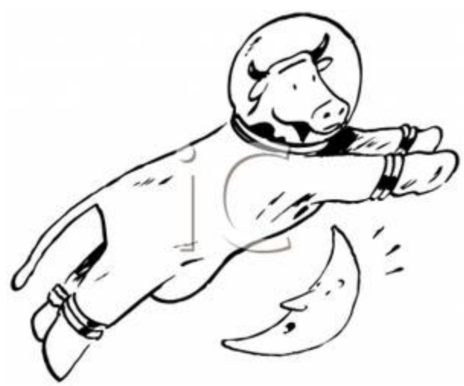
S.P.A.C.E. & COWS & SOFT. ENG.

TIM MENZIES

WVU

DEC 2011



THE COW DOCTRINE

- Seek the fence where the grass is greener on the other side.
  - Learn from there
  - Test on here
- Don't rely on trite definitions of "there" and "here"
  - Cluster to find "here" and "there"

THE AGE OF "PREDICTION" IS OVER

OLDE WORLDE

Porter & Selby, 1990

- Evaluating Techniques for Generating Metric-Based Classification Trees, JSS.
- Empirically Guided Software Development Using Metric-Based Classification Trees. IEEE Software
- Learning from Examples: Generation and Evaluation of Decision Trees for Software Resource Analysis. IEEE TSE

In 2011, Hall et al. (TSE, pre-print)

- reported 100s of similar studies.
- L learners on D data sets in a M*N cross-val

The times, they are a changing: harder now to publish D*L*M*N

NEW WORLD

Time to lift our game

No more: D*L*M*N

Time to look at the bigger picture

Topics at COW not studied, not publishable, previously:

- data quality
- user studies
- local learning
- conclusion instability,

What is your next paper?

- Hopefully not D*L*M*N

REALIZING AI IN SE (RAISE'12)

An ICSE'12 workshop submission

- Organizers: Rachel Harrison, Daniel Rodriguez, Me

AI in SE research

- To much focus on low-hanging fruit;
- SE only exploring small fraction of AI technologies.

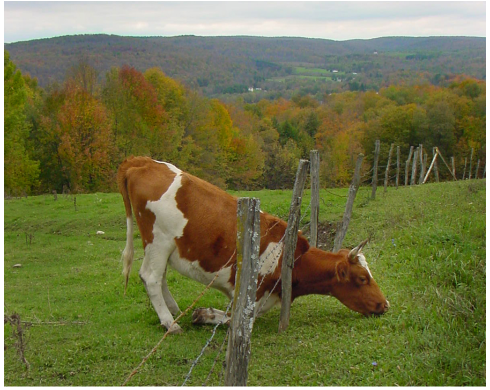
Goal:

- database of sample problems that both SE and AI researchers can explore, together

Success criteria

- ICSE'13: meet to report papers written by teams of authors from SE &AI community

ROADMAP

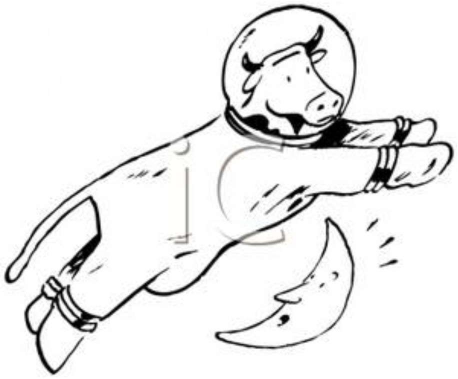


Some comments on the state of the art

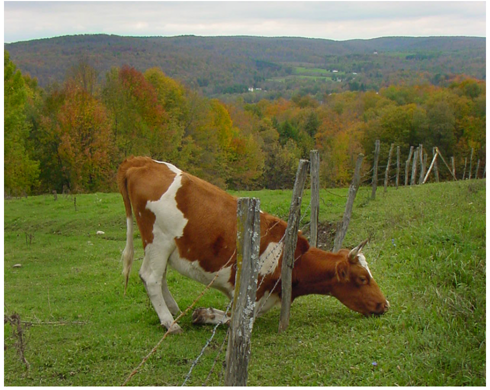
- Why so much SE + data mining?
- Why research SE + data mining
- But is data mining relevant to industry
- The problem of conclusion instability

Learning local

- Globalism: learn from all data
- Localism: learn from local samples
- Learning locality with clustering (S.P.A.C.E.)
- Implications



ROADMAP

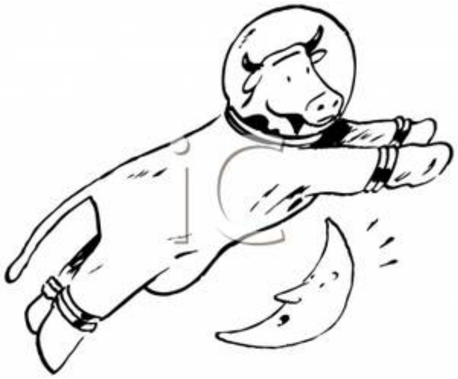


Some comments on the state of the art

- **Why so much SE + data mining?**
- Why research SE + data mining
- But is data mining relevant to industry
- The problem of conclusion instability

Learning local

- Globalism: learn from all data
- Localism: learn from local samples
- Learning locality with clustering (S.P.A.C.E.)
- Implications



Q1: WHY SO MUCH SE + DATA MINING?
A: INFORMATION EXPLOSION

http://CIA.vc

- Monitors 10K projects
- one commit every 17 secs

SourceForge.Net:

- hosts over 300K projects,

Github.com:

- 2.9M GIT repositories

Mozilla Firefox projects :

- 700K reports

Q1: WHY SO MUCH SE + DATA MINING? A: WELCOME TO DATA-DRIVEN SE

Olde worlde: large “applications” (e.g. MsOffice)

- slow to change, user-community locked in

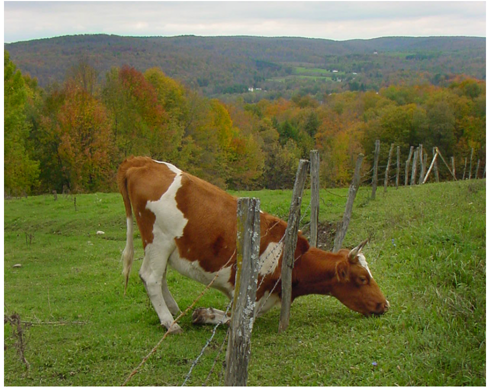
New world: cloud-based apps

- “applications” now 100s of services
    - offered by different vendors
- The user zeitgeist can dump you and move on
    - Thanks for nothing, Simon Cowell
- This change the release planning problem
    - What to release next…
    - … that most attracts and retains market share

Must mine your population

- To keep your population

ROADMAP

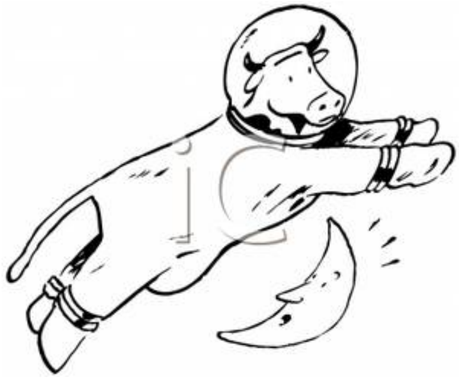


Some comments on the state of the art

- Why so much SE + data mining?
- **Why research SE + data mining**
- But is data mining relevant to industry
- The problem of conclusion instability

Learning local

- Globalism: learn from all data
- Localism: learn from local samples
- Learning locality with clustering (S.P.A.C.E.)
- Implications



Q2: WHY RESEARCH SE + DATA MINING?
A: NEED TO BETTER UNDERSTAND TOOLS

Q: What causes the variance in our results?

- Who does the data mining?
- What data is mined?
- How the data is mined (the algorithms)?
- Etc

Q2: WHY RESEARCH SE + DATA MINING?
A: NEED TO BETTER UNDERSTAND TOOLS

Q: What causes the variance in our results?

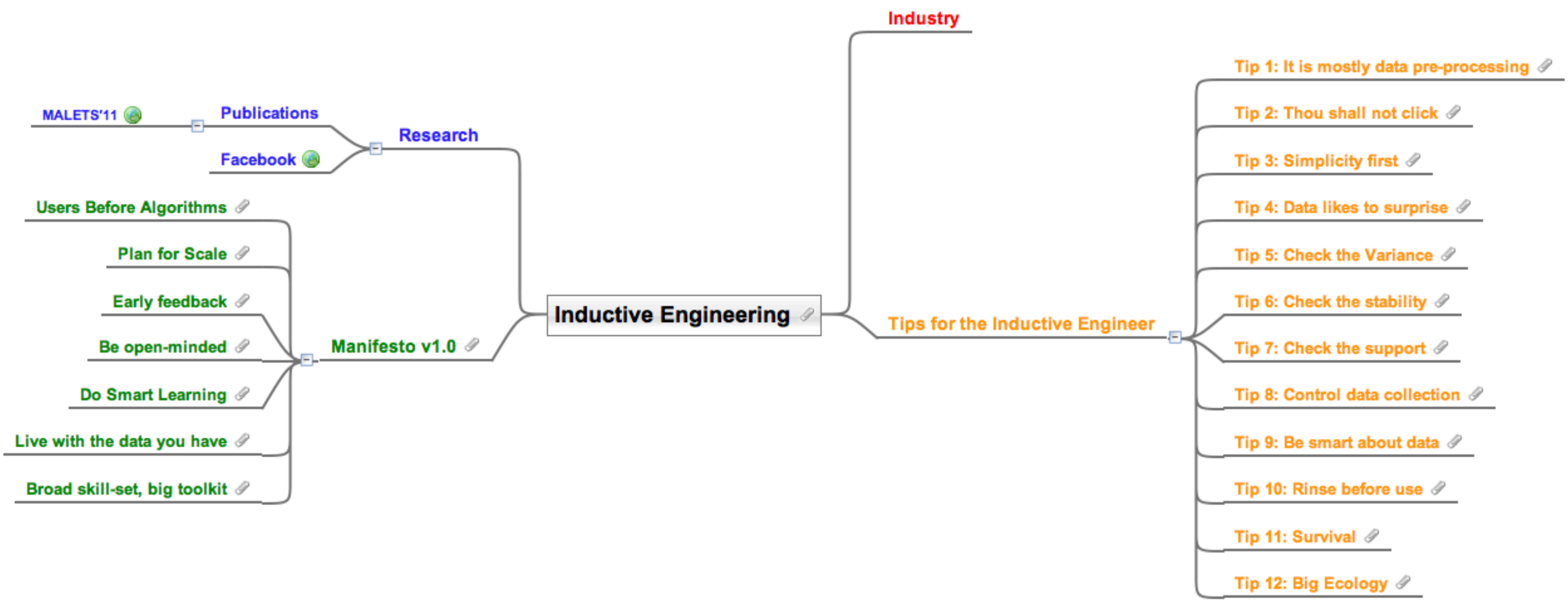
- Who does the data mining?
- What data is mined?
- How the data is mined (the algorithms)?
- Etc

Conclusions depend on who does the looking?

- Reduce the skills gap between user skills and tool capabilities
- Inductive Engineering: Zimmermann, Bird, Menzies (MALETS'11)
  - Reflections on active projects
  - Documenting the analysis patterns

Inductive Engineering:

Understanding user goals to inductively generate the models that most matter to the user.



Q2: WHY RESEARCH SE + DATA MINING?
A: NEED TO UNDERSTAND INDUSTRY

You are a university educator designing graduate classes for prospective industrial inductive engineers

- Q: what do you teach them?

You are an industrial practitioner hiring consultants for an in-house inductive engineering team

- Q: what skills do you advertise for?

You a professional accreditation body asked to certify an graduate program in “analytics”

- Q: what material should be covered?

Q2: WHY RESEARCH SE + DATA MINING? A: BECAUSE WE FORGET TOO MUCH

Basili

- Story of how folks misread NASA SEL data
- Required researchers to visit for a week
    - before they could use SEL data

But now, the SEL is no more:

- that data is lost

The only data is the stuff we can touch via its collectors?

- That's not how physics, biology, maths, chemistry, the rest of science does it.
- Need some lessons that survive after the institutions pass

Its not as if we can embalm those researchers, keep them with us forever

Unless you are from University College

PROMISE PROJECT

1) Conference,

2) Repository to store data from the conference: promisedata.org/data

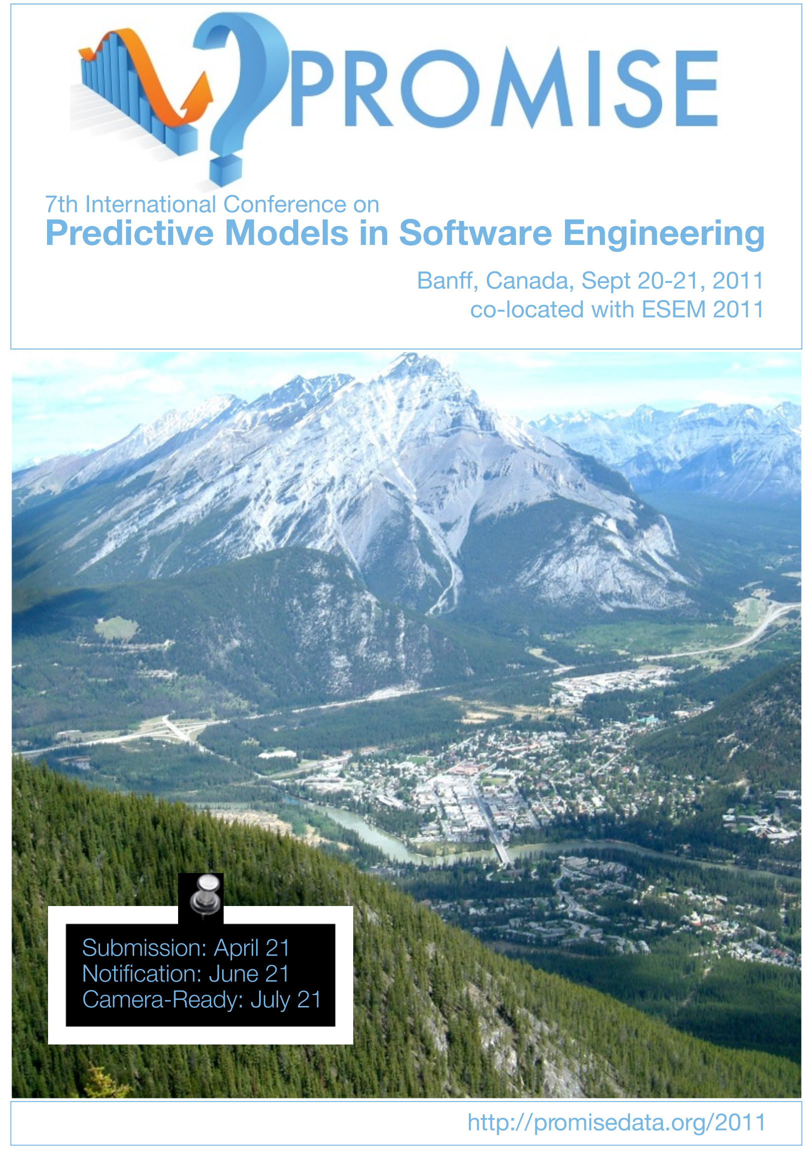
Steering committee:

- Founders: me, Jelber Sayyad
- Former: Gary Boetticher, Tom Ostrand, Guntheur Ruhe,
- Current: Ayse Bener, me, Burak Turhan, Stefan Wagner, Ye Yang, Du Zhang

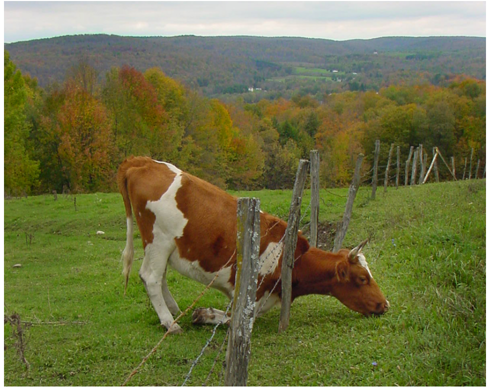
Open issues

- Conclusion instability
- Privacy: share, without reveal;
    - E.g. Peters & me ICSE'12
- Data quality issues:
    - see talks at EASE'11 and COW'11

See also SIR (U. Nebraska) and ISBSG



ROADMAP

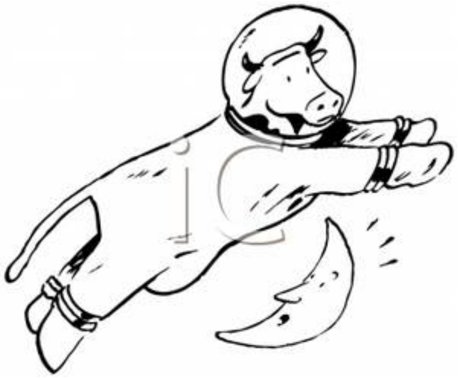


Some comments on the state of the art

- Why so much SE + data mining?
- Why research SE + data mining
- **But is data mining relevant to industry**
- The problem of conclusion instability

Learning local

- Globalism: learn from all data
- Localism: learn from local samples
- Learning locality with clustering (S.P.A.C.E.)
- Implications

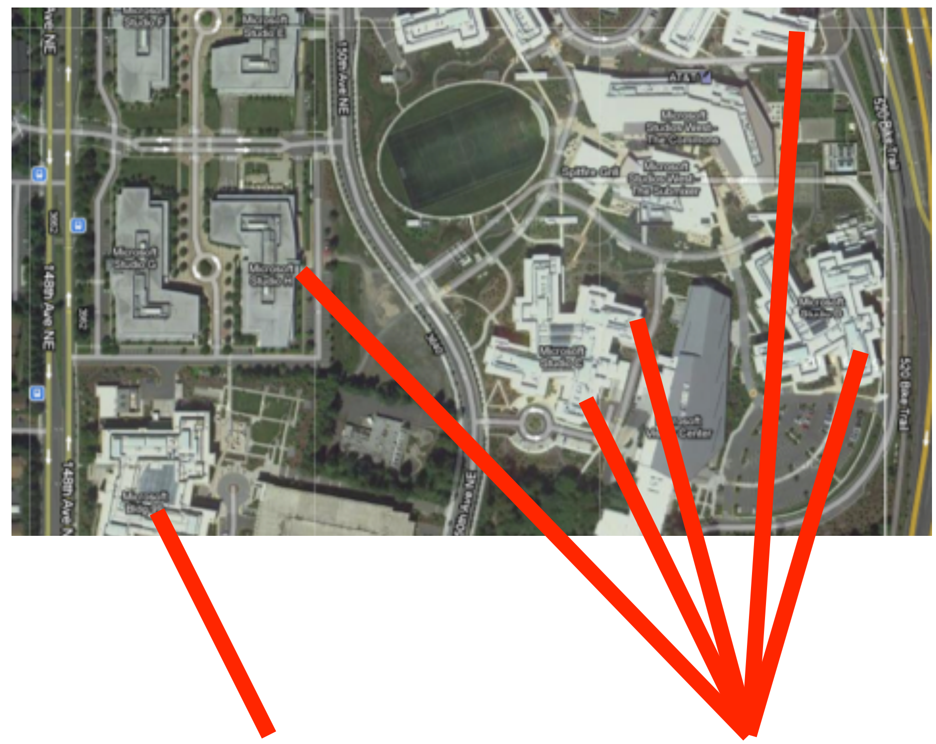


Q3: BUT IS DATA MINING RELEVANT TO INDUSTRY?

A: Which bit of industry?

Different sectors of (say) Microsoft need different kinds of solutions

As an educator and researchers, I ask “what can I do to make me and my students readier for the next business group I meet?”



Microsoft research, Redmond, Building 99

Other studios, many other projects

Q3: BUT IS IT RELEVANT TO INDUSTRY? A: YES, MUCH RECENT INTEREST

Business intelligence

Predictive analytics

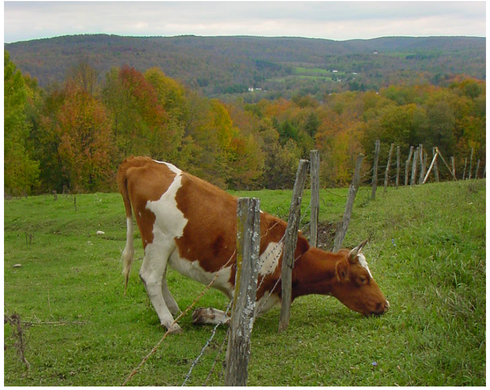
NC state: Masters in Analytics

MSA Class	2011	2010	2009	2008
graduates:	39	39	35	23
%multiple job offers by graduation:	97	91	90	91
Range of salary offers	70K- 140K	65K – 150K	60K- 115K	65K – 135K

POSITIONS OFFERED TO MSA GRADUATES:

Credit Risk Analyst
Data Mining Analyst
E-Commerce Business Analyst
Fraud Analyst
Informatics Analyst
Marketing Database Analyst
Risk Analyst
Display Ads Optimization
Senior Decision Science Analyst
Senior Health Outcomes Analyst
Life Sciences Consultant
Senior Scientist
Forecasting and Analytics
Sales Analytics
Pricing and Analytics
Strategy and Analytics
Quantitative Analytics
Director, Web Analytics
Analytic Infrastructure
Chief, Quantitative Methods Section

ROADMAP

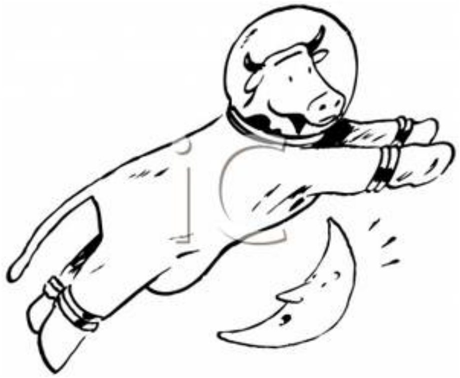


Some comments on the state of the art

- Why so much SE + data mining?
- Why research SE + data mining
- But is data mining relevant to industry
- **The problem of conclusion instability**

Learning local

- Globalism: learn from all data
- Localism: learn from local samples
- Learning locality with clustering (S.P.A.C.E.)
- Implications



The Problem of Conclusion Instability

Learning from software projects

- only viable inside industrial development organizations?
- e.g Basili at SEL
- e.g. Briand at Simula
- e.g Mockus at Avaya
- e.g Nachi at Microsoft
- e.g. Ostrand/Weyuker at AT&T

Conclusion instability is a repeated observation.

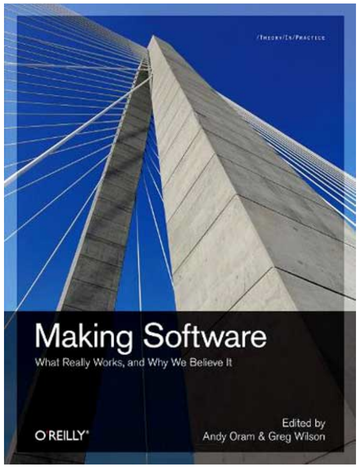
- What works here, may not work there
- Shull & Menzies, in “Making Software”, 2010
- Sheppered & Menzies: speial issue, ESE, conclusion instability

So we can’t take on conclusions from one site verbatim

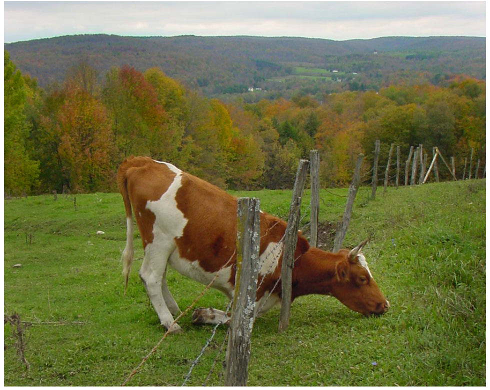
- Need sanity checks +certification envelopes + anomaly detectors
- check if “their” conclusions work “here”

Even “one” site, has many projects.

- Can one project can use another’s conclusion?
- Finding local lessons in a cost-effective manner



ROADMAP

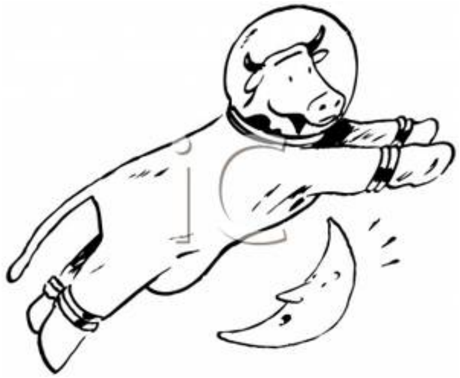


Some comments on the state of the art

- Why so much SE + data mining?
- Why research SE + data mining
- But is data mining relevant to industry
- The problem of conclusion instability

Learning local

- **Globalism: learn from all data**
- Localism: learn from local samples
- Learning locality with clustering (S.P.A.C.E.)
- Implications



GLOBALISM:
BIGGER SAMPLE IS BETTER

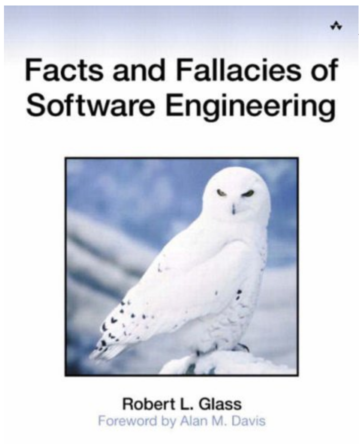
E.g. examples from 2 sources about 2 application types

Source	Gui apps	Web apps
Green Software Inc	gui1, gui2	web1, web2,
Blue Sky Ltd	gui3, gui4	web3, web4

To learn lessons relevant to “gui1”

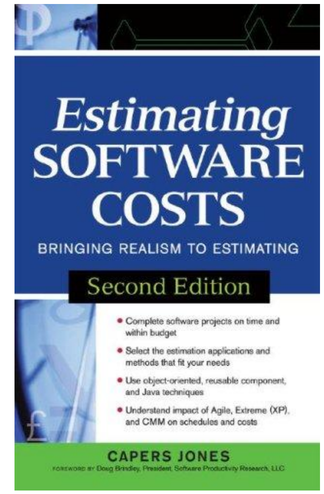
- Use all of {gui2, web1, web2} + {gui3, gui4, web3, web4}

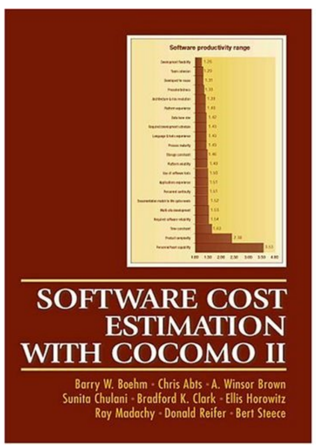
GLOBALISM & RESEARCHERS



R. Glass, *Facts and Falllacies of Software Engineering.* Addison- Wesley, 2002.

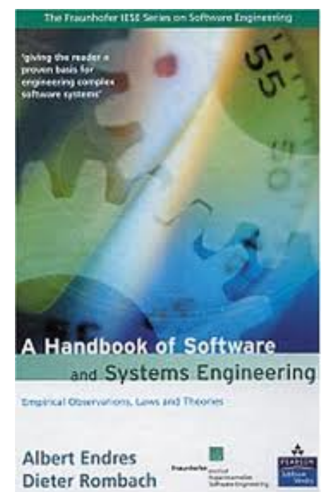
C. Jones, *Estimating Software Costs, 2nd Edition.* McGraw-Hill, 2007.





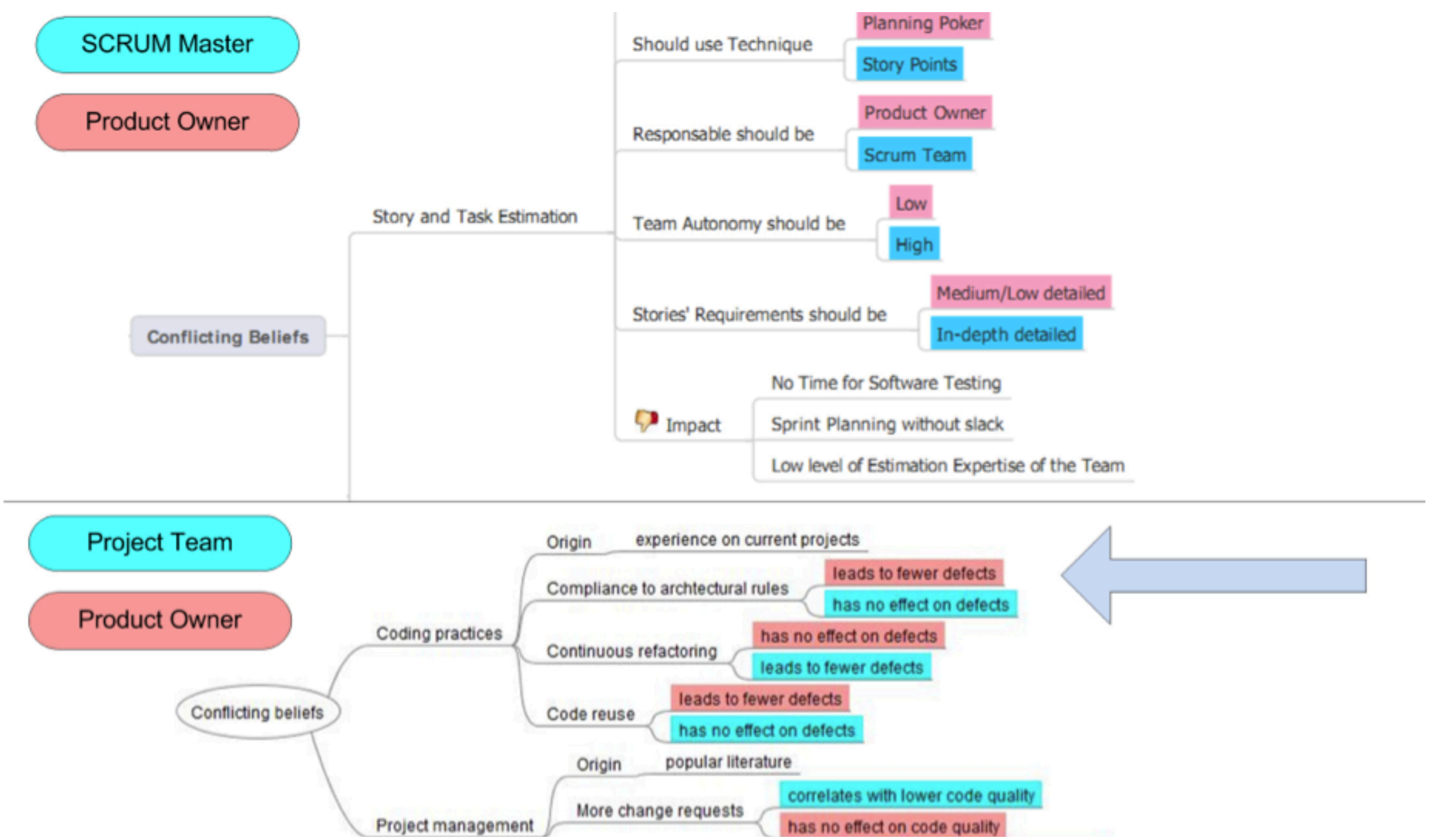
B. Boehm, E. Horowitz, R. Madachy, D. Reifer, B. K. Clark, B. Steece, A. W. Brown, S. Chulani, and C. Abts, *Software Cost Estimation with Cocomo II.* Prentice Hall, 2000.

R. A. Endres, D. Rombach, *A Handbook of Software and Systems Engi- neering: Empirical Observations, Laws and Theories.* Addison Wesley, 2003.



- 50 laws:
- “the nuggets that must be captured to improve future performance” [p3]

GLOBALISM
& INDUSTRIAL ENGINEERS



Mind maps of developers

Brazil (top) from PASSOS et al 20011

USA (bottom)

See also, Jorgensen, TSE, 2009

(NOT) GLOBALISM & DEFECT PREDICTION

ref	cbo	rfc	lcom	dit	noc	wmc	#projects	size	type
[15]	+	+	+	-	-	+	6	95-201 classes	6 versions of rhino (java)
[16]	+	+	+	-	-	+	12	86 classess (3-12kloc)	student
[17]	+	+	-				1	1700 classes (110kloc)	commercial telecom
[18]	+	+	-	+	+	+	8	113 classes	student
[19]	+	+	-	+	+	+	8	114 classes	student
[20]	+	+	+	+	-		1	83 classes	commercial: lalo (c++)
[21]				+	+		1	32 classes	commercial: telecom c++
[22]				+	-		1	42-69 classes	commercial java word proc.
[23]	+	-	-	-	-	-	1	85 classes	telecom c++
[24]	-	+		-	-	+	3	92 classes	3 c++ subsystems,commercial
[25]	+	+	+	-	+	+	1	123 classes (34kloc)	java commercial
[26]	+			+		+	1	706 classes	commercial c++ and java
[27]	+	+	+	-	+	+	1	145 classes	kc1-nasa
[28]	+	+	+	+	-	+	1	3677 classes	open source:mozilla
[29]	+	+	+			+	1	?	java (sap) commercial
[30]	+	+	+	+	+	+	3	?	eclipse 2.0, 2.1, 3.0
[31]	-	+	+	-	-	+	8	113 classes	student
[32]		+	+	+	+		2	64 classes	?sales and cd-selection system
[33]		-		-	-	-	1	3344 modules (2mloc)	commercial telecom c++
[34]	+	+	+	-	-	+	5	395 classes	commercial telecom c++
[35]	+	+		-	-	+	1	1412 classes	open source:jdt
[36]	+	+		-	-	+	2	9713 classes	eclipse 2.0, 2.1
[37]	+	+	-	-	-	+	1	145 classes	kc1-nasa
[38]				+	-		1	145 classes	commercial java xml editor
[39]	-	-	-	-	-	-	1	174 classes	commercial telecom c++
[40]	-					-	0	50 classes	student
[41]	+	+	-	-	-	+	1	145 classes	kc1-nasa
[42]		+		+	+		2	294 classes	commercial c++
total +	18	20	11	11	8	17			
total -	4	3	7	14	16	4			

Total percents:	"*" denotes majority conclusion in each column					
+	* 64%	* 71%	* 39%	39%	29%	* 61%
-	14%	11%	25%	* 50%	* 57%	14%

KEY:

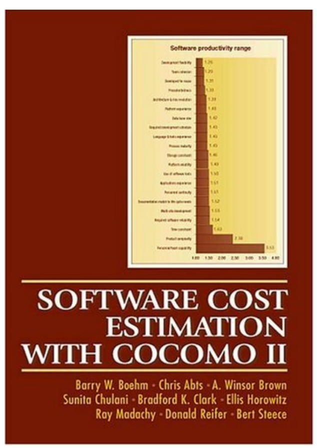
- Strong consensus (over 2/3rds)
- Some consensus (less than 2/3rds)
- Weak consensus (about half)
- No consensus

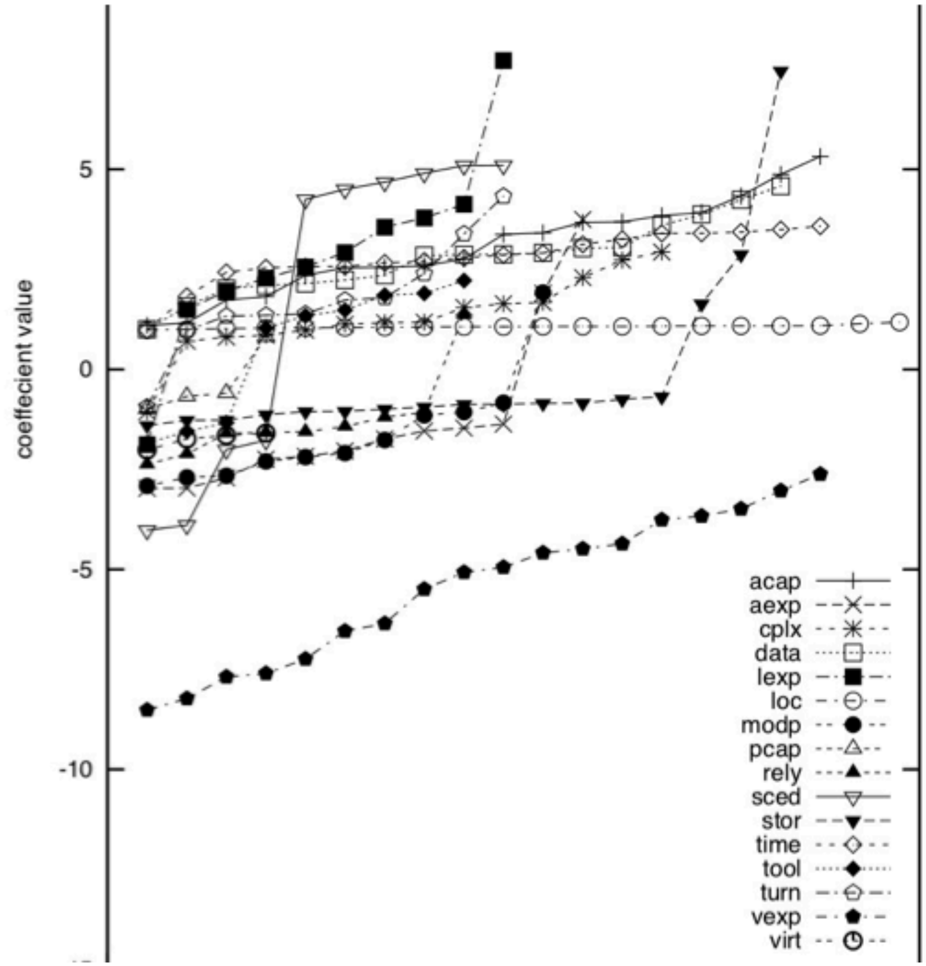
Fig. 3. Contradictory conclusions from OO-metrics studies for defect prediction. Studies report significant ("+") or irrelevant ("-") metrics verified by univariate prediction models. Blank entries indicate that the corresponding metric is not evaluated in that particular study. Colors comment on the most frequent conclusion of each column. CBO= coupling between objects; RFC= response for class (#methods executed by arriving messages); LCOM= lack of cohesion (pairs of methods referencing one instance variable, different definitions of LCOM are aggregated); NOC= number of children (immediate subclasses); WMC= #methods per class.

(NOT) GLOBALISM & EFFORT ESTIMATION

Effort = a . loc^x . y

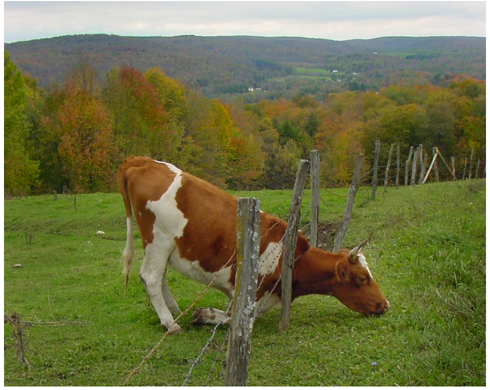
- learned using Boehm's methods
- 20*66% of NASA93
- COCOMO attributes
- Linear regression (log pre-processor)
- Sort the co-efficients found for each member of x,y





CONCLUSION (ON GLOBALISM)

ROADMAP

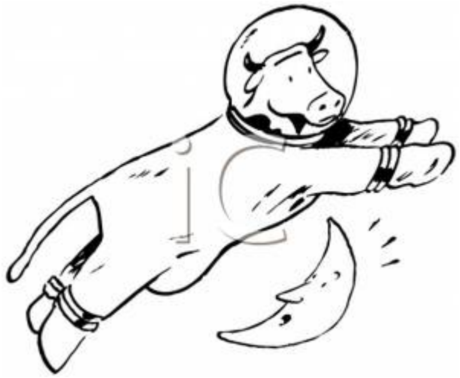


Some comments on the state of the art

- Why so much SE + data mining?
- Why research SE + data mining
- But is data mining relevant to industry
- The problem of conclusion instability

Learning local

- Globalism: learn from all data
- **Localism: learn from local samples**
- Learning locality with clustering (S.P.A.C.E.)
- Implications



LOCALISM: SAMPLE ONLY FROM SAME CONTEXT

E.g. examples from 2 sources about 2 application types

Source	Gui apps	Web apps
Green Software Inc	gui1, gui2	web1, web2,
Blue Sky Ltd	gui3, gui4	web3, web4

To learn lessons relevant to "gui1"

- Restrict to just this the gui tools {gui2, gui3, gui4 }
- Restrict to just this company {gui2,web1, web2}

Er… hang on

- How to find the right local context?

DELPHI LOCALIZATION

Ask an expert to find the right local context

- Are we sure they're right?
- Posnett at al. 2011:
    - What is right level for learning?
    - Files or packages?
    - Methods or classes?
    - Changes from study to study

And even if they are "right":

- should we use those contexts?
- E.g. need at least 10 examples to learn a defect model (Valerdi's rule, IEEE Trans, 2009)
- 17/147 = 11% of this data

	OPERATING ENVIRONMENT							
APPLICATION DOMAIN	avionics	fixed ground	missile	mobile ground	shipboard	unmanned airborne	unmanned space	total
business systems		6		4	2			12
command & control	1	41		16	35			93
communications	4	77			17		2	100
controls & display	8	6		2	5			21
executive		4			3			7
information assurance		1						1
infrastructure		11			23			34
maintenance & diagnostics	1				5			6
mission management	42	2	3	2		1		50
mission planning	1	17						18
modeling & simulation		1						1
process control		3		6	1			10
scientific systems					3			3
sensor control & processing	12	15			18			45
simulation & modeling		19			17			36
spacecraft BUS							9	9
spacecraft payload							16	16
test & evaluation		2			2			4
tool & tool systems		6	1					7
training				2	6			8
weaps delivery & control	11		19		9			39
totals	80	211	23	32	146	1	27	520

Fig. 1. Delphi localizations of 520 US Defense Department software projects; from Madachy et al. [12].

CLUSTERING TO FIND "LOCAL"

TEAK: estimates from "k" nearest-neighbors

- "k" auto-selected per test case
- Pre-processor to cluster data, remove worrisome regions
- IEEE TSE, Jan'11
  T = Tim
  E = Ekrem Kocaguneli
  A = Ayse Bener
  K= Jacky Keung

Dataset	Criterion	Subsets	Subsets Size
cocomo81	project type	cocomo81e	28
		cocomo81o	24
		cocomo81s	11
nasa93	development center	nasa93_center_1	12
		nasa93_center_2	37
		nasa93_center_5	39
desharnais	language type	desharnaisL1	46
		desharnaisL2	25
		desharnaisL3	10
finnish	application type	finnishAppType1	17
		finnishAppType2345	18
kemerer	hardware	kemererHardware1	7
		kemererHardware23456	8
maxwell	application type	maxwellAppType1	10
		maxwellAppType2	29
		maxwellAppType3	18
maxwell	hardware	maxwellHardware2	37
		maxwellHardware3	16
		maxwellHardware5	7
maxwell	source	maxwellSource1	8
		maxwellSource2	54

ESEM'11

- Train within one delphi localization
- Or train on all and see what it picks
- Results #1: usually, cross as good as within

Results #2: 20 times, estimate for x in S_i. TEAK picked across as picked within

Test Set	From S1	From S2	From S3
S1: cocomo81e (28)	1.0 (3.6%)	1.1 (4.8%)	1.6 (14.4%)
S2: cocomo81o (24)	1.8 (6.6%)	1.3 (5.6%)	1.1 (10.4%)
S3: cocomo81s (11)	1.4 (5.1%)	1.7 (7.0%)	1.0 (9.4%)
S1: nasa93_center_1 (12)	1.0 (8.1%)	2.9 (7.9%)	1.7 (4.3%)
S2: nasa93_center_2 (37)	1.6 (13.0%)	4.6 (12.4%)	3.8 (9.8%)
S3: nasa93_center_5 (39)	0.8 (6.7%)	2.2 (6.0%)	2.1 (5.4%)
S1: desharnaisL1 (46)	2.5 (5.5%)	1.7 (7.0%)	0.8 (7.9%)
S2: desharnaisL2 (25)	2.6 (5.6%)	1.5 (6.1%)	0.7 (6.7%)
S3: desharnaisL3 (10)	1.9 (4.1%)	1.3 (5.0%)	0.4 (4.0%)
S1: finnishAppType1 (17)	1.6 (9.1%)	1.6 (8.8%)	
S2: finnishAppType2345 (18)	1.4 (8.2%)	1.6 (8.8%)	
S1: kemererHardware1 (7)	0.6 (8.8%)	0.9 (10.7%)	
S2: kemererHardware23456 (8)	0.5 (7.3%)	0.8 (10.6%)	
S1: maxwellAppType1 (10)	0.7 (7.1%)	1.7 (5.9%)	1.0 (5.8%)
S2: maxwellAppType2 (29)	0.4 (3.7%)	1.8 (6.2%)	1.0 (5.5%)
S3: maxwellAppType3 (18)	0.6 (6.3%)	0.9 (3.2%)	1.0 (5.6%)
S1: maxwellHardware2 (37)	1.7 (4.6%)	0.8 (4.9%)	0.4 (6.0%)
S2: maxwellHardware3 (16)	2.5 (6.8%)	1.1 (6.8%)	0.3 (4.3%)
S3: maxwellHardware5 (7)	2.3 (6.2%)	0.8 (5.0%)	0.3 (4.5%)
S1: maxwellSource1 (8)	0.1 (1.6%)	2.8 (5.2%)	
S2: maxwellSource2 (54)	0.4 (4.6%)	2.8 (5.3%)	

CONCLUSION (ON LOCALIZATION)

Delphi localizations

- Can restrict sample size
- Don't know how to check if your delphi localizations are "right"
- How to learn delphi localizations for new domains?
- Not essential to inference

Auto-learned localizations
(learned via nearest neighbor methods)

- Works just as well as delphi
- Can select data from many sources
- Can be auto-generated for new domains
- Can hunt out relevant samples from data from multiple sources

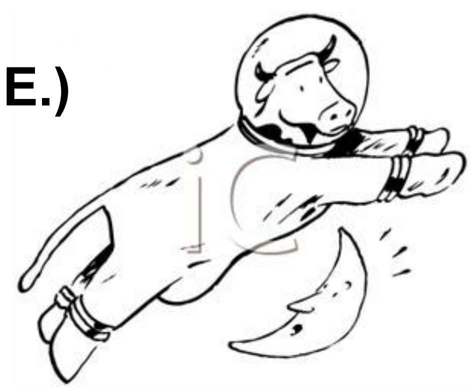
ROADMAP

Some comments on the state of the art

- Why so much SE + data mining?
- Why research SE + data mining
- But is data mining relevant to industry
- The problem of conclusion instability

Learning local

- Globalism: learn from all data
- Localism: learn from local samples
- **Learning locality with clustering (S.P.A.C.E.)**
- Implications



CLUSTERING + LEARNING

Turhan, Me, Bener, ESE journal '09

- Nearest neighbor, defect prediction
  - Combine data from other sources
  - Prune to just the 10 nearest examples to each test instance
  - Naïve Bayes on the pruned set

Turhan et al. (2009)	Me et al, ASE, 2011
Not scalable	Near linear time processing
No generalization to report to users	Use rule learning

CLUSTERING + LEARNING ON SE DATA

Cuadrado, Gallego, Rodriguez, Sicilia, Rubio, Crespo. Journal Computer Science and Technology (May07)

- EM on to 4 Delphi localizations
    - case tool = yes, no
    - methodology used = yes, no
- Regression models, learned per cluster, do better than global

Table 3. *MMRE* and *Pred* Comparison of a Single Model vs. Multiple Models

	MMRE	*Pred* (< 0.3) (%)
Single Model	2.17	26.75
Using Clustering	1.03	35.60

But why train on your own clusters?

- If your neighbors get better results…
- … train on neighbors…
- … test on local
- Training data similar to test
- No need for N*M-way cross val

MUST DO BETTER

Turhan et al. (2009)	Me et al, ASE, 2011
Not scalable	Near linear time processing
No generalization to report to users	Use rule learning

Cuadrado et .al (2007)	Me et al, ASE, 2011
Only one data set	Need more experiments
Just effort estimation	Why not effort and defect?
Delphi and automatic localizations ?	Seek fully automated procedure
Returns regression models	Our users want actions, not trends. Navigators, not maps
Clusters on naturally dimensions	What about synthesized dimensions?
Train and test on local clusters	Why not train on superior neighbors (the envy principle)
Tested via cross-val	Train on neighbor, test on self. No 10*10-way cross val

S.P.A.C.E = SPLIT, PRUNE

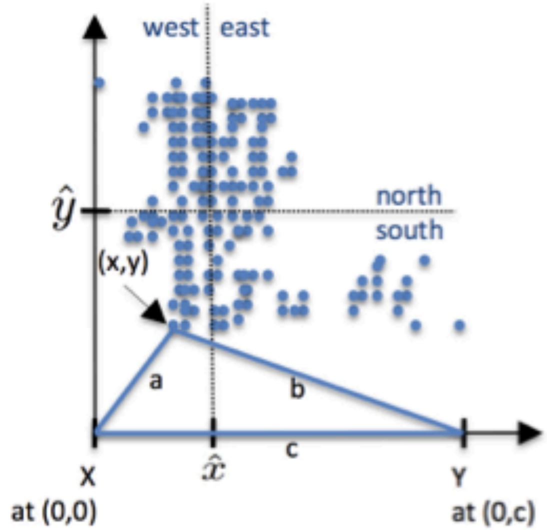
SPLIT: quadtree generation

Pick any point W; find X furthest from W, find Y furthest from Y.

XY is like PCA's first component; found in O(2N) time, note O(N^2) time

All points have distance a,b to (X,Y)
$x = (a^2 + c^2 - b^2)/2c$; $y = \text{sqrt}(a^2 - x^2)$

Recurse on four quadrants formed



PRUNE: FORM CLUSTERS

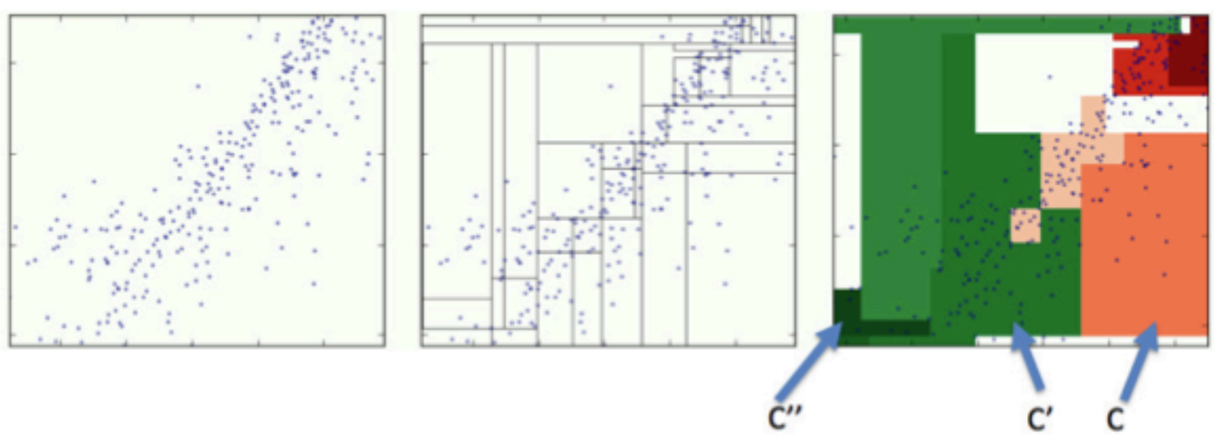
Combine quadtree leaves with similar densities

Score each cluster by median score of class variable

Find envious neighbors (C1,C2)

- score(C2) better than score(C1)

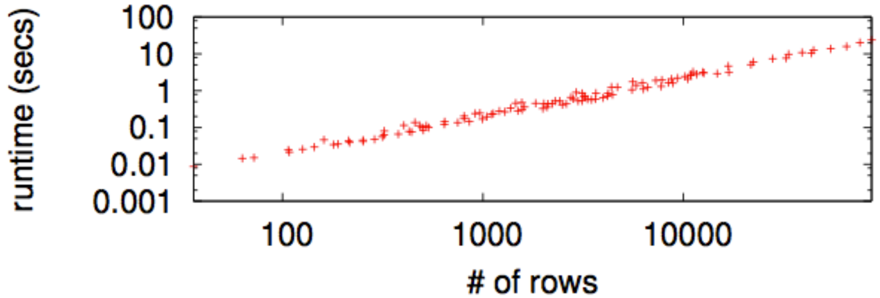
Train on C2 , test on C2



WHY SPLIT, PRUNE?

Unlike Turhan'09:

LogLinear clustering time:
i.e. fast and scales



Turhan et al. (2009)	Me et al, ASE, 2011	S.P.
Not scalable	Near linear time processing	✔
No generalization to report to users	Use rule learning	

Cuadrado et .al (2007)	Me et al, ASE, 2011	S. P.
Only one data set	Need more experiments	
Just effort estimation	Why not effort and defect?	
Delphi & automatic localizations ?	Seek fully automated procedure	✔
Returns regression models	Our users want actions, not trends. Navigators, not maps	
Clusters on naturally dimensions	What about synthesized dimensions?	✔
Train and test on local clusters	Why not train on superior neighbors (the envy principle)	✔
Tested via cross-val	Train on neighbor, test on self. No 10*10-way cross val	✔

S.P.A.C.E = S.P. ADD CONTRAST ENVY (A.C.E.)

Contrast set learning (WHICH)

Fuzzy beam search

First Stack = one rule for each discretized range of each attribute

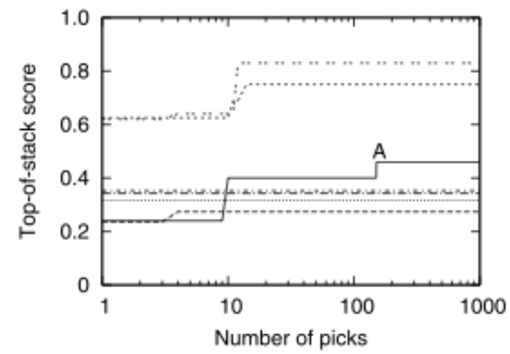
Repeat. Make next stack as follows:

- Score stack entries by lift (ability to select better examples)
- Sort stack entries by score
- Next stack = old stack
    - plus combinations of randomly selected pairs of existing rules
    - Selection biased towards high scoring rules

Halt when top of stack's score stabilizes

Return top of stack

WHY ADD CONSTRAST ENVY?



Search criteria is adjustable

- See Menzies et al ASE journal 2010

Early termination

Turhan et al. (2009)	Me et al, ASE, 2011	S.P.	A.C.E
Not scalable	Near linear time processing	✔	✔
No generalization to report to users	Use rule learning		✔

Cuadrado et .al (2007)	Me et al, ASE, 2011	S.P.	A.C.E.
Only one data set	Need more experiments		
Just effort estimation	Why not effort and defect?		
Delphi & automatic localizations ?	Seek fully automated procedure	✔	
Returns regression models	Our users want actions, not trends. Navigators, not maps		✔
Clusters on naturally dimensions	What about synthesized dimensions?	✔	
Train and test on local clusters	Why not train on superior neighbors (the envy principle)	✔	
Tested via cross-val	Train on neighbor, test on self. No 10*10-way cross val	✔	

DATA FROM HTTP://PROMISEDATA.ORG/DATA

Find (25,50,75,100)th percentiles of class values

- in examples of test set selected by *global* or *local*

Express those percentiles as ratios of max values in **all**.

Effort reduction = { NasaCoc, China } : COCOMO or function points

Defect reduction = { lucene, xalan, jedit, synapse,etc } : CK metrics(OO)

		effort		defect							
		NasaCoc	china	lucene2.4	xalan2.6	jedit4.0	synapse1.2	velocity1.6	tomcat	xerces1.4	$\frac{\sum local}{\sum global}$
median (50th percentile)	global	17	4	3	12	0	0	0	0	0	0.64
	local	7	3	0	12	0	0	0	0	1	
stability (75th-25th percentile)	global	16	7	10	12	0	11	8	0	1	0.37
	local	6	6	3	0	0	0	8	0	1	
worst-case (100th percentile)	global	100	100	100	100	100	100	100	100	100	0.39
	local	9	100	23	62	8	33	50	33	32	

When the same learner was applied globally or locally

- Local did better than global
- Death to generalism

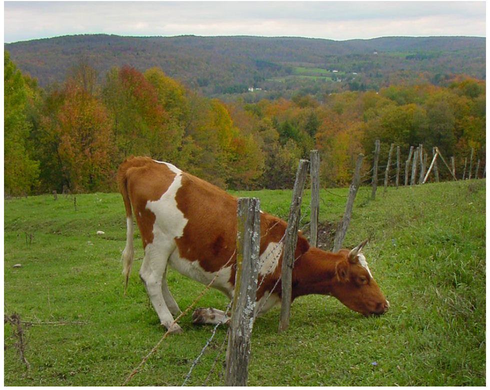
As with Cuadrado '07: local better than global (but for multiple effort and defect data sets and no delphi-localizations)

EVALUATION

Turhan et al. (2009)	Me et al, ASE, 2011	S.P.	A.C.E	COW
Not scalable	Near linear time processing	✔	✔	
No generalization to report to users	Use rule learning		✔	

Cuadrado et .al (2007)	Me et al, ASE, 2011	S.P.	A.C.E.	COW
Only one data set	Need more experiments			✔
Just effort estimation	Why not effort and defect?			✔
Delphi & automatic localizations ?	Seek fully automated procedure	✔		
Returns regression models	Our users want actions, not trends. Navigators, not maps		✔	
Clusters on naturally dimensions	What about synthesized dimensions?	✔		
Train and test on local clusters	Why not train on superior neighbors (the envy principle)	✔		
Tested via cross-val	Train on neighbor, test on self. No 10*10-way cross val	✔		

ROADMAP

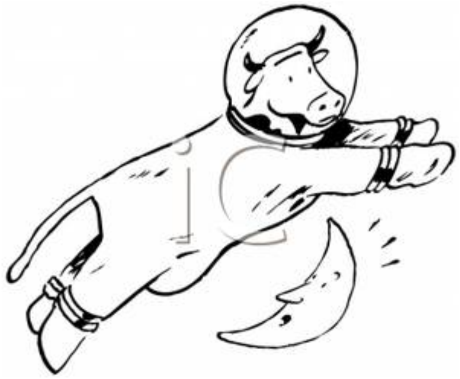


Some comments on the state of the art

- Why so much SE + data mining?
- Why research SE + data mining
- But is data mining relevant to industry
- The problem of conclusion instability

Learning local

- Globalism: learn from all data
- Localism: learn from local samples
- Learning locality with clustering (S.P.A.C.E.)
- **Implications**



IMPLICATIONS: GLOABLISM

Simon says, no

IMPLICATIONS: DELPHI LOCALISM

Simon says, no

IMPLICATIONS: CLUSTER-BASED LOCALISM

Simon says, yes

IMPLICATIONS: CONCLUSION INSTABILITY

From this work

- Misguided to try and tame conclusion instability
- Inherent in the data

	effort		defect						
cluster	NasaCoc	china	lucene2.4	xalan2.6	jedit4.0	velocity1.6	synapse1.2	tomcat	xerces1.4
global	kloc=1	afp=1	rfc=2	loc=1	rfc=2	cam=7	amc=1	loc=2	cbo=1
C0									
C1	rely=n	added=4	amc=7	amc=1	ic=7	noc=1	dit=4	cbm=1	dit=1
C2	prec=h	deleted=1	ca=1	cam=2	noc=1	dam=1 or 5		dam=1	dam=1
C3		deleted=1	dam=5	cam=3	amc=6	avg_cc=4		noc=1	ca=1 or 7
C4			mfa=1	dit=2 or 4	noc=1	moa=1		rfc=5	**cbo=1**
C5			moa=1	**loc=1**				lcom3=5	
C6				**loc =1 or 2**				max_cc=1	
C7				moa=1				cbm=1	

- Don't tame it, use it
  - Built lots of local models

IMPLICATIONS: OUTLIER REMOVAL

Remove odd training items
Examples:

- Keung & Kitchenham, IEEE TSE, 2008: effort estimation
- Kim et al., ICSE'11, defect prediction
  - case-based reasoning
  - prune neighboring rows containing too many contradictory conclusions.
- Yoon & Bae, IST journal, 2010, defect prediction
  - association rule learning methods to find frequent item sets.
  - Remove rows with too few frequent items.
  - Prunes 20% to 30% of rows.

Assumed, assumes a general pattern, muddle by some outliers

But my works says "its all outliers".

IMPLICATIONS: STRATIFIED CROSS-VALIDATION

Best to test on hold-out data

- That is similar to what will be seen in the future
- E.g. stratified cross validation

This work: “similar” is not a simple matter

- select cross-val bins via clustering
    - Train on neighboring cluster
    - Test on local cluster

Why learn from yourself?

- If the grass is greener on the other side of the fence
- Learn from your better neighbors

IMPLICATIONS: STRUCTURE LITERATURE REVIEWS

?

IMPLICATIONS: SBSE-1 (A.K.A. LEAP, THEN LOOK)

When faced with a new problem

- Jump off a cliff with roller skates and see where you stop.

That is:

- Define objective function and use it to guide a search engine.

IMPLICATIONS:
SBSE-2 (LOOK BEFORE YOU LEAP)

- Split
  - data on independent variables
- Prune
  - leaf quadrants using dependent variables
- Contrast.
  - Sort data in each cluster
  - Contrast intra-cluster data between good and bad examples
- Add Envy:
  - For each cluster C1…
  - Find C2; i.e. the neighboring clustering you most envy
  - Apply C2's rules to C1

THE COW DOCTRINE

- Seek the fence where the grass is greener on the other side.
  - Learn from there
  - Test on here
- Don't rely on trite definitions of "there" and "here"
  - Cluster to find "here" and "there"