






Using Kaggle to become a better data analyst

by Françoise Provencher
PyLadies Montreal, November 20th, 2014

Active Competitions

		Click-Through Rate Prediction Predict whether a mobile ad will be clicked	2 months 120 teams \$15,000
		Sentiment Analysis on Movie Reviews Classify the sentiment of sentences from the Rotten Tomatoes dataset	3 months 552 teams Knowledge
		Finding Elo Predict a chess player's FIDE Elo rating from one game	4 months 64 teams Knowledge
	Insert (noun?) here?	Billion Word Imputation Find and impute missing words in the billion word corpus	5 months 36 teams Knowledge

Kaggle is...

a website for data analysis competitions

Learning from Kaggle

- Nice 101 level « competitions »
 - Titanic
 - Data science London + Scikit-learn
- Tutorials
- Forums

Why I like it

- **Clean datasets** that everyone can use
- Learn **new techniques** (e.g. random forests)
- Fun **weekend project** that I can work on with friends
- It boosts my **motivation** and my **curiosity**
- Take a peak at **other people's code**

Getting started

- Install useful **Python libraries**, such as Scikit-learn, Pandas, Numpy and Scipy
- Create an account on **Kaggle**
- Choose a problem you would like to tackle and **download the dataset**
- **Explore**, try things, break things and share your discoveries!

Good resources

- Andrew Ng's class on Coursera about machine learning
 - <https://www.coursera.org/course/ml>
- Scikit-learn doc:
 - <http://scikit-learn.org/>

Demo time!

Announcements

Next PyLadies Meetup

- December 11th
- Hacking RaspberryPi with Nicole Parrot
- Ethernet cable + laptop (vnc viewer) + earbuds
- Bring a breadboard if you have one



Montreal Python



- December 1st at Notman house
- All about data science!
- Cameron will give a talk!
- <http://montrealpython.org/en/2014/11/mp50/>



Coffee and Coding



- Meet once a month in a café
- Code together at the same table
- Work on individual projects or group projects
- Helping each other out and having a good time

Project idea

- Use **Git** to clone this existing project:
 - <https://github.com/scrapinghub/pycon-speakers>
- Use **web scraping** for gathering the names of speakers at big conferences (e.g. PyCon)
- Use **machine learning** to guess the speakers' gender
- **Plot** the gender ratios for a given conference over time
 - Are there more women speakers now than 5 years ago?
 - North America vs Europe vs Asia?
 - Python vs Java?
 - Computer science vs Chemistry vs Physics?