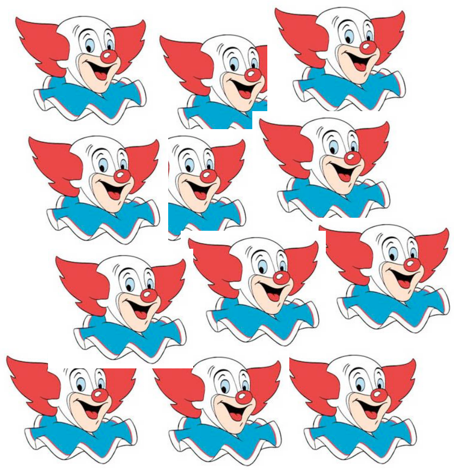
Real World Data is Ugly and Hard; What to Do?

Modern data is:
huge, relentless, deeply skewed, ill-suited, noisy,
and riddled with error and ambiguity.

So: give up on the craftsman model of pattern recognition.

"Ensembles" enable a *commodity* model:

- Accepts data as it is.
- No user tuning required.
- Robust in the face of noise.
- Scales to terabytes of data.
- Always improves accuracy.



Hordes of Bozos for Robust Prediction

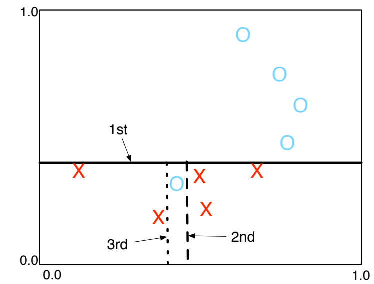
Groundtruth Data for Detection of Ideology

File	Ideo? Truth	Language a_1	CS-1 a_2	CS-2 a_3	…	CS-K a_K
f_1	Yes	12	1003	0.97	…	0.12
f_2	Yes	99	2	0.33	…	0.03
f_3	No	3	27	0.12	…	0.13
f_4	Yes	16	183	0.08	…	0.58
f_5	No	17	665	0.36	…	0.64
f_6	No	44	1212	0.29	…	0.42
f_7	No	42	24	0.33	…	0.88
f_8	Yes	78	42	0.44	…	0.52
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$		$\vdots$
f_N	No	12	3141	0.92	…	0.17

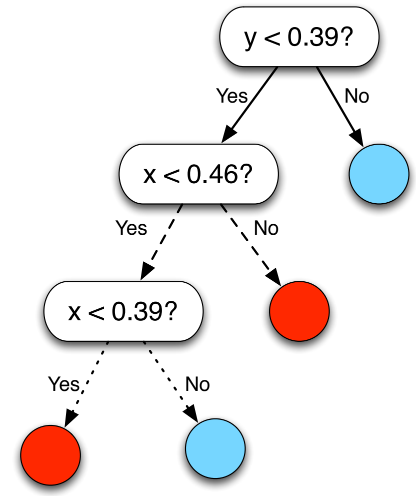
Use Groundtruth to Learn Label Predictor

Also known as: pattern recognition, statistical inference, data mining.

- Input: "ground truth" data.
  - Samples, with attributes, and *labels*.
  - Example: detect ideology in text
    - Samples: a document
    - Attributes: features of the text
    - Labels: "yes", "no"
- Apply suitable method:
  decision trees, neural nets, SVMs.
- Output:
  rules for labeling new, *unlabeled* data.
  Equivalently:
  a partitioning of attribute space.



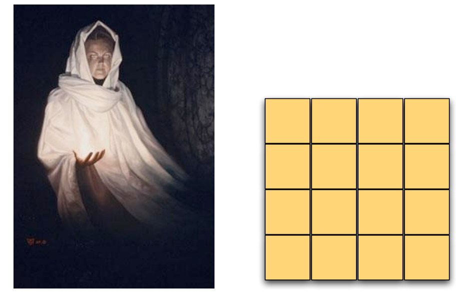
Attribute space partitioned.



Decision tree representation.

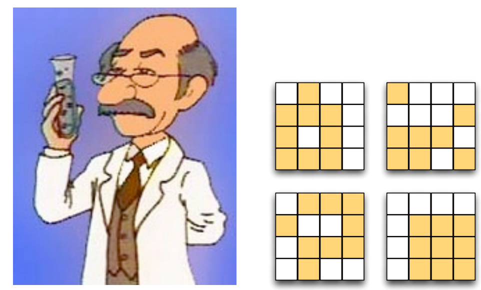
Ensembles: Efficient, Robust, Optimal Accuracy

Traditional: Use 100% of training data to build a sage.



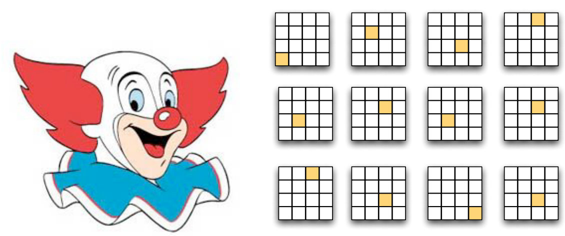
Sage sees all the data.

Ensembles: Use randomized 100% of training data to build an expert. Repeat to build many experts. Vote them.



Each expert sees 2/3rds of the data.

Sandia: Use a semi-random 1% of the training data to build a "bozo". Repeat to build very many bozos. Vote them.



Each bozo sees a tiny fraction.

The experts beat the sage[1].
The bozos beat the experts[4].

Groundtruth is Key! But ...

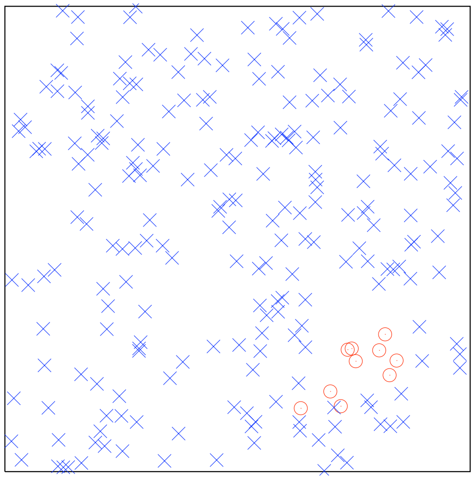
Groundtruth is also ...

- expensive,
- time-consuming,
- and often under-represents the most important class.

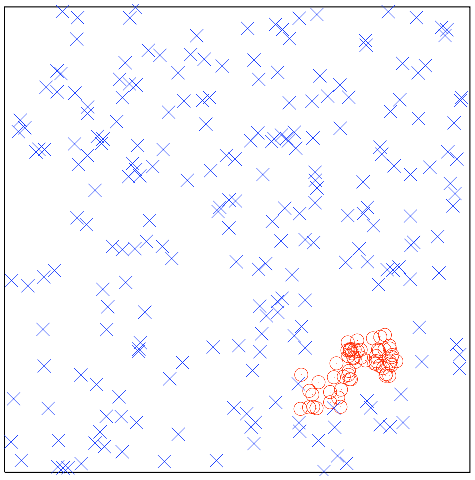
What to do?

SMOTE for Under-Represented Data

- Synthetic Minority Oversampling TEchnique[3].
- Oversample the minority population, but . . .
  . . . simple oversampling induces pathologies.
  So: add *synthetic* samples.
- Method:
  - Pick minority sample.
  - Pick a nearby neighbor.
  - Add new minority sample at a random point between them.
  - Repeat.

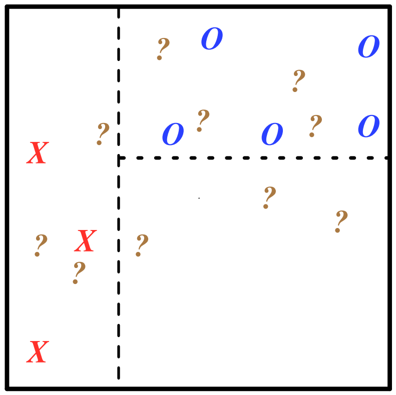


Minority class overwhelmed.

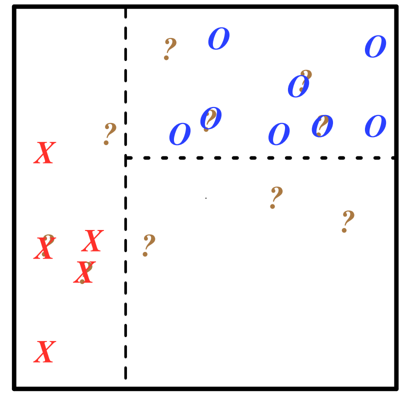


Minority class filled out by SMOTE.

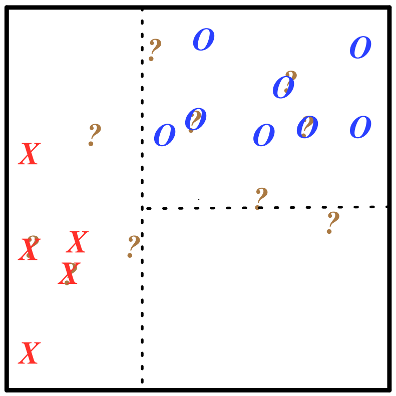
Semi-Supervised Learning[2], To Bootstrap Groundtruth



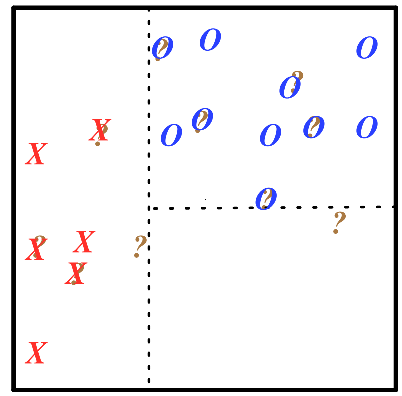
Many unlabeled points.



Re-label the most confident,

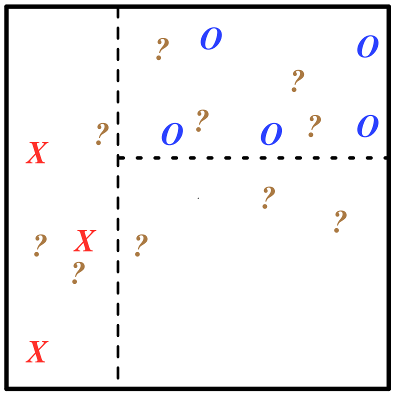


Which changes the decision boundaries,

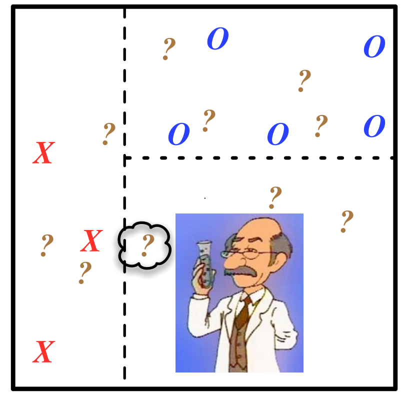


So more points can be labeled.

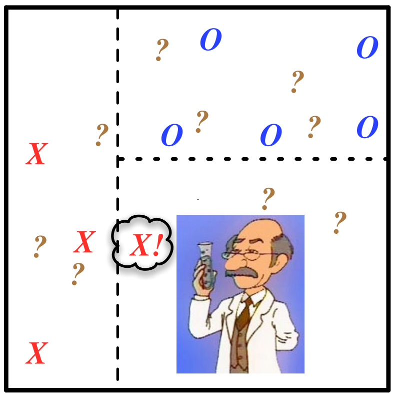
Active Learning[5], To Sharpen Groundtruth



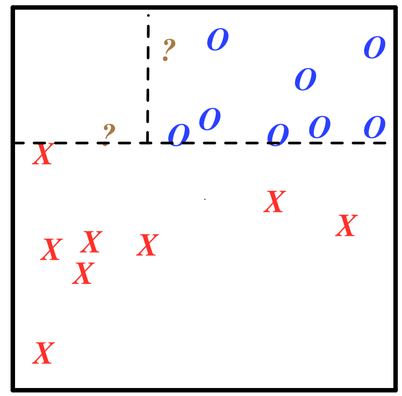
Many ambiguous points.



Find the most telling one to label.



Ask an expert to label it.



More points can be accurately labeled.

Example Applications, Existing or In Development

- Detection of steganography in audio signals.
- Malware classification.
- Search by example in NW simulation data.
- Determine friend or foe from body movement.
- Word classification for entity extraction, for building graphs.
- Predict successful gene expression process parameters.
- Detecting and identifying "ideology" in documents.

References

[1] Banfield, R. E., Hall, L. O., Bowyer, K. W., Bhadoria, D., Kegelmeyer, W. P., and Eschrich, S. A comparison of ensemble creation techniques. In *Proceedings of the Fifth International Conference on Multiple Classifier Systems, MCS2004* (2004), J. K. F. Roli and T. Windeatt, Eds., vol. 3077 of *Lecture Notes in Computer Science*, Springer-Verlag.

[2] Chapelle, O., Schölkopf, B., and Zien, A., Eds. *Semi-Supervised Learning*. MIT Press, Cambridge, MA, 2006.

[3] Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research 16* (2002), 321–357.

[4] Chawla, N. V., Hall, L. O., Bowyer, K. W., and Kegelmeyer, W. P. Learning ensembles from bites: A scalable and accurate approach. *Journal of Machine Learning Research 5* (2004), 421–451.

[5] Cohn, D. A., Ghahramani, Z., and Jordan, M. I. Active learning with statistical models. In *Advances in Neural Information Processing Systems* (1995), G. Tesauro, D. Touretzky, and T. Leen, Eds., vol. 7, The MIT Press, pp. 705–712.