Estatística, ciência de dados, machine learning, big data e outros gueri-gueris

Hedibert Freitas Lopes
Insper Institute of Education and Research
www.hedibert.org

3o Encontro de Data Science & Big Data
Centro Politénico, UFPR, 12/2024

É tempo de Botafogo! Glória Eterna

BOTAFOGO **CAMPEÃO DA LIBERTADORES 2024**

ALEJANDRO PAGNI / AFP

FOLHA DE S.PAULO

Outline

What is Statistics?

`https://www.stat.uci.edu/what-is-statistics`

Statistics is the science concerned with developing and studying methods for

- Collecting
- Analyzing
- Interpreting
- Presenting

empirical data.

Statistics is a highly interdisciplinary field.

Two fundamental idea: uncertainty and variation.

Probability: mathematical language to discuss uncertainty.

Statisticians attempt to understand/control variation.

What is data science?

`https://www.ibm.com/topics/data-science`

Data science combines

- Mathematics
- Statistics
- Specialized Programming
- Advanced Analytics
- Artificial Intelligence
- Machine Learning

with specific subject matter expertise to uncover actionable insights hidden in an organization's data.

These insights can be used to guide decision making and strategic planning.

20 Fastest Growing Occupations

US Bureau of Labor Statistics

`https://www.bls.gov/ooh/fastest-growing.htm`

Occupation	Growth rate 2023-2033	Annual median pay
Wind turbine service technicians	60%	61,770
Solar photovoltaic installers	48%	48,800
Nurse practitioners	46%	126,260
Data scientists	36%	**108,020**
Information security analysts	33%	120,360
Medical and health services managers	29%	110,680
Physician assistants	28%	130,020
Computer and information research scientists	26%	**145,080**
Physical therapist assistants	25%	64,080
Operations research analysts	23%	**83,640**
Occupational therapy assistants	22%	67,010
Actuaries	22%	**120,000**
Financial examiners	21%	84,300
Home health and personal care aides	21%	33,530
Veterinary assistants and laboratory animal caretakers	19%	36,440
Veterinary technologists and technicians	19%	43,740
Logisticians	19%	79,400
Veterinarians	19%	119,100
Substance abuse, behavioral disorder, and mental health counselors	19%	53,710
Epidemiologists	19%	81,390

Best Jobs

https://money.usnews.com/careers/best-jobs/rankings/the-100-best-jobs?sort=median-salary

Occupation	Projected Jobs	Annual median pay	Education Needed
Nurse Practitioner	118000	121000	Master's
Financial Manager	127000	140000	Bachelor's
Software Developer	410000	127000	Bachelor's
IT Manager	86000	164000	Bachelor's
Physician Assistant	39000	126000	Master's
Medical and Health Services Manager	145000	105000	Bachelor's
Information Security Analyst	53000	112000	Bachelor's
Data Scientist	**59400**	**104000**	**Bachelor's**
Actuary	**7000**	**114000**	**Bachelor's**
Speech-Language Pathologist	33000	84000	Master's
Marketing Manager	23700	140000	Bachelor's
Statistician	**10500**	**99000**	**Master's**
Management Analyst	96000	95000	Bachelor's
Genetic Counselor	600	90000	Master's
Operations Research Analyst	**25000**	**86000**	**Bachelor's**

Best Paying Jobs

Occupation	Projected Jobs	Annual median pay	Education needed
Anesthesiologist	1000	239000	Doctorate
Obstetrician and Gynecologist	500	239000	Doctorate
Oral and Maxillofacial Surgeon	200	239000	Doctorate
Psychiatrist	1900	227000	Doctorate
Nurse Anesthetist	4500	203000	Master's
Pediatrician	300	190000	Doctorate
Orthodontist	300	174000	Doctorate
IT Manager	86000	164000	Bachelor's
Dentist	6000	155000	Doctorate
Podiatrist	100	148000	Doctorate
Marketing Manager	23700	140000	Bachelor's
Financial Manager	127000	140000	Bachelor's
Industrial Psychologist	600	140000	Master's
Lawyer	62000	136000	Doctorate
Pharmacist	8700	133000	Doctorate
Sales Manager	22500	130000	Bachelor's
Political Scientist	400	128000	Master's
Software Developer	410000	127000	Bachelor's
Computer Network Architect	6300	127000	Bachelor's
Physician Assistant	39000	126000	Master's
Optometrist	3800	125000	Doctorate
Nurse Practitioner	118000	121000	Master's
Nurse Midwife	500	120000	Master's
Actuary	7000	114000	Bachelor's
Information Security Analyst	53000	112000	Bachelor's
Psychologist	2900	106000	Master's
Art Director	8200	105000	Bachelor's
Medical and Health Services Manager	145000	105000	Bachelor's
Biochemist	2300	104000	Doctorate
Data Scientist	59400	104000	Bachelor's
Veterinarian	18000	103000	Doctorate
Computer Systems Analyst	51000	102000	Bachelor's
Statistician	10500	99000	Master's

Jobs for Statisticians

(Project/Associate/Principal/Senior) Statistician
Data Analyst
Statistical modeler
(Senior) Statistical Programmer
Director, Statistics
Senior Scientist, Statistical Programming
Senior Statistics Manager
Statistics and Digital Innovations Lead
Predictive Analytics Lead
Senior Manager, Statistical Modeling
Associate Principal Scientist Statistical Programming
Senior Scientists, Statistical Programming

Jobs for Data Scientists

(Staff/Senior/Lead) Data Scientist
Analytics Assets, Data Scientist
Cognitive Data Scientist
Research Data Scientist, Senior
Senior Scientist
Manager - Cognitive Data Scientist Natural Language Processing
Manager, Data Scientist, NLP, Financial Services
Sr. Associate, NLP, Data Scientist
Associate, Data Scientist, Financial Services
Principal Data & Applied Scientist
Data Scientist - Machine Learning
Corporate Data Scientist Program Assessor

Statistician (2022)

`https://money.usnews.com/careers/best-jobs/statistician`

- Overall score 6.3/10
- #4 in best business jobs
- #9 in best STEM jobs
- #12 in 100 best jobs
- Number of jobs: 11000
- Annual salary quartiles ($000s): $(80, 100, 130)$
- Upward mobility: High
- Stress level: Below average
- Flexibility: Low

Data Scientist (2022)

`https://money.usnews.com/careers/best-jobs/data-scientist`

- Overall score 6.6/10
- #4 in best technology jobs
- #7 in best STEM jobs
- #8 in 100 best jobs
- Number of jobs: 60000
- Annual salary quartiles ($000s): $(80, 100, 140)$
- Upward mobility: High
- Stress level: Below average
- Flexibility: Average

Três profissões mais valiosas até o final da década

URL aqui.

Foram ouvidos 477 profissionais – entre estagiários e CEOs – de diferentes áreas sobre questões de trabalho no Brasil e suas reflexões para o futuro.

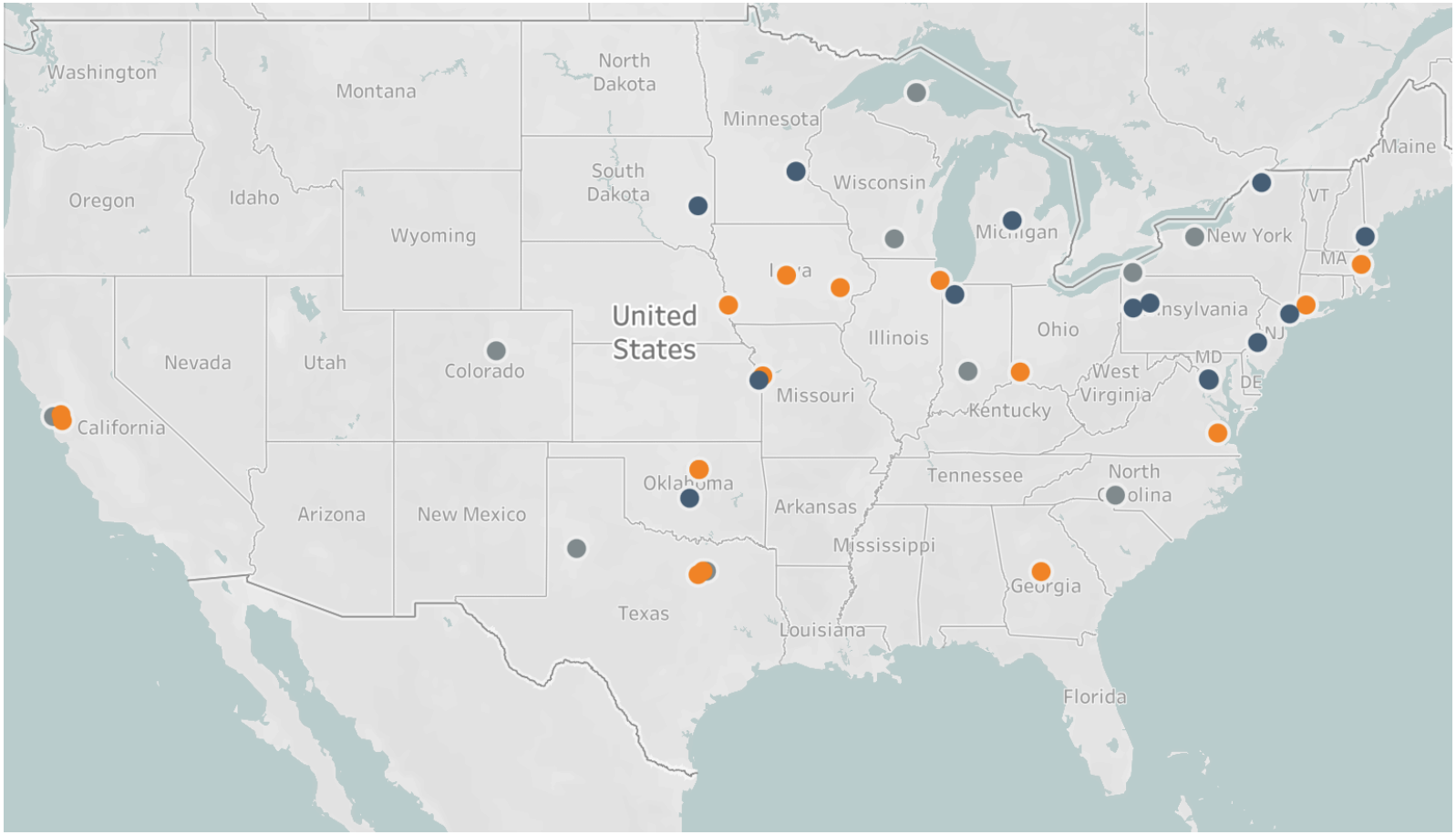
Segundo os entrevistados do setor de tecnologia, carreiras como

- Analista e Cientista de Dados (38%).
- Especialista em IA e Machine Learning (35%), and
- Analista de Segurança da Informação (31%) ,

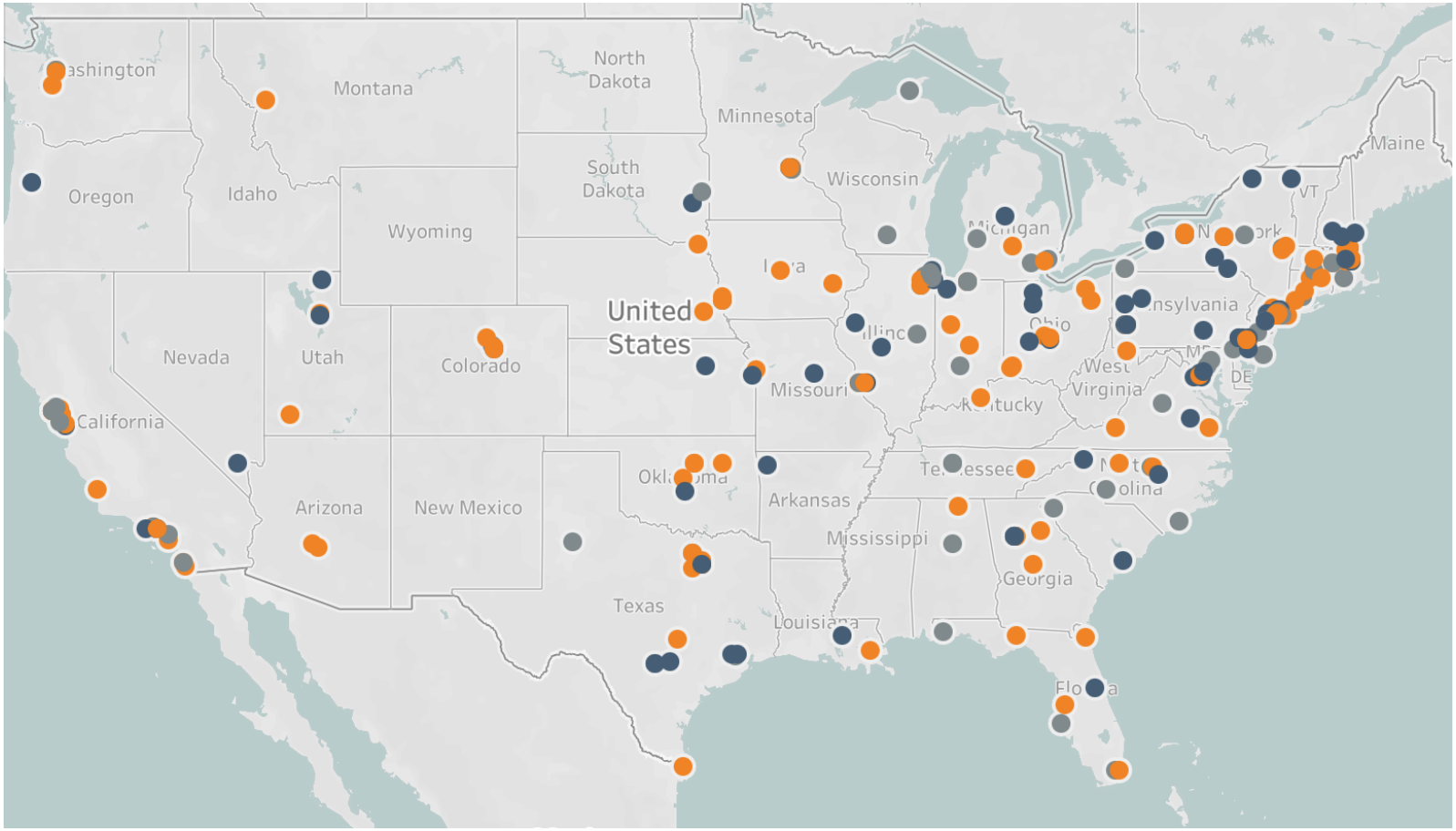
são as que mais se destacam.

2010: Analytics, Business Analytics & Data Science

2015


Washington
Montana
North Dakota
Minnesota
Maine
Oregon
Idaho
South Dakota
Wisconsin
VT
Wyoming
Michigan
New York
MA
United States
Illinois
Ohio
Nevada
Utah
Colorado
Missouri
West Virginia
MD
DE
NJ
California
Kentucky
Tennessee
Oklahoma
Arkansas
Arizona
New Mexico
Mississippi
Georgia
Texas
Louisiana
Florida

All


North Dakota
Montana
Minnesota
Maine
South Dakota
Wisconsin
Oregon
Idaho
Wyoming
United States
Nevada
Utah
Colorado
Missouri
California
Arizona
New Mexico
Arkansas
Mississippi
Georgia
Texas
Louisiana

Master in Data Science: 2007-2012 (12)

University	Degree	Credit	Established
North Carolina State University	Analytics	30	2007
University of Tennessee at Knoxville	Business Analytics	39	2010
Saint Joseph's University	Business Intelligence and Analytics	30	2010
Louisiana State University at Baton Rouge	Analytics	39	2011
University of Cincinnati	Business Analytics	35	2011
Northwestern University	Predictive Analytics	11	2011
Northwestern University	Analytics	11	2012
University of San Francisco	Analytics	35	2012
Drexel University	Business Analytics	45	2012
Fordham University	Business Analytics	30	2012
University of Michigan at Dearborn	Business Analytics	30	2012
Stevens Institute of Technology	Business Intelligence and Analytics	36	2012

Master in Data Science: 2013 (18)

University	Degree	Credit	Established
Harrisburg University of Science and Technology	Analytics	36	2013
Texas A&M University	Analytics	36	2013
Southern Methodist University	Applied Statistics and Data Analytics	36	2013
Arizona State University	Business Analytics	30	2013
Benedictine University	Business Analytics	64	2013
George Washington University	Business Analytics	33	2013
Michigan State University	Business Analytics	30	2013
New York University	Business Analytics	14	2013
Rensselaer Polytechnic Institute	Business Analytics	30	2013
University of Texas at Austin	Business Analytics	36	2013
Carnegie Mellon University	Computational Data Science	9	2013
Washington University in St. Louis	Customer Analytics	30	2013
Pace University	Customer Intelligence and Analytics	36	2013
City University of New York	Data Analytics	36	2013
Southern New Hampshire University	Data Analytics	12	2013
University of Maryland	Data Analytics	39	2013
Illinois Institute of Technology	Data Science	34	2013
New York University	Data Science	36	2013

Master in Data Science: 2014 (33)

University	Degree	Credit	Established
Bowling Green State University	Analytics	33	2014
Dakota State University	Analytics	30	2014
Georgia Institute of Technology	Analytics	36	2014
Georgia State University	Analytics	32	2014
University of Chicago	Analytics	11	2014
Villanova University	Analytics	33	2014
Saint Louis University	Applied Analytics	36	2014
Maryville University	Applied Statistics and Data Analytics	36	2014
Bentley University	Business Analytics	30	2014
Indiana University	Business Analytics	30	2014
Quinnipiac University	Business Analytics	33	2014
Southern Methodist University	Business Analytics	33	2014
University of Colorado Denver	Business Analytics	30	2014
University of Denver	Business Analytics	58	2014
University of Miami	Business Analytics	16	2014
University of Minnesota	Business Analytics	45	2014
University of Rochester	Business Analytics	41	2014
University of Southern California	Business Analytics	27	2014
University of Texas at Dallas	Business Analytics	36	2014
Creighton University	Business Intelligence and Analytics	33	2014
St. John's University	Data Mining and Predictive Analytics	30	2014
Elmhurst College	Data Science	30	2014
South Dakota State University	Data Science	30	2014
University of St. Thomas	Data Science	36	2014
University of Virginia	Data Science	11	2014
West Virginia University	Data Science	30	2014
Worcester Polytechnic Institute	Data Science	33	2014
Johns Hopkins University	Government Analytics	12	2014
University of California at Berkeley	Information and Data Science	27	2014
Philadelphia University	Modeling, Simulation and Data Analytics	30	2014
University of Arkansas	Statistics and Analytics	30	2014
Brandeis University	Strategic Analytics	30	2014
University of California, San Diego	Data Science and Engineering	38	2014

Master in Data Science: 2015 (25)

University	Degree	Credit	Established
Capella University	Analytics	48	2015
Georgetown University	Analytics	30	2015
University of New Hampshire	Analytics	36	2015
University of the Pacific	Analytics	30	2015
American University	Analytics Online	33	2015
Valparaiso University	Analytics and Modeling	36	2015
College of William&Mary	Business Analytics	30	2015
Fairfield University	Business Analytics	30	2015
Iowa State University	Business Analytics	30	2015
Mercer University	Business Analytics	30	2015
Northeastern University	Business Analytics	30	2015
University of Dallas	Business Analytics	30	2015
University of Iowa	Business Analytics	30	2015
University of Notre Dame	Business Analytics	30	2015
University of Texas at Arlington	Business Analytics	36	2015
Xavier University	Customer Analytics	30	2015
Clarkson University	Data Analytics	33	2015
Slippery Rock University	Data Analytics	33	2015
Columbia University	Data Science	30	2015
Indiana University Bloomington	Data Science	30	2015
Southern Methodist University	Data Science	31	2015
University of Rochester	Data Science	30	2015
University of Wisconsin's Extension	Data Science	36	2015
University of North Carolina at Charlotte	Data Science	33	2015
Penn State Great Valley	Data Analytics	30	2015

Bachelor Degree in Data Science

https://www.discoverdatascience.org/programs/bachelors-in-data-science

Bowling Green State University	Data Science Specialization
Brigham Young University	Data Science Major/Minor
Case Western Reserve University	Data Science and Analytics Major
Colorado State University	Data Science Major
Columbia University	BA in Data Science
DePaul University	Bachelor of Science Data Science
Drexel University	B.S. in Data Science
George Mason University	Bachelor of Science in Computational and Data Sciences
Indiana University	Bachelor of Science in Health Data Science
Marquette University	Data Science Major
Pennsylvania State University	Data Sciences - Intercollege Undergraduate Major
Purdue University	Data Science Major
Temple University	Data Science with Concentration in Computation and Modeling, B.S.
University of California Irvine	B.S. in Data Science
University of California San Diego	Data Science Major
University of Houston	Bachelor of Science in Data Science
University of Massachusetts Dartmouth	Major in Data Science
University of Michigan	Undergraduate Program in Data Science
University of New Hampshire at Manchester	Analytics and Data Science, B.S.
University of Rochester	BA and B.S. Data Science Degree
Valparaiso University	B.S. in Data Science
Yale University	Undergraduate Degree in Statistics and Data Science

DS Initiatives and Stats/DS departments

Brown Data Science Initiative
Columbia Data Science Initiative
Harvard Data Science Initiative
Irvine Data Science Initiative
Northwestern Data Science Initiative
Rice Data Science Initiative
Santa Barbara Data Science Initiative
Stanford Data Science Initiative
Toronto Data Science Initiative
Wisconsin-Madison Data Science Initiative

UT Austin - Department of Statistics and Data Sciences
Carnegie Mellon - Department of Statistics & Data Science
Cornell - Department of Statistics and Data Science
MIT - Statistics and Data Science Center
Wharton - Department of Statistics and Data Science
Yale - Department of Statistics and Data Science

Table I

Rankings vs Realities[1]

This study examines the competencies of data science graduate programs offered in the United States. It investigates whether there are any variations in competencies based on college and major rankings.

TABLE I. UNIVERSITY RANKING GROUPS

Group	U.S. News Ranking	N	Percentage
G1	1-75	77	46%
G2	76-171	45	27%
G3	172-225	12	7%
G4	226-300	21	13%
G5	>299	11	7%
Total		**166**	**100%**

[1]Li, Milonas and Zhang (2023) Rankings vs Realities: Exploring Competency Differences in Graduate Data Science Programs, *IEEE Frontiers in Education Conference*, pp. 1-4, `https://ieeexplore.ieee.org/document/10343290`.

Table II

The data science competencies employed in this study rely on the research conducted by the Association for Computing Machinary (ACM) Data Science Task Force, establishing a solid theoretical framework, which identified the following 11 data science competencies:

- 1) analysis and presentation,
- 2) artificial intelligence,
- 3) big data systems,
- 4) computing and computer fundamentals,
- 5) data acquisition, management and governance,
- 6) data mining,
- 7) data privacy, security, integrity, and analysis for security,
- 8) machine learning,
- 9) programming, data structures and algorithms,
- 10) software development and maintenance, and
- 11) professionalism.

TABLE II. COMPETENCIES, THEIR EQUIVALENT ACM TASK FORCE REPORT COMPETENCIES, AND SAMPLE COURSES

Our Competency Clusters	**ACM Data Science Competencies**	**Sample Courses**
Computing Fundamentals	4. Computing and Computer Fundamentals 9. Programming, data structures and algorithms 10. Software development and maintenance	SQL Programming, Introduction to Programming, Algorithms, Data Structures, Object Oriented Programming, Software Engineering, Systems Analysis and Design, Human-Computer Interaction
Data Management, Governance, Privacy	5. Data Acquisition, Management, and Governance 7. Data Privacy, Security, Integrity, and Analysis for Security 11.Professionalism	Data Warehousing, SQL, Databases, Security, Fraud Detection, Network Security, Ethics
Data Visualization	1. Analysis and Presentation	Data Visualization
Machine Learning	2. Artificial Intelligence 8. Machine learning	Machine Learning, Data Modeling, Artificial Intelligence, Deep Learning
Data Mining, Big Data	3. Big Data Systems 6. Data Mining	Data mining, Data modeling, systems analysis, Big Data, Data munging
Data Science in Context	11. Professionalism	Capstone, Internship, Senior Project, Courses in disciplines (physics, biology, chemistry, humanity, etc)
Math and Statistics		Calculus, discrete structures, probability theory, elementary statistics, advanced topics in statistics, and linear algebra.
Sensor and Sensor Networks		Sensor works, Foundamental of Sensors, Sensors and sensor systems

Table III

TABLE III. DATA SCIENCE MAJORS BY UNIVERSITY RANKINGS

	Major	G1	G2	G3	G4	G5	Total
1	Data Science	11%	8%	3%	4%	1%	**27%**
2	Business Analytics	10%	7%	1%	2%	1%	**21%**
3	Data Analytics	6%	3%	1%	1%	1%	**11%**
4	Biomedical	4%	0%	1%	0%	0%	**6%**
5	Math/Statistics	4%	2%	1%	1%	1%	**8%**
6	Computer Science	3%	3%	0%	1%	1%	**8%**
7	Info Science Tech	3%	3%	1%	3%	2%	**10%**
8	Public Policy	2%	0%	0%	0%	0%	**2%**
9	Health Informatics	1%	3%	0%	1%	0%	**4%**
10	Business Intelligence	1%	1%	0%	1%	0%	**3%**
11	Big Data	0%	0%	0%	1%	0%	**1%**
	Total	**44%**	**29%**	**8%**	**13%**	**7%**	**100%**

Table IV

TABLE IV. DATA SCIENCE DEPARTMENTS BY UNIVERSITY RANKINGS

	Departments /Schools	G1	G2	G3	G4	G5	Total
1	Business	12.2%	9%	3%	4%	3%	**30%**
2	Computer Science	5.8%	2%	1%	1%	1%	**10%**
3	Interdisciplinary	4.5%	3%	1%	3%	1%	**12%**
4	Info Science Tech	3.2%	3%	1%	2%	2%	**11%**
5	Health	4.5%	3%	1%	1%	0%	**10%**
6	Data Science	3.8%	3%	0%	1%	1%	**8%**
7	Engineering	3.8%	3%	0%	1%	0%	**7%**
8	Math/Statistics	3.8%	3%	1%	0%	0%	**8%**
9	Arts & Science	2.6%	1%	1%	0%	0%	**4%**

Table V

TABLE V. DATA SCIENCE COMPETENCIES BY UNIVERSITY RANKINGS

	Competency	G1	G2	G3	G4	G5	Total
1	Computing Fundamentals	34%	22%	5%	8%	6%	**76%**
2	Data Management, Governance, Privacy	28%	20%	7%	11%	6%	**72%**
3	Math and Statistics	27%	20%	5%	10%	5%	**67%**
4	Data Science in context	22%	19%	4%	8%	2%	**54%**
5	Data Mining, Big Data	16%	17%	3%	10%	4%	**51%**
6	Data Visualization	8%	5%	1%	4%	1%	**18%**
7	Machine Learning	7%	6%	1%	1%	0%	**16%**
8	Sensors and Sensor Networks	1%	1%	0%	1%	0%	**3%**

2023 ASA Statement on The Role of Statistics in Data Science and Artificial Intelligence - **URL here**

DS+AI have, in recent years, captured the attention of the world:

- the development of self-driving cars
- machines to recognize speech
- machines to generate human-like text
- technology that can accurately detect cancer

Interdisciplinary nature of DS+AI: statisticians – *who themselves are data scientists* – should be extensively involved in DS+AI initiatives to realize their full potential for productivity, innovation, and problem-solving.

In the past 20 years, DS+AI have rapidly evolved, fueled by the explosion of data and advancements in computing power.

Statistical and Computing

Development of sophisticated tools

- ML algorithms; Deep learning NN, and generative AI (+LLM)

that have revolutionized industries such as

- health care, finance, and marketing

plus new areas of research in statistics & computing.

The big tent of statistics has grown massively bigger, with the role of statisticians in this new and evolving world requiring adaptation.

The boundaries between statistical and computational methods have become more blurred, given the advances in ML and deep learning algorithms, which often require a deep understanding of both statistical theory and computer science.

This has led to a rethinking of the traditional role of statisticians in data science and a recognition that statisticians need to be equipped with a wider range of skills and expertise to remain effective and, indeed, relevant.

Statistical methods strengths

The central dogma of statistical inference, that there is a component of randomness in data, enables researchers to formulate questions in terms of underlying processes, quantify uncertainty in their answers, and separate signal from noise.

A statistical framework allows researchers to distinguish between causation and correlation, and thus to identify interventions that will cause changes in outcomes.

It also allows them to establish methods for prediction and estimation, to quantify their degree of certainty, and to do it all using algorithms that exhibit predictable and reproducible behavior.

Simply put, statistical methods enhance researchers' abilities to accumulate knowledge.

Collaborative efforts

Expertise in data organization, distributed computation, and model lifecycle management.

Statisticians must work with them, learn from them, and teach them.

Engagement must occur at all levels – with individuals, groups of researchers, academic departments, and the profession as a whole.

New problem-solving strategies are needed to develop end-to-end data science and ML operations pipelines, from

- raw data collection and management to
- model monitoring/retraining and governance to
- user-friendly implementations of principled statistical methods and
- the communication of substantive results.

Machine learning toolbox

- Linear regression
- Logistic regression
- Decision tree
- Support vector machines
- Naive Bayes
- K nearest neighbours
- K-means
- Random forest
- Dimensionality reduction algorithms
- Gradient boost & adaboost

Source: Analytics Vidhya

Machine learning toolbox

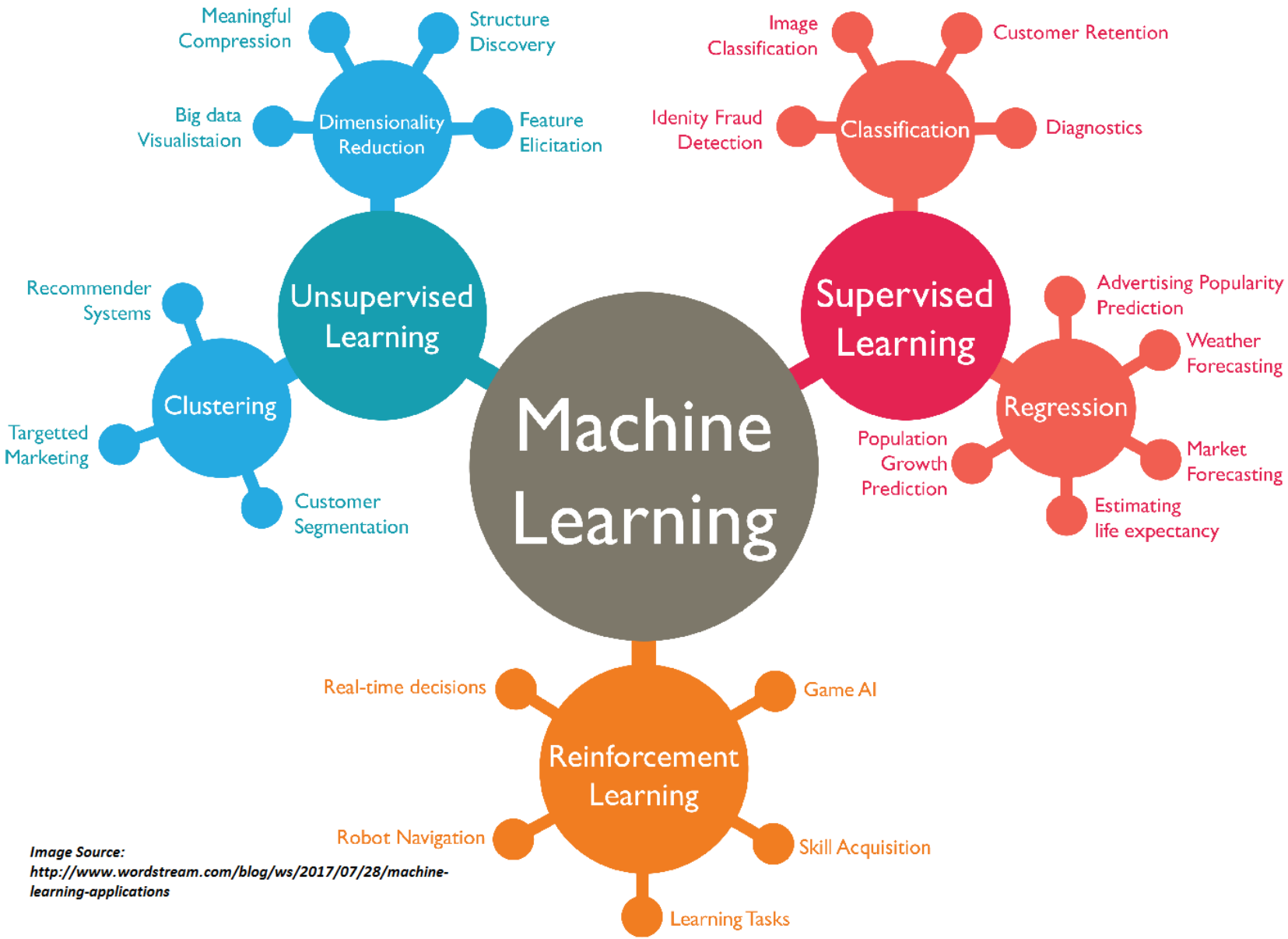


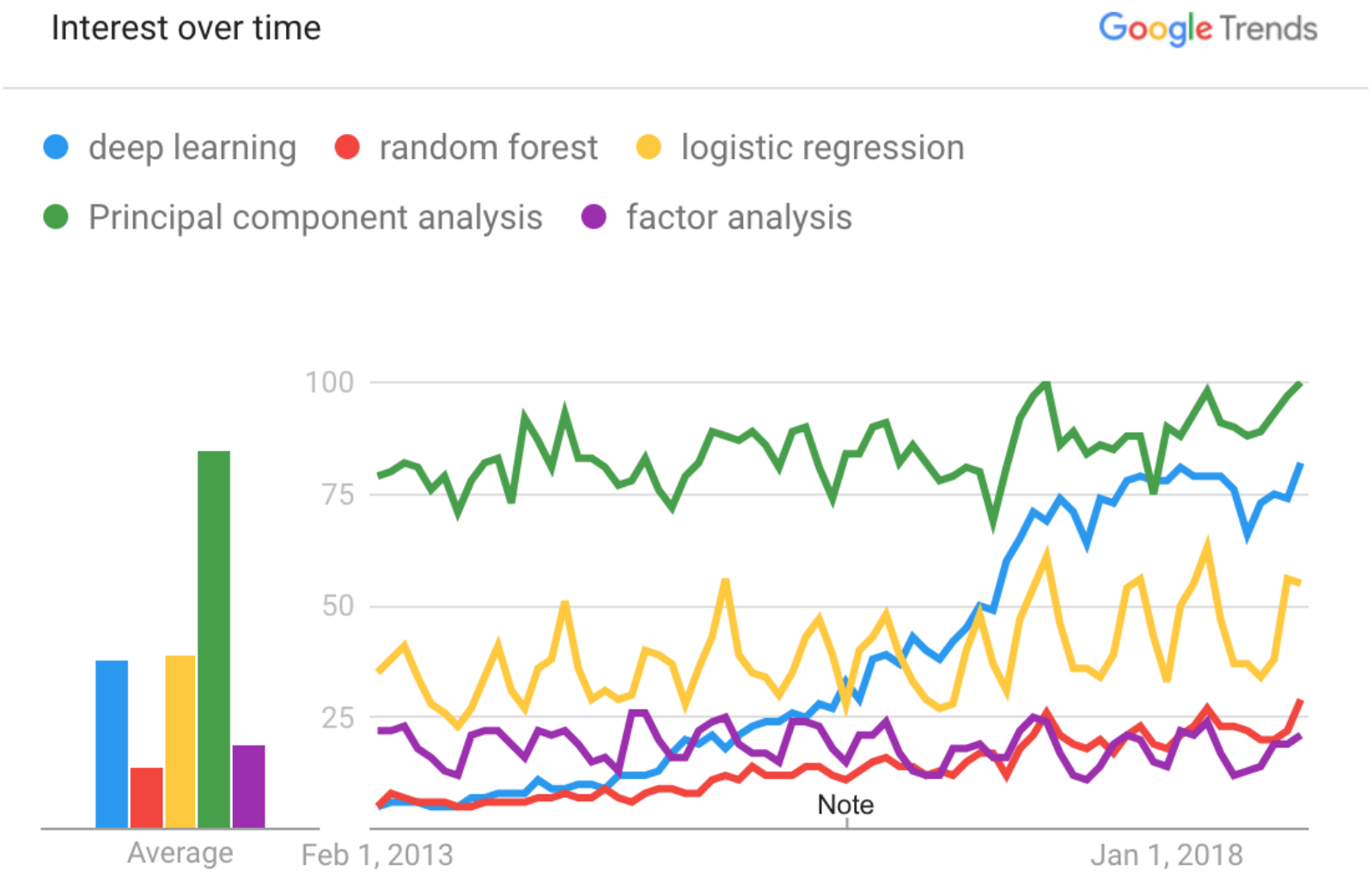
Image Source:
http://www.wordstream.com/blog/ws/2017/07/28/machine-learning-applications

Funny Glossary

Glossary

Machine learning	Statistics
network, graphs	model
weights	parameters
learning	fitting
generalization	test set performance
supervised learning	regression/classification
unsupervised learning	density estimation, clustering
large grant = \$1,000,000	large grant= \$50,000
nice place to have a meeting: Snowbird, Utah, French Alps	nice place to have a meeting: Las Vegas in August

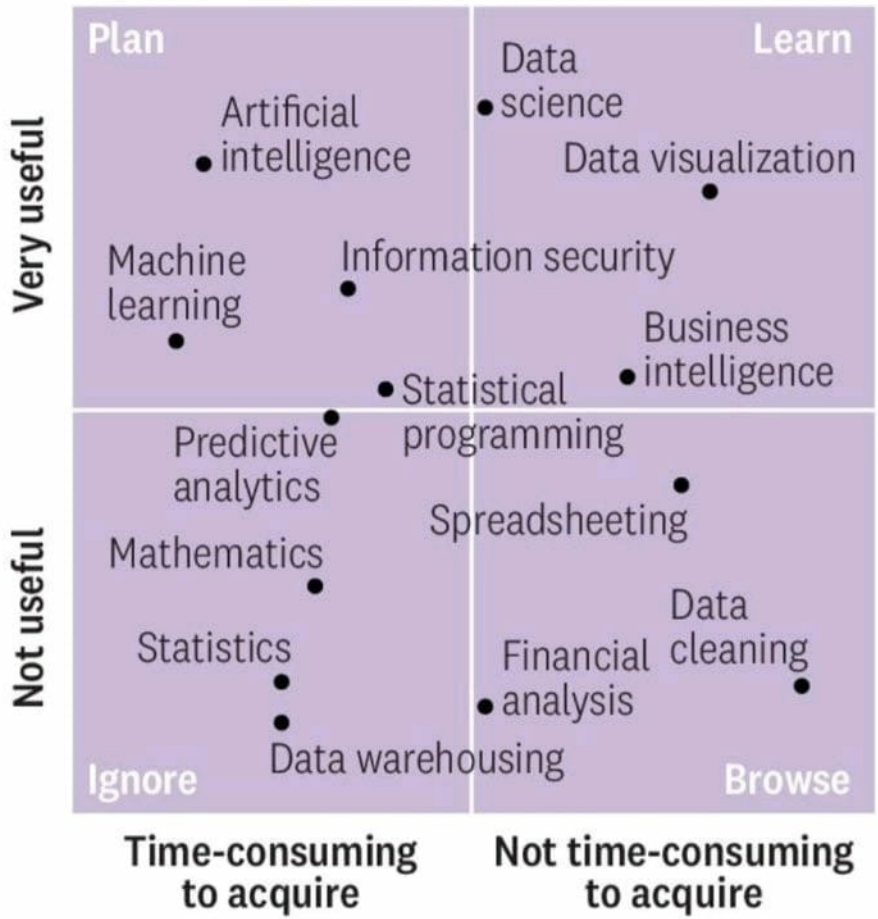
Popular statistical methods



Harvard Business Review

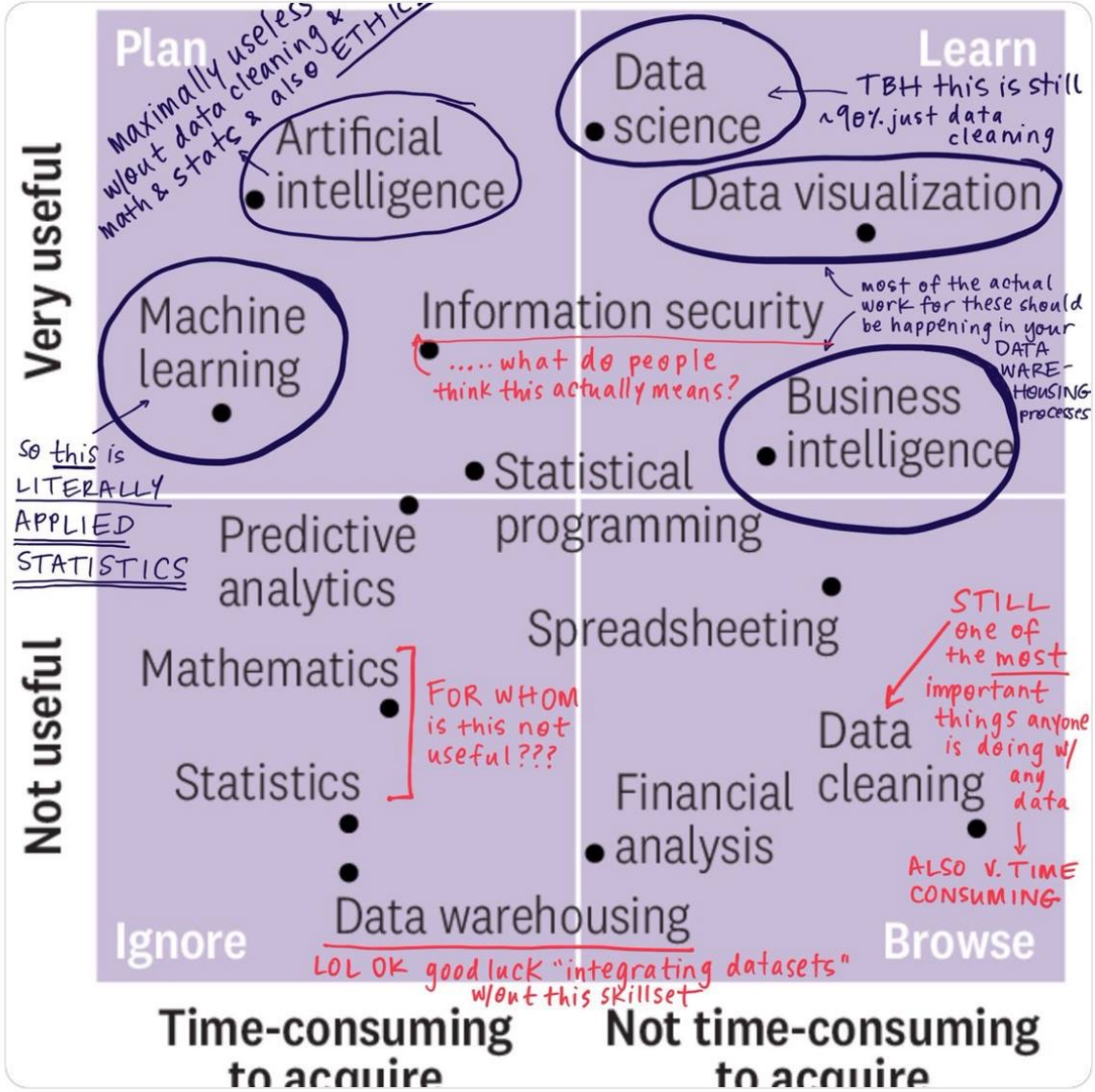
An Example of How to Plot Data Skills on a 2x2 Learning Matrix

How one company mapped its own internal learning needs.



Idk who made @HarvardBiz's "data skills" matrix but it's mostly terrible so I made some updates (hbr.org/2018/10/which- …) (cc: @KansasCityEric)

Traducir Tweet



The Role of Statistics in the Era of Big Data

Statistics & Probability Letters

SUPPORTS OPEN ACCESS

Articles in press Latest issue Special issues All iss

The role of Statistics in the era of big data

Edited by Laura Sangalli

Volume 136, Pages 1-170 (May 2018)

Statistics in the big data era: Failures of the machine

This special issue has been stimulated by a plenary lecture by David Dunson (Duke University), at the 2016 Meeting of the Italian Statistical Society.

Dunson's abstract:

> *There is vast interest in automated methods for complex data analysis.*
>
> *However, there is a lack of consideration of*
>
> *(1) interpretability,*
> *(2) uncertainty quantification,*
> *(3) applications with limited training data, and*
> *(4) selection bias.*
>
> *Statistical methods can achieve (1)-(4) with a change in focus.*

Greater (DS) and Lesser Statistics (Methodology)[2]

Greater statistics can be defined ... as everything related to learning from data, from the first planning or collection to the last presentation or report.

Lesser statistics is the body of specifically statistical methodology that has evolved within the profession – roughly, statistics as defined by texts, journals, and doctoral dissertations.

Greater statistics tend to be inclusive, eclectic with respect to methodology, closely associated with other disciplines, and practiced by many outside of academia and often outside professional statistics.

Lesser statistics tends to be exclusive, oriented to mathematical techniques, less frequently collaborative with other disciplines, and primarily practiced by members of university departments of statistics.

---

[2]Chambers (1993) Greater or lesser statistics: A choice for future research. *Statistics and Computing*, 3(4), 182-184.

Data science vs. statistics: two cultures?[3]

[W]e define data science as the union of six areas of greater data science, based on Donoho (2017) 50 years of data science. Journal of Computational and Graphical Statistics, *26(4), 745-766:*

1. *Data gathering, preparation, and exploration.*
2. *Data representation and transformation.*
3. *Computing with data.*
4. *Data modeling.*
5. *Data visualization and presentation.*

We take the position that data science is a reaction to the narrow understanding of lesser statistics; simply put, data science has come to mean a broader view of statistics.

[3] Carmichael and Marron (2018) *Japanese Journal of Statistics and Data Science*, 1, 117-138. `https://doi.org/10.1007/s42081-018-0009-3`

Michael Jordan on ML vs Statistics

Throughout the eighties and nineties, it was striking how many times people working within the "ML community" realized that their ideas had had a lengthy pre-history in statistics.

Decision trees, nearest neighboor, logistic regression, kernels, PCA, canonical correlation, graphical models, K-means and discriminant analysis come to mind, and also many general methodological principles (e.g., method of moments, Bayesian inference methods of all kinds, M estimation, bootstrap, cross-validation, EM, ROC, and stochastic gradient descent), and many many theoretical tools (large deviations, concentrations, empirical processes, Bernstein-von Mises, U statistics, etc).

Source: reddit machine learning blog

Michael Jordan (more)

When Leo Breiman developed random forests, was he being a statistician or a machine learner?

When my colleagues and I developed latent Dirichlet allocation, were we being statisticians or machine learners?

Are the SVM and boosting machine learning while logistic regression is statistics, even though they're solving essentially the same optimization problems?

I think the ML community has been exceedingly creative at taking existing ideas across many fields, and mixing and matching them to solve problems in emerging problem domains, and I think that the community has excelled at making creative use of new computing architectures.

I would view all of this as the proto emergence of an engineering counterpart to the more purely theoretical investigations that have classically taken place within statistics and optimization.

Michael Jordan (a bit more)

But one shouldn't definitely not equate statistics or optimization with theory and machine learning with applications.

The "statistics community" has also been very applied, it's just that for historical reasons their collaborations have tended to focus on science, medicine and policy rather than engineering.

The emergence of the "ML community" has helped to enlargen the scope of "applied statistical inference". It has begun to break down some barriers between engineering thinking (e.g., computer systems thinking) and inferential thinking. And of course it has engendered new theoretical questions.

Statistics, data sciences, machine learning, big data

John Tukey (1962) The future of data analysis
David Hand (2013) Data mining: statistics and more?
Marie Davidian (2013) Aren't we data science?
Hal Varian (2014) Big data: new tricks for econometrics
Einav and Levin (2014) Economics in the age of big data
Athey and Imbens (2015) Lectures on machine learning
David Donoho (2015) 50 years of data science
Peter Diggle (2015) Statistics: a data science for the 21st century
van Dyk *et al.* (2015) Role of statistics in data science
Francis Diebold (2016) Machine learning versus econometrics
Uchicago (2016) Machine learning: what's in it for economics?
Coveney, Dougherty, Highfield (2016) Big data need big theory too
Franke *et al.* (2016) Statistical Inference, Learning and Models in Big Data

AMSTAT NEWS

Davidian (1 jul 2013) Aren't we data science?

Bartlett (1 oct 2013) We are data science

Matloff (1 nov 2014) Statistics losing ground to computer science

van Dyk *et al.* (1 oct 2015) Role of statistics in data science

Jones (1 nov 2015) The identity of statistics in data science

Priestley (1 jan 2016) Data science: the evolution or the extinction of statistics?

See also Press (28 may 2013) A very short history of data science

Doris Fontes: Twitter discussion

Num post de 15/11/2024, Doris Fontes[4] divide conosco uma frustração da área de estatística:

> *Estamos ouvindo vários comentários sobre o aumento significativo de procura por cursos de Ciência da Computação, ou Engenharia de Dados... Já pelo Bacharelado em Estatística...*
>
> *Que pena! Curso que poderia ter muito mais formandos, mas, infelizmente bastante desconhecido pelos alunos de Ensino Médio.*
>
> *Nas últimas feiras de profissão que participei, havia muitos curiosos sobre Ciência da Computação, mas pouquíssimos sabiam que existia um Bacharelado em Estatística.*
>
> *No ano passado, durante a Feira de Profissões, o IME e o ICMC ficaram juntos no mesmo estande. O impacto do nome* **Estatística e Ciência de Dados** *foi evidente. Muitos pais de jovens estudantes do ensino médio nos procuraram para saber mais sobre essa tal de "ciência de dados".*

[4] Tesoureira do CONRE-3, Conselho Regional de Estatística (SP)

Francisco Louzada[5] complementou:

Realmente é uma pena que o Bacharelado em Estatística não tenha a mesma visibilidade que outros cursos.

Seria excelente se mais cursos de estatística do país aderissem ao movimento de modernização do nome, chamando-se Bacharelado em Estatística e Ciência de Dados ou Bacharelado em Ciência de Dados e Estatística.

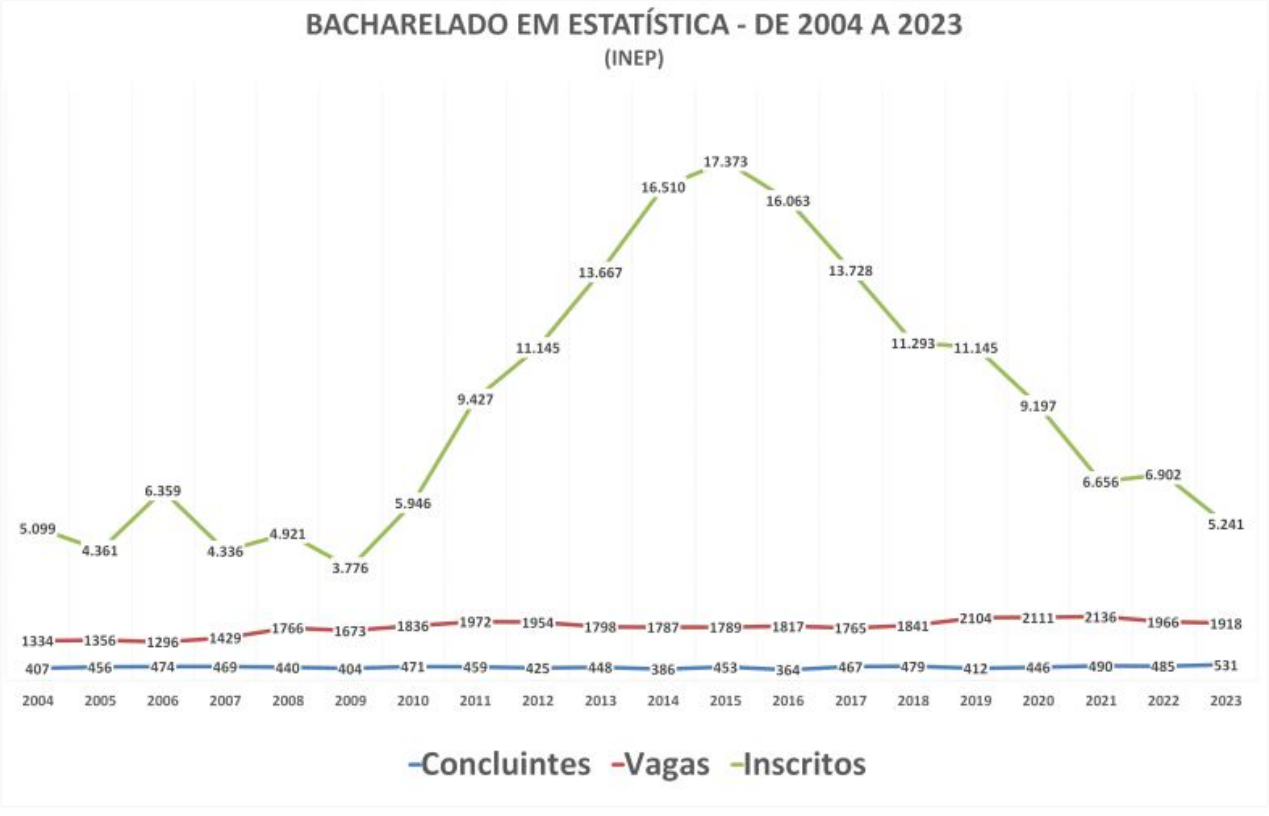
Por hora só temos dois: ICMC-USP *e* UFPR*.*

Isso poderia atrair mais atenção dos alunos do Ensino Médio, destacando a relevância e a modernidade do curso, além de abrir novas oportunidades para os formandos em um mercado cada vez mais orientado por dados.

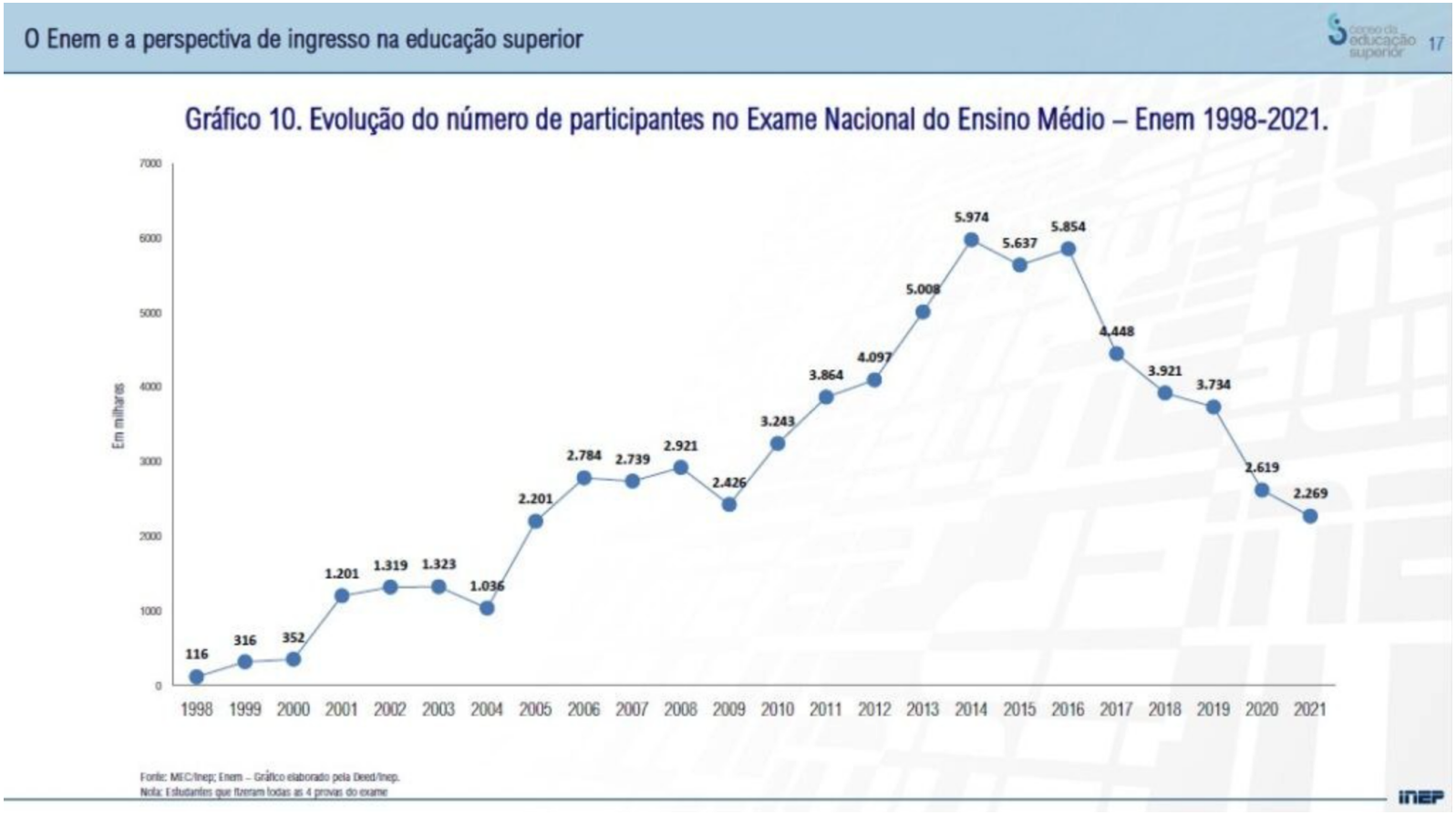
[5] Diretor do MBA em Ciência de Dados na CeMEAI, ICMC/USP

Doris Fontes - Gráfico 1

2004–2009: 5000 inscritos.
2009–2015: 5000 $\longrightarrow$ 17000
2015–2023: 17000 $\longrightarrow$ 5000
1300–2000 vagas por ano.
400–550 concluintes por ano.

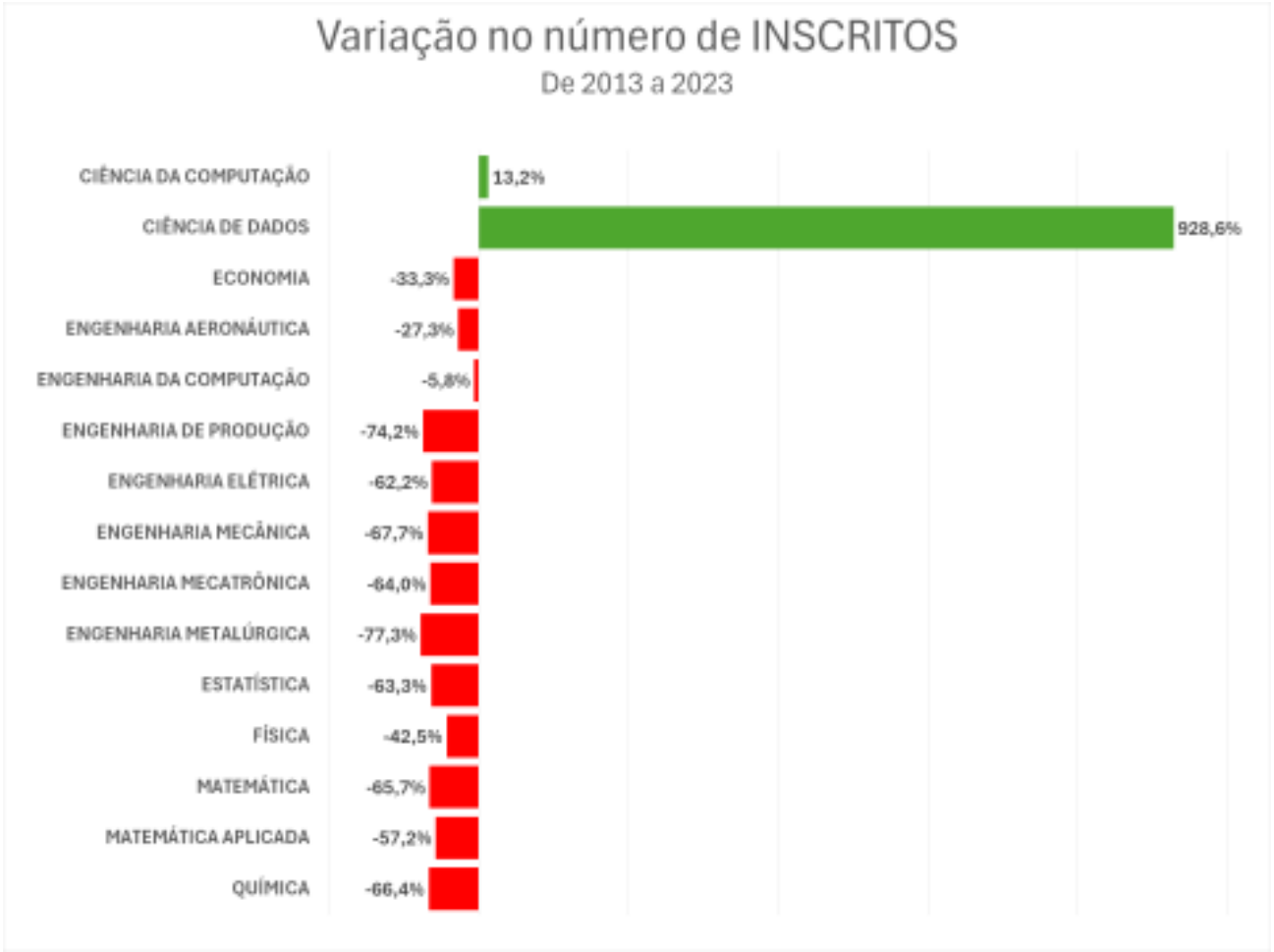


Felipe Bocca observa que “Pelo menos sobre o padrão de inscritos, eu diria que segue a mesma tendência de participantes do ENEM”.



Doris Fontes - Gráfico 2

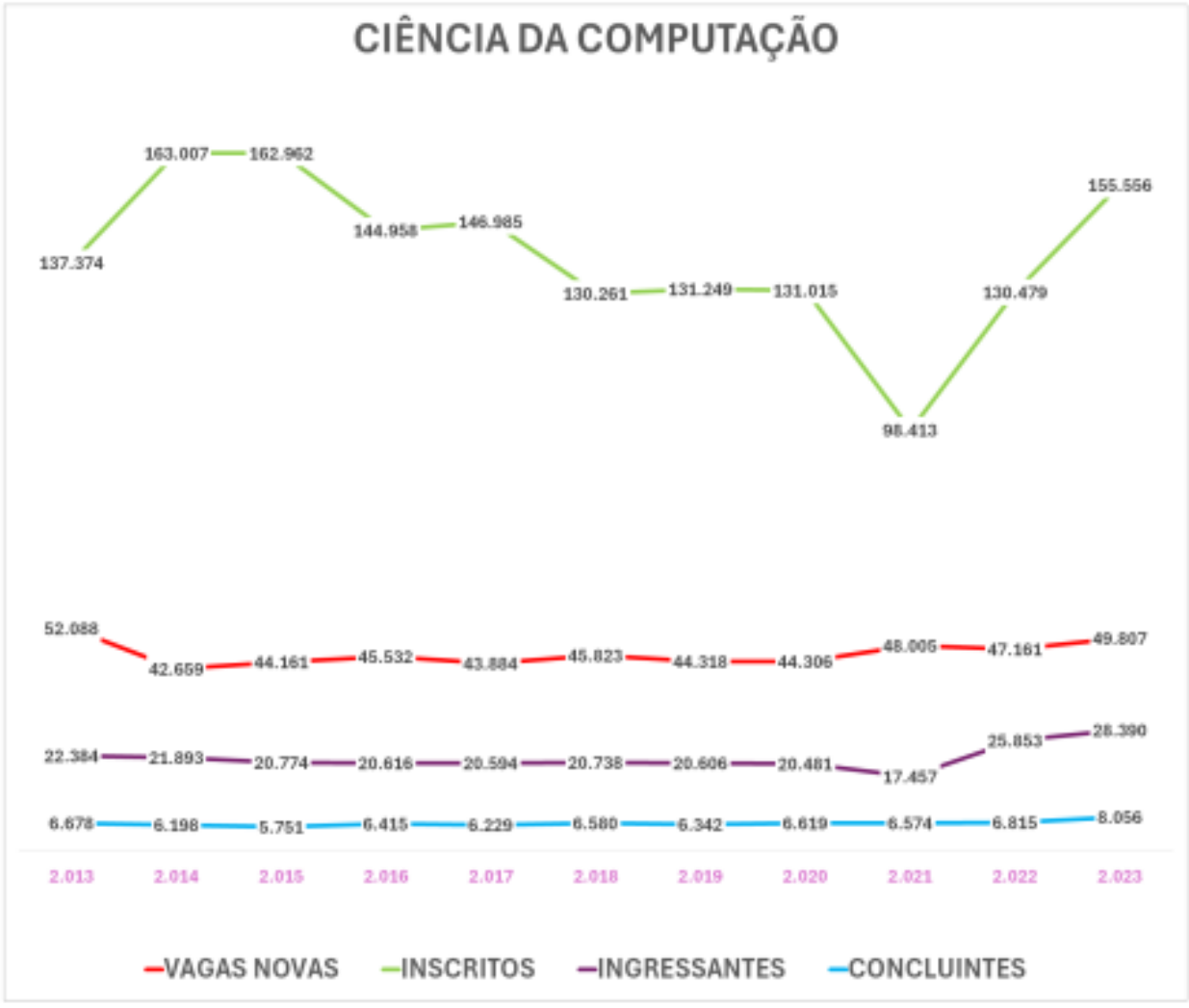
Obviamente 928% é efeito de muito poucas vagas em 2013!
Entretanto, as quedas nas várias outras áreas é marcante.



Doris Fontes - Gráfico 3

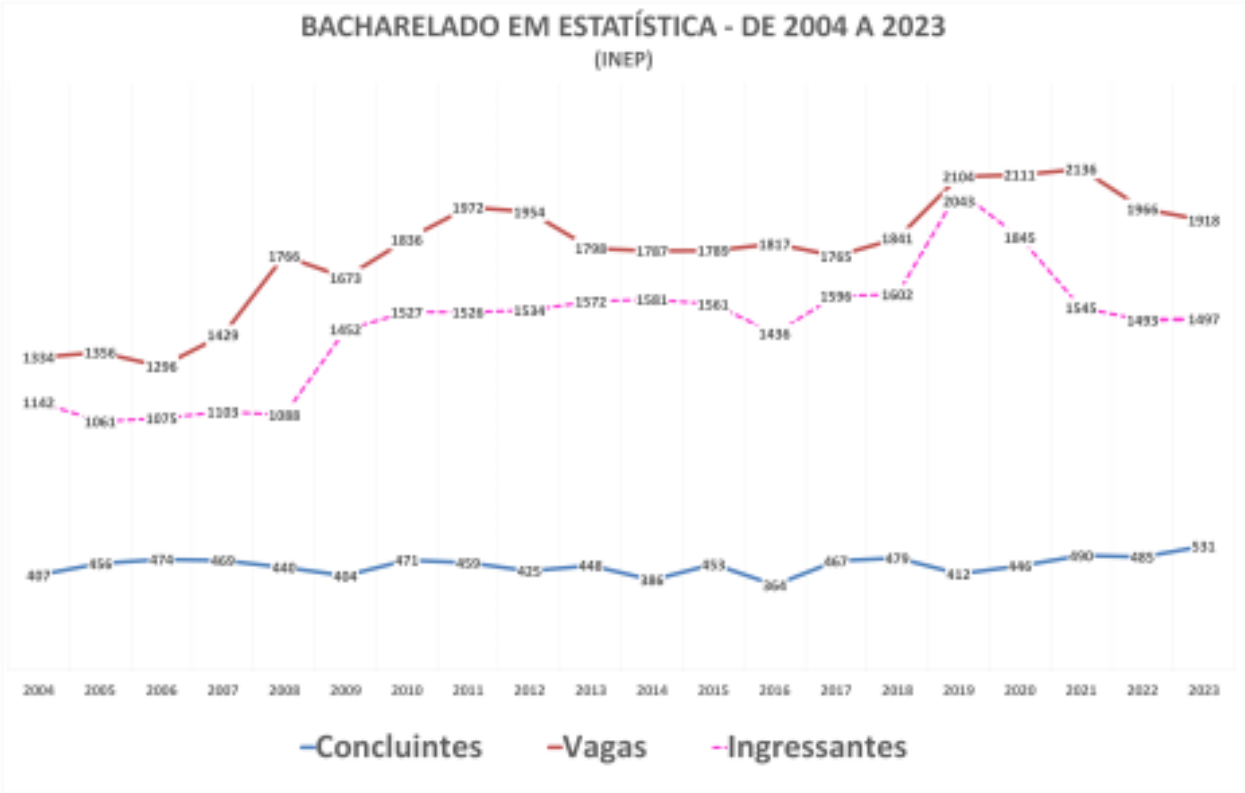
Entre 6000 e 8000 concluintes por ano.

15 vezes mais concluintes por ano do que na estatística!



Doris Fontes - Gráfico 4

Mesmo comparando concluintes em t com ingressantes em $t-4$, nota-se que somente 1/3 dos alunos concluem o bacharelado em estatística. O mesmo ocorre para ciências da computação.



Núcleo de Ciências de Dados e Decisão do INSPER

URL aqui: Pesquisa teórica e prática de métodos estatísticos e econométricos para oferecer respaldo de excelência em ciência de dados e auxiliar na tomada de decisão na presença de incerteza.

Seminário de Ciências de Dados e Decisão: Encontro mensal e gratuito aberto a todos os interessados no tema, que conta com a participação de pesquisadores em Estatística, Ciência de Dados e áreas correlatas.

Trilha de Ciência de Dados na graduação: duração de 320 horas: modelagem preditiva e R4DS e Py4DS, informação georreferenciada, big data para dados públicos e outros temas avançados.

Hedibert Lopes e Paulo Marques foram responsáveis pela elaboração do **Programa Avançado em Data Science (PADS)**, curso de pós-graduação lato sensu do Insper.

Membros

Coordenador:

Hedibert F. Lopes - Ph.D. em Estatística - Duke University

Demais membros:

- Rinaldo Artes - Doutor em Estatística - USP
- Adriana Bruscato Bortoluzzo - Doutora em Estatística - USP
- Paulo Marques - Doutor em Estatística - IME-USP
- José Heleno Faro - Doutor em Economia Matemática - IMPA

Jovens doutores em estatística do IME-USP

- Tiago Mendonça
- Yasmin Cavalieri
- Julio Trecenti
- Magno Severino

Programa Avançado em Data Science (PADS)[6]

1° trimestre:

Aprendizagem Estatística de Máquina I

Computação para Ciência de Dados

2° trimestre:

Aprendizagem Estatística de Máquina II

Big Data e Computação em Nuvem

3° trimestre:

Prática Avançada de Data Science e Visualization

Data Science Deploy

4° trimestre:

Financial analytics

Marketing analytics

[6] https://www.insper.edu.br/pt/cursos/pos-graduacao/programas-avancados/programas-avancados-data-science-e-decisao

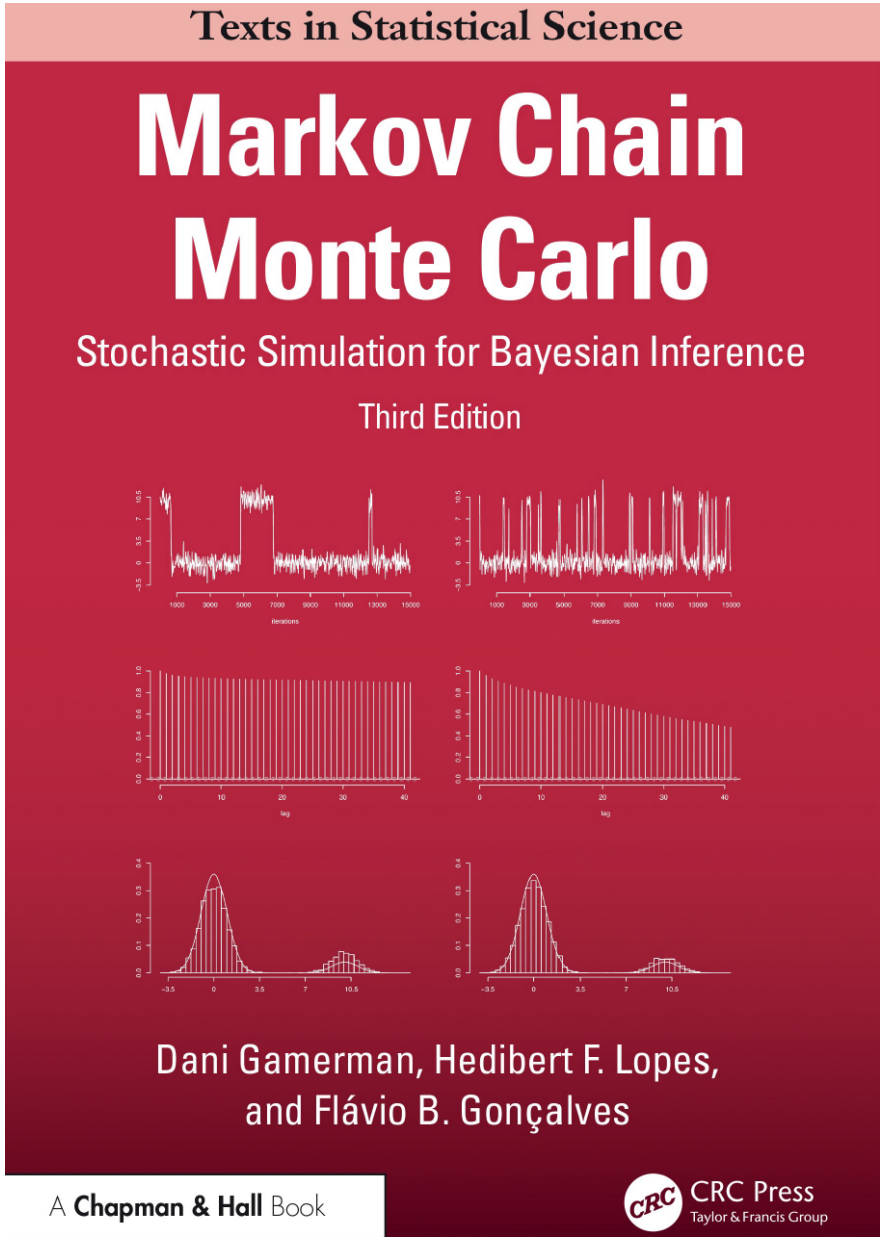
Mestrado Profissional em Ciências de Dados e Decisão (MPCDD) - Lançamento em 2026/2027

- Disciplinas obrigatórias
  - 1 Inferência Estatística
  - 2 Probabilidade e Processos Estocásticos
  - 3 Processamento de Dados em Escala
  - 4 Sistemas de base de dados distribuídos
  - 5 Introdução a Aprendizagem Estatística de Máquina
- Concentração: Modelagem Estatística de Máquina
  - 6 Estatística Bayesiana
  - 7 Estatística Computational
  - 8 Análise de Séries Temporais
  - 9 Inferência Causal
- Concentração: Computação e Decision Analytics
  - 6 Otimização
  - 7 Aprendizagem Estatística de Máquina Avançada
  - 8 Sistemas de Administração Base de Dados Avançado

Inspiração: MS Data Science, Analytics and Engineering (DSAE), Arizona State University (ASU)

MCMC 3rd edition coming in 2026!

`https://www.dme.ufrj.br/mcmc`



Muito obrigado!