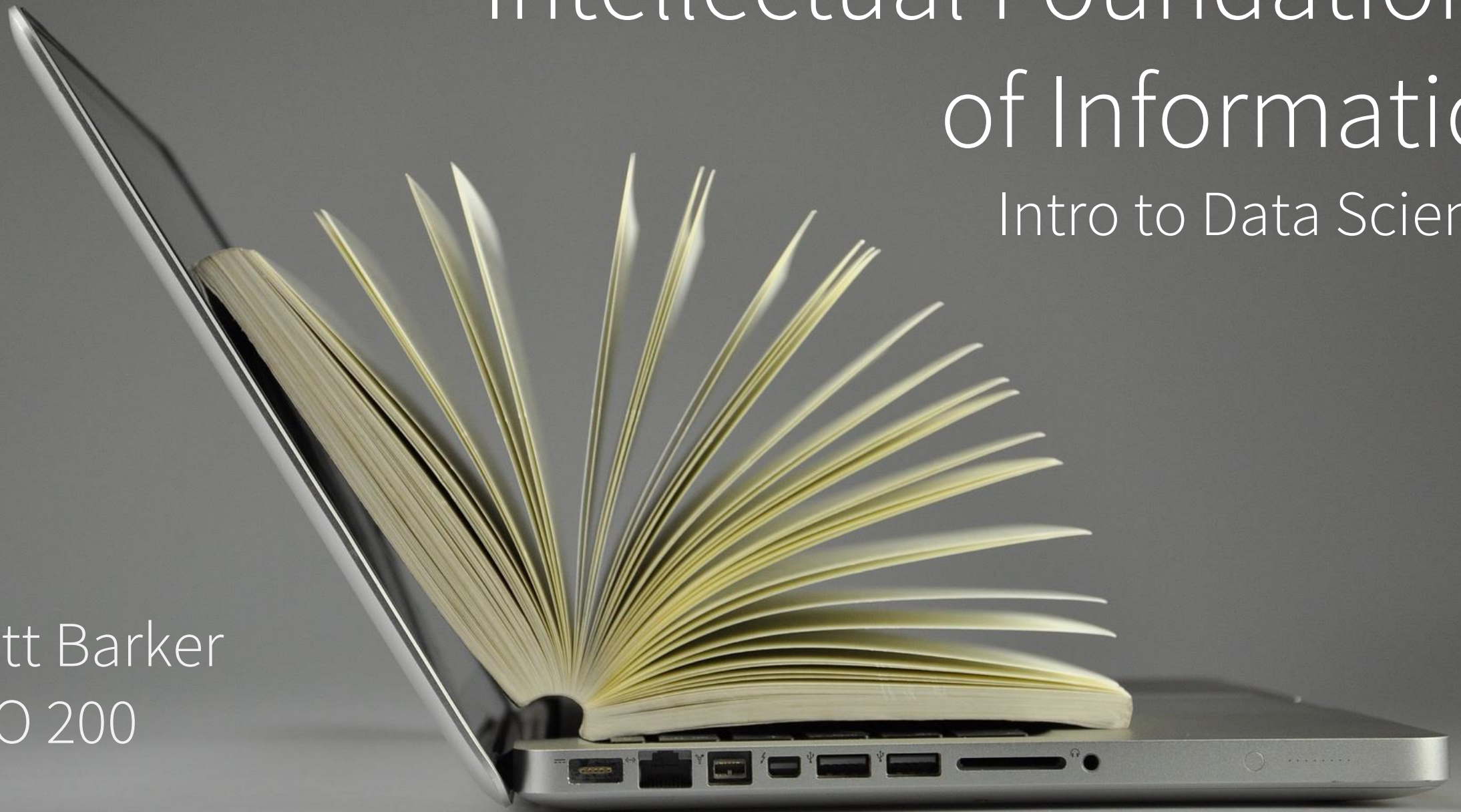
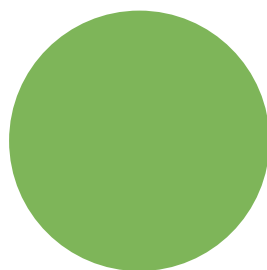


Intellectual Foundations of Informatics

Intro to Data Science

Scott Barker
INFO 200





DETECTION
MULTIMEDIA
NETWORK
PREDICTIVE
PROGRAM
ANALYTICS
MACHINE LEARNING
VISION
ENGINEERING
RESEARCH
PROBABILITY
COMPUTING
BIG DATA
STATISTICS
TARGET
INFORMATION
DIGITAL
CODING
SEGMENTATION
SOCIAL NETWORKS
SOCIAL MEDIA
SERVICES
PROJECTS
CONTENT
CONSUMER
ORGANIZATION
PLANNING
EVENTS
PROGRAMMING
MODELS
BRANDING
CONSUMER DEMAND MARKETS
WEB MARKETING
DATA MINING
DATA
BIG
KDD
WORLDWIDE
PRICING
STRATEGY
WEB DEV
SERVICE
PRO
MOBILE
INFORMATION
DIGITAL
CODING
SEGMENTATION
SOCIAL NETWORKS
BIG DATA
COMPUTING
PROMOTION
TECHNOLOGY
E-MARKETING
COMMUNICATION
COMPUTER

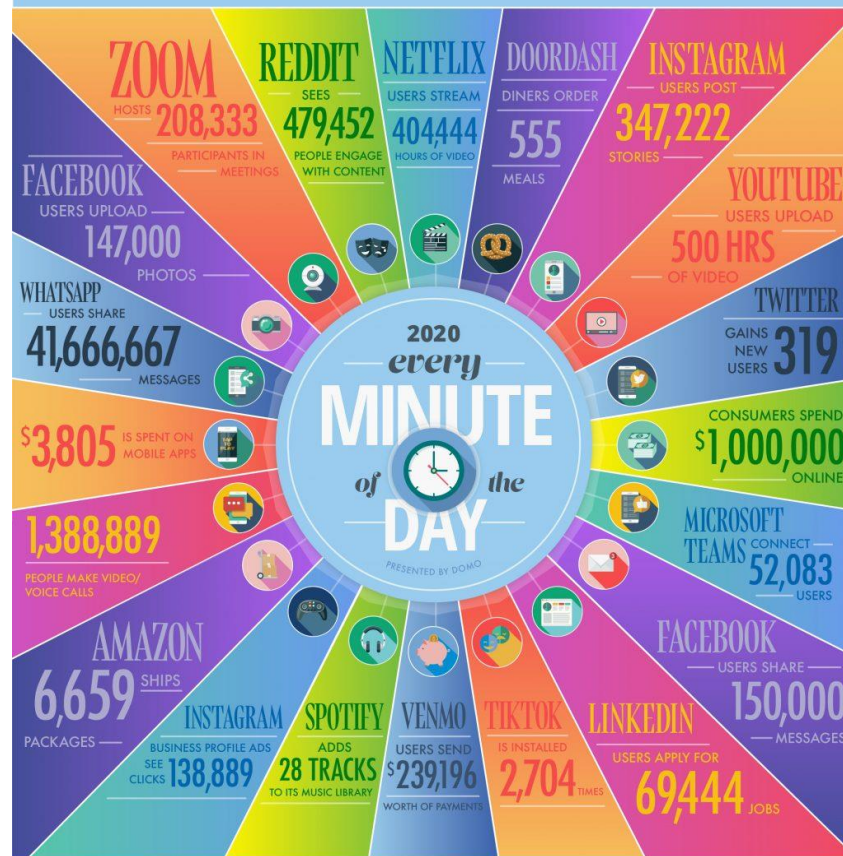
DATA SCIENCE

DOMO

DATA NEVER SLEEPS 8.0

How much data is generated *every minute*?

In 2020, the world changed fundamentally—and so did the data that makes the world go round. As COVID-19 swept the globe, nearly every aspect of life—from work to working out—moved online, and people depended more and more on apps and the Internet to socialize, educate and entertain ourselves. Before quarantine, just 15% of Americans worked from home. Now over half do. And that's not the only big shift. In our 8th edition of Data Never Sleeps, we bring you the latest stats on how much data is being created in every digital minute—a trend that shows no sign of stopping.



The world's internet population is growing significantly year over year. As of April 2020, the internet reaches 59% of the world's population and now represents 4.57 billion people — a 6% increase from January 2019.



GLOBAL INTERNET POPULATION GROWTH 2014–2020
(IN BILLIONS)

As the world changes, businesses need to change with the times—and that requires data. Every click, swipe, share or like tells you something about your customers and what they want, and Domo is here to help your business make sense of all of it. Domo gives you the power to make data-driven decisions at any moment, on any device, so you can make smart choices in a rapidly changing world.

Learn more at domo.com

SOURCES: STATISTA, VITAL CAPITALIST, BUSINESS INSIDER, GAMASPOT, TECHCRUNCH, OMNICORE AGENCY, DOORDASH, BUSINESS OF APPS, NEW YORK TIMES, MUSIC BUSINESS WORLDWIDE, INC., THE VERGE, INC., HOODPULSE, BUSTON STORY, REDDIT, YOUTUBE, AMAZON, VOA

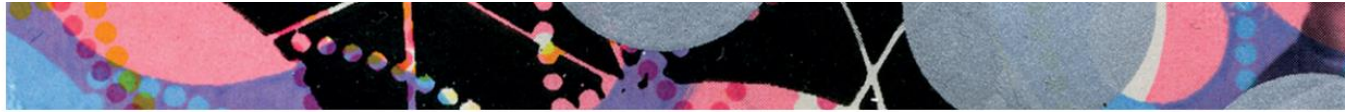


“Data is the new oil”

Clive Humby
Chief Scientist, Starcount



Why this analogy?



DATA

Data Scientist: The Sexiest Job of the 21st Century

by **Thomas H. Davenport** and **D.J. Patil**

FROM THE OCTOBER 2012 ISSUE



SUMMARY



SAVE



SHARE



COMMENT



TEXT SIZE



PRINT

\$8.95 BUY COPIES

When Jonathan Goldman arrived for work in June 2006 at LinkedIn, the business networking site, the place still felt like a start-up. The company had just under 8 million accounts, and the number was growing quickly as existing members invited their friends and colleagues to join. But users weren't seeking out connections with the people who were already on the site at the rate executives had expected. Something was apparently missing in the social experience. As one LinkedIn manager put it, "It was like arriving at a conference reception and realizing you don't know anyone. So you just stand in the corner sipping your drink—and you probably leave early."

Data scientist

Explore career information by location

United States

📍

Search

- Overview
- Salaries
- Career advice
- Jobs
- Companies
- Questions
- Articles

How much does a Data Scientist make in the United States?

Per year

▼

4,123 salaries reported ?

Average base salary

\$123,716

per year

The average salary for a data scientist is \$123,716 per year in the United States.

Most common benefits

💰 Stock options

🍴 Food provided

🚌 Commuter assistance

🛂 Green card sponsorship

💪 Gym membership

[View more benefits](#)

This is all great, but what exactly is “Data Science?”



“**Data Science** refers to an emerging area of work concerned with the collection, preparation, analysis, visualization, management and preservation of large collections of information”

Jeffrey Stanton

Syracuse University School of Information Studies



“**Data Science** is a set of methods for answering questions and making decisions based on heterogeneous data”



Josh Blumenstock
UC Berkeley Information School
(formerly UW Information School)

“**Data Science** is about answering questions using large, noisy, and heterogeneous datasets”

Bill Howe

UW Information School, UW eSciences Institute



60 second video – What is a Data Scientist?

Insight



What is an insight?

From dictionary.com

“an instance of apprehending the true nature of a thing, especially through intuitive understanding;

an understanding of the motivational forces behind one's actions, thoughts, or behavior”

Assume you are the iSchool Dean, the Chair of Informatics, or a student interested in taking INFO 200

What are a couple valuable hypothetical “insights” we might be able to discover about this course if we acted like a data scientist?

The Data Science Method

1. Start with a question
2. Gather, clean, restructure, transform, filter, load, integrate, and verify the data necessary to answer that question
3. Design and run a statistical model that can answer your question
4. Interpret and communicate the results, noting the limitations of your conclusions



80% of the work



The other 80% of the work

“Data Science is Statistics re-branded”

Common statement by some that work in Statistics!

Our view is that statistics is very important, but Data Scientists need to know much more than just statistics including information visualization, data ethics, and how to communicate results to stakeholders

What do Data Scientists do?

- Most work to solve problems with high impact for their organizations or for society. Useful for companies, government organizations, scientific research, non-profits. **Having domain specific knowledge is typically very helpful.**
 - There is lots of data to potentially analyze, knowing the important questions to ask or how that data can have an impact or make a difference is key
 - One good source of US Government data is data.gov
- Data Scientists are inquisitive, they like asking questions.
- Data Scientists often explore data from different sources, not just one. Often they use existing data.
- Data Scientists must be able to communicate findings well to leaders to help them make decisions or the public to inform policy. Data/Information Visualization is one way to do this.
- Real-time Information “Dash boards” may frequently be used to display/report information to users
- Data Scientists are often able to spot trends or patterns. This might be done using statistical analysis, by building prediction models, or [utilizing machine learning](#) technologies.

Short Intro to Machine Learning Video

Typical activities related to Data Science

- Data Wrangling, Data Munging, Data Jujitsu
 - Gathering, extracting, cleaning, and storing data
 - Making sure the data is correct/accurate
 - Pulling data from multiple sources, in multiple formats or with multiple representations together and making it ready for analysis
 - Knowledge of data file formats and how to convert between them – e.g. XML, HTML, CSV files
 - Some estimates say up to 80% of what Data Scientists do is “data wrangling”
- Data Analytics
 - Using Statistical Analysis, Statistical modeling, or machine learning to gain insight
 - Machine learning can be supervised where the machine is “taught” based on existing data and outcomes, or unsupervised where the machine uses algorithms to look for patterns to discover previously unknown relationships
 - Don’t forget – “Garbage in, Garbage out!” – video in a sec...
- Data Management
 - Representing and storing the data in some system (for example a relational database or NoSQL)
 - Ensuring the system has capacity or can scale in terms of processing/storage
 - Ensuring data is backed-up, properly retained, secured/protected

Garbage In, Garbage Out

“Calling Bullshit” Video

MODERN DATA SCIENTIST

Data Scientist, the sexiest job of 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 package 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



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

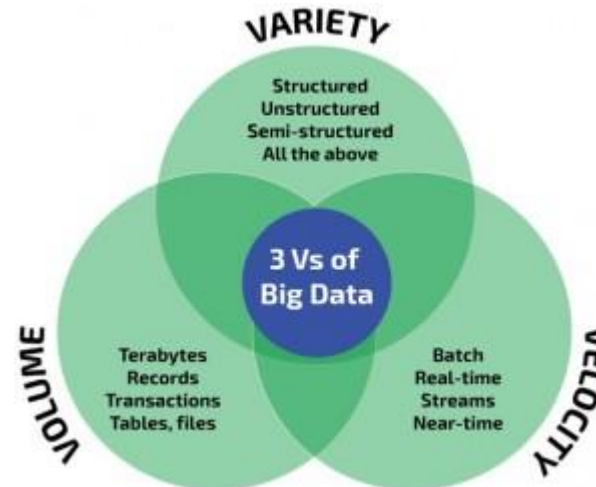
How is “Big Data” related to Data Science?

The three V's:

Volume – amount, cannot be stored or analyzed using traditional means – typically requiring multiple machines, high-end storage, or special techniques to distribute the processing

Velocity – pace of change is much faster. Data can change in seconds or less

Variety – data could be structured, semi-structured, or unstructured - can't pre-define what it looks like in advance

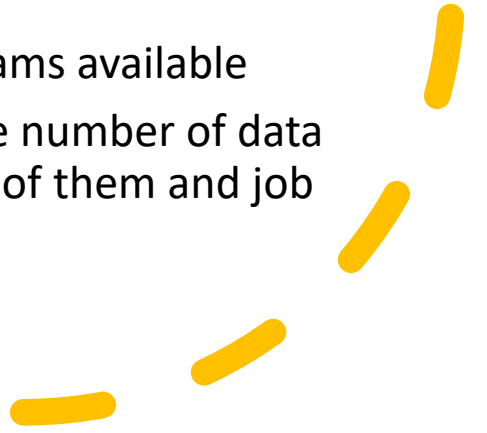


Data Science tools to potentially learn

- Excel – easy starting place, can do the basics
- [Tableau Desktop](#) – [free for students](#)
- [Microsoft Power BI](#) – can start for free
- [Azure Machine Learning Studio](#)
- Relational databases – SQL Server, MySQL, Oracle, Postgress – note Microsoft Access is not considered a “serious” Data Science/DBMS tool by most
- NoSQL databases – for unstructured data, example MongoDB
- Statistical Analysis software - **R and RStudio**, SPSS, SAS
 - R is the leading tool for data analysis, thousands of user contributed packages to extend like functionality such as ggplot2, open source, easy to share scripts with others
- Programming – Python and Pandas (Python Data Analysis Library), Java, Javascript (including the D3 Javascript Library for interactive visualizations)
- Amazon or Microsoft Azure Cloud services for storage and compute
- Social Networking API's – such as Twitter



Studying Data Science

- Data Science undergrad option in iSchool
 - INFO 201 now is a good intro
New INFO 180 Intro to Data Science
 - INFO 270 Data Reasoning in a Digital World
(formerly INFO 198 Calling Bullshit in the Age of Digital Data)
 - Data Science undergrad option in CSE
 - Data Science undergrad option from HCDE
 - ACMS – Applied and Computational Math Sciences major
 - Statistics major
 - UW eScience institute – many disciplines represented such as physics, astronomy, math
 - New Data Science Minor at UW now approved
 - Multiple grad programs, e.g. iSchool MSIM program, other certificates available
 - Outside UW, many code camps and for-profit programs available
 - Note: graduate degrees are often preferred, and the number of data science programs are exploding – may be too many of them and job prospects not as certain as a result
- 

End Part 1