A SYSTEMATIC APPROACH TO DATA SCIENCE

M. TAMER ÖZSU
UNIVERSITY OF WATERLOO
WORLD’S MOST VALUABLE RESOURCE

“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.
“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 NEEDS POSITIONING
AGENDA

- What is Data Science
- Data Science Applications
- Data Science Ecosystem
- Data Science Lifecycle
- Data Science System Architecture
- Who Owns Data Science
What Is Data Science?

“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.”

“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.”

Chikio Hayashi
1998

“Data science combines multiple fields, including statistics, scientific methods, artificial intelligence (AI), and data analysis, to extract value from data. … Data science encompasses preparing data for analysis, including cleansing, aggregating, and manipulating the data to perform advanced data analysis.”

“Data science is an umbrella term to describe the entire complex and multistep processes used to extract value from data.”

Rafael A. Irizarry
2020-01-31

“Data science combines multiple fields, including statistics, scientific methods, artificial intelligence (AI), and data analysis, to extract value from data. … In turn, these systems generate insights which analysts and business users can translate into tangible business value.”

DataRobot

“Data science is the field of study that combines domain expertise, programming skills, and knowledge of mathematics and statistics to extract meaningful insights from data. … change of all sciences moving from observational, to theoretical, to computational and now to the 4th Paradigm – Data-Intensive Scientific Discovery”

Gordon Bell
2009

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

Gordon Bell
2009

“… 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.”

John Kelleher & Brendan Tierney
2018

“Data science encompasses a set of principles, problem definitions, algorithms, and processes for extracting non-obvious and useful patterns from large data sets.”

“Data science encompasses a set of principles, problem definitions, algorithms, and processes for extracting non-obvious and useful patterns from large data sets.”

“Implications for all sciences moving from observational, to theoretical, to computational and now to the 4th Paradigm – Data-Intensive Scientific Discovery”

Gordon Bell
2009

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

Gordon Bell
2009

“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.”

Chikio Hayashi
1998

“Data science is a multidisciplinary approach to extracting actionable insights from the large and ever-increasing volumes of data collected and created by today’s organizations. Data science encompasses preparing data for analysis and processing, performing advanced data analysis, and presenting the results to reveal patterns and enable stakeholders to draw informed conclusions.”

“Data science is the field of study that combines domain expertise, programming skills, and knowledge of mathematics and statistics to extract meaningful insights from data.”

DataRobot
WHAT IS DATA SCIENCE?

“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.”

“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.”

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

“Data science combines multiple fields, including statistics, scientific methods, artificial intelligence (AI), and data analysis, to extract value from data. … Data science encompasses preparing data for analysis, including cleansing, aggregating, and manipulating the data to perform advanced data analysis.”

Data science is the field of study that combines domain expertise, programming skills, and knowledge of mathematics and statistics to extract meaningful insights from data. … In turn, these systems generate insights which analysts and business users can translate into tangible business value.”

“… 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.”

“Data science encompasses a set of principles, problem definitions, algorithms, and processes for extracting non-obvious and useful patterns from large data sets.”

“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.”

“data science is an umbrella term to describe the entire complex and multistep processes used to extract value from data.”

“Data science is a multidisciplinary approach to extracting actionable insights from the large and ever-increasing volumes of data collected and created by today’s organizations. Data science encompasses preparing data for analysis and processing, performing advanced data analysis, and presenting the results to reveal patterns and enable stakeholders to draw informed conclusions.”

“Data science is the field of study that combines domain expertise, programming skills, and knowledge of mathematics and statistics to extract meaningful insights from data. … In turn, these systems generate insights which analysts and business users can translate into tangible business value.”

“Data science encompasses a set of principles, problem definitions, algorithms, and processes for extracting non-obvious and useful patterns from large data sets.”

“Data science is a multidisciplinary approach to extracting actionable insights from the large and ever-increasing volumes of data collected and created by today’s organizations. Data science encompasses preparing data for analysis and processing, performing advanced data analysis, and presenting the results to reveal patterns and enable stakeholders to draw informed conclusions.”

“Data science combines multiple fields, including statistics, scientific methods, artificial intelligence (AI), and data analysis, to extract value from data. … Data science encompasses preparing data for analysis, including cleansing, aggregating, and manipulating the data to perform advanced data analysis.”

Data science is the field of study that combines domain expertise, programming skills, and knowledge of mathematics and statistics to extract meaningful insights from data. … In turn, these systems generate insights which analysts and business users can translate into tangible business value.”

“data science is an umbrella term to describe the entire complex and multistep processes used to extract value from data.”

“… the terms data science, machine learning, and data mining are often used interchangeably.”
A WORKING DEFINITION

A data-driven approach to problem solving by analyzing and exploring large volumes of possibly multi-modal data, extracting from it knowledge and insight that is used for better decision-making.
A WORKING DEFINITION

A data-driven approach to problem solving by analyzing and exploring large volumes of possibly multi-modal data, extracting from it knowledge and insight that is used for better decision-making.

It involves the process of collecting, preparing, managing, analyzing, and explaining the data and analysis results.
DATA SCIENCE AS A UNIFIER
WHO IS A DATA SCIENTIST?

The Real Data Scientists of the Enterprise

What I could be doing  What people think I do  What I actually do
WHO IS A DATA SCIENTIST?

To be revealed at the end…
TWO MYTHS...
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 $\neq$ Big data
- Big data is like a raw material
- Processing it leads to data science & better understanding
- Applications are important
  - No applications $\rightarrow$ no data science

- Data science $\neq$ Machine learning $\neq$ 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
AGENDA

What is Data Science
Data Science Applications
Data Science Ecosystem
Data Science Lifecycle
Data Science System Architecture
Who Owns Data Science
DATA SCIENCE APPLICATIONS

• Data science is about applications
  • Applications give purpose
  • Applications inform core technologies

• Almost any domain 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

Sustainability

• Climate variability and change
• Ecology
• FEW
• Large data sources
  • Earth observation data
  • Remote sensing data
  • Citizen-science data
  • Ground-based observational data
  • High spatial and temporal resolution data from mobile devices
DATA SCIENCE APPLICATION EXAMPLES

**Biological & Biomedical**

- Bioinformatics
- Genomics
- Transcriptomics
- Proteomics
- Computational systems biology
- Mathematical and computational medicine

Explosion of data
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 ECOSYSTEM

Data Science Building Blocks

Data Engineering
- Big data management
- Data preparation

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

Data Protection
- Security for data science
- Data privacy

Data Ethics
- Impact on individuals, organizations & society
- Ethical & normative concerns
- Bias in data
- Algorithmic bias
- Regulatory issues
# DATA SCIENCE ECOSYSTEM

## Applications

### Data Science Building Blocks

<table>
<thead>
<tr>
<th>Data Engineering</th>
<th>Data Analytics</th>
<th>Data Protection</th>
<th>Data Ethics</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Big data management</td>
<td>• Explore data (data mining)</td>
<td>• Security for data science</td>
<td>• Impact on individuals, organizations &amp; society</td>
</tr>
<tr>
<td>• Data preparation</td>
<td>• Build models &amp; algorithms (machine learning)</td>
<td>• Data privacy</td>
<td>• Ethical &amp; normative concerns</td>
</tr>
<tr>
<td></td>
<td>• Visualizations &amp; visual analytics</td>
<td></td>
<td>• Bias in data</td>
</tr>
</tbody>
</table>

## Social and Policy Context

- Ethical & normative concerns
- Bias in data
- Algorithmic bias
- Regulatory issues
DATA SCIENCE ECOSYSTEM

Data Science Building Blocks

Data Engineering
- Big data management
- Data preparation

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

Data Protection
- Security for data science
- Data privacy

Data Ethics
- Impact on individuals, organizations & society
- Ethical & normative concerns
- Bias in data
- Algorithmic bias
- Regulatory issues
DATA ENGINEERING
Big data management

- Data enrichment, integration and storage
  - ETL/ELT process (?)
  - Data lakes

- Storage and management of big datasets
- Data processing platforms

Data preparation
DATA ENGINEERING

- Big data management
  - Data enrichment, integration and storage
    - ETL/ELT process (?)
    - Data lakes
  - Storage and management of big datasets
  - Data processing platforms
- Data preparation
  - Data acquisition/gathering
  - Data cleaning
  - Data provenance & lineage
DATA ENGINEERING IS ESSENTIAL
DATA ENGINEERING IS ESSENTIAL
DATA UNDERLYING DATA SCIENCE: BIG DATA – FOUR VS
“refers to large, diverse, complex, longitudinal, and/or distributed data sets generated from instruments, sensors, Internet transactions, email, video, click streams, and/or all other digital sources available today and in the future.”

NSF BIGDATA Solicitation
DATA UNDERLYING DATA SCIENCE: 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/incorrecness in data
• Data quality
DATA PREPARATION

Data Acquisition
Find data sources appropriate for the problem

Dataset Selection
Determine which datasets are most useful and appropriate

Data Integration
Integrate multi-modal data from multiple sources

Data Quality
Address all impurities and errors in the integrated data
DATA INTEGRATION

Big Data Sources

Federated Data Store

Transactional Data

Historical Data

Data Warehouse

ETL/ELT

Downstream Tasks

Schema-on-write
DATA INTEGRATION

Big Data Sources

Transaction Data

Historical Data

Federated Data Store

Downstream Tasks

Schema-on-read

ETL/ELT

Data Lake

Big Data Sources

Transaction Data

Historical Data

Federated Data Store

Downstream Tasks

Schema-on-read

ETL/ELT

Data Lake

Big Data Sources

Transaction Data

Historical Data

Federated Data Store

Downstream Tasks

Schema-on-read

ETL/ELT

Data Lake

Big Data Sources

Transaction Data

Historical Data

Federated Data Store

Downstream Tasks

Schema-on-read

ETL/ELT

Data Lake
“massive collection of datasets that:

• may be hosted in different storage systems;

• may vary in their formats;

• may not be accompanied by any useful metadata or may use different formats to describe their metadata; and

• may change autonomously over time.”

Margesian et al, Data Lake Management: Challenges and Opportunities, PVLDB, 2019.
DATA WAREHOUSES VS DATA LAKES

- Simpler to architect
- Single store
- Centralized analytics
- Privacy concerns

- Complexity of dealing with autonomous systems
- Distributed
- Federated/distributed analytics
- Maintain original ownership of 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 INTEGRATION ⇒ DATA QUALITY ISSUES

ARE YOU SURE THE DATA YOU GAVE ME IS CORRECT?

I'VE BEEN GIVING YOU INCORRECT DATA FOR YEARS. THIS IS THE FIRST TIME YOU'VE ASKED.

WHAT?

I SAID THE DATA IS TOTALLY ACCURATE.
DATA QUALITY DIMENSIONS

Accuracy
Completeness
Consistency
Timeliness
Validity
Uniqueness

DAMA UK Working Group, 2013
## DATA SCIENCE ECOSYSTEM

### Data Science Building Blocks

<table>
<thead>
<tr>
<th>Data Engineering</th>
<th>Data Analytics</th>
<th>Data Protection</th>
<th>Data Ethics</th>
</tr>
</thead>
</table>
| • Big data management  
• Data preparation  | • Explore data (data mining)  
• Build models & algorithms (machine learning)  
• Visualizations & visual analytics | • Security for data science  
• Data privacy | • Impact on individuals, organizations & society  
• Ethical & normative concerns  
• Bias in data  
• Algorithmic bias  
• Regulatory issues |
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.
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 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 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 reveals 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

DATA ANALYTICS TASKS/METHODS

Fayyad et al, From data mining to knowledge discovery in databases, AI Magazine, 1996.
## DATA ANALYTICS TASKS/METHODS

<table>
<thead>
<tr>
<th>Task</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Clustering</strong></td>
<td>• Discovering groups &amp; structures of data that are “similar”</td>
</tr>
<tr>
<td><strong>Outlier detection</strong></td>
<td>• Detection of anomalous (rare) data items</td>
</tr>
<tr>
<td><strong>Association rule learning</strong></td>
<td>• Detecting relations between variables</td>
</tr>
<tr>
<td><strong>Classification</strong></td>
<td>• Generalizing known structure to new data</td>
</tr>
<tr>
<td><strong>Regression</strong></td>
<td>• Find model that fits data with least error</td>
</tr>
<tr>
<td><strong>Summarization</strong></td>
<td>• More compact representation of the data set</td>
</tr>
</tbody>
</table>

Fayyad et al, From data mining to knowledge discovery in databases, *AI Magazine*, 1996.
### DATA ANALYTICS TASKS/METHODS

<table>
<thead>
<tr>
<th>Task</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clustering</td>
<td>Discovering groups &amp; structures of data that are “similar”</td>
</tr>
<tr>
<td>Outlier detection</td>
<td>Detection of anomalous (rare) data items</td>
</tr>
<tr>
<td>Association rule learning</td>
<td>Detecting relations between variables</td>
</tr>
<tr>
<td>Classification</td>
<td>Generalizing known structure to new data</td>
</tr>
<tr>
<td>Regression</td>
<td>Find model that fits data with least error</td>
</tr>
<tr>
<td>Summarization</td>
<td>More compact representation of the data set</td>
</tr>
</tbody>
</table>

DATA ANALYTICS TASKS/METHODS

- **Clustering**
  - Discovering groups & structures of data that are “similar”

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

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

- **Classification**
  - Generalizing known structure to new data

- **Regression**
  - Find model that fits data with least error

- **Summarization**
  - More compact representation of the data set

Fayyad et al, From data mining to knowledge discovery in databases, *AI Magazine*, 1996.
DATA ANALYTICS TASKS/METHODS

<table>
<thead>
<tr>
<th><strong>Clustering</strong></th>
<th>Discovering groups &amp; structures of data that are “similar”</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Outlier detection</strong></td>
<td>Detection of anomalous (rare) data items</td>
</tr>
<tr>
<td><strong>Association rule learning</strong></td>
<td>Detecting relations between variables</td>
</tr>
<tr>
<td><strong>Classification</strong></td>
<td>Generalizing known structure to new data</td>
</tr>
<tr>
<td><strong>Regression</strong></td>
<td>Find model that fits data with least error</td>
</tr>
<tr>
<td><strong>Summarization</strong></td>
<td>More compact representation of the data set</td>
</tr>
</tbody>
</table>

Fayyad et al, From data mining to knowledge discovery in databases, *AI Magazine*, 1996.
## DATA ANALYTICS TASKS/METHODS

<table>
<thead>
<tr>
<th>Method</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Clustering</strong></td>
<td>• Discovering groups &amp; structures of data that are “similar”</td>
</tr>
<tr>
<td><strong>Outlier detection</strong></td>
<td>• Detection of anomalous (rare) data items</td>
</tr>
<tr>
<td><strong>Association rule learning</strong></td>
<td>• Detecting relations between variables</td>
</tr>
<tr>
<td><strong>Classification</strong></td>
<td>• Generalizing known structure to new data</td>
</tr>
<tr>
<td><strong>Regression</strong></td>
<td>• Find model that fits data with least error</td>
</tr>
<tr>
<td><strong>Summarization</strong></td>
<td>• More compact representation of the data set</td>
</tr>
</tbody>
</table>

Fayyad et al, *From data mining to knowledge discovery in databases, AI Magazine, 1996.*
DATA ANALYTICS TASKS/METHODS

- **Clustering**
  - Discovering groups & structures of data that are “similar”

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

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

- **Classification**
  - Generalizing known structure to new data

- **Regression**
  - Find model that fits data with least error

- **Summarization**
  - More compact representation of the data set

Fayyad et al, From data mining to knowledge discovery in databases, AI Magazine, 1996.
ANALYTICS ARCHITECTURES
ANALYTICS ARCHITECTURES

Batch Analytics

Integrated Data

Analytics
ANALYTICS ARCHITECTURES

Batch Analytics

Federated Analytics

Integrated Data

Analytics

Analytics
ANALYTICS ARCHITECTURES

Batch Analytics

- Integrated Data
- Analytics

Federated Analytics

- Analytics

Realtime Analytics

- Realtime Data Store
- Analytics
DATA SCIENCE ECOSYSTEM

Data Science Building Blocks

Data Engineering
- Big data management
- Data preparation

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

Data Protection
- Security for data science
- Data privacy

Data Ethics
- Impact on individuals, organizations & society
- Ethical & normative concerns
- Bias in data
- Algorithmic bias
- Regulatory issues
DATA PROTECTION – DATA SECURITY & PRIVACY
DIMENSIONS OF DATA PROTECTION

- Proper handling, processing, storage and usage of information
- Privacy policies
- Data retention & deletion policies
- DSARs
- Third-party management
- User consent
- PETs

- Protecting information from any unauthorized access or malicious attacks
- Encryption
- TEEs
- Infrastructure security
- Access control
- Monitoring
- DLP
CHANGING CONCEPTS OF DATA PROTECTION

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
BIG DATA PRIVACY & SECURITY THREATS
DATA PROTECTION  ⇒  CYBERSECURITY

- Platform
- Network
- Software
- Data
CLOUD SECURE ALLIANCE RECOMMENDATIONS

- Infrastructure security
  - Distributed processing of data
  - Non-relational databases
- Data privacy
  - Privacy-preserving analysis
  - Cryptography
  - Granular access control
- Data management & integrity
  - Secure data storage & tx logs
  - Granular audits
  - Data provenance
- Reactive security
  - End-to-end filtering & validation
  - Real-time supervision of security
DATA SCIENCE ECOSYSTEM

Data Science Building Blocks

Data Engineering
- Big data management
- Data preparation

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

Data Protection
- Security for data science
- Data privacy

Data Ethics
- Impact on individuals, organizations & society
- Ethical & normative concerns
- Bias in data
- Algorithmic bias
- Regulatory issues
DATA ETHICS

“… 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.”

Bias

“inclination or prejudice for or against one person or group, especially in a way considered to be unfair a concentration on or interest in one area or subject a systematic distortion of a statistical result due to a factor not allowed for in its derivation”

Oxford English Dictionary

Bias is inherent in human decision-making
- Accuracy
- Speed
- Efficiency
TYPES OF BIAS IN HUMANS

Action-Oriented Biases
- Speedy decision-making
- van Restorff effect, bizarreness effect, overconfidence

Stability Biases
- Preference for the status quo
- Anchoring effect

Pattern Recognition Biases
- Recognizing patterns to fill-in gaps
- Educated guess, confirmation bias

Interest Biases
- What do I want?
- Social biases
  - groupthink
  - go along

Tobias Baer, Understand, Manage and Prevent Algorithmic Bias, 2019
TYPES OF BIAS IN DATA SCIENCE
TYPES OF BIAS IN DATA SCIENCE

Bias in Data

• Historical or representational bias
TYPES OF BIAS IN DATA SCIENCE

Bias in Data

• Historical or representational bias

Bias in Algorithms

• Inclusion or omission of features will introduce bias
TYPES OF BIAS IN DATA SCIENCE

Bias in Data

• Historical or representational bias

Bias in Algorithms

• Inclusion or omission of features will introduce bias
• Unmeasurable outcomes & use of proxies will introduce bias
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
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? [FAIRNESS]

• 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]
What is Data Science
Data Science Applications
Data Science Ecosystem
Data Science Lifecycle
Data Science System Architecture
Who Owns Data Science
DATA LIFECYCLE

Data: 
Acquire: Create, capture, gather from: Lab, Fieldwork, Surveys, Devices, Simulations, More.
Clean: Organize, Filter, Annotate, Clean.
Publish: Share: Data, Code, Workflows, Disseminate, Aggregate, Collect, Create portals, databases, and more, Couple with literature.
Preserve/Destroy: Store to: Preserve, Replicate, Ignore, Subset, compress, Index, Curate, Destroy.

F. Berman et al., Realizing the Potential of Data Science, Comm. ACM, 2018
Variations

- D. Agrawal et al., Challenges and Opportunities with Big Data, White paper for CCC of CRA, 2012.
- H.V. Jagadish, Big Data and Science: Myths and Reality, Big Data Research, 2015.

F. Berman et al., Realizing the Potential of Data Science, Comm. ACM, 2018
DATA SCIENCE LIFECYCLE

Similar to:
- CRISP-DM Model (C. Shearer, The CRISP-DM Model, J. Data Warehousing, 2000)
DATA SCIENCE LIFECYCLE

Similar to:
• CRISP-DM Model (C. Shearer, The CRISP-DM Model, J. Data Warehousing, 2000)

Problem Definition
Application requirements

Research Question

Deployment

Data Preparation

Data Analysis

Data Storage & Management

Dissemination of results
Explanation of data & results

Dataset determination
Dataset selection
Data ingest
Data quality issues
Data provenance issues

Statistical & ML models
Feature engineering
Model validation

Data integration
Large-scale distributed storage
Access interfaces
Data provenance mgmt
DATA SCIENCE LIFECYCLE

Similar to:
- CRISP-DM Model (C. Shearer, The CRISP-DM Model, J. Data Warehousing, 2000)

**Research Question**

**Data Preparation**
- Dataset determination
- Dataset selection
- Data ingest
- Data quality issues
- Data provenance issues

**Data Storage & Management**
- Data integration
- Large-scale distributed storage
- Access interfaces
- Data provenance mgmt

**Data Analysis**
- Statistical & ML models
- Feature engineering
- Model validation

**Deployment**
- Dissemination of results
- Explanation of data & results

**Data Security & Privacy**

**Data Ethics, Social & Policy Issues**

---

**Problem Definition**
- Application requirements

---

**Research Question**

**Data Preparation**

**Data Storage & Management**

**Data Analysis**

**Deployment**

---

**Data Ethics, Social & Policy Issues**

**Data Security & Privacy**
DATA SCIENCE LIFECYCLE

Research Question

Data Preparation

Data Storage & Management

Data Analysis

Deployment

Do sources provide data? Do collectors abuse data?

Can data be hidden in DBMS? Do aggregate data reveal things?

How to secure infrastructure? How to release data without info leak?

Does model reveal info about data? Does model behave as intended?

Similar to:
- CRISP-DM Model (C. Shearer, The CRISP-DM Model, J. Data Warehousing, 2000)
• 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
  • Data management support for provenance
  • Trustworthy data management
  • Privacy & security – Ethics
  • …
AGENDA

What is Data Science
Data Science Applications
Data Science Ecosystem
Data Science Lifecycle
Data Science System Architecture
Who Owns Data Science
NIST REFERENCE ARCHITECTURE (NBDRA)
NBDRA MAPPING TO NATIONAL SECURITY APPLICATIONS

Big Data Framework Provider
- Infrastructure
- Data Storage
- Messaging
- Processing
- Security
- Management
- Federation

Big Data Application Provider
- Application Orchestration
- Collection
- Preparation
- Analytics
- Visualization
- Access

Data Consumers

Data Providers

CONCRETE ARCHITECTURE – SOFTWARE STACK

Application Platform
- Data Integration System/Tools: Tamr, Talend
- Data Preparation Tools: Altair, Trifacta
- Analytics Systems/Tools: SAS, Mahout
- BI: Tableau, Power BI

Processing Platform
- Big Data Management Interface(s): Apache PIG, Apache SQL, Spark
- Cloud Data Processing: Acumenix, Hadoop
- Data Ingest/Integration: Talend, Taelend
- Data Preparation Tools: SAS, Mahout
- Analytics Systems/Tools: Tableau, Power BI
- Dissemination Tools: OpenRefine, Apache hive

System Platform
- Kafka: Apache Kafka, Apache Storm
- Clusters: Apache Mesos, Kubernetes

Computing Infrastructure
- Data Storage: HDFS, Ceph

Security/Privacy Technologies
- Adversarial Robustness Toolbox
- HElib
- Acumenix
- Microsoft SEAL
- OpenDP
ARCHITECTURE – PROCESS VIEW

Data Sources
- Streaming Data Ingest
- Static Data Storage

Processing
- Streaming Data Processing
- Batch Data Processing

Analytics
- Real-time Analytics
- Combined Analytics
- Batch Analytics

Reporting and Dissemination
AGENDA

What is Data Science
Data Science Applications
Data Science Ecosystem
Data Science Lifecycle
Data Science System Architecture
Who Owns Data Science
“many academic statisticians perceive the new programs as ‘cultural appropriation’ …

‘Insightful statisticians have for at least 50 years been laying the groundwork for constructing [data science] as an enlargement of traditional academic statistics.”

Aren’t We Data Science?

50 Years of Data Science
David Donoho
2017

WHO OWNS DATA SCIENCE?
TUG OF WAR BETWEEN CS & STATS
WHO OWNS DATA SCIENCE?
WHO OWNS DATA SCIENCE?

Statistics – Conway Diagram

• CS part is just hacking

WHO OWNS DATA SCIENCE?

Statistics – Conway Diagram
• CS part is just hacking

CS – Ullman Diagram
• Major CS role


WHO OWNS DATA SCIENCE?

Statistics – Conway Diagram
- CS part is just hacking

CS – Ullman Diagram
- Major CS role

CS Internal
- It is all AI


WHO OWNS DATA SCIENCE?

Statistics – Conway Diagram
• CS part is just hacking

CS – Ullman Diagram
• Major CS role


WHO OWNS DATA SCIENCE?

Statistics – Conway Diagram
- CS part is just hacking

CS – Ullman Diagram
- Major CS role


WHO ARE THE STAKEHOLDERS?
WHO ARE THE STAKEHOLDERS?

Core Technology

STEM people who are involved in developing the core technologies
WHO ARE THE STAKEHOLDERS?

Core Technology
STEM people who are involved in developing the core technologies

Application
People in STEM, social sciences or humanities who are involved in data science applications in some domain
WHO ARE THE STAKEHOLDERS?

Core Technology
STEM people who are involved in developing the core technologies

Application
People in STEM, social sciences or humanities who are involved in data science applications in some domain

Ethicists, Social, Policy
People in social sciences and humanities who are concerned with and work on data science ethics or social impact of data science or policy issues
WHO ARE THE STAKEHOLDERS?

- Data Science
- Humanities
- Machine/Statistical Learning
- Application Domain Expertise
- Visualization
- Mathematical Optimization
- Social Sciences
- Law
- Data Management
WHO IS A DATA SCIENTIST?
WHO IS A DATA SCIENTIST?

Core competencies
WHO IS A DATA SCIENTIST?

Core competencies

• In-depth knowledge of at least one of data engineering or data analytics pillars (expert level)
WHO IS A DATA SCIENTIST?

Core competencies

• In-depth knowledge of at least one of data engineering or data analytics pillars (expert level)

• Working knowledge of the other three pillars
WHO IS A DATA SCIENTIST?

Core competencies

• In-depth knowledge of at least one of data engineering or data analytics pillars (expert level)
• Working knowledge of the other three pillars
• In-depth knowledge of at least one, preferably multiple, application areas (almost expert level)
WHO IS A DATA SCIENTIST?

Core competencies

- In-depth knowledge of at least one of data engineering or data analytics pillars (expert level)
- Working knowledge of the other three pillars
- In-depth knowledge of at least one, preferably multiple, application areas (almost expert level)
- Ability to work in a team & communicate
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 to many colleagues who contributed to various initiatives I’ve led and who contributed to my understanding of data science.