LTAT.05.008: Software Analytics

Course Introduction

Dietmar Pfahl
email: dietmar.pfahl@ut.ee

Spring 2018
Software Analytics – Teachers

• Dietmar Pfahl
• Lectures 1-4 & 8
• Labs 1-3

• Ezequiel Scott
• Lectures 5-7 & 8
• Labs 4-6
About me

• Professor at UT (Software Engineering)
• Adjunct Prof. at University of Calgary, Canada (since 2005)
  • Senior Member of ACM & IEEE
  • Certified SCRUM Product Owner
• Group Leader & Department Head at Fraunhofer Inst. of Experimental SW Engineering (1996-2005)
• Siemens Corporate Research (1987-1995)
TODAY: ‘Software Analytics’ – Why and What?
‘Software Analytics’ – Why?

• To support decision making at all stages of the software development and operations process
‘Software Analytics’ – Why?

- Many decisions have to be made …

- Engineering
  - Product
  - Process
  - People

Source: SEI at CMU, Donald Firesmith
‘Software Analytics’ – How?
‘Software Analytics’ – How?

• Motivation example: Test-Driven Development (TDD)
Test-Driven Development (TDD)

Possible Research Questions (RQs):
• ... ?
Test-Driven Development (TDD)

Possible Research Questions (RQs):
• Does TDD improve quality?
• Does TDD speed-up development?
• Does TDD save effort?
• …
Test-Driven Development (TDD)

There could be two possible settings time-wise
Imagine you asked the RQs in
• Year 2000
• Year 2018

What difference would that make?

Let’s discuss!
Test-Driven Development (TDD)

There could be two possible settings time-wise
Imagine you asked the RQs in
• Year 2000
  – No data/experience yet available
  – Do experiments/case studies to get evidence
• Year 2018
  – Lots of data from projects that used TDD
  – Lots of literature about TDD
Test-Driven Development (TDD)

Example Study 1 – Controlled Experiment
Setting of year 2000: no past data on TDD available

What is better?

• Iterative Test-Last (ITL) - traditional
• Test-Driven Development (TDD) - new

Controlled Experiment with developers of an Estonian company
Test-Driven Development (TDD)

Example Study 1 – Controlled Experiment

Credits for the content of the following slides go to:

Oscar Diente – UPM (Experiment Design, Training, Data Collection Instrumentation & Data Measurement)

Natalia Juristo – UPM, OYO (Research & Project Leader)

Ayse Tosun - OYO (Management & Assistance)

Burak Turham – OYO (Selection of Experimental Tasks)

Fernando Uyaguari – UPM (Data Collection Instrumentation & Data Measurement)

Sira Vegas - UPM (Data Analysis & Report Editor)
Test-Driven Development (TDD)

Example Study 1 – Controlled Experiment

Research Questions:

- **RQ1**: How does TDD affect the *amount of work done* compared to ITL?
- **RQ2**: How does TDD affect external *code quality* compared to ITL?
- **RQ3**: How does TDD affect *developer productivity* compared to ITL?
**Test-Driven Development (TDD)**

Example Study 1 – Controlled Experiment

**Measurement:**

- **Amount of work done:** Number of tackled user stories
  - A user story is tackled if at least one assert statement in the set of asserts associated with that user story passes

- **Code quality:** Conformance to requirements by means of acceptance tests
  - Calculated as the sum of the quality of each user story divided by the total number of user stories
  - Quality of user story = number of asserts passed by the user story divided by the number of asserts of the user story

- **Developer productivity:** Amount of work successfully delivered
  - Calculated as the number of asserts passed, divided by the total number of asserts
Test-Driven Development (TDD)

Example Study 1 – Controlled Experiment

Measurement:

• Amount of work done:

\[ TUS = \sum_{i=1}^{\#us} \#ASSERT_i(PASS) \geq 0 \rightarrow TRUE, \] where \#us is the total number of user stories composing the task (11 for Mars Rover and Music Phone and 13 for Bowling Score Keeper).

• Code quality:

\[ QLTY = \frac{\sum_{i=1}^{\#us} QLTY_i}{\#us}, \] where \( QLTY_i \) is the quality of the i-th tackled user story, and is defined as \( QLTY_i = \frac{\#ASSERT_i(PASS)}{\#ASSERT_i(ALL)} \). The values for \#ASSERT_i(PASS) represents the number of JUnit assertions passed in the acceptance test suite associated with the i-th user story.

• Developer productivity:

\[ PROD = \frac{\#ASSERT(PASS)}{\#ASSERT(ALL)} \]
Test-Driven Development (TDD)

Example Study 1 – Controlled Experiment

Experimental Design:
Two-factor three-level factorial within-subject design

Table 1. Temporal sequence of the experiment.

<table>
<thead>
<tr>
<th>DAY 1</th>
<th>DAY 2</th>
<th>DAY 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>T1</td>
<td>T2</td>
<td>T3</td>
</tr>
<tr>
<td>Training ITL</td>
<td>Experimental session ITL</td>
<td>Training TDD</td>
</tr>
<tr>
<td>T3</td>
<td>T4</td>
<td>T5</td>
</tr>
<tr>
<td>Experimental session TDD</td>
<td>Experimental session TDD-easy</td>
<td>Training ITL+TDD</td>
</tr>
<tr>
<td>T5</td>
<td>T6</td>
<td>T6</td>
</tr>
<tr>
<td>Experimental session TDD-complex</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 2. Experimental design.

<table>
<thead>
<tr>
<th>Day 1</th>
<th>Day 2</th>
<th>Day 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>ITL</td>
<td>TDD-easy</td>
<td>TDD-complex</td>
</tr>
<tr>
<td>Group 1</td>
<td>MarsRover</td>
<td>BSK</td>
</tr>
<tr>
<td>Group 2</td>
<td>BSK</td>
<td>MarsRover</td>
</tr>
</tbody>
</table>

Factors:

Levels:
Test-Driven Development (TDD)

Example Study 1 – Controlled Experiment

Experimental Design:
Two-factor three-level factorial within-subject design

<table>
<thead>
<tr>
<th>DAY 1</th>
<th>DAY 2</th>
<th>DAY 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>T1 Training ITL</td>
<td>Experimental session ITL</td>
<td>Training TDD</td>
</tr>
</tbody>
</table>

**Table 2. Experimental design.**

<table>
<thead>
<tr>
<th>Day 1</th>
<th>Day 2</th>
<th>Day 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>ITL</td>
<td>TDD-easy</td>
<td>TDD-complex</td>
</tr>
<tr>
<td>Group 1</td>
<td>MarsRover</td>
<td>BSK</td>
</tr>
<tr>
<td>Group 2</td>
<td>BSK</td>
<td>MarsRover</td>
</tr>
</tbody>
</table>

2 Factors:
- Dev. Approach, Task
3 Levels (each):
- ITL, TDD-easy, TDD-comp
- MR, BSK, MP
Test-Driven Development (TDD)

Result of RQ1: How does TDD affect the amount of work done compared to ITL?

The performance of developers when using TDD seems to be (slightly) lower than when using ITL.
Test-Driven Development (TDD)

**Result of RQ1:** How does TDD affect the amount of work done compared to ITL?

The performance of TDD-easy seems to depend on some (yet unknown) factor related to the system at hand.
Test-Driven Development (TDD)

Result of RQ2: How does TDD affect external code quality compared to ITL?

For easy tasks (systems) the performance of developers when using TDD seems not to be significantly different than when using ITL.
Test-Driven Development (TDD)

Result of RQ2: How does TDD affect external code quality compared to ITL?

The performance of TDD-easy seems to depend on some (yet unknown) factor related to the system at hand?
Test-Driven Development (TDD)

Result of RQ3: How does TDD affect developer productivity compared to ITL?

The performance of developers when using TDD seems to be lower than when using ITL.
Test-Driven Development (TDD)

Result of RQ3: How does TDD affect developer productivity compared to ITL?

This time the effect seems to be consistent across tasks.
Test-Driven Development (TDD)

Example Study 1 – Controlled Experiment

Summary:

- TDD does not seem to be a promising approach. In general, developers get better results on productivity, and small improvements on amount of work done and quality with ITL.
- These results are valid for practitioners with little or no previous experience in TDD, and with very little training.
- Results are dependent on experimental tasks used. However we did not try to benefit TDD, since TDD complex task was far more complex than ITL task. Since the tasks were not real, results need confirmation with tasks similar to those of developers’ daily work.
- Notice which is the control vs. which we compare: Incremental Test Last! It is not the pure test last approach where the whole system is first coded and then tested at the very end.
- Additionally, there seems to be other variables influencing the effect of TDD that need to be further investigated.
Test-Driven Development (TDD)

Example Study 2 – Data Mining

• Comparative analysis of GitHub repositories that adopt TDD in order to determine how TDD affects software development productivity and software quality.

• Classified GitHub repositories archived in 2015 in terms of how rigorously they practiced TDD, thus creating a TDD spectrum.

• Then matched and compared various subsets of these repositories on this TDD spectrum with control sets of equal size.

• The control sets were samples from all GitHub repositories that matched certain characteristics, and that contained at least one test file.

Source:
Test-Driven Development (TDD)

Example Study 2 – Data Mining

• Then compared how the **TDD sets** differed from the **control sets** on the following characteristics:
  – number of test files,
  – average commit velocity,
  – number of bug-referencing commits,
  – number of issues recorded,
  – usage of continuous integration,
  – number of pull requests,
  – distribution of commits per author
Test-Driven Development (TDD)

Example Study 2 – Data Mining

Research Questions:

• **RQ1:** Does the adoption of TDD improve *commit velocity*?
• **RQ2:** Does the adoption of TDD reduce the *number of bug-fixing commits*?
• **RQ3:** Does the adoption of TDD affect the *number of issues reported*?
• **RQ4:** Is *continuous integration* more prevalent in TDD development?
• **RQ5:** Does the adoption of TDD affect *developer collaboration*?
Test-Driven Development (TDD)

How were TDD-projects identified/characterized?

- Must have tests with class coverage:
  - Coverage up to: 10%, 20%, ..., 100%

- Must have test code committed before production code with a certain amount of relaxation allowed:
  - Time windows: none, up to: 1 hour, 1 day, 1 week

- Other criteria:
  - minimum size, minimum number of developers, ...
Test-Driven Development (TDD)

Heatmap:

Gray shades indicate whether TDD is better, worse, or equal to comparable control sets.

In total: 40 time window x coverage combinations
Test-Driven Development (TDD)

Result of Q1: Does the adoption of TDD improve commit velocity?

ACV = Average Commit Velocity

Wilcoxon Tests
(non-parametric version of t-test)
Test-Driven Development (TDD)

• **Result of RQ2**: Does the adoption of TDD reduce the number of bug-fixing commits?

![Graph showing the relationship between coverage and time window with NBFC (Number of Bug-Fixing Commits)]

NBFC = Number of Bug-Fixing Commits

Wilcoxon Tests
Test-Driven Development (TDD)

- **Result of RQ3:** Does the adoption of TDD affect the number of issues reported?

\[
\text{TNI} = \text{Total Number of Issues}
\]

Wilcoxon Tests
Test-Driven Development (TDD)

- **Result of RQ4**: Is continuous integration more prevalent in TDD development?

---

TRAVIS = TravisCI (travis-ci.com)

Wilcoxon Tests
Test-Driven Development (TDD)

• **Result of RQ5**: Does the adoption of TDD affect *developer collaboration*?

\[ \text{TNPR} = \text{Total Number of Pull Requests} \]

Wilcoxon Tests
Test-Driven Development (TDD)

No statistically significant difference between TDD and non-TDD projects!
Test-Driven Development (TDD)

The Death of TDD?

david.putman@agil8.com
Context:
Evidence-Based SE

• Knowledge in SE: Anecdotal vs. Evidence-based
• Evidence in Science -> Data
• Data Sources in SE?
  – Surveys, Case Studies, Experiments, Project Repos, Reviews, Dedicated Collections:
    http://promise.site.uottawa.ca/SERepository/datasets-page.html

• Tip: Link to Lecture by Gregory Wilson:
  https://vimeo.com/9270320
The Wallace Model

Research Questions – Taxonomy

- Exploratory Question
  - Existence Question
  - Description and Classification Question
  - Descriptive Comparative Question
  - Frequency and Distribution Question

- Design Question
  - Base-Rate Question
  - Descriptive-Process Question
  - Relationship Question
  - Simple Causality Question

- Knowledge Question
  - Causality Question
  - Causality-Comparative Question
  - Causality-Comparative Interaction Question

- Research Question
Exploratory Questions

• **Existence questions** -> Does X exist?
  – Example: *Do issue reports actually exist?*

• **Description and classification questions** -> What is X like? / What are its properties? / How can it be categorized? / How can we measure it? / What is its purpose? / What are its components? / How do the components relate to each other?
  – Example: *What are all the types of issue reports?*

• **Descriptive comparative questions** -> How does X differ from Y?
  – Example: *How do issue report formats differ between open source and closed source development projects?*
Knowledge and Design Questions

• **Knowledge Questions**: focusing on the way the world is
  – Questions about the normal pattern of occurrence of a phenomenon (*Base-rate Questions*)
  – Questions about relationships between two different phenomena (*Relationship Questions*)
  – Questions about causality between two phenomena (*Causality Questions*)

• **Design Questions**: concerned with how to do things better
Knowledge Questions

• Base-rate:
  – **Frequency and Distribution Questions** --> How often does X occur? / What is an average amount of X?
    Example: *How many distinct issue reports per issue report type are created in large software development projects?*
  – **Descriptive-Process Questions** --> How does X normally work? / What is the process by which X happens? / In what sequence do the events of X occur?
    Example: *How do software developers use issue reports?*
Knowledge Questions (cont’d)

• Relationship:
  – Relationship Questions -> Are X and Y related? / Do occurrences of X correlate with occurrences of Y?

Example: Do project managers’ claims about how often their teams use test tool X correlate with the actual use of test tool X?
Knowledge Questions (cont’d)

• Causality:
  – Simple Causality Questions -> Does X cause Y? / Does X prevent Y? / What causes Y? / What are all the factors that cause Y? / What effect does X have on Y?
    Example: Does the use of GUI test tool X improve software quality?
  – Causality-Comparative Questions -> Does X cause more Y than does Z? / Is X better at preventing Y than Z?
    Example: Does the use of GUI test tool X improve software quality more than other GUI test tools?
  – Causality-Comparative Interaction Questions
Knowledge Questions (cont’d)

• Causality:
  – Causality-Comparative Interaction Questions -> Does X or Z cause more Y under one condition but not others?

Example: Does the use of GUI test tool X improve software quality more than GUI test tools in web application projects, but not in genuine mobile applications?
Design Questions

-> "What is an effective way to achieve X?" / What strategies help to achieve X?"

Examples:

What is an effective way for teams to test mobile applications in order to improve quality without increasing cost?

or

What is an effective way for teams to design mobile applications in order to improve energy efficiency?
The Wallace Model

Data Collection & Research Methods

• Survey
  – Questionnaire-based (primary study)
  – Literature-based (secondary / tertiary study)

• Case Study
  – Exploratory
  – Descriptive
  – Confirmatory/Explanatory
  – Improving

• Experiment
  – Controlled Experiment
  – Quasi-Experiment

• Many other …
  – Action Research
  – Ethnography
  – Longitudinal Studies
  – Design Science
Data Collection & Research Methods in Data Science

• Survey
  – Questionnaire-based (primary study)
  – Literature-based (secondary / tertiary study)

• Case Study
  – Exploratory
  – Descriptive
  – Confirmatory/Explanatory
  – Improving

• Experiment
  – Controlled Experiment
  – Quasi-Experiment

• Many other...
  – Action Research
  – Ethnography
  – Longitudinal Studies
  – Design Science
Example of Questionnaire-based Survey Research
Survey – Characterisation

• A survey is a data collection method or tool used to gather information about individuals in order to identify the characteristics of a broad population.
• The defining characteristic is the selection of a representative sample from a well-defined population with the aim to generalise from the sample to the population.
• Usually conducted with questionnaires, but can also involve structured interviews or data logging techniques.
• Example:
  – Investigate to what extent, how, by which companies, and by whom within the companies, TDD is used.
Survey – Characterisation (cont'd)

When to use it?

- Either at start of research to get an understanding of the current situation …
- or at the end of a research phase to see the impact/acceptance/etc. of a new method/technique/tool

Issues:

- 'Superficial' --> no explanation / no causality --> not suitable for hypothesis testing
- 'Generalisability' of results depends on the choice of population and 'response rate', as well as validity and reliability of the data collection instrument
Survey – Example

What?
Research Questions:
- How is Agile practiced at Microsoft?
  - i.e. What do engineers do?
- How do engineers feel about it?
  - i.e. Do they like it?

Who, Where, and When?
Microsoft (worldwide, 2006)
Anonymous survey sent to 2821 engineers
  - 10% random sampling of all developers, testers, program managers at Microsoft in October 2006
  - 487 valid responses
    - 44% developers, 28% testers, 17% program managers

Why?
Many agile approaches exist – what's in it for Microsoft?

Survey – Example (cont'd)

Agile practice penetration at Microsoft

- Team coding standards
- Continuous integration of code
- System metaphor
- Simple design
- Sustainable pace
- User stories
- Small releases
- Direct interaction with customer
- Design improvement
- Collective code ownership
- Acceptance testing
- Whole team daily stand-up meeting
- Test-driven development
- Pair programming
Survey – Example (cont'd)

Quantitative Results (Highlights)

- 33% of respondents (spread across divisions) report their team uses Agile methodologies.
- They mainly use Scrum (68%).
- Used for many legacy products.
- Agile usage does not appear to depend on team co-location.
- Test-driven development and pair programming are not very common.

Qualitative Results (Highlights)

- MS engineers who have used Agile like it for their local team, but not necessarily for their organization.
- They worry about scale, overhead, and management buy-in.

Perceived benefits (687 comments, 44 themes)

<table>
<thead>
<tr>
<th>Benefit</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Improved Communication and Coordination</td>
<td>121</td>
</tr>
<tr>
<td>2. Quick Releases</td>
<td>101</td>
</tr>
<tr>
<td>3. Flexibility of Design – Quicker Response to Changes</td>
<td>86</td>
</tr>
<tr>
<td>4. More Reasonable Process</td>
<td>65</td>
</tr>
<tr>
<td>5. Increased Quality</td>
<td>62</td>
</tr>
</tbody>
</table>

Perceived problems (565 comments, 58 themes)

<table>
<thead>
<tr>
<th>Problem</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Does not scale to larger projects</td>
<td>52</td>
</tr>
<tr>
<td>2. Too many meetings</td>
<td>44</td>
</tr>
<tr>
<td>3. Management buy-in</td>
<td>37</td>
</tr>
<tr>
<td>4. Unfamiliar with Agile</td>
<td>36</td>
</tr>
<tr>
<td>5. Coordination with other teams</td>
<td>29</td>
</tr>
<tr>
<td>6. Losing sight of the big picture</td>
<td>29</td>
</tr>
</tbody>
</table>
Case Study Research

(examples given in Lecture 3)
Case Study – Variants

• Explorative
  – Finding out what is happening, seeking new insights and generating ideas and hypotheses for new research

• Descriptive
  – Portraying a situation or phenomenon

• Explanatory/Confirmatory
  – Start out with a given theory and try to refute it, ideally with a series of case studies covering various contexts

• Improving
  – Trying to improve a certain aspect of the studied phenomenon.

More on Case Study design (SE Group at Lund University):
http://serg.cs.lth.se/education/case_study_research/
Case Study – Characterisation

- Definition:
  - an empirical enquiry that investigates a contemporary phenomenon within its real-life context (in-vivo=in the living)
  - especially when the boundaries between phenomenon and context are not clearly evident.

- Examples:
  - Investigation on how a company takes advantage of ‘Open Innovation’
  - Investigation on how a company practices mobile app testing
  - Investigation on how and why a company practices TDD

- Characteristics:
  - When to use? --> When 'rich' information is requested
  - Often focus on qualitative data --> allows for better understanding of conditions under which a technique/tool works

- Issues:
  - Important: Proper case selection / clearly stated research question(s) / clearly defined framework for interpreting the observations
  - 'Generalisability' (1 case --> only 1 context)
Experimental Research
(examples given in Lecture 4)
Controlled Experiment – Characterisation

• An investigation of a testable hypothesis where one or more independent variables are manipulated to measure their effect on one or more dependent variables.

• In Software Engineering, typically, experiments require human subjects to perform some task.
Controlled Experiment – Simple Example

- Independent Variable: Tool used (Levels: X and Y)
- Dependent Variable: Design Quality
- Treatments: E = use the new Tool X / C = use the old Tool Y

Treatments (1 Factor / 2 Levels)

NB: Design can be within-subject or between-subject
Controlled Experiment vs. Quasi-Experiment

Randomization is a prerequisite for a controlled experiment!

- is random assignment used?
  - yes: randomized or true experiment
  - no: is there a control group or multiple measures?
    - yes: quasi-experiment
    - no: non-experiment
The Wallace Model

Research Methodology
– Overview of primary characteristics

<table>
<thead>
<tr>
<th>Methodology</th>
<th>Primary objective</th>
<th>Primary data</th>
<th>Design</th>
</tr>
</thead>
<tbody>
<tr>
<td>Survey</td>
<td>Descriptive</td>
<td>Quantitative</td>
<td>Fixed</td>
</tr>
<tr>
<td>Case study</td>
<td>Exploratory</td>
<td>Qualitative</td>
<td>Flexible</td>
</tr>
<tr>
<td>Experiment</td>
<td>Explanatory</td>
<td>Quantitative</td>
<td>Fixed</td>
</tr>
<tr>
<td>Action research</td>
<td>Improving</td>
<td>Qualitative</td>
<td>Flexible</td>
</tr>
</tbody>
</table>

Source:

LTAT.05.008 / Software Analytics / © Dietmar Pfahl 2018
Research Methodology
– Overview of primary characteristics

<table>
<thead>
<tr>
<th>Methodology</th>
<th>Primary objective</th>
<th>Primary data</th>
<th>Design</th>
</tr>
</thead>
<tbody>
<tr>
<td>Survey</td>
<td>Descriptive</td>
<td>Quantitative</td>
<td>Fixed</td>
</tr>
<tr>
<td>Case study</td>
<td>Exploratory</td>
<td>Qualitative</td>
<td>Flexible</td>
</tr>
<tr>
<td>Experiment</td>
<td>Explanatory</td>
<td>Quantitative</td>
<td>Fixed</td>
</tr>
<tr>
<td>Action research</td>
<td>Improving</td>
<td>Qualitative</td>
<td>Flexible</td>
</tr>
</tbody>
</table>

Source:
## Research Methodology – Overview of primary characteristics

<table>
<thead>
<tr>
<th>Methodology</th>
<th>Primary objective</th>
<th>Primary data</th>
<th>Design</th>
</tr>
</thead>
<tbody>
<tr>
<td>Survey</td>
<td>Descriptive</td>
<td>Quantitative</td>
<td>Fixed</td>
</tr>
<tr>
<td>Case study</td>
<td>Exploratory</td>
<td>Qualitative</td>
<td>Flexible</td>
</tr>
<tr>
<td>Experiment</td>
<td>Explanatory</td>
<td>Quantitative</td>
<td>Fixed</td>
</tr>
<tr>
<td>Action research</td>
<td>Improving</td>
<td>Qualitative</td>
<td>Flexible</td>
</tr>
</tbody>
</table>

Source:
## Research Methodology
- Overview of primary characteristics

<table>
<thead>
<tr>
<th>Methodology</th>
<th>Primary objective</th>
<th>Primary data</th>
<th>Design</th>
</tr>
</thead>
<tbody>
<tr>
<td>Survey</td>
<td>Descriptive</td>
<td>Quantitative</td>
<td>Fixed</td>
</tr>
<tr>
<td>Case study</td>
<td>Exploratory</td>
<td>Qualitative</td>
<td>Flexible</td>
</tr>
<tr>
<td>Experiment</td>
<td>Explanatory</td>
<td>Quantitative</td>
<td>Fixed</td>
</tr>
<tr>
<td>Action research</td>
<td>Improving</td>
<td>Qualitative</td>
<td>Flexible</td>
</tr>
</tbody>
</table>

Source:
Research Methodology
– Overview of primary characteristics

<table>
<thead>
<tr>
<th>Methodology</th>
<th>Primary objective</th>
<th>Primary data</th>
<th>Design</th>
</tr>
</thead>
<tbody>
<tr>
<td>Survey</td>
<td>Descriptive</td>
<td>Quantitative</td>
<td>Fixed</td>
</tr>
<tr>
<td>Case study</td>
<td>Exploratory</td>
<td>Qualitative</td>
<td>Flexible</td>
</tr>
<tr>
<td>Experiment</td>
<td>Explanatory</td>
<td>Quantitative</td>
<td>Fixed</td>
</tr>
<tr>
<td>Action research</td>
<td>Improving</td>
<td>Qualitative</td>
<td>Flexible</td>
</tr>
</tbody>
</table>

Source:
Data Mining of SW Repositories – Why and What?

• To support decision making at all stages of the software development process

• To complement other sources of evidence
  – Surveys, Case Studies, Experiments
What data can be extracted?

- Commits
- Pull requests
- Issues…
Administrative Info

- 8 Lectures
- 6 Labs

- 7 Quizzes (during lectures) -> 10 marks [best 5 count]
- 6 Homeworks (start during labs) -> 60 marks
- 1 Exam (written & open book) -> 30 marks
- Need 50% in each category to not fail

Course wiki:

https://courses.cs.ut.ee/2018/SWA/spring/Main/HomePage
Quiz 0: Data Mining Recap

(10 min -- 0 marks)
Quiz 0: Data Mining Recap – Q1

What order of steps in a data science project is correct?

a) Context understanding and goal setting -> Data collection -> Data understanding and validation -> Data preparation/processing -> Modeling -> Evaluation -> Deployment

b) Data collection -> Context understanding and goal setting -> Data understanding and validation -> Data preparation/processing -> Modeling -> Evaluation -> Deployment

c) Context understanding and goal setting -> Data collection -> Data preparation/processing -> Data understanding and validation -> Modeling -> Evaluation -> Deployment

d) Context understanding and goal setting -> Data collection -> Data understanding and validation -> Data preparation/processing -> Modeling -> Deployment -> Evaluation
Quiz 0: Data Mining Recap – Q1

What order of steps in a data science project is correct?

a) Context understanding and goal setting -> Data collection -> Data understanding and validation -> Data preparation/processing -> Modeling -> Evaluation -> Deployment

b) Data collection -> Context understanding and goal setting -> Data understanding and validation -> Data preparation/processing -> Modeling -> Evaluation -> Deployment

c) Context understanding and goal setting -> Data collection -> Data preparation/processing -> Data understanding and validation -> Modeling -> Evaluation -> Deployment

d) Context understanding and goal setting -> Data collection -> Data understanding and validation -> Data preparation/processing -> Modeling -> Deployment -> Evaluation
Quiz 0: Data Mining Recap – Q2

What data is quantitative?

- Color of car: blue, yellow, red, …
- Height of person: tall, average, short, …
- Temperature [in degrees Celsius]: 12, -8, 15
- Book code in a library: 500.12, 300.17, 312.13, …
Quiz 0: Data Mining Recap – Q2

What data is quantitative?

• Color of car: blue, yellow, red, …
• Height of person: tall, average, short, …
• **Temperature [in degrees Celsius]: 12, -8, 15 → Interval Scale**
• Book code in a library: 500.12, 300.17, 312.13, …
Quiz 0: Data Mining Recap – Q3

What are mean, median and mode of the following numbers?

• 10, 5, 0, 5, 5, 10, 10, 10, 0, 5

• Answer: Mean: ______ Median: ______ Mode: ______
Quiz 0: Data Mining Recap – Q3

What are mean, median and mode of the following numbers?

• 10, 5, 0, 5, 5, 10, 10, 0, 5

• Answer: Mean: ______ Median: ______ Mode: ______

• Mean: 60/10 = 6
• Median: 10, 10, 10, 10, 5, 5, 5, 5, 0, 0 -> 5 (point in the middle)
• Mode: ‘10’ and ‘5’ equally often -> modes: 5 and 10
Quiz 0: Data Mining Recap – Q4

Assume you have the data in the table below given and want to predict the value of ‘??’ using Machine Learning.

<table>
<thead>
<tr>
<th>P</th>
<th>Y</th>
<th>H</th>
<th>A</th>
<th>R</th>
<th>L</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>??</td>
</tr>
</tbody>
</table>

- Clustering
- Regression
- Classification
- Correlation
Quiz 0: Data Mining Recap – Q4

Assume you have the data in the table below given and want to predict the value of ‘??’ using Machine Learning.

- Clustering
- Regression
- Classification
- Correlation

[assuming ‘0’ & ‘1’ are nominal]

<table>
<thead>
<tr>
<th>P</th>
<th>Y</th>
<th>H</th>
<th>A</th>
<th>R</th>
<th>L</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
| 0 | 1 | 1 | 1 | 0 | ???
Assume you are trying to improve prediction accuracy by making your prediction model more complex. If you do so, you may face the problem of overfitting (as illustrated in the figure below).

What is a correct explanation of the increasing gap between training and test error on the right-hand side of the blue line ("just right")?

a) On test data the model fits both signal (meaningful information) as well as noise (randomness)
b) On test data the noise is the same as on training data but signal is different and cannot be predicted
c) On test data the signal is the same as on training data but noise is different and cannot be predicted

a) Model predicts noise well on training data but not on the test data, causing a gap in error rates
Quiz 0: Data Mining Recap – Q5

Assume you are trying to improve prediction accuracy by making your prediction model more complex. If you do so, you may face the problem of overfitting (as illustrated in the figure below).

What is a correct explanation of the increasing gap between training and test error on the right-hand side of the blue line ("just right")

a) On test data the model fits both signal (meaningful information) as well as noise (randomness)
b) On test data the noise is the same as on training data but signal is different and cannot be predicted
c) On test data the signal is the same as on training data but noise is different and cannot be predicted

a) Model predicts noise well on training data but not on the test data, causing a gap in error rates

What is a correct explanation of the increasing gap between training and test error on the right-hand side of the blue line ("just right")

a) On training test data the model fits both signal (meaningful information) as well as noise (randomness)
b) On test data the signal noise is the same as on training data but noise signal is different and cannot be predicted
c) On test data the signal is the same as on training data but noise is different and cannot be predicted

d) Model predicts signal noise well on training data but not on the test data, causing a gap in error rates
Data Mining Refresher (if needed)

• Look at Meelis Kull’s re-designed Data Mining course

• MTAT.03.183 Course Wiki – Fall 2017: https://courses.cs.ut.ee/2017/DM/fall
Prepare for Quiz 1

Read article: