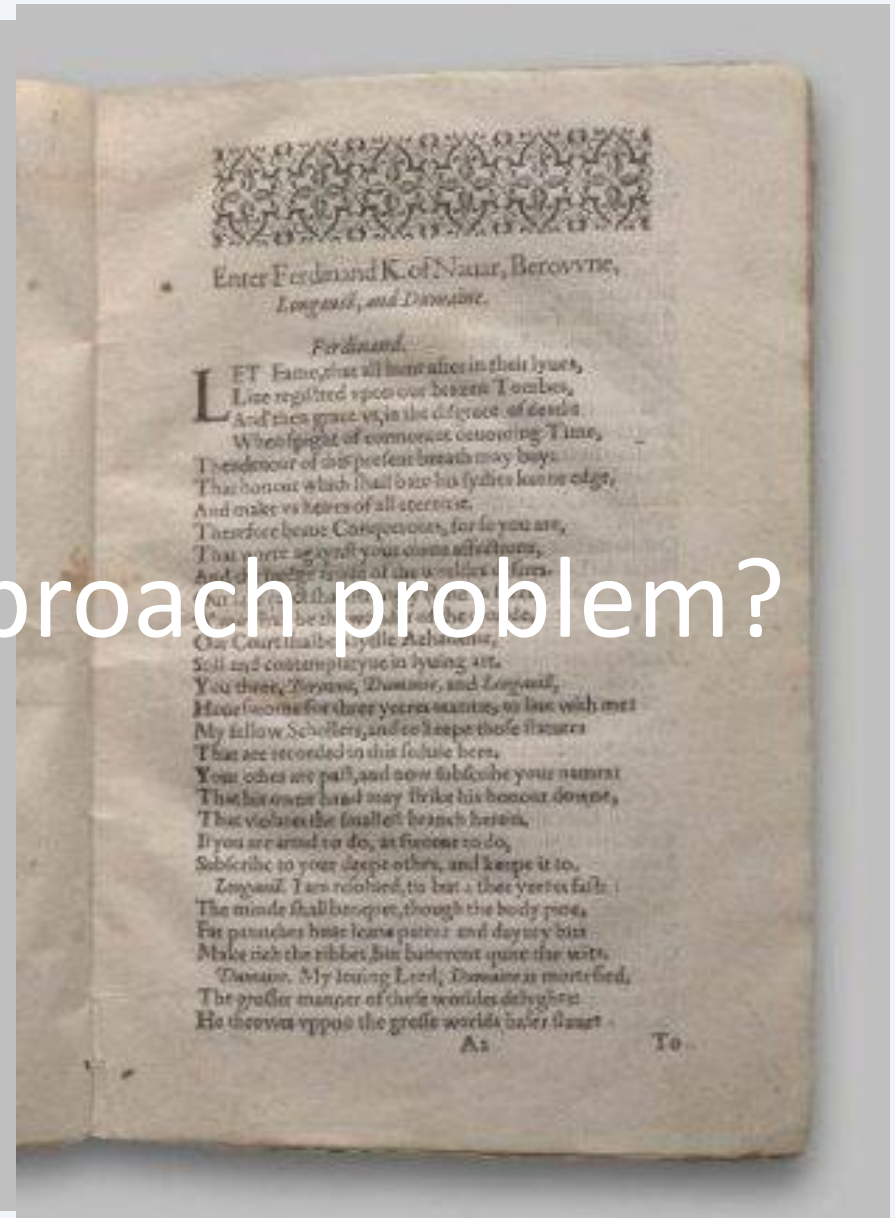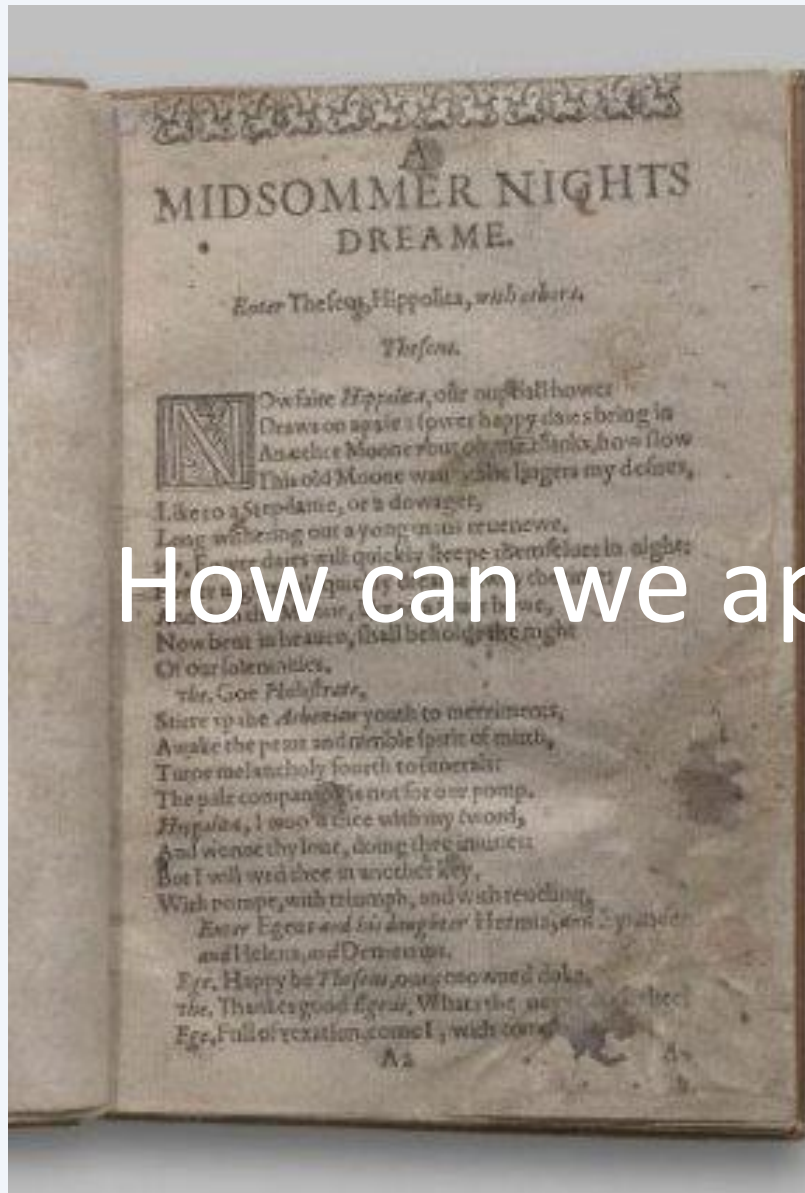# Text Searching, Analysis, & Generation
COMP 200 & COMP 130

Do Two Texts Share Authors?

How can we approach problem?
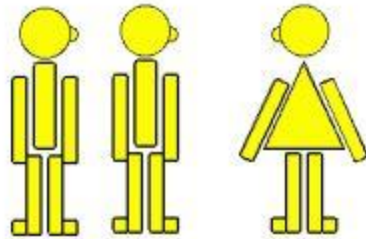
Some Possible Analyses

- Handwriting

- Chemical

- Historical

- Literary

- Linguistic -- *stylometry*

Hacker Factor

Hacker Factor
Gender Guesser

As demonstrated at:

Black Hat Briefings, USA 2006

Gender Guesser

The words you use can disclose identifying features. This tool attempts to determine an author's gender based on the words used.

Submitted text is evaluated based on two types of writing: formal and informal. Formal writing includes fiction and non-fiction stories, articles, and news reports. Informal writing includes blog and chat-room text. (Email can be formal, informal, or some combination.) You should view the results based on the appropriate type of writing.
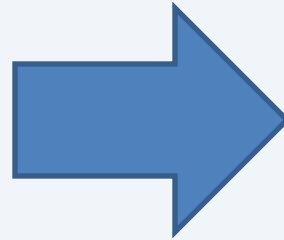
Analyze

Type or paste a writing sample for gender analysis. Then click on "Analyze" to see the results. For best performance, use at least 300 words -- more words is generally more accurate.

Analyze   Clear   About

Problem Decomposition – 1ˢᵗ Stage

Python data

File Input Overview

A typical text file: [Gettysburg Address](#)

- Multiple *lines* – read one at a time
- Words separated *delimiters* – spaces, punctuation
- Capitalization

Searching in a String

Can you **help** me?

She's my **help**er.

You are so un**help**ful.

Help!

Are those your dog's w**help**s?

Searching the British National Corpus

Some Suggested Readings

- Wikipedia: Stylometry

- Freeware & data: Signature
  - Blog: Applying it to purported Ron Paul newsletters

- Freeware: Java Graphical Authorship Attribution Program (JGAAP)

- Technical paper: Biblical authorship – includes overview of approaches

- Technical paper: Profiling anonymous authors

- Gender Guesser – guesses author's gender