Data Science:
The Good, the Bad and the Ugly

Jayant Haritsa
Database Systems Lab
Indian Institute of Science

The Evolution of Science

- Theory (explain data)
  - e.g. gravitational law
- Experiment (generate data)
  - e.g. LHC collider
- Computation (simulate data)
  - e.g. fluid dynamics
- Information (manage and analyse data)
  - e.g. black hole identification from space images

Data Science Definition

Really?

Data science is a multi-disciplinary field that uses scientific methods, processes, algorithms and systems to extract knowledge and insights from structured and un (Wikipedia)
[Dhanurjay Patil]

BLACK MAGIC!

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

Data Science: Predictive, automated, iterative, Big Data

(i.e. statistics on steroids)

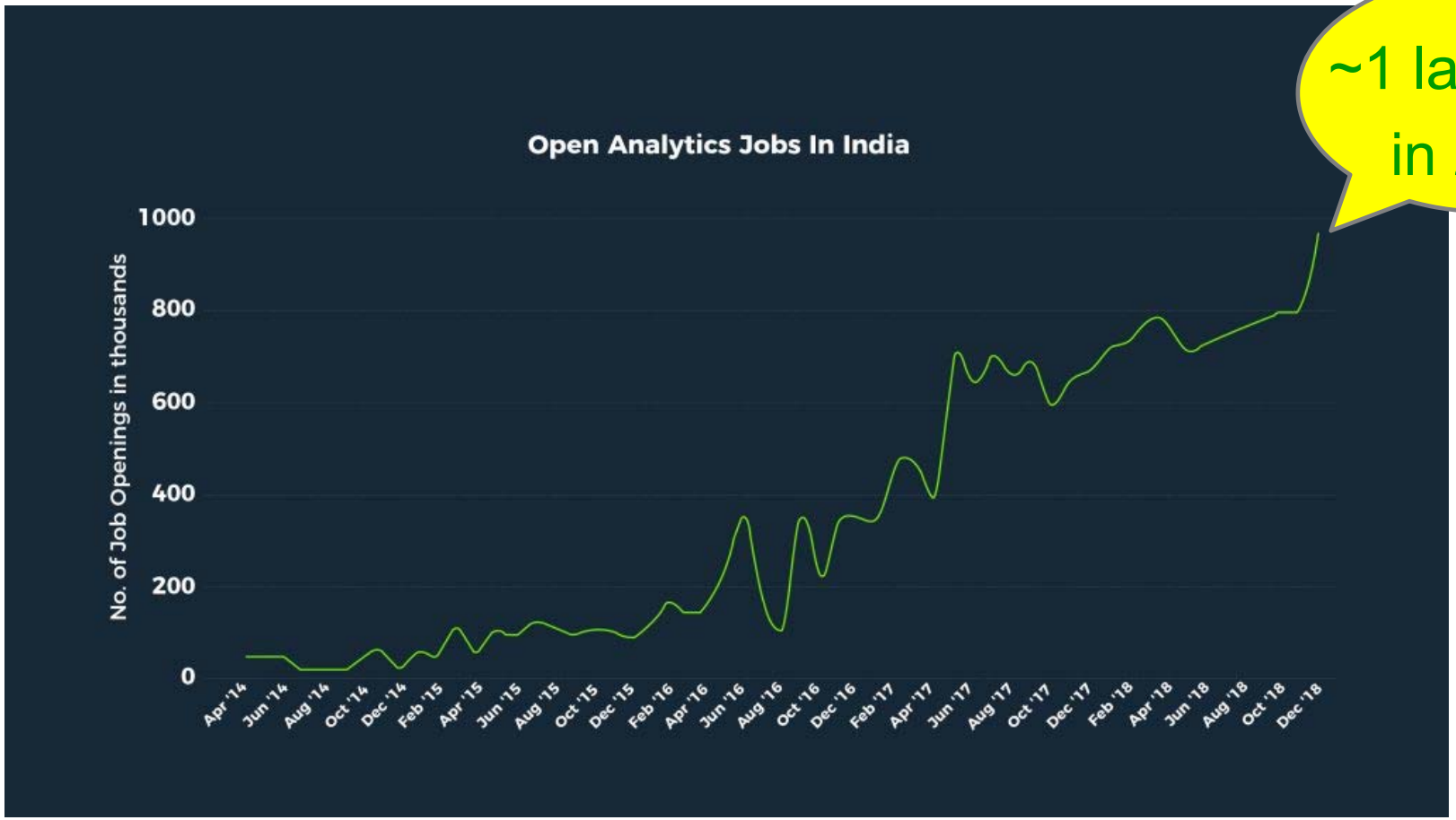
Data Science Applications

Finance, Manufacturing, Energy, Transport, Weather, Agriculture, Advertising, E-commerce, Communications, E-commerce, Web Search, Social Networks, Aadhaar …

Information Society
→ Knowledge Economy
→ Digital Citizens

Data Science: The Hype

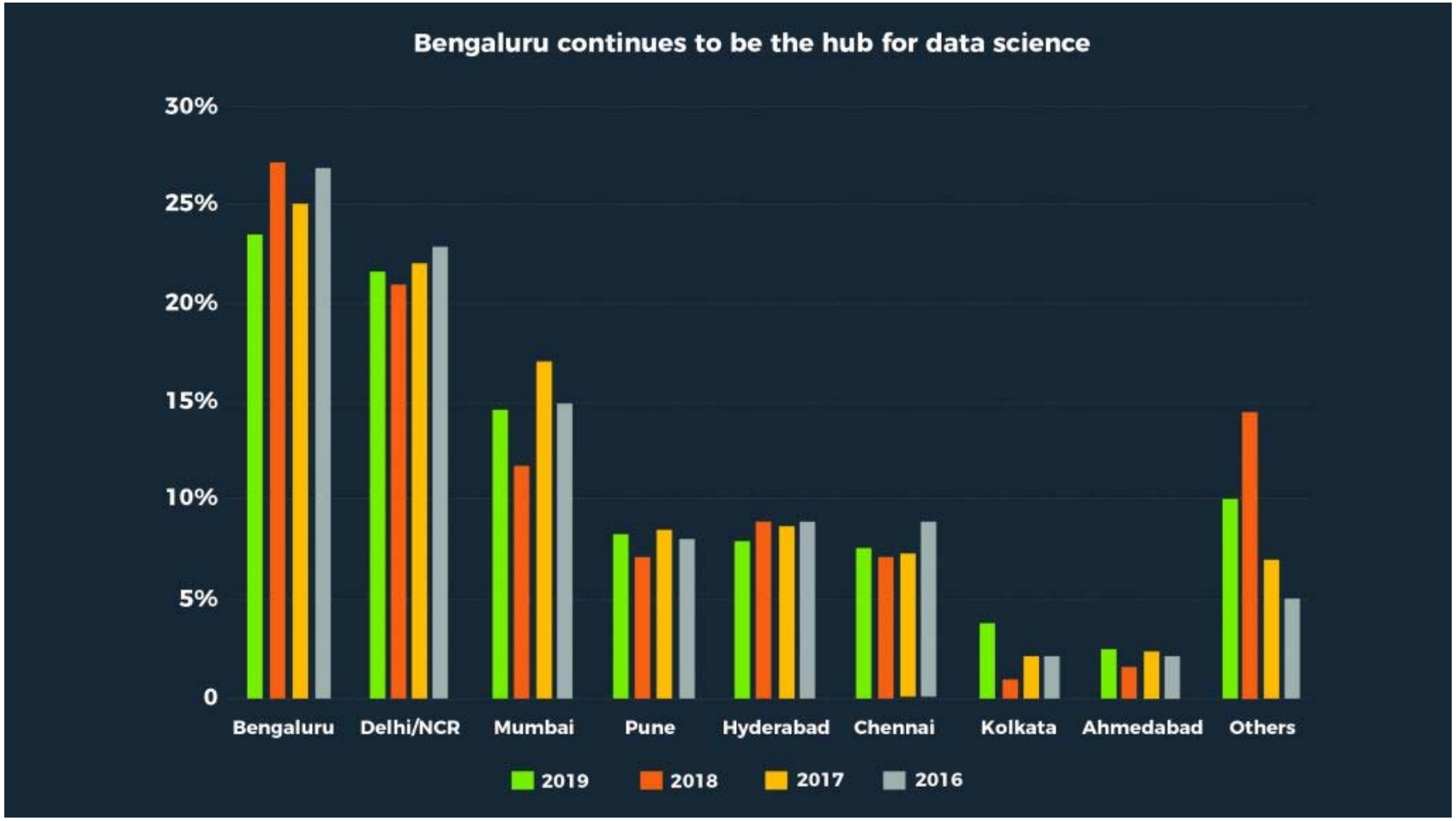
DS: Job Prospects



~1 lakh jobs in 2019

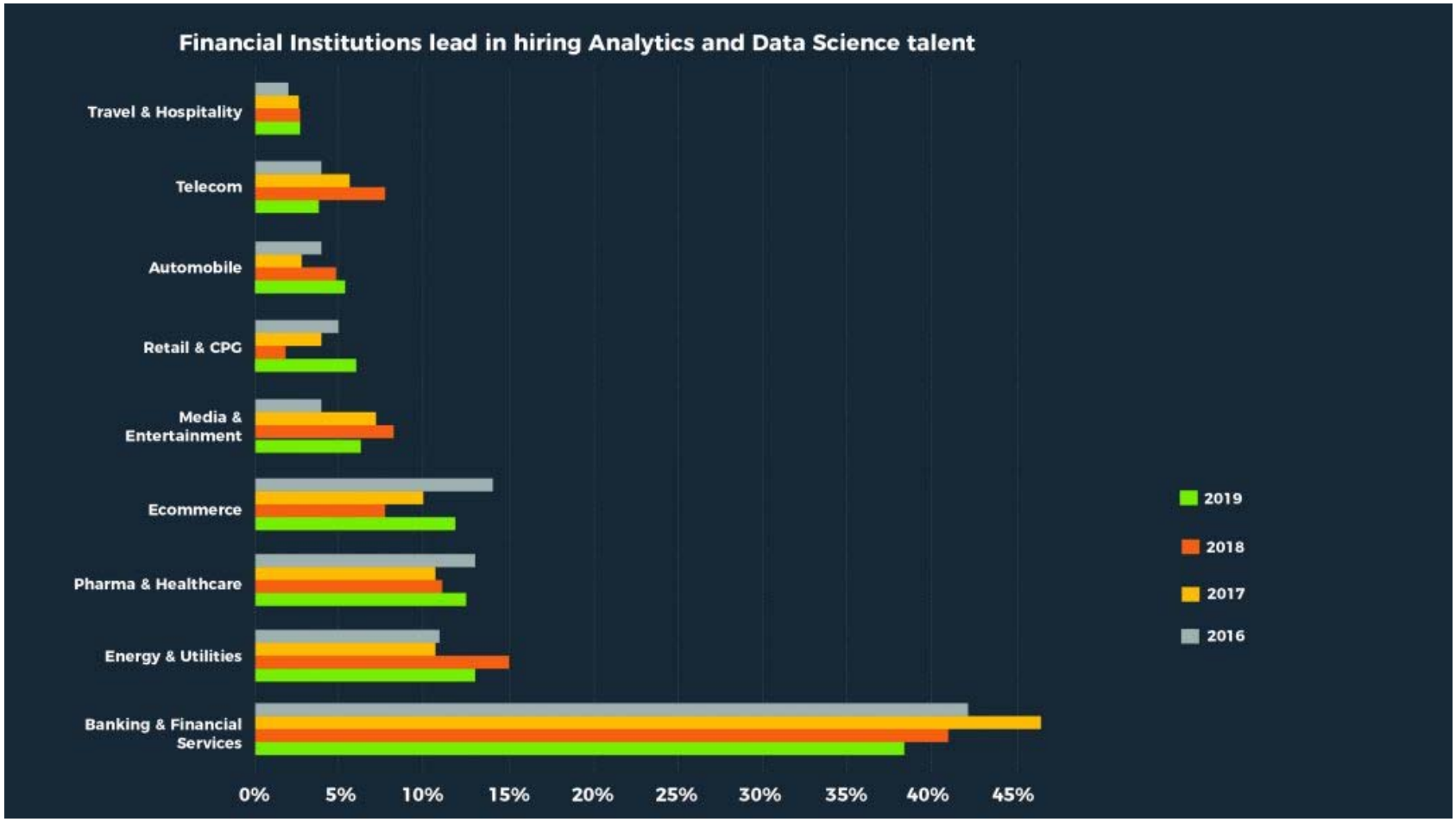
Source: www.analyticsindiamag.com/study-analytics-and-data-science-jobs-in-india-2019-by-great-learning-aim/

DS: Job Locations



Source: www.analyticsindiamag.com/study-analytics-and-data-science-jobs-in-india-2019-by-great-learning-aim/

DS: Job Domains



Source: www.analyticsindiamag.com/study-analytics-and-data-science-jobs-in-india-2019-by-great-learning-aim/

DS: No Background Required!

- Graduates from **Non-STEM background** can also become data scientists.
- **Beginners & Non-programmers** ideally would need to start learning programming - **Python** would be a good starting point. Simultaneously you can pursue learning Data Visualization, Analytics & Database tools like **Tableau, Excel, My SQL**.
- After which you can follow your **advanced learning** in these steps:
  - **First Step:** Try to advance your Programming capabilities (Python or R).
  - **Second Step:** Gain knowledge of Intermediate Statistics & Probability applications in business scenarios, how ML algorithms and methods are applied
    - ***CAUTION:*** Focus on few ML algorithms to begin with - *linear regression, logistic regression, K-Means*, and try to understand **(in depth)** how they are applied.
  - **Third Step:** Actual test of your advanced analytics and machine learning skills. Work with independent projects with data sets from platforms like Kaggle.

Source: www.quora.com/Can-I-become-a-data-scientist-with-no-STEM-background

DS: Education in a Jiffy!

- Data Science Bootcamps
  - 3 months on weekends (~Rs 1 lakh)
- 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

Source: www.jigsawacademy.com/analytics-classroom-training/

DS: Skillset

- 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 our study, the following skills were crucial:
  - Thorough knowledge of Python, as 44% of professionals use it the most
  - Knowledge of Tableau, as 51% of the data scientists use it
  - RStudio as an IDE
  - And in-depth knowledge of Hadoop

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

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

- Data is the new Snake Oil !
- Richard Feynman: "If you think you have understood quantum mechanics, you don't understand quantum mechanics" !
- Data Science requires deep understanding of physical/mathematical principles and the data domain, not just programming tools and environments.

Data Science: The Good

Computational Positivism

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

Energy: Failure prediction

- Shell built analytics platform to run predictive models to anticipate when more than 3,000 different oil drilling machine parts might fail.

- Reduced inventory analysis from 2 days to less than an hour, shaving millions of dollars a year off the cost of inventory management

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

Transportation: Customer Behavior

- Airlines Reporting Corp. (ARC) settles close to $100 billion worth of airline transactions.
- In the process, it knows where travellers are going, when they travel and how much they are paying for the 2+ billion flights/year.
- ARC captures the data, ingests it into analytics engines, refines it, and builds custom reports for its airline clients.

Source: 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 7 days in advance with 99 % accuracy.

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

India: Consumer Electronics

- Croma observed that customers found it difficult to make speedy purchase decisions due to the plethora of categories, products, and SKUs.

- Deep understanding of site visitors' preferences by studying their social profiles, brand affinities, recent activities, and household and macro-economic data, with the help of analytics.

Source: www.bloombergquint.com/labs/intel/transform-to-scale/analytics_for_business_success.html

India: Banking Sector

- Yes Bank uses analytics to design targeted campaigns and cashback offers for its customers. Automating the monitoring-to-settlement process, their solution credits cashback amounts to customer savings accounts or wallets.

Source: www.bloombergquint.com/labs/intel/transform-to-scale/analytics_for_business_success.html

India: Power Sector

- Reliance Power employs analytics for condition monitoring to detect and avoid critical equipment failures, which not only cause maintenance hassles but also huge losses and severe reputation damage. They deployed a condition monitoring and diagnostic system, a predictive analytics solution, helping its managers take informed maintenance decisions quickly.

Source: www.bloombergquint.com/labs/intel/transform-to-scale/analytics_for_business_success.html

Data Science: The Bad

Big Data: Old Concept, New Hype

- VLDB: Premier international database conference, started in 1975
  - Very Large Data Bases
- Large ≈ Big, ∴ Very Large ≫ Big
- Only a few enterprises really have big data, the others just say so for bragging rights!
  - World Data Center for Climate ~ 1 petabyte

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!)

We need to ensure that Big Data does not wind up becoming Huge Nonsense …

CallingBullshit.org [2017]

- Univ. of Washington, Seattle, USA
- Profs. Carl Bergstrom and Jevin West
- 1 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."

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!

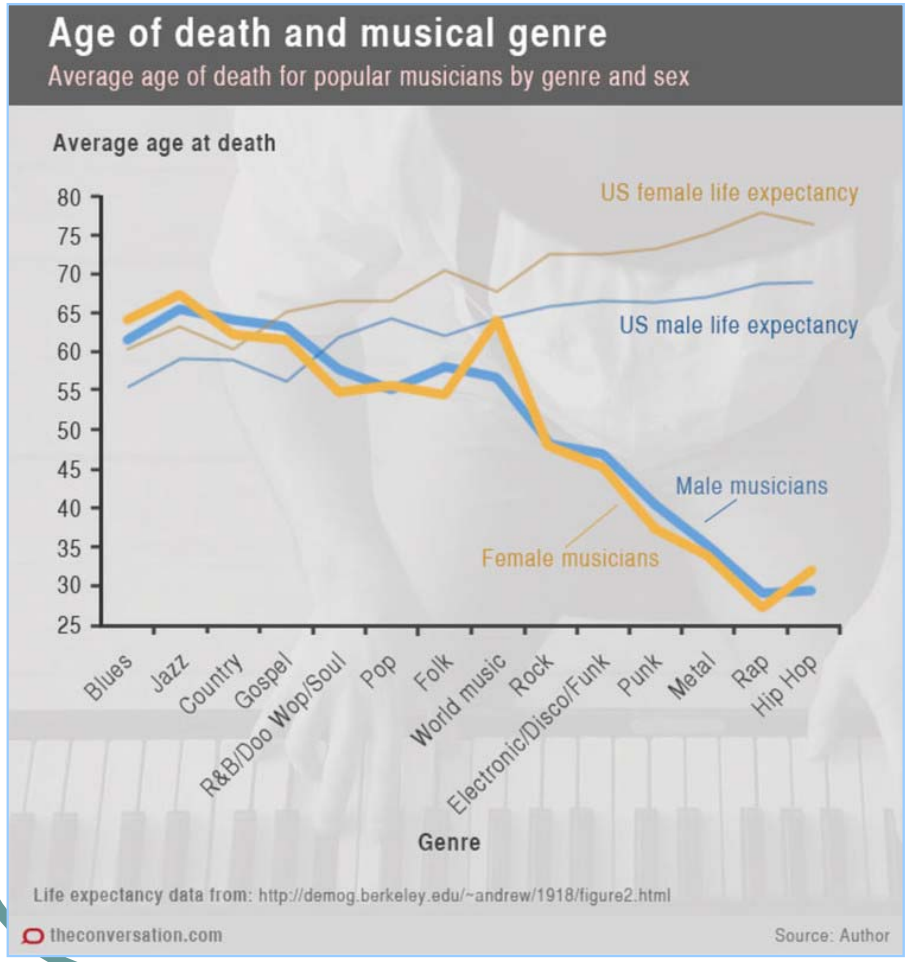
Source: callingbullshit.org/case_studies/

Women faster than men in sprint!

- Linear regression to fit Olympic gold medal times for men and women in 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.

Source: callingbullshit.org/case_studies/

Music to Die For! [The Conversation]



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

Source: callingbullshit.org/case_studies/

Music to Die For! [The Conversation]

Cause of death by genre

Various causes of death for musicians of different genres

	Accidental	Suicide	Homicide	Heart-related	Cancer
% deaths per cause	19.5%	6.8%	6.0%	17.4%	23.4%
Blues	9.2%	2.0%	3.5%	28.0%	24.2%
Jazz	10.6%	2.7%	1.9%	20.7%	30.6%
Country	15.8%	4.7%	1.6%	23.5%	25.1%
Gospel	13.3%	0.9%	3.6%	18.5%	23.0%
R&B	11.5%	1.6%	5.0%	23.2%	26.8%
Pop	19.0%	6.4%	2.9%	16.4%	26.7%
Folk	15.9%	5.5%	4.4%	15.3%	32.3%
World music	12.7%	3.4%	9.6%	17.8%	19.9%
Rock	24.4%	7.2%	3.6%	15.4%	24.7%
Electronic	16.7%	5.0%	10.0%	15.0%	25.0%
Punk	30.0%	11.0%	8.2%	12.6%	18.3%
Metal	36.2%	19.3%	5.9%	11.0%	14.1%
Rap	15.9%	6.2%	51.0%	6.9%	7.6%
Hip Hop	18.3%	7.4%	51.5%	6.1%	6.1%

Note: not all causes shown

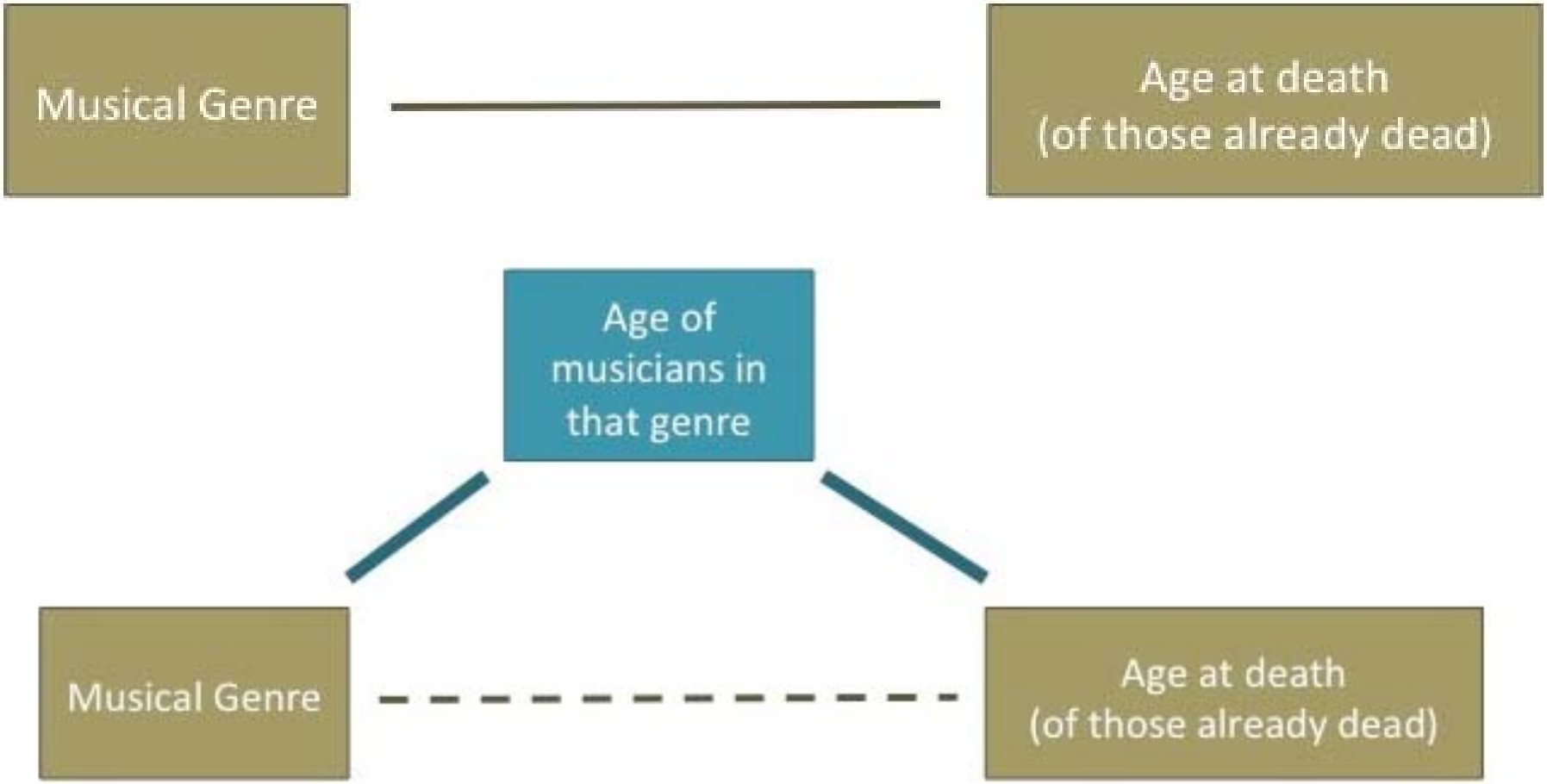
Red: significantly above the overall average rate for cause of death
Blue: above the overall average rate for cause of death
Green: significantly below the overall average rate for cause of death

theconversation.com Source: Author

- Not the lifetime probability of each cause of death, but the probability of death conditioned on occurrence at time of study.
- Genre-specific mortality is confounded with the age distribution of musicians in each genre.

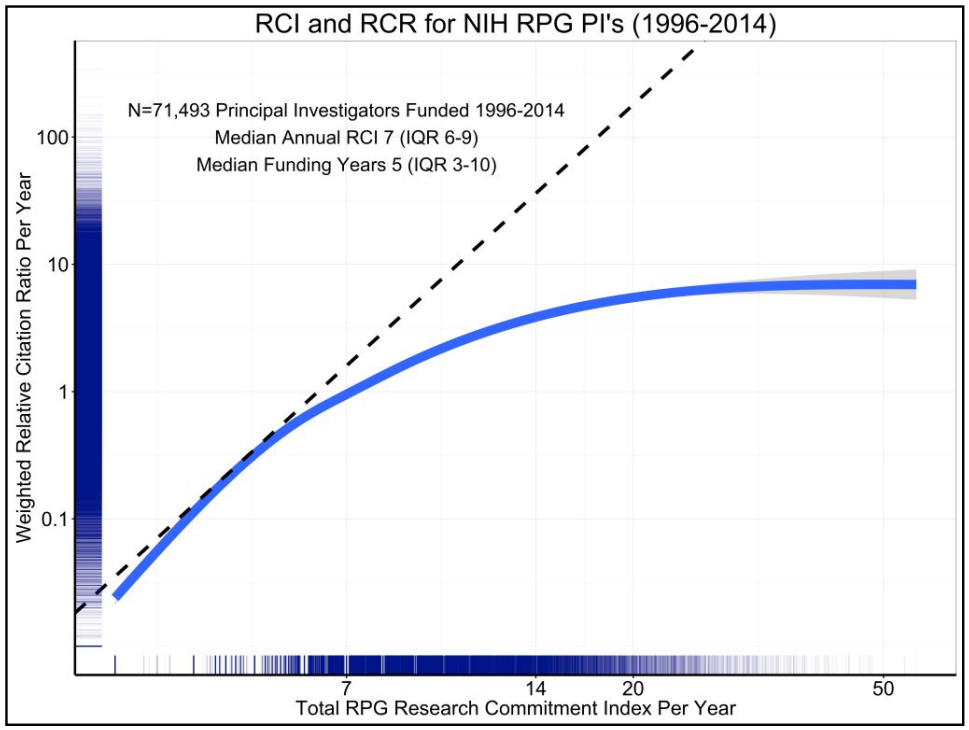
Source: callingbullshit.org/case_studies/

Correlation versus Causality



Source: callingbullshit.org/case_studies/

NIH Grant Allocation [2017]



- Decreasing returns on more grants to single investigator, since time is split across more projects
- Proof: concavity of ROI graph
- Log-log concavity X⇒ linear concavity !
- Composite results do not predict individual behavior

Design Disasters

1. UK Immigration [2013]

A Home Office text message campaign accusing people of being illegal immigrants has received numerous complaints after several people were contacted in error. Officials have sent messages to almost 40,000 people they suspect of not having a right to be in the UK, instructing them to contact border officials to discuss their immigration status. Government commissioned Capita, the outsourcing company, to trace people believed to have outstayed their visas.

UK Immigration (contd)

- In a few months, Capita was accused of mishandling cases and getting just as mixed up as the bureaucrats it was supposed to be replacing!
- In November, Capita admitted a backlog of 150,000 notifications to foreign students it hadn't been able to process and therefore determine if they should or shouldn't still be in the country.

  In IT terms, it's been at the center of a billion dollar botched "e-borders" system, which has been missing deadlines and delivery dates since the middle of the last decade and which may not even be legal under European Union legislation!

2. Obama HealthCare.gov [2013]

- Severe problems were caused by unexpected high volume when the site drew 250,000 simultaneous users instead of the 50,000-60,000 expected. More than 8 million people visited the site from October 1 to 4. White House officials subsequently conceded that it was not just an issue of volume, but involved software and systems design issues. Also, stress tests done by the contractors one day before the launch date revealed that the site became too slow with only 1,100 simultaneous users !
- HealthCare.gov problems persisted even weeks after the launch. For example, a networking error at the related data services hub killed the website's functionality. This occurred the exact day after Health & Human Services head Kathleen Sebelius had highlighted designing that data hub as a government success.

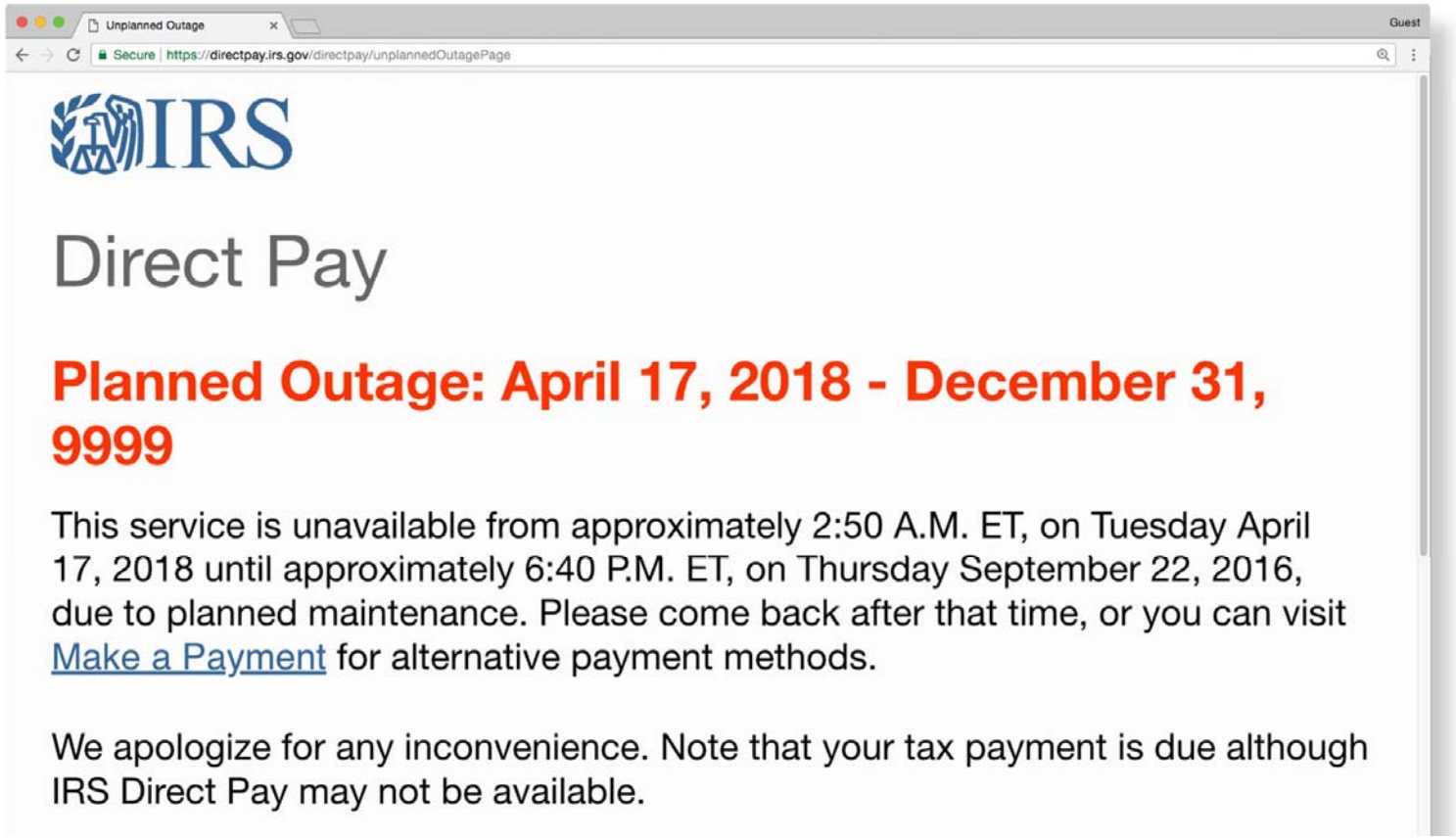
3. Flipkart → Flopkart [2014]

- Deccan Herald: Big Apology Day follows Flipkart's Big Billion Day
  - After its Big Billion Day on Monday, which fetched Flipkart.com $100 million by way of sales and the ire of hordes of angry customers who complained of technical glitches and false promises on discounts, the Bangalore-based online giant was quick to apologise for its drawbacks on Tuesday.
  - "Though we saw unprecedented interest in our products and traffic like never before, we also realised that we were not adequately prepared for the sheer scale of the event. We didn't source enough products and deals in advance to cater to your requirements. To add to this, the load on our server led to intermittent outages, further impacting your shopping experience on our site," the Bansals said.
  - Noting that it took enormous effort from everyone at Flipkart, many months of preparation and pushing its "capabilities and systems to the limit" for the big day, the Bansals said that they were looking at deals and offers painstakingly put together for months.

4. Electronic Health Records [2018]

- IEEE Spectrum: A US $4.3 billion dollars electronic health records program for the U.S. Department of Defense is "neither operationally effective nor operationally suitable," according to a recently released memo and report from the agency's director of operational test and evaluation.

5. Income Tax [2018]



Data Science: The Ugly

Weapons of Math Destruction [2016]

- Authored by Cathy O'Neil from Wall Street
- Longlisted for the National Book Award
- How data science increases inequality and threatens democracy
- Our lives increasingly depend on our ability to make our case to machines.

Source: 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.
- Made strong and incorrect assumptions regarding income distribution across the year - unfairly penalized legitimate benefit recipients.

Source: 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 [Starbucks].

Source: weaponsofmathdestructionbook.com/

Directed Behavior

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

Source: weaponsofmathdestructionbook.com/

The Orwellian State

China is using data science techniques to identify Hong Kong residents protesting a proposed extradition law by employing mass data collection and sophisticated facial recognition technology. So protestors are coming out wearing masks, buying transport tickets with cash.

Chinese authorities have also begun deploying a new surveillance tool that uses people's body shapes and how they walk to identify them, even when their faces are hidden from cameras.

Source: www.apnews.com/028636932a874675a3a5749b7a533969

CONCLUSION

Ethical Data Science

Data science codifies the past, but does not invent the future. Doing that requires moral imagination, and that's something that only humans can provide. We have to explicitly embed better values into our algorithms, creating models that follow good ethics. Sometimes that will mean putting fairness ahead of profit.

Source: weaponsofmathdestructionbook.com/

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 …