BROAD INSTITUTE

Weight of Evidence Inference and Bayesian Nomograms Using Genomic Feature Pairs

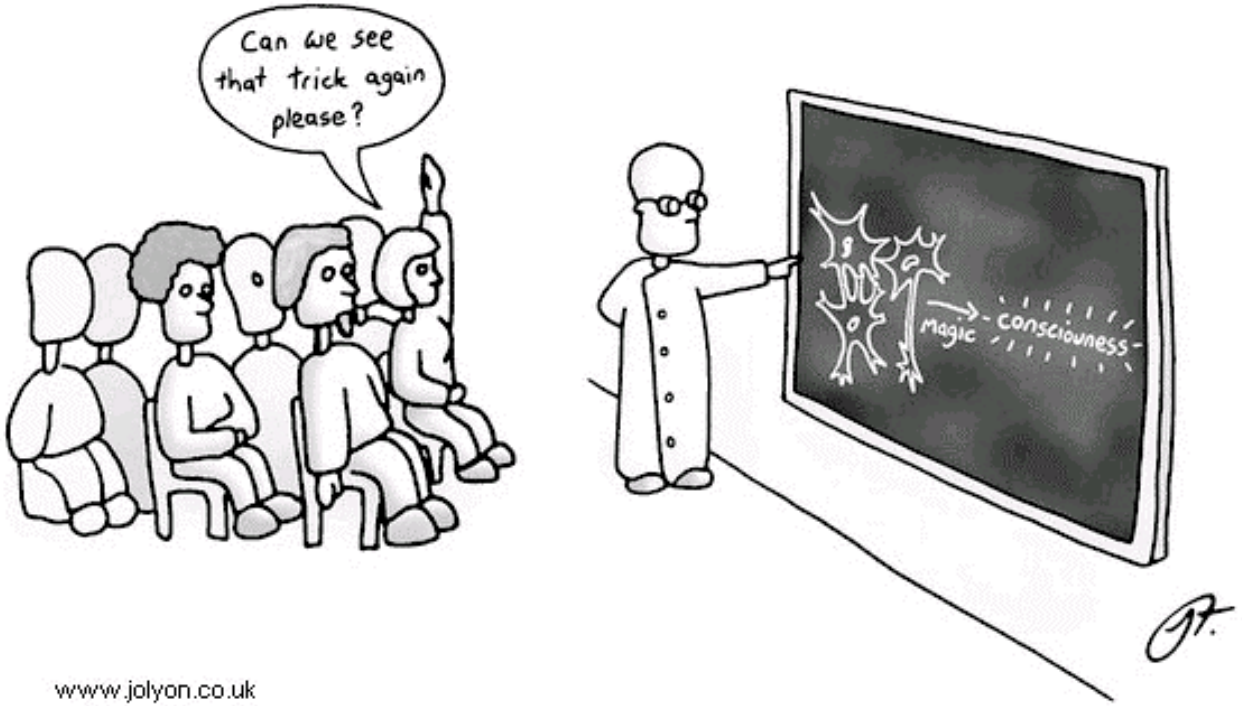
Sebastian Gomez
Mentor: Dr. Pablo Tamayo

BROAD INSTITUTE

Why do we care?

•Despite significant progress in recent years clinical prediction and stratification of risk in patients remains a challenge

•The focus has shifted from clinical parameters to molecular markers. e.g. expression of specific genes and selected genomic abnormalities

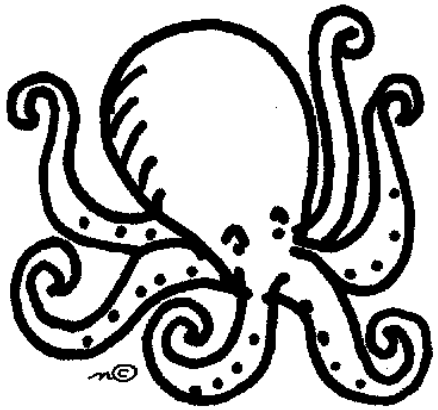


Purpose of the research

•Develop a way of predicting a particular outcome based on several input features (risk factors)

Examples:

•Clinical Prognosis: Predicting the probability of responding to treatment in different cancer types



+ OUR MODEL =

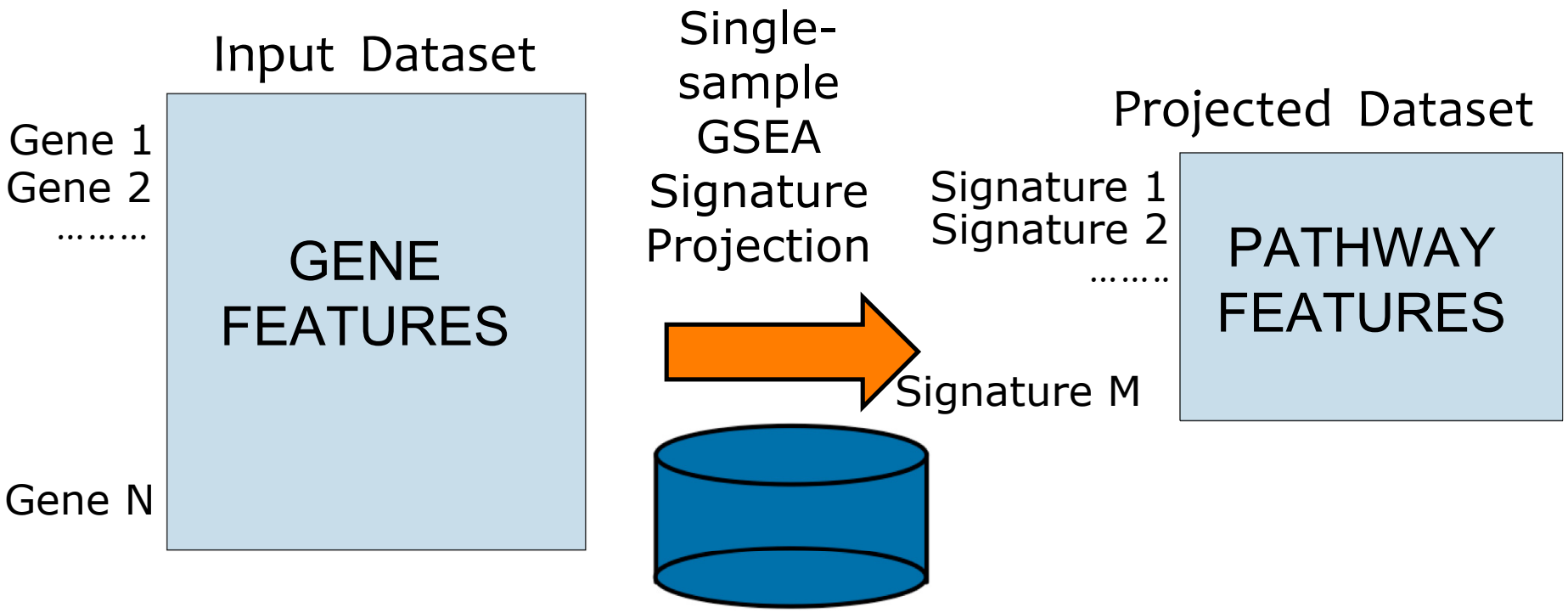
Our predictions are based on evidence provided by features

- Over-expression of gene or pathway
- Genomic abnormality: amplification
- Genomic abnormality: deletion

Types of outcomes we want to predict

- Response to treatment (e.g. chemo)
- Platinum Status (platinum sensitive or resistant)

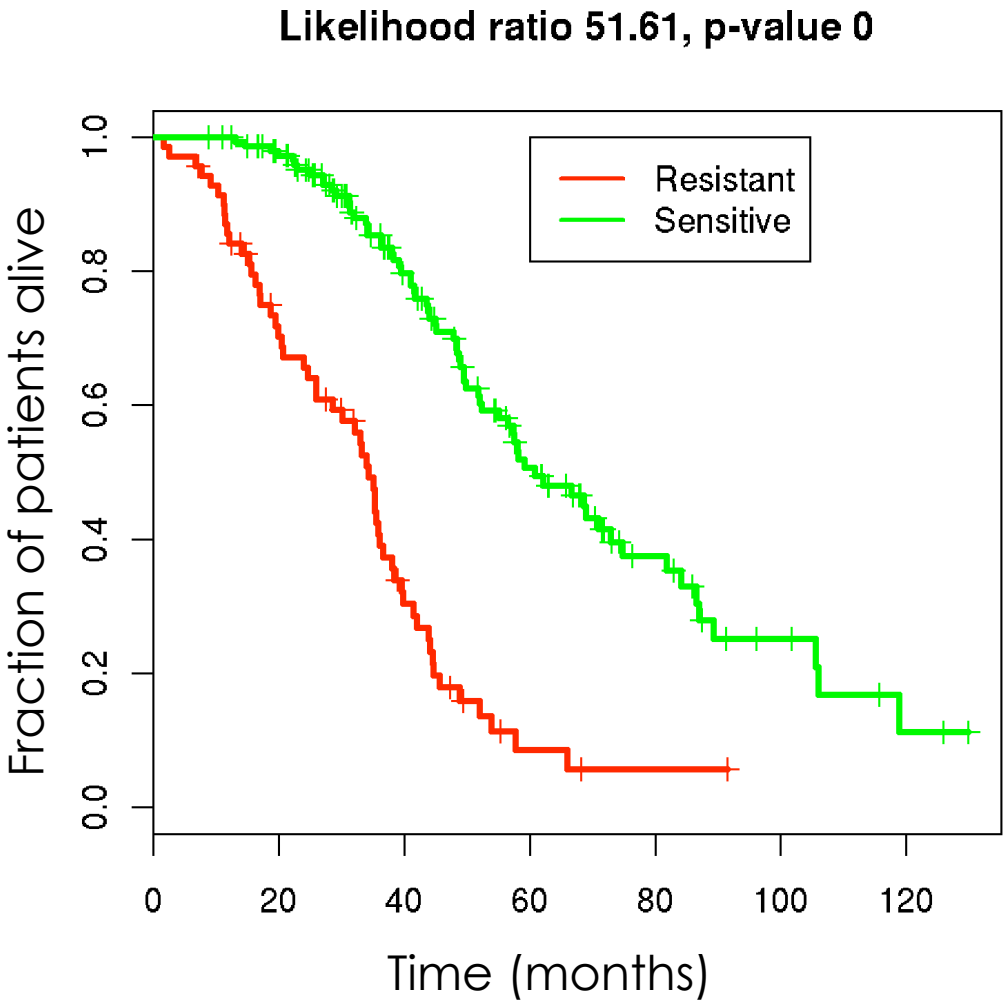
Pathway rather than gene features are used in model



Molecular Signatures database (MSigDB)
(www.broadinstitute.org/msigdb/)

Survival of platinum resistant patients versus sensitive patients in ovarian cancer

70 Resistant patients vs 156 Sensitive patients

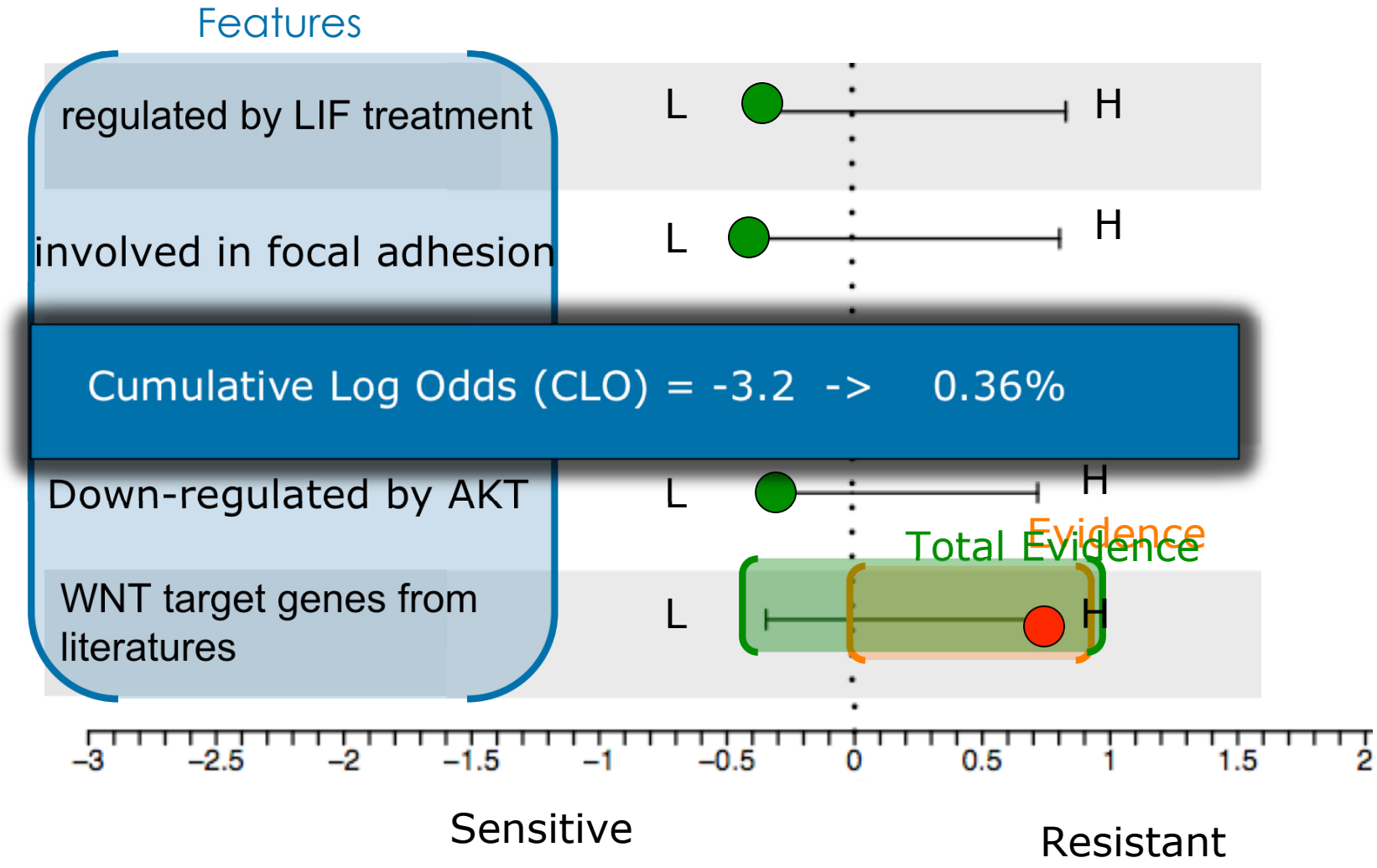


Platinum Resistant: recurrence of cancer within the first 6 months after platinum therapy.

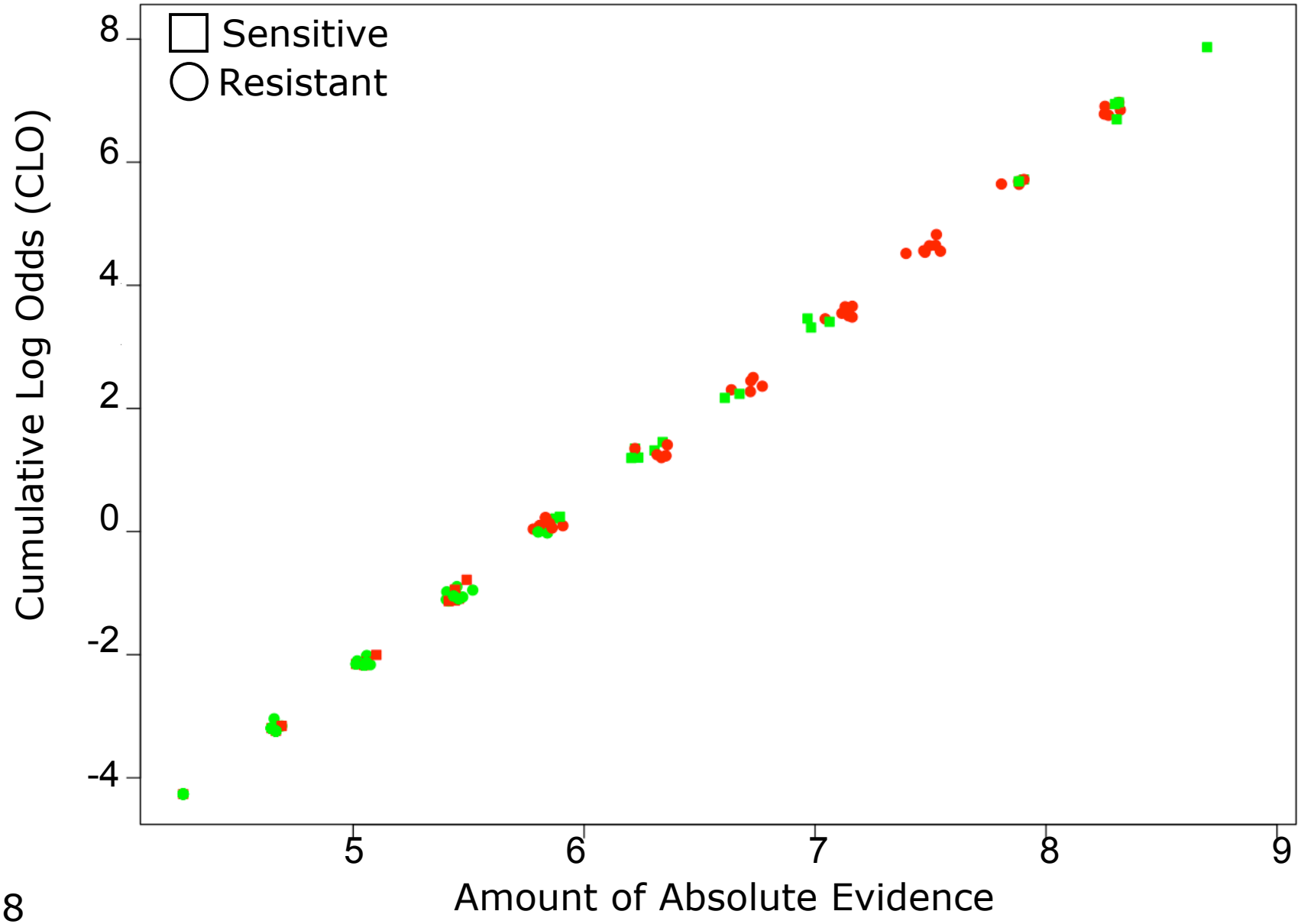
Platinum Sensitive: No recurrence of cancer within the first 6 months after treatment

Nomograms are a graphical way of representing a classification model

Ovarian cancer Bayesian nomogram for a specific patient. Showing 5 out of 12 total single features (current model used)



Current model for ovarian cancer has two main issues: features are strongly correlated and there are a lot of errors made even when sufficient evidence is available.

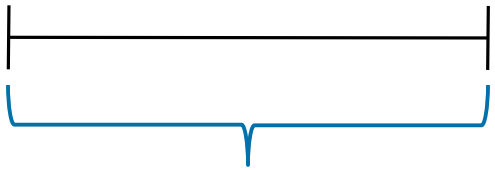


One way of improving these results is using pairs of features instead of single features

- Important things to know about dealing with pairs of features:
  - The evidence of a pair is not the same as the sum of the individual evidences.
  - Before, we had n features, now we have $(n^2 - n)/2$ possible pairs of features
  - Pairs not only perform better, they also have the potential to uncover interesting

How did I approach this problem?

Find the top 100 features
(out of ~3 million) based on

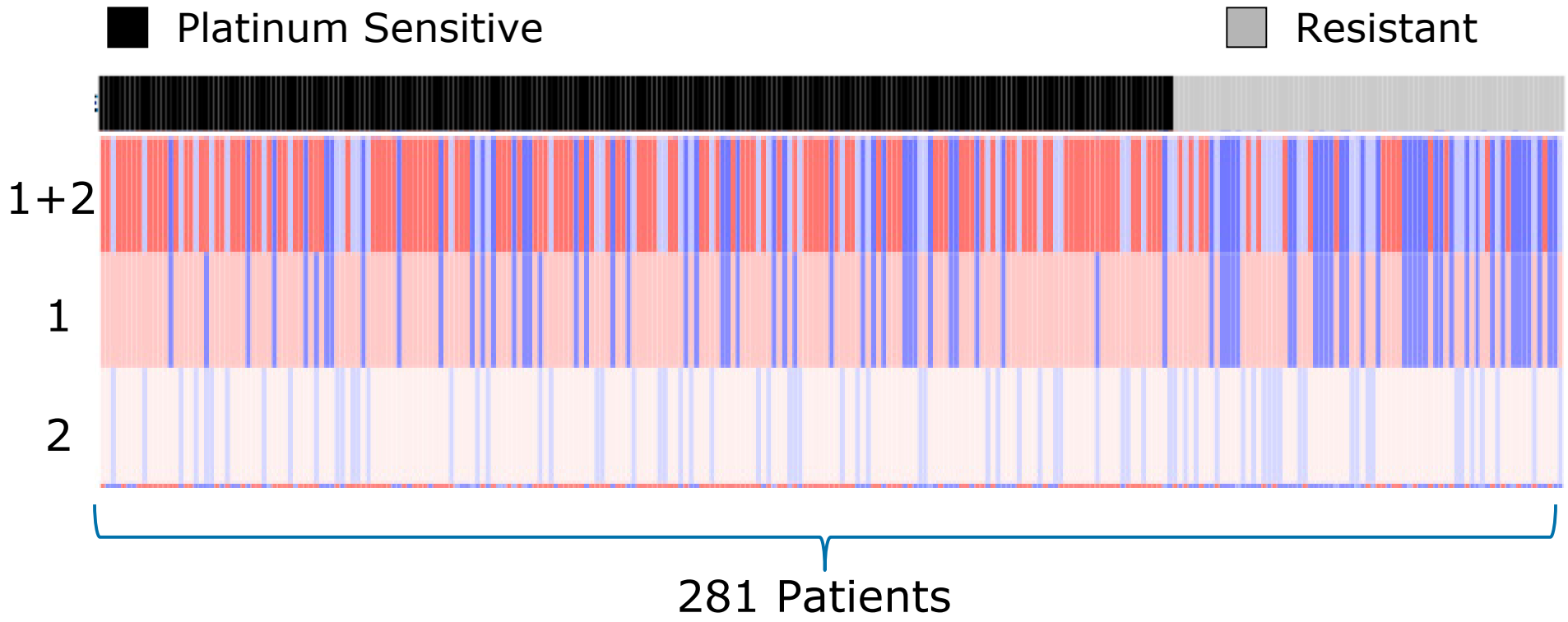


The longer the arms of the nomogram, the better the feature

subset based on their quality
and unambiguous annotation

Create the nomogram based
on selected features

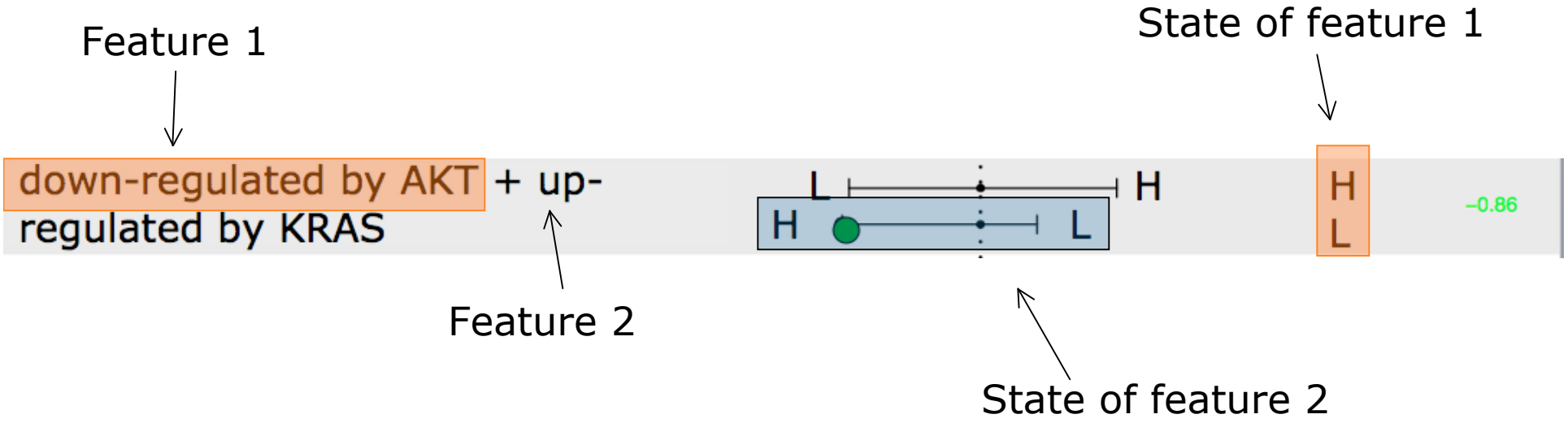
Ovarian cancer pairs work well because they correct each other's mistakes and they reinforce each other when correct.

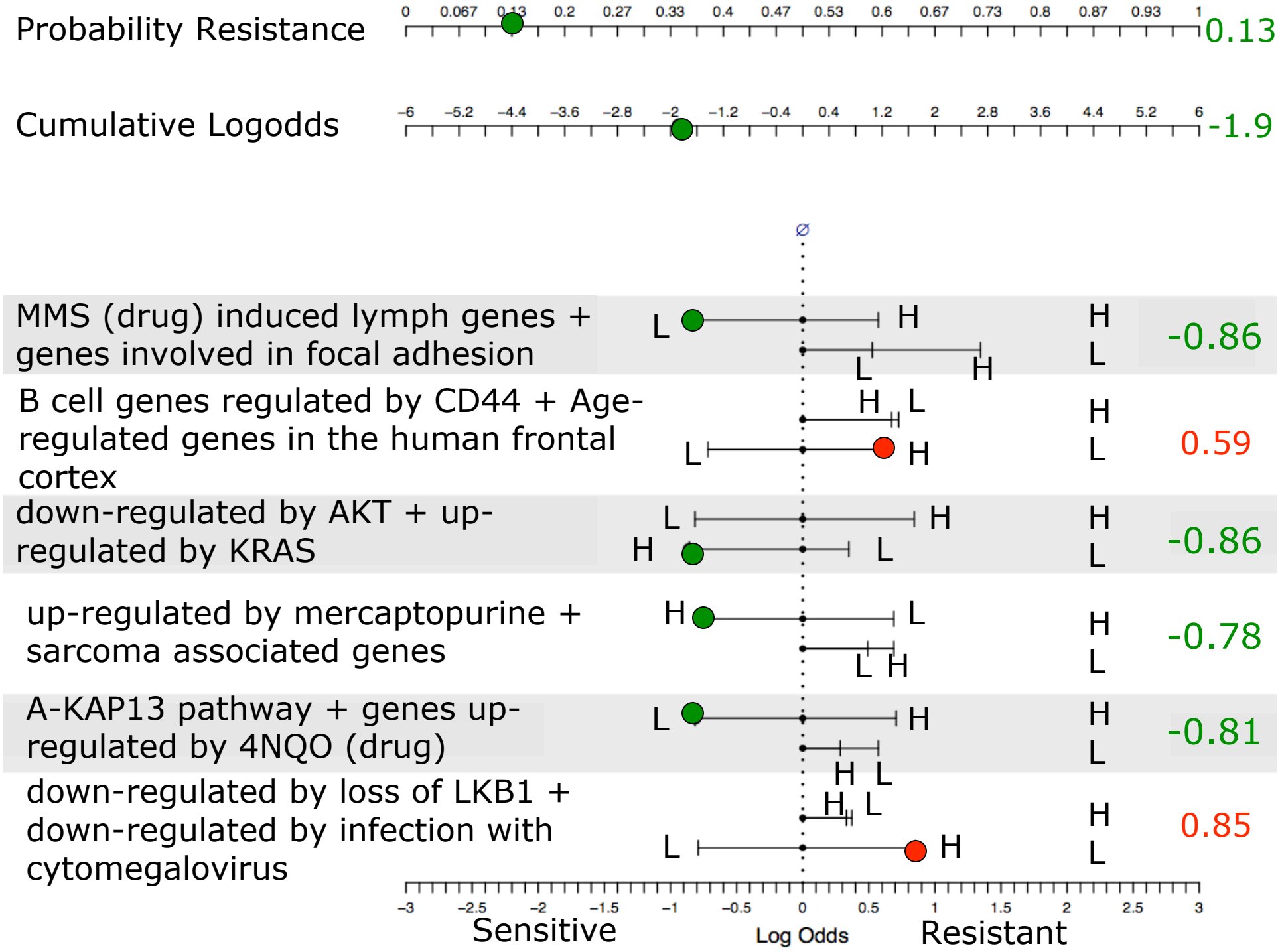


Feature 1: genes down regulated by AKT (oncogene)
Feature 2: genes up regulated by KRAS (oncogene)

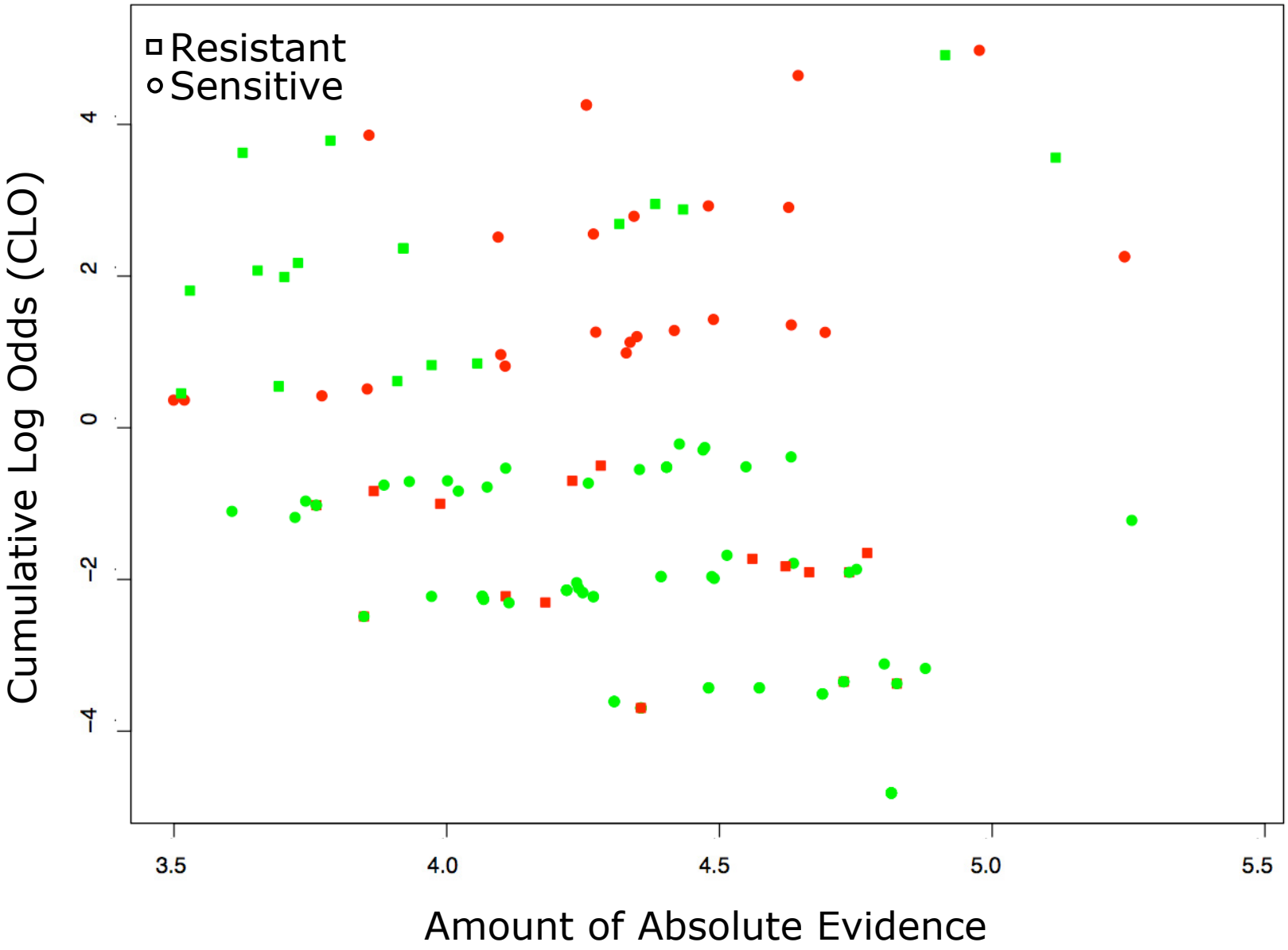
Intensity of color represents the certainty of the prediction. Red represents sensitive prediction and blue represents resistant prediction

Here's how we visualize pairs of features in a nomogram




Probability Resistance
0 0.067 0.13 0.2 0.27 0.33 0.4 0.47 0.53 0.6 0.67 0.73 0.8 0.87 0.93 1
0.13
Cumulative Logodds
−6 −5.2 −4.4 −3.6 −2.8 −2 −1.2 −0.4 0.4 1.2 2 2.8 3.6 4.4 5.2 6
-1.9
MMS (drug) induced lymph genes + genes involved in focal adhesion
-0.86
B cell genes regulated by CD44 + Age-regulated genes in the human frontal cortex
0.59
down-regulated by AKT + up-regulated by KRAS
-0.86
up-regulated by mercaptopurine + sarcoma associated genes
-0.78
A-KAP13 pathway + genes up-regulated by 4NQO (drug)
-0.81
down-regulated by loss of LKB1 + down-regulated by infection with cytomegalovirus
0.85
−3 −2.5 −2 −1.5 −1 −0.5 0 0.5 1 1.5 2 2.5 3
Sensitive
Log Odds
Resistant

Ovarian cancer pairs show a lesser degree of correlation than single features



Comparison of models for ovarian cancer. Both ROC and Error rate improvement from using pairs

Model	au-ROC	Error Rate	# of features
Singles	0.595	42.3%	12
Pairs	0.681	30.7%	6 pairs

Prediction

Real Outcome: Platinum Status	Correct	Incorrect
Resistant	20	15
Sensitive	73	29

Comparison of models for Medulloblastoma

Model	au-ROC	Error Rate	# of features
A	0.727	32.1%	2
B	0.747	34.6%	3
C	0.802	25.6%	46
D	0.782	25.6%	53
Pairs	0.803	29.5%	5 pairs + 2 singles

Conclusion

- Pairs are better at predicting outcome
- The fact that pairs "work together" might uncover interesting interactions between the features
- Pairs dramatically reduce the correlation in the model
- Naïve Bayesian is a simple approach that physicians can understand and is as accurate as more complex models

Data Acknowledgement

Datasets:

- Medulloblastoma Dataset
  - Training Set: 94 samples multi-institutional
    - Children's Hospital (Boston)
    - Children Oncology Group (COG)
    - U of Washington Medical Center
    - Children's Hospital (Texas)
    - The Johns Hopkins Medical Center
  - Independent test set: 78 samples multi-institutional
    - 47 samples, Pomeroy et al. 2002
    - 16 samples, Kool et al. 2008
    - 15 samples, COG 2009

BROAD INSTITUTE

Data Acknowledgement

Datasets:

- Ovarian Cancer Datasets
  - Toothill
  - TCGA

Acknowledgements

I would like to thank the following people for their contributions:
Roel Verhaak
Jill Mesirov
Todd Golub
Scott Pomeroy

I would also like to thank the following people for all their support.
Eboney Smith
Jacqueline Nkuebe
Bruce Birren

BROAD INSTITUTE

Probabilistic Model

- Conditional Probabilities & Bayes Theorem for independent features:

$$P(r \mid x_1, x_2, \ldots, x_n) = \frac{P(r, x_1, x_2, \ldots, x_n)}{P(x_1, x_2, \ldots, x_n)} = P(r \mid x_1)P(r \mid x_2)\ldots P(r \mid x_n)$$

- Weight of Evidence for target r and state x:

$$Ev(r=yes \mid x=female) = \log \left[\frac{P(r = yes \mid x = female) / P(r = no \mid x = female)}{P(r = yes) / P(r = no)} \right]$$

- Total evidence that feature x provides:

$$AvEv(r \mid x) = \sum_{i=1}^{|X|} P(x = X_i)\,|Ev(r \mid x = X_i)| \qquad \text{Where } X = \{Male, Female\}$$

Computing evidence for pairs

- Very similar to computing the evidence for singles

$$Ev(r=yes \mid x=female, y=adult) = \log \left[\frac{P(r = yes \mid x = female, y = adult) / P(r = no \mid x = female, y = adult)}{P(r = yes) / P(r = no)} \right]$$

- Remember Bayes Theorem:

$$P(r=yes \mid x=female, y=adult) = \frac{P(r=yes, x= female, y = adult)}{P(x=female, y = adult)}$$

- Total evidence that pair of features x, y provides:

$$AvEv(r \mid x, y) = \sum_{i=1}^{|X|} \sum_{j=1}^{|Y|} P(x = X_i, y=Y_j) \left| Ev(r \mid x = X_i, y=Y_j) \right|$$

Where X = {Male, Female}

Where Y = {Adult, Child}