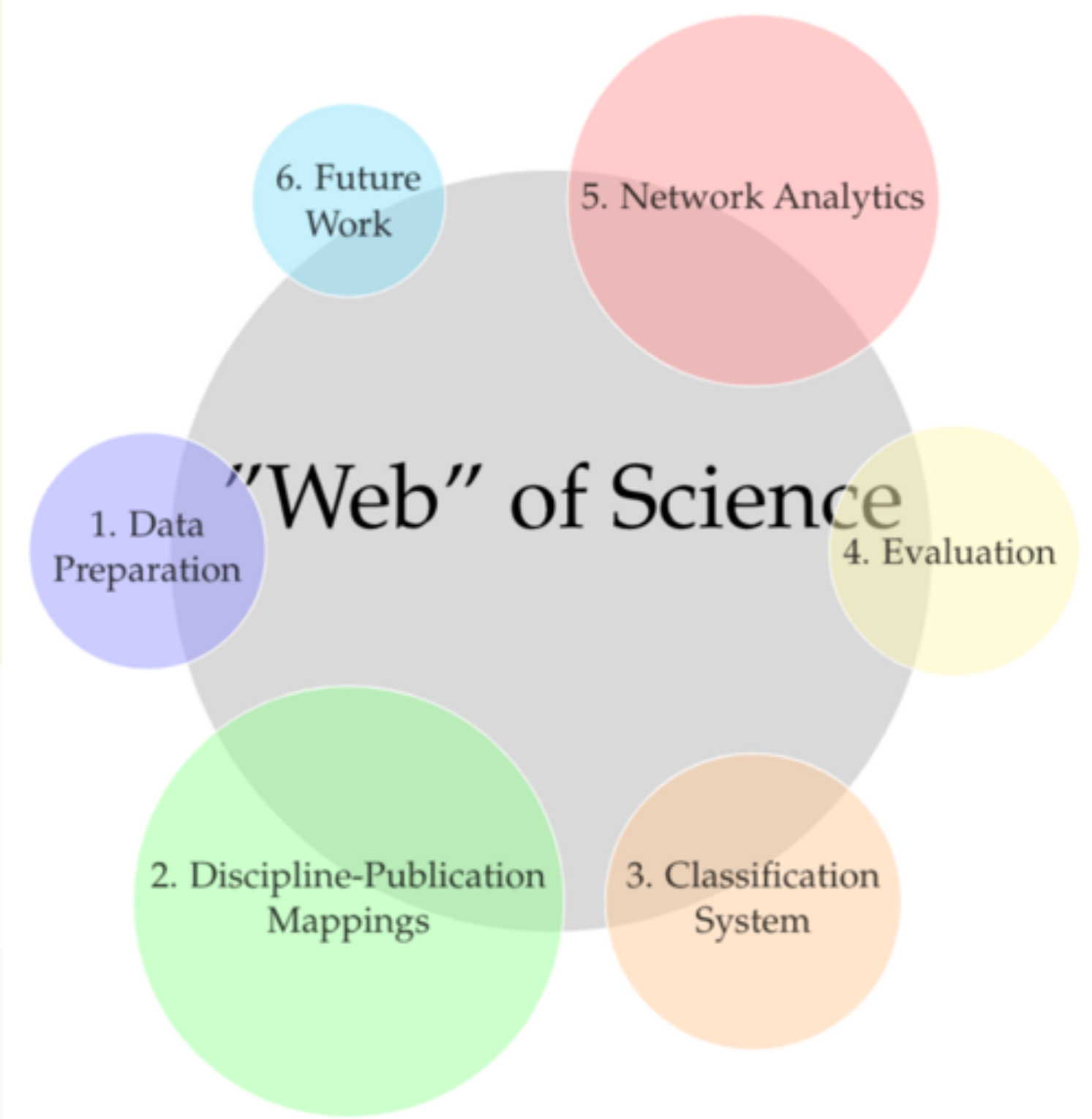
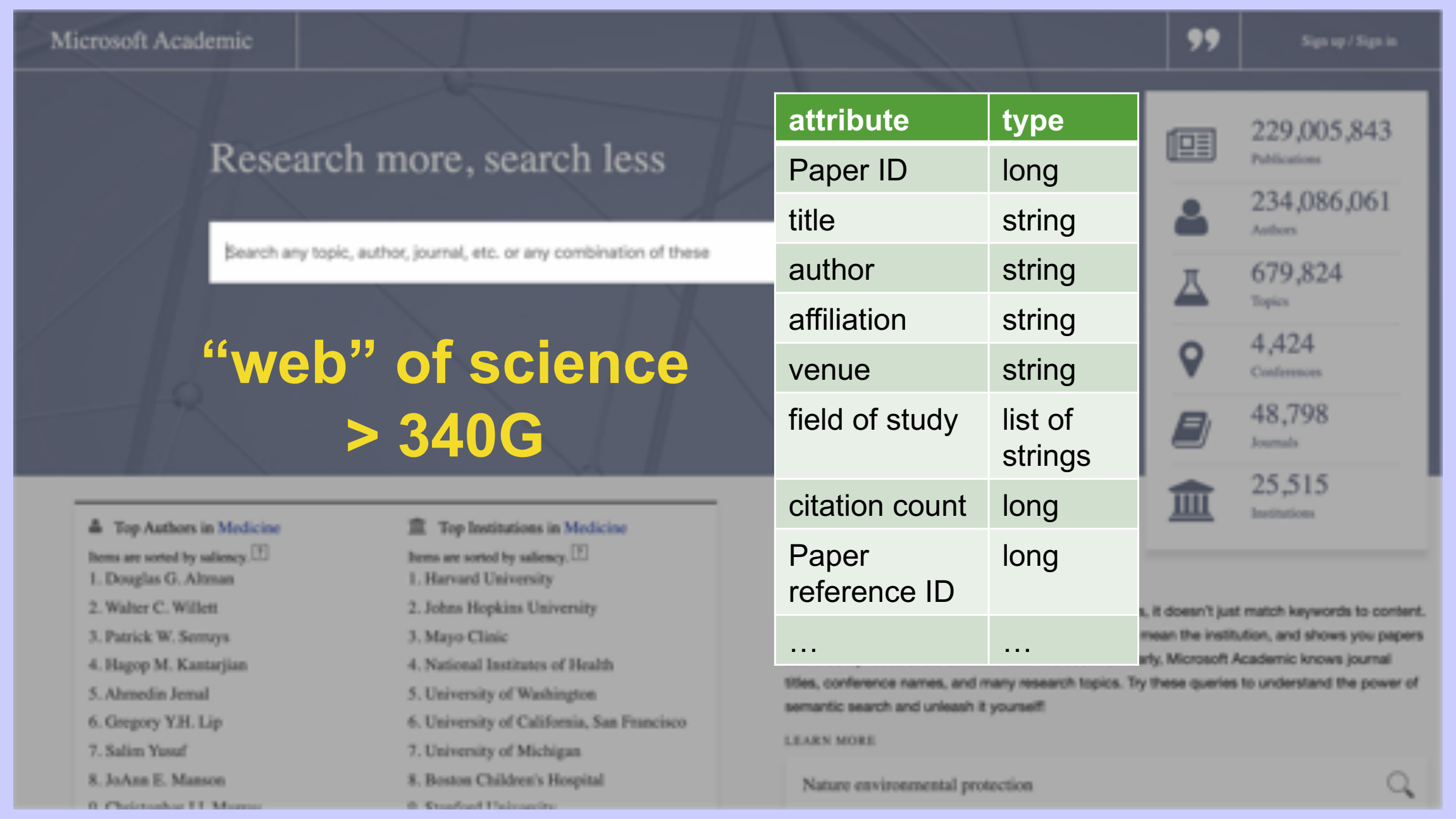
Understanding the Web of Science Using Deep Learning

Susie Xi Rao[1, 2], Ce Zhang[1], Peter H. Egger[2]
[1] DS3Lab, ETH Zurich; [2] KOF Swiss Economic Institute
Systems Group Retreat 2020
rao@kof.ethz.ch



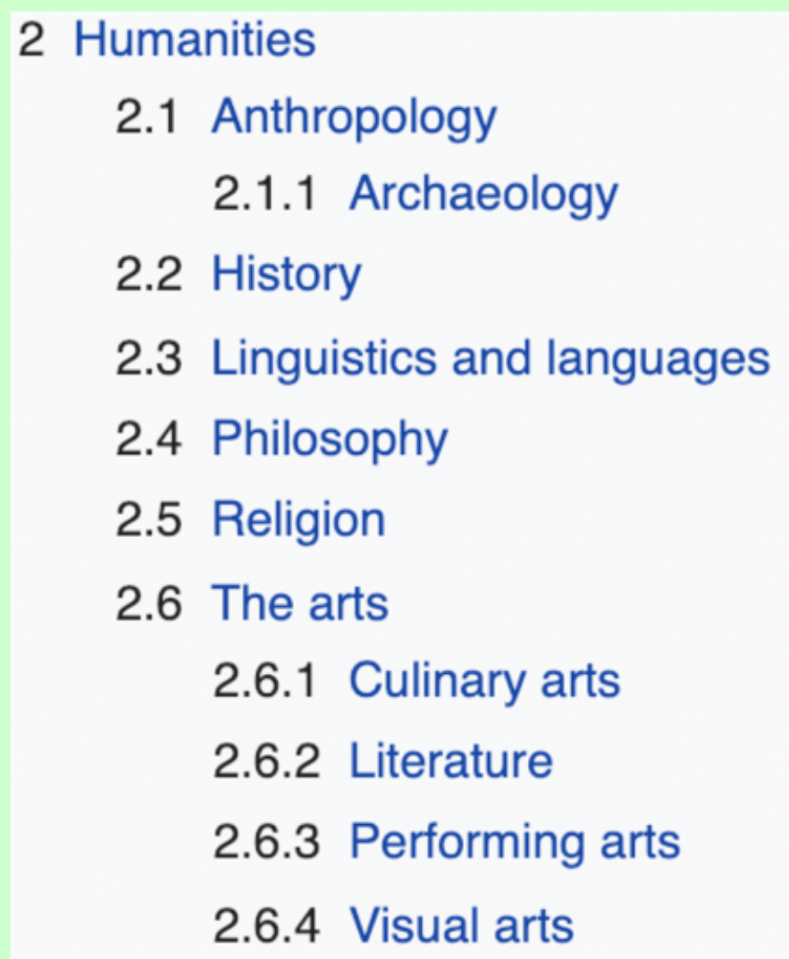
Research Goal:
1) A system that takes a growing amount of scholarly publications from each discipline and tells you to which discipline it belongs to.
2) Creation of networks that depict the publication behaviors of authors and institutions and how they impact the innovations and economic developments around the globe.



attribute	type
Paper ID	long
title	string
author	string
affiliation	string
venue	string
field of study	list of strings
citation count	long
Paper reference ID	long
…	…

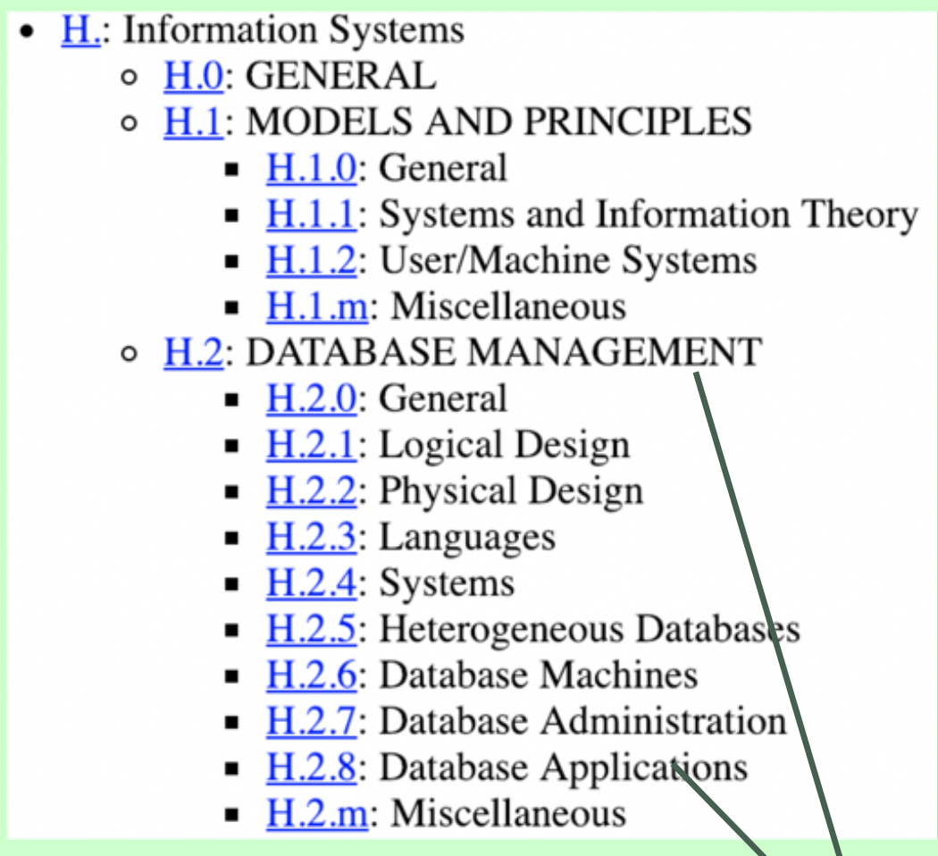
1 Data Preparation (X)

Microsoft Academic Service, frequent update, topics on different granularity for each discipline



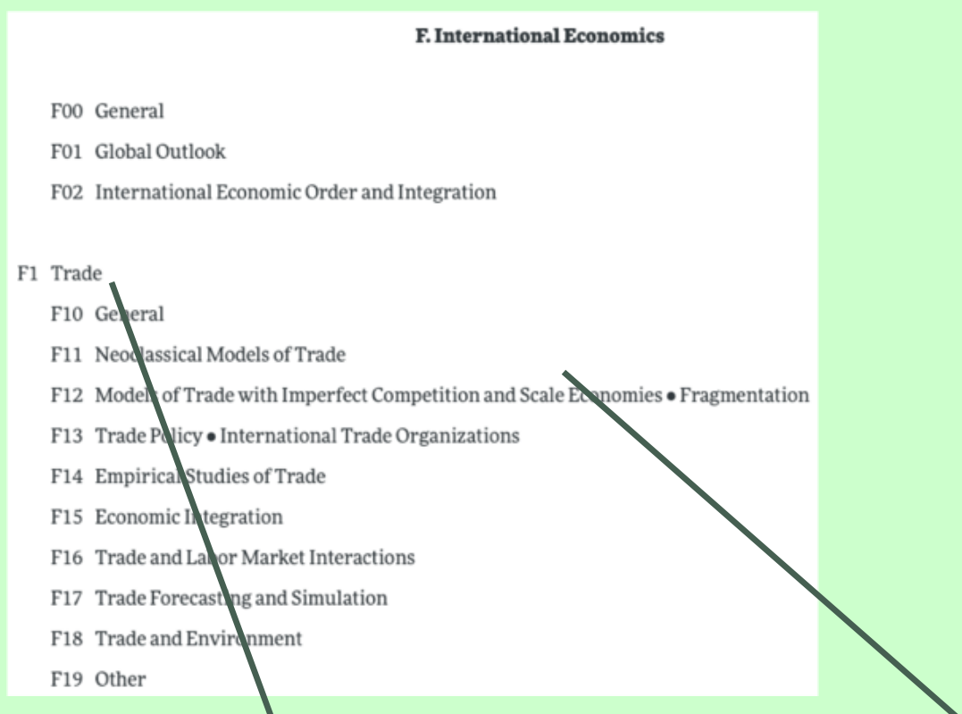
2 Humanities
- 2.1 Anthropology
  - 2.1.1 Archaeology
- 2.2 History
- 2.3 Linguistics and languages
- 2.4 Philosophy
- 2.5 Religion
- 2.6 The arts
  - 2.6.1 Culinary arts
  - 2.6.2 Literature
  - 2.6.3 Performing arts
  - 2.6.4 Visual arts

56 discipline by Wikipedia



- H.: Information Systems
  - H.0: GENERAL
  - H.1: MODELS AND PRINCIPLES
    - H.1.0: General
    - H.1.1: Systems and Information Theory
    - H.1.2: User/Machine Systems
    - H.1.m: Miscellaneous
  - H.2: DATABASE MANAGEMENT
    - H.2.0: General
    - H.2.1: Logical Design
    - H.2.2: Physical Design
    - H.2.3: Languages
    - H.2.4: Systems
    - H.2.5: Heterogeneous Databases
    - H.2.6: Database Machines
    - H.2.7: Database Administration
    - H.2.8: Database Applications
    - H.2.m: Miscellaneous

ACM classification for CS

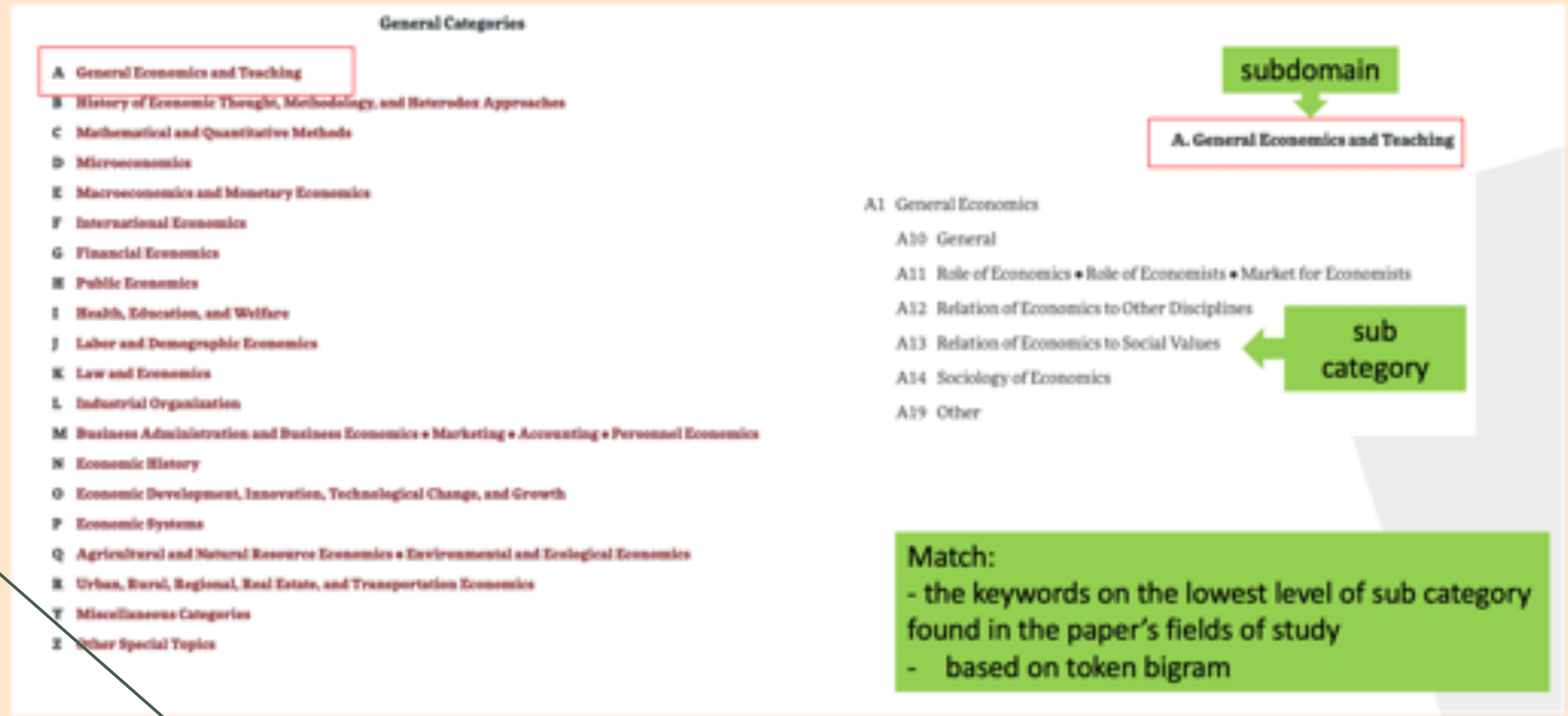


F. International Economics

F00	General
F01	Global Outlook
F02	International Economic Order and Integration
F1	Trade
F10	General
F11	Neoclassical Models of Trade
F12	Models of Trade with Imperfect Competition and Scale Economies • Fragmentation
F13	Trade Policy • International Trade Organizations
F14	Empirical Studies of Trade
F15	Economic Integration
F16	Trade and Labor Market Interactions
F17	Trade Forecasting and Simulation
F18	Trade and Environment
F19	Other

JEL classification for econ

2 Discipline-Publication Mappings (YL1 – YL2 – YL3)



General Categories

A	General Economics and Teaching
B	History of Economic Thought, Methodology, and Heterodox Approaches
C	Mathematical and Quantitative Methods
D	Microeconomics
E	Macroeconomics and Monetary Economics
F	International Economics
G	Financial Economics
H	Public Economics
I	Health, Education, and Welfare
J	Labor and Demographic Economics
K	Law and Economics
L	Industrial Organization
M	Business Administration and Business Economics • Marketing • Accounting • Personnel Economics
N	Economic History
O	Economic Development, Innovation, Technological Change, and Growth
P	Economic Systems
Q	Agricultural and Natural Resource Economics • Environmental and Ecological Economics
R	Urban, Rural, Regional, Real Estate, and Transportation Economics
Y	Miscellaneous Categories
Z	Other Special Topics

subdomain

A. General Economics and Teaching

A1	General Economics
A10	General
A11	Role of Economics • Role of Economists • Market for Economists
A12	Relation of Economics to Other Disciplines
A13	Relation of Economics to Social Values
A14	Sociology of Economics
A19	Other

sub category

Match:
- the keywords on the lowest level of sub category found in the paper's fields of study
- based on token bigram

3 Classification Systems (Generating training/testing material)

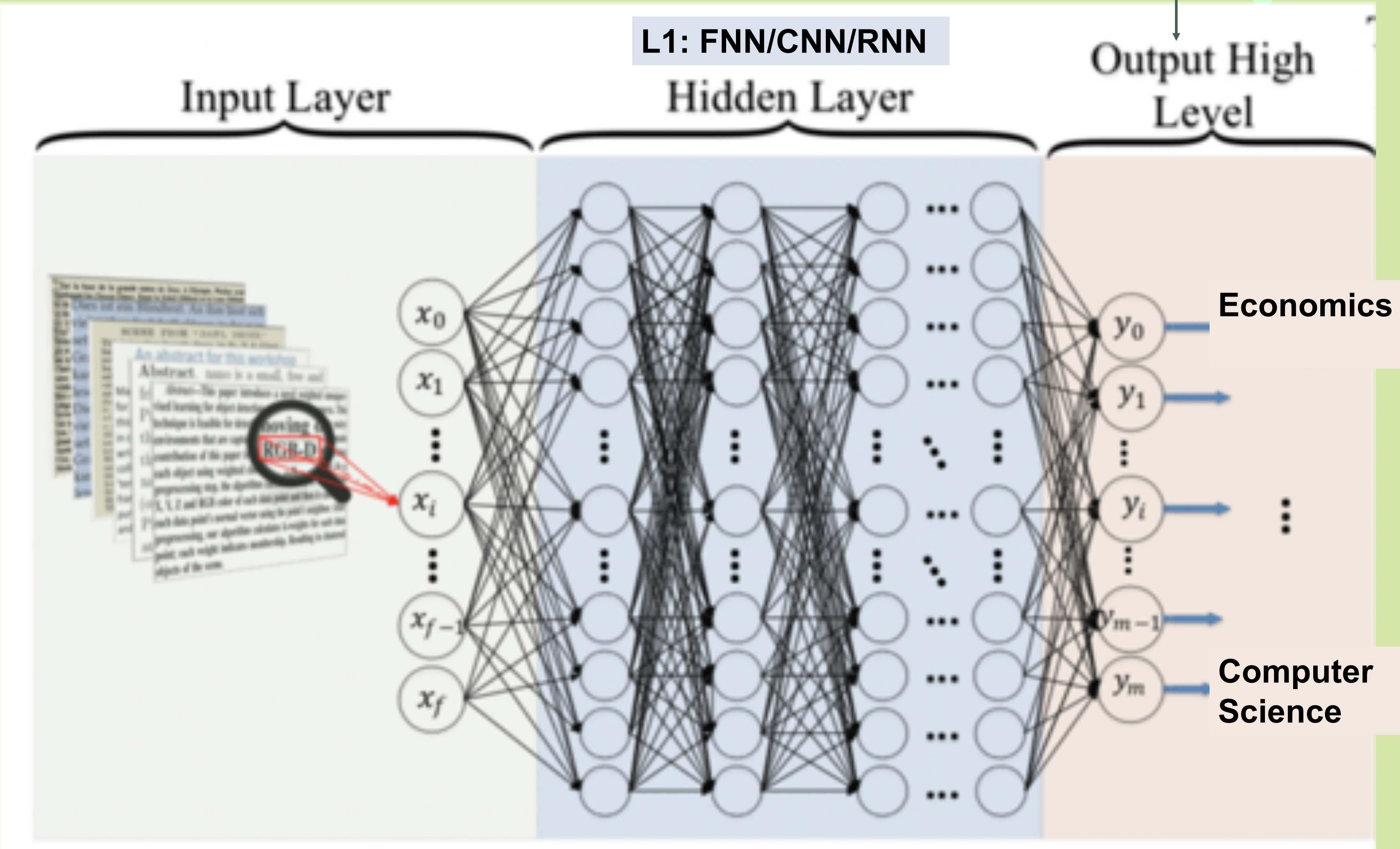
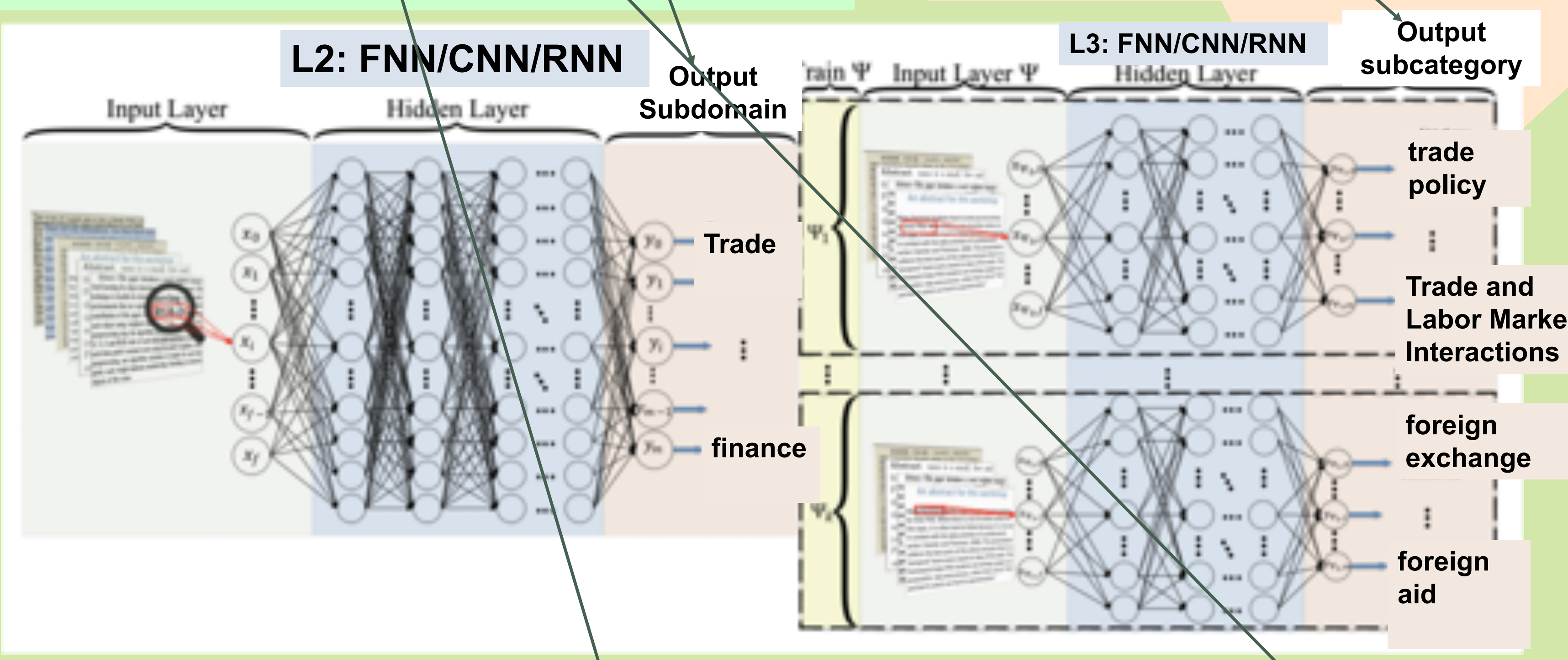
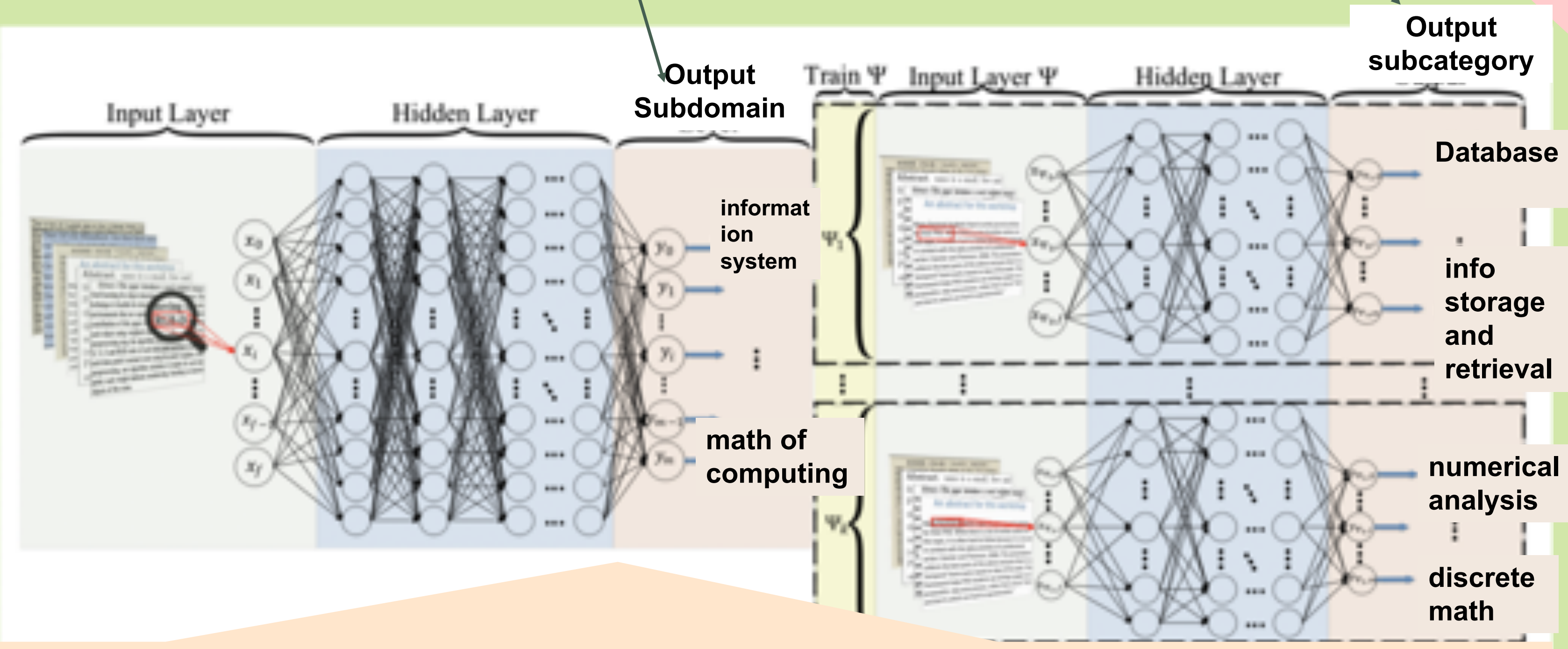


Fig. 1: HDLTex: Hierarchical Deep Learning for Text Classification. This is our Deep Neural Network (DNN) approach for text classification. The left figure depicts the parent-level of our model, and the right figure depicts child-level models defined by Ψ_i as input documents in the parent level.

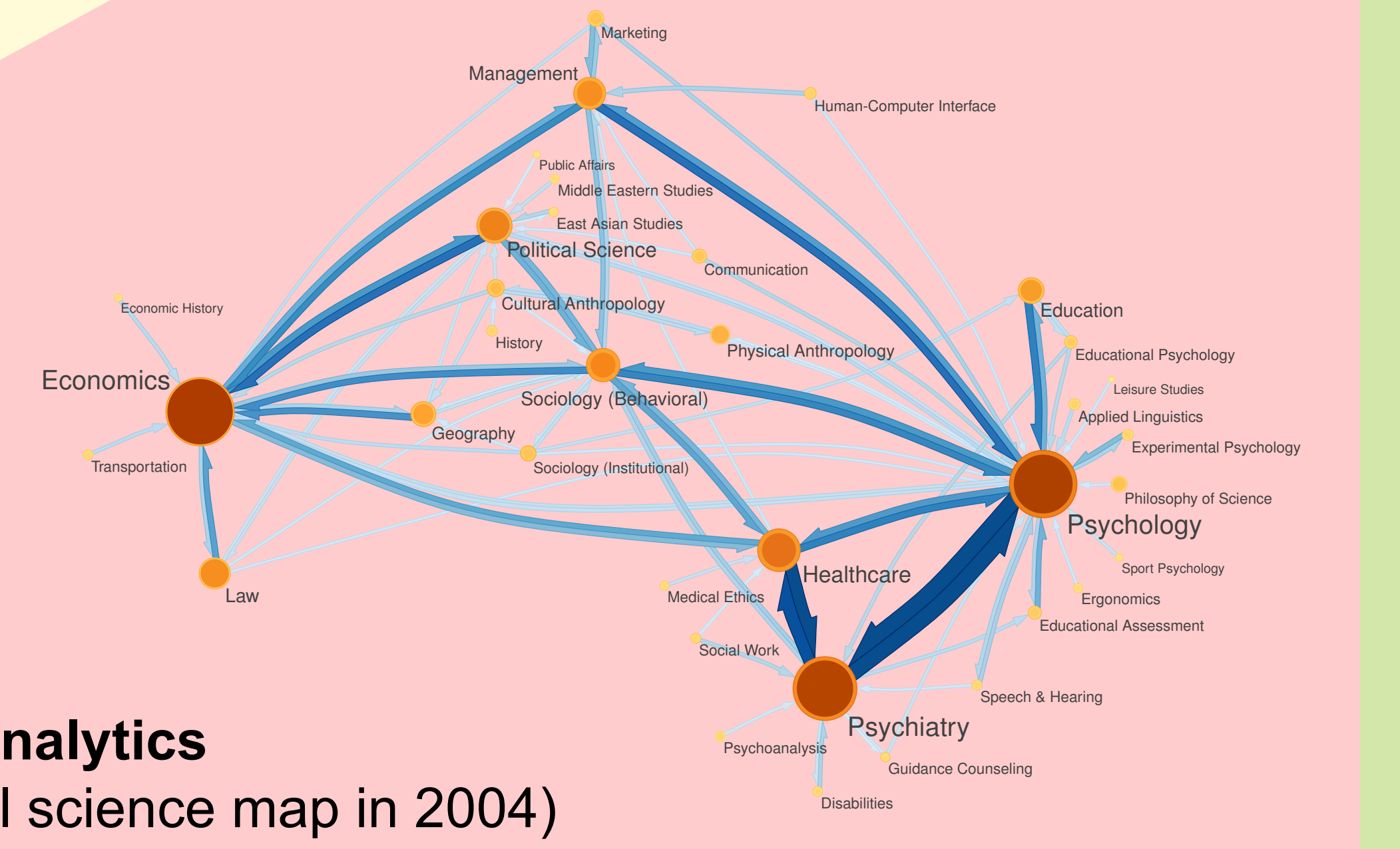
- Input: article abstracts (ca. 300 words)
- Output: three labels (discipline, subdomain, subcategory)





4 Evaluation (architecture search)

- For L1 + L2, econ + CS: RNN + CNN works the best, > 90% accuracies for output subcategories
- For L1 + L2 + L3, econ + CS: CNN + CNN + CNN works the best, to be reported



5 Network Analytics

(e.g., a social science map in 2004)
Other interesting questions:
- Rise and fall of fields: Using our classification, what are the fields that became important sources of spillovers to other fields.
- Rise and fall of institutions
- We could "invent" a taxonomy of break-through innovations?

3 Classification Systems (System optimization)

- Capability to take in large datasets: precomputation of word index for large training corpus using MapReduce (6 min. for 24 mio. abstracts), precomputation of one-hot encodings of the abstracts in each discipline and arbitrary combination of training sets across disciplines are possible, precomputation of training/validation partitions on all the levels.
- Distributed learning using *tensorflow.*
- A hierarchical classification system application to other sorts of hierarchies.
- Softmax function at each classification step renders probabilities of discipline membership given one abstract → to what extent one publication belongs to one discipline/subdomain/subcategory.

References
•[1] KHABSA, M., AND GILES, C. L. The number of scholarly documents on the public web. PloS one 9, 5 (2014), e93949.
•[2] KOWSARI, K., BROWN, D. E., HEIDARYSAFA, M., MEIMANDI, K. J., GERBER, M. S., AND BARNES, L. E. HDLTex: Hierarchical deep learning for text classification. In Machine Learning and Applications (ICMLA), 2017 16th IEEE International Conference on (2017), IEEE, pp. 364–371.
•[3] SHEN, Z., MA, H., AND WANG, K. A web-scale system for scientific knowledge exploration. arXiv preprint arXiv:1805.12216 (2018).
•[4] SINHA, A., SHEN, Z., SONG, Y., MA, H., EIDE, D., HSU, B.-J. P., AND WANG, K. An overview of Microsoft academic service (MAS) and applications. In Proceedings of the 24th international conference on world wide web (2015), ACM, pp. 243–246.
•[5] TANG, J., ZHANG, J., YAO, L., LI, J., ZHANG, L., AND SU, Z. Arnetminer: extraction and mining of academic social networks. In Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining (2008), ACM, pp. 990–998.
•[6] WIKIPEDIA. List of academic fields, 2018.
•[7] GOYAL, P., AND FERRARA, E. Graph embedding techniques, applications, and performance: A survey. Knowledge-Based Systems (2018), 151, 78-94.
•[8] a social science map in 2004, see http://www.eigenfactor.org/map/maps.php.