Topics Over Time: Into Darwin's Mind

Jaimie Murdock

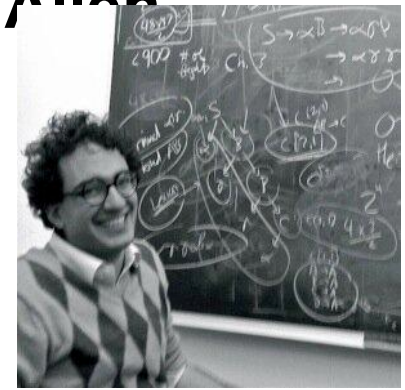
NetSci @ IU Talks

March 9, 2014

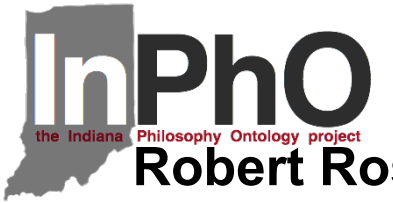
Slides: http://jamr.am/DarwinIUNetSci

Acknowledgements

Simon DeDeo
Allen



Colin



Robert Rose – VSM module
Tom Murphy – Data entry

Slides: http://jamr.am/DarwinIUNetSci

Presentation Goals

What will be discussed?

- Why study Darwin?
  - Individual vs. Collective SoS
- Methods talk for computational humanities.
  - HathiTrust
  - LDA Topic Modeling
  - Ordered Corpora
    - Lag, Burst, Focus
- Darwin's Reading Habits
- Darwin & Wallace's Writing

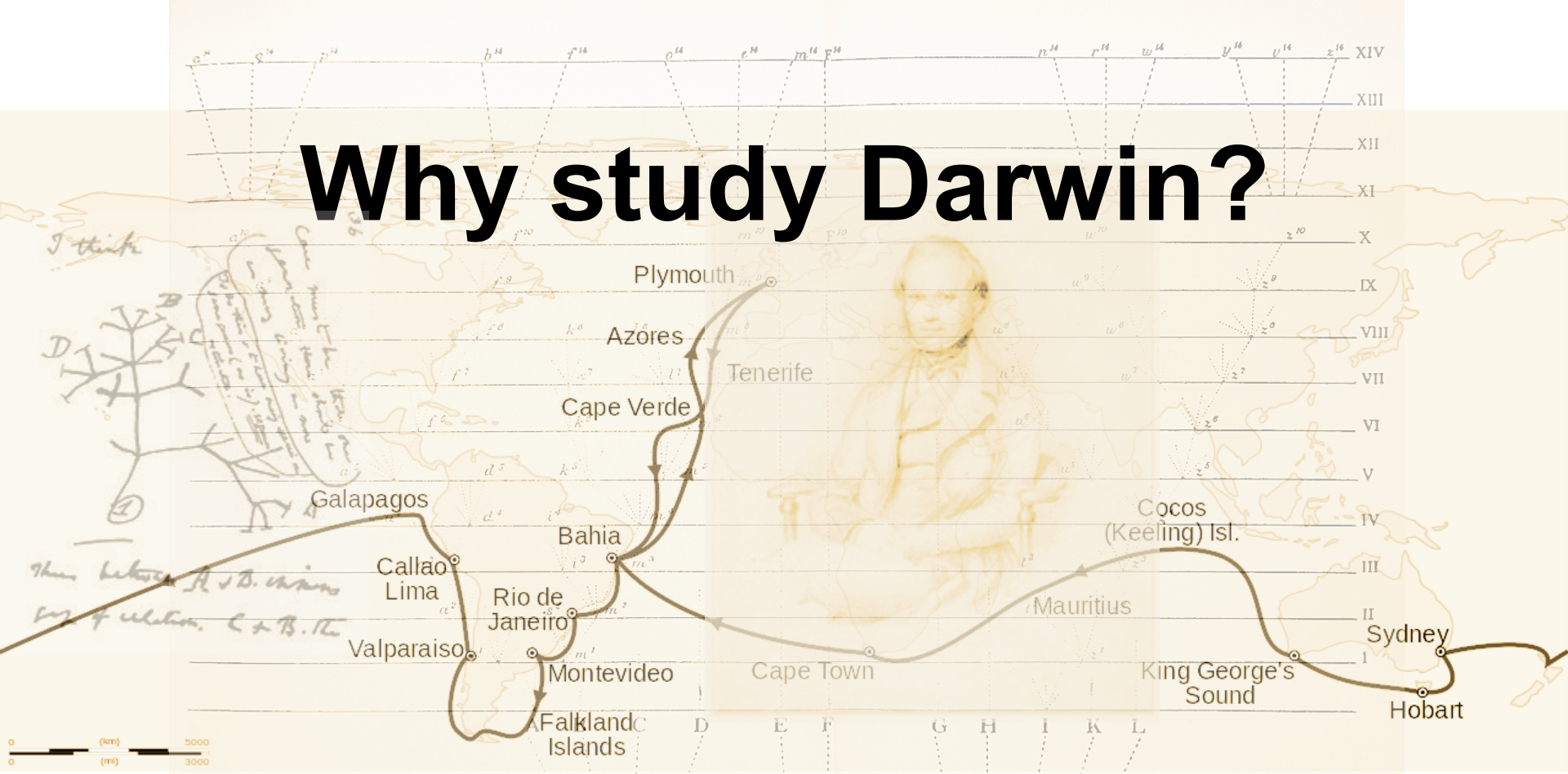
Take Away Messages

- Many ways to explore topics
- New tools enable rapid analysis
- Topic models can detect "facts" about reading structure

Work-in-progress talk!

Slides: http://jamr.am/DarwinIUNetSci

Why study Darwin?



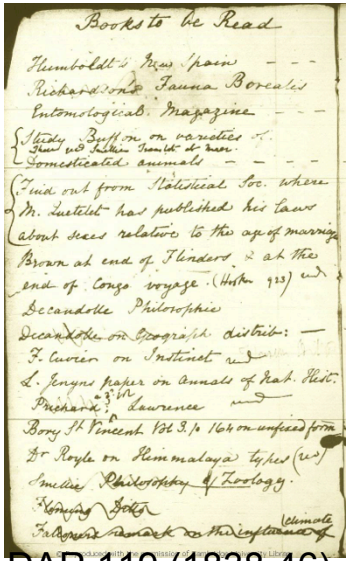
Slides: http://jamr.am/DarwinIUNetSci

Charles Darwin (1809-1882)

Abridged Timeline:

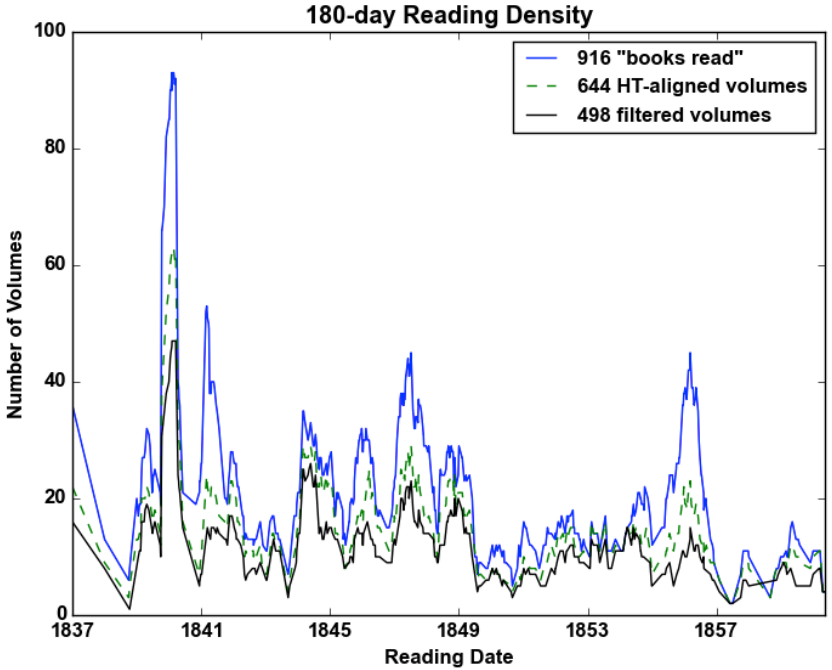
1809 – Born
1825-27 – Med school (Edinburgh)
1828-31 – Undergrad (Cambridge)
1831-36 – Voyage of the Beagle
1838 – Beagle *Journal & Remarks* pub'd
1842-44 – First *Origin* Abstracts
1854 – Barnacles
1858 – Joint *Species* pub w/Wallace
1859 – *Origin* published
1871 – *Descent of Man* published
1882 – Death

Darwin’s Reading Notebooks



- DAR 119 (1838-46)
- DAR 128 (1846-60)
- DAR *128 indicates “to be read”, excluded
  - 1248 total citations
  - 916 “read” citations

Retrieved from [Darwin Corresponence Project]

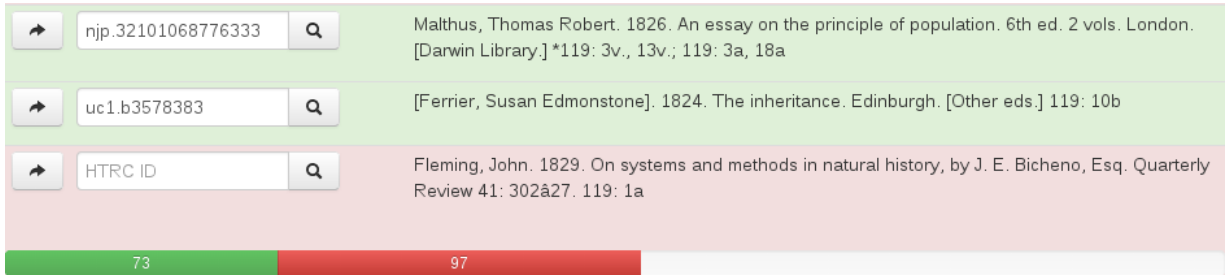


498/916 Filtered Sample

- English non-fiction
- LDA models: 500 iter, 20-80 topics, nltk stopwords; freq filter: 5 < n < 12k

HATHI TRUST Digital Library & InPhO Corpus Builder

- Darwin's Reading Notebooks citations identified at Darwin Correspondence Project
- Parsed using Anystyle.io
- Aligned using InPhO Corpus Builder



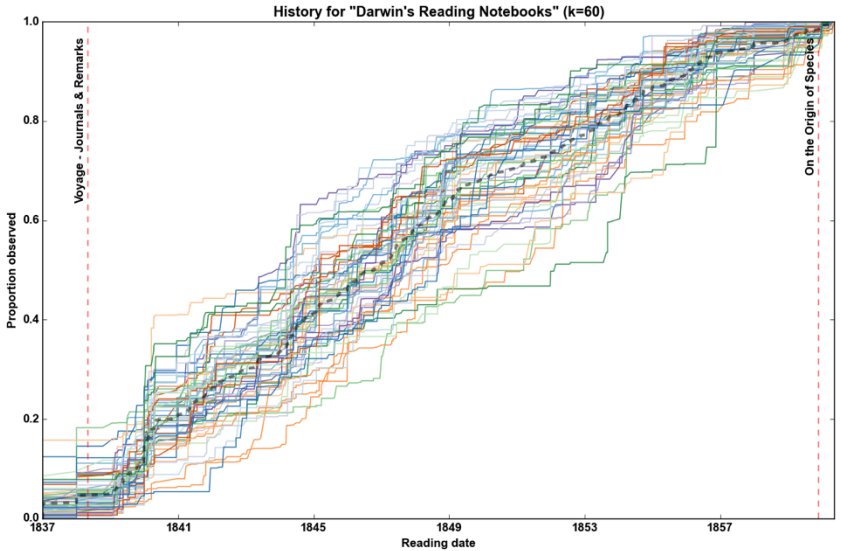
Source: https://github.com/inpho/corpus-builder

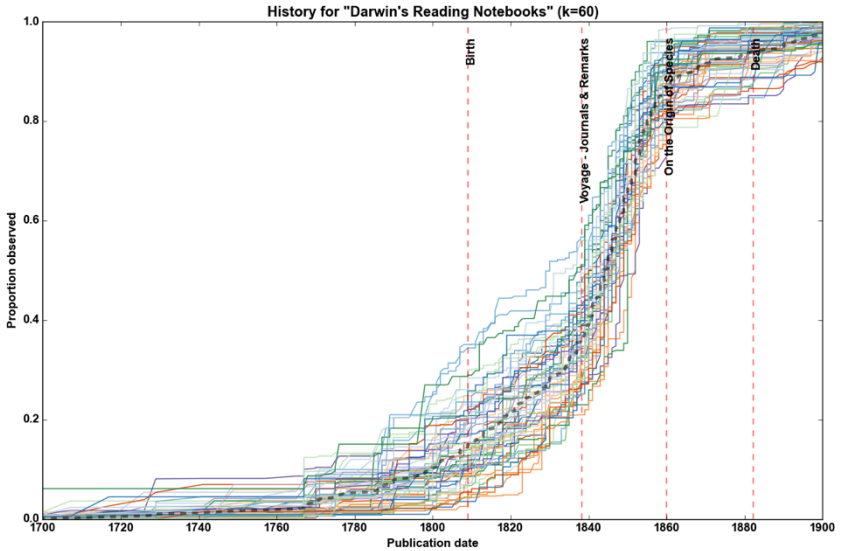
- **Fulltext access to 644 volumes Darwin read**

“The interplay between individual and collective phenomena where innovation takes place.” — The Dynamics of Correlated Novelties

Reading Date

Publication Date





Slides: http://jamr.am/DarwinIUNetSci

Methods

Topic Modeling
Visualization & Evaluation
Characterizing Topics
Lag, Burst, Focus

Latent Dirichlet Allocation (LDA)

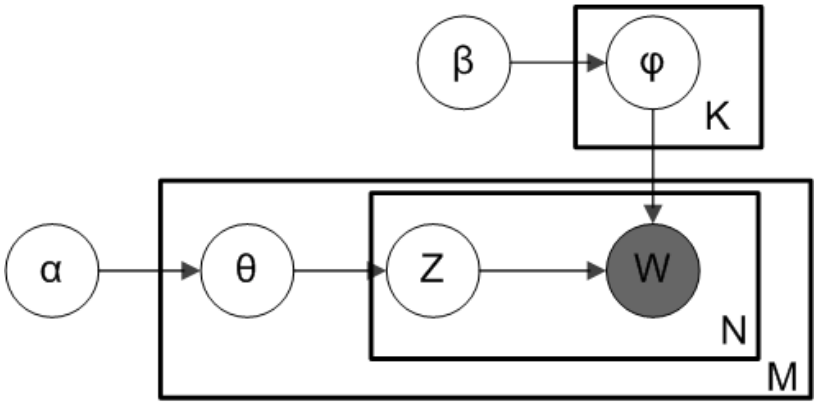


Plate Notation for Smoothed LDA

- $\boldsymbol{\alpha}$ – Dirichlet prior for per-doc topic dist
- $\boldsymbol{\beta}$ – Dirichlet prior for per-topic word dist
- $\boldsymbol{\theta}_i$ – topic distribution for doc *i*
- $\boldsymbol{\Phi}_j$ – word distribution for topic *k*
- $\boldsymbol{z}_{i,j}$ – topic for *j*th word in doc *i*
- $\boldsymbol{w}_{i,j}$ – the actual word

Generative Model

1. Choose $\boldsymbol{\theta}_i$ **~ Dir(**$\boldsymbol{\alpha}$**)** (*i* is doc)
2. Choose $\boldsymbol{\Phi}_k$ **~ Dir(**$\boldsymbol{\beta}$**)** (*k* is topic)
3. For each word position
   a. Choose a topic
      $\boldsymbol{z}_{i,j}$ **~ Multinomial(**$\boldsymbol{\theta}_i$**)**
   b. Choose a word
      $\boldsymbol{w}_{i,j}$ **~ Multinomial(**$\boldsymbol{\Phi}_{z\_i,j}$**)**

Training on a Corpus

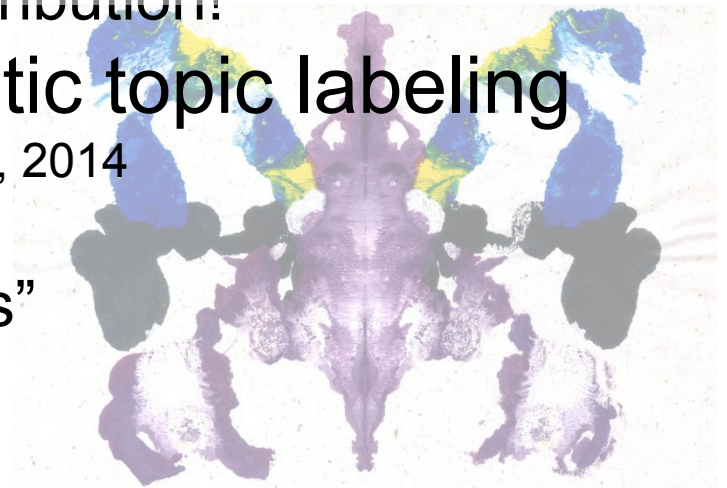
Bayesian inference on $\boldsymbol{\theta}$ and $\boldsymbol{\Phi}$

Training on a new Document (d)

Fix P(w|z) to infer P(z|d)

Topic Interpretation

- Topics are *distributions* over words.
  - Often displayed as top N words.
  - Remember: all words in corpus are in every topic distribution!
- Methods exist for automatic topic labeling
  - Aletaras et al, 2014, Sievert & Shirley, 2014
  - Human interpretation
    - “Reading the Tea Leaves”



Topic Explorer

Live demo: http://inphodata.cogs.indiana.edu:21020/
Source: http://github.com/inpho/topic-explorer/

Evaluating Topic Models

Log Likelihood

$$\mathcal{L}(w) = \log p(w, z|\phi, \theta)$$

Perplexity per Word

$$\text{Perplexity}(w) = \exp\left\{-\frac{\mathcal{L}(w)}{|\text{tokens}|}\right\}$$

N-way Jensen-Shannon Divergence

$$\text{KL}(P|Q) = \sum_i P(i) \log \frac{P(i)}{Q(i)}$$

$$\text{JS}_{div}(D) = \sum_i^{|D|} \frac{1}{|D|} \text{KL}(D_i|M)$$

$$M = \sum_i^{|D|} \frac{1}{|D|} D_i$$

Darwin Model Eval

k	JS_{div}	$\mathscr{L}(w)$	$\mathscr{L}_{avg}(w)$	Perp
20	2.081	-2.575e8	-10.810	49518
40	2.737	-2.571e8	-10.793	48667
60	3.048	-2.565e8	-10.769	47544
80	3.270	-2.556e8	-10.730	45695

Characterizing Topics

Oscillation

$$\mathrm{Osc}(k) = \max(\theta_{i,k}) - \min(\theta_{i,k})$$

Entropy by Word

$$H_w(k) = -\sum_w p(w|k) \log p(w|k)$$

Lag

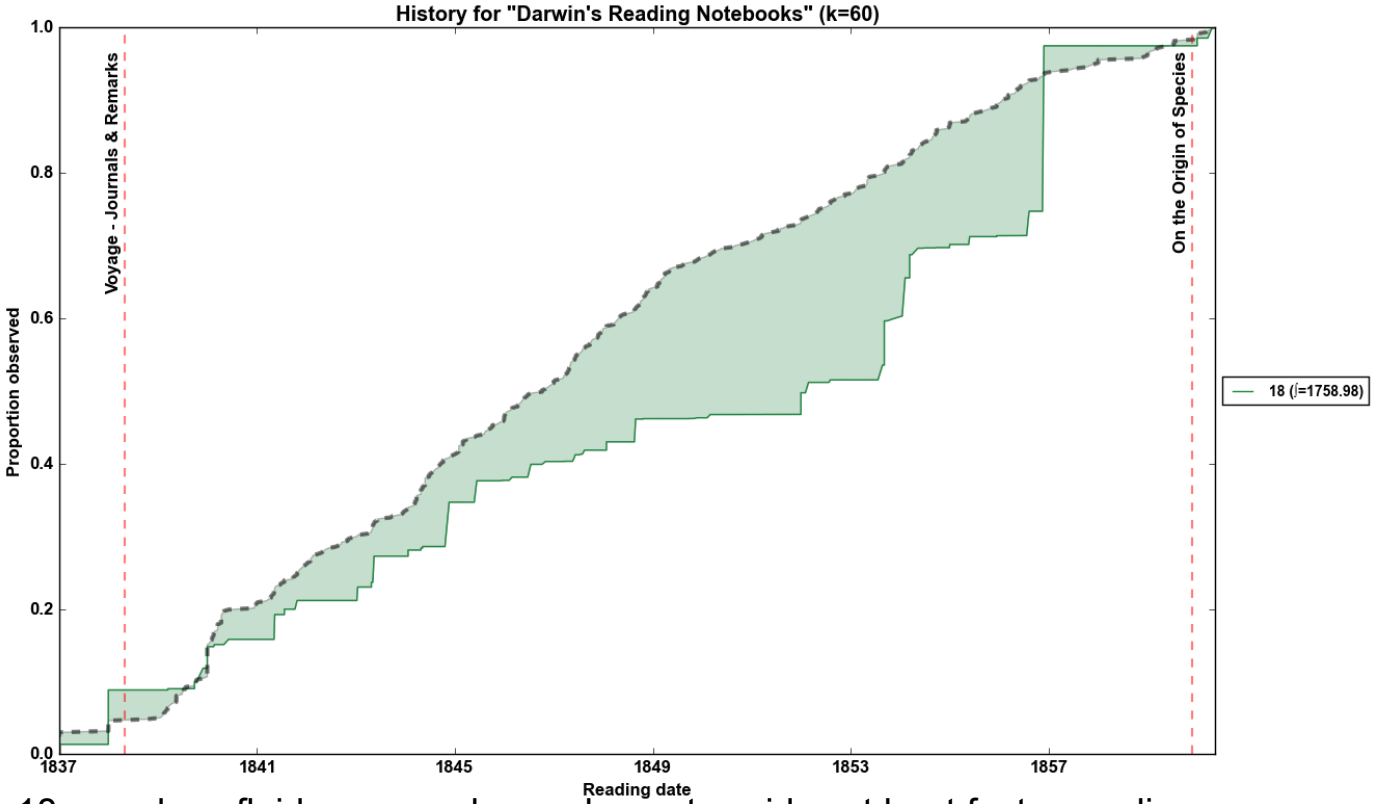
$$\mathrm{Lag}(k) = \int p_D(t) - p_k(t) dt = \sum_t t \cdot \mathrm{p}_t(D) - \mathrm{p}_t(k)$$

Burst

$$\mathrm{Burst}(k) = \int |p_D(t) - p_k(t)| dt = \sum_t t \cdot |\,\mathrm{p}_t(D) - \mathrm{p}_t(k)|$$

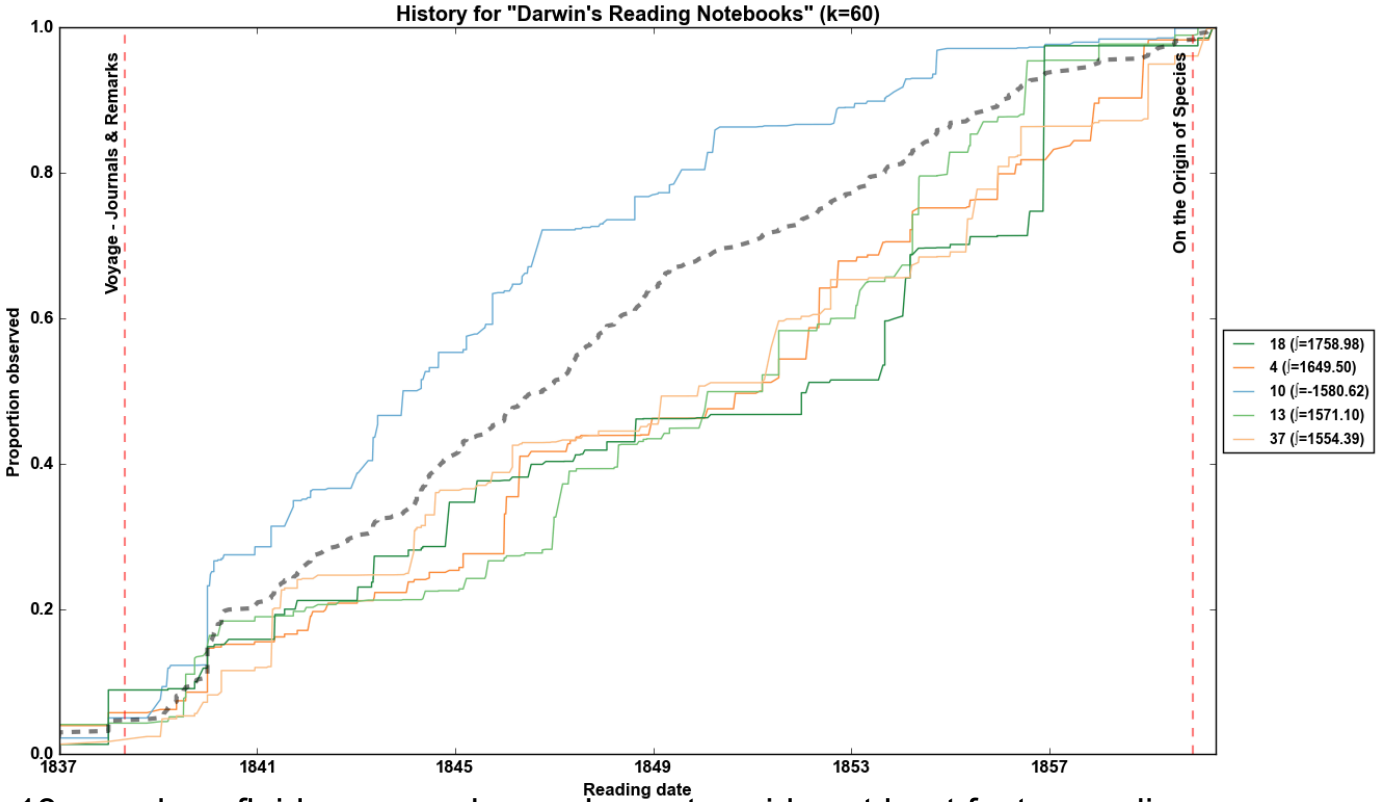
Topics Sorted by Oscillation	
Topic	**Words**
Topic 28	genera, norway, limit, indian, creek, 10, temperature, plant, indians, 12
Topic 38	peace, states, read, bill, parliament, sydney, beds, lake, york, correspondence
Topic 21	beauty, mean, fo, provinces, cicero, bear, counties, native, distribution, area
Topic 0	les, plus, pl, cette, valve, meme, deux, coquille, espece, tres
Topic 18	produce, fluid, progress, bone, elements, acid, sect, heat, facts, peculiar

Topics Sorted by Entropy	
Topic	**Words**
Topic 45	parliament, india, wool, major, tlie, june, governor, camp, military, political
Topic 39	fly, particularly, trout, lieut, valley, hook, follen, fishing, sport, practice
Topic 51	india, felt, hills, pitt, cape, written, hours, season, going, pleasure
Topic 20	flesh, yellow, medium, juicy, roundish, stalk, rich, vigorous, calyx, tender
Topic 25	henry, cancer, tumors, skin, muscles, organic, cases, nott, tumor, vessels



Lag

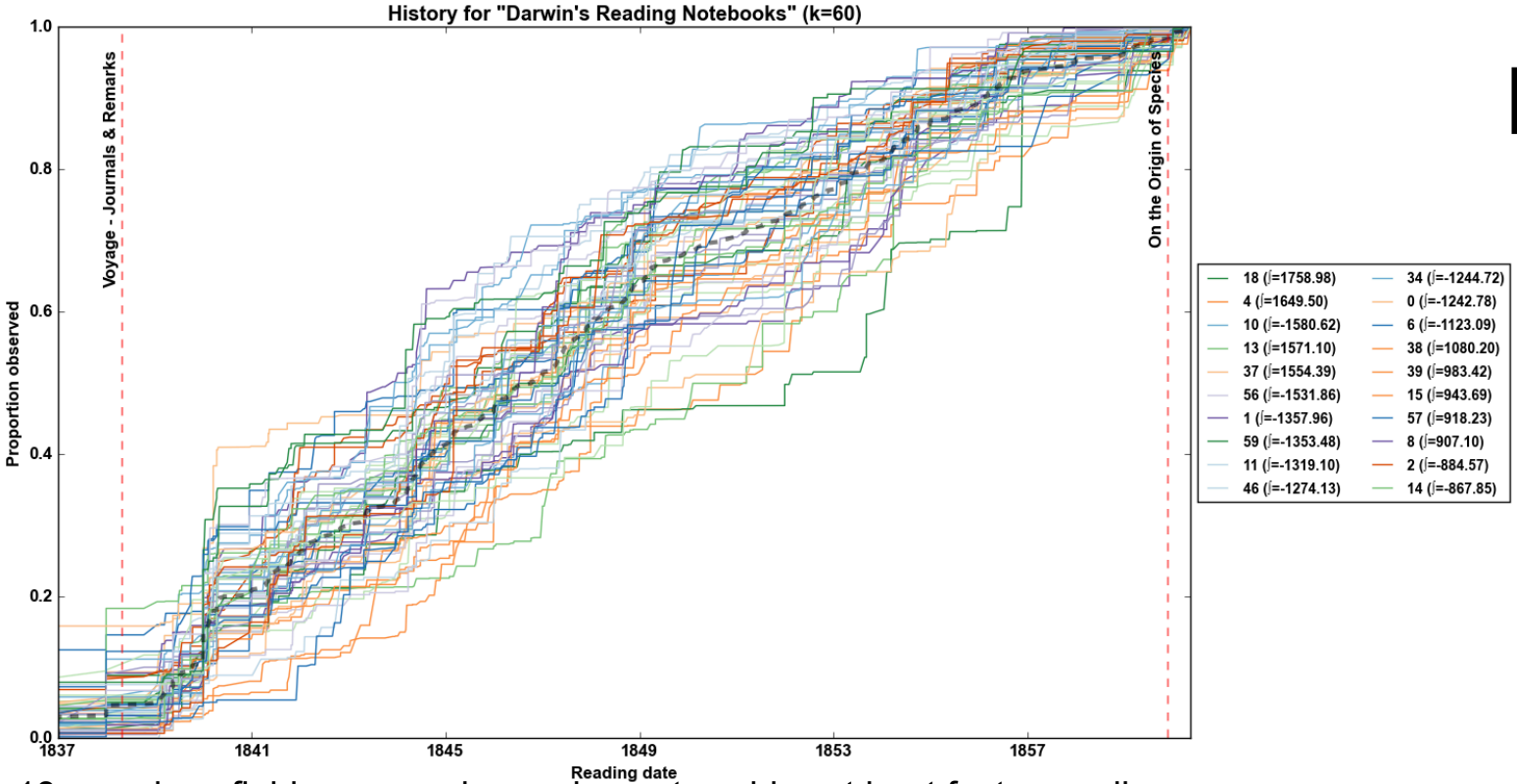
18 - produce fluid progress bone elements acid sect heat facts peculiar



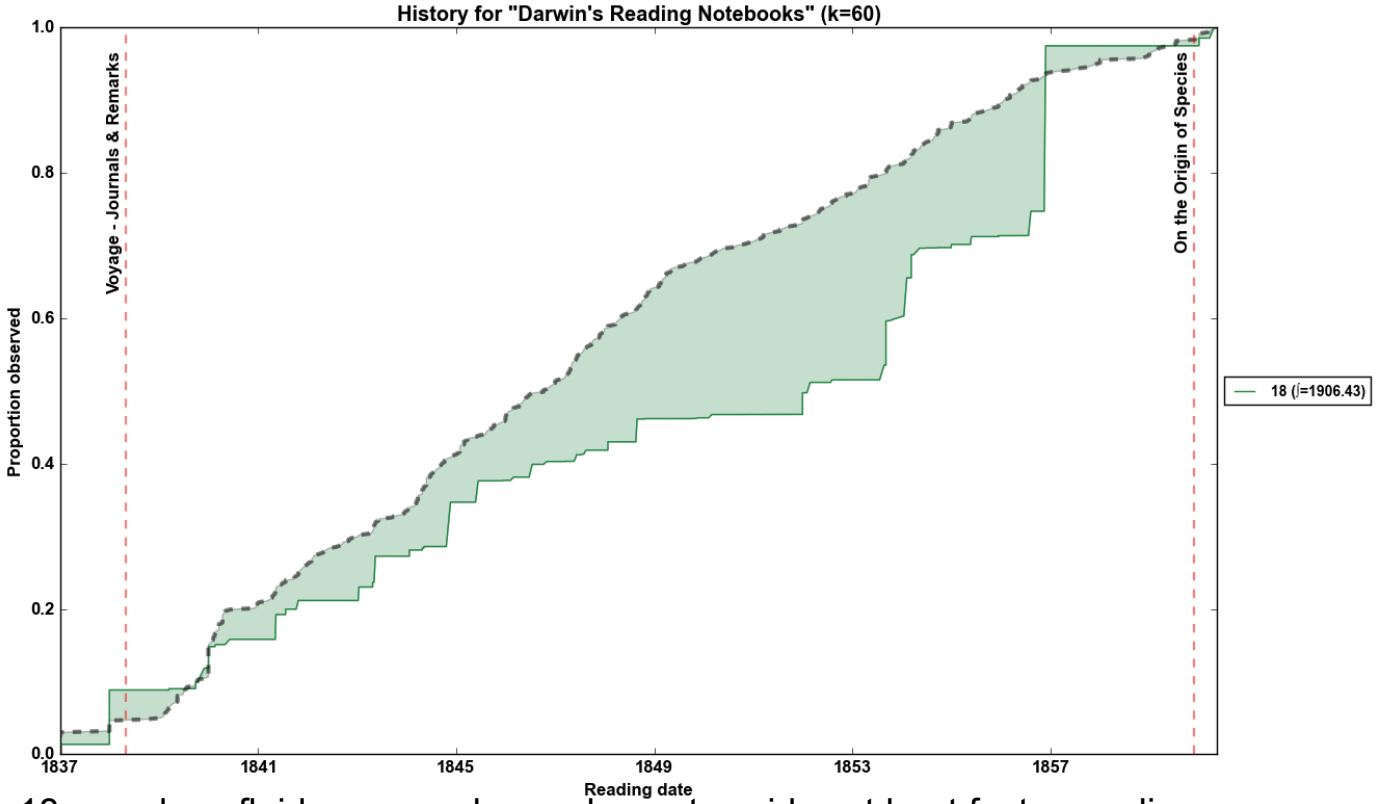
Lag

18 - produce fluid progress bone elements acid sect heat facts peculiar
4 - race dog egyptian bor alva bounderby fowl hoofd orange diary
10 - travels fome dogs hunting love les mrs napoleon lapland gun
13 - bird pigeons cuv mozart stomach q structure dorsal nerve pl
37 - cattle breed quantity milk esq yellow disease base cow africa

Lag



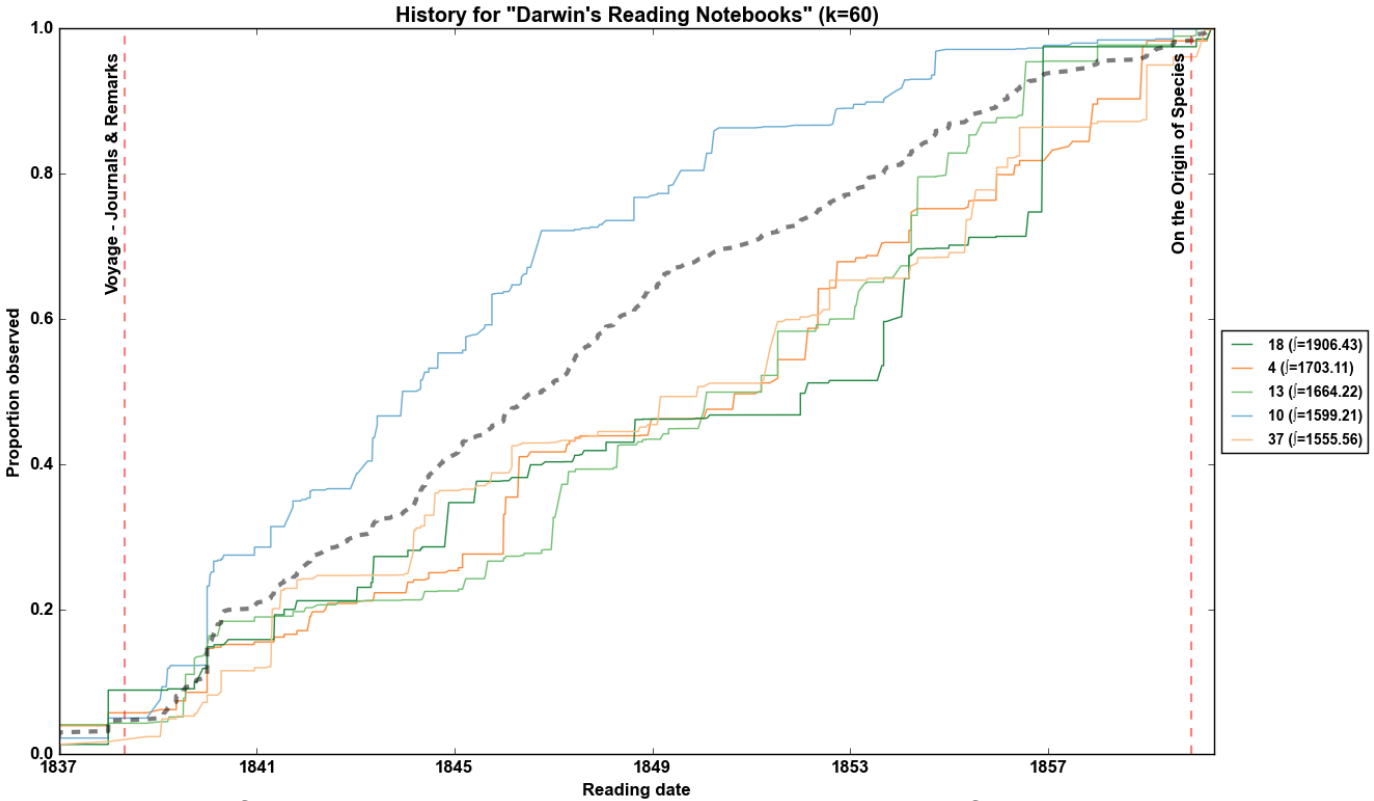
18 - produce fluid progress bone elements acid sect heat facts peculiar
4 - race dog egyptian bor alva bounderby fowl hoofd orange diary
10 - travels fome dogs hunting love les mrs napoleon lapland gun
13 - bird pigeons cuv mozart stomach q structure dorsal nerve pl
37 - cattle breed quantity milk esq yellow disease base cow africa



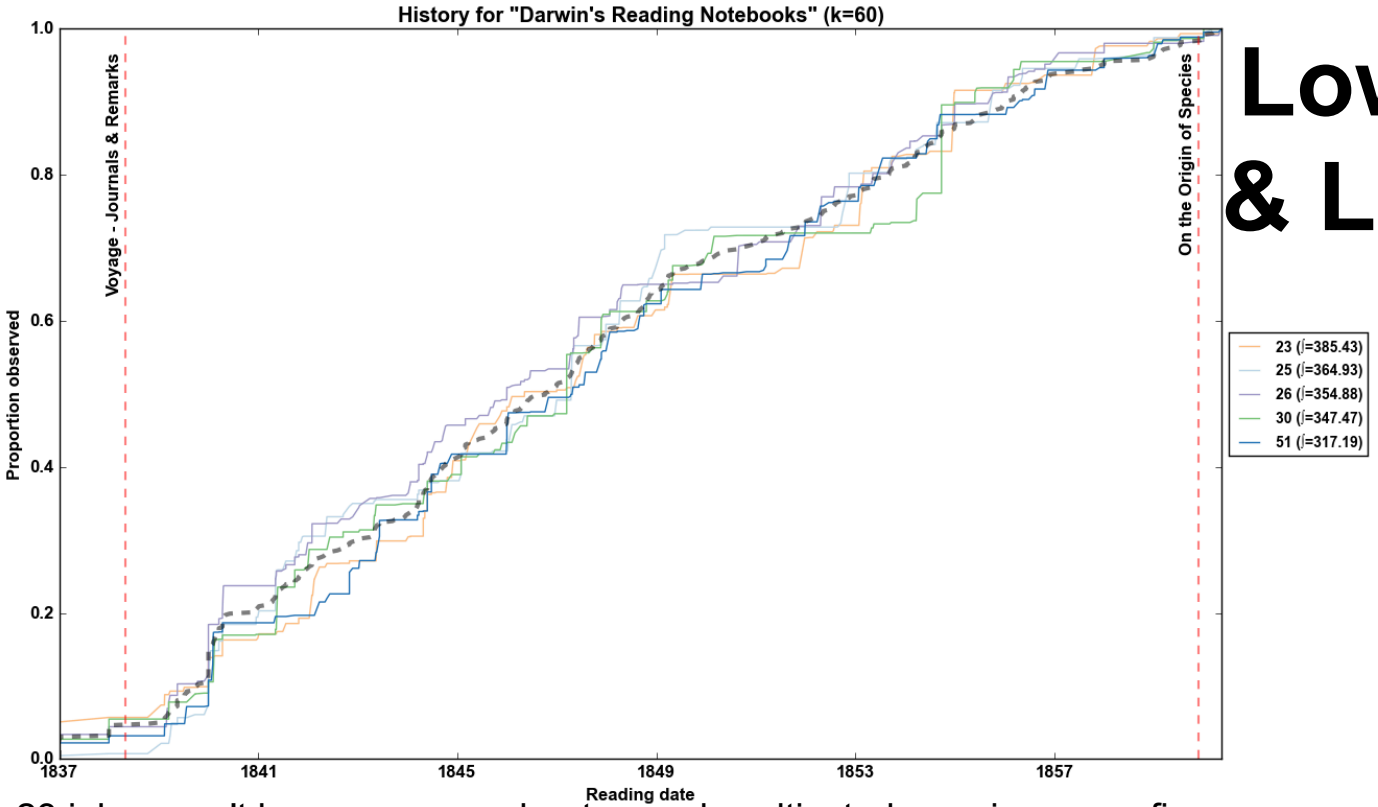
Burst

18 - produce fluid progress bone elements acid sect heat facts peculiar

Burst

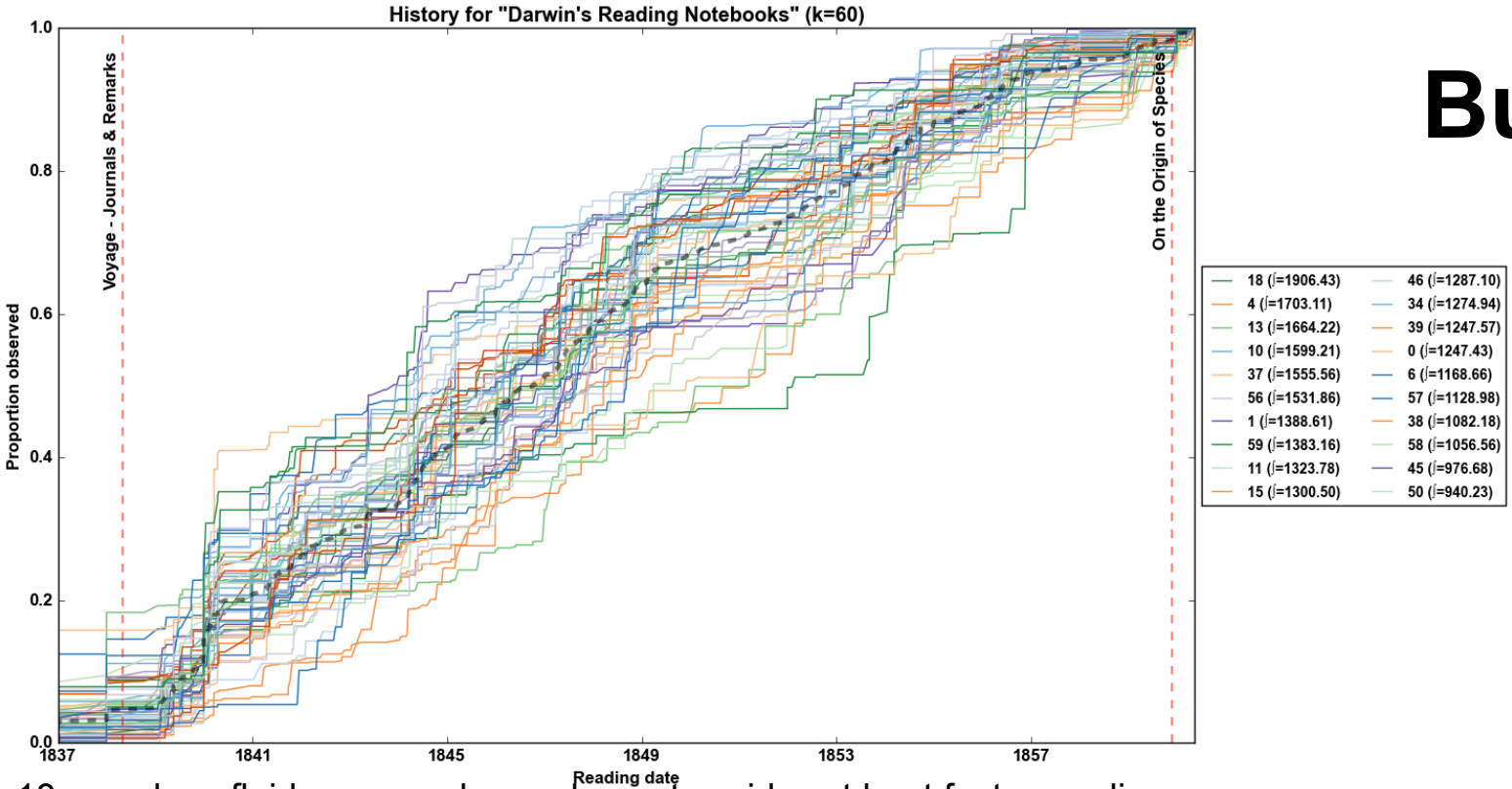


18 - produce fluid progress bone elements acid sect heat facts peculiar
4 - race dog egyptian bor alva bounderby fowl hoofd orange diary
13 - bird pigeons cuv mozart stomach q structure dorsal nerve pl
10 - travels fome dogs hunting love les mrs napoleon lapland gun
37 - cattle breed quantity milk esq yellow disease base cow africa



Low Burst & Low Lag

23 johnson n't hungary snow chapter yards cultivated america corps fire
25 henry cancer tumors skin muscles organic cases nott tumor vessels
26 child mrs count henry lady james dear smith sect art
30 sp insects fishes pi linn abdomen larva genus palpi antennae
51 india felt hills pitt cape written hours season going pleasure



Burst

18 - produce fluid progress bone elements acid sect heat facts peculiar
4 - race dog egyptian bor alva bounderby fowl hoofd orange diary
13 - bird pigeons cuv mozart stomach q structure dorsal nerve pl
10 - travels fome dogs hunting love les mrs napoleon lapland gun
37 - cattle breed quantity milk esq yellow disease base cow africa

Inquiries

Operationalizing

Directed Search

Do Darwin's readings exhibit directed search?

- Directed to what?
  i. Directed to the creation of *The Origin*
  ii. Darwin had sketch in 1842 and 1844,
  didn't start writing the "abstract" until 1856
  never published the "full" theory
  iii. Can we gain insight into how theory evolved?

Navigating Topic Space

Kullback-Leibler Divergence (KL):

$$\mathrm{KL}(P|Q) = \sum_i P(i) \log \frac{P(i)}{Q(i)}$$

Jensen-Shannon Divergence (JSD):

$$\mathrm{JS}_{div}(P|Q) = \frac{1}{2}\mathrm{KL}(P|M) + \frac{1}{2}\mathrm{KL}(Q|M)$$

$$M = \frac{P+Q}{2}$$

Jensen-Shannon Distance (JS Distance):

$$\mathrm{JS}_{dist}(P|Q) = \sqrt{\mathrm{JS}_{div}(P|Q)}$$

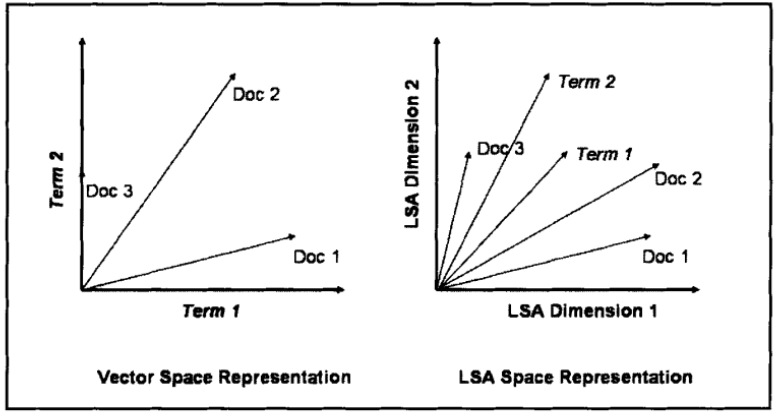


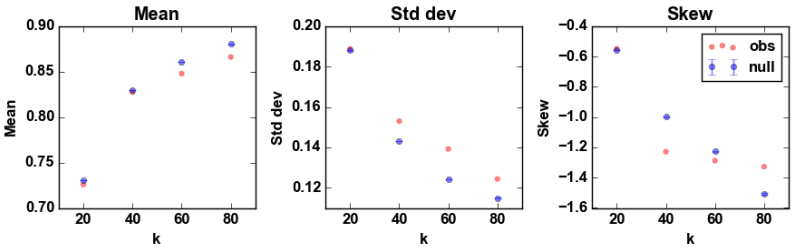
Figure 4.1 Comparison of Vector Space (left) and LSA (right) representations

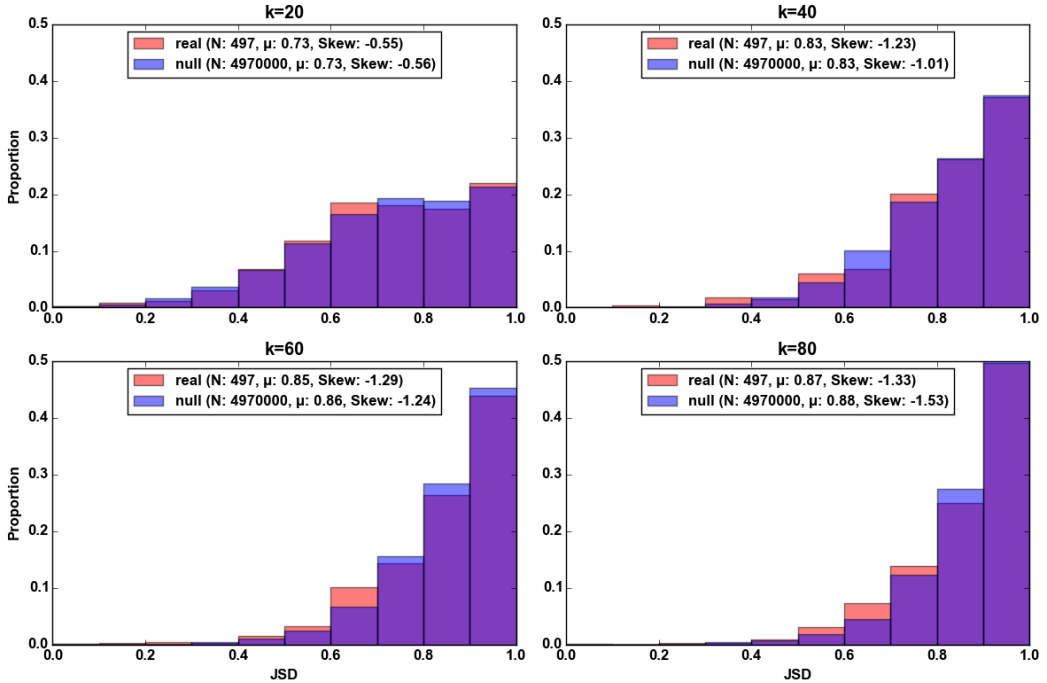
Graphic from Dumais, *LSA,* 2005

Null Reading Models

Randomly shuffle 498 volumes, 10,000 times

Compare distributions

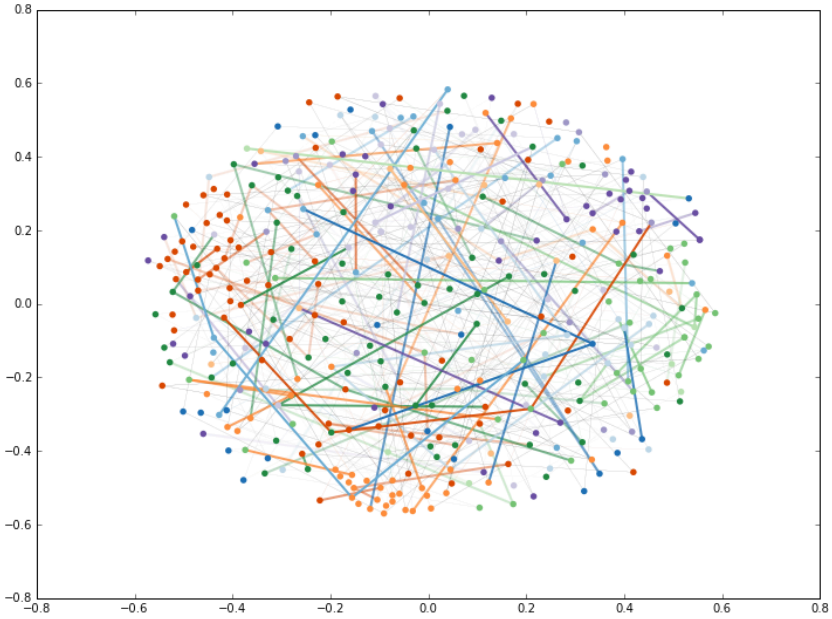


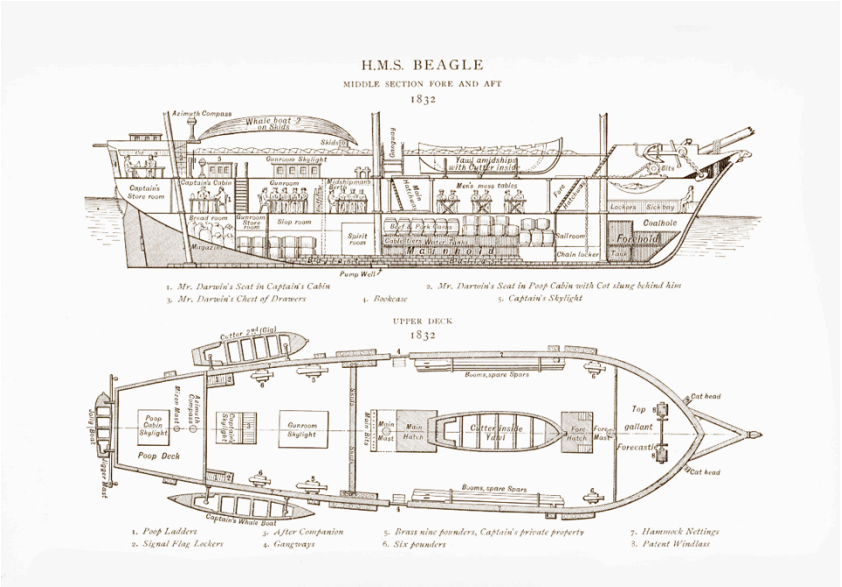


Slides: http://jamr.am/DarwinIUNetSci

The Travelling Reader Problem

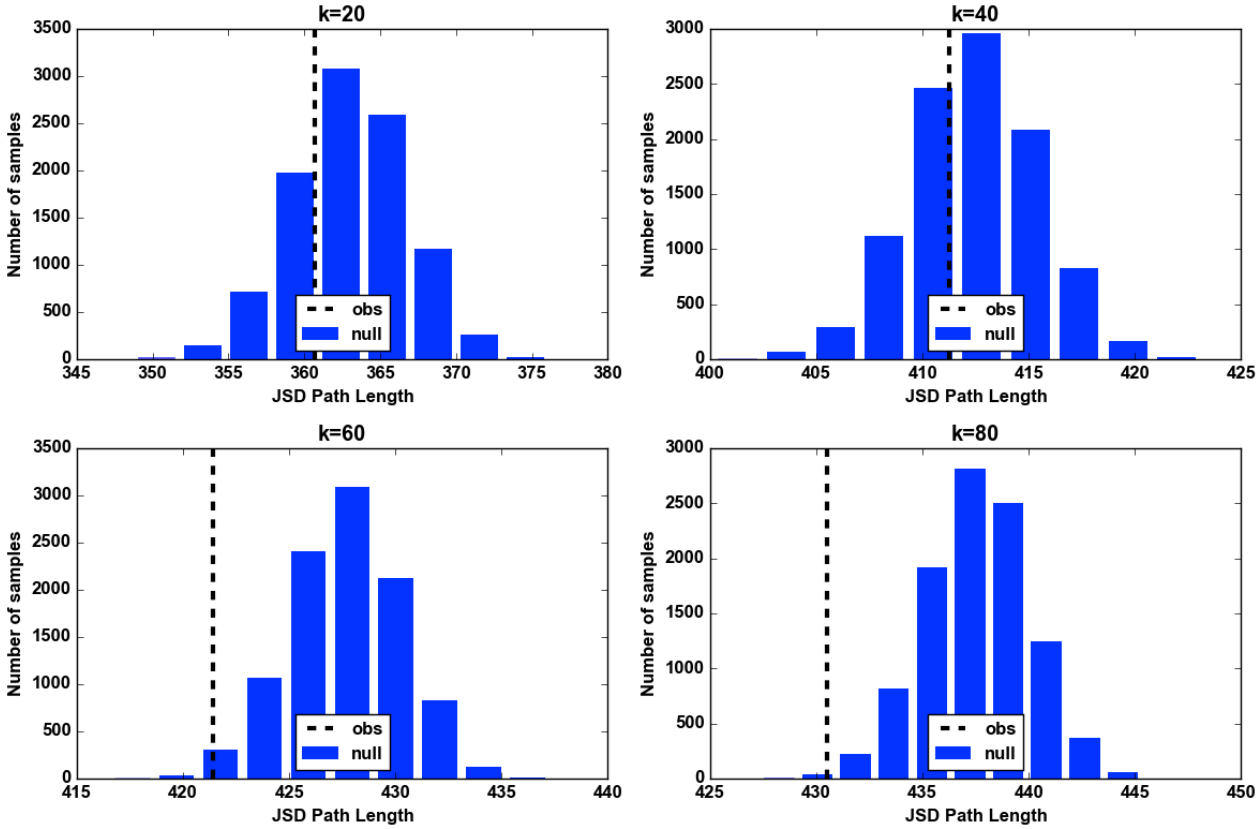
Was Darwin's path through topic space efficient?





Darwin had 42 linear-feet
of library space on the Beagle

Reading Path Length



Focus

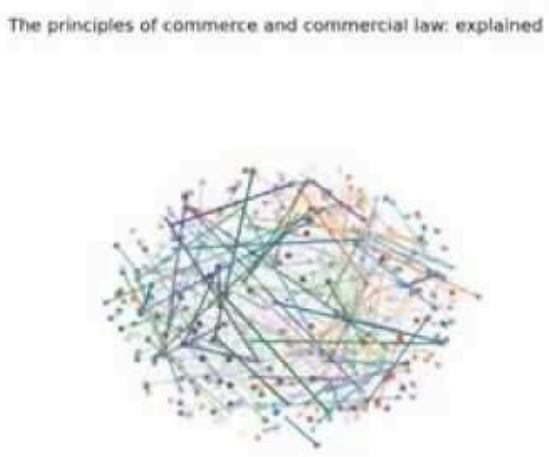
For each volume:

1. Take JSD to all other vols
2. If JSD(v_t,v_{t+1}) < 95% conf.,
   reading is in “focus”

Focus windows (v_i,v_j) has
agg topic proportion

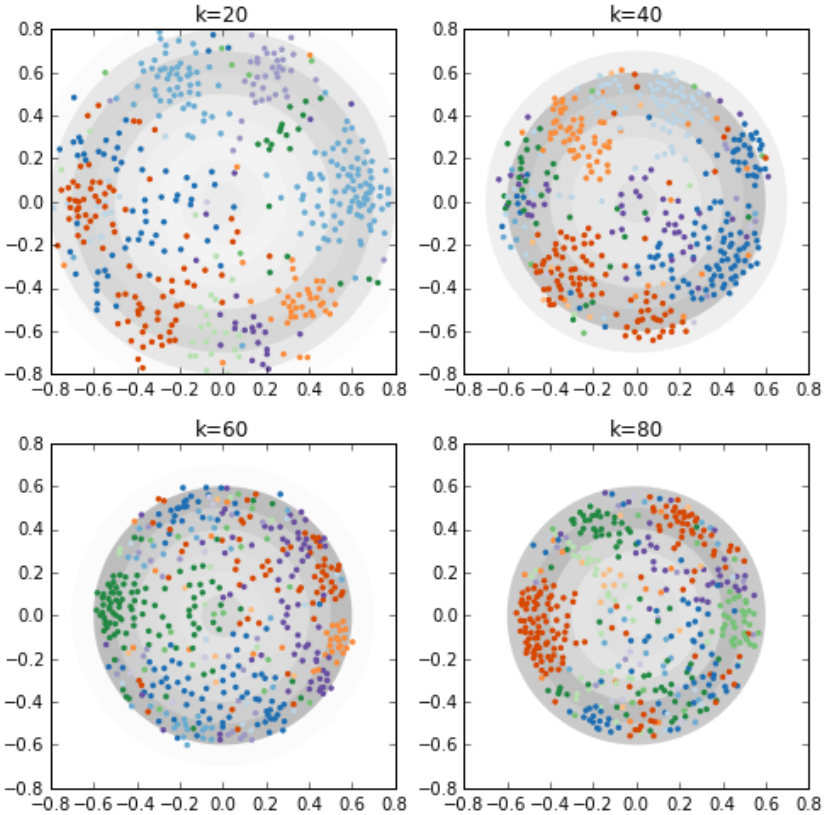
Topic-driven Foraging

- Goldstone, Todd, Landy Lab
- Friday, April 10 — 9-10a
- MSB II Gill Conference Room



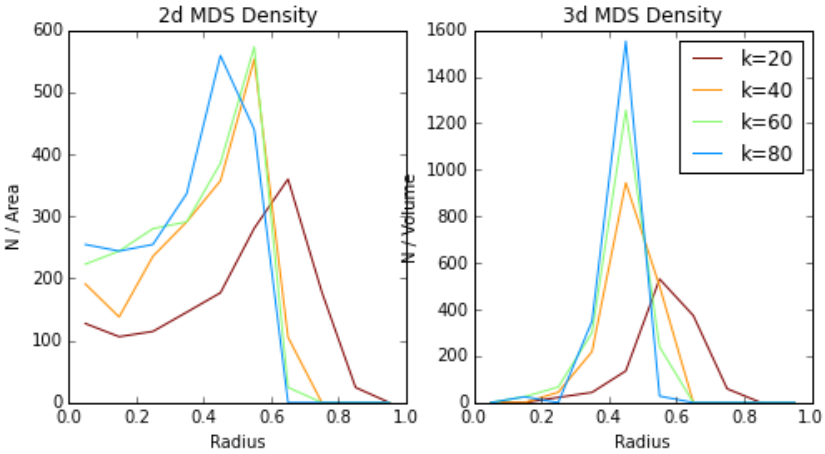
Dots show greatest proportionate topic
Solid lines show focus by largest agg topic

An Unusual Structure?



2D & 3D MDS show ring/ shell structure

more to come…



Reading & Writing

Use LDA Query Sampling on:

- Essay of ‘42
- Essay of ‘44
- *On the Origin of Species* (1st ed)
- *On the Tendency of Species to form Varieties* (A. Wallace)

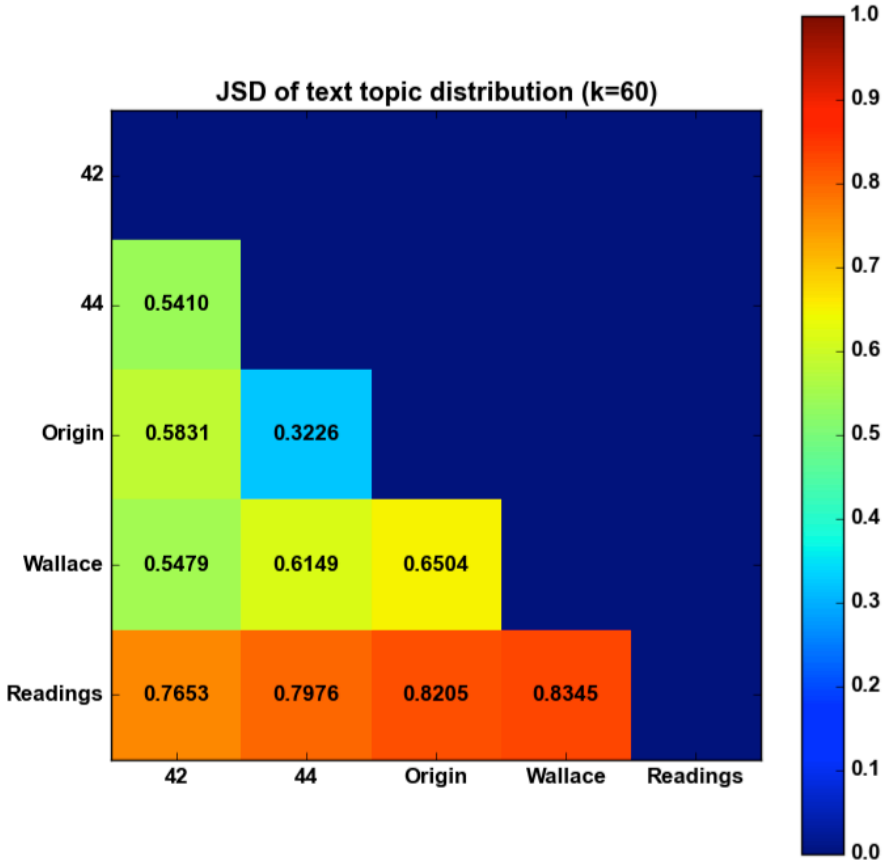
What does the model illuminate?

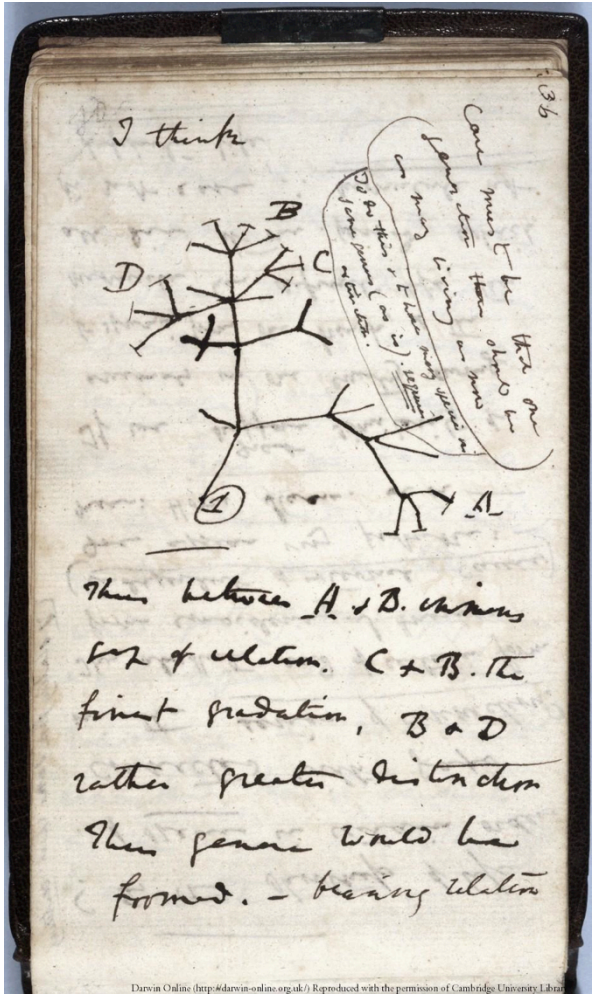
Reading & Writing

Use LDA Query Sampling on:

- Essay of ‘42
- Essay of ‘44
- *On the Origin of Species* (1st ed)
- *On the Tendency of Species to form Varieties* (A. Wallace)

What does the model illuminate?





Findings

"If Wallace had my MS. sketch written out in 1842, he could not have made a better short abstract!" - DCP 2285

Confounds

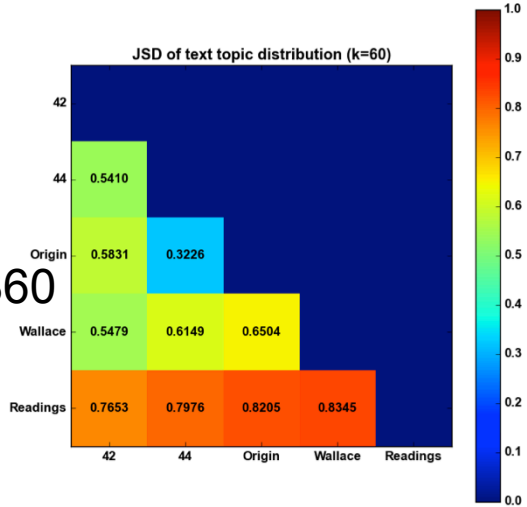
Corpus Cleanup

- Essays include editorial notes by Francis Darwin

Topic model goes through 1860

- readings until 1842
  readings until 1844
  readings until 1859

Bad reference model?

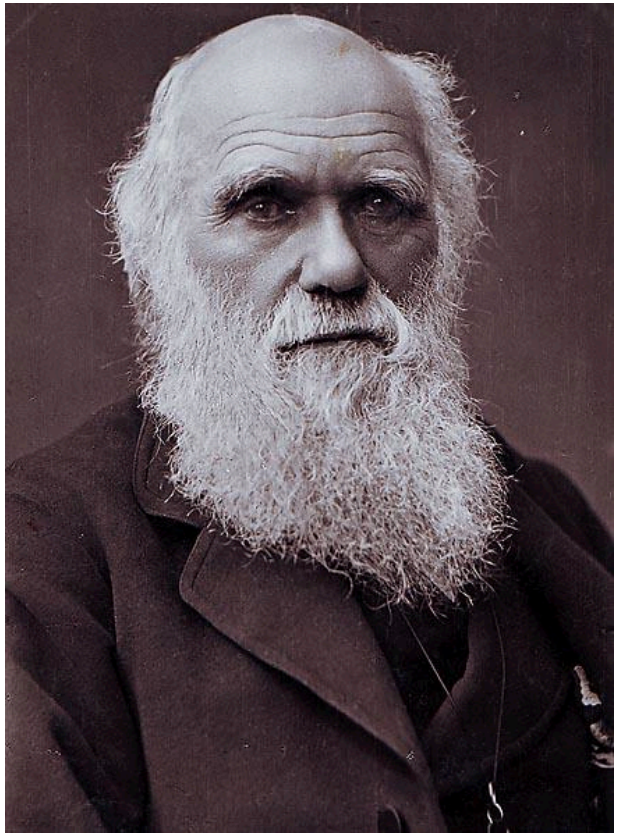


Wallace on Darwin

"This vast, this totally unprecedented change in public opinion has been the result of the work of one man, and was brought about in the short space of twenty years!"

The Guardian: Darwin did not cheat Wallace out of his rightful place in history

Darwin on Darwin



When I see the list of books of all kinds which I read and abstracted, including whole series of Journals and Transactions, I am surprised at my industry.

— *Autobiography*

Questions & Upcoming Events

Local: Topic-driven Foraging
- Goldstone, Todd, Landy Lab
- Friday, April 10 — 9-10a
- MSB II Gill Conference Room

Local: Visualization Techniques for LDA
- Cognitive Science 25th Anniversary Interactive Systems Open House
- Friday, April 17 — 3:30-5:15pm
- Location TBD

Local: Topic Modeling & Network Analysis
- Catapult Center Workshops
- Friday, April 24 — 1-4pm
- Wells Library E159
- Presenter: Colin Allen
- http://www.indiana.edu/~catapult/workshops.shtml

HT Corpus Builder & Topic Explorer
- HathiTrust UnCamp 2015
- Monday, March 30
- Ann Arbor, MI
- Presenters: Jaimie Murdock, Colin Allen
- http://www.hathitrust.org/htrc_uncamp2015

HT Data Capsule & Topic Modeling for Non-consumptive Research
- JCDL 2015 Tutorial
- Sunday, June 21 — 9am-noon
- Knoxville, TN
- Presenters: Jaimie Murdock, Jiaan Zeng, Robert MacDonald
- http://www.jcdl2015.org/

Slides: http://jamr.am/DarwinIUNetSci