Outline

- Why to study Data Mining?
- Why do we need Data Mining?
- What is Knowledge Discovery in Databases (KDD) and Data Mining?
- Main data mining tasks
- What’s next

Why to study Data Mining – famous quotes*

- **Data Mining** is often associated with **Machine Learning.** It is an area that has taken much of its inspiration and techniques from machine learning (and some, also, from statistics), but is put to different *ends*.
- **Data Mining** is about using statistics as well as other programming methods to find patterns hidden in the data so that you can *explain* some phenomenon.
- **Machine Learning** uses **Data Mining** techniques and other learning algorithms to build models of what is happening behind some data so that it can *predict* future outcomes.
- “*A breakthrough in machine learning would be worth ten Microsofts*” (Bill Gates, Microsoft)
- “*Machine learning is the next Internet*” (Tony Tether, Director, DARPA)
- “*Machine learning is the hot new thing*” (John Hennessy, President, Stanford)
- “*Machine learning is going to result in a real revolution*” (Greg Papadopoulos, Former CTO, Sun)
- “*Machine learning today is one of the hottest aspects of computer science*” (Steve Ballmer, CEO, Microsoft)

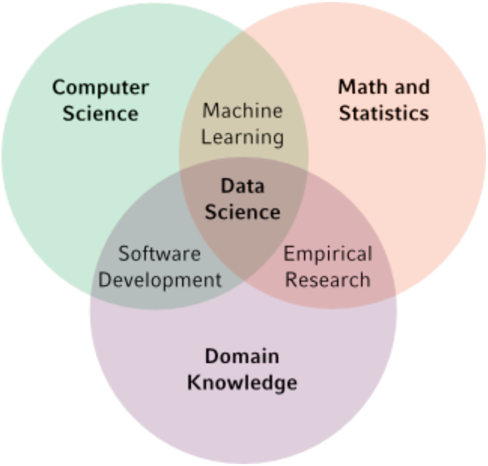
*Source: Pedro Domingos http://courses.cs.washington.edu/courses/cse446/15sp/slides/intro.pdf

Why to study Data Mining - Data Scientist: The sexiest job of 21st century

"If "sexy" means having rare qualities that are much in demand, data scientists are already there. They are difficult and expensive to hire and, given the very competitive market for their services, difficult to retain. There simply aren't a lot of people with their combination of scientific background and computational and analytical skills."

Source: Harvard Business Review. Data Scientist: The Sexiest Job of the 21st Century. October 2012 link

Key disciplines in Data Science:



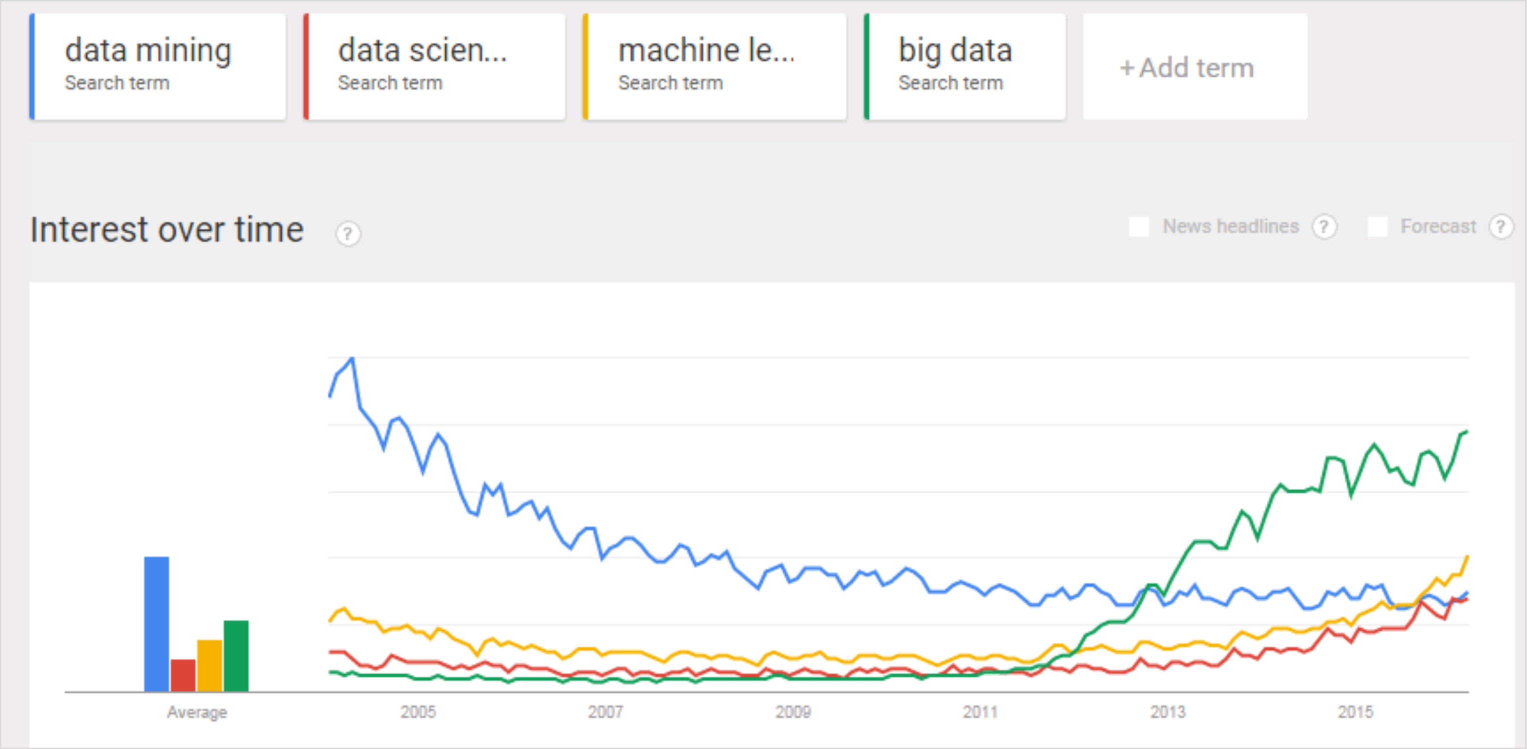
Data Mining – Data Science – Big Data – Machine Learning – Analytics …

- New fancy words for knowledge discovery from data
  - Data mining, machine learning have been focusing on knowledge discovery from data for decades
  - Well defined set of tasks and solutions
- Big data and analytics are more business terms and ill-defined

> *"Big data is like teenage sex: everyone talks about it, nobody really knows how to do it, everyone thinks everyone else is doing it, so everyone claims they are doing it."*
>
> Source: Dan Ariely, Duke University

- Though nowadays we have more data than ever and the infrastructure to deal with it
  - → more opportunities and challenges for data mining and machine learning

Interest over time



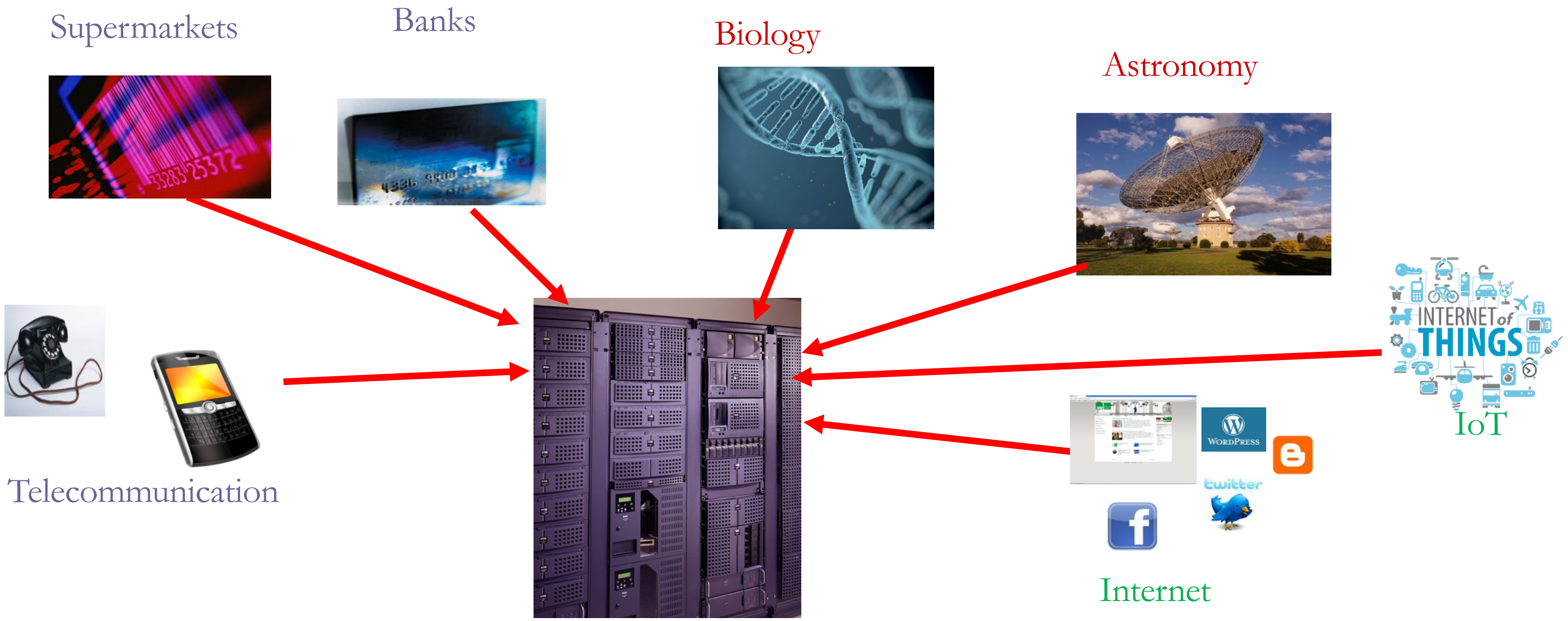
Source: Google trends, query on 18.3.2016

Outline

- Why to study Data Mining?
- Why do we need Data Mining?
- What is Knowledge Discovery in Databases (KDD) and Data Mining?
- Main data mining tasks
- What’s next

Why do we need Data Mining

- Huge amounts of data are collected nowadays from different application domains
- “*We are drowning in information but starving for knowledge*” John Naibett link
- The amount and the complexity of the collected data does not allow for manual analysis.



Supermarkets

Banks

Biology

Astronomy

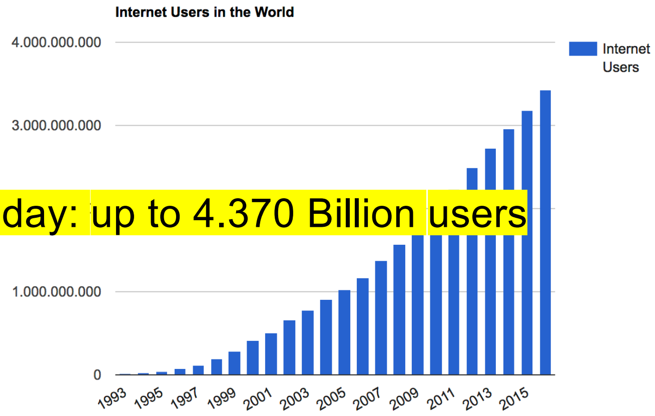
Telecommunication

IoT

Internet

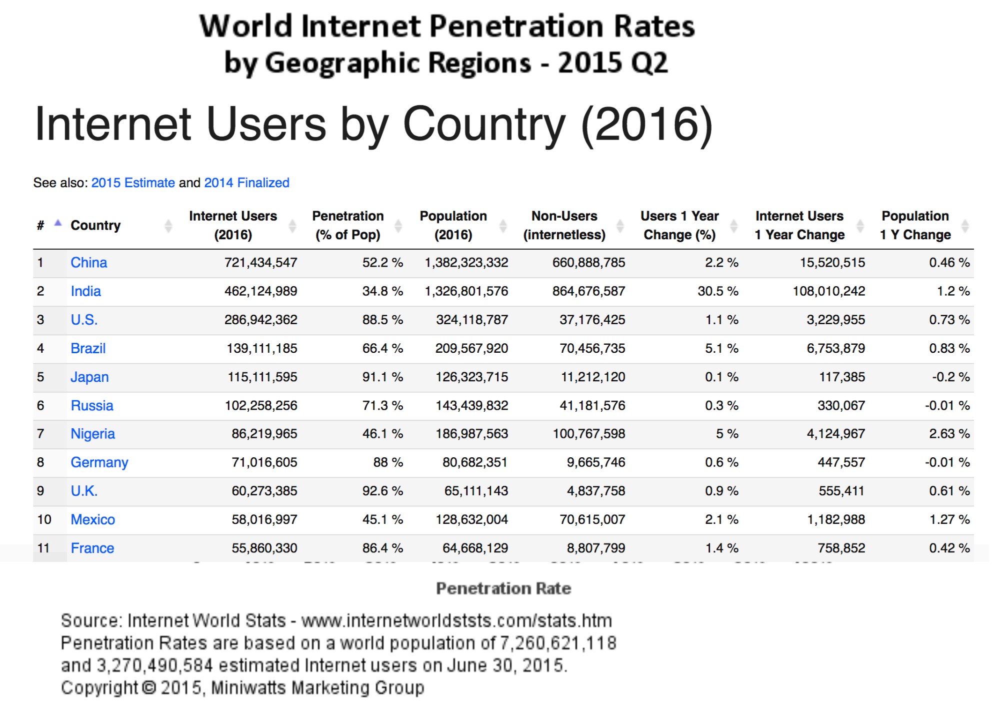
Examples of data sources: The Internet

- Internet users (Source: http://www.internetlivestats.com/internet-users/)



Today: up to 4.370 Billion users

Web 2.0: A world of opinions



World Internet Penetration Rates by Geographic Regions - 2015 Q2

Internet Users by Country (2016)

See also: 2015 Estimate and 2014 Finalized

#	Country	Internet Users (2016)	Penetration (% of Pop)	Population (2016)	Non-Users (internetless)	Users 1 Year Change (%)	Internet Users 1 Year Change	Population 1 Y Change
1	China	721,434,547	52.2 %	1,382,323,332	660,888,785	2.2 %	15,520,515	0.46 %
2	India	462,124,989	34.8 %	1,326,801,576	864,676,587	30.5 %	108,010,242	1.2 %
3	U.S.	286,942,362	88.5 %	324,118,787	37,176,425	1.1 %	3,229,955	0.73 %
4	Brazil	139,111,185	66.4 %	209,567,920	70,456,735	5.1 %	6,753,879	0.83 %
5	Japan	115,111,595	91.1 %	126,323,715	11,212,120	0.1 %	117,385	-0.2 %
6	Russia	102,258,256	71.3 %	143,439,832	41,181,576	0.3 %	330,067	-0.01 %
7	Nigeria	86,219,965	46.1 %	186,987,563	100,767,598	5 %	4,124,967	2.63 %
8	Germany	71,016,605	88 %	80,682,351	9,665,746	0.6 %	447,557	-0.01 %
9	U.K.	60,273,385	92.6 %	65,111,143	4,837,758	0.9 %	555,411	0.61 %
10	Mexico	58,016,997	45.1 %	128,632,004	70,615,007	2.1 %	1,182,988	1.27 %
11	France	55,860,330	86.4 %	64,668,129	8,807,799	1.4 %	758,852	0.42 %

Source: Internet World Stats - www.internetworldststs.com/stats.htm
Penetration Rates are based on a world population of 7,260,621,118
and 3,270,490,584 estimated Internet users on June 30, 2015.
Copyright © 2015, Miniwatts Marketing Group

Examples of data sources: Internet of things

- The Internet of Things (IoT) is the network of physical objects or "things" embedded with electronics, software, sensors, and network connectivity, which enables these objects to collect and exchange data.

Source: https://en.wikipedia.org/wiki/Internet_of_Things

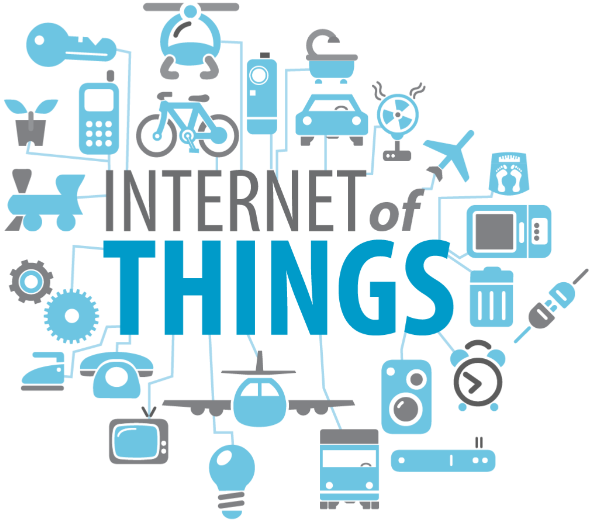


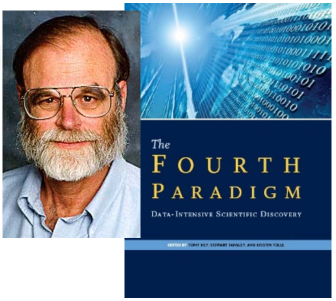
Image source:http://tinyurl.com/prtfqxf

During 2008, the number of things connected to the internet surpassed the number of people on earth… By 2020 there will be 50 billion … vs 7.3 billion people (2015).

These things are everything, smartphones, tablets, refrigerators …. cattle.

Source: http://blogs.cisco.com/diversity/the-internet-of-things-infographic

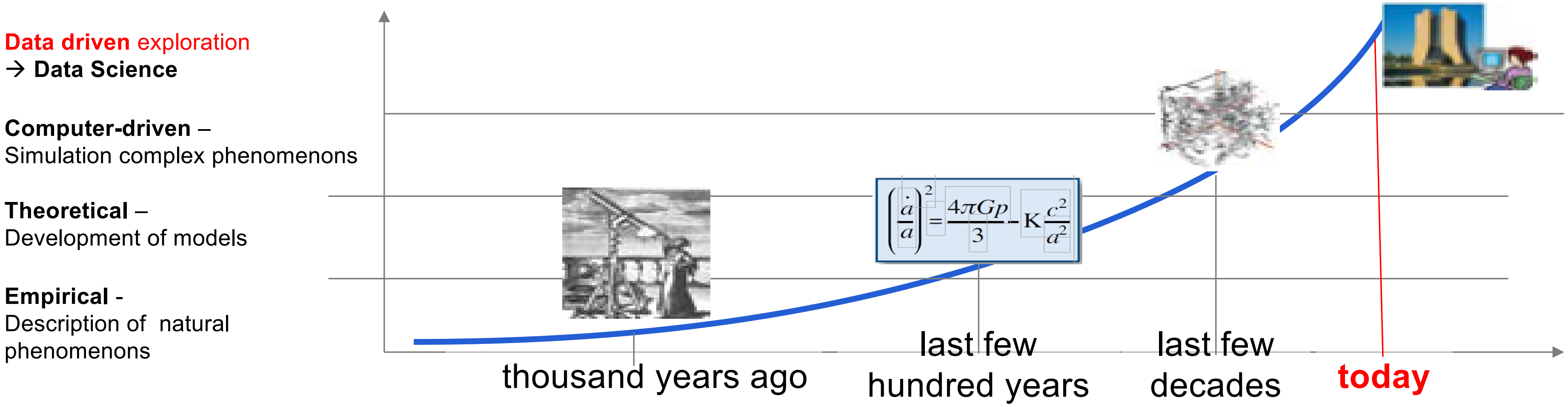
Examples of data sources: data intensive science



[Comp. Science Pionier Jim Gray]

- The Fourth Paradigm:
  Age of data driven exploration
  → Data Science

- Science Paradigms



Data driven exploration
→ **Data Science**

Computer-driven –
Simulation complex phenomenons

Theoretical –
Development of models

Empirical -
Description of natural
phenomenons

$$\left(\frac{\dot{a}}{a}\right)^2=\frac{4\pi Gp}{3}-K\frac{c^2}{a^2}$$

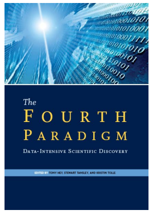
source:http://research.microsoft.com/en-us/um/people/gray/talks/nrc-cstb_escience.ppt

Examples of data sources: data intensive science

“Increasingly, scientific breakthroughs will be powered by advanced computing capabilities that help researchers manipulate and explore massive datasets.”

“*Modern science increasingly relies on integrated information technologies and computation to collect, process, and analyze complex data.*”

-The Fourth Paradigm – Microsoft



Examples of e-science applications:

- Earth and environment
- Health and wellbeing
  - E.g., The Human Genome Project (HGP)
- Citizen science
- Scholarly communication
- Basic science
  - E.g., CERN

Slide from:http://research.microsoft.com/en-us/um/people/gray/talks/nrc-cstb_escience.ppt

From data to knowledge

	Data	Methods	Knowledge
	Call records	Outlier Detection	Detect fraud cases
	Bank transactions	Classification	Customer credibility for loan applications
	Customer transactions from supermarkets/ online stores	Association rules	Which products people tend to buy together?
	Telescope images	Classification	What is the class of a star? E.g., early, intermediate or late formation

Outline

- Why to study Data Mining?
- Why do we need Data Mining?
- What is Knowledge Discovery in Databases (KDD) and Data Mining?
- Main data mining tasks
- What’s next