

# A scientometric analysis of diversity in HEP over the past three decades

Maria Grazia Pia (*INFN Genova, Italy*)

[maria.grazia.pia@cern.ch](mailto:maria.grazia.pia@cern.ch)

Tullio Basaglia (*CERN*), Zane W. Bell (*ORNL*),

Arnold Burger (*Fisk. Univ.*), Paul V. Dressendorfer (*IEEE*)

**ICHEP 2018**

Seoul, 4-11 July 2018

# Diversity from a scientometric perspective

## Scientometrics

The study of **measuring** and **analysing** science

## Scope of this study

**Publications** in scholarly journals, 1985-2017

Particle physics, Nuclear physics, Astrophysics, HEP technology

## Data sources

Web of Science (*Clarivate Analytics*)

UN and OECD data

<http://data.un.org/>

## Observables

Countries, organizations, authors, journals

Descriptive statistics

Measures of **diversity**

Measures of **inequality**

Statistical inference: **trend tests**, **correlation tests**

# Caveat



- CERN cancelled the subscription to the Web of Science (1970→) in July 2017
- It is no longer possible to do any scientometric studies at CERN
- WoS access through INFN (1990→) and Univ. of Genova (1985→)

The WoS is affected by several known problems

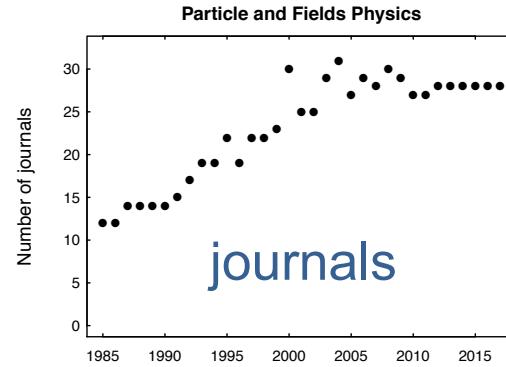
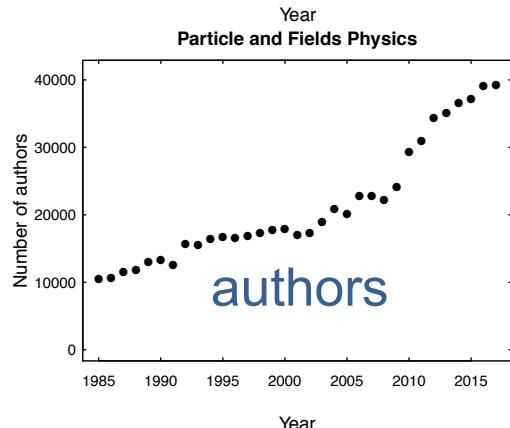
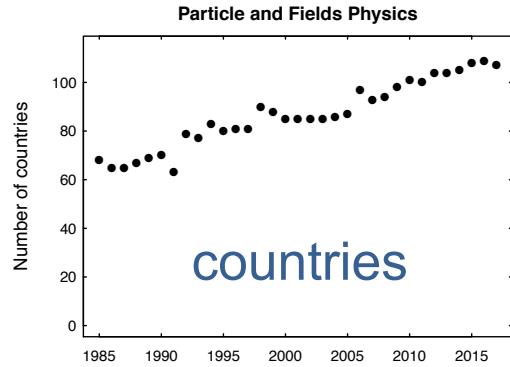
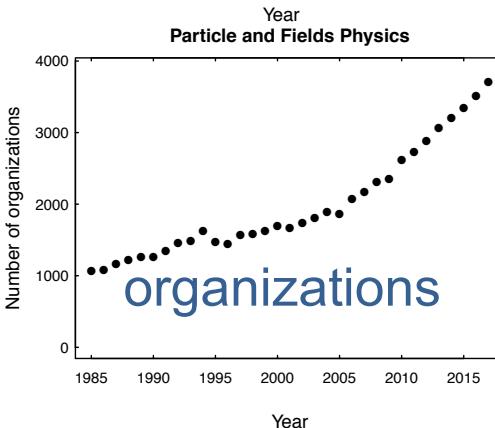
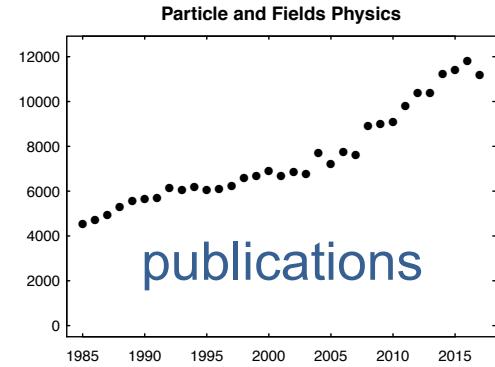
too many to mention,  
specific warnings in the course  
of the presentation

## WoS categories

journals  
~no conference papers

- Physics, **Particles & Fields** *(excluding NIM A)*
  - Physics, **Nuclear**
  - **Astronomy & Astrophysics** } *excluding intersection with Particle Physics, NIM*
- + **HEP technology** journals: NIM A/B, IEEE TNS, JINST

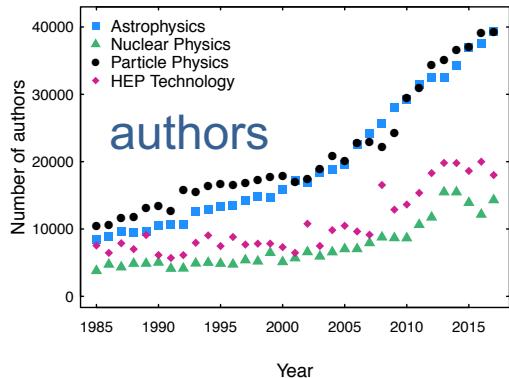
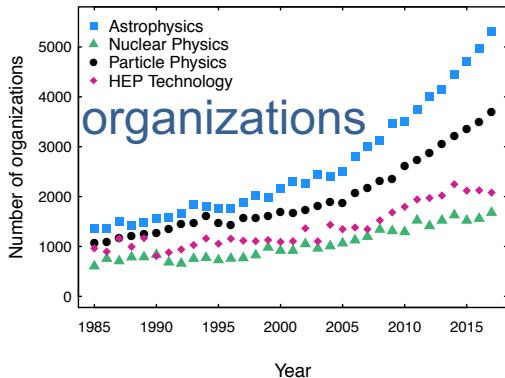
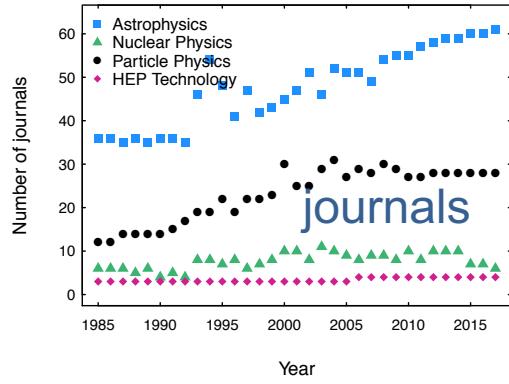
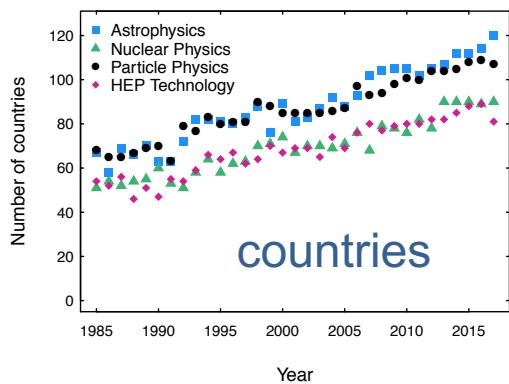
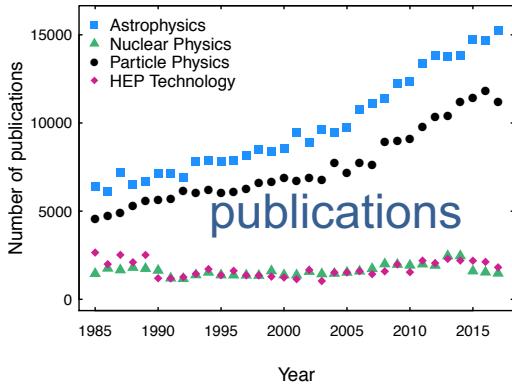
# 33 years of HEP publications



Year

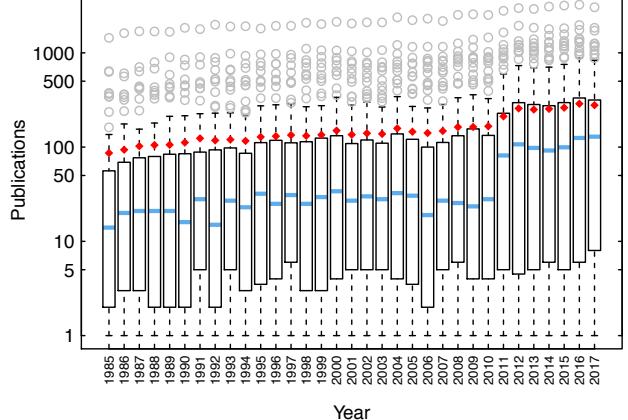
General increase of the number of publications, of journals, of participating countries, organizations and authors

# Compared to other research domains

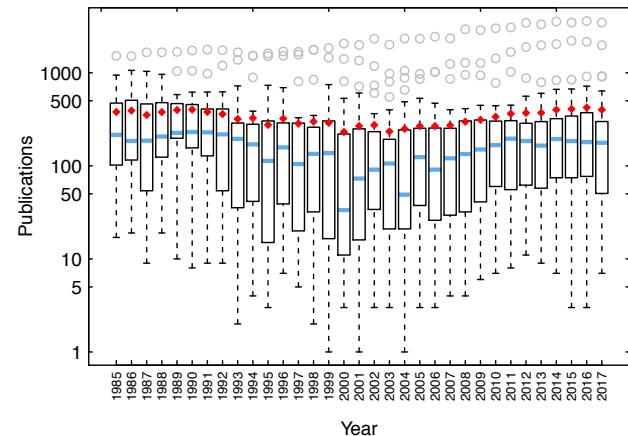


Beware:  
nuclear physics and  
astrophysics data sets  
exclude journals also  
classified as particle physics

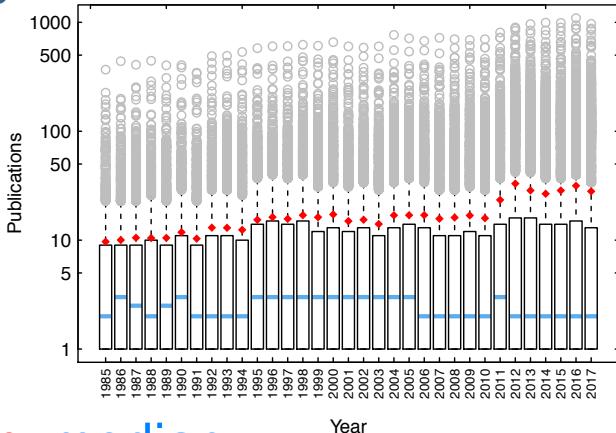
countries Particle: Country



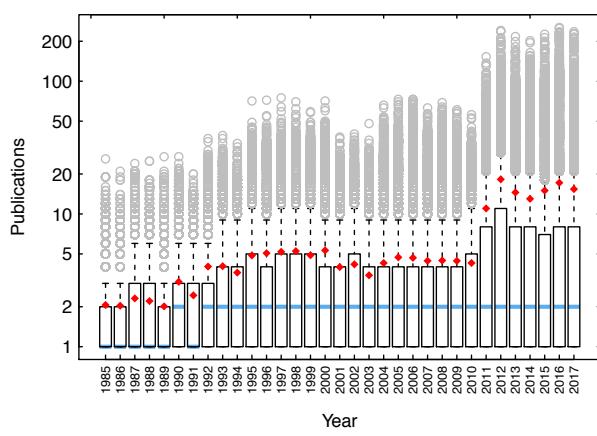
journals Particle: Journal



organizations Particle: Organization



Particle: Author



# Data distributions

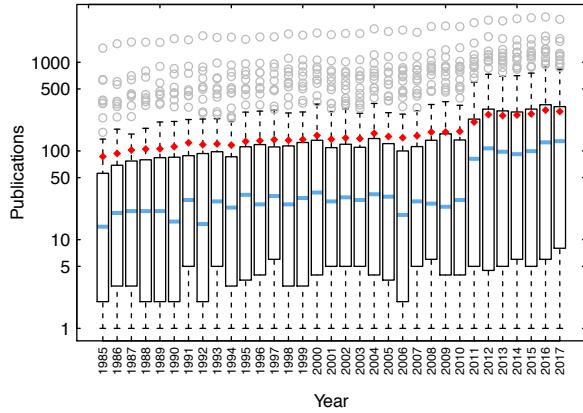
**Contradiction:**  
low and approximately  
constant **median**,  
**outliers** extending up  
to very large number of  
publications

**Food for thought:**  
scientific and sociological  
implications

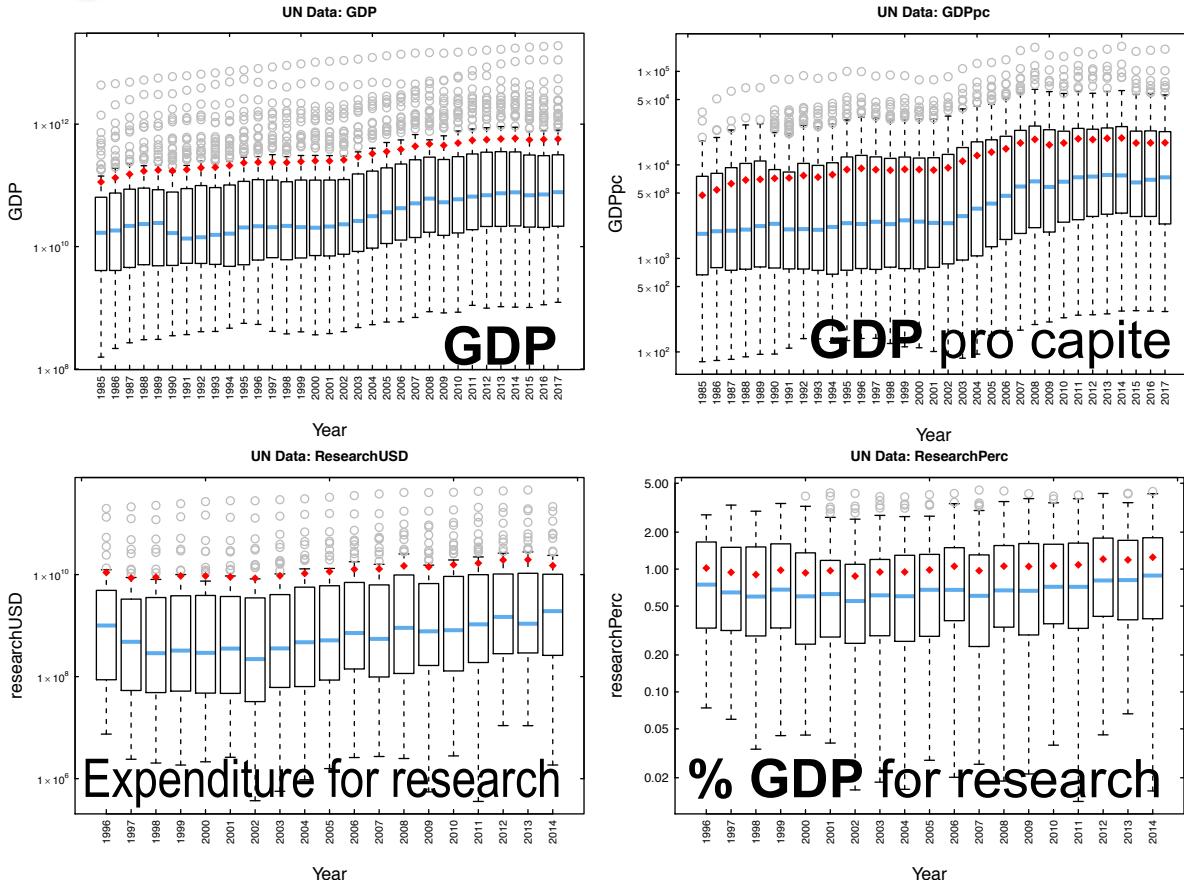
# The world changes...

## Particle physics publications

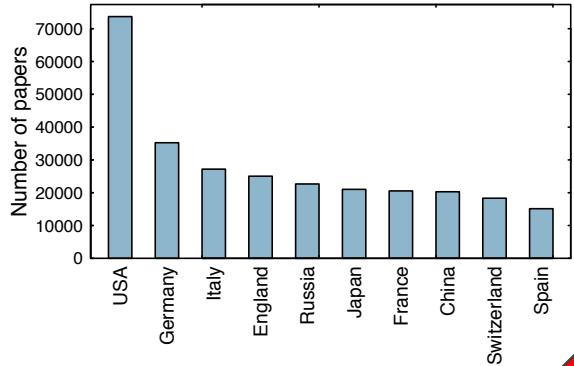
Particle: Country



Same set of countries  
as the WoS particle  
physics data



**Country 1985 – 2017**

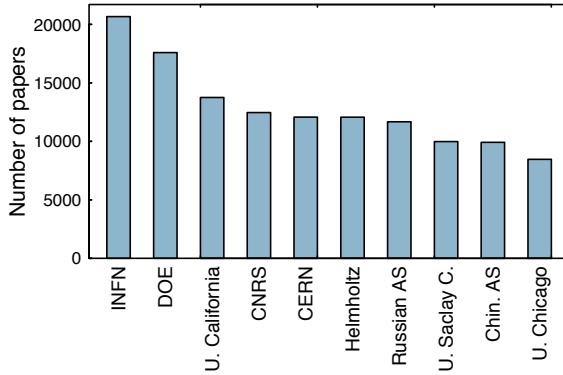


Particle physics

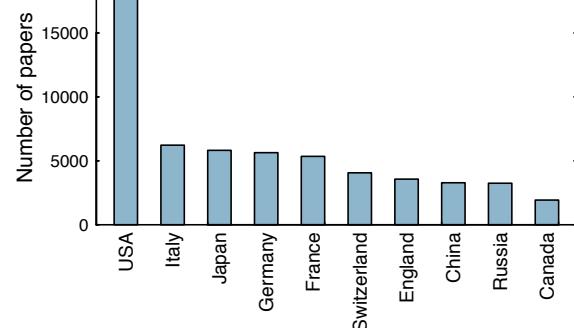
Top 10

Particle/nuclear technology

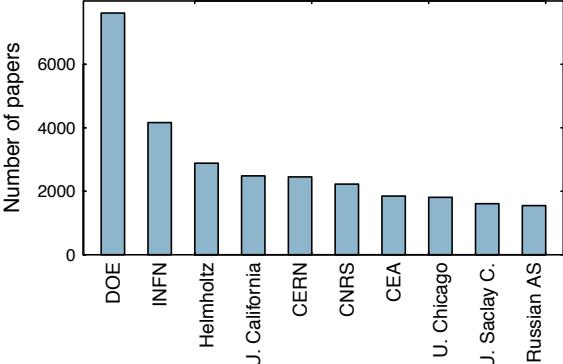
**Organization 1985 – 2017**



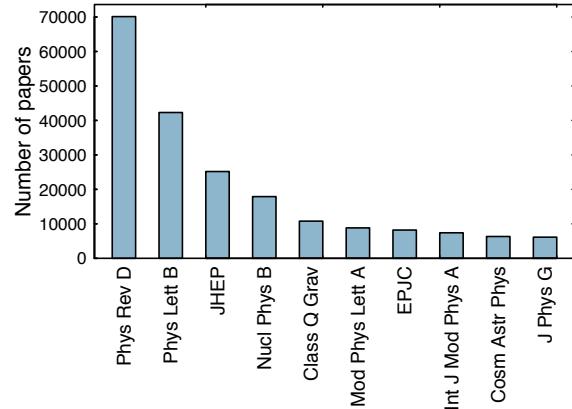
**Country 1985 – 2017**



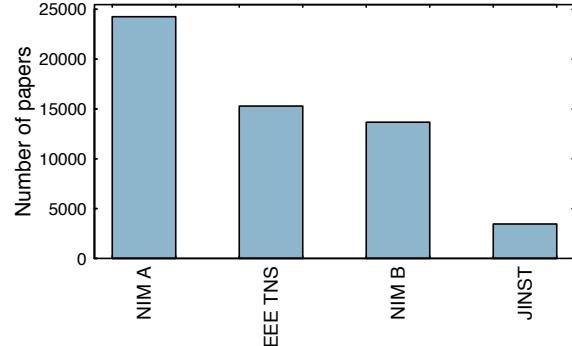
**Organization 1985 – 2017**

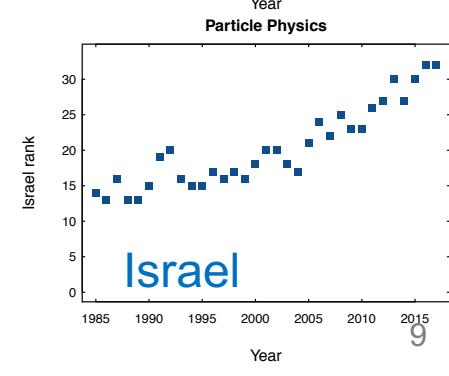
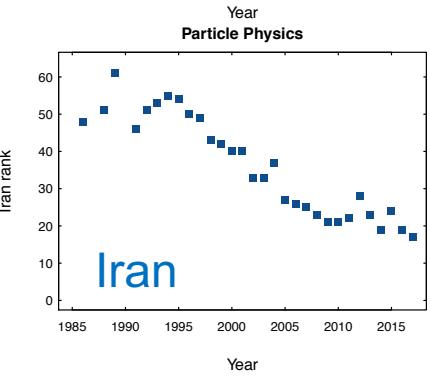
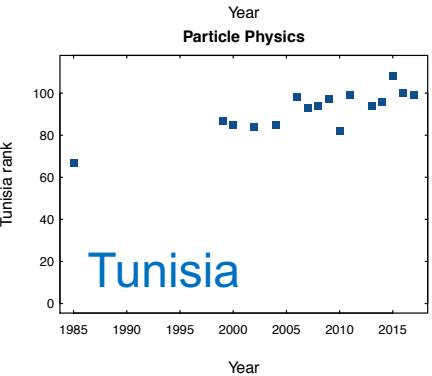
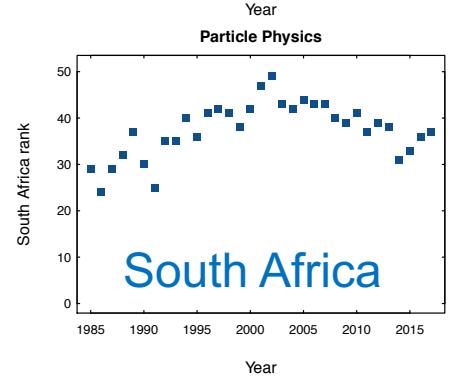
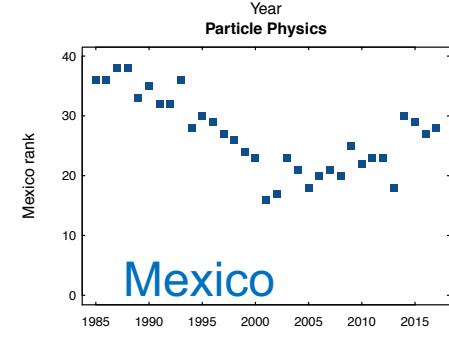
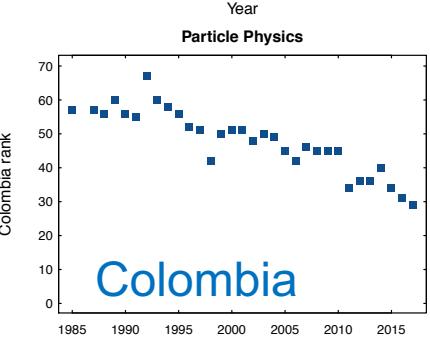
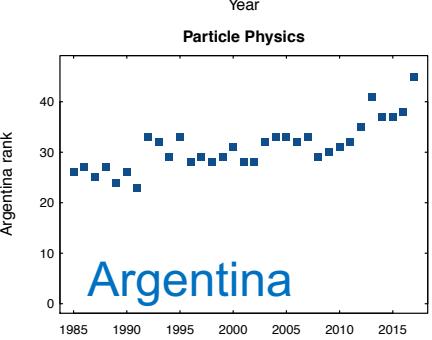
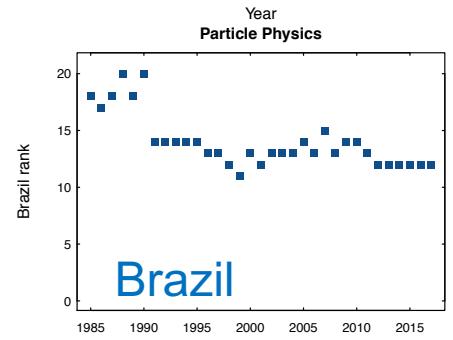
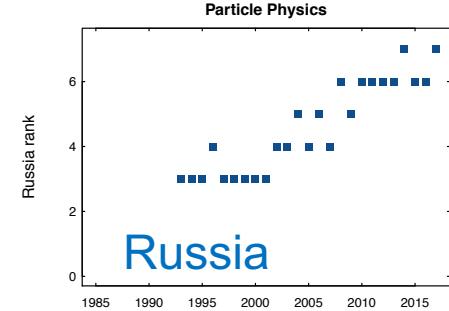
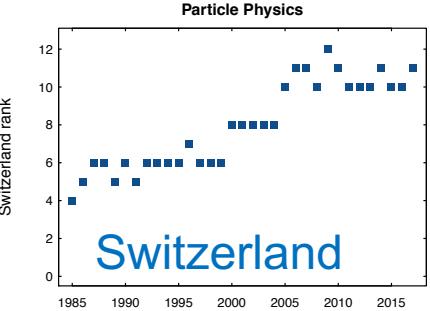
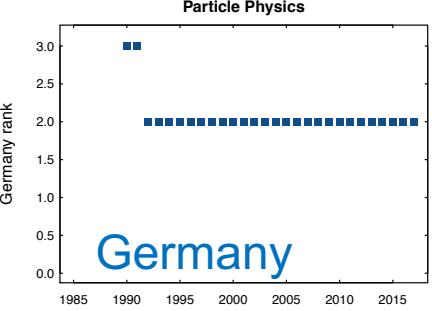
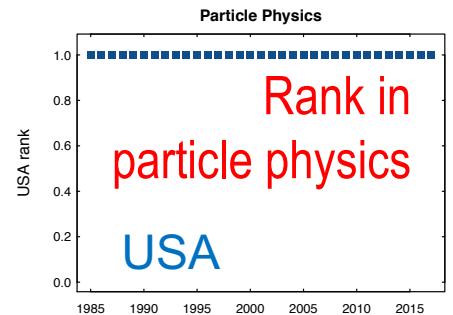


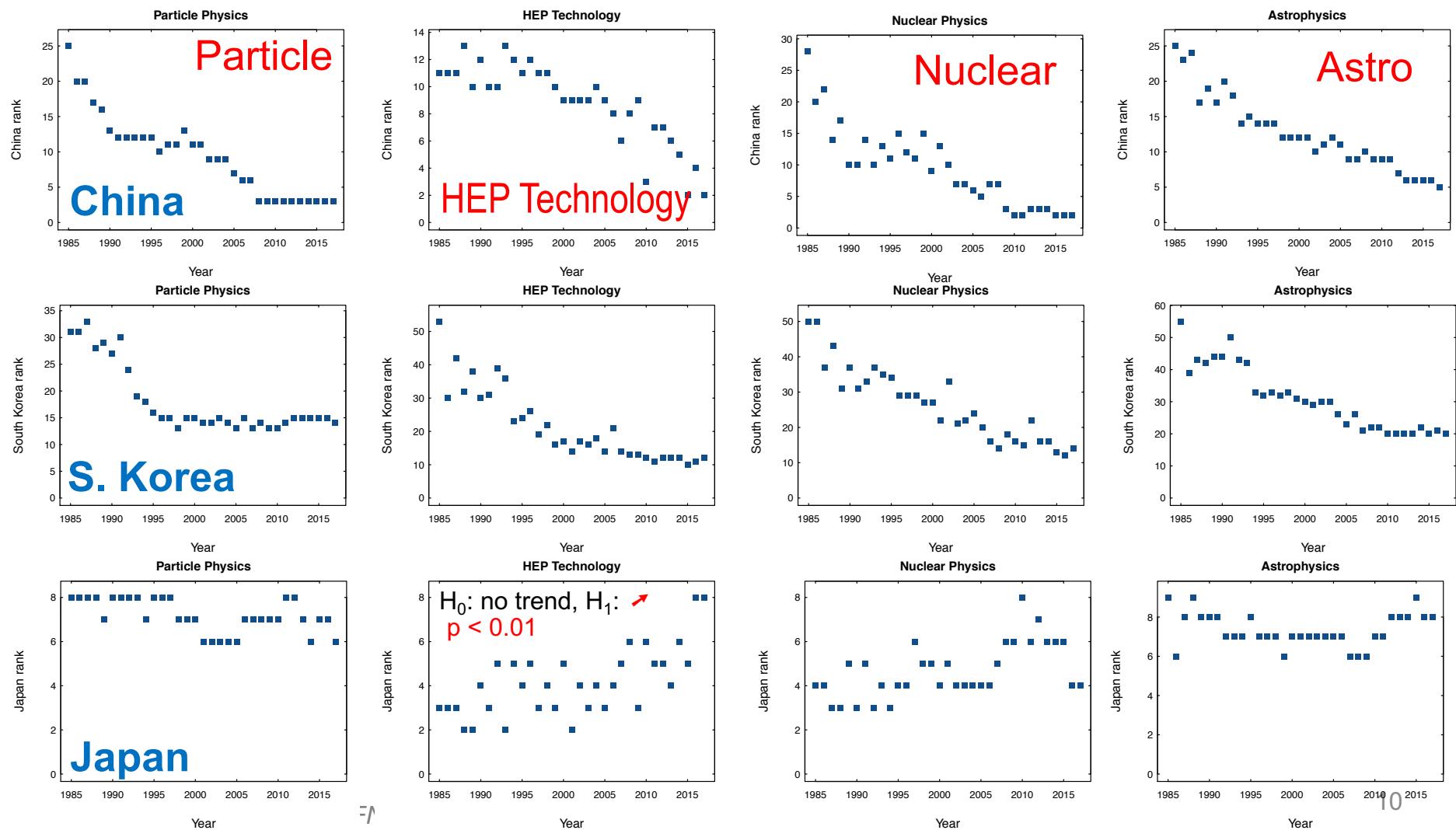
**Journal 1985 – 2017**

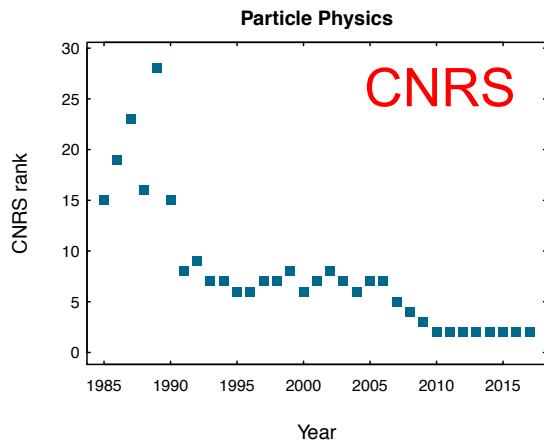
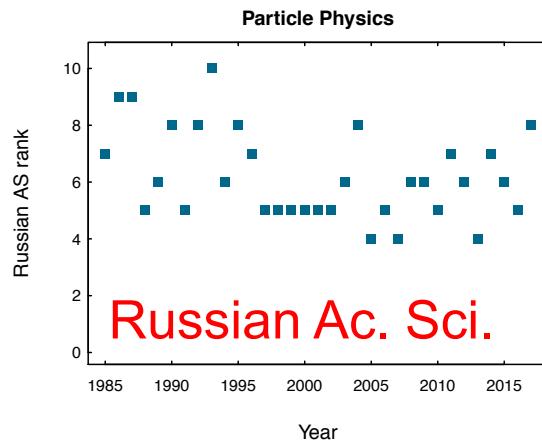
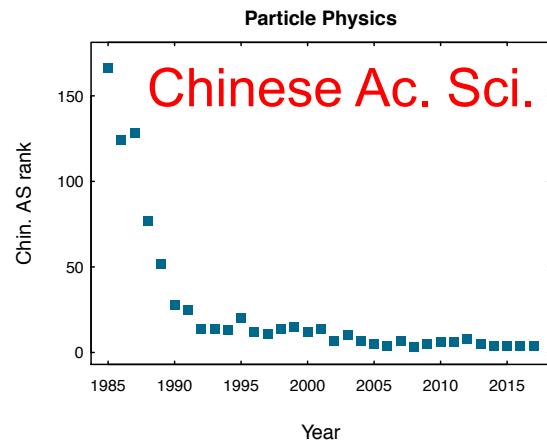
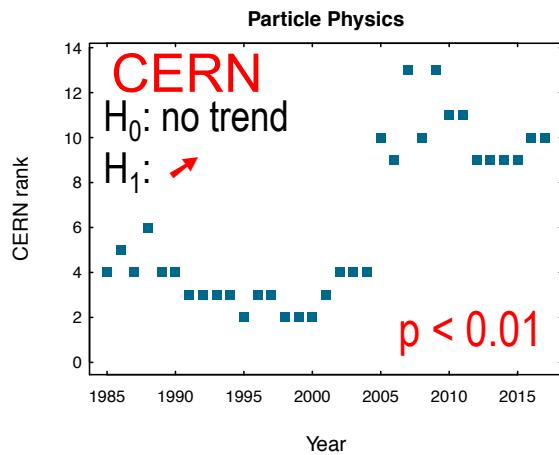
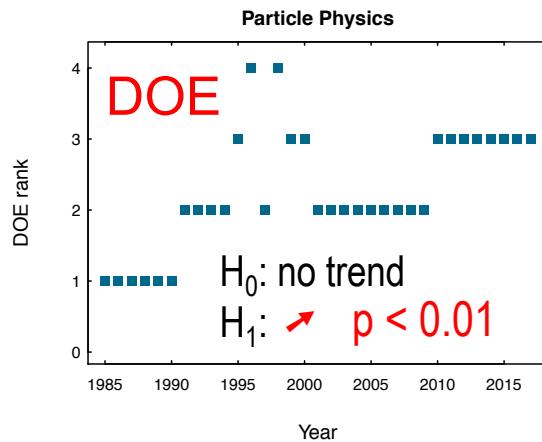
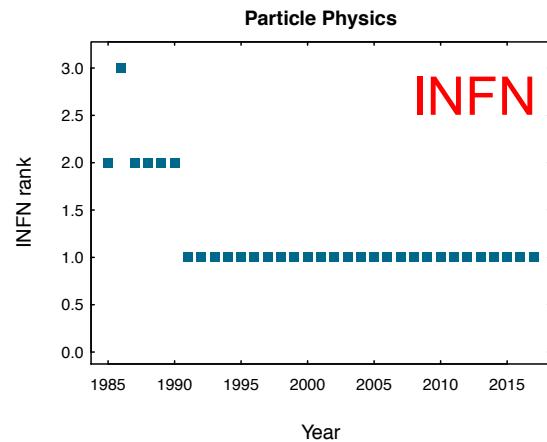


**Journal 1985 – 2017**









# Diversity

Concept drawn from **ecology**

**Species diversity** measures the different kinds of organisms living in a community

number of species  
and their abundance

- Related to the concept of **entropy** in information theory
- Measured by several indices, with different sensitivity to rare species
  - Mathematical functions that combine **richness** and **evenness** in a single measure
  - Correlations between indices due to their common background
  - Different measures may be appropriate, depending on the context

# Diversity measures

**S** = # of different species  
**N** = total # of individuals  
**p** = proportion of species

Margalef index

$$D_{Mg} = (S - 1) / \log N$$

Simpson index

$$\lambda = \sum_{i=1}^R p_i^2$$

Gini-Simpson

$$D_1 = 1 - \sum_{i=1}^S p_i^2$$

Shannon index

$$H = -\sum_{i=1}^S p_i \log p_i$$

Shannon entropy

Renyi diversity

$${}^q H_{Renyi} = \frac{1}{1-q} \log \left( \sum_{i=1}^S p_i^q \right)$$

extension of  
Shannon entropy

Tsallis diversity

$${}^q H_{Tsallis} = \frac{1}{1-q} \left( \sum_{i=1}^S p_i^q - 1 \right)$$

Hill numbers

$${}^q D = \left( \sum_{i=1}^S p_i^q \right)^{1/(1-q)}$$

0

species richness

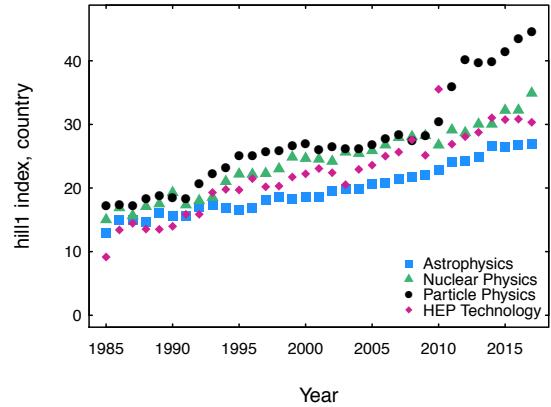
1

$$\lim_{q \rightarrow 1} {}^q D \equiv {}^1 D = \exp \left( - \sum_{i=1}^S p_i \log p_i \right) \exp(\text{Shannon})$$

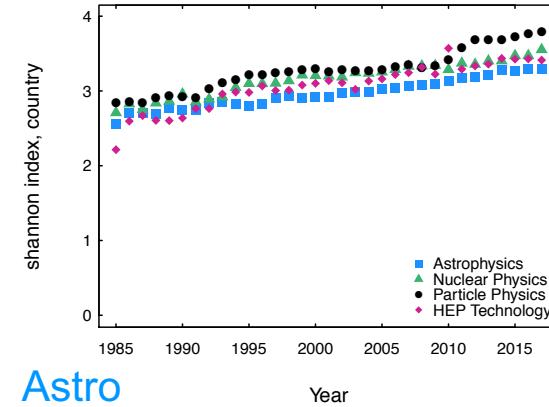
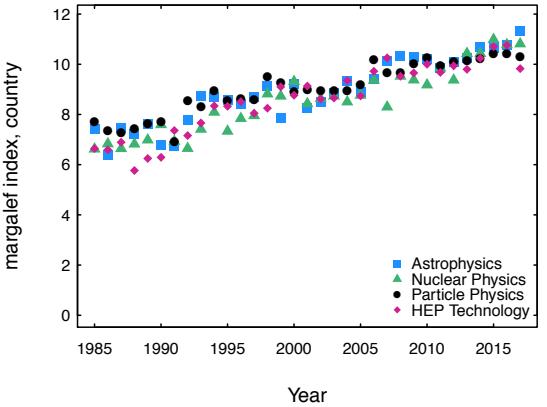
2

$${}^2 D = 1 / \sum_{i=1}^S p_i^2$$

# Country

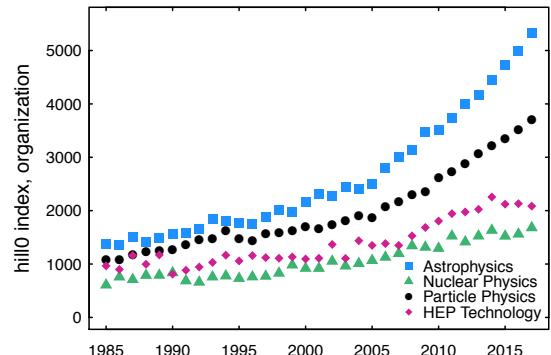


**Species richness  
(Hill<sub>0</sub>)**

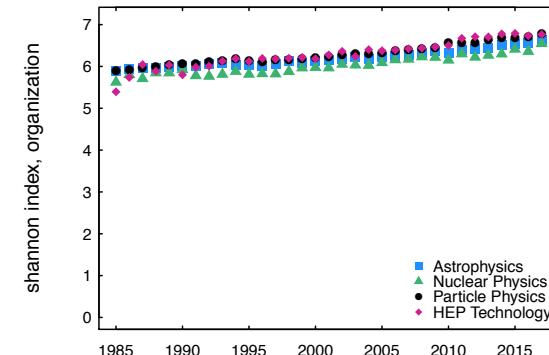
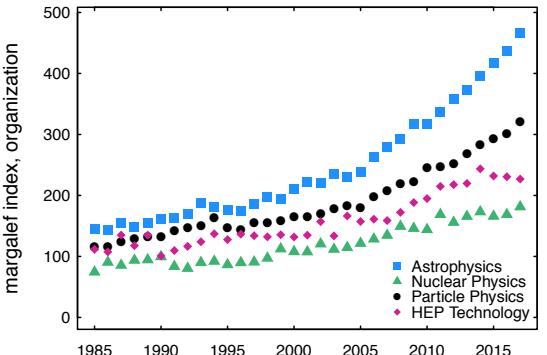


**Astro**  
**Nuclear**  
**Particle**  
**HEP technology**

# Organization



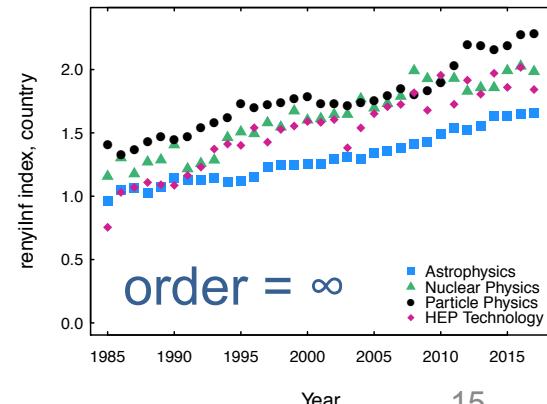
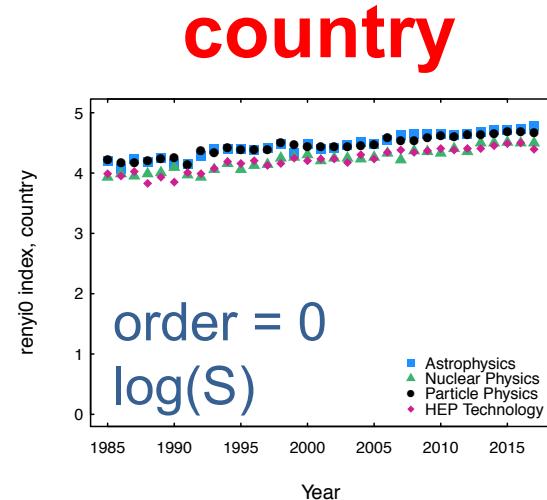
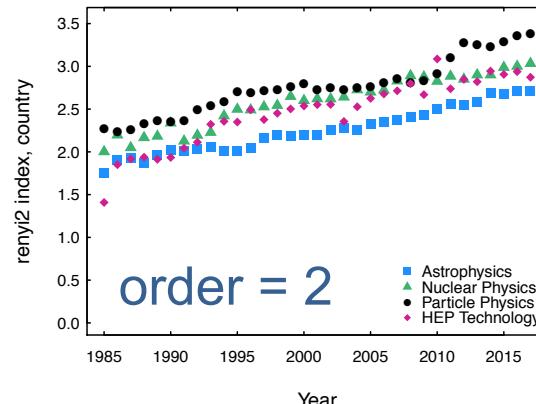
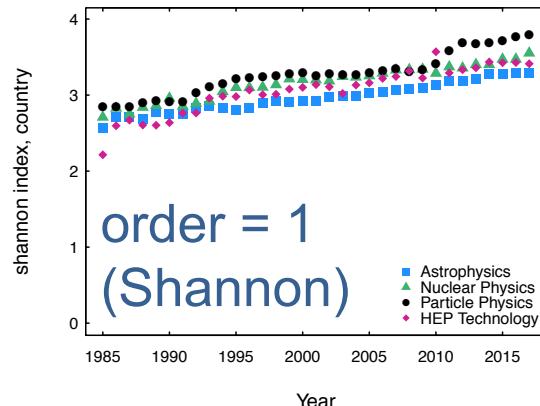
Maria Grazia Pia, INFN Genova



# Renyi diversity

Extension of Shannon entropy

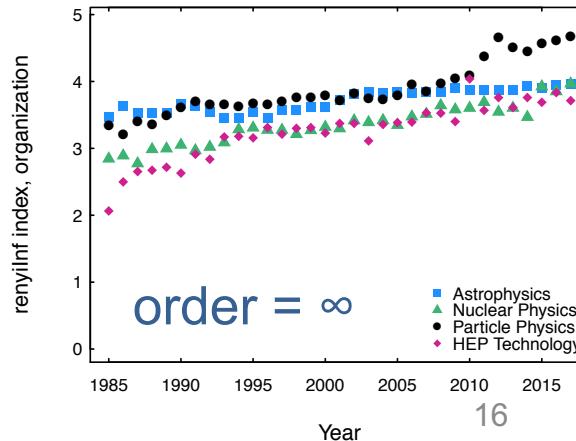
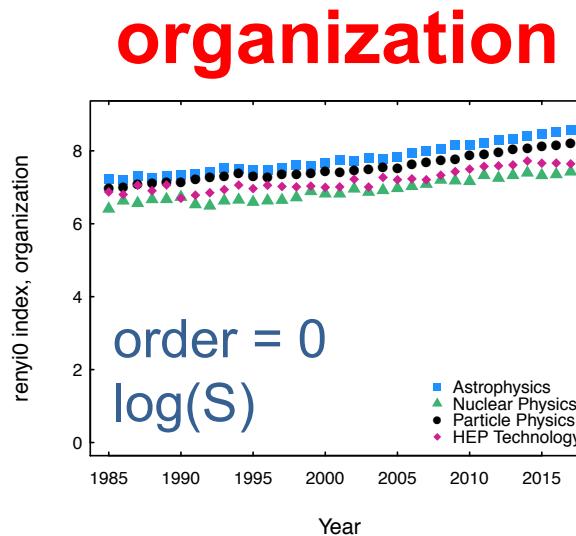
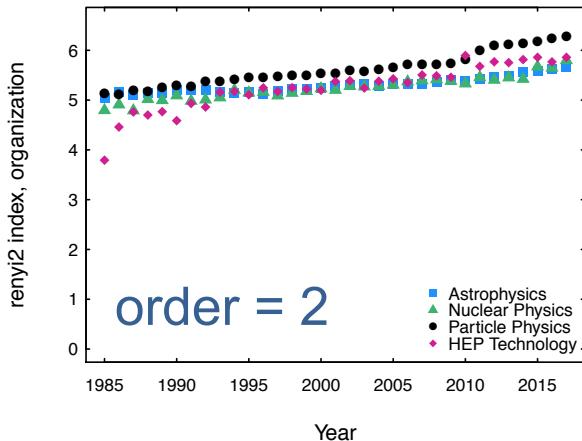
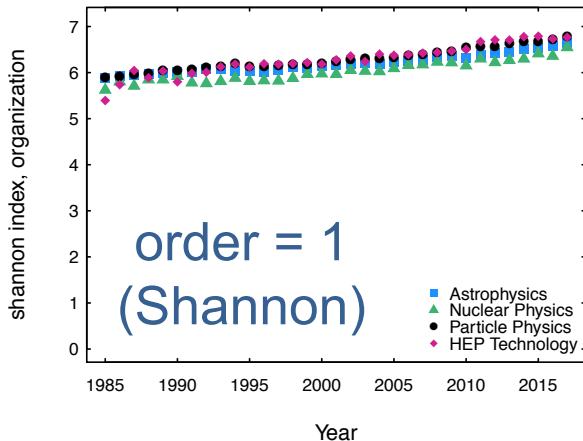
The higher the order ( $q$ ),  
the lower the sensitivity to rare species



# Renyi diversity

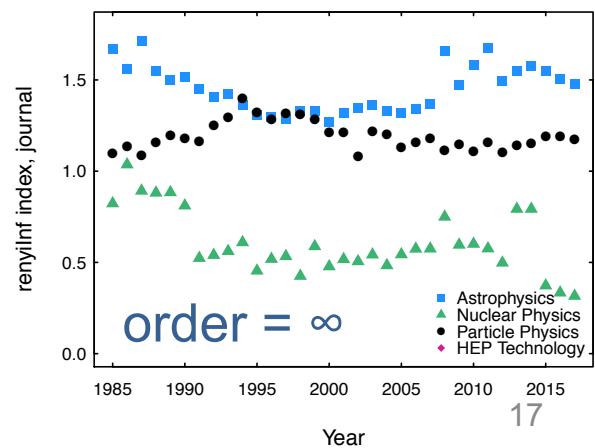
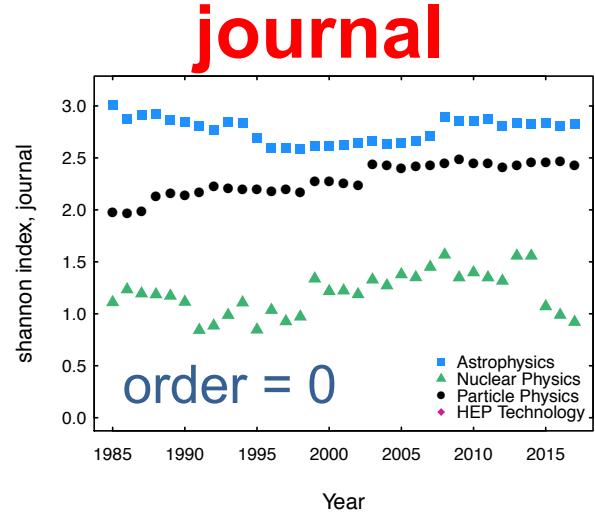
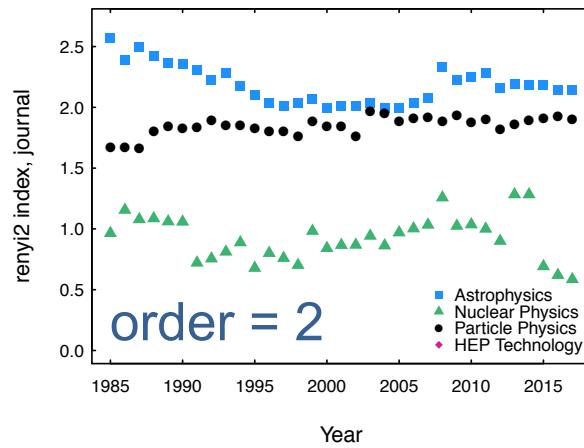
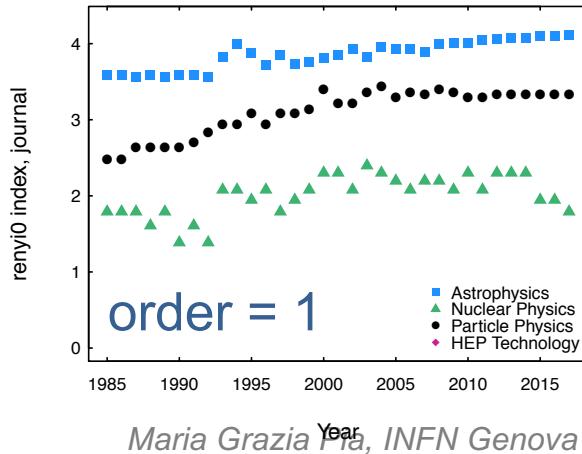
Similarity across all disciplines

Lower values at higher order, where rare species (*organizations*) have lower impact  
Trend: slowly increasing



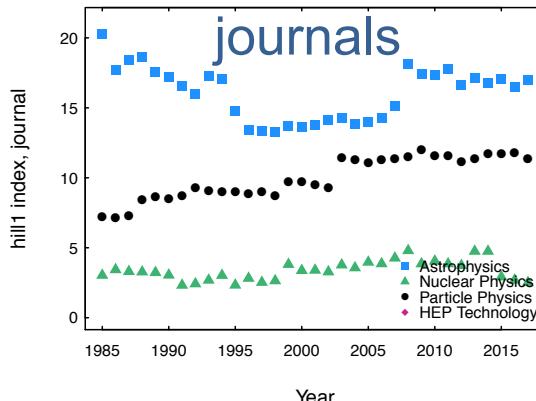
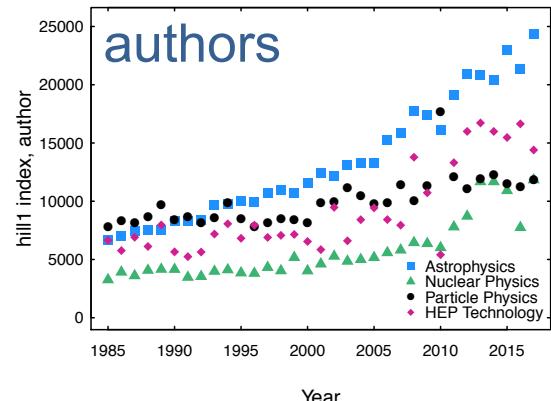
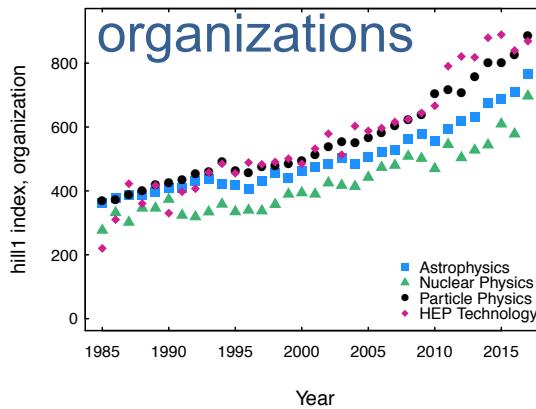
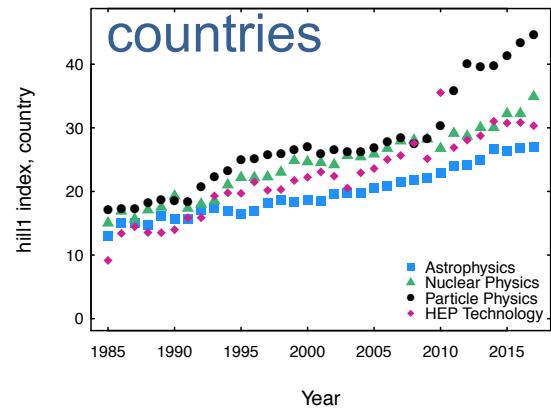
# Renyi diversity

Lower values at higher order, where rare species (*journals*) have lower impact  
Trend



# Hill number of order 1

$\exp(\text{Shannon})$



General consensus  
reached relatively  
recently on Hill numbers  
as measures of diversity

# All that glitters...

*Large and generally increasing diversity in HEP publications,  
but...*

How fairly are scholarly  
publications distributed within the  
scientific community?

**Econometric analysis**

**Inequality**

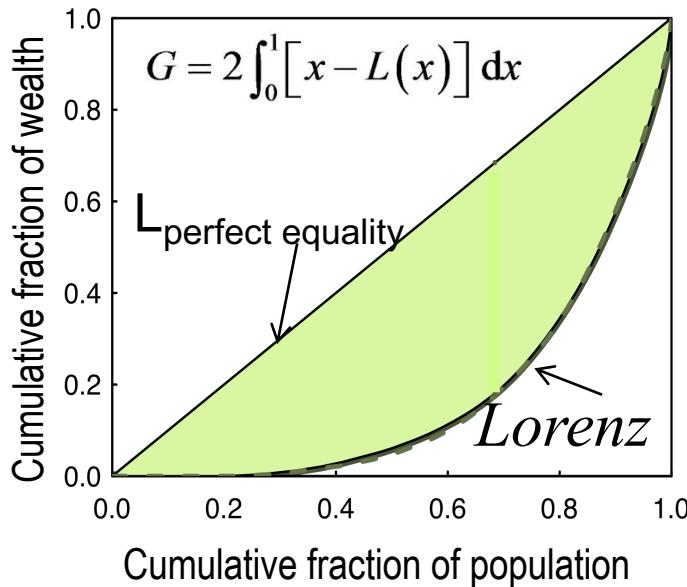
How does their distribution evolve  
as a function of time?

**Trend**

# Gini index

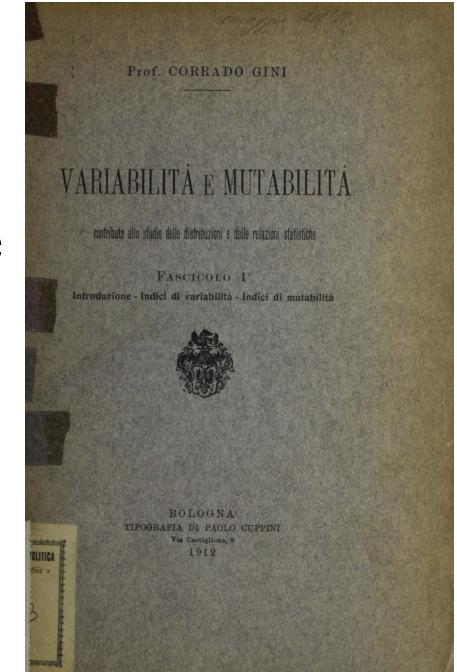
C. Gini, Variabilità e mutabilità : contributo allo studio delle distribuzioni e delle relazioni statistiche, 1912

Most common measure of inequality



“The  $x$  richest people in the world are worth more than the poorest  $y\%$ ”

0  $\leq$  **G**  $\leq$  1  
more unequal society

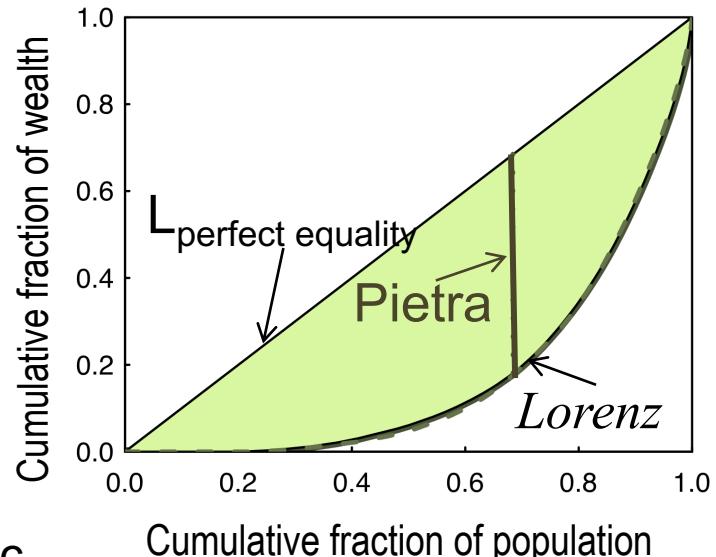


# Pietra index

AKA Ricci-Schutz index, Hoover index,  
Robin Hood index

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

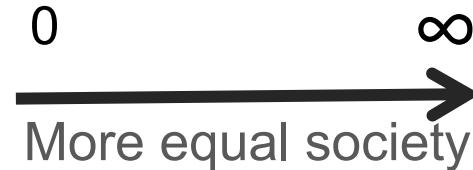
- Used in derivative markets as a benchmark measure of **statistical heterogeneity**
- Counterpart of Kolmogorov-Smirnov statistic
- It can be interpreted as the proportion of income that has to be transferred from those above the mean to those below the mean in order to achieve an equal distribution



# Other inequality measures

## Theil index

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



$s_i$  = share of the  $i^{\text{th}}$  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

$$I = 1 - \pi_e / \mu \quad 0 \leq I \leq 1$$

e = sensitivity parameter

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

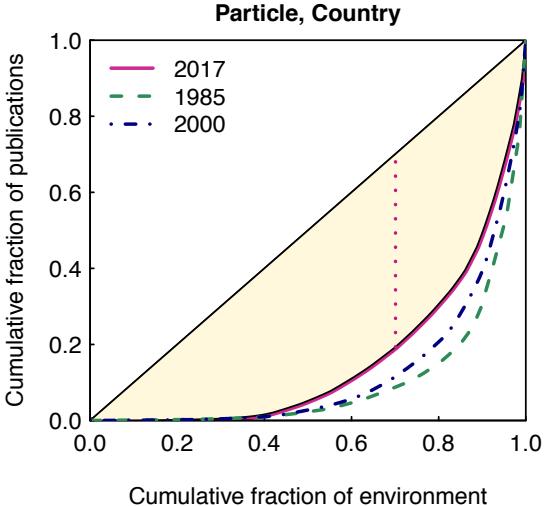
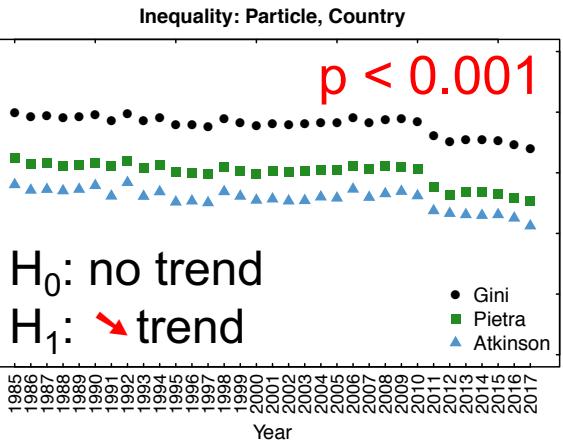
*Theil I, Theil II,  
Kolm index,  
coefficient of variation,  
generalized entropy  
and more...*

# Inequality evolution

**Trend tests**

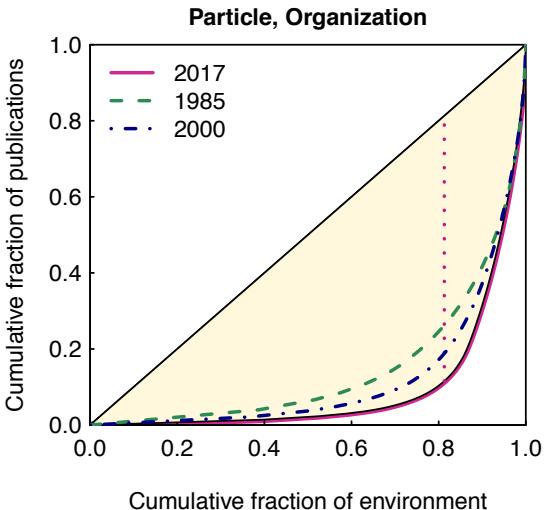
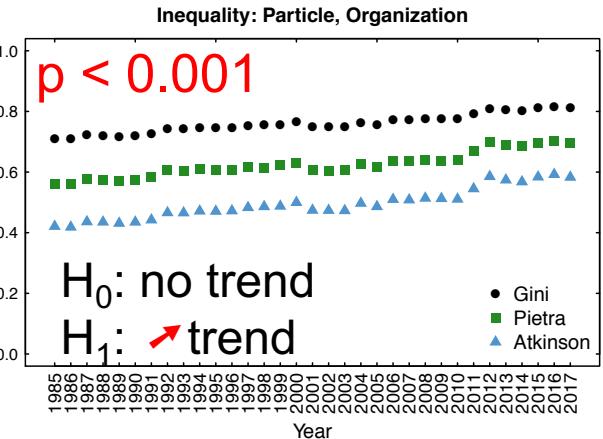
Mann-Kendall, Cox-Stuart  
 $\alpha = 0.01$

entropy index



Similar trends in Gini,  
Pietra and Atkinson indices

entropy index

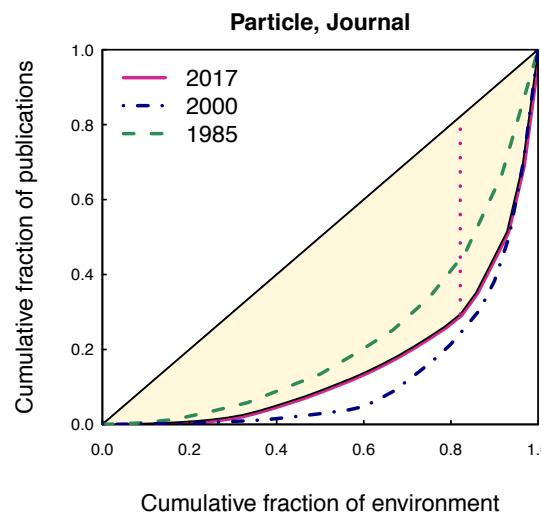
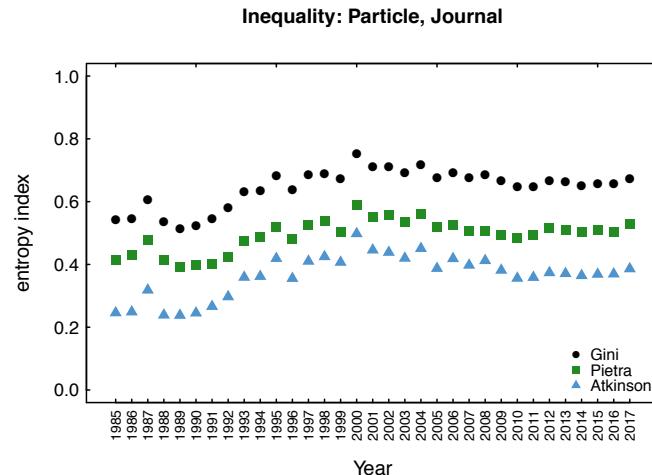
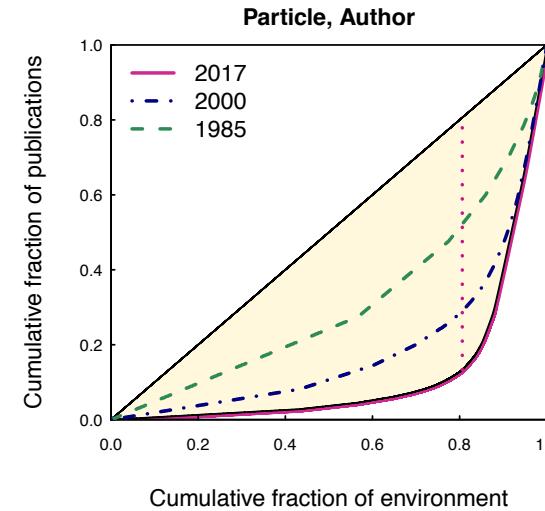
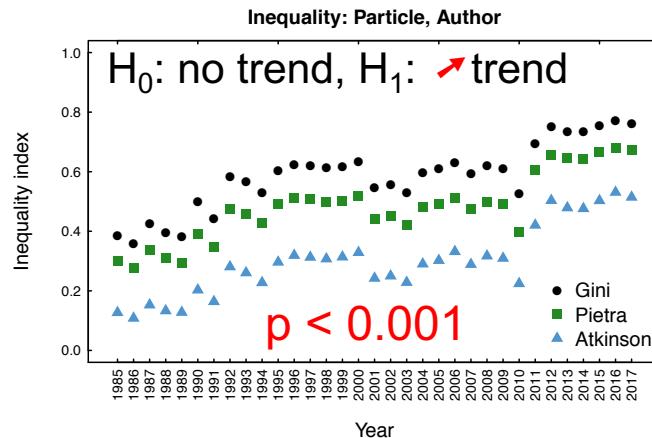


Inequality decreases  
across countries, but  
HEP publications are  
more and more  
concentrated within a  
small number of  
organizations

# Inequality evolution

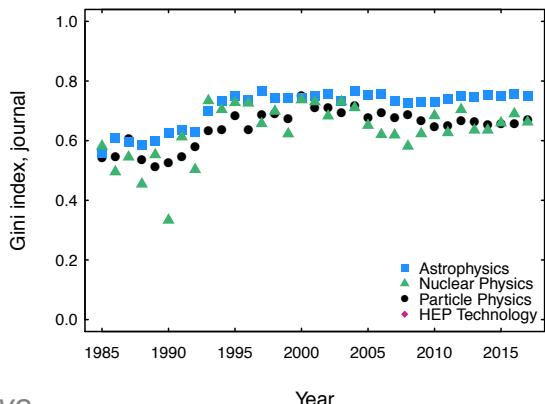
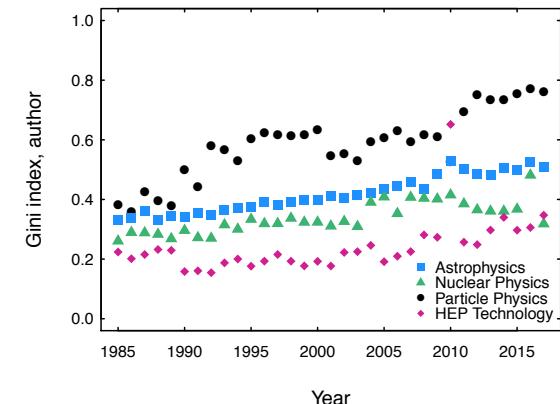
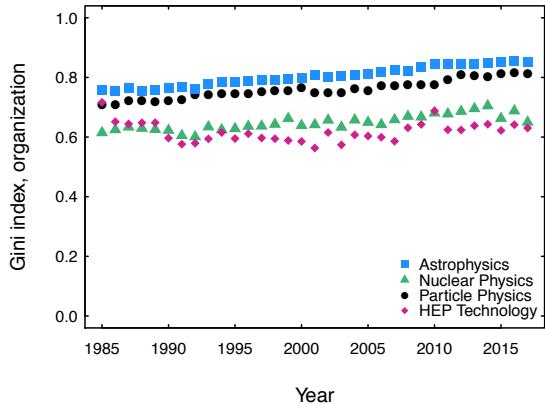
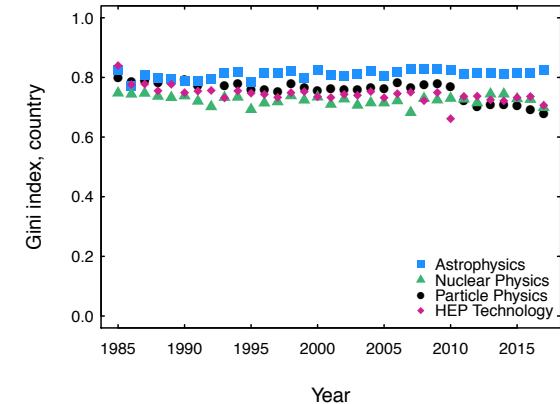
Inequality among HEP authors has been increasing over time

No univocal indication of general trend regarding HEP journals



# Gini index

Organization and author inequality in HEP technological publications is lower than in physics publications



Particle physics exhibits the largest inequality among authors,  
HEP technology the lowest

$p < 0.01$

$H_0$ : no trend	$H_1$ : $\nearrow$ trend	$H_1$ : $\searrow$ trend
Particle	organization author	country
HEP technology	author	country
Astro	country organization author journal	
Nuclear	organization author	

# Conclusions

Publication patterns in particle physics and related disciplines are studied with **statistical analysis** methods derived from ecology and econometrics

**objective  
quantitative**

evaluation of status and evolution

In general, evolution towards **greater diversity** in HEP, but **increasing concentration** in a small number of organizations

**Food for thought:** evolution of authorship and its meaning

“Biodiversity is the totality of all inherited variation in the life forms of Earth, of which we are one species. We study and save it to our great benefit.

We ignore and degrade it to our great peril.” *E.O. Wilson*