CS886 Guest Lecture

John A. Doucette

David R. Cheriton School of Computer Science
University of Waterloo

October 15, 2014

About Me

- UW PhD Student in AI
- CS886 Alumnus
- 7 years doing ML research

This Talk

Outline

1. My CS886 Experience
2. Short Break?
3. Other Studies and Questions

The Point

- See an example of a "good" project
- Understand my thought process at the time
- Improve **your** projects, and avoid pitfalls

The Topic

Finding a Great Topic

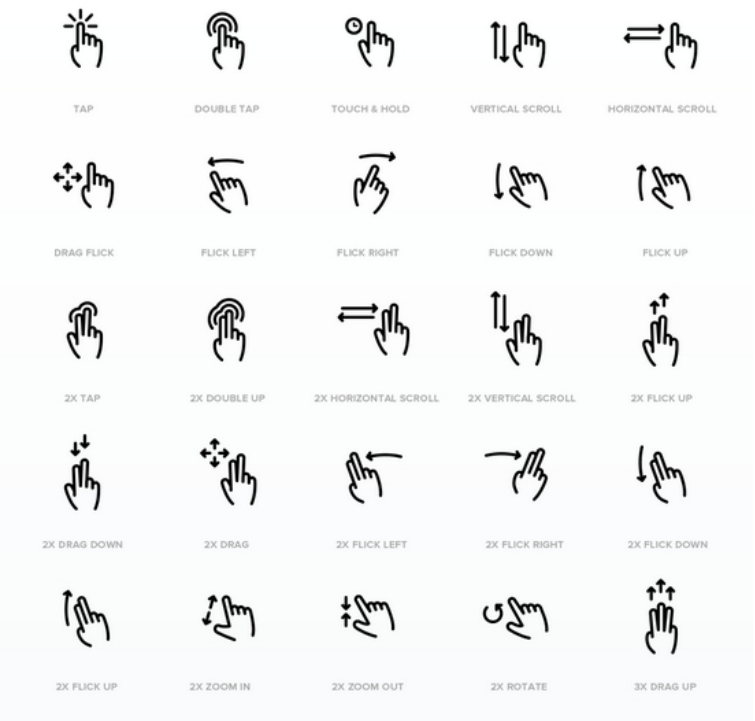
Pick a topic that:

- you know about

Finding a Great Topic

Pick a topic that:

- you know about
- you're excited about

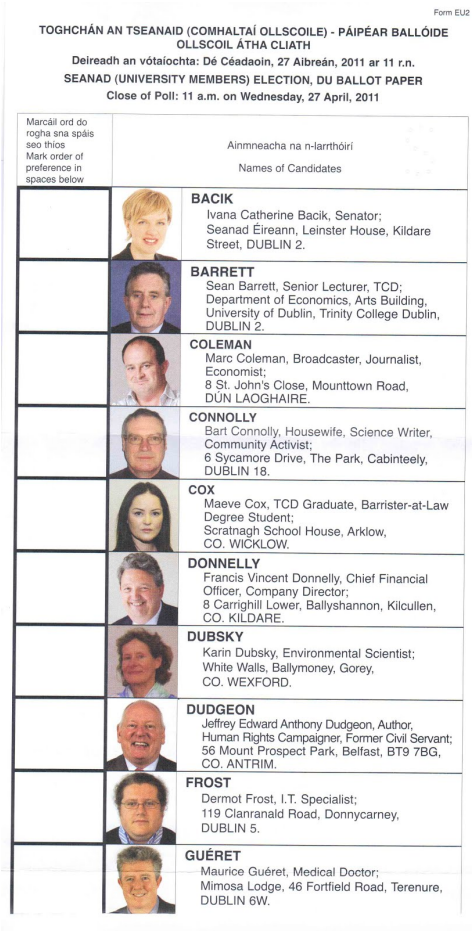


Finding a Great Topic

Pick a topic that:

- you know about
- you're excited about
- is self contained and testable

My Topic



Form EU2

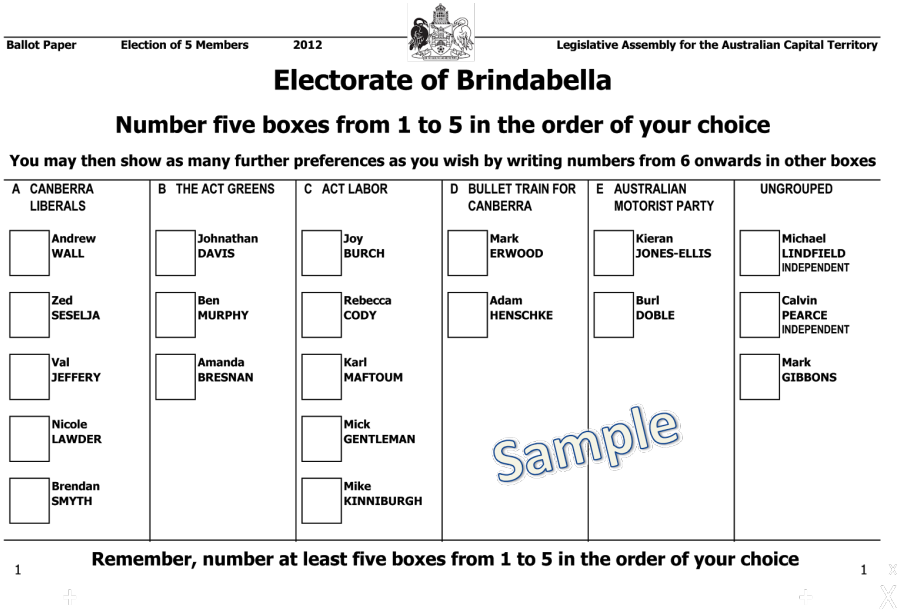
TOGHCHÁN AN TSEANAID (COMHALTAÍ OLLSCOILE) - PÁIPÉAR BALLÓIDE
OLLSCOIL ÁTHA CLIATH

Deireadh an vótaíochta: Dé Céadaoin, 27 Aibreán, 2011 ar 11 r.n.

SEANAD (UNIVERSITY MEMBERS) ELECTION, DU BALLOT PAPER

Close of Poll: 11 a.m. on Wednesday, 27 April, 2011

Marcáil ord do rogha sna spáis seo thíos Mark order of preference in spaces below	Ainmneacha na n-Iarrthóirí Names of Candidates
	BACIK Ivana Catherine Bacik, Senator; Seanad Éireann, Leinster House, Kildare Street, DUBLIN 2.
	BARRETT Sean Barrett, Senior Lecturer, TCD; Department of Economics, Arts Building, University of Dublin, Trinity College Dublin, DUBLIN 2.
	COLEMAN Marc Coleman, Broadcaster, Journalist, Economist; 8 St. John's Close, Mounttown Road, DÚN LAOGHAIRE.
	CONNOLLY Bart Connolly, Housewife, Science Writer, Community Activist; 6 Sycamore Drive, The Park, Cabinteely, DUBLIN 18.
	COX Maeve Cox, TCD Graduate, Barrister-at-Law Degree Student; Scratnagh School House, Arklow, CO. WICKLOW.
	DONNELLY Francis Vincent Donnelly, Chief Financial Officer, Company Director; 8 Carrighill Lower, Ballyshannon, Kilcullen, CO. KILDARE.
	DUBSKY Karin Dubsky, Environmental Scientist; White Walls, Ballymoney, Gorey, CO. WEXFORD.
	DUDGEON Jeffrey Edward Anthony Dudgeon, Author, Human Rights Campaigner, Former Civil Servant; 56 Mount Prospect Park, Belfast, BT9 7BG, CO. ANTRIM.
	FROST Dermot Frost, I.T. Specialist; 119 Clanranald Road, Donnycarney, DUBLIN 5.
	GUÉRET Maurice Guéret, Medical Doctor; Mimosa Lodge, 46 Fortfield Road, Terenure, DUBLIN 6W.



Ballot Paper Election of 5 Members 2012 Legislative Assembly for the Australian Capital Territory

Electorate of Brindabella

Number five boxes from 1 to 5 in the order of your choice

You may then show as many further preferences as you wish by writing numbers from 6 onwards in other boxes

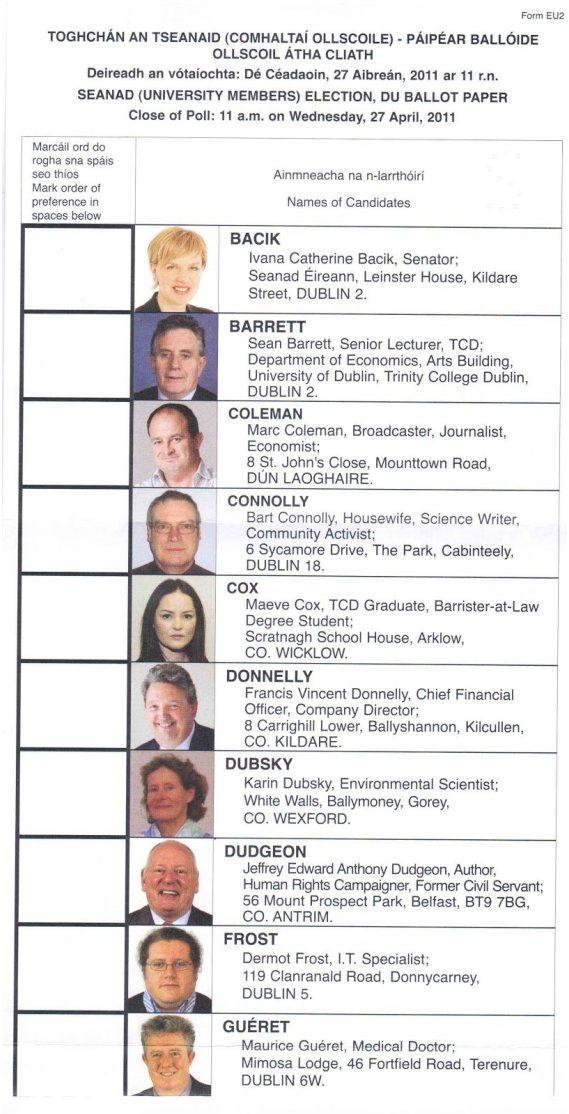
A CANBERRA LIBERALS	B THE ACT GREENS	C ACT LABOR	D BULLET TRAIN FOR CANBERRA	E AUSTRALIAN MOTORIST PARTY	UNGROUPED
Andrew WALL	Johnathan DAVIS	Joy BURCH	Mark ERWOOD	Kieran JONES-ELLIS	Michael LINDFIELD INDEPENDENT
Zed SESELJA	Ben MURPHY	Rebecca CODY	Adam HENSCHKE	Burl DOBLE	Calvin PEARCE INDEPENDENT
Val JEFFERY	Amanda BRESNAN	Karl MAFTOUM			Mark GIBBONS
Nicole LAWDER		Mick GENTLEMAN			
Brendan SMYTH		Mike KINNIBURGH			

Sample

1 **Remember, number at least five boxes from 1 to 5 in the order of your choice** 1

Voting

- A way to take collective actions
- Incorporates everyone's opinion
- (Potentially) produces a fair outcome



Form EU2

TOGHCHÁN AN TSEANAID (COMHALTAÍ OLLSCOILE) - PÁIPÉAR BALLÓIDE
OLLSCOIL ÁTHA CLIATH

Deireadh an vótaíochta: Dé Céadaoin, 27 Aibreán, 2011 ar 11 r.n.

SEANAD (UNIVERSITY MEMBERS) ELECTION, DU BALLOT PAPER

Close of Poll: 11 a.m. on Wednesday, 27 April, 2011

Marcáil ord do rogha sna spáis seo thíos Mark order of preference in spaces below	Ainmneacha na n-Iarrthóirí Names of Candidates
	BACIK Ivana Catherine Bacik, Senator; Seanad Éireann, Leinster House, Kildare Street, DUBLIN 2.
	BARRETT Sean Barrett, Senior Lecturer, TCD; Department of Economics, Arts Building, University of Dublin, Trinity College Dublin, DUBLIN 2.
	COLEMAN Marc Coleman, Broadcaster, Journalist, Economist; 8 St. John's Close, Mounttown Road, DÚN LAOGHAIRE.
	CONNOLLY Bart Connolly, Housewife, Science Writer, Community Activist; 6 Sycamore Drive, The Park, Cabinteely, DUBLIN 18.
	COX Maeve Cox, TCD Graduate, Barrister-at-Law Degree Student; Scratnagh School House, Arklow, CO. WICKLOW.
	DONNELLY Francis Vincent Donnelly, Chief Financial Officer, Company Director; 8 Carrighill Lower, Ballyshannon, Kilcullen, CO. KILDARE.
	DUBSKY Karin Dubsky, Environmental Scientist; White Walls, Ballymoney, Gorey, CO. WEXFORD.
	DUDGEON Jeffrey Edward Anthony Dudgeon, Author, Human Rights Campaigner, Former Civil Servant; 56 Mount Prospect Park, Belfast, BT9 7BG, CO. ANTRIM.
	FROST Dermot Frost, I.T. Specialist; 119 Clanranald Road, Donnycarney, DUBLIN 5.
	GUÉRET Maurice Guéret, Medical Doctor; Mimosa Lodge, 46 Fortfield Road, Terenure, DUBLIN 6W.

Applications of Voting

Form EU2

TOGHCHÁN AN TSEANAID (COMHALTAÍ OLLSCOILE) - PÁIPÉAR BALLÓIDE
OLLSCOIL ÁTHA CLIATH

Deireadh an vótaíochta: Dé Céadaoin, 27 Aibreán, 2011 ar 11 r.n.

SEANAD (UNIVERSITY MEMBERS) ELECTION, DU BALLOT PAPER

Close of Poll: 11 a.m. on Wednesday, 27 April, 2011

Marcáil ord do rogha sna spáis seo thíos Mark order of preference in spaces below	Ainmneacha na n-larrthóirí Names of Candidates
	BACIK Ivana Catherine Bacik, Senator; Seanad Éireann, Leinster House, Kildare Street, DUBLIN 2.
	BARRETT Sean Barrett, Senior Lecturer, TCD; Department of Economics, Arts Building, University of Dublin, Trinity College Dublin, DUBLIN 2.
	COLEMAN Marc Coleman, Broadcaster, Journalist, Economist; 8 St. John's Close, Mounttown Road, DÚN LAOGHAIRE.
	CONNOLLY Bart Connolly, Housewife, Science Writer, Community Activist; 6 Sycamore Drive, The Park, Cabinteely, DUBLIN 18.
	COX Maeve Cox, TCD Graduate, Barrister-at-Law Degree Student; Scratnagh School House, Arklow, CO. WICKLOW.
	DONNELLY Francis Vincent Donnelly, Chief Financial Officer, Company Director; 8 Carrighill Lower, Ballyshannon, Kilcullen, CO. KILDARE.
	DUBSKY Karin Dubsky, Environmental Scientist; White Walls, Ballymoney, Gorey, CO. WEXFORD.
	DUDGEON Jeffrey Edward Anthony Dudgeon, Author, Human Rights Campaigner, Former Civil Servant; 56 Mount Prospect Park, Belfast, BT9 7BG, CO. ANTRIM.
	FROST Dermot Frost, I.T. Specialist; 119 Clanranald Road, Donnycarney, DUBLIN 5.
	GUÉRET Maurice Guéret, Medical Doctor; Mimosa Lodge, 46 Fortfield Road, Terenure, DUBLIN 6W.

Ranked Ballot Voting

Ballot Paper Election of 5 Members 2012 Legislative Assembly for the Australian Capital Territory

Electorate of Brindabella

Number five boxes from 1 to 5 in the order of your choice

You may then show as many further preferences as you wish by writing numbers from 6 onwards in other boxes

A CANBERRA LIBERALS	B THE ACT GREENS	C ACT LABOR	D BULLET TRAIN FOR CANBERRA	E AUSTRALIAN MOTORIST PARTY	UNGROUPED
Andrew WALL	Johnathan DAVIS	Joy BURCH	Mark ERWOOD	Kieran JONES-ELLIS	Michael LINDFIELD INDEPENDENT
Zed SESELJA	Ben MURPHY	Rebecca CODY	Adam HENSCHKE	Burl DOBLE	Calvin PEARCE INDEPENDENT
Val JEFFERY	Amanda BRESNAN	Karl MAFTOUM			Mark GIBBONS
Nicole LAWDER		Mick GENTLEMAN			
Brendan SMYTH		Mike KINNIBURGH			

Sample

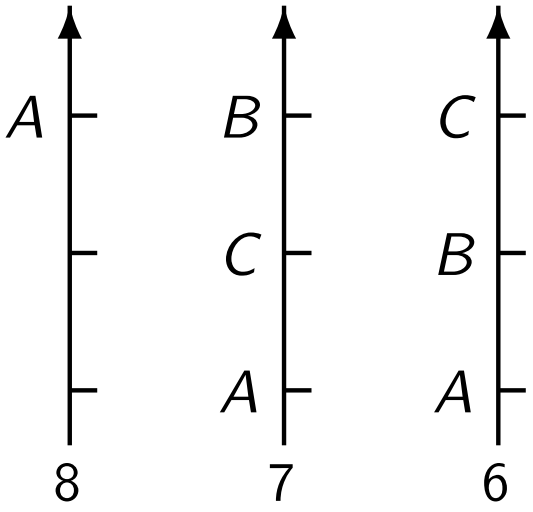
Remember, number at least five boxes from 1 to 5 in the order of your choice

1 1

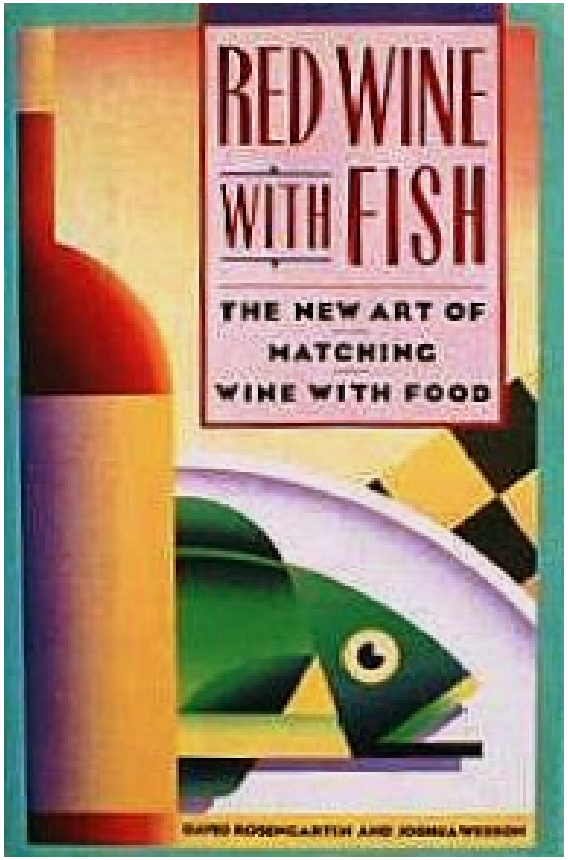
Why Use Ranked Ballot Voting?

Advantages

- Easier to be fair
- Easy to find nth-place
- Lots of systems



Why Use Ranked Ballot Voting?



Meta: Topic Selection

What I was thinking:

- Voting is exciting!
- New to me, but some experience
- Need a self-contained topic

A Self-Contained Topic

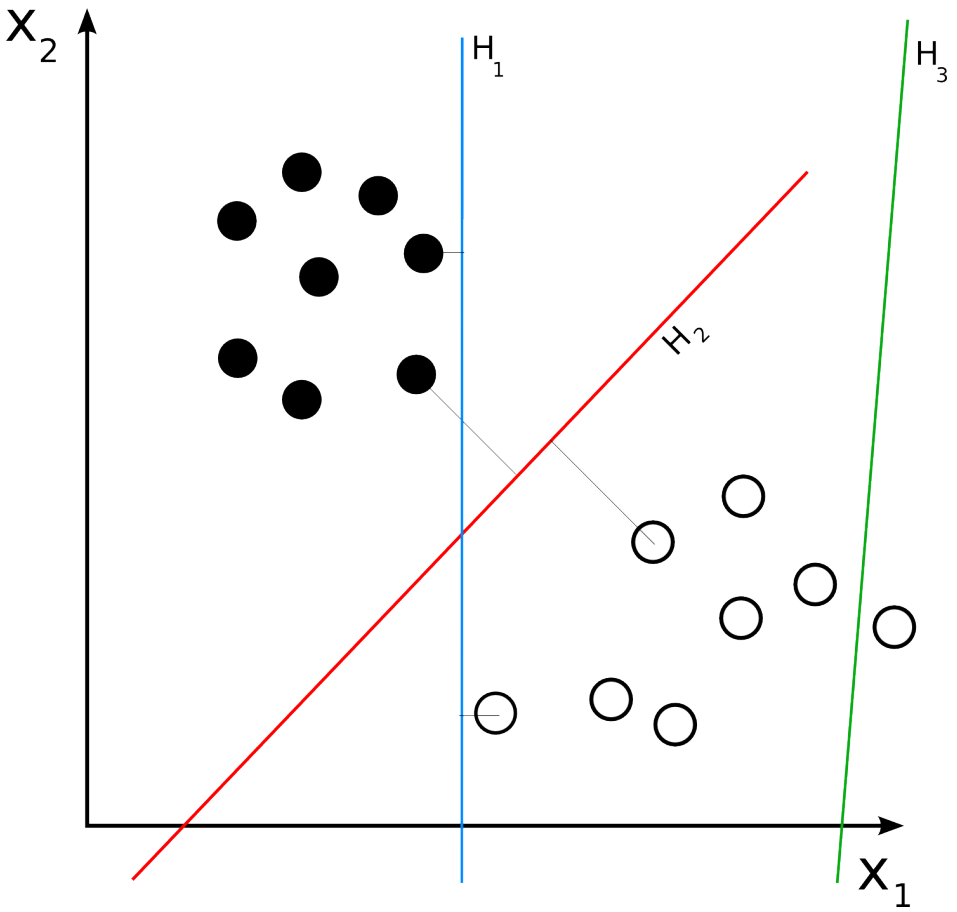
Completing a ranked ballot is hard:

- Expensive to learn about candidates
- Too many candidates
- Too many interdependent values

A Self-Contained Topic

Could we use learning to help?

- Candidates often *ordered*
- Voters often predictable
- Voters think alike



Can ML Help?

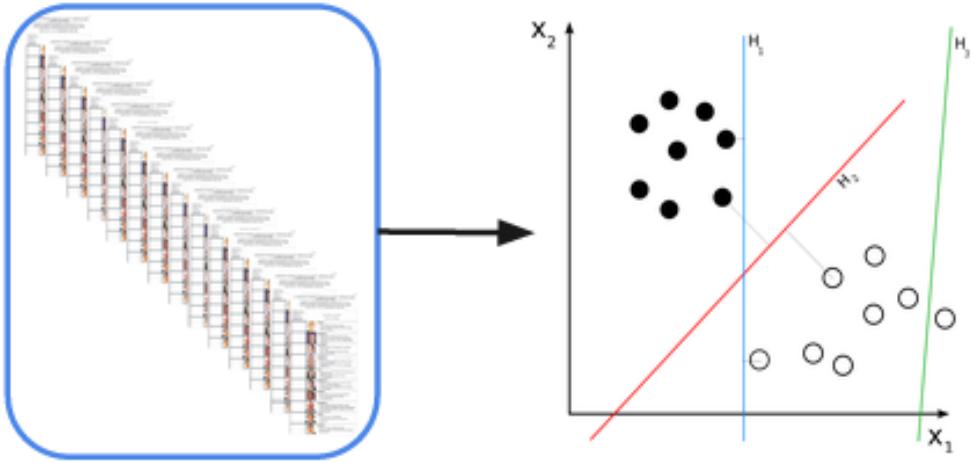
TOGHCHÁN AN TSEANAID (COMHALTAÍ OLLSCOILE) - PÁIPÉAR BALLÓIDE
OLLSCOIL ÁTHA CLIATH

Deireadh an vótaíochta: Dé Céadaoin, 27 Aibreán, 2011 ar 11 r.n.

SEANAD (UNIVERSITY MEMBERS) ELECTION, DU BALLOT PAPER

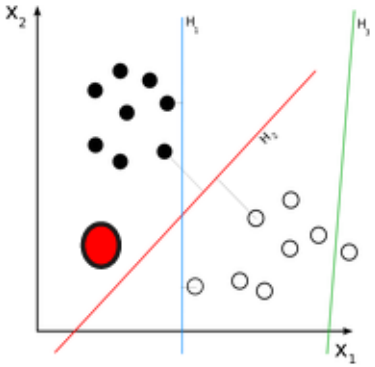
Close of Poll: 11 a.m. on Wednesday, 27 April, 2011

Marcáil ord do rogha sna spás seo thíos / Mark order of preference in spaces below	Ainmneacha na n-iarrthóirí / Names of Candidates
5	**BACIK** Ivana Catherine Bacik, Senator; Seanad Éireann, Leinster House, Kildare Street, DUBLIN 2.
?	**BARRETT** Sean Barrett, Senior Lecturer, TCD; Department of Economics, Arts Building, University of Dublin, Trinity College Dublin, DUBLIN 2.
?	**COLEMAN** Marc Coleman, Broadcaster, Journalist, Economist; 8 St. John's Close, Mounttown Road, DUN LAOGHAIRE.
3	**CONNOLLY** Bari Connolly, Housewife, Science Writer, Community Activist; 6 Sycamore Drive, The Park, Cabinteely, DUBLIN 18.
4	**COX** Maeve Cox, TCD Graduate, Barrister-at-Law Degree Student; Scrahagh School House, Arklow, CO. WICKLOW.
2	**DONNELLY** Francis Vincent Donnelly, Chief Financial Officer, Company Director; 8 Carrighill Lower, Ballyshannon, Kilcullen, CO. KILDARE.
1	**DUBSKY** Karin Dubsky, Environmental Scientist; White Walls, Ballymoney, Gorey, CO. WEXFORD.
?	**DUDGEON** Jeffrey Edward Anthony Dudgeon, Author, Human Rights Campaigner, Former Civil Servant; 56 Mount Prospect Park, Belfast, BT9 7BG, CO. ANTRIM.
?	**FROST** Dermot Frost, I.T. Specialist; 119 Clanranald Road, Donnycarney, DUBLIN 5.
?	**GUÉRET** Maurice Guéret, Medical Doctor; Mimosa Lodge, 46 Fortfield Road, Terenure, DUBLIN 6W.



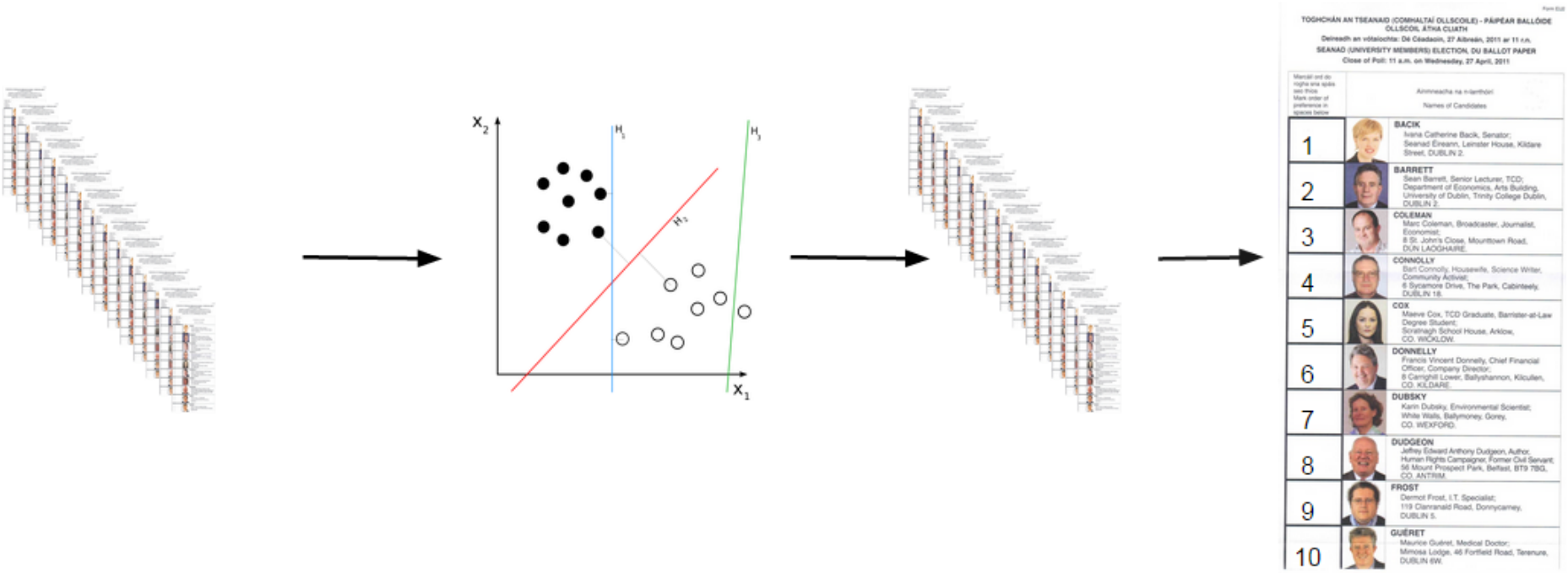
Can ML Help?


TOGHCHÁN AN TSEANAID (COMHALTAÍ OLLSCOILE) - PÁIPÉAR BALLÓIDE
OLLSCOIL ÁTHA CLIATH
SEANAD (UNIVERSITY MEMBERS) ELECTION, DU BALLOT PAPER
Close of Poll: 11 a.m. on Wednesday, 27 April, 2011
Ainmneacha na n-iarthóirí
Names of Candidates
5
BACIK
Ivana Catherine Bacik, Senator;
Seanad Éireann, Leinster House, Kildare Street, DUBLIN 2.
?
BARRETT
Sean Barrett, Senior Lecturer, TCD;
Department of Economics, Arts Building, University of Dublin, Trinity College Dublin, DUBLIN 2.
?
COLEMAN
Marc Coleman, Broadcaster, Journalist, Economist;
8 St. John's Close, Mounttown Road, DUN LAOGHAIRE.
3
CONNOLLY
Bari Connolly, Housewife, Science Writer, Community Activist;
6 Sycamore Drive, The Park, Cabinteely, DUBLIN 18.
4
COX
Maeve Cox, TCD Graduate, Barrister-at-Law Degree Student;
Scratnagh School House, Arklow, CO. WICKLOW.
2
DONNELLY
Francis Vincent Donnelly, Chief Financial Officer, Company Director;
8 Carrighill Lower, Ballyshannon, Kilcullen, CO. KILDARE.
1
DUBSKY
Karin Dubsky, Environmental Scientist;
White Walls, Ballymoney, Gorey, CO. WEXFORD.
?
DUDGEON
Jeffrey Edward Anthony Dudgeon, Author, Human Rights Campaigner, Former Civil Servant;
56 Mount Prospect Park, Belfast, BT9 7BQ, CO. ANTRIM.
?
FROST
Dermot Frost, I.T. Specialist;
119 Clanranald Road, Donnycarney, DUBLIN 5.
?
GUÉRET
Maurice Guéret, Medical Doctor;
Mimosa Lodge, 46 Fortfield Road, Terenure, DUBLIN 6W.


x_2
x_1
H_1
H_2
H_3


TOGHCHÁN AN TSEANAID (COMHALTAÍ OLLSCOILE) - PÁIPÉAR BALLÓIDE
OLLSCOIL ÁTHA CLIATH
SEANAD (UNIVERSITY MEMBERS) ELECTION, DU BALLOT PAPER
Close of Poll: 11 a.m. on Wednesday, 27 April, 2011
Ainmneacha na n-iarthóirí
Names of Candidates
5
BACIK
Ivana Catherine Bacik, Senator;
Seanad Éireann, Leinster House, Kildare Street, DUBLIN 2.
6
BARRETT
Sean Barrett, Senior Lecturer, TCD;
Department of Economics, Arts Building, University of Dublin, Trinity College Dublin, DUBLIN 2.
7
COLEMAN
Marc Coleman, Broadcaster, Journalist, Economist;
8 St. John's Close, Mounttown Road, DUN LAOGHAIRE.
3
CONNOLLY
Bari Connolly, Housewife, Science Writer, Community Activist;
6 Sycamore Drive, The Park, Cabinteely, DUBLIN 18.
4
COX
Maeve Cox, TCD Graduate, Barrister-at-Law Degree Student;
Scratnagh School House, Arklow, CO. WICKLOW.
2
DONNELLY
Francis Vincent Donnelly, Chief Financial Officer, Company Director;
8 Carrighill Lower, Ballyshannon, Kilcullen, CO. KILDARE.
1
DUBSKY
Karin Dubsky, Environmental Scientist;
White Walls, Ballymoney, Gorey, CO. WEXFORD.
9
DUDGEON
Jeffrey Edward Anthony Dudgeon, Author, Human Rights Campaigner, Former Civil Servant;
56 Mount Prospect Park, Belfast, BT9 7BQ, CO. ANTRIM.
10
FROST
Dermot Frost, I.T. Specialist;
119 Clanranald Road, Donnycarney, DUBLIN 5.
8
GUÉRET
Maurice Guéret, Medical Doctor;
Mimosa Lodge, 46 Fortfield Road, Terenure, DUBLIN 6W.

Can ML Help?


x_2
x_1
H_1
H_2
H_3
TOGHCHÁN AN TSEANAID (COMHALTAÍ OLLSCOILE) - PÁIPÉAR BALLÓIDE
OLLSCOIL ÁTHA CLIATH
Deireadh an vótaíochta: Dé Céadaoin, 27 Aibreán, 2011 ar 11 r.n.
SEANAD (UNIVERSITY MEMBERS) ELECTION, DU BALLOT PAPER
Close of Poll: 11 a.m. on Wednesday, 27 April, 2011
Ainmneacha na n-iarrthóirí
Names of Candidates
1
BACIK
Ivana Catherine Bacik, Senator;
Seanad Éireann, Leinster House, Kildare Street, DUBLIN 2.
2
BARRETT
Sean Barrett, Senior Lecturer, TCD;
Department of Economics, Arts Building, University of Dublin, Trinity College Dublin, DUBLIN 2.
3
COLEMAN
Marc Coleman, Broadcaster, Journalist, Economist;
8 St. John's Close, Mounttown Road, DUN LAOGHAIRE.
4
CONNOLLY
Bari Connolly, Housewife, Science Writer, Community Activist;
6 Sycamore Drive, The Park, Cabinteely, DUBLIN 18.
5
COX
Maeve Cox, TCD Graduate, Barrister-at-Law Degree Student;
Scoilnagh School House, Arklow, CO. WICKLOW.
6
DONNELLY
Francis Vincent Donnelly, Chief Financial Officer, Company Director;
8 Carrighill Lower, Ballyshannon, Kilcullen, CO. KILDARE.
7
DUBSKY
Karin Dubsky, Environmental Scientist;
White Walls, Ballymoney, Gorey, CO. WEXFORD.
8
DUDGEON
Jeffrey Edward Anthony Dudgeon, Author, Human Rights Campaigner, Former Civil Servant;
56 Mount Prospect Park, Belfast, BT9 7BQ, CO. ANTRIM.
9
FROST
Dermot Frost, I.T. Specialist;
119 Clanranald Road, Donnycarney, DUBLIN 5.
10
GUÉRET
Maurice Guéret, Medical Doctor;
Mimosa Lodge, 46 Fortfield Road, Terenure, DUBLIN 6W.

Prior Work

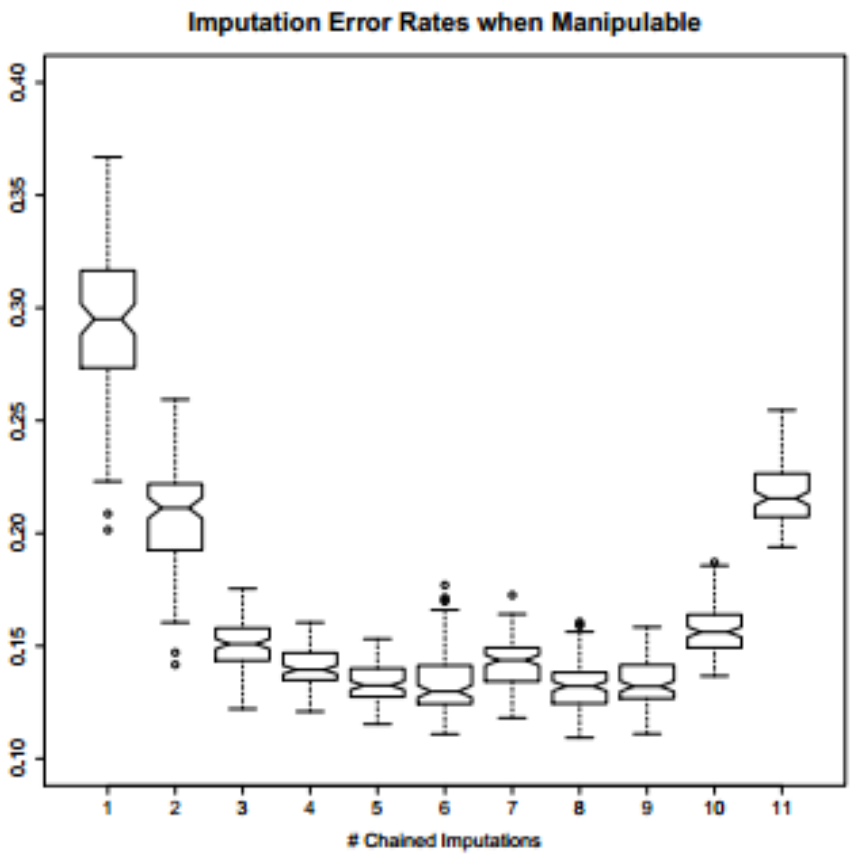
Previously Shown

- "Safe" when done carefully
- "Possible" to learn preferences
- Lots of bias toward popular candidates.

Outstanding Questions

Questions:

- Which algorithm works best?
- Which features are predictive?
- Are we fair to all candidates?



The Final Topic

- Find the most accurate combination of features and algorithms
- Find the least unfair combination of features and algorithms
- Compare across many datasets to eliminate bias

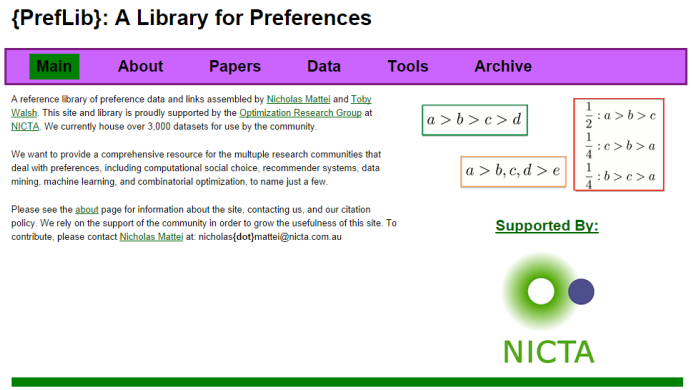
Narrowing Your Topic

- Find an expert and ask for ideas
- Find papers that solve problems without ML
- Make sure no one's solved your problem already!

How I Found Data

Finding Voting Data

1. Read papers in the area, found "standard" data
2. Browsed government databases, found some data, but low quality
3. Asked around, heard about a new repository



Meta: Tips for Finding Data

Finding Data

- Pitch to everyone
- If a paper uses the dataset you want, email the authors!
- Don't be afraid to spend some time finding the best data you can

Aside: Data Authenticity

Issues with data repositories:

- Research optimizes against the repo
- Repo data is not always realistic
- Active debate in many areas, be aware

Preprocessing

My Data

- Clean, uniform format.
- Not all usable (poor ground truth)
- Format is non-standard, and poorly documented

```
12
1,Cathal Boland F.G.
2,Clare Daly S.P.
3,Mick Davis S.F.
4,Jim Glennon F.F.
5,Ciaran Goulding Non-P
6,Michael Kennedy F.F.
7,Nora Owen F.G.
8,Eamonn Quinn Non-P
9,Sean Ryan Lab
10,Trevor Sargent G.P.
11,David Henry Walshe C.C. Csp
12,G.V. Wright F.F.
43942,43942,19299
800,12,6,4
680,4,6,12
506,6,12,4
486,12,4,6
429,6,4,12
367,4,12,6
343,10
```

Preprocessing

Cleaning

- Removed unusable data
- Is what's left representative?

```
4
1,Branden Robinson
2,Raphael Hertzog
3,Bdale Garbee
4,None Of The Above
475,475,41
60,3,1,2,4
50,1,3,2,4
40,3,1,2
34,3,2,1,4
31,3,2,4,1
29,2,3,1,4
29,1,3,2
24,2,1,3,4
```

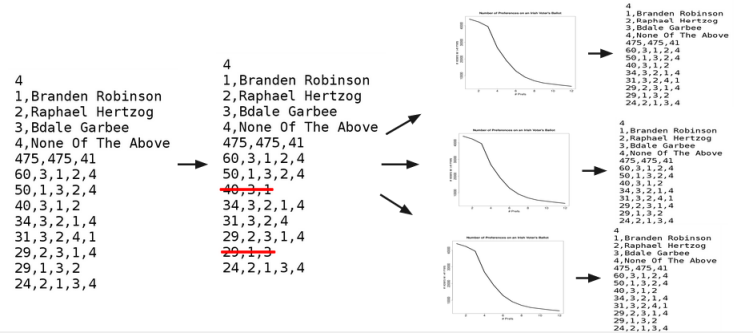
→

```
4
1,Branden Robinson
2,Raphael Hertzog
3,Bdale Garbee
4,None Of The Above
475,475,41
60,3,1,2,4
50,1,3,2,4
~~40,3,1~~
34,3,2,1,4
31,3,2,4
29,2,3,1,4
~~29,1,3~~
24,2,1,3,4
```

Preprocessing

Ablation Studies

1. Measure missingness in raw data
2. Clean
3. Randomly ablate cleaned data using measured distribution
4. Repeat 3 many times



Aside: Preprocessing

Problems

- Do voters with complete ballots have different views?
- Ablation study doesn't account for this
- **Check pertinent aggregates in raw and ablated sets**

Preprocessing

Feature Extraction

1. Absolute Positions
2. Relative Positions
3. Indicators
4. Triplet relations

$a \succ b \succ c$

1. Pos(a) $= 1$
2. Pos(a) - Pos(b) $=$ -1
3. $\mathrm{I}(a \succ b) = 1$
4. $\mathrm{I}(a \succ b \wedge a \succ c) = 1$

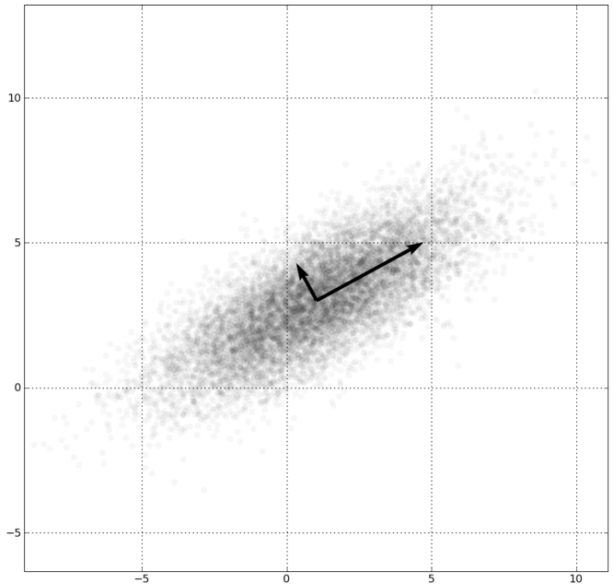
Picking Good Features

- Information *you* use to answer the question
- Information that your algorithm can work with
- Pitch to others, ask how what they'd use

Preprocessing

Feature Selection

- Small scale testing to find potentially useful techniques
- Settled on IG and PCA
- Parametrization matters



Tools

- Bash/Perl/Python scripts to turn raw data into matrix of features (fast prototyping, easy to verify correctness)
- Heavier tools (mostly R) to run feature selection
- Not always easy to use

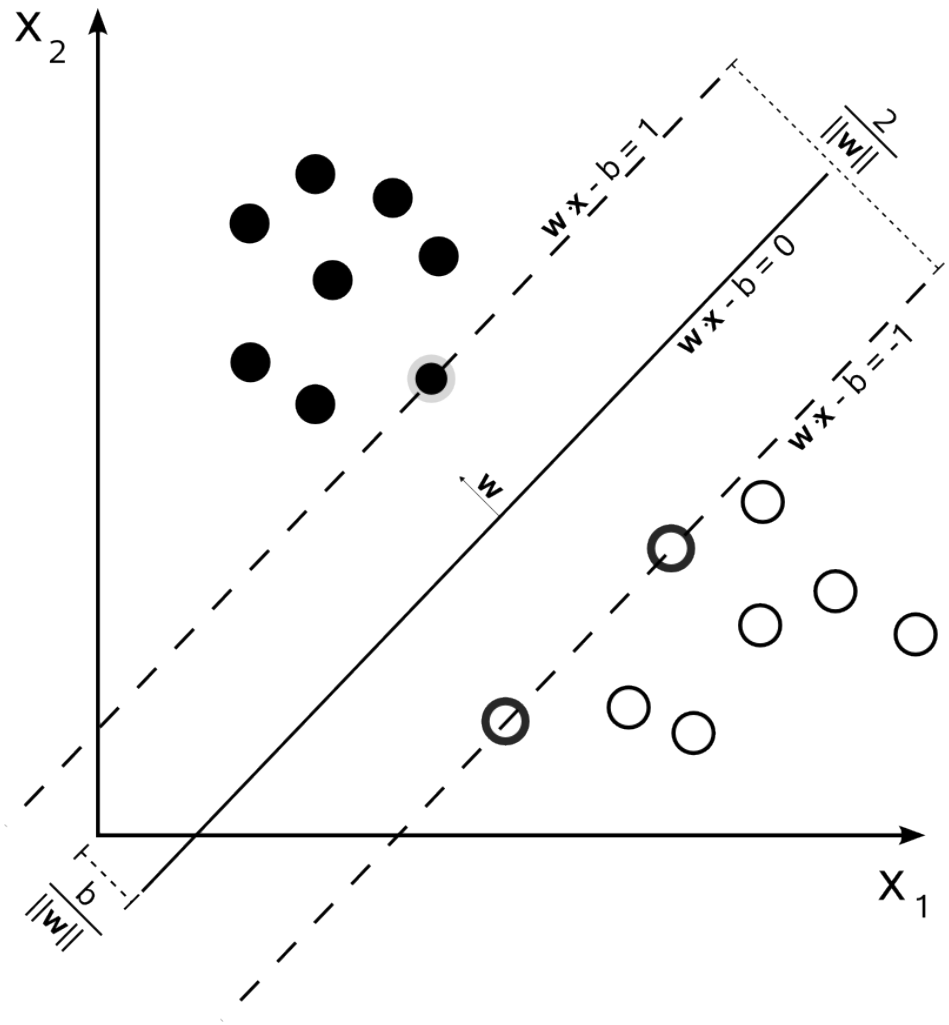
Development Choices

- Fast development is hard to maintain, and bugs are hard to find
- More is learned from rolling your own, but get it right!
- No one thinks "I wish I'd spent less time designing my experiment."

Classifier Selection

Classifiers for Voting

- Multinomial Dist. Learner (multi-class)
- Naive Bayes (baseline)
- SVM (higher order relations)



Finding the Right ML Algorithms

- May not find the "right" algorithm
- Should prefer available/working alg to "right" alg
- Workshop and tutorial proceedings to find esoteric algs

Meta: Classifier Selection

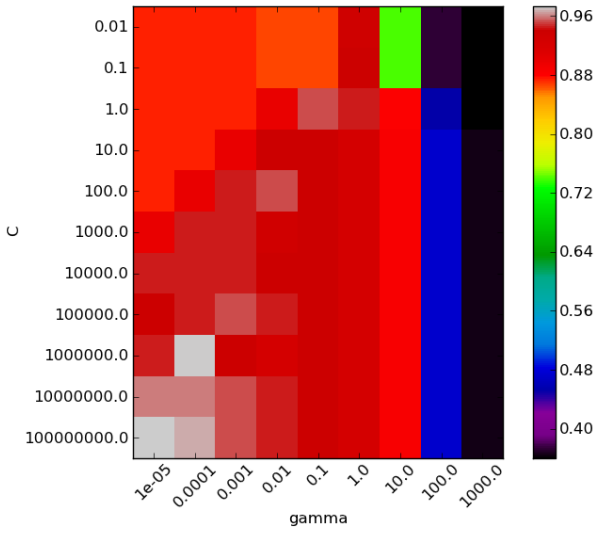
Use or Build?

- Build to learn more about how the algorithm works
- Use for faster speeds and more features
- Build if custom data formats are important

Meta: Using Classifiers

Know your Tools!

- **Parameters matter**
- Implementations are wrong or broken
- Implementations are "temperamental"
- Algorithms are "temperamental"



Pedregosa et al. 2011

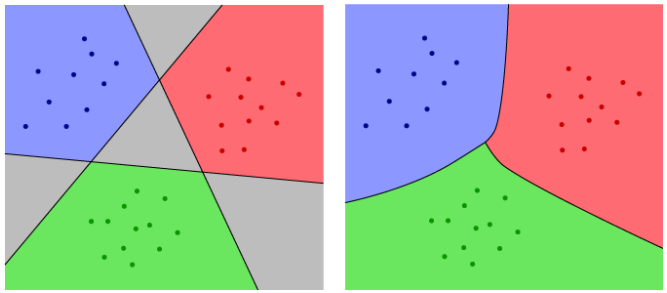
My Problems

- Multinomial implementation broke with redundant features
- NB implementation didn't smooth by default
- SVM needed lots and lots of expensive parameter tuning

Using Classifiers

Multi-class Classification

- Not all algorithms support > 2 classes
- Lots don't handle *constrained* MCC
- More than doubled implementation time



Budney 2013

Experiment Design

The Basics

- 10 sets
- 3 Feature Sets
- 3 Classifiers
- 100 Samples
- Over **9000** runs!

Additional Concerns

- Bootstrapping with training/test, not CV
- **Identical** data to allow paired tests
- Control of FDR

Performance Measures

BC Error

$$\frac{\sum_{c \in C} |count(c) - G(c)|}{\sum_{c \in C} G(c)}$$

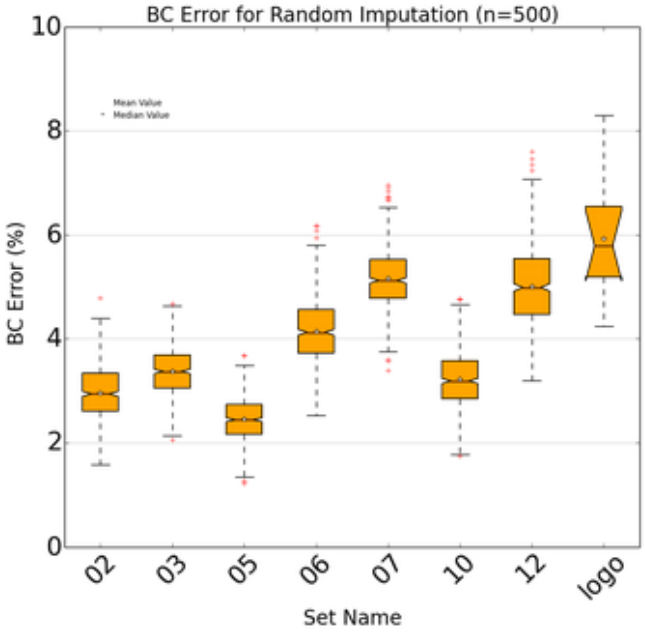
Bias

$$R^2 = cor(count(c) - G(c), G(c))^2$$

Meta: Comparisons

Designing Performance Comparisons

- What is the system for?
- What's the simplest technique?
- What's the best technique?



Meta: Computational Limits

CPU Limited Experiments

- Write or use multi-threaded code
- Select implementation languages with this in mind
- Get some extra CPU power
- Reduce experiment size

```
void trainModel(ProblemInstance pi, ulong length, immutable Candidate[ulong] alts){
        T[] results = new T[alts.length];
        foreach(ulong current_class ; parallel(iota(1,alts.length+1))){
                Vector y_i;
                foreach(ulong j; pi.y.keys)
                        y_i[j] = pi.y[j]==current_class?1.0:-1.0;
                results[current_class-1] = new T();
                results[current_class-1].train(pi.x, assumeUnique(y_i));
        }
        foreach(ulong i, T model; results){
                this.models[length][alts[i+1].dup] = model;
        }
```

Analysis

Processing Results

- R is great (e.g. pairwise.t.test)
- Tables often better than graphs

	SVM		
	PCA	IG	plain
% Borda Error			
Debian_1	0.008	0.007	0.008
Debian_2	0.009	**0.005**	0.009
Debian_3	**0.014**	**0.013**	**0.016**
Debian_4	**0.013**	**0.013**	**0.015**
Debian_5	**0.033**	**0.031**	**0.033**
Debian_6	**0.004**	**0.004**	**0.004**
Debian_7	0.011	0.011	0.011
Debian_8	**0.006**	**0.008**	**0.006**
Dublin_North	1.084	**0.923**	1.08
Dublin_West	**0.549**	**0.458**	0.575

Results

- SVM + IG is the best combination of accuracy and fairness
- Multinomial & NB generally do not perform well

Writing Tips

- Start now: you have everything but the results
- Read some other papers for style/layout
- Get as much feedback as possible

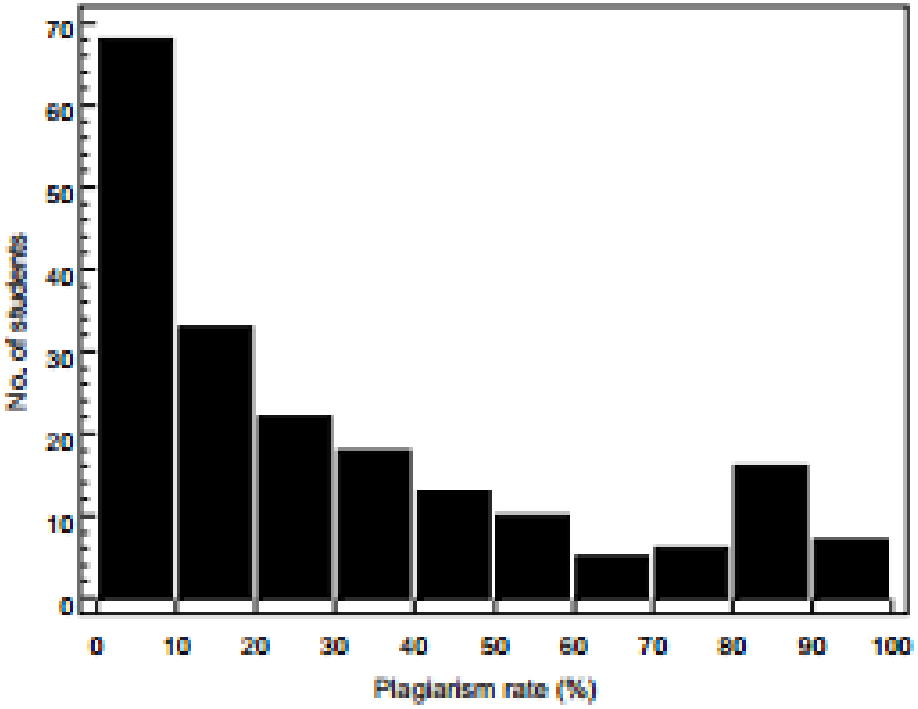
Tips for Reviewing Papers

- Don't prioritize spelling & grammar
- Think carefully about the paper's methodology
- Does the paper do what it claims?

Meta: Academic Honesty

Plagiarism

- Assume about 20% plagiarism rate (yes, even here)
- Breaks in text, changes in voice, stylistic glitches
- Not necessarily malicious



Biliæ-Zulle et al. 2005

Conclusions

- Project wrapped up nicely, despite many hiccups
- Concrete question answered, definitively
- More questions generated, but the constructed testbed made them easy to answer too

Future WorkL

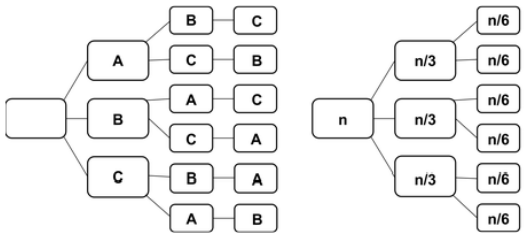
ots of New Questions

- IG parameters never optimized
- Does this actually select good outcomes?
- Would a tailored method work better?

Meta: Where am I now?

Since then:

- Instant thesis topic
- Much improved mining system
- Custom ML models and better metrics



$$|\hat{\theta}_i - \theta_i| < \frac{t\sqrt{\theta_i(1-\theta_i)}}{\sqrt{n_i} + \frac{|C|}{\sqrt{n_i}}} + \frac{\theta_i|C|+1}{n_i+|C|}$$

$$n_i = n_{i-1}(\theta_i - \epsilon)$$

Break

10 minute break / come talk to me.

Discussion and/or Other Topics

We can:

- Hangout / do Q & A
- Talk about working with temporal data at edX
- Talk about evolutionary approaches, multi-objective learning, and embedded feature selection
- Talk about Compression classifiers, the stock market, and ML on unstructured data

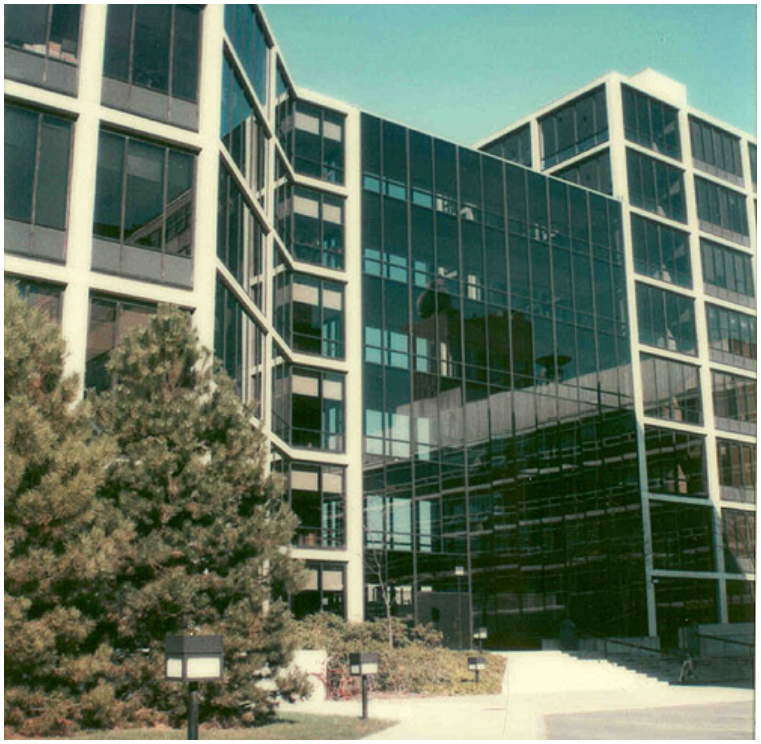
edX

EdX:

- Massive Online Classroom service
- Thousands of students per class
- Behavior tracked to the millisecond
- Demographics, behavior and grades are linked

The Data

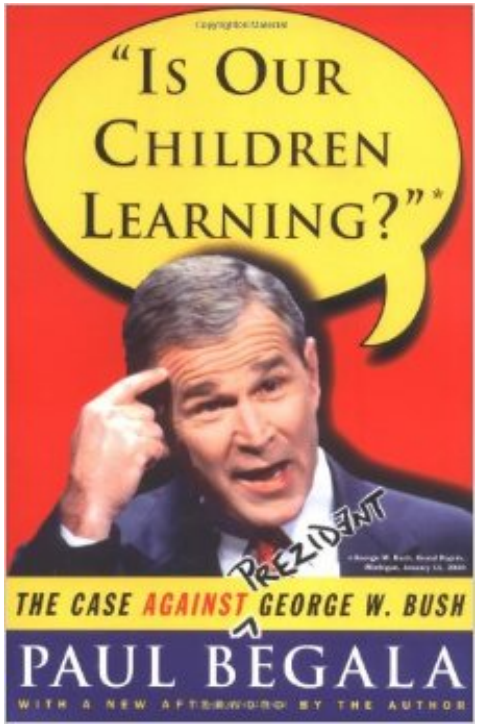
- Held at MIT on airgapped machines
- Strict usage protocol
- Haphazard processing
- Stratified sampling



The Question

Research Questions

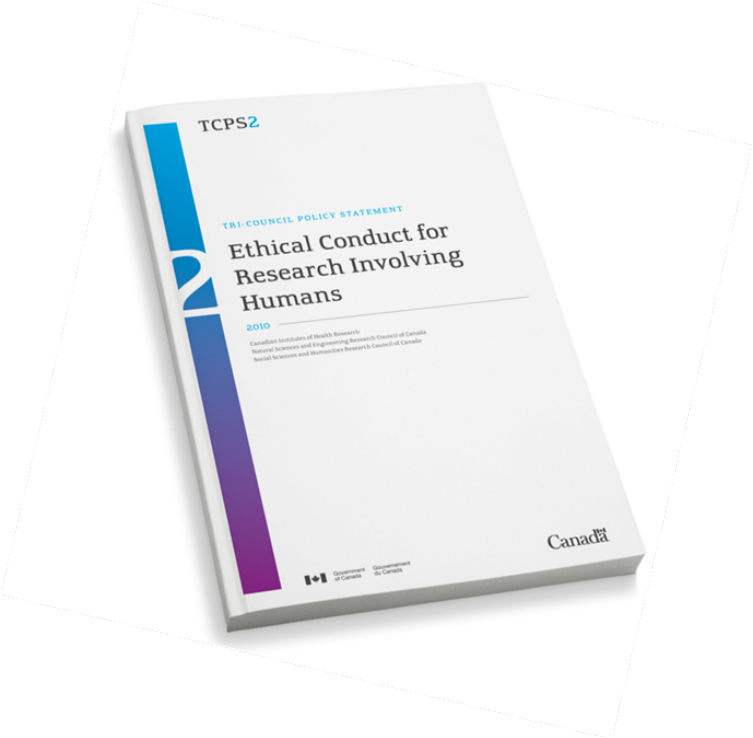
- Who is learning?
- How much do they learn?
- **What makes them learn better?**



Ethics Training

Training

- 12 hour US training
- 6 hour Canadian training
- IRB approval as additional researcher



Processing and Confounds

Processing

- Demographics explain many effects when present
- Many non-numeric fields
- Many students DNF

Table 1

Average Ability of Various Cohorts

Cohort	% of 8.MReV	Average ability in 8.MReV: Mean	Standard deviation	Standard error	Relative improvement (SE)
PhD	8	0.67	0.93	.10	0.16 (0.06)
Masters	19	0.26	0.91	.06	-0.06 (0.05)
College	29	-0.08	0.99	.06	-0.11 (0.04)
High school	11	-0.20	0.93	.07	-0.11 (0.06)
Less than HS	6	-0.05	0.84	.10	-0.21 (0.10)
No response	23	0.02	1.04	.07	0.01 (0.07)
Physics teachers	17	0.39	0.97	.07	0.00 (0.05)
8.011 students	3	-1.05	0.50	.08	-

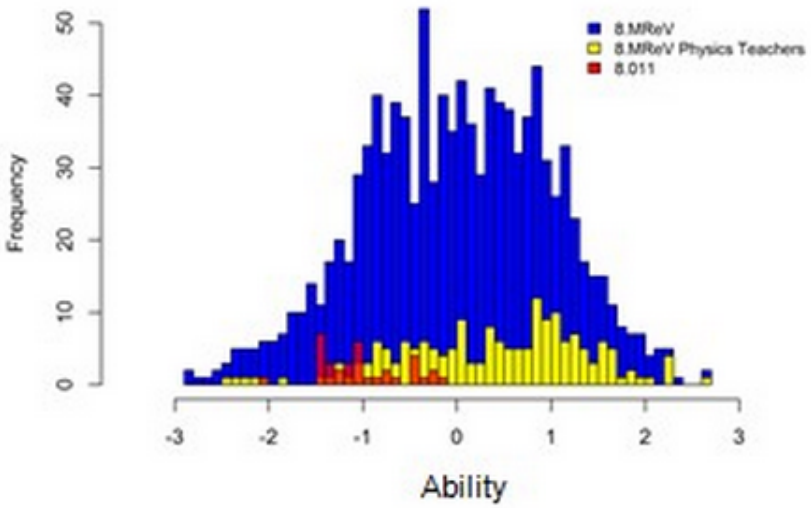
Note. Degree listed is highest degree attained. For example, "High School" refers to students who have obtained a high school diploma and may be enrolled in college.

Colvin et al. 2014

First Attempts

Classification

- Predict over/under performance
- Aggregate time spent studying + demographics
- Only generally predictive features are pretest and slope

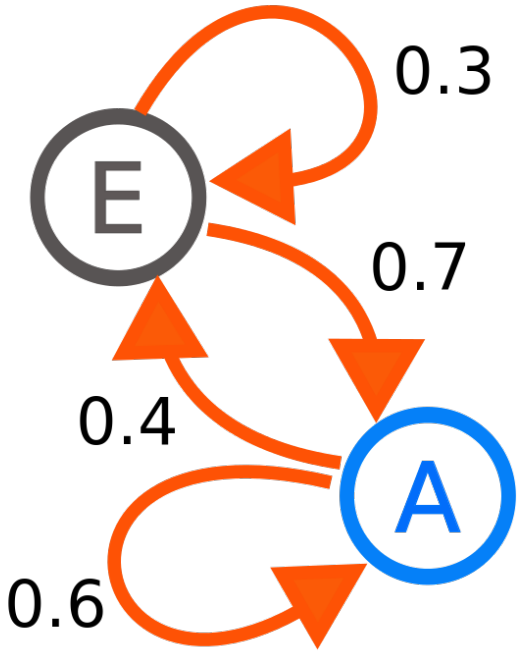


Colvin et al. 2014

Transition Model

Idea

- Not what you do, but how you do it
- Order of interaction could be predictive
- Learn as a Markov Chain



Joxemai and Falcorian 2014

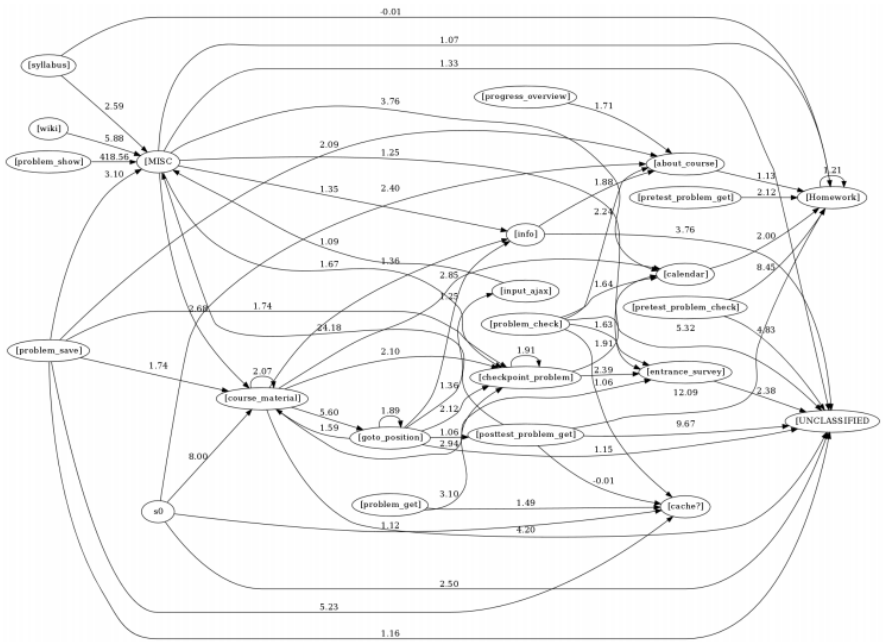
Data Issues

- Time window or event counts?
- Window size and automatic-transition events
- Normalized or raw counts?

Initial Analysis

Analysis Issues

- Support levels for transitions
- Effect sizes
- Accounting for FDR



More Careful Analysis

Useful Results

- Build "typical" models for over-/under-performers
- Pooled t-tests for parameter differences
- Controlled with support & FDR

From	To	Relative Risk (sample size)
[problem_check]	[problem_show]	1.682 (56)
[save_problem_check]	[goto_position]	1.300 (180)
[checkpoint_problem]	[goto_position]	0.979 (682)
[problem_get]	[course_material]	0.978 (679)
[course_material]	[problem_get]	0.966 (682)
[course_material]	[goto_position]	0.961 (682)
[checkpoint_problem]	[course_material]	0.961 (681)
[goto_position]	[problem_check]	0.940 (665)
[problem_get]	[about_course]	0.937 (603)
[course_material]	[posttest_problem_get]	0.932 (554)
[pretest_save_problem_check]	[pretest_save_problem_check]	0.913 (89)

Example Results

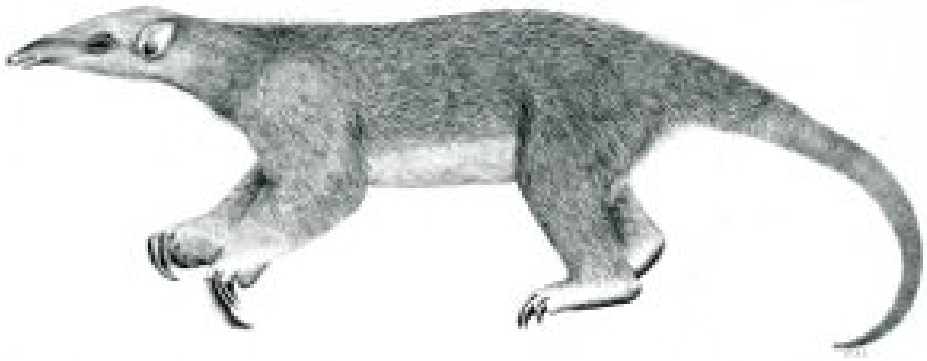
- Frequently looking at your grades $\downarrow$
- Doing many homework problems in a row $\uparrow$
- Frequent homework interruptions $\downarrow$

Embedded Feature Selection and Evolutionary Approaches

Problem Domain

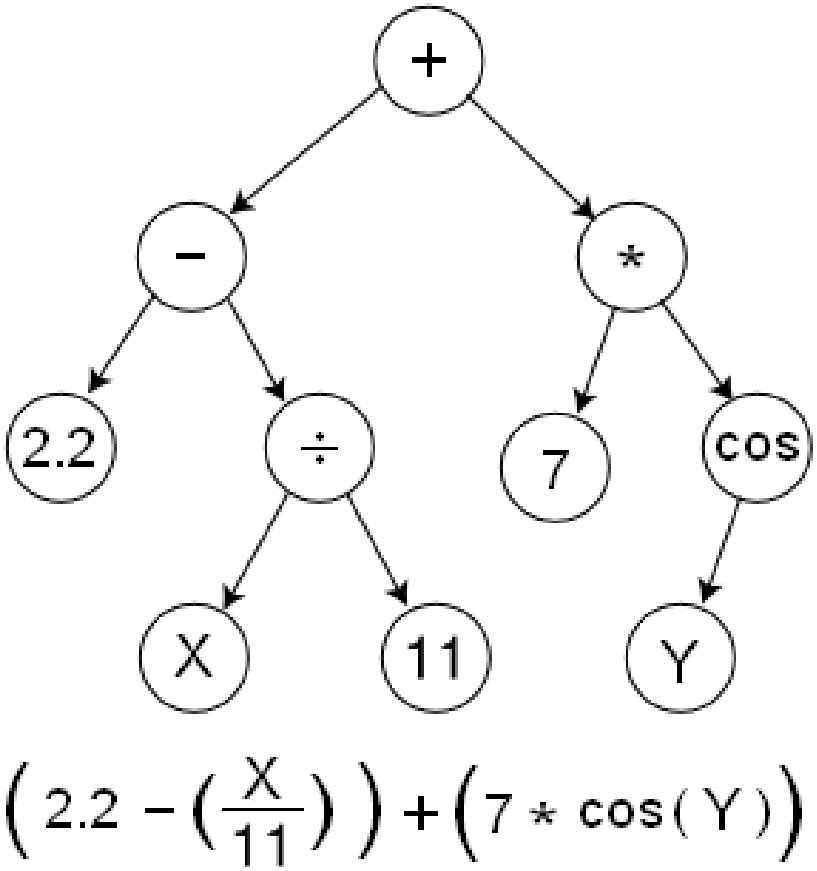
Traffic Analysis

- *Very* fast models
- Minimal, easily computed features
- Potentially frequent retraining



Genetic Programming

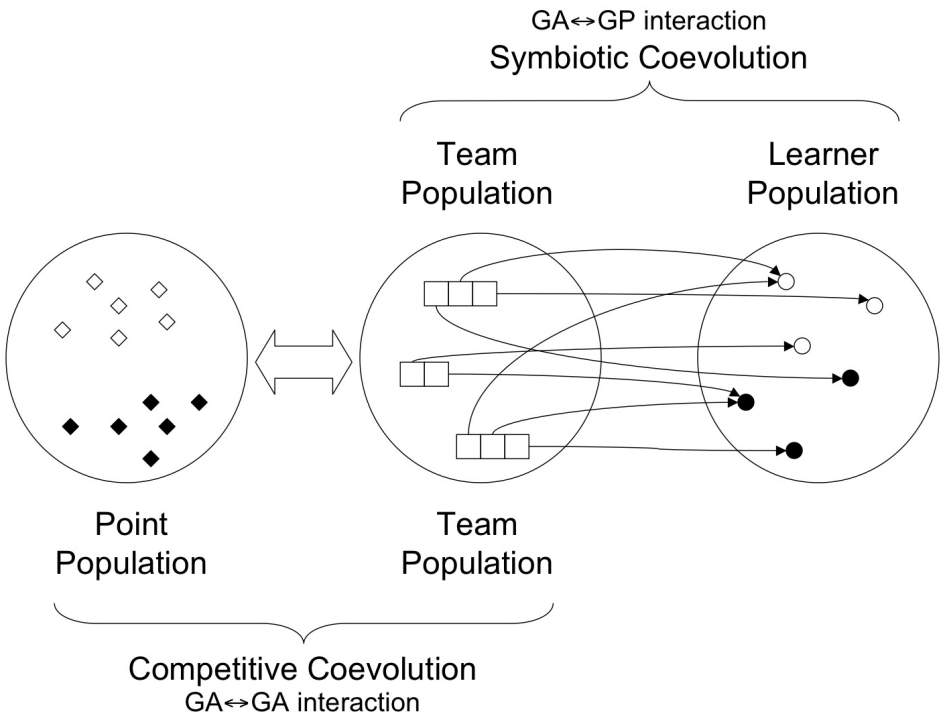
- 'Evolve' a program mapping inputs to outputs
- Pros: Programs can be *complex* and *arbitrary* mappings.
- Cons: Weak guarantees; 'artful' parameter selection; very large search spaces



Tools

Symbiotic Bid Based GP

- One modern alg
- Evolves very simple programs
- Simple programs as instruction set for complex programs
- Can layer indefinitely.



Data

Reuters + FS Challenge Data

- Old, but 'real' data
- *Many* features, very sparse
- **Can we get a tiny and powerful model?**

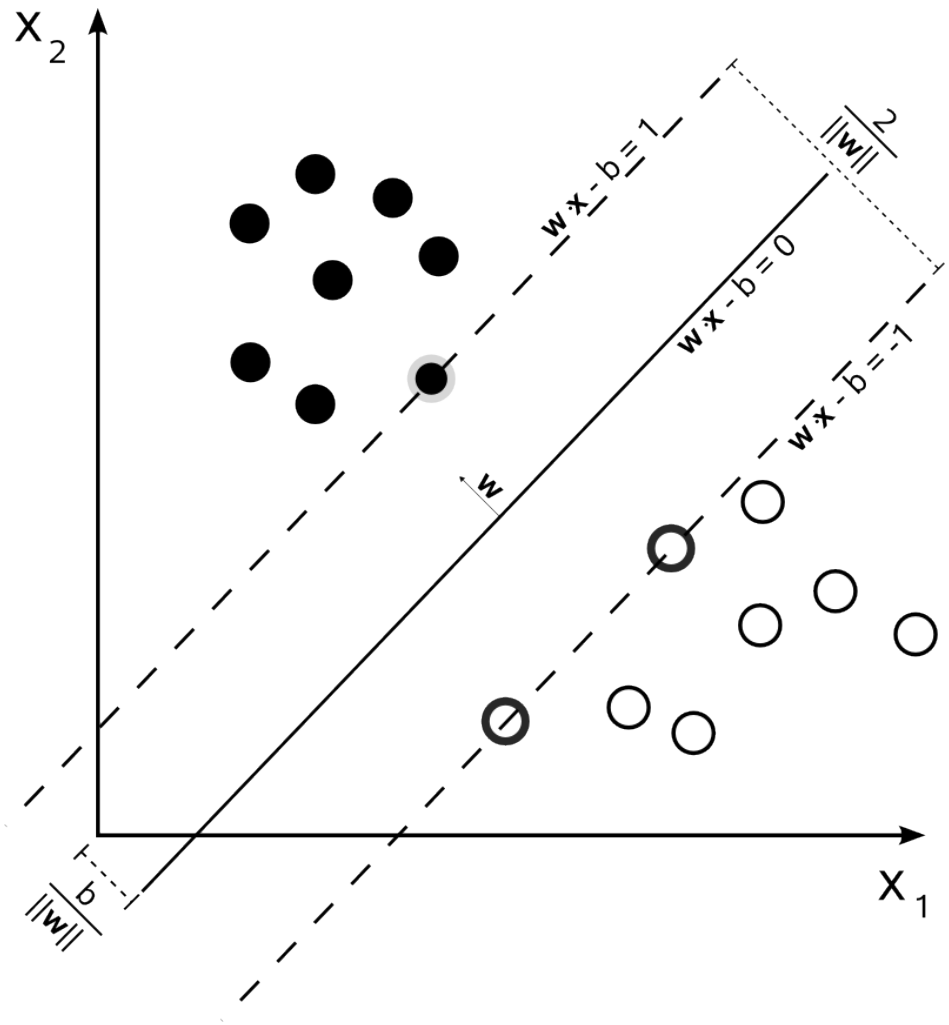
Table 1: Data set properties.

Data set	Exemplar Count train (test)	Feature Count
Handwritten character recognition		
Multifeature	1,510 (490)	649
Gisette	6,000 (1,000)	5,000
Document Classification: Bag-of-words		
NIPS	7,000 (3,500)	12,419
Enron	7,000 (3,500)	28,102
NY Times	7,000 (3,500)	102,660

Comparisons

'Out-of-the-Box' Approaches

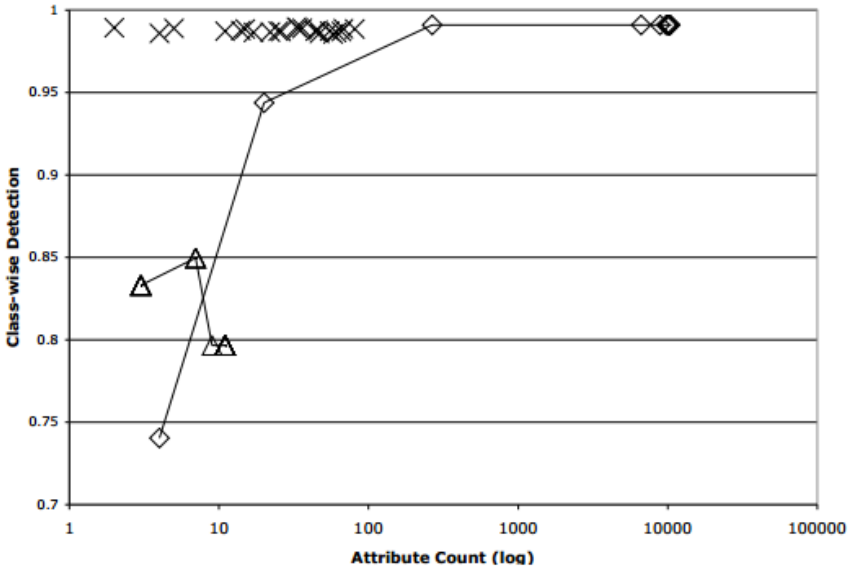
- Logres + Regularization
- F + SVM + Regularization
- Decision Tree + Confidence Factor



Results

SBB vs. Simple models

- SBB solutions are reliably simpler and/or better
- Often we can find great solutions with fewer than 10 features.

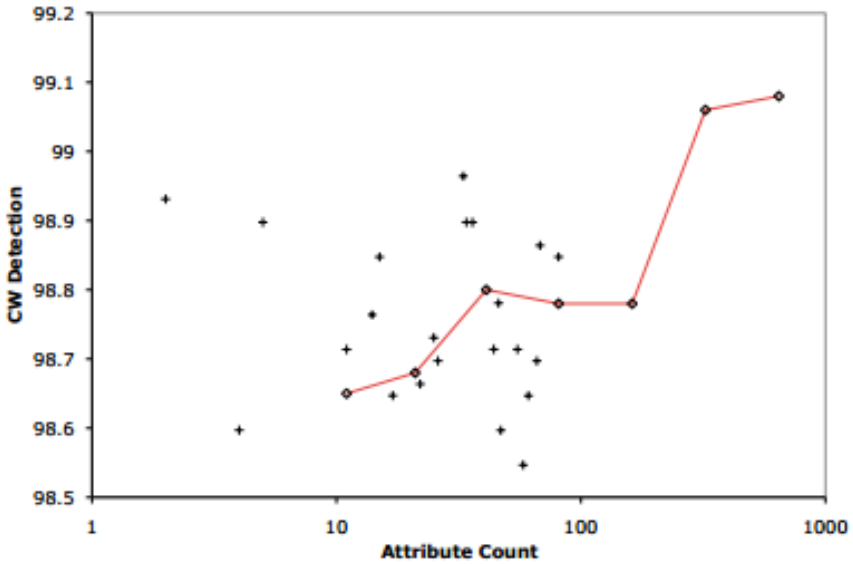


(a) NIPS

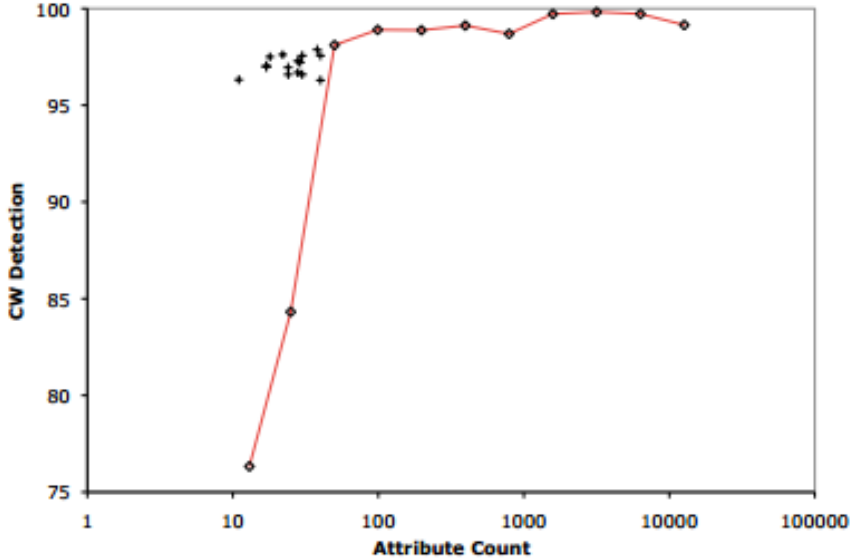
Results

SBB vs. F+SVM

- Similar Results
- But what about processing times?



(a) NIPS

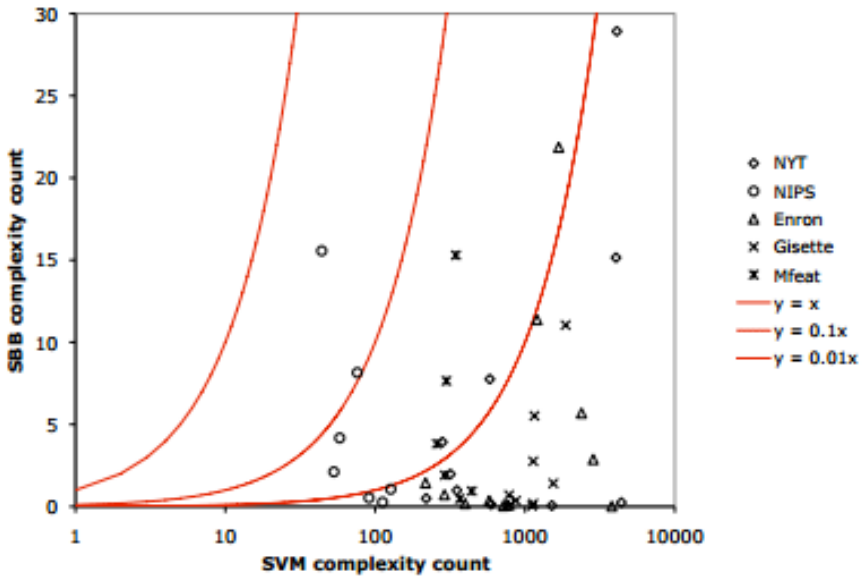


(b) Enron

Results

SBB vs. F+SVM

- SBB: Program size (# ops)
- SVM: # ops to classify new point
- Exponential advantage for SBB.



Conclusions

- Great results, great models
- Technique now used in ML for security
- Careful though: GP hard to use, not the best tool for every domain.

My Get Rich Quick Scheme

Idea

- Annual financial statements contain information.
- Formats are non-uniform, hard to process
- **Can we predict future stock performance using unprocessed reports?**

Data

Source Data

- All Form 10k's for all DJIA companies, 1992-2011
- Manually collected (never again)
- Manually labeled (never again)

UNITED STATES
SECURITIES AND EXCHANGE COMMISSION
Washington, D.C. 20549

Form 10-K

ANNUAL REPORT PURSUANT TO SECTION 13 OR 15(d) OF THE SECURITIES EXCHANGE ACT OF 1934
For the fiscal year ended September 27, 2008
or

TRANSITION REPORT PURSUANT TO SECTION 13 OR 15(d) OF THE SECURITIES EXCHANGE ACT OF 1934
For the transition period from_____ to_____
Commission file number: 000-10030

APPLE INC.
(Exact name of registrant as specified in its charter)

Compression Classifiers

Compression as Classification

- Idea: Similar data compresses well together
- Build compression codes for all data from each class
- Classify by best compression ratio
- **No preprocessing required!**

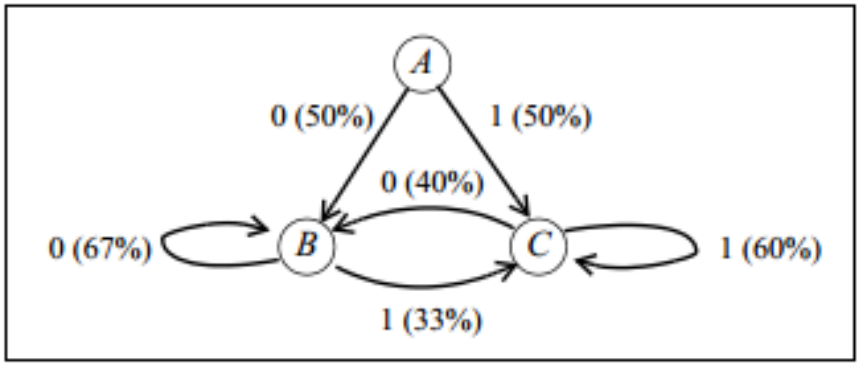


Figure 1. An example Markov model.

Table 1. Huffman coding for 3-bit groups

Source form	Probability	Huffman code
000	0.2424	10
001	0.1212	010
010	0.0727	1101
011	0.1091	000
100	0.1212	011
101	0.0606	1100
110	0.1091	001
111	0.1636	111

Cormack and Horspool 1987

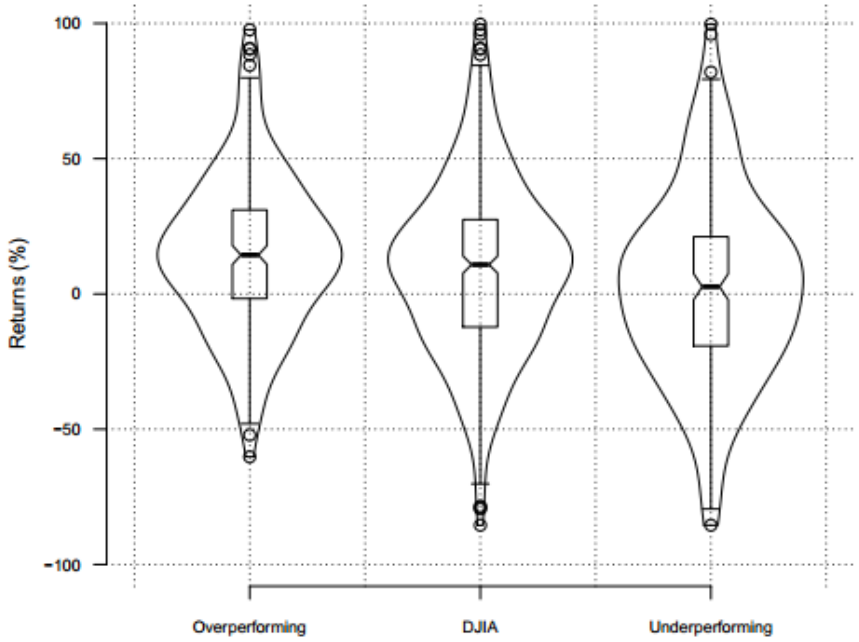
Difficulties

- Expensive training time, especially with later reports
- Many, many, many, problems with the data
- Original DMC code needed modifications

Results

Classification of Stocks

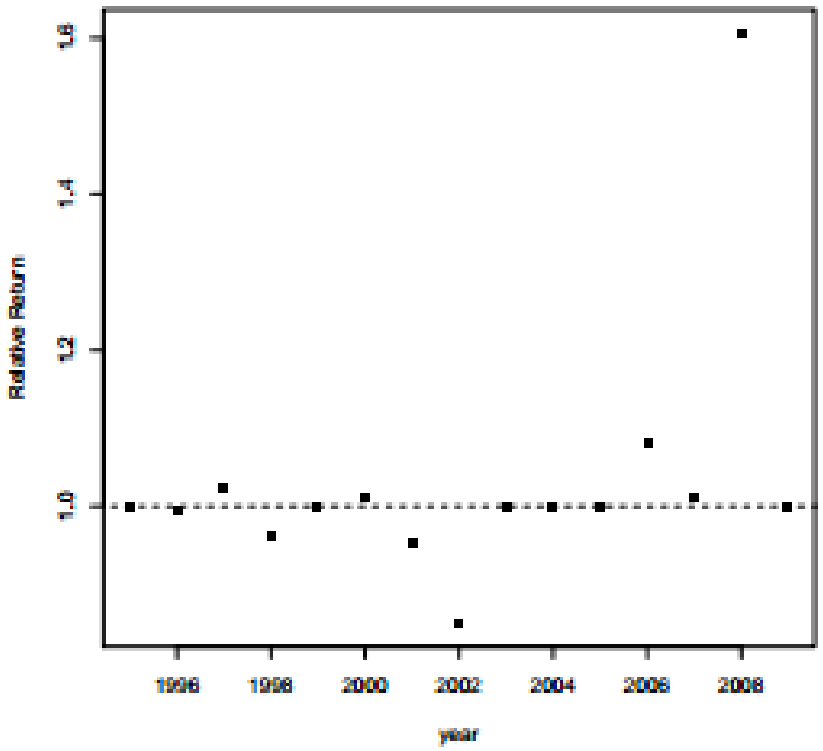
- Beats the market in expectation (4% higher returns)
- Considerable variance



Results

Prediction of Bubbles?

- Do we really need data from the last n years?
- Possibly allows crash prediction?



Conclusions

- Neat idea, surprising that it works at all
- Many possible confounding issues, needs more work
- It's finance, so it's hard to say what would happen if you used this in practice.