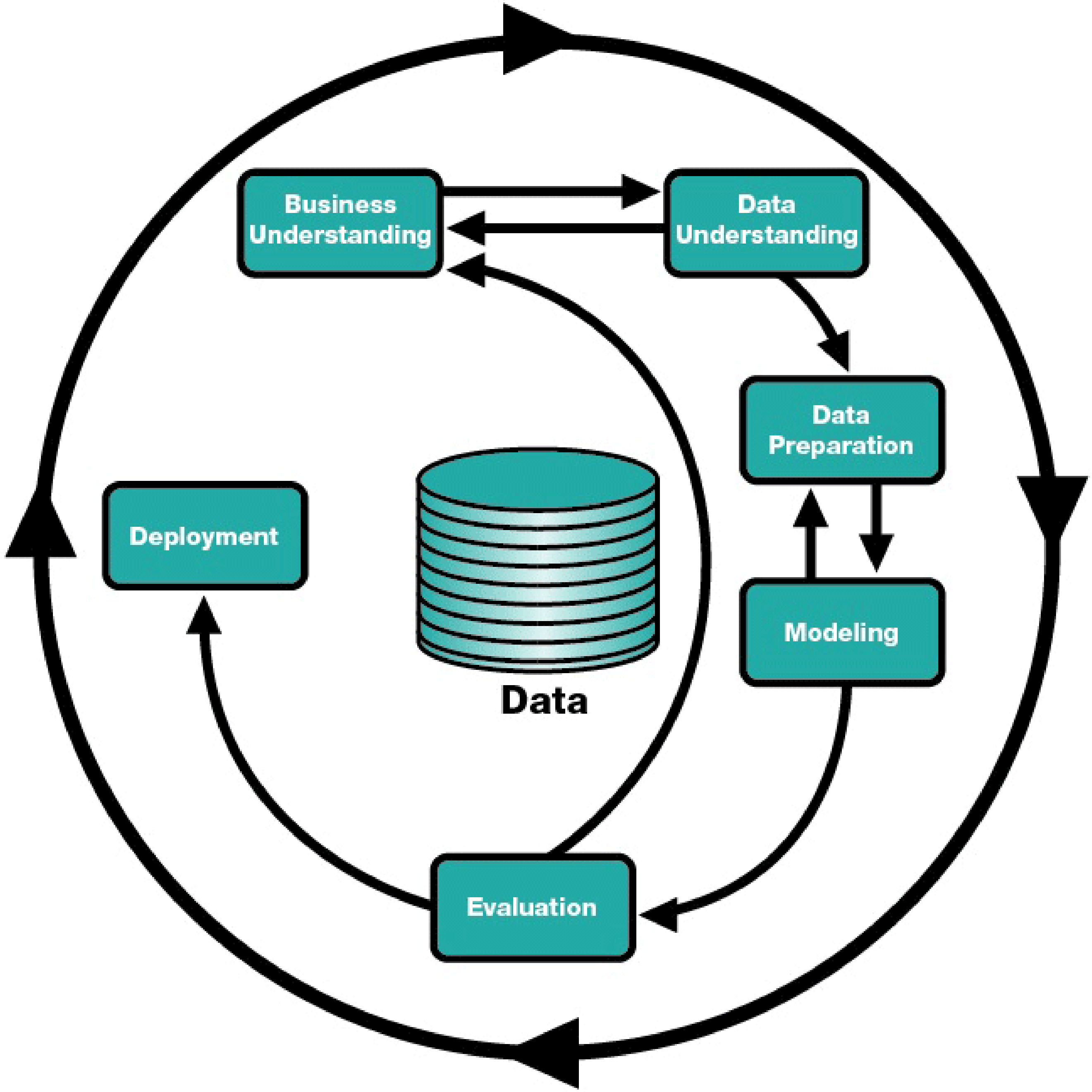
Machine Learning Techniques

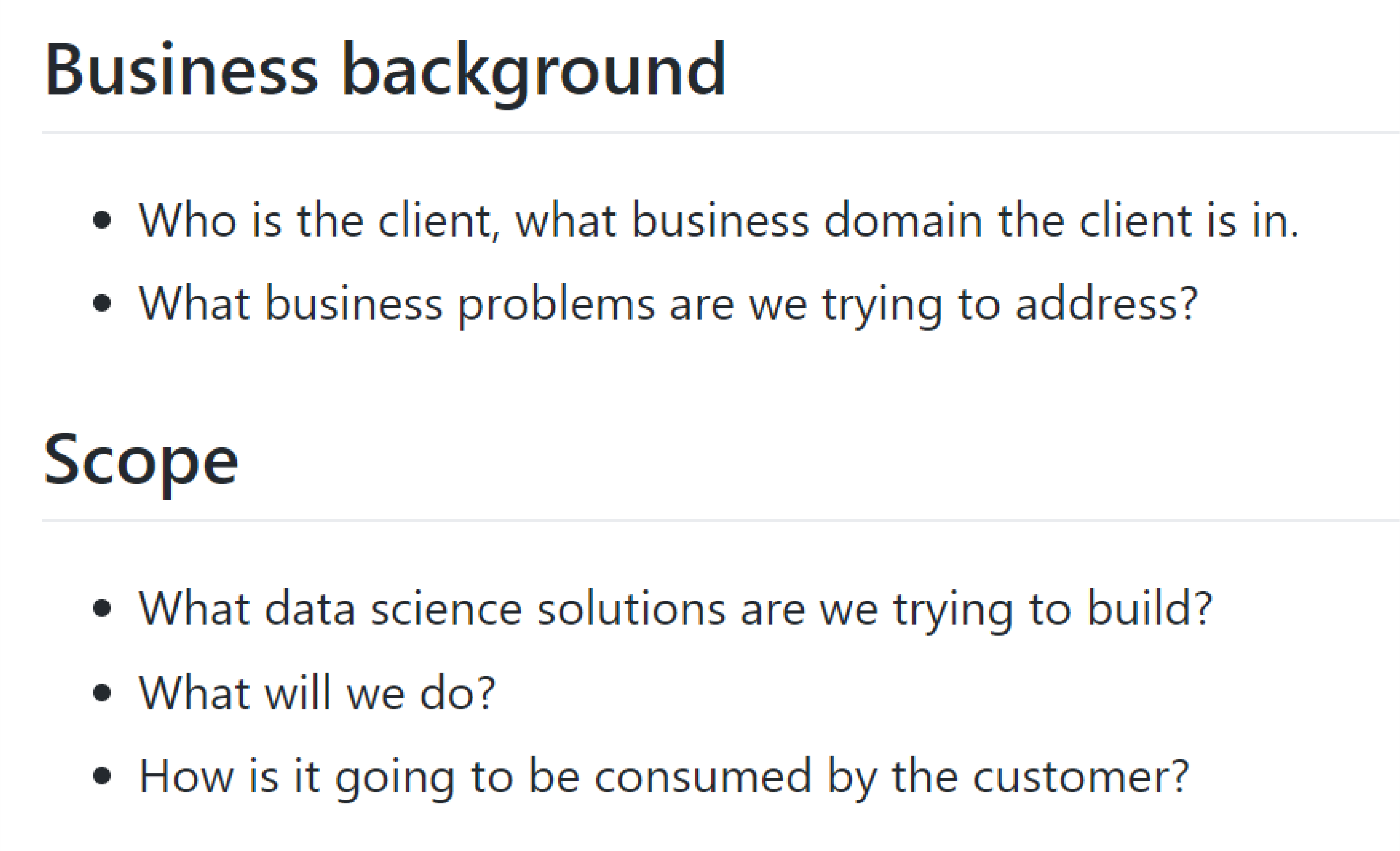
DATASCI 420

Lesson 01-2 CRISP-DM

CRISP-DM


Business Understanding
Data Understanding
Data Preparation
Modeling
Evaluation
Deployment
Data

Business Understanding



Business background

- Who is the client, what business domain the client is in.
- What business problems are we trying to address?

Scope

- What data science solutions are we trying to build?
- What will we do?
- How is it going to be consumed by the customer?

Examples of Scoping Down A Data Science Project (1)

Example 1: An F1 Racing Car Team

“I want to optimize my racing strategy” – Customer’s request

Questions to ask:

- What is the current racing strategy?
- How do you determine whether a racing car needs to change tires or refuel?
- What are the decision factors that are qualitative or experience-based?
- Whether there is any decision factors that can be data-driven?

Project #1:

Build a machine learning model to predict the tire surface temp

Examples of Scoping Down A Data Science Project (2)

Example 2: An Speaker Manufacturer

"I want to improve the efficiency of my customer service" – Customer's request

Questions to ask:

- What is the major pain point of you customer service department or your customers
- What are you doing now when you are tackling this major pain point

Project #1:

Build a machine learning model to do automatic fault diagnosis and propose solutions to customers

Success Criteria and Deployment Plan

- What is the performance of the current system?
- What is the expected performance of the data science solution
  - Be optimistic
  - But never over-commit
- If the expected performance is achieved, what is the plan of deployment?
  - It might take 2-3 months to complete a data science project
  - Management priority might have changed
  - But you should still be able to claim your project a successful as long as you reach the expected performance

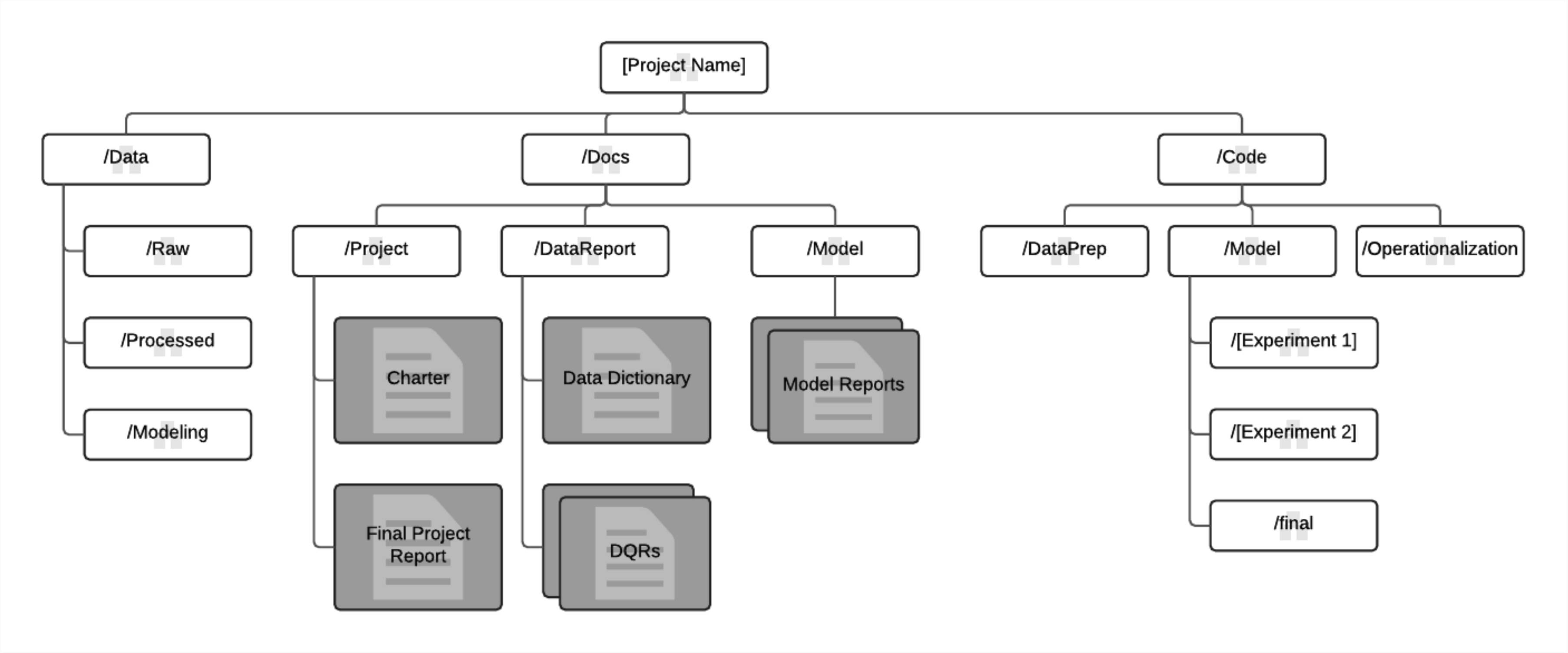
Data Understanding

- What possibly relevant data are there?
- Which is the most relevant set of data?
- Are they relevant to he business problem we want to address?
- With basic feature engineering on the most relevant set of data, how good the machine learning model can perform? (baseline model)
- If no relevant data available, any other business problem can be addressed by this data? Alternatively, any other data source might be relevant to the selected business problem?

Standardized Git Repository and Docs, Shared Productivity Utilities

- One git repository per project
- Standardized git repository directory structure
- A set of standardized document templates
- A shared data science utility repository, and a process to enrich it over time

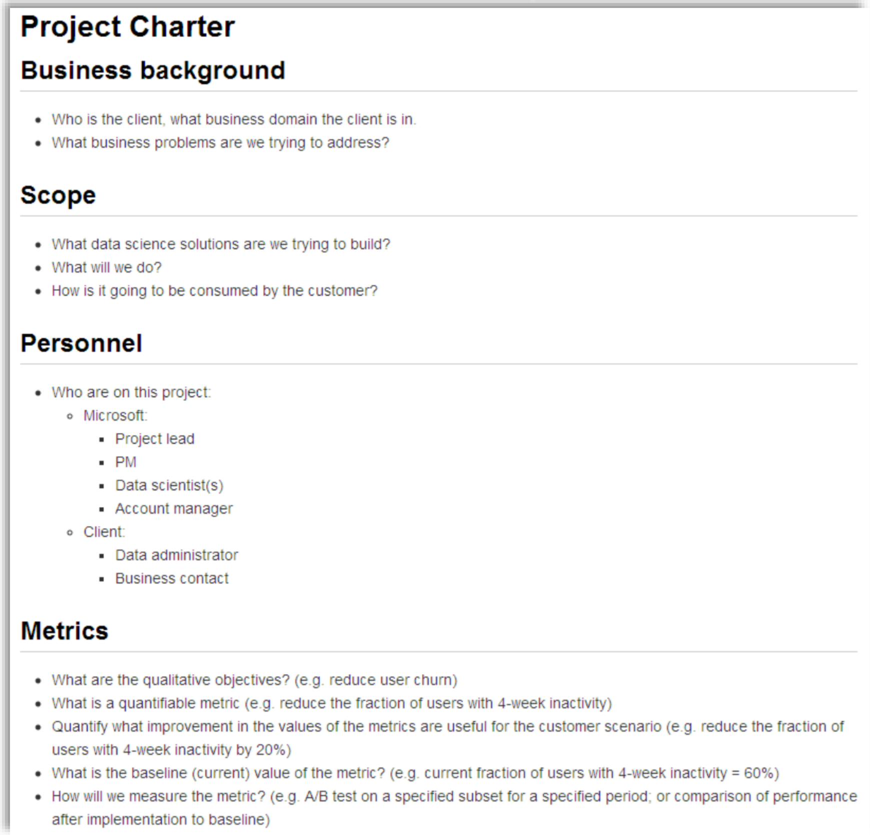
Templates



Standardized Git Repository and Docs, Shared Productivity Utilities (2)

- One git repository per project
- Standardized git repository directory structure
- A set of standardized document templates
- A shared data science utility repository, and a process to enrich it over time

Templates



Project Charter

Business background

- Who is the client, what business domain the client is in.
- What business problems are we trying to address?

Scope

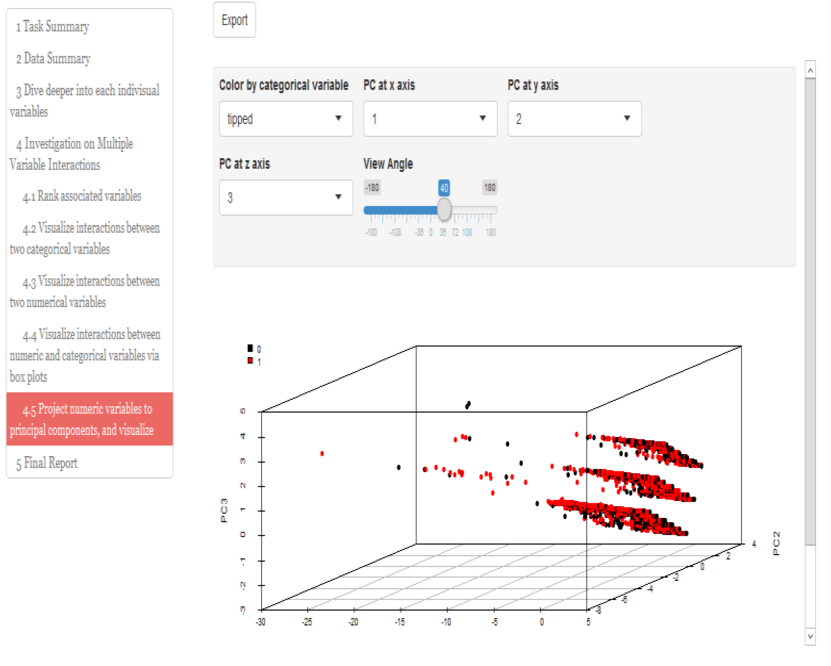
- What data science solutions are we trying to build?
- What will we do?
- How is it going to be consumed by the customer?

Personnel

- Who are on this project:
  - Microsoft:
    - Project lead
    - PM
    - Data scientist(s)
    - Account manager
  - Client:
    - Data administrator
    - Business contact

Metrics

- What are the qualitative objectives? (e.g. reduce user churn)
- What is a quantifiable metric (e.g. reduce the fraction of users with 4-week inactivity)
- Quantify what improvement in the values of the metrics are useful for the customer scenario (e.g. reduce the fraction of users with 4-week inactivity by 20%)
- What is the baseline (current) value of the metric? (e.g. current fraction of users with 4-week inactivity = 60%)
- How will we measure the metric? (e.g. A/B test on a specified subset for a specified period; or comparison of performance after implementation to baseline)



1 Task Summary
2 Data Summary
3 Dive deeper into each indivisual variables
4 Investigation on Multiple Variable Interactions
4.1 Rank associated variables
4.2 Visualize interactions between two categorical variables
4.3 Visualize interactions between two numerical variables
4.4 Visualize interactions between numeric and categorical variables via box plots
4.5 Project numeric variables to principal components, and visualize
5 Final Report
Export

Color by categorical variable: tipped

PC at x axis: 1

PC at y axis: 2

PC at z axis: 3

View Angle: 40

-180
180
-180 -108 -36 0 36 72 108 180
0
1
PC3
-3 -2 -1 0 1 2 3 4 5
-30 -25 -20 -15 -10 -5 0 5
PC2
-8 -6 -4 -2 0 2 4

Why we recommend separate git repositories for different data science projects?

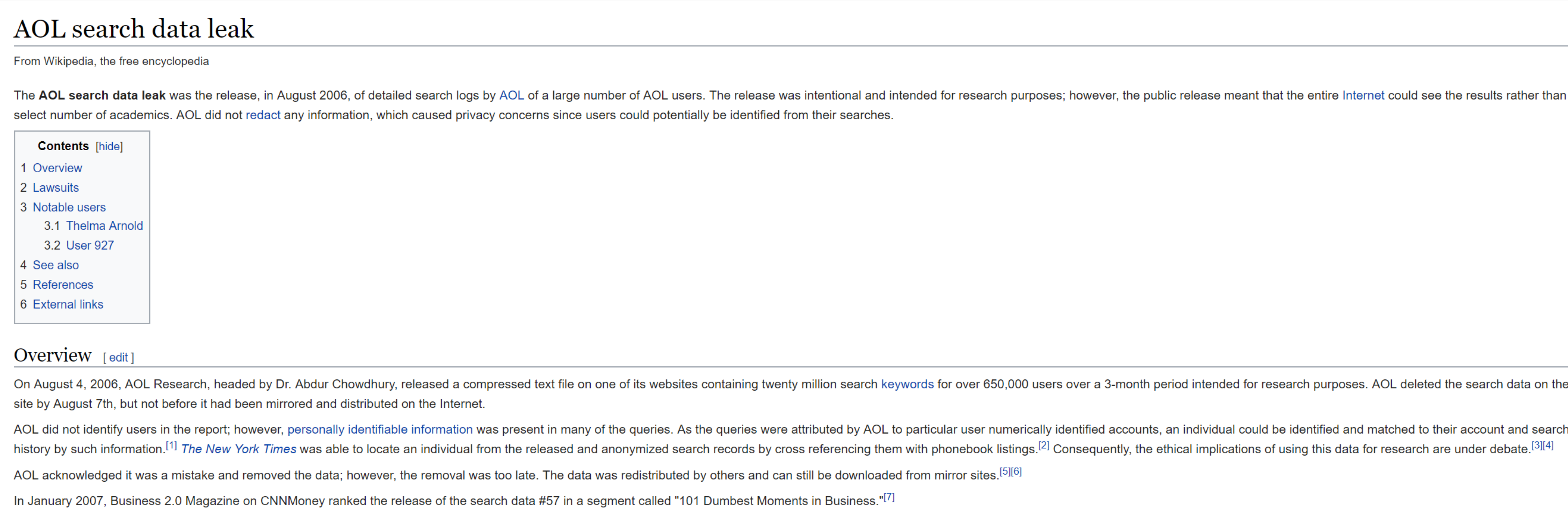
- Track all changes in your code
- Rollback bad code
- Git operations are reversible
- Facilitate collaboration
  - Work on the same code simultaneously
  - Code reviews
  - Identify and resolve conflicts
- Recover lost repos
- Operations are local
  - No need to repeatedly download from server for every operation
  - Can work offline

Three Things to Keep in Mind when Doing Data Science

- Keep the data/model ***SECURED*** as required by your customer

- Keep the data/model ***SECURED*** as required by your customer

- Keep the data/model ***SECURED*** as required by your customer



AOL search data leak

From Wikipedia, the free encyclopedia

The **AOL search data leak** was the release, in August 2006, of detailed search logs by AOL of a large number of AOL users. The release was intentional and intended for research purposes; however, the public release meant that the entire Internet could see the results rather than select number of academics. AOL did not redact any information, which caused privacy concerns since users could potentially be identified from their searches.

Contents [hide]

Overview [edit]

On August 4, 2006, AOL Research, headed by Dr. Abdur Chowdhury, released a compressed text file on one of its websites containing twenty million search keywords for over 650,000 users over a 3-month period intended for research purposes. AOL deleted the search data on the site by August 7th, but not before it had been mirrored and distributed on the Internet.

AOL did not identify users in the report; however, personally identifiable information was present in many of the queries. As the queries were attributed by AOL to particular user numerically identified accounts, an individual could be identified and matched to their account and search history by such information.[1] *The New York Times* was able to locate an individual from the released and anonymized search records by cross referencing them with phonebook listings.[2] Consequently, the ethical implications of using this data for research are under debate.[3][4]

AOL acknowledged it was a mistake and removed the data; however, the removal was too late. The data was redistributed by others and can still be downloaded from mirror sites.[5][6]

In January 2007, Business 2.0 Magazine on CNNMoney ranked the release of the search data #57 in a segment called "101 Dumbest Moments in Business."[7]

https://en.wikipedia.org/wiki/AOL_search_data_leak