Data / Text Mining

Renaud Richardet, PhD
Brain Mind Institute
EPFL, Switzerland
renaud.richardet@epfl.ch

Agenda

1) Introduction
2) Basics
    a) quick start with UNIX tools
    b) regular expressions
    c) know your text editor
3) Data acquisition
    a) strategy, considerations, copyrights
    b*) crawling data from an API (PubMed)
    c*) scraping data from a website
4) Data processing
    a*) text processing with NLTK
    b*) full-text search with Elasticsearch
    c*) a sense of machine learning
    d*) semantic similarity with word2vec
5) Your usecases

About me

- PhD from EPFL's Blue Brain project
- working on large scale natural language processing applications for the biomedical domain
- research scientist and senior developer in Switzerland, India and in the USA
- Primary fields of interests:
  - natural language processing (NLP)
  - deep learning
  - big data
  - open source software

About you: put your hands up if you know about:

- writing a simple program in an interpreted language
- creating an html page
- with some Javascript
- using regular expressions
- programatically accessing a public API
- part-of-speech tagging
- stemming
- inverted index
- ElasticSearch
- word2vec

Today: a broad overview

Everything there is about text mining

What we will work on today

source: http://worrydream.com/ClimateChange/

Introduction: a few principles

Introduction: a few principles

move fast and break things.
unless you are breaking stuff
you are not moving fast
enough

M. Zuckerberg

Introduction: a few principles

don't be evil

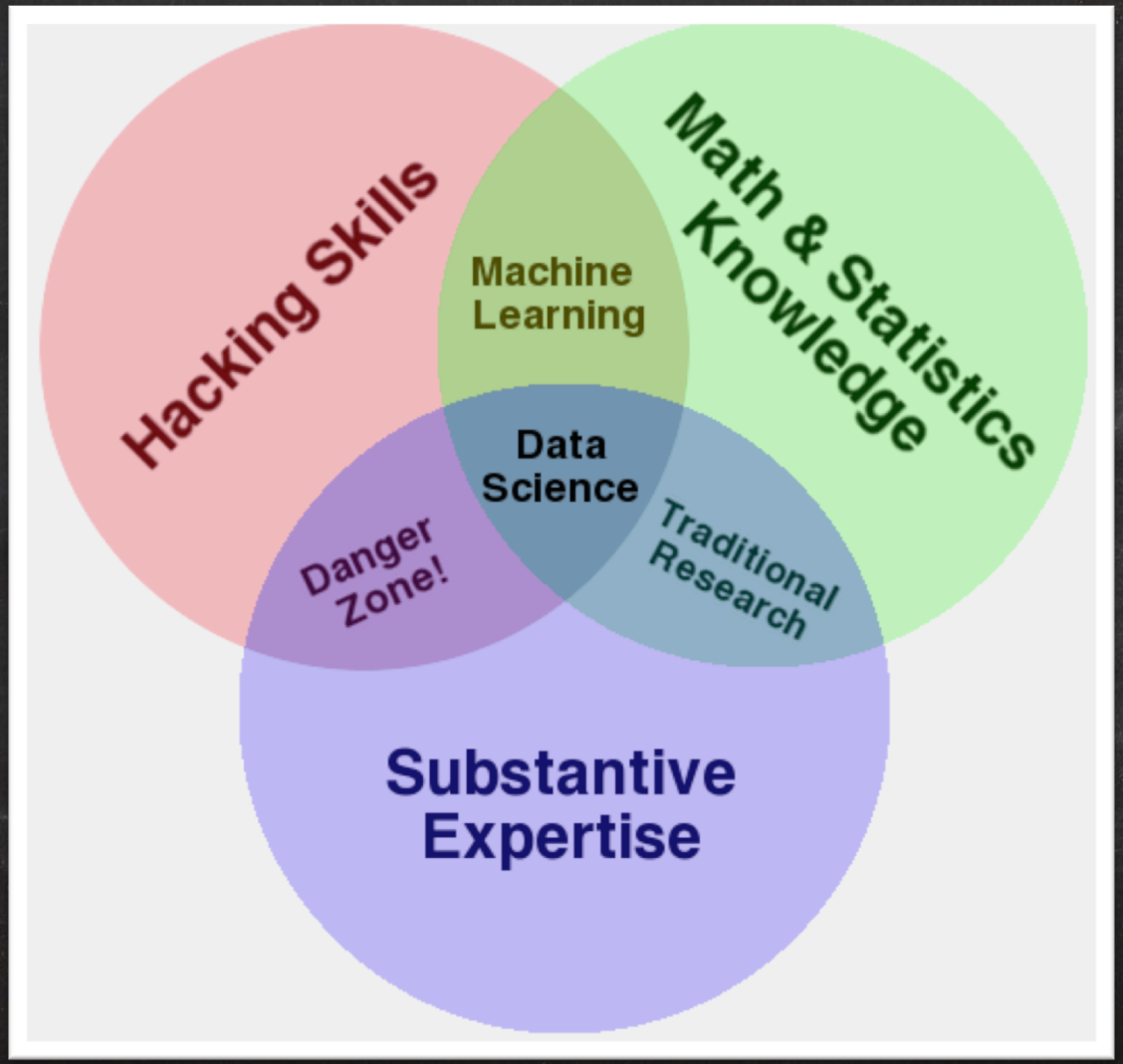
Google

Introduction: a few principles

if all else fails, you can write a program

Unknown

Introduction: Data Science



source: http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram

Introduction: Minimum Viable Product