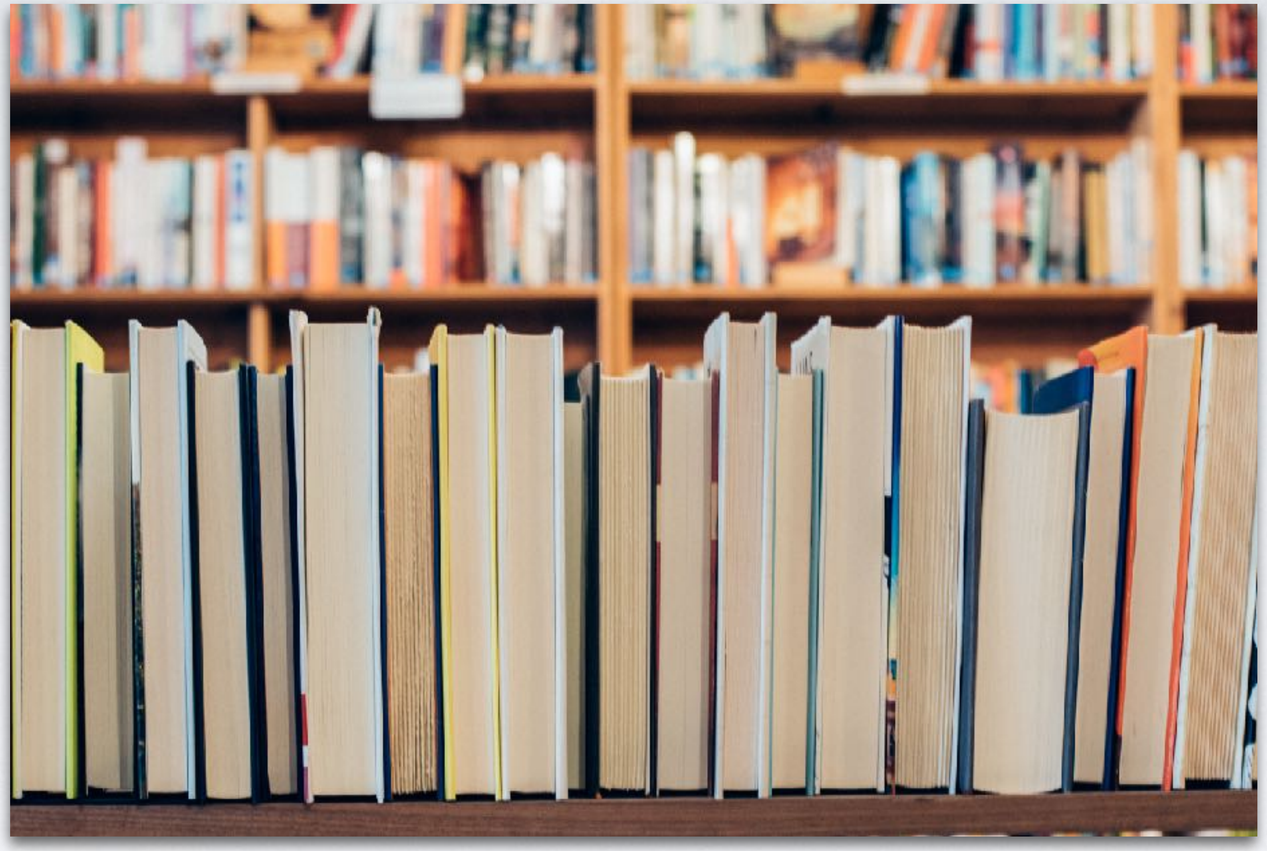
Northeastern University

Northeastern University
Network Science Institute

Barabási Lab



SUCCESS OF BOOKS AND AUTHORS

Xindi Wang
Northeastern University, Boston, USA | Center for Complex Network Research (CCNR)

More than 3 Million books are published every year in the US.

More than 3 Million books are published every year in the US.

Only 500 of those 3 Million become bestsellers!

Why some books become successful while others fail?

DATA SOURCES

ALL THE DATA!!!

DATA SOURCES

DATA SOURCES

- Covers approximately 85% of print trade book sales, over 500,000 ISBNs tracked in a week for 15 years
- Author's career history

npd Bookscan

DATA SOURCES

- Covers approximately 85% of print trade book sales, over 500,000 ISBNs tracked in a week for 15 years
- Author's career history

- Best seller lists with many categories
- Updated weekly from 2008
- Well known measure of success for books

npd Bookscan

The New York Times

DATA SOURCES

- Covers approximately 85% of print trade book sales, over 500,000 ISBNs tracked in a week for 15 years
- Author's career history

- Best seller lists with many categories
- Updated weekly from 2008
- Well known measure of success for books

npd Bookscan

The New York Times

goodreads
Meet your next favorite book.

- Detailed info of books and authors
- Ratings and reviews of books and authors
- Readers data: books read, friend, etc.

DATA SOURCES

- Covers approximately 85% of print trade book sales, over 500,000 ISBNs tracked in a week for 15 years
- Author's career history

- Best seller lists with many categories
- Updated weekly from 2008
- Well known measure of success for books

npd Bookscan

The New York Times

goodreads
Meet your next favorite book.

WIKIPEDIA
The Free Encyclopedia

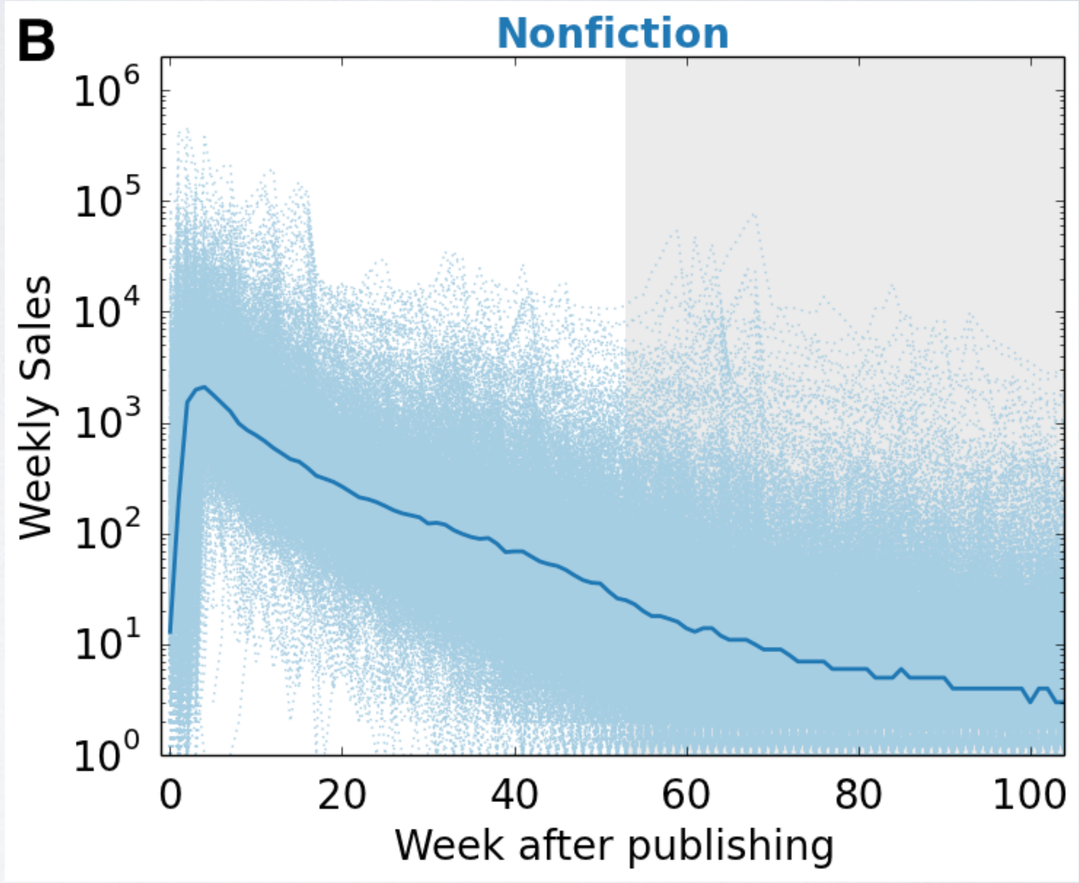
- Detailed info of books and authors
- Ratings and reviews of books and authors
- Readers data: books read, friend, etc.

- Page-views: how many clicks for a page
- Indicates global interest in an author/book
- Info about famous books & authors

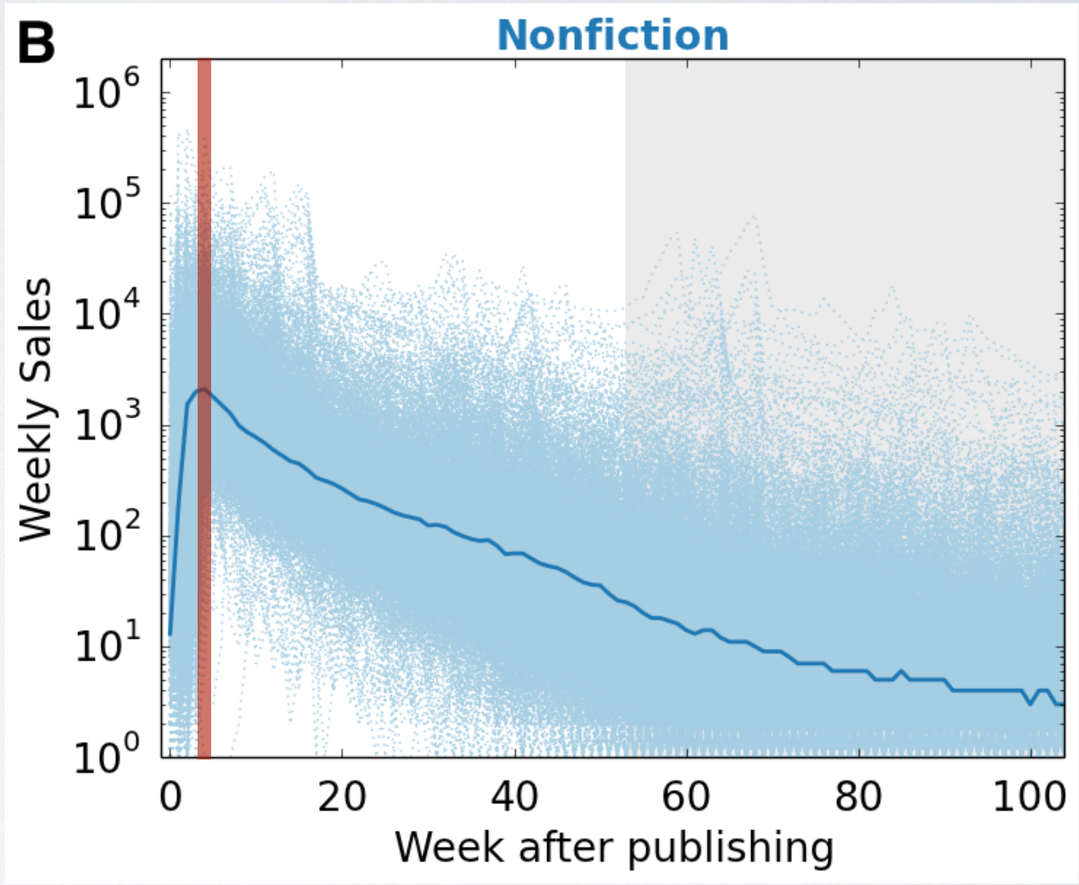
OUTLINE

- Bestsellers study, specifically sales pattern
- Predict success before the book published

SALES PATTERN



SALES PATTERN



Most books follow a universal "quick rise, slower decay" pattern
Peak arrives within 10 weeks after publishing

SALES PATTERN MODELING

- **Fitness** η_i : the book's ability to respond to the taste of a wide readership.
- **Preferential attachment**: the higher the up-to-date sales of a book S_i^t , the more likely it is to be purchased again.
- **Aging**: $A_i(t) = \frac{1}{\sqrt{2\pi}\sigma_i t} \exp[-\frac{(\ln t - \mu_i)^2}{2\sigma_i^2}]$

  μ_i : the book's immediacy determined by the time the sales reach its peak

  σ_i : the decay rate capturing the longevity.

Dashun Wang, Chaoming Song, and Albert-László Barabási, "Quantifying long-term scientific impact," Science 342, 127–132 (2013).

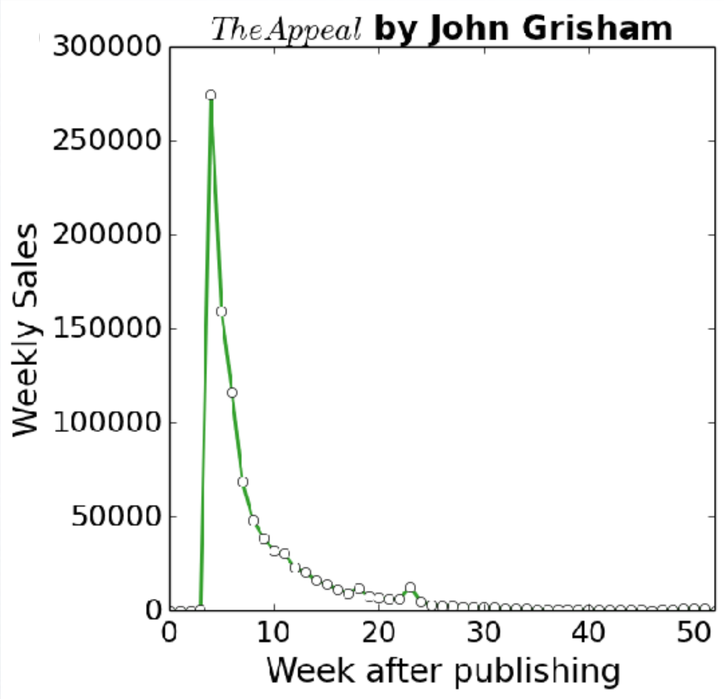
SALES PATTERN MODELING

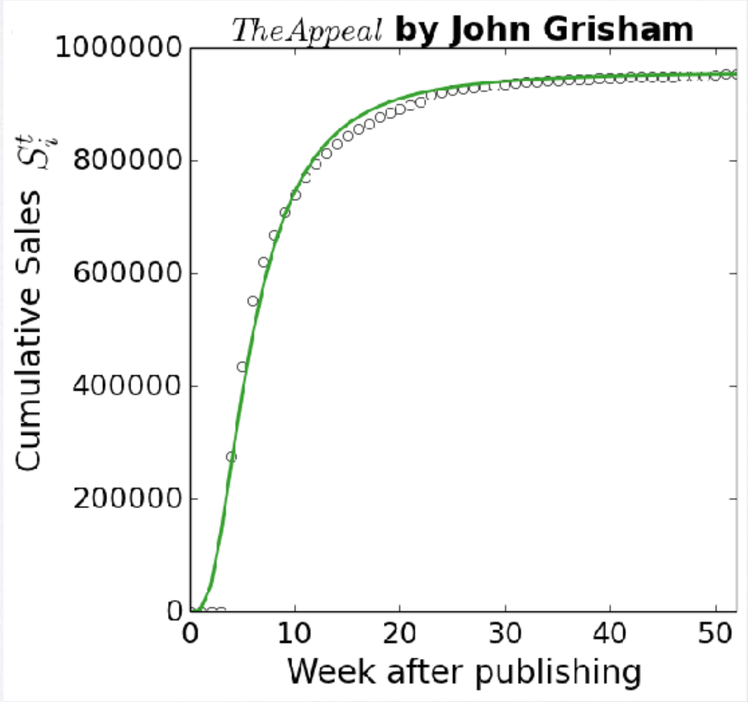
The cumulative sales of a book i at a time t after publication:

$$S_i^t = m[e^{\lambda_i \Phi(\frac{\ln t - \mu_i}{\sigma_i})} - 1]$$

λ_i is the relative fitness proportional to η_i

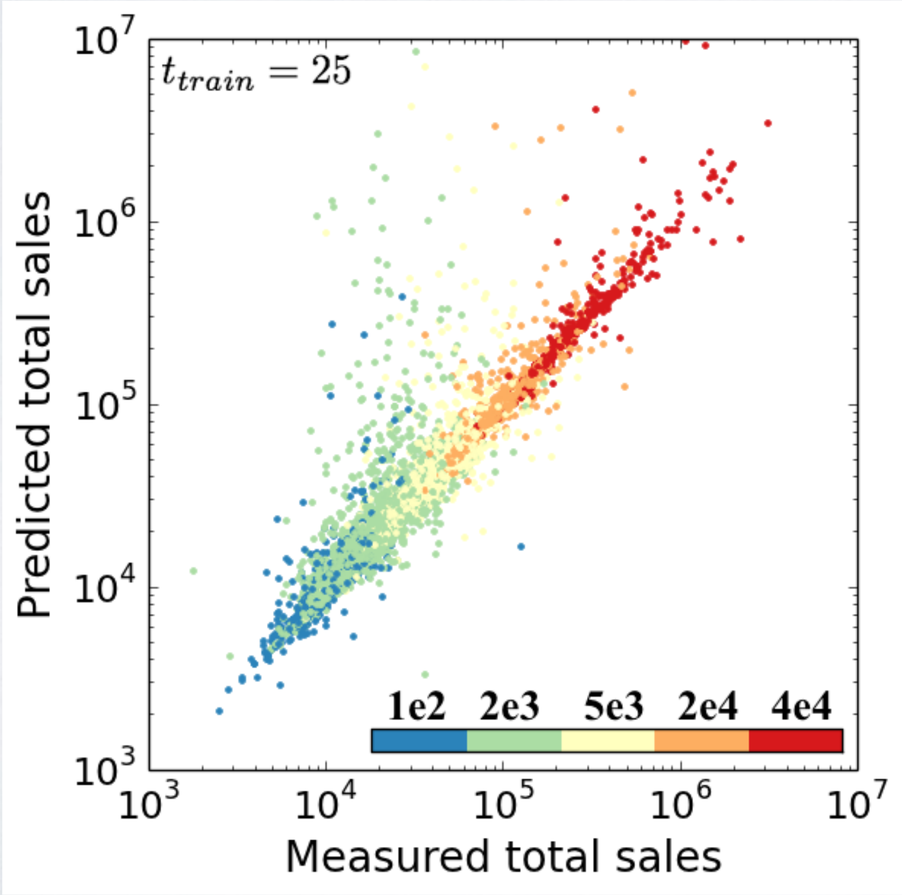
$$\Phi(x) = (2\pi)^{-1/2} \int_{-\infty}^{x} e^{-y^2/2} dy$$

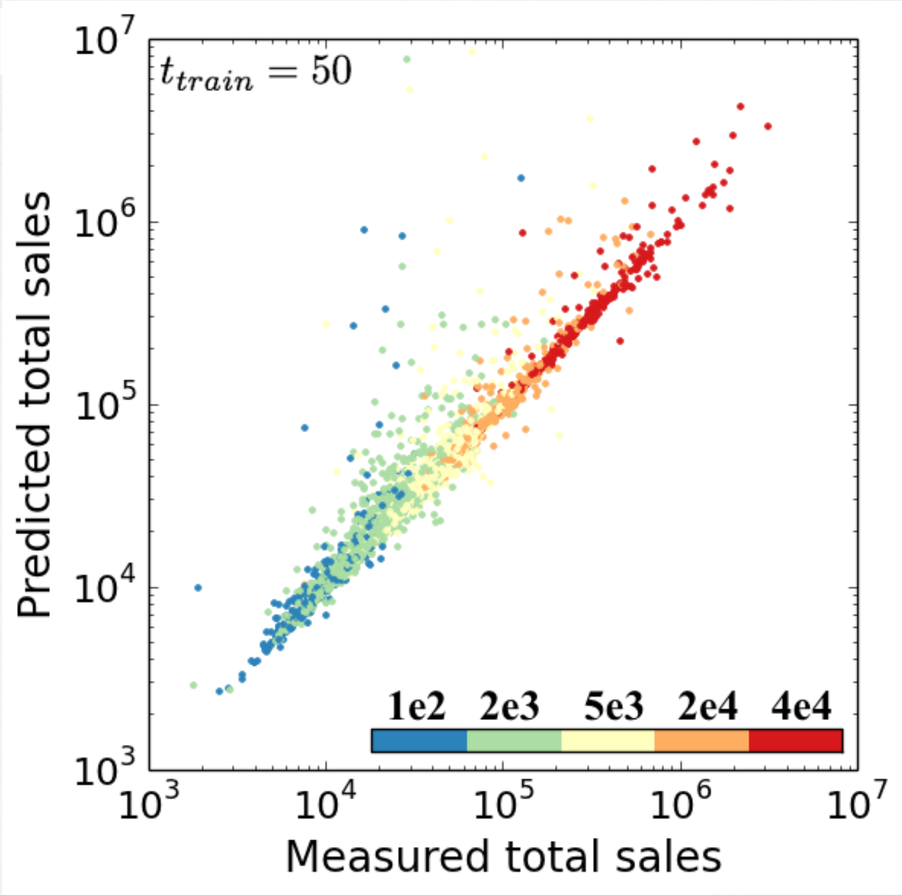




PREDICT THE FUTURE

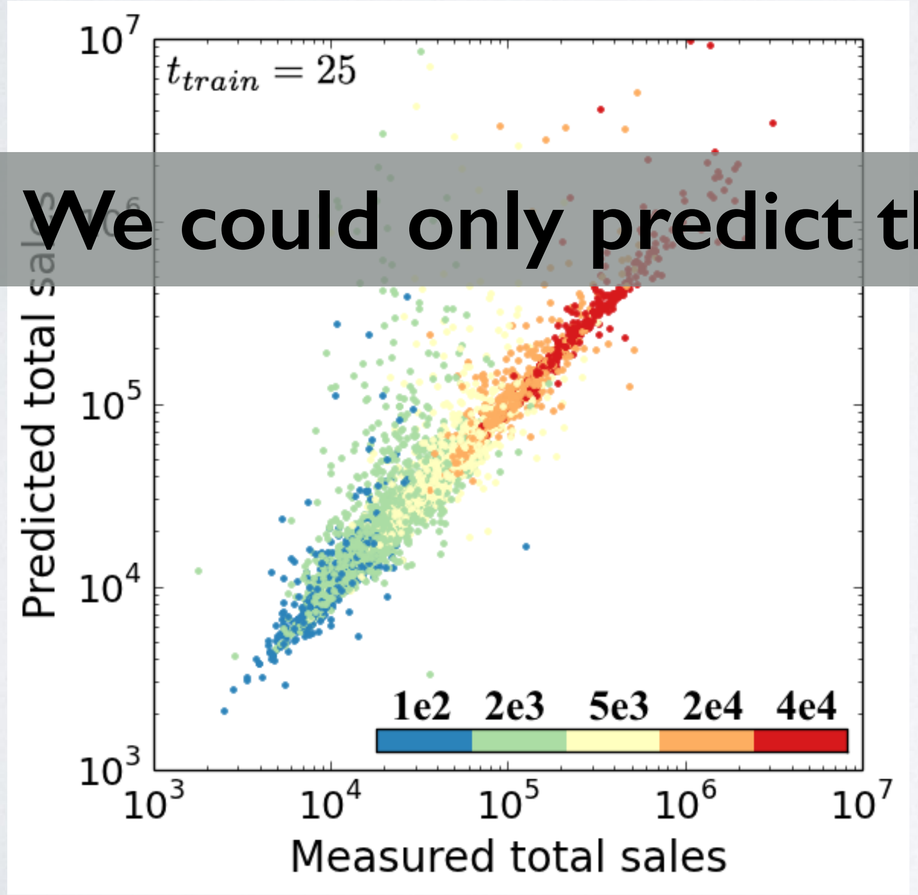
$$S_i^{\infty} = m(e^{\lambda_i} - 1)$$

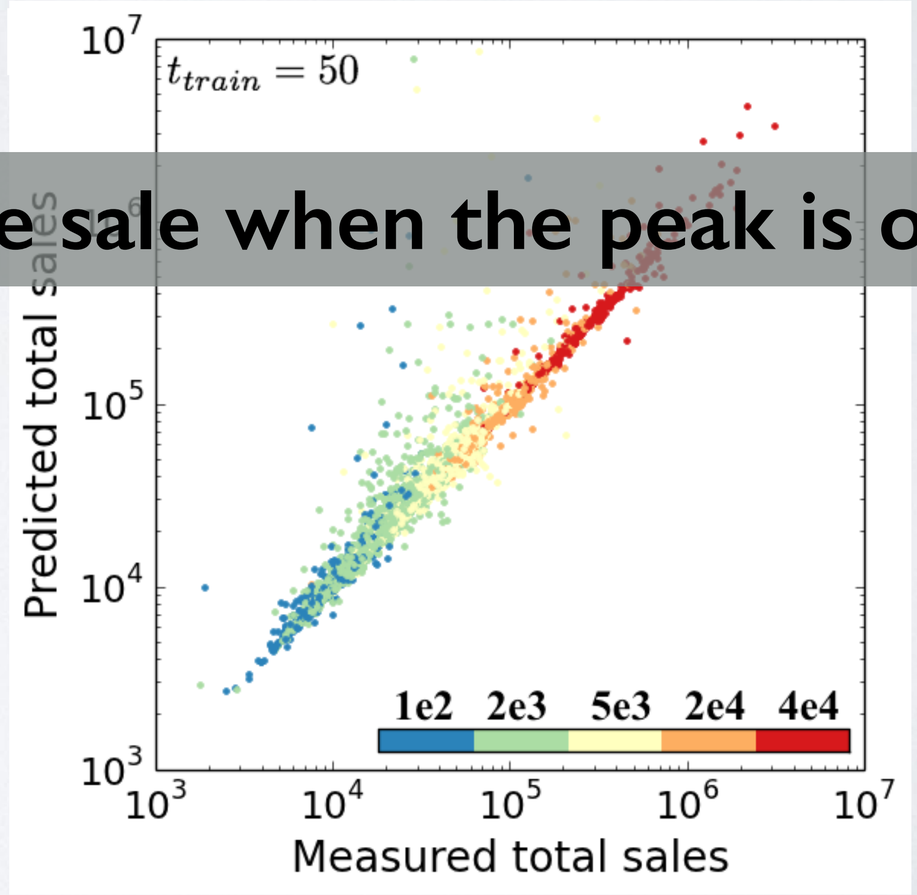




PREDICT THE FUTURE

$$S_i^{\infty} = m(e^{\lambda_i} - 1)$$





We could only predict the sale when the peak is over

CONCLUSION TILL NOW

CONCLUSION TILL NOW

- Sales pattern follows a “quick rise, slow decay” pattern, the peak usually arrives within 10 weeks.

CONCLUSION TILL NOW

- Sales pattern follows a “quick rise, slow decay” pattern, the peak usually arrives within 10 weeks.
- We can infer future sales from past sales, but this inference needs a lot of sales data

CONCLUSION TILL NOW

- Sales pattern follows a "quick rise, slow decay" pattern, the peak usually arrives within 10 weeks.
- We can infer future sales from past sales, but this inference needs a lot of sales data
- More investigations about bestseller books and authors[1]

Success in Books: A Big Data Approach to Bestsellers, Burcu Yucesoy, Xindi Wang, Junming Huang, Albert-László Barabási, EPJ Data Science (Submitted)

Can we predict sales before the book is published?

FEATURES

FEATURES

Author

- Fame
- Publishing History

FEATURES

Author

- **Fame**
- **Publishing History**

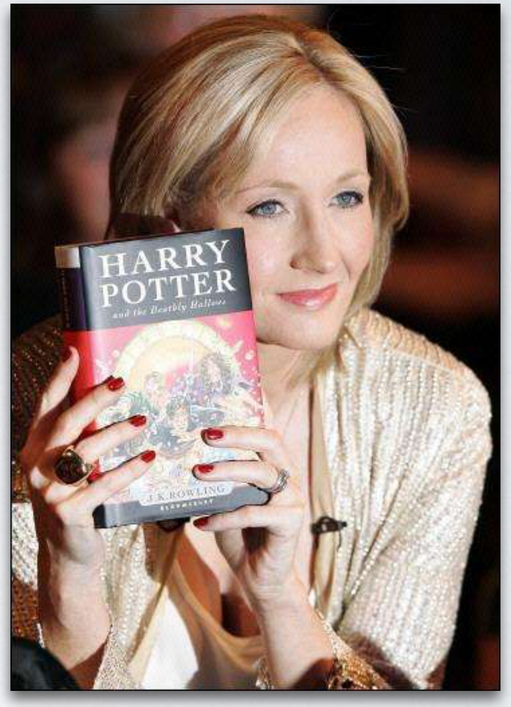
Famous because he was president

FEATURES

Author

- Fame
- Publishing History

Famous because he was president



Famous because of her books

FEATURES

Author

- Fame
- Publishing History

FEATURES

Author

- Fame
- Publishing History

Book

- Genre
- Topic
- Publishing Month

FEATURES

Author

- Fame
- Publishing History

Book

- Genre
- Topic
- Publishing Month

Publisher

- Publisher (Imprint) Prominence

FEATURES

Author

- **Fame**
- **Publishing History**

WIKIPEDIA
The Free Encyclopedia

- Fame
  - Cumulative Fame
  - Fame Establishing Days
  - Normalized Cumulative Fame
  - Recent Fame

npd Bookscan

- Publishing History
  - Total previous sales
  - Previous sales in this genre
  - Previous sales in other genres
  - Normalized previous sales

FEATURES

FEATURES

Book

- **Genre**
- **Topic**
- **Publishing Month**

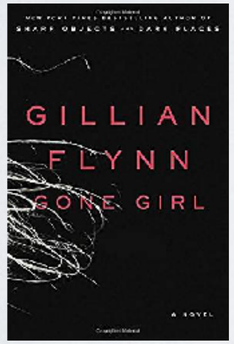
FEATURES

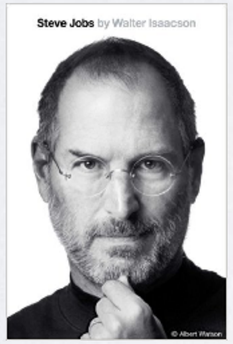
Book

- **Genre**
- **Topic**
- **Publishing Month**

npd Bookscan

- Genre
  - Bisac Code





FIC031000 BIO003000

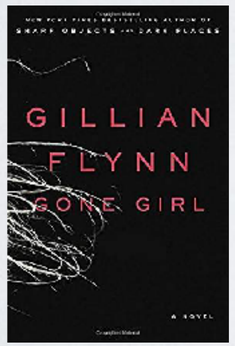
Book

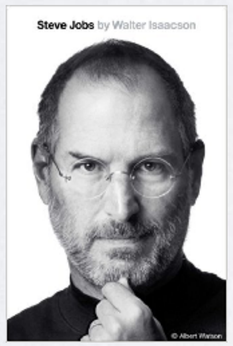
- Genre
- Topic
- Publishing Month

FEATURES

npd Bookscan

- Genre
  - Bisac Code

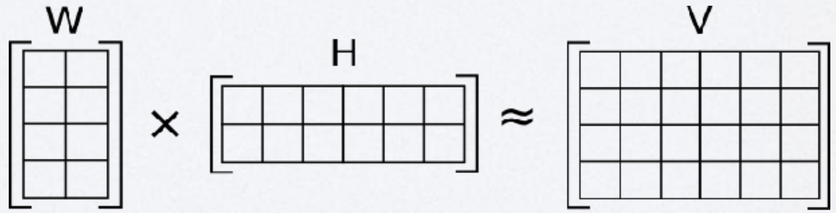




FIC031000 BIO003000

goodreads
Meet your next favorite book.

- Topic
  - Nonnegative matrix factorization (NMF) on book summary



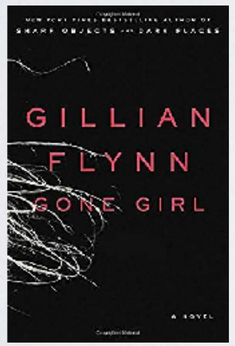
FEATURES

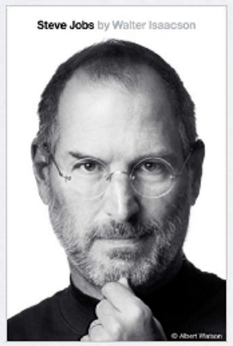
Book

- **Genre**
- **Topic**
- **Publishing Month**

npd Bookscan

- Genre
  - Bisac Code

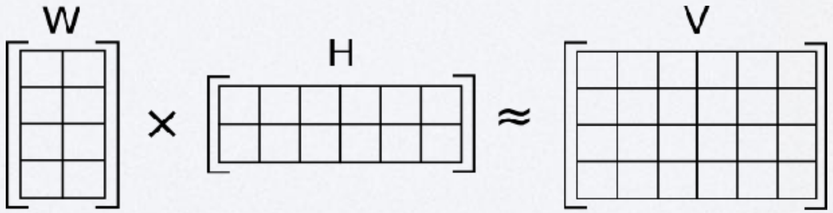




FIC031000 BIO003000

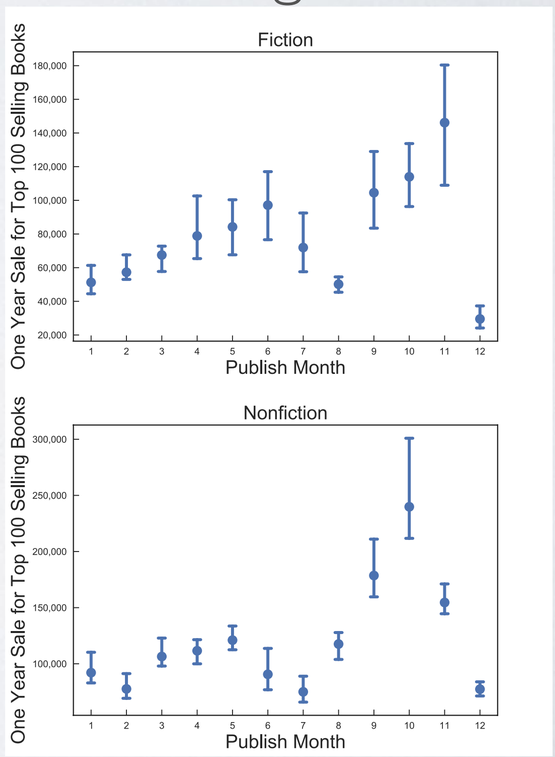
goodreads
Meet your next favorite book.

- Topic
  - Nonnegative matrix factorization (NMF) on book summary



npd Bookscan

- Publishing Month

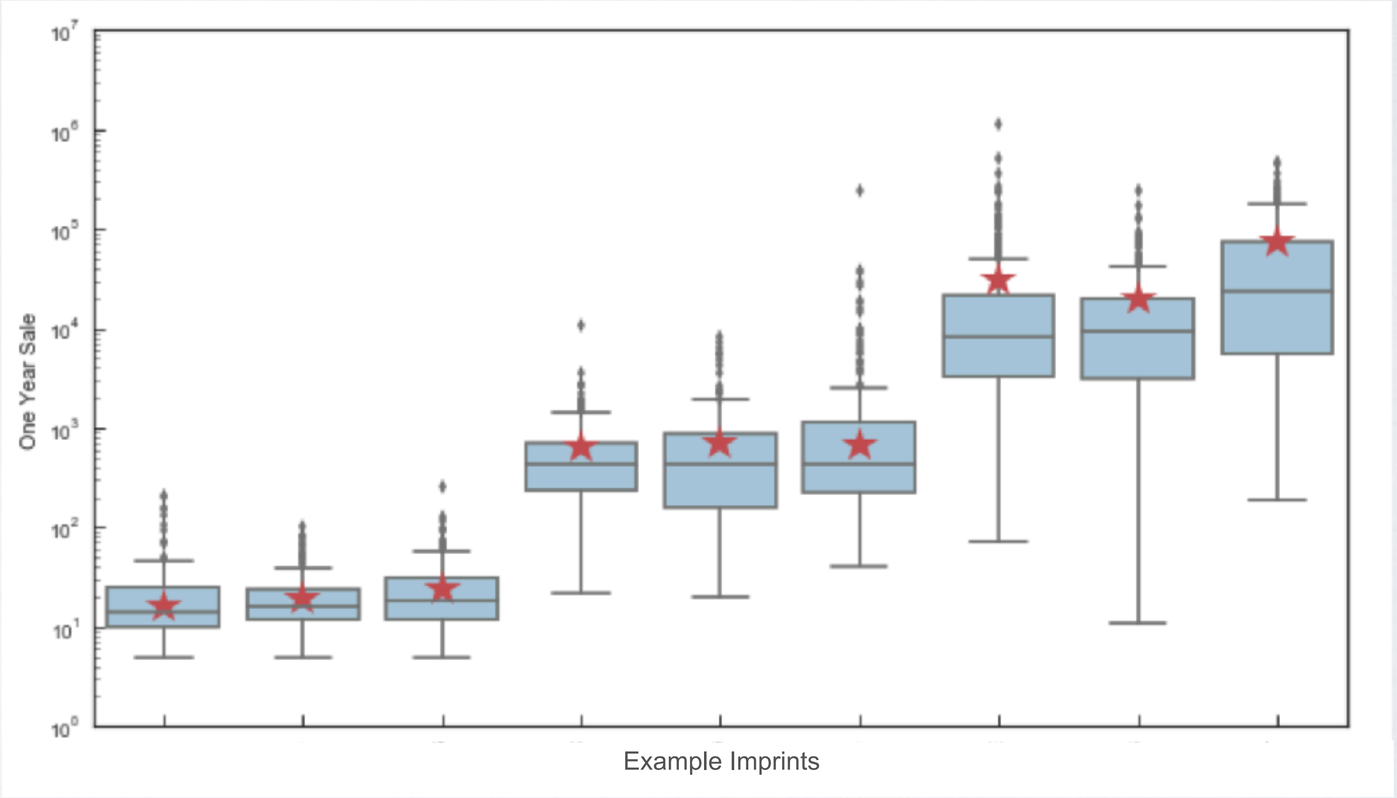


FEATURES

npd Bookscan

Publisher

- Publisher (Imprint) Prominence



HOW TO BUILD THE MODEL?

- Linear regression?
  - Problem of linear regression: unbalanced data

HOW TO BUILD THE MODEL?

- Linear regression?
  - Problem of linear regression: unbalanced data

Systematic Underprediction

Given a sequence of previously published books ranked by their sales, where would we place a new book in this sequence?

LEARNING TO PLACE

- Pairwise Comparison
- Placing objects

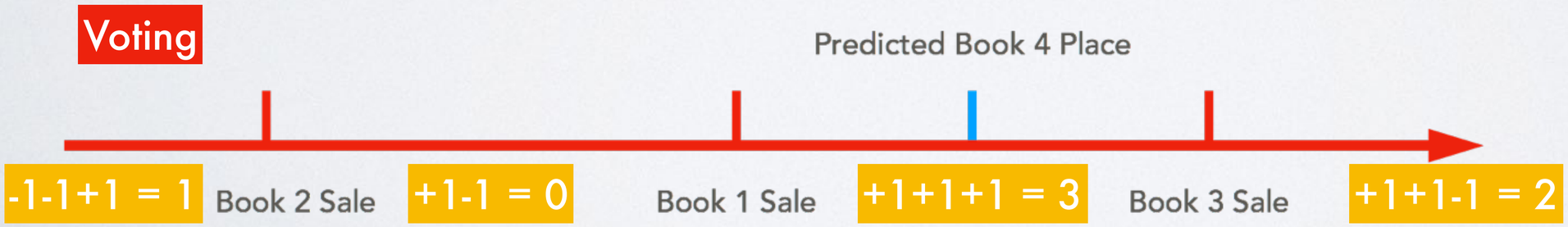
TRAINING PHASE

Book 1 Features — Book 2 Features = Feature Difference(Book 1 and 2)

Features	Label (based on peak sale)
Feature Difference(Book 1 and 2)	greater
Feature Difference(Book 1 and 3)	Smaller
Feature Difference(Book 2 and 3)	smaller

→ Random Forest Classifier

TESTING PHASE



TESTING PHASE

Book 4 Features

Features

Feature Difference(Book 4 and 1)

Feature Difference(Book 4 and 2)

Feature Difference(Book 4 and 3)

Random Forest Classifier

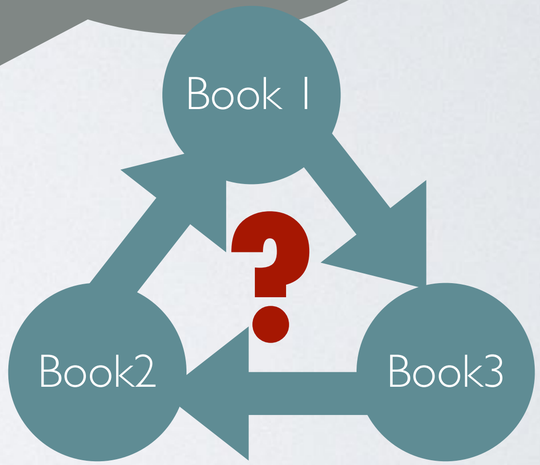
Predicted Label (based on peak sale)

greater

greater

Smaller

There maybe conflicts among predicted labels



Voting

Predicted Book 4 Place

-1-1+1 = 1

Book 2 Sale

+1-1 = 0

Book 1 Sale

+1+1+1 = 3

Book 3 Sale

+1+1-1 = 2

TESTING PHASE

Book 4 Features

Features

Feature Difference(Book 4 and 1)

Feature Difference(Book 4 and 2)

Feature Difference(Book 4 and 3)

Random Forest Classifier

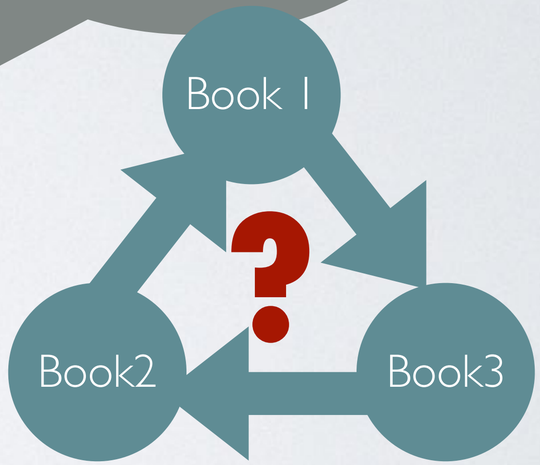
Predicted Label (based on peak sale)

greater

greater

Smaller

There maybe conflicts among predicted labels



Voting

Predicted Book 4 Place

-1-1+1 = 1 Book 2 Sale

+1-1 = 0

Book 1 Sale

+1+1+1 = 3

Book 3 Sale

+1+1-1 = 2

TESTING PHASE

Book 4 Features

Features

Feature Difference(Book 4 and 1)

Feature Difference(Book 4 and 2)

Feature Difference(Book 4 and 3)

Random Forest Classifier

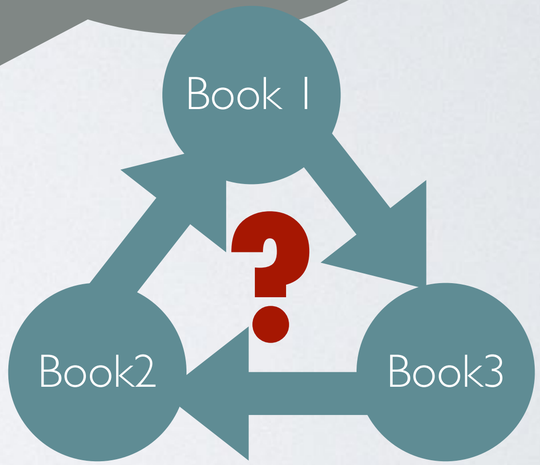
Predicted Label (based on peak sale)

greater

greater

Smaller

There maybe conflicts among predicted labels



Voting

Predicted Book 4 Place

-1-1+1 = 1 Book 2 Sale +1-1 = 0 Book 1 Sale +1+1+1 = 3 Book 3 Sale +1+1-1 = 2

TESTING PHASE

Book 4 Features

Features

Feature Difference(Book 4 and 1)

Feature Difference(Book 4 and 2)

Feature Difference(Book 4 and 3)

Random Forest Classifier

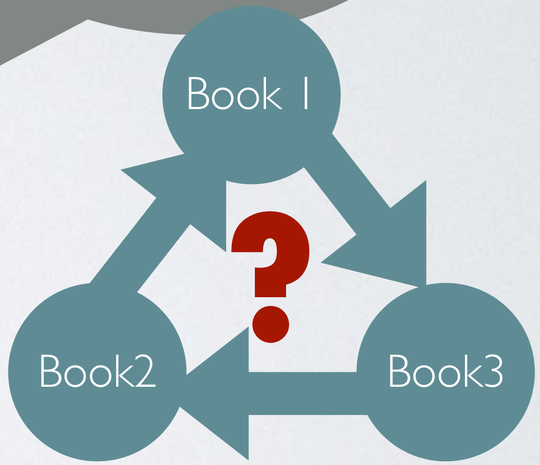
Predicted Label (based on peak sale)

greater

greater

Smaller

There maybe conflicts among predicted labels



Voting

Predicted Book 4 Place

-1-1+1 = 1 Book 2 Sale

+1-1 = 0

Book 1 Sale

+1+1+1 = 3

Book 3 Sale

+1+1-1 = 2

TESTING PHASE

Book 4 Features

Features

Feature Difference(Book 4 and 1)

Feature Difference(Book 4 and 2)

Feature Difference(Book 4 and 3)

Random Forest Classifier

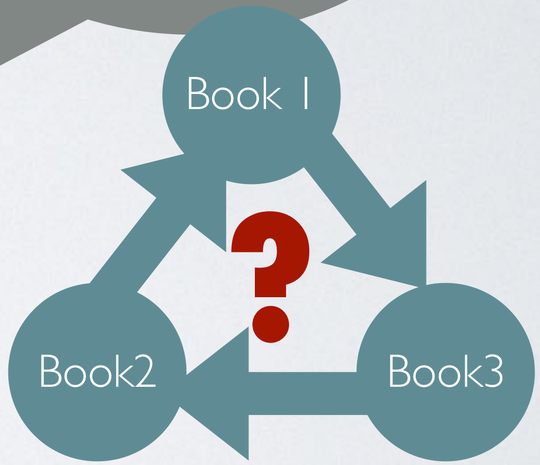
Predicted Label (based on peak sale)

greater

greater

Smaller

There maybe conflicts among predicted labels



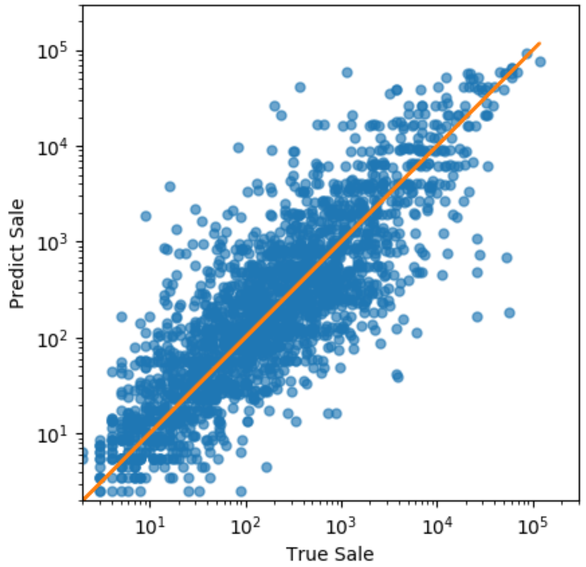
Voting

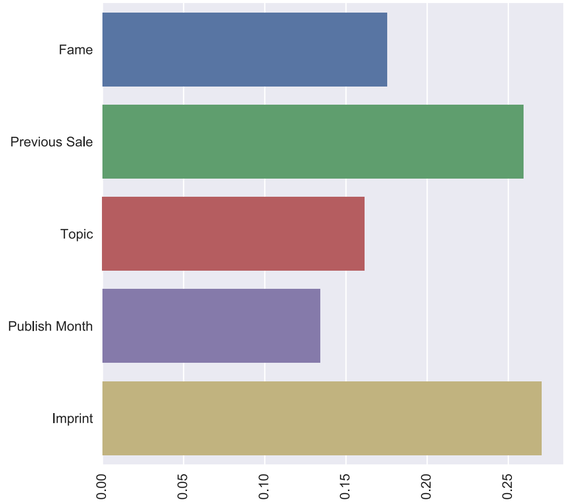
Predicted Book 4 Place

-1+1+1 = 1 Book 2 Sale +1-1 = 0 Book 1 Sale +1+1+1 = 3 Book 3 Sale +1+1-1 = 2

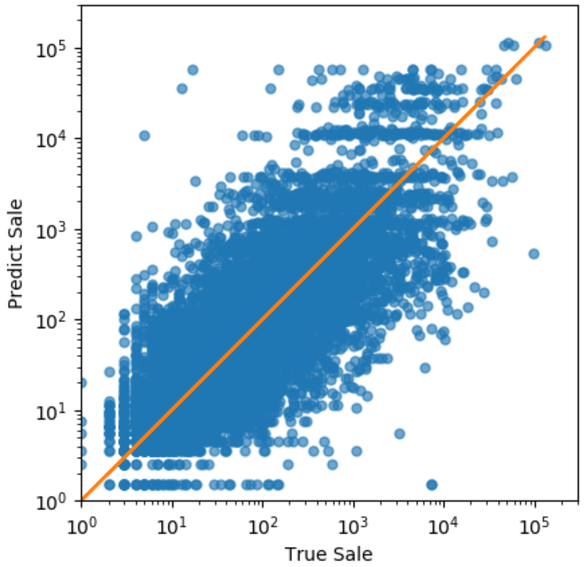
RESULTS

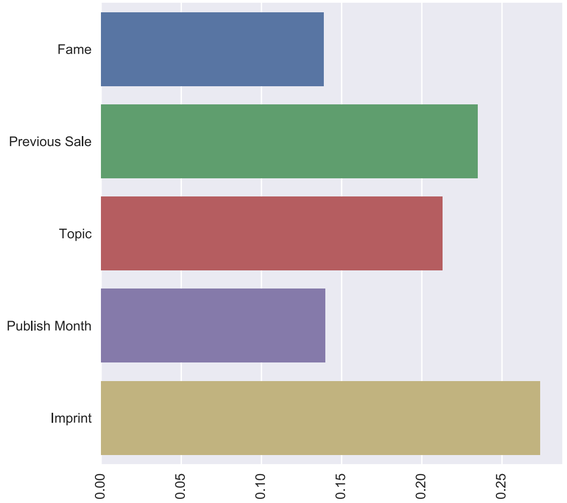
Fiction





Nonfiction

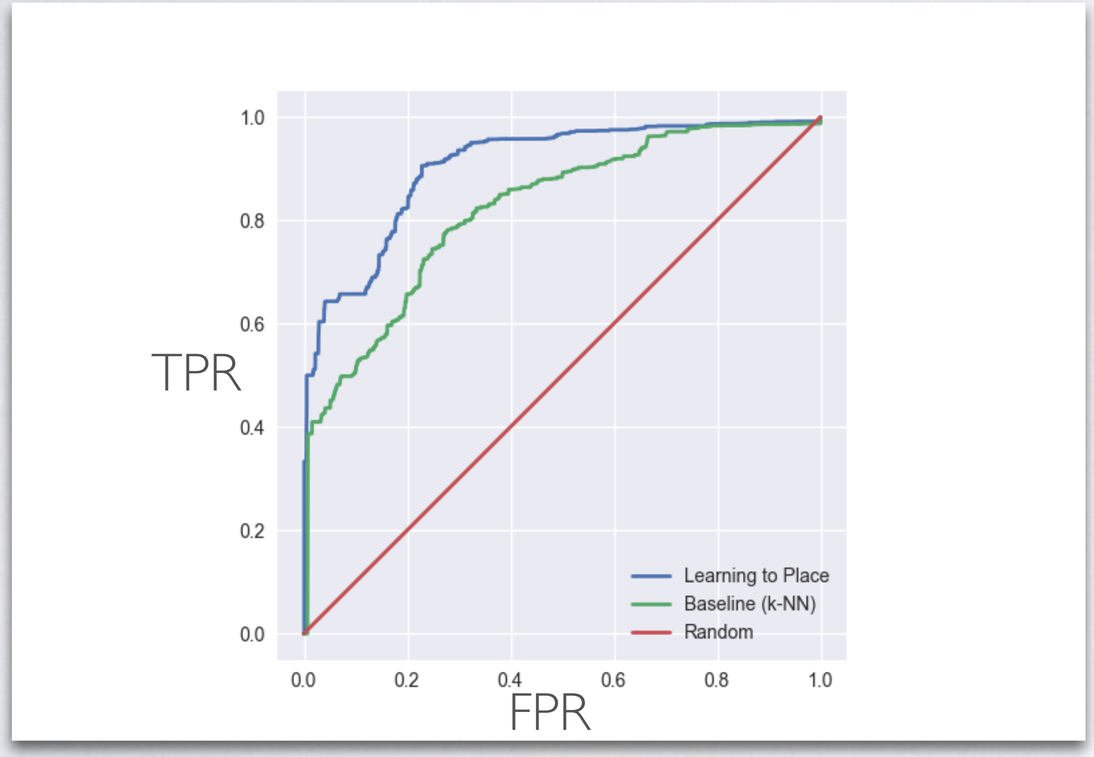




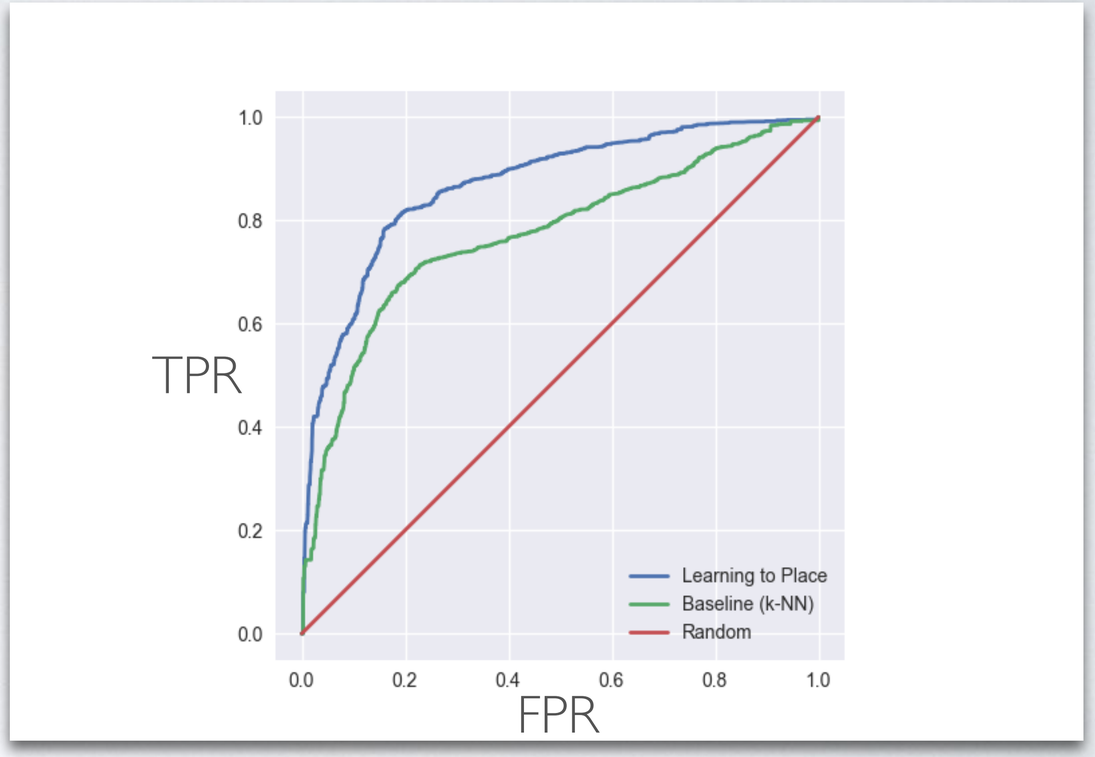
Predicted Sale is the middle point of most voted interval

RESULTS

Fiction



Nonfiction



CONCLUSION

CONCLUSION

- We build a model *Learning to Place* that could predict the success of books pretty accurately

CONCLUSION

- We build a model *Learning to Place* that could predict the success of books pretty accurately
- We know the feature importance for different categories (Fiction and Nonfiction)

WHAT'S NEXT

WHAT'S NEXT

- Improvement on Nonfiction

WHAT'S NEXT

- Improvement on Nonfiction
- Does social media play a role in publishing: Information cascading

WHAT'S NEXT

- Improvement on Nonfiction
- Does social media play a role in publishing: Information cascading
- Author career

Northeastern University
Network Science Institute

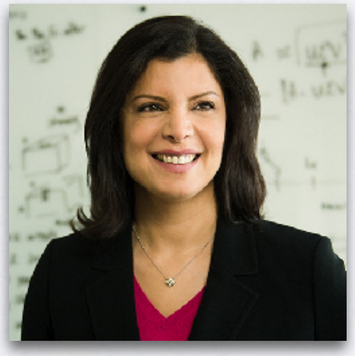
Barabási Lab

Northeastern University

Burcu Yucesoy

Onur Varol

Junming Huang



Tina Eliassi-Rad

Albert-László Barabási

Email: xindi.w1993@gmail.com

SALES PATTERN MODELING

- **Fitness** η_i : the book's ability to respond to the taste of a wide readership.
- **Preferential attachment**: the higher the up-to-date sales of a book S_i^t , the more likely it is to be purchased again.
- **Aging**: $A_i(t) = \frac{1}{\sqrt{2\pi}\sigma_i t} \exp[-\frac{(\ln t - \mu_i)^2}{2\sigma_i^2}]$

  μ_i : the book's immediacy determined by the time the sales reach its peak

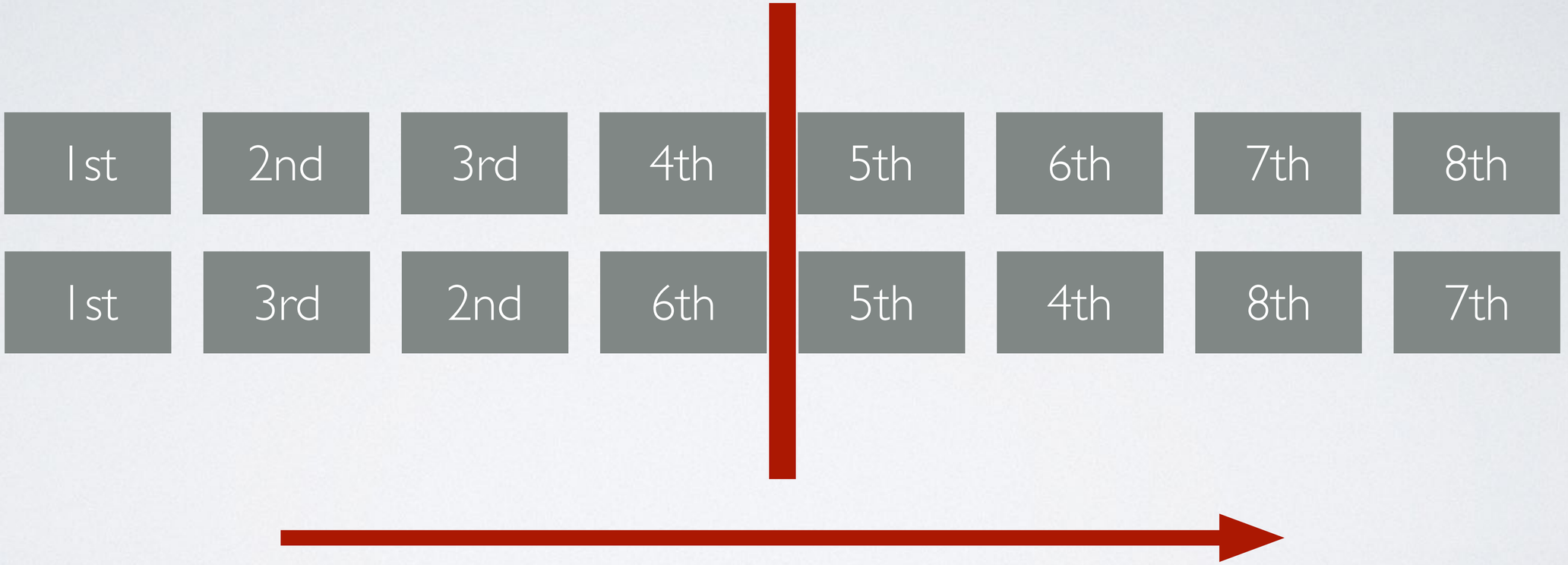
  σ_i : the decay rate capturing the longevity.

The probability of a book i to be purchased at a time t after publication

$$\Pi_i(t) \sim \eta_i S_i^t A_i$$

Dashun Wang, Chaoming Song, and Albert-László Barabási, "Quantifying long-term scientific impact," Science 342, 127–132 (2013).

HOW TO CALCULATE ROC



At each threshold, how many false positive and true positive