

(Semi-) Automatic Categorization of Natural Language Requirements

Eric Knauss and Daniel Ott

eric.knauss@cse.gu.se
@oerich
oerich.wordpress.com



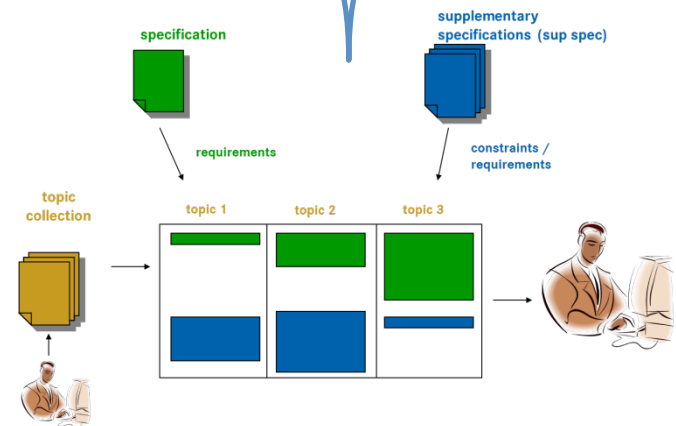
Importance of Requirement Categorization

Requirement	Security	Signal	Functional	Non-functional	Temperature	Test
R1	yes		yes		yes	
R2		yes		yes		
R3	yes	yes	yes			
R4	yes	yes		yes		yes
R5				yes	yes	yes

Recognising security issues [1]

Identification of special kinds [2] for:

- Architectural decisions
- Needed equipment
- Identification of dependencies for risk detection



Improving Review efficiency[3]

[1] E. Knauss, S. Houmb, K. Schneider, S. Islam, and J. Jürjens. Supporting requirements engineers in recognising security issues. Requirements Engineering: Foundation for Software Quality, pages 4-18, 2011.

[2] X. Song and B. Hwong. Categorizing requirements for a contract-based system integration project. In Requirements Engineering Conference (RE), pages 279-284. IEEE, 2012.

[3] Daniel Ott. Automatic requirement categorization of large natural language specifications at mercedes-benz for review improvements. In Requirements Engineering: Foundation for Software Quality, 2013.

Minimizing Efforts by Automation

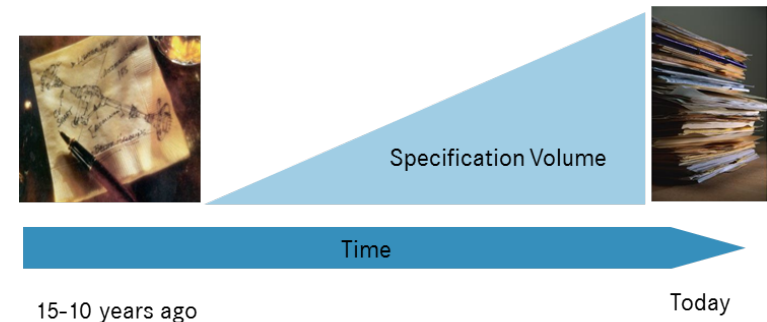
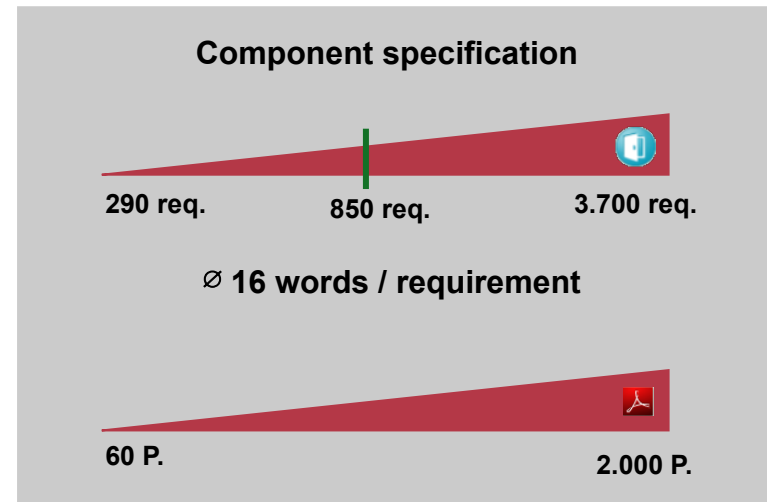
Today's specifications are large and complex.

Minimalize manual efforts for a categorization of requirements and at the same time maximize the quality of the resulting categorization

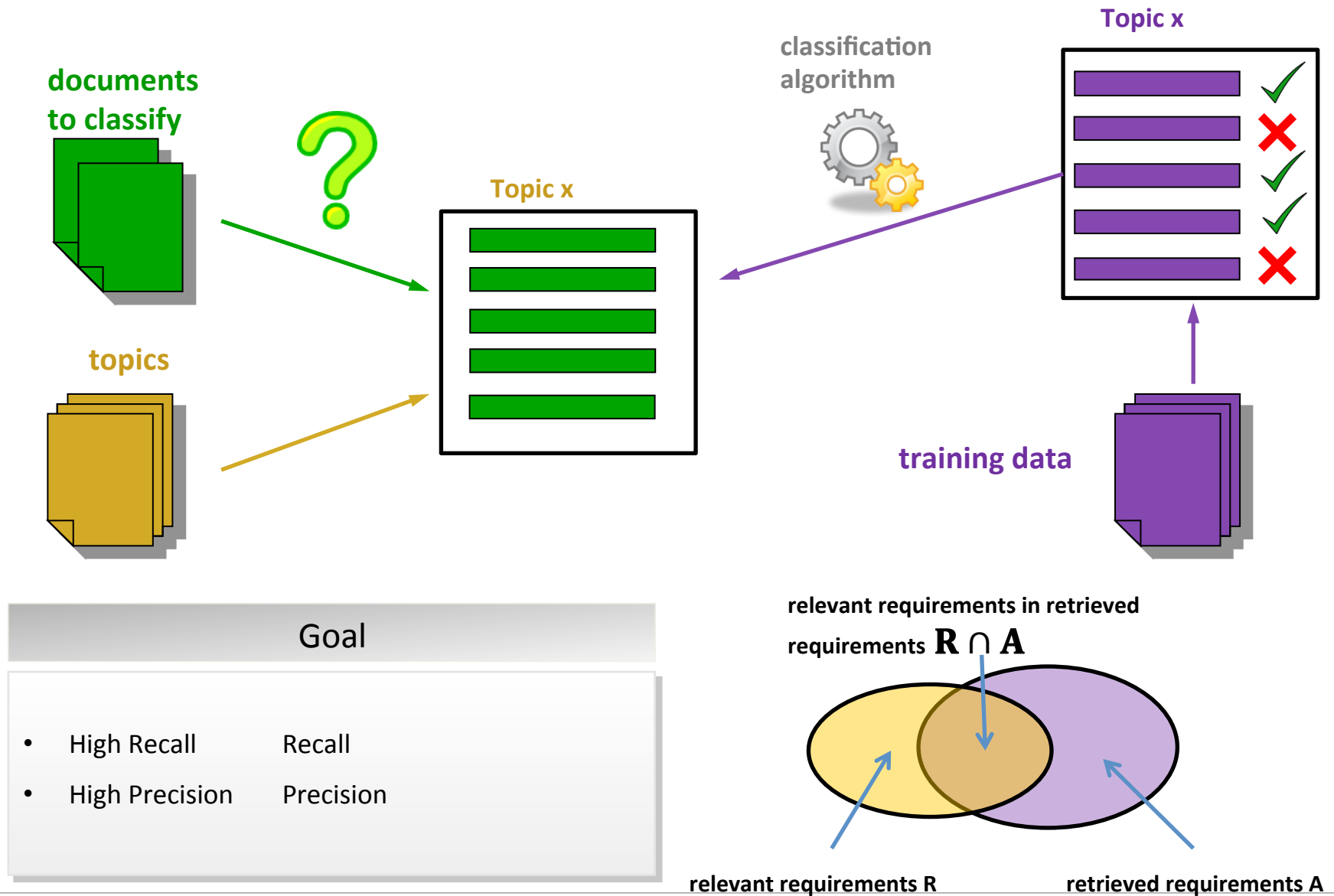
Research Goal

- Develop a **socio-technical system** for requirements categorization during the specification writing process
- Evaluate at Mercedes Benz with **full-automatic, semi-automatic and manual classification**

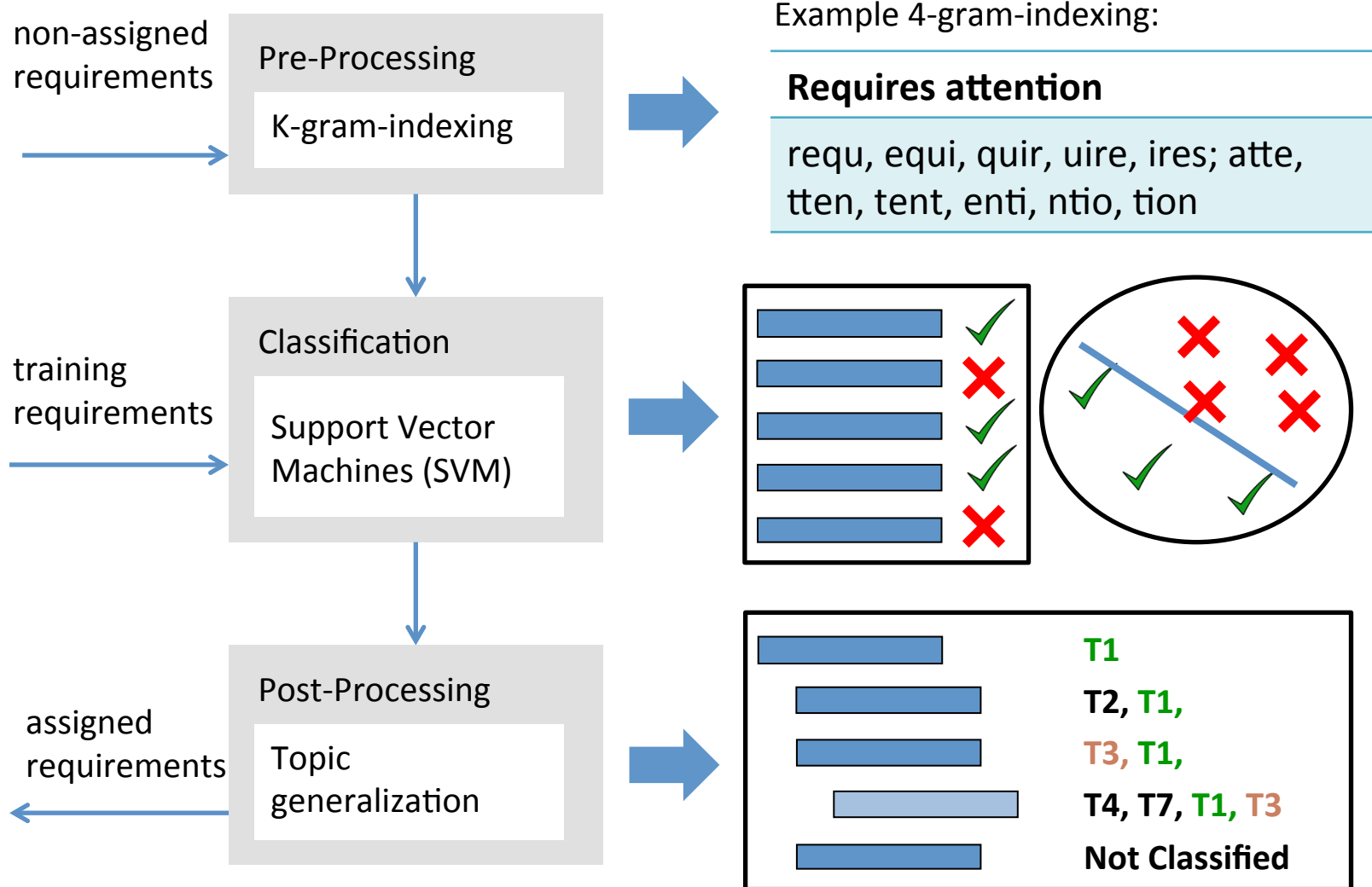
For example at Mercedes Benz



Automatic Classification - Basics



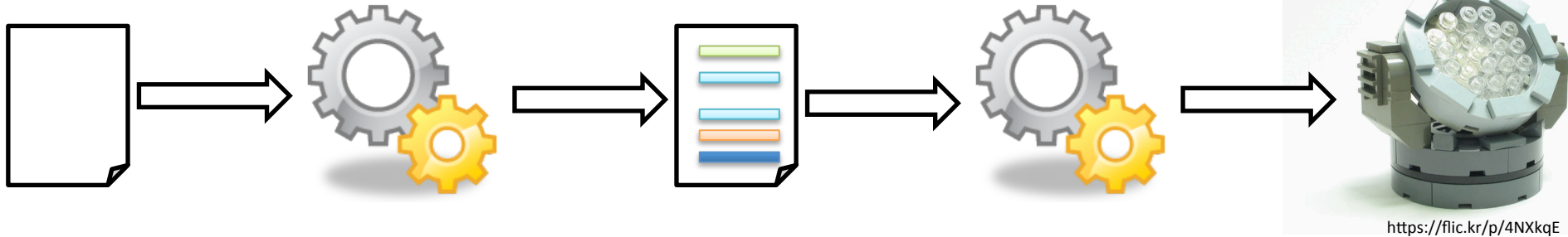
Machine Learning – Processing steps[1]



[1] D. Ott , „Automatic Requirement Categorization of Large Natural Language Specifications at Mercedes-Benz for Review Improvements”, REFSQ 2013

Idea

Step 1: Manually classify, train automatic classifier, gain a “spotlight”

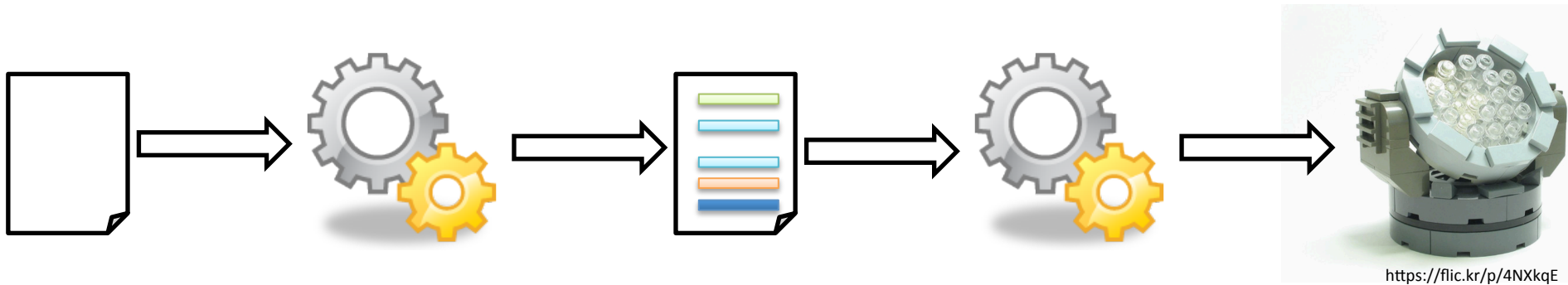


Step 2-n: Use “spotlight” to automatically categorize requirements ever after

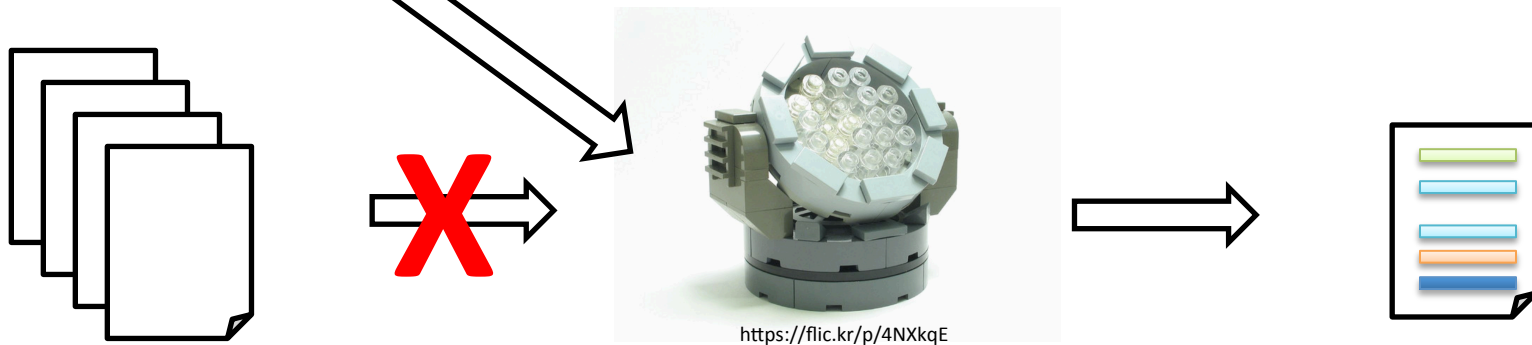


Reality

Step 1: Manually classify, train automatic classifier, gain a “spotlight”



Step 2: “Spotlight” only applicable to initial specification



Why?



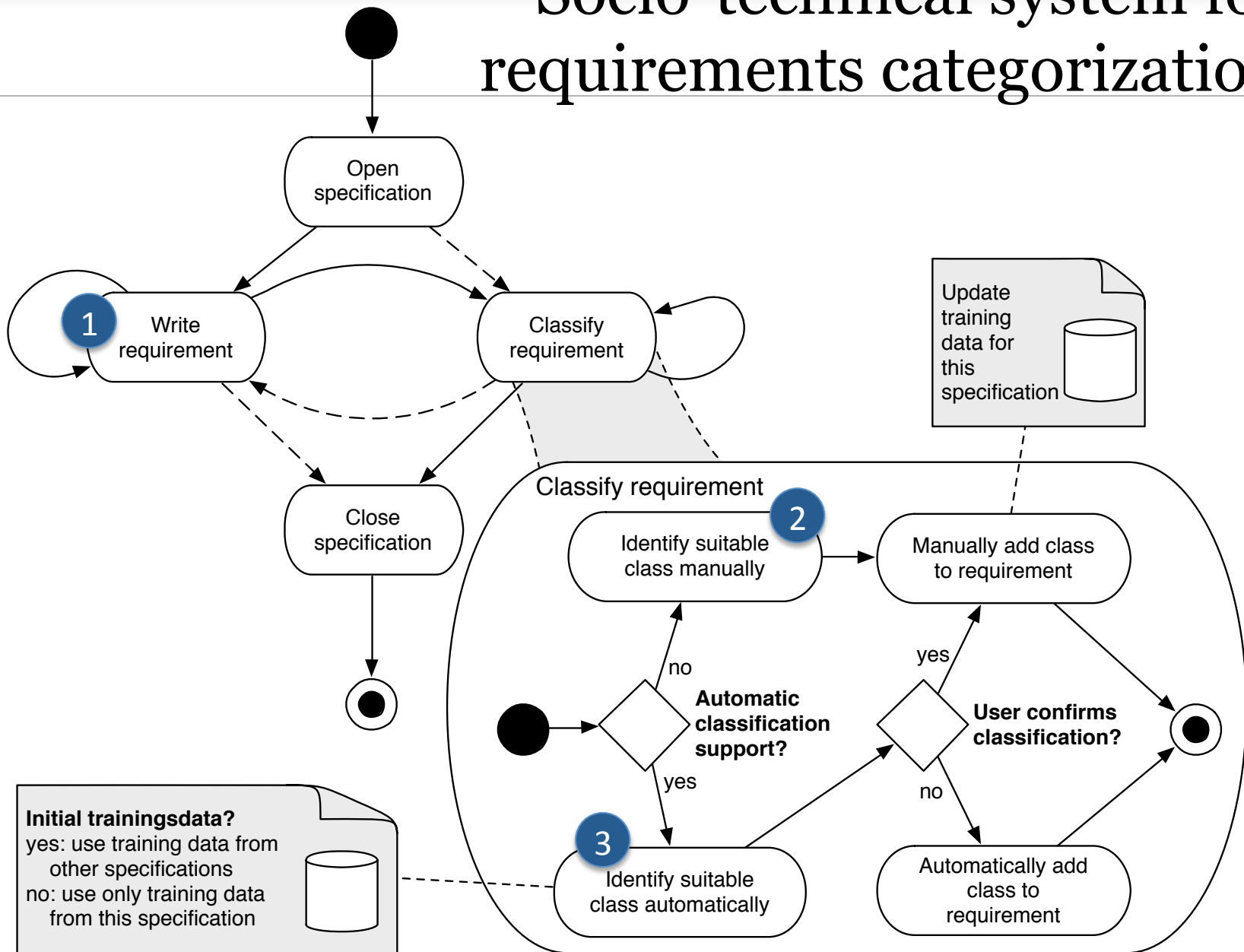
- Writing style
- Project culture
- Domain specific words
- ...
- Amount of training data

→ “Useful in product line and software evolution scenarios”

Shiny, but special purpose

- Our vision was a bit different...

Socio-technical system for requirements categorization



Infrastructure for Evaluation

Interactive Topic Classification

File Edit

- Spec: KLH_Türsteuergerät 1
 - 529: Vorwort
 - 531: Allgemeine Projektvorgab
 - 566: Projektmanagement und...
 - 576: Systemumgebung
 - 583: Technische Anforderungen
 - 1088: Die in diesem Laster
 - 3940: Benennung und Teile
 - 3941: Folgende...
 - 3942: Mercedes-Ber
 - 588: Funktionen
 - 1087: In diesem Abschr
 - 589: Funktionsbeschrei
 - 2718: Aktivieren und
 - 2722: Verhalten
 - 2750: Aktivie
 - 2751: Da
 - 2752
 - 2753
 - 2754
 - 2755
 - 2756: Ab
 - 2757: Deakti
 - 2758: Da
 - 2759
 - 2760
 - 2761: Da
 - 2762: Ist
 - 2723: Initialisier
 - 2771: Allgemeines V
 - 1091: Sitzeinstellun
 - 1089: Benutzerman
 - 1094: Türschloß
 - 1096: Fensterheber
 - 1098: Innenraumbel
 - 1100: Außenspiege
 - 599: Diagnose
 - 603: Steuergerätekonzept
 - 605: Elektrische Schnittstel
 - 618: Sicherheitsanforderun

| ID | Text | Topics |
|------|--|---|
| 1105 | Abbildung 1: Schematische Darstellung des TSG mit seinen penpheren Komponenten | |
| 583 | Technische Anforderungen | |
| 1088 | Die in diesem Lastenheft beschriebene Komponente wird als TSG. | |
| 3940 | Benennung und Teile-Ident-Nr. | |
| 3941 | Folgende Teile-Ident-Nummer wird festgelegt: | |
| 3942 | Mercedes-Benz Teile-Ident-Nr.: BO-TSG-5432
Benennung:TSG
Menge: 1
Lieferant: Fahrzeugteile Jackson AG
Lieferant-Bestell-Nr.: 12093478-1 | |
| 588 | Funktionen | |
| 1087 | In diesem Abschnitt wird ein kurzer Überblick über die Funktionalitäten und Eigenschaften des TSG aus Benutzersicht gegeben. Eine detaillierte Beschreibung der einzelnen Funktionalitäten findet sich in Abschnitt 5.2. In Zweifelsfällen gelten die Beschreibungen in Abschnitt 5.2. | |
| 589 | Funktionsbeschreibung | |
| 2718 | Aktivieren und Deaktivieren des TSG | |
| 2722 | Verhalten | ##NO_TOPICS## |
| 2750 | Aktivieren | |
| 2751 | Das Tür-SG (Türsteuergerät) wird geweckt, wenn eines der folgenden Ereignisse eintritt: | Betriebszustandsübergänge |
| 2752 | - Beifahrer-Tür wird geöffnet (Eingang T_OFFEN wird gegen Masse geschaltet) | Beifahrertür, Beifahrerseite |
| 2753 | - CAN-Bus wird geweckt | |
| 2754 | - Anheben der Türgriffeiste (Eingang T_GRIFF wird gegen Masse geschaltet) | |
| 2755 | - Schloßnußschalter schaltet (KEY_STATE wird auf Aufschließen oder Abschließen geschaltet) | |
| 2756 | Ab dem Zeitpunkt, zu dem das Tür-SG geweckt wird, ist die Batteriespannung (BATT) zu überwachen. Ist die Batteriespannung innerhalb der ersten 2 sec. immer unter 8 V, so legt sich das Tür-SG nach diesen 2 sec. sofort wieder schlafen. Ist ausreichend Spannung vorhanden (d.h. BATT >= 8 V), wird der CAN-Bus geweckt (aktiviert) und damit werden auch die anderen Steuergeräte am Innenraumbus geweckt. Dazu gehören z.B. die Außen- und Innenlichtsteuerung, die Steuerung der Sitze usw. | Innenlicht, Betriebszustandsübergänge, Spannung, Elektrik, Zeit |
| 2757 | Deaktivieren | Betriebszustandsübergänge |
| 2758 | Das Tür-SG wird deaktiviert (d.h., es geht in den Ruhezustand), wenn die Zündung aus ist (d.h. keine Zündung, also KL15_x = 0, siehe Tabelle 5) und folgende Bedingungen zutreffen: | Betriebszustandsübergänge, Zündung |
| 2759 | - Seit mehr als 10 min. wurde keine Tür mehr geöffnet oder geschlossen | Zeit, Öffnen |
| 2760 | - CAN-Signal SLEEP_NOW | |
| 2761 | Das zwischenzeitliche Öffnen oder Schließen der Tür setzt die Wartezeit wieder zurück. | |
| 2762 | Ist der Motor für mehr als 60 min. aus (MOTOR_LFT = 0), so wird nach Ablauf der 60 min. nach dem letzten Aktivieren das Tür-SG deaktiviert, um die Fahrzeugbatterie zu schonen (unabhängig vom Öffnen oder Schließen der Türen). | |
| 2723 | Initialisierung | |
| 2763 | Grundinitialisierung | |
| 2766 | Bei Inbetriebnahme prüft das Tür-SG, ob sich die Funktionssoftware in einem integen Zustand befindet (z.B. durch Prüfsummenberechnung). Kann die Integrität der Funktionssoftware nicht festgestellt werden, setzt das Tür-SG eine interne Fehlermarke (siehe auch Abschnitt 5.2.2.1) und legt sich sofort wieder schlafen. | |

Topic recommendation for #2759

[##NO_TOPICS##]

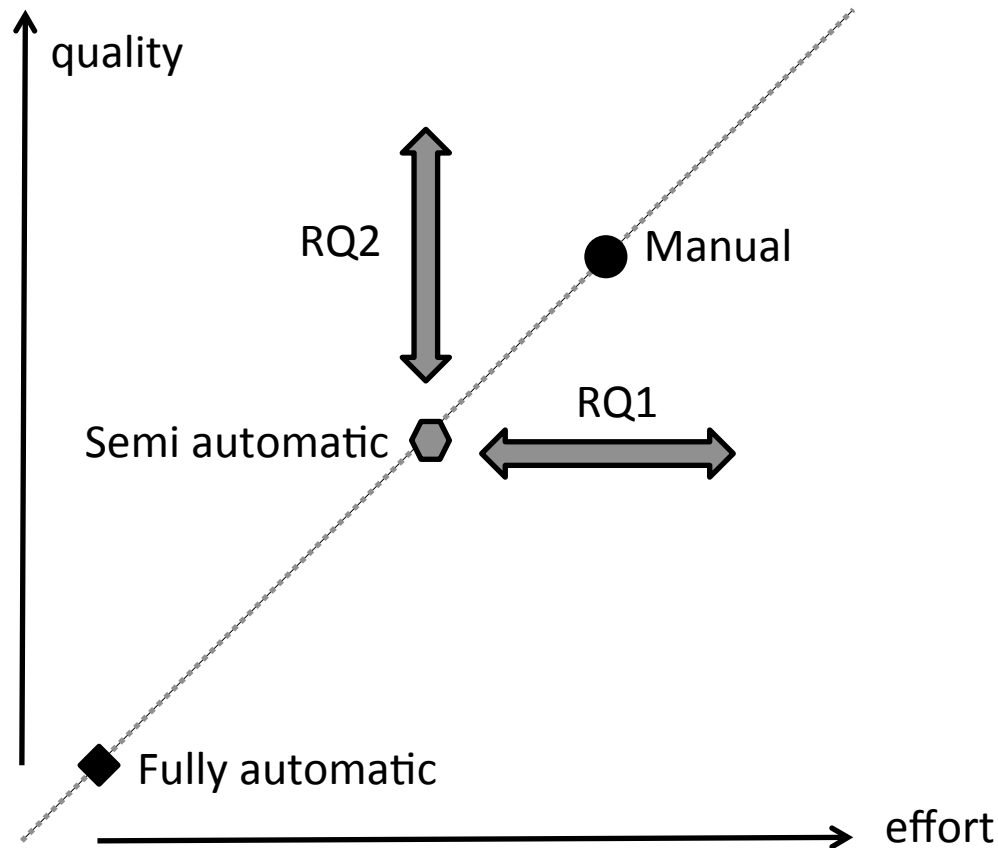
[Zeit](#)

[Betriebszustandsübergänge](#)

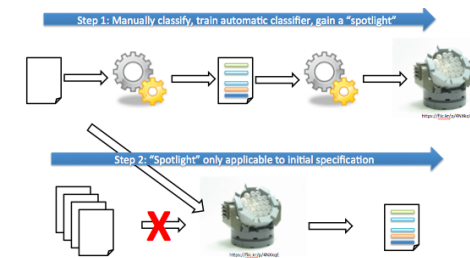
[Öffnen](#)

[Zündung](#)

What do we want to learn?



+ Ability to adjust to new domain



Four relevant treatments

| Treatment | | Automatic classification support | Initial training | User confirms classification |
|-----------|-----------|----------------------------------|------------------|------------------------------|
| T1 | manual | No | No | Yes |
| T2 | automatic | Yes | Yes | No |
| T3 | semi- | Yes | No | Yes |
| T4 | automatic | Yes | Yes | Yes |

Initial training

Data sources



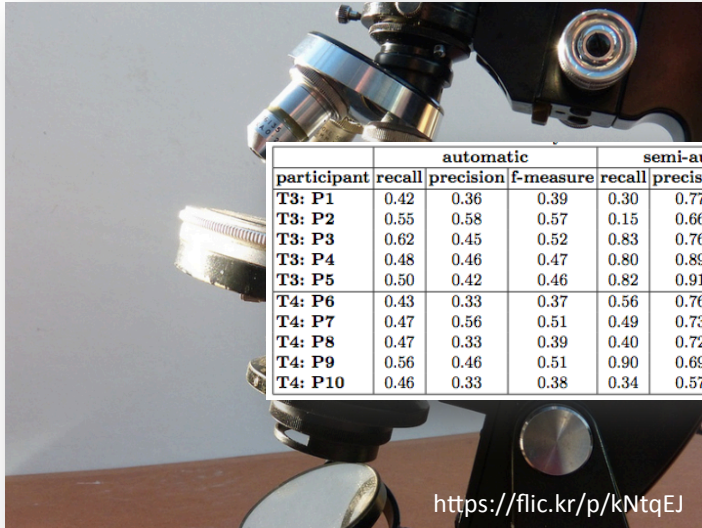
End-result vs. Ground truth

Telemetry



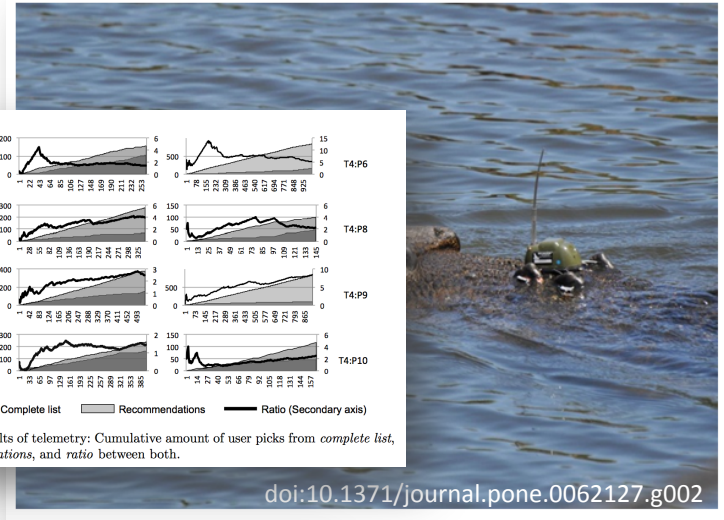
Questionnaire

Data sources



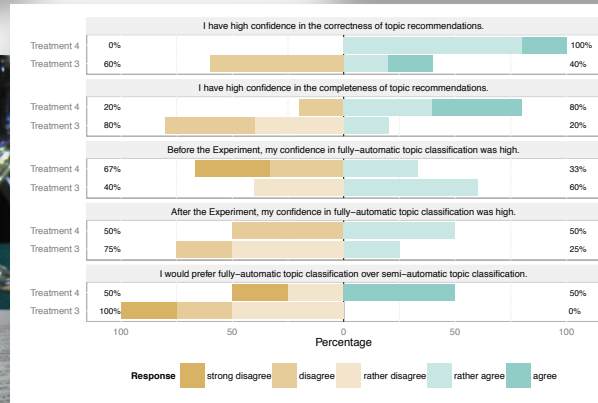
| participant | automatic | | | semi-automatic | | |
|-------------|-----------|-----------|-----------|----------------|-----------|-----------|
| | recall | precision | f-measure | recall | precision | f-measure |
| T3: P1 | 0.42 | 0.36 | 0.39 | 0.30 | 0.77 | 0.43 |
| T3: P2 | 0.55 | 0.58 | 0.57 | 0.15 | 0.66 | 0.25 |
| T3: P3 | 0.62 | 0.45 | 0.52 | 0.83 | 0.76 | 0.79 |
| T3: P4 | 0.48 | 0.46 | 0.47 | 0.80 | 0.89 | 0.84 |
| T3: P5 | 0.50 | 0.42 | 0.46 | 0.82 | 0.91 | 0.87 |
| T4: P6 | 0.43 | 0.33 | 0.37 | 0.56 | 0.76 | 0.65 |
| T4: P7 | 0.47 | 0.56 | 0.51 | 0.49 | 0.73 | 0.59 |
| T4: P8 | 0.47 | 0.33 | 0.39 | 0.40 | 0.72 | 0.51 |
| T4: P9 | 0.56 | 0.46 | 0.51 | 0.90 | 0.69 | 0.78 |
| T4: P10 | 0.46 | 0.33 | 0.38 | 0.34 | 0.57 | 0.43 |

<https://flic.kr/p/kNtqEJ>



Telemetry

End-result vs. Ground truth

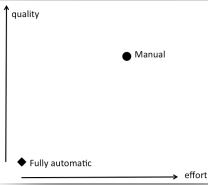


<https://flic.kr/p/4HU1ta>

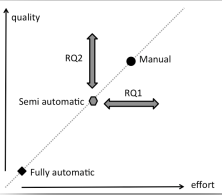
Questionnaire

Performance of fully auto. classif.

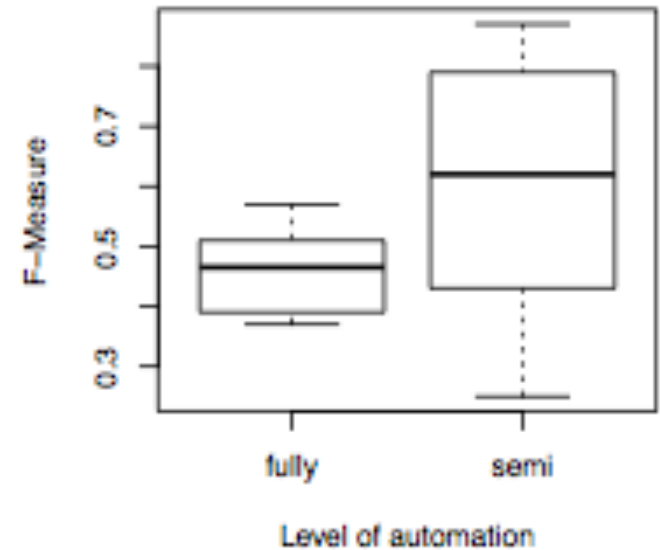
- **Quality: Not good enough**
 - (typical goal: recall > 0.7, precision > 0.6)
 - Used training data from one domain in a different domain
- **Effort: Naïve answer is “Great!”**
 - Automatic: init = 5min, classification < 1min
 - Manual: >>1hr
- **But:**
 - We need high quality training data
 - Two raters, Cohen’s Kappa, 2000-3000 requirements
>150hrs
 - This effort might never pay off!



Performance of semi-auto. approach



Better quality than full automatic?



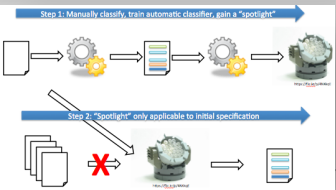
Only classification:

1.6min/reqt

Less effort than manual?

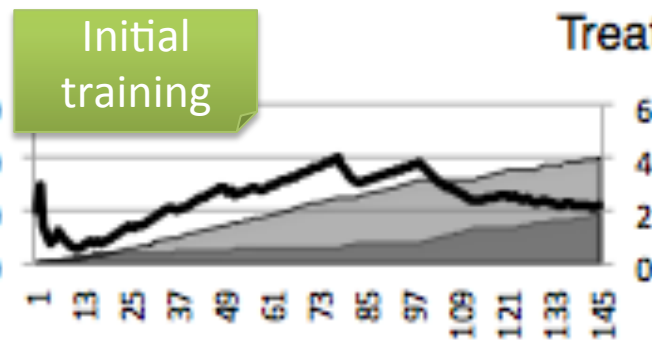
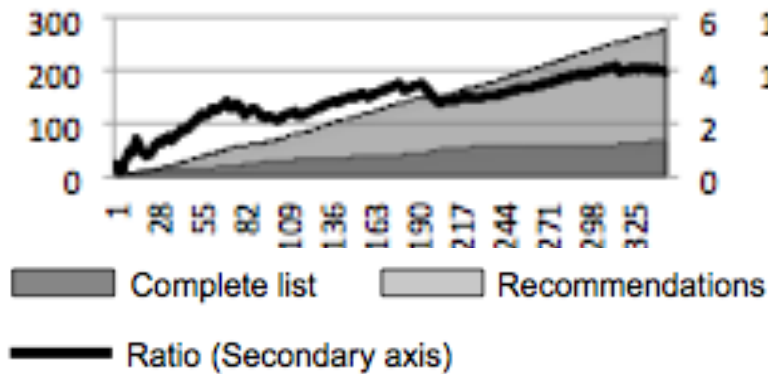
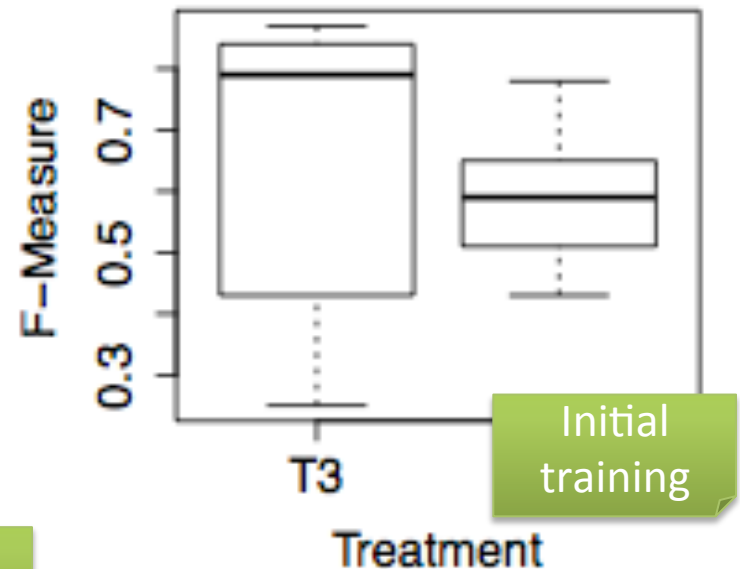
Reported time by participants:

3.76min/reqt = (2.99 + .77)min/reqt



Cold start: adjust to new domain

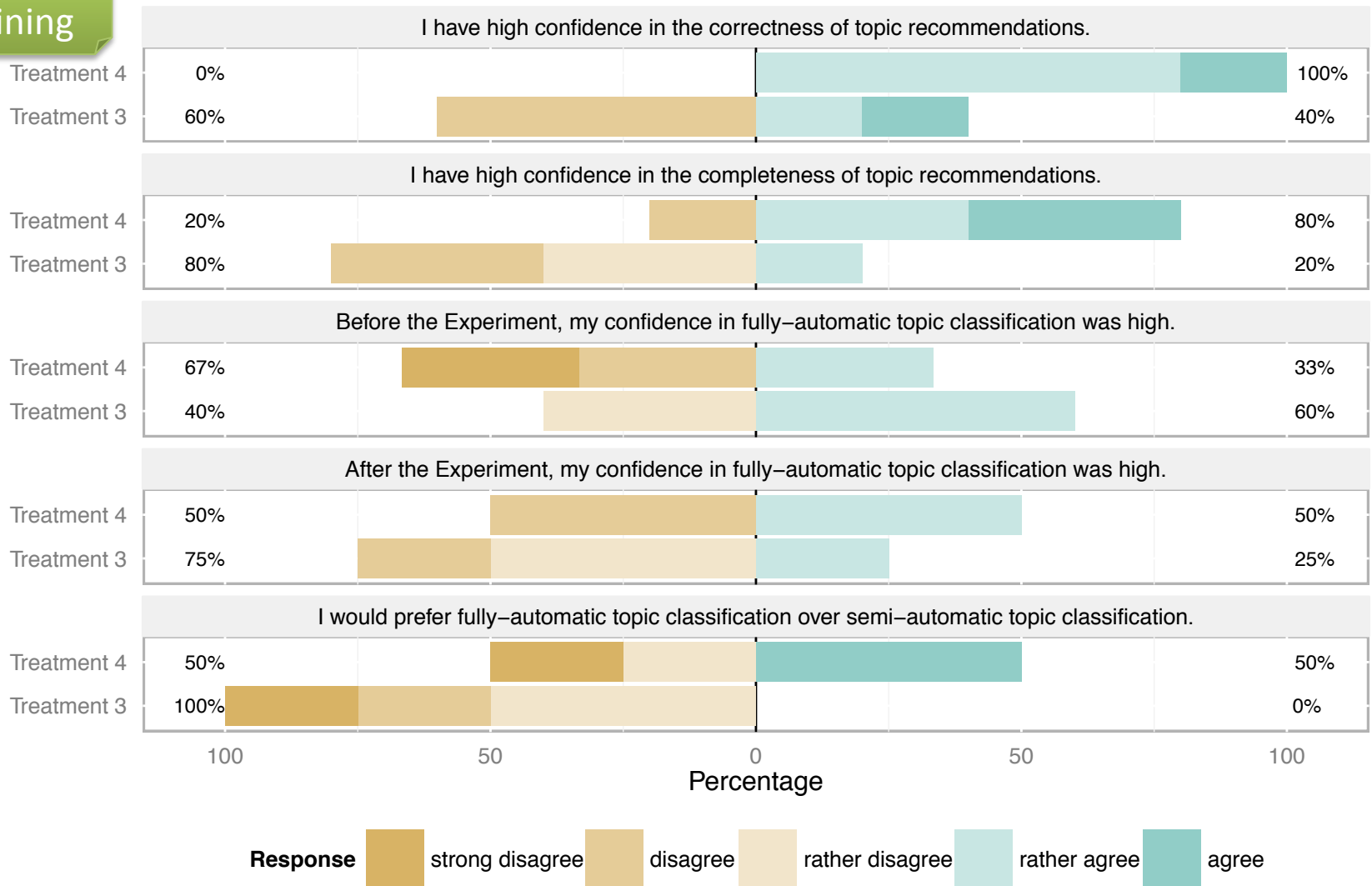
Does initial training lead to better results?



Does the classifier adopt quicker without initial training?

Result: Questionnaire

Initial
training



Provocative statement

“Without initial training,
the socio-technical system
of engineers and automatic classifier
performs better – but engineers hate it.”

Conclusion

- Semi-automatic classification can mitigate some major challenges in automatic requirements classification
 - Reasonable ratio of effort/quality
 - Supports organizational learning
 - Mitigates lack of high quality training data
- Plus
 - Recommendations can carry some shared knowledge of the team
 - ...can facilitate discussion about how to classify

Thank you!

Eric Knauss and Daniel Ott

eric.knauss@cse.gu.se

@oerich

oerich.wordpress.com

Results: End-result

| | automatic | | | semi-automatic | | |
|----------------|-----------|-----------|-----------|----------------|-----------|-----------|
| participant | recall | precision | f-measure | recall | precision | f-measure |
| T3: P1 | 0.42 | 0.36 | 0.39 | 0.30 | 0.77 | 0.43 |
| T3: P2 | 0.55 | 0.58 | 0.57 | 0.15 | 0.66 | 0.25 |
| T3: P3 | 0.62 | 0.45 | 0.52 | 0.83 | 0.76 | 0.79 |
| T3: P4 | 0.48 | 0.46 | 0.47 | 0.80 | 0.89 | 0.84 |
| T3: P5 | 0.50 | 0.42 | 0.46 | 0.82 | 0.91 | 0.87 |
| T4: P6 | 0.43 | 0.33 | 0.37 | 0.56 | 0.76 | 0.65 |
| T4: P7 | 0.47 | 0.56 | 0.51 | 0.49 | 0.73 | 0.59 |
| T4: P8 | 0.47 | 0.33 | 0.39 | 0.40 | 0.72 | 0.51 |
| T4: P9 | 0.56 | 0.46 | 0.51 | 0.90 | 0.69 | 0.78 |
| T4: P10 | 0.46 | 0.33 | 0.38 | 0.34 | 0.57 | 0.43 |

T2

Results: Telemetry

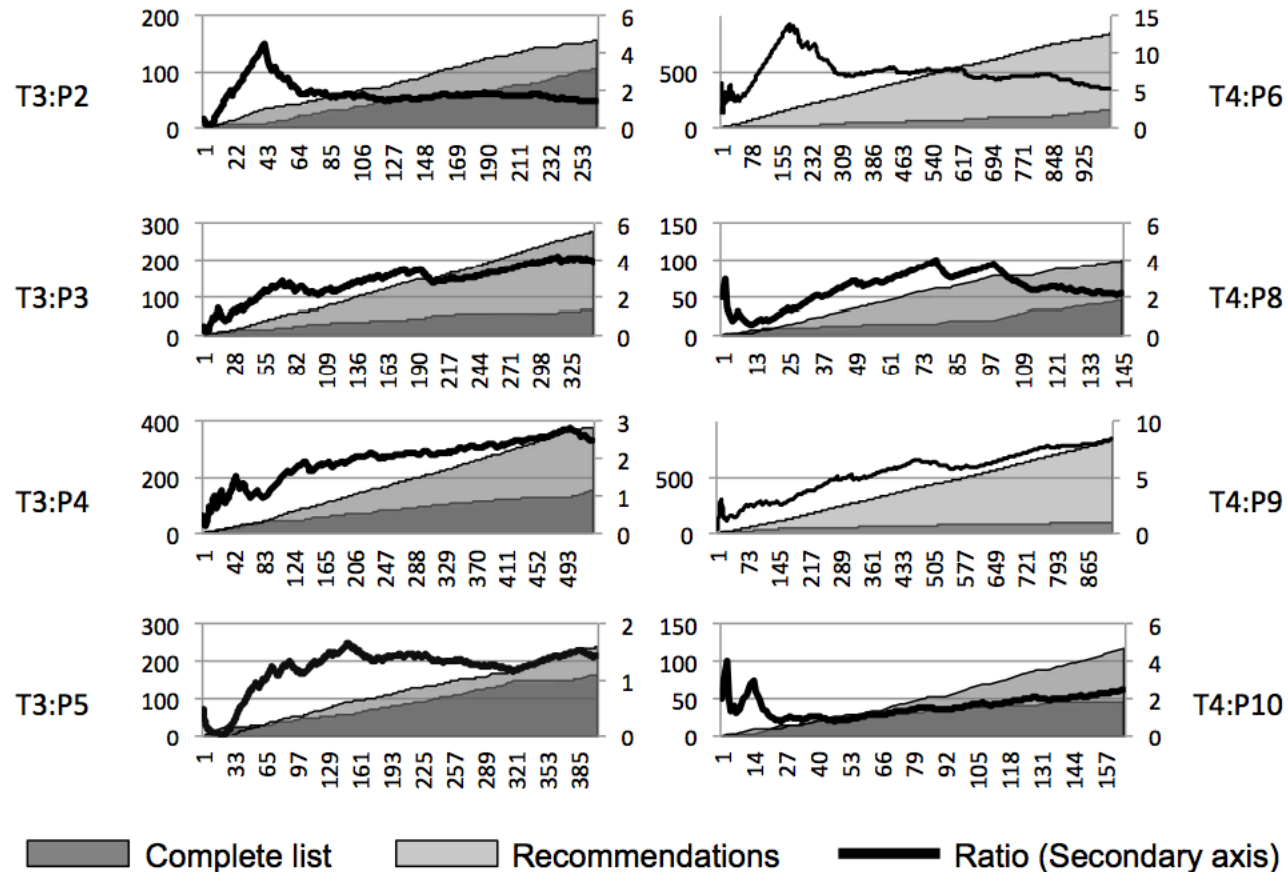


Fig. 4: Results of telemetry: Cumulative amount of user picks from *complete list*, *recommendations*, and *ratio* between both.

Result: Questionnaire

