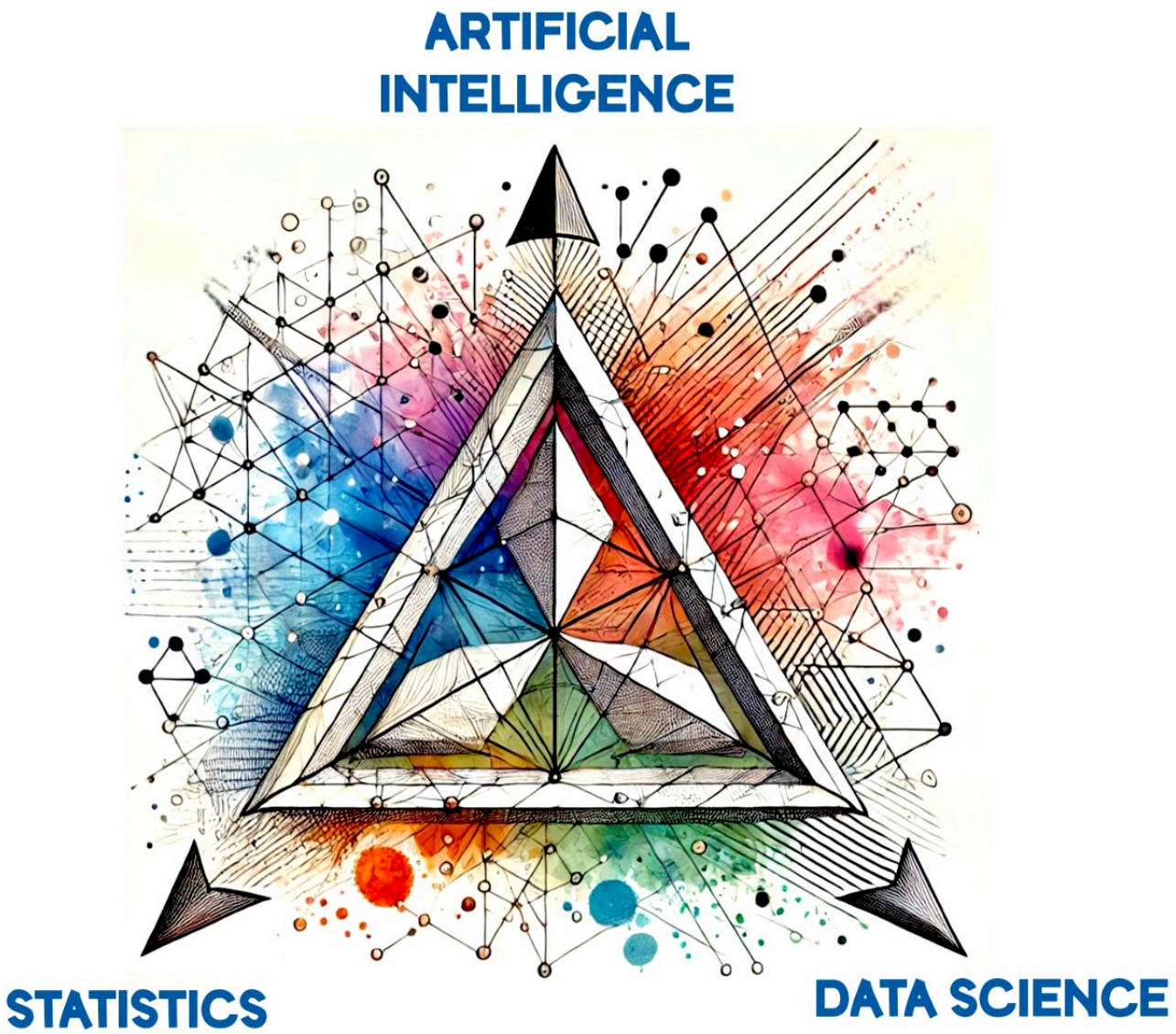
Data Science, Artificial Intelligence, and challenges for Statistics

Fabrizio Cipollini
Anna Gottard

November 14, 2024


ARTIFICIAL INTELLIGENCE
STATISTICS
DATA SCIENCE

Statistics vs Data Science

DATA SCIENCE CODE OF PROFESSIONAL CONDUCT

(b) "Data Science" means the scientific study of the creation, validation and transformation of data to create meaning.

(o) "Statistics" means the practice or science of collecting and analyzing numerical data in large quantities.

Are they really different?

Thumbs up!

- *Statistics emphasizes quantitative data and description. In contrast, data science deals with quantitative and qualitative data (e.g., from images, text, sensors, transactions, customer information, etc.) and emphasizes prediction and action.*
  (Vasant Dhar, Data Scientist, NYU; "Data Science and Prediction", Communications of the ACM, 2013)
- *Data Science without statistics is possible, even desirable.*
  (Vincent Granville, Statistician, `https://mltechniques.com/`; Quora Blog, 2013)
- *Statistics is the least important part of data science.*
  (Andrew Gelman, Statistician, Columbia University; Blog 2013)

Thumbs down!

- *I think data-scientist is a sexed up term for a statistician. Data scientist is slightly redundant, in some way.* (Nate Silver, Statistician, 2013)
- *A grand debate: is data science just a "rebranding" of statistics?* (Martin Goodson, Royal Statistical Society, 2015)
- *Aren't we Data Science?* (ASA President Marie Davidian, AmStat News, 2013)
- *Let us own Data Science* (IMS President Bin Yu, IMS bulletin, 2014)
- *Why Do We Need Data Science When We've Had Statistics for Centuries?* (Irving Wladawsky-Berger, Physicist, Wall Street Journal, 2014)
- *Data Science is statistics!* (Karl Broman, Statistician, Univ. Wisconsin, 2013)

What happened?

- **John Tukey** (Mathematician and Statistician, Princeton University) 1962: The Future of Data Analysis ▸ Tukey, 1962
- **John Chambers** (Statistician, A&T Bell Labs; S language) 1993: Greater or Lesser Statistics: A Choice for Future Research ▸ Chambers, 1993
- **C. F. Jeff Wu** (Statistician, University of Michigan) 1993: Statistics = Data Science? ▸ Wu, 1993
- **Ross Ihaka, Robert Gentleman** (Statisticians, University of Auckland) 1996: R: A Language for Data Analysis and Graphics ▸ Ihaka and Gentleman, 1996
- **William S. Cleveland** (Statistician, A&T Bell Labs) 2001: Data Science: An Action Plan for Expanding the Technical Areas of the Field of Statistics ▸ Cleveland, 2001
- **Leo Breiman** (Matematician and Statistician, University of California, Berkeley) 2001: Statistical Modeling: The Two Cultures ▸ Breiman, 2001

▸ Next chapter

Tukey, 1962

Tukey, John (1962). The Future of Data Analysis. **Annals of Mathematical Statistics**, 33, 1-67.

Tukey, 1962

For a long time I have thought I was a statistician, interested in inferences from the particular to the general. But as I have watched mathematical statistics evolve, I have had cause to wonder and to doubt. All in all I have come to feel that my central interest is in data analysis, which I take to include, among other things: procedures for analyzing data, techniques for interpreting the results of such procedures, ways of planning the gathering of data to make its analysis easier, more precise or more accurate, and all the machinery and results of (mathematical) statistics which apply to analyzing data.

Back

Chambers, 1993

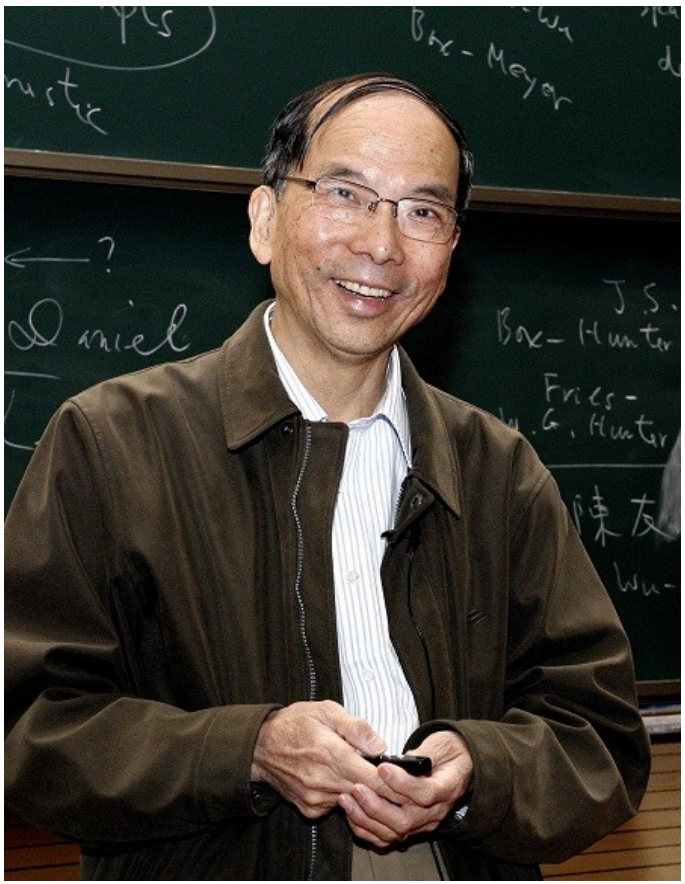
Chambers, John M. (1993). Greater or lesser statistics: a choice for future research. **Statistics and Computing**, 3, 182-184.

Chambers, 1993

The statistics profession faces a choice in its future research between continuing concentration on traditional topics, based largely on data analysis supported by mathematical statistic, and a broader viewpoint, based on an inclusive concept of learning from data. The latter course presents severe challenges as well as exciting opportunities. The former risks seeing statistics become increasingly marginal.

Back

Wu, 1993



Wu, C. F. Jeff (1993). Statistics = Data Science? Inauguration lecture as Carver Professor of Statistics at University of Michigan.

Wu, 1993

- Current state of statistics: trilogy of data collection, data modeling and analysis, and decision making.
- Promising current/future directions: large/complex data; use of Neural Networks; *"Think big, learn from others!"*
- Statistics to be renamed data science, and statisticians data scientists

Back

Ihaka and Gentleman, 1996

Ihaka, Ross, Gentleman, Robert (1996). R: A Language for Data Analysis and Graphics. **Journal of Computational and Graphical Statistics**, 5, 299-314. Back

Cleveland, 2001

Cleveland, William S. (2001). Data Science: An Action Plan for Expanding the Technical Areas of the Field of Statistics. **International Statistical Review**, 69, 21-26.

Cleveland, 2001

An action plan to expand the technical areas of statistics focuses on the data analyst. The plan sets out six technical areas of work for a university department and advocates a specific allocation of resources devoted to research in each area and to courses in each area. The value of technical work is judged by the extent to which it benefits the data analyst, either directly or indirectly.

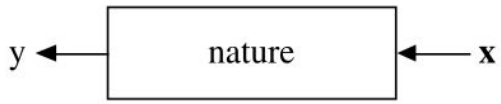
Back

Breiman, 2001

Breiman, Leo (2001). Statistical Modeling: The Two Cultures. **Statistical Science**, 16, 199-231.

Breiman, 2001

Think of the data as being generated by



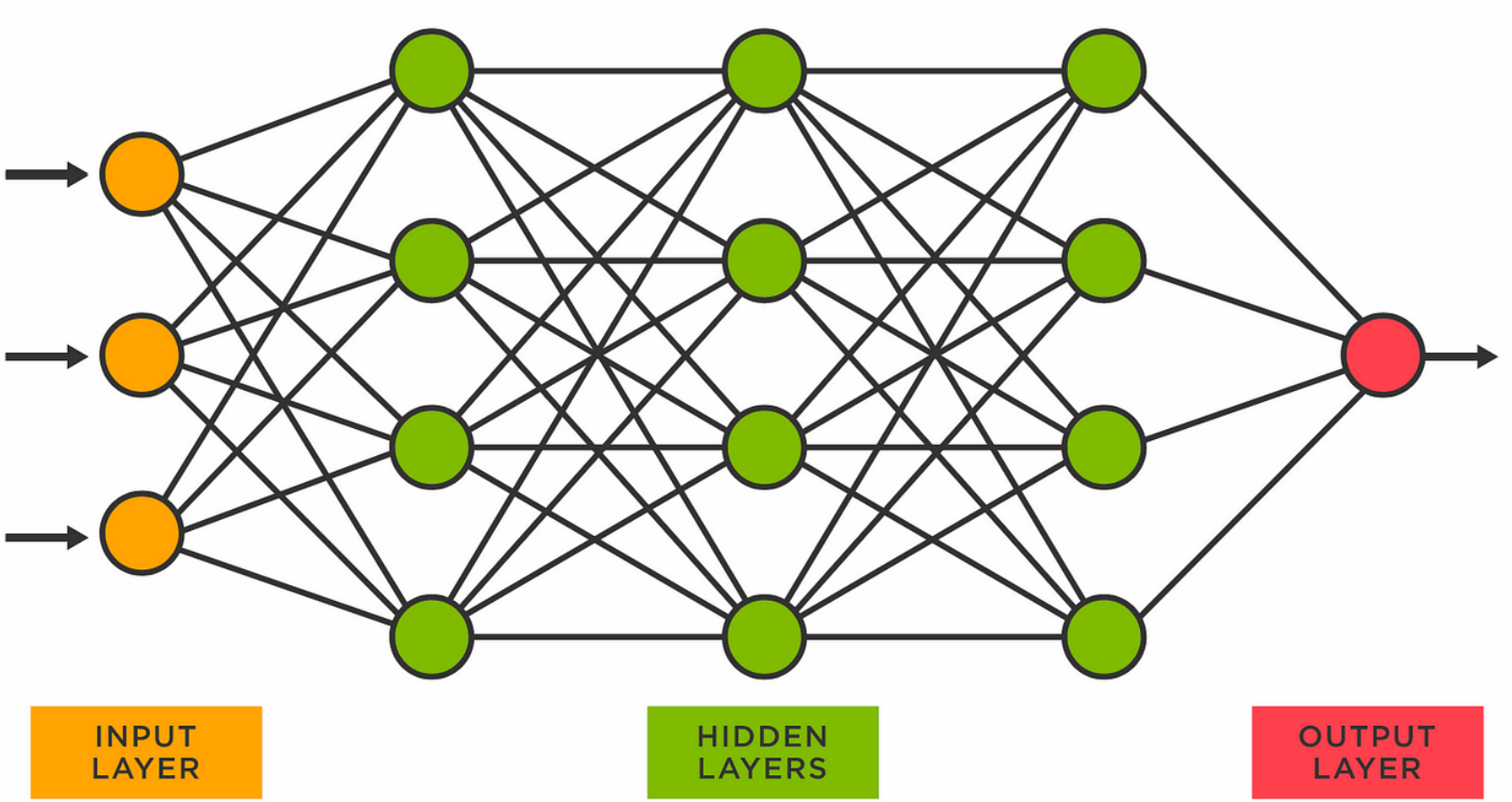
The "data modeling" culture assumes some parametric model for the "nature". Its parameters are estimated from the data and the model then used for inference and/or prediction. Model validation: goodness-of-fit tests and residual examination. Culture prevalence: 98% of all statisticians.

The "algorithmic modeling" culture considers the "nature" unknown, and estimates it by an algorithm. Model validation: Measured by predictive accuracy. Culture prevalence: 2% of statisticians, many in other fields.

This almost exclusive commitment of the statistical community to data models has led to irrelevant theory, questionable conclusions, and has kept statisticians from working on a large range of interesting current problems. Algorithmic modeling, both in theory and practice, has developed rapidly in fields outside statistics.

Back

Artificial Intelligence (AI)



Artificial Intelligence (AI)

- **Definition**: in its broadest sense, it is **intelligence exhibited by machines**. Sub-capabilities:
  - perception
  - learning
  - knowledge representation
  - natural language processing
  - social intelligence
  - reasoning and problem-solving
  - planning and decision-making
- **Applications**:
  - generative and creative tools (ChatGPT)
  - interacting via human speech (Google Assistant, Siri)
  - advanced web search engines (Google Search)
  - autonomous vehicles; drones; robotics
  - strategy games (chess, Go)
  - recommendation systems (Netflix, Amazon)
  - health and medicine, agriculture (image recognition)

Timeline

- (1950) **Turing** investigates the theoretical possibility of *machine intelligence*
- (1956) **Dartmouth workshop**: AI officially starts as an academic discipline; the term *Artificial Intelligence* is introduced
- (1956-1974) **Early successes**: perceptron (single-layer **neural network**) by Frank Rosenblatt (1958); optimism that a fully intelligent machine will be built in less than 20 years
- (1974-1980) **First AI Winter**: some approaches did not work; limited computing power; too small datasets
- (1980-1987) **Boom**: revival of neural networks; Hopfield net (John Hopfield); **backpropagation** (Geoffrey Hinton); probabilistic reasoning (Bayesian networks, hidden Markov models, stochastic models); reinforcement learning
- (1987-2005) **Second AI Winter**: the term Artificial Intelligence is intentionally avoided; narrow AI; Big Blue beats Garry Kasparov

Timeline (continued)

- (2005-2017) **AI takes off**:
  - **big data** (Labeled Faces in the Wild; ImageNet; word2vec; internet data); abundant and fast **computational resources** (GPU's, cloud computing)
    ▸ Hinton
  - in 2012, a **deep learning** model wins the ImageNet Large Scale Visual Recognition Challenge by a large margin
  - as of 2013, Geoffrey Hinton works part time for Google
  - Artificial General Intelligence (AGI: AI that matches or surpasses human cognitive capabilities across a wide range of cognitive tasks); DeepMind, OpenAI, Anthropic
- (2017-present) **AI-boom**:
  - large language models, **transformer architecture** (attention mechanism)
  - in 2023, Geoffrey Hinton resigns from Google
  - in 2024, John Hopfield and Geoffrey Hinton receive the Nobel prize

▸ Hinton

Geoffrey Hinton

Back in the 90s, our datasets were thousands of times too small, and our computers millions of times too slow

Back

Geoffrey Hinton (the "Godfather of AI")

- Born in 1947
- BA in experimental psychology in 1970
- PhD in artificial intelligence in 1978
- Paper on backpropagation to train multi-layer neural networks in 1986 (Nature)
- University of Toronto since 1987
- Behind the success of deep learning in 2012 (paper on Nature in 2015)
- Part time at Google Brain since 2013
- Resignation from Google in 2023 to be able to *"freely speak out about the risks of AI". "It is hard to see how you can prevent bad actors from using it for bad things"*
- Nobel prize in 2024

Premises

- **Which Statistics**?

  Narrow Statistics

  or

  Statistics?

- **Current situation**:
  - large datasets
  - sometimes more variables than observations (e.g. genomics, image recognition, text analysis)
  - prediction as main goal

  $\Rightarrow$ non-parametric approaches (Machine Learning, ML)

  $\Rightarrow$ importance of computational resources

Challenges

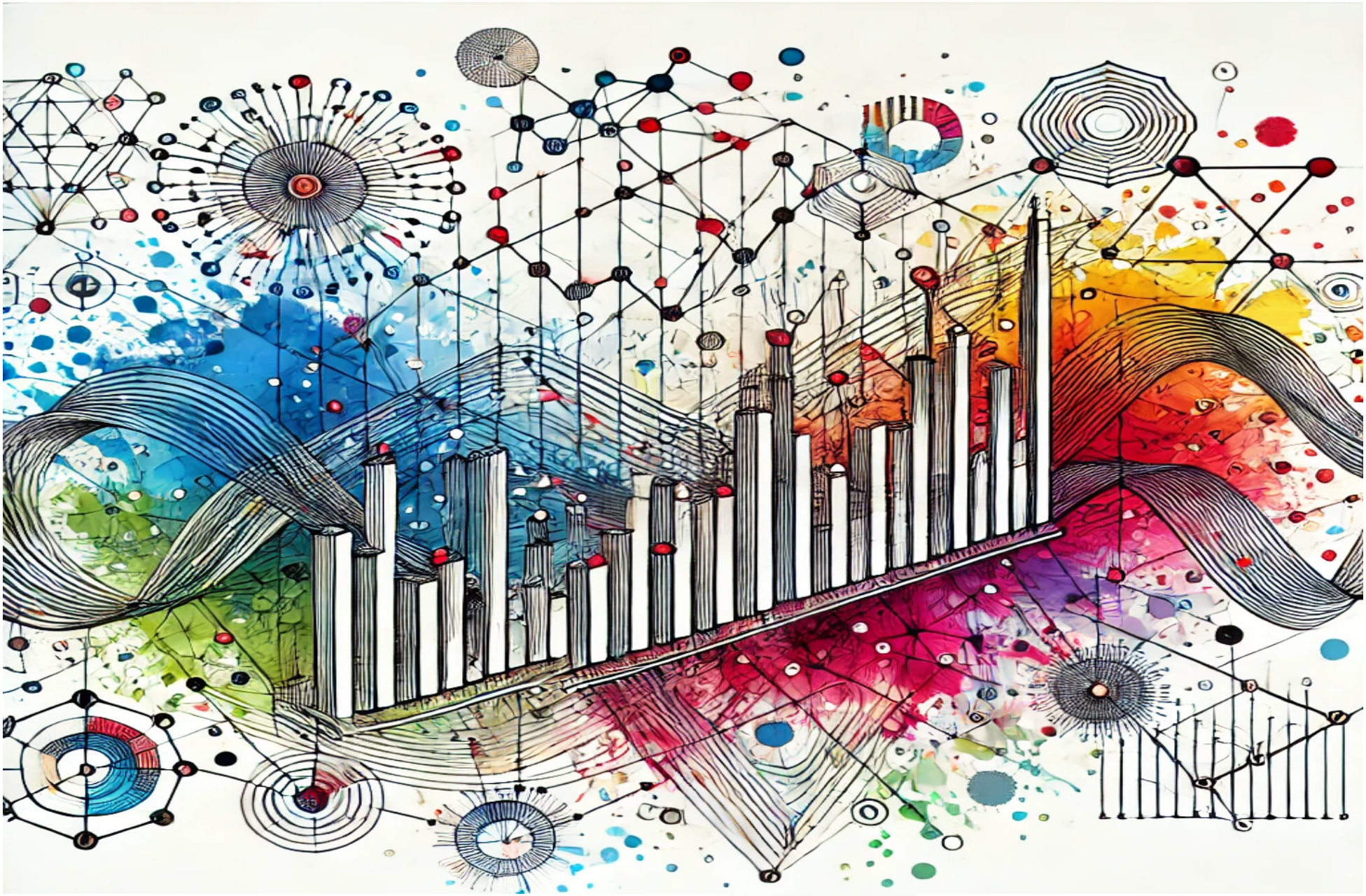
1. **Statisticians on the job market**: competition with computer scientists and engineers
2. **Statistician skills**: in light of Chat-GPT and friends, more focus on the data science process, how to guide and correct the AI answers, how to interpret the AI generated results; less focus on programming
3. **Role of domain knowledge**: in Statistics, domain knowledge guides model design and selection, as well as interpretation of results; AI/ML methods favor purely data-driven approaches, where domain knowledge is not considered in the modeling process (cf the PQRS workflow by Yu and Kumbier, 2018)
4. **Missing values, contamination by noise, outliers**: statistics has a long tradition in coping with these issues: it may contribute to go beyond crude treatments

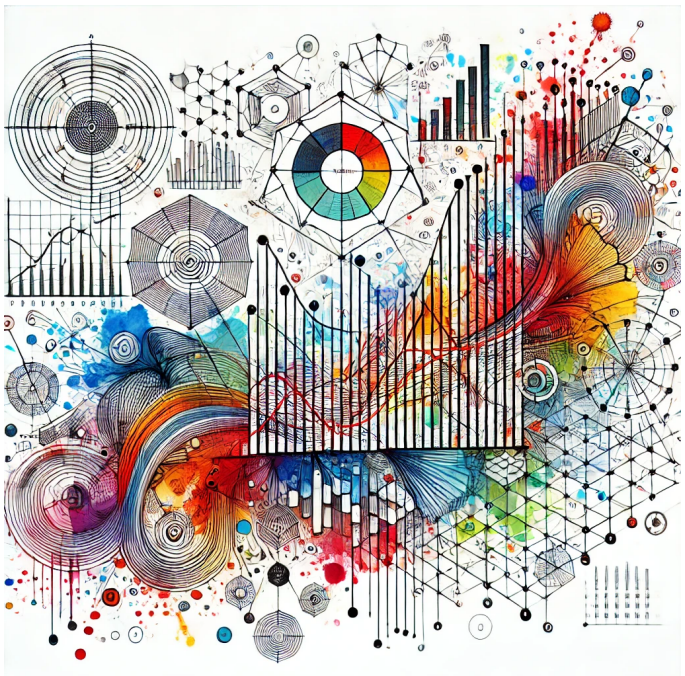
Challenges (continued)

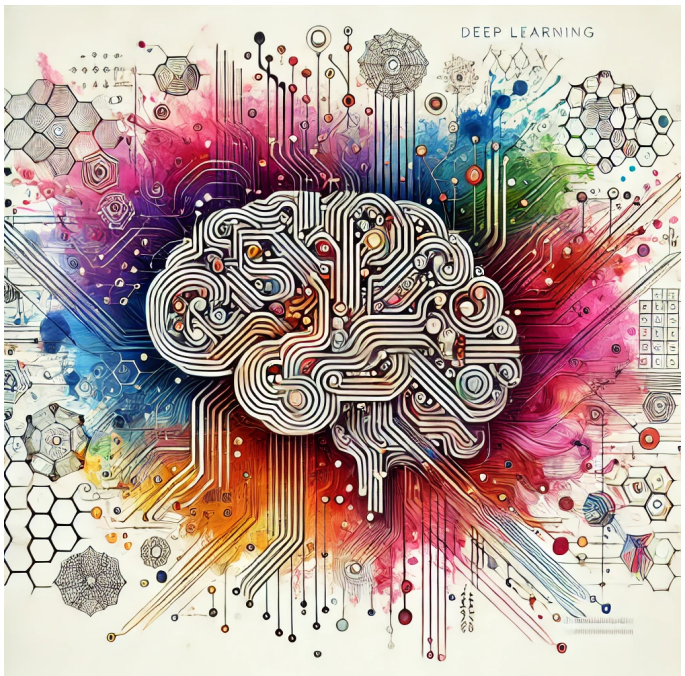
5. **Uncertainty quantification**: AI/ML methods struggle to quantify uncertainty about predictions → Statistics has well-established methods for quantifying uncertainty
6. **Prediction vs causality**: ML techniques tend to focus on prediction, but correlation does not imply causation; incorporating causal inference in ML (black box) approaches requires the winsdom and the imagination of the best statistical minds
7. **Explainability** (XAI, XML): provide transparency while maintaining the predictive power of AI/ML methods ⟶ new statistical methods to make the decision based on AI/ML understandable
8. **Ethical considerations; fairness**: particularly relevant in criminal justice, hiring, and lending; AI/ML methods amplify biases present in the training data, leading to unfair or discriminatory outcomes → new statistical approaches to reduce the impact of bias in the data and increase outcome fairness

Conclusions

- Challenges for **Statisticians**: competition with people coming from other domains (engineers, computer scientists), more numerous and more connected with the business
- Challenges for **Statistics**: more opportunities than challenges ⟵ cultural familiarity with the concepts of randomness, connection with the domain context (and the idea of a *data generating mechanism*), causality
- **Collaboration, collaboration, collaboration** (ASA Statement on "The Role of Statistics in Data Science and Artificial Intelligence", August 4, 2023)






DEEP LEARNING