No more excuse for missing significance tests and confidence intervals!

**MT-ComparEval**
- Try it now (WMT 2014-2016 results) http://wmt.ufal.cz
- Upload and analyze your translations http://mt-compareval.ufal.cz
- Install it (and report issues or contribute): https://github.com/choko/MT-ComparEval

- web-based tool for MT developers
- check progress of a system over time or compare several MT systems
- focus on analyzing system differences
- can be integrated into your workflow (import translations: disk/git/REST-API)

1. Select an "Experiment", e.g. English-Czech WMT15
2. See BLEU, F-measure,...
   Select two "Tasks", i.e. system translations
3. Sentences pane with various diffs highlighted
4. Statistics pane (paired bootstrap resampling,...)
5. Sentences sorted by diffBLEU
6. Confirmed n-grams pane

- confirmed n-gram = occurs in the reference (light yellow and blue highlight)
- improving n-gram = confirmed n-gram occurring in only one of the systems (dark yellow and blue highlight)
- worsening n-gram = unconfirmed, occurring in only one of the systems (red highlighting)
- diff (LCS underlined in green)

- See the sentences with biggest improvement/worsening
- See the 1-grams...4-grams with biggest improvement/worsening
- Hints for improving the systems

MQM (Multidimensional Quality Metrics)
explicit error annotation in translate5 software

- Detailed error analysis when needed
- Bridge to quality assurance in industry
- Enables "profiling" of systems/languages, etc.