Project Management for Data Science

Prof. Jochen L. Leidner, M.A. M.Phil. Ph.D.

University of Sheffield / Refinitiv

IEEE DSAA 2018 · 3 October 2018

Objectives

In this tutorial, I aim to:

- introduce you to some basic concepts and realities of project management as practiced in commercial and governmental organizations
- describe the Data-to-Value methodology for project management of data science projects (especially for those using NLP & ML)
- convey some best practices for data-centric projects

Methodology – Motivation (1/2)

When I started teaching big data and data science, I discovered

- there were no papers on methodology for data science (unlike software project management methodology)

When I had to mentor new team members in industry, I discovered

- there were no papers on methodology for data science (unlike software project management methodology)

$\Rightarrow$Need to try and fix that!

Fundamentals of Project Management

Methodology – Motivation (2/2)

- Why care about methodology?
  - (Relatively higher) consistency of outcome
  - Guidance to less experienced engineers
  - Provides a common set of assumptions, expectations & shared vocabulary in a team
  - Clarity of process reduces necessary coordination/communications (alignment)
  - Codification of best practices leads to a culture of continuous self-improvement

Project: A Definition

Project

A project is a *time-limited* activity to deploy *defined resources* to *effect change* with a *defined scope* with the aim to *benefit*.

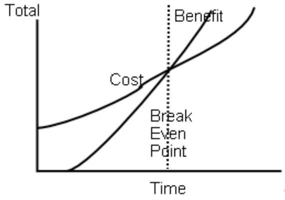
Project Success: 4 Criteria by the Project Management Institute (PMI)

A project is completed successfully if it is completed:

- on **time**,
- on **budget**,
- at performance level/**to specification**, and
- with **customer acceptance**.

Payback Period, Break-Even Point & ROI:

- **Payback Period**: time period until **break-even point (BEP)** is reached
- **Return on Investment (ROI)**: a measure, per period, of interest rate of return on money invested in an entity in order to decide whether to undertake an investment
- **break-even point (BEP)**: the point in time for which the gain from an investment less the cost of investment to obtain that gain equals zero
- $ROI = (\text{gain from investment} - \text{cost of investment})/\text{cost of investment}$

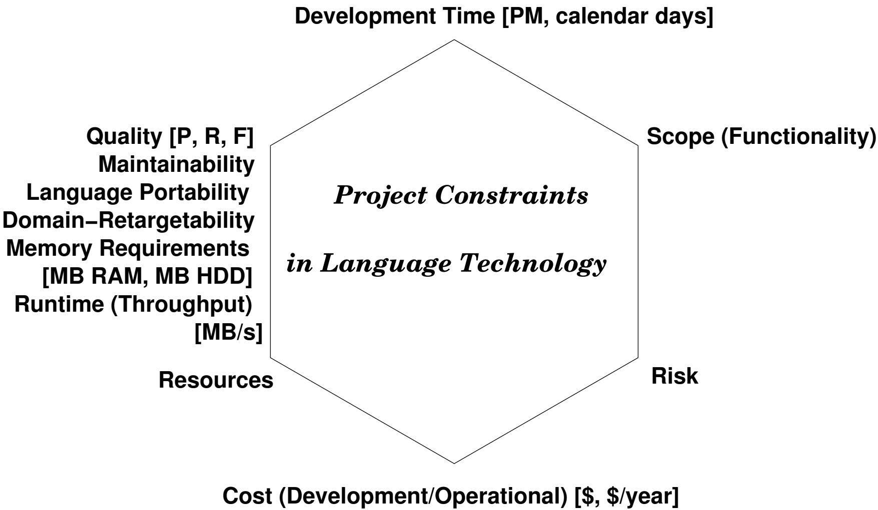


Estimating the Cost of Systems (Leidner, in prep.)

$$C_{Total} = C_{PM} + C_{Res} + C_{Dev} + C_{Comp} + C_{Data} + C_{KM}$$

- C_{PM}: the cost of project management, i.e. the cost of planning, initiating, executing/controlling and closing the project
- C_{Res}: the cost of research activities required to develop the system (prior art, evaluative comparison of existing systems, determining features, regular ongoing quantitative evaluation)
- C_{Dev}: the cost of developing and qualitative testing of the software and rules (e.g. "lingware") that constitute the system
- C_{Comp}: the cost of licensing in existing components to develop the system
- C_{Data}: the cost of licensing in existing data plus the cost of curating new data and/or meta-data (annotation layers, tags)
- C_{KM}: the cost of knowledge management (internal and externally facing: authoring customer documentation, authoring internal maintenance documentation, API documentation, training materials)

Data Science Project Constraints



The Five Project Phases (PMI, 2013)

We can distinguish between 5 clearly separate phases of every project:

- **Initiating**
- **Planning**
- **Executing**
- **Monitoring and Controlling**
- **Closing**

Project Management Areas (PMI, 2013)

We can distinguish between 10 project management sub-areas:

- Project **Integration** Management
- Project **Scope** Management
- Project **Time** Management
- Project **Cost** Management
- Project **Quality** Management
- Project **Human Resources** Management
- Project **Communication** Management
- Project **Risk** Management
- Project **Procurement** Management
- Project **Stakeholder** Management

Plan to Succeed: Contents of a (Good) Project Plan

In general, “what will be done and how?”

- Objectives, Motivation, background, terminology
- Work packages (Work Breakdown Structure, WBS) and schedule (GANTT)
- Life cycle for processes and the project
- Answers to these questions:
  - How will objectives be achieved?
  - How will change be monitored/controlled?
  - How will configuration management be performed?
  - How will integrity of performance measurement baseline be maintained?
  - How will open issues be addressed?
- Tailoring of results

Work Breakdown Structure and Dependency Analysis

- **Work Breakdown Structure**: hierarchical task decomposition
  1. Specify scope
  2. Obtain and process data
  2.1 obtain data
  2.2 pre-process data
  3. implement system
  4. process data
  5. test system
  6. analyse results
- **Dependency Analysis**: determine temporal sequencing Any two WPs can depend on each other or not (identify WPs that need to be completed before other WPs can be started)
- Example: "2.2 data pre-processing" depends on: "2.1 obtain data"

Risk Response Strategies

- **Avoid**: change project plan to eliminate risk entirely
- **Transfer**: shift responsibility to a third party (externalize, insure)
- **Mitigate**: reduce probability or impact
- **Exploit**: make an opportunity happen (for opportunity = positive risk)
- **Share**: allocate ownership to third party
- **Enhance**: modify size of probability/impact of opportunity (for positive risk)
- **Accept**: Accept the risk may happen and create contingency reserves or response plan ("what if")
- **Contingent Response**: Plan to execute under certain circumstances

Scientific Evaluation vs. Business Evaluation

- Scientific evaluation: e.g. $P = R = F1 = .88, \kappa = .8$
- Scientific evaluation compares the actual performance of the system to its potential maximum performance.
- It is fair, because it takes into account what is actually possible in the best case, based on the coverage of the data and the quality of gold data.
- Scientific evaluation is successful if the state of the art is statistically significantly outperformed by the proposed method or developed system.
- In reality, more factors are considered than a system's output quality (e.g. in terms of F-score): often, a faster/cheaper-to-build system with slightly lower quality is the prefered option.
- Similarly, a system that is easier to maintain/extend but has a 2% lower F-score can be a better choice over a method that is statistically superior but lacks these desirable properties.

Typical Challenges: Systems in the So-Called "Real World"

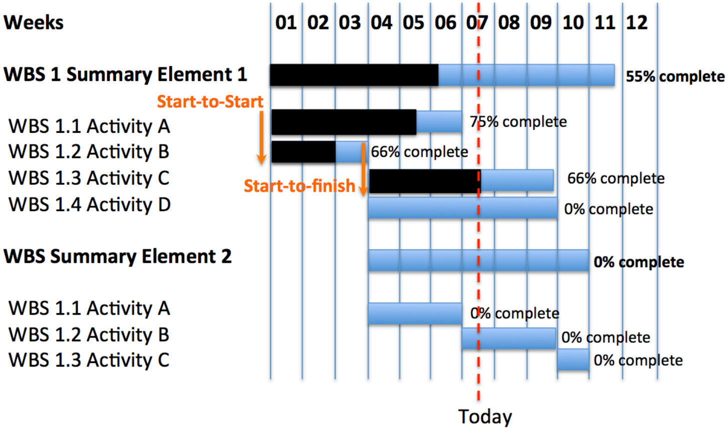
- Often-encountered issues:
  - Customers are typically unable to answer questions about the needed quality levels;
  - system needed "as soon as possible" (no dependency analysis);
  - requires "99% accuracy" (without investigation about the impact of errors on the business process);
  - budgets and time lines are often set using arbitrary guesswork;
  - particular vendors are often chosen without a systematic quantitative evaluation of their solutions' accuracy or accuracy/price ratio.
- Recommended behavior:
  - need to educate management and customers (planning, need/value of evaluation);
  - need to push back on unreasonable time lines (cf. Yourdon's excellent book *Death March*, 2003);
  - give estimates with baked-in contingency buffers (size proportional to the similarity of a project to other projects successfully delivered in the past).

Time Estimation: Program Evaluation and Review Technique (PERT)

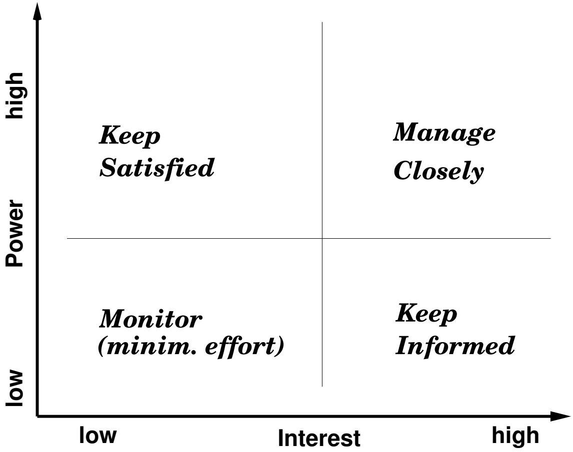
- Problem: how long will teach task take to complete?
- One solution: PERT, also known as: **Three-Point Estimate**
- Common technique to estimate the time for a piece of work
- Weighted Average:

$$t_{Est} = \frac{t_{optimist.} + 4 \cdot t_{mostlikely} + t_{pessimist.}}{6}$$

- Nota bene: *un*related to PERT Chart


Weeks
01 02 03 04 05 06 07 08 09 10 11 12
WBS 1 Summary Element 1
55% complete
Start-to-Start
WBS 1.1 Activity A
75% complete
WBS 1.2 Activity B
66% complete
WBS 1.3 Activity C
Start-to-finish
66% complete
WBS 1.4 Activity D
0% complete
WBS Summary Element 2
0% complete
WBS 1.1 Activity A
0% complete
WBS 1.2 Activity B
0% complete
WBS 1.3 Activity C
0% complete
Today

Stakeholder Classification



Responsibility Assignment Matrix (RAM) (also: RACI Table)

Types of Responsibilities for a Team Member per Work Package:

- **R** – Responsible,
- **A** – Accountable,
- **C** – Consulted or
- **I** – Informed

WP	Comp.Ling.	Ling.	Devel.	PM
1.2.4.1	A	R	C	I
1.2.4.2	I	A	I	I
1.2.4.3	I	C	A	I
1.3.1	I	C	A	I

Dealing with Management

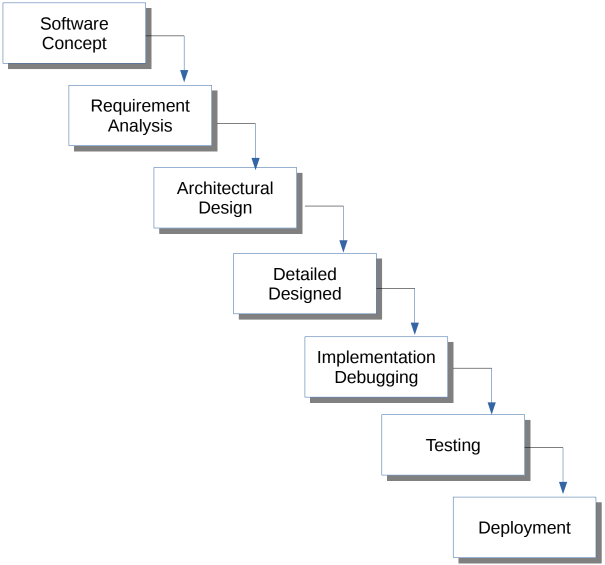
- Getting buy-in/commitment
- Setting and managing expectations
- Regular, proactive reporting
- Communicating results (milestones, roadblocks, success story)

D2V Data Science Methodology

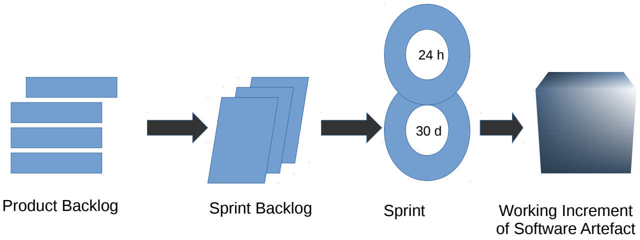
Existing Methodologies

- Project Management Methodologies
  - PMI
  - Prince2
- Software Development Methodologies
  - Waterfall Model
  - Agile Model
- *Data Mining* Methodologies
  - CRISP-DM
  - KDD
  - SEMMA
- Data Science Methodologies
  - **D2V**

The Waterfall Model of System Development



Agile Development Model with Sprints



The CRISP DM Process

- The **CRISP-DM** methodology (8; 48; 3)
- "CRoss Industry Standard Process for Data Mining"
- developed by: DamilerChrysler, SPSS, NCR and OHRA
- 6 phases:
  - Business Understanding
  - Data Understanding
  - Data Preparation
  - Modeling
  - Evaluation
  - Deployment

The CRISP DM Process (from Chapman et al., 2000)

Business Understanding	Data Understanding	Data Preparation	Modeling	Evaluation	Deployment
Determine Business Objectives *Background* *Business Objectives* *Business Success Criteria*	**Collect Initial Data** *Initial Data Collection Report*	**Select Data** *Rationale for Inclusion/ Exclusion*	**Select Modeling Techniques** *Modeling Technique* *Modeling Assumptions*	**Evaluate Results** *Assessment of Data Mining Results w.r.t. Business Success Criteria* *Approved Models*	**Plan Deployment** *Deployment Plan*
Assess Situation *Inventory of Resources* *Requirements, Assumptions, and Constraints* *Risks and Contingencies* *Terminology* *Costs and Benefits*	**Describe Data** *Data Description Report*	**Clean Data** *Data Cleaning Report*	**Generate Test Design** *Test Design*	**Review Process** *Review of Process*	**Plan Monitoring and Maintenance** *Monitoring and Maintenance Plan*
Determine Data Mining Goals *Data Mining Goals* *Data Mining Success Criteria*	**Explore Data** *Data Exploration Report*	**Construct Data** *Derived Attributes* *Generated Records*	**Build Model** *Parameter Settings* *Models* *Model Descriptions*	**Determine Next Steps** *List of Possible Actions* *Decision*	**Produce Final Report** *Final Report* *Final Presentation*
Produce Project Plan *Project Plan* *Initial Assessment of Tools and Techniques*	**Verify Data Quality** *Data Quality Report*	**Integrate Data** *Merged Data*	**Assess Model** *Model Assessment* *Revised Parameter Settings*		**Review Project** *Experience Documentation*
		Format Data *Reformatted Data*			
		Dataset *Dataset Description*			

The KDD Process (Fayyad et al., 1996)

- **KDD Process** (17; 18; 14): emerged from KDD (Knowledge Discovery in Databases) research community
- Sequence of 5-9 steps:
  - Selection
  - Pre-Processing
  - Transformation
  - Data Mining
  - Interpretation/Evaluation

SEMMA (SAS)

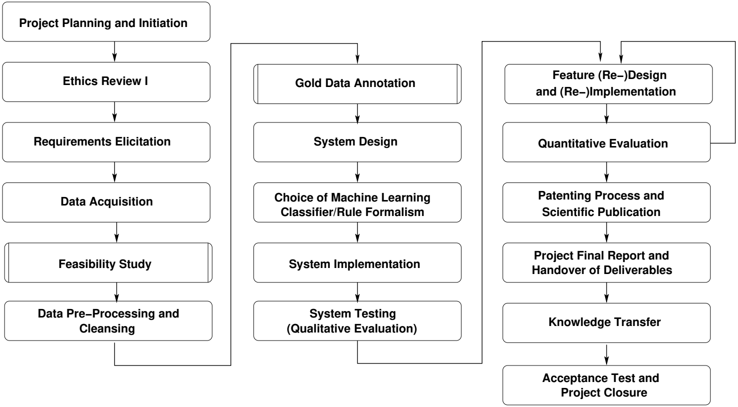
- **SEMMA** methodology (47): originally developed by SAS Institute Inc.
- Acronym: "Sample, Explore, Modify, Model, and Assess."
- Sample: extract a portion of a large data set big enough to contain the significant information, yet small enough to manipulate quickly;
- Explore: exploration of the data by searching for unanticipated trends and anomalies in order to gain understanding and ideas;
- Modify: creating, selecting, and transforming the variables to focus the model selection process
- Model: modeling the data by allowing the software to search
- Assess: assessing the data by evaluating

Relative Popularity of Methodologies

In a survey twice conducted by the KDNuggets Web site (kdnuggets.com):

Methodology	Used	by	Respondents
CRISP-DM	42%	-	43%
SEMMA	8%	-	13%
KDD	7%	-	8%

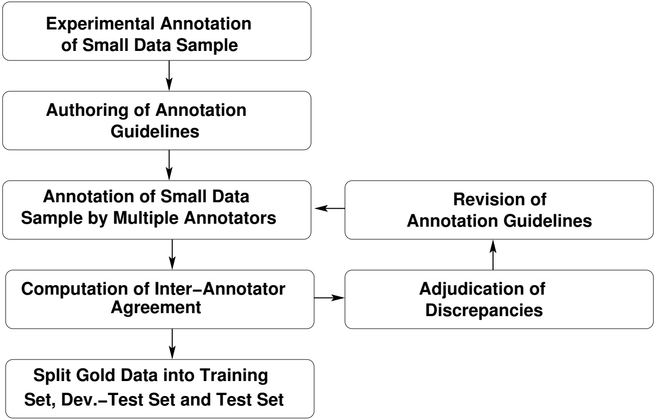
Data to Value (D2V) (Leidner, in prep.) (1/4): Phases



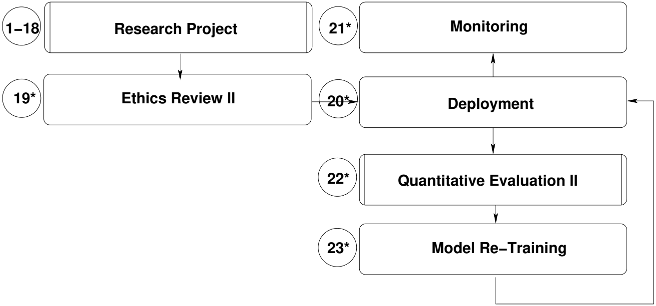
Data to Value (D2V) (Leidner, in prep.) (2/4): Phases

- One of the most common project management mistake: not to identify success metric
- We should define our success criteria before we even start
- Ask the question (to stakeholders to who the work is done):
  - Q: What are your current paint points?
  - Q: How would success look like for you?
  - Q: How would failure look like?
  - Q: Why are success/failure seen as they are (context, impact)?
  - Q: If we wanted to quantify success in a numeric metric, how would we do that?
- Often, the technical lead or project manager need to choose a suitable metric
- Defining evaluation metric and installing automatic code scaffolding for repeatable (daily, weekly) measurement aligns the team
- May have to be more than one metric (P, R, bias, confusion matrix, learning curve gradient)

Data to Value (D2V) (3/4): Gold Data Annotation



Data to Value (D2V) (4/4): Deployment Phases



D2V Phases Explained (1/4)

Project Planning and Initiation 1 phase
Ethics Review I 2 phase
Requirements Elicitation 3 phase
Data Acquisition 4 phase
Feasibility Study 5 phase
Evaluation Design 6 phase
Data Pre-Processing and Cleansing 7 phase

D2V Phases Explained (2/4)

Experimental Annotation of a Small Data Sample 8.1 phase

Authoring of Annotation Guidelines 8.2 phase

Computation of Inter-Annotator Agreement 8.4 phase

Gold Data Annotation 8 phase

Adjudication of Discrepancies 8.5 phase

Revision of Annotation Guidelines 8.6 phase

Split Gold Data into Training Set, Dev-Test Set and Test Set 8.7 phase

D2V Phases Explained (3/4)

System Architecture Design 9 phase

Choice of Machine Learning Classifier/Rule Formalism 10 phase

System Implementation 11 phase

System Testing 12 phase

Feature Design and Implementation 13 phase

Quantitative Evaluation I 14 phase

D2V Phases Explained (4/4)

Patenting and Publishing 15 phase

Final Report Authoring 16 phase

Knowledge Transfer 17 phase

Acceptance and Closure 18 phase

Deployment 20$*$ phase

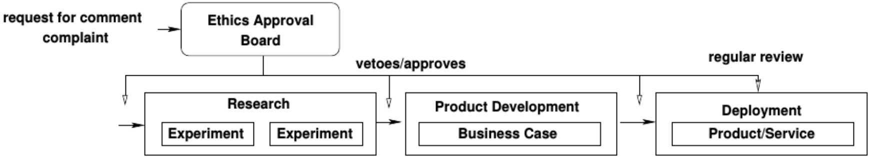
Ethics Review II 19$*$ phase

Monitoring 21$*$ phase

Quantitative Evaluation II 22$*$ phase

Model Re-Training 23$*$ phase

Ethics & Data Science – Ethics Reviews and ERBs (Leidner & Plachouras, 2017)



Legal and Ethical Questions for Data Science Practitioners

- **Privacy** Is an individual's right to self-determination of their data violated? Does a project work with PII information? Does the work respect the privacy rights of all individuals involved?
- **Abstraction** It is helpful to characterize a human with data, however a reduction of a human being to a mere set of data points is unethical (human dignity is also enshrined in some constitutions).
- **Algorithmic Bias** Is the big data method fair and unbiased to the whole population, intentionally or accidentally?
- **Copyright** Are copyright and the moral right to be recognized as author respected?
- **Competence** Does the experimenter have the statistical knowledge to conduct a big data experiment in a methodologically sound way?
- **Transparency** Can the method be inspected (in a code audit) to guarantee that what is said about what is done is actually what is done by the code? Can the user inspect what information the system holds about him or her, and correct errors in the data?

Ethics & Data Science – Responding to Ethical Issues (Leidner & Plachouras, 2017)

Demonstration	to effect a change in society by public activism
Disclosure	to document/to reveal injustice to regulators, the police, investigative journalists ("Look what they do!", "Stop what they do!")
Resignation	to distance oneself III ("I should not/cannot be part of this.')
Persuasion	to influence in order to halt non-ethical activity ("Our organization should not do this.")
Rejection	to distance oneself II; to deny participation; conscientious objection ("I can't do this.")
Escalation	raise with senior management/ethics boards ("You may not know what is going on here.")
Voicing dissent	to distance oneself I ("This project is wrong.")
Documentation	ensure all the facts, plans and potential and actual issues are preserved.

Handover and Closing the Project

- Knowledge transfer: author and share written documentation, but still recommended to hold Q&A session;
- Physical handover: all deliverables have been transfered to the customer (and confirmation of receipt has been obtained);
- Formal **sign-off**: receive formal approval (email, signed closure document) that the deliverables have been received, have been found acceptable and the knowledge transfer has been completed.

D2V Knowledge Areas

Knowledge Area	Relev. Literature	Relev. to Phases	# Q.
Value Analysis & Business Considerations	(11; 43; 39; 19)	1, 3, 18	26
Project Management	(28; 45; 51; 11; 54)	1, 16-18	8
Ethics	(24; 35; 33)	1-3, 19	6
Evaluation	(26; 30)	3, 6, 8, 14, 21-22	19
Data Management & Information Architecture	(15; 1; 44)	7-8	18
System Architecture	(50; 12; 9; 34)	1, 8-9, 15-16	1
Implementation & Testing	(47; 12; 50; 10; 5)	11-12, 15-16	1
Linguistic Resources (incl. I18N & L10N)	(27; 20)	6, 8-9, 11, 13-17	5
Scale Management (Processing & Storage)	(13; 55; 36)	3, 9, 20-21, 23	4
Legal, Privacy & Intellectual Property	(7; 42; 52; 25)	1, 3, 15	8
Deployment & Operations	(29; 46)	19-23	1

Some Questions from the D2V Guidance Question Catalog

No.	Sample Question	Area
Q9	How correct, truthful, reliable and complete is the data in the data set?	Veracity
Q10	How quickly does the data grow (in byte/s)?	Velocity
Q37	How structured/formalized is the data?	Data Management
Q46	What are the hypotheses that could be tested using this data set?	Value
Q51	What workflow is this data part of (in my organization, at my customers' sites)?	Workflow
Q65	Is it morally right to build the planned application?	Ethics
Q67	What licensing entitlements apply to the data set under consideration?	Legal
Q72	Will the system to be built need to support multiple languages?	Linguistics

Comparison of Methodologies

Process Model	Phases	unstruct. data	rule-b. approaches	learning-b. approaches	guidance questions	'eva first
CRISP-DM	6	no	yes	(yes)	n/a	no
SEMMA	4-5	no	yes	(yes)	n/a	no
KDD	5-9	no	yes	(yes)	n/a	no
D2V	**30**	**yes**	**yes**	**yes**	**96**	**yes**

Discussion: D2V – Claims and Limitations

- Only methodology which is evaluation first (to de-risk projects)
- Only methodology which features ethics check-points
- Only methodology which guides on gold standard creation
- Designed to give the practitioner comprehensive guidance
- Detailed; not aimed to be easily memorable
- Not all elements may be needed for each project
- Experienced project managers can adjust process to project complexity
- Informed by industry practice, Used in teaching (U Zurich, U Essex, GU Frankfurt, U Sheffield)
- No long-standing community experience/quant. evaluation available to date

D2V Methodology Summary

- New process model for the systematic pursuit of big data projects
- In particular:
  - ethically informed
  - "evaluation-first"
  - specific provisions for working with textual data
  - specific provisions for supervised learning (gold data annotation)
  - specific provisions for big data
  - informed by a catalog of guiding questions
- Like previous process models: iterative approach (but: acknowledges reality of diminishing returns)
- Future work:
  - forecasting-oriented modeling: predict time, cost and quality
  - tool support
  - gathering experimental data (project management databases gathered by practitioners)

Summary

In this tutorial, we have covered:

- what a project is and how success is defined;
- the 5 phases and 10 knowledge areas of project management;
- how to plan projects using WBS, PERT and GANTT;
- evaluation first: the importance of measuring;
- *Data-to-Value*: a process model for data science projects and best practices

Contact

Get in touch

I'd be interested in feedback, don't hesitate to get in touch.

E-mail me:	Jochen L. Leidner ⟨leidner@acm.org⟩
Twitter:	@jochenleidner

That's It, Folks...

Thank you!

[1] Abiteboul, S., Manolescu, I., Rigaux, P., Rousset, M.C., Senellart, P.: Web Data Management. Cambridge University Press, Cambridge, England, UK (2012)

[2] et al., S.: Machine learning: The high-interest credit card of technical debt. In: SE4ML – Software Engineering for Machine Learning

[3] Anand, S.S., Grobelnik, M., Herrmann, F., Hornick, M., Lingenfelder, C., Rooney, N., Wettschereck, D.: Knowledge discovery standards. Artif. Intell. Rev. 27(1), 21–56 (2007)

[4] Azevedo, A., Santos, M.F.: KDD, SEMMA and CRISP-DM: a parallel overview. In: Proceedings of the IADIS European Conference on Data Mining 2008. pp. 182–185 (2008)

[5] Beesley, K.R., Karttunen, L.: Finite State Morphology. CSLI Studies in Computational Linguistics, Chicago University Press and CSLI Publications, Chicago, IL, USA (2003)

[6] Benoît, G.: Data mining. Annual Review of Information Science and Technology 36(1), 265–310 (2002)

[7] Bygrave, L.A.: Data Privacy Law: An International Perspective. Oxford University Press, Oxford, England, UK (2014)

[8] Chapman, P., ...: CRISP-DM 1.0 – step-by-step data mining guide. Tech. rep., The CRISP-DM Consortium (2000), accessed 2008-05-01, http://www.crisp-dm.org/CRISPWP-0800.pdf

[9] Coad, P., Yourdon, E.: Object-Oriented Analysis. Prentice Hall and Yourdon Press, Englewood Cliffs, New Jersey, USA, 2nd edn. (1991)

[10] Coad, P., Yourdon, E.: Object-Oriented Design. Prentice Hall and Yourdon Press, Englewood Cliffs, New Jersey, USA (1991)

[11] Cockburn, A.: Writing Effective Use Cases. Addison-Wesley, Boston, MA, USA (2001)

[12] Cunningham et al., H.: Developing Language Processing Components with GATE Version 8.5 (A User Guide). University of Sheffield, Sheffield (2017), online, http://gate.ac.uk/sale/tao/, accessed 2017-05-28

[13] Dean, J., Ghemawat, S.: MapReduce: Simplified data processing on large clusters. In: Proceedings of the Sixth Symposium on Operating System Design and Implementation (OSDI), December 6-8, 2004. pp. 137–150. San Francisco, California, USA (2004)

[14] Debuse, J.C.W., de la Iglesia, B., Howard, C.M., Rayward-Smith, V.J.: Building the KDD roadmap: A methodology for knowledge discovery. In: Roy, R. (ed.) Industrial Knowledge Management, pp. 179–196. Springer (2001)

[15] Doan, A., Halevy, A., Ives, Z.: Principles of Data Integration. Morgan Kaufmann, Waltham, MA, USA (2012)

[16] Fayyad, U.M.: Data mining and knowledge discovery: making sense out of data. IEEE Expert 11(5), 20–25 (1996)

[17] Fayyad, U.M., Piatetsky-Shapiro, G., Smyth, P.: From data mining to knowledge discovery: an overview. In: Fayyad, U.M. (ed.) Advances in Knowledge Discovery and Data Mining. MIT Press, Cambridge, MA, USA (1996)

[18] Fayyad, U., Piatetsky-Shapiro, G., Smyth, P.: The kdd process for extracting useful knowledge from volumes of data. Commun. ACM 39(11), 27–34 (1996)

[19] Ferrari, M.: ROI in text mining projects. In: Zanasi, A. (ed.) Text Mining and its Applications to Intelligence, CRM and Knowledge Management, chap. 7, pp. 155–183. WITPress, Southampton, England, UK (2007)

[20] Gibbon, D., Moore, R., Winski, R.: Handbook of Standards and Resources for Spoken Language Systems. de Gruyter (1997)

[21] Hachey, G.: Instant OpenNMS Starter. Packt (2013)

[22] Halevy, A., Norvig, P., Pereira, F.: The unreasonable effectiveness of data. IEEE Intelligent Systems 24(2), 8–12 (2009)

[23] Hirji, K.K.: A proposed process for performing data mining projects. In: Pendharkar, P.C. (ed.) Managing Data Mining Technologies in Organizations, pp. 88–105. IGI Global,

Hershey, PA, USA (2003),
http://dl.acm.org/citation.cfm?id=778595.778609

[24] Hovy, D., Spruit, S.L.: The social impact of natural language processing. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics. pp. 591–598. ACL 2016, Berlin, Germany (2016)

[25] Howell, C., Farrand, B.: Law Express: Intellectual Property Law. Pearson, 5th edn. (2016)

[26] James, P., Stubbs, A.: Natural Language Annotation for Machine Learning. O'Reilly, Sebastopol, CA, USA (2012)

[27] Jurafsky, D., Martin, J.H.: Speech and Language Procesing. Prentice Hall, Upper Saddle River, NJ, USA, 2nd edn. (2008)

[28] Kerzner, H.: Project Management: A Systems Approach to Planning, Scheduling, and Controlling. Wiley, Hoboken, NJ, USA, 10th edn. (2009)

[29] King, R., Churchill, E.F.: Designing with Data: Improving User Experience with Large Scale User Testing. O'Reilly, Sebastopol, CA, USA, 1st edn. (2016)

[30] Krippendorff, K.: Content Analysis: An Introduction to Its Methodology. Sage, Thousand Oaks, CA, USA, 3rd edn. (2013)

[31] Kurgan, L., Musilek, P.: A survey of knowledge discovery and data mining process models. Knowledge Engineering Review 21(1), 1–24 (2006)

[32] Kurgan, L.A., Musilek, P.: A survey of knowledge discovery and data mining process models. Knowl. Eng. Rev. 21(1), 1–24 (2006)

[33] Lane, J., Stodden, V., Bender, S., Nissenbaum, H. (eds.): Privacy, Big Data and the Public Good. Cambridge University Press, New York, NY, USA (2014)

[34] Leidner, J.L.: Current issues in software engineering for natural language processing. In: Proceedings of the Workshop on Software Engineering and Architecture of Language Technology Systems (SEALTS) held at HLT-NAACL. pp. 45–50. ACL, Edmonton, Alberta, Canada (2003)

[35] Leidner, J.L., Plachouras, V.: Ethical by design: Ethics best practices for natural language processing. In: Proceedings of the First Workshop on Ethics and Natural Language Processing, April 4, 2007. pp. 30–40. Valencia, Spain (2017)

[36] Leskovec, J., Rajaraman, A., Ullman, J.D.: Mining Massive Datasets. Cambridge University Press, New York, NY, USA, 2nd edn. (2014)

[37] Lin, J., Dyer, C.: Data-Intensive Text Processing with MapReduce. Morgan & Claypool (2010)

[38] Liu, B.: Web Data Mining: Exploring Hyperlinks, Contents and Usage Data. Springer, Heidelberg, Germany, 2nd edn. (2011)

[39] Marr, B.: Big Data: Using Smart Big Data Analytics and Metrics to Make Better Decisions and Improve Performance. Wiley, Chichester, England, UK (2015)

[40] Mayer-Schönberger, V., Cukier, K.: Big Data. John Murray, London, England, UK (2013)

[41] Mehta, V.: Icinga Network Monitoring. Packt (2013)

[42] Moreham, N., Warby, S.M. (eds.): Tugendhat and Christie: The Law of Privacy and The Media. Oxford University Press, Oxford, England, UK, 3rd edn. (2016)

[43] Ohlhorst, F.J.: Big Data Analytics. Wiley, Hoboken, NJ, USA (2013)

[44] Pearson, R.K.: Mining Imperfect Data: Dealing with Contamination and Incomplete Records. SIAM, Philadelpha, PA, USA (2005)

[45] PMI (ed.): A Guide to the Project Management Body of Knowledge Project (PMBOK Guide). The Project Management Institute Inc., New York, NY, USA, 5th edn. (2013)

[46] Ryder, T.: Nagios Core Administration Cookbook. Packt Publishing, 2nd edn. (2016)

[47] SAS Institute Inc.: Data Mining Using SAS Enterprise Miner: A Case Study Approach. SAS Institute Inc., Cary, NC, USA, 3rd edn. (2013)

[48] Shearer, C.: The CRISP-DM model: the new blueprint for data mining. Journal of Data Warehousing pp. 13–22 (2000)

[49] Sommerville, I.: Software Engineering. Addison-Wesley, Boston, MA, USA, 10th edn. (2015)

[50] The Apache UIMA Development Community: UIMA Tutorial and Developers' Guides. Apache (2010), http://uima.apache.org, online, accessed 2017-02-22

[51] Wiegers, K.: Software Requirements. Microsoft Press, Redmond, WA, USA, 3rd edn. (2013)

[52] Williams, A.: Drafting Agreements for the Digital Media Industry Paperback. Oxford University Press, Oxford, England, UK, 2nd edn. (2016)

[53] Witten, I.H., Frank, E.: Data Mining: Practical Machine Learning Tools and Techniques. Morgan Kaufmann, 2nd edn. (2005)

[54] Yourdon, E.: Death March. Prentice Hall, Englewood Cliffs, NJ, 2nd edn. (2003)

[55] Zaharia, M., Chowdhury, M., Das, T., Dave, A., Ma, J., McCauley, M., Franklin, M.J., Shenker, S., Stoica, I.: Resilient distributed datasets: A fault-tolerant abstraction for in-memory cluster computing. In: Proceedings of the 9th USENIX Conference on Networked Systems Design and Implementation. pp. 2–2. NSDI'12, Berkeley, CA, USA (2012)

[56] Zinkevich, M.: Rules of machine learning: Best practices for ml engineering. Tech. rep., Google Inc. (nD)