

# Programming for Data Analysis: Introduction

---

**Kasthuri Kannan, PhD**  
**Associate Professor**  
**Translational Molecular Pathology**  
**UT MD Anderson Cancer Center**

**Envisioned to empower**



**Not about R**

**Using R all the time**

**Not about programming**

**Program all the time**

**Not about big data, Hadoop, distributed comp. etc.**

**Do data science all the time**

**Not about machine learning, stat. inference etc.**

**Course on machine learning offered in Spring**

# Contents

**Data science, a hype?**

**Why am I here?**

**Why data science?**

**Hopefully, there is no penalty drop-out date**

**Ok, so tell me what is data science?**

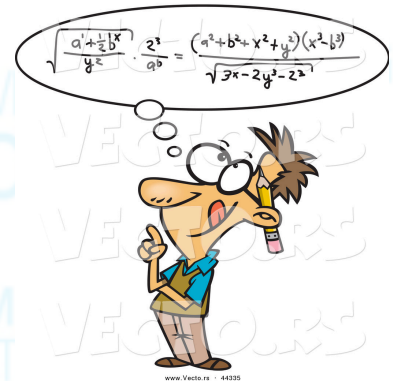
**I'll stay put if you don't kill us with equations**

I HAVE NO IDEA  
WHAT'S GOING  
TO HAPPEN.



AND I LOVE IT.

**What is this course about?**





# Data science in media

## The New York Times

Harvard Business Review

BIG DATA

### Less Noise but More Money in Data Science

BY STEVE LOHR APRIL 28, 2015 9:30 AM 10

Email

Share

Tweet

Save

More

The outlook for data scientists: less hype, more hiring.

The exuberance surrounding big data has passed its peak and is trending down, the technology research firm Gartner declared last August in its annual "hype cycle" report on perceptions of technology.

Perhaps, but it remains a rising market for data scientists. Salaries rose 8 percent on average in the last year, with bonuses adding \$56,000, according to a salary and employment survey released on Tuesday by [Burtch Works](#), a recruiter of professionals with quantitative skills.



ARTWORK: TAMAR COHEN, ANDREW J. BOWLE, STEVE BIRCHER ON A PHOTO FROM A HIGH-SCHOOL YEARBOOK, 8.7 X 11"

### Data Scientist: The Sexiest Job of the 21st Century

by Thomas H. Davenport and D.J. Patil  
FROM THE OCTOBER 2012 ISSUE

WHAT TO READ NEXT



Big Data: The Management Revolution

Bloomberg Markets Tech Pursuits Politics Opinion Businessweek



## Here's a Retail Job That's Still in High Demand: Data Scientist

By Taylor Cromwell

August 21, 2017, 7:00 AM EDT



FEATURE

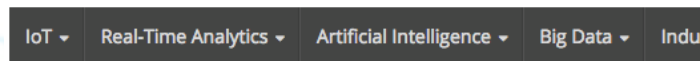
### What is a data scientist? A key data analytics role and a lucrative career

Becoming a data scientist varies depending on industry, but there are common skills, experience, education and training that will give you the leg up in starting your data science career.



By Sarah K. White  
Senior Writer, CIO | AUG 18, 2017 3:00 AM PT

RTInsights.com  
Accelerate Your Business with Real-Time Insights



Home / Analytics / Why You May Want a Career in Data Science

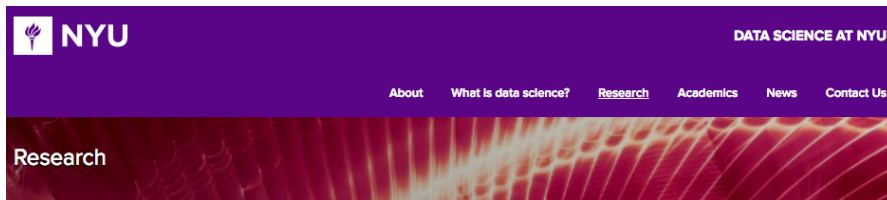
### Why You May Want a Career in Data Science

By Dan Muse | August 17, 2017

# Data science in academia

## NSF Announces \$17.7 Million Funding for Data Science Projects

August 25, 2017 by [staff](#) [Leave a Comment](#) 



### 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.

# 500k

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

### Big data will be focus of new UW research institute

By Erik Lorenzsonn Sep 2, 2017

## KSU launches Analytics and Data Science Institute

Staff reports Aug 25, 2017 Comments

## NSF awards \$1.5 million grant for data science research at UC Santa Cruz

A cross-disciplinary team of computer scientists, statisticians, and mathematicians is developing the theoretical foundations of the emerging field of data science

PUBLIC RELEASE: 24-AUG-2017

Brown awarded \$1.5M to establish data science research institute

BROWN UNIVERSITY

# Demand is likely to outpace supply

## Demand in data science

At its core, data science involves using automated methods to analyze massive amounts of data and to extract knowledge from them. With such automated methods turning up everywhere from genomics to high-energy physics, data science is helping to create new branches of science, and influencing areas of social science and the humanities.



# 50X in 2020

The world will generate 50 times more data than was generated in 2011.

The U.S. alone is going to face a **shortage of 140,000 to 200,000** professionals with data science skills by 2018.

Source: McKinsey Global Institute

No. **1** The position of "data scientist" on the list of the 25 best jobs in America in 2016.

**\$116,840**  
median salary in the US.

**94%**  
of the US graduates have found jobs, averaging **\$114,000** since 2011.

Over 2/3 believe demand for talent will outpace the supply of data scientists

OVER THE NEXT FIVE YEARS, DEMAND FOR DATA SCIENTISTS WILL:



Only 12% see today's BI professional as the best source for new data scientists

## JOB GROWTH AND DEMAND

**NAMED THE TOP JOB IN AMERICA FOR 2016** BY GLASSDOOR

**SEXIEST JOB OF THE 21ST CENTURY** BY HARVARD BUSINESS REVIEW

**79.7%** OF DATA SCIENTISTS REPORT THERE IS A **SHORTAGE** IN THEIR FIELD

**11%** PROJECTED GROWTH FROM 2014 TO 2024 **7%** FASTER THAN GROWTH FOR ALL OCCUPATIONS

# Extremely well-paid career

The average salary for Data Scientists is \$189K.★

Sign in to see how much you should be making »



## 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

## Data Scientist salaries in San Francisco, CA

**\$140,703 per year**

Based on 4,567 salaries



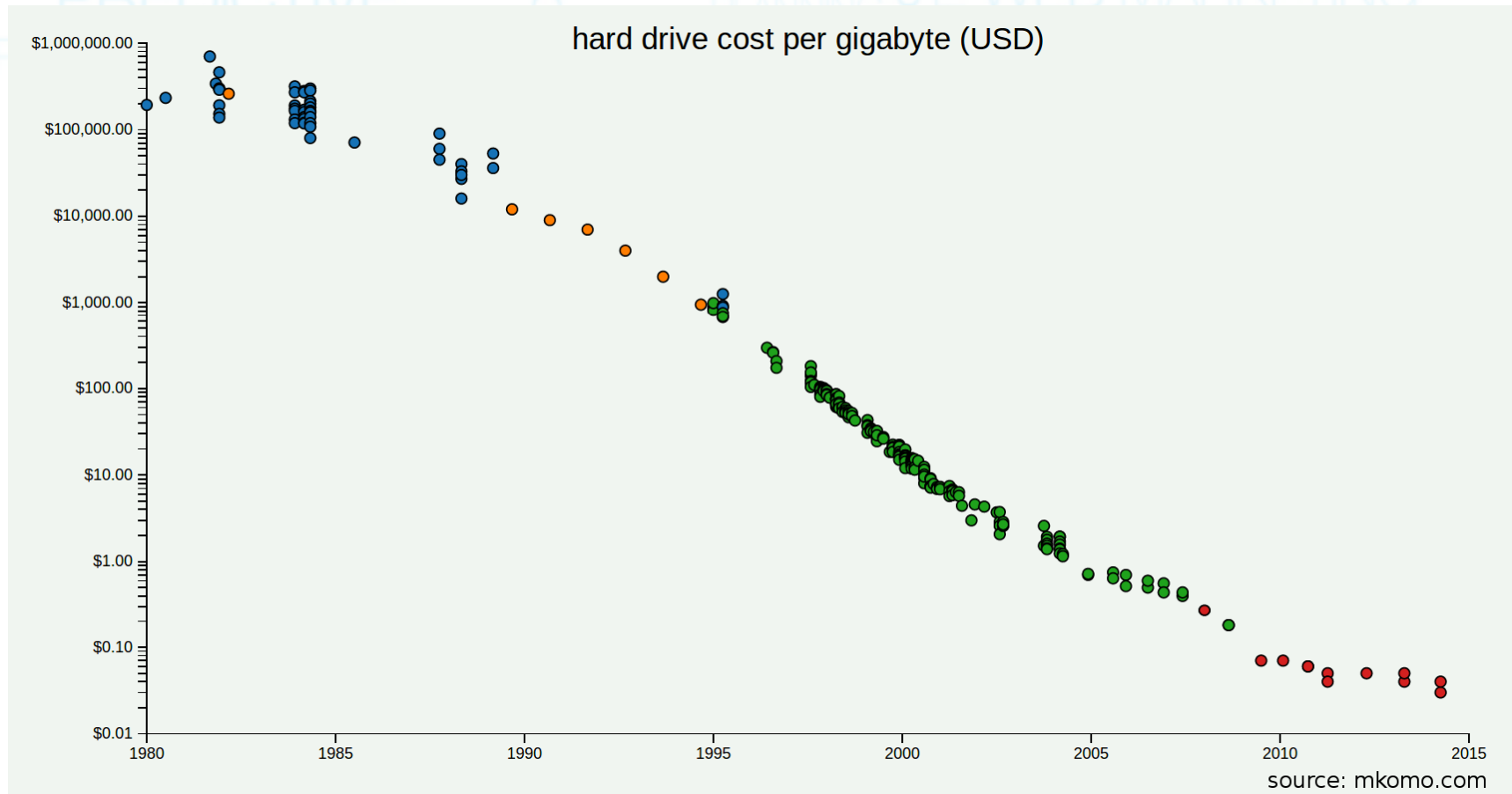
Data Scientist salaries by company in San Francisco, CA





# Drastic reduction in storage costs

**Technology = Reduction in storage cost**



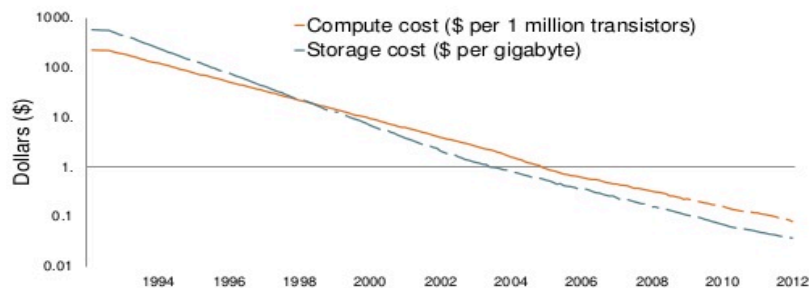
# Drastic reduction in computation costs

## Technology = Reduction in computing cost

Moore's law refers to an observation made by Intel co-founder Gordon Moore in 1965. He noticed that the number of transistors per square inch on integrated circuits had doubled every year since their invention. Moore's law predicted that this trend will tend to continue into the foreseeable future. It is almost ending now, though.

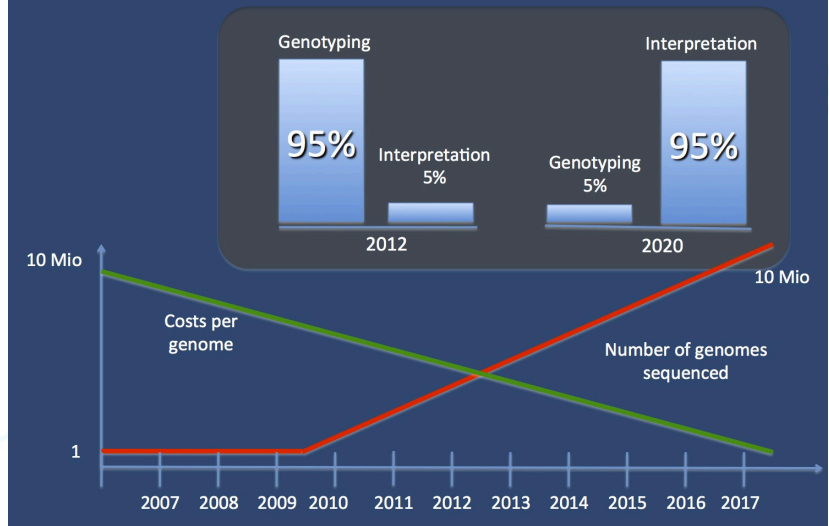
### Moore's law: cost of storage, compute $\Rightarrow$ zero

Storage cost-performance and computing cost-performance



ANDREESSEN HOROWITZ

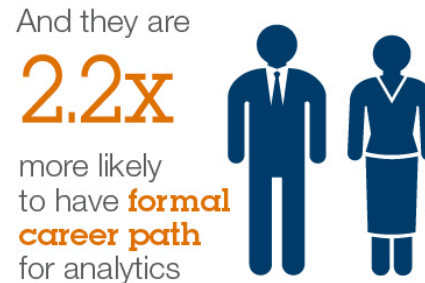
### Genome vs. genome analysis



# Big data has immense value

## Capitalizing on Big Data:

Strategies outperforming companies are taking to deliver results



Leaders **measure the impact** of analytics investments



Leaders have **predictive analytics** capabilities



Leaders have some form of **shared analytics resources**

Join the conversation on Twitter at #ibmanalytics and follow @IBMIBV

# Raising expectations

## Cognitive computing

People expect systems to behave like humans

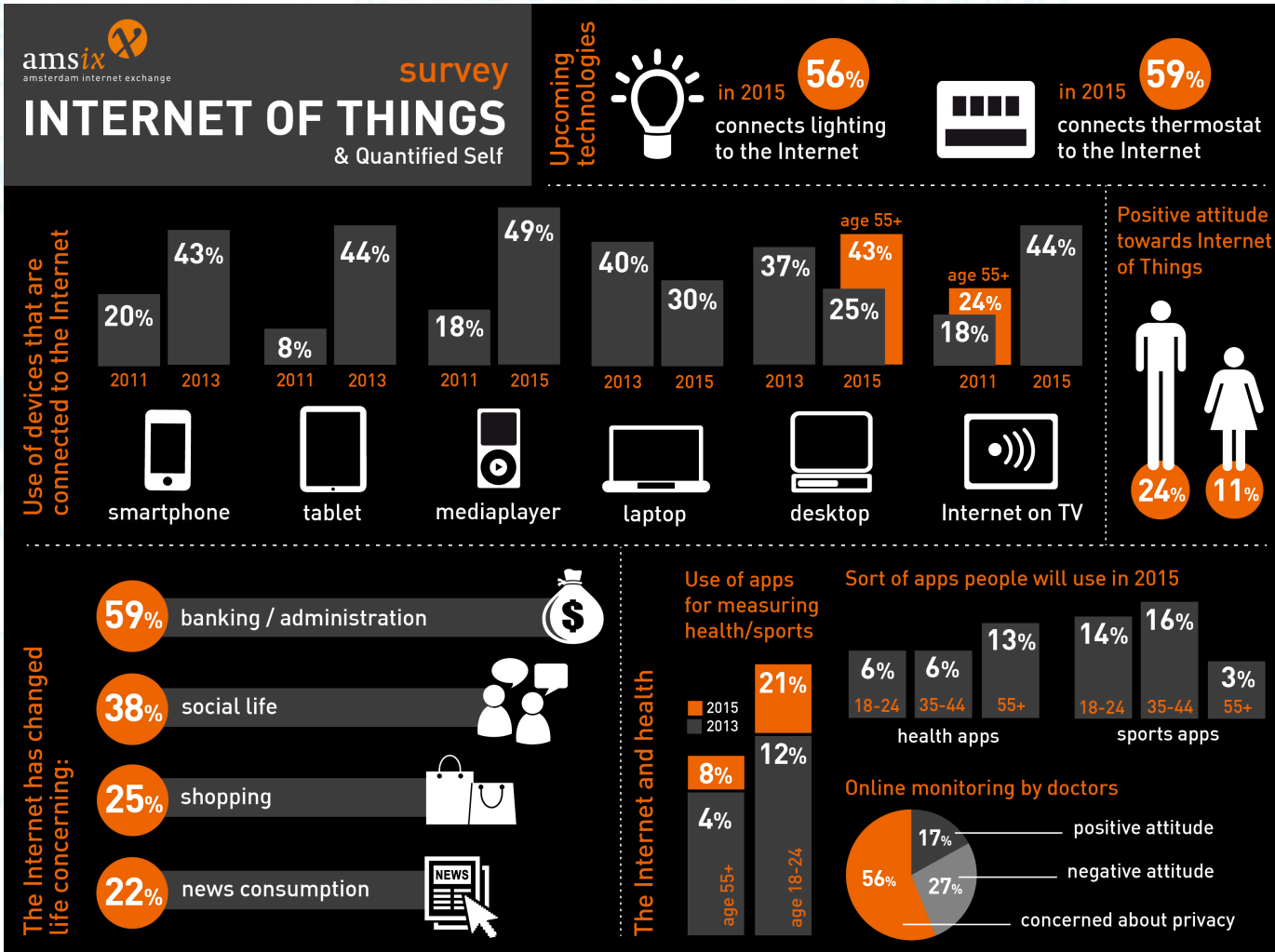
Adaptive (learning as the information changes)

Interactive (communicating with other humans/systems)

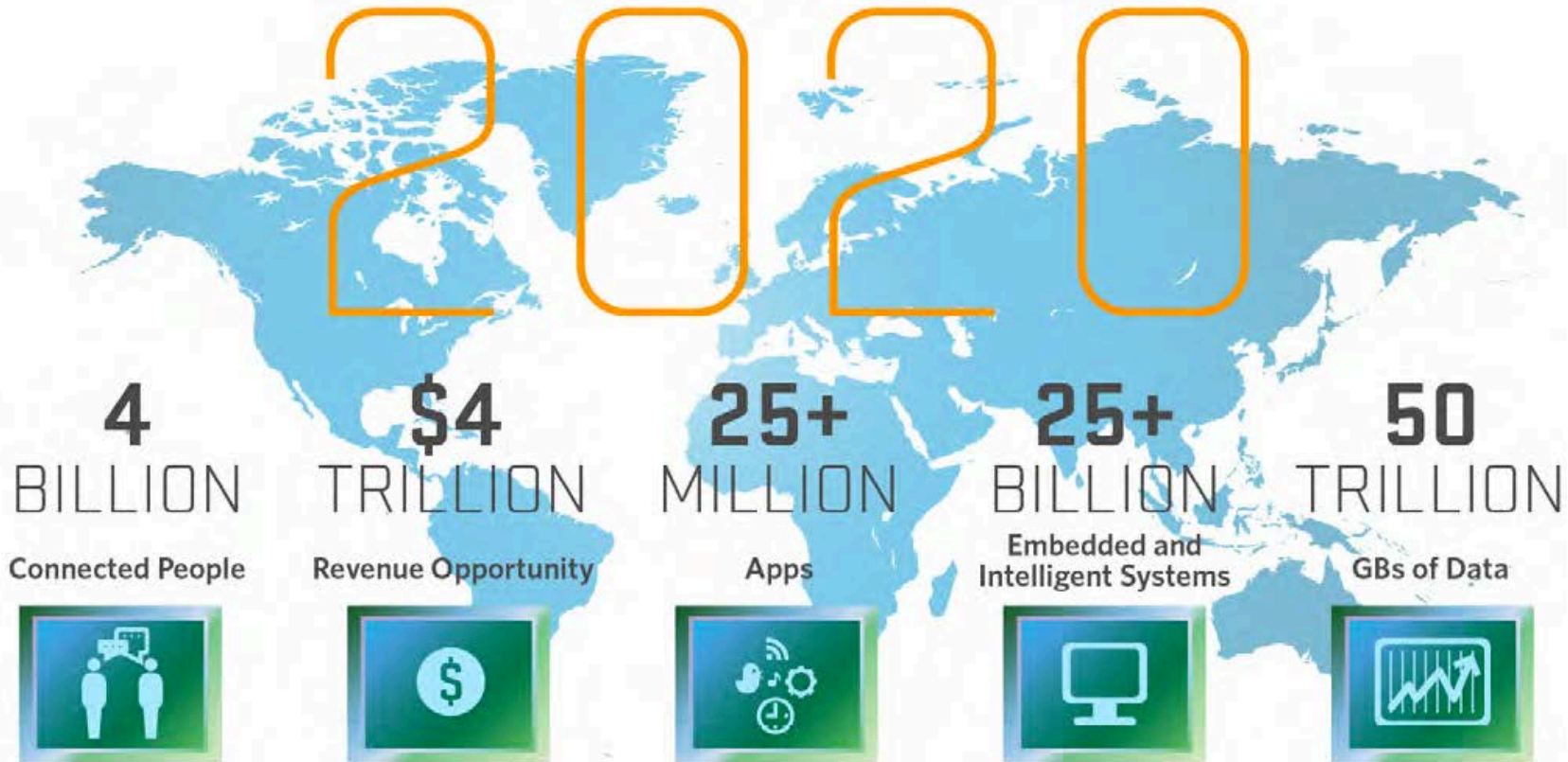
Contextual (understanding meanings, integrate other info)

Processing large datasets of differing types like text, voice, sensors and images.

# Internet of Things: The next frontier



# Internet of Things: Projection



Source: Mario Morales, IDC

# Why biomedical data science?

## THE INTERNET OF (MEDICAL) THINGS TECHNOLOGY

**3.7M** Medical devices in use today connecting to & monitoring various parts of the body

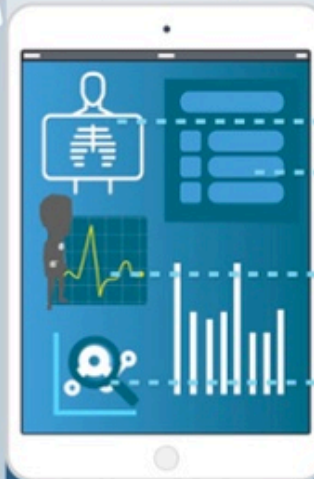
Active implantable medical devices control stimulation &/or precision medicine therapy to treat disease & improve patient quality of life.



Monitors medical conditions specific to patient's disease & other systemic conditions such as heart rate, blood sugar, exercise, etc.

**Closed-Loop System**  
"Smart" software supports device iteration based on data inputs to deliver best patient therapy

One IOMT system solution collecting data from medical devices, medications, & biometrics to modify therapeutic window towards best care option



**97%** Wi-Fi adoption rates in hospitals  
**10%** Medical devices enabled with Wi-Fi

### OPTIMIZED RESULTS FOR:

#### PATIENTS...



Receive **individually-optimized care** faster, with few doctor office visits, and decreased overall time "thinking" about the disease

#### HEALTHCARE PROFESSIONALS...



Monitor patient status, disease progression, & device performance. This allows for:

- **Enhanced patient support**
- **Reduced risk**
- **Feedback on device design improve opportunities**

#### PATIENT FAMILIES...



Can be included in regular communications to **help monitor or reassure assurance of patient wellness.**

#### HEALTHCARE SYSTEM...



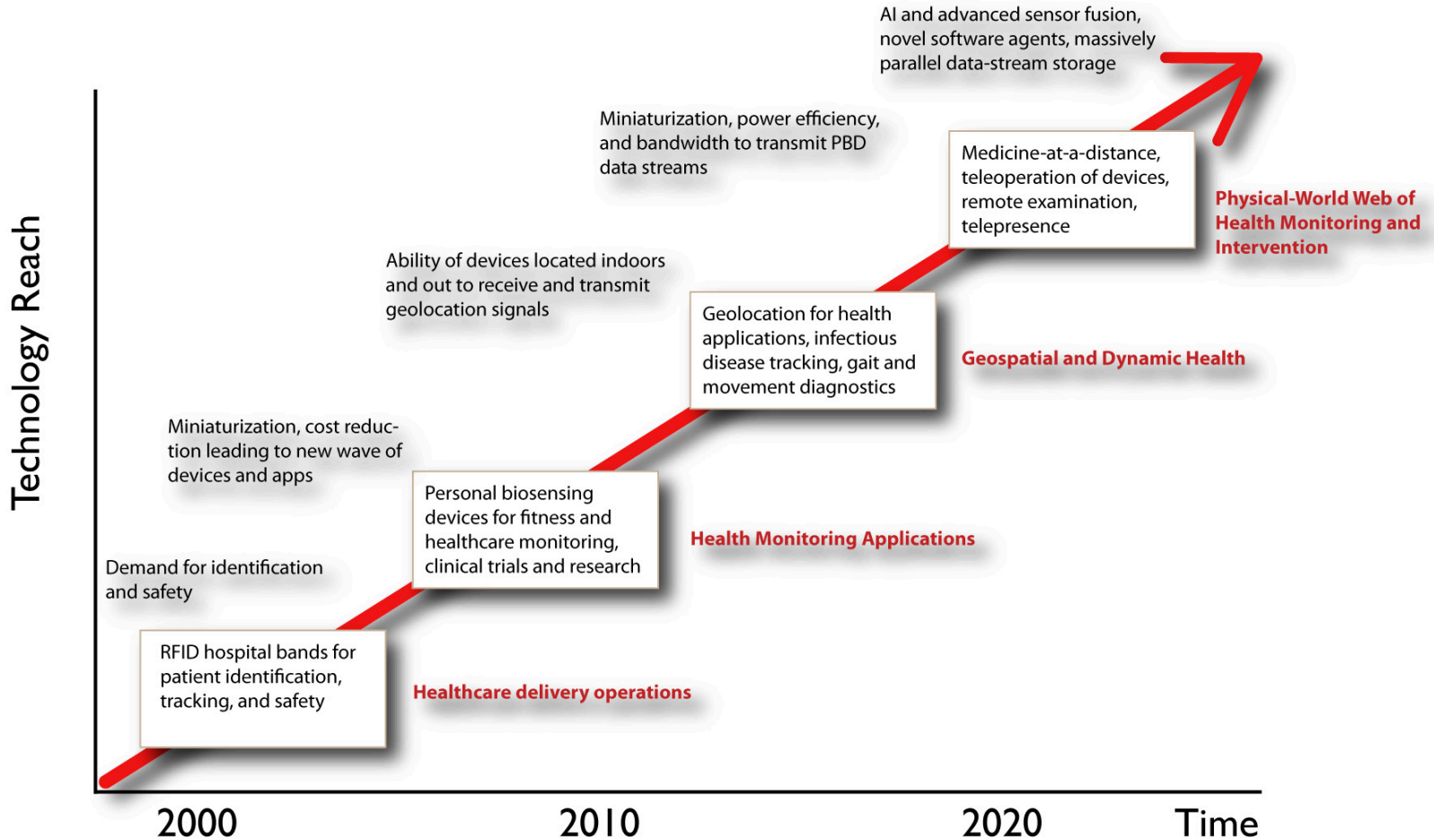
Automated monitoring & verification of advanced products to **eliminate human error & falsification**

**NEXEON**



# Io(M)T: Tech vs. Time

## Roadmap: The Internet of Medical Things.....





# A mash up of disciplines

**An umbrella term for techniques used when trying to extract insights and information from the data.**

## 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.

# Example: Data science in healthcare

## Survival analysis

Analyze survival statistics for different patient attributes (like age, gender, blood type) and treatments.

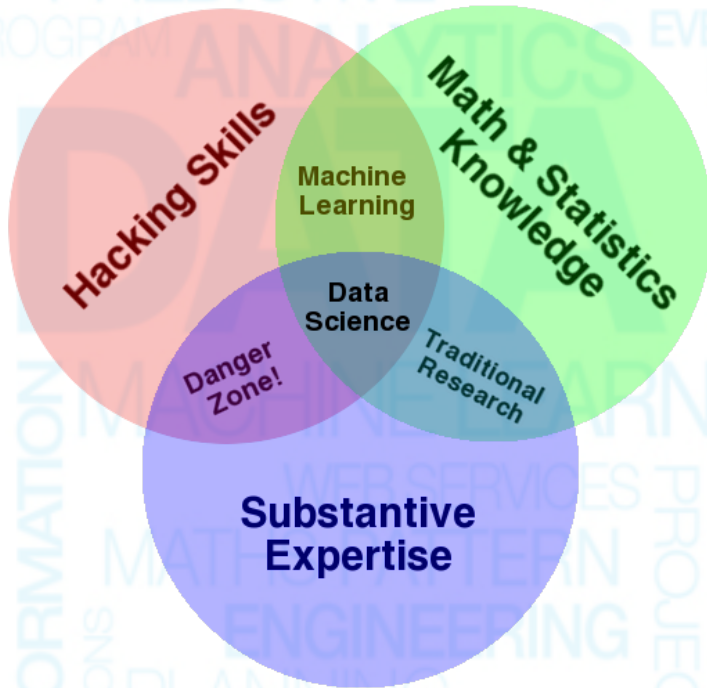
## Dosage effectiveness

Study the effectiveness of the dosage from the measured variables based on the medication for a disease.

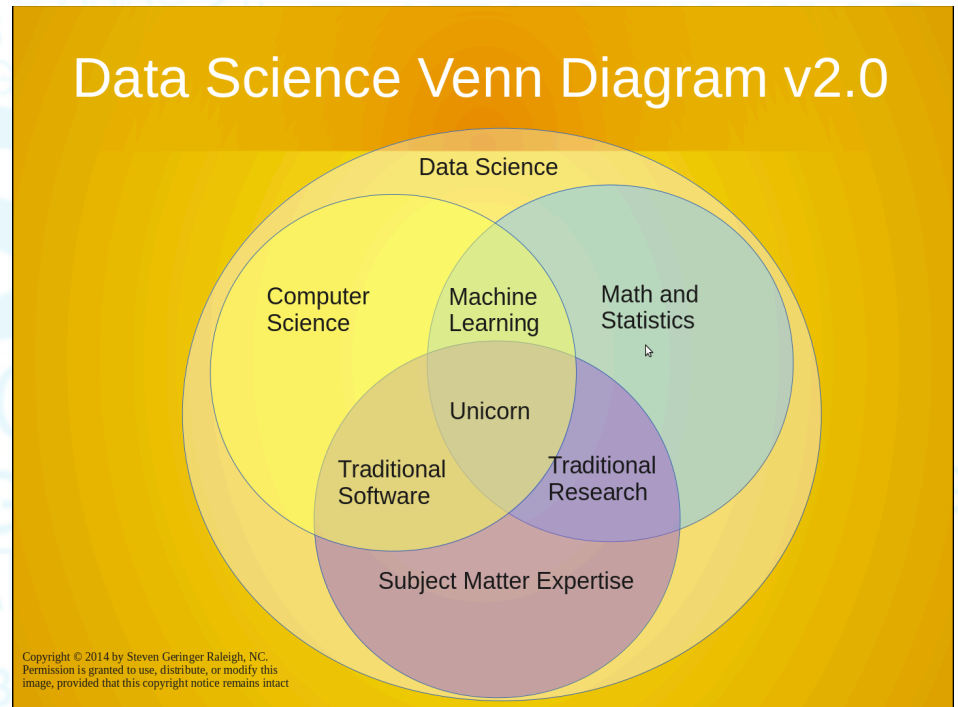
## Readmission risk

Predict the risk of readmission based on patient attributes, medical history, diagnosis and treatment.

# A note on Venn diagrams

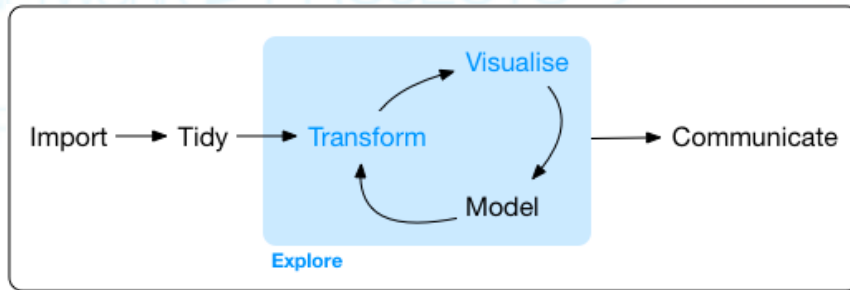


**Drew Conway's definition**

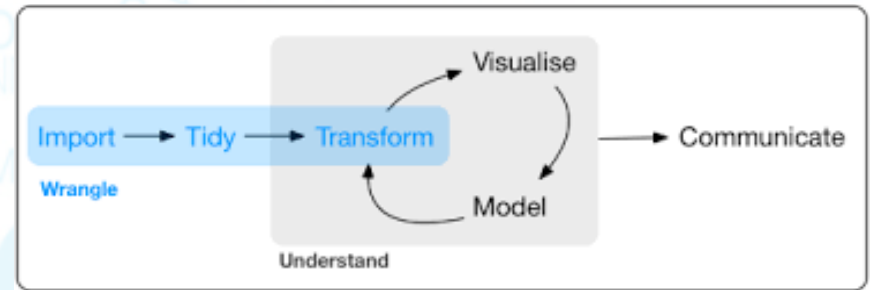


**Steven Geringer's definition**

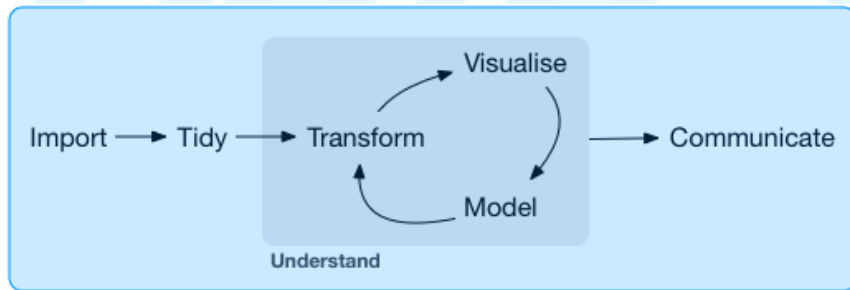
# Doing data science: Explore, Wrangle, Program, Model and Communicate



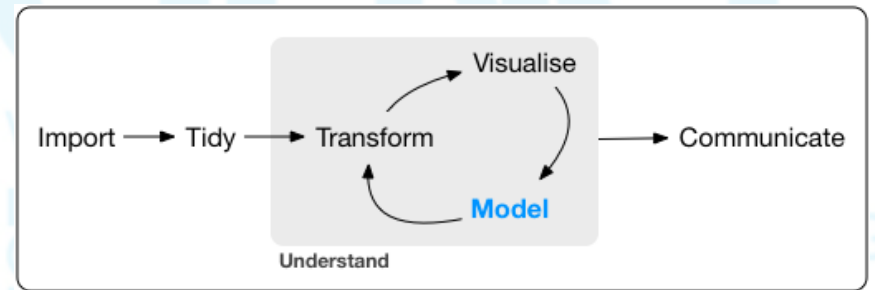
Program



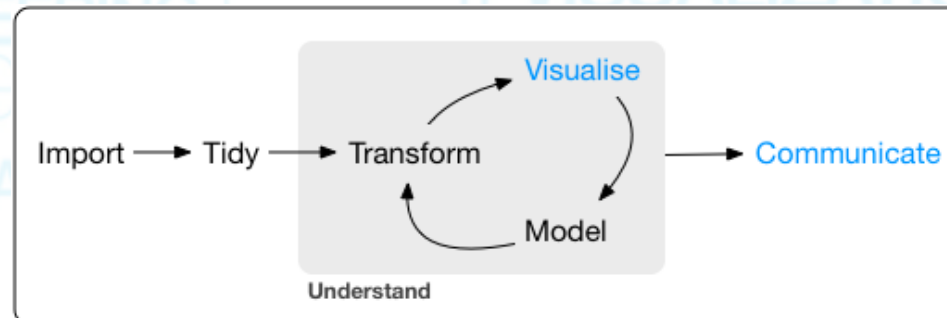
Program



Program



Program



Program

# Data scientist: Amalgamation of skills

## MODERN DATA SCIENTIST

Data Scientist, the sexiest job of the 21st century, requires a mixture of multidisciplinary skills ranging from an intersection of mathematics, statistics, computer science, communication and business. Finding a data scientist is hard. Finding people who understand who a data scientist is, is equally hard. So here is a little cheat sheet on who the modern data scientist really is.

### MATH & STATISTICS

- ☆ Machine learning
- ☆ Statistical modeling
- ☆ Experiment design
- ☆ Bayesian inference
- ☆ Supervised learning: decision trees, random forests, logistic regression
- ☆ Unsupervised learning: clustering, dimensionality reduction
- ☆ Optimization: gradient descent and variants



### PROGRAMMING & DATABASE

- ☆ Computer science fundamentals
- ☆ Scripting language e.g. Python
- ☆ Statistical computing packages, e.g., R
- ☆ Databases: SQL and NoSQL
- ☆ Relational algebra
- ☆ Parallel databases and parallel query processing
- ☆ MapReduce concepts
- ☆ Hadoop and Hive/Pig
- ☆ Custom reducers
- ☆ Experience with xaaS like AWS

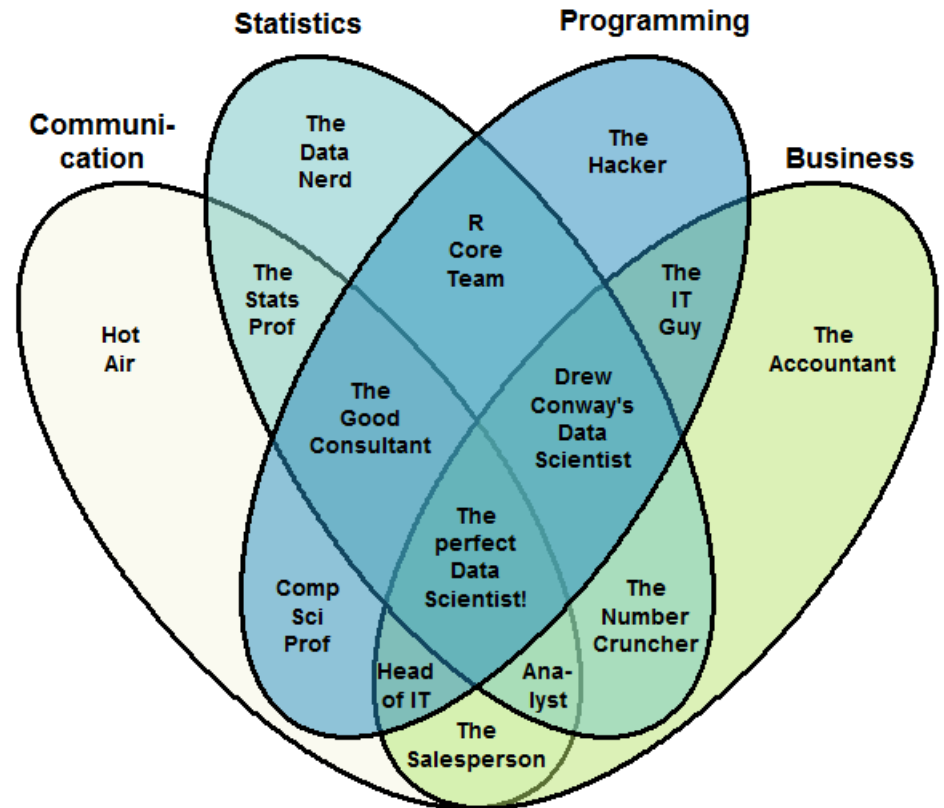
### DOMAIN KNOWLEDGE & SOFT SKILLS

- ☆ Passionate about the business
- ☆ Curious about data
- ☆ Influence without authority
- ☆ Hacker mindset
- ☆ Problem solver
- ☆ Strategic, proactive, creative, innovative and collaborative

### COMMUNICATION & VISUALIZATION


- ☆ Able to engage with senior management
- ☆ Story telling skills
- ☆ Translate data-driven insights into decisions and actions
- ☆ Visual art design
- ☆ R packages like ggplot or lattice
- ☆ Knowledge of any of visualization tools e.g. Flare, D3.js, Tableau

## The Data Scientist Venn Diagram



MarketingDistillery.com is a group of practitioners in the area of e-commerce marketing. Our fields of expertise include: marketing strategy and optimization; customer tracking and on-site analytics; predictive analytics and econometrics; data warehousing and big data systems; marketing channel insights in Paid Search, SEO, Social, CRM and brand.

Marketing  
DISTILLERY  
© Krzysztof Zawadzki

 <https://datascience.stackexchange.com/questions/2403/d-ata-science-without-knowledge-of-a-specific-topic-is-it-worth-pursuing-as-a-ca> #83642646

# Data analysts vs. Data scientists

	ANALYSTS	DATA SCIENTISTS
<b>Types of data</b>	Structured and semistructured, mostly numeric data	All types, including unstructured, numeric and nonnumeric data (such as images, sound, text)
<b>Preferred tools</b>	Statistical and modeling tools, usually contained in a data repository	Mathematical languages (such as R and Python), machine learning, natural language processing and open-source tools that access and manipulate data on multiple servers (such as Hadoop)
<b>Nature of work</b>	Report, predict, prescribe and optimize	Explore, discover, investigate and visualize
<b>Typical educational background</b>	Operations research, statistics, applied mathematics, predictive analytics	Computer science, data science, symbolic systems, cognitive science
<b>Mind-set</b>	Percentage who say they: <ul style="list-style-type: none"> <li>•are entrepreneurial: 69%</li> <li>•explore new ideas: 58%</li> <li>•gain insights outside of formal projects: 54%</li> </ul>	Percentage who say they: <ul style="list-style-type: none"> <li>•are entrepreneurial: 96%</li> <li>•explore new ideas: 85%</li> <li>•gain insights outside of formal projects: 89%</li> </ul>



# Roles and paychecks

## DATA ENGINEER

SOFTWARE ENGINEERS BY TRADE

### Role

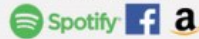
Develops, constructs, tests and maintains architectures (such as databases and large-scale processing systems)

### Mindset

All-purpose everyman



HIRED BY



### Languages

SQL, Hive, Pig, R, Matlab, SAS, SPSS, Python, Java, Ruby, C++, Perl

### Skills & Talents

- ✓ Database systems (SQL & NO SQL based)
- ✓ Data modeling & ETL tools
- ✓ Data APIs
- ✓ Data warehousing solutions

## DATABASE ADMINISTRATOR

DATABASE CARETAKER

### Role

Ensures that the database is available to all relevant users, is performing properly and is being kept safe

### Mindset

Master of Disaster Prevention



HIRED BY



### Languages

SQL, Java, Ruby on Rails, XML, C#, Python

### Skills & Talents

- ✓ Backup & recovery
- ✓ Data modeling and design
- ✓ Distributed Computing (Hadoop)
- ✓ Database systems (SQL and NO SQL based)
- ✓ Data security
- ✓ ERP & business knowledge

## BUSINESS ANALYST

CHANGE AGENT

### Languages

SQL

### Skills & Talents

- ✓ Basic tools (e.g. MS Office)
- ✓ Data visualization tools (e.g. Tableau)
- ✓ Conscious listening and storytelling
- ✓ Business Intelligence understanding
- ✓ Data modeling



HIRED BY



### Role

Improves business processes as intermediary between business and IT

### Mindset

Resilient project juggler

## DATA SCIENTIST

AS RARE AS UNICORNS

### Role

Cleans, massages and organizes (big) data

### Mindset

Curious data wizard



HIRED BY



### Languages

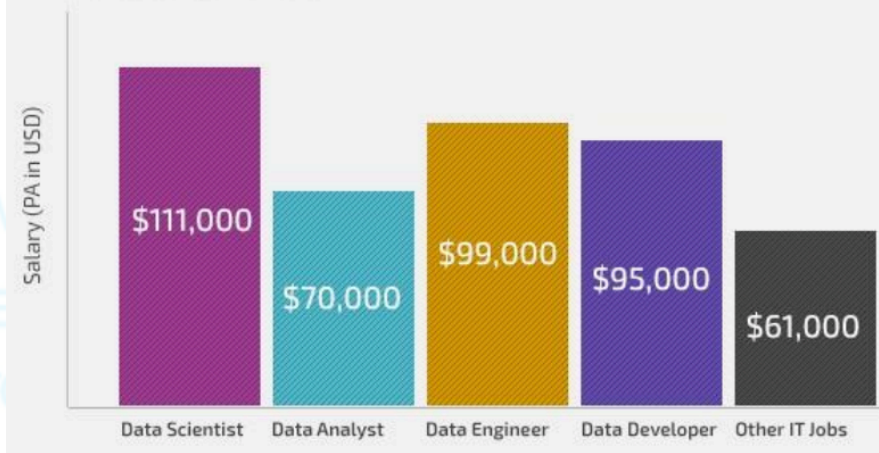
R, SAS, Python, Matlab, SQL, Hive, Pig, Spark

### Skills & Talents

- ✓ Distributed computing
- ✓ Predictive modeling
- ✓ Story-telling and visualizing
- ✓ Math, Stats, Machine Learning

## Salary Trend – Data Scientist Experts

wwwIndeed.com (2013-2014)





# ABOUT THIS COURSE

# More practice than theory

**A very basic course (introductory but not elementary)**

Designed with biology and medical majors in mind.

Theory to bare minimum (only in the context of programming).

**Focus on practical aspects (means programming)**

Adapting to advancement in technologies (Rstudio etc.)

Emphasis on communication, reproducibility and version control (Rmarkdown, Git etc.)

Working with databases

**Focus on biomedical and population health data sets**

# 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 encouraged to know (not essential)

## Data is usually represented as a matrix

A little linear algebra will go a long way (you may want to attend my linear algebra lecture on Sept. 27)

## In data science, most problems are vague

Probability, statistics & machine learning are very important

## Fundamental concepts in algorithms and data structures

Big data requires non-trivial data structures and algorithms (you may want to attend my algorithms lecture on Sept. 25)

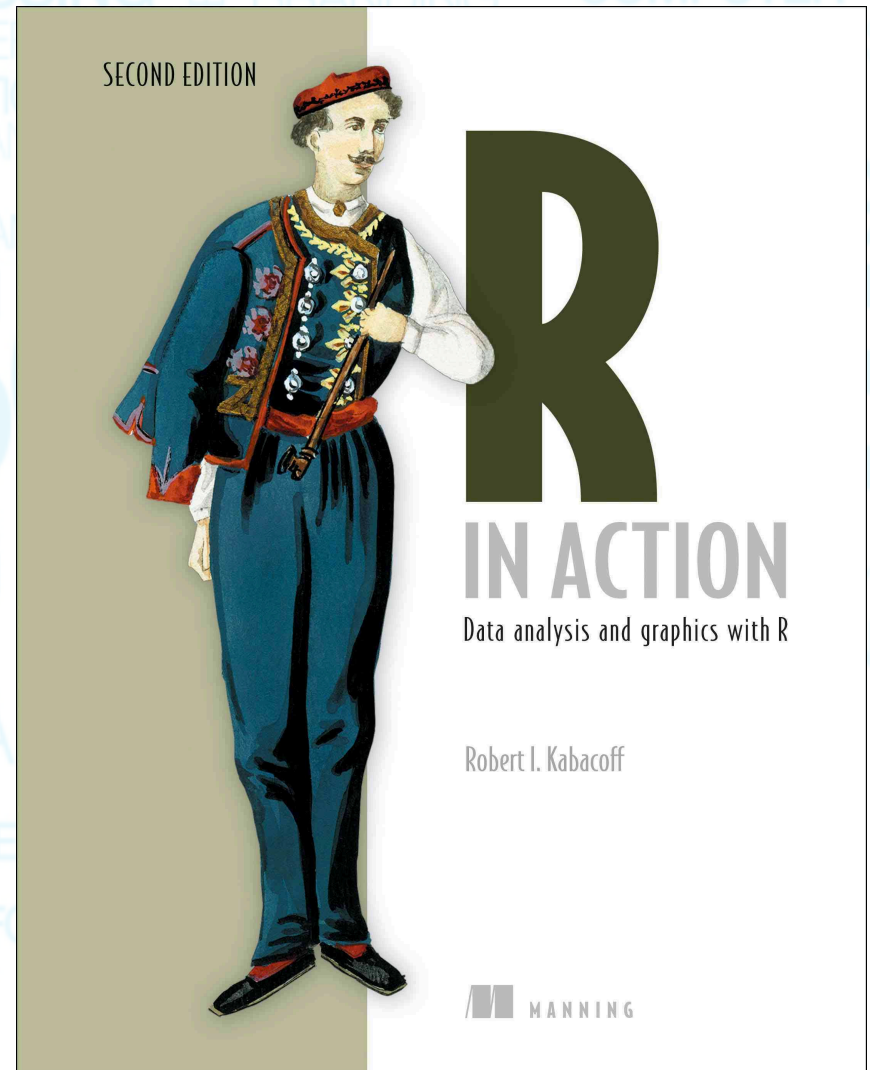
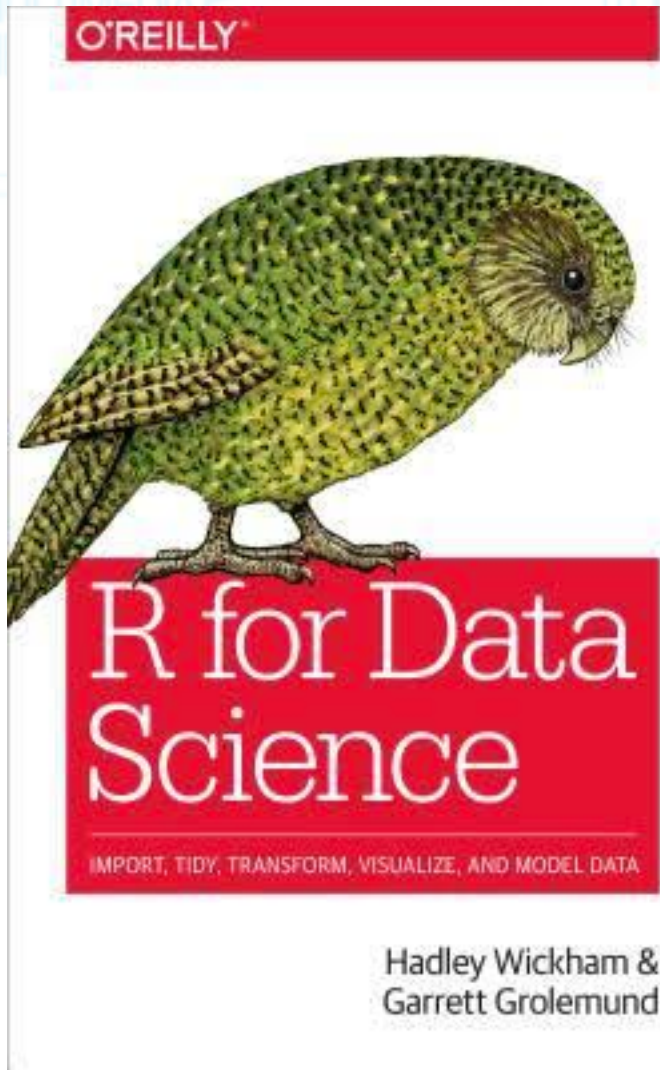
## High-performance computing

We have a very good facility here for HPC

## Genomics/Epigenetics

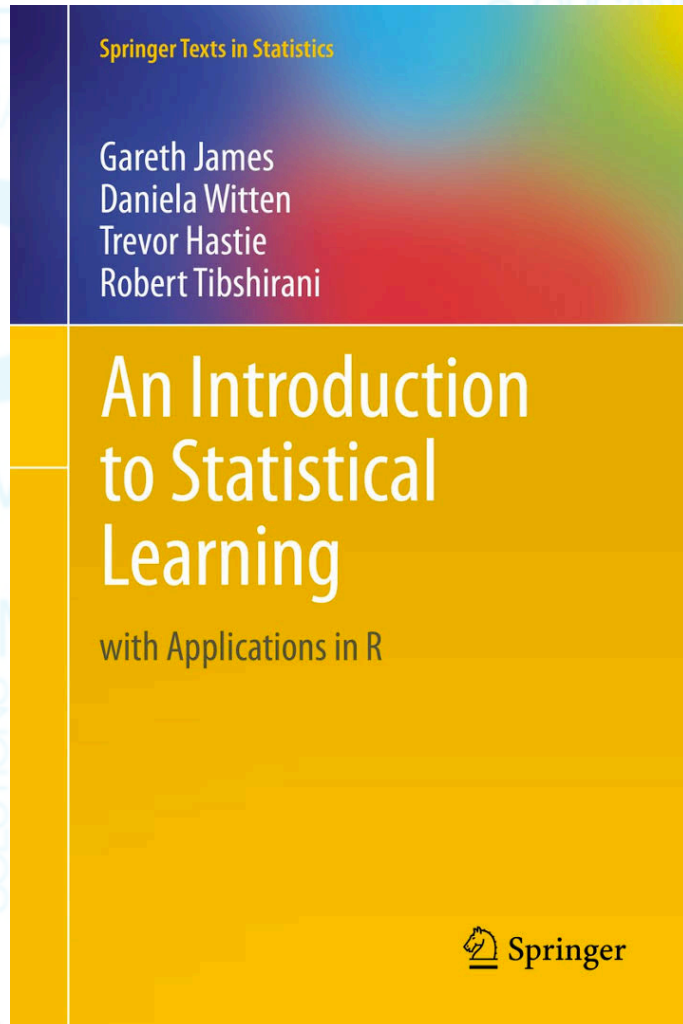
Io(M)T will integrate genomics/epigenetics as we progress

# Textbooks (recommended)

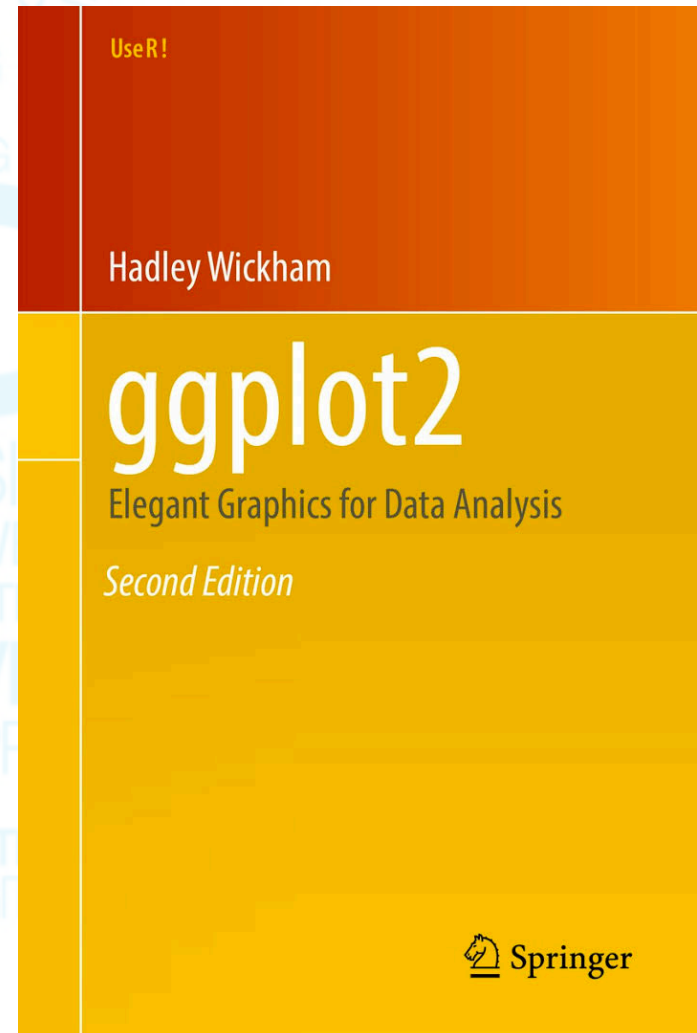


# Other selected data science books

## Machine Learning



## Visualization



# Let's rock and roll...

- Sign-in sheet
- Registered vs. non-registered students
- Course webpage: <https://kannan-kasthuri.github.io/pda.html>
- GitHub registration
- Email associated with GitHub account
- First homework assigned (due Sept. 27)