DRR XX

• Slides used in Panel Discussion on Evaluation (Thursday, Feb. 7th, 2013)
• Slides prepared by R. Zanibbi, B. Coüasnon, L. Likforman-Sulem and D. Lopresti
Competitions – Value?

Large Number of Competitions at ICDAR 2013

- Is this a good use of our energy for our community?
- **Problem Definition:** data collection expensive; data sets tend to be non-representative (and small)
- Many have a small number of participants
- Requires significant effort to solve, but importance/generalization to “real” problems unclear
- **Solving the data set vs. solving the problem?**

Is there an *alternative* to competitions and/or benchmarks?
OpenHart evaluation

• huge database: 43 K pages

• writing factors + noise
  – underlined paper
  – clean/noisy images

• large vocabulary: 50K

→ a challenge!
  (computer ressources, modelling....)
Performance Evaluation

Why is it so important to get it right?
- Measure progress ⇒ prevent wasted effort.
- Scientific respectability.

Why are we so bad at it?
- Everyone believes his/her own problem is unique.
- Disconnect between problems and real-world tasks.
- Desperate need to generate publications.
- As a community, we are too polite.

Some statements here that should have provoked fights*:
- Ray Smith: “Test data ≈ training data proves nothing.”
- Gady Agam: “DIBCO at ICDAR 2011 was bogus.”

* I take full responsibility for any errors in interpretation or generalization.
More Generally...

• Should we be using statistical hypothesis tests (chi-squared, ANOVA) to compare systems?
TC-11 Website

Excellent resource for document datasets and tools.

http://www.iapr-tc11.org