MODULE 2: DATA ANALYSIS, DATA SCIENCE, AND BUSINESS INTELLIGENCE

CT ACADEMY | DATA ACTION LAB

6. ASKING QUESTIONS & NON-TECHNICAL ASPECTS OF DATA WORK

DATA ANALYSIS, DATA SCIENCE, AND BUSINESS INTELLIGENCE

ASKING THE RIGHT QUESTIONS

Data science is about asking and answering questions:

- **Analytics:** “How many clicks did this link get?”
- **Data Science:** “Based on this user’s previous purchasing history, can I predict what links they will click on the next time they access the site?”

Data mining/science models are usually **predictive** (not **explanatory**): they show connections, but don't reveal why these exist.

Warning: not every situation calls for data science, artificial intelligence, machine learning, statistics, or analytics.

THE WRONG QUESTIONS

Too often, analysts are asking the **wrong questions:**

- questions that are **too broad** or **too narrow**
- questions that **no amount of data could ever answer**
- questions for which **data cannot reasonably be obtained**

The **best-case scenario** is that stakeholders will recognize the answers as irrelevant.

The **worst-case scenario** is that they will erroneously implement policies or make decisions based on answers that have not been identified as misleading or useless.

ROADMAP TO FRAMING QUESTIONS

Understand the problem (opportunity vs. problem)

What initial assumptions do I have about the situation?

How will the results be used?

What are the risks and/or benefits of answering this question?

What stakeholder questions might arise based on the answer(s)?

Do I have access to the data necessary to answering this question?

How will I measure my ‘success’ criteria?

YES/NO TRAP

Examples of **bad** questions:

- Are our revenues **increasing** over time? **Has it** increased year-over-year?
- Are most of our customers from **this demographic**?
- **Does this project have** valuable ambitions to the broader department?
- **How great** is our hard-working customer success team?
- How often do you **triple check** your work?

Examples of **good** questions:

- What's the **distribution** of our revenues over the past three months?
- Where are our **top 5** high-spending cohorts from?
- What are the **different benefits** of pursuing this project?
- What are **three good** *and* **bad traits** of our customer success team?
- Do you **tend to** do quality assurance testing on your deliverables?

QUESTION AUDIT CHECKLIST

1. Did I avoid creating any **yes/no** questions?
2. Would **anyone** in my team/department understand the question irrespective of their backgrounds?
3. Does the question need more than one sentence to express?
4. Is the question '**balanced**' – scope is not too broad that the question will never truly be answered, or too small that the resulting impact is minimal?
5. Is the question being **skewed** to what may be easier to answer for my/my team's particular skillset(s)?

EXERCISE

Are the following examples of good questions? Are they vague or specific? What are the ranges of answers we could expect? How would you improve them?

1. How does rain affect goal percentage at a soccer match?
2. Did the Toronto Maple Leafs beat the Edmonton Oilers?
3. Did you like watching the Tokyo Olympics?
4. What types of recovery drinks do hockey players drink?
5. How many medals will Canada win at the Paris 2024 Olympics?
6. Should we fund the Canadian Basketball team more than the Canadian Hockey team?

Try to build “bad” questions as well, to get a handle on the difference.

MULTIPLE I'S APPROACH TO DATA WORK

Technical and quantitative proficiency (or expertise) is **necessary** to do good quantitative work *in the real world*, but it is **not sufficient** – optimal real-world solutions may not always be the optimal academic or analytical solutions.

This might be the biggest surprise for those transitioning out of academia.

What works for one person, one job application, one project, one client, etc. may not work for another – **beware the tyranny of past success**!

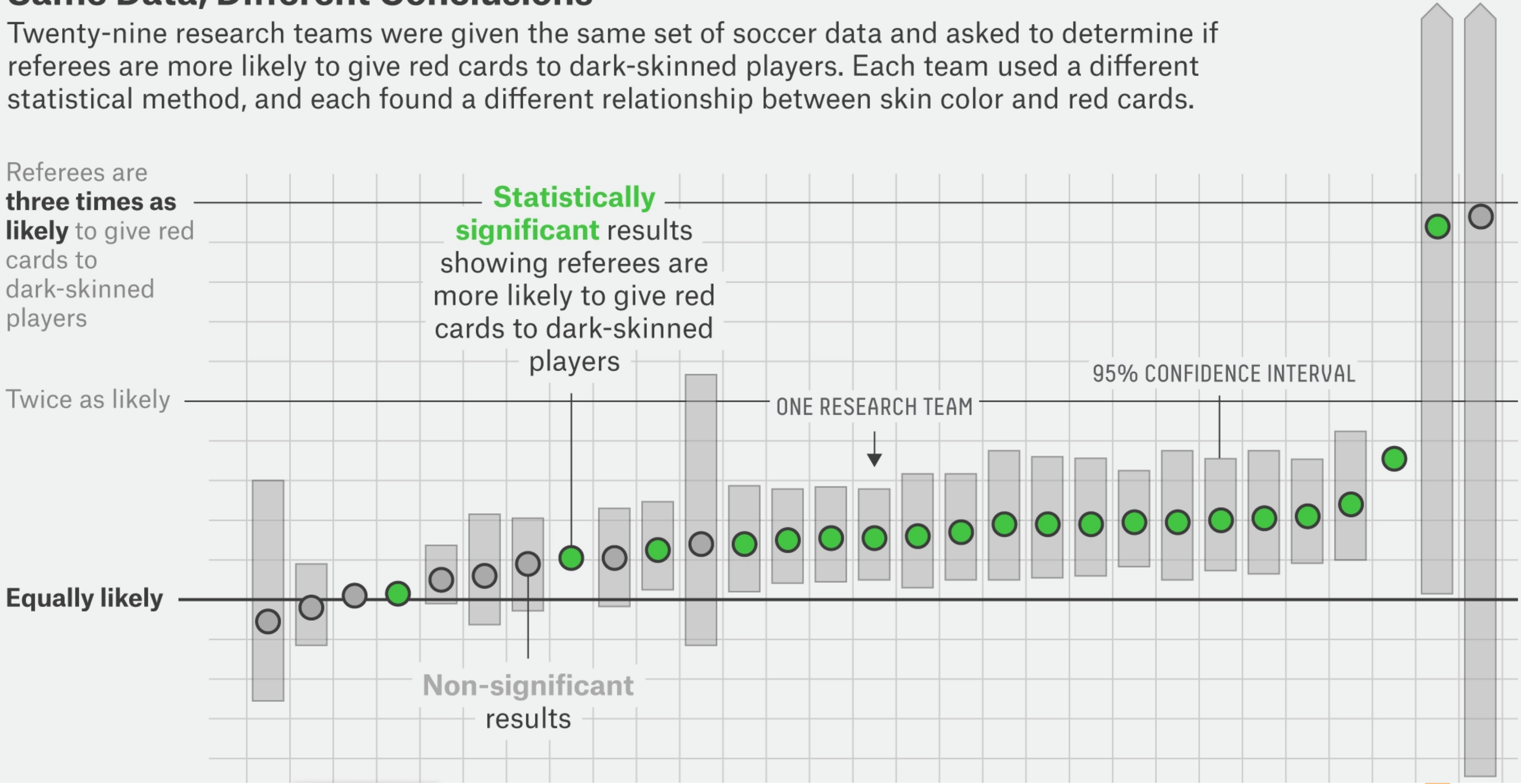
The focus of quantitative work must include the delivery of **useful analyses/ products** (Multiple "I"s).

MULTIPLE I’S APPROACH TO DATA WORK

- **intuition**
  understanding context
- **initiative**
  establishing an analysis plan
- **innovation**
  new ways to obtain results, if required
- **insurance**
  trying multiple approaches
- **interpretability**
  providing explainable results
- **inquisitiveness**
  not only asking the same questions repeatedly
- **integrity**
  staying true to objectives and results
- **independence**
  self-learning and self-teaching
- **interactions**
  strong analyses through teamwork
- **interest**
  finding and reporting on interesting results
- **intangibles**
  thinking “outside the box”
- **insights**
  providing actionable results

Same Data, Different Conclusions

Twenty-nine research teams were given the same set of soccer data and asked to determine if referees are more likely to give red cards to dark-skinned players. Each team used a different statistical method, and each found a different relationship between skin color and red cards.



From: Science isn't broken - It's just a hell of a lot harder than we give it credit for. [**Christie Aschwanden, 2015**]

ANALYSIS CHEAT SHEET

1. Business solutions are not always academic solutions.
2. Data and models don't always support the stakeholder's hopes, wants, needs.
3. Timely communication is key – externally and internally.
4. Data scientists need to be flexible (within reason), and willing and able to learn something new, quickly.
5. Not every problem calls for data science methods.
6. We should learn from both our good and our bad experiences.

ANALYSIS CHEAT SHEET

7. Manage projects and expectations.
8. Maintain a healthy work-life balance.
9. Respect the stakeholders, the project, the methods, and the team.
10. Data science is not about how smart we are; it is about how we can provide actionable insight.
11. When what the client wants can't be done, offer alternatives.
12. "There ain't no such thing as a free lunch."

EXERCISES

1. Have you encountered the Analysis Cheat Sheet lessons in your work? Have you encountered others?

2. Find examples of recent “Data in the News” stories. Were they successes or failures? What social consequences could emerge from the technologies described in the stories?

SUPPLEMENTAL MATERIAL

6. ASKING QUESTIONS & NON-TECHNICAL ASPECTS OF DATA WORK

MULTIPLE I’S APPROACH TO DATA WORK

Prospective employees/analysts are not solely gauged on technical know-how, but also on the ability to **contribute positively** to the workplace/project:

- communication
- team work and multi-disciplinary abilities
- social niceties and flexibility
- non-technical interests

Employers rarely chose robots when human beings are available; stakeholders are more likely to accept data recommendations from **well-rounded people**.

You should also evaluate eventual employers/clients on these axes.

ROLES AND RESPONSIBILITIES

A data analyst or a data scientist (in the **singular**) is unlikely to get meaningful results – there are simply too many moving parts to any data project.

Successful projects require **teams** of highly-skilled individuals who understand the **data**, the **context**, and the **challenges**.

Team *size* could vary from a few to several dozens; typically easier to manage small-ish teams (with 1-4 members, say, with **role overlaps**).

Domain Experts / SMEs

- are authorities in a particular area or topic
- guide team through unexpected complications and knowledge gaps

ROLES AND RESPONSIBILITIES

Project Managers / Team Leads

- understand the process enough to recognize whether what is being done makes sense
- provide realistic estimates of the time and effort required to complete tasks
- act as intermediary between team and clients/stakeholders
- responsible for project deliverables.

Data Translators

- have a good grasp on the data and the data dictionary
- help SMEs transmit the underlying context to the data science team

Data Engineers / Database Specialists

- work with clients and stakeholders to acquire useable data sources
- may participate in the analyses, but are not necessarily specialists

ROLES AND RESPONSIBILITIES

Data Analysts

- clean and process data
- prepare initial visualizations
- have a decent understanding of quantitative methods (at most 1 area of expertise)
- conduct preliminary analyses

Data Scientists

- work with processed data to build sophisticated models
- focus on actionable insights
- have a sound understanding of algorithms/quantitative methods (2 or 3 areas of expertise)
- can apply them to a variety of data scenarios
- can be counted on to catch up on new material quickly

ROLES AND RESPONSIBILITIES

Computer Engineers

- design and build computer systems and pipelines
- are involved in software development and deployment of data science solutions

AI/ML Quality Assurance/Quality Control Specialists

- design testing plans for solutions that implement AI/ML models
- help the team determine whether the models are able to learn

Communication Specialists

- communicate actionable insights to managers, policy analysts, decision-makers, stake holders
- may participate in the analyses, but are not necessarily specialists (often data translators)
- keep abreast of popular accounts of quantitative results and developments

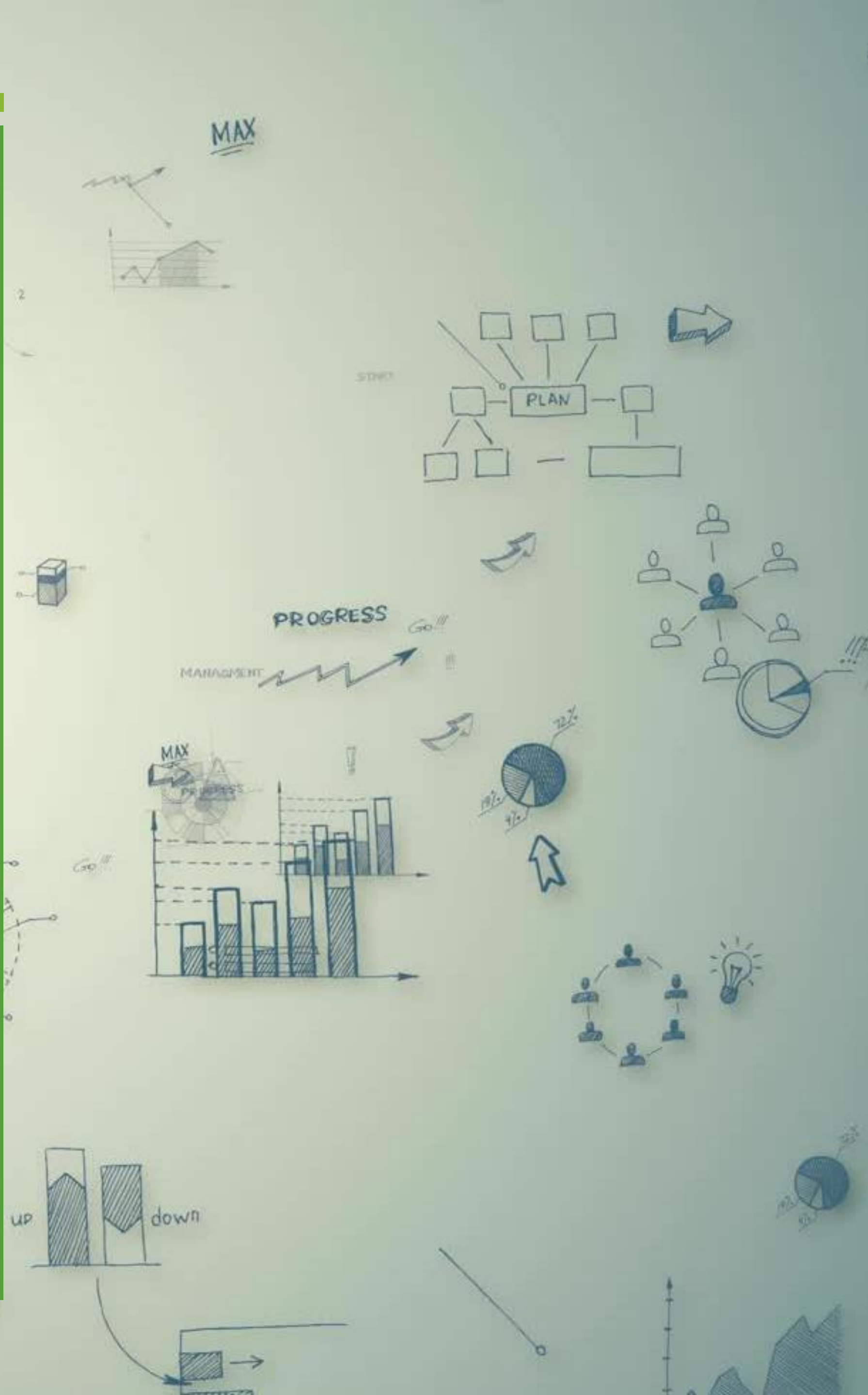
THE DIGITAL/ANALOG DATA DICHOTOMY

Humans have been collecting data for a long time; J.C. Scott argues that data collection was a major enabler of the modern nation-state.

For most of the history of data collection, we have lived in the **analogue world** (understanding grounded in continuous experience of **physical reality**).

Our data collection activities were the first steps towards a different strategy for understanding and interacting with the world.

Data leads us to conceptualize the world in a way that is **more discrete than continuous**..



THE DIGITAL/ANALOG DATA DICHOTOMY

Translating our experiences into numbers and categories, we create **sharper** and more definable boundaries than experience might suggest.

This discretization strategy leads to the **digital computer** (1;0), which is surprisingly successful at representing our physical world: the **digital world** takes on a reality as pervasive and important as the physical one.

This digital world is built on top of the physical world, but it **does not operate under the same set of rules:**

- in the physical world, the default is to **forget**; in the digital world, it is to **remember**;
- in the physical world, the default is **private**; in the digital world, the default is **public**;
- in the physical world, copying is **hard**; in the digital world, it is **easy**.

THE DIGITAL/ANALOG DATA DICHOTOMY

Digitization is making things that were **once hidden, visible; once veiled, transparent**.

Data scientists are study the **digital world**. They seek to understand:

- the **fundamental principles of data**
- how these fundamental principles manifest themselves in different digital phenomena

Ultimately, data and the digital world are **tied to the physical world**. What is done with data has repercussions in the physical world; and data scientists must have a solid grasp of the fundamentals/ context of data work before leaping into the tools and techniques that drive it forward.

DATA IN THE NEWS

Here is a sample of headlines and article titles showcasing the growing role of **data science** (DS), **machine learning** (ML), and **artificial/augmented intelligence** (AI) in different domains of society.

While these demonstrate some of the functionality/capabilities of DS/ML/AI technologies, it is important to remain aware that new technologies are always accompanied by emerging social consequences (not always positive).

DATA IN THE NEWS

- “Robots are better than doctors at diagnosing some cancers, major study finds”
- “Deep-learning-assisted diagnosis for knee magnetic resonance imaging: Development and retrospective validation of MRNet”
- “Google AI claims 99% accuracy in metastatic breast cancer detection”
- “Data scientists find connections between birth month and health”
- “Scientists using GPS tracking on endangered Dhole wild dogs”
- “These AI-invented paint color names are so bad they’re good”
- “We tried teaching an AI to write Christmas movie plots. Hilarity ensued. Eventually.”
- “Math model determines who wrote Beatles’ "In My Life": Lennon or McCartney?”

DATA IN THE NEWS

- “Scientists use Instagram data to forecast top models at New York Fashion Week”
- “How big data will solve your email problem”
- “Artificial intelligence better than physicists at designing quantum science experiments”
- “This researcher studied 400,000 knitters and discovered what turns a hobby into a business”
- “Wait, have we really wiped out 60% of animals?”
- “Amazon scraps secret AI recruiting tool that showed bias against women”
- “Facebook documents seized by MPs investigating privacy breach”
- “Firm led by Google veterans uses A.I. to ‘nudge’ workers toward happiness”
- “At Netflix, who wins when it’s Hollywood vs.the algorithm?”

DATA IN THE NEWS

- “AlphaGo vanquishes world’s top Go player, marking A.I.’s superiority over human mind”
- “An AI-written novella almost won a literary prize”
- “Elon Musk: Artificial intelligence may spark World War III”
- “A.I. hype has peaked so what’s next?”

Opinions on the topic are varied – to some, DS/ML/AI provide examples of **brilliant successes**, while to others it is the **dangerous failures** that are at the forefront.

What do you think?

Are you a glass half-full or glass half-empty sort of person when it comes to data and applications?