IMPRO-3:
Big Data Analytics Project

Course Introduction

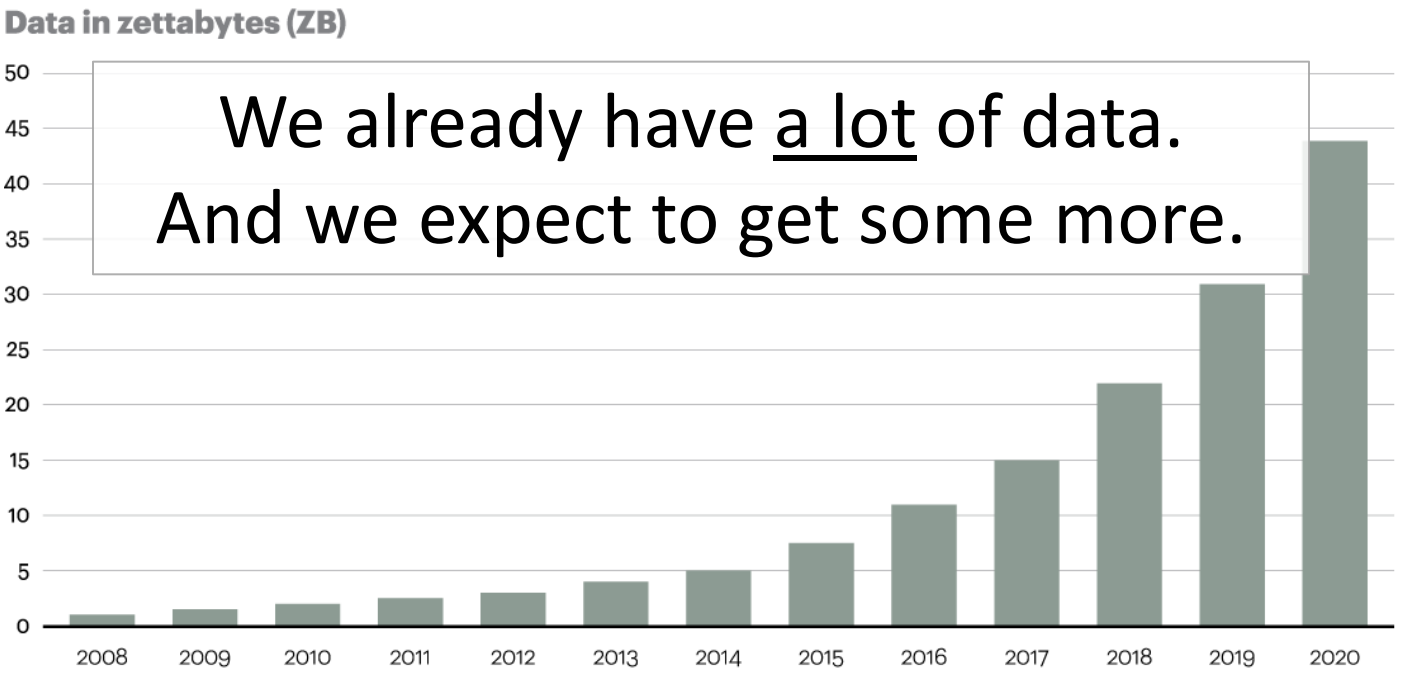
It’s so damn early…

AND YET YOU MADE IT!

And I know why…

Figure 1

Data is growing at a 40 percent compound annual rate, reaching nearly 45 ZB by 2020



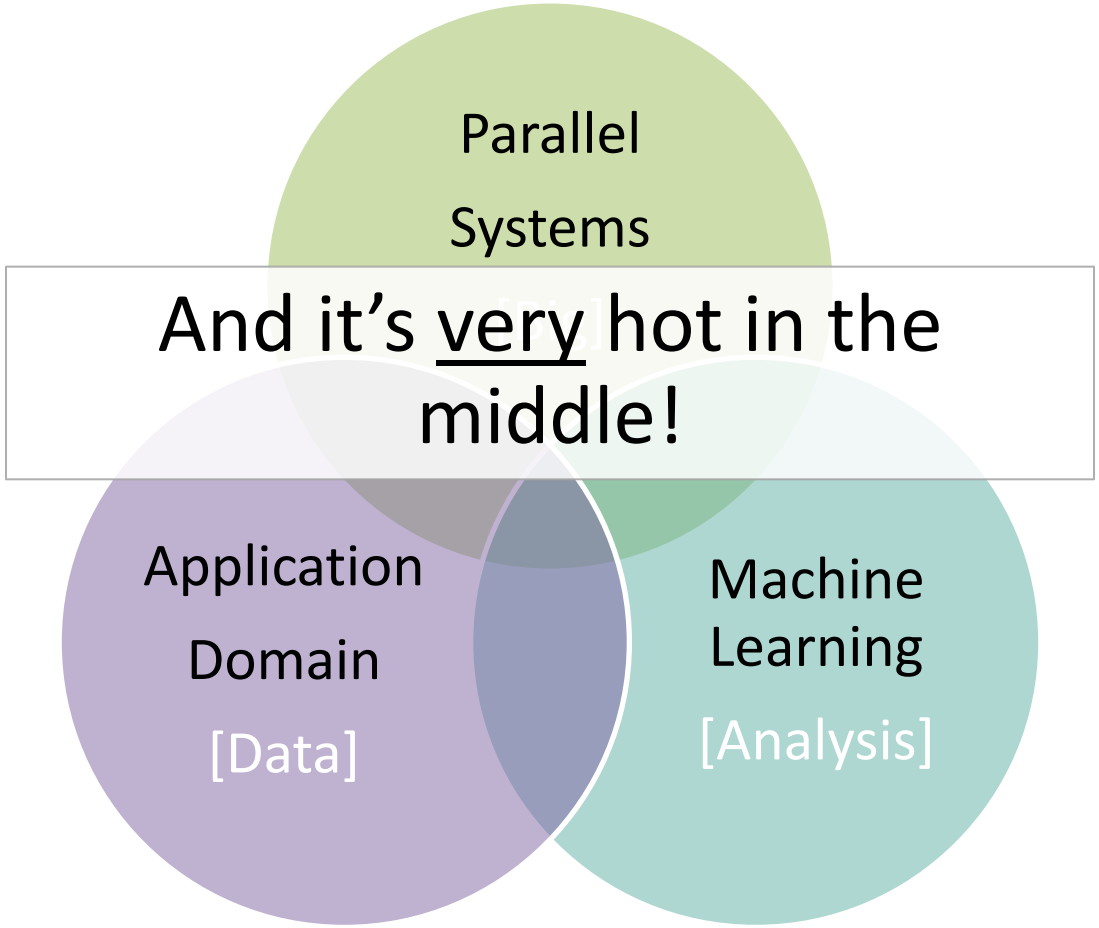
Source: Oracle, 2012

Source: http://shar.es/TvYIZ

And I know why…

Source: http://goo.gl/7rBPW7

A unique set of skills is required…



…But It's a Long Way To The Top (If You Wanna Rock N' Roll)

PROJECT GOALS

What will you get ?

- Learn a Machine Learning algorithm
  - Classification (Naïve Bayes, Logistic Regression, Random Forest, Hidden Markov Models)
  - Clustering (K-Means, Canopy Clustering, K-Means++)
- Learn about Parallel Data Processing with open-source software
  - Theoretical Background
  - [Apache] Stratosphere
  - [Apache] Apache Spark
- Get hands-on experience in [Big] [Data] [Analysis]
- Learn how to test and evaluate scalable data analysis programs

What will we get?

A control group of students

Some feedback

ROADMAP

Preliminary Plan

1. Pick a Machine Learning algorithm — W01
2. Read about it! — W01 – W02
3. Hack it in Scala — W03 – W04
4. Learn about Parallel Data Processing — W04
5. Implement a scalable versions of your algorithm — W05 – W12
   - In Stratosphere — W05 – W09
   - In Spark — W10 – W12
6. Test and analyze your implementations — W13
7. Summarize your experiences in a final presentation! — W15

Course Organizers

- Andreas Kunft
  - andreas.kunft@tu-berlin.de
- Alexander Alexandrov
  - alexander.alexandrov@tu-berlin.de
- Asterios Katsifodimos
  - asterios.katsifodimos@tu-berlin.de

Tasks for Today

- Organize yourselves in groups of three
- Pick an algorithm
- Create a Git account

Next Week: Introduction in Scrum!