Data sources and risks

DATA SCIENCE FOR BUSINESS

Michael Chow
Data Scientist

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

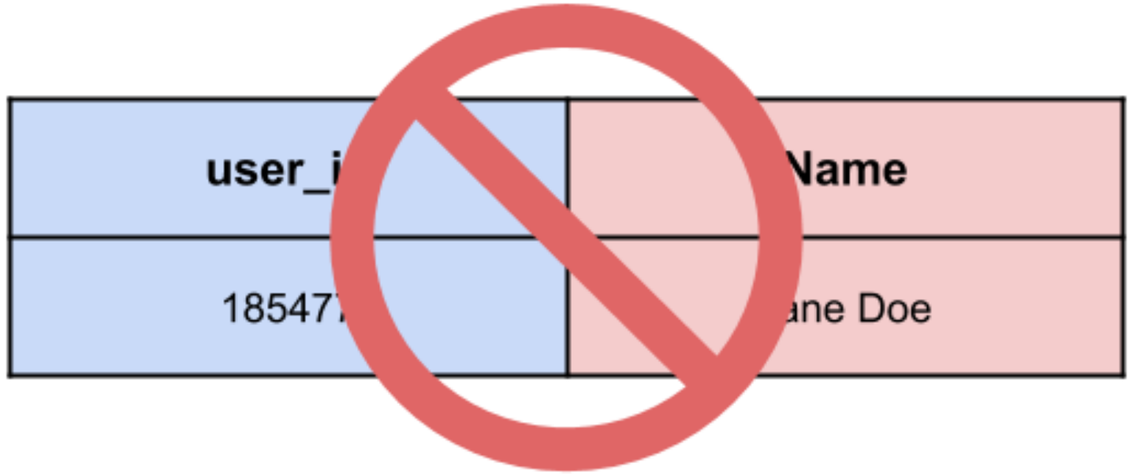
user_id	Timestamp	Object Clicked
185477	2019-01-20 12:05:00	Like Button

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
Data Scientist

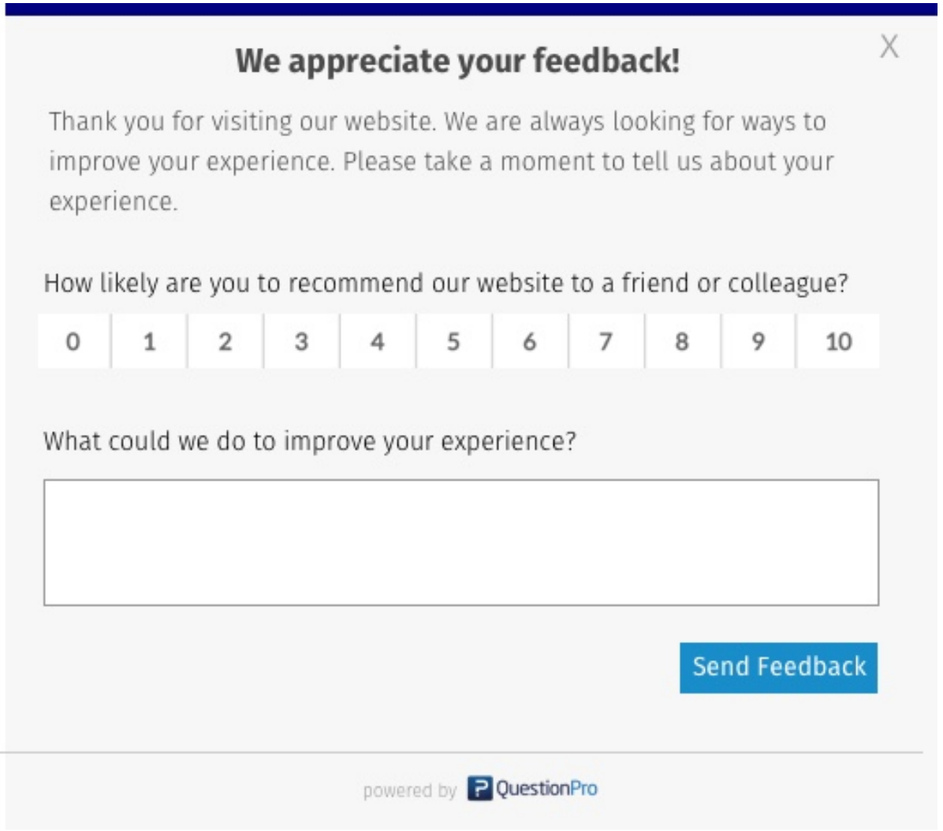
datacamp

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



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 **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
Data Scientist

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

DataFramed

DataCamp's official podcast. Presented by Hugo Bowne-Anderson.

Data Science is one of the fastest growing industries and has been called the « Sexiest job of the 21st Century ». But what exactly is Data Science? In the podcast by DataCamp, Hugo Bowne-Anderson approaches this question from the perspective of what problems Data Science tries to solve instead of what definition fits it best. From automated medical diagnosis and self-driving cars to recommendation systems and climate change, come on a journey with industry and academic experts to explore the inner workings of the industry that will color the 21st century.

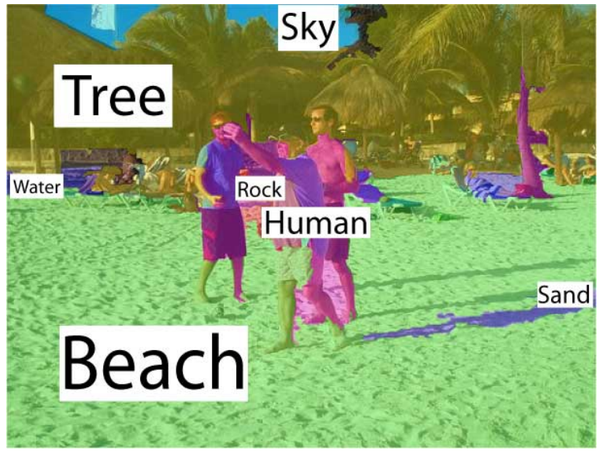
Get it on iTunes

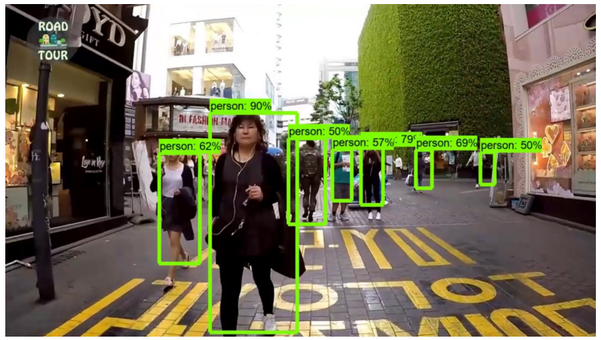
Listen on Google Play Music

Public records

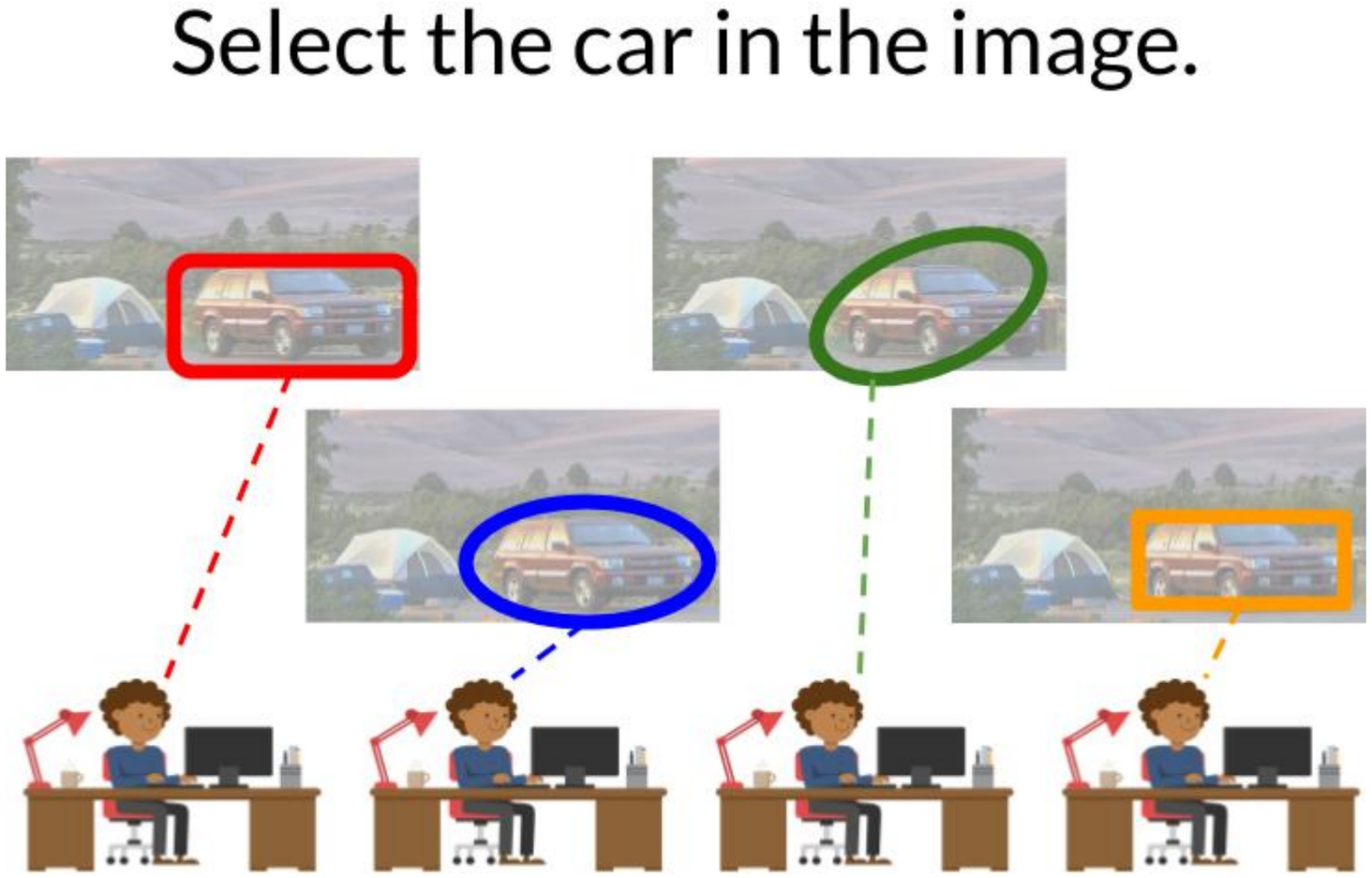
- For the US, **data.gov**
- For the EU, **data.europa.eu**

Building a training set



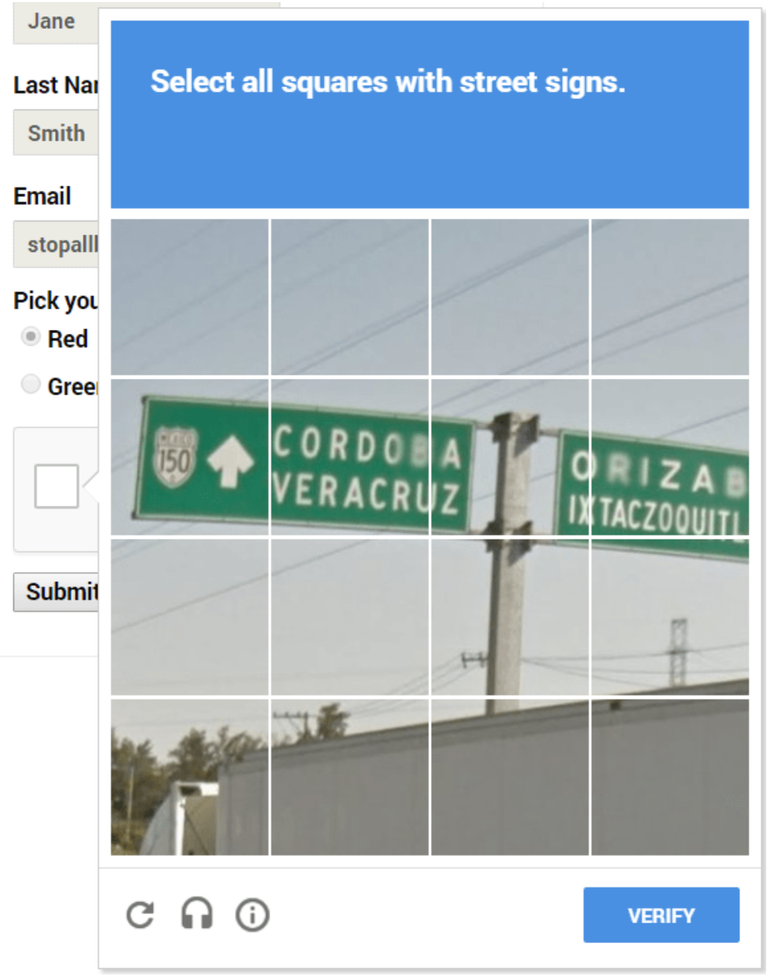


Mechanical Turk



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

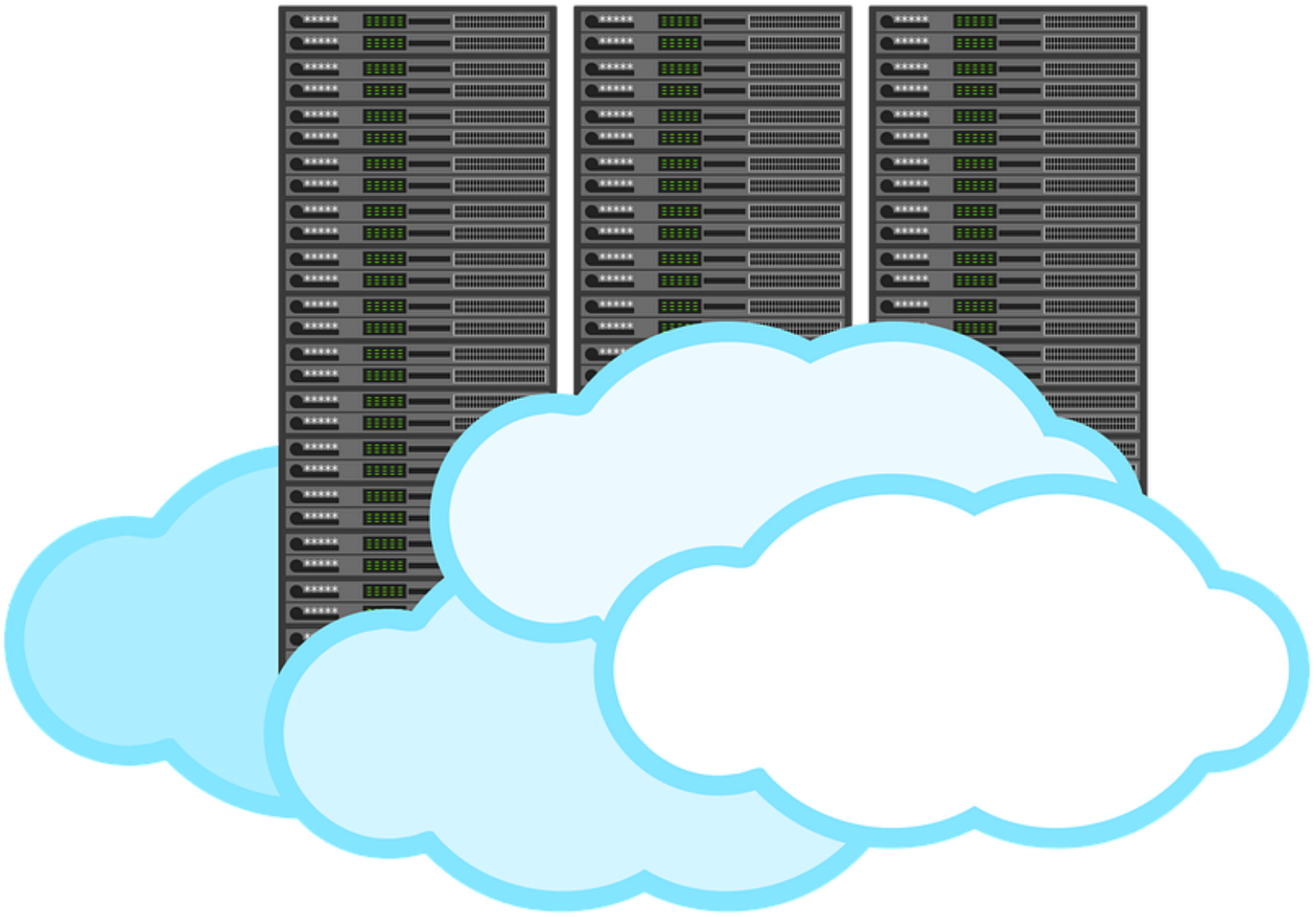
DATA SCIENCE FOR BUSINESS

Michael Chow
Data Scientist

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
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