Semantic Analysis

Statistical Analysis

WAR OF THE WORDS

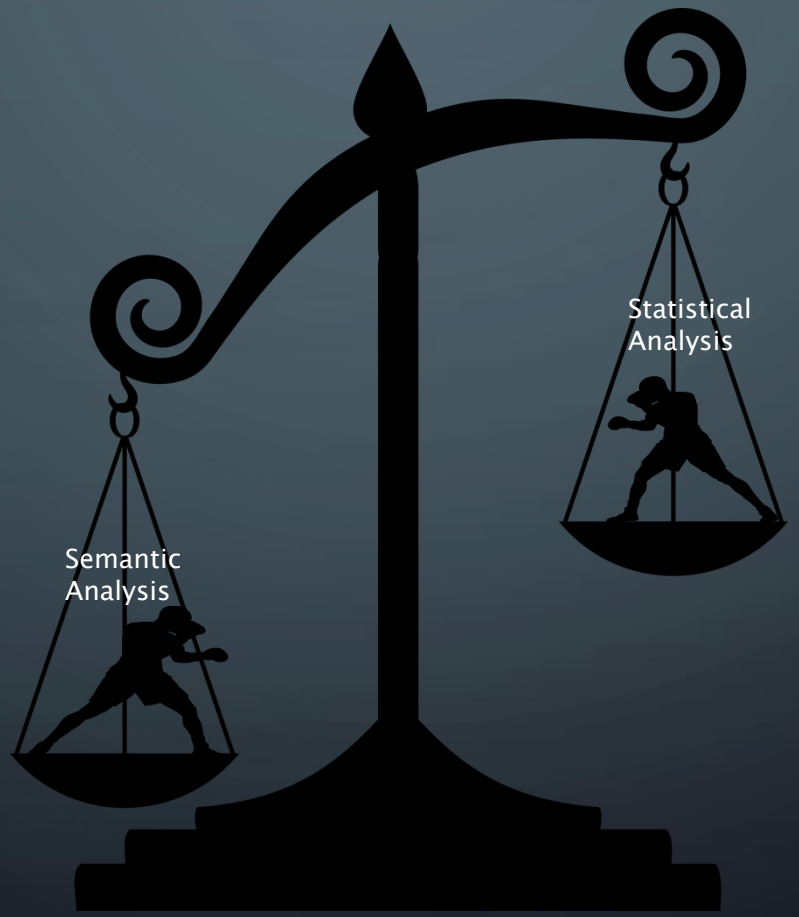
SYSTEMS I HAVE WORKED WITH

- MicroTac – Direct Transfer (semantic)
- Globalink MT – Rules based transfer system (semantic)
- IBM WebSphere Translation Server – Rules based transfer system (semantic)
- IBM WebSphere Voice Server (statistical)
- LinguaSys – Interlingua (semantic)
- Cognitive Code SILVIA (statistical)

OTHER WELL KNOWN SYSTEMS

- Google Translate
- AT&T (now Interactions)
- Nuance
- Amazon (Yap)
- Every Speech Reco engine
- Almost every MT engine

THE SCALES ARE TILTED, BUT ARE THEY RIGHT?



SILVIA
Intelligence On Command

OVERLY SIMPLISTIC DEFINITIONS

SEMANTIC

Let's use our knowledge of each language's semantic and syntactic rules to figure out the semantic meaning of each word, usually at the sentence level, based on the words around it, identify parts of speech, and the relationships between all of the words to detect correct meaning.

STATISTICAL

Let's take thousands of sentences or utterances, tag them with meta-data, and then use proprietary algorithms, such as HMMs, to statistically decide on the highest probability of correct meaning.

SEMANTIC ANALYSIS

- PROS
  - Work done by Computational Linguists, not data scientists
  - Easier to manipulate and correct mistakes
  - Is not dependent on finding thousands of examples
- CONS
  - Always requires Computational Linguist level people
  - Requires constant maintenance to keep up with new words/phrases/colloquialisms
  - Doesn't take advantage of world knowledge

STATISTICAL ANALYSIS

- PROS
  - Takes advantage of world knowledge
  - Does not require hard to find linguistic skills
  - Hands off work – heuristic
- CONS
  - Doesn't seem to work on Asian Languages
  - Prone to hacking with "purposeful mistakes"
  - Requires very expensive skill set
  - Won't work for low data languages

- IF "more data is good data", then why isn't google translate much better?

A BROADER ARRAY OF DATA ANALYSIS TOOLS

- 2012-2013 Google Flu Trends missed by 100% from CDC actual numbers
- Missing was CDC reporting data from the field
- With changes, still missed in 2014 by 30%
- A “Mashup” of the two streams would be best – conclusion

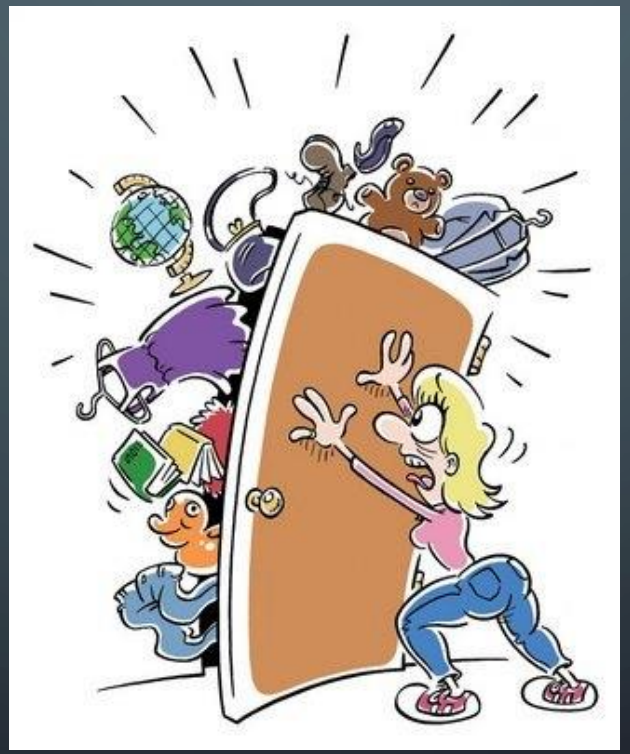
TYPICAL SEMANTIC ISSUES

Republicans turned off by size of Obama's package

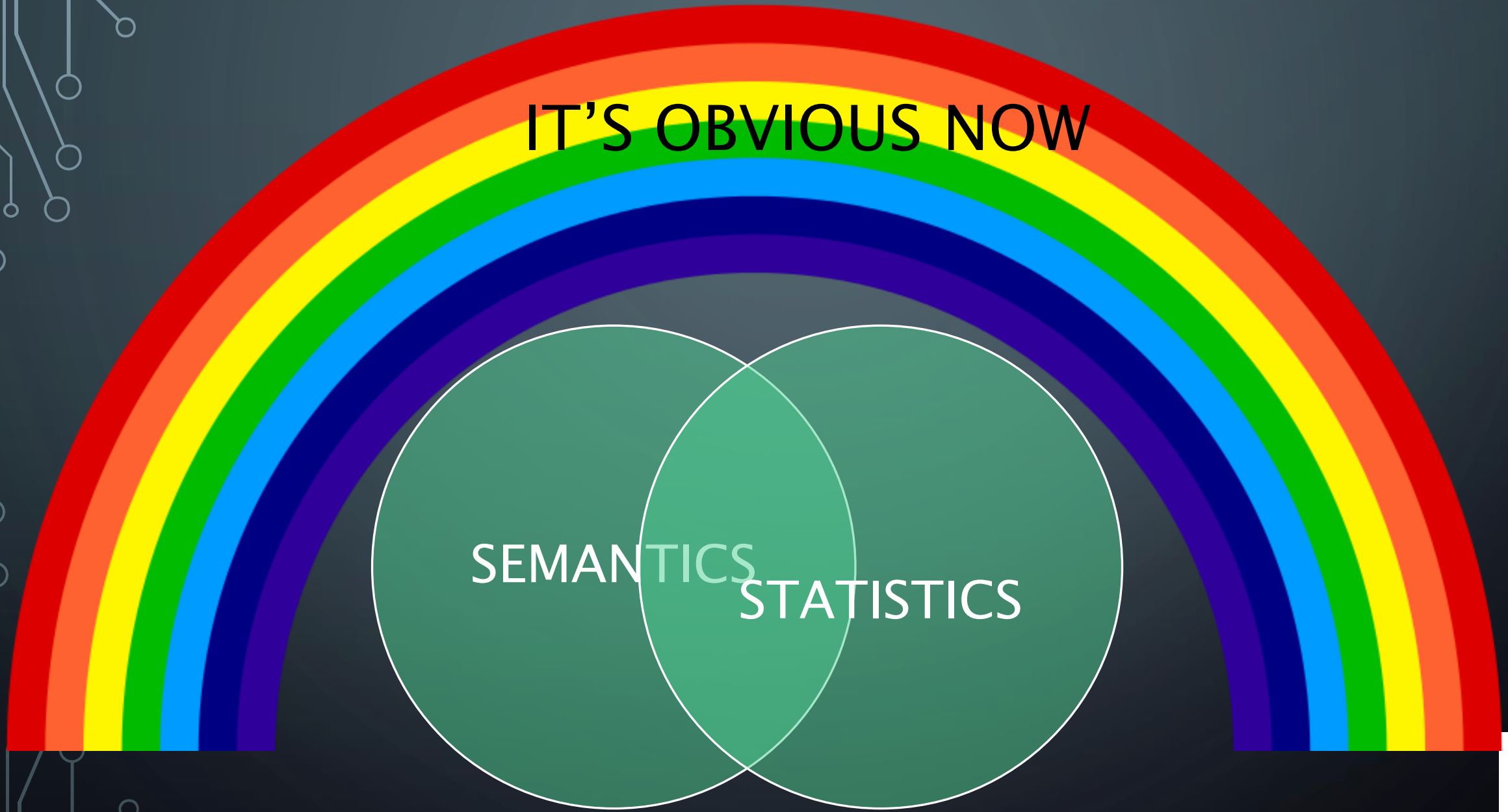
Tom Brandt

- Semantics
- “Children make nutritious snacks.“
- The thief was sentenced to six months in the violin case.”
- “Police shoot man with crossbow”

HISTORICALLY



IT'S OBVIOUS NOW



SILVIA
Intelligence On Command