CHEP 2016

10-14 October 2016, San Francisco, CA, USA

Application of econometric and ecology analysis methods in physics software

Maria Grazia Pia, *INFN Genova, Italy*

M. C. Han, G. Hoff, C. H. Kim, S. H. Kim, E. Ronchieri, P. Saracco

Hanyang University, Seoul, Korea - INFN CNAF, Bologna, Italy - CAPES, Brasilia, Brazil

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

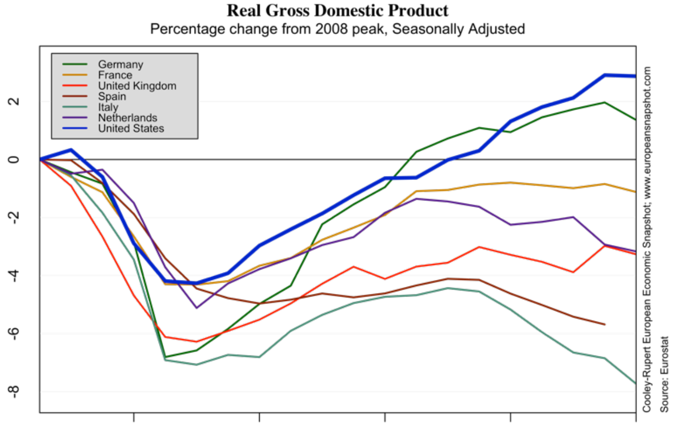
Software development environment

Observables produced by the software

Quantitative analysis:

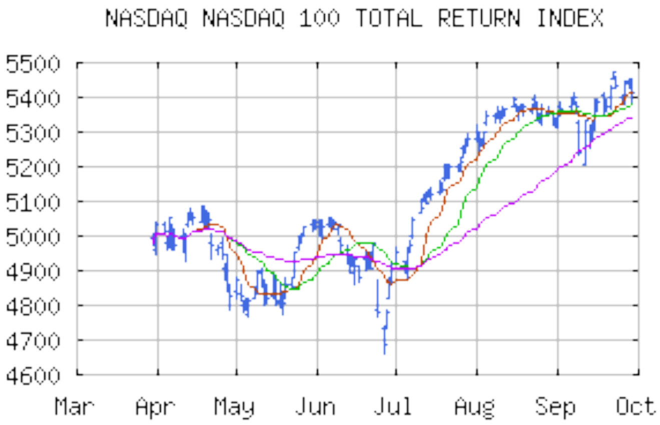
- **Inference**
- **Measures**

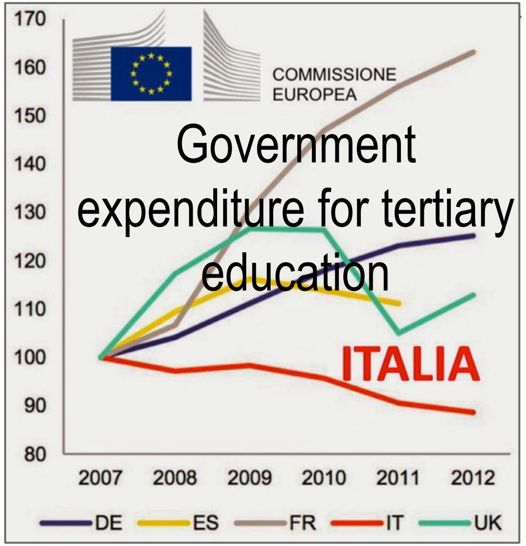
Trend



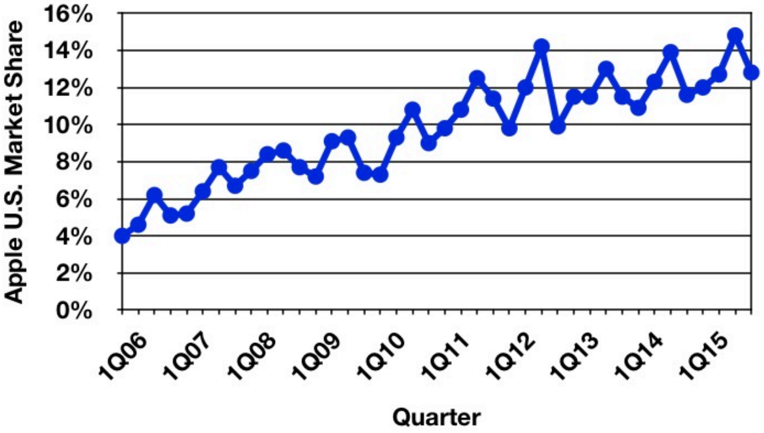
Cooley-Rupert European Economic Snapshot; www.europeansnapshot.com

Source: Eurostat





Government expenditure for tertiary education



Trend analysis

- Statistical techniques to identify **patterns** in a **series of data**
  - Ability to deal with noise
- Used to forecast the future *(although it do<u>es not</u> 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

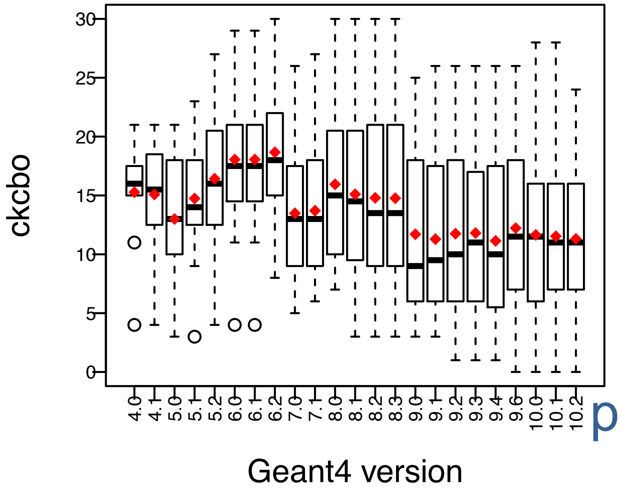
Chidamber and Kemerer **CBO**

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

processes/electromagnetic/standard: child



Leaf

H_0: random

H_1: downward

Abstract

H_0: random

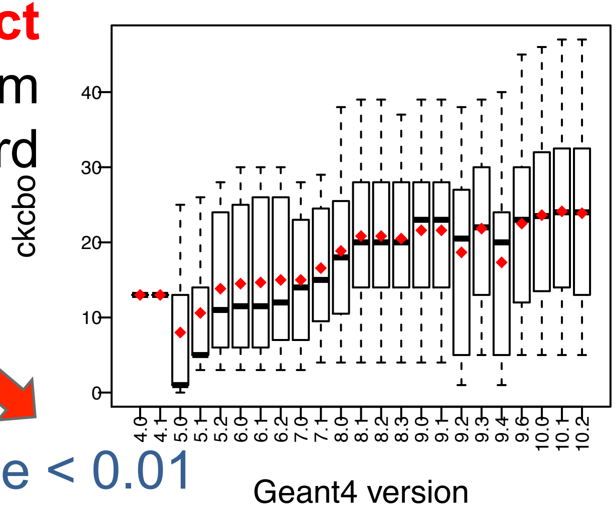
H_1: upward

Mann-Kendall test

p-value < 0.01

p-value < 0.01

processes/electromagnetic/utils: abstract



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?

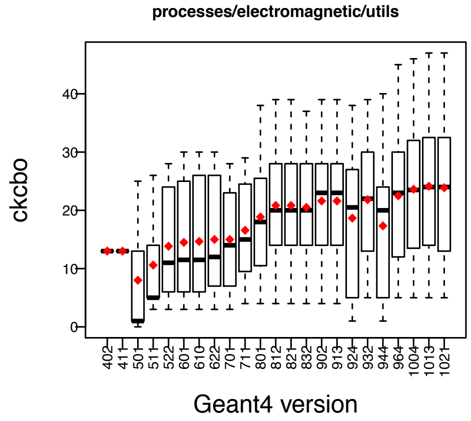
Chidamber and Kemerer OO metrics

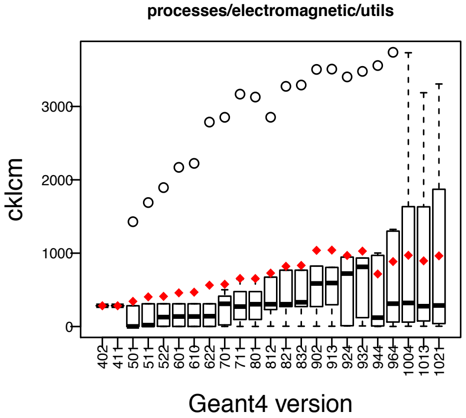
Abstract classes

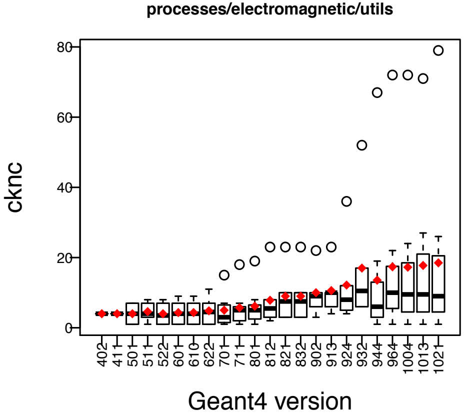
H_0: random

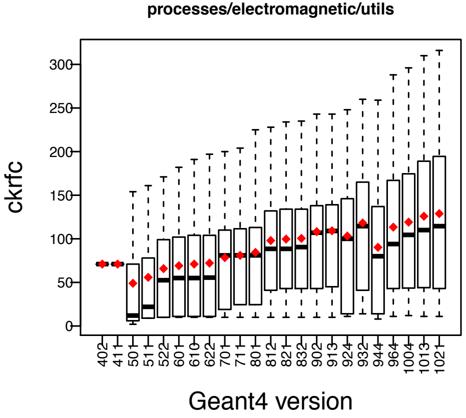
H_1: upward

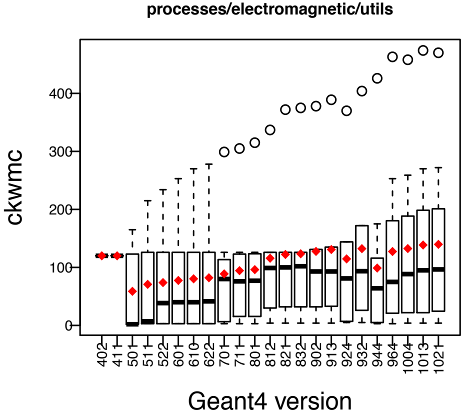
p-value < 0.01

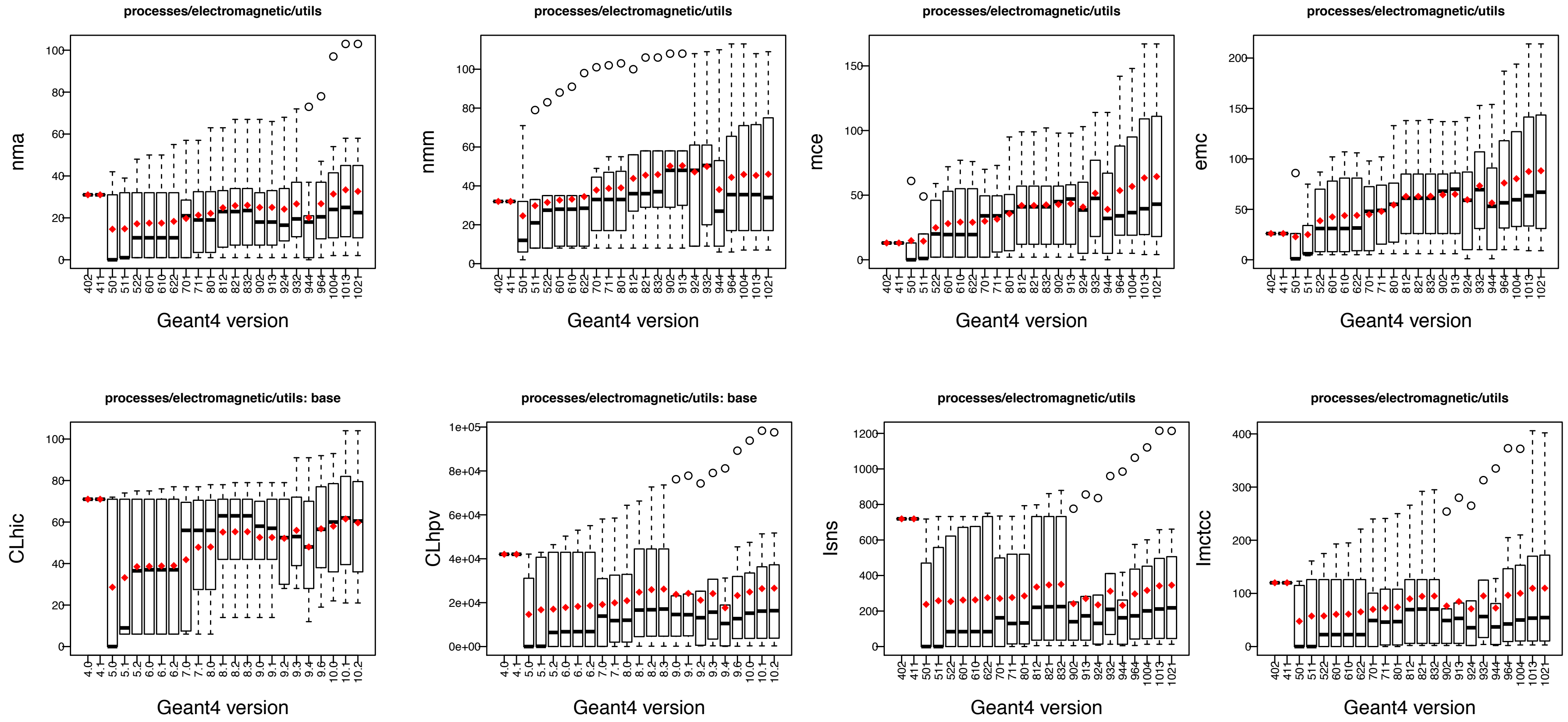

processes/electromagnetic/utils


processes/electromagnetic/utils


processes/electromagnetic/utils


processes/electromagnetic/utils

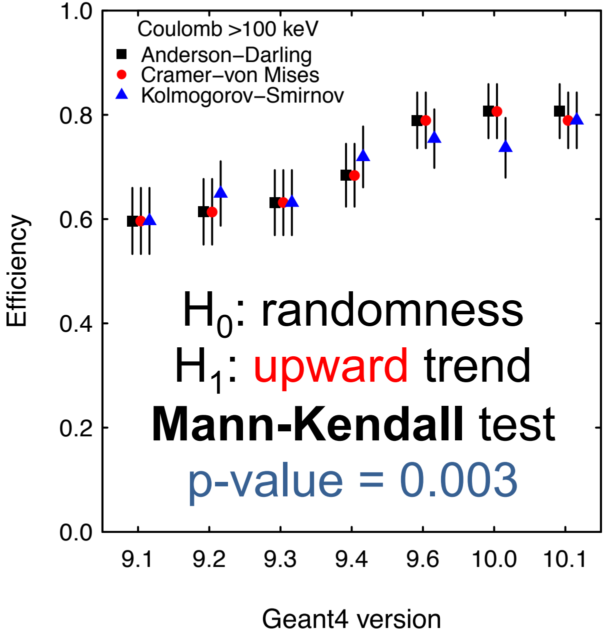

processes/electromagnetic/utils



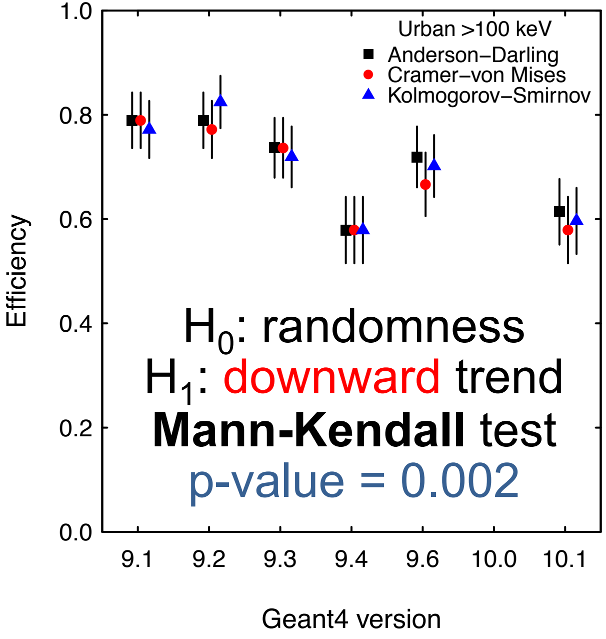
H_0: random – H_1: upward → **p-value < 0.01**

Trends in software functionality

Electron backscattering simulation with Geant4



H_0: randomness
H_1: upward trend
Mann-Kendall test
p-value = 0.003



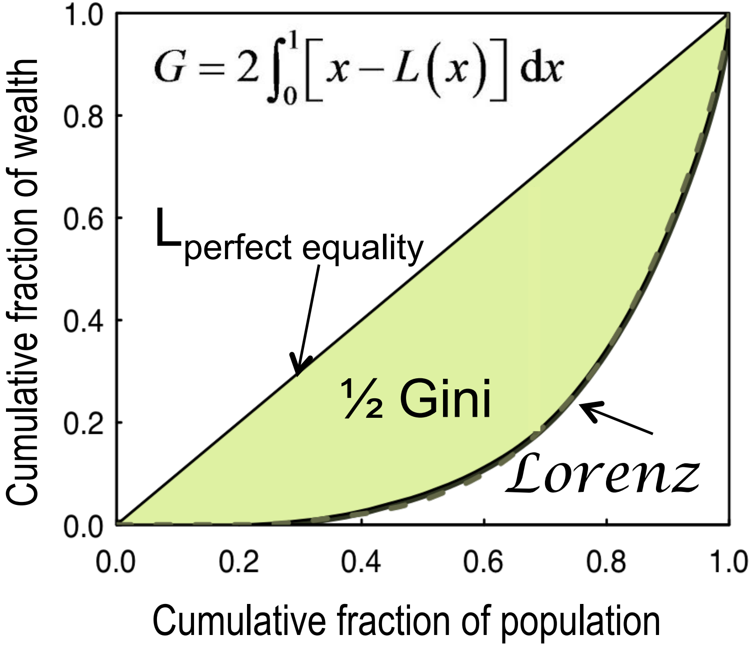
H_0: randomness
H_1: downward trend
Mann-Kendall test
p-value = 0.002

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

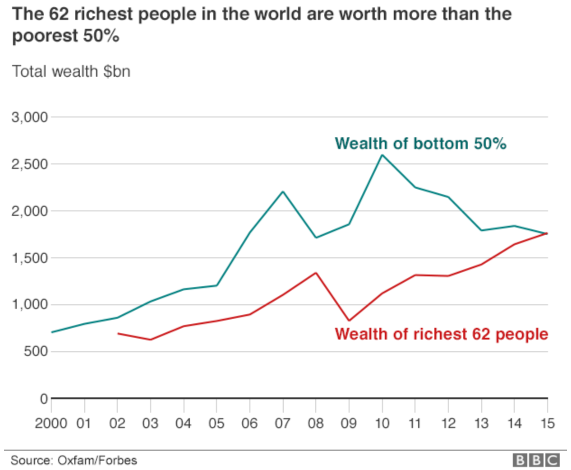
Helpful guidance in algorithm development, optimization, regression testing, software maintenance…

Income inequality measures

Gini index



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



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

Total wealth $bn

Source: Oxfam/Forbes

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

0 **more unequal society** 1

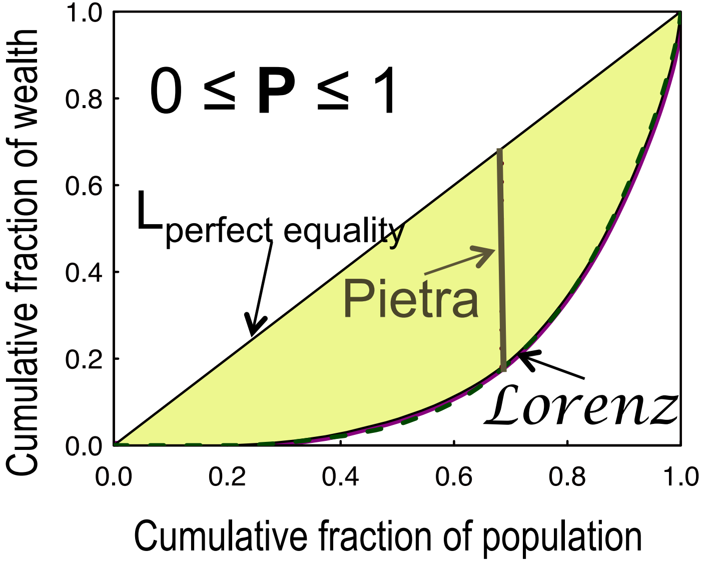
C. Gini, Variabilità e mutabilità : contributo allo studio delle distribuzioni e delle relazioni statistiche, 1912

Pietra index

AKA Ricci-Schutz index, Hoover index

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

- 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



Maria Grazia Pia, *INFN Genova*

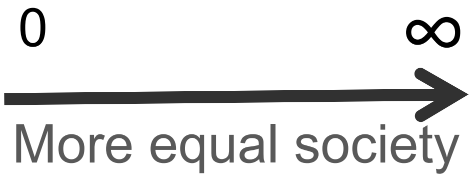
Other inequality measures

Theil index

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

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

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

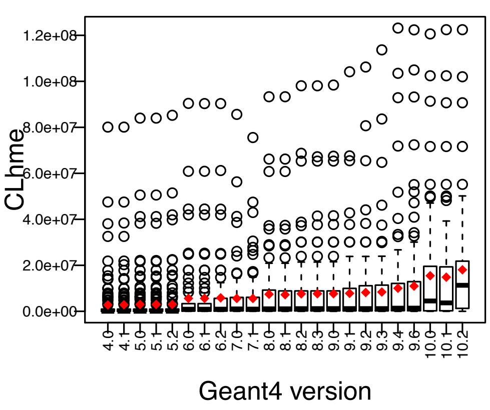
e = sensitivity parameter

0 ≤ I ≤ 1

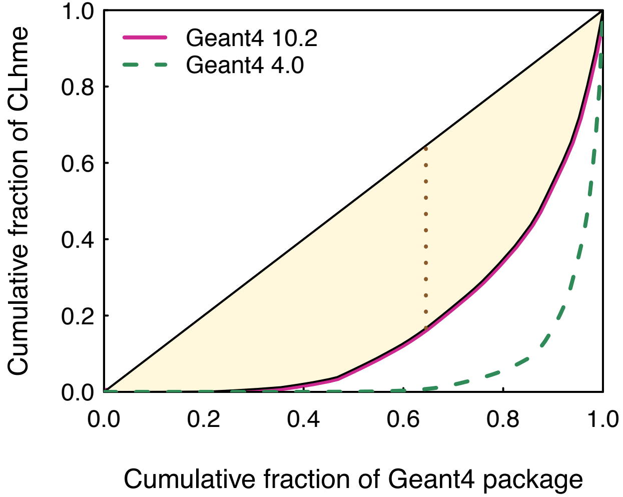
Used to calculate the proportion of total income that would be required to achieve an equal level of social welfare as at present, if incomes were perfectly distributed

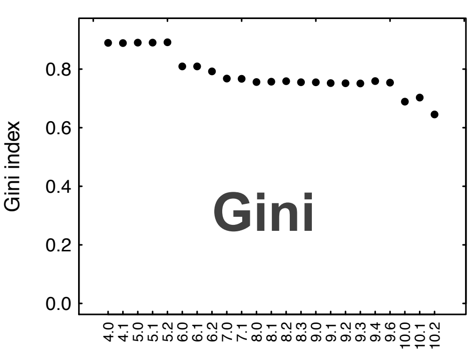
Theil I, Theil II, Kolm index, coefficient of variation, generalized entropy and more…

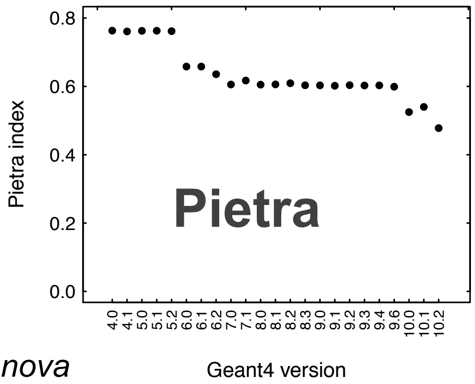
Halstead mental effort

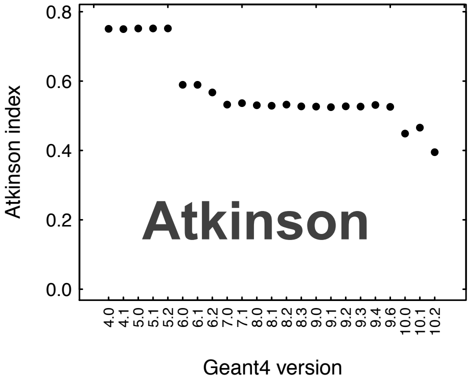

geometry/solids

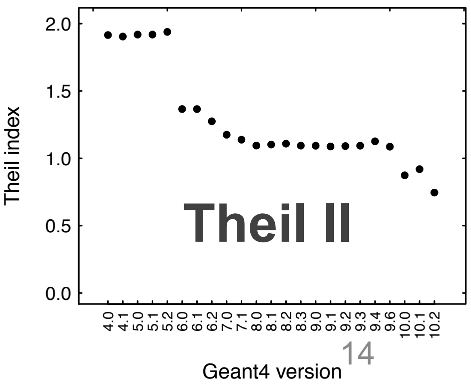
Measure of the number of elemental mental discriminations necessary to create or understand a class

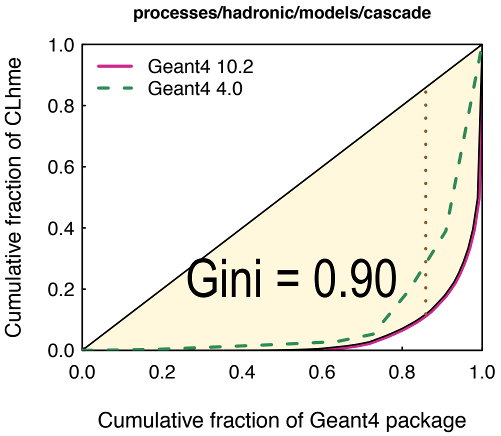

geometry/solids

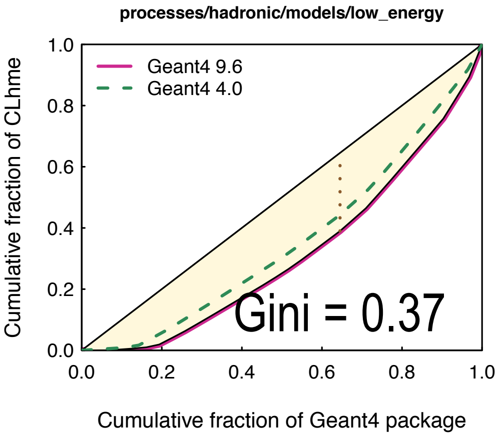

geometry/solids

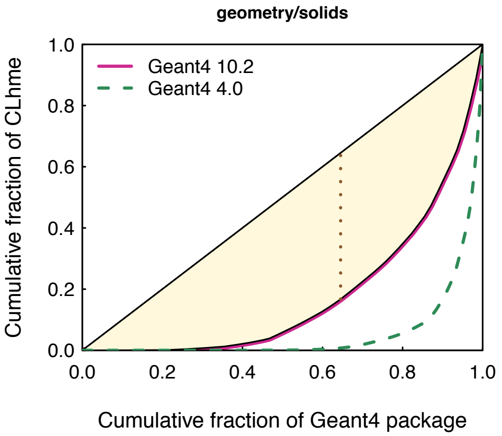

geometry/solids


geometry/solids


geometry/solids


processes/hadronic/models/cascade

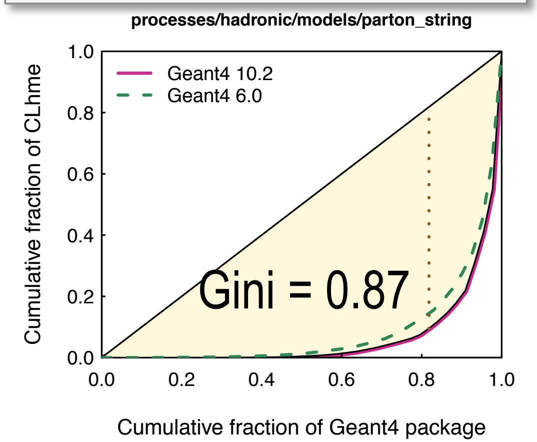

processes/hadronic/models/low_energy

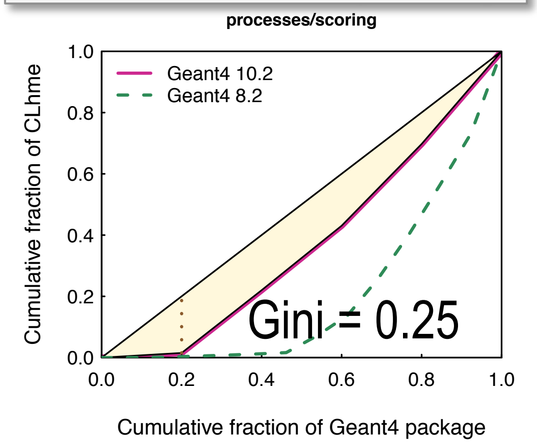

geometry/solids

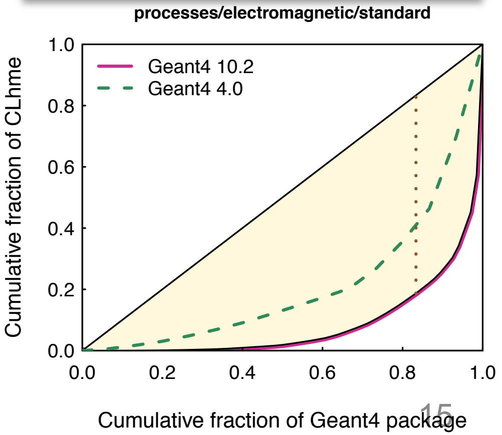
concentrated software complexity

distributed software complexity

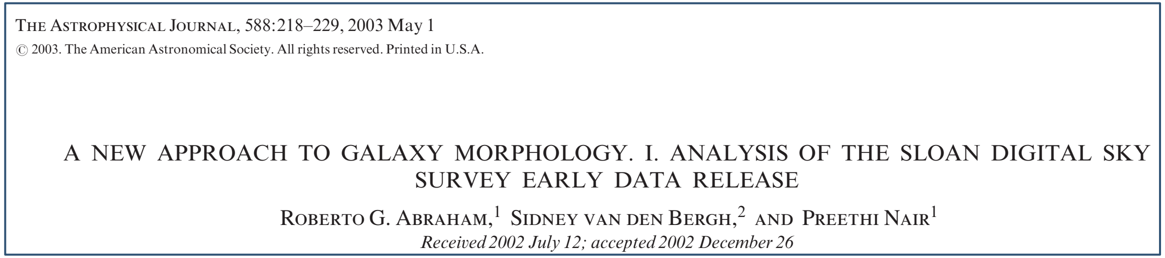
evolution of concentration


processes/hadronic/models/parton_string


processes/scoring


processes/electromagnetic/standard

Gini and galaxies



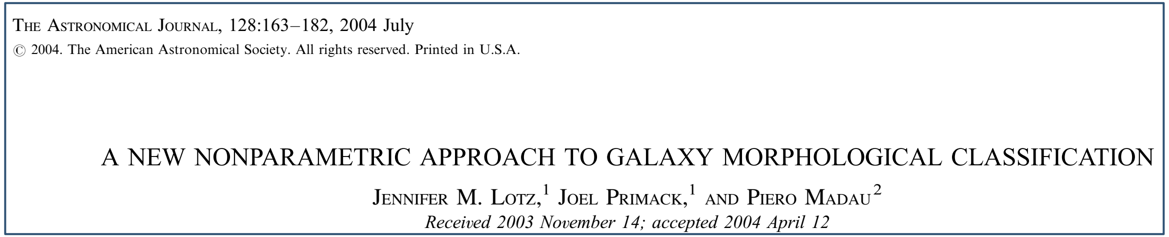
THE ASTROPHYSICAL JOURNAL, 588:218–229, 2003 May 1

© 2003. The American Astronomical Society. All rights reserved. Printed in U.S.A.

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



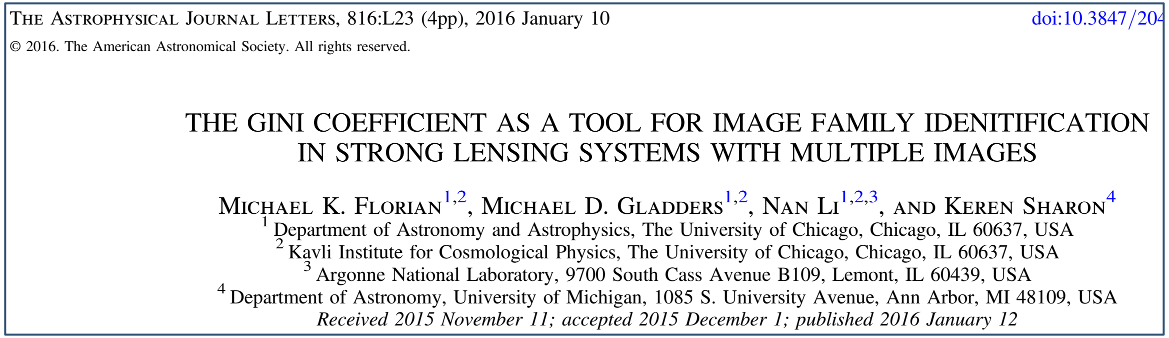
THE ASTRONOMICAL JOURNAL, 128:163–182, 2004 July

© 2004. The American Astronomical Society. All rights reserved. Printed in U.S.A.

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:10.3847/204

© 2016. The American Astronomical Society. All rights reserved.

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

Comparative evaluation of measures and tests

Decomposition of inequality measures by subgroups

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

Conclusion

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