Natural Language Processing(NLP)

معالجة اللغات الطبيعية

تطبيق عملي-الدرس الاول

SMS Spam Filtering (تصفية الرسائل غير الرغوب فيها)

Natural Language Processing

NLP

معالجة اللغات الطبيعية

Ahmad Shhadeh

From the previous session

1. What is NLP?
2. Structured Data vs. Unstructured Data
3. NLP Components
4. NLP, Artificial intelligence & Machine Learning
5. Main approaches in NLP(TimeLine)
6. Why NLP is very important?
7. Natural Language Processing Applications
8. Areas That Leverage NLP Technology
9. Why is NLP so difficult?
10. What are the techniques used in NLP?
11. Libraries and tools
12. Arabic Natural Language Processing
13. The Future of NLP

Before we start !!!

What you should know before you start

1- Understanding of some of the key concepts in **natural language processing** and **machine learning algorithms**

2- Basic Knowledge in **Python**

3- Some experience using the **NumPy, pandas and scikit-learn** libraries.

Environment setup

Installing Python(IDE like anaconda) and **Jupyter Notebooks**

OR

Google Colab

colab

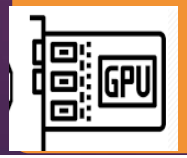
Colab is a free cloud service based on Jupyter Notebooks for machine learning education and research. It provides a runtime fully configured for deep learning and **free-of-charge access to a robust GPU**.

01 **Free-Of-Charge**

02 **Cloud Service**

03 **Jupyter Notebook Environment**

04 **Zero Configuration Required**



05 **Access to GPU /TPU**

Python Natural Language Processing (NLP) libraries

Natural Language Toolkit (NLTK)	TextBlob	CoreNLP	Gensim
spaCy	polyglot	scikit–learn	Pattern

Libraries and tools

- **NLTK**
  - Small but useful datasets with markup
  - Preprocessing tools: tokenization, normalization…
  - Pre-trained models for POS-tagging, parsing…
- **Stanford parser**
- **spaCy:**
  - python and cython library for NLP
- **Gensim**
  - python library for text analysis, e.g. for word
  - embeddings and topic modeling
- **MALLET**
  - Java-based library, e.g. for classification,
  - sequence tagging, and topic modeling
- •

Ahmad Shhadeh

Python Natural Language Processing (NLP) libraries

1) **Natural Language Toolkit (NLTK)**
2) TextBlob
3) CoreNLP
4) Gensim
5) spaCy
6) polyglot
7) scikit–learn
8) Pattern

Natural Language Toolkit(**NLTK)**

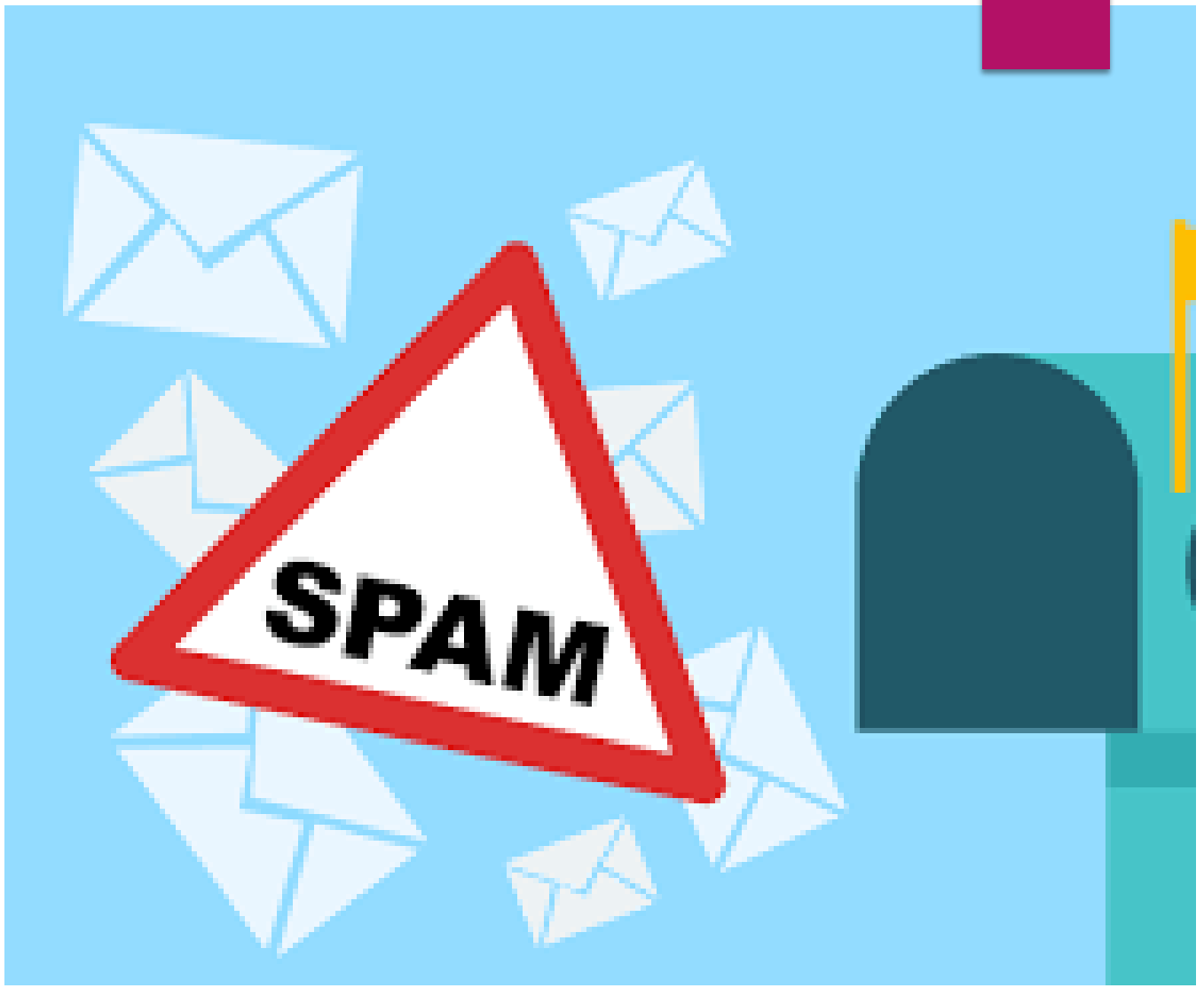
- **The natural language toolkit** is the most utilized package for handling natural language processing tasks in Python. Usually called NLTK for short.
- NLTK is a **leading platform for building Python programs** to work with human language data. It provides easy-to-use interfaces to over 100 corpora and lexical resources such as WordNet.
- Providing a suite of text processing libraries for **classification, tokenization, stemming, tagging, parsing, and semantic reasoning,** wrappers for industrial-strength NLP libraries
- NLTK has been called "**a wonderful tool for teaching, and working in, computational linguistics using Python**," and "**an amazing library to play with natural language**."
- It is a suite of open-source tools originally created in **2001 at the University of Pennsylvan**ia for the purpose of making building NLP processes in Python easier.
- This package has been **expanded** through the extensive **contributions of open-source users** in the years since its original development.

NLP Tool Kit installation

- How to install NLTK on your local machine:
  - We assumed that python and anaconda both are installed
  - `pip install nltk` ➔installing nltk
  - `dir(nltk)` ➔ check all installed packages under NLTK


LET'S
START!

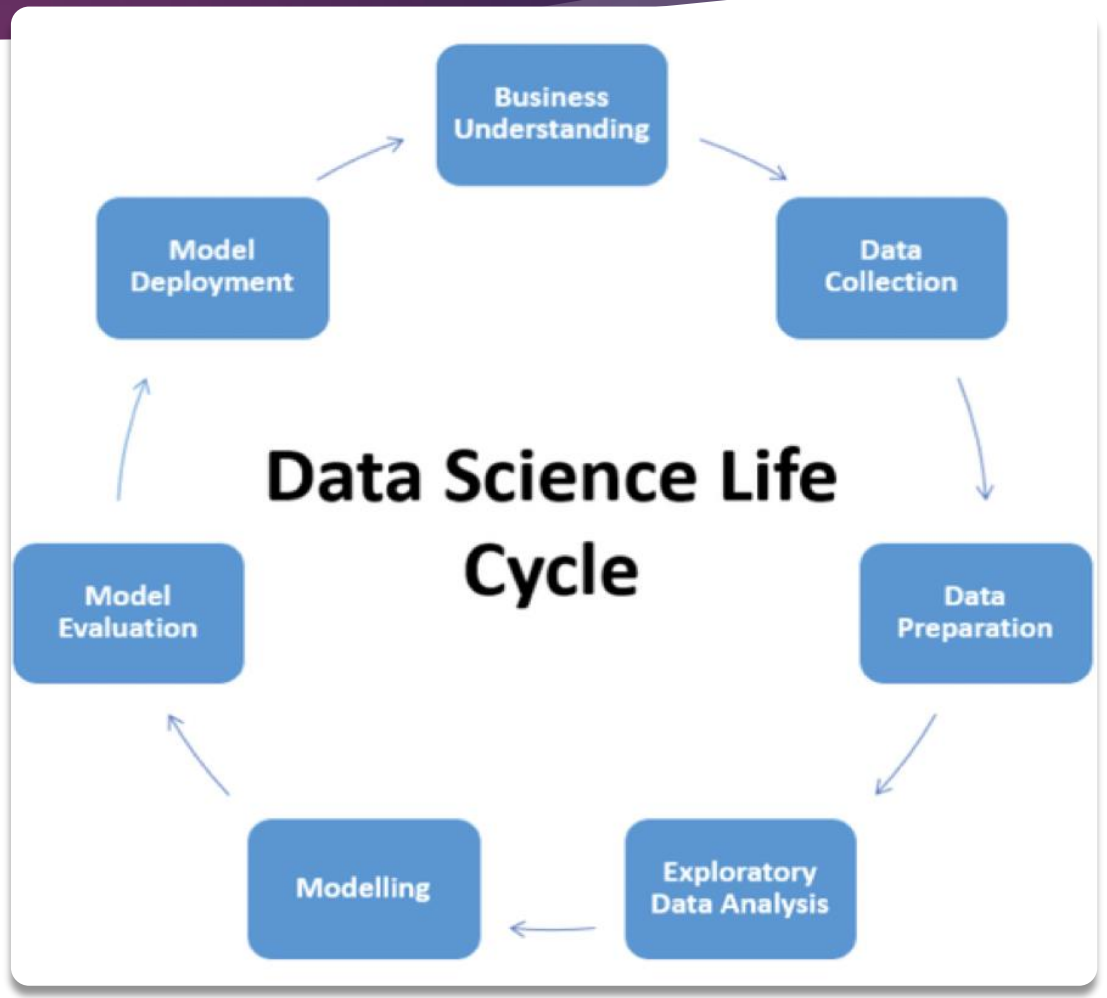
Email/SMS Spam Filtering Use case



Ahmad Shhadeh

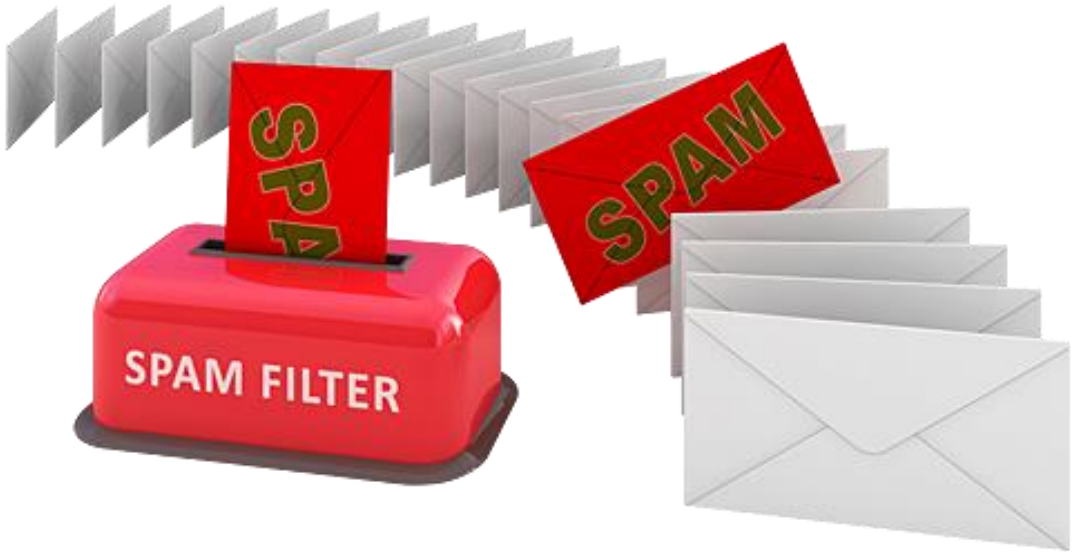
Data Science life cycle

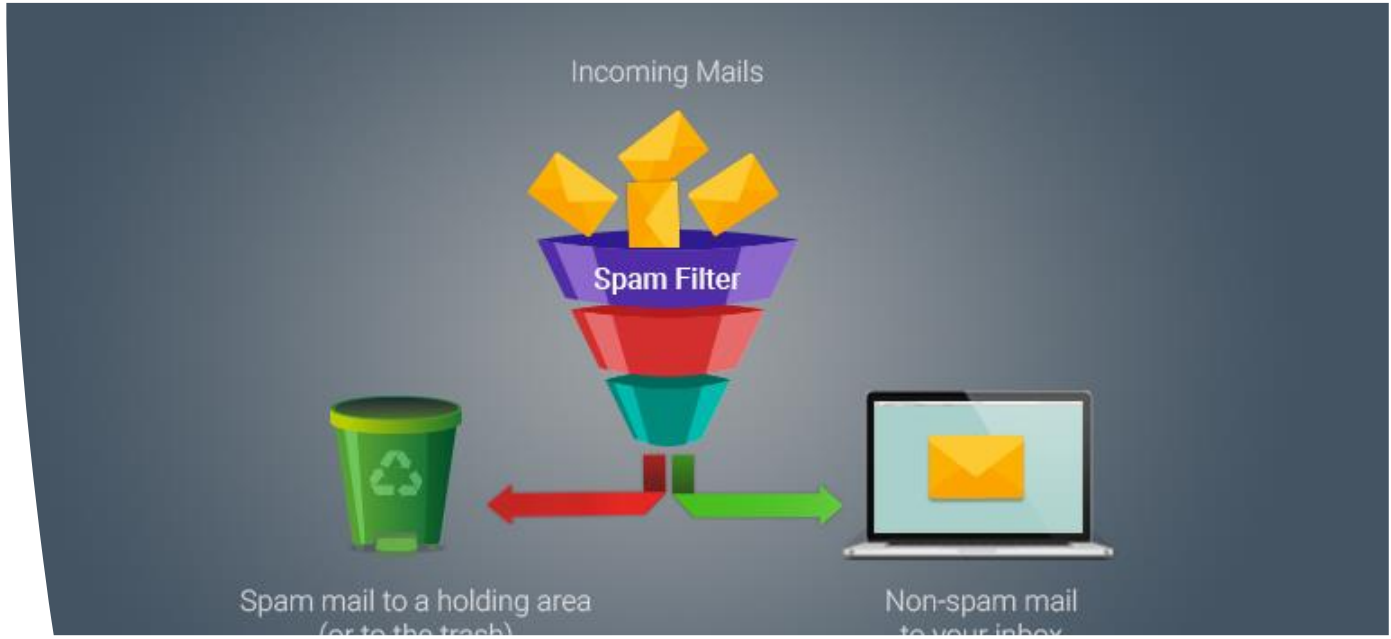
1. **Business Understanding**
2. Data Collection
3. Data Preparation
4. Exploratory data analytics(EDA)
5. Data Modelling
6. Model Evaluation
7. Model Deployment



Email/SMS Spam Filtering

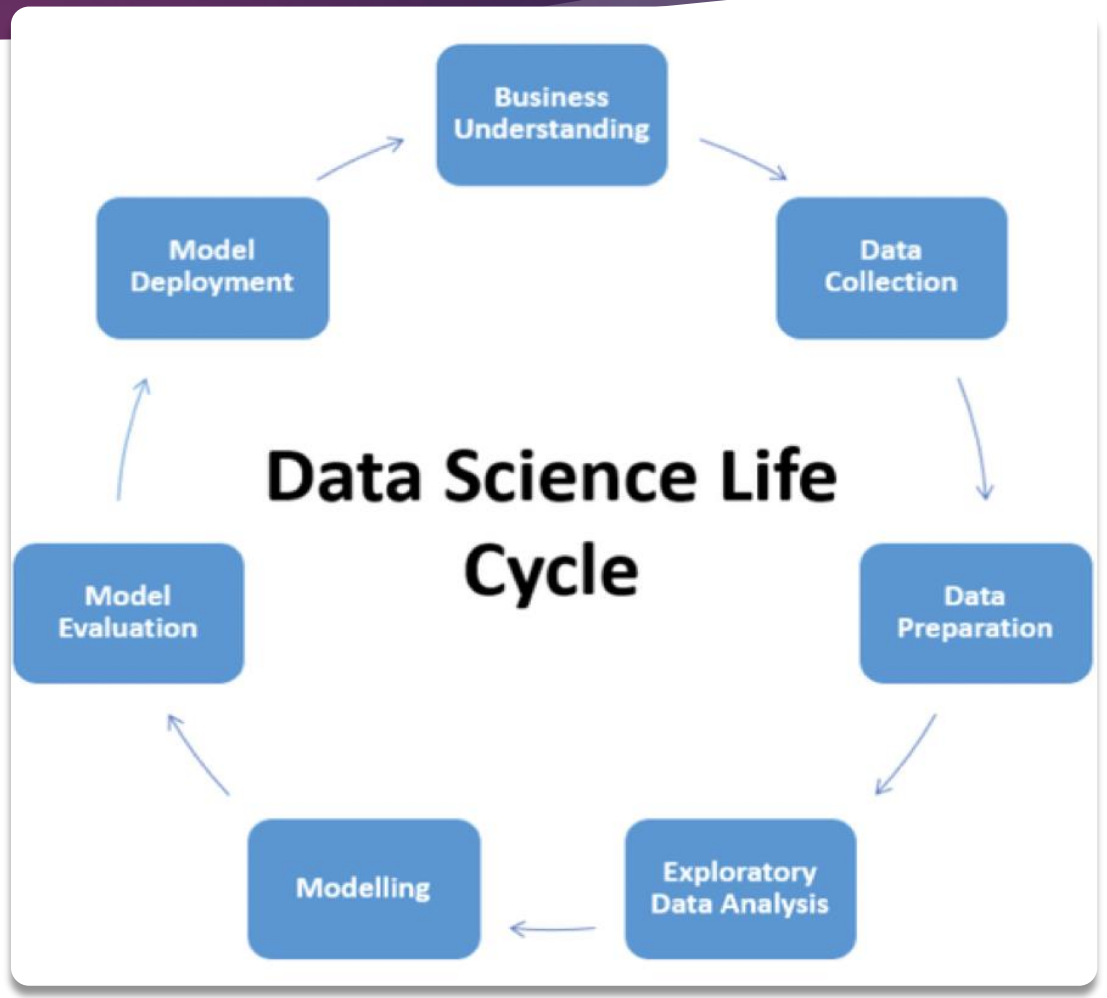
- Spam filtering is a beginner's example of document classification task which involves classifying an email as spam or non-spam (a.k.a. ham) mail.
- Spam box in your Gmail account is the best example of this.
- So lets get started in building a spam filter on a publicly available mail corpus.





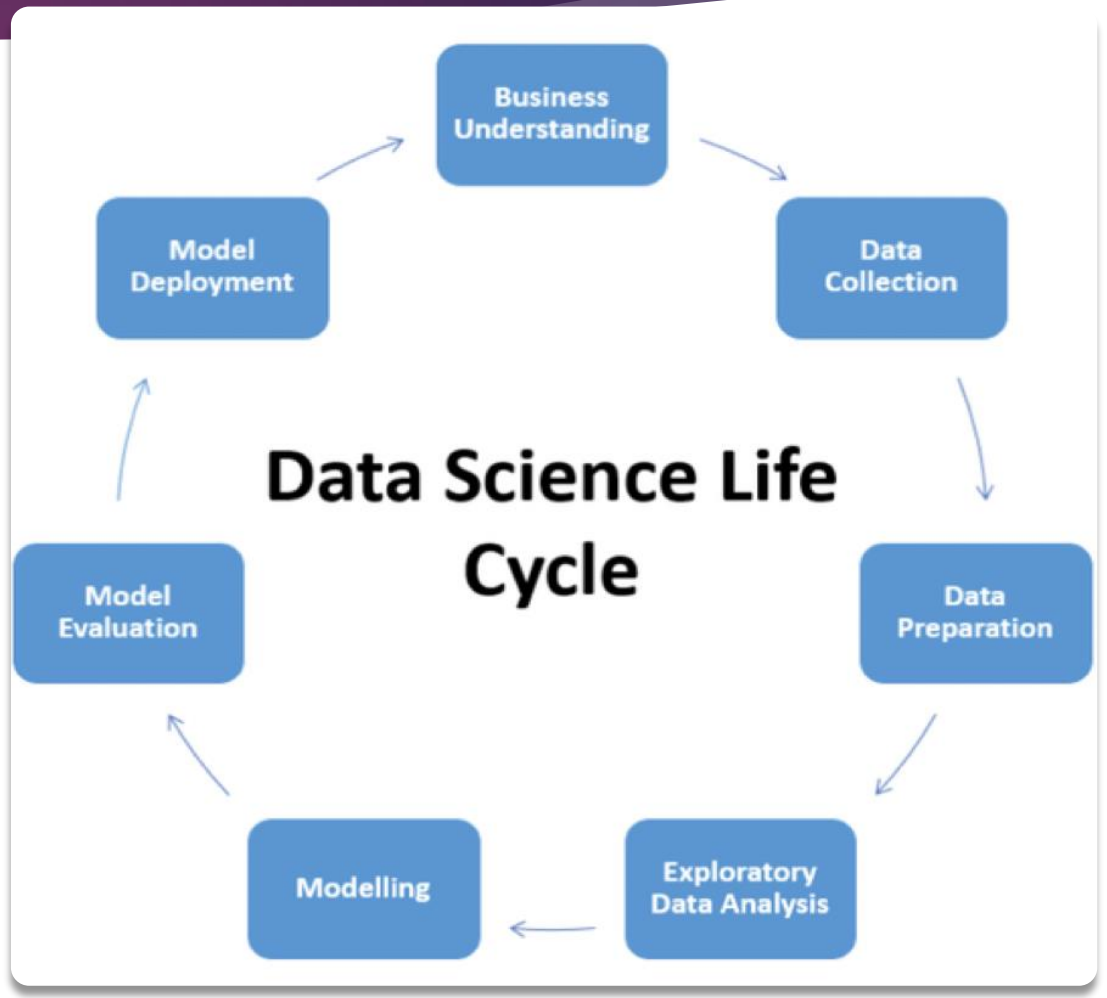
Data Science life cycle

- Business Understanding
- **Data Collection**
- Data Preparation
- Exploratory data analytics(EDA)
- Data Modelling
- Model Evaluation
- Model Deployment



Data Science life cycle

- Business Understanding
- **Data Collection**
- Data Preparation
- Exploratory data analytics(EDA)
- Data Modelling
- Model Evaluation
- Model Deployment



Public and Private Data

- **Private data**: it is **private and belongs to an organization**, and there are certain **security and privacy** concerns attached to it. It is used for the companies' **internal analysis** purposes in order to gain business and growth insights. Some examples of such organizational private data are telecom data, retail data, and banking and medical data.
- **Public data:** This is the data that is **available for public** use and is offered by many sites such as government websites and public agencies for the purpose of research. Accessing this data does not require any special permission or approval.
- **Open data** is the idea that some data should be freely available to everyone to use and republish as they wish, **without restrictions from copyright**,the goals of the open-source data movement are similar to those of other "open(-source)" movements such as **open-source software**

Data Collection

- **SMS Spam Collection Data Set**
  Download: Data Folder, Data Set Description
- Abstract: The SMS Spam Collection is a public set of SMS labeled messages that have been collected for mobile phone spam research.

Data Set Characteristics:	Multivariate, Text, Domain-Theory	**Number of Instances:**	**5574**	**Area:**	Computer
Attribute Characteristics:	Real	**Number of Attributes:**	N/A	**Date Donated**	2012-06-22
Associated Tasks: Ahmad Shhadeh	Classification, Clustering	**Missing Values?**	N/A	**Number of Web Hits:**	331230

NEXT LESSON

Reading and exploring the data