Predicting What Follows Predictive Modeling

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

UCL, Crest Open Workshop, Nov 23,24 2015
view : tiny.cc/timcow15
discuss: tiny.cc/timcow15discuss

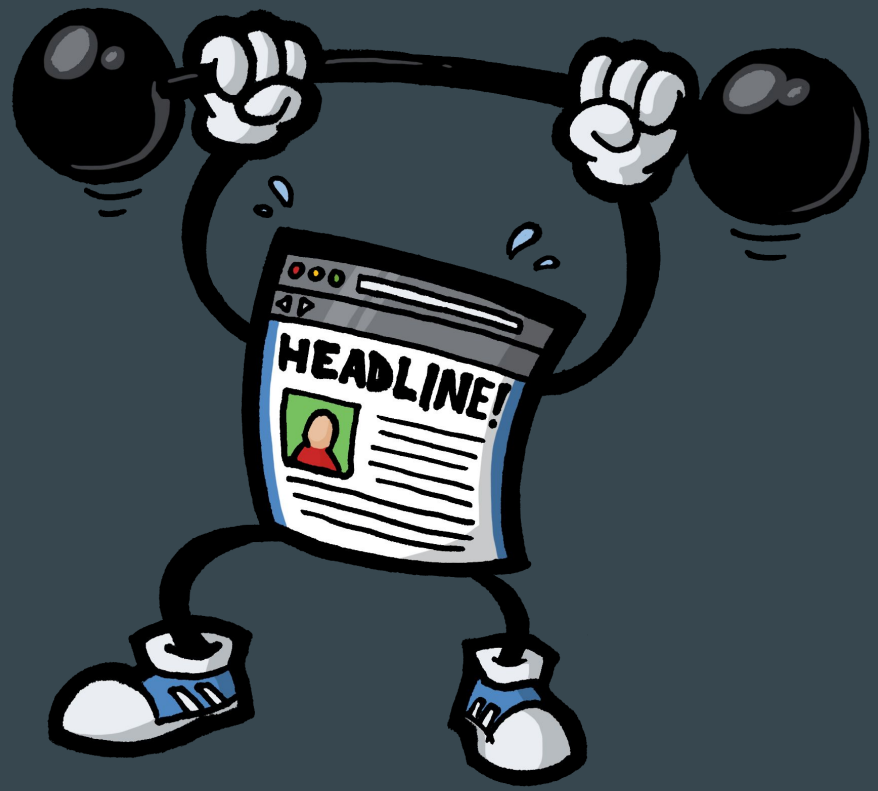
Tools to let other people run data miners… better

Sound bites

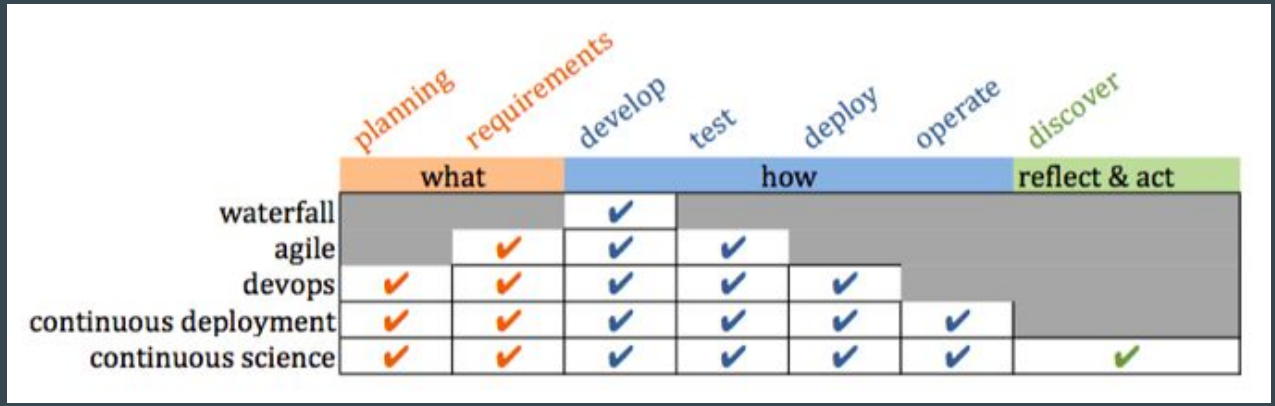
- "Prediction" = combination of many things
  - Can remix, reuse in novel ways

- Is it "prediction"?
  - Or "optimization" ?
    or "spectral learning" ?
    or "response surface methods" ?
    or "surrogate modeling"?
    or "local search"? or ...
  - or "finding useful quirks in the data"?

- Call it anything:
  - But expand your mind,
  - Refactor your tools,
  - Expand your role



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

Tools to let other people run data miners… better

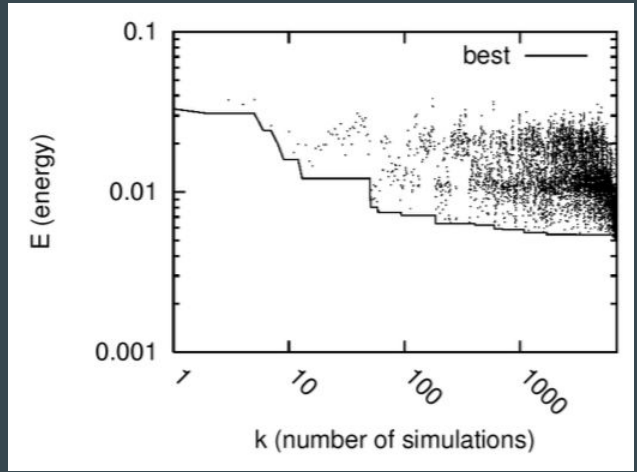
eg #1 : Helping Magne

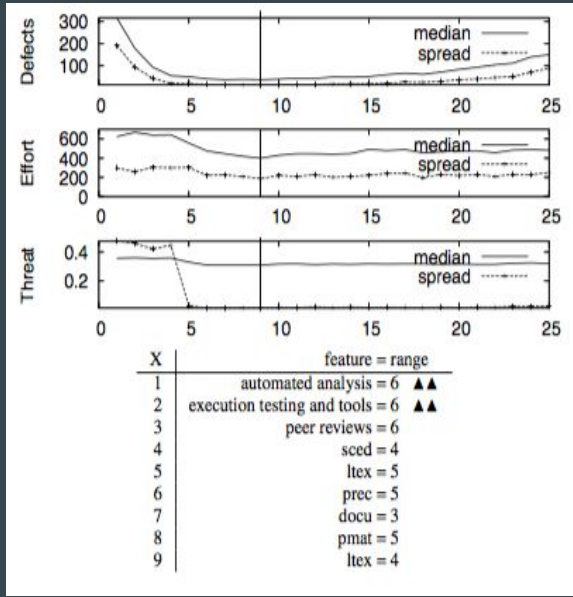
Models : useful for exploring uncertainty:

Menzies ASE'07, Gay ASE journal March'10

- The COQUALMO defect predictor [3, p254-268];
- The COCOMO effort predictor [3, p29-57];
- The THREAT predictor for effort & schedule overrun [3, 284-291].

id	features	relative weight
1	Personnel/team capability	3.53
2	Product complexity	2.38
3	Time constraint	1.63
4	Required software reliability	1.54
5	Multi-site development	1.53
6	Doc. match to life cycle	1.52
7	Personnel continuity	1.51
8	Applications experience	1.51
9	Use of software tools	1.50
10	Platform volatility	1.49
11	Storage constraint	1.46
12	Process maturity	1.43
13	Language & tools experience	1.43
14	Required dev. schedule	1.43
15	Data base size	1.42
16	Platform experience	1.40
17	Arch. & risk resolution	1.39
18	Precedentedness	1.33
19	Developed for reuse	1.31
20	Team cohesion	1.29
21	Development mode	1.32
22	Development flexibility	1.26





X	feature = range
1	automated analysis = 6 ▲▲
2	execution testing and tools = 6 ▲▲
3	peer reviews = 6
4	sced = 4
5	ltex = 5
6	prec = 5
7	docu = 3
8	pmat = 5
9	ltex = 4

eg #2: Helping Queens

Yesterday: 30 mins per optimizer?

Can we do better than that?

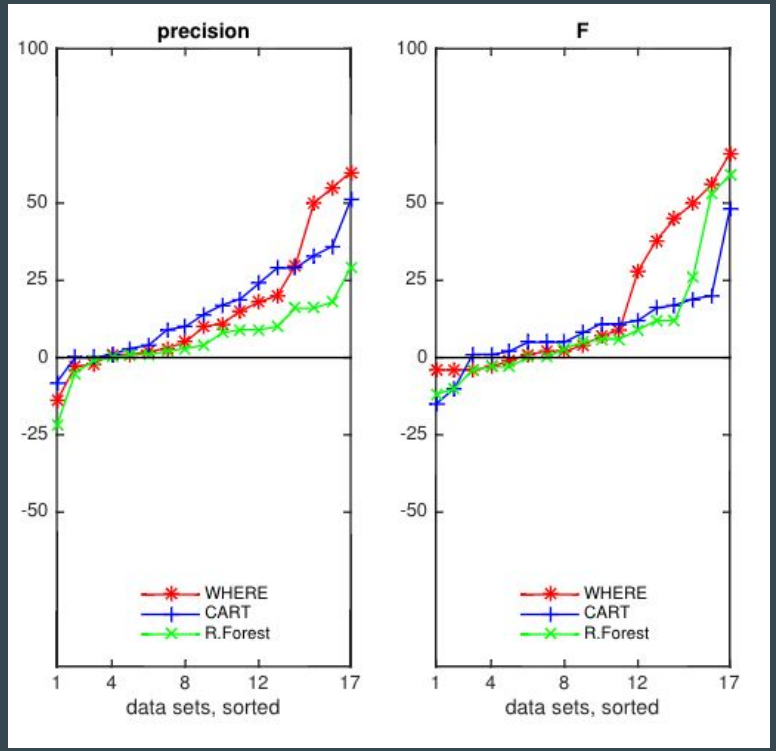
eg #2 : Helping Queens

Decision tree options =
#examples too split; #examples to stop, etc
(usually 6 settings per learner)

- Differential evolution (Storn 1995)

- frontier = Pick N options at random # e.g. N = 5
  R times repeat: # e.g. R = 10
  for Parent in frontier:
  - j,k,l = three other frontier items
  - Candidate = j + f * (k - l) # ish
  - if Candidate "better", replaces Parent

- Large improvements in defect prediction (Xalan, Jedit, Lucene, etc)
- For astonishingly little effort: seconds to run



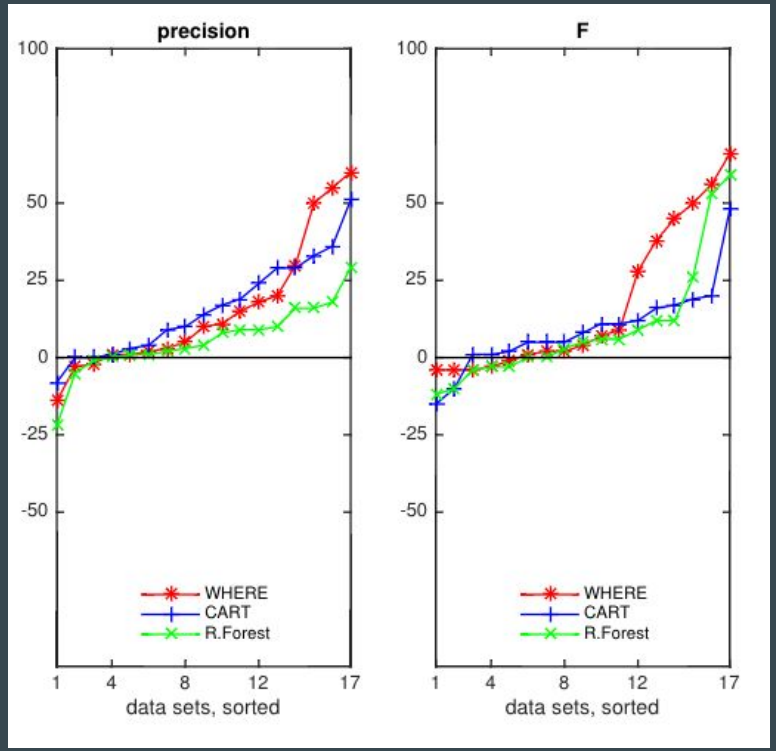
eg #2 : Helping Queens

Decision tree options =
#examples too split; #examples to stop, etc
(usually 6 settings per learner)

- Differential evolution (Storn 1995)

- frontier = Pick N options at random # e.g. N = 5
  R times repeat: # e.g. R = 10
  for Parent in frontier:
  - j,k,l = three other frontier items
  - Candidate = j + f * (k - l) # ish
  - if Candidate "better", replaces Parent

- Large improvements in defect prediction (Xalan, Jedit, Lucene, etc)
- For astonishingly little effort: seconds to run
-



No more prediction without pre-tuning study

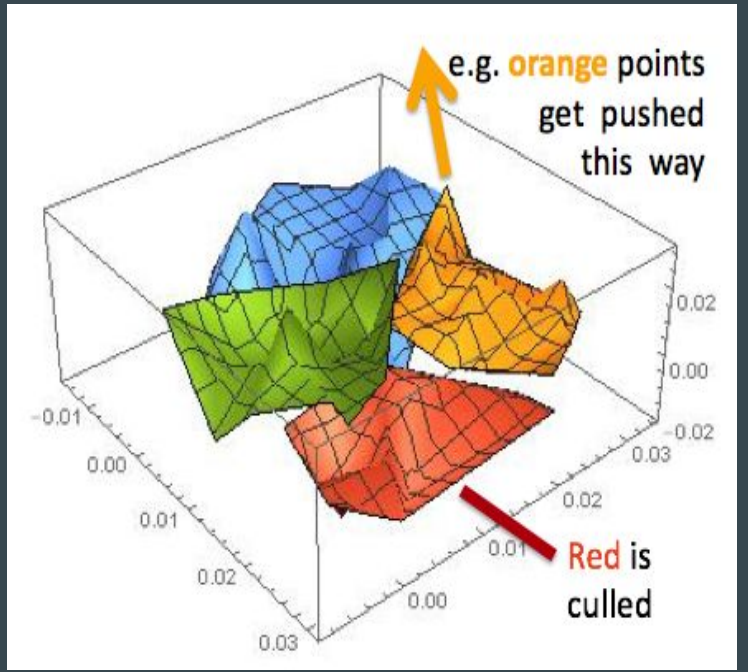
eg #3: helping tuning for HARDER problems

- 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.



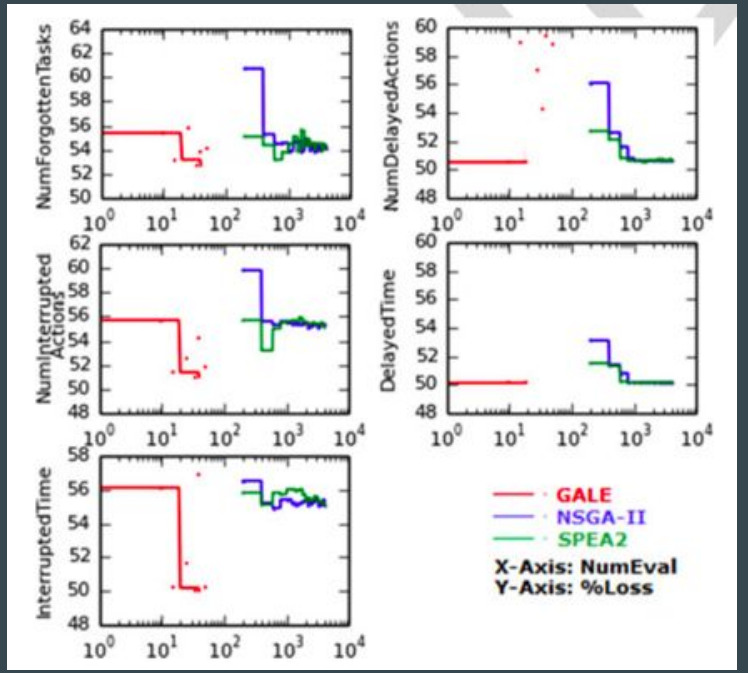
eg #3: helping tuning for HARDER problems

- 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

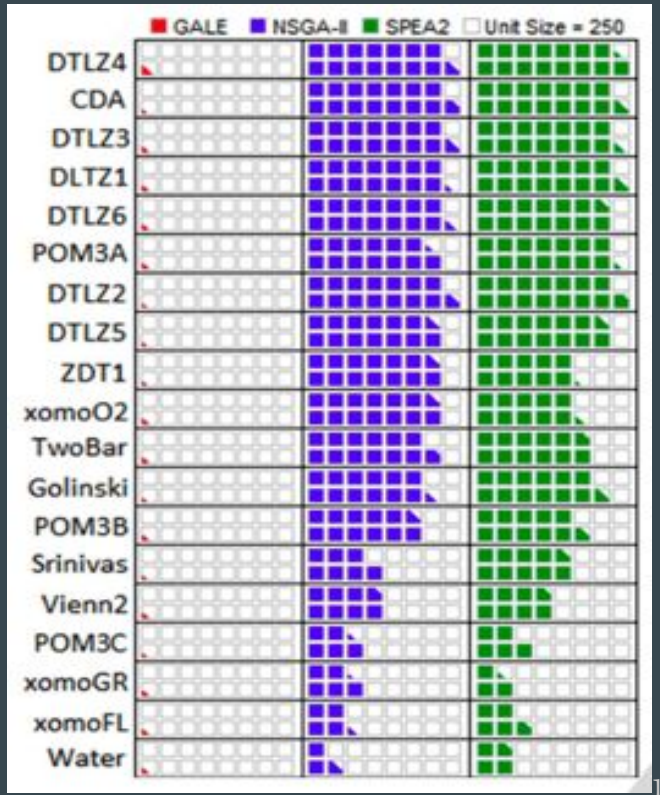
eg #3: helping tuning for HARDER problems

- 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

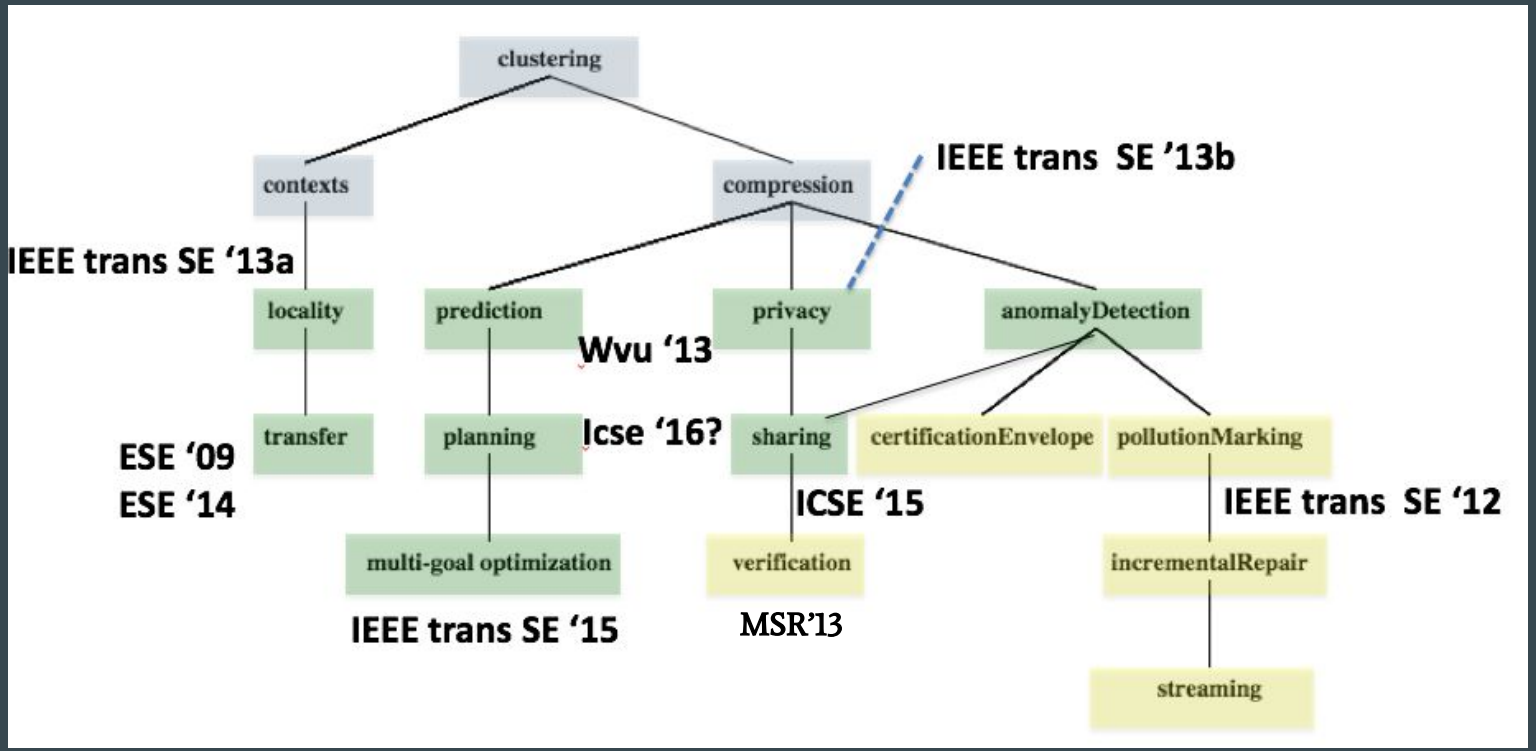
Only $\log_2 N$ evaluations.



And more...

...

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

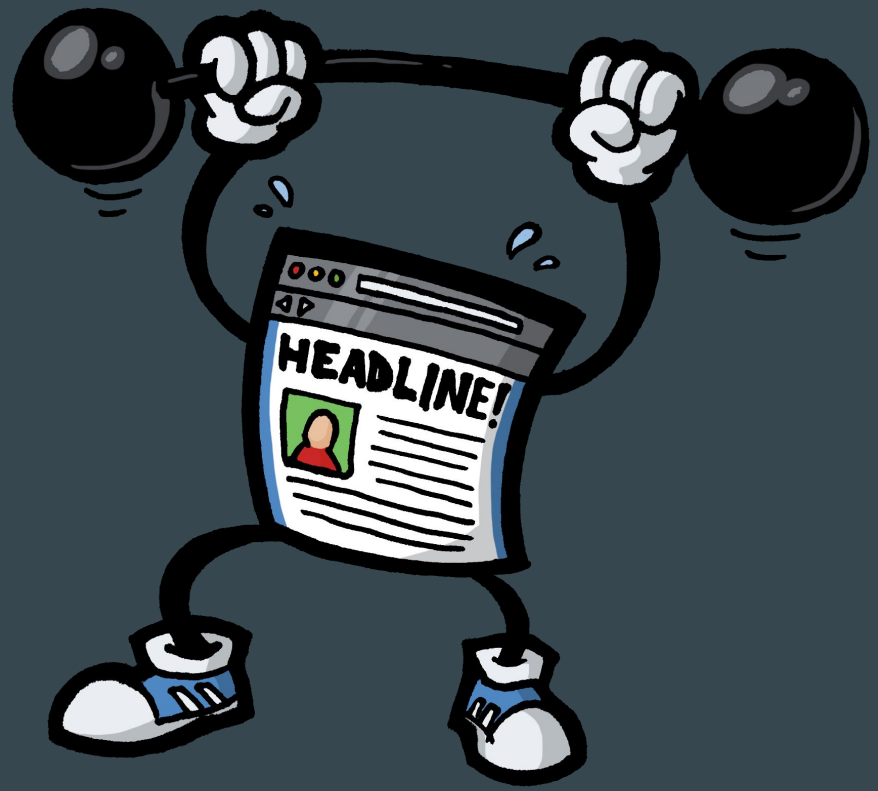


Sound bites

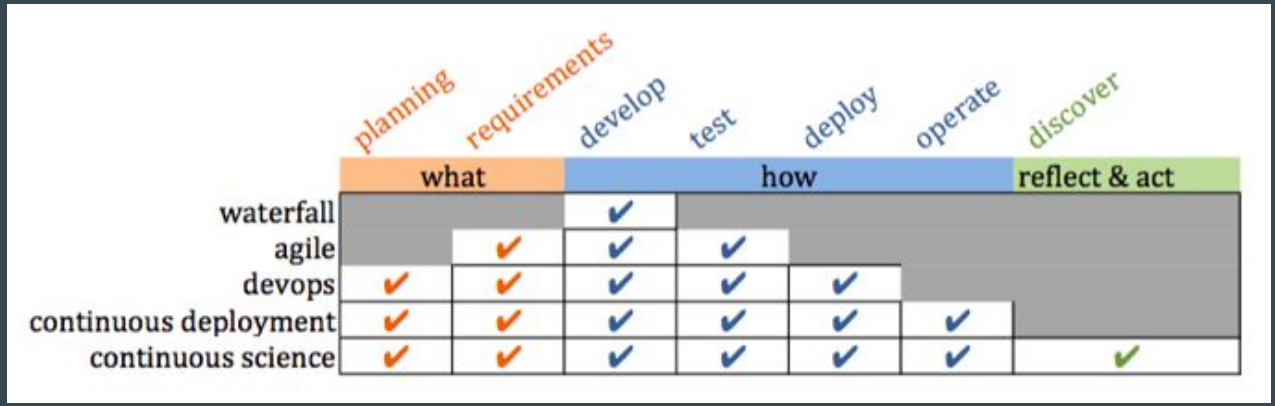
- "Prediction" = combination of many things
  - Can remix, reuse in novel ways

- Is it "prediction"?
  - Or "optimization" ?
    or "spectral learning" ?
    or "response surface methods" ?
    or "surrogate modeling"?
    or "local search"? or ...
  - or "finding useful quirks in the data"?

- Call it anything:
  - But expand your mind,
  - Refactor your tools,
  - Expand your role



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

Tools to let other people run data miners… better

tiny.cc/timcow15

Back up slides

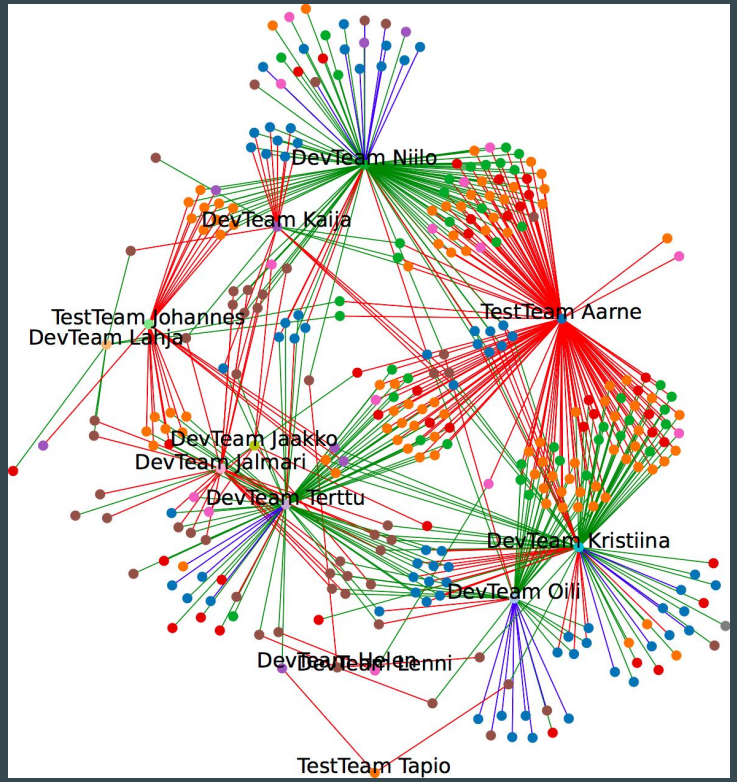
Next gen3: Insight generators

Less numbers, more insight

- Burak Turhan's "The graph"
  - circle = reported to
  - red = error report
  - green = error fix
  - blue = report+fix in the same team

More coarse grain control

- ("ontime", "aLittleLate", "wayOverdue")
- E.g.. Predicting delays in software projects using networked classification
- Choetkiertikul et al. ASE'15



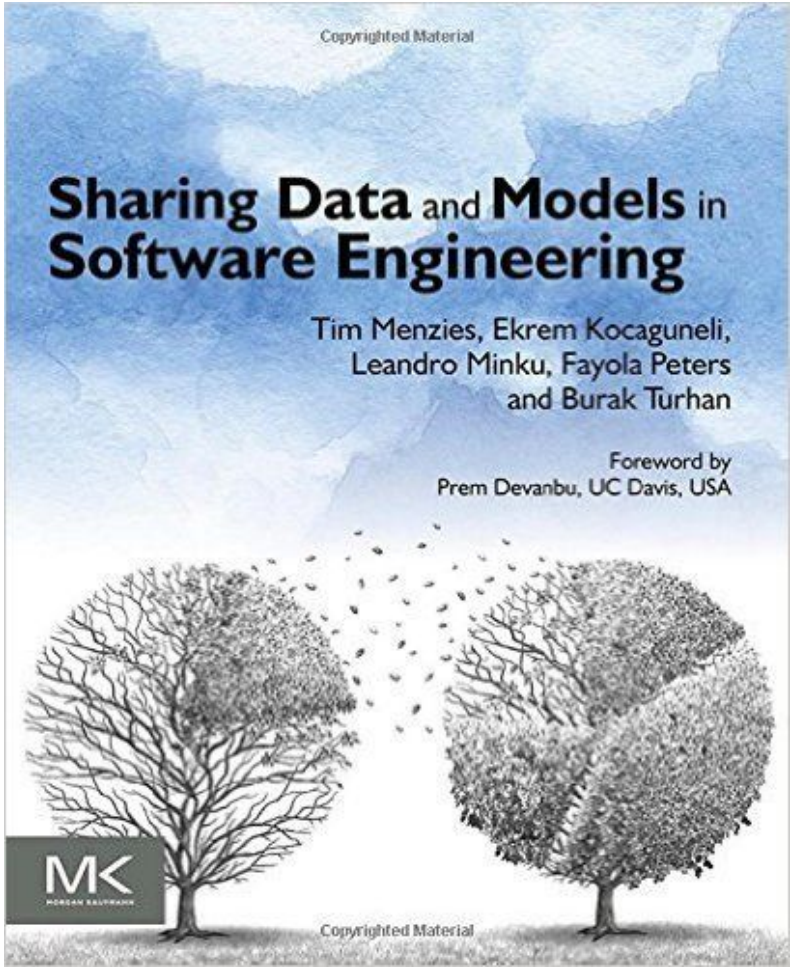
tiny.cc/timcow15

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



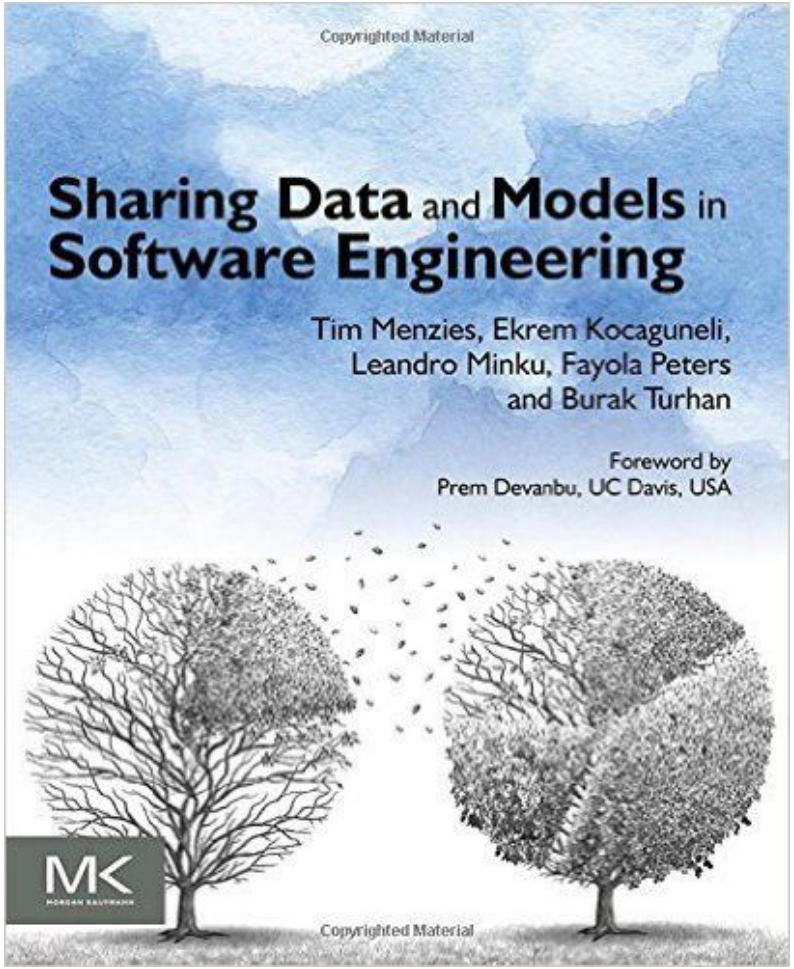
tiny.cc/timcow15

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

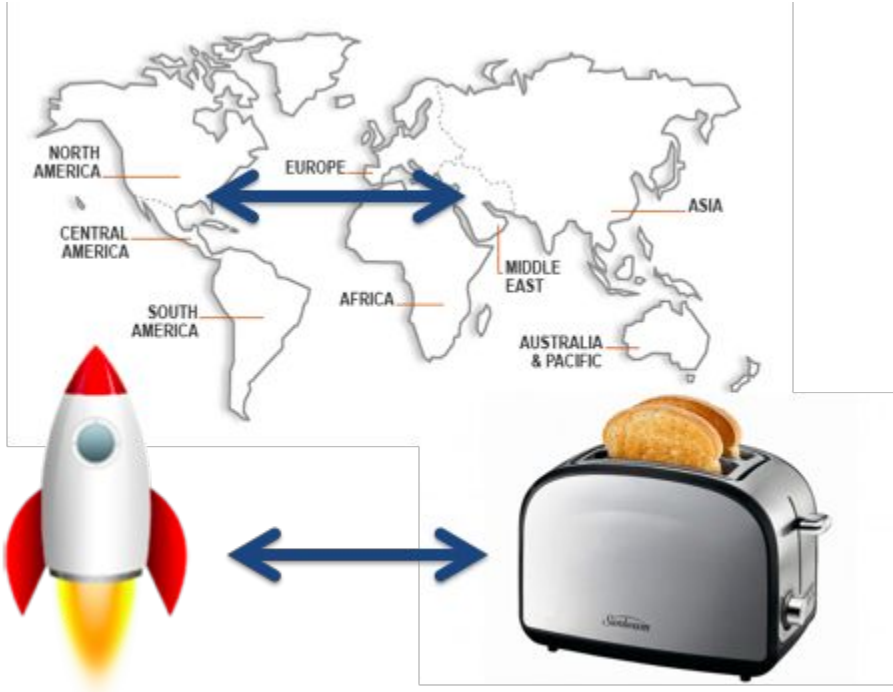


Gooder news: Transfer learning

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)



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

Scales up to massive studies

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

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

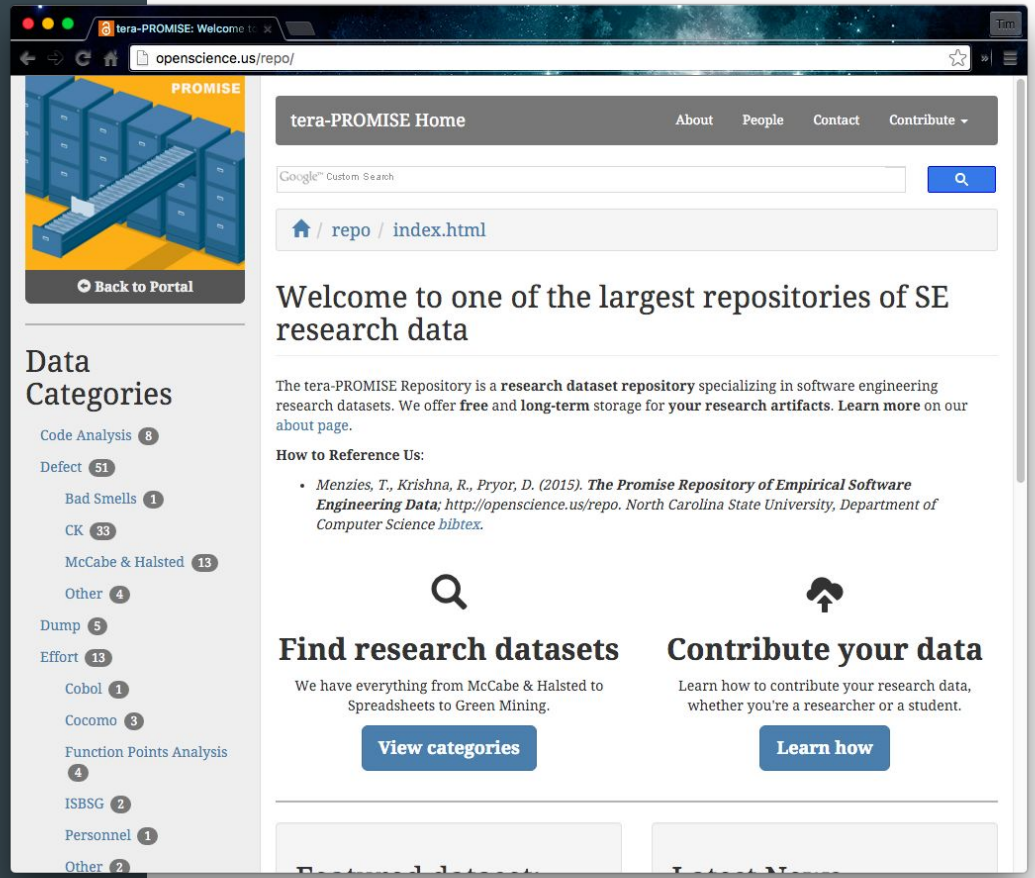
FSE'14 November 16âĂŞ22, 2014, Hong Kong, China

A little advertisement

...

Let's all share more data

- openscience.us/repo



tiny.cc/timcow15

(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

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

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

goals

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

Next gen challenges

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" model

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

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

no "best" model generator

Dramatic improvements to learner performance via data-set-dependent tunings
- See next slide.

Hyper-parameter optimization
- Maybe, N papers at ICSE'16

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

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