What Follows Predictive Modeling?

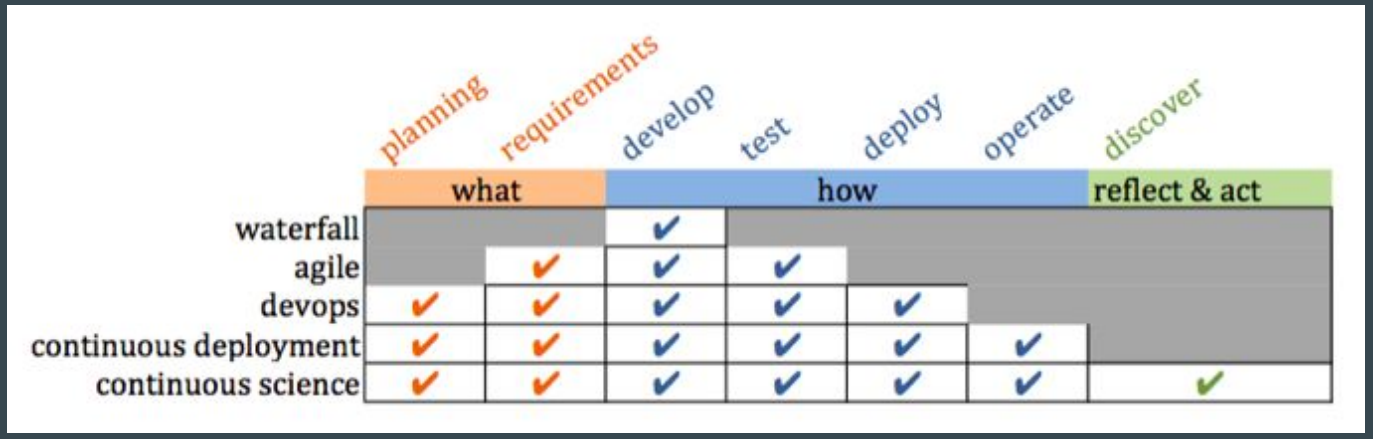
TIM MENZIES, CS, NC State, USA
tim.menzies@gmail.com

Tools to let other people run data miners… better

Why expand our role?

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

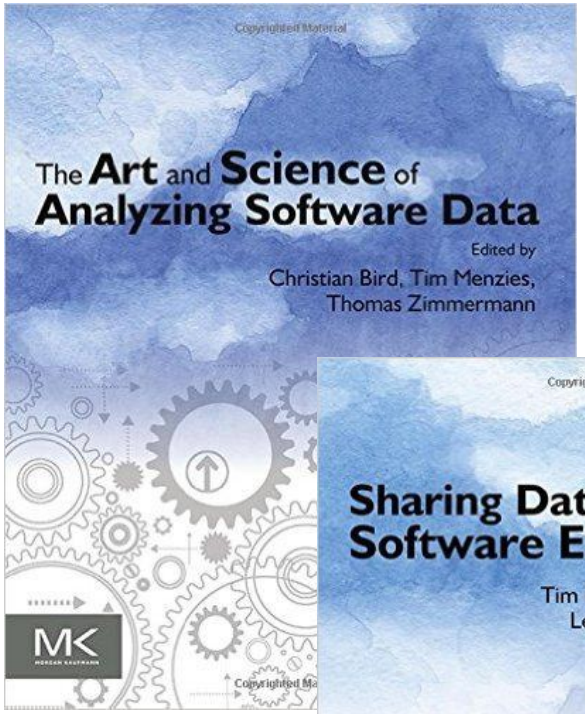
- NOW: we run the data miners
  - NEXT: we write tools that let other people run data miners... better



Good News

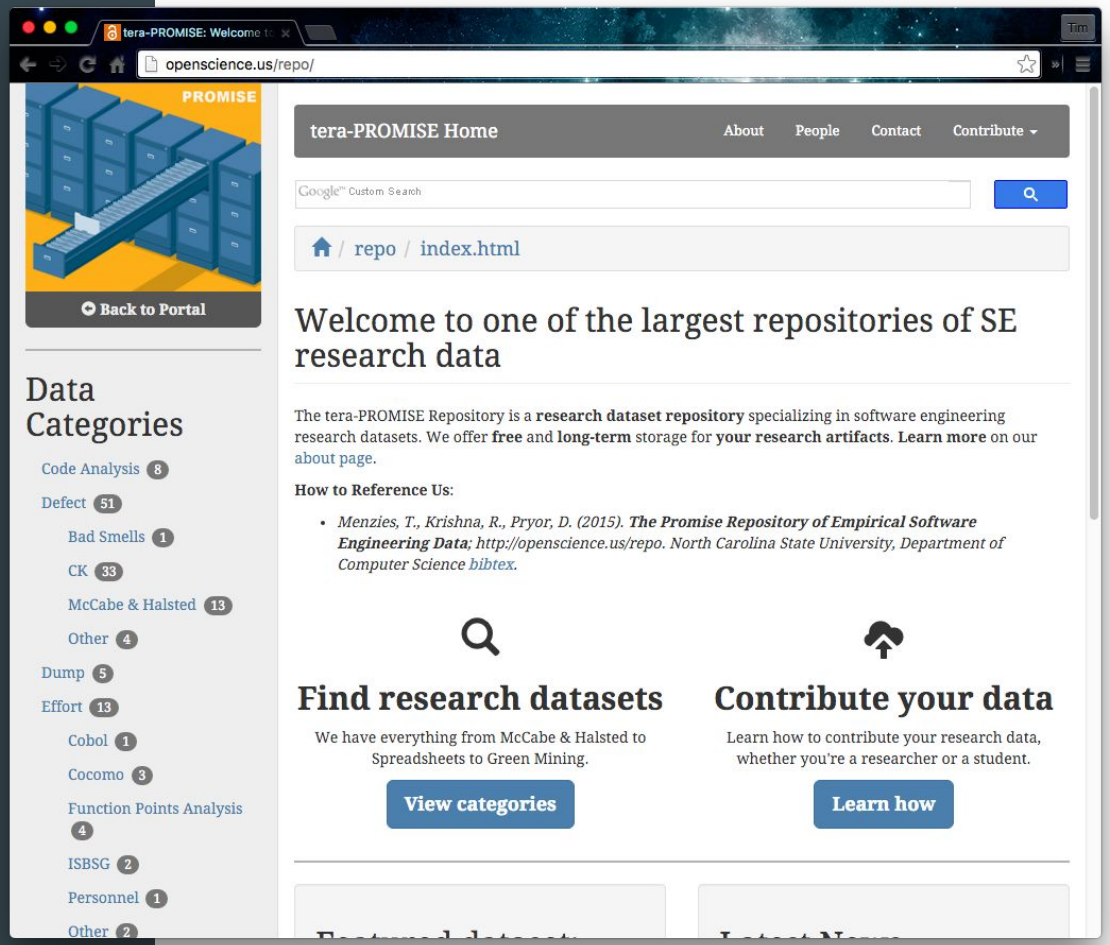
Software project data can

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



Let's all share more data

- openscience.us/repo



(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

(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

(My) Lessons from the PROMISE project

Always re-learning

New data?
- Then, maybe, new model.

Predictors anomaly detectors
- to recognize when old model needs updating

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

All learners are 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 ASE conf (2009)
- Menzies, Bener et al. ASE journal, 2010, 17(4)
- Krall, TSE 2015
- Minku, TOSEM'13

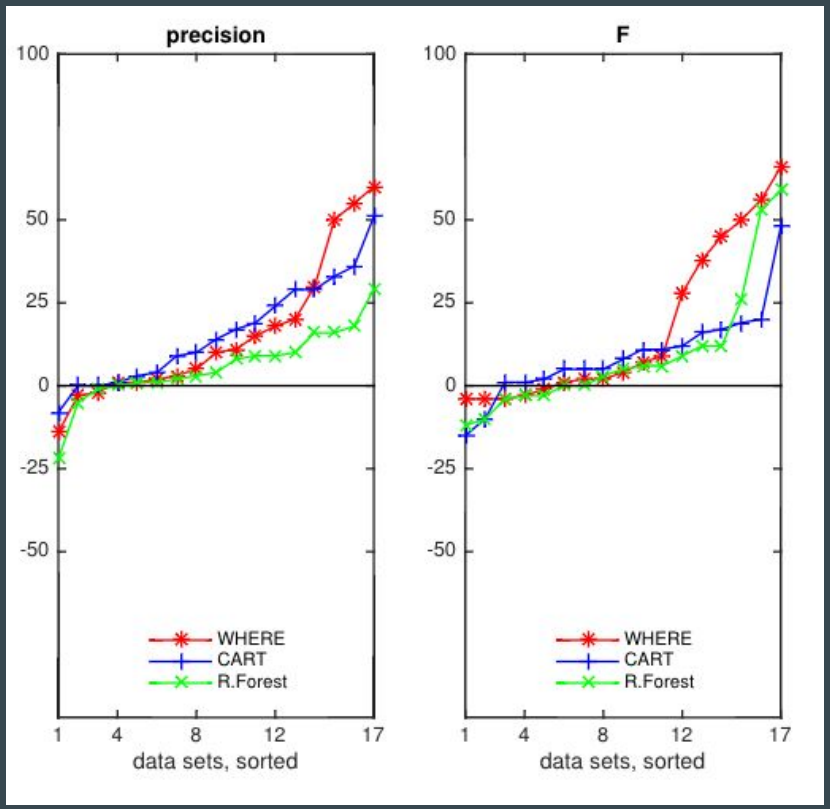
Optimizers for data miners

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



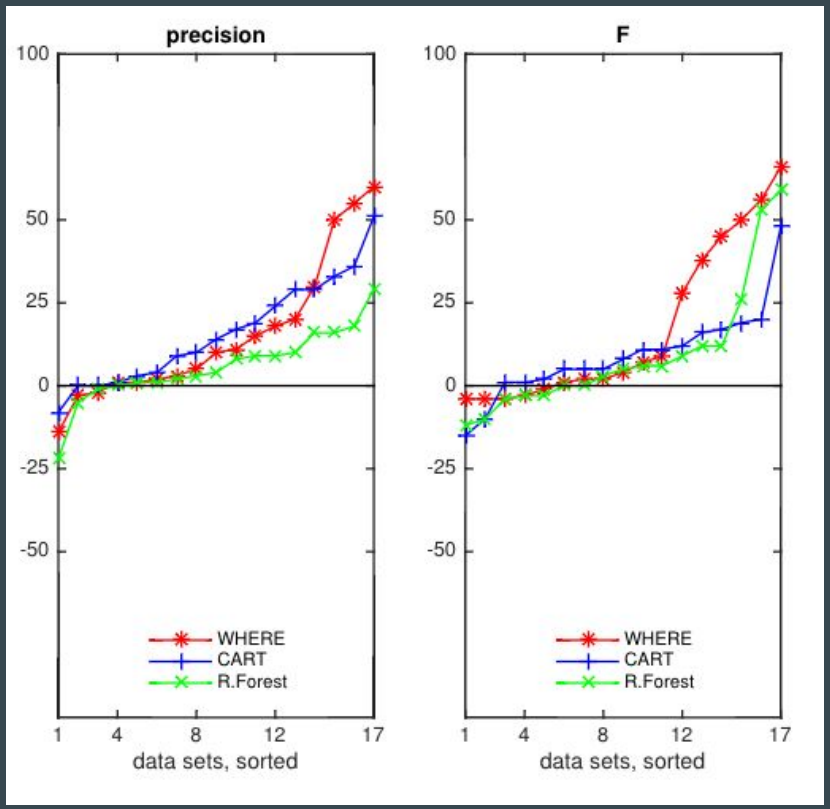
Optimizers for data miners

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**



Tools to let other people run data miners… better

Next Generation Optimizers

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

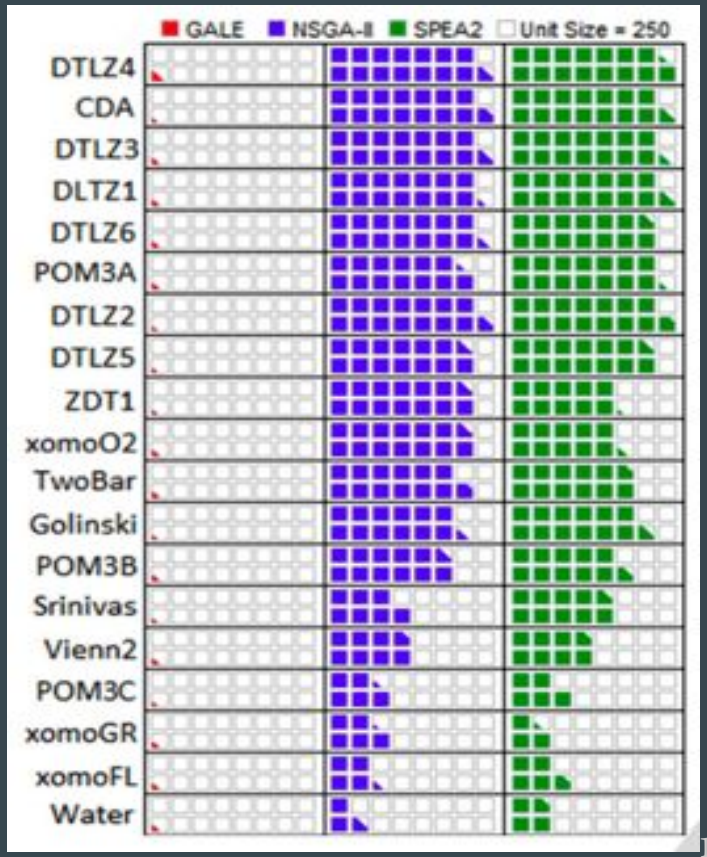
function GALE():

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.



Tools to let other people run data miners… better


QUESTIONS?
COMMENTS?

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