Data Science Introduction

Dr.-Ing. Achim Basermann
German Aerospace Center (DLR)
Institute for Software Technology
Department Head “High-Performance Computing”
German Aerospace Center

• Research Institution
  • Research and development in aeronautics, space, energy, transportation, digitalization and security
  • National and international cooperations

• Space Agency
  • Planning and implementation of German space activities

• Project Management Agency
  • Research promotion
Approx. 9000 employees across 51 institutes and facilities at 28 sites.

Institute for Software Technology

- stands for **innovative software engineering**,  

- develops **challenging individual software solutions** for DLR, and  

- is partner in **scientific projects** in the area of simulation and software technology.
DLR Institute for Software Technology
Scientific Themes and Working Groups

Departments

- Intelligent and Distributed Systems
- High-Performance Computing
- Software for Space Systems and Interactive Visualization

Working Groups

- Distributed Software Systems
- Software Engineering
- Intelligent Systems
- Intelligent Algorithms and Optimization
- Parallel Numerics
- Quantum Computing
- Onboard Software Systems
- Modeling and Simulation
- Scientific Visualization
- Virtual and Extended Reality
High Performance Computing Teams

**Department**
High Performance Computing
Head: Dr. Achim Basermann
Deputy: Dr. Margrit Klitz

- **Intelligent Algorithms & Optimization**
  Dr. Martin Siggel

- **Quantum Computing**
  Dr. Tobias Stollenwerk

- **Parallel Numerics**
  Dr. Jonas Thies
Go to www.menti.com and use the code 96 37 79 (SciML: 88 94 91)

- Use your mobile
- Use your laptop
- Use your tablet
- Use your desktop

for input to mentimeter questions
Survey

• What characterizes a Data Scientist?

• Data Science Basics

• The Data Science Process

• AI and ML

• Common Applications

• Scientific Machine Learning

What characterizes a Data Scientist?
What characterizes a Data Scientist?

A data scientist is a person who has the knowledge and skills to conduct sophisticated and systematic analyses of data. A data scientist extracts insights from data sets for product development, and evaluates and identifies strategic opportunities.

• “There’s a joke running around on Twitter that the definition of a data scientist is ‘a data analyst who lives in California’,“ — Malcolm Chisholm

• “Data scientists are involved with gathering data, massaging it into a tractable form, making it tell its story, and presenting that story to others,” — Mike Loukides

• “A data scientist is a rare hybrid, a computer scientist with the programming abilities to build software to scrape, combine, and manage data from a variety of sources and a statistician who knows how to derive insights from the information within. S/he combines the skills to create new prototypes with the creativity and thoroughness to ask and answer the deepest questions about the data and what secrets it holds,” — Jake Porway
What characterizes a Data Scientist? (cont.)

- Data scientists are “analytically-minded, statistically and mathematically sophisticated data engineers who can infer insights into business and other complex systems out of large quantities of data,” — Steve Hillion

- “A data scientist is someone who blends math, algorithms, and an understanding of human behavior with the ability to hack systems together to get answers to interesting human questions from data,” — Hilary Mason

- Data scientist is a “change agent.” “A data scientist is part digital trendspotter and part storyteller stitching various pieces of information together.” — Anjul Bhambhri

- “The definition of “data scientist” could be broadened to cover almost everyone who works with data in an organization. At the most basic level, you are a data scientist if you have the analytical skills and the tools to ‘get’ data, manipulate it and make decisions with it.” — Pat Hanrahan

- “By definition all scientists are data scientists. In my opinion, they are half hacker, half analyst, they use data to build products and find insights. It’s Columbus meet Columbo – starry eyed explorers and skeptical detectives.” — Monica Rogati
What characterizes a Data Scientist?

• “A data scientist is someone who can obtain, scrub, explore, model and interpret data, blending hacking, statistics and machine learning. Data scientists not only are adept at working with data, but appreciate data itself as a first-class product.” — Daniel Tunkelang

• An ideal data scientist is “someone who has both the engineering skills to acquire and manage large data sets, and also has the statistician’s skills to extract value from the large data sets and present that data to a large audience.” — John Rauser

• Data scientist is “someone who can bridge the raw data and the analysis – and make it accessible. It’s a democratising role; by bringing the data to the people, you make the world just a little bit better,” — Simon Rogers
What characterizes a Data Scientist? (cont.)

• “A data scientist is an engineer who employs the scientific method and applies data-discovery tools to find new insights in data. The scientific method — the formulation of a hypothesis, the testing, the careful design of experiments, the verification by others — is something they take from their knowledge of statistics and their training in scientific disciplines. The application (and tweaking) of tools comes from their engineering, or more specifically, computer science and programming background. The best data scientists are product and process innovators and sometimes, developers of new data-discovery tools,” — Gil Press

• “A data scientist represents an evolution from the business or data analyst role. The formal training is similar, with a solid foundation typically in computer science and applications, modeling, statistics, analytics and math. What sets the data scientist apart is strong business acumen, coupled with the ability to communicate findings to both business and IT leaders in a way that can influence how an organization approaches a business challenge. Good data scientists will not just address business problems, they will pick the right problems that have the most value to the organization,” — IBM researchers

What characterizes a Data Scientist? (cont.)

**MATH & STATISTICS**
- Machine learning
- Statistical modeling
- Experiment design
- Bayesian inference
- Supervised learning: decision trees, random forests, logistic regression
- Unsupervised learning: clustering, dimensionality reduction
- Optimization: gradient descent and variants

**PROGRAMMING & DATABASE**
- Computer science fundamentals
- Scripting language e.g. Python
- Statistical computing packages, e.g., R
- Databases: SQL and NoSQL
- Relational algebra
- Parallel data access and parallel query processing
- MapReduce concepts
- Hadoop and Hive/Py
- Custom reducers
- Experience with IaaS like AWS

**DOMAIN KNOWLEDGE & SOFT SKILLS**
- Passionate about the business
- Curious about data
- Influence without authority
- Hacker mindset
- Problem solver
- Strategic, proactive, creative, innovative and collaborative

**COMMUNICATION & VISUALIZATION**
- Able to engage with senior management
- Story telling skills
- Translate data driven insights into decisions and actions
- Visual art design
- R packages like ggplot2, Shiny
- Knowledge of any visualization tools e.g. Flare, Tableau

**MODERN DATA SCIENTIST**

Data scientist, the sexiest job of the 21st century, requires a mixture of multidisciplinary skills ranging from an intersection of mathematics, statistics, computer science, communication and business. Finding a data scientist is hard. Finding people who understand who a data scientist is, is equally hard. So here is a little cheat sheet on who the modern data scientist really is.
What characterizes a Data Scientist? (cont.)
Data Science Basics
What is Data Science?
Models

![Diagram showing data processing and machine learning models](chart.png)
Models (cont.)
Data Science Tasks

- Regression
- Classification
- Clustering
- Feature selection
- Text mining
- Time series forecasting
- Association analysis
- Anomaly detection
- Recommendation Engines
- Deep Learning
<table>
<thead>
<tr>
<th>Tasks</th>
<th>Description</th>
<th>Algorithms</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Classification</td>
<td>Predict if a data point belongs to one of predefined classes. The prediction will be based on learning from known data set.</td>
<td>Decision Trees, Neural networks, Bayesian models, Induction rules, K nearest neighbors</td>
<td>Assigning voters into known buckets by political parties eg: soccer moms. Bucketing new customers into one of known customer groups.</td>
</tr>
<tr>
<td>Regression</td>
<td>Predict the numeric target label of a data point. The prediction will be based on learning from known data set.</td>
<td>Linear regression, Logistic regression</td>
<td>Predicting unemployment rate for next year. Estimating insurance premium.</td>
</tr>
<tr>
<td>Anomaly detection</td>
<td>Predict if a data point is an outlier compared to other data points in the data set.</td>
<td>Distance based, Density based, LOF</td>
<td>Fraud transaction detection in credit cards. Network intrusion detection.</td>
</tr>
<tr>
<td>Time series</td>
<td>Predict if the value of the target variable for future time frame based on history values.</td>
<td>Exponential smoothing, ARIMA, regression</td>
<td>Sales forecasting, production forecasting, virtually any growth phenomenon that needs to be extrapolated</td>
</tr>
<tr>
<td>Clustering</td>
<td>Identify natural clusters within the data set based on inherit properties within the data set.</td>
<td>K means, density based clustering - DBSCAN</td>
<td>Finding customer segments in a company based on transaction, web and customer call data.</td>
</tr>
<tr>
<td>Association analysis</td>
<td>Identify relationships within an item set based on transaction data.</td>
<td>FP Growth, Apriori</td>
<td>Find cross selling opportunities for a retailer based on transaction purchase history.</td>
</tr>
</tbody>
</table>
The Data Science Process
Data Science Process

- CRISP-DM
Process

1. Prior Knowledge
2. Preparation
3. Modeling
4. Application
5. Knowledge
1. Prior Knowledge

Gaining information on:

- Objective of the problem
- Subject area of the problem
- Data

Table 2.1 Data Set

<table>
<thead>
<tr>
<th>Borrower ID</th>
<th>Credit Score</th>
<th>Interest Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>01</td>
<td>500</td>
<td>7.31%</td>
</tr>
<tr>
<td>02</td>
<td>600</td>
<td>6.70%</td>
</tr>
<tr>
<td>03</td>
<td>700</td>
<td>5.95%</td>
</tr>
<tr>
<td>04</td>
<td>700</td>
<td>6.40%</td>
</tr>
<tr>
<td>05</td>
<td>800</td>
<td>5.40%</td>
</tr>
<tr>
<td>06</td>
<td>800</td>
<td>5.70%</td>
</tr>
<tr>
<td>07</td>
<td>750</td>
<td>5.90%</td>
</tr>
<tr>
<td>08</td>
<td>550</td>
<td>7.00%</td>
</tr>
<tr>
<td>09</td>
<td>650</td>
<td>6.50%</td>
</tr>
<tr>
<td>10</td>
<td>825</td>
<td>5.70%</td>
</tr>
</tbody>
</table>
2. Data Preparation

Data Exploration
Data quality
Handling missing values
Data type conversion
Transformation
Outliers
Feature selection
Sampling
3. Modeling

- Training Data → Build model
- Test Data → Evaluation
- Evaluation → Final Model
3. Modeling (cont.)

Spliting training and test data sets

### Table 2.3 Training Data Set

<table>
<thead>
<tr>
<th>Borrower</th>
<th>Credit Score (X)</th>
<th>Interest Rate (Y)</th>
</tr>
</thead>
<tbody>
<tr>
<td>01</td>
<td>500</td>
<td>7.31%</td>
</tr>
<tr>
<td>02</td>
<td>600</td>
<td>6.70%</td>
</tr>
<tr>
<td>03</td>
<td>700</td>
<td>5.95%</td>
</tr>
<tr>
<td>05</td>
<td>800</td>
<td>5.40%</td>
</tr>
<tr>
<td>06</td>
<td>800</td>
<td>5.70%</td>
</tr>
<tr>
<td>08</td>
<td>550</td>
<td>7.00%</td>
</tr>
<tr>
<td>09</td>
<td>650</td>
<td>6.50%</td>
</tr>
</tbody>
</table>

### Table 2.4 Test Data Set

<table>
<thead>
<tr>
<th>Borrower</th>
<th>Credit Score (X)</th>
<th>Interest Rate (Y)</th>
</tr>
</thead>
<tbody>
<tr>
<td>04</td>
<td>700</td>
<td>6.40%</td>
</tr>
<tr>
<td>07</td>
<td>750</td>
<td>5.90%</td>
</tr>
<tr>
<td>10</td>
<td>825</td>
<td>5.70%</td>
</tr>
</tbody>
</table>
3. Modelling (cont.)

Spliting training and test data sets

- Training Data
- Test Data
3. Modeling (cont.)

\[ y = 0.1036 - \frac{6.1}{100,000} x \]
3. Modeling (cont.)

Evaluation of test dataset

<table>
<thead>
<tr>
<th>Borrower</th>
<th>Credit Score (X)</th>
<th>Interest Rate (Y)</th>
<th>Model Predicted (Y)</th>
<th>Model Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>04</td>
<td>700</td>
<td>6.40%</td>
<td>6.09%</td>
<td>-0.31%</td>
</tr>
<tr>
<td>07</td>
<td>750</td>
<td>5.90%</td>
<td>5.79%</td>
<td>-0.11%</td>
</tr>
<tr>
<td>10</td>
<td>825</td>
<td>5.70%</td>
<td>5.33%</td>
<td>-0.37%</td>
</tr>
</tbody>
</table>
4. Application

- Product readiness
- Technical integration
- Model response time
- Remodeling
- Assimilation
5. Knowledge

Posterior knowledge

Data Science Pipeline

1. Raw data → Data wrangling → Data cleansing → Data preparation → Data engineering
2. Prepared data → Model learning → Model validation → Machine learning
3. Training data → Model learning → Model validation → Machine learning
4. Test data → Model learning → Model validation → Machine learning
5. Model deployment → Data visualization → Operations
Models of Data

- **Structured data**
  - Databases

- **Semi-structured data**
  - XML / JSON data
  - Email
  - Web pages

- **Unstructured data**
  - Audio
  - Video
  - Image data
  - Natural language
  - Documents
Data Science Workflow

1. Raw Data Collected
2. Data Is Processed
3. Clean Dataset
4. Models & Algorithms
5. Exploratory Data Analysis
6. Data Product
7. Communicate Visualize Report
8. Make Decisions
AI and ML
Artificial Intelligence (AI)

- Artificial Intelligence is the science of getting machines to think and make decisions like human beings do.
- Theory and development of computer systems
Importance of AI

- AI automates Repetitive Learning and discovery through data.
- AI adds intelligence to existing products.
- AI adapts through progressive learning algorithms to let the data do the programming.
- AI analyzes more and deeper data using neural networks that have many hidden layers.
- AI achieves incredible accuracy through deep neural networks.
Applications of AI

- **AI in Health Care**
Applications of AI (cont.)

- AI in Agriculture
Selected Domains of AI

- Neural Networks
  - Neural Networks are a class of models within the general machine learning literature.
Selected Domains of AI (cont.)

- **Robotics**
  - A branch of AI, which is composed of different branches and applications of robots
Selected Domains of AI (cont.)

- **Expert System**
  - A computer system that emulates the decision-making ability of a human expert
**Selected AI Technologies**

- **Machine Learning (ML)**
Selected AI Technologies (cont.)

- Deep Learning Platforms
Selected AI Technologies (cont.)

- Virtual Agents / Virtual Assistants
Classification of AI

- **Weak AI**
  - AI system that is designed and trained for a specific type of task
  - Also known as Narrow AI
Classification of AI (cont.)

➢ **Strong AI**

➢ AI system with generalized human cognitive abilities so that when presented with an unfamiliar task, it has enough intelligence to find a solution

➢ Also known as Artificial General Intelligence
AI, ML and DL

- It covers anything which enables the computers to behave like humans.
- It deals with the extraction of patterns from large data sets.
- It deals to train Deep Neural Networks so as to achieve better accuracy in those cases where former technologies were not performing up to the mark.
Need for Machine Learning (ML)

Increase in Data Generation

Uncover patterns & trends in data

Improve Decision Making

Solve complex problems
ML Process

Define Objective

Prediction

Data Gathering

Preparing Data

Model Evaluation

Building a Model

Data Exploration
ML Types

1. Supervised Learning
2. Unsupervised Learning
3. Reinforcement Learning
Understanding Supervised Learning

- A technique in which we teach or train the machine using data which is well labeled
Understanding Supervised Learning (cont.)
Understanding Unsupervised Learning

• Involves training by using unlabeled data and allowing the model to act on that information without guidance
Understanding Unsupervised Learning (cont.)

Unlabelled Data

Data Cleaning

Algorithm

Model Evaluation

Unlabelled Output

Clusters formed based on feature similarity

Unknown output

Understand patterns & discover output
Understanding Reinforcement Learning

- A part of ML where an agent is put in an environment and he learns to behave in this environment by performing certain actions and observing the rewards which it gets from those actions.
Understanding Reinforcement Learning (cont.)

- Hit and trial method of learning
- All about the interaction between the environment and the learning agent
- Exploration and exploitation
Understanding Reinforcement Learning (cont.)

1. Before Conditioning

Food \[\text{Response}\] \[\text{Salivation}\]

Unconditioned Stimulus \[\text{Unconditioned Response}\]
Understanding Reinforcement Learning (cont.)

2. Before Conditioning

Bell
Neutral stimulus
Response
No Salivation
No Conditioned Response
Understanding Reinforcement Learning (cont.)

3. During Conditioning

Bell + Food → Response

Salivation → Unconditioned Response

Doggie
Understanding Reinforcement Learning (cont.)

4. After Conditioning

Bell
Conditioned Stimulus
Response
Salivation
Conditioned Response
ML Tasks from a Statistical View

**Regression** – Looking for a statistical relationship across variables that may give us an estimate of a particular outcome.

**Classification** – Similar to regression but looking for separations in the data given predefined classes. (Supervised)

**Clustering** – Do not have predefined classes but trying to find groups or sets based upon data at hand. (Unsupervised)

**Anomaly Detection** – Identification of outliers based upon expected ranges of data.
Common Applications
Applications of Data Science

• Security
Applications of Data Science (cont.)

- Sports
Applications of Data Science (cont.)

- Banking and Finance
Applications of Data Science (cont.)

• Internet Search
Applications of Data Science (cont.)

- Digital Advertisements

2018 Digital Advertising Trends

- Facebook (54%) and Google (44%) remain the most influential platforms for advertising followed by Instagram, Spotify, and Pandora.
- 72% of consumers do not prefer video ads over other types of online advertisements, despite video being a major predicted trend in 2018.
- 45% of respondents said they are most likely to click an ad on their mobile device.
- Despite the focus on mobile-first design, 41 percent of respondents say that they are still most likely to click on an advertisement on a desktop device.
Applications of Data Science (cont.)

- Recommender Systems
Applications of Data Science (cont.)

- Image Processing
Applications of Data Science (cont.)

- Speech Recognition
Applications of Data Science (cont.)

- Gaming
Applications of Data Science (cont.)

- Price Comparison Websites

10 Price Comparison Websites

- mysmartprice
- compareindia
- priceDekho
- smartprix
- pricepanda
- buyhatke!
- junglee
- CompareKya
- naaptol
- India Book Store
Applications of Data Science (cont.)

- Airline Routing Planning
Applications of Data Science (cont.)

- Fraud and Risk Detection
Applications of Data Science (cont.)

• Delivery Logistics
Applications of Data Science (cont.)

- Internet of Things (IoT)
Applications of Data Science (cont.)

- Health Care
Applications of Data Science (cont.)

- Augmented Reality
Applications of Data Science (cont.)

- Self-Driving Cars
Applications of Data Science (cont.)

• Robots
Impact of Data Science on Society

- Saving Energy
Impact of Data Science on Society (cont.)

- Data-Driven Hospitals
Impact of Data Science on Society (cont.)

- A Cleaner Environment
Data Analytics Trends
Scientific Machine Learning
Scientific Machine Learning

“Scientific machine learning (SciML) is a core component of artificial intelligence (AI) and a computational technology that can be trained, with scientific data, to augment or automate human skills.

Across the Department of Energy (DOE), SciML has the potential to transform science and energy research. Breakthroughs and major progress will be enabled by harnessing DOE investments in massive data from scientific user facilities, software for predictive models and algorithms, high-performance computing platforms, and the national workforce.”
SciML: Role for Model Reduction

What role for model reduction?
1 reduce the cost of training 2 foundational shift in ML perspectives

Embed domain knowledge
Integrate heterogeneous, noisy & incomplete data

Respect physical constraints
Bring interpretability to results
Get predictions with quantified uncertainties
SciML: Predictive Digital Twin

via component-based ROMs and interpretable machine learning

ROMs embed predictive modeling and reduce the cost of training

**Offline:**

Construct library of ROMs representing different asset states

Use model library to train a classifier that predicts asset state based on sensor data

**Online:**

sensor data

current Digital Twin

[Analysis Prediction Optimization]

updated Digital Twin

[Kapteyn, Knezevic, W. AIAA Scitech 2020]
SciML: ML and MOR

Machine learning
“The scientific study of algorithms & statistical models that computer systems use to perform a specific task without using explicit instructions, relying on patterns & inference instead.” [Wikipedia]

Reduced-order modeling
“Model order reduction (MOR) is a technique for reducing the computational complexity of mathematical models in numerical simulations.” [Wikipedia]

What is the connection between reduced-order modeling and machine learning?

Model reduction methods have grown from Computational Science & Engineering, with focus on reducing high-dimensional models that arise from physics-based modeling, whereas machine learning has grown from Computer Science, with a focus on creating low-dimensional models from black-box data streams. [Swischuk et al., Computers & Fluids, 2019]
SciML: ML and MOR (cont.)

**Machine learning**
“The scientific study of algorithms & statistical models that computer systems use to perform a specific task without using explicit instructions, relying on patterns & inference instead.” [Wikipedia]

**Reduced-order modeling**
“Model order reduction (MOR) is a technique for reducing the computational complexity of mathematical models in numerical simulations.” [Wikipedia]

**Reduced-order modeling and machine learning:**
Can we get the best of both worlds?

- Discover hidden structure
- Non-intrusive implementation
- Black-box & flexible
- Accessible & available
- Embed governing equations
- Structure-preserving
- Predictive (error estimators)
- Stability-preserving
Learning from data through the lens of models is a way to exploit structure in an otherwise intractable problem.

- Embed domain knowledge
- Integrate heterogeneous, noisy & incomplete data
- Respect physical constraints
- Bring interpretability to results
- Get predictions with quantified uncertainties
SciML@DLR: Space Applications of Statistics and Machine Learning

Tools:
- Clustering, statistics
- Neural networks + feedback loops
- Optimal control theory
- Uncertainty quantification

Datafusion
- Measurements
- Simulation

Dimensional Reduction
- Feature Extraction
- Reduced Order Modelling

Real-time Prediction & Feedback
- Health Monitoring
- Robust Optimal Control
SciML@DLR: Software HeAT

- **HeAT** = Helmholtz Analytics Toolkit
- A framework for data analysis and Machine Learning, jointly developed by six Helmholtz Centers.
- Open Source with MIT License
- Available at [https://github.com/helmholtz-analytics](https://github.com/helmholtz-analytics)
Many thanks for your attention!

Questions?

Dr.-Ing. Achim Basermann
German Aerospace Center (DLR)
Simulation and Software Technology
Department Head High-Performance Computing
Achim.Basermann@dlr.de
http://www.DLR.de/sc

We are hiring: http://www.dlr.de/jobs/