Is there a Quick Fix for Open-ended Questions? A Comparison of Qualitative Analysis Techniques

Casey Langer Tesfaye

American Institute of Physics
Georgetown University
Free Range Research

Qualitative Coding

- Cannot easily quantify qualitative results
  - Numbers not representative
    - Few responses per domain
    - What does a nonmention mean?
  - Descriptive value doesn't seem worth the expense
    - Time consuming
    - Different training or software
    - Not often done
- Can undermine more reliable quantitative results
  - Stories are compelling

Alternative Method: Text Analysis

- Range
  - Automated
  - Assisted
  - Manual
- Tools
  - Software packages or inhouse programming
  - Web interfaces, service providers

What do they do?

- Patterning in Language
  - Aboutness
  - How we communicate

Patterns in How we Communicate

- Fact vs Opinion
- Agency vs Passivity
- Specialized Vocabulary
- Sentiment
- Temporality
- Repetition

The question is:

How do we use these patterns?

Units of Analysis

- Respondents

Units of Analysis

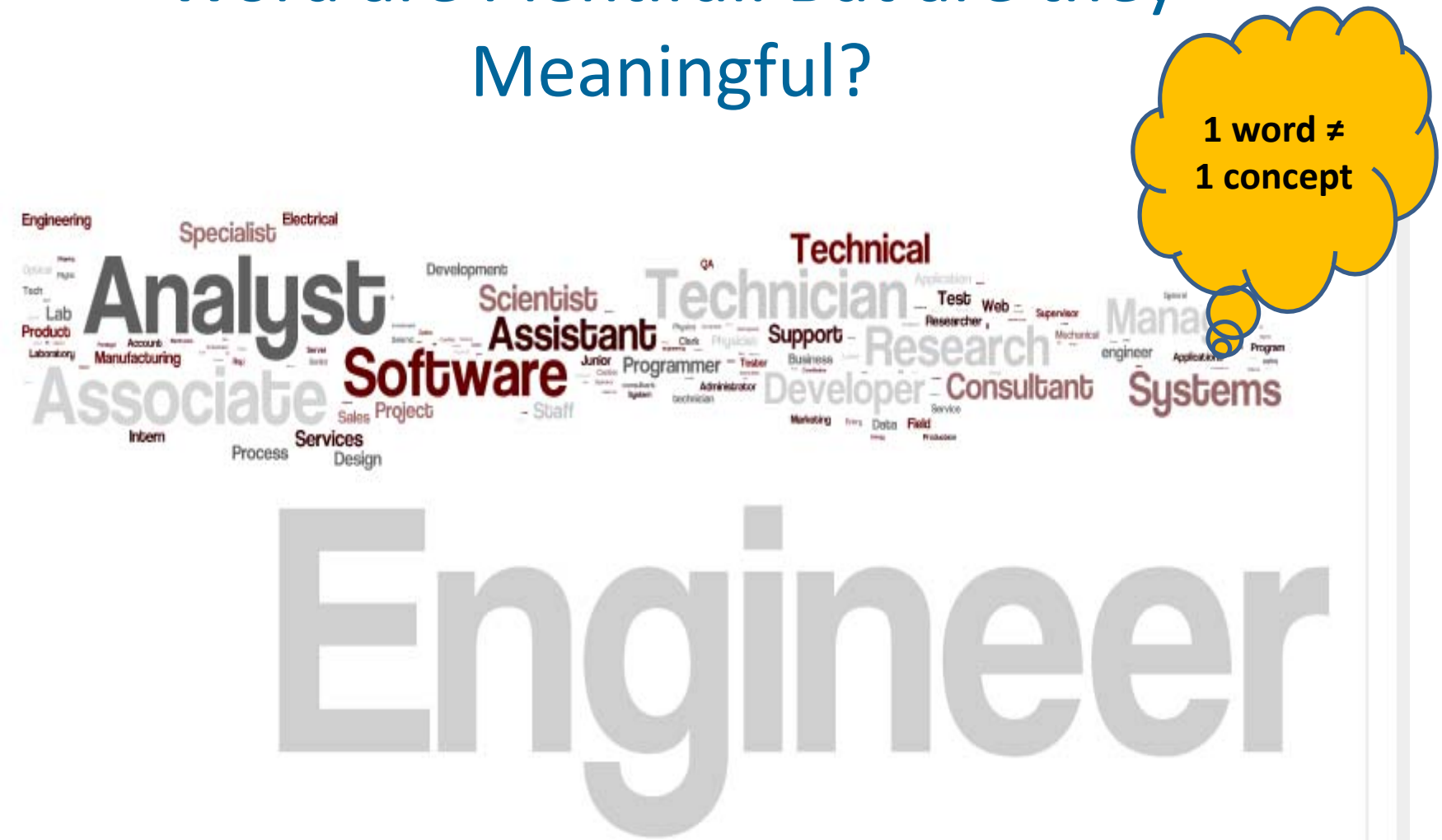
- ~~Respondents~~
- **Bag of Words**
  - Sentence
  - Word groups or types
    - Phrase
    - Common combinations
    - Concepts
  - Words

Common Analyses

- Frequencies (including bigrams, etc.)
  - List or cloud
- Concept extraction
- Sentiment analysis

- … More to come …

Word are Plentiful. But are they Meaningful?



Language is complicated

- Concepts have more than one word
- Words have multiple meanings
  - "That movie was sick!"
  - "Our health care system is sick."
- Indirect references
  - "That one is better than this."
  - "I would follow this movie to the moon and back."
  - "The one with the keyboard is more fun."

Different Strategy, different controls

- Precision
  - % of correct hits
- Recall
  - % of target hit

Methods

- Descriptive
- Comparative
  - Over time
  - Other groupings
    - Accessibility
      - Link to dataset (unit of analysis problem)
      - Derived from text (potentially complicated)

How has the No Child Left Behind Act affected the physics program or your classes at your school?

- Past:
  - "We have not seen much change"
  - "I don't know what changes to attribute to NCLB"
- Present:
  - "Not many students take physics because it's not required."
  - "Students not very enthusiastic, because it's required"
- Future:
  - "not at all yet--in two years all students will be required to take physics"

Methods

- Automated
- Assisted
- Manual

- Varying degrees of customization

Tools

- Software packages or programming inhouse
  - Training/ Learning curve
  - Repeatability, tweakability, documentation, transparency

- Web interfaces (API's), Service providers
  - Field constantly evolving
    - "Bake - off mentality"
  - Hot commodity
    - Proprietary knowledge, secret formulas, less transparency
    - Black box

Analysis Potential

- Wider array of data sources
  - Social media
  - Journals
  - Interview or Focus Group transcripts

In Summary

- Language is patterned grammatically, not topically
- To fully take advantage of text data, we need to think carefully about the meaningful patterns in our data
- Isolating these patterns requires a complicated balance between precision and recall
- The field of text analysis is fast growing and changing and may broadly influence our field in the future

Further Resources

- Free Range Research blog: http://www.freerangeresearch.com
- King, Gary, and Will Lowe. *An Automated Information Extraction Tool For International Conflict Data with Performance as Good as Human Coders: A Rare Events Evaluation Design.* International Organization 57 (2003): 617-642. copy at http://j.mp/lxhNuB
- NLTK: http://www.nltk.org/
- Pang, Bo, and Lee, Lillian (2008). *Opinion Mining and Sentiment Analysis.* Foundations and Trends in Information Retrieval. Vol 2(1-2), pp. 1-135
- My paper: http://www.mapor.org/2011_papers/6b5Tesfaye.pdf
  - Final analysis available: http://www.aip.org/statistics

Some providers

- Packages
  - Provalis QDA miner and wordstat (http://www.provalisresearch.com/simstat/simstw.html)
  - Textpack (Gesis) (manual and assisted, develop & validate dictionary)
  - Language Logic Ascribe (automated (sentiment, concepts), hybrid, manual)
  - nVivo
  - Atlas (mostly or entirely manual coding necessary)
  - Excel, word, etc. (entirely manual)
  - SPSS
    - Dictionary (standard or custom)
    - Sentiment tagged
    - Steep learning curve
- API's and Service Providers
  - Crimson Hexagon (assisted)
  - Open Amplify (manual)
  - Radian6 (uses some Open Amplify technology)
  - Vigiglobe (integrates cloud functionality)
  - Revelation (proprietary SNS)
- Programming Languages
  - R (see: http://f.cl.ly/items/12070m0e1y1g0S3h1k1O/paper.pdf)
  - Python (NLTK available)

To follow up with me:

Casey Langer Tesfaye

clanger@aip.org

http://www.freerangeresearch.com

Twitter: FreeRangeRsch