INTRODUCTION TO DATA SCIENCE

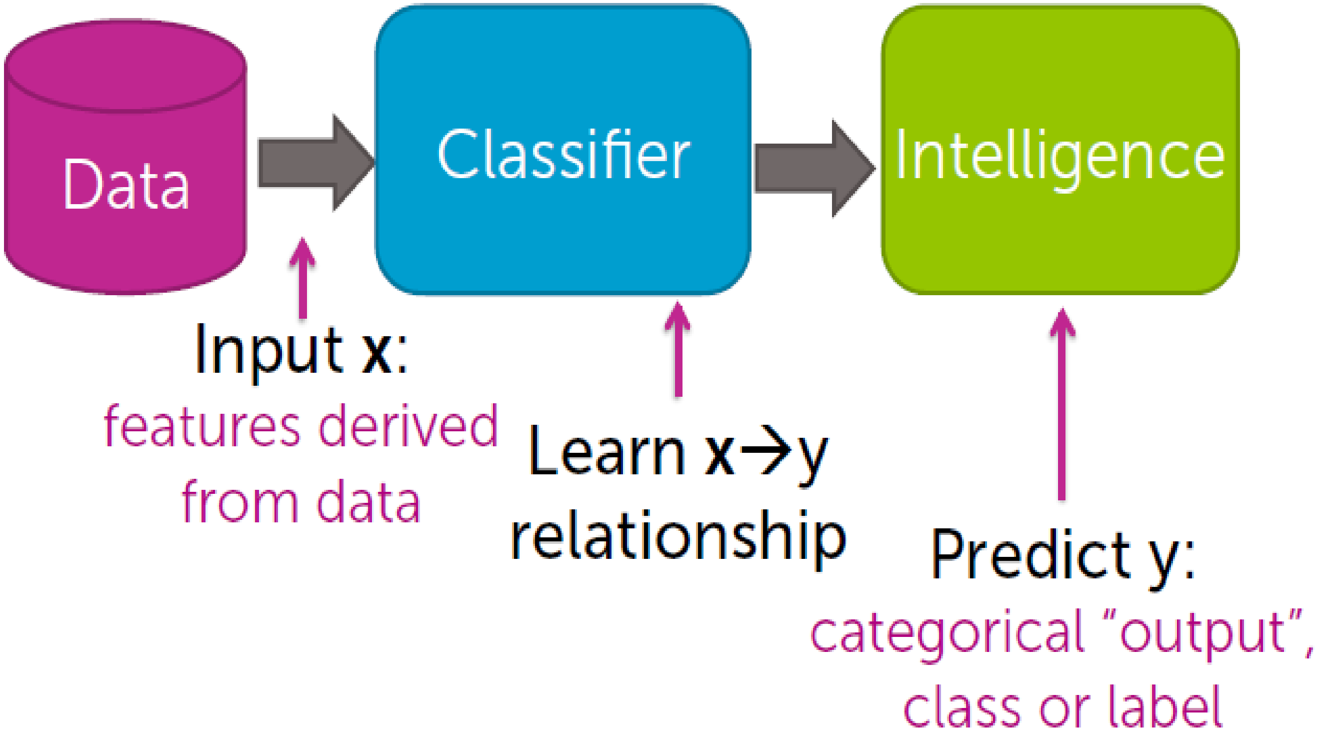
This lecture is
based on course by E. Fox and C. Guestrin, Univ of Washington

14/11/2017

WFAiS UJ, Informatyka Stosowana
II stopień studiów

What is classification?

From features to predictions



Data

Classifier

Intelligence

Input x:

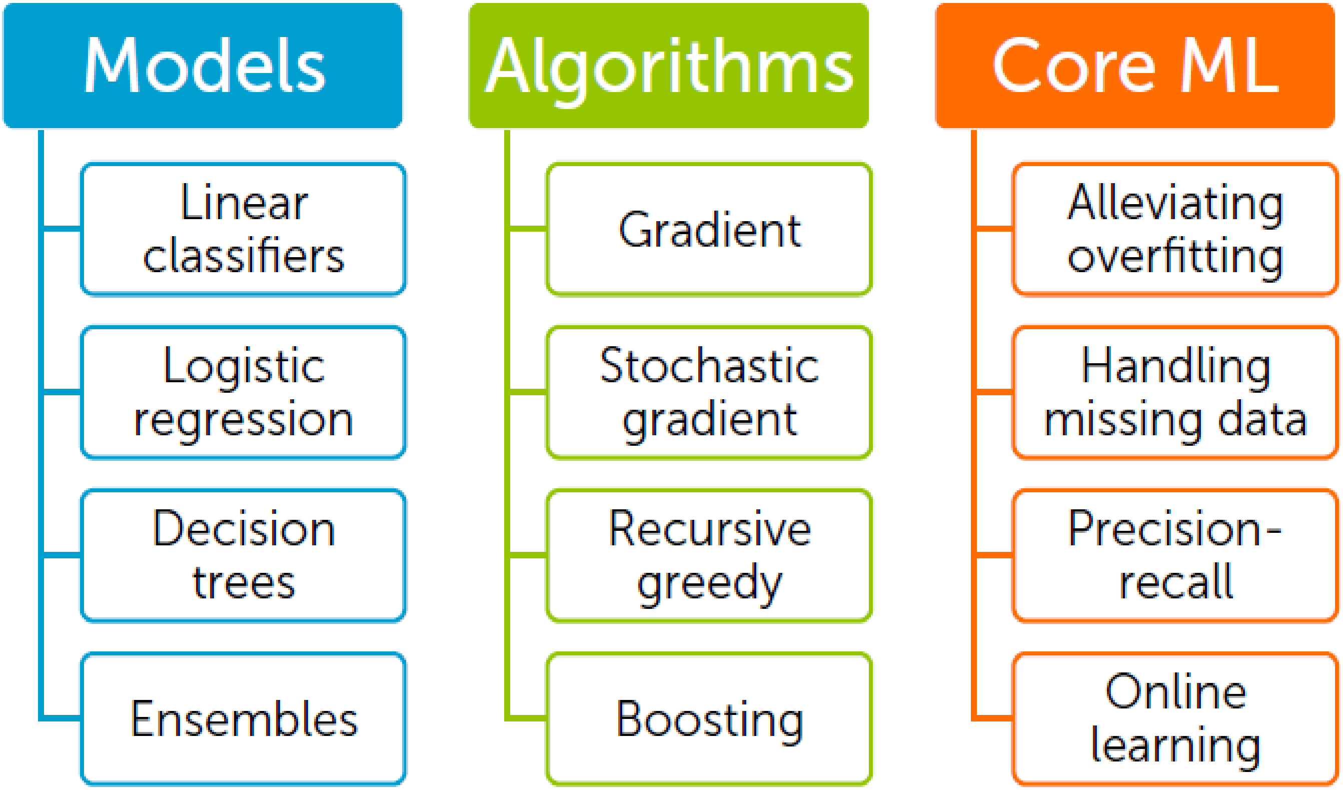
features derived from data

Learn x→y relationship

Predict y:

categorical "output", class or label

Overwiew of content



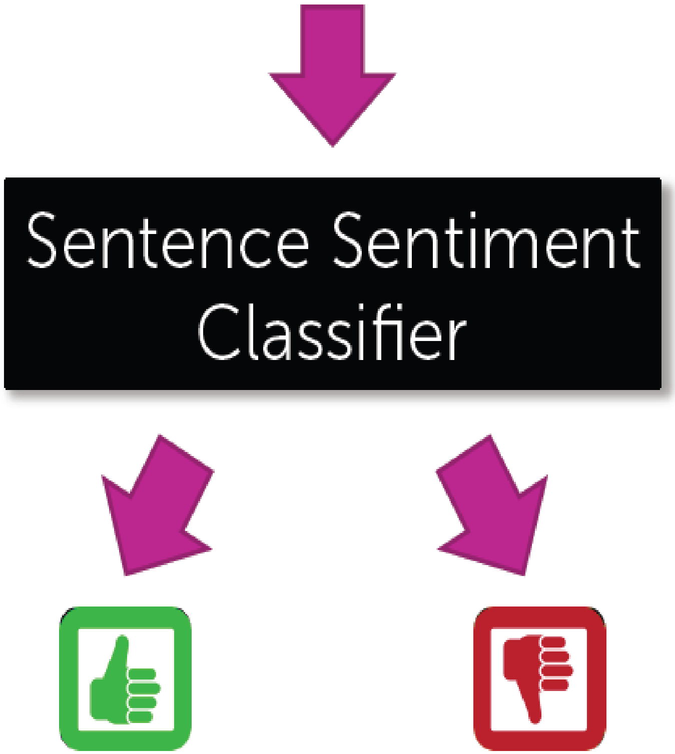
Models
- Linear classifiers
- Logistic regression
- Decision trees
- Ensembles

Algorithms
- Gradient
- Stochastic gradient
- Recursive greedy
- Boosting

Core ML
- Alleviating overfitting
- Handling missing data
- Precision-recall
- Online learning

Sentiment classifier

Input **x**: Easily best sushi in Seattle.

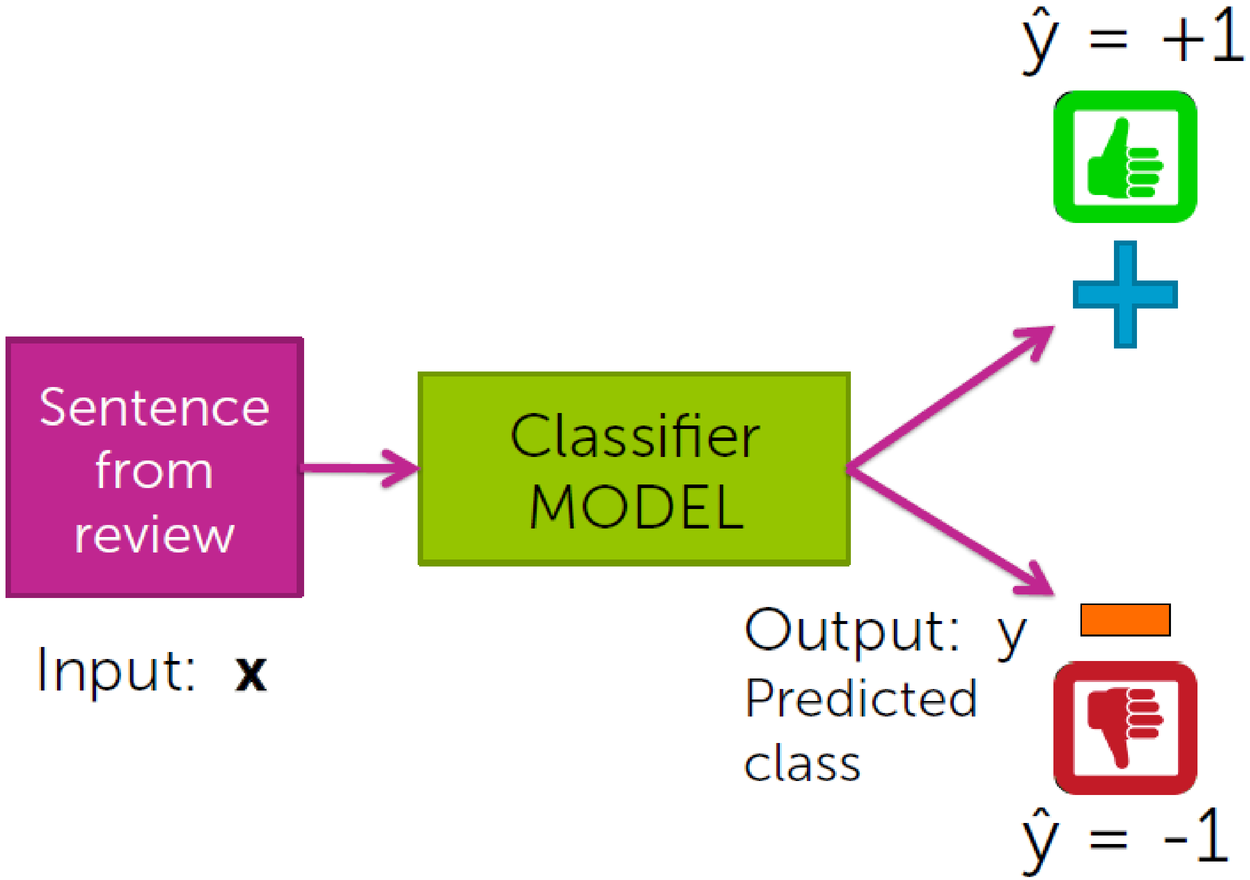


Sentence Sentiment Classifier

Output: y
Sentiment

👍 👎

Classifier

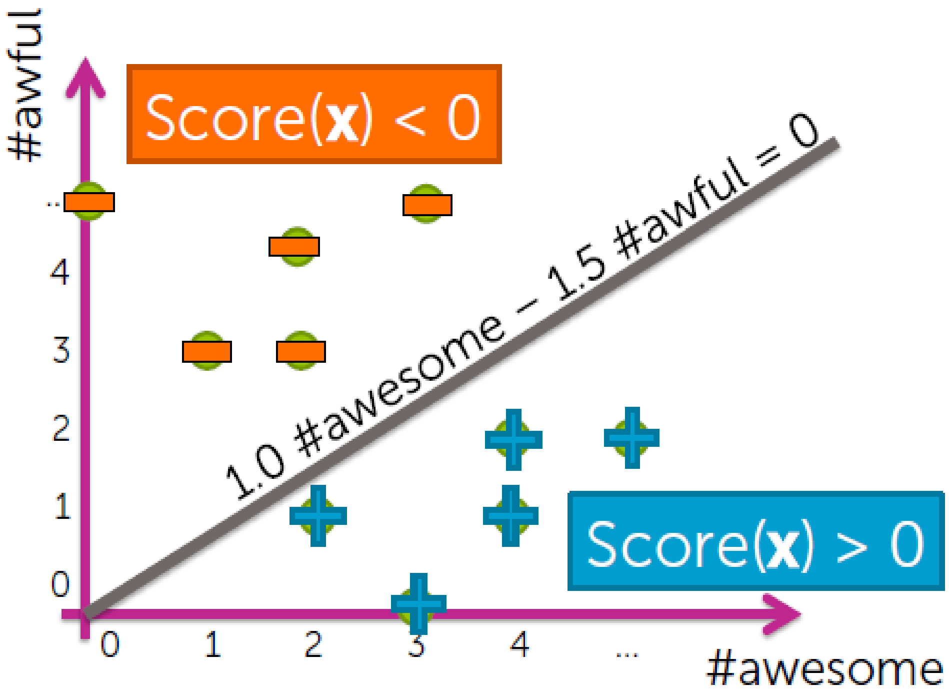


Note: we'll start talking about 2 classes, and address multiclass later

Linear classifiers

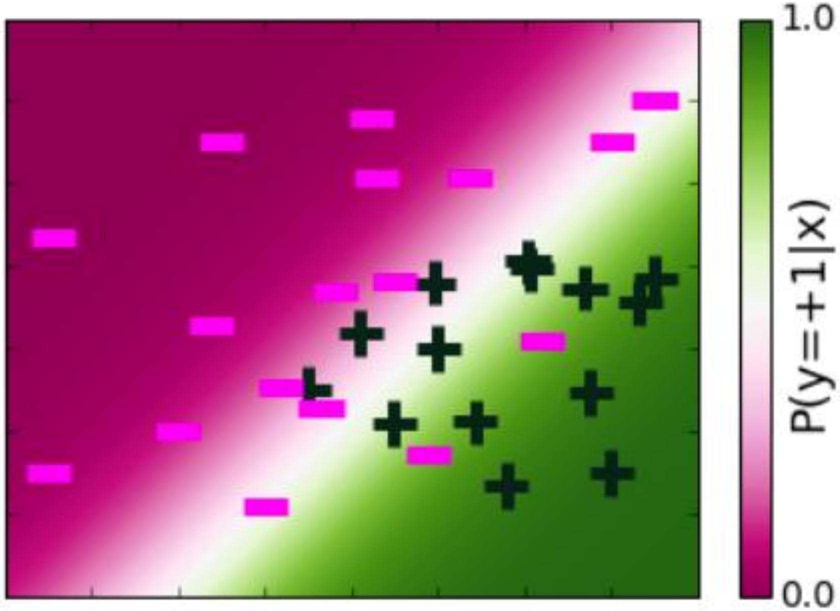
Word	Coefficient
#awesome	1.0
#awful	-1.5

→ Score(x) = 1.0 #awesome – 1.5 #awful



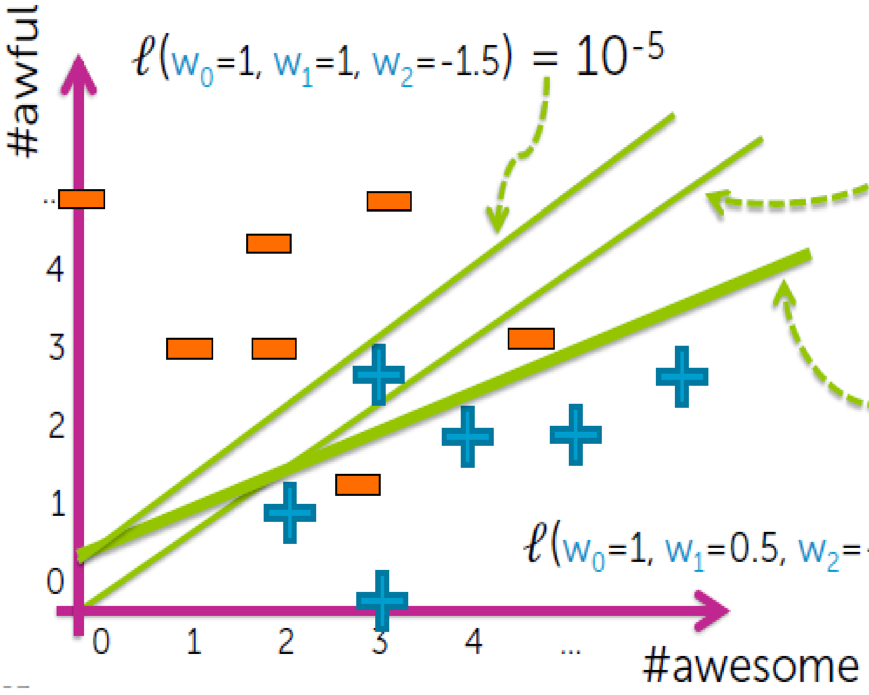
Logistic regression represents probabilities

$$\hat{P}(y=+1|\mathbf{x},\hat{\mathbf{w}}) = \frac{1}{1+e^{-\hat{\mathbf{w}}^T h(\mathbf{x})}}$$



Learning „best" classifier

Maximize likelihood over all possible w_0,w_1,w_2



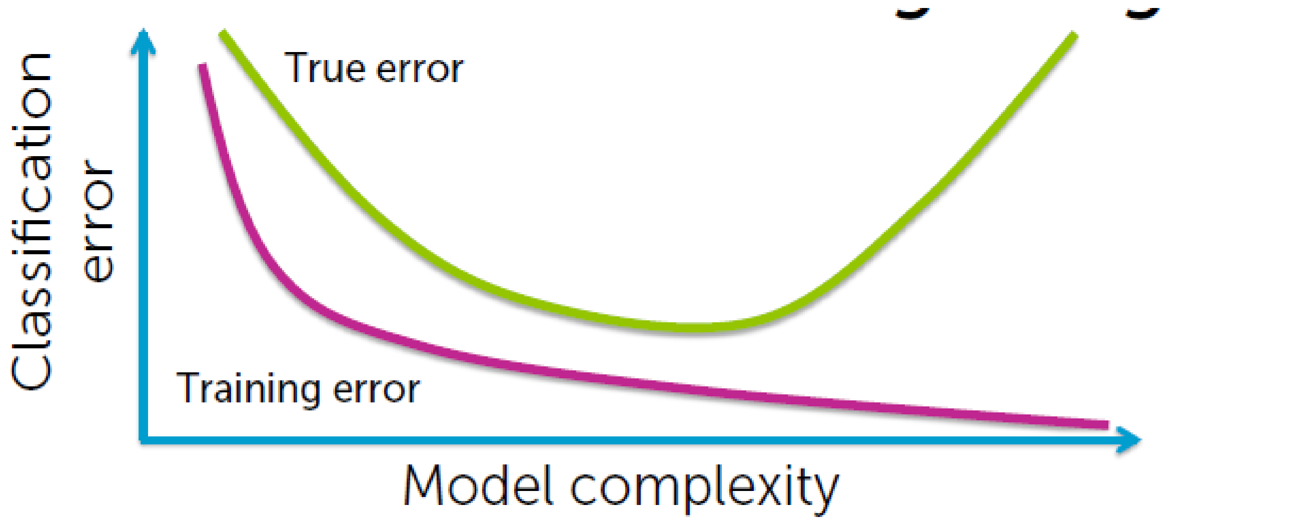
$\ell(w_0=0, w_1=1, w_2=-1.5) = 10^{-6}$

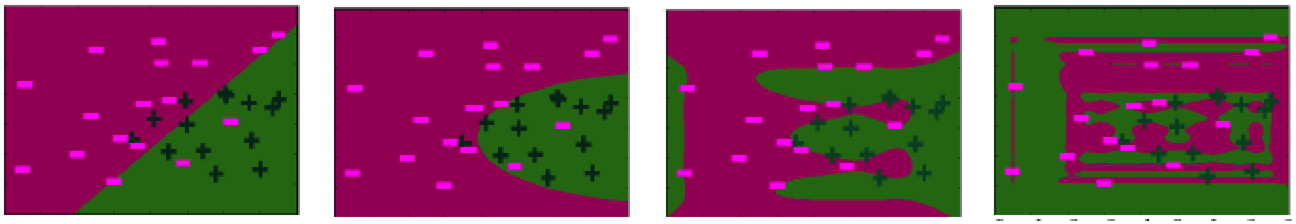
$\ell(w_0=1, w_1=1, w_2=-1.5) = 10^{-5}$

$\ell(w_0=1, w_1=0.5, w_2=-1.5) = 10^{-4}$

Best model with ***gradient ascent****:*
Highest likelihood $\ell(\mathbf{w})$
$\hat{\mathbf{w}} = (w_0=1, w_1=0.5, w_2=-1.5)$

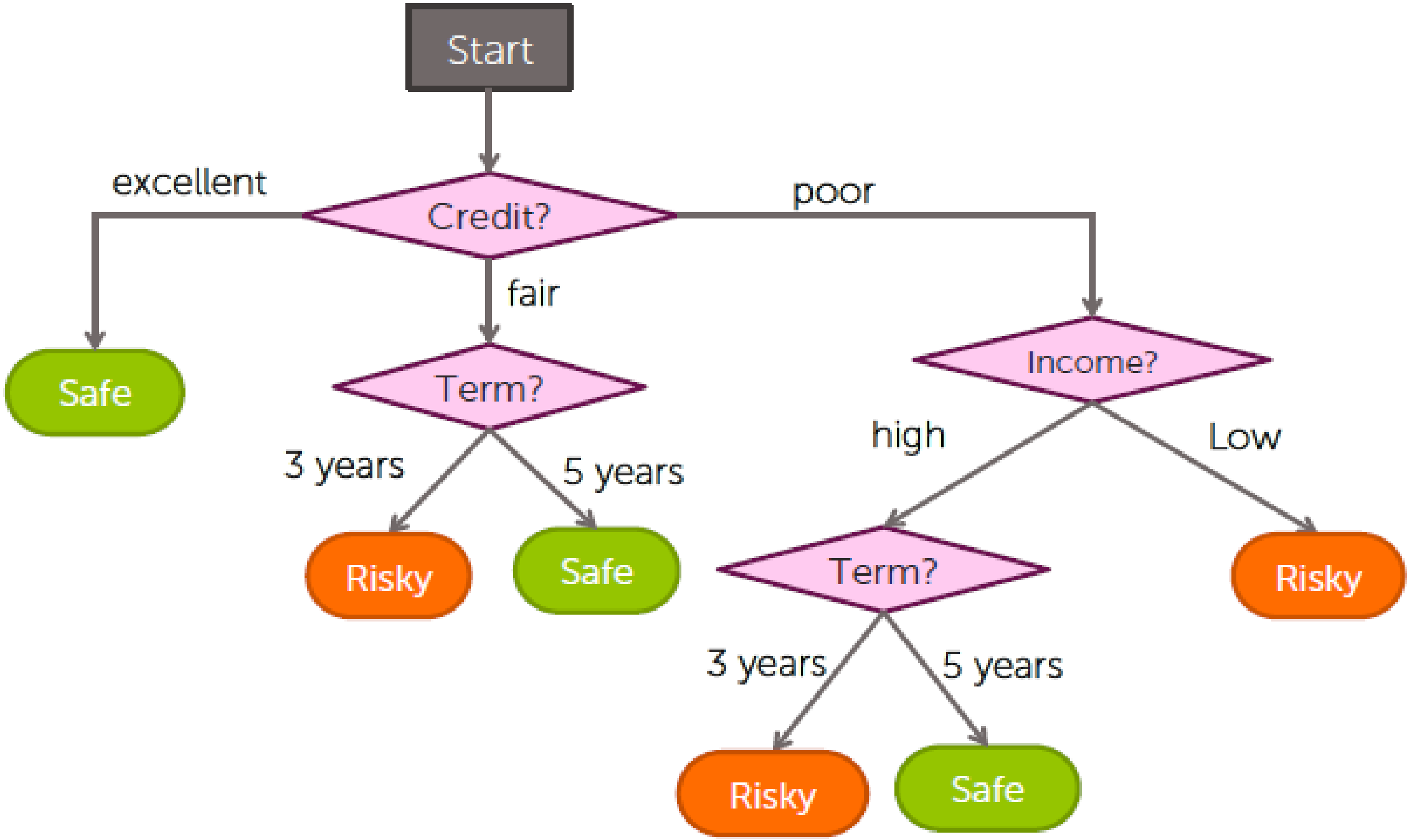
Overfitting & regularisation



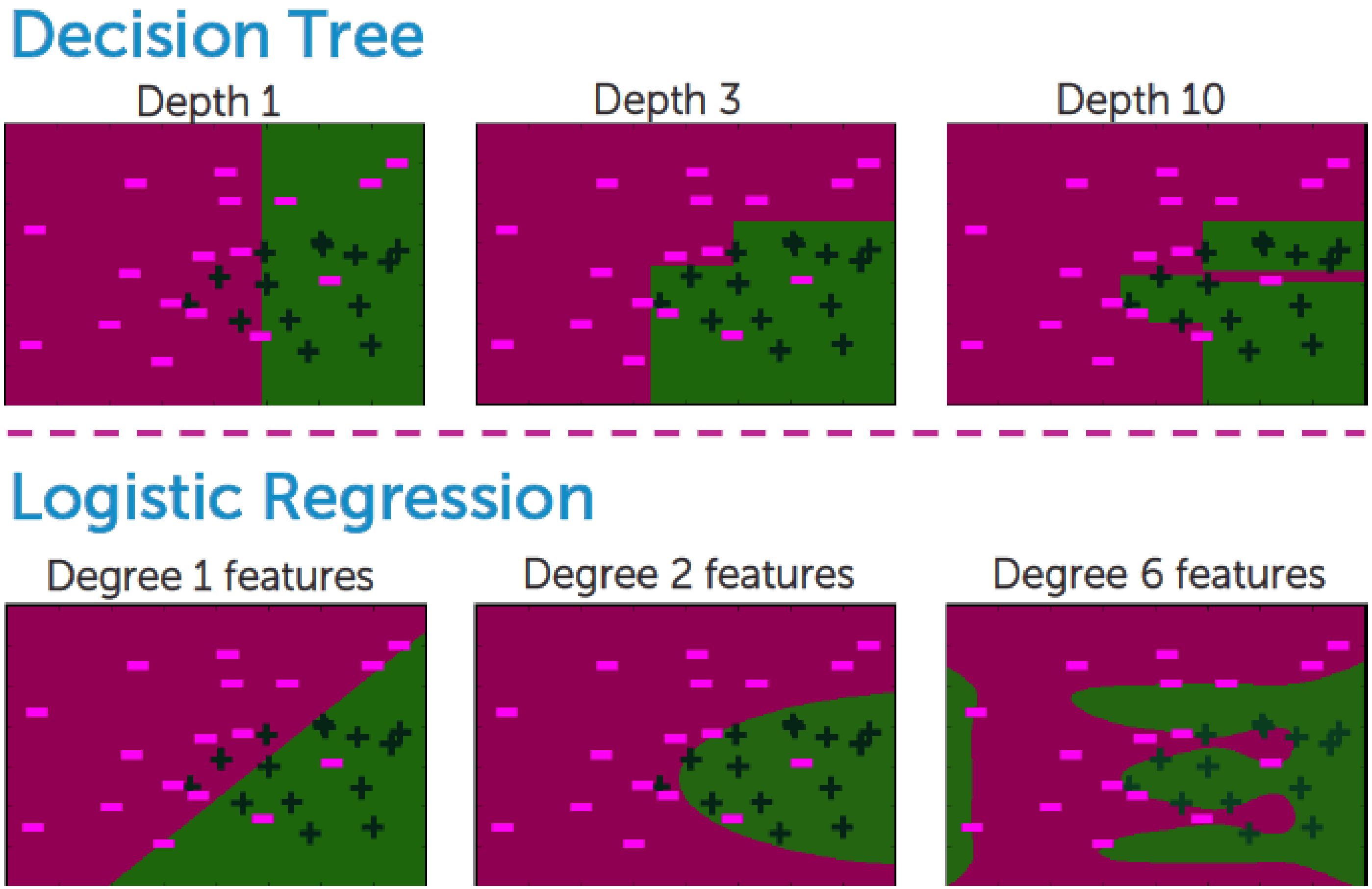


Use regularization penalty to mitigate overfitting $\ell(\mathbf{w}) - \lambda||\mathbf{w}||_2^2$

Decision trees

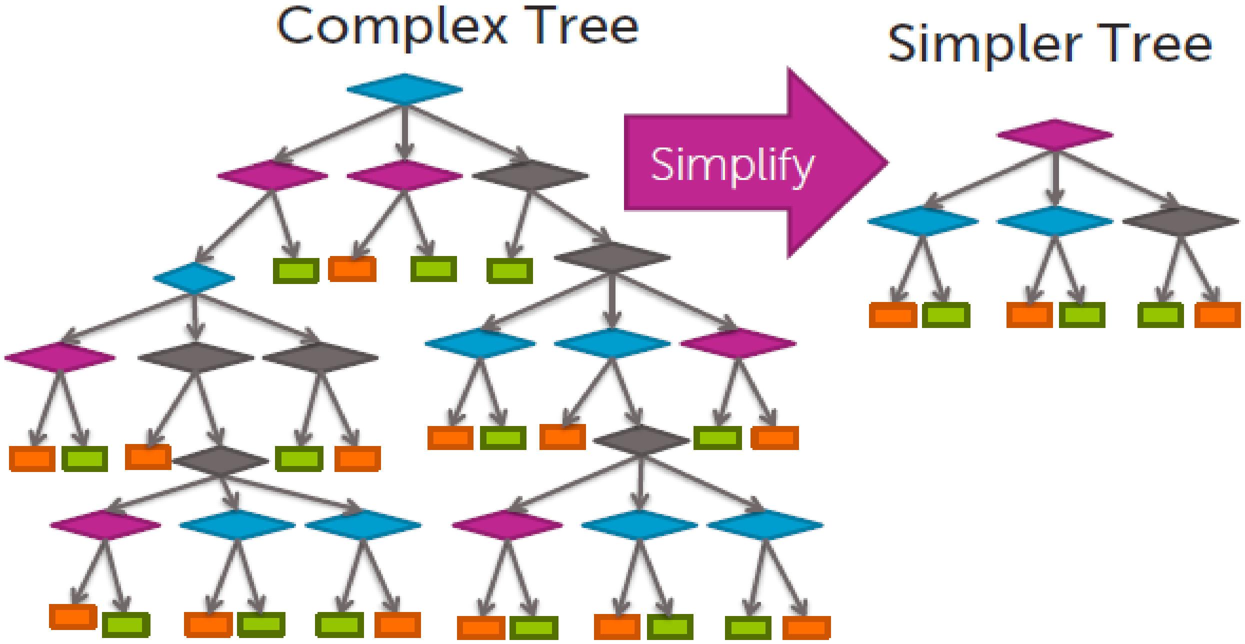


Overfitting & decision trees

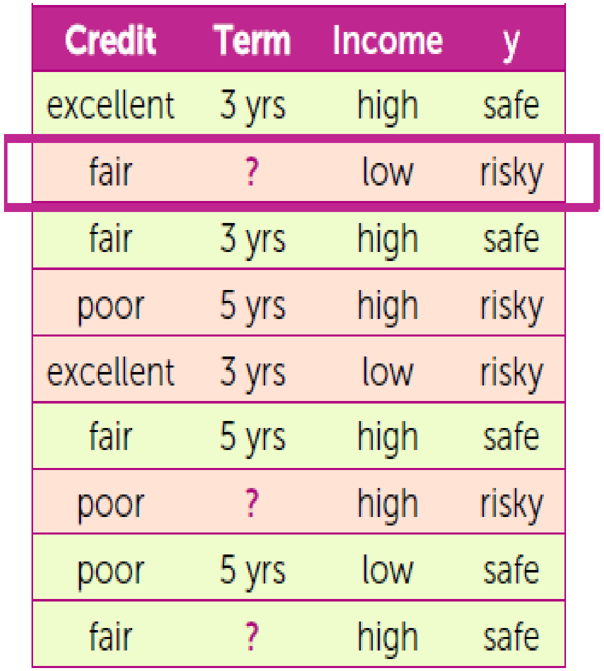


Alleviate overfitting by learning simpler trees

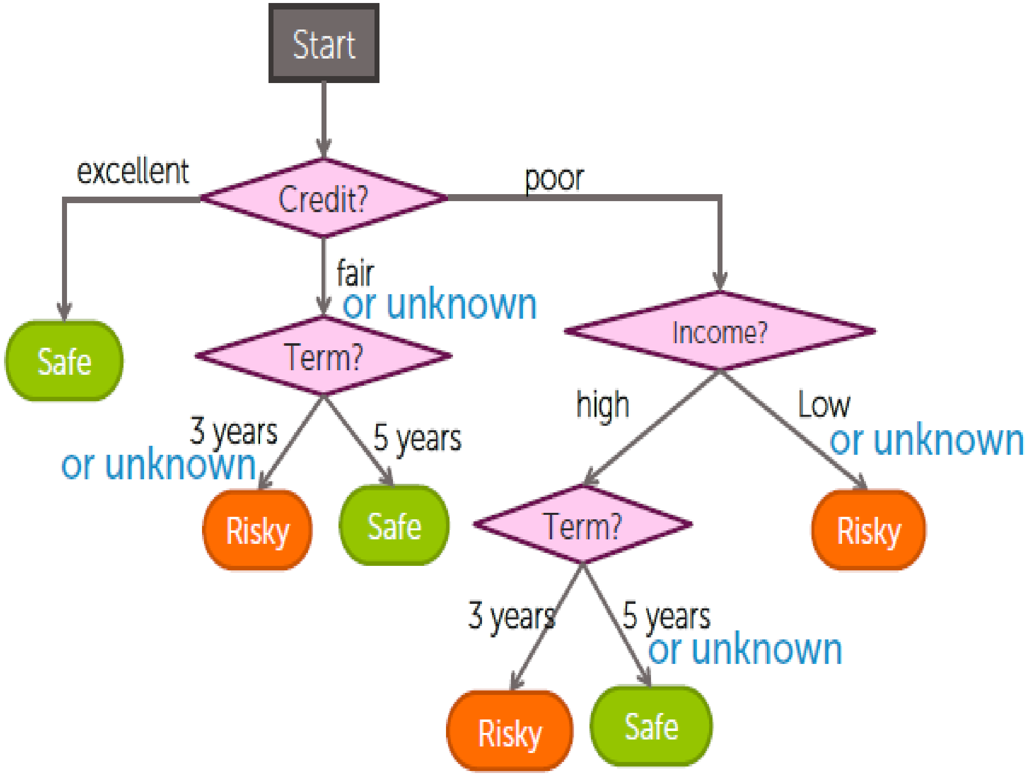
Occam's Razor: *"Among competing hypotheses, the one with fewest assumptions should be selected", William of Occam, 13th Century*



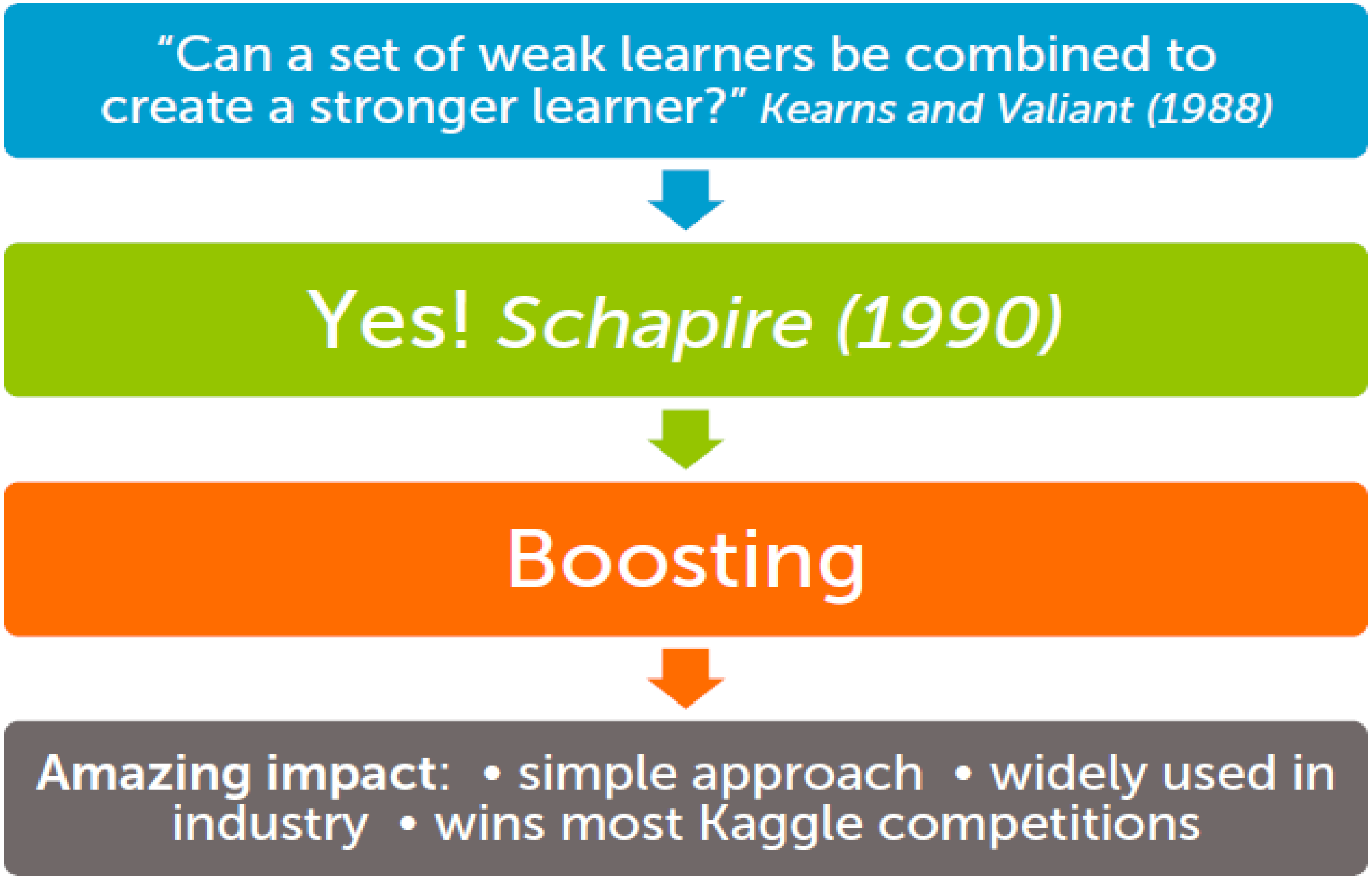
Handling missing data



Credit	Term	Income	y
excellent	3 yrs	high	safe
fair	?	low	risky
fair	3 yrs	high	safe
poor	5 yrs	high	risky
excellent	3 yrs	low	risky
fair	5 yrs	high	safe
poor	?	high	risky
poor	5 yrs	low	safe
fair	?	high	safe



Boosting questions



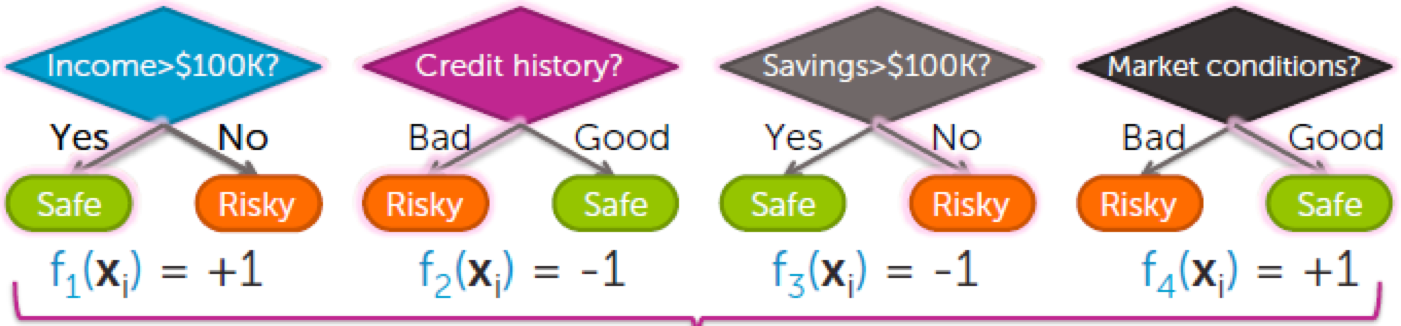
"Can a set of weak learners be combined to create a stronger learner?" *Kearns and Valiant (1988)*

Yes! *Schapire (1990)*

Boosting

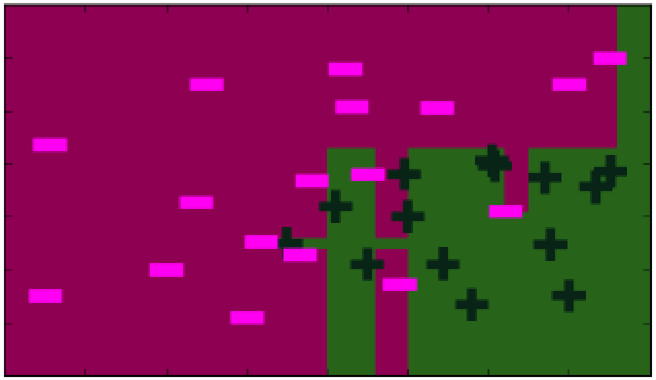
Amazing impact: • simple approach • widely used in industry • wins most Kaggle competitions

Boosting using AdaBoost

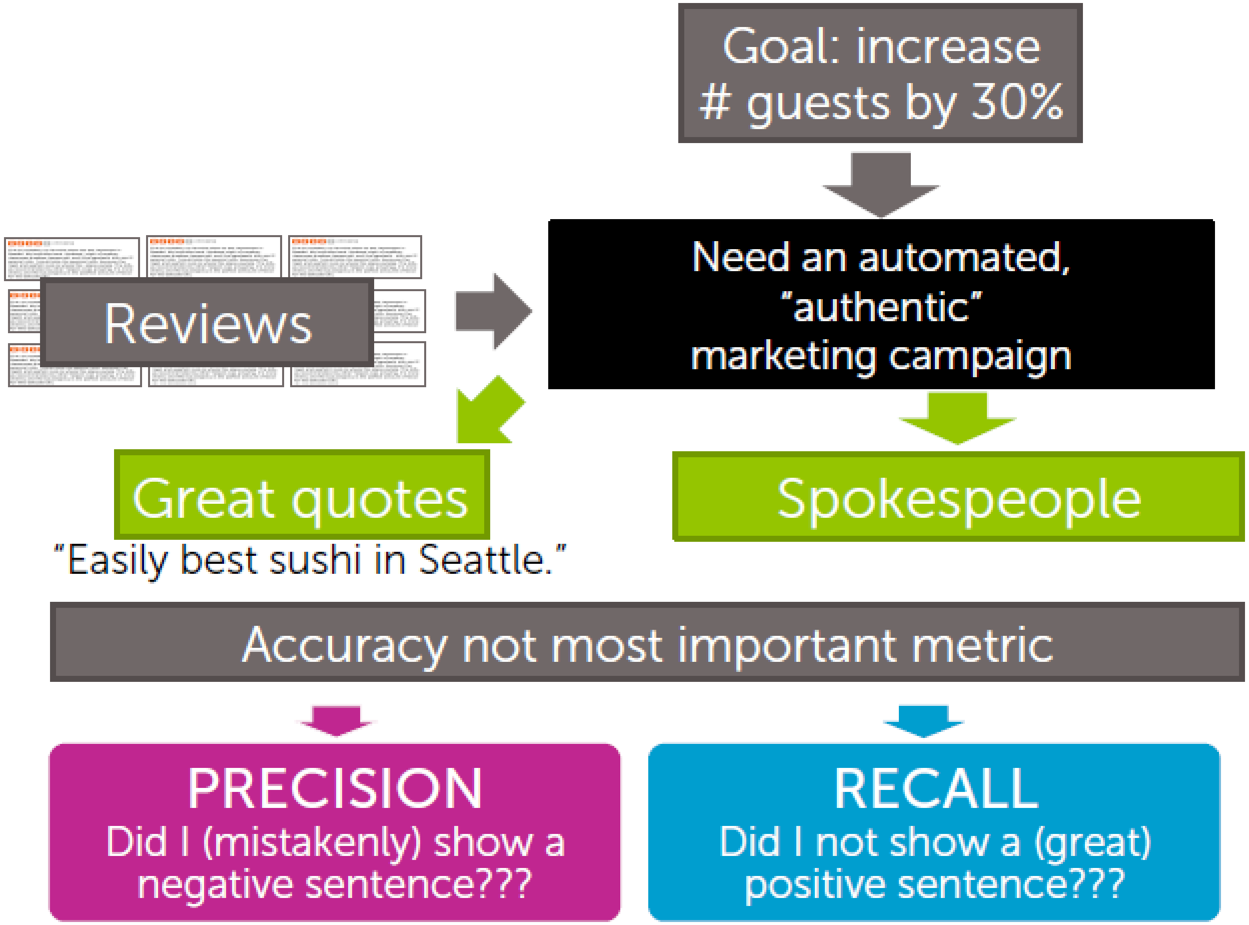


$f_1(\mathbf{x}_i) = +1$ $f_2(\mathbf{x}_i) = -1$ $f_3(\mathbf{x}_i) = -1$ $f_4(\mathbf{x}_i) = +1$

Ensemble: Combine votes from many simple classifiers to learn complex classifiers



Precision - recall



Goal: increase # guests by 30%

Reviews

Need an automated, "authentic" marketing campaign

Great quotes

"Easily best sushi in Seattle."

Spokespeople

Accuracy not most important metric

PRECISION

Did I (mistakenly) show a negative sentence???

RECALL

Did I not show a (great) positive sentence???

Scalling to huge dataset & on-line learning

4.8B webpages

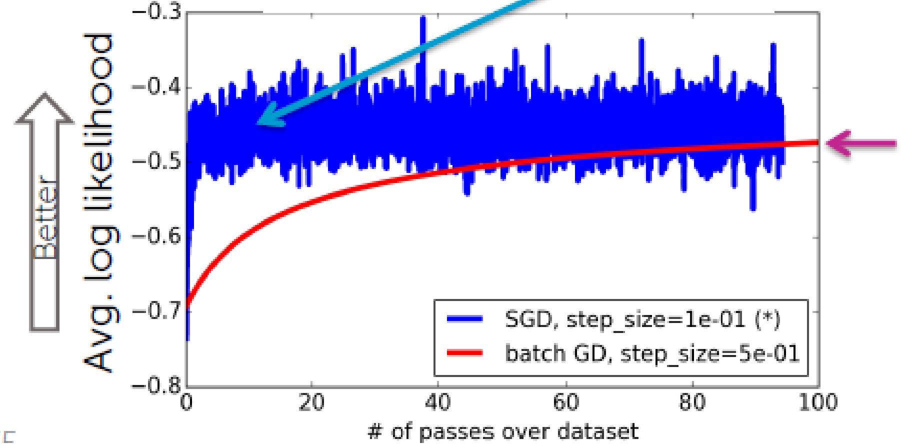
twitter

500M Tweets/day

YouTube

5B views/day

Stochastic gradient: tiny modification to gradient, a lot faster, but annoying in practice



Gradient