SOMAIYA
VIDYAVIHAR UNIVERSITY
K J Somaiya College of Engineering

Introduction to Applied Data Science

Vaibhav P. Vasani

Assistant Professor
Department of Computer Engineering
K. J. Somaiya College of Engineering
Somaiya Vidyavihar University

vaibhav.vasani@gmail.com

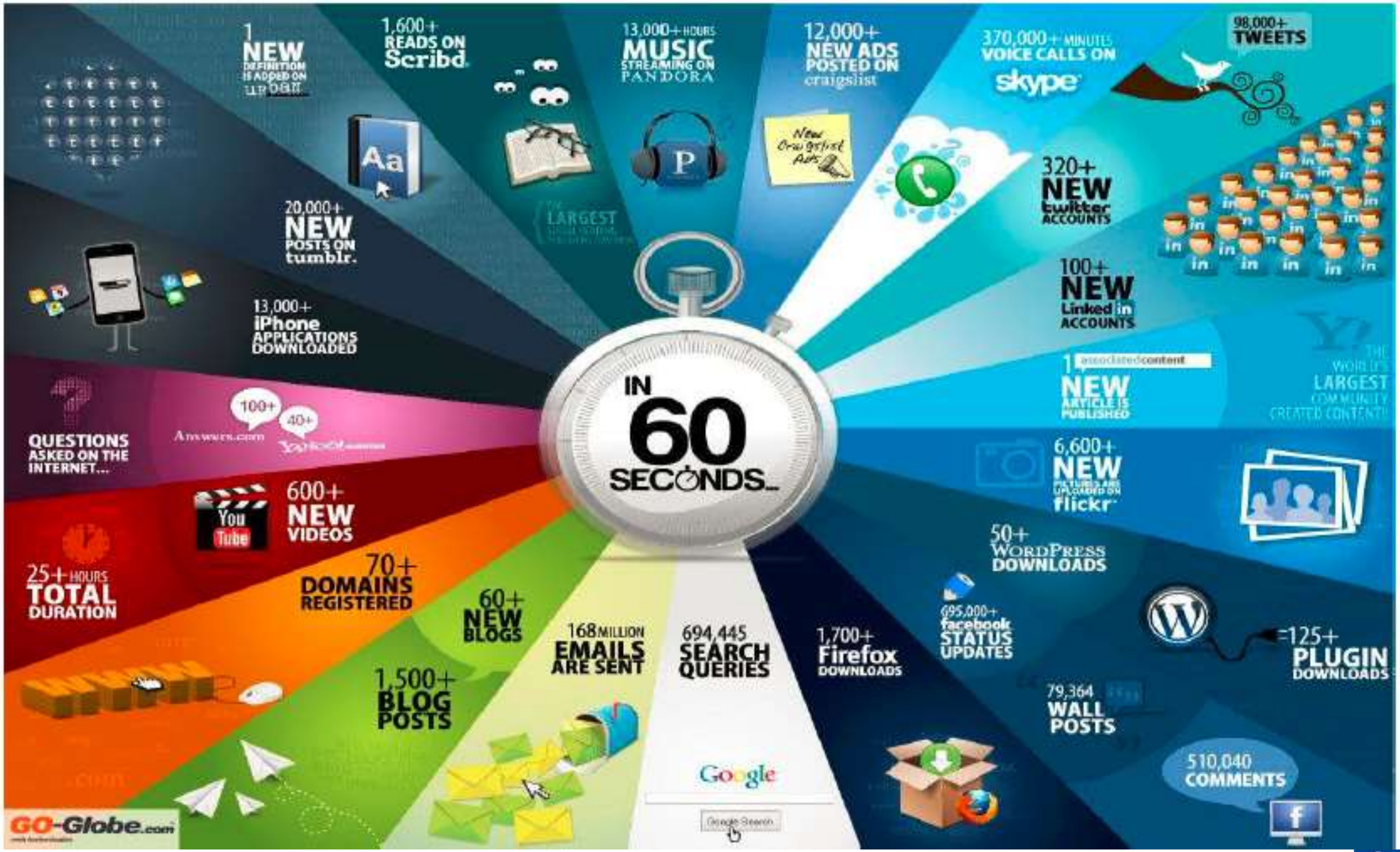
Introduction to Applied Data Science

Outline

- Datafication- Data everywhere
- Big Data
- What is Data Science?
- Big Data and Data Science
- Current landscape of perspectives
- Data Scientist Skill sets
- Challenges and skill Sets needed and various applications areas.
- Impact of applying Data Science in business scenario
- Estimation and validation for added value due to data science

Data is every where

Datafication


1 NEW DEFINITION IS ADDED ON urban
1,600+ READS ON Scribd
13,000+ HOURS MUSIC STREAMING ON PANDORA
12,000+ NEW ADS POSTED ON craigslist
New Craigslist Ads
370,000+ MINUTES VOICE CALLS ON skype
98,000+ TWEETS
20,000+ NEW POSTS ON tumblr.
LARGEST
320+ NEW twitter ACCOUNTS
13,000+ iPhone APPLICATIONS DOWNLOADED
100+ NEW Linked in ACCOUNTS
1 associatedcontent NEW ARTICLE IS PUBLISHED
THE WORLD'S LARGEST COMMUNITY CREATED CONTENT!
QUESTIONS ASKED ON THE INTERNET...
100+ Answers.com
40+ Yahoo! Answers
IN 60 SECONDS
6,600+ NEW PICTURES ARE UPLOADED ON flickr
You Tube
600+ NEW VIDEOS
50+ WordPress DOWNLOADS
25+ HOURS TOTAL DURATION
70+ DOMAINS REGISTERED
60+ NEW BLOGS
168 MILLION EMAILS ARE SENT
694,445 SEARCH QUERIES
1,700+ Firefox DOWNLOADS
695,000+ facebook STATUS UPDATES
125+ PLUGIN DOWNLOADS
79,364 WALL POSTS
1,500+ BLOG POSTS
Google
Google Search
510,040 COMMENTS
GO-Globe.com

Datafication

Definition

- Datification is about taking a process or activity that was previously invisible and turning it into data. That data can then be tracked, monitored, and optimized, leading to new opportunities — and new challenges.

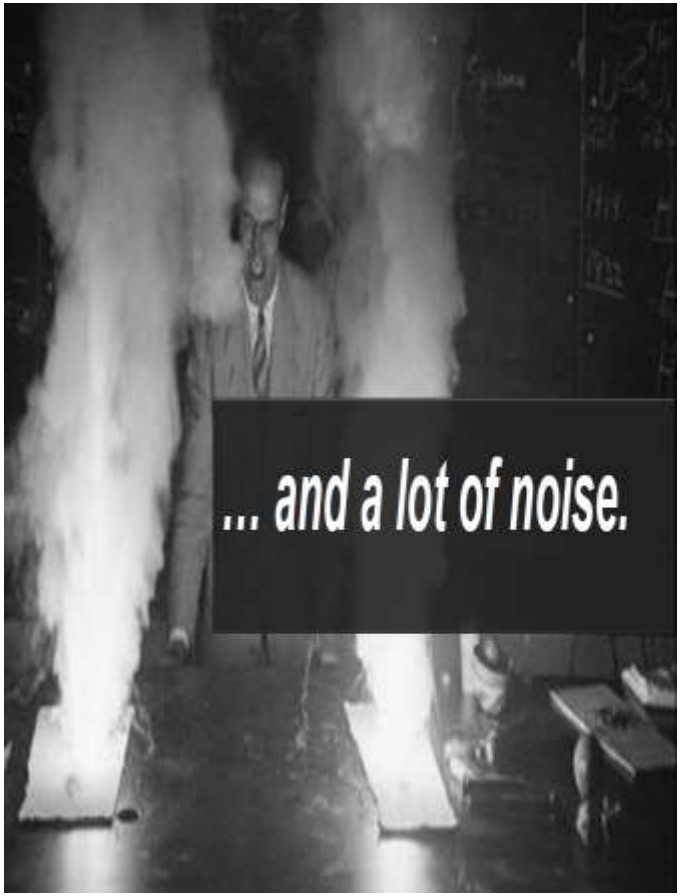
Example

- Quantify friends with ‘likes’
- Googles augmented reality glass to quantify the gaze
- Twitter datify the thoughts
- LinkedIn datify our professional networks
- Browsing web, unintentionaly with cookies
- Walk in store, street we are datafied via sensors, cameras, google glasses
- Taking part of social media experiment

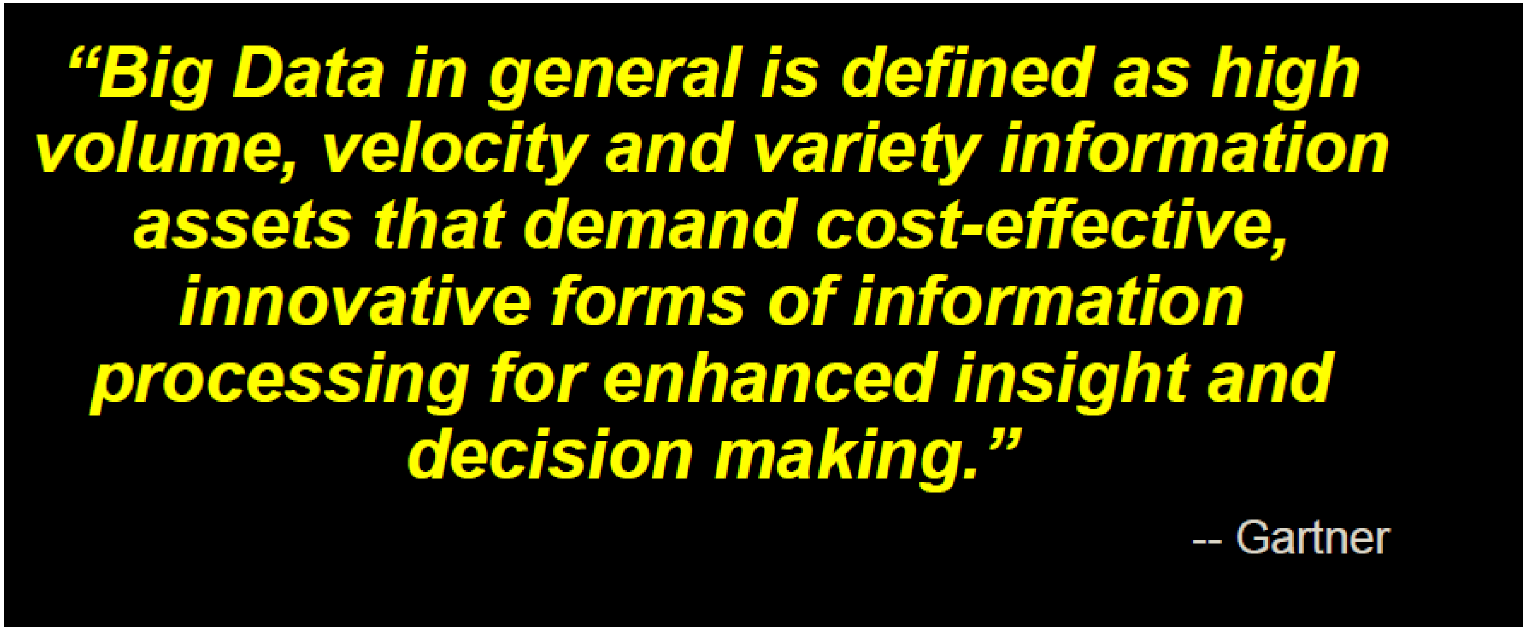
Big Data

Volume
Velocity
Variety

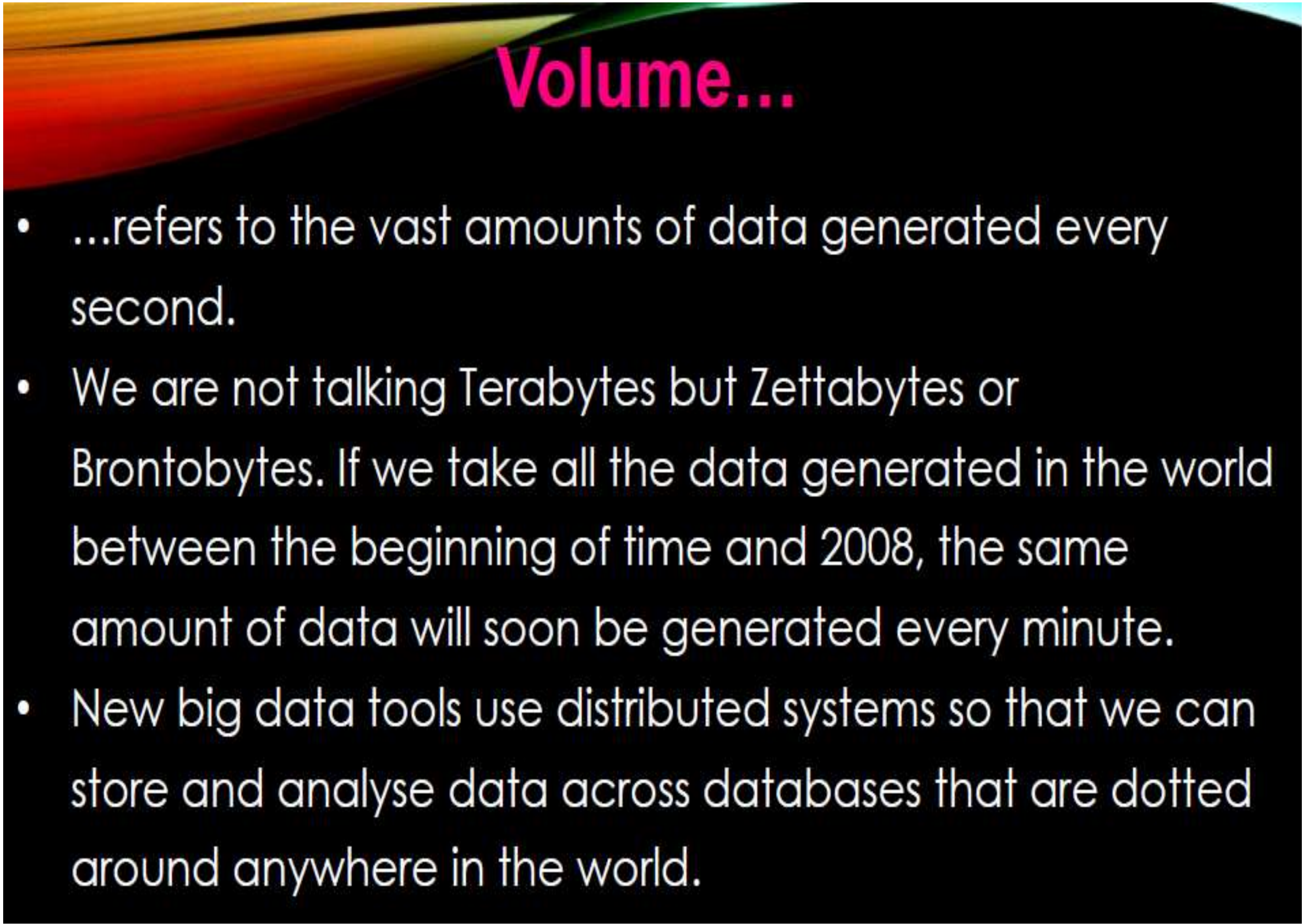
Additional V- Veracity



Big Data Definition

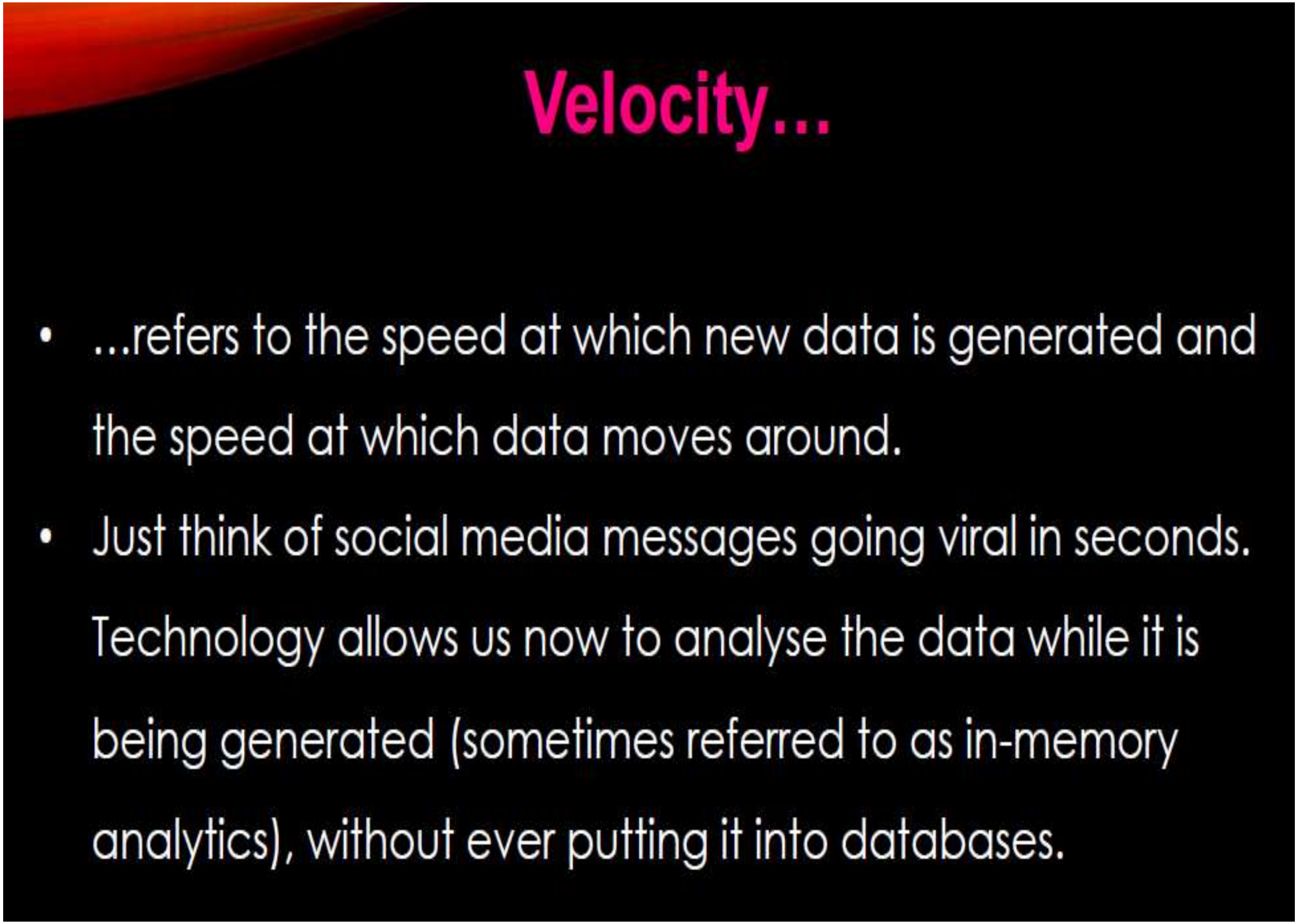


> “Big Data in general is defined as high volume, velocity and variety information assets that demand cost-effective, innovative forms of information processing for enhanced insight and decision making.”
>
> -- Gartner



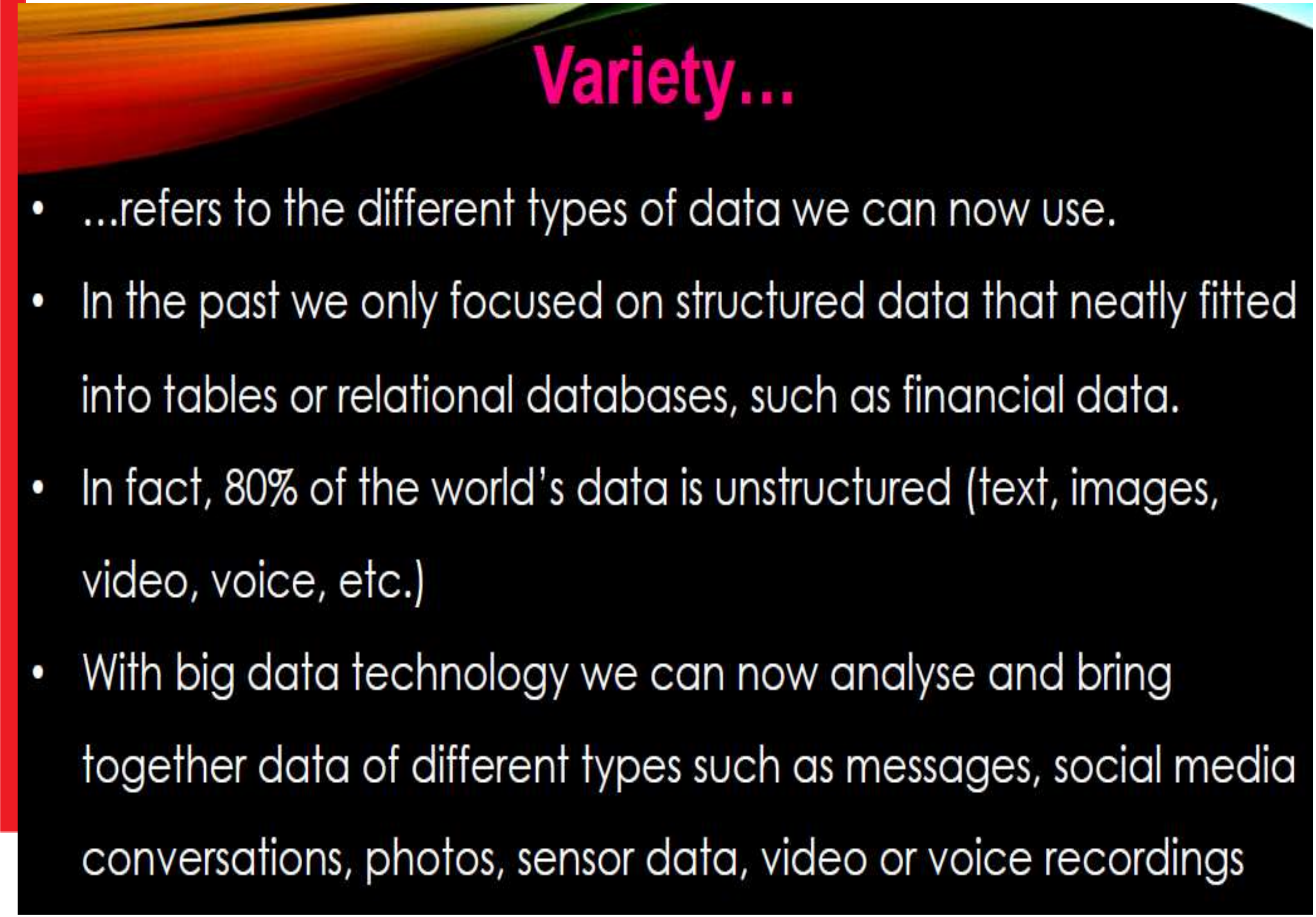
Volume…

- …refers to the vast amounts of data generated every second.
- We are not talking Terabytes but Zettabytes or Brontobytes. If we take all the data generated in the world between the beginning of time and 2008, the same amount of data will soon be generated every minute.
- New big data tools use distributed systems so that we can store and analyse data across databases that are dotted around anywhere in the world.



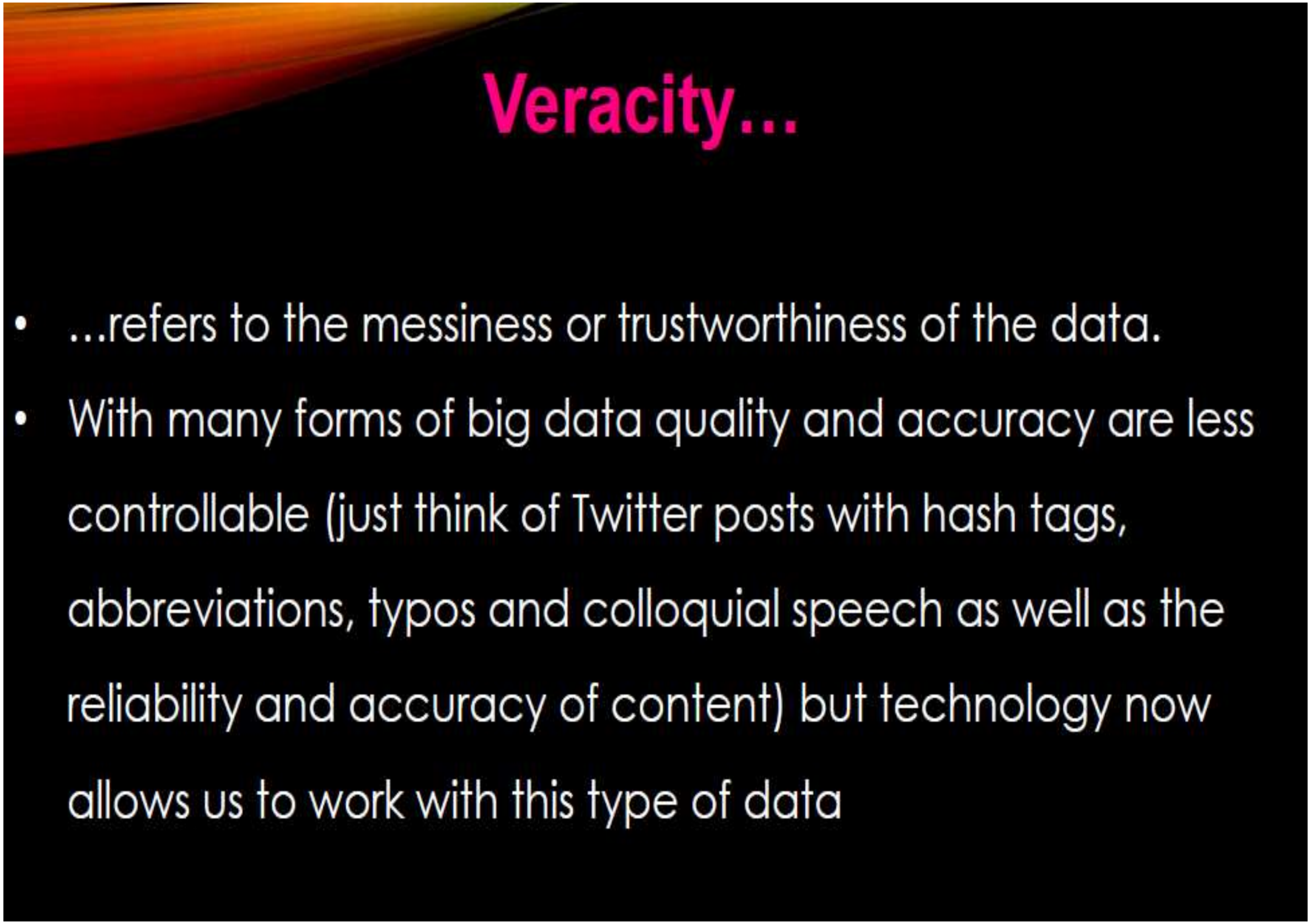
Velocity…

- …refers to the speed at which new data is generated and the speed at which data moves around.
- Just think of social media messages going viral in seconds. Technology allows us now to analyse the data while it is being generated (sometimes referred to as in-memory analytics), without ever putting it into databases.



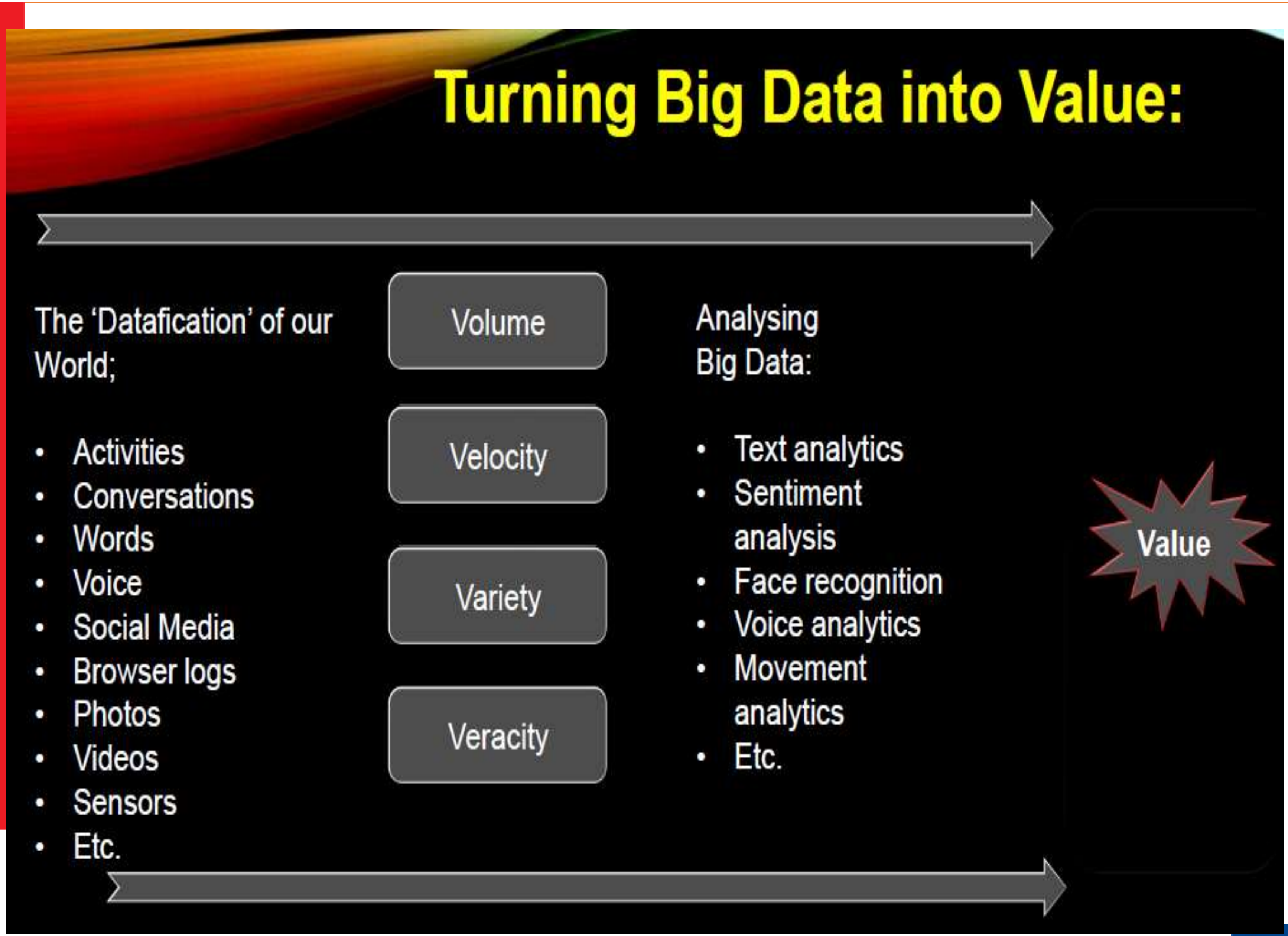
Variety…

- …refers to the different types of data we can now use.
- In the past we only focused on structured data that neatly fitted into tables or relational databases, such as financial data.
- In fact, 80% of the world's data is unstructured (text, images, video, voice, etc.)
- With big data technology we can now analyse and bring together data of different types such as messages, social media conversations, photos, sensor data, video or voice recordings



Veracity…

- …refers to the messiness or trustworthiness of the data.
- With many forms of big data quality and accuracy are less controllable (just think of Twitter posts with hash tags, abbreviations, typos and colloquial speech as well as the reliability and accuracy of content) but technology now allows us to work with this type of data



Turning Big Data into Value:

The 'Datafication' of our World;

- Activities
- Conversations
- Words
- Voice
- Social Media
- Browser logs
- Photos
- Videos
- Sensors
- Etc.

Analysing Big Data:

- Text analytics
- Sentiment analysis
- Face recognition
- Voice analytics
- Movement analytics
- Etc.

What is Data Science

The application of **data centric**, **computational**, and **inferential thinking** to

understand the world **&** *solve problems*

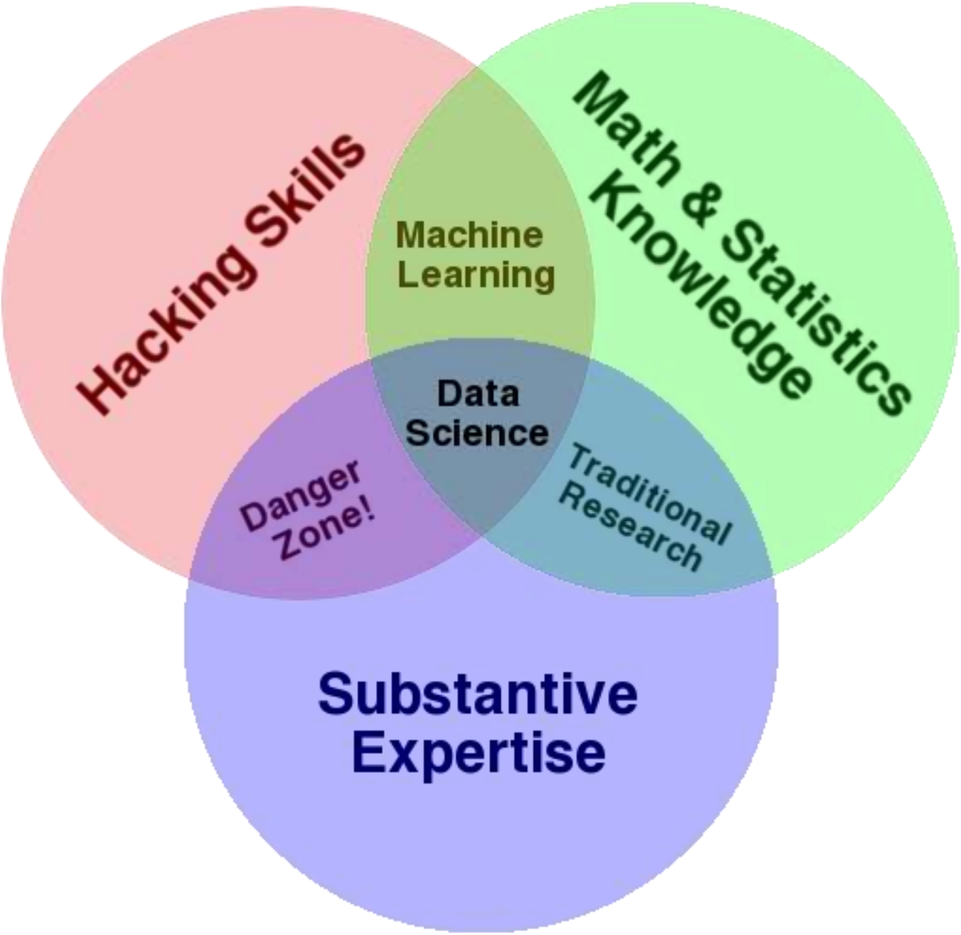
Science **Engineering**

- ➢ *Data science is fundamentally* *interdisciplinary*

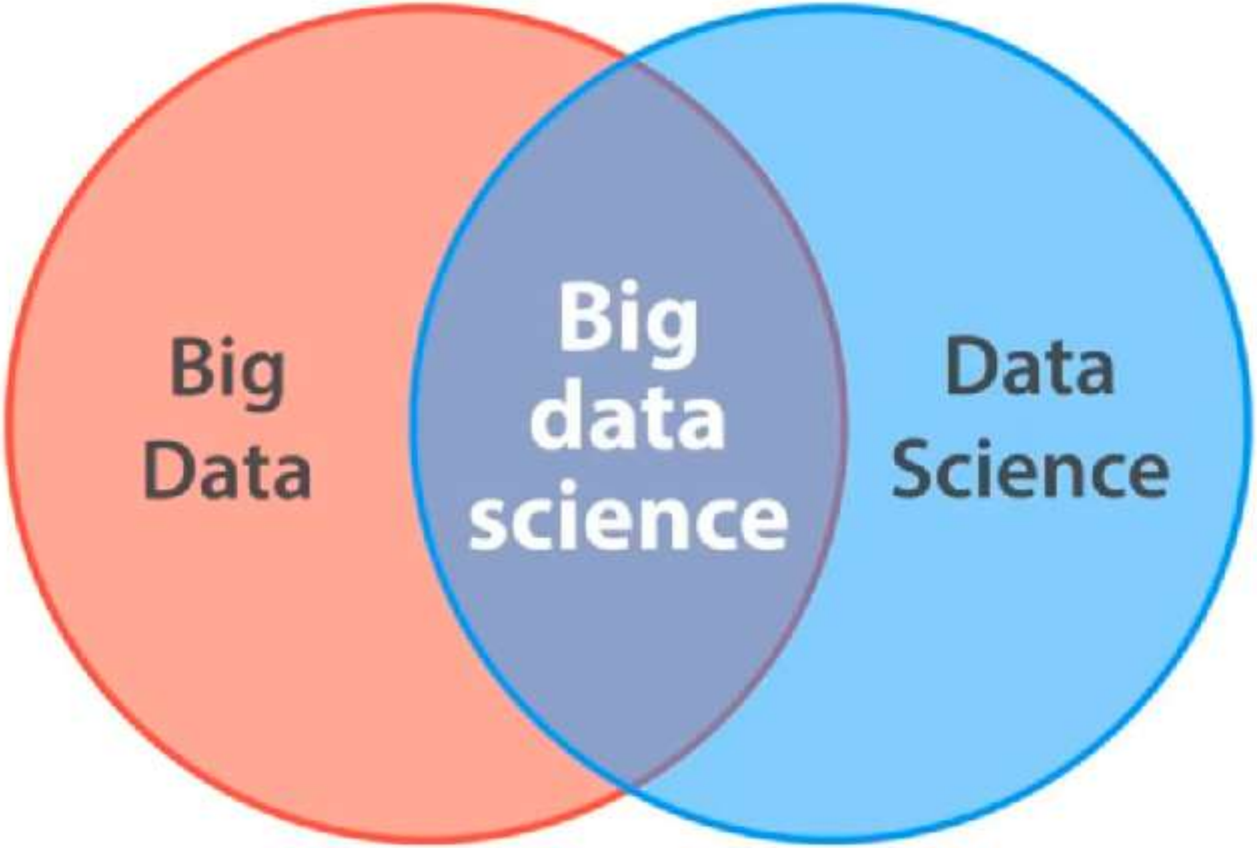
What is Data Science?

- Data Science is a science of analyzing **raw data** using statistics and machine learning with the purpose of drawing conclusion about the information.
- Data Science is used in many industries to allow them to make better business decisions and in sciences to test model and theories.
- This requires process of inspecting, cleaning, transforming, modeling, analyzing an interpreting raw data.

Drew Conway's Venn diagram of data science



Big Data vs Data science

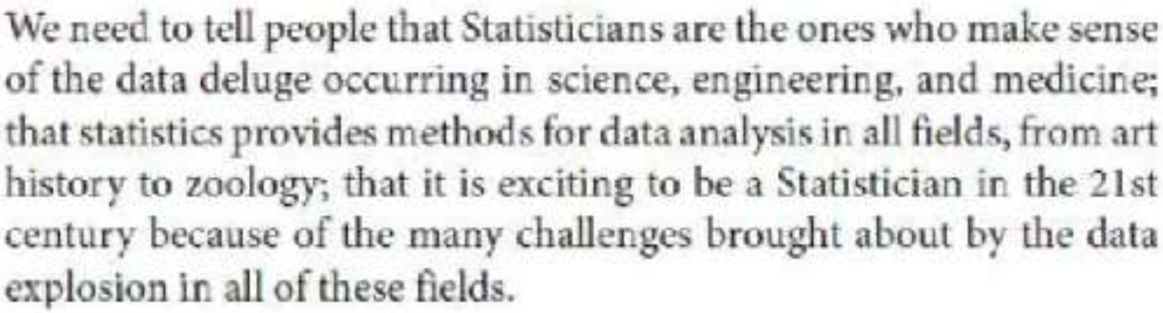


Current landscape of perspectives

- Nathan Yau's 2009 post, "Rise of the Data Scientist", which include:

1. **Statistics (traditional analysis you're used to thinking about)**
2. **Data munging (parsing, scraping, and formatting data)**
3. **Visualization (graphs, tools, etc.)**

- ASA President Nancy Geller's 2011 Amstat News article, "Don't shun the 'S' word", in which she defends statistics:



> We need to tell people that Statisticians are the ones who make sense of the data deluge occurring in science, engineering, and medicine; that statistics provides methods for data analysis in all fields, from art history to zoology; that it is exciting to be a Statistician in the 21st century because of the many challenges brought about by the data explosion in all of these fields.

- Then at LinkedIn and Facebook, respectively—coined the term "data scientist" in 2008.
- Wikipedia finally gained an entry on data science in 2012.

Current landscape of perspectives

- In 2001, William Cleveland wrote a position paper about data science called "Data Science: An action plan to expand the field of statistics."
- Harvard Business Review declared data scientist to be the "Sexiest Job of the 21st Century".

So data science existed before data scientists? Is this semantics, or does it make sense?

A Data Science Profile

- Computer Science
- Math
- Statistics
- Machine learning
- Domain expertise
- Communication and presentation skills
- Data Visualization

proschool
An ims Initiative

DATA SCIENTIST
MUST-HAVE SKILLS

MATH & STATISTICS

- Machine Learning
- Statistical Modeling
- Exploratory Analysis
- Clustering
- Regression Analysis

PROGRAMMING & DATABASE

- Computer Science Fundamentals
- Database Management System
- Data Visualization
- Python
- Big Data

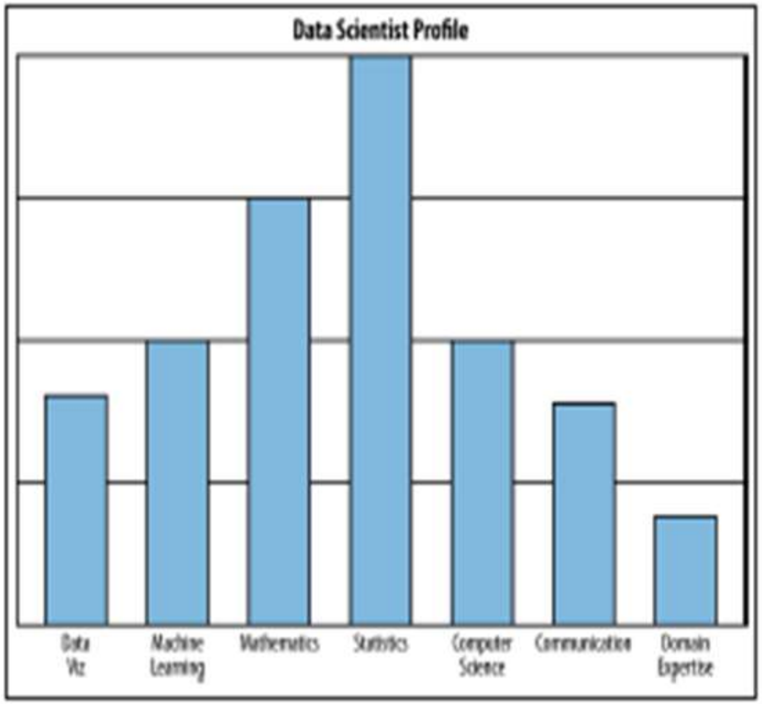
DOMAIN KNOWLEDGE & SOFT SKILLS

- Inclination towards business operations
- Keen on working with data
- Problem solver
- Strategic, proactive, and cooperative
- Interested in hacking

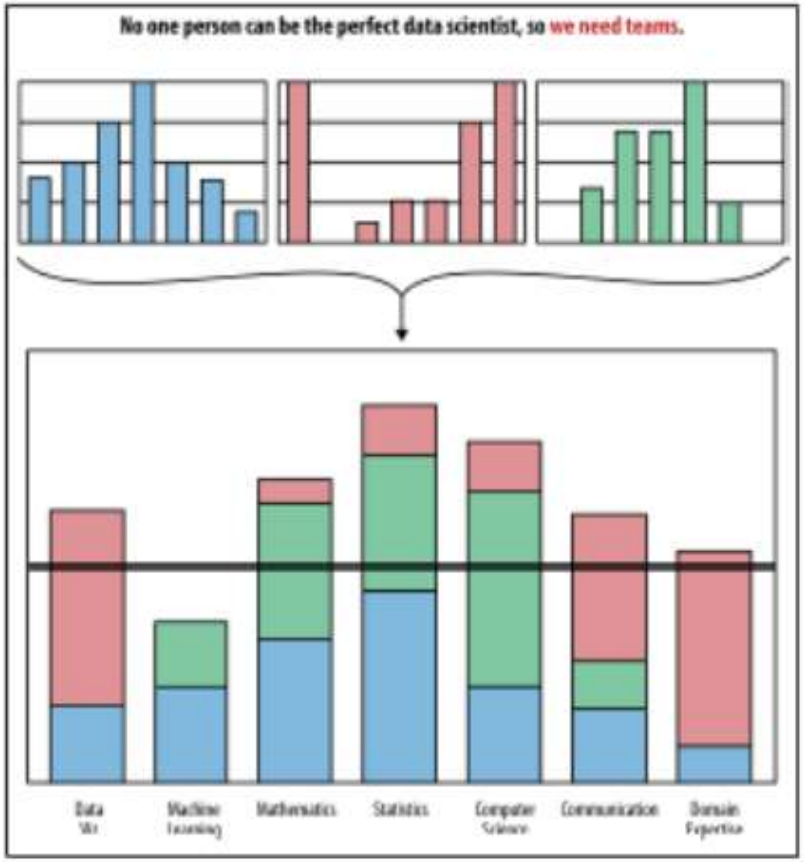
COMMUNICATION & VISUALIZATION

- Storytelling skills
- Convert data-based insights into decisions
- Collaborative with Sr. Management
- Knowledge of tools like Tableau
- Visual art design

Data Science team profile

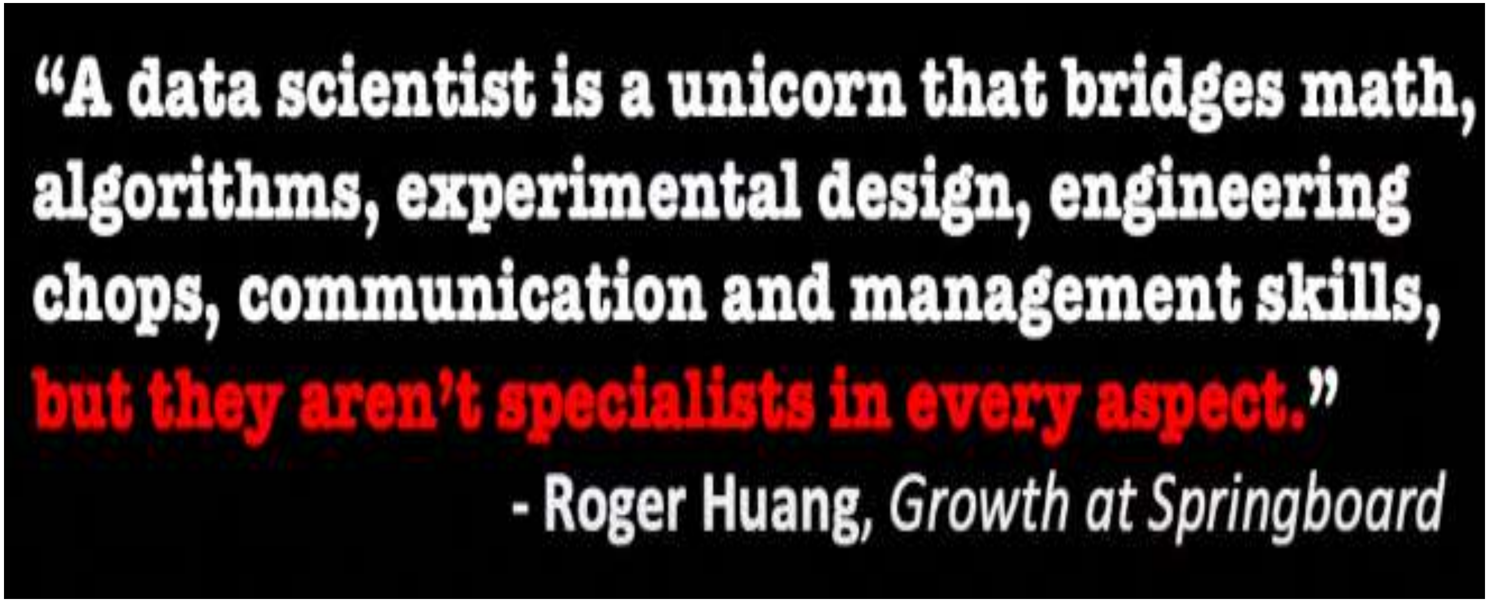


A Data Scientist Profile



Data Science Team profile

Data Scientist

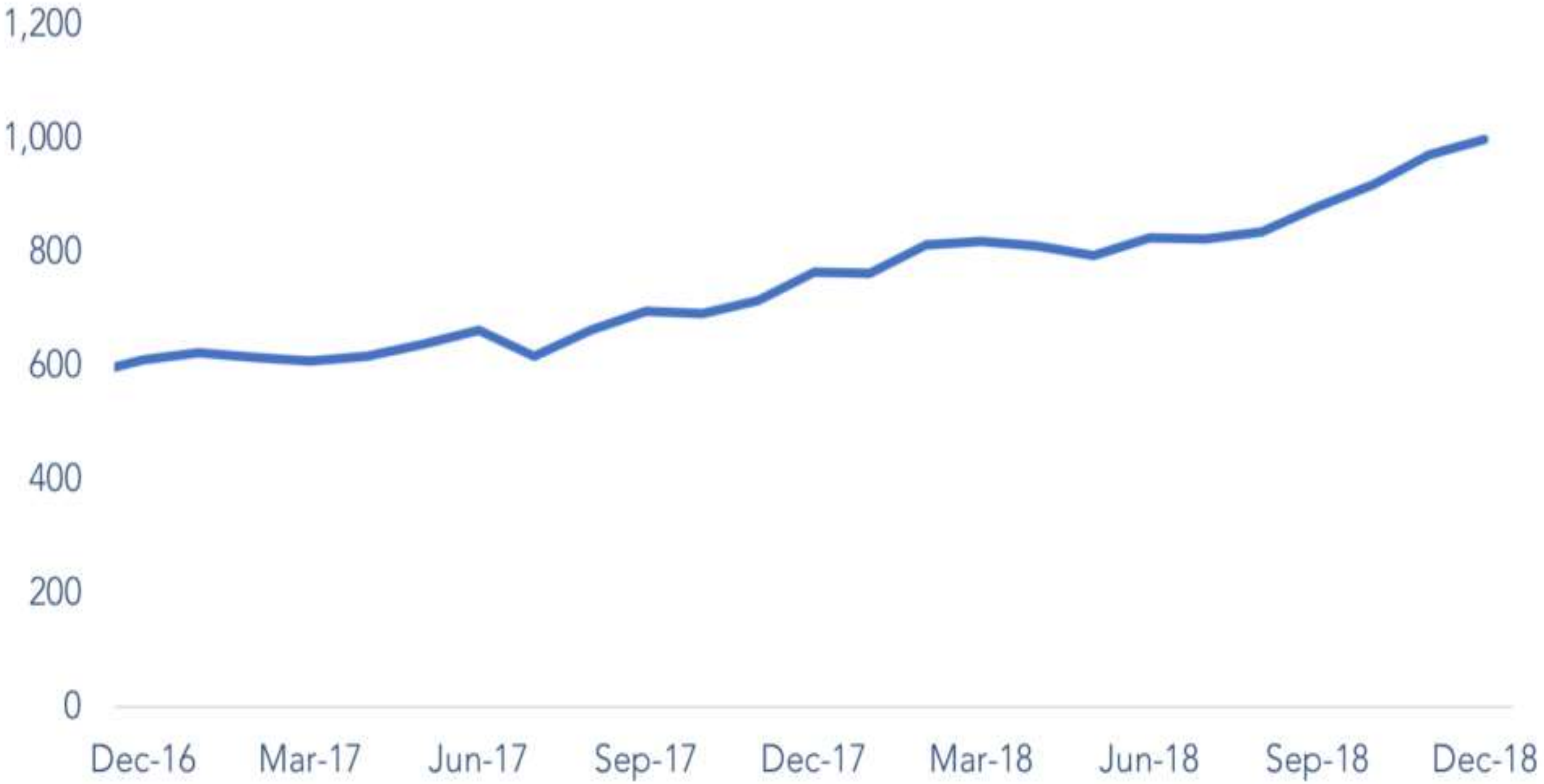


"A data scientist is a unicorn that bridges math, algorithms, experimental design, engineering chops, communication and management skills, but they aren't specialists in every aspect."

- Roger Huang, *Growth at Springboard*

Data scientists are in high demand

Data scientist job postings, per 1 million postings on Indeed



Data Science Process

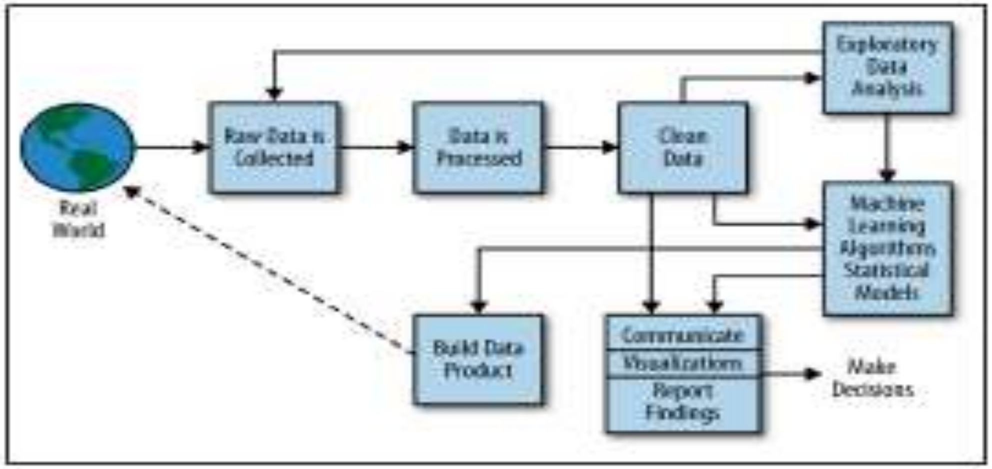


Figure 2-2. The data science process

Role of Data scientist in the process

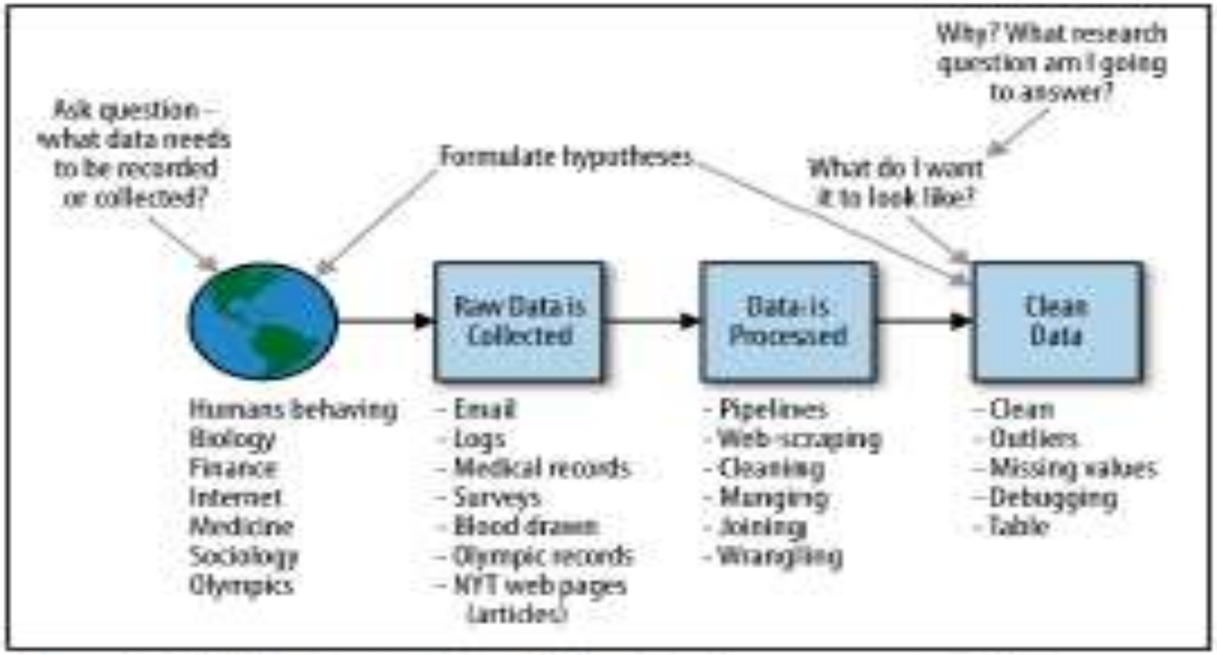


Figure 2-3. The data scientist is involved in every part of this process

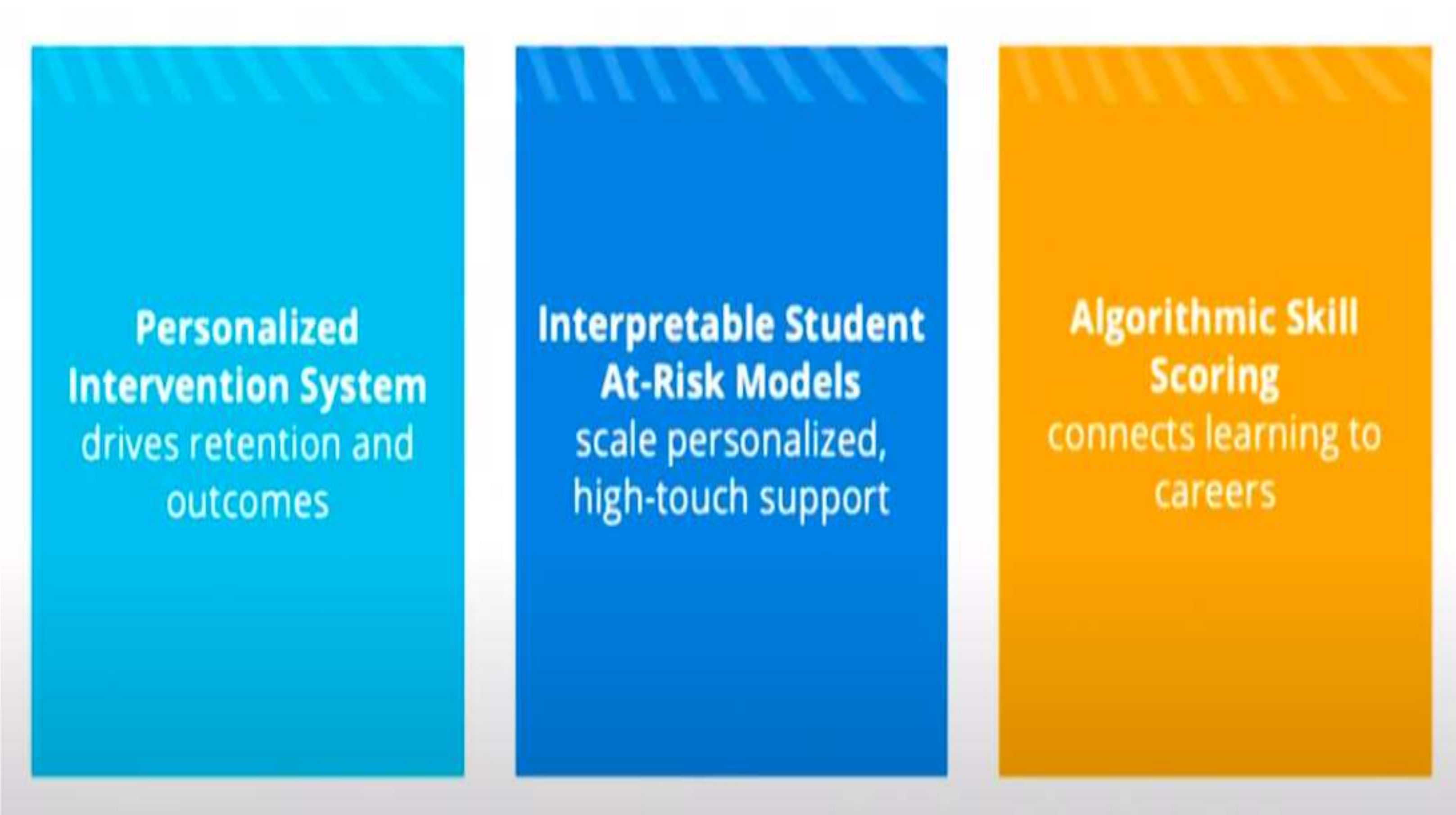
OK so what does Data Scientist do Really?

In Industry:

More generally someone who knows:

- How to design experiments?
- Knows the process of collecting, cleaning and munging data
- Skills that are necessary for understanding the biases in the data and for debugging logging output from code
- Exploratory data analysis which combines visualization and data sense.
- Finding patterns, build models and algorithms
- Use analysis for decision making

Data Science: Case study Unlocking Teaching learning at scale- Coursera

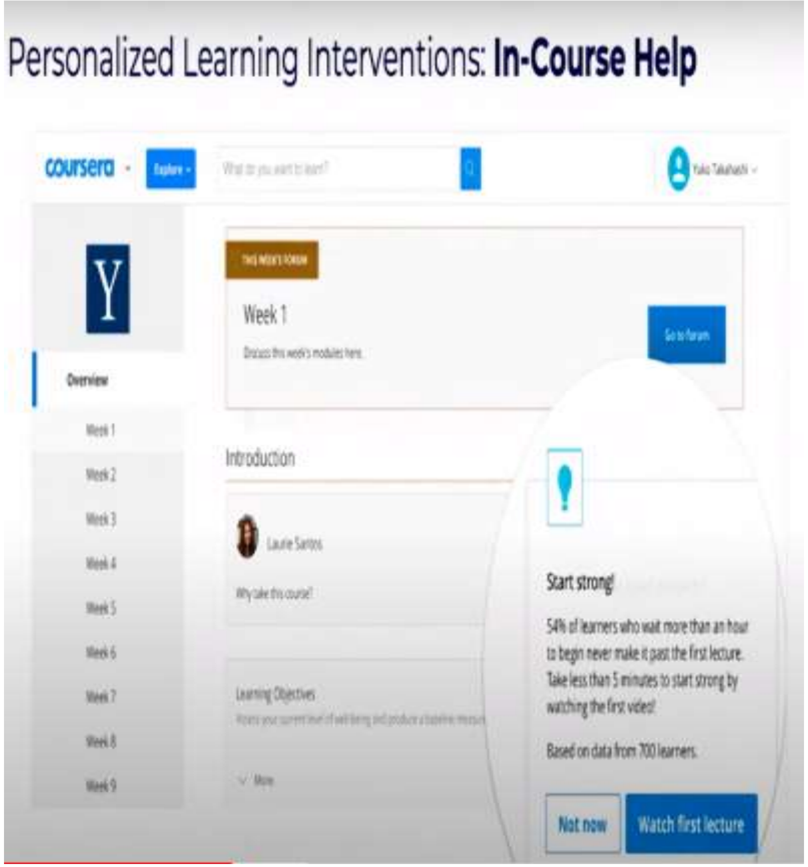


Personalized Intervention System drives retention and outcomes

Interpretable Student At-Risk Models scale personalized, high-touch support

Algorithmic Skill Scoring connects learning to careers

Personalized Intervention system



Personalized Learning Interventions: **In-Course Help**

Week 1

Discuss this week's modules here.

Why take this course?

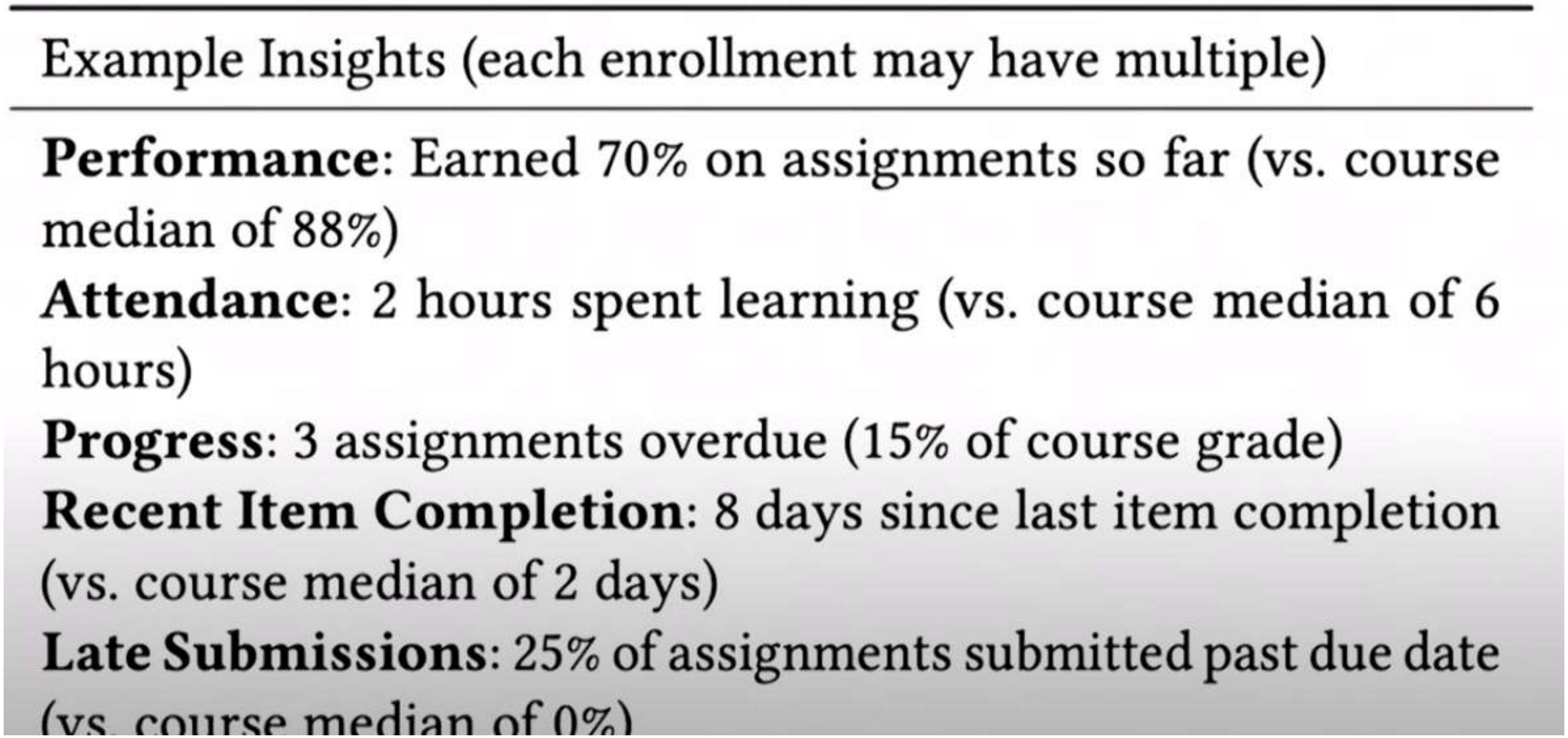
Start strong!

54% of learners who wait more than an hour to begin never make it past the first lecture. Take less than 5 minutes to start strong by watching the first video!

Based on data from 700 learners.

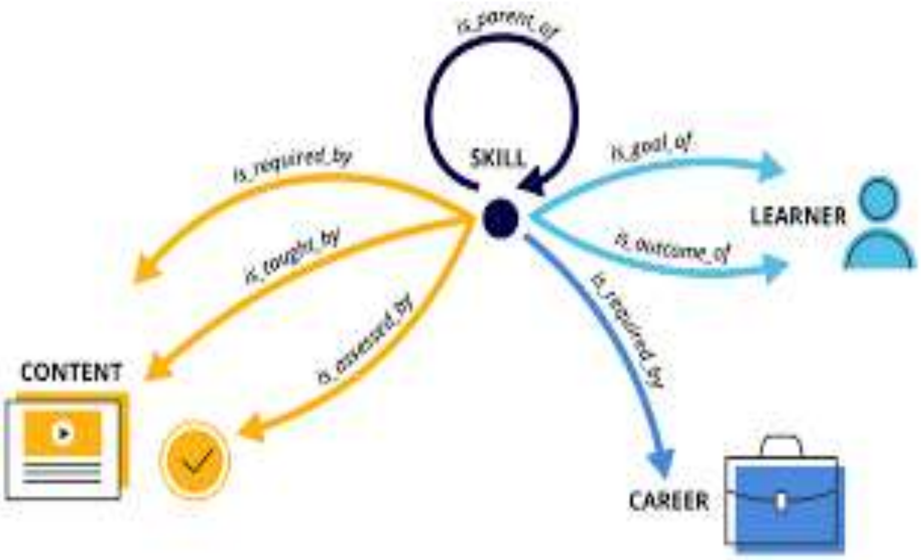
Under the Hood | **Identifying the Right Review Material**

Machine assisted student support dash board



Example Insights (each enrollment may have multiple)
Performance: Earned 70% on assignments so far (vs. course median of 88%)
Attendance: 2 hours spent learning (vs. course median of 6 hours)
Progress: 3 assignments overdue (15% of course grade)
Recent Item Completion: 8 days since last item completion (vs. course median of 2 days)
Late Submissions: 25% of assignments submitted past due date (vs. course median of 0%)

Skill Scoring



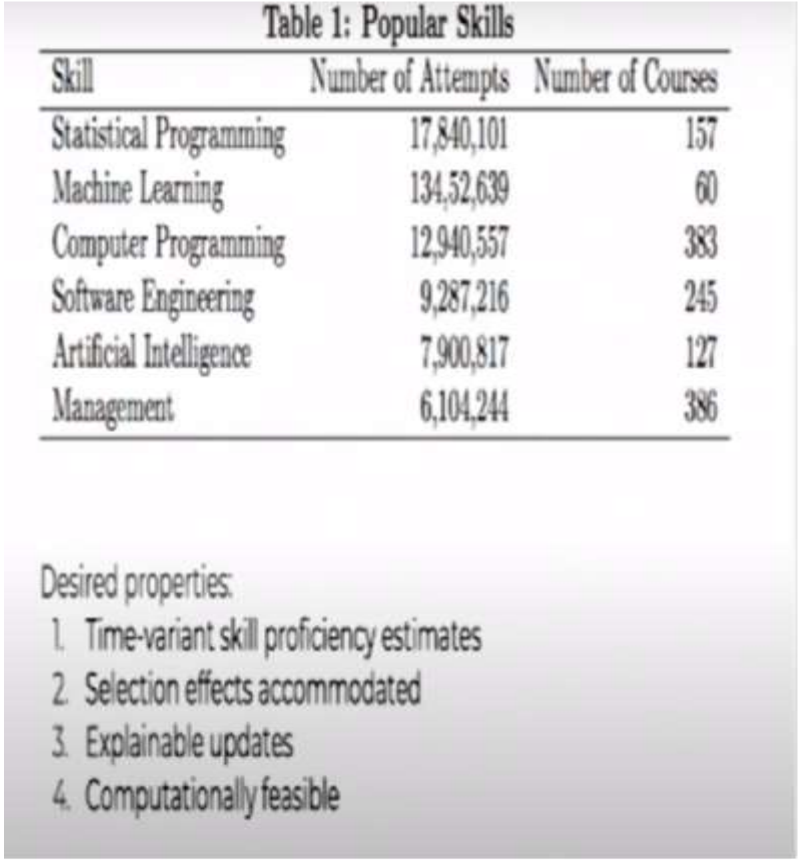
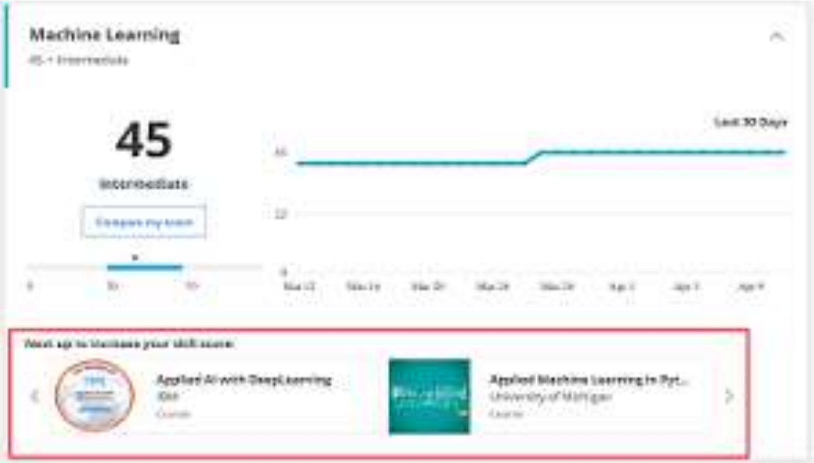


Table 1: Popular Skills

Skill	Number of Attempts	Number of Courses
Statistical Programming	17,840,101	157
Machine Learning	134,52,639	60
Computer Programming	12,940,557	383
Software Engineering	9,287,216	245
Artificial Intelligence	7,900,817	127
Management	6,104,244	386

Desired properties:

1. Time-variant skill proficiency estimates
2. Selection effects accommodated
3. Explainable updates
4. Computationally feasible



Machine Learning

45 • Intermediate

45

Intermediate

Compare my score

Last 30 Days

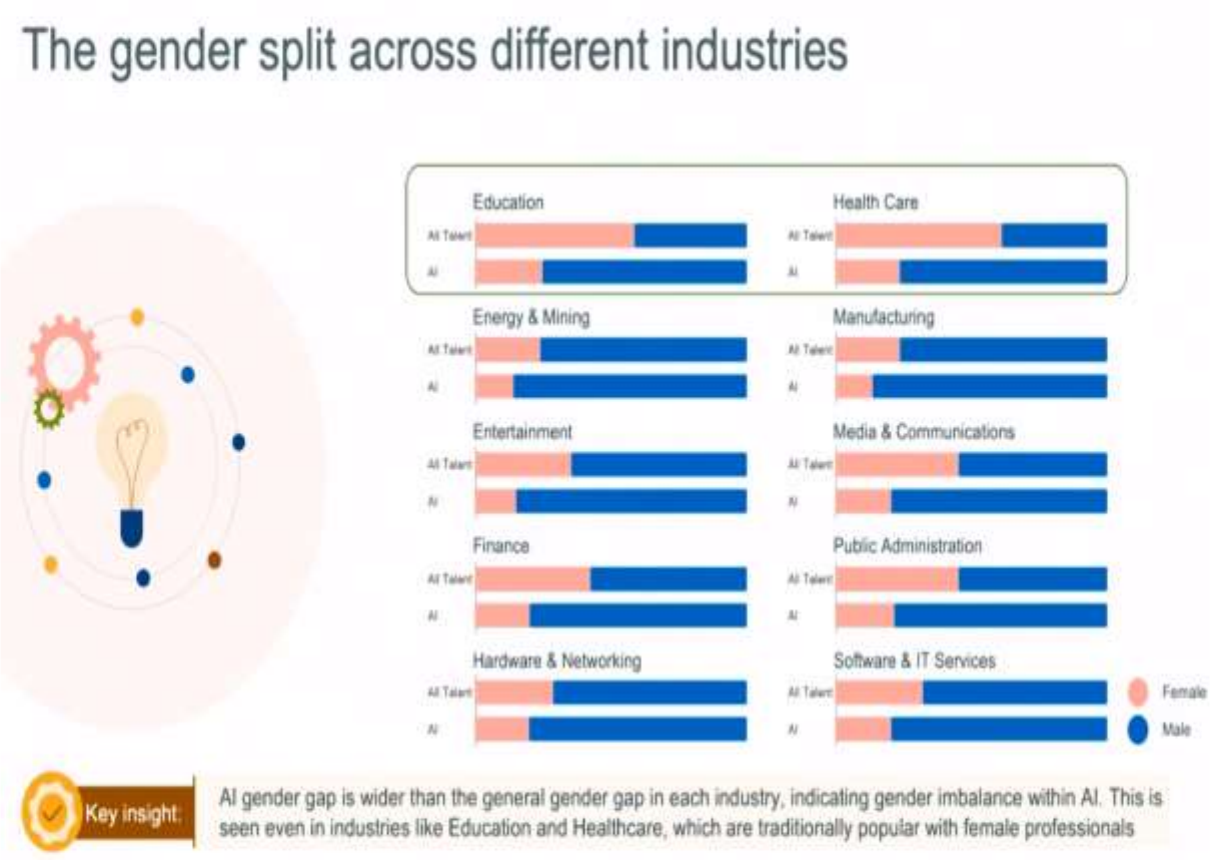
Next up to increase your skill score

Applied AI with DeepLearning
IBM
Course

Applied Machine Learning in Pyt...
University of Michigan
Course

Creating global economic Opportunity

LinkedIn

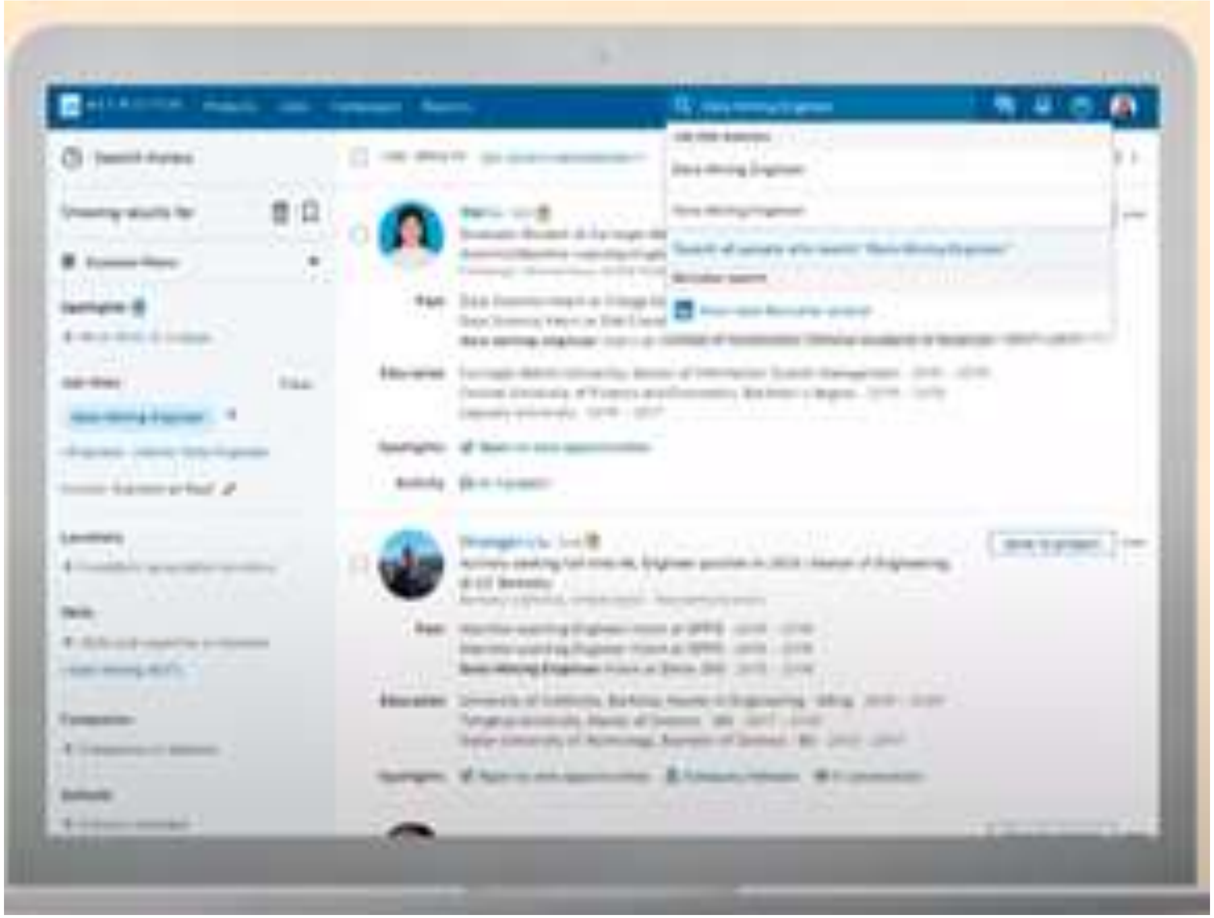


AI gender gap is wider than the general gender gap in each industry, indicating gender imbalance within AI. This is seen even in industries like Education and Healthcare, which are traditionally popular with female professionals

Help students increase diversity of network for better career opportunities

Representatives in search

- Ensure everyone is visible to recruiters



Getting equal opportunities to equally talented peoples

1 Get equally good job recommendations

Job quality, seniority, etc.

2 Apply to jobs equally

At the same applying rate

3 Are hired to jobs equally

At the same hiring rate

4 Are hired equally to equally good jobs with equally good pay

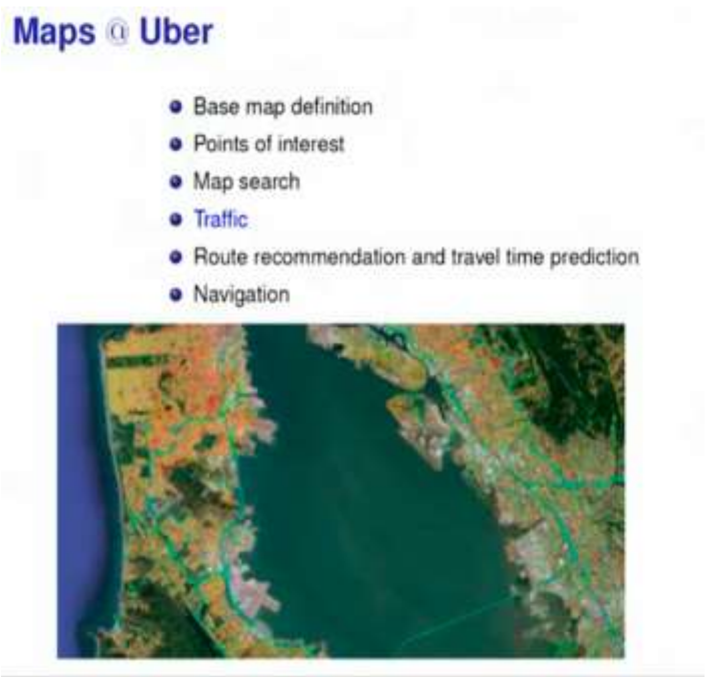
What is a good job? One that they value? Best hours? Best income?

Data Science @UBER

- Intelligent matching lowers waiting time
- Dynamic pricing is a key tor reliability

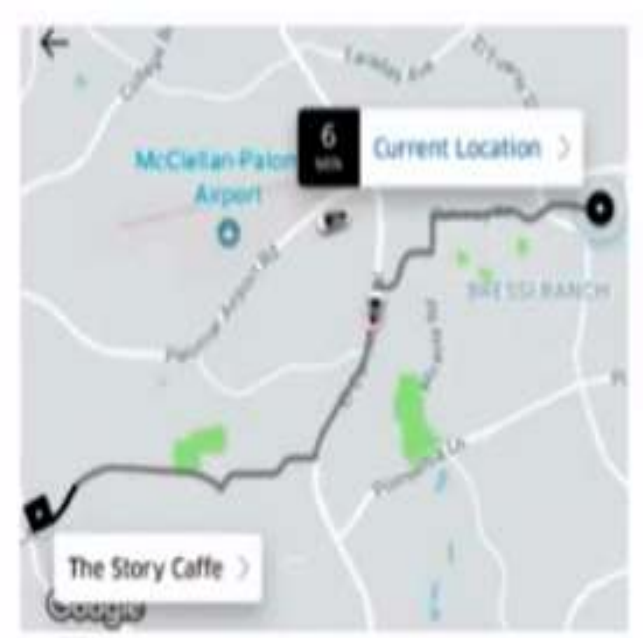
Data Science @UBER

- Uber is physical logistic system – every aspect of experience of rider and driver is attributed to the road network.



Maps @ Uber

- Base map definition
- Points of interest
- Map search
- Traffic
- Route recommendation and travel time prediction
- Navigation

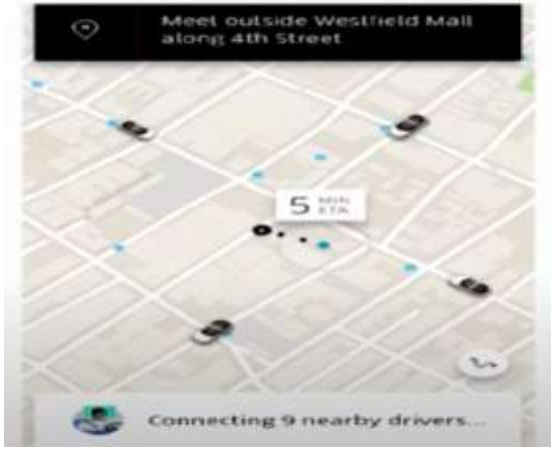


Matching

- How should the riders match with the drivers?
- How should carpool riders be matched with each other?

Simple mechanism:

- ○ Immediately dispatch the Driver with shortest pick up Time
- ○ Can be improved with the Mechanism like Trip-upgrade



Meet outside Westfield Mall along 4th Street

Connecting 9 nearby drivers...

Predicting Travel time

- Matching requires predicting the travel time between two points

Lower wait time

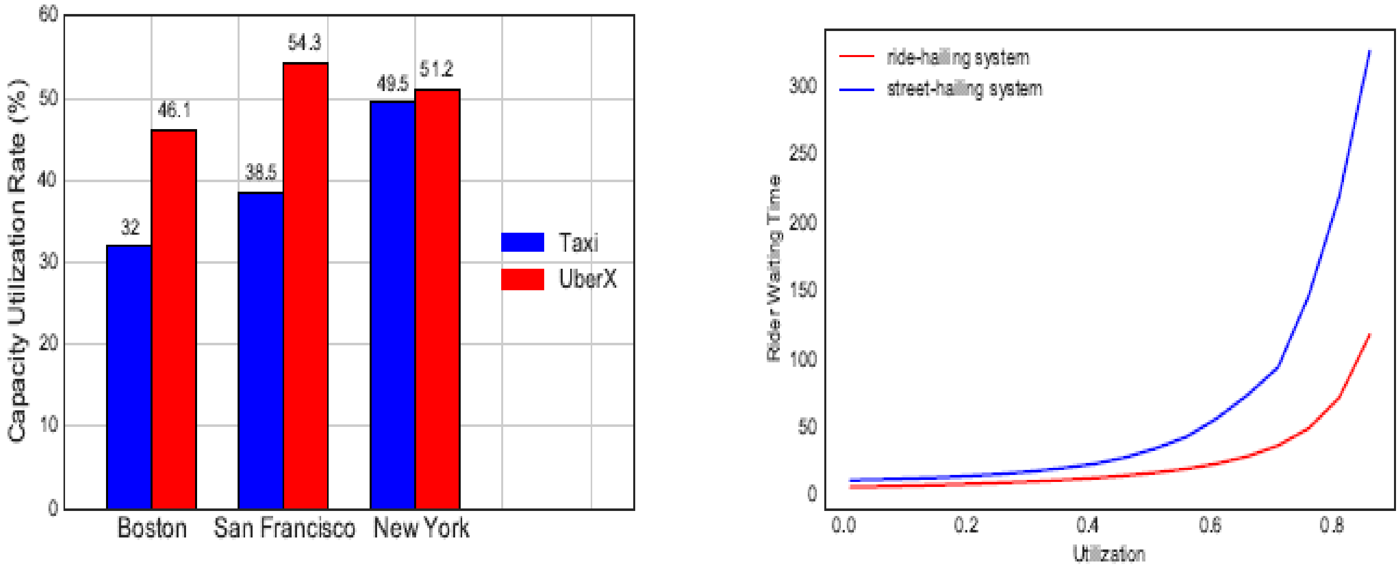


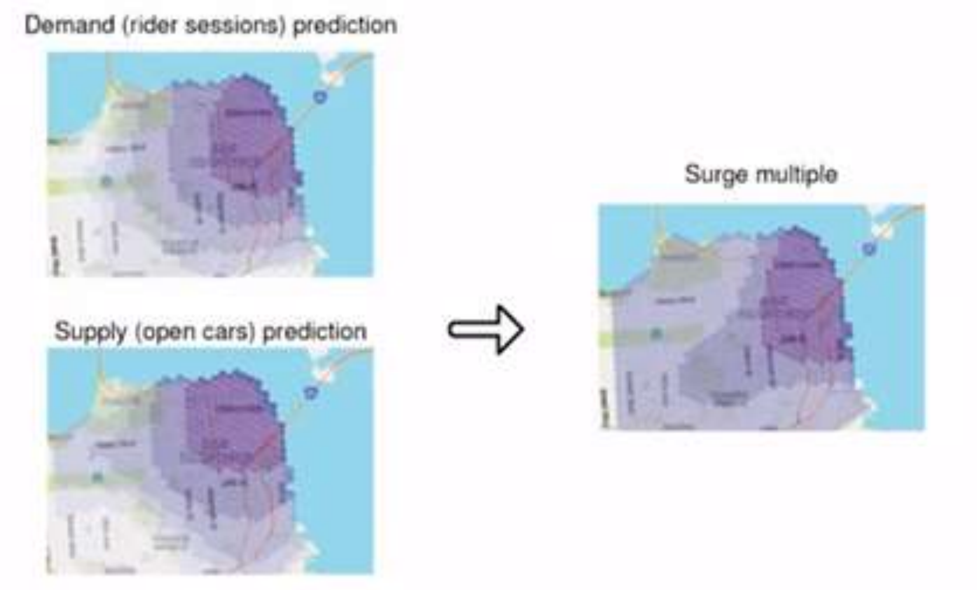
Figure 1 Comparisons for capacity utilization rate and rider waiting time. Left: capacity utilization rate comparison between taxi services and UberX (Cramer and Krueger 2016); Right: rider waiting time comparison between ride-hailing and street-hailing (Feng et al. 2017).

Intelligent matching lowers wait time

Dynamic pricing is a key to reliability

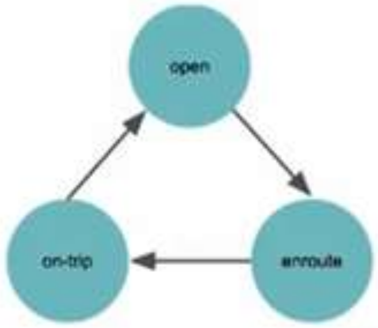
- Price is based on short-term prediction of demand and supply.

NYC taxi demand data
Pickup density - 0:00

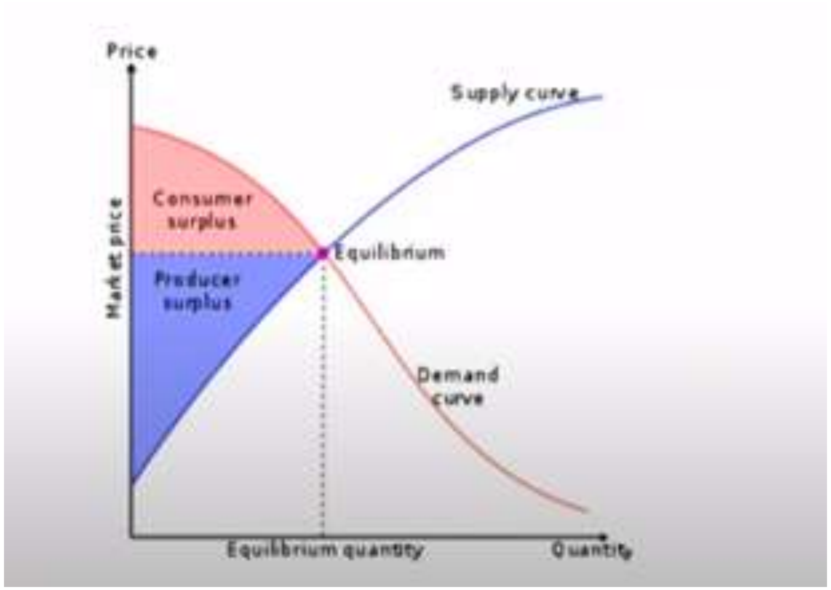

Demand (rider sessions) prediction
Supply (open cars) prediction
Surge multiple

Dynamic pricing

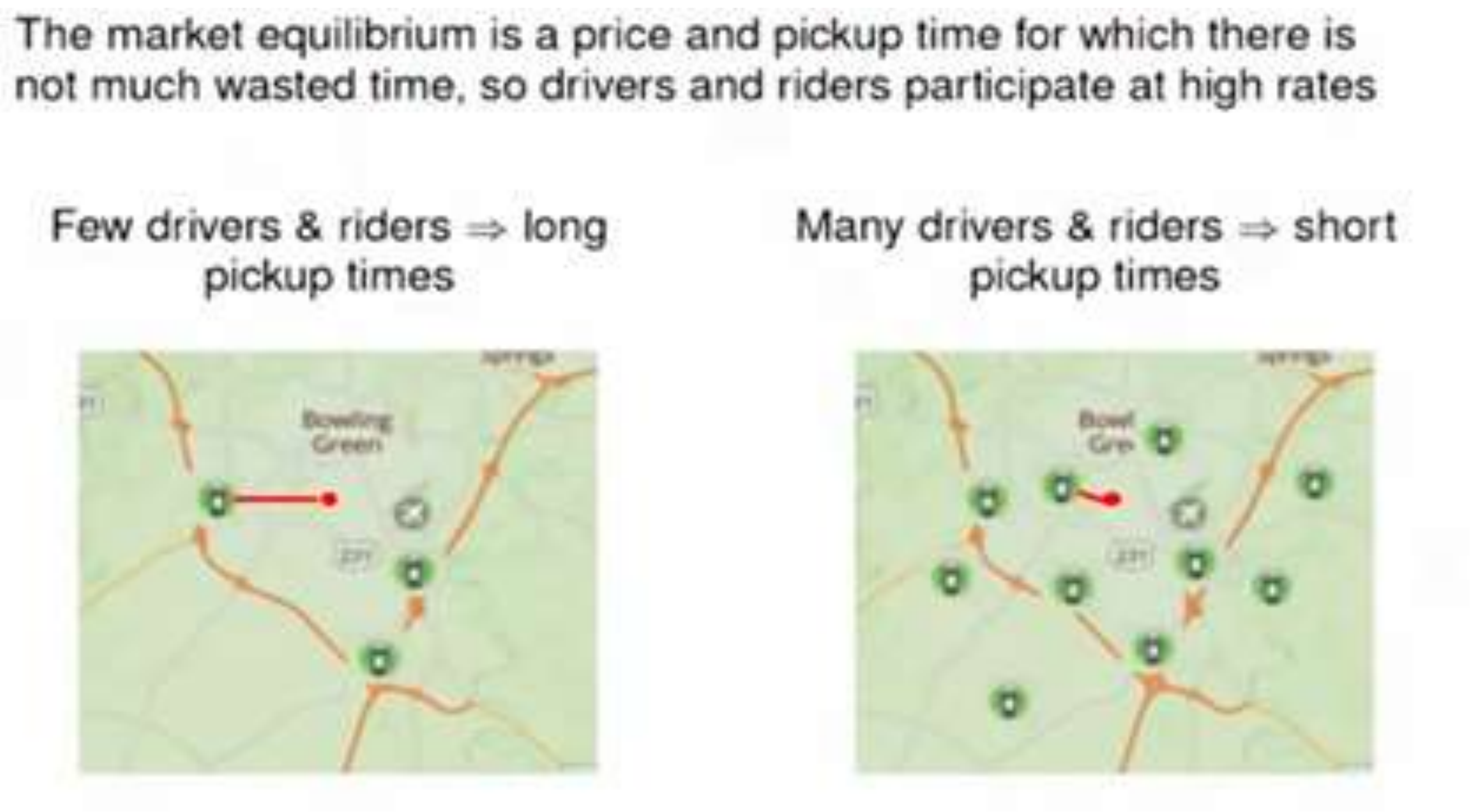
Drivers cycle between three states:



- Riders are sensitive to pickup time
- Drivers are sensitive to pickup + open time



Dynamic Pricing

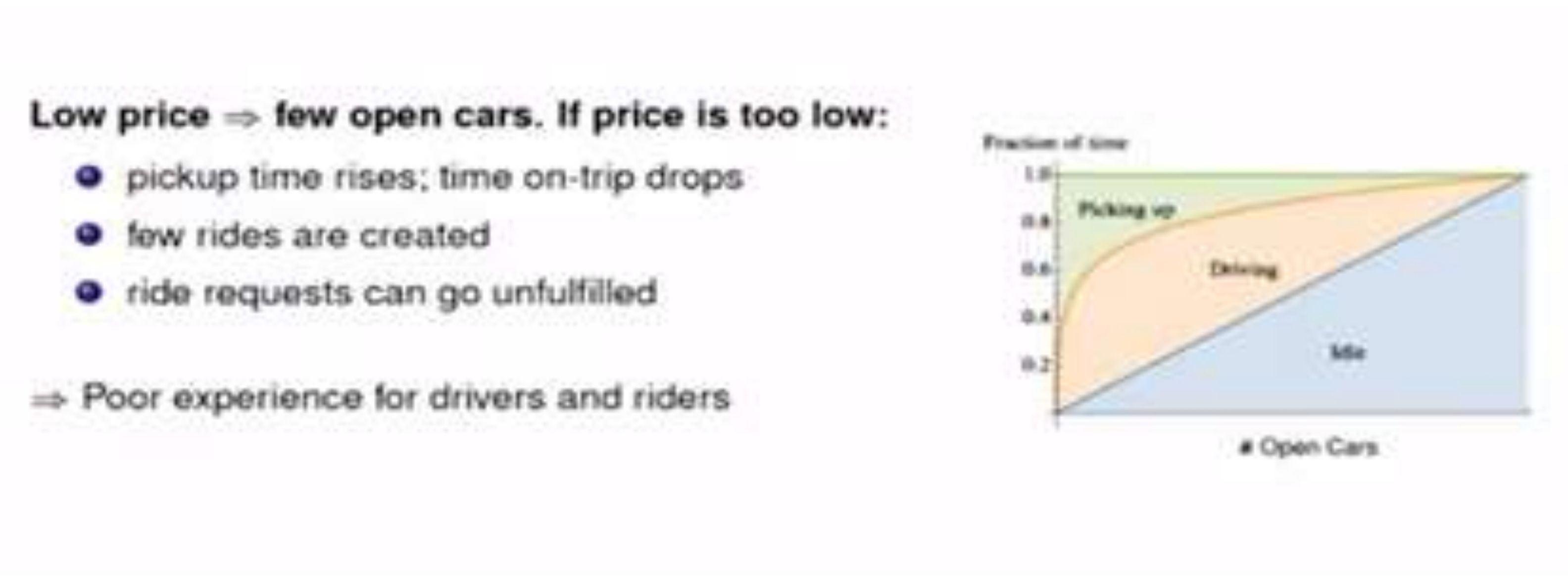


The market equilibrium is a price and pickup time for which there is not much wasted time, so drivers and riders participate at high rates

Few drivers & riders ⇒ long pickup times

Many drivers & riders ⇒ short pickup times

Dynamic Pricing

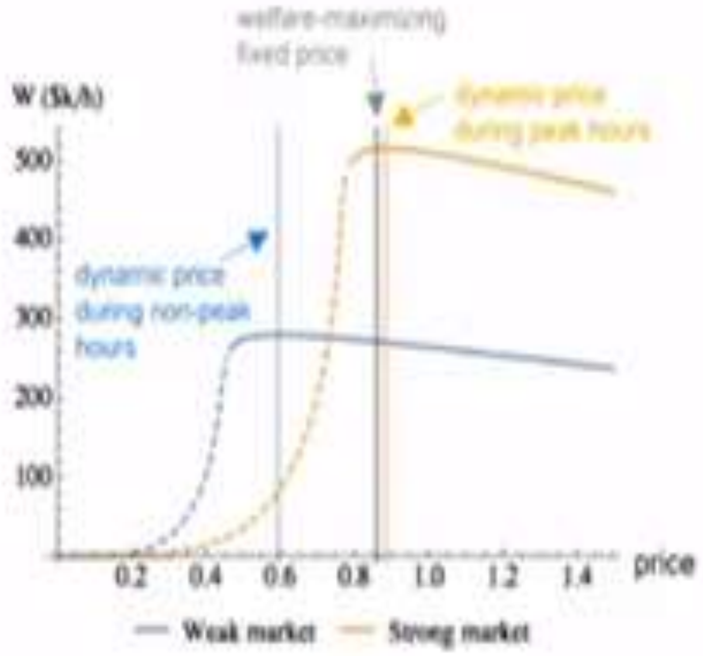


Low price ⇒ few open cars. If price is too low:

- pickup time rises; time on-trip drops
- few rides are created
- ride requests can go unfulfilled

⇒ Poor experience for drivers and riders

Dynamic Pricing



- Welfare is measure of value created for both rider and driver
- When the price is below a threshold welfare drops
- Threshold changes over time as demand and supply change
- When dynamic pricing is disallowed both welfare is reduced

Data Science@Uber

- Growth of ride-sharing services is based on data driven matching and pricing
- Intelligent dispatch reduces the wait time
- Dynamic pricing is important to realiability
- Both matching and pricing requires forecasting demand, supply and travel time

Question ?