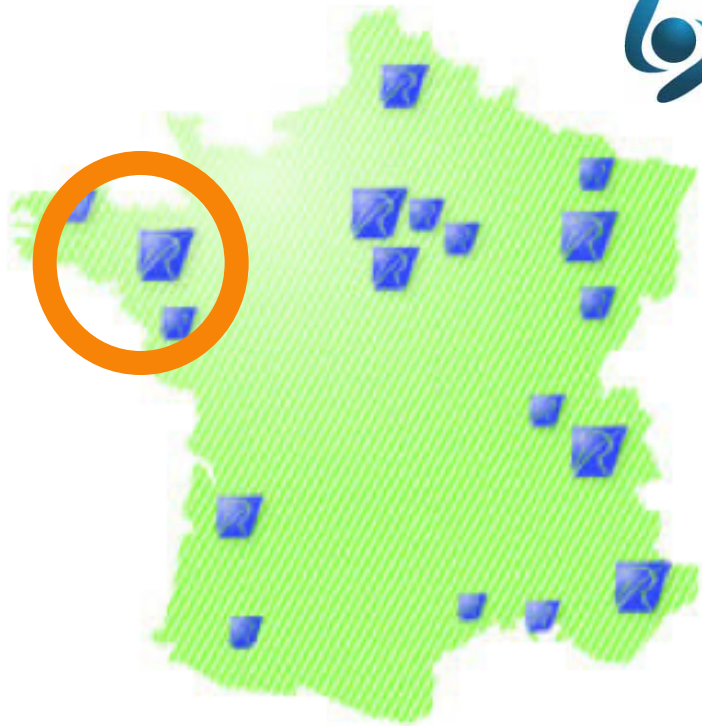


# Modeling Variability in Natural Language Documents: Two Case Studies

Sana Ben Nasr, Mathieu Acher and Benoit Baudry

GDR GPL, 09 October 2015  
Paris

# DiverSE team



Research in software engineering.

- 8 faculty members
- 40 researchers and engineers on projects
- SPL, MDE, distributed systems and verification



There are numerous artefacts that exhibit features and their dependencies.



**Modeling  
variability  
is crucial**

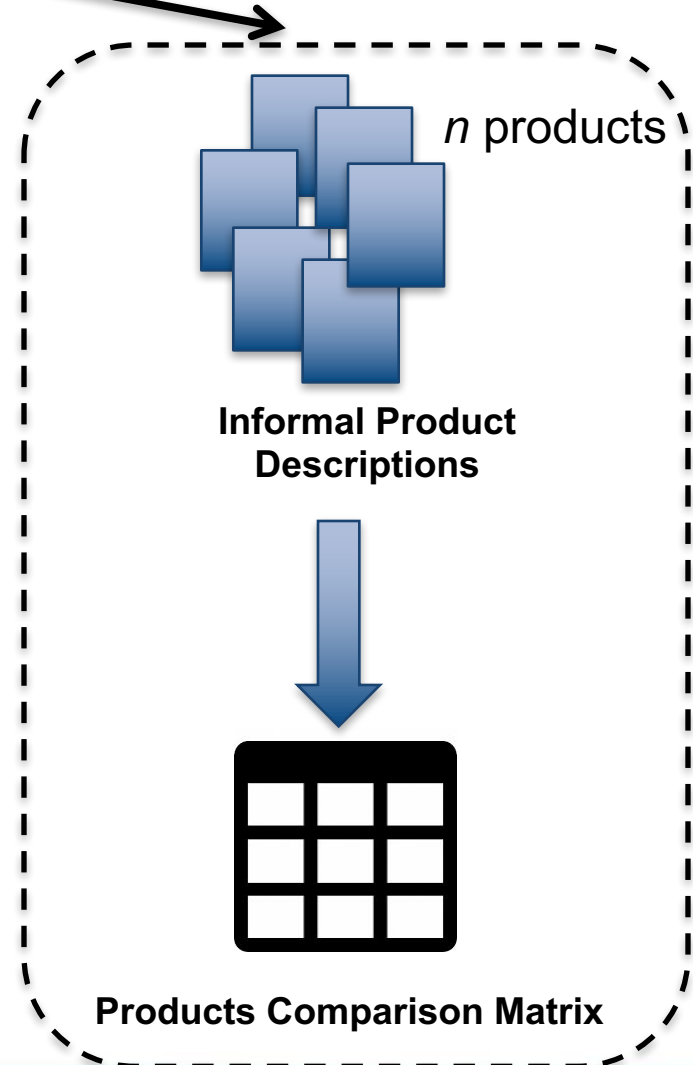
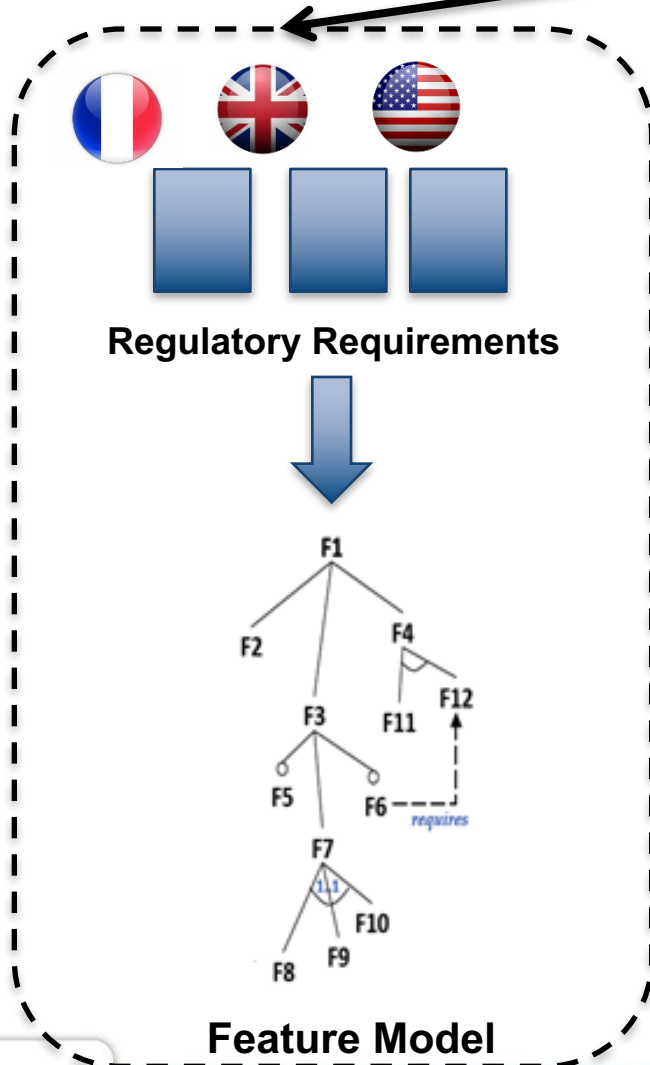


Product Line

# Mining and Modeling Variability from Informal, Textual Documents

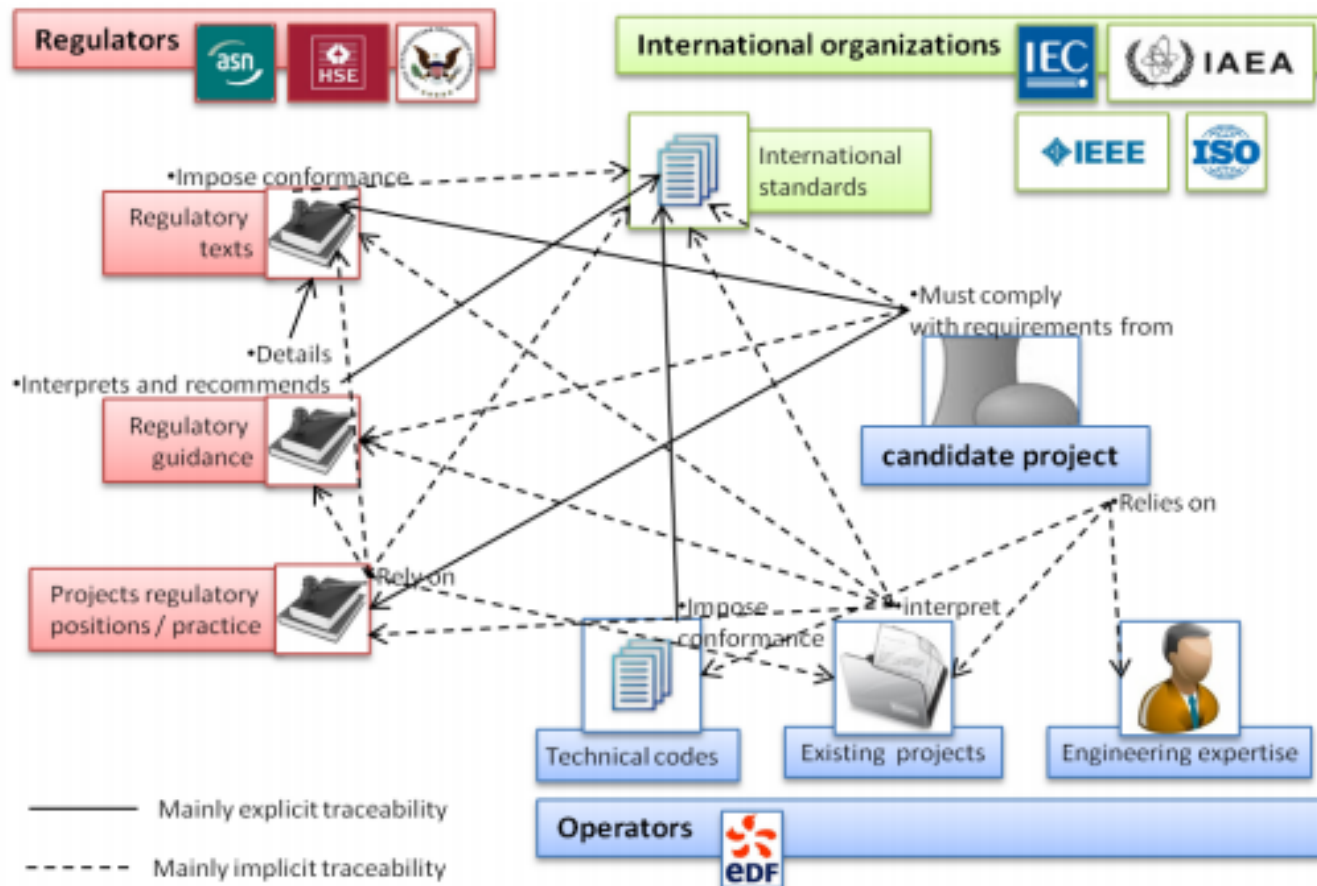
## Case Study 1: Nuclear power plants

## Case Study 2: Manufactured products



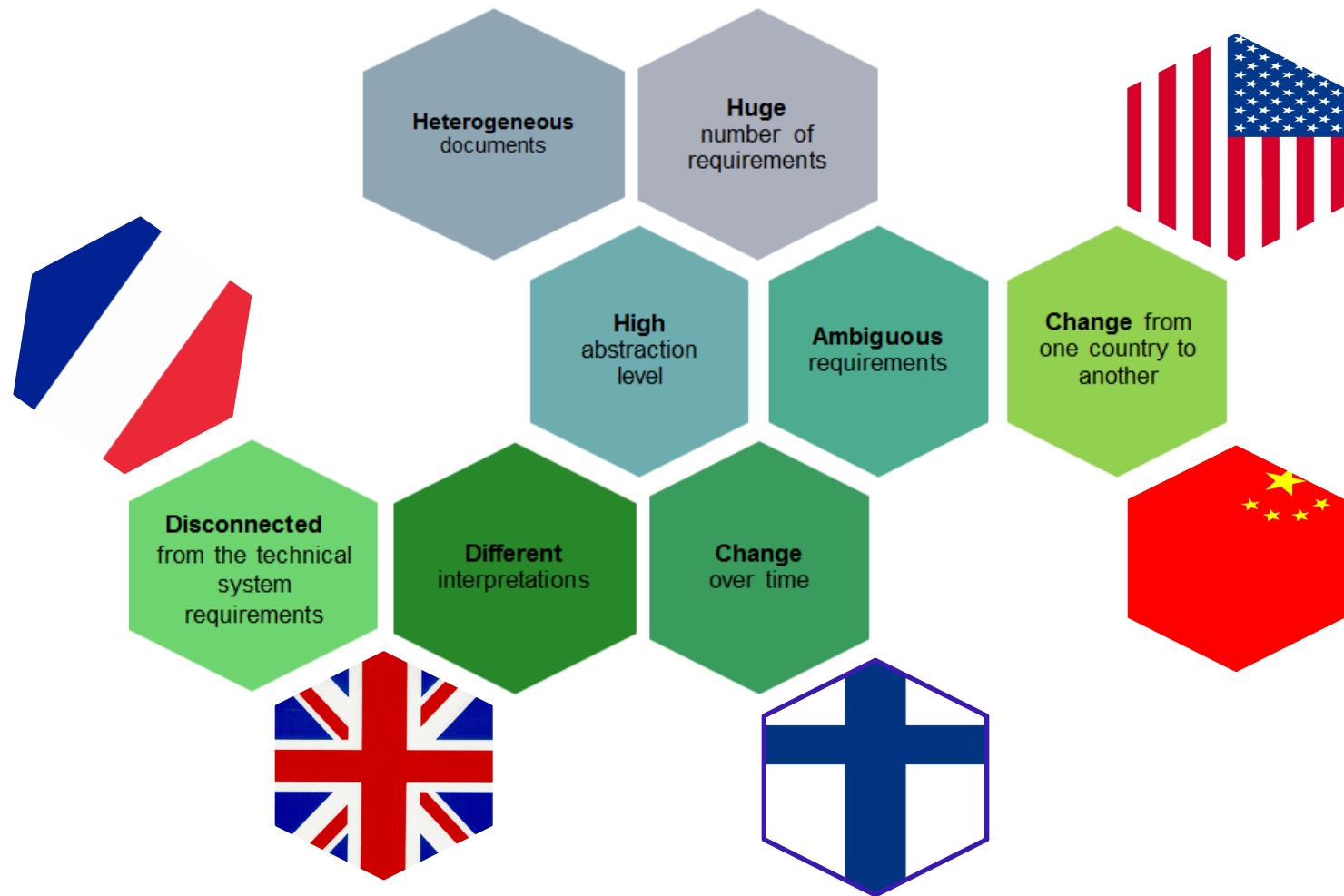
# Case Study 1: Modeling Variability in Nuclear Domain

# CONNEXION Project : Overview of the nuclear regulatory landscape



RELAW 2012 - Sannier 2012/09/25

# Complexity of regulatory requirements

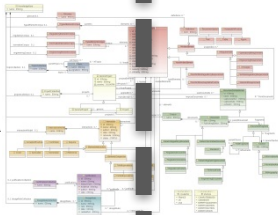


# Lack of Product Line Culture

## Requirements repository



instantiated by project



## Design rules catalogue



instantiated by project

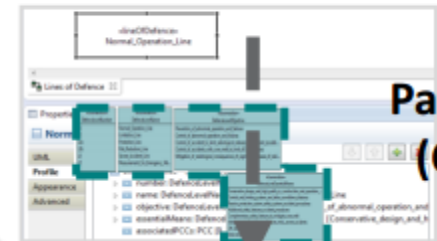
conformant to

conformant to

## Design partner of I&C Architecture



instantiated by project



**Papyrus (CEA)**

**Kermeta (Inria)**



regulatory requirements and practices

Design rules

I&C Architecture

traceability demonstration / variability impact

traceability demonstration / variability impact



# Overview of the Approach

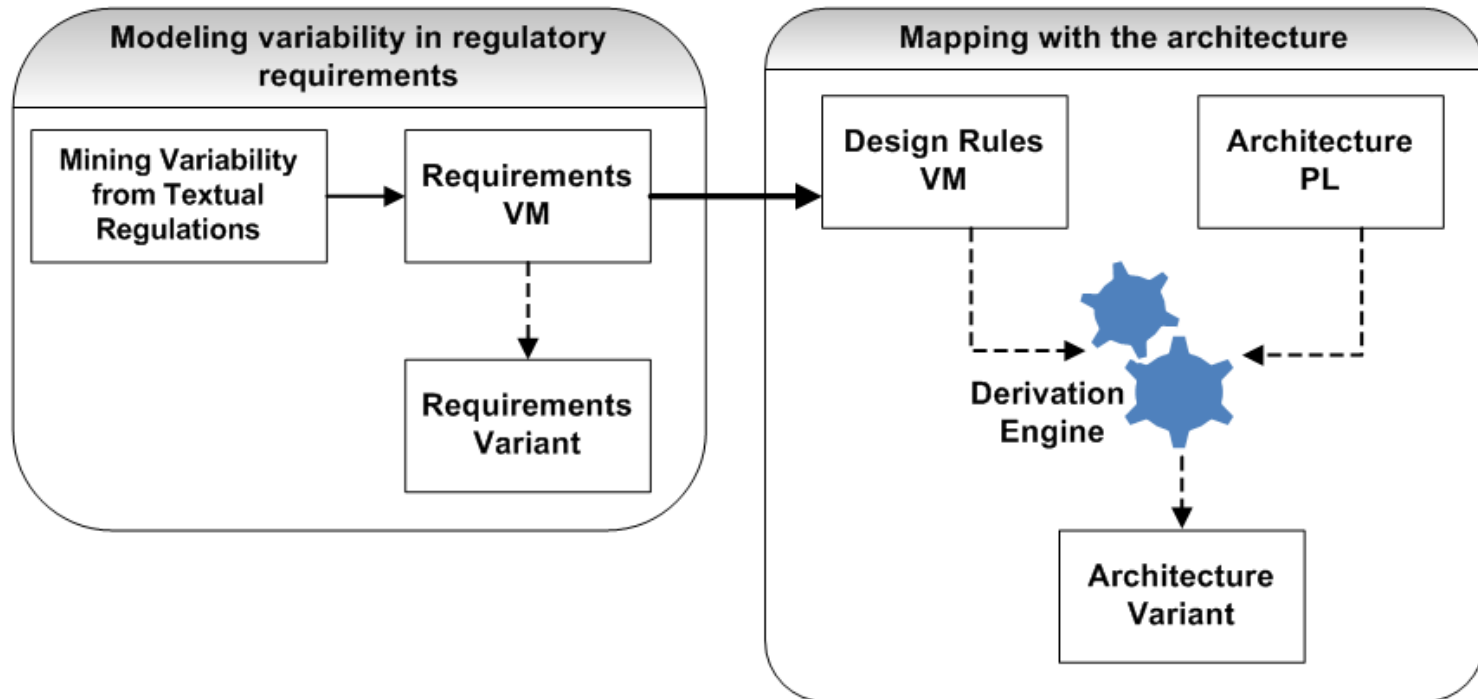


Fig. A two-stages approach (overview)

Sana Ben Nasr, Nicolas Sannier, Mathieu Acher, and Benoit Baudry. **Moving Toward Product Line Engineering in a Nuclear Industry Consortium.** In *18th International Software Product Line Conference (SPLC'2014)*, Florence, Italy, September 2014.

## Topic: Independence

Index	Verbatim	Countries
IEC 60709.1	Systems performing category A functions shall be designed so as to prevent physical effects caused by faults and normal action of those systems, and shall ensure independence between redundant parts.	France and UK
IEC 60709.11	Failures and malfunctions of category B or C systems shall be prevented by design, such as noise, or of a system which might impair the ability of the system to perform its function.	France and UK
IEC 60709.12	Where signals are extracted from category B or C systems for use in lower category systems, isolation devices may not be required; however, good engineering practices should be followed.	France and UK
IEEE 384.1	Physical separation and electrical isolation shall be provided to ensure independence of Class 1E circuits and equipment during and following any design basis event can be accomplished.	US
IEEE 384.5	1) Non-Class 1E circuits shall be physically separated from Class 1E circuits and associated components. 2) Non-Class 1E circuits shall be electrically isolated from Class 1E circuits and associated circuits by the use of isolation devices, such as isolating transformers or separation distance.	US

## Requirements similarity identification

Requirements are considered related if they concern **similar matters**.

## Feature clustering

Requirements which are **semantically similar** are "**clustered**" to form a **feature**.

### Example:

IEC 60709.1, IEC 60709.11 and IEC 60709.12 => **Prevent System Degradation**

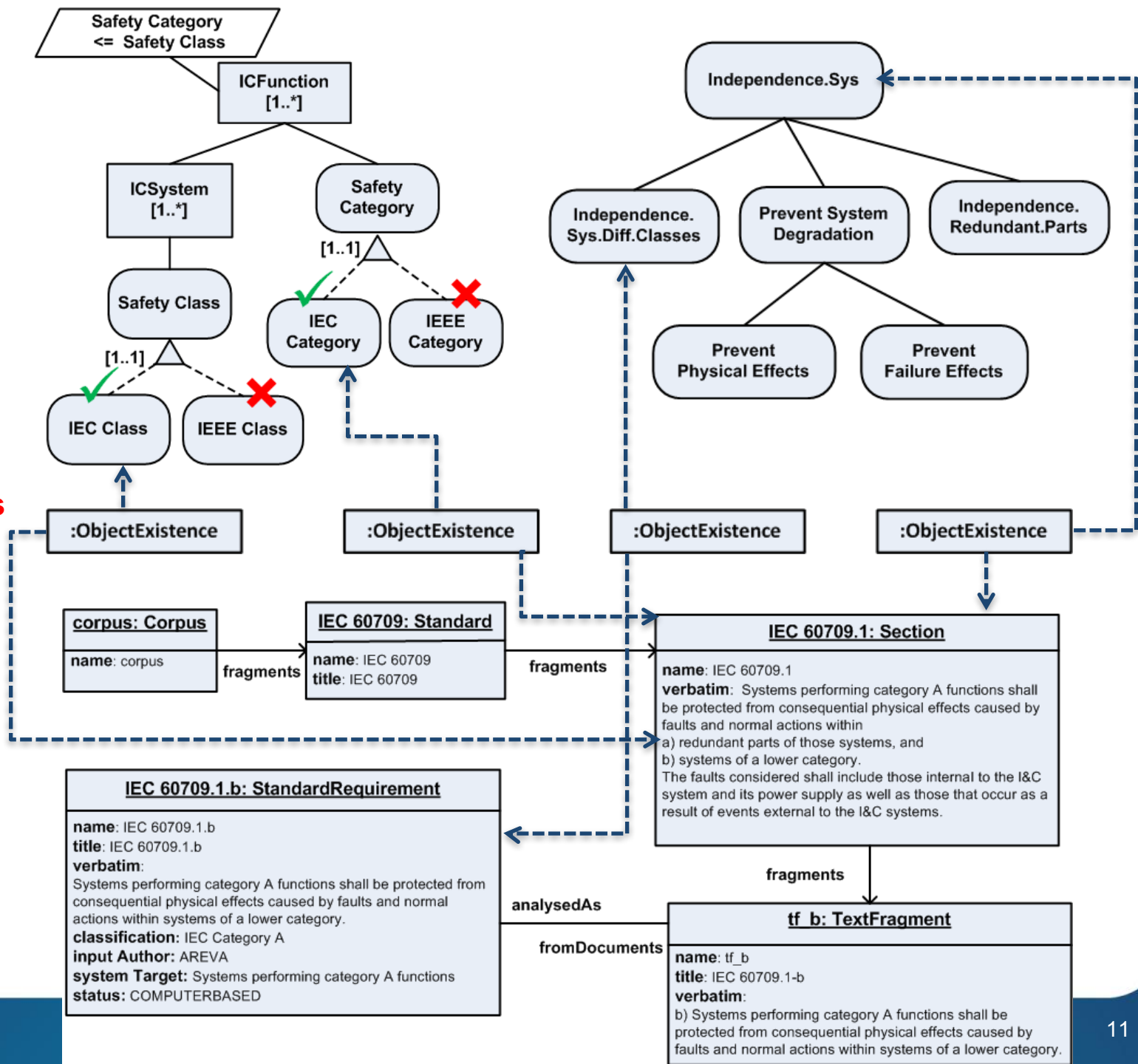
IEEE 384.1 and IEEE 384.5 => **Communication Separation**

IEC 60709.1, IEC 60709.11, IEC 60709.12, IEEE 384.1 and IEEE 384.5 => ***Independence between Systems of Different Classes***

## Standards VAM

## Variation points

## Standards BM



## Mapping between Design Rules VAM and I&C Architecture

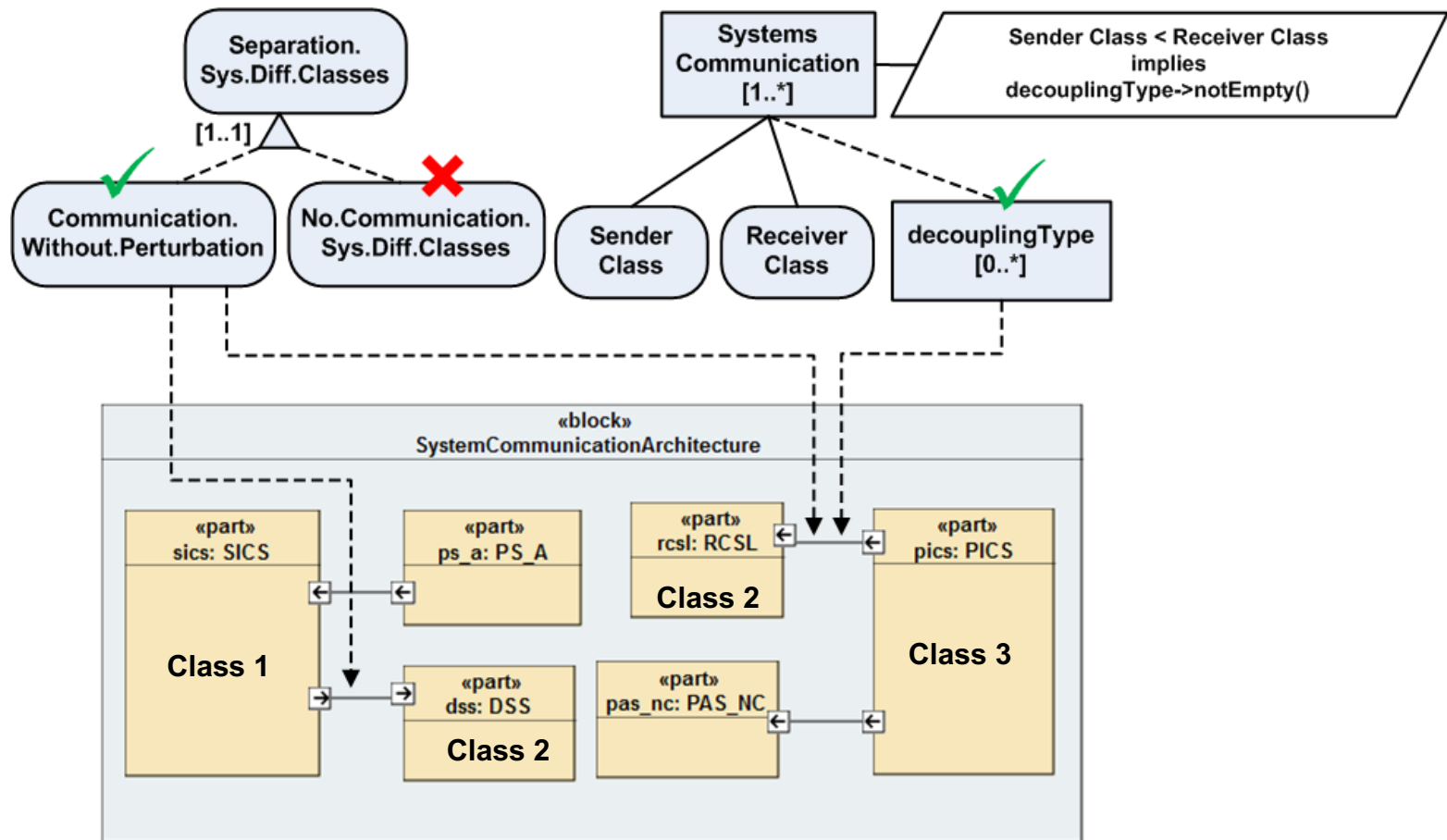


Fig. Mapping Design Rules VAM with Systems Communication

# Lessons Learned and Discussion

## **RQ1: How to Introduce Variability Modeling to Nuclear Engineers?**

- We set up several half to full-day workshops
- We proposed a dedicated variability training performed by one of the author
- We localized where there will be valuable variability to model
- We elicited what our industry partners expect from the variability modeling.

## **RQ2: What is the Best Variability Modeling Paradigm Accordingly to the Domain Practices?**

- We presented several different modeling paradigms, analysis capabilities, current limitations.
- We discussed the choice of the modeling paradigm and the format of data that we will have to handle.
- CVL is domain independent, it supports a broad range of types, including multiple instances, and a constraint language for expressing dependencies over these types.

# Lessons Learned and Discussion

## **RQ3: How to Deal with the Search Space in Regulatory Corpora and What Is the Right Granularity Level for Modeling Requirements Variability?**

- To narrow the problem space, the first idea was to analyze variability in regulatory documents by topic on different corpora and on the same abstraction level.
- Using traceability matrix improves the understandability and the maintainability of variability models.

## **RQ4: How to Address Requirements Variability on the Architecture?**

- There is no direct mapping from requirements to the architecture.
- We propose heading toward modeling variability in design rules since they act as a pivot between requirements and the architecture
- This idea allowed us to investigate the robustness of the architecture against regulatory requirements.

## **Case Study 2: Extraction of Product Comparison Matrices From Informal Product Descriptions**



# Extract Variability from Informal Product Descriptions





# Challenge: Automatic Extraction of PCMs

**HP - ENVY 17.3" Touch-Screen Laptop - Intel Core i7 12GB Memory 1TB Hard Drive - Red**



PRICE MATCH GUARANTEE

Add to Cart

**\$999.99**

FREE SHIPPING

Enjoy stunning images with this HP ENVY m7-k211dx laptop's 17.3" high-definition **touch screen** and NVIDIA GeForce 840M graphics, with 2GB dedicated video memory, which display games, movies and streaming media in brilliant clarity and detail.

**5th Gen Intel® Core™ i7-5500U processor.** Features a 4MB L3 cache and **2.4GHz processor speed.**

Intel® Core™ i7 processor. Delivers efficient multiway processing for ultimate smart performance. Intel® Turbo Boost Technology delivers extra performance when you need it and increased energy efficiency when you don't.

**12GB DDR3L SDRAM memory.** For multitasking power. Multiformat DVD ± RW/CD-RW drive. Create custom DVDs and CDs. 17.3" WLED-backlit high-definition touch-screen display With **1920 x 1080 resolution** enables simple navigation of content and showcases your favorite media in arresting clarity.

**1TB hard drive** (5400 rpm). Offers spacious storage. ProtectSmart hard drive protection safeguards your stored data.

NVIDIA GeForce 840M graphics. Feature 2GB DDR3 dedicated video memory and up to 4096MB total video memory for **high-quality images.** HDMI output for connection to an HDTV.

**Front-facing TrueVision high-definition webcam.** With an integrated dual-array digital microphone makes it easy to video chat with friends and family members.

**Digital media reader.** Supports SD format.

**Microsoft Windows 8.1 64-bit operating system** preinstalled. Provides a stable platform for word processing, Web navigation, gaming, media storage and more.

Software package included. With CyberLink PhotoDirector, CyberLink PowerDirector and more. Includes 1-month trial of Microsoft Office 365.

## Features

HP - ENVY

Lenovo - Yoga

Apple® - MacBook

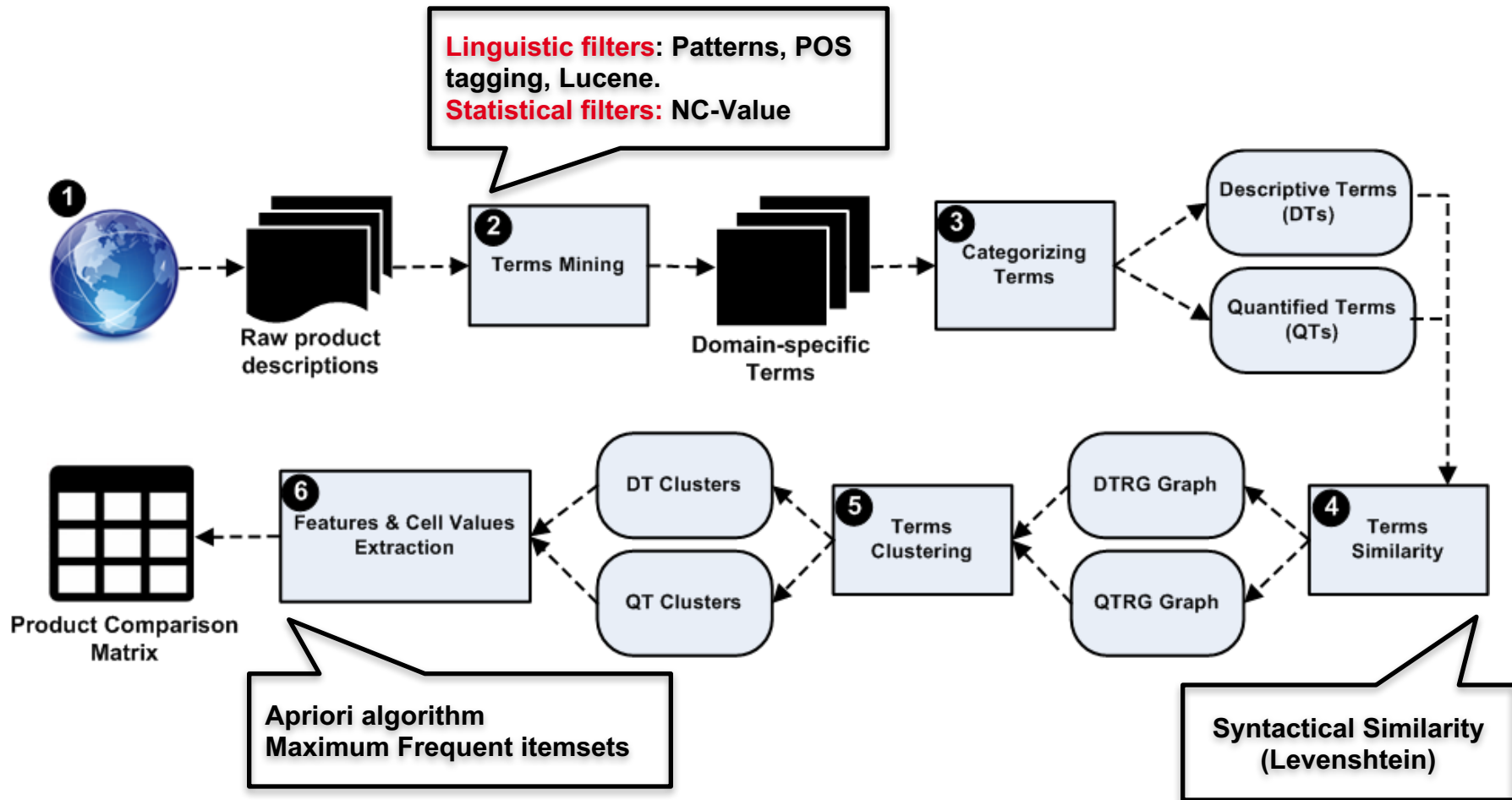
Dell - Inspiron

Features	HP - ENVY	Lenovo - Yoga	Apple® - MacBook	Dell - Inspiron
High-quality images	YES	YES	YES	YES
Voice recognition software	NO	YES	NO	YES
Touch Screen	YES	YES	NO	YES
Media Reader	Digital	Multiformat	Digital	5-in-1
Webcam	Front-facing TrueVision high-definition	N/A	Built-in high-definition	Built-in high-definition
Hard Drive	1TB	500 GB	750 GB	1TB
Resolution	1920 x 1080	1366 x 768	2560 x 1600	1920 x 1080
Processor Speed	2.4 Ghz	1.5 Ghz	2.8 Ghz	3.0 Ghz
Memory	12GB DDR3L SDRAM	4GB DDR3L	16 GB	16GB DDR3L
Processor	5th Gen Intel Core i7-5500U	4th Gen Intel Core i3-4012Y	Fourth-generation quad-core Intel Core i7	5th Gen Intel Core i7-5500U
Operating System	Microsoft Windows 8.1 64-bit	Microsoft Windows 8.1 64-bit	Mac OS X 10.9 Mavericks	Microsoft Windows 8.1 64-bit


(exp. overviews)  
**Product Overview**

**PCM (Products Comparison Matrix)**

# Overview of the Approach



Sana Ben Nasr, Guillaume Bécane, Mathieu Acher, João Bosco Ferreira Filho, Benoit Baudry, Nicolas Sannier, and Jean-Marc Davril. **MatrixMiner : A Red Pill to Architect Informal Product Descriptions in the Matrix**. In *ESEC/FSE'15*, Bergamo, Italy, August 2015.



**HP - ENVY 17.3" Touch-Screen Laptop - Intel Core i7 12GB Memory 1TB Hard Drive - Red**

**\$999.99**

FREE SHIPPING

Enjoy stunning images with this HP ENVY m7-4211dx laptop's 17.3" high-definition **touch screen** and NVIDIA GeForce 840M graphics, with 2GB dedicated video memory, which display games, movies and streaming media in brilliant clarity and detail.

**5th Gen Intel® Core™ i7-5500U processor**, Features a 4MB L3 cache and **2.4GHz processor speed**

Intel® Core™ i7 processor. Delivers efficient multitask processing for ultimate smart performance. Intel® Turbo Boost Technology delivers extra performance when you need it and increased energy efficiency when you don't.

**12GB DDR3L SDRAM memory**, For multitasking power.

Multiformat DVD-R, RW drive. Create custom DVDs and CDs.

17.3" WLED-backlit high-definition touch-screen display

Win **1920 x 1080 resolution** enables simple navigation of content and showcases your favorite media in arresting clarity.


**1TB hard drive** (5400 rpm). Offers spacious storage. ProtectSmart hard drive protection safeguards your stored data.

NVIDIA GeForce 840M graphics. Feature 2GB DDR3 dedicated video memory and up to 4096MB total video memory for **high-quality images**. HDMI output for connection to an HDTV.

**Front-facing TrueVision high-definition webcam**, With an integrated dual-array digital microphone makes it easy to video chat with friends and family members.

**Digital media reader**, Supports SD format.

**Microsoft Windows 8.1 64-bit operating system** preinstalled. Provides a stable platform for word processing, Web navigation, gaming, media storage and more.



**Lenovo - Yoga 2 2-in-1 11.6" Touch-Screen Laptop Intel Core i3 - 4GB Memory 500GB Hard Drive Silver/Black**

**\$579.99**

FREE SHIPPING

With 4 usage modes, this Lenovo YOGA2 11-59430714 laptop makes it easy to choose between laptop, stand, tent and tablet settings to accommodate your fluctuating lifestyle. The 11.6" 10-point capacitive multitouch screen allows precise navigation.

**4th Gen Intel® Core™ i3-4012Y processor**

Features a 3MB cache and **1.5GHz processor speed**.

For multitasking power.

11.6" high-definition 10-point capacitive multitouch-screen display. With IPS technology and **1366 x 768 resolution** presents visual media in striking detail and makes it simple to navigate features.

**500GB hybrid hard drive**

Incorporates an 8GB solid state drive as a cache to offer fast start-up times and plus plenty of room to store pictures, videos, music and other important files.

**Lenovo Smart Voice voice recognition software**

Lets you perform basic commands using just your voice.


Intel® HD Graphics 4200

Feature shared video memory for lush images with rich, intricate detail. A micro HDMI output enables simple connection to an external high-definition display or other HDMI-enabled device.

**Multiformat media reader**

Supports Secure Digital and MultiMediaCard formats.

**Microsoft Windows 8.1 64-bit operating system** preinstalled. Provides a stable platform for word processing, Web navigation, gaming, media storage and more.



**Apple® - MacBook Pro with Retina display - 15.4" Display - 16GB Memory - 750GB Flash Storage - Silver**

**\$849.99**

CLEARANCE

Save \$100 (reg. \$949.99)

FREE SHIPPING

With fourth-generation Intel Core processors, the latest graphics, and faster flash storage, the incredibly advanced MacBook Pro with Retina display moves even further ahead in performance and battery life.\*

\*Compared with the previous generation.

Retina display with **2560 x 1600 resolution** on the 13-inch model and 2880-by-1800 resolution on the 15-inch model.

**Fourth-generation quad-core Intel Core i7 processor**

Features **2.8GHz processor speed**

Intel Iris Graphics in the 13-inch model and Intel Iris Pro Graphics or Iris Pro plus NVIDIA GeForce GT 750M graphics with 2GB video memory in the 15-inch model.

**16GB Memory**

Up to 9 hours (13-inch model) or 8 hours (15-inch model) of battery life†

**750GB hard drive**

Offers spacious storage. ProtectSmart hard drive protection safeguards your stored data.

Faster flash storage†

802.11ac Wi-Fi

**Digital media reader**, Supports SD format.

Use Thunderbolt 2 ports for connecting high-performance devices and transferring data at lightning speed

Two USB 3 ports (compatible with USB 2 devices) and HDMI

**Built-in high-definition webcam** and integrated microphone


FaceTime HD camera

Phone, Numbers, Messages, iPhoto, iMovie, GarageBand included

**Mac OS X 10.9 Mavericks** operating system, the world's most advanced desktop operating system

† Battery life varies by use and configuration. See [www.apple.com/batteries](#) for more information

† 2 Compared with the previous generation.



**Dell - Inspiron 15.6" Touch-Screen Laptop - Intel Core i7 - 16GB Memory - 1TB + 8GB Hybrid Hard Drive - Black**

**\$949.99**

FREE SHIPPING

Stay connected with this Dell Inspiron i5548S33SLV laptop, which features a **touch-screen** display for simple navigation and built-in wireless networking for easy Web connection. AMD Radeon graphics with 4GB of video memory ensure smooth performance.

**5th Gen Intel® Core™ i7-5500U processor**, Features a 4MB cache and up to **3.0GHz processor speed** with Turbo Boost.

Intel® Core™ i7 processor

Delivers efficient multitask processing for ultimate smart performance. Intel® Turbo Boost Technology delivers extra performance when you need it and increased energy efficiency when you don't.

**16GB DDR3L memory**, For multitasking power.

15.6" LED backlit high-definition touch-screen display.

With IPS technology and **1920 x 1080 resolution** for intense detail and clarity and simple navigation.

**1TB hybrid hard drive**

Incorporates an 8GB solid state drive as a cache to offer fast start-up times plus plenty of room to store pictures, music, videos and other important files.

AMD Radeon R7 M270 graphics

Win 4GB DDR3L video memory for smooth video performance. An HDMI output enables flexible connectivity to digital devices.

**Built-in high-definition webcam** and integrated microphone

Lets you video chat with friends and family in 720p resolution.

**5-in-1 media reader**

Supports SD, SDHC, SD 3.0, MS and MS PRO formats.

**Microsoft Windows 8.1 64-bit operating system** preinstalled. Provides a stable computing platform.

Specifications	
Specs:	Details:
Height	1.22 inches
Width	16.52 inches
Depth	10.98 inches
Touch Screen	YES
Media Card Reader	YES
Built-In Webcam	YES
Hard Drive Capacity	1TB
Screen Resolution	1920 x 1080
Processor Speed	2.4 Ghz
System Memory (RAM)	12 GB
Processor Model	Intel 5th Generation Core i7
Operating System	Windows 8.1

Specifications	
Specs:	Details:
Height	0.7 inches
Width	11.7 inches
Depth	8.1 inches
Touch Screen	YES
Media Card Reader	YES
Built-In Webcam	YES
Hard Drive Capacity	500 GB
Screen Resolution	1366 x 768
Processor Speed	1.5 Ghz
System Memory (RAM)	4 GB
Processor Model	Intel 4th Generation Core i3
Operating System	Windows 8.1

Specifications	
Specs:	Details:
Height	0.71 inches
Width	14.13 inches
Depth	9.73 inches
Touch Screen	NO
Media Card Reader	YES
Built-In Webcam	NO
Hard Drive Capacity	750 GB
Screen Resolution	2560 x 1600
Processor Speed	2.8 Ghz
System Memory (RAM)	16 GB
Processor Model	Intel 4th Generation Core i7
Operating System	Mac OS X 10.9 Mavericks

Specifications	
Specs:	Details:
Height	0.9 inches
Width	14.9 inches
Depth	10 inches
Touch Screen	YES
Media Card Reader	YES
Built-In Webcam	YES
Hard Drive Capacity	1TB
Screen Resolution	1920 x 1080
Processor Speed	3.0 Ghz
System Memory (RAM)	16 GB
Processor Model	Intel 5th Generation Core i7
Operating System	Windows 8.1

Merge

Synthesize

Features	HP - ENVY	Lenovo - Yoga	Apple® - MacBook	Dell - Inspiron
High-quality images	YES	YES	YES	YES
Voice recognition software	NO	YES	NO	YES
Touch Screen	YES	YES	NO	YES
Media Reader	Digital	Multiformat	Digital	5-in-1
Webcam	Front-facing TrueVision high-definition	N/A	Built-in high-definition	Built-in high-definition
Hard Drive	1TB	500 GB	750 GB	1TB
Resolution	1920 x 1080	1366 x 768	2560 x 1600	1920 x 1080
Processor Speed	2.4 Ghz	1.5 Ghz	2.8 Ghz	3.0 Ghz
Memory	12GB DDR3L SDRAM	4GB DDR3L	16 GB	16GB DDR3L
Processor	5th Gen Intel Core i7-5500U	4th Gen Intel Core i3-4012Y	Fourth-generation quad-core Intel Core i7	5th Gen Intel Core i7-5500U
Operating System	Microsoft Windows 8.1 64-bit	Microsoft Windows 8.1 64-bit	Mac OS X 10.9 Mavericks	Microsoft Windows 8.1 64-bit

Features	HP - ENVY	Lenovo - Yoga	Apple® - MacBook	Dell - Inspiron
Height	1.22 inches	0.7 inches	0.71 inches	0.9 inches
Width	16.52 inches	11.7 inches	14.13 inches	14.9 inches
Depth	10.98 inches	8.1 inches	9.73 inches	10 inches
Touch Screen	YES	YES	NO	YES
Media Card Reader	YES	YES	YES	YES
Built-In Webcam	YES	YES	NO	YES
Hard Drive Capacity	1TB	500 GB	750 GB	1TB
Screen Resolution	1920 x 1080	1366 x 768	2560 x 1600	1920 x 1080
Processor Speed	2.4 Ghz	1.5 Ghz	2.8 Ghz	3.0 Ghz
System Memory (RAM)	12 GB	4 GB	16 GB	16 GB
Processor Model	Intel 5th Generation Core i7	Intel 4th Generation Core i3	Intel 4th Generation Core i7	Intel 5th Generation Core i7
Operating System	Windows 8.1	Windows 8.1	Mac OS X 10.9 Mavericks	Windows 8.1

Overview PCM

Specification PCM



# MATRIX MINER

Dataset: manual-dataset Category: Laptops Filter 1: Filter-Brand-Category Filter 2: Lenovo-2-in-1 Load

Prioritize Features

Select a Filter (brand, sub category)

Filter Cell Values

Sort Cell Values

Traceability with the original product descriptions

Traceability with the technical specifications

Product	Resolution	Operating System	Memory	Hard Drive	Flip-And-Fold De...	Media Reader
Lenovo - Miix 2 2-in-1 11.6"	1920 x 1080	microsoft windows 8.1 64-bit	4gb ddr3l			
Lenovo - Miix 2 2-in-1 11.6"	1920 x 1080	microsoft windows 8.1 64-bit	6gb ddr3l			
Lenovo - Miix 2 2-in-1 11.6"	1920 x 1080	microsoft windows 8.1 64-bit	6gb ddr3l	1tb		
Lenovo - Edge 15 2-in-1 15.6"	1920 x 1080	microsoft windows 8.1 64-bit	6gb ddr3l	1tb		
Lenovo - Flex 2 2-in-1 15.6"	1920 x 1080	microsoft windows 8.1 64-bit	8gb ddr3l	1tb	NO	2-in-1
Lenovo - Geek Squad Certified	1920 x 1080	microsoft windows 8.1 64-bit	8gb ddr3l	1tb hybrid	NO	2-in-1
Lenovo - Edge 15 2-in-1 15.6"	1920 x 1080	microsoft windows 8.1 64-bit	8gb ddr3l	1tb	NO	2-in-1
Lenovo - Yoga 2 2-in-1 11.6"	1366 x 768 hd	microsoft windows 8	6gb ddr3l	500gb	YES	2-in-1
Lenovo - IdeaPad Flex 2 2-in-1	1366 x 768	microsoft windows 8.1 64-bit	4gb ddr3l	500gb	NO	2-in-1
Lenovo - Geek Squad Certified	1366 x 768	microsoft windows 8.1 64-bit	4gb ddr3l	500gb	NO	2-in-1

## Textual overview

### 11.6" 10-point multitouch screen

Capacitive display responds to finger touches instead of pressure, recognizing a light swipe but not a standard stylus. IPS technology offers wide viewing angles. 1366 x 768 HD resolution. LED backlight.

### 4GB system memory for basic multitasking

Adequate high-bandwidth RAM to smoothly run multiple applications and browser tabs all at once.

**500GB hard drive** for serviceable file storage space

## Specification

Feature	Value
Hard Drive Capacity	500 gigabytes
Hard Drive Type	SATA
Hard Drive RPM	5400 revolutions per minute

<http://matrix-miner.variability.io/>

# Evaluation

**2 main datasets:** overviews dataset and specifications dataset.

We mined 2692 raw product overviews.

**9 products categories:** cover a very large spectrum of domains from BestBuy.

Products Category	#Products Overviews	#Words per Overview (Avg)
Laptops	425	350
Cell Phones	99	225
Cameras	141	279
Printers	183	277
TVs	253	283
Refrigerators	708	187
Ranges	538	275
Washing Machines	107	255
Dishwashers	238	263
Total	2692	897,020

# Empirical Results

## RQ1 Impact of Selected Products on the Synthesized PCM:

- A supervised scoping of the input products reduces the complexity (in average 107.9 of features and 1079.7 of cells) and increases the homogeneity and the compactness (only 13.0% of empty cells) of the synthesized PCMs.
- Our approach is capable of extracting numerous quantitative and comparable information (12.5% of quantified features and 15.6% of descriptive features).

## RQ2. Complementarity of Overview PCM and Specification PCM:

- A significant portion of features (49.7%) and cell values (26.2%) is recovered in the technical specifications, showing the usefulness of our approach.
- The proportion of overlap of overview PCMs regarding specification PCMs is significantly greater than the overlap of the latter regarding overview matrices.

## RQ3. Compact Overview PCM vs Compact Specification PCM:


- An automatic clustering of comparable products regarding technical specification can help to reduce the complexity of synthesized overview PCMs. However, it does not lead always to an homogeneous overview matrices (49.0% of quantified empty cells and 52.2% of descriptive empty cells). Thus, a compact specification PCM does not guarantee a compact overview PCM.

# Summary: Two Case Studies

Criteria	Case Study 1	Case Study 2
<b>Context</b>	Nuclear power plants	Manufactured products
<b>Target</b>	Modeling variability to improve certification and safety of I&C systems in different countries	Modeling variability to compare products
<b>Input</b>	Safety requirements	Informal product descriptions
<b>Output</b>	Feature Models	Product Comparison Matrices
<b>Corpus Size</b>	Huge number of requirements	Medium amount of text
<b>Number of Products</b>	[3-7] Countries	$n$ Products
<b>Automation Level</b>	Semi-Automatic	Automatic
<b>Feature</b>	Similar Requirements	Similar Terms
<b>Clustering</b>	Requirements Clustering	Terms Clustering
<b>Similarity</b>	Semantic (LSA)	Syntactical (Levenshtein)
<b>Traceability</b>	Traceability with the original regulations  Mapping with the architecture elements.	Traceability with the original product descriptions.  Traceability with the technical specifications.



# Extract Variability from Informal Product Descriptions



The image is a cartoon illustration of a man in a blue shirt and yellow tie, holding a baseball bat, standing next to a computer monitor. He is surrounded by a chaotic pile of papers, each containing a product description and a small image of a product. The papers are scattered around him, with some labeled "Features", "Values", "Alternatives", "Commonalities", and "Difference". There are also several large question marks scattered among the papers. The man has a surprised or excited expression on his face.

**CONNECTION Project : Overview of the nuclear regulatory landscape**

**Questions**

The diagram illustrates the regulatory landscape for the CONNECTION project. It shows the relationship between various entities and the project's regulatory guidance.

**Regulators:** ASN, IRSN, BPS.

**International organizations:** IEC, IAEA, ISO, INFC, IAEA.

**Regulatory guidance:** Interprets and recommends, Details.

**candidate project:** Relies on.

**Safety Category & ICFunction:** Safety Category <= Safety Class, ICFunction [1..\*].

**Independence.Sys:** Independence.Sys.Diff.Classes, Prevent System Degradation, Independence.Redundant.Parts.

**IEC & IRSN:** IEC, IRSN.

**Artise:** Artise.

# Questions

