Sudipta Dasmohapatra (sd345@duke.edu)

Trends in Data Science: Education, Research and Practice

July 20-24, 2020

Artificial Intelligence Workshop: Strategies to Train and Engage Students

About Me

Market Research Consulting (3 yrs)

NC State University (10 yrs)
Teaching/Mentoring + Research

→

GEORGETOWN UNIVERSITY McDonough
SCHOOL of BUSINESS

Duke Statistics (3 yrs) – Administration + Teaching (+ Research)
Duke Fuqua (3 yrs) – Teaching

~ Others

SAMSI
ASA
DAA
Vertaeon
Kloutix

Introduction and Key Thoughts

- There will never be a better time to be in the field of Data Science and AI
  - **Democratization** : data analysis as a basic skill will give you a competitive edge in any discipline
  - **Interdisciplinary nature** more pronounced
  - **Collaboration** abound between industry and academia (industry boards, data fests, research, etc.)
  - Data Science and related **academic training proliferate** (various modes, differing content, across various schools, etc.)
  - **Big data** (and complexity) and technologies will present **opportunities and challenges**

Origin of Data Science

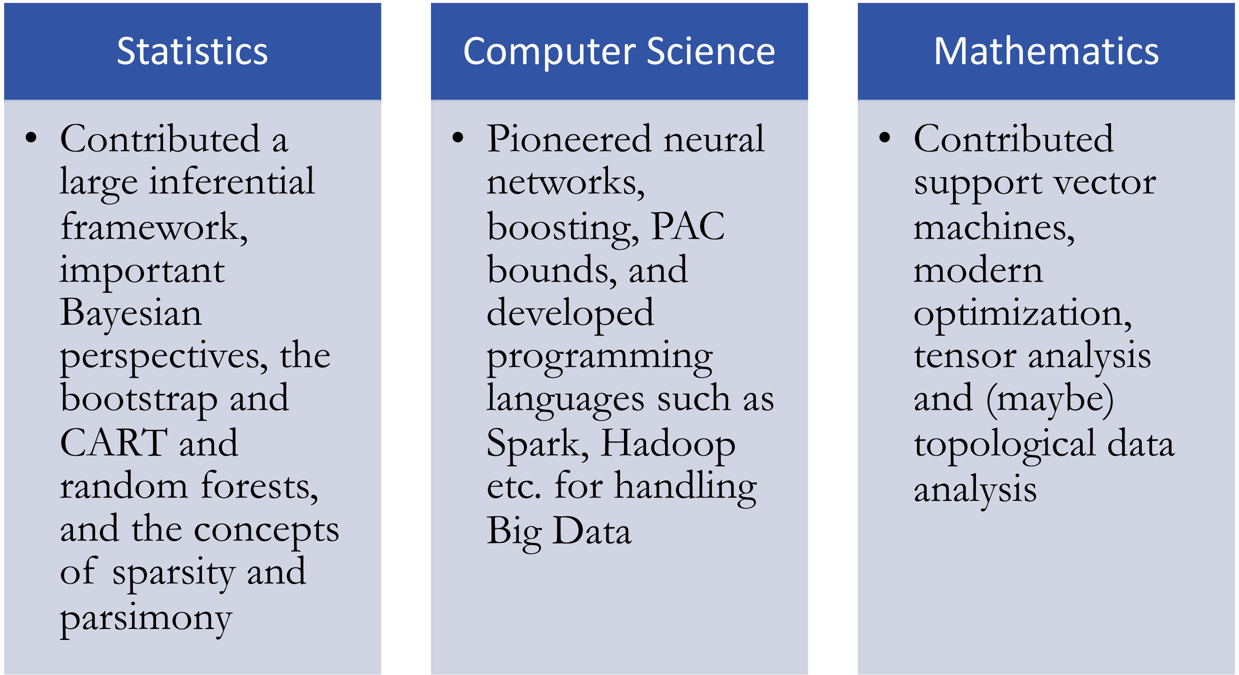
More than 50 years ago, John Tukey in "The Future of Data Analysis," pointed to the existence of an as-yet unrecognized *science*, whose subject of interest was learning from data, or "data analysis."

The term "data science" was suggested by Bill Cleveland and Jeff Wu

Beyond theoretical statistics and modeling

What is Data Science?

Data Science is an area at the interface of statistics, computer science, and mathematics



Statistics

- Contributed a large inferential framework, important Bayesian perspectives, the bootstrap and CART and random forests, and the concepts of sparsity and parsimony

Computer Science

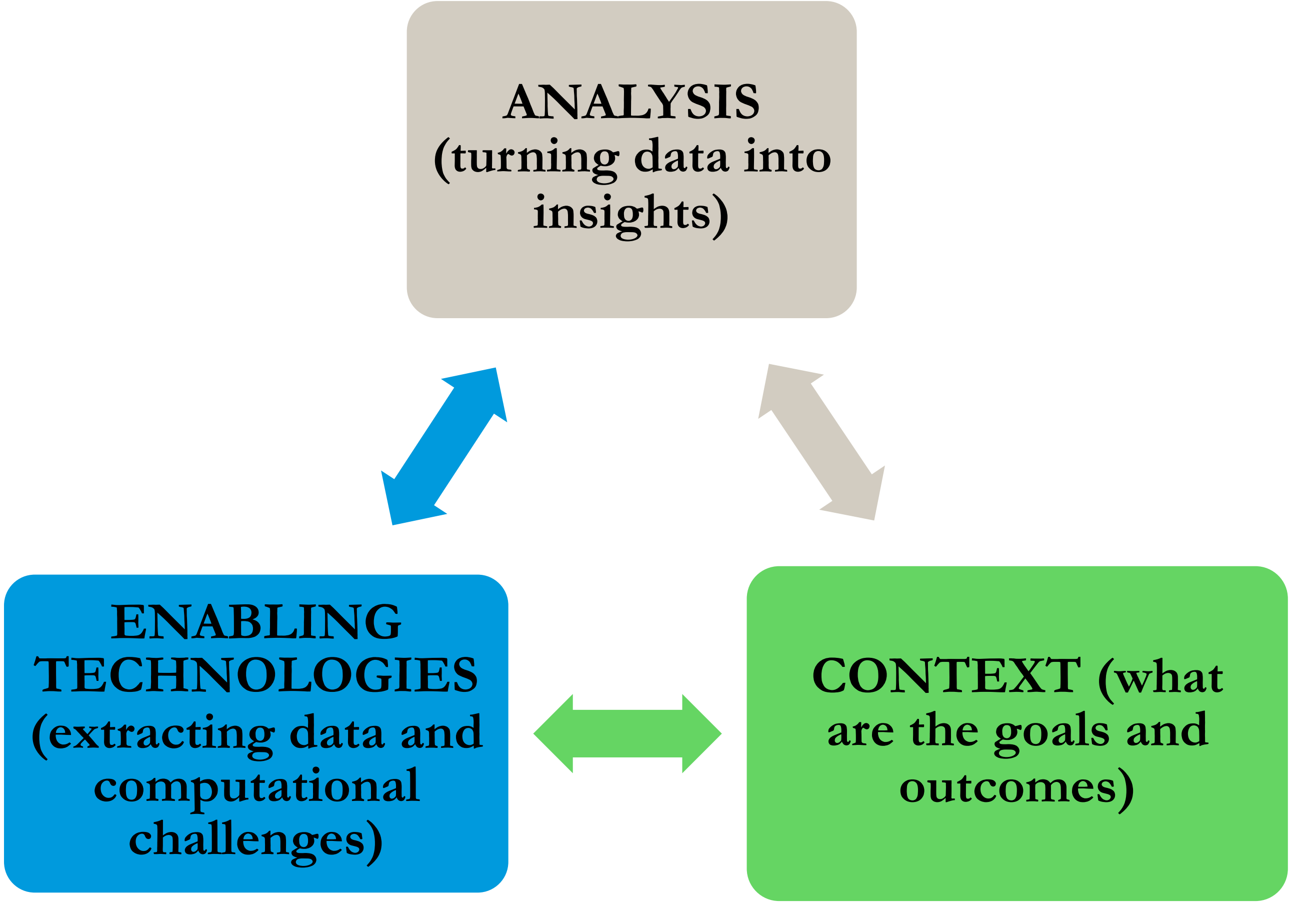
- Pioneered neural networks, boosting, PAC bounds, and developed programming languages such as Spark, Hadoop etc. for handling Big Data

Mathematics

- Contributed support vector machines, modern optimization, tensor analysis and (maybe) topological data analysis

Elements of Data Science

Scientific thinking, methods and approaches to spur development of intelligence to solve problems

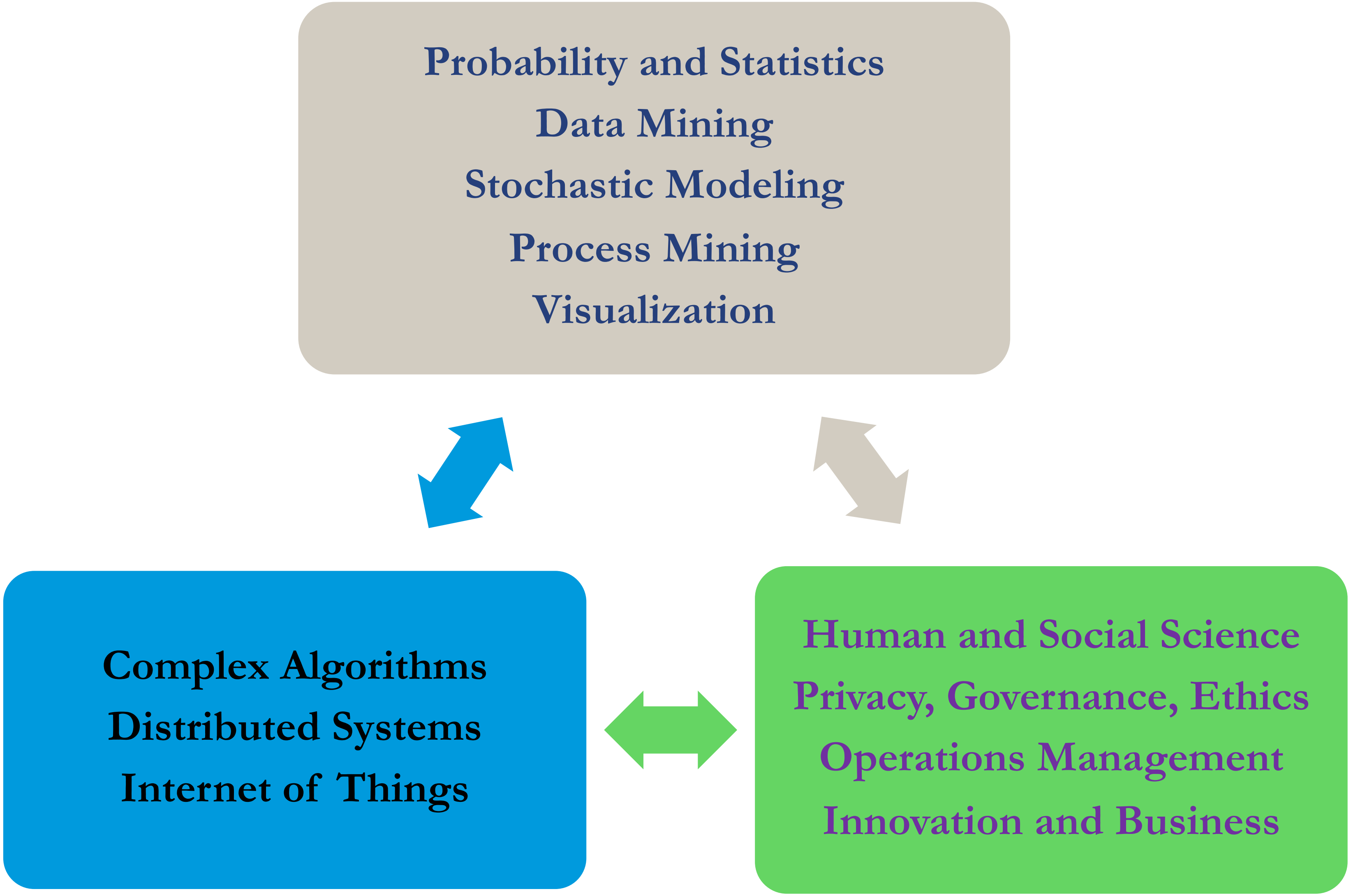


ANALYSIS (turning data into insights)

ENABLING TECHNOLOGIES (extracting data and computational challenges)

CONTEXT (what are the goals and outcomes)

Elements of Data Science



Probability and Statistics
Data Mining
Stochastic Modeling
Process Mining
Visualization

Complex Algorithms
Distributed Systems
Internet of Things

Human and Social Science
Privacy, Governance, Ethics
Operations Management
Innovation and Business

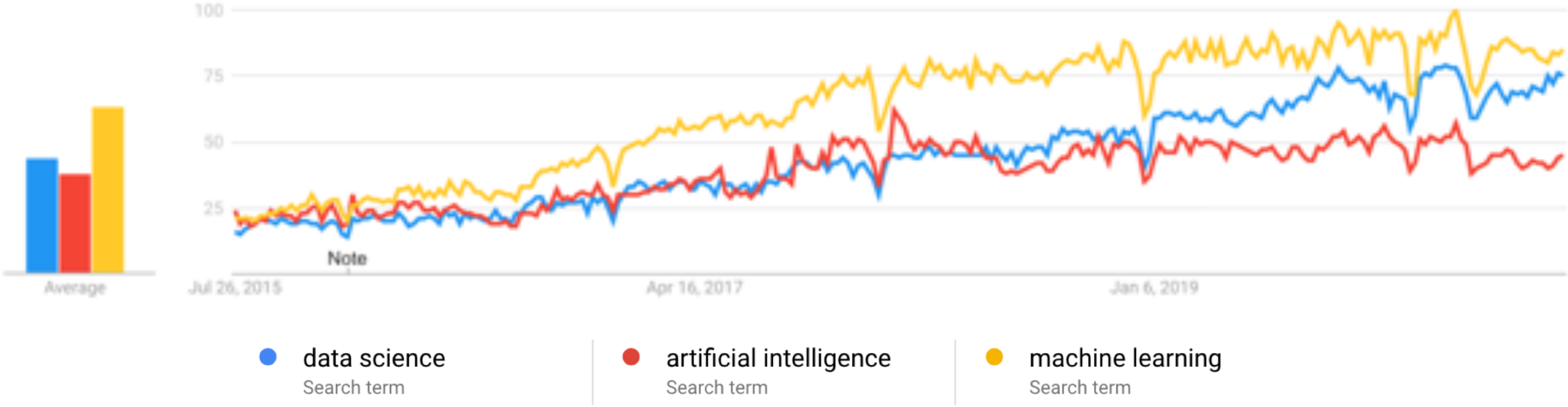
Artificial Intelligence and ML as Subsets of Data Science

AI: Information systems and algorithms designed to make decisions commonly associated with human intelligence (learning, problem solving, pattern recognition), generally with real-time data (Automated Cars, Assist AI in Healthcare)

ML: Machine learning focuses on the development of computer programs that can access data and use it to learn for themselves (GMail: content modeling; Amazon: product recommendations; Uber: *ubereats* to estimate time to deliver food; Youtube: improve search results, etc.)

Source: HBR, 2018 Thomas Davenport and Rajeev Ronanki

General Interest (Data Science and Related Topics)



what is data science
computer science
data science masters
data science jobs
data science online
python data science
data analytics
analytics
python
data scientist
data science salary
data science degree
data science definition
machine learning

python
google machine learning
what is machine learning
ai
ai machine learning
deep machine learning
deep learning
machine learning algorithms
machine learning tutorial
machine learning coursera
machine learning jobs
machine learning model
machine learning course

ai artificial intelligence
what is artificial intelligence
movie artificial intelligence
artificial intelligence learning
google artificial intelligence
artificial intelligence machine learning
machine learning
definition artificial intelligence
artificial intelligence robot
artificial intelligence technology
artificial intelligence jobs
artificial intelligence companies

Global State of Enterprise Analytics (2020)

- Over past five years, globally 11% of organizations (24% of US firms) spent over $100,000 on analytics products (includes ML and Cloud Computing) (State of Data Science and ML study, Kaggle 2019)
- 65% of global organizations (with >$100 million annual revenue) say that they plan to increase their analytics spending in 2020 (telecommunications, hospitality, and retail lead all spending) (MicroStrategy Report on Enterprise Analytics 2020)
- Cloud Computing (AWS, Google Cloud and Microsoft Azure dominate), IoT, AI/ML will have the greatest Impact on enterprise's analytics initiatives (over the next five years)
- Advanced and Predictive Analytics dominate the analytics initiatives today (Linear/logistic regression, Decision Trees, CNN, GBM, Bayesian Approaches, NN important in that order)

Source:: Global State of Enterprise Analytics Report by MicroStrategy (2020), >500 organizations; State of Data Science and ML Study 2019 by Kaggle (>19500 professionals)

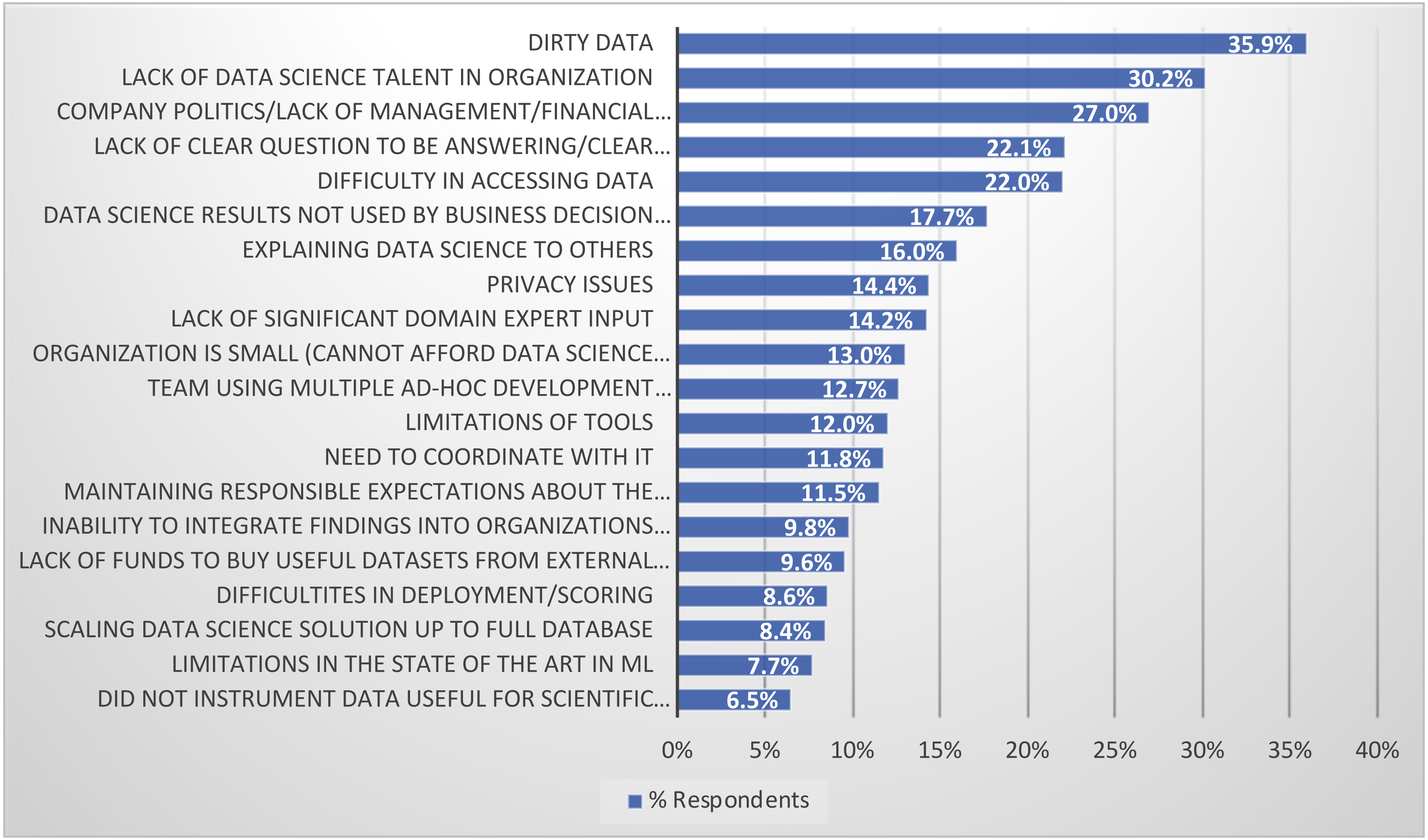
Industry Outlook and Opportunities

- **Availability of data across multiple channels, modes**
  - Videos (will pull 80% web traffic) and voice data (most search queries) will dominate (vlogs instead of blogs?)
  - Citizen data science
- **Technology:**
  - Transitional mode: advanced data technologies to replace routine business processes (e.g., low code software development platforms with user-friendly structures)
  - Augmented Analytics and Edge Computing: driven by sensors (streaming data stored close to the data source for real-time analytics)
  - Open source tools that make reproducibility easier
  - Integration of big data across multiple sources

Industry Outlook and Opportunities

- **Intelligent data mesh**: connected network of people, processes, devices, apps (self learning intelligent systems)
- **Collaboration** with academia – research and education
- **Bring data science roles to main stream** (diverse roles): Need for data literacy at all levels and functions
- **Governance regulations**/data privacy and security will be key and will guide data use (GDPR)

Common Data Science Challenges at Work (Industry Perspective)

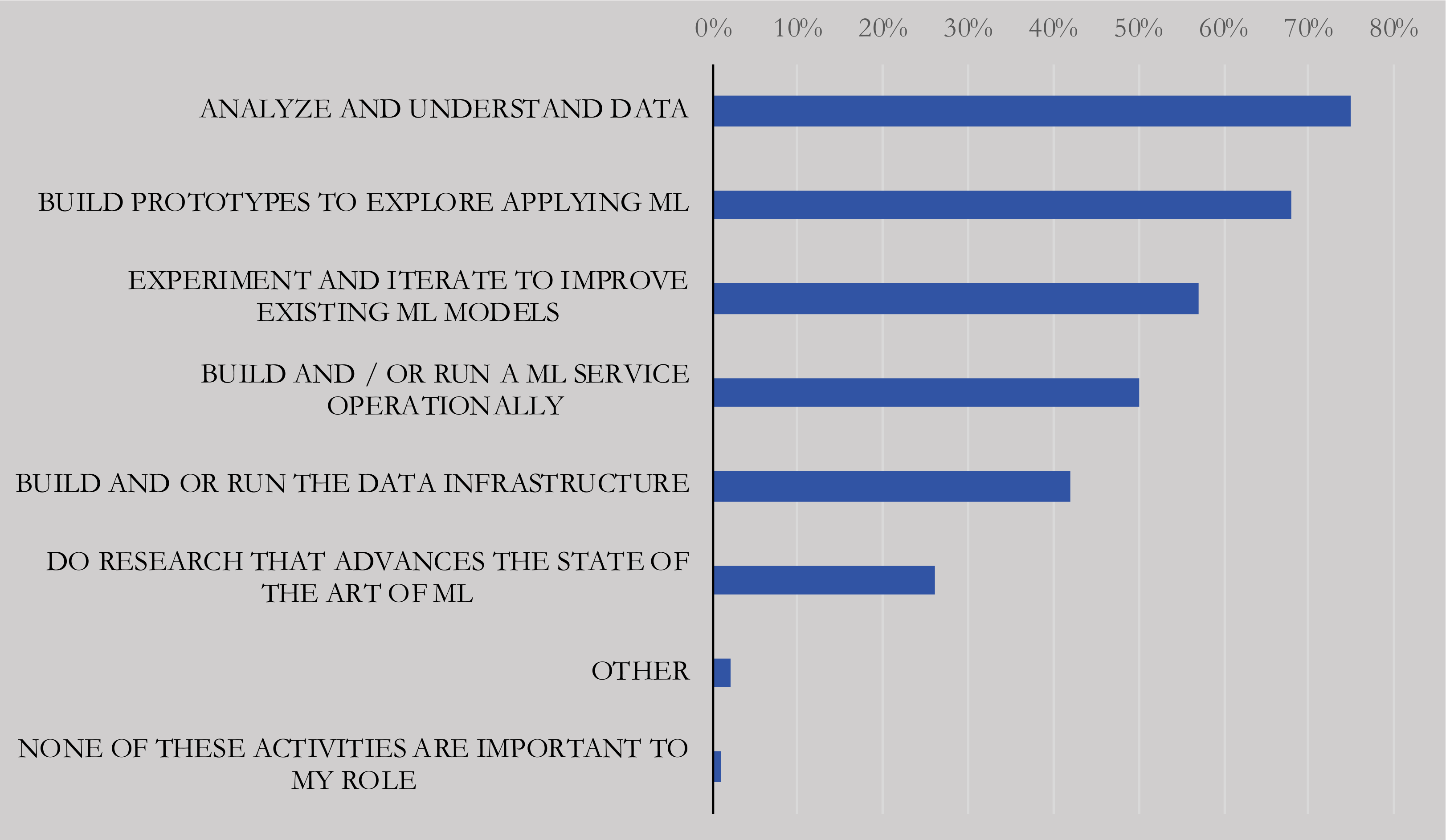


Source: State of Data Science and ML study, Kaggle 2017 (16000 professionals)

Common Data Science Challenges in Organizations

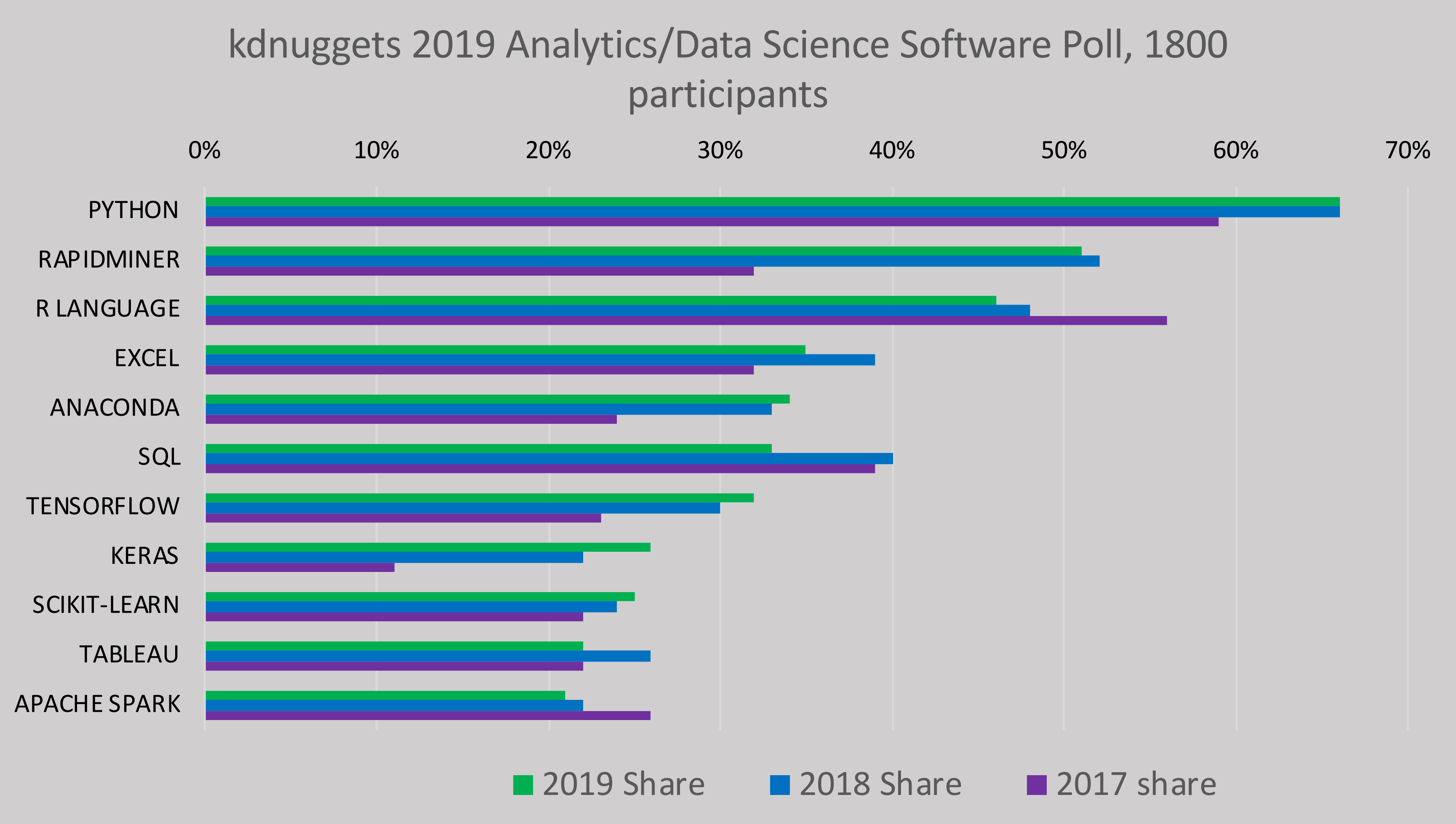
- Lack of understanding of the value of data across organization
- Too much data across multiple sources – lack of consolidation or integration
- Issues of data quality
- Turf war over who owns data
- Lack of skilled staff and access to skilled personnel
- Small and mid-size companies cannot afford to hire specialists?
- Training needs for existing employees
- Finding the right partner
- Data security and privacy

How Data Scientists Spend their Time?



Source: State of Data Science and ML study, Kaggle 2019 (>19500 professionals)

Common Data Science Tools



Kaggle survey -19,500 participants in 2019 (Jupyter Notebook, Rstudio, Python in that order)

Degrees that Industry Seeks

- Over 52% of employees in the data science industry have a MS degree and 19% have PhD's according to Kaggle Data Science professionals (2019 data, >19,500 professionals)

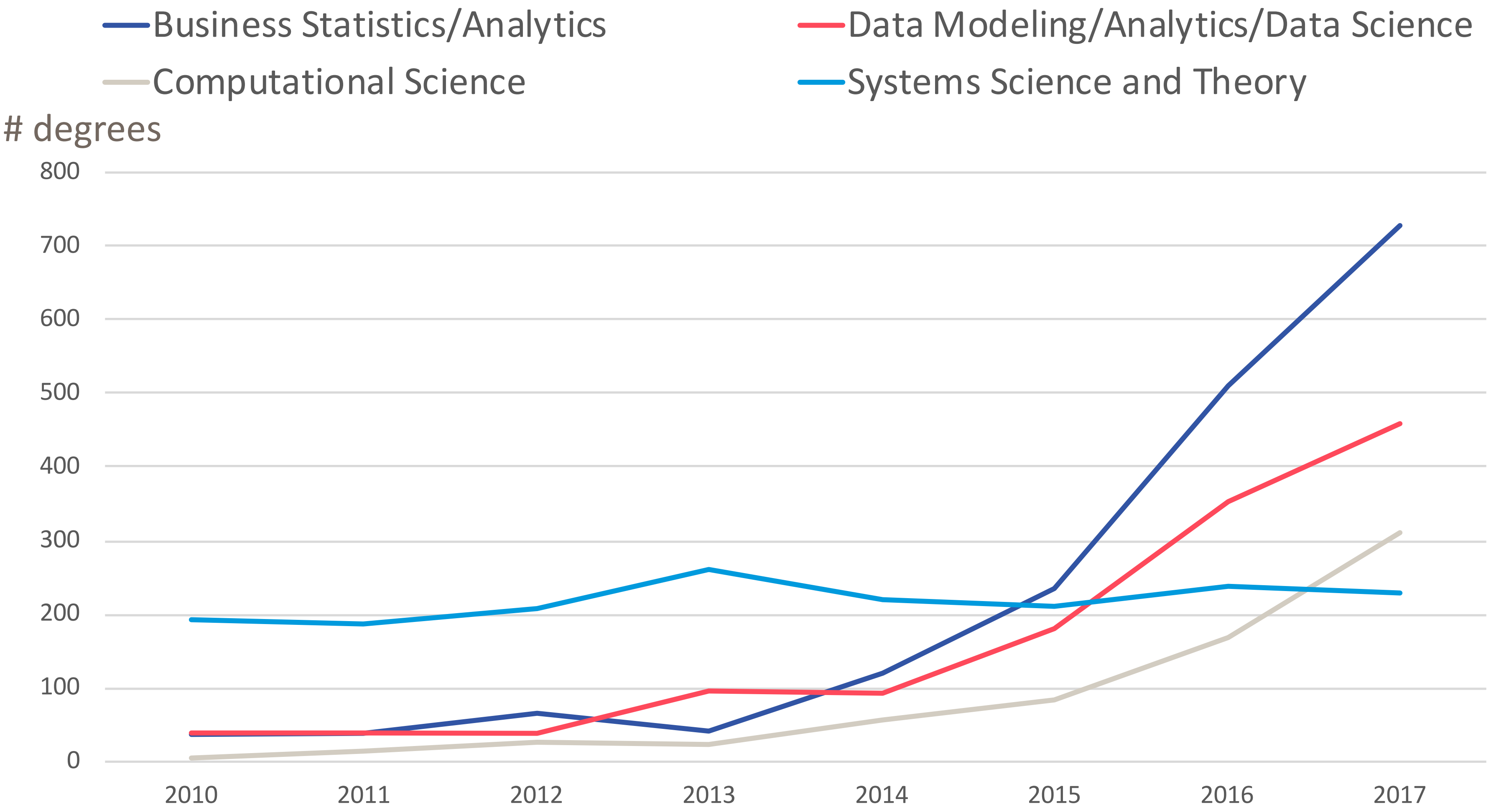
Type	Data Scientist
Number of jobs	10,996
Employer Category*	Technology: 64.2% Consulting: 20.3% Financial Services: 10.5% Telecommunication: 2.6% Retail: 2.3%
Level of position Entry (<=BS degree) Mid (MS or BS+) Senior (PhD or MS+)	 22.3% 53.8% 23.9%

Indeed Search Query (Data Scientist in Job Title) – March 2020)

Note: Full time jobs in the US, all open jobs; *Academic jobs not represented well here

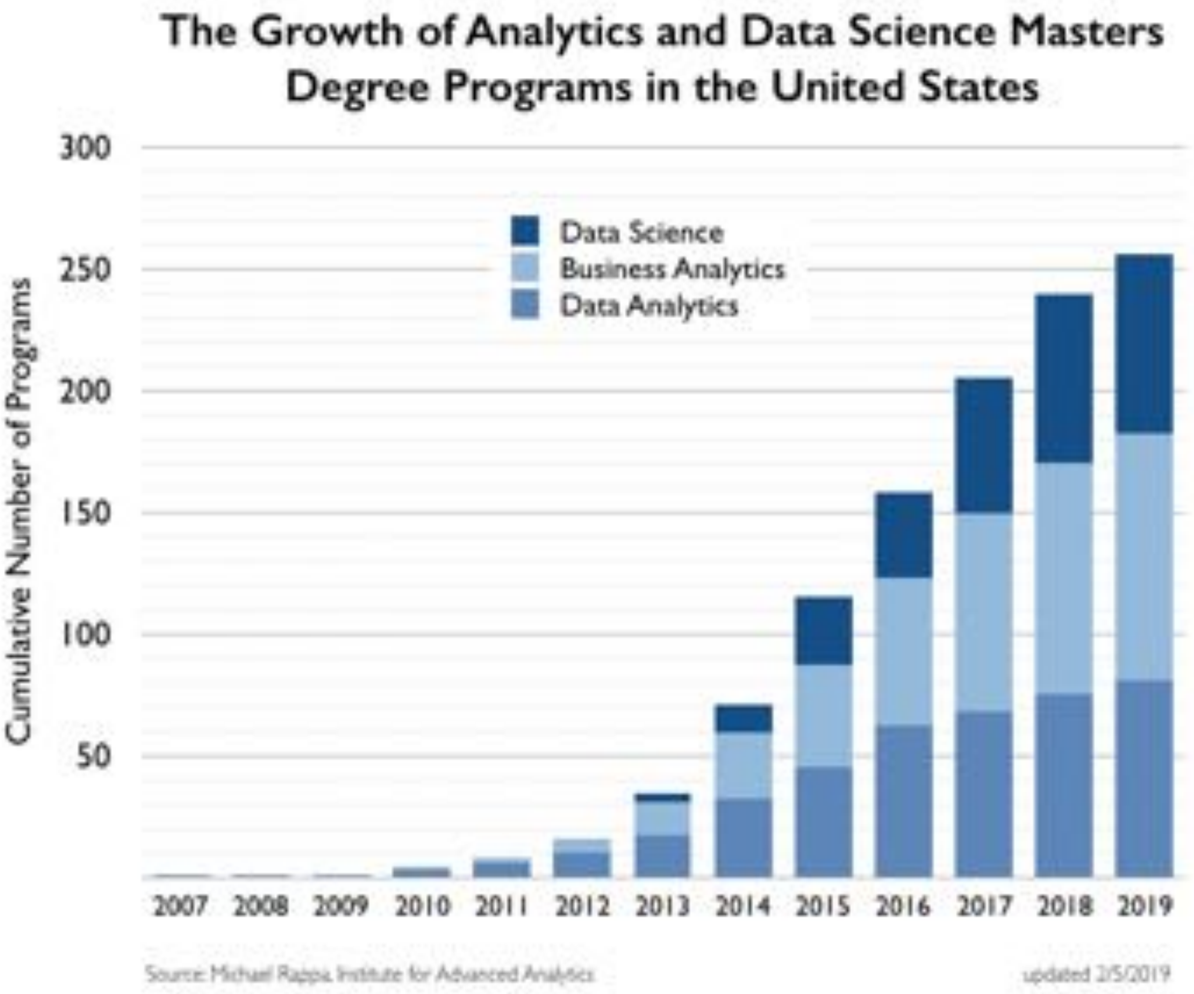
Glassdoor shows >24,000 jobs with data scientist query

Rise in MS in Data Science and Analytics Degrees



Source: amstat.org

Growth of MS Degree Programs



Data Science Education

- Curriculum (1 -2 YEARS)
  - Basic/Advanced Statistics
  - Machine Learning
  - Computing & Visualization
  - Business Analytics
  - Communication
  - Capstone/Internship/Practicum
- Tool skills:
  - R, Python, SQL, Tableau/Shiny/Power BI, Hadoop- Map Reduce, C++

- Industry (Forbes, 2018)
  - Specialist, not a generalist
  - Inform choices based on business goals
  - Get cross-departmental expertise (domain)
  - Data management, integration and posturing is key
  - Communication
  - Collaboration and flexibility

Prevalent Themes for the Future

Ethics in data science (ownership, privacy, bias, misuse, transparency, right of access, etc.)

Integration of experiential learning (internships, capstone projects, case studies, etc.)

Recognition of ”power skills”(communication, leadership, team work, visualization, critical thinking)

Collaboration between Industry and Academia (Where and Benefits)

Collaboration Areas

Research and Consulting

Collaboration on lectures, internships/ practicums

Data-fest/Kaggle, Consulting case studies, Kaggle class projects

Tool set reconfiguration based on industry needs

Industry board

BENEFITS

Academia

- Exposure to the real world
- Well prepared students
- Research topics
- Guidance into application of theory…

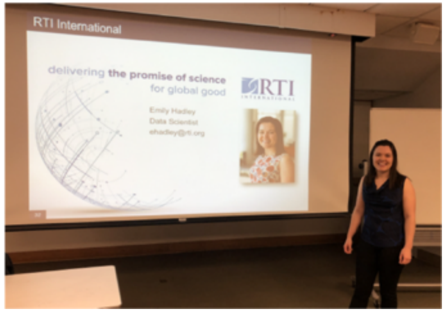
Industry

Input into future of education

Talent (~need less training)

Get innovative research projects off the ground (trial and error)

…



Future of Analytics Education

- High demand for analytics talent (US alone is projected to face a shortfall of ~250,000 data scientists by 2024)
- Education will continue to be embedded into various disciplines (business, finance, social science, healthcare, etc.): bring different skills and problem solving from varied perspectives)
- Partnerships with industry/government/non-government (Business/ Functional expertise important for making "business decisions from data")
- Accelerated learning: Data science and analytics related training courses and boot-camps will proliferate
- Various delivery options and packaging:
  - On-Campus
  - Online
  - Blended learning
  - Corporate training
  - Partnerships with online learning platforms (e.g., data camp)
  - Part-time and full-time

Data Science Education Challenges

- Still a very young discipline (new programs very different from first programs)
- No consensus on where it should be housed
- Varied quality, content and focus (how do advisors guide students?)
- Various platforms and delivery options : (off-line, on-line, certifications vs degrees, Coursera / Data Camp, MOOCs, …)
- How to collaborate with Industry and Other Disciplines?
- Numerous conferences around this topic (e.g., Open Data Science Conference)
- Computing tools and automated platforms

Research Opportunities (GRANTS.GOV)

1581/2785 (posted)
1829 /2.912 (closed)

- Harnessing the Data Revolution (HDR): Institutes for Data-Intensive Research in Science and Engineering - Ideas Labs (NSF)
- NLM Research Grants in Biomedical Informatics and Data Science (R01 Clinical Trial Optional) (HHS, NIH)
- Predoctoral Training in Advanced Data Analytics for Behavioral and Social Sciences Research (BSSR) - Institutional Research Training Program [T32] (HHS)
- Short-term Mentored Career Enhancement Awards in Mobile and Wireless Health Technology and Data Analytics: Cross-Training at the intersection of Behavioral and Social Sciences and STEM Disciplines (DOI-NPS)
- Computational and Data-Enabled Science and Engineering (NSF)
- ROSES 2019: New Frontiers Data Analysis (NASA)
- U.S. Army Research Institute for the Behavioral and Social Sciences Broad Agency Announcement for Basic, Applied, and Advanced Research (Fiscal Years 2018-2023)
- …

Majority Discipline : Healthcare/Biological Science

Current Research in Data Science

- Emerging as cross and inter-disciplinary field
- Most publications concern concepts and topics in statistics, data mining, machine learning but many now in social science, policy, business
- Fusion of data from heterogeneous sources (devices, online and offline, across different divisions in companies to improve predictions)
- NLP and Deep Learning gaining ground in cutting edge research (healthcare informatics/analytics)
- Interplay between big data/data streams and complex data types (multidimensional data to social networks, graphs, social streams)
- Compact storage systems and efficient processors for data processing
- Data ethics and privacy
- …

Data Science (Scientific Research Opportunities)

Exploration of complexities inherently trapped in data, business and systems

Mathematical and statistical foundation challenges: Existing theories extended or substantially redeveloped

Development of domain specific analytics theories, tools and systems (e.g., self monitoring devices)

Identify, specify and respect social issues in various domains (e.g., privacy, security preserving algorithms)

Identify and quantify value associated with data (e.g., developing measurement for actionability)

Decision support theory and system development (e.g., tools for transforming findings to actionable strategies)

Source: Adapted from Data Science: Challenges and Directions, 2016, ACM, Longbing Cao and Usama Fayyad

Future Potential in Research and Practice areas of Data Science

- Digital Era and Big Data (complex Data Types)
- Privacy and Ethics
- Reproducibility
- Interdisciplinary and Collaborative
- Computing and Automation

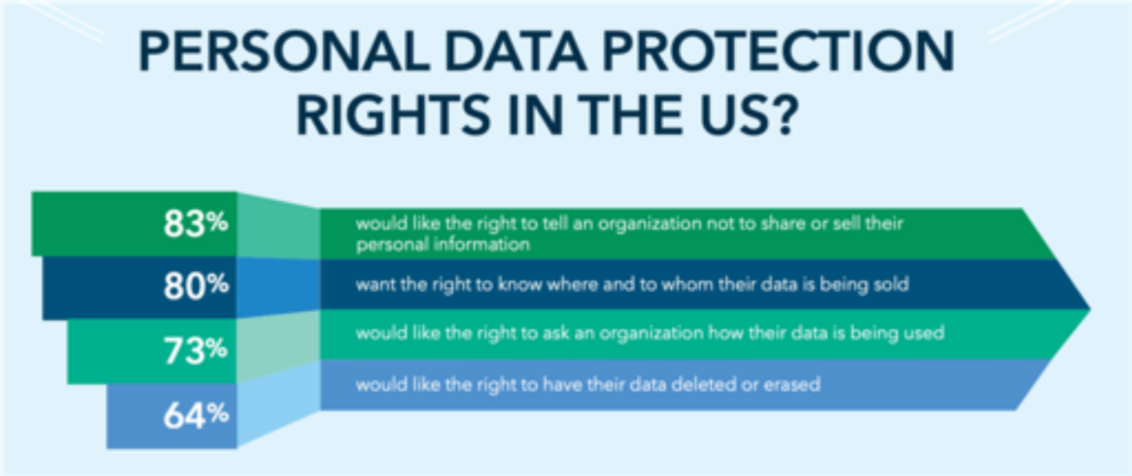
Digital Era and Big Data

- Large data sets with complex data types: sparse data, non-linear distributions, unstructured data, high-dimensional, etc.- Efficient procedures (with higher precision) for performing analysis and drawing large scale inference on trillions of data points

University of Chicago, Stat/Math department sophomores present gene network analysis tool : its basis is Statistics (parametric and nonlinear correlation among genes to generate network)

Privacy and Ethics

- Privacy implications and the human element



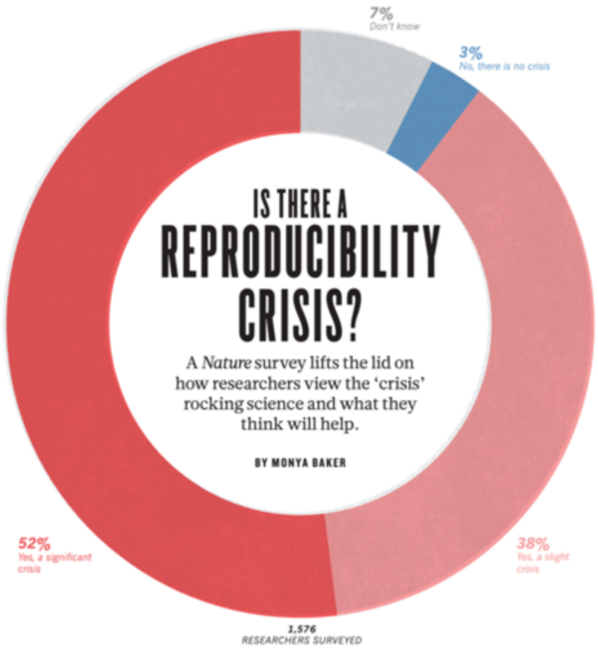
N=525 adult consumers

Development of statistical methods to provide meaningful inferences from anonymized data

Source: Data privacy: Are you Concerned? 2018 report by SAS Institute, www.sas.com

Reproducibility

- Reproducibility and Transparency
  - Data collection
  - Statistical methodology and algorithm accountability (Methods to produce reliable and consistent results)
    - o Insights into how and why?
    - o Trade-off between error and transparency

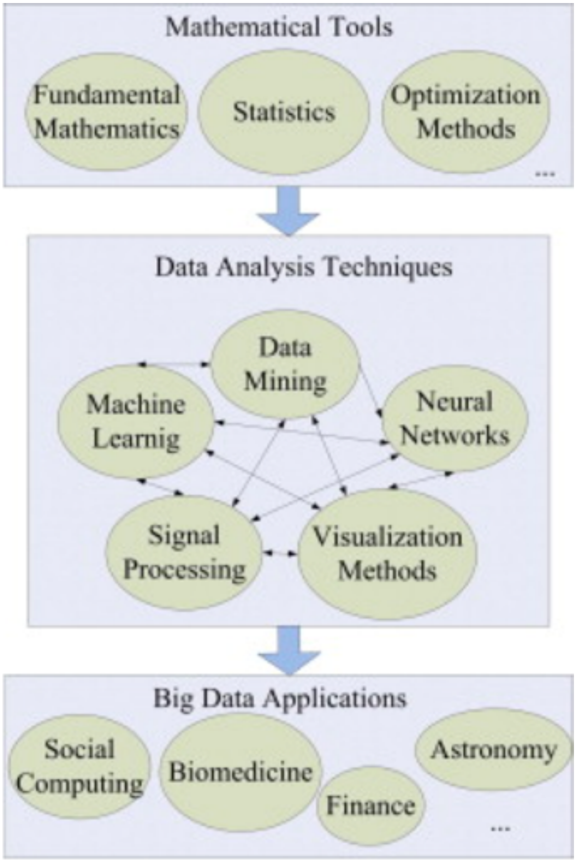


One of the most important factors driving the reproducibility crisis is insufficient statistical knowledge

Source: Collection of Papers from Statistics and Probability Letters Journal on role of statistics in Big Data (2018); Is there a reproducibility crisis? Monya Baker, 2016, Nature

Interdisciplinary and Stronger Collaborations

- Robust analysis of complex data sets will require better collaborations among various disciplines



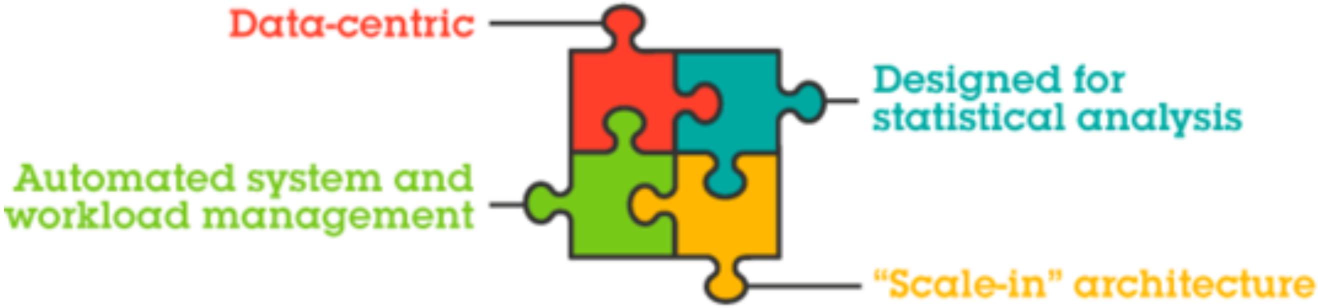
→ Statistics
+ ML Community
Big Data Models

Interdisciplinary degrees offered across many universities

Source: Data intensive applications, challenges, techniques, and technologies: A survey of big data, Chen, P. and Zhang, C., Information Sciences, 10(2014): 314-347

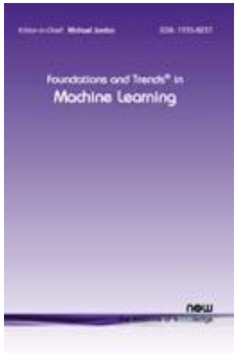
Computing and Automation

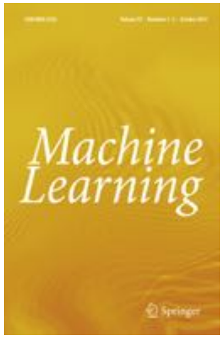
- Increasing focus on computing as a big part of statistics and technology and tools will drive the science
  - Scalable statistical procedures for massive datasets
  - Almost all statistics degree programs include some level of computation
    - Tabulation
    - Programming
    - Automation (e.g., distributed systems, self learning software)


Data-centric
Designed for statistical analysis
Automated system and workload management
"Scale-in" architecture

Cognitive Systems – New Era of Computing (IBM)

Popular Data Science Journals

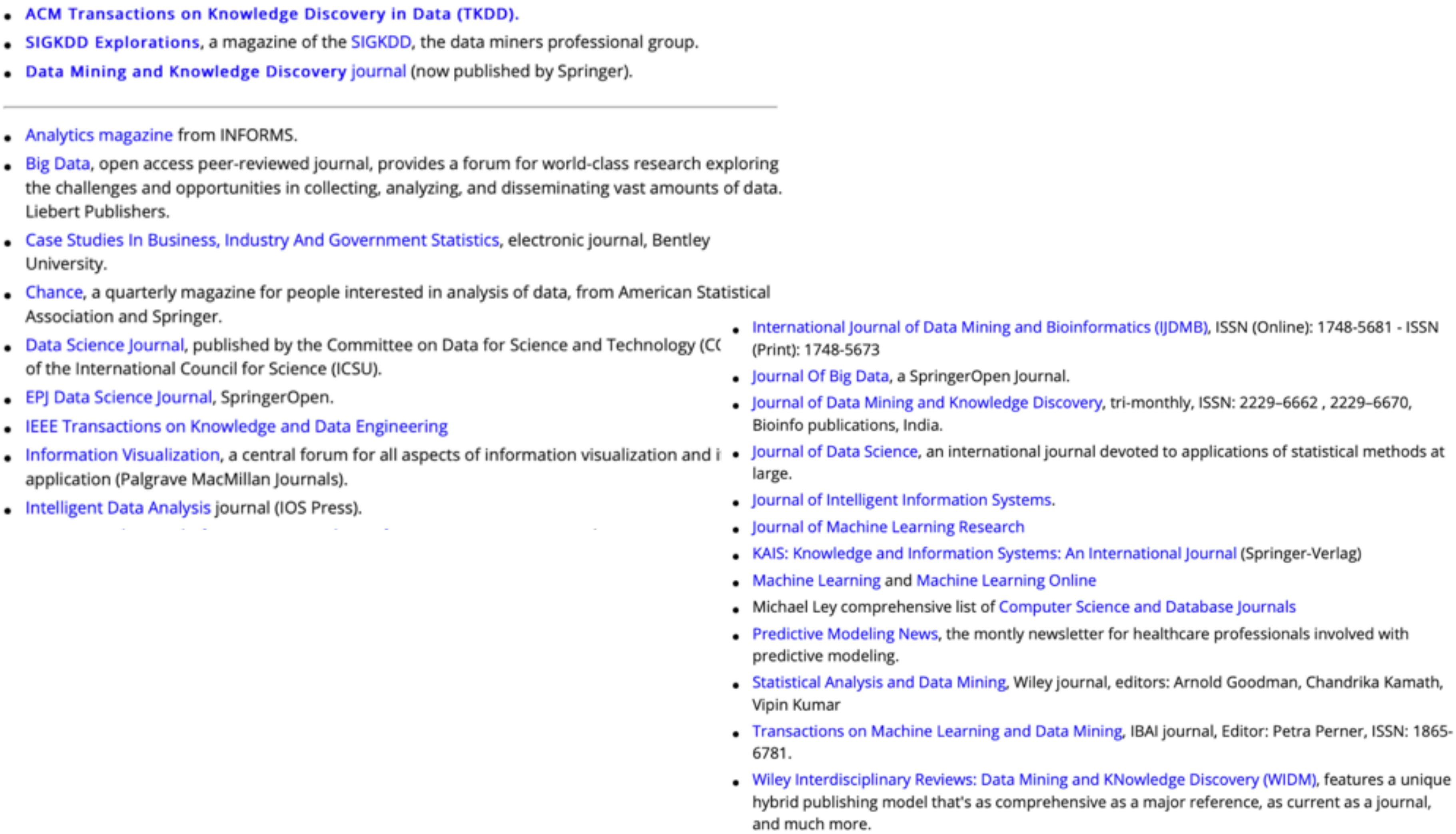




Source: Analyticsindiamag.com (2016)

Combination of factors including Impact factor of the journal, how relevant, active & up to date the journals are for a data scientist.

Journals, Magazines in Analytics, Big Data, Data Mining



- ACM Transactions on Knowledge Discovery in Data (TKDD).
- SIGKDD Explorations, a magazine of the SIGKDD, the data miners professional group.
- Data Mining and Knowledge Discovery journal (now published by Springer).

- Analytics magazine from INFORMS.
- Big Data, open access peer-reviewed journal, provides a forum for world-class research exploring the challenges and opportunities in collecting, analyzing, and disseminating vast amounts of data. Liebert Publishers.
- Case Studies In Business, Industry And Government Statistics, electronic journal, Bentley University.
- Chance, a quarterly magazine for people interested in analysis of data, from American Statistical Association and Springer.
- Data Science Journal, published by the Committee on Data for Science and Technology (C(of the International Council for Science (ICSU).
- EPJ Data Science Journal, SpringerOpen.
- IEEE Transactions on Knowledge and Data Engineering
- Information Visualization, a central forum for all aspects of information visualization and i application (Palgrave MacMillan Journals).
- Intelligent Data Analysis journal (IOS Press).

- International Journal of Data Mining and Bioinformatics (IJDMB), ISSN (Online): 1748-5681 - ISSN (Print): 1748-5673
- Journal Of Big Data, a SpringerOpen Journal.
- Journal of Data Mining and Knowledge Discovery, tri-monthly, ISSN: 2229–6662 , 2229–6670, Bioinfo publications, India.
- Journal of Data Science, an international journal devoted to applications of statistical methods at large.
- Journal of Intelligent Information Systems.
- Journal of Machine Learning Research
- KAIS: Knowledge and Information Systems: An International Journal (Springer-Verlag)
- Machine Learning and Machine Learning Online
- Michael Ley comprehensive list of Computer Science and Database Journals
- Predictive Modeling News, the montly newsletter for healthcare professionals involved with predictive modeling.
- Statistical Analysis and Data Mining, Wiley journal, editors: Arnold Goodman, Chandrika Kamath, Vipin Kumar
- Transactions on Machine Learning and Data Mining, IBAI journal, Editor: Petra Perner, ISSN: 1865-6781.
- Wiley Interdisciplinary Reviews: Data Mining and KNowledge Discovery (WIDM), features a unique hybrid publishing model that's as comprehensive as a major reference, as current as a journal, and much more.

Conclusion

- Increasingly complex data science algorithms will continue to evolve in technologies and tools that will be easier to deploy (exponentially faster)
- Competitive organizations will adopt tools and create technologies to solve business problems
- Academic programs will expose students to data science principles (statistics, math, computer science, engineering and other related disciplines)
- Routine works will be automated/ semi-automated (transferred to less highly trained workers with just enough exposure to build basic models)

Questions?