# Introduction to Software Analytics

Dongmei Zhang Software Analytics Group Microsoft Research

November 26, 2014

# Outline

- Overview of Software Analytics
- Selected projects
- Experience sharing on Software Analytics in practice

# New Era...Software itself is changing...





#### facebook









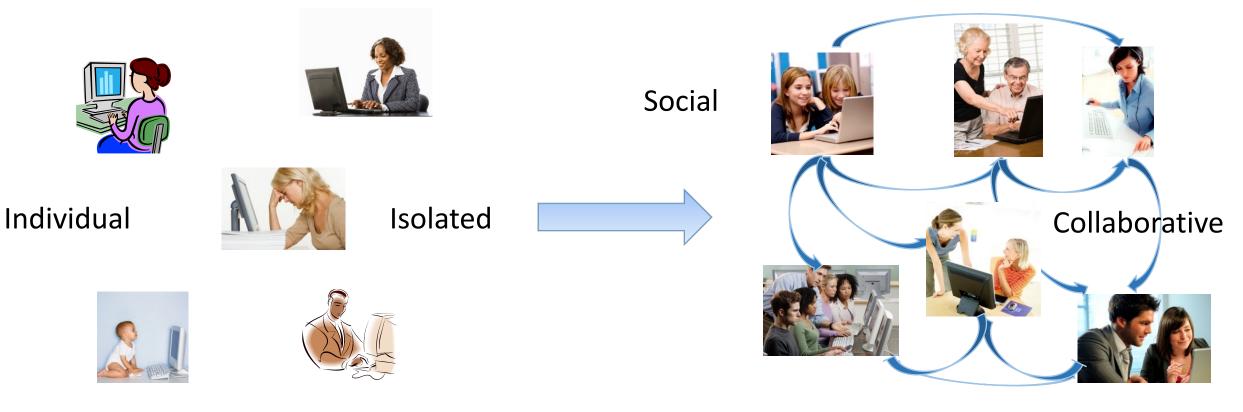
amazon



Software

Services

### How people use software is changing...



Not much content generation

Huge amount of artifacts generated anywhere anytime

### How software is built & operated is changing...

Code centric

In-lab testing

Experience & gut-feeling

**Centralized development** 

Long product cycle

Data pervasive

Debugging in the large

Informed decision making

**Distributed development** 

Continuous release

. . .

# Software Analytics

Software analytics is to enable software practitioners to perform data exploration and analysis in order to obtain insightful and actionable information for datadriven tasks around software and services.

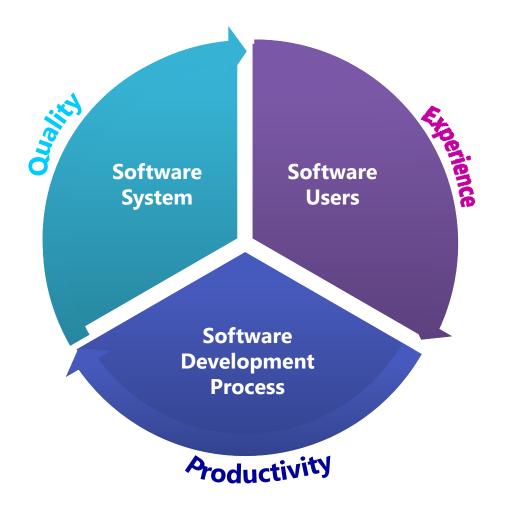
# Software Analytics

Software analytics is to enable *software practitioners* to perform data exploration and analysis in order to obtain *insightful and actionable information* for *data-driven tasks* around software and services.

### Five dimensions



# Research topics



- Covering different areas of software domain
- Throughout entire development cycle
- Enabling practitioners to obtain insights

### Data sources

0	

Runtime traces Program logs System events Perf counters



Usage log User surveys Online forum posts Blog & Twitter



Source code Bug history Check-in history Test cases

10

# Target audience – software practitioners





Management personnel



Designer



Tester



### Support engineer



**Operation engineer** 



**Usability engineer** 



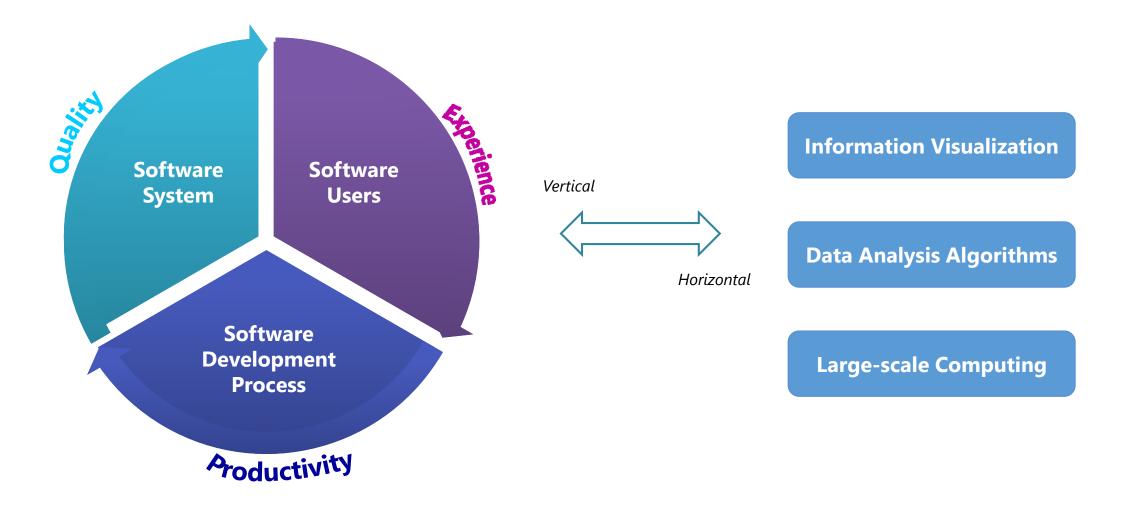
# Output – insightful information

- Conveys meaningful and useful understanding or knowledge towards completing the target task
- Not easily attainable via directly investigating raw data without aid of analytics technologies
- Examples
  - It is easy to count the number of re-opened bugs, but how to find out the primary reasons for these re-opened bugs?
  - When the availability of an online service drops below a threshold, how to localize the problem?

# Output – actionable information

- Enables software practitioners to come up with concrete solutions towards completing the target task
- Examples
  - Why bugs were re-opened?
    - A list of bug groups each with the same reason of re-opening
  - Why availability of online services dropped?
    - A list of problematic areas with associated confidence values
  - Which part of my code should be refactored?
    - A list of cloned code snippets easily explored from different perspectives

# Research topics and technology pillars

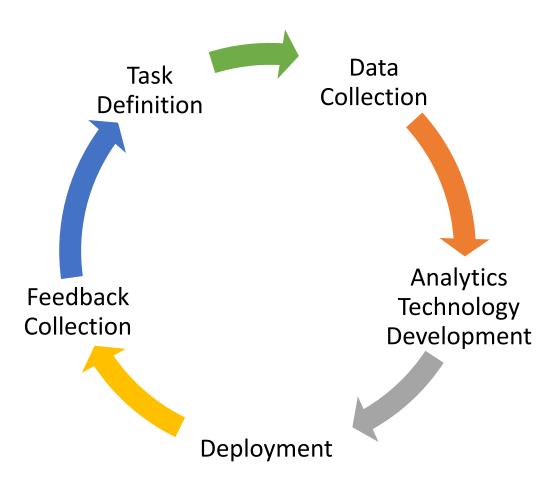


# Connection to practice

- Software Analytics is naturally tied with software development practice
- Getting real



# Approach



# Various related efforts...

- Mining Software Repositories (MSR)
- Software Intelligence
- Software Development Analytics



http://www.msrconf.org/

A. E. Hassan and T. Xie. Software intelligence: Future of mining software engineering data. In Proc. FSE/SDP Workshop on Future of Software Engineering Research (FoSER 2010), pages 161–166, 2010.

R. P. Buse and T. Zimmermann. Analytics for software development. In Proc. FSE/SDP Workshop on Future of Software Engineering Research (FoSER 2010), pages 77–80, 2010.

# Outline

- Overview of Software Analytics
- Selected projects
- Experience sharing on Software Analytics in practice

# Selected projects

### **XIAO** Scalable code clone analysis



StackMine – Performance debugging in the large via mining millions of stack traces



Service Analysis Studio: Incident management for online services

# XIAO Scalable code clone analysis

Yingnong Dang, Dongmei Zhang, Song Ge, Chengyun Chu, Yingjun Qiu, Tao Xie, XIAO: Tuning Code Clones at Hands of Engineers in Practice, in Proceedings of Annual Computer Security Applications Conference 2012, (ACSAC 2012), Orlando, Florida, USA, December, 2012.

# Code clone research

- Tons of papers published in the past decade
- 8 years of International Workshop on Software Clones (<u>IWSC</u>) since 2006



Source: http://www.dagstuhl.de/12071

- Dagstuhl Seminar
  - <u>Software Clone Management towards Industrial Application (2012)</u>
  - <u>Duplication, Redundancy, and Similarity in Software</u> (2006)

# XIAO: Code clone analysis

- Motivation
  - Copy-and-paste is a common developer behavior
  - A real tool widely adopted internally and externally
- XIAO enables code clone analysis in the following way
  - High tunability
  - High scalability
  - High compatibility
  - High explorability



#### Code Clone Detection Experience at Microsoft

Yingnong Dang, Song Ge, Ray Huang and Dongmei Zhang Microsoft Research Asia yidang;songge;rayhuang;dongmeiz@microsoft.com

#### ABSTRACT

Chaning source code is a common practice in the software development process. In general, the number of code closes increases in proportion to the growth of the code base. It is challenging to proscively keep closes consistent and remove unnecessary clones during the entire software development process of large-scale commercial obstrue. In this position paper, we briefly thate some typical usage scenarios of code close discuss our experimence on building XIAO. a code clone detection tool, and the feedback we have received from Microsoft engineers on using XIAO are ald development settings. Fix Bugs Once If a bug is identified in a piece of code with duplicated copies, it is desirable to have the ability to fix all of them at once. This scenario is beneficial to multiple stages of the development process as long as there are bug fixing tasks; for example, during the feature implementation stage, stabilization stage and post-release maintenance stage.

Footprint Reduction Code clones can be found at various degrees for different product teams we have worked with in Microsoft. Some teams are keen on reducing the memory footprint of their components; they look for every possible opportunity to achieve this soal a Removing code clones is one of the important actions they want to take.

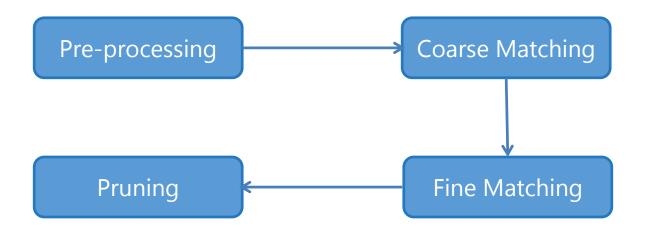
[IWSC'11 Dang et.al.]

### High tunability – what you tune is what you get

- Intuitive similarity metric
  - Effective control of the degree of syntactical differences between two code snippets
- Tunable at fine granularity
  - Statement similarity
  - % of inserted/deleted/modified statements
  - Balance between code structure and disordered statements

# High scalability

• Four-step analysis process



• Easily parallelizable based on source code partition

# High compatibility

- Compiler independent
- Light-weight built-in parsers for C/C++ and C#
- Open architecture for plug-in parsers to support different languages
- Easy adoption by product teams
  - Different build environment
  - Almost zero cost for trial

# High explorability

minute       minut       minute       minute	🚰 CloneExplorer -				🖄 CloneExplorer -		- 1 <del>- X</del>
Image: Note: Interviewed and the second s	File						$\bigcirc$
Image: provide the server's meet th			9	- 6		(6)	(4)
<ul> <li>I. Clone navigation based on source tree hierarchy</li> <li>Pivoting of folder level statistics</li> <li>Folder level statistics</li> <li>Clone function filters</li> <li>Source function filters</li>     &lt;</ul>							
<ul> <li>I. Clone navigation based on source tree hierarchy</li> <li>P. Pivoting of folder level statistics</li> <li>S. Folder level statistics</li> <li>S. Clone function filters</li> <li>S. Sorting by bug or refactoring potential</li> <li>T. acquing</li> </ul>	Show statistics by Bug potential Clone Function View			Show as 💿 List 💿 Group	Clone Group Group Index: 6003 Clone number: 2 Components: outlook, mso		
<ul> <li>I. Clone navigation based on source tree hierarchy</li> <li>Privoting of folder level statistics</li> <li>S. Folder level statistics</li> <li>S. Clone function filters</li> <li>S. Sorting by bug or refactoring potential</li> <li>T. agoing</li> </ul>	Folder High Moderate Low Zero						3
<ul> <li>I. Clone navigation based on source tree hierarchy</li> <li>Proteing of folder level statistics</li> <li>S. Folder level statistics</li> <li>Clone function list in selected folder</li> <li>C. Clone function list in selected folder</li> <li>S. Clone function list in selected folder</li> <li>S. Clone function filters</li> <li>S. Soring by bug or refactoring potential</li> <li>T. agoing</li> </ul>	Win7 20090508     14591 22182 35798 52246     Location scope     ✓ Inside file	Cross file, inside folder	Cross folder	$\bigcirc$			
<ul> <li>I. Clone navigation based on source tree hierarchy</li> <li>Pivoting of folder level statistics</li> <li>S. Folder level statistics</li> <li>C. Clone function list in selected folder</li> <li>S. Clone function filters</li> <li>S. Bug filing</li> <li>S. Tagging</li> </ul>				▶(7)			
<ul> <li>I. Clone navigation based on source tree hierarchy</li> <li>Pivoting of folder level statistics</li> <li>Folder level statistics</li> <li>Clone function list in selected folder</li> <li>S. Clone functio</li></ul>		w 📝 Moderate 📝 High		$\smile$			$\sim$
<ul> <li>I. Clone navigation based on source tree hierarchy</li> <li>Pivoting of folder level statistics</li> <li>S. Folder level statistics</li> <li>S. Clone function list in selected folder</li> <li>S. Clone function filters</li> <li>S. Clone function filters&lt;</li></ul>		40.1	Refa	ctor _ View Detail ^			(5)
I. Clone navigation based on source tree hierarchy Audional dial dial dial dial dial dial dial di			Bug Potential V Pote	ntial Tags & Tag			
<ul> <li>I. Clone navigation based on source tree hierarchy</li> <li>Pivoting of folder level statistics</li> <li>Souting by bug or refactoring potential</li> <li>Conse function list in selected folder</li> <li>Clone function list in selected folder</li></ul>	1402 Z FISHCOD CENSIMAL	-					
<ul> <li>I. Clone navigation based on source tree hierarchy</li> <li>Pivoting of folder level statistics</li> <li>Folder level statistics</li> <li>Sorting by bug or refactoring potential</li> <li>Conse function filters</li> <li>Sorting by bug or refactoring potential</li> <li>Tagging</li> </ul>					Clone Function Comparison		
<ul> <li>I. Clone navigation based on source tree hierarchy</li> <li>Pivoting of folder level statistics</li> <li>Folder level statistics</li> <li>Clone function list in selected folder</li> <li>Clone function filters</li> <li>Sorting by bug or refactoring potential</li> <li>Tagging</li> </ul>	2 Uni_inder.cpp Cler.vex.c	ostreamPart 80			Change - 1: FlsDark	Dhange v 2: FlsDark	
<ul> <li>I. Clone navigation based on source tree hierarchy</li> <li>Pivoting of folder level statistics</li> <li>Folder level statistics</li> <li>Clone function list in selected folder</li> <li>Sorting by bug or refactoring potential</li> <li>Tagging</li> </ul>	1513 Z audiotrace.cpp Audiovvin						-1
I. Clone navigation based on source tree hierarchy Pivoting of folder level statistics 3. Folder level statistics 4. Clone function list in selected folder 5. Clone function list in selected folder 5. Clone function list in selected folder 5. Clone function filters 6. Sorting by bug or refactoring potential 7. Taggging	1004 Z _voice.pp Cvoice.pp						
<ul> <li>Notice of the second of the sec</li></ul>	1455 2 endpoint.cpp Criveria				1536 const unsigned char ucNibblesize = 4;	2365 const unsigned char ucNibblesize = 4;	
1. Clone navigation based on source tree hierarchy 2. Pivoting of folder level statistics 3. Folder level statistics 4. Clone function filters 5. Clone function filters 6. Sorting by bug or refactoring potential 7. Tagging					1538	2367	Next Block
	1434 Z Policyconligc CPolicyco	nfigClient::ResetDeviceFor 37	High High	Detail	1540 {	2369 {	Action
<ul> <li>In the second of the second of</li></ul>	1435 2 endpoint.cpp Criveno				<pre>1542 palEntry = *(pPal + *(BYTE *)pPixel);</pre>	<pre>2371 palEntry = *(pPal + *(BYTE *)pPixel);</pre>	
A clone function list in selected folder         5. Clone function filters         6. Sorting by bug or refactoring potential         7. Tagging         7. T	1406 Z am_contros.cpp CDevicer				1544 case 4: // 4bpp: Extract nibble and deference into a 16	6 pal 2373 case 4: // 4bpp: Extract nibble and deference	into a 16 🔹 Gopy right function
File and the second of the					1546 palEntry = *(pPal + ((*(BYTE *)pPixel) & ucPixe	elmas 2375 palEntry = *(pPal + ((*(BYTE *)pPixel	.) & ucPixel 🍙 Copy both
How mids 47 28 40 00 register plate allower backet backet back back back back back back back back					1548 palEntry = *(pPal + ((*(BYTE *)pPixel) >> ucNik	bbles 2377 palEntry = *(pPal + ((*(BYTE *)pPixel	.) >> ucNibk
1. Clone navigation based on source tree hierarchy hgh Hgh Qual to statistics 2. Pivoting of folder level statistics 3. Folder level statistics 4. Clone function list in selected folder 5. Clone function filters 6. Sorting by bug or refactoring potential 7. Tagging	Exact leader leade	oCreate 46	High High	Detail	1550 case 24: // 24bpp: Read colours from pixel	2379 case 24: // 24bpp: Read colours from pixel	
<ul> <li>Hoh</li> <li>Hoh</li></ul>		IgBar::OnNcCalcSize 51			3 1552 break;	2 2381 palEntry.rgbRed = ((RGBQUAD *)pPixel)-	>rgbRed; Block type
<ul> <li>1. Clone havigation based on source tree hierarchy</li> <li>2. Pivoting of folder level statistics</li> <li>3. Folder level statistics</li> <li>4. Clone function list in selected folder</li> <li>5. Clone function filters</li> <li>6. Sorting by bug or refactoring potential</li> <li>7. Tagging</li> </ul>						2383 palEntry rdbBlue = ((RGBOUAD *))pEixel)	->rgbBlue; Exactly same
<ul> <li>2. Pivoting of folder level statistics</li> <li>3. Folder level statistics</li> <li>4. Clone function list in selected folder</li> <li>5. Clone function filters</li> <li>6. Sorting by bug or refactoring potential</li> <li>7. Tagging</li> </ul>	1. Clone navigation based on source tree l	hierarchy				pal 2385 default: // What else could it be? 2386 return 0;	
3. Folder level statistics         4. Clone function list in selected folder         5. Clone function filters         6. Sorting by bug or refactoring potential         7. Tagging		,			1. Block correspondence	4 2388 if (palEntry.rgbRed >= 0xFE 55 palEntry.rgbGreen	>= 0xFE 66
3. Folder level statistics         4. Clone function list in selected folder         5. Clone function filters         6. Sorting by bug or refactoring potential         7. Tagging			High High	Detail	2 Block types	5 2389 ((palEntry.rgbRed >= 0xbf && palEntry.rgbGreen 2390 (palEntry.rgbRed <= 0xc1 && palEntry.rgbGreen	<= 0xc1 44
4. Clone function list in selected folder         5. Clone function filters         6. Sorting by bug or refactoring potential         7. Tagging             8. Clone function filters             9. Clone function filters             9. Clone function filters             9. Sorting by bug or refactoring potential             9. Tagging             9. Tagging	3 Folder level statistics		-			2392 return TRUE;	(DIS) Disordered
4. Clone function list in selected folder         4. Clone function list in selected folder         5. Clone function filters         6. Sorting by bug or refactoring potential         7. Tagging					3. Block navigation	pix 2394	
5. Clone function filters       Hgh       Moderate       Detail         Hgh       Hgh       Detail	4. Clone function list in selected folder		-		C C	2396 /*	
6. Sorting by bug or refactoring potential       Hgh       Mgh       Detail       5. Bug filing         7. Tagging       Hgh       Hgh       Detail       6. Tagging       1000000000000000000000000000000000000	E Clana function filters				4. Copying	2398	
6. Softing by bug of relacioning potential         Han       Han         Han       Detail         Hon       Han         Hon       Detail         Hon       Detail         Hon       Han         Detail       6. Tagging	5. CIONE IUNCION IIILEIS		High Mode			2400 not get touched), into the rectangle at prc using	the HDC hc
$\frac{H_{gh}}{H_{gh}} + \frac{H_{gh}}{H_{gh}} = 0$	6 Sorting by bug or refactoring potential					BYT 2402 as bit-fields.	are given j
					6 Tagging	2404 %%Owner: TCoon	
	7. lagging					2406	
			Go to page: <<	< 1 💠 > >>	1579 Case R:	2407 Void DrawLustomBitmap(HDC ndc, kBCT *prc, HANDLE hDIB     2408 MSOCEO *prceo int thbex)	, namula nr -

# Scenarios and solutions

#### Quality gates at milestones

- Architecture refactoring
- Code clone clean up
- Bug fixing

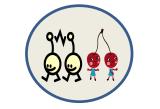
#### Post-release maintenance

- Security bug investigation
- Bug investigation for sustained engineering

#### Development and testing

- Checking for similar issues before check-in
- Reference info for code review
- Supporting tool for bug triage

Online code clone search



Offline code clone analysis

# Benefiting developer community



# Available in Visual Studio 2012

ething-Microsoft Vande Bando Edit Vene Antone Papate Bandi Sabing Team Oata Taulo Architecture Teat Analyse Window Help	
- (1) - + basy - (3, 1) - 2 - C - C - C - C - C - C - C - C - C	
Feren Taglaver Cina View Deletras Centeras Perding/Indu/Antworkers & 20 Centered/India	• Salution Egylawy •
<pre>itema-bandwidted.html:phalm = signal provide bandwidted.html:phalm = signal phalm = signal phalm</pre>	Company of the second sec
One Onep Court	# I ProfileDALFactory
Original     NephralitheoridanSateScoundaria - Cillian/yidang@acument/(iniual Studie Dev/23/highed).NET Petr Stop AUT/SQLPetRisdRX.UH2RepPrefile/hooider.cs Inius 17-302	📷 References 📲 Accemblylinfis.cs
Drang Math. (). File)     1.0xe	Distancements     PetDispunk
	e (B) SQU-balak I (B) Anomalytics (B)
1 Owned Shippets	Ren.cs     Order.cs
Toward and the	A DIMED

Searching similar snippets for fixing bug once

	top - (14) To 2 - P - Q - Passel - Q - (2 - (2 - (2 - (2 - (2 - (2 - (2			
een Eglorer Clau	u View 🛛 Ordenza 🕨 🗶 Ordenza – Petitikapike Referencielenza – Cunternalisika	- 2	Aution Explorer	
PetShop.SQLServerEAL.0	Order - 19 GatCridec(et anderlig		8 🖉 🔁 🕼 🖗 🖉 🖉 🖉 🖉 🖉	oh
//Cre	eate a parameter		lig Petihop.ank	
	harameter parm = new SqlParameter(PARP_GRDEM_ID, SqlDbType.Int);	1 A	PetShopQueue.co	
pare.	.Vilue = orderid;		# Crashdini.	
//***	ecute a query to read the order		Assembliefs.cs	
using	g (tqDatabasier rdr = SqDatiper.ExecuteBeader(tqDatiper.ConnectionStringOrderDistributedTransaction, ConnectionStringErderDistributedTransaction, Connection, C		Category CI	
			inventory.ci	
	]f (rdr.kead()) {		and Barriers	
	//demerate an order header from the first row		Cieder.cs	
	AddressInfo billingAddress = new AddressInfo(rdr.GetString(6), rdr.GetString(6), rdr.GetString(7), rdr.GetString(8), rdr	(11), rdr.	Order,Incert.sequence	
	Addressinfs shippingAddress = new AddressInfs(rdr.GetString(13), rdr.GetString(14), rdr.GetString(13), rdr.GetString(16), rdr.GetString(17), rdr.GetString(18), rdr.Get	Ptring(19)	Even Petilisep.orik	
	erder - new Orderinfs(orderid, rdr.GetDateTime(#), rdr.GetDitring(1), null, billingAddress, shippingAddress, rdr.GetDetimal(21), null, null);		Traductics	
	erer a neu unartaniquimera, recuenterae(a), recuentring(a), nut, satangueress, antpingueress, recuenterae(a), mats mats)		<ul> <li>OracleProfileDAL</li> </ul>	
	<pre>ilistellecteminfus limeftems + res listellecteminfus[];</pre>		<ul> <li>Faferences</li> <li>Assembly/info.cs</li> </ul>	
	limited its = mult	1	Accentight Forci Petilinop.ork	
	//Create the lineitees from the first row and subsequent rows		Tet DepPrefielrowider co	
	ACTENT OF ANTENESS FOR ONE TAILS FOR AND SANSAGETS FORS		4 G OrderProcessor	
	<pre>item = new ilselicelefo(rdr.GetString(22), string.Empty, rdr.GetInt32(23), rdr.GetInt32(24), rdr.GetBecimal(25));</pre>		Properties	
	liseitees.Add(itee);		E E Faferences	
	) while (rdr.Read()):		🚵 App.config	
	order.LineStens = new LineStenDito[lineStens.Count];		1 Assemblyänfo.cz	
	limeltems.CopyTs(order.Limeltems, 0);		E Patibop.ork	
)	3		Program.cs	
1			<ul> <li>Image: PaymentSatzway</li> </ul>	
8% + F			Expertis	
N No		- 1		
ode Clore Analysis Results	i e X Dapa		Properties Properties PaymentProcessor.cs PrtShop.orik	
ode Core Analysis Results Clone Group	i in X Depet		Im Properties     Im Properties     Im Pathemas     Province/Processor.cs     Professor.ex     ImProfessor.ex	
ode Clone Analysis Results Dane Group B Exact Metch 1/2 File	t n x Count Count Rd Tolers		Properties     Properties     Proventhecesserves     Proventhecesserves     Profileses     Profileses     Properties	
ode Clove Analysis Results Onne Orang B. Exoct Metch 1 (2 File B. Strang Metch 1 (2 File	1 ≤ x ≤ β λέγκε Rel Count Rel J Obere		Properties	
ode Clove Analysis Results Dane Group B Exect Metch 1 (2 File B Strang Metch 1 (2 File B Strang Metch 1 (2 File	Coart Coart Ref 1 Onen Ref 1 Onen Ref 1 Onen		Projekti     Projekti     Provestinosi     Poymestinosisiussi     Provestinosis     Professi     Projekti     Popekti     Popekti     Popekti     Popekti     Popekti     Popektisi	
ode Clove Analyse Results Clove Group B ExectMetch 1(2 File B Strong Metch 1(2 File Drang Metch 2 (2 Fil Order:GetOxfer - CS	A Start     Source     Sourc	-	Properties	
ada Clove Analyse Results Cane Orang B. Exect Metch 1 (2 File B. Strong Metch 1 (2 File B. Strong Metch 2 (2 Fil Onder GetOxder - C1 Onder GetOxder - C1	Count         Count           Rel         1 Oneon           Tanta         1 Oneon           Standy Langebourneet/Data State Schlangebourneet/Data Schlange	-	Properter     Properter	
olo Clove Analyse Result Date Group E ExectMetch 1 (2 File Strong Match 2 (2 Fil Onder Get2044 - C1 Onder Get2044 - C1 E Strong Match 2 (2 Fil	1         Count           No         Count           No         Count           Status         Count           Status         Count           Variation         Count	-	Properties     Properties     Propries     Provent/Inscissors.cs     Profiles on     Profiles     Profiles     Profiles     Profiles.cs     Profiles.cs	
old Obre Analyse Results Date Orang 8 Eact Match 1 (2 File 9 Drang Match 2 (2 Fil 0 nder GetOxder - C1 0 nder GetOxder - C1 8 Drang Match 2 (2 Fil 8 Drang Match 4 (2 Fil 8 Drang Match 4 (2 Fil	k ≥ K Drayet Rel Coart Rel J Down Tadj J Down	-	Properties	
alte Clear d'au Jon Routit Disse Group B. Exect Metch 102 File B. Streng Metch 20 Fil Order/SetDider - Cl Order/SetDider - Cl B. Streng Metch 20 Fil B. Streng Metch 20 Fil B. Streng Metch 20 Fil B. Streng Metch 20 Fil	Image: Source         Count           bit         10 mm		Popeter     P	
Alle Clove Andyos Reads Clove Group B Eact March 10 Fee 9 Strong Match 10 (Fe 0 Orbor GetDider - Cl 0 Orbor GetDider - Cl 9 Strong Match 10 (Fe 9 Strong Match 10 (Fe 9 Strong Match 10 (Fe 9 Strong Match 10 (Fe) 9 Strong Match 10 (Fe) 9 Strong Match 10 (Fe) 9 Strong Match 10 (Fe)			Propeter	
Alle Clove des des des Results Deses Braus Match 102 File Deses Match 102 Fil Deses Match 102 Fil			Propeter	
obe Clove Analysis Result: Clove Gimag B Exect Match 1 (2 Fee 9 Strong Match 1 (2 Fee 9 Strong Match 1 (2 Fe 0 detr GetMate - (1) 0 detr GetMate - (2) 0 detr GetMate - (2) 8 Strong Match 1 (2 Fe 8 Strong Match 1 (2 Fe 8 Strong Match 1 (2 Fe 8 Mediam Match 2 (2 9 Mediam Match 2 (2)	i i i i i i i i i i i i i i i i i i i		Propeter     Propeter     Propeter     Propeter     Provembrossmics     Provembrossmics     Propeter	
Other Adapts Read?           Disc Match 10 File           BiochMatch 10 File           Drong Match 10 File           Drong Match 10 File           OrberGelforder - CI           OrberGelforder - CI           Drong Match 10 File           Drong Matc	Conf Conf		In Propertie     I	
der Claim Zutrigen Result? Diese Greun Berreg Match 1:2 File Dereig Match 1:2 File Order Gefülder - Cl Order Gefülder - Cl Order Gefülder - Cl Dereig Match 1:2 Fil Direng Match 1:2 Fil Direng Match 1:2 Fil Berreg Match 1:2 Fil Berlegen Match 1:2 Fil Berlegen Match 1:2 Fil Mediann Match 1:2 Fil Mediann Match 1:2 Fil	a - 32 / Dright           HB         Conet           HB         20mm           State         Down		Constant and a second and	
Ad Class Radys Reads Date Even J Back Match 10 File Date gene Match 20 File Date Match 20 File Date Match 20 File Median Match 30 File Median Match 30 File Median Match 30 File Match 20 File Wash Match 30 File Wash Match 30 File Wash Match 30 File			Enclose     E	
At Clove Action Result. Direct Grag. Electritudes 10:2FeC Brendy Math 10:FE Direct GetMath 20: Direct GetMath 20: Direct GetMath 20: Breng Math 20:FE Breng Math 20:FE			Constant of the second of	
Alt Clove Active Insult Direct Dress B Sect Nets 1: 0 File Derey Methics 1: 0 File Mediane Nets 1: 2 File Mediane Nets 1: 2 File Mediane Methics 1: 0 File Weak	i i i i i i i i i i i i i i i i i i i		Construction	
At Clove Active Result Direct Gray Exect Name 5 12 File Street States 12 File Direct States 12 File Direct States 12 File Direct States 12 Direct States 12 Direct States 12 Direct States 12 Direct Match 12 Direct Match 12 Direct Match 12 Direct Match 12 Direct Match 12 Match 12 Mat	i i i i i i i i i i i i i i i i i i i		Construction	
Alt Clove Analysis broads Direct Dress B Sect Meds 10, 2 File Derey Medis 10, 2 File Mediaes Medis 10, 2 File Mediaes Medis 10, 2 File Mediaes Medis 10, 2 File Mediaes Medis 10, 2 File Weak M	i i i i i i i i i i i i i i i i i i i		Construction     C	
Alt Clove Analysis broads Direct Dress B Sect Meds 10, 2 File Derey Medis 10, 2 File Mediaes Medis 10, 2 File Mediaes Medis 10, 2 File Mediaes Medis 10, 2 File Mediaes Medis 10, 2 File Weak M	i i i i i i i i i i i i i i i i i i i		Construction     C	
Alt Clove Analysis broads Direct Dress B Sect Meds 10, 2 File Derey Medis 10, 2 File Mediaes Medis 10, 2 File Mediaes Medis 10, 2 File Mediaes Medis 10, 2 File Mediaes Medis 10, 2 File Weak M	i i i i i i i i i i i i i i i i i i i		Construction     C	
Alt Clove Analysis broads Direct Dress B Sect Meds 10, 2 File Derey Medis 10, 2 File Mediaes Medis 10, 2 File Mediaes Medis 10, 2 File Mediaes Medis 10, 2 File Mediaes Medis 10, 2 File Weak M	i i i i i i i i i i i i i i i i i i i		Constant of the second se	
Alt Clove Analysis broads Direct Dress B Sect Meds 10, 2 File Derey Medis 10, 2 File Mediaes Medis 10, 2 File Mediaes Medis 10, 2 File Mediaes Medis 10, 2 File Mediaes Medis 10, 2 File Weak M	i i i i i i i i i i i i i i i i i i i		Construction     C	
da Cherc Analysis Na July Deer Creat Boots Heat 3 (2) Fee Boots Heat 3 (2) Fee Boots Heat 3 (2) Fee Deer Marks 3 (2) Fo Order Definition - (2) Boots Heat 3 (2) Fee Deer Marks 3 (2) Fee Boots Heat 3 (2) Fee Heat Heat 3 (2) Fee Boots Heat 3 (2) Fee	i i i i i i i i i i i i i i i i i i i		Comparison     C	

Finding refactoring opportunity

# More secure Microsoft products



Code Clone Search service integrated into workflow of Microsoft Security Response Center



Over hundreds of million lines of code indexed across multiple products



Real security issues proactively identified and addressed

# Example – MS security bulletin MS12-034

Combined Security Update for Microsoft Office, Windows, .NET Framework, and Silverlight, published: Tuesday, May 08, 2012

3 publicly disclosed vulnerabilities and seven privately reported involved. Specifically, one is exploited by the <u>Duqu malware</u> to execute arbitrary code when a user opened a malicious Office document

Insufficient bounds check within the font parsing subsystem of win32k.sys Cloned copy in gdiplus.dll, ogl.dll (office), Silver Light, Windows Journal viewer

### Microsoft Technet Blog about this bulletin

"However, we wanted to be sure to address the vulnerable code wherever it appeared across the Microsoft code base. To that end, we have been working with Microsoft Research to develop a "Cloned Code Detection" system that we can run for every MSRC case to find any instance of the vulnerable code in any shipping product. This system is the one that found several of the copies of CVE-2011-3402 that we are now addressing with MS12-034."

# Three years of effort

### Prototype development

- Problem formulation
- Algorithm research
- Prototype
   development

### Early adoption

- Algorithm
   improvement
- System / UX improvement

### Tech transfer

- System integration
- Process integration

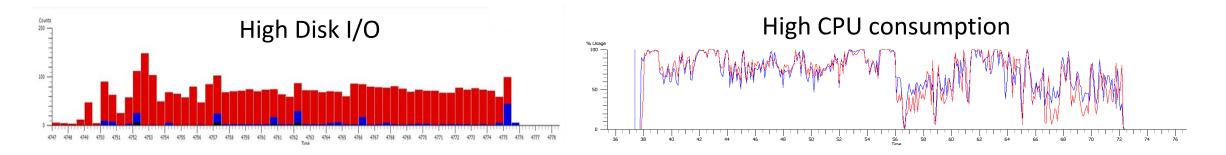
# StackMine

# Performance debugging in the large via mining millions of stack traces

Shi Han, Yingnong Dang, Song Ge, Dongmei Zhang, and Tao Xie, <u>Performance Debugging in the Large via Mining Millions of Stack Traces</u>, in *Proceedings of the 34th International Conference on Software Engineering (<u>ICSE 2012</u>), Zurich, Switzerland, June 2012.* 

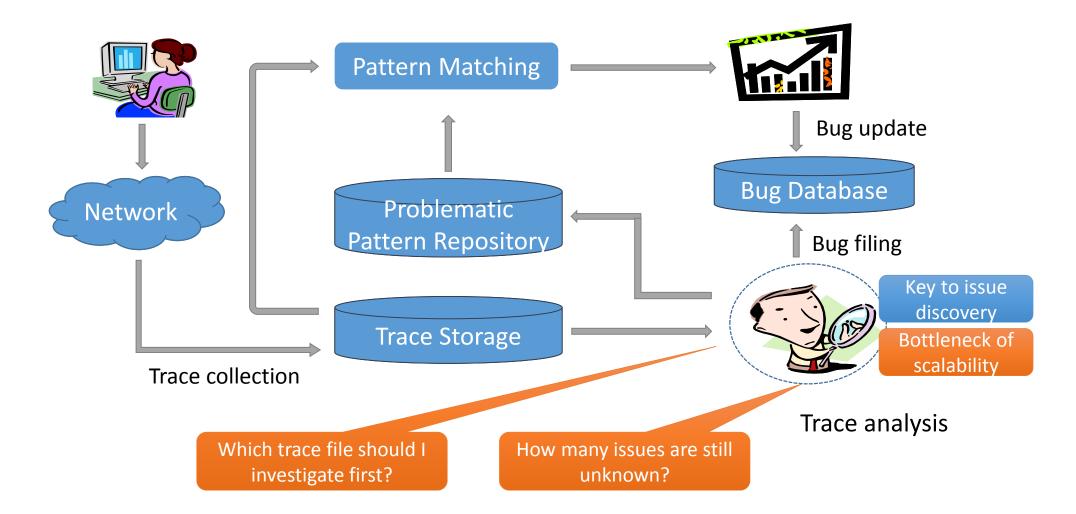
# Performance issues in the real world

- One of top user complaints
- Impacting large number of users every day
- High impact on usability and productivity



As modern *software systems* tend to get more and more *complex*, given *limited* time and resource *before* software *release*, *development-site* testing and debugging become more and more *insufficient* to ensure satisfactory software performance.

# Performance debugging in the large



# Problem definition

Given operating system traces collected from tens of thousands (potentially millions) of users, how to help domain experts identify the program execution patterns that cause the *most impactful* underlying performance problems with *limited time and resource*?

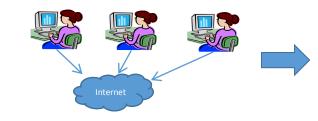


# Goal

Systematic analysis of OS trace sets that enables

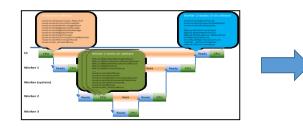
- Efficient handling of large-scale trace sets
- Automatic discovery of new program execution patterns
- Effective prioritization of performance investigation

## Challenges



#### Large-scale trace data

- TBs of trace files and increasing
- Millions of events in single trace stream



#### Highly complex analysis

- Numerous program runtime combinations triggering performance problems
- Multi-layer runtime components from application to kernel intertwined

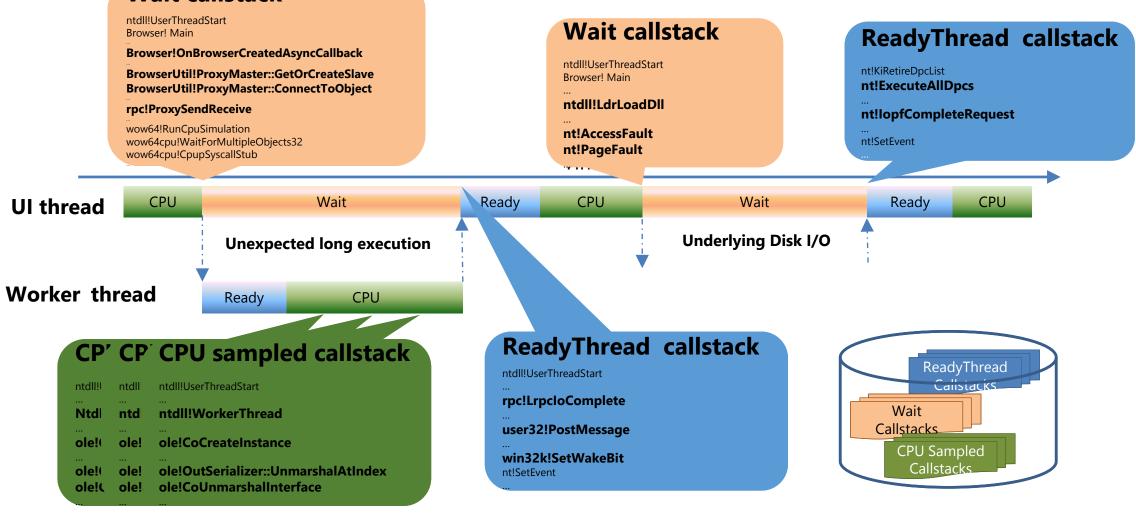


#### Combination of expertise

• Generic machine learning tools without domain knowledge guidance do not work well

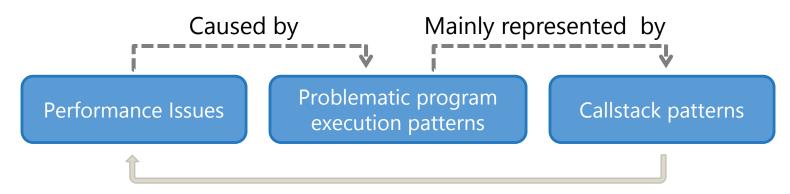
## Intuition

What happens behind a typical UI-delay? An example of delayed browser tab creation - Wait callstack





#### Formulate as a callstack mining and clustering problem



Discovered by mining & clustering costly patterns

## Technical highlights

- Machine learning for system domain
  - Formulate the discovery of problematic execution patterns as callstack mining and clustering
  - Systematic mechanism to incorporate domain knowledge
- Interactive performance analysis system
  - Parallel mining infrastructure based on HPC+MPI
  - Visualization aided interactive exploration

### Impact



"We believe that the MSRA tool is highly valuable and much more efficient for mass trace (100+ traces) analysis. For 1000 traces, we believe the tool saves us 4-6 weeks of time to create new signatures, which is quite a significant productivity boost."



Highly effective new issue discovery on Windows mini-hang



Continuous impact on future Windows versions

## Service Analysis Studio

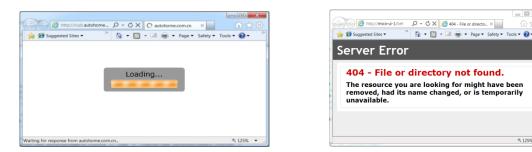
#### Incident management for online services

Jian-Guang Lou, Qingwei Lin, Rui Ding, Qiang Fu, Dongmei Zhang and Tao Xie, <u>Software Analytics for Incident Management of Online Services: An Experience Report</u>, in *Proceedings of the 28<sup>th</sup> IEEE/ACM* International Conference on Automated Software Engineering (<u>ASE 2013</u>), Experience papers, Palo Alto, California, November 2013.

### Motivation

- Online services are increasingly popular and important
- High service quality is the key
- Incident management is a critical task to ensure service quality

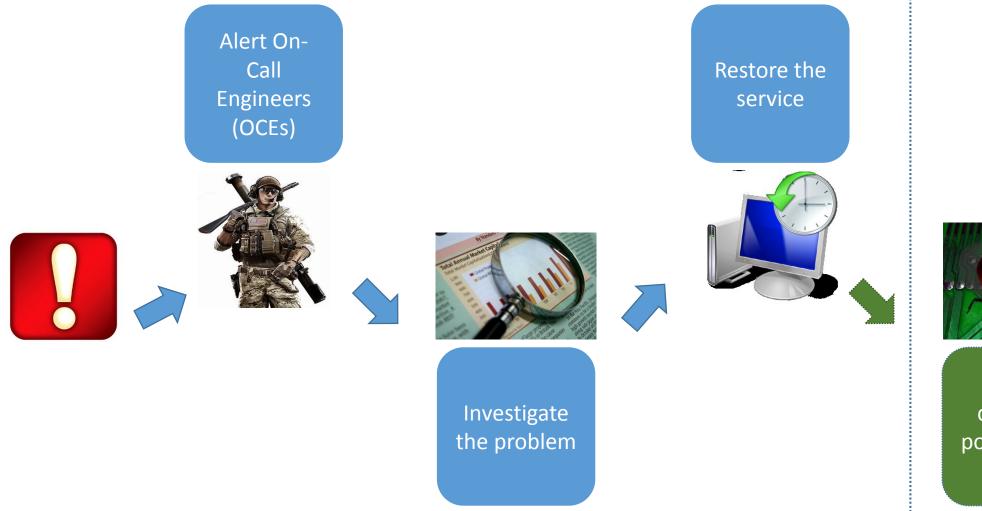






125% •

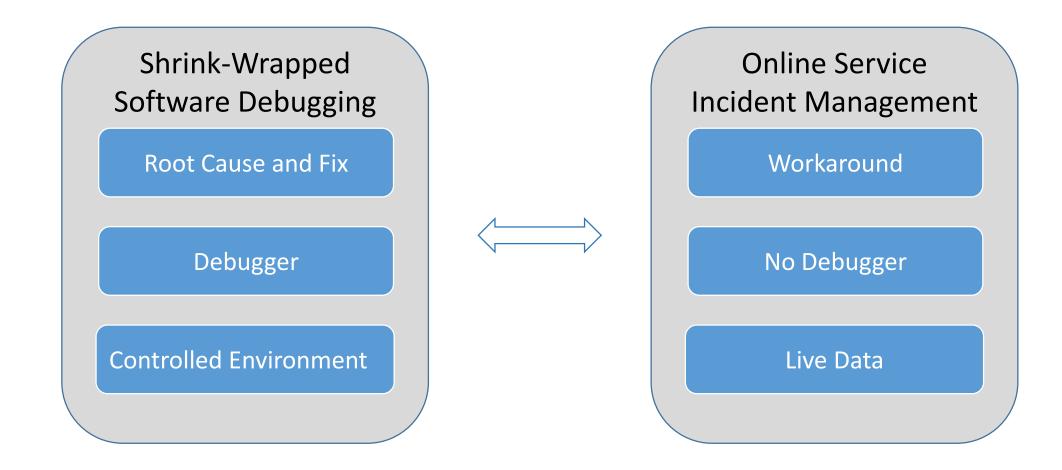
### Incident management: workflow





Fix root cause via postmortem analysis

### Incident management: characteristics



### Incident management: challenges



Large-volume and noisy data Highly complex problem space



Knowledge scattered and not well organized Few people with knowledge of entire system

### Data sources

Name	Description	Examples
Key Performance Indicators (KPI)	Measurements indicating the major quality perspectives of an online service	Request failure rate, average request latency, etc.
Performance counters and system events	Measurements and events indicating the status of the underlying system and applications	CPU, disk queue length, I/O, request workload, SQL-related metrics, and application-specific metrics, etc.
User requests	Information on user requests	Request return status, processing time, consumed resources, etc.
Transaction logs	Generated during execution, recording system runtime behaviors when processing requests	Timestamp, request ID, thread ID, event ID, and detailed text message, etc.
Incident repository	Historical records of service incidents	Incident description, investigation details, restoration solution, etc.

## Service Analysis Studio (SAS)

• Goal

Given an incident in an online service, effectively helping service engineers reduce Mean Time To Restore (MTTR).

- Design principals
  - Automating data analysis
  - Handling heterogeneous data sources
  - Accumulating knowledge
  - Supporting human-in-the-loop (HITL)



### Data analysis techniques

Mining suspicious execution patterns from transaction logs

Identifying incident beacons from system metrics

> Data-driven service analytics

Mining resolution solutions from historical incidents

### Impact

#### Deployment

- SAS deployed to worldwide datacenters of Service X in June 2011
- Five more updates since first deployment

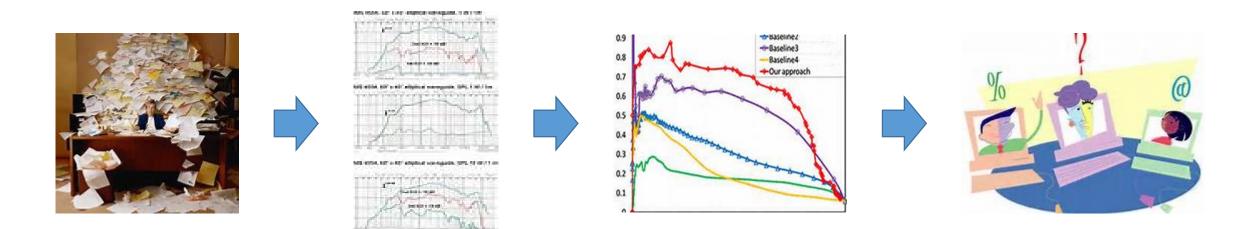
#### Usage

- Heavily used by On-Call Engineers of Service X for about 2 years
- Helped successfully diagnose ~76% of service incidents

### Lessons learned

- Understanding and solving real problems
- Understanding data and system
- Handling data issues
- Making SAS highly usable
- Achieving high availability and performance
- Delivering step-by-step

## Understanding and solving real problems



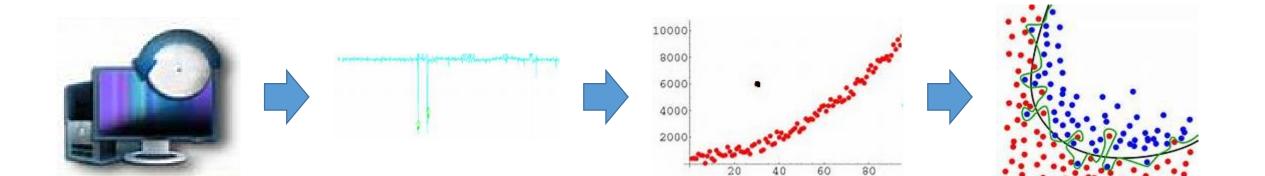
- Working side-by-side with On-Call Engineers
- Targeting at reducing MTTR
- Focusing on addressing challenges in real-world scenarios

### Understanding data and system

#### Techniques



#### **Practical Problems**



### Handling data issues

Data issues	Approach	Experience	
<ul><li>(1) Missing/duplicated</li><li>(2) Buggy</li><li>(3) Disordered</li></ul>	<ul><li>(1) Preprocessing</li><li>(2) Designing robust algorithms</li></ul>	Data preprocessing cannot be perfect. Robust algorithms are in great need.	

## Making SAS highly usable

There is an internal server error related issue.

Datacenter: DC1

Start time: 9/4/2012 3:48:00 AM End time: 9/4/2012 3:58:00 AM

#### Impact:

Influenced requests	1000
Influenced end users	100

#### Diagnosis:

This issue is a problem of "Credential loss". The source of the issue mainly locates at Front End Server—

#### <u>"FE001"</u>.

Here are similar previous occurrences of the issue:

- Incident ID 91236: 3/14/2012 10:49:00 AM (see detail)
- Incident ID 91271: 7/26/2012 14:25:00 AM (see detail)

#### See also:

Malfunctioned Frontend Servers973 of 1000 failed requests related to FE001.Malfunctioned SQL ServersNo malfunctioned SQL servers detected.Suspicious MetricsNo highly correlated metrics found.Suspicious Execution Patterns1 major pattern in the logs covers 973 of 1000 failed requests.

Suggested actions based on similar past incident (ID 91236): Reset the IIS service on the front end server FE001.

#### Actionable

#### Understandable

Easy to navigate

## Achieving high availability and performance

- SAS is also a service
  - To serve On-Call Engineers at any time with high performance
  - Critical to reducing MTTR of services
- Auto recovery
  - Continuously monitored
  - Check-point mechanism adopted
- Backend service + On-demand analysis

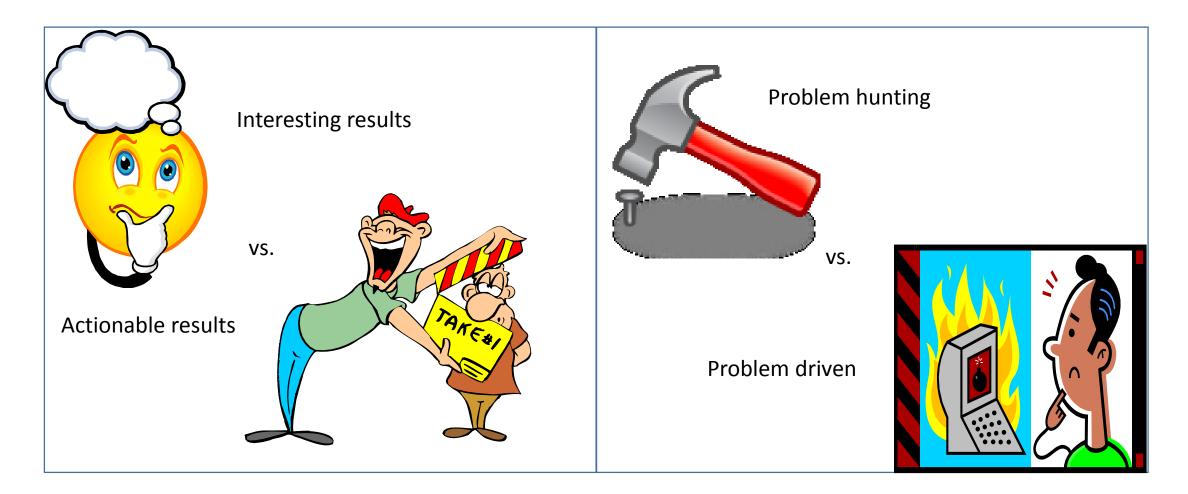
## Delivering step-by-step

- Demonstrating value and building trust
  - Deployment in production has cost and risk
  - In-house  $\rightarrow$  dogfood  $\rightarrow$  one datacenter  $\rightarrow$  worldwide datacenters
- Getting timely feedback
  - Requirements may not be clear early on and requirements may change
  - Gaining troubleshooting experiences from On-Call Engineers
  - Understanding how SAS was used
  - Identifying direction of improvement

## Outline

- Overview of Software Analytics
- Selected projects
- Experience sharing on Software Analytics in Practice

### Analytics is the means to the end



## Beyond the "usual" mining

Mining vs. matching

- Automatic vs. interactive

Researchers vs. practitioners

## Keys to making real impact

- Engagement of practitioners
  - Solving their problem
     Champions in product teams
     Timing
     Culture
- Walking the last mile
  - Targeting at real scenariosTrying out tool has cost"It works" is not enoughGetting engineering support
- Combination of expertise
  - **¤** Research capabilities
  - Xisualization & design

- **¤** Engineering skills to build systems
- X Communication

## Summary

#### Software Analytics

Software analytics is to enable *software practitioners* to perform data exploration and analysis in order to obtain *insightful and actionable* information for *data-driven tasks* around software and services.

#### Target audience - software practitioners



#### Five dimensions



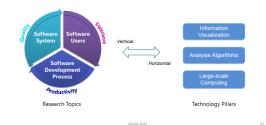
#### Output – insightful information

Meaningful and useful understanding or knowledge towards completing target tasks

#### Output - actionable information

- Enabling software practitioners to come up with concrete solutions towards completing target tasks
- Examples
- How to reduce regression (re-opened bugs) in development? Groups of re-opened bugs each with the same reason of re-opening
- How to recover an online service when an incident occurs?
   Problematic areas identified with reasons and confidence values
- How to decide which part of my code should be refactored?
   Detected cloned code snippets easily explored from different perspectives

Research topics and technology pillars



#### Connection to practice

Software Analytics is naturally tied with software development practice

#### Getting real



#### Together let us walk the exciting journey to make great impact!

# Q&A

http://research.microsoft.com/groups/sa/