Strengthening University-Industry Partnerships

Data Science for All: An Open-Source Approach

UIDPVirtual 2020

Andre de Waal, IBM

March 26, 2020

Disclaimer

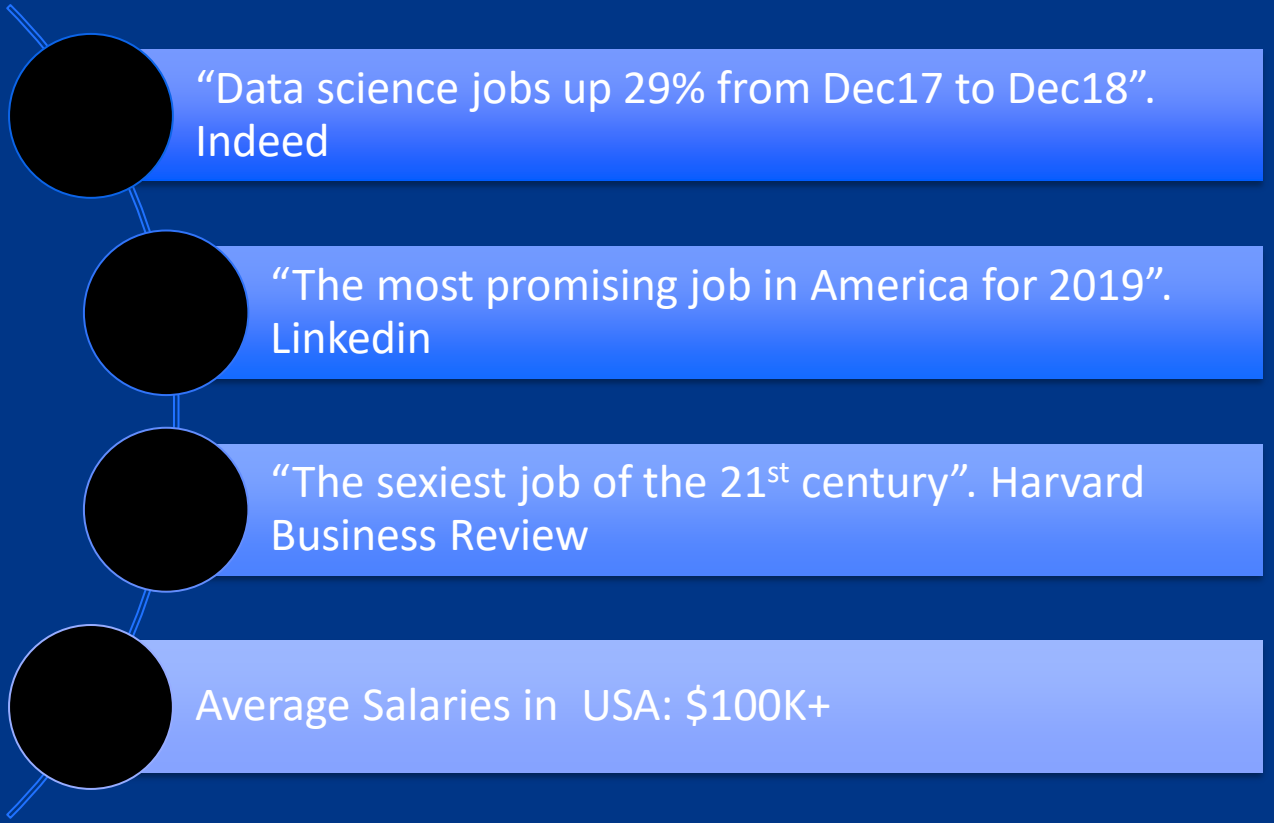
UIDP materials, which include publications, webinars, videos, and presentations, reflect an amalgamation of the experiences and knowledge of those who participate in UIDP activities. The views and opinions expressed in UIDP materials do not necessarily reflect the official policy or position of any individual organization or the UIDP. At no time should any UIDP materials be used as a replacement for an individual organization's policy, procedures, or legal counsel. UIDP is not a lobbying organization and UIDP materials are not intended to be used to influence government decisions.

OpenDS4All

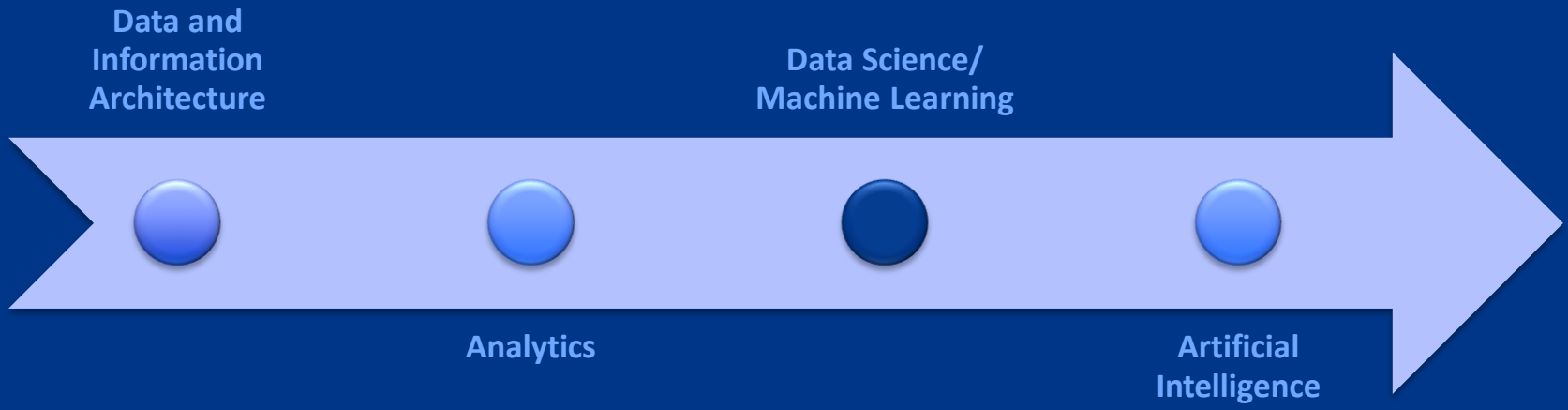
An open source approach to accelerate the creation of data science curricula at academic institutions

Andre de Waal

Data Science And Machine Learning LinkedIn's Fastest-Growing Jobs



On the road to AI, Data Science skills are essential

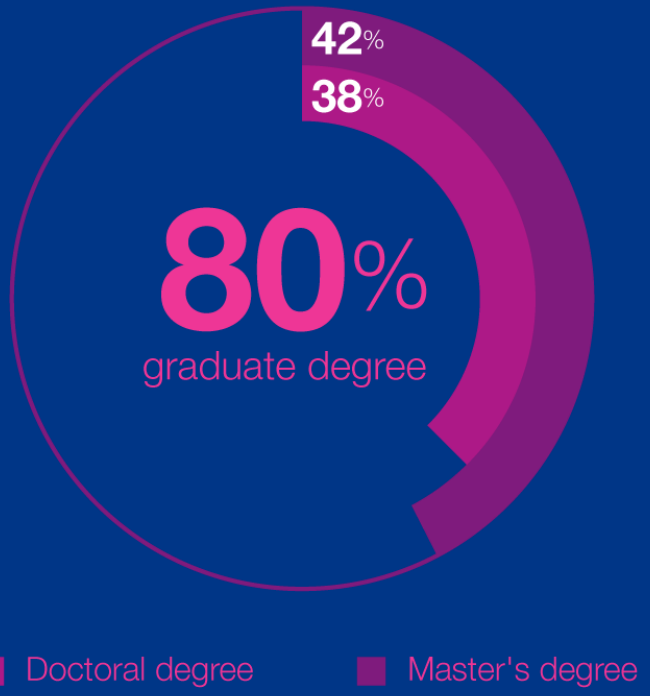


Multiple organizations are building F2F Data Science Education Programs

K-12	Community Colleges	Universities	Services Organizations	Corporate Training

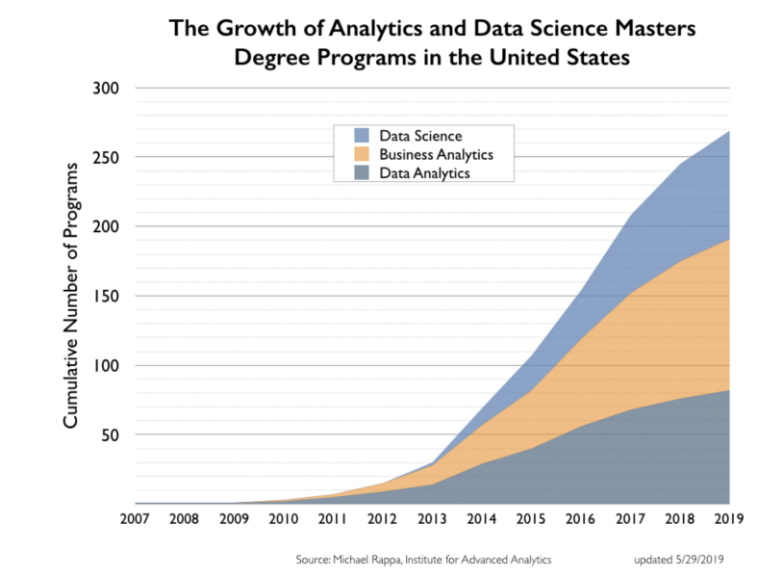
In addition to digital offerings

80% of Data Scientists have a Graduate degree



Source: *Mastering the art of data science*. IBM Institute for Business Value. © 2016 IBM. ibm.biz/artofdata

Only 250+ Master Programs in the US. Small # undergraduate programs



Growth in demand for Data Science skills currently outpaces the ability of academic institutions around the world to build data science programs

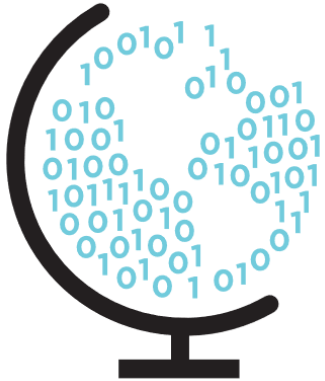
OpenDS4All - An Open Source Approach to Data Science Education

- A Data Science Curriculum Kit built on Open Source (Python)
- Built by professors at the University of Pennsylvania
- Launched as an Open Source Project under the Linux Foundation Governance to ensure updates and growth
- Live in February 2020

GOAL

Enable Organizations to build their own Data Science Educational Programs

https://github.com/odpi/OpenDS4All



ODPi

OpenDS4All

GitH

Description

OpenDS4All is a project created to accelerate the creation of data science curricula at academic institutions. While a great deal of online material is available for data science, **including** online courses, we recognize that the **best** way for many students to learn (and for many institutions to deliver) content is through a combination of lectures, recitation or flipped classroom activities, and hands-on assignments.

OpenDS4All attempts to fill this important niche. Our goal is to provide recommendations, slide sets, sample Jupyter

Audience (Instructor and Student)

The initial modules were designed to target a broad, cross-university audience at both the undergraduate and graduate levels. Modules contain instructor notes and comments intended to aid in the delivery of the material; the expectation is that instructors will be generally fluent in basic database and machine learning concepts.

The perspective of the materials largely comes from computer science, with an emphasis on data wrangling and engineering as well as machine learning and validation. However, prior versions of the content have been used to teach students ranging from freshmen to PhD students, across a wide range of fields. The emphasis is largely on core concepts and algorithms with grounding in today's technologies and best practices.

Students are expected to come in with two major prerequisites:

- Comfort and familiarity with programming in Python (writing small functions, importing and calling library functions, using Python data structures).
- Familiarity with probability theory and very basic statistical notions.

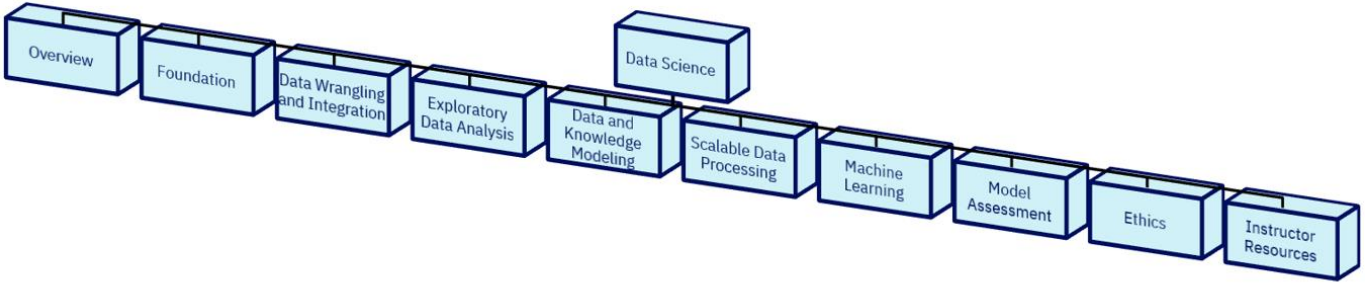
To some extent, students with a limited background *can* follow along with this material, but they will likely need to supplement extensively.

How to use

The following topology shows how content is **currently** organized around categories. This is a living/dynamic taxonomy that

How to use

The following topology shows how content is **currently** organized around categories. This is a living/dynamic taxonomy that is updated as new content is added to the project.



Each category contains modules and each module consists of one or more of the following components:

- instructor notes (`Instructor_Notes.md`) and guide to files
- a set of PowerPoint slides (with presenter notes) ending in `.pptx`
- companion Jupyter notebooks, for students to see the lecture materials "in context" and to be able to experiment
- sample quiz materials (where applicable)
- sample homework assignments (where applicable)
- additional documentation (where applicable)

There are many ways to interact with this repository:

- browse the repository in search of content (use the '**Find file**' search functionality)
- download content (PowerPoint slides, Jupiter notebooks, etc.)
- contribute content (become a contributor to the project)
- become involved in the day-to-day management of the project (become a committer)
- provide overall direction and leadership to the project (become a Technical Steering Committee member)

The project's governance principles clarifies the different roles and describes the processes for becoming a contributor, a committer or a TSC member.

Contributing

Anyone can contribute to this repository - learn more at CONTRIBUTING.md. Follow the step-by-step instructions COMMUNITY-GUIDE.md to submit a module for possible inclusion into to repository.

Governance

OpenDS4All is a project hosted by ODPi. This project has established it's own processes for managing day-to-day processes in the project at GOVERNANCE.md.

Reporting Issues

Contribution Guidelines

Educational Modules

The building blocks of this repository are **modules**. Each module covers one or more lessons that can be taught at undergraduate or graduate level (at any higher educational institution).

Modules should be:

- mostly independent of other modules
- cover a limited number of topics
- the coverage of a topic should be substantial and thorough if it is not an introductory or an overview module

The components of a module are:

- a set of PowerPoint slides (with presenter notes)
- a Jupyter notebook
- a quiz
- a homework assignment
- instructor notes
- additional documentation (where applicable)

The **minimum requirement** for a module to be considered for inclusion in this repository is that it contains:

- a set of PowerPoint slides (with presenter notes)
  - 30 or more slides are recommended
  - there must be enough substance in the slide deck to cover at least a 50-minute lecture
- a Jupyter notebook (illustrating how material covered in the slides are applied to one or more data sets)
  - use public data sets that are available for download or accessible through a hyperlink
  - do not assume dependent packages are pre-installed in the user's Jupyter environment
  - import all modules needed to run the code cells successfully
  - keep the markdown cells as simple as possible
    NB! The Jupyter notebook my be omitted in special cases, such as in Foundational modules where no accompanying data sets exist. But, this should be the exception rather than the rule.
- a short summary of the module with a set of learning outcomes (in a text or a markdown file)
  - 300 or less words are recommended (for the summary)
  - use active verbs when formulating outcomes
  - make sure the the outcomes are measurable
  - examples of learning outcomes are
    - understand sampling, probability theory, and probability distributions
    - implement descriptive and inferential statistics using Python
    - demonstrate ability to visualize data and extract insight

Read the specifications in the NAMING-CONVENTIONS.md file to learn home to name your modules to facilitate search.

Naming Conventions for Modules

Modules are identified by a descriptive name, keywords, skill level and file extension

A module name (file name) consists of four parts:

- **XXX-Yyy-zzz.abc**

where

- **XXX** represents the descriptive name of the module, e.g.
  - INTRODUCTION
  - WRANGLING-DATA
  - UNSUPERVISED-MACHINE-LEARNING
    (multiple words in the descriptive name are separated by dashes '-', with a maximum of 3 words allowed)
- **Yyy** represents keywords, e.g.
  - Access-Integration
  - Big-Data-Analytics
  - Convolutional-Neural-Networks
    (multiple keywords are separated by dashes '-', with a with maximum of 3 keywords allowed)
- **zzz** represents the skill level

- **zzz** represents the skill level
  - basic
  - intermediate
  - advanced
    (the skill level can only have one value)
- **abc** represents the file extension, e.g.
  - txt
  - pptx
  - md
  - ipynb

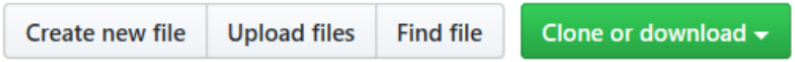
Examples

- INTRODUCTION-Big-Data-Analytics-basic.pptx
- INTRODUCTION-Big-Data-Analytics-basic.md
- WRANGLING-DATA-Access-Integration-intermediate.pptx
- WRANGLING-DATA-II-Access-Integration-intermediate.pptx
- SUPERVISED-MACHINE-LEARNING-Neural-Networks-advanced.pptx
- SUPERVISED-MACHINE-LEARNING-Neural-Networks-advanced.txt
- SUPERVISED-MACHINE-LEARNING-Neural-Networks-intermediate.ipynb

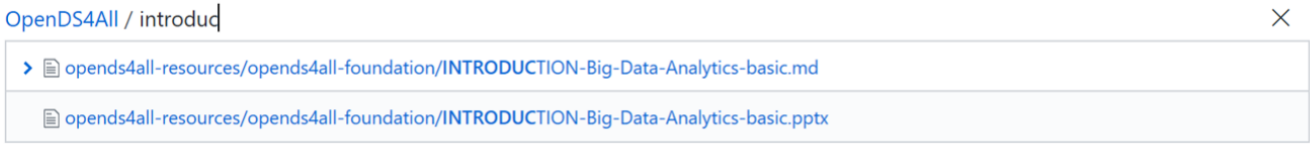
The benefit of this naming convention is that modules are searchable on name, keywords, skill level and file extension.

Searching for Content

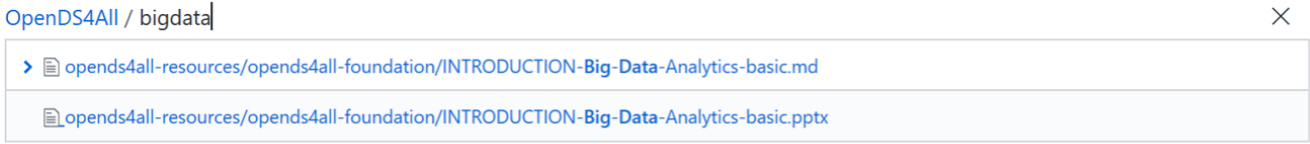
The file name is searchable with the standard **Find file** search functionality already in GitHub



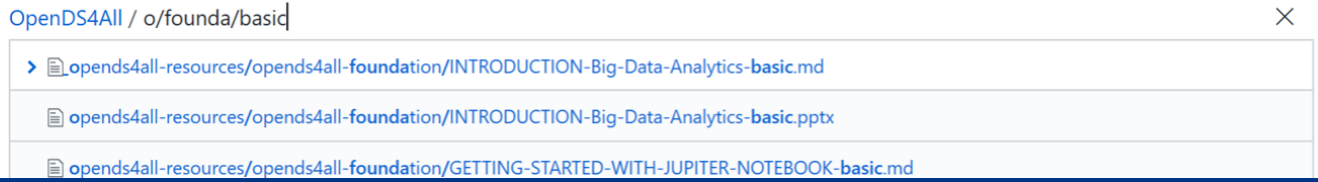
- Search by descriptive name, e.g. find INTRODUCTION modules



- Search by topic, e.g. find Big Data modules



- Search by skill level inside a category, e.g. basic modules



Instructor Notes

Readings, Texts, and References

For the overall course, we recommend the following books as potentially being useful:

- Data Science from Scratch: First Principles with Python, 2nd ed, by Joel Grus, published by O'Reilly.
- Python for Data Analysis: Data Wrangling with Pandas, NumPy, and IPython, 2nd ed, by Wes McKinney, published by O'Reilly.

Additionally, we recommend Towards Data Science as a useful resource for this space.

Courses Using OpenDS4All Materials

- The University of Pennsylvania's CIS 545, Big Data Analytics, www.cis.upenn.edu/~cis545

Background Material

Students may find the following resources to be useful as background:

- Google's Python class (free): https://developers.google.com/edu/python

Suggested Configuration of Modules

The OpenDS4All modules can be "mixed and matched" at the discretion of the instructor, according to preferences, time constraints, and the target audience. However, certain elements do have dependencies. We suggest a "core" outline as follows:

1. Overview, 1.5 lecture hours (basic)
    - *Optional recitation*: review of Python basics, including data structures
2. Acquiring, wrangling, integrating, and cleaning data, 3-4 lecture hours (basic-intermediate)
    - *Optional recitation*: basics of HTML and the Document Object Model
    - *Optional recitation*: basics of regular expressions (often used for pattern matching) and XPath (which builds on some ideas from regular expressions and traverses XML trees)
3. Modeling data: types, graphs, schemas, 2-4 lecture hours
    - *Optional recitation*: encoding tree- or graph-structured data in relations, and traversing the data
4. Performance:
    - Foundations: Computer architecture basics, 1 hour (basic, provides an overview of CPU and memory)
    - Efficient data processing, 3-7 lecture hours (intermediate, appropriate for a more computational and big data audience)

    - Foundations: Computer architecture basics, 1 hour (basic, provides an overview of CPU and memory)
    - Efficient data processing, 3-7 lecture hours (intermediate, appropriate for a more computational and big data audience)
    - *Optional recitation*: Use `merge` and `merge_map` algorithms from Lecture Notebook to study performance of alternative strategies. Use `%%time` and SQLite to study performance of database indices.

5. Building machine learning models
    - **Overview and Unsupervised Models**, 1 lecture hour, basic.
    - **Supervised Models, Decision Trees, Random Forests**, 1-1.5 lecture hours, basic.
    - **Linear and Logistic Regression**, 1-1.5 lecture hours, basic.
    - **Neural Networks**, builds upon linear and logistic regression, 2-4 lecture hours, intermediate [requires understanding of calculus].
6. Validating and tuning models, 1.5-3 hours, basic

Additional and advanced topics:

- Data ethics, 1-2 hours, basic, most appropriately covered after a discussion of machine learning models.
- Data exploration and visualization, 1-2 hours, basic.
- Big data and the cloud, 3-5 hours, intermediate. Most appropriate after a discussion of performance.

Branch: master | New pull request | Create new file | Upload files | Find file | Clone or download

zackives Merge pull request #62 from odpi/Frenchhorn005-patch-3 ... Latest commit 86c36b5 7 days ago

assets	Add files via upload	13 days ago
opends4all-resources	Merge pull request #59 from odpi/penn-processing-zgi	15 days ago
tsc/meetings	Update 20200306.md	11 days ago
CODE_OF_CONDUCT.md	Updates for spellcheck and MarkDown lint	21 days ago
COMMITTERS.csv	Update COMMITTERS.csv	21 days ago
COMMUNITY-GUIDE.md	Updates for spellcheck and MarkDown lint	21 days ago
CONTENT-ORGANIZATION.md	Updates for spellcheck and MarkDown lint	21 days ago
CONTRIBUTING.md	Update CONTRIBUTING.md	21 days ago
GOVERNANCE.md	Updates for spellcheck and MarkDown lint	21 days ago
Instructor_Notes.md	Update Instructor_Notes.md	13 days ago
LICENSE	Initial commit	7 months ago
NAMING-CONVENTIONS.md	Updates for spellcheck and MarkDown lint	21 days ago
README.md	Updates for spellcheck and MarkDown lint	21 days ago

Branch: master ▾ OpenDS4All / opends4all-resources / Create new file Upload files Find file History

SusanBDavidson Merge pull request #59 from odpi/penn-processing-zgi ... Latest commit 50c9753 15 days ago

..		
opends4all-data-and-knowledge-...	Update Instructor_Notes.md	20 days ago
opends4all-data-wrangling-and-int...	Update Instructor_Notes.md	20 days ago
opends4all-ethics	added notes	21 days ago
opends4all-exploratory-data-analysis	More detail on presentation, added preview for ML, removed comments	16 days ago
opends4all-foundation	changed README	21 days ago
opends4all-instructor-resources	Initial commit of basic materials for Module 0 / Overview	21 days ago
opends4all-machine-learning	One more link fix.	21 days ago
opends4all-model-assessment	Model assessment and robustness	21 days ago
opends4all-overview	test	21 days ago
opends4all-scalable-data-processing	Merge	21 days ago
README.md	Fixed link to instructor's guide	21 days ago

README.md

Branch: master ▾ OpenDS4All / opends4all-resources / opends4all-overview / Create new file | Upload files | Find file | History

SusanBDavidson and zackives test ... Latest commit 82bec3a 25 days ago

..		
INTRODUCTION-Data-Science-basic.pptx	Susan's changes based on Ana's feedback	last month
Instructor_Notes.md	Updates for spellcheck and MarkDown lint	last month
Quiz.md	Fixed line breaks.	28 days ago
README.md	test	25 days ago

README.md

Overview: What is Data Science?

This module provides a broad overview of modern data science, data analytics / data engineering, and big data. It sets the context for the OpenDS4All curriculum.

Directory Contents

Motivation for the Course: Data Is Driving Everything

1. Modern data acquisition is inexpensive!
   - Smartphones, embedded systems, inexpensive sensors,
   - Medical devices, simulators, ...
2. Data storage is inexpensive!
3. Parallel (compute cluster) computation is inexpensive
   - The Cloud, clusters of computers, GPUs, tensor processors, ...

Can we use **algorithms + data** to understand phenomena? Build or augment **models**? Build **detectors**? Make **diagnoses**?

There is data everywhere. It is constantly being collected (e.g. through smartphones, IoT sensors, Alexa) and contributed (e.g. through Facebook and Twitter). Storage is cheap, and compute power readily available and rapidly increasing. Computer clusters are available to enable parallelism, and special architectures developed for ML operations. All this is available not just for those who can afford to own them but through the commodity of the cloud..

This increased ability to collect, store and process data has led to a revolution in how we do things in almost every field, from biological sciences to social sciences, humanities, political science, policy, marketing and so on. But just because we have the data doesn't imply we understand it. It must be processed to derive information, make models, enable predictions, and make diagnoses. We want to use the data to develop (or augment) models using machine learning and other algorithms.

That is the essence of this course.

Branch: master ▾ | Create new file | Upload files | Find file | History

OpenDS4All / opends4all-resources / opends4all-exploratory-data-analysis /

zackives More detail on presentation, added preview for ML, removed comments — ✓ Latest commit 7803f46 16 days ago

..		
README.md	Exploratory data analysis - visual analytics	21 days ago
VISUAL-ANALYTICS-visualization.ip...	added crawled dataset for module 1 wikipedia etc	20 days ago
VISUAL-ANALYTICS-visualization.p...	More detail on presentation, added preview for ML, removed comments	16 days ago

README.md

Exploratory Data Analysis

The topic of exploratory data analysis is extremely important in discovering and formulating hypotheses. A key aspect of that process is *visual analytics* or *information visualization*, which allows one to see certain aspects of the data holistically. This module is focused on such topics.

Directory Contents

1 lines (1 sloc) | 200 KB

Raw Blame History

Lecture 1: Wrangling

In this lecture notebook, we will test the hypothesis that the CEOs of major companies are typically in their 40s or older, i.e. that they were born after 1980. To do this, we will identify top companies, find their CEOs, extract their ages, and look at the distribution of ages.

In [1]:
```
# Let's start by installing some libraries that are useful for processing web data.

# For crawling pages
!pip3 install scrapy

# # Optional, for parallel execution
!pip3 install swifter

# # For string similarity
!pip3 install py_stringsimjoin

# # lxml to parse xml tree
!pip3 install lxml
```

```
Collecting scrapy
  Downloading https://files.pythonhosted.org/packages/3b/e4/69b87d7827abf03dea2ea984230d50f347b00a7a3897bc
93f6ec3dafa494/Scrapy-1.8.0-py2.py3-none-any.whl (238kB)
     |████████████████████████████████| 245kB 2.7MB/s
Collecting w3lib>=1.17.0
  Downloading https://files.pythonhosted.org/packages/6a/45/1ba17c50a0bb16bd950c9c2b92ec60d40c8ebda9f3371a
e4230c437120b6/w3lib-1.21.0-py2.py3-none-any.whl
Collecting protego>=0.1.15
```

In [0]:

```
# Here are some imports we'll use through the notebook, collected here for simplicity

# For parsing dates and being able to compare
import datetime

# For fetching remote data
import urllib

# Pandas dataframes and operations
import pandas as pd

# Numpy matrix and array operations
import numpy as np

# Sqlite is a simplistic database
import sqlite3

# Crawler for multiple web pages at once
import scrapy
from scrapy.crawler import CrawlerProcess

# Can use dataframe.swifter.apply() instead of dataframe.apply()
# to try to parallelize the computation!
import swifter

# Approximate string matching, see
import py_stringsimjoin as ssj
import py_stringmatching as sm

# Data visualization
import matplotlib

# import urllib and etree for download and parsing
```

1. Acquiring data about companies and CEOs.

We'll start by loading a (remote) CSV file containing information about companies into a dataframe (company_data_df), and then persist it to an SQLite database. We'll also read an HTML file to obtain information about the CEOs of companies, and put in another dataframe (company_ceos_df).

In [0]:
```
data = urllib.request.urlopen(\
       'https://gist.githubusercontent.com/jvilledieu/c3afe5bc21da28880a30/raw/a344034b82a11433ba6f149afa47e57567d4a18f/Companies.csv')

company_data_df = pd.read_csv(data)
```

In [4]:
```
# Persist it to an SQLite database, and read it back.

conn = sqlite3.connect('local.db')

company_data_df.to_sql("companies", conn, if_exists="replace", index=False)

pd.read_sql_query('select * from companies', conn)
```

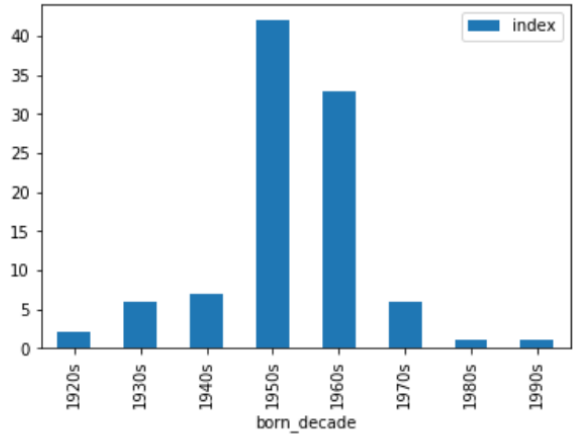
Out[4]:

	permalink	name	homepage_url	category_list	market	funding_total_us
0	/organization/waywire	#waywire	http://www.waywire.com	|Entertainment|Politics|Social Media|News|	News	1 750 000
1	/organization/tv-communications	&TV Communications	http://enjoyandtv.com	|Games|	Games	4 000 000
2	/organization/rock-your-paper	'Rock' Your Paper	http://www.rockyourpaper.org	|Publishing|Education|	Publishing	40 000

In [35]:
```
# Plot the result.

exec_df[['born_decade']].reset_index().groupby('born_decade').count().plot(kind='bar')
```

Out[35]: <matplotlib.axes._subplots.AxesSubplot at 0x7fa0045d84e0>



So are most CEO's born before 1980s?

7. An Exercise

Take this one step further, and link the companies table with another dataset to determine the market for each company.

Is there a correlation between the kind of company and the age of the CEO?

Branch: master ▾ OpenDS4All / opends4all-resources / opends4all-machine-learning / Create new file Upload files Find file History


zackives One more link fix. ✔ Latest commit 98a24f3 21 days ago

..

MACHINE-LEARNING-Homework.ipynb	Eliminated outputs	22 days ago
README.md	One more link fix.	21 days ago
SUPERVISED-ML-NN-neural-nets-intermediate.ipynb	Neural nets	22 days ago
SUPERVISED-ML-NN-neural-nets-intermediate.pptx	Neural nets	22 days ago
SUPERVISED-ML-OVERVIEW-trees-forests-basic.ipynb	Neural nets	22 days ago
SUPERVISED-ML-OVERVIEW-trees-forests-basic.pptx	Neural nets	22 days ago
SUPERVISED-ML-regression-basic.ipynb	Neural nets	22 days ago
SUPERVISED-ML-regression-basic.pptx	Neural nets	22 days ago
UNSUPERVISED-ML-matrices-pca-clustering-basic.ipynb	Neural nets	22 days ago
UNSUPERVISED-ML-matrices-pca-clustering-basic.pptx	Neural nets	22 days ago

README.md

OpenDS4All – Collaborating with the project – Help us Find…..

- Contributors
  - Faculty to contribute curricula
- Committers
  - Faculty to ensure curricula integrity

- Adopters
  - Organizations building Data Science academic programs

https://github.com/odpi/OpenDS4All

Interested in U-I

Partnerships?

Sign up for information about UIDP news, webinars, projects, and more at https://uidp.org/listserv-signup/.

Member Webinar

WEDNESDAY, APRIL 8, 2020

12 to 1 p.m. EDT

Jim Bray
Northwestern University
Moderator

UIDP Strengthening University-Industry Partnerships

How Companies Approach Academic Research Engagement in these Disruptive Times

Join us to learn how our industry members, in diverse sectors, are evaluating and reframing their current approaches to academic collaborations.

Panelists

Gaylene Anderson
Boehringer Ingelheim Pharmaceuticals, Inc

Kent Foster
Microsoft

Austin Kozman
PepsiCo