Digital humanities: modeling semi-structured data from traditional scholarship

Tom Lippincott
IntroHLT Fall 2019
Human Language Technology Center of Excellence
Center for Language and Speech Processing

Outline

Intro: A few thoughts on “Digital humanities”

Motivating study: Post-Atlantic Slave Trade

Model: Graph-Entity Autoencoders

Bonus study: Authorship attribution of ancient documents

Ongoing work

Intro: A few thoughts on “Digital humanities”

What is “digital humanities”?

Some responses:

- “an idea that will increasingly become invisible” -Stanford
- “a term of tactical convenience” -UMD
- “I don’t: I’m sick of trying to define it” -GMU
- “a convenient label, but fundamentally I dont believe in it” -NYU
- “an unfortunate neologism” -Library of Congress

What is “digital humanities”?

Themes at DH2019

- Visualization
- Geographic information systems
- Social and ethical issues
- Education
- VR, maker spaces
- OCR
- Machine learning

Working definitions

Digital humanities

Traditional researcher

(Traditional) scholarly dataset

Computational researcher

Working definitions

Digital humanities

Traditional inquiries enabled by computational intelligence

Traditional researcher

(Traditional) scholarly dataset

Computational researcher

Working definitions

Digital humanities
Traditional inquiries enabled by computational intelligence

Traditional researcher
Academic from field that doesn’t typically employ quantitative methods (History, Literary Criticism, ...

(Traditional) scholarly dataset

Computational researcher

Working definitions

Digital humanities

Traditional inquiries enabled by computational intelligence

Traditional researcher

Academic from field that doesn't typically employ quantitative methods (History, Literary Criticism, . . .

(Traditional) scholarly dataset

Data assembled by a traditional researcher in the field

Computational researcher

Working definitions

Digital humanities

Traditional inquiries enabled by computational intelligence

Traditional researcher

Academic from field that doesn't typically employ quantitative methods (History, Literary Criticism, . . .

(Traditional) scholarly dataset

Data assembled by a traditional researcher in the field

Computational researcher

Design and bring machine learning models to bear on datasets

Why is collaboration rare?

Traditional researchers have insight into the *data*

Machine learning researchers can pair *data* with appropriate models

Why is collaboration rare?

Traditional researchers have insight into the *data*

- Data is painstakingly gathered and coveted
- Hypotheses are *subtle* but not numerically evaluated
- May publish one or two papers during PhD, but *dissertation* is primary focus

Machine learning researchers can pair *data* with appropriate models

Why is collaboration rare?

Traditional researchers have insight into the *data*

- Data is painstakingly gathered and coveted
- Hypotheses are *subtle* but not numerically evaluated
- May publish one or two papers during PhD, but *dissertation* is primary focus

Machine learning researchers can pair *data* with appropriate models

- Data is aggressively shared to encourage rigorous evaluation
- Tasks are often *shallow* and *prespecified*
- Publish multiple papers per year

Topic models: the rare success story

Topic models: the rare success story

Widely used

- Low barrier to entry: everyone has "documents"
- Little expertise required
- Output easy to visualize and interpret

Topic models: the rare success story

Widely used

- Low barrier to entry: everyone has “documents”
- Little expertise required
- Output easy to visualize and interpret

Widely abused

- Deceptively easy to use: it will give you *something*
- You can always find “patterns”: confirmation bias abounds
- Older than some undergrads: LDA from early 2000s

A guiding challenge:

Can we leverage sophisticated modeling techniques without losing the advantages that popularize topic models and recreating some of the same bad community practices?

Aside: Traditional Researchers are Knowledge Workers

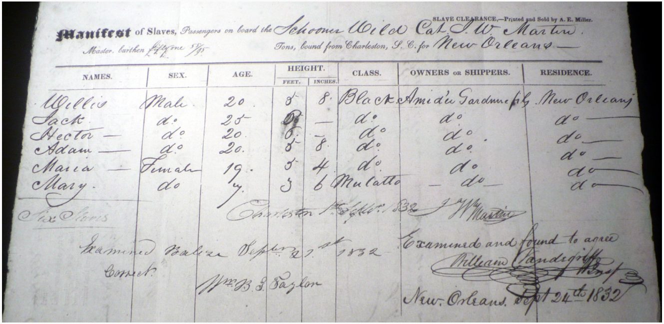
Financial analysts, investigative reporters . . .

- Concerned with specific domains
- Need to gather, build, and understand datasets
- Wide range of technical abilities
- The DH story is relevant to industry, government, etc

Motivating study: Post-Atlantic Slave Trade

Shipping manifests

Shipping manifests



SLAVE CLEARANCE,—Printed and Sold by A. E. Miller.

Manifest of Slaves, Passengers on board the Schooner Wild Cat J. W. Martin, Master, burthen fifty one 5/95 Tons, bound from Charleston, S. C. for New Orleans

NAMES.	SEX.	AGE.	HEIGHT. FEET.	HEIGHT. INCHES.	CLASS.	OWNERS OR SHIPPERS.	RESIDENCE.
Willis	Male	20	5	8	Black	Amidée Gardner fils	New Orleans
Jack	do	25	6	—	do	do	do
Hector	do	20	6	—	do	do	do
Adam	do	20	5	8	do	do	do
Maria	Female	19	5	4	do	do	do
Mary	do	7	3	6	Mulatto	do	do

Six Slaves

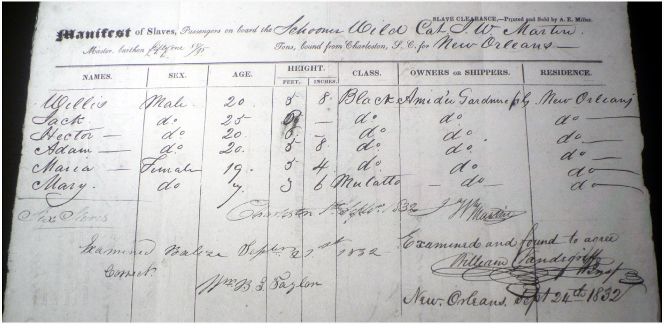
Charleston 1st Septr. 1832 J. W. Martin

Examined Balize Septr. 27th 1832 Correct. Wm. B. G. Taylor

Examined and found to agree William Vandegrift

New Orleans Sept 24th 1832

Shipping manifests


SLAVE CLEARANCE.—Printed and Sold by A. E. Miller.

Manifest of Slaves, Passengers on board the Schooner Wild Cat J. W. Martin Master, burthen fifty one 5/95 Tons, bound from Charleston, S. C. for New Orleans

NAMES.	SEX.	AGE.	HEIGHT. FEET.	HEIGHT. INCHES.	CLASS.	OWNERS or SHIPPERS.	RESIDENCE.
Willis	Male	20	5	8	Black	Ami du Gardner fils	New Orleans
Jack	do	25	[illegible]	—	do	do	do
Hector	do	20	6	—	do	do	do
Adam	do	20	5	8	do	do	do
Maria	Female	19	5	4	do	do	do
Mary	do	7	3	6	Mulatto	do	do

Six Slaves

Charleston 1st Septr 1832 J. W. Martin

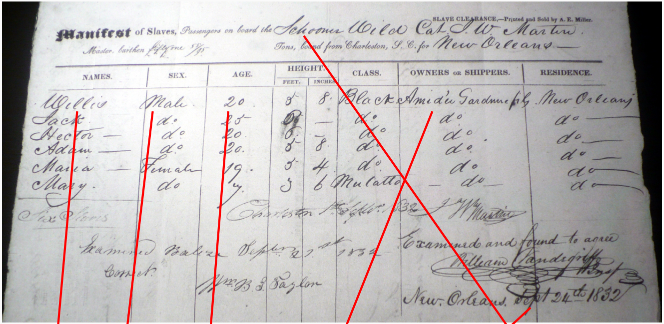
Examined Balize Septr 27th 1832 Correct. Wm B. G. Taylor

Examined and found to agree William Vandegriff

New Orleans Sept 24th 1832

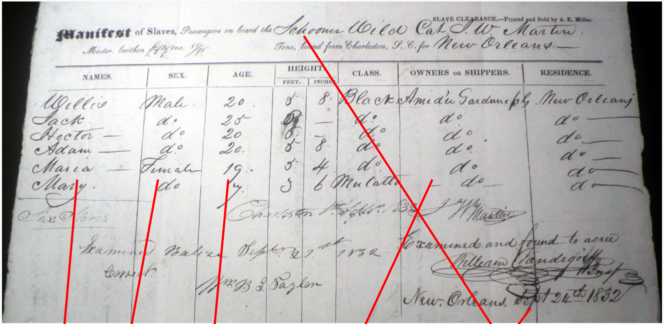
slave name	slave sex	slave age	owner name	journey date	vessel type

Shipping manifests



slave name	slave sex	slave age	owner name	journey date	vessel type
Willis	m	20	Amidu	1832/9/24	Schooner

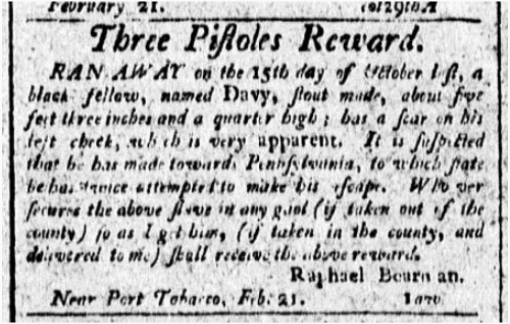
Shipping manifests



slave name	slave sex	slave age	owner name	journey date	vessel type
Willis	m	20	Amidu	1832/9/24	Schooner
Maria	f	19	Amidu	1832/09/24	Schooner

Fugitive notices

Fugitive notices



February 21.

Three Pistoles Reward.

RAN AWAY on the 15th day of October last, a black fellow, named Davy, stout made, about five feet three inches and a quarter high; has a scar on his left cheek, which is very apparent. It is suspected that he has made towards Pennsylvania, to which state he has twice attempted to make his escape. Whoever secures the above slave in any gaol (if taken out of the county) so as I get him, (if taken in the county, and delivered to me) shall receive the above reward.

Raphael Bourman.

Near Port Tobacco, Feb. 21. [illegible]

Fugitive notices

February 21.

Three Pistoles Reward.

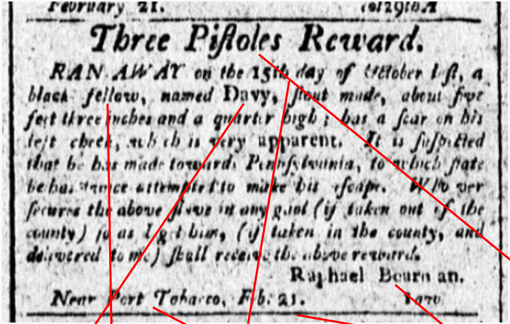
RAN AWAY on the 15th day of October last, a black fellow, named Davy, stout made, about five feet three inches and a quarter high; has a scar on his left cheek, which is very apparent. It is suspected that he has made towards Pennsylvania, to which state he has twice attempted to make his escape. Whoever secures the above slave in any gaol (if taken out of the county) so as I get him, (if taken in the county, and delivered to me) shall receive the above reward.

Raphael Boarman.

Near Port Tobacco, Feb. 21. [illegible]

slave name	slave sex	escape date	escape location	owner name	notice reward	notice date

Fugitive notices



February 21.

Three Pistoles Reward.

RAN AWAY on the 15th day of October last, a black fellow, named Davy, stout made, about five feet three inches and a quarter high; has a scar on his left cheek, which is very apparent. It is suspected that he has made towards Pennsylvania, to which state he has twice attempted to make his escape. Whoever secures the above slave in any gaol (if taken out of the county) so as I get him, (if taken in the county, and delivered to me) shall receive the above reward.

Raphael Bourman.

Near Port Tobacco, Feb. 21.

slave name	slave sex	escape date	escape location	owner name	notice reward	notice date
Davy	m	1795/10/15	Port Tobacco	Bourman	3 Pistoles	1796/02/21

Some numbers

- 45k manifest entries spanning five cities
- 11k fugitive notices from 70 gazettes
- 28k unique slave names
- 7k unique owner names
- Not big data, but thousands of studies like this at a research university!

Difficulties with data in the wild

Difficulties with data in the wild

- Unnormalized
  - People/places/things recorded *many* times
  - “What’s the age/height/sex distribution of escapees?”

Difficulties with data in the wild

- Unnormalized
  - People/places/things recorded *many* times
  - “What’s the age/height/sex distribution of escapees?”
- Noisy
  - Vessel type: Bark, Barke, BArque, Barque, Barques
  - Slave name: “Nelly’?, Nelly’s child”, “not visible”
  - Owner sex: 3k missing

Difficulties with data in the wild

- Unnormalized
  - People/places/things recorded *many* times
  - “What’s the age/height/sex distribution of escapees?”
- Noisy
  - Vessel type: Bark, Barke, BArque, Barque, Barques
  - Slave name: “Nelly’?, Nelly’s child”, “not visible”
  - Owner sex: 3k missing
- Underspecified entities
  - *Majority* of slaves have no last name
  - Can’t tell if two “Johns” are the same person

What might a historian want to do with this data?

- Follow one slave throughout their life
- Group owners according to the nature of their workforce
- Determine what drove valuation in transactions and rewards
- Reconstruct slave families when there are no last names
- Map out trade “ecosystems” of sellers, shippers, owners, etc

Fundamental observation

There is an implicit *database schema* here

- **Field**: a recorded *value* with a clear interpretation (age, name, manufacturer . . .)
- **Entity-type**: a coherent *bundle of fields* (person, location, object . . .)
- Entity-types and fields have been determined by traditional scholars and common sense
- Relations *between* entities are also (conservatively) implied by the tabular format

Fundamental observation

There is an implicit *database schema* here

- **Field**: a recorded *value* with a clear interpretation (age, name, manufacturer . . .)
- **Entity-type**: a coherent *bundle of fields* (person, location, object . . .)
- Entity-types and fields have been determined by traditional scholars and common sense
- Relations *between* entities are also (conservatively) implied by the tabular format

This sets things up so we (ML researchers) can tackle the *general* problem

Entities, field types, and relations

Traditional scholarly data

slave_name	*Jim*
slave_age	20
owner_name	*Jane*
owner_sex	*F*
vessel_name	*Uncas*
vessel_type	*Brig*
voyage_date	6/2/1823
voyage_dest	29.9, 90.0
...	...

Numbers

slave_name	*Jim*
slave_age	20
owner_name	*Jane*
owner_sex	*F*
vessel_name	*Uncas*
vessel_type	*Brig*
voyage_date	6/2/1823
voyage_dest	29.9, 90.0
...	...

Entities, field types, and relations

Categories

slave_name	*Jim*
slave_age	20
owner_name	*Jane*
owner_sex	*F*
vessel_name	*Uncas*
vessel_type	*Brig*
voyage_date	6/2/1823
voyage_dest	29.9, 90.0
...	...

Entities, field types, and relations

Strings

slave_name	*Jim*
slave_age	20
owner_name	*Jane*
owner_sex	*F*
vessel_name	*Uncas*
vessel_type	*Brig*
voyage_date	6/2/1823
voyage_dest	29.9, 90.0
...	...

More complex fields

slave_name	*Jim*
slave_age	20
owner_name	*Jane*
owner_sex	*F*
vessel_name	*Uncas*
vessel_type	*Brig*
voyage_date	6/2/1823
voyage_dest	29.9, 90.0
...	...

Entities, field types, and relations

Entities

$slave\_name$	Jim
$slave\_age$	20
$owner\_name$	$Jane$
$owner\_sex$	F
$vessel\_name$	$Uncas$
$vessel\_type$	$Brig$
$voyage\_date$	$6/2/1823$
$voyage\_dest$	$29.9, 90.0$
...	...

Entities, field types, and relations

Entities

slave_name	*Jim*
slave_age	20
owner_name	*Jane*
owner_sex	*F*
vessel_name	*Uncas*
vessel_type	*Brig*
voyage_date	6/2/1823
voyage_dest	29.9, 90.0
...	...

Entities, field types, and relations

Slave-to-owner

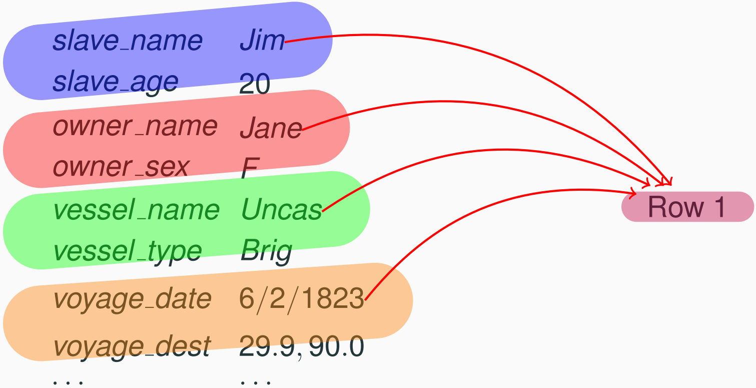
slave_name	*Jim*
slave_age	20
owner_name	*Jane*
owner_sex	*F*
vessel_name	*Uncas*
vessel_type	*Brig*
voyage_date	6/2/1823
voyage_dest	29.9, 90.0
...	...

Vessel-to-voyage, slave-to-voyage

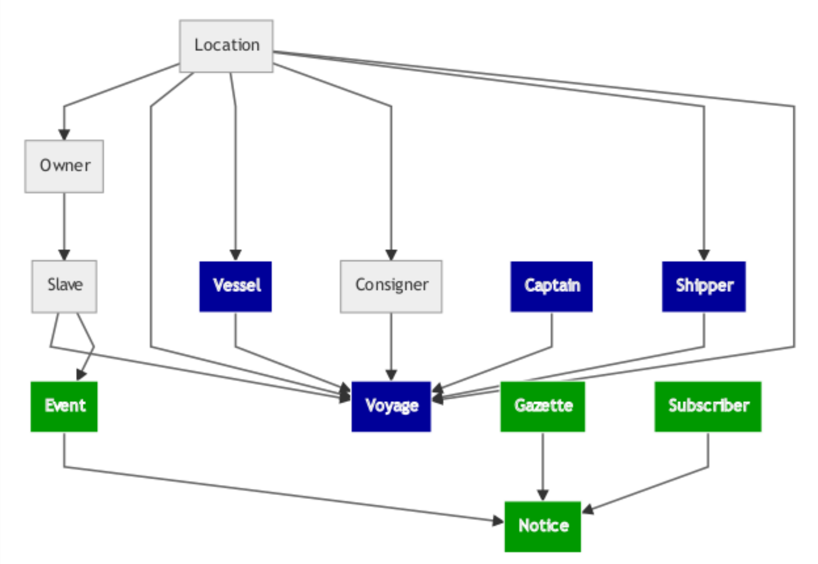
slave_name	*Jim*
slave_age	20
owner_name	*Jane*
owner_sex	*F*
vessel_name	*Uncas*
vessel_type	*Brig*
voyage_date	6/2/1823
voyage_dest	29.9, 90.0
...	...

Entities, field types, and relations

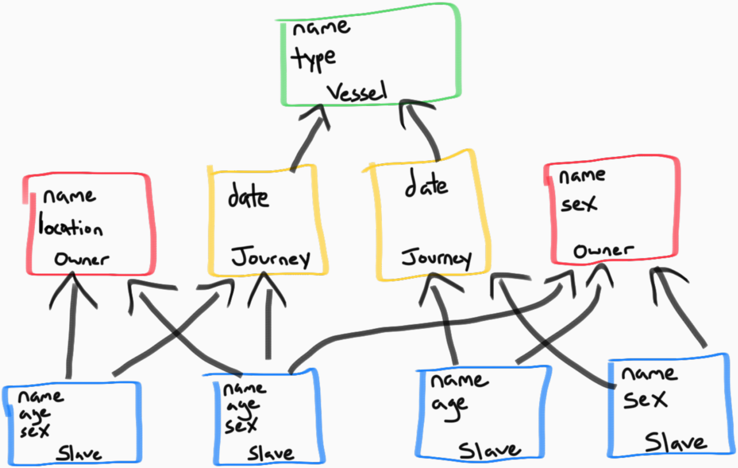
Fewer assumptions



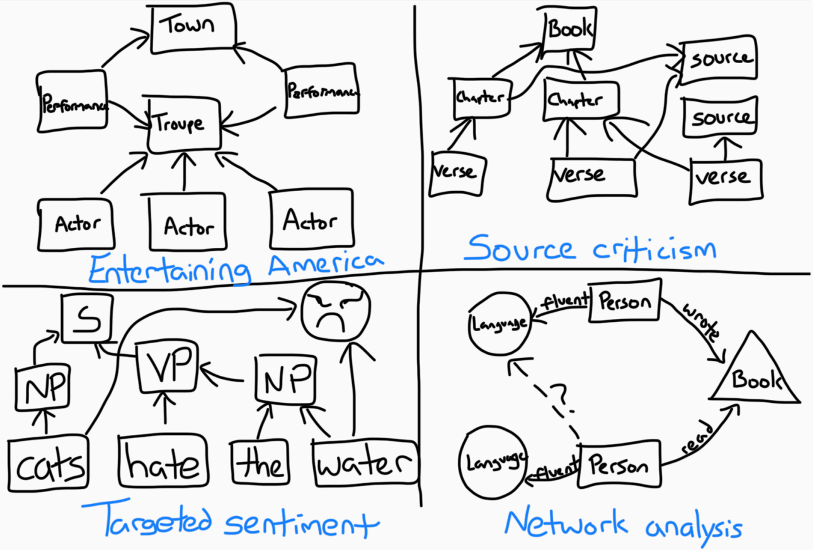
Full schema of possible entity relationships



Example data point: one graph component



Subsumes studies from a wide range of domains



Model: Graph-Entity Autoencoders

General scholarly questions

- What is the overall picture of a *particular entity*?
- In what ways can we *group* entities?
- How are fields *correlated*?
- What *missing fields and relationships* can be recovered?

General scholarly questions

- What is the overall picture of a *particular entity*?
- In what ways can we *group* entities?
- How are fields *correlated*?
- What *missing fields and relationships* can be recovered?

Three basic operations we'd like:

- Measure similarity of two entities
- Generate plausible field-values
- Score a proposed relationship between two entities

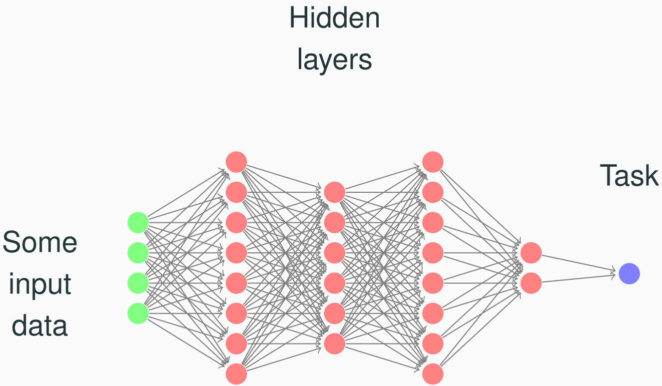
Building-blocks of the model

Encoders, decoders, and autoencoders

Graph convolutional networks

; ; ; ;

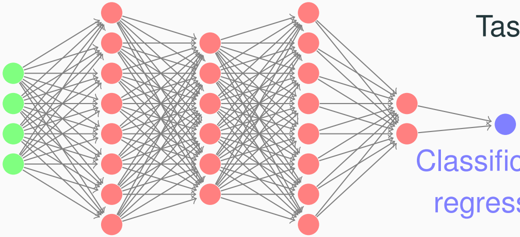
Basic feed-forward neural model



Basic feed-forward neural model

Hidden
layers

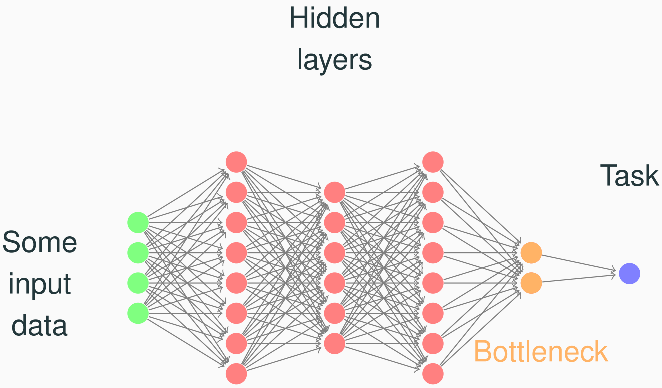
Some
input
data



Task

Classification,
regression,
generation,
some combo ...

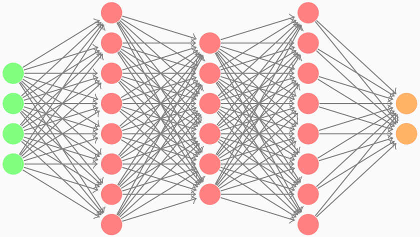
Basic feed-forward neural model



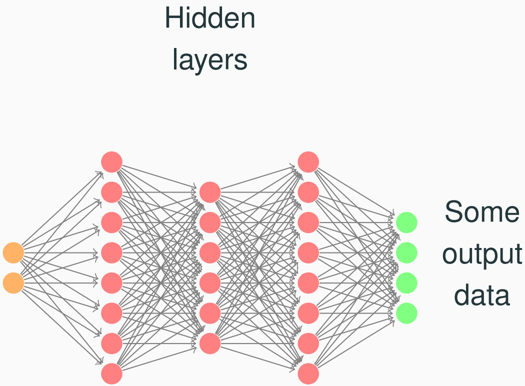
Basic feed-forward neural model (an “encoder”)

Hidden
layers

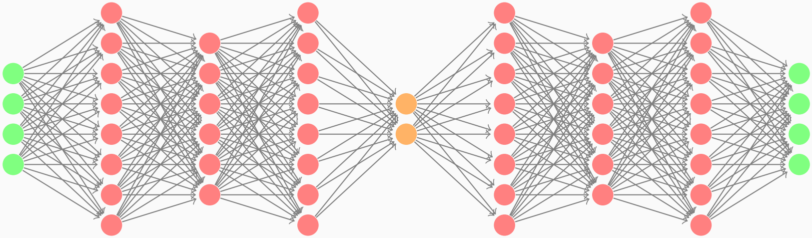
Some
input
data



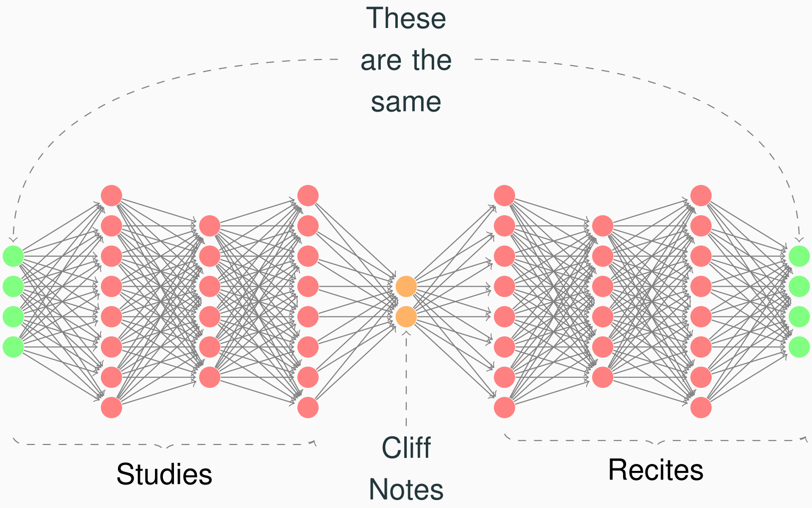
A “decoder” goes in the opposite direction



Encoders and decoders are often paired



If the goal is to reconstruct the input, it's an "autoencoder"

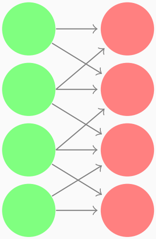


Convolutional networks (CNNs)

Grid
(image,
text . . .)

Convolutional networks (CNNs)

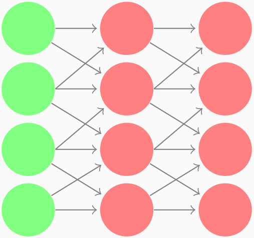
Grid (image, text . . .)



Each position incorporates its “receptive field”

Convolutional networks (CNNs)

Grid (image, text . . .)



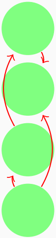
Repeat process, expand field

Graph-convolutional networks (GCNs)

Graph nodes (e.g. entities)

Graph-convolutional networks (GCNs)

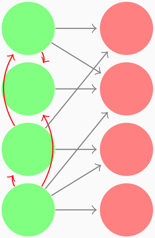
Graph nodes (e.g. entities)



Adjacent nodes (related entities)

Graph-convolutional networks (GCNs)

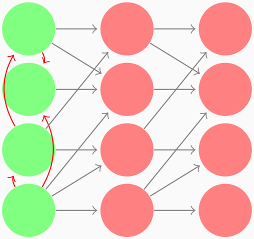
Graph nodes (e.g. entities)



Each node incorporates its neighbors

Graph-convolutional networks (GCNs)

Graph nodes (e.g. entities)

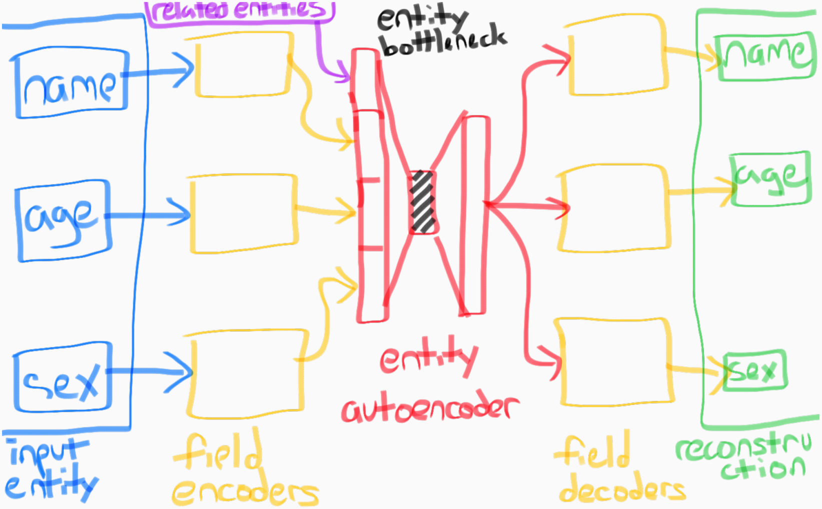


Info spreads according to graph

Full model summary

- Read data, determine:
  - Fields and field-types
  - Entity-types
  - Relationships
- Each field allocated appropriate encoder-decoder pair
- Each entity-type allocated autoencoder
- Autoencoders use GCN-like mechanism to incorporate adjacent bottlenecks

Model sketch



Training is a complex process

- Random field dropout
- Graph component subselection
- Ways to combine loss functions
- …

How can we use a trained model?

- Compute distance between two entities
- Find flat or hierarchical clusters of entities
- Generate likely value of missing field
- Detect an improbable value of a present field
- Observe response of one field to another

Example insights looking at most-similar entities

Mistranscriptions

Baltiomre, Austin Woolfolk	$\Longleftrightarrow$	Baltimore, Austin Woolfolk
New Orleans, William Wiliams	$\Longleftrightarrow$	New Orleans, William Williams

Semantically-equivalent variants

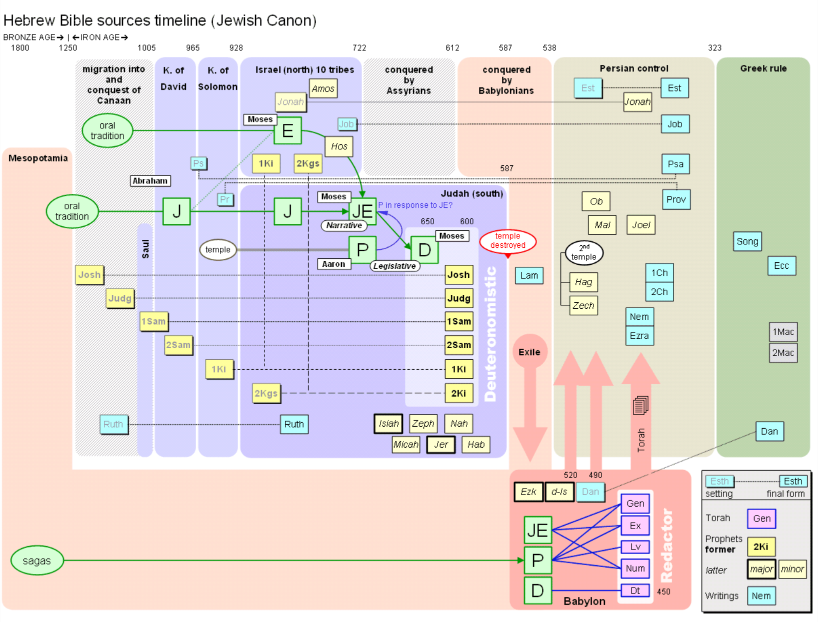
Baltimore, George Y. Kelso	$\Leftrightarrow$	Baltimore, Kelso & Ferguson
New Orleans, Leon Chabert	$\Leftrightarrow$	Louisiana, Leon Chabert

Same slave transported multiple times

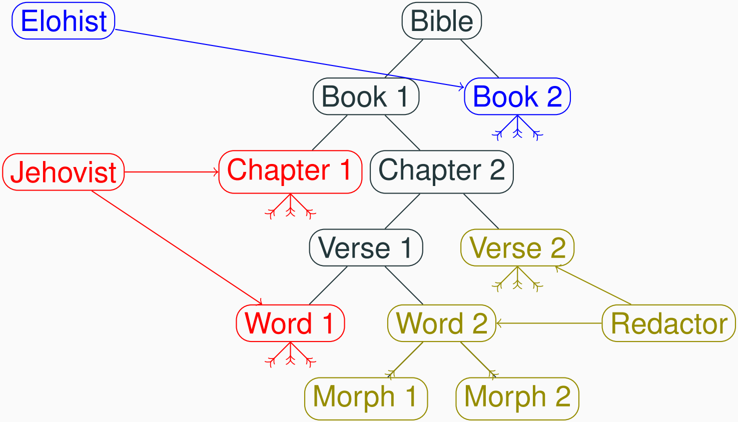
Louisa, F, 16yo	$\Leftrightarrow$	Louisa, F, 17yo
Waters, F, 14yo	$\Leftrightarrow$	Waters, F, 15yo
Kesiah, F, 20yo	$\Leftrightarrow$	Kesiah, F, 22yo
Taylor, F, 15yo	$\Leftrightarrow$	Taylor, F, 16yo

Bonus study: Authorship attribution of ancient documents

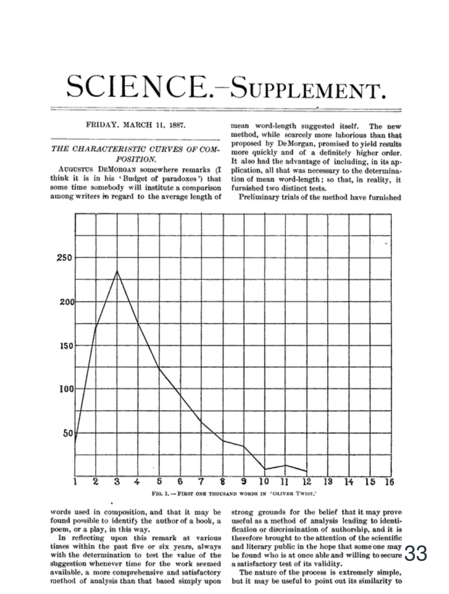
Transmission of a text: the "Documentary Hypothesis"



Hypothesis as pointers into document structure



Thomas Mendenhall: The Characteristic Curves of Composition



SCIENCE.—SUPPLEMENT.

FRIDAY, MARCH 11, 1887.

THE CHARACTERISTIC CURVES OF COMPOSITION.

AUGUSTUS DEMORGAN somewhere remarks (I think it is in his 'Budget of paradoxes') that some time somebody will institute a comparison among writers in regard to the average length of words used in composition, and that it may be found possible to identify the author of a book, a poem, or a play, in this way.

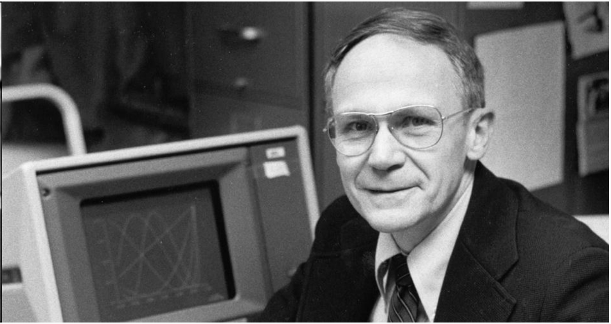
In reflecting upon this remark at various times within the past five or six years, always with the determination to test the value of the suggestion whenever time for the work seemed available, a more comprehensive and satisfactory method of analysis than that based simply upon mean word-length suggested itself. The new method, while scarcely more laborious than that proposed by DeMorgan, promised to yield results more quickly and of a definitely higher order. It also had the advantage of including, in its application, all that was necessary to the determination of mean word-length; so that, in reality, it furnished two distinct tests.

Preliminary trials of the method have furnished strong grounds for the belief that it may prove useful as a method of analysis leading to identification or discrimination of authorship, and it is therefore brought to the attention of the scientific and literary public in the hope that some one may be found who is at once able and willing to secure a satisfactory test of its validity.

The nature of the process is extremely simple, but it may be useful to point out its similarity to

FIG. 1.—FIRST ONE THOUSAND WORDS IN 'OLIVER TWIST.'

Mosteller and Wallace: Inference in an Authorship Problem



The Federalist papers

- 85 articles written by Hamilton, Madison, and Jay
- 12 are unattributed
- Frequency analysis of *function words* determined Madison as author

Back to the Documentary Hypothesis

Problems

- The “authors” are also editors, redactors, synthesizers . . . they interact in context-dependent ways
- There is no predefined segmentation into “articles”
- We *know* more than function-words are important (e.g. name of God)

Solutions

- Limit vocabulary to words that are used frequently by all authors
- Employ a GCN to exploit the document structure
- Take the DH for granted (for now)

GEA predicts the author slightly better ...

Model	F-score
LR	41.39
MLP	47.45
GEA	48.60

Gold	Guess							
	J	E	P	1D	2D	nD	R	O
J	100	8	7	0	0	0	3	0
E	22	53	8	0	0	0	0	0
P	13	5	77	0	1	0	4	0
1D	2	0	2	7	1	0	0	0
2D	2	2	1	0	5	0	0	0
nD	0	0	0	1	0	0	0	0
R	3	3	11	0	0	0	33	0
O	2	0	1	0	0	0	1	0

Error analysis

Sentiment and in-context word senses

- “wife” shows up as polygamous in older but monogamous in newer sources
- Redactor’s positive view of Aaron+Moses, violent story of rebellion

Narrative continuity

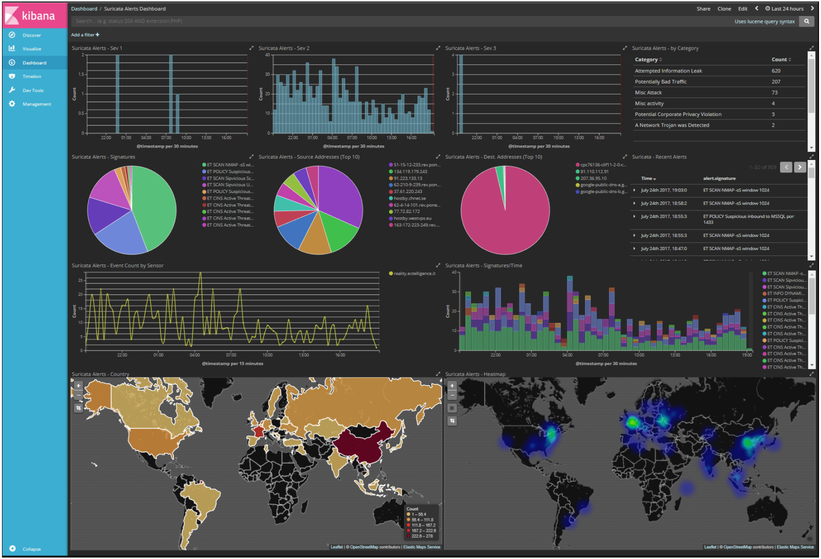
- Abraham and Isaac story thought to *originally end with sacrifice*, changed by the Redactor
- “it was the season for grapes. ¡travel and geographic locations¿ They broke off some grapes.”

Recipe for literary criticism

- Collect or construct useful resources from traditional scholarship
- Determine fit of potential “compositional actions” to observed document tree
- Choose the actions that are high-scoring *and* parsimonious
- Put the hypothesis in front of domain experts for verification/annotation

Ongoing work

Visualizing results



Assembling other example studies

- JHU history department's "Entertaining America" (tabular)
- Northeastern U's Women Writers Collection (XML/TEI)
- Targeted sentiment analysis (JSON)
- Tennyson's poetic development (unconstrained text)

Thanks!

Quick plug: come to David Mimno's talk!

- Nov. 15 at noon (Hackerman B17)
- CS Professor at Cornell
- Rare CS faculty working in DH (topic modeling)