Training needs for data stewards

Workshop

Lucas Grus, WDCC

1st Data Steward @WUR network meeting, November 8, 2018

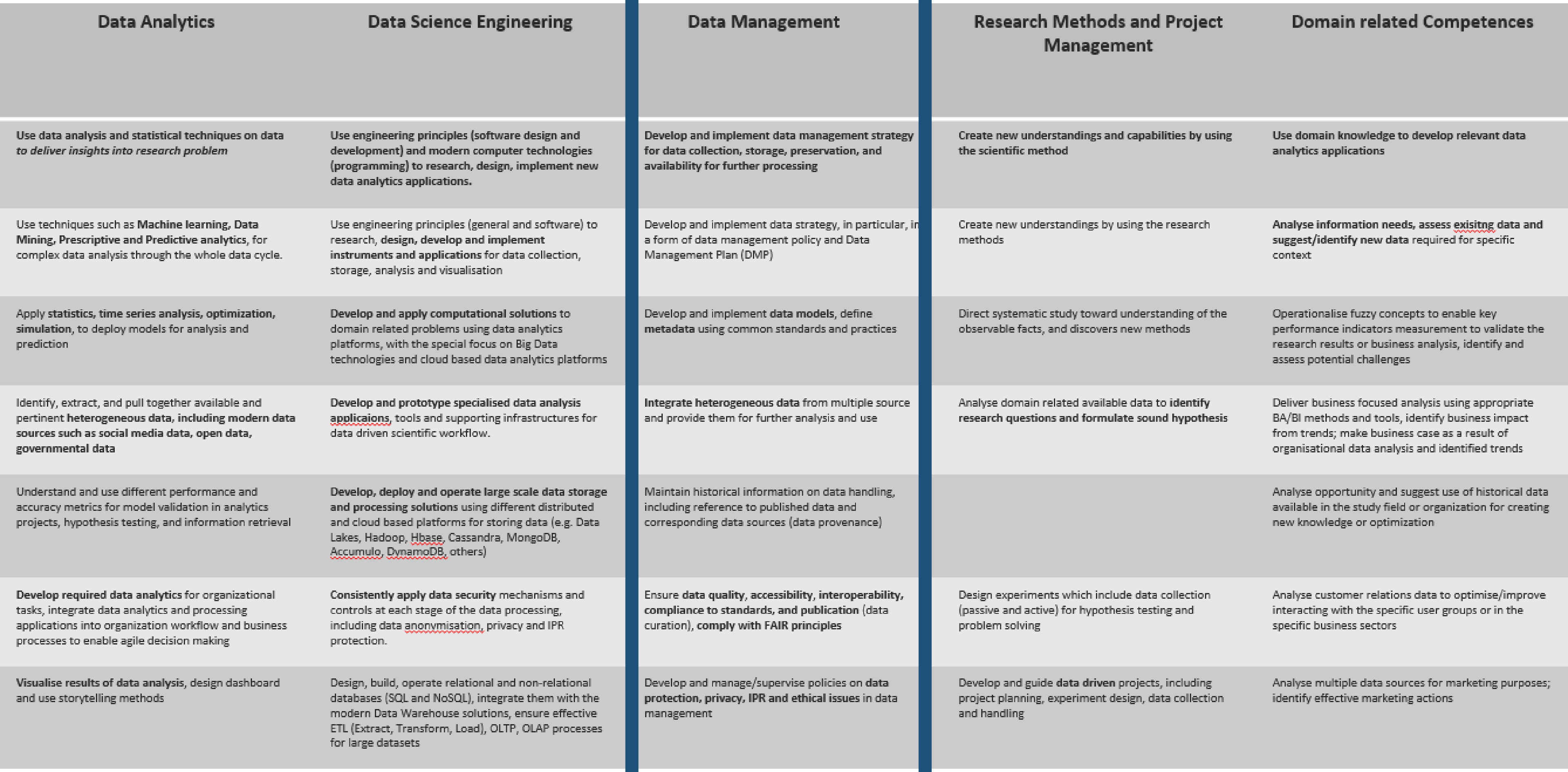
Een Data Steward binnen een eenheid heeft tot *doel* om:

1. Onderzoekers te adviseren en ondersteunen op het gebied RDM in de hele onderzoekscyclus
2. Het generieke RDM beleid te 'vertalen' naar praktische invulling binnen het wetenschappelijk domein.
3. De verbinding te vormen tussen Data Management Support en de onderzoekers
4. Kennisoverdracht over oplossingen van DMS naar het onderzoek

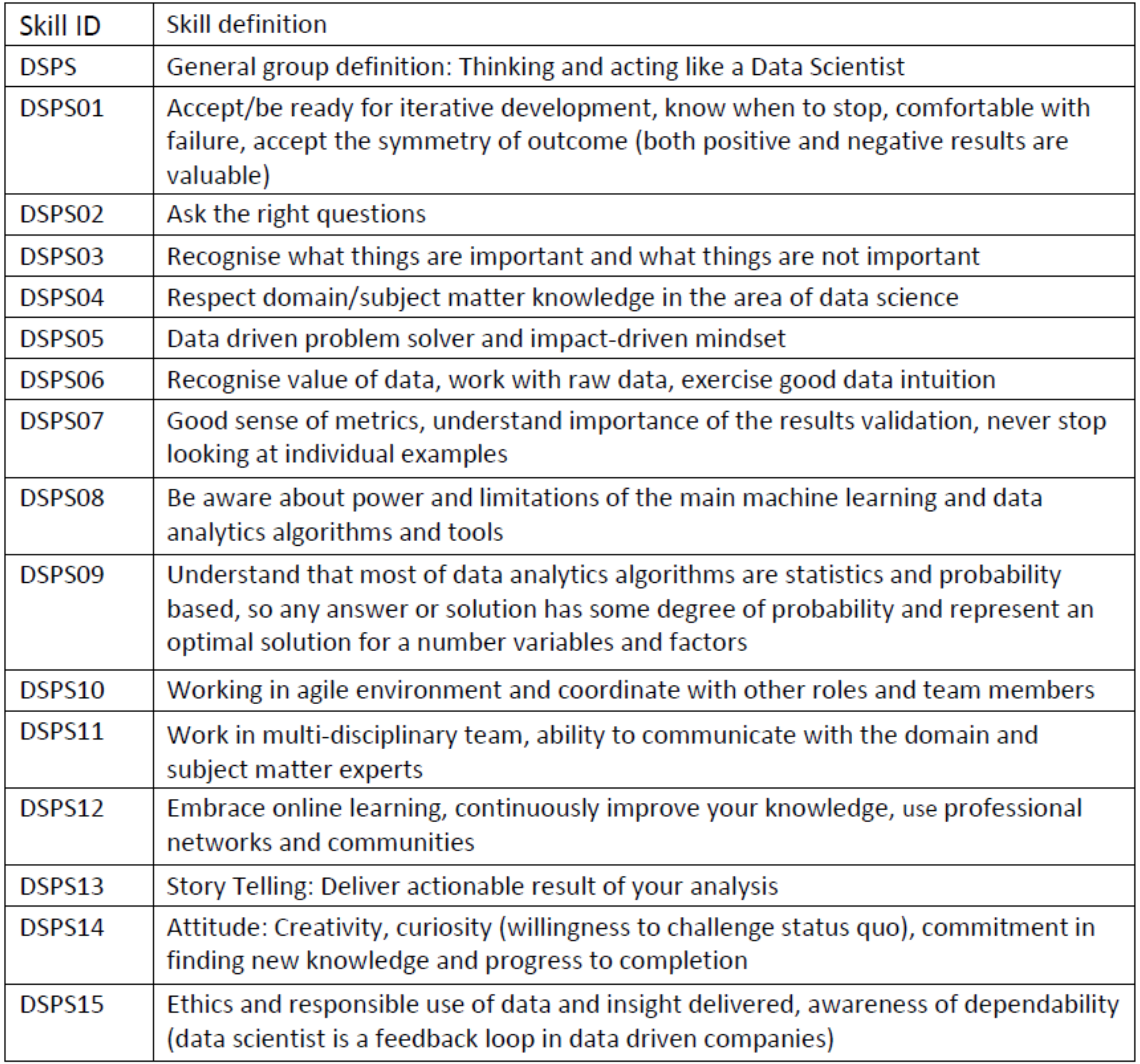
Profile Data scientist – competences (adapted from EDISON project)

Data Analytics	Data Science Engineering	Data Management	Research Methods and Project Management	Domain related Competences
Use data analysis and statistical techniques on data *to deliver insights into research problem*	**Use engineering principles (software design and development) and modern computer technologies (programming) to research, design, implement new data analytics applications.**	**Develop and implement data management strategy for data collection, storage, preservation, and availability for further processing**	**Create new understandings and capabilities by using the scientific method**	**Use domain knowledge to develop relevant data analytics applications**
Use techniques such as **Machine learning, Data Mining, Prescriptive and Predictive analytics**, for complex data analysis through the whole data cycle.	Use engineering principles (general and software) to research, **design, develop and implement instruments and applications** for data collection, storage, analysis and visualisation	Develop and implement data strategy, in particular, in a form of data management policy and Data Management Plan (DMP)	Create new understandings by using the research methods	**Analyse information needs, assess exisitng data and suggest/identify new data** required for specific context
Apply **statistics, time series analysis, optimization, simulation**, to deploy models for analysis and prediction	**Develop and apply computational solutions** to domain related problems using data analytics platforms, with the special focus on Big Data technologies and cloud based data analytics platforms	Develop and implement **data models**, define **metadata** using common standards and practices	Direct systematic study toward understanding of the observable facts, and discovers new methods	Operationalise fuzzy concepts to enable key performance indicators measurement to validate the research results or business analysis, identify and assess potential challenges
Identify, extract, and pull together available and pertinent **heterogeneous data, including modern data sources such as social media data, open data, governmental data**	**Develop and prototype specialised data analysis applicaions**, tools and supporting infrastructures for data driven scientific workflow.	**Integrate heterogeneous data** from multiple source and provide them for further analysis and use	Analyse domain related available data to **identify research questions and formulate sound hypothesis**	Deliver business focused analysis using appropriate BA/BI methods and tools, identify business impact from trends; make business case as a result of organisational data analysis and identified trends
Understand and use different performance and accuracy metrics for model validation in analytics projects, hypothesis testing, and information retrieval	**Develop, deploy and operate large scale data storage and processing solutions** using different distributed and cloud based platforms for storing data (e.g. Data Lakes, Hadoop, Hbase, Cassandra, MongoDB, Accumulo, DynamoDB, others)	Maintain historical information on data handling, including reference to published data and corresponding data sources (data provenance)		Analyse opportunity and suggest use of historical data available in the study field or organization for creating new knowledge or optimization
Develop required data analytics for organizational tasks, integrate data analytics and processing applications into organization workflow and business processes to enable agile decision making	**Consistently apply data security** mechanisms and controls at each stage of the data processing, including data anonymisation, privacy and IPR protection.	Ensure **data quality, accessibility, interoperability, compliance to standards, and publication** (data curation), **comply with FAIR principles**	Design experiments which include data collection (passive and active) for hypothesis testing and problem solving	Analyse customer relations data to optimise/improve interacting with the specific user groups or in the specific business sectors
Visualise results of data analysis, design dashboard and use storytelling methods	Design, build, operate relational and non-relational databases (SQL and NoSQL), integrate them with the modern Data Warehouse solutions, ensure effective ETL (Extract, Transform, Load), OLTP, OLAP processes for large datasets	Develop and manage/supervise policies on **data protection, privacy, IPR and ethical issues** in data management	Develop and guide **data driven** projects, including project planning, experiment design, data collection and handling	Analyse multiple data sources for marketing purposes; identify effective marketing actions

EDISON building the data science profession

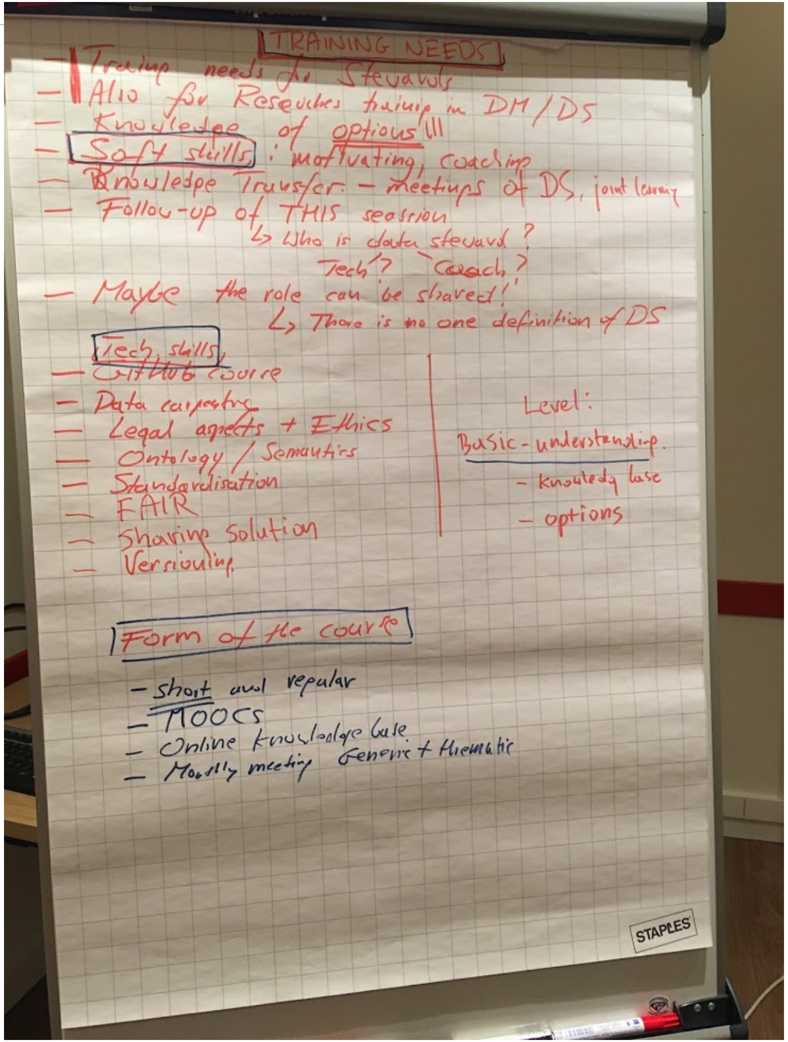


Data Analytics	Data Science Engineering	Data Management	Research Methods and Project Management	Domain related Competences
Use data analysis and statistical techniques on data *to deliver insights into research problem*	**Use engineering principles (software design and development) and modern computer technologies (programming) to research, design, implement new data analytics applications.**	**Develop and implement data management strategy for data collection, storage, preservation, and availability for further processing**	**Create new understandings and capabilities by using the scientific method**	**Use domain knowledge to develop relevant data analytics applications**
Use techniques such as **Machine learning, Data Mining, Prescriptive and Predictive analytics**, for complex data analysis through the whole data cycle.	Use engineering principles (general and software) to research, **design, develop and implement instruments and applications** for data collection, storage, analysis and visualisation	Develop and implement data strategy, in particular, in a form of data management policy and Data Management Plan (DMP)	Create new understandings by using the research methods	**Analyse information needs, assess exisitng data and suggest/identify new data** required for specific context
Apply **statistics, time series analysis, optimization, simulation**, to deploy models for analysis and prediction	**Develop and apply computational solutions** to domain related problems using data analytics platforms, with the special focus on Big Data technologies and cloud based data analytics platforms	Develop and implement **data models**, define **metadata** using common standards and practices	Direct systematic study toward understanding of the observable facts, and discovers new methods	Operationalise fuzzy concepts to enable key performance indicators measurement to validate the research results or business analysis, identify and assess potential challenges
Identify, extract, and pull together available and pertinent **heterogeneous data, including modern data sources such as social media data, open data, governmental data**	**Develop and prototype specialised data analysis applicaions**, tools and supporting infrastructures for data driven scientific workflow.	**Integrate heterogeneous data** from multiple source and provide them for further analysis and use	Analyse domain related available data to **identify research questions and formulate sound hypothesis**	Deliver business focused analysis using appropriate BA/BI methods and tools, identify business impact from trends; make business case as a result of organisational data analysis and identified trends
Understand and use different performance and accuracy metrics for model validation in analytics projects, hypothesis testing, and information retrieval	**Develop, deploy and operate large scale data storage and processing solutions** using different distributed and cloud based platforms for storing data (e.g. Data Lakes, Hadoop, Hbase, Cassandra, MongoDB, Accumulo, DynamoDB, others)	Maintain historical information on data handling, including reference to published data and corresponding data sources (data provenance)		Analyse opportunity and suggest use of historical data available in the study field or organization for creating new knowledge or optimization
Develop required data analytics for organizational tasks, integrate data analytics and processing applications into organization workflow and business processes to enable agile decision making	**Consistently apply data security** mechanisms and controls at each stage of the data processing, including data anonymisation, privacy and IPR protection.	Ensure **data quality, accessibility, interoperability, compliance to standards, and publication** (data curation), **comply with FAIR principles**	Design experiments which include data collection (passive and active) for hypothesis testing and problem solving	Analyse customer relations data to optimise/improve interacting with the specific user groups or in the specific business sectors
Visualise results of data analysis, design dashboard and use storytelling methods	Design, build, operate relational and non-relational databases (SQL and NoSQL), integrate them with the modern Data Warehouse solutions, ensure effective ETL (Extract, Transform, Load), OLTP, OLAP processes for large datasets	Develop and manage/supervise policies on **data protection, privacy, IPR and ethical issues** in data management	Develop and guide **data driven** projects, including project planning, experiment design, data collection and handling	Analyse multiple data sources for marketing purposes; identify effective marketing actions



Skill ID	Skill definition
DSPS	General group definition: Thinking and acting like a Data Scientist
DSPS01	Accept/be ready for iterative development, know when to stop, comfortable with failure, accept the symmetry of outcome (both positive and negative results are valuable)
DSPS02	Ask the right questions
DSPS03	Recognise what things are important and what things are not important
DSPS04	Respect domain/subject matter knowledge in the area of data science
DSPS05	Data driven problem solver and impact-driven mindset
DSPS06	Recognise value of data, work with raw data, exercise good data intuition
DSPS07	Good sense of metrics, understand importance of the results validation, never stop looking at individual examples
DSPS08	Be aware about power and limitations of the main machine learning and data analytics algorithms and tools
DSPS09	Understand that most of data analytics algorithms are statistics and probability based, so any answer or solution has some degree of probability and represent an optimal solution for a number variables and factors
DSPS10	Working in agile environment and coordinate with other roles and team members
DSPS11	Work in multi-disciplinary team, ability to communicate with the domain and subject matter experts
DSPS12	Embrace online learning, continuously improve your knowledge, use professional networks and communities
DSPS13	Story Telling: Deliver actionable result of your analysis
DSPS14	Attitude: Creativity, curiosity (willingness to challenge status quo), commitment in finding new knowledge and progress to completion
DSPS15	Ethics and responsible use of data and insight delivered, awareness of dependability (data scientist is a feedback loop in data driven companies)

Results from the workshop



TRAINING NEEDS

- Training needs for Stewards
- Also for Researches training in DM / DS
- Knowledge of options !!!
- Soft skills: motivating, coaching
- Knowledge Transfer: - meetups of DS, joint learning
- Follow-up of THIS seassion
  - ↳ Who is data steward ? Tech ? Coach ?
- Maybe the role can 'be shared!'
  - ↳ There is no one definition of DS

Tech skills

- GitHub course
- Data carpentry
- Legal aspects + Ethics
- Ontology / Semantics
- Standardisation
- FAIR
- Sharing solution
- Versioning

Level:

Basic-understanding.

- knowledg base
- options

Form of the course

- short and regular
- MOOCs
- Online knowledge base
- Monthly meeting Generic + thematic