Measuring Dataset Impact: Data Citation as an Economic Process

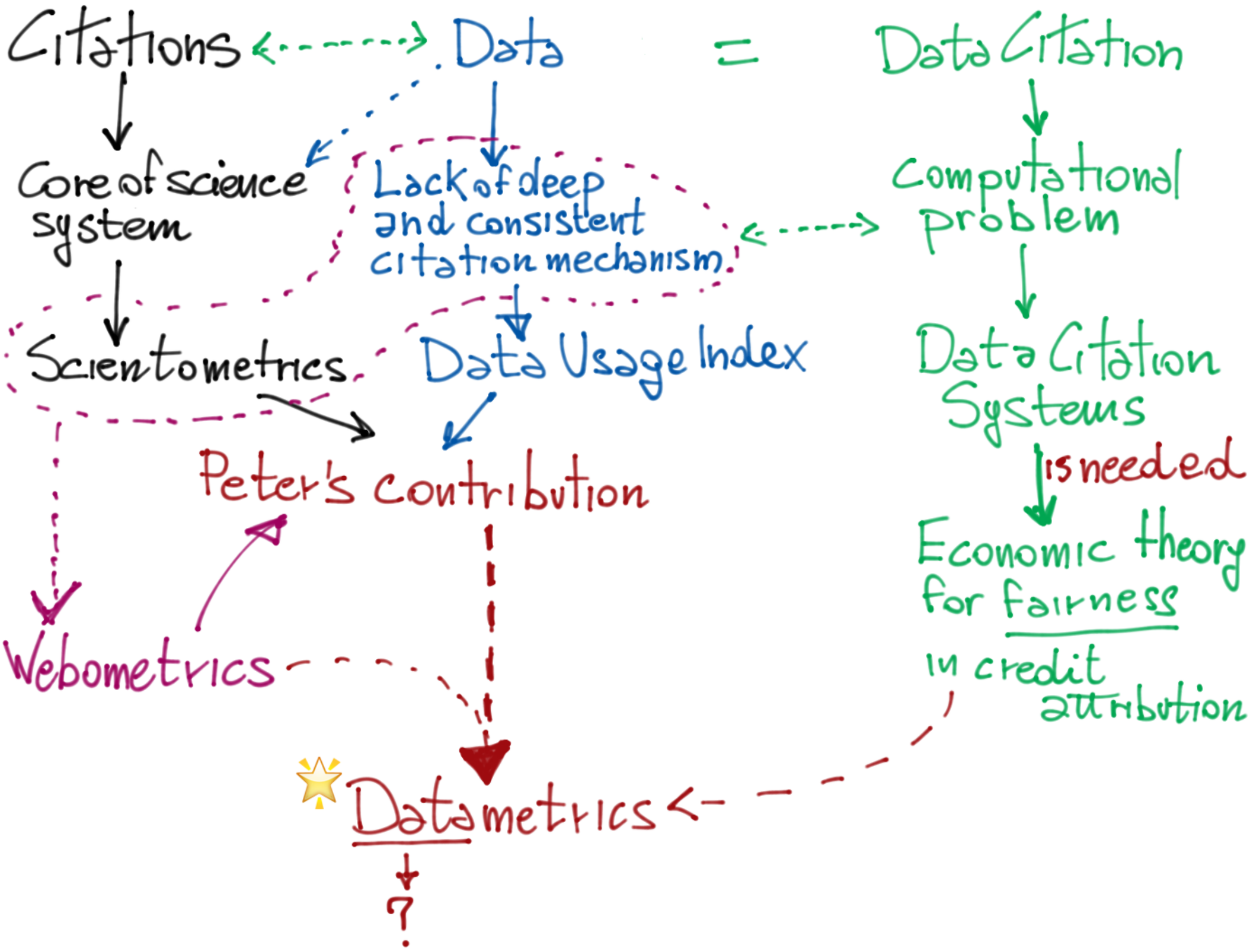
Gianmaria Silvello
@giansilv

Information Management Systems Research Group
Department of Information Engineering
University of Padua, Italy

Information Retrieval and Interaction Fest in Honour of Peter Ingwersen

Copenhagen, Denmark, 20 October 2016

Outline

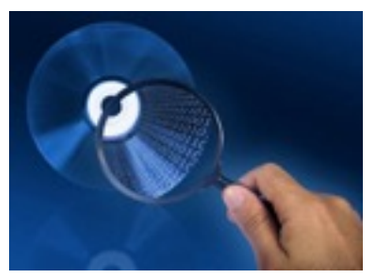

Citations
Data
=
Data Citation
Core of science system
Lack of deep and consistent citation mechanism
Computational problem
Scientometrics
Data Usage Index
Data Citation Systems
Peter's contribution
is needed
Economic theory for fairness
Webometrics
in credit attribution
Datametrics
?

Why Data Citation is Important?

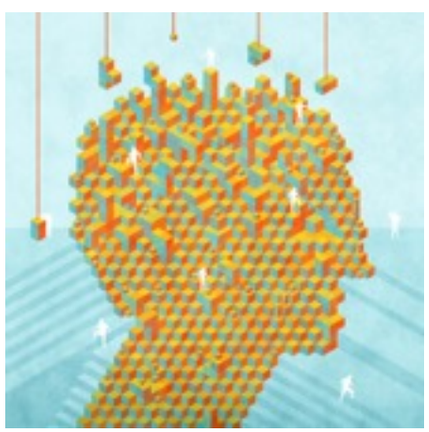
- Give credit to data creators and curators (and institutions)

- Repeatability, reproducibility and generalizability of research

- Referencing data in order to identify, discover and retrieve them



- Building and propagating knowledge



~~Why~~ Data Citation is Important ~~?~~

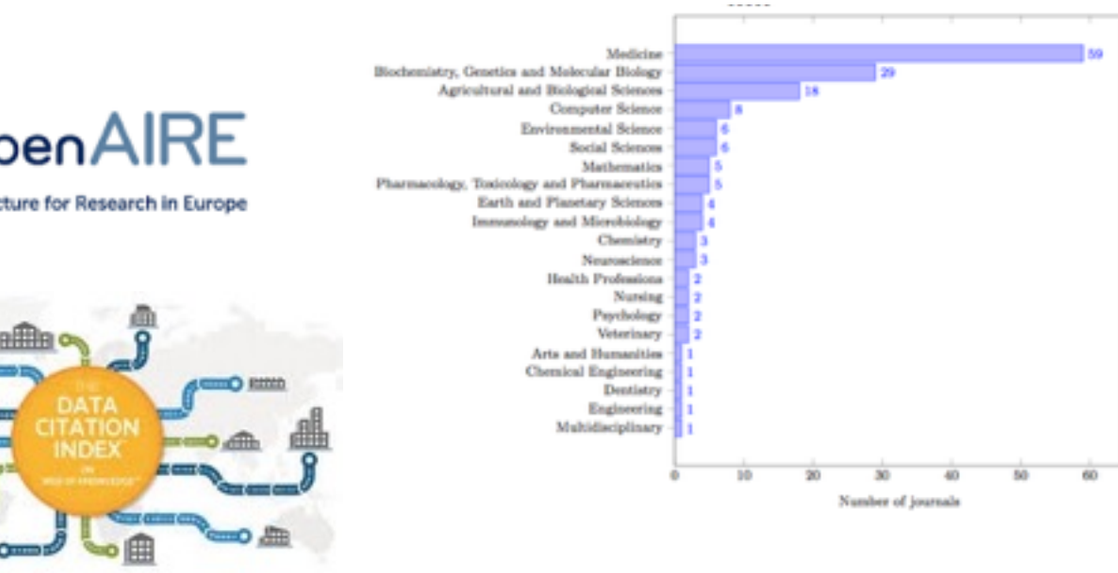
A lot of work has been done…

- Principles of data citation

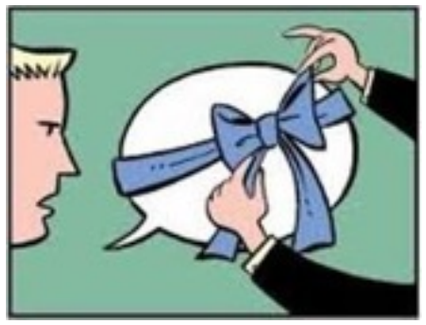
- Recommendations for data citation systems

- Data publishing infrastructures and data journals

- Indexes and dataset impact



~~Why~~ Data Citation is Important ~~?~~



(euphemism)

The practice of citing data is still not pervasive in scientific publishing

We need a deep and permanent data citation mechanism

[Ingwersen and Chavan, 2011]

From the user perspective

From the user perspective

- The generation of human- and machine-readable citations should be automatic

- Cited data should be uniquely identified: use DOI (?)

- Citing data should be easy: click, generate, copy and paste

- Setting up and maintaining a citation system should require low (no) effort to data creators/curators

From the computer scientist perspective

From the computer scientist perspective

- Data is not (always) fixed, it changes

- Persistent identifiers are (only) part of the solution

- Variable granularity (deep citations)

- Automatic generation of citation snippets (yes, but how?)

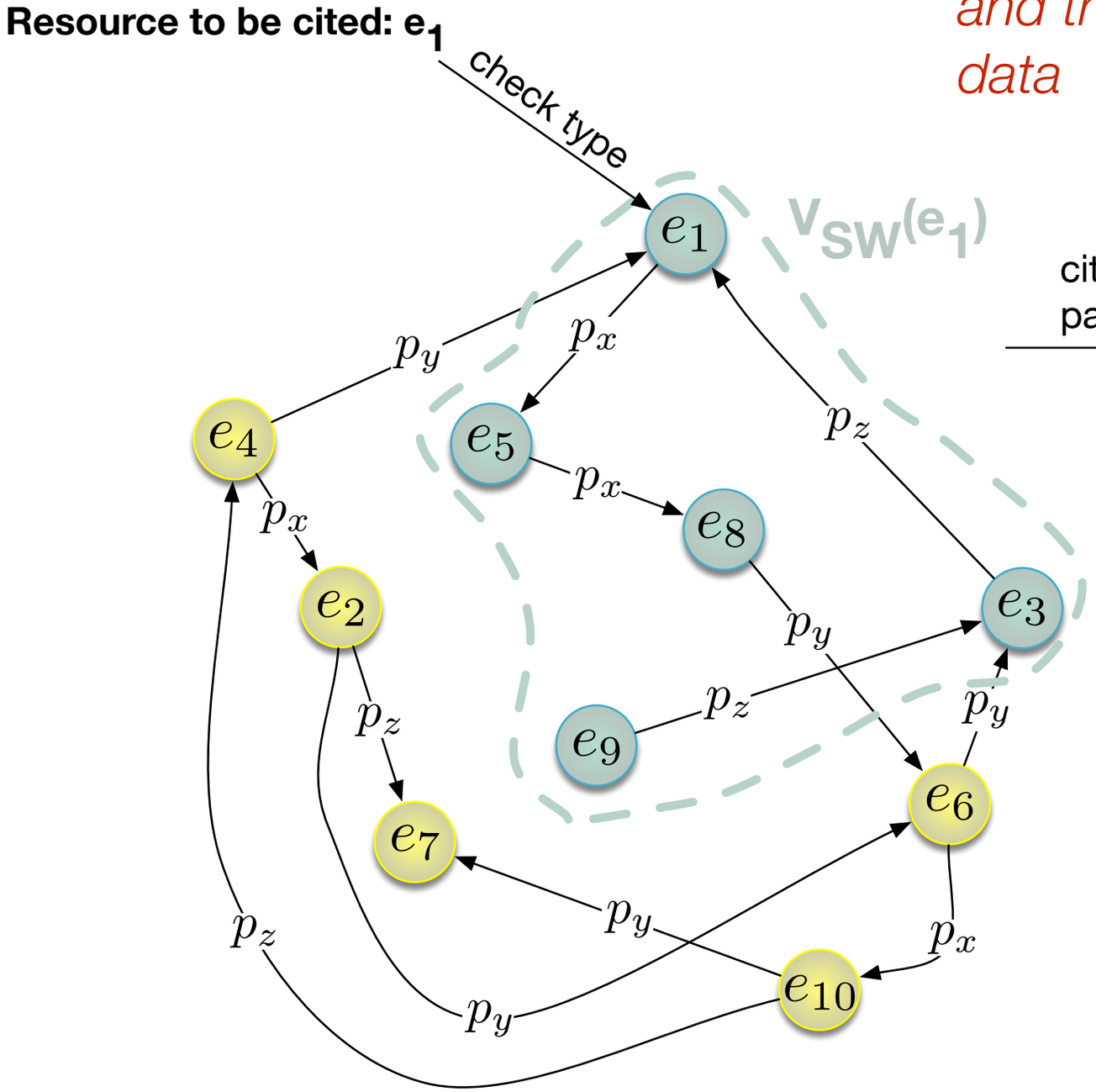
- Different data types and formats

We started to build automatic techniques

- XML: Rule-based system [Buneman&Silvello, 2010]
- XML: View-based system [Buneman et al., 2016]
- XML: Learning to cite framework [Silvello, 2016]
- Relational DB: View-based model [Davidson et al., 2017]
- Relational DB: Queries as proxies for data [Rauber et al., 2016]
- RDF: Named graphs based method (again views) [Silvello, 2015]
- RDF: View-based method [almost ready to be submitted and (maybe) accepted in 2017]

Views seem to be central

Citation views are the basis of the citation model and they define who is gaining credit for which data



citation query parametrized by e_1

$C_{SW}(e_1,s,v,d,t,o,u)$

Citation Function

Final citation

{eagle-id: “eagle-id: e1",
name: `”Significance Tester",
developers: {“Grant, G.", “Lazar, M.l", “Manduchi, E."},
url: “http://www.cbil.upenn.edu/STAR/ " }

Fairness and Truthfulness of the Models

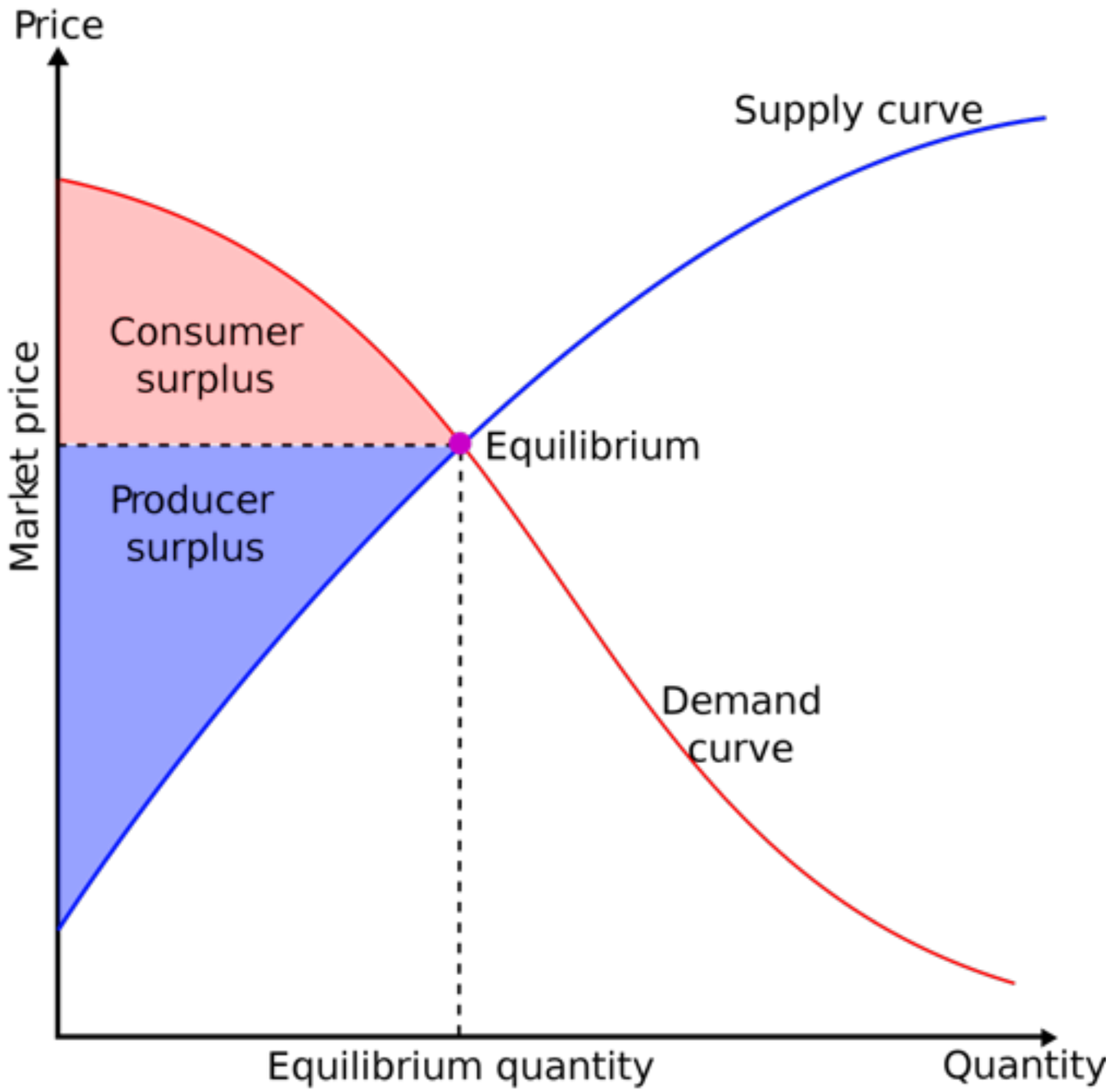
Are there systematic biases introduced by these system?

Is the credit attribution mechanism:

- *fair:* credit attribution is not inflated
- *truthful*: scientists are motivated to honestly report their contributions
- *efficient:* credit needs to be computed from the data in polynomial time

An Economic Perspective

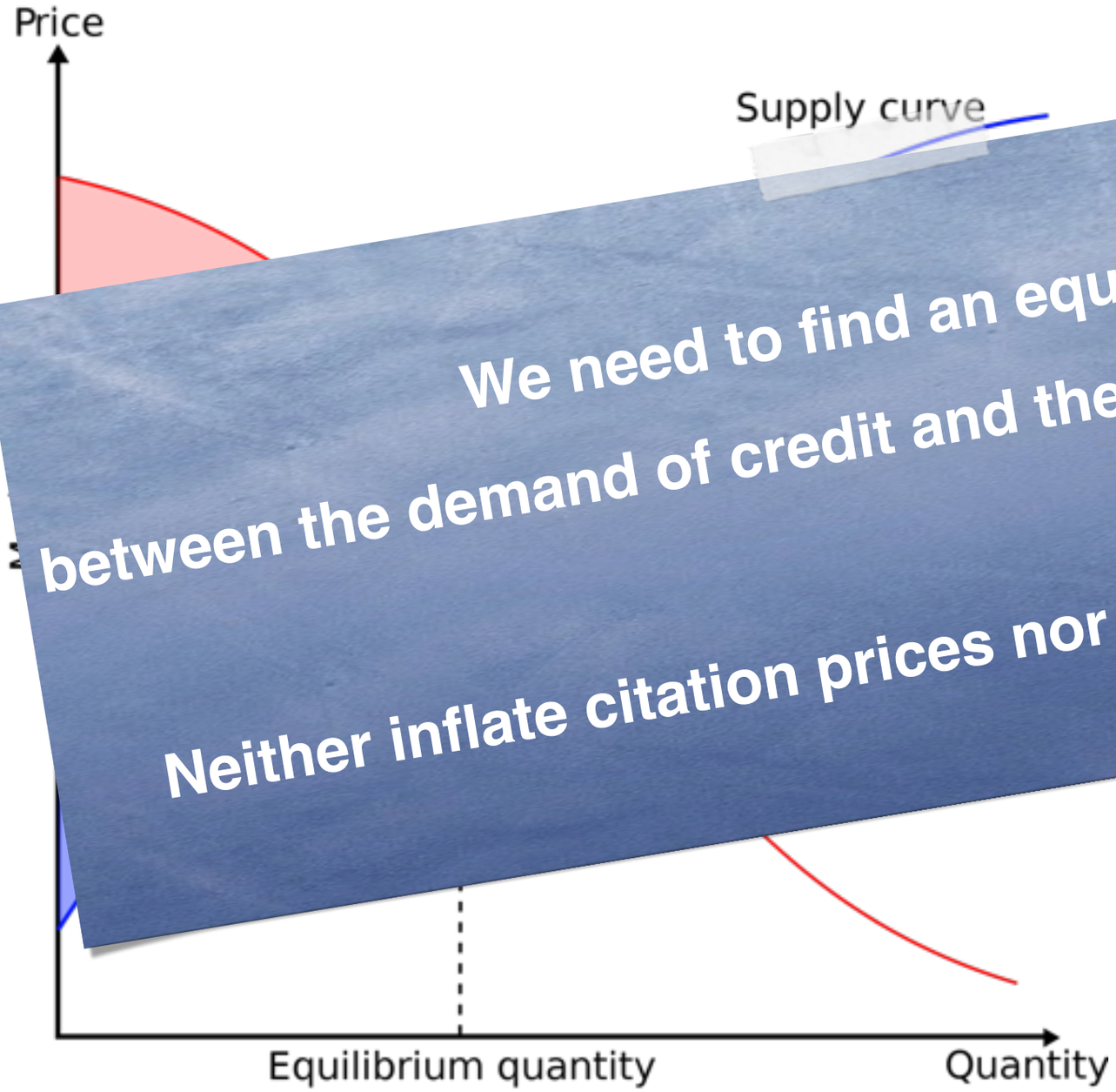
Citations are the currency of the system of sciences



A viable economic assumption is: credit is a finite yet divisible resource, and the higher the credit provided by a citation, the higher the price of the citation

An Economic Perspective

Citations are the currency of the system of sciences



We need to find an equilibrium between the demand of credit and the production of citations

Neither inflate citation prices nor discount contribution

ic

t

nd

y a citation,

the higher the price of the citation

Future

New impact indicators based on reliable and fair data citation mechanisms

Datametrics