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
Copyright © 2018 by Jochen L. Leidner · All rights reserved.
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
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)

⇒ Need to try and fix that!
Fundamentals of Project Management
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
A project is a *time-limited* activity to deploy *defined resources* to *effect change* with a *defined scope* with the aim to *benefit*. 
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** = \( \frac{\text{gain from investment} - \text{cost of investment}}{\text{cost of investment}} \)
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

**Project Constraints in Language Technology**

- Development Time [PM, calendar days]
- Quality [P, R, F] Scope (Functionality)
- Cost (Development/Operational) [$, $/year]
- Maintainability
- Language Portability
- Domain-Retargetability
- Memory Requirements [MB RAM, MB HDD]
- Runtime (Throughput) [MB/s]
- Resources
- Risk
We can distinguish between 5 clearly separate phases of every project:

- Initiating
- Planning
- Executing
- Monitoring and Controlling
- Closing
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
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 preferred 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).
- 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:

\[ t_{\text{Est}} = \frac{t_{\text{optimist.}} + 4 \cdot t_{\text{mostlikely}} + t_{\text{pessimist.}}}{6} \]

- Nota bene: unrelated to PERT Chart
Stakeholder Classification

- **Keep Satisfied**
- **Manage Closely**
- **Monitor (minim. effort)**
- **Keep Informed**

- Low Power, Low Interest: Monitor (minim. effort)
- Low Power, High Interest: Keep Satisfied
- High Power, Low Interest: Keep Informed
- High Power, High Interest: Manage Closely
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

<table>
<thead>
<tr>
<th>WP</th>
<th>Comp.Ling.</th>
<th>Ling.</th>
<th>Devel.</th>
<th>PM</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.2.4.1</td>
<td>A</td>
<td>R</td>
<td>C</td>
<td>I</td>
</tr>
<tr>
<td>1.2.4.2</td>
<td>I</td>
<td>A</td>
<td>I</td>
<td>I</td>
</tr>
<tr>
<td>1.2.4.3</td>
<td>I</td>
<td>C</td>
<td>A</td>
<td>I</td>
</tr>
<tr>
<td>1.3.1</td>
<td>I</td>
<td>C</td>
<td>A</td>
<td>I</td>
</tr>
</tbody>
</table>
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

1. Software Concept
2. Requirement Analysis
3. Architectural Design
4. Detailed Designed
5. Implementation Debugging
6. Testing
7. Deployment

References: Jochen L. Leidner - Project Management for Data Science
Agile Development Model with Sprints

- Product Backlog
- Sprint Backlog
- Sprint
- Working Increment of Software Artefact

- 24 hours
- 30 days

Jochen L. Leidner
Project Management for Data Science
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**
- Determine Business Objectives
  - Background
  - Business Objectives
  - Business Success Criteria
- Assess Situation
  - Inventory of Resources
  - Requirements
  - Assumptions
  - Constraints
  - Risks and Contingencies
  - Terminology
  - Costs and Benefits
- Determine Data Mining Goals
  - Data Mining Goals
  - Data Mining Success Criteria
- Produce Project Plan
  - Project Plan
  - Initial Assessment of Tools and Techniques

**Data Understanding**
- Collect Initial Data
  - Initial Data Collection Report
- Describe Data
  - Data Description Report
- Explore Data
  - Data Exploration Report
- Verify Data Quality
  - Data Quality Report

**Data Preparation**
- Select Data
  - Rationale for Inclusion/Exclusion
- Clean Data
  - Data Cleaning Report
- Construct Data
  - Derived Attributes
  - Generated Records
- Integrate Data
  - Merged Data
- Format Data
  - Reformatted Data
  - Dataset
  - Dataset Description

**Modeling**
- Select Modeling Techniques
  - Modeling Technique
  - Modeling Assumptions
- Generate Test Design
  - Test Design
- Build Model
  - Parameter Settings
  - Models
  - Model Descriptions
- Assess Model
  - Model Assessment
  - Revised Parameter Settings

**Evaluation**
- Evaluate Results
  - Assessment of Data Mining Results w.r.t. Business Success Criteria
  - Approved Models
- Review Process
  - Review of Process
- Determine Next Steps
  - List of Possible Actions
  - Decision

**Deployment**
- Plan Deployment
  - Deployment Plan
- Plan Monitoring and Maintenance
  - Monitoring and Maintenance Plan
- Produce Final Report
  - Final Report
  - Final Presentation
- Review Project
  - Experience
  - Documentation
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
In a survey twice conducted by the KDNuggets Web site (kdnuggets.com):

<table>
<thead>
<tr>
<th>Methodology</th>
<th>Used by Respondents</th>
</tr>
</thead>
<tbody>
<tr>
<td>CRISP-DM</td>
<td>42% - 43%</td>
</tr>
<tr>
<td>SEMMA</td>
<td>8% - 13%</td>
</tr>
<tr>
<td>KDD</td>
<td>7% - 8%</td>
</tr>
</tbody>
</table>
Data to Value (D2V) (Leidner, in prep.) (1/4): Phases

Project Planning and Initiation
- Ethics Review I
- Requirements Elicitation
- Data Acquisition
- Feasibility Study
- Data Pre-Processing and Cleansing

Gold Data Annotation
- System Design
- Choice of Machine Learning Classifier/Rule Formalism
- System Implementation
- System Testing (Qualitative Evaluation)

Feature (Re-)Design and (Re-)Implementation
- Quantitative Evaluation
- Patenting Process and Scientific Publication
- Project Final Report and Handover of Deliverables
- Knowledge Transfer
- Acceptance Test and Project Closure

References

Jochen L. Leidner
Project Management for Data Science
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

- Experimental Annotation of Small Data Sample
- Authoring of Annotation Guidelines
- Annotation of Small Data Sample by Multiple Annotators
- Computation of Inter–Annotator Agreement
- Split Gold Data into Training Set, Dev.–Test Set and Test Set
- Revision of Annotation Guidelines
- Adjudication of Discrepancies
Data to Value (D2V) (4/4): Deployment Phases

1–18
Research Project

19*
Ethics Review II

20*
Deployment

21*
Monitoring

Quantitative Evaluation II

22*
Model Re-Training

23*
**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
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
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
Patenting and Publishing  phase
Final Report Authoring  phase
Knowledge Transfer  phase
Acceptance and Closure  phase
Deployment  phase
Ethics Review II  phase
Monitoring  phase
Quantitative Evaluation II  phase
Model Re-Training  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)

<table>
<thead>
<tr>
<th>Strategy</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Demonstration</td>
<td>to effect a change in society by public activism</td>
</tr>
<tr>
<td>Disclosure</td>
<td>to document/to reveal injustice to regulators, the police, investigative journalists (“Look what they do!”; “Stop what they do!”)</td>
</tr>
<tr>
<td>Resignation</td>
<td>to distance oneself III (“I should not/cannot be part of this.”)</td>
</tr>
<tr>
<td>Persuasion</td>
<td>to influence in order to halt non-ethical activity (“Our organization should not do this.”)</td>
</tr>
<tr>
<td>Rejection</td>
<td>to distance oneself II; to deny participation; conscientious objection (“I can’t do this.”)</td>
</tr>
<tr>
<td>Escalation</td>
<td>raise with senior management/ethics boards (“You may not know what is going on here.”)</td>
</tr>
<tr>
<td>Voicing dissent</td>
<td>to distance oneself I (“This project is wrong.”)</td>
</tr>
<tr>
<td>Documentation</td>
<td>ensure all the facts, plans and potential and actual issues are preserved.</td>
</tr>
</tbody>
</table>
Knowledge transfer: author and share written documentation, but still recommended to hold Q&A session;

Physical handover: all deliverables have been transferred 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

<table>
<thead>
<tr>
<th>Knowledge Area</th>
<th>Relev. Literature</th>
<th>Relev. to Phases</th>
<th># Q.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Value Analysis &amp; Business Considerations</td>
<td>(11; 43; 39; 19)</td>
<td>1, 3, 18</td>
<td>26</td>
</tr>
<tr>
<td>Project Management</td>
<td>(28; 45; 51; 11; 54)</td>
<td>1, 16-18</td>
<td>8</td>
</tr>
<tr>
<td>Ethics</td>
<td>(24; 35; 33)</td>
<td>1-3, 19</td>
<td>6</td>
</tr>
<tr>
<td>Evaluation</td>
<td>(26; 30)</td>
<td>3, 6, 8, 14, 21-22</td>
<td>19</td>
</tr>
<tr>
<td>Data Management &amp; Information Architecture</td>
<td>(15; 1; 44)</td>
<td>7-8</td>
<td>18</td>
</tr>
<tr>
<td>System Architecture</td>
<td>(50; 12; 9; 34)</td>
<td>1, 8-9, 15-16</td>
<td>1</td>
</tr>
<tr>
<td>Implementation &amp; Testing</td>
<td>(47; 12; 50; 10; 5)</td>
<td>11-12, 15-16</td>
<td>1</td>
</tr>
<tr>
<td>Linguistic Resources</td>
<td>(27; 20)</td>
<td>6, 8-9, 11, 13-17</td>
<td>5</td>
</tr>
<tr>
<td>(incl. I18N &amp; L10N)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Scale Management</td>
<td>(13; 55; 36)</td>
<td>3, 9, 20-21, 23</td>
<td>4</td>
</tr>
<tr>
<td>(Processing &amp; Storage)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Legal, Privacy &amp; Intellectual Property</td>
<td>(7; 42; 52; 25)</td>
<td>1, 3, 15</td>
<td>8</td>
</tr>
<tr>
<td>Deployment &amp; Operations</td>
<td>(29; 46)</td>
<td>19-23</td>
<td>1</td>
</tr>
</tbody>
</table>
### Some Questions from the D2V Guidance Question Catalog

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

<table>
<thead>
<tr>
<th>Process Model</th>
<th>Phases</th>
<th>unstruct. data</th>
<th>rule-b. approaches</th>
<th>learning-b. approaches</th>
<th>guidance questions</th>
<th>‘eva first’</th>
</tr>
</thead>
<tbody>
<tr>
<td>CRISP-DM</td>
<td>6</td>
<td>no</td>
<td>yes</td>
<td>(yes)</td>
<td>n/a</td>
<td>no</td>
</tr>
<tr>
<td>SEMMA</td>
<td>4-5</td>
<td>no</td>
<td>yes</td>
<td>(yes)</td>
<td>n/a</td>
<td>no</td>
</tr>
<tr>
<td>KDD</td>
<td>5-9</td>
<td>no</td>
<td>yes</td>
<td>(yes)</td>
<td>n/a</td>
<td>no</td>
</tr>
<tr>
<td>D2V</td>
<td>30</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>96</td>
<td>yes</td>
</tr>
</tbody>
</table>
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)
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
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
Thank you!


References


