

# Improving predictions of Ames mutagenicity by correcting unbalanced training data

Jonathan Vessey

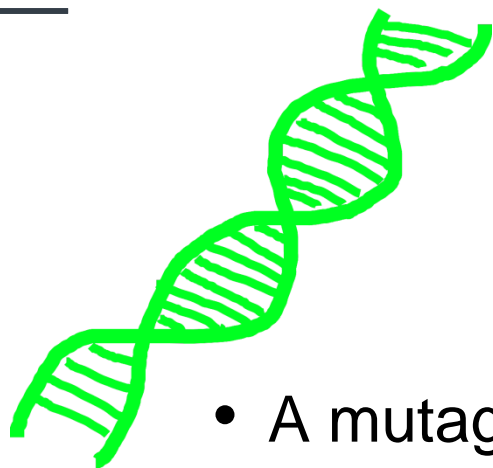
Research Leader

[jonathan.vessey@lhasalimited.org](mailto:jonathan.vessey@lhasalimited.org)



Leaders in the development of expert chemoinformatic systems  
and trusted curators of proprietary data.

# Introduction: Mutagenicity



- A mutagen produces change in DNA which leads to increased mutation
- Considered a precursor to carcinogenicity
- Prediction is important especially for potentially genotoxic impurities subject to [ICH M7 guidelines](#)

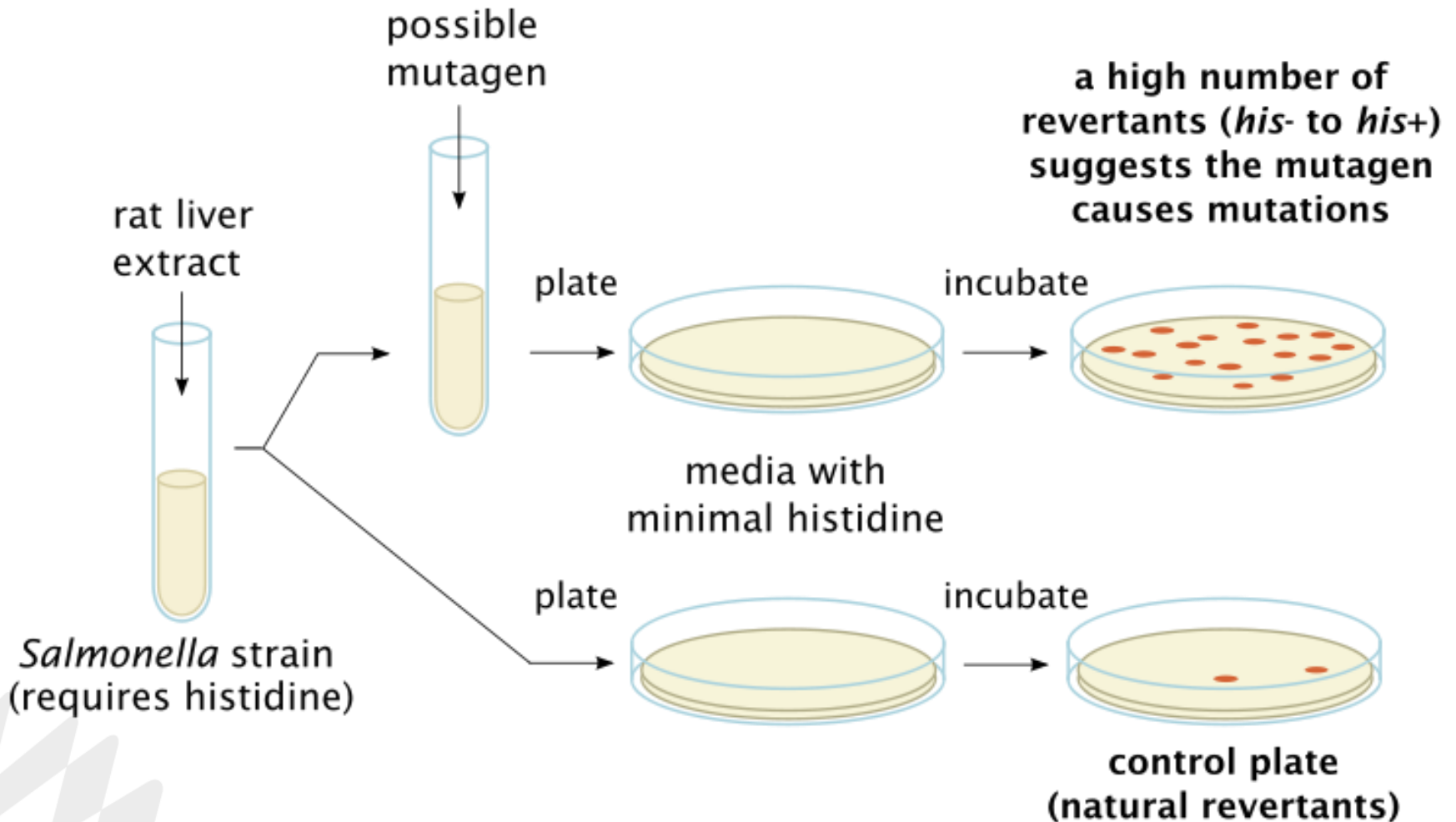


# Introduction: Ames test for mutagenicity


- An *in vitro* assay for mutagenicity
- Identifies compounds which cause mutation in strains of *Salmonella typhimurium*
- Different strains show different sorts of mutation
- Considered about 85% reproducible<sup>1</sup>
- Often Ames test data is used as the basis for prediction of mutagenicity
- Predictions of mutagenicity are frequently predictions of Ames test

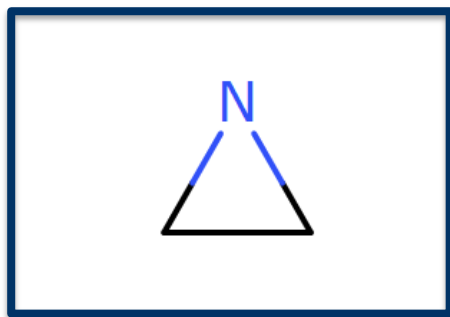
<sup>1</sup>Piegorsch WW, Zeiger E . *Measuring intra-assay agreement for the Ames Salmonella assay*. In: Rienhoff O, Lindberg DAB, editors. *Statistical Methods in Toxicology*. Heidelberg, Germany: Springer-Verlag; 1991. p. 35-41.

# Introduction: Ames test for mutagenicity



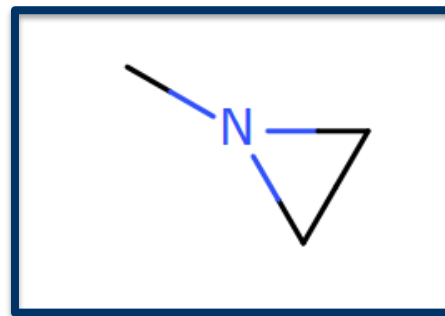
# Introduction: SOHN models

- Self-organising hypothesis network
- Used as the basis of predictions in Sarah Nexus 
- In Sarah Nexus all hypotheses are substructure matches
- The support set of hypotheses is used to organise them as being a subset or superset of each other



A

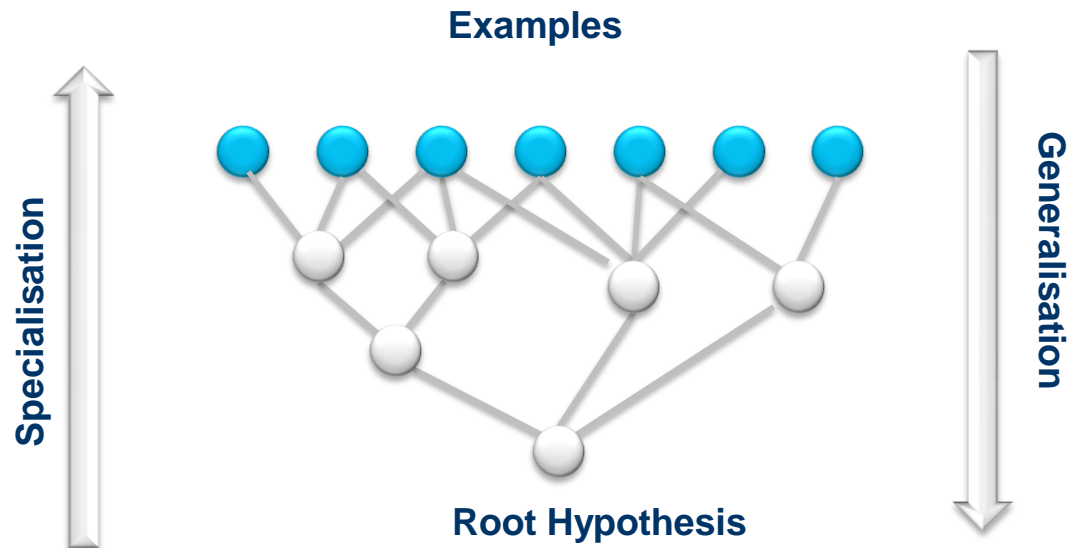
  
more general than



B

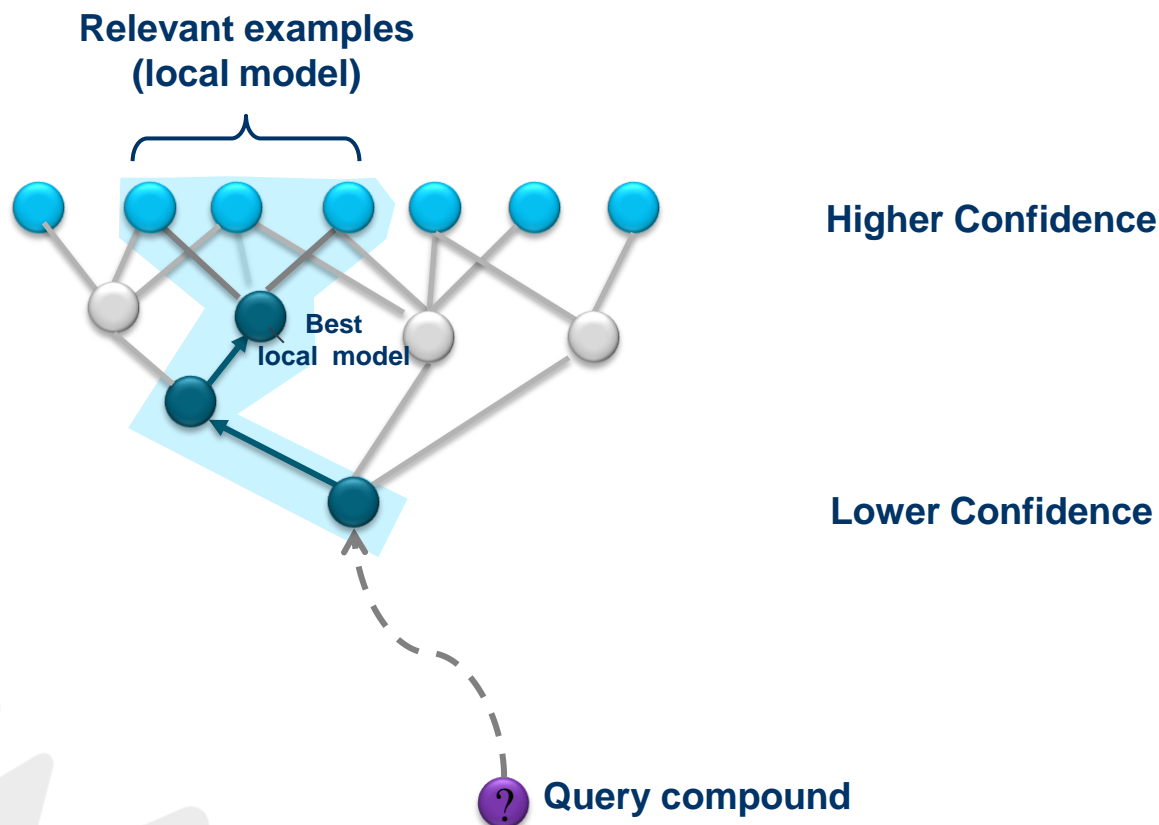
# Introduction: SOHN models

- The organisation of the hypotheses produces a network from the most general to the most specific



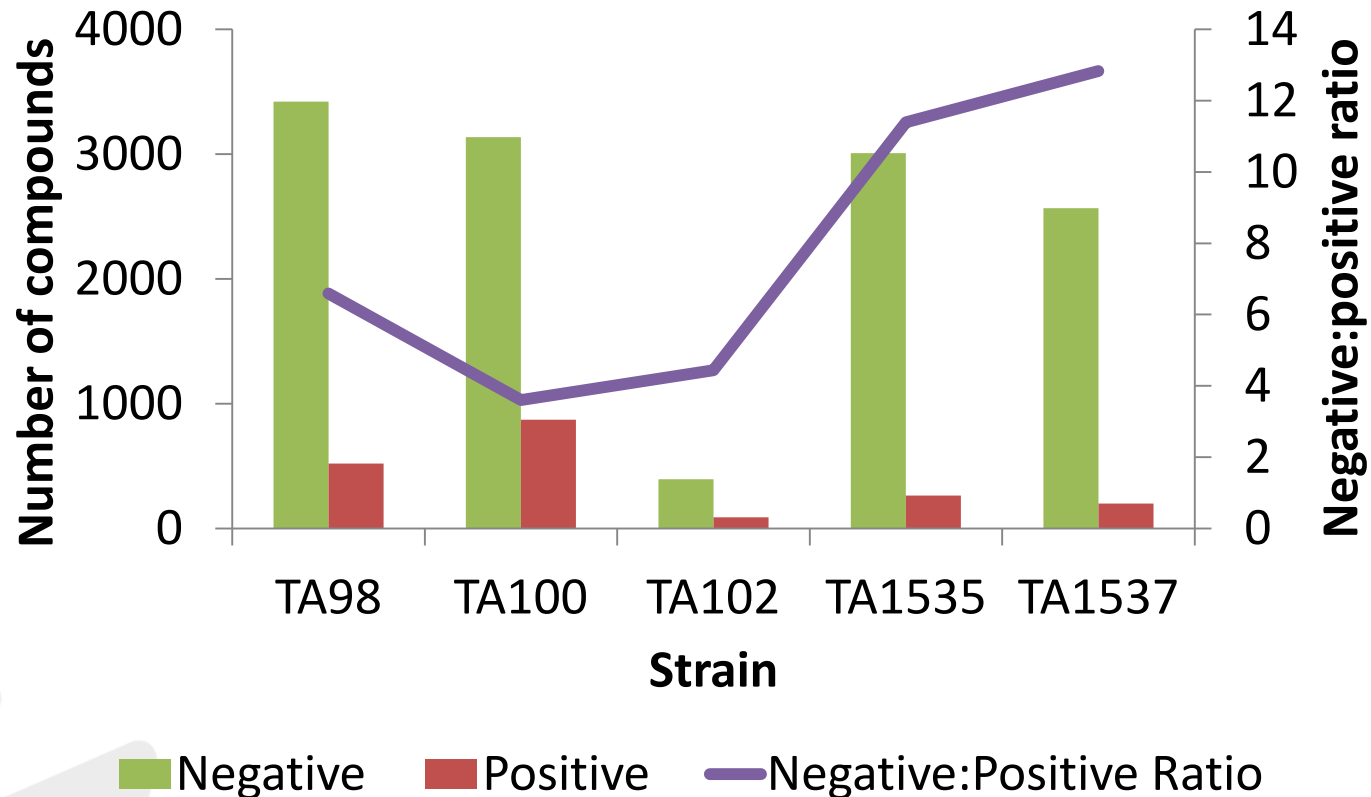
# Introduction: SOHN models

- Predictions are made from the most relevant hypotheses
- Predictions are made with a corresponding confidence



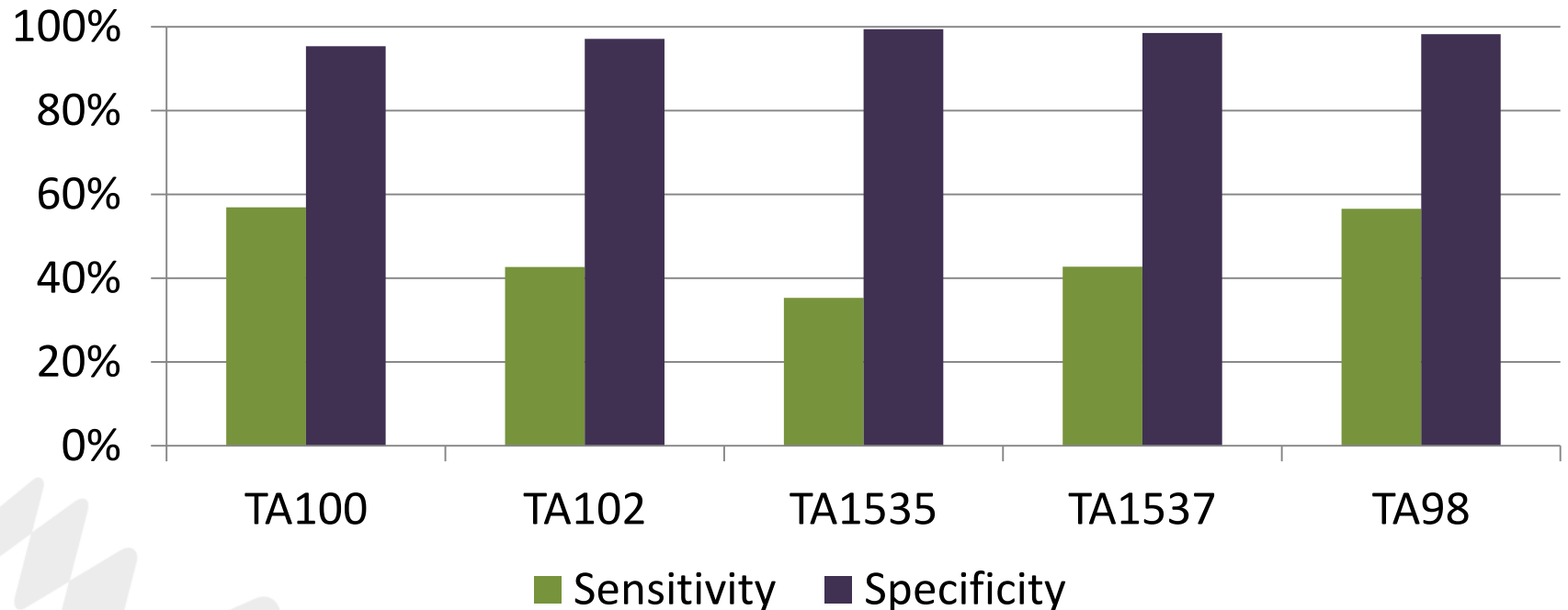
## Data to be modelled

- Data modelled from 5 different strains
  - Different numbers of compounds in each strain
  - Different bias in each strain: but negatives > positives



# SOHN Prediction accuracy without correction

- Overlearns the major class
- Where the minor class is positives, leads to poor sensitivity and high specificity



# Different methods of attempting to correct the bias



- Oversample the minor class
- Undersample the major class
- Weight the learning by penalising incorrect predictions of the minor class



# Different methods of attempting to correct the bias

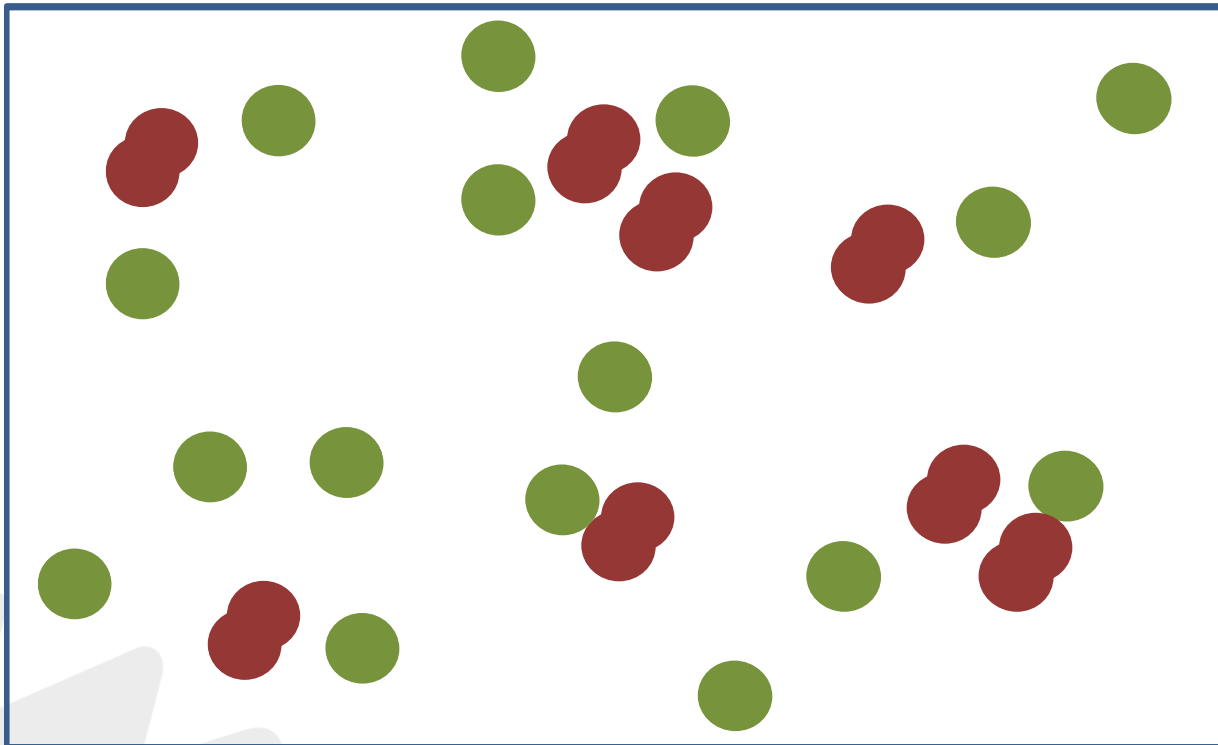


- Oversample the minor class
- Undersample the major class
- Weight the learning by penalising incorrect predictions of the minor class



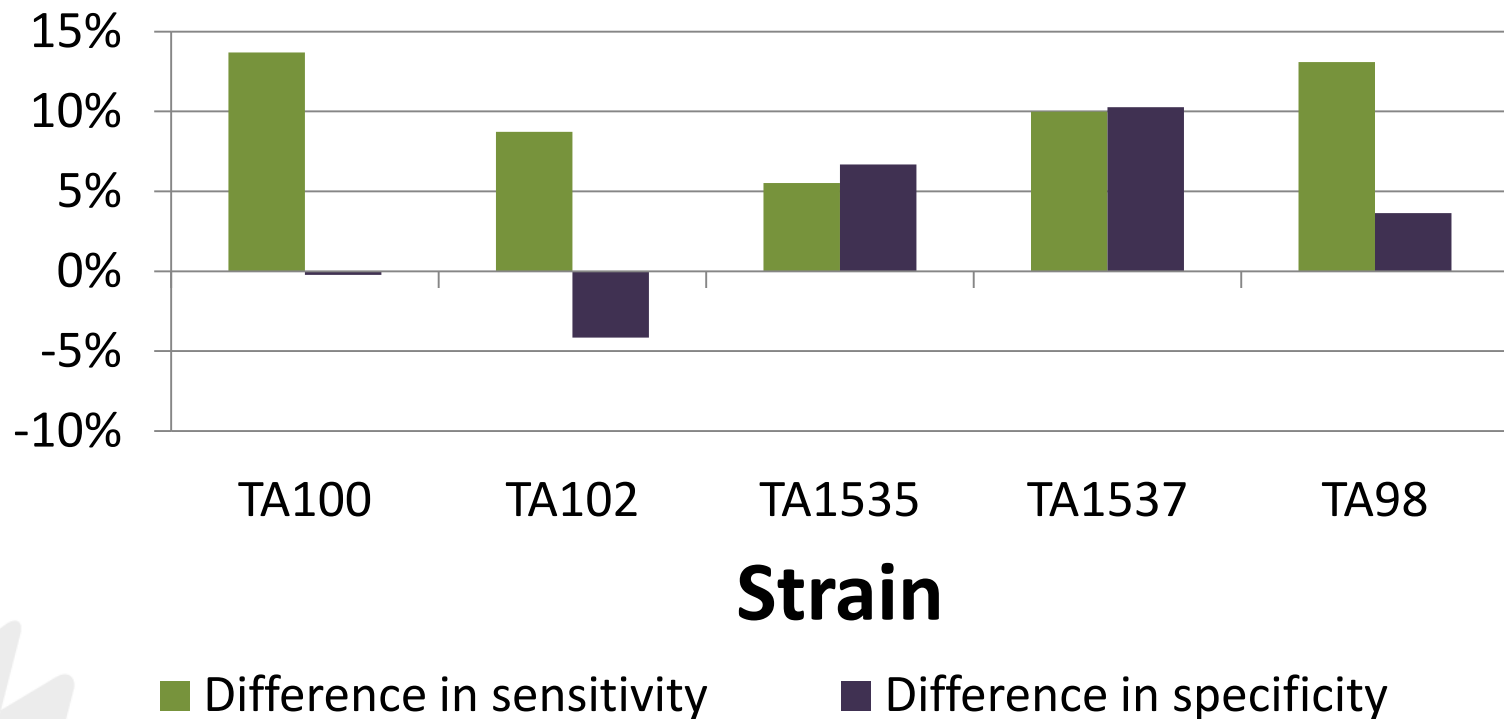
## Oversample the minor class

- In this simple case of a training set where the bias is 2 : 1 in favour of Ames negatives
- We can correct the bias by sampling each element of the minor class twice



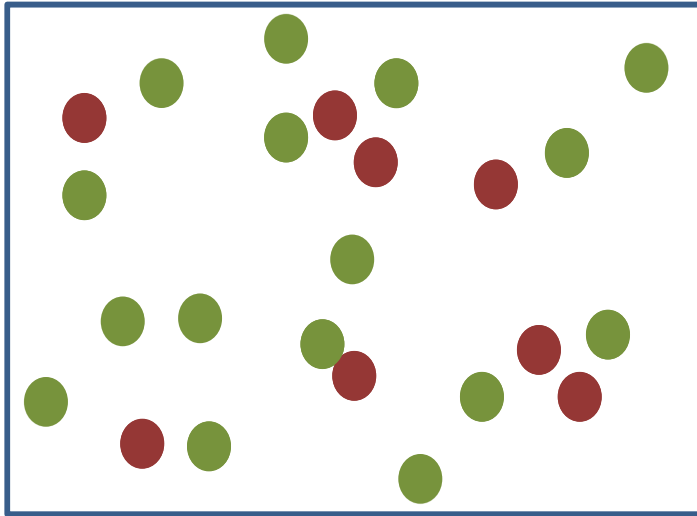
## Oversample the minor class

- Change in prediction metrics from SOHN models: increase in sensitivity, sometimes a decrease in specificity

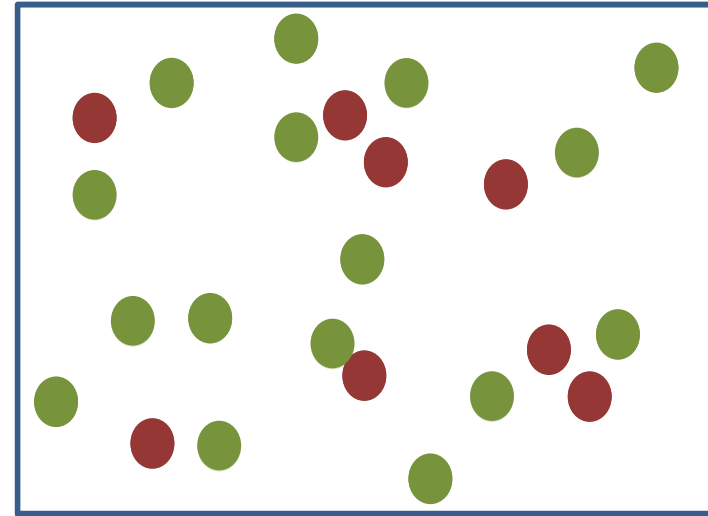


# Undersample the major class: deterministic approaches

- Will always lose information from the major class



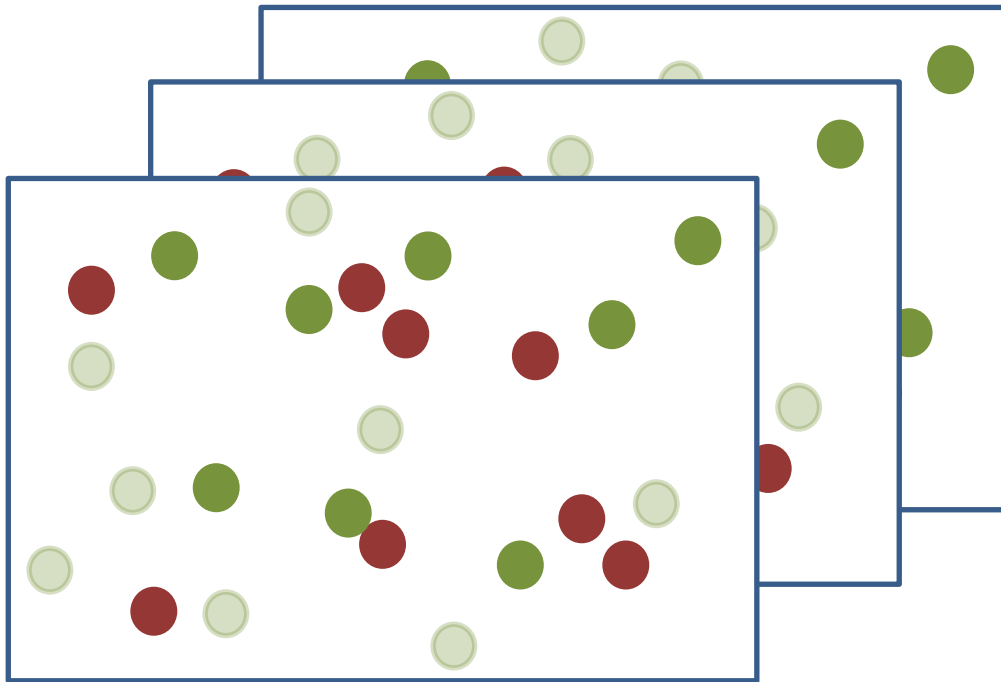
Major class data points closest to minor class data points



The most diverse set of data from the major class

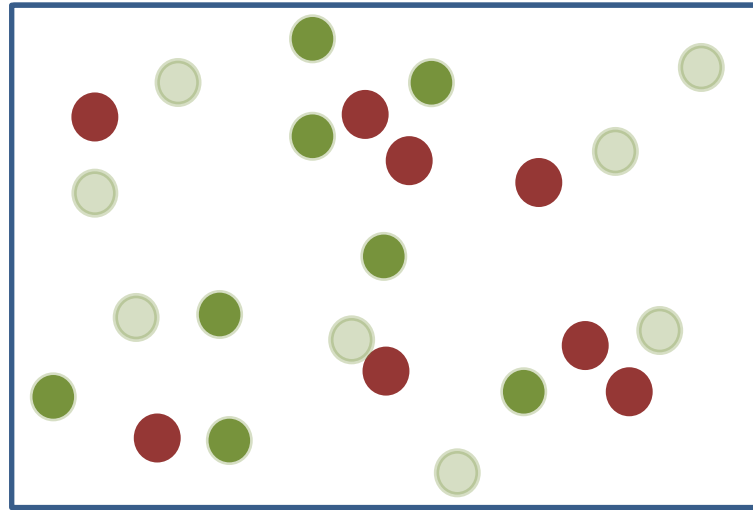
# Undersample the major class: stochastic approaches

- Repeatedly sample at random



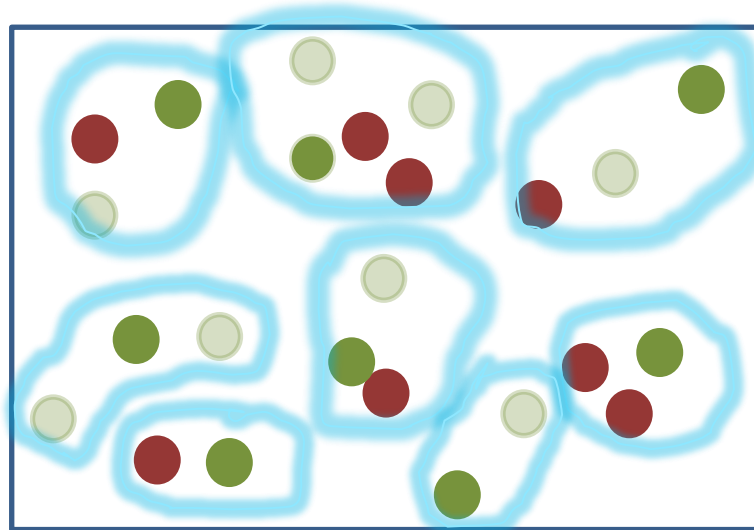
# Undersample the major class: stochastic approaches

- Repeatedly sample ensuring all are selected



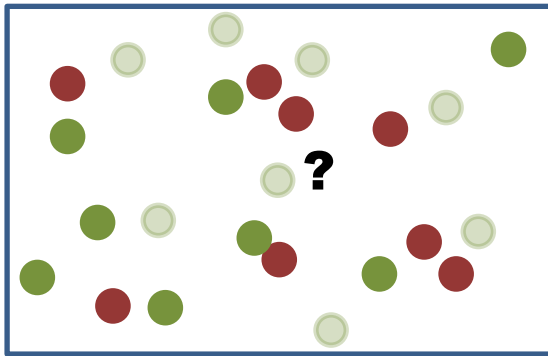
# Undersample the major class: stochastic approaches

- Randomly sample from clusters

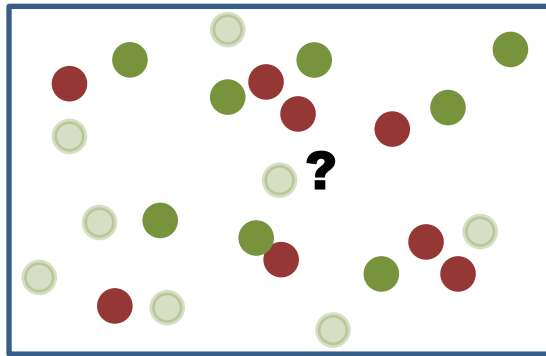


# Undersampling: Combining Multiple models

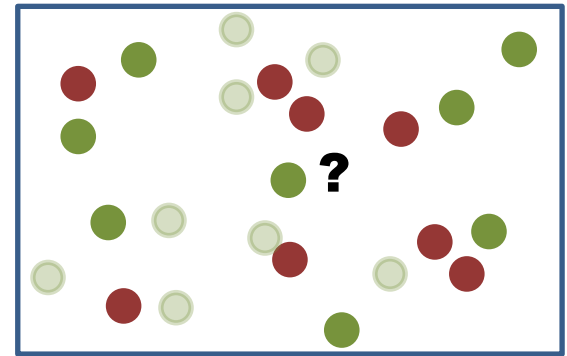
- Majority vote



? = ●



? = ●

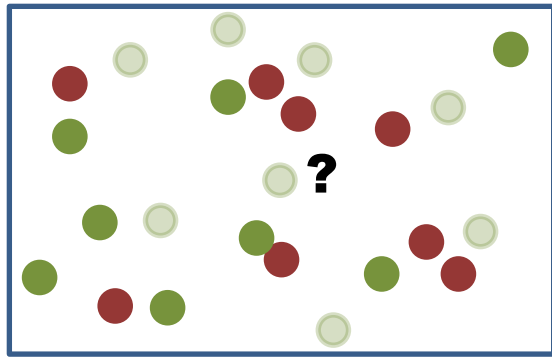


? = ●

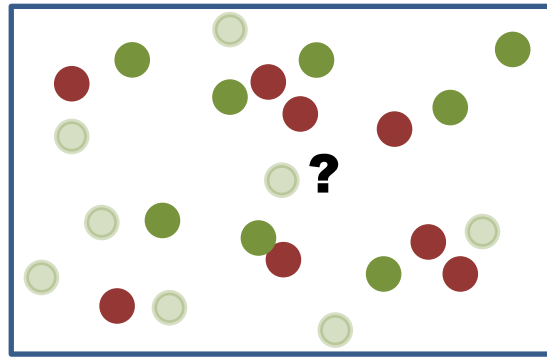
Positive wins 2 votes to 1

# Undersampling: Combining Multiple models

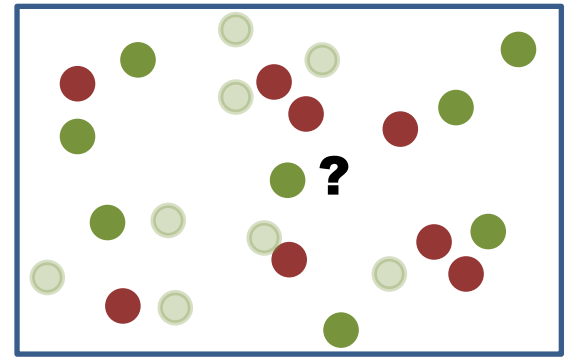
- Single most confident



? = ●



? = ●



? = ●

The most confident single prediction is negative



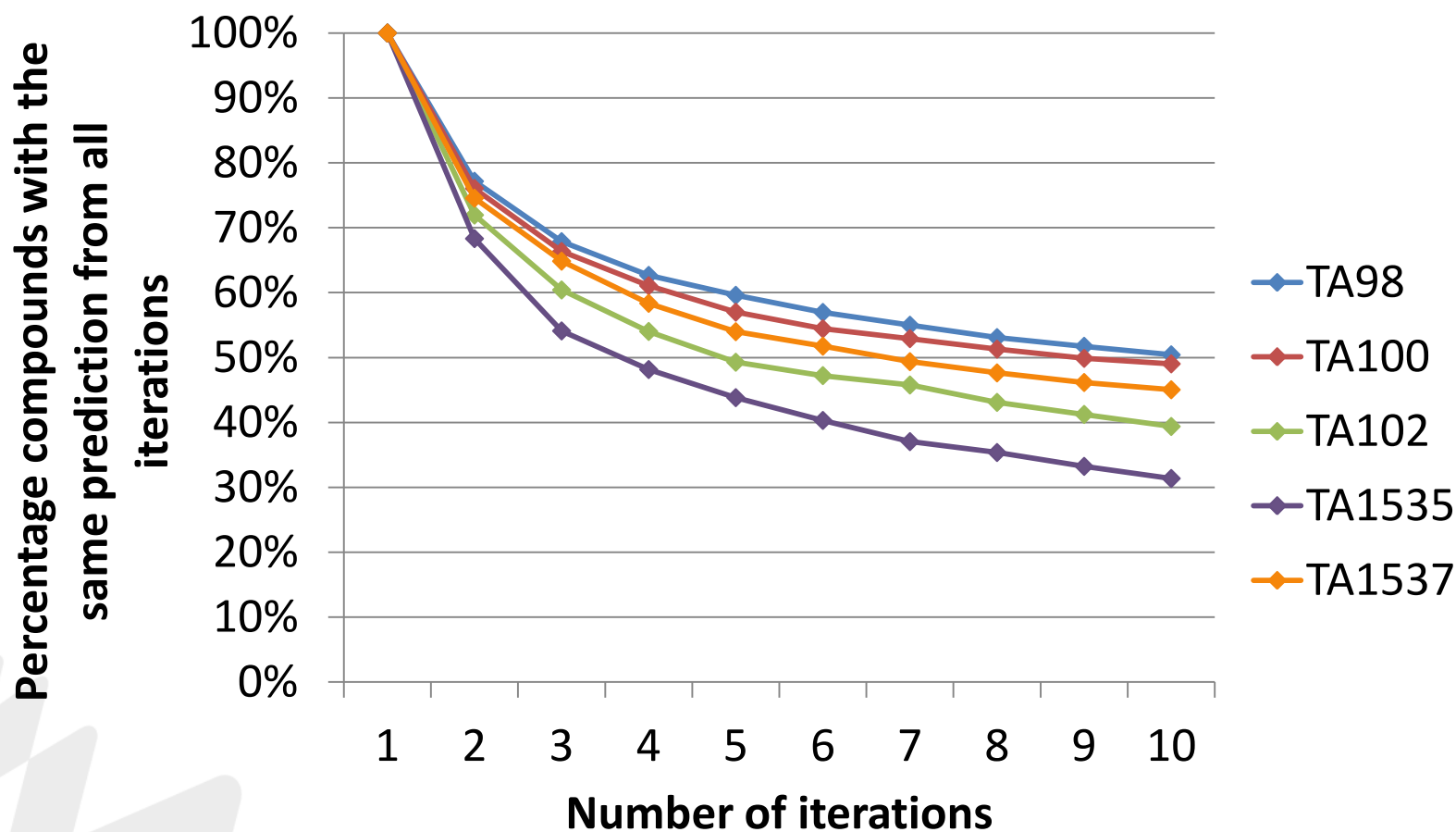
# Sarah Nexus' confidence score correlates with accuracy

Confidence Score	0-20%	20-40%	40-60%	60-80%	80-100%
PPV	58%	74%	85%	93%	92%
NPV	62%	80%	95%	96%	97%

Sarah vs. external dataset of confidential data from Lhasa members

# Undersampling: Variation from Multiple models

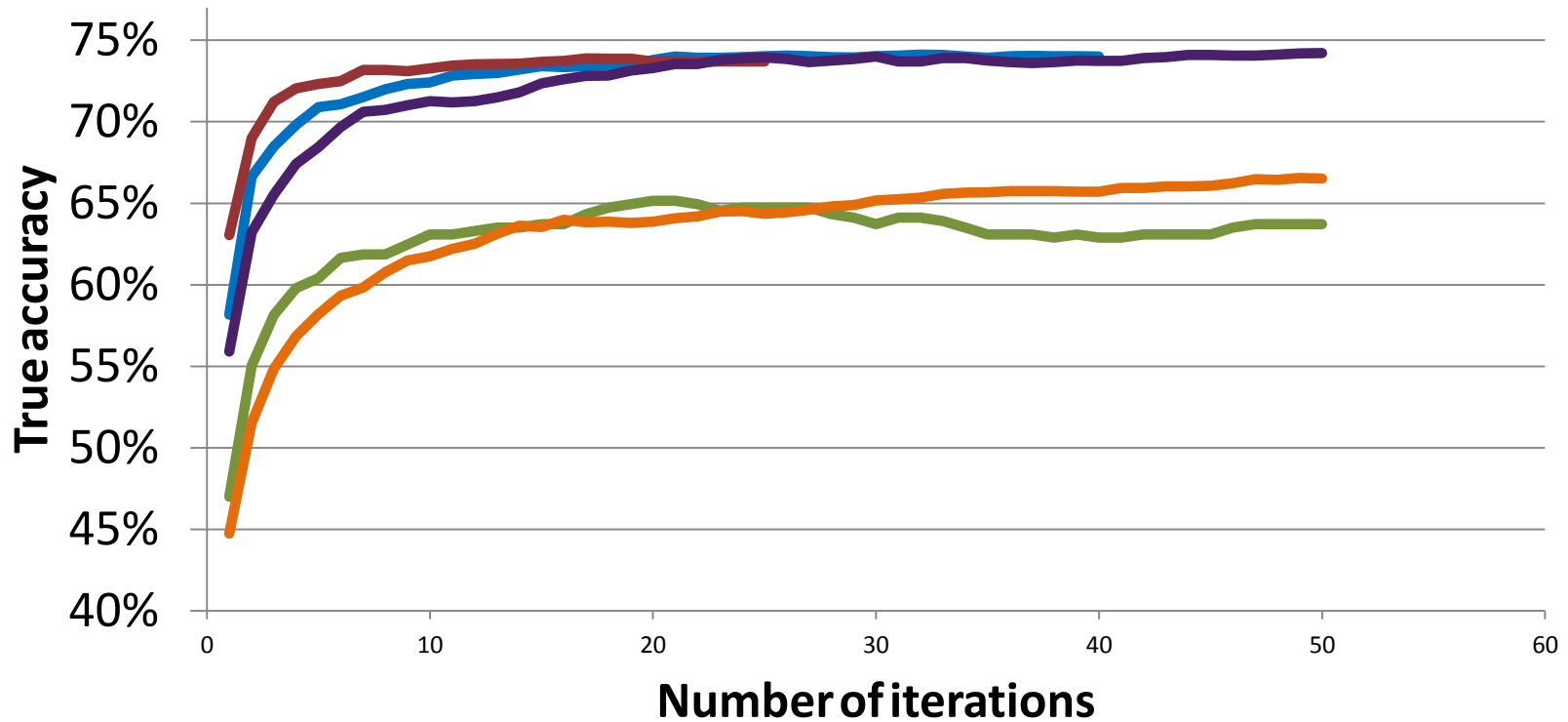
- With more than one model we get the possibility of variation between predictions





Leaders in the development of expert chemoinformatic systems  
and trusted curators of proprietary data.

# Undersampling: Reaching the best performance



— TA98, selected at random

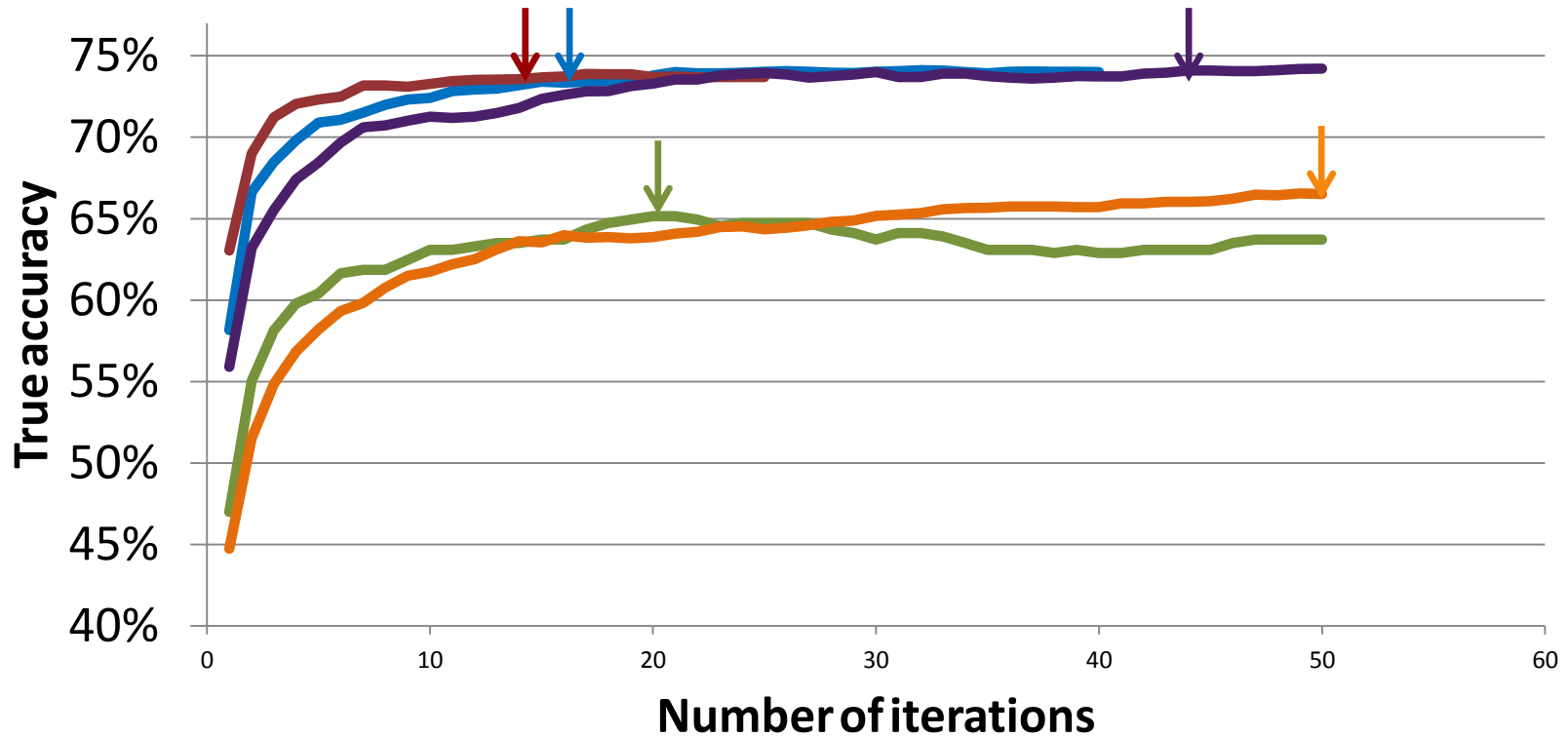
— TA102, selected at random

— TA1537, selected at random

— TA100, selected at random

— TA1535, selected at random

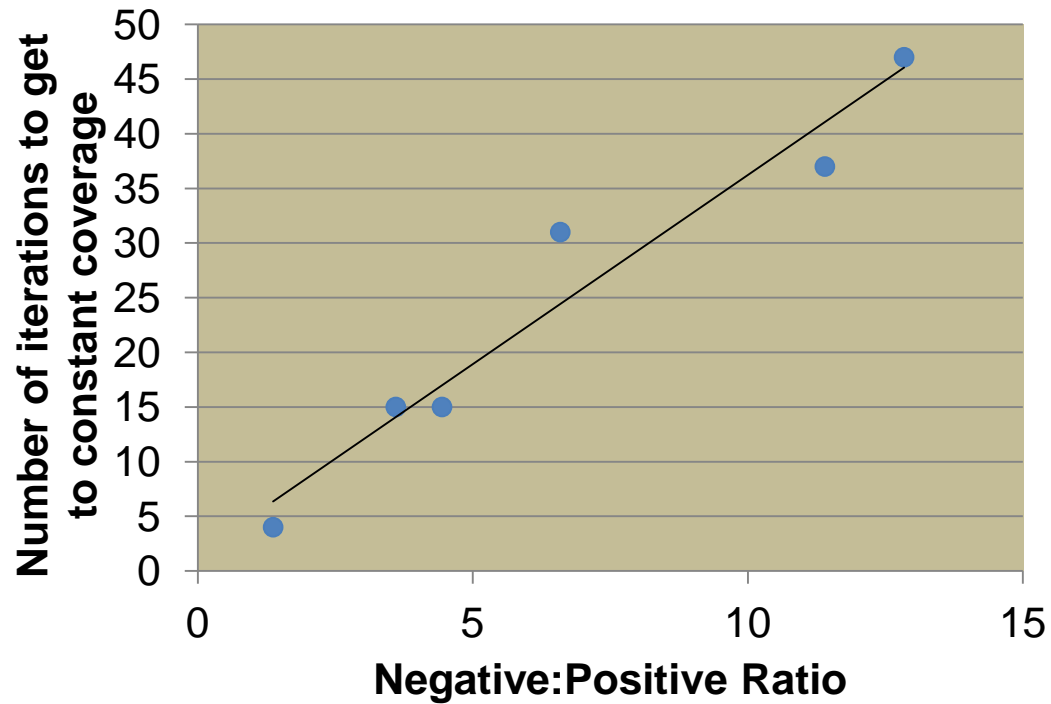
# Undersampling: Reaching the best performance



- TA98, selected at random
- TA102, selected at random
- TA1537, selected at random

- TA100, selected at random
- TA1535, selected at random

# Undersampling: Variation with data bias



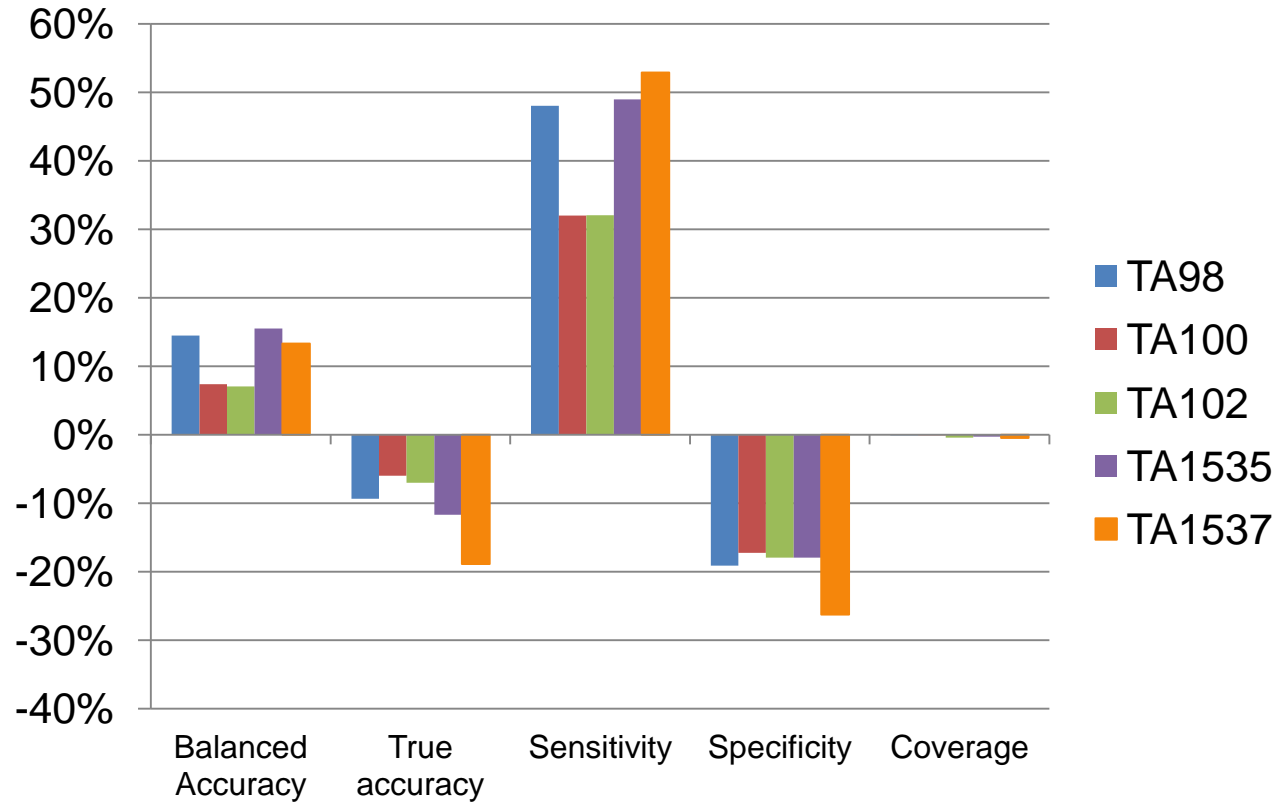
# Undersampling method summary



- Sample at random from the major class
- Repeat the sampling to ca. 5x the dataset bias, i.e. a 3:1 bias in the dataset is sampled 15 times
- Combine predictions from different models by selecting the single most confident prediction



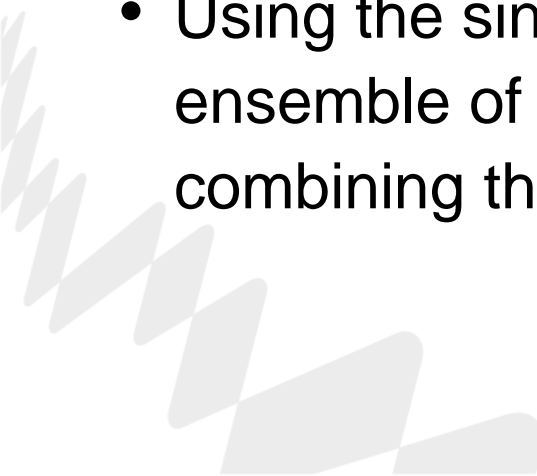
# Undersampling: Performance metrics





## Conclusions


---

- Learning from balanced datasets gives better overall performance than from biased datasets
  - Correcting for the bias in the data by repeated undersampling of the major class allows the ML to learn from all the data in the major class
  - Repeating the undersampling up to ~5 times the bias in the dataset gives the best improvement
  - Using the single most confident prediction from the ensemble of models is the most successful method of combining the predictions
- 



# Acknowledgements

---

- Data and data curation
    - Andrew Thresher
    - Lhasa Knowledge Base Team
    - Sam Webb
  - Development of SOHN methodology
    - Thierry Hanser
    - Ed Rosser
    - Stephane Werner
  - Discussion
    - Chris Barber
    - Naomi Kruhlak
    - Lidiya Stavitskaya
- 

**Thank you!**  
**Any questions?**



shared **knowledge** • shared **progress**

Lhasa Limited Registered Office  
Granary Wharf House, 2 Canal Wharf, Leeds LS11 5PS  
UK Registered Charity (290866)

+44 (0)113 394 6020  
info@lhasalimited.org  
[www.lhasalimited.org](http://www.lhasalimited.org)

