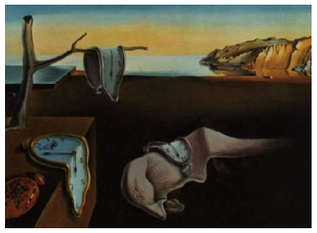
Statistics and Biomechanics

Todd Pataky
Department of Human Health Sciences
Kyoto University

- Statistics and Biomechanics are separate fields
- Statistics is under-represented in Biomechanics
- Statistics does not give us biomechanical meaning

“60% of the time, it works every time.”

–Brian Fantana, Anchorman



Journal of Statistical Software

July 2016, Volume 71, Issue 7. *doi: 10.18637/jss.v071.i07*

rft1d: Smooth One-Dimensional Random Field Upcrossing Probabilities in Python

Todd Pataky
Shinshu University

Abstract

Through topological expectations regarding smooth, thresholded n-dimensional Gaussian continua, random field theory (RFT) describes probabilities associated with both the field-wide maximum and threshold-surviving upcrossing geometry. A key application of RFT is a correction for multiple comparisons which affords field-level hypothesis testing for both univariate and multivariate fields. For unbroken isotropic fields just one parameter in addition to the mean and variance is required: the ratio of a field's size to its smoothness. Ironically the simplest manifestation of RFT (1D unbroken fields) has rarely surfaced in the literature, even during its foundational development in the late 1970s. This Python package implements 1D RFT primarily for exploring and validating RFT expectations, but also describes how it can be applied to yield statistical inferences regarding sets of experimental 1D fields.

Keywords: random field theory, Gaussian random fields, multivariate analysis, time series, continuum analysis.

Overview

- History
- The p value
- Classical techniques

Questions

- Emerging techniques
- Controversies
- The future

Questions

Goals

- Obtain something practical
- Actively (re-) learn
- New perspectives on Statistics topics

github.com

Search or jump to... / Pull requests Issues Marketplace Explore

0todd0000 / ISB2019StatisticsTutorial

Watch 0 Star 0 Fork 0

Code Issues 0 Pull requests 0 Projects 0 Wiki Security Insights Settings

Materials for the "Statistics and Biomechanics" tutorial at ISB 2019

Edit

Manage topics

14 commits 1 branch 0 releases 1 contributor MIT

Branch: master New pull request Create new file Upload files Find File Clone or download

Todd Pataky Update example Slides notebooks Latest commit 1c561c6 1 minute ago

Notebooks	Add rendered HTML files	yesterday
Slides	Update example Slides notebooks	1 minute ago
Supplementary	Add supplementary PDFs	yesterday
.gitignore	Add gitignore and README	3 days ago
LICENSE	Initial commit	3 days ago
README.md	Translate notebooks to Matlab	yesterday

README.md

ISB2019StatisticsTutorial

This repository contains materials for the tutorial "Statistics and Biomechanics", held on 31 July 2019 at XXVII Congress of the International Society of Biomechanics in Calgary.

- ISB image: © International Society of Biomechanics. Linked from Wikipedia.
- Coin images: © Royal Canadian Mint. Linked from Wikipedia (heads and tails). Links are provided for educational purposes (to illustrate coin flipping probability) and for promoting Canada to an international congress audience. Consult the Royal Canadian Mint's Intellectual Property Guidelines for details regarding fair use.


jupyter


Kahoot!


jupyter

localhost

File Edit View Run Kernel Tabs Settings Help

Launcher | ExampleNotebookMatlab.i

Markdown | Matlab

Name	Last Modified
_example	14 days ago
dt	a month ago
ExampleNotebookMatl...	2 minutes ago
Dorn2012.csv	a month ago
mwarp1d.npz	20 days ago
randn1d.m	10 months ago
Screen Shot 2019-07-...	a minute ago

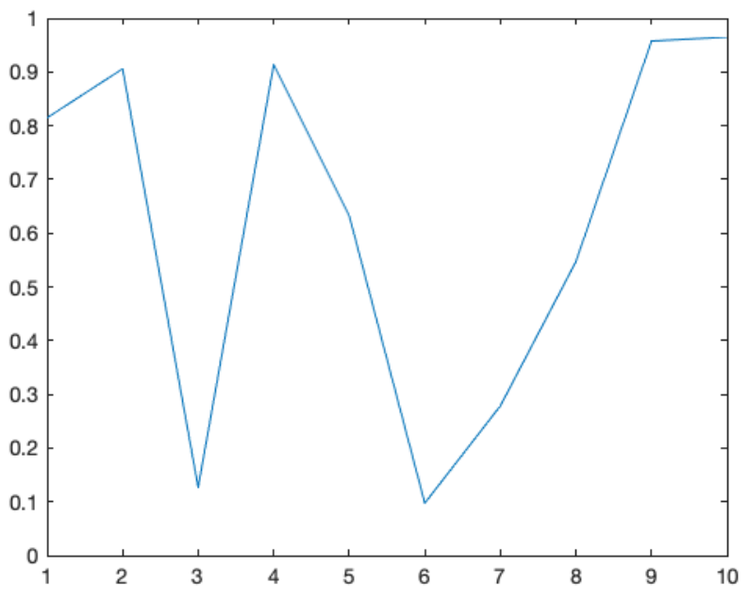
Eample MATLAB Notebook

- Markdown (easy-to use HTML)
- Interactive MATLAB!

- ISB image: © International Society of Biomechanics. Linked from Wikipedia

[1]:
```
x = rand(10, 1);
plot(x)
```

Output View



DOWNLOAD AND INSTALL THE APP!!!

Introductory Quiz

Overview

- History
- The p value
- Classical techniques
- Emerging techniques
- Controversies
- The future

Questions

Questions

“It is easy to lie with statistics.
It is hard to tell the truth without statistics.”

–Andrejs Dunkels, Prof. of Mathematics

What is Statistics?

The science of uncertainty.

The mathematics of Science.

Why statistics?

Because our intuition is poor.

Alan Smith, Financial Times

Why you should love statistics

(TED Talk, 2017)

UK Survey:

What is the UK teenage pregnancy rate?

Survey result: 15%

Reality: 0.6%

Alan Smith, Financial Times

Why you should love statistics

(TED Talk, 2017)

Japan Survey:

What proportion of the population lives in rural areas in Japan?

Survey result: 56%

Reality: 7%

Alan Smith, Financial Times

Why you should love statistics

(TED Talk, 2017)

Why statistics?

Because objectivity is difficult.

“Machines have objectivity, humans have passion.”

–Garry Kasparov, Chess Grandmaster

Don’t fear intellegent machines. Work with them.
(TED Talk, 2017)

5

Your statistics experience

History of the t test

Timeline of statistics

Statistics is about gathering data and working out what the numbers can tell us. From the earliest farmer estimating whether he had enough grain to last the winter to the scientists of the Large Hadron Collider confirming the probable existence of new particles, people have always been making inferences from data. Statistical tools like the mean or average summarise data, and standard deviations measure how much variation there is within a set of numbers. Frequency distributions - the patterns within the numbers or the shapes they make when drawn on a graph - can help predict future events. Knowing how sure or how uncertain your estimates are is a key part of statistics.

Today vast amounts of digital data are transforming the world and the way we live in it. Statistical methods and theories are used everywhere, from health, science and business to managing traffic and studying sustainability and climate change. No sensible decision is made without analysing the data. The way we handle that data and draw conclusions from it uses methods whose origins and progress are charted here.

Julian Champkin
Significance magazine

Early beginnings

450 BC Hippias of Elis uses the average value of the length of a king's reign (the mean) to work out the date of the first Olympic Games, some 300 years before his time.

400 BC In the Indian epic the *Mahabharata*, King Rtuparna estimates the number of fruit and leaves (2095 fruit and 50 000 000 leaves) on two great branches of a vibhitaka tree by counting the number on a single twig, then multiplying by the number of twigs. The estimate is found to be very close to the actual number. This is the first recorded example of sampling – "but this knowledge is kept secret", says the account.

AD 7 Census by Quirinus, governor of the Roman province of Judea, is mentioned in Luke's Gospel as causing Joseph and Mary to travel to Bethlehem to be taxed.

10th century The earliest known graph, in a commentary on a book by Cicero, shows the movements of the planets through the zodiac. It is apparently intended for use in monastery schools.

1188 Gerald of Wales completed the first population census of Wales.

1303 A Chinese diagram entitled "The Old Method Chart of the Seven Multiplying Squares" shows the binomial coefficients up to the eighth power – the numbers that are fundamental to the mathematics of probability, and that appeared five hundred years later in the west as Pascal's triangle.

1346 Giovanni Villani's *Nuova Cronica* gives statistical information on the population and trade of Florence.

431 BC Attackers besieging Plataea in the Peloponnesian war calculate the height of the wall by counting the number of bricks. The count was repeated several times by different soldiers. The most frequent value (the mode) was taken to be the most likely. Multiplying it by the height of one brick allowed them to calculate the length of the ladders needed to scale the walls.

AD 2 Chinese census under the Han dynasty finds 57.67 million people in 12.36 million households – the first census from which data survives, and still considered by scholars to have been accurate.

840 Islamic mathematician Al-Kindi uses frequency analysis – the most common symbols in a coded message will stand for the most common letters – to break secret codes. Al-Kindi also introduces Arabic numerals to Europe.

1069 Domesday Book: survey for William the Conqueror of farms, villages and livestock in his new kingdom – the start of official statistics in England.

1150 Trial of the Pyx, an annual test of the purity of coins from the Royal Mint, begins. Coins are drawn at random, in fixed proportions to the number minted. It continues to this day.

Mathematical foundations

1560 Gerolamo Cardano calculates probabilities of different dice throws for gamblers.

1654 Pascal and Fermat correspond about dividing stakes in gambling games and together create the mathematical theory of probability.

1663 John Graunt uses parish records to estimate the population of London.

1713 Jacob Bernoulli's *Ars conjectandi* derives the law of large numbers – the more often you repeat an experiment, the more accurately you can predict the result.

1749 Gottfried Achenwall coins the word "statistics" (in German, *Statistik*); he means the information you need to run a nation state.

1761 The Rev. Thomas Bayes proves Bayes' theorem – the cornerstone of conditional probability and the testing of beliefs and hypotheses.

1789 Gilbert White and other clergymen-naturalists keep records of temperatures, dates of first snowdrops and cuckoos, etc; the data is later useful for study of climate change.

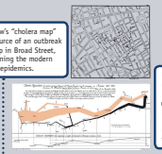
1791 First use of the word "statistics" in English, by Sir John Sinclair in his *Statistical Account of Scotland*.

1808 Gauss, with contributions from Laplace, derives the normal distribution – the bell-shaped curve fundamental to the study of variation and error.

1835 Belgian Adolphe Quetelet's *Treatise on Man* introduces social science statistics and the concept of the "average man" – his height, body mass index, and earnings.

1840 William Farr sets up the official system for recording causes of death in England and Wales. This allows epidemics to be tracked and diseases compared – the start of medical statistics.

1854 John Snow's "cholera map" pins down the source of an outbreak as a water pump in Broad Street, London, beginning the modern study of epidemics.



1868 Minard's graphic diagram of Napoleon's March on Moscow shows on one diagram the distance covered, the number of men still alive at each kilometre of the march, and the temperatures they encountered on the way.

1886 Philanthropist Charles Booth begins his survey of the London poor, to produce his "poverty map of London". Areas were coloured black, for the poorest, through to yellow for the upper-middle class and wealthy.

1898 Von Bortkiewicz's data on deaths of soldiers in the Prussian army from horse kicks shows that apparently rare events follow a predictable pattern, the Poisson distribution.

1570 Astronomer Tycho Brahe uses the arithmetic mean to reduce errors in his estimates of the locations of stars and planets.

1644 Michael van Langren draws the first known graph of statistical data that shows the size of possible errors. It is of different estimates of the distance between Toledo and Rome.

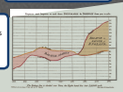
1657 Huygens's *On Reasoning in Games of Chance* is the first book on probability theory. He also invented the pendulum clock.

1693 Edmund Halley prepares the first mortality tables statistically relating death rates to age – the foundation of life insurance. He also drew a stylised map of the path of a solar eclipse over England – one of the first data visualisation maps.

1728 Voltaire and his mathematician friend de la Condamine spot that a Paris bond lottery is offering more in prize money than the total cost of the tickets; they corner the market and win themselves a fortune.

1757 Casanova becomes a trustee of, and may have had a hand in devising, the French national lottery.

1786 William Playfair introduces graphs and bar charts to show economic data.



1790 First US census, taken by men on horseback directed by Thomas Jefferson, counts 3.9 million Americans.

1805 Adrien-Marie Legendre introduces the method of least squares for fitting a curve to a given set of observations.

1833 The British Association for the Advancement of Science sets up a statistics section. Thomas Malthus, who analysed population growth, and Charles Babbage are members. It later becomes the Royal Statistical Society.

1839: The American Statistical Association is formed. Alexander Graham Bell, Andrew Carnegie and President Martin Van Buren will become members.

1849 Charles Babbage designs his "difference engine", embodying the ideas of data handling and the modern computer. Ada Lovelace, Lord Byron's niece, writes the world's first computer program for it.

1859 Florence Nightingale uses statistics of Crimean War casualties to influence public opinion and the War Office. She shows casualties month by month on a circular chart she devises, the "Nightingale rose", forerunner of the pie chart. She is the first woman member of the Royal Statistical Society and the first overseas member of the American Statistical Association.

1877 Francis Galton, Darwin's cousin, describes regression to the mean. In 1888 he introduces the concept of correlation. At a "Guess the weight of an Ox" contest in Devon he describes the "Wisdom of Crowds" – that the average of many uninformed guesses is close to the correct value.

1894 Karl Pearson introduces the term "standard deviation". If errors are normally distributed, 68% of samples will lie within one standard deviation of the mean. Later he develops chi-squared tests for whether two variables are independent of each other.

1979 Bradley Efron introduces bootstrapping, a simple way to estimate the distribution of almost any sample of data.

1997 The term "Big Data" first appears in print.

2002 Paul DePodesta uses statistics – "sabermetrics" – to transform the fortunes of the Oakland Athletics baseball team; the film *Moneyball* tells the story.

2012 Nate Silver, statistician, successfully predicts the result in all 50 states in the US Presidential election. He becomes a media star.

1900 Louis Bachelier shows that fluctuations in stock market prices behave in the same way as the random Brownian motion of molecules – the start of financial mathematics.

1916 During the First World War car designer Frederick Lanchester develops statistical laws to predict the outcomes of aerial battles: if you double their size land armies are only twice as strong, but air forces are four times as powerful.

1924 Walter Shewhart invents the control chart to aid industrial production and management

1935 R. A. Fisher revolutionises modern statistics. His *Design of Experiments* gives ways of deciding which results of scientific experiments are significant and which are not.

1940-45 Alan Turing at Bletchley Park cracks the German wartime Enigma code, using advanced Bayesian statistics and Colossus, the first programmable electronic computer.

1946 Cox's theorem derives the axioms of probability from simple logical assumptions.

1948-53 The Kinsey Report gathers objective data on human sexual behaviour. A large-scale survey of 5000 men and, later, 5000 women, it causes outrage.

1950s Genichi Taguchi's statistical methods to improve the quality of automobile and electronics components revolutionise Japanese industry, which far overtakes western European rivals.

Modern era

1908 William Sealy Gosset, chief brewer for Guinness in Dublin, describes the t-test. It uses a small number of samples to ensure that every brew tastes equally good.

1911 Herman Hollerith, inventor of punch-card devices used to analyse data in US censuses, merges his company to form what will become IBM, pioneers of machines to handle business data and of early computers.

1935 George Zipf finds that many phenomena – river lengths, city populations – obey a power law so that the largest is twice the size of the second largest, three times the size of the third, and so on.

1937 Jerzy Neyman introduces confidence intervals in statistical testing. His work leads to modern scientific sampling.

1944 The German tank problem: the Allies desperately need to know how many Panther tanks they will face in France on D-Day. Statistical analysis of the serial numbers on gearboxes from captured tanks indicates how many of each are being produced. Statisticians predict 270 a month; reports from intelligence sources predict many fewer. The total turned out to be 276. Statistics had outperformed spies.

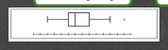
1948 Claude Shannon introduces information theory and the "bit" – fundamental to the digital age.

1950 Richard Doll and Bradford Hill establish the link between cigarette smoking and lung cancer. Despite fierce opposition the result is conclusively proved, to huge public health benefit.

1958 The Kaplan–Meier estimator gives doctors a simple statistical way of judging which treatments work best. It has saved millions of lives.

1972 David Cox's proportional hazard model and the concept of partial likelihood.

1977 John Tukey introduces the box-plot or box-and-whisker diagram, which shows the quartiles, medians and spread of data in a single image.



1982 Edward Tufte self-publishes *The Visual Display of Quantitative Information*, setting new standards for graphic visualisation of data.

1988 Margaret Thatcher becomes the first world leader to call for action on climate change.

1993 The statistical programming language "R" is released, now a standard statistical tool.

2002 The amount of information stored digitally surpasses non-digital.

2004 Launch of *Significance* magazine.

2008 Hal Varian, chief economist at Google, says that statistics will be "the sexy profession of the next ten years".

2012 The Large Hadron Collider confirms existence of a Higgs boson-like particle with probability of five standard deviations – around one chance in 3.5 million that all they are seeing is coincidence.

Photo: Sami Keinänen

1908 William Sealy Gosset, chief brewer for Guinness in Dublin, describes the *t*-test. It uses a small number of samples to ensure that every brew tastes equally good.

Object of Study

- Not the small sample
- The distribution from which the small sample was drawn

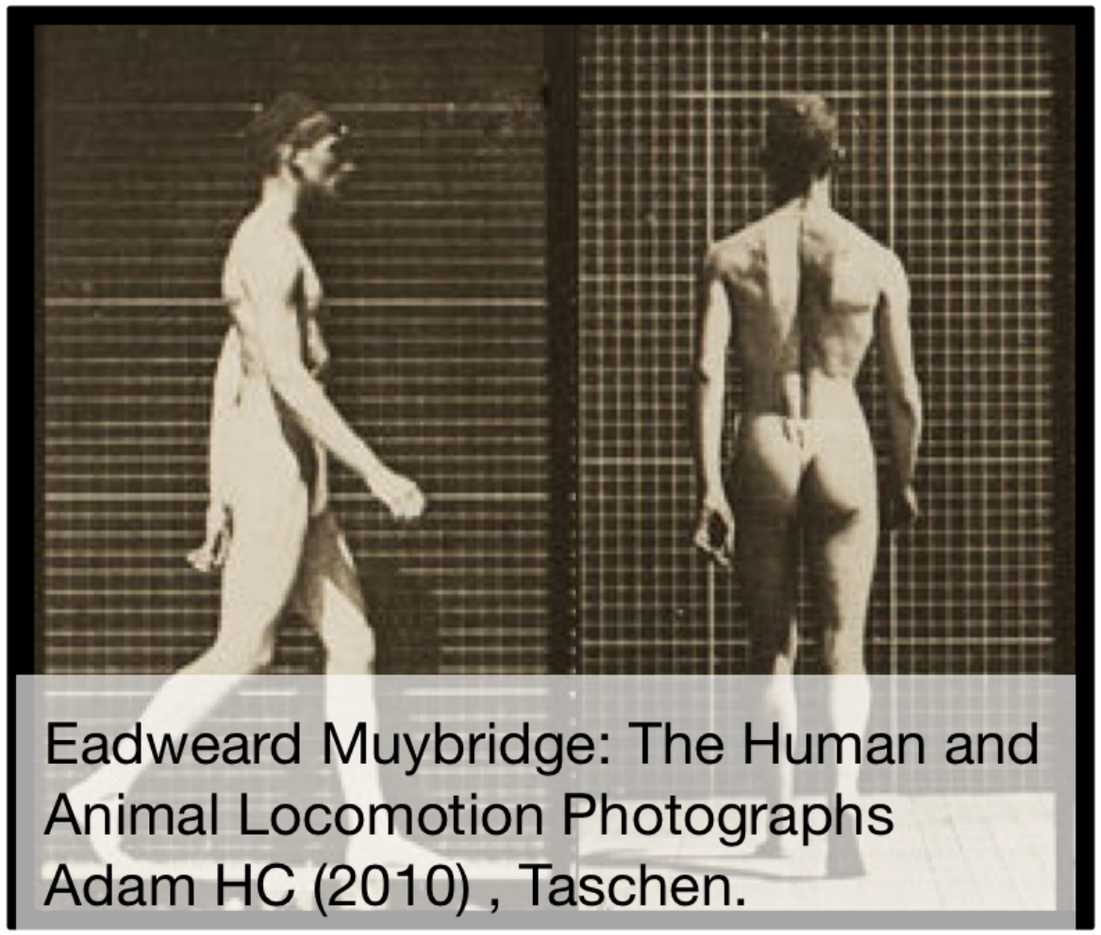
Key developments

- Pearson (1901): PCA
- Gosset (1908): t tests
- Hotelling (1933): Multivariate analysis
- Fisher (1935): ANOVA, design
- Pearson/Neyman (1937): Power + hypothesis testing

A serious problem for Biomechanics:

Multivariate Data
Univariate Analysis

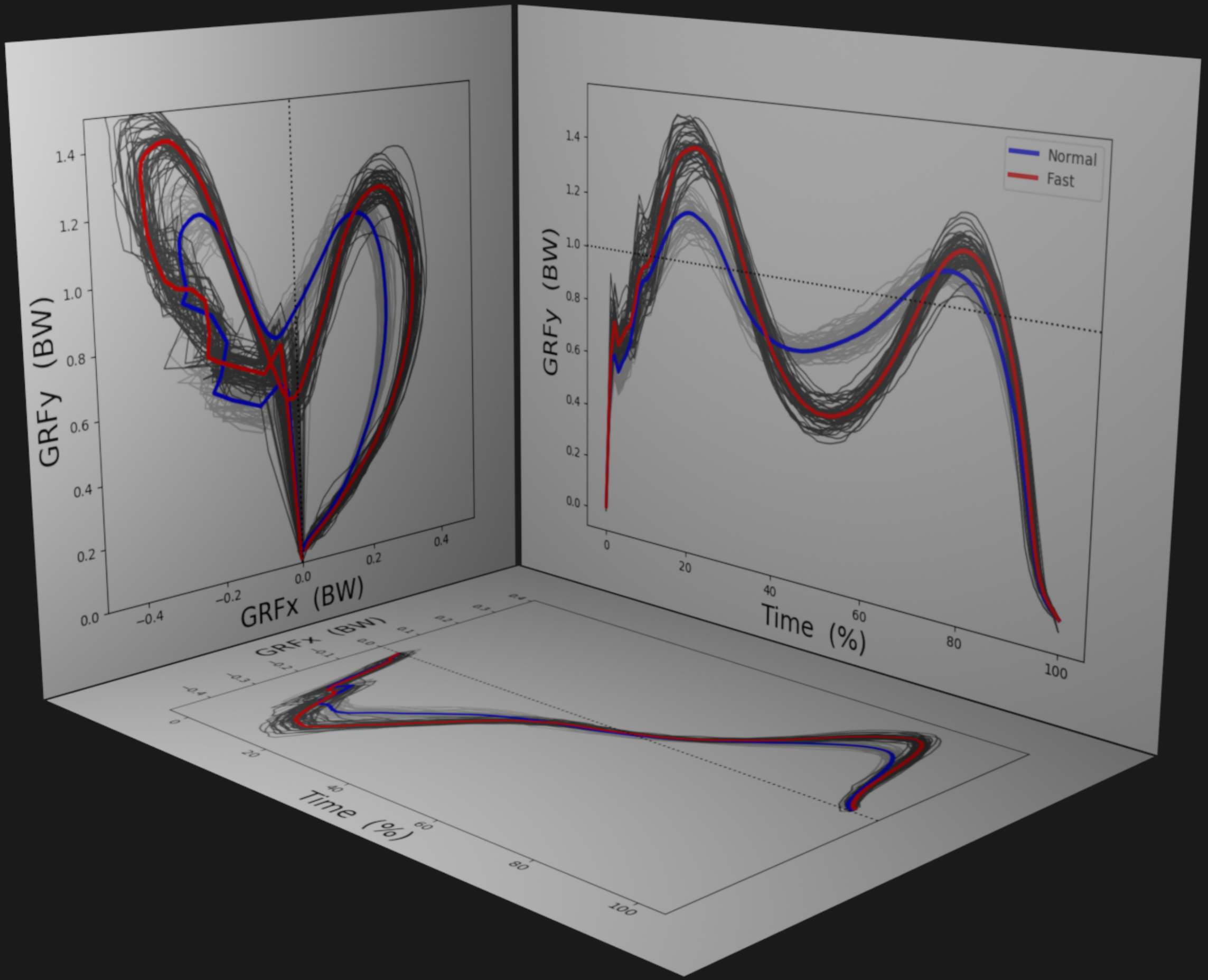
(a) 2D tradition

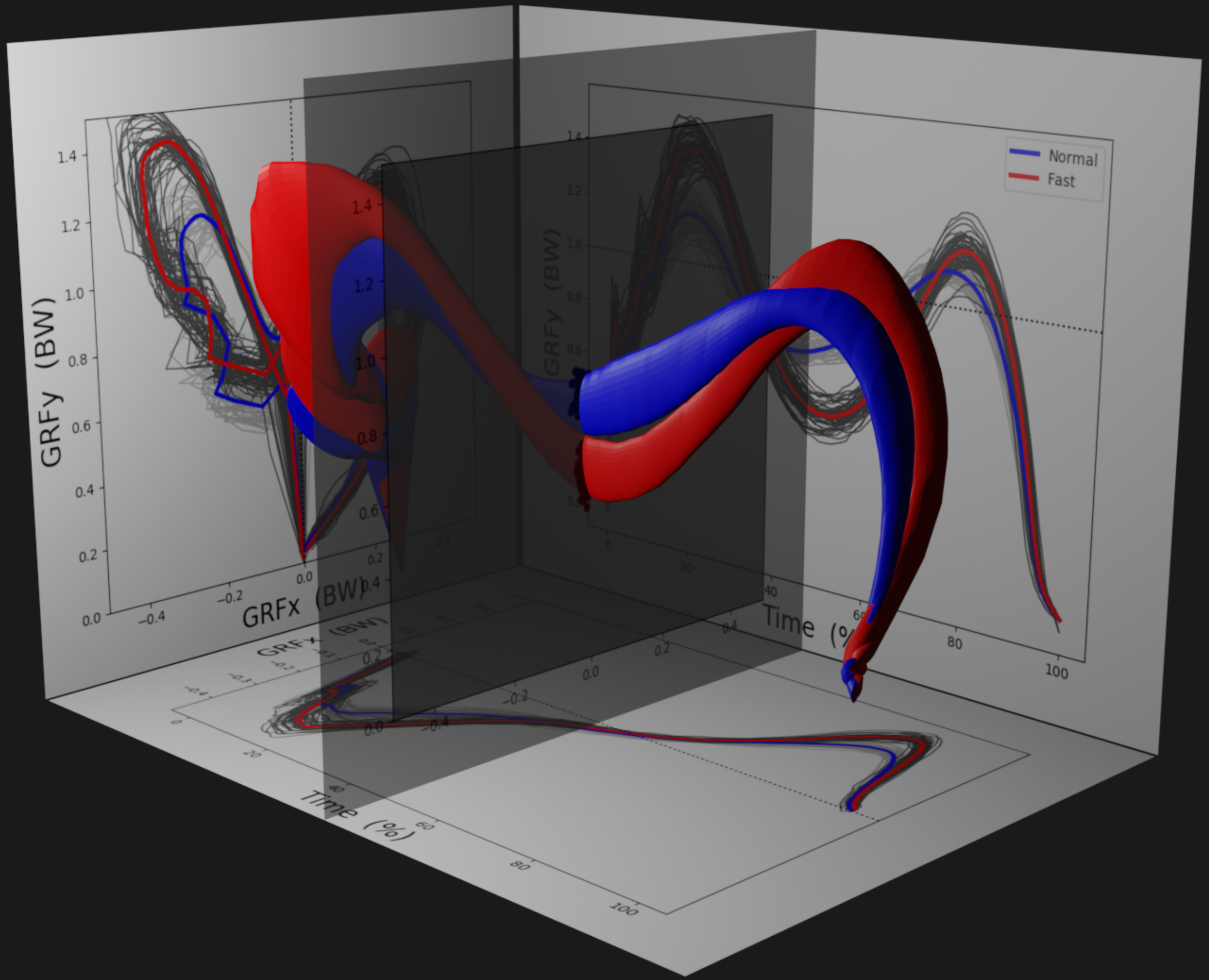


Eadweard Muybridge: The Human and Animal Locomotion Photographs
Adam HC (2010) , Taschen.

(b) 3D real-time dynamics

Separate-plane analysis → Separate-plane analysis


GRFy (BW)
GRFx (BW)
Time (%)
Normal
Fast


Normal
Fast
GRFy (BW)
GRFx (BW)
Time (%)

Significance

Colloquial Uses

(Everyday, Non-scientific)

TERM	COLLOQUIAL USE	SCIENTIFIC USE
Statistic	Data Number Summary statistic (mean, percentage, etc.)	Probability, uncertainty
Theory / hypothesis	Guess	Prediction generator / prediction
Significant	Meaningful, Important	

Overview

- History
- The p value
- Classical techniques
- Emerging techniques
- Controversies
- The future

Questions

Questions

“The p value was never meant to be used the way it is used today.”

–Steven Goodman, Stanford University

from Nuzzo (2014) *Nature*

Body mass (kg)

(Group A vs. Group B)

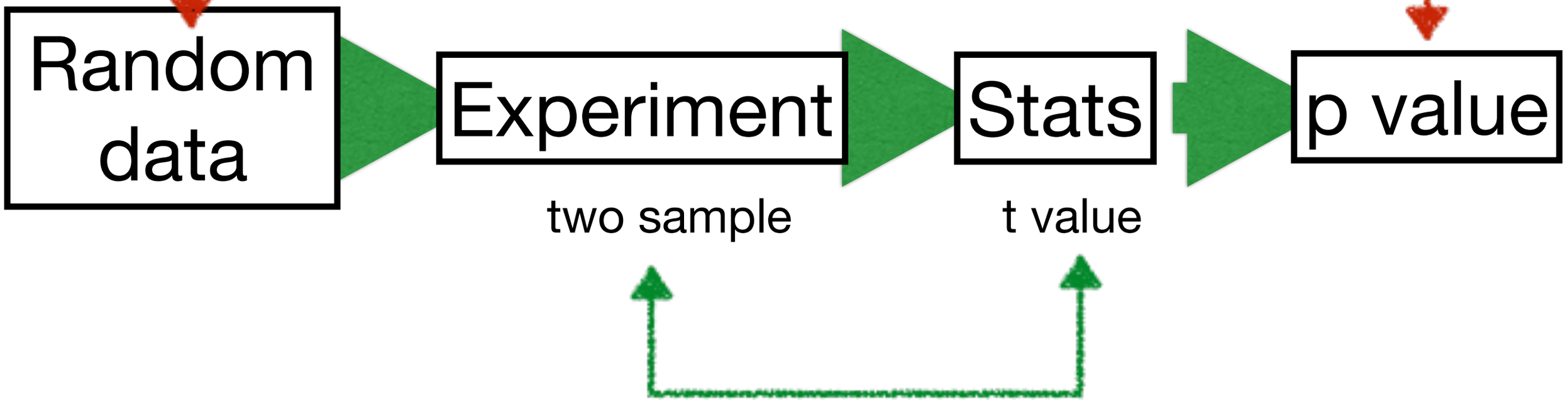
$p = 0.03$

What does the p value mean?

What does "*p*=0.03" mean?

The probability that **random sampling** will produce a more extreme result than the observed result (**when the null hypothesis is true**) is 3%.

Infinite set of experiments



One experiment

Problem 1

The p value is a **conditional probability**

$$p(\text{ data } | H_0)$$

but often interpreted as:

$$p(H_0 | \text{ data })$$

p(sunny | 30 deg) 0.98 ?

p(30 deg | sunny) 0.04 ?

$$P(A|B) \neq P(B|A)$$

Descriptive statistics

Problem 2

p values are used as descriptive statistics

- Descriptive statistics:
  - Sample mean, SD, min, max, IQR, etc.
  - Pertain to a specific dataset

Probabilities DO NOT pertain to single datasets !!!

	Females	Males	p-value
Age (years)	34.8 (12.7)	34.8 (12.1)	0.997
Height (m)	1.62 (0.08)	1.78 (0.06)	< 0.001
Mass (kg)	62.5 (14.4)	83.0 (13.5)	< 0.001

Gait & Posture (2019)

	Females	Males	p-value
Mass (kg)	62.5 (14.4)	83.0 (13.5)	< 0.001

$H_0: \overline{mass}_{females} = \overline{mass}_{males}$

H_0: is unjustified

- Non-ideal use of p values
- Promotes incorrect interpretations of p values

p values describe H_0

The scientific quality of the p value depends on the scientific quality of H_0

“The p value was never meant to be used the way it is used today.”

–Steven Goodman, Stanford University

from Nuzzo (2014) *Nature*

Overview

- History
- The p value
- Classical techniques

Questions

- Emerging techniques
- Controversies
- The future

Questions

“If your experiment needs statistics, you ought to have done a better experiment.”

–Ernest Rutherford

Classical techniques

Classical hypothesis testing

- t tests
- Regression
- ANOVA
- ...
- MANCOVA

Power analysis

Fisher

Pearson-Neyman

t test True or False

Example

Part 2: Classical hypothesis testing: critical thresholds and p values

Consider the following dataset set and the one-sample hypothesis testing results:

In [10]:
```
rng(58)       %seed=58 chosen to obtain a p value close to 0.05
J     = 8;    %sample size
mu    = 80;   %true population mean
sigma = 10;   %true population SD

y     = mu + sigma * randn(J,1);
[h,p,ci,stats] = ttest(y, mu);
fprintf('One-sample t test results:  t = %.3f, p = %.3f', stats.tstat, p)
```

```
One-sample t test results:  t = -1.999, p = 0.086
```

Example

H_0: μ = 80 kg

Sample mean (SD) : 73.9 (8.5) kg

One sample t value : t = -1.999

Two-tailed p value : p = 0.086
One-tailed p value : p = 0.043

Simulation

Part 2: Classical hypothesis testing: critical thresholds and p values

Consider the following dataset set and the one-sample hypothesis testing results:

In [10]:
```
rng(58)       %seed=58 chosen to obtain a p value close to 0.05
J     = 8;    %sample size
mu    = 80;   %true population mean
sigma = 10;   %true population SD

y     = mu + sigma * randn(J,1);
[h,p,ci,stats] = ttest(y, mu);
fprintf('One-sample t test results:  t = %.3f, p = %.3f', stats.tstat, p)
```

```
One-sample t test results:  t = -1.999, p = 0.086
```

In simulation we know the population reality.

In experiments we never know the population reality.

Simulations are cheap.
Experiments are expensive.

Ideal vs. non-ideal

Hypothesis Testing

3

False positives & negatives

Ideal Hypothesis Testing

"Side-effect" approach

- Medicine
- H_0: No side-effect
  - e.g. a new drug
- Good result: failure to reject H_0

Predictive approach

- Physics
- H_0: A specific prediction
  - e.g. H_0: F = ma
- Good result: failure to reject H_0

Non-Ideal Hypothesis Testing

Exploratory approach

- Biomechanics and many other fields
- H_0: Null prediction
  - e.g. Group A = Group B
- Good result: reject H_0

Problem 1

Hypotheses stated as inequalities

e.g. “We hypothesize that Group A exhibits greater knee valgus moment than Group B.”

Statements of inequality are irrefutable

Examples

Exploratory

H_0: Quadriceps weakness produces no change in walking speed.

Predictive

H_0: Based on musculoskeletal modeling optimization results involving simulated quadriceps weakness, we predict a walking speed reduction of 0.1 m/s in older adults

Problem 2

Ambiguous dependent variables

e.g. “The purpose of this study was to investigate knee kinematics in Group A vs. Group B.”

(null) H_0: ? = 0

(alternative) H_1: ? > 0

Irrefutable

Politicians:

“The economy is stronger now than last year”

Irrefutable

Devising a non-rejectable hypothesis is scientifically move valuable than rejecting a poor hypothesis

From the literature

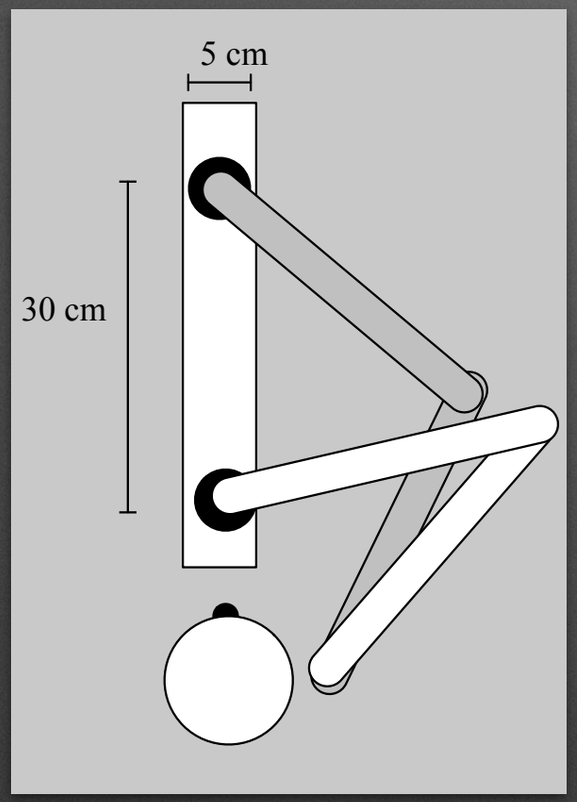
- “The purpose of our investigation was to evaluate the influence of hamstrings musculotendinous stiffness on lower extremity kinematics and kinetics during landing.”
- “The objective of this study was to examine the kinetic and kinematic differences between the first and second landings in a DVJ.”
- “We hypothesise that the series elastic tissues of the MG muscle have a compliance which allows the muscle to operate with optimal efficiency during normal walking and running.”

2

Power Analysis

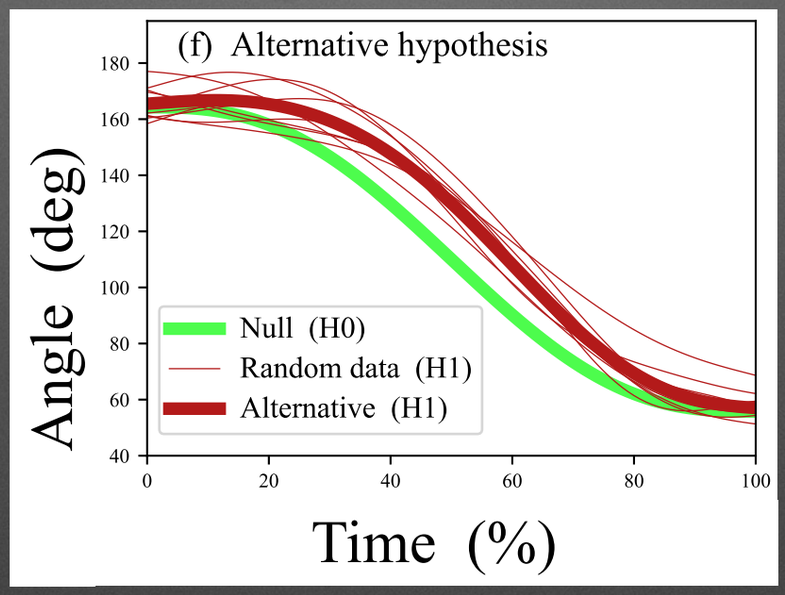
- Define effect
- Set α (usually 0.05)
- Set target power (usually 0.8)
- Calculate sample size

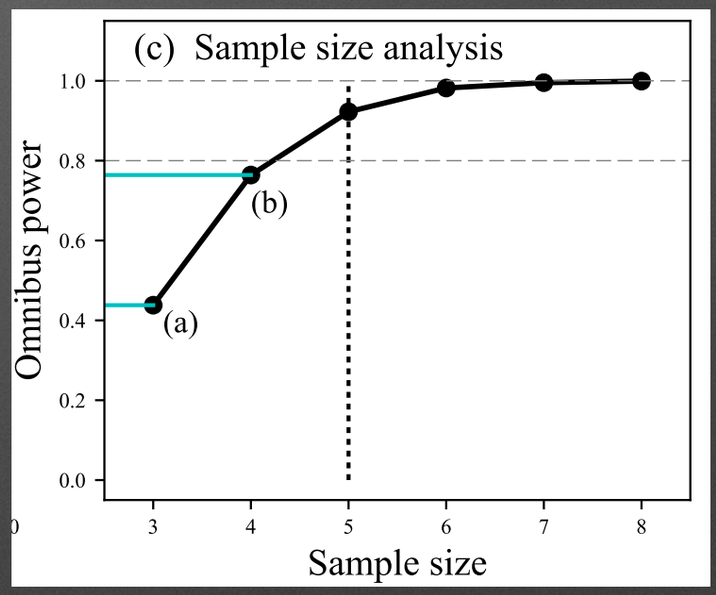
Example power analysis



H_0: Min jerk trajectory

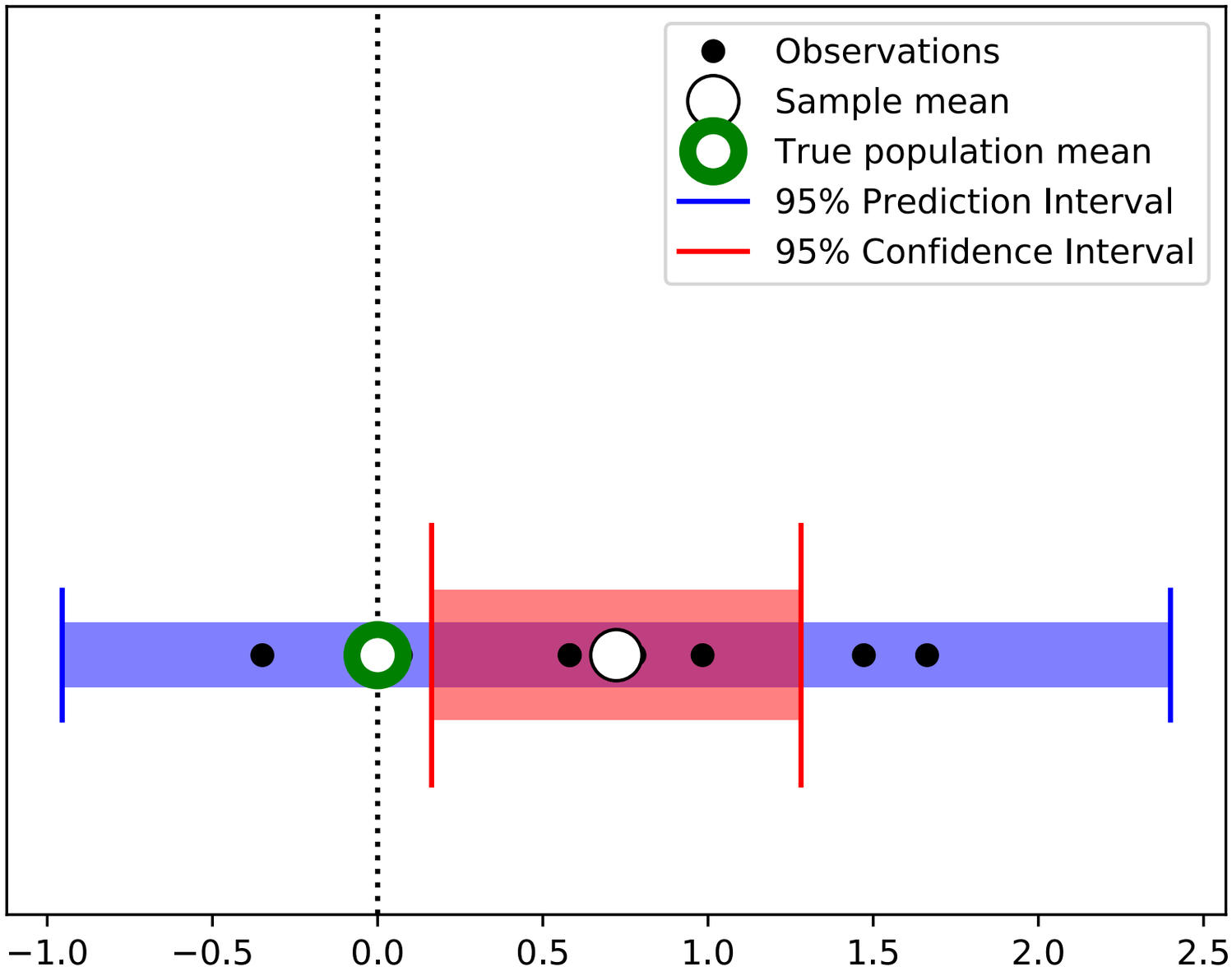
H_1: 10% greater muscular work

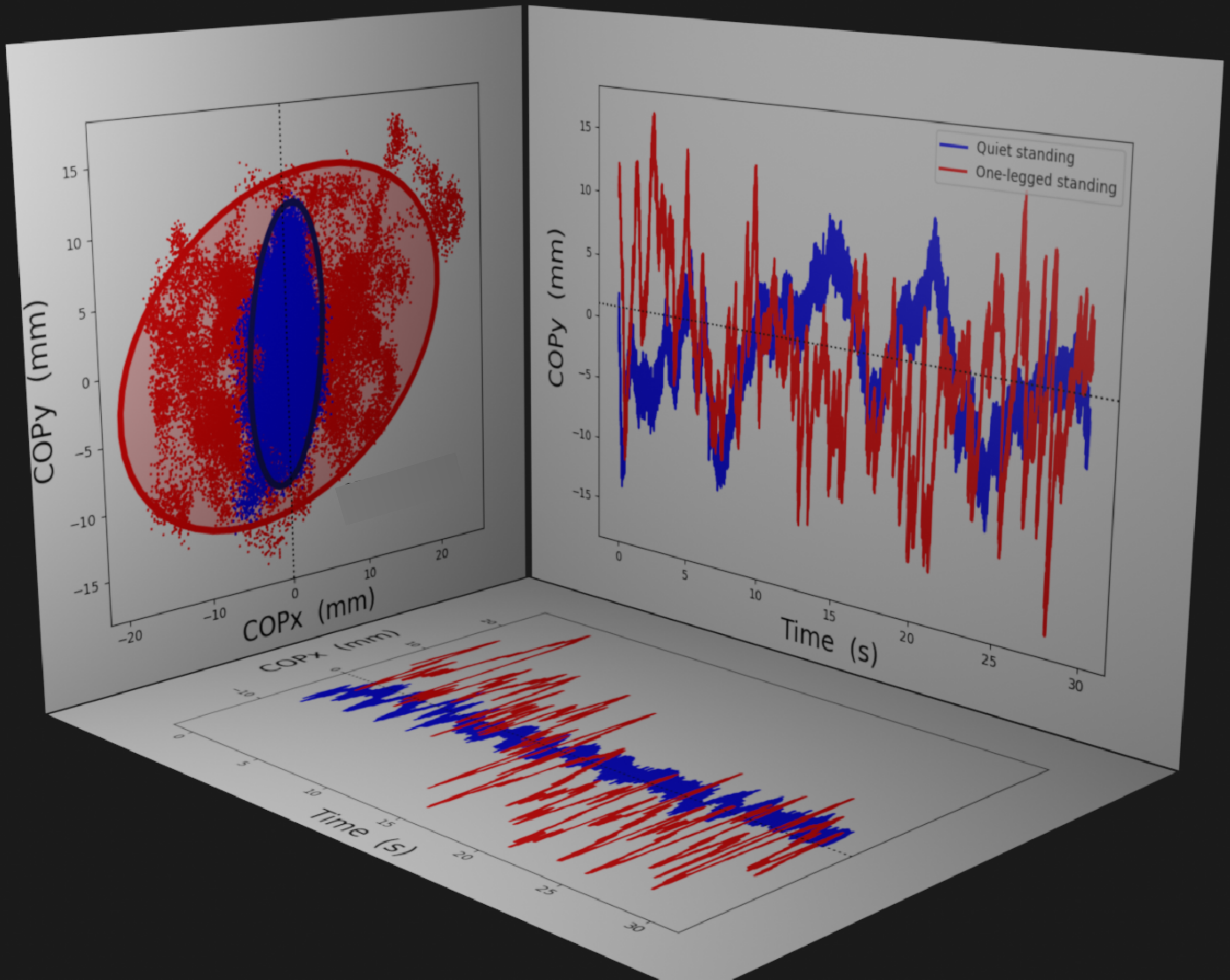




Giving biomechanical meaning to hypotheses is better than trying to infer meaning from biomechanically meaningless hypotheses

95% regions


Observations
Sample mean
True population mean
95% Prediction Interval
95% Confidence Interval
−1.0
−0.5
0.0
0.5
1.0
1.5
2.0
2.5


COPy (mm)
COPx (mm)
Time (s)
Quiet standing
One-legged standing

Predictive

Exploratory

We should search for non-rejectable hypotheses

We should not search for significance

Overview

- History
- The p value
- Classical techniques
- Emerging techniques
- Controversies
- The future

Questions

Questions

Overview

- History
- The p value
- Classical techniques

Questions

- Emerging techniques
- Controversies
- The future

Questions

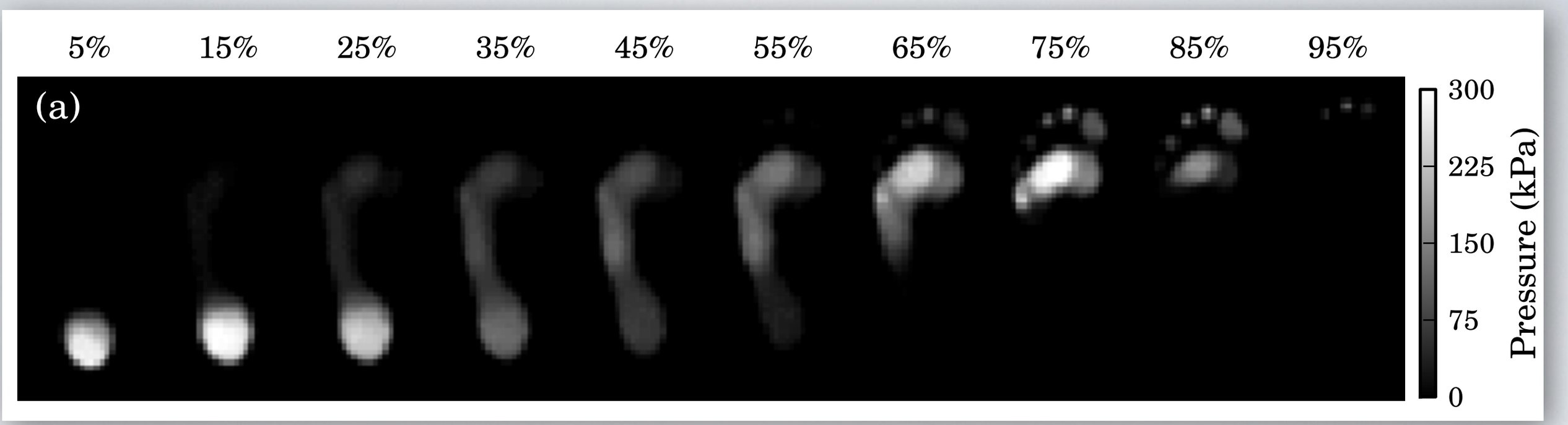
“We are drowning in information but starved for knowledge.”

–John Naisbitt, Megatrends

Motivation

Data volume
Data complexity


zebris


5%
15%
25%
35%
45%
55%
65%
75%
85%
95%
(a)
300
225
150
75
0
Pressure (kPa)
(b)
Pressure (kPa)
Time (s)
(c)
P100
P90
P80
P70
P60
P50
Tmax
PIm
CDurn
Tfirst
Time (s)

1 variable (pressure)

0.6 seconds

500 sensors contacted

200 Hz

60000 data points every step

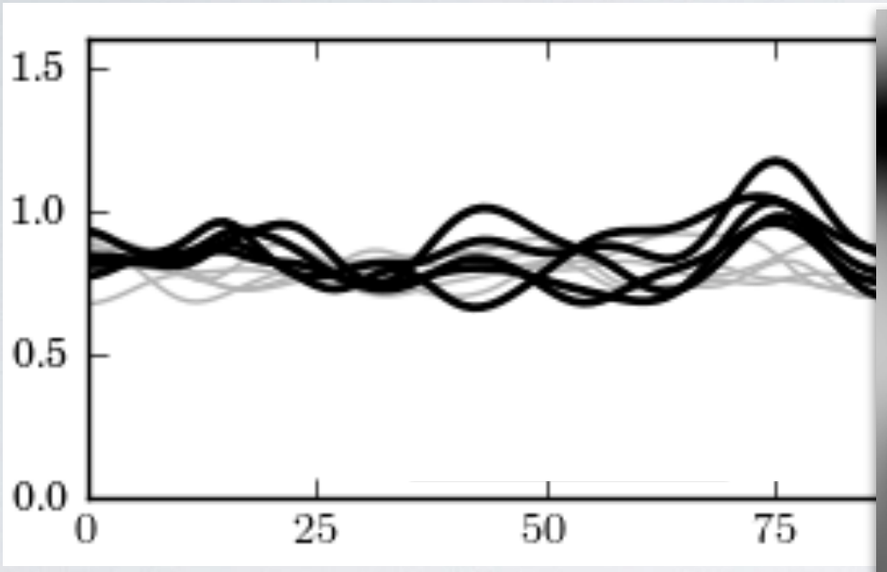
Motivation

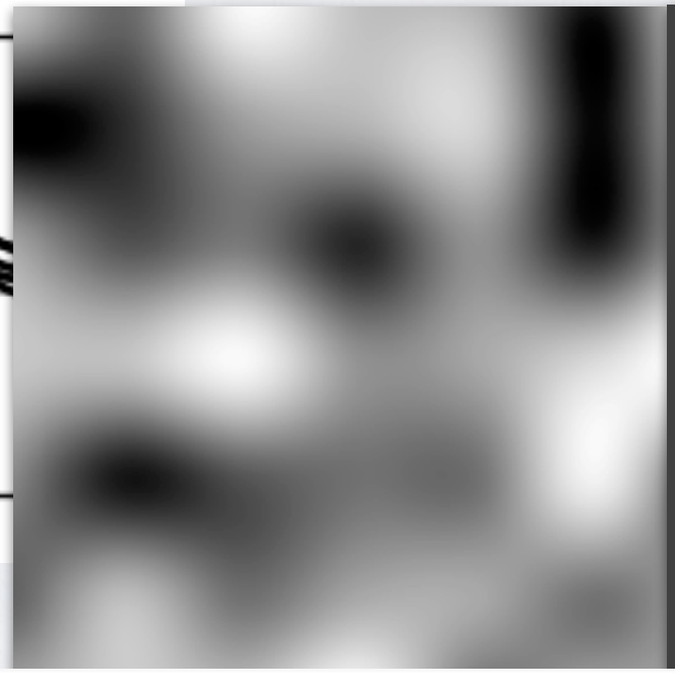
Continuous nature

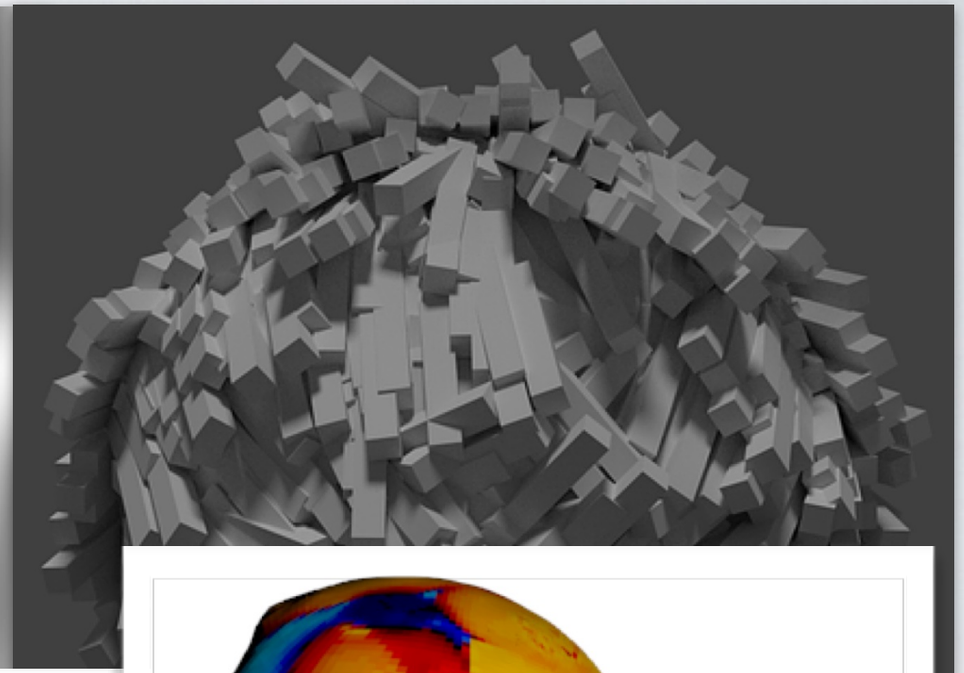
1D

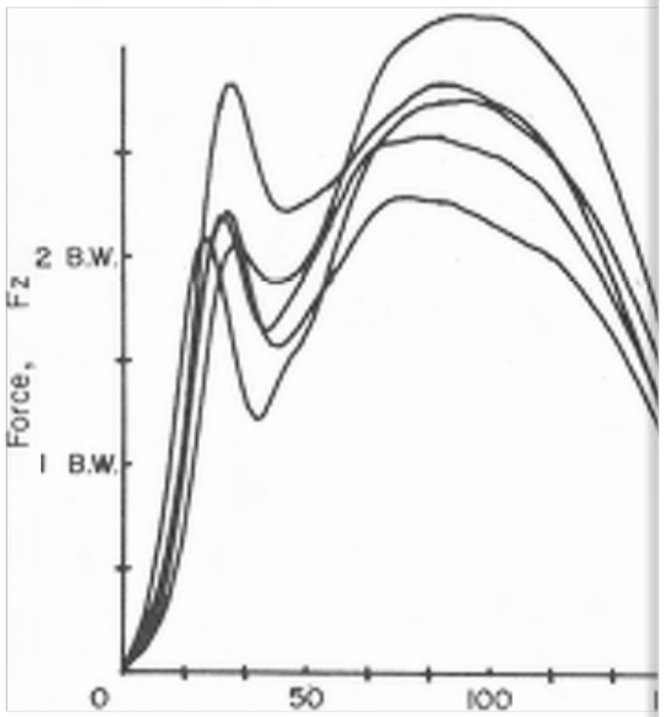
2D

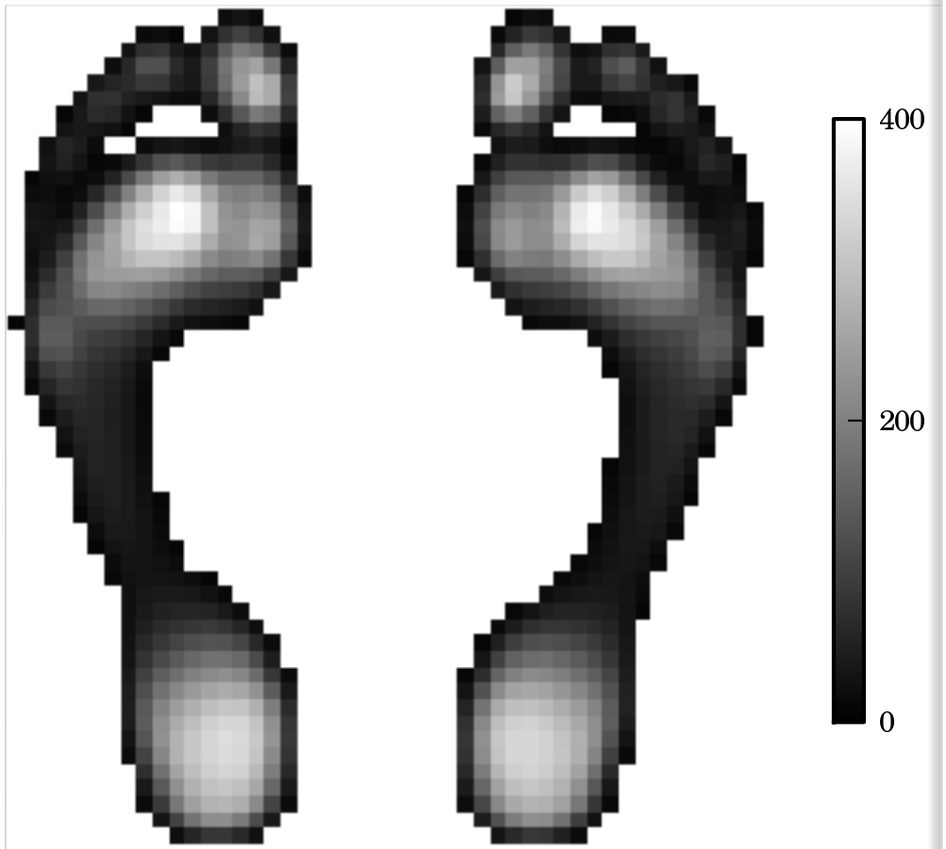
3D

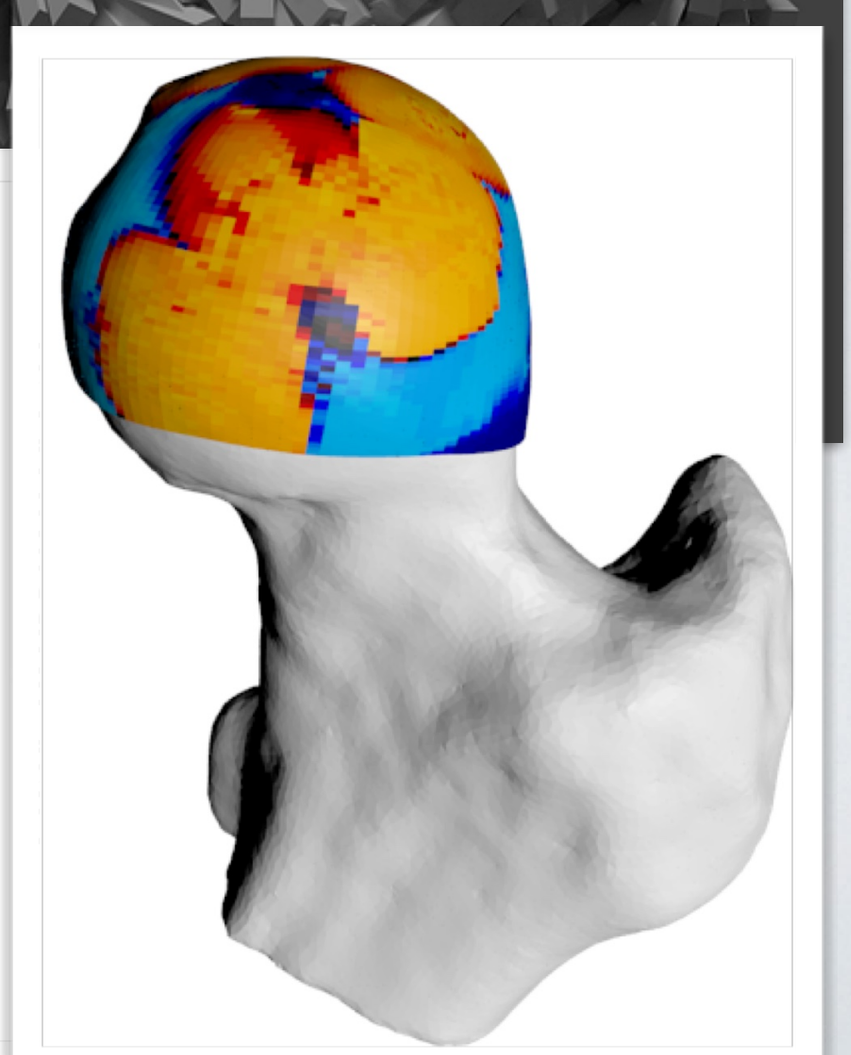




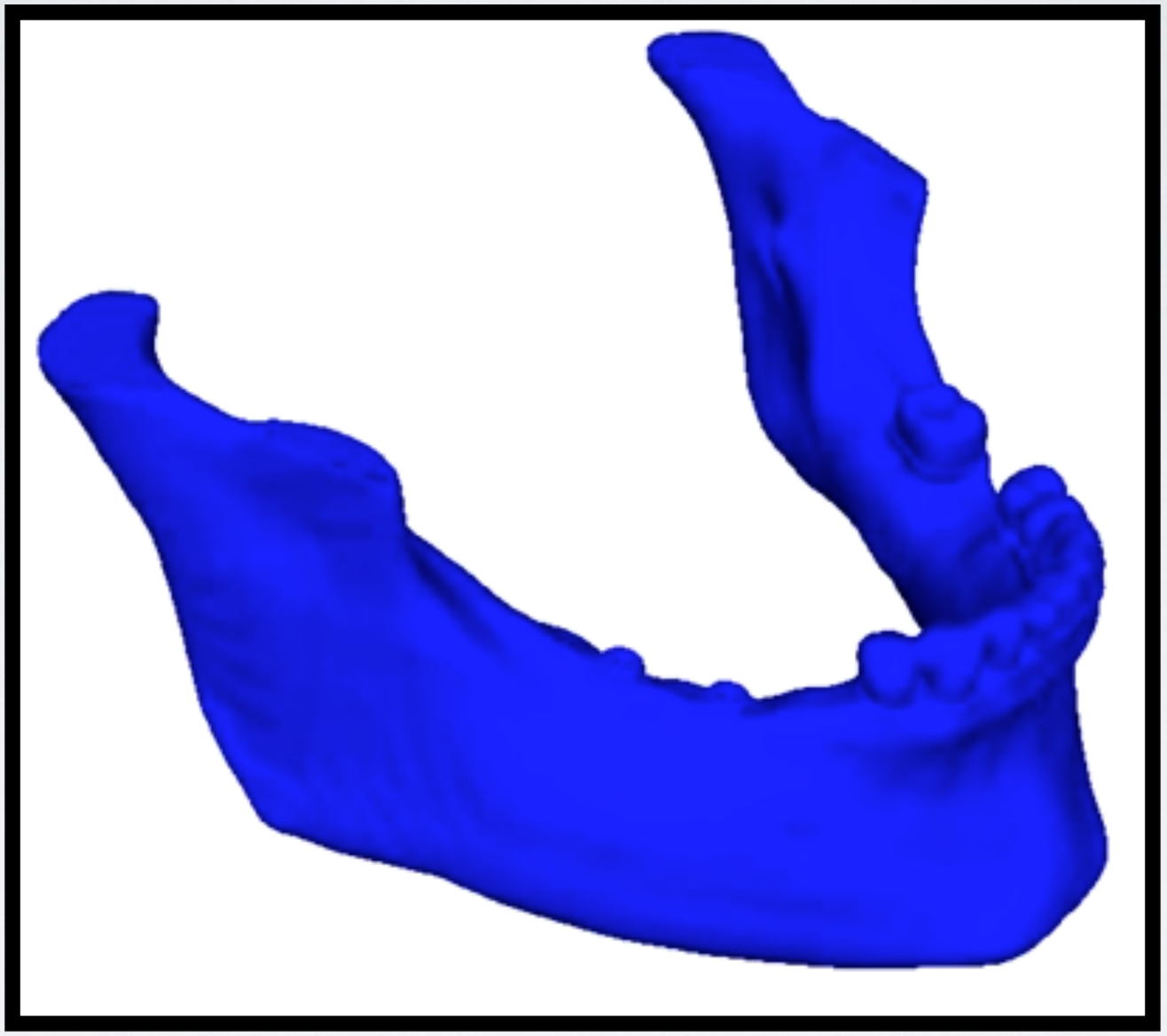


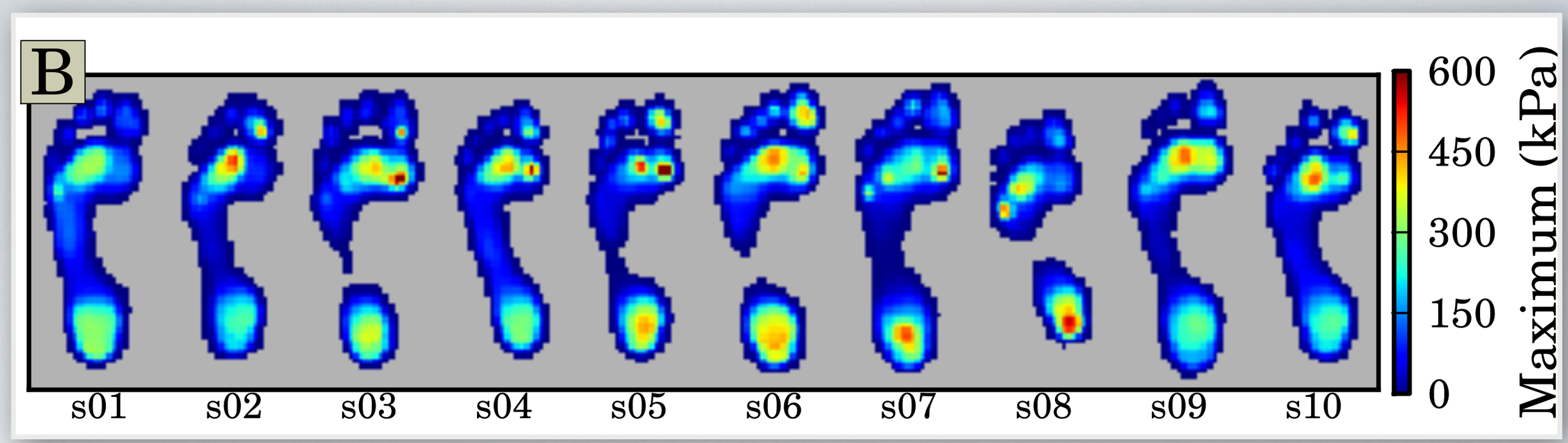


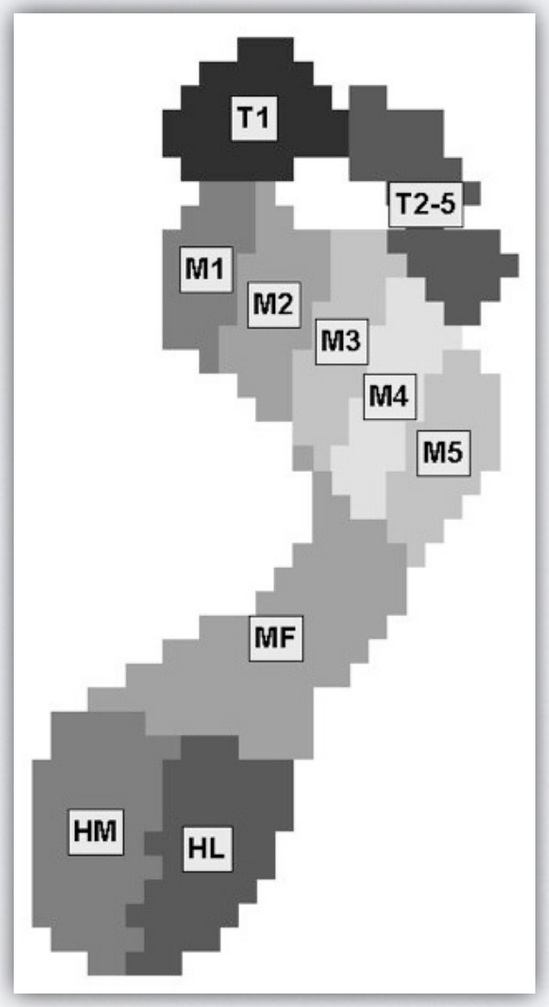




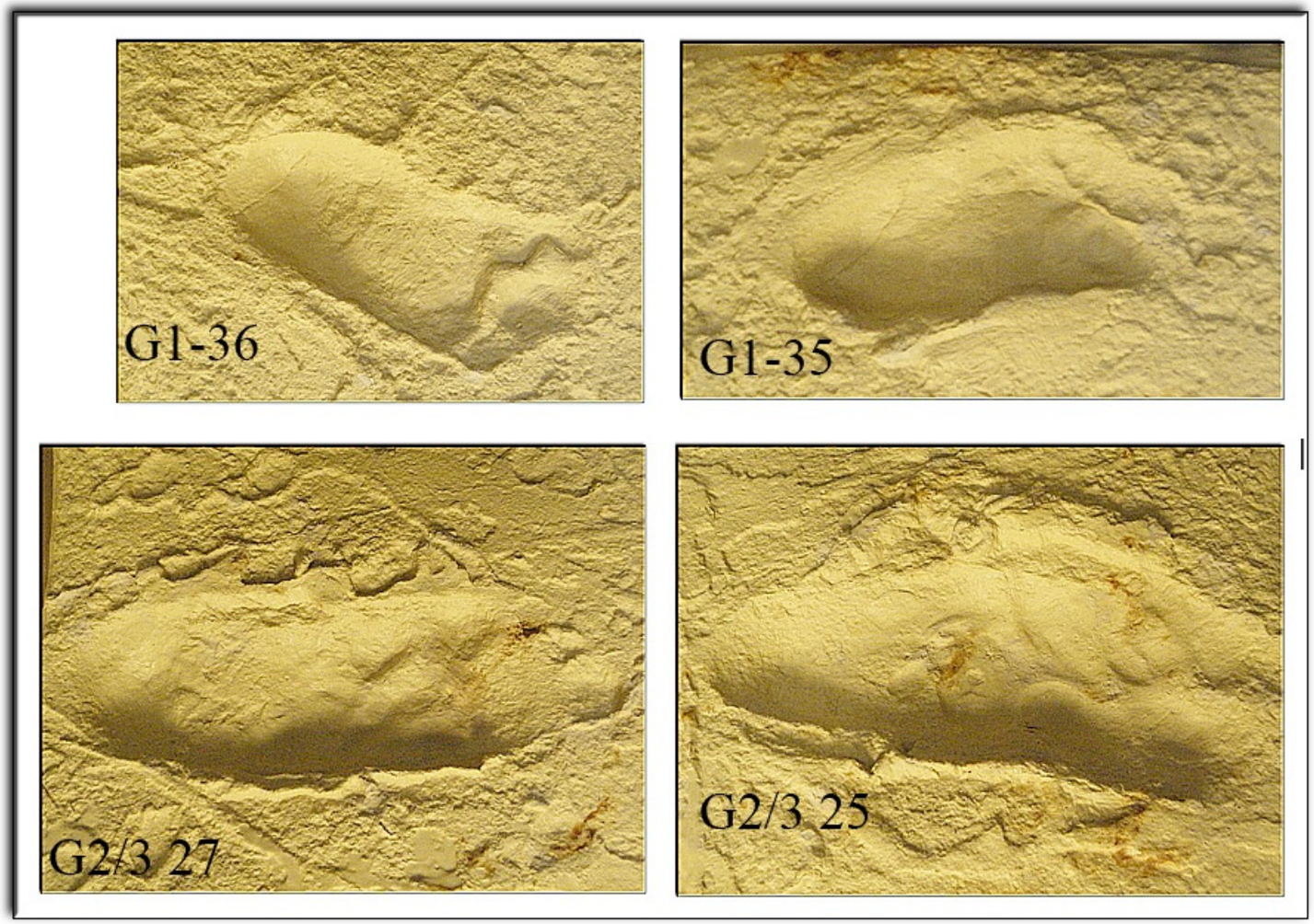
4D




B
s01
s02
s03
s04
s05
s06
s07
s08
s09
s10
600
450
300
150
0
Maximum (kPa)


T1
T2-5
M1
M2
M3
M4
M5
MF
HM
HL

Laetoli footprints



Leakey & Hay (1979)

- FDA
- SPM
- Dimensionality reduction (PCA, ICA)
- Machine Learning

(This list is not comprehensive)

SPM

3

SPM

S

STATISTICAL

P

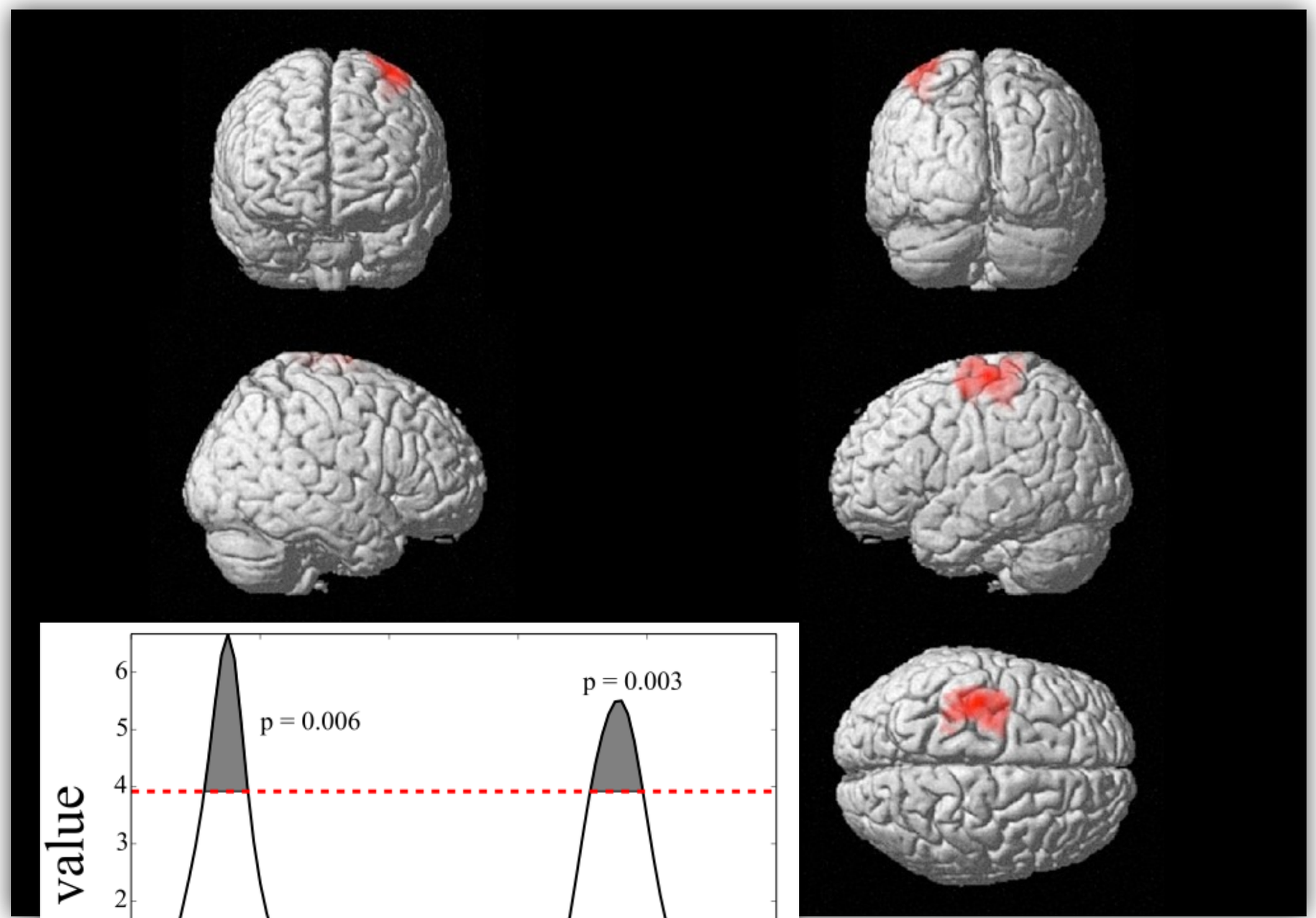
PARAMETRIC

- Based on **mean** & **SD** & **sample size**
- Also non-parametric (SnPM)
- Parameterized model of cerebral blood flow

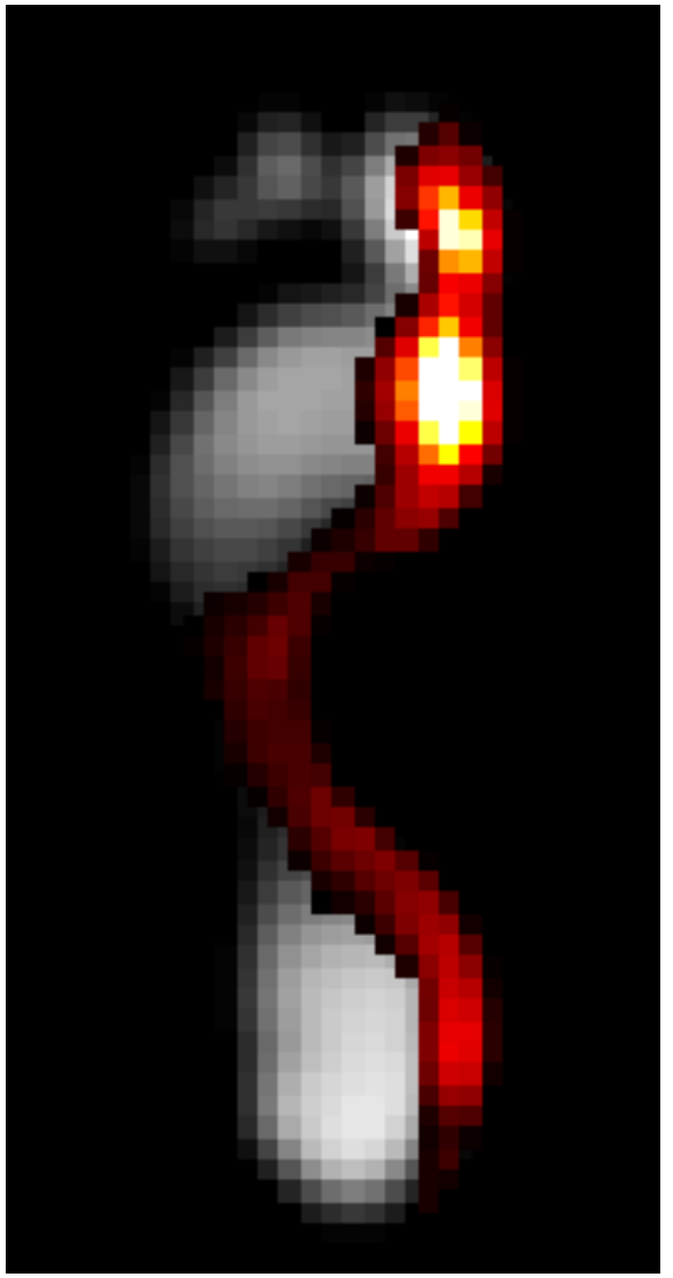
M

MAPPING

- Results form an *n*-Dimensional “map” in the same time / space as the original data
- Test statistics [t and F] are continuous in time / space



p = 0.006
p = 0.003
t value
Time (%)
0
20
40
60
80
100
−1
0
1
2
3
4
5
6



A brief history of SPM

1976 Adler & Hasofer, Annals of Prob.

1990 Friston et al. J Cerebral Blood Flow

1995 Friston et al. Human Brain Mapping

9489 citations

2004 Worsley et al. NeuroImage

2008 Pataky et al. J Biomech

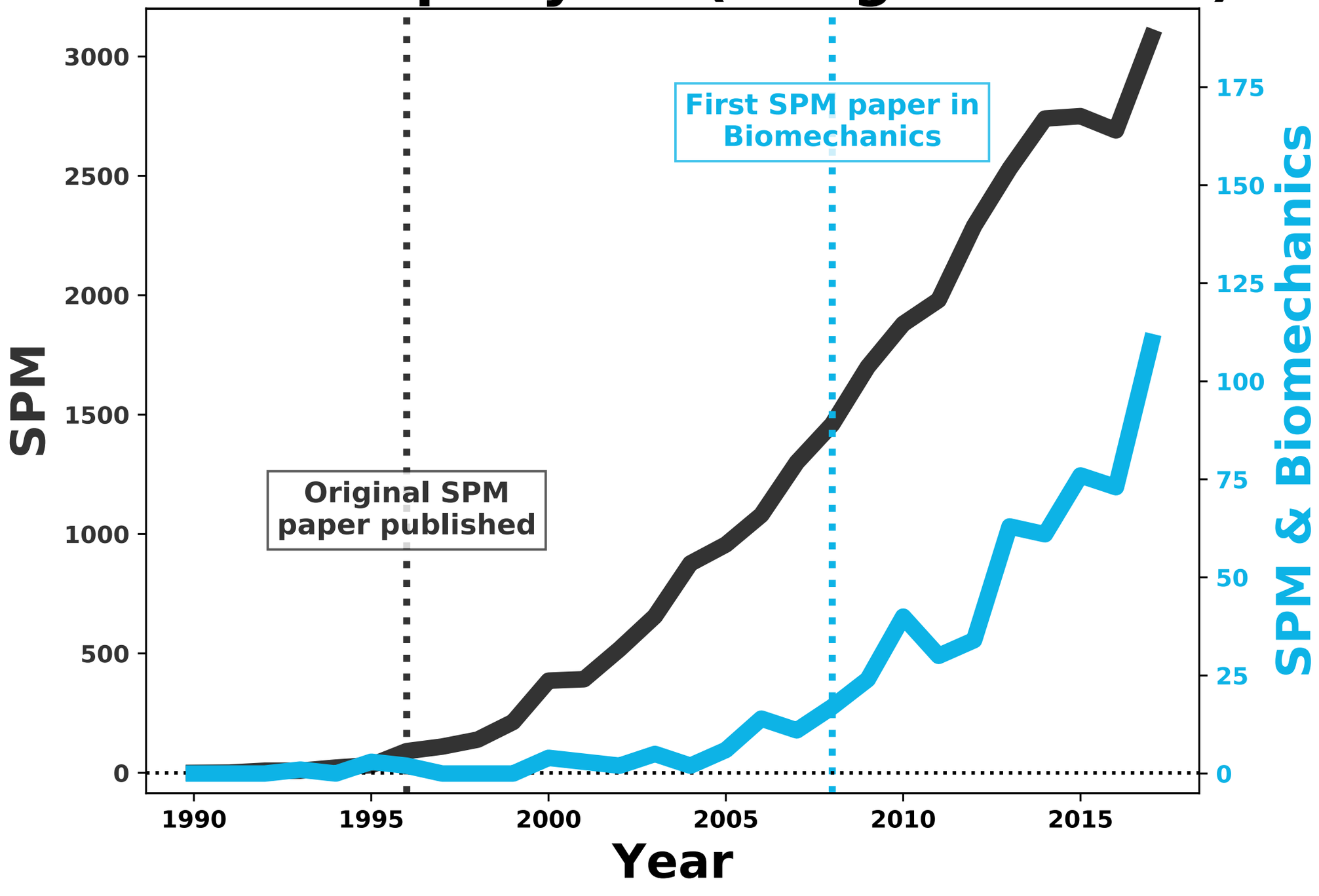
2009 Li et al. Bone 44: 596-602

Karl Friston

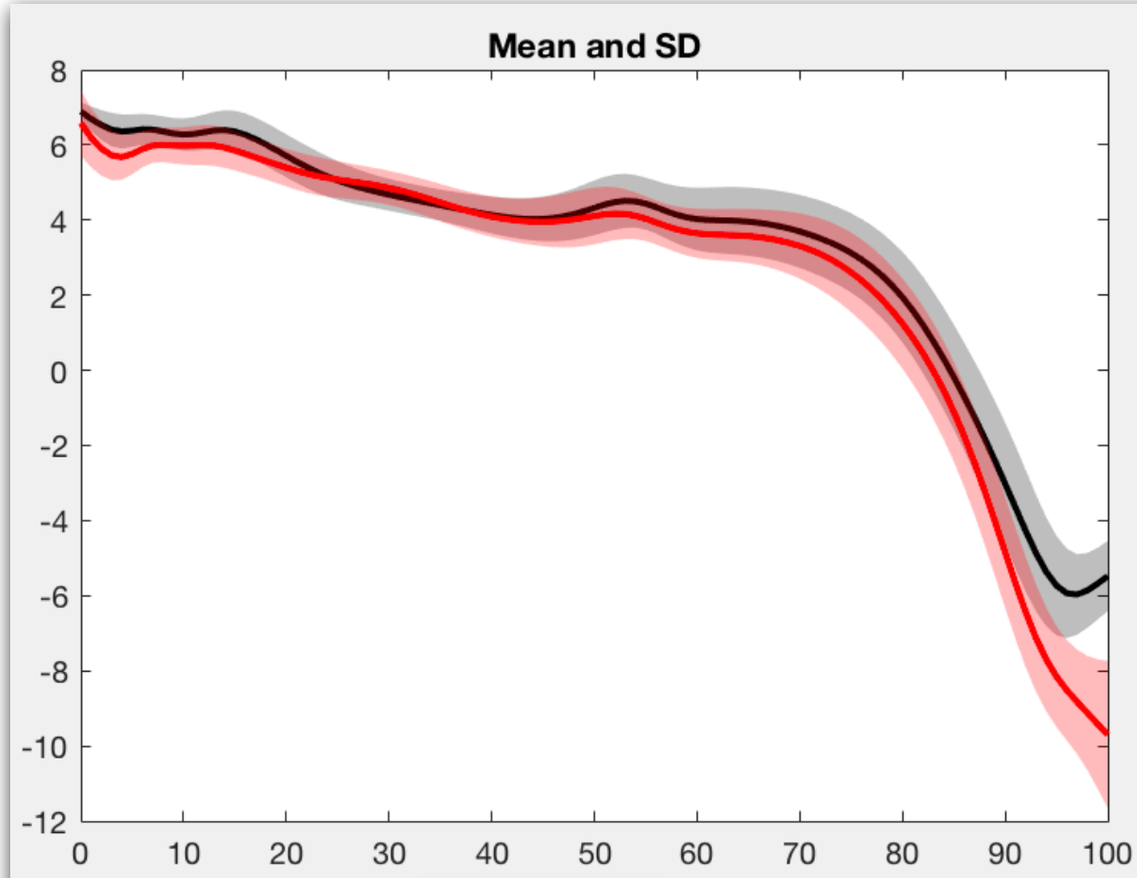
H-index: 220

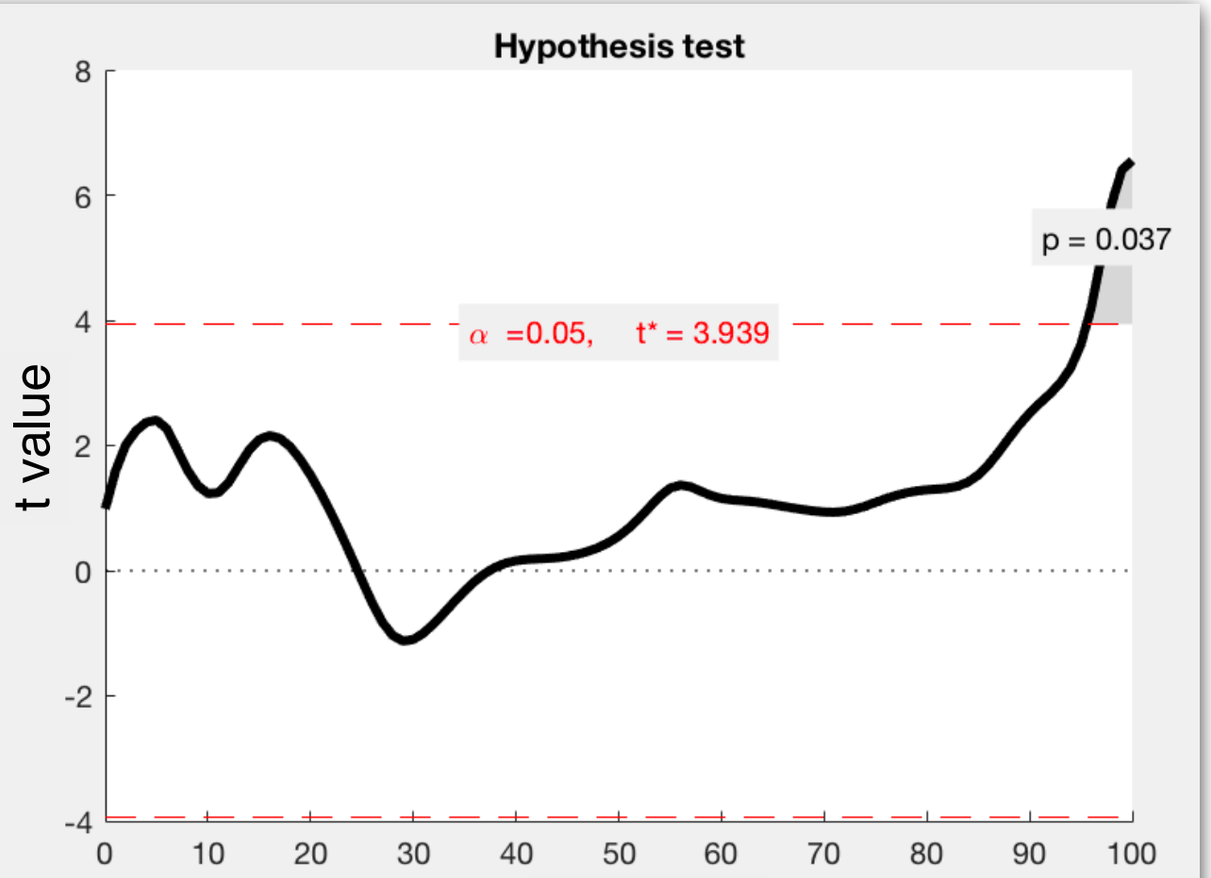
i10-index: 889

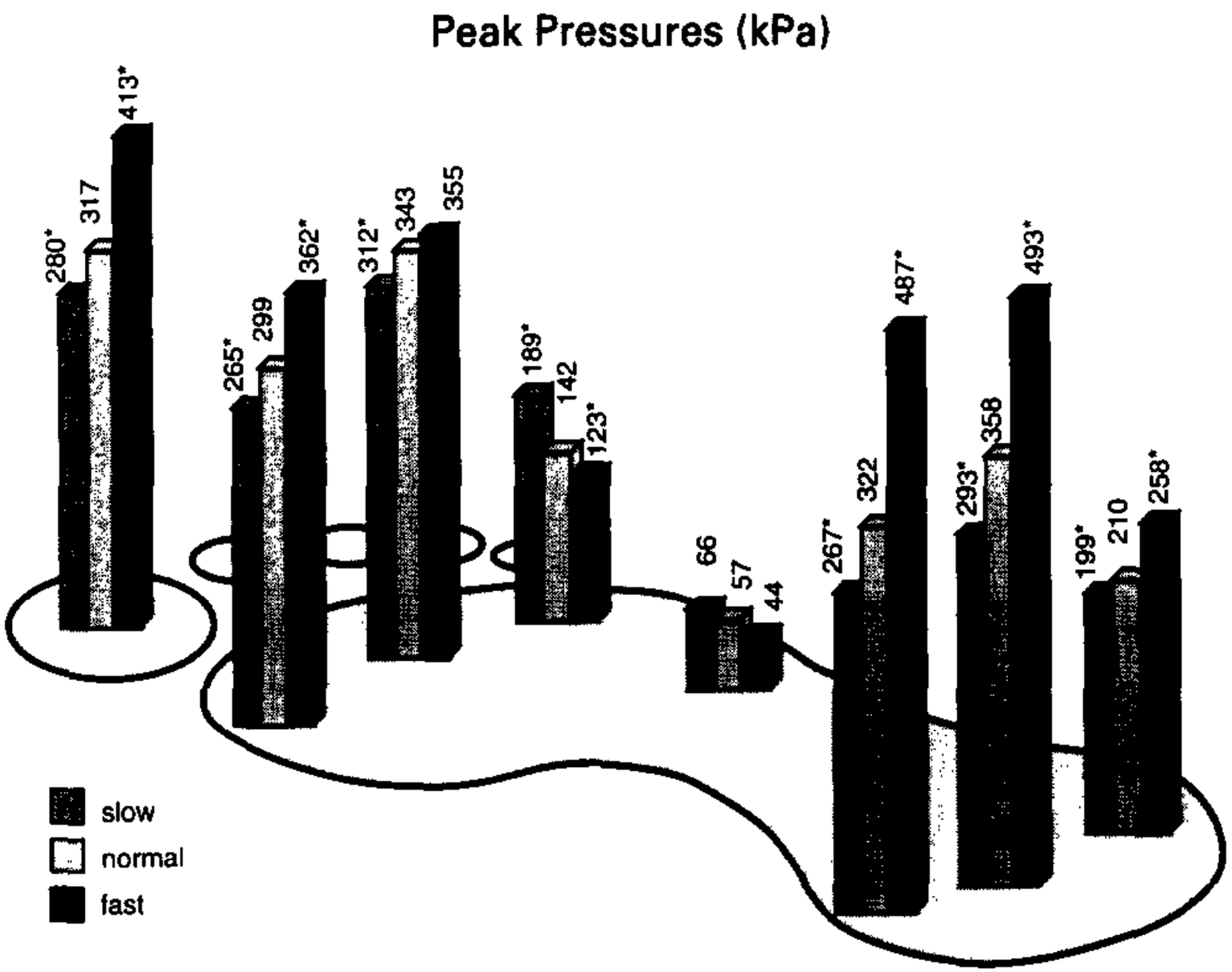
Citations per year (Google Scholar)



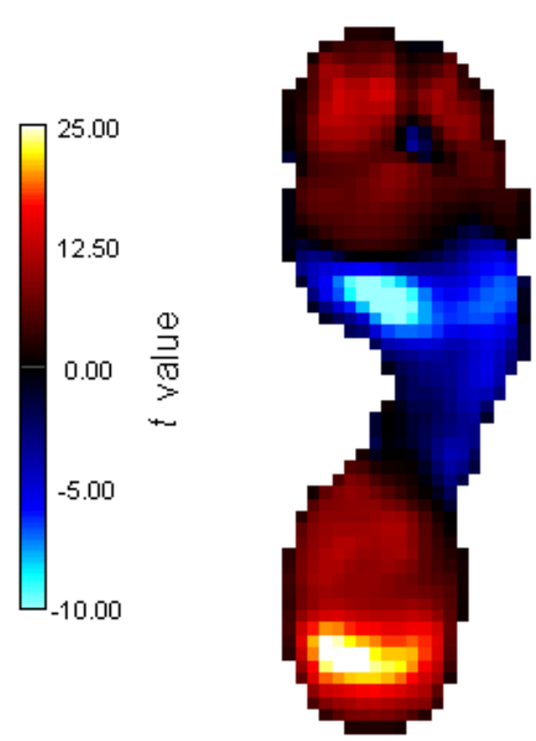
Two-sample test



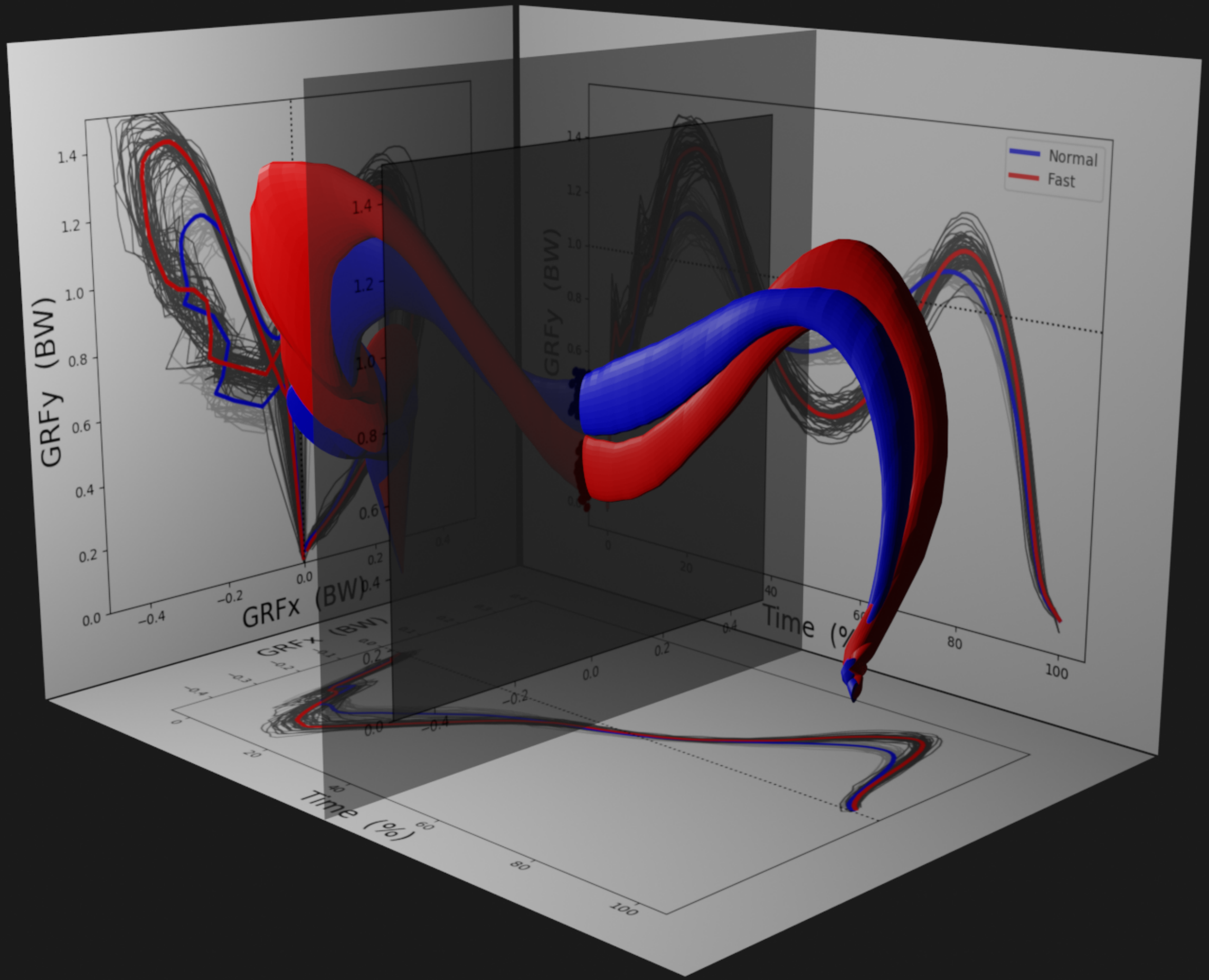




Gait & Posture 1994,2:191-7.



Pataky 2008. J Biomech 41:1987-1994.


Normal
Fast
GRFy (BW)
GRFx (BW)
Time (%)

FDA

F **FUNCTIONAL**

Data can be represented using continuous mathematical “basis” functions

D **DATA**

A **ANALYSIS**

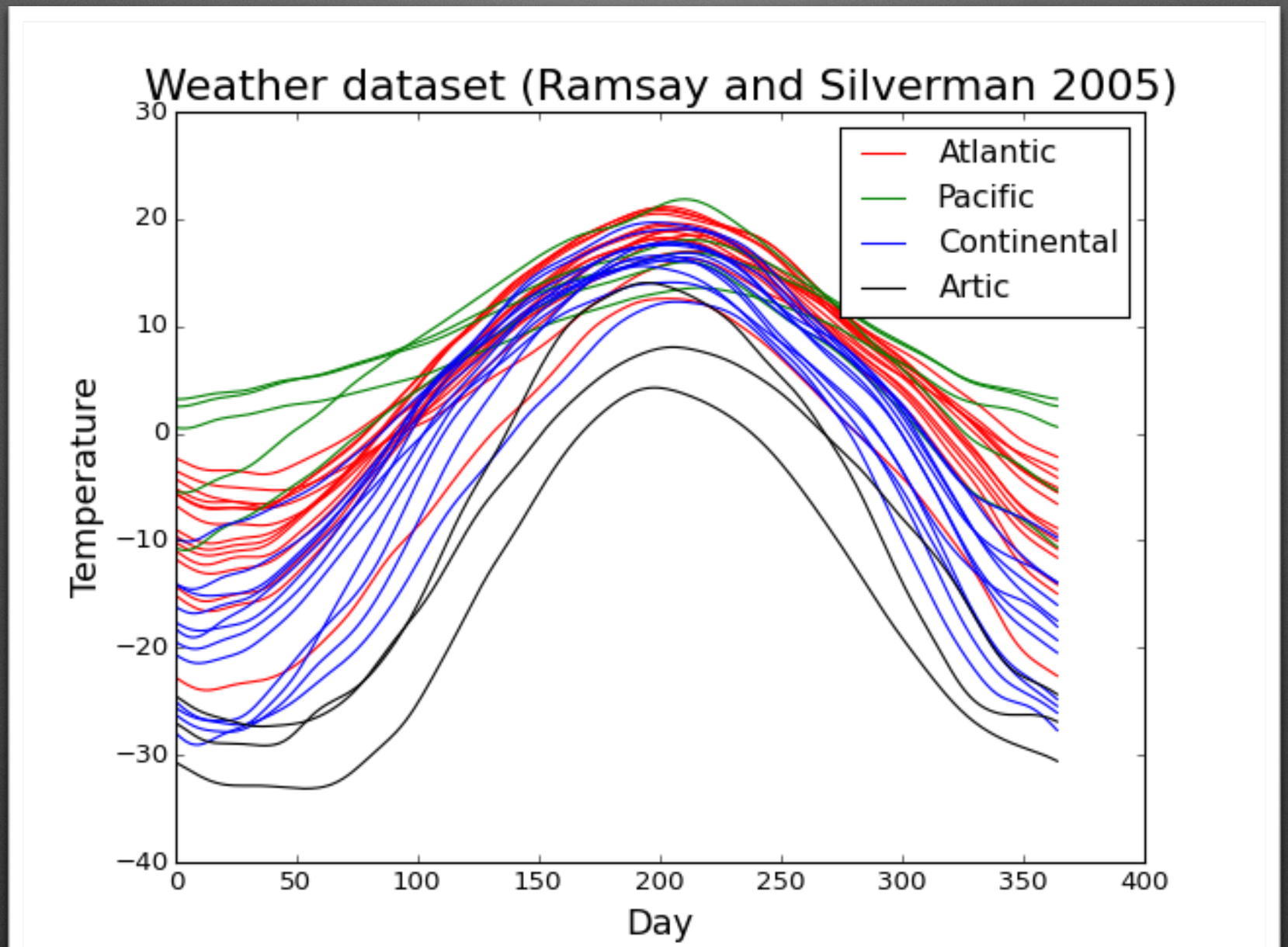
4

SPM & FDA

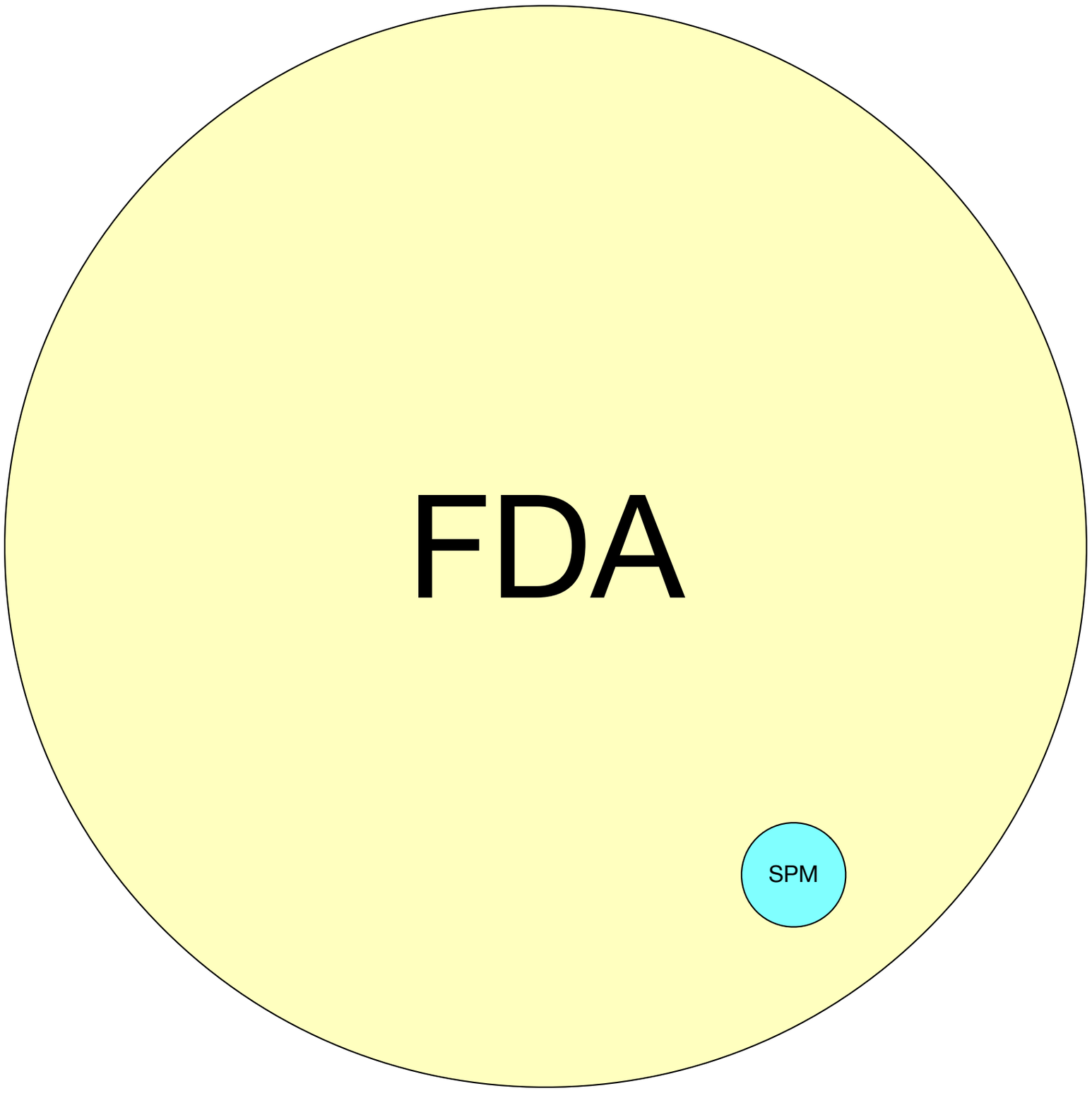
What is FDA?

"Functional data analysis refers to a collection of methods for analyzing data over a curve, surface or continuum."

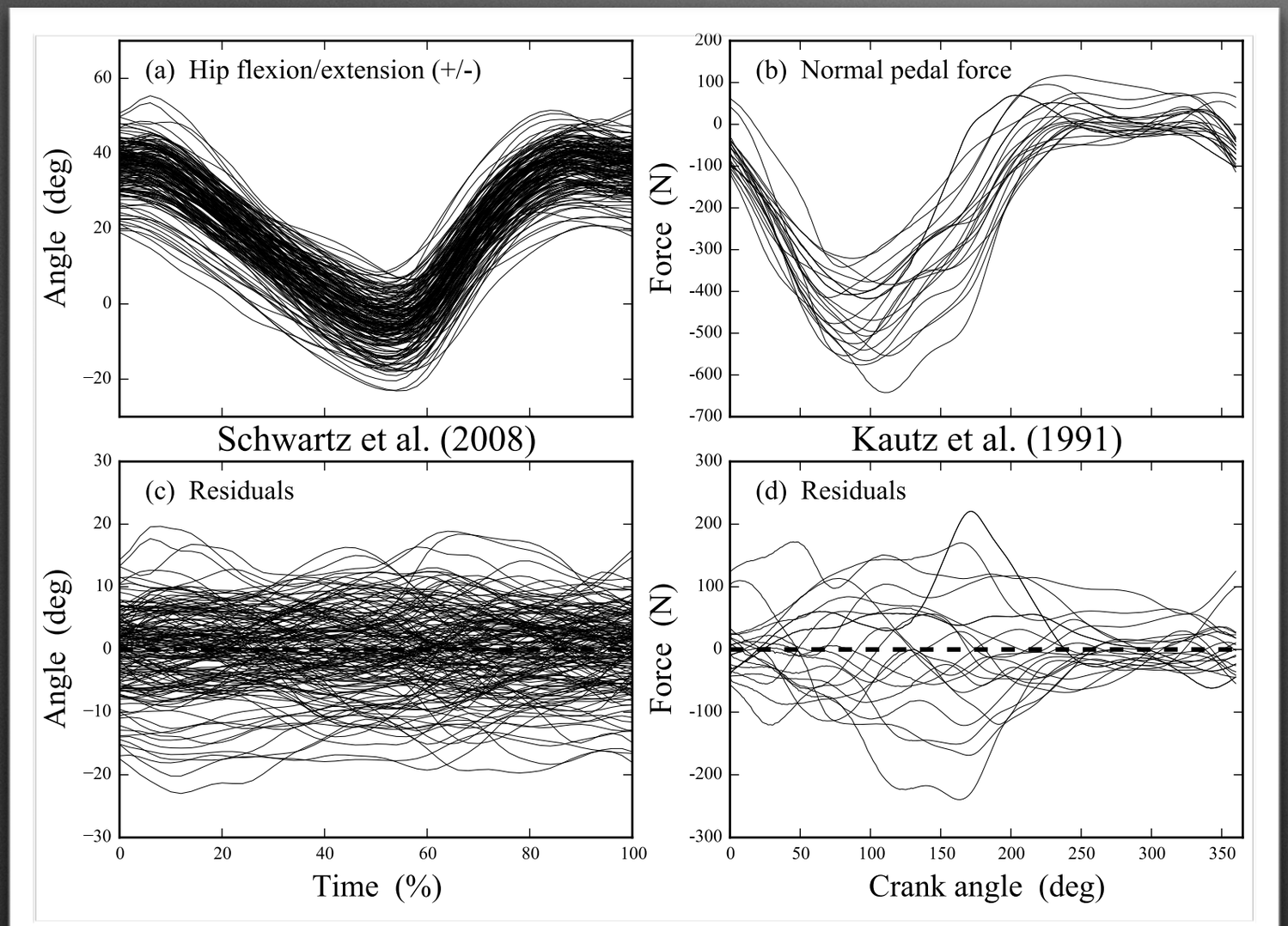
Mathew McLean, displayr.com


Weather dataset (Ramsay and Silverman 2005)
Atlantic
Pacific
Continental
Artic
Temperature
Day
30
20
10
0
−10
−20
−30
−40
0
50
100
150
200
250
300
350
400

An FDA Statistician's perspective



My perspective

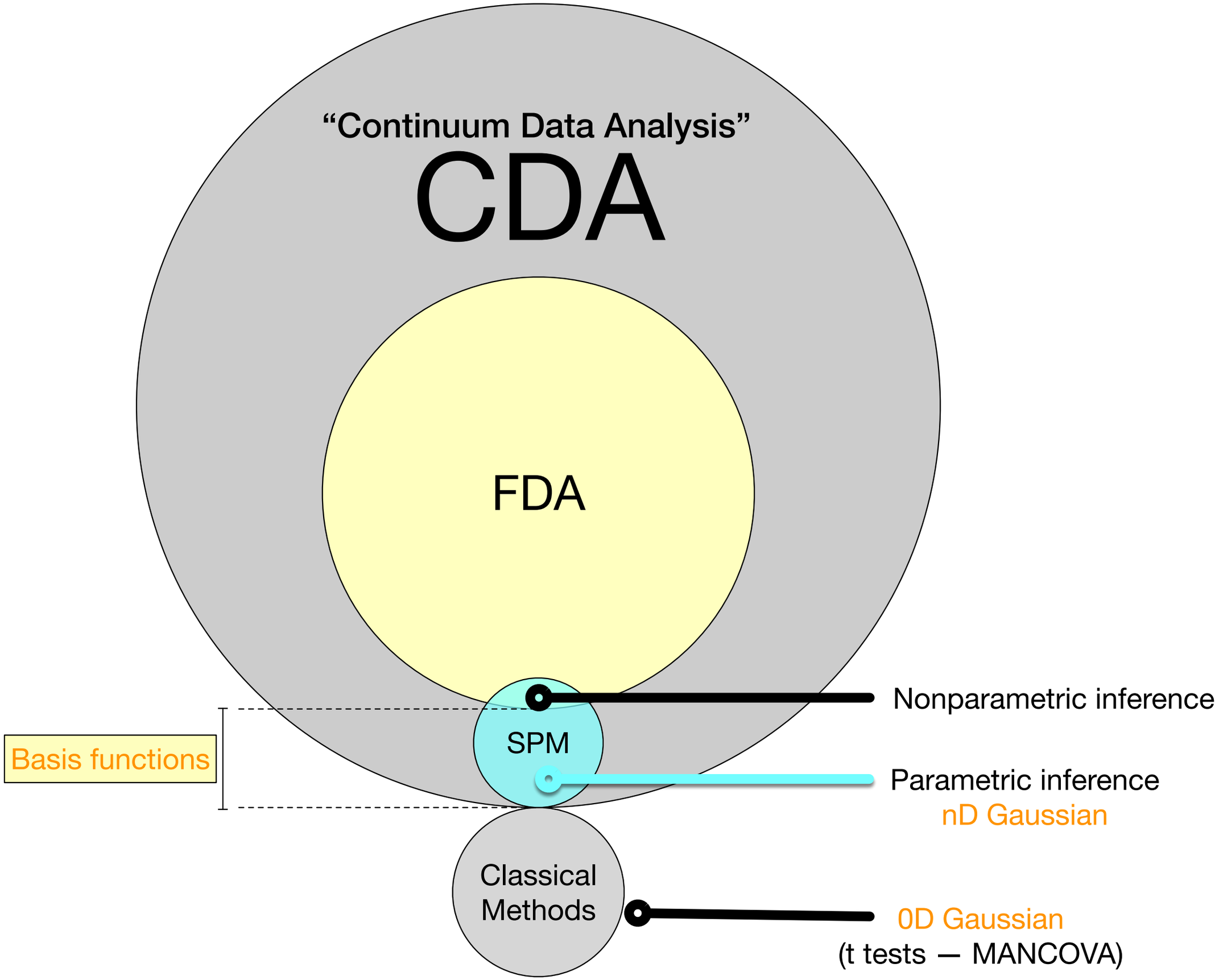


FDA models this using “basis functions”

SPM models this as Gaussian random fields

SPM shows that you don’t need basis functions to analyze continuous data

My perspective



How serious is the 0D vs. 1D problem?

Probability of finding 0D significance when no effect actually exists:

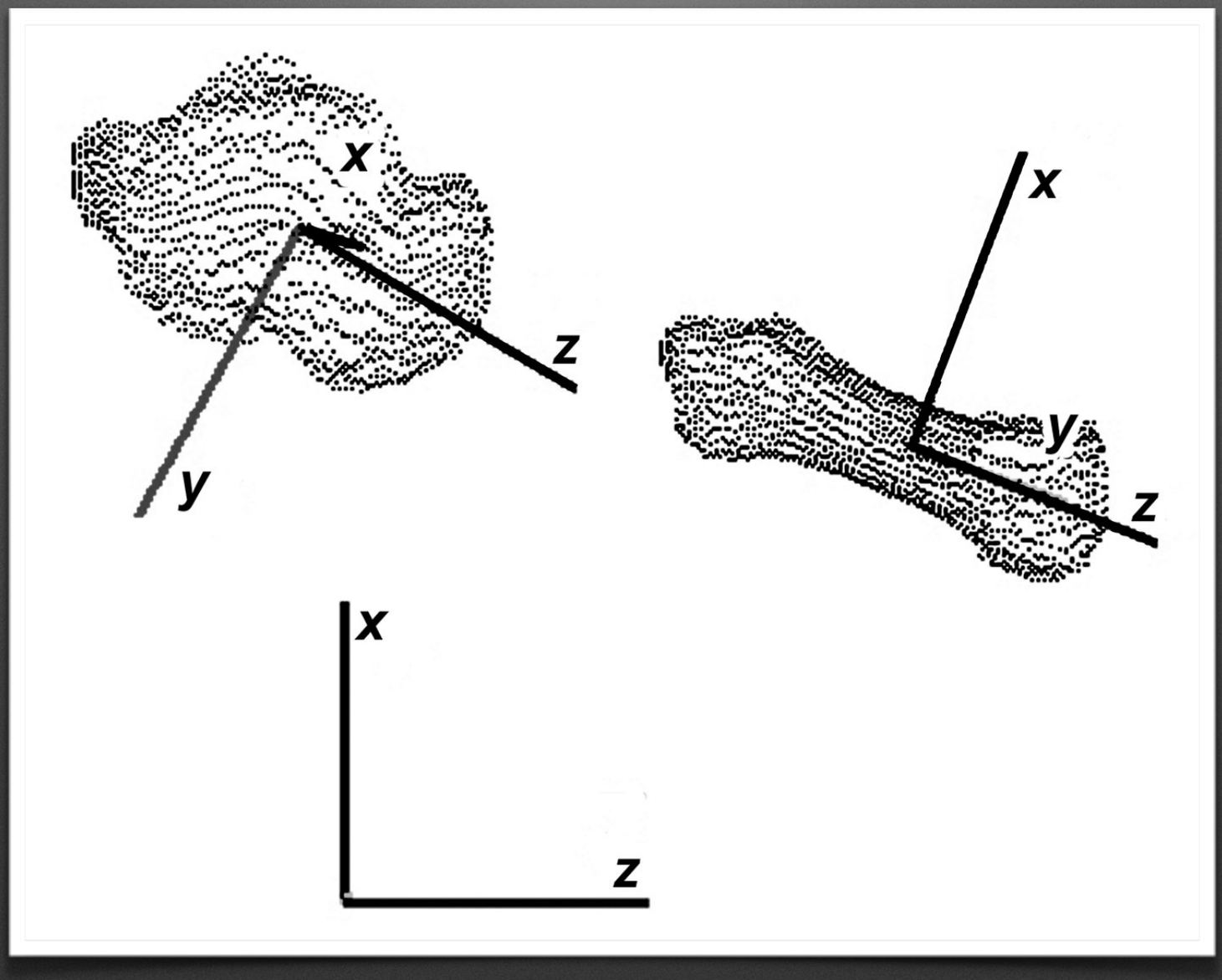
1 scalar trajectory:	34%
1 vector trajectory:	76%
2 vector trajectories:	94.5%
6 vector trajectories:	99.9%

PCA

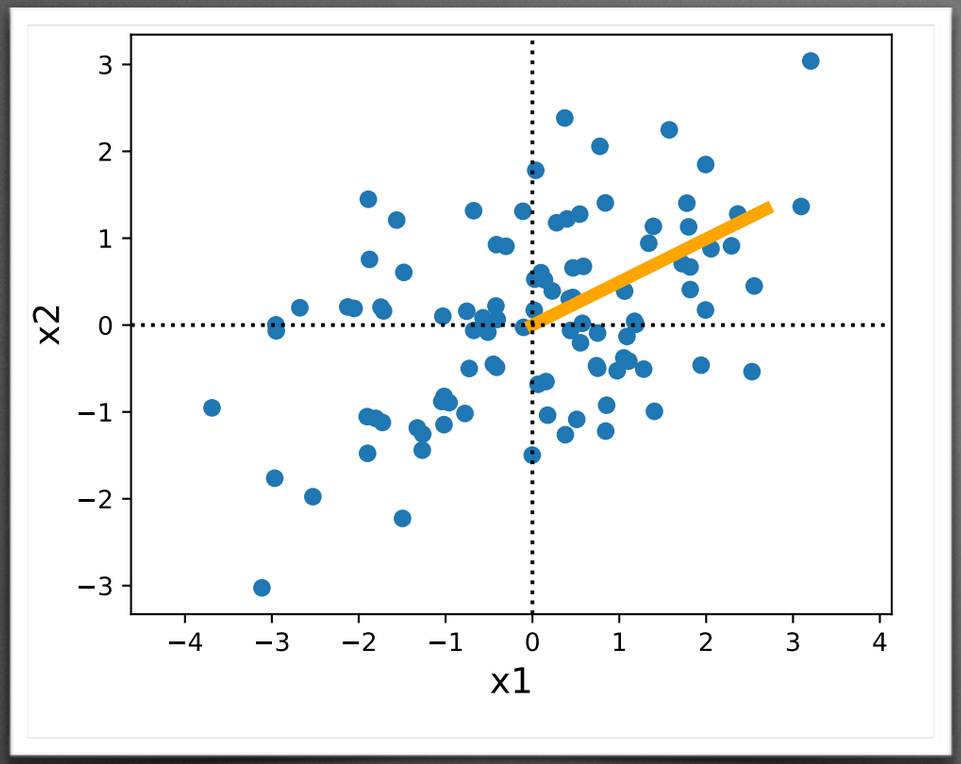
What is PCA?

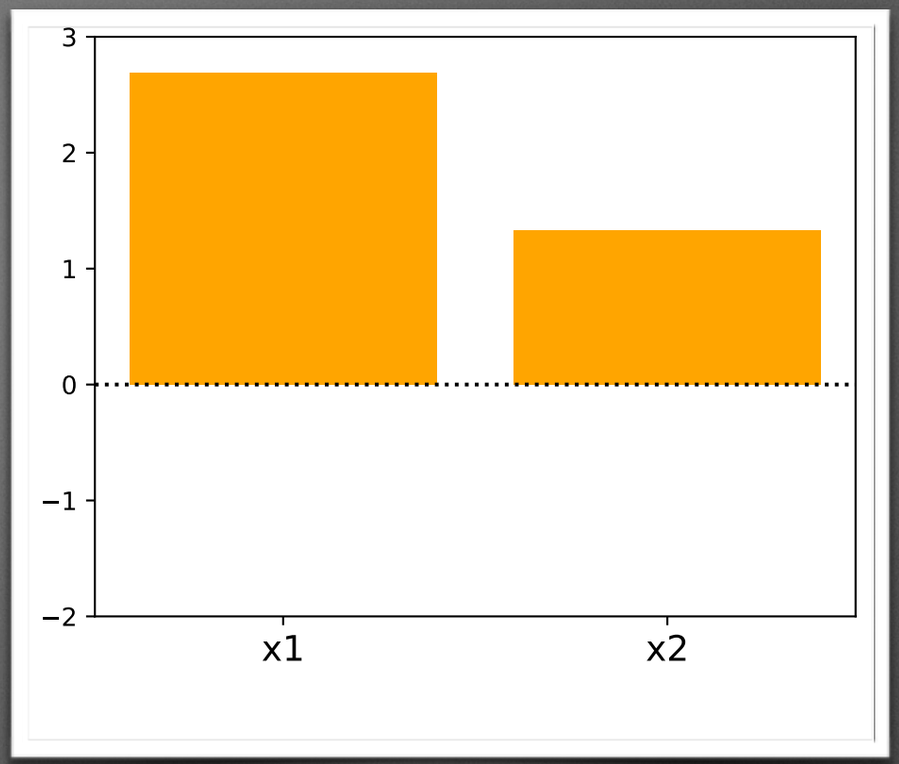
- A dimensionality reduction technique
- A coordinate system rotation
- A linear combination of multiple variables
- Reveals the directions along which maximum variance occurs

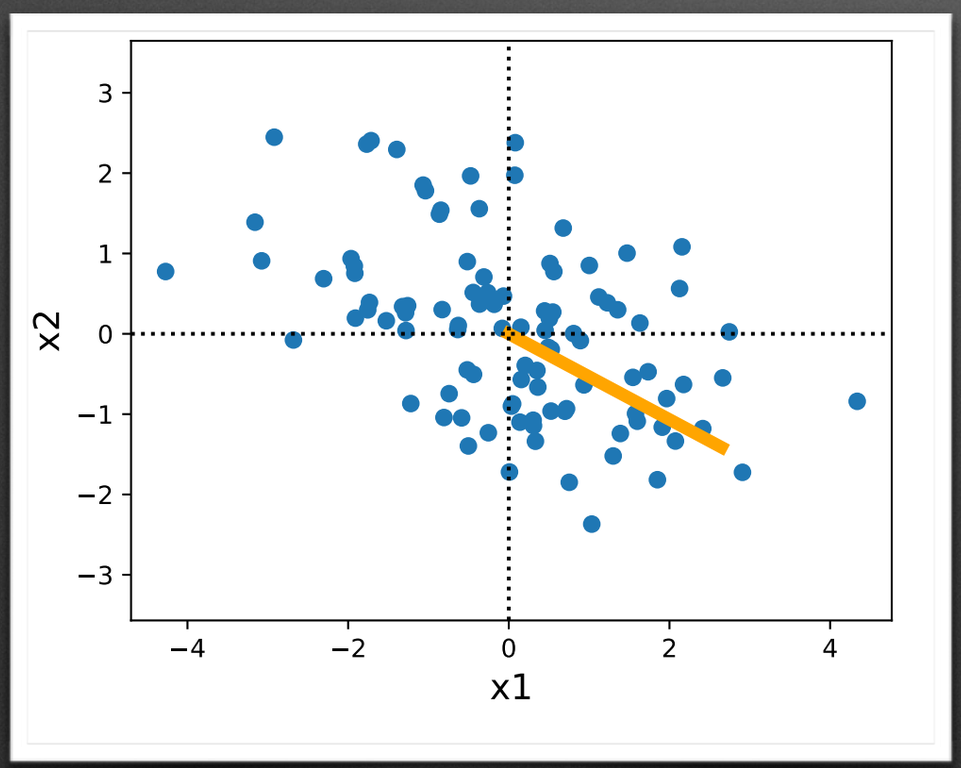
Principal axes of inertia (3D bone point masses)

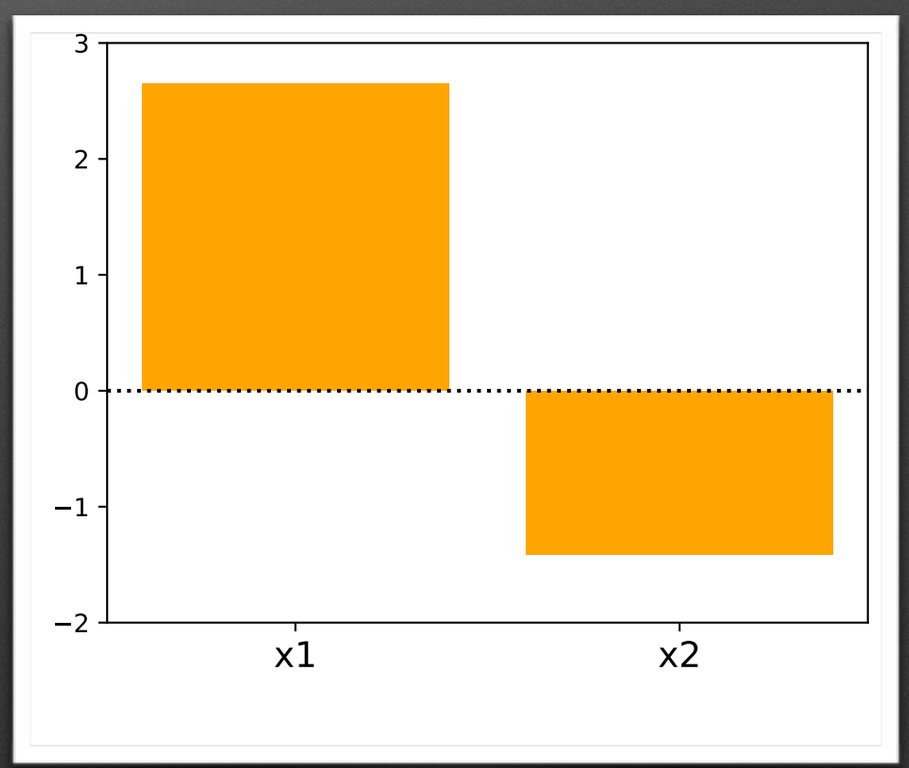


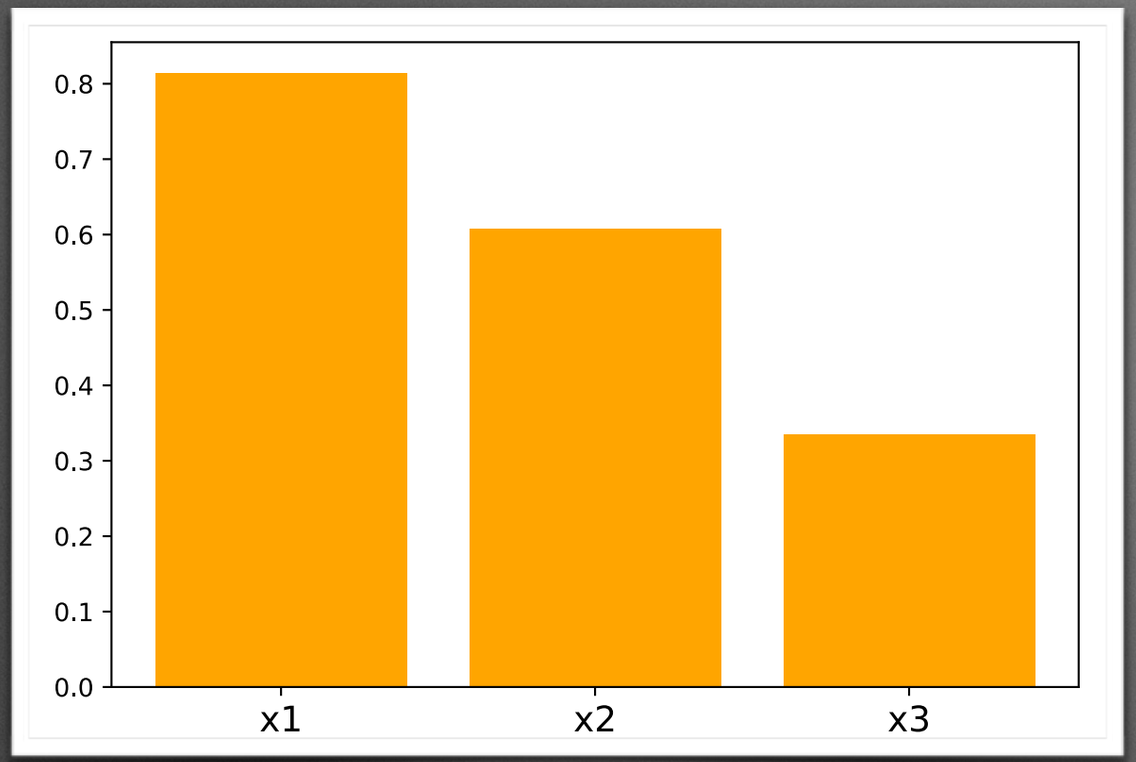
Camacho et al. (2002) *Journal of Rehabilitation Research and Development* 39: 401-410.

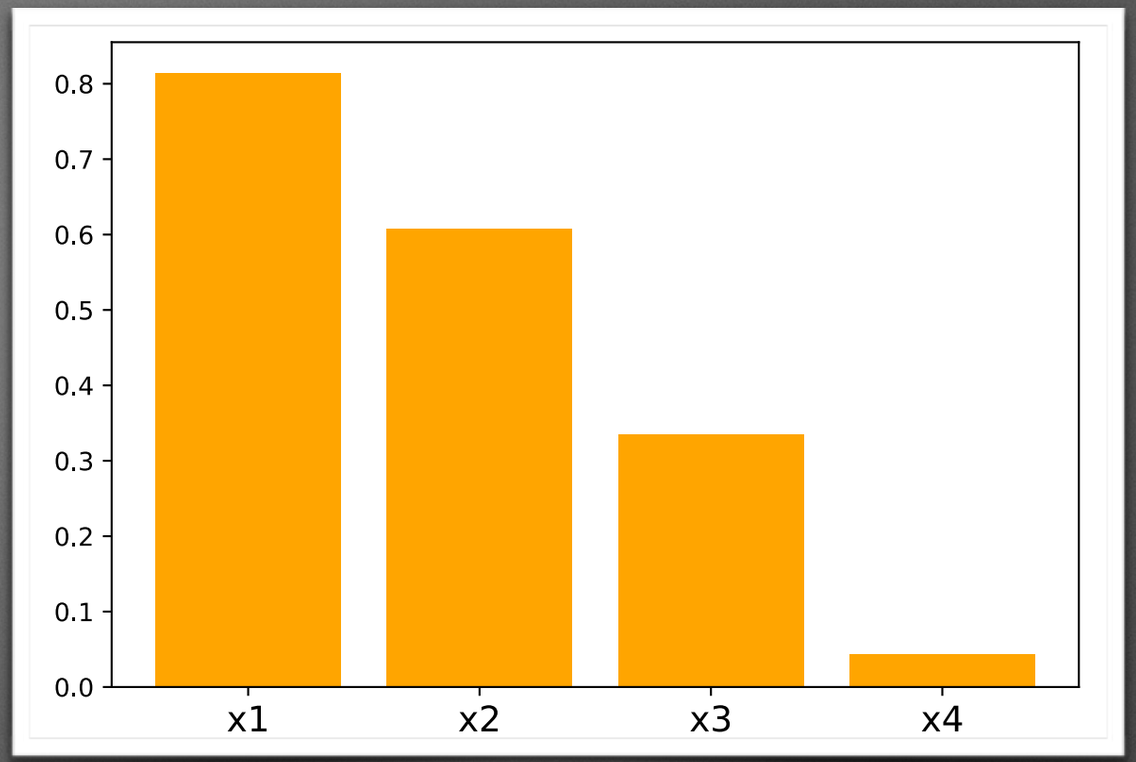

x2
x1
3
2
1
0
−1
−2
−3
−4
−3
−2
−1
0
1
2
3
4

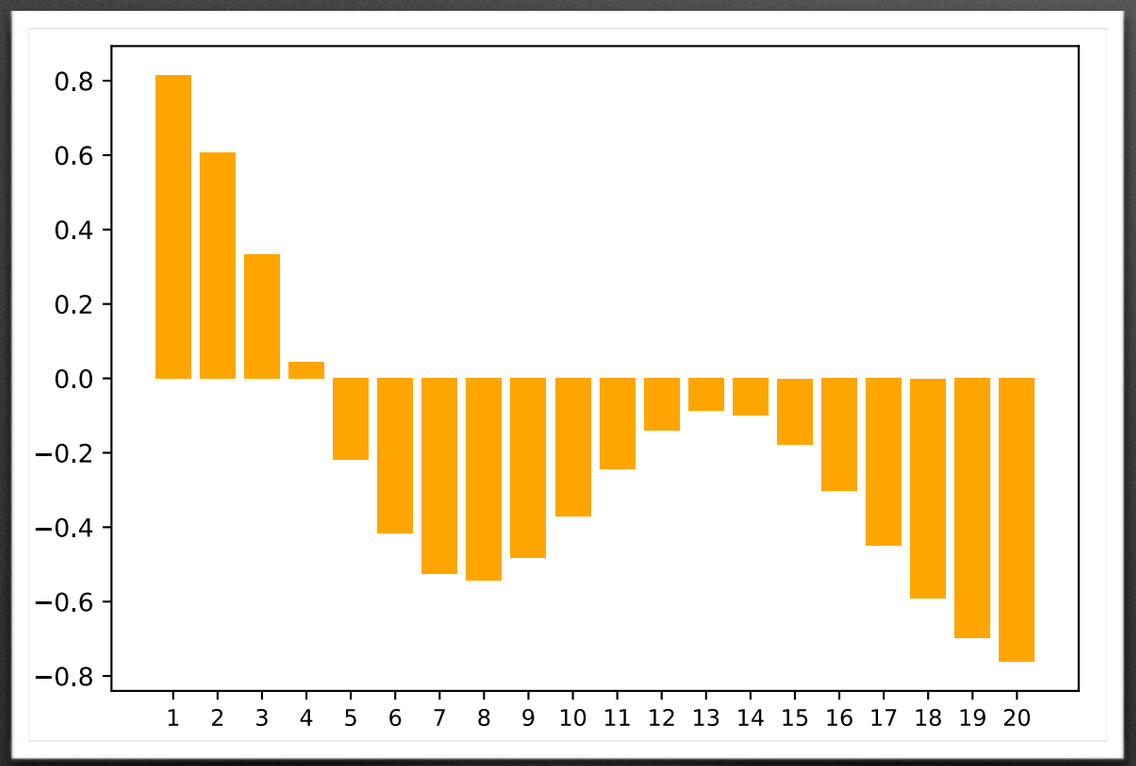

3
2
1
0
−1
−2
x1
x2

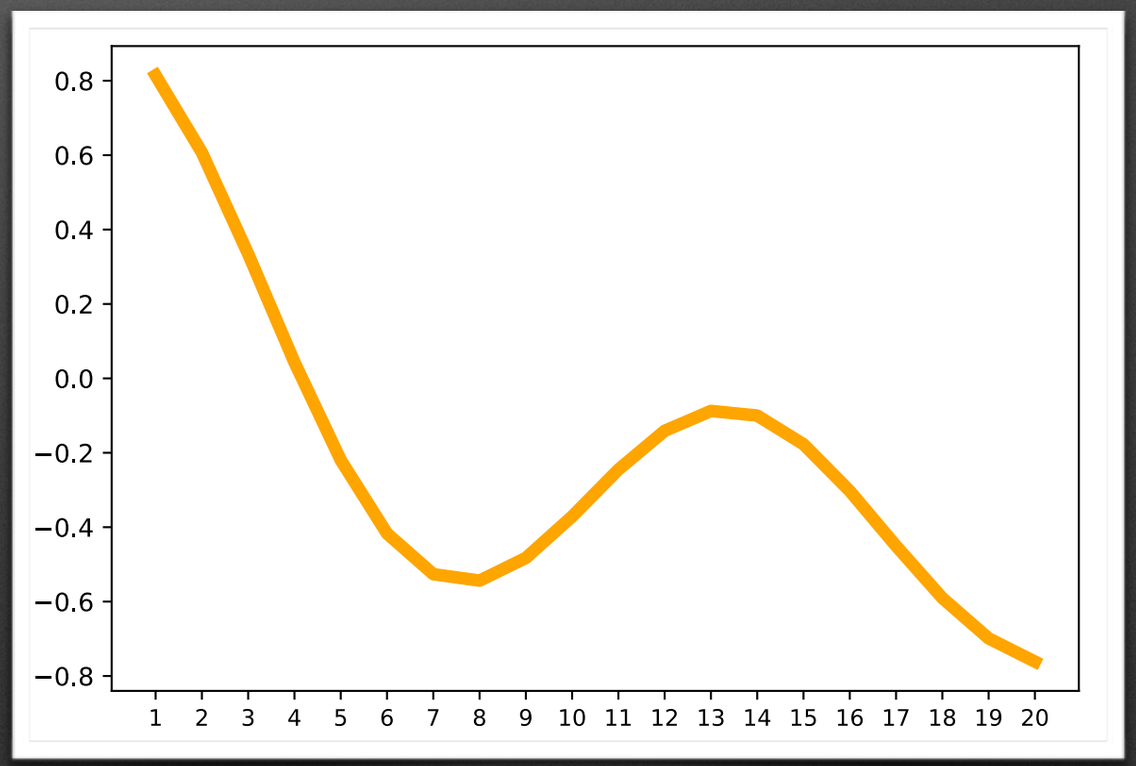

x2
x1
3
2
1
0
−1
−2
−3
−4
−2
0
2
4


3
2
1
0
−1
−2
x1
x2


0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
x1
x2
x3


0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
x1
x2
x3
x4


0.8
0.6
0.4
0.2
0.0
−0.2
−0.4
−0.6
−0.8
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20


0.8
0.6
0.4
0.2
0.0
−0.2
−0.4
−0.6
−0.8
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20

PCA, ICA, kPCA

- PCA: Orthogonal components (i.e., rotation)
- ICA: Independent components, usually non-orthogonal
- kPCA: Higher-dimensional PCA, usually for clustering

- In Biomechanics: usually used for exploration (pattern discovery)

Machine Learning

What is Machine Learning?

- A collection of methods for objectively identifying patterns in complex datasets
- Usually dimensionality reduction is applied before and/or along with ML techniques

Purpose of ML

- Usually to solve an engineering problem
- e.g. Self-driving cars

Common ML algorithms

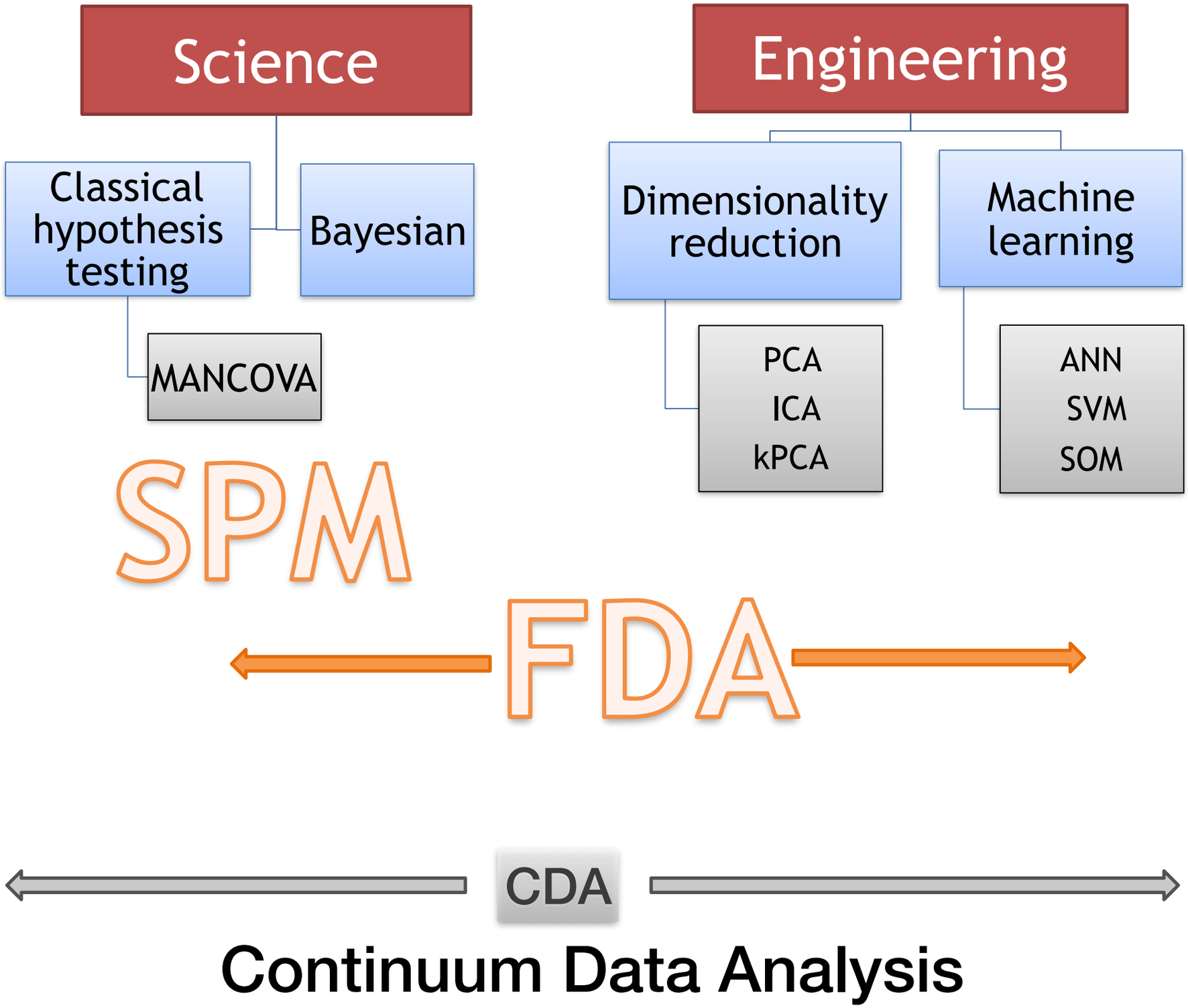
- Clustering (e.g. k-means clustering)
- Neural networks
- Support vector machines
- Self-organizing maps

ML uses

- In Biomechanics: usually used for exploration (pattern discovery)

Test theoretical predictions

Find / exploit patterns



SPM

FDA

CDA

Continuum Data Analysis

Common issues

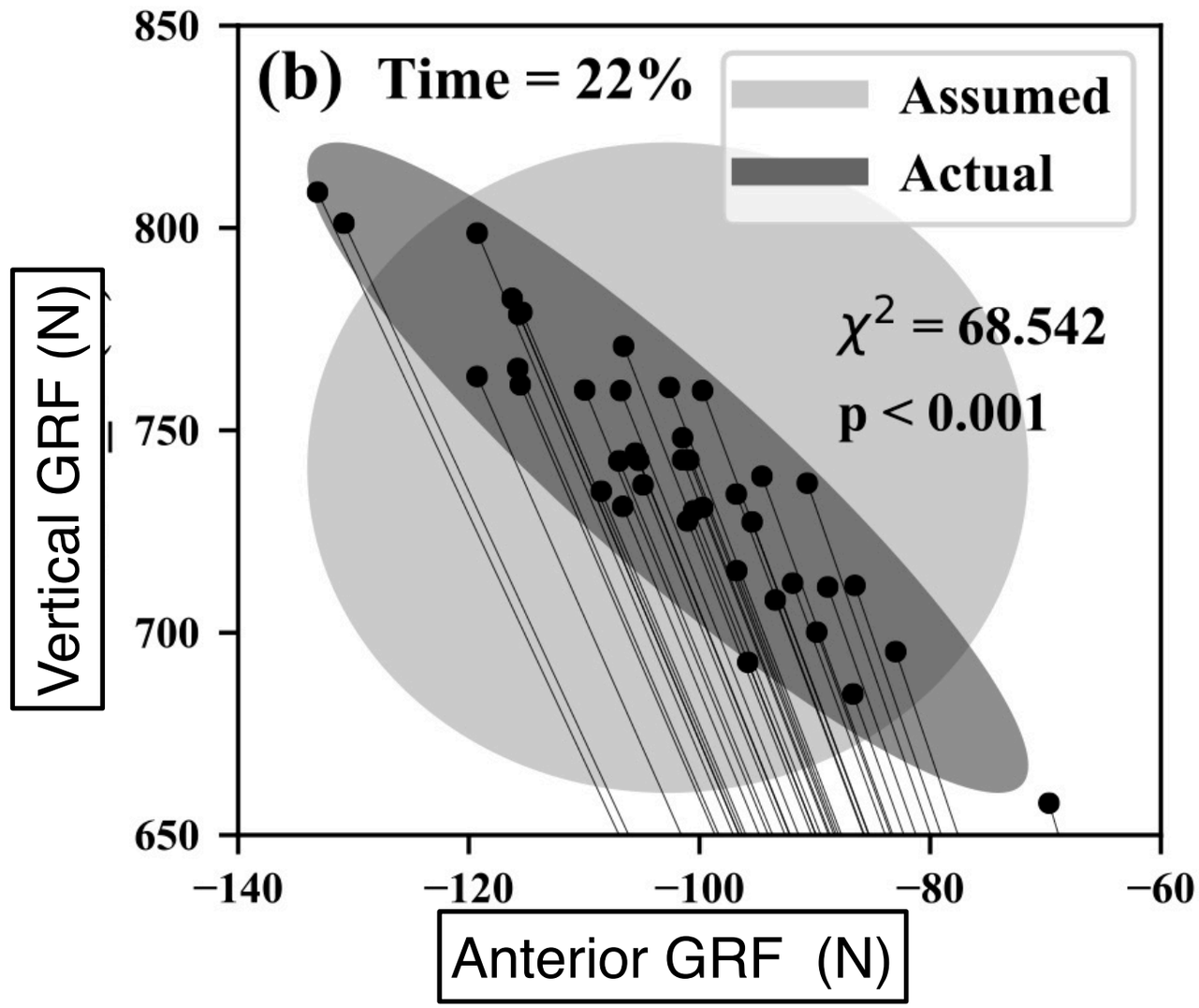
Common Issues

- Segmentation
- Registration
- Normality
- Processing Sensitivity
- Covariance
- Nonuniform smoothness
- etc.

Covariance

Methods

Statistical test



Test for the equality of covariance matrices

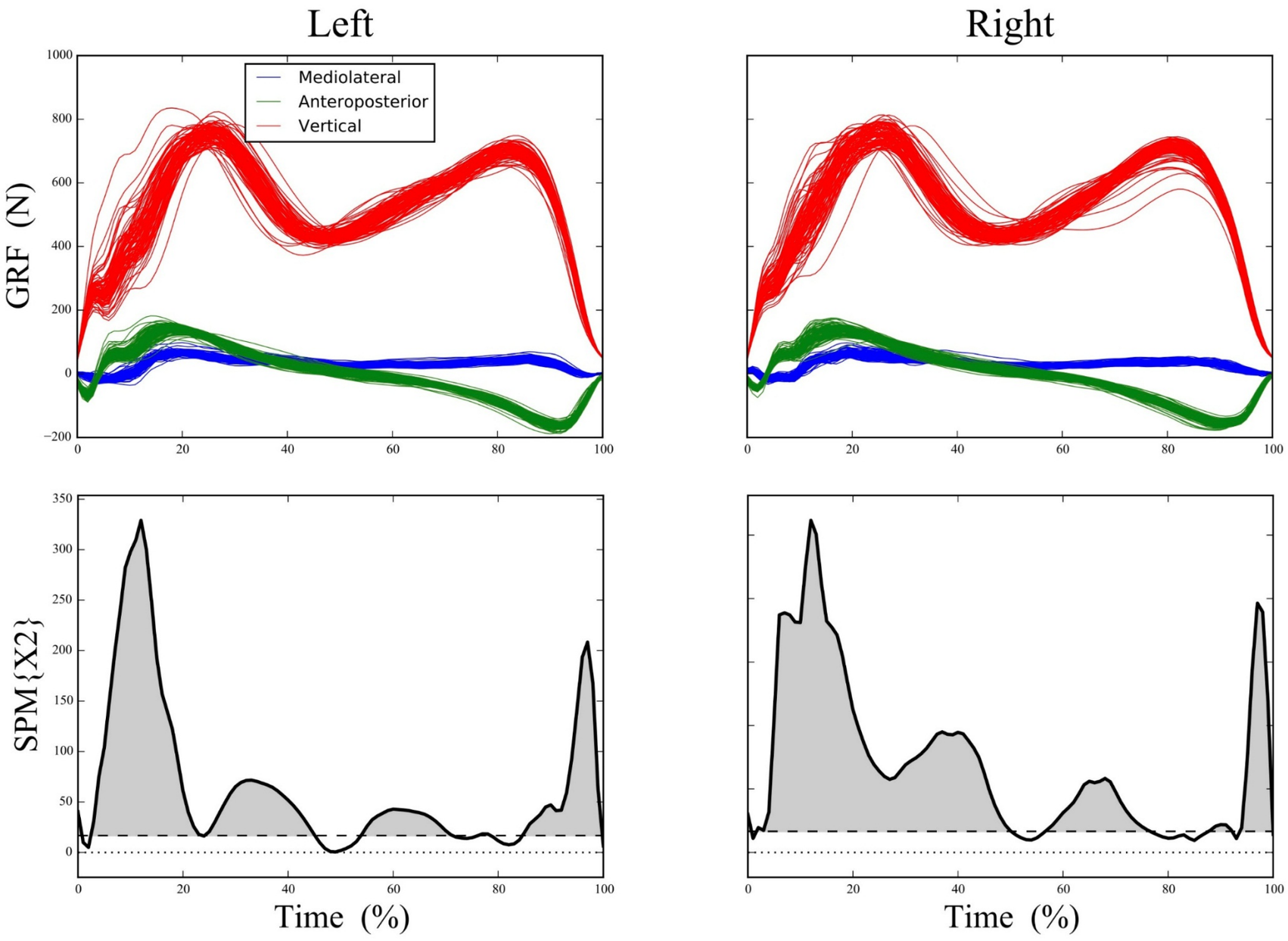
Rencher and Christensen, Methods of Multivariate Analysis 2012, Wiley; 261–262.

Adapted for time series analysis using Random Field Theory

Adler and Taylor, Random Fields and Geometry 2007, Springer-Verlag.

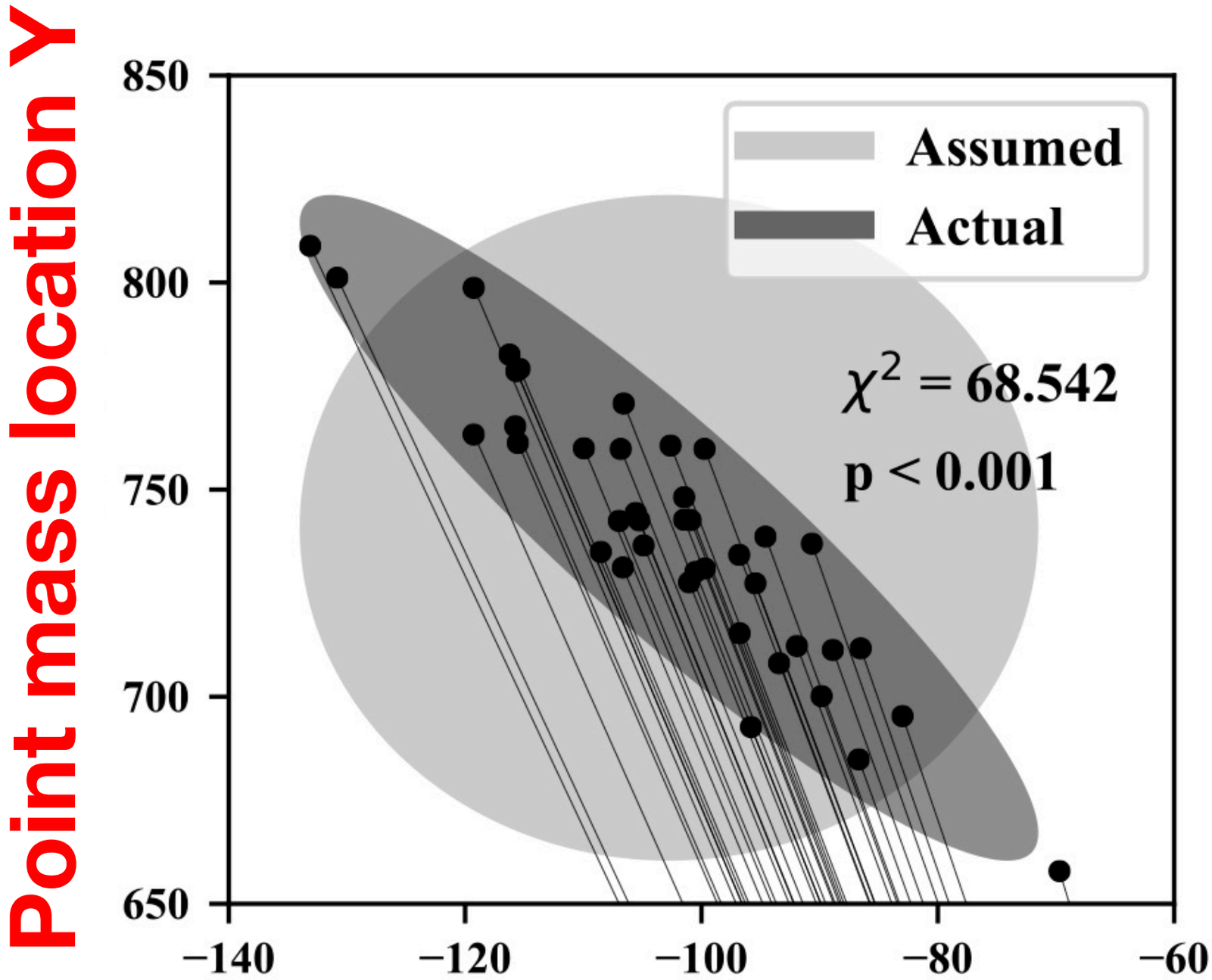
Results

Example subject, Walking



Separate plane analysis is statistically invalid

Discussion


Point mass location Y
850
800
750
700
650
−140
−120
−100
−80
−60
Assumed
Actual
$\chi^2 = 68.542$
$p < 0.001$
Point mass location X

Overview

- History
- The p value
- Classical techniques

Questions

- Emerging techniques
- Controversies
- The future

Questions

“Despite frequent calls for the overhaul of null hypothesis significance testing (NHST), this controversial procedure remains ubiquitous in behavioral, social and biomedical teaching and research.”

–Jose D. Perezgonzalez (Frontiers in Psychology, 2015)

SPM

“Many correlations reported in recent (SPM) literature are impossibly high.”

–Edward Vul et al. (2009) Perspectives on Psychological Science

(Lots of reviews)

"This has not been validated for biomechanical data."

- Skeptical Reviewer

p value

The fickle *P* value generates irreproducible results

Lewis G Halsey, Douglas Curran-Everett, Sarah L Vowler & Gordon B Drummond

The reliability and reproducibility of science are under scrutiny. However, a major cause of this lack of repeatability is not being considered: the wide sample-to-sample variability in the *P* value. We explain why *P* is fickle to discourage the ill-informed practice of interpreting analyses based predominantly on this statistic.

In defense of *P* values

Paul A. Murtaugh[1]

Department of Statistics, Oregon State University, Corvallis, Oregon 97331 USA

Abstract. Statistical hypothesis testing has been widely criticized by ecologists in recent years. I review some of the more persistent criticisms of *P* values and argue that most stem from misunderstandings or incorrect interpretations, rather than from intrinsic shortcomings of the *P* value. I show that *P* values are intimately linked to confidence intervals and to differences in Akaike's information criterion (ΔAIC), two metrics that have been advocated as replacements for the *P* value. The choice of a threshold value of ΔAIC that breaks ties among competing models is as arbitrary as the choice of the probability of a Type I error in hypothesis testing, and several other criticisms of the *P* value apply equally to ΔAIC. Since *P* values, confidence intervals, and ΔAIC are based on the same statistical information, all have their places in modern statistical practice. The choice of which to use should be stylistic, dictated by details of the application rather than by dogmatic, a priori considerations.

Key words: *AIC; confidence interval; null hypothesis;* P *value*; *significance testing*.

STATISTICAL ERRORS

P values, the 'gold standard' of statistical validity, are not as reliable as many scientists assume.

BY REGINA NUZZO

For a brief moment in 2010, Matt Motyl was on the brink of scientific glory: he had discovered that extremists quite literally see the world in black and white.

The results were "plain as day", recalls Motyl, a psychology PhD student at the University of Virginia in Charlottesville. Data from a study of nearly 2,000 people seemed to show that political moderates saw shades of grey more accurately than did either left-wing or right-wing extremists. "The hypothesis was sexy," he says, "and the data provided clear support." The *P* value, a common index for the strength of evidence, was 0.01 — usually interpreted as 'very significant'. Publication in a high-impact journal seemed within Motyl's grasp.

But then reality intervened. Sensitive to controversies over reproducibility, Motyl and his adviser, Brian Nosek, decided to replicate the study. With extra data, the *P* value came out as 0.59 — not even close to the conventional level of significance, 0.05. The effect had disappeared, and with it, Motyl's dreams of youthful fame[1].

It turned out that the problem was not in the data or in Motyl's analyses. It lay in the surprisingly slippery nature of the *P* value, which is neither as reliable nor as objective as most scientists assume. "*P* values are not doing their job, because they can't," says Stephen Ziliak, an economist at Roosevelt University in Chicago, Illinois, and a frequent critic of the way statistics are used.

For many scientists, this is especially worrying in light of the reproducibility concerns. In 2005, epidemiologist John Ioannidis of Stanford University in California suggested that most published findings are false[2]; since then, a string of high-profile replication problems has forced scientists to rethink how they evaluate results.

At the same time, statisticians are looking for better ways of thinking about data, to help scientists to avoid missing important information or acting on false alarms. "Change your statistical philosophy and all of a sudden different things become important," says Steven Goodman, a physician and statistician at Stanford. "Then 'laws' handed down from God are no longer handed down from God. They're actually handed down to us by ourselves, through the methodology we adopt."

DALE EDWIN MURRAY

OUT OF CONTEXT

P values have always had critics. In their almost nine decades of existence, they have been likened to mosquitoes (annoying and impossible to swat away), the emperor's new clothes (fraught with obvious problems that everyone ignores) and the tool of a "sterile intellectual rake" who ravishes science but leaves it with no progeny[3]. One researcher suggested rechristening the methodology "statistical hypothesis inference testing"[3], presumably for the acronym it would yield.

The irony is that when UK statistician Ronald Fisher introduced the *P* value in the 1920s, he did not mean it to be a definitive test. He intended it simply as an informal way to judge whether evidence was significant in the

Historical Controversies

Null hypothesis significance testing

Fisher vs. Neyman-Pearson (1940s)

Exploratory vs. confirmatory analysis

Tukey (1980)

Multiple comparisons problem

Tukey, Scheffe (1950s), Bonferroni (1970s)

Magnitude-Based Inference

MBI

SPORTSCIENCE sportsci.org
Perspectives / Research Resources

Making Meaningful Inferences About Magnitudes

Alan M Batterham, Will G Hopkins

Sportscience 9, 6-13, 2005 (sportsci.org/jour/05/ambwgh.htm)
School of Health and Social Care, University of Teesside, Middlesbrough, UK; Sport and Recreation, AUT University, Auckland 1020, New Zealand. Email. Reviewer: Stephen W Marshall, Dept of Epidemiology, University of North Carolina, Chapel Hill Chapel Hill, NC 27599-7435, USA.

A study of a sample provides only an estimate of the true (population) value of an outcome statistic. A report of the study therefore usually includes an inference about the true value. Traditionally, a researcher makes an inference by declaring the value of the statistic statistically significant or non-significant on the basis of a p value derived from a null hypothesis test. This approach is confusing and can be misleading, depending on the magnitude of the statistic, error of measurement, and sample size. We use a more intuitive and practical approach based directly on uncertainty in the true value of the statistic. First we express the uncertainty as confidence limits, which define the likely range of the true value. We then deal with the real-world relevance of this uncertainty by taking into account values of the statistic that are substantial in some positive and negative sense, such as beneficial and harmful. If the likely range overlaps substantially positive and negative values, we infer that the outcome is unclear; otherwise, we infer that the true value has the magnitude of the observed value: substantially positive, trivial, or substantially negative. We refine this crude inference by stating qualitatively the likelihood that the true value will have the observed magnitude (e.g., very likely beneficial). Quantitative or qualitative probabilities that the true value has the other two magnitudes or more finely graded magnitudes (such as trivial, small, moderate, and large) can also be estimated to guide a decision about the utility of the outcome.

KEYWORDS: clinical significance, confidence limits, statistical significance.

Reprint pdf · Reprint doc · Commentary by Stephen Marshall · Update

Researchers usually conduct a study by selecting a sample of subjects from some population, collecting the data, then calculating the value of a statistic that summarizes the outcome. In almost every imaginable study, a different sample would produce a different value for the outcome statistic, and of course none would be the value the researchers are most interested in–the value obtained by studying the entire population. Researchers are therefore expected to make an inference about the population value of the statistic when they report their findings in a scientific journal. In this article we first critique the traditional approach to inferential statistics, the null-hypothesis test. Next we explain confidence limits, which have begun to appear in publications in response to a growing awareness that the null-hypothesis test fails to deal with the real-world significance of an outcome. We then show that confidence limits alone also fail, before outlining our own approach and other approaches to making inferences based on meaningful magnitudes.

The Null-Hypothesis Test

The almost universal approach to inferential statistics has been the null hypothesis test, in which the researcher uses a statistical package to produce a p value for an outcome statistic. The p value is the probability of obtaining any

OPEN

The Problem with "Magnitude-based Inference"

KRISTIN L. SAINANI

Division of Epidemiology, Department of Health Research and Policy, Stanford University, Stanford, CA

ABSTRACT

SAINANI, K. L. The Problem with "Magnitude-based Inference." *Med. Sci. Sports Exerc.*, Vol. 50, No. 10, pp. 2166–2176, 2018. **Purpose**: A statistical method called "magnitude-based inference" (MBI) has gained a following in the sports science literature, despite concerns voiced by statisticians. Its proponents have claimed that MBI exhibits superior type I and type II error rates compared with standard null hypothesis testing for most cases. I have performed a reanalysis to evaluate this claim. **Methods**: Using simulation code provided by MBI's proponents, I estimated type I and type II error rates for clinical and nonclinical MBI for a range of effect sizes, sample sizes, and smallest important effects. I plotted these results in a way that makes transparent the empirical behavior of MBI. I also reran the simulations after correcting mistakes in the definitions of type I and type II error provided by MBI's proponents. Finally, I confirmed the findings mathematically; and I provide general equations for calculating MBI's error rates without the need for simulation. **Results**: Contrary to what MBI's proponents have claimed, MBI does not exhibit "superior" type I and type II error rates to standard null hypothesis testing. As expected, there is a tradeoff between type I and type II error. At precisely the small-to-moderate sample sizes that MBI's proponents deem "optimal," MBI reduces the type II error rate at the cost of greatly inflating the type I error rate—to two to six times that of standard hypothesis testing. **Conclusions**: Magnitude-based inference exhibits worrisome empirical behavior. In contrast to standard null hypothesis testing, which has predictable type I error rates, the type I error rates for MBI vary widely depending on the sample size and choice of smallest important effect, and are often unacceptably high. Magnitude-based inference should not be used. **Key Words:** TYPE I ERROR, TYPE II ERROR, HYPOTHESIS TESTING, CONFIDENCE INTERVALS, STATISTICS

I was recently asked to weigh in on a statistical debate that has been brewing in the sports science literature. Some researchers have advocated the use of a new statistical method they are calling "magnitude-based inference" (MBI) as an alternative to standard hypothesis testing (1–5). The method is being used in practice in the sports science literature (4,6), which makes it imperative to resolve this debate.

Several statisticians have criticized MBI due to its lack of a sound theoretical framework (7–9). In a 2015 article in *Medicine & Science in Sports & Exercise*, Welsh and Knight provided a statistical review of MBI in which they identified theoretical problems with the method, including that it creates unacceptably high false-positive rates (8). In response, MBI's proponents, Hopkins and Batterham (4), published a rebuttal in *Sports Medicine* 2016 in which they claim that MBI "outperforms" standard null hypothesis testing in terms of both type I (false-positive) and type II (false-negative) error rates for most cases.

At face value, this conclusion is dubious. There is a tradeoff between type I and type II error: when you improve one, you sacrifice the other. Thus, you do not need to be a statistician to immediately be skeptical of their paper. Indeed, their article is flawed in both its methods and conclusions.

First, Hopkins and Batterham (4) have obscured the systematic behavior of MBI in the way they presented their results. I have reproduced the exact numbers they report in their article, but have regraphed them in a more transparent and informative way. This single change reveals the fundamental problem with MBI. Second, Hopkins and Batterham have incorrectly defined type I and type II error. When I correct these mistakes, I show that the problem with MBI holds for all cases. Finally, I derive general mathematical equations for the type I and type II error rates for MBI; these equations confirm the findings from the simulations.

The problem boils down to this: MBI creates peaks of false positives at specific sample sizes; and MBI's creators provide sample size calculators (1) that specifically find these peaks. For example, for a particular statistical comparison, Hopkins and Batterham (4) conclude that 50 participants per group is the "optimal" sample size when using MBI. It turns out that 50 per group is precisely where the false-positive rate peaks for that case.

Address for correspondence: Kristin L. Sainani, Ph.D., Department of Health Research and Policy, 150 Governor's Lane, HRP Redwood Bldg, Stanford, CA 94305; E-mail: kcobb@stanford.edu.
Submitted for publication December 2017.
Accepted for publication April 2018.
Supplemental digital content is available for this article. Direct URL citations appear in the printed text and are provided in the HTML and PDF versions of this article on the journal's Web site (www.acsm-msse.org).

0195-9131/18/5010-2166/0
MEDICINE & SCIENCE IN SPORTS & EXERCISE®

DOI: 10.1249/MSS.0000000000001645

SPECIAL COMMUNICATIONS

1563 citations

A Flawed Statistical Method Was Just Banned From A Major Sports Science Journal

By Christie Aschwanden
Filed under Meta-Science
Published Jun. 27, 2018

SPORTSCIENCE · sportsci.org
Perspectives / Research Resources

A Battle Worth Fighting: a Comment on *The Vindication of Magnitude-Based Inference*

Martin Buchheit

Sportscience 22, sportsci.org/2018/CommentsOnMBI/mb.htm, 2018
Paris Saint-Germain, 78100 Saint-Germain-en-Laye, Paris, France. mb@martin-buchheit.net

Summary: MBI has changed the research and practice of thousands of sport scientists, but it is now under undeserved attack by defenders of p values. I trust the analytical foundations of MBI, and I consider it irreplaceable for research with samples and for monitoring individual athletes. I encourage all sport and exercise scientists to fight for MBI. Start by retweeting #supportMBI.

Magnitude-Based Decisions

Will G Hopkins, Institute for Health and Sport, Victoria University, Melbourne, Australia. Email.
Reviewer: Ross D Neville, School of Public Health, Physiotherapy and Sports Science, University College Dublin, Dublin, Ireland. Sportscience 23, i-iii, 2019 (sportsci.org/2019/inbrief.htm#decisions. Published June 2019.

Earlier this year I contacted the statistician Sander Greenland for clarification of a remark he had made in a discussion about Bayesian priors on the datamethods.org site. In the subsequent interactions, Greenland provided extensive advice on how to present MBI to a skeptical statistics community. He is opposed to the use of the term *inference*, unless it includes consideration not only of the sampling uncertainty in the magnitude (regardless of the frequentist, Bayesian or other interpretation of the uncertainty) but also of all the other potential biases arising from violation of assumptions about sampling and the analytic model. He agrees that it would be appropriate to rebrand MBI as a method for making magnitude-based *decisions* (MBD). He also prefers *compatibility* to *confidence* limits and intervals, in the sense that the interval defines a range of values interpreted as being compatible with the data and the statistical model (Greenland, 2019). *MBD* and *compatibility* have now been edited into the spreadsheets at *Sportscience*.

SPORTSCIENCE · sportsci.org Latest issue
Perspectives / Research Resources This issue

A Spreadsheet for Bayesian Posterior Compatibility Intervals and Magnitude-Based Decisions

Will G Hopkins

Sportscience 23, 5-7, 2019 (sportsci.org/2019/bayes.htm)
Institute for Health and Sport, Victoria University, Melbourne, Australia. Email. Reviewer: Ross D Neville, School of Public Health, Physiotherapy and Sports Science, University College Dublin, Dublin, Ireland.

The usual compatibility (confidence) interval for an effect in a sample can be modified to a Bayesian posterior compatibility (credibility) interval by combining the value of the effect and its interval with a prior belief in the effect expressed as its own value and interval. The spreadsheet accompanying this article provides such analyses for four kinds of effect: differences in means and other t-distributed estimates; percent or factor effects for such means derived from analyses of log-transformed dependent variables; ratios of risks, odds, hazards, and counts derived from generalized linear models; and Pearson correlation coefficients. Inclusion of a smallest important value for the effect allows the spreadsheet to provide a probabilistic magnitude-based decision about implementation of a clinically or practically relevant effect and about adequate precision for a non-clinical effect. The spreadsheet shows that realistic weakly informative priors applied to compatibility intervals from typically small samples produce posterior intervals that are practically the same as the original intervals. The minimally informative prior implicit in the magnitude-based decision method therefore provides acceptable Bayesian probabilistic estimates of the true magnitude of effects. Weakly informative priors should nevertheless be used to shrink unrealistically large compatibility limits arising from very small sample sizes and to reduce bias in effect magnitudes from generalized linear models with sparse data. Use of more-informative priors is problematic, owing to the difficulty of quantifying a belief and to bias in the belief. KEYWORDS: bias, clinical decisions, confidence, inference, probability, sample.

Reprint pdf · Reprint docx · Spreadsheet (the *Bayes* tab)

Positives

- Identifies real problems with hypothesis testing
- Provides an easy-to-implement solution
- Suggests a graded, non-binary interpretation of results
- More intuitive results than standard hypothesis testing

Criticisms

- Inappropriately claims lower rates of both Type I and Type II error
    - Considerably higher Type I errors (20%)
- Inappropriate Bayesian interpretations (mixture of Bayesian and frequentist interpretations)
- Other, similar methods which are theoretically stronger

My opinion

- Excellent motivation
- Gets people thinking critically about hypothesis testing
- Empathize

- Moot for small and large effects
- Promotes exploratory studies
- Hypothesis testing itself is not the problem. The real problems are:
    - Non-ideal use of hypothesis testing
    - Gross imbalance between exploratory and predictive studies

Overview

- History
- The p value
- Classical techniques
- Emerging techniques
- Controversies
- The future

Questions

Questions

“To call in the statistician after the experiment is done may be no more than asking him to perform a postmortem examination … he may be able to say what the experiment died of.”

–R.A. Fisher, 1938

Quick search

Go

Introduction

power1d is a software package for numerically estimating statistical power in experiments involving one-dimensional continua.

Dependencies:

- Python 3
- NumPy
- SciPy
- Matplotlib

Please consider citing:

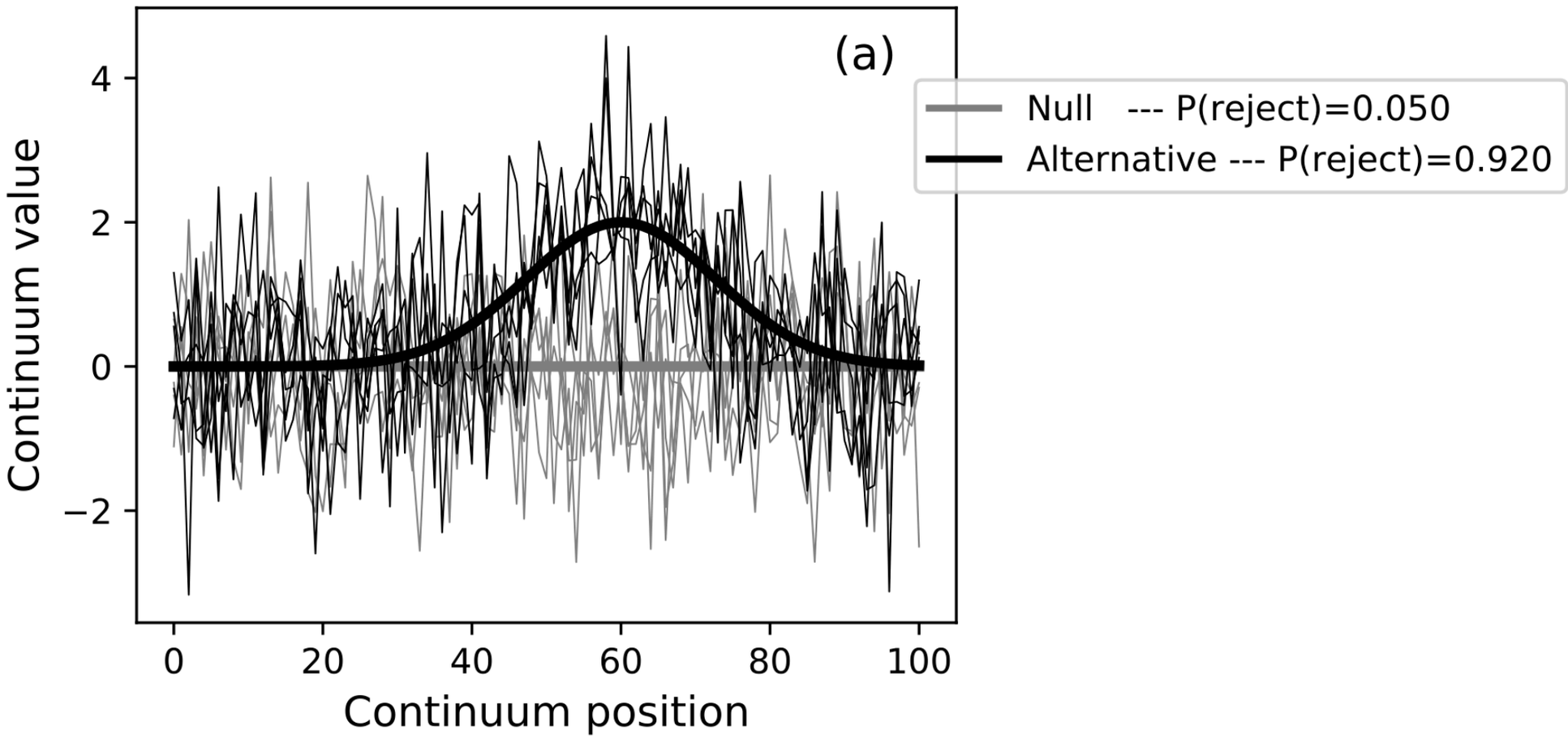
Pataky TC (in review).
Power1D: A Python toolbox for numerical power estimates in experiments involving one-dimensional continua.
PeerJ Computer Science.

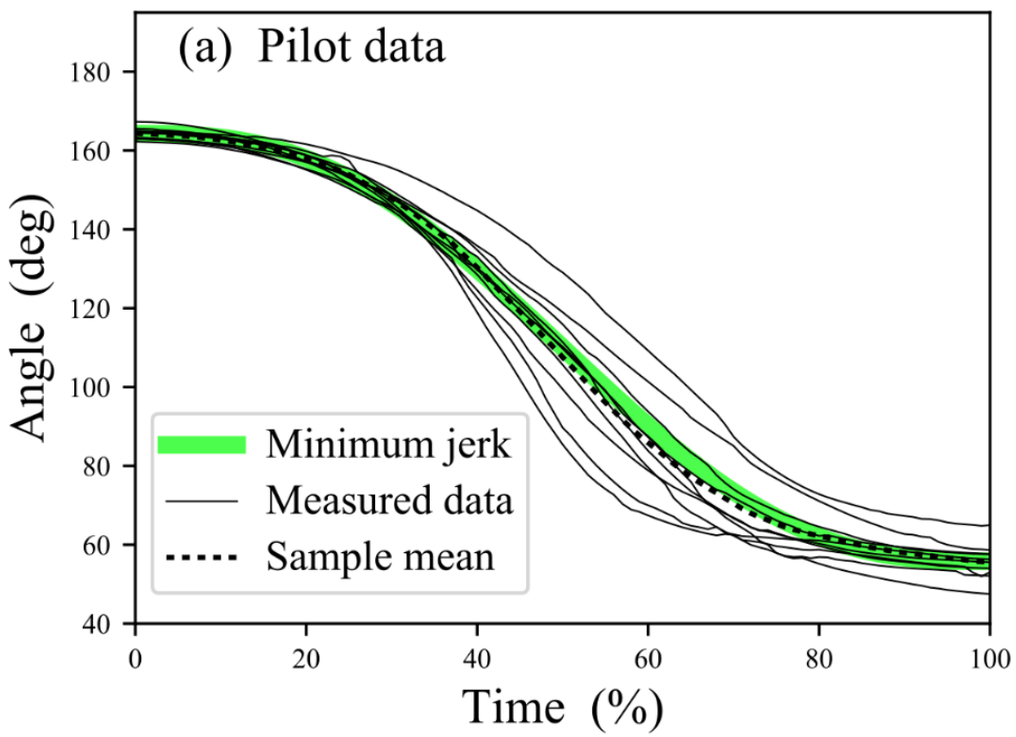
Download

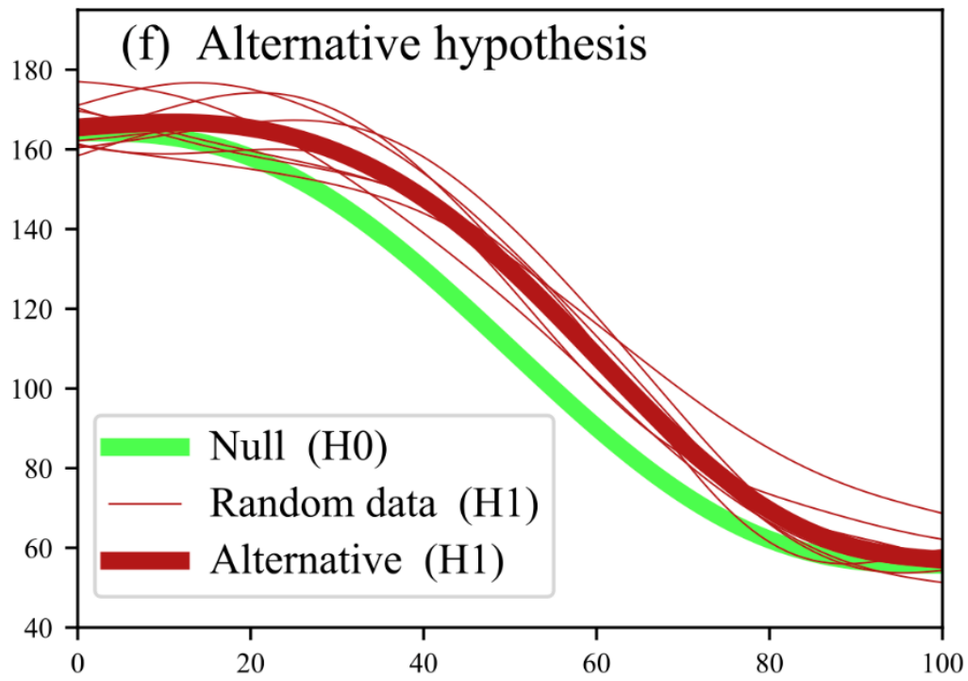
Python source code

Contents

spm1d.org/power1d


(a)
Null --- P(reject)=0.050
Alternative --- P(reject)=0.920
Continuum value
Continuum position
4
2
0
−2
0
20
40
60
80
100


(a) Pilot data
Angle (deg)
Time (%)
Minimum jerk
Measured data
Sample mean
180
160
140
120
100
80
60
40
0
20
40
60
80
100


(f) Alternative hypothesis
Null (H0)
Random data (H1)
Alternative (H1)
180
160
140
120
100
80
60
40
0
20
40
60
80
100

Bayesian

p(sunny | 30 deg)

p(30 deg | sunny)

Imagine an ESP experiment involving coin flips.
A subject guesses correctly on 60 flips.
Do they have ESP?

Classical hypothesis testing:

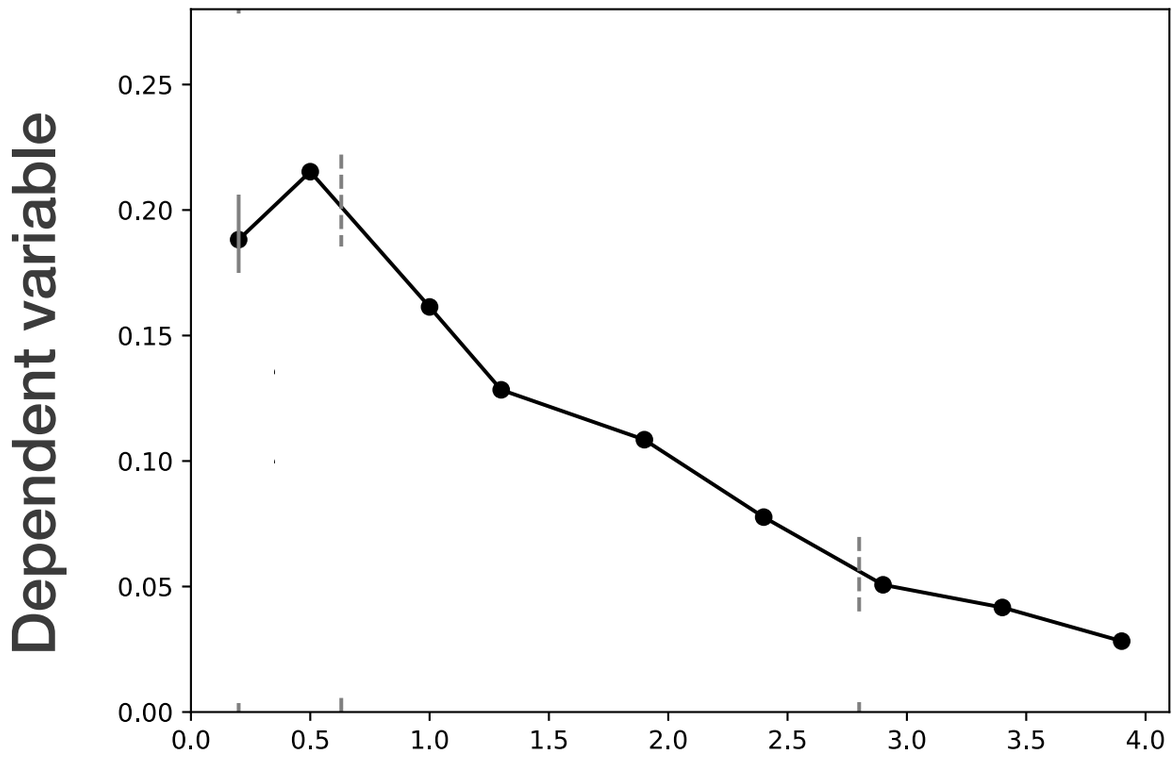
$$p(\text{60\% correct} \mid \text{ESP}) = 0.0176$$

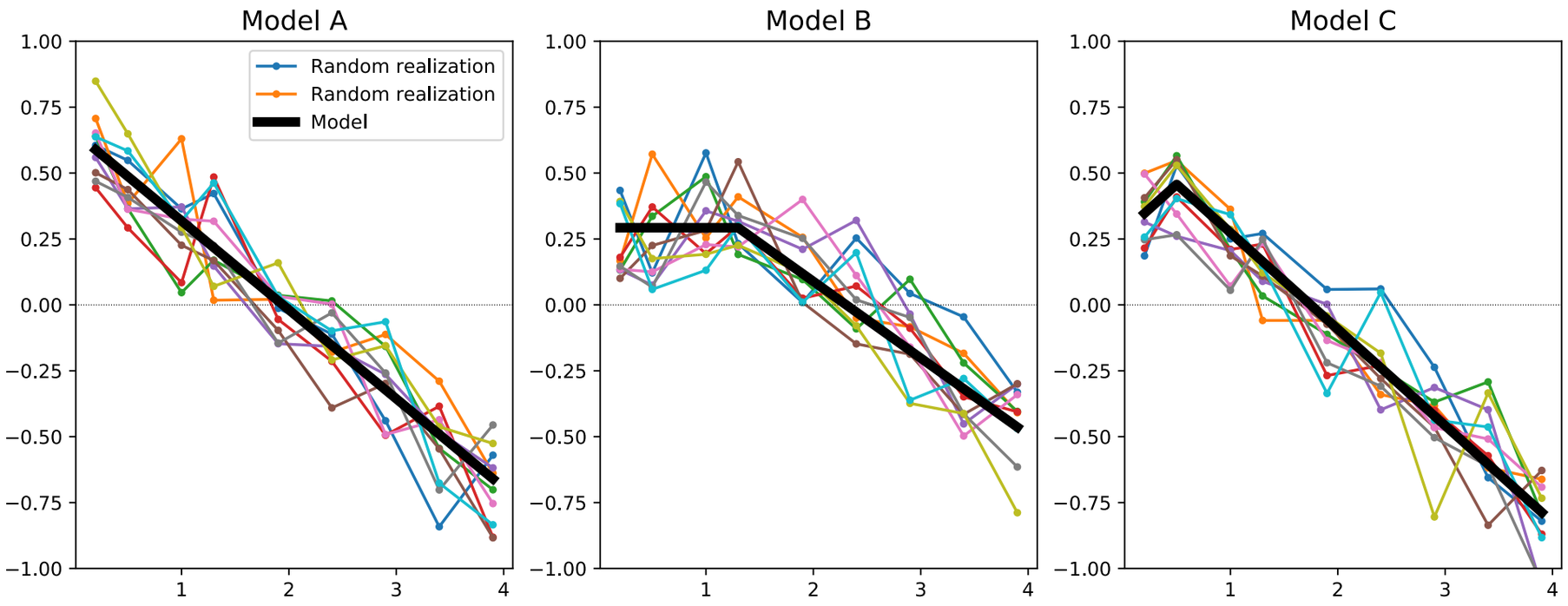
Relatively strong evidence that this person has ESP

Bayesian:

$$P(\text{ESP} \mid \text{60\% correct}) \propto P(\text{60\% correct} \mid \text{ESP})\ P(\text{ESP})$$

A Bayesian perspective on linear regression





Summary

- Biomechanical data are complex
- Emerging techniques are starting to comprehensively handle this complexity

Summary

- The biggest problem in Biomechanics: exploration
    - Imbalance between predictive and exploratory studies

Summary

Hypotheses should:

- identify specific variable(s)
- make non-null predictions (usually)

Otherwise hypotheses are scientifically poor

and statistical results are scientifically poor

Summary

- Statistics can’t give us biomechanical meaning
- We must give biomechanical meaning to our statistical methods

Thank you

thunss.com