# Exploration and exploitation of Victorian science in Darwin's reading notebooks

Authors: Jaimie Murdock, Colin Allen, and Simon DeDeo

Presented by Zachary K. Stine | March 2, 2018

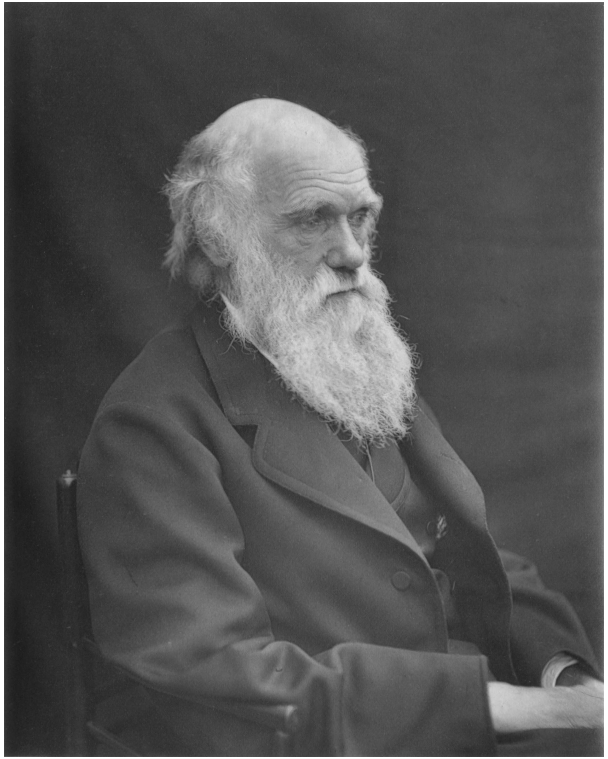
# I. Introduction

# Information-foraging strategies

- Exploitation -- researching in domains in which an individual is an expert.
- Exploration -- researching in domains that are novel to an individual.
- Cognitive searching requires some balance of exploiting existing resources while also exploring new resources.

# Charles Darwin (1809-1882)

- Case study of an individual’s information-foraging strategies
- A qualitatively well-studied individual
- Left reading lists during some of the most productive years of his life



source: Wikipedia

# II. Materials and Methods

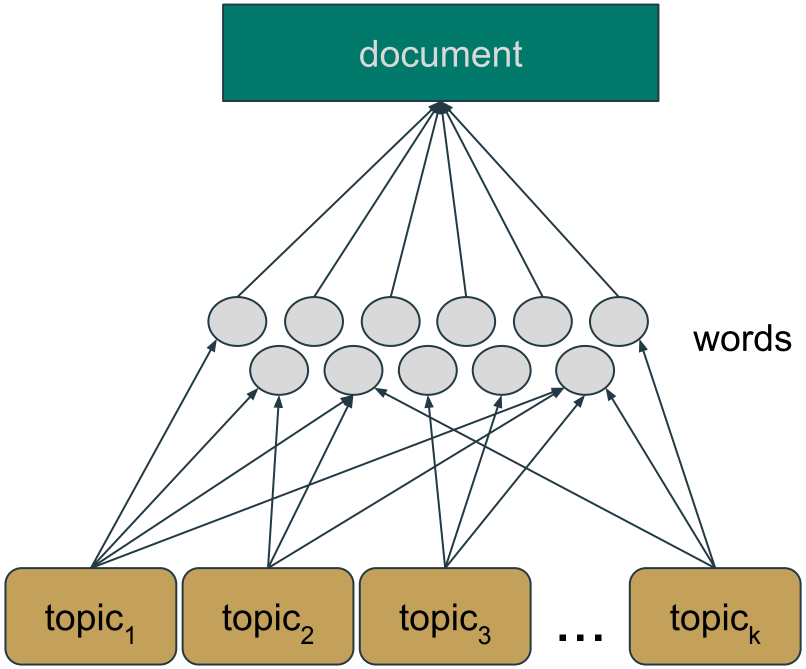
# Data

- Full text of 665 (out of 687) English non-fiction books read by Darwin

- Covers 23 years (1837 - 1860)

# Probabilistic topic models (LDA)

- Each document represented as a bag of words generated by a mixture of “topics”
- Have to choose the number of topics, *k*
- Lets us think of a document as a distribution over the *k* topics



# Cognitive surprise: Kullback-Leibler divergence

- Given what one has already read, how much are we surprised by the next book read?
- KL divergence: given a distribution, *p*, how much surprise is in distribution, *q*?
    - *p* ⟵ distribution of topics already observed
    - *q* ⟵ distribution of topics in the next book
- Two versions
    - Text-to-text surprise (T2T)
        - *p* is the distribution of topics in the last book read only
        - Local surprise
    - Text-to-past surprise (T2P)
        - *p* is the distribution of topics in all books previously read
        - Global surprise

# Cultural production and null reading models

Cultural production

- Uses the same texts
- Ordered by publication date
- I.e., the order that the broader culture produced them
- How does Darwin’s foraging compare with foraging of culture?

Null model

- Permutations of possible reading orders
- Gives us an idea of average, expected surprise
- All results are relative to this null model

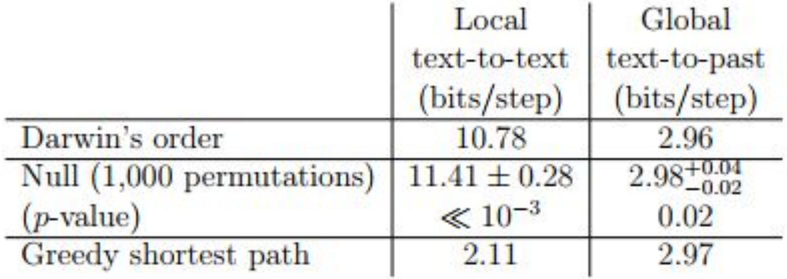
# Bayesian epoch estimation (BEE)

- Unsupervised approach for identifying sustained periods of exploitation or exploration
- Each epoch is defined by
  - A beginning point (either beginning of data or the end of the previous epoch)
  - An average level of surprise
  - The variance around that average
- Requires *n* number of epochs to be chosen
  - For larger *n*, better fit but likely to cause overfitting
  - Akaike Information Criterion (AIC) used to constrain *n*
- Minimum epoch length restricted to 5 years
- Will these information-foraging epochs align with salient events in Darwin’s life?

# III. Results

# Exploration and exploitation

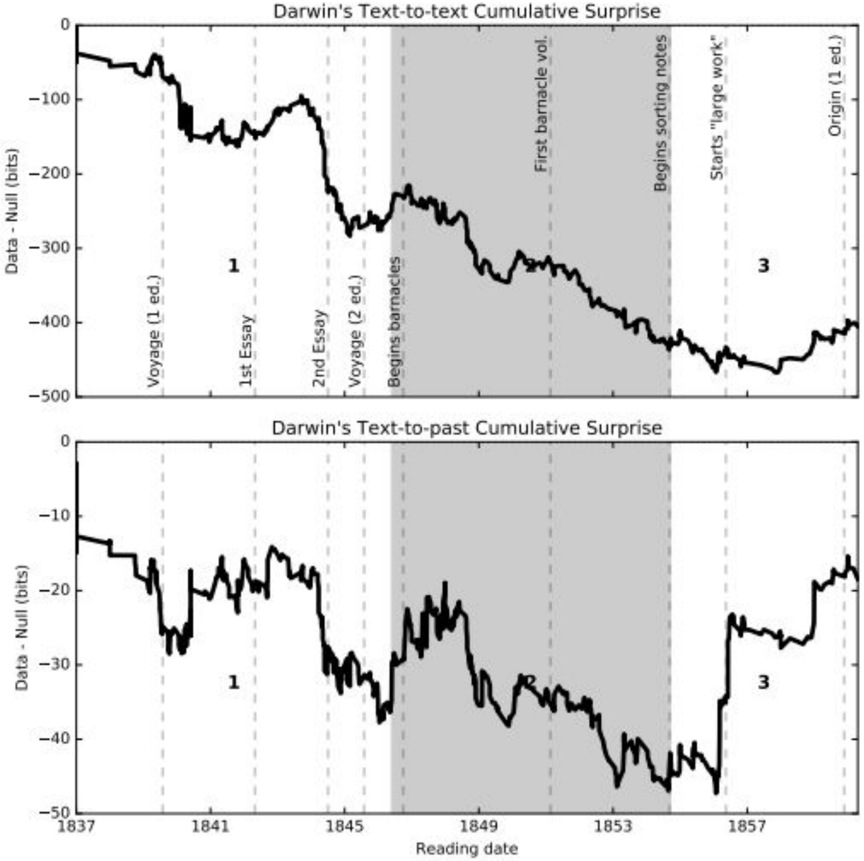
- Darwin is more exploitative than the null model for both T2T and T2P
- Compared with a greedy path that minimizes surprise through documents:
  - Darwin is much more exploratory in T2T
  - But surprisingly less exploratory in T2P



| | Local text-to-text (bits/step) | Global text-to-past (bits/step) |
|---|---|---|
| Darwin's order | 10.78 | 2.96 |
| Null (1,000 permutations) | $11.41 \pm 0.28$ | $2.98^{+0.04}_{-0.02}$ |
| (*p*-value) | $\ll 10^{-3}$ | 0.02 |
| Greedy shortest path | 2.11 | 2.97 |

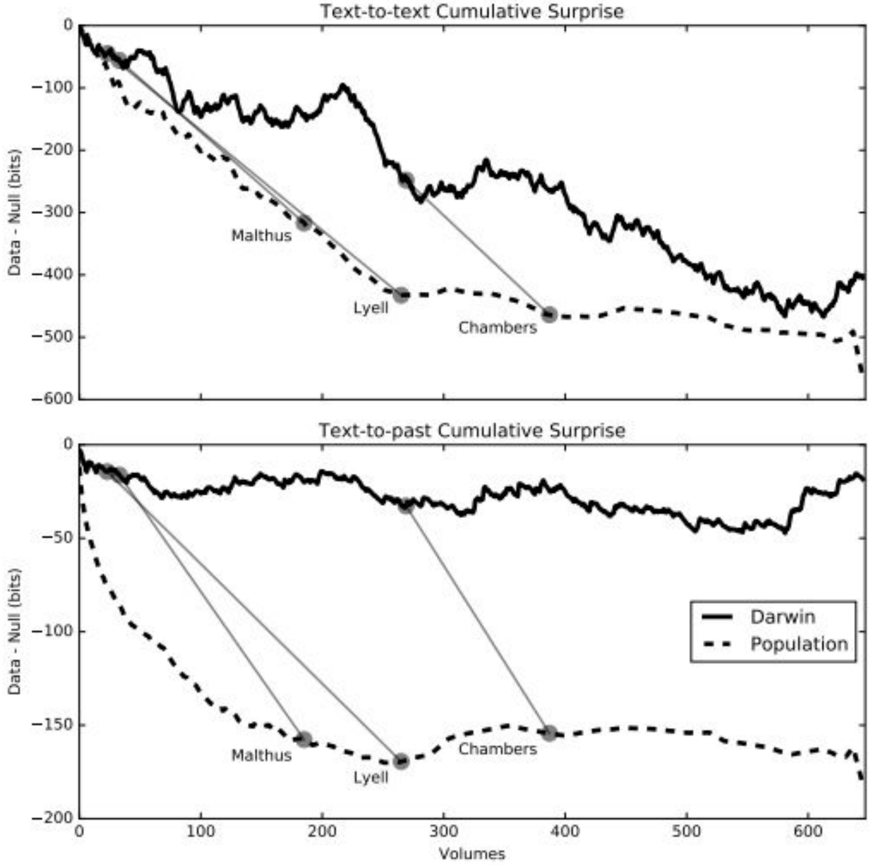
# Readings over time

- Negative slope means less surprise than null model (exploitation)
- Positive slope means more surprise than null model (exploration)
- Three epochs shown by white and grey bands inferred using BEE



# Individual and collective

- Both T2T and T2P surprise is greater in the ordering of texts as Darwin read them compared with the order they were produced
- Darwin is sampling texts in a way that juxtaposes their themes to a greater degree than the population producing them



# Strategy shifts between biographically significant epochs

Major events in Darwin’s life during this time (qualitatively identified):

- **2 October 1836 - 30 September 1846**
  - beginning of data - last volumes from *HMS Beagle* studies published
  - T2T and T2P: exploitation
- **1 October 1846 - 8 September 1854**
  - start of work on barnacles - last volume of barnacles work published
  - T2T: exploitation, T2P: exploration
- **9 September 1854 - 1860**
  - start notes on species - Origin of Species and end of data
  - T2T and T2P: exploration

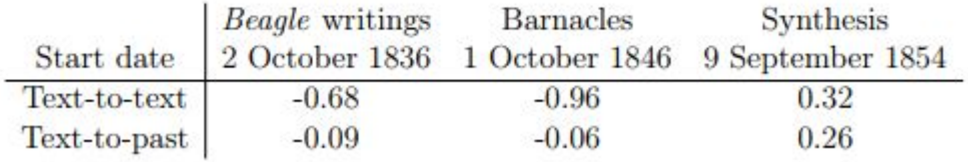
# Unsupervised detection of strategy shifts

Without knowing the dates of the qualitatively important events, inferred epochs that align very closely with those events:

- Start of the first epoch does not need to be inferred (just the beginning of the data): **2 October 1836**
- Start of the second epoch: **27 May 1846**
- Start of the third epoch: **16 September 1854**

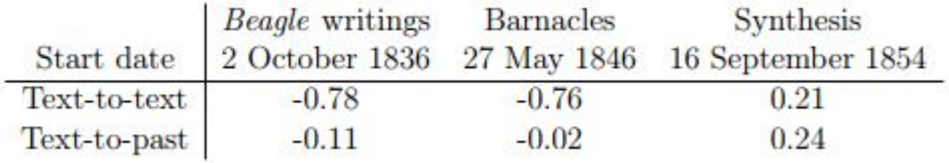
# Qualitative vs. quantitative epoch dates

Qualitative epochs:



| | *Beagle* writings | Barnacles | Synthesis |
|---|---|---|---|
| Start date | 2 October 1836 | 1 October 1846 | 9 September 1854 |
| Text-to-text | -0.68 | -0.96 | 0.32 |
| Text-to-past | -0.09 | -0.06 | 0.26 |

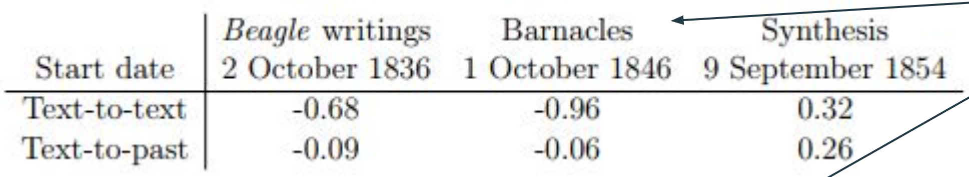
Quantitative epochs:



| | *Beagle* writings | Barnacles | Synthesis |
|---|---|---|---|
| Start date | 2 October 1836 | 27 May 1846 | 16 September 1854 |
| Text-to-text | -0.78 | -0.76 | 0.21 |
| Text-to-past | -0.11 | -0.02 | 0.24 |

# Qualitative vs. quantitative epoch dates

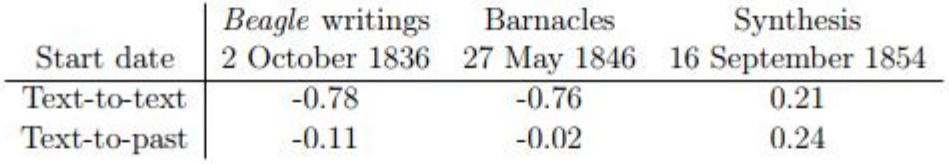
Qualitative epochs:



| | *Beagle* writings | Barnacles | Synthesis |
|---|---|---|---|
| Start date | 2 October 1836 | 1 October 1846 | 9 September 1854 |
| Text-to-text | -0.68 | -0.96 | 0.32 |
| Text-to-past | -0.09 | -0.06 | 0.26 |

~4 months apart

Quantitative epochs:



| | *Beagle* writings | Barnacles | Synthesis |
|---|---|---|---|
| Start date | 2 October 1836 | 27 May 1846 | 16 September 1854 |
| Text-to-text | -0.78 | -0.76 | 0.21 |
| Text-to-past | -0.11 | -0.02 | 0.24 |

1 week apart

# IV. Discussion

# Discussion

- Many studies exist that focus solely on population-level cultural change
- Understanding mechanisms behind population-level changes requires understanding cognitive processes at individual-level as well
- Extending beyond Darwin, can look at reading records of others
  - UK Reading Experience Database (1450-1945)
  - 50 million users on Goodreads

# Questions?

Link to paper on journal’s website: https://www.sciencedirect.com/science/article/pii/S0010027716302840

Link to paper on arXiv: https://arxiv.org/pdf/1509.07175.pdf