Ronald Fisher and Maximum Likelihood Estimation

By
John F. McGowan, Ph.D.
jmcgowan79@gmail.com

August 20, 2016
Bay Area Entrepreneurs in Statistics
Symation, Richmond, CA

Ronald Fisher (1890 – 1962) in 1913

Outline of Presentation

- Bio
- Maximum Likelihood
- Refresher on Probability and Statistics
- Mathematical Modeling Before Fisher
- Ronald Fisher and Maximum Likelihood (1912-1922)
- Modern Maximum Likelihood Estimation (MLE)
- Conclusion
- Q&A

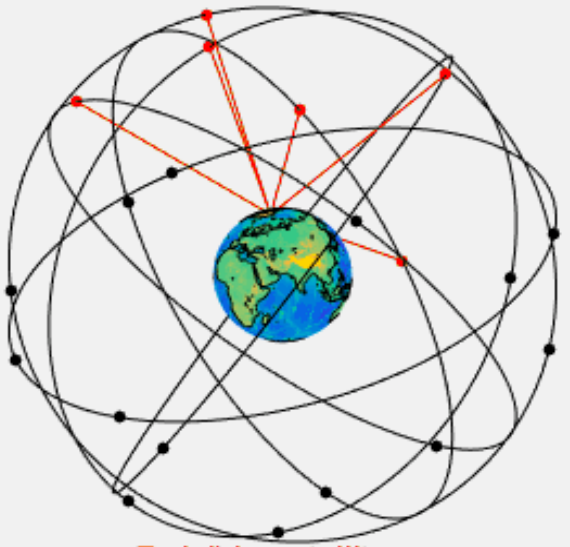
BIO

- **John F. McGowan, Ph.D.**
  - B.S. in Physics, Caltech
  - Ph.D. in Physics, University of Illinois at Urbana-Champaign (MLE)
- MPEG Video Compression (first software MPEG-2/DVD player at Compcore, acquired by Zoran for $60 M in stock in 1997, Zoran acquired by CSR in 2011, CSR acquired by Qualcomm in 2015)
- Image and Video Compression, Quality and Processing Research and Development at NASA Ames Research Center. Worked with Dr. Andrew B. “Beau” Watson, a noted Vision Science researcher at NASA (now at Apple).
- Speech recognition application software development and research for own business. (MLE)
- Visiting Scholar at HP Labs, working on mobile software. (MLE)
- Research and development of smart keyboard algorithms and program at Nod Labs (MLE)
- Human Interface Devices Algorithm Engineer at Apple Inc. (2014 to present)
  - Work on signal processing for Apple Pencil (accessory for iPad Pro)
  - Work on touch processing for the iPad Pro Mini

Maximum Likelihood

- Powerful, widely used statistical technique for parameter estimation and classification.
  - Global Positioning System (GPS) is a widely used example of maximum likelihood parameter estimation
  - Used by speech recognition programs such as Dragon Naturally Speaking, Carnegie Mellon open source SPHINX speech recognizer, and many others.
  - Swype and other smart keyboard programs for iPhones and Android smartphones. Widely used. Sold to Nuance for $102.5 Million in cash (October 2011)
  - Used for detection and measurement of the properties of the Higgs Boson at CERN (Large Hadron Collider)
  - Many other uses in agriculture, economics, finance, physics, and many other fields.

Global Positioning System (GPS)



A visual example of a 24 satellite GPS constellation in motion with the earth rotating. Notice how the number of *satellites in view* from a given point on the earth's surface, in this example in Golden CO (39.7469° N, 105.2108° W), changes with time.

GPS maximum likelihood

All Videos Shopping Images News More Search tools

About 400,000 results (0.36 seconds)

Scholarly articles for **GPS maximum likelihood**

High dynamic **GPS** receiver using **maximum likelihood** … - Hurd - Cited by 128
Maximum likelihood estimates of linear dynamic … - Rauch - Cited by 1106
Maximum likelihood multiple-source localization using … - Sheng - Cited by 588

Maximum-Likelihood GPS Parameter Estimation, NAVIGATION ...

https://www.ion.org/publications/abstract.cfm?jp=j...2409 ▾ Institute of Navigation ▾
Title: **Maximum-Likelihood GPS** Parameter Estimation. Author(s): Ilir F. Progri, Matthew C. Bromberg, and William R. Michalson. Published in: NAVIGATION ...

gps signal tracking using maximum-likelihood parameter estimation

https://www.ion.org/publications/abstract.cfm?jp=j...2237 ▾ Institute of Navigation ▾
Abstract: This paper considers the problem of **GPS** carrier tracking in the ... The modulation parameter is estimated using a **maximum-likelihood** method in an ...

Maximum-Likelihood GPS Parameter Estimation - PROGRI - 2014 ...

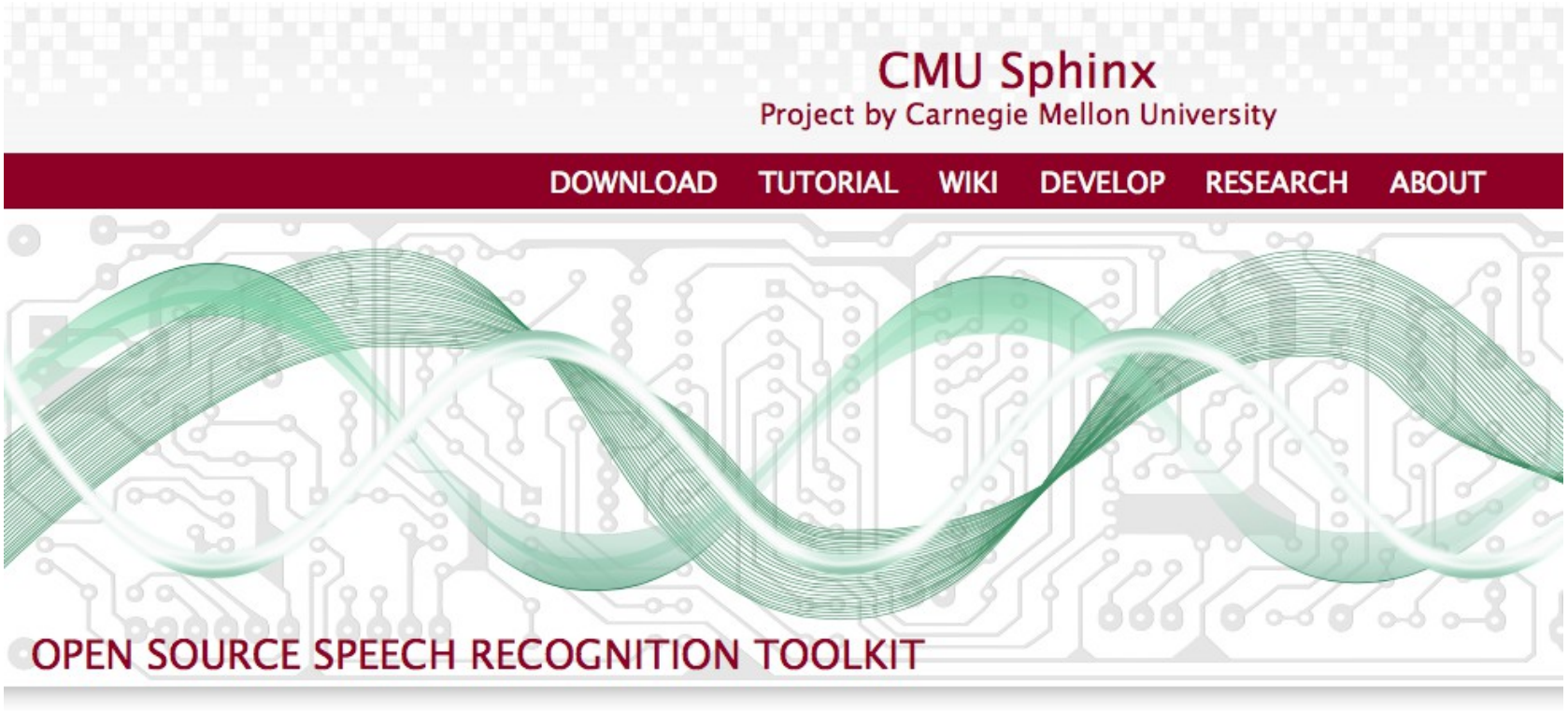
onlinelibrary.wiley.com › ... › Navigation › Vol 52 Issue 4 ▾ John Wiley & Sons ▾
by IF PROGRI - 2005 - Cited by 17 - Related articles
Aug 29, 2014 - ABSTRACT: Recently we proposed an acquisition process for a **maximum- likelihood GPS** receiver that considers the joint processing of all ...

GPS Signal Tracking Using Maximum-Likelihood Parameter Estimation

onlinelibrary.wiley.com › ... › Navigation › Vol 45 Issue 4 ▾ John Wiley & Sons ▾
by DE GUSTAFSON - 1998 - Cited by 2 - Related articles
GPS Signal Tracking Using **Maximum-Likelihood** Parameter Estimation. DONALD E.

Dragon Naturally Speaking

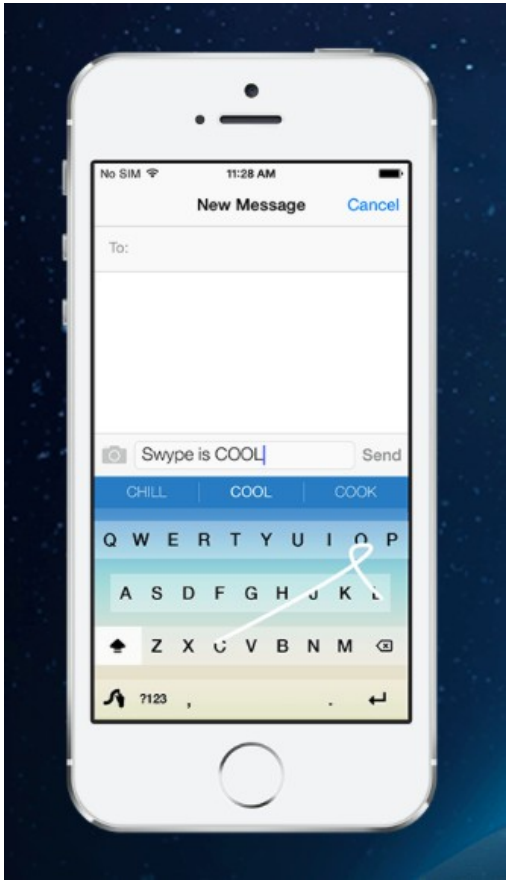
CMU Sphinx

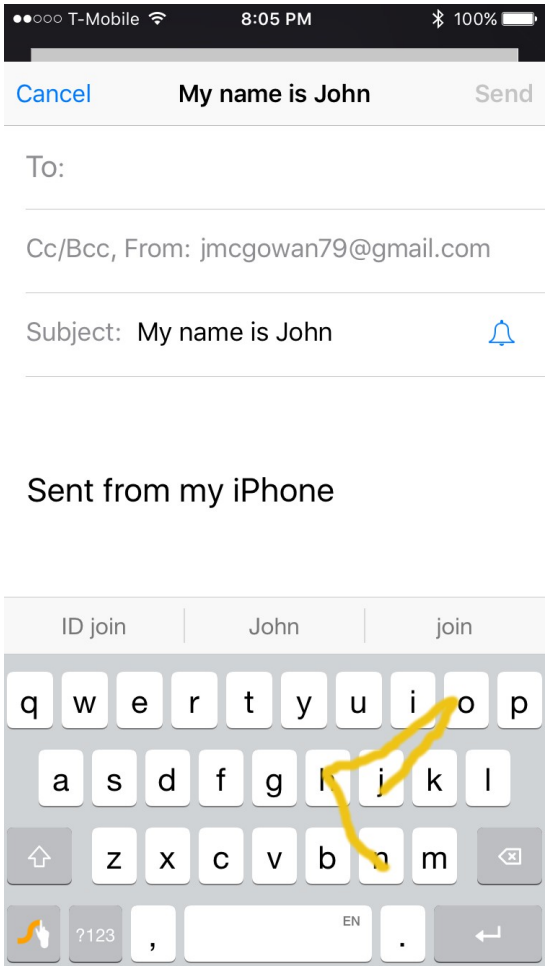


CMUSphinx is a project of the week on Sourceforge
June 9th, 2016

Polls

Swype


No SIM
11:28 AM
New Message
Cancel
To:
Swype is COOL
Send
CHILL
COOL
COOK
Q W E R T Y U I O P
A S D F G H J K L
Z X C V B N M
?123


T-Mobile
8:05 PM
100%
Cancel
My name is John
Send
To:
Cc/Bcc, From: jmcgowan79@gmail.com
Subject: My name is John
Sent from my iPhone
ID join
John
join
q w e r t y u i o p
a s d f g h j k l
z x c v b n m
?123
EN

Swype

- Highly successful software keyboard
- Widely used on both Android and iPhones
- Acquired by Nuance for $102.5 Million in cash in October, 2011
- Still market leader.

The Higgs Boson

The Nobel Prize in Physics 2013

Photo: A. Mahmoud

François Englert

Prize share: 1/2

Photo: A. Mahmoud

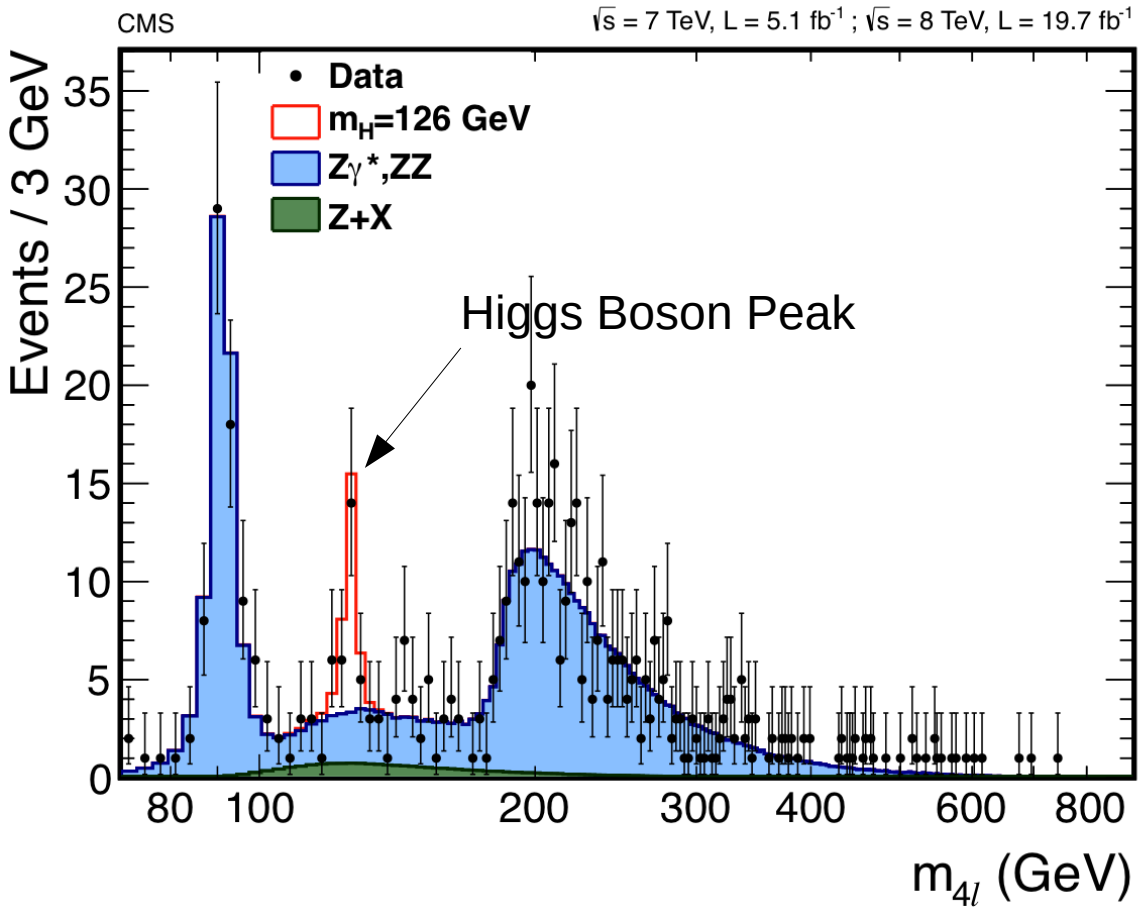
Peter W. Higgs

Prize share: 1/2

The Higgs Boson

Nobel Prize in Physics 2013 to Francois Englert and Peter Higgs for theory of Higgs Boson, following discovery of the particle at Large Hadron Collider (LHC) using Maximum Likelihood methods.

Higgs Boson Discovery



Refresher
Probability and Statistics

What is Probability?

prob·a·bil·i·ty (noun)

the extent to which something is probable; the likelihood of something happening or being the case.

"the rain will make the probability of their arrival even greater"

synonyms: likelihood, prospect, expectation, chance, chances, odds

What is probability?

- MATHEMATICS
- the extent to which an event is likely to occur, measured by the ratio of the favorable cases to the whole number of cases possible. (*frequentist*)
- "the area under the curve represents probability"
- Real number from 0.0 to 1.0
- Interpreted as a frequency or rate.

What is probability?

- A coin with heads and tails
- We say the coin has a probability of heads of 0.5 (50 percent) if as the number of times we flip the coin increases, the number of heads over the total number of coin flips tends to 0.5
- The probability is a rate or frequency that the coin lands heads up or tails up!
- Not well defined for a single coin flip
- What about a small number of flips: 5 for example?

Bayesian Probability

- Bayesian probability is a quantity that is assigned to represent *a state of knowledge*, or *a state of belief*.
- Also a real number from 0.0 to 1.0
- Defined for a single event or assertion
- What is the probability that someone you know likes you?
- Bayesian Probability has enjoyed a renaissance in the last thirty years, but has a number of problems that led Fisher and others to reject it a century ago – that remain unresolved.

Frequentists, Bayesians, Fisher and Maximum Likelihood

- Ronald Fisher was not, technically, a frequentist
- Maximum Likelihood can be formulated in a Bayesian way
- However, both Fisher's ideas on probability and statistics and Maximum Likelihood Estimation are largely frequentist.
- I will keep to the frequentist or quasi-frequentist tradition in discussing Maximum Likelihood Estimation (MLE).

Maximum Likelihood

What is Maximum Likelihood?

A statistical method for estimating population parameters (as the mean and variance) from sample data that selects as estimates those parameter values maximizing the probability of obtaining the observed data.

Merriam Webster

Binomial (Coin Flipping) Example

$$f(k;n,p) = \Pr(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

for $k = 0, 1, 2, \ldots, n$, where

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}$$

Symbols*:* *n* number of coin flips *k* number of heads
p probability of a head

Binomial (Coin Flipping) Example

- I flip my coin 100 times (n = 100)
- I get 53 heads and 47 tails (k=53)
- What is the best estimate of the parameter *p* in the Binomial Formula?
- The Maximum Likelihood Estimator (MLE) for *p* is the value of *p* that maximizes *f(k;n,p)*
- Easier to solve working with *log(f(k;n,p))*, the natural logarithm of the likelihood!
- Simple first year calculus

Binomial (Coin Flipping) Example

- $d(\log(f(k;n,p))/dp = 0$ (Calculus 101)
- Get: $d\,(k \log(p) + (n-k) \log(1-p)) = 0$
- Recall: $d \log(x) = 1/x$
- Get: $k/p + (n-k)/(1 - p) = 0$
- Next: $k\,(1-p) = (n-k)\,p$
- Then: $k - k\,p = n\,p - k\,p$
- *Then:* $k = n\,p$
- Therefore, the Maximum Likelihood Estimator (MLE) $p = k/n$

$$P = 0.53\ (53/100)$$

Maximum Likelihood Estimator

- “Common Sense”
- Basic concept predates Fisher
  - Chauvenet's derivation of least squares in 1891
- Fisher generalizes maximum likelihood to all estimation problems not just least squares.
- Usually right where frequentist assumptions are valid.
  - Fisher's rivals found a few pathological cases where the Maximum Likelihood Estimator (MLE) is actually *wrong*!

Maximum Likelihood

- Binomial (Coin Flipping) Example is Simple
- Becomes more complex, subtle issues with multi-dimensional data, multiple parameters in the model etc.
- More on this to come

Maximum Likelihood and Classification

- the action or process of classifying something according to shared qualities or characteristics.
  - Male versus Female
  - Cat versus Dog
  - Friend versus Foe
  - Paying Customer versus Deadbeat
  - “I scream” versus “ice cream”
- The maximum likelihood classifier is one of the most popular methods of classification, in which a data sample with the maximum likelihood is classified into the corresponding class.

Two Classes of Coins

- Fair Coins – probability of heads is 0.5
- Biased Coins – probability of heads is 0.9
- Equal Chance of fair or biased coin
- How to classify a coin based on coin flipping data?
- Based on one coin flip?
  - Heads (Maximum Likelihood classifier is Biased)
  - Tails (Maximum Likelihood classifier is Fair)

Probability of Heads

- Fair Coin (p = 0.5 times 0.5 = 0.25)
- Biased Coin (p = 0.5 times 0.9 = 0.45)
- Bayes Formula
  - p(Biased|Heads) = 0.45/ (0.25 + 0.45) = 0.69 (69 percent)
  - p(Fair|Heads) = 0.25/ (0.25 + 0.45) = 0.38 (38 percent)

Probability of Tails

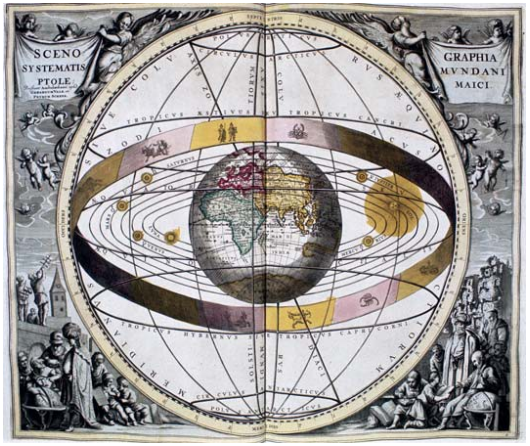
- Fair Coin (p = 0.5 times 0.5 = 0.25)
- Biased Coin (p = 0.5 times 0.1 = 0.05)
- Bayes Formula
  - p(Fair|Tails) = 0.25/(0.25 + 0.05) = 0.83 (83 percent)
  - p(Biased|Tails) = 0.05/(0.25 + 0.05) = 0.17 (17 percent)

More Advanced Classification Problems

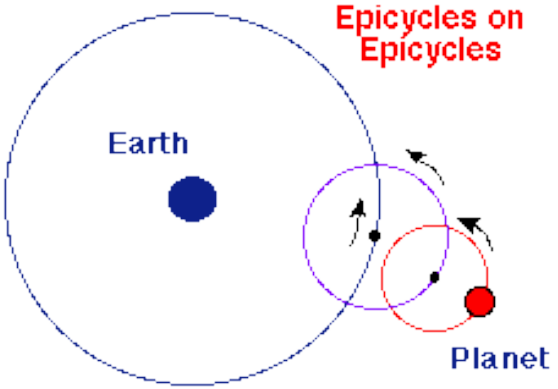
- Is a peak the Higgs particle, “background,” or something else (e.g. supersymmetry or SUSY)?
- Is an utterance “I scream” or “ice cream”?
- Is an utterance “media rights” or “meteorites”?
- Is an utterance “to,' “too,” or two?
- Is a swipe on smart keyboard “yes” or “Yrs,” the abbreviation for Yours or Years? Or “us,” very similar on standard QWERTY keyboard layout?

Early Mathematical Modeling

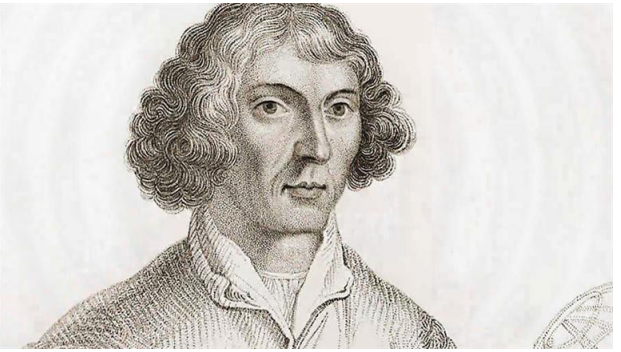
Early Mathematical Modeling



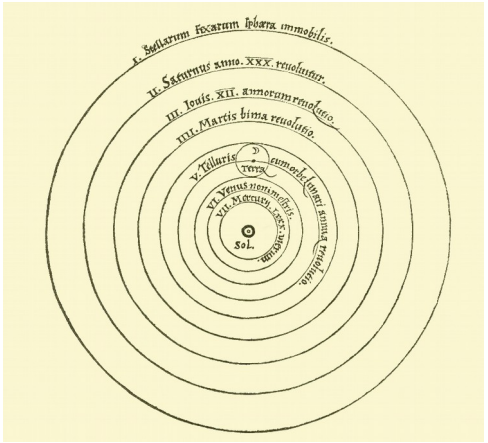
Geocentric Model



Uniform Circular Motion

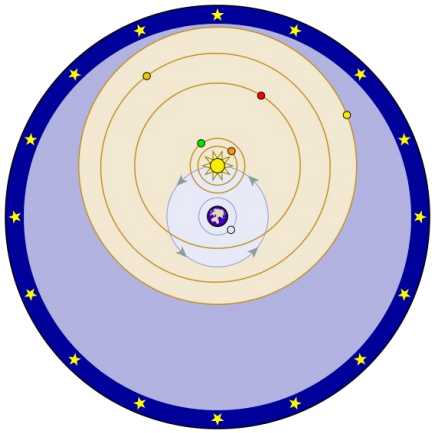


Copernicus



Heliocentric Model

Tyco Brahe

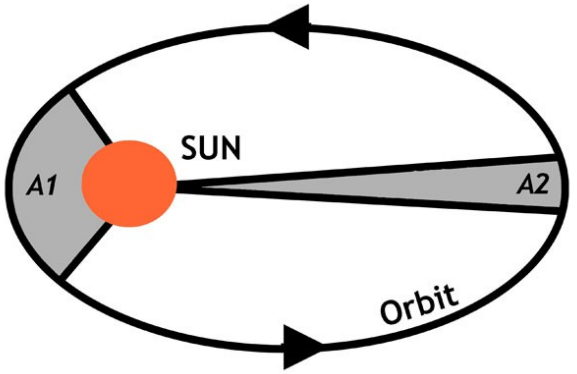


Hybrid Model

Early Mathematical Modeling

Rudolf II

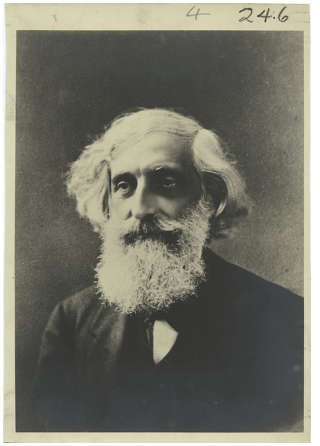
Kepler



Non-Uniform Elliptical Orbits

Legendre

Gauss

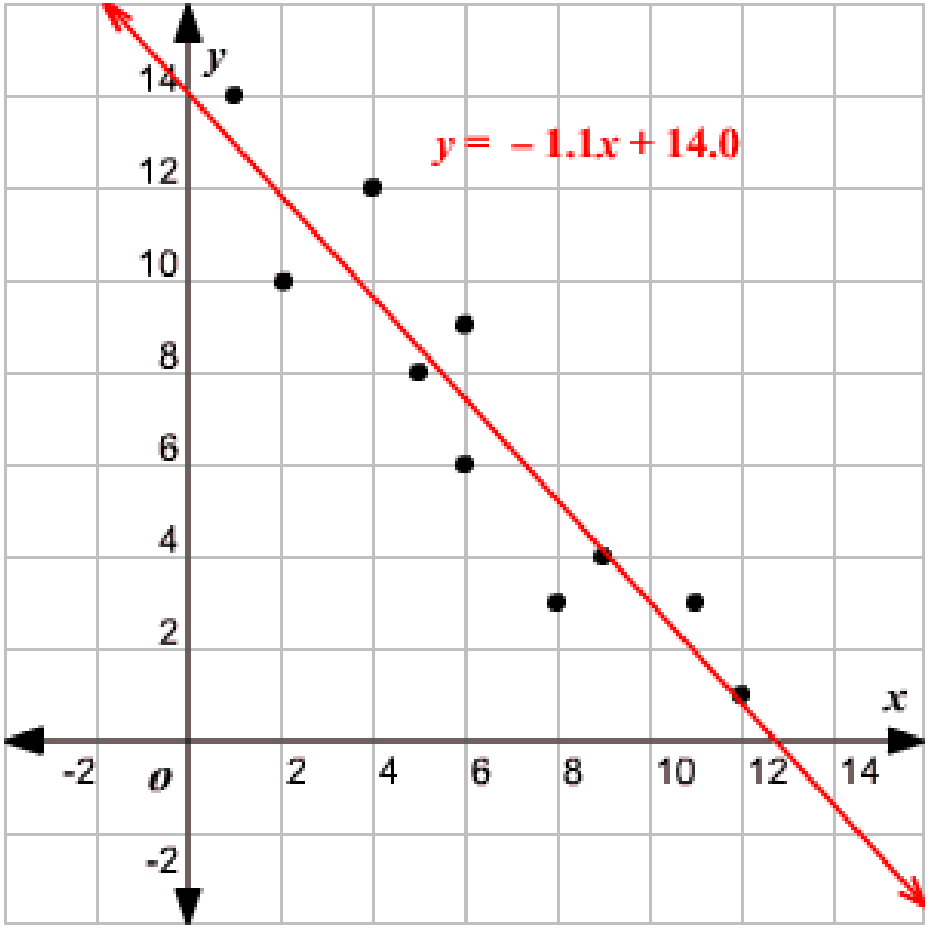


Chauvenet

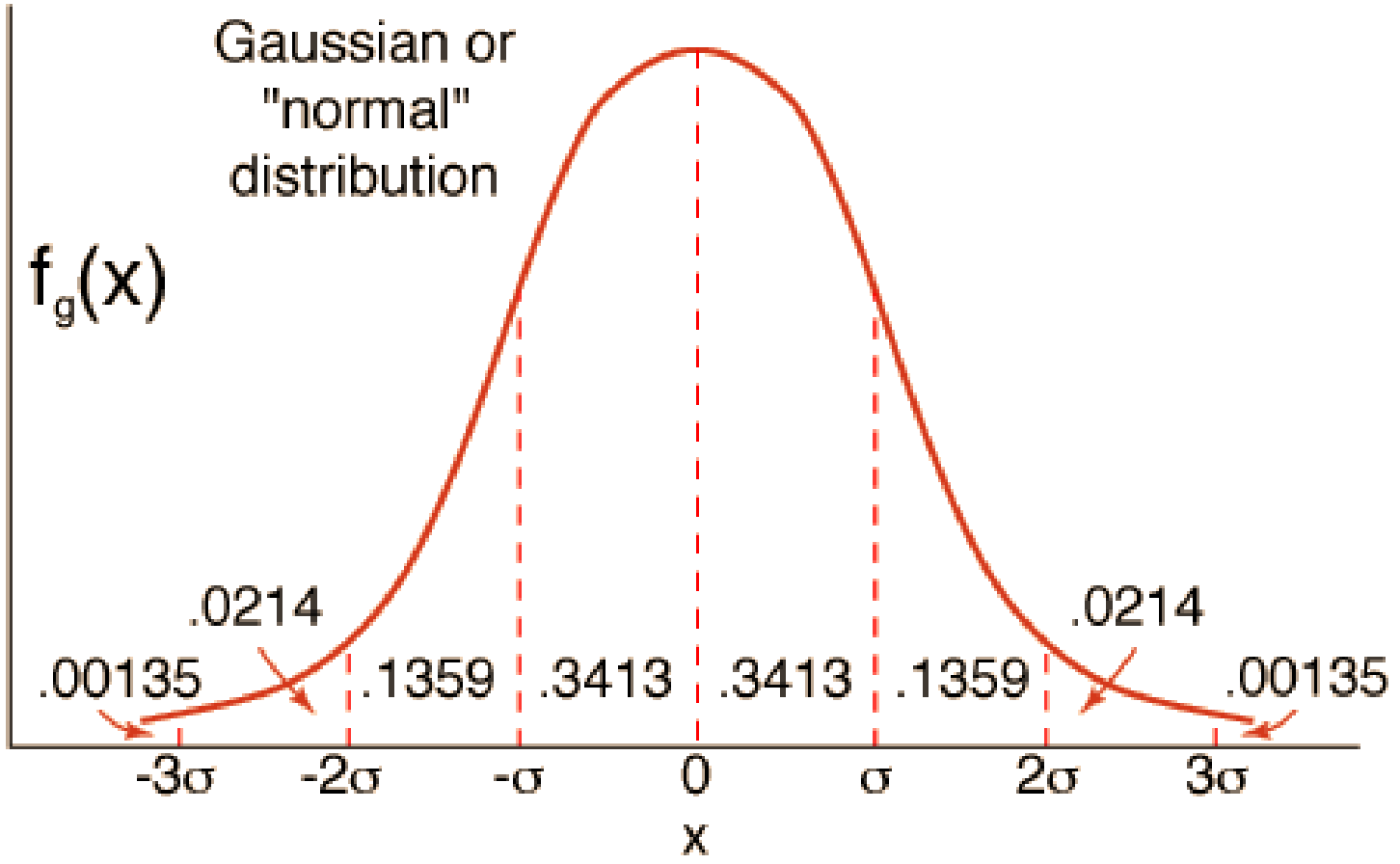
Least Squares Fitting

- In 1800s, scientists and mathematicians including Carl Friedrich Gauss and Adrien-Marie Legendre (1805) develop *method of least squares* to fit orbital models to the motion of various planets, moons, and asteroids.
- Minimize *the sum of squares of the errors* between the position of the planet predicted by the model and the actual measured position.
- Will turn out to be equivalent to Maximum Likelihood Estimation (MLE) when the measurement errors follow a strictly Gaussian/Normal/Bell Curve distribution. *Often not a realistic assumption!*
- Least Squares Fitting can be made to work for astronomical and physical data by removing non-Gaussian outliers! Succeeds in astronomy and other fields.

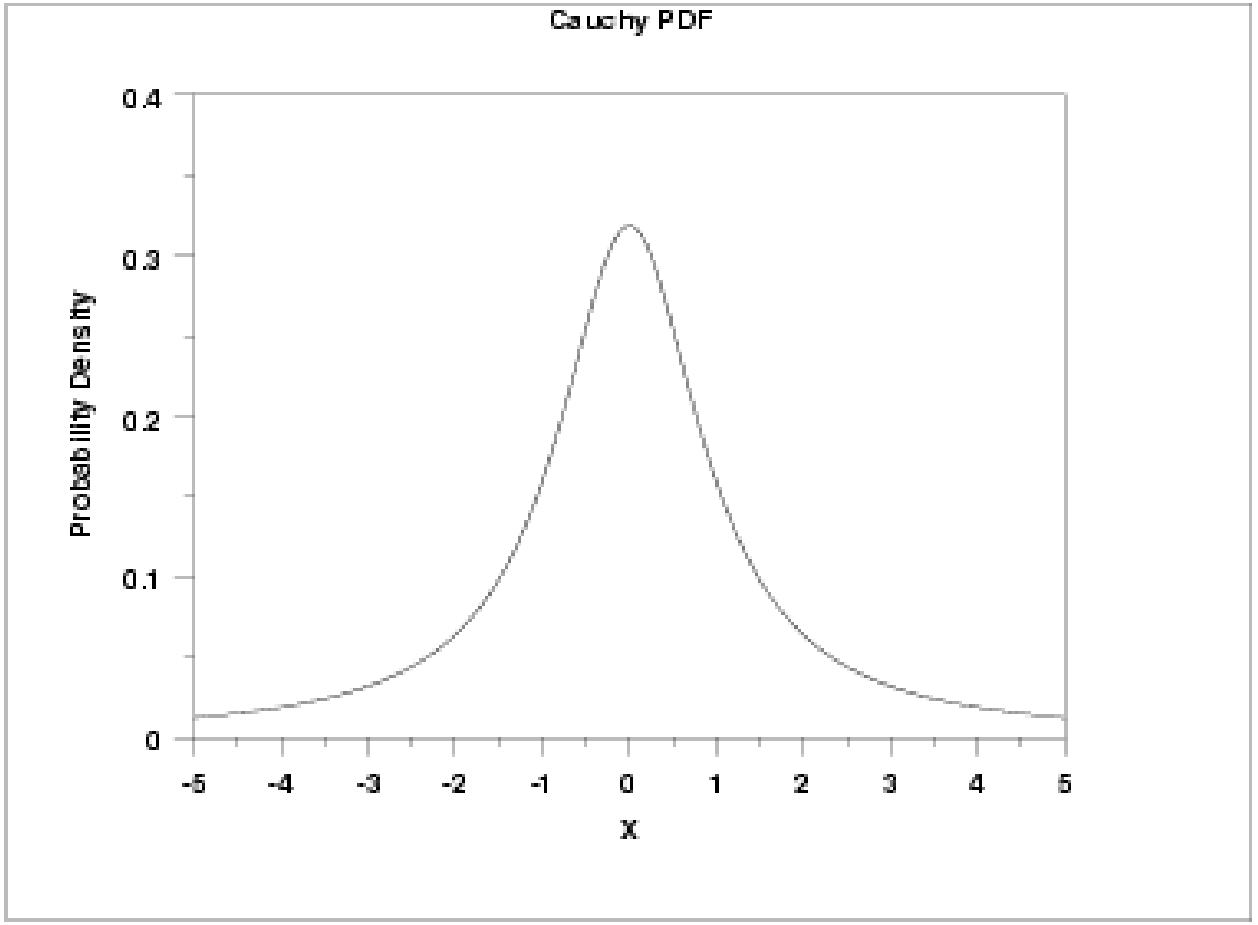
Least Squares Fit



Gaussian/Normal/Bell Curve



Cauchy-Lorentz Distribution

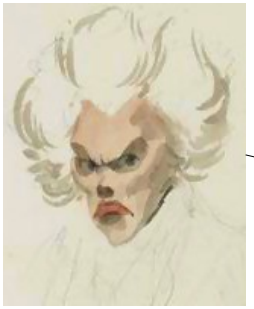


Ronald Fisher

Ronald Aylmer Fisher (1890 - 1962)

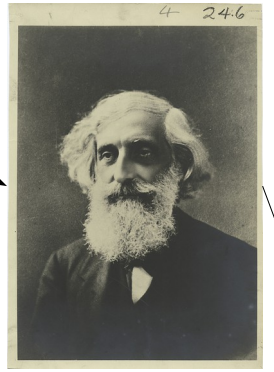
Fisher in 1913 at the start of his career.

Astronomy



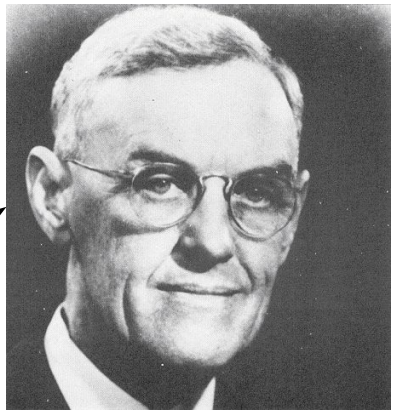
Legendre

Gauss



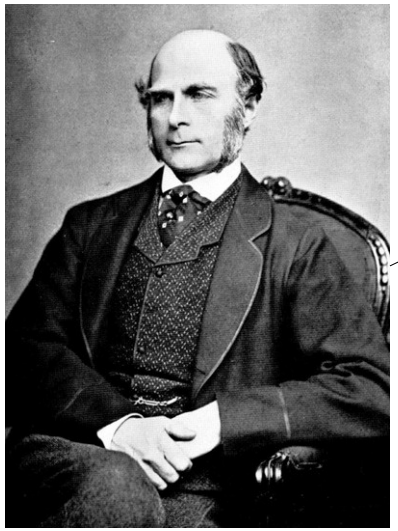
Chauvenet

Ronald A. Fisher

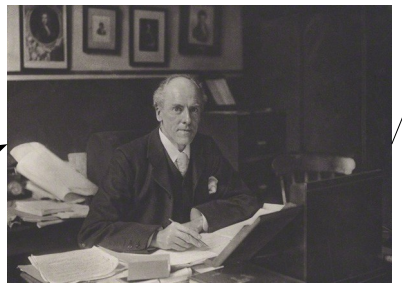


Snedecor (USA)

- Agriculture
- Economics
- Finance
- Physics
- Medicine



Francis Galton



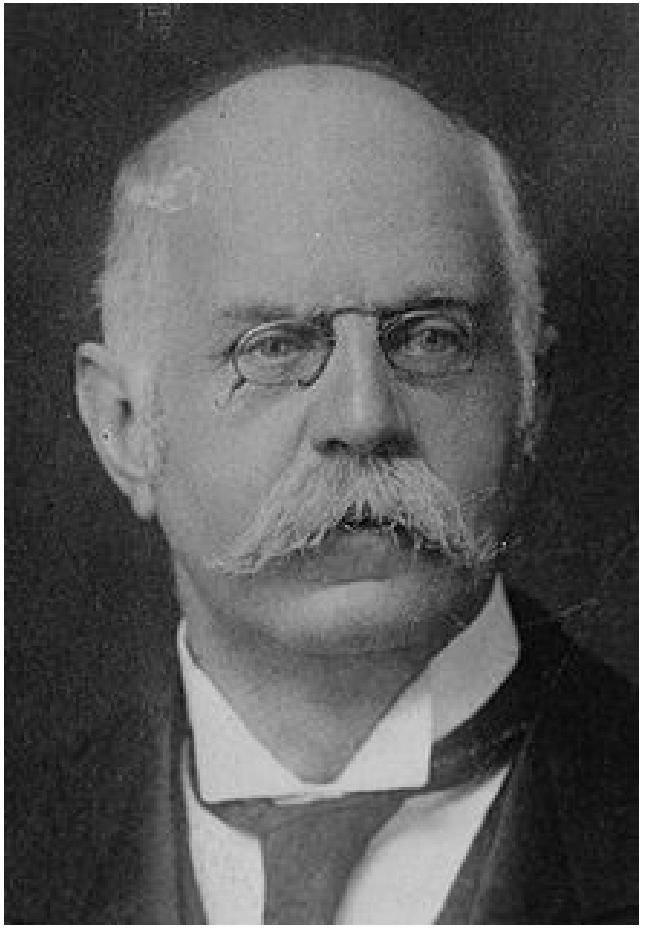
Karl Pearson

Eugenics

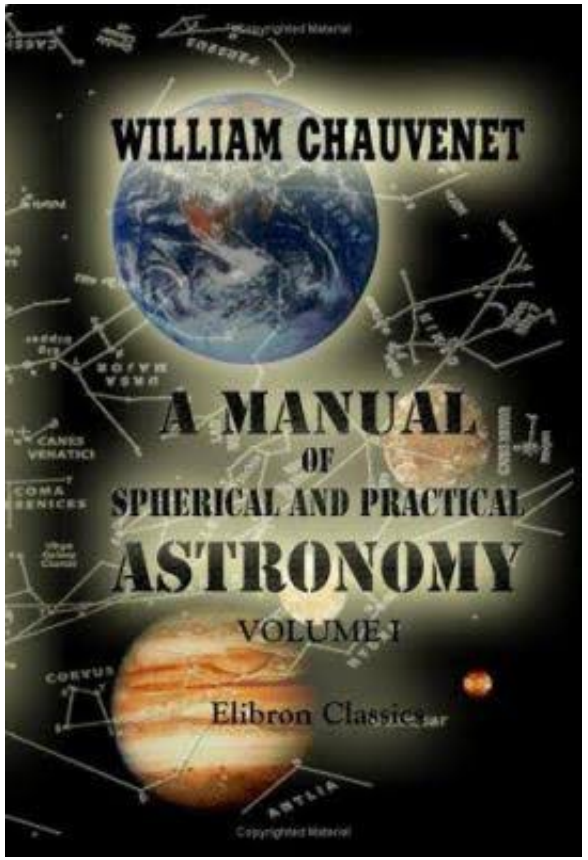
Ronald Fisher

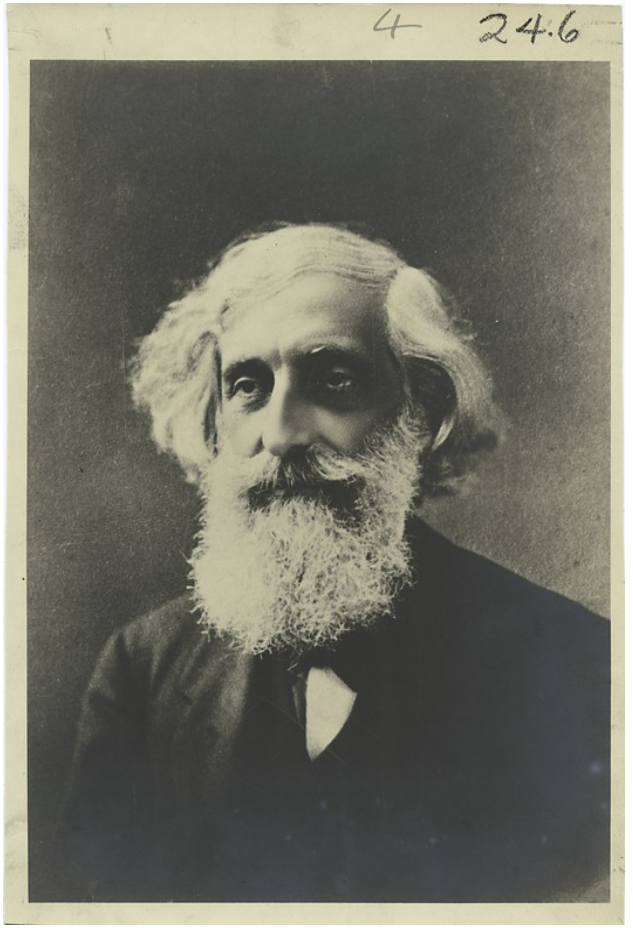
- Born 1890
- Attends the elite *Harrow School*, British “Public School,” from about 1904 to 1909
- Attends *Gonville and Caius College* at *Cambridge University*, graduates in 1912 (First in Astronomy)
- Volunteers for World War I but bad eyesight keeps him out of army.
- Fisher worked for six years as a statistician in the City of London and taught physics and maths at a sequence of public schools, and at the *Thames Nautical Training College*, and *Bradfield College*
- Takes Job at *Rothamsted Research* in 1919 until 1933
- Seminal works on Maximum Likelihood from 1912 to 1922

Major Leonard Darwin



Fisher Starts with Astronomy



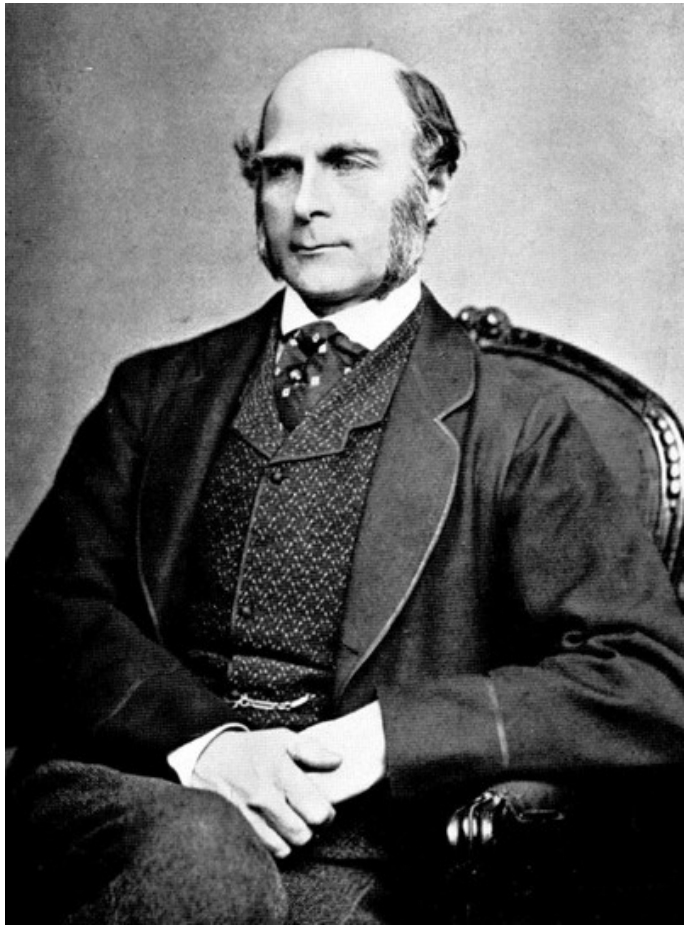


William Chauvenet

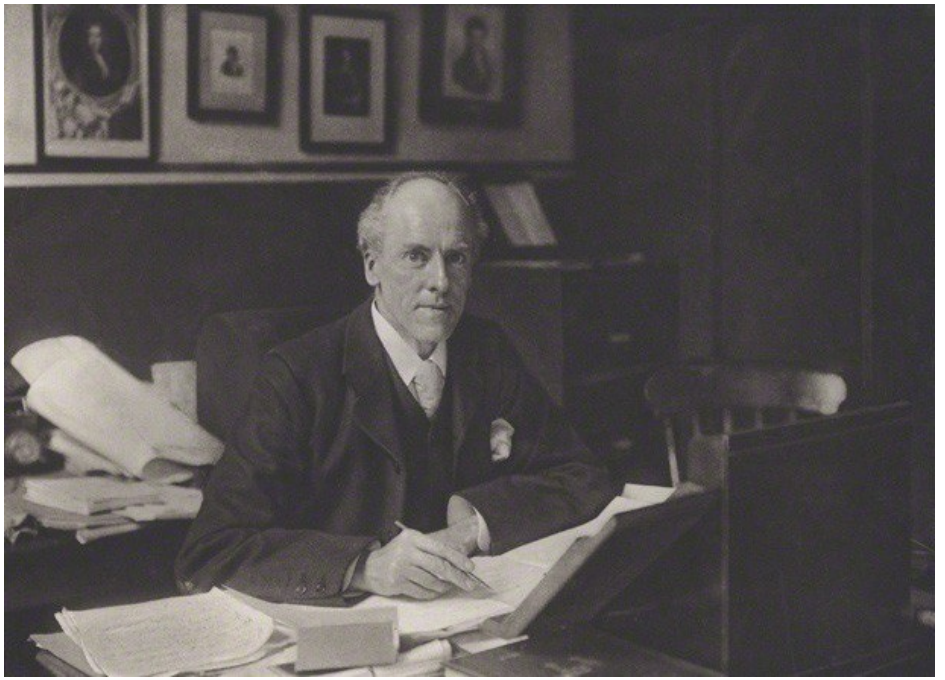
Fisher's Maximum Likelihood Papers

- FISHER, R. A. (1912) **On an absolute criterion for fitting frequency curves.** *Messenger of Mathematics* 41 155-160
- FISHER, R. A. (1915) **Frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population.** *Biometrika* 10 507-521
- FISHER, R. A. (1918). **The correlation between relatives on the supposition of Mendelian inheritance.** *Transactions of the Royal Society of Edinburgh* 52 399-433.
- FISHER, R. A. (1920) **A mathematical examination of the methods of determining the accuracy of an observation by the mean error, and by the mean square error.** *Monthly Notices of the Royal Astronomical Society* 80 758-770
- FISHER, R. A. (1921) **On the "probable error" of a coefficient of correlation deduced from a small sample.** *Metron* 1 3-32.
- FISHER, R. A. (1922) **On the mathematical foundations of theoretical statistics.** *Philos. Trans. Roy. Soc. London* Ser. A 222 309-368.
- FISHER, R. A. (1922b) **The goodness of fit of regression formulae, and the distribution of regression coefficients.** *J. Roy. Statist. Soc.* 85, 597-612

Francis Galton and Karl Pearson



Francis Galton



Karl Pearson

Egon Pearson and Jerzy Neyman

Egon Pearson

Jerzy Neyman

The Bitter Feud

- Karl Pearson, the Old Lion, and Ronald Fisher, the young upstart, become embroiled in a bitter feud.
  - Fisher revives and improves what he eventually calls “Maximum Likelihood”
  - Fisher rejects Pearson's “Inverse Probability” and what Fisher later calls “Bayesian Probability”
  - Fisher revives and modernizes what is now called “frequentist” or classical statistics
  - Pearson criticizes and in some cases rejects some of Fisher's key papers; Fisher cries foul.

The Bitter Feud II

- The infamous feud continues even after Karl Pearson's death in 1936 as a feud between Fisher and Pearson's son Egon Pearson and Karl Pearson's protege Jerzy Neyman, who became a professor at Berkeley in 1938:
  - Colors development of statistics with arcane disputes between Fisher and Pearson-Neyman over foundations of probability and statistics, terminology, and *credit*.
  - Elaborate efforts by Jerzy Neyman to find obscure cases where Maximum Likelihood Estimators are incorrect.
  - *Can be hard to distinguish genuine scholarship from petty squabbling and infighting in their works.*

Fisher and His Foes

- Fisher had a bad temper.
- Fisher was self-righteous, arrogant, rude.
- *“Did not suffer fools lightly.” (Joan Fisher, daughter)*
- Fisher was phenomenally successful in the small world of probability and statistics.
- His success engendered considerable *jealousy* and resentment among his rivals.

Maximum Likelihood in the Computer Era

Maximum Likelihood: Recap

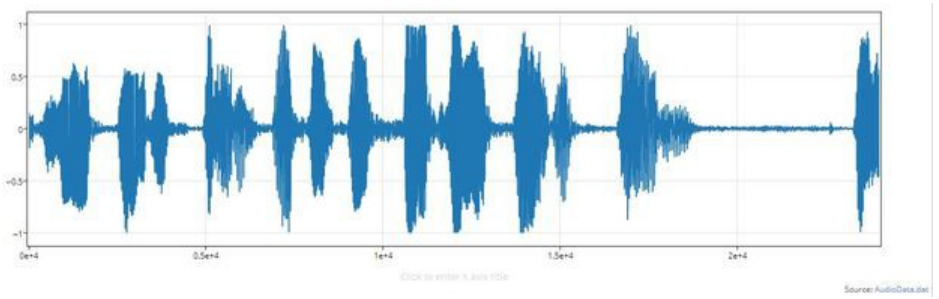
A statistical method for estimating population parameters (as the mean and variance) from sample data that selects as estimates those parameter values maximizing the probability of obtaining the observed data.

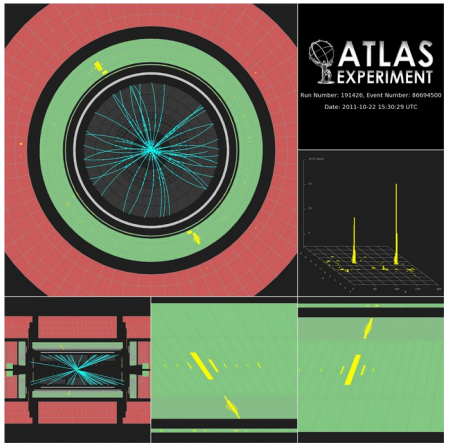
Merriam Webster

Maximum Likelihood

- Works well for simple cases like the Binomial Distribution (Coin Flipping)
- One dimensional data
- One or a few parameters in the model (location and dispersion, for example)
- Clean data (no outliers, other problems)
- Usually computationally tractable for these cases
- Often real-time on modern CPU's, DSP's

Higher Dimensional Data





Maximum Likelihood Parameter Estimation

Data

Pre Process Data

Model Selection

Fit Model to Data using Maximum Likelihood

Evaluate Goodness of Fit

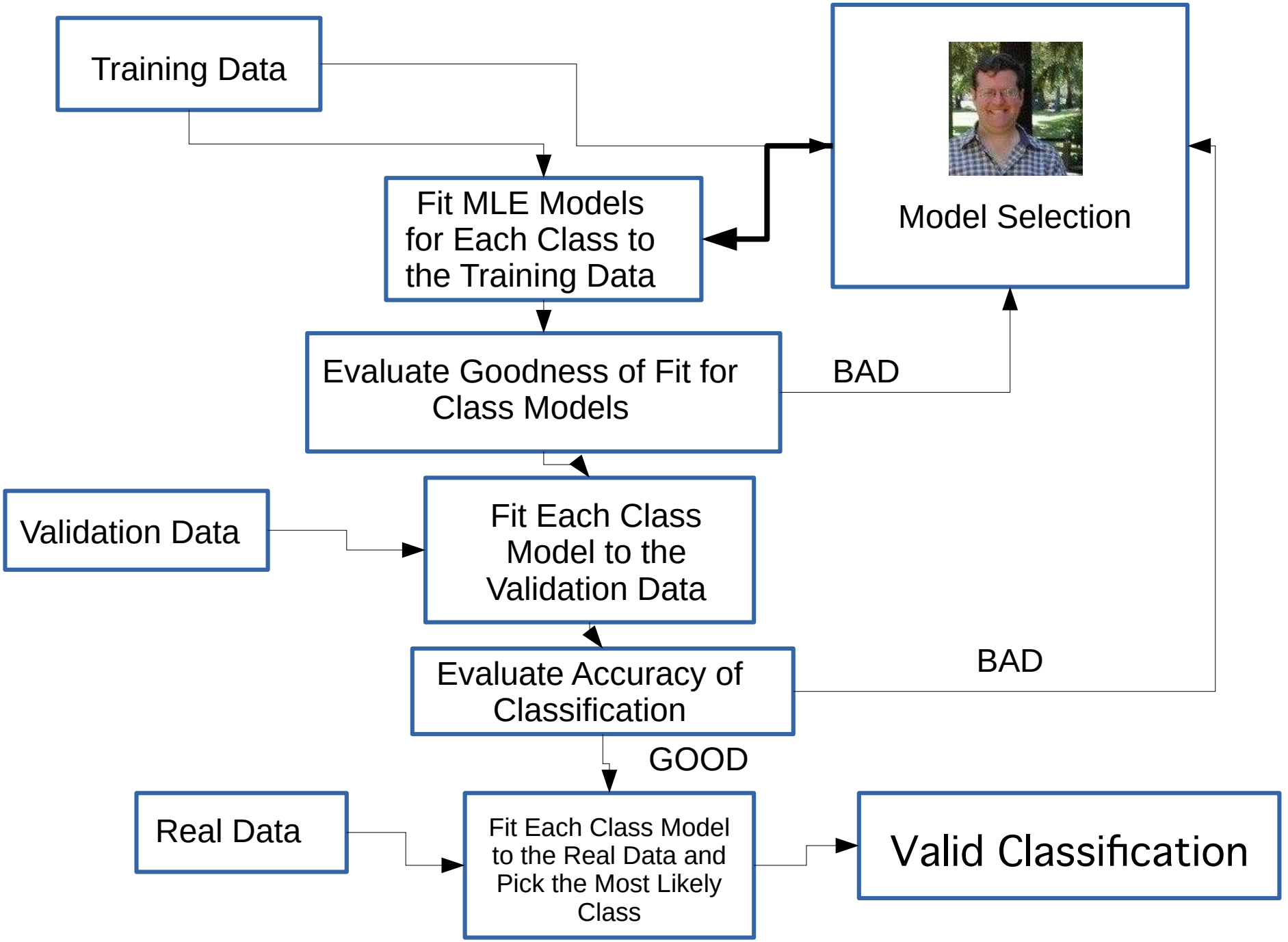
BAD

GOOD

Parameter Estimates

Maximum Likelihood Classifiers

Maximum Likelihood Classifier



Maximum Likelihood Classifiers

CAT

DOG

Class Models (e.g. “DOG”)
Typically have Fit Parameters
(e.g. “SIZE OF DOG”)
Determined During Recognition

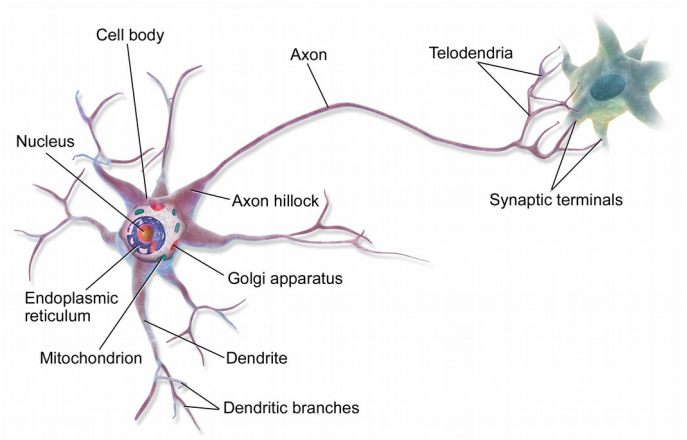
Mammal_Model(Pointiness_of_Ears, Length_of_Whiskers, Eye_shape, Height, Weight, Hair_Color,...)

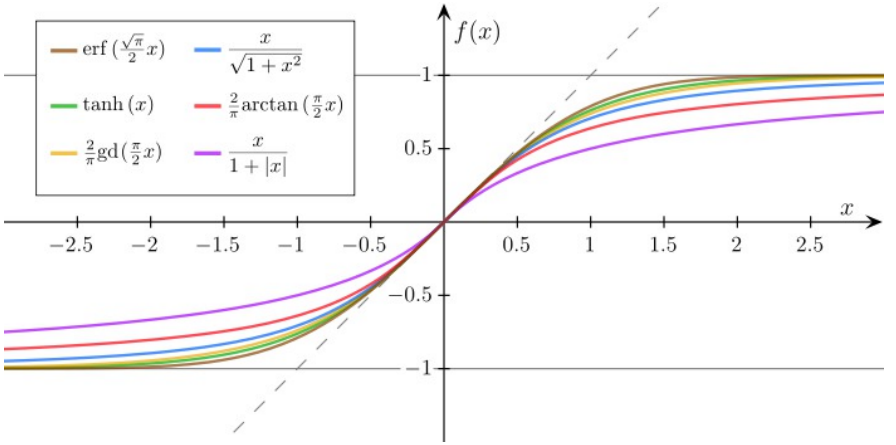
Fit and Fix Parameters Associated with Class Membership in ***Training***

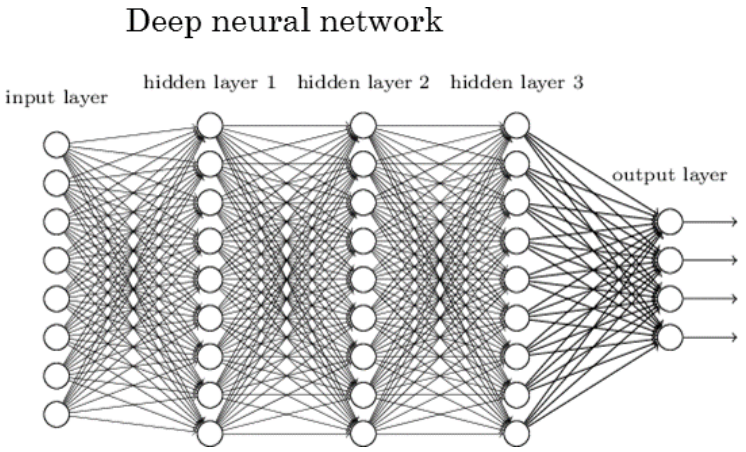
Fit for Independent Parameters such as Size During ***Recognition***

How Does Maximum Likelihood Compare to Deep Learning?

Deep Learning







- Formerly known as Artificial Neural Networks
- Based on a *very* simplified model of neurons in the human brain.

MLE

- Conceptual Understanding of Class (What makes a Dog a Dog?)
- Closer to Semantics (Meaning) in Human Mind
- Basis in Probability Theory
- Can be simple (Newton's Law of Gravitation)
- Route to simplification if complex (epicycles → Newton's Law of Gravitation)
- May not need to retrain if circumstances change, e.g. new sensor
- Human expert chooses models based on theory and experiment

Deep Learning

- Black Box
- Arguably closer to low level structure of Human Brain (physical neurons)
- Huge Number of Parameters (connection weights between neurons)
- No obvious connection between connection weights and physical reality
- Often need to retrain if circumstances change
- Human expert chooses input features, wiring of the neural network, transfer functions of neurons, and other low level characteristics and prays!

Problem Areas for All AI

Consciousness

Flashes of Insight and Changes of World View

Human Beings can often explain an intuition under close questioning, during dialog with other people, or after some contemplation.

Conceptual understanding seems to precede ability to verbalize: may need to find or invent new words or phrases to explain the new concept or concepts.

Does the Human Brain and Its Neurons Work the Way Mainstream Science Thinks?

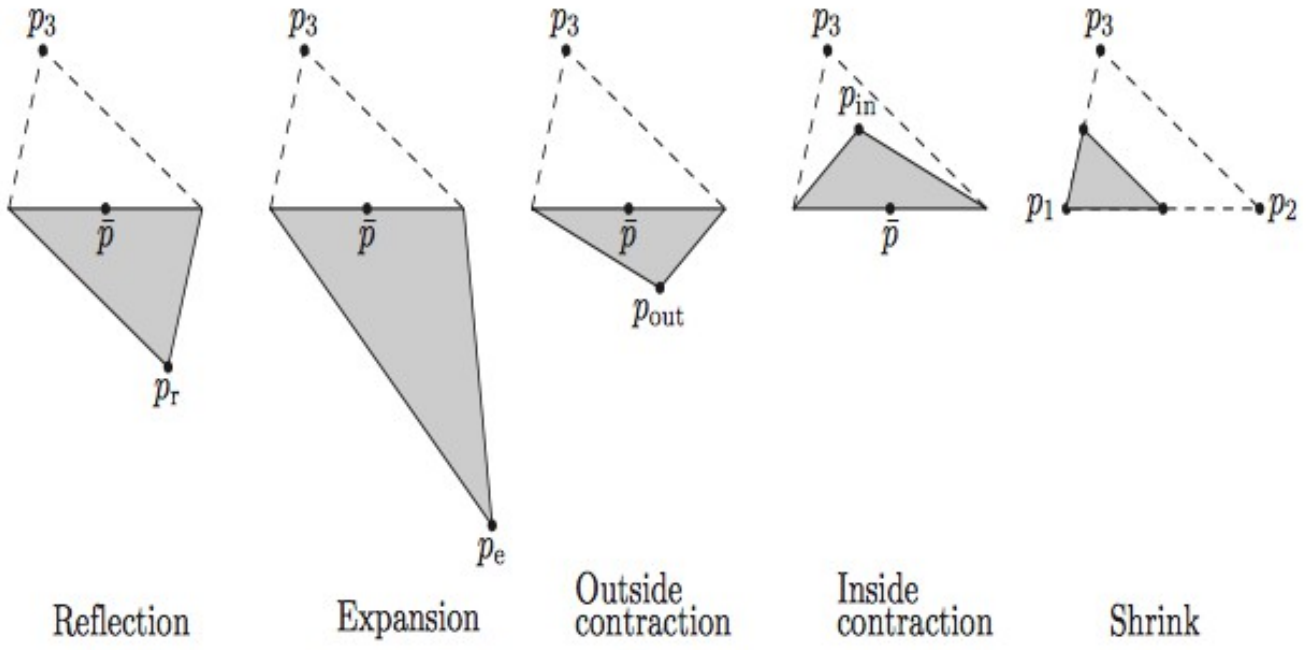
Complex Multiparameter Models

- Mass, width, form factors, coupling factors in particle physics.
- Hidden Markov Model (HMM) speech recognition models have hundreds of thousands of parameters fitted by the training/learning process.
- Often combined with large data sets ranging from Gigabytes (speech/audio, video) to Petabytes (one million gigabytes)
  - One DVD movie is about 4 Gigabytes (compressed)
  - A Petabyte is about 250,000 movies.

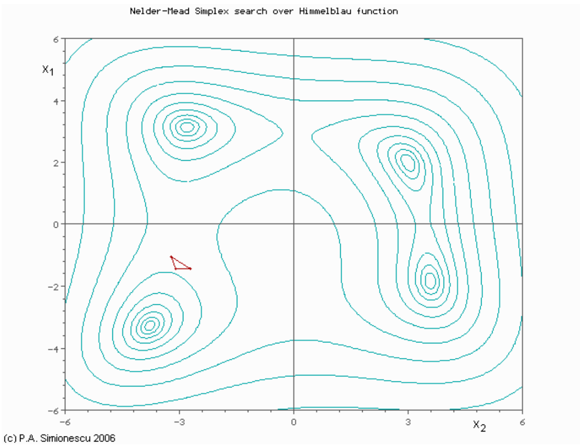
Numerical Methods and Powerful Computers Needed

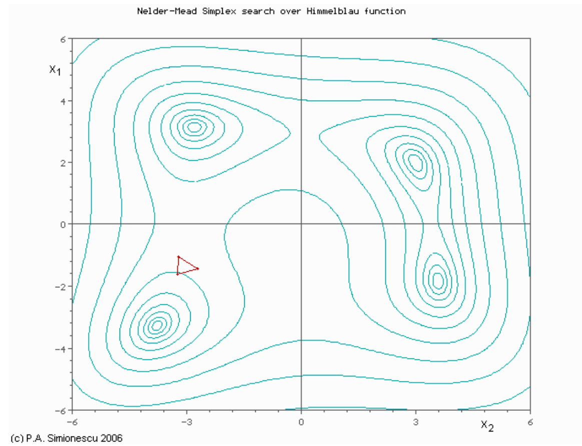
- Closed form solutions from calculus are impossible or very difficult to derive – unlike simple Binomial Distribution case.
  - Probability *p* = (number of heads)/(number of coin flips)
- Need to use non-linear optimization methods applied to data to find the maximum likelihood estimates of the model parameters. Search for set of model parameters that maximize likelihood for the data set.
  - Nelder-Mead/polytope/”simplex” method
  - Levenberg-Marquardt
  - Other methods

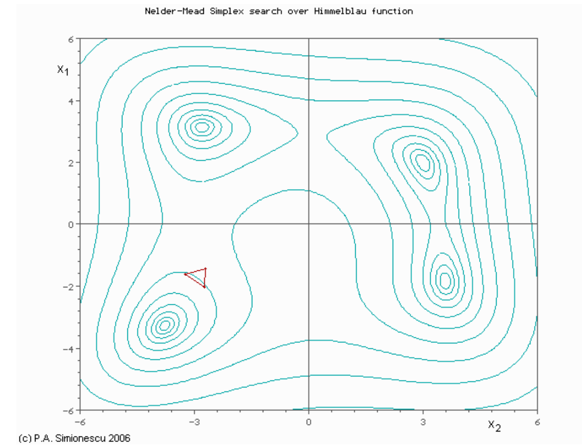
Nelder-Mead

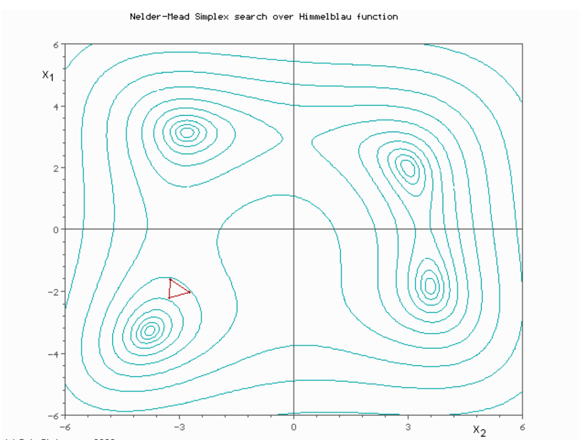

p_3
$\bar{p}$
p_r
Reflection
p_3
$\bar{p}$
p_e
Expansion
p_3
$\bar{p}$
p_{out}
Outside contraction
p_3
p_{in}
$\bar{p}$
Inside contraction
p_3
p_1
p_2
Shrink

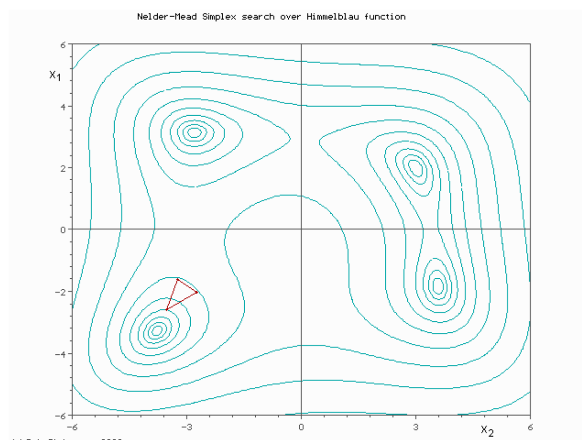
Nelder-Mead Fitting

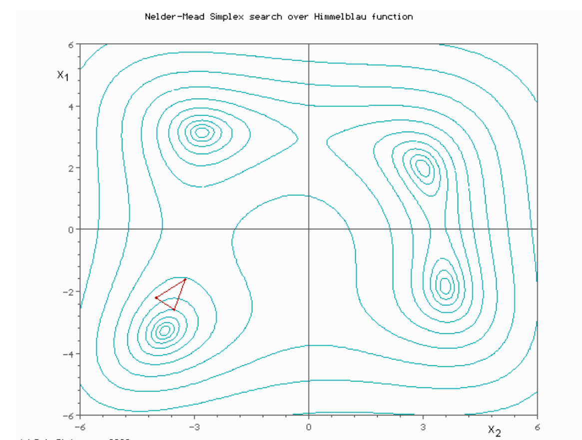




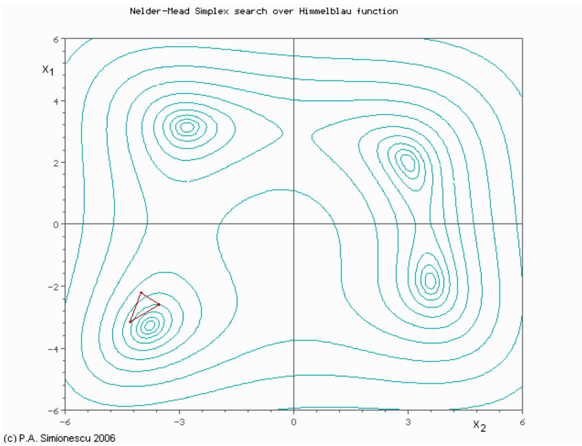


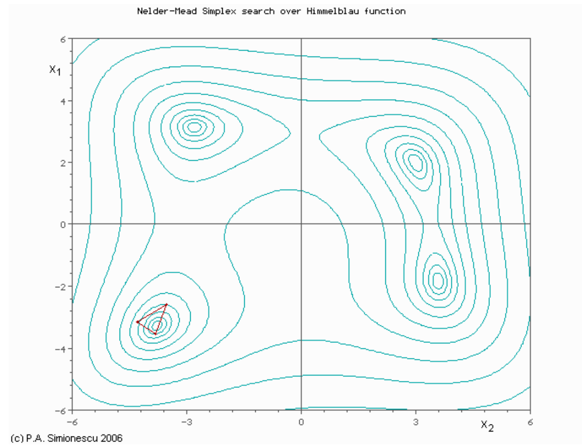


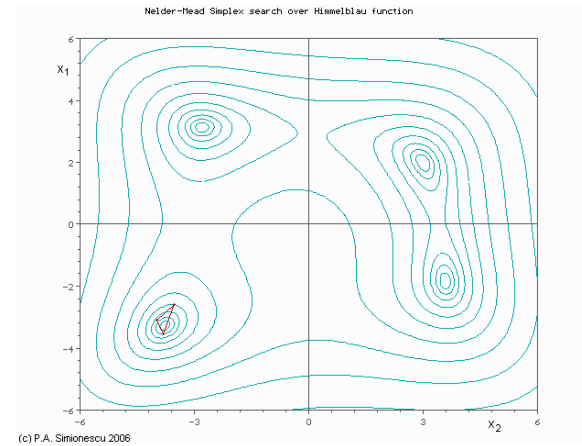


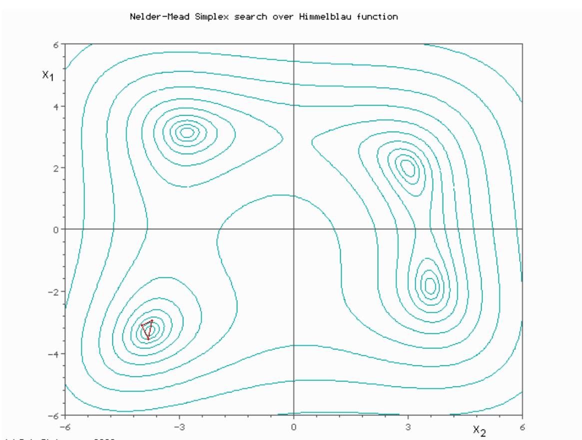


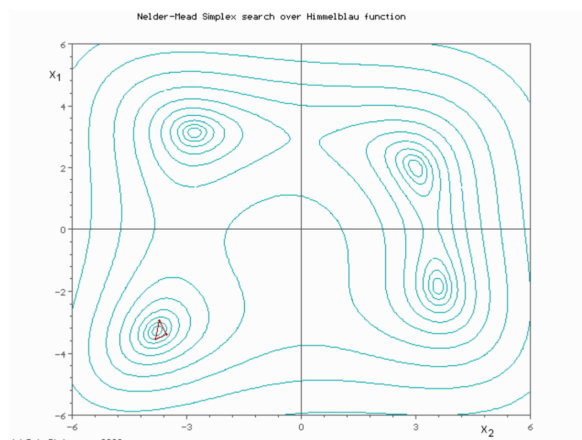
Nelder-Mead Fitting

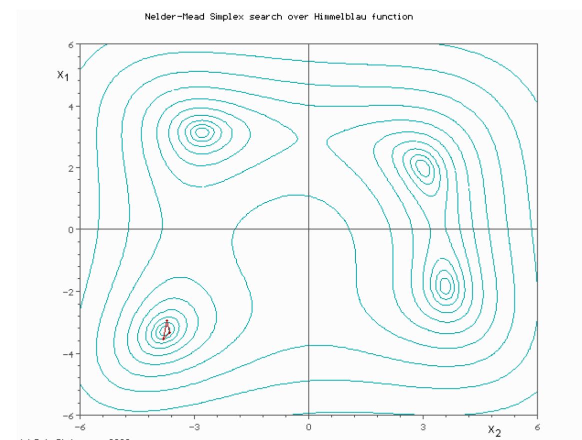




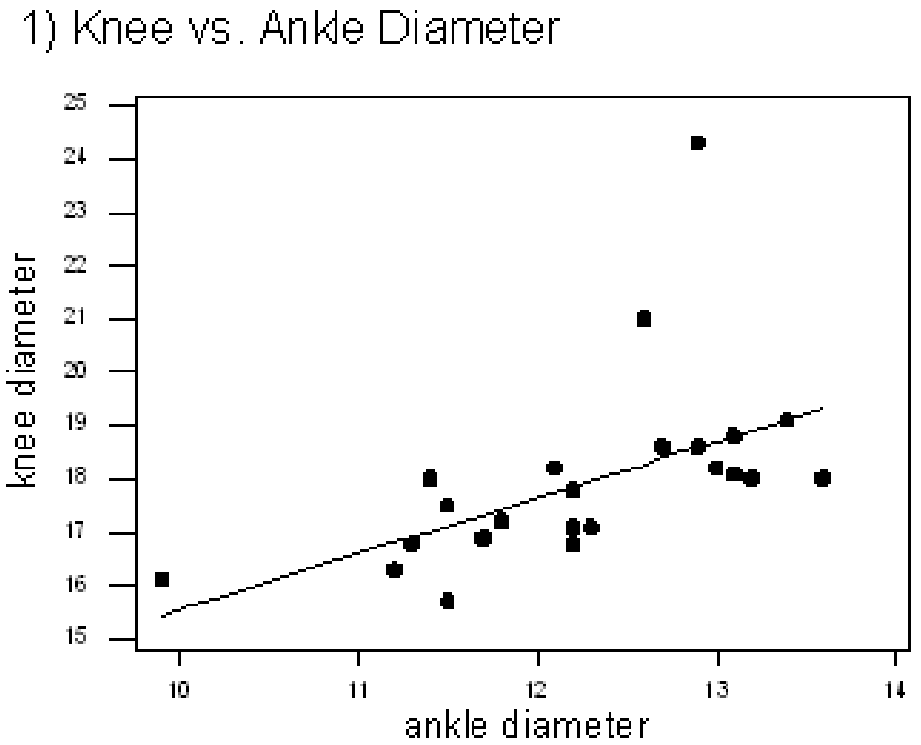








MLE is not Robust



MLE is not Robust

- MLE is not robust. Vulnerable to outliers, data samples in regions where model predicts no or very little data.
- Solutions
  - Cleaning data
  - Robust Maximum Likelihood methods

Often Requires Optimization for Speed and Memory

- Multidimensional models, multi-parameter fits are often computationally intensive.
- Often not real-time.
- Can just take too long even with super computers
- May need to develop efficient, fast algorithms to get useful results.

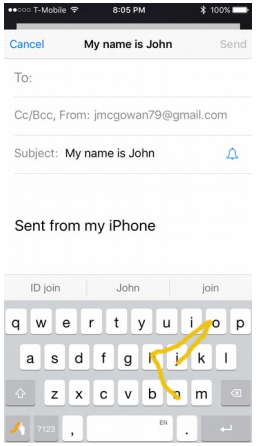
Distinguishing Between Models with Similar Data Distributions

- MLE can be poor at distinguishing between models with very similar distributions and observable consequences.
- The “nature” versus “nurture” problem that Galton, Pearson (Karl), and Fisher were never able to solve.
  - If exceptional intelligence is hereditary and intelligence leads to wealth, power, success, then intelligence will tend to run in wealthy families (natural nobility).
  - If exceptional intelligence is a consequence of expensive education and training, intelligence will tend to run in wealthy families (the Matthew effect).

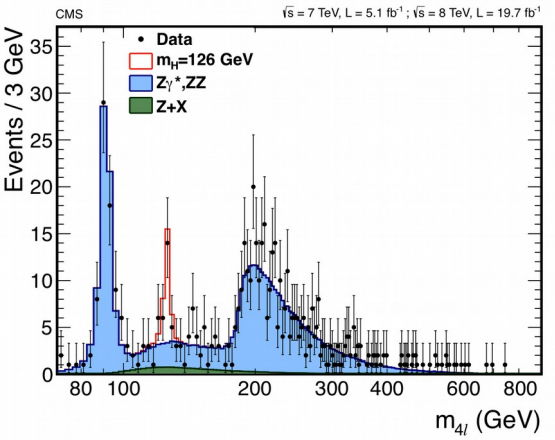
Technically Challenging

- MLE for high dimensional data and complex multi-parameter models requires substantial effort by experts.
  - Typically 1-20 man years for more advanced projects.
- Painstaking, nitpicky work. Requires attention to detail.
- “Black Art”
- Yields results – GPS, software keyboards for smart phones, speech recognition, Higgs Boson

Successes


T-Mobile
8:05 PM
100%
Cancel
My name is John
Send
To:
Cc/Bcc, From: jmcgowan79@gmail.com
Subject: My name is John
Sent from my iPhone
ID join
John
join


Nuance
Dragon
Professional
Individual
NUANCE


CMS
√s = 7 TeV, L = 5.1 fb-1 ; √s = 8 TeV, L = 19.7 fb-1
Data
mH=126 GeV
Zγ*,ZZ
Z+X
Events / 3 GeV
0
5
10
15
20
25
30
35
80
100
200
300
400
600
800
m4l (GeV)

Conclusion

- Maximum Likelihood Estimation (MLE) is a powerful, widely used statistical method.
- Many successes in both scientific and commercial worlds: GPS, Dragon Naturally Speaking, SPHINX, Swype, discovery of Higgs Boson.
- Remains technically challenging for multi-dimensional data and complex multi-parameter models – many real world problems.
- Some foundational issues such as the definition of probability remain unresolved. (e.g. Bayesian interpretation and prior can change results)

Questions?