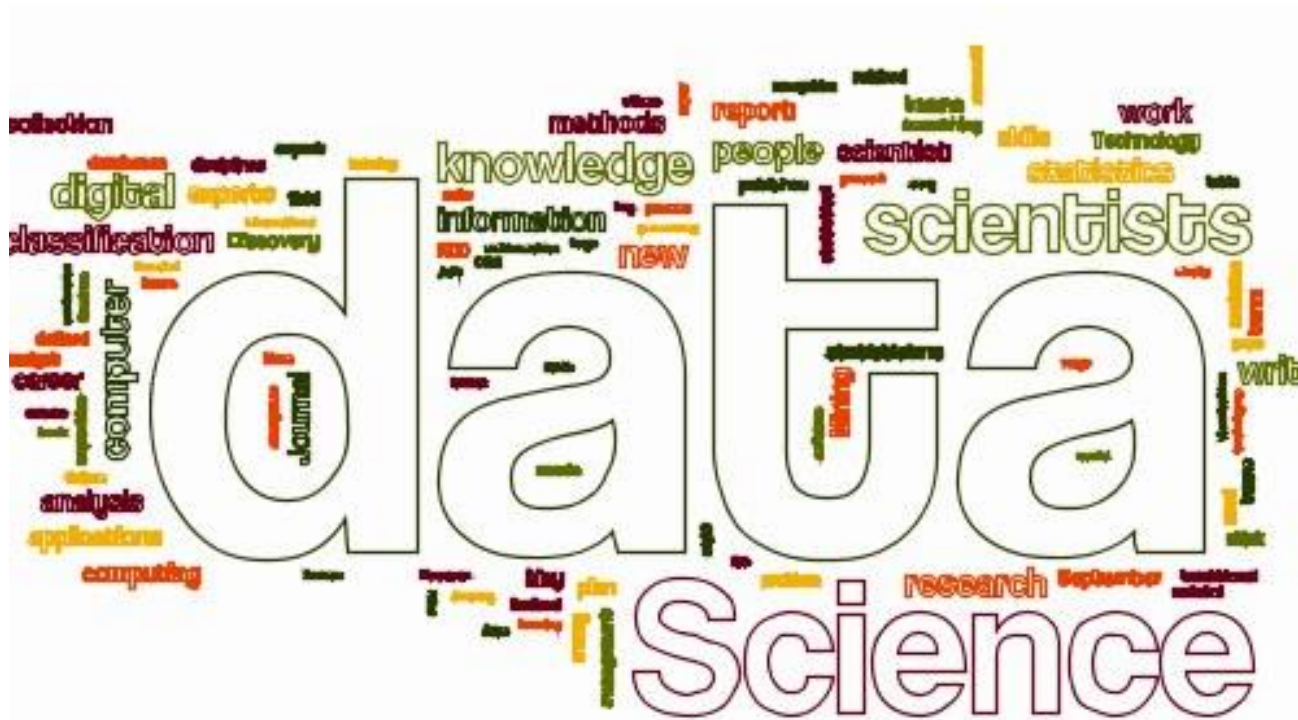


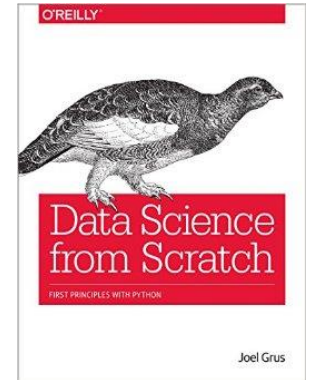
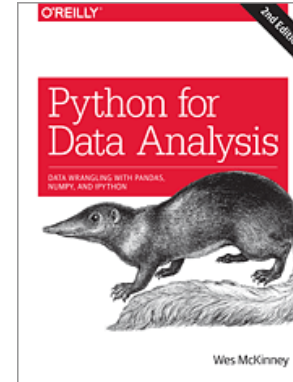
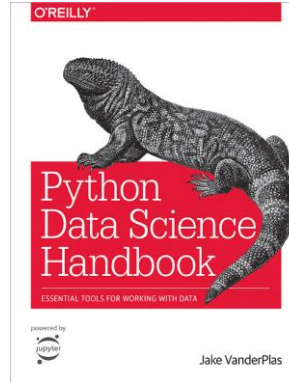
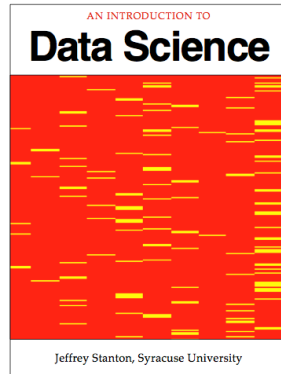
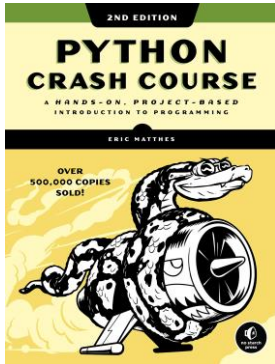
Welcome to STQD6014!

Data Science



Details

Textbook



Programming: Python

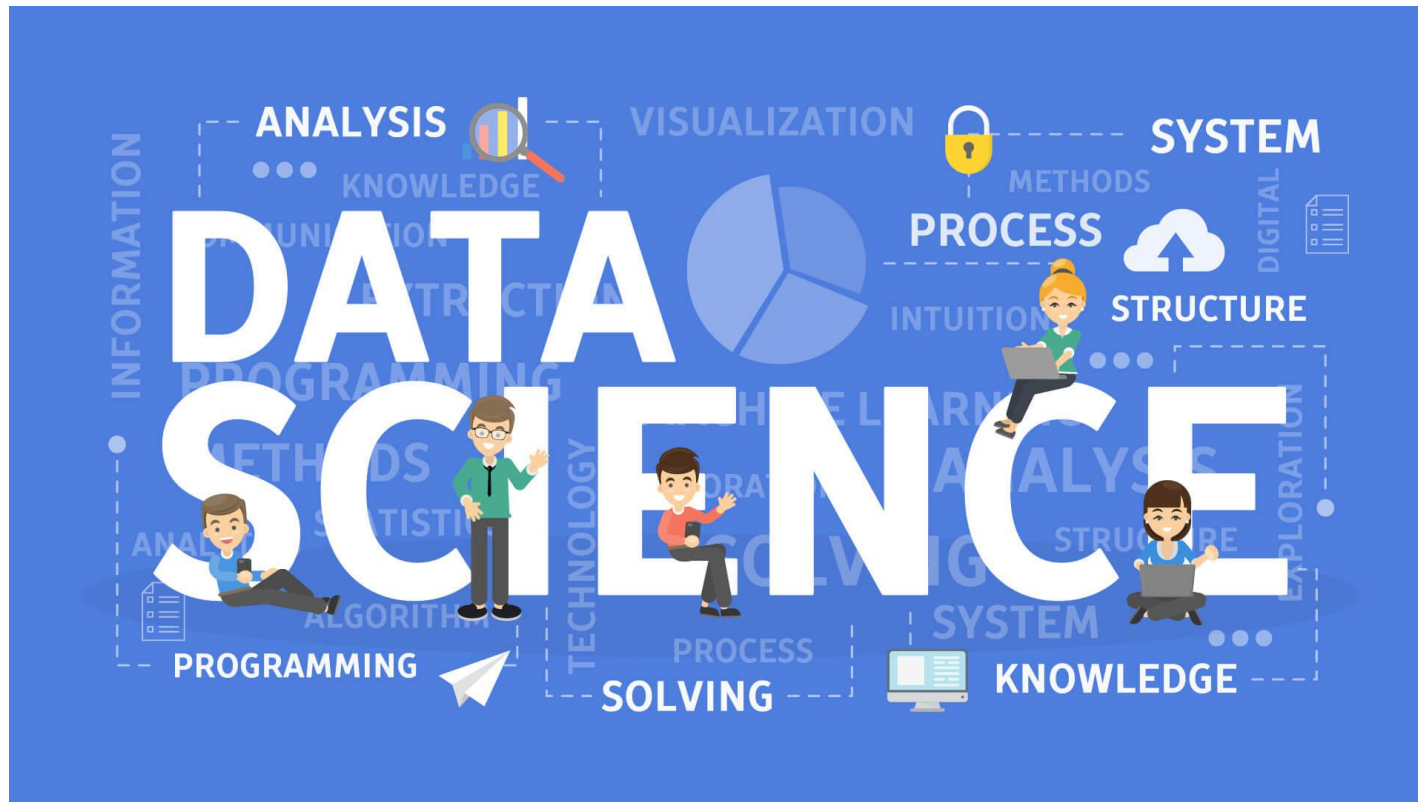
<https://www.python.org/>

Python distribution:

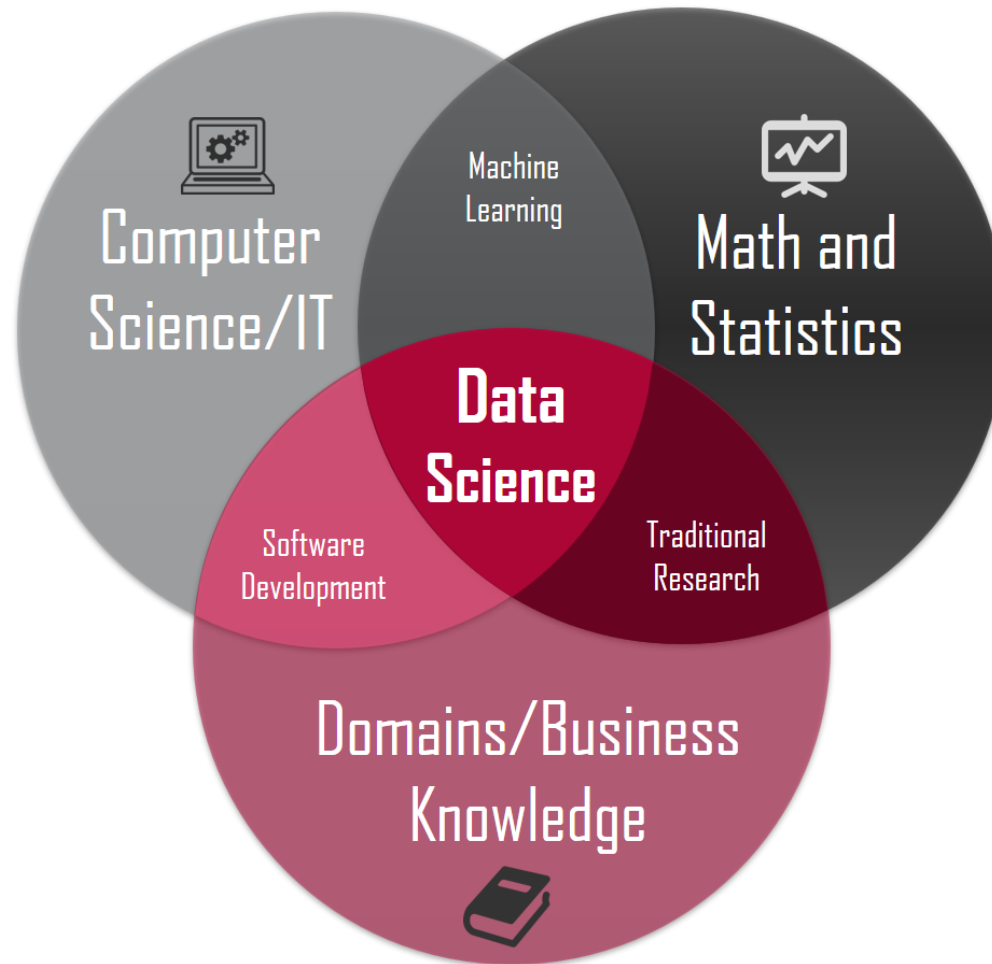
<https://www.anaconda.com/>

Why are you here???

FIRST LOOK



FIRST LOOK



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

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.

DEFINITION

- an emerging area of work concerned with the collection, preparation, analysis, visualization, management, and preservation of large collections of information
- includes data analysis as an important component of the skill set required for many jobs in this area, but is not the only necessary skill
- Data scientists play active roles in the design and implementation work of four related areas:
 - data architecture,
 - data acquisition,
 - data analysis, and
 - data archiving.

EXAMPLE

- Let's consider this idea by thinking about some of the data involved in buying a box of cereal.
- Whatever your cereal preferences - fruity, chocolaty, fibrous, or nutty - you prepare for the purchase by writing "cereal" on your grocery list.
- Already your planned purchase is a piece of data, albeit a pencil scribble on the back on an envelope that only you can read.
- When you get to the grocery store, you use your data as a reminder to grab that jumbo box of FruityChocoBoms off the shelf and put it in your cart.
- At the checkout line the cashier scans the barcode on your box and the cash register logs the price.
- Back in the warehouse, a computer tells the stock manager that it is time to request another order from the distributor, as your purchase was one of the last boxes in the store.
- You also have a coupon for your big box and the cashier scans that, giving you a predetermined discount.

EXAMPLE

- At the end of the week, a report of all the scanned manufacturer coupons gets uploaded to the cereal company so that they can issue a reimbursement to the grocery store for all of the coupon discounts they have handed out to customers.
- Finally, at the end of the month, a store manager looks at a colorful collection of pie charts showing all of the different kinds of cereal that were sold, and on the basis of strong sales of fruity cereals, decides to offer more varieties of these on the store's limited shelf space next month.

THE PROCESS

- Computer/barcode scanner: collecting, manipulating, transmitting, and storing the data
- Softwares: organize, aggregate, visualize, and present the data
- Human: involved in working with the data. People decided which systems to buy and install, who should get access to what kinds of data, and what would happen to the data after its immediate purpose was fulfilled. The personnel of the grocery chain and its partners made a thousand other detailed decisions and negotiations before the scenario described above could become reality.
- Q: Is data scientist involved in all of these phases?

THE ACTIVE ROLES OF DATA SCIENTIST

- data architecture

help the system architect by providing input on how the data would need to be routed and organized to support the analysis, visualization, and presentation of the data to the appropriate people.

- data acquisition

how the data are collected, and, importantly, how the data are represented prior to analysis and presentation. Representing, transforming, grouping, and linking the data are all tasks that need to occur before the data can be profitably analyzed, and these are all tasks in which the data scientist is actively involved.

THE ACTIVE ROLES OF DATA SCIENTIST

- data analysis

summarization of the data, using portions of data (samples) to make inferences about the larger context, and visualization of the data by presenting it in tables, graphs, and even animations. Although there are many technical, mathematical, and statistical aspects to these activities, keep in mind that the ultimate audience for data analysis is always a person or people, hence excellent communication skills are needed.

- data archiving

Preservation of collected data in a form that makes it highly reusable - what you might think of as "data curation" - is a difficult challenge because it is so hard to anticipate all of the future uses of the data.

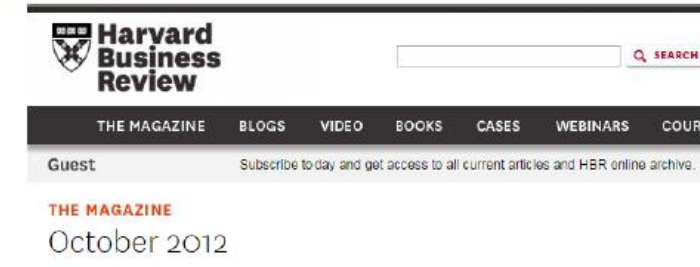
ESSENTIAL SKILLS NEEDED

- Learning the application domain - The data scientist must quickly learn how the data will be used in a particular context.
- Communicating with data users - A data scientist must possess strong skills for learning the needs and preferences of users. Translating back and forth between the technical terms of computing and statistics and the vocabulary of the application domain is a critical skill.
- Seeing the big picture of a complex system - After developing an understanding of the application domain, the data scientist must imagine how data will move around among all of the relevant systems and people.
- Knowing how data can be represented - Data scientists must have a clear understanding about how data can be stored and linked, as well as about "metadata" (data that describes how other data are arranged).

ESSENTIAL SKILLS NEEDED

- Data transformation and analysis - When data become available for the use of decision makers, data scientists must know how to transform, summarize, and make inferences from the data. As noted above, being able to communicate the results of analyses to users is also a critical skill here.
- Visualization and presentation - Although numbers often have the edge in precision and detail, a good data display (e.g., a bar chart) can often be a more effective means of communicating results to data users.
- Attention to quality - No matter how good a set of data may be, there is no such thing as perfect data. Data scientists must know the limitations of the data they work with, know how to quantify its accuracy, and be able to make suggestions for improving the quality of the data in the future.
- Ethical reasoning - If data are important enough to collect, they are often important enough to affect people's lives. Data scientists must understand important ethical issues such as privacy and must be able to communicate the limitations of data to try to prevent misuse of data or analytical results.

Data Scientists are in high demand



Harvard Business Review

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



Forbes

Now Posts Most Popular Lists Video 2 Free issues of Forbes Search magazines, people

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: Top Of People To Up



הדגם המסחר
המובילה על פי בארנס 2014



EMC²

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



25 CNBC

Enter Symbols GO Enter Keywords GO

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

NEW SHOW **SQUAWK**alley The Intersection Wall St. & Tech

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



TechRepublic / U.S.

All Topics Newsletters Photos Forums Resource Library Reso

CXO Software Startups Cloud Data Center Mobile Microsoft Apple Google Search TechRepub

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 | HOME

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

Press 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

Search ScientificAmerican.com

Subscribe

News & Features

Topics

Blogs

Videos & Podcasts

Education

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



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: Kaspersky)

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: Kaspersky)

UNIVERSITY of WASHINGTON

eScience Institute
Supporting Data-Driven Discovery In All Fields

WHO WE ARE

New Ph.D. Tracks in "Big Data"

Pays well

Big Data, Big Paycheck

Median salary for analytics professionals and those specifically within data science, by level of experience.



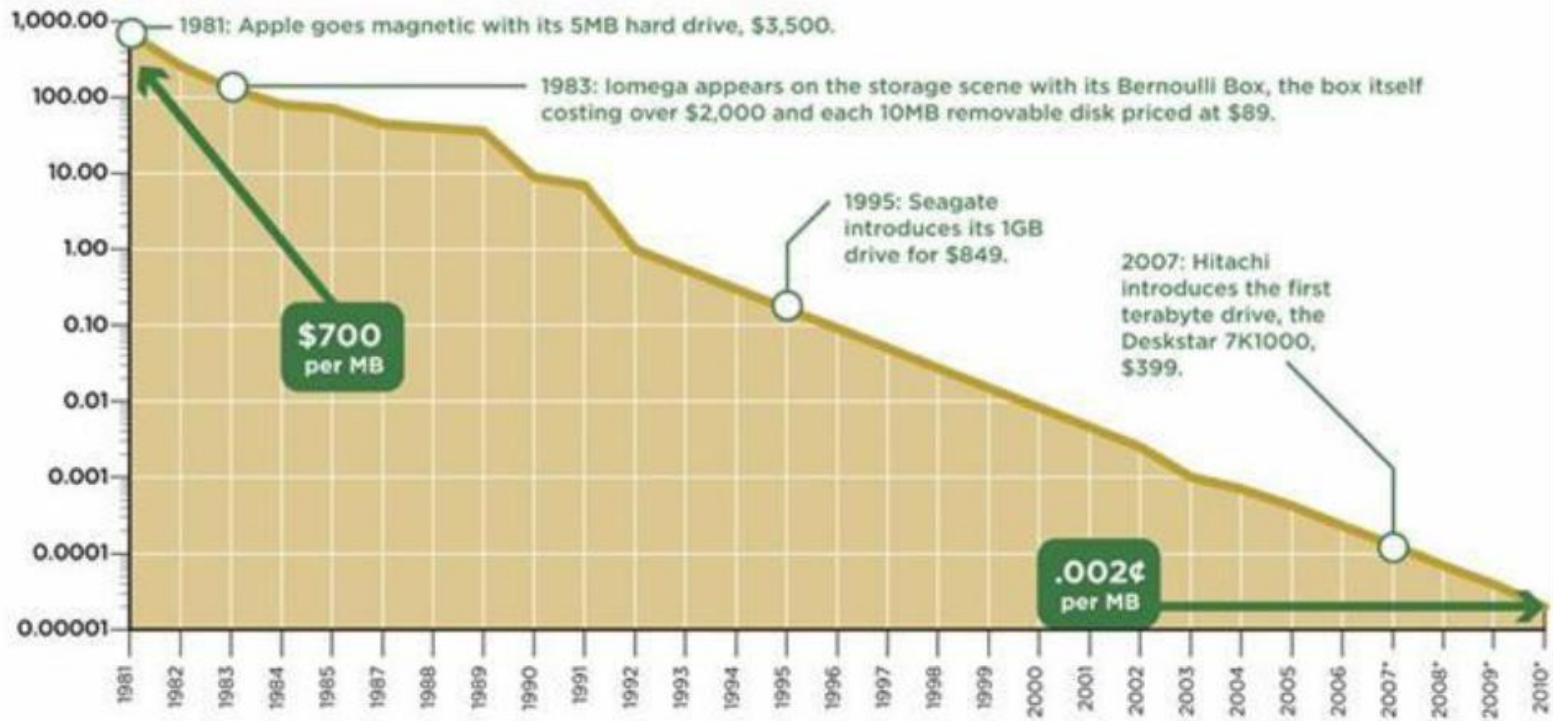
Note: Data do not include managers Source: Burtch Works

The Wall Street Journal

Declining cost of storage

STORAGE: FROM HIGHWAY ROBBERY TO RUNAWAY BARGAIN

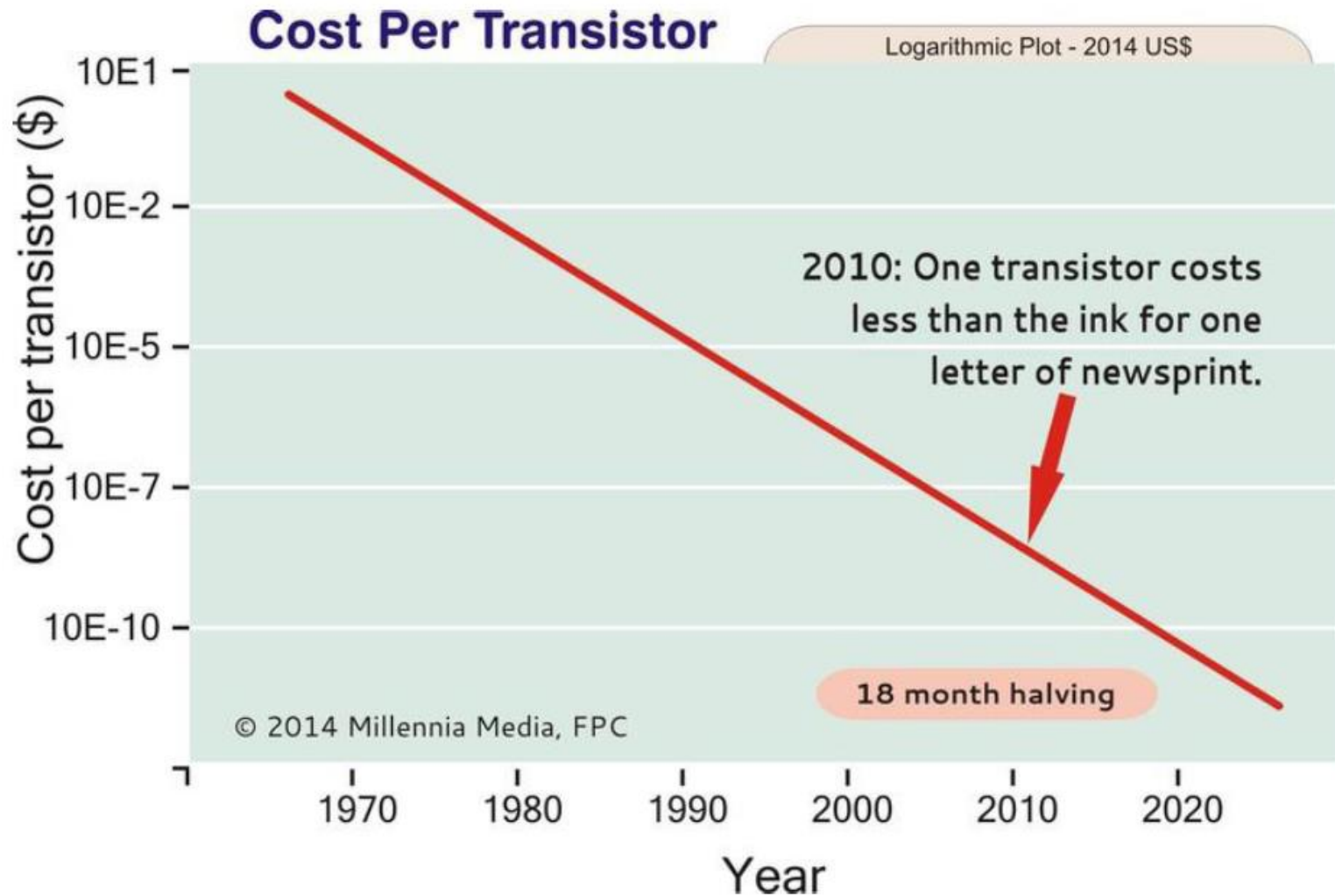
\$ per megabyte



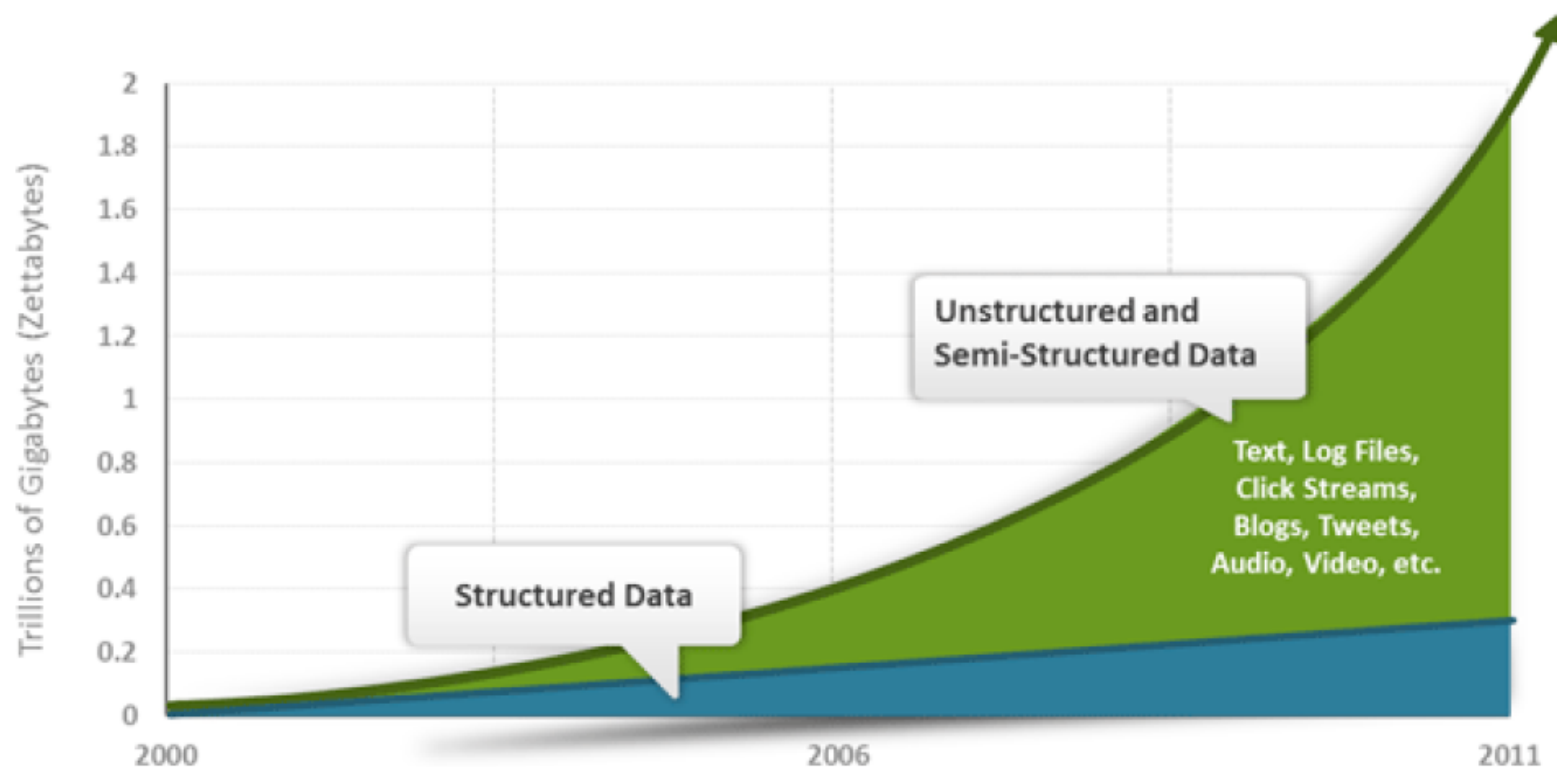
*Projected. No data is available for 1986.

Sources: Ars Technica, Little Tech Shoppe, Steve Gilheany, ExtremeTech

Declining cost of computing



More data can be stored and processed

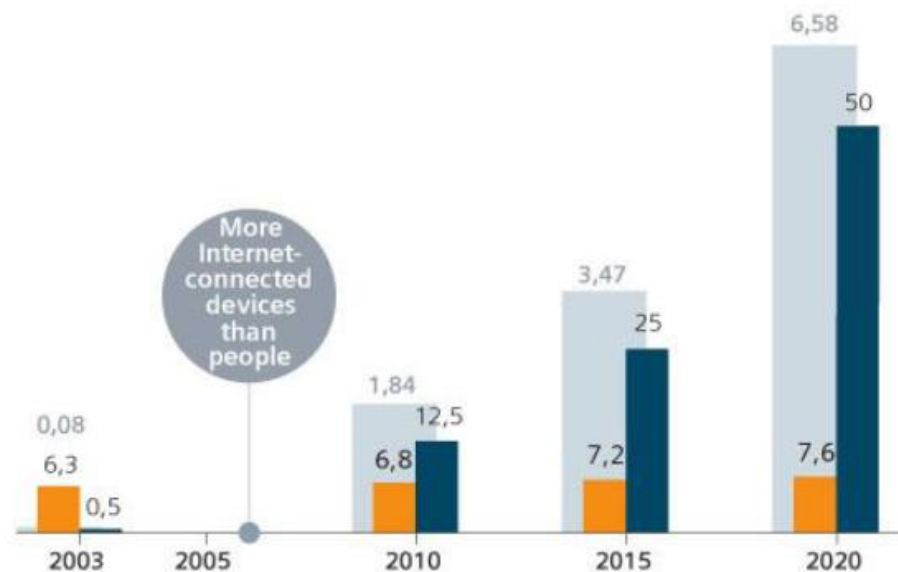


Source: IDC 2011 Digital Universe Study (<http://www.emc.com/collateral/demos/microsites/emc-digital-universe-2011/index.htm>)

Devices vs. People

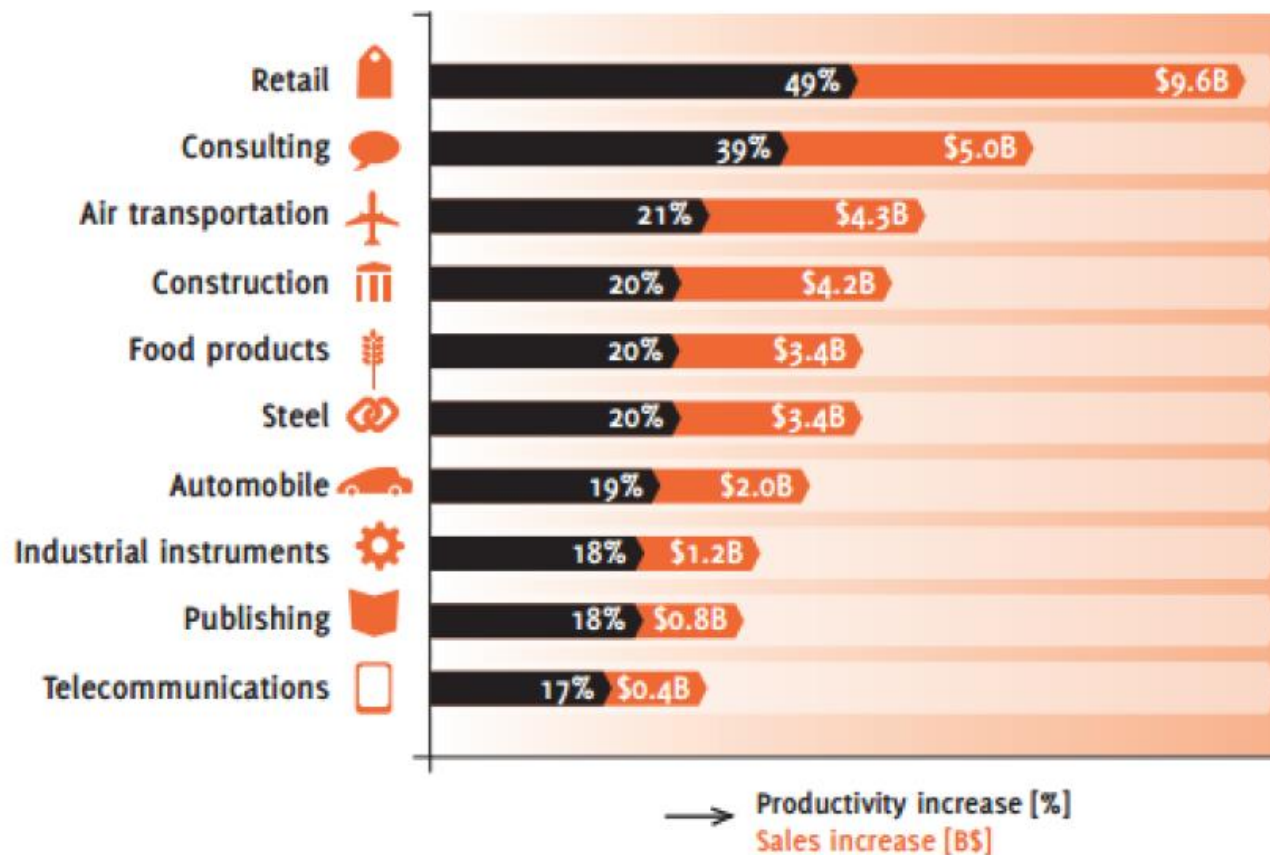
Growth in Internet-Connected Devices by 2020

- World population (in billions)
- Internet-connected devices in (billions)
- Internet-connected devices per person



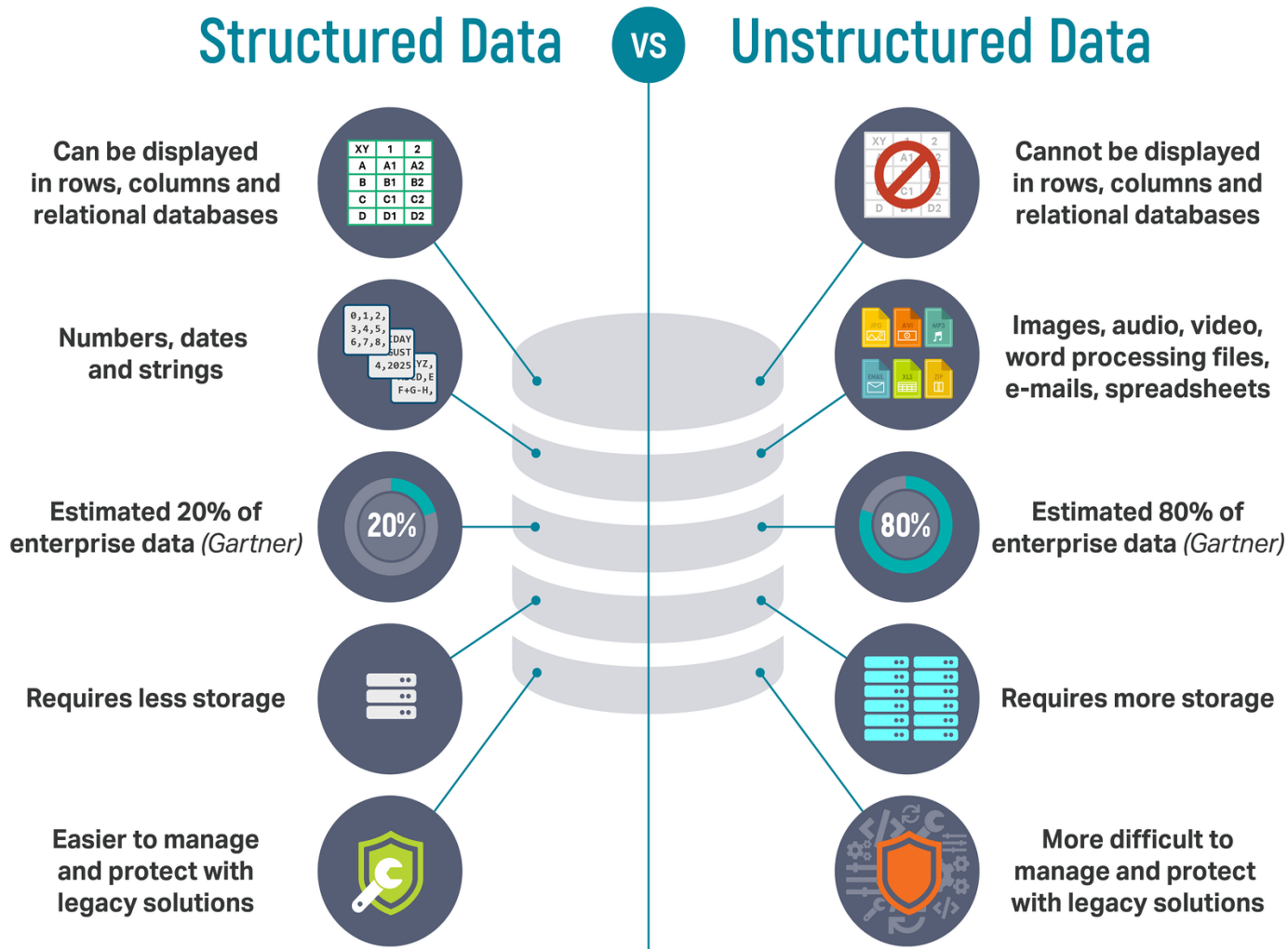
Source: Cisco IBSG, April 2011

Value of Big Data



Source: University of Texas (2011)

Unstructured vs structured data



A few examples...



Recommender Systems



Doctor Who: The End of Time

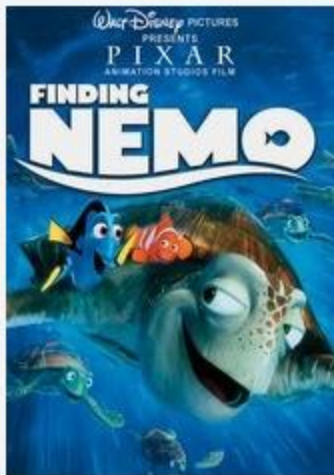
The Tenth Doctor's story comes to a close in this two-part film on the malevolent Master and the rest of the Time Lords as

Cast: David Tennant, John Simm, Bernard Cribbins, Timothy Bloom, June Whitfield, David Harewood, Tracy Iffachor

Genre: TV Sci-Fi & Fantasy, TV Action & Adventure, British



Choose Discs



Finding Nemo 2003 G 10

In this Oscar-winning animated adventure, find the missing son, Nemo, who's been scooped up

Cast: Albert Brooks, Ellen DeGeneres, Alexander Pendleton, Stephen Root, Vicki Lewis, Joe Ranney, Bob Peterson, John Ratzenberger

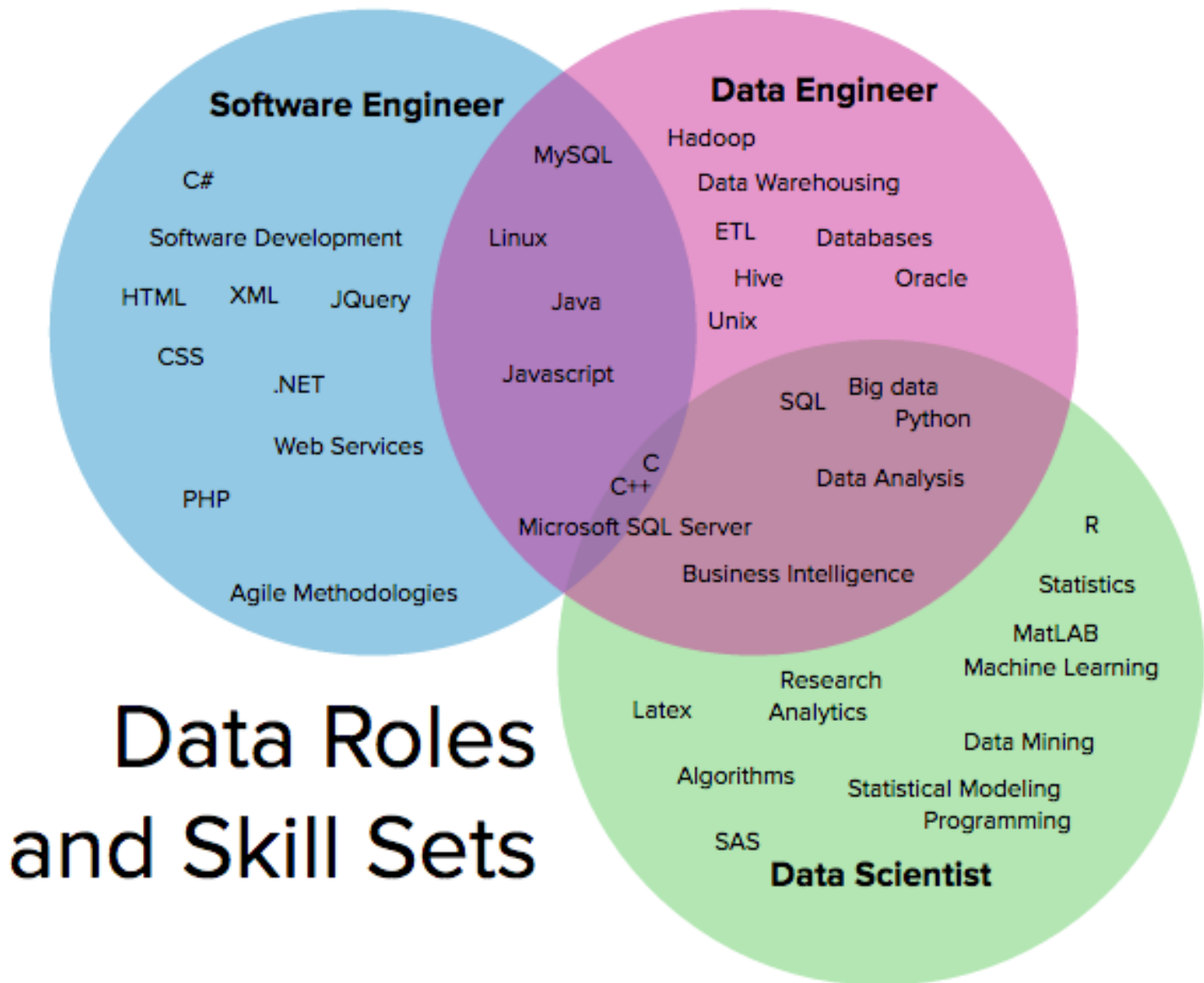
Genre: Family Animation, Family Adventures



Add



predicting
movie ratings





Data Scientist

vs



Data Engineer

vs



Statistician

These people use their analytical and technical capabilities to extract meaning insights from data.

These people ensure uninterrupted flow of data between servers and applications. They are responsible for data architecture.

These people understand statistics theoretically and apply them to real life problems.

Responsibilities

Develop and plan required analytic projects in response to business needs.

Contribute to data mining architectures, modeling standards, reporting, and data analysis methodologies.

Collaborate with stakeholders to integrate data mining results with existing systems.

Monitor data mining system performance and implement efficiency improvements.

Design, construct, install, test and maintain highly scalable data management systems

Improve data foundational procedures, guidelines and standards

Integrate new data management technologies and software engineering tools into existing structures

Create custom software components (e.g. specialized UDFs) and analytics applications

Apply statistical theories and methods to solve practical problems of various industries

Determine methods for finding or collecting data

Design surveys or experiments or opinion polls to collect data

Analyze, interpret & undertake data analysis

Report conclusions from their analyses

Skills

Programming, Mathematics, Business Understanding, Statistics, Data Visualization, Machine Learning, Attention to detail

Database design, Production coding, Data collection, data warehousing, Data transformation, Work diligently with data

Technical and Analytics Skills, Mathematics, Operational Research, Writing skills, Ability to Analyze, Model and interpret data, Flair of explaining difficult concepts in simple manner

Tools





Why STQD6014 ?

Specific skills:

Phyton programming language - 60%

Experience with several *statistical analyses* (descriptive statistics) and visualization – 30%

Experience with predictive statistics (modeling) and *machine learning* algorithms - 10%

Why STQD6014 ?

Specific skills:

Python programming language - 60%

Experience with several *statistical analyses* (descriptive statistics) and visualization – 30%

Experience with predictive statistics (modeling) and *machine learning* algorithms - 10%

Broad background:

You'll be confident and capable with whatever datasets you encounter in the future – on your own or as part of a team.

Key takeaway: <https://www.youtube.com/watch?v=X3paOmcTjQ>