

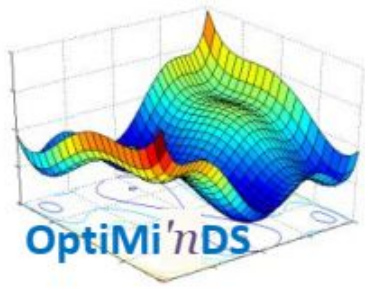
# Data Science

Dr. Irfan Younas

1

## Introduction to Data Science

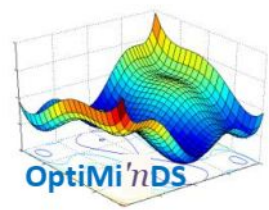
Lecture 1 – Spring 2021



# Optimization and Data Science (OptiMi'nDS)

2

- ❑ **Research Themes:** Evolutionary Computation, Swarm Intelligence, Evolutionary Deep Learning, Multi/Many-objective Optimization, Artificial Intelligence, Data Science, Machine Learning, Natural Language Processing and Information Retrieval
- ❑ **Past and Ongoing Projects (<http://lhr.nu.edu.pk/fsc/research/>)**
  - ❑ Designing novel Socio-inspired Optimization Algorithms for Global Optimization
  - ❑ Developing Transfer Learning Based Classifier System for Image Classification
  - ❑ Evolving Deep Neural Networks using Evolutionary Computation
  - ❑ Automatic Cardiac Segmentation using Metaheuristics and Active Contours
  - ❑ Large scale optimization of Assignment, Planning and Scheduling Problems
  - ❑ Distributed Large Scale Many Objective Optimization
  - ❑ Multi and Many Objective Optimization Algorithms
  - ❑ Many Objective Optimization for IoT
  - ❑ Learning Regular Expressions using Learning Classifier Systems
  - ❑ Solving Large-scale Optimization Problems using Evolutionary Computation and Machine Learning
  - ❑ Solving Classification and Learning Problems using Evolutionary Machine Learning
  - ❑ Predicting Future News Events and Crimes using Data Science



## Recent Publications

3

- Saba Kanwal<sup>+</sup>, Irfan Younas, and Maryam Bashir. "Evolving Convolutional Autoencoders Using Multi-Objective Particle Swarm Optimization." *Computers and Electrical Engineering* [Special issue: Deep Learning-based Intelligent Systems: Theories, Algorithms, and Applications] (2021): Accepted. (**Impact Factor = 2.663**)
- Qamar Askari<sup>+</sup>, and Irfan Younas. "Political Optimizer Based Feedforward Neural Network for Classification and Function Approximation." *Neural Processing Letters* (2021): 1-30. (**Impact Factor = 2.891**)
- Shah Bano<sup>+</sup>, Maryam Bashir, and Irfan Younas. "A Many-Objective Memetic Generalized Differential Evolution Algorithm for DNA Sequence Design." *IEEE Access* 8 (2020): 222684-222699. (**Impact Factor = 3.745**)
- Qamar Askari<sup>+</sup>, Irfan Younas, and Mehreen Saeed. "Political Optimizer: A novel socio-inspired meta-heuristic for global optimization." *Knowledge-Based Systems* (2020): 105709. (**Impact Factor = 5.921**)
- Qamar Askari<sup>+</sup>, Mehreen Saeed, and Irfan Younas. "Heap-based optimizer inspired by corporate rank hierarchy for global optimization." *Expert Systems with Applications* (2020): 113702. (**Impact Factor = 5.452**)
- Irfan Younas<sup>+</sup>, Uzman Perwaiz<sup>++</sup>, and Adeem Ali Anwar<sup>+</sup>. "Many-objective BAT algorithm." *Plos one* 15, no. 6 (2020): e0234625. (**Impact Factor = 2.740**)
- Qamar Askari<sup>+</sup>, Irfan Younas, and Mehreen Saeed. "Critical evaluation of sine cosine algorithm and a few recommendations." In *Proceedings of the 2020 Genetic and Evolutionary Computation Conference Companion*, pp. 319-320. 2020. (**CORE rank A**)
- Adeem Ali Anwar<sup>+</sup>, and Irfan Younas. "Optimization of Many Objective Pickup and Delivery Problem with Delay Time of Vehicle Using Memetic Decomposition Based Evolutionary Algorithm." *International Journal on Artificial Intelligence Tools* 29, no. 01 (2020): 2050003. (**Impact Factor = 0.689**)
- Hafiz Asadul Rehman<sup>+</sup>, Muhammad Iqbal, Irfan Younas, and Maryam Bashir. "Learning Regular Expressions Using XCS-Based Classifier System." *International Journal of Pattern Recognition and Artificial Intelligence* (2019): 2051011. (**Impact Factor = 1.375**)
- Irfan Younas, Farzad Kamrani, Maryam Bashir, and Johan Schubert. "Efficient genetic algorithms for optimal assignment of tasks to teams of agents." *Neurocomputing* 314 (2018): 409-42 (**Impact Factor = 4.438**)

## □ Top Journals

- IEEE Transactions on Evolutionary Computation
- Evolutionary Computation
- IEEE Computational Intelligence Magazine
- Knowledge-based Systems
- Expert Systems with Applications
- Applied Soft Computing
- Swarm & Evolutionary Computation
- Neurocomputing
- there are other related journals also (like soft computing, computational intelligence, European journal of operational research, Computers and Operations Research etc.)

## □ Top conferences

- Genetic and Evolutionary Computation Conference (GECCO)
- IEEE Congress on Evolutionary Computation

# Outline

5

- What?
- Why?
- How?

# What is Data Science?

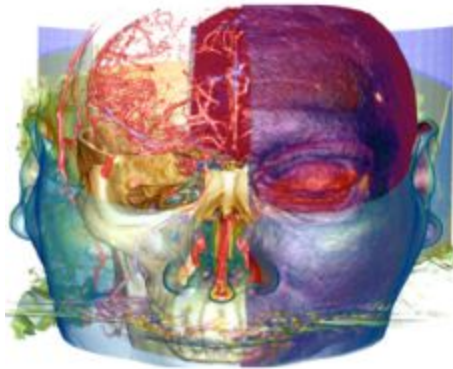
6

- Like any emerging field, it isn't yet well defined, but incorporates elements of:
  - Data Wrangling
  - Exploratory Data Analysis and Visualization
  - Machine Learning and Statistics
  - High-Performance Computing technologies for dealing with scale.

# What is Data Science?

7

- To gain insights into data through computation, statistics, and visualization



- New technology makes it possible to capture vast amounts of logging / sensor data.
- Computing advances make it possible to analyze data on ever increasing scales.
- Prominent role models (Google, Nate Silver, ...) have proven the power of modern data analytics.



# A Data Scientist is ...

9

- “A data scientist is someone who knows more statistics than a computer scientist and more computer science than a statistician.”

- Josh Blumenstock

- “Data Scientist = statistician + programmer + coach + artist”

- Shlomo Aragon

# Genius vs Wisdom?

10

- Software developers are hired to produce code.
- Data Scientists are hired to produce insights.
- Genius shows in finding the right answer!!!
- Wisdom shows in avoiding the wrong answers.
- Data science (like most things) benefits more from wisdom than from genius.

# Asking Good Questions

11

- Software developers are not encouraged to ask questions, but data scientists are:
- What exciting things might you be able to learn from a given data set?
- What things do you/your people really want to know?
- What data sets might get you there?

# Baseball Questions

12

- How to best measure individual player's skill, value or performance?
- How fair do trades between teams work out?
- What is the trajectory of player's performances as they mature and age?
- To what extent does batting performance correlate with the position played?

# Nate Silver



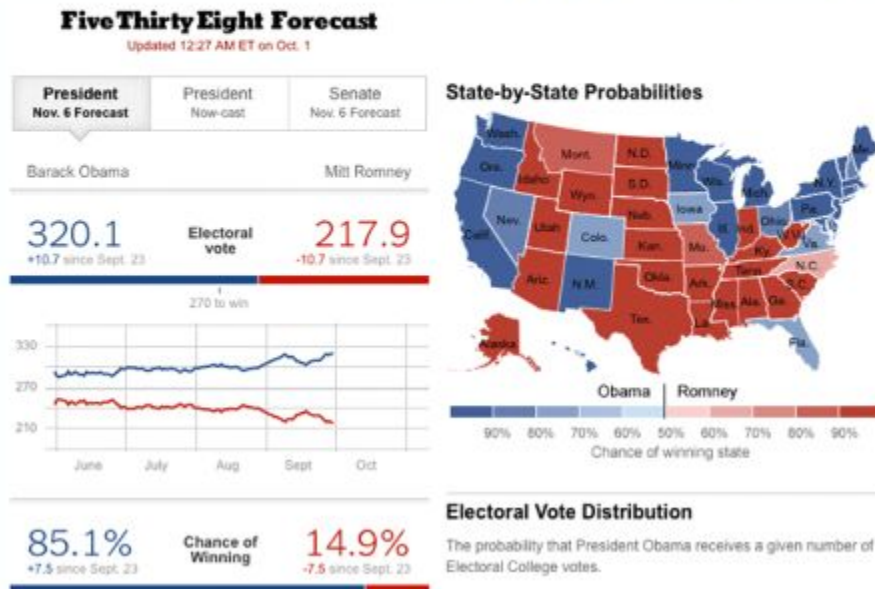
# Nate Silver

14

- American Data Scientist who analyzes elections and baseball.
- PECOTA: a system for forecasting the performance and career development of Major League Baseball players.
- 2008 U.S. Presidential election
  - successfully called the outcome of 49 of the 50 states.

- 2012 U.S. Presidential election
  - Correctly predicted the winners of all the states.

“Nate Silver won the election”  
– Harvard Business Review



## Is Election Predictor Nate Silver A Witch? Probably. And Quantified Self Data Will Make You One Too



JOSH CONSTINE

Wednesday, November 7th, 2012

7 Comments



Scientists are yesterday's wizards and demigods. And Nate Silver is a scientist. One whose ability to **predict the outcome of elections** is so precise, it's nearly indistinguishable from magic. That's why **isNateSilverAWitch.com** is so funny. But really what his flawless prediction of the presidential election signifies is the coming of age of the quantified universe.

<http://techcrunch.com/2012/11/07/nate-silver-as-software/>

---



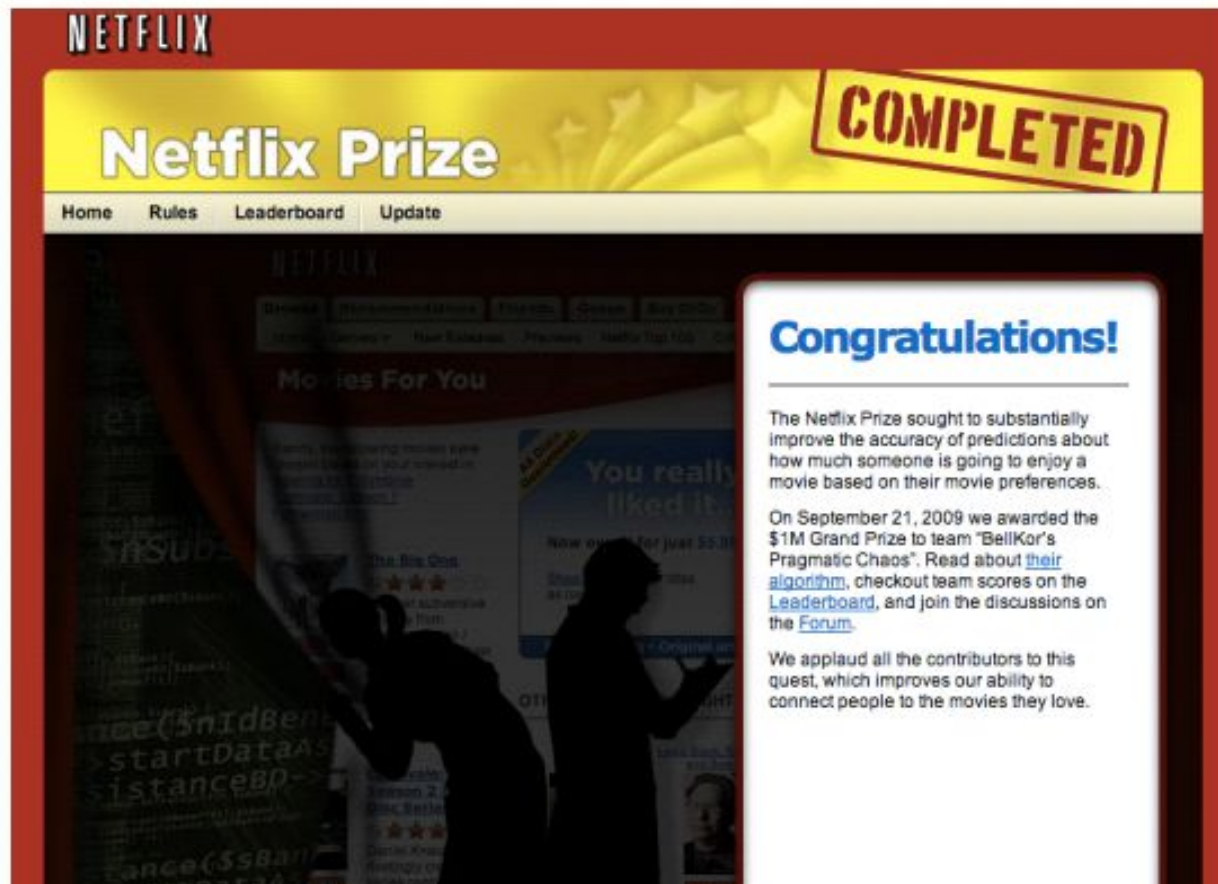
# Some Key Principles

17

- use many data sources
- understand how the data were collected (sampling is essential)
- weight the data thoughtfully (not all polls are equally good)
- use statistical models (not just hacking around in Excel)
- understand correlations (e.g., states that trend similarly)
- have good communication skills (What does a 60% probability even mean? How can we visualize, validate, and understand the conclusions?)

# Netflix Prize

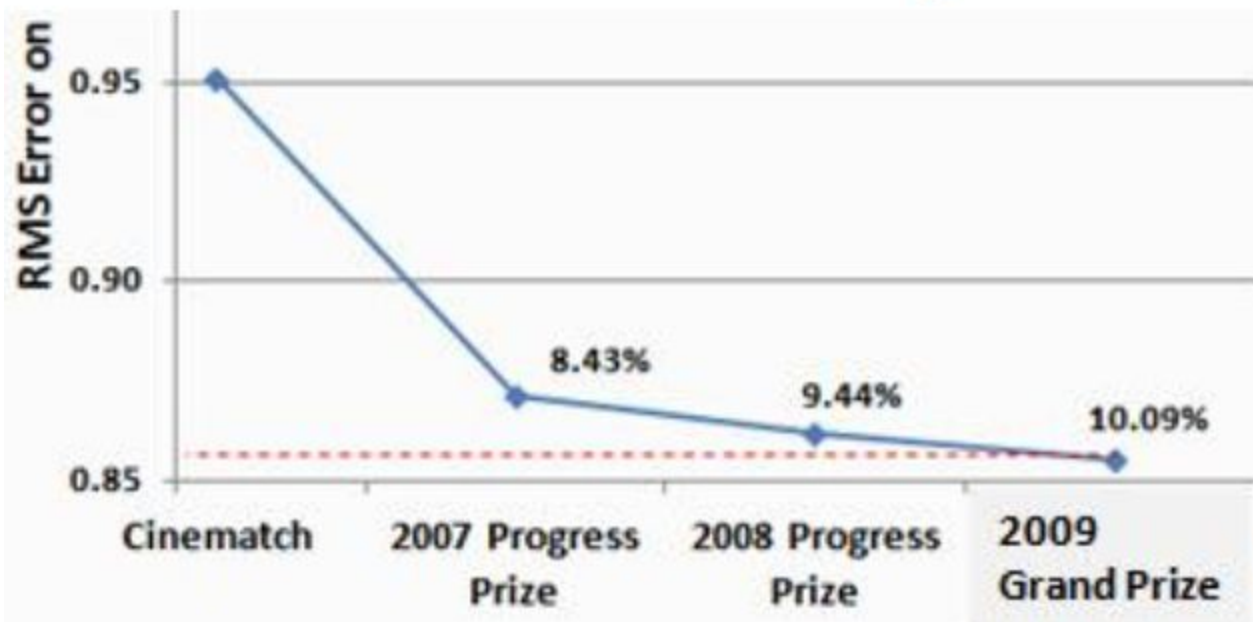
18



# Netflix Prize

19

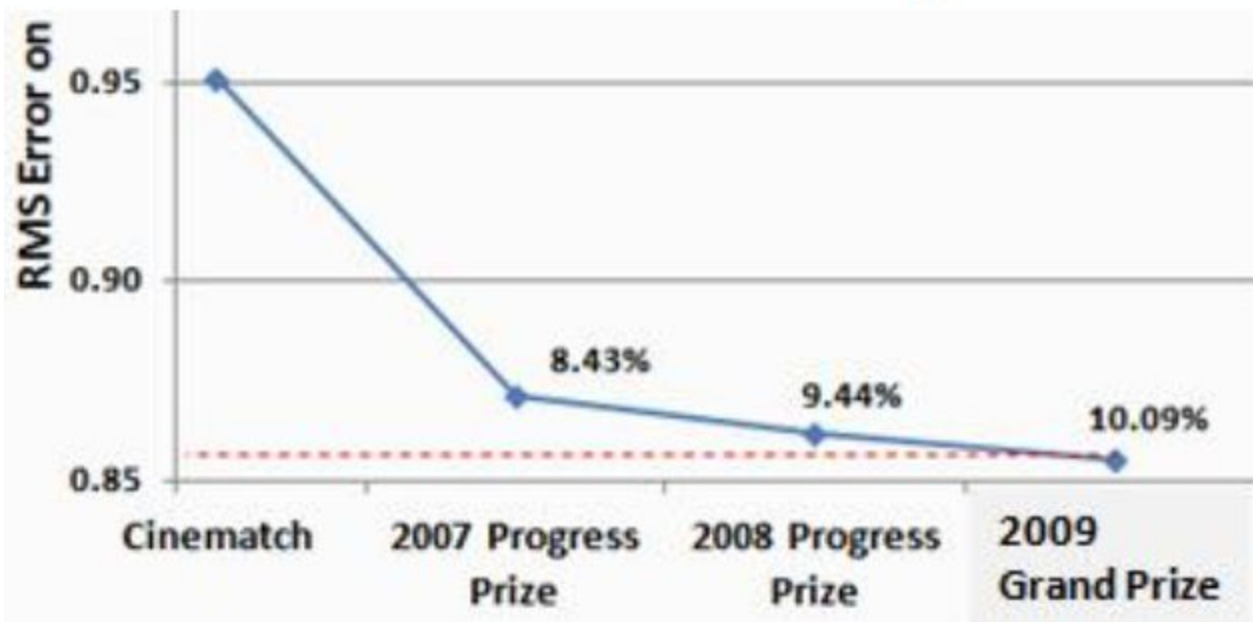
## Netflix Prize Progress



# Netflix Prize

20

## Netflix Prize Progress



# Kaggle

21

← → ↻ 🏠

Secure | https://www.kaggle.com/competitions

🔍 ☆ 🛡️ 📄

☰ kaggle

🕒 Home

🏆 Compete

📊 Data

⏏ Notebooks

💬 Discuss

📖 Courses

💼 Jobs

⌵ More


🔍 Search

Sign In Register

All Competitions

Active Completed InClass

All Categories ▾ Default Sort ▾




OSIC Pulmonary Fibrosis Progression

Predict lung function decline

Featured • a month to go • Code Competition • 1304 Teams

\$55,000




Lyft Motion Prediction for Autonomous Vehicles

Build motion prediction models for self-driving vehicles

Featured • 3 months to go • Code Competition • 189 Teams

\$30,000




Cornell Birdcall Identification

Build tools for bird population monitoring

Research • 16 days to go • Code Competition • 1150 Teams

\$25,000



Google Landmark Recognition 2020

Label famous (and not-so-famous) landmarks in images

\$25,000

# Why do we need Data Science?

22

## □ Big Data

- “Between the dawn of civilization and 2003, we only created five exabytes of information; now we’re creating that amount every two days.”
  - Eric Schmidt, Google (and others)

# Why do we need Data Science?

23

travers808, Visual.ly

## THE BIG V'S OF BIG DATA

*Turning Information Overload Into Big Sales*

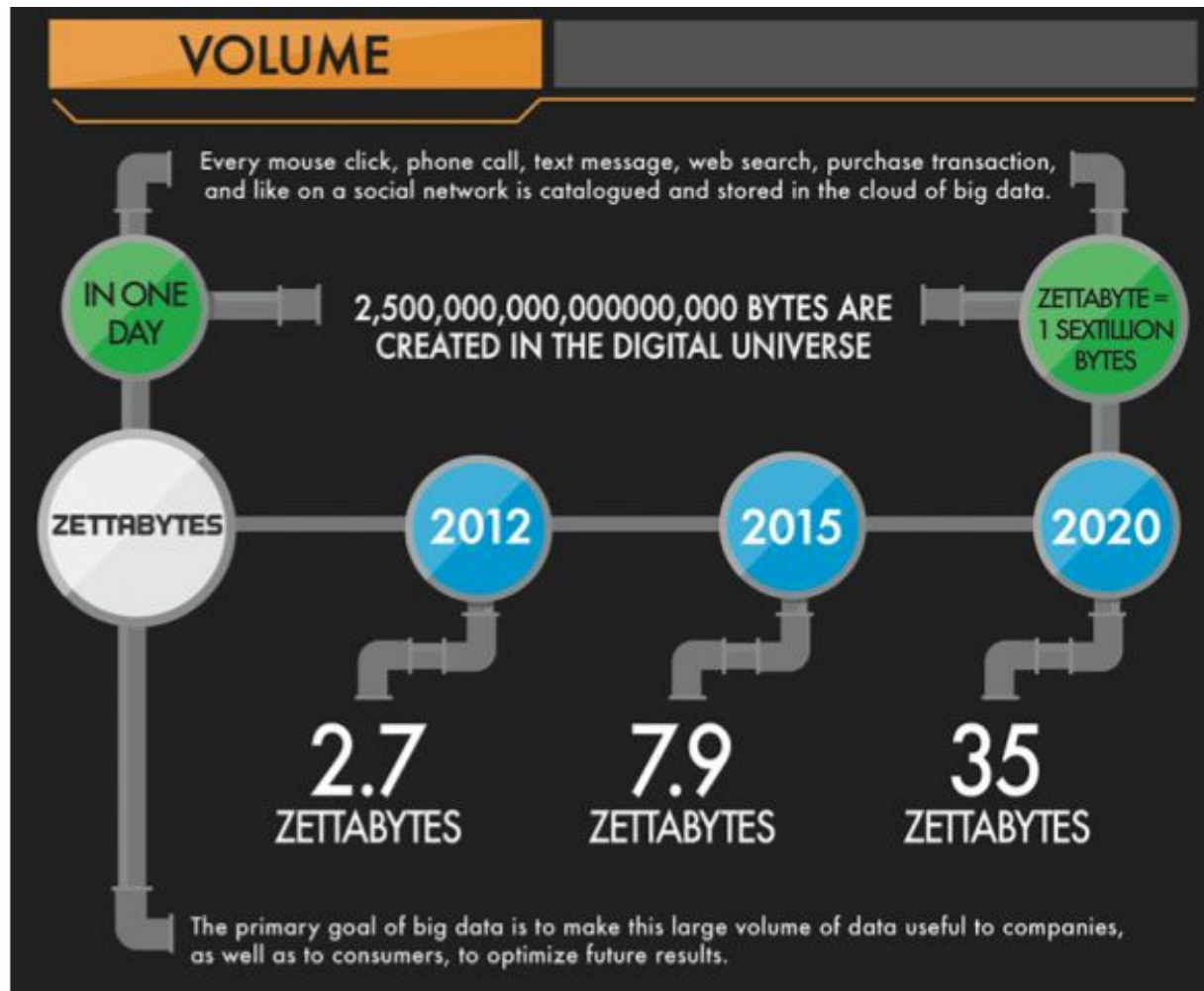
In the emerging market of Big Data, three "V" words have often been used to describe the issues at hand with information overload in our digital world.

### THE EXISTING V'S

Big data has brought both great opportunity and change to the technological industry. Data scientists traditionally look at the existing V's, the ones that have classically been utilized to understand key variables of any data set.

# Why do we need Data Science?

24





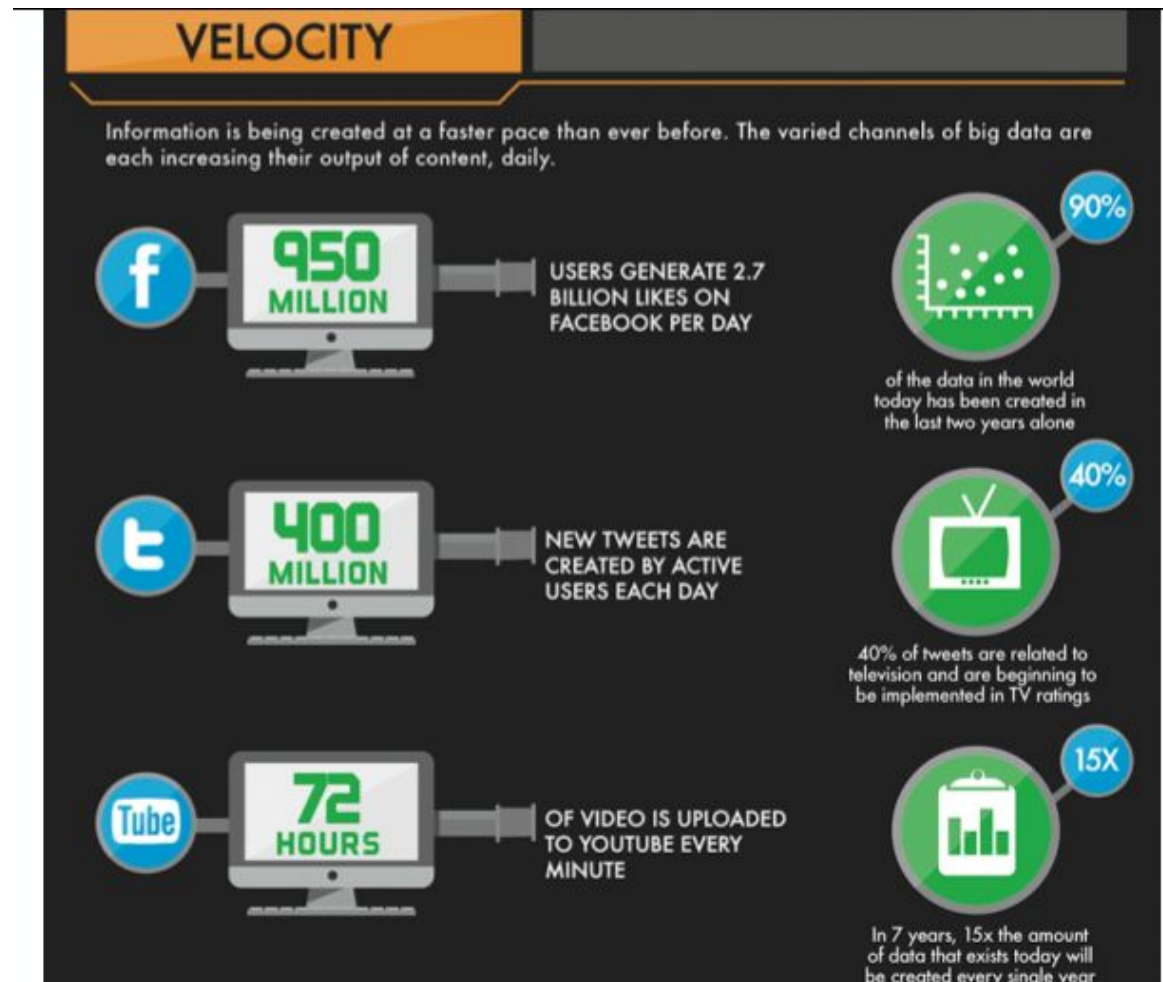
# Why do we need Data Science?

25



# Why do we need Data Science?

26



# Science Paradigms

27

## Science Paradigms

- Thousand years ago:  
science was **empirical**  
*describing natural phenomena*
- Last few hundred years:  
**theoretical** branch  
*using models, generalizations*
- Last few decades:  
a **computational** branch  
*simulating complex phenomena*
- Today: **data exploration** (eScience)  
*unify theory, experiment, and simulation*
  - Data captured by instruments  
or generated by simulator
  - Processed by software
  - Information/knowledge stored in computer
  - Scientist analyzes database/files  
using data management and statistics



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



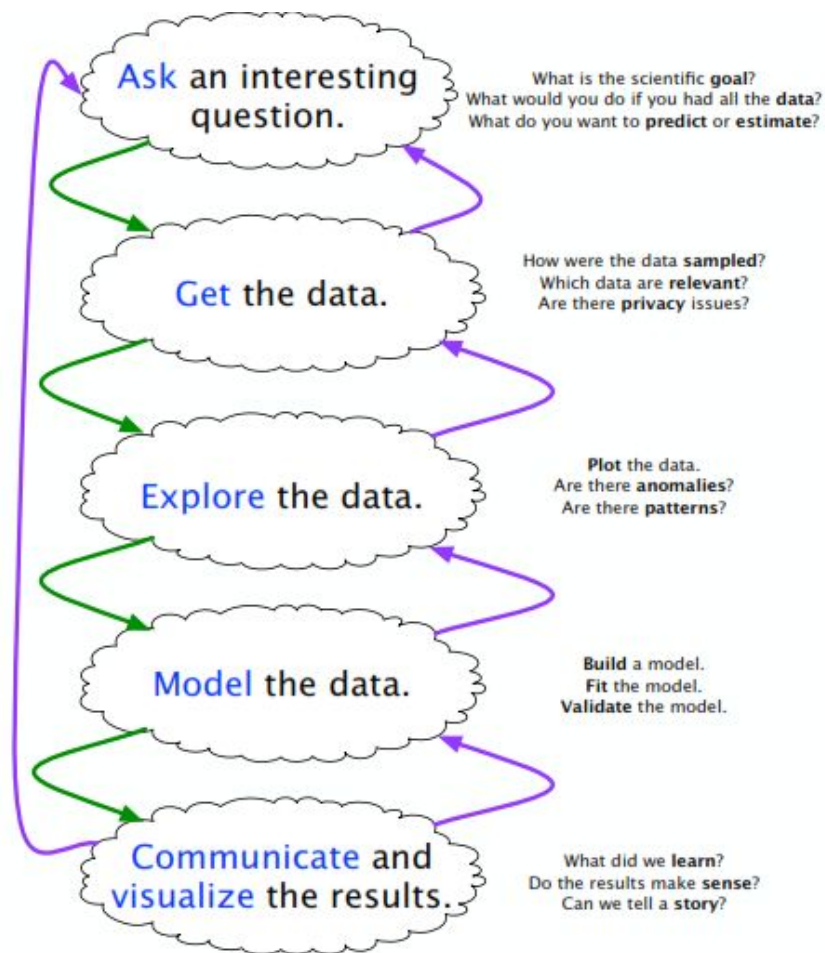
□ The best job in the next 10 years will be Data Scientist. (Not statisticians)

- Hal Varian, (Prof. UC Berkeley, Chief Economist, Google)

# Hal Varian says...

29

- The ability to take **data** – to be able to **understand** it, to **process** it, to **extract value** from it, to **visualize** it, to **communicate** it's going to be a hugely important skill in the next decades, not only at the professional level but even at the educational level for elementary school kids, for high school kids, for college kids. Because now we really do have essentially free **data**.” — Hal Varian



# How?

31

- data wrangling/scraping/sampling/cleaning in order to get an informative, manageable data set;
- Other data preprocessing: Dimensionality reduction, feature subset selection, feature creation etc.
- data storage and management in order to be able to access data quickly and reliably during subsequent analysis;
- exploratory data analysis to generate hypotheses and intuition about the data;
- prediction based on statistical tools such as regression, classification, and clustering;
- communication of results through visualization, and interpretable summaries.

# Language and Tools

32

- Language: Python
- Tools: Jupyter Notebook, NumPy, Pandas

Other packages when needed: SciPy, matplotlib, scikit-learn, and SymPy

See IPython Tutorial: <http://cs231n.github.io/ipython-tutorial/>