



INDIAN INSTITUTE OF
INFORMATION
TECHNOLOGY

Data Science, Knowledge Discovery, and Data mining

Dr. Animesh Chaturvedi

Assistant Professor: IIIT Dharwad

Young Researcher: Heidelberg Laureate Forum

Postdoc: King's College London & The Alan Turing Institute

PhD: IIT Indore MTech: IIITDM Jabalpur



Indian Institute of Technology Indore
भारतीय प्रौद्योगिकी संस्थान इंदौर



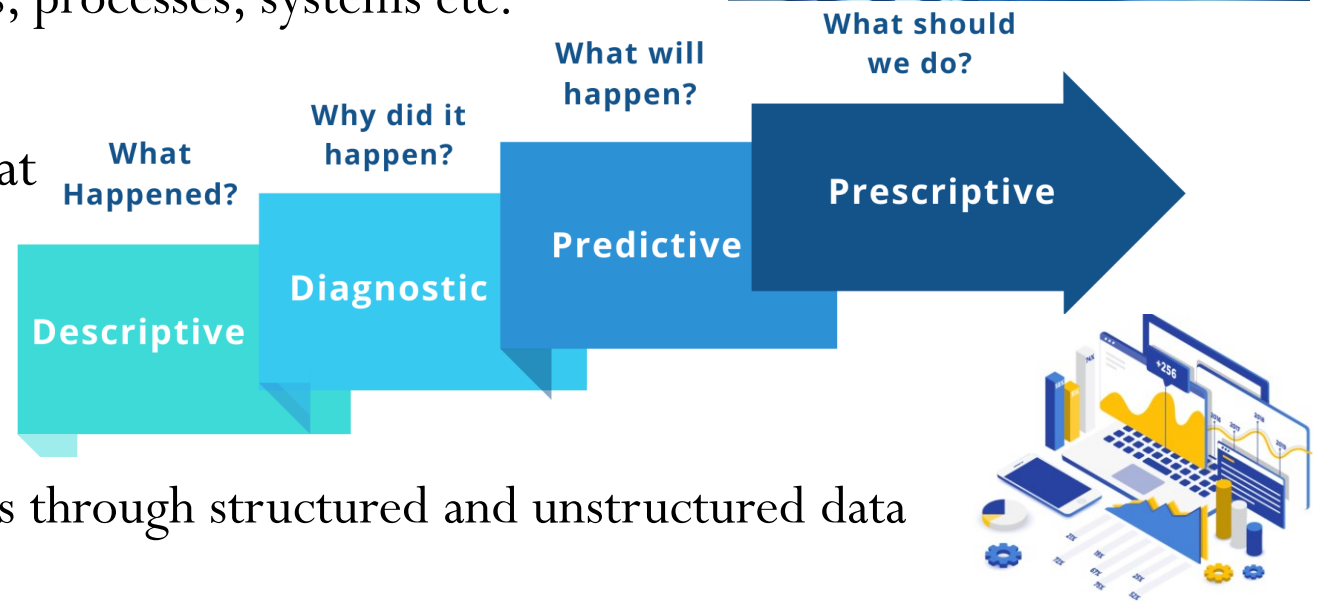
PDPM

Indian Institute of Information Technology,
Design and Manufacturing, Jabalpur



Data Science and Data Analytics

- **Data Science:** interdisciplinary science that
 - deals with data: methods, algorithms, processes, systems etc.
 - Theory oriented
- **Data Analytics:** analysis of data that
 - discovers trends, graph, tables etc.
 - Technology oriented
- Both
 - extracts knowledge and apply actions through structured and unstructured data insights
 - deals with data mining, machine learning, data management, and big data.
- **Data Scientist** and **Data Analyst** applies Data Science and Data Analytics



Evolution to Data Science Societies

- Institute of Radio Engineers (**IRE**) in 1951
- Institute of Electrical and Electronics Engineers (**IEEE**), **IEEE Computer Society** (1946; 1963)
- Association for Computing Machinery (**ACM**) in 1947
- Technology Engineering and Management Society (**TEMS**) 1955
- IEEE Systems, Man, and Cybernetics Society (**SMC**) (1958; 1972)
- Association for Computational Linguistics (**ACL**) 1962
- International Joint Conf. on Artificial Intelligence (**IJCAI**) 1969
- Special Interest Group on Management of Data (**SIG-MOD**), 1975
- Special Interest Group on Information Retrieval (**SIG-IR**) 1978
- Association for the Advancement of Artificial Intelligence (**AAAI**) 1979
- International Conf. on Machine Learning (**ICML**) 1980 in Pittsburgh
- Conf. on Neural Information Processing Systems (**NeurIPS**) 1986 -1987
- 38th IEEE International Conf. on Data Engineering (**IEEE ICDE**)*
- International World Wide Web Conf. (**WWW**) - Web Conf. 1994
- 28th International Conf. on Knowledge Discovery and Data Mining (**SIG-KDD**)*
- 22nd IEEE International Conf. on Data Mining (**IEEE ICDM**)*
- 10th IEEE International Conf. on Big Data (**IEEE Big Data**)*
- 9th IEEE International Conf. on Data Science and Advanced Analytics (**IEEE DSAA**)*

IRE

IEEE

ACM

TEMS

SMC

ACL

IJCAI

SIG-MOD & IR

AAAI, ICML

NeurIPS

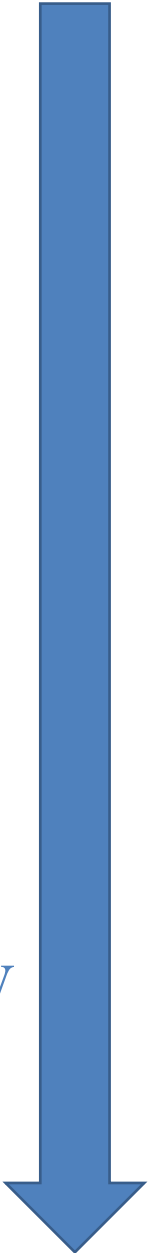
ICDE, WWW

KDD, ICDM

Big Data,

DSAA

* in 2022



DBLP & Google Scholar


- “Data” in DBLP search*
 - There are 459+ Venues matched in the DBLP, world most referred computer science bibliography website.
 - DBLP launched in 1993
- “Data” in Google Scholar metrics
 - Top 20 publications venues matching “data”

*2022

| | Publication* | <u>h5-index</u> | <u>h5-median</u> |
|-----|---|---------------------|------------------|
| 1. | ACM SIGKDD International Conference on Knowledge Discovery & Data Mining | 114 | 196 |
| 2. | IEEE Transactions on Knowledge and Data Engineering | 88 | 147 |
| 3. | International Conference on Artificial Intelligence and Statistics | 85 | 119 |
| 4. | ACM International Conference on Web Search and Data Mining | 69 | 133 |
| 5. | Journal of Big Data | 55 | 104 |
| 6. | IEEE International Conference on Data Mining | 53 | 81 |
| 7. | IEEE International Conference on Big Data | 52 | 93 |
| 8. | Knowledge and Information Systems | 51 | 76 |
| 9. | ACM Conference on Recommender Systems | 47 | 111 |
| 10. | Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery | 47 | 89 |
| 11. | European Conference on Machine Learning and Knowledge Discovery in Databases | 40 | 57 |
| 12. | ACM Transactions on Intelligent Systems and Technology (TIST) | 38 | 72 |
| 13. | Data Mining and Knowledge Discovery | 37 | 72 |
| 14. | SIAM International Conference on Data Mining (SDM) | 35 | 60 |
| 15. | ACM Transactions on Knowledge Discovery from Data (TKDD) | 35 | 53 |
| 16. | International Conference on Advances in Social Networks Analysis and Mining | 34 | 65 |
| 17. | Big Data Mining and Analytics | 31 | 39 |
| 18. | International Journal of Data Science and Analytics | 30 | 52 |
| 19. | Social Network Analysis and Mining | 30 | 46 |
| 20. | Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD) | 29 | 45 |

CSRankings.org for India

CSRankings: Computer Science Rankings

CSRankings is a metrics-based ranking of top computer science institutions around the world. Click on a triangle (▶) to expand areas or institutions. Click on a name to go to a faculty member's home page. Click on a chart icon (the  after a name or institution) to see the distribution of their publication areas as a . Click on a Google Scholar icon (🔍) to see publications, and click on the DBLP logo (📄) to go to a DBLP entry. Applying to grad school? Read this first. Do you find CSRankings useful? Sponsor CSRankings on GitHub.

Rank institutions in by publications from to

All Areas ☐ off | ☐ on

AI ☐ off | ☐ on

- ▶ Artificial intelligence ☒
- ▶ Computer vision ☒
- ▼ Machine learning & data mining ☒
- ACM SIGKDD, IMLS, NEURIPS/NIPS
- ICML ☒
- KDD ☒
- NeurIPS/NIPS ☒
- ▶ Natural language processing ☒
- ▶ The Web & information retrieval ☒

Systems ☐ off | ☐ on

- ▶ Computer architecture ☒
- ▶ Computer networks ☒
- ▶ Computer security ☒
- ▶ Databases ☒
- ▶ Design automation ☒
- ▶ Embedded & real-time systems ☒
- ▶ High-performance computing ☒
- ▶ Mobile computing ☒

| # | Institution | Count | Faculty |
|----|--|-------|---------|
| 1 | ▶ IISc Bangalore  | 2.6 | 28 |
| 2 | ▶ IIT Bombay  | 2.1 | 24 |
| 2 | ▶ IIT Delhi  | 2.1 | 22 |
| 4 | ▶ IIT Kanpur  | 2.0 | 19 |
| 5 | ▶ IIT Madras  | 1.9 | 23 |
| 6 | ▶ IIIT Hyderabad  | 1.6 | 19 |
| 6 | ▶ IIT Kharagpur  | 1.6 | 21 |
| 8 | ▶ IIIT Delhi  | 1.5 | 24 |
| 9 | ▶ IIT Gandhinagar  | 1.3 | 9 |
| 9 | ▶ IIT Hyderabad  | 1.3 | 10 |
| 11 | ▶ IIIT Bangalore  | 1.2 | 5 |
| 11 | ▶ IIT Patna  | 1.2 | 4 |
| 11 | ▶ IMSc  | 1.2 | 5 |
| 11 | ▶ Tata Inst. of Fundamental Research  | 1.2 | 9 |
| 15 | ▶ CMI  | 1.1 | 8 |
| 15 | ▶ IIT Goa  | 1.1 | 4 |

| #S | Institution | Count | Faculty |
|----|--------------------------------------|-------|---------|
| 1 | ▶ IISc Bangalore | 2.6 | 28 |
| 2 | ▶ IIT Bombay | 2.1 | 24 |
| 2 | ▶ IIT Delhi | 2.1 | 22 |
| 4 | ▶ IIT Kanpur | 2.0 | 19 |
| 5 | ▶ IIT Madras | 1.9 | 23 |
| 6 | ▶ IIIT Hyderabad | 1.6 | 19 |
| 6 | ▶ IIT Kharagpur | 1.6 | 21 |
| 8 | ▶ IIIT Delhi | 1.5 | 24 |
| 9 | ▶ IIT Gandhinagar | 1.3 | 9 |
| 9 | ▶ IIT Hyderabad | 1.3 | 10 |
| 11 | ▶ IIIT Bangalore | 1.2 | 5 |
| 11 | ▶ IIT Patna | 1.2 | 4 |
| 11 | ▶ IMSc | 1.2 | 5 |
| 11 | ▶ Tata Inst. of Fundamental Research | 1.2 | 9 |
| 15 | ▶ CMI | 1.1 | 8 |
| 15 | ▶ IIT Goa | 1.1 | 4 |
| 15 | ▶ IIT Guwahati | 1.1 | 7 |
| 15 | ▶ IIT Jodhpur | 1.1 | 5 |
| 15 | ▶ IIT Ropar | 1.1 | 2 |
| 20 | ▶ BITS Pilani | 1.0 | 1 |
| 20 | ▶ BITS Pilani-Goa | 1.0 | 1 |
| 20 | ▶ DAIICT | 1.0 | 2 |
| 20 | ▶ IIT (BHU) Varanasi | 1.0 | 1 |
| 20 | ▶ IIT Indore | 1.0 | 3 |
| 20 | ▶ IIT Mandi | 1.0 | 1 |
| 20 | ▶ IIT Roorkee | 1.0 | 3 |

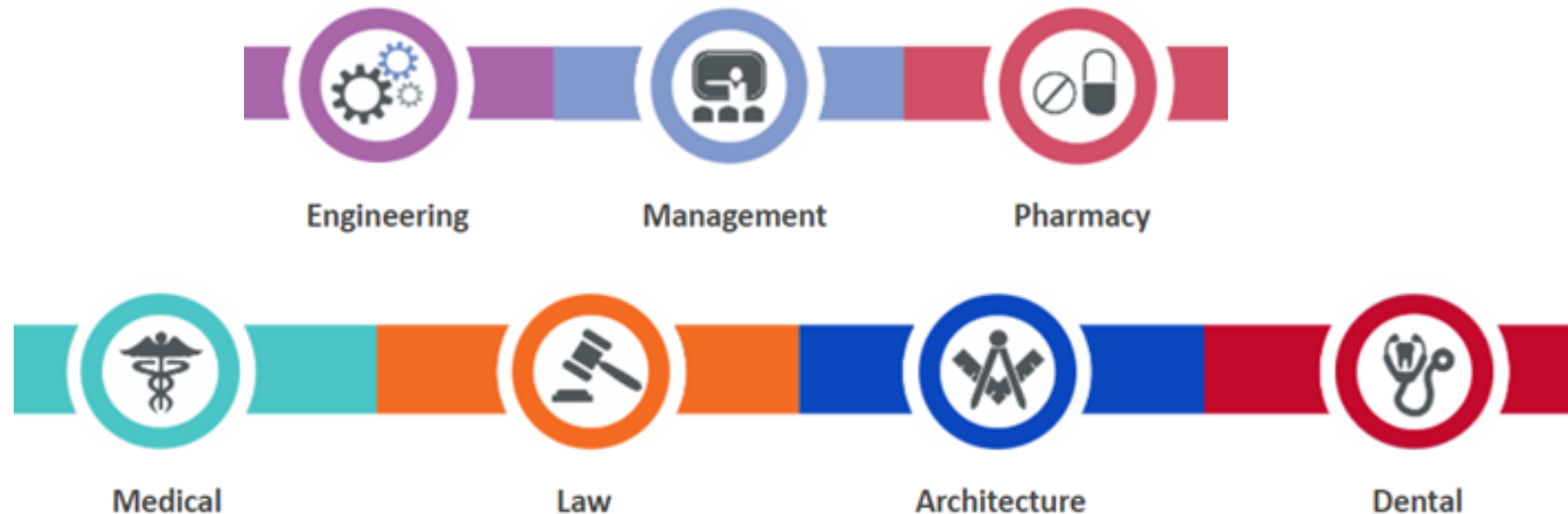
CS, EC, and IT and other related
branch based institute ranking

Institutes from India listed for A* CS events like KDD in 2022

Data Analytics for Information Technology (IT)

- IT based category parameter are different from Engineering, Maths, and Sciences
- IT means (CS-EC-DS-AI-Bio/Chem-informatics-etc)
 - research, jobs, in DBLP, Scholar Metrics
 - Competitions like ICPC, KDD Cup, ACM Student Research, etc.

IT is applied as advanced field for



Knowledge Discovery and Data Mining

- “a unifying framework for Knowledge Discovery in Database (KDD)”
- links between data mining, knowledge discovery, and other related field

Selection, Preprocessing, Transformation, Data Mining, Interpretation/Evaluation

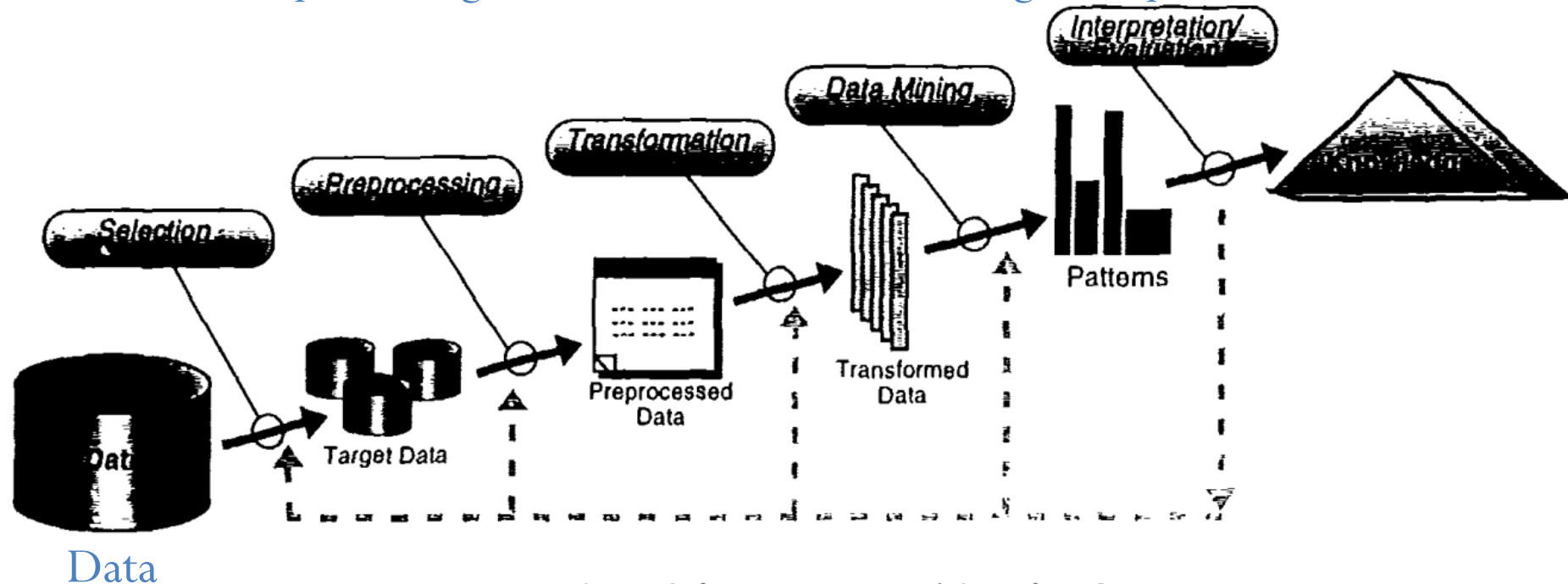


Figure 1: An overview of the steps comprising the KDD process.

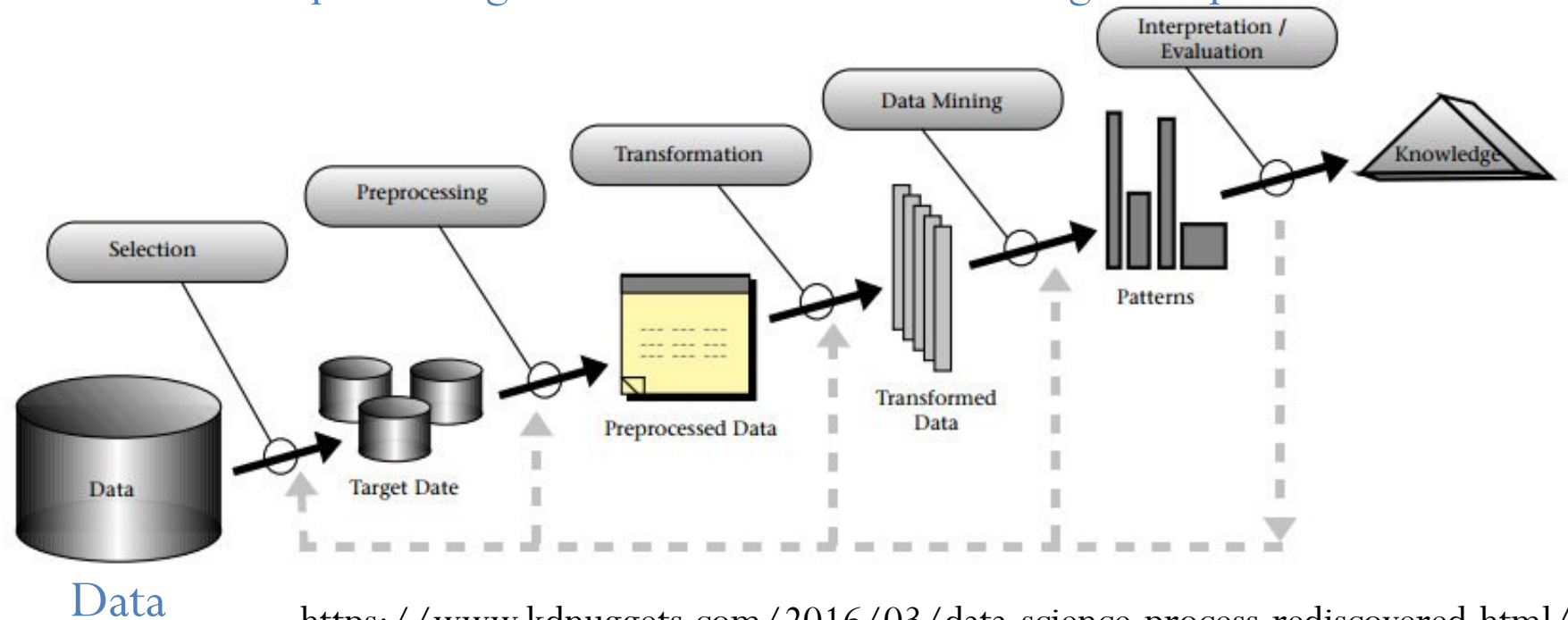
Fayyad, Usama M., Gregory Piatetsky-Shapiro, and Padhraic Smyth.

"Knowledge Discovery and Data Mining: Towards a Unifying Framework." *KDD*. 1996.

Knowledge Discovery and Data Mining

- “a unifying framework for Knowledge Discovery in Database (KDD)”
- links between data mining, knowledge discovery, and other related field

Selection, Preprocessing, Transformation, Data Mining, Interpretation/Evaluation



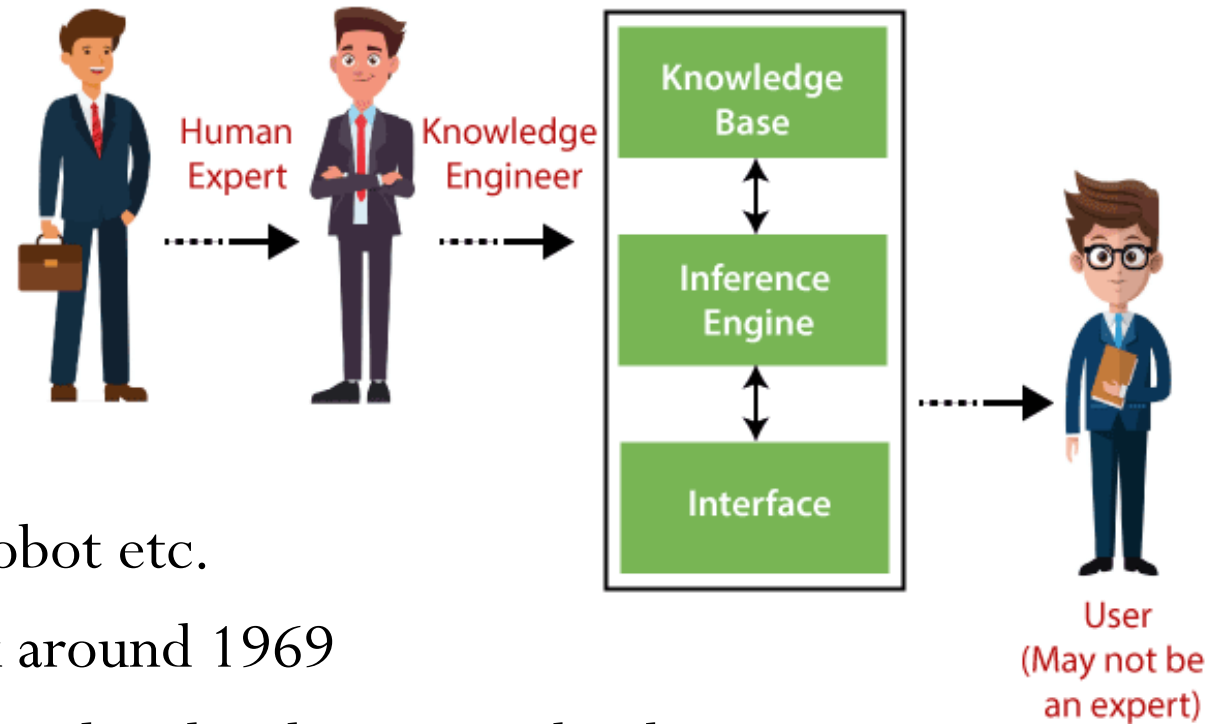
<https://www.kdnuggets.com/2016/03/data-science-process-rediscovered.html/2>

Fayyad, Usama M., Gregory Piatetsky-Shapiro, and Padhraic Smyth.

"Knowledge Discovery and Data Mining: Towards a Unifying Framework." **KDD**. 1996.

Expert, System, and Knowledge Engineer

- Expert Systems (1960-74)
 - Explicit rules
- Playing Chess,
- Organic and Biology recommendations
- Solving word problem in Algebra
- Natural Language processing, Mobile Robot etc.
- Backpropagation based Neural Network around 1969
- Association Rule Mining by Rakesh Agrawal and Srikant Ramakrishnan 1993-95
- Big Files and Google File System in 1995-2005 by Larry Page, Sergey Brin and Sanjay Ghemawat, et al.



ขอบคุณ

Thai

Grazie
Italian

תודה רבה
Hebrew

धन्यवादः
Sanskrit

ಧನ್ಯವಾದಗಳು
Kannada

Ευχαριστώ
Greek

Thank You
English

Gracias
Spanish

Спасибо
Russian

Obrigado
Portuguese

شكراً
Arabic

<https://sites.google.com/site/animeshchaturvedi07>

Merci
French

多謝
Traditional
Chinese

धन्यवाद
Hindi

Danke
German

多谢
Simplified
Chinese

நன்றி
Tamil

ありがとうございました
Japanese

감사합니다
Korean