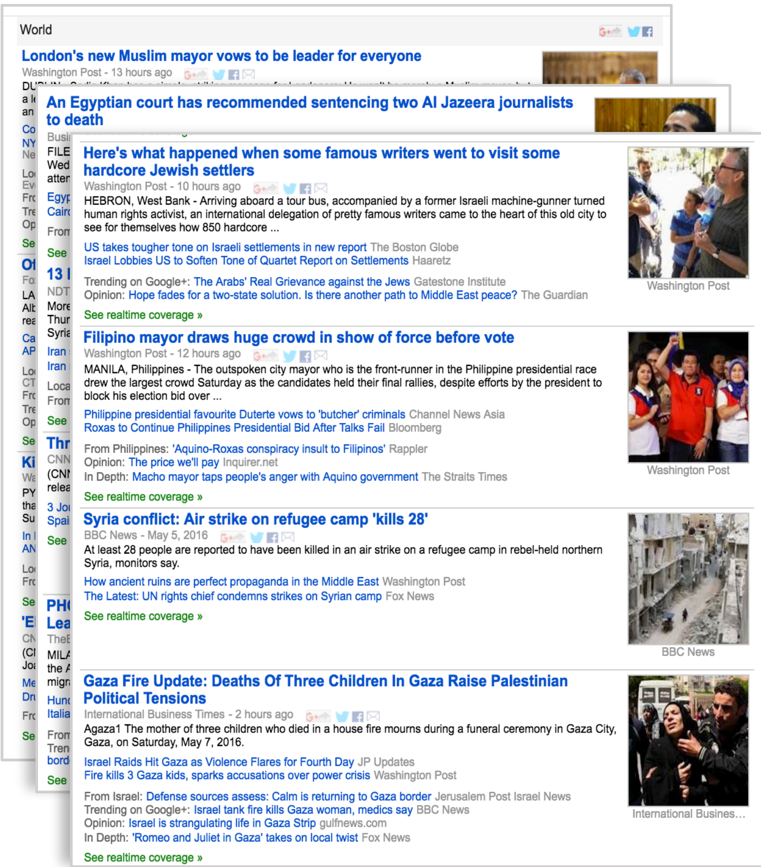
Interpretable Machine Learning: Lessons from Topic Modeling

Michael Paul
University of Colorado Boulder

CHI HCML
May 8, 2016

Topic Modeling

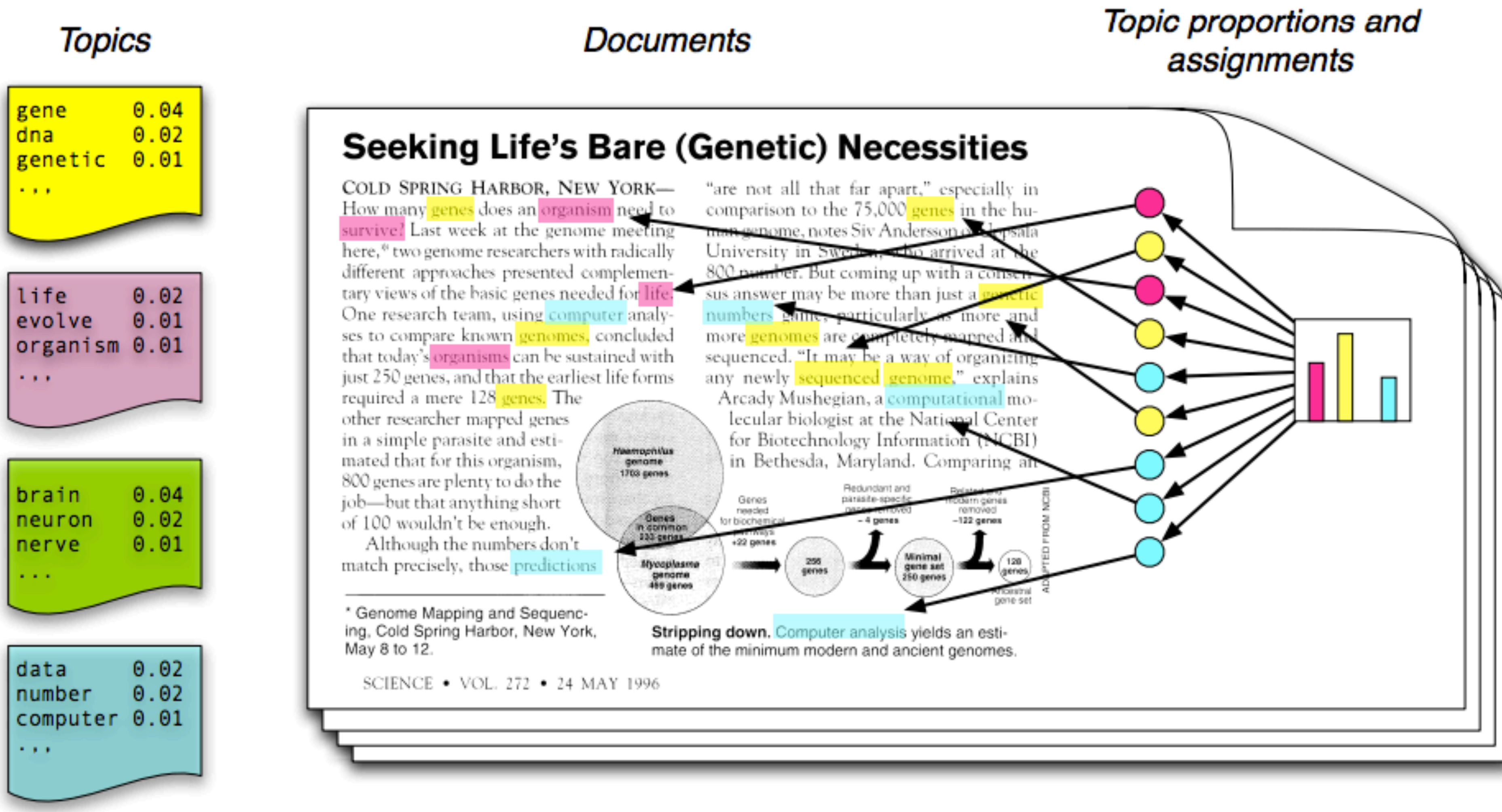


attorney	air	united
federal	flight	states
charged	plane	trade
filed	airlines	nations
court	pilots	world
fraud	eastern	countries
indictment	airline	european
investigation	airport	international

voters	nicaragua	sales
vote	government	percent
votes	rebels	billion
campaign	contras	share
democratic	sandinista	quarter
candidate	ortega	earnings
state	sandinistas	last
election	chamorro	first
...	...	...

from: http://www.cs.princeton.edu/~blei/lda-c/ap-topics.pdf

Topic Modeling

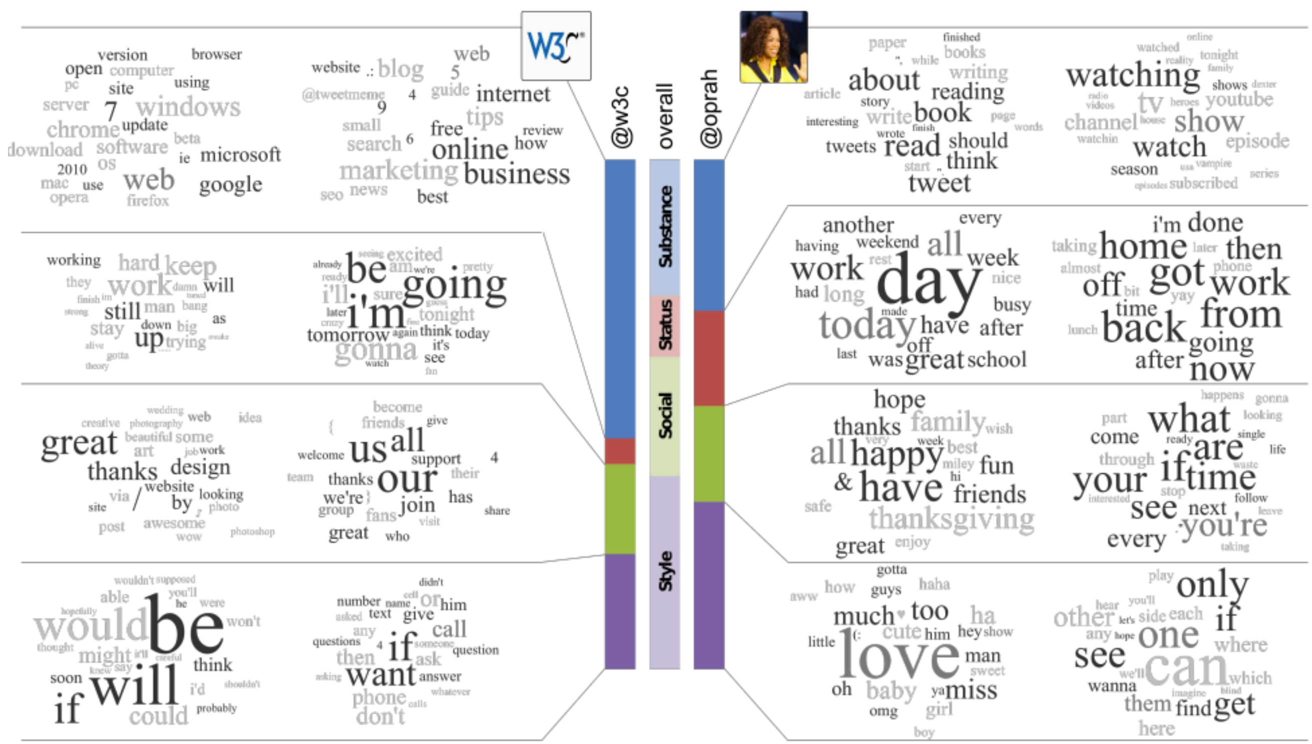


David Blei (2012) Probabilistic Topic Models. *CACM*.

Topic Modeling: Uses

- Identifying main themes in text

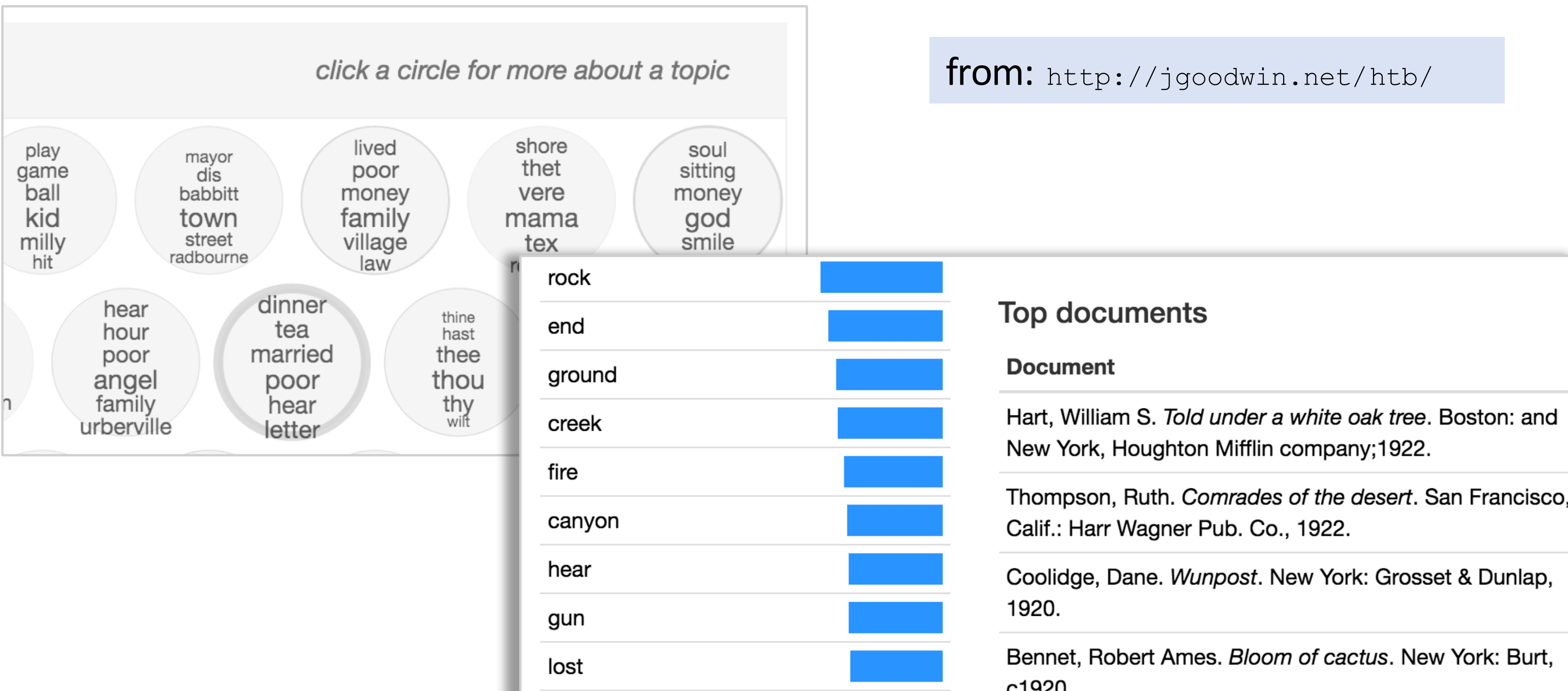
"What do people talk about on Twitter?"



Ramage. Dumais. Liebling (2010) Characterizing Microblogs with Topic Models. *ICWSM*.

Topic Modeling: Uses

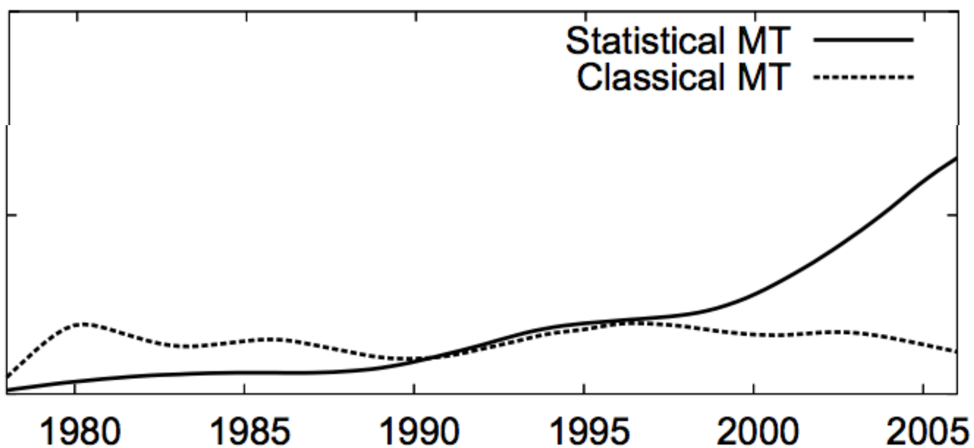
- Identifying main themes in text
- Exploring/navigating corpora



from: `http://jgoodwin.net/htb/`

Topic Modeling: Uses

- Identifying main themes in text
- Exploring/navigating corpora
- Revealing trends in content



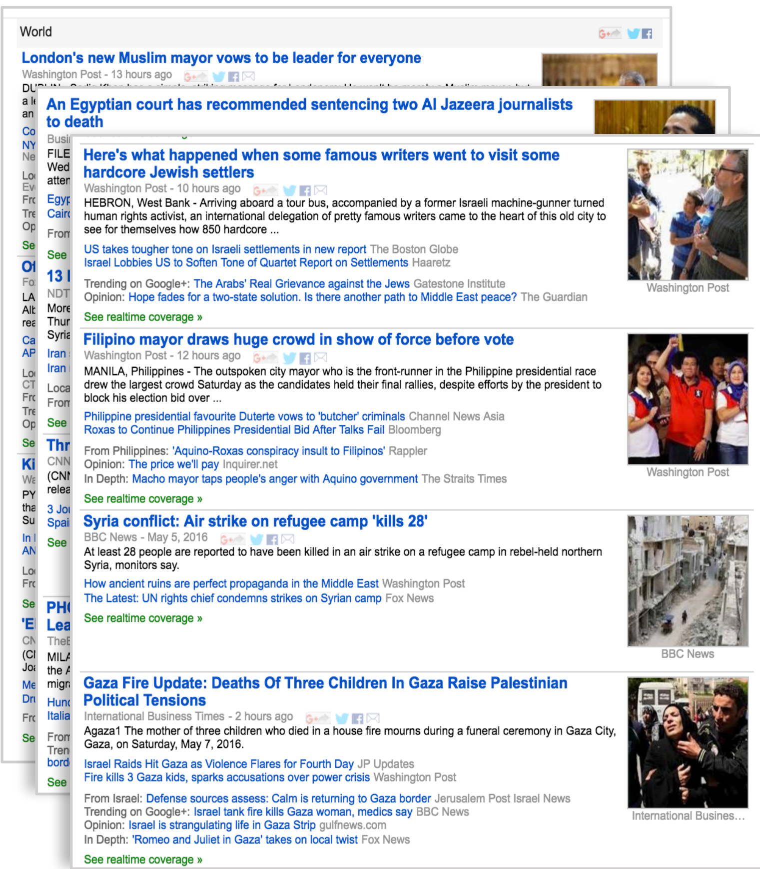
Hall, Jurafsky, Manning (2008) Studying the History of Ideas Using Topic Models. *EMNLP*.

Topic Modeling: Uses

- Identifying main themes in text
- Exploring/navigating corpora
- Revealing trends in content

These uses assume topic models are human-interpretable

Interpretability in Topic Models



Topics aren't always easy to decipher…

door	hes	ms
peace	i	thompson
table	frank	day
i	years	jackpot
california	time	hospital
day	good	eight
hand	tree	pm
fast	calls	time
…	…	…

from: http://www.cs.princeton.edu/~blei/lda-c/ap-topics.pdf

Interpretability in Topic Models

How does the topic modeling community deal with interpretability?

Two key areas:

- **Evaluation**
- **Training**

Interpretability in Topic Models: Evaluation

How to evaluate the quality of a topic?

Commonly used concept: *coherence*

Coherent	
space	health
earth	disease
moon	aids
science	virus
scientist	vaccine
light	infection
nasa	hiv
mission	cases
planet	infected
mars	asthma

Incoherent	
dog	king
moment	bond
hand	berry
face	bill
love	ray
self	rate
eye	james
turn	treas
young	byrd
character	key

Newman, Lau, Grieser, Baldwin (2010) Automatic Evaluation of Topic Coherence. *NAACL*.

Interpretability in Topic Models: Evaluation

How to evaluate the quality of a topic?

- Coherence

- Human judgments: *word intrusion*
  - Spot the out-of-place word:

    nasa
    mission
    planet
    dog
    mars

Chang, Gerrish, Wang, Boyd-Graber, Blei (2009) Reading Tea Leaves: How Humans Interpret Topic Models. *NIPS*.

Interpretability in Topic Models: Evaluation

How to evaluate the quality of a topic?

- Coherence

- Human judgments: *word intrusion*
- Metrics based on co-occurrence statistics
  - Similarity of all pairs of words in a topic

$$C(t; V^{(t)}) = \sum_{m=2}^{M} \sum_{l=1}^{m-1} \log \frac{D(v_m^{(t)}, v_l^{(t)}) + 1}{D(v_l^{(t)})}$$

Mimno, Wallach, Talley, Leenders, McCallum (2011)
Optimizing Semantic Coherence in Topic Models. *EMNLP*.

Interpretability in Topic Models: Training

How to train topic models in a way that will be more interpretable?

- Priors to incorporate human preferences/constraints

  Andrzejewski, Zhu, Craven (2009) Incorporating Domain Knowledge into Topic Modeling via Dirichlet Forest Priors. *ICML*.

- Priors to encourage co-occurrence patterns

  Mimno, Wallach, Talley, Leenders, McCallum (2011) Optimizing Semantic Coherence in Topic Models. *EMNLP*.

- Interactive topic modeling

  Hu, Boyd-Graber, Satinoff, Smith (2013) Interactive Topic Modeling. *Machine Learning*.

Beyond Topic Modeling

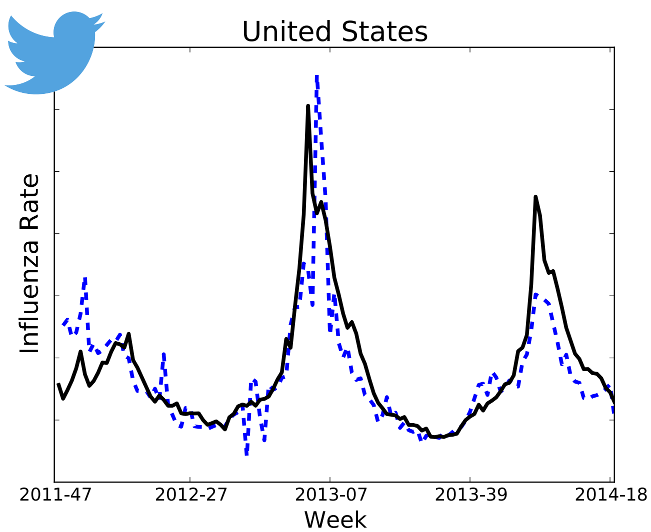
Can other areas of machine learning benefit from the topic modeling approach to interpretability?

Is coherence a desirable property in other types of models?

Beyond Topic Modeling

Poor coherence = poor predictive performance?

Example: **Modeling flu prevalence from Twitter**



christmas
sick
flu
strong
processing
snow
new
want
hard
better
body
best
coughing
festivities
eve

Top predictors of regression model

Conclusion

Takeaway:

The topic modeling community has a growing body of research on making models more interpretable

This research should be incorporated into the larger context of interpretable machine learning

What's next?

- Coherence as an evaluation metric for other machine learning models
- Training models to be coherent