INTRODUCTION TO DATA SCIENCE

Introduction and Administration

Plan

Why data science is important?	• "Why are you here"
What is data science?	• Mashup of disciplines
What this course is about?	• Hopefully right mix of theory and practical skills
Course requirements	• Syllabus • Grade ,exam, homework assignments • Homepage, contact details

1. Why are you here?

Introduction: Media Buzz

Data Scientists are in high demand

Harvard Business Review

SEARCH

THE MAGAZINE | BLOGS | VIDEO | BOOKS | CASES | WEBINARS | COUR

Guest — Subscribe today and get access to all current articles and HBR online archive.

THE MAGAZINE

October 2012

ARTICLE PREVIEW To read the full article, sign-in or register. HBR subscribers, click here to register for FREE access »

Data Scientist: The Sexiest Job of the 21st Century

by Thomas H. Davenport and D.J. Patil

CNBC

Enter Symbols | GO | Enter Keywords | GO

HOME U.S. | NEWS | MARKETS | INVESTING | TECH | SMALL BIZ | VIDEO | SHOWS | PRIM

NEW SHOW SQUAWKalley — The Intersectio Wall St. & Techr

BIG DATA | A CNBC SPECIAL REPORT

Why your kids will want to be data scientists

John Phillips | @J_Phillips_IV

Tuesday, 3 Jun 2014 | 7:05 PM ET

CNBC

Forbes

New Posts | Most Popular | Lists | Video | 2 Free issues of Forbes

COVER STORY The World's Richest Doctor

Cash Kings 2014: The World's Highest-Paid Hip-Hop Acts

Shanghai Scores As Top New Tech Hub In The World As Silicon Valley Gap Grows

Chart: T Of Peop To Upg

Log in | Sign up | Connect | Privacy

פלטפורמת המסחר הטובה ביותר לפי דירוג 2014 — Interactive Brokers

EMC²

ForbesBrandVoice

The Hottest Jobs In IT: Training Tomorrow's Data Scientists

TechRepublic / U.S.

All Topics | Newsletters | Photos | Forums | Resource Library | Rese

CXO | Software | Startups | Cloud | Data Center | Mobile | Microsoft | Apple | Google | Search TechRepul

SAP — Is your business making the most out of today's technologies?

BIG DATA

Big data skills: Should data scientist be your next job?

Also in Academia

WHITE HOUSE TO UNIVERSITIES: WE NEED MORE DATA SCIENTISTS

NEW YORK UNIVERSITY, UNIVERSITY OF CALIFORNIA-BERKELEY, AND THE UNIVERSITY OF WASHINGTON ARE LAUNCHING A $37.8 MILLION PROJECT TO BOOST THE NUMBERS OF AMERICAN DATA SCIENTISTS.

BY NEAL UNGERLEIDER

It's official: America needs more data scientists. This week, a $37.8 million project

Berkeley Research
UNIVERSITY OF CALIFORNIA

CONTACT US | HOM

RESEARCH HIGHLIGHTS | NEWS | ABOUT US | RESEARCH UNITS | FACULTY EXPERTISE | RESEARCH POLICIES & ADMINISTRATION | TECH TRANSFER | FUND YOUR RESEARCH

HOME ▸ DATA SCIENCE

Data Science

DATA SCIENCE
OVERVIEW
INSTITUTE FOR DATA SCIENCE ▸
News Release
Past Events
PEOPLE
CAREER OPPORTUNITIES
2013-14 LECTURE SERIES
CAMPUS EVENTS ▸
Archive
NEWS
INSTITUTES AND PROGRAMS ▸

Data Science at UC Ber

SCIENTIFIC AMERICAN™

Sign In | Register 0

Search ScientificAmerican.com

Subscribe | News & Features | Topics | Blogs | Videos & Podcasts | Education | C

More Science » Scientific American Volume 309, Issue 4 — 4 :: Email :: Print

How Big Data Can Transform Society for the Better

The digital traces we leave behind each day reveal more about us than we know. This could become a privacy nightmare—or it could be the foundation of a healthier, more prosperous world

By Alex Pentland

NYU — DATA SCIENCE AT NYU

About | What is data science? | Research | Academics | News | Contact Us

Research

RESEARCH CENTERS IN THE FIELD OF DATA SCIENCE

Center for Data Science (CDS)

The NYU Center for Data Science (CDS) is a focal point for New York University's university-wide initiative in data science. It was established to help advance NYU's goal of creating the country's leading data science training and research facilities, arming researchers and professionals with tools to harness the power of big data.

LEARN MORE

Center for the Promotion of Research Involving Innovative Statistical Methodology (PRIISM)

The Center for the Promotion of Research Involving Innovative Statistical Methodology (PRIISM) is a new center dedicated to improving the caliber of research in quantitative social, educational, behavioral, allied health and policy science.

500k

The world's 500,000+ data centres are large enough to fill 5,955 football fields. *(Source: Kurtosys)*

75%

75% of digital information is generated by individuals, whilst enterprises have liability for 80% of digital data at some point in its life. *(Source: Kurtosys)*

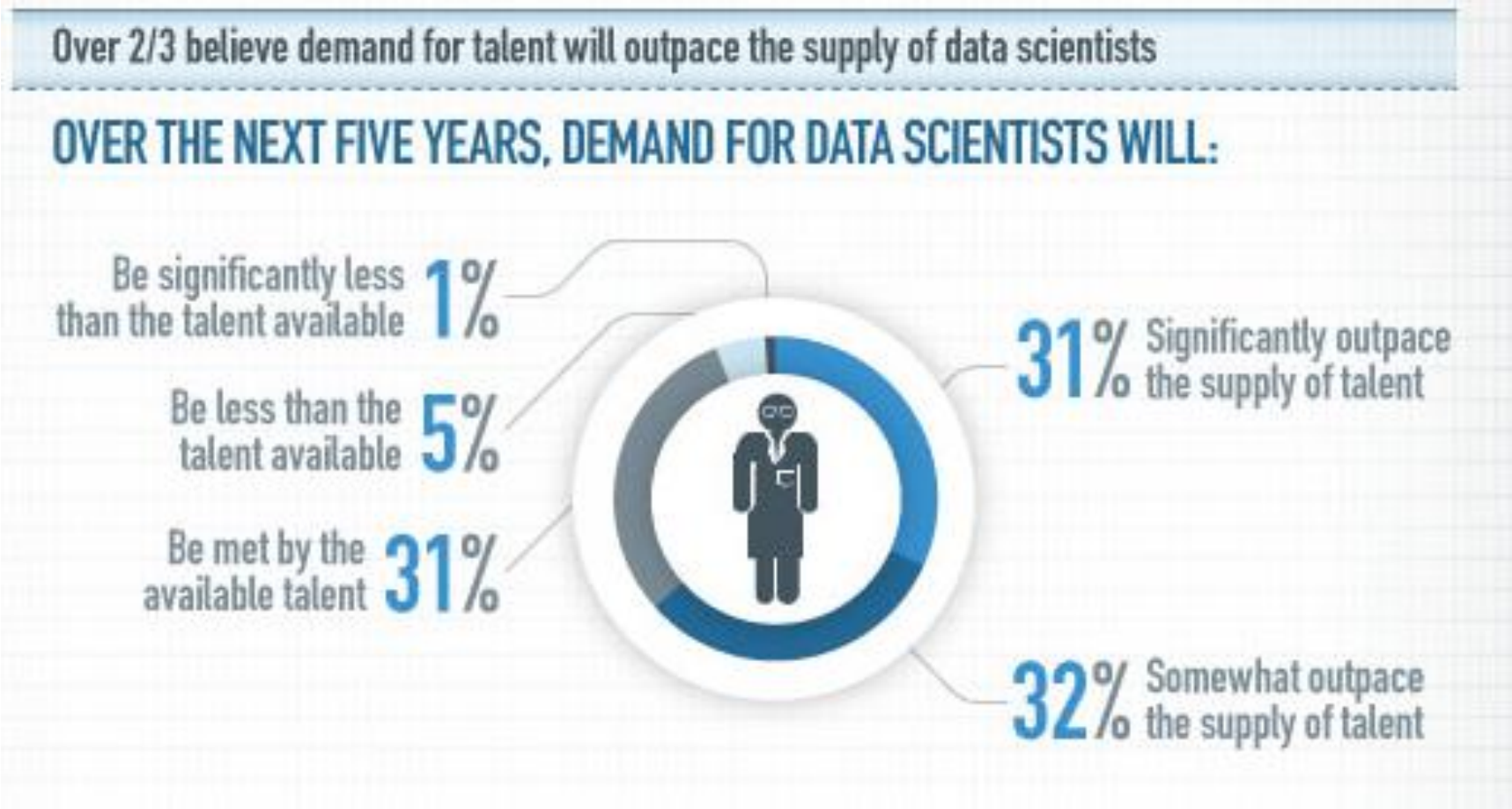
UNIVERSITY of WASHINGTON

eScience Institute
Supporting Data-Driven Discovery In All Fields

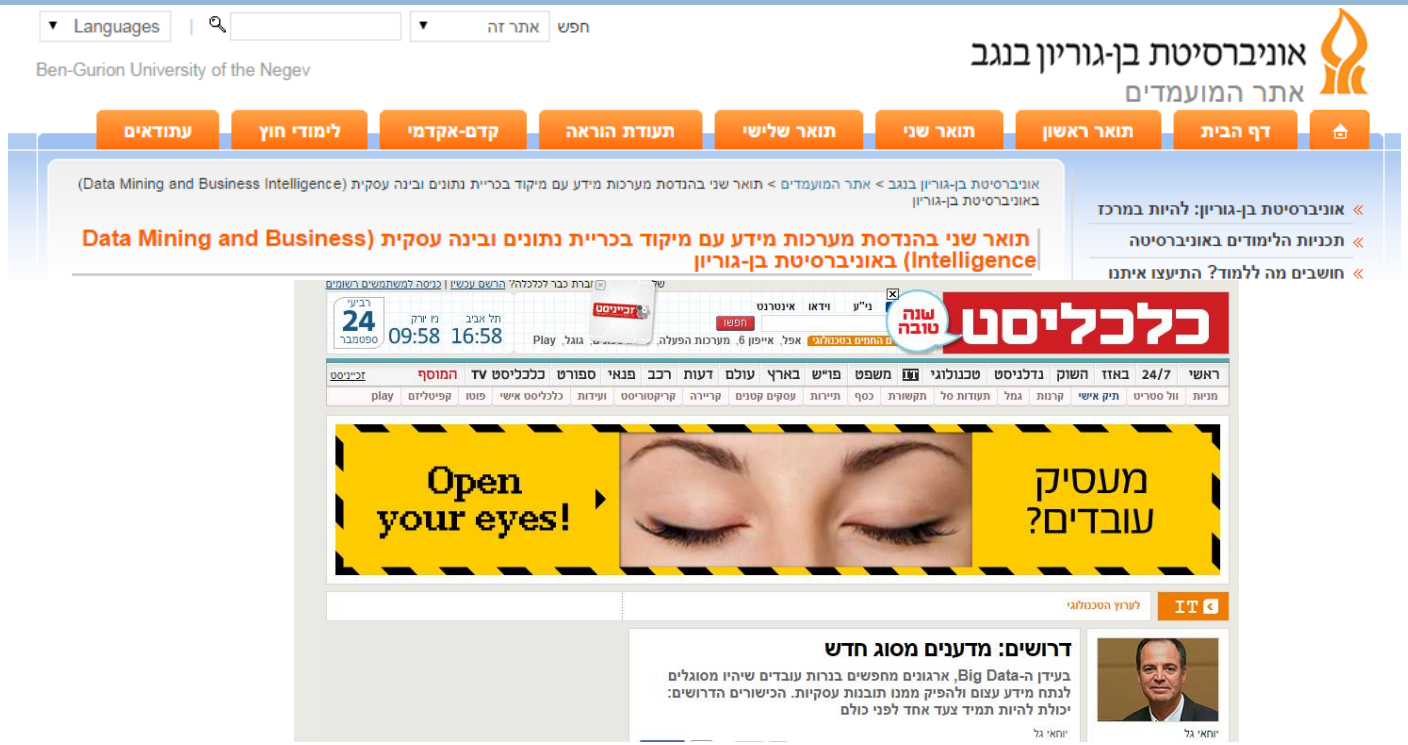
WHO WE ARE

New Ph.D. Tracks in "Big Data"

Demand will outpace the supply

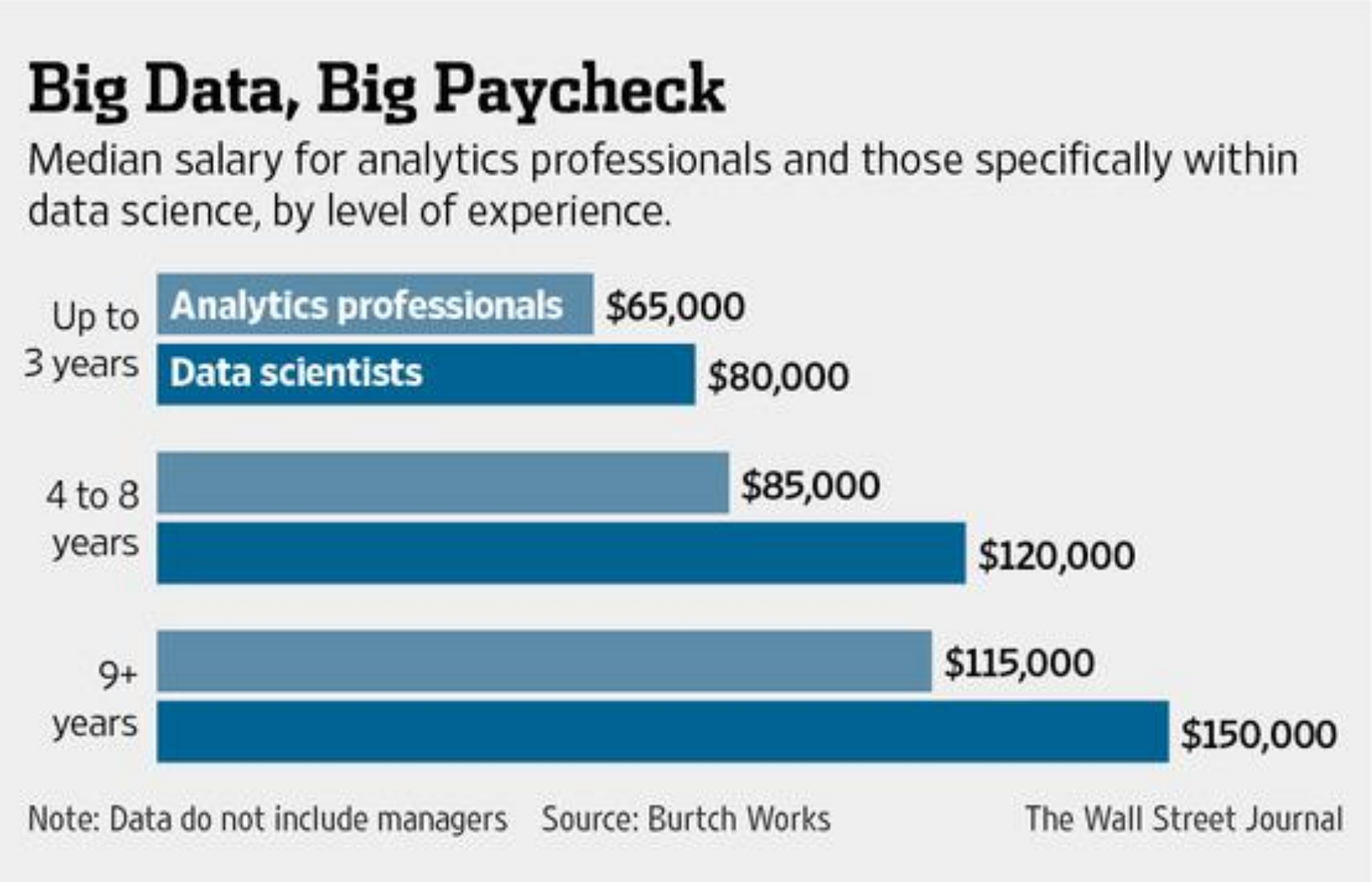


Israel


Languages
חפש
אתר זה
Ben-Gurion University of the Negev
אוניברסיטת בן-גוריון בנגב
אתר המועמדים
עתודאים
לימודי חוץ
קדם-אקדמי
תעודת הוראה
תואר שלישי
תואר שני
תואר ראשון
דף הבית
אוניברסיטת בן-גוריון: להיות במרכז
תכניות הלימודים באוניברסיטה
חושבים מה ללמוד? התיעצו איתנו
תואר שני בהנדסת מערכות מידע עם מיקוד בכריית נתונים ובינה עסקית (Data Mining and Business Intelligence) באוניברסיטת בן-גוריון
כלכליסט
שנה טובה
24
09:58
16:58
Open your eyes!
מעסיק עובדים?
דרושים: מדענים מסוג חדש
בעידן ה-Big Data, ארגונים מחפשים בנרות עובדים שיהיו מסוגלים לנתח מידע עצום ולהפיק ממנו תובנות עסקיות. הכישורים הדרושים: יכולת להיות תמיד צעד אחד לפני כולם


English
מפת אתר
באתר
חיפוש
חברת טכנולוגיית המידע המובילה בישראל
matrix
מאחלת שנה טובה
צור קשר
לעבוד במטריקס
מגזרים ולקוחות
מרכזי התמחות
שירותי רוחב
אודות
דרושים
דף הבית
מטריקס > חדשות >
Matrix BI הקימה תחום חדש של Data Science ומינתה את איתי בליצר למנהל ולמוביל התחום
Matrix BI הקימה תחום חדש של Data Science ומינתה את

Pays well


Big Data, Big Paycheck
Median salary for analytics professionals and those specifically within data science, by level of experience.
Up to 3 years
Analytics professionals
$65,000
Data scientists
$80,000
4 to 8 years
$85,000
$120,000
9+ years
$115,000
$150,000
Note: Data do not include managers
Source: Burtch Works
The Wall Street Journal

2. What is data science?

Technology and raising expectations

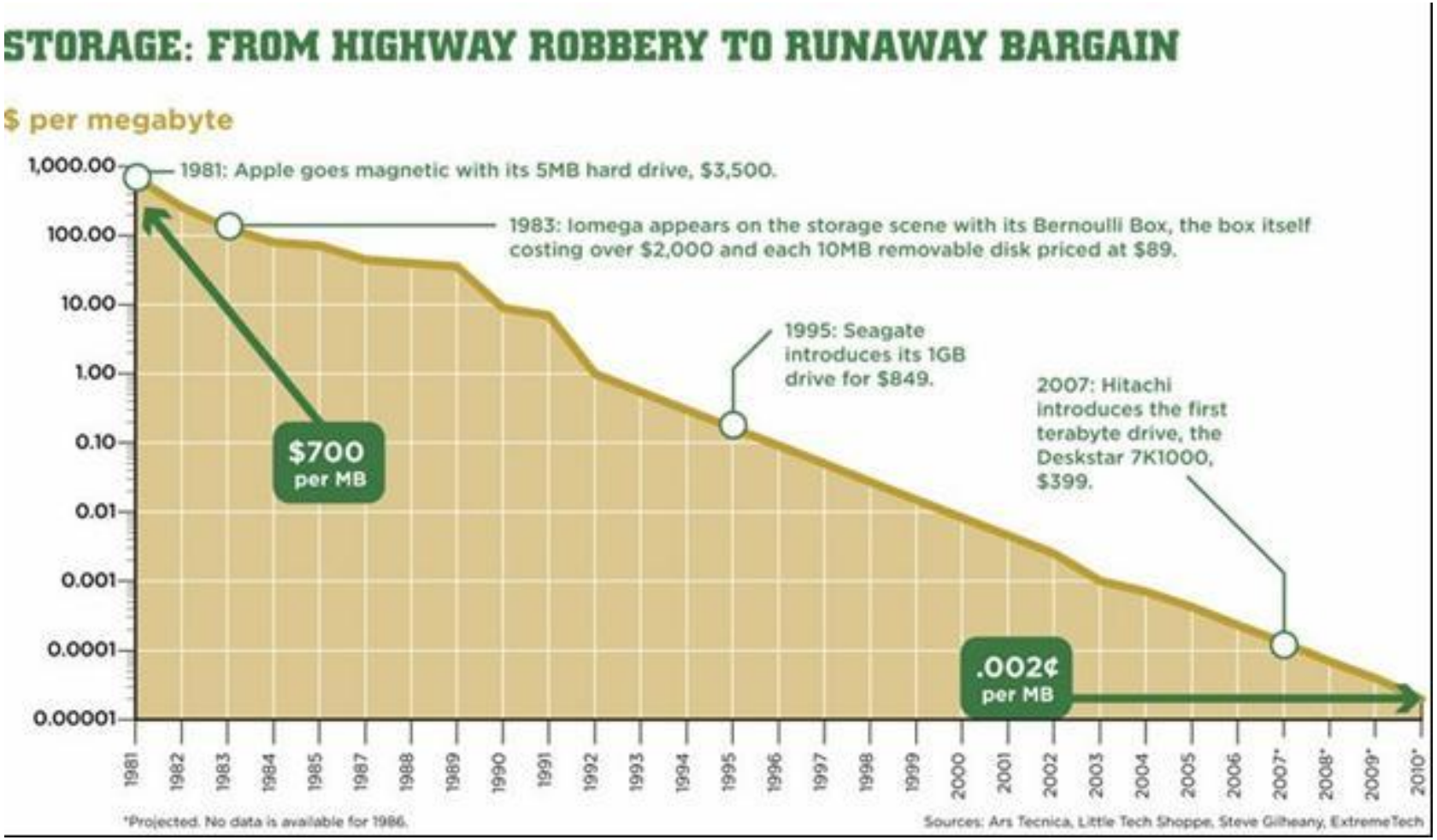
Data Science

- New Discipline
- Very little/none textbooks/courses covering the discipline as a whole
  - Compare to Software Engineering/Compute Science during 70-80^{s} of the last century
  - Data Science is what data scientists do
- Why data science and data scientists are needed?
  - Development of enabling technology
  - Raising Expectations from customers

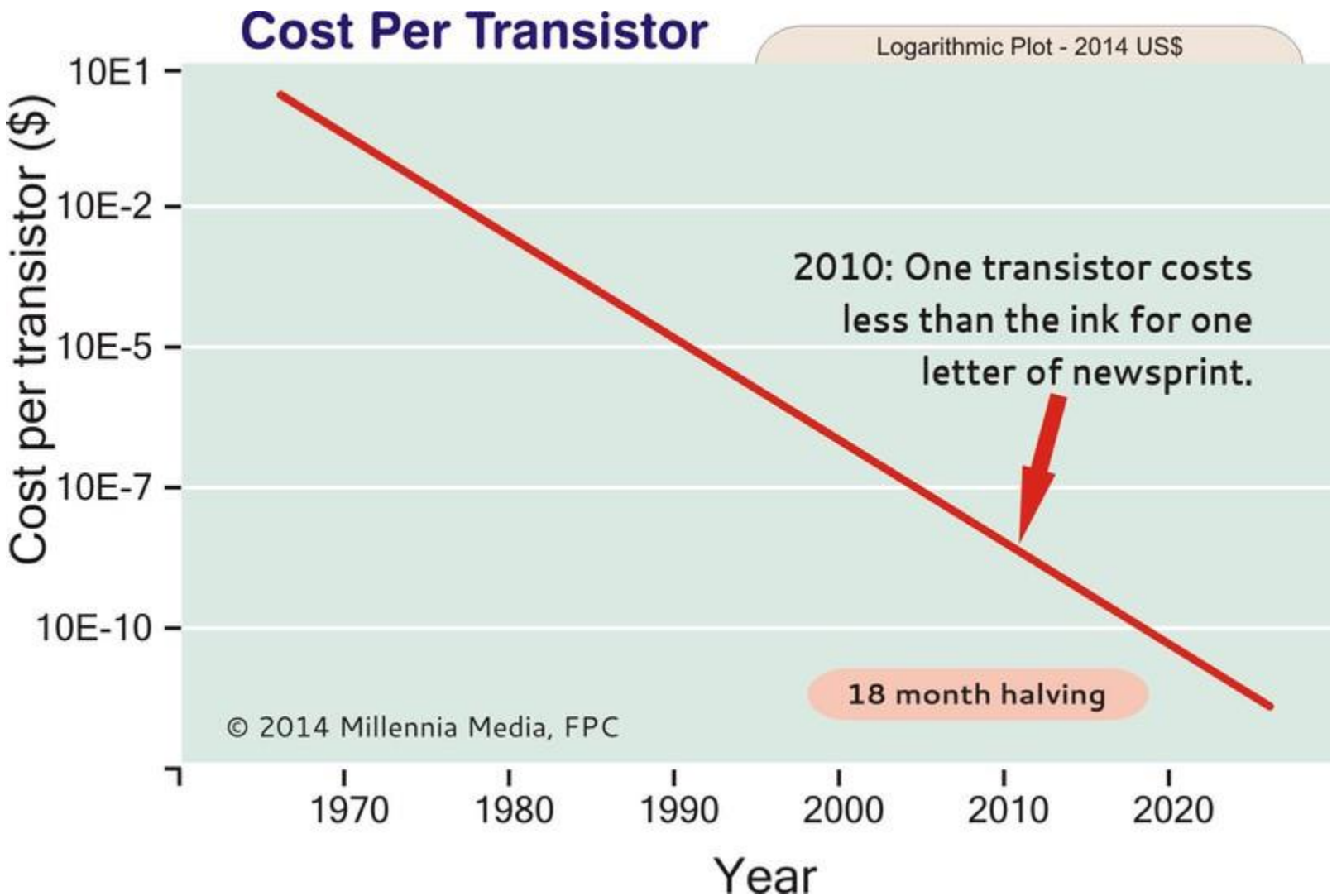
2. What is data science?

Technological developments

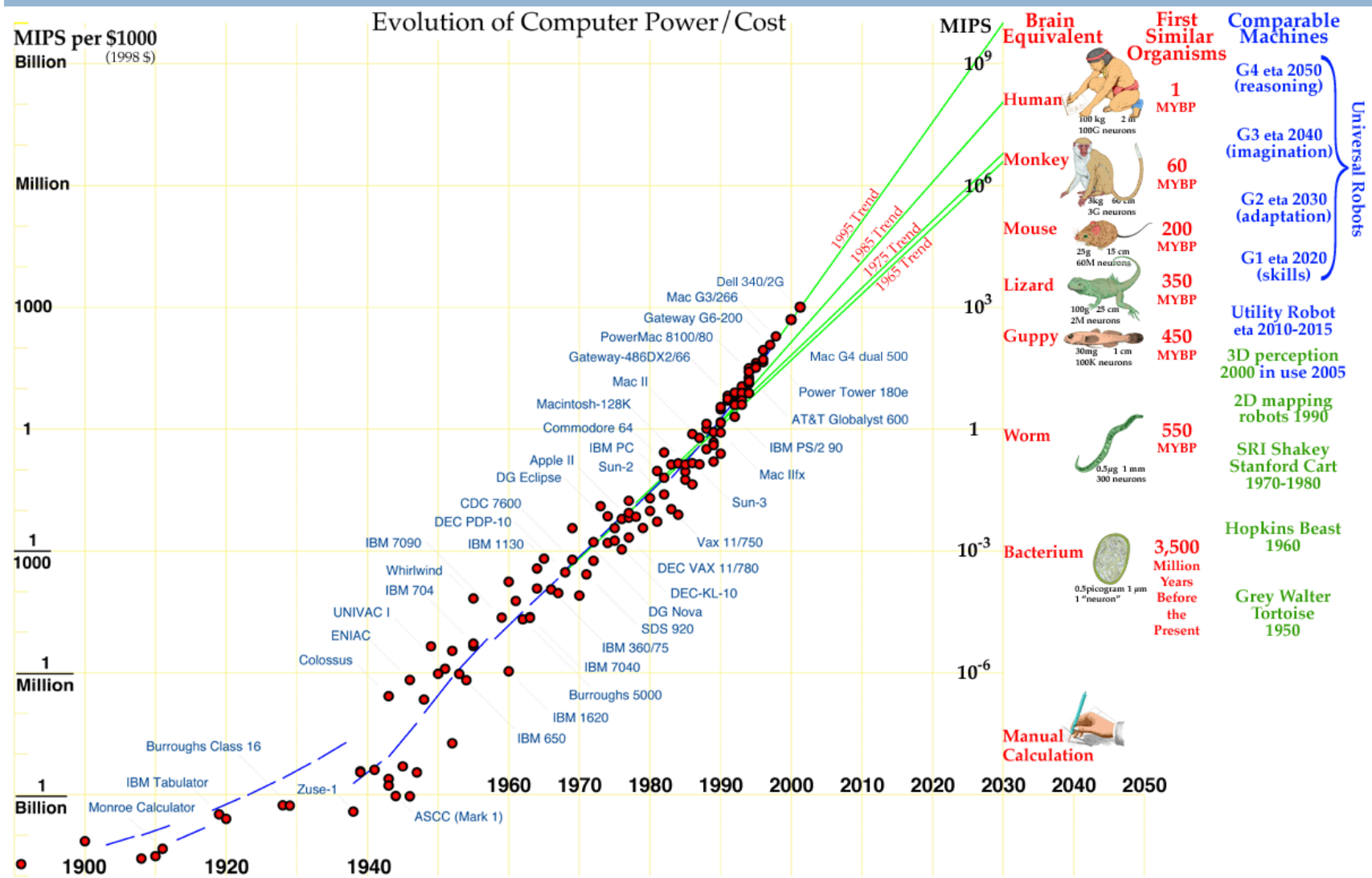
Declining cost of storage



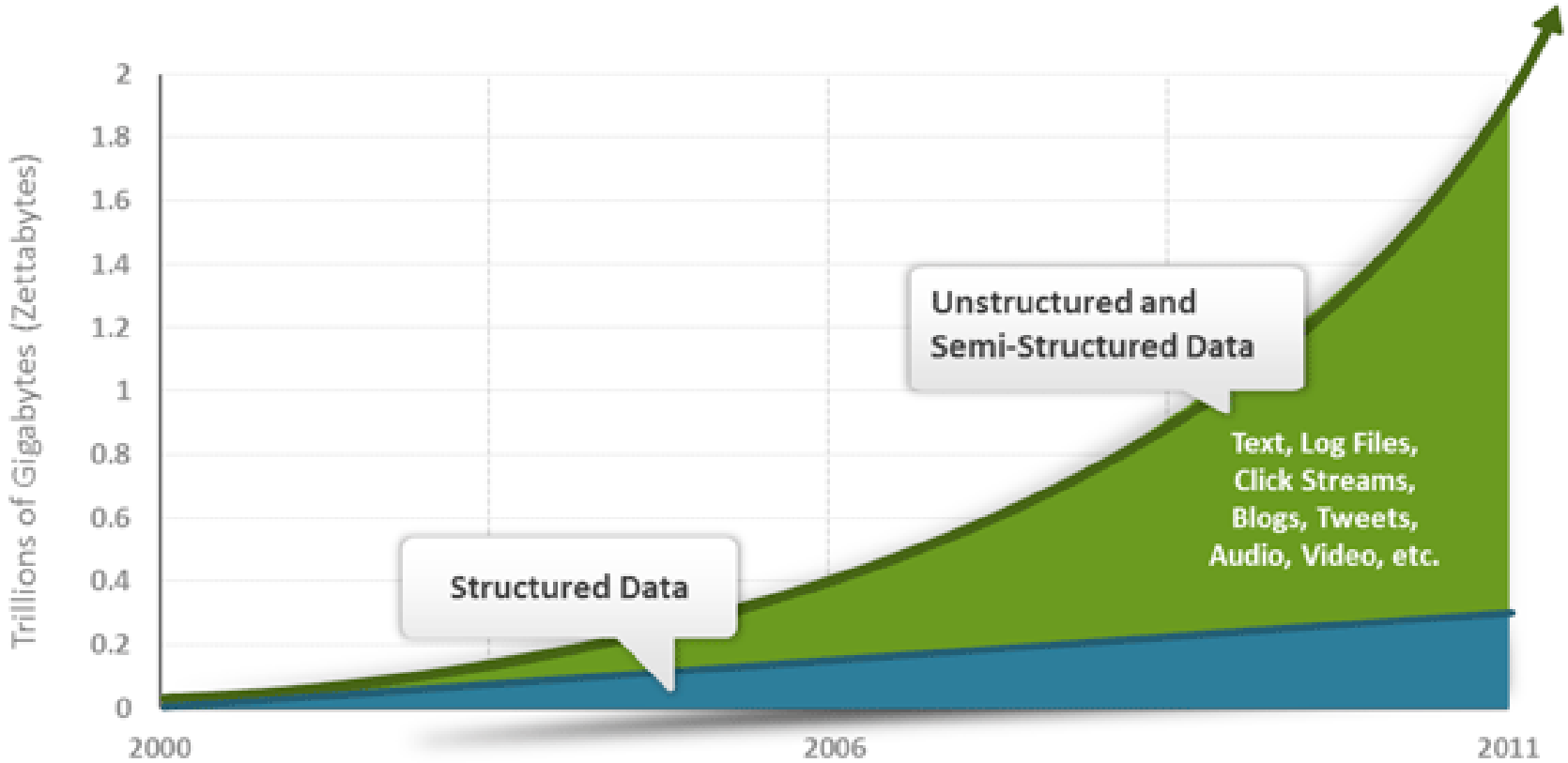
Declining cost of computing



Surpassing the brain

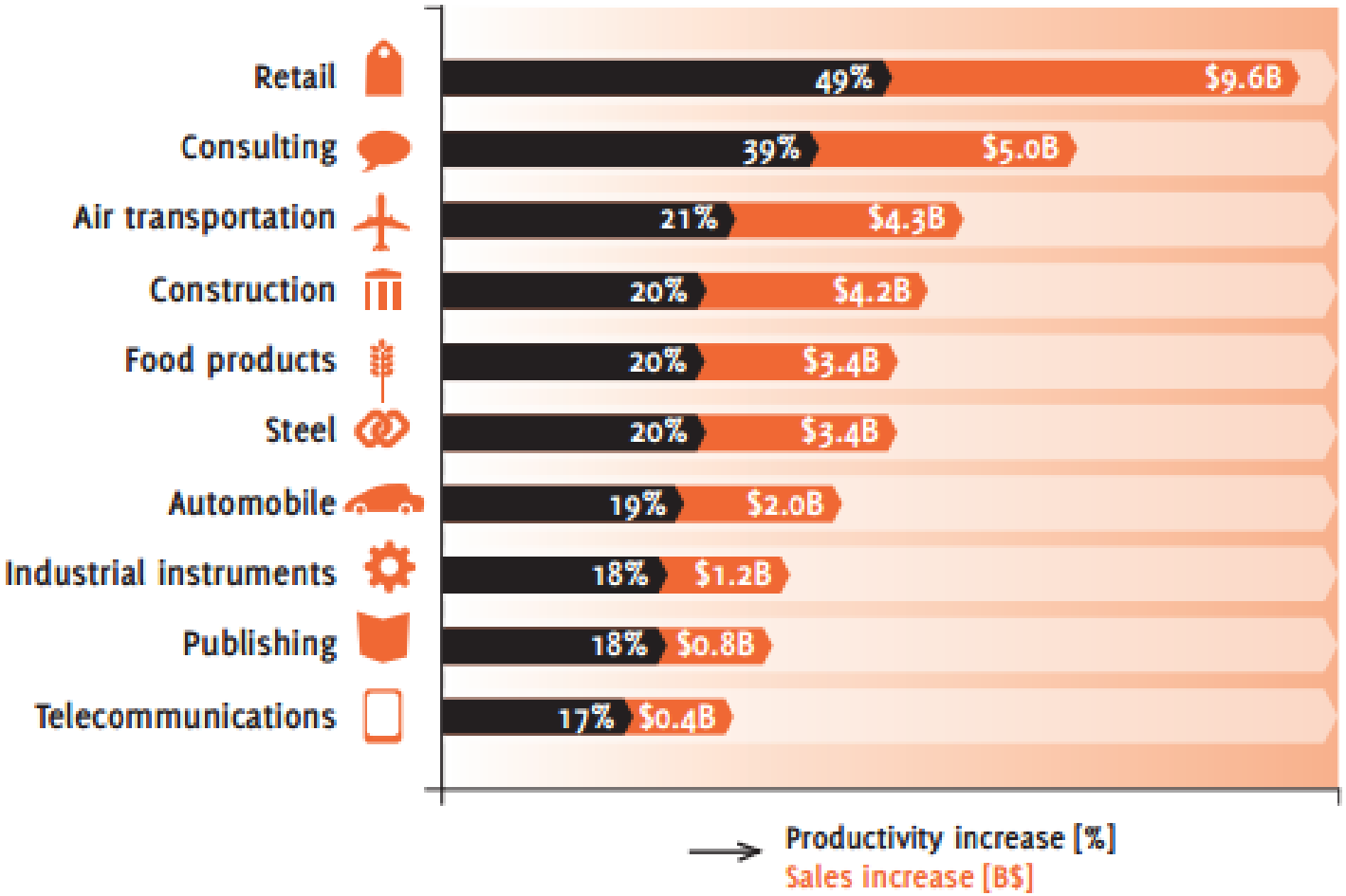


More data can be stored and processed



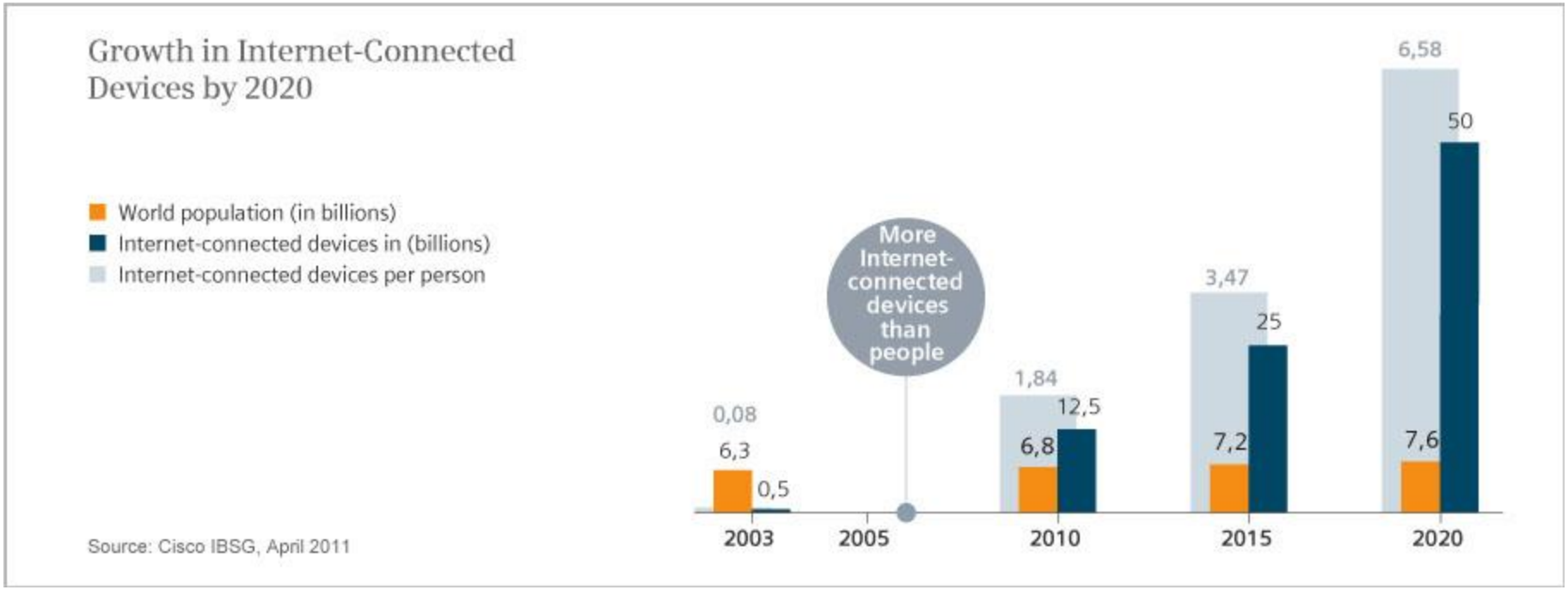
Source: IDC 2011 Digital Universe Study (http://www.emc.com/collateral/demos/microsites/emc-digital-universe-2011/index.htm)

Value of Big Data

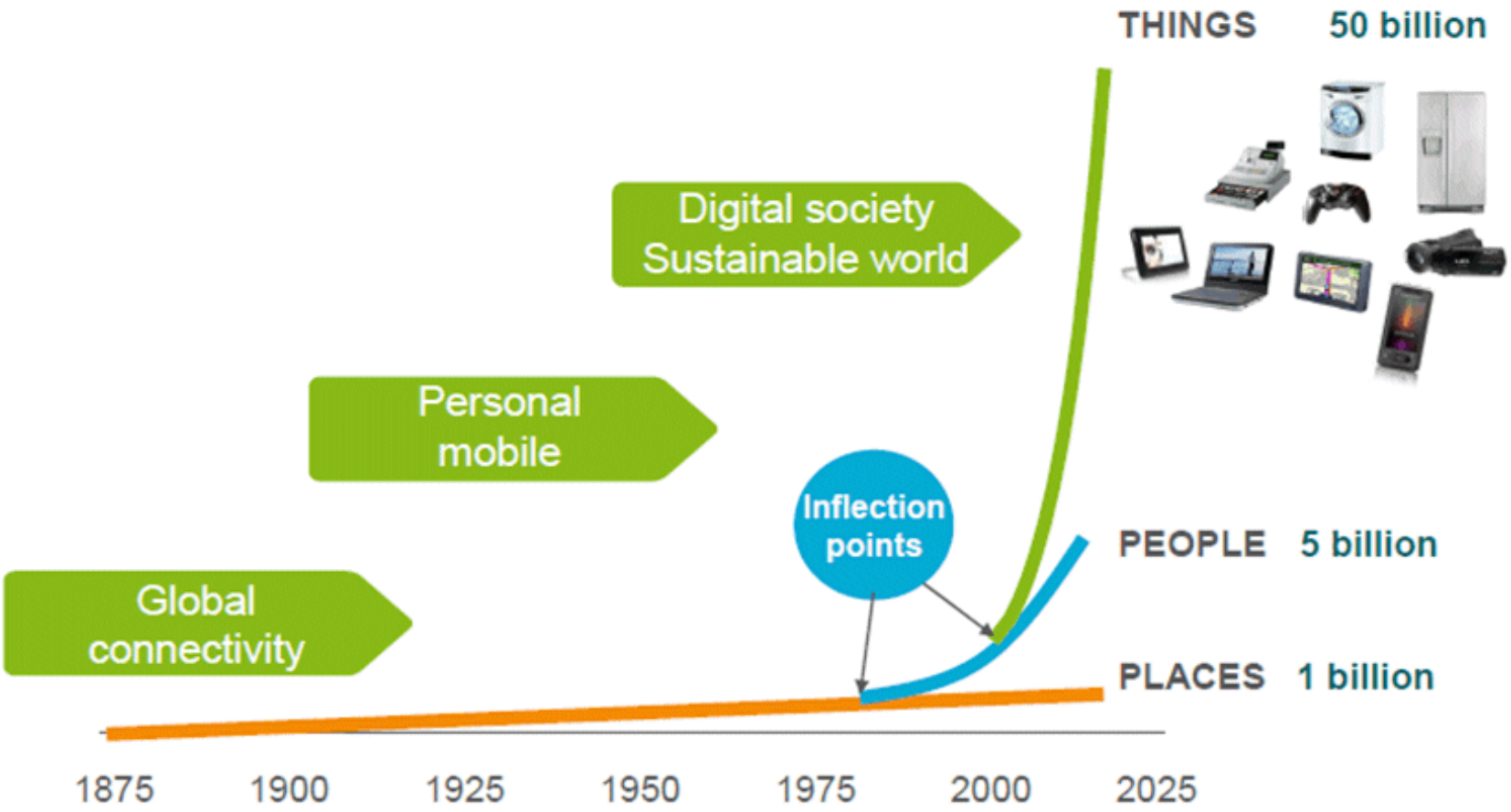


Source: University of Texas (2011)

Devices vs. People

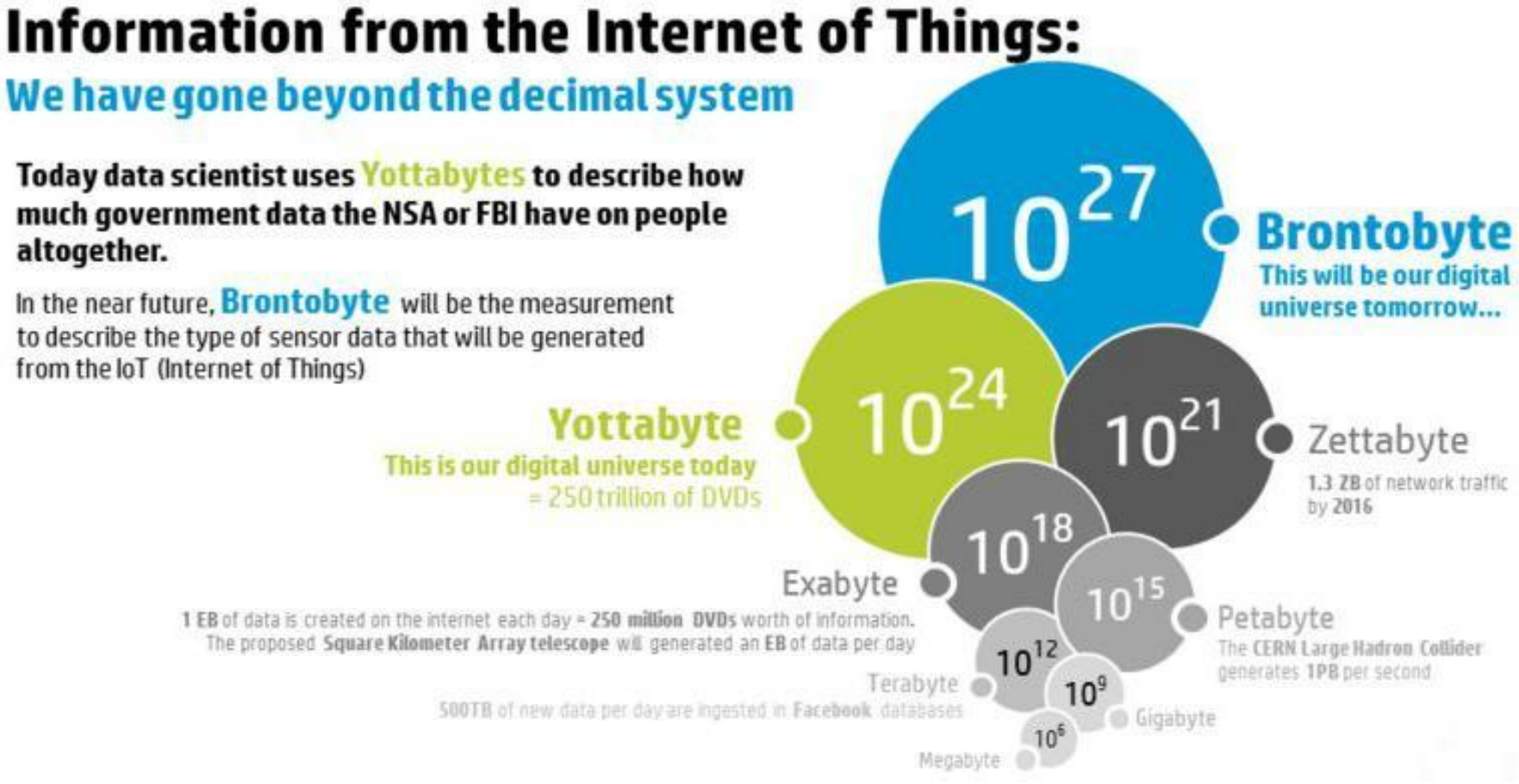


Internet of Things



Source: Ericsson AB, "Infrastructure Innovation - Can the Challenge be met?," Sept 2010

Next frontier: IoT



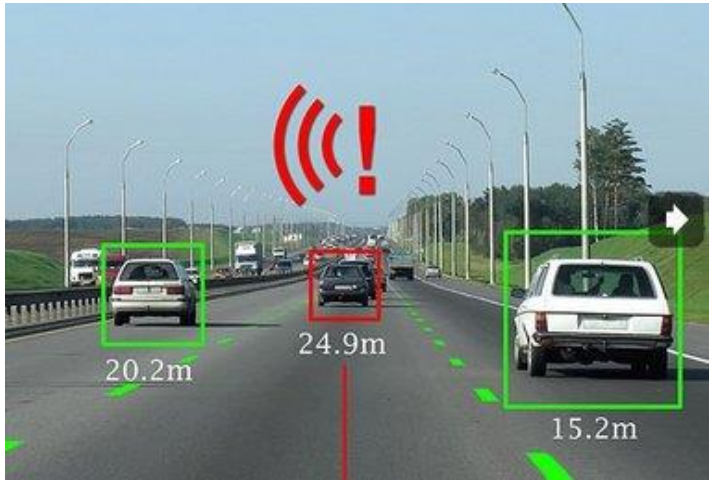
2. What is data science?

Raising expectations

Cognitive Computing

- People expect systems to behave like humans
  - Be Adaptive
    - Learn as information and goals change
  - Be Interactive
    - Interact easily with people and other systems
  - Be Contextual
    - Understand meaning, exploit additional sources of information
- Need to process large quantities of uncertain data of different types (text, speech, sensors, images etc.)

Cognitive Computing



Cognitive Computing in 5 Years

Mashable MUST READS SOCIAL MEDIA TECH BUSINESS ENTERTAINMENT US & WORLD WATERCOOLER

Tech

AdChoices ▶ Smell Taste ▶ IBM Computers ▶ Taste Buds ▶ Smell Machine

IBM: Computers Will See, Hear, Taste, Smell and Touch in 5 Years

Cognitive and Data Science

- People want their systems/devices to behave smarter
  - Personal devices
  - Industrial systems

- More data to acquire and analyze using more complex algorithms and technologies

3. What is data science

Some examples

Example I: Marketing

- Predicting Lifetime Value (LTV)
  - **what for:** if you can predict the characteristics of high LTV customers, this supports customer segmentation, identifies upsell opportunities and supports other marketing initiatives
  - **usage:** can be both an online algorithm and a static report showing the characteristics of high LTV customers

Example II: Logistics

- Demand forecasting
  - How many of what thing do you need and where will we need them? (Enables lean inventory and prevents out of stock situations.)
  - revenue impact: supports growth and militates against revenue leakage
  - usage: online algorithm and static report

Example III: Healthcare

- Survival analysis
  - Analyze survival statistics for different patient attributes (age, blood type, gender, etc) and treatments
- Medication (dosage) effectiveness
  - Analyze effects of admitting different types and dosage of medication for a disease
- Readmission risk
  - Predict risk of re-admittance based on patient attributes, medical history, diagnose & treatment

Example IV: Wearable Health and Fitness

Example V: Brain Computer Interface

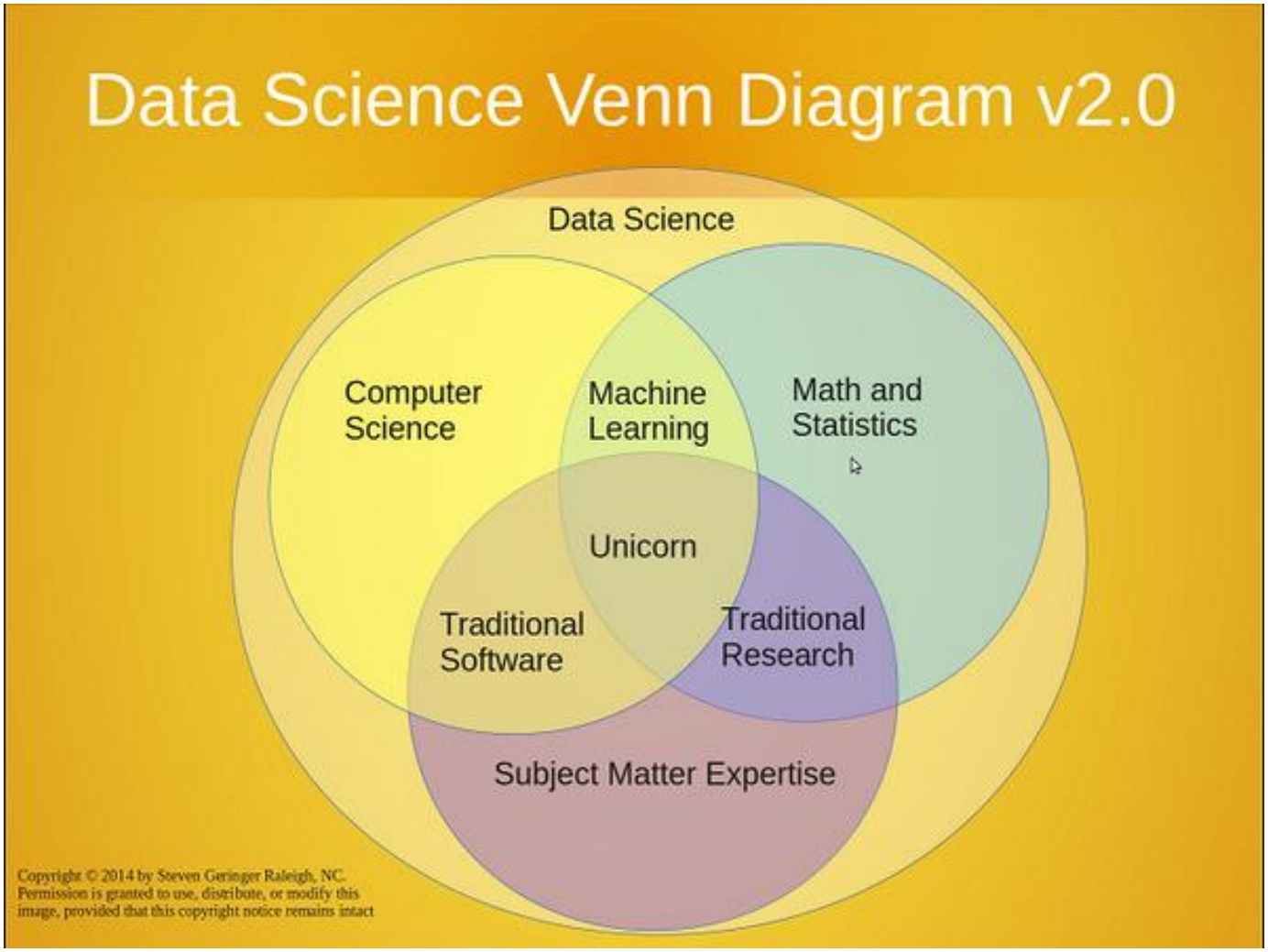
2. What is data science?

A Mashup of disciplines

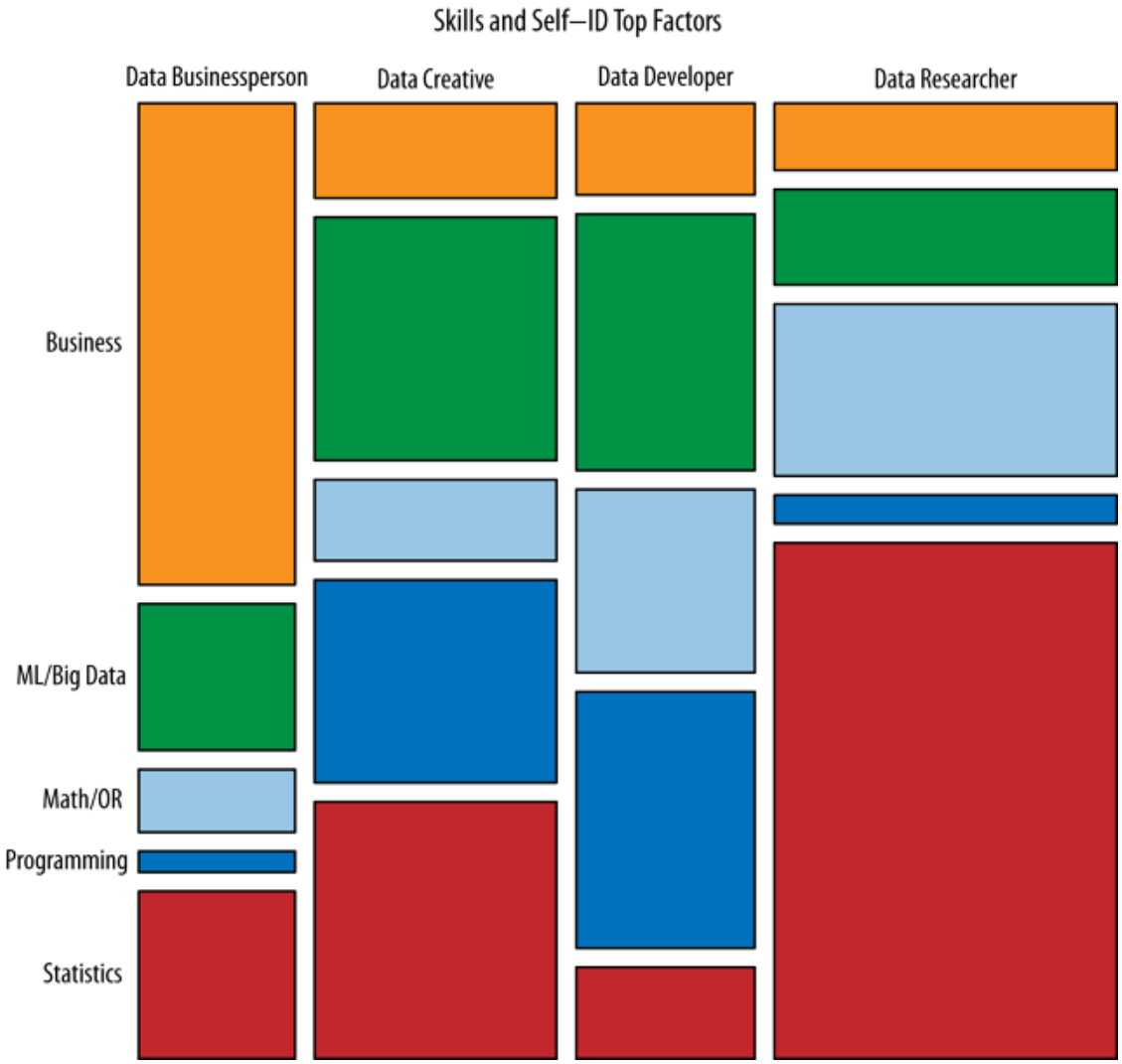
A mashup of disciplines

Math and Theory	• Statistics, Linear Algebra, Optimization, Time Series, etc.
Applied Algorithms	• Machine Learning, Data Structures, Parallel Algorithms, etc.
Engineering and Technologies	• Storage and computing platforms, statistical tools ,etc.
Domain Expertise	• Text, Finance, Images, Econometrics etc.
Art	• Visualization, Infographics
Best practices and hacks	• Handle missed values in data, transform and represent data, etc.

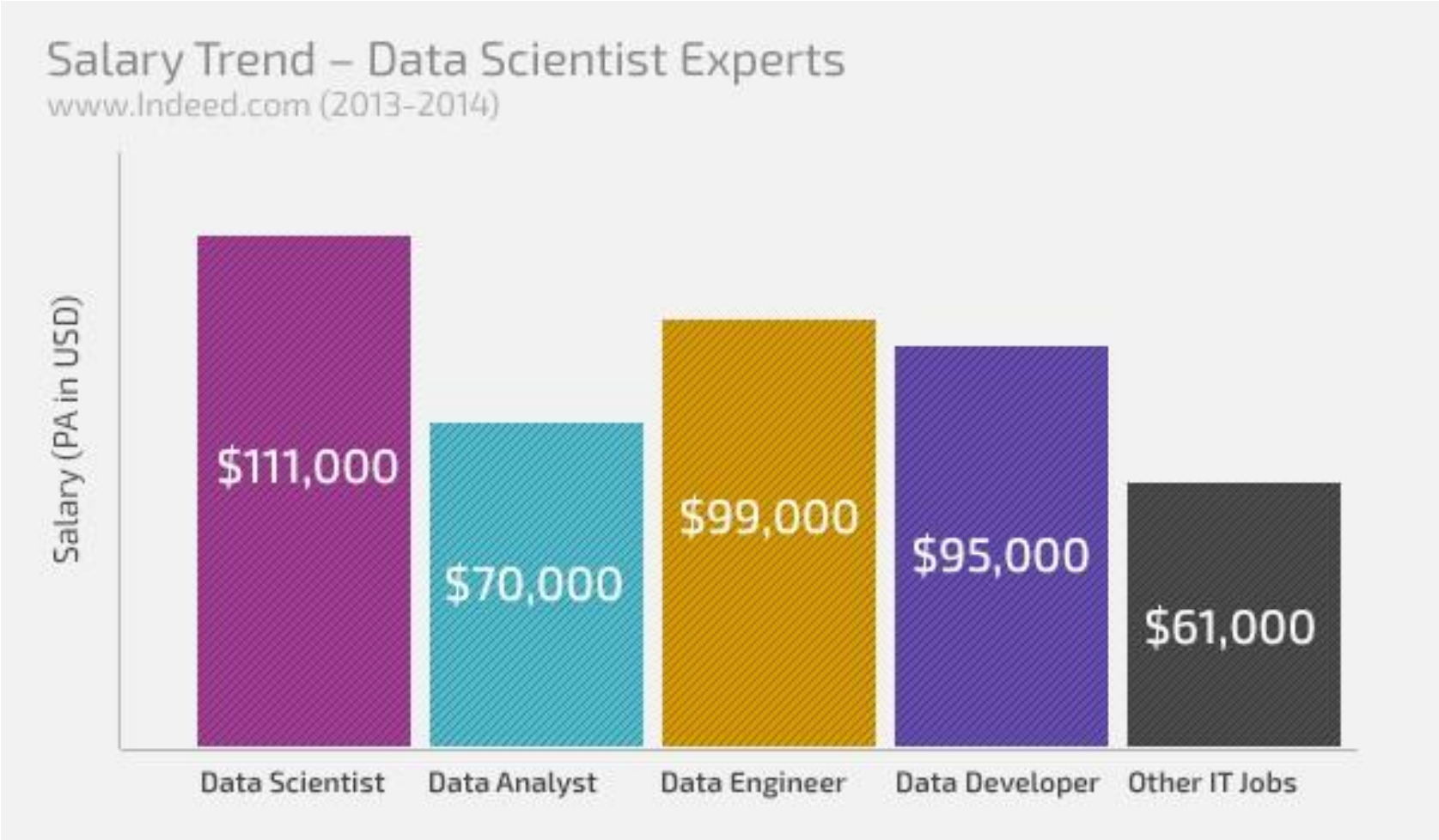
Yet Another View



Types of Data Scientists



Roles and Paycheck



3. About this course

A mix of theory and practice

General

- Introductory course
  - But for advanced undergrads
- Broad overview of subjects
  - But deep enough to have an exam
- Focus on practical aspects
  - But not on ever-changing technology and tools

Tentative content(subject to change)

- 70% Statistical Machine Learning (7 weeks)
  - Focus on practical aspects
  - Classes
    - Necessary theoretical background
    - Basic R programming lab
- 20% Big Data Algorithms (2 weeks)
  - Focus on algorithms not on big data technologies
- 10% Data Visualization (1 weeks)
  - Grammar of graphics in R

This course is **not**

- About big data tools or technologies
  - No: Hadoop technical details
  - Yes: Basic R programming
- About statistical learning theory
  - No: Theoretical low bounds or other proofs
  - Yes: Some theory is necessary
- About a specific domain
  - No: Deep discussions on Text, Finance, BI etc.
  - Yes: Some examples will be presented

Some case studies we will cover

PREDICTION OF FUTURE MOVEMENTS IN THE STOCK MARKET:	• What is the next move of S&P 500?
PREDICTING INSURANCE PURCHASE	• Will a potential customer purchase?
DIRECT MARKETING	• Who will respond?
HOUSING VALUATIONS	• What affect the price of a house?
MARKETING OF ORANGE JUICE	• What brand a customer will buy?
EMAIL SPAM	• Is this a spam message?

The course's language of choice: R

AVERAGE SALARY FOR High Paying Skills and Experience

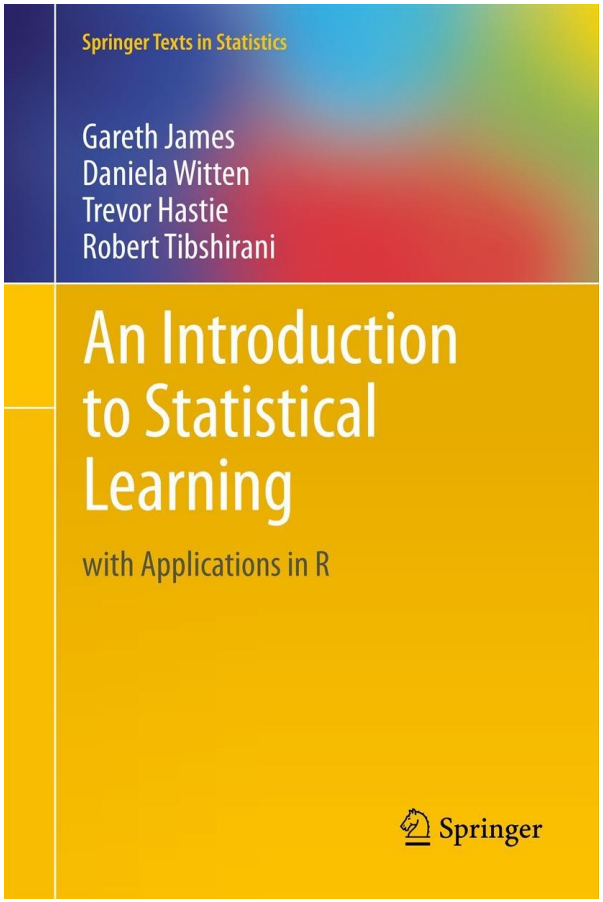
SKILL	2013	YR/YR CHANGE
R	$ 115,531	n/a
NoSQL	$ 114,796	1.6%
MapReduce	$ 114,396	n/a
PMBok	$ 112,382	1.3%
Cassandra	$ 112,382	n/a
Omnigraffle	$ 111,039	0.3%
Pig	$ 109,561	n/a
SOA (Service Oriented Architecture)	$ 108,997	-0.5%
Hadoop	$ 108,669	-5.6%
Mongo DB	$ 107,825	-0.4%

What you are expected to know

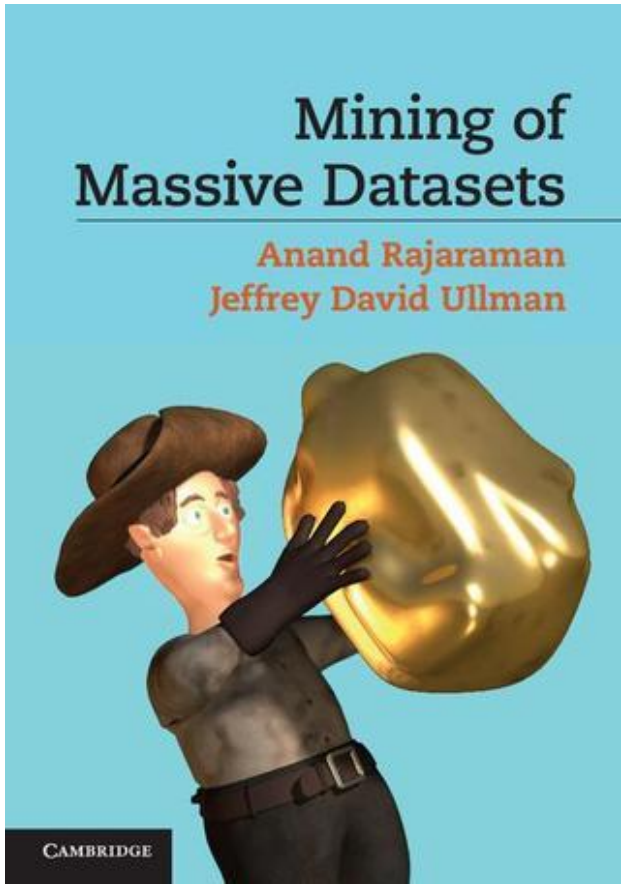
- Data is represented as a matrix
  - Basic linear algebra
- Most problems are not well-defined/uncertain
  - Basic probability and statistics
- Big data requires non-trivial data structures and algorithms
  - Basic data structures and algorithms concepts
- Practical means programming
  - Basic Programming skills

Textbooks are available online

Machine Learning and R

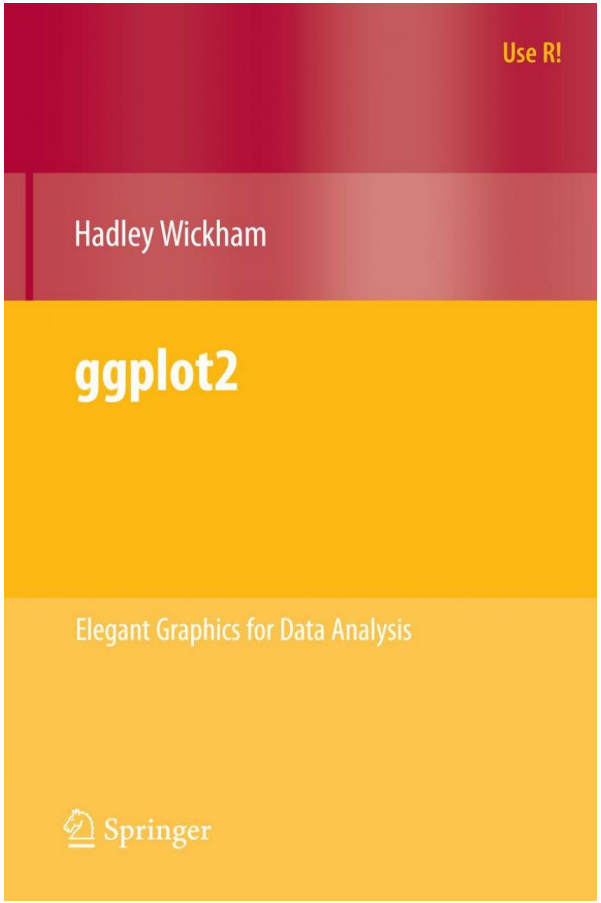


Big Data Algorithms

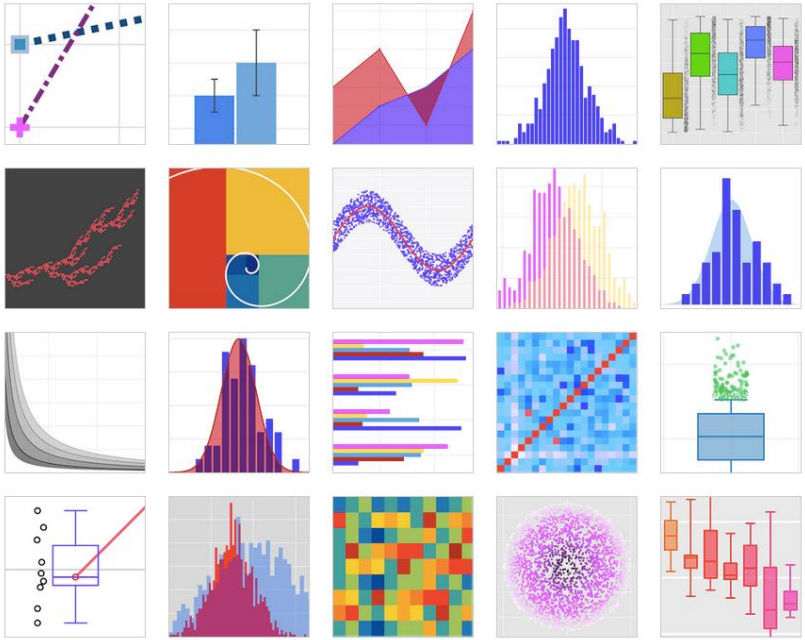


Visualization

Introduction from

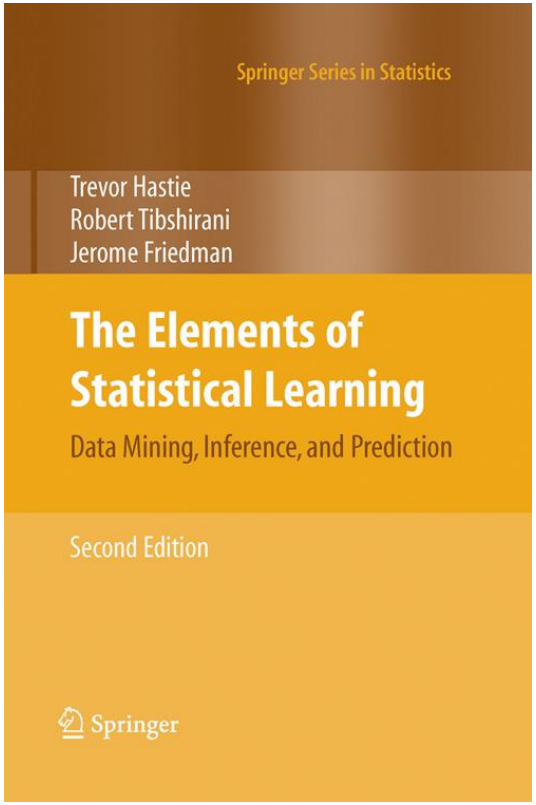


On-going examples

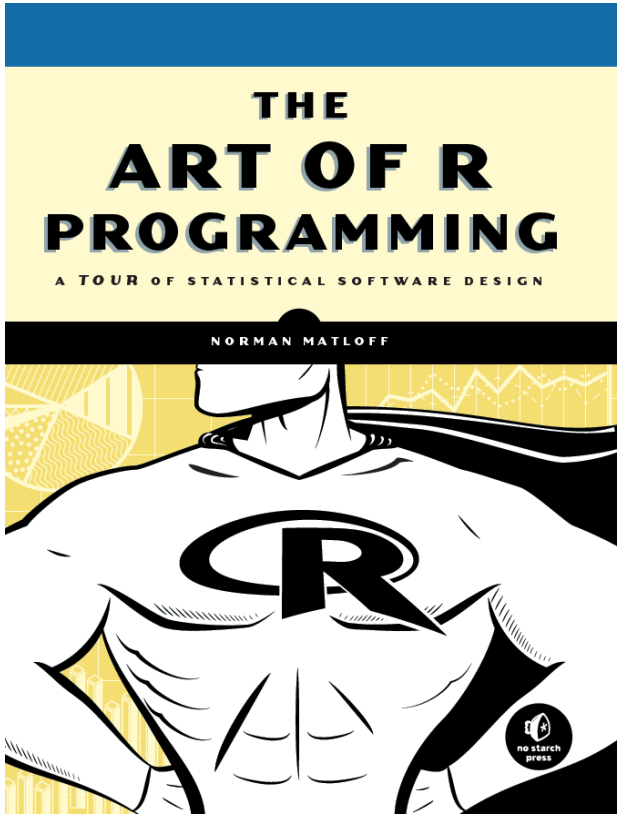


For curious minds

More on Machine Learning

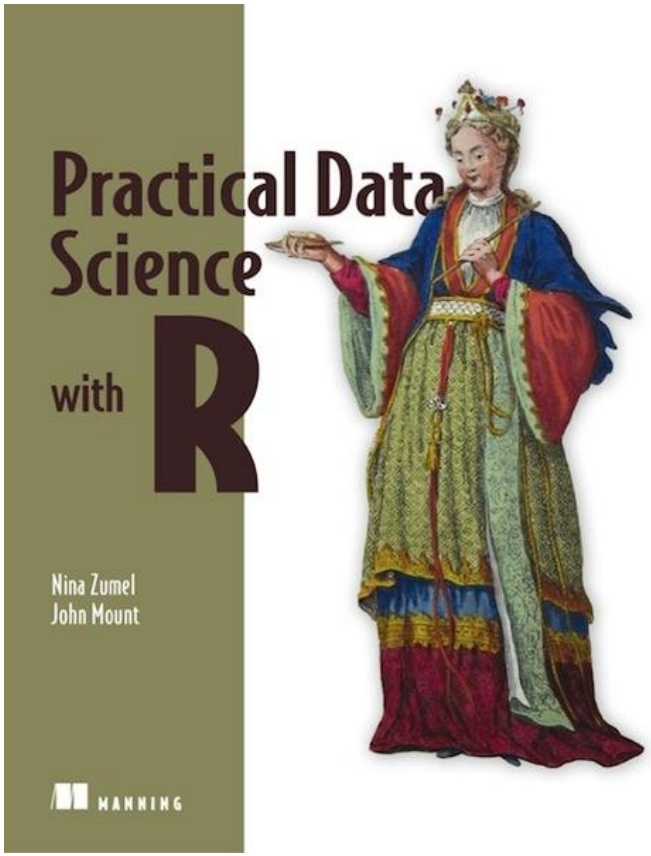


More on R Programming

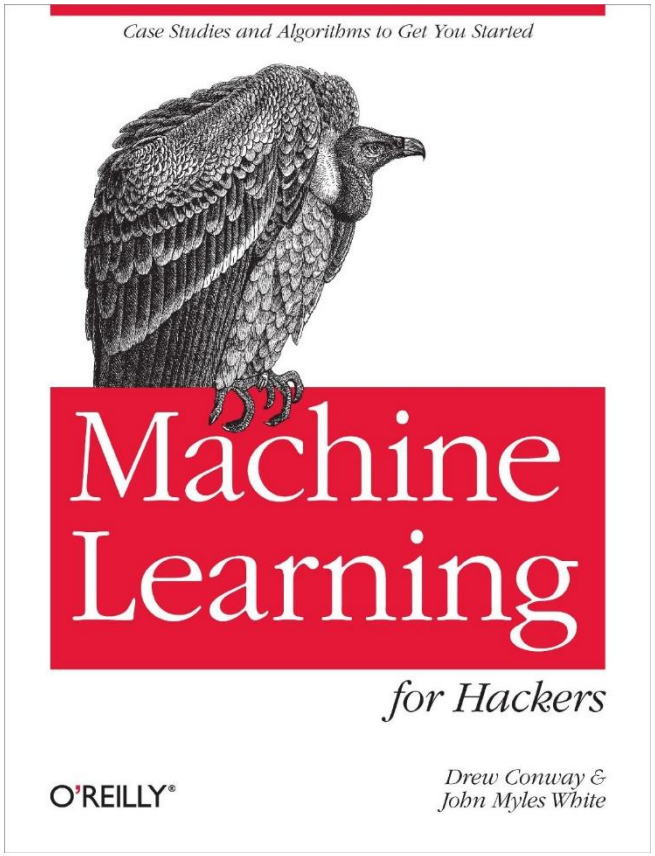


Becoming a data scientist

Data Scientist Skills



Quick Hacks/Examples



4. Course requirements

Requirements

- Grade
  - 100% closed material exam
- No previous year exams
  - Both textbooks have after chapter exercises
  - Exam questions (and HW assignments) will be very similar to these questions
- See course homepage for HW submission guielines

Contacts

- Lecturer: Dr. Sasha Apartsin (apartisn@gmail.com)
- Course homepage: http://www.cs.tau.ac.il/~apartzin/ds2015
- Office hours: By appointment
- Course forum :
  - **groups.google.com/d/forum/tau-data-science-course-2015s**

Plan

Why data science is important?	• "Why are you here"
What is data science?	• Mashup of disciplines
What this course is about?	• Hopefully right mix of theory and practical skills
Course requirements	• Syllabus • Grade ,exam, homework assignments • Homepage, contact details

Few More Disclaimers

Very inaccurate explanation

- **Statistics:** take a sample (data), answer questions about the **process** that produced this sample
  - Is it a normal distribution? **Estimate** it's mean.
- **Machine Learning:** take a sample(data), build a model to answer questions about **future samples**
  - Given a sample of named faces, **design a model** for naming a new unseen face.
- **Data Mining:** mine huge data store for interesting patterns or relationships
  - Given DB of transactions, **apply** tools and algorithms to find frequent product bundles
- **Data Science:** do whatever necessary to **extract value** from the data
  - Use data to improve book sales: mine patterns, engineer recommender systems, suggest improvements, estimate impact

No clear-cut boundaries!

Disclaimer: Math in the course

- All the computation are performed by computer
- You are in charge for interpretation of numbers
- So you'll have to understand the logic behind the number
  - You'll see significant amount formulas during the course
  - Mostly arithmetic, matrices and probability
- You are not expected to memorize or derive each formula (with exceptions), but you are expected to
  - Understand its meaning and use