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

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... Error rate is the same only on the dataset class distribution

- 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)

<table>
<thead>
<tr>
<th>problem type</th>
<th>CHC</th>
</tr>
</thead>
<tbody>
<tr>
<td>path</td>
<td>98</td>
</tr>
<tr>
<td>match</td>
<td>99</td>
</tr>
<tr>
<td>sched</td>
<td>96</td>
</tr>
<tr>
<td>r75P</td>
<td>83</td>
</tr>
<tr>
<td>r90P</td>
<td>90</td>
</tr>
<tr>
<td>r90N</td>
<td>89</td>
</tr>
<tr>
<td>arb</td>
<td>87</td>
</tr>
</tbody>
</table>
Is CHC better than random HC?

Percentage of CHC solutions better than random HC solutions

<table>
<thead>
<tr>
<th>problem type</th>
<th>% better</th>
</tr>
</thead>
<tbody>
<tr>
<td>path</td>
<td>100</td>
</tr>
<tr>
<td>match</td>
<td>100</td>
</tr>
<tr>
<td>sched</td>
<td>100</td>
</tr>
<tr>
<td>r75P</td>
<td>63</td>
</tr>
<tr>
<td>r90P</td>
<td>7</td>
</tr>
<tr>
<td>r90N</td>
<td>6</td>
</tr>
<tr>
<td>arb</td>
<td>20</td>
</tr>
</tbody>
</table>
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.