Présentation de la Data Science

Data Venture - Afterwork #1 - 06/03/2018

Sylvain Marchienne

kaggle

2017

The State of Data Science & Machine Learning

This year, for the first time, we conducted an industry-wide survey to establish a comprehensive view of the state of data science and machine learning. We received over **16,000 responses** and learned a ton about who is working with data, what's happening at the cutting edge of machine learning across industries, and how new data scientists can best break into the field. The below report shares some of our

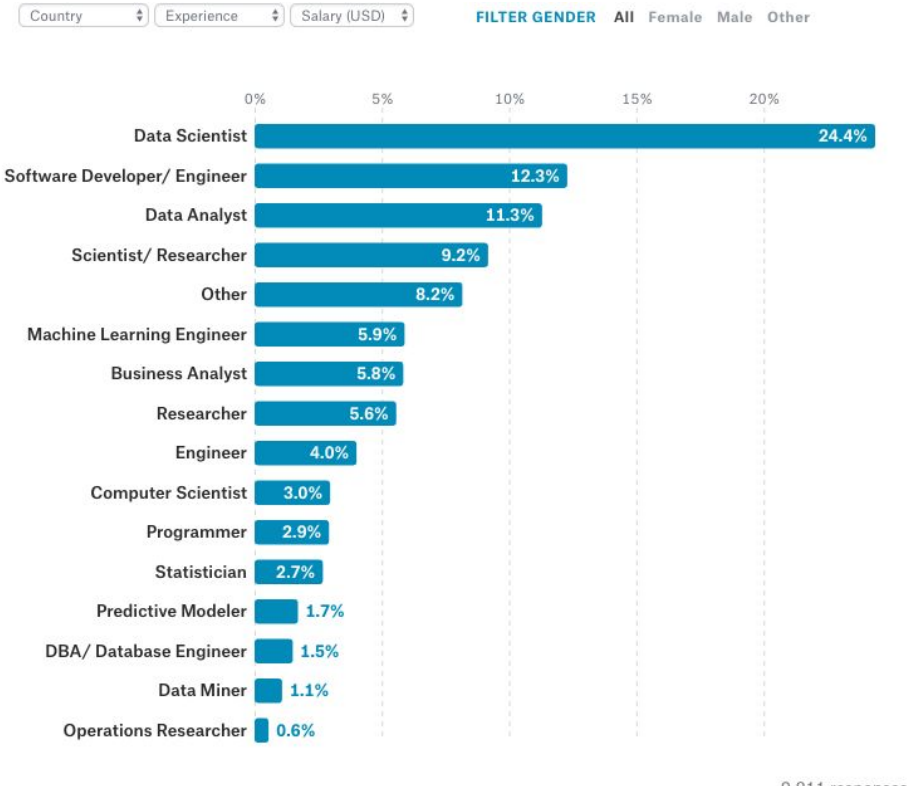
The State of Data Science 2017

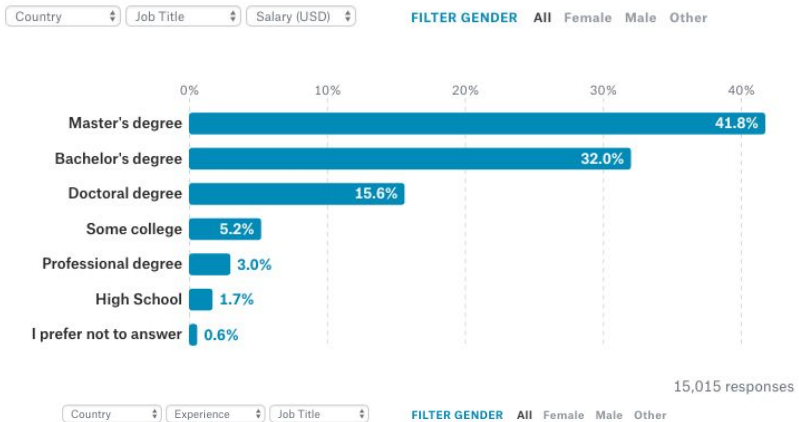
www.kaggle.com/surveys/2017

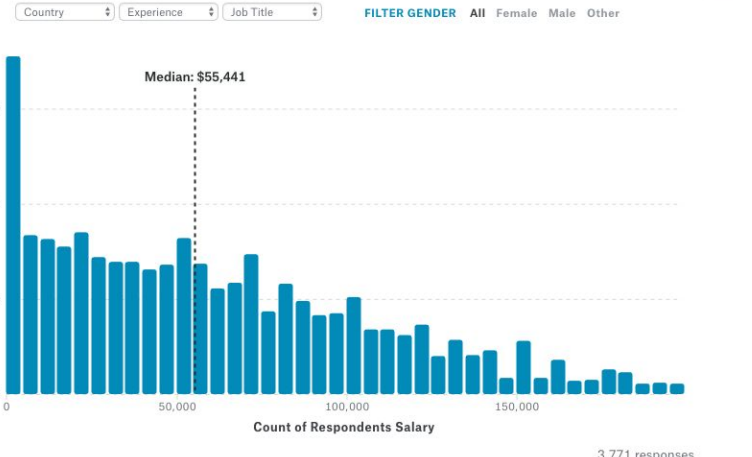
Qui travaille avec de la Data ?

Que font les Data Scientists ?

Qui travaille avec de la Data ? 1/2

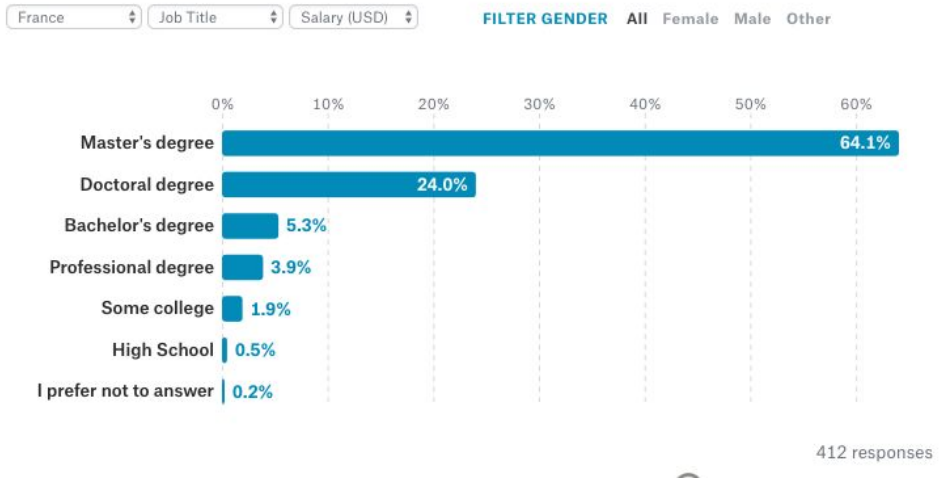






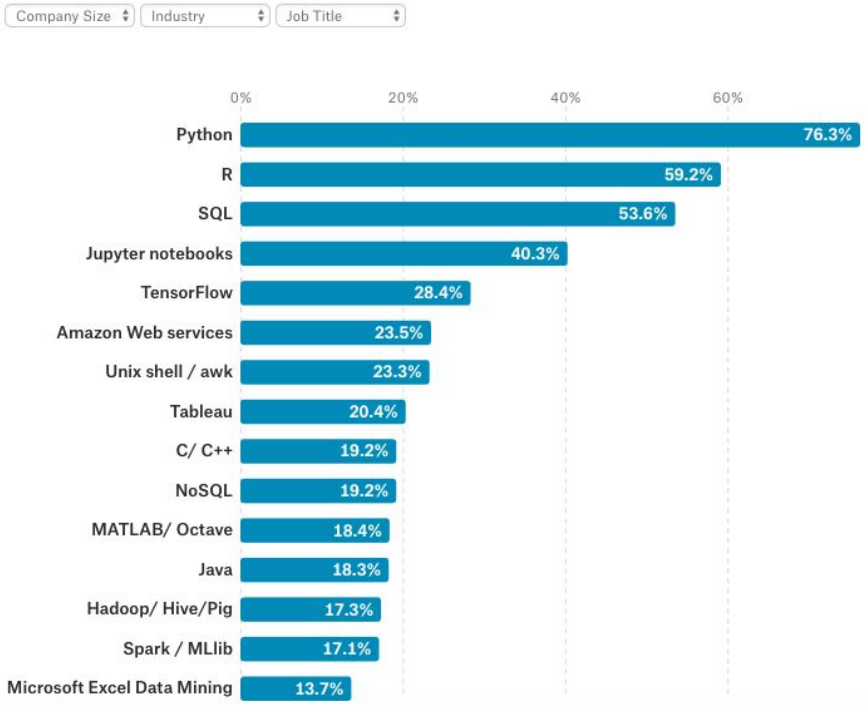
Qui travaille avec de la Data ? 2/2

Le cas de la France

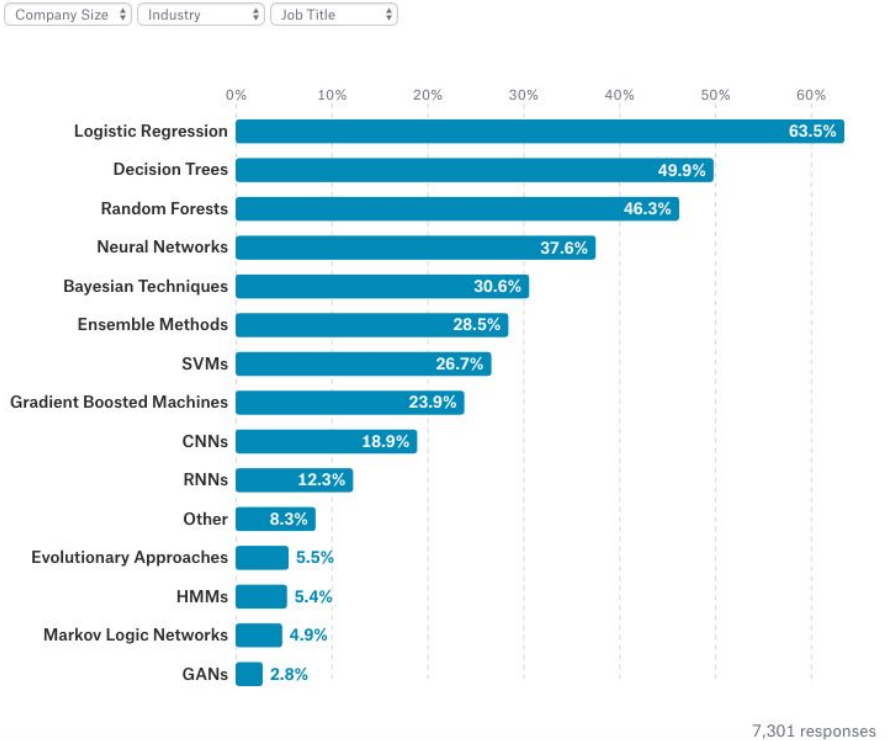


Que font les Data Scientists ? 1/2

Langages

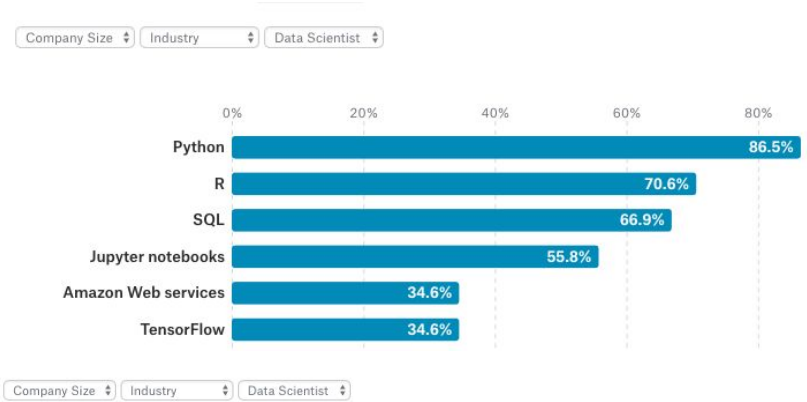


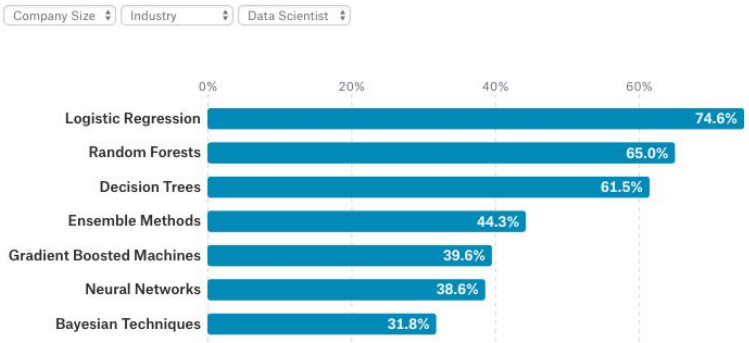
Algorithmes



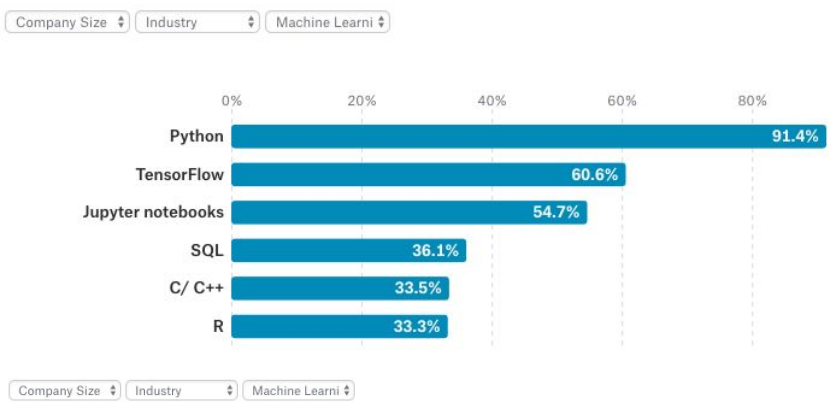
Que font les Data Scientists ? 2/2

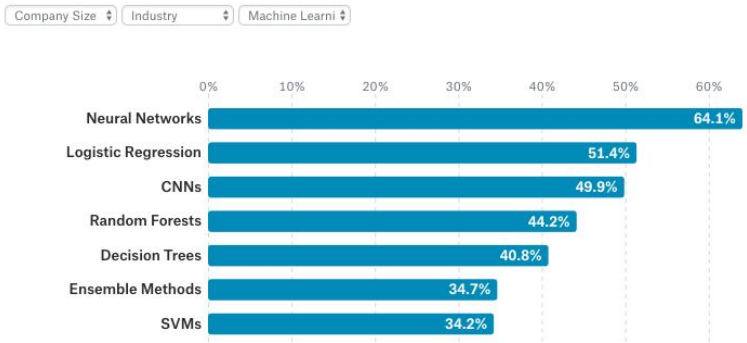
Data Scientist





Machine Learning Eng.





We define a data scientist as someone who “*(writes code to)* analyze data”.

Kaggle

Les outils du Data Scientist

De la manipulation à la prédiction en passant par la visualisation

Jupyter Notebook / Lab

Gephi

PyData

Scikit-Learn / TensorFlow

AWS

jupyter spectrogram (autosaved)

File Edit View Insert Cell Kernel Help Python 3

Markdown CellToolbar

Simple spectral analysis

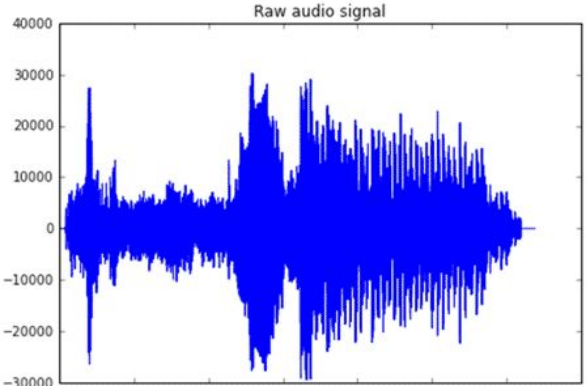
An illustration of the Discrete Fourier Transform

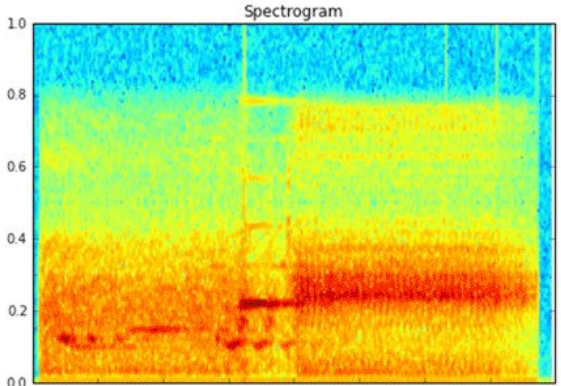
$$X_k = \sum_{n=0}^{N-1} x_n exp^{\frac{-2\pi i}{N} kn} \quad k = 0, \ldots, N-1$$

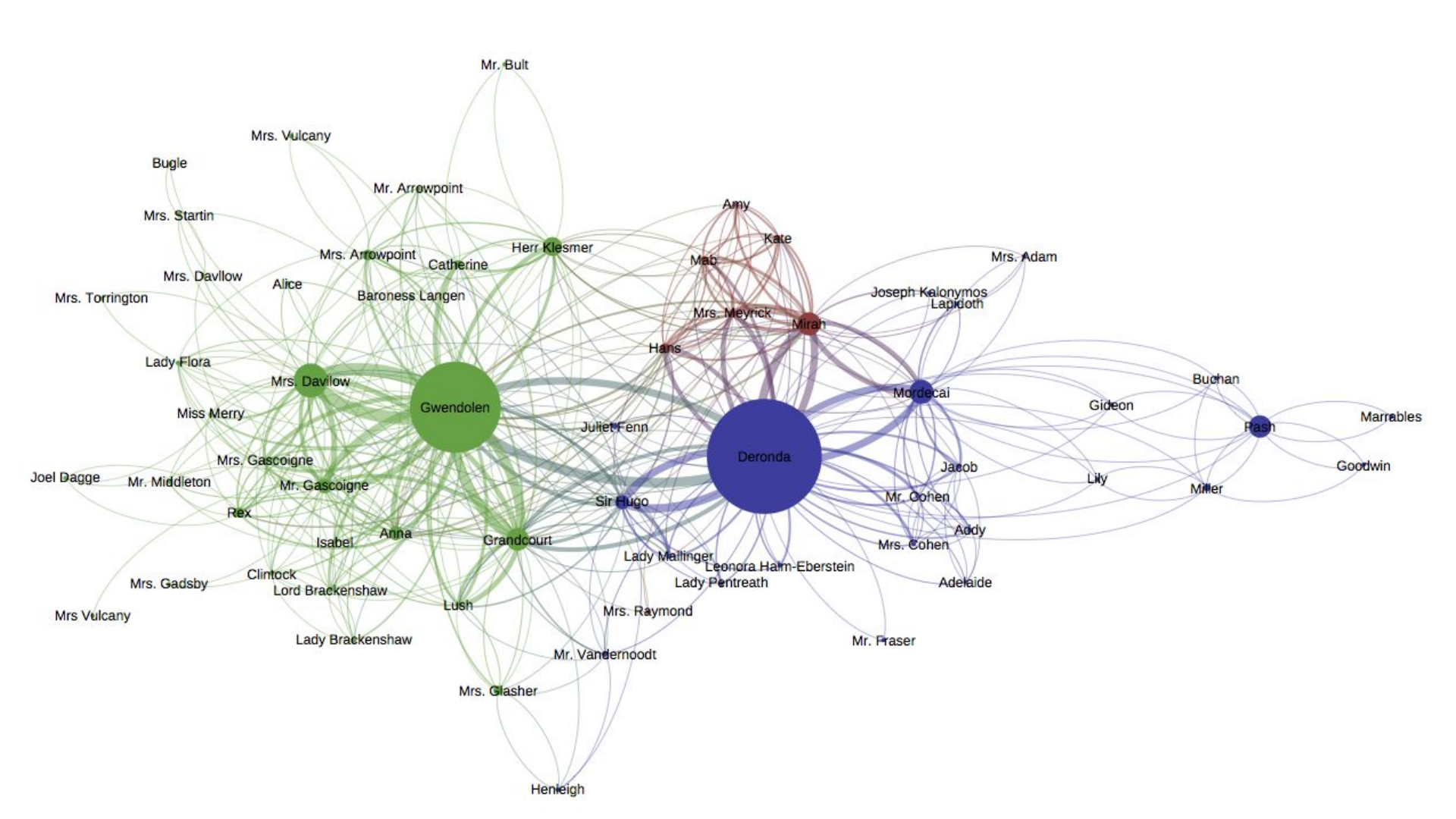
In [2]:
```
from scipy.io import wavfile
rate, x = wavfile.read('test_mono.wav')
```

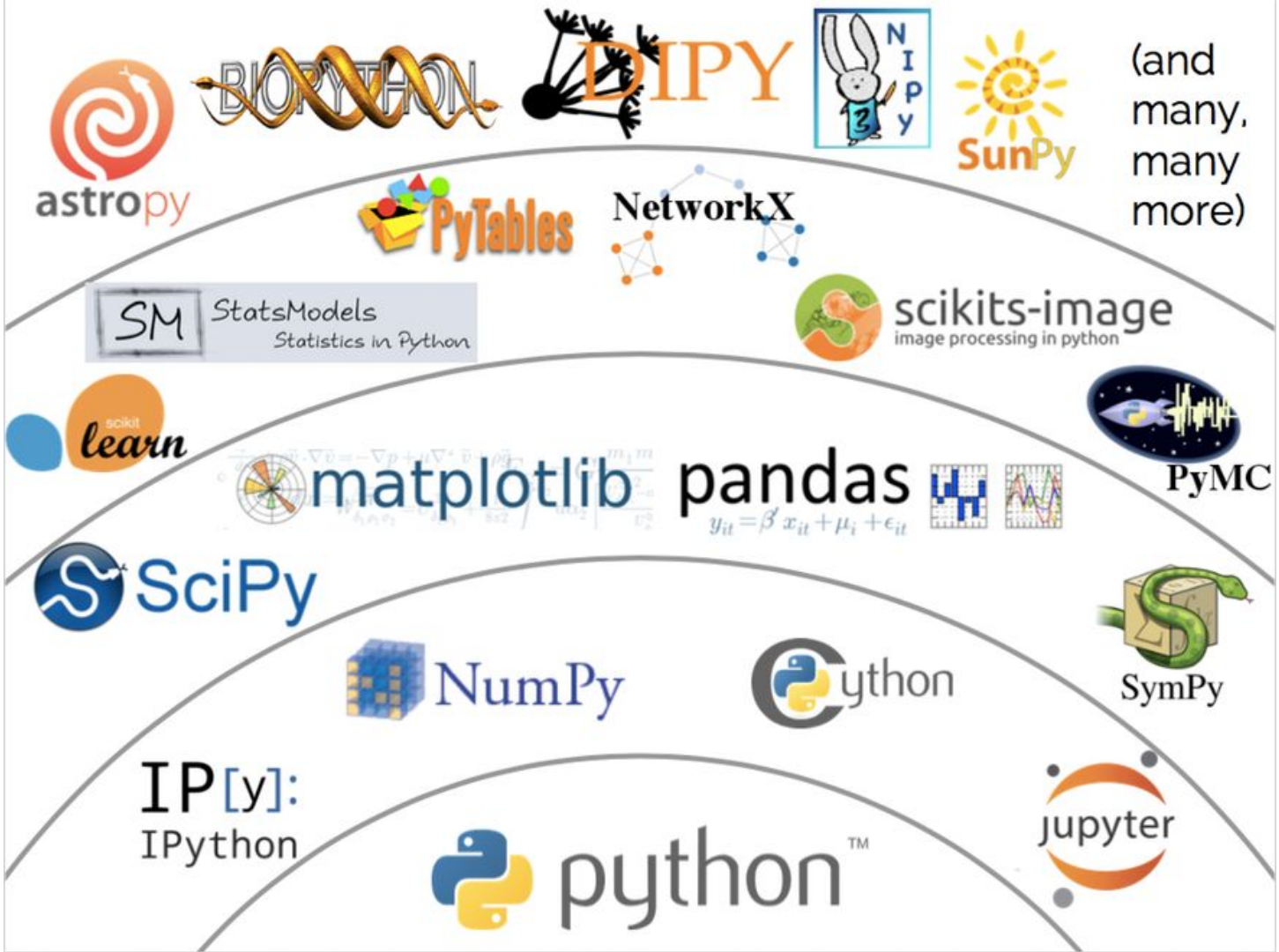
And we can easily view it's spectral structure using matplotlib's builtin specgram routine:

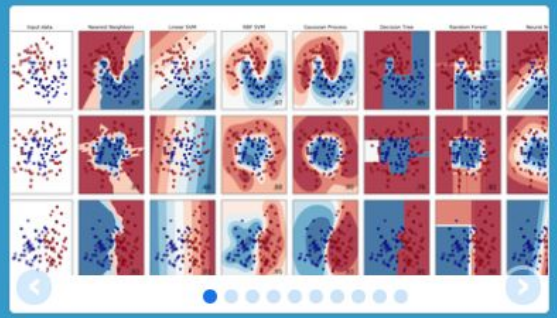
In [5]:
```
fig, (ax1, ax2) = plt.subplots(1,2,figsize(16,5))
ax1.plot(x); ax1.set_title('Raw audio signal')
ax2.specgram(x); ax2.set_title('Spectrogram');
```






Mr. Bult
Mrs. Vulcany
Bugle
Mr. Arrowpoint
Mrs. Startin
Amy
Kate
Herr Klesmer
Mrs. Arrowpoint
Catherine
Mab
Mrs. Adam
Mrs. Davllow
Alice
Joseph Kalonymos
Lapidoth
Mrs. Torrington
Baroness Langen
Mrs. Meyrick
Mirah
Hans
Lady Flora
Mrs. Davilow
Gwendolen
Mordecai
Buchan
Miss Merry
Juliet Fenn
Deronda
Gideon
Rash
Marrables
Mrs. Gascoigne
Joel Dagge
Mr. Middleton
Mr. Gascoigne
Jacob
Lily
Goodwin
Rex
Sir Hugo
Mr. Cohen
Miller
Anna
Addy
Isabel
Grandcourt
Mrs. Cohen
Lady Mallinger
Leonora Halm-Eberstein
Clintock
Mrs. Gadsby
Lady Pentreath
Adelaide
Lord Brackenshaw
Lush
Mrs. Raymond
Mrs Vulcany
Lady Brackenshaw
Mr. Fraser
Mr. Vandernoodt
Mrs. Glasher
Henleigh


astropy
BIOPYTHON
DIPY
NIPY
SunPy
(and many, many more)
PyTables
NetworkX
SM StatsModels Statistics in Python
scikits-image image processing in python
scikit learn
matplotlib
pandas
PyMC
SciPy
NumPy
Cython
SymPy
IP[y]: IPython
python™
jupyter



scikit-learn

Machine Learning in Python

- Simple and efficient tools for data mining and data analysis
- Accessible to everybody, and reusable in various contexts
- Built on NumPy, SciPy, and matplotlib
- Open source, commercially usable - BSD license

Classification

Identifying to which category an object belongs to.

Applications: Spam detection, Image recognition.
Algorithms: SVM, nearest neighbors, random forest, … — Examples

Regression

Predicting a continuous-valued attribute associated with an object.

Applications: Drug response, Stock prices.
Algorithms: SVR, ridge regression, Lasso, … — Examples

Clustering

Automatic grouping of similar objects into sets.

Applications: Customer segmentation, Grouping experiment outcomes
Algorithms: k-Means, spectral clustering, mean-shift, … — Examples

Dimensionality reduction

Reducing the number of random variables to consider.

Applications: Visualization, Increased efficiency
Algorithms: PCA, feature selection, non-negative matrix factorization. — Examples

Model selection

Comparing, validating and choosing parameters and models.

Goal: Improved accuracy via parameter tuning
Modules: grid search, cross validation, metrics. — Examples

Preprocessing

Feature extraction and normalization.

Application: Transforming input data such as text for use with machine learning algorithms.
Modules: preprocessing, feature extraction.

— Examples

Tutorials

Images
MNIST
Image Recognition
Image Retraining
Convolutional Neural Networks

Sequences
Recurrent Neural Networks
Neural Machine Translation
Drawing Classification
Simple Audio Recognition

Data Representation
Linear Models
Wide & Deep Learning
Vector Representations of Words
Kernel Methods

Non-ML

A Guide to TF Layers: Building a Convolutional Neural Network

The TensorFlow `layers` module provides a high-level API that makes it easy to construct a neural network. It provides methods that facilitate the creation of dense (fully connected) layers and convolutional layers, adding activation functions, and applying dropout regularization. In this tutorial, you'll learn how to use `layers` to build a convolutional neural network model to recognize the handwritten digits in the MNIST data set.

The MNIST dataset comprises 60,000 training examples and 10,000 test examples of the handwritten digits 0–9, formatted as 28x28-pixel monochrome images.

Getting Started

Let's set up the skeleton for our TensorFlow program. Create a file called `cnn_mnist.py`, and add the following code:

```
from __future__ import absolute_import
from __future__ division
```

Sommaire
Getting Started
Intro to Convolutional Neural Networks
Building the CNN MNIST Classifier
Input Layer
Convolutional Layer #1
Pooling Layer #1
Convolutional Layer #2 and Pooling Layer #2
Dense Layer
Logits Layer
Generate Predictions
Calculate Loss
Configure the Training Op
Add evaluation metrics
Training and Evaluating the CNN MNIST Classifier
Load Training and

AWS services

Find a service by name or feature (for example, EC2, S3 or VM, storage).

Recently visited services

EC2

All services

Compute
EC2
EC2 Container Service
Lightsail
Elastic Beanstalk
Lambda
Batch

Developer Tools
CodeStar
CodeCommit
CodeBuild
CodeDeploy
CodePipeline
X-Ray

Internet of Things
AWS IoT
AWS Greengrass

Contact Center
Amazon Connect

Storage
S3
EFS
Glacier
Storage Gateway

Management Tools
CloudWatch
CloudFormation
CloudTrail
Config
OpsWorks
Service Catalog
Trusted Advisor
Managed Services

Game Development
Amazon GameLift

Mobile Services
Mobile Hub
Cognito
Device Farm
Mobile Analytics
Pinpoint

Database
RDS
DynamoDB
ElastiCache
Amazon Redshift

Security, Identity &

Application Services

Helpful tips

Manage your costs

Get real-time billing alerts based on your cost and usage budgets. Start now

Create an organization

Use AWS Organizations for policy-based management of multiple AWS accounts. Start now

Explore AWS

Amazon Relational Database Service (RDS)

RDS manages and scales your database for you. RDS supports Aurora, MySQL, PostgreSQL, MariaDB, Oracle, and SQL Server. Learn more.

Real-Time Analytics with Amazon Kinesis

Stream and analyze real-time data, so you can get timely insights and react quickly. Learn more.

Get Started with Containers on AWS

Amazon ECS helps you build and scale containers for any size application. Learn more.

Se former à la Data Science

Quelques pistes

A l'UTC

MOOC (Coursera)

Kaggle

Kdnuggets

Data Venture

Se former à l’UTC

GI - Fouille De Données (FDD)

- SY09
  - Analyse de données et apprentissage automatique
- SY19
  - Apprentissage automatique
- NF26
  - Data warehouse

Autres

- MT09 / RO04
  - Analyse numérique
  - Optimisation
- IC05
  - Data Visualisation (Gephi)
- SY10
  - Logique floue

Machine Learning

Vue d'ensemble

Programme de cours

FAQ

Créateurs

Notation et examens

S'inscrire
Commence le Mar 05

Apply for Financial Aid

À propos de ce cours : Machine learning is the science of getting computers to act without being explicitly programmed. In the past decade, machine learning has given us self-driving cars, practical speech recognition, effective web search, and a vastly improved understanding of the human genome. Machine learning is so pervasive today that you probably use it dozens of times a day without knowing it. Many

Plus

Créé par : Stanford University

Stanford University

Enseigné par : Andrew Ng, Co-founder, Coursera; Adjunct Professor, Stanford University; formerly head of Baidu AI Group/Google Brain

Langue

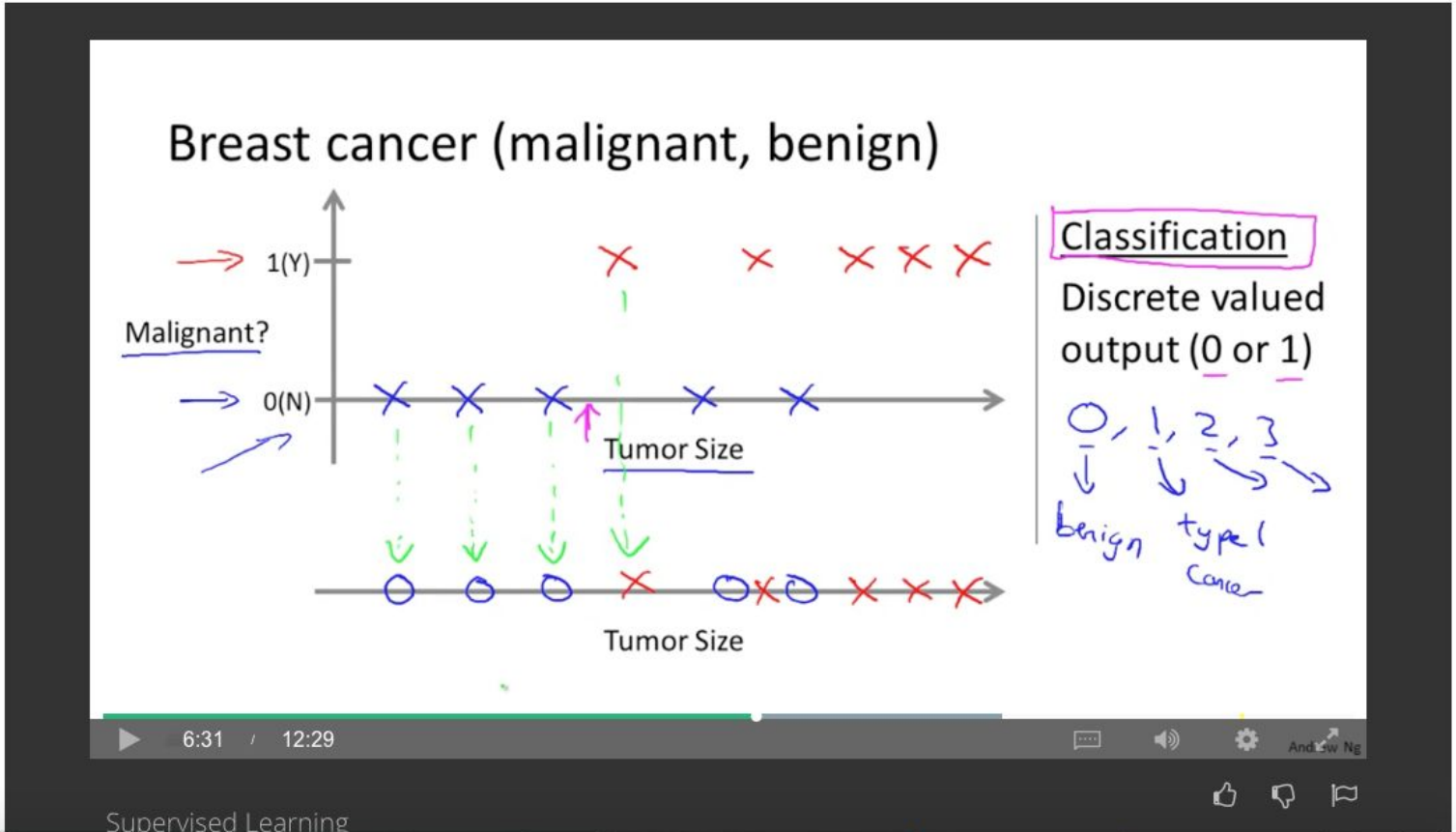
English, **Sous-titres :** Spanish, Hindi, Japanese, Chinese (Simplified)

Welcome

Introduction

- Welcome 6 min
- What is Machine Learning? 7 min
- What is Machine Learning? 5 min
- How to Use Discussion Forums 4 min
- Supervised Learning 12 min
- Supervised Learning 4 min
- Unsupervised Learning 14 min
- Unsupervised Learning 3 min
- Who are Mentors? 3 min
- Get to Know Your Classmates 8 min
- Frequently Asked Questions 11 min

Review



Supervised Learning

Sécurisé | https://www.kaggle.com/competitions

kaggle Search kaggle Competitions Datasets Kernels Discussion Jobs ...

Welcome to Kaggle Competitions

Challenge yourself with real-world machine learning problems

New to Data Science?
Get started with a tutorial on our most popular competition for beginners, Titanic: Machine Learning from Disaster.

Build a Model
Get the data & use whatever tools or methods you prefer to make predictions.

Make a Submission
Upload your prediction file for real-time scoring & a spot on the leaderboard.

Learn more | InClass

× Dismiss

General | InClass

Sort by Grouped

Machine Learning, Data Science, Data Mining, Big Data, Analytics, AI

Software (Suites, Text, Visualization)
Jobs - Industry | Academic
Meetings, Conferences
Companies (Consulting, Products)
Courses in Big Data, Data Science
Datasets (APIs/Markets, Gov)
Data Mining Course | Gregory Piatetsky
Education (online, USA, Europe, cert)
FAQ | Polls | Publications (Books)
Solutions (Fraud, Data Cleaning)
Webcasts | Websites (Blogs, Cartoons, Podcasts)

KDnuggets News

Latest News & Stories
Upcoming Schedule
Subscribe to KDnuggets News

Most Recent

- TDWI Chicago, May 6-11: Get Your Hands Dirty With Data ...
- Upcoming Meetings in AI, Analytics, Big Data, Data Scie...
- For GPU Databases of today, the big challenge is doing ...
- Data Science in Fashion
- Data Science for Javascript Developers
- Unleash a faster Python on your data

Latest poll results: Artificial General Intelligence (AGI) in less than 50 years, say KDnuggets readers

Latest
News | Tutorials

- For GPU Databases of today, the big challenge...
- Data Science in Fashion
- Data Science for Javascript Developers
- Unleash a faster Python on your data
- Is Google Tensorflow Object Detection API the...

Opinions, Interviews, Reports

Tweets by @kdnuggets

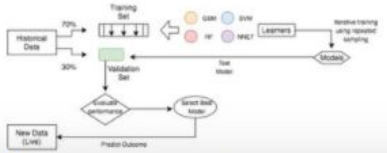
KDnuggets @kdnuggets
Unleash a faster #Python on your data buff.ly/2GXmo78

12m

KDnuggets @kdnuggets
How Docker Can Help You Become A More Effective #DataScientist #KDN ow.ly/MvN730iz5qn

29m

KDnuggets @kdnuggets
How #DataScience can improve retail buff.ly/2t9ecP7



AfterWorks