IEEE NSS 2016

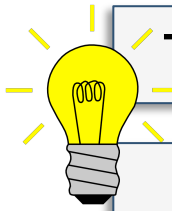
Strasbourg, France
3 November 2016

Application of Econometric Data Analysis Methods to Physics Software

Maria Grazia Pia, *INFN Genova, Italy*
Elisabetta Ronchieri, *INFN CNAF, Bologna, Italy*

Foreword

Due to limited time allocation, there is room only to highlight some basic concepts and to illustrate them in a few examples of application



Treat a software system as a **sociosystem/ecosystem**

Apply data analysis
concepts, **methods** and **techniques**
developed in **economy/ecology**

multiple perspectives

Software development environment

Observables produced by the software

Quantitative analysis:
- **Inference**
- **Measures**

Evolution

Trend analysis

Gain insight into the past

Use knowledge to shape the future

Aggregation

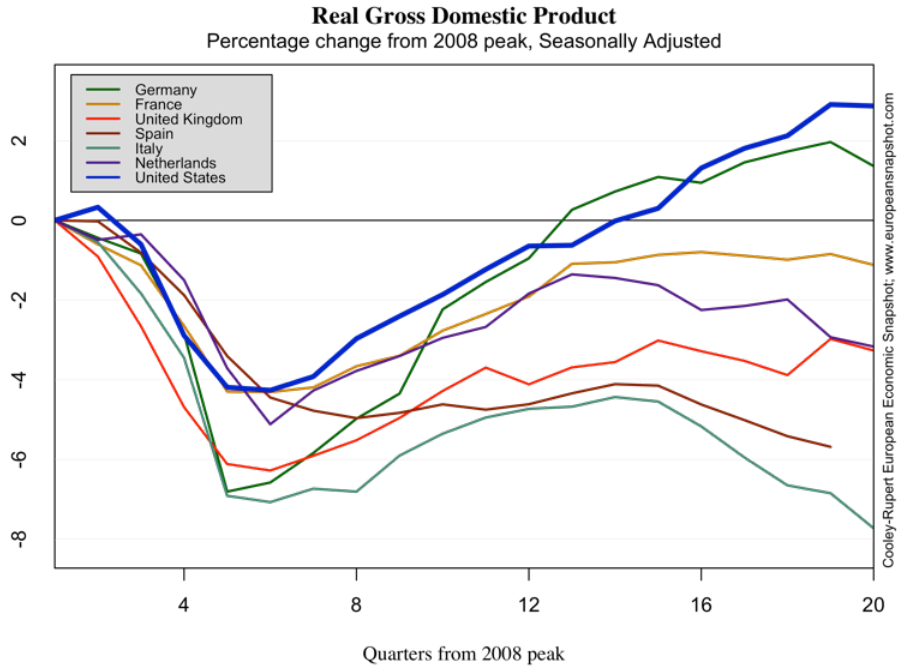
Inequality analysis

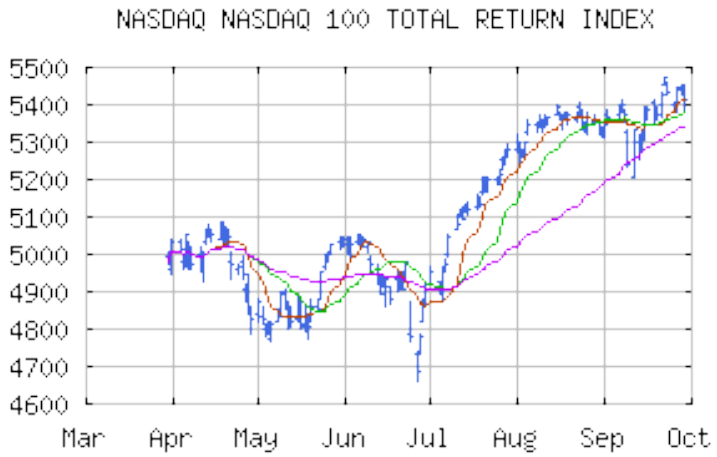
Large number of heterogeneous data

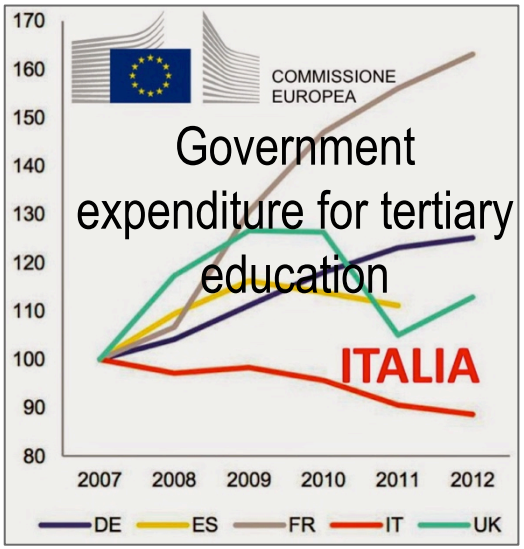
Distribution of their properties

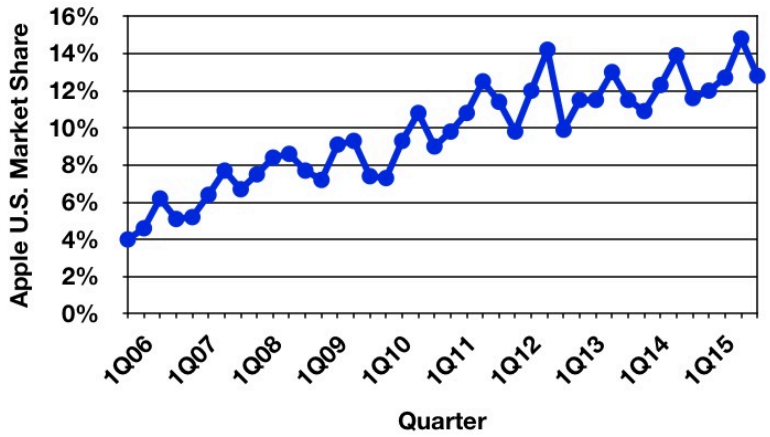
Highlights, no time for details and for other analysis methods

Trend


Real Gross Domestic Product
Percentage change from 2008 peak, Seasonally Adjusted
Germany
France
United Kingdom
Spain
Italy
Netherlands
United States
2
0
-2
-4
-6
-8
4
8
12
16
20
Quarters from 2008 peak
Cooley-Rupert European Economic Snapshot; www.europeansnapshot.com
Source: Eurostat


NASDAQ NASDAQ 100 TOTAL RETURN INDEX
5500
5400
5300
5200
5100
5000
4900
4800
4700
4600
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct


COMMISSIONE
EUROPEA
Government expenditure for tertiary education
ITALIA
170
160
150
140
130
120
110
100
90
80
2007
2008
2009
2010
2011
2012
DE
ES
FR
IT
UK


Apple U.S. Market Share
16%
14%
12%
10%
8%
6%
4%
2%
0%
1Q06
1Q07
1Q08
1Q09
1Q10
1Q11
1Q12
1Q13
1Q14
1Q15
Quarter

Trend analysis

- Statistical techniques to identify **patterns** in a **series of data**
  - Ability to deal with noise
- Used to forecast the future
  (although it does not predict the future)
  - But also to analyze past events
- Tests for **statistical inference**
  - Parametric and non parametric
  - Test for randomness: H_0 = random, H_1 = monotonic trend/upward/downward
  - **Mann-Kendall** test, **Cox-Stuart** test, **Bartels** test etc.
- Related: **change point detection**

Lehman laws

M. M. Lehman,
Programs, Life Cycles, and Laws of Software Evolution,
Proc. IEEE, vol. 68, no. 9, pp. 1060-1076, 1980

1. Continuing Change

A program that is used and that as an implementation of its specification reflects some other reality, **undergoes continual change** or **becomes progressively less useful**. The change or decay process continues until it is judged more cost effective to replace the system with a recreated version.

2. Increasing Complexity

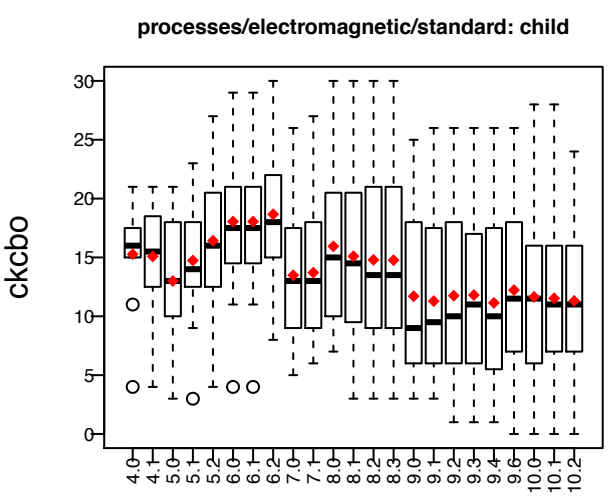
As an evolving program is continually changed, **its complexity**, *reflecting deteriorating structure*, **increases** unless work is done to maintain or reduce it.

Coupling between classes

Chidamber and Kemerer **CBO**

High CBO is undesirable

Excessive coupling between object classes
is detrimental to modular design and prevents reuse
A high coupling has been found to indicate fault-proneness



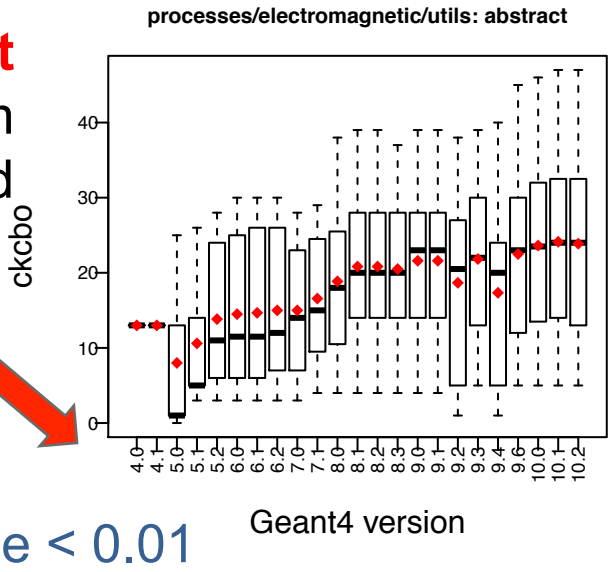
Leaf
H_0: random
H_1: downward

p-value < 0.01

Abstract
H_0: random
H_1: upward

Mann-Kendall test

p-value < 0.01



How high is too high? CBO>14

H. Sahraoui et al., "Can Metrics Help to Bridge the Gap Between the Improvement of OO Design Quality and Its Automation?" *Proc. Int. Conf. Software Maintenance*, pp. 154-262, 2000

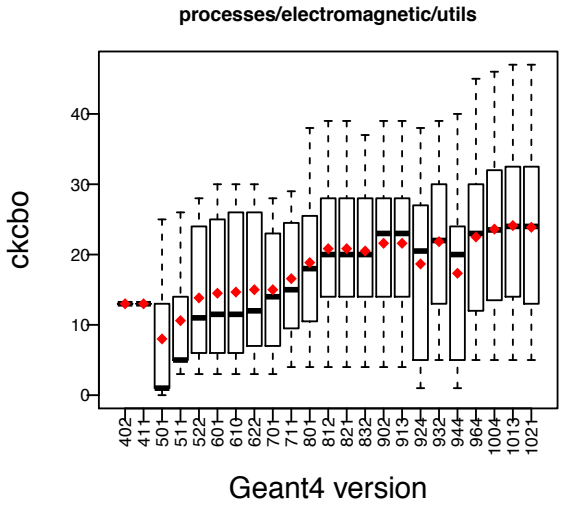
Do I really need a statistical test to see a trend?

I can see a trend just by looking at the plot!

What about seeing trends in **26581** plots?

How to objectively quantify what different eyes see?
How to aggregate the trends observed in various plots?

Chidamber and Kemerer OO metrics

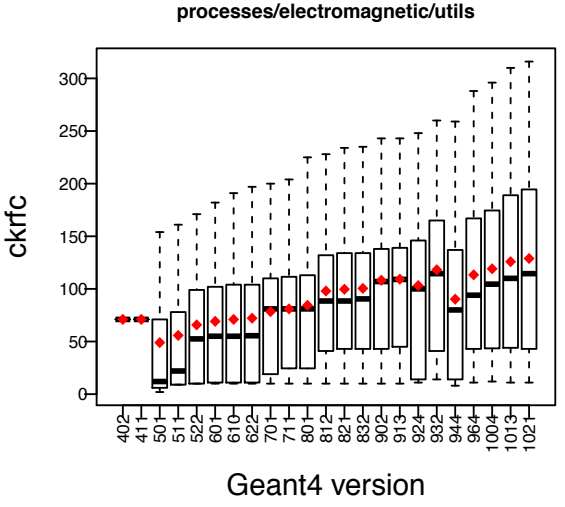


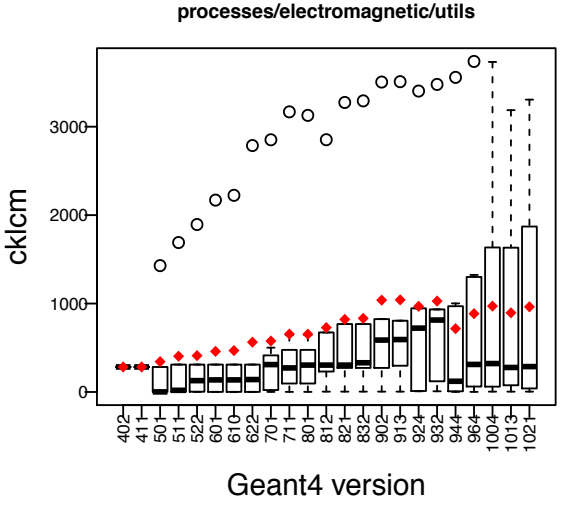
Abstract classes

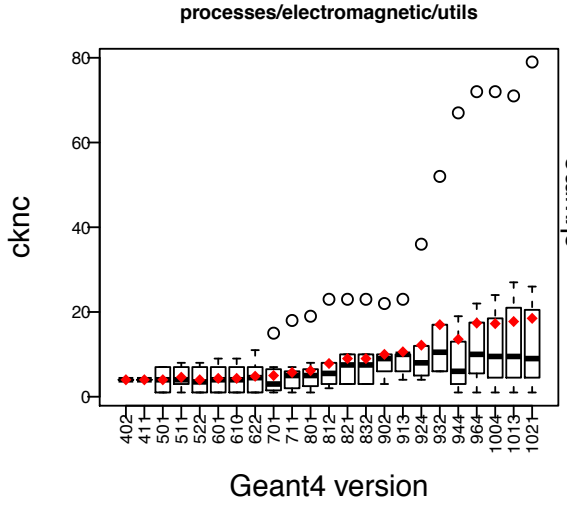
H_0: random

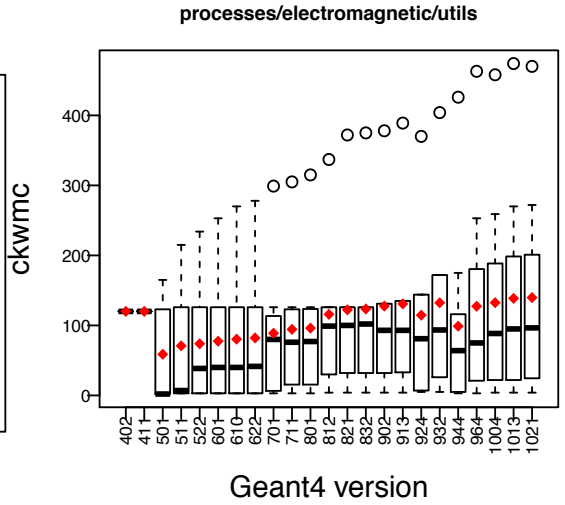
H_1: upward

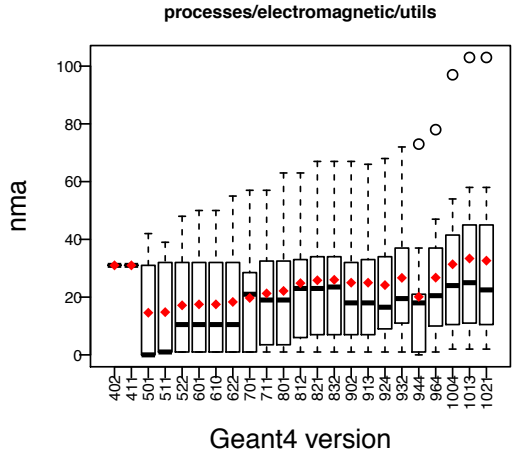
p-value < 0.01

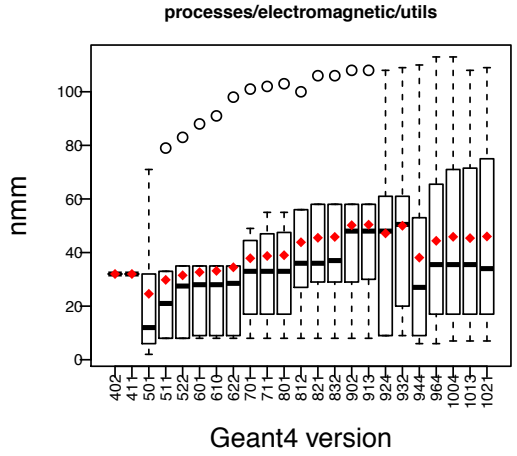


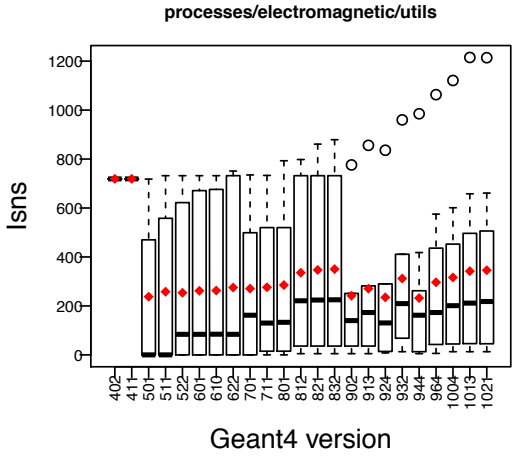


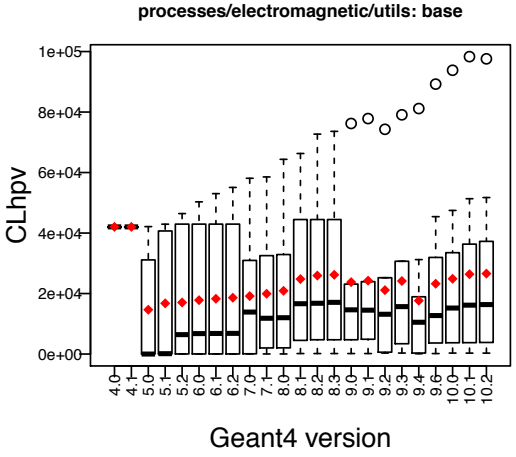


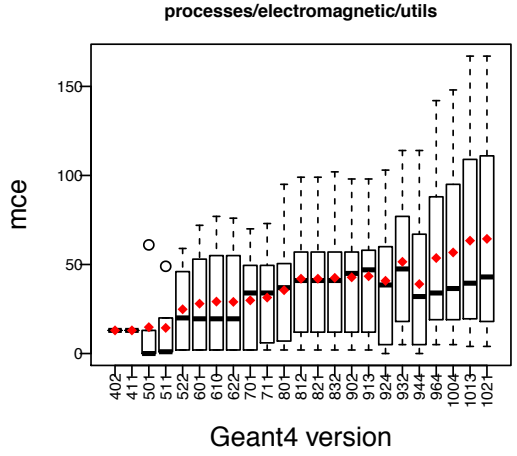


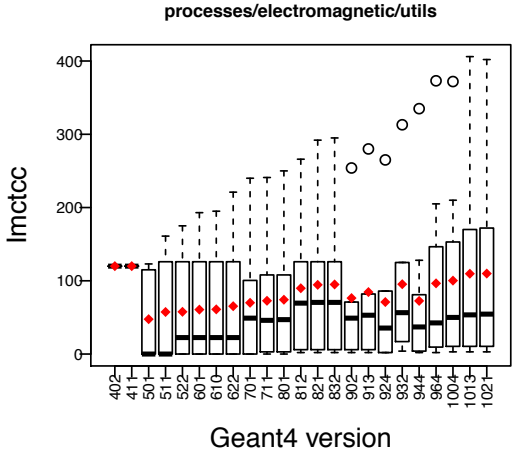


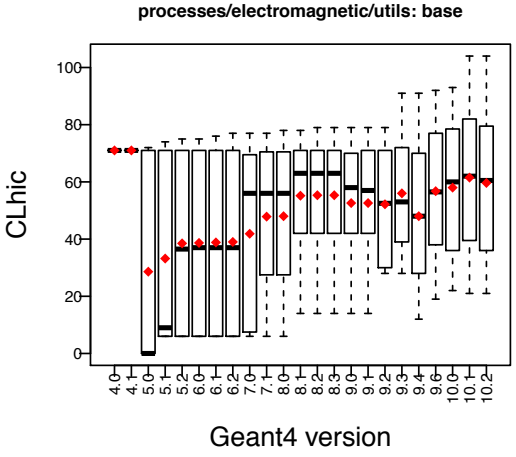


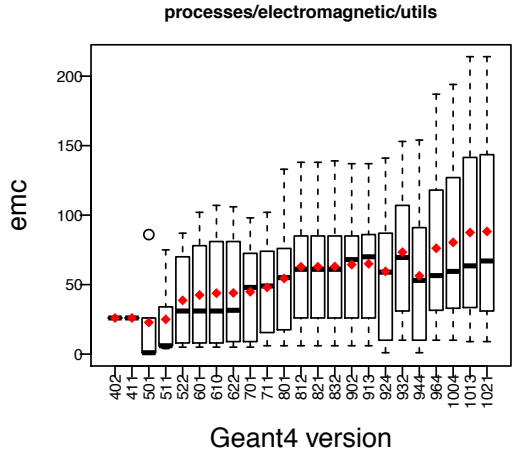












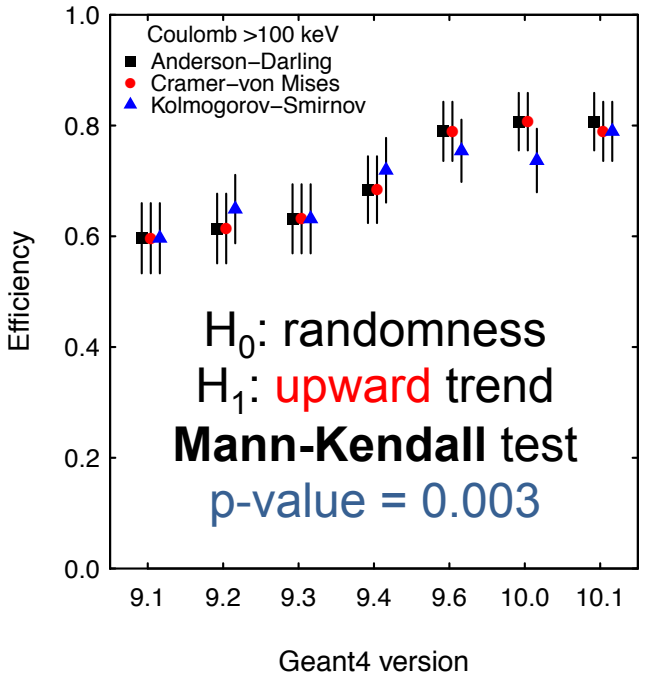
H_0: random

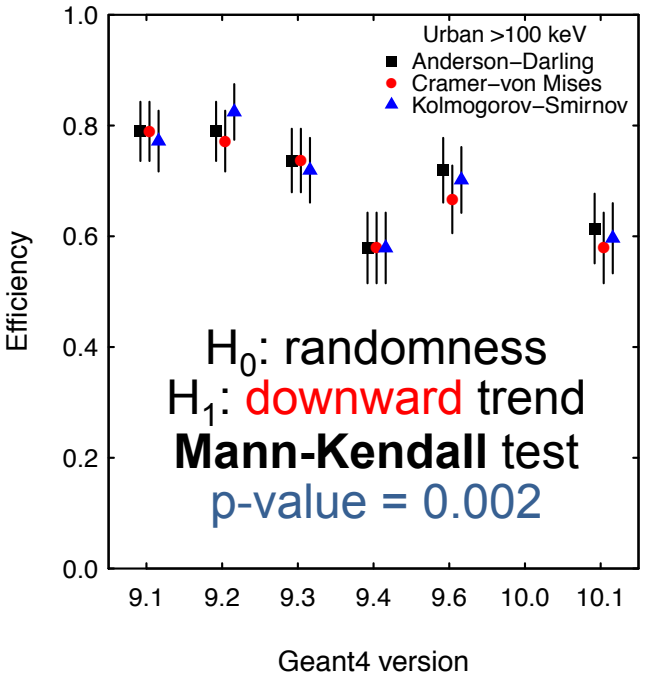
H_1: upward

p-value < 0.01

Trends in software functionality

Electron backscattering simulation with Geant4



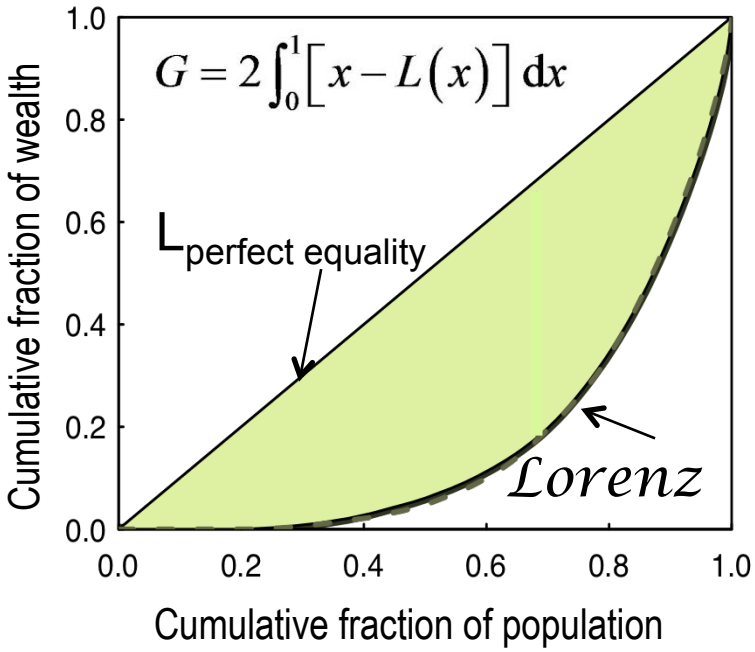


Trend of compatibility with experiment as a function of Geant4 version for different physics configurations

Helpful guidance in algorithm development, optimization, regression testing, software maintenance…

Income inequality measures

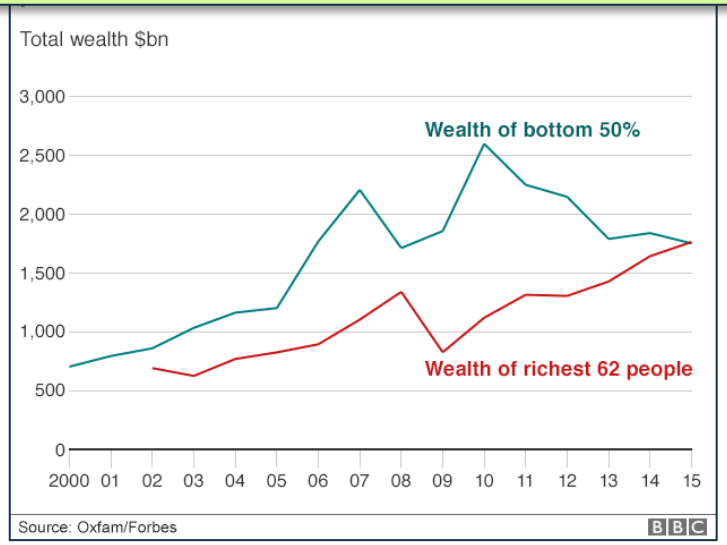
Gini index

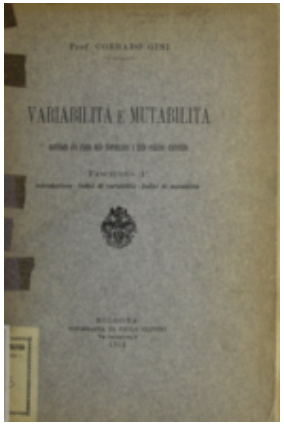


$0 \le \mathbf{P} \le 1$

0 **more unequal society** 1

The 62 richest people in the world are worth more than the poorest 50%



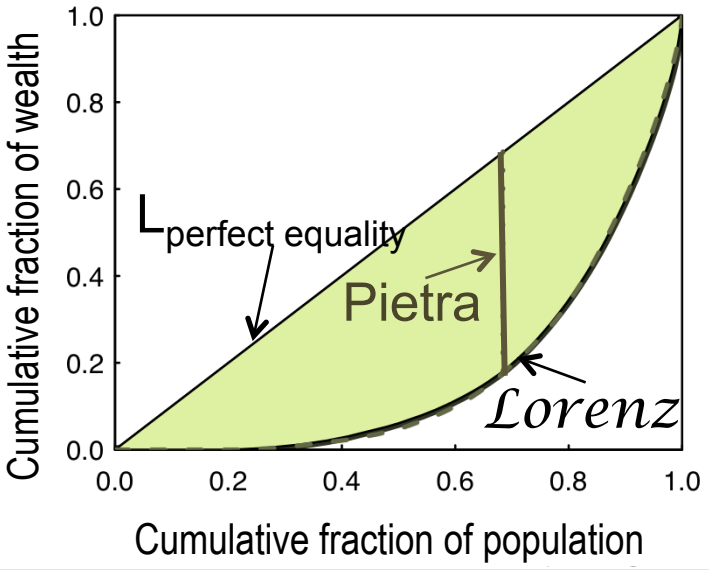


C. Gini, Variabilità e mutabilità : contributo allo studio delle distribuzioni e delle relazioni statistiche, 1912

Pietra index

AKA Ricci-Schutz index, Hoover index

$$\mathbf{P} = \max(L_{pe}(x) - \mathcal{L}(x))$$



- Used in derivative markets as a benchmark measure of **statistical heterogeneity**
- Counterpart of Kolmogorov-Smirnov statistic
- It can be interpreted as the proportion of income that has to be transferred from those above the mean to those below the mean in order to achieve an equal distribution
  - Emphasis on individual-mean interaction

Other inequality measures

Theil index

0 → ∞

More equal society

$$T = \sum_{i=1}^{n} s_i \left[\log s_i - \log(\frac{1}{n}) \right]$$

s_i = share of the i^{th} group in total income
n = total number of income groups

The same as **redundancy** in information theory:
the maximum possible entropy of the data minus the observed entropy

Atkinson index

$$I = 1 - \pi_e / \mu$$

e = sensitivity parameter

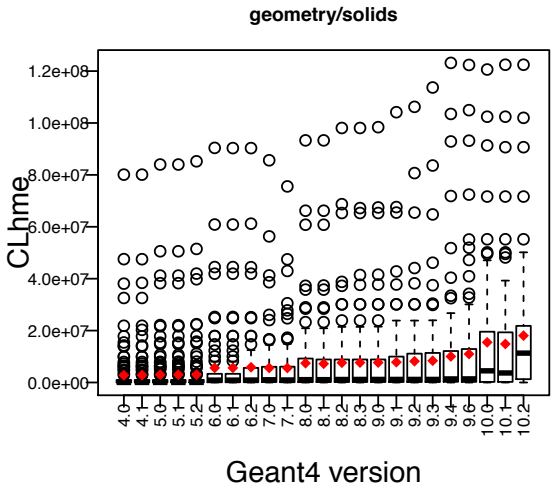
$0 \leq \mathbf{I} \leq 1$

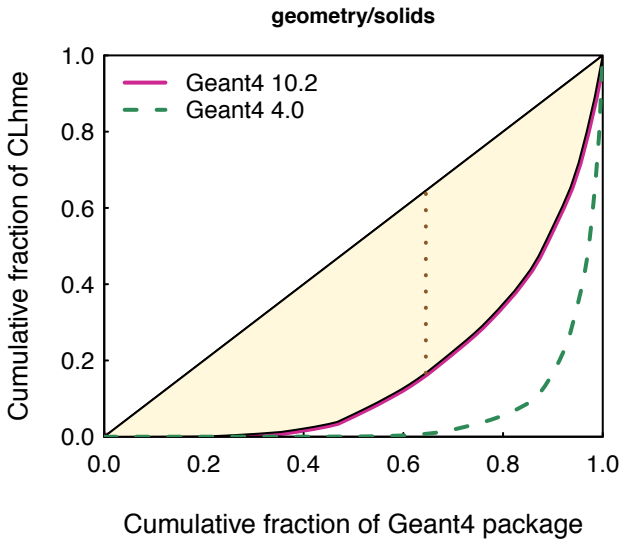
Used to calculate the proportion of total income that would be required to achieve an equal level of social welfare as at present, if incomes were perfectly distributed

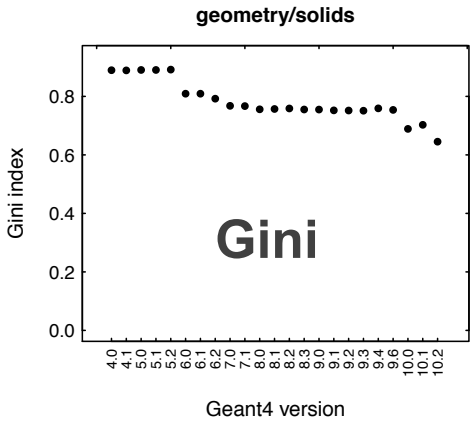
Theil I, Theil II, Kolm index, coefficient of variation, generalized entropy and more…

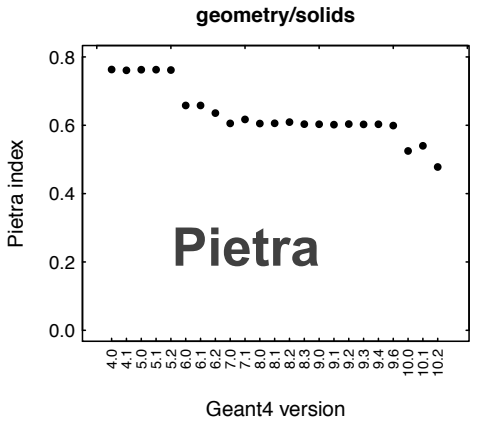
Halstead mental effort

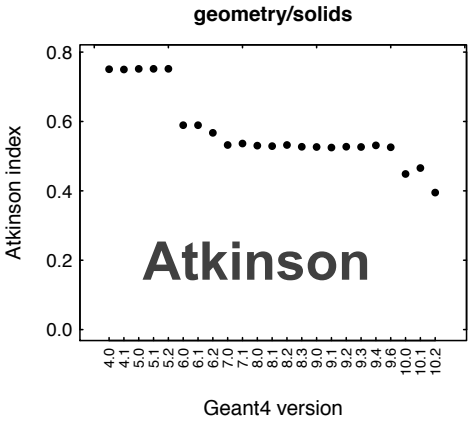
Measure of the number of elemental mental discriminations necessary to create or understand a class

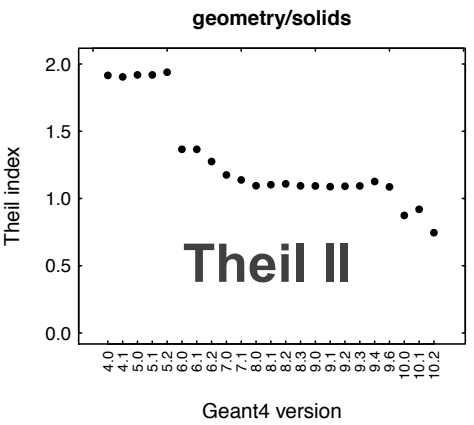


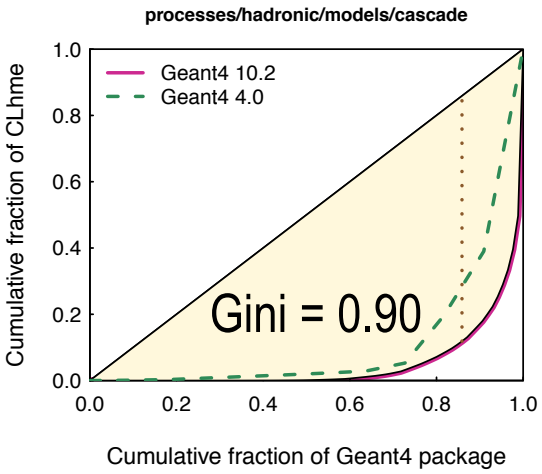


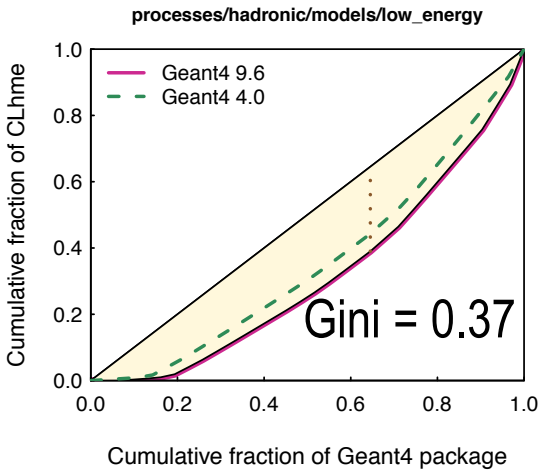


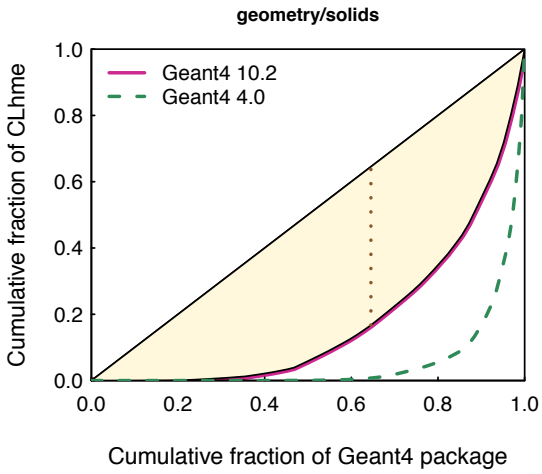








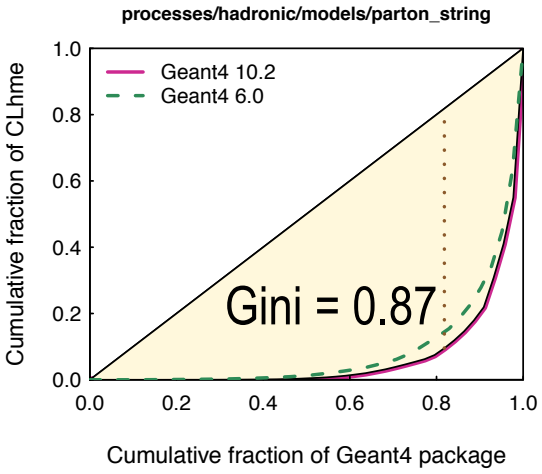


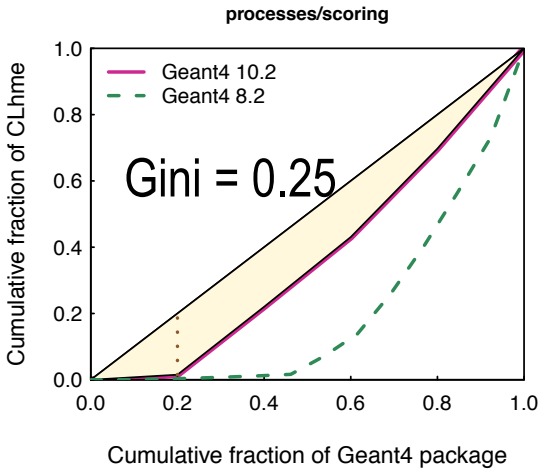


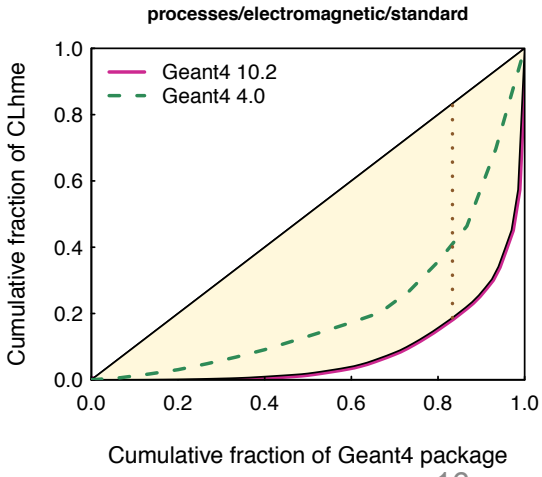
concentrated software complexity

distributed software complexity

evolution of concentration







Gini and galaxies

THE ASTROPHYSICAL JOURNAL, 588:218–229, 2003 May 1

A NEW APPROACH TO GALAXY MORPHOLOGY. I. ANALYSIS OF THE SLOAN DIGITAL SKY SURVEY EARLY DATA RELEASE

ROBERTO G. ABRAHAM,[1] SIDNEY VAN DEN BERGH,[2] AND PREETHI NAIR[1]
Received 2002 July 12; accepted 2002 December 26

THE ASTRONOMICAL JOURNAL, 128:163–182, 2004 July

A NEW NONPARAMETRIC APPROACH TO GALAXY MORPHOLOGICAL CLASSIFICATION

JENNIFER M. LOTZ,[1] JOEL PRIMACK,[1] AND PIERO MADAU[2]
Received 2003 November 14; accepted 2004 April 12

THE ASTROPHYSICAL JOURNAL LETTERS, 816:L23 (4pp), 2016 January 10 doi:10.3847/20

THE GINI COEFFICIENT AS A TOOL FOR IMAGE FAMILY IDENITIFICATION IN STRONG LENSING SYSTEMS WITH MULTIPLE IMAGES

MICHAEL K. FLORIAN[1,2], MICHAEL D. GLADDERS[1,2], NAN LI[1,2,3], AND KEREN SHARON[4]
[1] Department of Astronomy and Astrophysics, The University of Chicago, Chicago, IL 60637, USA
[2] Kavli Institute for Cosmological Physics, The University of Chicago, Chicago, IL 60637, USA
[3] Argonne National Laboratory, 9700 South Cass Avenue B109, Lemont, IL 60439, USA
[4] Department of Astronomy, University of Michigan, 1085 S. University Avenue, Ann Arbor, MI 48109, USA
Received 2015 November 11; accepted 2015 December 1; published 2016 January 12

Aggregate the capabilities of Geant4 PhysicsLists to reproduce experimental observables

Other econometric analysis methods:
Concentration, **Change point**

Relation with methods used in ecology
(e.g. **analysis of diversity**)

Information theory background

Decomposition of inequality measures by subgroups

Comparative evaluation
of measures and tests

Methods, applications to physics software and results will be documented in forthcoming papers

Conclusion

Aggregation of information
Quantitative measurements

- Statistical methods commonly used in other disciplines can be valuable in software and physics analysis
- Rich variety of econometric concepts and techniques
  - Trend, inequality, concentration, diversity, changepoint…
- Similar/complementary concepts in quantitative ecology

- Ongoing R&D to explore applications in physics software
  - To characterize software properties
  - To evaluate the behaviour of physics models
- A few highlights, no time for extensive details