Data Science and Astrology: Is There a Difference?

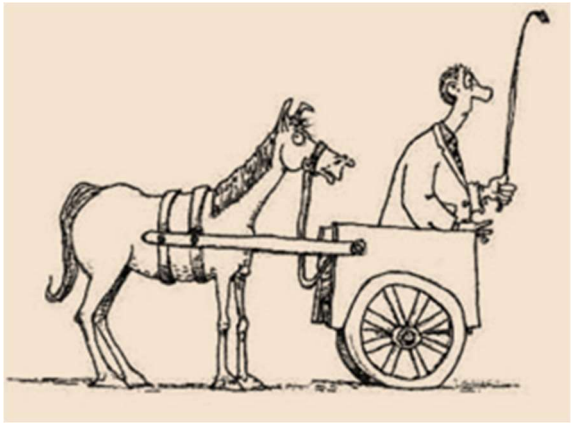
Jayant Haritsa
Database Systems Lab
Indian Institute of Science

Disclaimer: The opinions expressed in this talk are entirely personal and based on publicly available material. They do not represent the views of ACM India or Indian Institute of Science.

Data Science: Use with (Great) Care

The Evolution of Science

- Empirical (observe data)
  - e.g. chemical reactions
- Theoretical (explain data)
  - e.g. gravitational law
- Simulation (create data)
  - e.g. fluid dynamics
- Information (explore data)
  - e.g. black hole identification from space images



Output

Input

Data Science Definition

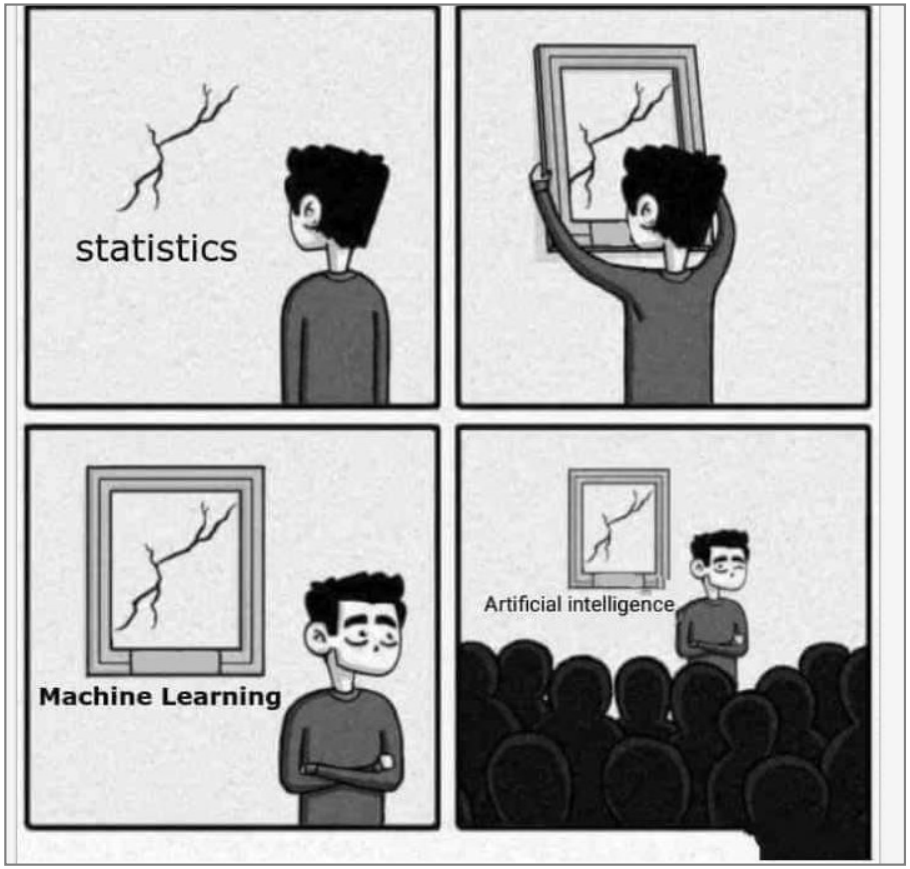
Data science is a multi-disciplinary field that uses scientific methods, processes, algorithms and systems to extract knowledge and insights from structured and unstructured data. (Wikipedia)

Statistics: Explanatory, manual, one-shot, limited data

Data Science: Predictive, automated, iterative, Big Data

(i.e. statistics on steroids)

Statistics → Data Science



When you're implementing,
it's logistic regression.
When you're hiring, it's ML.
When you're fundraising, it's AI.

www.facebook.com/statguy88/posts/488474608382395

Computational Positivism

- Studied by (late) Prof. Roddam Narasimha
  - "Axiomatism and Computational Positivism: Two Mathematical Cultures in Pursuit of Exact Sciences" [EPW, 38(35), 2003]
- Make computation match with observation
  - दृग्गणित ऐक्य
  - ancient Indian approach for astronomy
- No models, deductions, theories, philosophies
  - diametrically different to the Hellenic approach
  - Aryabhata vs Ptolemy

Data Science: The Good

Energy: Failure prediction

- Royal Dutch Shell built analytics platform to run predictive models to anticipate when oil drilling machine parts, numbering in the thousands, might fail.
- Reduced inventory analysis from days to hours, shaving millions of dollars on cost of inventory management.

www.cio.com/article/3221621/6-data-analytics-success-stories-an-inside-look.html

Transportation: Rate Determination

- RRD, the communications giant needs to find optimum shipping rates.
- Variables such as weather, geography, and drivers cost its business.
- Able to predict freight rates in real-time a week days in advance with 99% accuracy.

www.cio.com/article/3221621/6-data-analytics-success-stories-an-inside-look.html

Health: Medical Imaging

- Imaging techniques like X-Ray, MRI and CT Scan visualize inner parts of the human body.
- Traditionally, doctors manually inspect images to find irregularities. However, difficult to find microscopic deformities making it hard for proper diagnosis.
- With deep learning technologies in data science, such as image segmentation, possible to find microscopic deformities in the scanned images.

data-flair.training/blogs/data-science-in-healthcare/

Data Science: The Hype

DS: Skillset Required [Magazine Survey]

- What are the most valuable skills for a data scientist?
- Data Science is now being integrated with industries across all sectors, so data scientists are expected to have a broad set of skills. According to the study, the following skills were crucial:
  - Thorough knowledge of Python, as 44% of professionals use it heavily
  - Knowledge of Tableau, as 51% of data scientists use it
  - RStudio as an IDE
  - And in-depth knowledge of Hadoop

www.analyticsindiamag.com/top-8-faqs-about-data-scientists-in-india-answered/

DS: Education in a Jiffy [Coaching Academy]

- Data Science Bootcamps
  - 3 months on weekends
- Bootcamp Outcomes
  - Master all three elements of Data Science: Statistics, Tools, and Business Knowledge
  - Professional assistance and guidance on how to craft your CV and identify the right job opportunities

DS: The Myth

- Data is the new Oil !
- Data Science is the new Quantum Mechanics !
- ABCD: Any body can do Data Science!

DS: The Reality

- The key word in "Data Science" is not Data, it is Science! †
- Data Science requires deep understanding of physical/mathematical principles and data domain, not just programming environments.
- If you torture the data long enough, it will confess to anything! [Ronald Coase, Nobel 91]

† simplystatistics.org/2013/12/12/the-key-word-in-data-science-is-not-data-it-is-science/

Data Science: The Bad

NYT Op-ed Article [April 2014]

- Eight (No, Nine!) Problems With Big Data
  - Gary Marcus, Ernest Davis (NYU faculty)
    "big data is prone to giving scientific-sounding solutions to hopelessly imprecise questions"

Who's Bigger? Where Historical Figures Really Rank

(Book by MIT/Google: Hitler ranks higher than Aristotle!)

Need to ensure Big Data does not become Huge Nonsense …

www.nytimes.com/2014/04/07/opinion/eight-no-nine-problems-with-big-data.html

CallingBullshit.org [2017]

- Univ. of Washington, Seattle, USA
- Profs. Carl Bergstrom and Jevin West
- 3 credit course: Calling Bullshit in the Age of Big Data

"We will focus on bullshit that comes clad in the trappings of scholarly discourse. Traditionally, such highbrow nonsense has come couched in big words and fancy rhetoric, but more and more we see it presented instead in the guise of big data and fancy algorithms — and these quantitative, statistical, and computational forms of bullshit are those that we will be addressing in the present course."

www.callingbullshit.org

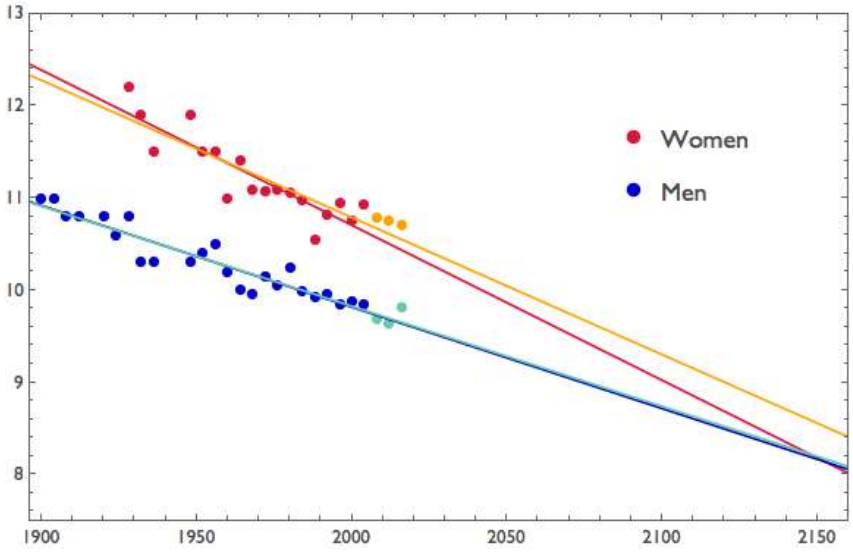
99.9% caffeine-free!

- Strong coffee (e.g. Starbucks) is also 99.9% caffeine-free!
- Because caffeine is a very potent drug even in small quantities!

www.callingbullshit.org/case_studies.html

Women faster than men in sprint!

- Linear regression to fit Olympic gold medal times for men and women in the 100 metre dash
  [Nature 2004, Tatem et al]

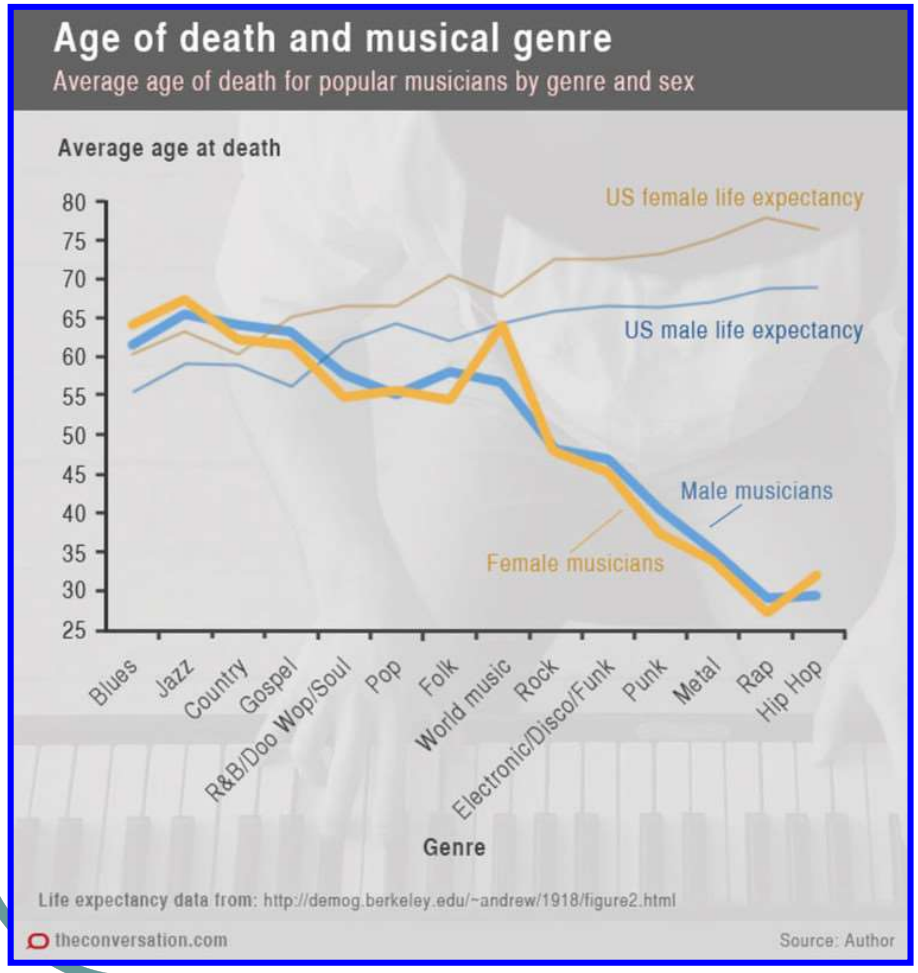


- In 2156, women will beat men for the first time
- In 2636, times of < 0 seconds will be recorded ☺
- Ignore reality, simply work backwards from data

www.callingbullshit.org/case_studies.html

Music to Die For!

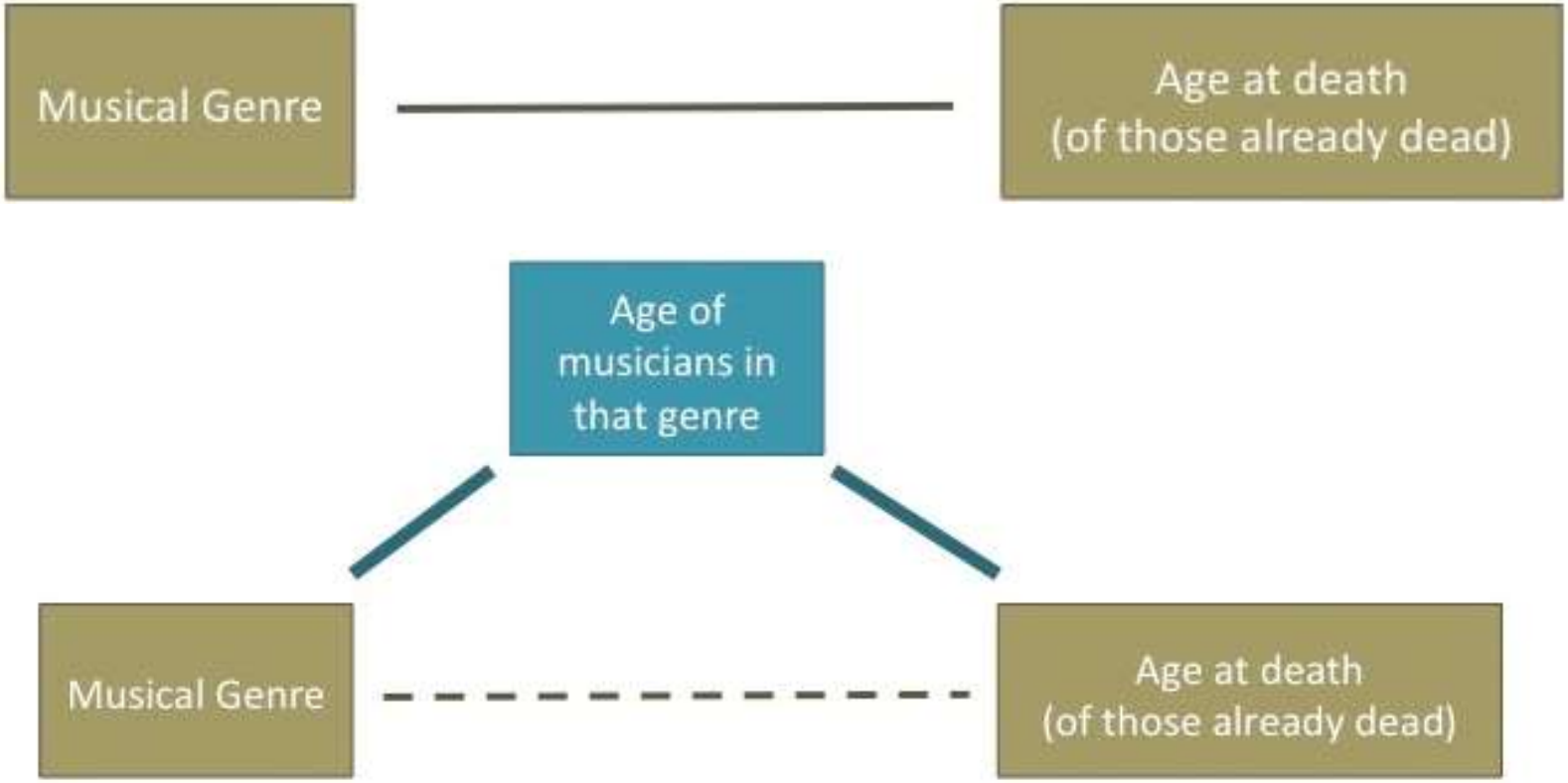
[The Conversation, Prof. Dianna Kenny, U Sydney]



- Worse than war – we don't lose half our army in a battle!
- Rap and hip-hop music are new genres! So, the data does not showcase the long-term age at death, but only those of the premature deaths in the genres.
- Average age at death converted to life expectancy in media.

www.callingbullshit.org/case_studies.html

Correlation versus Causality



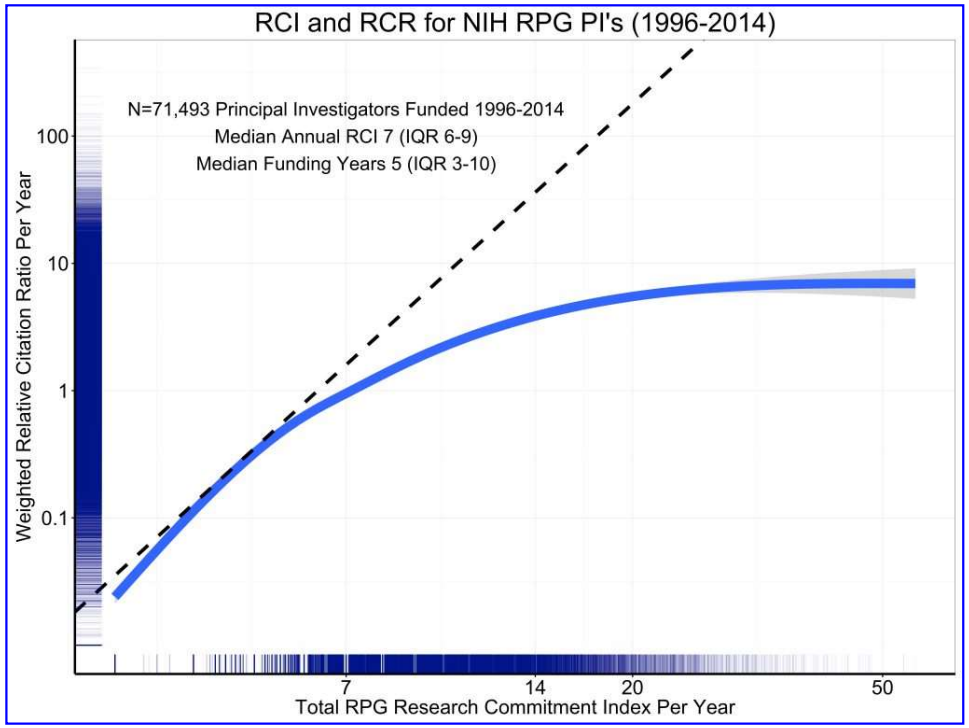
www.callingbullshit.org/case_studies.html

US Airline Rankings

- Best (American) and worst (Allegiant) airlines in 2019 were reversed in 2020
- Baggage safety parameter as #losses/#carried_bags instead of #losses/#carried_passengers
- Allegiant lost 1.75 bags/1000, Industry average is 5.57
- Allegiant passengers rarely check in bags because it charges huge extra fees for baggage that can exceed the price of the ticket! So there are no bags to lose!

www.forbes.com/sites/christopherelliott/2020/05/04/the-best-and-worst-airlines-of-2020/?sh=6cf8714e1a8e

NIH Grant Allocation [2017]



- Decreasing returns on more grants to single investigator, as time split across projects
- Proof: concavity of ROI graph
- Log-log concavity
  $\not\Rightarrow$ linear concavity !
- Composite results do not predict individual behaviour
  - If a and x are correlated, $f(x,a)\,\pi(a|x) \neq f(x,a)\,\pi(a)$

www.callingbullshit.org/case_studies.html

Chemistry Example

- **Predicting reaction performance in C-N cross-coupling using ML**
  Ahneman et al, Princeton University
  Science, 13 April 2018

- Studied impact of isoxazole additives on reaction poisoning during Buchwald-Hartwig coupling, a popular technique in pharmaceutical synthesis.
- Advanced ML technique – Decision Forest – worked much better than linear regression; successfully applied to sparse training sets and out-of-sample prediction.

Chemistry Example (contd)

- **Comment by Chuang and Keiser, UC San Francisco**
  Science, 16 November 2018

- Replacing the measured values of the chemical features (NMR shifts, dipole moments, etc.) with random values produced a model that had similar performance to the Princeton outcomes!
- Importance of high additive features cannot be distinguished from hidden structure within the dataset.

Covid: SUTRA Model [Hindu, 4/5/2021]

- Unlike many epidemiological models that extrapolated cases based on the existing number of cases, the behaviour of the virus and manner of spread, the SUTRA model chose a "data centric approach". The equation that gave out estimates of what the number of future infections might be and the likelihood of when a peak might occur, needed certain 'constants'. These numbers kept changing and their values relied on the number of infections being reported at various intervals. However, the equation couldn't tell when a constant changed. A rapid acceleration of cases couldn't be predicted in advance.
- SUTRA model was problematic as it relied on too many parameters, and recalibrated those parameters whenever its predictions "broke down". The more parameters you have, the more you are in danger of "overfitting".
- A combination of good epidemiologists, data-centric modelling like SUTRA and time-series models would have worked best.

www.thehindu.com/news/national/government-backed-model-to-predict-pandemic-rise-and-ebb-lacks-foresight-scientists/article34479503.ece

Covid (Pro)Test

- Sensitivity: 0.7 (true +ve)
- Specificity: 0.95 (true –ve)
- Serosurvey India Infected Population: 0.2

P (Infected/PositiveTest)

= 0.7 * 0.2 / (0.7*0.2 + 0.05 * 0.8)

= 0.75 (one in four incorrectly identified!)

Data Science: The Ugly

Weapons of Math Destruction [2016]

- Book by Cathy O'Neil
- How data science increases inequality and threatens democracy
- Our lives increasingly depend on our ability to make our case to machines

weaponsofmathdestructionbook.com/

Flawed Models

- Centrelink is Australian organization for administering welfare.
- Automated compliance system compares income self-reported by clients to information held by the taxation office.
- Strong and incorrect assumptions regarding income distribution across the year – unfairly penalized legitimate benefit recipients.

weaponsofmathdestructionbook.com/

Self-fulfilling prophecies

- A loan denial by a faulty risk model is more likely to be denied again when applying elsewhere, because it is on record that they have been refused credit before.
- Predictive policing based on demographics can alienate innocent targets to where they actually start behaving the way they are suspected to be.
- Software-driven just-in-time scheduling practices by companies resulted in people being treated like machine parts.

weaponsofmathdestructionbook.com/

Directed Behavior

- Modulate news feed algorithms to selectively push an opinion slant or pander to a particular section of society
- Confuse the issues with fake news
- Exert peer pressure (e.g. Facebook likes)

weaponsofmathdestructionbook.com/

CONCLUSION

Data Science Usage

- Tool of Last Resort to Validate a Hypothesis, not First
- Tool is a Support, not Substitute, for Domain Expertise
- Tool outputs should be compliant with Science, not biases

TakeAway

Data Science, like nuclear power, has enormous potential for benefiting mankind, and equally destructive power for ruining society …

Answer to Intro Question

Yes, there is a difference between Data Science and Astrology …

Astrology is more accurate ☺