Intelligent System
Logging and Monitoring
ISyLM Research Lab

Using AI to Empower System Development and Operations Team

Wahab Hamou-Lhadj, PhD

ECE, Concordia University

wahab.hamou-lhadj@concordia.ca

TNSBC, Montréal, QC

May 23, 2018

Software-intensive systems are everywhere!

- Health
- Energy
- Finance
- Manufacturing
- Education
- Public safety
- Telecom
- Aerospace
- Entertainment
- Hospitality
- Public administration
- Social interactions

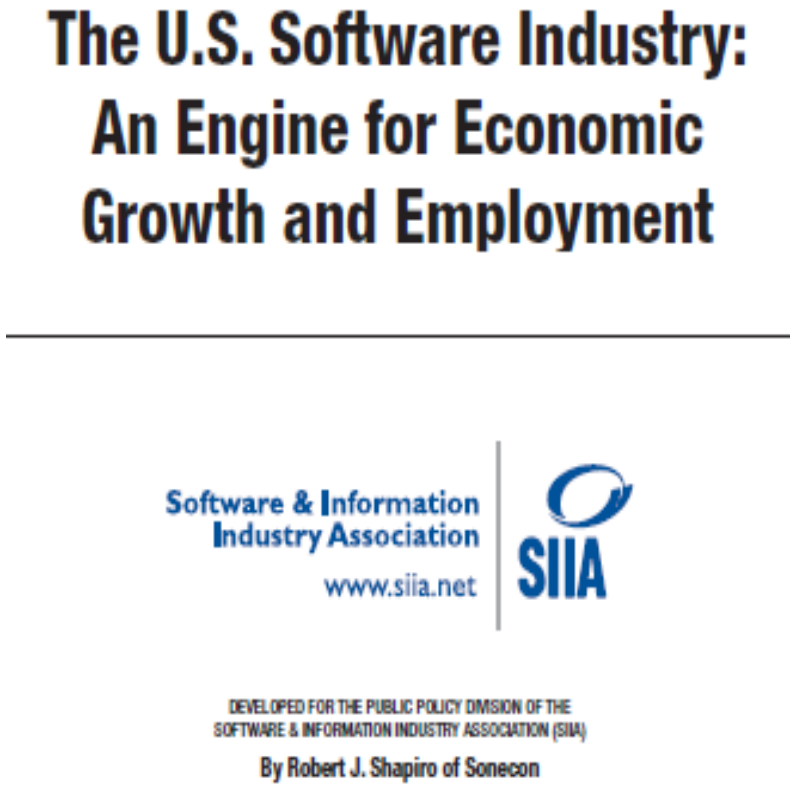
"Our civilization runs on software"

B. Stroustrup

Dr. Wahab Hamou-Lhadj (wahab.hamou-lhadj@concordia.ca)

Facts

- From 1997 to 2012, software industry production grew from $149 billion to $425 billion.
- The software industry's direct share of U.S. GDP went from 1.7% to 2.6%.
- Software accounted for 12.1% percent of all U.S. labor productivity gains from 1995 to 2004 and 15.4% from 2004 to 2012.



Dr. Wahab Hamou-Lhadj (wahab.hamou-lhadj@concordia.ca)

Software Development

A set of activities for creating a software system including requirements analysis, architectural design, detailed design, coding, testing, maintenance, integration, acceptance testing, etc.

Software Operations

A set of activities for supporting end users of a software product in an operational environment. Typical activities include: installation, upgrade, monitoring, configuration, etc.

Source: Definitions adapted from ISO/IEC24748-1 2011, ISO/IEC15288 2008, ISO/IEC12207 2008).

Dr. Wahab Hamou-Lhadj (wahab.hamou-lhadj@concordia.ca)

Software Development

A set of activities for creating a software system including requirements analysis, architectural design, detailed design, coding, testing, maintenance, integration, acceptance testing, etc.

Software Operations

A set of activities for supporting end users of a software product in an operational environment. Typical activities include: installation, upgrade, monitoring, configuration, etc.

Source: Definitions adapted from ISO/IEC24748-1 2011, ISO/IEC15288 2008, ISO/IEC12207 2008).

Dr. Wahab Hamou-Lhadj (wahab.hamou-lhadj@concordia.ca)

SW Development Challenges

- Increased complexity
- High cost
- Heavy reliance on people
- Lack of automated tools
- Time to market pressure
- Maintaining quality

Dr. Wahab Hamou-Lhadj (wahab.hamou-lhadj@concordia.ca)

A NIST study* shows that defects in software cost the U.S. economy **$56 billion annually**.

A large percentage of software development costs are **spent on identifying and correcting defects.**

There is a need to **invest in automated and intelligent solutions**.

*Source: Research Triangle Institute, *The Economic Impacts of Inadequate Infrastructure for Software Testing*, NIST Planning Report 02-3, May 2002.

Dr. Wahab Hamou-Lhadj (wahab.hamou-lhadj@concordia.ca)

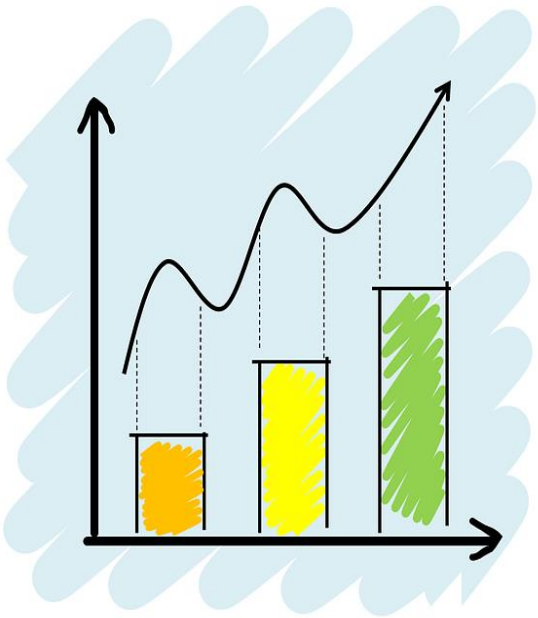
Active Research Community

- Change and defect management
- Continuous integration/deployment
- Release engineering
- Reverse engineering and re-engineering
- Run-time evolution and dynamic configuration
- Software and system comprehension
- Software migration and renovation
- Software refactoring and restructuring
- Empirical studies
- Evolution of non-code artefacts
- Human aspects of software maintenance and evolution
- Maintenance and evolution of model-based methods
- Maintenance and evolution processes
- Maintenance and evolution of mobile apps
- Maintenance versus release process
- Mining software repositories
- Etc.

Dr. Wahab Hamou-Lhadj (wahab.hamou-lhadj@concordia.ca)

Emergence of Software Analytics

- Data-driven software development and maintenance
- Big Data: source code, bug reports, test cases, logs, user feedback, etc.
- Predictive analytics using ML, DL, CI, and PR
- Information visualization of large-scale data



Dr. Wahab Hamou-Lhadj (wahab.hamou-lhadj@concordia.ca)

The Commit Assistant Project

- An NSERC project in collaboration with Ubisoft.
- Goal: To empower SW developers with an intelligent tool that detects defects as they write code, and proposes fixes.

Dr. Wahab Hamou-Lhadj (wahab.hamou-lhadj@concordia.ca)

Open Technologies behind CommitAssistant

Bug Metarepository Search Engine for Developers and Reseachers

BIANCA

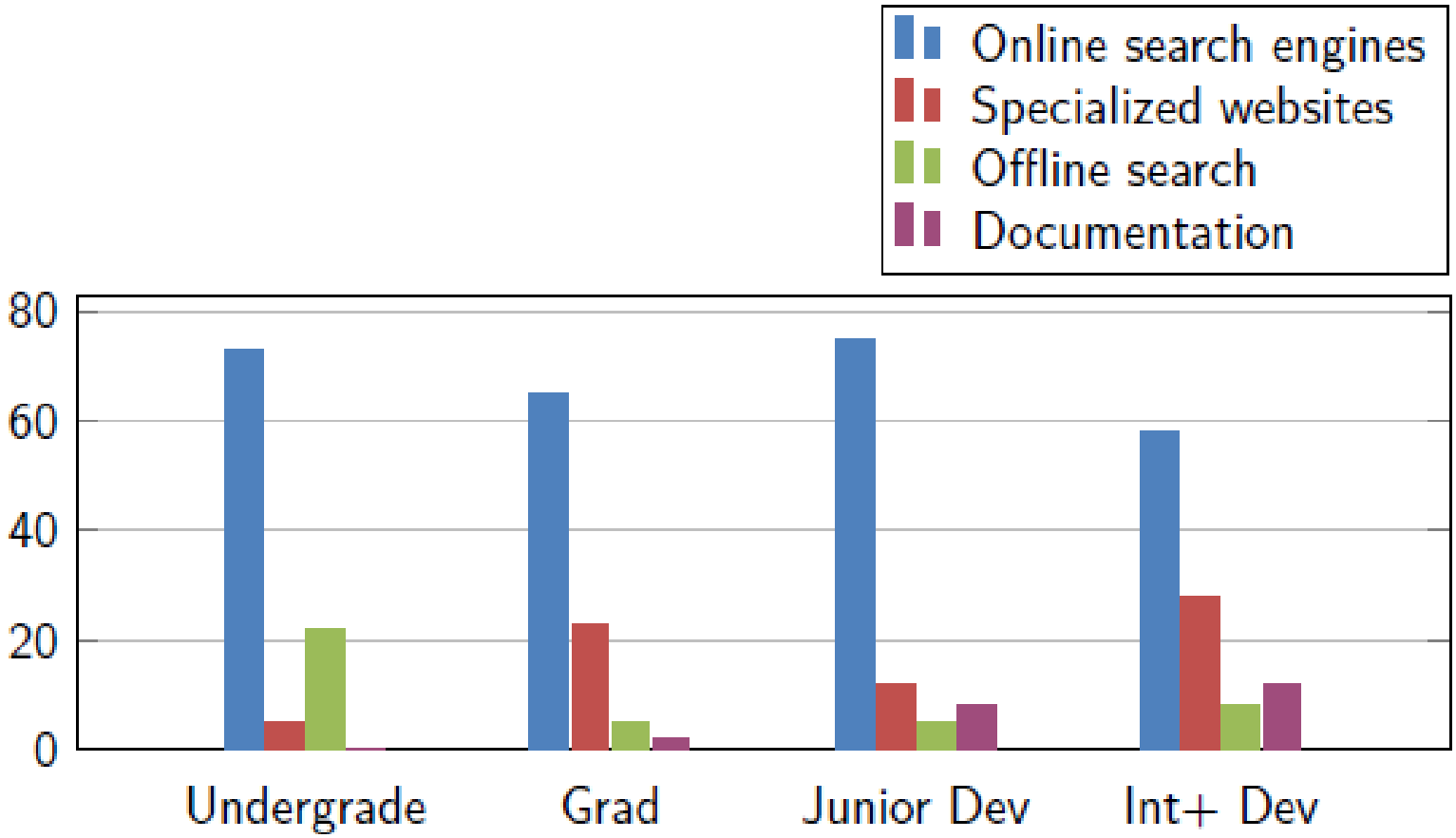
Preventing Bug Insertion at Commit-Time Using Clone Detection

CLEVER

Combining Levels of Bug Prevention and Resolution Techniques

Dr. Wahab Hamou-Lhadj (wahab.hamou-lhadj@concordia.ca)

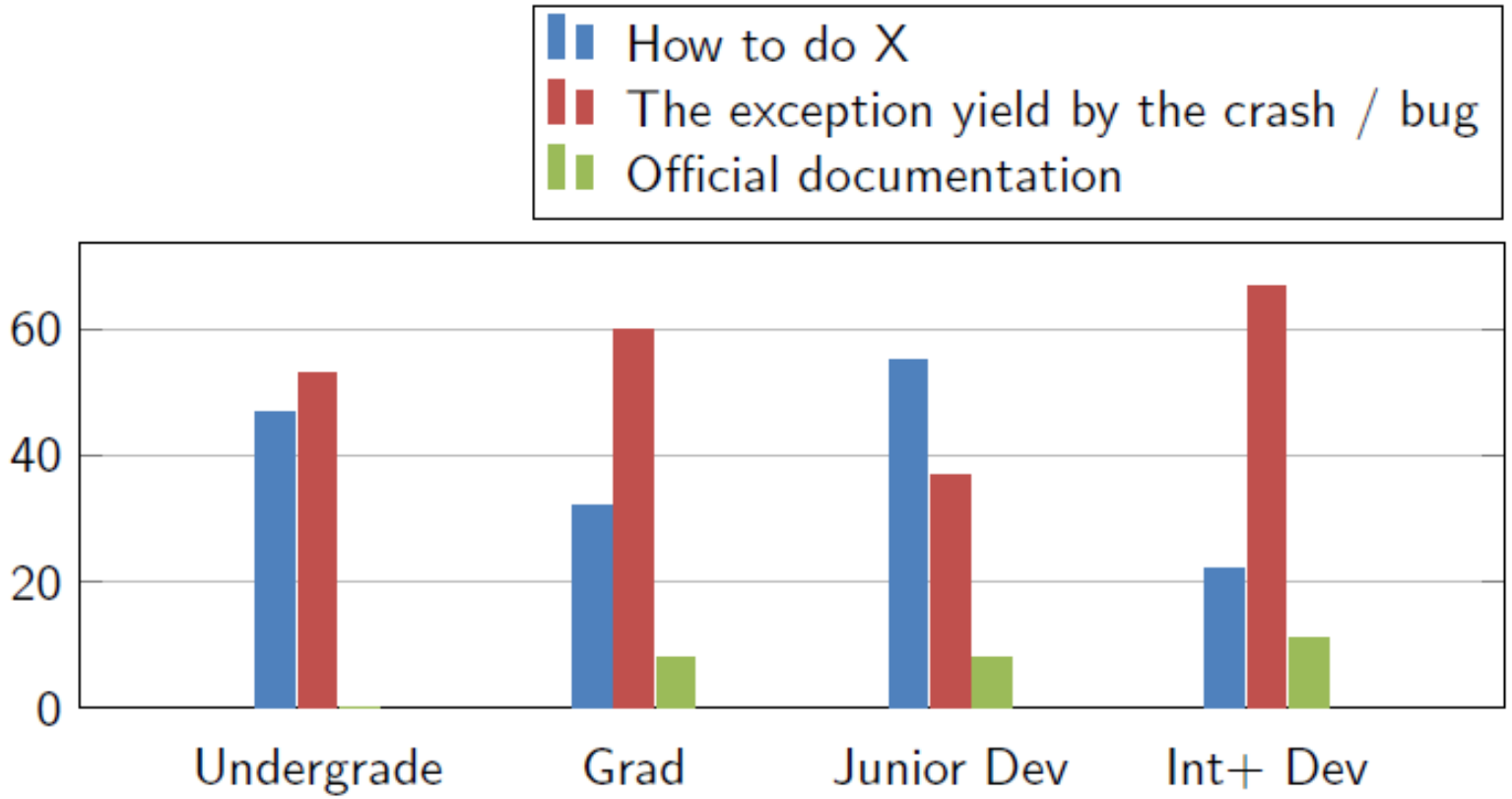
Where do developers look for information when facing an unknown bug/crash?



89 participants

Dr. Wahab Hamou-Lhadj (wahab.hamou-lhadj@concordia.ca)

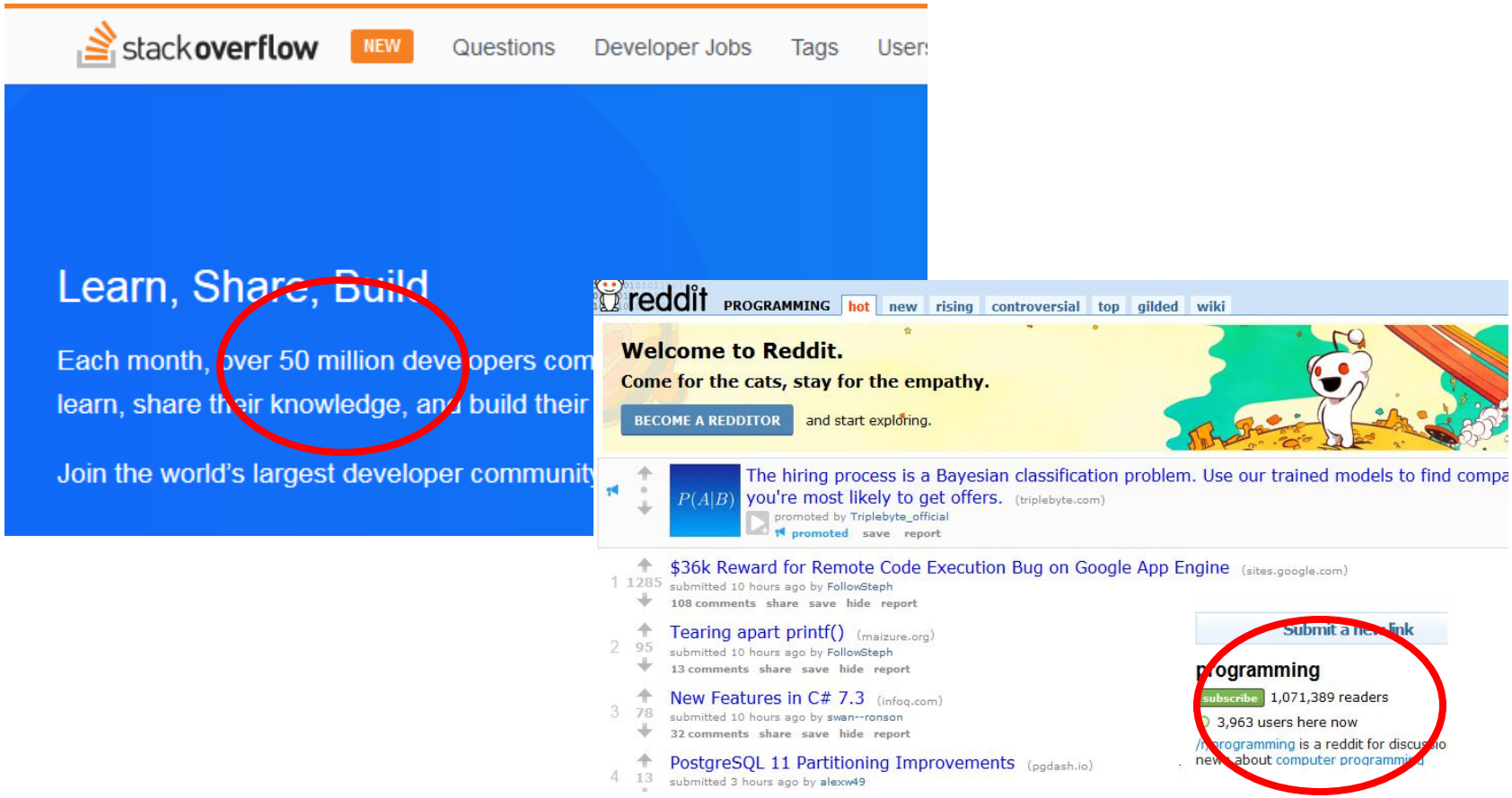
What do developers search for when facing an unknown bug/crash?



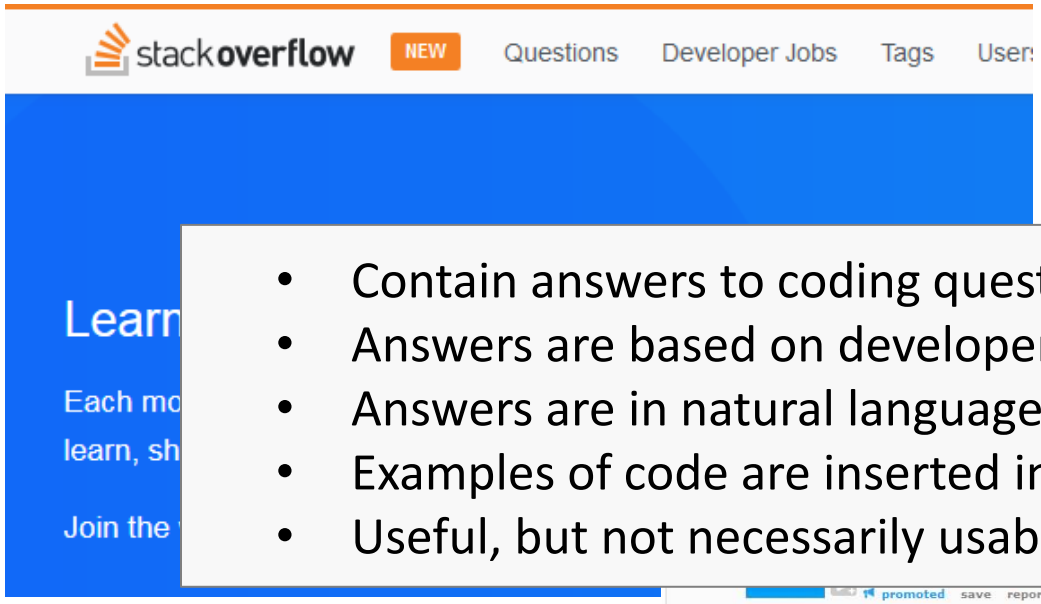
89 participants

Dr. Wahab Hamou-Lhadj (wahab.hamou-lhadj@concordia.ca)

Examples of Coding Websites



Examples of Coding Websites



- Contain answers to coding questions
- Answers are based on developers' experience
- Answers are in natural language
- Examples of code are inserted into text
- Useful, but not necessarily usable

Dr. Wahab Hamou-Lhadj (wahab.hamou-lhadj@concordia.ca)

BUMPER: Bug Metarepository Search Engine for Developers and Researchers

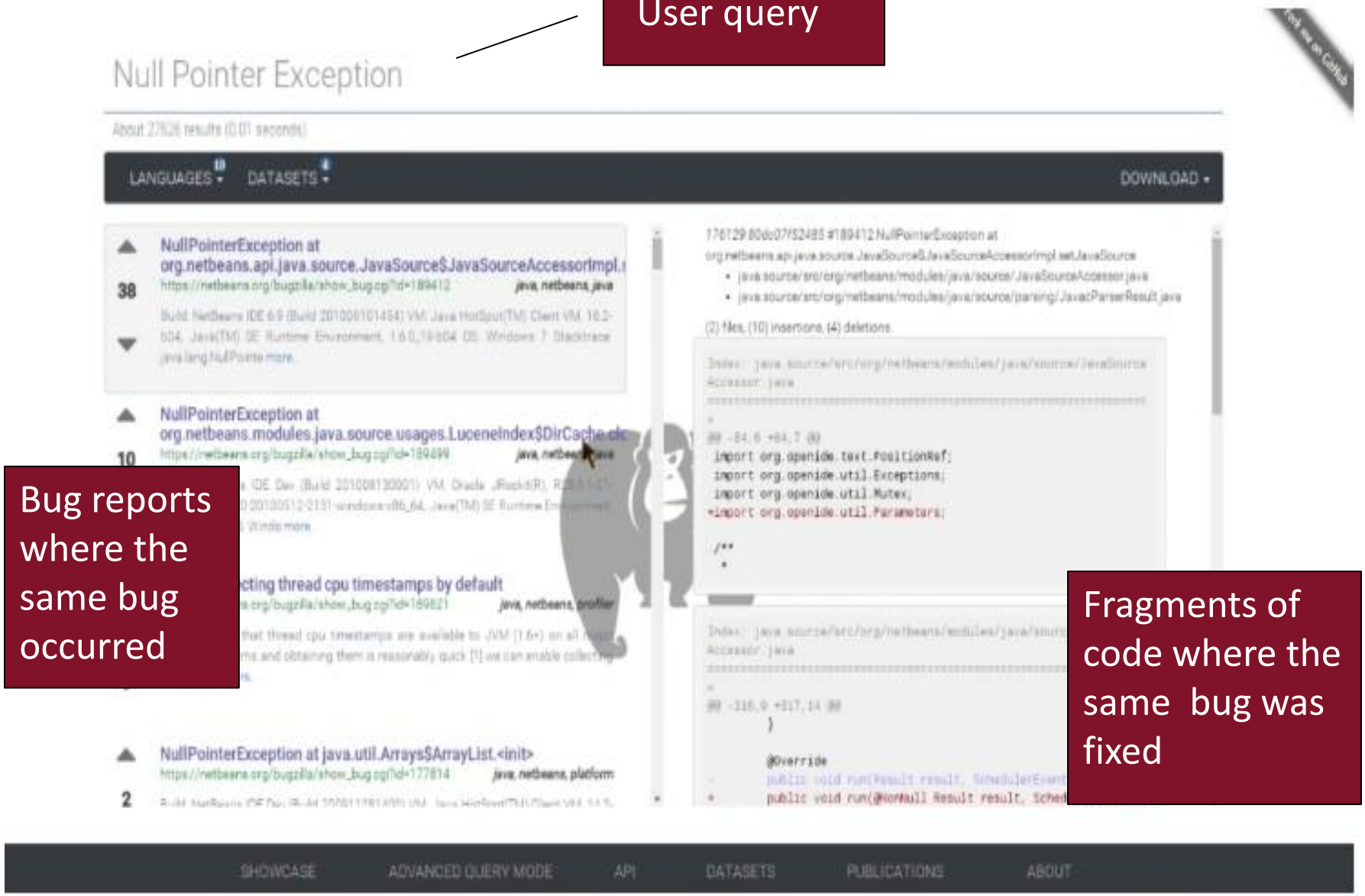
- Aggregates information from many bug report and code versioning systems
- Is an online search engine to millions of bug reports and fixes from open-source repositories
- Uses a query system for developers and advanced API for researchers
- Leverages the concept of collective coding → collective intelligence

Dr. Wahab Hamou-Lhadj (wahab.hamou-lhadj@concordia.ca)

Developers can search millions of lines of code and bug reports for a bug or crash they encountered.

Dr. Wahab Hamou-Lhadj (wahab.hamou-lhadj@concordia.ca)

User query



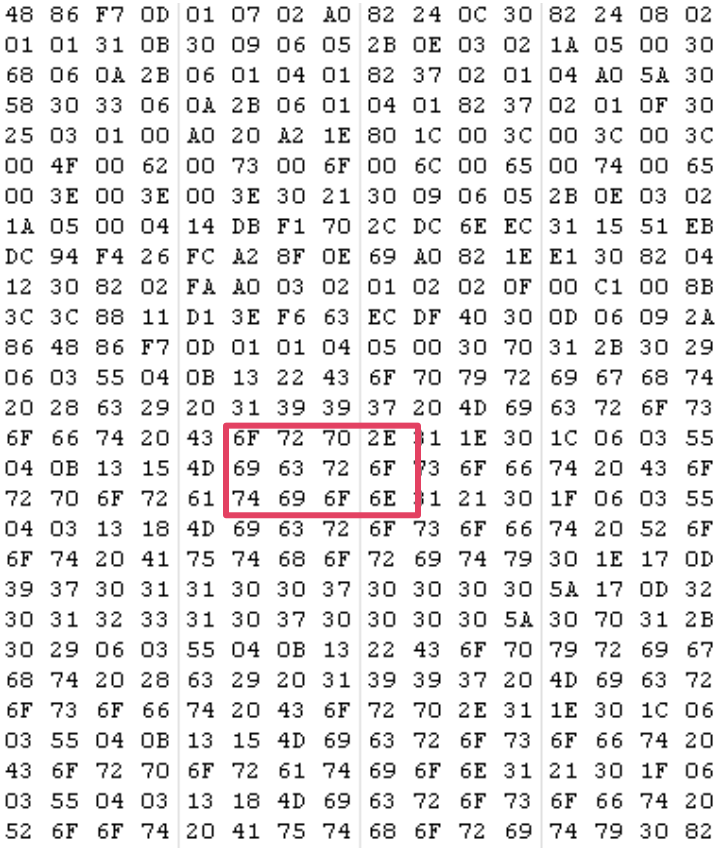
Bug reports where the same bug occurred

Fragments of code where the same bug was fixed

Dr. Wahab Hamou-Lhadj (wahab.hamou-lhadj@concordia.ca)

BIANCA: Preventing Bug Insertion at Commit-Time Using Clone Detection

- BIANCA learns known defects by mining BUMPER-indexed systems.
- It intercepts developer's code and compares it to signatures of known defects.
- If a match exists, a flag is raised and a fix is proposed.



Dr. Wahab Hamou-Lhadj (wahab.hamou-lhadj@concordia.ca)

TABLE 3: BIANCA results in terms of organization, project name, a short description, number of class, number of commits, number of defect introducing commits, number of risky commit detected, precision (%), recall (%), F_1-measure (%), the average similarity of first 3 and 5 proposed fixes with the actual fix and the average time difference between detected and original.

Organization	Project Name	Short Description	NoC	#Commits	Bug Introducing Commit	Detected	Precision	Recall	F_1	Top 5 Fixes Similarity	Top 3 Fixes Similarity
Alibaba	druid	Database connection pool	3,309	4,775	1,260	787	88.44	62.46	73.21	39.97	46.69
	dubbo	RPC framework	1,715	1,836	119	61	96.72	51.26	67.01	60.01	57.14
	fastjson	JSON parser/ generator	2,002	1,749	516	373	95.71	72.29	82.37	18.19	15.23
	jstorm	Stream Process	1,492	215	24	21	90.48	87.50	88.96	22.38	30.48
Apache	hadoop	Distributed processing	9,108	14,154	3,678	851	86.84	23.14	36.54	38.94	47.68
	storm	Realtime system	2,209	7,208	951	444	86.26	46.69	60.58	53.03	61.10
Clojure	clojure	Programming language	335	2,996	596	46	86.96	7.72	14.18	53.61	59.52
Dropwizard	dropwizard	RESTful web services	964	3,809	581	179	96.65	30.81	46.72	47.54	53.56
	metrics	JVM metrics	335	1,948	331	129	95.35	38.97	55.33	22.53	31.82
Eclipse	che	Eclipse IDE	7,818	1,826	169	9	88.89	5.33	10.05	31.01	39.04
E											
F											
G											
G											
G											
Ja											
Jf											
L											
N											
O											
OpenZipkin	zipkin	Distributed tracing system	397	799	176	73	87.67	41.48	56.31	33.92	31.90
Orfjackal	retrolambda	Backport of Java 8's lambda	171	447	97	35	94.29	36.08	52.19	34.69	42.06
OrientTechnologie	orientdb	Multi-Model DBMS	2,907	13,907	7,441	2,894	86.77	38.89	53.71	62.20	70.00
Perwendel	spark	Sinatra for java	205	703	125	82	97.56	65.60	78.45	21.88	28.00
PrestoDb	presto	Distributed SQL query	4,381	8,065	2,112	991	90.62	46.92	61.83	23.34	20.64
RoboGuice	roboguice	Google Guice on Android	1,193	1,053	229	70	91.43	30.57	45.82	53.81	56.55
Lombok	lombok	Additions to the Java language	1,146	1,872	560	212	91.98	37.86	53.64	58.94	57.49
Scribejava	scribejava	OAuth library	218	609	72	16	93.75	22.22	35.93	30.05	38.16
Square	dagger	Dependency injector	232	697	144	84	90.48	58.33	70.93	64.29	64.97
	javapoet	Java API	66	650	163	113	100.00	69.33	81.88	51.04	53.20
	okhttp	HTTP+HTTP/2 client	344	2,649	592	474	93.04	80.07	86.07	29.09	24.91
	okio	I/O API for Java	90	433	40	24	100.00	60.00	75.00	31.51	35.50
	otto	Guava-based event bus	84	201	15	15	93.33	100.00	96.55	54.11	49.94
	retrofit	Type-safe HTTP client	202	1,349	151	111	99.10	73.51	84.41	49.88	45.46
StephaneNicolas	robospice	Android library	461	865	113	39	87.18	34.51	49.45	60.90	65.04
ThinkAurelius	titan	Graph Database	2,015	4,434	1,634	527	90.13	32.25	47.51	48.64	50.59
Xetorthio	jedis	Redis client	203	1,370	295	226	92.04	76.61	83.62	25.69	29.45
Yahoo	anthelion	Plugin for Apache Nutch	1,620	7	0	-	-	-	-	-	-
Zxing	zxing	1D/2D barcode image	3,030	3,253	791	123	94.31	15.55	26.70	29.35	37.96
Total			**96,003**	**165,912**	**41,225**	**15316**	**90.75**	**37.15**	**52.72**	**40.78**	**44.17**

- **Subject systems: 42 open source projects**
- **Precision = 90% and Recall: 37%**
- **BIANCA fixes are accurate in 79% of the cases**

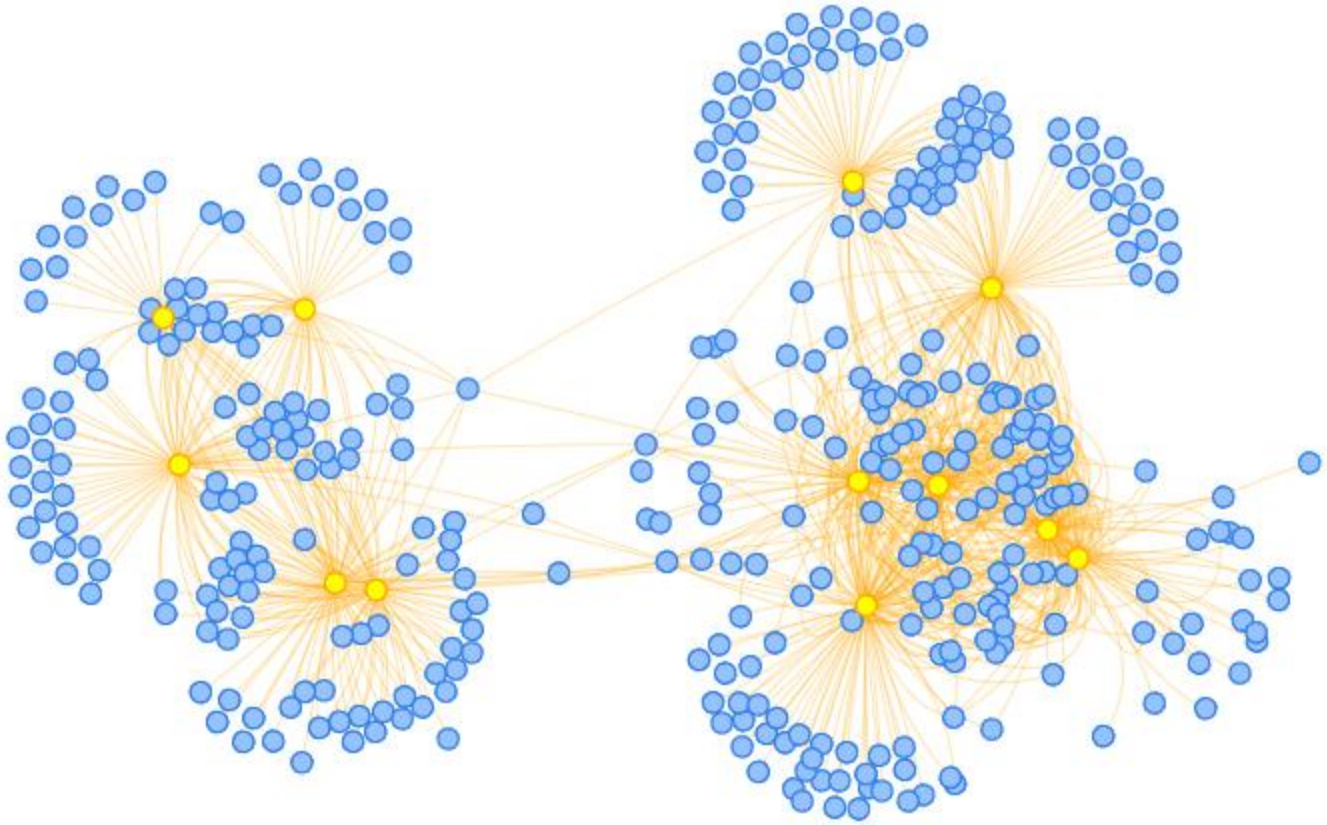
Dr. Wahab Hamou-Lhadj (wahab.hamou-lhadj@concordia.ca)

CLEVER: Combining Levels of Bug Prevention and Resolution Techniques

- Combines multiple features to determine the defect signatures
- Uses domain expertise to create clusters of projects for improved accuracy
- Uses better code matching techniques
- Is evaluated on 12 Ubisoft systems

Dr. Wahab Hamou-Lhadj (wahab.hamou-lhadj@concordia.ca)

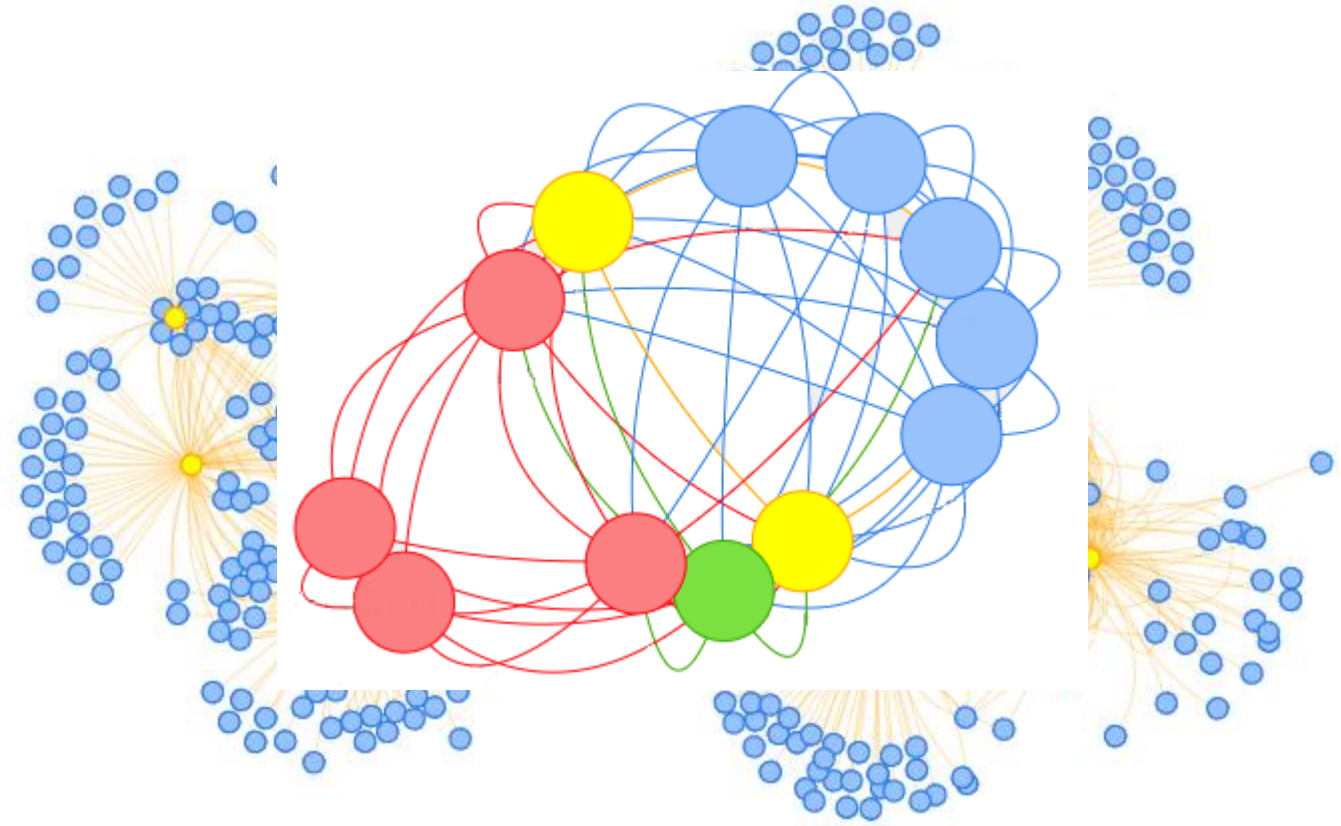
CLEVER Project Clustering



We can improve the detection accuracy if we search within inter-related projects

Dr. Wahab Hamou-Lhadj (wahab.hamou-lhadj@concordia.ca)

CLEVER Project Clustering (Cont.)



We can improve the detection accuracy if we search within inter-related projects

Dr. Wahab Hamou-Lhadj (wahab.hamou-lhadj@concordia.ca)

Evaluation of CLEVER at Ubisoft

- **Results**:
  - Subject systems: 12 Ubisoft systems
  - Precision = 79% and Recall = 65%
  - CLEVER recommends fixes in 67% of the cases
- **Impact on productivity:**
  - CommitAssistant (internal implementation of CLEVER) is designed to integrate well with developers' workflow
  - Ubisoft announced in a press release that CommitAssistant can cut the bug fixing time by 20%

Dr. Wahab Hamou-Lhadj (wahab.hamou-lhadj@concordia.ca)

Software Development

A set of activities for creating a software system including requirements analysis, architectural design, detailed design, coding, testing, maintenance, integration, acceptance testing, etc.

Software Operations

A set of activities for supporting end users of a software product in an operational environment. Typical activities include: installation, upgrade, monitoring, configuration, etc.

Source: Definitions adapted from ISO/IEC24748-1 2011, ISO/IEC15288 2008, ISO/IEC12207 2008).

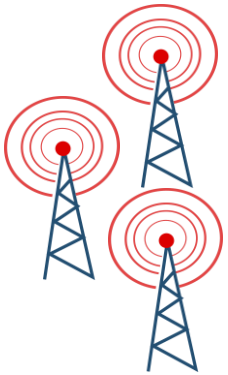
Dr. Wahab Hamou-Lhadj (wahab.hamou-lhadj@concordia.ca)

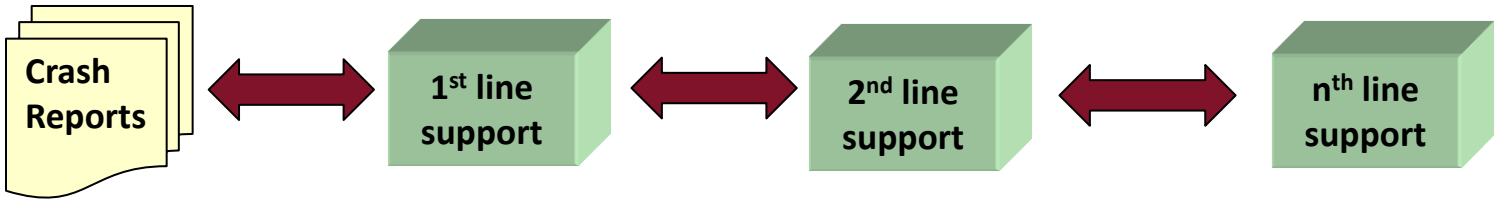
Software Operations

- D2K Project: From Data To Knowledge for Better System Maintenance
- Collaborators: Ericsson, NSERC, and MITACS
- Objectives:
  - Improve the crash report (CR) handling process
  - Investigate automated solutions
  - Provide analysis capabilities to operators
  - Provide data governance solutions

Dr. Wahab Hamou-Lhadj (wahab.hamou-lhadj@concordia.ca)

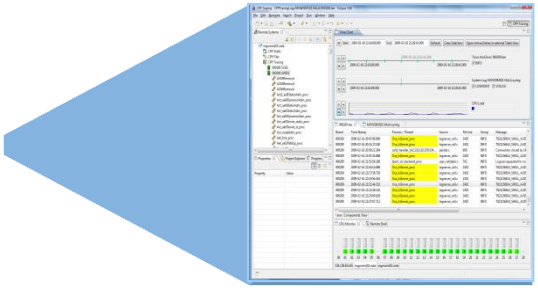
Problem





Operators/Analysts

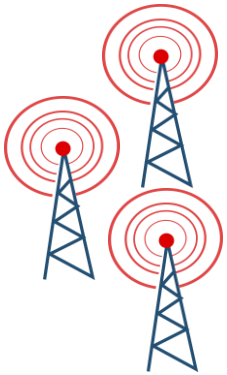
- A large number of CRs
- Meeting contractual WLAs
- Severe problems have to be fixed quickly
- Cost is also an issue



Log data
Dumps
Configuration Files

Dr. Wahab Hamou-Lhadj (wahab.hamou-lhadj@concordia.ca)

Solution



Crash Reports

1st line support

2nd line support

nth line support

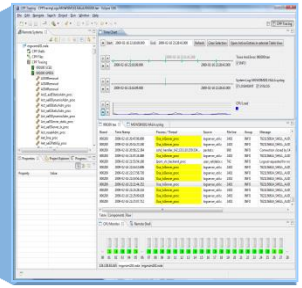
Operators/Analysts

- ML model of CRs
- Domain knowledge

D2K Database

Crash report

Log data
Dumps
Configuration Files

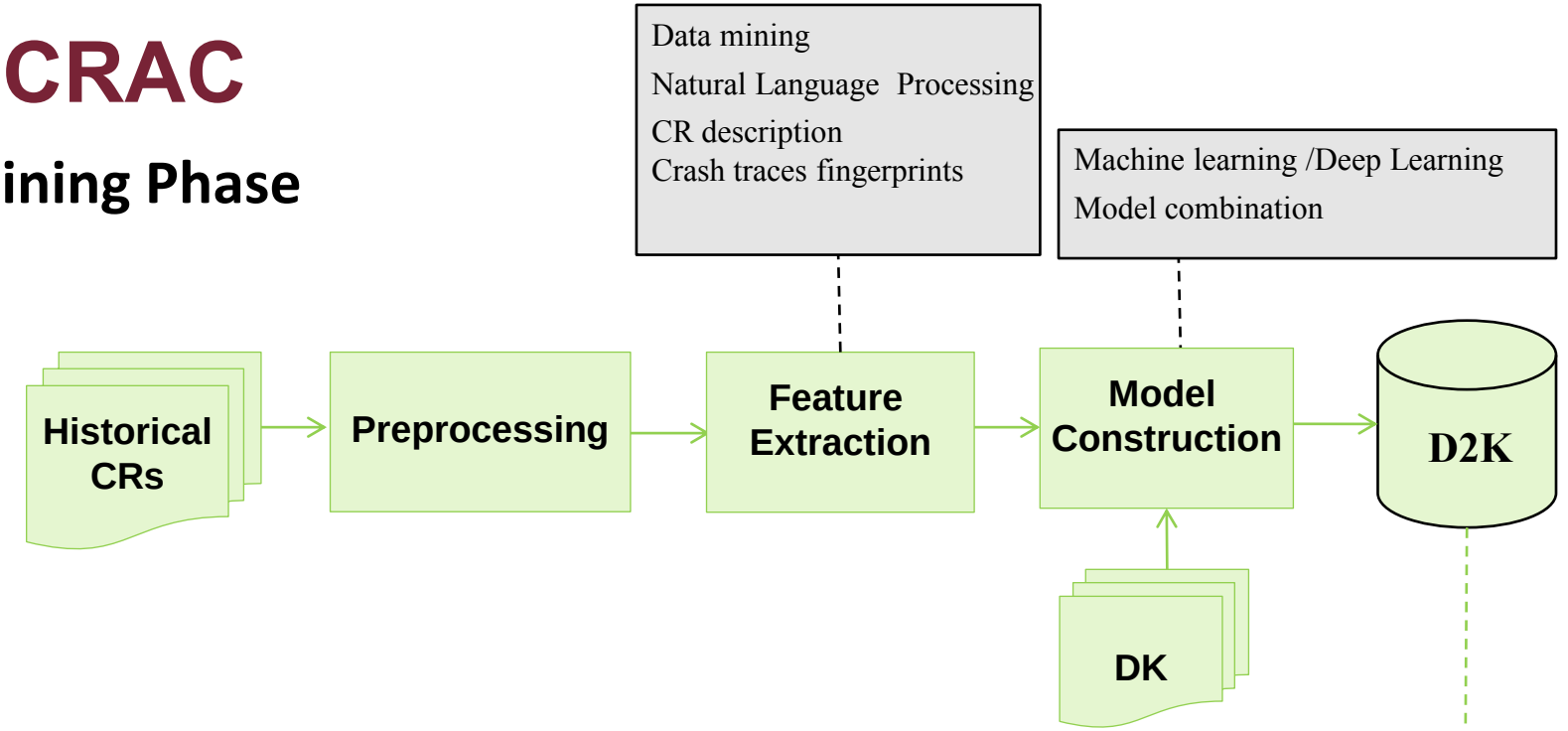


- **ReCRAC: CR Duplicate Detection Tool**
- **TotalADS: Anomaly Detection System**

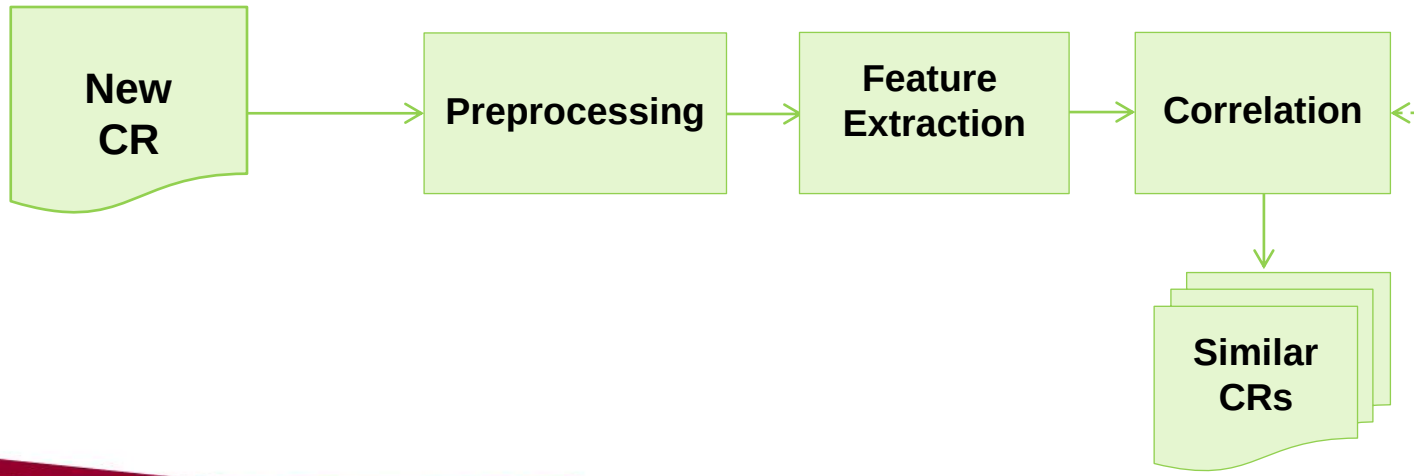
Dr. Wahab Hamou-Lhadj (wahab.hamou-lhadj@concordia.ca)

ReCRAC

Training Phase



Testing Phase (ReCRAC in operation)



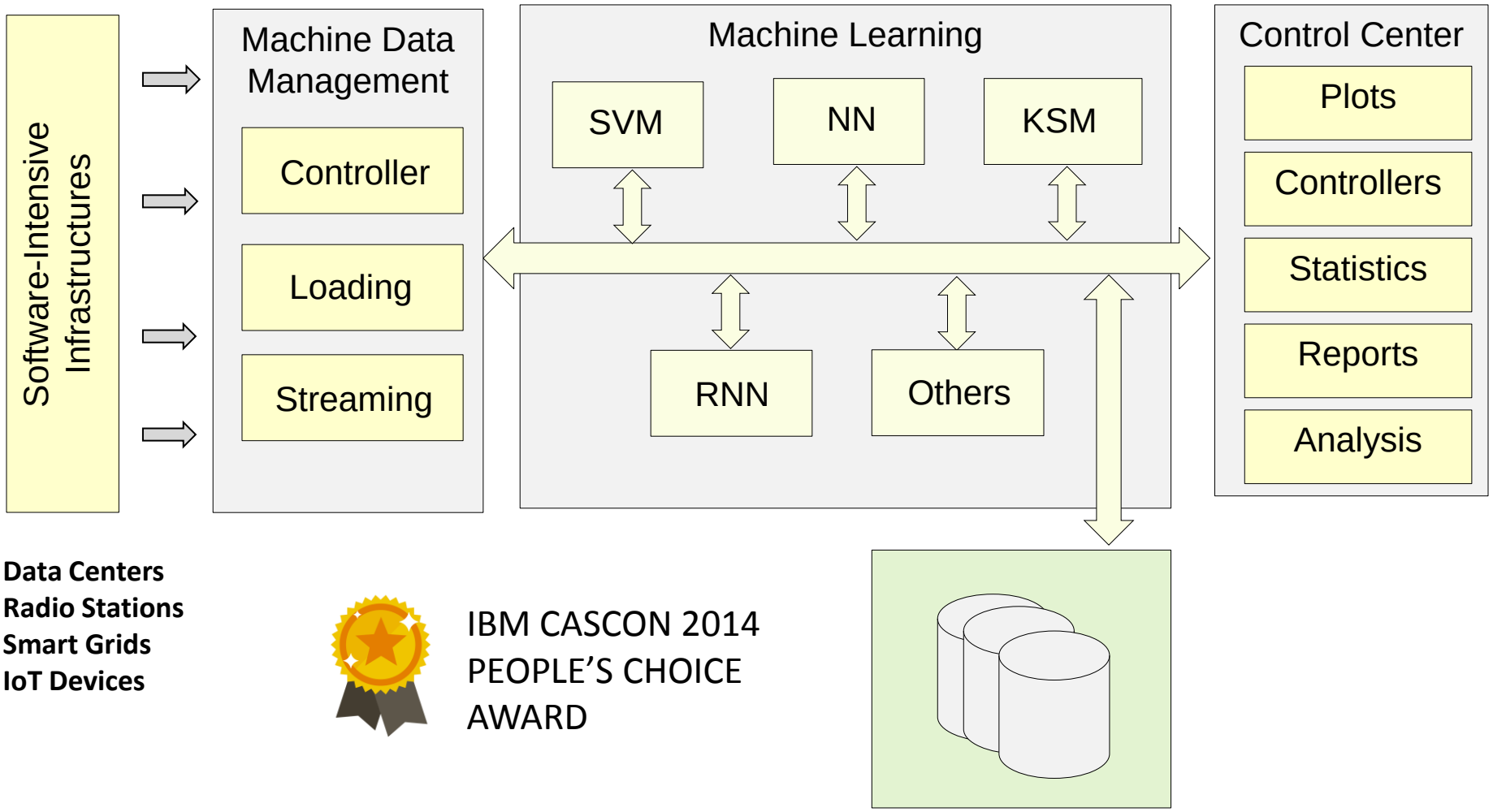
Dr. Wahab Hamou-Lhadj (wahab.hamou-lhadj@concordia.ca)

TotalADS: Total Anomaly Detection System

- Developed in an NSERC project with Defence R&D Canada and Ericsson
- Objectives:
  - Detection of abnormal behavior in computer hosts through the analysis of machine data
  - Combination of multiple machine learning techniques
  - Leverage of data abstraction, model combination, adaptive learning, and online learning
  - Tool development and integration

Dr. Wahab Hamou-Lhadj (wahab.hamou-lhadj@concordia.ca)

TotalADS: Total Anomaly Detection System Architecture



Data Centers
Radio Stations
Smart Grids
IoT Devices

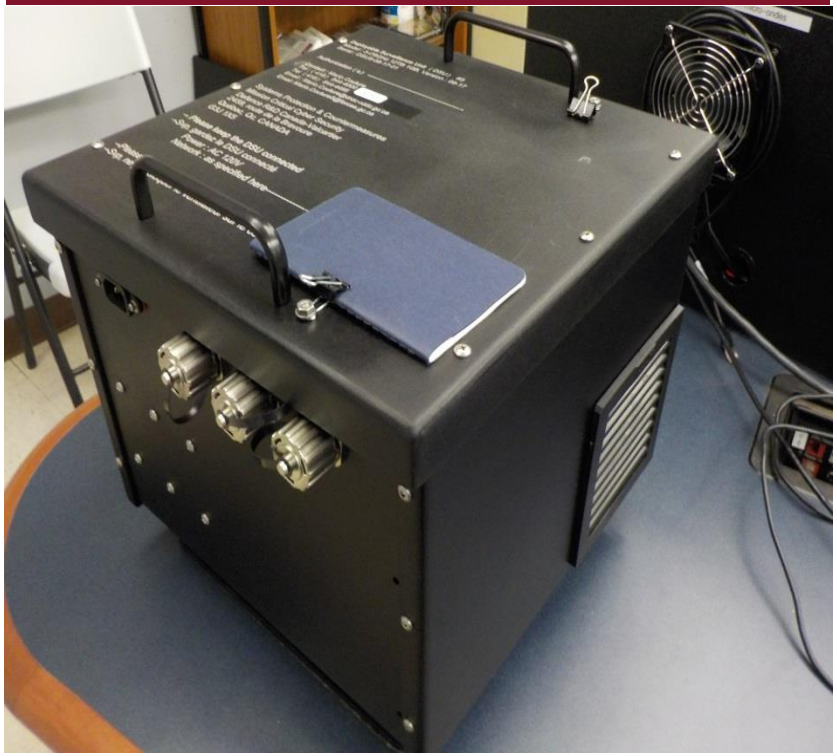
IBM CASCON 2014
PEOPLE'S CHOICE
AWARD

Dr. Wahab Hamou-Lhadj (wahab.hamou-lhadj@concordia.ca)

TotalADS and Deployable System Units (DSUs)

The 4th DSU prototype (PoC1)

The 5th DSU prototype (PoC2)



Six Jetway industrial Mini-ITX computers + one manageable GB switch + six 4-TB hard disks
(Intel's Haswell Core i7-4770TE 2.3 GHz processor, 8 GiB DDR3, 6 GB/sec mSATA, dual LAN)
(The whole DSU needs less than 350 watts when used at full capacity)

(Next technology to be considered: the new NVIDIA Jetson TX-2 AI computing board)

Couture, M., Fattahi, J., The use of Deployable Surveillance Units (DSUs) for online cyber surveillance—Proof of concept, Phase 1. Defence Research and Development Canada (DRDC), Report number: DRDC-RDDC-2018-R021, DRDC Valcartier Research Center, April 2018, Unclass.

Some thoughts on the use of DL/ML in SW Development and Operations


Powerful tool suite


Healthy analytics


Context matters


Domain expertise


Education


Impact on society

THANK YOU!

CONCORDIA.CA