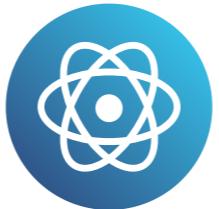


# Data sources and risks

DATA SCIENCE FOR MANAGERS



**Michael Chow**

Assessment Research Lead, DataCamp

# Common sources of data

- Web events
- Customer data
- Logistics data
- Financial transactions

# Web data

- Events
- Timestamps
- User information

<b>user_id</b>	<b>event_name</b>	<b>timestamp</b>
1234	homepage_visit	2019-01-01 12:01:01

# Personally Identifiable Information (PII)



Name	Timestamp	Object Clicked
Jane Doe	2019-01-20 12:05:00	Like Button

"Jane Doe" = Personally Identifiable Information (PII)

# Data pseudonymization



user_id	Timestamp	Object Clicked
184577	2019-01-20 12:05:00	Like Button

user_id	Name
184577	Jane Doe

- Restricted access
- Audit logs

# Data anonymization



user_id	Timestamp	Object Clicked
184577	2019-01-20 12:05:00	Like Button

A red circle with a diagonal slash through it is overlaid on a table. The table has two columns: "user\_id" and "Name". The first row shows the column headers. The second row contains the data: "184577" in the "user\_id" column and "Jane Doe" in the "Name" column. The "Name" column is highlighted with a pink background.

user_id	Name
184577	Jane Doe

# General Data Protection Regulation (GDPR)

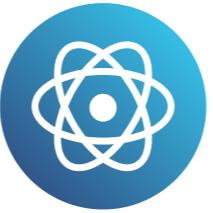
- Applies to all data inside of the EU
- Give individuals control over their personal data
- Regulates how long data can be stored
- Mandates appropriate anonymization
- Disclose data collection and gain consent

# Let's practice!

DATA SCIENCE FOR MANAGERS

# Solicited data

DATA SCIENCE FOR MANAGERS



**Michael Chow**

Assessment Research Lead, DataCamp

# Why do we solicit data?

- Create marketing collateral
- De-risk decision making
- Monitor quality



# Types of solicited data

- Surveys
- Customer reviews
- In-app questionnaires
- Focus groups

**We appreciate your feedback!** X

Thank you for visiting our website. We are always looking for ways to improve your experience. Please take a moment to tell us about your experience.

How likely are you to recommend our website to a friend or colleague?

0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10

What could we do to improve your experience?

**Send Feedback**

powered by  QuestionPro

# Types of solicited data

## Qualitative

- Conversations
- Open-ended questions

## Quantitative

- Multiple choice
- Rating scale

# Revealed and stated preferences

## Stated preference

- Hypothetical
- Subjective



## Revealed preference

- Actions
- Purchasing decisions



# Best practices

## Be specific

Do this	Not that
On a scale of 1 - 5, how would you rate the <b>quality of content</b> on DataCamp?	How would you rate DataCamp?

# Best practices

## Be specific

### Do this

On a scale of 1 - 5, how would you rate the **quality of content** on DataCamp?

### Not that

How would you rate DataCamp?

## Avoid loaded language

### Do this

Which of the following political issues is most important to you?

### Not that

Which of the following **controversial** political issues is most important to you?

# Best practices

## Calibrate

Do this	Not that
Rate your interest in each of the following products at DataCamp.	Are you interested in Skill Assessment at DataCamp?

# Best practices

## Calibrate

<b>Do this</b>	<b>Not that</b>
Are you interested in Skill Assessment at DataCamp?	Rate your interest in each of the following products at DataCamp.

## Require actionable results

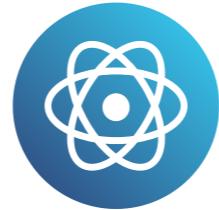
<b>Do this</b>	<b>Not that</b>
Have a hypothesis for each question	Ask a question just because it's interesting

# Let's practice!

DATA SCIENCE FOR MANAGERS

# Collecting additional data

DATA SCIENCE FOR MANAGERS



**Michael Chow**

Assessment Research Lead, DataCamp

# Even more data

- APIs
- Public records
- Mechanical Turk



# Data APIs

- Application Programming Interface
- Request data over the internet
- Twitter
- Wikipedia
- Yahoo! Finance
- Google Maps
- Many more!

# Tracking a hashtag

- All tweets with **#DataFramed** (DataCamp's podcast!)
- Use Twitter API

Hugo Bowne-Anderson @hugobowne · Mar 15

Coming at your ears next Monday -- [@jseabold](#) will break down for you the current and looming credibility crisis in [#datascience](#) on [#DataFramed](#), the [@DataCamp](#) pod.

« What is it that we do as data scientists? How do we provide value? What is our process for working? »

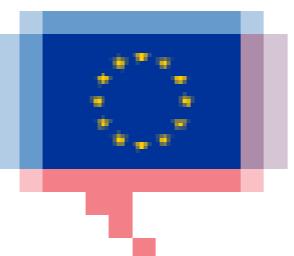
SKIPPER SEABOLD

Data  
Framed  
BY DataCamp

1 4 21

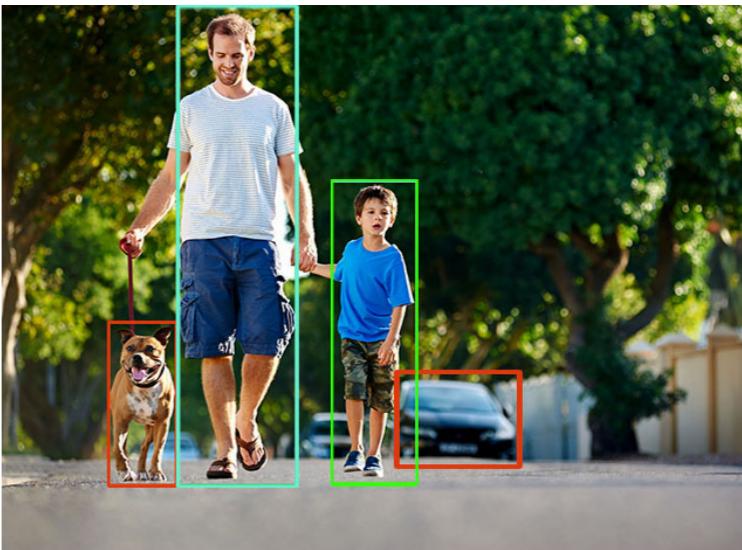
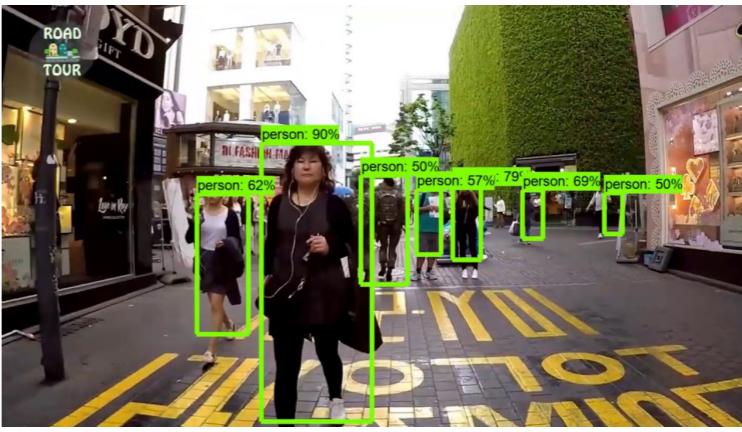
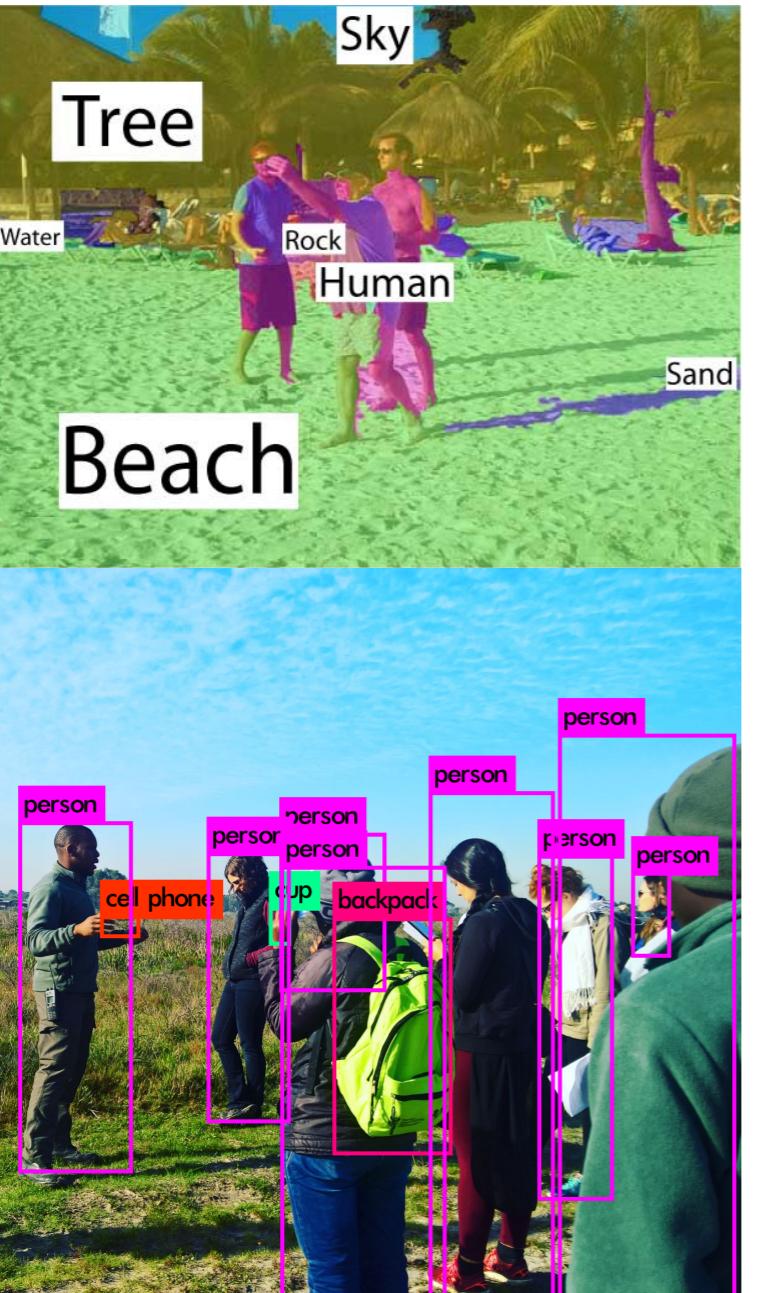
# Public records

- For the US, [data.gov](https://www.data.gov)
- For the EU, [data.europa.eu](https://data.europa.eu)



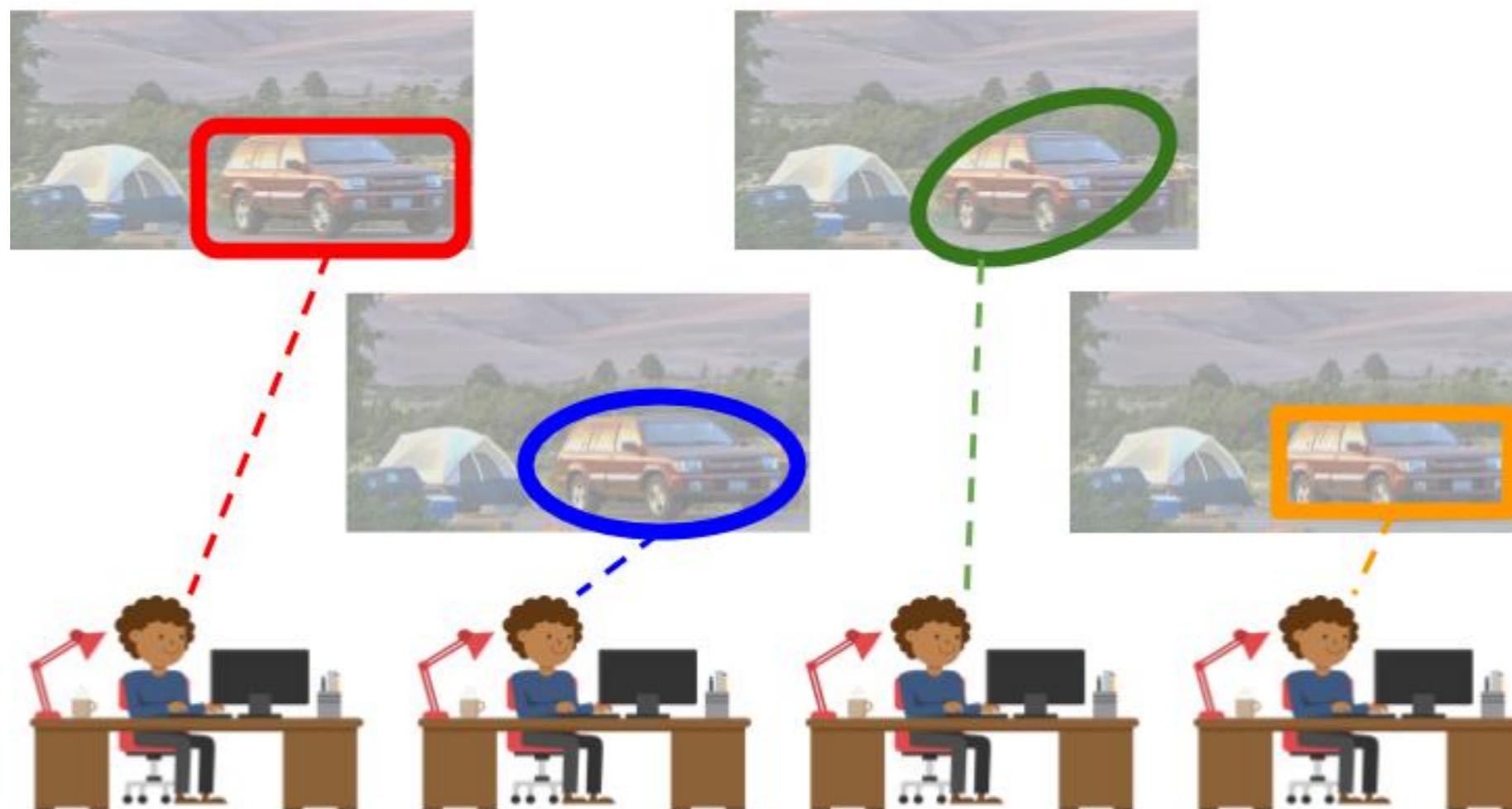
EU **Open Data** Portal

# Building a training set



# Mechanical Turk

Select the car in the image.



# Mechanical Turk

- Resource: AWS MTurk
- Label customer reviews
- Extract text from a form
- Highlight key words in a sentence

Jane  
Last Name  
Smith  
Email  
stopall11  
Pick your color:  
 Red  
 Green

Select all squares with street signs.



Submit

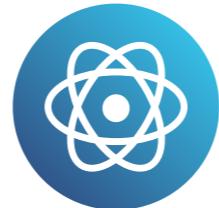
VERIFY

# Let's practice!

DATA SCIENCE FOR MANAGERS

# Data storage and retrieval

DATA SCIENCE FOR MANAGERS



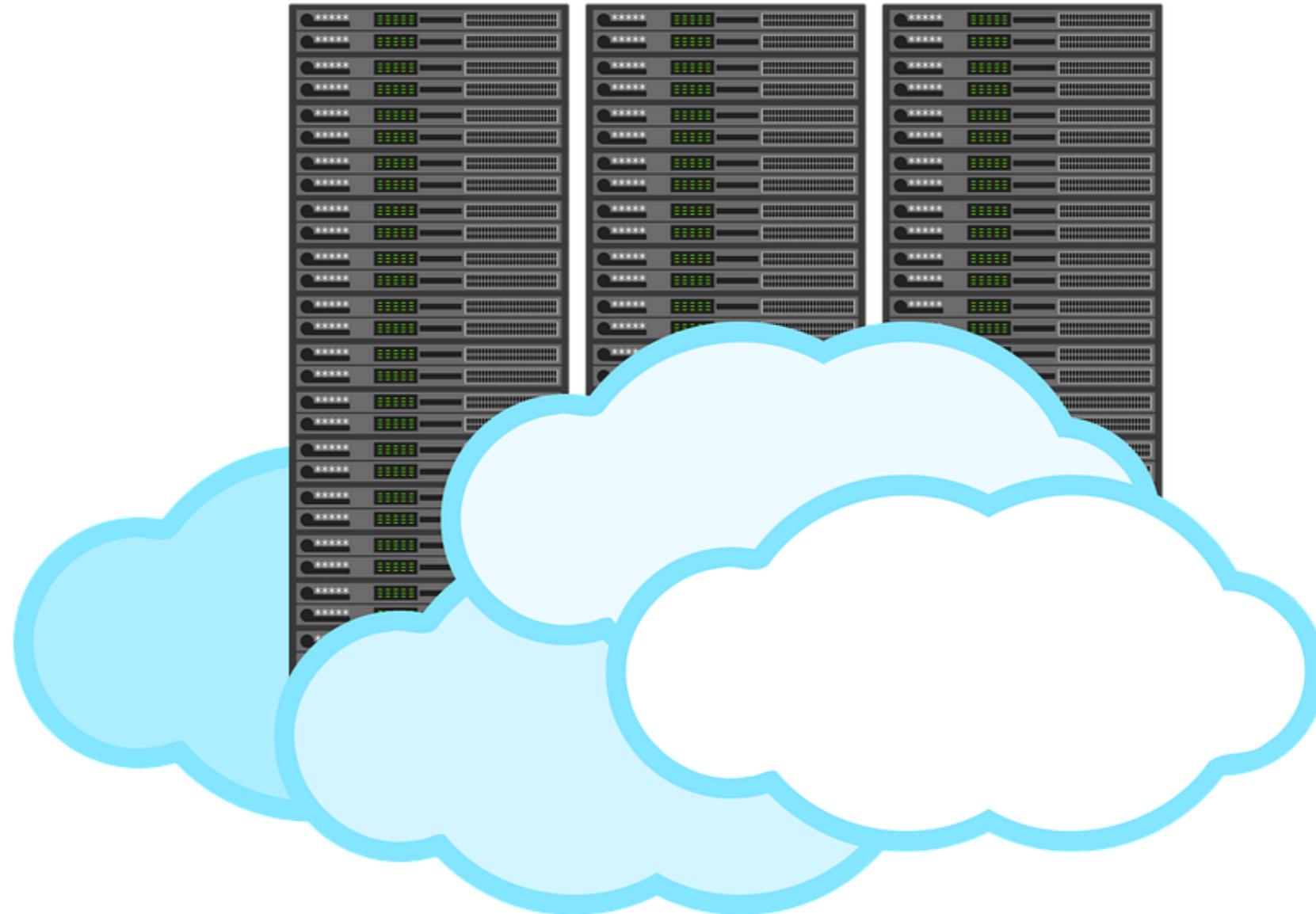
**Michael Chow**

Assessment Research Lead, DataCamp

# Parallel storage solutions



# The cloud



# Types of data storage

## Unstructured

- Email
- Text
- Video and audio files
- Web pages
- Social media

## Document Database

# Types of data storage

## Unstructured

- Email
- Text
- Video and audio files
- Web pages
- Social media

## Tabular

Customer Name	Customer Address	...
Jane Doe	123 Maple St.	...

## Relational Database

## Document Database

# Data querying



# Data querying



Data Type	Query Language
Document Database	NoSQL
Relational Database	SQL

# Putting it all together: Location



- On-premises cluster
- Cloud provider:
  - Azure
  - AWS
  - Google Cloud

# Putting it all together: Data type



# Putting it all together: Data type

Data Type	Storage Solution
Unstructured	Document Database
Tabular	Relational Database



# Putting it all together: Queries



# Putting it all together: Queries



Data Type	Query Language
Document Database	NoSQL
Relational Database	SQL

# Let's practice!

DATA SCIENCE FOR MANAGERS