A Systematic Approach to Data Science

M. Tamer Ö兹su
University of Waterloo
tamer.ozsu@uwaterloo.ca
“Data is the new oil.”
Clive Robert Humby
mathematician, entrepreneur, and
Chief Data Scientist, Starcount

“Data is the new currency.”
Antonio Neri, President
Hewlett Packard Enterprise

“Data is a commodity like gold.”
Matt Shepherd
Head of Data Strategy, BBH London

“At the heart of the digital economy and society is the explosion of insight, intelligence and information – data. Data is the lifeblood of the digital economy.”
World Economic Forum
A New Paradigm for Business of Data
BRIEFING PAPER - JULY 2020
Data Science/Big Data in the News

Big Brother meets Big Data, in an office near you.

The Atlantic
Sponsor Content: What’s this?

How Big Data And The Internet Of Things Improve Public Transport In London

Forbes / Tech
MAY 27, 2015 @ 10:20 AM
34,550

BIG DATA AND
HOLLYWOOD: A LOVE STORY

The Big Idea Behind Big Data

npr
SCIENCE

New York Times Adapts Data Science Tools for Advertisers

Carnival Strategy Chief Bets That Big Data Will Optimize Prices

Team will help lure marketers with tools to predict which articles will resonate with certain readers to better target advertising

Data Veracity is Critical for Insurers to Make Better Business Decisions, according to Accenture Report

The Little Black Book of Billionaires
Data Science Everywhere!

“No candy? No flowers? No cards?
Big Data predicted that 67.53% of you would remember!”

“I don’t like the look of this. Searches for gravy and turkey stuffing are going through the roof!”

“You can’t keep adjusting the data to prove that you would be the best Valentine’s date for Scarlett Johansson.”
Data Science Everywhere!...

“Do we have any actionable analytics from our big data in the cloud?”

“YES, THE DATA SHOWS THAT MY PRODUCTIVITY PLUNGES WHENEVER YOU LEARN NEW JARGON.”

“MAYBE IN-MEMORY COMPUTING WILL ACCELERATE YOUR APPLICATIONS.”

“Plunge, plunge, plunge.”

“You can’t keep adjusting the data to prove that you would be the best Valentine’s date for Scarlett Johansson.”
Data Science Everywhere!

[Comic:]

**1.** DO WE HAVE ANY ACTIONABLE ANALYTICS FROM OUR BIG DATA IN THE CLOUD?

**2.** YES, THE DATA SHOWS THAT MY PRODUCTIVITY PLUNGES WHENEVER YOU LEARN NEW JARGON.

**3.** MAYBE IN-MEMORY COMPUTING WILL ACCELERATE YOUR APPLICATIONS.

**4.** WHAT DOES THE DATA TELL US TO DO?

**5.** WE ONLY HAVE BAD DATA ON THIS.

**6.** DOES THE BAD DATA SUGGEST WE SHOULD DO WHAT WE WANTED TO DO ANYWAY?

**7.** WELL, YES.

**8.** THAT’S CALLED “GOOD DATA.”

Valentine’s date for Scarlett Johansson.
Data Science Needs Positioning
“Data science, also known as data-driven science, is an interdisciplinary field of scientific methods, processes, algorithms and systems to extract knowledge or insights from data in various forms, either structured or unstructured, similar to data mining.”
What is Data Science?

“Data science intends to analyze and understand actual phenomena with ‘data’. In other words, the aim of data science is to reveal the features or the hidden structure of complicated natural, human, and social phenomena with data from a different point of view from the established or traditional theory and method.”
What is Data Science?

“What is Data Science? The book is about Data Science, Classification, and Related Methods. The author is Chikio Hayashi, Keiji Takama, Hiroshi Bock, Nobuyuki Shumi, and Yasutaka Yamaoka. The book was published by Springer in 2002. The book has 214 pages and is 22.8 cm long and 15.2 cm wide. The cover of the book is blue and white. The title of the book is "Data Science, Classification, and Related Methods.""
What is Data Science?

- Fourth paradigm
  - “… change of all sciences moving from observational, to theoretical, to computational and now to the 4th Paradigm – Data-Intensive Scientific Discovery”
What is Data Science?

- “Data science encompasses a set of principles, problem definitions, algorithms, and processes for extracting non-obvious and useful patterns from large data sets.”
- “…the terms data science, machine learning, and data mining are often used interchangeably.”
- “…although data science borrows from these other fields, it is broader in scope.”
A Working Definition

A data-driven approach to problem solving that involves the process of collecting, managing, analyzing, explaining and visualizing data and analysis results.
Data Science as a Unifier
Who is a Data Scientist?

To be revealed at the end…
Two Myths…

- Data science = Big data
Two Myths…

- Data science ≠ Big data
- Big data is like a raw material
- Processing it leads to data science & better understanding
- Applications are important
  - No applications → no data science
Two Myths…

- Data science ≠ Big data
- Big data is like a raw material
- Processing it leads to data science & better understanding
- Applications are important
  - No applications → no data science

- Data science ⊆ Machine learning ⊆ AI
Two Myths…

- Data science ≠ Big data
- Big data is like a raw material
- Processing it leads to data science & better understanding
- Applications are important
  - No applications → no data science

- Data science ≠ Machine learning ≠ AI

- They are related but not the same
Data Science Applications

- Data science is about applications
  - Applications give purpose
  - Applications inform core technologies
- Almost any field with large data sets are good candidates
- Some examples
  - Fraud detection
  - Biological & biomedical applications
  - Recommender systems
  - Health sciences & health informatics applications
  - Sustainability
  - Finance & insurance
  - Smart cities
  - Sports
  - …
Data Science Application Examples

- Fraud detection
  - Investigate fraud patterns in past data
  - Early detection is important
    - Before damage propagates
    - Harder than late detection
  - Precision is important
    - False positive and false negative are both bad
  - Real-time analytics
Data Science Application Examples

- Recommender systems
  - The ability to offer unique personalized service
  - Increase sales, click-through rates, conversions, …
  - Collaborative filtering at scale
Data Science Application Examples

- **Sustainability**
  - Climate variability and change
  - Ecology
  - FEW
    - Food
    - Energy
    - Water
Data Science Application Examples

- Moneyball
  - How to build a baseball team on a very low budget by relying on data
  - *Sabermetrics*: the statistical analysis of baseball data to objectively evaluate performance
  - 2002 record of 103-59 was joint best in MLB
    - Team salary budget: $40 million
  - Other team: Yankees
    - Team salary budget: $120 million
Data Science Ecosystem

Data Science Building Blocks

**Data Engineering**
- Data quality
- Big Data storage and computing solutions
- Data pipelines (ETL)

**Data Analytics**
- Explore data (data mining)
- Build models & algorithms (machine learning)
- Visualizations & visual analytics

**Data Security & Privacy**
- Differential privacy
- Applications of cryptography
- Data integrity

**Data Ethics**
- Impact on individuals, organizations & society
- Ethical & normative concerns
- Regulatory issues
Data Engineering

Big data management
(Four Vs)
Data Engineering

Big data management

(Four Vs)

- Data processing platforms
- Data integration
  - ETL process
  - Data lakes
- Data quality issues
- Data provenance
Data Engineering Essential
Data Engineering Essential

GARbage DATA → GREAT MODEL → GARbage RESULTS
Excel spreadsheet error blamed for UK’s 16,000 missing coronavirus cases

The case went missing after the spreadsheet hit its filesize limit

By James Vincent | Oct 5, 2020, 9:41am EDT
Excel spreadsheet error blamed for UK’s 16,000 missing coronavirus cases

The case went missing after the spreadsheet hit its filesize limit

By James Vincent | Oct 5, 2020, 9:41am EDT

"THE ISSUE WAS CAUSED BY THE FACT THAT SOME FILES CONTAINING POSITIVE TEST RESULTS EXCEEDED THEIR MAXIMUM FILE SIZE"
Excel spreadsheet error blamed for UK’s 16,000 missing coronavirus cases

The case went missing after the spreadsheet hit its filesize limit

By James Vincent | Oct 5, 2020, 9:41am EDT

Under-reported figures
From 25 Sept to 2 Oct

50,786 Cases initially reported by PHE
15,841 Unreported cases, missed due to IT error

8 days of incomplete data
1,980 cases per day, on average, were missed in that time
48 hours Ideal time limit for tracing contacts after positive test

Source: PHE and gov.uk
Excel spreadsheet error blamed for UK’s 16,000 missing coronavirus cases

The case went missing after the spreadsheet hit its filesize limit

By James Vincent | Oct 5, 2020, 9:41am EDT

Under-reported figures
From 25 Sept to 2 Oct

50,786
Cases initially reported by PHE

15,841
Unreported cases, missed due to IT error

8 days
of incomplete data

1,980
Cases per day, on average, were missed in that time

48 hours
Ideal time limit for tracing contacts after positive test

Source: PHE and gov.uk
Big Data – Four Vs

**Volume**

- Scale of data
- Data at rest
Big Data – Four Vs

Volume

• Scale of data
• Data at rest

There were 5 exabytes of information created between the dawn of civilization through 2003, but that much information is now created every 2 days.

- Eric Schmidt
  Executive Chairman of Google
Big Data – Four Vs

Volume

• Scale of data
• Data at rest

Volume in Exabytes

1 Exabyte (EB) = 1,000,000,000,000,000,000 Bytes

2010

2020

50x Growth from 2010 to 2020

90% of the world’s data was created in the last 2 years

Source: Infosys
Big Data – Four Vs

Volume
- Scale of data
- Data at rest

Variety
- Forms of data
- Unstructured challenges
Big Data – Four Vs

**Volume**
- Scale of data
- Data at rest

**Variety**
- Forms of data
- Unstructured challenges

![Graph showing increasing data volume](image)

![Illustration of data sources](image)
Big Data – Four Vs

Volume
- Scale of data
- Data at rest

Variety
- Forms of data
- Unstructured challenges

Velocity
- Streaming data
- Data in motion
Big Data – Four Vs

**Volume**
- Scale of data
- Data at rest

**Velocity**
- Streaming data
- Data in motion

**Variety**
- Forms of data
- Unstructured challenges

Global Video Streaming Software Market, by Region

*Chart showing the growth of global video streaming software market by region from 2018 to 2026.*
Big Data – Four Vs

**Volume**
- Scale of data
- Data at rest

**Velocity**
- Streaming data
- Data in motion

**Variety**
- Forms of data
- Unstructured challenges

---

**Global Video Streaming Software Market, by Region**

**Growth in Internet of Things Devices**
- Billions of IoT devices according to NCTA

Data source: NCTA
Big Data – Four Vs

**Volume**
- Scale of data
- Data at rest

**Variety**
- Forms of data
- Unstructured challenges

**Velocity**
- Streaming data
- Data in motion

**Veracity**
- Uncertainty/incorrectness in data
- Data quality
Data Integration – Data Lakes

Analysis  Access
Data Quality in Big Data

89% of executives believe that data quality issues impact the quality of customer service they provide (2017)

Only 33% of senior executives have a high level of trust in the accuracy of their big data analytics (2016)

59% of executives do not believe their company has capabilities to generate business insights from their data (2016)
Data Quality in Big Data

89% of executives believe that data quality issues impact the quality of customer service they provide (2017).

Only 33% of senior executives have a high level of trust in the accuracy of their big data analytics (2016).

59% of executives do not believe their company has capabilities to generate business insights from their data (2016).

(Dilbert cartoon by Scott Adams; © 2014)
Data Quality Dimensions

- Accuracy
- Completeness
- Consistency
- Validity
- Timeliness
- Uniqueness

DAMA UK Working Group, 2013
Data Quality Problems & Techniques

- **Data unification**
  - Schema mapping (if schemas exist)
  - Deduplicating records
  - Classification and mastering

- **Data repair**
  - Spotting errors and violations (e.g., outliers)
  - Repairing incorrect values
  - Missing value imputation
Data Analytics

The application of statistical and machine learning techniques to draw insights from data under study and to make predictions about the behaviour of the system under study.
Data Analytics

The application of statistical and machine learning techniques to draw insights from data under study and to make predictions about the behaviour of the system under study

- Statistics
- Computer Science (DM/ML)
Data Analytics

The application of statistical and machine learning techniques to draw insights from data under study and to make predictions about the behaviour of the system under study.

- Statistics
- Computer Science (DM/ML)
Data Analytics

The application of statistical and machine learning techniques to draw insights from data under study and to make predictions about the behaviour of the system under study.

- Statistics
- Computer Science (DM/ML)
- The lines between the two disciplines have blurred
Data Analytics Types

**Descriptive**
- What does the data reveal about what is happening?
- Exploratory analysis

**Diagnostic**
- Why is it happening?
- What does the data suggest about the reasons?

**Predictive**
- What is likely to happen?
- Decisions are affected
- Machine learning fits here

**Prescriptive**
- Recommended actions

[Diagram showing the types of data analytics: Descriptive, Diagnostic, Predictive, Prescriptive, with a value and complexity axis.]

[Link: https://www.kdnuggets.com/2017/07/4-types-data-analytics.html]
Data Analytics Tasks

- **Clustering**
  - Grouping objects into clusters

- **Outlier detection**
  - Detection of anomalous (rare) data items

- **Association rule mining**
  - Detecting relations between variables

- **Prediction**
  - Classification and regression
Data Security & Privacy
Dimensions of Data Protection

DATA PROTECTION

SECURITY
- Encryption
- Network Security
- Access Control
- Activity Monitoring
- Breach Response
- DLP/CASB

How those policies got enforced

PRIVACY
- Discovery & Classification
- DSARs
- 3rd-party management
- Data Removal
- Consents
- Policies

What data is important and why

https://dataprivacymanager.net/security-vs-privacy/
# Challenges

## Human-in-the-loop
- Many data science processes involve humans, but controlling information in humans is different than in computer systems

## Unintended side effects
- Traces of raw data persist into the latest steps of the data science process
- The combination of two data sources may reveal more than their “sum”

## Distinct application requirements
- Aggregate data analysis is different from transaction analysis and different security and privacy mechanisms are needed

## Inherent limitations
- Cannot have performance, accuracy (or utility) and security (or privacy) at the same time. At least one needs to go.
Different Concepts of Security

Traditional Security & Privacy
- Confidentiality
  - Do not reveal data to unauthorized users
- Integrity
  - Unauthorized users should not be able to modify data

Data Security & Privacy in Data Science
- Privacy
  - Enable users to control their data usage by others
- Veracity
  - Data provided should be true and current
“… the branch of ethics that studies and evaluates moral problems related to data, … algorithms, … and corresponding practices, in order to formulate and support morally good solutions.”
“… the branch of ethics that studies and evaluates moral problems related to data, … algorithms, … and corresponding practices, in order to formulate and support morally good solutions.”

Ethics of Data

Ownership
- Who has ownership of data?
- Typically, individuals should have ownership

Transparency
- Subjects should know that data about them is being collected, stored and will be processed and how
- Consent

Privacy
- Personal identifiable information

Intention
- What are you planning to do with the data?
- Secondary use
Ethics of Algorithms – Algorithmic Bias

Training data set → Data collection

Feature engineering → Outcome & Conclusions

Actual data set

Training phase
Ethics of Algorithms – Algorithmic Bias

Data collection

Feature engineering

Outcome & Conclusions

Training data set

Actual data set

Training phase

Deployment phase
Ethics of Algorithms – Algorithmic Bias

1. Data collection
   - Historical or representational bias
2. Feature engineering
3. Outcome & Conclusions

Training phase:
- Training data set

Deployment phase:
- Actual data set
Ethics of Algorithms – Algorithmic Bias

Data collection

Training data set

Feature engineering

Historical or representational bias

Inclusion or omission of features will introduce bias

Outcome & Conclusions

Actual data set

Training phase

Deployment phase
Ethics of Algorithms – Algorithmic Bias

- **Data collection**
  - Training data set
  - Historical or representational bias

- **Feature engineering**
  - Inclusion or omission of features will introduce bias

- **Outcome & Conclusions**
  - Actual data set
  - Unmeasurable outcomes & use of proxies will introduce bias

- **Phases**
  - Training phase
  - Deployment phase
Examples of Algorithmic Bias
Examples of Algorithmic Bias

Dissecting racial bias in an algorithm used to manage the health of populations

Ziad Obermeyer1,2,*, Brian Powers3, Christine Vogel3, Sendhil Mullainathan1,2,3,4

* See all authors and affiliations

DOI: 10.1126/science.aax2342

Racial bias in health algorithms

The U.S. health care system uses commercial algorithms to guide health decisions. Obermeyer et al. find evidence of racial bias in one widely used algorithm, such that Black patients assigned the same level of risk by the algorithm are sicker than White patients (see the Perspective by Benjamin). The authors estimated that this racial bias reduces the number of Black patients identified for extra care by more than half. Bias occurs because the algorithm uses health costs as a proxy for health needs. Less money is spent on Black patients who have the same level of need, and the algorithm thus falsely concludes that Black patients are healthier than equally sick White patients. Reformulating the algorithm so that it no longer uses costs as a proxy for needs eliminates the racial bias in predicting who needs extra care.

Science, this issue p. 447; see also p. 421
Examples of Algorithmic Bias

Machine Bias
There's software used across the country to predict future criminals. And it's biased against blacks.

by Julia Angwin, Jeff Larson, Surya Mattu and Lauren Kirchner, ProPublica
May 23, 2016

On a spring afternoon in 2014, Brisha Borden was running late to pick up her god-sister from school when she spotted an unlocked kid's blue Huffy bicycle and a silver Razor scooter. Borden and a friend grabbed the bike and scooter and tried to ride them down the street in the Fort Lauderdale suburb of Coral Springs.
Examples of Algorithmic Bias

AI Is the Future—But Where Are the Women?

Just 12 percent of machine learning researchers are women—a worrying statistic for a field supposedly reshaping society.
Examples of Algorithmic Bias – Gender Shades
Data Ethics Checklist

- Have we listed how this technology can be attacked or abused? [SECURITY]
- Have we tested our training data to ensure it is fair and representative? [FAIRNESS]
- Have we studied and understood possible sources of bias in our data? [FAIRNESS]
- Does our team reflect diversity of opinions, backgrounds, and kinds of thought? [FAIRNESS]
- What kind of user consent do we need to collect to use the data? [PRIVACY/TRANSPARENCY]
- Do we have a mechanism for gathering consent from users? [TRANSPARENCY]
- Have we explained clearly what users are consenting to? [TRANSPARENCY]
- Do we have a mechanism for redress if people are harmed by the results? [TRANSPARENCY]
- Can we shut down this software in production if it is behaving badly? [TRANSPARENCY]
- Have we tested for fairness with respect to different user groups? [FAIRNESS]
- Have we tested for disparate error rates among different user groups? [FAIRNESS]
- Do we test and monitor for model drift to ensure our software remains fair over time? [FAIRNESS]
- Do we have a plan to protect and secure user data? [SECURITY]
Issues at the Intersections

- Data science components should not be siloed
- Many important problems at the intersections remain to be solved
- Examples
  - Data visualization – Visual analytics
  - Data management – Machine Learning
    - DM for ML
    - ML for DM
  - Privacy & security – Ethics
Data Science Lifecycle

Data Acquisition → Data Preparation → Data Integration & Management → Modeling & Analysis → Explanation & Dissemination → Improved Decisions

Core:
- Security & Privacy
- Ethics, Policy & Social Impact

Applications:
- Application
- ...
- Application
Core Research Issues and Interactions

- Data lakes
- Big data processing
- Data platforms
- Metadata management

- Visualization for wider audience
- Visualization for data exploration
- Open data technologies

- Data cleaning
- Sampling
- Data provenance

- Data analysis
- Machine learning techniques for data analysis
Core Research Issues and Interactions

Data Preparation
- DM support for provenance
- Data preparation for big data management
- Cleaning for data analysis
- DM for ML
- ML for DM
- Visual analytics
- ... 

Big Data Management
- Data lakes
- Big data processing
- Data platforms
- Metadata management
- Visualization for wider audience
- Visualization for data exploration
- Open data technologies

Dissemination
- Data cleaning
- Sampling
- Data provenance

• Data analysis
• Machine learning techniques for data analysis
Data Science Lifecycle – Alternative
Data Science Lifecycle – Alternative
Data Science Lifecycle – Alternative

C. Shearer, The CRISP-DM Model, J. Data Warehousing, 2000
Data Science Lifecycle – Alternative

- Problem Definition
- Data Acquisition
- Data Preparation
- Data Modeling & Analysis
- Deployment (Explanation & Dissemination)
- Monitoring
Reference Architecture I

Concrete Architecture – Data Science Software Stack

Distributed Data Store

- Resource Management
  - NoSQL
  - Search
  - Stream Processing

- Data Analysis Tools
  - Script, SQL
  - Data Processing Framework

- Security/Privacy Technologies

Data
Concrete Architecture – Data Science Software Stack

- **Data Analysis Tools**: Mahout, Tableau
- **NoSQL**: MongoDB
- **Search**: Elasticsearch
- **Stream Processing**: Flink
- **Data Processing Framework**: Spark
- **Resource Management**: Hadoop, GlusterFS
- **Distributed Data Store**: HDFS
- **Security/Privacy Technologies**: SQL, DynamoDB, Neo4j

Data flow diagram with Data layers pointing to Data Analysis Tools and distributed data stores.
Who Owns Data Science?

Computer Science
- It is all AI

Statistics – Conway Diagram
- CS part is just hacking
Who Owns Data Science?

**Computer Science**
- It is all AI

**Statistics – Conway Diagram**
- CS part is just hacking
Who Owns Data Science

There seems to be great interest in this argument & in these diagrams

Who are the Constituents?
Who are the Constituents?

STEM – Core
People who are involved in developing the core technologies
Who are the Constituents?

**STEM – Core**
People who are involved in developing the core technologies

**STEM – Application**
People who are involved in data science applications in some domain
Who are the Constituents?

**STEM – Core**
People who are involved in developing the core technologies

**STEM – Application**
People who are involved in data science applications in some domain

**Non-STEM**
People in social sciences and humanities who might be involved in applications or data ethics or social aspects or policy issues
Who are the Constituents?

Data Science

- Humanities
- Machine/Statistical Learning
- Application Domain Expertise
- Visualization
- Mathematical Optimization
- Social Sciences
- Law
- Data Management
Who is a Data Scientist?

Core competencies

- In-depth knowledge of at least one of data engineering or data analytics pillars (expert level)
- Knowledge of the other two pillars of data security & privacy and data ethics (acquaintance)
- In-depth knowledge of at least one, preferably two, application areas (almost expert level)
- Ability to work in a team & communicate
Battle of the Venn Diagrams
Final Thoughts

- Data is central and it is increasing in volume and complexity
- Treat the data properly and it will tell a story
- Data science is multifaceted and multidisciplinary
- Data science may not yet be a discipline, but can become one
- The view I presented is from STEM (Computer Science) perspective
  - There is much more
Thank you