Interpretable Modeling in Machine Learning

Cynthia Rudin

Associate Professor

Computer Science Department

Electrical and Computer Engineering Department

Duke University

Cynthia’s Principles of Machine Learning

1) The Rashomon Effect: There is no "best" model for a finite dataset.

2) Thou shalt not make up a model using domain expertise alone.

3) People do not like to trust models that they don't understand.

4) Thou shalt not mistake “computationally fast” for “better.”

CHADS$_2$	
Risk factors	Points
CHF	1
HTN	1
Age≥ 75	1
DM	1
Stroke/TIA/embolism	2
	Max 6

Data

X: patient histories
Y: whether patient had stroke next year

Machine Learning Algorithm

Model

$CHADS_2$	
Risk factors	Points
CHF	1
HTN	1
Age≥ 75	1
DM	1
Stroke/TIA/embolism	2
	Max 6

Data

X: patient histories
Y: whether patient had stroke next year

Model

CHADS$_2$	
Risk factors	Points
CHF	1
HTN	1
Age≥ 75	1
DM	1
Stroke/TIA/embolism	2
	Max 6

TIMI Risk Score for STEMI

<u>Historical</u>

Age 65-74	**2 points**
$\geq$ 75	**3 points**
DM/HTN or angina	**1 point**

<u>Exam</u>

SBP < 100	**3 points**
HR >100	**2 points**
Killip II-IV	**2 points**
Weight < 67 kg	**1 point**

<u>Presentation</u>

Anterior STE or LBBB	**1 point**
Time to rx > 4 hrs	**1 point**

Risk Score = Total (0 -14)

(FRONT)

STATIC-99 – TALLY SHEET

Subject Name: ______________________________________

Place of Scoring: ____________________________________

Date of Scoring: __________ **Name of Assessor:** _________

Question Number	Risk Factor	Codes	Score
1	Young	Aged 25 or older Aged 18 – 24.99	0 1
2	Ever Lived With	Ever lived with lover for at least two years? Yes No	 0 1
3	Index non-sexual violence - Any Convictions?	No Yes	0 1
4	Prior non-sexual violence - Any Convictions?	No Yes	0 1
5	Prior Sex Offences	Charges / Convictions None / None 1-2 / 1 3-5 / 2-3 6 + / 4+	 0 1 2 3
6	Prior sentencing dates (excluding index)	3 or less 4 or more	0 1
7	Any convictions for non-contact sex offences	No Yes	0 1
8	Any Unrelated Victims	No Yes	0 1
9	Any Stranger Victims	No Yes	0 1
10	Any Male Victims	No Yes	0 1
	Total Score	**Add up scores from individual risk factors**	

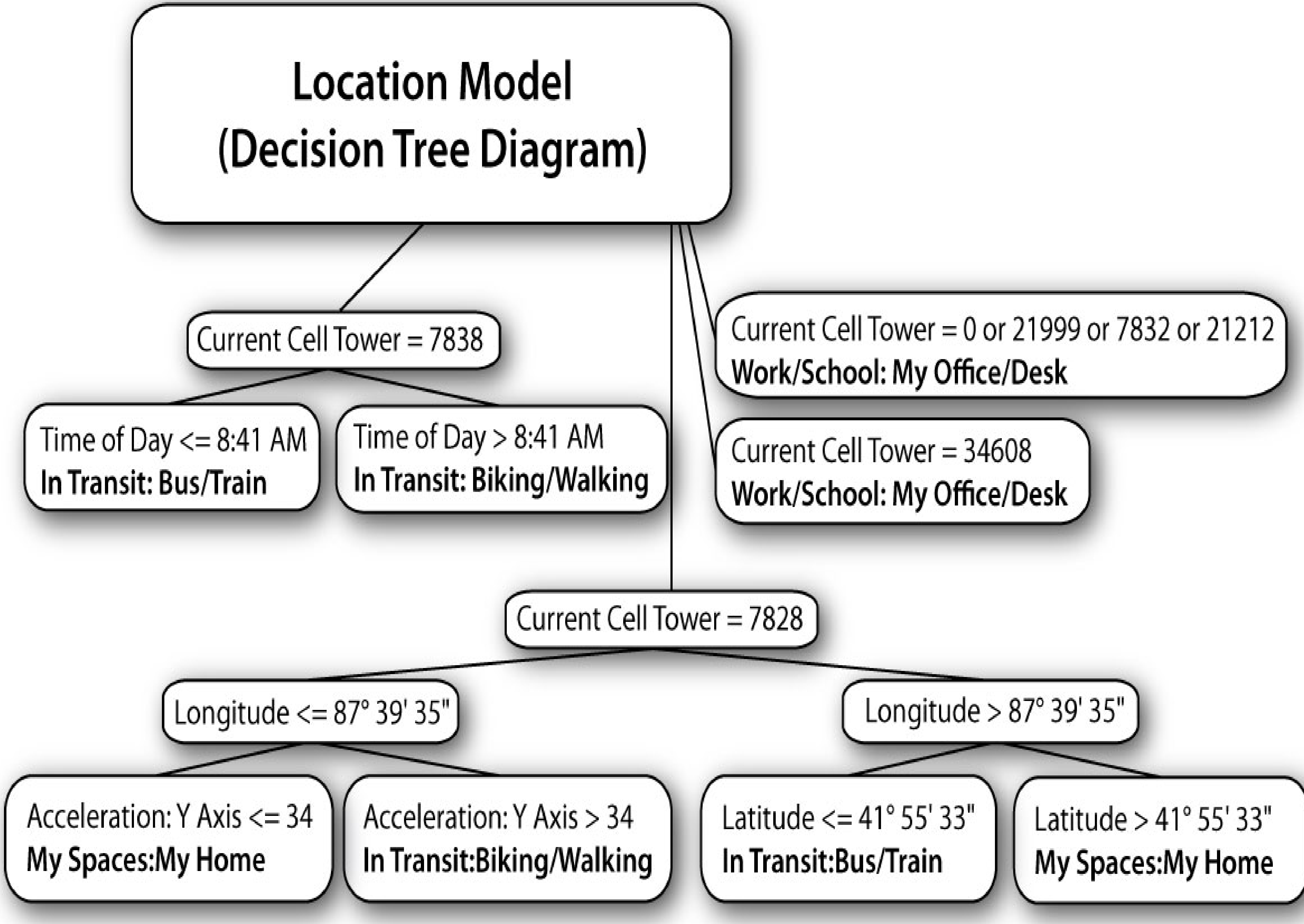
	POINTS	Risk Category
Suggested Nominal Risk Categories	**0,1**	**Low**
	2,3	**Moderate-Low**
	4,5	**Moderate-High**
	6+	**High**

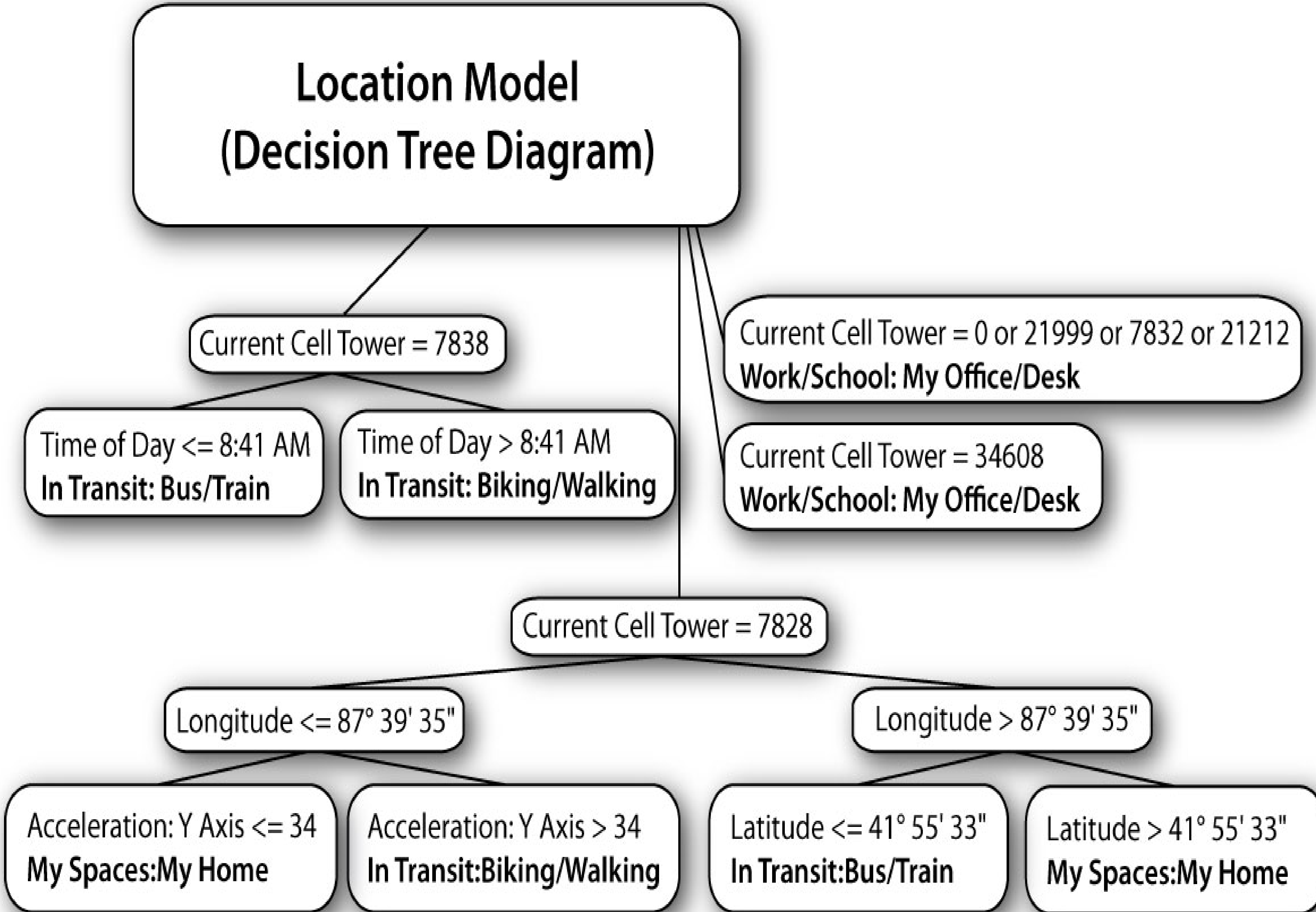
Date of Scoring: ___________ **Name of Assessor:** __________

Question Number	Risk Factor	Codes	Score
1	Young	Aged 25 or older Aged 18 – 24.99	0 1
2	Ever Lived With	Ever lived with lover for at least two years? Yes No	 0 1
3	Index non-sexual violence - Any Convictions?	No Yes	0 1
4	Prior non-sexual violence - Any Convictions?	No Yes	0 1
5	Prior Sex Offences	Charges / Convictions None / None 1-2 / 1 3-5 / 2-3 6 + / 4+	 0 1 2 3
6	Prior sentencing dates (excluding index)	3 or less 4 or more	0 1
7	Any convictions for non-contact sex offences	No Yes	0 1
8	Any Unrelated Victims	No Yes	0 1
9	Any Stranger Victims	No Yes	0 1

		1-2 1 3-5 2-3 6 + 4+	1 2 3
6	Prior sentencing dates (excluding index)	3 or less 4 or more	0 1
7	Any convictions for non-contact sex offences	No Yes	0 1
8	Any Unrelated Victims	No Yes	0 1
9	Any Stranger Victims	No Yes	0 1
10	Any Male Victims	No Yes	0 1
	Total Score	**Add up scores from individual risk factors**	

	POINTS	**Risk Category**
Suggested Nominal Risk Categories	**0,1**	**Low**
	2,3	**Moderate-Low**
	4,5	**Moderate-High**
	6+	**High**


Location Model
(Decision Tree Diagram)
Current Cell Tower = 7838
Time of Day <= 8:41 AM
In Transit: Bus/Train
Time of Day > 8:41 AM
In Transit: Biking/Walking
Current Cell Tower = 0 or 21999 or 7832 or 21212
Work/School: My Office/Desk
Current Cell Tower = 34608
Work/School: My Office/Desk
Current Cell Tower = 7828
Longitude <= 87° 39' 35"
Longitude > 87° 39' 35"
Acceleration: Y Axis <= 34
My Spaces:My Home
Acceleration: Y Axis > 34
In Transit:Biking/Walking
Latitude <= 41° 55' 33"
In Transit:Bus/Train
Latitude > 41° 55' 33"
My Spaces:My Home


Location Model
(Decision Tree Diagram)
Current Cell Tower = 7838
Time of Day <= 8:41 AM
In Transit: Bus/Train
Time of Day > 8:41 AM
In Transit: Biking/Walking
Current Cell Tower = 0 or 21999 or 7832 or 21212
Work/School: My Office/Desk
Current Cell Tower = 34608
Work/School: My Office/Desk
Current Cell Tower = 7828
Longitude <= 87° 39' 35"
Longitude > 87° 39' 35"
Acceleration: Y Axis <= 34
My Spaces:My Home
Acceleration: Y Axis > 34
In Transit:Biking/Walking
Latitude <= 41° 55' 33"
In Transit:Bus/Train
Latitude > 41° 55' 33"
My Spaces:My Home

Decision Trees

- Why trees?

- CART (Breiman 1993) is arguably the most widely used predictive modeling method used in industry currently.

Decision Trees

- Example: Will the customer wait for a table at a restaurant?
  - OthOptions: Other options, True if there are restaurants nearby.
  - Weekend: This is true if it is Friday, Saturday or Sunday.
  - Area: Does it have a bar or other nice waiting area to wait in?
  - Plans: Does the customer have plans just after dinner?
  - Price: This is either $, $$, $$$, or $$$$
  - Precip: Is it raining or snowing?
  - Genre: French, Mexican, Thai, or Pizza
  - Wait: Wait time estimate: 0-5 min, 6-15 min, 16-30 min, or 30+
  - Crowded: Whether there are other customers (no, some, or full)

Credit: Adapted from Russell and Norvig

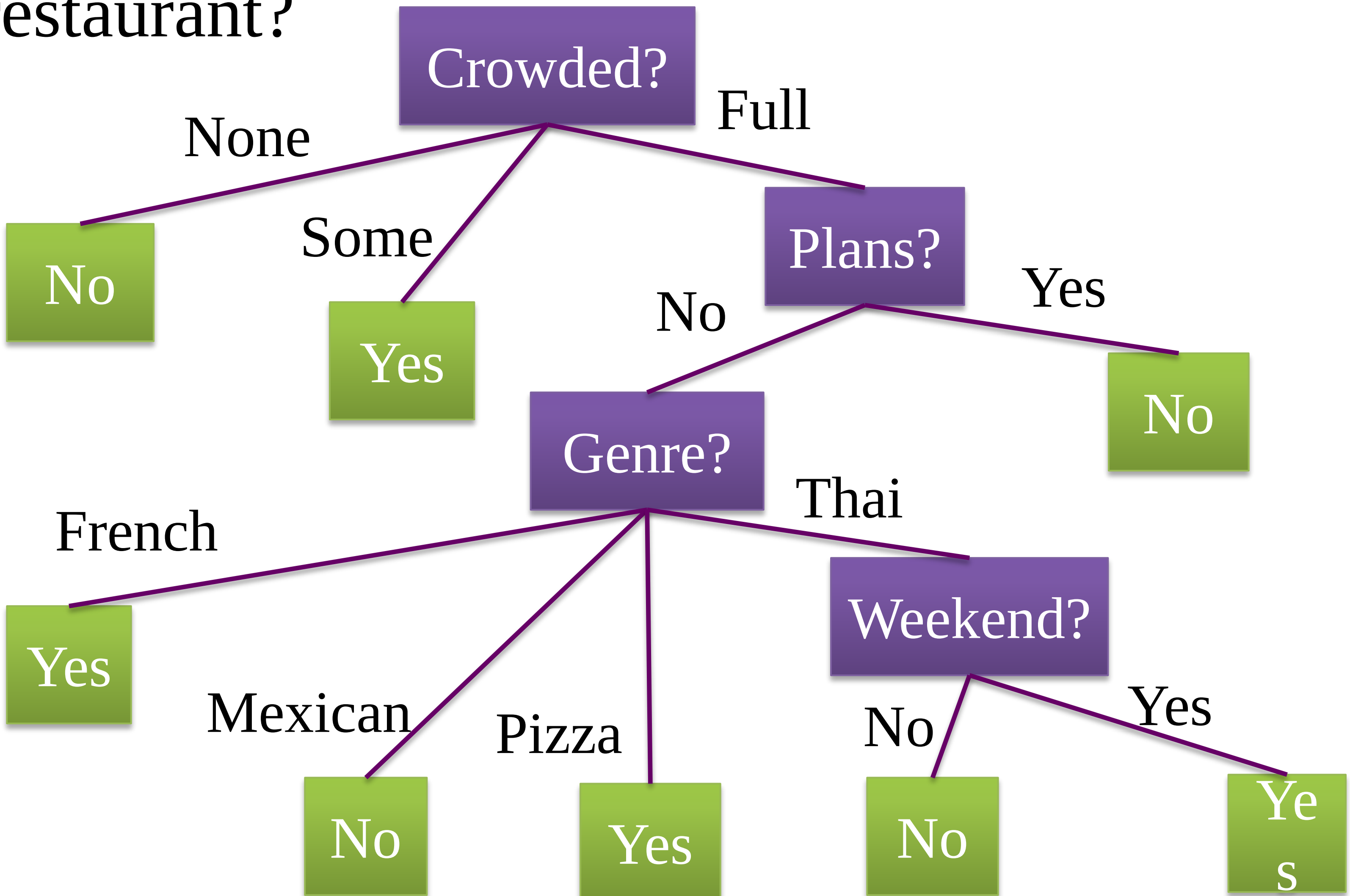
Decision Trees

Example: Will the customer wait for a table at a restaurant?

	OthOptions	Weekend	Area	Plans	Price	Precip	Genre	Wait	Crowded	Stay?
x_1	Yes	No	No	Yes	$$$	No	French	0-5	some	Yes
x_2	Yes	No	No	Yes	$	No	Thai	16-30	full	No
x_3	No	No	Yes	No	$	No	Pizza	0-5	some	Yes
x_4	Yes	Yes	No	Yes	$	No	Thai	6-15	full	Yes
x_5	Yes	Yes	No	No	$$$	No	French	30+	full	No
x_6	No	No	Yes	Yes	$$	Yes	Mexican	0-5	some	Yes
x_7	No	No	Yes	No	$	Yes	Pizza	0-5	none	No
x_8	No	No	No	Yes	$$	Yes	Thai	0-5	some	Yes
x_9	No	Yes	Yes	No	$	Yes	Pizza	30+	full	No
x_{10}	Yes	Yes	Yes	Yes	$$$	No	Mexican	6-15	full	No
x_{11}	No	No	No	No	$	No	Thai	0-5	none	No
x_{12}	Yes	Yes	Yes	Yes	$	No	Pizza	16-30	full	Yes

Decision Trees

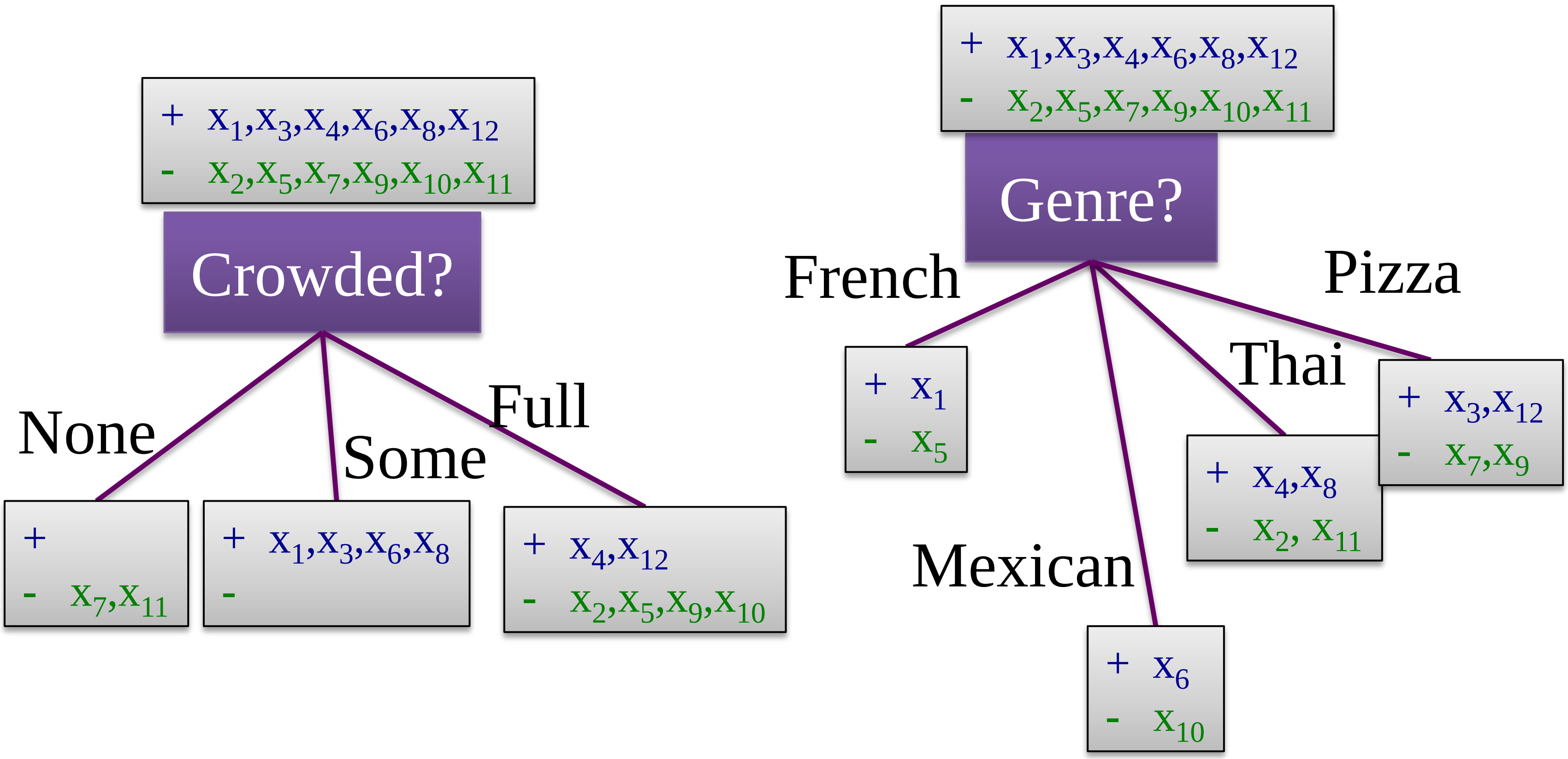
- Example: Will the customer wait for a table at a restaurant?



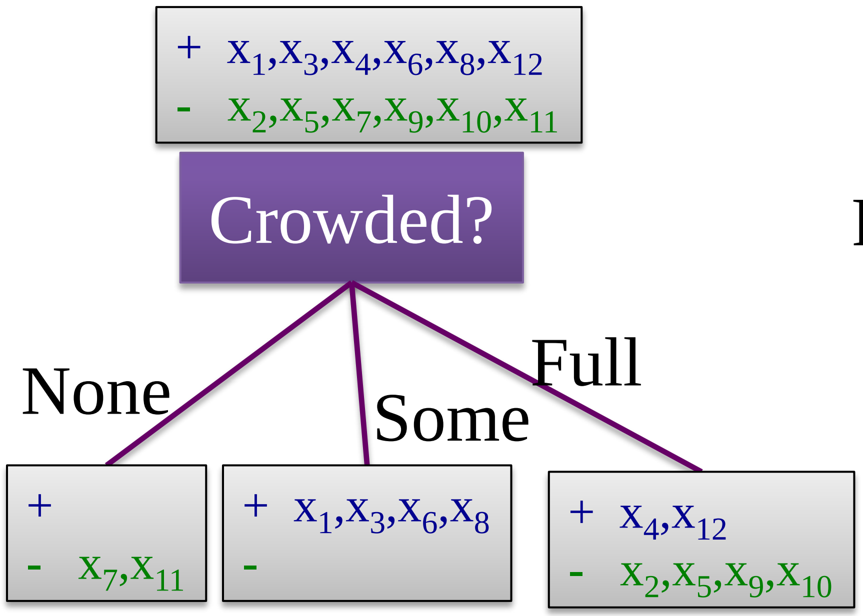
Standard Way to Build a Decision Tree

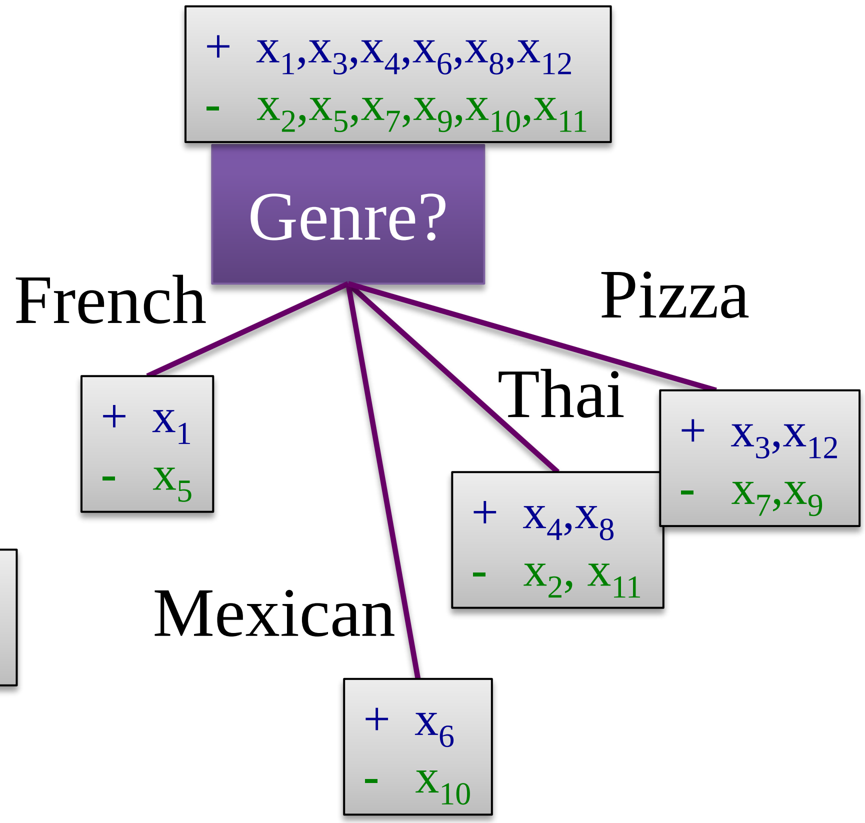
- Start at the top of the tree.
- Grow it by “splitting” features one by one. To split, look at how “impure” the node is.
- Assign leaf nodes the majority vote in the leaf.

- Which of these two features should we split on?



- Which of these two features should we split on?

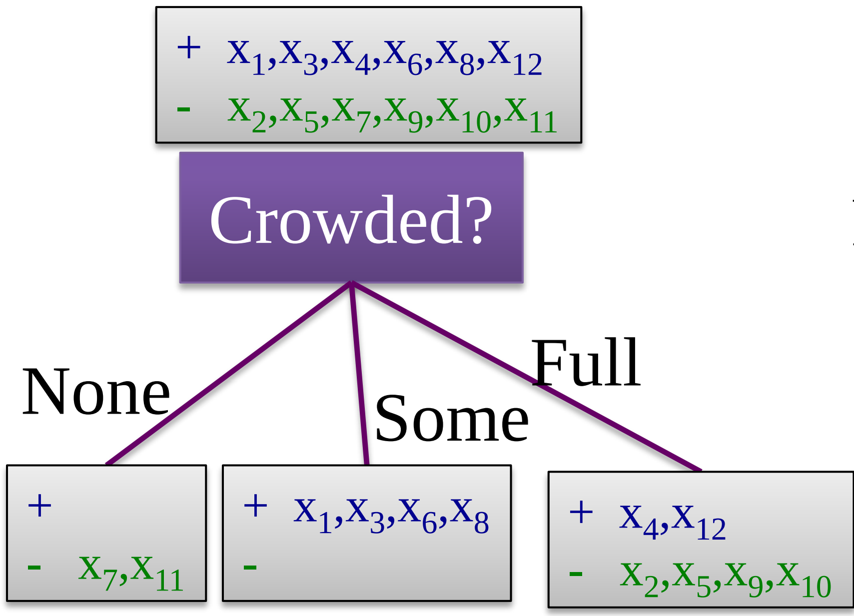


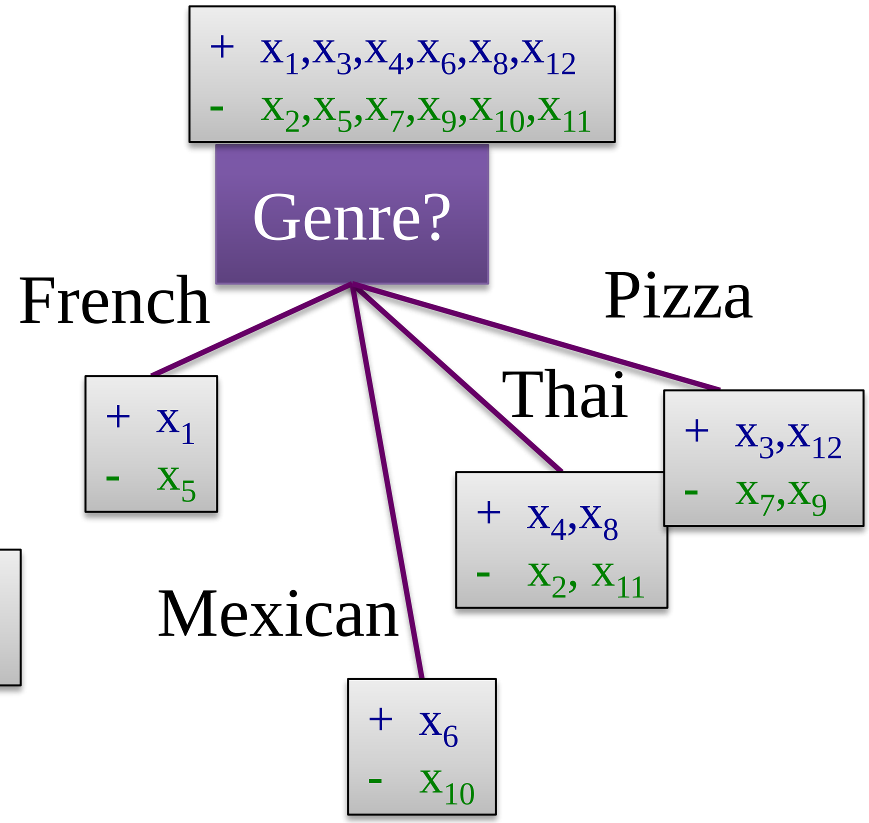


[.08,.99] is good

[.50,.50] is bad

- Which of these two features should we split on?





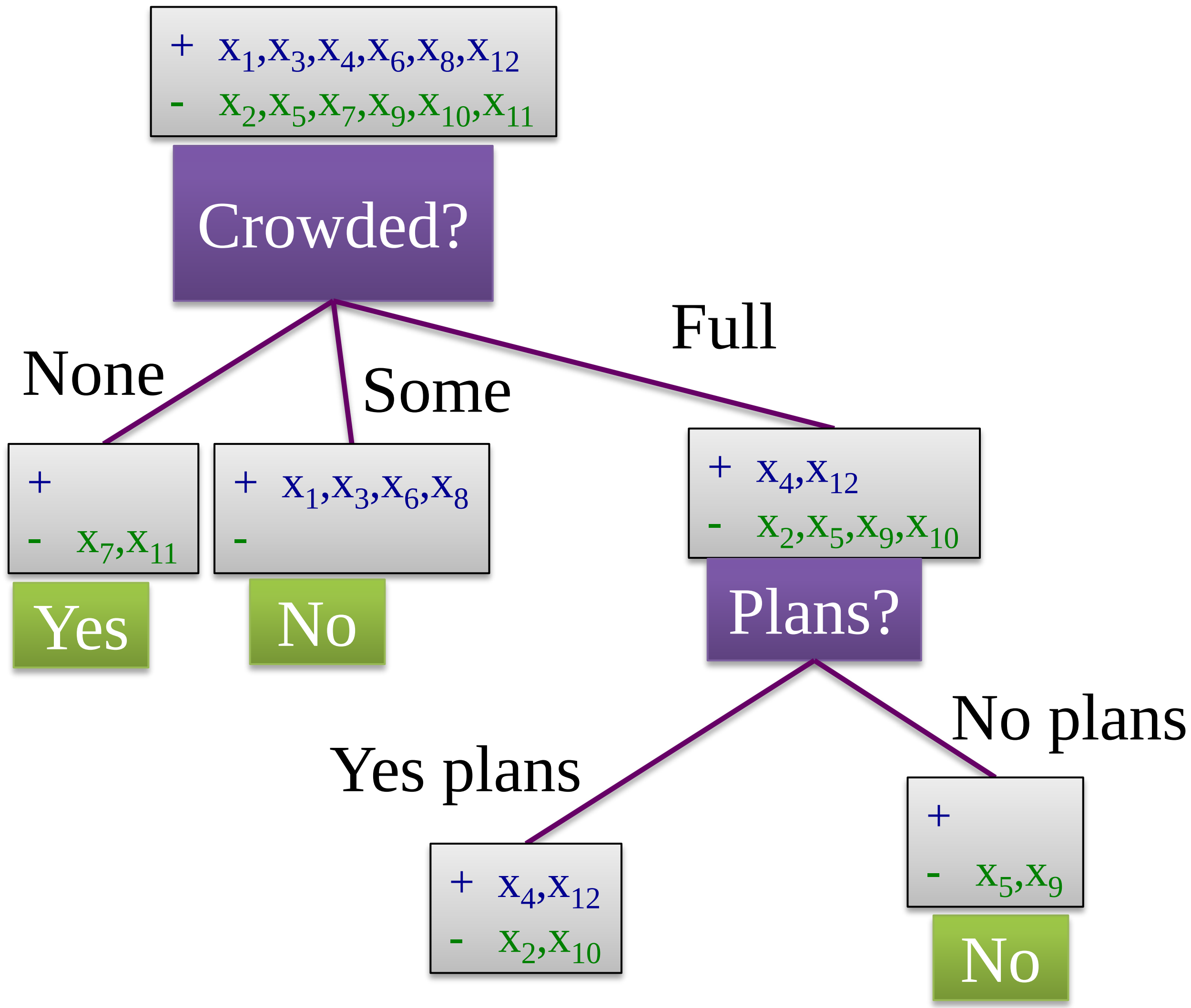
[.08,.99] is good

low entropy is good

[.50,.50] is bad

high entropy is bad

- Next we'll split on Plans



Standard Way to Build a Decision Tree

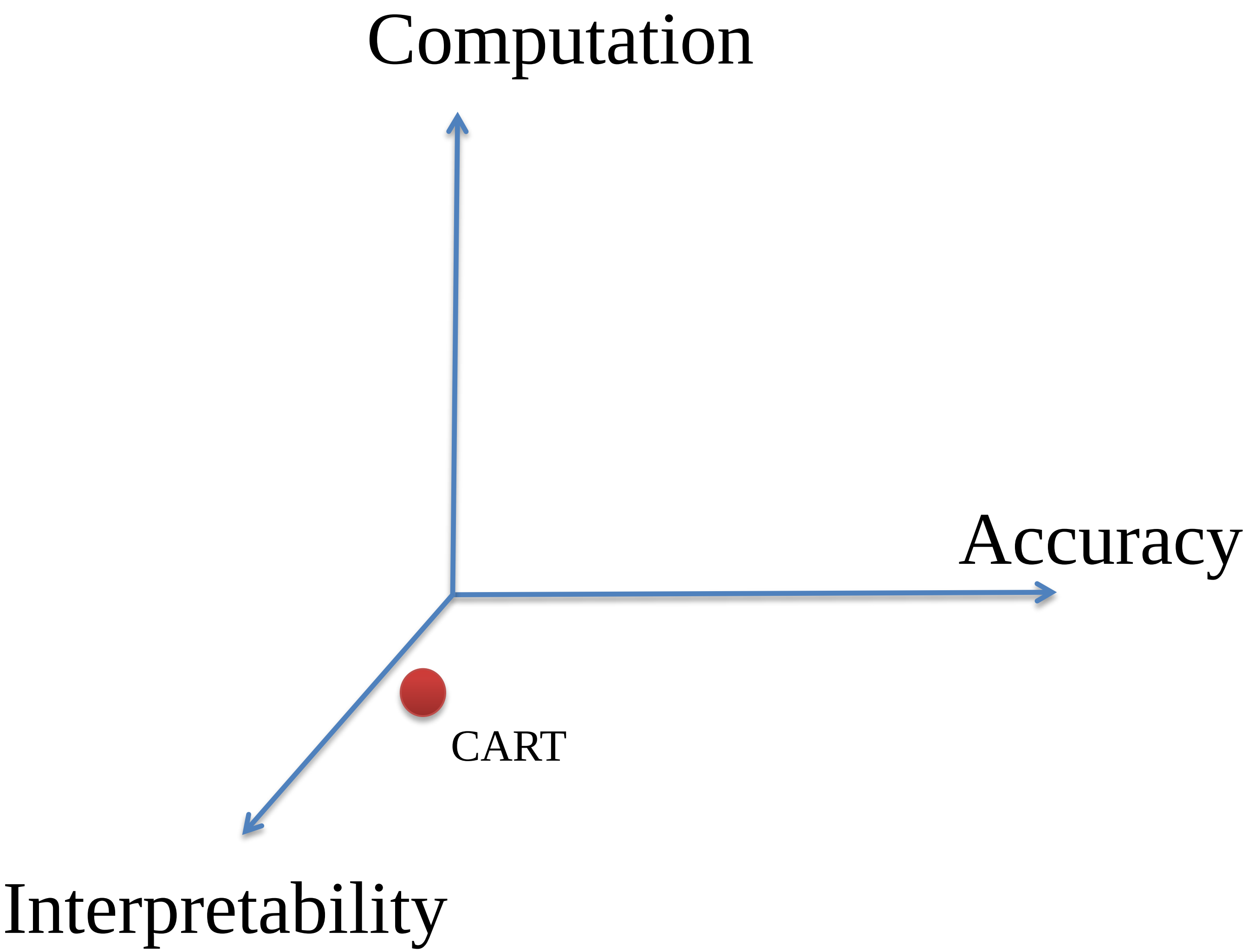
- Start at the top of the tree.
- Grow it by "splitting" features one by one. To split, look at how "impure" the node is.
- Assign leaf nodes the majority vote in the leaf.

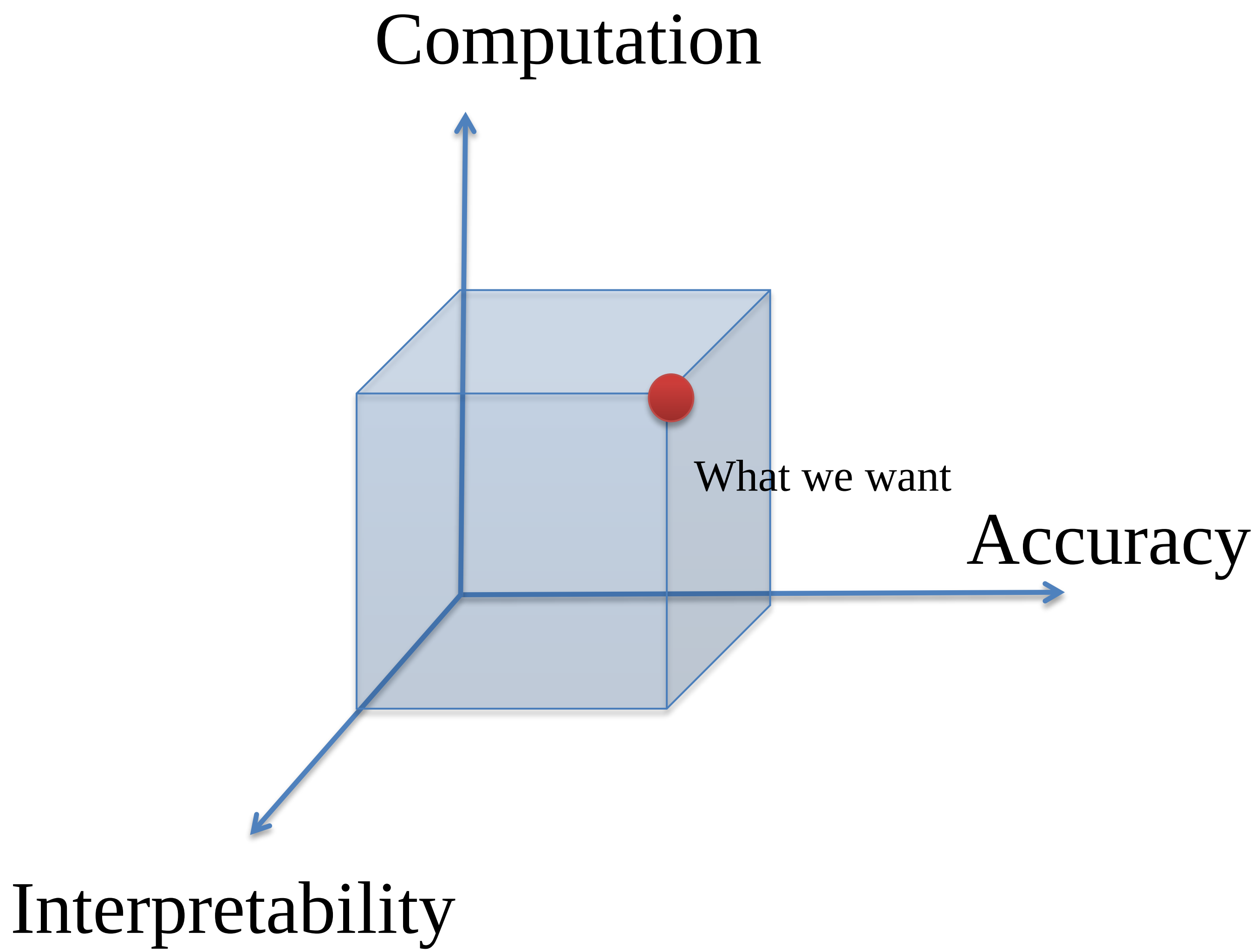
- At the end, go back and prune leaves to reduce overfitting.

Why is this a good way to build a tree?

- Bottom line: Decision trees don’t optimize anything

- No wonder there’s a tradeoff between accuracy and interpretability in predictive modeling…


Computation
Accuracy
CART
Interpretability


Computation
What we want
Accuracy
Interpretability

CART

Not a CART

Scalable Bayesian Rule Lists

- accurate
- principled
- interpretable
- scalable

Step 1 of Scalable Bayesian Rule Lists: Mine Frequent Patterns

If age<50 and aspirin ---> no stroke

If past stroke and warfarin ---> stroke

If age<40 ---> no stroke

If diabetes and hypertension ---> stroke

Step 2 of Scalable Bayesian Rule Lists: Choose and Assemble Patterns into a Decision List

Example coming…

An Example of a Model for Predicting Customer Churn (IBM Watson Telco Customer Churn Data)

Input: data about each customer, and whether they churned

Output: predictive model on the next slide

An Example of a Model for Predicting Customer Churn (IBM Watson Telco Customer Churn Data)

if Contract= 1 Year & StreamingMovies=Yes) -> P(churn) = 0.20

else if (Contract= 1 Year) -> P(churn) = 0.05

else if (Tenure<1 year & InternetService=FiberOptic) -> P (churn) = 0.70

else if (Contract=2 year), -> P(churn) = 0.03

else if (InternetService=FiberOptic & OnlineSecurity=No) -> P(churn) = 0.48

else if (OnlineBackup=No & TechSupport=No), -> P(churn) = 0.41

else -> P(churn) = 0.22

Stroke Prediction Model (work in progress)

Input: data about each patient, and whether they had a stroke later

Output: predictive model

IF past history of stroke	THEN P(stroke) = 40.5%
ELSE IF patient takes warfarin reliably	THEN P(stroke) = 5.6%
ELSE IF age<70	THEN P(stroke) = 7.2%
ELSE IF Blood Pressure>110	THEN P(stroke) = 45%
ELSE IF age<75	THEN P(stroke) = 6.8%
ELSE IF high BMI	THEN P(stroke) = 18.4%
ELSE	P(stroke) = 7.2%

Step 2 of Scalable Bayesian Rule Lists: Choose and Assemble Patterns into a Decision List

Solves a special optimization problem over decision lists.

maximize$_{\text{models}}$ Posterior(model):

pile of rules   size of list

$$\text{Posterior}(\{c_j\}_j, m, \{N_{jl}\}_{jl}) = \left(\prod_{j=0}^{m} \frac{\prod_{l=1}^{L} \Gamma(N_{jl} + \alpha_l)}{\Gamma\left(\sum_{l=1}^{L} N_{jl} + \alpha_l \right)} \right) \frac{\left(\lambda^m / m! \right)}{\sum_{j=0}^{|A|} \left(\lambda^j / j! \right)} \prod_{j=1}^{m} \frac{\left(\eta^{c_j} / c_j! \right)}{\sum_{k \in R_{j-1}(c_{<j})} \left(\eta^k / k! \right)},$$

What do you want in a model anyway?

- Accuracy
  - Data should look like it could have been generated by the model
- Gorgeousness
  - Sparsity, Your own beliefs about the truth

$$P(\text{data} \mid \text{model})$$

Likelihood of data to come from model

$$P(\text{model}) \times P(\text{data} \mid \text{model})$$

Prior
of model

Likelihood of data
to come from
model

$$P(\text{model} \mid \text{data}) \propto P(\text{model}) \times P(\text{data} \mid \text{model})$$

Posterior
of model

Prior
of model

Likelihood of data
to come from
model

Step 2 of Scalable Bayesian Rule Lists: Choose and Assemble Patterns into a Decision List

Solves a special optimization problem over decision lists.

$$P(\text{model} \mid \text{data}) \propto P(\text{model}) \times P(\text{data} \mid \text{model})$$

Posterior of model

Prior of model

Likelihood of data to come from model

Step 2 of Scalable Bayesian Rule Lists: Choose and Assemble Patterns into a Decision List

Solves a special optimization problem over decision lists.

maximize$_{\text{models}}$ Posterior(model):

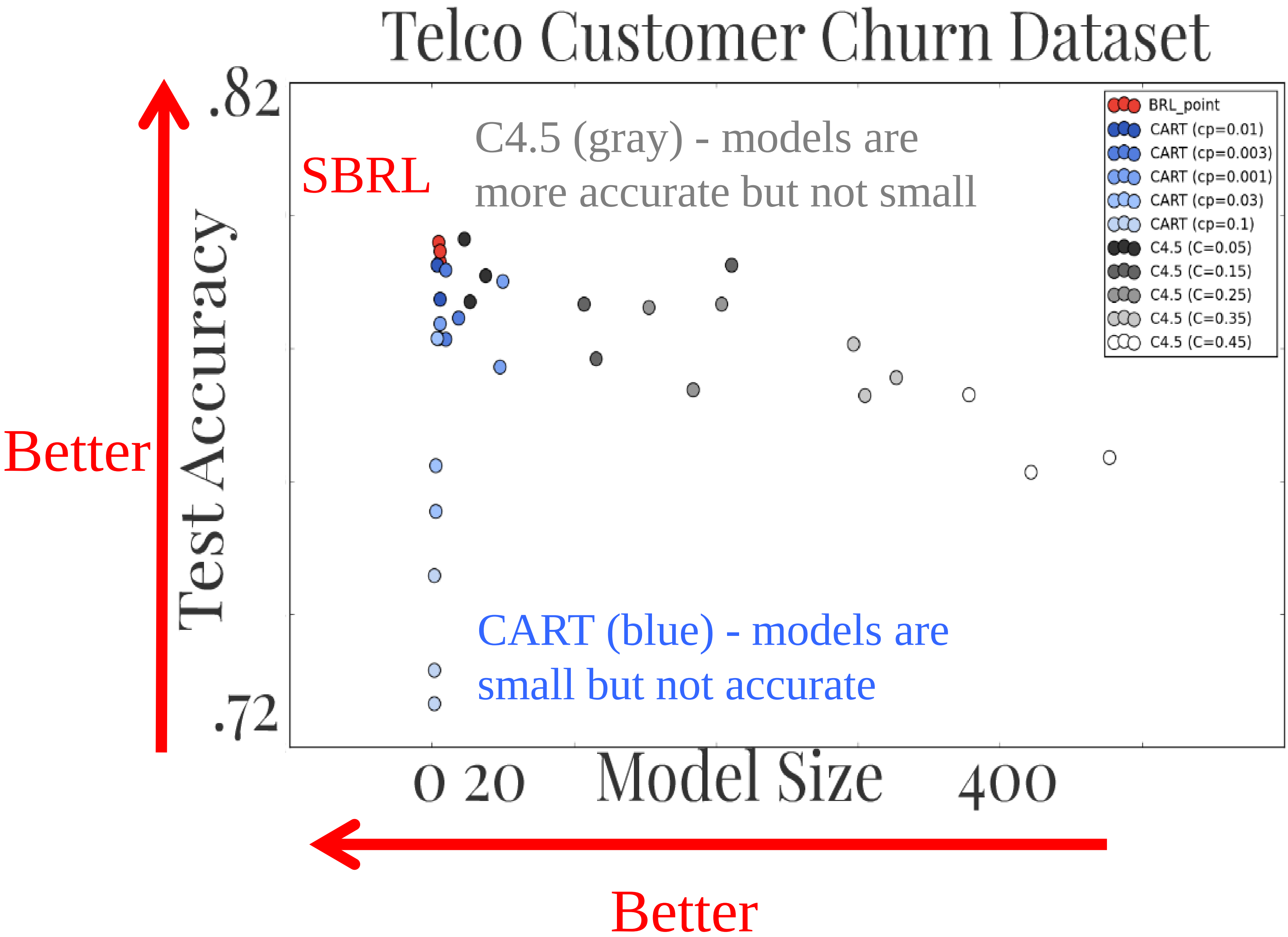
pile of rules size of list

$$\text{Posterior}(\{c_j\}_j, m, \{N_{jl}\}_{jl}) = \left(\prod_{j=0}^{m} \frac{\prod_{l=1}^{L} \Gamma(N_{jl}+\alpha_l)}{\Gamma\left(\sum_{l=1}^{L} N_{jl}+\alpha_l\right)} \right) \frac{\left(\lambda^m / m!\right)}{\sum_{j=0}^{|A|} \left(\lambda^j / j!\right)} \prod_{j=1}^{m} \frac{\left(\eta^{c_j} / c_j!\right)}{\sum_{k \in R_{j-1}(c_{<A})} \left(\eta^k / k!\right)},$$

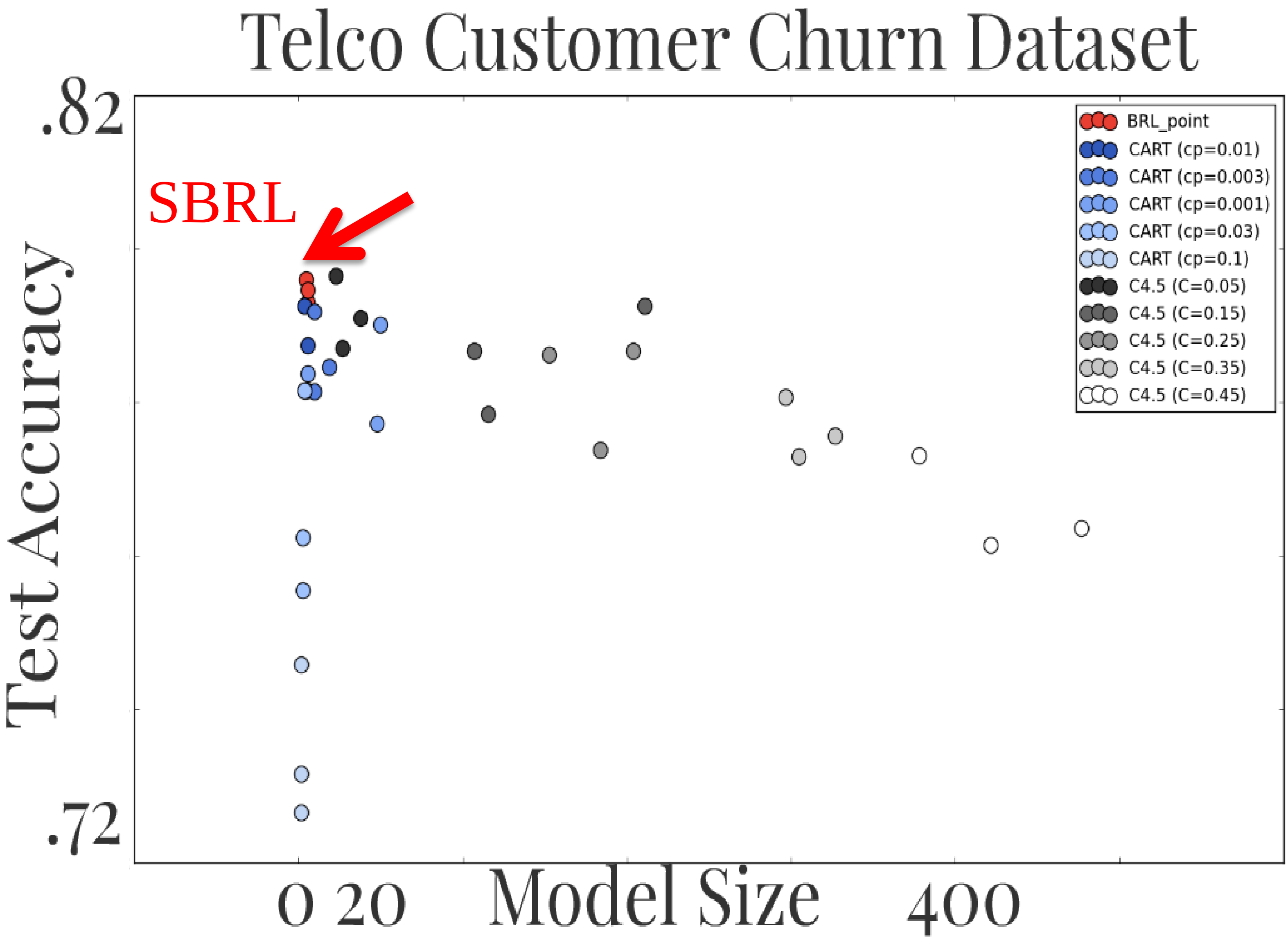
counts for label l, rule j

Other SBRL ingredients

- Very fast bit-vector manipulation. Computational reuse. Pre-processing expensive computation.
- Theoretical bounds that are used as optimization cuts.



Telco Customer Churn Dataset
.82
.72
Test Accuracy
Model Size
0 20
400
Better
Better
SBRL
C4.5 (gray) - models are
more accurate but not small
CART (blue) - models are
small but not accurate
BRL_point
CART (cp=0.01)
CART (cp=0.003)
CART (cp=0.001)
CART (cp=0.03)
CART (cp=0.1)
C4.5 (C=0.05)
C4.5 (C=0.15)
C4.5 (C=0.25)
C4.5 (C=0.35)
C4.5 (C=0.45)



Telco Customer Churn Dataset
.82
.72
Test Accuracy
0 20
Model Size
400
SBRL
BRL_point
CART (cp=0.01)
CART (cp=0.003)
CART (cp=0.001)
CART (cp=0.03)
CART (cp=0.1)
C4.5 (C=0.05)
C4.5 (C=0.15)
C4.5 (C=0.25)
C4.5 (C=0.35)
C4.5 (C=0.45)

An Example of a Decision List for Predicting Customer Churn (IBM Watson Telco Customer Churn Data)

Model from Fold 1

if Contract= 1 Year & StreamingMovies=Yes) -> P(churn) = 0.20

else if (Contract= 1 Year) -> P(churn) = 0.05

else if (Tenure<1 year & InternetService=FiberOptic) -> P (churn) = 0.70

else if (Contract=2 year), -> P(churn) = 0.03

else if (InternetService=FiberOptic & OnlineSecurity=No) -> P(churn) = 0.48

else if (OnlineBackup=No & TechSupport=No), -> P(churn) = 0.41

else -> P(churn) = 0.22

Model from Fold 2

if (Contract=One_year & StreamingMovies=Yes) -> P(churn) = 0.20

else if (Tenure<1 year & InternetService=Fiber_optic) -> P(churn) = 0.70

else if (Tenure<1year&OnlineBackup=No) -> P(churn) = 0.44

else if (InternetService=Fiber_optic&Contract=Month-to-month) -> P(churn) = 0.43

else if (Contract=Month-to-month) -> P(churn) = 0.22

else -> P(churn) = 0.03

Model from Fold 3

if (Contract=One_year & StreamingMovies=Yes) -> P(churn) = 0.25

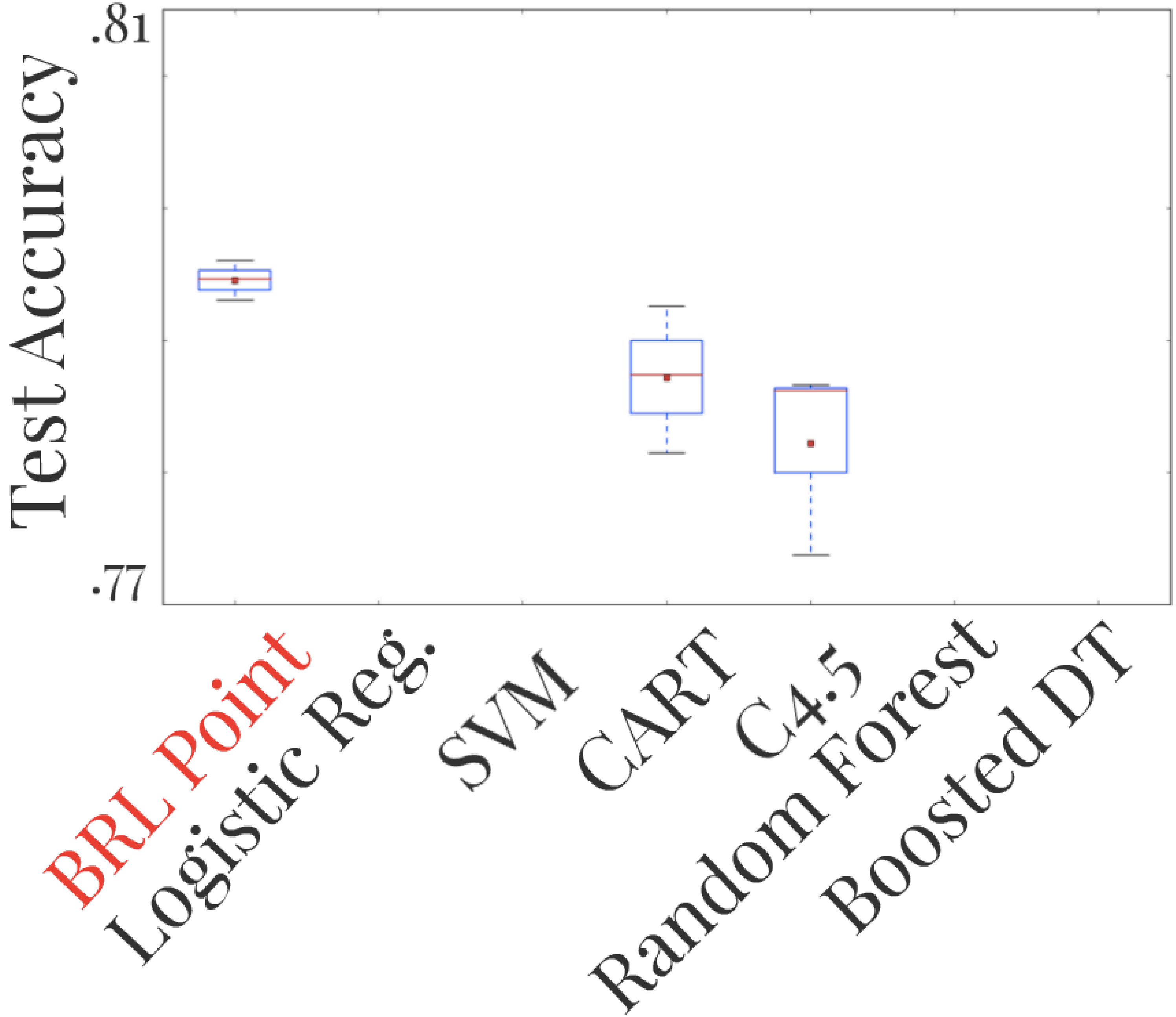
else if (Contract=One Year) -> P(churn) = 0.03

else if (Contract=Two Year) -> P(churn) = 0.05

else if (Tenure<1year & InternetService=Fiber_optic) -> P(churn) = 0.69

else if (TechSupport=No&OnlineSecurity=No) -> P(churn) = 0.45

else -> P(churn) = 0.25


Test Accuracy
.81
.77
BRL Point
Logistic Reg.
SVM
CART
C4.5
Random Forest
Boosted DT

Runtime in seconds

	BRL	LR	SVM	CART	C4.5	RF	ADA
fold 1	0.81	0.30	2.45	0.12	0.42	2.70	6.20
fold 2	0.85	0.19	2.28	0.12	0.22	2.56	6.07
fold 3	0.87	0.18	2.34	0.11	0.23	2.60	6.05

Limits: 1 million data points, hundreds of rules: 2700 sec
50K data points, 50K rules: 9000 sec

Cynthia's Principles of Machine Learning

1) The Rashomon Effect: There is no "best" model for a finite dataset.

2) Thou shalt not make up a model using domain expertise alone.

3) People do not like to trust models that they don't understand.

4) Thou shalt not mistake "computationally fast" for "better."

5) If you want both accuracy and interpretability….

optimize for them

Implications in

- Energy (equipment failure prediction)
- Healthcare (scoring systems, diagnose/predict)
- Criminology (who will commit a crime?)
- Marketing (understanding customer preferences)
- :

*Code for SBRL is publicly available on CRAN and on my website. Creative Commons License.

Thanks

*code for SBRL is publicly available on CRAN and on my website