Experimental Methodology

Rob Holte
University of Alberta
holte@cs.ualberta.ca

Experiments serve a purpose

- They provide evidence for claims, design them accordingly
- Choose appropriate test datasets, consider using artificial data
- Record measurements directly related to your claims

Establish a need

- Try very simple approaches before complex ones
- Try off-the-shelf approaches before inventing new ones
- Try a wide range of alternatives not just ones most similar to yours

- Make sure comparisons are fair

Explore limitations

- Under what conditions does your system work poorly ? When does it work well ?
- What are the sources of variance ?
  - Eliminate as many as possible
  - Explain the rest

Explore anomalies

Superlinear speedup

IDA* on N processors is more than N times faster than on 1 processor

"…we were surprised to obtain superlinear speedups on average… our first reaction was to assume that our sample size was too small …" - V. Rao & V. Kumar

Superlinear Speedup in Parallel State-Space Search
Technical Report AI88-80, 1988
CS Dept., U of Texas - Austin

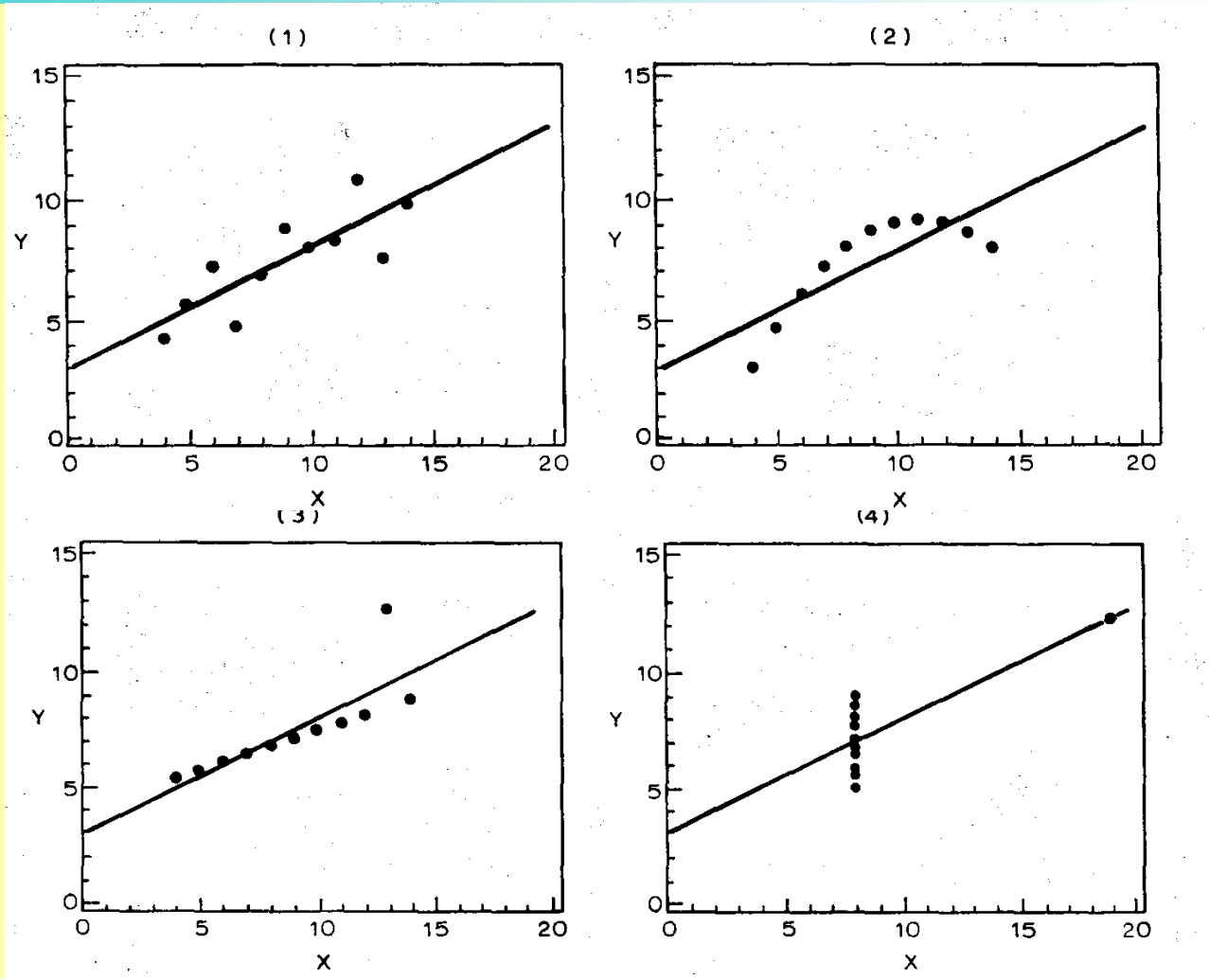
Look at your data

4 x-y datasets, all with the same statistics.
Are they similar ? Are they linear ?

- mean of the x values = 9.0
- mean of the y values = 7.5
- equation of the least-squared regression line is: y = 3 + 0.5x
- sums of squared errors (about the mean) = 110.0
- regression sums of squared errors = 27.5
- residual sums of squared errors (about the regression line) = 13.75
- correlation coefficient = 0.82
- coefficient of determination = 0.67

F.J. Anscombe (1973), "Graphs in Statistical Analysis," *American Statistician,* 27, 17-21

Anscombe datasets plotted

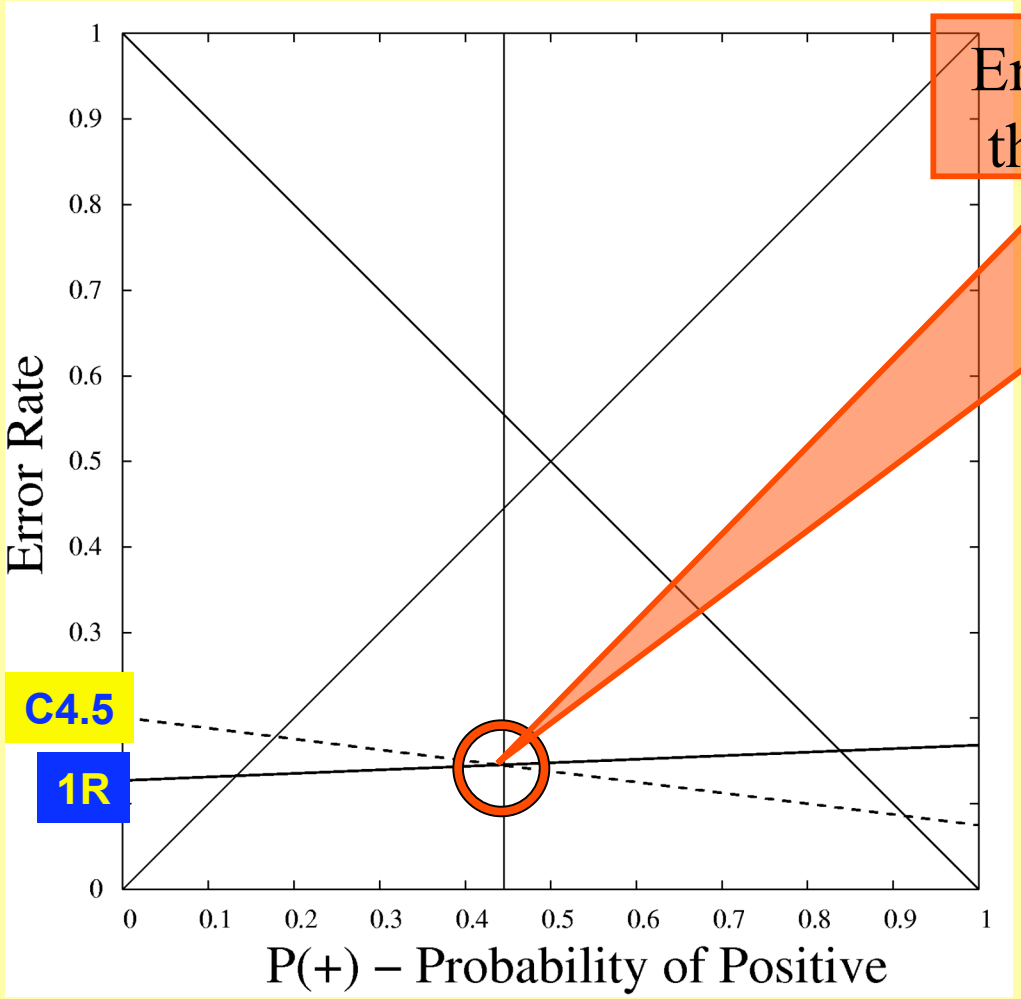


Look at your data, again

- Japanese credit card dataset (UCI)
- Cross-validation error rate is identical for C4.5 and 1R

Is their performance the same ?

Closer analysis reveals…



- ROC curves
- Cost curves
- REC curves
- Learning curves

Test alternative explanations

Combinatorial auction problems

CHC = hill-climbing with a clever new heuristic

Solution Quality (% of optimal)

problem type	CHC
path	98
match	99
sched	96
r75P	83
r90P	90
r90N	89
arb	87

Is CHC better than random HC ?

Percentage of CHC solutions better than random HC solutions

problem type	% better
path	100
match	100
sched	100
r75P	63
r90P	7
r90N	6
arb	20

Avoid "Overtuning"

- Overtuning = using all your data for system development. Final system likely overfits.

- David Lubinksy, PhD thesis
  - Held out ~10 UCI datasets for final testing
- Chris Drummond, PhD thesis
  - Fresh data used for final testing

Too obvious to mention ? (no)

- Debug and test your code thoroughly
- Keep track of parameter settings, versions of program and dataset, etc.