Fiction’s Functions:
Three Data-Driven Hypotheses

Andrew Piper, McGill University

How can we use data to UNDERSTAND literature?

Three Hypotheses

- Legibility
- Sensibility
- Immutability

Three Hypotheses

- Legibility
- Sensibility
- Immutability
- Heteronormativity
- Social Hierarchy

Key Terms

- Predictive Modeling
- Machine Learning
- Feature Space
- Inference v. Observation

Data

Collection	Key	Description	Documents
19C Canon	EN_FIC	English Fiction	100
	EN_NOV	English Novels	100
	EN_NOV_3P	English Novels 3-Person	107
	EN_NON	English Non-Fiction	100
	EN_HIST	English Histories	85
	DE_NOV	German Novels	100
	DE_NOV_3P	German Novels 3-Person	110
	DE_NON	German Non-Fiction	100
	DE_HIST	German Histories	75
Hathi Trust 19C	HATHI_FIC	Hathi Trust Fiction	9,426
	HATHI_NON	Hathi Trust Non-Fiction	11,732
	HATHI_TALES	Hathi Trust Fiction Minus Novels	428
1790-1990	STAN_KLAB	English Novels	6,421
Contemporary	CONT_NOV	Contemporary Novels	200
	CONT_NOV_3P	Cont. Novels 3-Person	210
	CONT_NON	Contemporary Non-Fiction	200
	CONT_HIST	Contemporary Histories	200

How do we know something is a work of fiction?

A

On the short ferry ride from Buckley Bay to Denman Island, Juliet got out of her car and stood at the front of the boat, in the summer breeze. A woman standing there recognized her, and they began to talk. It is not unusual for people to take a second look at Juliet and wonder where they've seen her before, and sometimes, to remember.

B

Jeff is 24, tall and fit, with shaggy brown hair and an easy smile. After graduating from Brown three years ago, with an honors degree in history and anthropology, he moved back home to the Boston suburbs and started looking for a job. After several months, he found one, as a sales representative for a small Internet provider. He stays in touch with friends from college by text message and email, and still heads downtown on weekends to hang out at Boston's "Brown bars." "It's kinda like I never left college," he says, with a mixture of resignation and pleasure. "Same friends, same aimlessness."

The Feature Space

LIWC (Linguistic Inquiry and Word Count)

- Linguistic Process
  - Pronouns, Verb Tense, Punctuation, etc.
- Social Process
  - Family, Friends, Humans
- Cognitive Process
  - Insight (think, know), Causation, Discrepancy, Certainty
- Perceptual Process
  - See, Hear, Feel
- Affective Process
  - Positive / Negative Emotion, Sadness, Anxiety, Fear
- Biological Concerns
  - Bodies, Health, Sex, Eating
- Relativity
  - Motion, Time, Space
- Thematic
  - Work, Achievement, Leisure, Money, Religion, Death, Home

Legibility

Legibility

- “There is no textual property, syntactical or semantic, that will identify a text as a work of fiction.” John Searle, “The logical status of fictional discourse”
- “It is almost universally accepted today that no distinguishing features separate literary from non-literary texts.” Benjamin Hrushovski, *Fictionality and Fields of Reference*
- “This is the hypothesis I would like to test and submit to your discussion. There is no essence or substance of literature: literature is not. It does not exist.” Jaques Derrida, *Demeure: Fiction and Testimony*

Legibility

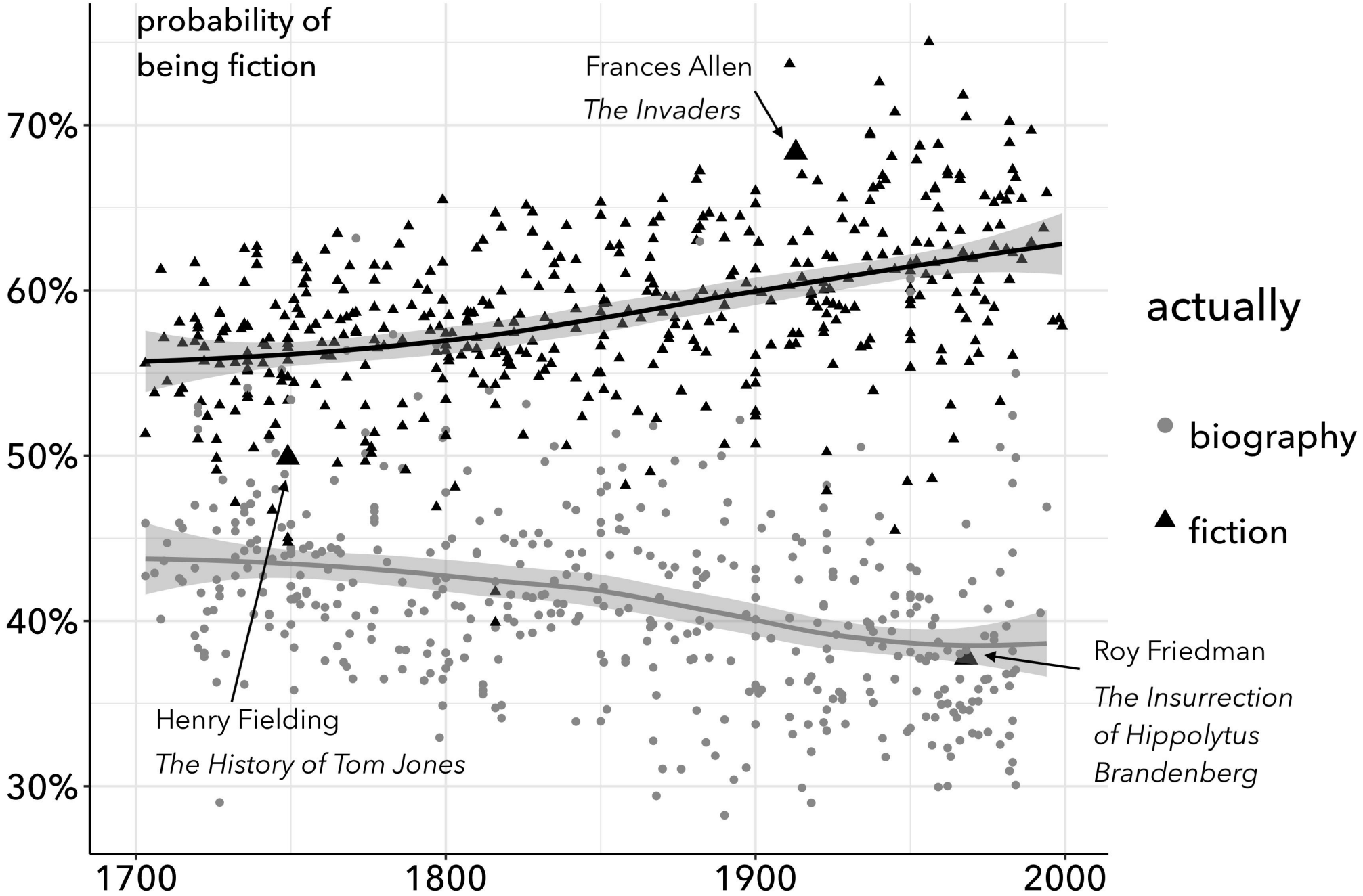
Classification results for predicting fictional texts using tenfold cross-validation with an SVM classifier

Corpus1	Corpus2	Avg. Accuracy (F1)	No. Docs
Fiction (EN_FIC)	Non-Fiction (EN_NON)	0.94	100/100
English Novel (EN_NOV)	Non-Fiction (EN_NON)	0.96	100/100
German Novel (DE_NOV)	Non-Fiction (DE_NON)	0.95	100/100
English Novel 3P (EN_NOV_3P)	History (EN_HIST)	0.99	95/86
Germ Novel 3P (DE_NOV_3P)	History (DE_HIST)	0.99	88/75
Cont. Novel (CONT_NOV)	Non-Fiction (CONT_NON)	0.96	193/200
Cont. Novel 3P (CONT_NOV_3P)	History (CONT_HIST)	0.99	210/200
19C Fiction (HATHI) (Trained)	Cont. Novel (CONT) (Tested)	0.91	21,158/400

Legibility

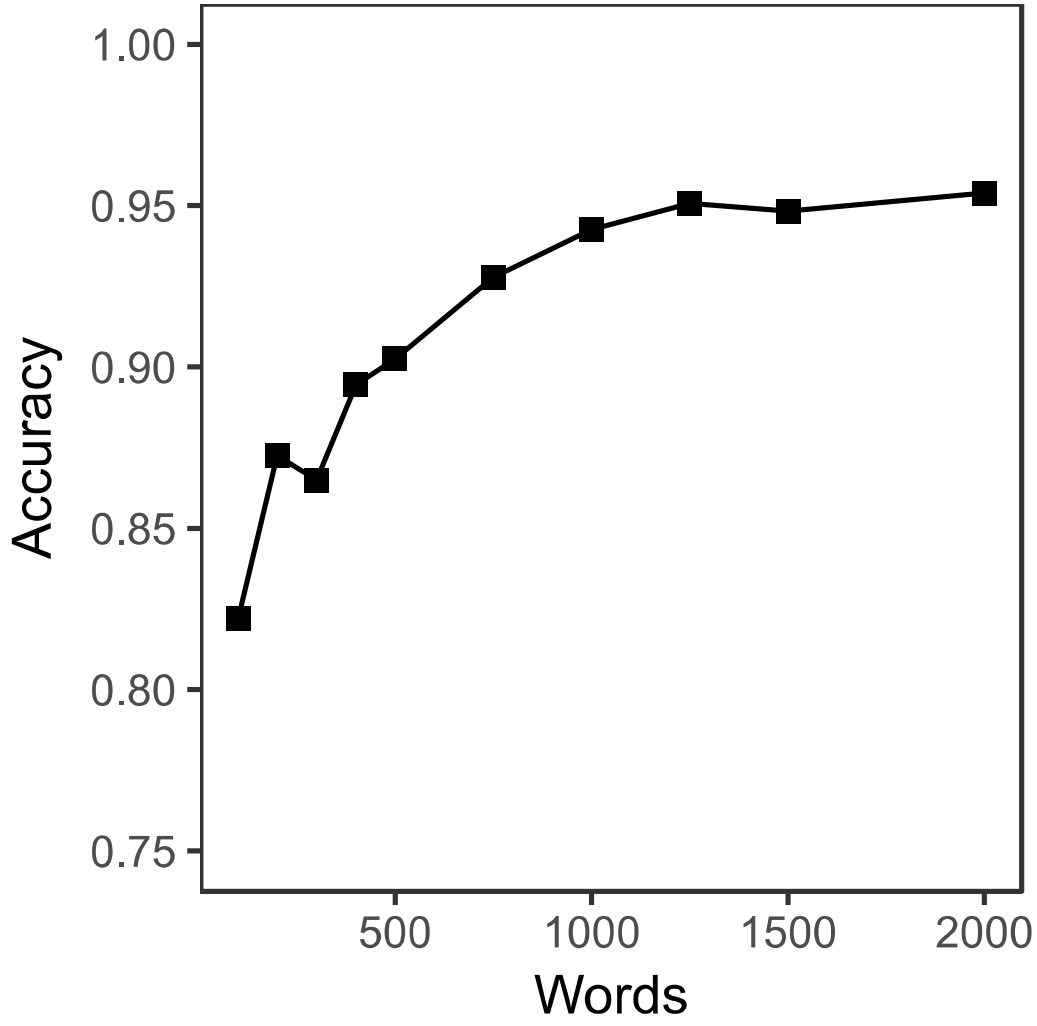
Classification results for predicting fictional texts using tenfold cross-validation with an SVM classifier

Corpus1	Corpus2	Avg. Accuracy (F1)	No. Docs
Fiction (EN_FIC)	Non-Fiction (EN_NON)	0.94	100/100
English Novel (EN_NOV)	Non-Fiction (EN_NON)	0.96	100/100
German Novel (DE_NOV)	Non-Fiction (DE_NON)	0.95	100/100
English Novel 3P (EN_NOV_3P)	History (EN_HIST)	0.99	95/86
Germ Novel 3P (DE_NOV_3P)	History (DE_HIST)	0.99	88/75
Cont. Novel (CONT_NOV)	Non-Fiction (CONT_NON)	0.96	193/200
Cont. Novel 3P (CONT_NOV_3P)	History (CONT_HIST)	0.99	210/200
19C Fiction (HATHI) (Trained)	**Cont. Novel (CONT) (Tested)**	**0.91**	**21,158/400**



Credit: Ted Underwood, *Distant Horizons*

Legibility



Accuracy of predicting fictional texts using an increasing number of words from the beginning of the document

Sensibility

Decision Tree Rules

Rule	No_Docs	Accuracy	Leading
12	9997	80.8%	ppron
41	6524	98.9%	ppron
43	5989	98.6%	anxiety_perception
38	5594	98.9%	ppron
8	5459	95.4%	pronoun
27	4736	99.3%	ppron
36	4689	99.0%	past_quote
25	4542	99.4%	you_percept
20	4514	99.5%	anx_percept
21	4187	99.4%	ppron_past
19	4169	99.5%	pronoun
29	4146	99.2%	function_ppron
28	4076	99.2%	pronoun
45	4017	98.5%	social
23	3742	99.4%	ppron
37	3729	99.0%	past_quote
7	3564	95.4%	verb

Data Set: HATHI_FIC + HATHI_NON (n=20,344)

Rule 41: (6524/68, lift 1.8)
ppron <= 7.23
verb <= 11
Exclam <= 0.16
-> class non [0.989]

Rule 43: (5989/83, lift 1.8)
anx <= 0.47
percept <= 1.56
-> class non [0.986]

Rule 8: (5459/252, lift 2.1)
pronoun > 10.1
past > 3.37
anx > 0.33
see > 0.62
feel > 0.43
Exclam > 0.16
Parenth <= 0.17
OtherP <= 0.31
-> class fic [0.954]

Overall Model Accuracy

Precision	Recall	F1
0.913	0.945	0.929

Data Set: HATHI_FIC + HATHI_NON (n=20,344)

Removing pronouns and dialogue markers

Rule 6: (10223/2310, lift 1.7)
percept > 2.01
-> class fic [0.774]

Rule 4: (5504/493, lift 2.0)
past > 3.41
future > 0.77
friend > 0.16
anx > 0.33
-> class fic [0.910]

fiction

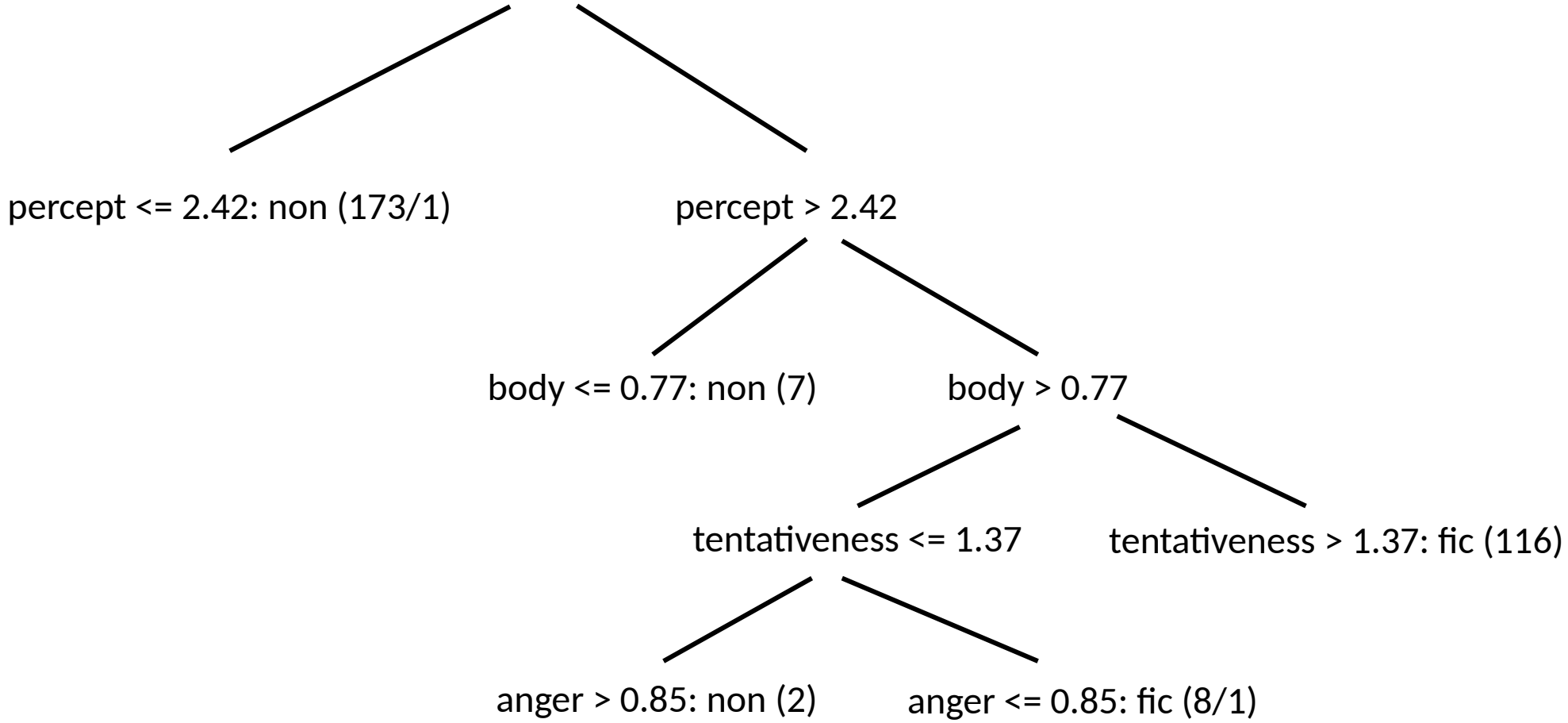
Rule 41: (4961/77, lift 1.8)
past <= 3.41
percept <= 2.01
-> class non [0.984]

Rule 21: (4919/37, lift 1.8)
friend <= 0.11
percept <= 1.78
-> class non [0.992]

non

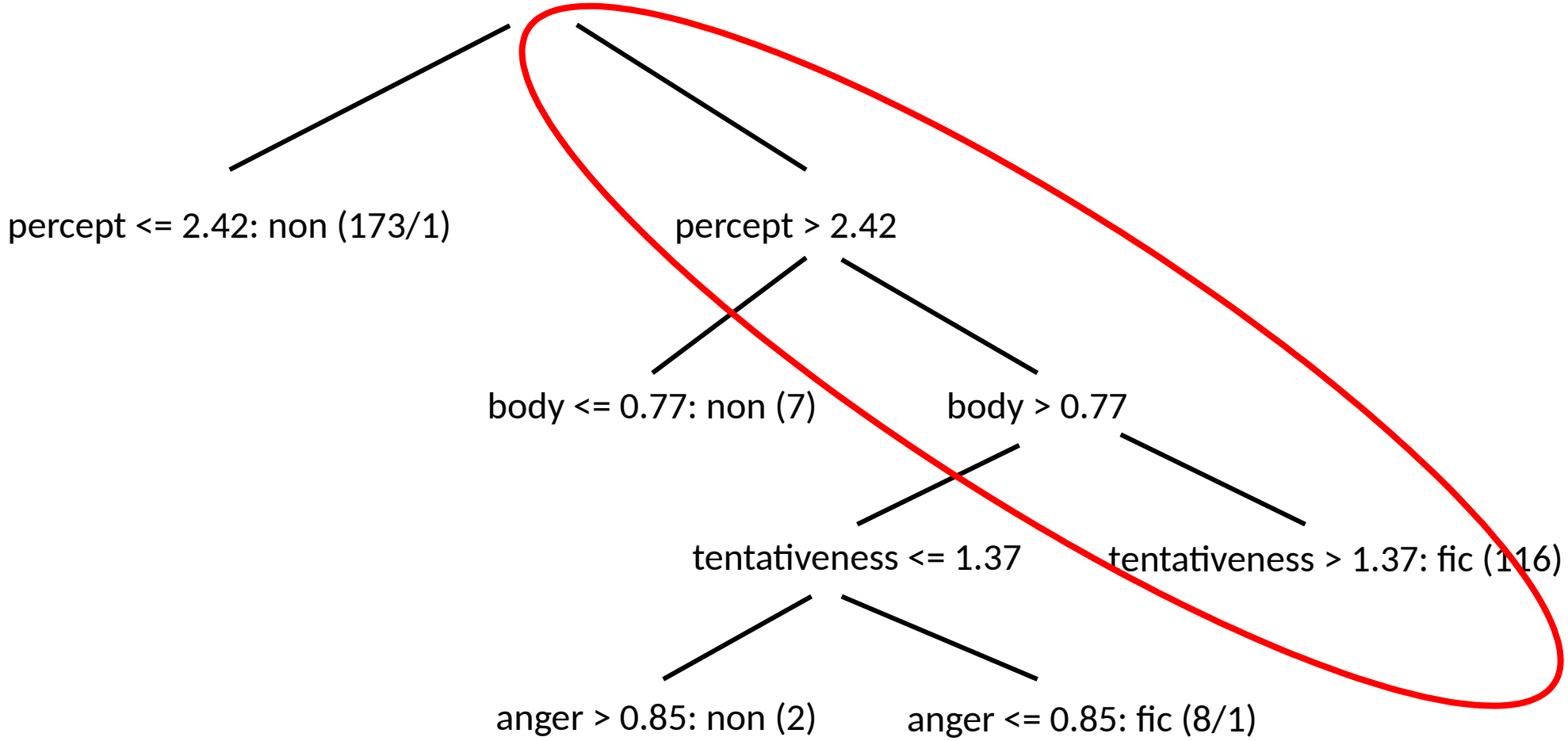
Data Set: HATHI_FIC + HATHI_NON (n=20,344)

Contemporary Literature



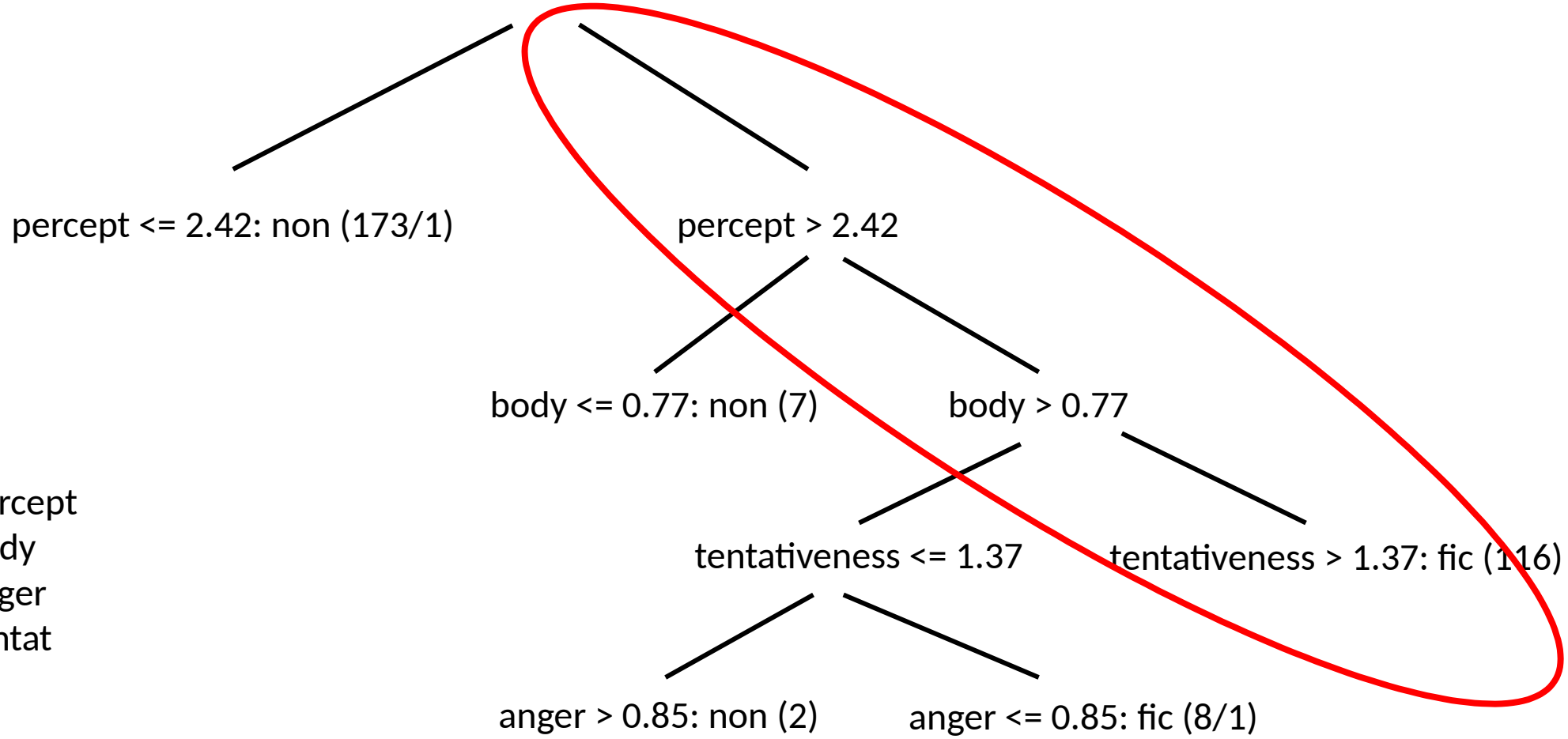
Data Set: CONT_NOV_3P + CONT_HIST (n=306)

Contemporary Literature



Data Set: CONT_NOV_3P + CONT_HIST (n=306)

Contemporary Literature



Attribute usage:

97.06% percept
93.46% body
48.37% anger
47.39% tentat

Data Set: CONT_NOV_3P + CONT_HIST (n=306)

Implications

- Beyond realism
- Beyond theories of mind
- Toward a phenomenological theory of fiction’s function

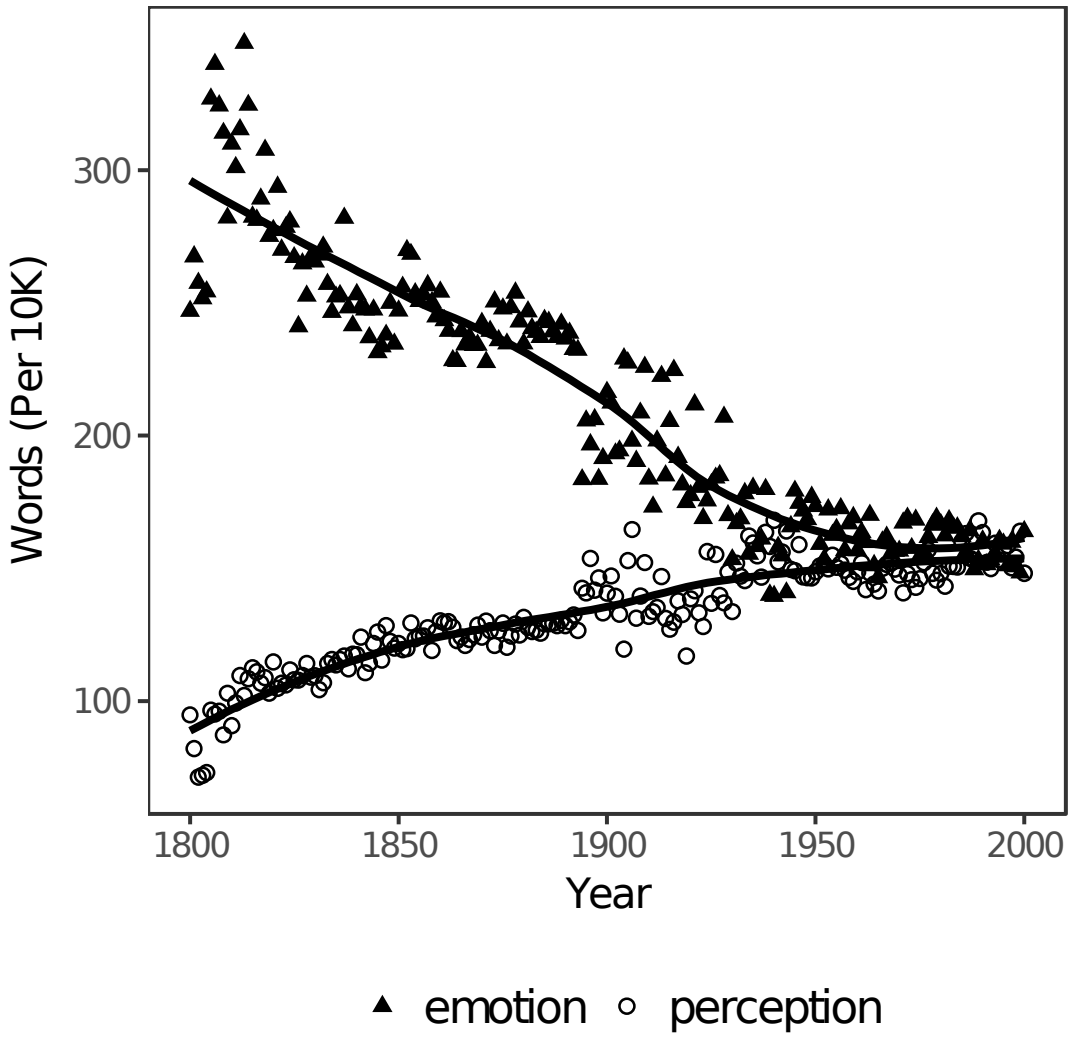
Immutability

Immutability

Classification results for predicting fictional texts using tenfold cross-validation with an SVM classifier

Corpus1	Corpus2	Avg. Accuracy (F1)	No. Docs
Fiction (EN_FIC)	Non-Fiction (EN_NON)	0.94	100/100
English Novel (EN_NOV)	Non-Fiction (EN_NON)	0.96	100/100
German Novel (DE_NOV)	Non-Fiction (DE_NON)	0.95	100/100
English Novel 3P (EN_NOV_3P)	History (EN_HIST)	0.99	95/86
Germ Novel 3P (DE_NOV_3P)	History (DE_HIST)	0.99	88/75
Cont. Novel (CONT_NOV)	Non-Fiction (CONT_NON)	0.96	193/200
Cont. Novel 3P (CONT_NOV_3P)	History (CONT_HIST)	0.99	210/200
19C Fiction (HATHI) (Trained)	**Cont. Novel (CONT) (Tested)**	**0.91**	**21,158/400**

The Great Convergence, or Redefining Feeling



Frequency of words related to emotions and perception in 6,421 English-language novels