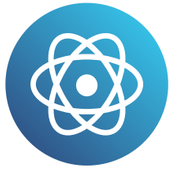
# Data sources and risks

DATA SCIENCE FOR BUSINESS



**Michael Chow**
Assessment Research Lead, DataCamp

# Common sources of data

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

# Web data

- Events
- Timestamps
- User information

| user_id | event_name | timestamp |
|---|---|---|
| 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

Jane can't be identified by the events table alone, but she can be identified if we combine information from the users table with the events table.
To protect Jane, we'll want to make sure that access to the users table is restricted to only folks who need Jane's identity.
We'll also want to periodically audit who has accessed this data and how they have used it to ensure that Jane's data is respected.

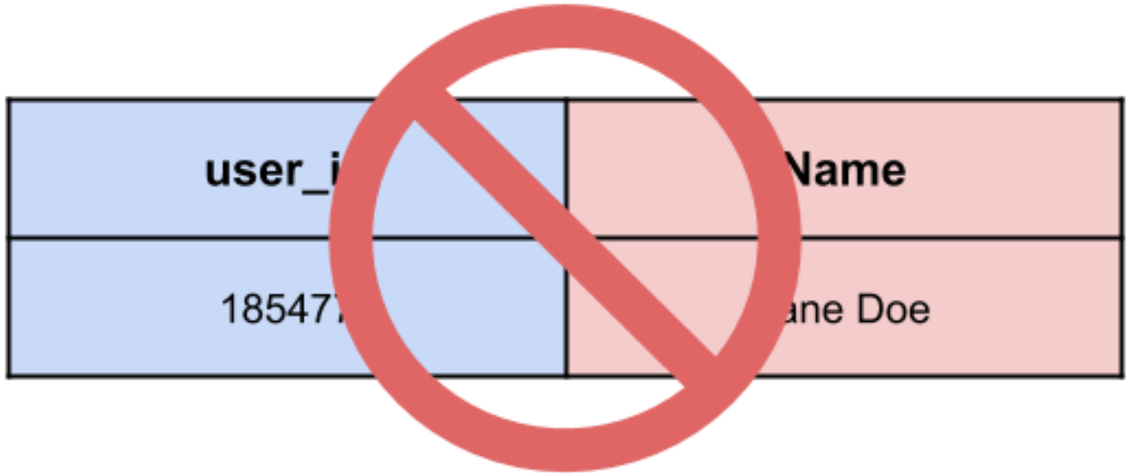
| user_id | Timestamp | Object Clicked |
|---|---|---|
| 185477 | 2019-01-20 12:05:00 | Like Button |

| user_id | Name |
|---|---|
| 185477 | Jane Doe |

- Restricted access
- Audit logs

# Data anonymization

| user_id | Timestamp | Object Clicked |
| --- | --- | --- |
| 185477 | 2019-01-20 12:05:00 | Like Button |



# 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 BUSINESS

# Solicited data

DATA SCIENCE FOR BUSINESS

**Michael Chow**
Assessment Research Lead, DataCamp

# Why do we solicit data?

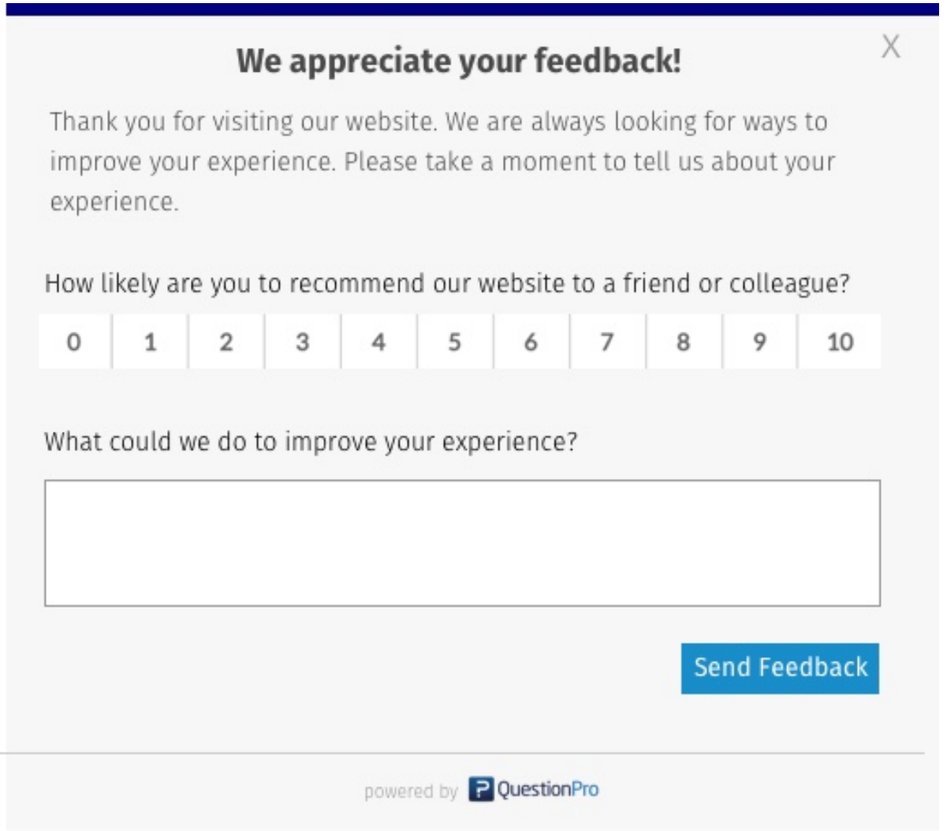
obtain by asking your customers for their opinions

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

Biased questions are only appropriate for marketing collateral, while a simple rating is great for quality monitoring. Comparisons to existing products and gauging interest can help de-risk the decision to introduce a new product.

# Types of solicited data

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



# Types of solicited data

## Qualitative

- Conversations
- Open-ended questions

## Quantitative

- Multiple choice
- Rating scale

In general, collection of small-scale qualitative data is good for generating hypotheses. For example, a focus group might provide some ideas about what features we might want to build.
Larger scale quantitative collection is needed to validate these hypotheses. For example, we can ask users to rank a list of features from most desirable to least desirable.

# Revealed and stated preferences

## Stated preference

- Hypothetical
- Subjective

Net Promoter Score (or NPS) is a common metric used to track the success of a product or website. It's measured by asking a simple question:
How likely are you to recommend this product or website to a friend or colleague?
Users respond on a scale of 0 - 10 with 0 being not at all likely to recommend and 10 being extremely likely to recommend.

## Revealed preference

- Actions
- Purchasing decisions

# Best practices

## Be specific

| Do this | Not that |
| --- | --- |
| On a scale of 1 - 5, how would you rate the **quality of content** on DataCamp? | How would you rate DataCamp? |

# Best practices

## Be specific

| Do this | Not that |
| --- | --- |
| On a scale of 1 - 5, how would you rate the **quality of content** on DataCamp? | How would you rate DataCamp? |

## Avoid loaded language

| Do this | Not that |
| --- | --- |
| Which of the following political issues is most important to you? | 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

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

## Require actionable results

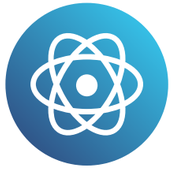
| Do this | Not that |
| --- | --- |
| Have a hypothesis for each question. | Ask a question just because it's interesting. |

# Let's practice!

DATA SCIENCE FOR BUSINESS

# Collecting additional data

DATA SCIENCE FOR BUSINESS



**Michael Chow**
Assessment Research Lead, DataCamp

# Even more data

- APIs
- Public records
- Mechanical Turk

# Data APIs

- **A**pplication **P**rogramming **I**nterface
- Request data over the internet

- Twitter
- Wikipedia
- Yahoo! Finance
- Google Maps
- Many more!

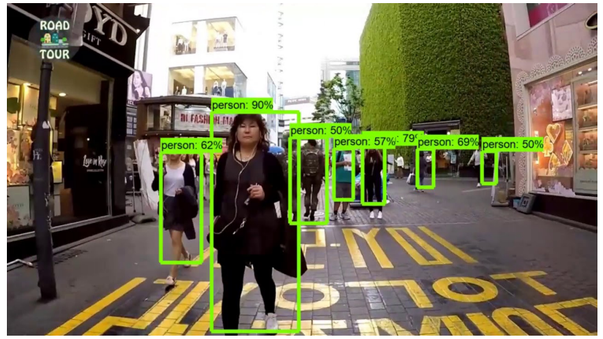
# Tracking a hashtag

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

# Public records

- For the US, **data.gov**
- For the EU, **data.europa.eu**

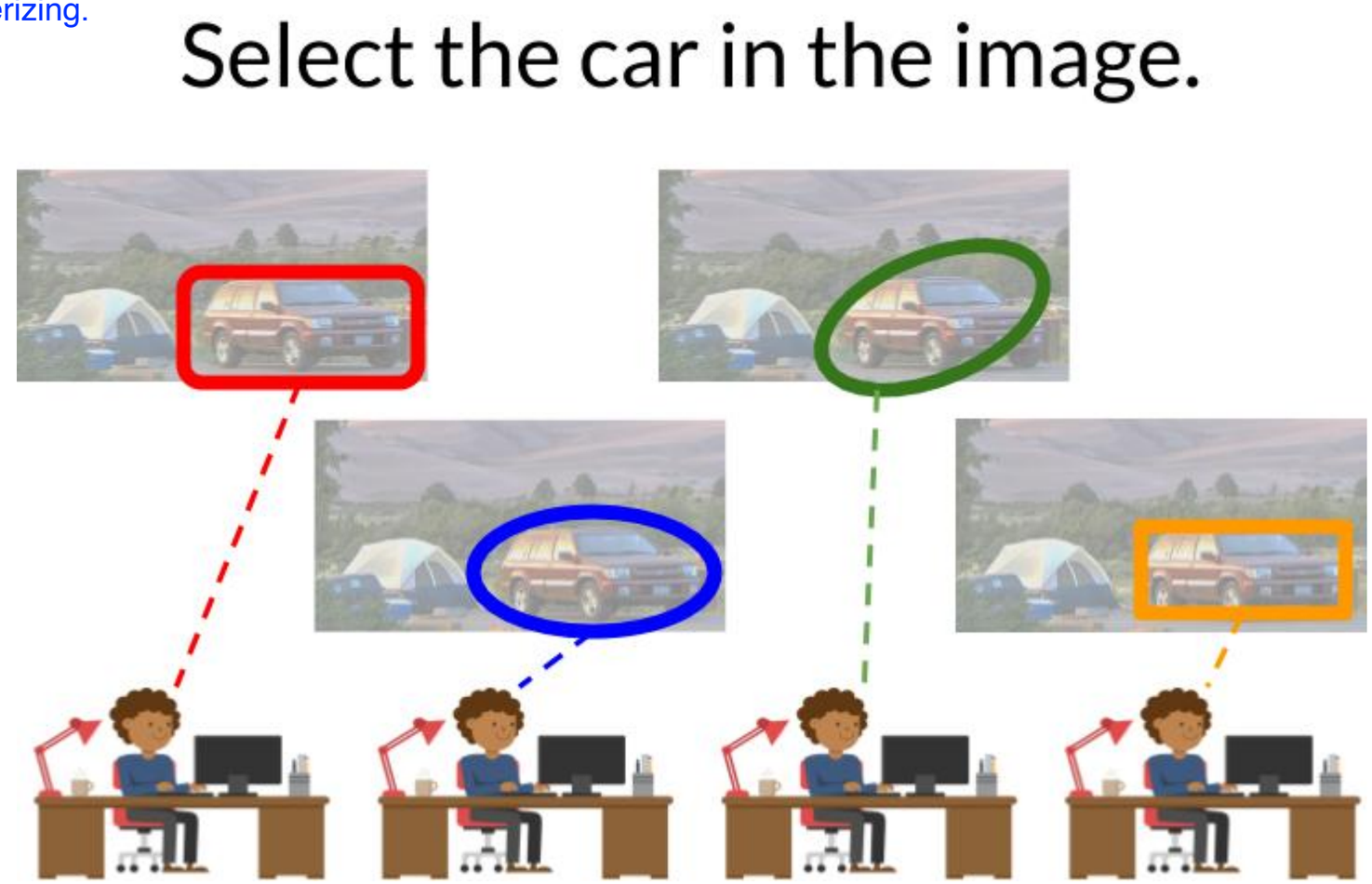

# Building a training set



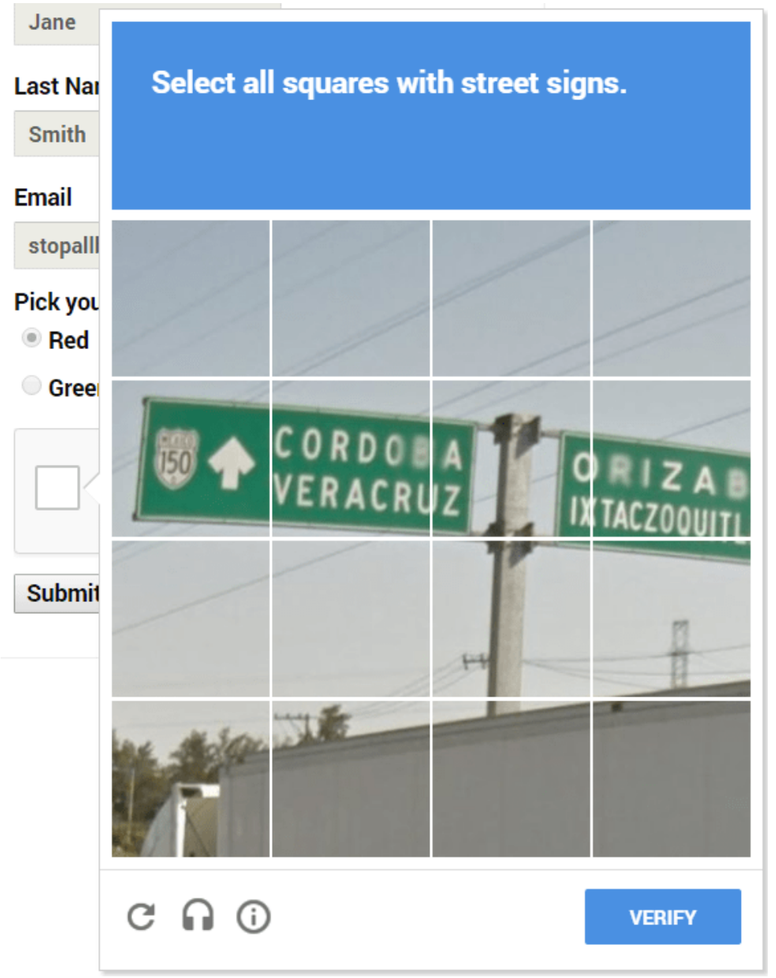
# Mechanical Turk

MTurk means asking humans to complete a task that we eventually plan on computerizing.



# Mechanical Turk

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

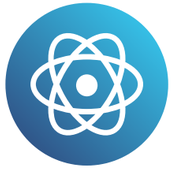


# Let's practice!

DATA SCIENCE FOR BUSINESS

# Data storage and retrieval

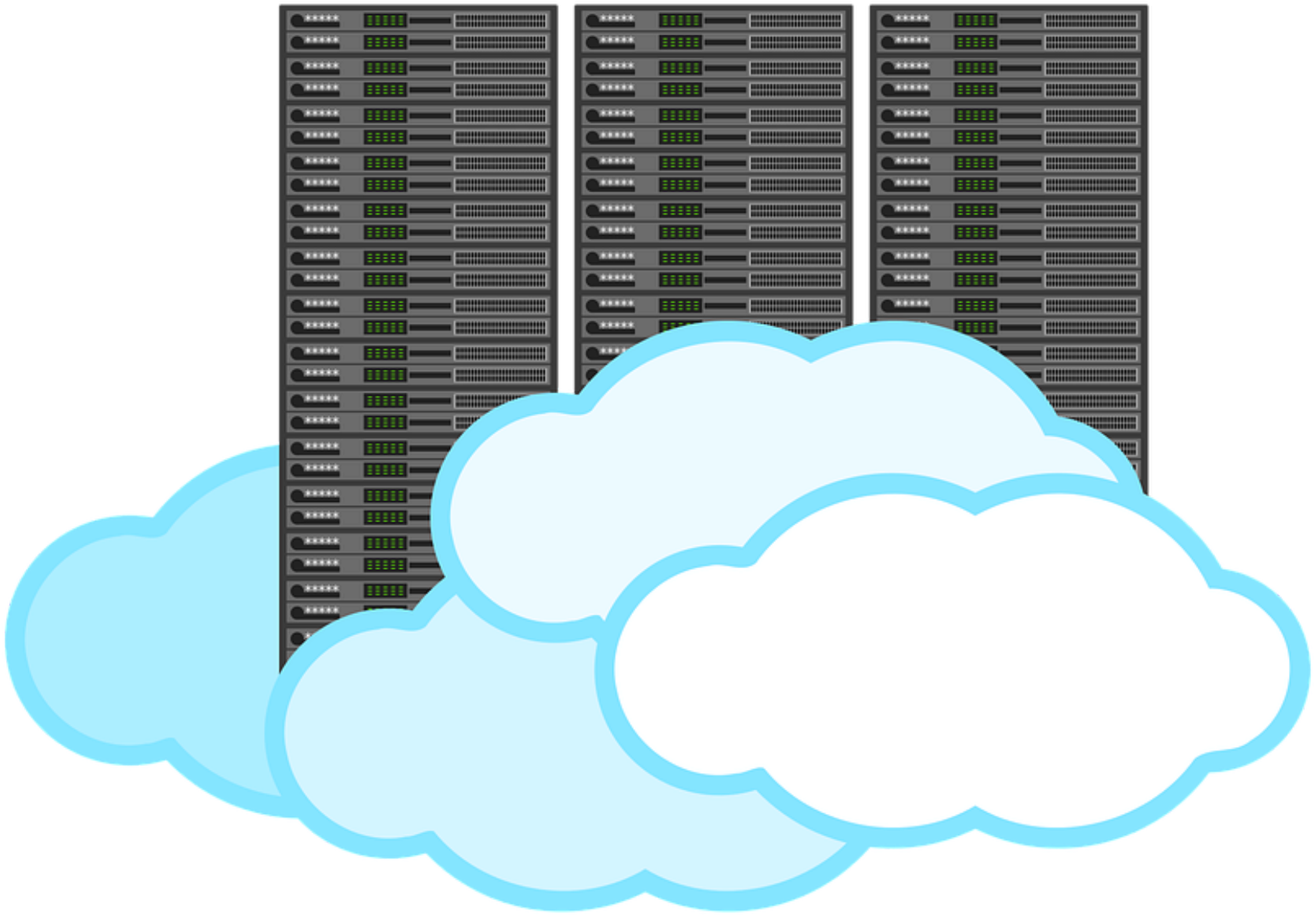
DATA SCIENCE FOR BUSINESS



**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

**Document Database**

Tabular

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

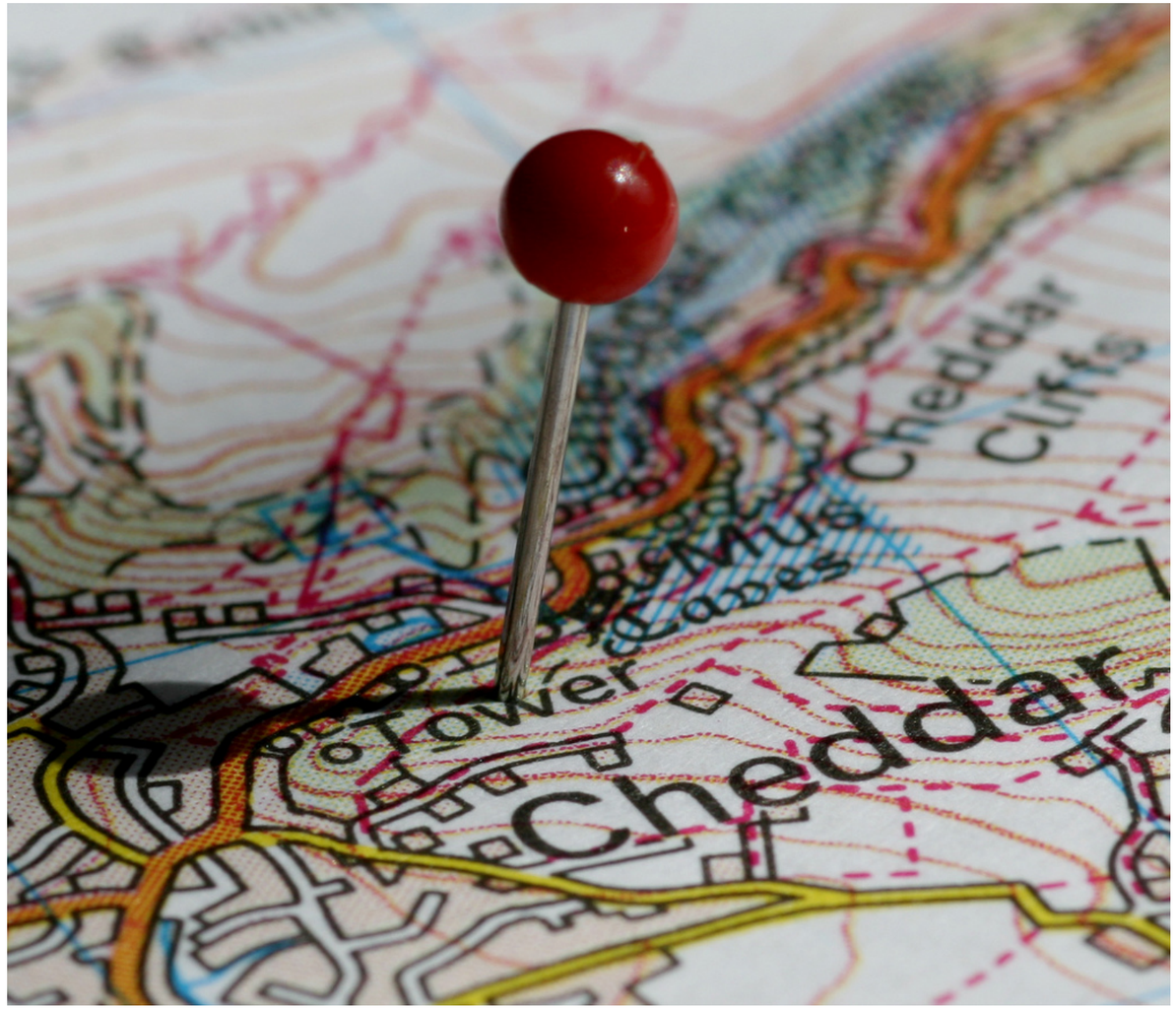
**Relational Database**

# Data querying

# Data querying

| Data Type | Query Language |
| --- | --- |
| Document Database | NoSQL Not only SQL |
| 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 BUSINESS