2. Preliminaries

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 our raw experience might suggest.

This discretization strategy leads to the **digital computer** (series of 1s and 0s), which is surprisingly successful at representing our physical world: the **digital world** is taking 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, copying is **easy**

The Digital/Analog Data Dichotomy

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

Data scientists are scientists of 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 it is crucial for data scientists to have a solid grasp of the fundamentals and context of data work before leaping into the tools and techniques that drive it forward.

What is Data?

It is difficult to give a clear-cut definition of **data** (is it singular or plural?).

Linguistically, a *datum* is "a piece of information", so **data** means "pieces of information," or **collection** of "pieces of information".

Data represents the whole (potentially greater than the sum of its parts) or simply the idealized concept.

Is that clear?

What is Data?

Is the following data?

4,529 red 25.782 Y

Why? Why not? What, if anything is missing?

The Stewart approach: “we know it when we see it.”

Pragmatically, we think of data as a collection of facts about **objects** and their **attributes**.

Objects and Attributes

Object: *apple*

- **Shape:** spherical
- **Colour:** red
- **Function:** food
- **Location:** fridge
- **Owner:** Jen

Object: *sandwich*

- **Shape:** rectangle
- **Colour:** brown
- **Function:** food
- **Location:** office
- **Owner:** Pat

Remember: an object is not simply **the sum of its attributes**.

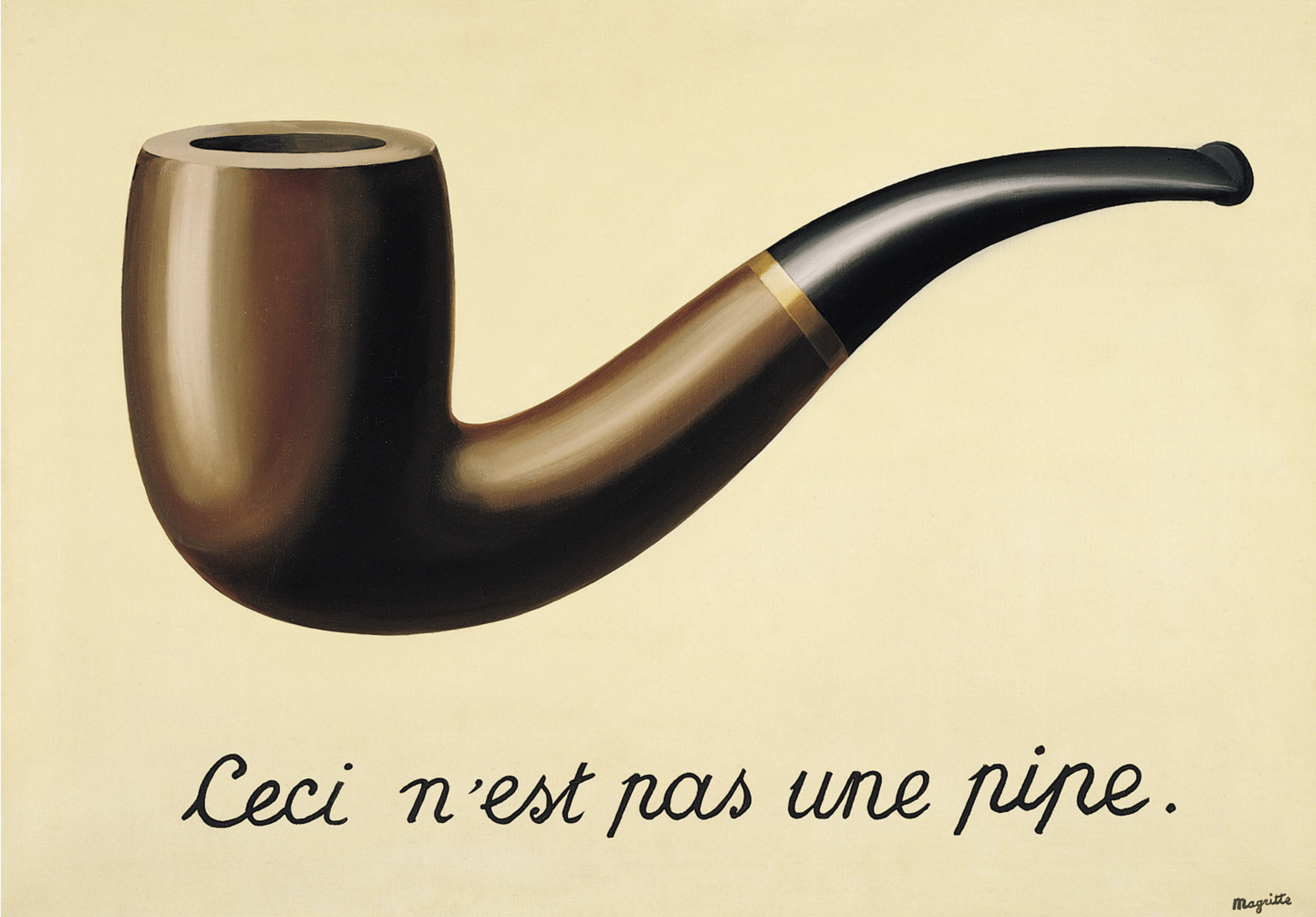
Objects and Attributes

Ambiguities when it comes to **measuring** (and **recording**) the attributes:

- apple picture is a 2-dimensional representation of a 3-dimensional object
- overall shape of the sandwich is vaguely rectangular, it is not exact (**measurement error**?)
- insignificant for most, but not necessarily all, analytical purposes
- apple's shape = volume, sandwich's shape = area (**incompatible measurements)**
- a number of potential attributes are not mentioned: size, weight, time, etc.
- are there other issues?

Measurement errors and incomplete lists are always part of the picture; is this collection of attributes providing a reasonable **description** of the objects?

[R. Magritte, La trahison des images (ceci n’est pas une pipe)]



From Objects and Attributes to Datasets

Raw data may exist in any format.

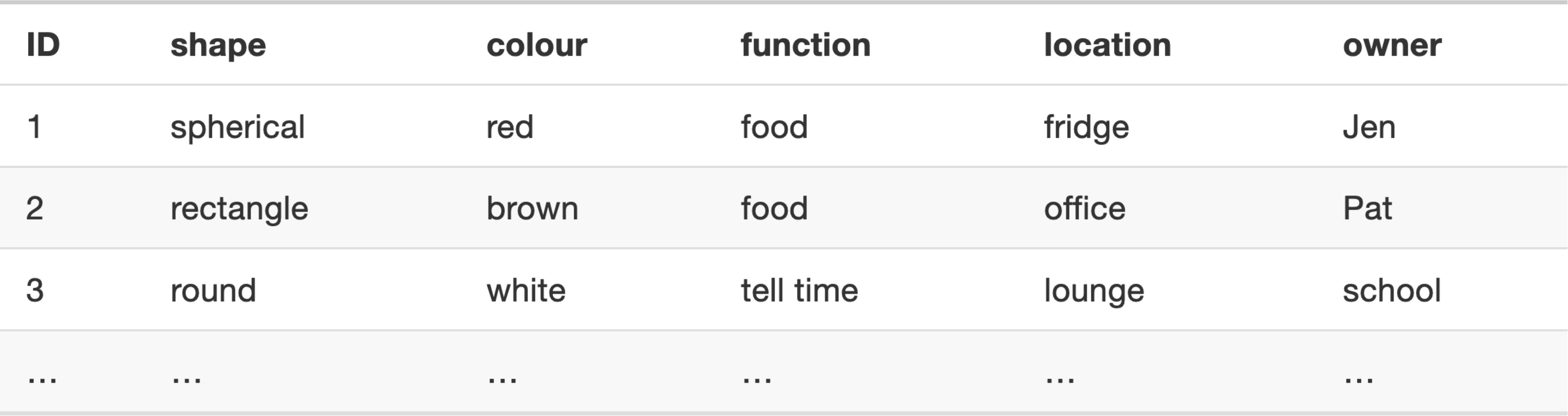
A **dataset** represents a collection of data that could conceivably be fed into algorithms for analytical purposes.

Datasets appear in a **table** format, with rows and columns; attributes are the **fields** (or columns, variables); objects are **instances** (or cases, rows, records).

Objects are described by their **feature vector** (observation's signature) – the collection of attributes associated with value(s) of interest.

From Objects and Attributes to Datasets

The dataset of physical objects could start with:

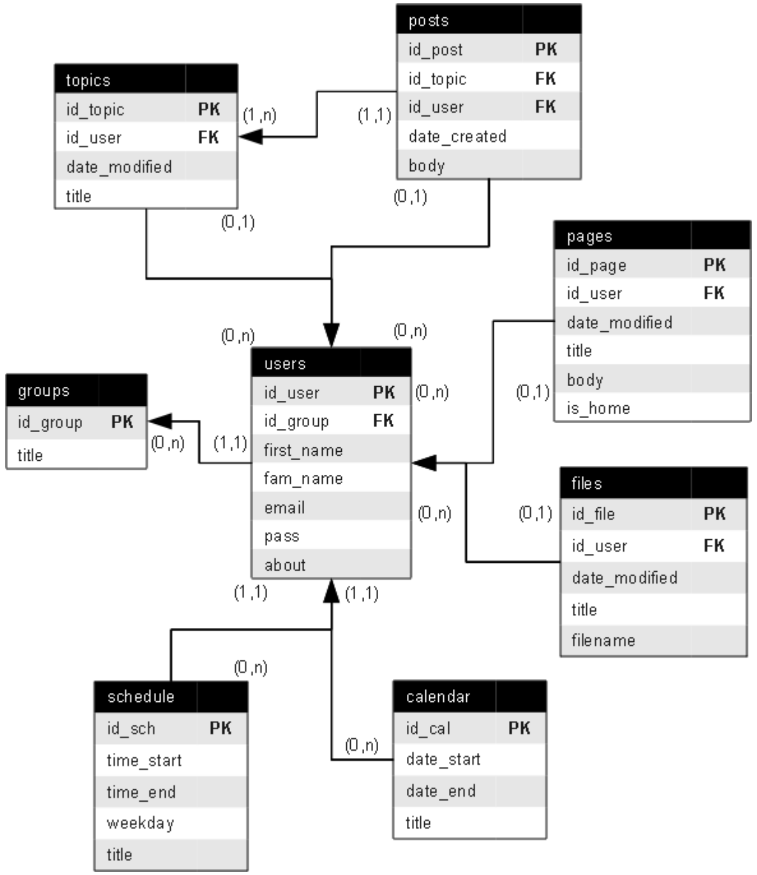


ID	shape	colour	function	location	owner
1	spherical	red	food	fridge	Jen
2	rectangle	brown	food	office	Pat
3	round	white	tell time	lounge	school
...	...	...	...	...	...

[Wooptoo, “Entity - relationship model.” Wikimedia]

From Objects and Attributes to Data

In practice, more complex **databases** are used, for a variety of reasons that we briefly discuss at a later stage.



posts	
id_post	PK
id_topic	FK
id_user	FK
date_created	
body	

topics	
id_topic	PK
id_user	FK
date_modified	
title	

pages	
id_page	PK
id_user	FK
date_modified	
title	
body	
is_home	

users	
id_user	PK
id_group	FK
first_name	
fam_name	
email	
pass	
about	

groups	
id_group	PK
title	

files	
id_file	PK
id_user	FK
date_modified	
title	
filename	

schedule	
id_sch	PK
time_start	
time_end	
weekday	
title	

calendar	
id_cal	PK
date_start	
date_end	
title	

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?”

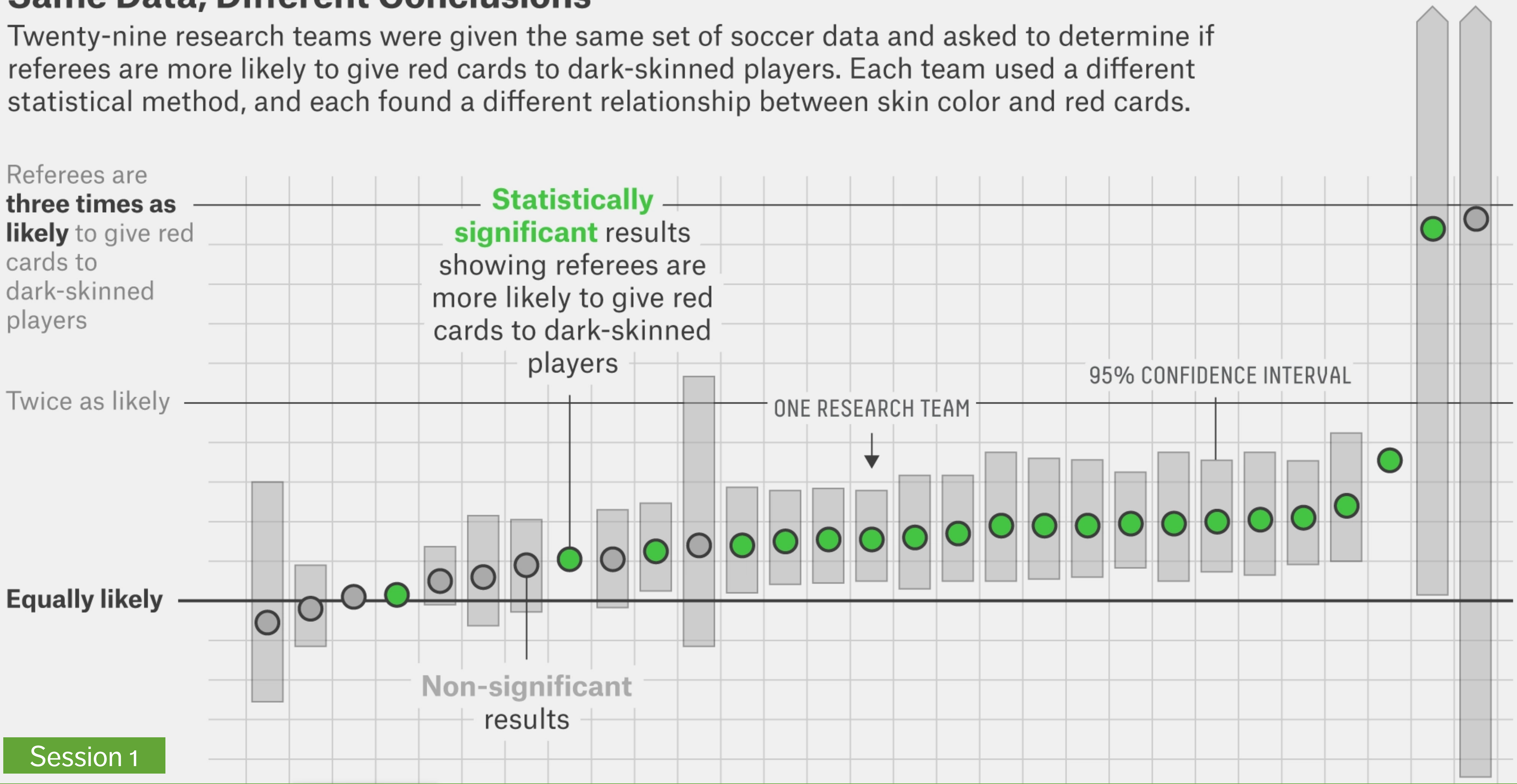
Opinions on the topic are varied – to some, DS/ML/AI provide examples of **brilliant successes**, while to others it is the **dangerous failures** that are at the forefront.

What do you think?

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

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.



Session 1

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

Suggested Reading

Preliminaries

Data Understanding, Data Analysis, Data Science
Data Science Basics

Introduction

- What is Data?
- From Objects and Attributes to Datasets
- Data in the News
- The Analog/Digital Data Dichotomy

Exercises

Preliminaries

1. 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?
2. In what format is your organization’s data available? Are you able to access it easily? Is it updated regularly? Are there data dictionaries? Have you read them?