The Secret of Big Data is….

…

TIM MENZIES, CS, NC State, USA
tim.menzies@gmail.com

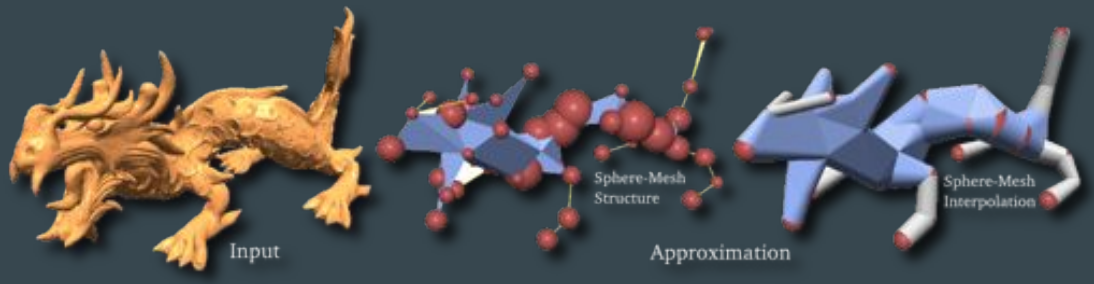
on line: http://tiny.cc/tim15research

Is the secret of big data little data?

tiny.cc/tim15research

When you have lots of data… throw most of it away?

In Computer Graphics



In Machine learning

- Variable subset selection (Kohavi, 1997)
- Instance selection (Chen, 1975)
- Active learning

In Theorem proving

- Narrows (Amarel, 1986)
- Master variables (Crawford 1995)
- Back doors (Selman 2002).

In Software Eng.

- Saturation in mutation testing (Budd, 1980 and many others

Data mining = data carving?

1. Find some cr*p
2. Cut the cr*p
3. Goto step 1

Some background

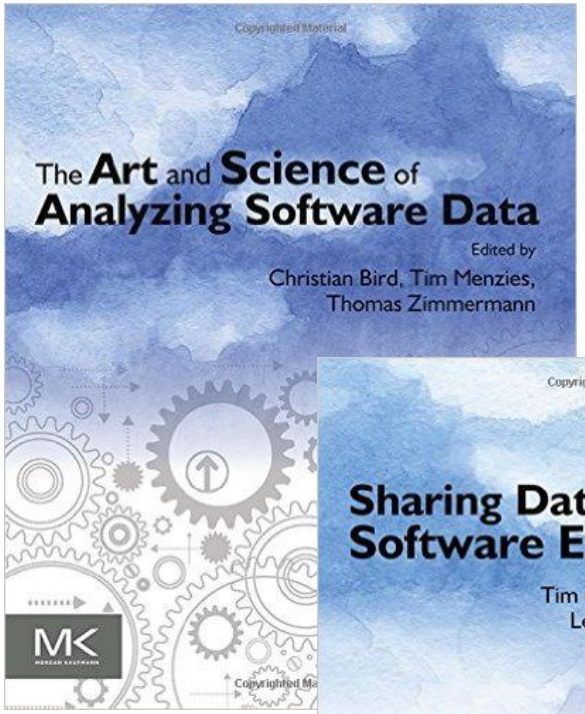
Good News

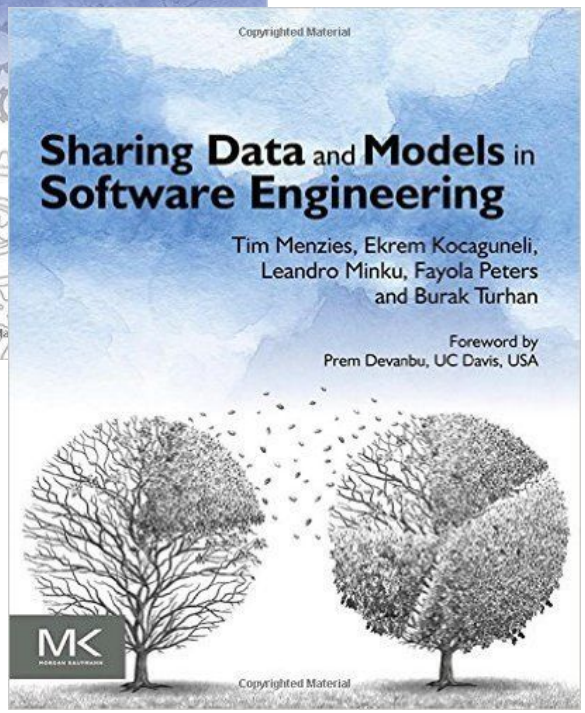
Software project data can

- be shared
- still be private
- still be used to build predictors

- Peters ICSE'12
- Peters TSE'13
- Peters ICSE'15

tiny.cc/tim15research





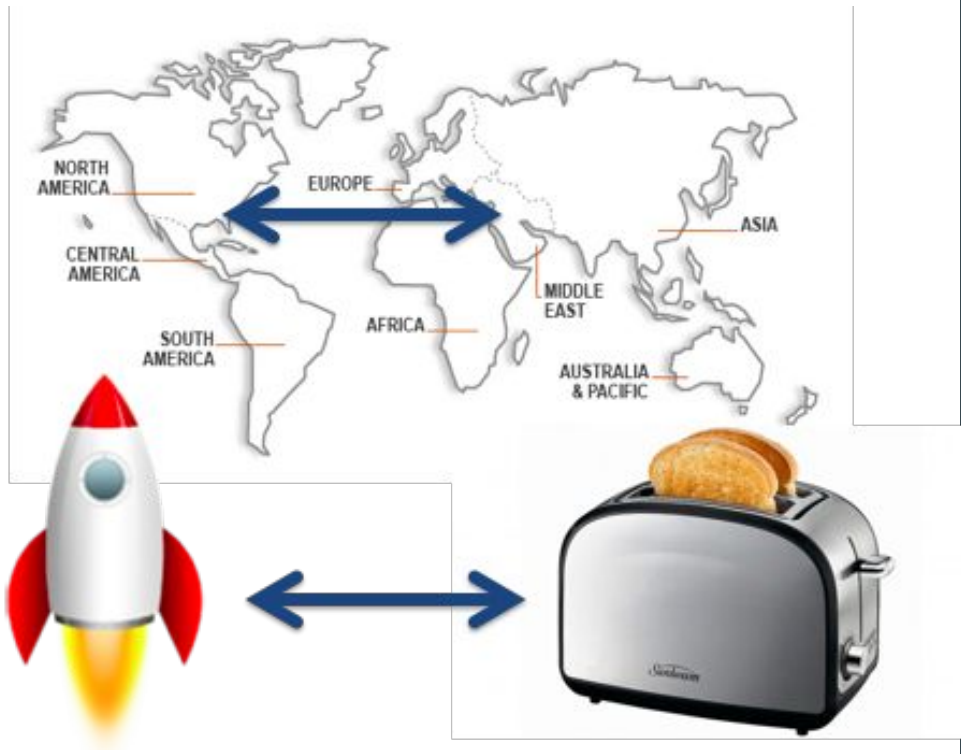
Good news: Transfer learning

Cross-company learning works:

- even proprietary to open source,
- even better data with different column names

- Turhan, Menzies, Bener ESE'09
- He et al. ESEM'13
- Peters ICSE'15
- Nam FSE'15 (Heterogeneous)

tiny.cc/tim15research



Turhan'09: Turkey to Texas. Toasters to rocket ships.

Scales up to massive studies

e.g. every Devanbu et al. study of Github

tiny.cc/tim15research

A Large Scale Study of Programming Languages and Code Quality in Github

Baishakhi Ray, Daryl Posnett, Vladimir Filkov, Premkumar Devanbu
{bairay@, dpposnett@, filkov@cs., devanbu@cs.}ucdavis.edu
Department of Computer Science, University of California, Davis, CA, 95616, USA

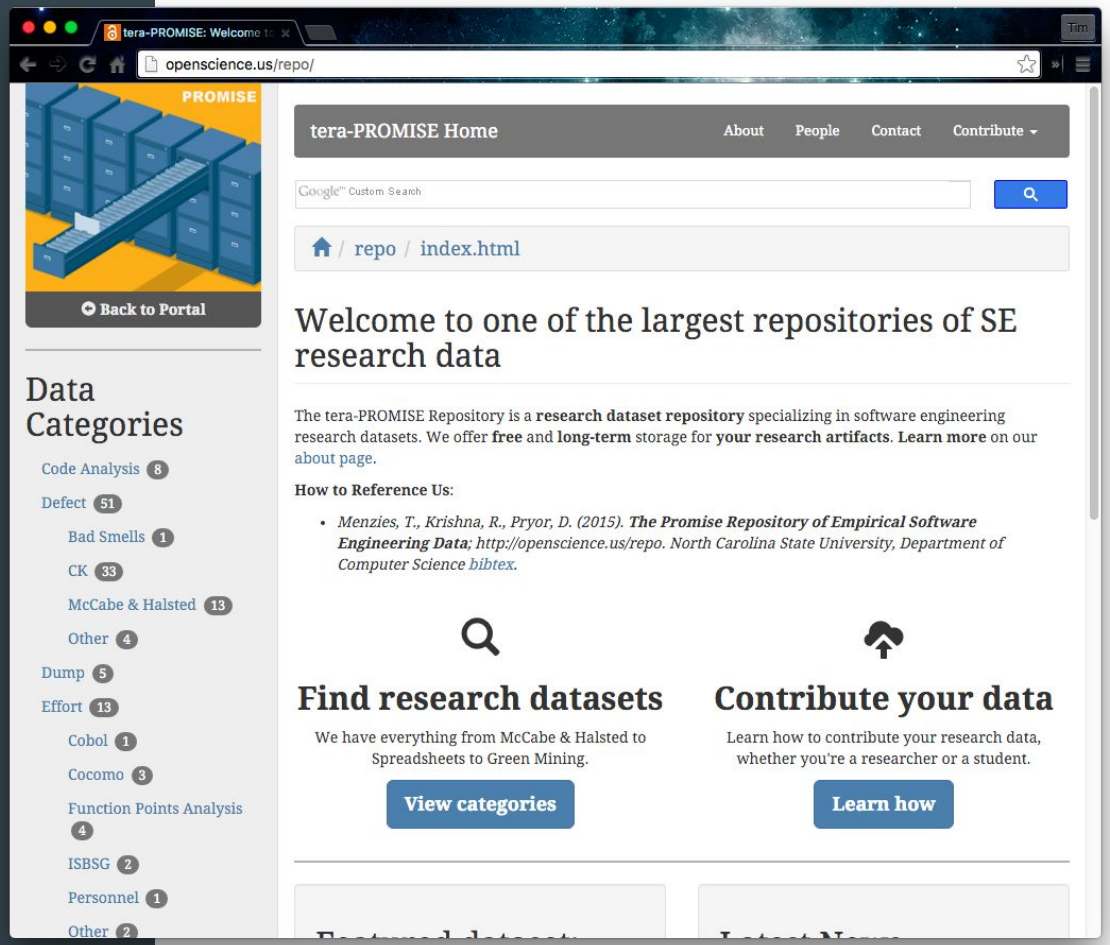
What is the effect of programming languages on software quality? This question has been a topic of much debate for a very long time. In this study, we gather a very large data set from GitHub (728 projects, 63 Million SLOC, 29,000 authors, 1.5 million commits, in 17 languages) in an attempt to shed some empirical light on this question. This reasonably large sample size allows us to use

FSE'14 November 16âĂŞ22, 2014, Hong Kong, China

tiny.cc/tim15research

Let's all share more data

- openscience.us/repo



Data mining = data carving

1. Find some cr*p
2. Cut the cr*p
3. Goto step 1

Example1: trusted data sharing

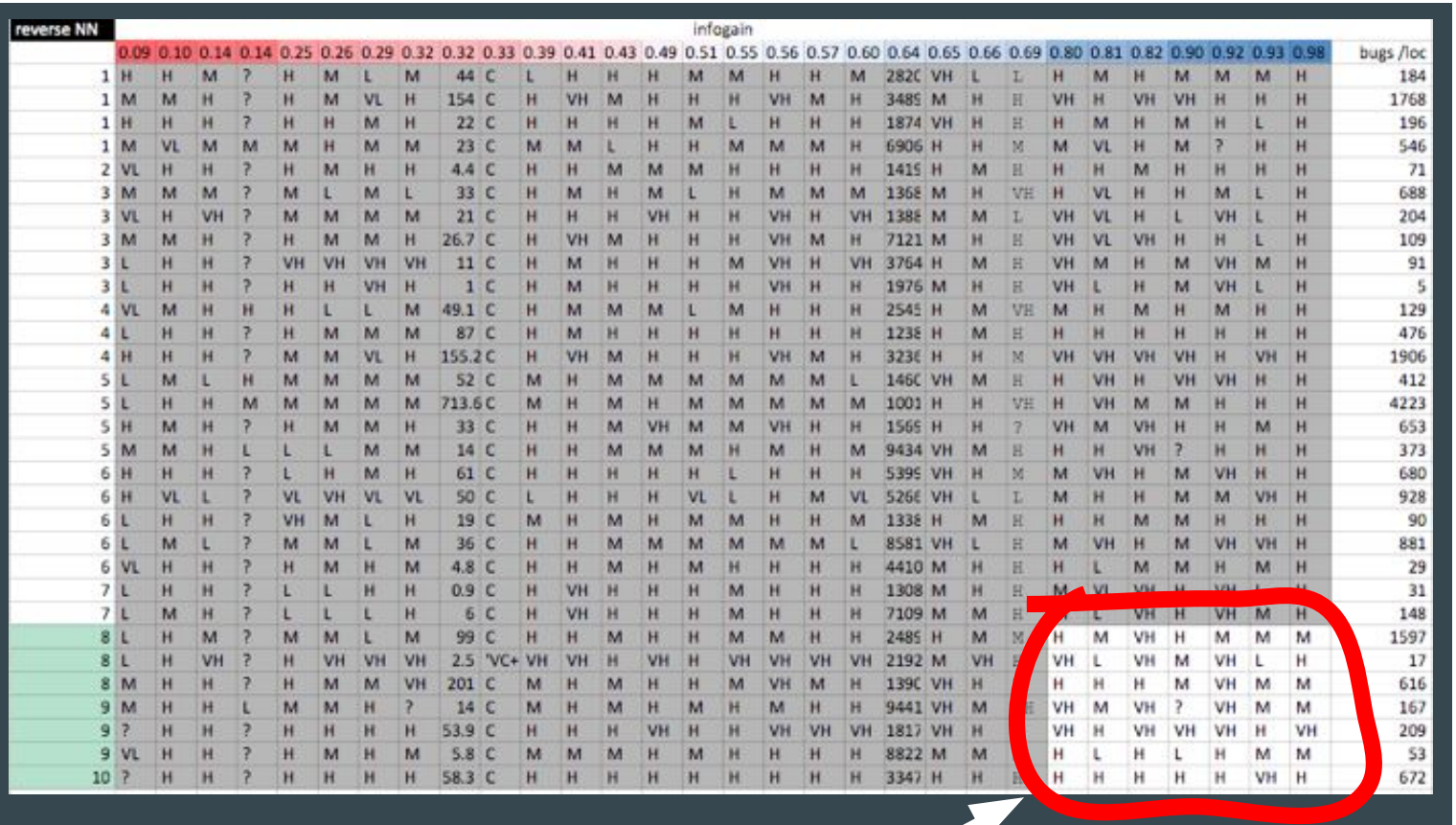
3 laws of trusted data sharing

1) only share corners
2) mutate corners before sharing
3) never mutate beyond halfway to nearest neighbor

reverse NN	infogain																															
	0.09	0.10	0.14	0.14	0.25	0.26	0.29	0.32	0.32	0.33	0.39	0.41	0.43	0.49	0.51	0.55	0.56	0.57	0.60	0.64	0.65	0.66	0.69	0.80	0.81	0.82	0.90	0.92	0.93	0.98	bugs /loc
1	H	H	M	?	H	M	L	M	44	C	L	H	H	H	M	M	H	H	M	282(	VH	L	L	H	M	H	M	M	M	H	184
1	M	M	H	?	H	M	VL	H	154	C	H	VH	M	H	H	H	VH	M	H	348(	M	H	H	VH	H	VH	VH	H	H	H	1768
1	H	H	H	?	H	H	M	H	22	C	H	H	H	H	M	L	H	H	H	1874	VH	H	H	H	M	H	M	H	L	H	196
1	M	VL	M	M	M	H	M	M	23	C	M	M	L	H	H	M	M	M	H	6906	H	H	M	M	VL	H	M	?	H	H	546
2	VL	H	H	?	H	M	H	H	4.4	C	H	H	M	M	M	H	H	H	H	1415	H	M	H	H	H	M	H	H	H	H	71
3	M	M	M	?	M	L	M	L	33	C	H	M	H	M	L	H	M	M	M	1368	M	H	VH	H	VL	H	H	M	L	H	688
3	VL	H	VH	?	M	M	M	M	21	C	H	H	H	VH	H	H	VH	H	VH	1388	M	M	L	VH	VL	H	L	VH	L	H	204
3	M	M	H	?	H	M	M	H	26.7	C	H	VH	M	H	H	H	VH	M	H	7121	M	H	H	VH	VL	VH	H	H	L	H	109
3	L	H	H	?	VH	VH	VH	VH	11	C	H	M	H	H	H	M	VH	H	VH	3764	H	M	H	VH	M	H	M	VH	M	H	91
3	L	H	H	?	H	H	VH	H	1	C	H	M	H	H	H	H	VH	H	H	1976	M	H	H	VH	L	H	M	VH	L	H	5
4	VL	M	H	H	H	L	L	M	49.1	C	H	M	M	M	L	M	H	H	H	2545	H	M	VH	M	H	M	H	M	H	H	129
4	L	H	H	?	H	M	M	M	87	C	H	M	H	H	H	H	H	H	H	1238	H	M	H	H	H	H	H	H	H	H	476
4	H	H	H	?	M	M	VL	H	155.2	C	H	VH	M	H	H	H	VH	M	H	323(	H	H	M	VH	VH	VH	VH	H	VH	H	1906
5	L	M	L	H	M	M	M	M	52	C	M	H	M	M	M	M	M	M	L	146(	VH	M	H	H	VH	H	VH	VH	H	H	412
5	L	H	H	M	M	M	M	M	713.6	C	M	H	M	H	M	M	M	M	M	1001	H	H	VH	H	VH	M	M	H	H	H	4223
5	H	M	H	?	H	M	M	H	33	C	H	H	M	VH	M	M	VH	H	H	156(	H	H	?	VH	M	VH	H	H	M	H	653
5	M	M	H	L	L	L	M	M	14	C	H	H	M	M	M	H	M	H	M	9434	VH	M	H	H	H	VH	?	H	H	H	373
6	H	H	H	?	L	H	M	H	61	C	H	H	H	H	H	L	H	H	H	539(	VH	H	M	M	VH	H	M	VH	H	H	680
6	H	VL	L	?	VL	VH	VL	VL	50	C	L	H	H	H	VL	L	H	M	VL	526(	VH	L	L	M	H	H	M	M	VH	H	928
6	L	H	H	?	VH	M	L	H	19	C	M	H	M	H	M	M	H	H	M	133(	H	M	H	H	H	M	M	H	H	H	90
6	L	M	L	?	M	M	L	M	36	C	H	H	M	M	M	M	M	M	L	8581	VH	L	H	M	VH	H	M	VH	VH	H	881
6	VL	H	H	?	H	M	H	M	4.8	C	H	H	M	H	M	H	H	H	H	4410	M	H	H	H	L	M	M	H	M	H	29
7	L	H	H	?	L	L	H	H	0.9	C	H	VH	H	H	H	M	H	H	H	1308	M	H	H	M	VL	VH	H	VH	L	H	31
7	L	M	H	?	L	L	L	H	6	C	H	VH	H	H	H	M	H	H	H	7109	M	M	H	M	L	VH	H	VH	M	H	148
8	L	H	M	?	M	M	L	M	99	C	H	H	M	H	H	M	M	H	H	248(	H	M	M	H	M	VH	H	M	M	M	1597
8	L	H	VH	?	H	VH	VH	VH	2.5	'VC+	VH	VH	H	VH	H	VH	VH	VH	VH	2192	M	VH	H	VH	L	VH	M	VH	L	H	17
8	M	H	H	?	H	M	M	VH	201	C	M	H	M	H	H	M	VH	M	H	139(	VH	H	H	H	H	H	M	VH	M	M	616
9	M	H	H	L	M	M	H	?	14	C	M	H	M	H	M	H	M	H	H	9441	VH	M	VH	VH	M	VH	?	VH	M	M	167
9	?	H	H	?	H	H	H	H	53.9	C	H	H	H	VH	H	H	VH	VH	VH	1817	VH	H	H	VH	H	VH	VH	VH	H	VH	209
9	VL	H	H	?	H	M	H	M	5.8	C	M	M	M	H	M	H	H	H	H	8822	M	M	M	H	L	H	L	H	M	M	53
10	?	H	H	?	H	H	H	H	58.3	C	H	H	H	H	H	H	H	H	H	3347	H	H	H	H	H	H	H	H	VH	H	672

3 laws of trusted data sharing

1) only share corners
2) mutate corners before sharing
3) never mutate beyond halfway to nearest neighbor

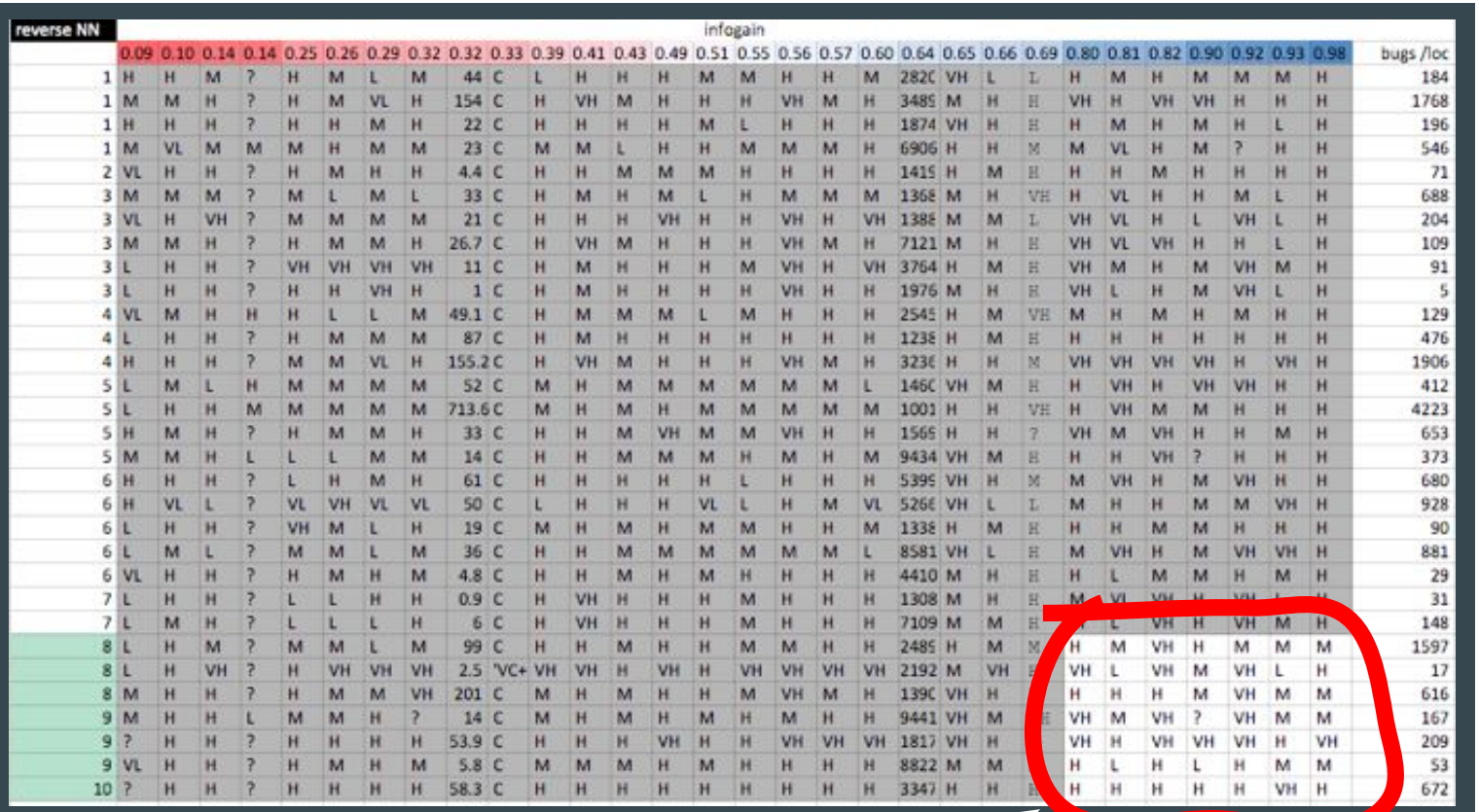


reverse NN	infogain																															
	0.09	0.10	0.14	0.14	0.25	0.26	0.29	0.32	0.32	0.33	0.39	0.41	0.43	0.49	0.51	0.55	0.56	0.57	0.60	0.64	0.65	0.66	0.69	0.80	0.81	0.82	0.90	0.92	0.93	0.98	bugs /loc
1	H	H	M	?	H	M	L	M	44	C	L	H	H	H	M	M	H	H	M	282(	VH	L	L	H	M	H	M	M	M	H	184
1	M	M	H	?	H	M	VL	H	154	C	H	VH	M	H	H	H	VH	M	H	348(	M	H	H	VH	H	VH	VH	H	H	H	1768
1	H	H	H	?	H	H	M	H	22	C	H	H	H	H	M	L	H	H	H	1874	VH	H	H	H	M	H	M	H	L	H	196
1	M	VL	M	M	M	H	M	M	23	C	M	M	L	H	H	M	M	M	H	6906	H	H	M	M	VL	H	M	?	H	H	546
2	VL	H	H	?	H	M	H	H	4.4	C	H	H	M	M	M	H	H	H	H	1415	H	M	H	H	H	M	H	H	H	H	71
3	M	M	M	?	M	L	M	L	33	C	H	M	H	M	L	H	M	M	M	1368	M	H	VH	H	VL	H	H	M	L	H	688
3	VL	H	VH	?	M	M	M	M	21	C	H	H	H	VH	H	H	VH	H	VH	1388	M	M	L	VH	VL	H	L	VH	L	H	204
3	M	M	H	?	H	M	M	H	26.7	C	H	VH	M	H	H	H	VH	M	H	7121	M	H	H	VH	VL	VH	H	H	L	H	109
3	L	H	H	?	VH	VH	VH	VH	11	C	H	M	H	H	H	M	VH	H	VH	3764	H	M	H	VH	M	H	M	VH	M	H	91
3	L	H	H	?	H	H	VH	H	1	C	H	M	H	H	H	H	VH	H	H	1976	M	H	H	VH	L	H	M	VH	L	H	5
4	VL	M	H	H	H	L	L	M	49.1	C	H	M	M	M	L	M	H	H	H	2545	H	M	VH	M	H	M	H	M	H	H	129
4	L	H	H	?	H	M	M	M	87	C	H	M	H	H	H	H	H	H	H	1238	H	M	H	H	H	H	H	H	H	H	476
4	H	H	H	?	M	M	VL	H	155.2	C	H	VH	M	H	H	H	VH	M	H	323(	H	H	M	VH	VH	VH	VH	H	VH	H	1906
5	L	M	L	H	M	M	M	M	52	C	M	H	M	M	M	M	M	M	L	146(	VH	M	H	H	VH	H	VH	VH	H	H	412
5	L	H	H	M	M	M	M	M	713.6	C	M	H	M	H	M	M	M	M	M	1001	H	H	VH	H	VH	M	M	H	H	H	4223
5	H	M	H	?	H	M	M	H	33	C	H	H	M	VH	M	M	VH	H	H	156(	H	H	?	VH	M	VH	H	H	M	H	653
5	M	M	H	L	L	L	M	M	14	C	H	H	M	M	M	H	M	H	M	9434	VH	M	H	H	H	VH	?	H	H	H	373
6	H	H	H	?	L	H	M	H	61	C	H	H	H	H	H	L	H	H	H	539(	VH	H	M	M	VH	H	M	VH	H	H	680
6	H	VL	L	?	VL	VH	VL	VL	50	C	L	H	H	H	VL	L	H	M	VL	526(	VH	L	L	M	H	H	M	M	VH	H	928
6	L	H	H	?	VH	M	L	H	19	C	M	H	M	H	M	M	H	H	M	133(	H	M	H	H	H	M	M	H	H	H	90
6	L	M	L	?	M	M	L	M	36	C	H	H	M	M	M	M	M	M	L	8581	VH	L	H	M	VH	H	M	VH	VH	H	881
6	VL	H	H	?	H	M	H	M	4.8	C	H	H	M	H	M	H	H	H	H	4410	M	H	H	H	L	M	M	H	M	H	29
7	L	H	H	?	L	L	H	H	0.9	C	H	VH	H	H	H	M	H	H	H	1308	M	H	H	M	VL	VH	H	VH	L	H	31
7	L	M	H	?	L	L	L	H	6	C	H	VH	H	H	H	M	H	H	H	7109	M	M	H	?	L	VH	H	VH	M	H	148
8	L	H	M	?	M	M	L	M	99	C	H	H	M	H	H	M	M	H	H	248(	H	M	M	H	M	VH	H	M	M	M	1597
8	L	H	VH	?	H	VH	VH	VH	2.5	'VC+	VH	VH	H	VH	H	VH	VH	VH	VH	2192	M	VH	H	VH	L	VH	M	VH	L	H	17
8	M	H	H	?	H	M	M	VH	201	C	M	H	M	H	H	M	VH	M	H	139(	VH	H		H	H	H	M	VH	M	M	616
9	M	H	H	L	M	M	H	?	14	C	M	H	M	H	M	H	M	H	H	9441	VH	M	H	VH	M	VH	?	VH	M	M	167
9	?	H	H	?	H	H	H	H	53.9	C	H	H	H	VH	H	H	VH	VH	VH	1817	VH	H		VH	H	VH	VH	VH	H	VH	209
9	VL	H	H	?	H	M	H	M	5.8	C	M	M	M	H	M	H	H	H	H	8822	M	M		H	L	H	L	H	M	M	53
10	?	H	H	?	H	H	H	H	58.3	C	H	H	H	H	H	H	H	H	H	3347	H	H	H	H	H	H	H	H	VH	H	672

Repeated result: only sqrt(cols) and row/10
e.g. 900 cells (total) but 64 cells (in the "corner")
900 - 64 / 900 = 93% data with 100% privacy

3 laws of trusted data sharing

1) only share corners
2) mutate corners before sharing
3) never mutate beyond halfway to nearest neighbor



reverse NN	infogain																																	
	0.09	0.10	0.14	0.14	0.25	0.26	0.29	0.32	0.32	0.33	0.39	0.41	0.43	0.49	0.51	0.55	0.56	0.57	0.60	0.64	0.65	0.66	0.69	0.80	0.81	0.82	0.90	0.92	0.93	0.98	bugs /loc
1	H	H	M	?	H	M	L	M	44	C	L	H	H	H	M	M	H	H	M	282(	VH	L	L	H	M	H	M	M	M	H	184
1	M	M	H	?	H	M	VL	H	154	C	H	VH	M	H	H	H	VH	M	H	348(	M	H	H	VH	H	VH	VH	H	H	H	1768
1	H	H	H	?	H	H	M	H	22	C	H	H	H	H	M	L	H	H	H	1874	VH	H	H	H	M	H	M	H	L	H	196
1	M	VL	M	M	M	H	M	M	23	C	M	M	L	H	H	M	M	M	H	6906	H	H	M	M	VL	H	M	?	H	H	546
2	VL	H	H	?	H	M	H	H	4.4	C	H	H	M	M	M	H	H	H	H	1415	H	M	H	H	H	M	H	H	H	H	71
3	M	M	M	?	M	L	M	L	33	C	H	M	H	M	L	H	M	M	M	1368	M	H	VH	H	VL	H	H	M	L	H	688
3	VL	H	VH	?	M	M	M	M	21	C	H	H	H	VH	H	H	VH	H	VH	1388	M	M	L	VH	VL	H	L	VH	L	H	204
3	M	M	H	?	H	M	M	H	26.7	C	H	VH	M	H	H	H	VH	M	H	7121	M	H	H	VH	VL	VH	H	H	L	H	109
3	L	H	H	?	VH	VH	VH	VH	11	C	H	M	H	H	H	M	VH	H	VH	3764	H	M	H	VH	M	H	M	VH	M	H	91
3	L	H	H	?	H	H	VH	H	1	C	H	M	H	H	H	H	VH	H	H	1976	M	H	H	VH	L	H	M	VH	L	H	5
4	VL	M	H	H	H	L	L	M	49.1	C	H	M	M	M	L	M	H	H	H	2545	H	M	VH	M	H	M	H	M	H	H	129
4	L	H	H	?	H	M	M	M	87	C	H	M	H	H	H	H	H	H	H	1238	H	M	H	H	H	H	H	H	H	H	476
4	H	H	H	?	M	M	VL	H	155.2	C	H	VH	M	H	H	H	VH	M	H	323(	H	H	M	VH	VH	VH	VH	H	VH	H	1906
5	L	M	L	H	M	M	M	M	52	C	M	H	M	M	M	M	M	M	L	146(	VH	M	H	H	VH	H	VH	VH	H	H	412
5	L	H	H	M	M	M	M	M	713.6	C	M	H	M	H	M	M	M	M	M	1001	H	H	VH	H	VH	M	M	H	H	H	4223
5	H	M	H	?	H	M	M	H	33	C	H	H	M	VH	M	M	VH	H	H	156(	H	H	?	VH	M	VH	H	H	M	H	653
5	M	M	H	L	L	L	M	M	14	C	H	H	M	M	M	H	M	H	M	9434	VH	M	H	H	H	VH	?	H	H	H	373
6	H	H	H	?	L	H	M	H	61	C	H	H	H	H	H	L	H	H	H	539(	VH	H	M	M	VH	H	M	VH	H	H	680
6	H	VL	L	?	VL	VH	VL	VL	50	C	L	H	H	H	VL	L	H	M	VL	526(	VH	L	L	M	H	H	M	M	VH	H	928
6	L	H	H	?	VH	M	L	H	19	C	M	H	M	H	M	M	H	H	M	133(	H	M	H	H	H	M	M	H	H	H	90
6	L	M	L	?	M	M	L	M	36	C	H	H	M	M	M	M	M	M	L	8581	VH	L	H	M	VH	H	M	VH	VH	H	881
6	VL	H	H	?	H	M	H	M	4.8	C	H	H	M	H	M	H	H	H	H	4410	M	H	H	H	L	M	M	H	M	H	29
7	L	H	H	?	L	L	H	H	0.9	C	H	VH	H	H	H	M	H	H	H	1308	M	H	H	M	VL	VH	H	VH	L	H	31
7	L	M	H	?	L	L	L	H	6	C	H	VH	H	H	H	M	H	H	H	7109	M	M	H	?	L	VH	H	VH	M	H	148
8	L	H	M	?	M	M	L	M	99	C	H	H	M	H	H	M	M	H	H	2485	H	M	M	H	M	VH	H	M	M	M	1597
8	L	H	VH	?	H	VH	VH	VH	2.5	'VC+	VH	VH	H	VH	H	VH	VH	VH	VH	2192	M	VH	H	VH	L	VH	M	VH	L	H	17
8	M	H	H	?	H	M	M	VH	201	C	M	H	M	H	H	M	VH	M	H	139(	VH	H		H	H	H	M	VH	M	M	616
9	M	H	H	L	M	M	H	?	14	C	M	H	M	H	M	H	M	H	H	9441	VH	M	H	VH	M	VH	?	VH	M	M	167
9	?	H	H	?	H	H	H	H	53.9	C	H	H	H	VH	H	H	VH	VH	VH	1817	VH	H		VH	H	VH	VH	VH	H	VH	209
9	VL	H	H	?	H	M	H	M	5.8	C	M	M	M	H	M	H	H	H	H	8822	M	M		H	L	H	L	H	M	M	53
10	?	H	H	?	H	H	H	H	58.3	C	H	H	H	H	H	H	H	H	H	3347	H	H	H	H	H	H	H	H	VH	H	672

Repeated result: only sqrt(cols) and row/10
e.g. 900 cells (total) but 64 cells (in the "corner")
900 - 64 / 900 = 93% data with 100% privacy

FYI, predictions from "corners" = predictions from the whole (for SE data) Vasil, WVU, 2013

3 laws of trusted data sharing

1) only share corners
2) mutate corners before sharing
3) never mutate beyond halfway to nearest neighbor

The Sharing Experiment

Peters, ICSE'15
20 stakeholders

1. Pass around a cache, in random order
2. Stakeholders add data not in cache
3. Before sharing, apply 3 laws

Results

4. Shared: 5% of data
5. Predictions: good

reverse NN	infogain																																	
	0.09	0.10	0.14	0.14	0.25	0.26	0.29	0.32	0.32	0.33	0.39	0.41	0.43	0.49	0.51	0.55	0.56	0.57	0.60	0.64	0.65	0.66	0.69	0.80	0.81	0.82	0.90	0.92	0.93	0.98	bugs /loc
1	H	H	M	?	H	M	L	M	44	C	L	H	H	H	M	M	H	H	M	282(	VH	L	L	H	M	H	M	M	M	H	184
1	M	M	H	?	H	M	VL	H	154	C	H	VH	M	H	H	H	VH	M	H	348(	M	H	H	VH	H	VH	VH	H	H	H	1768
1	H	H	H	?	H	H	M	H	22	C	H	H	H	H	M	L	H	H	H	1874	VH	H	H	H	M	H	M	H	L	H	196
1	M	VL	M	M	M	H	M	M	23	C	M	M	L	H	H	M	M	M	H	6906	H	H	M	M	VL	H	M	?	H	H	546
2	VL	H	H	?	H	M	H	H	4.4	C	H	H	M	M	M	H	H	H	H	1415	H	M	H	H	H	M	H	H	H	H	71
3	M	M	M	?	M	L	M	L	33	C	H	M	H	M	L	H	M	M	M	1368	M	H	VH	H	VL	H	H	M	L	H	688
3	VL	H	VH	?	M	M	M	M	21	C	H	H	H	VH	H	H	VH	H	VH	1388	M	M	L	VH	VL	H	L	VH	L	H	204
3	M	M	H	?	H	M	M	H	26.7	C	H	VH	M	H	H	H	VH	M	H	7121	M	H	H	VH	VL	VH	H	H	L	H	109
3	L	H	H	?	VH	VH	VH	VH	11	C	H	M	H	H	H	M	VH	H	VH	3764	H	M	H	VH	M	H	M	VH	M	H	91
3	L	H	H	?	H	H	VH	H	1	C	H	M	H	H	H	H	VH	H	H	1976	M	H	H	VH	L	H	M	VH	L	H	5
4	VL	M	H	H	H	L	L	M	49.1	C	H	M	M	M	L	M	H	H	H	2545	H	M	VH	M	H	M	H	M	H	H	129
4	L	H	H	?	H	M	M	M	87	C	H	M	H	H	H	H	H	H	H	1238	H	M	H	H	H	H	H	H	H	H	476
4	H	H	H	?	M	M	VL	H	155.2	C	H	VH	M	H	H	H	VH	M	H	323(	H	H	M	VH	VH	VH	VH	H	VH	H	1906
5	L	M	L	H	M	M	M	M	52	C	M	H	M	M	M	M	M	M	L	146(	VH	M	H	H	VH	H	VH	VH	H	H	412
5	L	H	H	M	M	M	M	M	713.6	C	M	H	M	H	M	M	M	M	M	1001	H	H	VH	H	VH	M	M	H	H	H	4223
5	H	M	H	?	H	M	M	H	33	C	H	H	M	VH	M	M	VH	H	H	1565	H	H	?	VH	M	VH	H	H	M	H	653
5	M	M	H	L	L	L	M	M	14	C	H	H	M	M	M	H	M	H	M	9434	VH	M	H	H	H	VH	?	H	H	H	373
6	H	H	H	?	L	H	M	H	61	C	H	H	H	H	H	L	H	H	H	5395	VH	H	M	M	VH	H	M	VH	H	H	680
6	H	VL	L	?	VL	VH	VL	VL	50	C	L	H	H	H	VL	L	H	M	VL	526(	VH	L	L	M	H	H	M	M	VH	H	928
6	L	H	H	?	VH	M	L	H	19	C	M	H	M	H	M	M	H	H	M	1338	H	M	H	H	H	M	M	H	H	H	90
6	L	M	L	?	M	M	L	M	36	C	H	H	M	M	M	M	M	M	L	8581	VH	L	H	M	VH	H	M	VH	VH	H	881
6	VL	H	H	?	H	M	H	M	4.8	C	H	H	M	H	M	H	H	H	H	4410	M	H	H	H	L	M	M	H	M	H	29
7	L	H	H	?	L	L	H	H	0.9	C	H	VH	H	H	H	M	H	H	H	1308	M	H	H	M	VL	VH	H	VH	L	H	31
7	L	M	H	?	L	L	L	H	6	C	H	VH	H	H	H	M	H	H	H	7109	M	M	H	L	L	VH	H	VH	M	H	148
8	L	H	M	?	M	M	L	M	99	C	H	H	M	H	H	M	M	H	H	2485	H	M	M	H	M	VH	H	M	M	M	1597
8	L	H	VH	?	H	VH	VH	VH	2.5	'VC+	VH	VH	H	VH	H	VH	VH	VH	VH	2192	M	VH	H	VH	L	VH	M	VH	L	H	17
8	M	H	H	?	H	M	M	VH	201	C	M	H	M	H	H	M	VH	M	H	139(	VH	H	H	H	H	H	M	VH	M	M	616
9	M	H	H	L	M	M	H	?	14	C	M	H	M	H	M	H	M	H	H	9441	VH	M	H	VH	M	VH	?	VH	M	M	167
9	?	H	H	?	H	H	H	H	53.9	C	H	H	H	VH	H	H	VH	VH	VH	1817	VH	H	H	VH	H	VH	VH	VH	H	VH	209
9	VL	H	H	?	H	M	H	M	5.8	C	M	M	M	H	M	H	H	H	H	8822	M	M	H	H	L	H	L	H	M	M	53
10	?	H	H	?	H	H	H	H	58.3	C	H	H	H	H	H	H	H	H	H	3347	H	H	H	H	H	H	H	H	VH	H	672

Repeated result: only sqrt(cols) and row/10
e.g. 900 cells (total) but 64 cells (in the "corner")
900 - 64 / 900 = 93% data with 100% privacy

FYI, predictions from "corners" = predictions from the whole (for SE data) Vasil, WVU, 2013

In summary, to share data, use little data (the corners)

- Advantage: community evolves and shares lessons learned
- Research question: will this work for other kinds of data?

Data mining = data carving?

1. Find some cr*p
2. Cut the cr*p
3. Goto step 1

Example2: data “mining” for “optimization”

Motivation: Why are we Talking about Optimization?

tiny.cc/timcow15

Many SE activities are like optimization problems [Harman,Jones'01].

Due to computational complexity, exact optimization methods impractical for SE.

Use evolutionary metaheuristic methods to find near optimal or good-enough solutions

tiny.cc/timcow15

Motivation: Why are we Talking about Optimization?

1. Requirements	Menzies, Feather, Bagnall, Mansouri, Zhang
2. Transformation	Cooper, Ryan, Schielke, Subramanian, Fatiregun, Williams
3.Effort prediction	Aguilar-Ruiz, Burgess, Dolado, Lefley, Shepperd
4. Management	Alba, Antoniol, Chicano, Di Pentam Greer, Ruhe
5. Heap allocation	Cohen, Kooi, Srisa-an
6. Regression test	Li, Yoo, Elbaum, Rothermel, Walcott, Soffa, Kampfhamer
7. SOA	Canfora, Di Penta, Esposito, Villani
8. Refactoring	Antoniol, Briand, Cinneide, O'Keeffe, Merlo, Seng, Tratt
9. Test Generation	Alba, Binkley, Bottaci, Briand, Chicano, Clark, Cohen, Gutjahr, Harrold, Holcombe, Jones, Korel, Pargass, Reformat, Roper, McMinn, Michael, Sthamer, Tracy, Tonella,Xanthakis, Xiao, Wegener, Wilkins
10. Maintenance	Antoniol, Lutz, Di Penta, Madhavi, Mancoridis, Mitchell, Swift
11. Model checking	Alba, Chicano, Godefroid
12. Probing	Cohen, Elbaum
13. UIOs	Derderian, Guo, Hierons
14. Comprehension	Gold, Li, Mahdavi
15. Protocols	Alba, Clark, Jacob, Troya
16. Component sel	Baker, Skaliotis, Steinhofel, Yoo
17. Agent Oriented	Haas, Peysakhov, Sinclair, Shami, Mancoridis

Many SE activities are like optimization problems [Harman,Jones'01].

Due to computational complexity, exact optimization methods impractical for SE.

Use evolutionary metaheuristic methods to find near optimal or good-enough solutions

tiny.cc/timcow15

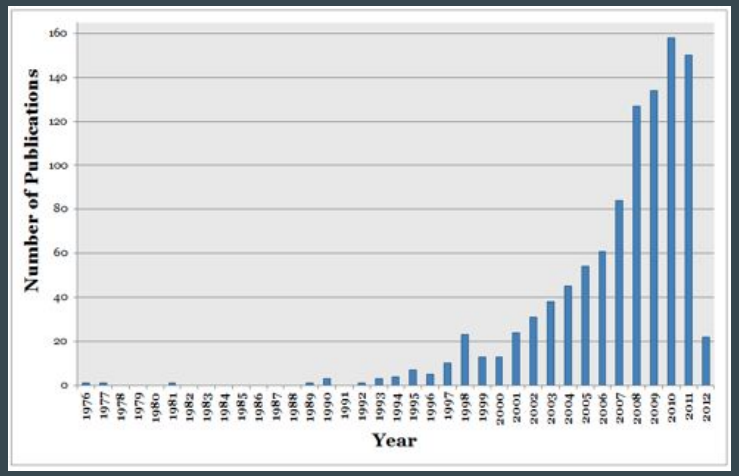
Motivation: Why are we Talking about Optimization?

1. Requirements	Menzies, Feather, Bagnall, Mansouri, Zhang
2. Transformation	Cooper, Ryan, Schielke, Subramanian, Fatiregun, Williams
3.Effort prediction	Aguilar-Ruiz, Burgess, Dolado, Lefley, Shepperd
4. Management	Alba, Antoniol, Chicano, Di Pentam Greer, Ruhe
5. Heap allocation	Cohen, Kooi, Srisa-an
6. Regression test	Li, Yoo, Elbaum, Rothermel, Walcott, Soffa, Kampfhamer
7. SOA	Canfora, Di Penta, Esposito, Villani
8. Refactoring	Antoniol, Briand, Cinneide, O'Keeffe, Merlo, Seng, Tratt
9. Test Generation	Alba, Binkley, Bottaci, Briand, Chicano, Clark, Cohen, Gutjahr, Harrold, Holcombe, Jones, Korel, Pargass, Reformat, Roper, McMinn, Michael, Sthamer, Tracy, Tonella,Xanthakis, Xiao, Wegener, Wilkins
10. Maintenance	Antoniol, Lutz, Di Penta, Madhavi, Mancoridis, Mitchell, Swift
11. Model checking	Alba, Chicano, Godefroid
12. Probing	Cohen, Elbaum
13. UIOs	Derderian, Guo, Hierons
14. Comprehension	Gold, Li, Mahdavi
15. Protocols	Alba, Clark, Jacob, Troya
16. Component sel	Baker, Skaliotis, Steinhofel, Yoo
17. Agent Oriented	Haas, Peysakhov, Sinclair, Shami, Mancoridis

Many SE activities are like optimization problems [Harman,Jones'01].

Due to computational complexity, exact optimization methods impractical for SE.

Use evolutionary metaheuristic methods to find near optimal or good-enough solutions



tiny.cc/tim15research

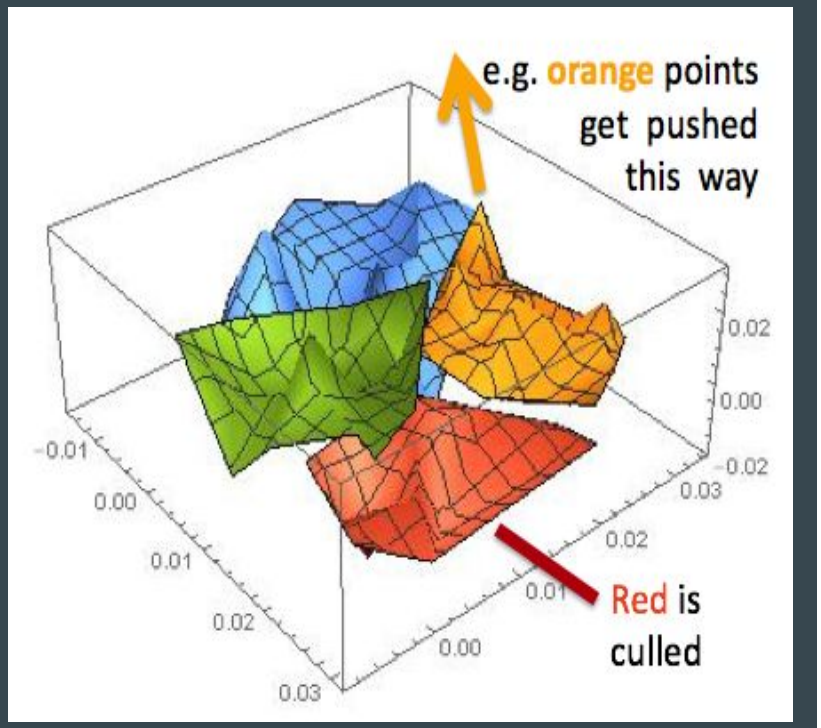
Optimization for Very Hard Problems

"Easy" if your problem continuous and differential and single goal
- Otherwise...

GALE: Krall, Menzies TSE 2015
- k=2 divisive clustering

function GALE():
1. (X,Y)= 2 very distant points found in O(2N) time
   - Euclidean distance in decision space
2. Evaluate only (X,Y)
3. If X "better" than Y
   If size(cluster) < sqrt(N) mutate towards X
   Else split, cull worst half, goto 1

Only $\log_2 N$ evaluations.



tiny.cc/tim15research

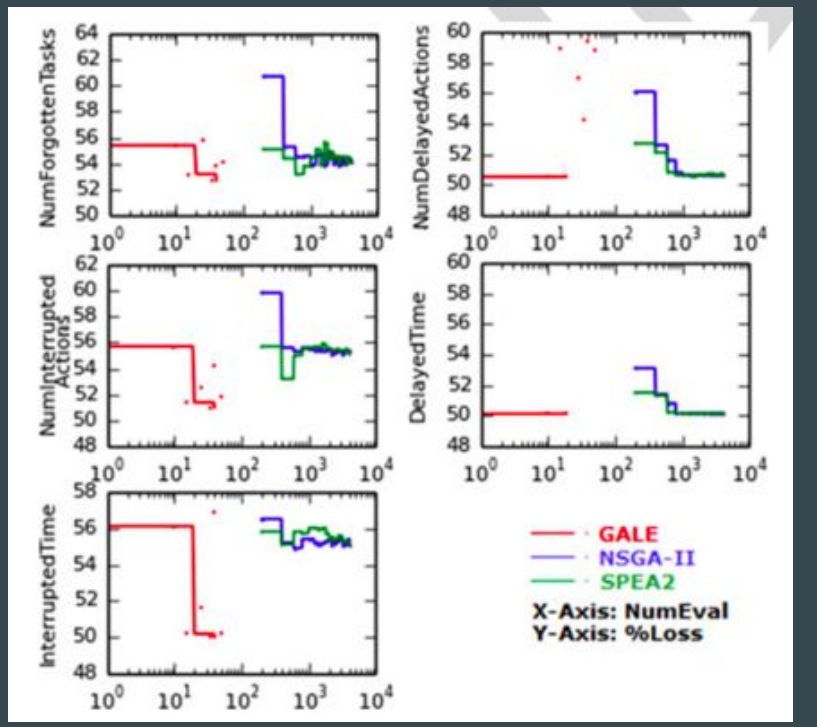
Optimization for Very Hard Problems

"Easy" if your problem continuous and differential and single goal
- Otherwise...

GALE: Krall, Menzies TSE 2015
- k=2 divisive clustering

function GALE():
1. (X,Y)= 2 very distant points found in O(2N) time
   - Euclidean distance in decision space
2. Evaluate only (X,Y)
3. If X "better" than Y
   If size(cluster) < sqrt(N) mutate towards X
   Else split, cull worst half, goto 1

Only $\log_2 N$ evaluations.



4 minutes, not 7 hours

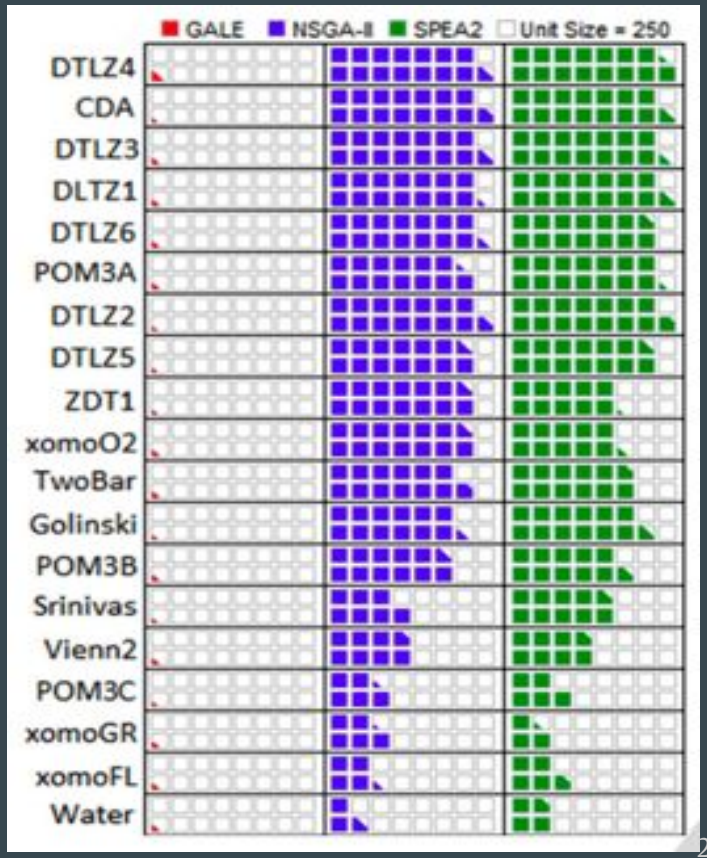
Optimization for Very Hard Problems

"Easy" if your problem continuous and differential and single goal
- Otherwise...

GALE: Krall, Menzies TSE 2015
- k=2 divisive clustering

function GALE():
1. (X,Y)= 2 very distant points found in O(2N) time
   - Euclidean distance in decision space
2. Evaluate only (X,Y)
3. If X "better" than Y
   If size(cluster) < sqrt(N) mutate towards X
   Else split, cull worst half, goto 1

Only $\log_2 N$ evaluations.



In summary, to optimize something faster, use little data

- Research question: will this work for other kinds of models?

Is the secret of big data little data?

Can we do “big data” different?

- Big Data
  - Volume
  - Velocity
  - Variety
- Use case
  - Central storage
  - Fast queries over all

- Little data
  - The amount of data you can conveniently store and process on a single machine

- Use case:
  - Everyone owns, stores, their own data
  - Passes out anonymised summaries, when asked
  - If you neighbor has something surprising then you talk more.

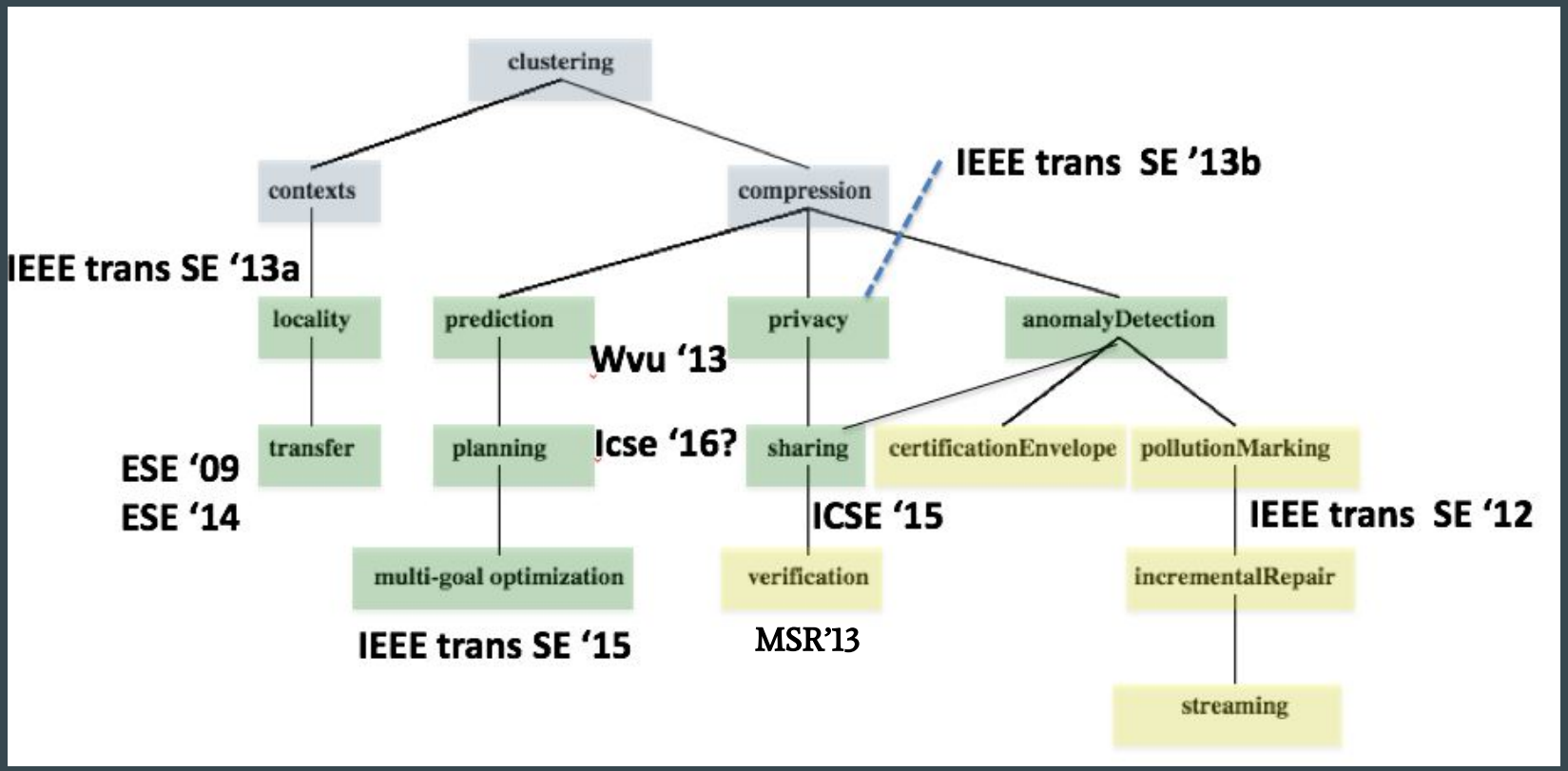
tiny.cc/tim15research

Back up slides

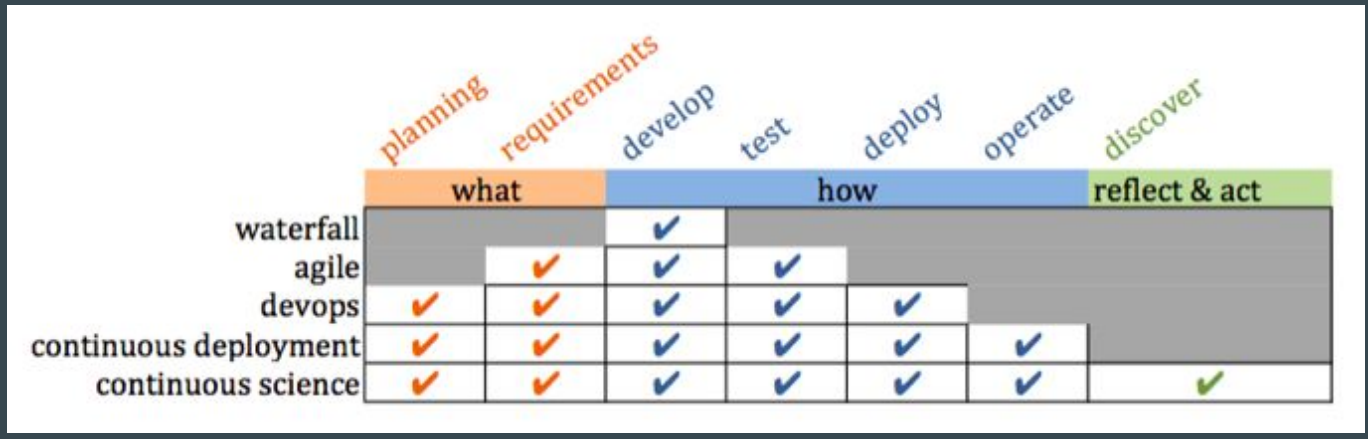
Engineering principles for small data

- Research question: will this work for other kinds of data?

http://www.slideshare.net/timmenzies/actionable-analytics-why-how
http://www.slideshare.net/timmenzies/future-se-oct15



Why expand our role?



- After continuous deployment:
  - Next gen SE = “continuous science”.
  - Services for data repositories supporting large teams running data miners

- NOW: we run the data miners
  - NEXT: we write tools that let other people run data miners... better

tiny.cc/tim15research

(My) Lessons from the PROMISE project

more data

More data does not actually help
- increases variance in conclusions
- need to reason within data clusters
- Menzies TSE'13 (local vs global)
- IST '13, 55(8), Promise issue

software project data

Conclusions that hold for all, may not hold for one (so beware SLRs)
- Posnett et al. ASE'11

Not general models, but general methods for finding local models
- Menzies TSE'13 (local vs global)
- IST '13, 55(8), Promise issue

Context best uncovered automatically, not specified manually.
- Menzies TSE'13 (local vs global)
- Kocaguneli ESEM'11

effort estimation

Humans rarely use lessons from past projects to improve their future reasoning
- Jørgensen TSE, 2009
- Passos ESEM'11

"Size" metrics useful, but not essential for accurate estimates
- Kocaguneli Promise'12

Model-based effort estimation, New high water mark:
- Choetkiertikul et al. ASE'15

tiny.cc/tim15research

(My) Lessons from the PROMISE project

no "best" model

Ensembles rule
(N models beat one)
- Kocageunli TSE'12 (Ensemble)
- Minku IST'13 55(8)

data mining

Poor method to confirm hypothesis

Good method to refute hypothesis
(when target not in any model)

Great way to generate hypotheses
(user meetings: heh... that's funny)
- Inductive SE Manifesto
  Menzies Malets'11

no "best" metrics

Best thing to do with data is to
throw most of it away
- Select sqrt(columns)
- Select sqrt(rows)
- So n^2 cells becomes $(n^{0.5})^2 = n$

Combine survivors, synthesize
dimensions (e.g. using WHERE).
Then cluster in synthesize space.
- Menzies TSE'13 (local vs global)

Can't assure that best models are
human comprehensible, or
contain initial expectations

All learners must be biased

No bias
⇒ no way to cull "dull" stuff
⇒ no summary
⇒ no model.
⇒ no predictions

So bias makes us blind, but bias
lets us see (the future).

Need learners that are biased
by the users' goals
- Menzies, Bener et al.
  ASE journal, 2010, 17(4)
- Krall, TSE 2015
- Minku, TOSEM'13

tiny.cc/tim15research

(My) Lessons from the PROMISE project

always re-learning

New data?
- Then, maybe, new model.

Not general models, but general methods for finding local models
- Menzies TSE'13 (local vs global)
- IST '13, 55(8), Promise issue

Conclusions that hold for all, may not hold for one (so beware SLRs)
- Posnett et al. ASE'11

no "best" prediction

Need to know range of outputs
- Then summarize the output
- Then try to pick inputs to minimize variance in output
- Jørgensen 2015, COW
- Menzies, ASE'07

goals, matter

Learners must be biased.

No bias? Then...
⇒ no way to cull "dull" stuff
⇒ no summary
⇒ no model.
⇒ no predictions

So bias makes us blind, but bias lets us see (the future).

Need learners that are biased by the users' goals
- Menzies, Bener et al. ASE journal, 2010, 17(4)
- Krall, TSE 2015
- Minku, TOSEM'13

Science has escaped the lab and roams free in the world

Every citizen can be a scientist (making generalizations from data)

1. Download a data mining toolkit
2. Run data miners to make conclusions

How to audit those conclusions?

Want to mistrust the conclusions of citizen scientists

- Just as we mistrust, evaluate, review, explore, evolve the conclusions of any other scientist.

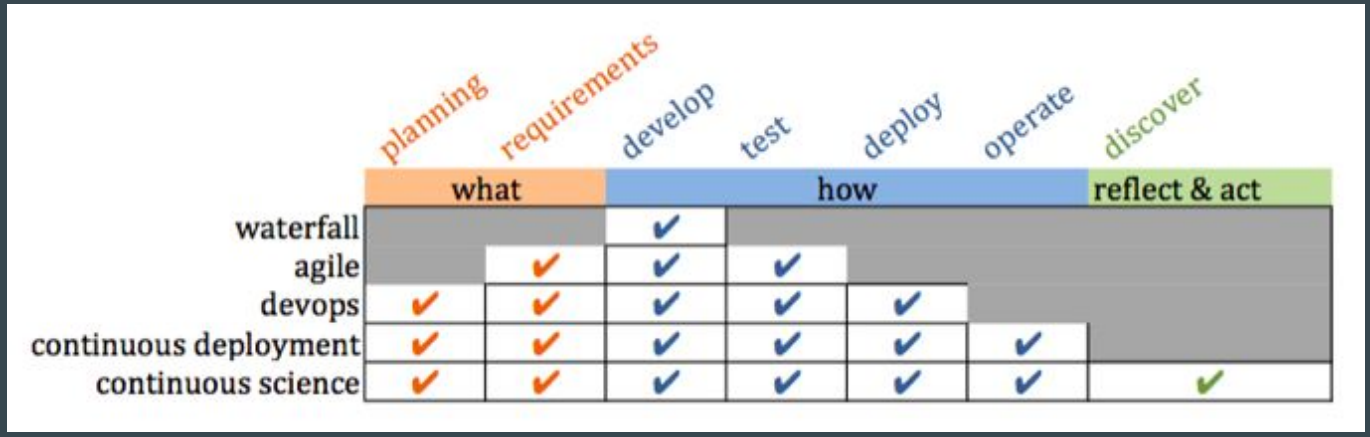
tiny.cc/tim15research

Q: What could these citizen scientists effect?
A: Everything

- Silicon valley developers view every new feature as an experiment, to be tested within some mash up.
- Chemists win Nobel Prize for software sims http://goo.gl/Lwensc
- Engineers use software to optical tweezers, radiation therapy, remote sensing, chip design, http://goo.gl/qBMyIZ
- Web analysts use software to analyze clickstreams to improve sales and marketing strategies; http://goo.gl/b26CfY
- Stock traders write software to simulate trading strategies http://www.quantopian.com
- Analysts write software to mine labor statistics data to review proposed gov policies http://goo.gl/X4kgnc
- Journalists use software to analyze economic data, make visualizations of their news stories http://fivethirtyeight.com
- In London or New York, ambulances wait for your call at a location determined by a software model http://goo.gl/8SMd1p
- Etc etc etc

tiny.cc/tim15research

Welcome to the next great challenge of SE (where SE = everything)



- After continuous deployment:
  - Next gen SE = "continuous science".
  - Services for data repositories supporting large teams running data miners

- NOW: we run the data miners
  - NEXT: we write software tools that let other people run data miners... better