EPSILON – Data Science for Social Good (DSSG) – Notes on the Video Lectures

Co-funded by the European Union

Project EPSILON was co-funded by the European Union (2021-1-DE01-KA220-HED-000029711). All views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or DAAD. Neither the European Union nor the granting authority can be held responsible for them.

All video lectures (voiced by Prof. Dr. Theo Berger) referred to in this script can be found as Open Educational Resources (OER) on YouTube: https://www.youtube.com/@epsilonproject2025

Contents

Introduction to the Learning Material

Overview and Objectives

Welcome to the initial session of our learning material. This session aims to offer an overview of the structure and usage of the resources available to you. We will also provide ideas on tailoring the materials to fit the needs of different target groups.

Our teaching material is designed to help you understand the fundamentals behind “Data for Good” projects in the European context. We aim to equip you with the necessary information and motivation to effectively engage with these resources. This material was born out of the EPSILON project, which stands for the **E**uropean **P**latform for **S**ocial Data Science **I**ncubation, **L**earning, **O**peration, and **N**etworking. This project involved partners from Germany, Portugal, Cyprus, and Lithuania, focusing on supporting European "Data for Good" initiatives. Our aim is to provide targeted learning materials for students, university instructors, and data science professionals. All materials are based on extensive research, interviews, benchmarking, and expert insights. Through this course, participants will gain knowledge and skills applicable to real-world challenges in Data Science for Good initiatives.

Figure 1: Logo of project EPSILON.

Four universities collaborated in developing this material: Harz University of Applied Sciences in Germany, Vytautas Magnus University in Lithuania, Nova School of Business and Economics in Portugal, and the University of Cyprus. These institutions were pivotal in both the conceptualization and creation of the learning content.

Structure and Target Groups

Our learning material is meticulously structured into six distinct sessions, catering to different levels of knowledge and expertise. It is designed for both self-study and classroom instruction. Each session can be consumed independently, allowing for flexibility in choosing topics of interest. Alternatively, students and teachers may opt to undertake the full course for a comprehensive understanding.

We have tailored our material to various target groups, including beginners, advanced data scientists, students, and enthusiasts for data science or social projects. This diverse approach ensures that each group has a specialized entry path into the topic, accommodating different learning needs and experience levels. The material is thoughtfully split into sections for beginners, who may lack programming skills or experience, and advanced learners with a background in computer science. This ensures that both ends of the spectrum are well-supported in their educational journey.

Detailed Content Overview

- Session 1: This module introduces fundamental concepts such as Big Data, data processing, and ethical considerations in data science.
- Session 2: What is – and who exactly does – Data Science for Social Good (DSSG)? A deep dive into how data science can contribute to social good, including an overview of key organizations, initiatives, and methodologies.

- Session 3: This module presents findings from a benchmark study conducted by the Nova School of Business & Economics, highlighting key indicators and insights concerning the Data Science for Good Landscape in the European Union.
- Session 4: This module provides a look at four projects showcasing DSSG Best Practices.
- Session 5: This module outlines essential aspects of successful DSSG projects, including project scoping, ethical concerns, data maturity, and project lifecycle management.
- Session 6: The final module recaps the key learnings and provides guidance for future engagement with Data Science for Good initiatives.

Our course includes detailed examinations of practical use cases. These examples provide insights into typical data science projects for social good. Beginners will find an accessible description of problem statements and derived solutions, while advanced users can explore the more technical aspects via GitHub repositories for practical engagement.

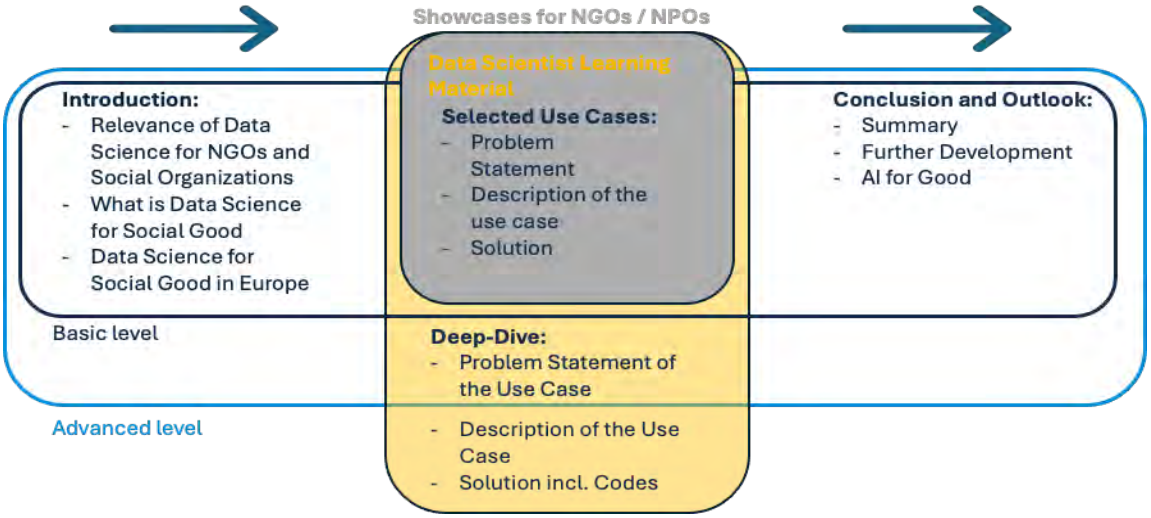

Figure 2: Structure of the EPSILON learning material.

Practical Application and Best Practices

We discuss four diverse projects:

- An intelligent data visualization tool from Paris.
- Tools for predicting and mitigating long-term unemployment by suggesting relevant training.
- A COVID-19 mortality surveillance platform in Portugal.
- A tool for visualizing domestic violence data, which involves complex data aggregation techniques.

Drawing from interviews, research, and identified benchmarks, we have compiled best practices for executing successful data science projects. This guidance includes project qualification criteria, data maturity concepts, and strategies for effective partnership and teamwork.

Project Lifecycle and Ethical Considerations

Understanding the typical lifecycle of a "data for good" project can enhance communication and organization. Frequent meetings and effective partner engagement are critical components emphasized in this cycle. Ethical and legal considerations are crucial when working with data. Our course provides insights into these issues, ensuring you are well-prepared to handle data responsibly in your projects.

Conclusion and Future Directions

In our final session, we offer a brief summary of the material covered and a forward-looking perspective on areas warranting further research and exploration. We invite you to engage with this material at your own pace and according to your learning preferences. We hope you find the sessions insightful and conducive to your objectives in data science for social good.

Introduction to Data Science

Understanding Data Science

In this introductory session to data science, we will explore the essential components that make up the field of data science and highlight the most significant aspects of this rapidly evolving discipline. The session is designed to provide a comprehensive overview, helping you understand the interconnected nature of data science.

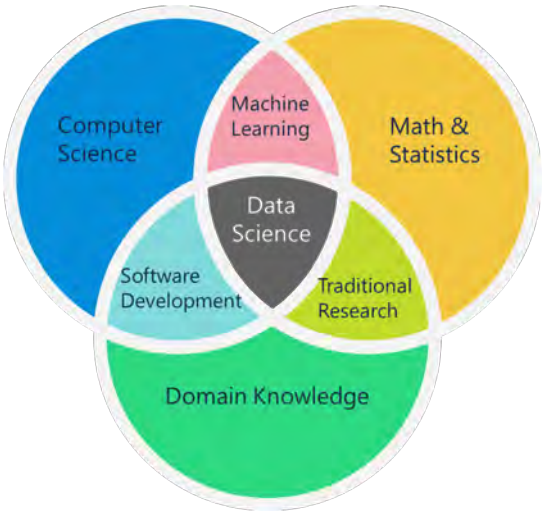

Figure 3: Elements of Data Science.

Data science is a **multifaceted field**, drawing from several disciplines. It combines elements of computer science, particularly software development and machine learning, with mathematical and statistical methodologies. This **interdisciplinary nature** allows data scientists to **tackle complex problems** by leveraging **domain-specific knowledge**. The term "Data Science" was introduced in the 1960s as a profession dedicated to handling and interpreting large datasets. Over time, it has evolved to become a **critical discipline** in business, research, and social initiatives.

Computer science contributes essential skills in software development and machine learning, which are crucial for creating algorithms capable of processing large datasets. **Mathematical and statistical knowledge** enables the development and refinement of models that can extract meaningful insights from data. **Machine learning** is particularly noteworthy as it represents a key component of artificial intelligence (AI). It involves training algorithms on large datasets to recognize patterns and make predictions, connecting traditional statistical methods with cutting-edge AI techniques.

Domain knowledge serves as the foundation for any data science project. It informs the formulation of research questions, which guide the entire data science process. These questions often stem from fields like economics, business administration, and increasingly, social sciences. Asking the right questions is crucial for ensuring that data-driven methods provide accurate and valuable answers.

In recent years, many data science questions have arisen from sectors such as business and computer science. However, as we will explore in subsequent sessions, social sciences are becoming a very rich source of inquiry, necessitating a solid understanding of these domains to frame questions effectively.

The Role of Big Data

The exponential increase in available data, driven by digital transformation and social media expansion, has made **big data** a central focus of data science. Analysing this vast quantity of information requires innovative approaches and robust processing capabilities.

Big data is generally characterized by five V's: Volume, Velocity, Variety, Veracity, and Value.

- **Volume** refers to the sheer quantity of data being generated, from diverse sources such as social media, transactions, and sensors.
- **Velocity** describes the speed at which data is generated and processed. In today's digital age, information is available almost instantaneously, requiring real-time processing capabilities.
- **Variety** includes the different types of data formats, both structured (like databases and spreadsheets) and unstructured (such as text and multimedia content).
- **Veracity** addresses the quality and reliability of the data, crucial for ensuring accurate data analysis and decision-making.
- **Value** emphasizes the importance of extracting meaningful insights from data, highlighting the potential business and societal benefits of data analysis.

Handling big data presents several challenges, including managing its volume and velocity while ensuring its quality (veracity). Additionally, the varied nature (variety) of data demands flexible processing strategies. Understanding the value of data often emerges only after it has been analysed, complicating initial data collection efforts.

Data Processing Workflow

Figure 4: Overview of the data processing workflow.

The data processing workflow begins with clear **problem framing**. This step involves articulating a specific research question or problem statement, informed by domain knowledge, to guide the data science project.

Data collection is a critical next step, involving the acquisition of relevant data from various sources. This could include APIs, sensors, or web scraping techniques. Once collected, data must be **cleaned and prepared** to address issues like missing values or outliers. This preparation is often the most time-consuming stage, demanding careful synchronization and integration of diverse data types.

During the **data analysis** phase, statistical and mathematical models are applied to distil insights from the prepared data. Selecting the appropriate models is crucial for the validity of any conclusions, with techniques ranging from traditional statistics to advanced machine learning algorithms.

Interpreting the results involves aligning the analysis outcomes with the original research question. Effective communication of these findings is essential, especially when sharing insights with stakeholders who may not have technical expertise. This requires skills in data visualization and results presentation to ensure clarity and relevance.

Ethical Considerations in Data Science

Ethical considerations are paramount in data science, particularly given the impact of data-driven decisions on individuals and society. Ethics guide the responsible use of data, helping to determine what is considered right or wrong in data collection, analysis, and application.

Several key ethical issues require attention:

- **Decision-Making Accountability**: Determining who is responsible for decisions made by algorithms, such as those in autonomous vehicles, is a critical ethical concern.
- **Data Privacy and Confidentiality**: It is essential to protect personal and sensitive data, ensuring it is used appropriately and with consent.
- **Data Ownership**: Understanding who owns the data and the implications of data sharing including questions of licensing and accountability are integral to ethical data management.

Addressing these ethical challenges involves setting clear guidelines and boundaries for data use as well as a balance between technological advancement and responsible usage to ensure fair and unbiased outcomes. This includes considering the direct and indirect impacts of data decisions, ensuring privacy and confidentiality, and clarifying data ownership responsibilities.

Conclusion

In this session, we have covered the components of data science, the challenges and characteristics of big data, the data processing workflow, and essential ethical considerations. These foundations provide a framework for understanding the complex and dynamic nature of data science.

As we continue in this course, we will delve into specific projects that illustrate these concepts in action, providing practical examples of data science at work. For further reading, we encourage you to explore the supplementary materials listed at the end of this script.

Social Data Science: Understanding its Role and Impact

Introduction to Social Data Science

Welcome to the second session of our EPSILON course on social data science. Building on our previous discussion about data science in general, in this session we will delve into the concept of data science for social good, often referred to as social data science. Our focus will be on defining social good, identifying relevant data, and understanding the various organizations and individuals involved in this impactful field. By the end of this session, you will have a clear understanding of how data science can be applied to address global challenges.

The session will follow this agenda:

- Define the concept of social good.
- Explore relevant datasets necessary for conducting social data science.
- Discuss the terminology and framework of data science for social good.
- Highlight the goals and key players in this space.

Understanding Social Good

Social good encompasses **services or products that enhance human well-being on a large scale**. It refers to **benefits that affect the largest number of people positively**, such as clean air, water, healthcare, and education. Social good addresses societal challenges, including climate change, human rights, food security, and housing crises.

Social data science aims to provide data-driven solutions to these societal challenges. Organizations dedicated to helping society play a crucial in tackling these issues through their focus on social good.

Data Sources for Social Good

Data plays a critical role in social good initiatives. Organizations rely on data from multiple sources to develop solutions for societal problems. These sources can be categorized into:

- **Internal Data Sources**: These are generated within an organization, such as operational data (staffing, client information) and outcome data (process documentation, service results, impact measurements). This data is often crucial for understanding internal processes and making informed decisions.
- **External Data Sources**: These include publicly accessible data and information from outside the organization, such as government statistics (e.g. unemployment rates) and other open data sources.

External data can be sourced from open data platforms such as Our World in Data, provided by the University of Oxford, offering insights into global health and demographics. Other sources include Statista and various national and regional databases that provide valuable context for social data science projects. By utilizing these datasets, organizations can analyse trends, predict outcomes, and develop data-driven solutions to social issues.

Data Science for Social Good

Data Science for Social Good (DSSG) bridges traditional data science with social impact, supporting projects that address social issues. It involves using data analytics, machine learning, and artificial intelligence in non-profit or public sectors to mitigate societal challenges.

The DSSG movement began in 2013 at the University of Chicago, aiming to train data scientists to use their skills for positive social impact. While initially academic, it has grown into a broader movement, engaging volunteers worldwide who offer data science services to NGOs and government bodies.

The primary goal of DSSG is to leverage data-driven decision-making to benefit non-profits, governments, and various social organizations. Volunteers, often professional data scientists and students, contribute their expertise to enhance public welfare projects.

Related Initiatives and Movements

Similar initiatives like DataKind and CorrelAid operate under the same philosophy of supporting social organizations through data science. These entities match volunteers with social projects, often through events like **hackathons** and **data dives**, where teams work on real-world challenges.

DSSG initiatives highlight a **vibrant community of data professionals dedicated to social good**. Volunteers typically work pro bono, using their skills to support initiatives in education, health, and public welfare, creating a broad positive impact on society.

Aside from academic roots, the DSSG movement has flourished globally, with numerous volunteer networks springing up across Europe. These networks are instrumental in fostering local collaborations between data scientists and social entities.

- DataKind: Founded in 2010/2011, DataKind was among the first organizations to match data scientists with nonprofits and small businesses. They organize "Datathons," which are 24-hour challenges where data scientists work intensively to solve social problems. They also offer summer fellowship programs to train data professionals.
- CorrelAid: A network of over 2,400 volunteer data scientists who offer pro bono services to nonprofit organizations. CorrelAid connects small teams of volunteers with NGOs in need of data expertise, supporting them in project planning, analysis, and consulting.

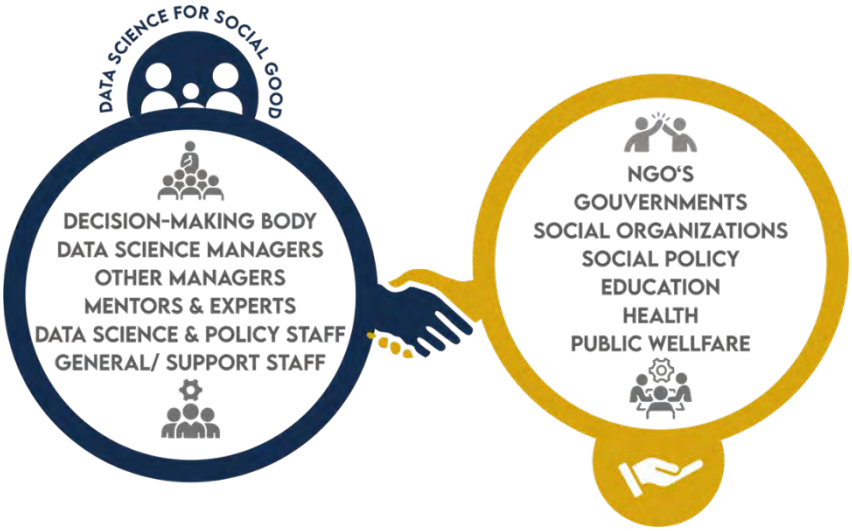

Figure 5: Who does Data Science for Social Good?

Conclusion

In this session, we've explored the foundations of social data science, defined social good, and examined the integral role of data in addressing societal challenges. We discussed the evolution of DSSG and the collaborative nature of volunteer-based initiatives like DataKind and CorrelAid. As we continue our exploration of social data science, this session provides context for understanding the powerful intersection of data science and social impact.

Comparative Analysis of Data for Good Initiatives

Introduction

In this third session on data science, we will look at a comparative analysis of "Data for Good" initiatives across Europe. In our previous session, we already explored the concept of social data science as a collaborative space filled with those diverse initiatives driven by volunteers who lend their data science expertise to social organizations.

The Landscape of Data for Good Initiatives

The European landscape of data for good initiatives is vast and varied, with many organizations and efforts dedicated to supporting social projects through data science. This overview builds on a benchmark report conducted as part of our European Union-funded project. The report, with major contributions from the Nova SBE Data, Operations & Technology Knowledge Center in Lisbon, provides an academic perspective on this growing field.

Our research identified 47 initiatives in Europe, with additional initiatives in the United States, Australia, Thailand, and other countries. While our focus is on Europe, it's important to note the global reach of data for good initiatives, even if the density of such initiatives is particularly high in Europe.

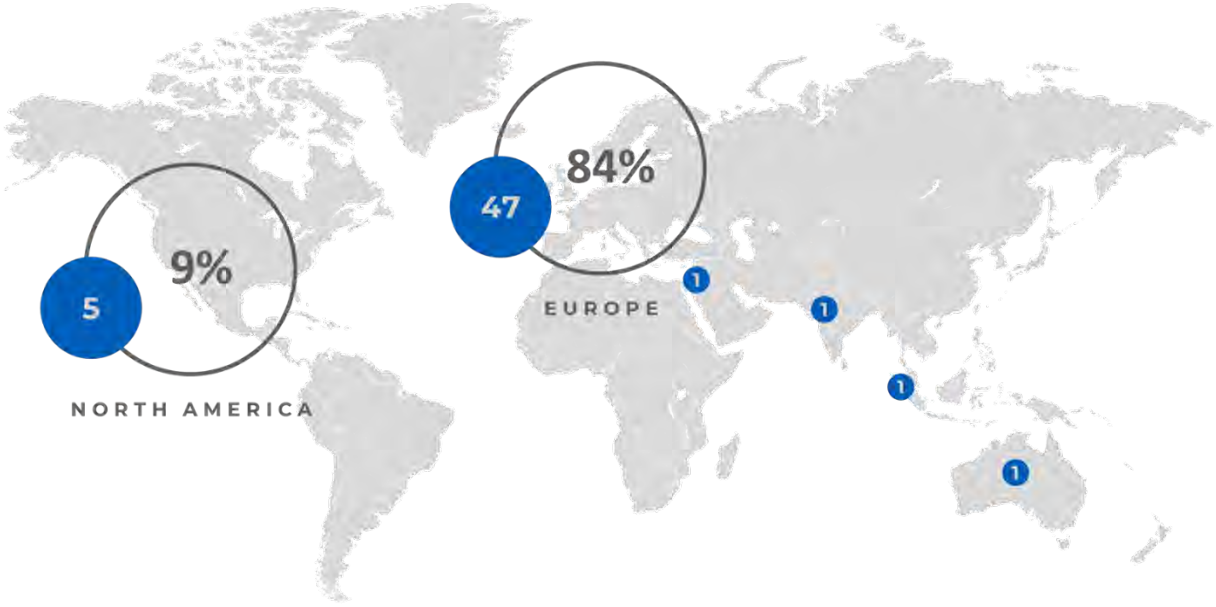

Figure 6: Global distribution of DSSG initiatives identified during project EPSILON.

Characterizing Initiatives

To better understand these initiatives, we identified five **key indicators**: Status and Legal Structure, General Operations, Working Methodology, Human Resources and Financing as well as Impact.

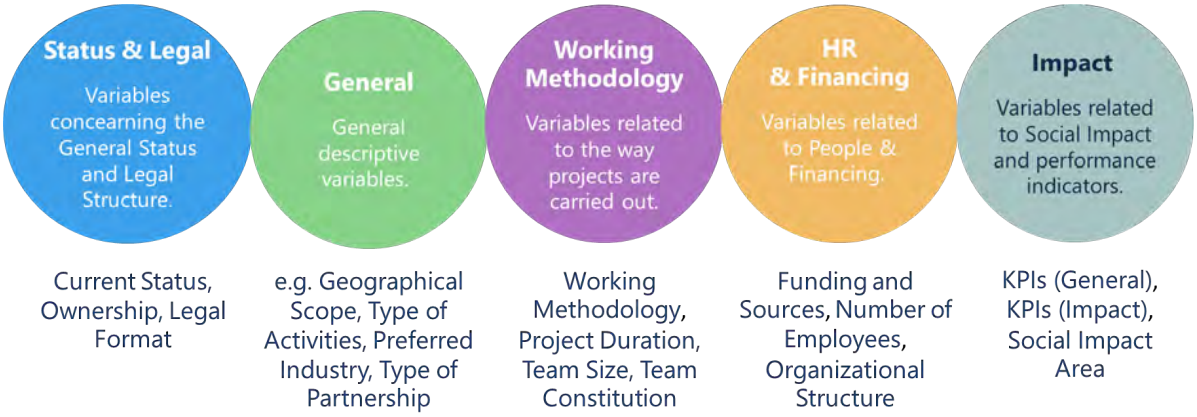

Figure 7: The key indicators used to characterize and compare the DSSG initiatives.

- **Status and Legal Structure**: Initiatives are largely volunteer-driven, with their legal formats varying by country and depending on individual preferences of the founders.
- **General Operations**: The type of services offered and partnerships formed heavily influence operations. These services vary based on the locality and the specific needs of beneficiaries.
- **Working Methodology**: Initiatives employ various methodologies, ranging from hackathons, where teams tackle problems in a short burst of activity, to longer projects requiring sustained commitment over several months. The choice of methodology often depends on volunteer availability and project requirements.
- **Human Resources and Financing**: Financing and human resources are critical to these initiatives. Funding sources may include corporate sponsors or be entirely volunteer-driven. This influences the legal responsibilities and organizational structure significantly.
- **Impact**: Impact is assessed through indicators like social impact measurements and project results. Key performance metrics include the number of projects completed and the community members involved, reflecting the initiative's reach and effectiveness.

Main Findings

Our survey found that **89% of the initiatives operate privately** and **93% are nonprofit**. These figures emphasize the **grassroots, volunteer-driven nature** of most organizations involved in social data science. Many initiatives have emerged in the last five years, highlighting a trend towards more active and engaged volunteer communities. The **high turnover of volunteers**, often due to personal commitments or professional changes, presents ongoing challenges in maintaining momentum and continuity. A common collaboration type is **data partnerships**, where social organizations provide datasets to volunteer data scientists who analyse and extract insights. This emphasizes the symbiotic relationship between data providers and analysts. **Volunteer motivations vary**, with many individuals aiming to do good, gain experience, or expand their professional networks. Initiatives leverage these motivations to attract participants and keep projects dynamic and impactful.

Typical projects last between four to six months, requiring careful alignment of volunteer availability with project demands. Coordination and project acquisition are crucial, as they involve nuanced understanding between social organizations and data scientists. Project **teams usually consist of six to eight members** with varied backgrounds, emphasizing the need for **diverse skill sets** and **collaboration among volunteers** to achieve successful outcomes.

Impact measurement remains a challenge, with many initiatives assessing their success based on project results. However, longer-term social impacts are less frequently measured due to resource constraints. The most common areas of focus are **health, well-being** and **climate action**, reflecting current societal challenges.

These areas drive the project themes and align with broader social development goals. In summary, data for good initiatives in Europe demonstrate a strong commitment to leveraging data science for societal benefit. Despite some clear challenges, the continued growth and engagement in this sector are promising, showing potential to tackle pressing social issues through data-driven approaches. By identifying key success factors and impact metrics, DSSG volunteers can support their organizations in driving meaningful social change.

Selected Use Cases in Social Data Science

Introduction

In this fourth session of our class on social data science, we will explore selected use cases to illustrate how data science can be applied effectively for social good. We will provide insights into four different projects, each showcasing a unique approach to addressing societal challenges through data science.

The goal of this session is to provide a detailed overview of typical data science projects in the realm of social good. By examining these use cases, you will gain a better understanding of the diverse applications of data science and how they can drive impactful decisions and solutions.

Use Case 1: Bicycle Parking Spots in Paris

Project Overview

Our first project focuses on an intelligent data visualization designed to enhance bicycle infrastructure in Paris. The objective was to support data-driven decision-making for the City of Paris Mobility Department to identify new bicycle parking locations.

Team and Stakeholders

The project involved several volunteers from CorrelAid, divided into data and research teams. The stakeholders were the City of Paris Mobility Department, aiming to support their transition to green mobility as part of an initiative to enhance cycling infrastructure.

Data and Methods

The team utilized open-government data, including existing bicycle parking locations, train station visitor statistics, traffic volume, and public spaces. Data aggregation, normalization, and cleansing were vital steps to ensure relevant, high-quality input for their intelligent dashboard design.

Results

A smart data visualization was created, enabling easy visualization of demand and supply indices across Paris. This tool helped identify areas with the highest parking demands, supporting efficient planning and resource allocation.

Figure 8: This dashboard provides an easy visual access to otherwise complex data (© CorrelAid).

Use Case 2: Predicting Long-Term Unemployment in Portugal

Project Overview

The second project aimed to develop tools for predicting long-term unemployment in Portugal, assisting counsellors in creating personalized action plans for job seekers.

Team and Stakeholders

Led by the Data Science for Social Good (DSSG) team in Portugal, the project was supported by experts from the Nova SBE Data, Operations & Technology Knowledge Center and involved collaboration with The Institute of Employment and Vocational Training.

Data and Methods

Using a comprehensive dataset of 3.1 million individuals, the team applied machine learning techniques to classify job seekers into high-risk and low-risk categories for long-term unemployment. A recommender system was also developed to suggest appropriate training interventions.

Results

The tools effectively identified high-risk individuals and recommended suitable training, thus enhancing the efficacy of counsellor interventions and reducing reliance on experience alone.

Use Case 3: COVID-19 Mortality Surveillance Platform

Project Overview

This project developed a COVID-19 mortality surveillance platform to streamline data collection and improve access to structured data during the pandemic.

Team and Stakeholders

Conducted by the DSSG Portugal team, the project was initiated without direct institutional sponsorship but responded to the urgent need for structured data solutions.

Data and Methods

The project addressed the lack of well-organized mortality data by creating a web scraping tool that transformed unstructured mortality rates into a downloadable and analysable format.

Results

A dynamic web scraping tool was developed, providing researchers with structured data, thereby aiding in timely data-driven decisions during the ongoing pandemic.

Use Case 4: Domestic Violence Data Observatory

Project Overview

The final project involved the creation of a Domestic Violence Data Observatory to consolidate various data sources related to domestic violence in Portugal.

Team and Stakeholders

Carried out by the DSSG Portugal team, the project aimed to address the fragmentation of existing data sources without direct organizational sponsorship.

Data and Methods

The team aggregated data from multiple reports and sources into a centralized repository, facilitating an accessible, user-friendly dashboard for stakeholders.

Results

An intuitive dashboard was developed, displaying the incidence of domestic violence across regions, thus simplifying data accessibility for policymakers and social workers.

Additional Projects and Conclusion

On our EPSILON platform, you will find further projects such as:

- Fraud detection in international development projects.
- Surveying interests of the elderly in rural areas.
- Optimizing waiting times in veteran hospitals.

This session provided an overview of diverse use cases in social data science. Each project demonstrated unique applications of data science to tackle societal challenges, offering valuable insights and solutions. For more information on any project, please visit our EPSILON platform, where you can access complete documentation, coding repositories, and points of contact.

Best Practices in Data Science for Social Good

Introduction

In this fifth and penultimate session of our course, we will dive into the best practices that have emerged from our research and practical experiences in this field. These practices are designed to guide you in creating impactful data for good projects that are both effective and sustainable.

Today's session will cover:

- Defining a data for good project and understanding its key characteristics.
- Necessary skills and resources for a successful project.
- Assessing data maturity and leveraging it for project success.
- Addressing ethical and legal considerations within projects.
- Exploring project lifecycle and management structures to ensure success.

Defining a Data for Good Project

A data for good project is distinguished by its **capacity to generate positive social impact while fostering learning and development for all parties involved**. Such projects typically involve partnerships with non-profit or public organizations that lack the resources to conduct rigorous data analysis independently.

When **selecting project partners**, it's essential to consider their legal structure and financial status, as well as their alignment with your organization's values. A strong partner relationship is built on shared goals and mutual benefits, ensuring that both volunteers and partners gain value from the collaboration.

Projects should be designed to provide a **win-win scenario**: volunteers can expand their skills and knowledge while the partner organization benefits from data-driven insights that enhance their operations or address societal challenges.

Project Requirements and Sustainability

Successful data for good projects necessitate a **team** composed of **diverse skill sets**, including expertise in data science, project management, and domain knowledge relevant to the project's focus area. It is crucial to match team members' strengths with project demands to maximize effectiveness.

Before initiating a project, assess the partner organization's **data maturity**. This assessment evaluates the availability, quality, and suitability of data for analysis. Understanding data maturity helps in planning the project's data-related requirements and ensuring that realistic goals are set.

Project outcomes should ideally be **sustainable**, providing lasting value for the partner organization. Deliverables should be comprehensively documented, ensuring that projects continue to benefit the organization after volunteers have moved on.

Ethical and Legal Concerns

Ethical considerations are paramount when dealing with data, especially personal or sensitive information. All projects should uphold high ethical standards and adhere to legal requirements, such as GDPR, to protect data privacy.

Wherever possible, **data should be anonymized to protect individual identities**. Volunteers must be cautious about data privacy and ensure that they have explicit permissions to use the data provided by partner organizations.

Organizations should have legal advisors to address any legal concerns that arise, including contract law and data privacy laws. Clear agreements with partners about data usage and project goals are essential to protect all parties involved.

Project Lifecycle and Management

The lifecycle of a data for good project typically includes:

- **Problem Identification**: Clearly understanding the issue that the project seeks to address.
- **Project Scoping**: Defining project goals, deliverables, and timelines in close collaboration with the partner.
- **Volunteer Recruitment and Onboarding**: Gathering a team with the necessary skills and orienting them to the project.
- **Project Execution**: Implementing project plans, continuously engaging with partners to ensure alignment.
- **Impact Assessment**: Evaluating the project's outcomes and the social impact achieved.

Projects benefit from a structured organization involving such roles as project managers, technical mentors, and data scientists. Clear responsibilities and communication pathways enhance team efficiency and project outcomes.

Common pitfalls

Data for Good projects may offer great transformative potential, but they also come with a unique set of challenges. Here are some common pitfalls that organizations might encounter:

- **Misaligned goals:** Lack of clarity on objectives can lead to unrealistic expectations between partners and volunteers. Projects without a clear "for-good" factor – where the social impact is not evident or where learning opportunities for volunteers are minimal – tend to fail.
- **Data challenges:** Organizations may struggle to obtain relevant data due to privacy concerns, technical limitations, or regulatory restrictions. Data may also lack completeness, accuracy, or relevance, hindering meaningful analysis.
- **Ethical concerns:** Insufficient focus on ethical issues such as privacy, discrimination or bias can harm affected groups or reinforce inequalities. Organizations may also fail to integrate ethical considerations throughout the project lifecycle.
- **Lack of sustainability:** Projects often lack mechanisms for continuity after volunteer teams leave. For instance, dashboards offering great real-time insights into data might be created but cannot be maintained by the organization due to technical or financial constraints.
- **Resource mismanagement:** Projects that require volunteers to perform monotonous tasks like data entry can lead to disengagement. Poor resource allocation or unclear roles within the team can result in inefficiency and delays.
- **Inadequate scoping:** Overly ambitious or underdeveloped project scopes can lead to failure. Examples for this problem are attempting to solve unsolvable problems or ignoring time and resource constraints.
- **Uncommitted or overwhelmed partners:** Partner organizations that are not fully committed or lack expertise can stall progress. Unrealistic expectations from partners, such as expecting volunteers to replace paid labour, can also derail a project.
- **Insufficient documentation:** Lack of thorough documentation can make it hard for stakeholders to understand or replicate outcomes, diminishing the project's long-term value.
- **Technical limitations:** Dependence on specific tools or technologies can exclude team members or create barriers for partner organizations. Over-reliance on automation may fail to capture human nuances crucial for social projects.
- **Ignoring legal regulations:** GDPR violations or similar legal oversights can result in meaningful penalties and reputational damage.

By addressing these pitfalls proactively – e.g. through proper scoping, ethical integration, data readiness, and sustainability planning – organizations can significantly enhance the success of their Data for Good projects.

Conclusion

In this session, we have explored the main components that make a data for good project successful – from selection and planning to execution and assessment. Emphasizing ethics and sustainability ensures that projects not only achieve immediate goals but also contribute to longer-term social benefits. By applying these best practices, you are equipped to contribute effectively to projects that harness data science for meaningful societal change. We conclude with four hands-on tips from the DSSG organisations interviewed by team EPSILON on how to maximise the efficiency and impact of DSSG projects:

- Define clear **objectives**, **timelines**, and **deliverables**.
- Use open tools like RMarkdown for **reproducibility** and interactive **reporting**.
- Document **processes** thoroughly, focusing on explanations rather than technical details.
- **Prioritize ethical considerations** at every stage to prevent harm and reinforce societal benefits.

Summary and Conclusion

Introduction

In this sixth and final session, we will try to summarize and conclude our exploration into the impactful world of data for good. Our goal has been to provide a comprehensive overview of how data science can be leveraged to address societal challenges and inspire meaningful change.

Course Overview and Key Takeaways

Throughout the course, we aimed to map the landscape of data for good, highlighting the robust ecosystem of initiatives and organizations active across Europe. We started by discussing foundational concepts in data science and illustrating their application in social good initiatives.

We encouraged you to independently initiate and engage in impactful projects by understanding the underlying motivations for data for good endeavours. The course emphasized the **power of data science as a catalyst for positive social change** and inspired students and professionals to take action.

Collaborative Networking and Best Practices

Our discussions underscored the **importance of networking and collaboration** within the data for good community. We promoted the establishment and expansion of initiatives, like the new Data Science Lab at Vilnius University in Lithuania, which connect volunteers with impactful opportunities.

We stressed the **importance of open-source solutions** in enabling data enthusiasts to learn and grow from shared experiences. By showcasing best practices, we highlighted the value of sharing project findings and code, thereby strengthening the broader data for good movement.

Challenges and Opportunities in the Data for Good Landscape

While the data for good landscape is rich and diverse, it is often **fragmented**, presenting challenges in coordinating and accessing vital resources. This fragmentation can hinder both volunteers seeking to join initiatives and organizations seeking consultation.

Many data for good organizations lack **visibility** outside specific communities, leading to underutilized resources. Effective **marketing** and **outreach strategies** are essential for maximizing the impact and reach of these initiatives.

Advancing Data for Good Initiatives

The success of data for good initiatives relies heavily on **effective knowledge management**. As volunteers come and go, maintaining continuity and sharing best practices are crucial for carrying forward organizational learning and impact.

Our course catered to various target groups:

- **Students**: Inspiring curiosity and showcasing how data science can solve real-world problems.
- **Data Enthusiasts**: Providing networking tools and resources to support project success.
- **Nonprofit Organizations**: Raising awareness of data science's potential for impactful decision-making and simplifying access to expertise.

Future Directions in Social Data Science

As we look forward, **integrating artificial intelligence into social data science** presents new opportunities and challenges. Ensuring AI solutions are fair, transparent, and free from bias is essential for maintaining public trust and integrity.

The focus should be on **developing scalable and sustainable solutions** that address complex social challenges across regions. Collaboration among sectors will be key to achieving significant, long-term impact.

Call to Action

We invite you to join the data science for good movement. Leveraging your skills to drive positive change in society can be incredibly rewarding. Visit our EPSILON platform to explore opportunities and connect with organizations making a difference.

Thank you for participating in this course. We hope to have sparked your interest and equipped you with the necessary tools to contribute meaningfully to the field of social data science. We look forward to seeing how you will make an impact and welcome you to continue this journey with us. Any questions or interests can be directed to our contacts at EPSILON.

Thank you for your attention. We are eager to hear from you.

Sources and Further Reading

1. Aula, V. & Bowles, J. (09.05.2023) Stepping back from Data and AI for Good – current trends and ways forward, Big Data & Society, p.2f, https://doi.org/10.1177/20539517231173901
2. Barak, M. (2020) The Practice and Science of Social Good: Emerging Paths to Positive Social Impact, an article in „Research on Social Work Practice Vol 30 (2)“ p.139-150 DOI: 10.1177/1049731517745600
3. Belford, G. & Tucker, A. (last updated 25.11.2023) Computer Science, https://www.britannica.com/science/computer-science
4. Bezuidenhout, L. & Ratti. E. (2021) What does it mean to embed ethics in data science? An integrative approach based on microethics and virtues, AI & Societies 36, p.939-953, https://doi.org/10.1007/s00146-020-01112-w
5. Cambridge Dictionary (last opened 23.11.2023), Research https://dictionary.cambridge.org/dictionary/english/research
6. Center for Data Science and Public Policy, University of Chicago (2018a): Data Maturity Framework, http://www.datasciencepublicpolicy.org/our-work/tools-guides/datamaturity/
7. Center for Data Science and Public Policy, University of Chicago (2018b): Data Science Project Scoping Guide, http://www.datasciencepublicpolicy.org/our-work/tools-guides/data-science-project-scoping-guide/
8. Coursera, 1 (last opened 23.11.2023), What is Machine Learning?, https://www.coursera.org/articles/what-is-machine-learning
9. Data Gouv France: Where to build new bicycle parking spots in Paris supporting data-driven decision making with open data. Available online at https://www.data.gouv.fr/en/reuses/where-to-build-new-bicycle-parking-spots-in-parissupporting-data-driven-decision-making-with-open-data/ (last accessed on 04.12.2023)
10. Data Orchard (2022): Data maturity framework for the not-for-profit sector, https://www.dataorchard.org.uk/resources/data-maturity-framework

11. Data Science for Social Good Foundation (2021): What Makes a Good Data Science for Social Good Project? https://www.dssgfellowship.org/2015/11/04/what-makes-a-good-dssg-project/
12. DSSG Portugal (a): Predicting long-term unemployment in Portugal. Available online at https://www.dssgfellowship.org/project/predicting-long-term-unemployment-in-continental-portugal/ (last accessed on 11.11.2024)
13. DSSG Portugal (b): Mortality Surveillance. Available online at https://www.dssg.pt/en/projects/mortality-surveillance/ (last accessed on 11.11.2024)
14. DSSG Portugal (c): Domestic Violence Data Observatory. Available online at https://www.dssg.pt/en/projects/domestic-violence-data-observatory/ (last accessed on 11.11.2024)
15. DSSG PT (n.d.): How we work, https://www.dssg.pt/en/how-we-work/
16. Egger, R. (2020) Applied Data Science in Tourism, Springer, ISBN 978-3-030-88389-8
17. Farmer, J., McCosker, A., Albury, K. & Aryani, A. (2023) Data for Social Good: Non-Profit Sector Data Projects, p. 10, https://doi.org/10.1007/978-981-19-5554-9
18. Foote, K. (16.10.2021) A Brief History of Data Science, https://www.dataversity.net/brief-history-data-science/
19. IBM (last opened 23.11.2023), Was ist Softwareentwicklung?, https://www.ibm.com/de-de/topics/software-development
20. International Association of Business Analytics Certification (last updated 14.05.2024): Data Science for Social Good: Making an Impact with Analytics, https://iabac.org/blog/data-science-for-social-good-making-an-impact-with-analytics
21. Kenton, W. (last updated 26.08.2022) Social Good: Definition, Benefits, Examples, https://www.investopedia.com/terms/s/social_good.asp
22. Kloppenburg & Moura, K. & Dietrich, J. (2023) CorrelAid/paris-bikes: Where to build new bicycle parking spots in Paris? Supporting data-driven decision making with open data. Available online at https://github.com/CorrelAid/paris-bikes/ (last accessed on 04.12.2023).
23. Majumder, P. (last updated 09.08.2023) Ethics in Data Science and Proper Privacy and Usage of Data, https://www.analyticsvidhya.com/blog/2022/02/ethics-in-data-science-and-proper-privacy-and-usage-of-data/
24. NOVA School of Business and Economics (2022), European Platform for Data Science: Incubation, Learning, Operations and Network – EPSILON Benchmark Report, Portugal
25. Scribblr (last opened 30.11.2023) Types of Bias in Research | Definition & Examples, https://www.scribbr.com/category/research-bias/

Contact Information

Prof. Dr. Philipp David Schaller
Phone +49 3943 – 297
Email pschaller@hs-harz.de
Friedrichstraße 57 – 59
38855 Wernigerode
Germany

With thanks to Prof. Dr. Theo Berger, Suntje Ehmann, Grit Lehmann, David Dommes, Stefan Apitz, Ellen Rabe, Christian Reinboth and our EPSILON partners at the Vytautas Magnus University in Lithuania, the Nova School of Business and Economics in Portugal, the University of Cyprus and DSSG Portugal for their contributions to this material as well as for editorial work and proofreading.