Mahidol University

Rensselaer

why not change the world?℠

TRENDS IN COMPUTER SCIENCE RESEARCH*

Present by

Apirak Hoonlor

Faculty of ICT, Mahidol University

at

E-Leader 2014, Bangkok

*Apirak Hoonlor, Boleslaw K. Szymanski, and Mohammed J. Zaki, Communications of the ACM, 56(10): 74-83, 2013

Introduction

- Computer science is an expanding research field.
- The number of research paper published increased over the past two decades.
- Trend analysis

20 TECH TRENDS FOR 2013

www.frogdesign.com

GET MORE PHYSICAL

MATTEO PENZO
TECHNOLOGY DIRECTOR, MILAN

In 2013, the combination of 20-nanometer processors (ARM, Intel, and Apple are planning launches for Q2/Q3) and 4G Networks becoming available in most countries will alter how we use our smart phones.

Higher computational power, reduced energy consumption, and faster data communication in our hands will accelerate the development of biometric applications, such as the authentication of the eye or fingerprints through a hand-held device's camera. This will play a big role in sensitive applications such as mobile banking or payments. Pairing biometric authentication with voice-based logins will start becoming the norm, granting us faster and more secure access to information. As a result, private databases storing bio-information will arise, fueling start-up and funding action in this area.

We can look forward to a time when the authentication layer won't be based on our human memory anymore. In 2013, we'll move closer to a time when we won't be forced to rely on easily forgettable (and not very secure) pass-

HUMAN-COMPUTER INTERACTION GETS MORE HUMANISTIC

MARK ROLSTON
CHIEF CREATIVE OFFICER, AUSTIN

WE LOSE CONTROL OF OUR CARS

KATIE DILL
CREATIVE DIRECTOR, SAN FRANCISCO

Our cars are becoming ever more automated. They are parallel parking themselves, monitoring our speed while in cruise control, and now

Frog Design, “20 Tech Trends for 2013”, http://frogdesign.com

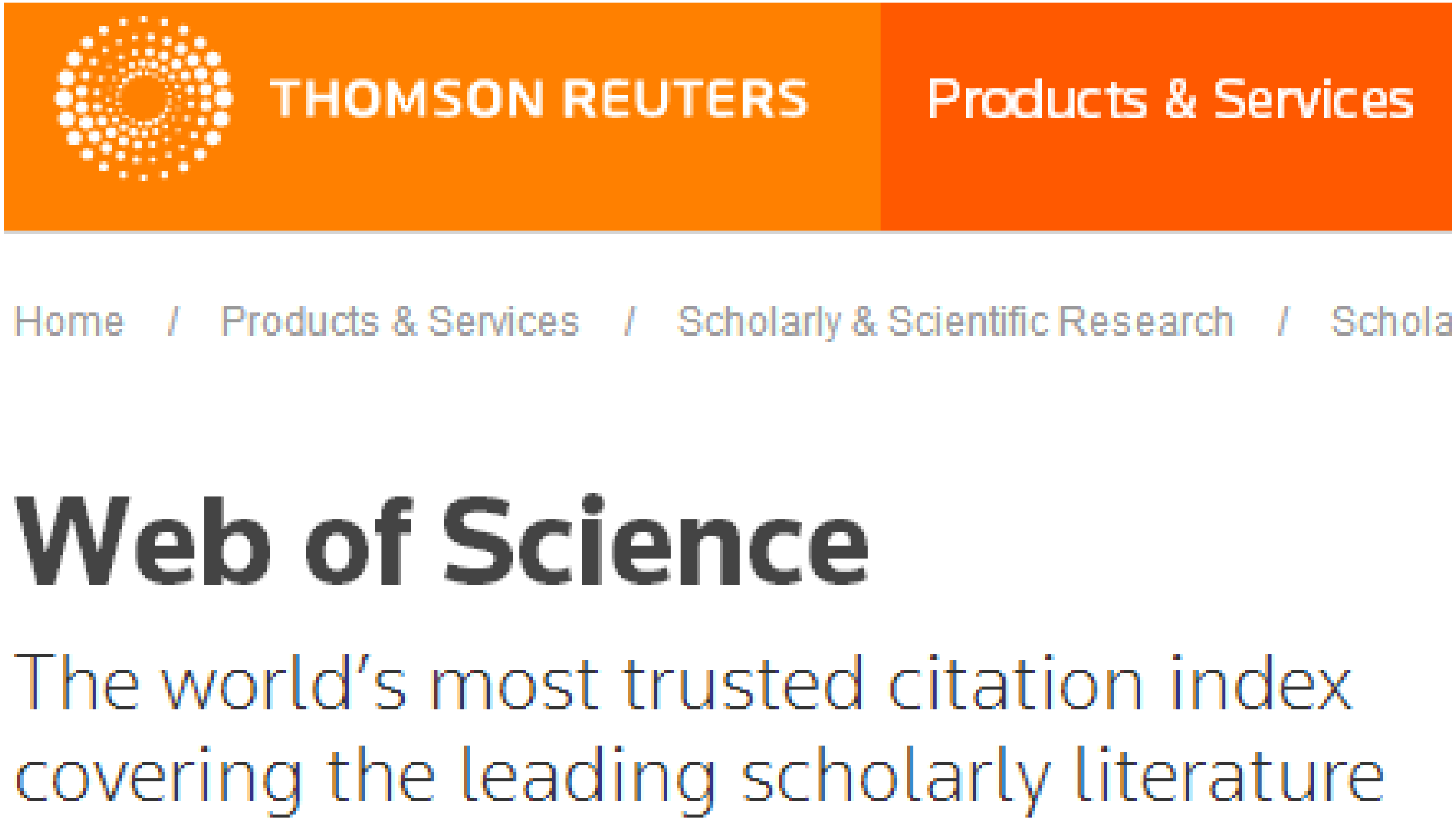
Introduction

Trend Analysis on Science datasets

- Web of Science – Thomson Reuters

Introduction

- Trend Analysis on Science datasets



Thomson Reuters, "Web Of Science", http://thomsonreuters.com/web-of-science/

Introduction

Trend Analysis on Science datasets

- Web of Science – Thomson Reuters
- MapEquation -- Rosvall and Bergstrom

Introduction

- Trend Analysis on Science datasets

“MapEquation”, http://www.mapequation.org [1]

Introduction

Trend Analysis on Science datasets

- Web of Science – Thomson Reuters
- MapEquation -- Rosvall and Bergstrom
- Other works
  - "Is science becoming more interdisciplinary?" – Porter and Rafols
  - "Gender and computing in conference papers" – Cohoon et al.
  - "The Claremont report on database research" – Agrawal et al.

Introduction

“What drives a research topic and turns it into a trend?”

$$$

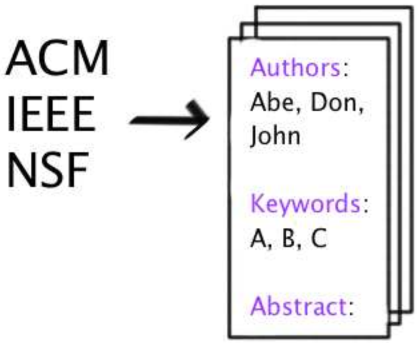
Analysis

Datasets: Collected from 1990 - 2010

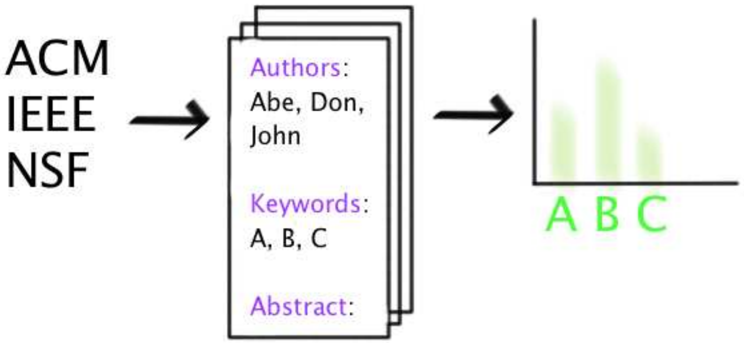
- ACM Dataset: ACM Digital Library
- IEEE Dataset: IEEE Xplore Digital Library
- NSF Dataset: Publicly available awarded grants from www.nsf.gov

Images were taken respectively from ACM Digital Library, IEEE Xplore, and NSF sites.

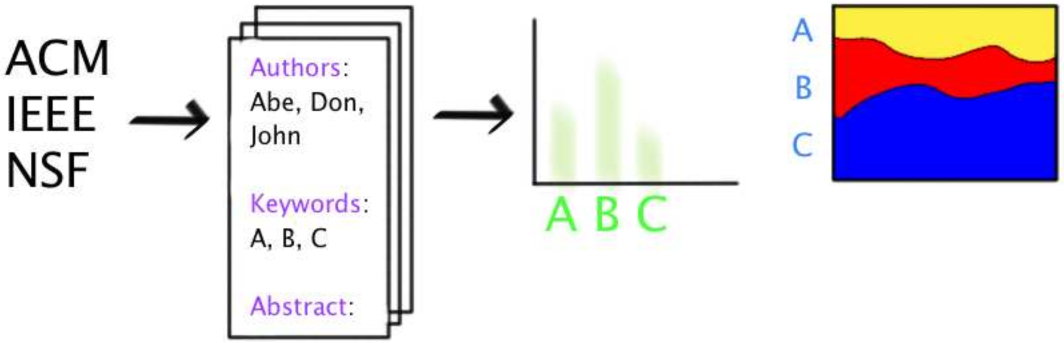
Analysis



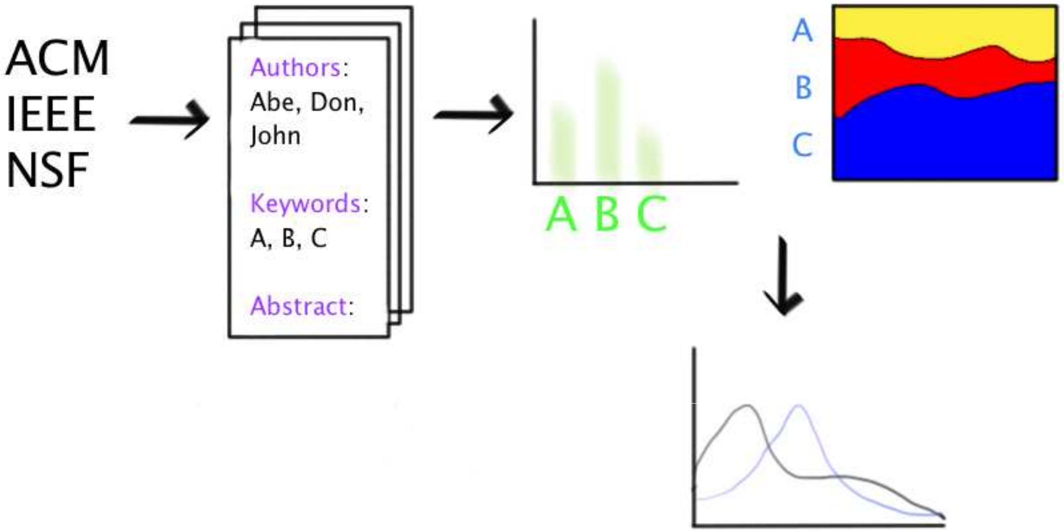
Analysis



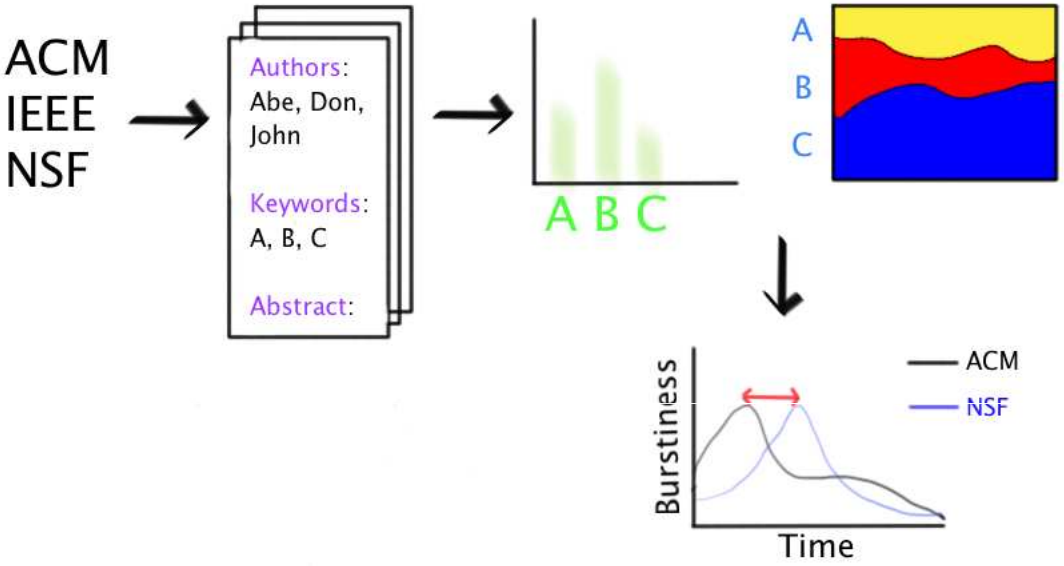
Analysis


ACM
IEEE
NSF
Authors:
Abe, Don,
John
Keywords:
A, B, C
Abstract:
A B C
A
B
C

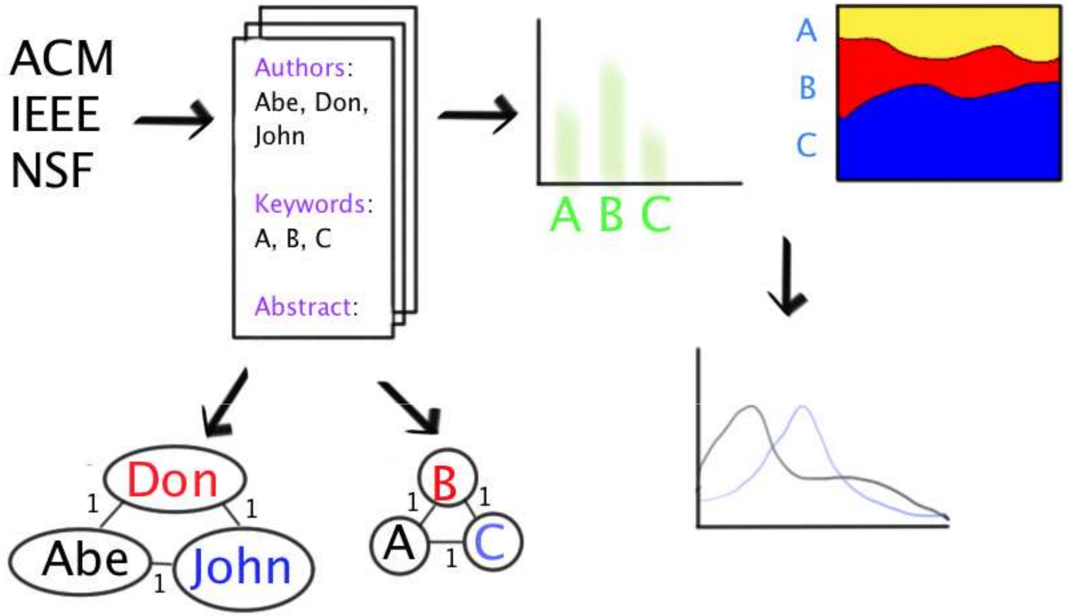
Analysis


ACM
IEEE
NSF
Authors:
Abe, Don,
John
Keywords:
A, B, C
Abstract:
A B C
A
B
C

Analysis

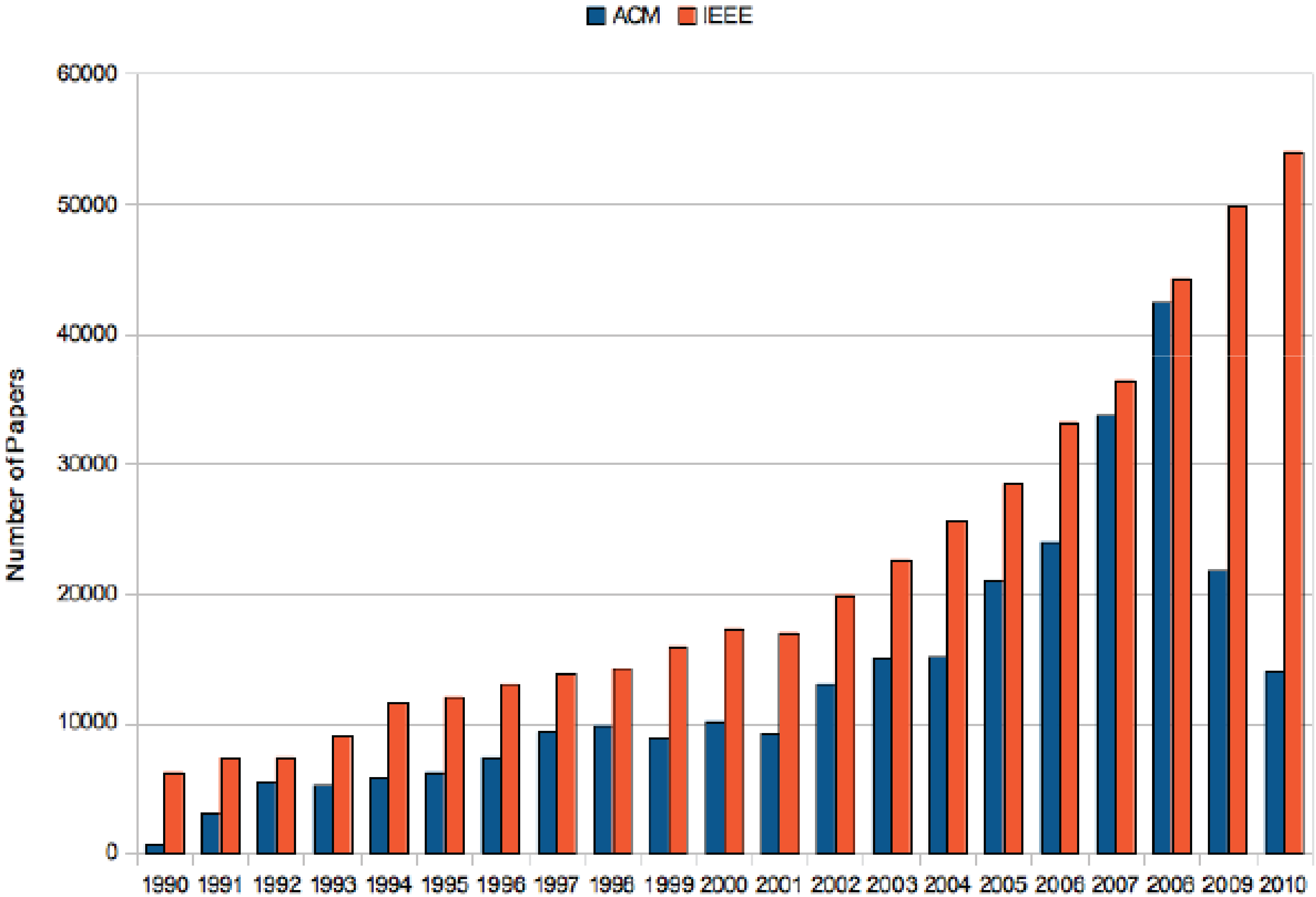

ACM
IEEE
NSF
Authors:
Abe, Don,
John
Keywords:
A, B, C
Abstract:
A B C
A
B
C
Burstiness
Time
ACM
NSF

Analysis

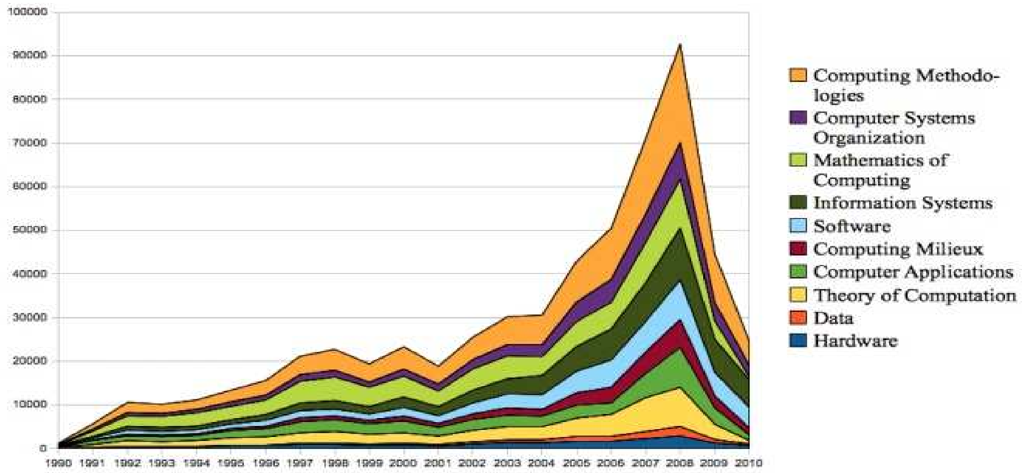

ACM
IEEE
NSF
Authors:
Abe, Don,
John
Keywords:
A, B, C
Abstract:
A B C
A
B
C
Don
Abe
John
1
1
1
B
A
C
1
1
1

Result

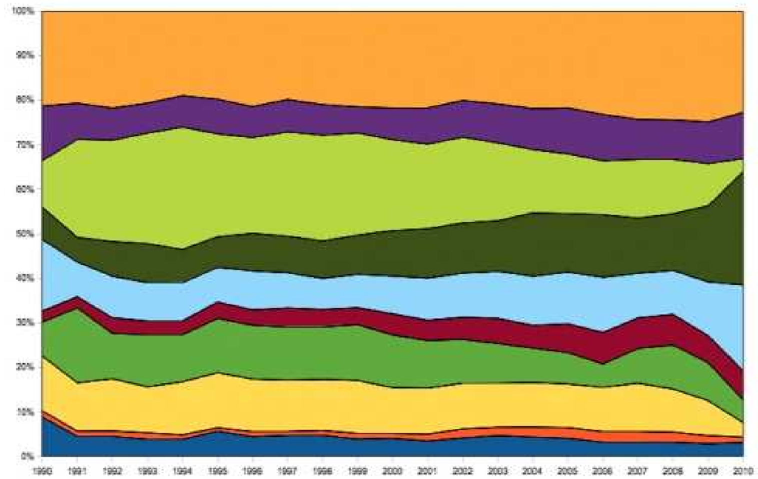
Datasets



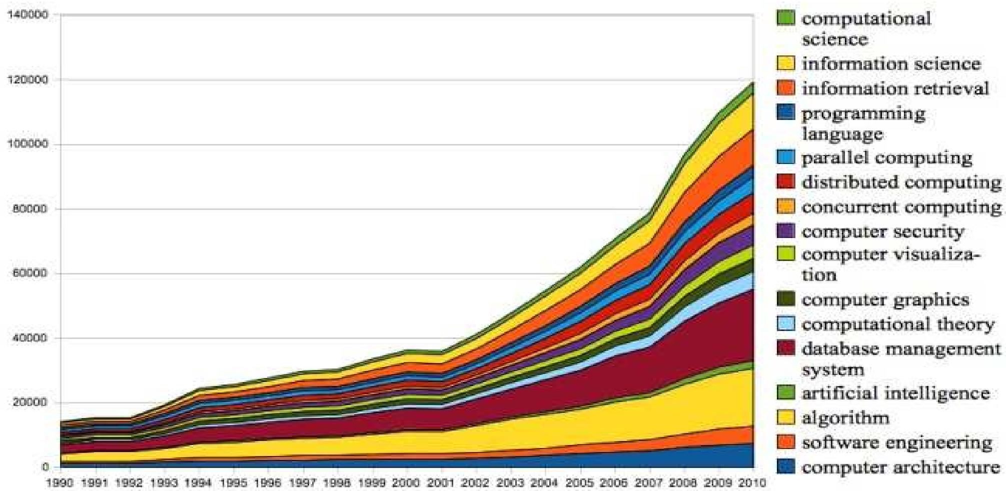
Landscapes of Computer Science research



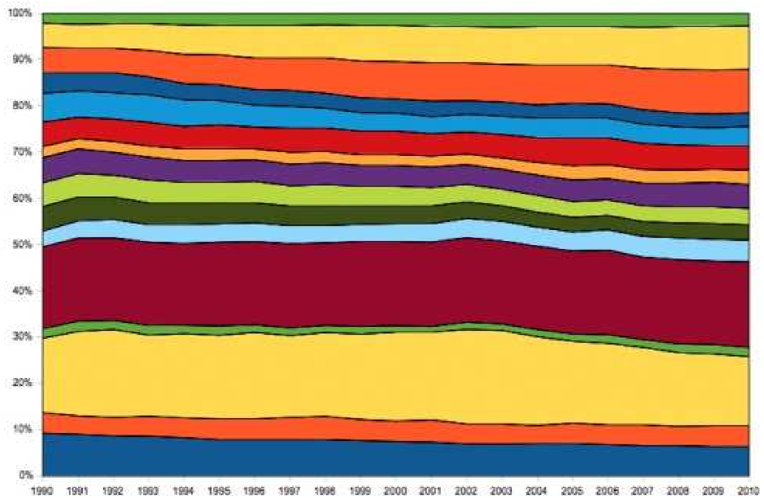
(a) ACM: Frequency



(b) ACM: Fraction



(c) IEEE: Frequency



(d) IEEE: Fraction

Landscapes of Computer Science research

Top rankings

DF – Document Frequency

TFIDF - Term frequency inverse document frequency

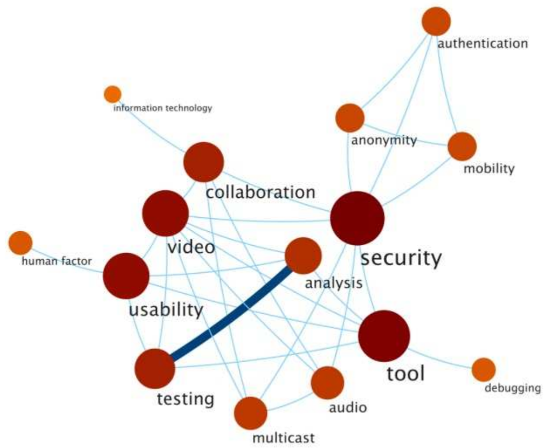
$$\frac{n_{k,d}}{\sum_{w \in d} n_{w,d}} \cdot \log \frac{|D|}{|j : k \in d_j|}$$

where $|D|$ is the number of documents and $n_{k,d}$ is the number of times k appears in d

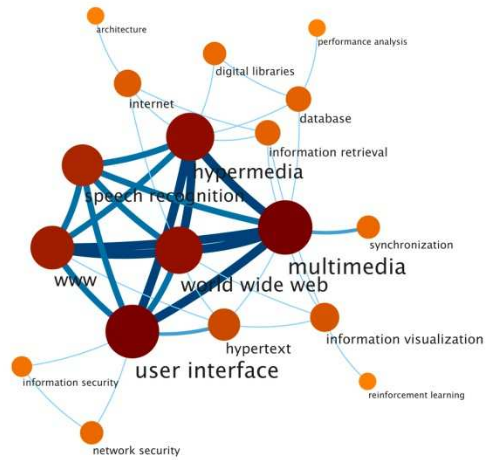
Interesting findings

- Most publications in collaboration, data mining, information retrieval, machine learning, privacy, and XML appeared 2000–2010
- Internet and World Wide Web did not appear in any publication until 1995
  - NSFNET
  - Net
  - Point-to-point
  - Internetworking
  - TCP/IP

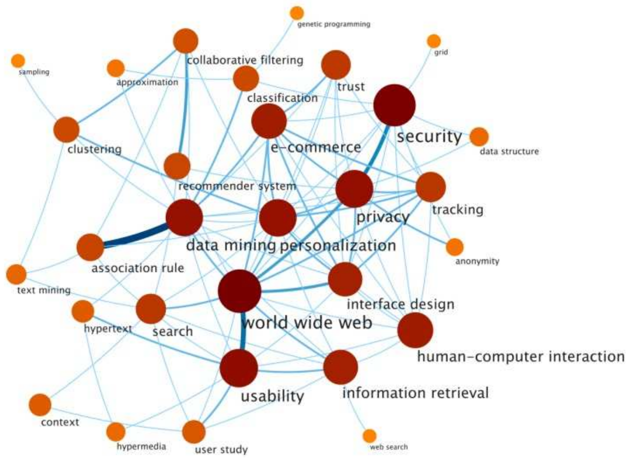
Networks of Computer Science Research



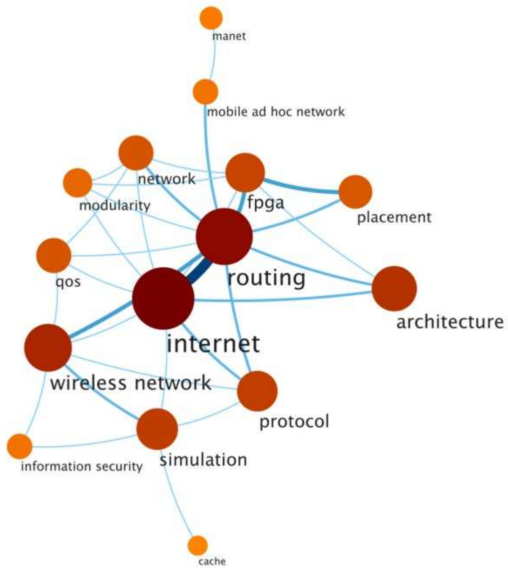
(a) Security Cluster: 1995



(b) Multimedia Cluster: 1995



(c) World Wide Web Cluster: 2001



d) Internet Cluster: 2001

Bursty-period Analysis

The beginning point

- Used to evaluate the influence of research funding on publications and vice versa
- "Bursty keywords" := The keywords appearing with uncommonly high frequency during some intervals as defined below

$$Burst(w,t) = \left(\frac{|d_t : w \in d_t|}{|d : w \in d|} - \frac{1}{T} \right)$$

where w is the keyword/topic of interest, t is a time period, d_t is a document created during time t, d is any document, and T is the total time over which all documents were created.

Bursty-period Analysis

ACM vs. NSF

- If a keyword became bursty in ACM data first, it became bursty in NSF 2.4 years later on average
- In the reverse case, the average delay was 4.8 years
- 75% of keyword became bursty in the NSF dataset before it did in the ACM dataset
- 16% of cases, it was the reverse

IEEE vs. NSF

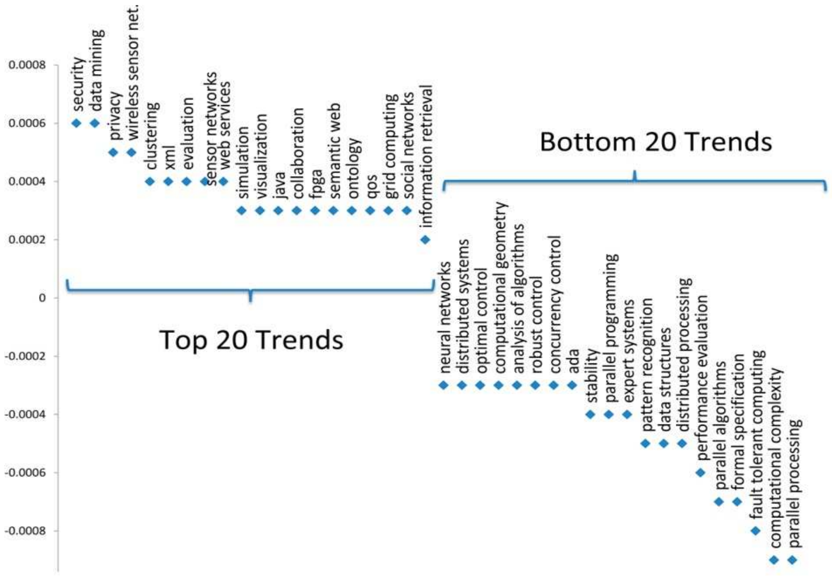
- IEEE became bursty in NSF 3.4 years later on average
- In the reverse case, the average delay was 5.7 years.
- 67% of keyword became bursty in the NSF dataset before it did in the IEEE dataset
- 16% of cases, it was the reverse

Trend analysis

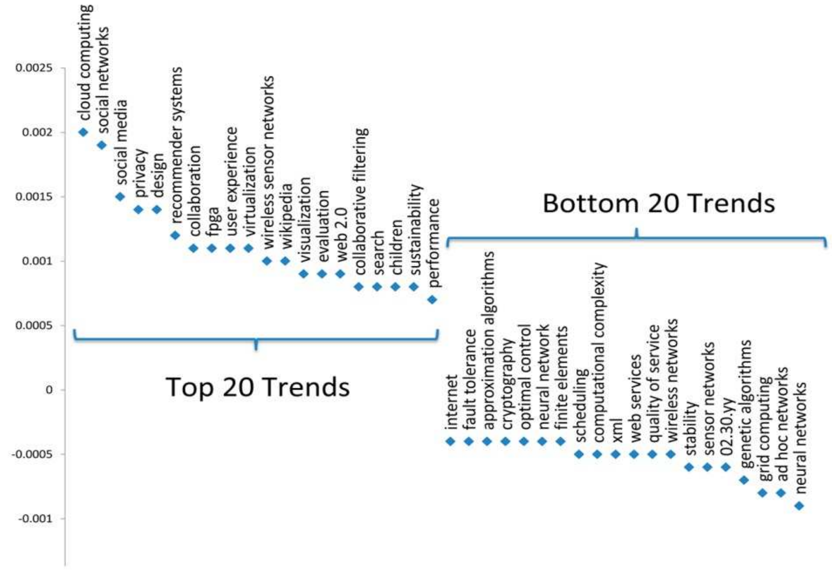
- The linear regression trend line.
- Plot a line based on the fraction of papers to see the trends.
- Look at two trends:

  (i) 1990-2010, and

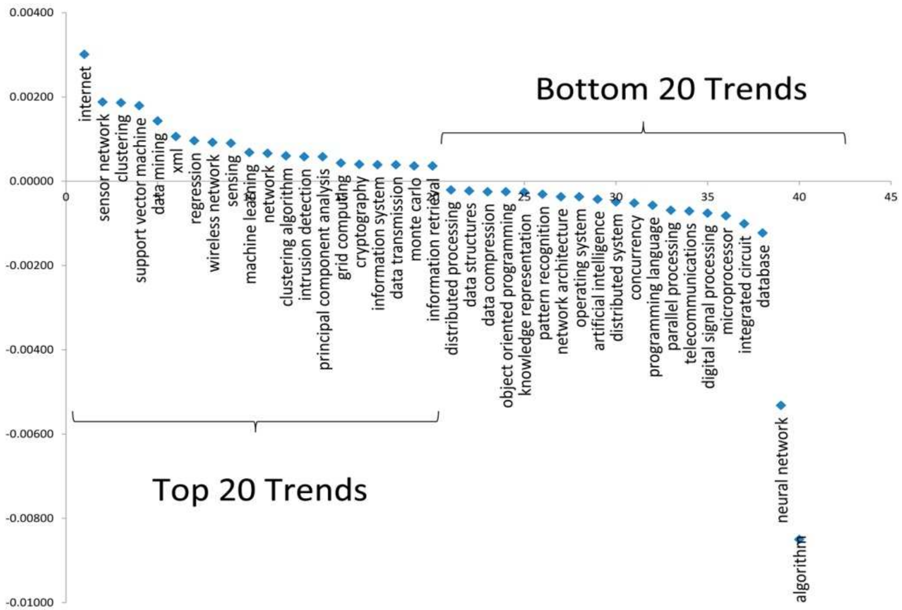
  (ii) 2006 - 2010



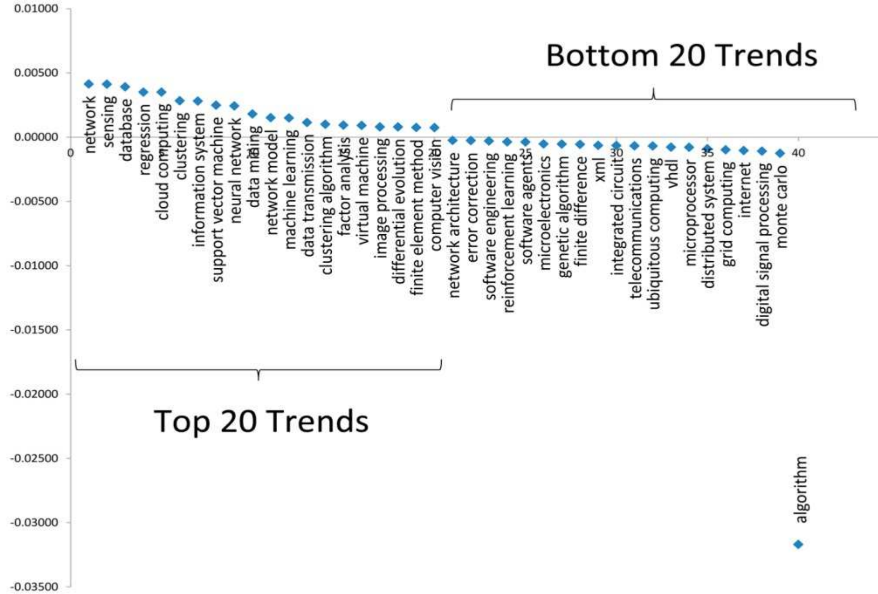
(a) ACM: past 21 years



(b) ACM: past 5 years

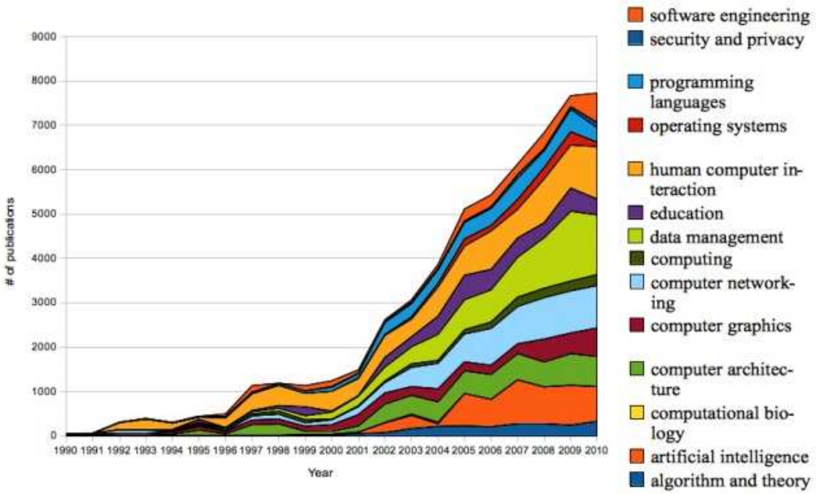


(c) IEEE: past 21 years

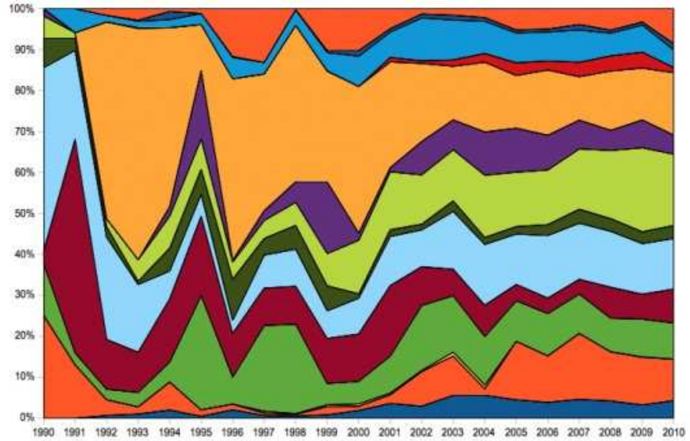


(d) IEEE: past 5 years

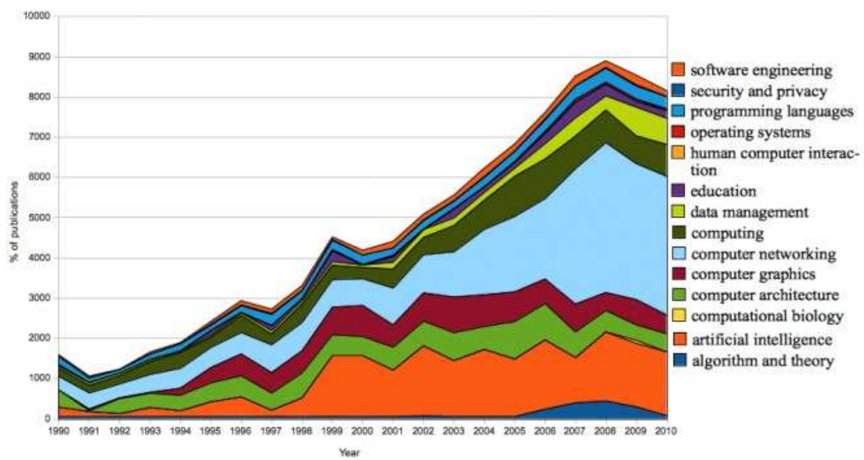
Venue of publications



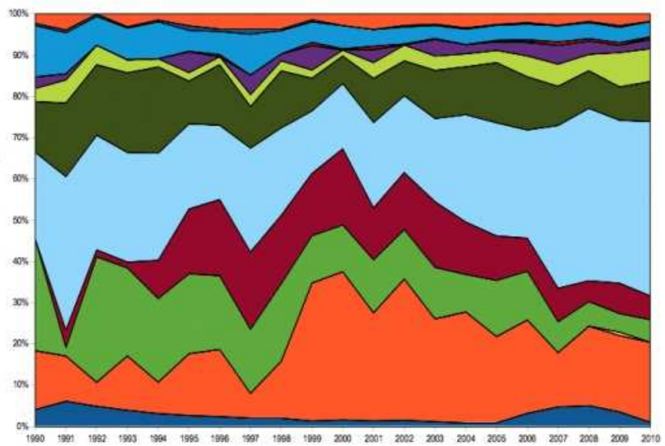
(a) ACM: Frequency



(b) ACM: Fraction



(c) IEEE: Frequency



(d) IEEE: Fraction

CS Researchers

How often do you publish in a certain area?

- Find the sequences of publications in the same major research category by the same author
- Frequent sequence mining
  - cSpade sequence mining algorithm
  - A one-year gap in publication dates
  - Min. Freq. at least 1% of documents
- Look at the half-life
  - How long does it take for the # of publication by the authors in certain topic to reduce by half.

CS Researchers

ACM Dataset

CCS	1-year gap			2-year gap		
	1st DR	$T_{\frac{1}{2}}$	Max. CL	1st DR	$T_{\frac{1}{2}}$	Max. CL
hardware	66%	0.94	5	59%	1.28	7
comp. sys. organization	54%	1.22	8	46%	1.49	9
software	52%	1.15	7	43%	1.47	9
data	81%	0.48	3	75%	0.59	3
theory of computation	60%	0.90	6	50%	1.27	8
mathematics of computing	51%	1.06	7	41%	1.58	10
information systems	48%	1.32	8	40%	1.70	11
computing methodologies	41%	1.26	8	32%	1.66	11
computer applications	72%	0.61	4	63%	0.83	5
computing milieu	68%	0.78	5	59%	0.99	6

CS Researchers

IEEE Dataset

Category	1-year gap			2-year gap		
	1st DR	$T_{\frac{1}{2}}$	Max. CL	1st DR	$T_{\frac{1}{2}}$	Max. CL
alg. and theory	70%	0.58	2	63%	0.83	3
programming language	83%	0.39	2	76%	0.55	3
computing	65%	0.86	5	57%	1.20	7
soft. eng.	82%	0.41	2	75%	0.50	2
operating systems	100%	N/A	1	100%	N/A	1
comp. arch	63%	0.95	6	54%	1.37	8
computer networking	48%	1.11	7	39%	1.47	9
security and privacy	N/A	N/A	N/A	N/A	N/A	N/A
data management	72%	0.65	3	65%	0.92	4
artificial intelligence	58%	0.88	5	47%	1.32	8
computer graphics	63%	0.89	5	57%	1.20	7
HCI	N/A	N/A	N/A	N/A	N/A	N/A

CS Researchers

Putting things in perspective

Why is there a significant drops?

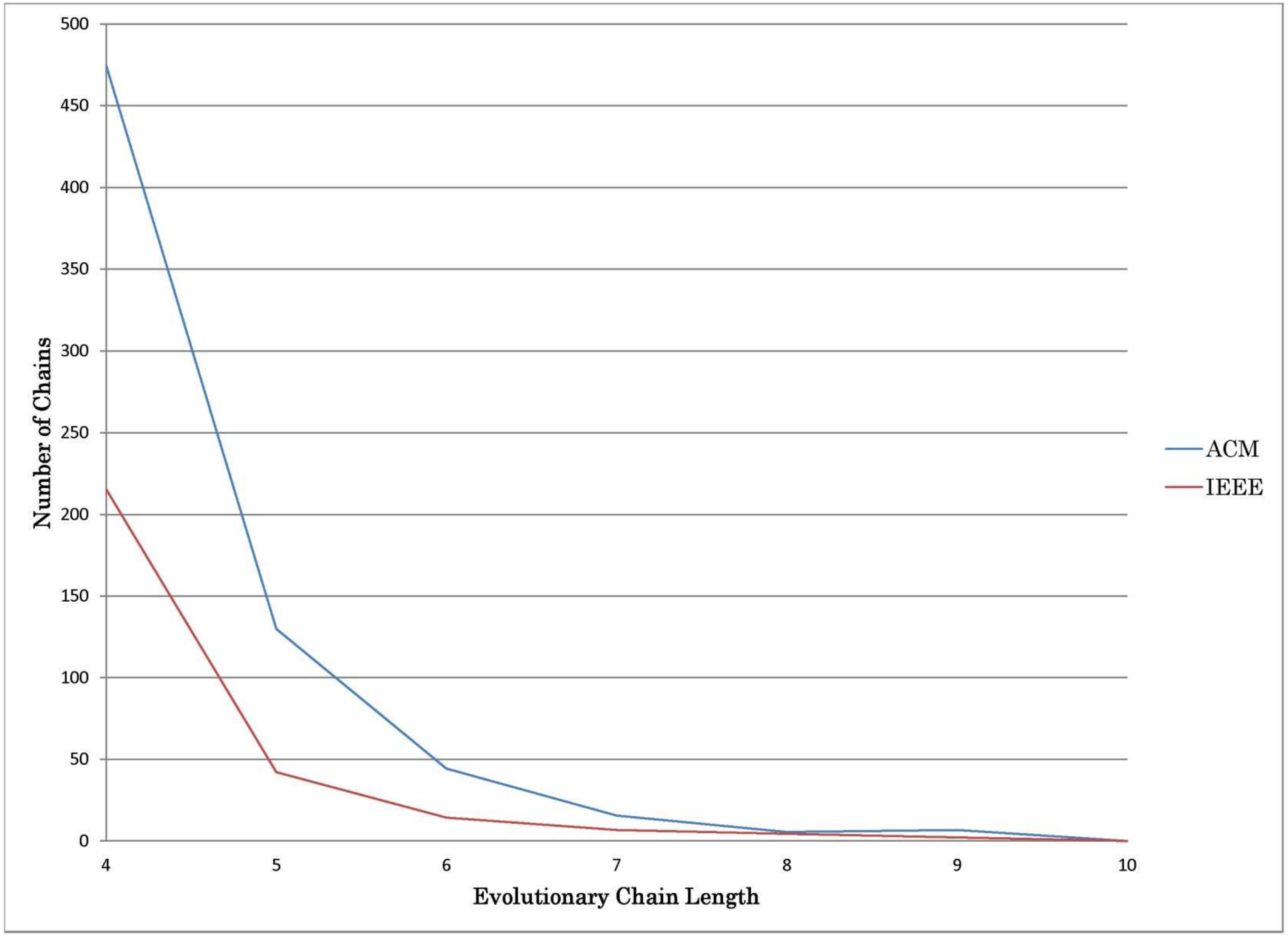
Why can’t one person publish in any given topics?

CS Researchers

Putting things in perspective

- Jack Dongarra – University of Tennessee, Knoxville
  - High-performance linear algebra software
  - His research interests have evolved continuously in response to challenges created by new computer technologies
- George Cybenko – Dartmouth College, USA
  - High-performance computing and classification by neural networks
  - Investigate each subject “in five-year (more or less) phases” then “discovers an open field often related to previous work.” One exception was a major shift in 1992 when moving from one university to another.

Communities of CS researchers



Conclusion

Key findings

- CS continues to experience continuous and fundamental transformation;
- A burst of new keywords in grants generally precedes their burst in publications; less than 1/3 of new keywords burst in publications first, reflecting the importance of funding for success of new CS fields.
- A typical scientist's research focus changes in roughly a 10-year cycle and often includes a once-in-a-career dramatic shift, likely in response to evolving technology creating new CS fields.

Key References

[1] Edler, D. and M. Rosvall, M. Map Generator software package. MapEquation, Umeå, Sweden, 2010; http://www.mapequation.org

[2] Goldberg, M., Magdon-Ismail, M., Nambirajan, S., and Thompson, J. Tracking and predicting evolution of social communities. In *Proceedings of the Third IEEE International Conference on Social Computing* (Boston, Oct. 9–11), IEEE, 2011, 780–783.

[3] Hoonlor, A., Szymanski, B.K., Zaki, M.J., and Thompson, J. *An Evolution of Computer Science Research*. RPI Technical Report 12-01, Rensselaer Polytechnic Institute, Troy, NY, 2012; http://www.cs.rpi.edu/research/tr.html

Key References

[4] Lappas, T. et al. On Burstiness-Aware Search for Document Sequences. In *Proceedings of the 15th ACM SIGKDD* (Paris, France, June 28–July 1). ACM Press, New York, 2009, 477-486.

[5] Zaki, M.J. Sequences mining in categorical domains: Incorporating constraints. In *Proceedings of the Ninth ACM Int. Conf. on Information and Knowledge Management* (Washington D.C., Nov. 6–11).

Thank You for Your Attention

Any Question?