Chapter 1:

Introduction to Big Data — the four V's

This chapter is mainly based on the Big Data script by Donald Kossmann and Nesime Tatbul (ETH Zürich)
Goal of Today

• **What is Big Data?**
  • introduce some major buzz words

• **What is not Big Data?**
  • get a feeling for opportunities & limitations
Answering Tough Questions

- **Problem:**
  - sales for lollipops are going down

- **Data:**
  - all sales data by customer, region, time, ...

- **Information:**
  - lollipops bought by people older than 25 (but eaten by people younger than 10)

- **Knowledge:**
  - moms believe: lollipops = bad teeth

- **Value:**
  - dentists advertise your lollipops
Why is this difficult?

- You need more data than your data warehouse.
  - you need more data that you have
  - logs, Twitter feeds, blogs, customer surveys, ...

- You need to ask the right questions.
  - data alone is silent

- You need technology and organization that help you concentrate on asking the right questions.
From “Small Data” to “Big Data”

• Step 1:

You! (TB)
• Step 2:

You! (PB)

Your Friends (TB)
• Step 3:

You! (PB)

Your Friends (TB/PB)

The World (EB)
Limitations of State of the Art

Business Processes

Storage Network

data is dead

Archive

does not scale

Manual Analysis

ETL = (Extract, Transform, Load)

ETL into RDBMS

inflexible + data loss
What needs to be done? (Technology)

- Take Steps 0 to 3
  - Step 0: Data Warehouses (relational Databases)
  - Step 1: Data Warehouses + Hadoop (HDFS)
  - Step 2: Business Processes + Analytics + Exchange
  - Step 3: BP + Analytics + Exchange + Real-Time
What needs to be done? (Technology)

- **Take Steps 0 to 3**
  - Step 0: Data Warehouses (relational Databases)
  - Step 1: Data Warehouses + Hadoop (HDFS)
  - Step 2: Business Processes + Analytics + Exchange
  - Step 3: BP + Analytics + Exchange + Real-Time
What needs to be done? (Technology)

• **Take Steps 0 to 3**
  • Step 0: Data Warehouses (relational Databases)
  • Step 1: Data Warehouses + Hadoop (HDFS)
  • **Step 2: Data Warehouses + Hadoop + XML (Standards)**
  • Step 3: BP + Analytics + Exchange + Real-Time
What needs to be done? (Technology)

• Take Steps 0 to 3
  • Step 0: Data Warehouses (relational Databases)
  • Step 1: Data Warehouses + Hadoop (HDFS)
  • Step 2: Data Warehouses + Hadoop + XML (Standards)
  • Step 3: Data Warehouses + Hadoop + XML + ?
What needs to be done? (Organisation)

• **Static Business Model -> Agile Business Model**
  • You and your customers adapt to each other
  • No more data silos (ownership of data is distributed)
  • You allocate resources on demand

• **Execute Business Process -> Data Science**
  • You think about experience you have made
What is Big Data?

• Three alternative perspectives
  • philosophical
  • business
  • technical

• (Ultimately, it is a buzz word for everybody.)
Philosophical

• What is more valuable, if you had to pick one?
  • experience or intelligence?

• Traditional (computer) science: logic!
  • understand the problem, build model / algorithm
  • answer question from implementation of model

• New twist in (computer) science: statistics!
  • collect data
  • answer question from data (what did others do?)
Statistics vs. Logic?

• Problems:
  • Find a spouse?
  • Should Adam bite into the apple?
  • $1 + 1$?
  • Cure for cancer?
  • How to treat a cough?
  • Should I give Matthias a loan?
  • Premium for life insurance?
  • When should my son come home?
  • Which book should I read next?
  • Translate from German to English.

What Type of solutions (statistic/logic)? What do you think?
Statistics vs. Logic?

• Problems:
  • Find a spouse?  
  • Should Adam bite into the apple?  
  • 1 + 1?  
  • Cure for cancer?  
  • How to treat a cough?  
  • Should I give Matthias a loan?  
  • Premium for life insurance?  
  • When should my son come home?  
  • Which book should I read next?  
  • Translate from German to English.

• Is there a solution?
  • I don’t want to know!  
  • If you believe…  
  • Yes (Definition)  
  • I don’t know, maybe.  
  • Yes (Google Insight)  
  • Yes (e.g. Schufa)  
  • YES (e.g. Alliance)  
  • No, but…  
  • Yes (e.g. Amazon)  
  • Yes (Google Transl.)
Data Science, 4th Paradigm

• New approach to do science
  • Step 1: Collect data
  • Step 2: Generate Hypotheses
  • Step 3: Validate Hypotheses
  • Step 4: (Goto Step 1 or 2)

• Why is this a good approach?
  • it can be automated: no thinking, less error

• Why is this a bad approach?
  • how do you debug without a ground truth?
Is bigger = smarter?

• Yes!
  • tolerate errors
  • discover the long tail and corner cases
  • machine learning works much better
**Is bigger = smarter?**

- **Yes!**
  - tolerate errors
  - discover the long tail and corner cases
  - machine learning works much better

- **But!**
  - more data, more error (e.g., semantic heterogeneity)
  - with enough data you can prove anything
  - still need humans to ask right questions
Big Data Success Story

• Google Translate
  • you collect snippets of translations
  • you match sentences to snippets
  • you continuously debug your system

• Why does it work?
  • there are tons of snippets on the Web
  • there is a ground truth that helps to debug system
Big Data Farce (only a joke)

- Which lane is fastest in a traffic jam?
  - you ask people where they go and whether happy
  - (maybe, you even use a GPS device)
  - you conclude that left lane is fastest

- Why is this stupid?
  - because there is no ground truth!
  - you will get a conclusion because Big Data always gives an answer. But, it does not make sense!
  - getting more data does not help either
How to play lottery in Napoli

• Step 1: You visit (and pay) “oracles”
  • they tell you which numbers to play

• Step 2: You visit (and pay) “interpreters”
  • they explain what oracles told you

• Step 3: After you lost, you visit (and pay) “analyst”
  • they explain why “oracles” and “interpreters” were right
  • goto Step 1

• Lessons learned
  • life is try and error; trying keeps the system running

[Luciano de Crescenzo: Thus Spake Bellavista]
What is Big Data?

• **Business Perspective**
  • it is a new business model

• **People pay with data**
  • e.g. Facebook, Google, Twitter:
    • use service, give data
    • Google sells your data to advertisers
    • (you pay advertisers indirectly)
  • e.g., 23andMe, Amazon:
    • pay service + give data
    • sells data and uses data to improve service
Business Perspective

- **Bank**
  - keeps your money securely (kind of...)
  - puts your money at work (lends it to others), interest
  - you keep ownership of money and take it when needed

- **Databank**
  - keeps your data securely (kind of...)
  - puts your data at work: interest or better service
  - (you keep ownership of data: hopefully to come)
Technical Perspective (us!)

• You collect all data
  • the more the better -> statistical relevance, long tail
  • keeping all is cheaper than deciding what to keep

• You decide independently what to do with data
  • run experiments on data when question arises

• Huge difference to traditional information systems
  • design upfront what data to keep and why!!!
  • (e.g., waterfall model of software engineering!)
Consequences

• **Volume:** data at rest
  • it is going to be a lot of data

• **Speed:** data in motion
  • it is going to arrive fast

• **Diversity:** data in many formats
  • it is going to come in different shapes
  • (e.g., different versions, different sources)

• **Complexity:** You want to do something interesting
  • SQL will not be enough
The 4 Vs of Big Data

- **Volume**: same as before
- **Velocity**: same as "speed"
- **Variety**: same as "diversity"
- **Veracity**: data in doubt
  - you do not know exactly what you have
Alternative Definition:

Literature does not agree upon the # of Vs defining Big Data

Examples:

• Laney 2001
  talks about 3 Vs: volume, velocity, and variety

• later in Van Rijmenam 2014 and Borne 2014
  it is pointed out that 3Vs are insufficient. 
  In addition to volume, velocity, and variety, further 7 Vs are identified: 
  veracity, validity, value, variability, venue, vocabulary, and vagueness
Four Vs of Big Data

The FOUR V’s of Big Data

Volume
- 40 Zettabytes (40 trillion gigabytes) of data will be created by 2020, an increase of 300 times from 2005
- 6 billion people have cell phones
- World population: 7 billion

High Speed Data
- The New York Stock Exchange captures 1 TB of trade information during each trading session
- Modern cars have close to 100 sensors that monitor items such as fuel level and tire pressure

By 2015, 4.4 million IT jobs will be created globally to support big data, with 1.9 million in the United States

Variety
- As of 2011, the global size of data in healthcare was estimated to be 150 exabytes (150 billion gigabytes)
- 30 billion pieces of content are shared on Facebook every month
- 420 million wearable, wireless health monitors
- 4 billion hours of video are watched on YouTube each month
- 400 million tweets are sent per day by about 200 million monthly active users

Veracity
- By 2014, it’s anticipated there will be 18.9 billion network connections
- Almost 2.5 connections per person on earth
- 1 in 3 business leaders don’t trust the information they use to make decisions

- Poor data quality costs the US economy around $3.1 trillion a year
- 27% of respondents in one survey were unsure of how much of their data was inaccurate

Sources: McKinsey Global Institute, Twitter, Cisco, Gartner, EMC, SAS, IBM, NIST, GAO
Overview

• Intro
• What is Big Data?
• NoSQL Systems
• Hadoop / HDFS / MapReduce & Applications
• Spark
• Data Streams & Applications
  Storm, ...
• Text Data
• High-Dimensional Data
• Graph Data
• Uncertain Data
Why now?

- **Mega-trend:** All data is digital, digitally born!
  - 70 years ago: computers for “+”
  - 15 years ago: disks cheaper than paper
  - 7 years ago: Internet has eyes and ears
- **Because we can**
  - 40 years of databases -> volume
  - 40 years of Moore’s law -> complexity
  - 2000+ years of statistics -> it is only counting
  - enough optimisms that we get the rest done, too
- **Because we reached dead end with logic (?)**
Because we can... Really?

- Yes!
  - all data is digitally born
  - storage capacity is increasing
  - counting is embarrassingly parallel
Because we can… Really?

• **Yes!**
  - all data is digitally born
  - storage capacity is increasing
  - counting is embarrassingly parallel

• **But,**
  - data grows faster than energy on chip
  - value / cost tradeoff unknown
  - ownership of data unclear (aggregate vs. individual)
What you have learnt today?

• a number of buzz words, some cool examples
  • you should survive any discussion with your boss

• motivation to come back next week
  • learn some of the technologies