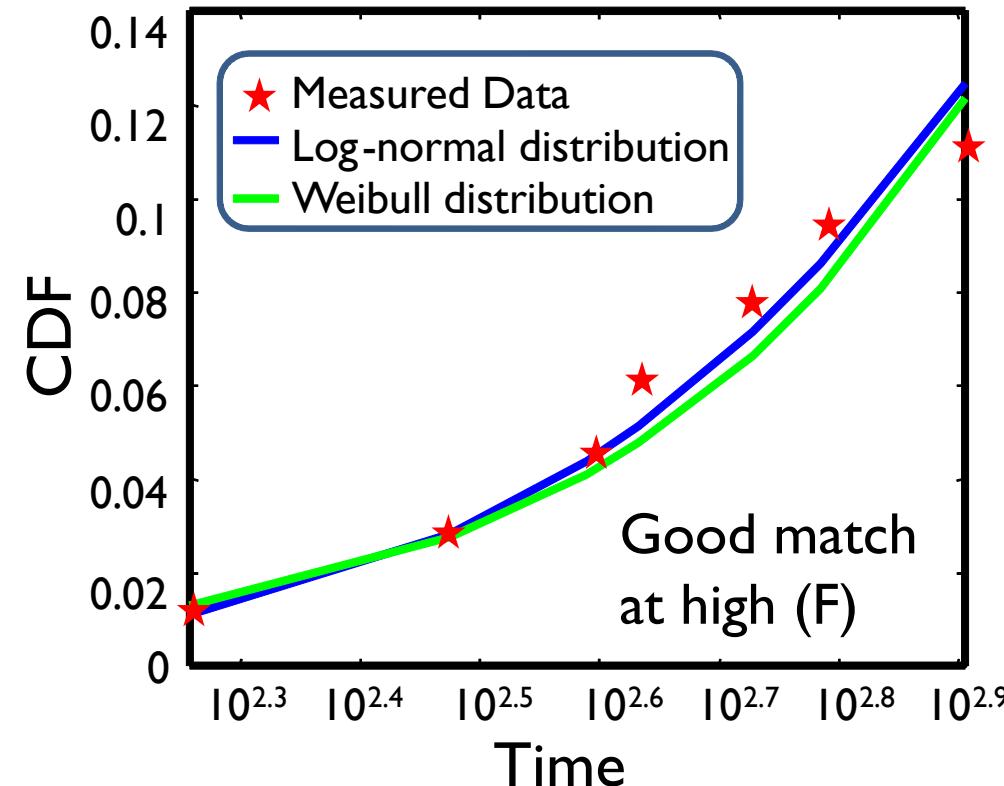
Rank	Lifetime	
1	181	0.012
2	299	0.028
3	389	0.045
4	430	0.061
5	535	0.078
6	610	0.094
7	805	0.111

	Normal	Log Normal	Weibull
PDF	$\frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(t-\mu)^2}{2\sigma^2}}$	$\frac{1}{t\sigma\sqrt{2\pi}} e^{-\frac{(\ln t - \ln \mu)^2}{2\sigma^2}}$	$\frac{\beta}{\alpha} \left(\frac{t}{\alpha}\right)^{\beta-1} \cdot e^{-\left(\frac{t}{\alpha}\right)^\beta}$
PDF			
CDF			
Moment 1 <sup>st</sup>	$\mu$	$\mu \cdot e^{-\frac{\sigma^2}{2}}$	$\alpha \sqrt{\frac{1}{\beta}}$
Moment 2 <sup>nd</sup>	$\sigma^2$	$2\mu \cdot (e^{\sigma^2} - 1) \cdot e^{\sigma^2}$	$\sqrt{\alpha^2 \sqrt{1 + \frac{2}{\beta}} - \alpha^2 \Gamma^2(1 + \frac{1}{\beta})}$

# Matching moments to distributions

Of 60 oxides, 7 failed in 1000 hrs

Rank	Lifetime	
1	181	0.012
2	299	0.028
3	389	0.045
4	430	0.061
5	535	0.078
6	610	0.094
7	805	0.111



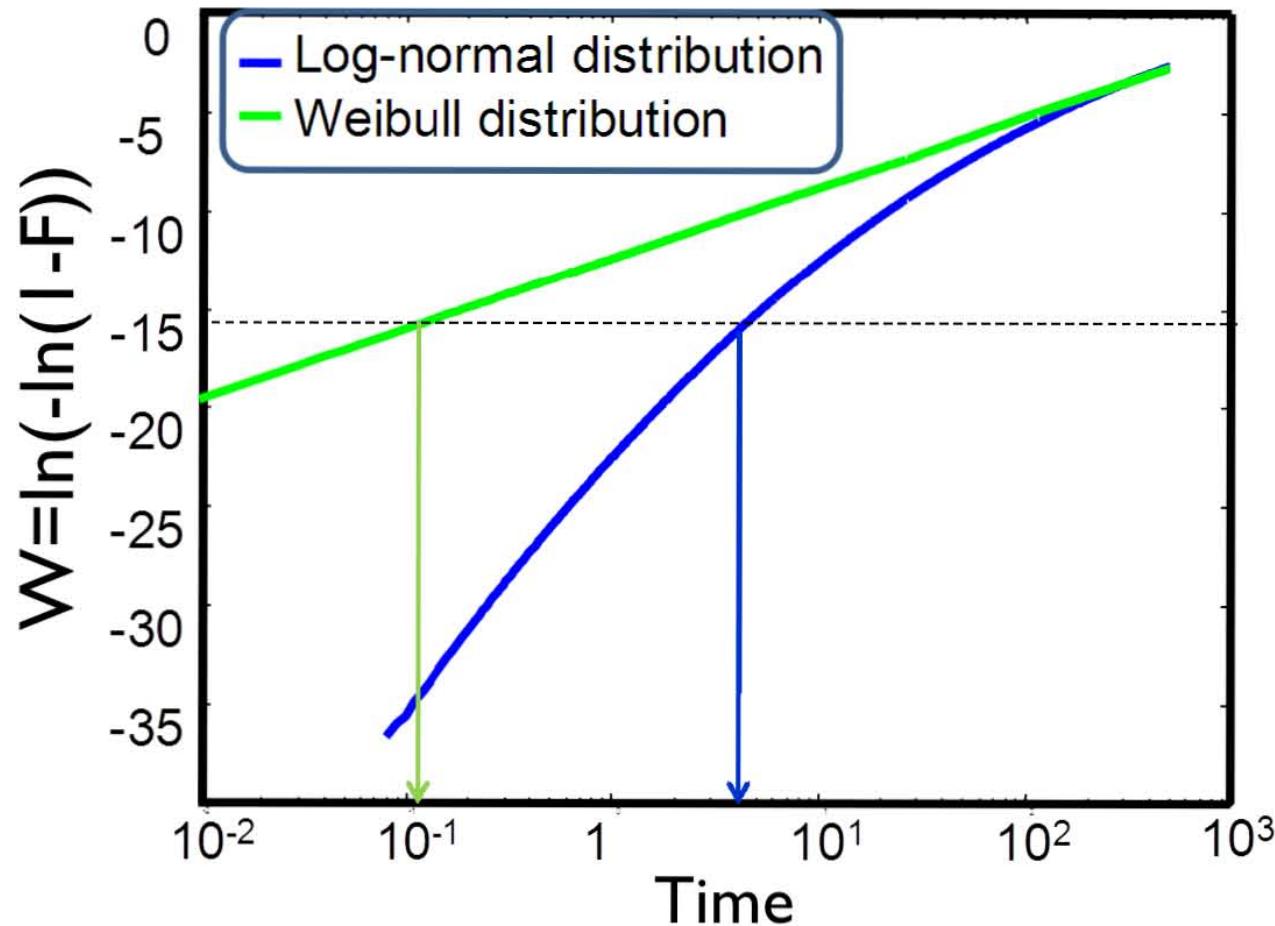
## Weibull Distribution Parameters

When  $t=\alpha$ ,  $\ln(1-F(t))=-1$ ,  $F(t)=0.632$ ,  $\alpha=2990$   
 $\beta$  estimated using parameter fitting as 1.56

## Log-Normal Distribution Parameters

$s=\ln(T_{50\%}/T_{15.9\%})$ ,  $\sigma=\ln(3600/980)=1.30$   
 $\mu=\ln(T_{50\%})=\ln(3600)=8.19$

# Problem of matching the moments



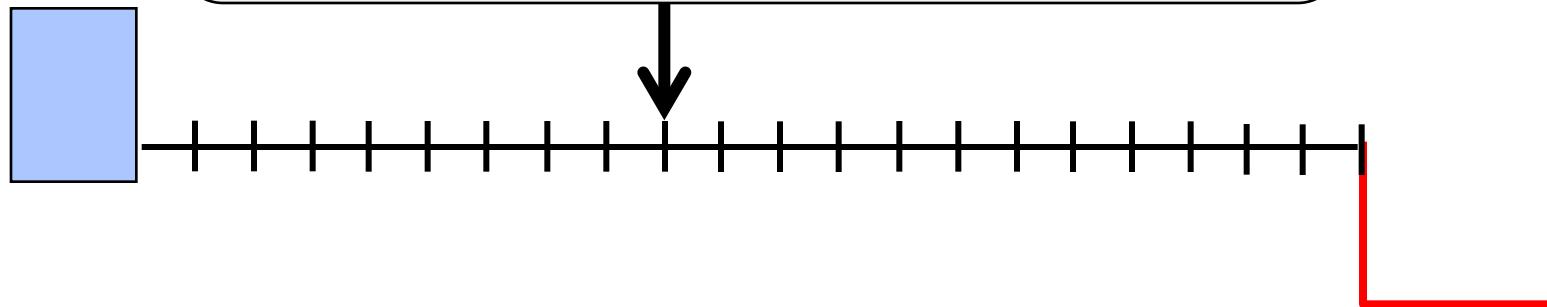
Log-normal distribution is considerably optimistic

# Problem of matching distribution: BFRW

$$\frac{\partial P}{\partial t} = D \frac{\partial^2 P}{\partial x^2}$$

$$P(x, t=0) = \delta(x - x_0)$$

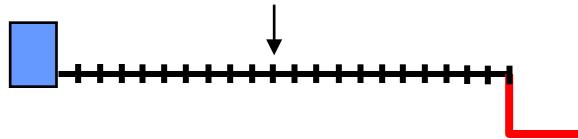
$$P(x=0, t) = 0$$



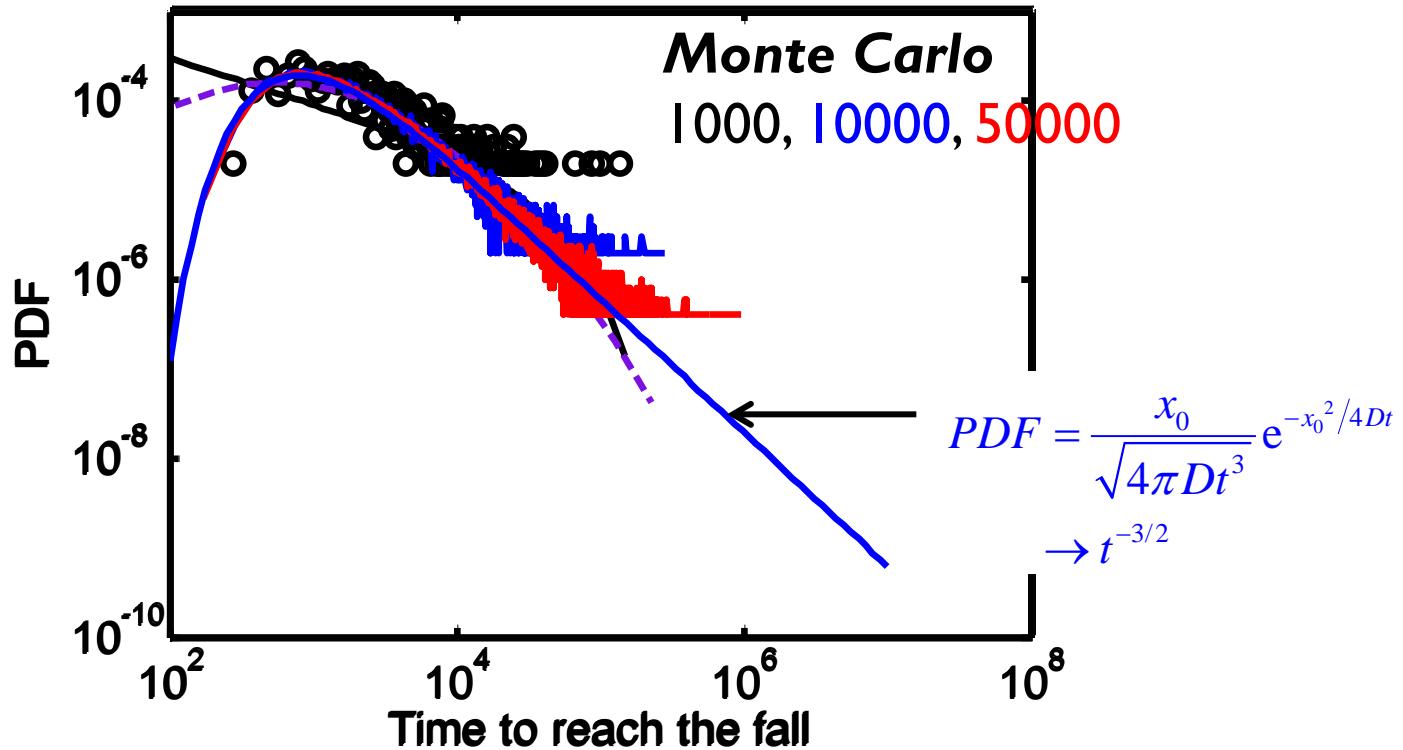
$$P(x, t) = (4\pi Dt)^{-1/2} \left[ e^{-(x - x_0)^2 / 4Dt} - e^{-(x + x_0)^2 / 4Dt} \right]$$

$$\int_0^t f(\tau) d\tau + \int_0^L P(x, t) dx = 1 \Rightarrow f(t) = \frac{x_0}{\sqrt{4\pi Dt^3}} e^{-x_0^2 / 4Dt}$$

# Match apparently reasonable, but wrong



$$f_G(t) = \frac{t^{k-1} e^{-t/\theta}}{\Gamma(k)\theta^k} \quad T_{avg} = k\theta$$



# Conclusions

1. Once the data is plotted using the principles discussed in the previous lectures, the phenomenon can be described by a statistical model.
2. If unsuccessful, one should choose functions with least number of variables that described the system. Many applications are described by 2-parameter distributions (e.g. log-normal, Weibull).
3. For an extreme value problem, one should pay particular attention to the tail of the distribution and choose sample size accordingly.
4. Moment-based methods are popular, but cannot distinguish between the tails of the distribution (associated with high moments)

# References

D. C. Hoaglen, F. Mosteller, and J.W.Tukey, “Understanding Robust and Exploratory Data Analysis”, Wiley Interscience, 1983. Explains the importance of Median based analysis when the dataset is small and the quality cannot be guaranteed.

Linda C. Wolsterholme, “Reliability Modeling – A Statistical Approach, Chapman Hall, CRC, 1999. Chapter 1-7 has excellent summary of ‘Goodness of Fit’ analysis.

R. H. Myers and D.C. Montgomery, “Response Surface Methodology”, Wiley Interscience, 2002. This book discusses design of experiment in great detail.

An excellent textbook that covers many topics discussed in this Lectures is Applied Statistics and Probability for Engineers, 3<sup>rd</sup> Edision, D.C. Montgomery and G. C. Runger, Wiley, 2003.

AT&T, “Statistical Quality Control Handbook”. Joan Fisher Box, “R. A. Fisher and the Design of Experiments, 1922-1926”, *The American Statistician*, vol. 34, no. 1, pp. 1-7, Feb. 1980.

F.Yates, “Sir Ronald Fisher and the Design of Experiments”, *Biometrics*, vol. 20, no. 2, In Memoriam: Ronald Aylmer Fisher, 1890-1962., pp. 307-321, (Jun. 1964).

Ranjith Roy, “A primer on the Taguchi Method”, Van Nostrand Reinhold International Co. Ltd., 1990.

Lloyd W Condra, “Reliability Improvement with design of experiments”, Marcel Dekker Inc., 1993.

# Review questions

- G1: Why do people use Normal, log-normal, Weibull distributions when they do not know the exact physical distribution?
- G2: What is the problem of using empirical distributions? What are the advantages?
- G3: If you must choose an empirical distribution, what should be your criteria? (Nos. of parameters, physical principles, etc.)
- G4: Why does everyone suggest the use of CDF for empirical data-fitting, rather than PDF? (Obviously one can go from one function to the other)
- G5: There are all sorts of distribution functions (e.g. survivability function) ? If everything is related to everything else, why do we need so many?
- G6: How would you determine the BFRW failure rates? Mean Hazard rate?

# Excellent resource at ....

## I. Statistics Online Computational Resource

<http://www.socr.ucla.edu/SOCR.html>

2. Excellent toolset within Excel
3. S and S-Plus software set
4. MATLAB has nearly everything!

# Parametric vs. non-parametric Bootstrap

0.2 -0.1 0.5 0.3 -0.6    Fit the distribution of your choice by  
Maximum likelihood estimators (MLE)  
(obtain parameters, i.e.  $\eta_0, \beta_0$ )

Generate synthetic samples based on the parametric distribution

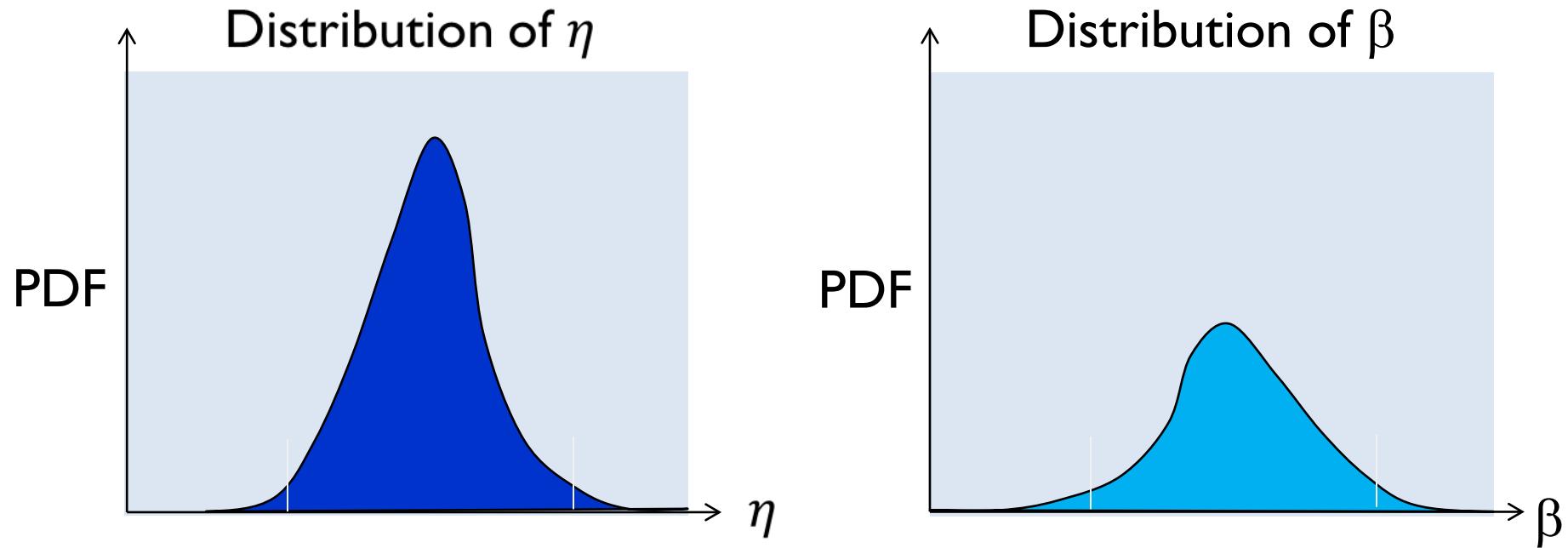
0.12 -0.17 -0.44 -0.71 0.52    Synthetic sample 1 (new  $\eta_1, \beta_1$ )

0.32 0.21 -0.69 0.23 0.58    Synthetic sample 2 (new  $\eta_2, \beta_2$ )



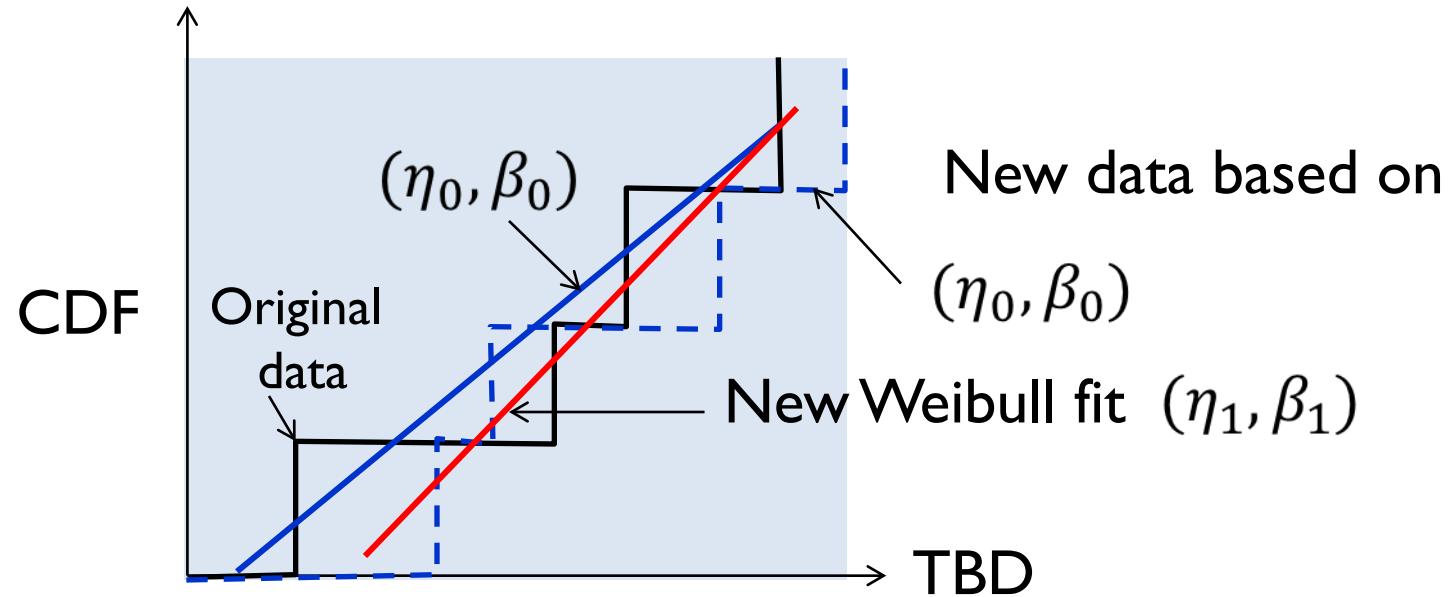
Plot distribution of statics  $\eta_i, \beta_i$

# Distribution of $\alpha$ and $\beta$



Same technique for polling and tenure rate of faculty!

# Why resampling from the same distribution generates new fit parameters



Samples taken from the same distribution  $(\eta_0, \beta_0)$  generates datapoints that are fitted with new  $(\eta_i, \beta_i)$

# References

1. “Detecting Novel Associations for large scale dataset”, D. Reshef et al., Science 334, p. 1418, 2011.
2. “Survival Analysis of Faculty Retention in Science and Engineering”, D. Kaminski et al., Science, 335, 864, 2012.
3. Clauset, C. R. Shalizi, and M. E. J. Newman, “Power-Law Distributions in Empirical Data,” SIAM Review, vol. 51, no. 4, p. 661, Nov. 2009.

# An Introduction to Data Analysis, Design of Experiment, and Machine Learning

## *Lecture 4. Model Selection and Goodness of Fit*

Muhammad A. Alam  
[alam@purdue.edu](mailto:alam@purdue.edu)



# Copyright 2018

This material is copyrighted by M. Alam under the  
following Creative Commons license:



Conditions for using these materials is described at

<http://creativecommons.org/licenses/by-nc-sa/2.5/>

# Course Outline

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

Lecture 1: Introduction

Lecture 2: Collecting and plotting  $x_1, x_2, \dots x_n$

Lecture 3: Physical and empirical  $f, F, df/dx, \dots$

Lecture 4: **Model selection between  $f_1, f_2, \dots$**

Lecture 5: Model Selection: Cross-validation and Bootstrapping method

Lecture 6: Scaling theory with known  $f$ ,  $f(\bar{x}) = f(\bar{X})$

Lecture 7: Scaling theory with unknown  $f$ ,  $\bar{x} \rightarrow X$

Lecture 8: Design of experiments to determine  $\bar{y}_{\max} = f(\bar{x})$

Lecture 9: DOE and ANOVA

Lecture 11: Principle component analysis for classifying  $\{y\}$ .

Lecture 12: Machine learning ... Statistical approach learn  $f$

Lecture 13: Machine learning ... Deep network, Karnaugh map, and other approaches

Lecture 14: Interpretable ML: Physics-based machine learning  $f = f_{\text{physics}} + \Delta f$

Lecture 15: Conclusions

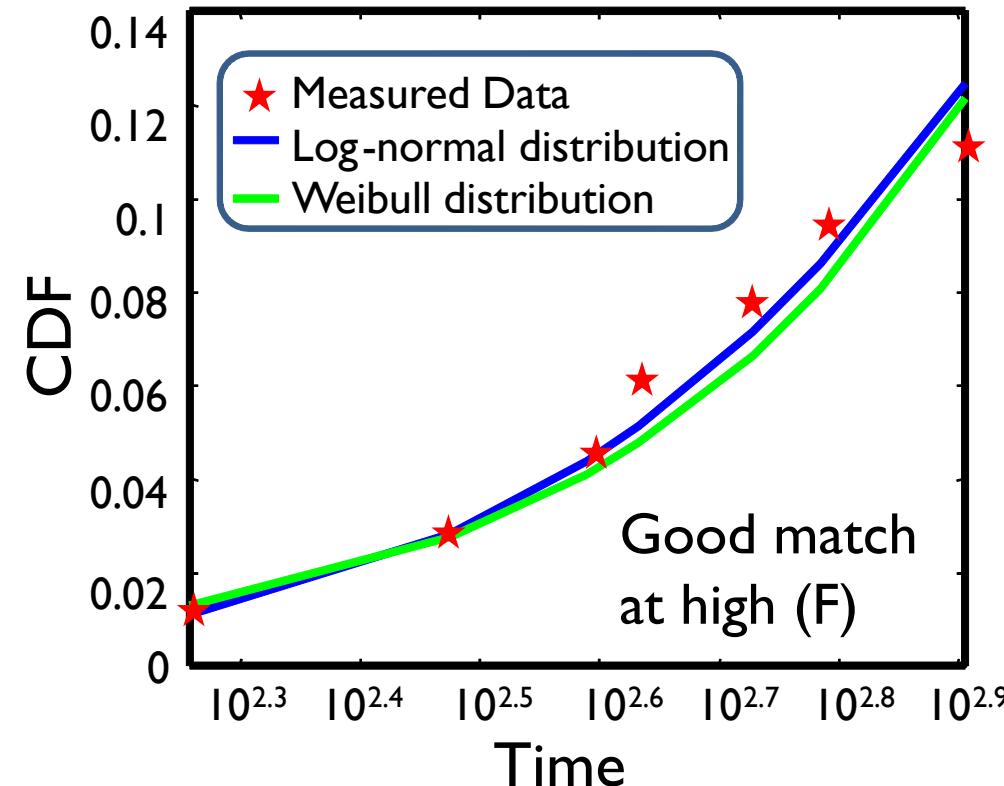
# Outline

1. The problem of matching data with theoretical distribution
2. Parameter extractions: Moments, linear regression, maximum likelihood
3. Goodness of fit: Residual, Pearson, Cox, Akika
4. Conclusion

# Matching moments to distributions

Of 60 oxides, 7 failed in 1000 hrs

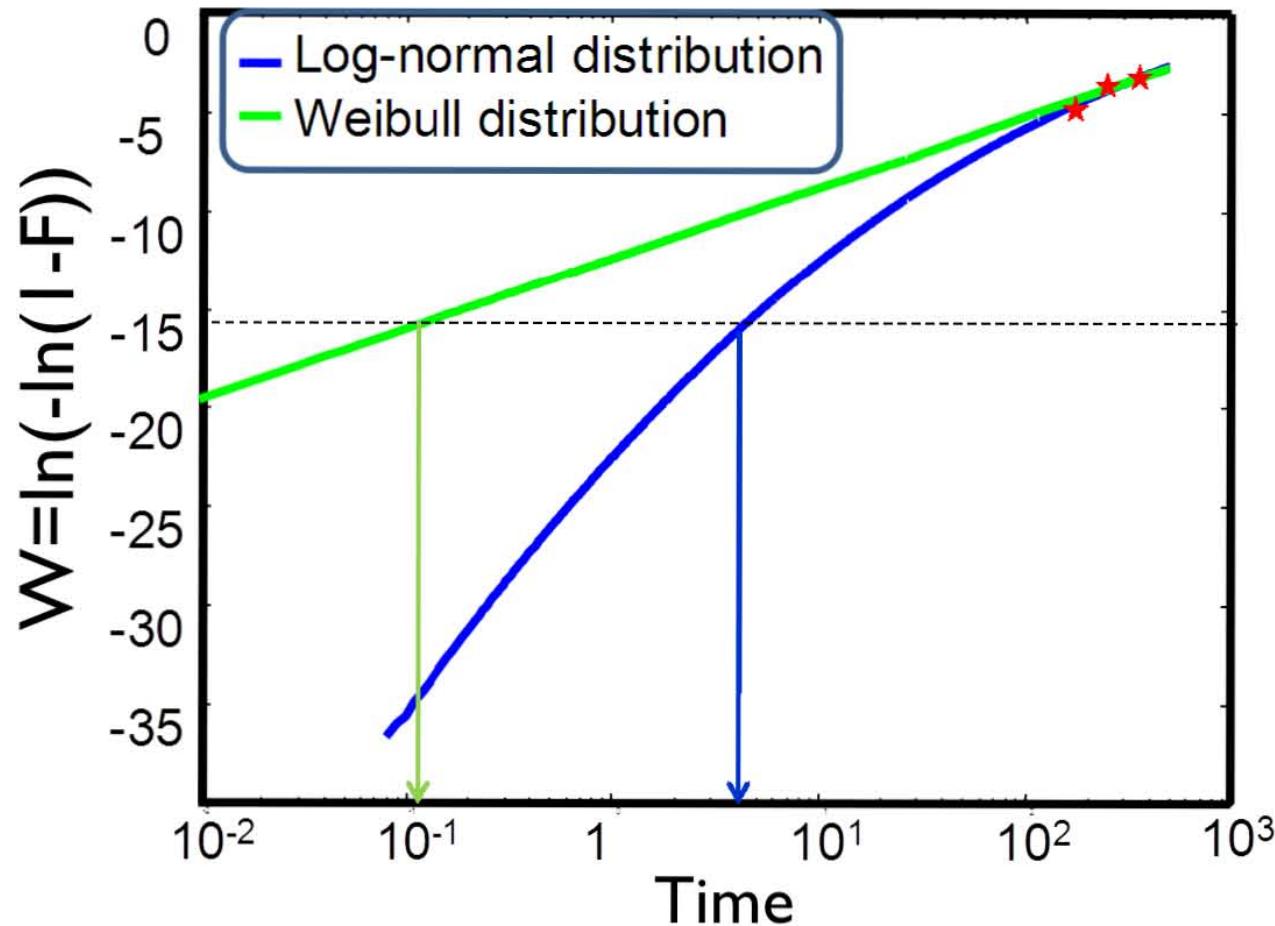
Rank	Lifetime	
1	181	0.012
2	299	0.028
3	389	0.045
4	430	0.061
5	535	0.078
6	610	0.094
7	805	0.111



Weibull Distribution Parameters  
When  $t=\alpha$ ,  $\ln(1-F(t))=-1$ ,  $F(t)=0.632$ ,  $\alpha=2990$   
 $\beta$  estimated using parameter fitting as 1.56

Log-Normal Distribution Parameters  
 $s=\ln(T_{50\%}/T_{15.9\%})$ ,  $\sigma=\ln(3600/980)=1.30$   
 $\mu=\ln(T_{50\%})=\ln(3600)=8.19$

# Problem of matching the moments



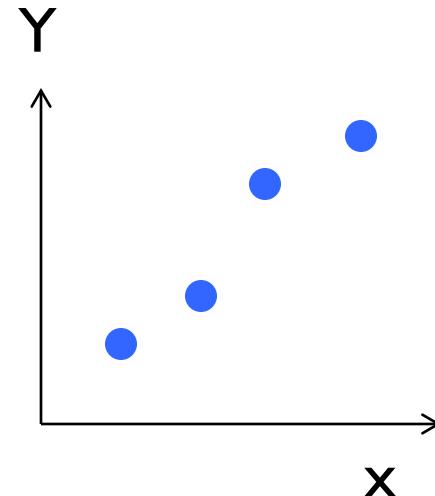
Log-normal distribution is considerably optimistic

# (1) Linear regression: balanced errors

$$W \equiv \ln(-\ln(1 - F)) = \beta \ln t + c$$

Theory:  $y = ax + b$

Data:  $y_i = ax_i + b$



Minimize  $SSR = \sum_i (y - y_i)^2$

$$a = \left( \sum x_i^2 \sum y_i - \sum x_i \sum x_i y_i \right) \times \textcolor{red}{D}^{-1}$$

$$b = \left( n \sum x_i y_i - \sum x_i \sum y_i \right) \times \textcolor{red}{D}^{-1}$$

$$\textcolor{red}{D} \equiv n \sum x_i^2 - \left( \sum x_i \right)^2$$

# Uncertainty in regression coefficients

Dependent variable subject to **random Gaussian Error** of same magnitude at each data point

Theory:  $y = ax + b$

$$\sigma_a^2 = s \left( \sum x_i^2 \right) \times \textcolor{red}{D}^{-1}$$

$$\sigma_b^2 = s \sqrt{n} \times \textcolor{red}{D}^{-1}$$

$$\textcolor{red}{D} \equiv n \sum x_i^2 - \left( \sum x_i \right)^2$$

$$s^2 = \sum (y - y_i)^2 / (n - 2)$$

t-distribution with (n-2) degree of freedom

$$a \pm t_{95\%,(n-2)} \sigma_a$$

$$b \pm t_{95\%,(n-2)} \sigma_b$$

\* Note s and (n-2) ... So called Bessel correction, Because we needed data to calculate a and b.

# Methods of least squares for weibull

$T_i$  obtained from measurement,

$F_i$  obtained from Hazen or Kaplan-Meier formula.

Define  $E(\alpha, \beta) = \sum_i (F_{i,\text{exp}}(t_i) - F_{i,\text{theroy}}(t_i, \alpha, \beta))^2$

Minimize  $\frac{dE}{d\alpha} = 0, \quad \frac{dE}{d\beta} = 0$

Error and Residual ...

$E(\alpha_0, \beta_0) = \sum_i (F_{i,\text{exp}}(t_i) - F_{i,\text{theroy}}(t_i, \alpha_0, \beta_0))^2$

# Fitting of physical models: challenges

Is the error in  $W$  Gaussian distributed ?

$$W \equiv \beta \ln t + c \quad \ln t \equiv \beta^{-1}W - \beta^{-1}c = a^*W + b^*$$

Inverse fitting is more appropriate ...  $x = a^* + b^*y$

$$a^* = \frac{n \sum x_i y_i - \sum x_i \sum y_i}{n \sum y_i^2 - (\sum y_i)^2}$$

$$b^* = \frac{n \sum x_i y_i - \sum x_i \sum y_i}{n \sum x_i^2 - (\sum x_i)^2}$$

$$\sigma_{\beta}^2 = \left( \frac{\delta \beta}{\delta a} \right)^2 \sigma_a^2 + \left( \frac{\delta \beta}{\delta b} \right)^2 \sigma_b^2$$

$$\sigma_c^2 = \left( \frac{\delta c}{\delta a} \right)^2 \sigma_a^2 + \left( \frac{\delta c}{\delta b} \right)^2 \sigma_b^2$$

# Method of correlation coefficient

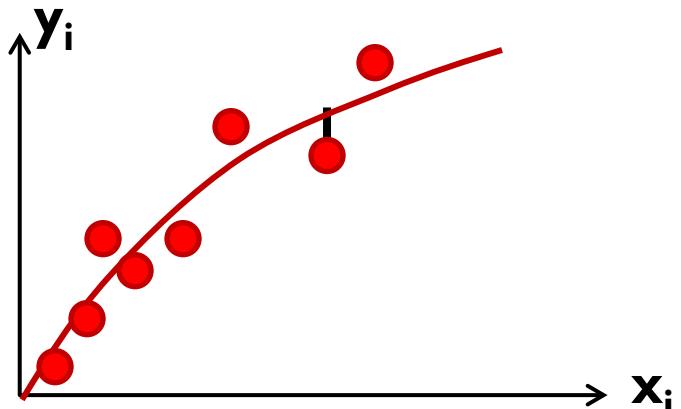
$$r = \sqrt{b \times b^*}$$

$$y = a + bx \quad x = a^* + b^* y$$

Prob. of r when x-y are uncorrelated

$$b = \frac{n \sum x_i y_i - \sum x_i \sum y_i}{n \sum y_i^2 - (\sum y_i)^2}$$

$$b^* = \frac{n \sum x_i y_i - \sum x_i \sum y_i}{n \sum x_i^2 - (\sum x_i)^2}$$



n/r	0.5	0.7	0.9
3	0.667	0.506	0.287
4			0.1
6			
7			
10	0.141	0.024	

Example. If  $r=0.9$  for  $n=4$ , there is only 10% chance (0.1 value) that this is accidental. If however  $r=0.5$  with  $n=10$ , there is 14.1% chance that it is accidental.

## (2) Fisher's Maximum Likelihood Method

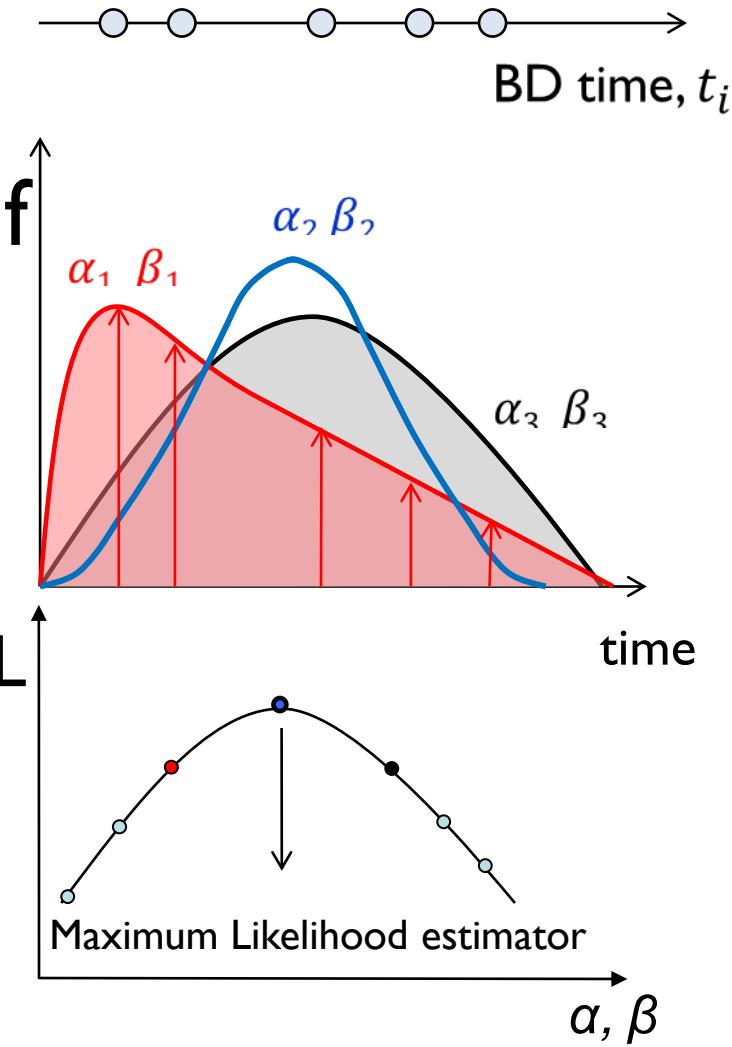
Towering figure, showed that Mendel manipulated data; MATLAB **fitdist** functions

$$f(t_i, \alpha, \beta)$$

$$L = \prod_{i=1}^n f(t_i, \alpha, \beta)$$

$$\ln L = \sum_{i=1}^n \ln f(t_i, \alpha, \beta)$$

$$\frac{d \ln L}{d \alpha} = 0 \quad \frac{d \ln L}{d \beta} = 0$$



# Example: origin of least square method

Let the error around each data point be distributed Normally ...

$$f(y_i, \mu) = \frac{1}{\sigma_i \sqrt{2\pi}} e^{-\frac{(y_i - \mu)^2}{2\sigma_i^2}}$$

Then the Likelihood function for this problem is :

$$\begin{aligned} L &= \prod_{i=1}^N f(y_i, \mu) = \prod_{i=1}^N \frac{1}{\sigma_i \sqrt{2\pi}} e^{-\frac{(y_i - \mu)^2}{2\sigma_i^2}} \\ &= \left[ \frac{1}{\sqrt{2\pi}} \right]^n \frac{1}{\prod \sigma_i} e^{-\frac{(y_1 - \mu)^2}{2\sigma_1^2} - \frac{(y_2 - \mu)^2}{2\sigma_2^2} - \dots - \frac{(y_n - \mu)^2}{2\sigma_n^2}} \\ &= \left[ \frac{1}{\sqrt{2\pi}} \right]^n \frac{1}{\prod \sigma_i} e^{-\sum \frac{(y_i - \mu)^2}{2\sigma_i^2}} \end{aligned}$$

## Example (continued)

$$\left. \frac{\partial \ln L(a,b)}{\partial a} \right| = \frac{\partial}{\partial a} \left[ n \ln(n \ln \left( \frac{1}{\sigma \sqrt{2\pi}} \right) - \sum \frac{(y_i - y(x_i))^2}{2\sigma_i^2} \right] = 0$$

$$\frac{\partial}{\partial a} \left[ \sum_{i=1}^n \frac{(y_i - y(x_i))^2}{2\sigma^2} \right] = 0$$
$$\frac{\partial}{\partial b} \left[ \sum_{i=1}^n \frac{(y_i - y(x_i))^2}{2\sigma_i^2} \right] = 0$$

$$a = \left( \sum x_i^2 \sum y_i - \sum x_i \sum x_i y_i \right) \times \textcolor{red}{D}^{-1}$$

$$b = \left( n \sum x_i y_i - \sum x_i \sum y_i \right) \times \textcolor{red}{D}^{-1}$$

$$\textcolor{red}{D} \equiv n \sum x_i^2 - \left( \sum x_i \right)^2$$

Linear fit is a special case of MLE with requirement  
that error is distributed normally ...

# Example: MLE estimator for one-parameter distribution

$$f(t; K) = K \times t \times e^{-Kt^2/2}$$

$$\begin{aligned} L &= \left( Kt_1 e^{-Kt_1^2/2} \right) \times \left( Kt_2 e^{-Kt_2^2/2} \right) \times \left( Kt_3 e^{-Kt_3^2/2} \right) \times \left( Kt_4 e^{-Kt_4^2/2} \right) \times \dots \\ &= K^n \left( \prod_{i=1}^n t_i \right) \exp \left( -\frac{K}{2} \sum_{i=1}^n t_i^2 \right) \end{aligned}$$

$$\ln L = n \ln K + \sum_{i=1}^n \ln t_i - [\text{3rd term ?}]$$

- (A)  $\frac{K}{2} \sum_{i=1}^n \ln t_i^2$     (B)  $\frac{K}{2} \sum_{i=1}^n t_i^2$     (C)  $K \sum_{i=1}^n t_i^2 / 2$

# Example: MLE estimator for one-parameter distribution

$$f(t; K) = K \times t \times e^{-Kt^2/2}$$

$$\begin{aligned} L &= \left( Kt_1 e^{-Kt_1^2/2} \right) \times \left( Kt_2 e^{-Kt_2^2/2} \right) \times \left( Kt_3 e^{-Kt_3^2/2} \right) \times \left( Kt_4 e^{-Kt_4^2/2} \right) \times \dots \\ &= K^n \left( \prod_{i=1}^n t_i \right) \exp \left( -\frac{K}{2} \sum_{i=1}^n t_i^2 \right) \end{aligned}$$

$$\ln L = n \ln K + \sum_{i=1}^n \ln t_i - \frac{K}{2} \sum_{i=1}^n t_i^2$$

$$\frac{d \ln L}{dK} = 0 \quad \Rightarrow \quad K = 2n \sqrt{\sum_{i=1}^n t_i^2}$$

# Example: MLE estimator for Weibull

Recall  $f(t; \alpha, \beta) = \frac{\beta}{\alpha^\beta} \cdot t^{\beta-1} \cdot e^{-\left(\frac{t}{\alpha}\right)^\beta}$

$$\begin{aligned}\ln L &= \sum_{i=1}^n \ln f(t_i, \alpha, \beta) \\ &= n \ln \beta - n \ln \alpha + (\beta - 1) \sum_{i=1}^n \ln t_i / \alpha - \sum_{i=1}^n (t_i / \alpha)^\beta\end{aligned}$$

$$\frac{d \ln L}{d \alpha} = 0 \quad \frac{d \ln L}{d \beta} = 0$$

$$\left( \sum_{i=1}^n t_i^\alpha \ln(t_i)^\beta \middle/ \sum_{i=1}^n t_i^\beta \right) - \frac{1}{n} \sum_{i=1}^n \ln(t_i)^\beta = 1 \quad \alpha = \left[ \frac{1}{n} \sum_{i=1}^n t_i^\beta \right]^{\frac{1}{\beta}}$$

Solve for unknowns  $\alpha, \beta$

# HW: MLE for Log-Normal

$$f(t; \mu, \sigma) = \frac{1}{t\sigma\sqrt{2\pi}} \cdot \exp\left[-\frac{\{\ln(t) - \ln(\mu)\}^2}{2\sigma^2}\right]$$

$$\ln L = \sum_{i=1}^n \ln f(t_i, \alpha, \beta)$$

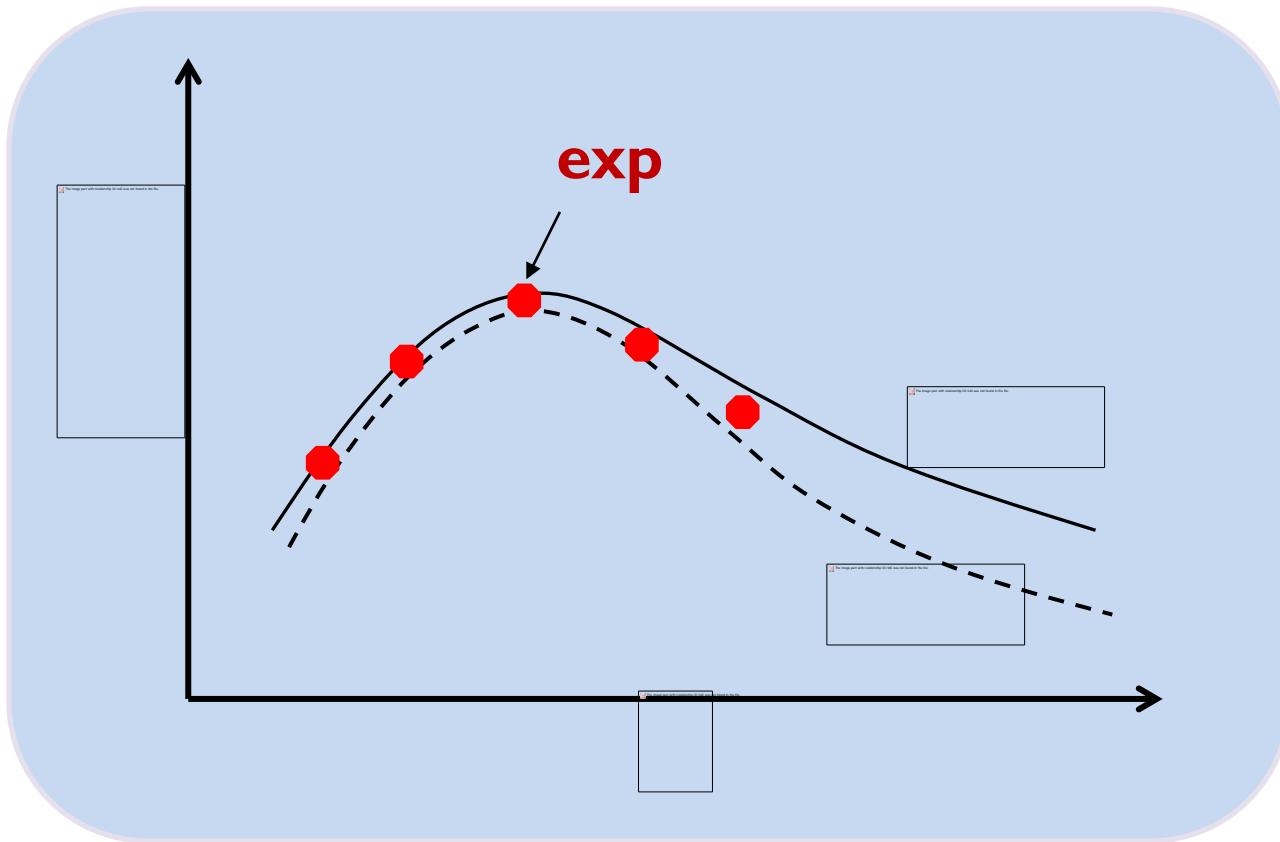
$$\frac{d \ln L}{d\alpha} = 0 \quad \frac{d \ln L}{d\beta} = 0$$

$$\frac{\sum_{i=1}^n t_i^\beta \ln(t_i)^\beta}{\sum_{i=1}^n t_i^\beta} - \frac{1}{n} \sum_{i=1}^n \ln(t_i)^\beta = 1 \quad \alpha = \left[ \frac{1}{n} \sum_{i=1}^n t_i^\beta \right]^{\frac{1}{\beta}}$$

# Outline

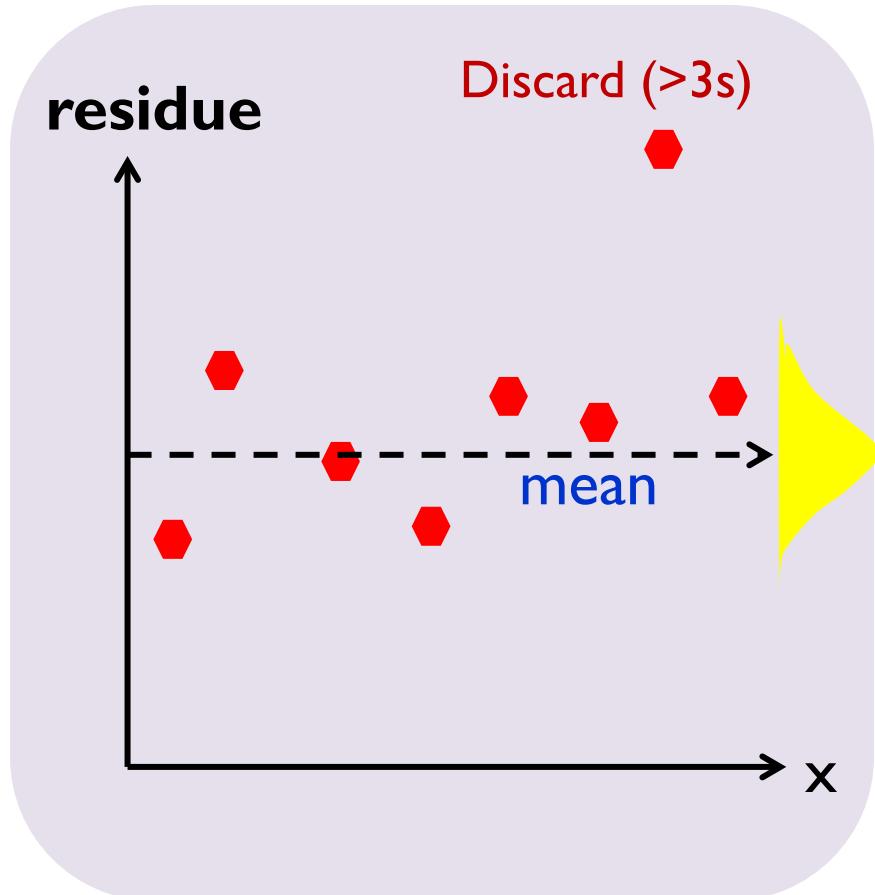
1. Introduction: The problem of matching data with theoretical distribution
2. Parameter extractions: Moments, linear regression, maximum likelihood
3. Goodness of fit: Residual, Pearson, Cox, Akika
4. Conclusion

# (1) Goodness of Fit: First check visually

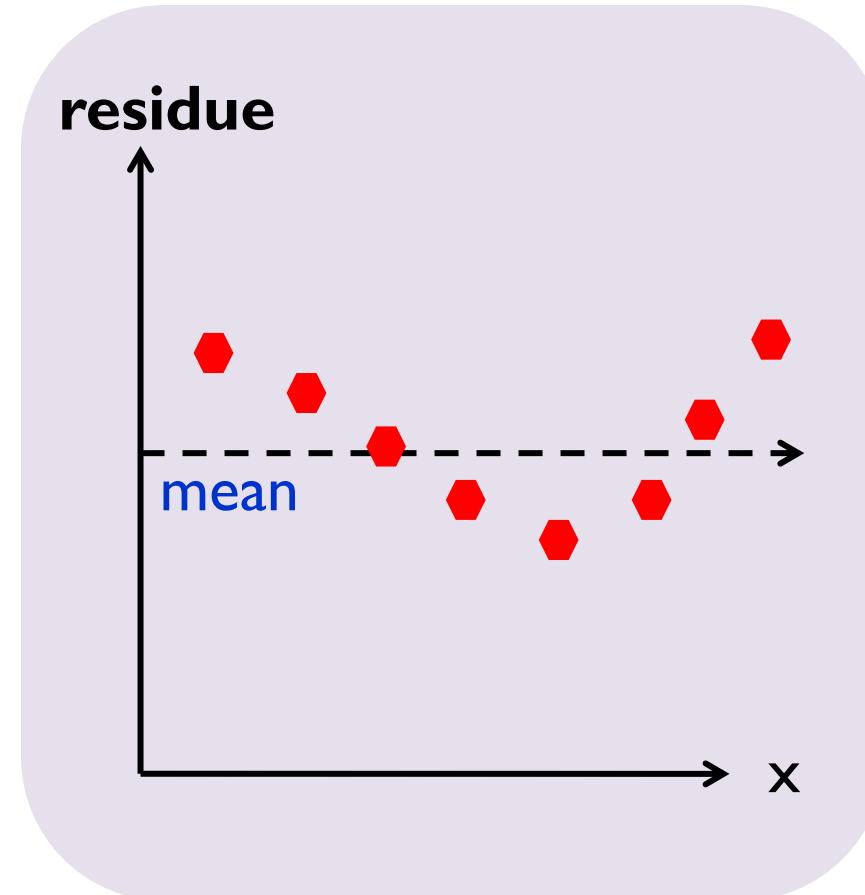


Statistical analysis is helpful only when there is  
an intuitive feel that the fit looks good ...

## (2) Goodness of Fit: Residual method



A good fit (normal distribution of residue))



A bad fit (systematic distribution in residual)

### (3) Q-Q Method: An example

Data: {3, 6, 7, 8, 8, 10, 13, 15, 16, 20}

What is the first quartile point? (A) 3 (B) 7 (C) 8 (D) 10

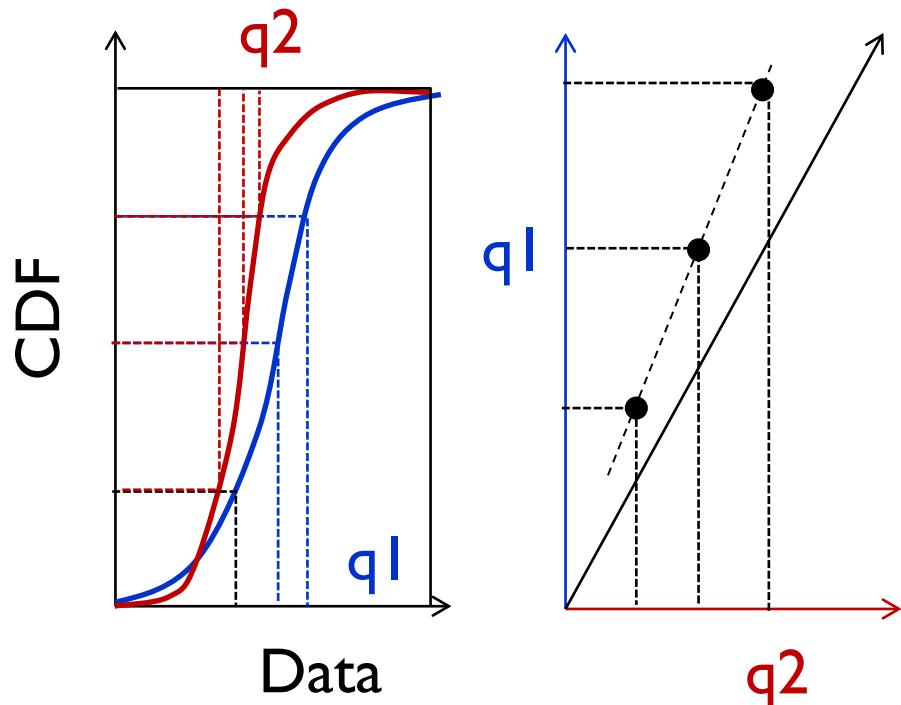
What is the median point? (A) 3 (B) 8 (C) 10 (D) 16

What is the third quartile point? (A) 13 (B) 15 (C) 16 (D) 20

Exponential distribution  $Q_2(p, K) = -\ln(1 - p)/K$

What is the 2nd quartile point? (A)  $\ln(1.33)/k$ , (B)  $\ln(2)/K$ , (C)  $\ln 4/k$

### (3) Goodness of fit: Q-Q Method



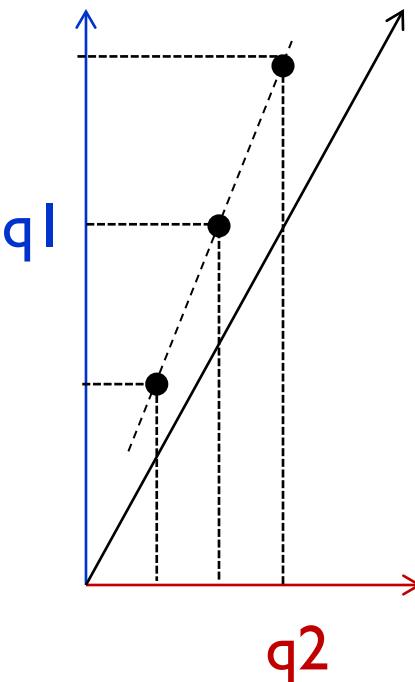
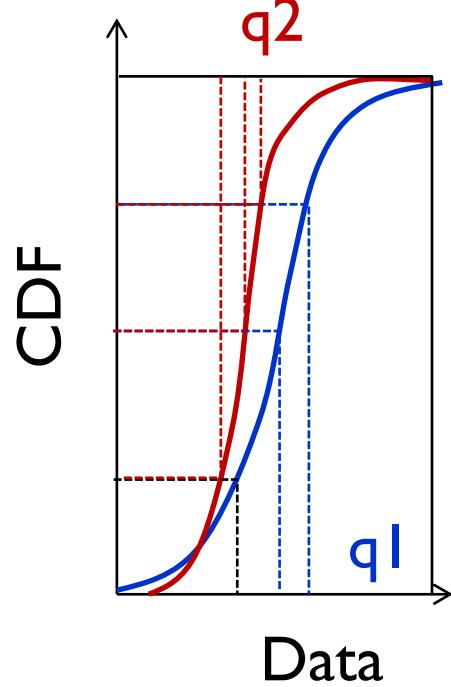
q-Quantile and quartile are different things. Median is 2-quantile, Quartile is 4-quantile, decile is a 10-quantile, percentile is 100-quantile, etc.

Take the q-quantile values of the original data and plot in the y-axis.

Take the q-quantile values of the test-distribution (i.e., calculate  $x = F^{-1}(q)$ ) to define the x-axis.

Visually inspect and establish deviation from linearity.

# Q-Q Method: An example



median  
↓

{3, 6, 7, 8, 8, 10, 13, 15, 16, 20}

$10 \times 1/4 \sim 3$      $10 \times 2/4 \sim 5$      $10 \times 3/4 \sim 8$

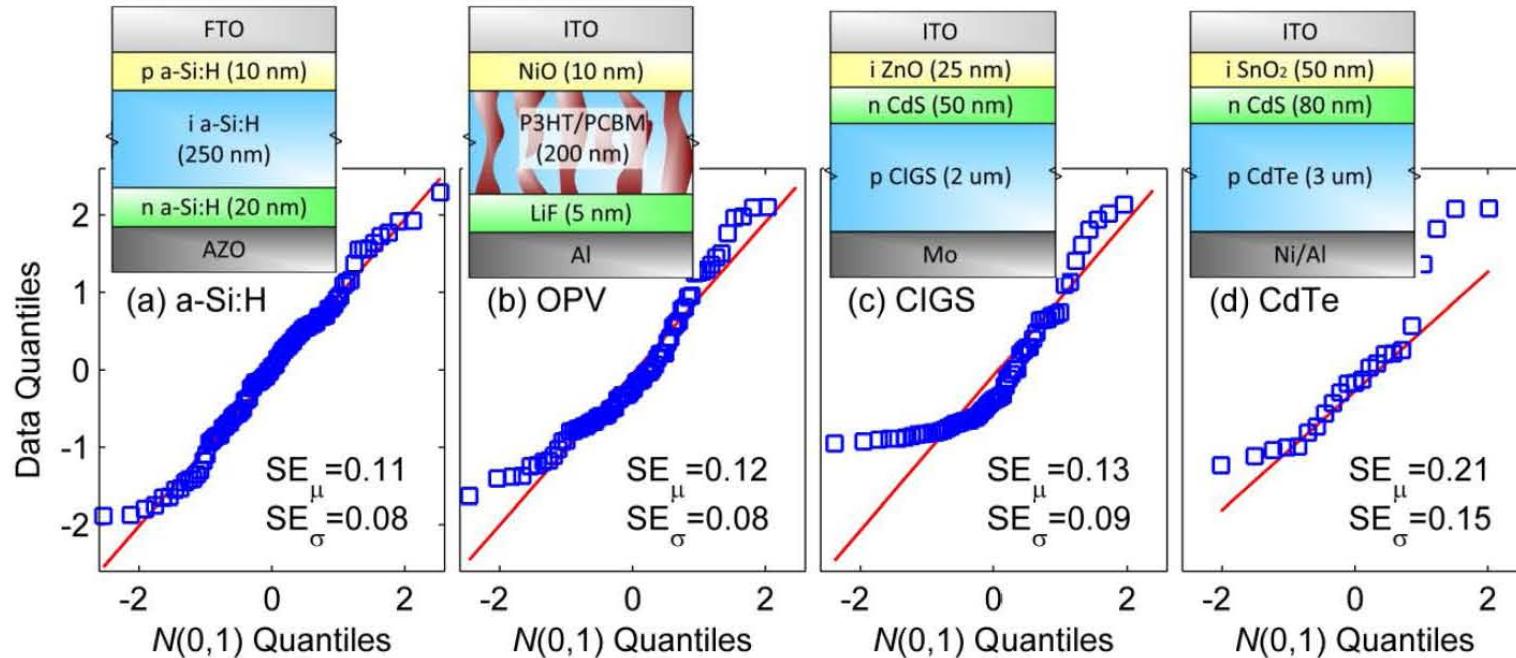
Q1 points = (7, 9, 15)

Q2: Exponential distribution

$$Q_2(p, K) = -\ln(1 - p)/K$$

$$\left. \begin{array}{l} \text{1}^{\text{st}} \text{ q: } \ln(4/2)/K \\ \text{2}^{\text{nd}} \text{ q: } \ln(4/3)/K \\ \text{3}^{\text{rd}} \text{ q: } \ln(4/1)/K \end{array} \right\} \text{ Given } K$$

# Q-Q method: an example



Data against log-normal plot: Optimize  $(\mu, \sigma)$

## (4) Goodness of Fit: Cox-Oakes measure

$$\mu_3^* = \frac{\mu_3}{\sigma^3}$$

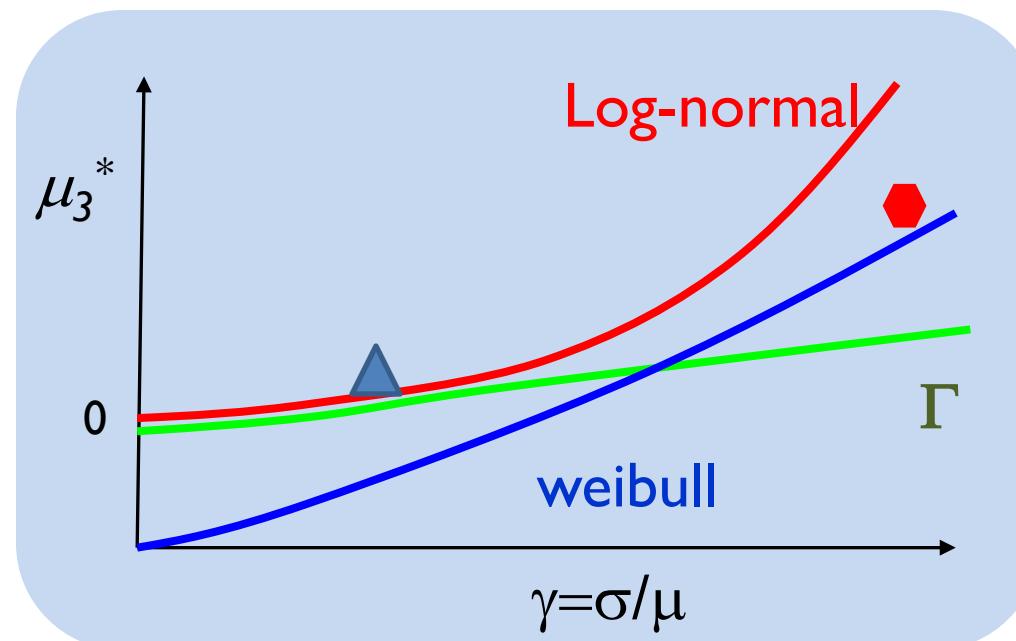
$$\mu = \frac{1}{n} \sum_{i=1}^n t_i$$

$$\sigma^2 = \frac{1}{n-1} \sum_{i=1}^n (t_i - \mu)^2$$

*Solid lines are known for various distributions.*

Example. For a given  $\alpha, \beta$ , Weibull has a specific  $\mu, \sigma, \mu_3$  (blue triangle)

Logic: Every distribution has different shape.



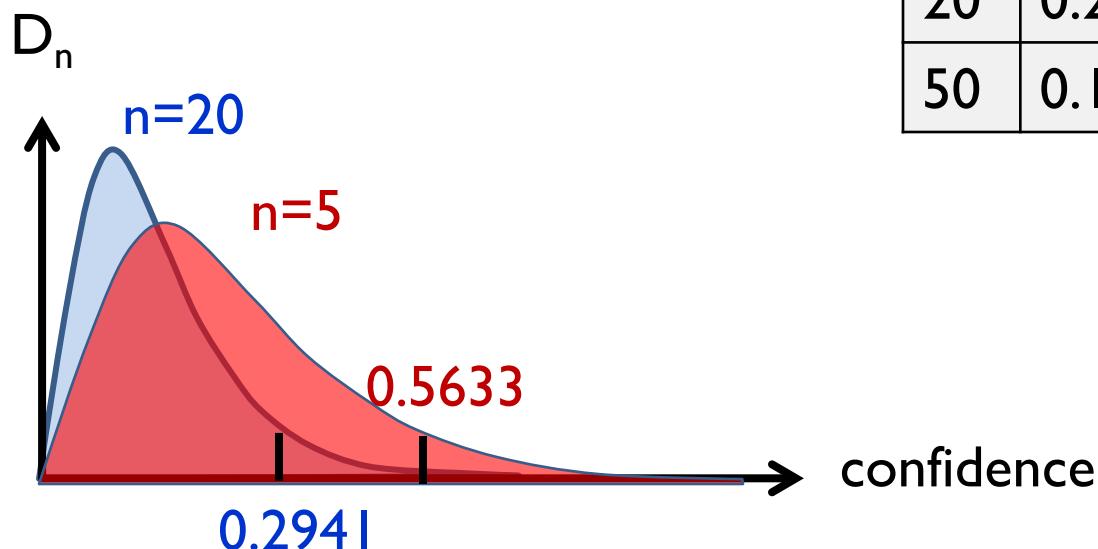
## (5) Kolmogorov-Smirnov algorithm

Compute ...  $D_n = \max |F_{obs}(t_i) - F_{theory}(t_i)|$       5% significance level

Sample size

If  $D_n > D_n^{crit}$ , fit is poor ...

n	D <sub>crit</sub> (n)
5	0.5633
10	0.4092
20	0.2941
50	0.1884



# Example: Kolmogorov-Smirnov Test

Compute ...  $D_n = \max \left| F_{obs}(t_i) - F_{theory}(t_i) \right|$

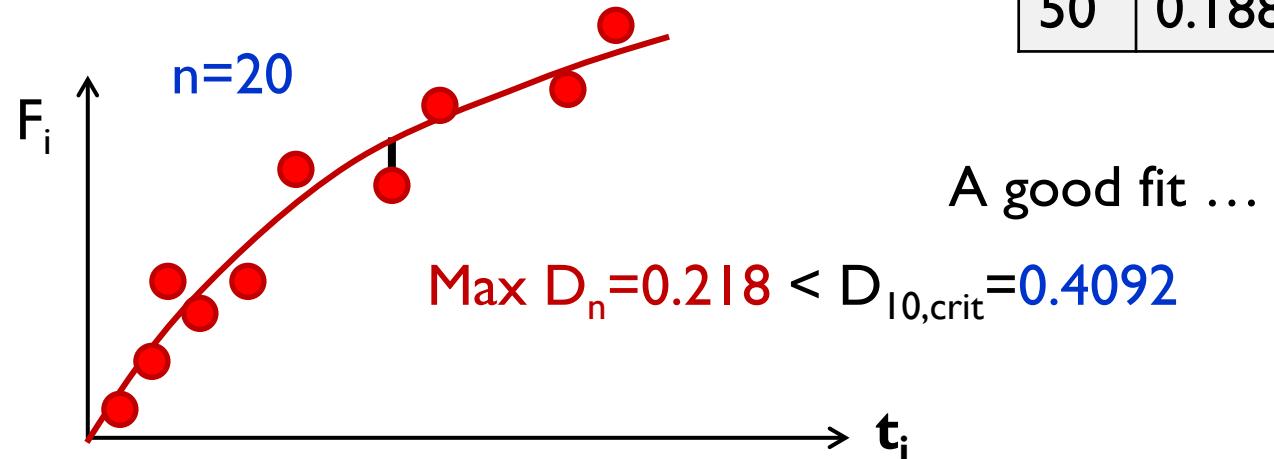
Sample size

If  $D_n > D_{n,crit}$ , fit is poor ...

	3	20	40	52	53	54	85	318	429	553
	0.067	0.164	0.260	0.356	0.452	0.548	0.644	0.740	0.837	0.933

n	D <sub>crit(n)</sub>
5	0.5633
10	0.4092
20	0.2941
50	0.1884

[h,p] = kstest(x,'CDF',test\_cdf,'Alpha',0.01)



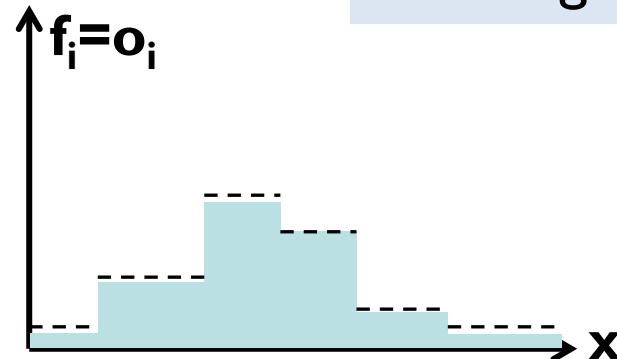
## (6) Pearson $\chi^2$ – test algorithm

Calculate ...

$$\chi_s^2 = \sum_{i=1}^{n^*} \frac{(o_i - e_i - 0.5)^2}{e_i} \quad e_i = n^* \times p_i$$

- o ... Observed
- e ... expected
- n\* ... datapoints
- P<sub>i</sub> .... probability
- v ... deg. of freedom

v	5% ( $\chi^2$ )
2	5.99
4	9.49
10	18.307
20	27.68



If the value observed  $\chi^2$  value exceeds critical value, the fit is poor.

# A famous example: Schon story

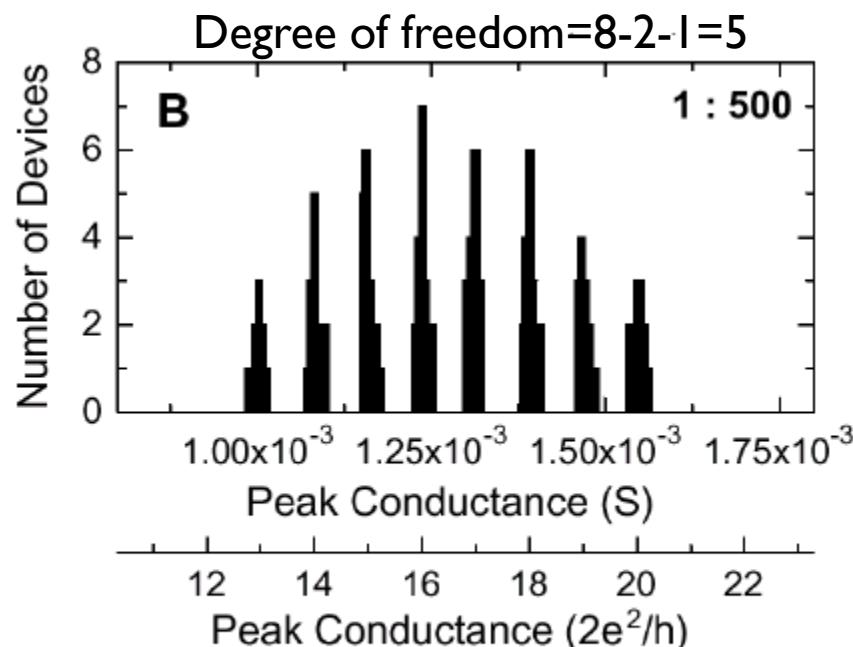


Figure 46. Figure 3(B) from "SingleMolecule" Paper (XIII), showing a histogram of conductances from diluted SAMFETs,

The data indicating conductance quantization did not arise from an objective measurement process. At a minimum, the assignment of conductance values was controlled by the expected shape of the final distribution. Such a biased process cannot provide convincing evidence for quantization. The response to this concern appears to be deliberately deceptive, suggesting that this misrepresentation was intentional.

The preponderance of evidence indicates that Hendrik Schon committed scientific misconduct, specifically data fabrication. In this case,

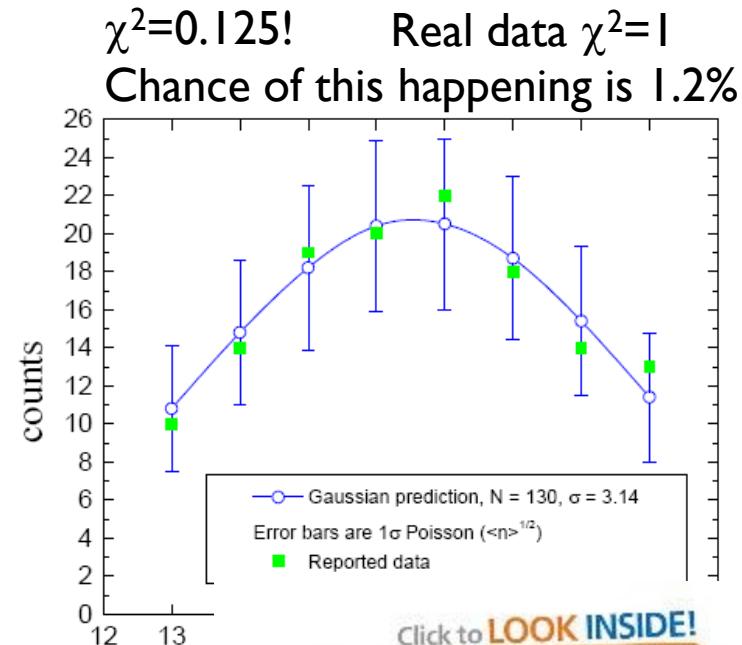
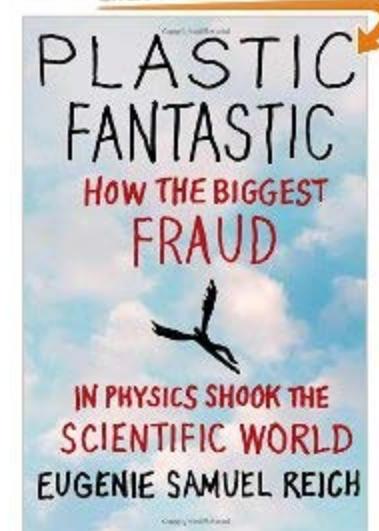


Figure 47. devices in ea



# An Introduction to Data Analysis, Design of Experiment, and Machine Learning

## *Lecture 5. Design of Experiments Scaling of Theory of Equations*

Muhammad A. Alam  
[alam@purdue.edu](mailto:alam@purdue.edu)



# Course Outline

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

Lecture 1: Introduction

Lecture 2: Collecting and plotting  $x_1, x_2, \dots x_n$

Lecture 3: Physical and empirical  $f, F, df/dx, \dots$

Lecture 4: Model selection between  $f_1, f_2, \dots$

Lecture 5: Model Selection: Cross-validation and Bootstrapping method

Lecture 6: **Scaling theory with known  $f$ ,  $f(\bar{x}) = f(\bar{X})$**

Lecture 7: Scaling theory with unknown  $f$ ,  $\bar{x} \rightarrow X$

Lecture 8: Design of experiments to determine  $\bar{y}_{\max} = f(\bar{x})$

Lecture 9: DOE and ANOVA

Lecture 11: Principle component analysis for classifying  $\{y\}$ .

Lecture 12: Machine learning ... Statistical approach learn  $f$

Lecture 13: Machine learning ... Deep network, Karnaugh map, and other approaches

Lecture 14: Interpretable ML: Physics-based machine learning  $f = f_{\text{physics}} + \Delta f$

Lecture 15: Conclusions

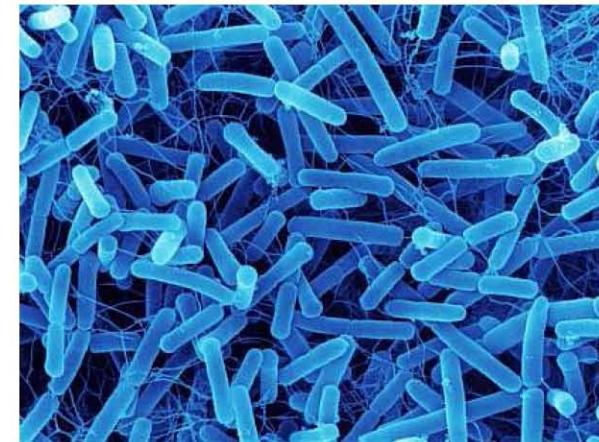
# Outline

1. Introduction
2. Rules of scaling or nondimensionalization
3. Scaling of ordinary differential equations
4. Scaling of partial differential equations
5. Equivalence of equations and solutions
6. Conclusions

# Stress-induced cell death

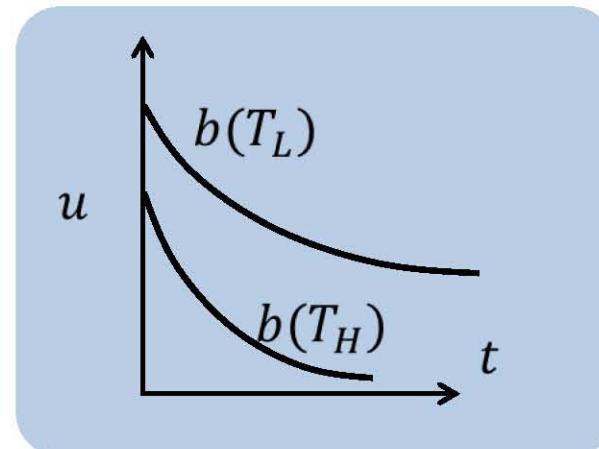
Equation:  $\frac{dn}{dt} = -b(T)n$

$$\Rightarrow n = n_0 e^{-b(T)t} \equiv f(n_0, b, t)$$



5 experiments each for  $n_0, b, t$   
... 125 measurements

If with multiple samples, hundreds  
of measurements required.



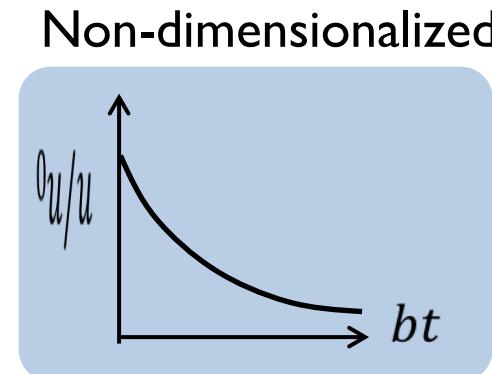
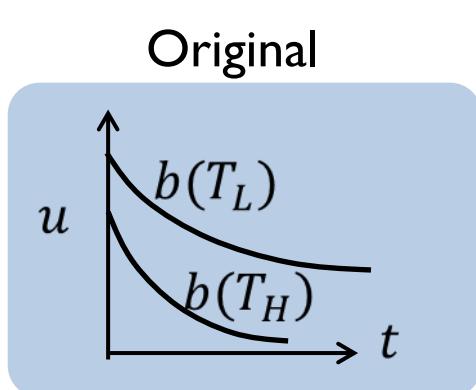
# Stress-induced cell death

$$n = n_0 e^{-b(T) t} \equiv f(n_0, b, t)$$

Three variables: 125 measurements

Normalized equation:  $\frac{n}{n_0} = g(bt) \Rightarrow N = g(\tau)$

Two variables: 25 experiments.



# Goals of Nondimensionalization

- Simplify differential equations
- Rescale variables to a unitless form
- Get rid of unnecessary parameters
- Reduce the number of experiments needed to test a hypothesis

# Rules for nondimensionalization

- Identify the **independent** and **dependent** variables;
- Replace each of them with a quantity **scaled** relative to a characteristic unit of measure to be determined;
- Divide through by the coefficient of the **highest order** polynomial or derivative term;
- Choose judiciously the definition of the characteristic unit for each variable **so that the coefficients** of as many terms as possible become 1;
- Rewrite the system of equations in terms of their **new dimensionless** quantities.

# Outline

1. Introduction
2. Rules of scaling or nondimensionalization
3. Scaling of ordinary differential equations
4. Scaling of partial differential equations
5. Equivalence of equations and solutions
6. Conclusions

# (1) Constant Coefficient 1st order Equation

I. Equation:  $a \frac{dy}{dt} + by = A\mathbf{f}(t)$

Define scaled variables:  $y = x y_c$ ,  $t = \tau t_c$

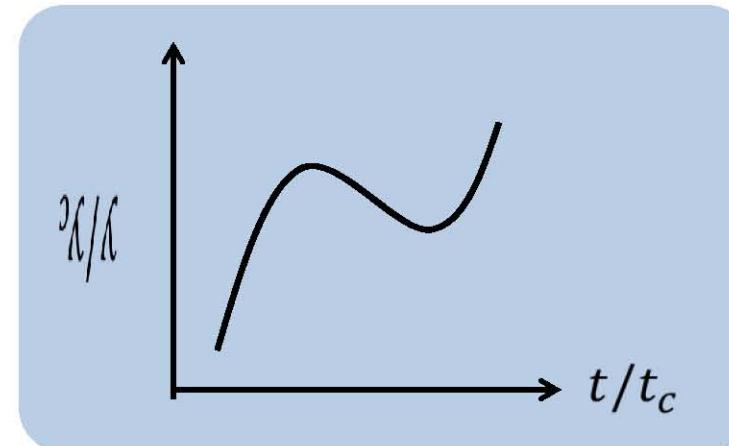
2. Normalized equation:  $a \frac{y_c dx}{t_c d\tau} + b y_c x = A\mathbf{f}(\tau t_c)$

$$\frac{dx}{d\tau} + \frac{bt_c}{a} x = \frac{At_c}{ay_c} f(\tau t_c) = BF(\tau)$$

3. Scale factors:  $\frac{bt_c}{a} \equiv 1 \Rightarrow t_c \equiv \frac{a}{b}$ ,  $B \equiv \frac{At_c}{ay_c} \equiv \frac{A}{by_c}$

4. Final Equation:  $\frac{dx}{d\tau} + x = BF(\tau)$

Note: What about  $y_c$ ; Undefined, but scales B.



# (1) ... Must scale the boundary conditions

Original Equation:  $a \frac{d\textcolor{blue}{y}}{d\textcolor{teal}{t}} + b\textcolor{blue}{y} = A\textcolor{brown}{f}(t)$

Final Equation:  $\frac{d\textcolor{blue}{x}}{d\tau} + \textcolor{blue}{x} = BF(\tau)$

Original boundary condition:  $y(t = 3) = y_0$

Scaled boundary condition:

$$\textcolor{blue}{y} = \textcolor{blue}{x} y_c, \textcolor{teal}{t} = \textcolor{teal}{\tau} t_c$$

$$\Rightarrow xy_c(\tau t_c = 3) = y_0 \Rightarrow x \left( \tau = \frac{3}{t_c} \right) = \frac{y_0}{y_c}$$

## (2) Higher order equations

1. Equation:  $a \frac{d^2y}{dt^2} + b \frac{dy}{dt} + cy = Af(t)$ . With  $y = x y_c$  and  $t = t_c \tau$ .

$$\frac{ay_c}{t_c^2} \frac{d^2x}{d\tau^2} + \frac{by_c}{t_c} \frac{dx}{d\tau} + cy_c x = Af(t_c \tau).$$

2. Normalized equation:  $\frac{d^2x}{d\tau^2} + \frac{bt_c}{a} \frac{dx}{d\tau} + \frac{ct_c^2}{a} x = \frac{At_c^2}{ay_c} f(t_c \tau) \equiv BF(\tau)$ .

3. Two parameters ( $t_c$  and  $y_c$ ), therefore we can two coefficients set to 1.

$t_c$  scaling: Either  $\frac{bt_c}{a} = 1 \Rightarrow t_c = \frac{a}{b}$  or

$\frac{ct_c^2}{a} = 1 \Rightarrow t_c = \sqrt{\frac{a}{c}}$  so that the 1st coefficient is  $b \frac{t_c}{a} = b \sqrt{\frac{a}{c}} = 2\xi$

## (2) ...Higher order equations

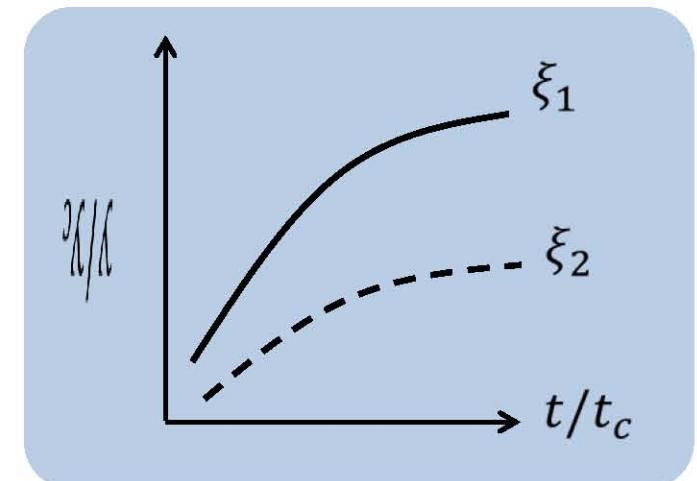
I. Equation:  $a \frac{d^2y}{dt^2} + b \frac{dy}{dt} + cy = Af(t)$ .

Two parameters ( $t_c$  and  $y_c$ ), therefore we can set two coefficients to unity.

$t_c = \sqrt{\frac{a}{c}}$  so that the first coefficient becomes,  $b \frac{t_c}{a} = b \sqrt{\frac{a}{c}} = 2\xi$

$y_c$  scaling:  $B = \frac{At_c^2}{ay_c} = 1$ , therefore  $y_c = A \frac{t_c^2}{a} = \frac{A}{c}$

4. Final equation:  $\frac{d^2x}{d\tau^2} + 2\xi \frac{dx}{d\tau} + x = \frac{A}{c} f(t_c \tau) \equiv BF(\tau)$ .



### (3) HW: Coupled Equations

$$\frac{dx}{dt} = \gamma x \left(1 - \frac{\alpha x + \beta y}{N}\right) \quad \frac{dy}{dt} = \theta y \left(1 - \frac{\alpha x + \beta y}{N}\right)$$

$x = x_c X, \quad y = y_c Y, \quad t = t_c \tau$

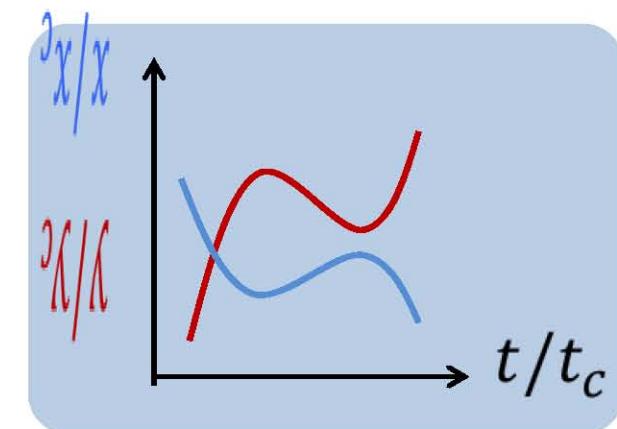
$$\frac{dX}{d\tau} = \frac{t_c}{x_c} x_c \gamma X \left(1 - \frac{x_c \alpha X + y_c \beta Y}{N}\right) \quad \frac{dY}{d\tau} = \frac{t_c}{y_c} y_c \theta Y \left(1 - \frac{x_c \alpha X + y_c \beta Y}{N}\right)$$

$$\frac{dX}{d\tau} = t_c \gamma X - \frac{t_c \gamma x_c \alpha X + t_c \gamma y_c \beta Y}{N} \quad \frac{dY}{d\tau} = t_c \theta Y - \frac{t_c \gamma x_c \alpha X + t_c \gamma y_c \beta Y}{N}$$

Three variables allows three coefficients to set to 1.

$$t_c \gamma = 1 \Rightarrow t_c = \gamma^{-1} \quad t_c \gamma x_c \alpha / N = 1 \Rightarrow x_c = N \alpha^{-1}, \quad t_c \gamma y_c \beta / N \Rightarrow y_c = N \beta^{-1}$$

$$\frac{dX}{d\tau} = -Y \quad \text{and} \quad \frac{dY}{d\tau} = \theta \gamma^{-1} Y - X - Y = (\kappa - 1)Y - X$$



# Outline

1. Introduction
2. Rules of scaling or nondimensionalization
3. Scaling of ordinary differential equations
4. Scaling of partial differential equations
5. Equivalence of equations and solutions
6. Conclusions

# Nondimensionalization: example

- Minority carrier diffusion equation:  $\frac{dn}{dt} = D \frac{d^2n}{dx^2} - \frac{n}{\tau} + G_L$ 
  - 3 parameters:  $D$ ,  $\tau$ , and  $G_L$
  - 3 variables:  $n$ ,  $x$ , and  $t$
- Goal: convert  $n$ ,  $x$ , and  $t$  to dimensionless  $\tilde{n}$ ,  $\tilde{x}$ ,  $\tilde{t}$  in terms of  $D$ ,  $\tau$ , and  $G_L$
- Technique:
  - Set  $\tilde{n} = \frac{n-n_r}{n_0}$ ,  $\tilde{x} = \frac{x-x_r}{x_0}$ ,  $\tilde{t} = \frac{t-t_r}{t_0}$   
( $n_r, x_r, t_r$ : reference values;  $n_0, x_0, t_0$ : scaling factors)
  - Assume  $n$ ,  $x$ ,  $t$  starts at  $n=0$ ,  $x=0$  and  $t=0$  so  $n_r = 0$ ,  $x_r = 0$ ,  $t_r = 0$

# Nondimensionalization: example

- Rewrite:  $\frac{dn}{dt} = D \frac{d}{dx} \left( \frac{dn}{dx} \right) - \frac{n}{\tau} + G_L$ , insert  $n = \tilde{n}n_0, x = \tilde{x}x_0, t = \tilde{t}t_0$

$$\Rightarrow \frac{d\tilde{n}n_0}{d\tilde{t}t_0} = D \frac{d}{d\tilde{x}x_0} \left( \frac{d\tilde{n}n_0}{d\tilde{x}x_0} \right) - \frac{\tilde{n}n_0}{\tau} + G_L$$

Divide by this factor  
on both sides

$$\Rightarrow \frac{n_0}{t_0} \frac{d\tilde{n}}{d\tilde{t}} = D \frac{n_0}{x_0^2} \frac{d}{d\tilde{x}} \left( \frac{d\tilde{n}}{d\tilde{x}} \right) - \frac{n_0}{\tau} \tilde{n} + G_L$$

$$\Rightarrow \frac{d\tilde{n}}{d\tilde{t}} = \boxed{D \frac{t_0}{x_0^2} \frac{d}{d\tilde{x}} \left( \frac{d\tilde{n}}{d\tilde{x}} \right)} - \boxed{\frac{t_0}{\tau} \tilde{n}} + \boxed{\frac{t_0}{n_0} G_L}$$

New coefficients

- Choose  $t_0 = \tau, x_0 = \sqrt{D\tau}, n_0 = \tau G_L$  so that PDE become as simple as possible:

$$d\tilde{n} - d^2\tilde{n} = 0$$

# Conclusions

1. Scaling of equations is a powerful concept.
2. Scaling of the equations involves very specific rules; the equations and the boundary conditions must be scaled simultaneously.
3. The power of scaling involves reducing the number of experiments or simulations needed to investigate a hypothesis.
4. The scaling makes numerical solution simpler by making the variables of similar magnitude.
5. The scaling also allows one to look up solutions from in differential equations handbook or websites.
6. Scaling allows one to compare equations from very different fields and solve the problem in one field by borrowing solution from a different field.

# References

Book on dimensional analysis including python code:

<https://hplgit.github.io/scaling-book/doc/pub/book/pdf/scaling-book-4screen-sol.pdf>

Nondimensionalized models produce physically universal and numerically robust results. The topic is easily learned from the following articles.

ODE:

<https://en.wikipedia.org/wiki/Nondimensionalization>

PDE:

<https://link.springer.com/article/10.1007/s11071-015-2233-8>

Examples:

[https://user.engineering.uiowa.edu/~fluids/Posting/Schedule/Example/Dimensional%20Analysis\\_11-03-2014.pdf](https://user.engineering.uiowa.edu/~fluids/Posting/Schedule/Example/Dimensional%20Analysis_11-03-2014.pdf)

Coupled Equation:

<https://math.stackexchange.com/questions/845891/nondimensionalization-of-coupled-ode>

References: R. W. Robinett,  
"Dimensional Analysis at the Other Language of Physics," American Journal of Physics, 83(4), 353, 2015.

# Review Questions

1. A non-dimensionalized equation is also called a scaled equation. Explain.
2. If there are two variables (one independent, the other dependent), how many scaled coefficients can be set to 1?
3. When scaling the differential equation, do you also need to scale the boundary conditions as well?
4. Why is it important to plot the experimental and simulation results in terms of scaled variables?
5. Why is it helpful to non-dimensionalize an equation before looking up the solution in a handbook?

# An Introduction to Data Analysis, Design of Experiment, and Machine Learning

## *Lecture 6. Equation-free Scaling Theory for Design of Experiments*

Muhammad A. Alam  
[alam@purdue.edu](mailto:alam@purdue.edu)



copyright 2018

This material is copyrighted by M.Alam under the  
following Creative Commons license:



**Attribution-NonCommercial-ShareAlike 2.5 Generic (CC BY-NC-SA 2.5)**

Conditions for using these materials is described at

<http://creativecommons.org/licenses/by-nc-sa/2.5/>

# Outline

1. Introduction
2. Buckingham PI Theorem
3. An Illustrative Example
4. Why does the method work
5. Conclusions

# Problem definition

Many temperature dependent degradation rate depend on temperature, barrier height, etc.

$$R = f(T, E_B; k_B, \hbar)$$

If I need to perform 10 experiment for  $T$  and 10 for  $E_B$ , etc. the number of experiments will be 100 – too expensive.

Can the same information be obtained with fewer experiments?

$$\frac{R}{(k_B T / \hbar)} = f_1\left(\frac{E_B}{k_B T}\right)$$

Variables do not matter individually,  
They only matter in combination.  
Fewer experiments are sufficient.

# Buckingham PI Theorem

Assume that a function  $g$  depends on parameters  $q_1, q_2, \dots, q_n$ , such that

$$g(q_1, q_2, \dots, q_n) = 0$$

Here  $g$  could a differential equation

$$q_1 \frac{d^2y}{dx^2} + q_2 \frac{dy}{dx} + q_3 y + q_4 = 0$$

Or, it could be a unknown blackbox, with control parameters  $q_1, q_2, q_3$ , etc.

# Buckingham PI Theorem

If the function  $g$  depends on parameters  $q_1, q_2, \dots, q_n$ , then

$$g(q_1, q_2, \dots, q_{\textcolor{red}{n}}) = 0$$

The same expression can be expressed in terms of  $(n-m)$  independent dimensionless ratios, or  $\Pi$  parameters.

$$G(\Pi_1, \Pi_2, \dots, \Pi_{\textcolor{red}{n}-\textcolor{blue}{m}}) = 0$$

$m$ = minimum number of independent dimensions typically given by  $r$ , where  $r$  is the rank of the matrix

# To determine PI ...

Determine

$$A = \begin{bmatrix} \textcolor{red}{P} & R \\ \textcolor{blue}{Q} & S \end{bmatrix} \quad \textcolor{red}{P} \text{ is a } r \times r \text{ nonsingular matrix}$$

Find the exponent matrix

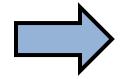
$$E = (-\textcolor{blue}{Q}\textcolor{red}{P}^{-1}, I)$$

Finally,

$$\Pi_i = q_1^{e_{i1}} q_2^{e_{i2}} \dots \dots q_N^{e_{iN}}$$

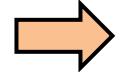
# Recall the dimensions of variables

Variable  $\rightarrow M^a \times L^b \times t^c \times \Theta^d$

  $E_B \rightarrow M^1 \times L^2 \times t^{-2} \times \Theta^0 \quad (0.5mv^2)$

$T \rightarrow M^0 \times L^0 \times t^0 \times \Theta^1 \quad (\text{kelvin})$

$R \rightarrow M^0 \times L^0 \times t^{-1} \times \Theta^0 \quad (\text{sec}^{-1})$

  $k_B \rightarrow M^1 \times L^2 \times t^{-2} \times \Theta^{-1} \quad (\text{energy/kelvin})$

$\hbar \rightarrow M^1 \times L^2 \times t^{-1} \times \Theta^0 \quad (\text{energy-sec})$

# Outline

1. Introduction
2. Buckingham PI Theorem
3. An Illustrative Example
4. Why does the method work
5. Conclusions

# Illustrative Example

$$R = f(T, \mathbf{E}_B; k_B, \hbar) \Rightarrow 0 = g(R, T, \mathbf{E}_B; k_B, \hbar)$$

$$A = \begin{bmatrix} M & L & t & \Theta \\ 0 & 0 & 0 & 1 \\ 1 & 2 & -2 & -1 \\ 1 & 2 & -1 & 0 \\ 0 & 0 & -1 & 0 \\ 1 & 2 & -2 & 0 \end{bmatrix} \begin{matrix} T \\ k_B \\ h \\ R \\ \mathbf{E}_B \end{matrix}$$

- Number of unknowns  
 $n = 5$
- Rank of the matrix  
 $r = 3$  (independent rows)
- Number of parameters ( $\pi_1, \pi_2$ )  
 $n - r = 2$
- Number of repeating variable  
 $r = m = 3$

## Example: Any nonzero determinant for P

$$\begin{array}{cccc}
 M & L & t & \Theta \\
 \left[ \begin{array}{cccc} 0 & 0 & 0 & 1 \\ 1 & 2 & -2 & -1 \\ 1 & 2 & -1 & 0 \\ 0 & 0 & -1 & 0 \\ 1 & 2 & -2 & 0 \end{array} \right] & \begin{array}{c} T \\ k_B \\ h \\ R \\ E_B \end{array} & \begin{array}{ccc} L & T & \Theta \\ \mathcal{P}_1 = \left[ \begin{array}{ccc} 0 & 0 & -1 \\ 2 & -2 & -1 \\ 2 & -1 & 0 \end{array} \right] T \\ k_B \\ h \end{array} & \text{Rank 3 ...}
 \end{array}$$

$$E = \left[ -Q P^{-1}, I \right] = \left[ \begin{array}{ccccc} T & k_B & h & R & E_B \\ -1 & -1 & 1 & 1 & 0 \\ -1 & -1 & 0 & 0 & 1 \end{array} \right] \Rightarrow \Pi_1 = \frac{\hbar R}{kT}, \Pi_2 = \frac{E_B}{kT}$$

# Physical Meaning of the exponent matrix

$$E = [-QP^{-1}, I] = \left[ \begin{array}{ccc|cc} T & k_B & h & R & E_B \\ -1 & -1 & 1 & 1 & 0 \\ -1 & -1 & 0 & 0 & 1 \end{array} \right] \Rightarrow \Pi_1 = \frac{\hbar R}{kT}, \Pi_2 = \frac{E_B}{kT}$$

$$T^a \times k_B^b \times h^c \times R = L^0 \times t^0 \times \Theta^0$$

$\Rightarrow$  3 equations for a,b,c

$$T^a \times k_B^b \times h^c \times E_B = L^0 \times t^0 \times \Theta^0$$

$\Rightarrow$  3 equations for a,b,c

Variable  $\rightarrow M^a \times L^b \times t^c \times \Theta^d$

$E_B \rightarrow M^1 \times L^2 \times t^{-2} \times \Theta^0$   $(0.5mv^2)$

$T \rightarrow M^0 \times L^0 \times t^0 \times \Theta^1$  (kelvin)

$R \rightarrow M^0 \times L^0 \times t^{-1} \times \Theta^0$  ( $\text{sec}^{-1}$ )

$k_B \rightarrow M^1 \times L^2 \times t^{-2} \times \Theta^{-1}$  (energy/kelvin)

$\hbar \rightarrow M^1 \times L^2 \times t^{-1} \times \Theta^0$  (energy-sec)

The dimensionless parameters are the Pi parameters

## Example continued ...

$$E = [-QP^{-1}, I] = \begin{bmatrix} T & k_B & h & R & E_B \\ -1 & -1 & 1 & 1 & 0 \\ -1 & -1 & 0 & 0 & 1 \end{bmatrix} \Rightarrow \Pi_1 = \frac{\hbar R}{kT}, \Pi_2 = \frac{E_B}{kT}$$

$$G\left(\Pi_1 \equiv \frac{\hbar R}{kT}, \Pi_2 \equiv \frac{E_B}{kT}\right) = 0$$

$$\Rightarrow \frac{\hbar R}{kT} = f\left(\frac{E_B}{kT}\right)$$

If you assumed  $\hbar$  to be absent (you did not know about it before Quantum mechanics), then

$$\frac{cR}{kT} = f\left(\frac{E_B}{kT}\right)$$

## Example: Any nonzero determinant for P

$$M \quad L \quad t \quad \Theta$$

$$A = \begin{bmatrix} 0 & 0 & 0 & 1 \\ 1 & 2 & -2 & -1 \\ 1 & 2 & -1 & 0 \\ 0 & 0 & -1 & 0 \\ 1 & 2 & -2 & 0 \end{bmatrix}$$

T  
 $k_B$   
h  
R  
 $E_B$

$$L \quad T \quad \Theta$$

$$P_2 = \begin{bmatrix} 2 & -2 & -1 \\ 2 & -1 & 0 \\ 0 & -1 & 0 \end{bmatrix} k_B \quad h \quad R$$

Rank 3 ...

$$Q_2 = \begin{bmatrix} 2 & -2 & 0 \\ 0 & 0 & 1 \end{bmatrix} E_B \quad T$$

$$E = [-QP^{-1}, I] = \begin{bmatrix} k_B & h & R & E_B & T \\ +0 & -1 & -1 & 1 & 0 \\ +1 & -1 & -1 & 0 & 1 \end{bmatrix} \Rightarrow \Pi_1 = \frac{E_B}{hR}, \Pi_2 = \frac{k_B T}{hR}$$

## Example: Any nonzero determinant for P

$$E = \begin{bmatrix} -QP^{-1}, I \end{bmatrix} = \begin{bmatrix} k_B & h & R & E_B & T \\ +0 & -1 & -1 & 1 & 0 \\ +1 & -1 & -1 & 0 & 1 \end{bmatrix} \Rightarrow \Pi_1 = \frac{E_B}{hR}, \Pi_2 = \frac{k_B T}{hR}$$

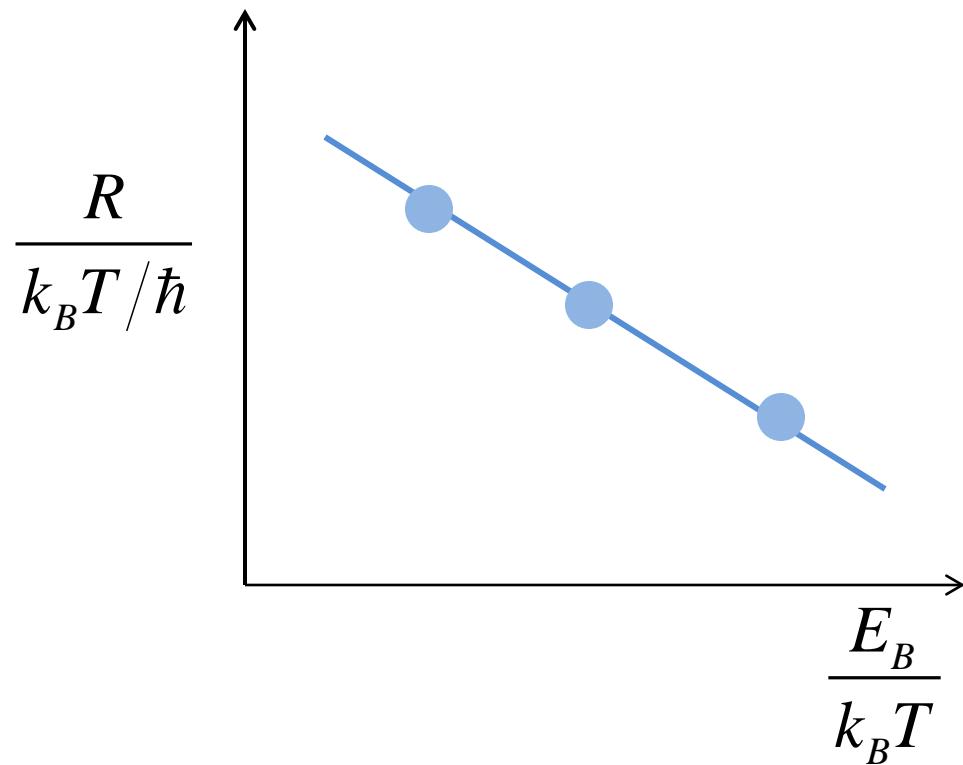
$$\Pi_3 \equiv \frac{\Pi_1}{\Pi_2} = \frac{E_B}{k_B T}$$

$$G\left(\Pi_3 \equiv \frac{E_B}{k_B T}, \Pi_2 \equiv \frac{k_B T}{hR}\right) = 0$$

$$\Rightarrow \frac{k_B T}{hR} = f_2\left(\frac{E_B}{k_B T}\right)$$

$$G\left(\Pi_1 \equiv \frac{E_B}{hR}, \Pi_2 \equiv \frac{k_B T}{hR}\right) = 0$$
$$\Rightarrow \frac{E_B}{k_B R} = f\left(\frac{k_B T}{\hbar R}\right)$$

# Plotting with dimensionless variables



# Example 2: Newton anticipates Einstein's results

**Use Bucking Pi theorem to analyze the remarkable fact that Newton foresaw bending of light by gravity long before Einstein!**

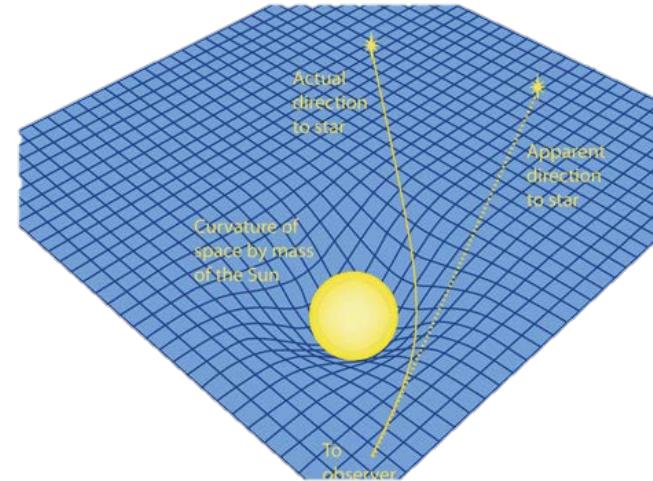
Assuming the sun can be treated as a point of mass  $m$  and the ray of light passes the mass with a distance of closest approach  $r$ , dimensional reasoning helps to predict the deflection angle  $\theta$ . The problem involves 3 of the 7 fundamental units: mass  $M$ , length  $L$ , and time  $T$ .

- a) Write the Buckingham matrix first by thinking about three variables,  $\theta, r, \text{ and } m$ . We have the following dimensions: angle  $[\theta] = L^0 T^0 M^0$ ,  $[r] = L^1 T^0 M^0$ , and  $[m] = L^0 T^0 M^1$ . By putting in Bucking Pi matrix, we find that there is no solution.
- b) Augment the matrix with two other variables, the gravitational constant,  $[G] = L^3 T^{-2} M^{-1}$  and velocity of light  $[c] = L^1 T^{-1} M^0$ . After all, we now light and gravity in the problem. Use the Buckingham pi theorem to show that the scaling factor involved.
- c) Show that the final result can be written in the form  $\theta = \alpha \frac{G m}{c^2 r}$  where  $\alpha$  is the dimensionless factor. (Using  $r = 6.96 \times 10^8 \text{ m}$ ,  $m = 1.99 \times 10^{30} \text{ kg}$ , and  $\alpha = 2$ , Newton anticipated the general relativity result within a factor of 2!)

Dynamic Similarity , the dimensionless science, A Sept. 2011 (p. 47) physics Today Article by D. Bolster. Also, see R. Kurth, "Dimensional Analysis and Group Theory in Astrophysics" Pergamon Press, Oxford, UK, 1972.

## Example 2: Newton anticipates Einstein?!

$$\theta = f(r, m) \Rightarrow 0 = g(\theta, r, m)$$



$$A = \begin{bmatrix} M & L & T \\ 0 & 0 & 0 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{bmatrix} \begin{matrix} \theta \\ r \\ m \end{matrix}$$

- The determinant is zero
- Top row is zero
- Two equations and three variables
- No solution
- Incomplete problem specification,
- Must have hidden variable(s)

## .... continued

$$\theta = f(G, c, r, m) \Rightarrow 0 = g(\theta, G, c, r, m)$$

	$M$	$L$	$T$	
$Q_1$	0	0	0	$\theta$
	0	1	0	$r$
$P_1$	1	0	0	$m$
	-1	3	-2	$G$
	0	1	-1	$c$

- Number of unknowns  
 $n = 5$
- Rank of the matrix  
 $r = 3$  (independent rows)
- Number of parameters ( $\pi_1, \pi_2$ )  
 $n - r = 2$
- Number of repeating variable  
 $r = m = 3$

$Q$  should contain the output variables,  $P$  the input variables

Muhammad A. Alain, Purdue University

## .... continued

$$\theta = f(G, c, r, m) \Rightarrow 0 = g(\theta, G, c, r, m)$$

	$M$	$L$	$T$		$E = (-Q_1 P_1^{-1}, I)$	
$Q_1$	0	0	0	$\theta$		
	0	1	0	$r$		
	1	0	0	$m$	$m \quad G \quad c \quad \theta \quad r$	
$P_1$	-1	3	-2	$G$	0 0 0 1 0	$\Pi_1 = \theta^1 m^0 G^0 c^0 r^0$
	0	1	-1	$c$	-1 -1 2 0 1	$\Pi_2 = m^{-1} G^{-1} c^2 \theta^0 r$ $= c^2 r / Gm$

$$\theta = f\left(\frac{Gm}{c^2r}\right) \sim \left(\frac{Gm}{c^2}r\right) \qquad \qquad \Pi_1 = f(\Pi_2)$$

# Outline

1. Introduction
2. Buckingham PI Theorem
3. An Illustrative Example
4. Why does the method work
5. Conclusions

## Example 3: The trick of the magic (Diffusion equation: Standard scaling)

$$D \frac{d^2 n}{dx^2} - \frac{n}{\tau} = 0$$

$$n = n_0 n^* \quad x = x_0 x^*$$

$$D \frac{n_0 d^2 n^*}{x_0^2 dx^{*2}} - \frac{n_0 n^*}{\tau} = 0$$

$$\frac{d^2 n^*}{dx^{*2}} - \frac{n_0 x_0^2 n^*}{D\tau} = 0$$

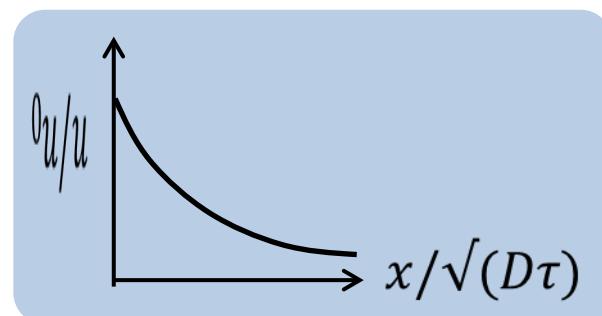
$$\frac{x_0^2}{D\tau} = 1, \quad \rightarrow \quad x_0 = \sqrt{D\tau}$$

$$\frac{d^2 n^*}{dx^{*2}} - n^* = 0$$

$$n^* = A e^{x^*} + B e^{-x^*}$$

$$\frac{n}{n_0} = A e^{\frac{x}{\sqrt{D\tau}}} + B e^{-\frac{x}{\sqrt{D\tau}}} \equiv f\left(\frac{x}{\sqrt{(D\tau)}}\right)$$

Non-dimensionalized



## Example 3: Buckingham Pi approach

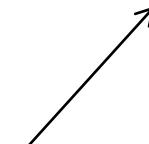
$$D \frac{d^2n}{dx^2} - \frac{n}{\tau} = 0$$

$Q_1$	$L$	$T$	
	-3	0	$n$
	1	0	$x$

$P_1$	$L$	$T$	
	2	-1	$D$
	0	1	$\tau$

$M$	$L$	$T$	
0	-3	0	$n$
0	1	0	$x$
0	2	-1	$D$
0	0	1	$\tau$



$$E = (-Q_1 P_1^{-1}, I)$$

$D$	$\tau$	$n$	$x$
-1.5	-1.5	1	0
-0.5	-0.5	0	1



$$\Pi_1 = n^1 D^{-1.5} \tau^{-1.5} \equiv \frac{n}{n_0}$$

$$\Pi_2 = x^1 D^{-0.5} \tau^{-0.5} \equiv \frac{x}{\sqrt{D\tau}}$$

Variables ...  $n=4$

Rank ...  $r=2$

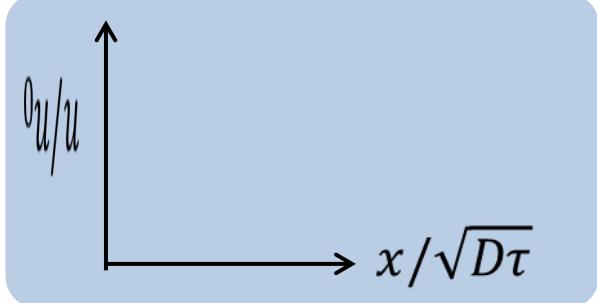
Scaled dimension ...  $(4-2)=2$

$D$  implies a second order term

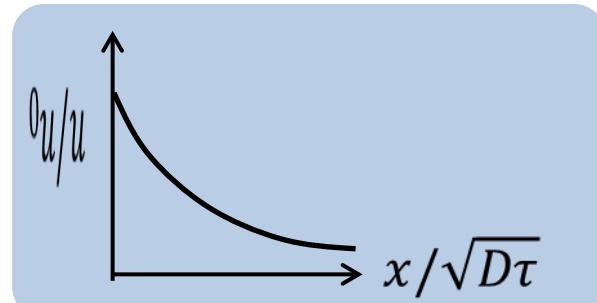
## Example 3: continued ...

$$\Pi_1 = f(\Pi_2) \equiv \frac{n}{n_0} = f\left(\frac{x}{\sqrt{D\tau}}\right)$$

D implies a second order term!



Scaling theory defines the axes



Experiments gets the function

Deep essence of machine learning ... you do not need to know the equation

# Discussion

- I. Widely used in fluid mechanics (Rayleigh, Reynold, Pandtl numbers), percolation theory, reliability problems, etc. Newton predicted bending of light by gravitational field simply by dimensional analysis. He was off by a factor of 2 compared to Einstein.

HW.  $F = f(D, V, \rho, \mu)$ , show that  $\frac{F}{\rho V^2 D^2} = f\left(\frac{\mu}{\rho V D}\right)$ .

We did not solve for Navier-Stokes equation.

Similitude explains why Wind-tunnels work. And why Wright brothers succeeded why others failed.

2. If you add extra variables, which are unimportant – they will either disappear or appear as normalized variable that will be shown to be irrelevant experimentally.
3. Related to the principle component analysis in a interesting way (e.g. ‘Recommended for you’ by Amazon and Netflix)

# Significant Dimensionless Group

$$\text{Viscous force: } \tau A = \mu \frac{du}{dy} A \propto \mu \frac{V}{L} L^2 = \mu V L$$

$$\text{Gravity force: } mg \propto g \rho L^3$$

$$\text{Pressure force: } (\Delta p)A \propto (\Delta p)L^2$$

$$\text{Surface tension force: } \sigma L$$

$$\text{Compressibility force: } E_v A \propto E_v L^2$$

$$Re = \frac{\rho \bar{V} D}{\mu} = \frac{\bar{V} D}{\nu}$$

$$Eu = \frac{\Delta p}{\frac{1}{2} \rho V^2}$$

$$M = \frac{V}{c} = \frac{V}{\sqrt{\frac{dp}{d\rho}}} = \frac{V}{\sqrt{\frac{E_v}{\rho}}}$$

$$Ca = \frac{p - p_v}{\frac{1}{2} \rho V^2}$$

$$Fr = \frac{V}{\sqrt{gL}}$$

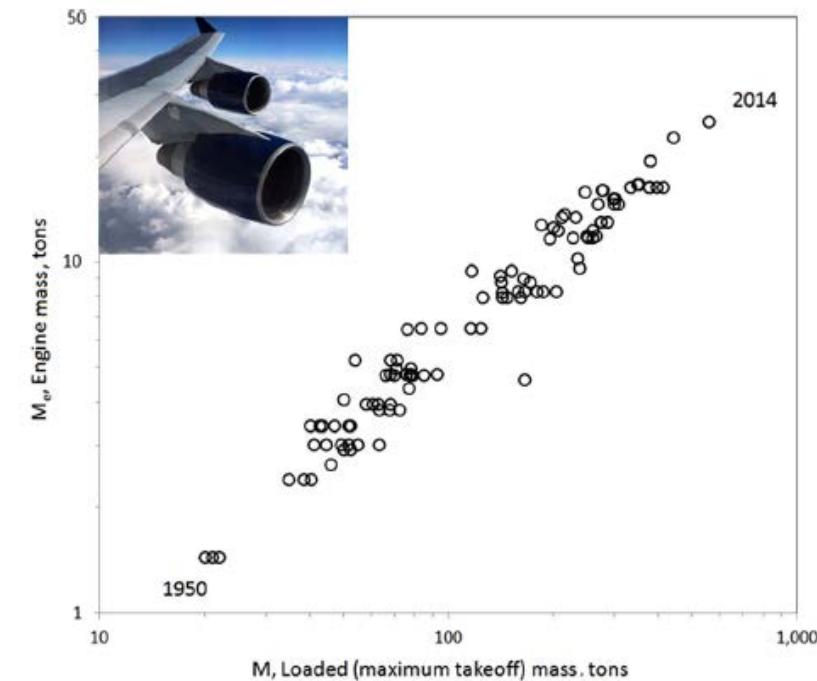
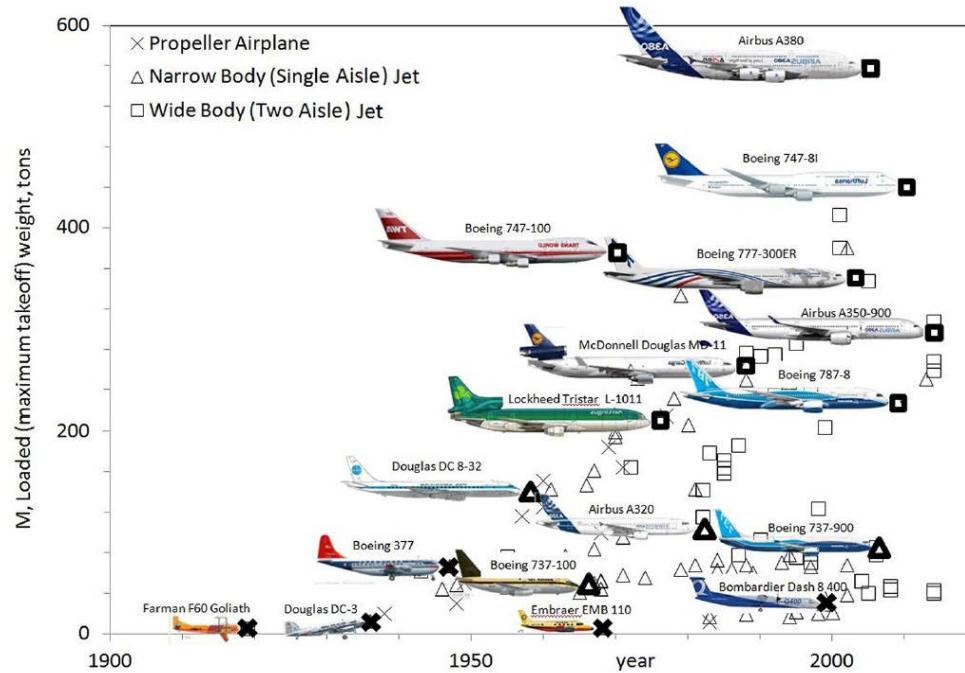
$$M^2 = \frac{\rho V^2 L^2}{E_v L^2}$$

$$We = \frac{\rho V^2 L}{\sigma}$$

$$Fr^2 = \frac{V^2}{gL} = \frac{\rho V^2 L^2}{\rho g L^3}$$

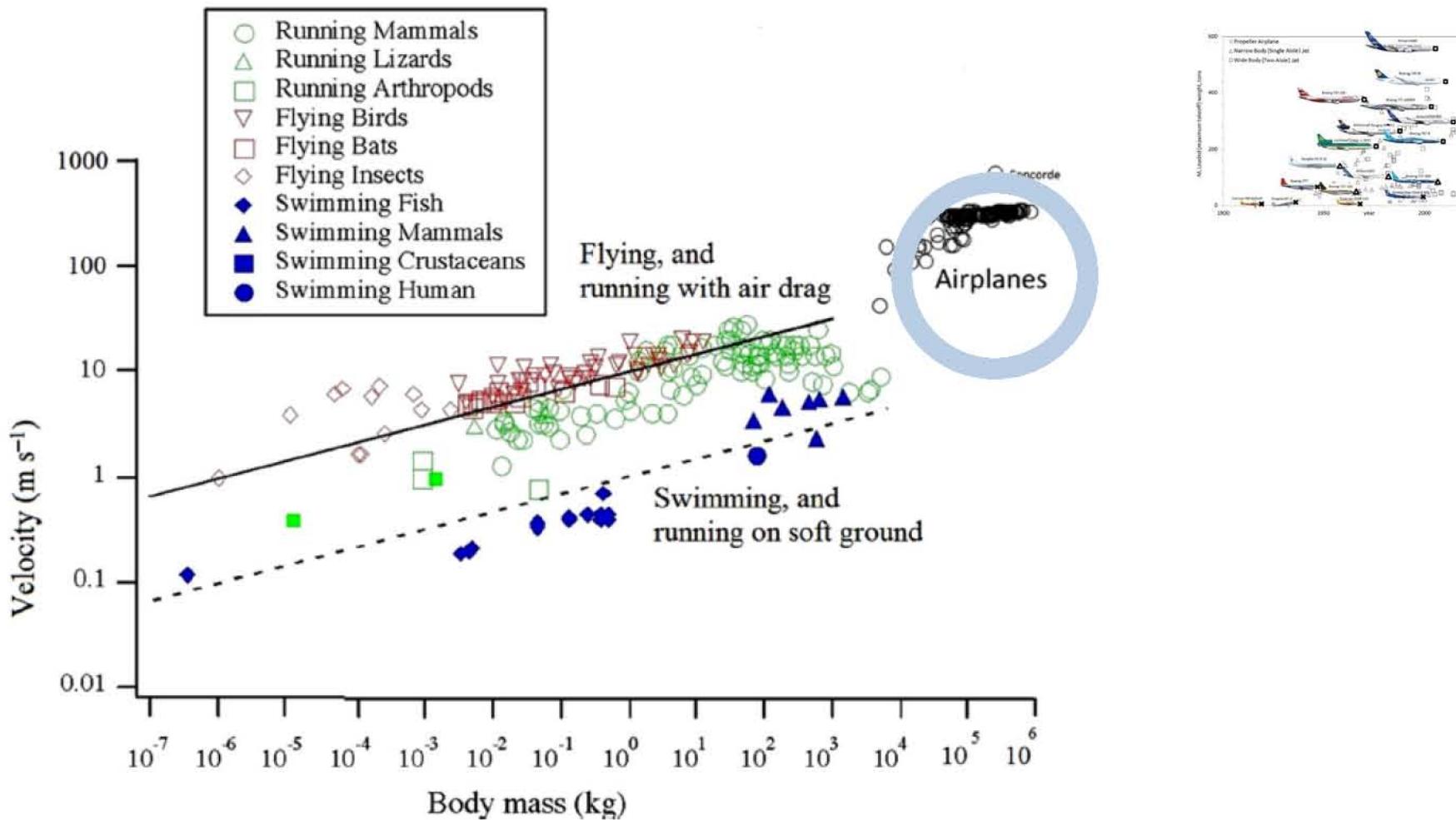
# The evolution of airplanes

A. Bejan et al. JAP, 116, 044901, 2014



$$\text{Engine lifts 10 times its mass: } M_e = 0.13 M^{0.83}$$

# Scaling theory of things that move



# Conclusions

1. The laws of physics must be non-dimensional, because otherwise laws will depend where you are in the universe.
2. Scaling of variables is a very important way of reducing the number of variables in an experiment. However, scaling requires that we have some idea about the key variables.
3. There are many applications of the scaling theory, especially in fluid mechanics, phase transition, percolation theory, etc. The problem is complex, similar to that of reliability, and therefore scaling provides enormous simplification.
4. Some of the problems may not be fully specified in terms of explicitly stated variables. The Fisher/Taguchi method help design those experiments.

# References

Dimensional Analysis:

Most books on Fluid mechanics has a chapter on “Dimensional Analysis”. See for example,[http://en.wikibooks.org/wiki/Fluid\\_Mechanics/Dimensional\\_Analysis](http://en.wikibooks.org/wiki/Fluid_Mechanics/Dimensional_Analysis)

One of the best articles on this topic is by D. Bolster, R. Hershberger, and R. J. Donnelly, Physics Today, p. 42, 2011.  
[http://astro.berkeley.edu/~eliot/Astro202/dimensional\\_PhysicsToday.pdf](http://astro.berkeley.edu/~eliot/Astro202/dimensional_PhysicsToday.pdf)

The reliability example I used is from a bookchapter on “Some Unifying Concepts in Reliability Physics, Mathematical Models, and Statistics” by R. E. Thomas,

Another excellent book is Introduction to Fluid mechanics, 5<sup>th</sup> Edition written by R.W. Fox and A.T. McDonald, John Wiley and Sons, Inc.

# Review Questions

1. How does one choose the variables in for Buckingham Pi analysis?
2. Convince yourself that Rayleigh, Reynold, Womersley, Prandtl numbers are dimensionless.
3. Use the example of wind-tunnel and Dennard scaling to argue that scaling reduces the number of experiments while simultaneously reducing the physical size of the experimental setup.
4. The scaling theory may give fundamentally wrong results, if the choice of the variables is not guided by physical insights of the problem. Explain.
5. What are the advantages and disadvantages of nature units?
6. What is the difference among geometric similarity, kinetic similarity and dynamic similarity?

# An Introduction to Data Analysis, Design of Experiment, and Machine Learning

## *Lecture 7. Bootstrap, Cross-Validation, and Goodness of Fit*

Muhammad A. Alam  
[alam@purdue.edu](mailto:alam@purdue.edu)



copyright 2018

This material is copyrighted by M. Alam under the following Creative Commons license:



**Attribution-NonCommercial-ShareAlike 2.5 Generic (CC BY-NC-SA 2.5)**

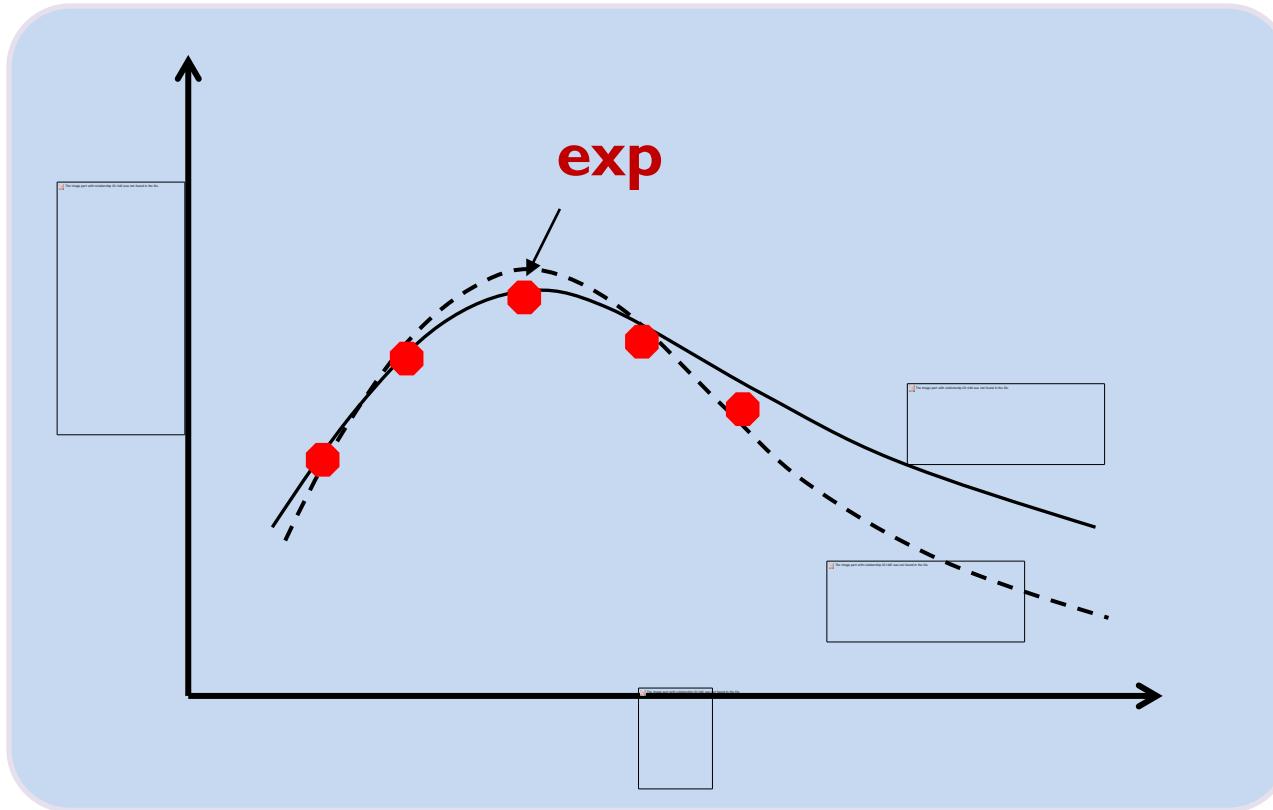
Conditions for using these materials is described at

<http://creativecommons.org/licenses/by-nc-sa/2.5/>

# Outline

1. Introduction
2. Goodness of Fit: Adjusted R-square, AIC methods, etc.
3. Cross-validation: Another way to compare models
4. Bootstrap method to generation population properties based on sample characteristics
5. Parametric vs. non-parametric distribution
6. Conclusions

# Recall: MLE can be used to fit any model to the data



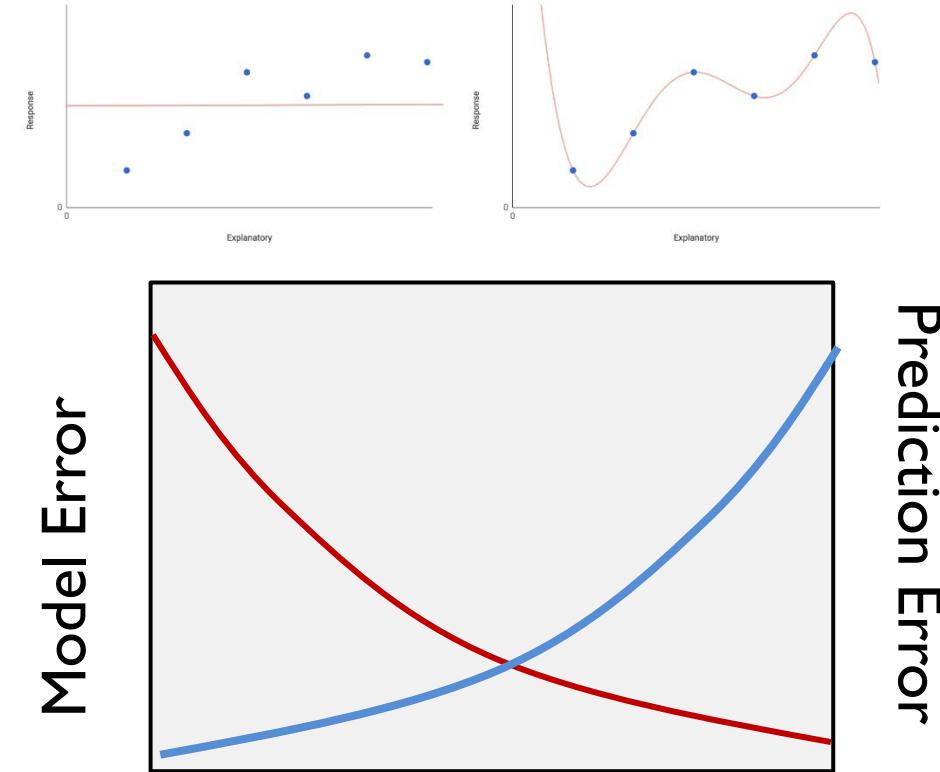
Each model can be checked for  $\chi^2$ ,  $KS$ , or  $QQ$  tests.  
What if two or more models passes the test. Which one is better?

# Principle of Parsimony

Aristotle: Nature operates in the shortest way possible.

George Box: All models are wrong, but some are useful.

Occam's Razor: “given two or more equally acceptable explanations for a phenomenon, work with the one which introduces the fewest assumptions.”



Model Complexity  
Defined by # parameters

# Parameter number vs. goodness of fit

$n$  = number of samples,  $M$ =number of parameters

I) Method of adjusted residual ...

$$R_{adj}^2 = \frac{(n-1)R^2 - (M-1)}{n-M}$$

$$M \rightarrow p + 1$$

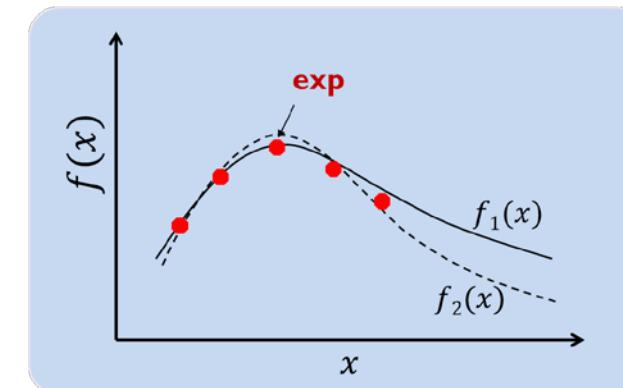
2) Akaike Information Criterion

$$AIC = n \times \ln(R^2/n) + 2M$$

2) Schwarz Information Criterion

$$BIC = n \times \ln(R^2/n) + M \times \ln n$$

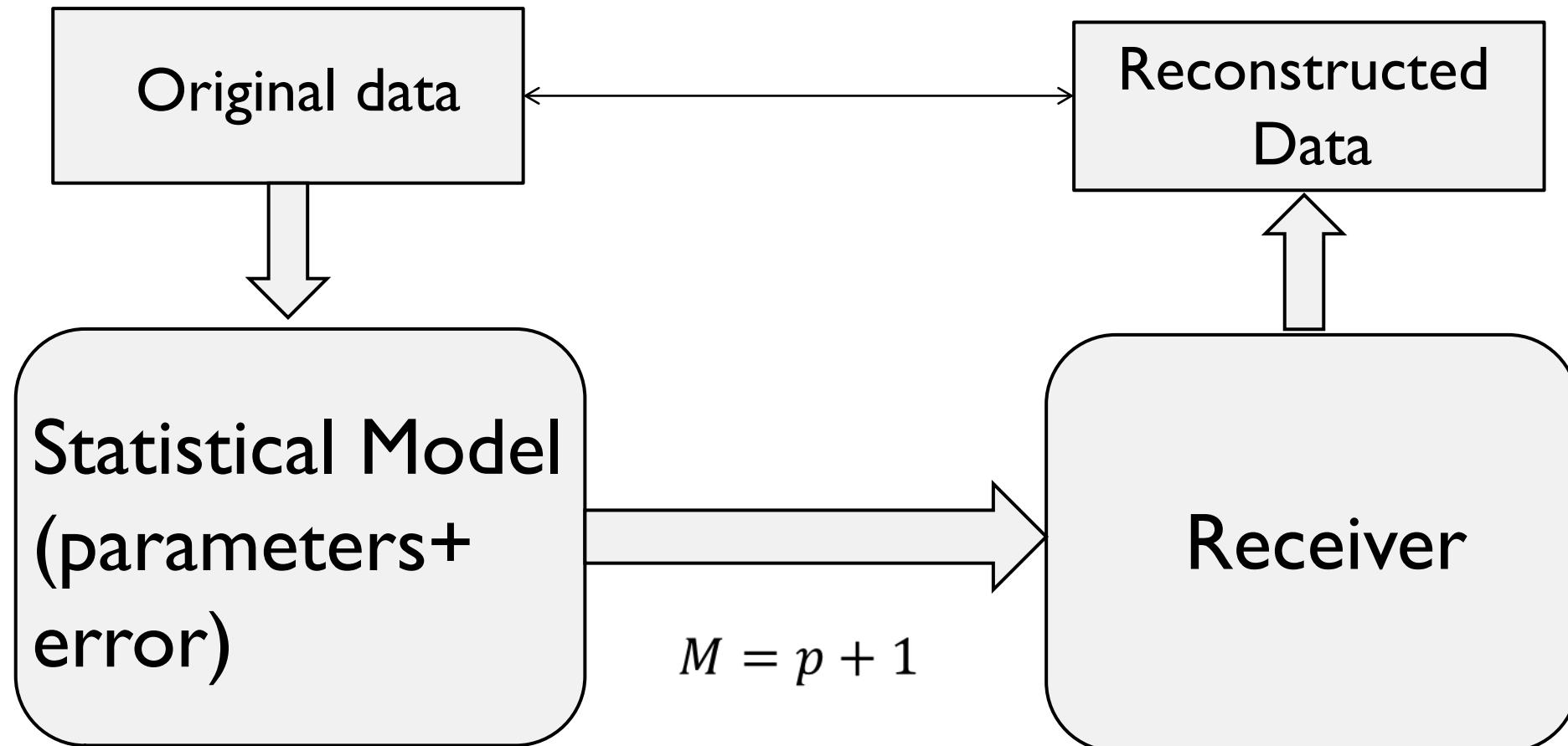
$$R \equiv \sum_{n=1}^n (t_i - t_{i,fit})^2$$



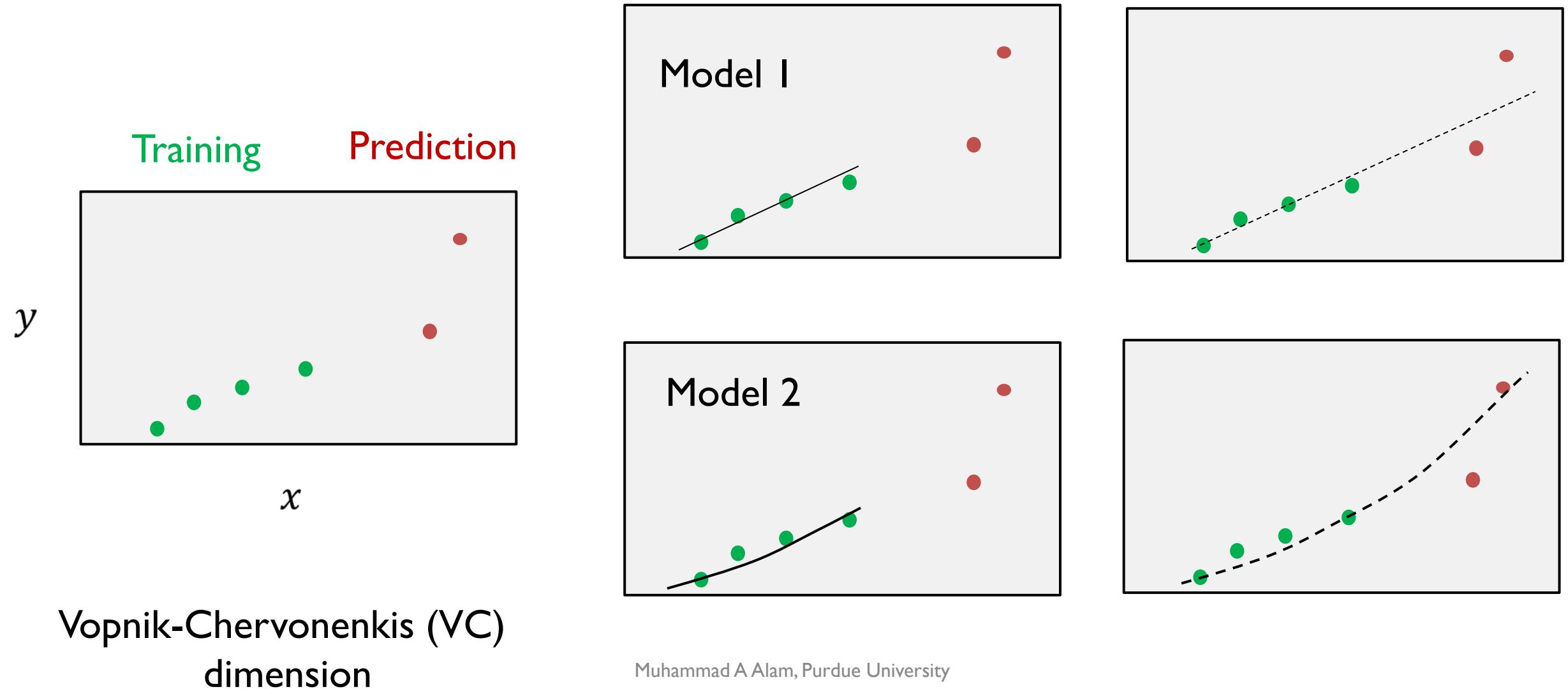
Error penalty  
Parameter Penalty

Ref. Les Kirkup, *Data Analysis with Excel*,  
Cambridge Univ. Press. P. 304

# Essence of the information theoretic approach



# Cross validation method



# Statistics of Sample vs. Population

Distribution-free statistical measure of data ....

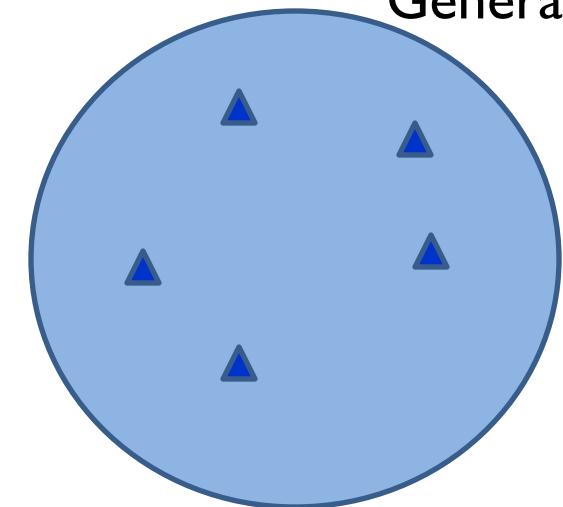
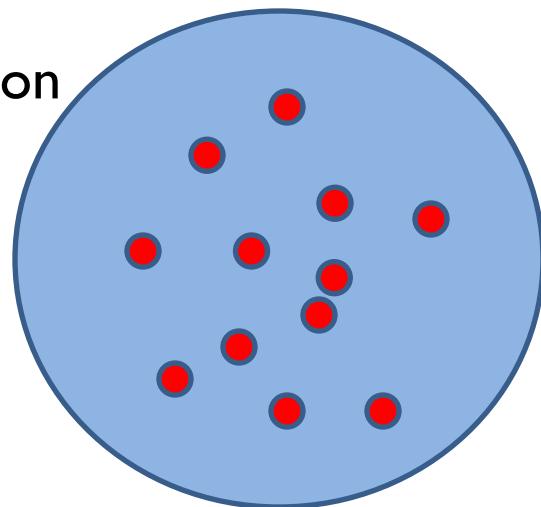
$$\langle t \rangle = \frac{\sum_{j=1,N} t_j}{N}$$
$$s^2 = \frac{\sum_{j=1,N} (t_j - \langle t \rangle)^2}{N-1}$$

Parameter-space

$$\delta_{T_k} = \sqrt[k]{\frac{\sum_{j=1}^N (t_i - \langle t \rangle)^k}{N-k+1}}$$

General formula

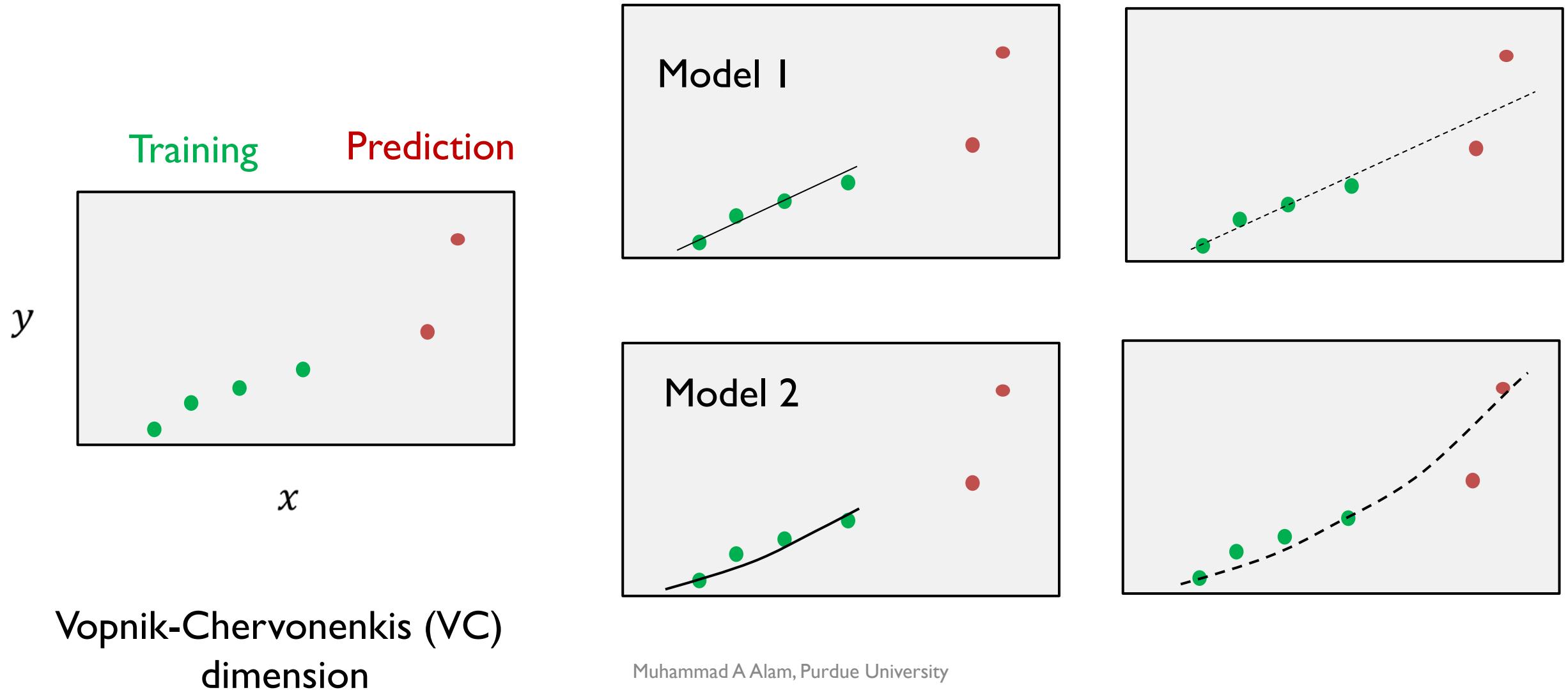
Population



Parameter

Similar to Fourier Series, First used by Brahe for Alpha Aretis  
Good for comparison, but not appropriate for projection

# Cross validation method



# Outline

1. Introduction
2. Goodness of Fit: Adjusted R-square, AIC methods, etc.
3. Cross-validation: Another way to compare models
4. **Bootstrap method** to generate population properties based on sample characteristics
5. Parametric vs. non-parametric distribution
6. Conclusions

# Statistics of Sample vs. Population

Distribution-free statistical measure of data ....

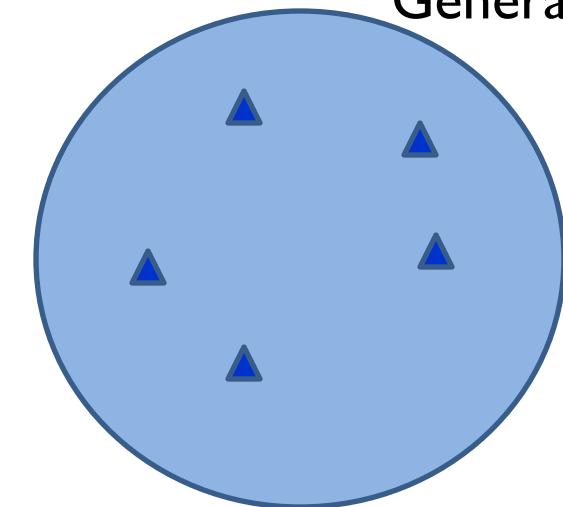
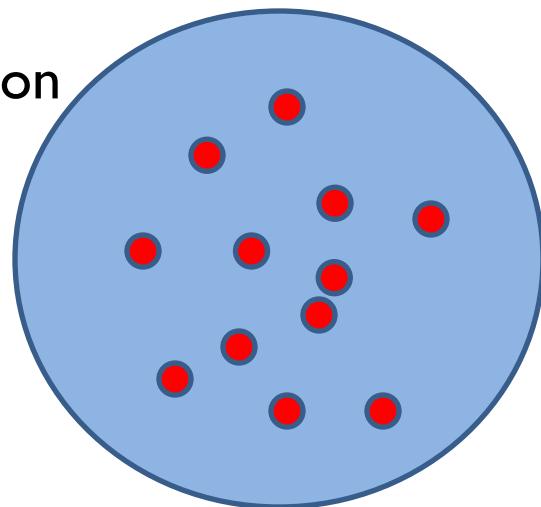
$$\langle t \rangle = \frac{\sum_{j=1,N} t_j}{N}$$
$$s^2 = \frac{\sum_{j=1,N} (t_j - \langle t \rangle)^2}{N-1}$$

Parameter-space

$$\delta_{T_k} = \sqrt[k]{\frac{\sum_{j=1}^N (t_i - \langle t \rangle)^k}{N-k+1}}$$

General formula

Population

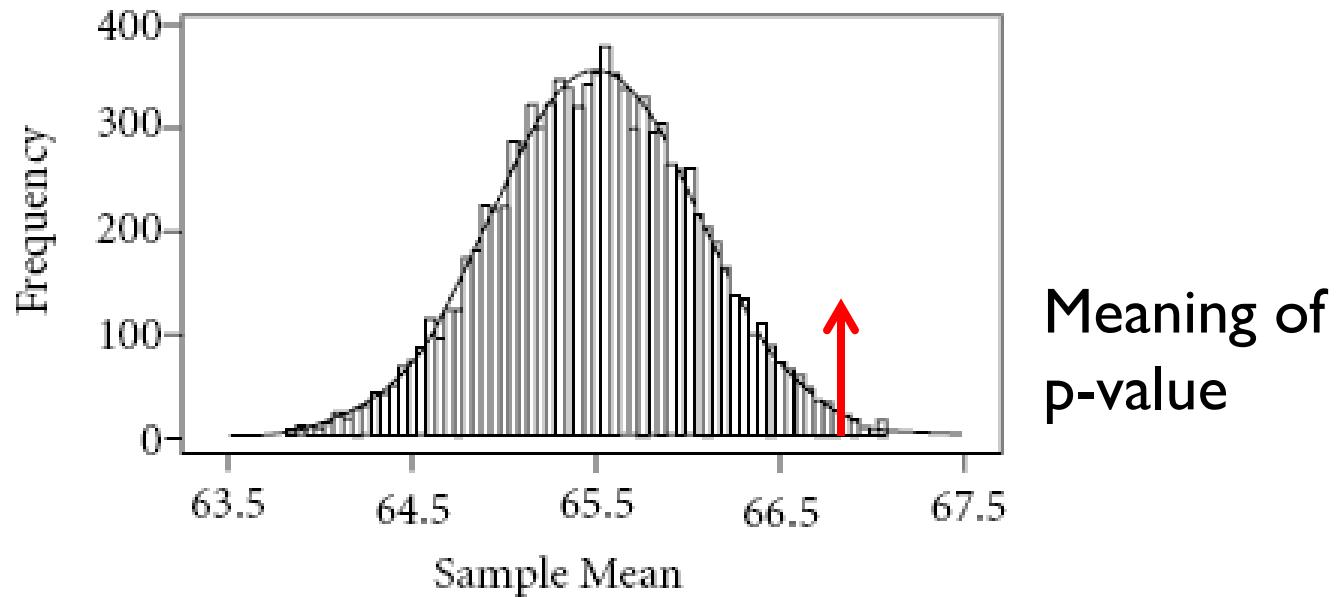


Parameter

Similar to Fourier Series, First used by Brahe for Alpha Aretis  
Good for comparison, but not appropriate for projection

# Distribution of the Sample Statistic/Moment (e.g. Mean)

Sample Size =20  
Number of samples=10k (from population)

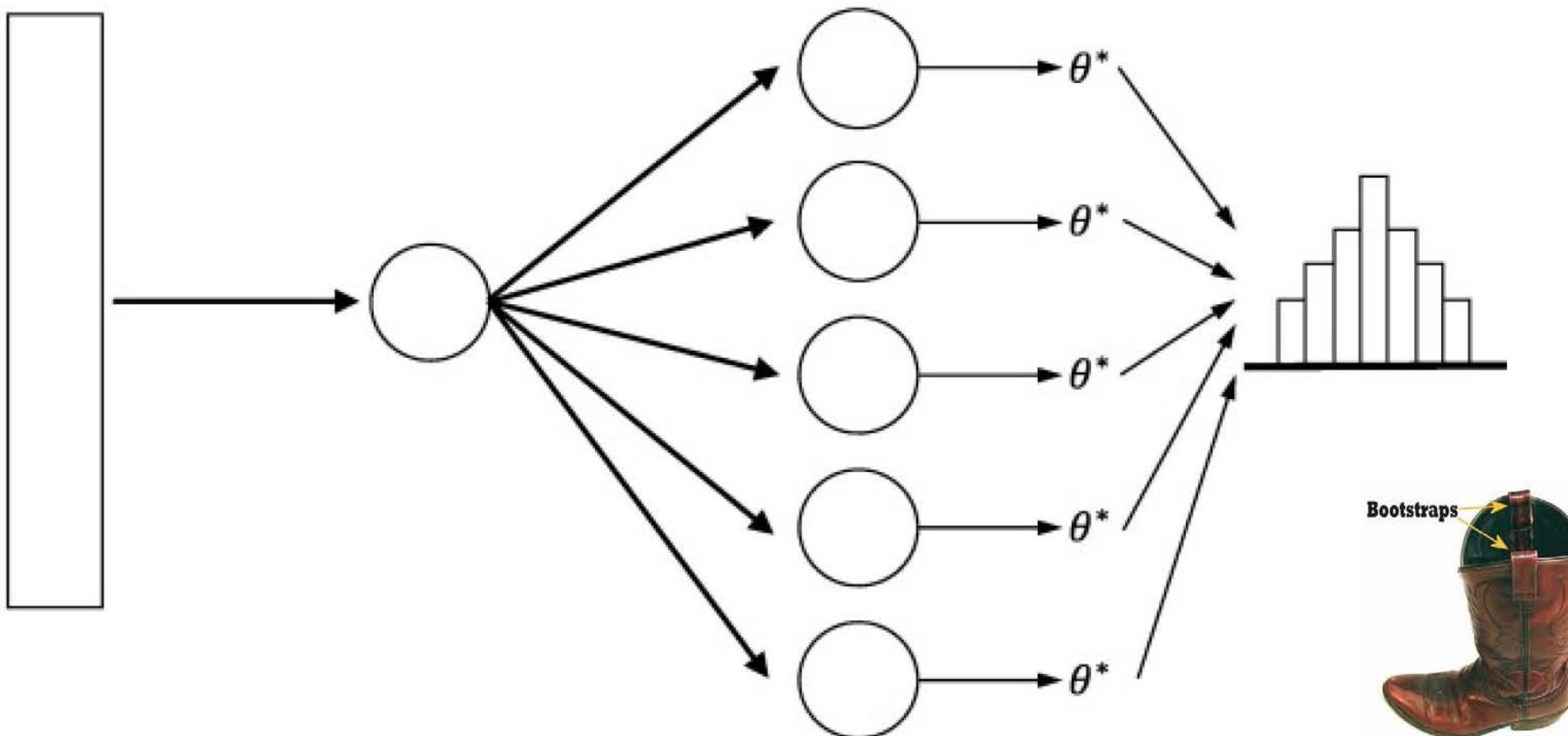


$$\begin{aligned}\mu_x &= \mu \\ \sigma_x &= \sigma / \sqrt{N}\end{aligned}$$

$$\begin{aligned}Z &= (X - \mu) / (\sigma / \sqrt{N}) & N > 30 \\ Z &= (X - \mu) / (s / \sqrt{N}) & N < 30\end{aligned}$$

# Overall algorithm for bootstrapping

population      sampling      measured sample      re-sampling with replacement      bootstrap samples      statistical quantities of interest      distribution analysis



# Working with a single sample

0.2 -0.1 0.5 0.3 -0.6

All you have is a single sample ..

Generate synthetic samples from the original (with replacement)

0.2 -0.1 -0.6 -0.1 0.5

Synthetic sample 1

0.3 0.2 -0.6 0.2 0.5

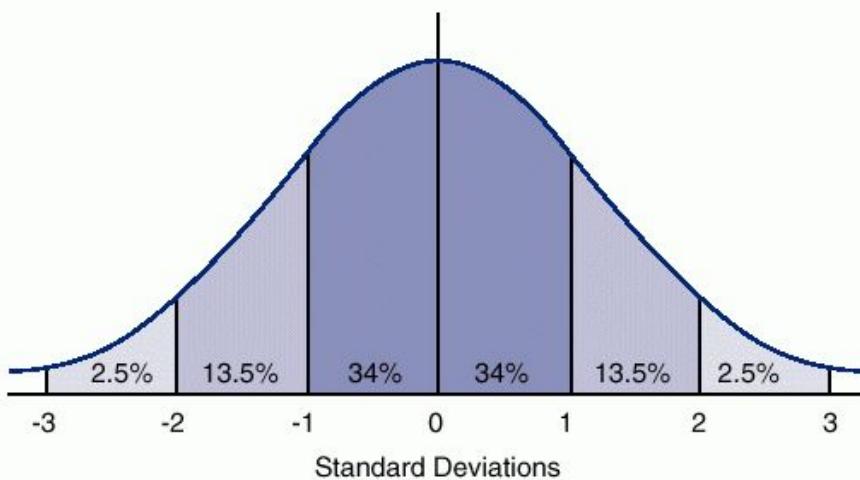
Synthetic sample 2

0.5 -0.1 0.5 0.2 0.3

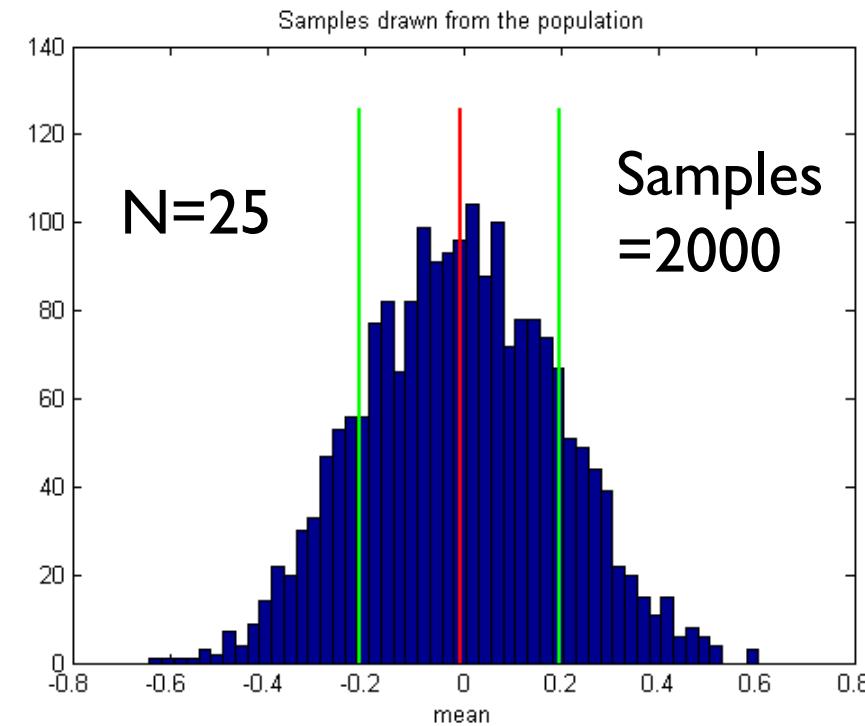
Synthetic sample 3



# Bootstrap method - Introduction

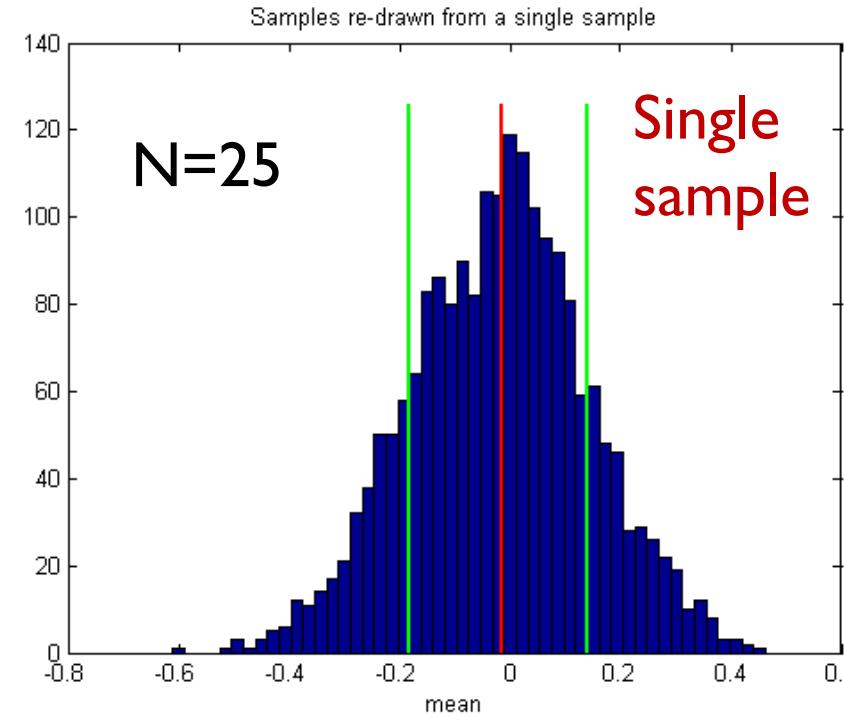
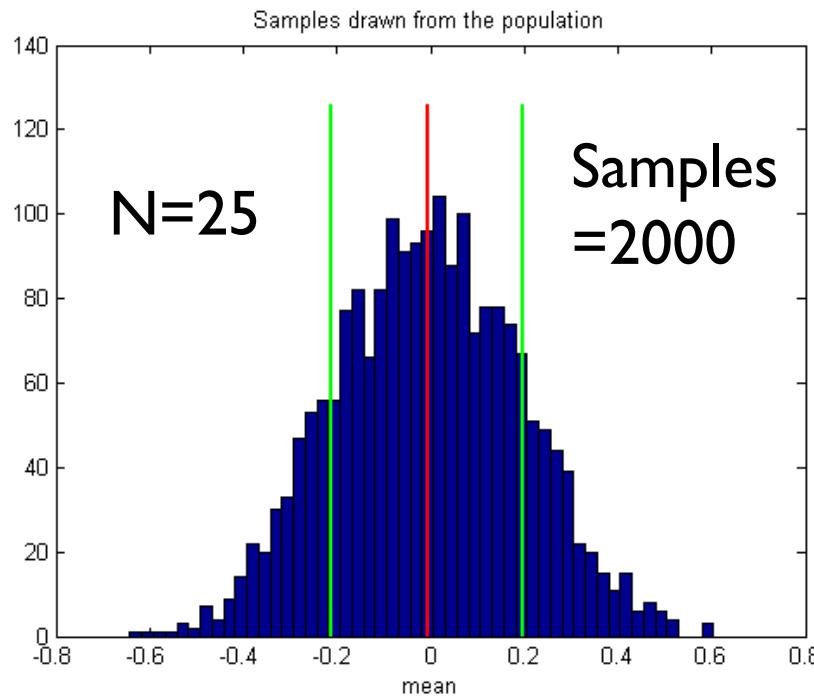


68% between +0.2 to -0.2  
95% between +0.4 to -0.4



$$s = \sqrt{\frac{\sum_{j=1, N=25} (t_j - \langle t \rangle)^2}{N-1}} \sim \sqrt{\frac{1}{24}} \sim 0.2$$

# Multiple sample vs. single sample



Bootstrap average is not zero!

And yet, the  $s \sim 0.18$ , just from a single sample.

The success of the method relies on precision measurement

# Parametric vs. non-parametric Bootstrap

0.2 -0.1 0.5 0.3 -0.6    Fit the distribution of your choice by  
Maximum likelihood estimators (MLE)  
(obtain parameters, i.e.  $\eta_0, \beta_0$ )

Generate synthetic samples based on the parametric distribution

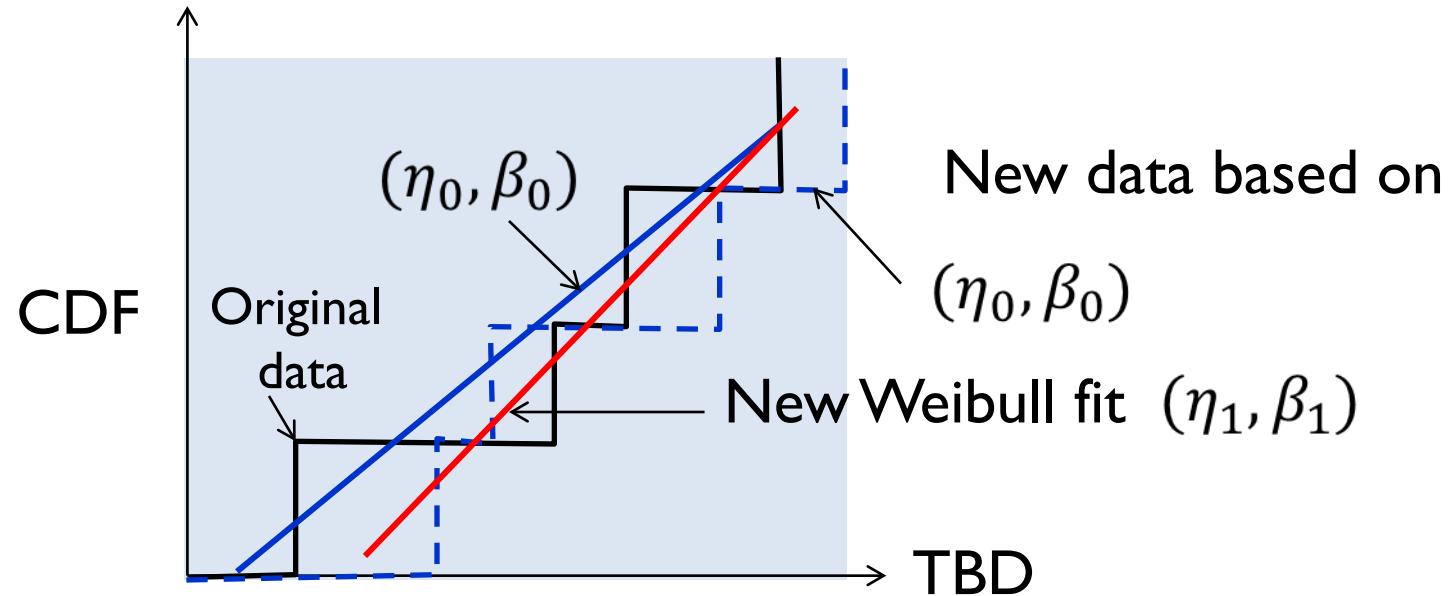
0.12 -0.17 -0.44 -0.71 0.52    Synthetic sample 1 (new  $\eta_1, \beta_1$ )

0.32 0.21 -0.69 0.23 0.58    Synthetic sample 2 (new  $\eta_2, \beta_2$ )



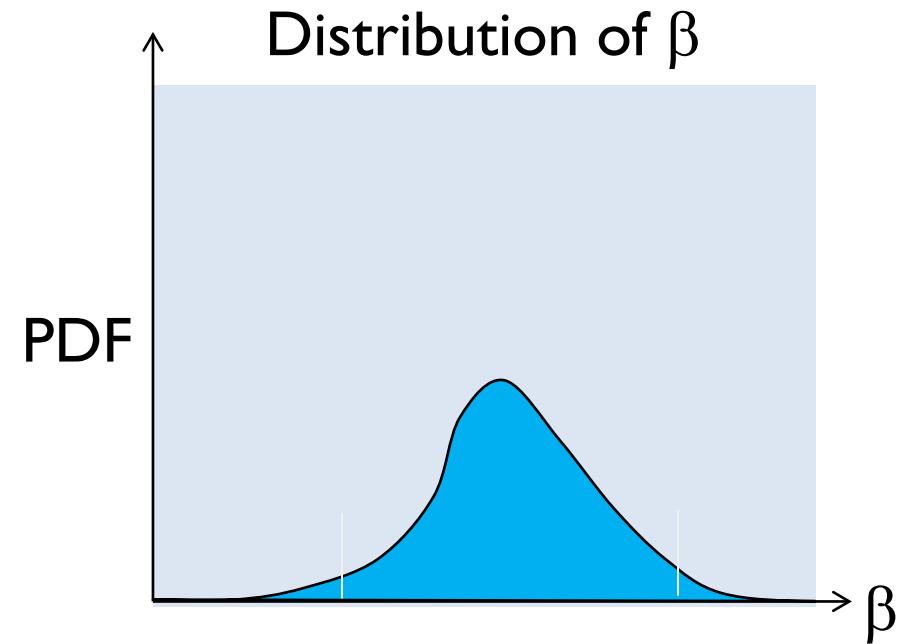
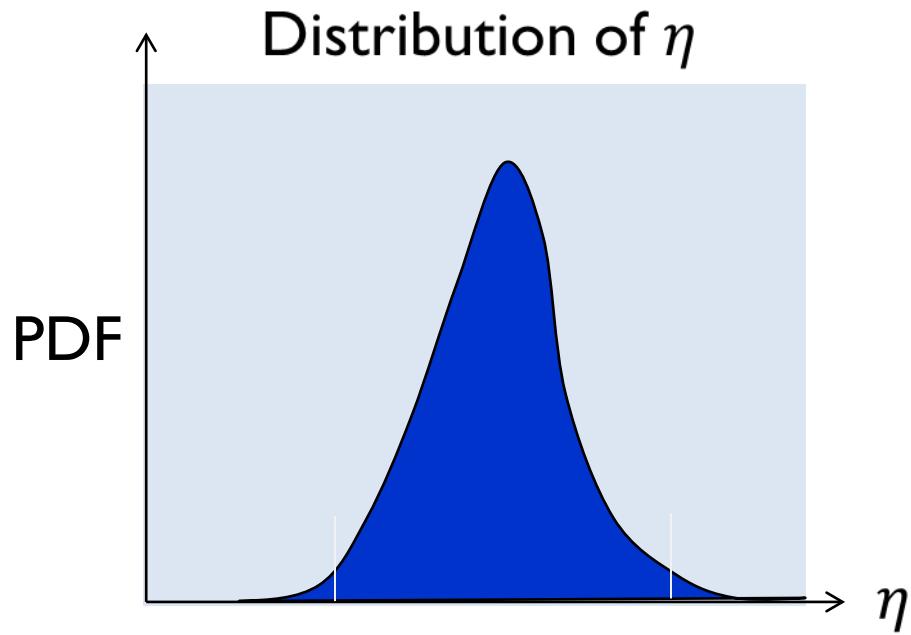
Plot distribution of statics  $\eta_i, \beta_i$

# Why resampling from the same distribution generates new fit parameters



Samples taken from the same distribution  $(\eta_0, \beta_0)$  generates datapoints that are fitted with new  $(\eta_i, \beta_i)$

# Distributions of $\alpha$ and $\beta$



Same technique for polling and tenure rate of faculty!

# Conclusions

1. If you have physics on your side, the game is over. If physics is unknown, then computer-based model comparison is helpful.
2. The approach selects the best among two or more models based on the principle of parsimony. It can be formulated in terms of information theoretic approach or cross validation approach.
3. The second approach relies on the Bootstrap or Monte Carlo approach. Here one generates the population statistics based on single sample. And then uses parametric or non-parametric approaches to define the goodness of fit.
4. In parametric bootstrap, you compute coefficient variability without having to calculate a complex function.

# References

Goutte, Cyril. "Note on free lunches and cross-validation." *Neural Computation* 9.6 (1997): 1245-1249.

Wolpert, David H., and William G. Macready. *No free lunch theorems for search*. Vol. 10. Technical Report SFI-TR-95-02-010, Santa Fe Institute, 1995.

Efron, Bradley, and Gail Gong. "A leisurely look at the bootstrap, the jackknife, and cross-validation." *The American Statistician* 37.1 (1983): 36-48.

Efron, Bradley. "How biased is the apparent error rate of a prediction rule?." *Journal of the American statistical Association* 81.394 (1986): 461-470.

Diaconis, Persi, and Bradley Efron. "Computer-intensive methods in statistics." *Scientific American* 248.5 (1983): 116-131.

A simple and gentle tutorial on Vapnik-Chervonenkis (VC) dimension part 1 and part 2 of 2 on [November 11, 2013](#) by [panthimanshu17](#)

Gershenfeld, Neil A., and Neil Gershenfeld. *The nature of mathematical modeling*. Cambridge university press, 1999.

Bollen, Kenneth A., and Robert A. Stine. "Bootstrapping goodness-of-fit measures in structural equation models." *Sociological Methods & Research* 21.2 (1992): 205-229.

Gershenfeld, Neil A. "Dimension measurement on high-dimensional systems." *Physica D: Nonlinear Phenomena* 55.1-2 (1992): 135-154.

Cross-validation was developed by S. Geisser (U. Minnesota), M. Stone (U. of London), G. Wahba (U. of Wisconsin)

Jack-knife was developed by Maurice Quenouille and J.W. Tukey (Princeton/Bell Labs)

All these are precursors to machine learning

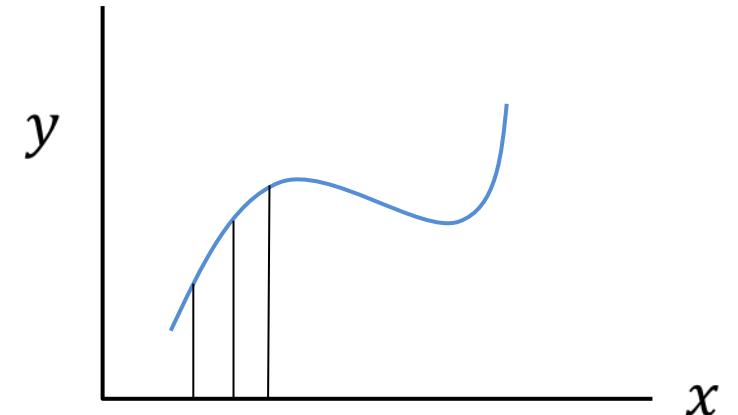
# An Introduction to Data Analysis, Design of Experiment, and Machine Learning

## *Lecture 8. Statistical Design of Experiments*

Muhammad A. Alam  
[alam@purdue.edu](mailto:alam@purdue.edu)



# The story so far ...



## Lecture 8-14

How to get a better  $f$

## Lectures 6-7

$\bar{x} = x_1, x_2, \dots x_n$   $\longrightarrow$

$$\bar{y} = f(\bar{x})$$

## Lectures I-2

$\bar{y} = y_1, y_2, \dots y_m$   $\longrightarrow$

## Lectures 3-5

How to fit multiple hypothetical function  $f$  to the same  $y$

# Design of Experiments

- Set of guidelines for designing, conducting and analyzing experiments for system optimization
- Foundations of DOE were laid by Sir. R.A. Fisher in early 1920s (Analysis of Farm data, output as good as input).
- Concepts of Orthogonal arrays were introduced by Taguchi in 1950s. (Formalized the whole analysis)
- DOE has revolutionized quality control/reliability in all fields of science and technology (Toyota was one of the early adopter, most semiconductor companies use the method).

# Philosophical shift with DOE

Before Fisher ...

Experimentalist  
determine what  
experiments to do

Results

Statisticians/  
Theorists/Expt  
collaborate to  
interpret results

After Fisher ...

Statisticians/  
Theorists/Expt  
plan what  
experiments to do

Results

Statisticians/  
Theorists/Expt  
collaborate to  
interpret results

Output cannot be greater than input .....

# Problem definition

(A) Oxide Thickness

(B) Doping

(C) Anneal temp.

(D) Junction depth

(E) Gate Overlap

(F) Halo implant

(G) Supply voltage

MOSFET

Drain current

7 parameter optimization for a single  
objective function

Could be car, farm, airport

# Definition of terms

Factor	Level	Run/trial/replicate
Tox	1, 2, 3 nm	$(2 \text{ nm}, 10^{17} \text{ cm}^{-3}, 4 \mu\text{m})_{\text{rep}}$
Doping	$10^{16}, 10^{17} \text{ cm}^{-3}$	
Lch	2, 3, 4 $\mu\text{m}$	

- 1 factor, 3 level, 4 replicate experiment
- 2 factor, 2 level, 3 replicate experiment
- 8 factor, 2 level, 1 replicate experiment

# Puzzle Analogy: Many factors, 2 levels

## Graeco-Latin Squares Euler Squares

Land type ....A,B,C,D (Latin)  
Fertilizer ... a,b,c,d (Greek)

Aa	Bc	Cd	Db
Bb	Ad	Dc	Ca
Cc	Da	Ab	Bd
Dd	Cb	Aa	Ac

Randomization and  
statistical content

Sudoku

	2		7		4		9	
		5	6		9	2		
I								7
5			4		8			2
		2				6		
8			3		7			4
9								I
		8	I		2	3		
	4		9		5		8	

30 filled cells vs. 81 cells

3	2	6	7	8	4	I	9	5
7	8	5	6	I	9	2	4	3
I	9	4	2	5	3	8	6	7
5	I	7	4	6	8	9	3	2
4	3	2	5	9	I	6	7	8
8	6	9	3	2	7	5	I	4
9	5	3	8	7	6	4	2	I
6	7	8	I	4	2	3	5	9
2	4	I	9	3	5	7	8	6

# Outline

- I. Context and background
- 2. Single factor and full factorial method**
3. Orthogonal vector analysis: Taguchi/Fisher model
4. Correlation in dependent parameters
5. Conclusions

# 7 Factor, 2 level: One factor at a time

	A	B	C	D	E	F	G	Output
Run 1	1	1	1	1	1	1	1	10
Run 2	2	1	1	1	1	1	1	15
Run 3	2	2	1	1	1	1	1	12
Run 4	2	1	2	1	1	1	1	9
Run 5	2	1	1	2	1	1	1	18
Run 6	2	1	1	2	2	1	1	19
Run 7	2	1	1	2	2	2	1	17
Run 8	2	1	1	2	2	1	2	13
Final	2	1	1	2	2	1	1	19

Simple, widely used, but non-optimum solutions

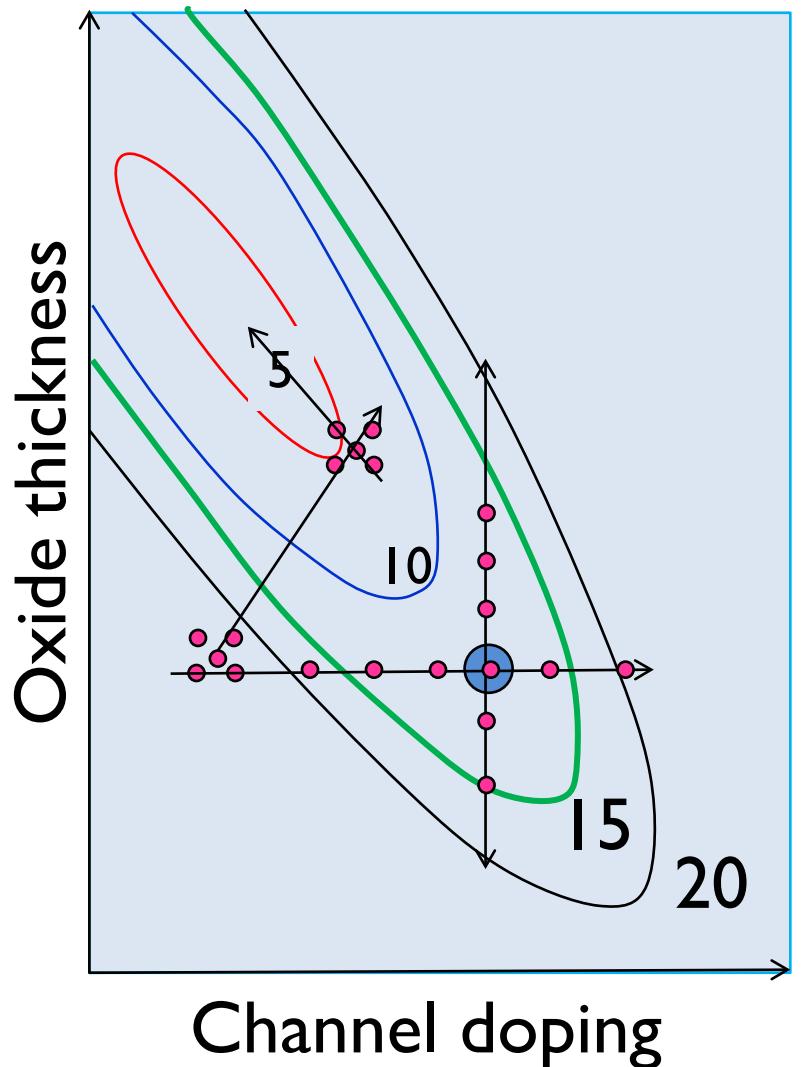
# 7 Factor, 2 Level: Full factorial analysis

				A <sub>1</sub>				A <sub>2</sub>			
				B <sub>1</sub>		B <sub>2</sub>		B <sub>1</sub>		B <sub>2</sub>	
				C <sub>1</sub>	C <sub>2</sub>						
D <sub>1</sub>	E <sub>1</sub>	F <sub>1</sub>	G <sub>1</sub>	R-1 (10)				R-2 (15)		R-3 (12)	
			G <sub>2</sub>								
		F <sub>2</sub>	G <sub>1</sub>					R-4 (9)			
			G <sub>2</sub>								
	E <sub>2</sub>	F <sub>1</sub>	G <sub>1</sub>								
			G <sub>2</sub>								
		F <sub>2</sub>	G <sub>1</sub>								
			G <sub>2</sub>								
D <sub>2</sub>	E <sub>1</sub>	F <sub>1</sub>	G <sub>1</sub>					R-5 (18)			
			G <sub>2</sub>								
		F <sub>2</sub>	G <sub>1</sub>								
			G <sub>2</sub>								
	E <sub>2</sub>	F <sub>1</sub>	G <sub>1</sub>					R-6 (19)			
			G <sub>2</sub>					R-8 (13)			
		F <sub>2</sub>	G <sub>1</sub>					R-7 (17)			
			G <sub>2</sub>								

Single parameter method is a fractional non-optimal factorial method: After A<sub>2</sub> win, will never visit A<sub>1</sub>. After B<sub>2</sub> loss, will never visit B<sub>2</sub>. Same for C<sub>2</sub> Column, etc.

$$\text{Level factor} = 2^7 = 128$$

# The problem with one-at-a-time approach



Response surface  
Orthogonal sampling

# Outline

- I. Context and background
2. Single factor and full factorial method
3. **Orthogonal vector analysis: Taguchi/Fisher model**
4. Correlation in dependent parameters
5. Conclusions

# Uncorrelated main effect (forward/backward)

4 factors for simplicity

	Level 1	Level 2
A	0	2
B	0	-6
C	0	4
D	0	-2

Full factorial (Only A)

		A <sub>1</sub>		A <sub>2</sub>	
		B <sub>1</sub>	B <sub>2</sub>	B <sub>1</sub>	B <sub>2</sub>
C <sub>1</sub>	D <sub>1</sub>	0	0	2	2
	D <sub>2</sub>	0	0	2	2
C <sub>2</sub>	D <sub>1</sub>	0	0	2	2
	D <sub>2</sub>	0	0	2	2

Full factorial (A and B)

		A <sub>1</sub>		A <sub>2</sub>	
		B <sub>1</sub>	B <sub>2</sub>	B <sub>1</sub>	B <sub>2</sub>
C <sub>1</sub>	D <sub>1</sub>	0	-6	2	-4
	D <sub>2</sub>	0	-6	2	-4
C <sub>2</sub>	D <sub>1</sub>	0	-6	2	-4
	D <sub>2</sub>	0	-6	2	-4



	Level 1	Level 2	L2-L1
A	-2	0	2
B	2	-4	-6
C	-3	1	4
D	0	-2	2

Full factorial (A, B, C and D)

		A <sub>1</sub>		A <sub>2</sub>	
		B <sub>1</sub>	B <sub>2</sub>	B <sub>1</sub>	B <sub>2</sub>
C <sub>1</sub>	D <sub>1</sub>	0	-6	2	-4
	D <sub>2</sub>	-2	-8	0	-6
C <sub>2</sub>	D <sub>1</sub>	4	-2	6	0
	D <sub>2</sub>	2	-4	4	-2

Full factorial (A, B and C)

		A <sub>1</sub>		A <sub>2</sub>	
		B <sub>1</sub>	B <sub>2</sub>	B <sub>1</sub>	B <sub>2</sub>
C <sub>1</sub>	D <sub>1</sub>	0	-6	2	-4
	D <sub>2</sub>	0	-6	2	-4
C <sub>2</sub>	D <sub>1</sub>	4	-2	6	0
	D <sub>2</sub>	4	-2	6	0

# Taguchi orthogonal array (L8 array)

	A	B	C	D	E	F	G
R-1	I	I	I	I	I	I	I
R-2	I	I	I	2	2	2	2
R-3	I	2	2	I	I	2	2
R-4	I	2	2	2	2	I	I
R-5	2	I	2	I	2	I	2
R-6	2	I	2	2	I	2	I
R-7	2	2	I	I	2	2	I
R-8	2	2	I	2	I	I	2

- 1) Check to see that for every factor, e.g. A, the rest of factors are fully randomized, e.g. every column sums to same number.
- 2) Does it remind you of Sudoku?
- 3) For smaller system (4 factors, 2 levels), choose the first four columns, ignore the remaining 3 – still need 8 experiments. For other systems, see ...

[http://www.freequality.org/sites/www\\_freequality\\_org/documents/tools/Tagarray\\_files/tamatrix.htm](http://www.freequality.org/sites/www_freequality_org/documents/tools/Tagarray_files/tamatrix.htm)

# Orthogonal measurements (uncorrelated)

				A <sub>1</sub>				A <sub>2</sub>			
				B <sub>1</sub>		B <sub>2</sub>		B <sub>1</sub>		B <sub>2</sub>	
				C <sub>1</sub>	C <sub>2</sub>						
D <sub>1</sub>	E <sub>1</sub>	F <sub>1</sub>	G <sub>1</sub>	R-1							
			G <sub>2</sub>								
		F <sub>2</sub>	G <sub>1</sub>								
			G <sub>2</sub>				R-3				
	E <sub>2</sub>	F <sub>1</sub>	G <sub>1</sub>								
			G <sub>2</sub>						R-5		
		F <sub>2</sub>	G <sub>1</sub>							R-7	
			G <sub>2</sub>								
D <sub>2</sub>	E <sub>1</sub>	F <sub>1</sub>	G <sub>1</sub>								
			G <sub>2</sub>							R-8	
		F <sub>2</sub>	G <sub>1</sub>					R-6			
			G <sub>2</sub>								
	E <sub>2</sub>	F <sub>1</sub>	G <sub>1</sub>				R-4				
			G <sub>2</sub>								
		F <sub>2</sub>	G <sub>1</sub>								
			G <sub>2</sub>	R-2							

$$Y_{A1} = (R1 + R3 + R2 + R4)/4 \quad Y_{C2} = (R3 + R4 + R5 + R6)/4$$

If the system optimizes for (A1 B2 C2 D2 E2 F1 G2)

$$Y = Y_M + (Y_{A1} - Y_M) + (Y_{B2} - Y_M) + (Y_{C2} - Y_M) + (Y_{D2} - Y_M) + \dots + (Y_{G2} - Y_M)$$

$$Y_M = (Y_{A1} + Y_{A2} + Y_{B1} + Y_{B2} + \dots + Y_{G1} + Y_{G2})/14.$$

# Outline

1. Context and background
2. Single factor and full factorial method
3. Orthogonal vector analysis: Taguchi/Fisher model
4. **Correlation in dependent parameters**
5. Conclusions

# Correlated effect & level factor

4 parameters for simplicity ...

	Level 1	Level 2
A	0	2
B	0	+6
C	0	4
D	0	-2
<b>B if A</b>	<b>0</b>	<b>-6</b>

Uncorrelated

		A <sub>1</sub>		A <sub>2</sub>	
		B <sub>1</sub>	B <sub>2</sub>	B <sub>1</sub>	B <sub>2</sub>
C <sub>1</sub>	D <sub>1</sub>	0	-6	2	-4
	D <sub>2</sub>	-2	-8	0	-6
C <sub>2</sub>	D <sub>1</sub>	4	-2	6	0
	D <sub>2</sub>	2	-4	4	-2

		A <sub>1</sub>		A <sub>2</sub>	
		B <sub>1</sub>	B <sub>2</sub>	B <sub>1</sub>	B <sub>2</sub>
C <sub>1</sub>	D <sub>1</sub>	0	6	2	-4
	D <sub>2</sub>	-2	4	0	-6
C <sub>2</sub>	D <sub>1</sub>	4	10	6	0
	D <sub>2</sub>	2	8	4	-2

$$A_2 + B_2(\text{Given } A_2) + C_2 + D_2$$

$$\text{e.g. } A_1 B_1 = (0-2+4+2)/4 = 1$$

Are the variables correlated?

Define Level factors  $A_m B_m = (A \ B)_1$

Define  $A_m B_n = (A \ B)_2$

# Correlated effect & level factor

		A <sub>1</sub>		A <sub>2</sub>	
		B <sub>1</sub>	B <sub>2</sub>	B <sub>1</sub>	B <sub>2</sub>
C <sub>1</sub>	D <sub>1</sub>	0	6	2	-4
	D <sub>2</sub>	-2	4	0	-6
C <sub>2</sub>	D <sub>1</sub>	4	10	6	0
	D <sub>2</sub>	2	8	4	-2

Example of pair correlation:

$$A_1 B_1 = 1, A_1 B_2 = 7$$

$$A_2 B_1 = 3, A_2 B_2 = -3$$

$$\text{Pair Corr} \dots \quad \text{Corr}_{AB} = \sum_{i,j=1,2} A_i B_j (-1)^{(i+j)}$$

$$\text{Third order} \dots \quad \text{Corr}_{ABC} = \sum_{i,j,k=1,2} A_i B_j C_k (-1)^{(i+j+k)}$$

$$\text{Fourth order} \dots \quad \text{Corr}_{ABCD} = \sum_{i,j,k,p=1,2} A_i B_j C_k D_p (-1)^{(i+j+k+p)}$$

$$\text{Corr}_{AB} = \sum_{i,j} A_i B_j (-1)^{(i+j)} = A_1 B_1 - A_1 B_2 - A_2 B_1 + A_2 B_2 = -12$$

$$\text{Corr}_{AC} = \sum_{i,j} A_i C_j (-1)^{(i+j)} = 0$$

$$\text{Corr}_{AD} = \sum_{i,j} A_i D_j (-1)^{(i+j)} = 0$$

$$\text{Corr}_{BC} = \sum_{i,j} B_i C_j (-1)^{(i+j)} = 0$$

$$\text{Corr}_{BD} = \sum_{i,j} B_i D_j (-1)^{(i+j)} = 0$$

$$\text{Corr}_{CD} = \sum_{i,j} C_i D_j (-1)^{(i+j)} = 0$$

A is correlated to B ... There are no other pair correlation

# Correlated effect & level factor

		A <sub>1</sub>		A <sub>2</sub>		
		B <sub>1</sub>	B <sub>2</sub>	B <sub>1</sub>	B <sub>2</sub>	
C <sub>1</sub>	D <sub>1</sub>	0	6	2	-4	
	D <sub>2</sub>	-2	4	0	-6	
C <sub>2</sub>	D <sub>1</sub>	4	10	6	0	
	D <sub>2</sub>	2	8	4	-2	

Pair Corr .....  $Corr_{AB} = \sum_{i,j=1,2} A_i B_j (-1)^{(i+j)}$

Third order ...  $Corr_{ABC} = \sum_{i,j,k=1,2} A_i B_j C_k (-1)^{(i+j+k)}$

Fourth order ...  $Corr_{ABCD} = \sum_{i,j,k,p=1,2} A_i B_j C_k D_p (-1)^{(i+j+k+p)}$

Third order correlation:

A<sub>2</sub>B<sub>2</sub>C<sub>2</sub> = -1, A<sub>2</sub>B<sub>2</sub>C<sub>1</sub> = -5, etc.

$$Corr_{ABC} = \sum_{i,j,k=1,2} A_i B_j C_k (-1)^{(i+j+k)} = A_2 B_2 C_2 - A_1 B_2 C_2 \dots$$

$$= -1 - (+5) + (1) - (-5) + (+3) - (+9) - (-1) + (+5) = 0$$

$$Corr_{ABCD} = \sum_{i,j,k,p=1,2} A_i B_j C_k D_p (-1)^{(i+j+k)} = A_2 B_2 C_2 D_2 - A_1 B_2 C_2 D_2 \dots = 0$$

No third or fourth order correlation ....

# How to fix for correlation

4 parameters for simplicity ...

	Level 1	Level 2
A	0	2
B	0	+6
C	0	4
D	0	-2
B given A	0	-6

$$\text{Corrected} = B_2 - B_1 - \frac{\sum_{A,C,D} \text{all B interaction}}{2}$$

$$= 2 - 2 - (-12 / 2) = +6$$

		A <sub>1</sub>		A <sub>2</sub>	
		B <sub>1</sub>	B <sub>2</sub>	B <sub>1</sub>	B <sub>2</sub>
C <sub>1</sub>	D <sub>1</sub>	0	6	2	-4
	D <sub>2</sub>	-2	4	0	-6
C <sub>2</sub>	D <sub>1</sub>	4	10	6	0
	D <sub>2</sub>	2	8	4	-2

$$\text{Corrected} = A_2 - A_1 - \frac{\sum_{B,C,D} \text{all A interaction}}{2}$$

$$= 0 - 4 - (-12 / 2) = +2$$

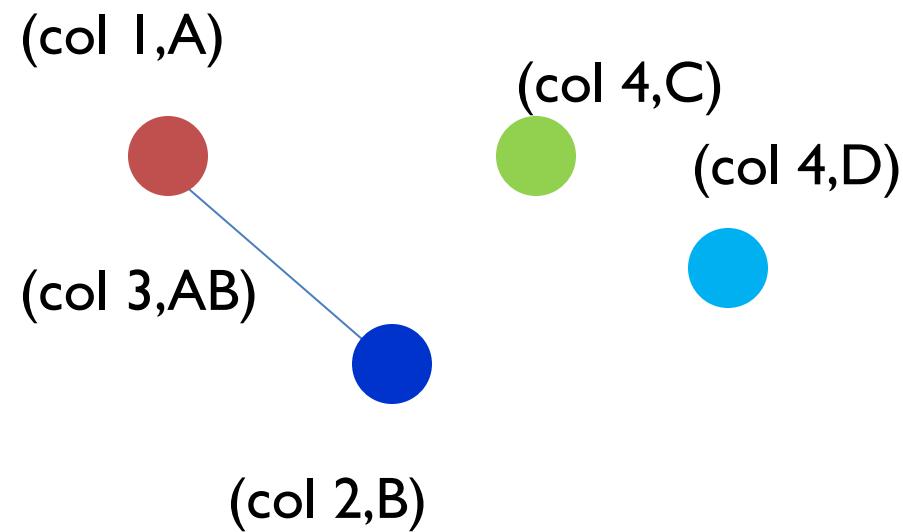
$$\text{Corr}_{AB} = \sum_{i,j} A_i B_j (-1)^{(i+j)} = -12$$

Only (AB) pair correlation found, no other correlation ....

# Aside: correlation linear graph

		A <sub>1</sub>		A <sub>2</sub>	
		B <sub>1</sub>	B <sub>2</sub>	B <sub>1</sub>	B <sub>2</sub>
C <sub>1</sub>	D <sub>1</sub>	0	6	2	-4
	D <sub>2</sub>	-2	4	0	-6
C <sub>2</sub>	D <sub>1</sub>	4	10	6	0
	D <sub>2</sub>	2	8	4	-2

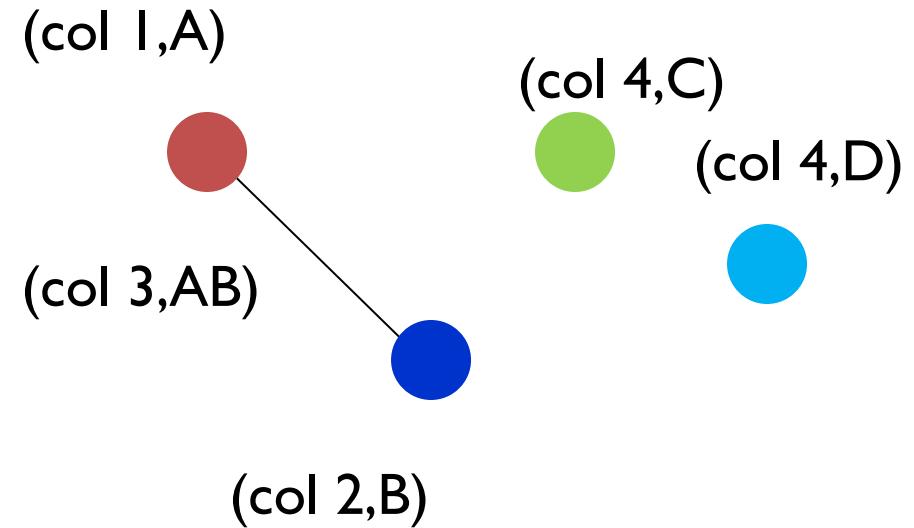
Run	Columns						
	1	2	3	4	5	6	7
1	1	1	1	1	1	1	1
2	1	1	1	2	2	2	2
3	1	2	2	1	1	2	2
4	1	2	2	2	2	1	1
5	2	1	2	1	2	1	2
6	2	1	2	2	1	2	1
7	2	2	1	1	2	2	1
8	2	2	1	2	1	1	2



Only (AB) pair correlation found, no other correlation ....

# Main effect and interactions

	A	B	AxB	C	D
R-1	1	1	1	1	1
R-2	1	1	1	2	2
R-3	1	2	2	1	1
R-4	1	2	2	2	2
R-5	2	1	2	1	2
R-6	2	1	2	2	1
R-7	2	2	1	1	2
R-8	2	2	1	2	1



$(AB)=I$  means  $A_1B_1=I$ ,  $(AB)_2 = A_1B_2$

Expanded basis set and orthogonal vector set ....

$(AB)$  is a dummy column, without it the  $C$  and  $D$  would have different arrangements ...

Still need L8 array (4-7), other two-level arrays L4 (1-3) and L12(8-11)

# An Introduction to Data Analysis, Design of Experiment, and Machine Learning

## *Lecture 9. DOE and Taguchi Experiments*

Muhammad A. Alam  
[alam@purdue.edu](mailto:alam@purdue.edu)



# Three representations of full factorial design

(All levels repeated equal times)

$$F = 2, L = 4;$$
$$R = L^F = 2^4 = 16$$

Aa (1)	Bc (8)	Cd (9)	Db (13)
Bb (2)	Ad (7)	Dc(10)	Ca (14)
Cc (3)	Da (6)	Ab (11)	Bd (15)
Dd (4)	Cb (5)	Ba (12)	Ac (16)

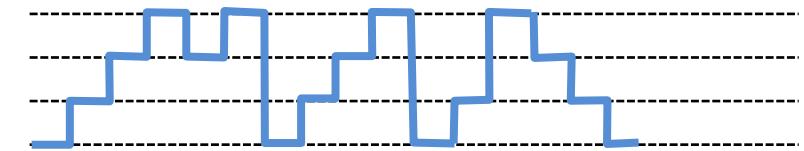
Field-view

Run	Name	Factors	
1	Aa	1	1
2	Bb	2	2
3	Cc	3	3
4	Dd	4	4
5	Cb	3	2
6	Da	4	1
7	Ad	1	4
8	Bc	2	3
9	Cd	3	4
10	Dc	4	3
11	Ab	1	2
12	Ba	2	1
13	Db	4	2
14	Ca	3	1
15	Bd	2	4
16	Ac	1	3

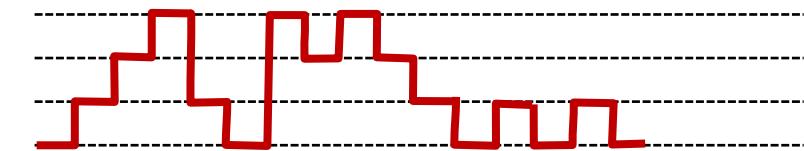
Run-view

4-level random code

$F_1$



$F_2$



Codes are orthogonal  
if their product is zero.

Code-view

# Uncorrelated linear graph (Field and Run views)

		A		A <sub>2</sub>	
		B <sub>1</sub>	B <sub>2</sub>	B <sub>1</sub>	B <sub>2</sub>
C <sub>1</sub>	D <sub>1</sub>	x	x	x	x
	D <sub>2</sub>	x	x	x	x
C <sub>2</sub>	D <sub>1</sub>	x	x	x	x
	D <sub>2</sub>	x	x	x	x

 (col 1,A)  
 (col 3,C)  
 (col 2,B)  
 (col 4,D)

A	B	C	D
0	0	0	0
0	0	0	1
0	0	1	0
0	0	1	1
0	1	0	0
0	1	0	1
0	1	1	0
0	1	1	1
1	0	0	0
1	0	0	1
1	0	1	0
1	0	1	1
1	2	1	1
1	2	1	2
1	2	2	1
1	2	2	2
2	1	1	1
2	1	1	2
2	1	2	1
2	1	2	2
2	2	1	1
2	2	1	2
2	2	2	1
2	2	2	2

$$L=2$$

$$F=4$$

$$L_n(L^F) = L_n(2^4)$$

What is n?

# Taguchi table: How to determine n (Run view)

$L_4(2^3)$

Run	Columns		
	1	2	3
1	1	1	1
2	1	2	2
3	2	1	2
4	2	2	1

$L_8(2^7)$

Run	Columns						
	1	2	3	4	5	6	7
1	1	1	1	1	1	1	1
2	1	1	1	2	2	2	2
3	1	2	2	1	1	2	2
4	1	2	2	2	2	1	1
5	2	1	2	1	2	1	2
6	2	1	2	2	1	2	1
7	2	2	1	1	2	2	1
8	2	2	1	2	1	1	2

$$DOF = 1 + F(L - 1)$$

$L_{12}(2^{11})$

Run	Columns										
	1	2	3	4	5	6	7	8	9	10	11
1	1	1	1	1	1	1	1	1	1	1	1
2	1	1	1	1	1	1	2	2	2	2	2
3	1	1	2	2	2	1	1	1	2	2	2
4	1	2	1	2	1	2	2	1	2	1	2
5	1	2	2	1	2	2	1	2	1	2	1
6	1	2	2	2	1	2	2	1	2	1	1
7	2	1	2	2	1	2	1	1	2	2	1
8	2	1	2	1	2	2	2	2	1	1	2
9	2	1	1	2	2	2	1	2	2	1	1
10	2	2	2	1	1	1	1	2	2	1	2
11	2	2	1	2	1	2	1	1	1	2	2
12	2	2	1	1	2	1	2	1	2	2	1

$L_n(L^F)$

$2^3$

$$DOF$$

$$1 + 3 \times (2 - 1) = 4$$

$n$

Close multiple of 2 = 4

Each col. Two 1's, Two 2's

$2^7$

$$1 + 7 \times (2 - 1) = 8$$

Close multiple of 2 = 8

Four 1's, Four 2's

$2^{11}$

$$1 + 11 \times (2 - 1) = 12$$

Close multiple of 2 = 12

six 1's, six 2's

# Aside: Taguchi Orthogonal Columns

$L_4(2^3)$

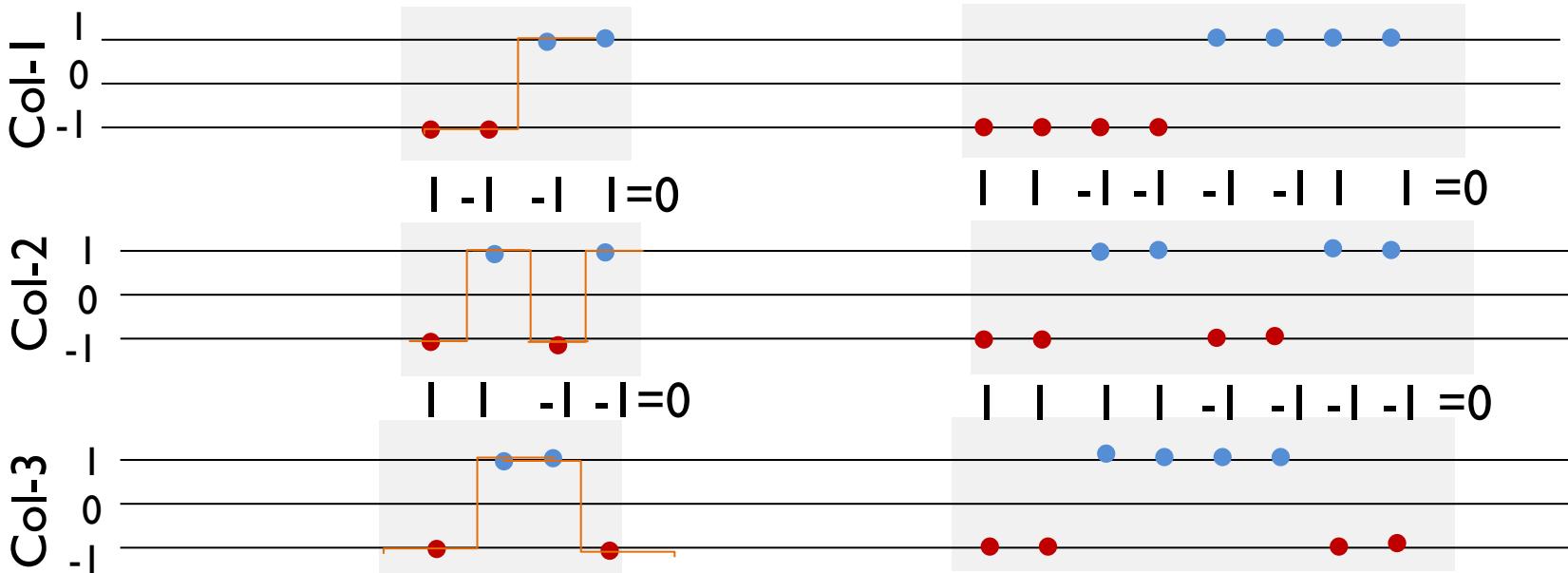
Run	Columns		
	1	2	3
1	1	1	1
2	1	2	2
3	2	1	2
4	2	2	1

$L_8(2^7)$

Run	Columns						
	1	2	3	4	5	6	7
1	1	1	1	1	1	1	1
2	1	1	1	2	2	2	2
3	1	2	2	1	1	2	2
4	1	2	2	2	2	1	1
5	2	1	2	1	2	1	2
6	2	1	2	2	1	2	1
7	2	2	1	1	2	2	1
8	2	2	1	2	1	1	2

$L_{12}(2^{11})$

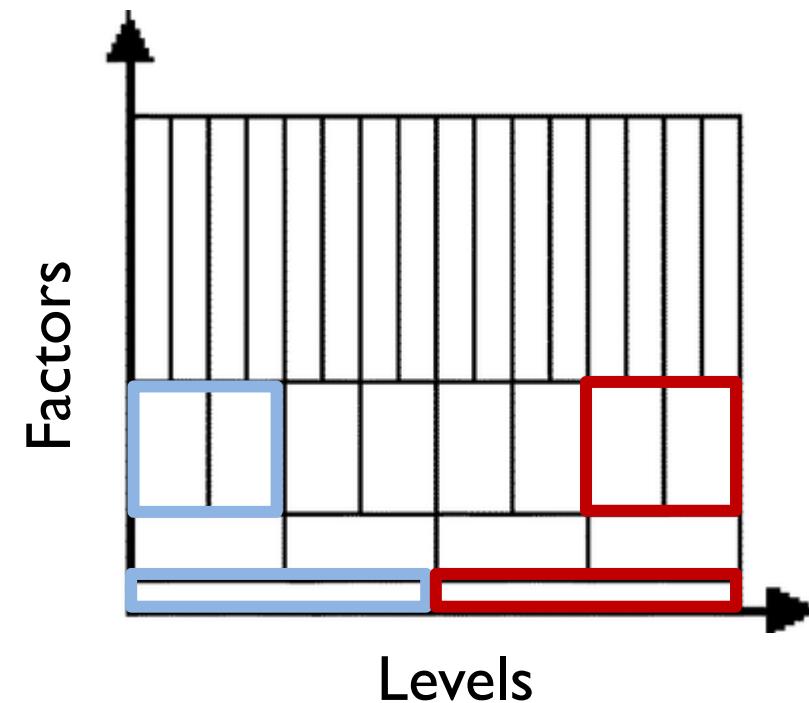
Run	Columns										
	1	2	3	4	5	6	7	8	9	10	11
1	1	1	1	1	1	1	1	1	1	1	1
2	1	1	1	1	1	1	2	2	2	2	2
3	1	1	2	2	2	1	1	1	2	2	2
4	1	2	1	2	1	2	2	1	2	2	1
5	1	2	2	1	2	2	1	2	1	2	1
6	1	2	2	2	1	2	2	1	2	1	1
7	2	1	2	2	2	1	1	2	2	1	2
8	2	1	2	1	2	2	2	1	1	1	2
9	2	1	1	2	2	2	1	2	2	1	1
10	2	2	2	1	1	1	1	2	2	1	2
11	2	2	1	2	1	2	1	1	1	2	2
12	2	2	1	1	2	1	2	1	2	2	1



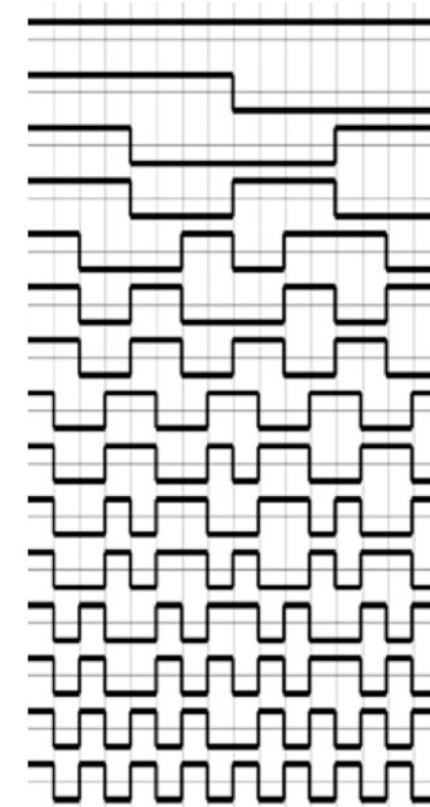
Take any two columns  
(i.e. factors), set 2 to 1  
And 1 to -1, then take  
Inner product and sum.  
The result is always zero.

# Generating Taguchi (orthogonal) Arrays

Run	Columns						
	1	2	3	4	5	6	7
1	1	1	1	1	1	1	1
2	1	1	1	2	2	2	2
3	1	2	2	1	1	2	2
4	1	2	2	2	2	1	1
5	2	1	2	1	2	1	2
6	2	1	2	2	1	2	1
7	2	2	1	1	2	2	1
8	2	2	1	2	1	1	2



CDMA Coding and Wavelet Transform



Reduce interference by Amplitude (AM), frequency (FM), and language (CDMA)

# Full Factorial to Taguchi Table

A	B	C	D
1	1	1	1
1	1	1	2
1	1	2	1
1	1	2	2
1	2	1	1
1	2	1	2
1	2	2	1
1	2	2	2
2	1	1	1
2	1	1	2
2	1	2	1
2	1	2	2
2	2	1	1
2	2	1	2
2	2	2	1
2	2	2	2

Null ... 1  
 Single 4  
 Pair ..... 6  
 Triple ... 6  
 Quad ... 1

A	B	C	D
1	1	1	1
1	1	1	2
1	1	2	1
1	1	2	2
1	2	1	1
1	2	1	2
1	2	2	1
1	2	2	2
2	1	1	1
2	1	1	2
2	1	2	1
2	1	2	2
2	2	1	1
2	2	1	2
2	2	2	1
2	2	2	2

Null ... 1  
 Single ... 1  
 Pair ..... 3  
 Triple ... 3  
 Quad ... 0

Run	Columns						
	1	2	3	4	5	6	7
1	1	1	1	1	1	1	1
2	1	1	1	1	2	2	2
3	1	2	2	1	1	2	2
4	1	2	2	2	2	1	1
5	2	1	2	1	2	1	2
6	2	1	2	2	1	2	1
7	2	2	1	1	2	2	1
8	2	2	1	2	1	1	2

$$S = (-1)^*(-1) + (-1)^*(-1) + (-1)^*(+1) + (-1)^*(+1) + \\ (+1)^*(-1) + (+1)^*(-1) + (+1)^*(+1) + (+1)^*(+1) = 0$$

# Main effect assuming no interactions (Run view)

	A	B	C	D	Y
R-1	1	1	1	1	
R-2	1	1	1	2	
R-3	1	2	2	1	
R-4	1	2	2	2	
R-5	2	1	2	1	
R-6	2	1	2	2	
R-7	2	2	1	1	
R-8	2	2	1	2	

(col 1,A)



(col 3,C)



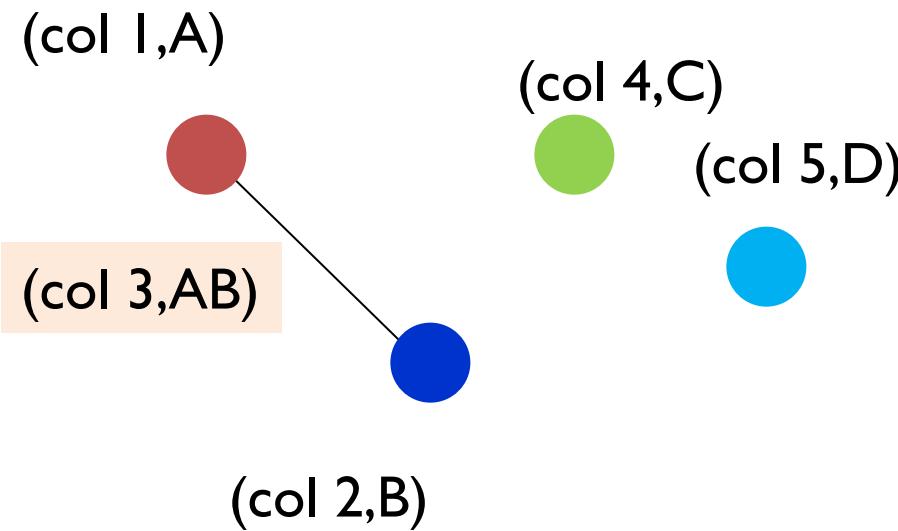
(col 4,D)



(col 2,B)

Let us say that the analysis of Y indicates **AB interaction**, but nothing else.  
We need to redo the experiment

# Main effect with interactions (Run view)



	A	B	AxB	C	D
R-1	I	I	I	I	I
R-2	I	I	I	2	2
R-3	I	2	2	I	I
R-4	I	2	2	2	2
R-5	2	I	2	I	2
R-6	2	I	2	2	I
R-7	2	2	I	I	2
R-8	2	2	I	2	I

Run	Columns						
	1	2	3	4	5	6	7
1	1	1	1	1	1	1	1
2	1	1	1	2	2	2	2
3	1	2	2	1	1	2	2
4	1	2	2	2	2	1	1
5	2	1	2	1	2	1	2
6	2	1	2	2	1	2	1
7	2	2	1	1	2	2	1
8	2	2	1	2	1	1	2

$$L_n(2^5) = 32$$

$$DOF = 1 + F(L - 1) = 1 + 5(2 - 1) = 6 \dots n = 8$$

(AB) is a dummy column, without it the C and D would have different arrangements ...

A	B	C	D
0	0	0	0
0	0	0	1
0	0	1	0
0	0	1	1
0	1	0	0
0	1	0	1
0	1	1	0
0	1	1	1
1	0	0	0
1	0	0	1
1	0	1	0
1	0	1	1
1	1	0	0
1	1	0	1
1	1	1	0
1	1	1	1

Run	Columns						
	1	2	3	4	5	6	7
1	1	1	1	1	1	1	1
2	1	1	1	2	2	2	2
3	1	2	2	1	1	2	2
4	1	2	2	2	2	1	1
5	2	1	2	1	2	1	2
6	2	1	2	2	1	2	1
7	2	2	1	1	2	2	1
8	2	2	1	2	1	1	2

A B C D

0 0 0 0  
0 0 0 1  
0 0 1 0  
0 0 1 1  
0 1 0 0  
0 1 0 1  
0 1 1 0  
0 1 1 1

Null ... I  
Single 4  
Pair ..... 6  
Triple ... 6  
Quad ... I

0 0 0 0  
0 0 0 1  
0 0 1 0  
0 0 1 1  
0 1 0 0  
0 1 0 1  
0 1 1 0  
0 1 1 1

+ 0 0 0  
+ 0 0 1  
+ 0 1 0  
+ 0 1 1  
+ 1 0 0  
+ 1 0 1  
+ 1 1 0  
+ 1 1 1

Run	Columns						
	1	2	3	4	5	6	7
1	1	1	1	1	1	1	1
2	1	1	1	2	2	2	2
3	1	2	1	1	1	2	2
4	1	2	2	2	2	1	1
5	2	1	1	2	1	1	2
6	2	1	2	1	2	1	1
7	2	2	2	1	2	2	1
8	2	2	2	1	1	1	2

A B C D

0 0 0 0  
0 0 0 1  
0 0 1 0  
0 0 1 1  
0 1 0 0  
0 1 0 1  
0 1 1 0  
0 1 1 1

0 0 0 0  
0 0 0 1  
0 0 1 0  
0 0 1 1  
0 1 0 0  
0 1 0 1  
0 1 1 0  
0 1 1 1

+ 0 0 0  
+ 0 0 1  
+ 0 1 0  
+ 0 1 1  
+ 1 0 0  
+ 1 0 1  
+ 1 1 0  
+ 1 1 1

Null ... I  
Single 1  
Pair ..... 3  
Triple ... 3  
Quad ... 0

The effect of B is  
now understood

# Web Design: 4 Factor, 5 level

4 factors (font, color, background, foreground) and 5 levels of each

Factors

x1	x2	x3	x4	x5	x6
1	1	1	1	1	1
1	2	2	2	2	2
1	3	3	3	3	3
1	4	4	4	4	4
1	5	5	5	5	5
2	1	2	3	4	5
2	2	3	4	5	1
2	3	4	5	1	2
2	4	5	1	2	3
2	5	1	2	3	4
3	1	3	5	2	4
3	2	4	1	3	5
3	3	5	2	4	1
3	4	1	3	5	2
3	5	2	4	1	3
4	1	4	2	5	3
4	2	5	3	1	4
4	3	1	4	2	5
4	4	2	5	3	1
4	5	3	1	4	2
5	1	5	4	3	2
5	2	1	5	4	3
5	3	2	1	5	4
5	4	3	2	1	5
5	5	4	3	2	1

Levels

$$L_n(L^F) = L_n(5^4) \\ = L_n(125), n = ?$$

$$DOF = 1 + F(L - 1) = \\ 4 \times 4 = 17.$$

Balanced design:

multiple of 5;  $n = 20, 25$

Partial factorial Design

Randomization

fjords	jawbox	phlegm	qiviut	zincky
zincky	fjords	jawbox	phlegm	qiviut
qiviut	zincky	fjords	jawbox	phlegm
phlegm	qiviut	zincky	fjords	jawbox
jawbox	phlegm	qiviut	zincky	fjords

[https://www.mne.psu.edu/cimbala/me345/Lectures/Taguchi\\_orthogonal\\_arrays.pdf](https://www.mne.psu.edu/cimbala/me345/Lectures/Taguchi_orthogonal_arrays.pdf)

<https://www.york.ac.uk/depts/maths/tables/orthogonal.htm>

# Conclusions

- I. Design of experiment is a powerful technique universally used in industry and in large scale field trials.
2. Taguchi/Fisher methods replace the older one-factor-at-a-time experiments with experiments based on orthogonal arrays; In this approach, only the effect of main factors remain; others are cancelled.
3. Understanding and analyzing correlation is important in design of experiments. Unless the correlation is well understood and incorporated through dummy variables, the analysis may lead to faulty conclusions.

# Review Questions

1. What role did Fisher play in developing the design of experiment?
2. If you have 3 variables at two levels, what Taguchi array would you choose?
3. How does one find correlation among variables in Full factorial method?
4. What is the role of linear graphs in Taguchi method?
5. In what ways Fisher philosophy of change the ways experiments are done?  
Is there a down side of such analysis?
6. What is dummy variable? What does dummy variable do in DOE?
7. Can you have 3<sup>rd</sup> or higher order correlation, if you do not have second order correlation?

# An Introduction to Data Analysis, Design of Experiment, and Machine Learning

## Lecture 9. *DOE Analysis by ANOVA*

Muhammad A. Alam  
[alam@purdue.edu](mailto:alam@purdue.edu)



# Course Outline

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

Lecture 1: Introduction

Lecture 2: Collecting and plotting  $x_1, x_2, \dots x_n$

Lecture 3: Physical and empirical  $f, F, df/dx, \dots$

Lecture 4: Model selection among  $f_1, f_2, \dots$

Lecture 5: Scaling theory with known  $f$ ,  $f(\bar{x}) = f(\bar{X})$

Lecture 6: Scaling theory with unknown  $f$ ,  $\bar{x} \rightarrow X$

**Lecture 7: Design of experiments to determine  $\bar{y}_{\max} = f(\bar{x})$**

Lecture 8: Machine learning ... Statistical approach to learn f

Lecture 9: Physics-based machine learning  $f = f_{\text{physics}} + \Delta f$

Lecture 10: Principle component analysis for classifying  $\{y\}$ .

Lecture 11: Conclusions

# Copyright 2018

This material is copyrighted by M. Alam under the following Creative Commons license:



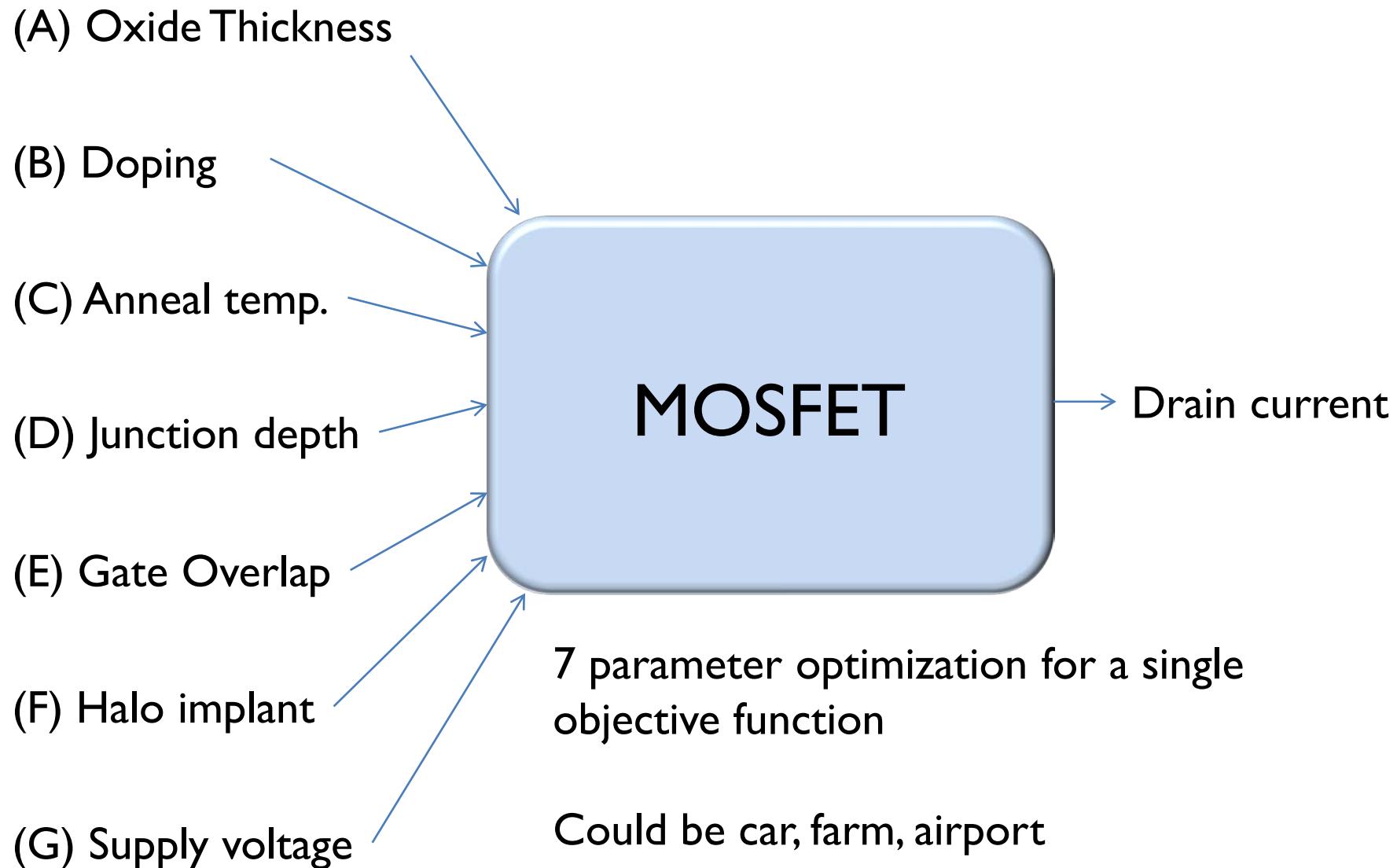
Conditions for using these materials is described at

<http://creativecommons.org/licenses/by-nc-sa/2.5/>

# Outline

1. Introduction to Analysis of Variance (Anova)
2. Single factor Analysis of Variance
3. Two factor Anova
4. Generalized Anova
5. Conclusions

# Another way to reduce the number of experiments

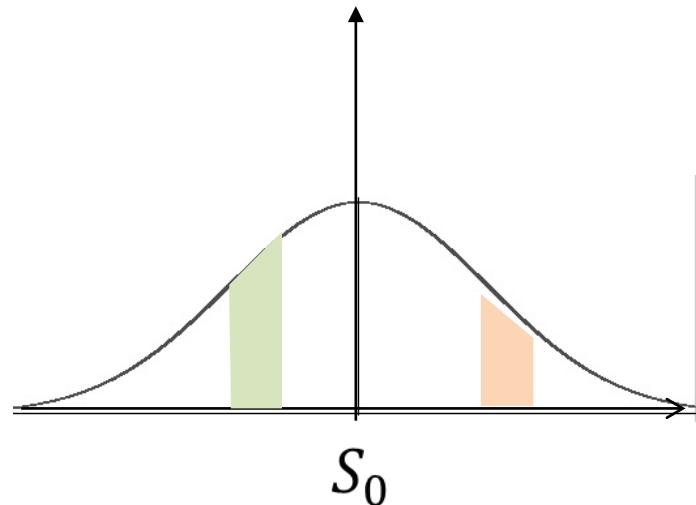


# Single factor ANOVA: Treatment

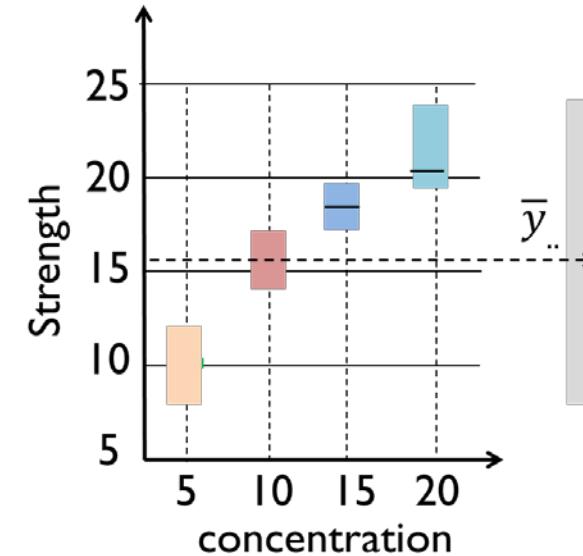
replicates

	1	2	3	4	5	6
5	7	8	15	11	9	10
10	12	17	13	18	19	15
15	14	18	19	17	16	18
20	19	25	22	23	18	20

Treatments (levels)

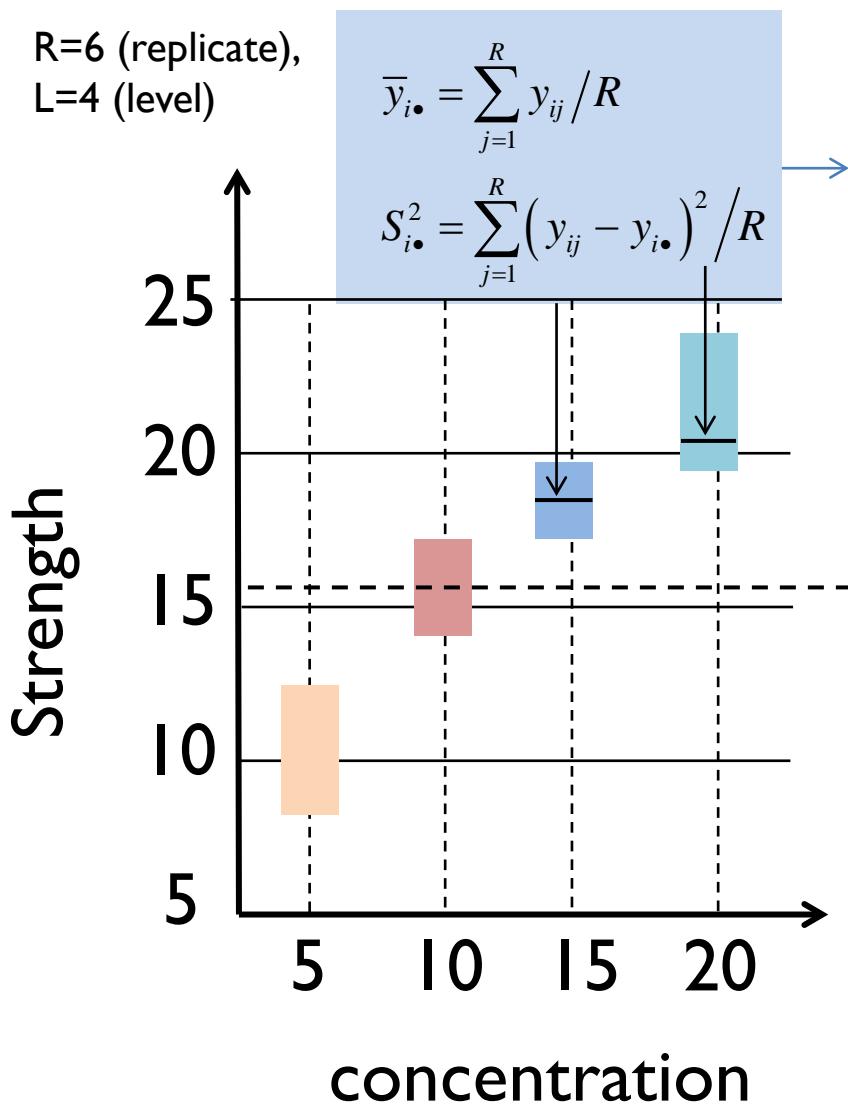


In essence, no effect



# Single factor Anova: Treatment Analysis

R=6 (replicate),  
L=4 (level)



$$SS_T^2 = \sum_{j=1}^L \sum_{i=1}^R (y_{ij} - \bar{y}_{..})^2$$

$$RL \times s_E^2 = SS_T^2 - SS_{Treatment}^2$$

$$\bar{y}_{..} = \sum_{i=1}^L \sum_{j=1}^R y_{ij} / RL$$

$$SS_{Treatment} = L\sigma^2 = R \sum_{i=1}^L (\bar{y}_{i\bullet} - \bar{y}_{..})^2$$

Recall:  $\chi^2$ , KS, etc.

$$F = \frac{\sigma^2 \times L/(L-1)}{s_E^2 \times RL/L(R-1)} : F_{crit}(L-1, L(R-1))$$

# Single factor ANOVA (continued)

	1	2	3	4	5	6
5	7	8	15	11	9	10
10	12	17	13	18	19	15
15	14	18	19	17	16	18
20	19	25	22	23	18	20

- Treatment number,  $a = 4$ ;  $dof_a = 3$ ; Sample number:  $n = 6$
- Global sample number:  $a \times n = 24$ ,  $dof_n = 23$ , global AVG = 15.96
- Total sum of square,  $SS_T = \sum_{24} (data - AVG)^2 = 512.96$
- Treatment sum:  $SS_{treatment} = n \times \sum_4 (treat.\ avg - AVG)^2 = 382.$
- $SS_{Error} = SS_T - SS_{treatment} = 130.62$
- $ME_{treatment} = SS_E/dof_a$ ,  $ME_E = SS_{error}/(dof_n - dof_a)$
- Finally,  $F = (ME_{treatment})/(ME_{error}) = 19.6$
- Compare:  $f(0.01, dof_a, dof_n)$ , or  $P(F_{3,20} > 19.6) = 3.59 \times 10^{-6}$

# Single factor ANOVA: Wood Treatment

	replicates					
	1	2	3	4	5	6
5	7	8	15	11	9	10
10	12	17	13	18	19	15
15	14	18	19	17	16	18
20	19	25	22	23	18	20

treatments

$$\sum(data - AVG)^2 = 512$$

$$6 \times 63.8 = 382.8$$

$S_{avg} (s-avg-AVG)^2$	
10.00	35.50174
15.67	0.085069
17.00	1.085069
21.17	27.12674
15.96	63.79861

Variation	SS	df	MS	F	P-value	F crit
Between Groups	382.7917	3	127.60	19.605	3.59E-06	4.94
Within Groups	130.1667	20	6.51			
Total	512.9583	23				

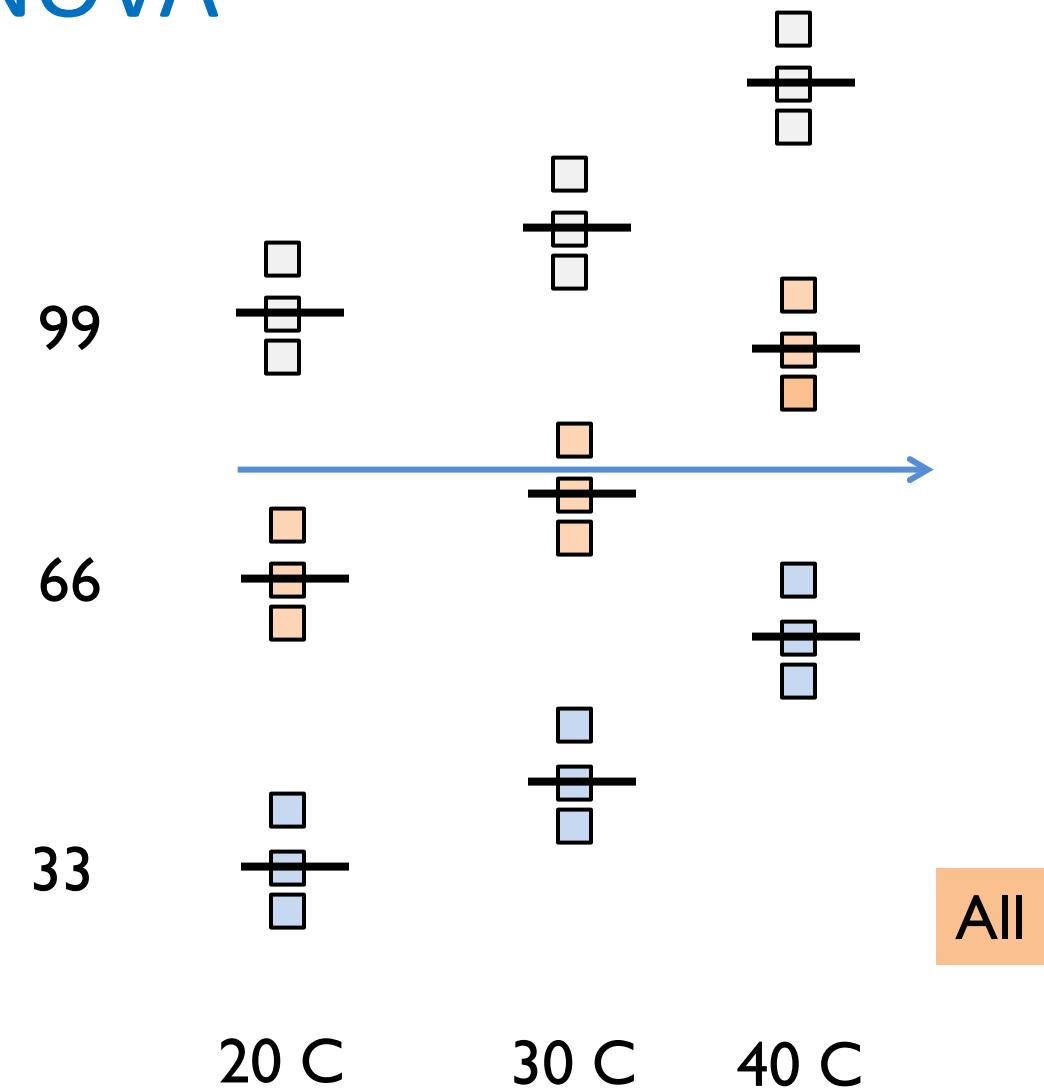
# Outline

- I. Introduction to Analysis of Variance (Anova)
2. Single factor Analysis of Variance
3. **Generalized Anova**
4. Conclusions

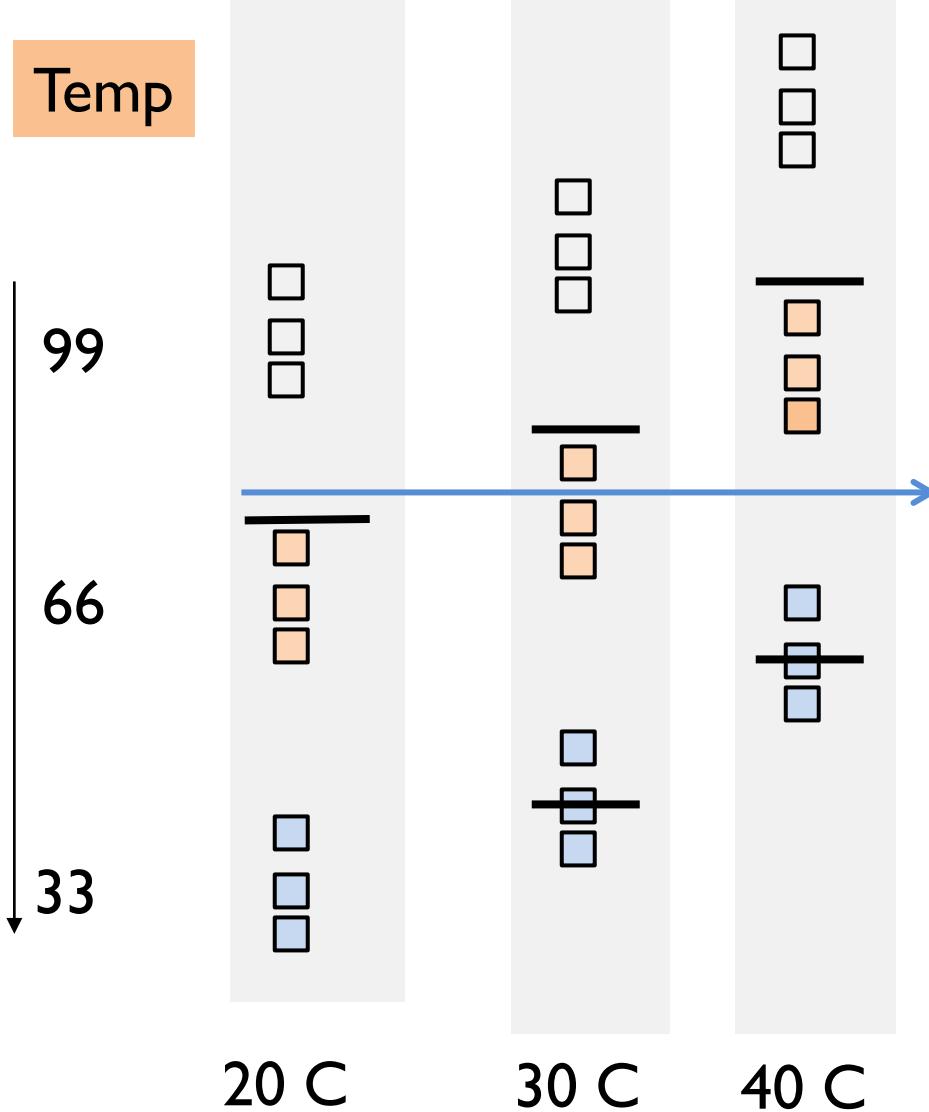
# Two factor ANOVA

Full factorial:  
2 factor, 3 level,  
3 replicate experiment

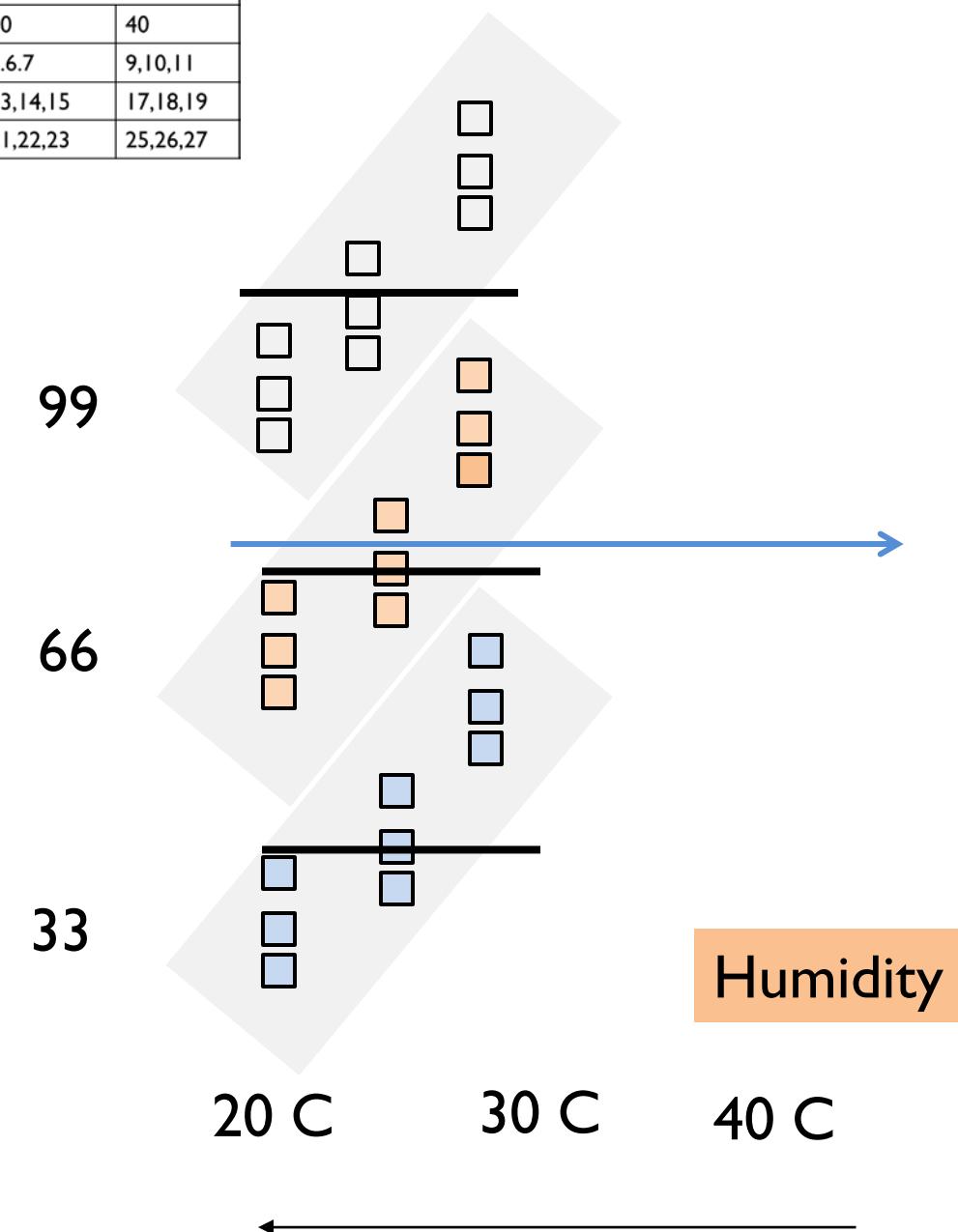
	Temperature (C)		
Humidity (%)	20	30	40
33	1,2,3	5,6,7	9,10,11
66	9,10,11	13,14,15	17,18,19
99	17,18,19	21,22,23	25,26,27



# Two factor ANOVA



Humidity (%)	Temperature (C)		
	20	30	40
33	1,2,3	5,6,7	9,10,11
66	9,10,11	13,14,15	17,18,19
99	17,18,19	21,22,23	25,26,27



# Two factor ANOVA

Full factorial:  
2 factor, 3 level,  
3 replicate experiment

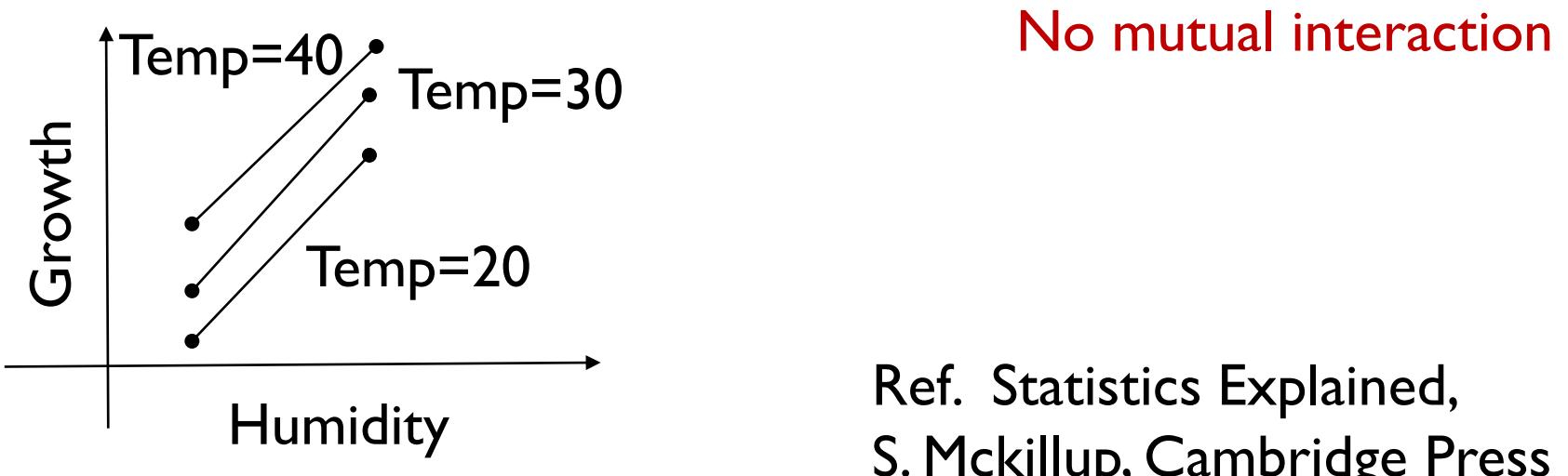
	Temperature (C)		
Humidity (%)	20	30	40
33	1,2,3	5,6,7	9,10,11
66	9,10,11	13,14,15	17,18,19
99	17,18,19	21,22,23	25,26,27

Excel/Minitab Analysis

	Sum of Squares	dof	Mean-square	F Ratio	Significance
Temp	312.66	3-1=2	312.66/2=156.33	156.33/1.0=156.33	0.000 (significant)
Humidity	1200.66	3-1=2	1200.66/2=600.33	600.66/1=600.33	0.000 (Significant)
Temp*Humidity	1.33	2x2=4	1.33/4=0.33	0.33/1.0=0.33	0.853 (insignificant)
Error	18.00	27-2-2-4=19	18.00/18=1		

# Two factor ANOVA (Excel/Minitab Analysis)

	Sum of Squares	dof	Mean-square	F Ratio	p-value
Temp	312.66	$3-1=2$	$312.66/2=156.33$	$156.33/1.0=156.33$	0.000 (significant)
Humidity	1200.66	$3-1=2$	$1200.66/2=600.33$	$600.66/1=600.33$	0.000 (Significant)
Temp*Humidity	1.33	$2\times 2=4$	$1.33/4=0.33$	$0.33/1.0=0.33$	0.853 (insignificant)
Error	18.00	$27-2-2-4=19$	$18.00/18=1$		



# An Introduction to Data Analysis, Design of Experiment, and Machine Learning

## *Lecture 10. Big Data Classification by Principal Component Analysis*

Muhammad A. Alam  
[alam@purdue.edu](mailto:alam@purdue.edu)



# copyright 2018

This material is copyrighted by M. Alam under the following Creative Commons license:



**Attribution-NonCommercial-ShareAlike 2.5 Generic (CC BY-NC-SA 2.5)**

Conditions for using these materials is described at

<http://creativecommons.org/licenses/by-nc-sa/2.5/>

# Course Outline

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

Lecture 1: Introduction

Lecture 2: Collecting and plotting  $x_1, x_2, \dots x_n$

Lecture 3: Physical and empirical  $f, F, df/dx, \dots$

Lecture 4: Model selection between  $f_1, f_2, \dots$

Lecture 5: Model Selection: Cross-validation and Bootstrapping method

Lecture 6: Scaling theory with known  $f$ ,  $f(\bar{x}) = f(\bar{X})$

Lecture 7: Scaling theory with unknown  $f$ ,  $\bar{x} \rightarrow X$

Lecture 8: Design of experiments to determine  $\bar{y}_{\max} = f(\bar{x})$

Lecture 9: DOE and ANOVA

Lecture 10: Principle component analysis for classifying  $\{y\}$ .

Lecture 11: Machine learning ... Statistical approach to learn  $f$

Lecture 12: Interpretable ML: Physics-based machine learning  $f = f_{\text{physics}} + \Delta f$

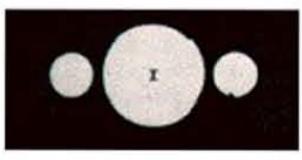
Lecture 13: Interpretable ML: System Equation Modeling

Lecture 14: 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.

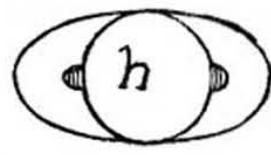
# Small vs. big data



Galileo first sketch  
1610



Better telescope  
1616

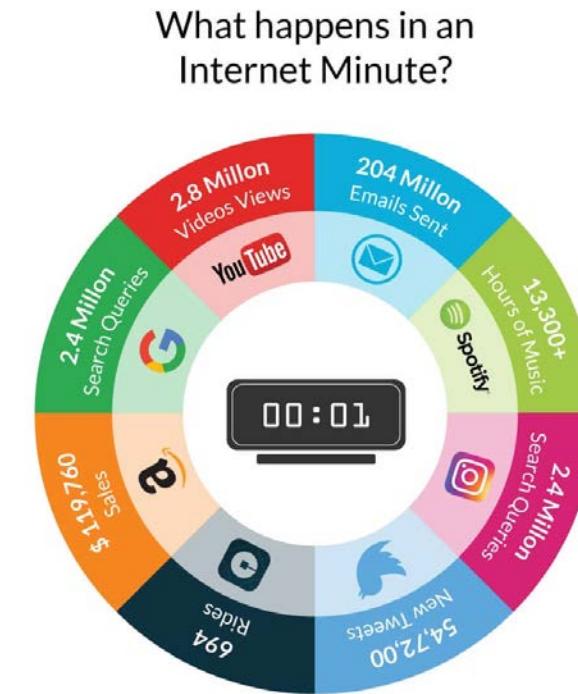


Published etch  
1623



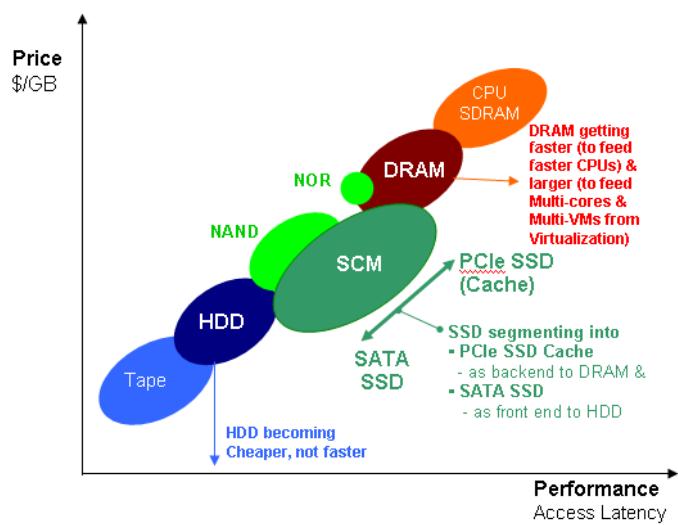
# Where do data come from?

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



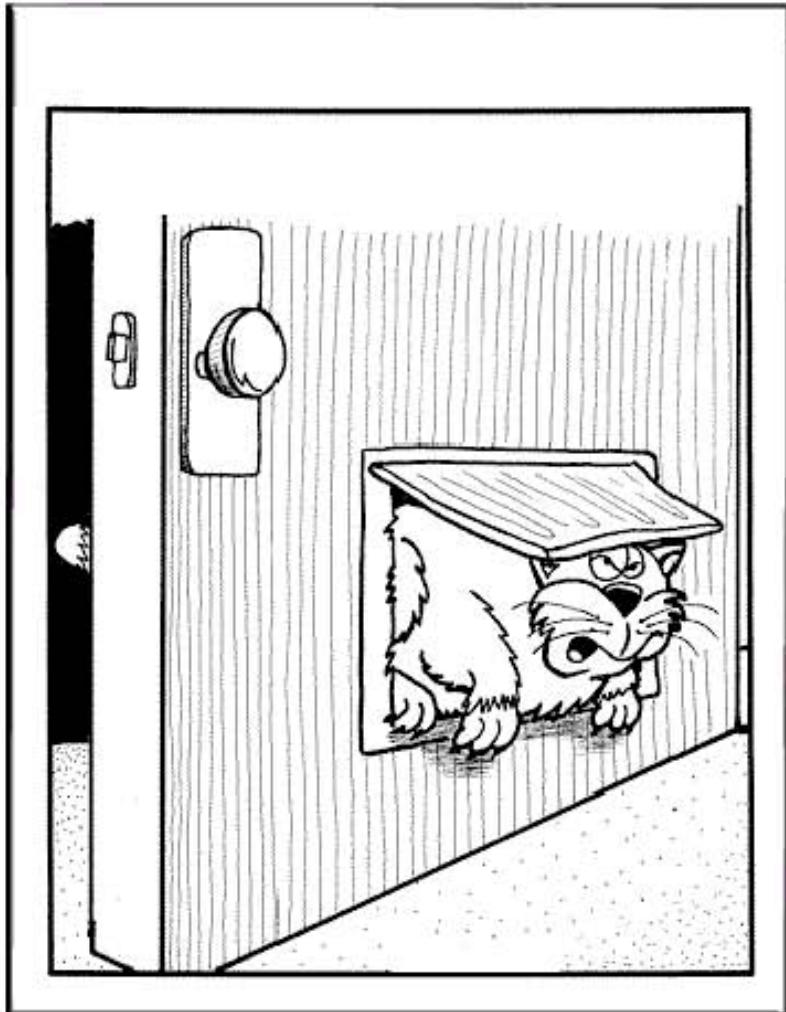
# ... driven by memory technology

- 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<sup>2</sup>, eight times the size of Manhattan
- Amazon Web services, Google Cloud, IBM Cloud, Microsoft Azure.



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

# “Big data” techniques apply to “little data” too



Isaac Newton, known as a physicist, mathematician and astronomer, may have also been the “cat door inventor”! According to an anecdote, Newton foolishly made a large hole for the mother cat and six small holes for her six kittens, not understanding

that the kittens could follow their mother through the large hole!



# Our goal for the next few lectures ...

## Lectures 8-14

How to get a better  $f$

## Lectures 6-7

$$\bar{x} = x_1, x_2, \dots x_n \longrightarrow$$

$$\bar{y} = f(\bar{x})$$

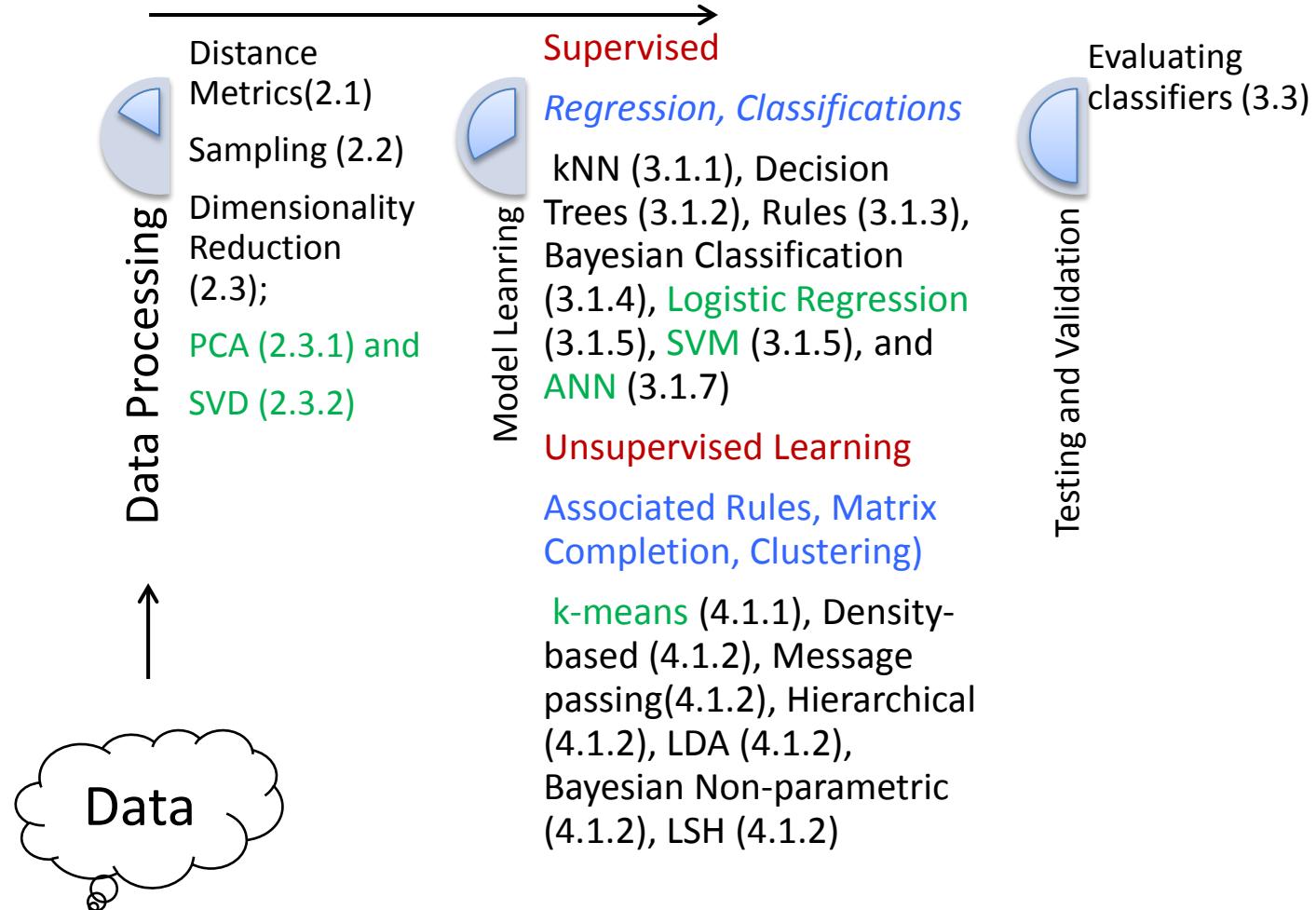
## Lectures 1-2

$$\bar{y} = y_1, y_2, \dots y_m$$

## Lectures 3-5

How to fit multiple hypothetical function  $f$  to the same  $y$

# Analysis of big data



# Outline

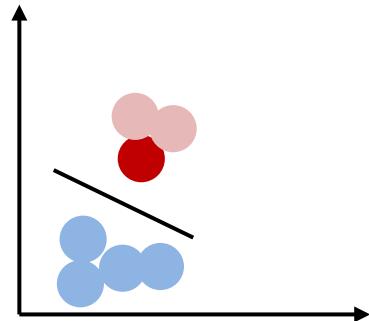
1. Introduction
2. Why do we need reduction in data dimension
2. Theory of Principle Component Analysis
3. Applications of Principle Component Analysis
4. Conclusions

# Classification problem in big data

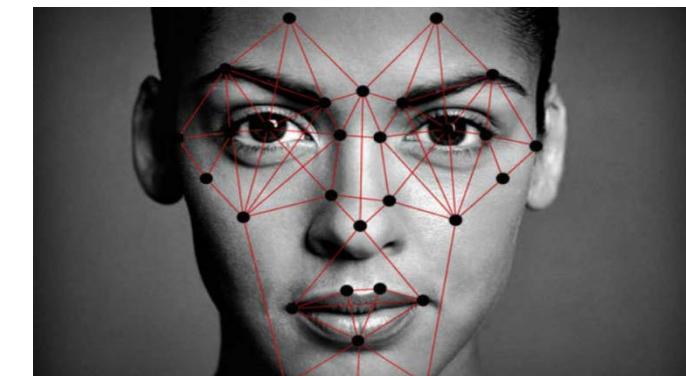
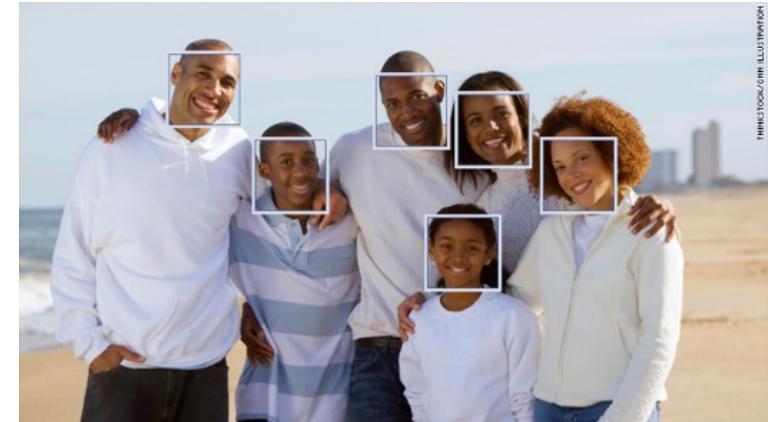
Advertisement  
Recommendation



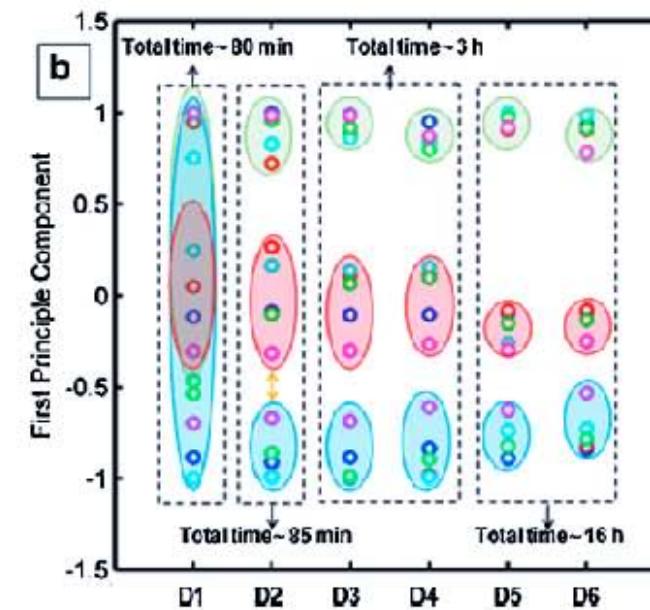
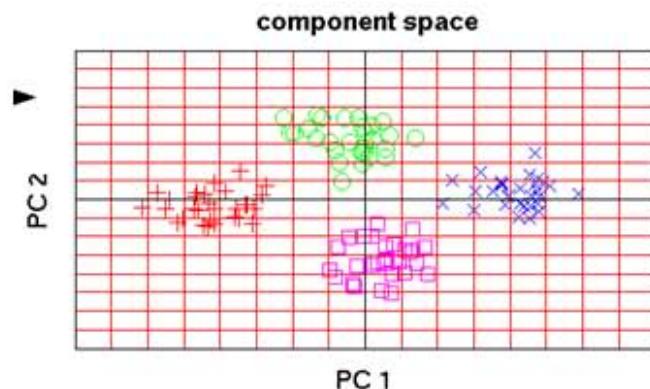
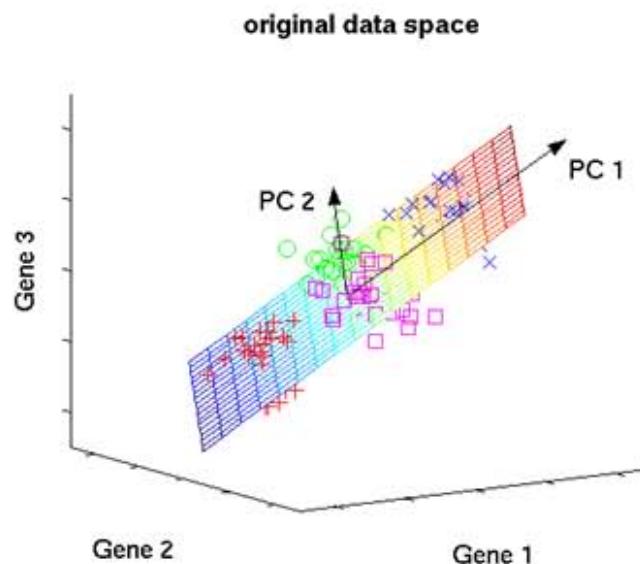
Everything is a Recommendation



Facial Recognition  
Voice Recognition  
Spam Filtering



# PCA helps classification



## PCA Also help in data compression

3D information projected onto a 2D plane



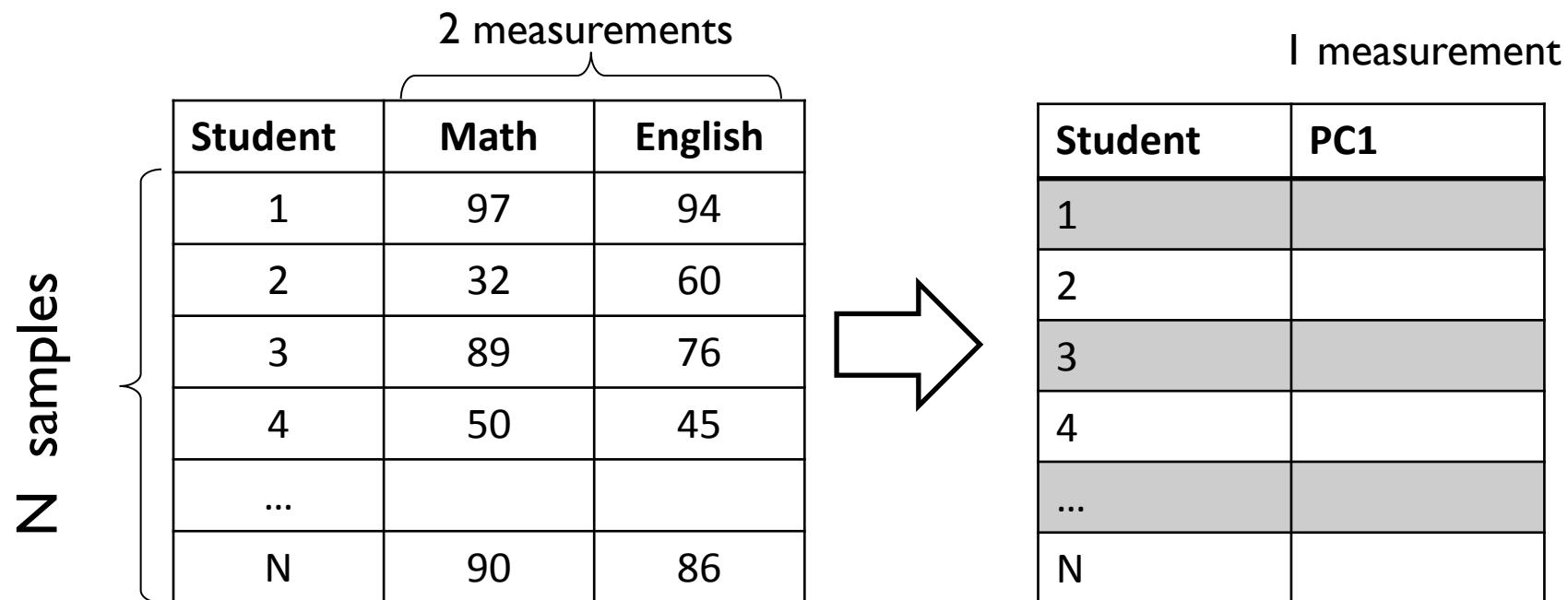
Perspective projection invented by Fillipo Brunelleschi & Masaccio

# Outline

- I. Why do we need reduction in data dimension
2. Theory of Principle Component Analysis
3. Applications of Principle Component Analysis
4. Conclusions

# Principle Component Analysis (PCA)

- Example: Are some students falling behind?  
Difficult to decide in a multidimensional data



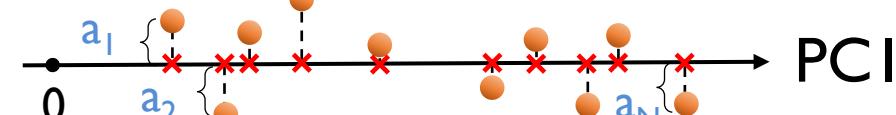
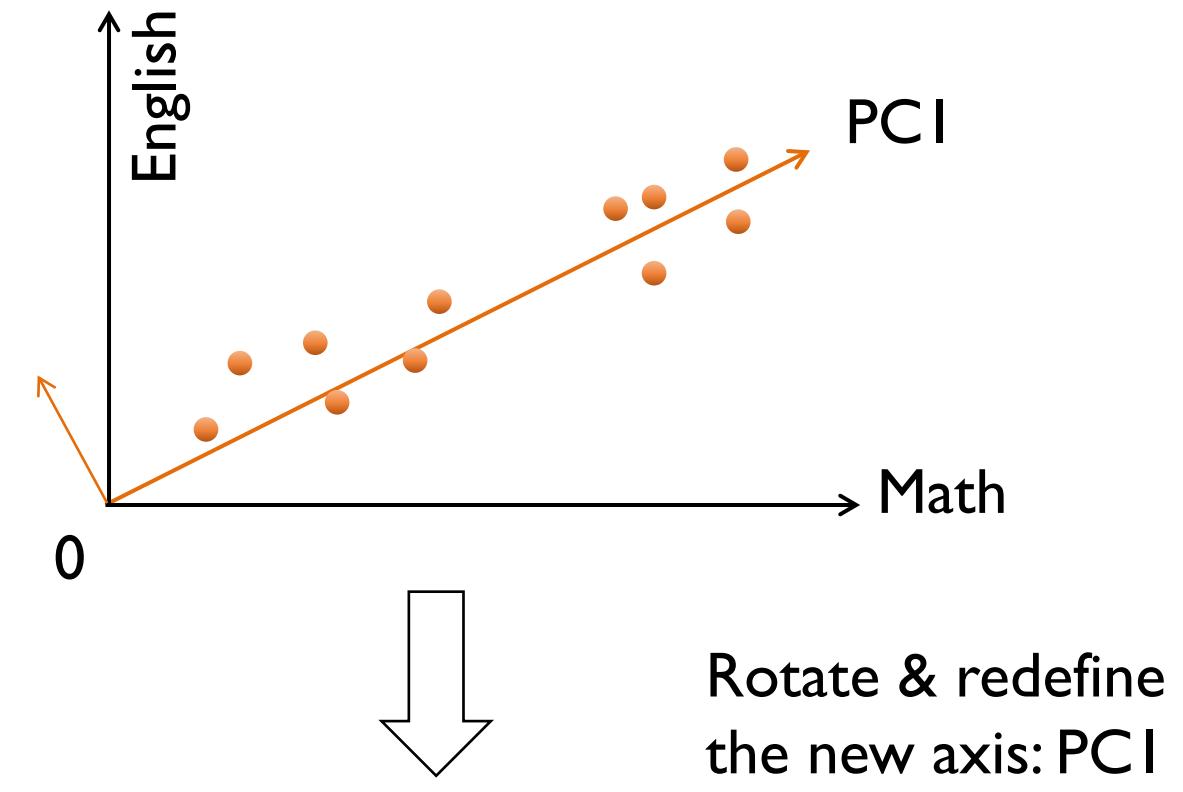
# Basic Concept of PCA

2 measurements

Student	Math	English
1	97	94
2	32	60
3	89	76
4	50	45
...		
N	90	86

N samples

To reduce 2-D data to 1-D data:  
find a direction onto which to  
project the data so as to  
*minimize the projection error*



$$\text{Projection error} = \sum_N a_N^2$$

# PCA through Singular Value Decomposition

2 measurements

Student	Math	English
1	97	94
2	32	60
3	89	76
4	50	45
...		
N	90	86

$m$  measurements

$X:$   
 $n \times m$   
matrix

$n$  samples

$$X = U\Sigma V^T$$

Weight of each  
principal components

$$X = \begin{bmatrix} u_{11} & u_{21} & \dots & u_{n1} \\ u_{12} & u_{22} & \dots & u_{n2} \\ \dots & \dots & \dots & \dots \\ u_{1n} & u_{2n} & \dots & u_{nn} \end{bmatrix} \begin{bmatrix} \sigma_1 & 0 & \dots & 0 \\ 0 & \sigma_2 & & \\ \dots & \dots & \dots & \\ 0 & 0 & \dots & \sigma_m \end{bmatrix} \begin{bmatrix} v_{11} & v_{21} & \dots & v_{m1} \\ v_{12} & v_{22} & & v_{m2} \\ \dots & \dots & \dots & \\ v_{1m} & v_{2m} & \dots & v_{mm} \end{bmatrix}$$

$U:$   $n \times n$  matrix

$\Sigma:$   
 $n \times m$  diagonal  
matrix

$V:$   $m \times m$  matrix

Direction of the new  
axis (PC1, PC2 ...)

# Reduce dimension by Singular Value Decomposition

$$A_{n \times d} = \hat{U}_{n \times r} \Sigma_{r \times r} \hat{V}^T_{r \times d}$$

$\hat{U}$        $\Sigma$        $\hat{V}^T$

$n \times d$      $n \times d$      $d \times d$

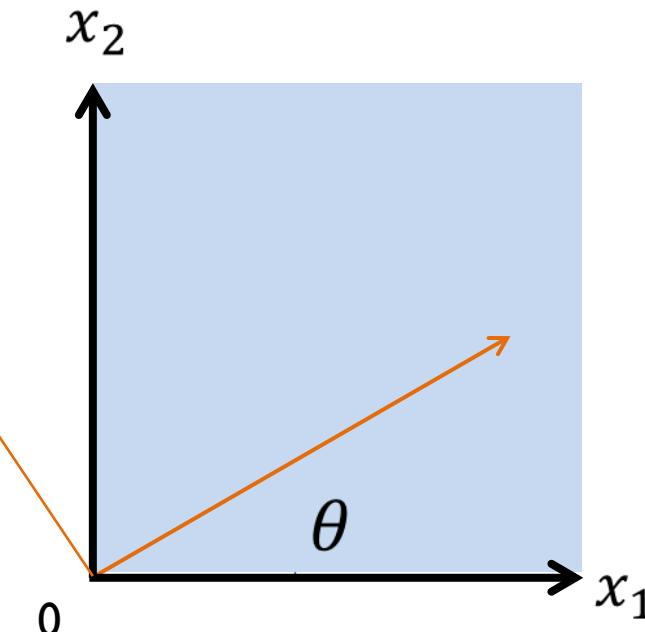
## Example 1: Rotation matrix

$$R = \begin{pmatrix} \cos(\theta) & -\sin(\theta) \\ \sin(\theta) & \cos(\theta) \end{pmatrix}$$

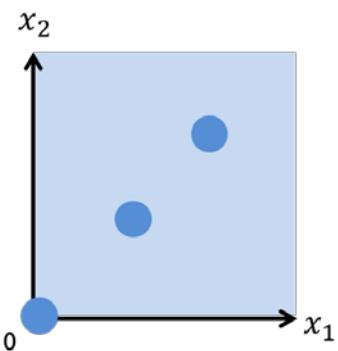
$$R^{90} = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}$$

$$R^{180} = \begin{pmatrix} -1 & 0 \\ 1 & -1 \end{pmatrix}$$

$$R^{270} = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}$$



# SVD rotates the axes optimally



(3x2 matrix)

$$X = \begin{pmatrix} 1 & 1 \\ 2 & 2 \\ 0 & 0 \end{pmatrix}$$

Three points (1,1), (2,2) and (0,0)  
SVD (<{1,1}, {2,2}, {0,0}>)

$$X = U \Sigma V^T$$

(2x2 matrix)

$$V^T = \begin{pmatrix} 1/\sqrt{2} & 1/\sqrt{2} \\ -1/\sqrt{2} & 1/\sqrt{2} \end{pmatrix}$$

Rotate by 45 degrees  
The PC is sufficient

$$U = \begin{pmatrix} 1/\sqrt{5} & 0 & -2/\sqrt{5} \\ 2/\sqrt{5} & 0 & 1/\sqrt{5} \\ 0 & 1 & 0 \end{pmatrix}$$

(3x3 matrix)

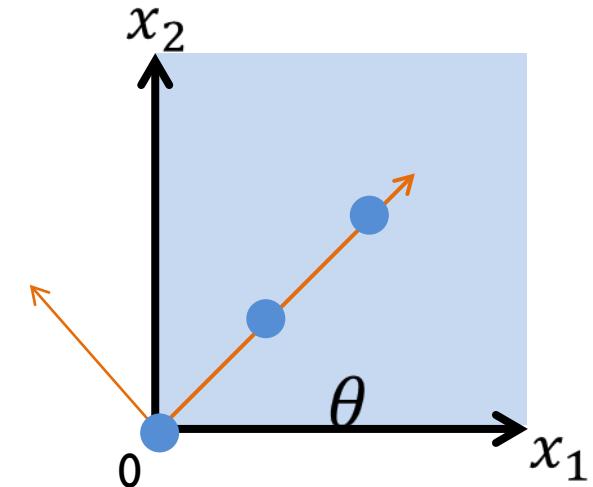
$$\Sigma = \begin{pmatrix} \sqrt{10} & 0 \\ 0 & 0 \\ 0 & 0 \end{pmatrix}$$

## SVD components allows reconstruction

$$X = \begin{pmatrix} 1 & 1 \\ 2 & 2 \\ 0 & 0 \end{pmatrix} = u_1 \sigma_1 v_1^T + u_2 \sigma_2 v_2^T$$

$$u_1 = \begin{pmatrix} 1/\sqrt{5} \\ 2/\sqrt{5} \\ 0 \end{pmatrix} \quad \sigma_1 = \sqrt{10} \quad v_1^T = (1/\sqrt{2} \quad 1/\sqrt{2})$$

$$X = \begin{pmatrix} 1 & 1 \\ 2 & 2 \\ 0 & 0 \end{pmatrix} \quad X' = u_1 \sigma_1 v_1^T = \begin{pmatrix} 1 & 1 \\ 2 & 2 \\ 0 & 0 \end{pmatrix}$$



$X' = X$  because the projection is exact

# Projection along PCs

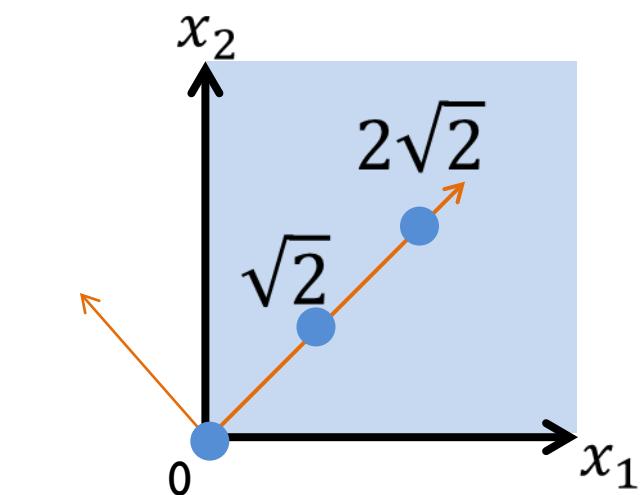
$$XV = U\Sigma V^T V = U\Sigma$$

(3x2 matrix)

$$X = \begin{pmatrix} 1 & 1 \\ 2 & 2 \\ 0 & 0 \end{pmatrix}$$

(2x2 matrix)

$$V = \begin{pmatrix} 1/\sqrt{2} & -1/\sqrt{2} \\ 1/\sqrt{2} & 1/\sqrt{2} \end{pmatrix}$$



$$X \cdot V = \begin{pmatrix} \sqrt{2} & 0 \\ 2\sqrt{2} & 0 \\ 0 & 0 \end{pmatrix}$$

$$U = \begin{pmatrix} 1/\sqrt{5} & 0 & -2/\sqrt{5} \\ 2/\sqrt{5} & 0 & 1/\sqrt{5} \\ 0 & 1 & 0 \end{pmatrix}$$

(3x3 matrix)

$$\Sigma = \begin{pmatrix} \sqrt{10} & 0 \\ 0 & 0 \\ 0 & 0 \end{pmatrix}$$

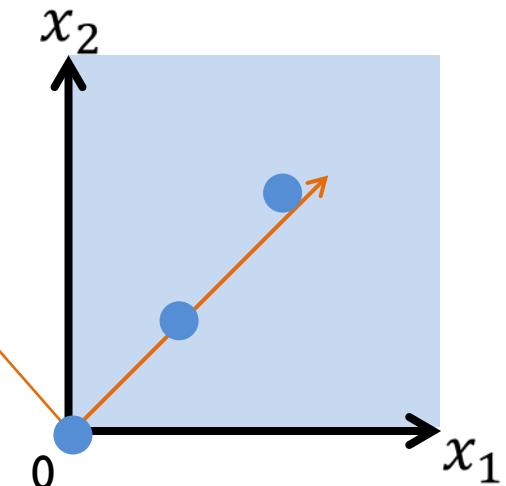
(3x2 matrix)

$$U \cdot \Sigma = \begin{pmatrix} \sqrt{2} & 0 \\ 2\sqrt{2} & 0 \\ 0 & 0 \end{pmatrix}$$

(3x2 matrix)

Projection along PC1

## Example 2: More general result



$$X = \begin{pmatrix} 0 & 0 \\ 1 & 1 \\ 2 & 2.1 \end{pmatrix}$$

(3x2 matrix)

Three points (0,0), (1,1), (2,2.1)  
SVD ({0,0}, {1,1}, {2,2.1})

$$V^T = \begin{pmatrix} 0.693 & 0.721 \\ 0.721 & -0.693 \end{pmatrix}$$

(2x2 matrix)

$$X = U \Sigma V^T$$

$$U = \begin{pmatrix} \sim 0 & \sim 0 & 1 \\ 0.438 & 0.899 & \sim 0 \\ 0.899 & -0.438 & 0 \end{pmatrix}$$

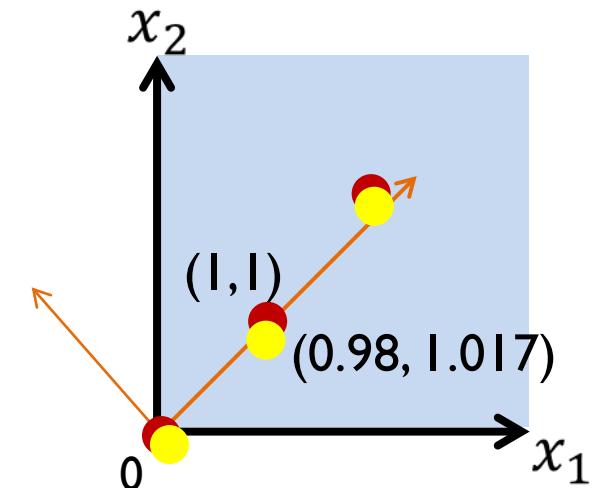
(3x3 matrix)

$$\Sigma = \begin{pmatrix} 3.226 & 0 \\ 0 & 0.031 \\ 0 & 0 \end{pmatrix}$$

(3x2 matrix)

## SVD approximates the exact result

$$X = \begin{pmatrix} 0 & 0 \\ 1 & 1 \\ 2 & 2.1 \end{pmatrix} = u_1 \sigma_1 v_1^T + u_2 \sigma_2 v_2^T$$



$$u_1 = \begin{pmatrix} 0 \\ -0.438 \\ -0.721 \end{pmatrix} \quad \sigma_1 = 3.226 \quad v_1^T = (-0.693 \quad -0.721)$$

$$X = \begin{pmatrix} 1 & 1 \\ 2 & 2 \\ 0 & 0 \end{pmatrix} \quad X' = u_1 \sigma_1 v_1^T = \begin{pmatrix} 0 & 0 \\ 0.98 & 1.017 \\ 2.01 & 2.08 \end{pmatrix}$$

$X' \sim X$  because the projection is approximate

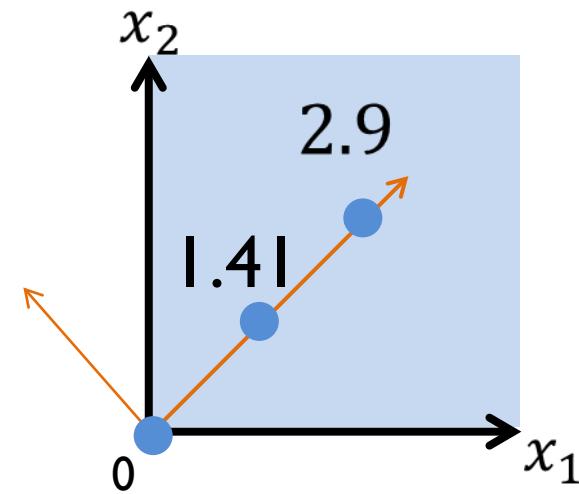
## (continued) Projection along PCs

$$XV = U\Sigma V^T V = U\Sigma$$

$$X = \begin{pmatrix} 0 & 0 \\ 1 & 1 \\ 2 & 2.1 \end{pmatrix} \quad V = \begin{pmatrix} -0.693 & -0.721 \\ -0.721 & 0.693 \end{pmatrix}$$

$$X.V = U.\Sigma = \begin{pmatrix} 0 & 0 \\ 1.414 & -0.028 \\ 2.9 & 0.0133 \end{pmatrix}$$

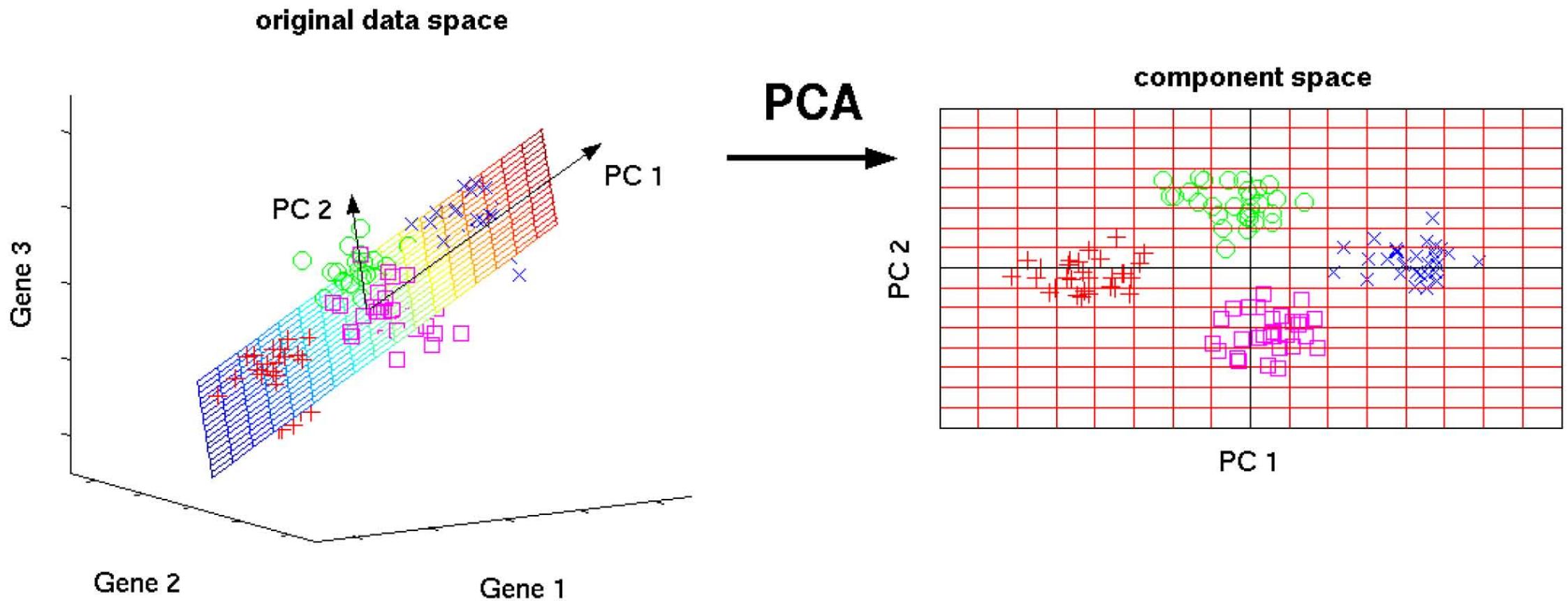
↑  
Projection along PC1



# Outline

- I. Why do we need reduction in data dimension
2. Theory of Principle Component Analysis
3. Applications of Principle Component Analysis
4. Conclusions

# Principle Component Analysis for classification



If you like this book, you will also like that book (because you belong to the same category)

# Image Transmission by Principle Component Analysis

$$X_{1000 \times 500} = u_1 \sigma_1 v_1^T + u_2 \sigma_2 v_2^T + \dots$$

$|x|$   
↓  
 $|1000 \times 1|$   
 $|1000 \times 500|$   
↑      ↑



# MATLAB code (by Camsari)

## **%% SVD - Image processing**

```
clearvars;clc;close all;  
%%  
  
IMM=imread('MonaLisa.jpg');  
IMM=im2double(IMM);  
imshow(IMM);  
%% Turn it into 2D (grayscale)  
IM2D=rgb2gray(IMM);  
%% Original in Grayscale  
figure(123)  
subplot(2,2,1)  
imshow(IM2D);  
title('Original')
```

## **%% SVD Decomposition**

```
[U,S,V]=svd(IM2D);  
%% Optional:Plot the diagonals  
%figure(234)  
%semilogy(diag(S))
```

## **%% Keep the first 50 dimensions.**

```
NR = 50;AR50=zeros(size(IM2D));  
for ii=1:NR  
    AR50=AR50+U(:,ii)*S(ii,ii)*V(:,ii)';  
end
```

## **%% Keep the first 15 dimensions.**

```
NR = 15;AR15=zeros(size(IM2D));  
for ii=1:NR  
    AR15=AR15+U(:,ii)*S(ii,ii)*V(:,ii)';  
end
```

# Conclusions

1. PCR is a powerful tool to classify multi-dimensional data (e.g. postal codes in handwritten envelops)
2. PCR decomposition by SVD provides both the rotated axes ( $V$ ) and the projection on the rotated axes ( $U\Sigma$ ). Each column of  $(U\Sigma)$  is the projection on that principal component.
3. If there are 100 dimensions and 5 key distinguishing features, then top five singular values may not align with the top five features. One should keep approximately 20 to preserve the top 5 features.
4. The desired accuracy is obtained by choosing a  $k$  such that  $p = r_k = \sum_{i=1}^k \lambda_i^2 / \sum_{i=1}^N \lambda_i^2$ .
5. Other techniques (e.g. Fisher linear discriminators) which finds the direction of the line that best separates two classes may be more accurate or efficient. For example, in Facial recognition, the PCA eigenvalues are called eigenfaces, while that from Fisher LDA is called Fisher's faces.

# Review Questions

1. What is “singular” about singular value decomposition?
2. What is the physical meaning of U and V?
3. How many Principal Components should we need to keep? How do you quantify it ?
4. What are the disadvantages of SVD-based classification? In what ways is machine-learning better?
5. What other methods of classification do we have?
6. What applications do we have SVD other than classification (e.g. data compression, etc.)?
7. Taken from your daily experience, Give several examples where SVD classification can be useful.
8. Can you do SVD with Excel? What about Wolfram alpha?

# References

## Principal Component Analysis

A tutorial on Principle Component Analysis, J. Shlens, arxiv 2009.  
([shlens@salk.edu](mailto:shlens@salk.edu))

For an interesting application in PCA, see “Recommended for you”, J. A. Konstan and J. Riedl, IEEE Spectrum, p. 55, Oc. 2012.

I understood the essence of the problem of randomization from the book “Nets, Puzzle, and Postman”, by Peter Higgins, Oxford University Press.

A wonderful set of lectures by Stuart Hunter is available in youtube, see ...  
[http://www.youtube.com/watch?v=AVUAt0Qly60&list=PLWQBDMTHPQVH3IUGF7EM\\_3XHJWFD2EIP](http://www.youtube.com/watch?v=AVUAt0Qly60&list=PLWQBDMTHPQVH3IUGF7EM_3XHJWFD2EIP)

For DOE based on Taguchi method, I liked Lloyd W Condra, “Reliability Improvement with design of experiments”, Marcel Dekker Inc., 1993. Another good book is by Ranjith Roy, “A primer on the Taguchi Method”, Van Nostrand Reinhold International Co. Ltd., 1990. Some of the examples are taken from AT&T, “Statistical Quality Control Handbook”.

Also, see the lectures on DOE by Hunter  
<http://www.youtube.com/watch?v=NoVIRaq0Uxs>  
<http://www.youtube.com/watch?v=hTviHGsl5ag>

The classical AVONA method is discussed in great detail in Chapter 13 and 14 of “Applied Statistics and Probability for Engineers, 3<sup>rd</sup> Edision, D.C. Montgomery and G. C. Runger, Wiley, 2003.

Hunter’s lectures on AVONA is also very enjoyable  
<http://www.youtube.com/watch?v=k3n9iSB6Cns>  
<http://www.youtube.com/watch?v=F05zZL3uyRo>

A slightly different approach that also reduces the number of experiments greatly is based on the response surface approach. It uses Newton-like algorithm to find the peaks/valleys of the response surface, see R. H. Myers and D.C. Montgomery, “Response Surface Methodology”, Wiley Interscience, 2002. This book discusses design of experiment in great detail.

For general reference see

Joan Fisher Box, “R.A. Fisher and the Design of Experiments, 1922-1926”, *The American Statistician*, vol. 34, no. 1, pp. 1-7, Feb. 1980.

F.Yates, “Sir Ronald Fisher and the Design of Experiments”, *Biometrics*, vol. 20, no. 2, In Memoriam: Ronald Aylmer Fisher, 1890-1962., pp. 307-321, (Jun. 1964).

# An Introduction to Data Analysis, Design of Experiment, and Machine Learning

## *Lecture 12. Basics of Machine Learning*

Muhammad A. Alam  
[alam@purdue.edu](mailto:alam@purdue.edu)



# Copyright 2018

This material is copyrighted by M. Alam under the  
following Creative Commons license:



Conditions for using these materials is described at

<http://creativecommons.org/licenses/by-nc-sa/2.5/>

Alam 2011

# Course Outline

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

Lecture 1: Introduction

Lecture 2: Collecting and plotting  $x_1, x_2, \dots x_n$

Lecture 3: Physical and empirical  $f, F, df/dx, \dots$

Lecture 4: Model selection between  $f_1, f_2, \dots$

Lecture 5: Model Selection: Cross-validation and Bootstrapping method

Lecture 6: Scaling theory with known  $f$ ,  $f(\bar{x}) = f(\bar{X})$

Lecture 7: Scaling theory with unknown  $f$ ,  $\bar{x} \rightarrow X$

Lecture 8: Design of experiments to determine  $\bar{y}_{\max} = f(\bar{x})$

Lecture 9: DOE and ANOVA

Lecture 11: Principle component analysis for classifying  $\{y\}$ .

Lecture 12-13: **Machine learning ... Statistical approach learn f**

Lecture 14: Interpretable ML: Physics-based machine learning  $f = f_{\text{physics}} + \Delta f$

Lecture 15: Conclusions

# Classification problem in big data

Advertisement  
Recommendation

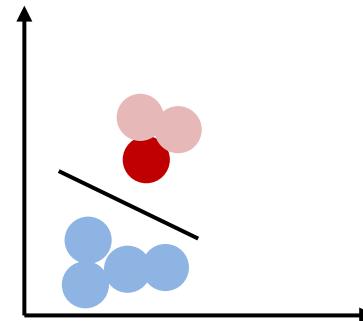


Everything is a Recommendation

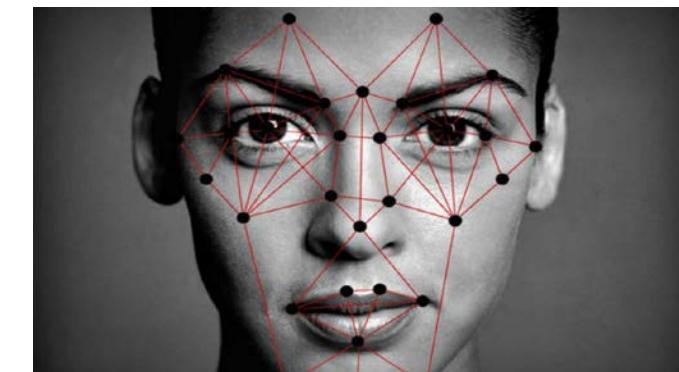
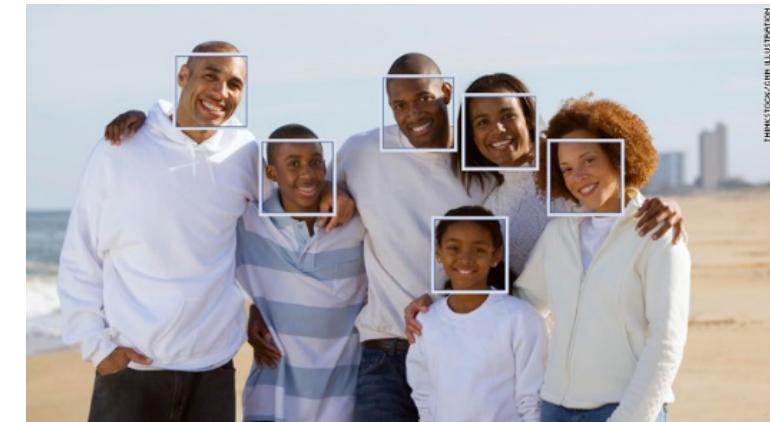


Over 75% of what people watch comes from our recommendations

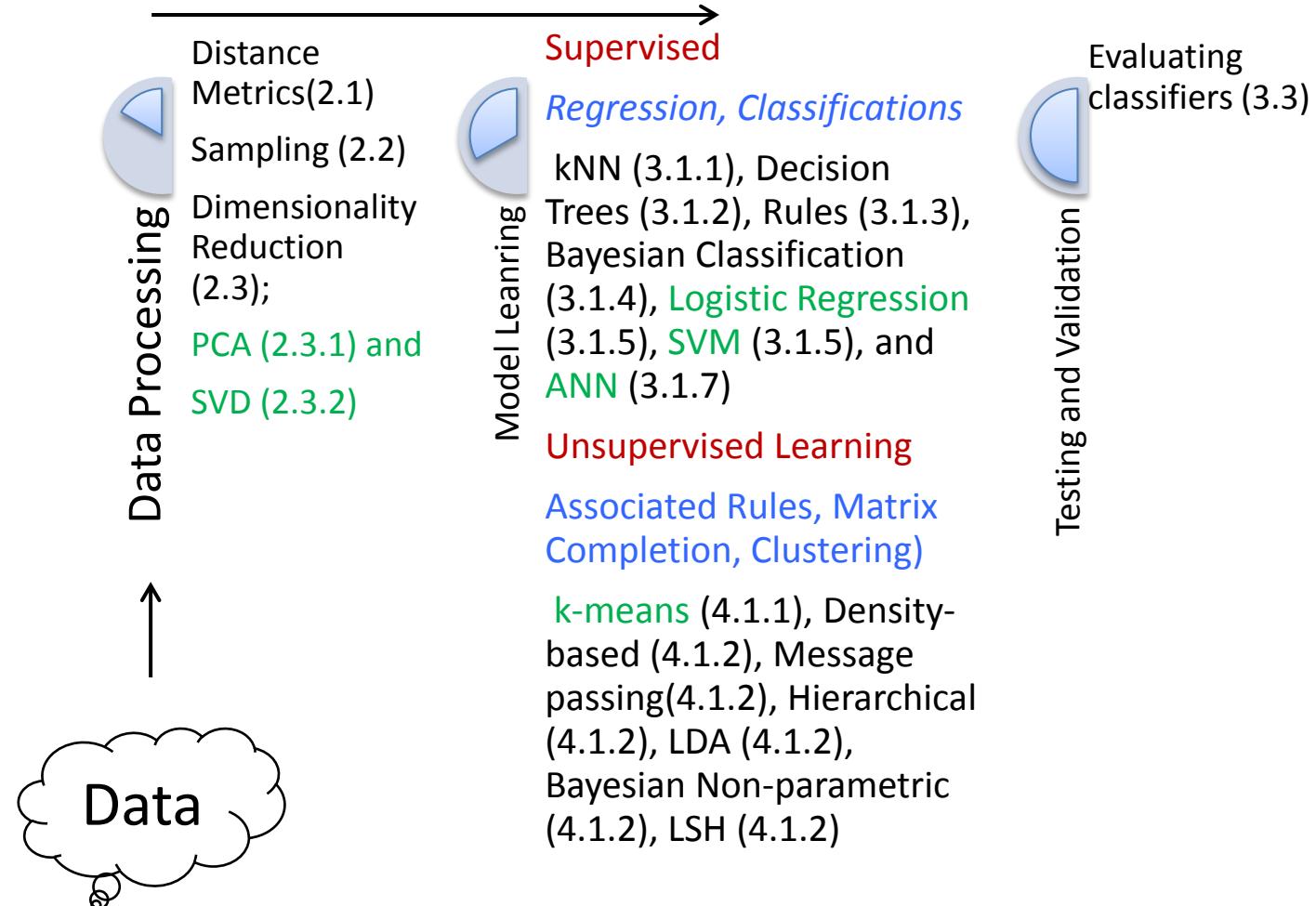
Recommendations are driven by Machine Learning



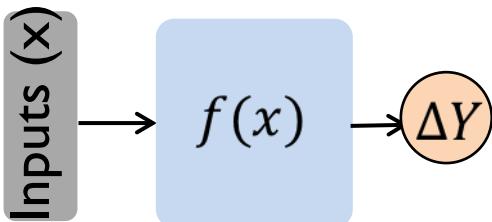
Facial Recognition  
Voice Recognition  
Spam Filtering



# Analysis of big data



# Machine Learning Introduced



$$y = f(x)$$

y: Pass, fail  
y: A, B, C, D, E  
Y: grade points.

$f(x)$  ... Physics  
 $f(x)$  ... Statistical curve fitting  
 $f_{\max}(x)$  .... Design of expt

## From the headlines

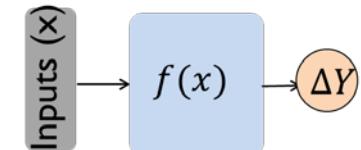
- Microsoft AI beats humans at speech recognition (TechNewsWorld)
- More accurate, fluent sentences in google translate (Barak Turovsky, lead Google Translate)
- AlphaGo: gaming that beats human (deepmind.com)
- Self driving cars (google, ....)
- Image recognition and so on ...

# Outline

- I. Machine learning is an algorithm for “fast” curve fitting
2. Machine learning and classification: Example I
3. Machine learning and classification: Example 2
4. Any function can be represented by machine learning approach
5. Conclusions

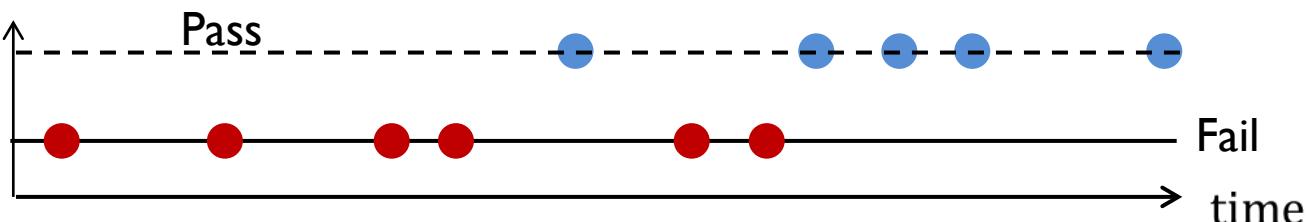
# A 1D classification problem

Input: How many hours studied;  
output: if they passed or failed  
Goal: A “machine learning” function  $f(\cdot)$

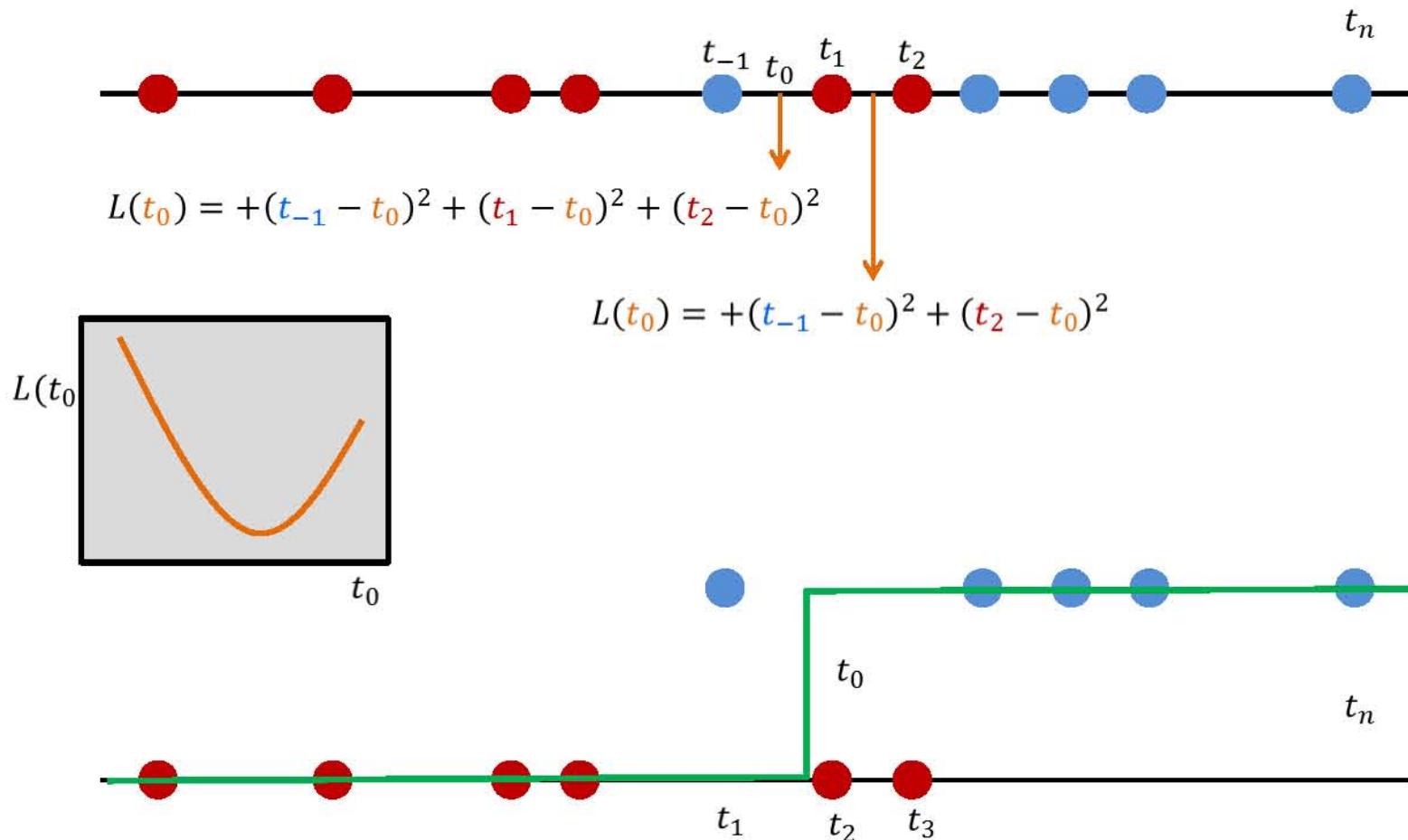


Hours	0.50	0.75	1.00	1.25	1.50	1.75	1.75	2.00	2.25	2.50
Pass	0	0	0	0	0	0	1	0	1	0

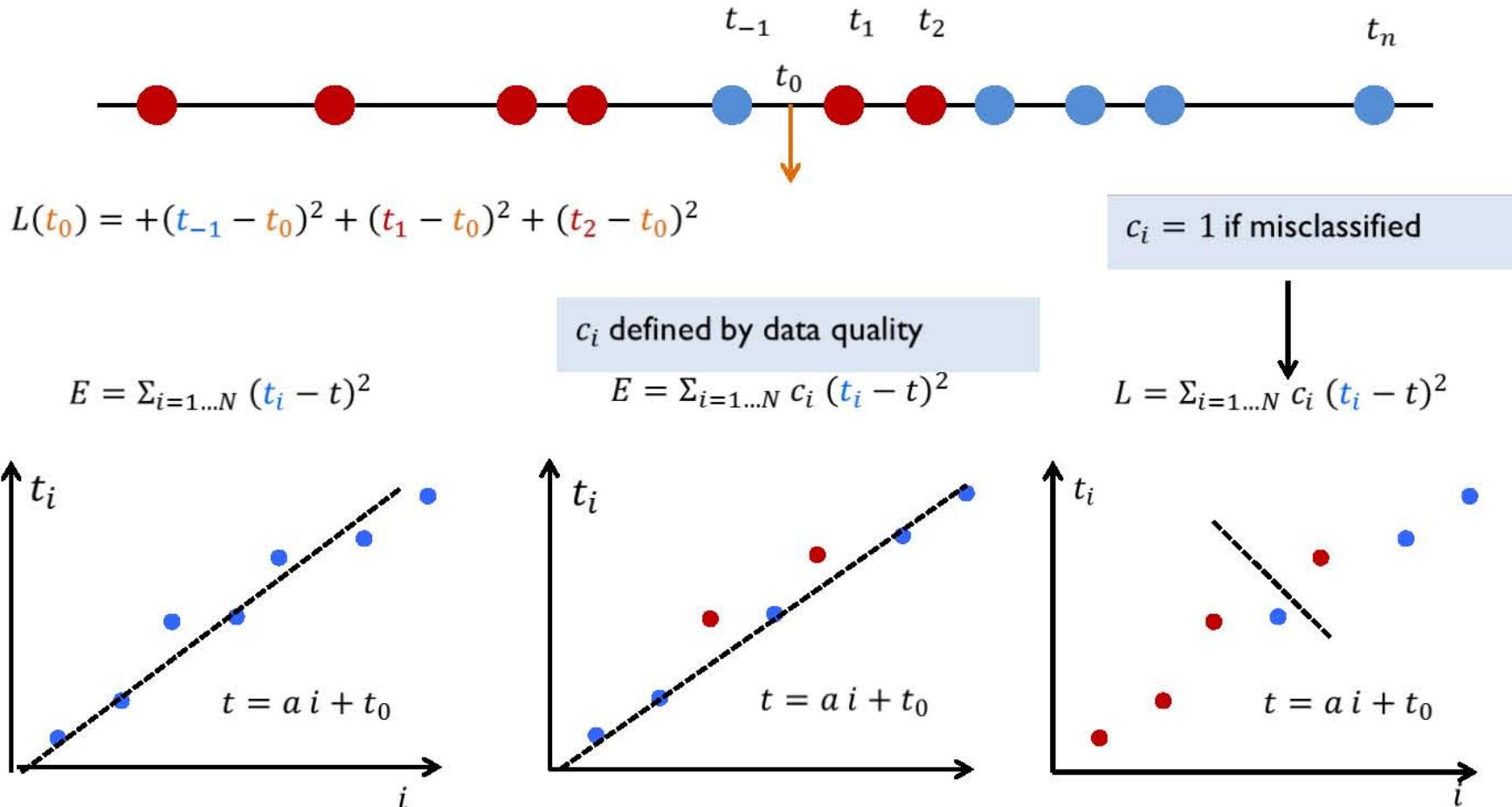
Hours	2.75	3.00	3.25	3.50	4.00	4.25	4.50	4.75	5.00	5.50
Pass	1	0	1	0	1	1	1	1	1	1



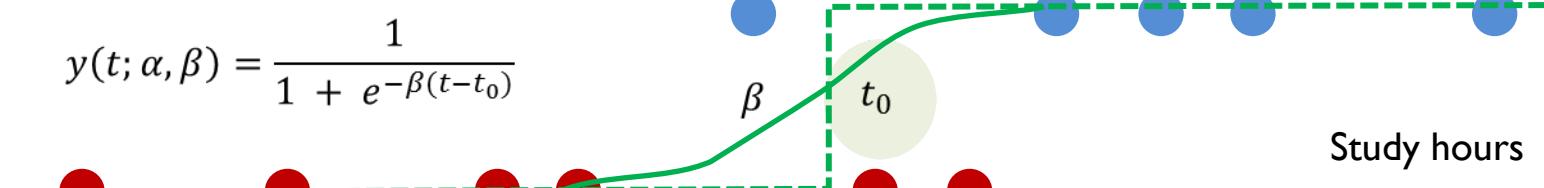
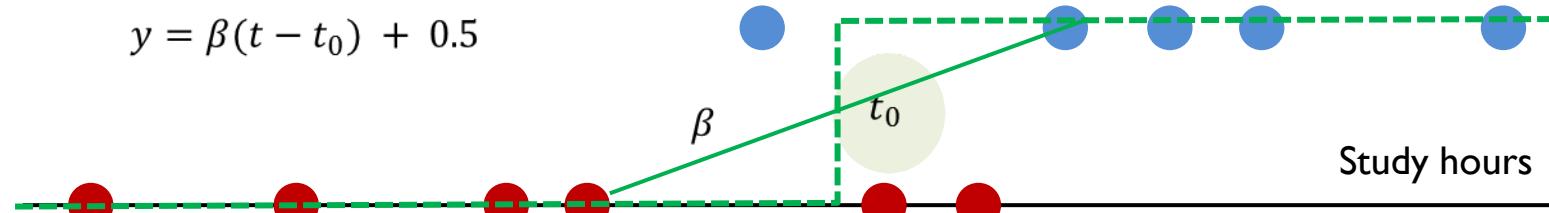
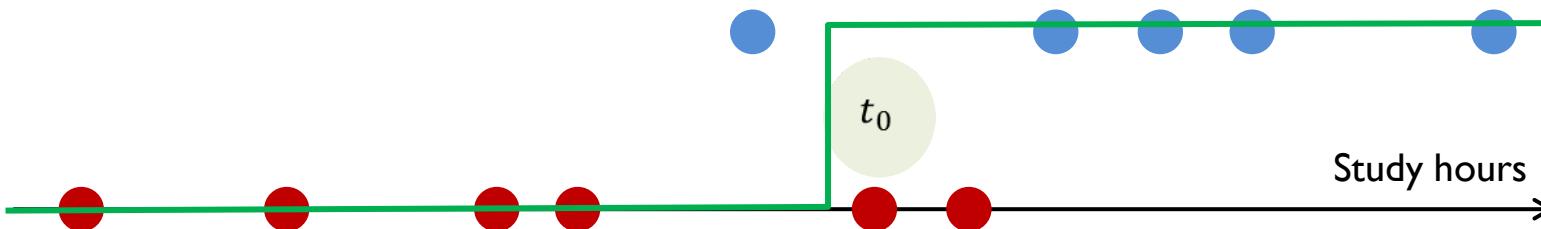
# Classification: Loss function



# Aside: Loss function vs. curve fitting



# Classification: fitting the function



# Classification by sigmoidal function

## A Wikipedia Example

$$\sigma(t; \alpha, \beta) == \frac{1}{1 + e^{-\beta(t-t_0)}} = \frac{1}{1 + e^{-(\alpha t + \beta)}}$$

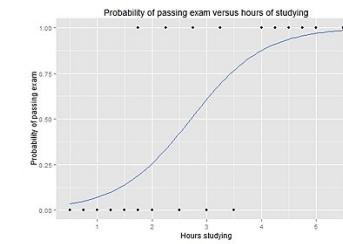
$$\sigma(t; \alpha, \beta) == \frac{1}{1 + e^{-1.505(t-2.71)}}$$

Hours	0.50	0.75	1.00	1.25	1.50	1.75	1.75	2.00	2.25	2.50
Pass	0	0	0	0	0	0	1	0	1	0

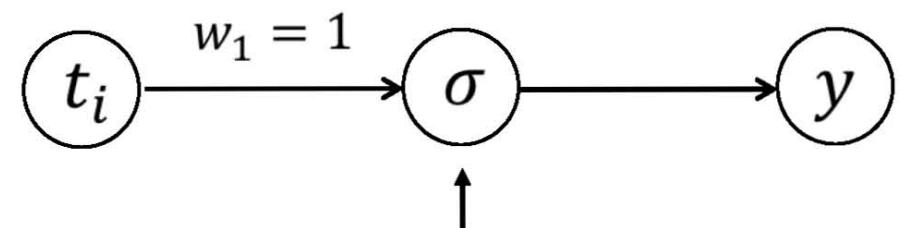
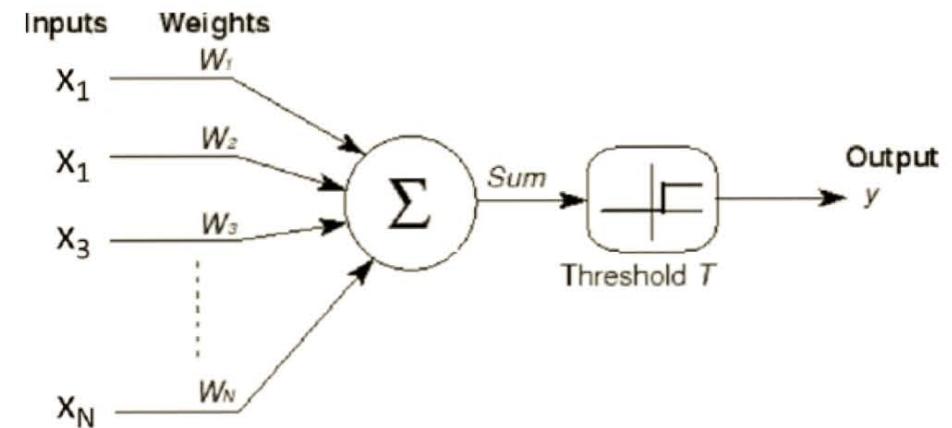
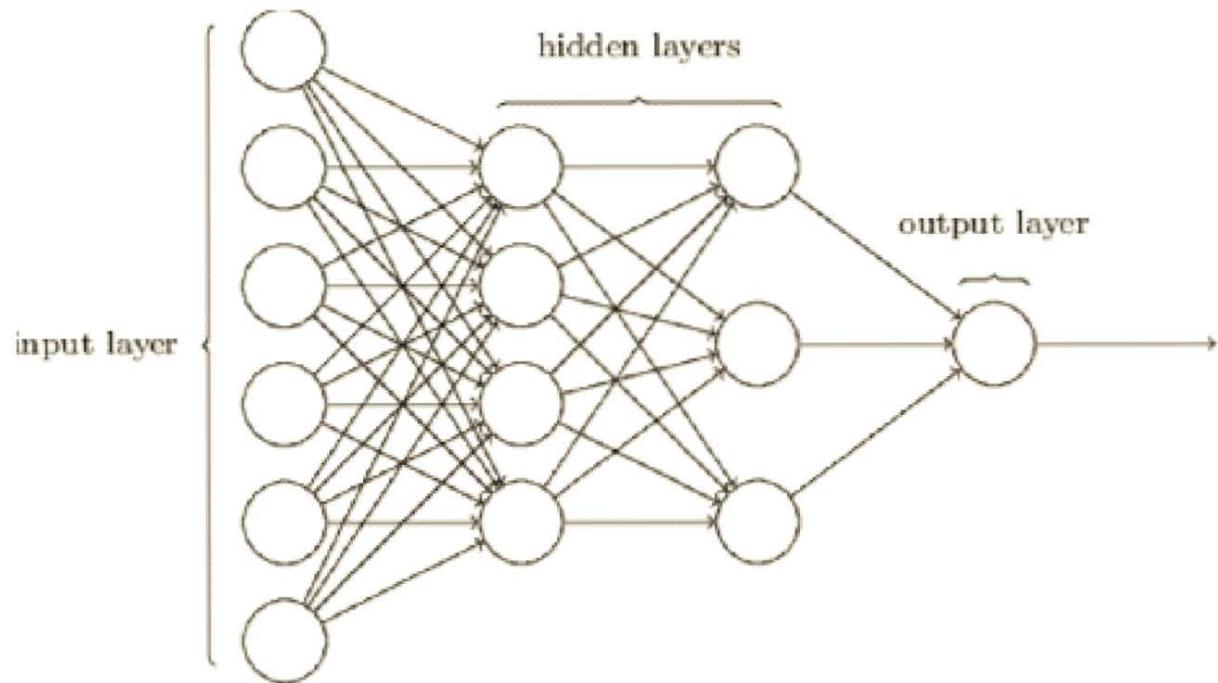
Hours	2.75	3.00	3.25	3.50	4.00	4.25	4.50	4.75	5.00	5.50
Pass	1	0	1	0	1	1	1	1	1	1

$$\ln(\sigma^{-1} - 1) = -1.505 t - 4.078 = -1.505 (t - 2.71)$$

	Coefficient	Std. Error	z-value	P-value
Intercept	-4.0777	1.7610	-2.316	0.0206
Hours	1.5046	0.6287	2.393	0.0167



# Our first machine learning circuit



$$\sigma(t; \alpha, \beta) == \frac{1}{1 + e^{-1.505(t-2.71)}}$$

# Deriving the Loss function Coefficients

$$L_0(\alpha, \beta) = \prod_{i=1 \dots N} \sigma_i(\alpha, \beta)^{y_i} \times (1 - \sigma_i(\alpha, \beta))^{1-y_i}$$

Compare with MLE where  $L_0 = \prod_{i=1 \dots N} f_i$

Appropriate for binary classification

$$-L(\alpha, \beta) = \sum_{i=1 \dots N} [y_i \ln(\sigma_i(\alpha, \beta)) + (1 - y_i) \ln(1 - \sigma_i(\alpha, \beta))]$$

$dL/d\alpha = 0$  and  $dL/d\beta = 0$  determines  $\alpha$  and  $\beta$

# One input: Numerical Example

$\alpha, \beta$			
alpha	1.5046	beta	4.077
y	x	s	L
0	0.5	0.034733	0.965267
0	0.75	0.049805	0.950195
0	1	0.070936	0.929064
0	1.25	0.100088	0.899912
0	1.5	0.139422	0.860578
0	1.75	0.190934	0.809066
1	1.75	0.190934	0.190934
0	2	0.255822	0.744178
1	2.25	0.333666	0.333666
0	2.5	0.421773	0.578227
1	2.75	0.515158	0.515158
0	3	0.607496	0.392504
1	3.25	0.692738	0.692738
0	3.5	0.76658	0.23342
1	4	0.874506	0.874506
1	4.25	0.91032	0.91032
1	4.5	0.936654	0.936654
1	4.75	0.955632	0.955632
1	5	0.969112	0.969112
1	5.5	0.985201	0.985201
		sum (L)	14.72633

$$\sigma(\alpha, \beta) = (1 + \exp(-(ax - \beta)))^{-1}$$

C5=1/(1+EXP(-(\$B\$1\*B5- \$D\$1))

$$L_i = [y_i \ln(\sigma_i(\alpha, \beta)) + (1 - y_i) \ln(1 - \sigma_i(\alpha, \beta))]$$

D5=(A5\*C5) +( (1-A5)\*(1-C5) )

$$-L(\alpha, \beta) = \sum_{i=1 \dots N} L_i(\alpha, \beta)$$

D5=SUM (D5:D24)

# Homework: One input optimization

I. Excel-based HW for understanding the fitting process.

2. Use logistic calculator to optimize the coefficient

<http://statpages.info/logistic.html>    <http://statpages.info/logistix.html>

Descriptives...						
<b>Data</b>						
10 cases have Y=0; 10 cases have Y=1.						
Variable Avg SD						
1 2.7875 1.4690						
Iteration History...						
-2 Log Likelihood = 27.7259 (Null Model)						
-2 Log Likelihood = 25.9205						
-2 Log Likelihood = 23.1187						
-2 Log Likelihood = 20.3710						
-2 Log Likelihood = 18.2717						
-2 Log Likelihood = 16.9599						
-2 Log Likelihood = 16.3181						
-2 Log Likelihood = 16.1022						
-2 Log Likelihood = 16.0626						
-2 Log Likelihood = 16.0598						
-2 Log Likelihood = 16.0598						
-2 Log Likelihood = 16.0598 (Converged)						
Overall Model Fit...						
Chi Square= 11.6661; df=1; p= 0.0006						
Coefficients, Standard Errors, Odds Ratios, and 95% Confidence Limits...						
Variable	Coeff.	StdErr	p	O.R.	Low	High
1	1.5046	0.6287	0.0167	4.5026	1.3131	15.4393
Intercept	-4.0777	1.7610	0.0206			
Predicted Probability of Outcome, with 95% Confidence Limits...						
X	Y	Prob	Low	--	High	
0.5000	0	0.0347	0.0020		0.3914	
0.7500	0	0.0498	0.0038		0.4157	
1.0000	0	0.0709	0.0073		0.4424	
1.2500	0	0.1000	0.0136		0.4722	
1.5000	0	0.1393	0.0249		0.5063	
1.7500	0	0.1908	0.0442		0.5460	
2.0000	0	0.2557	0.0749		0.5933	
2.2500	1	0.3335	0.1189		0.6498	
2.5000	0	0.4216	0.1745		0.7155	
2.7500	1	0.5150	0.2349		0.7860	
3.0000	0	0.6074	0.2930		0.8524	
3.2500	1	0.6926	0.3444		0.9062	
3.5000	0	0.7665	0.3885		0.9443	
3.7500	1	0.8444	0.4599		0.9827	
4.0000	1	0.9103	0.4897		0.9908	
4.2500	1	0.9366	0.5168		0.9951	
4.5000	1	0.9556	0.5418		0.9975	
4.7500	1	0.9691	0.5651		0.9987	
5.0000	1	0.9852	0.6079		0.9997	

# Outline

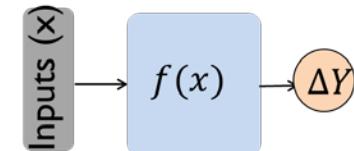
1. Machine learning is an algorithm for “fast” curve fitting
2. Machine learning and classification: Example 1
3. Machine learning and classification: Example 2
4. Any function can be represented by machine learning approach
5. Conclusions

# Generalized 1D classification problem

Input: How many hours studied;

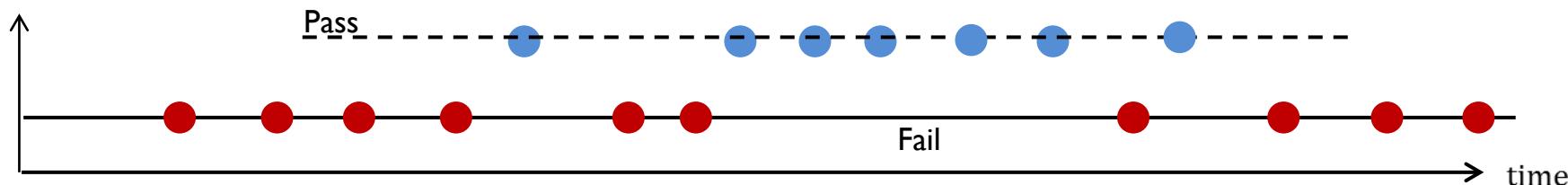
output: if they passed or failed

Goal: A “machine learning” function  $f(\cdot)$

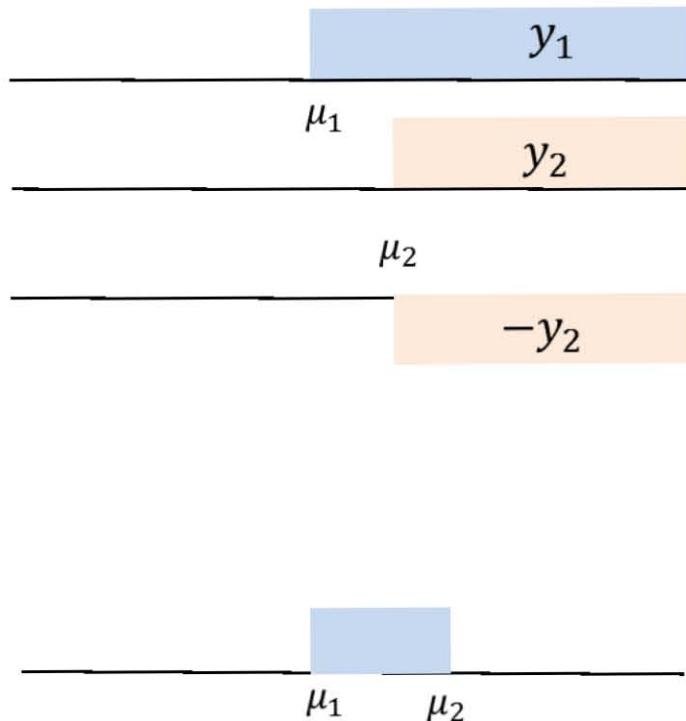


Hours	0.50	0.75	1.00	1.25	1.50	1.75	1.75	2.00	2.25	2.50
Pass	0	0	0	0	0	0	1	0	1	0

Hours	2.75	3.00	3.25	3.50	4.00	4.25	4.50	4.75	5.00	5.50	5.75	6.0	6.25	6.5	6.75	7.0
Pass	1	0	1	0	1	1	1	1	1	1	0	1	0	0	0	0



# A bit more complex classification



$$y_1 = 1 / (1 + \exp(-(w_1 x - \mu_1)/\sigma))$$

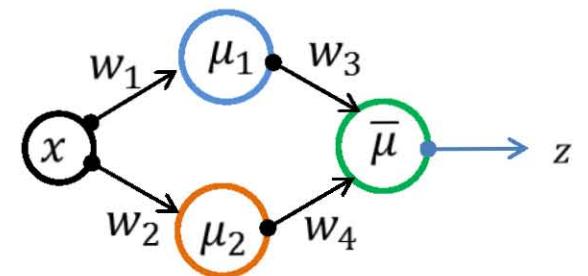
$$y_2 = 1 / (1 + \exp(-(w_2 x - \mu_2)/\sigma))$$

$$w_1 = 1, \quad w_2 = 1$$

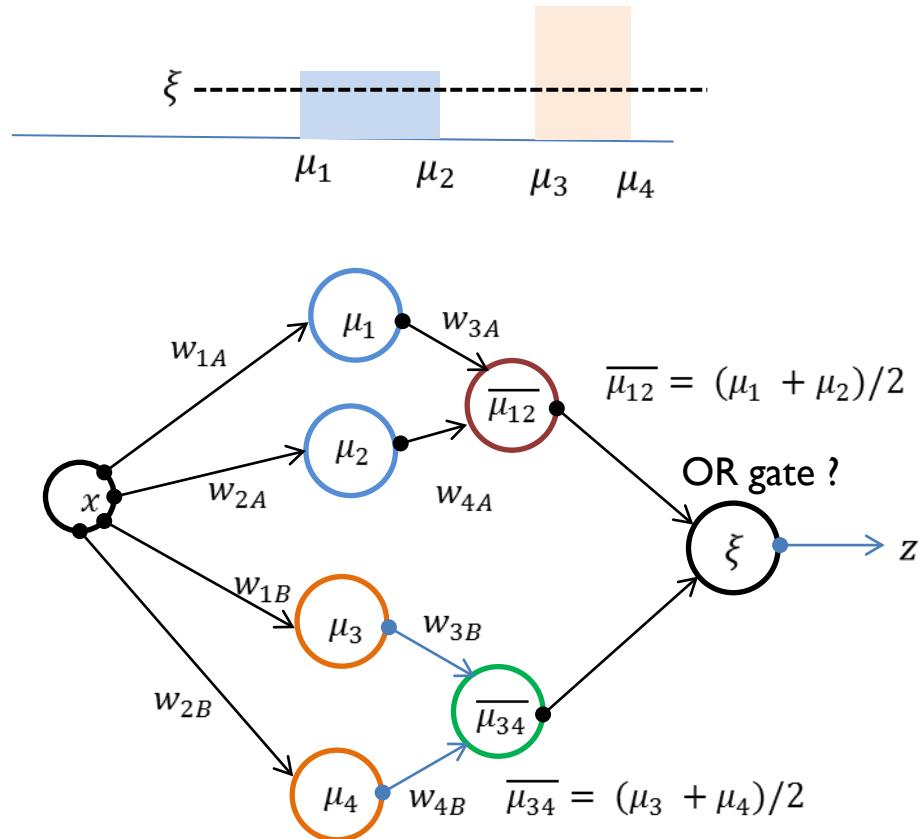
$$\bar{\mu} = (\mu_1 + \mu_2)/2$$

$$z = 1 / (1 + \exp(-(w_3 y_1 + w_4 y_2 - \mu_a)/\sigma))$$

$$w_3 = 1, w_4 = -1$$



# Any $f(x)$ can be represented by a ML network



$$y_{1A} = 1 / (1 + \exp(-(w_{1A}x - \mu_{1A})/\sigma))$$

$$y_{2A} = 1 / (1 + \exp(-(w_{2A}x - \mu_{2A})/\sigma))$$

$$z_A = 1 / (1 + \exp(-(w_{3A}y_1 + w_{4A}y_2 - \overline{\mu_{12}})/\sigma))$$

$$y_{1B} = 1 / (1 + \exp(-(w_{1B}x - \mu_{1B})/\sigma))$$

$$y_{2B} = 1 / (1 + \exp(-(w_{2B}x - \mu_{2B})/\sigma))$$

$$z_B = 1 / (1 + \exp(-(w_{3B}y_1 + w_{4B}y_2 - \overline{\mu_{34}})/\sigma))$$

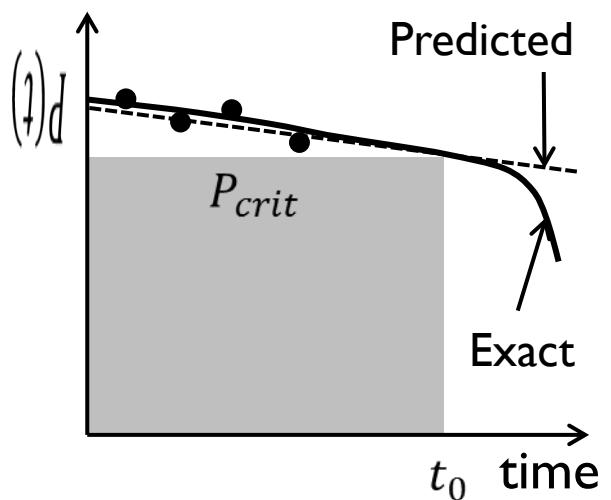
$$z = 1 / (1 + \exp(-(z_A + z_B - \xi)/\sigma))$$

$$w_{1A} = w_{2A} = w_{1B} = w_{2B} = 1, w_2 = 1$$

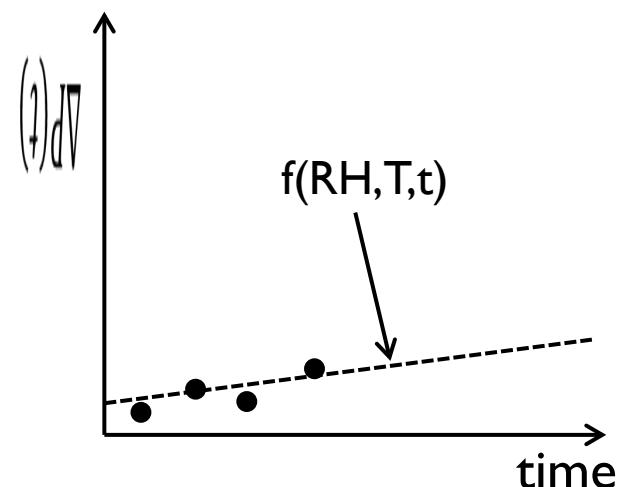
$$w_{3A} = w_{3B} = 1, \quad w_{4A} = w_{4B} = -1$$

# Reliability of Solar Farms ...

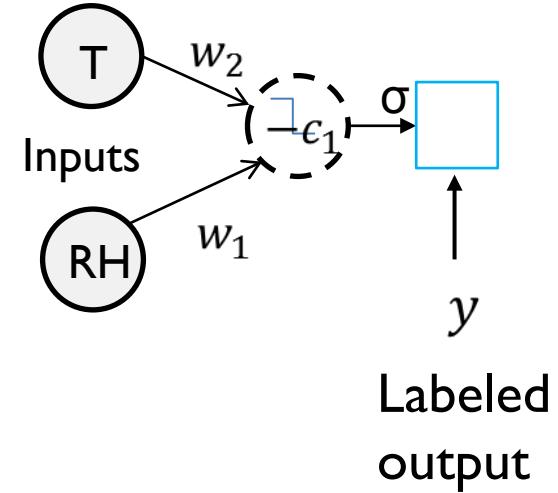
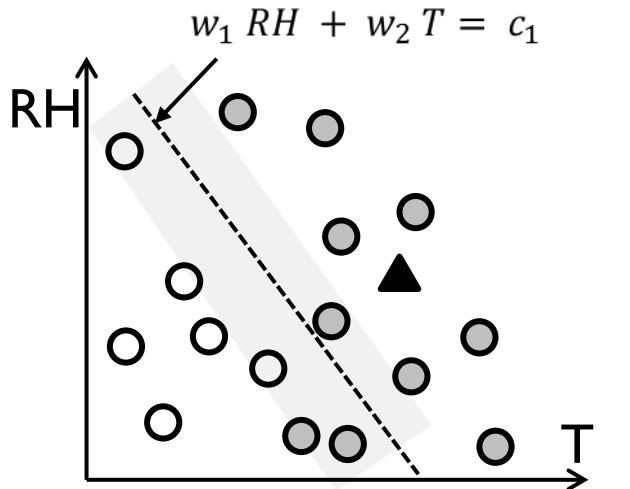
(a)



(b)

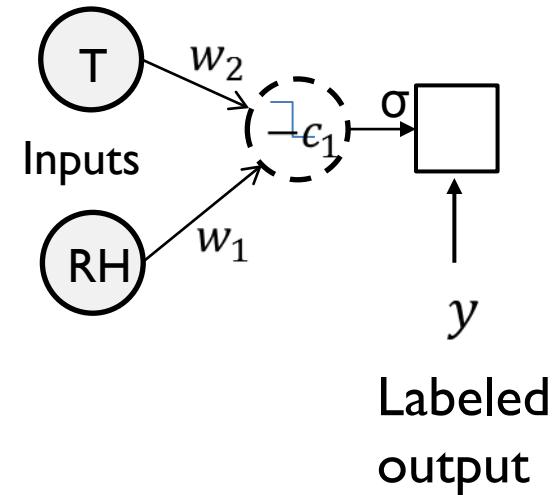
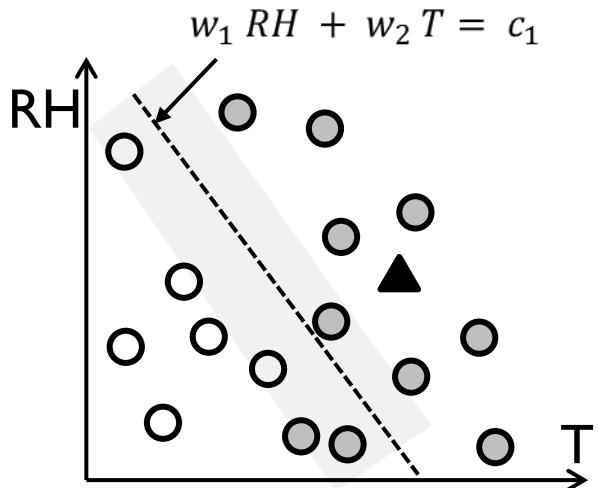


# .... represented by two input ANN



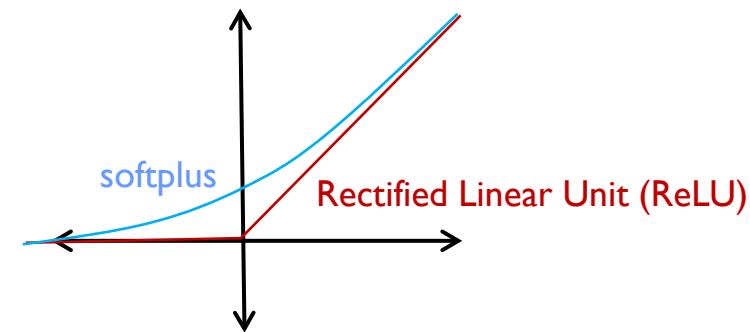
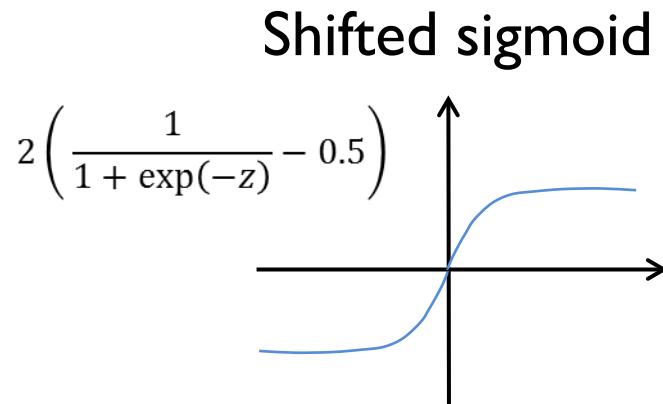
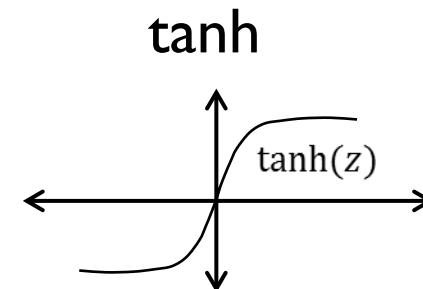
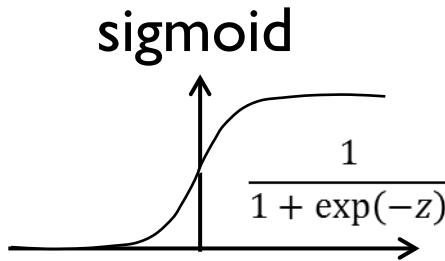
$$\sigma(w_1, w_2, c) = \frac{1}{1 + \exp(-(w_1 T + w_2 RH - c)/\sigma)}$$

# Training by backpropagation



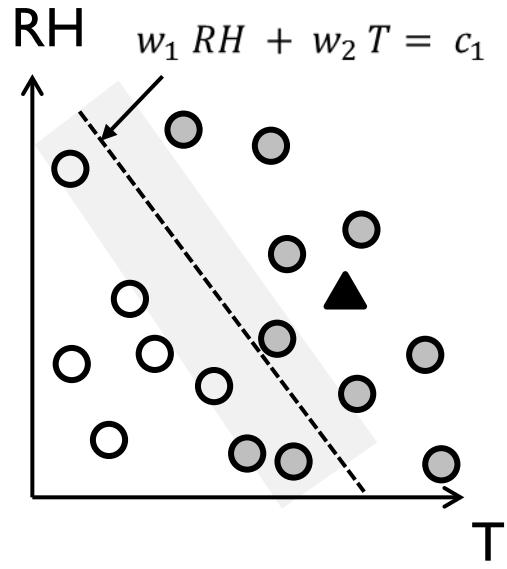
Algorithms by computer scientists  
We only have straight lines, hence many layers

# Aside: Transition Functions

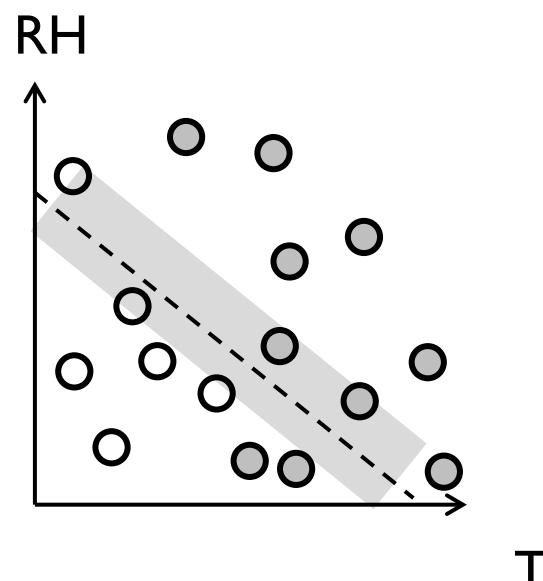


Sigmoid/tanh emphasizes points close to transition

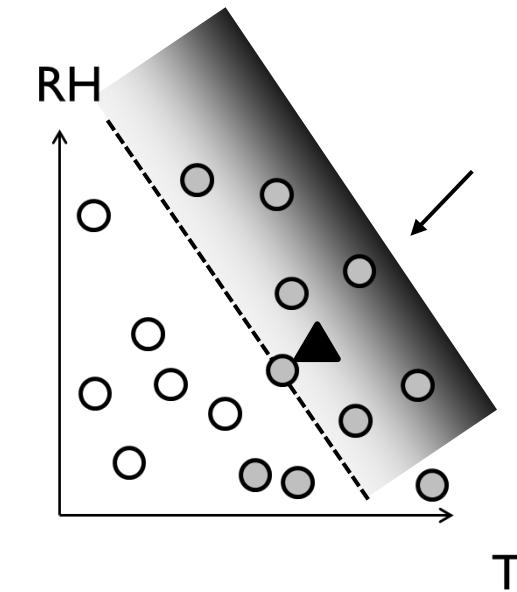
## Aside: Different transition functions



Sigmoidal



Support Vector Machine



LiRu

# Conclusions

1. Classification of data is an important statistical problem with applications in advertisement, recommendation, etc.
2. Machine learning is an empirical (and easily generalizable) multi-parameter curve fitting process. While SVD is more powerful, machine learning applies to larger datasets.
3. Any function can be represented by a machine learning algorithm. The definition of loss function and quick calculation of coefficients are the key issues.
4. We have focused on one or two input systems. The problem is easily generalized.

# Review Questions

1. What is the difference between a sigmoid function and a tanh function?
2. Why can a XOR not be implemented by a single neuron or perceptron?
3. If the weights of the input to a OR-neuron is 0.6 and 1.2, what should be its threshold?
4. What does the support vectors of a support vector machine (SVM) refer to?
5. In what ways is a SVM better than a sigmoidal transition function?
6. How does a random forest model compare with that of neural network model?
7. What is a loss function?
8. How does the sigmoid or tanh transformation of the original data reduces the sensitivity of accidental misclassification of the data?

# References

Machine Learning for Absolute Beginners, 2<sup>nd</sup> Edition, Oliver Theobald.

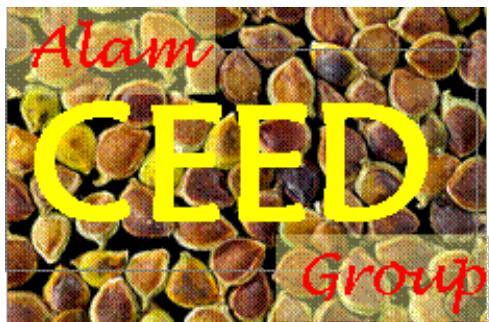
Kolmogorov, Andrei Nikolaevich. "On the representation of continuous functions of many variables by superposition of continuous functions of one variable and addition." Doklady Akademii Nauk. Vol. 114. No. 5. Russian Academy of Sciences, 1957.

Cybenko, George. "Approximation by superpositions of a sigmoidal function." Mathematics of control, signals and systems 2.4 (1989): 303-314

# An Introduction to Data Analysis, Design of Experiment, and Machine Learning

## *Lecture 13. Deep Learning, Karnaugh Mapping, and Unsupervised Classification*

Muhammad A. Alam  
[alam@purdue.edu](mailto:alam@purdue.edu)



Muhammad A. Alam, Purdue University

# Copyright 2018

This material is copyrighted by M. Alam under the  
following Creative Commons license:



Conditions for using these materials is described at

<http://creativecommons.org/licenses/by-nc-sa/2.5/>

# Course Outline

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

Lecture 1: Introduction

Lecture 2: Collecting and plotting  $x_1, x_2, \dots x_n$

Lecture 3: Physical and empirical  $f, F, df/dx, \dots$

Lecture 4: Model selection between  $f_1, f_2, \dots$

Lecture 5: Model Selection: Cross-validation and Bootstrapping method

Lecture 6: Scaling theory with known  $f$ ,  $f(\bar{x}) = f(\bar{X})$

Lecture 7: Scaling theory with unknown  $f$ ,  $\bar{x} \rightarrow X$

Lecture 8: Design of experiments to determine  $\bar{y}_{\max} = f(\bar{x})$

Lecture 9: DOE and ANOVA

Lecture 11: Principle component analysis for classifying  $\{y\}$ .

Lecture 12-13: **Machine learning ... Statistical approach learn f**

Lecture 14: Interpretable ML: Physics-based machine learning  $f = f_{\text{physics}} + \Delta f$

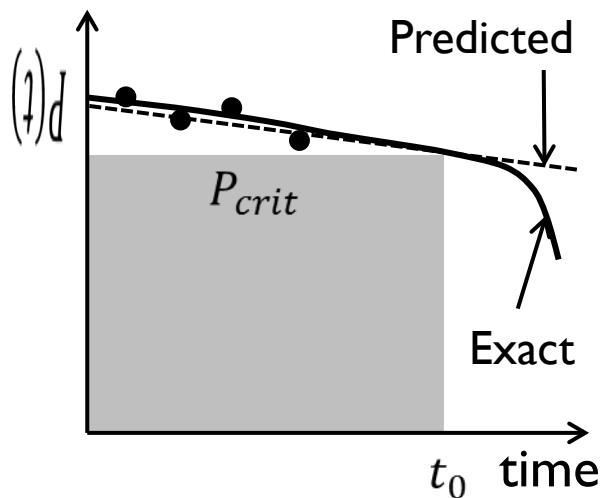
Lecture 15: Conclusions

# Outline

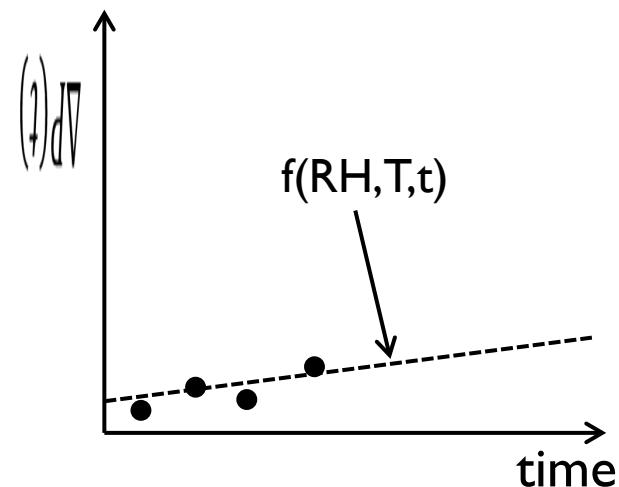
- I. Introduction
2. A two input, single and multiple perceptron problem
3. Backpropagation and coefficient fitting
4. Machine learning and Karnaugh mapping
5. Other forms of Machine Learning (Unsupervised, optical, quantum)
6. Conclusions

# Reliability of Solar Farms ...

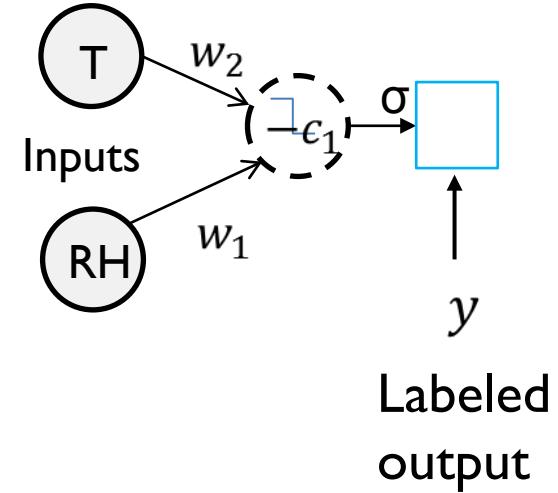
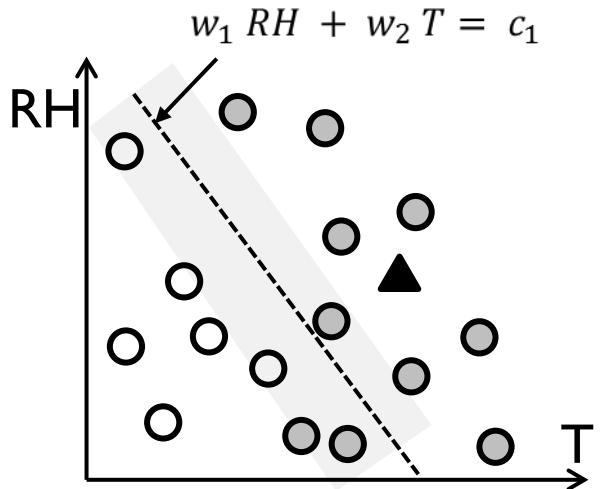
(a)



(b)

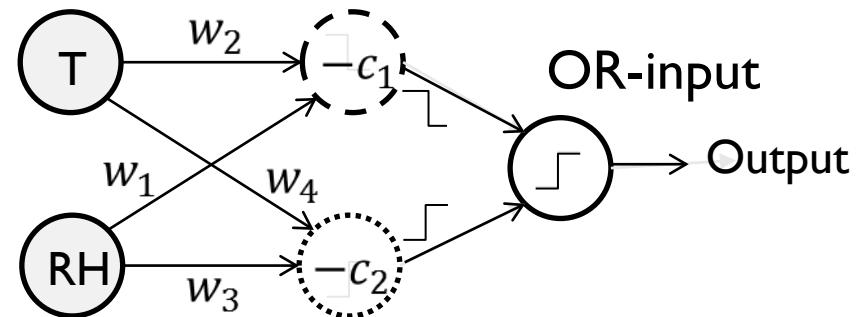
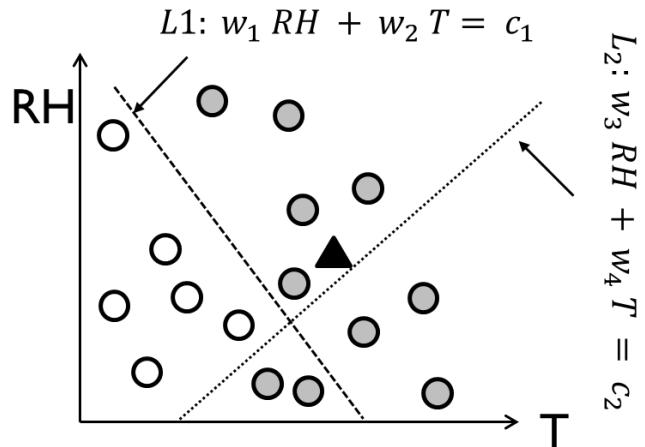


# .... represented by two input ANN

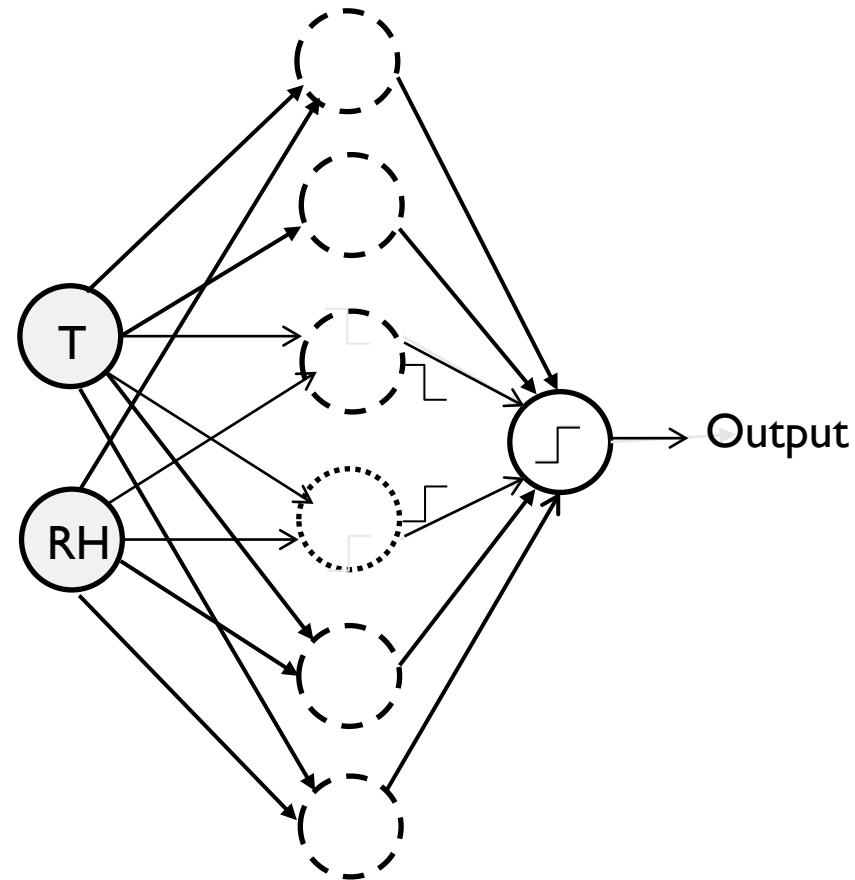
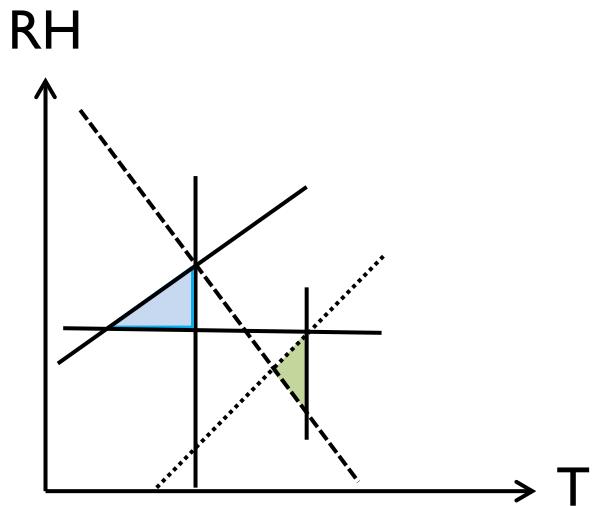


$$\sigma(w_1, w_2, c) = \frac{1}{1 + \exp(-(w_1 T + w_2 RH - c)/\sigma)}$$

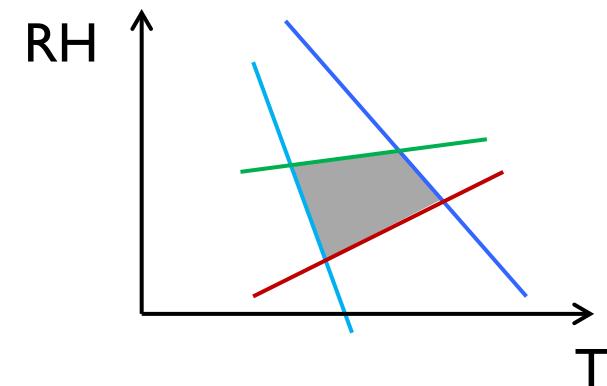
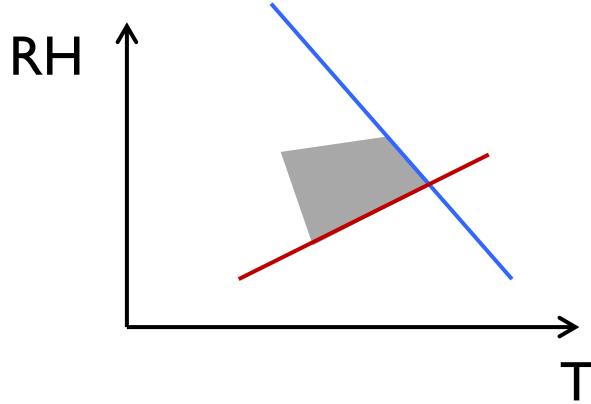
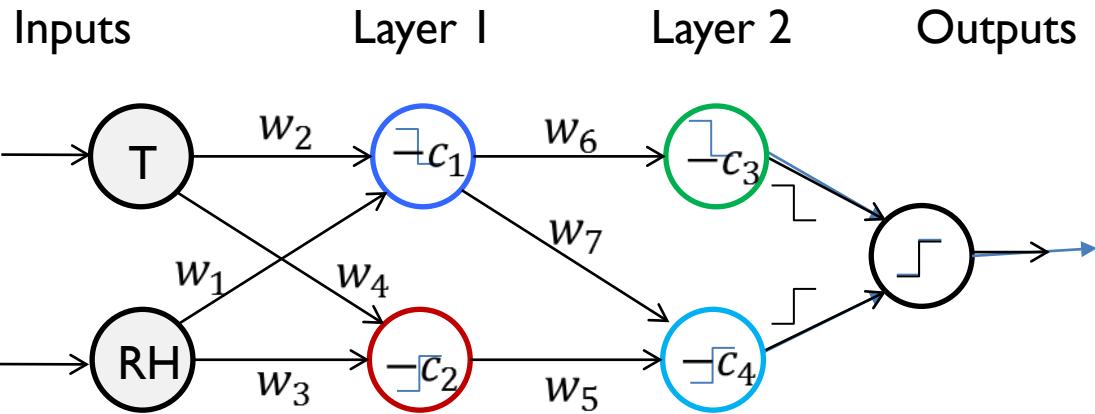
# Region defined by two lines



# Region defined by multiple lines

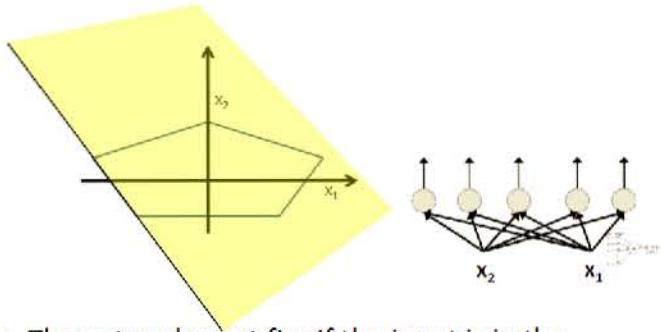


# Deep network



# 2D Image Recognition by Deep Network

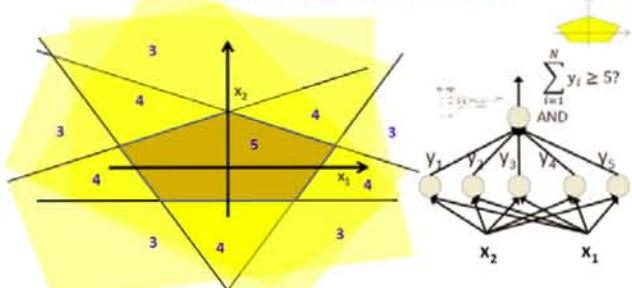
Booleans over the reals



- The network must fire if the input is in the coloured area

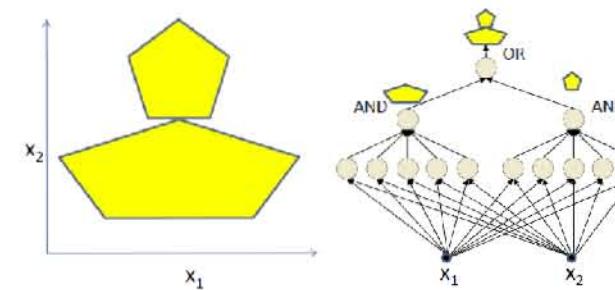
74

Booleans over the reals



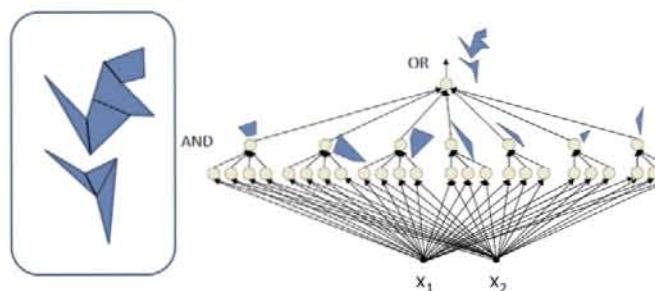
- The network must fire if the input is in the coloured area

More complex decision boundaries



- Network to fire if the input is in the yellow area
  - “OR” two polygons

Complex decision boundaries



- Can compose *arbitrarily* complex decision boundaries

75

# Outline

1. Introduction
2. A two input, single and multiple perceptron problem
3. Backpropagation and coefficient fitting
4. Machine learning and Karnaugh mapping
5. Other forms of Machine Learning (Unsupervised, optical, quantum)
6. Conclusions

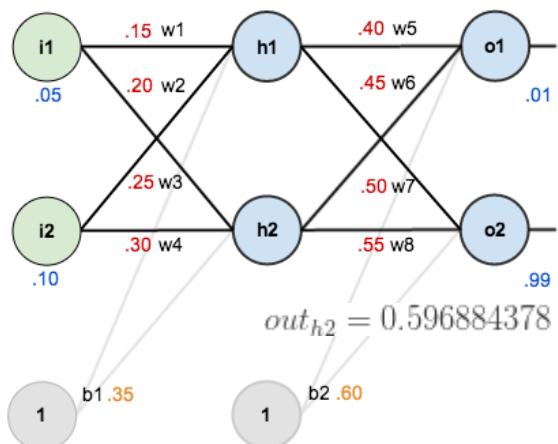
# Backpropagation algorithm

Input pair: 0.05, 0.10   Output pair 0.01, 0.99

$$net_{h1} = 0.15 * 0.05 + 0.2 * 0.1 + 0.35 * 1 = 0.3775$$

$$out_{h1} = \frac{1}{1+e^{-net_{h1}}} = \frac{1}{1+e^{-0.3775}} = 0.593269992$$

$$net_{o1} = 0.4 * 0.593269992 + 0.45 * 0.596884378 + 0.6 * 1 = 1.105905967$$



$$out_{o1} = \frac{1}{1+e^{-net_{o1}}} = \frac{1}{1+e^{-1.105905967}} = 0.75136507$$

$$E_{o1} = \frac{1}{2}(target_{o1} - out_{o1})^2 = \frac{1}{2}(0.01 - 0.75136507)^2 = 0.274811083$$

$$E_{o2} = 0.023560026$$

$$out_{o2} = 0.772928465$$

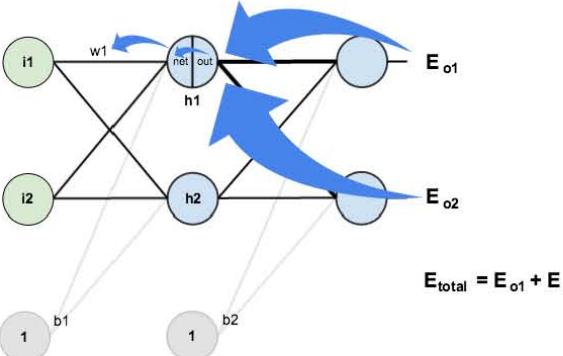


$$E_{total} = E_{o1} + E_{o2} = 0.274811083 + 0.023560026 = 0.298371109$$

<https://mattmazur.com/2015/03/17/a-step-by-step-backpropagation-example/>

# ... continued: Backpropagation algorithm

$$\begin{aligned}\frac{\partial E_{total}}{\partial w_1} &= \frac{\partial E_{total}}{\partial out_{h1}} * \frac{\partial out_{h1}}{\partial net_{h1}} * \frac{\partial net_{h1}}{\partial w_1} \\ \downarrow \\ \frac{\partial E_{total}}{\partial out_{h1}} &= \frac{\partial E_{o1}}{\partial out_{h1}} + \frac{\partial E_{o2}}{\partial out_{h1}}\end{aligned}$$



$$out_{h1} = \frac{1}{1+e^{-net_{h1}}}$$

$$\frac{\partial out_{h1}}{\partial net_{h1}} = out_{h1}(1 - out_{h1}) = 0.59326999(1 - 0.59326999) = 0.241300709$$

$$net_{h1} = w_1 * i_1 + w_3 * i_2 + b_1 * 1 \quad \frac{\partial net_{h1}}{\partial w_1} = i_1 = 0.05$$

$$\frac{\partial E_{total}}{\partial w_1} = \frac{\partial E_{total}}{\partial out_{h1}} * \frac{\partial out_{h1}}{\partial net_{h1}} * \frac{\partial net_{h1}}{\partial w_1}$$

$$\frac{\partial E_{total}}{\partial w_1} = \left( \sum_o \frac{\partial E_{total}}{\partial out_o} * \frac{\partial out_o}{\partial net_o} * \frac{\partial net_o}{\partial out_{h1}} \right) * \frac{\partial out_{h1}}{\partial net_{h1}} * \frac{\partial net_{h1}}{\partial w_1}$$

$$\frac{\partial E_{total}}{\partial out_{h1}} = \frac{\partial E_{o1}}{\partial out_{h1}} + \frac{\partial E_{o2}}{\partial out_{h1}}$$

$$w_1^+ = w_1 - \eta * \frac{\partial E_{total}}{\partial w_1} = 0.15 - 0.5 * 0.000438568 = 0.149780716$$

$$\frac{\partial E_{o1}}{\partial net_{o1}} = \frac{\partial E_{o1}}{\partial out_{o1}} * \frac{\partial out_{o1}}{\partial net_{o1}} = 0.74136507 * 0.186815602 = 0.138498562$$

$$w_2^+ = 0.19956143$$

$$net_{o1} = w_5 * out_{h1} + w_6 * out_{h2} + b_2 * 1 \quad \frac{\partial net_{o1}}{\partial out_{h1}} = w_5 = 0.40$$

$$w_3^+ = 0.24975114$$

$$\frac{\partial E_{o1}}{\partial out_{h1}} = \frac{\partial E_{o1}}{\partial net_{o1}} * \frac{\partial net_{o1}}{\partial out_{h1}} = 0.138498562 * 0.40 = 0.055399425$$

$$w_4^+ = 0.29950229$$

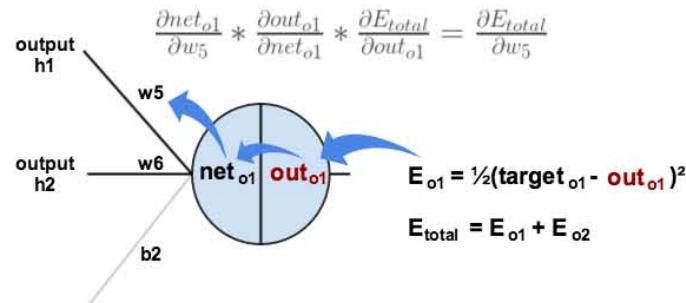
$$\frac{\partial E_{o2}}{\partial out_{h1}} = -0.019049119$$

$$\frac{\partial E_{total}}{\partial out_{h1}} = \frac{\partial E_{o1}}{\partial out_{h1}} + \frac{\partial E_{o2}}{\partial out_{h1}} = 0.055399425 + -0.019049119 = 0.036350306$$

Old error: 0.298371109

New error: 0.291027924

# Backpropagation algorithm



$$E_{total} = \frac{1}{2}(target_{o1} - out_{o1})^2 + \frac{1}{2}(target_{o2} - out_{o2})^2$$

$$\frac{\partial E_{total}}{\partial out_{o1}} = 2 * \frac{1}{2}(target_{o1} - out_{o1})^{2-1} * -1 + 0$$

$$\frac{\partial E_{total}}{\partial out_{o1}} = -(target_{o1} - out_{o1}) = -(0.01 - 0.75136507) = 0.74136507$$

$$w_6^+ = 0.408666186$$

$$\frac{\partial out_{o1}}{\partial net_{o1}} = out_{o1}(1 - out_{o1}) = 0.75136507(1 - 0.75136507) = 0.186815602$$

$$w_7^+ = 0.511301270$$

$$net_{o1} = w_5 * out_{h1} + w_6 * out_{h2} + b_2 * 1$$

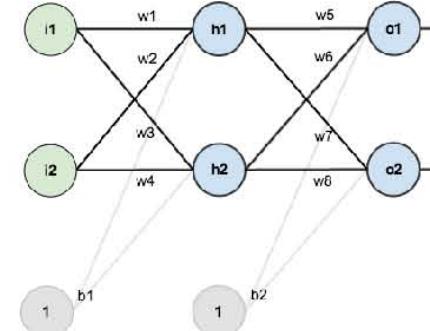
$$w_8^+ = 0.561370121$$

$$\frac{\partial net_{o1}}{\partial w_5} = 1 * out_{h1} * w_5^{(1-1)} + 0 + 0 = out_{h1} = 0.593269992$$

$$\frac{\partial E_{total}}{\partial w_5} = \frac{\partial E_{total}}{\partial out_{o1}} * \frac{\partial out_{o1}}{\partial net_{o1}} * \frac{\partial net_{o1}}{\partial w_5}$$

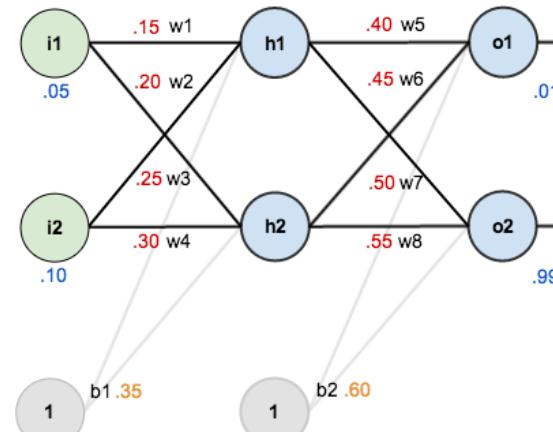
$$w_5^+ = w_5 - \eta * \frac{\partial E_{total}}{\partial w_5} = 0.4 - 0.5 * 0.082167041 = 0.35891648$$

$$\frac{\partial E_{total}}{\partial w_5} = 0.74136507 * 0.186815602 * 0.593269992 = 0.082167041$$



# ... continued: Updated Coefficients & Error

	Epoch 1	Epoch 2	Epoch
	0.1500	0.1498	
	0.2000	0.1996	
	0.2500	0.2498	
	0.3000	0.2995	
	0.4000	0.3589	
	0.4500	0.4087	
	0.5000	0.5113	
	0.5500	0.5614	
i1,i2	o1,o2	o1,o2	o1,o2
0.05	0.7514, 0.77293	xxxxxx xxxxxx	0.0156 0.9846
Err.	0.2983	0.2910	3.5e-5



After 10k iteration, gets within 1% of the final result (0.01, 0.99)

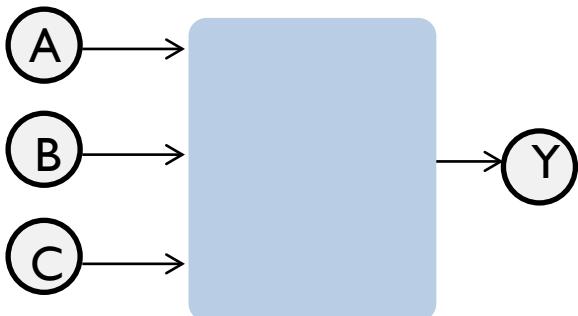
# Outline

1. Introduction
2. A two input, single and multiple perceptron problem
3. Backpropagation and coefficient fitting
4. Machine learning and Karnaugh mapping
5. Other forms of Machine Learning (Unsupervised, optical, quantum)
6. Conclusions

# Digital Synthesis has similar form

Online calculator:  
<http://www.32x8.com/>

In			Out
A	B	C	Y
0	0	0	0
0	0	1	0
0	1	0	0
0	1	1	1
1	0	0	0
1	0	1	1
1	1	0	1
1	1	1	1



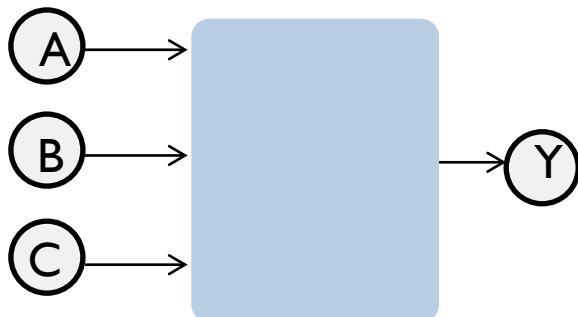
A	BC			
	00	01	11	10
0	0	0	1	0
1	0	1	1	1

$$Y = \overline{A} \cdot B \cdot C + A \cdot \overline{B} \cdot C + A \cdot B \cdot C + A \cdot B \cdot \overline{C}$$

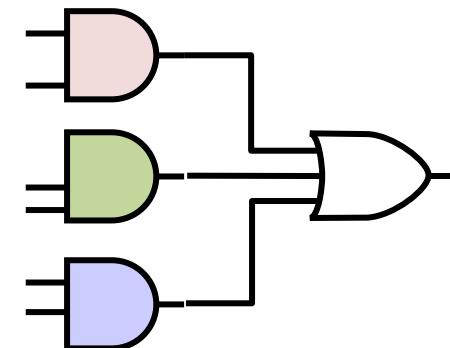
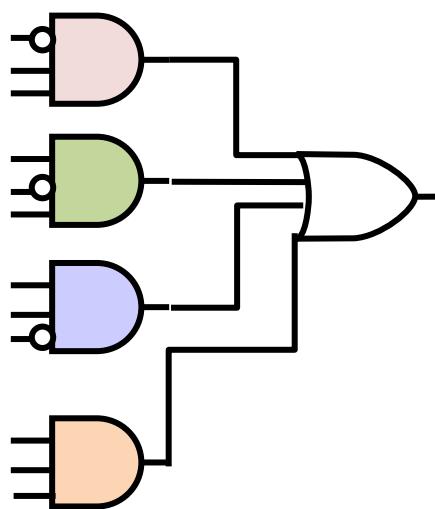
$$Y = \textcolor{red}{AC} + BC + \textcolor{violet}{AB}$$

# Neural Network and Digital logic Synthesis

$$Y = \overline{A} \cdot B \cdot C + A \cdot \overline{B} \cdot C + A \cdot B \cdot C + A \cdot B \cdot \overline{C}$$



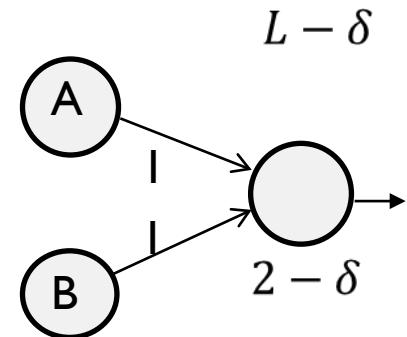
$$Y = AC + BC + AB$$



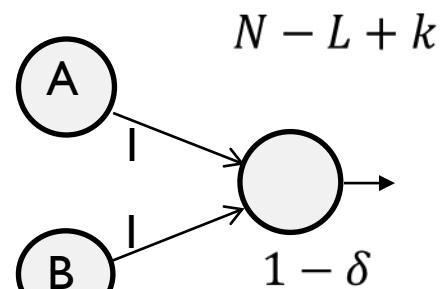
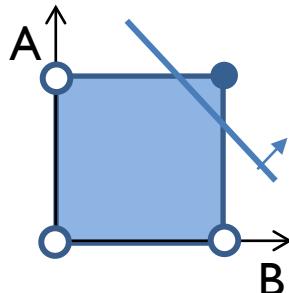
Any logic circuit can be synthesized by AND, OR, and NOT gates ...

# A perceptron implements AND, OR, NOT gates

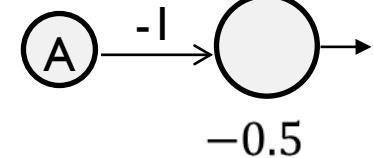
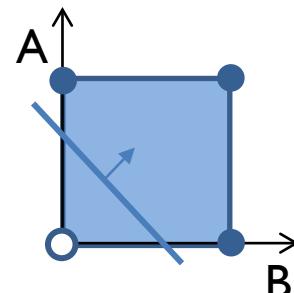
$L - N + k$  here  $L =$  (positive input),  $N =$  total number = 2,  $k = 1$  is the threshold for binary logic



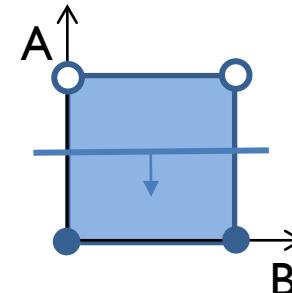
AND



OR

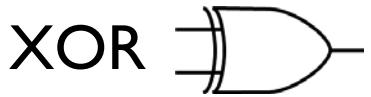


NOT



A	-A
1	0
0	1

# XOR cannot be represented by one layer

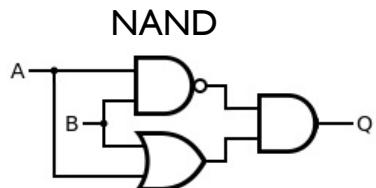
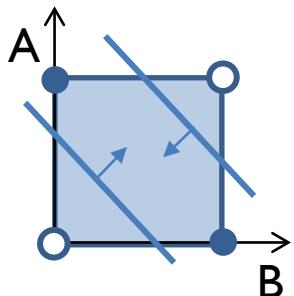


A	B	Z
0	0	0
1	0	1
0	1	1
1	1	0

$$\begin{aligned}Z &= \overline{A} \cdot B + A \cdot \overline{B} \\&= (A + B) \cdot (\overline{A} + \overline{B}) \\&= (A + B) \cdot (\overline{A} \cdot \overline{B})\end{aligned}$$

OR      NAND

	0	1
0	0	1
1	1	0



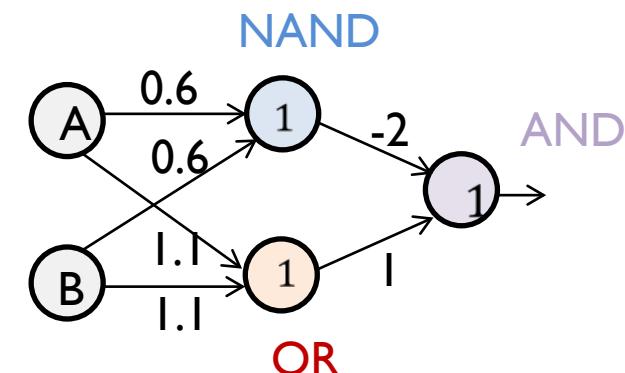
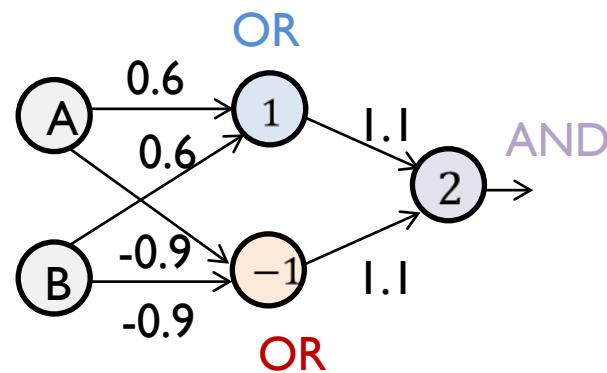
# “XOR” implementation by ANN

$$\begin{aligned} Z &= \overline{A} \cdot B + A \cdot \overline{B} \\ &= (A + B) \cdot (\overline{A} + \overline{B}) \end{aligned}$$

↑ OR      ↑ OR  
AND      AND

$$\begin{aligned} Z &= \overline{A} \cdot B + A \cdot \overline{B} \\ &= (A + B) \cdot (\overline{A} \cdot \overline{B}) \end{aligned}$$

↑ OR      ↑ NAND  
AND      AND



3 perceptrons, two layers , 6 weights and 3 threshold, 9 parameters.

# Another example ....single layer in disjunctive normal form can express any truth table

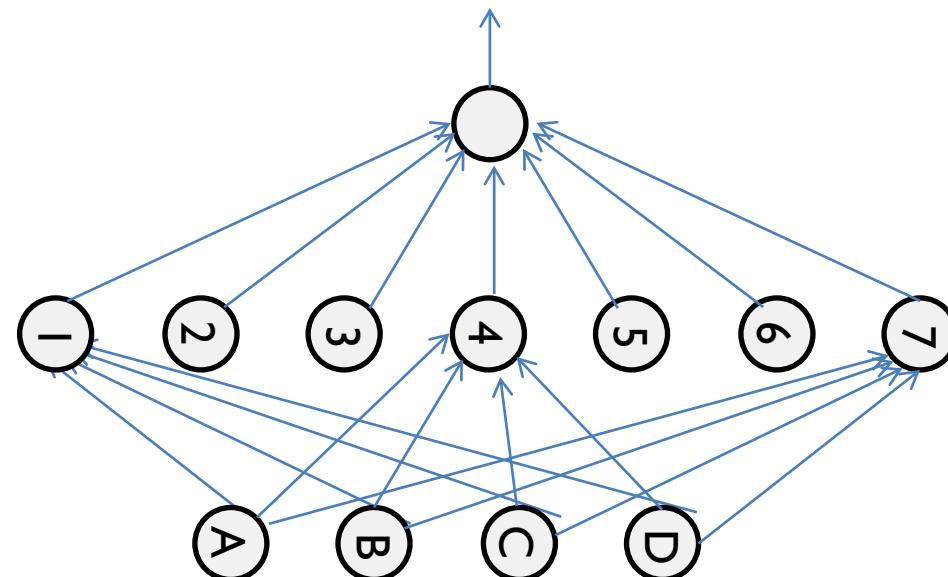
		AB			
		00	01	11	10
CD	00	1			1
	01	1	1		
	11	1			
	10	1			1

$2^{N-1}$  perceptron width in a single layer.

Exponential in  $N$

Requires  $(O N * 2^{N-1})$  weights superexponential in  $N$

$$Z = \overline{A} \cdot \overline{B} \cdot \overline{C} \cdot \overline{D} + \overline{A} \cdot \overline{B} \cdot C \cdot D + \overline{A} \cdot (B) \cdot C \cdot D + \overline{A} \cdot \overline{B} \cdot C \cdot \overline{D} \\ + \overline{A} \cdot B \cdot \overline{C} \cdot D + A \cdot \overline{B} \cdot \overline{C} \cdot D + A \cdot \overline{B} \cdot C \cdot \overline{D}$$



# Depth vs. width trade-off (Karnaugh map reduction)

Depth vs. width can be traded off.

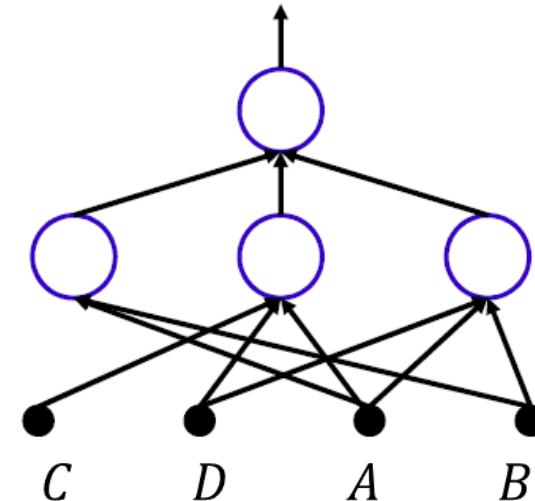
		AB			
		00	01	11	10
CD	00	1			1
	01	1	1		
	11	1			
	10	1			1

$$Z = \overline{A} \cdot \overline{B} + \overline{A} \cdot \overline{C} \cdot D + A \cdot \overline{B} \cdot \overline{D}$$

Deep network requires  
3(N-1) perceptrons with 9(N-1) parameters.

Linear in  $N$ , which can be arranged in  $2 \log_2 N$  layers.

Alternatively, Shannon limit for  $n$  input is at least  $2^n/n$



# Outline

1. Introduction
2. A two input, single and multiple perceptron problem
3. Backpropagation and coefficient fitting
4. Machine learning and Karnaugh mapping
5. Other forms of Machine Learning (Unsupervised, optical, quantum)
6. Conclusions

# Unsupervised k-mean clustering



Supervised



Unsupervised



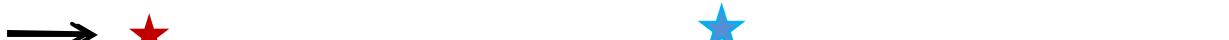
Two clusters  
Pick centroids manually



Calculate ALL centroid-to-point distances.  
Shorter distance wins the cluster

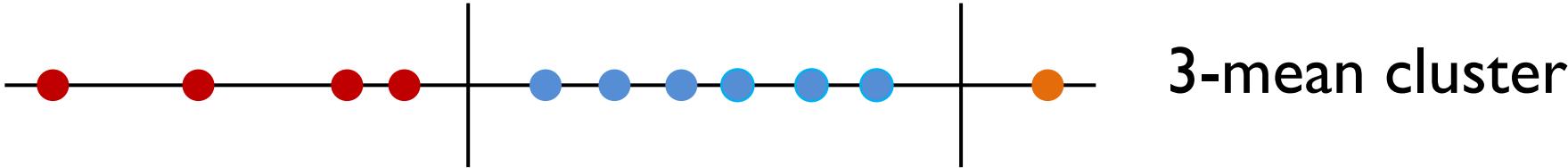
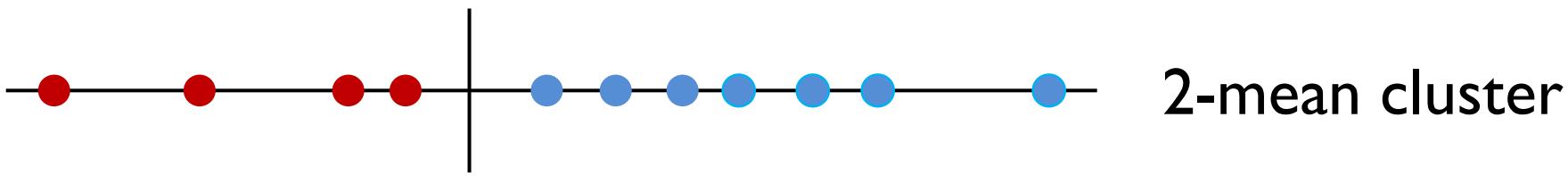
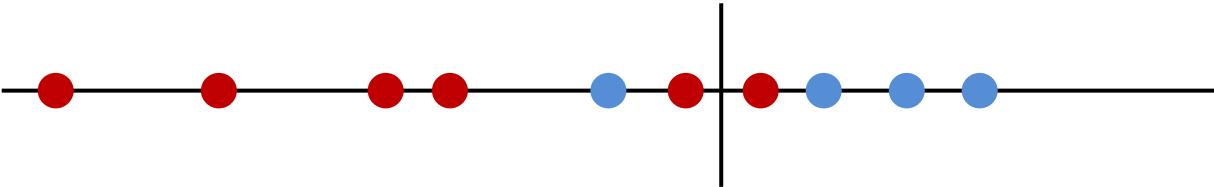


Move centroid to new average



Repeat until points do not change cluster

# Supervised vs. unsupervised clustering

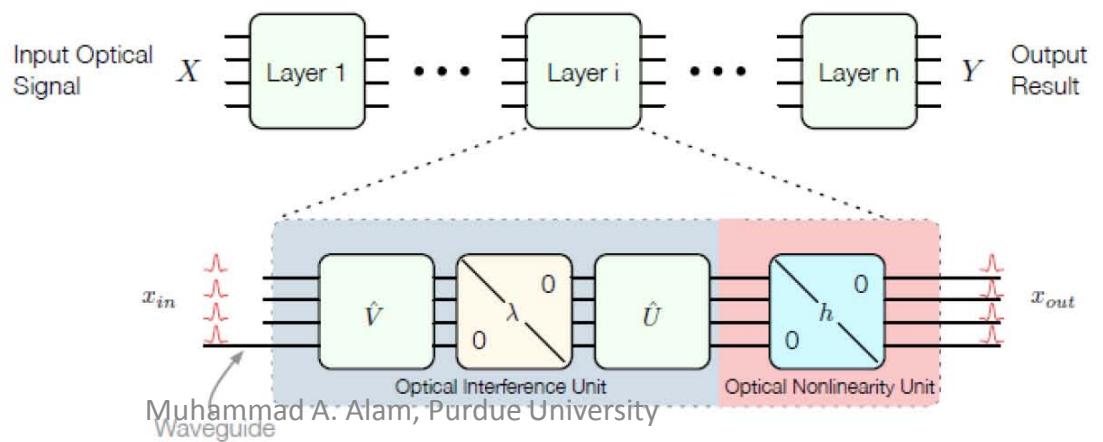
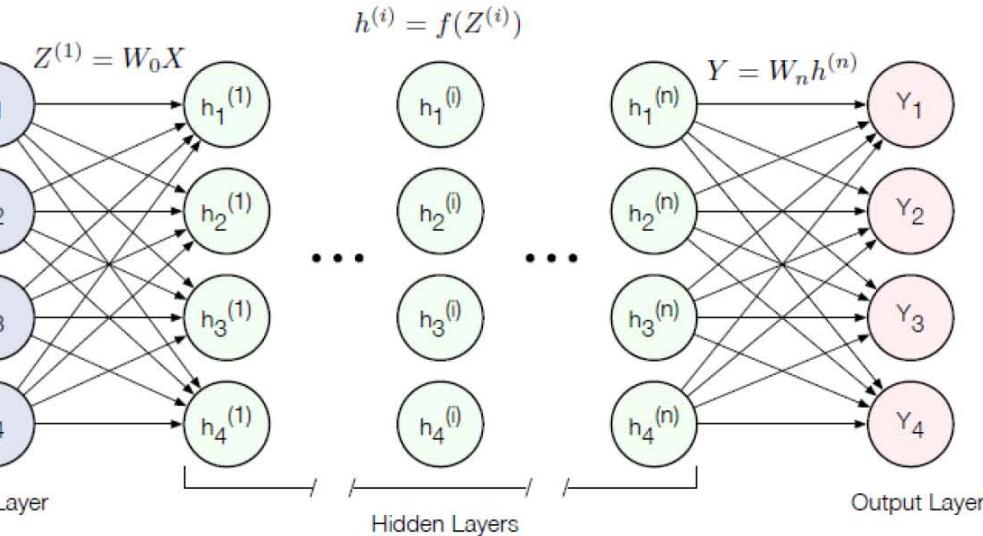


If you like this book, you may like this book  
(because the folks that belong to your cluster do)

# Deep Learning with Coherent Nanophotonic Circuits

## Deep Learning with Coherent Nanophotonic Circuits

Yichen Shen<sup>1\*</sup>, Nicholas C. Harris<sup>1\*</sup>, Scott Skirlo<sup>1</sup>, Mihika Prabhu<sup>1</sup>, Tom Baehr-Jones<sup>2</sup>, Michael Hochberg<sup>2</sup>, Xin Sun<sup>3</sup>, Shijie Zhao<sup>4</sup>, Hugo Larochelle<sup>5</sup>, Dirk Englund<sup>1</sup>, and Marin Soljačić<sup>1</sup>



Muhammad A. Alam, Purdue University

# ... continued: Nanophotonic Circuits

## Deep Learning with Coherent Nanophotonic Circuits

Yichen Shen<sup>1\*</sup>, Nicholas C. Harris<sup>1\*</sup>, Scott Skirlo<sup>1</sup>, Mihika Prabhu<sup>1</sup>, Tom Baehr-Jones<sup>2</sup>, Michael Hochberg<sup>2</sup>, Xin Sun<sup>3</sup>, Shijie Zhao<sup>4</sup>, Hugo Larochelle<sup>5</sup>, Dirk Englund<sup>1</sup>, and Marin Soljačić<sup>1</sup>

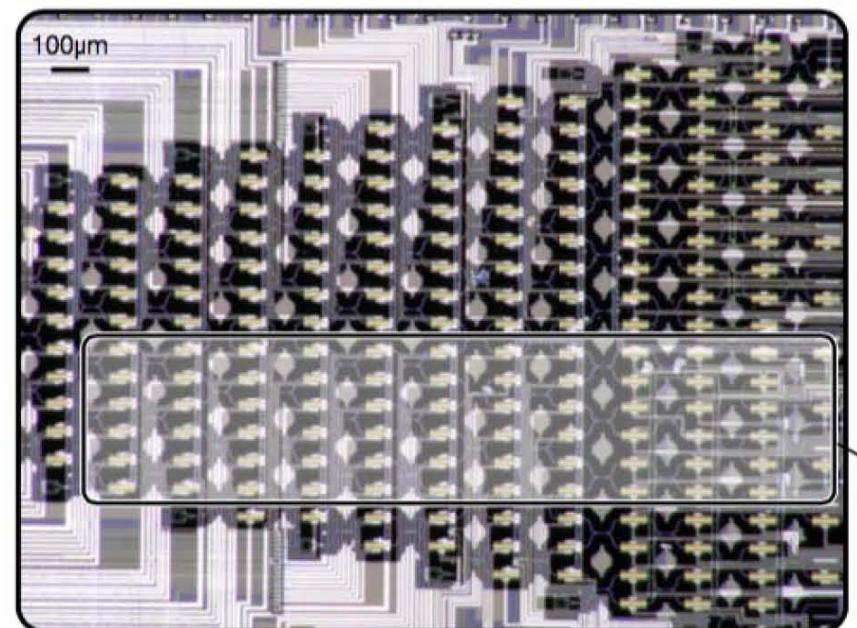
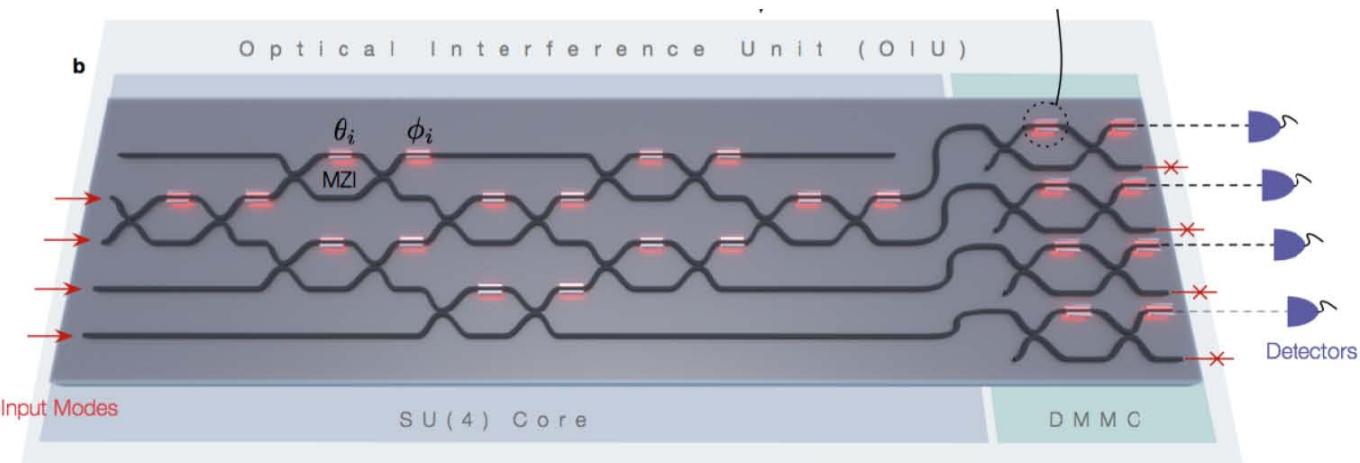


FIG. 2. Illustration of Optical Interference Unit a. Optical micrograph of an experimentally fabricated 22-mode on-chip optical interference unit; the physical region where the optical neural network program exists is highlighted in grey. The system acts as an optical field-programmable gate array—a test bed for optical experiments. b. Schematic illustration of the optical neural network program demonstrated here which realizes both matrix multiplication and amplification fully optically. c. Schematic illustration of a single phase shifter in the Mach-Zehnder Interferometer (MZI) and the transmission curve for tuning the internal phase shifter of the MZI

# Conclusions

1. Multi-input, multiple layer network can be used to represent complex functional forms.
2. Since the approach does not rely on physics, it can handle complex interpolation problem. The out-of-domain predictions are difficult and error prone.
3. The mapping onto the digital logic synthesis (i.e. Karnaugh mapping) answers some key questions regarding the depth and width of the network. It also suggests how the neural network synthesizes logic step-by-step.
4. There are variety of machine learning tools: Supervised vs. unsupervised, random forest methods, optical methodologies. All these address specific issues, such as speed of classification, energy cost of training, etc.

# References

The example involving passing probability vs. hours studied is taken from Real Statistics with Excel: Logistics Regression  
<http://www.real-statistics.com/logistic-regression>

Logistic Regression by Excel in Youtube:  
<https://www.youtube.com/watch?v=EKRjDurXau0>

Has a step by step analysis:  
<https://www.youtube.com/watch?v=jQI4pkKP9k4>

The logistic calculator is here:  
[http://astatsa.com/Logit\\_Probit/](http://astatsa.com/Logit_Probit/)

The corresponding Wikipedia page gives detailed information  
[https://en.wikipedia.org/wiki/Logistic\\_regression](https://en.wikipedia.org/wiki/Logistic_regression)

**Random Forest Model:**  
<https://www.youtube.com/watch?v=gmmV4drPTS4>

**IRandom Forest Tutorial:**  
<https://medium.com/@williamkoehrsen/random-forest-simple-explanation-377895a60d2d>

**Support Vector Machine**  
p. 67 Machine Learning for Absolute Beginners by Oliver Theobald

An excellent presentation by Xavier Amatriain on NETFlix Recommendation Systems presented at 2012 ACM Meeting <https://www.slideshare.net/xamat/netflix-recommendations-beyond-the-5-stars>

Also see  
<https://www.slideshare.net/xamat/kdd-2014-tutorial-the-recommender-problem-revisited>

An excellent tutorial in Kaggle regarding the need for multiple hidden layers (staircase problem)  
<http://blog.kaggle.com/2017/11/27/introduction-to-neural-networks/>  
<http://blog.kaggle.com/2017/12/06/introduction-to-neural-networks-2/>

**TesnsorFlow:** REF:  
<https://www.coursera.org/lecture/deep-learning-business/6-1-introduction-to-tensorflow-playground-ArFBs>  
<https://cloud.google.com/blog/products/gcp/understanding-neural-networks-with-tensorflow-playground>  
<https://developers.google.com/machine-learning/crash-course/introduction-to-neural-networks/playground-exercises>

# Machine Learning References

## Intuitive explanation:

<https://www.youtube.com/watch?v=nz-FrbAa8dY>  
<https://www.youtube.com/watch?v=eX2sY2La4Ew>

**Excel:** <https://www.youtube.com/watch?v=jQI4pkKP9k4>  
[https://www.youtube.com/watch?v=gNhogKJ\\_q7U](https://www.youtube.com/watch?v=gNhogKJ_q7U)

## Simple Visual Explanation

<https://www.youtube.com/watch?v=yIYKR4sgzl8>

## Derivation of the parameters

<https://www.youtube.com/watch?v=YMJtsYlp4kg>  
<https://www.youtube.com/watch?v=YMJtsYlp4kg>

## Step by step:

<https://www.youtube.com/watch?v=HQ7P-Ft7Cuc>

## Artificial Neural Network and Digital Logic Synthesis

<http://toritris.weebly.com/>  
<http://toritris.weebly.com/perceptron-5-xor-how--why-neurons-work-together.html>  
<https://www.youtube.com/watch?v=RALqlk7T4xc>  
<http://www-inst.eecs.berkeley.edu/~ee40/fa03/lecture/lecture29.pdf>  
<https://www.youtube.com/watch?v=FOf00W8WSBg>  
<https://www.youtube.com/watch?v=UdpV-ksadkQ> (must make the group in powers of 2)

How many hidden layers:

[https://cse.buffalo.edu/~hungngo/classes/2010/711/lectures/008\\_1.pdf](https://cse.buffalo.edu/~hungngo/classes/2010/711/lectures/008_1.pdf)

What size net gives valid generalization?

E. B. Baum and David Haussler

# Machine Learning References

[1] Wide Residual Networks

Sergey Zagoruyko, Nikos Komodakis

(Submitted on 23 May 2016 (v1), last revised 14 Jun 2017  
(this version, v4))

URL: <https://arxiv.org/abs/1605.07146>

[2] M. Bianchini and F. Scarselli, "On the Complexity of Neural Network Classifiers: A Comparison Between Shallow and Deep Architectures," in IEEE Transactions on Neural Networks and Learning Systems, vol. 25, no. 8, pp. 1553-1565, Aug. 2014.

[3] He, Kaiming, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. "Deep residual learning for image recognition." In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 770-778. 2016.

[4] <https://www.quora.com/Why-are-neural-networks-becoming-deeper-more-layers-but-not-wider-more-nodes-per-layer#>

Discrete Weights:

Baldassi, C., Borgs, C., Chayes, J.T., Ingrosso, A., Lucibello, C., Saglietti, L. and Zecchina, R., 2016. Unreasonable effectiveness of learning neural networks: From accessible states and robust ensembles to basic algorithmic schemes.

Proceedings of the National Academy of Sciences, 113(48), pp.E7655-E7662.

<http://www.pnas.org/content/113/48/E7655.short>

# Review Questions

1. Any function can be represented by a single layer neurons. If so, why does one use multiple layer “deep” network?
2. Explain why we use tanh, sigmoid, or LiRu other saturated function in machine learning.
3. What are the support vectors in a support vector machine?
4. What are ID3, decision tree, and random forest model? Explain the applicability of the model.
5. What is the difference between supervised vs. non-supervised learning?  
How does the random clustering model work?

# An Introduction to Data Analysis, Design of Experiment, and Machine Learning

## *Lecture 14. Physics-based Machine Learning*

Muhammad A. Alam

[alam@purdue.edu](mailto:alam@purdue.edu)



# Copyright 2018

This material is copyrighted by M. Alam under the  
following Creative Commons license:



Conditions for using these materials is described at

<http://creativecommons.org/licenses/by-nc-sa/2.5/>

# Course Outline

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

Lecture 1: Introduction

Lecture 2: Collecting and plotting  $x_1, x_2, \dots x_n$

Lecture 3: Physical and empirical  $f, F, df/dx, \dots$

Lecture 4: Model selection between  $f_1, f_2, \dots$

Lecture 5: Model Selection: Cross-validation and Bootstrapping method

Lecture 6: Scaling theory with known  $f$ ,  $f(\bar{x}) = f(\bar{X})$

Lecture 7: Scaling theory with unknown  $f$ ,  $\bar{x} \rightarrow X$

Lecture 8: Design of experiments to determine  $\bar{y}_{\max} = f(\bar{x})$

Lecture 9: DOE and ANOVA

Lecture 11: Principle component analysis for classifying  $\{y\}$ .

Lecture 12: Machine learning ... Statistical approach learn  $f$

Lecture 13: Machine learning ... Deep network, Karnaugh map, and other approaches

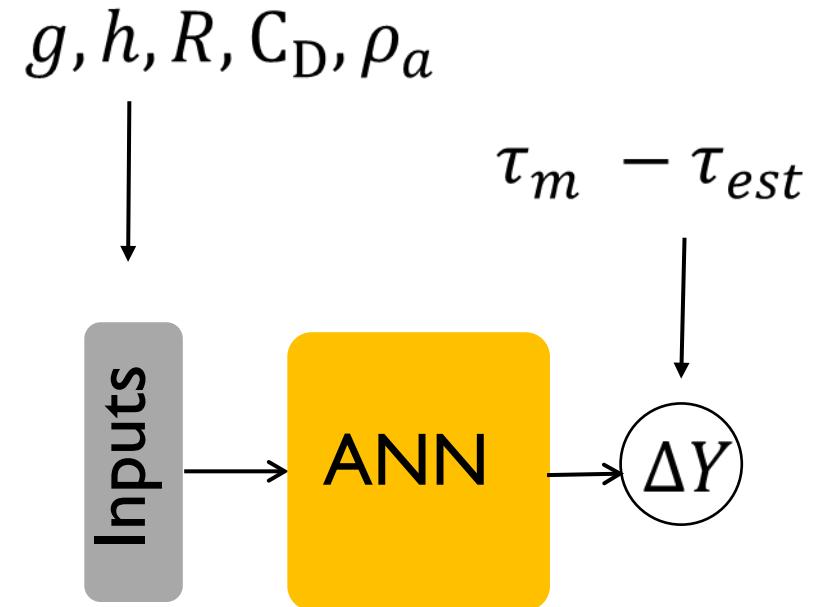
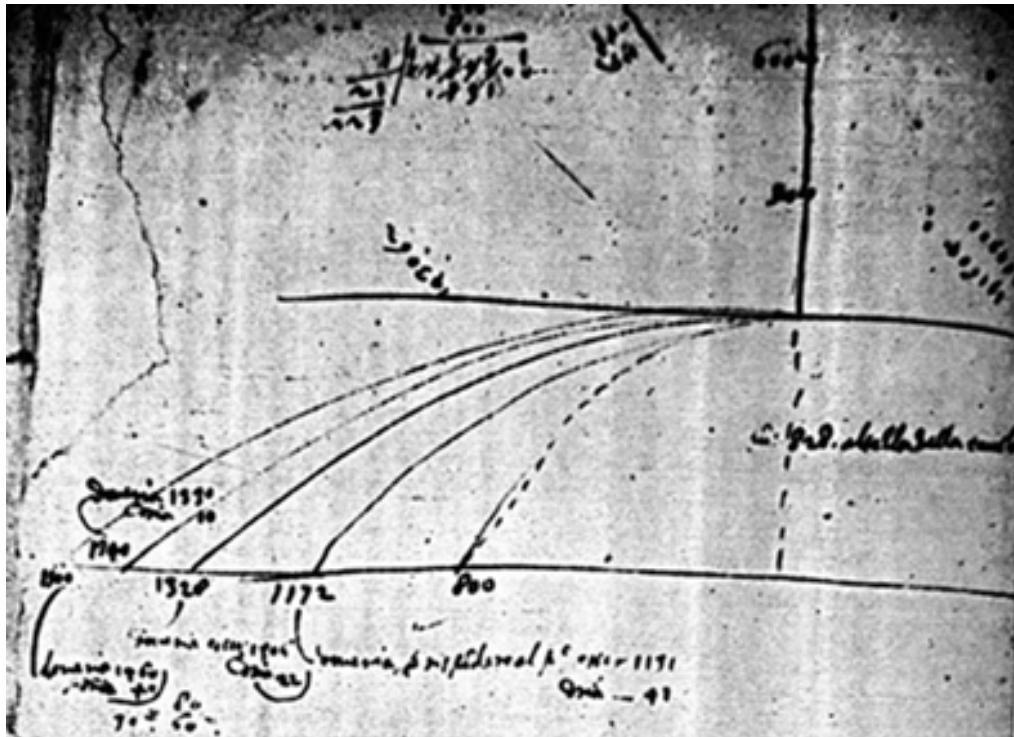
Lecture 14: **Interpretable ML: Physics-based machine learning**  $f = f_{\text{physics}} + \Delta f$   
**System Equation Modeling**

Lecture 15: Conclusions

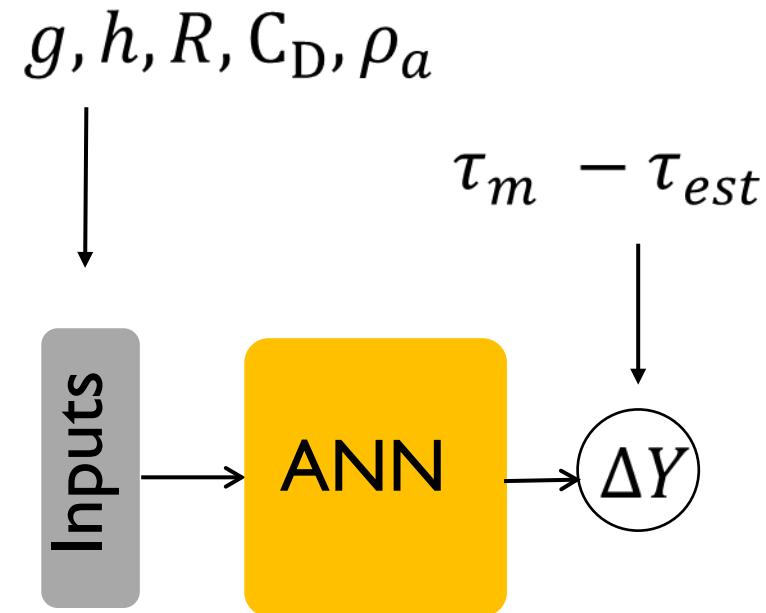
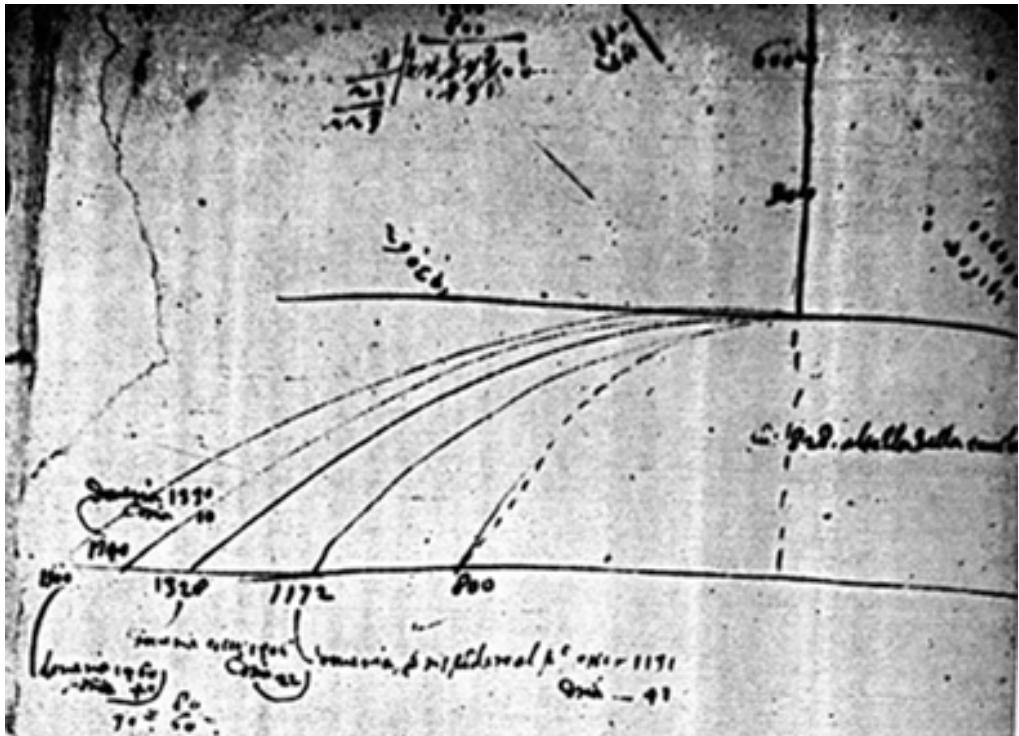
# Outline

- I. Why and what of physics-based machine learning
  2. Example I: Dropping a ball in the real world
  3. Example 2: Lake temperature distribution
  4. Approach 2: Structural Equation Modeling
- 
2. Conclusions

# Galileo experiments



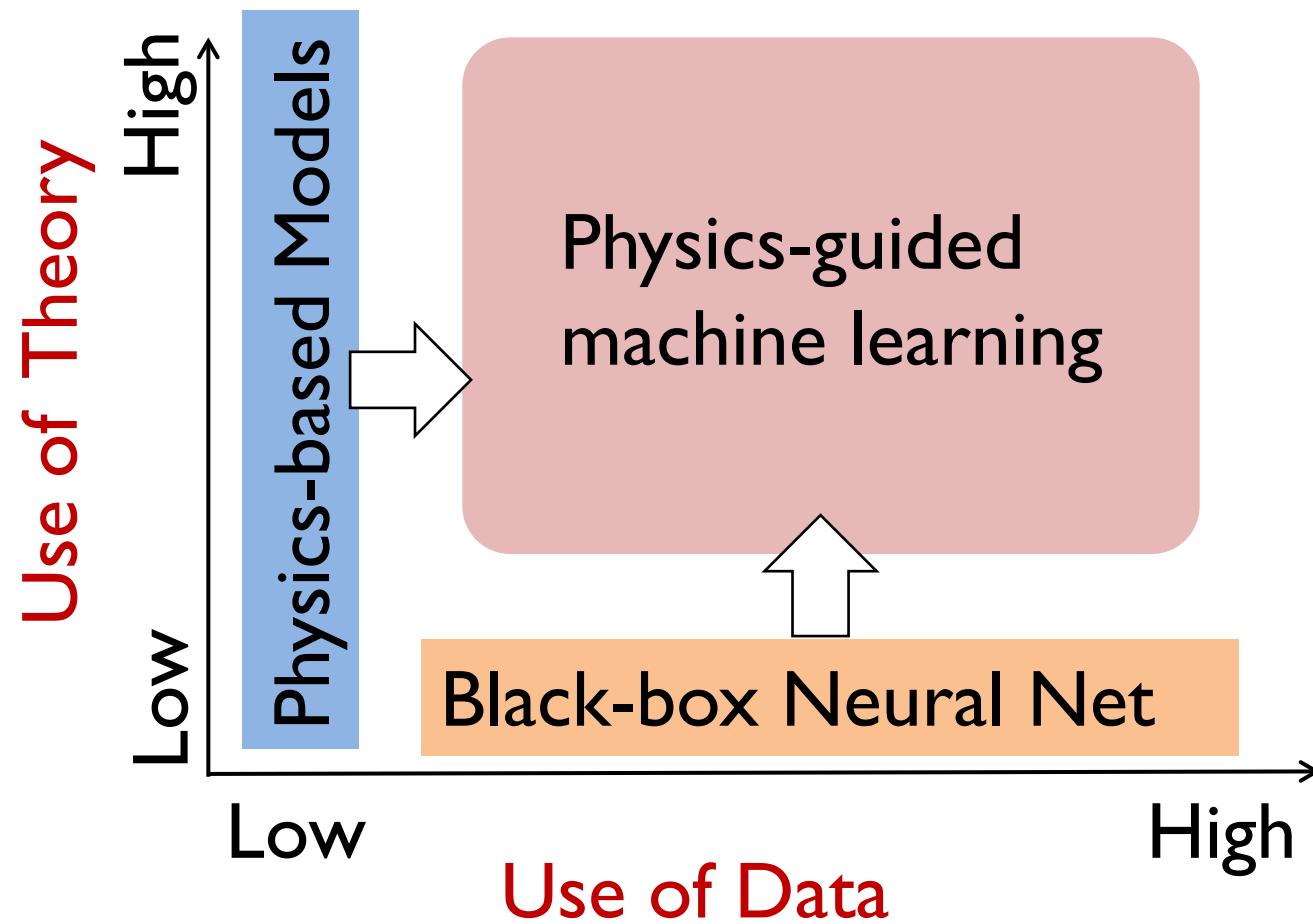
# Galileo vs. Newton



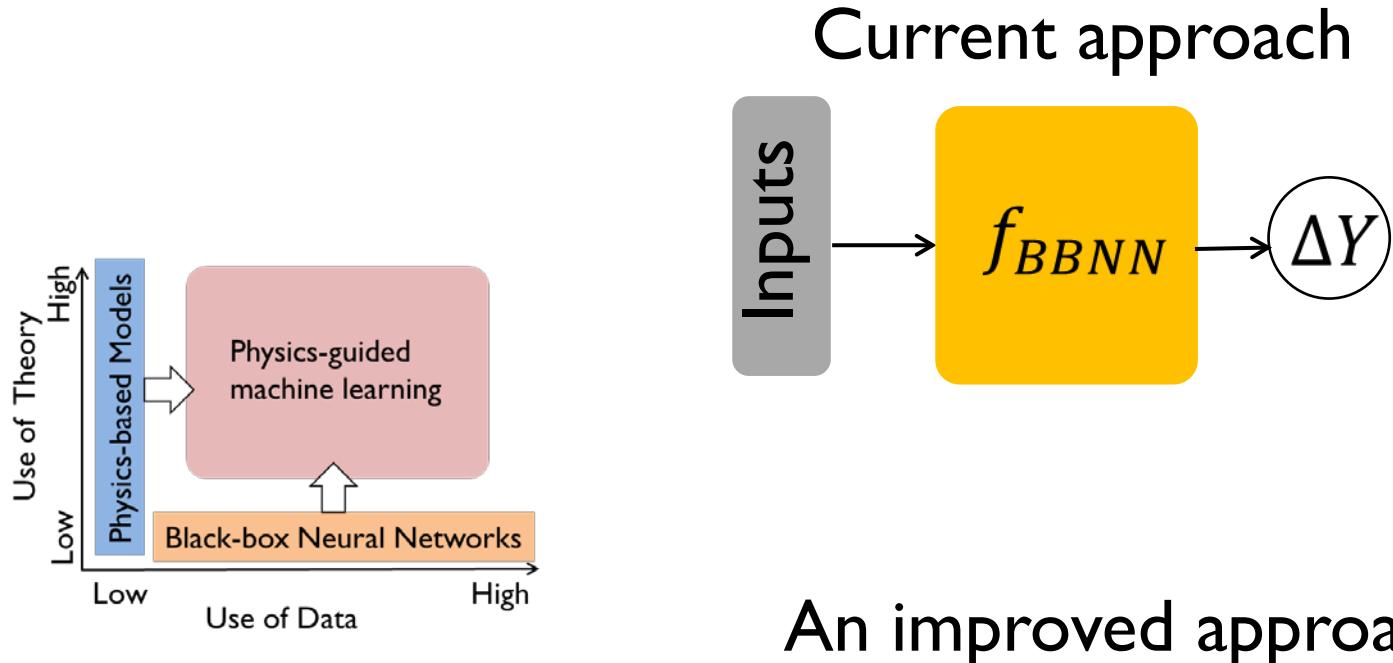
$$\frac{d^2 z}{dt^2} = -\frac{g}{m}$$

$$\tau_0 = \sqrt{\frac{2h}{g}}$$

# Physics-based machine learning approach



# Physics-based machine learning approach



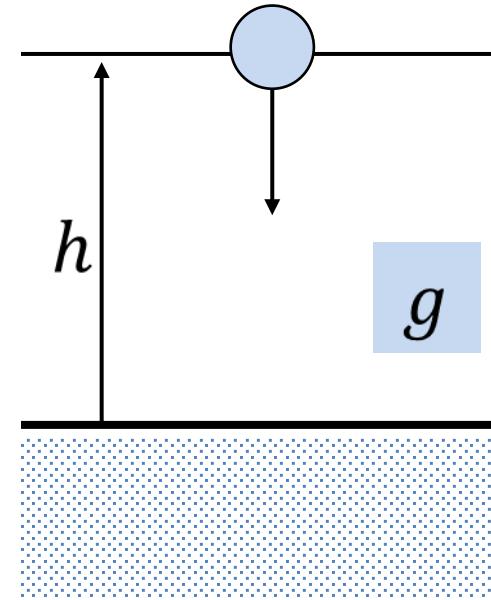
# Outline

- I. Why and what of physics-based machine learning
2. Example 1: Dropping a ball in the real world
3. Example 2: Lake temperature distribution
4. Example 3:
2. Conclusions

# A ball falling under gravity (idealized)

$$\begin{aligned}\frac{d^2z}{dt^2} &= -g \Rightarrow z = h - \frac{gt^2}{2} \\ \Rightarrow h &= \frac{g\tau_0^2}{2}\end{aligned}$$

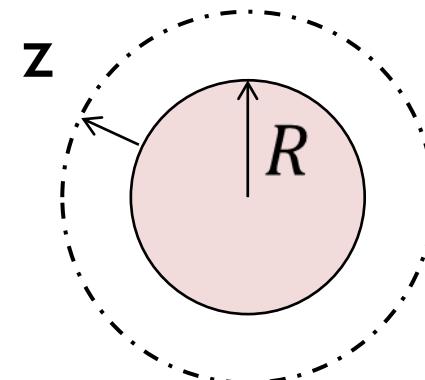
$$\tau_0 = \sqrt{\frac{2h}{g}}$$



The actual ball will experience height dependent gravity, air resistance that depends on temperature and humidity, etc.

# A falling ball with height dependent gravity (idealized)

$$\frac{d^2 z}{dt^2} = -\frac{g}{(1 + z/R)^2}$$



$$\tau_1 = \tau_0 \cosh^{-1}(1 - 2x) \quad x = z/R$$

$$\tau_1 \sim \tau_0 \left( 1 + \frac{5}{6} x + \frac{43}{40} x^2 + \frac{177}{112} x^3 + \dots \right)$$

C. F. Bohren, “Dimensional analysis, falling bodies,  
and the fine art of not solving differential equations,” 2003.

# A falling ball with air resistance

$$m \frac{dv}{dt} = -mg + \frac{1}{2} \rho_a (\pi r^2) C_D v^2 \rightarrow \frac{dv}{dt} = -g + bv^2$$

$\rho_a$  ... atmospheric density

$C_D$  ... drag coefficient

$R$  .... Radius of the ball

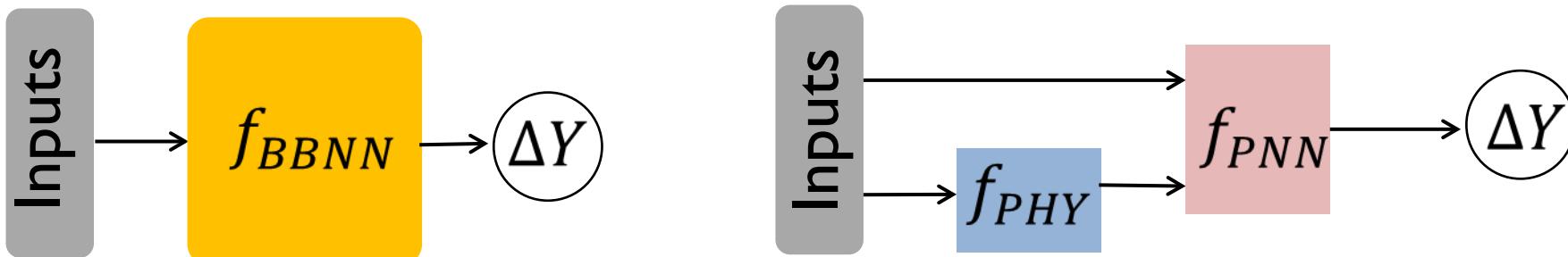
$$\cosh(t\sqrt{gb}) = \exp(hb) \quad z = h - \frac{\ln \cosh(t\sqrt{gb})}{b}$$

$$v = \frac{dz}{dt} = \sqrt{\frac{g}{b}} \tanh(t\sqrt{gb})$$

# A falling ball with gravity and resistance

$$\frac{dv}{dt} = -\frac{g}{(1 + z/R)^2} + bv^2$$

A numerical/perturbative solution may not be possible. In practice,  $b(z)$  is unknown function of humidity, temperature, etc. A machine learning approach is preferred.

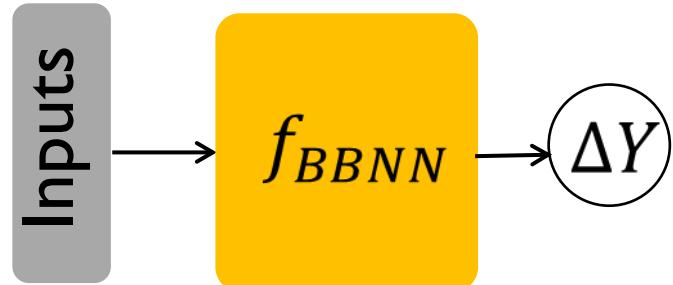


# Outline

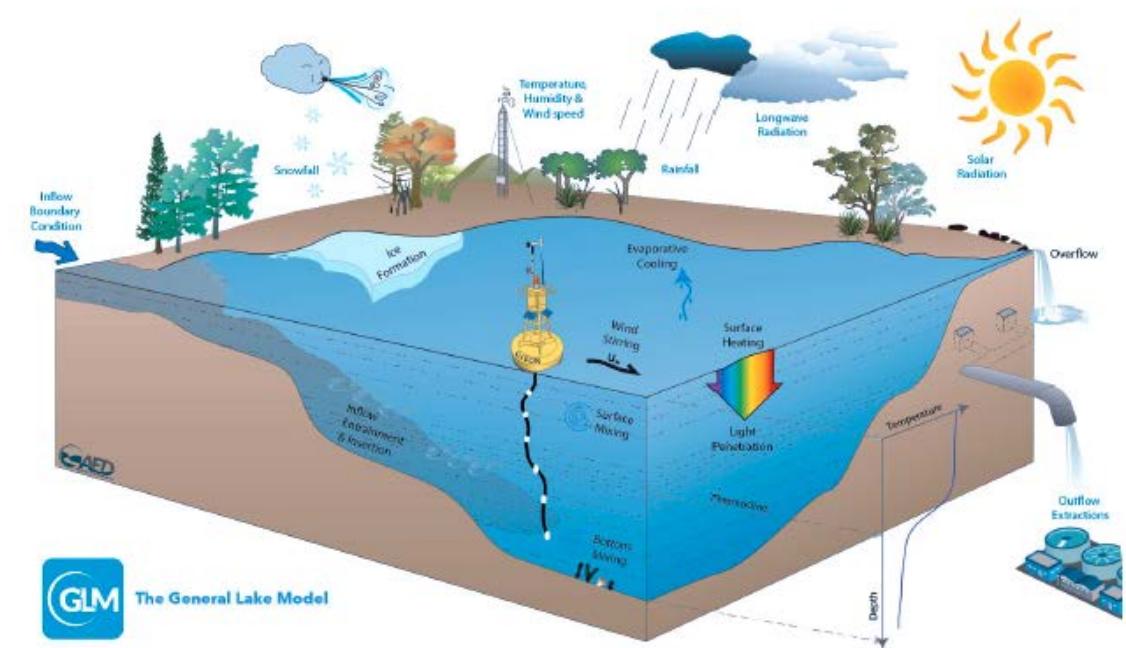
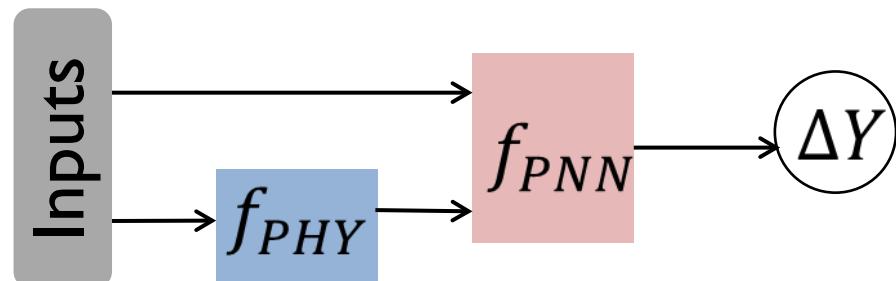
- I. Why and what of physics-based machine learning
  2. Example I: Dropping a ball in the real world
  3. Example 2: Lake temperature distribution
  4. Approach 2: Structural Equation Modeling
- 
2. Conclusions

# Physics-based machine learning approach

Current approach



An improved approach



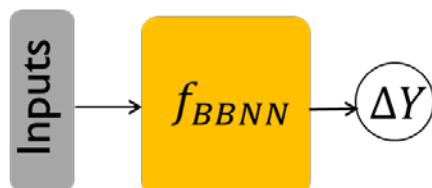
The General Lake Model

~ 10,000 T-measurement in two lakes

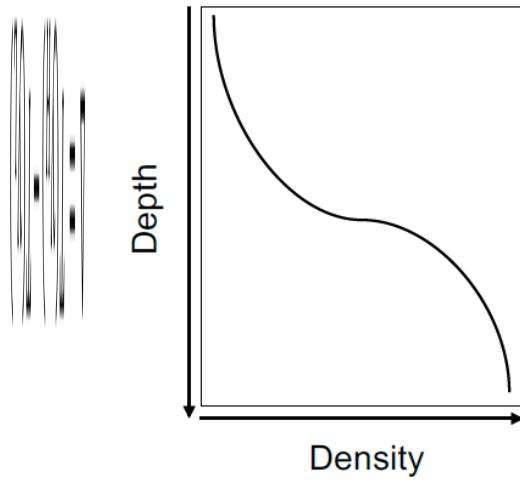
Physics-guided Neural Networks (PGNN): An application in Lake temperature Modeling Anuj Karpatne, 2016.

# An example of lake temperature: The master variable

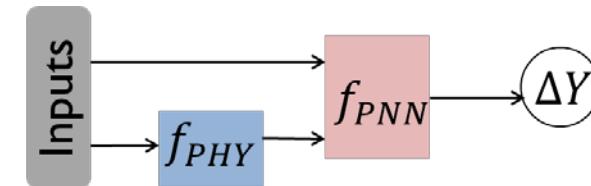
	Input Drivers
1	Day of Year (1 – 366)
2	Depth (in m)
3	Short-wave Radiation (in W/m <sup>2</sup> )
4	Long-wave Radiation (in W/m <sup>2</sup> )
5	Air Temperature (in °C)
6	Relative Humidity (0 – 100 %)
7	Wind Speed (in m/s)
8	Rain (in cm)
9	Growing Degree Days [14]
10	Is Freezing (True or False)
11	Is Snowing (True or False)



$$\arg \min_f Loss(\hat{Y}, Y) + \lambda R(f),$$



$$\frac{1 - \rho}{1000} = \frac{(Y + 289)(Y - 3.98)^2}{508929(Y + 68.13)}$$



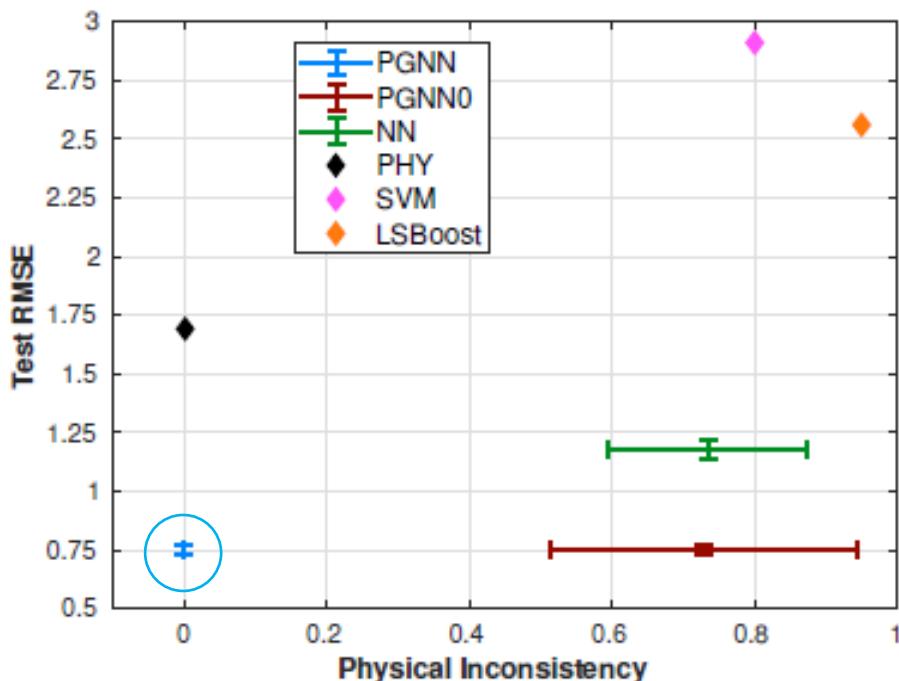
Lake temperature model  
with flux and energy model

$$\arg \min_f \underbrace{Loss(\hat{Y}, Y)}_{\text{Empirical Error}} + \underbrace{\lambda R(f)}_{\text{Structural Error}} + \underbrace{\lambda_{PHY} Loss.PHY(\hat{Y})}_{\text{Physical Inconsistency}},$$

# A example involving lake temperature

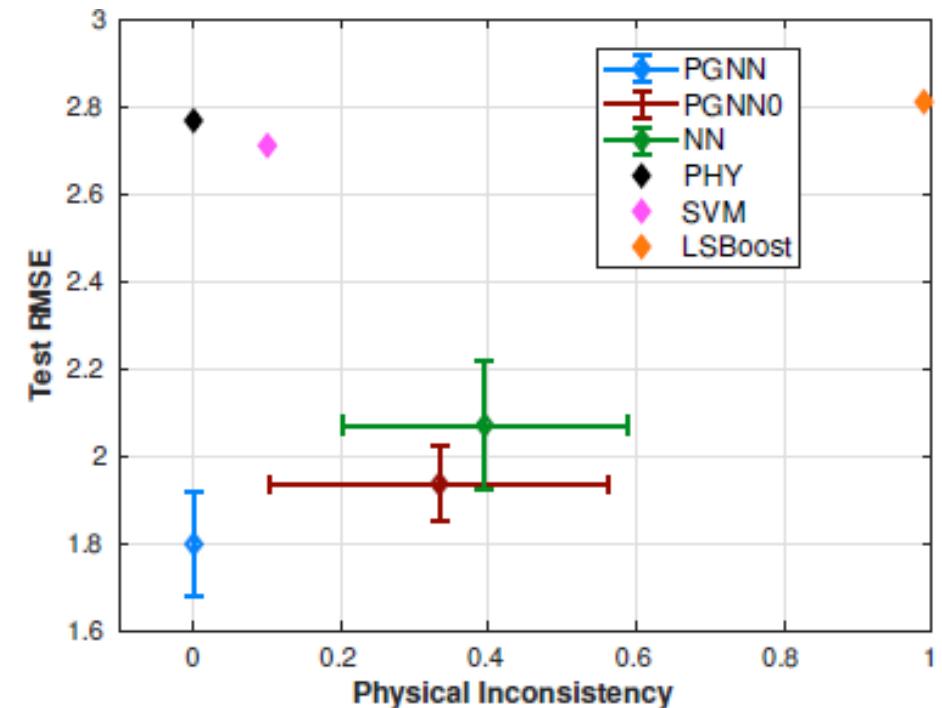
NN ... Neural network  
SVM ... Support vector machine  
RBF ... Radial basis function  
LSBoost ... Least square boosted tree

Lake 1



(a) Results on Lake Mille Lacs

Lake 2

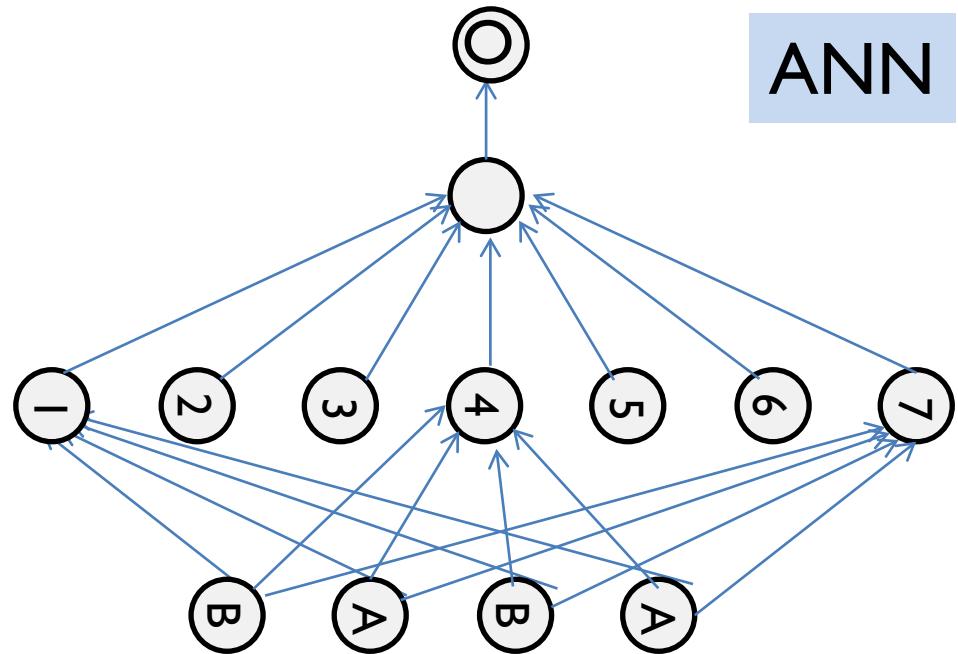


(b) Results on Lake Mendota

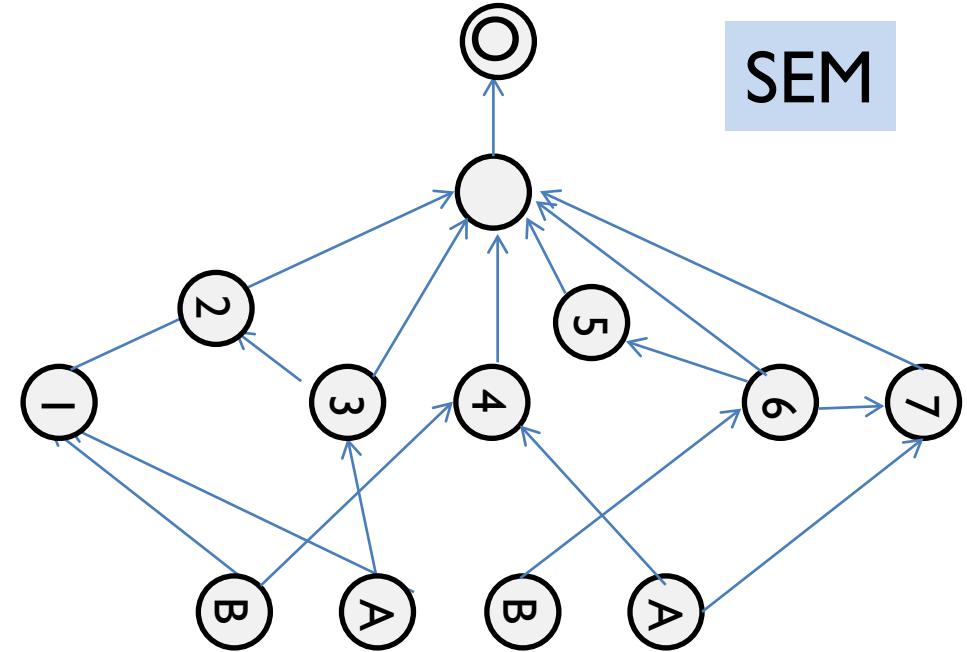
# Outline

- I. Why and what of physics-based machine learning
2. Example 1: Dropping a ball in the real world
3. Example 2: Lake temperature distribution
4. Approach 2: Structural Equation Modeling
2. Conclusions

# Structural Equation modeling: Motivation



ANN

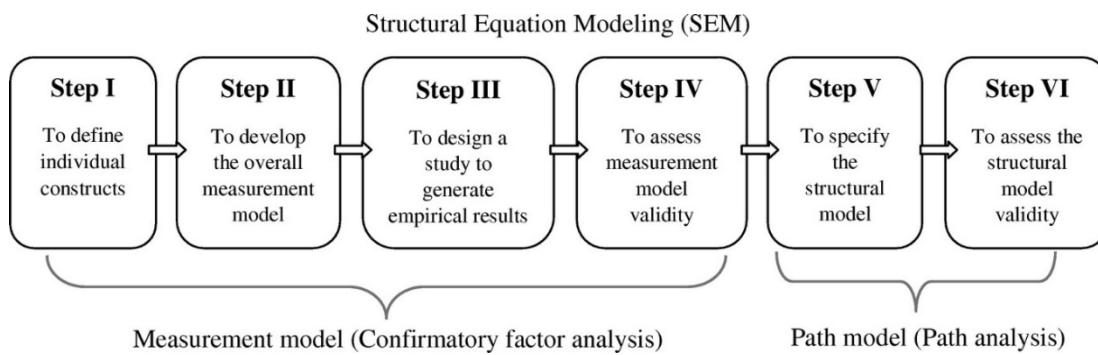


SEM

Nodes defined statistically  
not appropriate for extrapolation

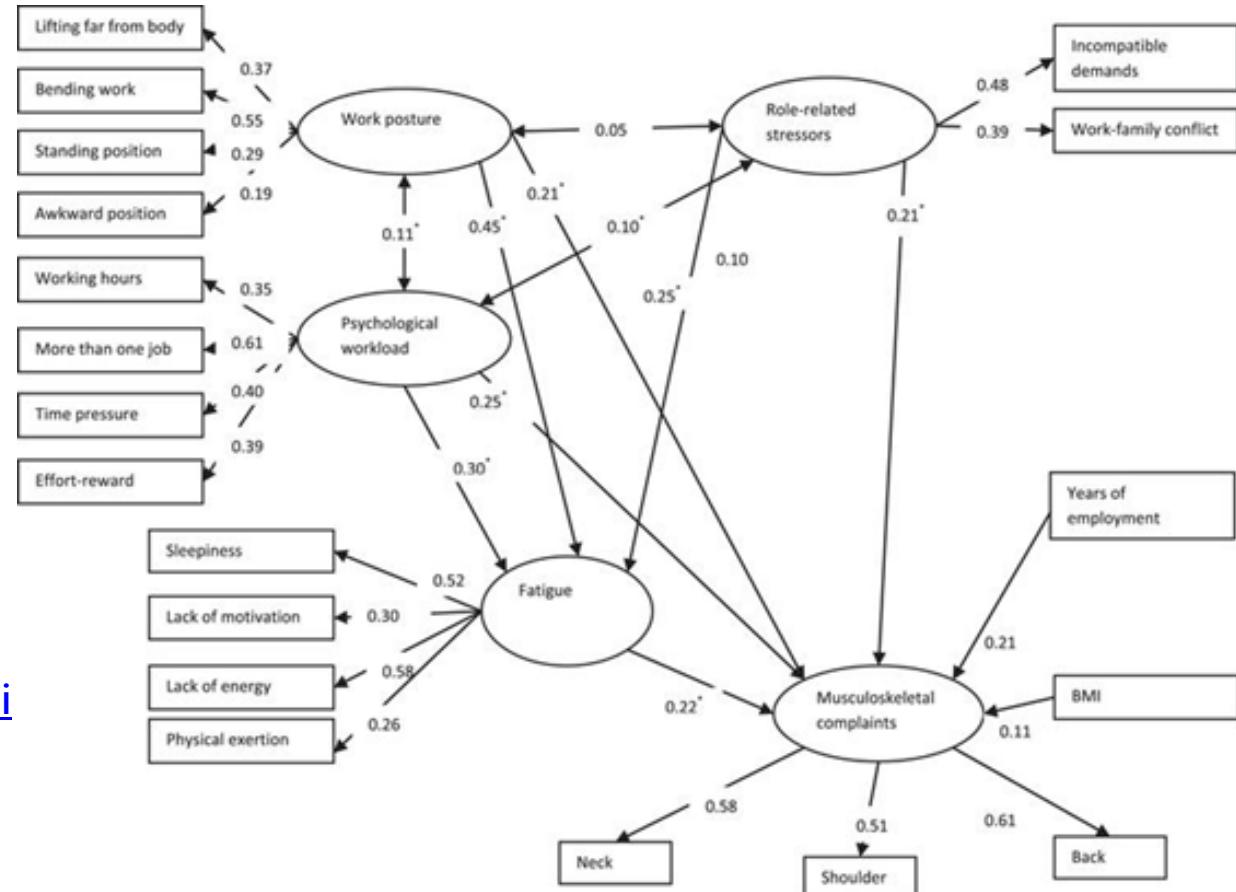
Nodes defined physically  
Interconnects are nonlinear  
Extrapolation possible

# Structural Equation modeling: Examples



Structural equation model of interactions between risk factors and work-related musculoskeletal complaints among Iranian hospital nurses Article type: Research Article  
Authors: [Mehralizadeh, Semira<sup>a</sup>](#) | [Dehdashti, Alireza<sup>b,\\*</sup>](#) | [Motalebi Kashani, Masoud<sup>c</sup>](#)

## Hidden variables



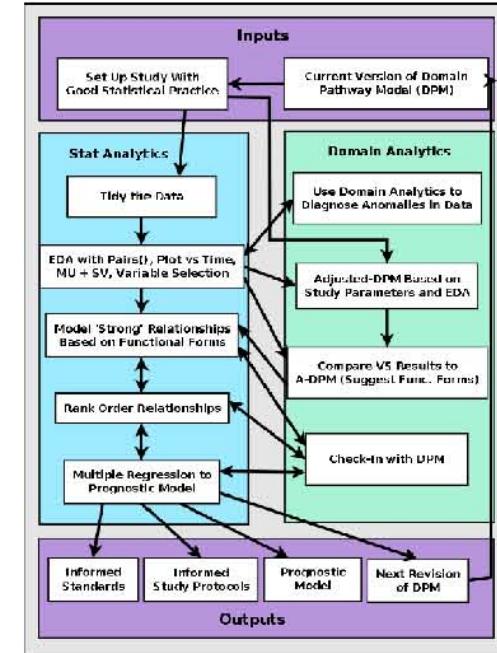
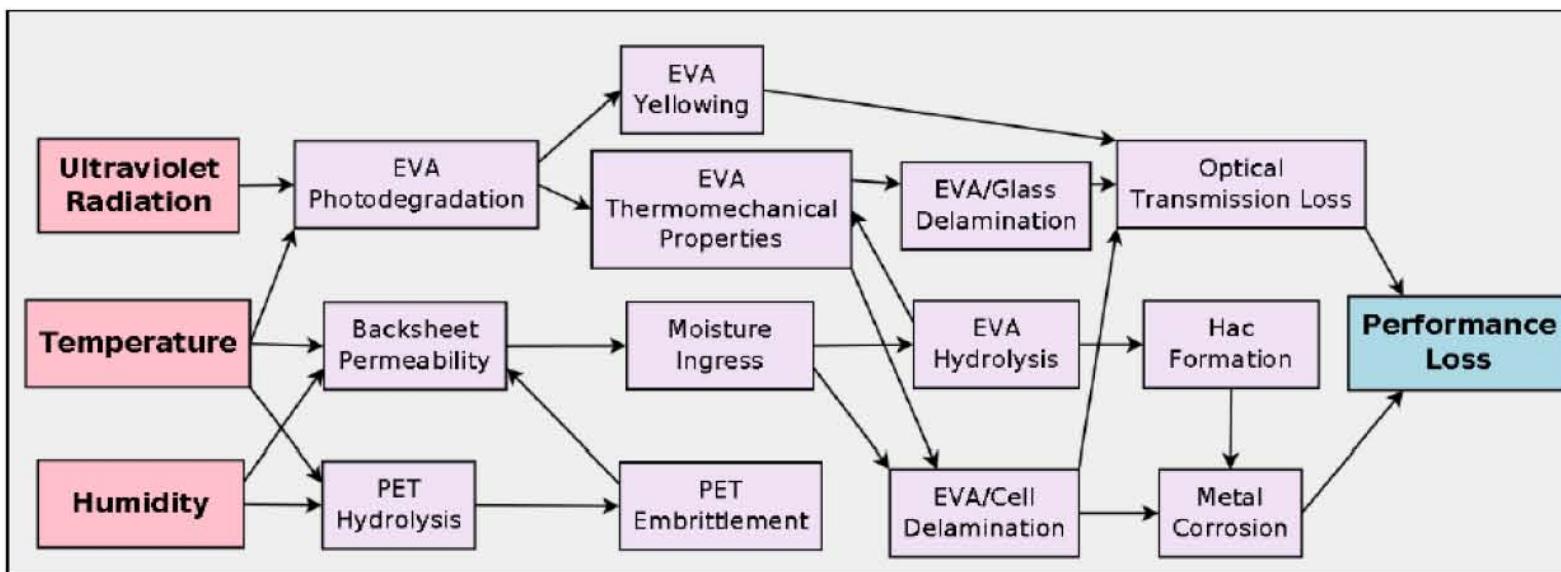
# Example: Degradation of Solar Cells

Received April 16, 2013, accepted April 23, 2013, date of publication June 10, 2013, date of current version June 25, 2013.

Digital Object Identifier 10.1109/ACCESS.2013.2267611

## Statistical and Domain Analytics Applied to PV Module Lifetime and Degradation Science

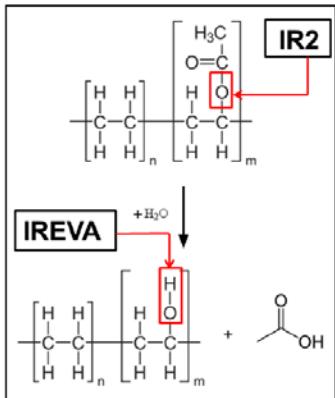
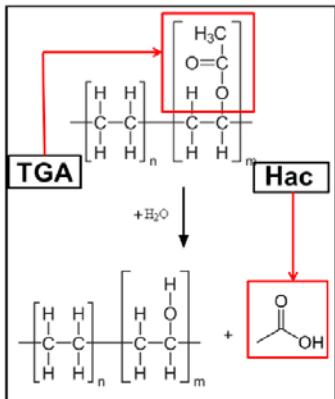
LAURA S. BRUCKMAN<sup>1</sup>, NICHOLAS R. WHEELER<sup>2</sup>, JUNHENG MA<sup>3</sup>, ETHAN WANG<sup>4</sup>,  
CARL K. WANG<sup>4</sup>, IVAN CHOU<sup>5</sup>, JIAYANG SUN<sup>3</sup>, AND ROGER H. FRENCH<sup>6</sup> (Member, IEEE)



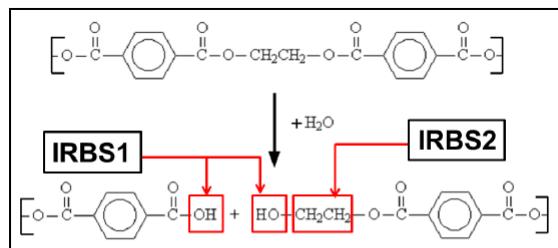
Variable Fits	Functional Form
Simple Linear (A)	$y = a + bx + \varepsilon$
Quadratic (B)	$y = a + bx + c * x^2 + \varepsilon$
Simple Quadratic (C)	$y = a + c * x^2 + \varepsilon$
Exponential (D)	$y = a + d * \exp^x + \varepsilon$
Logarithmic (E)	$y = a + f * (\log(x)) + \varepsilon$
Linear Change Point (F)	$a + d * (1 \pm \exp(g(x - h))) + \varepsilon$
Nonlinearizable Exponential (G-up, H-down)	$y = a + b * x + b_1 * (x - c)_+ + \varepsilon$

# PV Degradation (continued)

## Physical phenomenon of IRBS1, IRBS2, Hac, IR2,



**FIGURE 2.** Encapsulant degradation of EVA hydrolysis. (a): The mass decrease [48] of the sample was monitored during the heating process in dependence of the temperature. The decomposition of EVA and the appearance of acetic acid can be detected. Hydrolysis of vinyl-acetate monomers in EVA results in the generation of acetic acid [49]. (b): As the exposure time increases, the acetate C=O ( $1735\text{ cm}^{-1}$ ) peak decreases continuously, whereas the aldehyde/ketone C=O ( $1716\text{ cm}^{-1}$ ) and O-H (near  $3400\text{ cm}^{-1}$ ) peaks increase. This results from decomposition of vinyl acetate in the EVA and the formation of aldehydes, ketones and alcohols [50]. Ester ether C-O-C./overall CH absorbance ratios were calculated to measure the relative acetate content [47].

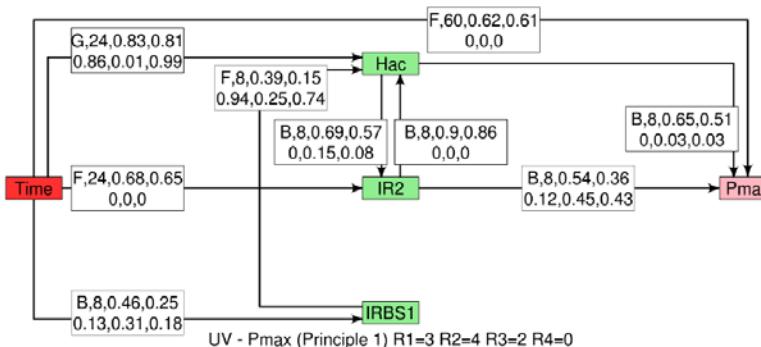


**FIGURE 3.** An IR peak at  $3373\text{ cm}^{-1}$  refers to a stretching vibration of hydroxyl groups which are related to hydrolysis [51]. The changes in the stretching vibration region of methylene group ( $\text{CH}_2$ ) at  $2927\text{ cm}^{-1}$  are attributed to chain scission due to hydrolysis [52].

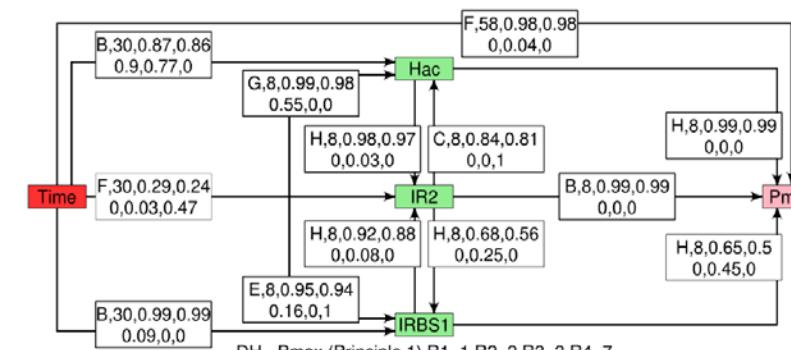
Variable Fits	Functional Form
Simple Linear (A)	$y = a + bx + \varepsilon$
Quadratic (B)	$y = a + bx + c * x^2 + \varepsilon$
Simple Quadratic (C)	$y = a + c * x^2 + \varepsilon$
Exponential (D)	$y = a + d * \exp^x + \varepsilon$
Logarithmic (E)	$y = a + f * (\log(x)) + \varepsilon$
Linear Change Point (F)	$a + d * (1 \pm \exp(g(x - h))) + \varepsilon$
Nonlinearizable Exponential (G-up, H-down)	$y = a + b * x + b_1 * (x - c)_+ + \varepsilon$

Muhammad A. Alam, Purdue University

UV radiation

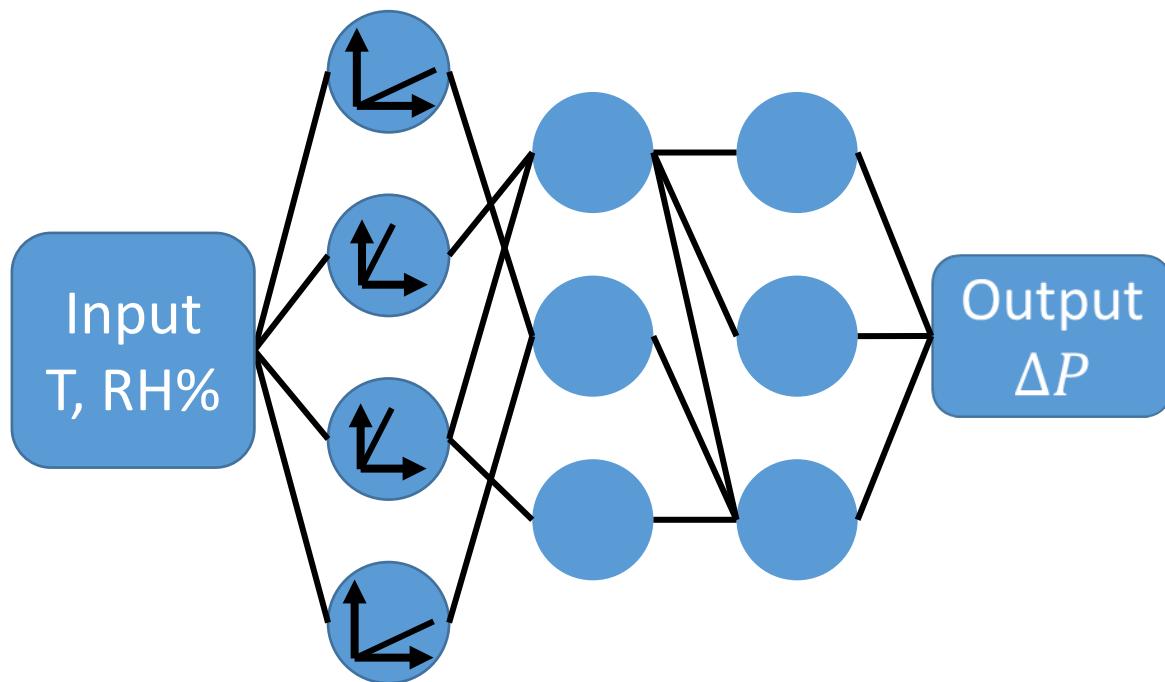


**FIGURE 8.** Model generated with the most relationships shown for Principle 1 for the UV exposure for the system response of  $P_{\text{max}}$  using the unit variables of  $\text{Hac}$ ,  $\text{IR2}$  and  $\text{IRBS1}$  shows 10 relationships. Information on each relationship is described in the box. The information contained is functional form, number of observations,  $R^2$ , adjusted  $R^2$ , P-value 1, P-value 2 and P-value 3, respectively. The strength of the SSR is summarized by the line width of the SSR border based on the  $R^2$  value to aide visualization (below 0.2 not shown, R1 has the thinnest border (0.2-0.5), R2 (0.5-0.7), R3 (0.7-0.9) and R4 the thickest ( $\geq 0.9$ )). The functional forms are designated as A (simple linear), B (quadratic), C (simple quadratic), D (exponential), E (logarithmic), F (linear change point), G (nonlinearizable exponential-up) and H (nonlinearizable exponential-down).

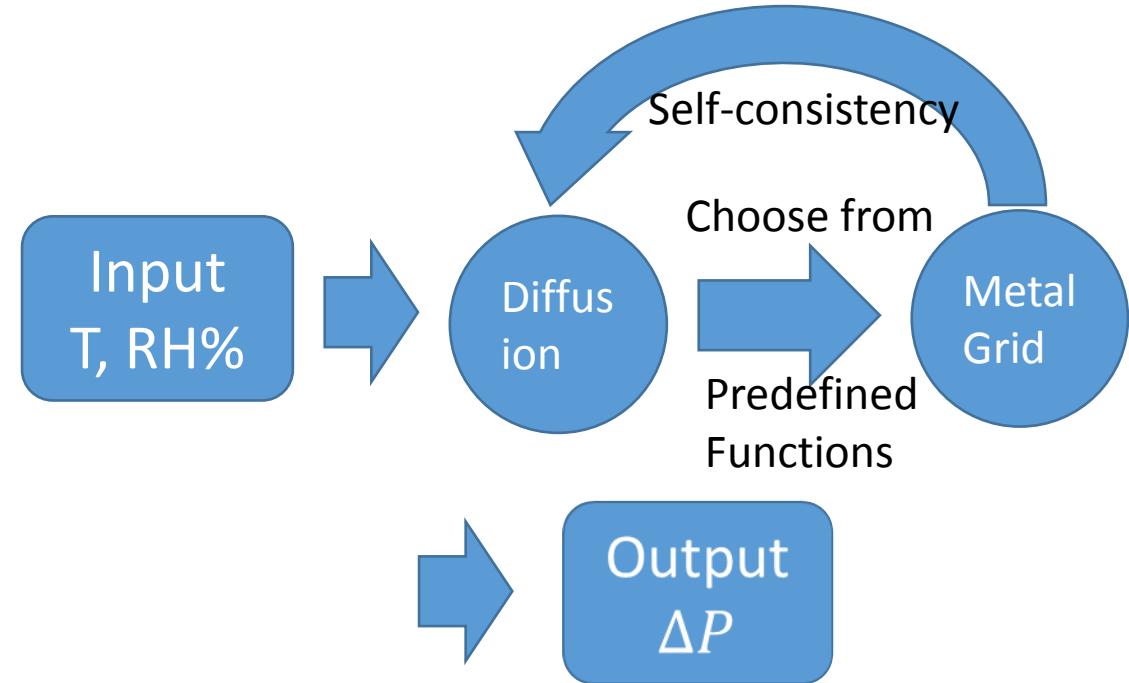


**FIGURE 7.** Model generated with the most relationships shown for Principle 1 for the damp heat exposure for the system response of  $P_{\text{max}}$  using the unit variables of  $\text{Hac}$ ,  $\text{IR2}$  and  $\text{IRBS1}$  shows 13 relationships. Information on each relationship is described in the box. The information contained is functional form, number of observations,  $R^2$ , adjusted  $R^2$ , P-value 1, P-value 2 and P-value 3, respectively. The strength of the SSR is summarized by the line width of the SSR border based on the  $R^2$  value to aide visualization (below 0.2 not shown, R1 has the thinnest border (0.2-0.5), R2 (0.5-0.7), R3 (0.7-0.9) and R4 the thickest ( $\geq 0.9$ )). The functional forms are designated as A (simple linear), B (quadratic), C (simple quadratic), D (exponential), E (logarithmic), F (linear change point), G (nonlinearizable exponential-up) and H (nonlinearizable exponential-down).

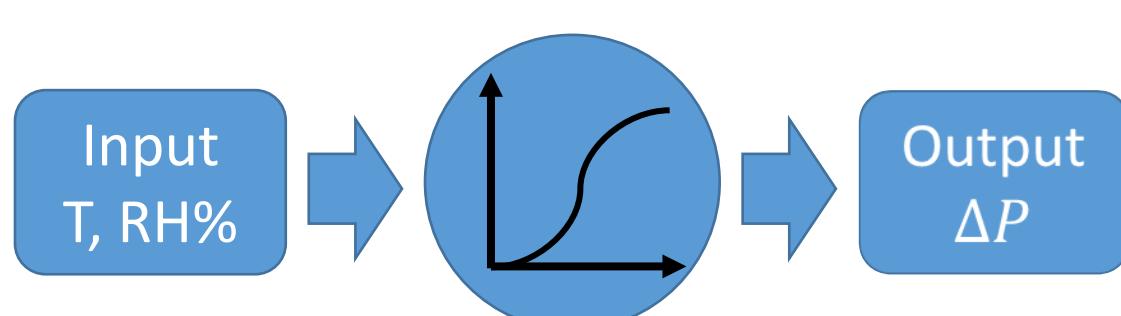
## Machine Learning (Linear connection)



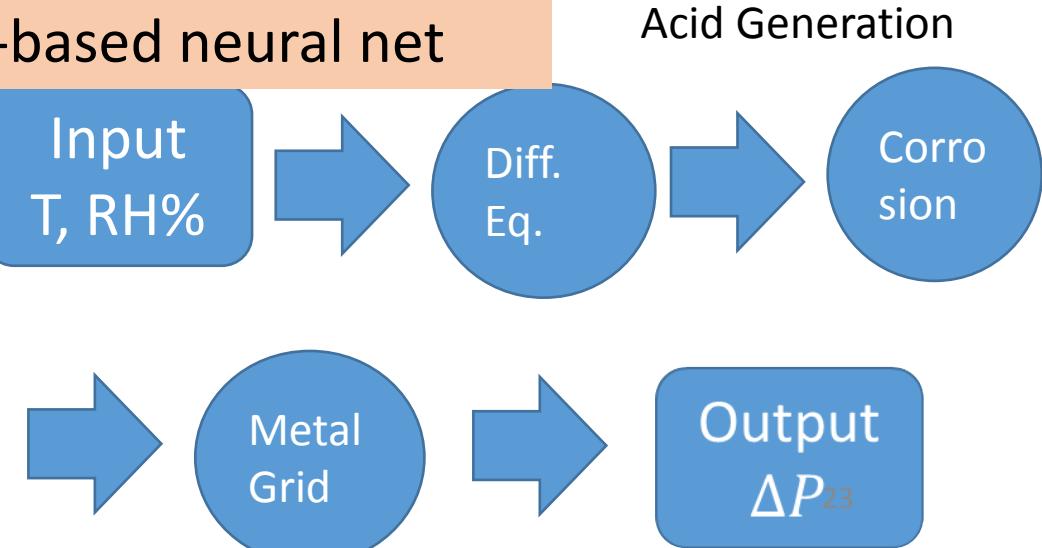
## Physics-based Structural Equation Modeling



## Machine Learning (Non-linear connection)



## Physics-based neural net



# Conclusions

1. Despite its demonstrated utility in variety of problems in commerce and business, the statistical nature of Machine Learning makes it difficult to use in engineering design or predictive modeling problems.
2. Physics-based approaches, such as physics-guided neural network or structural equation modeling offers opportunities to embed physical understanding of the model.
3. Some of the models involve non-linear coefficients – the computational efficiency of the model for very large systems are not fully understood.
4. One may be able to move from purely statistical model (e.g. Ptolemy) to physics-inspired model (e.g. Newton) one the physical understanding of the latent variable emerge over time.

# Review Questions

1. Statistical machine learning is not good for out-of-range extrapolation.  
Explain.
2. In what ways traditional machine learning similar to a “child learning a new language or Ptolemy’s mode of the solar system? In what sense are the comparisons accurate?
3. Deeper the learning, shallower is the understanding. Why?
4. Name two examples where physics-based machine learning could improve predictions.
5. Give one example where physics based machine learning can actually give wrong result.
6. What is the difference between Singular value decomposition and newer machine learning algorithms?

# References

## Physics based Machine Learning

C. F. Bohren, “Dimensional analysis, falling bodies, and the fine art of not solving differential equations,”

Galileo experiment on projectile motion:  
[http://galileo.rice.edu/lib/student\\_work/experiment95/paraintr.html](http://galileo.rice.edu/lib/student_work/experiment95/paraintr.html)

Physics-guided Neural Networks (PGNN): An application in Lake temperature Modeling Anuj Karpatne, 2016.

## Structured Equation Approach.

Professor Patrick Sturgis, NCRM director, in the first (of three) part of the Structural Equation Modeling NCRM online course. Structural Equation Modeling: what is it and what can we use it for? (part 1 of 6)  
<https://www.youtube.com/watch?v=eKkESdyMG9w>

<https://www.youtube.com/watch?v=-m4ag3WQcCw>

A talk about three principal advantages of structural equation models (SEMs) relative to more traditional analytic techniques, like the linear regression model. These include...

- (1) The ability to represent constructs as latent variables that are uncontaminated by measurement error
- (2) Falsification tests and indices of fit with to evaluate the tenability of a proposed theoretical model
- (3) Flexibility to allow researchers to specify statistical models that more closely match theory

Dan describes these advantages using a specific example on the factors that relate to young children's popularity with peers.

In addition to these three principal advantages of the SEM, there are many other ways that the model can be expanded and used to address interesting theoretical questions. For instance, a variety of SEMs exist for analyzing longitudinal data, including latent growth curve models and latent change score models. SEMs also provide a powerful framework within which to evaluate population heterogeneity, including differences over known groups (e.g., boys and girls) or latent groups (e.g., clusters of individuals for whom predictive relationships differ). For those interested in learning more, we offer summer training seminars on SEM and longitudinal SEM, see <http://www.curranbauer.org/training/>.

# An Introduction to Data Analysis, Design of Experiment, and Machine Learning

## *Lecture 15. Conclusions and Outlook*

Muhammad A. Alam  
[alam@purdue.edu](mailto:alam@purdue.edu)



# copyright 2018

This material is copyrighted by M. Alam under the following Creative Commons license:



**Attribution-NonCommercial-ShareAlike 2.5 Generic (CC BY-NC-SA 2.5)**

Conditions for using these materials is described at

<http://creativecommons.org/licenses/by-nc-sa/2.5/>

# Outline

- 1. Introduction**
- 2. Review of the lectures**
- 3. Conclusions**

# The topics covered ...

## Lectures 8-14

How to get a better  $f$

## Lectures 6-7

$\bar{x} = x_1, x_2, \dots x_n$  

$$\bar{y} = f(\bar{x})$$

## Lectures 1-2

$\bar{y} = y_1, y_2, \dots y_m$

## Lectures 3-5

How to fit multiple hypothetical function  $f$  to the same  $y$

# Course Outline

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

Lecture 1: Introduction

Lecture 2: Collecting and plotting  $x_1, x_2, \dots x_n$

Lecture 3: Physical and empirical  $f, F, df/dx, \dots$

Lecture 4: Model selection between  $f_1, f_2, \dots$

Lecture 5: Model Selection: Cross-validation and Bootstrapping method

Lecture 6: Scaling theory with known  $f$ ,  $f(\bar{x}) = f(\bar{X})$

Lecture 7: Scaling theory with unknown  $f$ ,  $\bar{x} \rightarrow X$

Lecture 8: Design of experiments to determine  $\bar{y}_{\max} = f(\bar{x})$

Lecture 9: DOE and ANOVA

Lecture 10: Principle component analysis for classifying  $\{y\}$ .

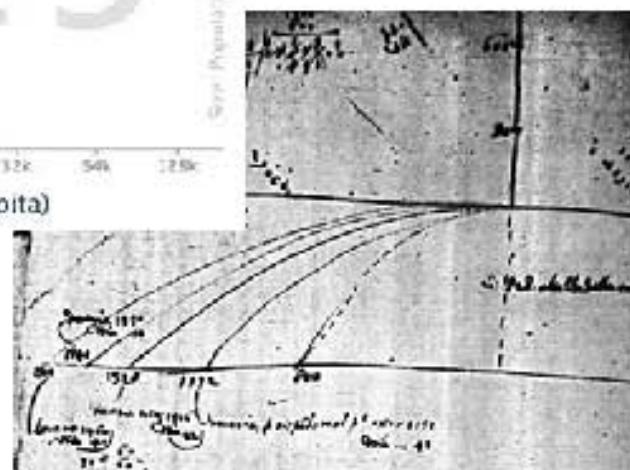
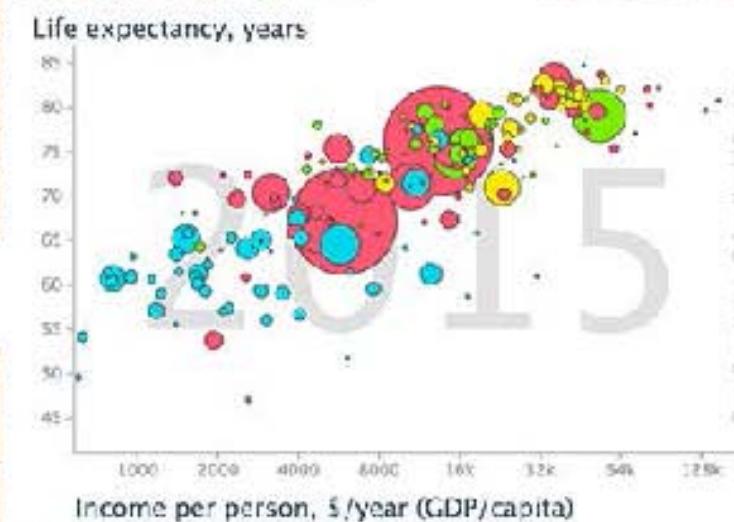
Lecture 11: Machine learning ... Statistical approach to learn  $f$

Lecture 12: Machine Learning .... Deep network, Karnaugh map, and other components

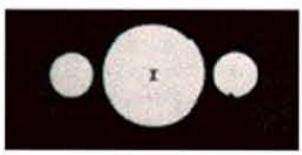
Lecture 13: Interpretable ML: Physics-based machine learning  $f = f_{\text{physics}} + \Delta f$

Lecture 14: Conclusions

# Lecture I: A short history of data



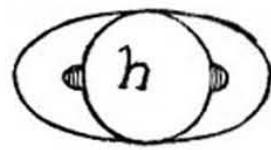
# Small vs. big data



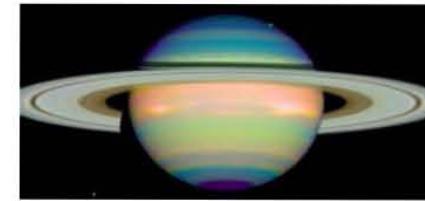
Galileo first sketch  
1610



Better telescope  
1616



Published etch  
1623



# Course Outline

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

Lecture 1: Introduction

Lecture 2: **Collecting and plotting  $x_1, x_2, \dots x_n$**

Lecture 3: Physical and empirical  $f, F, df/dx, \dots$

Lecture 4: Model selection between  $f_1, f_2, \dots$

Lecture 5: Model Selection: Cross-validation and Bootstrapping method

Lecture 6: Scaling theory with known  $f$ ,  $f(\bar{x}) = f(\bar{X})$

Lecture 7: Scaling theory with unknown  $f$ ,  $\bar{x} \rightarrow X$

Lecture 8: Design of experiments to determine  $\bar{y}_{\max} = f(\bar{x})$

Lecture 9: DOE and ANOVA

Lecture 10: Principle component analysis for classifying  $\{y\}$ .

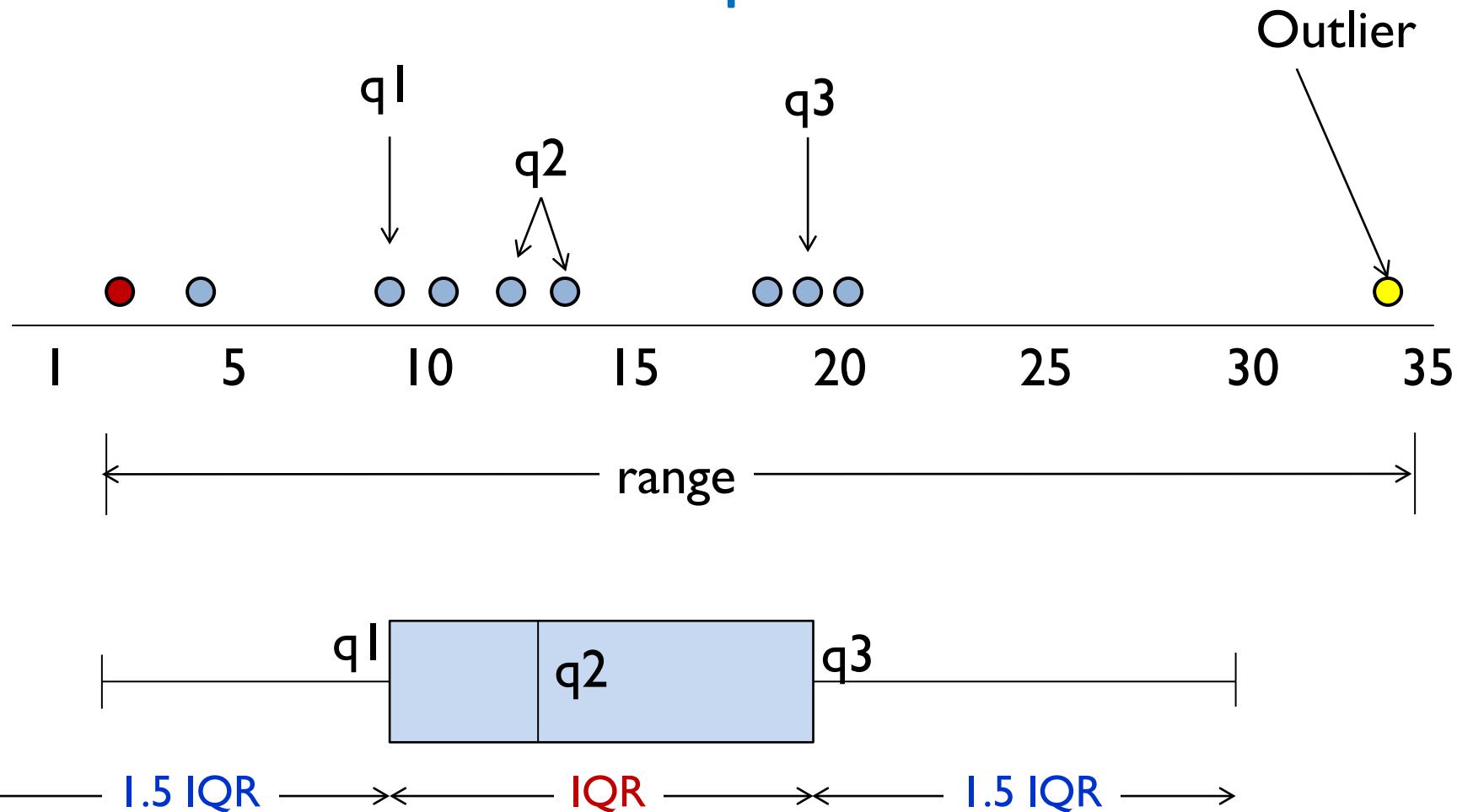
Lecture 11: Machine learning ... Statistical approach to learn  $f$

Lecture 12: Machine Learning .... Deep network, Karnaugh map, and other components

Lecture 13: Interpretable ML: Physics-based machine learning  $f = f_{\text{physics}} + \Delta f$

Lecture 14: Conclusions

# Box plot



# Stem and leaf display: Pre-histogram

Order data

44 46 47 49 63 64 66 68 68 72 72 75 76 81 84 88 106

n=17

4 | 4679 ← Leaf

5 |

6 | 34688

7 | 2256

8 | 148

9 |

10 | 6

$$L = [10 \times \log_{10} n] \sim 13$$

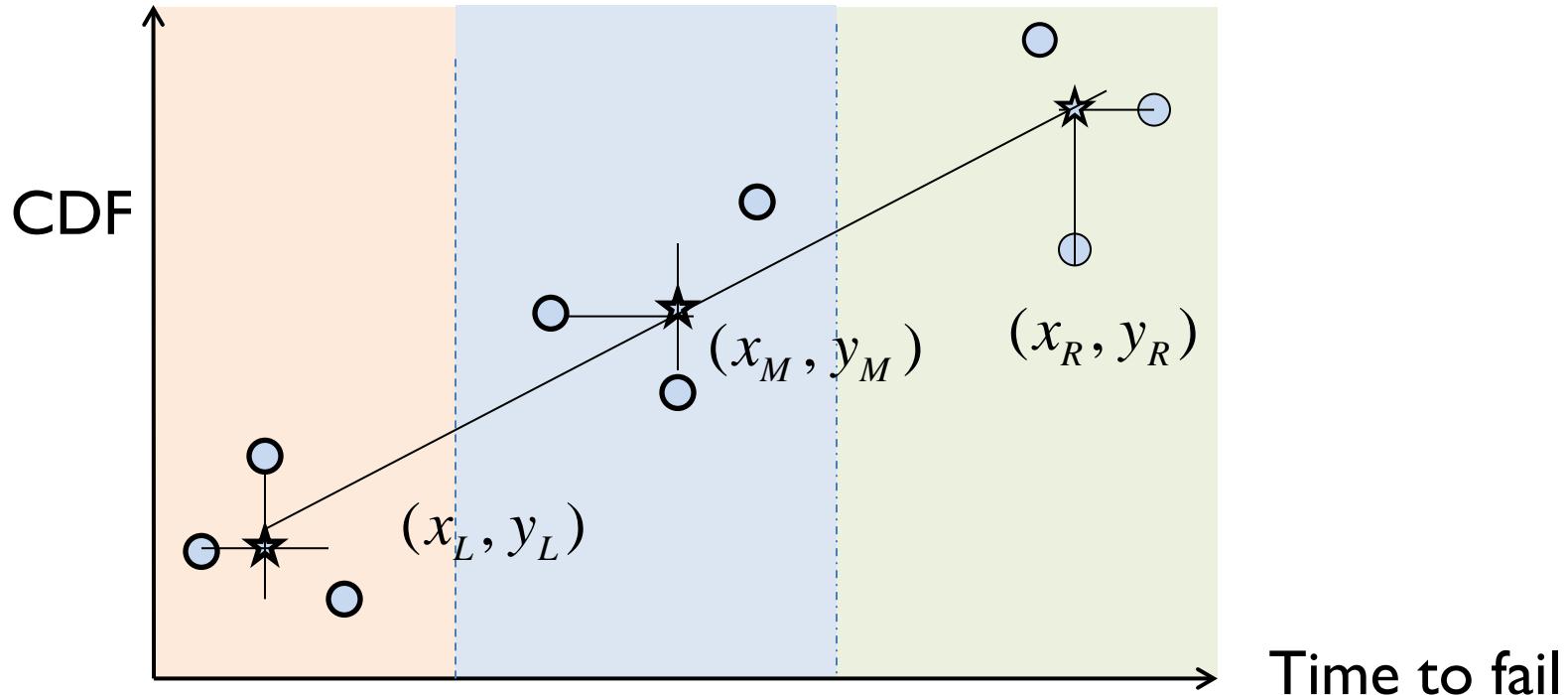
$h_n = (\text{Range}/L)$  to power of 10 (i.e. 4.77 → 10)

Therefore, 40, 50, 60 ... 90, 100 are stem values

Should use the same approach for histogram  
Histogram should not increase precision

↑  
stem

# Drawing lines resistant to outliers



$$y = b(x - x_M) + a$$

$$b_0 = (y_R - y_L) / (x_R - x_L)$$

$$3a_0 = [y_L - b_0(x_L - x_M)] + y_M + [y_R - b_0(x_R - x_M)]$$

$$r_i = y_i - [a_0 + b_0(x_i - x_0)]$$

$$a_1 = a_0 + \gamma_1 \quad b_1 = b_0 + \delta_1$$

# For censored data

Assume that at time  $t_3$ , one sample is taken out of the experiments (censored)

$$F_1 = 1 - \frac{4+1}{4+2} = \frac{1}{6}$$

$$F_2 = 1 - \frac{4+1}{4+2} \cdot \frac{3+1}{3+2} = \frac{2}{6}$$

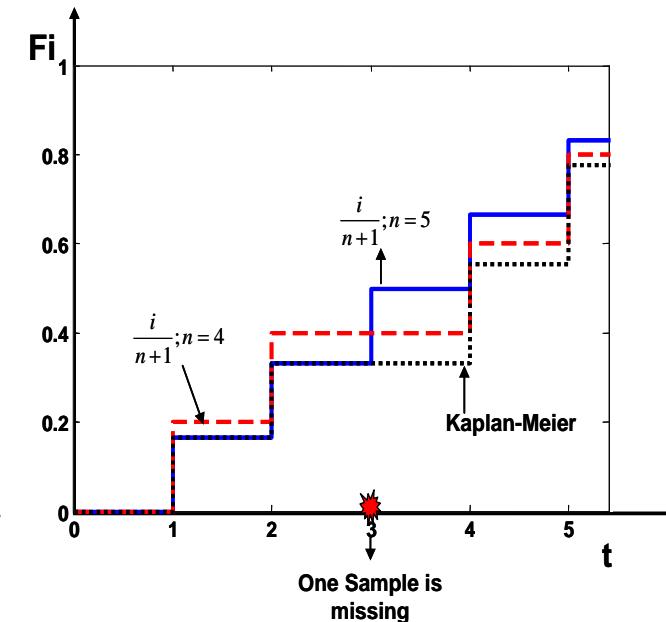
$$F_3 = 1 - \frac{4+1}{4+2} \cdot \frac{3+1}{3+2} = \frac{2}{6}$$

$$F_4 = 1 - \frac{4+1}{4+2} \cdot \frac{3+1}{3+2} \cdot \frac{1+1}{1+2} = 1 - \frac{5}{6} \cdot \frac{4}{5} \cdot \frac{2}{3} = \frac{5}{9}$$

$$F_5 = 1 - \frac{5}{6} \cdot \frac{4}{5} \cdot \frac{2}{3} \cdot \frac{1}{2} = \frac{7}{9}$$

← **3/4 missing ...**

$n_{si}$ before $t_i$	5	4	3	2	1
$n_{si}$ after $t_i$	4	3	2	1	0



# Course Outline

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

Lecture 1: Introduction

Lecture 2: Collecting and plotting  $x_1, x_2, \dots x_n$

Lecture 3: Physical and empirical  $f, F, df/dx, \dots$

Lecture 4: Model selection between  $f_1, f_2, \dots$

Lecture 5: Model Selection: Cross-validation and Bootstrapping method

Lecture 6: Scaling theory with known  $f$ ,  $f(\bar{x}) = f(\bar{X})$

Lecture 7: Scaling theory with unknown  $f$ ,  $\bar{x} \rightarrow X$

Lecture 8: Design of experiments to determine  $\bar{y}_{\max} = f(\bar{x})$

Lecture 9: DOE and ANOVA

Lecture 10: Principle component analysis for classifying  $\{y\}$ .

Lecture 11: Machine learning ... Statistical approach to learn  $f$

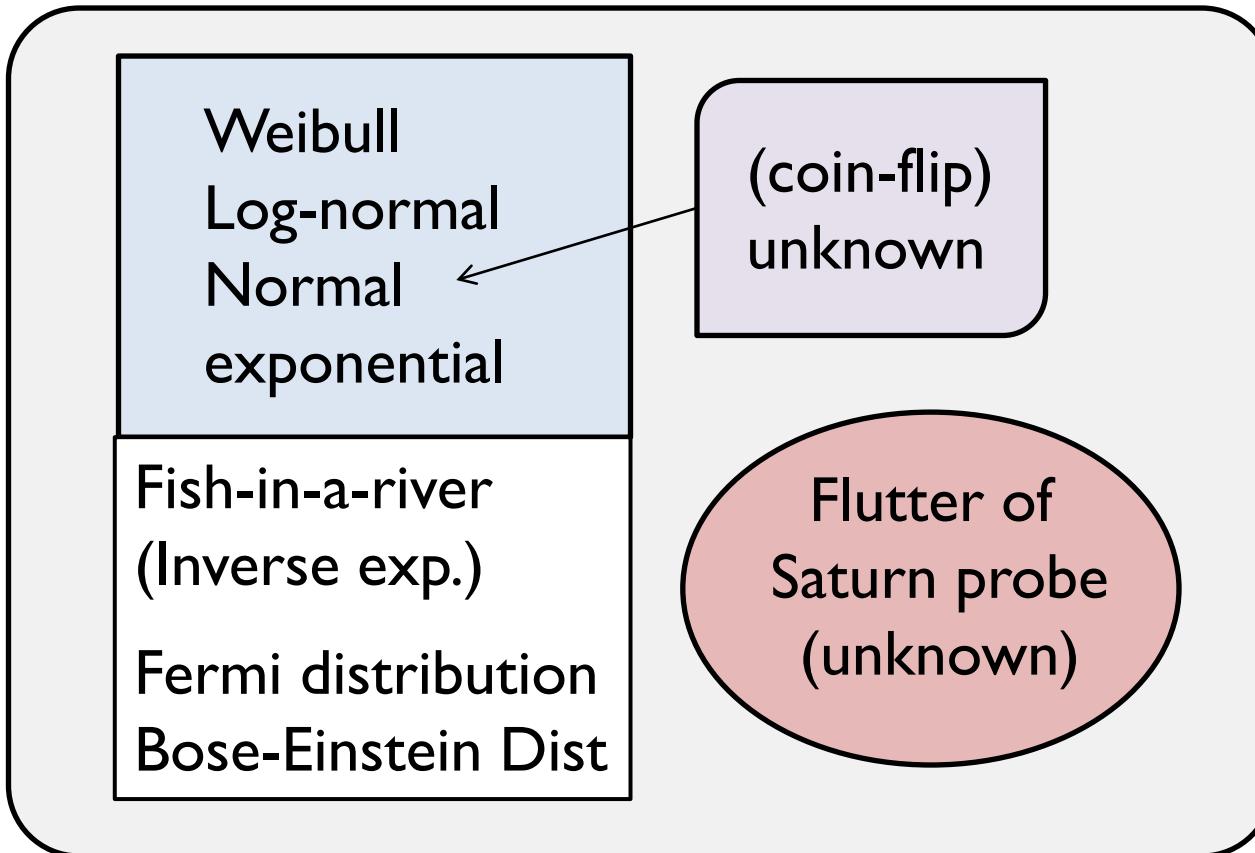
Lecture 12: Machine Learning .... Deep network, Karnaugh map, and other components

Lecture 13: Interpretable ML: Physics-based machine learning  $f = f_{\text{physics}} + \Delta f$

Lecture 14: Conclusions

# Statistical Distribution is Physical

## Experiments

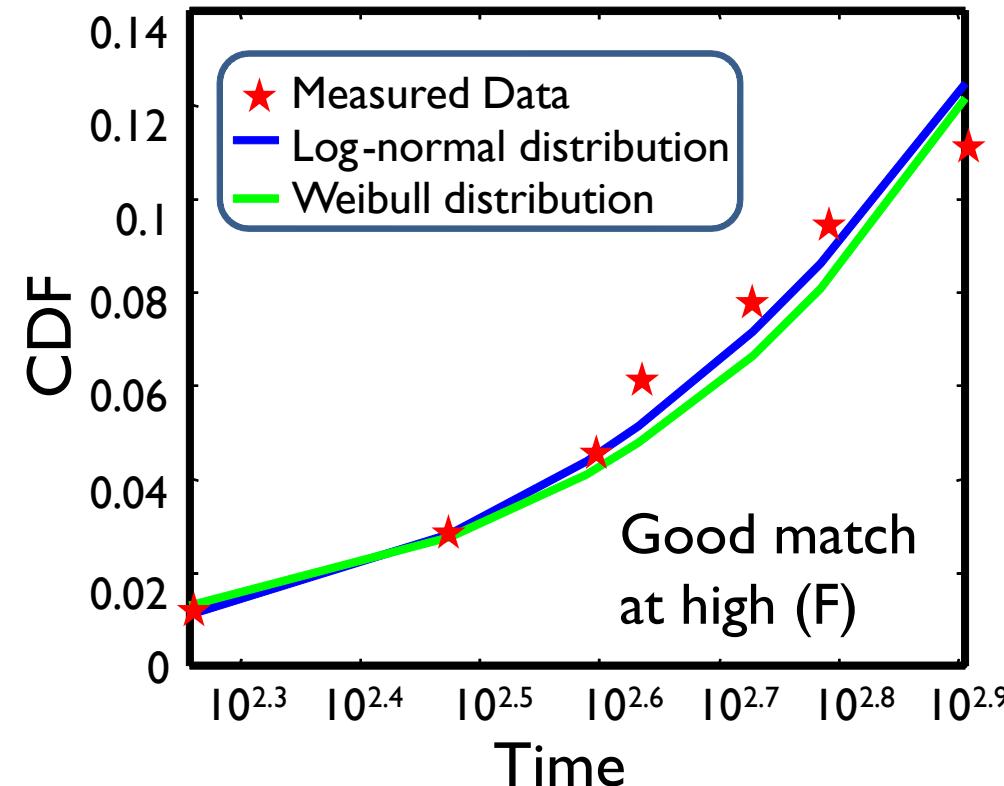


If a problem can be mapped into one of the well known family,  
large number of results are available.

# Matching moments to distributions

Of 60 oxides, 7 failed in 1000 hrs

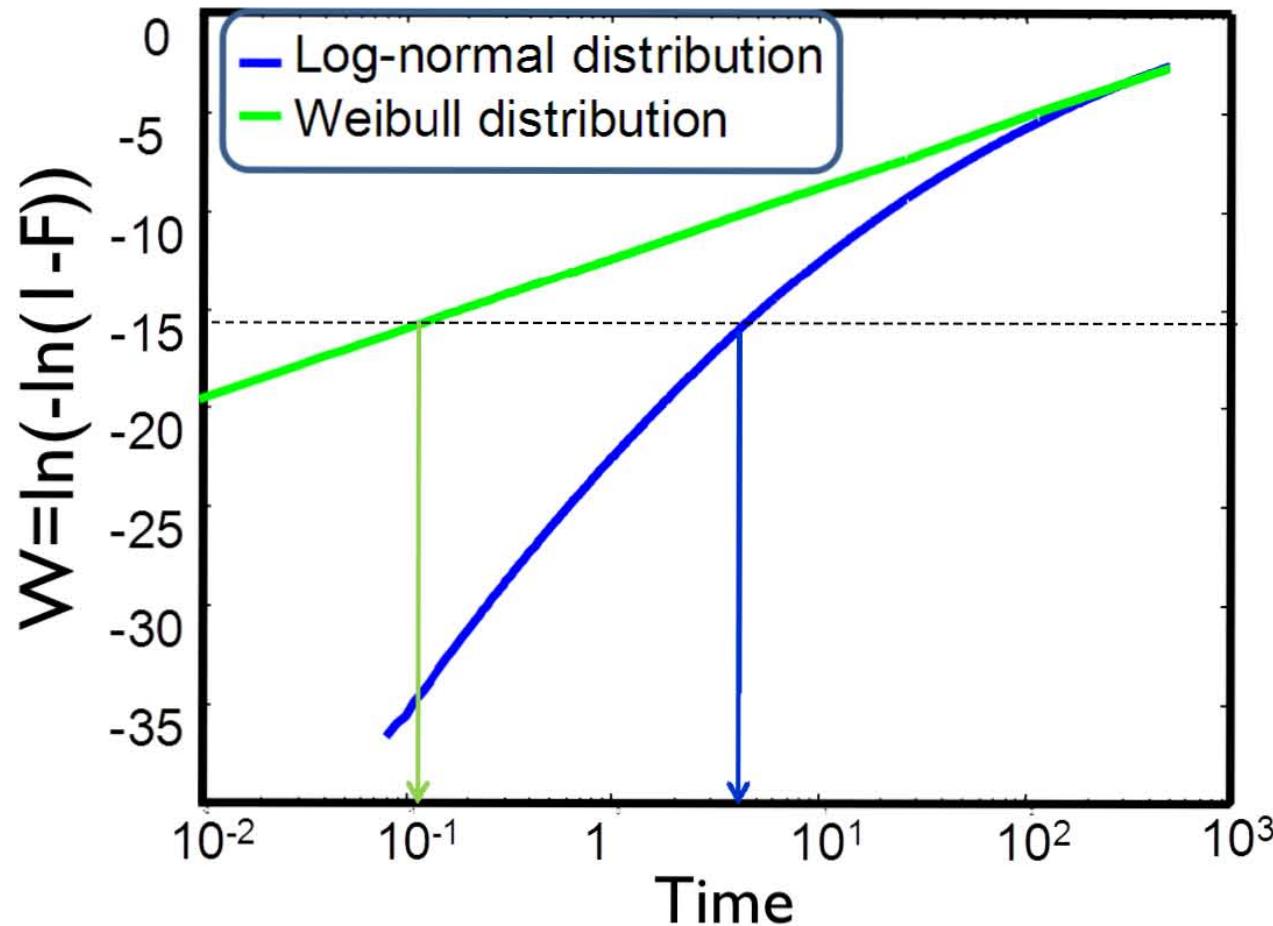
Rank	Lifetime	
1	181	0.012
2	299	0.028
3	389	0.045
4	430	0.061
5	535	0.078
6	610	0.094
7	805	0.111



Weibull Distribution Parameters  
When  $t=\alpha$ ,  $\ln(1-F(t))=-1$ ,  $F(t)=0.632$ ,  $\alpha=2990$   
 $\beta$  estimated using parameter fitting as 1.56

Log-Normal Distribution Parameters  
 $s=\ln(T_{50\%}/T_{15.9\%})$ ,  $\sigma=\ln(3600/980)=1.30$   
 $\mu=\ln(T_{50\%})=\ln(3600)=8.19$

# Problem of matching the moments



Log-normal distribution is considerably optimistic

# Fisher's Maximum Likelihood Method

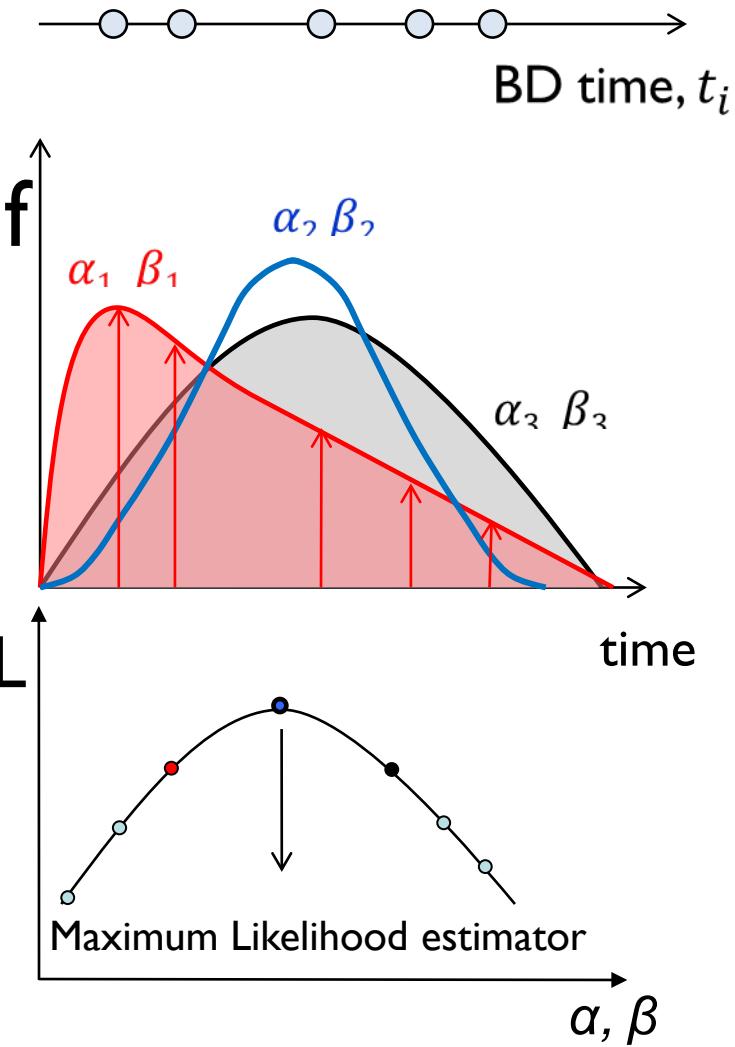
Towering figure, showed that Mendel manipulated data; MATLAB **fitdist** functions

$$f(t_i, \alpha, \beta)$$

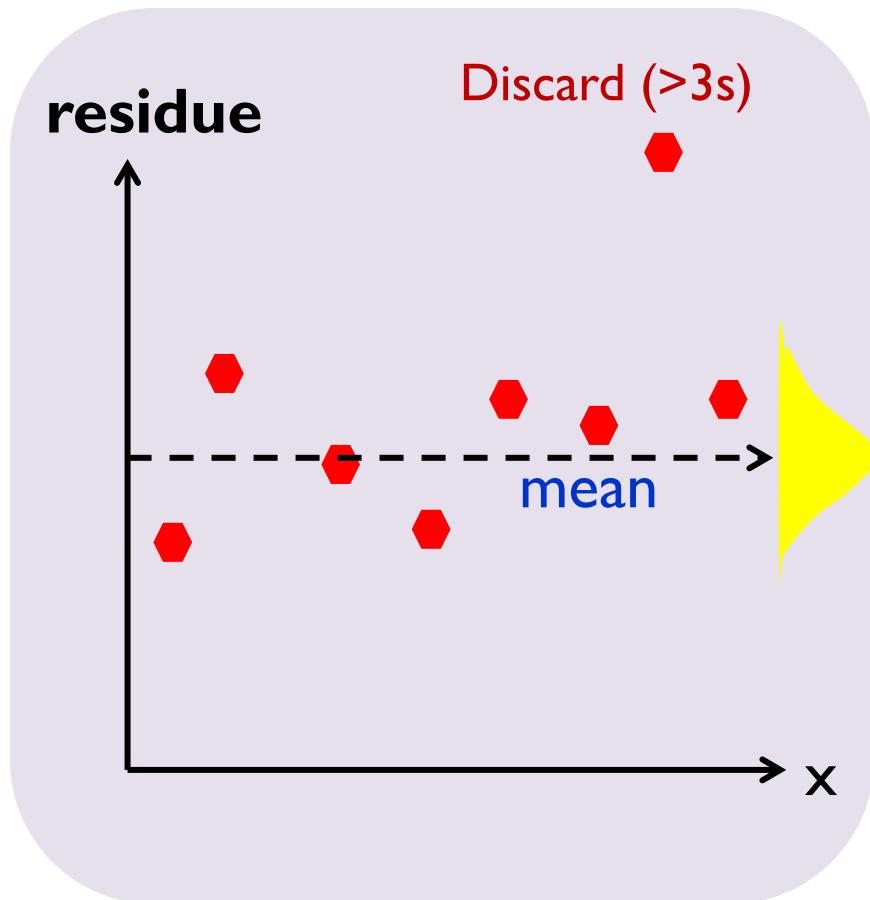
$$L = \prod_{i=1}^n f(t_i, \alpha, \beta)$$

$$\ln L = \sum_{i=1}^n \ln f(t_i, \alpha, \beta)$$

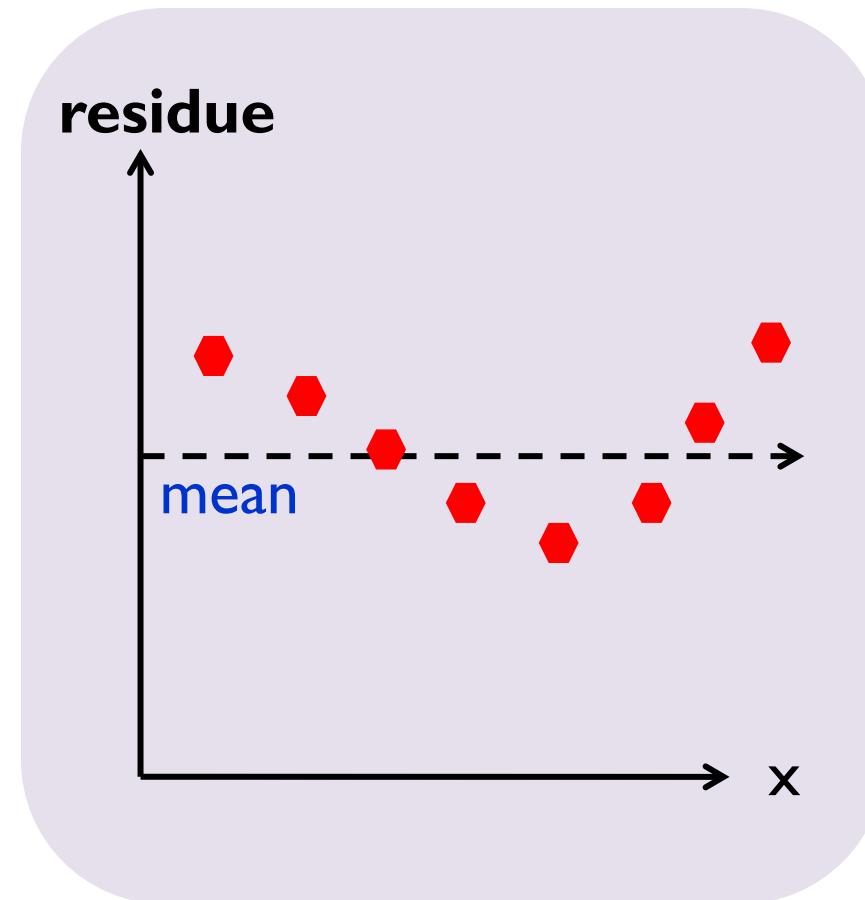
$$\frac{d \ln L}{d \alpha} = 0 \quad \frac{d \ln L}{d \beta} = 0$$



# Goodness of Fit: Residual method



A good fit (normal distribution of residue))



A bad fit (systematic distribution in residual)

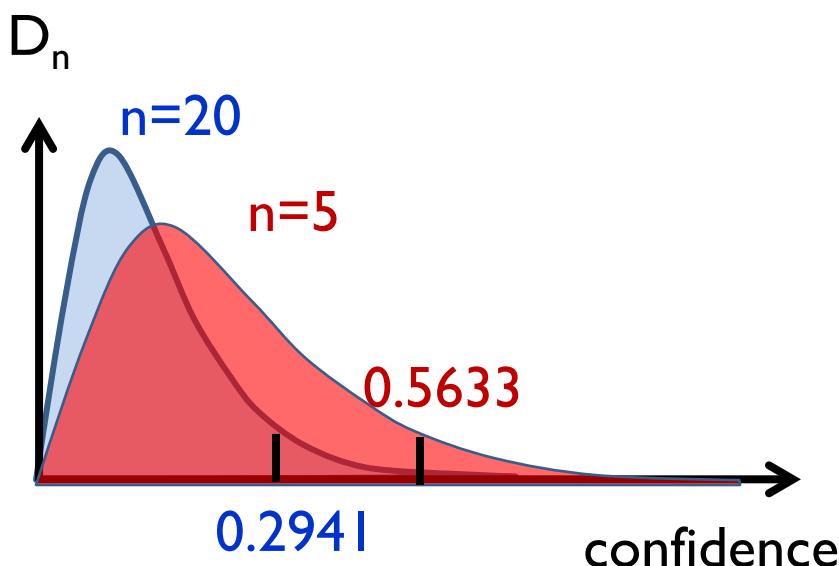
# Kolmogorov-Smirnov algorithm

Compute ...  $D_n = \max |F_{obs}(t_i) - F_{theory}(t_i)|$       5% significance level

Sample size

If  $D_n > D_n^{crit}$ , fit is poor ...

n	D <sub>crit</sub> (n)
5	0.5633
10	0.4092
20	0.2941
50	0.1884



# A famous example: Schon story

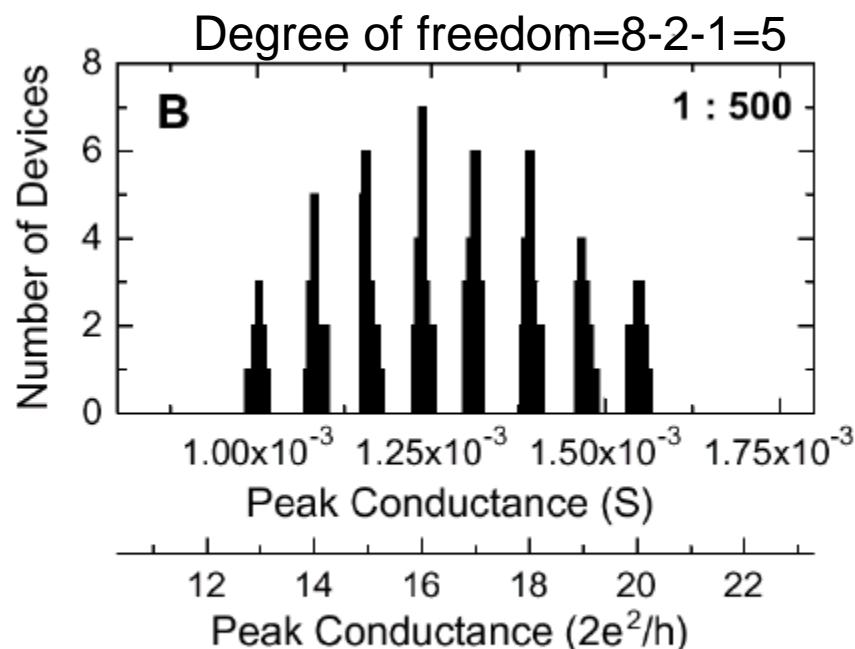


Figure 46. Figure 3(B) from "SingleMolecule" Paper (XIII), showing a histogram of conductances from diluted SAMFETs,

The data indicating conductance quantization did not arise from an objective measurement process. At a minimum, the assignment of conductance values was controlled by the expected shape of the final distribution. Such a biased process cannot provide convincing evidence for quantization. The response to this concern appears to be deliberately deceptive, suggesting that this misrepresentation was intentional.

The preponderance of evidence indicates that Hendrik Schon committed scientific misconduct, specifically data fabrication, in this case.

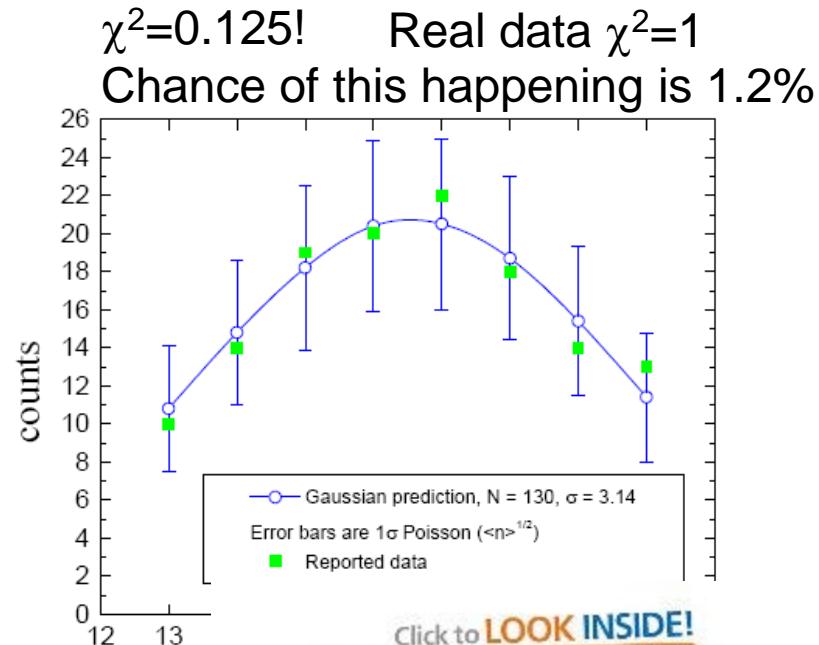
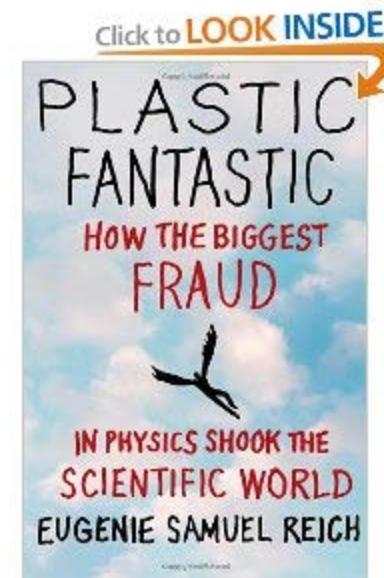
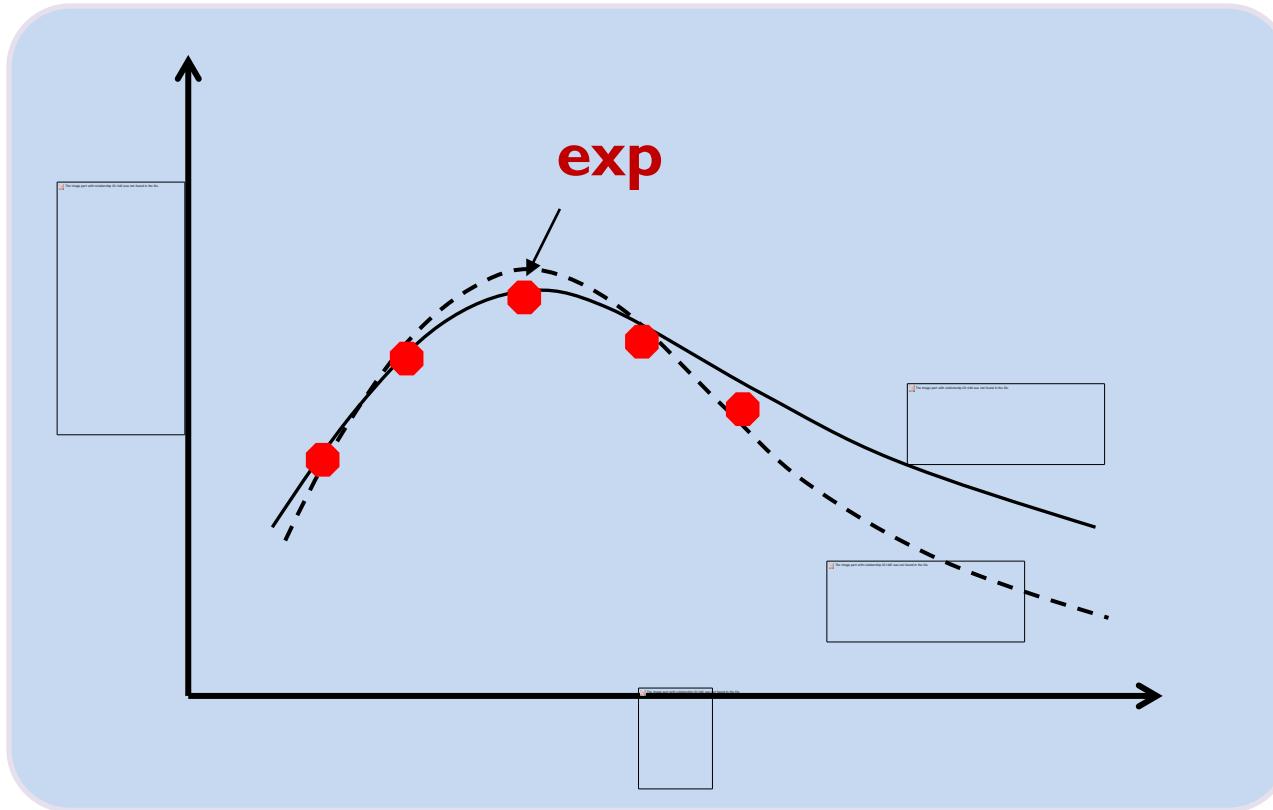


Figure 47. devices in ea



# Recall: MLE can be used to fit any model to the data



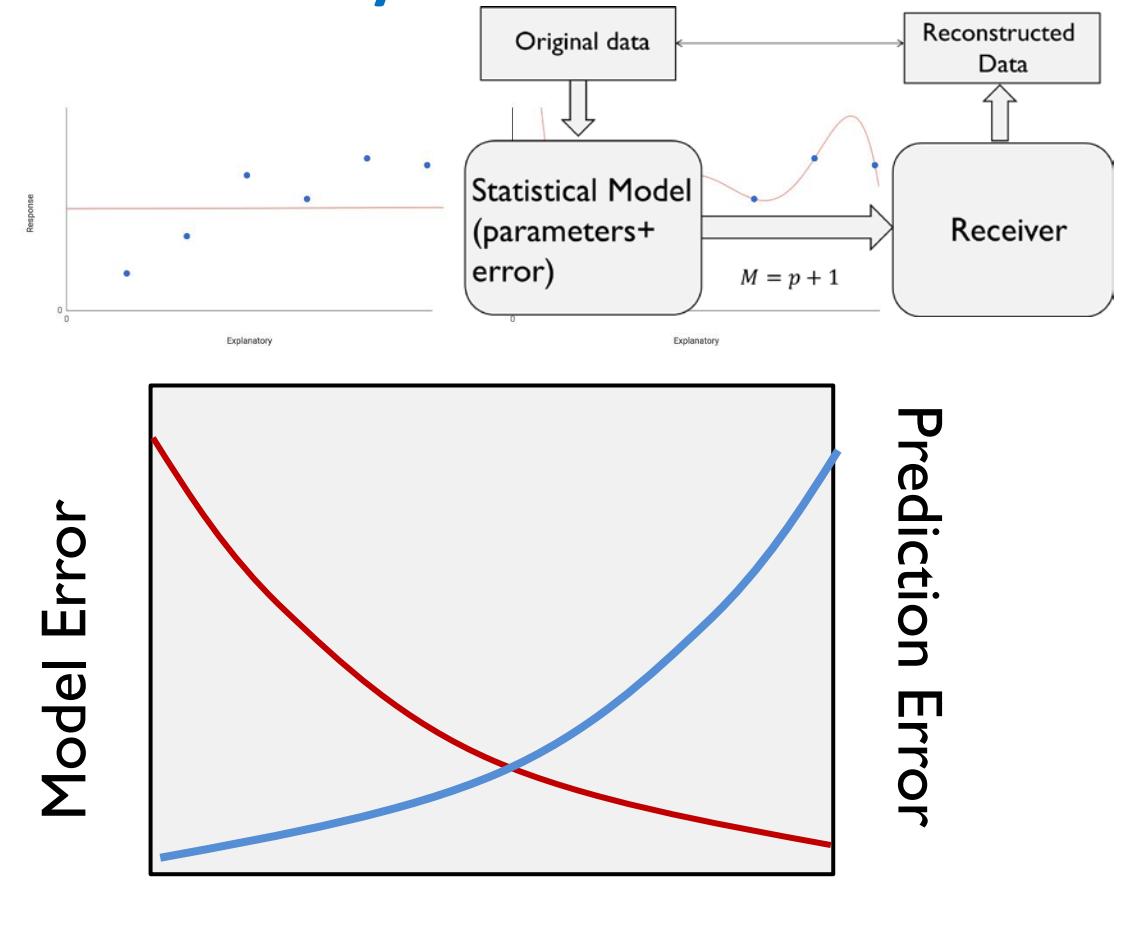
Each model can be checked for  $\chi^2$ ,  $KS$ , or  $QQ$  tests.  
What if two or more models passes the test. Which one is better?

# Principle of Parsimony

Aristotle: Nature operates in the shortest way possible.

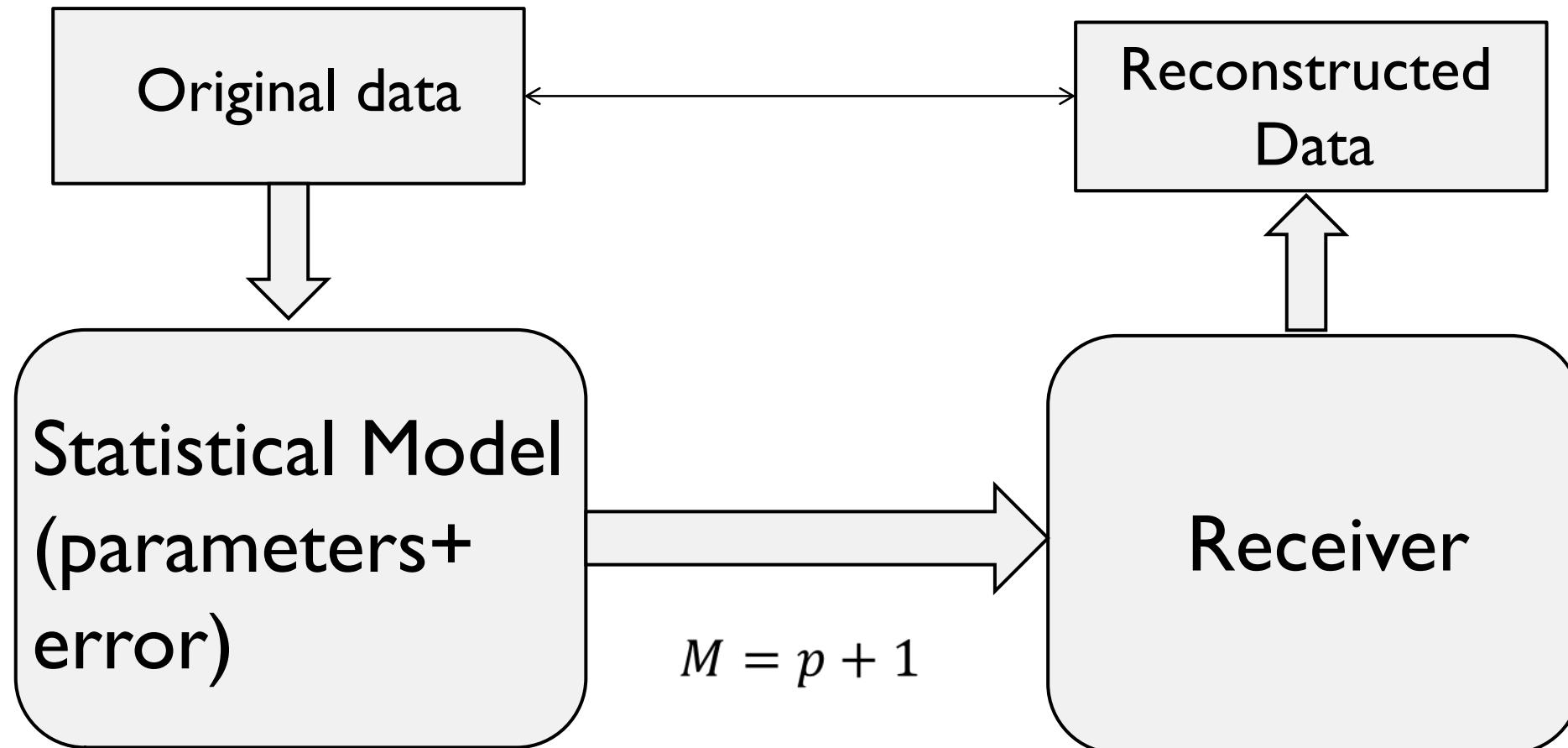
George Box: All models are wrong, but some are useful.

Occam's Razor: “given two or more equally acceptable explanations for a phenomenon, work with the one which introduces the fewest assumptions.”



Model Complexity  
Defined by # parameters

# Essence of the information theoretic approach



# Parameter number vs. goodness of fit

$n$  = number of samples,  $M$ =number of parameters

I) Method of adjusted residual ...

$$R_{adj}^2 = \frac{(n-1)R^2 - (M-1)}{n-M}$$

$$M \rightarrow p + 1$$

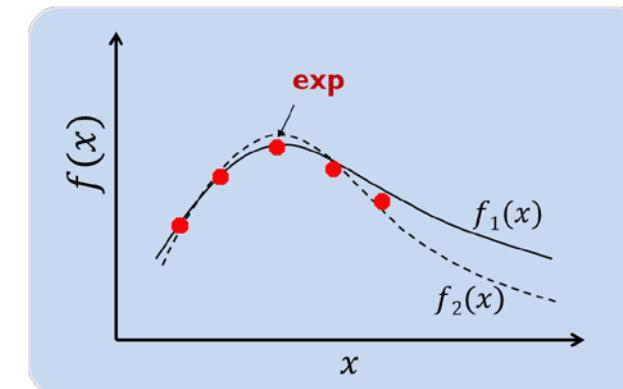
2) Akaike Information Criterion

$$AIC = n \times \ln(R^2/n) + 2M$$

2) Schwarz Information Criterion

$$BIC = n \times \ln(R^2/n) + M \times \ln n$$

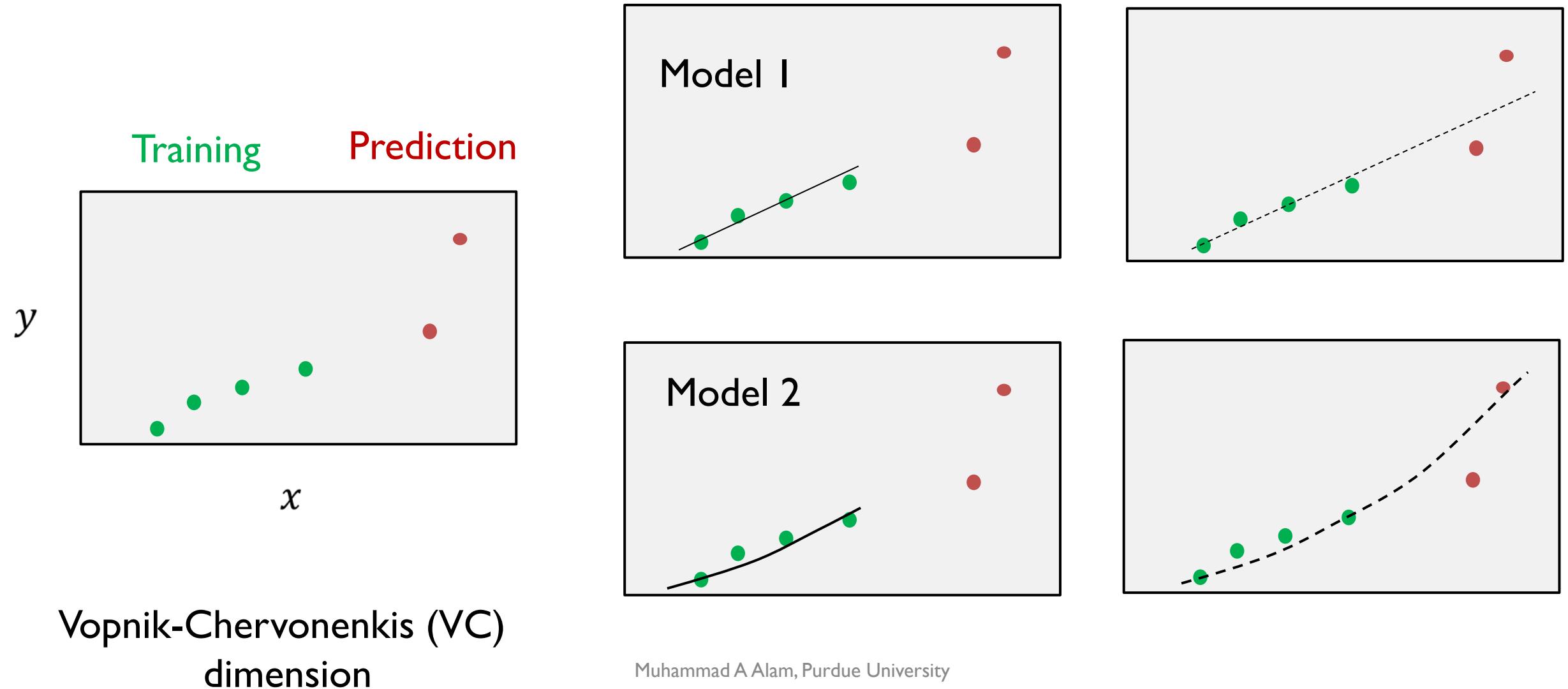
$$R \equiv \sum_{n=1}^n (t_i - t_{i,fit})^2$$



Error penalty  
Parameter Penalty

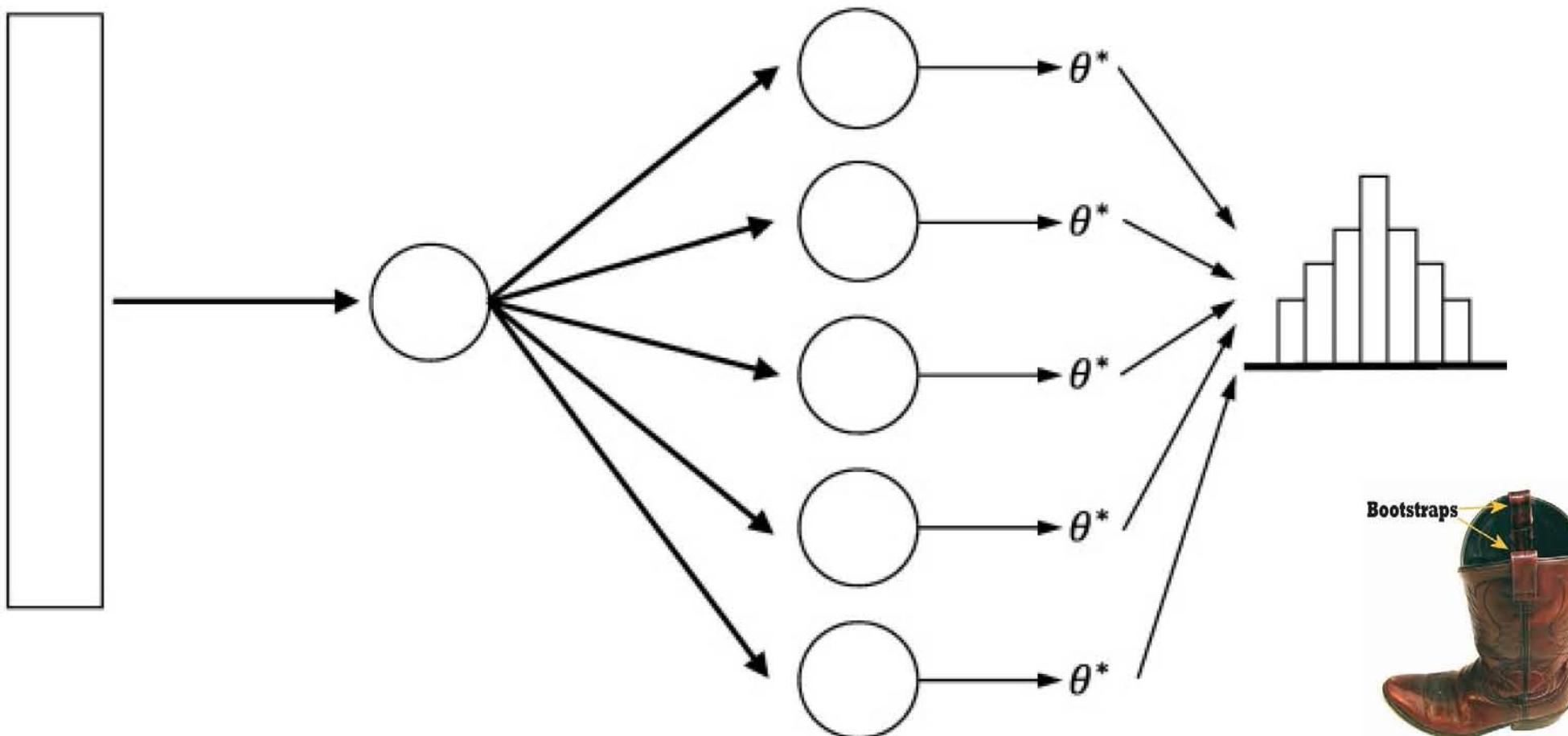
Ref. Les Kirkup, *Data Analysis with Excel*,  
Cambridge Univ. Press. P. 304

# Cross validation method

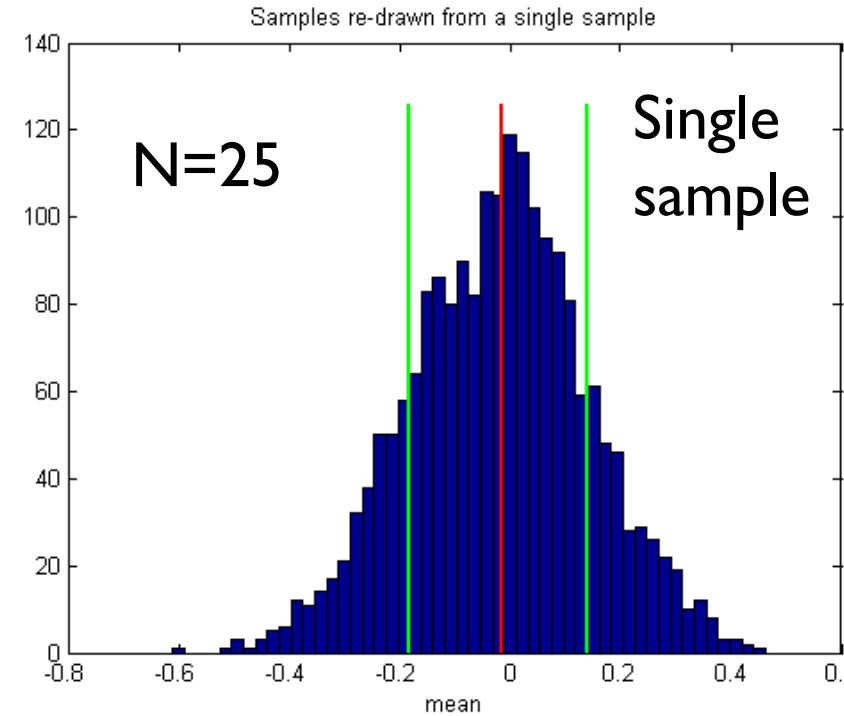
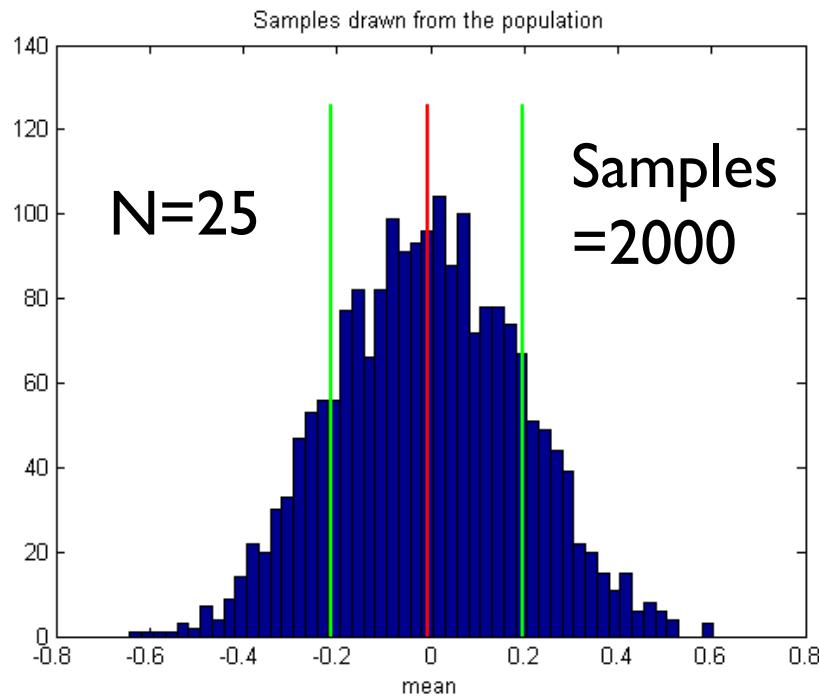


# Overall algorithm for bootstrapping

population      sampling      measured sample      re-sampling with replacement      bootstrap samples      statistical quantities of interest      distribution analysis



# Multiple sample vs. single sample



Bootstrap average is not zero!

And yet, the  $s \sim 0.18$ , just from a single sample.

The success of the method relies on precision measurement

# Course Outline

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

Lecture 1: Introduction

Lecture 2: Collecting and plotting  $x_1, x_2, \dots x_n$

Lecture 3: Physical and empirical  $f, F, df/dx, \dots$

Lecture 4: Model selection between  $f_1, f_2, \dots$

Lecture 5: Model Selection: Cross-validation and Bootstrapping method

Lecture 6: Scaling theory with known  $f$ ,  $f(\bar{x}) = f(\bar{X})$

Lecture 7: Scaling theory with unknown  $f$ ,  $\bar{x} \rightarrow X$

Lecture 8: Design of experiments to determine  $\bar{y}_{\max} = f(\bar{x})$

Lecture 9: DOE and ANOVA

Lecture 10: Principle component analysis for classifying  $\{y\}$ .

Lecture 11: Machine learning ... Statistical approach to learn  $f$

Lecture 12: Machine Learning .... Deep network, Karnaugh map, and other components

Lecture 13: Interpretable ML: Physics-based machine learning  $f = f_{\text{physics}} + \Delta f$

Lecture 14: Conclusions

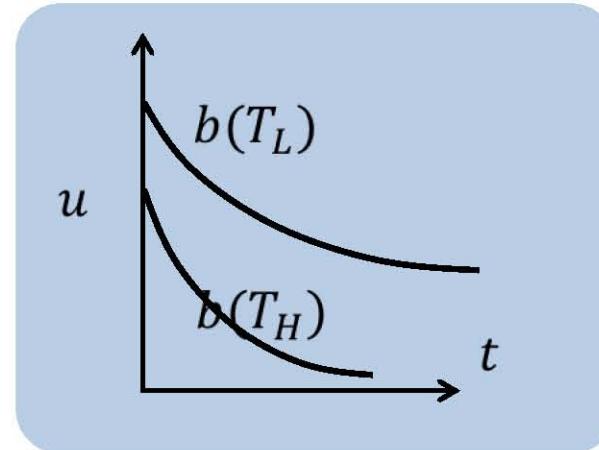
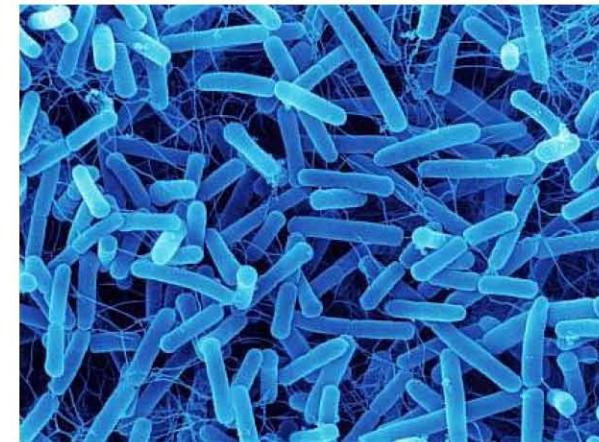
# Stress-induced cell death

Equation:  $\frac{dn}{dt} = -b(T)n$

$$\Rightarrow n = n_0 e^{-b(T)t} \equiv f(n_0, b, t)$$

5 experiments each for  $n_0, b, t$   
... 125 measurements

If with multiple samples, hundreds  
of measurements required.



# Buckingham PI Theorem

If the function  $g$  depends on parameters  $q_1, q_2, \dots, q_n$ , then

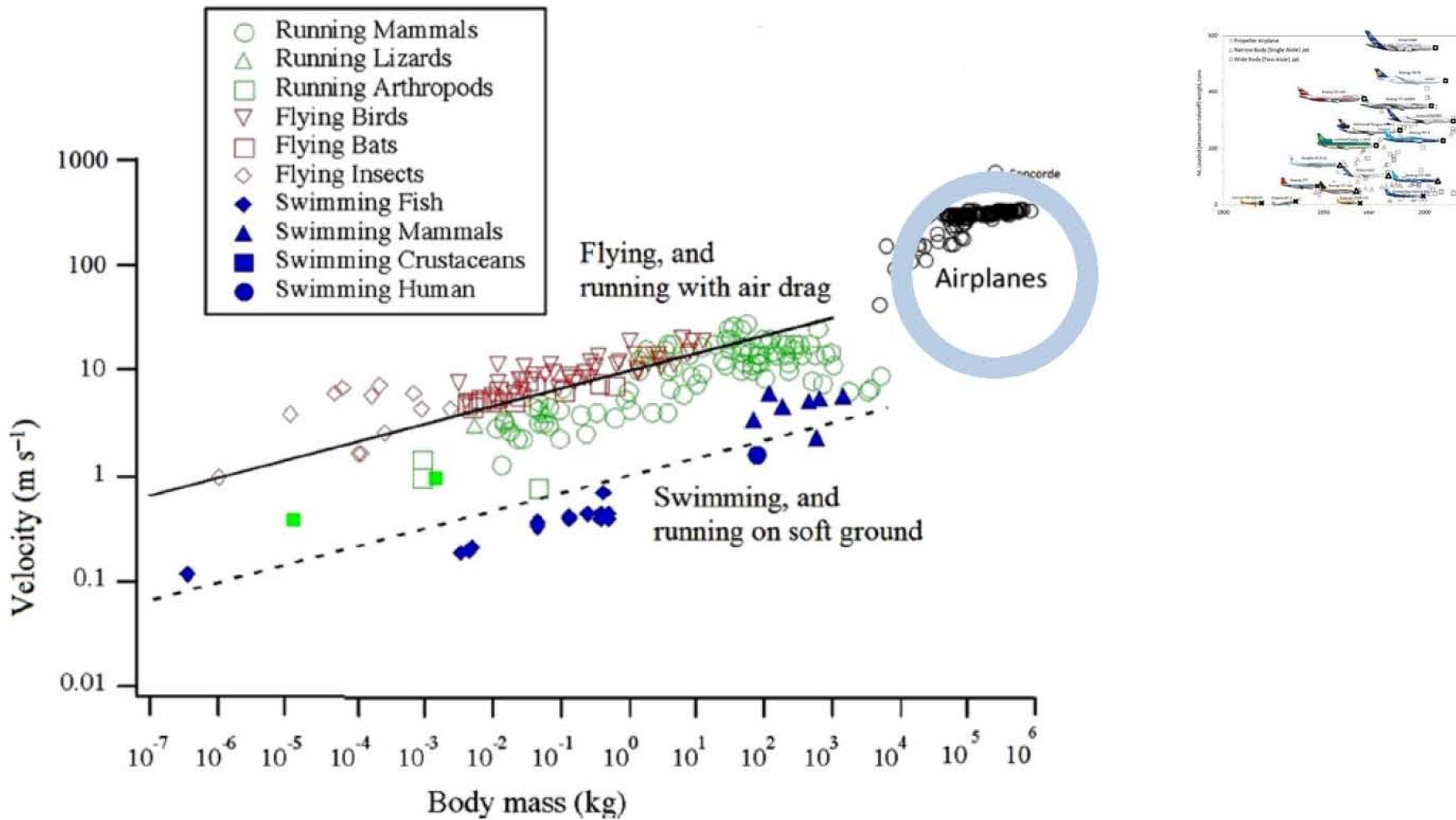
$$g(q_1, q_2, \dots, q_{\textcolor{red}{n}}) = 0$$

The same expression can be expressed in terms of  $(n-m)$  independent dimensionless ration, or  $\Pi$  parameters.

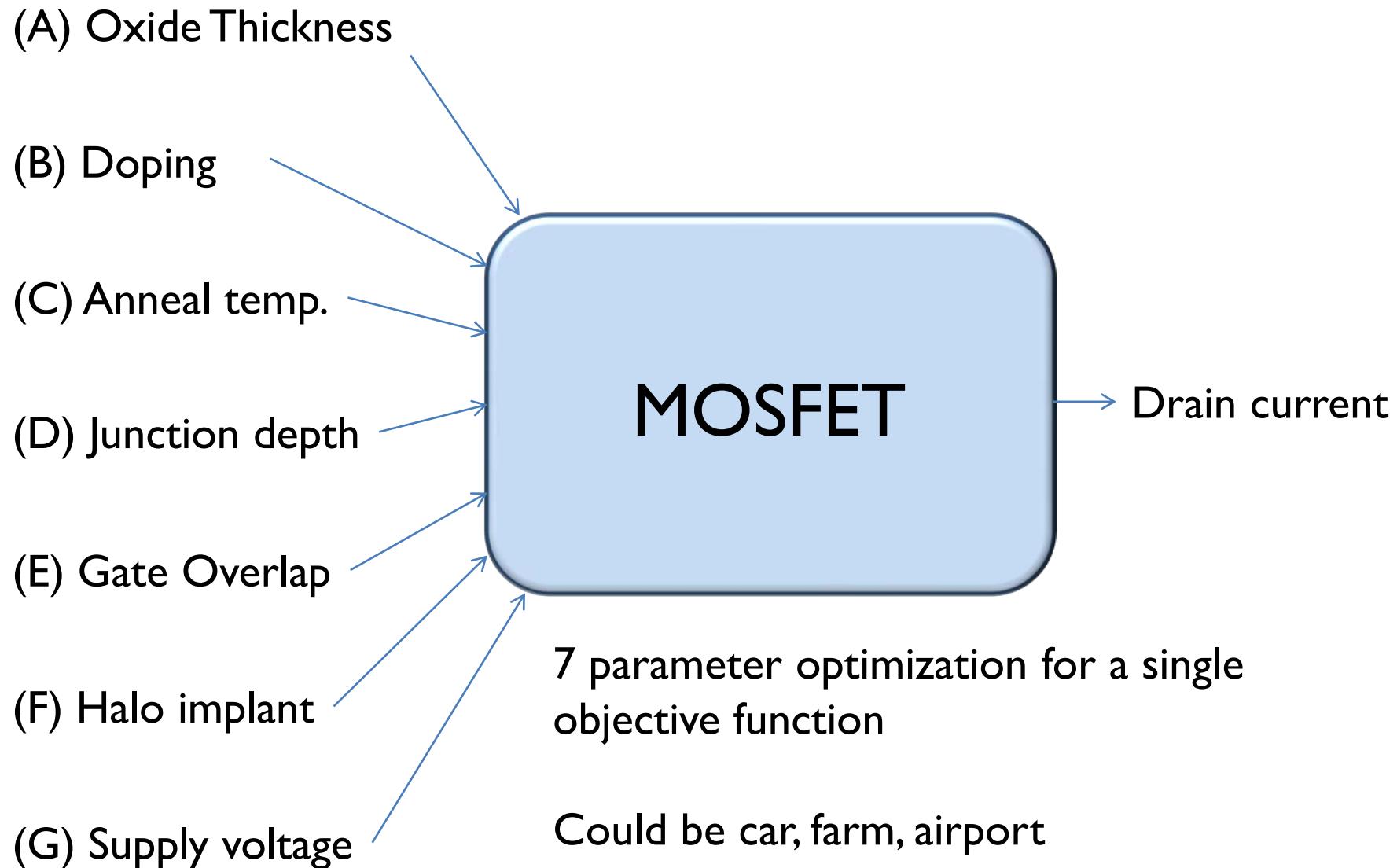
$$G(\Pi_1, \Pi_2, \dots, \Pi_{\textcolor{red}{n}-\textcolor{blue}{m}}) = 0$$

$m$ = minimum number of independent dimension typically given by  $r$ , where  $r$  is the rank of the matrix

# Scaling theory of things that move



# Fisher's design of experiment



# Philosophical shift with DOE

Before Fisher ...

Experimentalist  
determine what  
experiments to do

Results

Statisticians/  
Theorists/Expt  
collaborate to  
interpret results

After Fisher ...

Statisticians/  
Theorists/Expt  
plan what  
experiments to do

Results

Statisticians/  
Theorists/Expt  
collaborate to  
interpret results

Output cannot be greater than input .....

# Taguchi table: Continued

$L_4(2^3)$

Run	Columns		
	1	2	3
1	1	1	1
2	1	2	2
3	2	1	2
4	2	2	1

$L_8(2^7)$

Run	Columns						
	1	2	3	4	5	6	7
1	1	1	1	1	1	1	1
2	1	1	1	2	2	2	2
3	1	2	2	1	1	2	2
4	1	2	2	2	2	1	1
5	2	1	2	1	2	1	2
6	2	1	2	2	1	2	1
7	2	2	1	1	2	2	1
8	2	2	1	2	1	1	2

$L_{12}(2^{11})$

Run	Columns										
	1	2	3	4	5	6	7	8	9	10	11
1	1	1	1	1	1	1	1	1	1	1	1
2	1	1	1	1	1	1	2	2	2	2	2
3	1	1	2	2	2	1	1	1	2	2	2
4	1	2	1	2	2	1	2	2	1	1	2
5	1	2	2	1	2	2	1	2	1	2	1
6	1	2	2	2	1	2	2	1	2	1	1
7	2	1	2	2	1	1	2	2	1	2	1
8	2	1	2	1	2	2	2	1	1	1	2
9	2	1	1	2	2	2	1	2	2	1	1
10	2	2	2	1	1	1	1	2	2	1	2
11	2	2	1	2	1	2	1	1	1	2	2
12	2	2	1	1	2	1	2	1	2	2	1

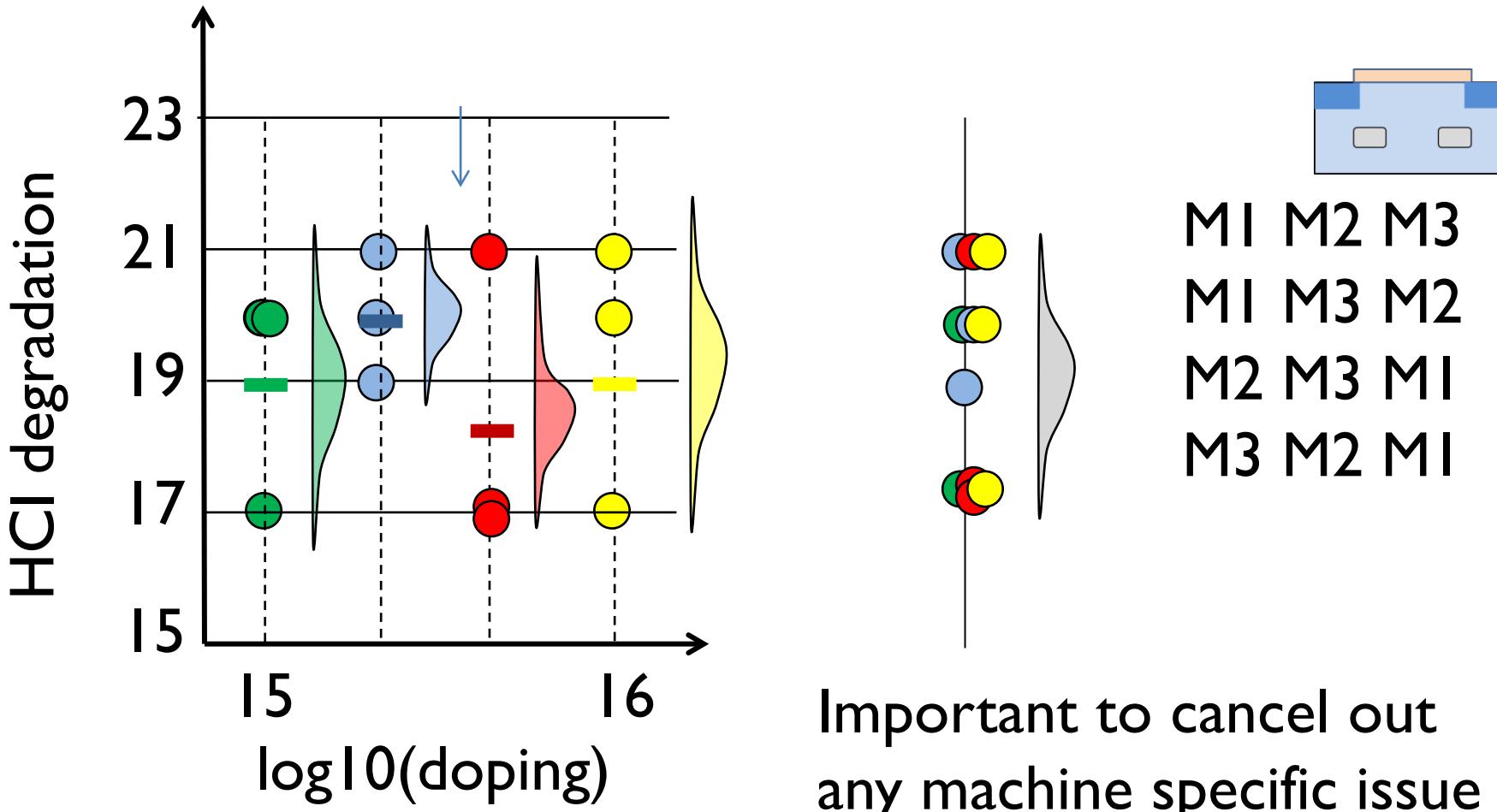
$L_N(S^M)$

M .. Variables

S ... Levels

N ... experiment > DOF = I + M(S-I)

# Analysis of Variance: Single/multiple factors Analysis



# Single factor ANOVA: Wood Treatment

	replicates					
	1	2	3	4	5	6
5	7	8	15	11	9	10
10	12	17	13	18	19	15
15	14	18	19	17	16	18
20	19	25	22	23	18	20

↓ treatments

$\sum(data - AVG)^2 = 512$        $6 \times 63.8 = 382.8$

s-avg	(s-avg-AVG) <sup>2</sup>
10.00	35.50174
15.67	0.085069
17.00	1.085069
21.17	27.12674
15.96	63.79861

Variation	SS	df	MS	F	P-value	F crit
Between Groups	382.7917	3	127.5972	19.60521	3.59E-06	4.938193
Within Groups	130.1667	20	6.508333			
Total	512.9583	23				

# Course Outline

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

Lecture 1: Introduction

Lecture 2: Collecting and plotting  $x_1, x_2, \dots x_n$

Lecture 3: Physical and empirical  $f, F, df/dx, \dots$

Lecture 4: Model selection between  $f_1, f_2, \dots$

Lecture 5: Model Selection: Cross-validation and Bootstrapping method

Lecture 6: Scaling theory with known  $f$ ,  $f(\bar{x}) = f(\bar{X})$

Lecture 7: Scaling theory with unknown  $f$ ,  $\bar{x} \rightarrow X$

Lecture 8: Design of experiments to determine  $\bar{y}_{\max} = f(\bar{x})$

Lecture 9: DOE and ANOVA

Lecture 10: Principle component analysis for classifying  $\{y\}$ .

Lecture 11: Machine learning ... Statistical approach to learn  $f$

Lecture 12: Machine Learning .... Deep network, Karnaugh map, and other components

Lecture 13: Interpretable ML: Physics-based machine learning  $f = f_{\text{physics}} + \Delta f$

Lecture 14: Conclusions

# Classification problem in big data

## Advertisement Recommendation

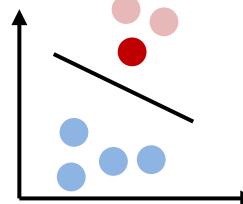


Everything is a Recommendation

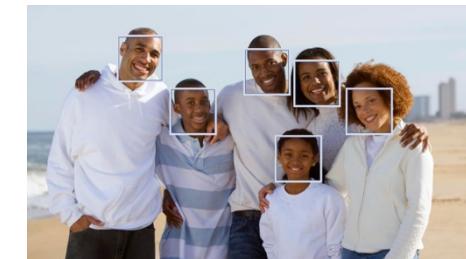
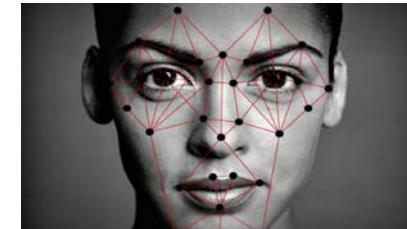


Over 75% of what people watch comes from our recommendations

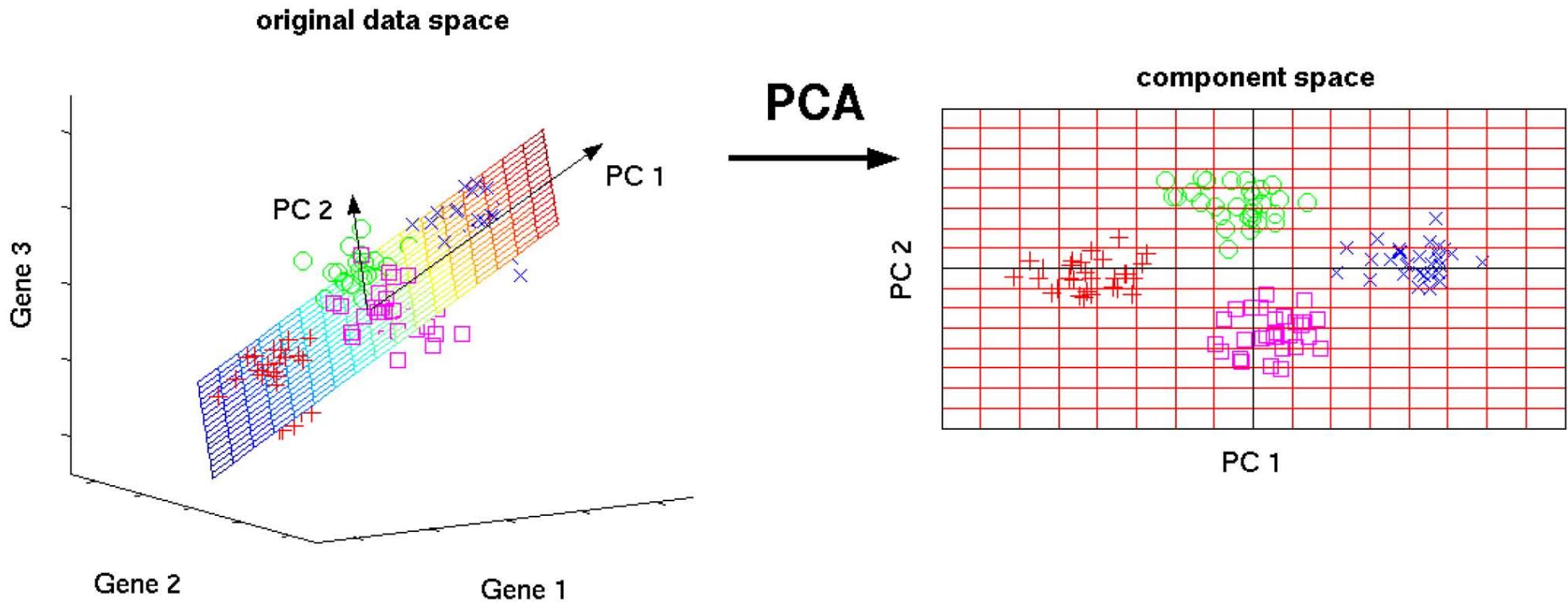
Recommendations are driven by Machine Learning



## Facial Recognition Voice Recognition Spam Filtering



# Principle Component Analysis: Classification

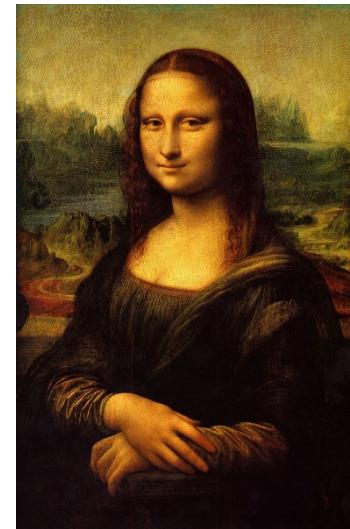


If you like this book, you will also like that book (because you belong to the same category)

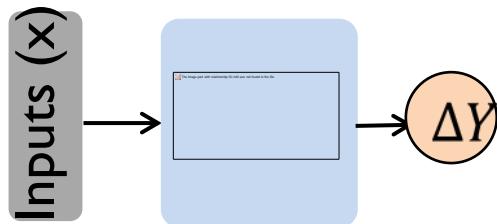
# Principal Component Analysis: Image Transimission

$$X_{1000 \times 500} = u_1 \sigma_1 v_1^T + u_2 \sigma_2 v_2^T + \dots$$

$|x|$   
↓  
 $|x500|$   
↑      ↑  
 $|1000x1|$



# Statistical Machine Learning

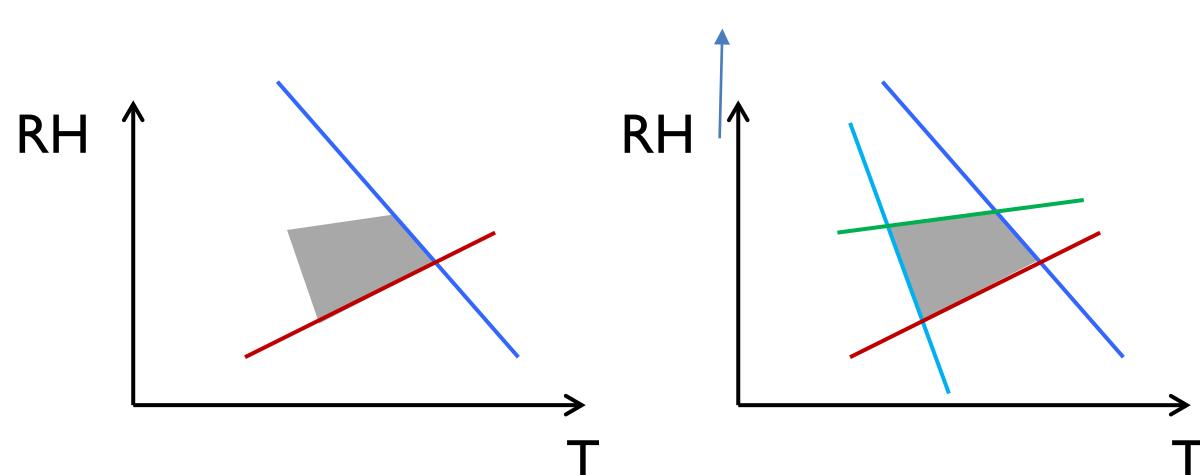
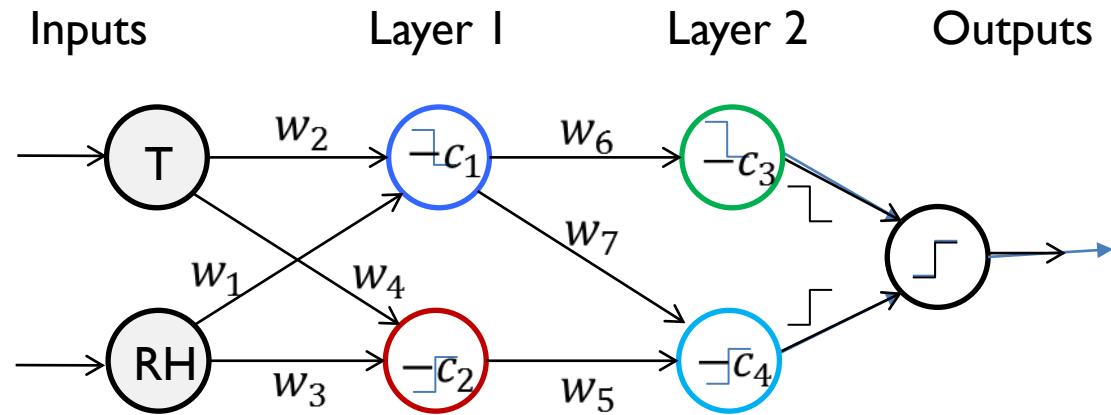


$$y = f(x)$$

y: pass, fail  
y: A, B, C, D, E  
Y= grade points.

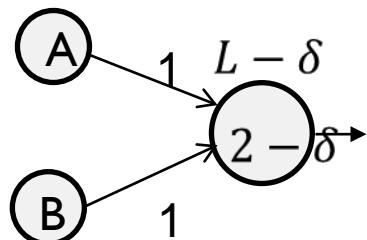
$f(x)$  ... Physics  
 $f(x)$  ... Statistical curve fitting  
 $f_{\max}(x)$  .... Design of expt

# Deep network

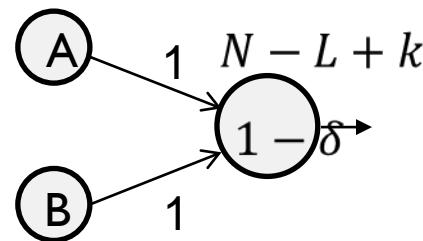
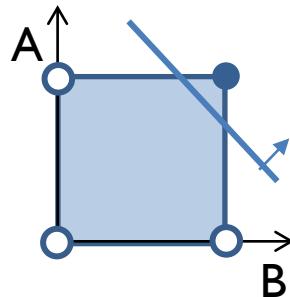


# A perceptron implements AND, OR, and NOT gates

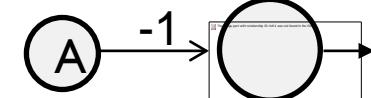
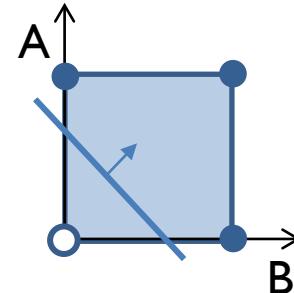
$L-N + k$  here L= (positive input), N= total number =2, k=1 is the threshold for binary logic



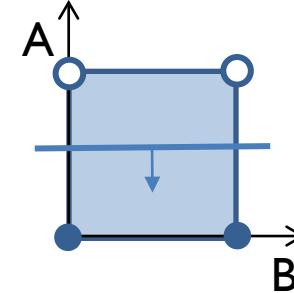
AND



OR



NOT

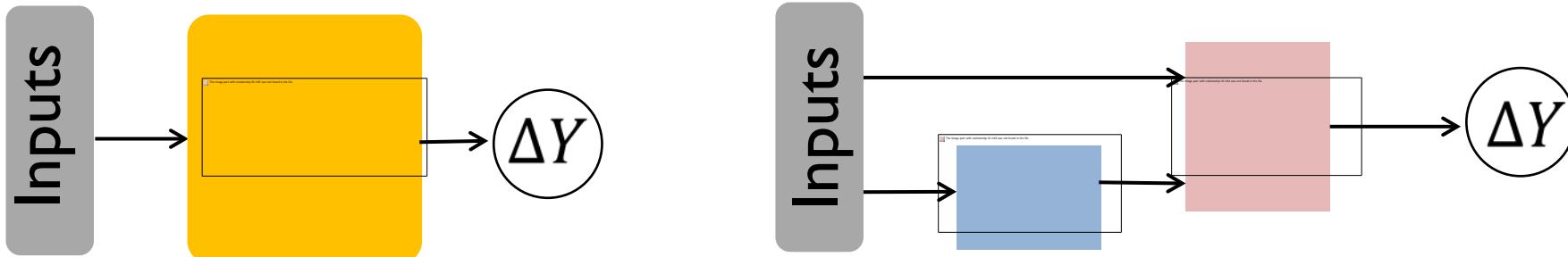


A	-
1	
0	

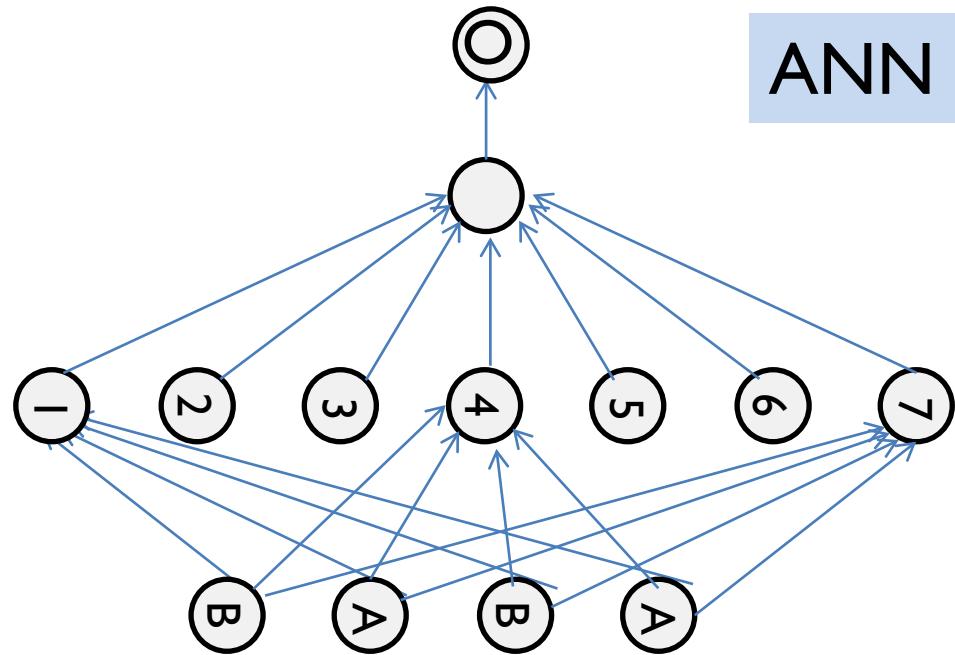
# A falling ball with gravity and resistance

$$\frac{dv}{dt} = -\frac{g}{(1 + z/R)^2} + bv^2$$

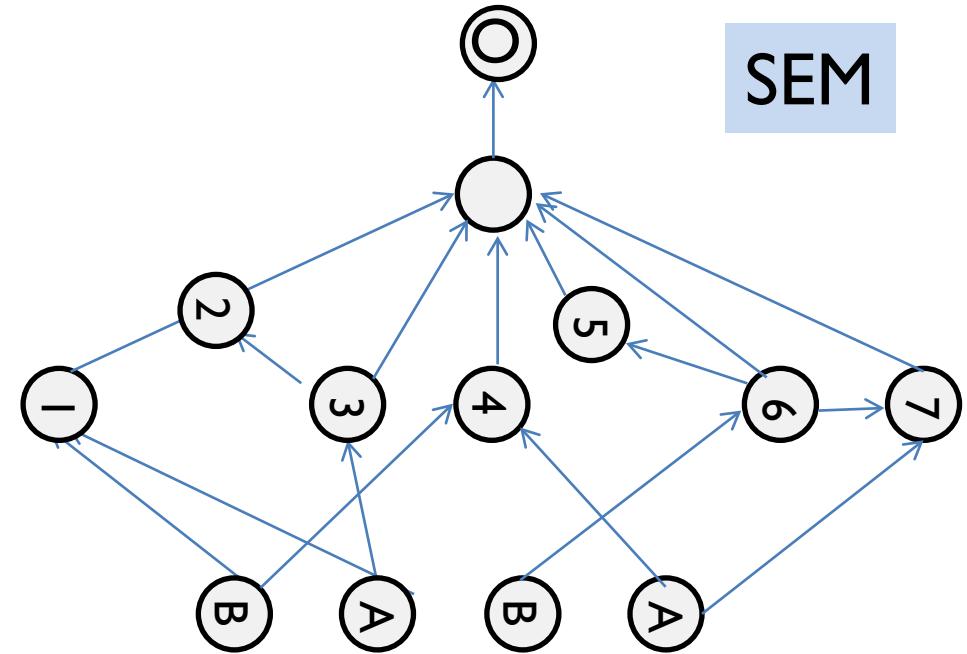
A numerical/perturbative solution may not be possible. In practice,  $b(z)$  is unknown function of humidity, temperature, etc. A machine learning approach is preferred.



# Structural Equation modeling: Motivation



ANN

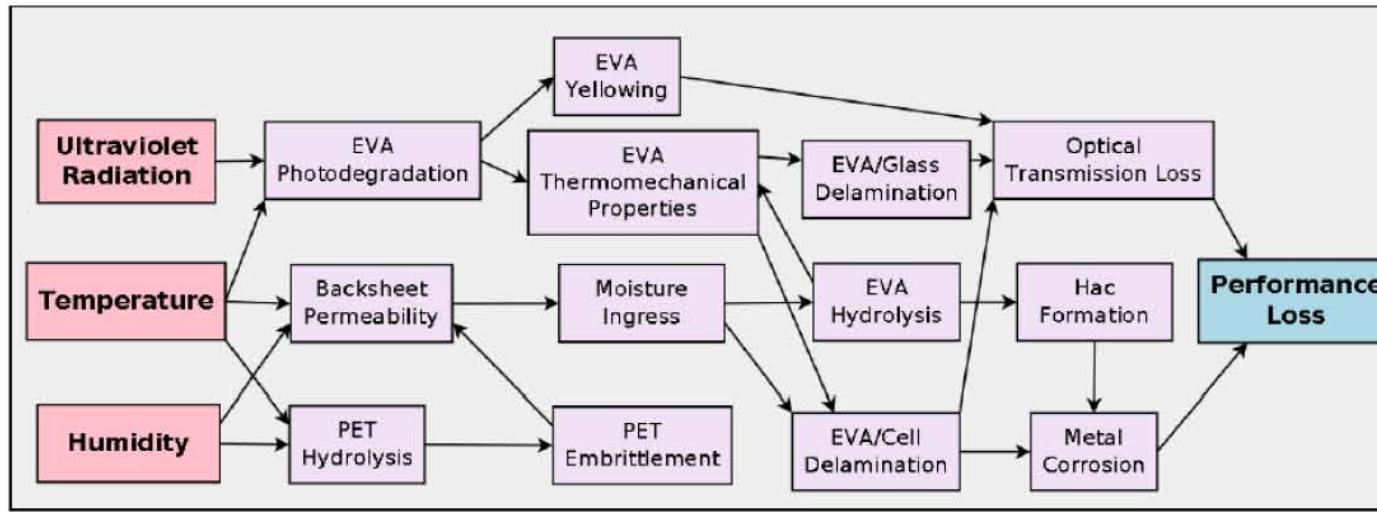


SEM

Nodes defined statistically  
not appropriate for extrapolation

Nodes defined physically  
Interconnects are nonlinear  
Extrapolation possible

# Example: Degradation of Solar Cells

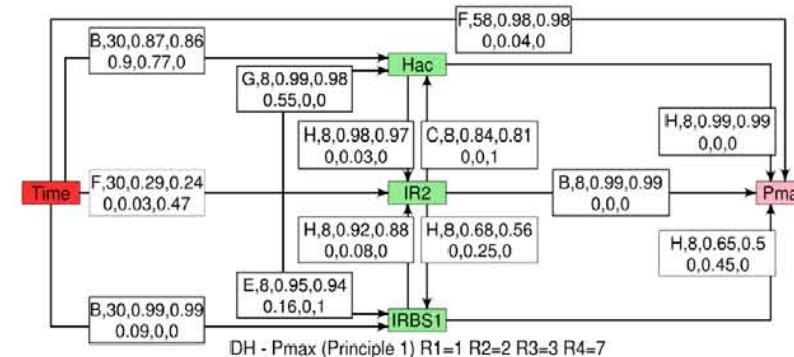


Received April 16, 2013, accepted April 23, 2013, date of publication June 10, 2013, date of current version June 25, 2013.

Digital Object Identifier 10.1109/ACCESS.2013.2267611

## Statistical and Domain Analytics Applied to PV Module Lifetime and Degradation Science

LAURA S. BRUCKMAN<sup>1</sup>, NICHOLAS R. WHEELER<sup>2</sup>, JUNHENG MA<sup>3</sup>, ETHAN WANG<sup>4</sup>,  
CARL K. WANG<sup>4</sup>, IVAN CHOU<sup>5</sup>, JIAYANG SUN<sup>3</sup>, AND ROGER H. FRENCH<sup>6</sup> (Member, IEEE)



**FIGURE 7.** Model generated with the most relationships shown for Principle 1 for the damp heat exposure for the system response of  $P_{max}$  using the unit variables of  $Hac$ ,  $IR2$  and  $IRBS1$  shows 13 relationships. Information on each relationship is described in the box. The information contained is functional form, number of observations,  $R^2$ , adjusted  $R^2$ , P-value 1, P-value 2 and P-value 3, respectively. The strength of the SSR is summarized by the line width of the SSR border based on the  $R^2$  value to aide visualization (below 0.2 not shown, R1 has the thinnest border (0.2-0.5), R2 (0.5-0.7), R3 (0.7-0.9) and R4 the thickest ( $\geq 0.9$ )). The functional forms are designated as A (simple linear), B (quadratic), C (simple quadratic), D (exponential), E (logarithmic), F (linear change point), G (nonlinearizable exponential-up) and H (nonlinearizable exponential-down).

# Conclusions

- I. Data is the lifeblood of modern science and technology. Learning to treat data with respect is an essential skill.
2. Modern data analysis is easy because they are embedded in systems. Once we become aware of the treasure-trove of functions available, data analysis is considerably simplified.
3. Design of experiments is a powerful approach for modern manufacturing. The scaling theory, Taguchi techniques, ANOVA all help reduce the dimensionality of the problem.
4. Machine learning is a modern way of curve-fitting, powered by modern computers. Its applications in advertisement and classification have been remarkable. One must develop the technique further for applications in science and engineering.

# Review Questions

1. Name three concepts that you are going to remember from this course.  
Explain why they were particularly important for you.
2. What is the essential difference between classical and modern statistics?
3. What are the topics from this course you can immediately apply to your research? What specific problems would you be able to apply them?
4. What are your over-all impression about the course? Did you find the concepts coherent or was the diversity of concepts distracting?
5. What additional topics would you have preferred that we cover in this course?