Data and Culture

Prof. Andrew Goldstone (andrew.goldstone@rutgers.edu)
Prof. Meredith McGill (meredith.mcgill@rutgers.edu)

November 10, 2022. Genre and literary classification.

lab 7, concluded

- Skip to section (4) and make sure everyone in your group can explain what `facet_grid` and `facet_wrap` do.
- Then do make sure you can all label axes with the code in section (5).
- the point: visualization is also a data sandwich
  - "reduction" is very useful for seeing patterns
  - put more information into plots to see more variation
  - …up to a point

the data sandwich (as always)

github.com/tedunderwood/horizon/blob/master/chapter2/-metadata/concatenatedmeta.csv

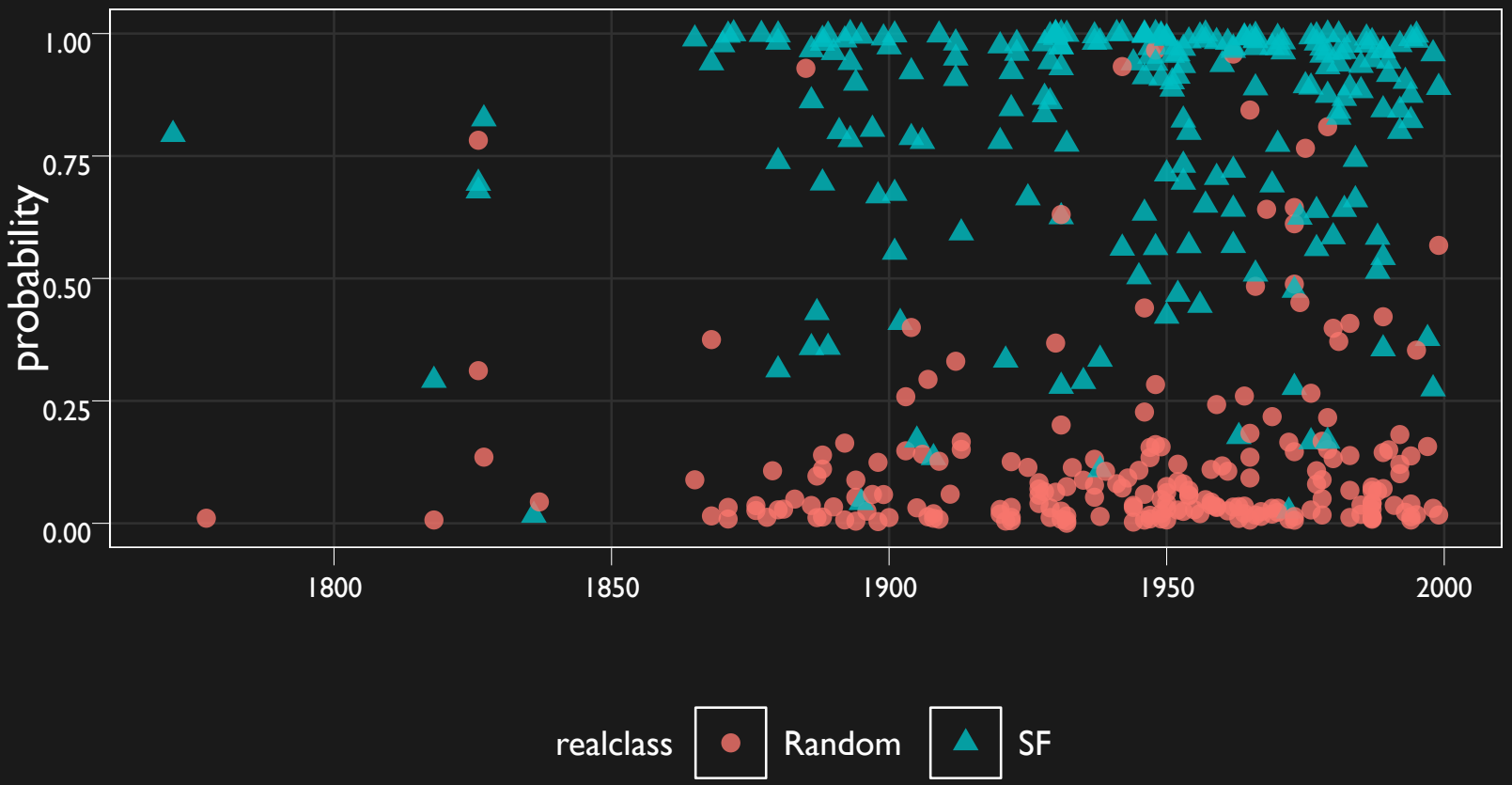
- cf. github.com/tedunderwood/horizon/blob/master/chapter2

look at the pictures

- Consider Underwood's Figure 2.2. What is the grammar of this graphic? Remember the layers of the grammar:
  1. Source data
  2. Data transformation, if any
  3. Choice of graphical marks (`geom_`)
  4. Aesthetic mappings (`aes`)
  5. Faceting, if any (`facet_`)
  6. Scales and coordinates (`scale_`)
- interpret: What does this visualization choice emphasize in the argument about genre, and what does it put in the background?

reproduction

```
sf <- read_csv("https://raw.githubusercontent.com/tedunderwood/h
# cosmetic relabeling
sf <- sf |> mutate(realclass=case_when(
    realclass == 1 ~ "SF",
    realclass == 0 ~ "Random"))
fig2_2 <- ggplot(sf) +
    geom_point(aes(x=firstpub,
                   y=logistic,
                   color=realclass,
                   shape=realclass),
               alpha=0.8)
```



Probability of being science fiction. Science fiction, 1771–1999, classified with 90.6% accuracy.

in-sample accuracy: 90.6%

figure 2.1, the grammar

- What is the grammar of Underwood's figure 2.1?

figure 2.1, the grammar

- What is the grammar of Underwood's figure 2.1?

```
read_csv("https://github.com/tedunderwood/horizon/.../etcetcetc.
    # ...then do a little munging... |>
    ggplot() +
    geom_point(aes(firstpub, logistic,
                   color=realclass, shape=genre)) +
    labs(shape="actually...not a multinomial model")
```

how Underwood models genres

- obtain digitized SF and random fiction texts
  - how and from where?
- "bag of words" model (almost)
  - convert texts to feature vectors $x_1, x_2, x_3, ...$
  - $x_1 =$ frequency of "the" in the text
  - $x_2 =$ frequency of "star" in the text
  - $x_3 =$ frequency of "child" in the text
  - ...about 4100 features (rare words ignored)
- for each text, record $y = 1$ if SF, $y = 0$ otherwise

"train a classifier"

- logistic regression: pretend every case is a (biased) coin flip
- bias of the coin assumed to depend systematically on x_i as:

$$P(y = 1|x_i) = \frac{1}{1 + \exp\left(-\left(b_0 + \sum_i b_i x_i\right)\right)}$$

- find best fit b_i using training data ("best"…)
- now you have an SF-detector: for any text $x_1, x_2, ...$
  - if $\hat{P}(x_1, x_2, ...) \geq 0.5$, guess it's SF

what is a genre?

Genre…is a set of conventional and highly organised constraints on the production and interpretation of meaning.

Genres are always complex structures which must be defined in terms of all three of these dimensions: the formal, the rhetorical, and the thematic.

John Frow, *Genre*, 2nd ed. (London: Routledge, 2014), 10, 83.

thematic

A quick glance at the words most predictive of detective fiction reveals the themes we would expect: *police*, *murder*, *investigation*, and *crime*. (48)

formal

Detective stories…possess a "specific device" of exceptional visibility and appeal: clues…. I speak of clues as a *formal* device because their narrative function (the encrypted reference to the criminal) remains constant, although their concrete embodiment changes. (Moretti, "Slaughterhouse," 212, 212n7)

rhetorical

If we restrict our model to one hundred extremely common words, we can still identify detective stories with 86% accuracy…the predictive words in this version of the model include vaguely interrogative signals appropriate for mysteries—*who, why, any, something,* and the question mark—but also more puzzling words like *have* and *was*. (Underwood, 49)

fuzzy thinking

it increasingly seems that a genre is not a single object we can observe and describe. It may instead be a mutable set of relations between works that are linked in different ways and resemble each other to different degrees. (Underwood, 41)

Usually, we tend to have a rather "Platonic" idea of genre…The tree suggests a different image: branches, formal choices, that don't replicate each other but rather move away from each other, turning the genre into a wide field of diverging moves. (Moretti, 217)

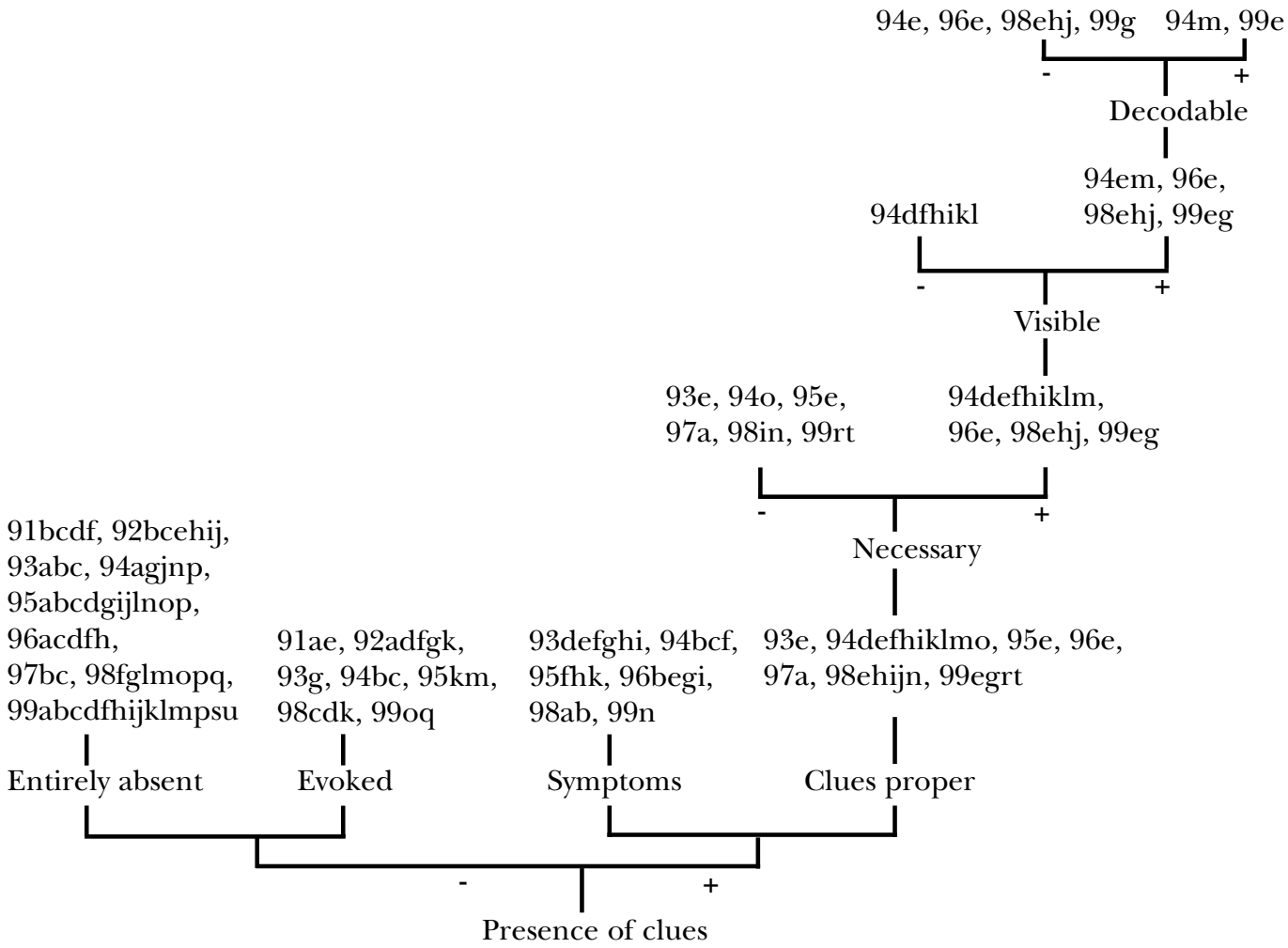


Figure 2 Clues in the Strand magazine, 1891–99

Moretti, 219, figure 2. What is the grammar of this graphic?

# of being numerous

And wrong moves, mostly: where nine writers out of ten (and half of the tenth) end up on dead branches. This was my initial question, remember: what happens to the 99.5 percent of published literature? This: it's caught in a morphological dead end. (Moretti, 217)

Our habit of equating brand identity [of SF] with literary consolidation seems not to be well founded. (Underwood, 65)

next

- look out for an e-mail about updating the `dataculture` package again
- strongly recommended: Underwood appendices A–B
- also recommended: browse Underwood's online "replication archive"