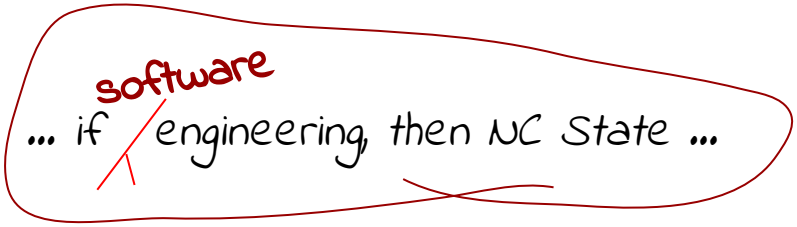


DSE = Data-Driven Search-Based SE

MSR'18, Gothenburg, Sweden

Vivek Nair, Amritanshu Agrawal, Jianfeng Chen
Wei Fu, George Mathew, **Tim Menzies**
Leandro Minku, **Markus Wagner**, Zhe Yu

UNIVERSITY OF LEICESTER
leandro.minku@le.ac.uk

THE UNIVERSITY of ADELAIDE
markus.wagner@adelaide.edu.au

NC STATE UNIVERSITY
timm@ieee.org

Why did these MSR people meet in Japan in Dec'17?

DSE = Data-Driven Search-based SE

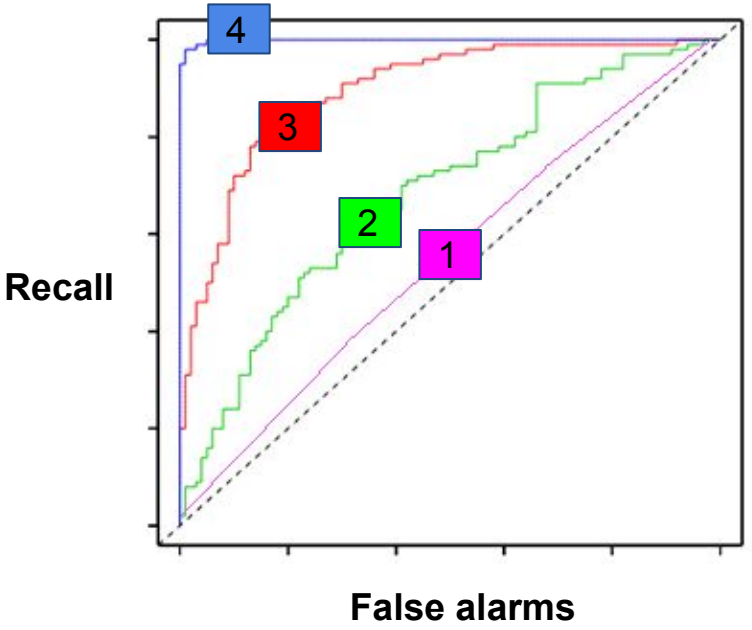
Search-based SE: highly acceptable at MSR

rank	topic	Comparing with the year 2017	submissions	accepted	acceptance rate
1	MSR on Multiple Projects	-	54	17	0.31
2	Tools and Techniques for MSR	-	46	14	0.3
3	Empirical studies	-	31	11	0.35
4	MSR with NLP	↑ 1	35	9	0.26
4	Change Patterns and Trends	↑ 4	23	9	0.39
5	Search-driven software development	↑ 3	20	7	0.35
5	Other	↑ 4	18	7	0.39
6	Software project evolution	↑ 3	19	6	0.32
6	Mining mobile app stores	↑ 2	12	6	0.5
7	Defect Analysis	↓ 3	23	5	0.22
7	Integrating Mined Data	↑ 2	12	5	0.42
8	Prediction with MSR	↓ 1	17	4	0.24
8	Social and development processes	↓ 1	14	4	0.29
9	PL features with MSR	↓ 2	11	4	0.36
9	Mining interaction data repositories	↓ 1	9	2	0.22
9	Bias in mining and guidelines	-	7	2	0.29
9	Sharing Data	-	6	2	0.33
9	Visualization	↑ 1	6	2	0.33
10	Mining code review repositories	↓ 3	9	1	0.11
10	Extracting New Forms of Data	↓ 4	9	1	0.11
10	Reliability and defect occurrences	↓ 1	6	1	0.17
10	Software licensing and copyrights	NEW	4	1	0.25
10	Energy aware mining	NEW	2	1	0.5
11	Mining execution traces and logs	↓ 4	4	0	0
11	Privacy and ethics	NEW	1	0	0

What is SBSE? (Search-based Software Engineering)

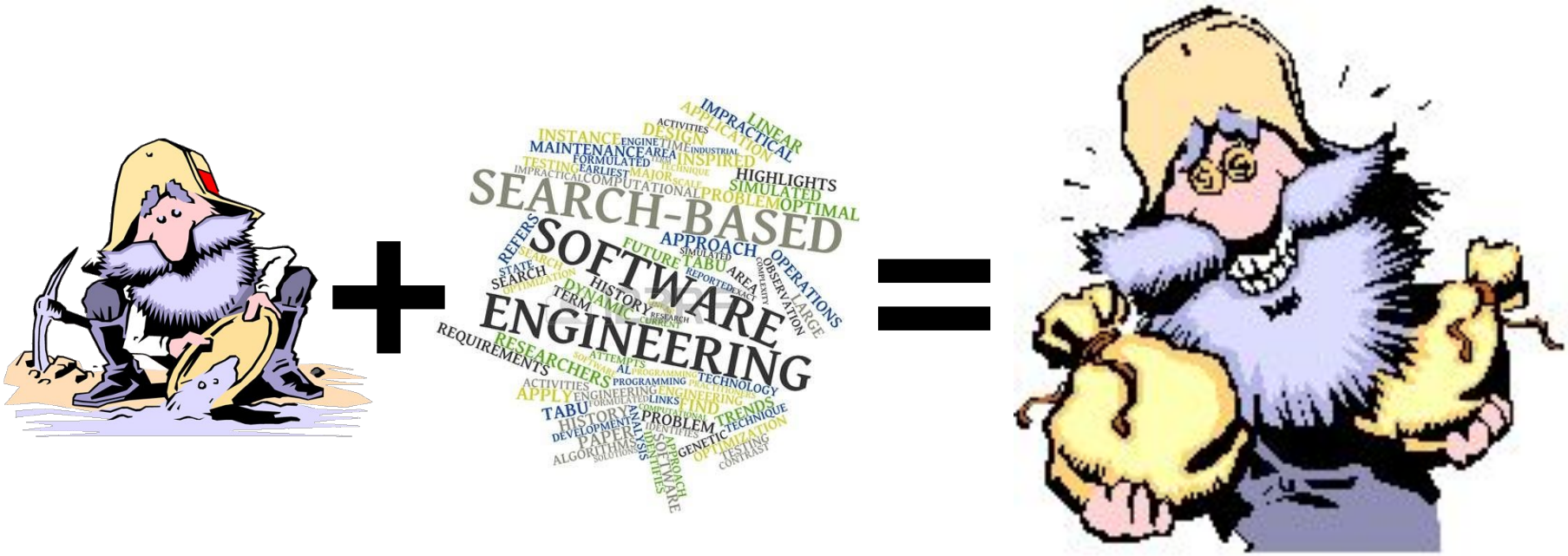
- Many SE activities are like optimization problems [Harman'02]

- Due to computational complexity, exact optimization methods are impractical



- Alternative: find good-enough solutions using meta-heuristic search as our optimizers
  - e.g. genetic algorithms
  - e.g simulated annealing
  - e.g. tabu search
  - e.g. NSGA-II, SPEA2, MOEA/D, Differential Evolution, Bayesian parameter optimization, etc etc

DSE = Data-Driven Search-based SE



- Conceptually, common higher level goal
  - supporting and giving insights to software engineers

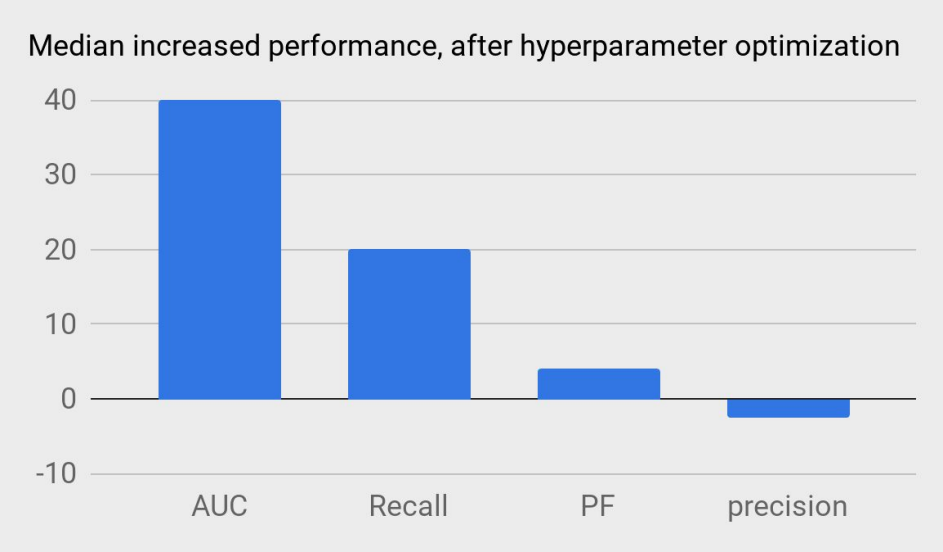
Data-Driven Search-based SE (DSE)

- To solve an SE problem:
  - Insert a data miner into an optimizer;
  - Or use an optimizer to improve a data miner.

- A new era for MSR (better MSR)
- A new era for SBSE (better SBSE)

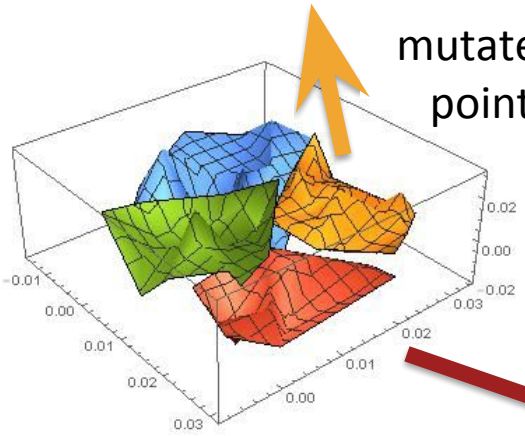
A new era for SBSE: Supercharging MSR

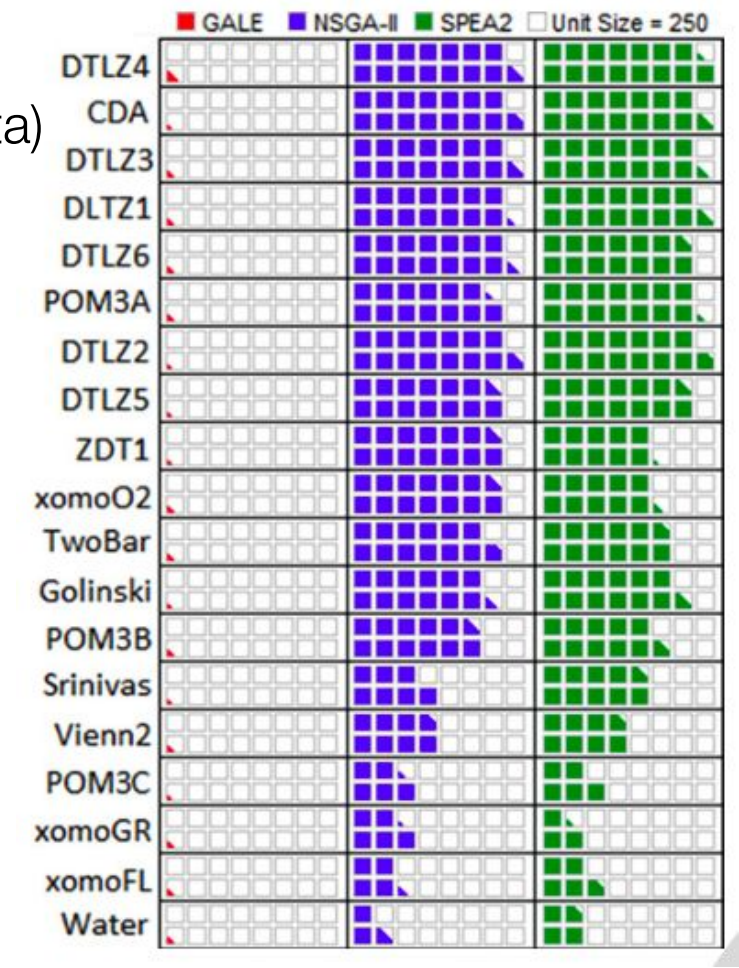
- Black art: hyperparameter optimization
- E.G. learning how many trees in a random forest
- E.G. learning how many “k” in kth-nearest neighbors
- Thanks to SBSE: massive improvements in, say, defect prediction
  - e.g. Agrawal & Menzies, ICSE 2018
  - performance details (after - before) tuning



A new era for SBSE: Let MSR help you run faster

- Landscape analysis
  - Find the lay of the land (shape of data)
  - Jump faster to better conclusions
  - e.g.. GALE, TSE 2015
- Note that this "optimizer" is really a "data miner"
  - clustering, PCA



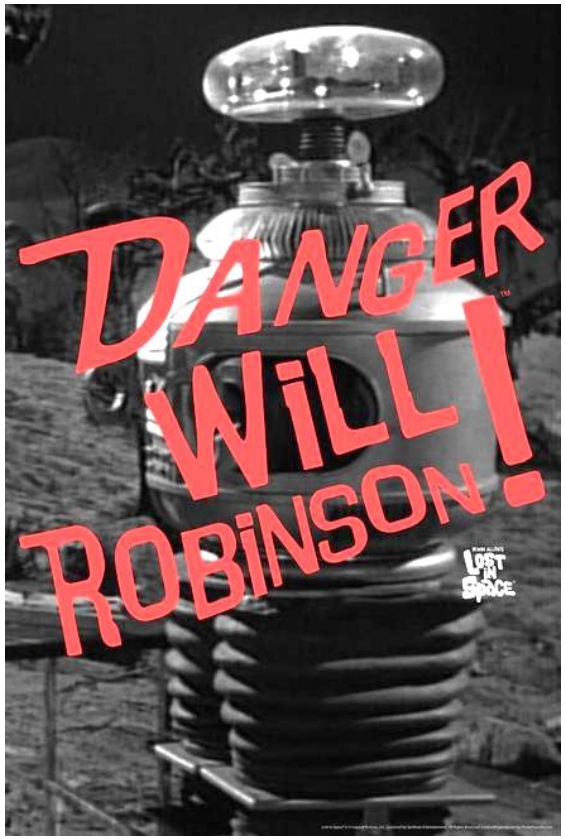


Q: Why explore MSR+SBSE?
A: So many application areas

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, rold, Holcombe, Jones, Korel, Pargass, Reformat, Rope, McMinn, M, r, Trac, Tonella,Xanthakis, Xiao, Wegener, Wilkin
10. Maintenance	Antoniol, Lutz, Di Penta, Madhavi, M
11. Model checking	Alba, Chicano, Godefroid
12. Probing	Cohen, Elbaum
13. Comprehension	Gold, Li, Mahdavi
14. Protocols	Alba, Clark, Jacob, Troya
15. Component sel	Baker, Skaliotis, Steinhofel, Yoo
16. Agent Oriented	Haas, Peysakhov, Sinclair, Shami, Mancoridis

so many novel contributions to so many areas

Q: Why explore MSR+SBSE?
A2: cause you got to



- How to get a paper rejected (in 2020):
  - Publish data mining results **without** hyper-parameter optimization
- Coming to the end of "merely mining"
  - See debates on "unsupervised learning"
    - Too easy to just chase precision, recall etc
- Complex problems need complex inference
  - e.g.minimizing #false alarms before first defect [Huang et al.ICSME'17]
  - Needed to reply to (e.g.) [Parnin, Orso, Issta'11]

http://tiny.cc/data-SE: A new resource for MSR researchers
89 DSE artifacts, in 13 groups
(e.g. RE,software product lines, software processes)

existing results; useful for testing new methods

Domain	Problem	Decision Space	C/D	Projects	Description	Links	Related Work
MSR	Defect Prediction	Numeric	D	10	CK Metric	raise_data_defect	[29]
	Text Classification	Text	-	1 6 1	Citemap Pits StackOverflow	raise_data_pits raise_data_pits SOProcess	[1]
	Performance Optimization	Mixed	D	22	Performance Configuration optimization	raise_data_perf	[66, 67, 69]
SBSE	Software Product Lines	Boolean	D	5	Product Lines	raise_data_SPL	[15]
	General Optimization	Numeric	C	7 6	DTLZ ZDT	raise_dtlz_zdt raise_dtlz_zdt	[65]
	Workflow	Numeric	D	20	Workflow	raise_gen_workflow	[14]
	Text Discovery	Text	-	4	Reading Faster	raise_data_fastread	[103]
	Software Processes	Numeric	C	5 4	Xomo POM3	raise_pom_xomo	[16, 65]
	Requirement Engineering	Numeric	D	8	Requirement Engineering	raise_short	[57]

So now we know why all these MSR people are so interested in SBSE

- Thanks to organizers the Dec'17 NII Shonan Meeting
  - Data-Driven Search-based SE, Dec 11-14, 2017
  - Markus Wagner, Leandro Minku , Ahmed Hassan, John Clark

DSE = Data-Driven Search-based SE

- To solve an SE problem:
  - Insert a data miner into an optimizer;
  - Or use an optimizer to improve a data miner.

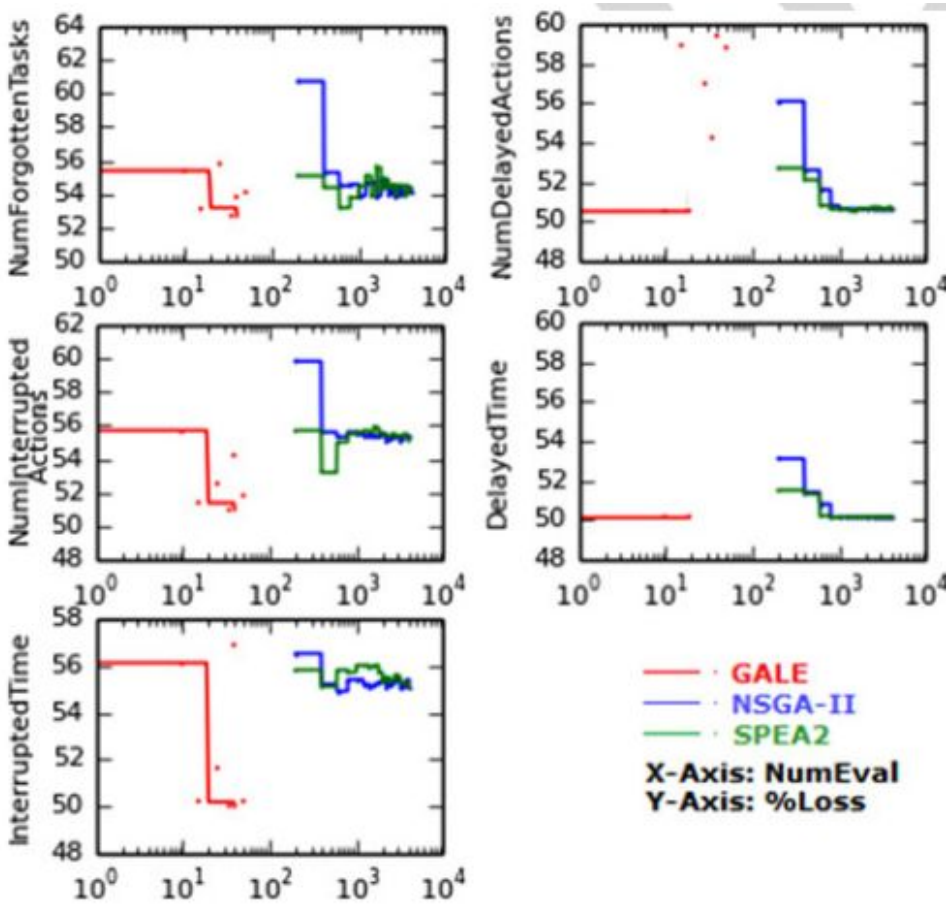
- A new era for MSR (better MSR)
- A new era for SBSE (better SBSE)

... if ~~software~~ engineering, then NC State ...

Back-up slides

A new era for MSR: Data farming (MSR + SBSE)

- Big data and massive Monte Carlo analysis
  - find important interactions
- domain intuitions ⇒
- model ⇒
  - generation += 1
  - simulation[i]
  - data
  - mining
  - insight
  - repeat



Q:Why explore farming data from models?
A: Cause models are everywhere

1. Silicon valley developers, new features are experiments, to be tested
2. Chemists win Nobel Prize for model sims http://goo.gl/Lwensc
3. Engineers test designs via models: radiation therapy, remote sensing, chip design, http://goo.gl/qBMylZ
4. Web analysts use models to analyze clickstreams to improve marketing: http://goo.gl/b26CfY
5. Stock traders use models to simulate trading strategies http://www.quantopian.com
6. Analysts review proposed gov policies via models of labor statistics data http://goo.gl/X4kgnc
7. Journalists use models to analyze economic data http://fivethirtyeight.com
8. In London or New York, ambulances wait at locations determined by a model http://goo.gl/8SMd1p
9. Etc etc etc

Why explore SBSE + MSR? (the carrot)

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, rold, Holcombe, Jones, Korel, Pargass, Reformat, Rope, McMinn, M, r, Trac, Tonella,Xanthakis, Xiao, Wegener, Wilkin
10. Maintenance Antoniol, Lutz, Di Penta, Madhavi, M
11. Model checking Alba, Chicano, Godefroid
12. Probing Cohen, Elbaum
13. Comprehension Gold, Li, Mahdavi
14. Protocols Alba, Clark, Jacob, Troya
15. Component sel Baker, Skaliotis, Steinhofel, Yoo
16. Agent Oriented Haas, Peysakhov, Sinclair, Shami, Mancoridis

so many novel contributions to so many areas

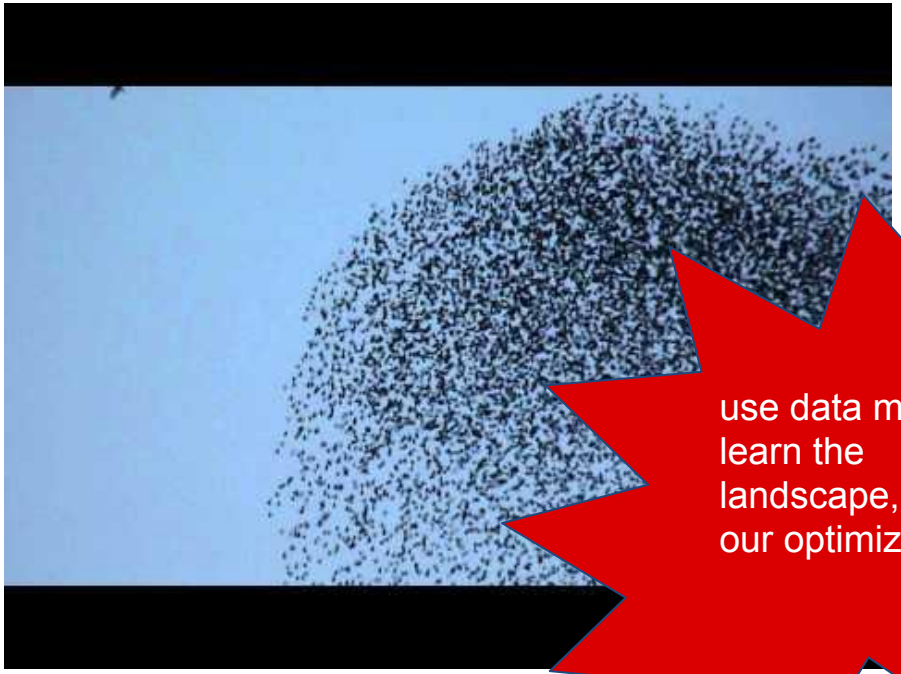
Some technical differences

	MSR	SBSE
Inference	induction, visualize	optimization
Speed	Faster, often more scalable	Becoming faster
Data	Collected before inference	Sampling controlled by inference
Tools	R, SciKitLearn, WEKA	jMetal, AutoWeka, AutoSklearn, Opt4j, DEAP
Example problems	• e.g. defect prediction; • StackOverflow mining	• minimize a test suite • configure software
Goals	e.g. just a few: recall, precision, MRE	• domain-specific goals. • meta-criteria (hypervolume, spread, IGD)

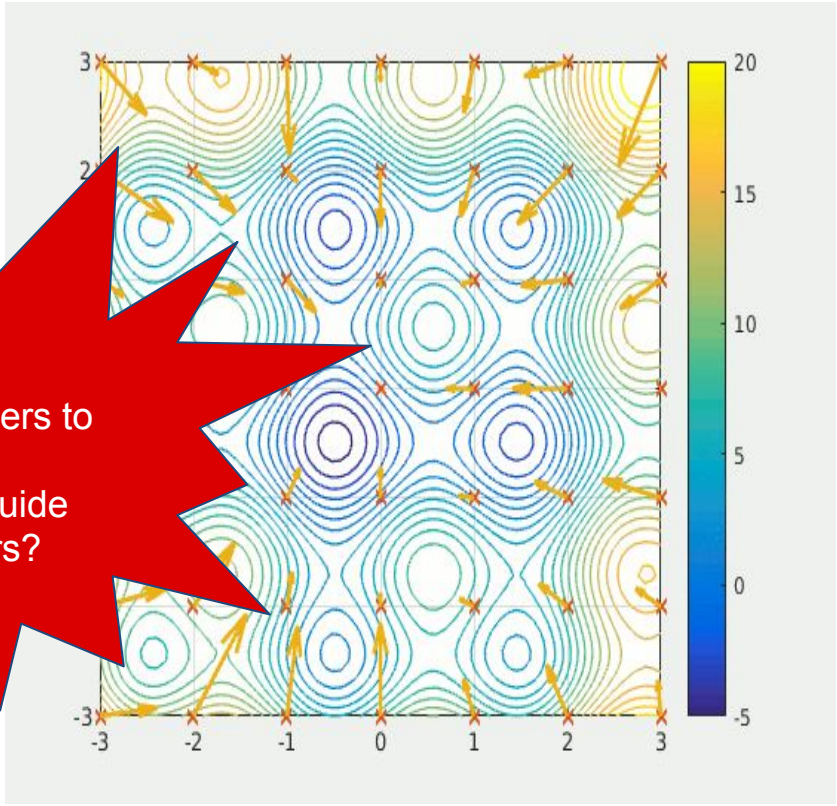
Optimization = surfing the landscape

murmuration of starlings
(learn safe "shapes" to avoid predators)

Particle Swarm Optimization:
new = old + φ1*rand(ourBest - now) ;; **social cognition**
+ φ2*rand(myBest - now)) ;; **private cognition**



use data miners to learn the landscape, guide our optimizers?

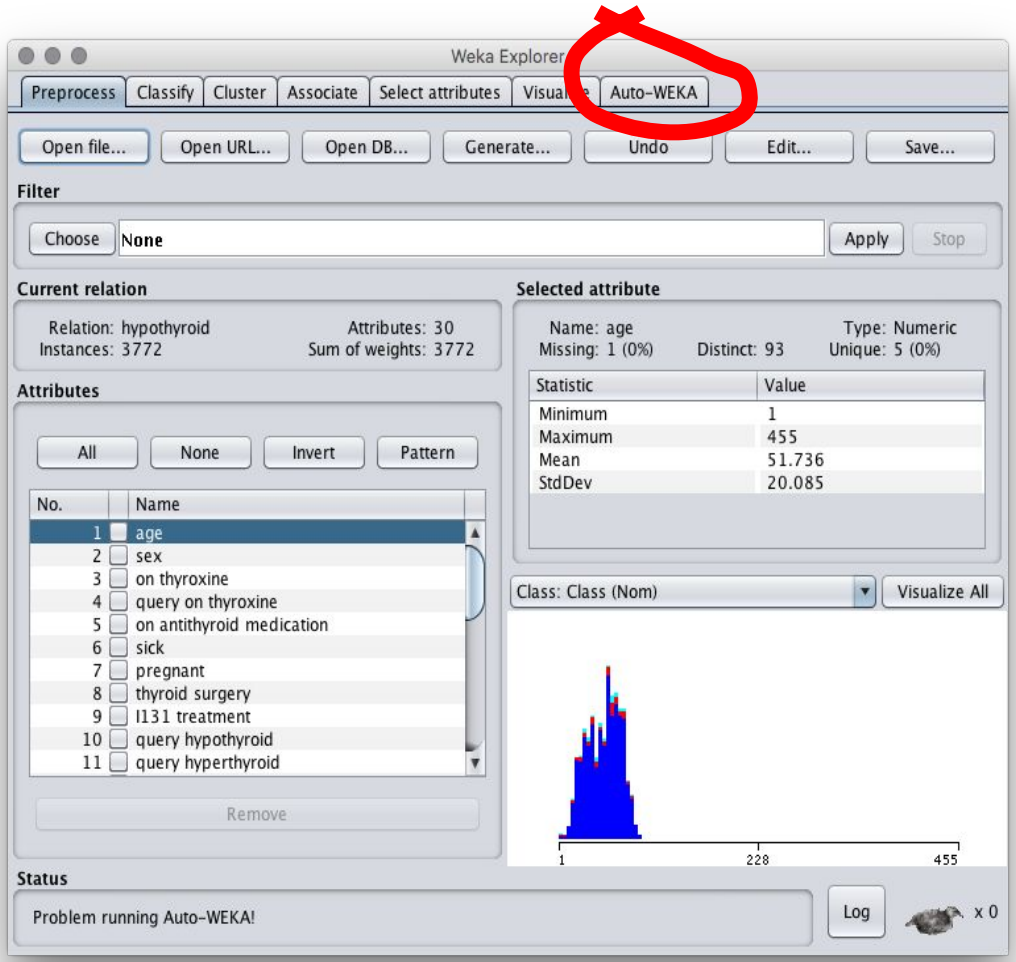


Something is changing. Things are …. different

Strange new words:

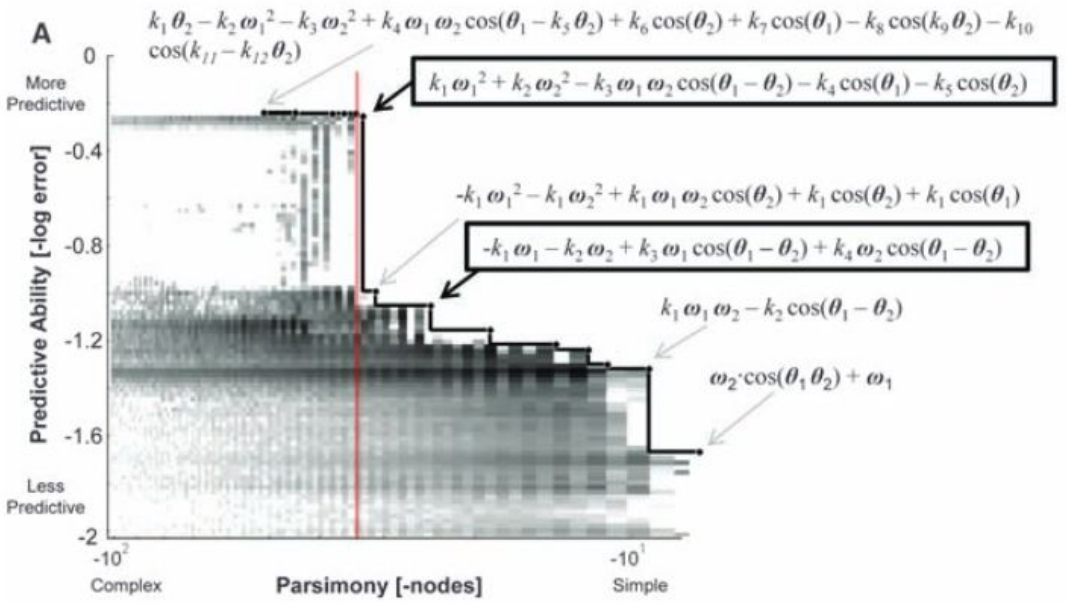
- "hyper-parameter optimization"
- "evolutionary algorithms"
- "differential evolution"
- "model-based reasoning"

What is going on?



MSR has much to gain from SBSE

- See paper, Fig4, long list of domain-specific goals
  - e.g. minimizing initial false before first defect [Huang et al., ICSME'17] [Parnin, Orso, Issta'11]
  - e.g. favor the shortest. most readable, model with least error
- Goals are domain-dependent
  - Need tools that adjust to different goals



Q: But why bother?
A: Cause much of SE is about choice

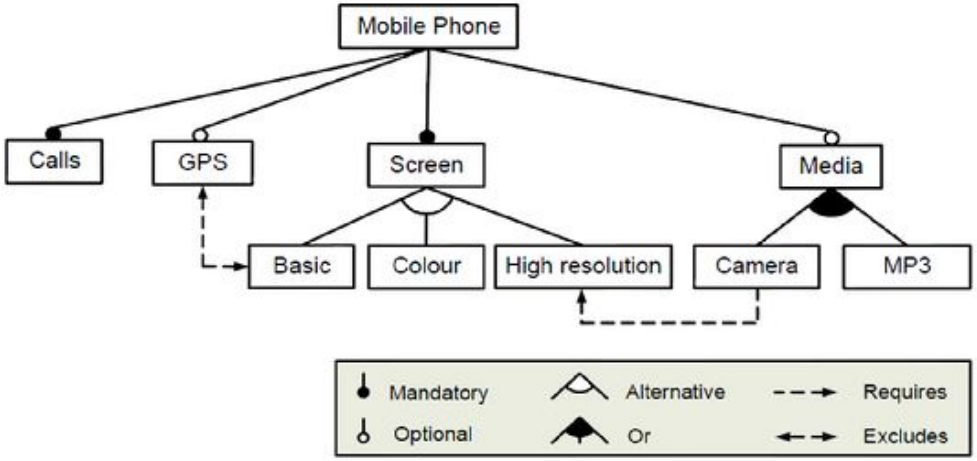


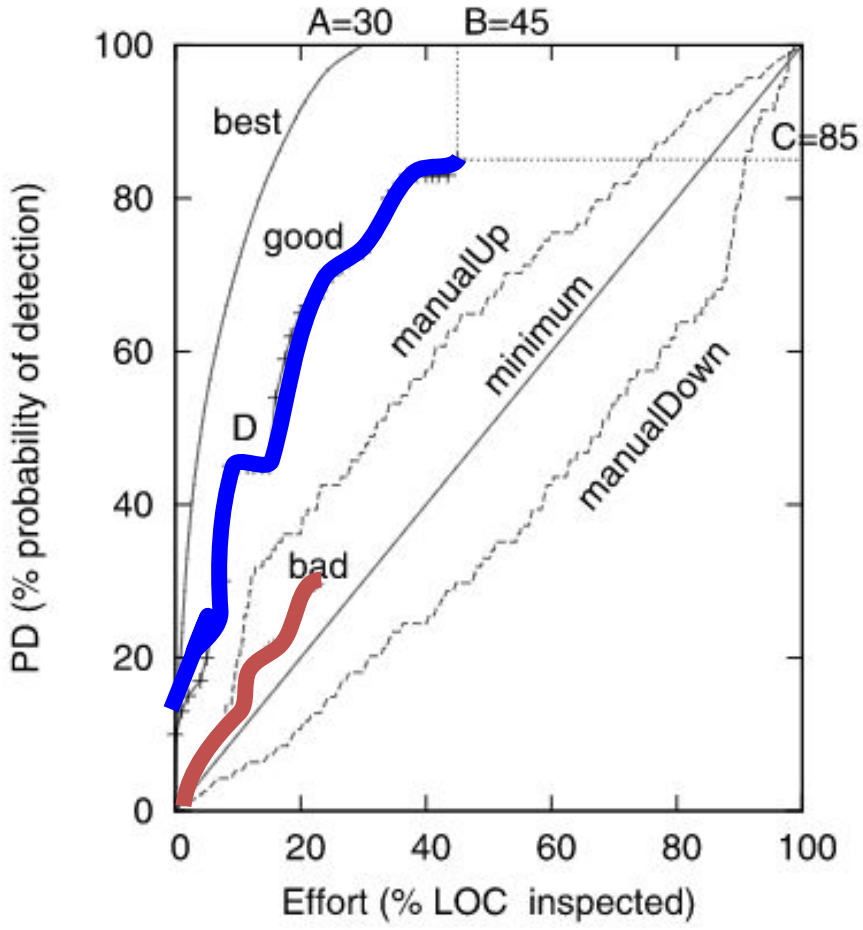
Figure 1: A sample feature model

¬Mobile Phone ∨ Calls
Mobile Phone ∨ ¬Calls
¬Mobile Phone ∨ Screen
Mobile Phone ∨ ¬Screen
Mobile Phone ∨ ¬GPS
Mobile Phone ∨ ¬Media
Media ∨ ¬Camera
Media ∨ ¬MP3
¬Media ∨ Camera ∨ MP3
Screen ∨ ¬Basic
Screen ∨ ¬Color
Screen ∨ ¬High resolution
¬Screen ∨ Basic ∨ Color ∨ High resolution
¬Basic ∨ ¬Color ∨ ¬High resolution
Basic ∨ ¬Color ∨ ¬High resolution
¬Basic ∨ Color ∨ ¬High resolution
¬Basic ∨ ¬Color ∨ High resolution
¬GPS ∨ ¬Basic
¬Camera ∨ High resolution

BTW: Linux kernel:
- 7000 terms
- 350,000 constraints

How does SBSE connect to MSR?

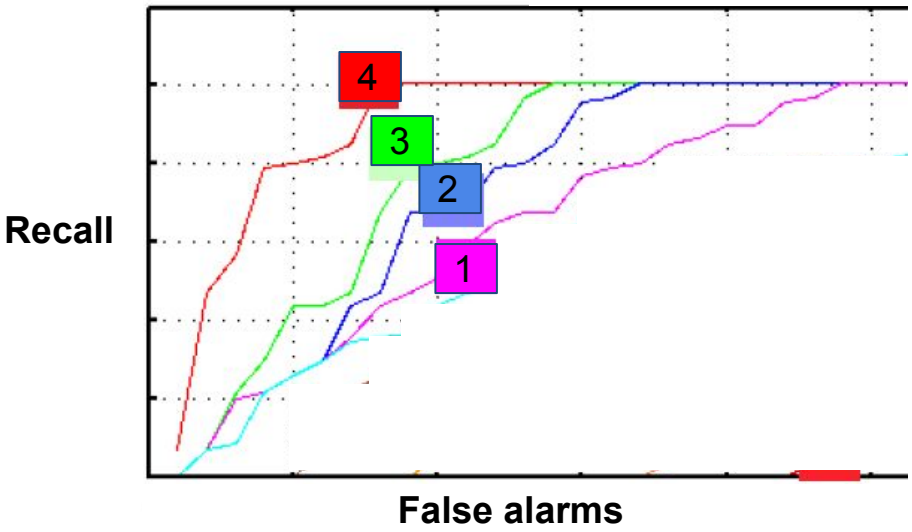
- Theoretically:
  - All learners build models that trade off competing goals
  - e.g. maximize recall, minimize false alarms
- Empirically:
  - better algorithms adjust themselves to the curves



What is SBSE? (Search-based Software Engineering)

- Many SE activities are like optimization problems [Harman'02]

- Due to computational complexity, exact optimization methods are impractical



- Alternative: find good-enough solutions using meta-heuristic search as our optimizers
  - e.g. genetic algorithms
  - e.g simulated annealing
  - e.g. tabu search
  - e.g. NSGA-II, SPEA2, MOEA/D, Differential Evolution, Bayesian parameter optimization, etc etc

A new era for MSR: Surfing cost-benefit decisions

- Exploring cost-benefit trade offs in software engineering
- e.g. learn tests that run fastest, most likely to fail
- e.g. as done manually by Elbaum et al, FSE'14
- e.g. as could be done automatically via SBSE