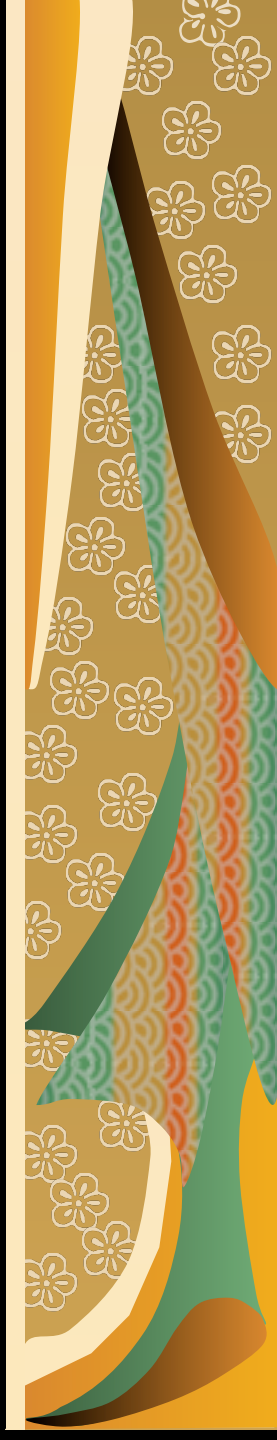
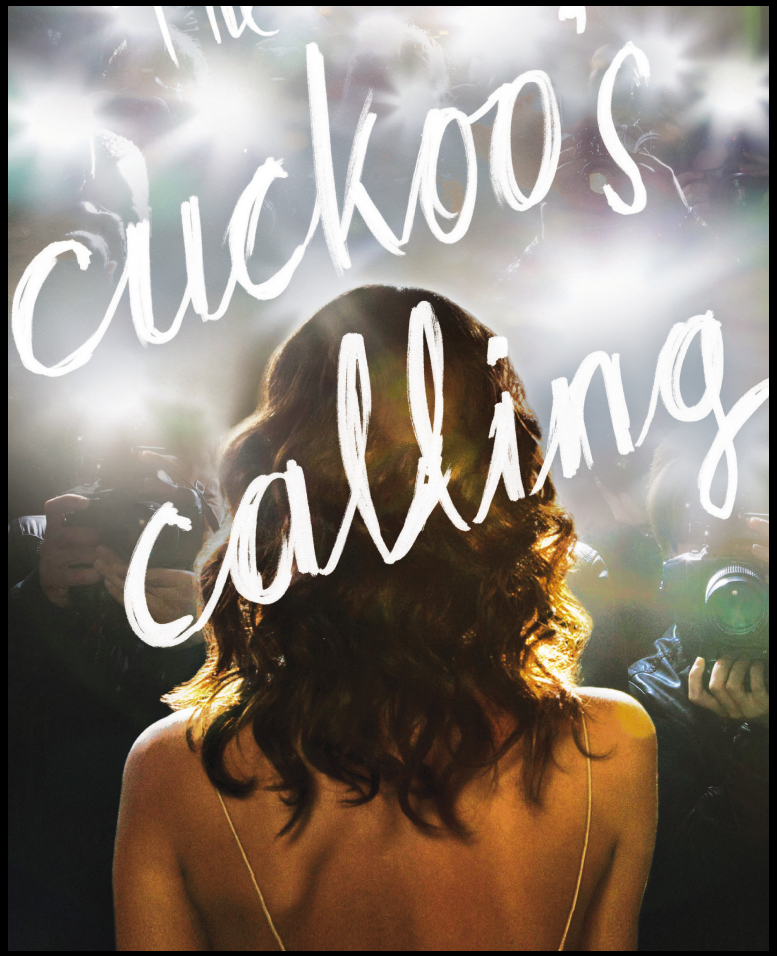
Merging Stylometry and Markup

Patrick Juola
Evaluating Variations in Language Lab
Duquesne University



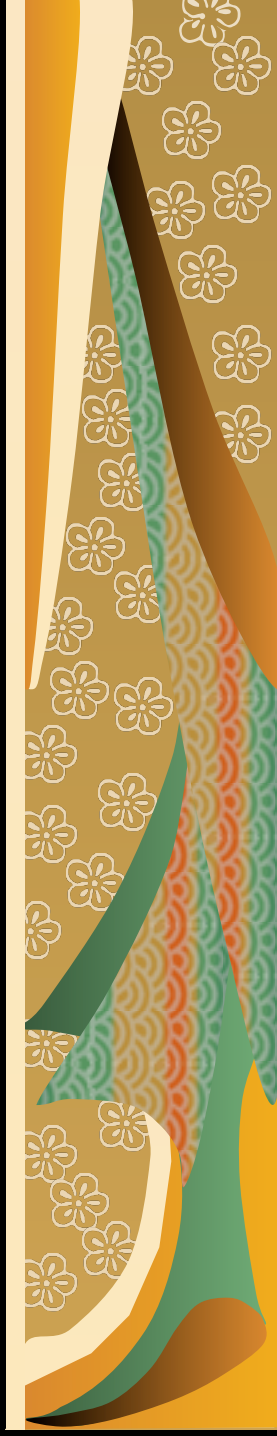
Stylometry: The Case of *The Cuckoo's Calling*

- Published April 30, 2013
- Author named as Robert Galbraith
  - Acknowledged pseudonym
- Author suggested on Twitter as J.K. Rowling



Stylometry is …

- … a classic problem in scholarship, including literature, forensic/legal, and historical scholarship
- Given a document, who wrote it?
  - Increasing research area with substantial body of work
- Sometimes called authorship attribution



Whodunit?

- Did Shakespeare really write those plays?
  - Or was it the Earl of Oxford?
  - Or Francis Bacon?
  - Or Roger Bacon?
  - Or Kevin Bacon?
    - et cetera
- How can we tell?
- How can we prove it?

How to find authors 101

Document prep

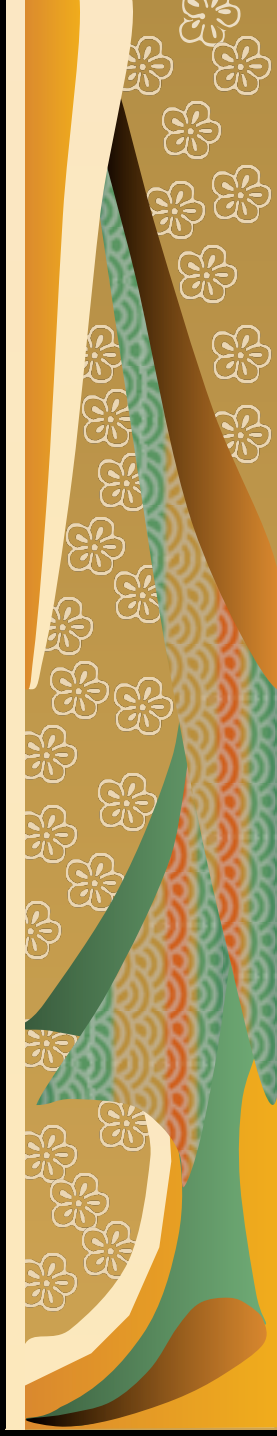
- Training documents (KD)
- Test documents (QD)

Feature Selection

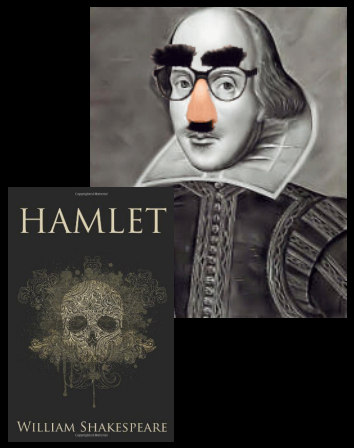
- Identify author-specific features
- Words, character clusters, &c.

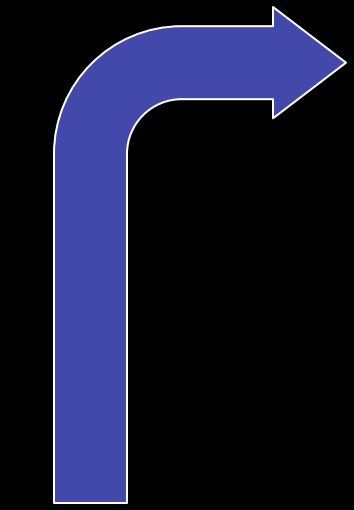
Classify

- Apply ML techniques
- k-NN, SVM, PCA, &c.



How it works

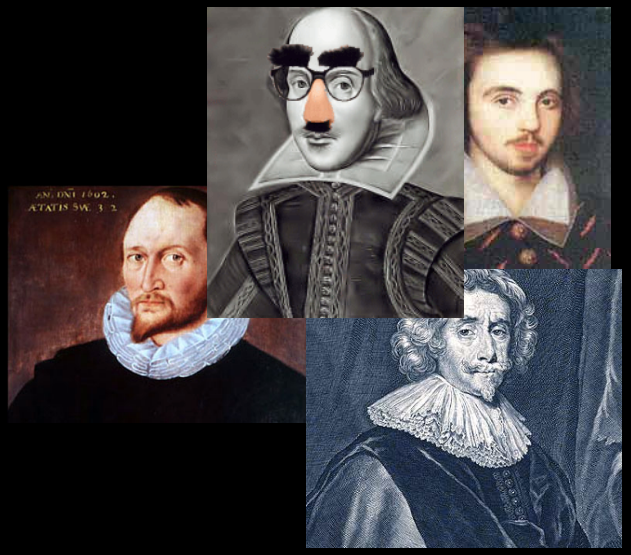

HAMLET
WILLIAM SHAKESPEARE

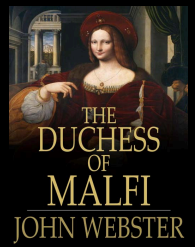



2 DISC SPECIAL EDITION
APOLLO 13


Footloose
THERE COMES A TIME TO CUT LOOSE

Works on groups, too

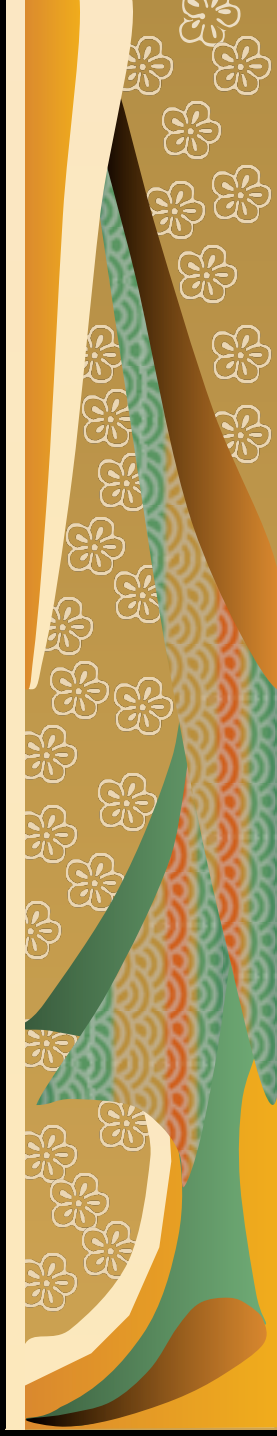




- This is called “profiling”

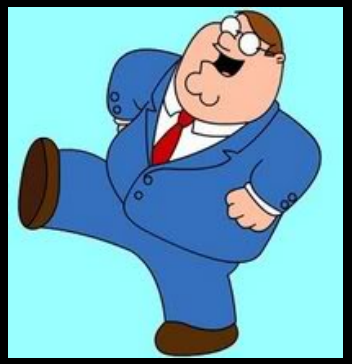
Lots of profiling options

- Gender
- Age
- Education level
- Socio-economic status
- Native language
- Mental traits like personality
- … really, only limited by imagination (and data)



What kind of “features”?

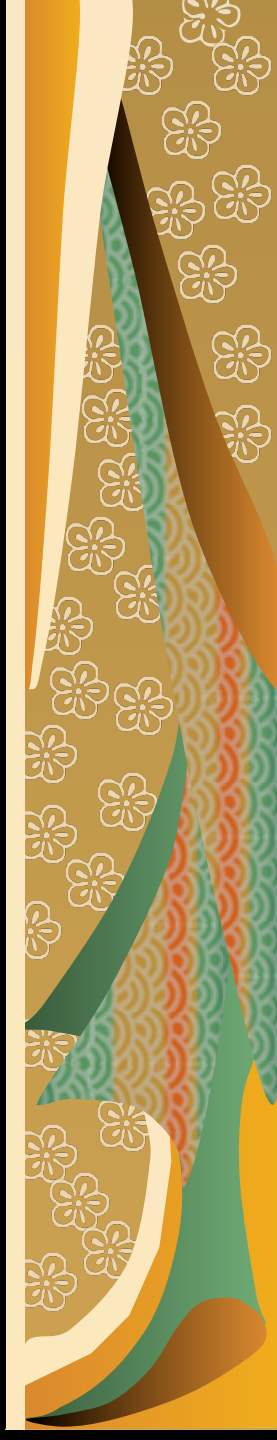
- The paradigmatic and systematic utilization of sesquipedalian lexical items can be an informative element of individual and idiosyncratic patterns of linguistic variation



Or, some people use big words

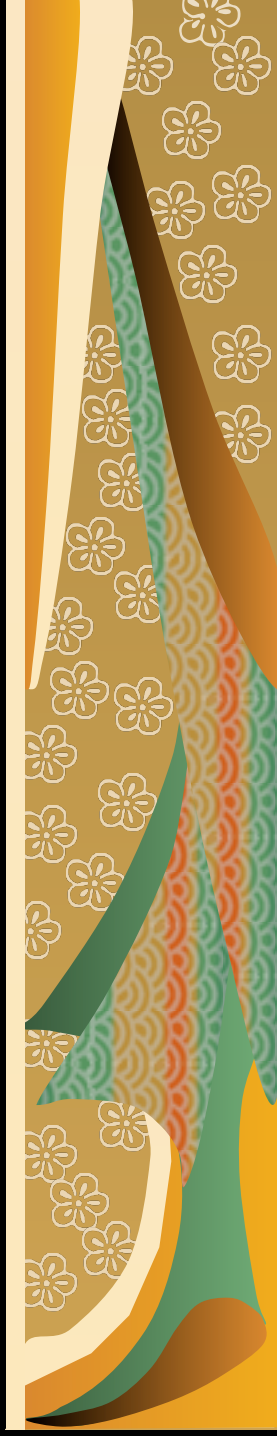
Word length as cue

- "Try to balance in your own mind the question whether the latter [text] does not deal in longer words than the former [text].... Some of these days spurious writings will be detected by this test." (August de Morgan, 1851)
- Statistically, calculate average word length and use ordinary stats (*t*-tests)



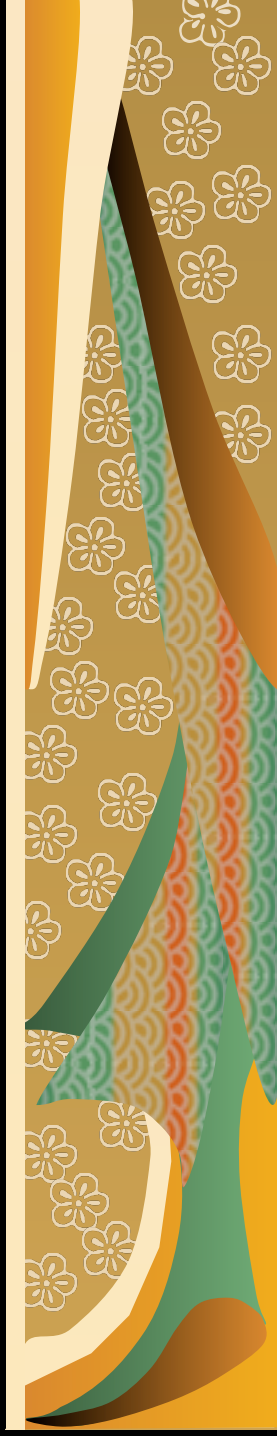
The issue

- The issue is that it doesn't work (well).
  - Like classifying people by height—almost everyone is average
- Word length isn't stable, either.
  - I write differently to friends than to journals.
- If word length won't work, what else will?



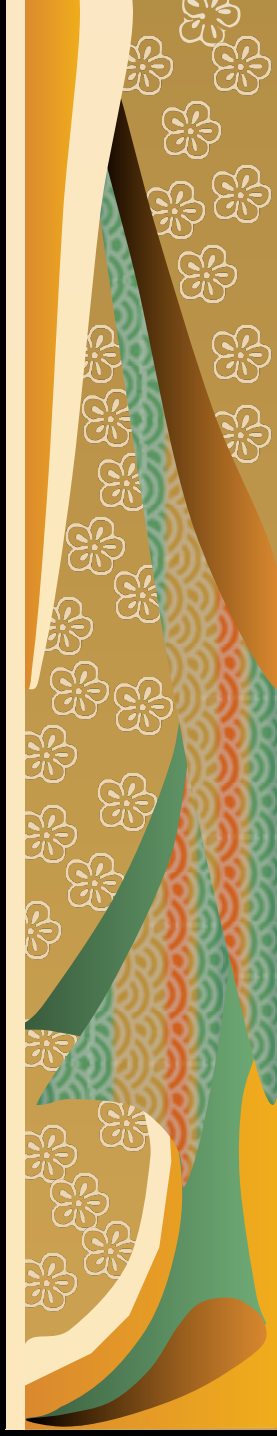
Successful feature sets

- A good feature set is
  - common
  - stable
  - informative
  - computationally tractable
- Most success so far with variations on words and character clusters.



Function words

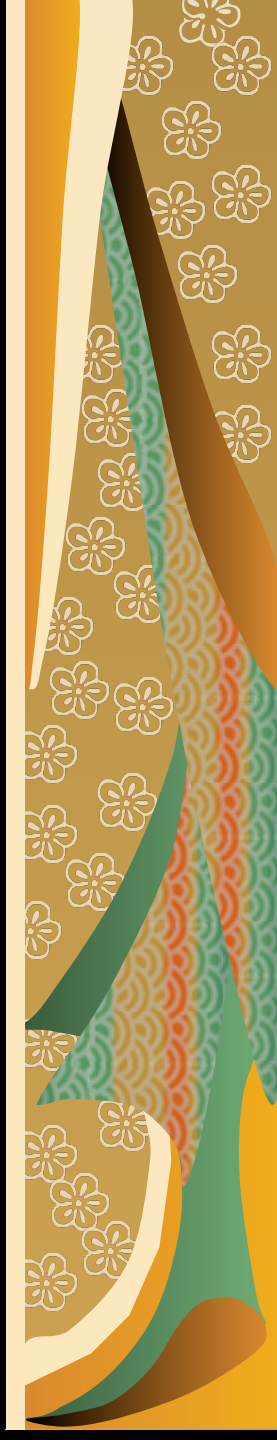
- One possibility : function words
  - *The, that, of, as, between, among*, etc.
  - Short, common, light on meaning
  - Stable across texts
  - Many types, allowing for multivariate statistics and high discriminative power



The Case of *The Royal Book of Oz*

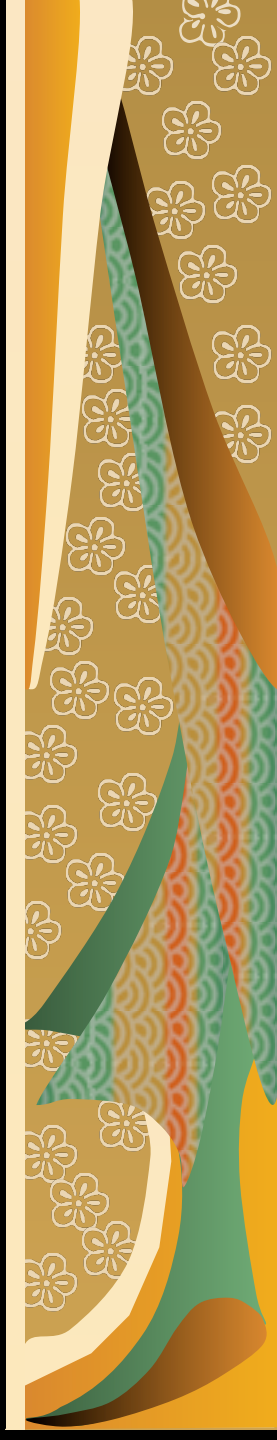
Historical synopsis

- Baum wrote “Wizard of Oz” in 1900; 13 more books by 1919.
- 15th book (“Royal Book”) unfinished at his death, Ruth Plumly Thompson took “fragmentary draft” and “unfinished notes” to complete and publish
- Published 1921 with Baum’s name on cover, “enlarged and edited” by Thompson

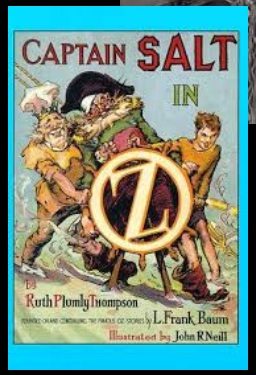


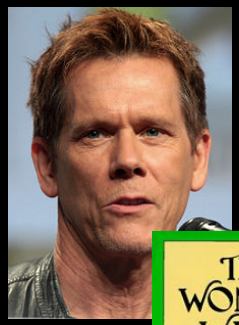
Historical revisionism

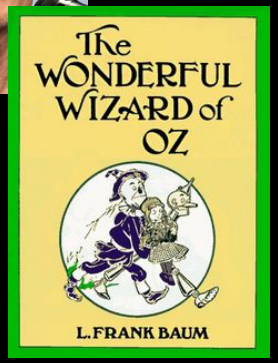
- Or did she?
- Alternate view : “Royal Book” was entirely Thompson’s work. Story about “unfinished notes” and Baum’s name used by publisher to strengthen sales.
- Thompson went on to publish nearly 20 more Oz books under her own name.



(Betcha didn't know that Baum was *really* Kevin Bacon!)

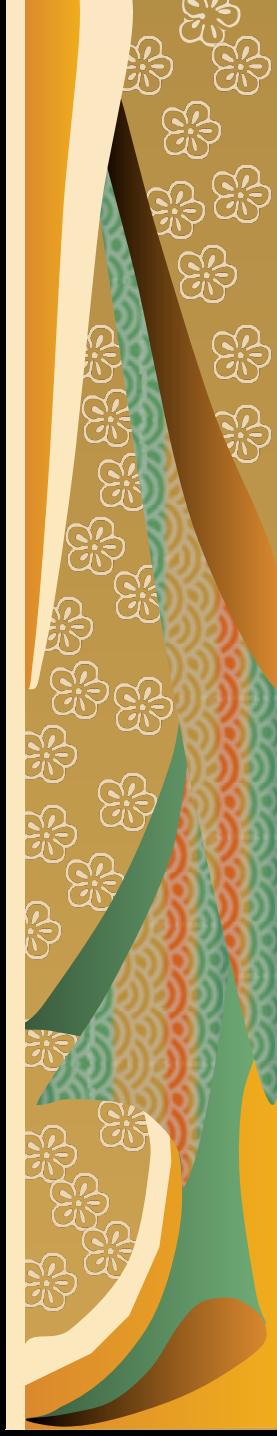




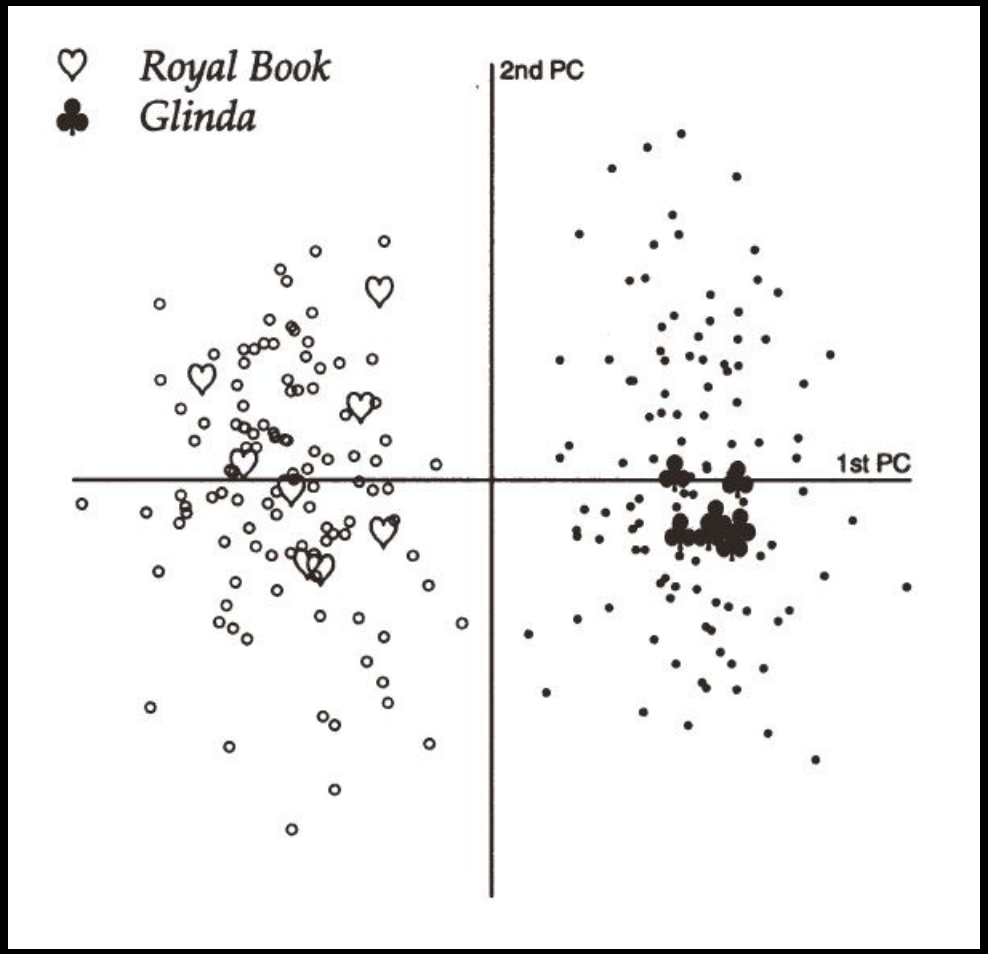


Binongo analysis

- Collect 50 “most frequent words”
- Candidates include *the, and, to, not, as, with, before, after, some, well, back, how*
- Apply stats (PCA) to perform “dimensionality reduction,” essentially making 2D image from 50D data.



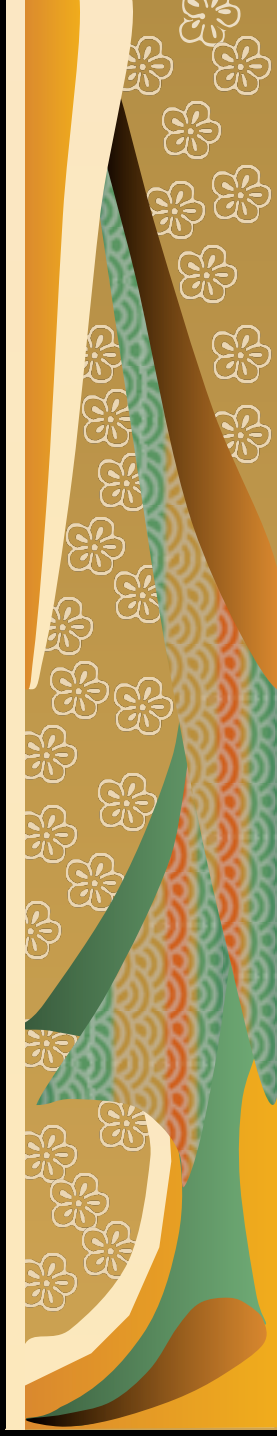
Baum is on right



So who wrote *Royal Book*?

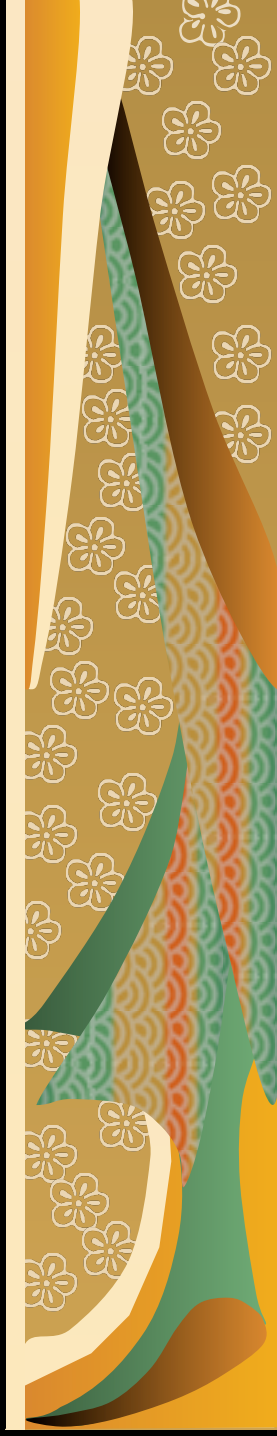
Which attributes? Which features?

- Tech can be deployed to address many types of questions….
- … but useful feature sets are limited
  - Characters
  - Words
  - Part-of-speech tags
- Tend to be low-level and uninteresting from a literary p-o-v.



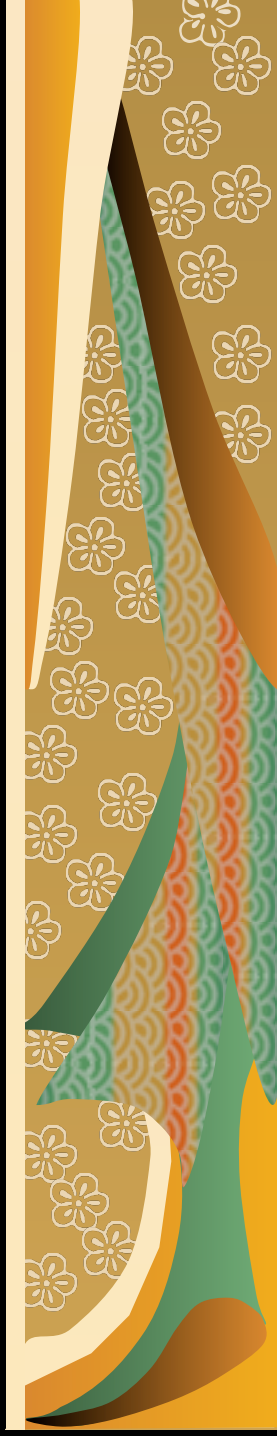
Markup and Stylometry

- … have not historically had much to do with each other
- … which is unfortunate
- Can markup help stylometrists?
- Can stylometry help the markup process?



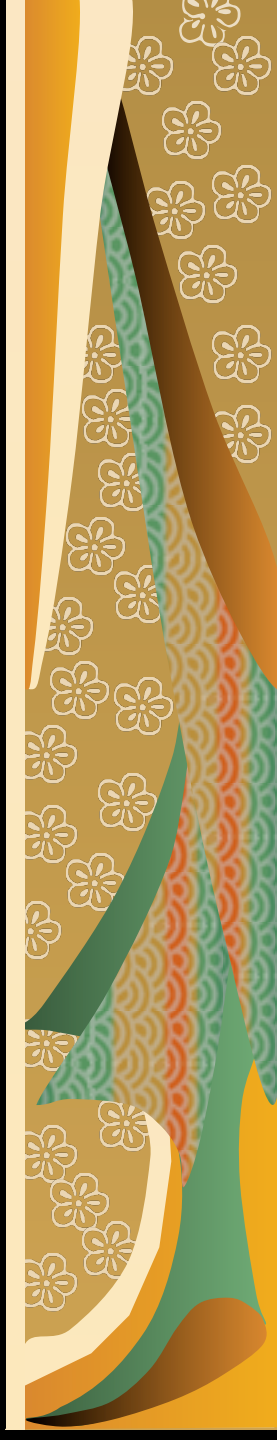
Markup is …

- … making subtle properties (“features”) obvious to computers
- Key problem with sophisticated feature sets is reliability



How confident are we?

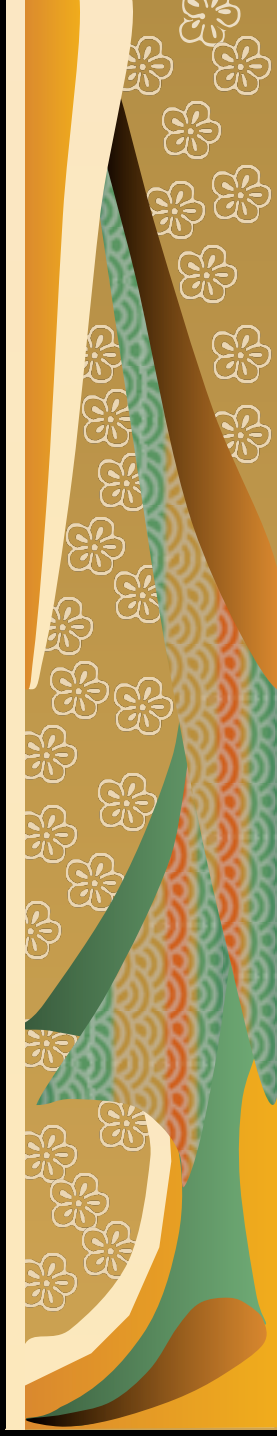
- We are 100% accurate of a UTF-8 ‘A’
- We are about 99% confident of a typed/scanned ‘A’
- POS tagging is 98% accurate (56% per sentence)
- Shallow parsing is 90% accurate
- … and it only goes downhill from there



Language is <whine>hard</whine>!

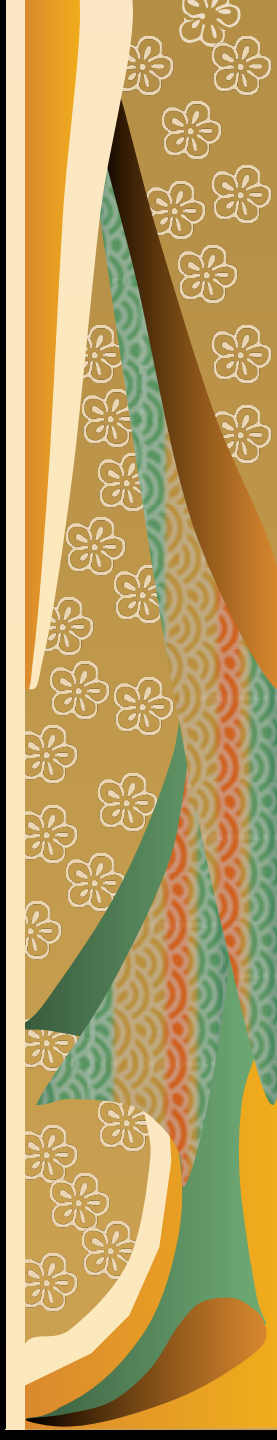
- The council refused a permit to the women because they advocated violence.
- The council refused a permit to the women because they feared violence.

- Good luck with that, Google Translate!



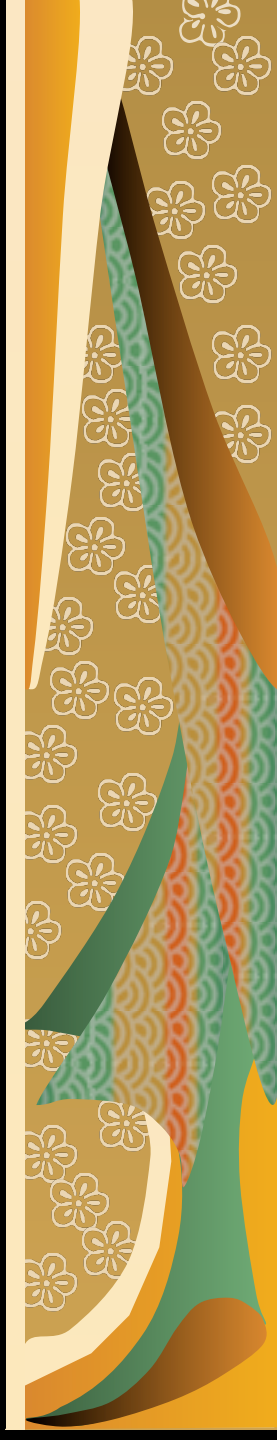
Yup, Google muffed it!

- The council refused a permit to the women because they feared violence.
- *Le conseil a refusé un permis aux femmes parce qu'elles craignaient la violence.*
  - Requires human expertise
  - Computer “just” applies model
  - But the computer’s mistakes are exactly where an individual writing differs from the computational model!



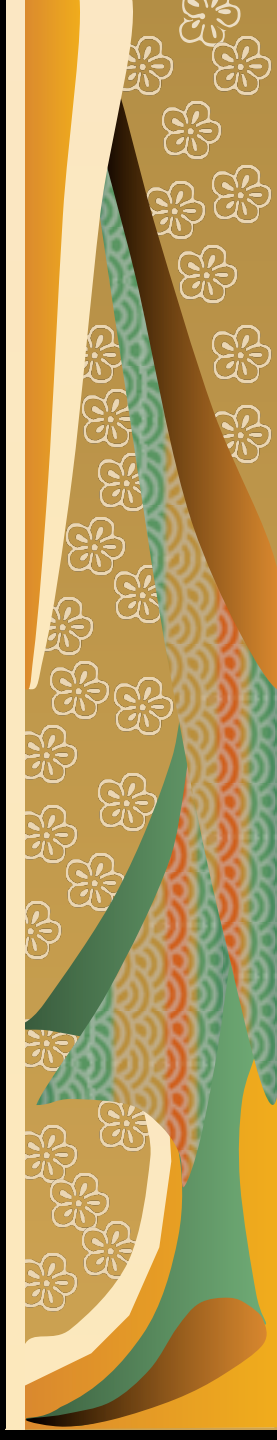
Where markup can help

- <epigram>Markup makes the inobvious obvious.</epigram>
- Allows stylometric focus on properties that are not “tractable”
  - Human accuracy is higher than machine
  - Example: Phonetics. *“The tough coughs as he ploughs through the dough.”*



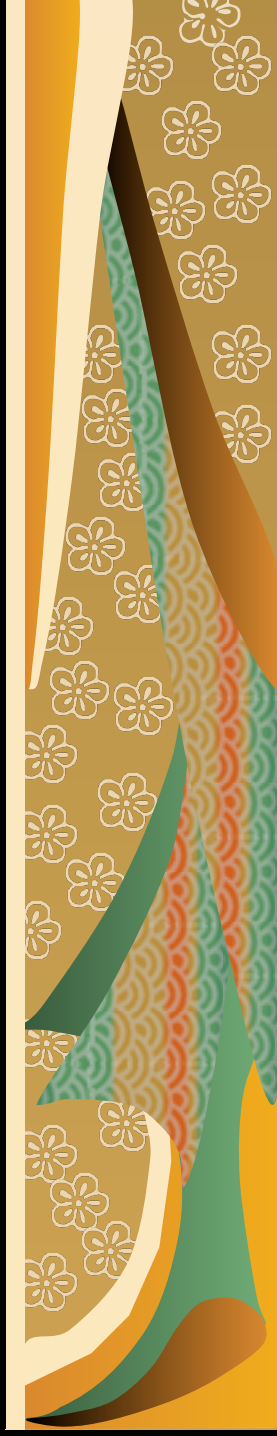
Pushing feature sets

- … into the realm of usefulness
- No one cares about Mitford's use of "that" vs. "which."
- What do Mitford scholars care about? Learning about useful properties of Mitford's world can improve our ability both to understand and to detect her.



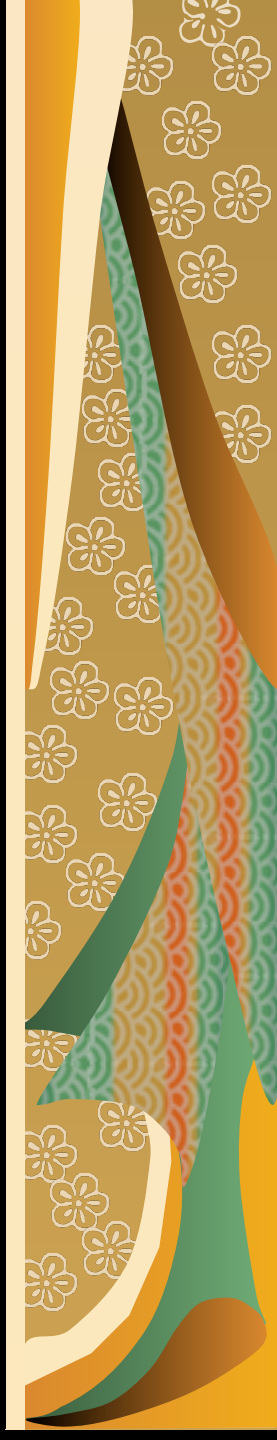
Upcoming workshop

- Practice problems to demonstrate stylometric analysis
- Will discuss approaches and limitations
- … and you get a copy of the software to play with if you like
  - Feel free to “bring your own data”



Summary and conclusions

- Stylometry is an important DH topic
- The stylometric and markup communities have not generally had much to do with each other
  - … which is unfortunate
- We need to discuss common interests and common tasks to help find common solutions



So let's start discussing!

Thank you very much!

Patrick Juola
juola@mathcs.duq.edu