# *Controversial (?) Position Statements*

Kevin Bowyer

University of Notre Dame

Two sources of “found data” are:

(1) scrubbed from the web

(2) logged in operational scenario

They have different essential problems, relative to “laboratory” or “planned” data collections.

**If image provenance is unknown, and there is no meta-data, you can’t be certain what problem you have solved.**

**E.g., did you pair images or by identity or by some other property: photoshop / no photoshop, compression / no compression?**

# The operational contradiction –

The better run the operational scenario (more user friendly, faster, …), the more problematic it is to use the data for research.

E.g., 1-to-first rather than 1-to-N is faster for the user, but seeds ground truth errors for research.

**From a researcher’s perspective, a planned dataset is preferable to a found dataset.**