Computational Intelligence for the Humanities

Tom Lippincott (tom@cs.jhu.edu)

Background: Working with humanists

Starcoder: A neural ensemble for humanities research

Example: The post-Atlantic slave trade

Ongoing work

Background: Working with humanists

Downstream users of NLP/ML (“knowledge workers”)

Downstream users of NLP/ML (“knowledge workers”)

Finance

Medicine

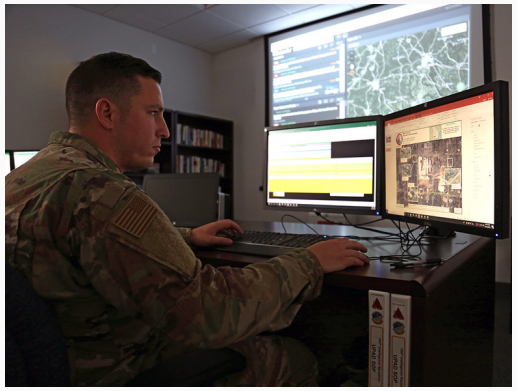
Military

Journalism

Downstream users of NLP/ML (“knowledge workers”)

Finance

Medicine



Military

Journalism

- Concerned with specific domains, esoteric questions
- Need to gather and structure targeted data
- Explore and reason over those structures
- Find and present a coherent, interesting narrative
- Wide range of technical abilities

Downstream users of NLP/ML ("knowledge workers")

Finance

Medicine

Military

Journalism

Humanist scholars

Why might computational intelligence help humanists?

Why might computational intelligence help humanists?

Scholars are capable of deep, focused reasoning, *but* . . .

- have baggage from centuries of received wisdom
- are often deeply invested in supporting/refuting a conclusion
- not possible to attend evenly to data and hypotheses

Why might computational intelligence help humanists?

Scholars are capable of deep, focused reasoning, *but* . . .

- have baggage from centuries of received wisdom
- are often deeply invested in supporting/refuting a conclusion
- not possible to attend evenly to data and hypotheses

AI has limited capacity for abstract reasoning, *but* . . .

- Easier to avoid bias towards a particular conclusion
- Can uniformly process data and explore hypothesis-spaces
- May produce well-formed probabilities, well-specified representations

Two broad categories of “help”

Supervised, focused tasks (“I can’t label/transcribe/etc all this data!”)

Unsupervised, exploratory (“What’s interesting about this data?”)

Two broad categories of “help”

Supervised, focused tasks (“I can’t label/transcribe/etc all this data!”)

- Starts with careful task definition
- Requires (maybe lots of) example pairs (datum, label/transcription)
- OCR, translation, any sort of predefined categorization, annotation

Unsupervised, exploratory (“What’s interesting about this data?”)

Two broad categories of “help”

Supervised, focused tasks (“I can’t label/transcribe/etc all this data!”)

- Starts with careful task definition
- Requires (maybe lots of) example pairs (datum, label/transcription)
- OCR, translation, any sort of predefined categorization, annotation

Unsupervised, exploratory (“What’s interesting about this data?”)

- Starts with careful, unbiased representations
- No specific task, minimal bias and annotation cost
- Anomaly detection, topic models, clustering, intrinsic graph metrics

Two broad categories of “help”

Supervised, focused tasks (“I can’t label/transcribe/etc all this data!”)

- Starts with careful task definition
- Requires (maybe lots of) example pairs (datum, label/transcription)
- OCR, translation, any sort of predefined categorization, annotation

Unsupervised, exploratory (“What’s interesting about this data?”)

- Starts with careful, unbiased representations
- No specific task, minimal bias and annotation cost
- Anomaly detection, topic models, clustering, intrinsic graph metrics

Broad goals for collaboration

Minimal disruption to existing practice

“Sustainability”

Broad goals for collaboration

Minimal disruption to existing practice

- Humanists can keep working directly with primary sources as they always have
- Computational scholars can develop and swap in new architectures and techniques

“**Sustainability**”

Broad goals for collaboration

Minimal disruption to existing practice

- Humanists can keep working directly with primary sources as they always have
- Computational scholars can develop and swap in new architectures and techniques

“**Sustainability**”

- Can’t manage dozens of bespoke collaborations
- Need to focus on the boundaries: data in, hypotheses out
- Start with simple model that’s easily expanded

Starcoder: A neural ensemble for humanities research

Observation: spreadsheets are everywhere

owner_name	owner_job	owner_age	car_make	car_price	dealer_name
John	Lawyer	23	Honda	25000	Crazy Ray's
Jane	Doctor	31	Ford	35000	Crazy Ray's
...	...	...	...	...	...

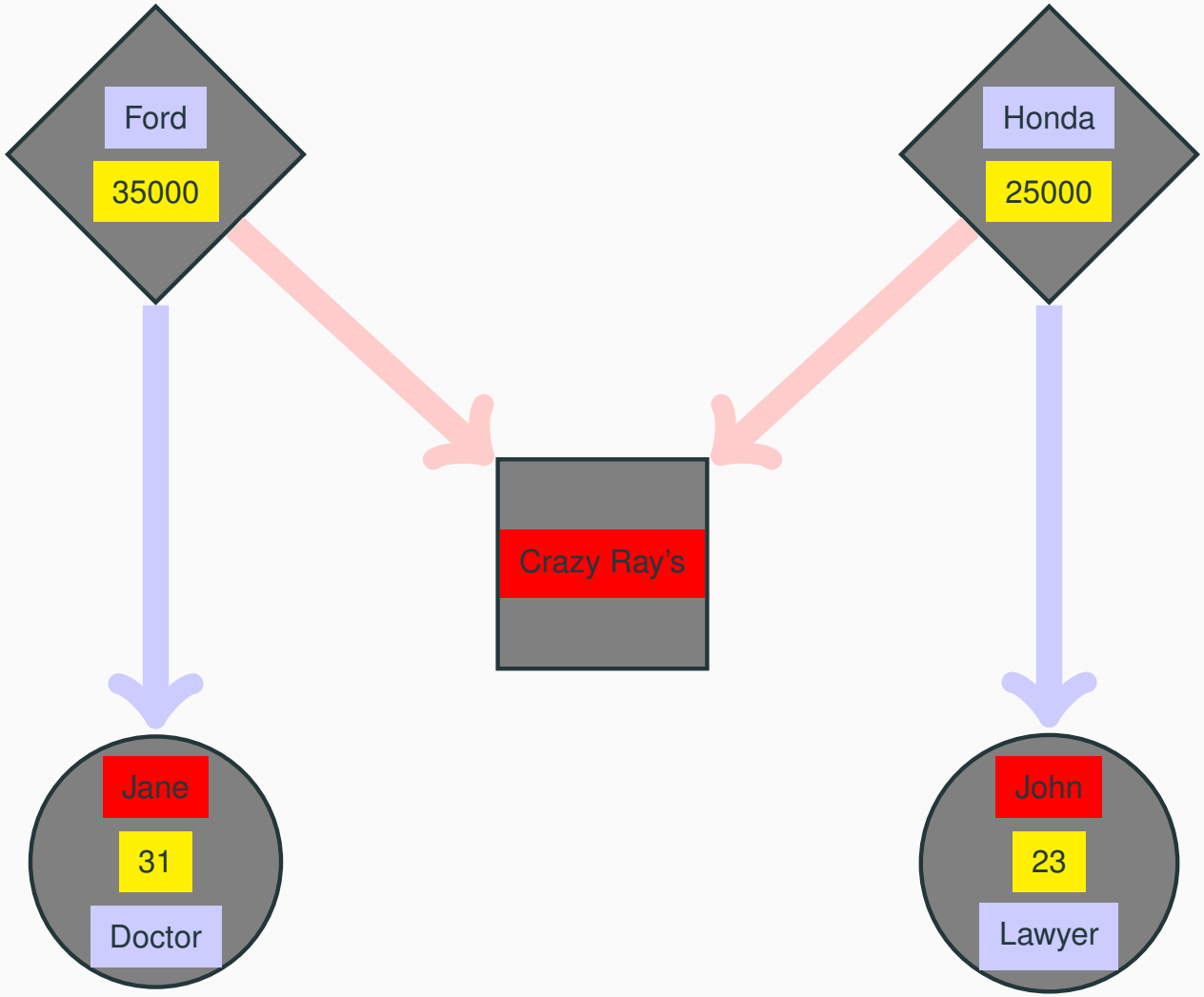
Observation: spreadsheets are everywhere

owner_name	owner_job	owner_age	car_make	car_price	dealer_name
John	Lawyer	23	Honda	25000	Crazy Ray's
Jane	Doctor	31	Ford	35000	Crazy Ray's
...	...	...	...	...	...

Entities with properties and relationships

- **Property-type** (e.g. text, category, number, image, location...)
- An **Entity-type** is a coherent *bundle of potential properties* (e.g. an *owner* has a name, job, and age)
- A **Relationship-type** is a predicate with a specific interpretation that might link entities of the appropriate entity-types (e.g. owner_of, sold_by)

Graph view of the same data



Humanist describes their domain in a *schema*

owner_name	owner_job	owner_age	car_make	car_price	dealer_name
John	Lawyer	23	Honda	25000	Crazy Ray's
Jane	Doctor	31	Ford	35000	Crazy Ray's
...	...	...	...	...	...

Humanist describes their domain in a *schema*

owner_name	owner_job	owner_age	car_make	car_price	dealer_name
John	Lawyer	23	Honda	25000	Crazy Ray's
Jane	Doctor	31	Ford	35000	Crazy Ray's
...	...	...	...	...	...

```
{
  "properties" : {
    "owner_age" : {"type" : "scalar", "meaning" : "human_age"},
    ...
  },
  "entity_types" : {
    "owner" : ["owner_name", "owner_job", ...]
    ...
  },
  "relationships" : {
    "owned_by" : {
      "source_entity_type" : "car",
      "target_entity_type" : "owner"
    },
    ...
  }
}
```

Schema combined with spreadsheet/XML/etc generates JSON entities

```
[
  {"type" : "owner", "name" : "John", "age" : 23,
      "job" : "Lawyer", "id" : 0},
  {"type" : "owner", "name" : "Jane", "age" : 31,
      "job" : "Doctor", "id" : 1},
  {"type" : "dealer", "dealer_name" : "Crazy Ray's",
      "id" : 2},
  {"type" : "car", "make" : "Honda", "price" : 25000,
      "owned_by" : 0, "sold_by" : 2},
  {"type" : "car", "make" : "Ford", "price" : 35000,
      "owned_by" : 1, "sold_by" : 2}
]
```

Design a model-generator that matches the schema

Building-blocks for the model

Encoder, decoder, and autoencoder mechanisms
Capture the *entities* and *fields*

Graph convolutional mechanism
Capture the *relationships*

Building-blocks for the model

Encoder, decoder, and autoencoder mechanisms
Capture the *entities* and *fields*

Graph convolutional mechanism
Capture the *relationships*

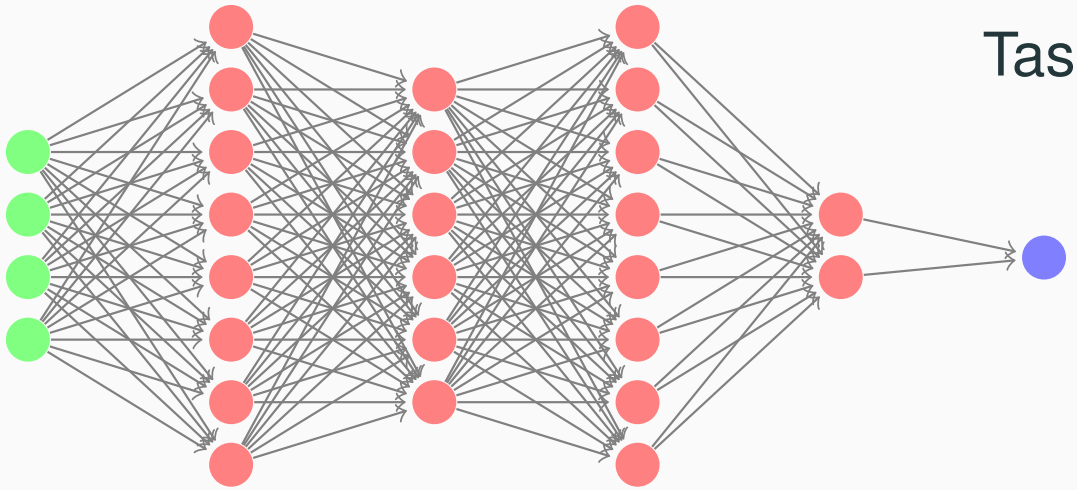
Starcoder, from the Kleene-closure (asterisk/wildcard)

Intro to encoders/decoders/autoencoders

Feed-forward network

Hidden
layers

Some
input
data

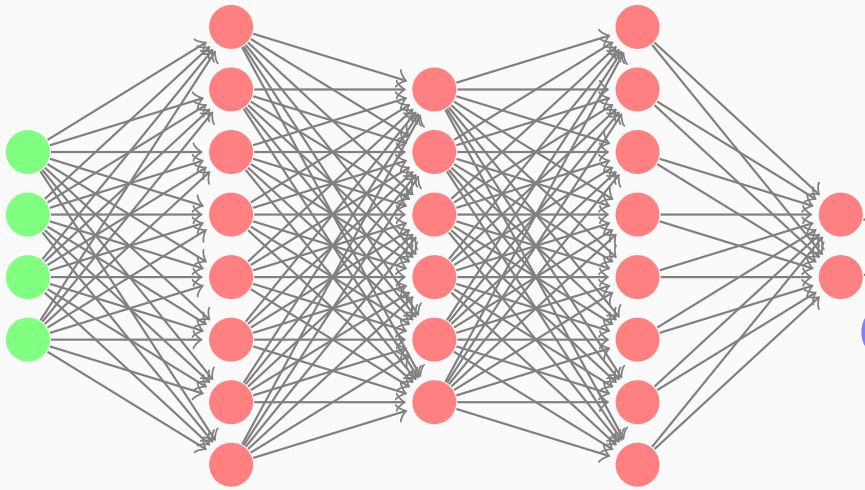


Task

Feed-forward network

Hidden
layers

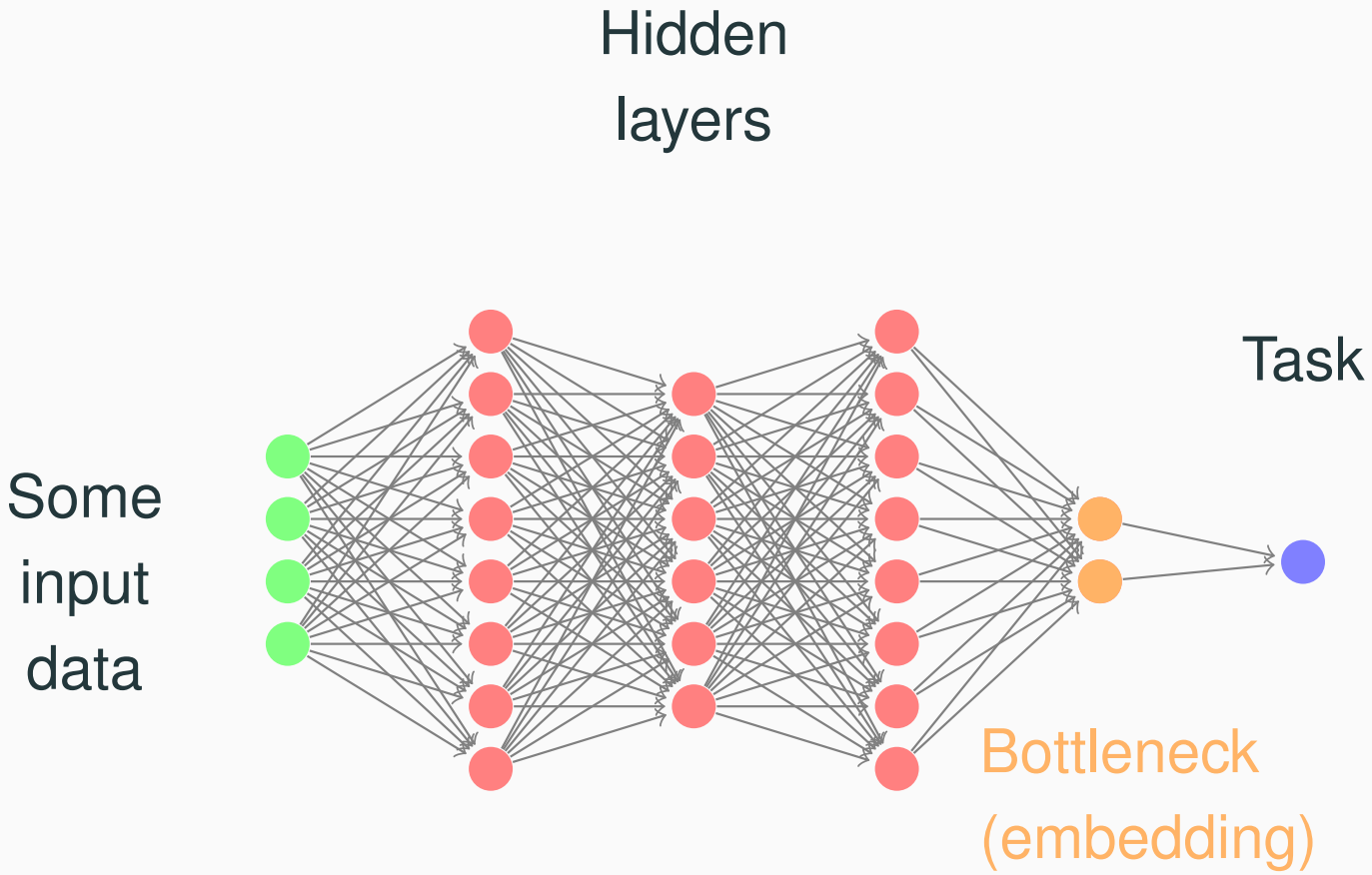
Some
input
data



Task

Classification,
regression,
generation,
some combo . . .

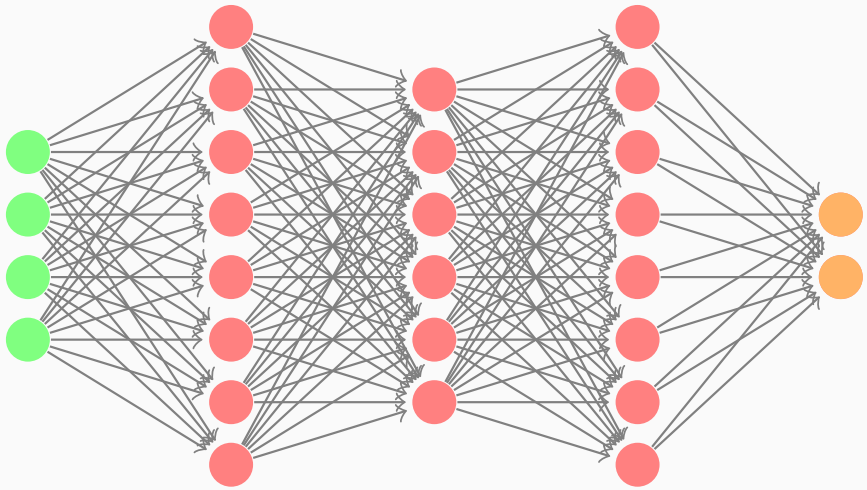
Feed-forward network



Encoder

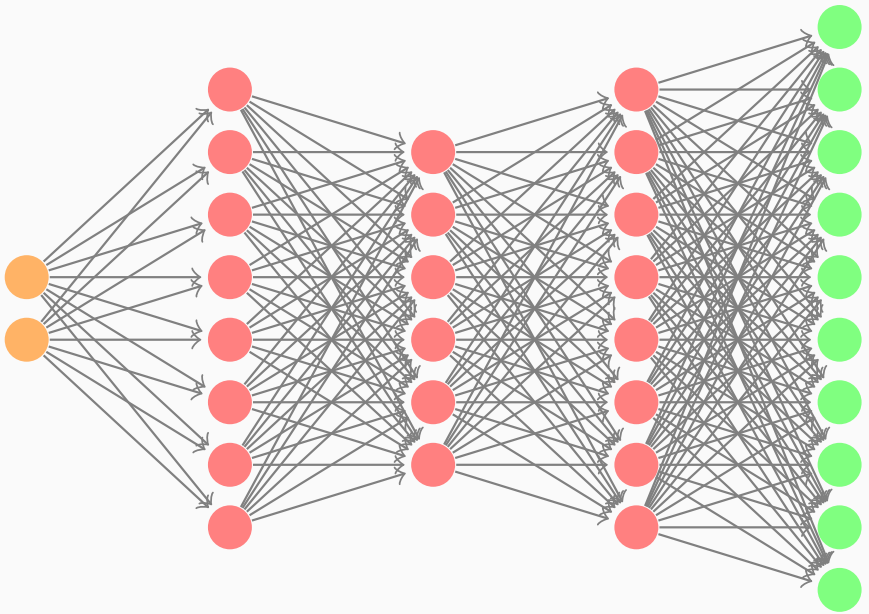
Hidden
layers

Some
input
data



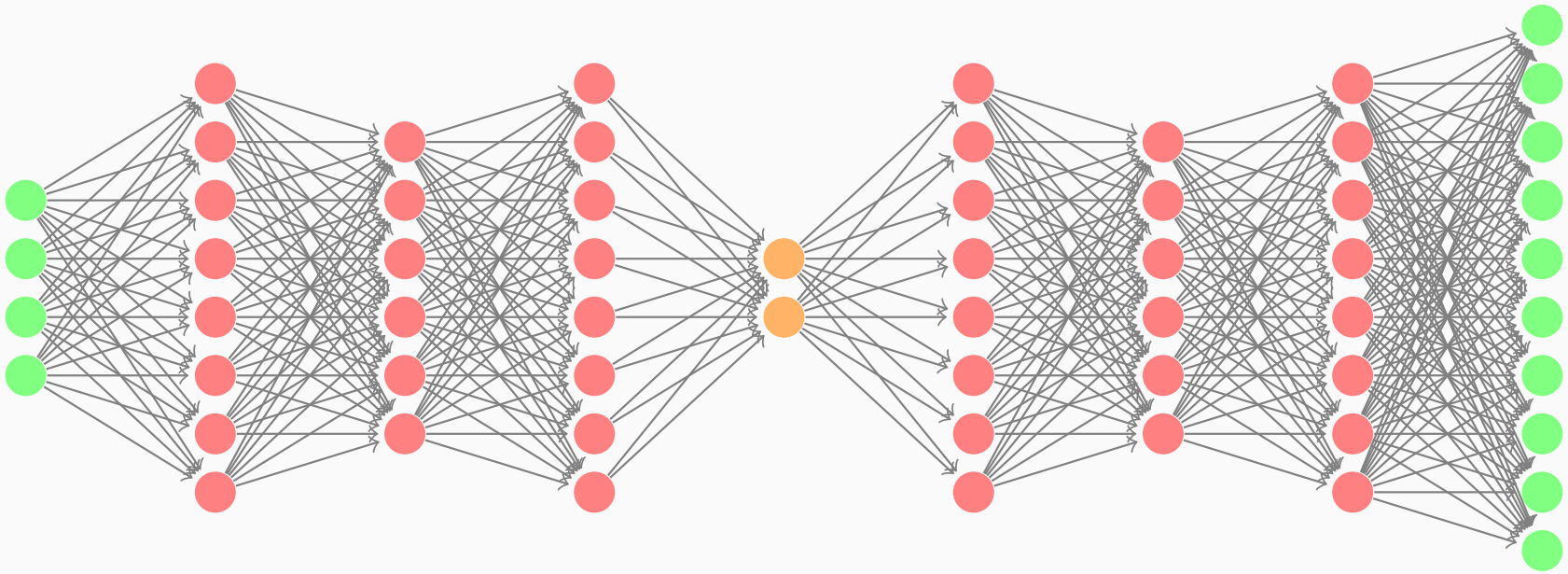
Decoder

Hidden layers

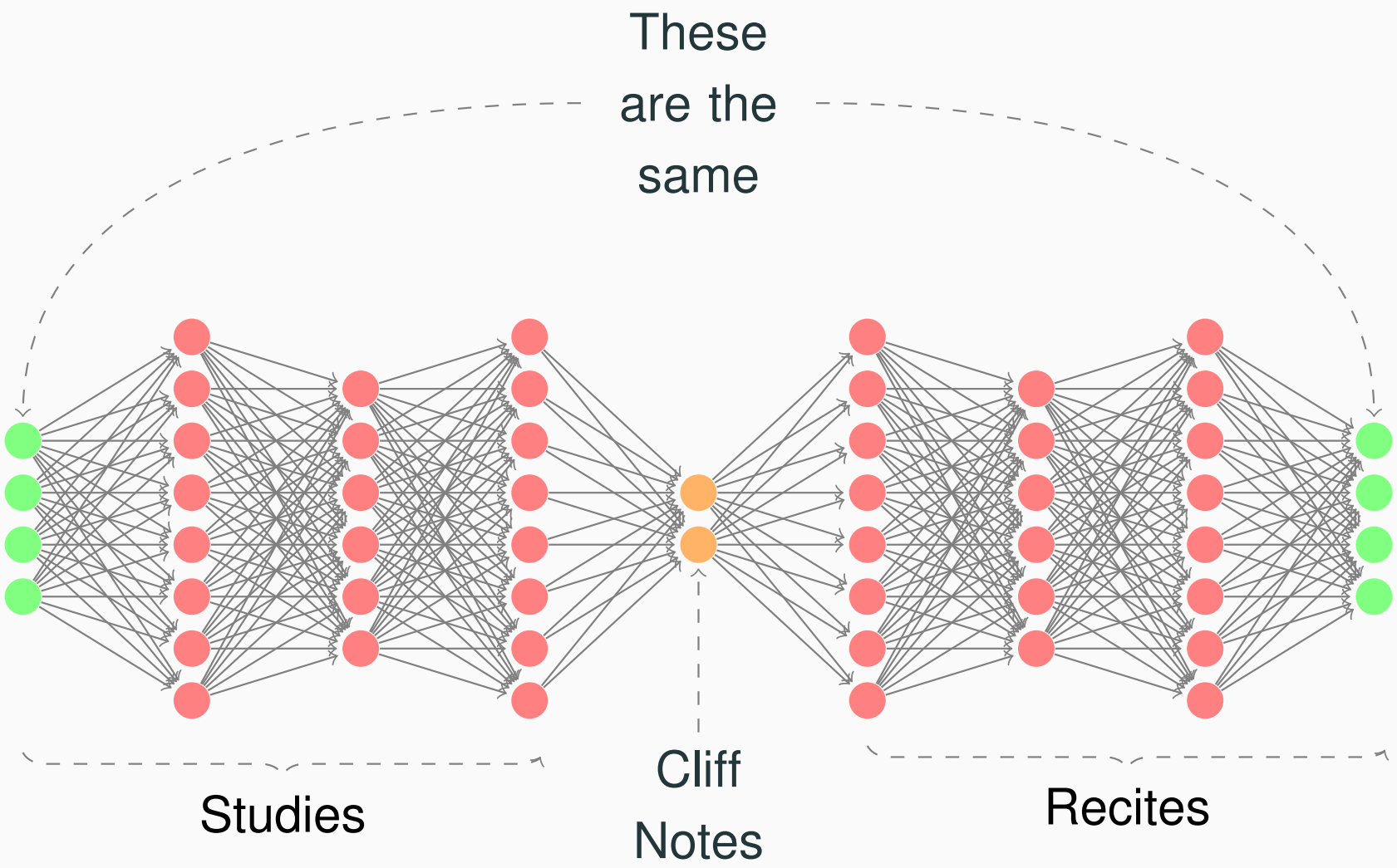


Some output data

Encoders and decoders are often paired



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



Summary of coder mechanisms

- An **encoder** transforms data into a fixed-length representation
- A **decoder** takes a fixed-length representation and generates data
- An **autoencoder** is an encoder and decoder working together to preserve data through a **bottleneck**

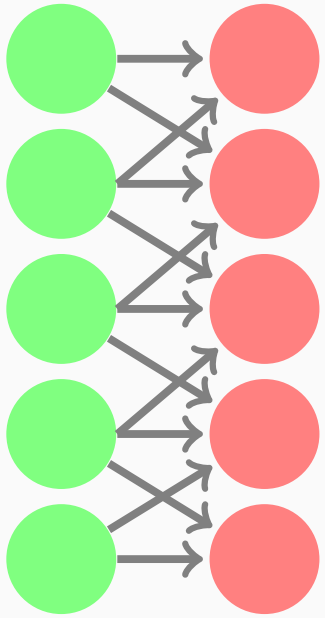
On to *graph convolutions*. . .

Normal 1D CNN

Grid (image, text ...)

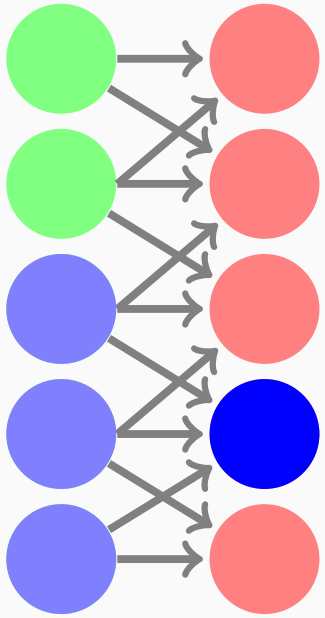
Normal 1D CNN

Grid (image, text . . .)



Normal 1D CNN

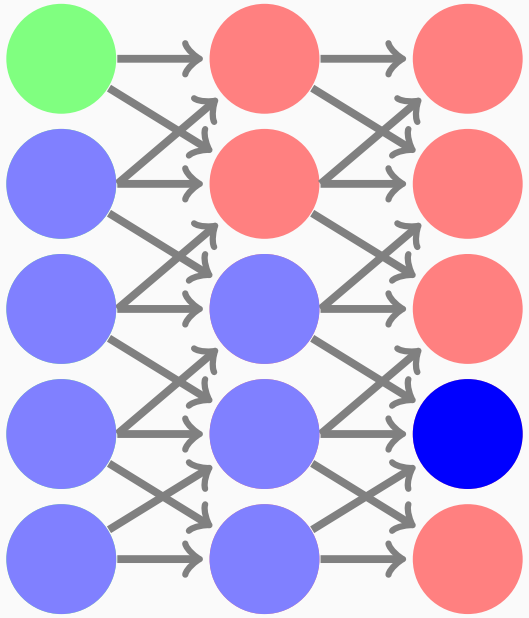
Grid (image, text . . .)



Each position incorporates its “receptive field”

Normal 1D CNN

Grid (image, text . . .)



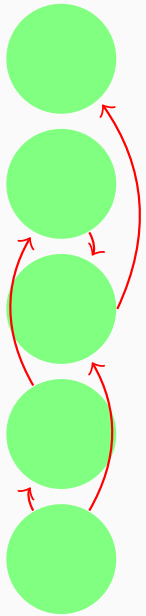
Repeat process, expand field

Graph convolutional network (GCN)

Graph nodes
(e.g. entities)

Graph convolutional network (GCN)

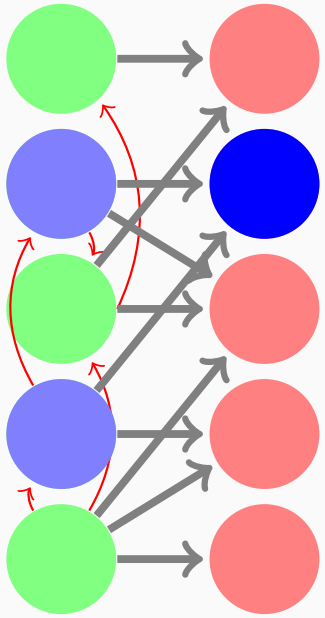
Graph nodes
(e.g. entities)



Adjacent
nodes
(related
entities)

Graph convolutional network (GCN)

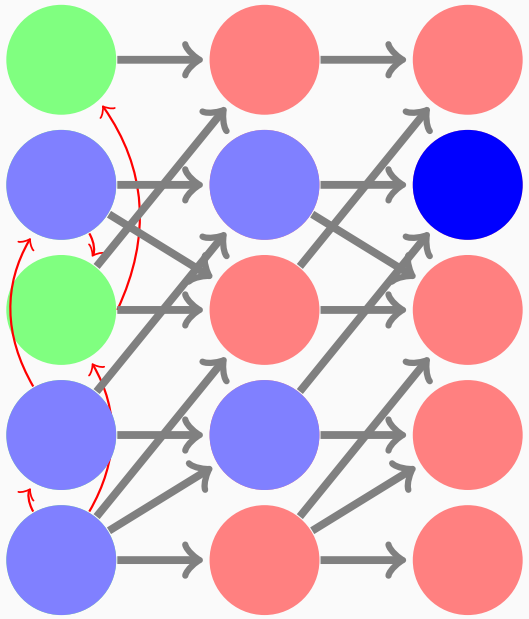
Graph nodes
(e.g. entities)



Each node
incorporates
its neighbors

Graph convolutional network (GCN)

Graph nodes
(e.g. entities)



Info spreads
according
to graph

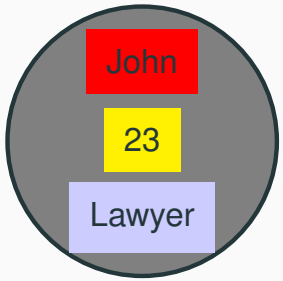
Summary of graph-convolutional mechanism

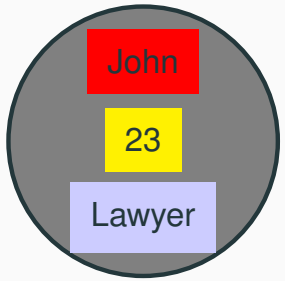
- Extends CNNs from grids to graphs
- Information can pass along edges
- Each layer allows nodes to see one further “hop”

Modeling relational data

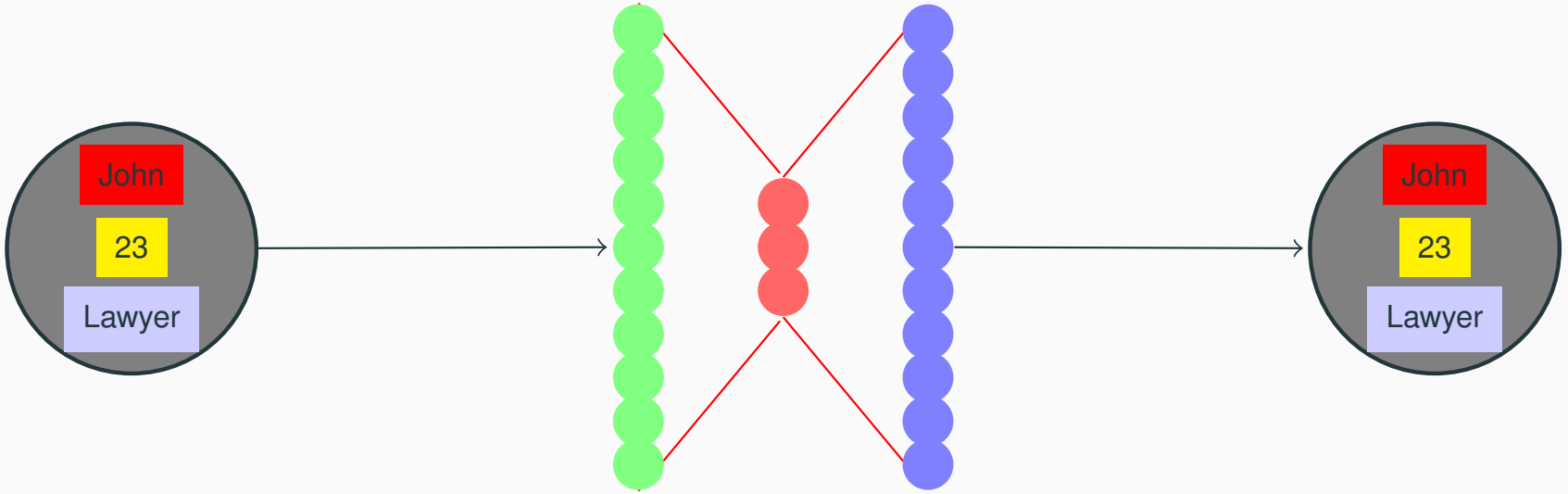
- Encoders, decoders, autoencoders
- Graph convolutional mechanism
- **Combine these to match the data being modeled**

Starcoder

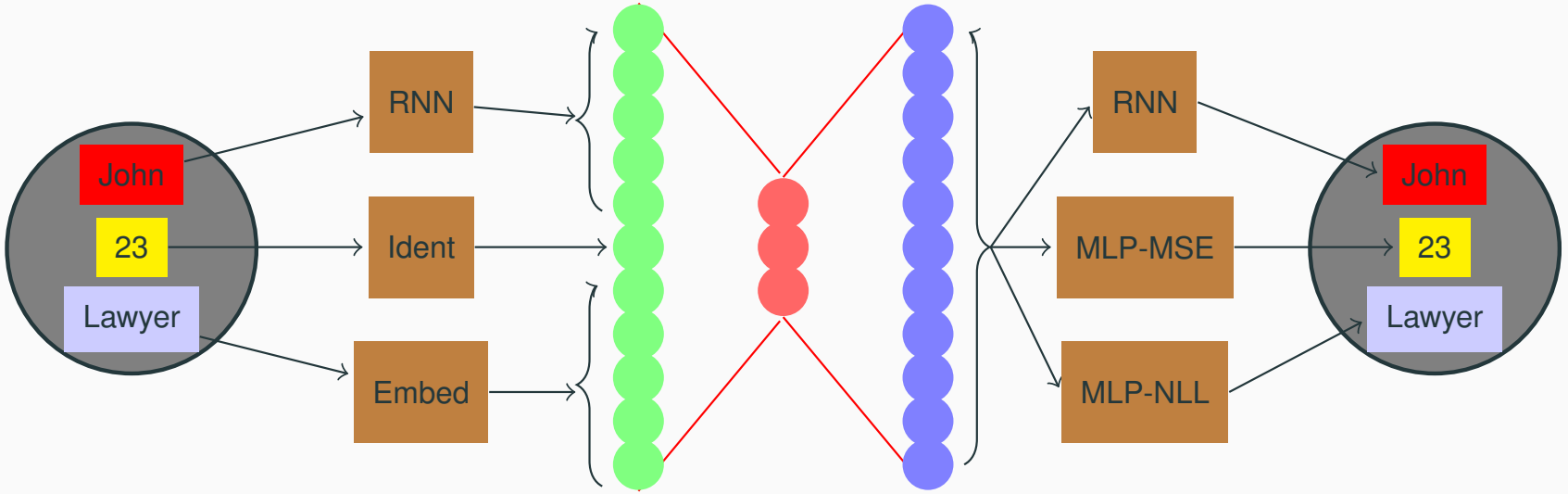




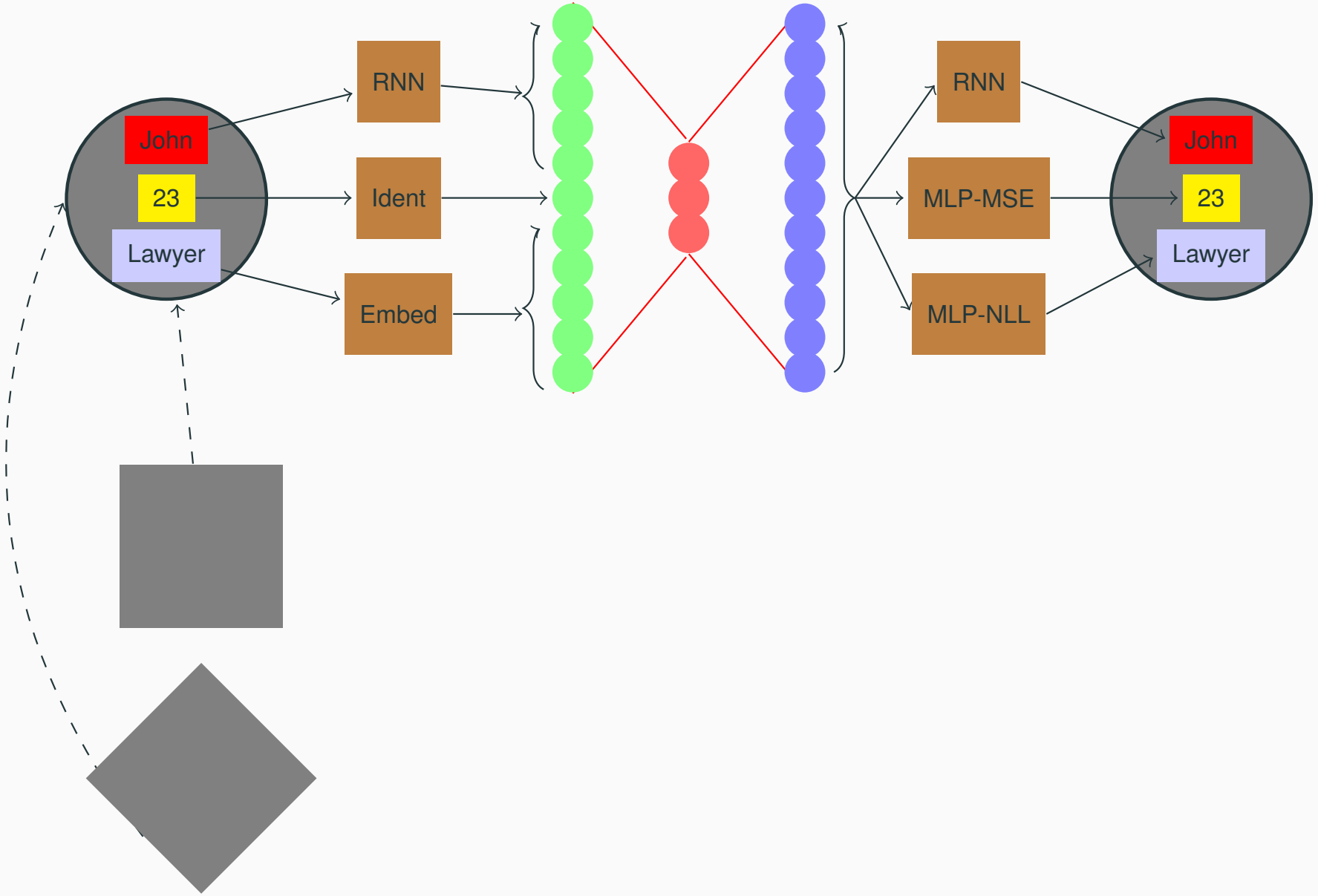
Starcoder



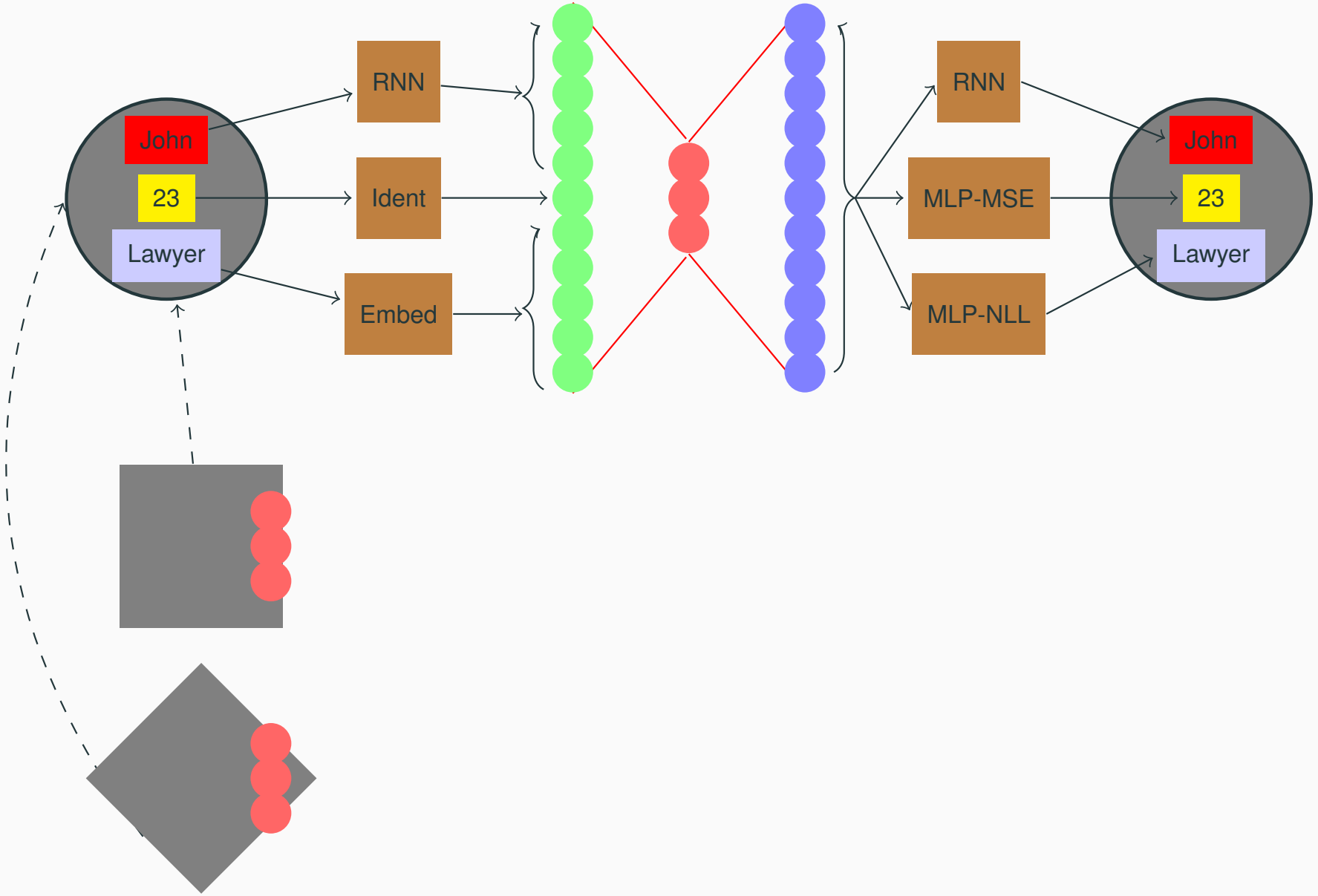
Starcoder


John
23
Lawyer
RNN
Ident
Embed
RNN
MLP-MSE
MLP-NLL
John
23
Lawyer

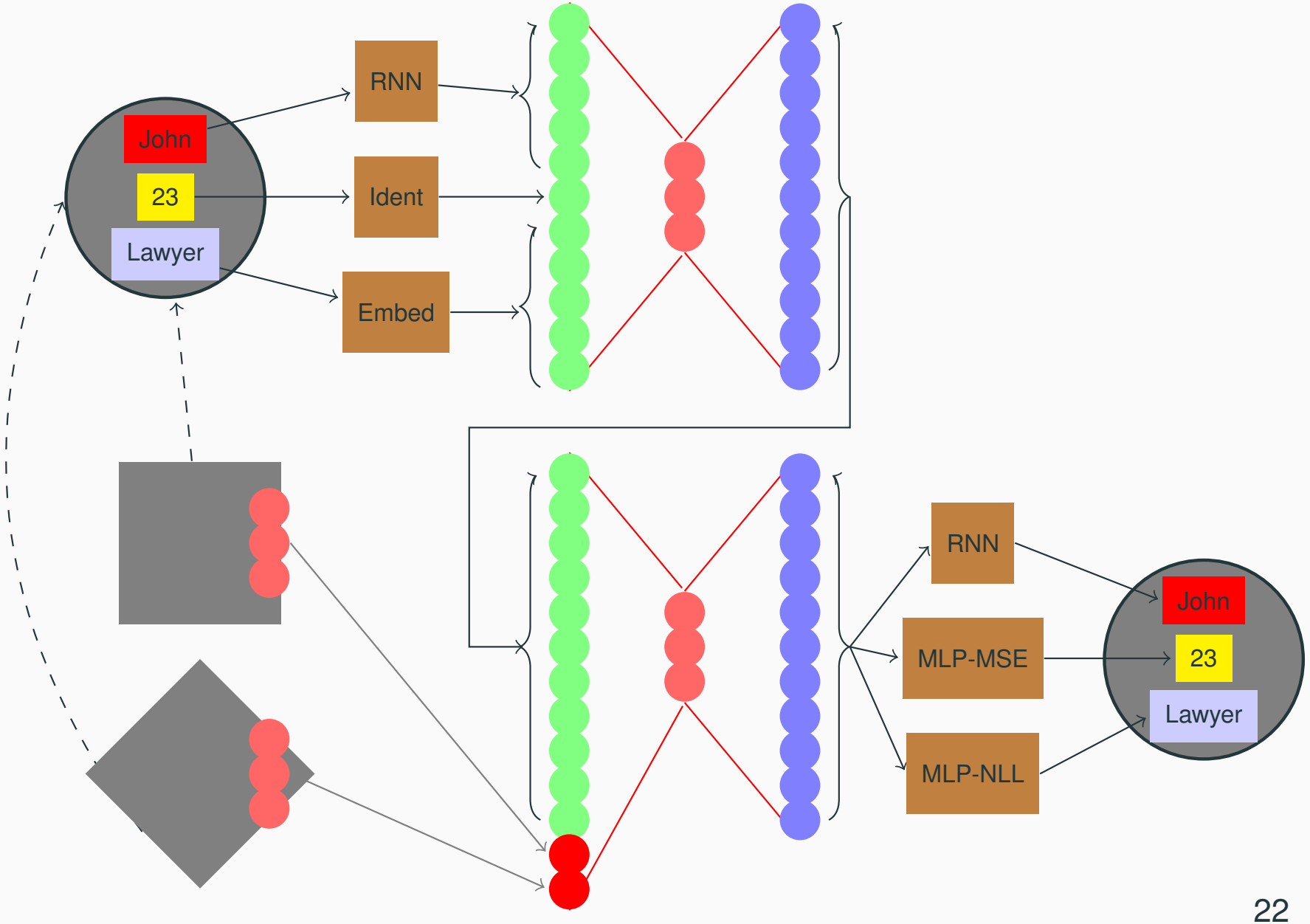
Starcoder


John
23
Lawyer
RNN
Ident
Embed
RNN
MLP-MSE
MLP-NLL
John
23
Lawyer

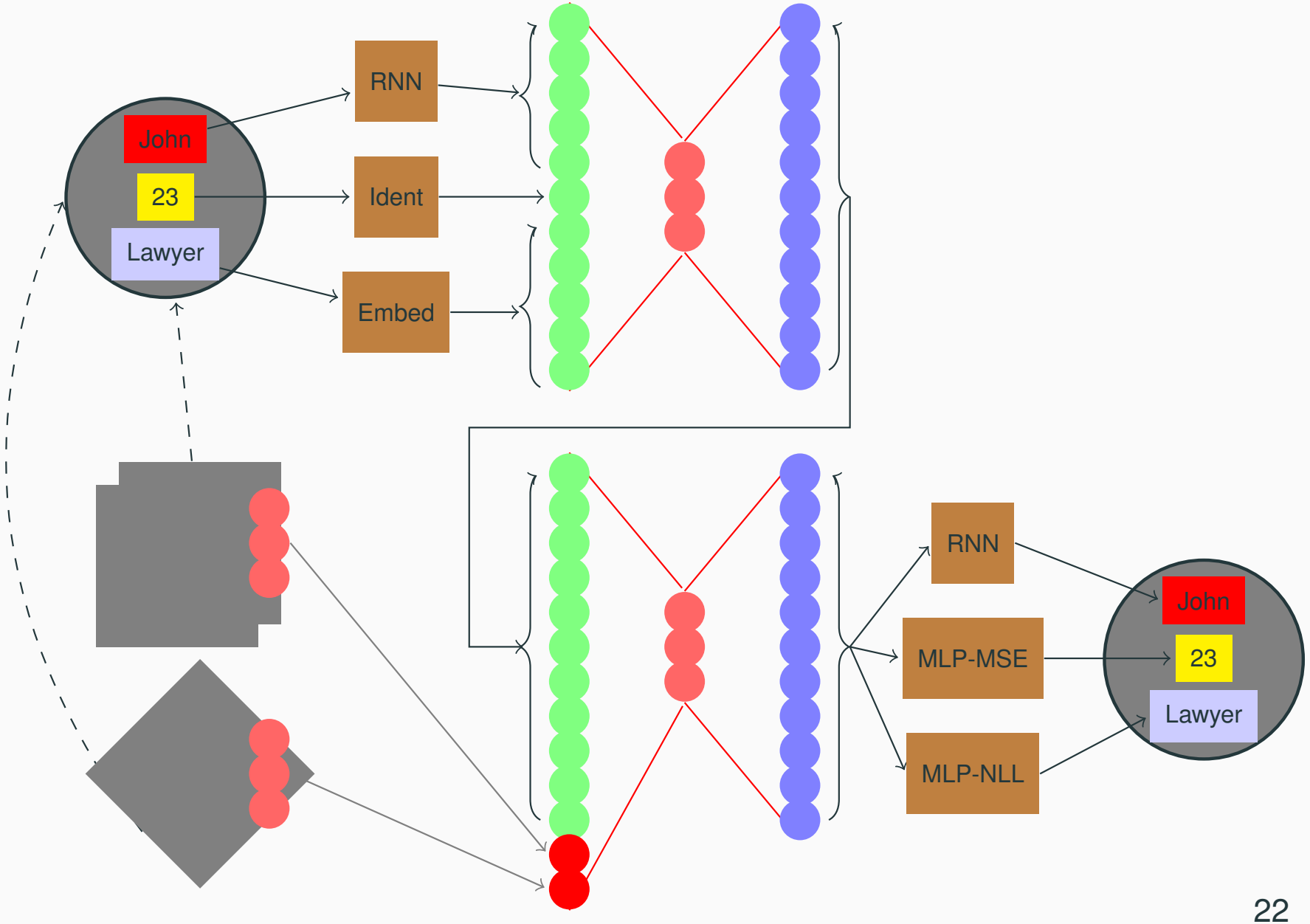
Starcoder


John
23
Lawyer
RNN
Ident
Embed
RNN
MLP-MSE
MLP-NLL
John
23
Lawyer

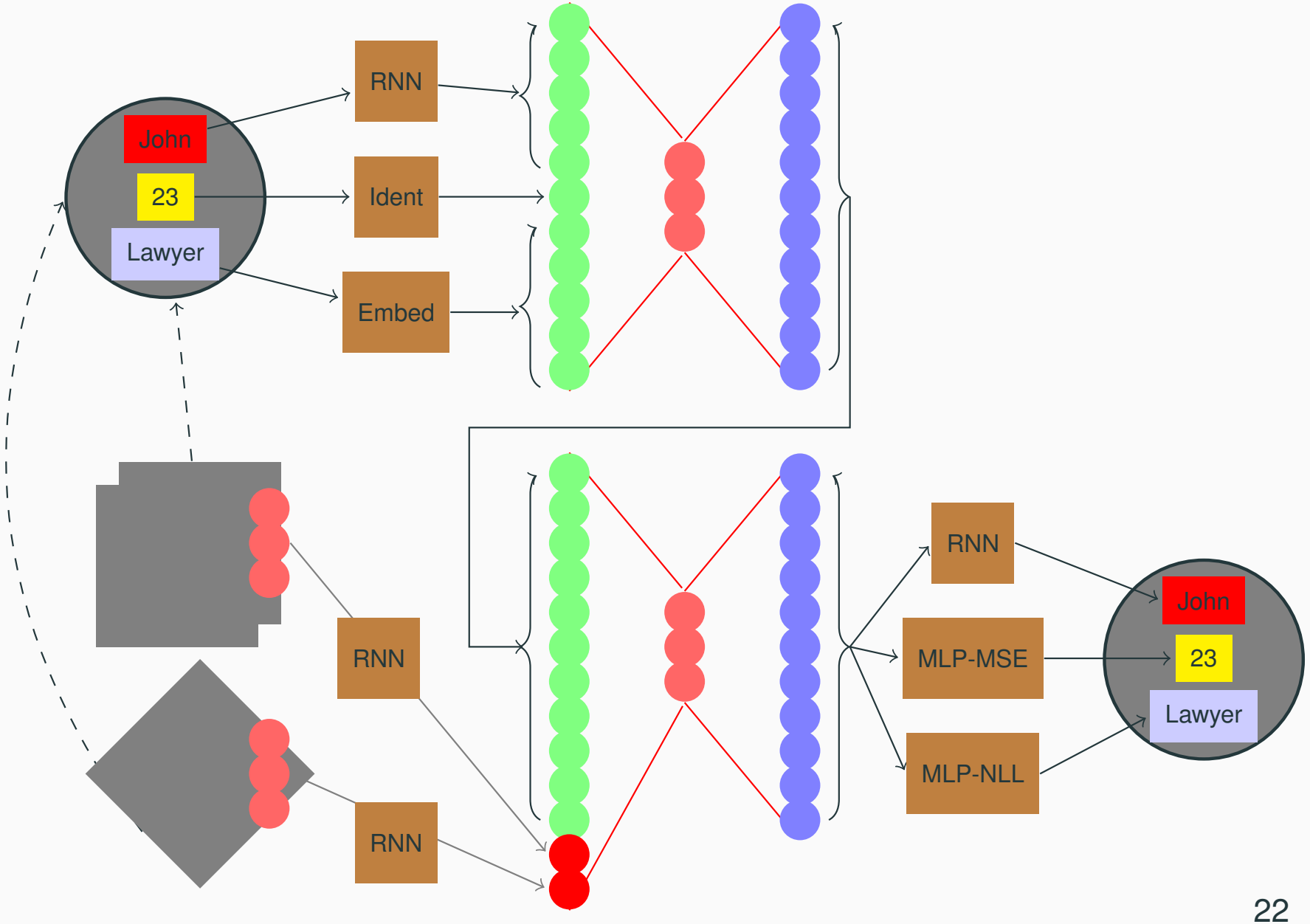
Starcoder


RNN
John
23
Ident
Lawyer
Embed
RNN
MLP-MSE
MLP-NLL
John
23
Lawyer

Starcoder


RNN
John
23
Ident
Lawyer
Embed
RNN
MLP-MSE
MLP-NLL
John
23
Lawyer

Starcoder


John
23
Lawyer
RNN
Ident
Embed
RNN
RNN
RNN
MLP-MSE
MLP-NLL
John
23
Lawyer

Training is a bit involved . . .

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

How can we use a trained model?

Bottleneck (embedding) similarity

- Compute distance between two entities
- Find flat or hierarchical clusters of entities

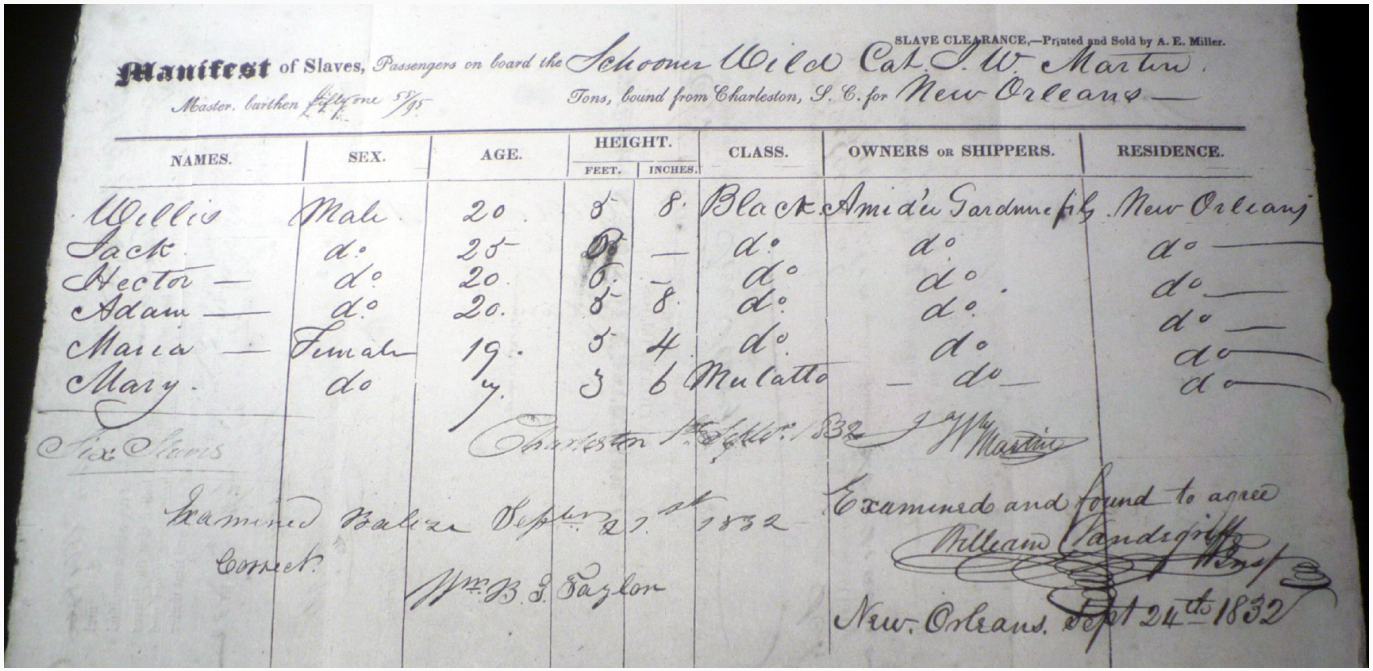
Field generation

- Generate likely value of missing field
- Detect an improbable value of a present field
- Observe response of one field to another

Example: The 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, barthen Lafayette 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	Amidee Gardner & Co	New Orleans
Jack	do	25	5	—	do	do	do
Hector	do	20	5	8	do	do	do
Adam	do	20	5	8	do	do	do
Maria	Female	19	5	4	do	do	do
Mary	do	4	3	6	Mulatto	do	do

Six Slaves

Charleston Sept. 1832 J. W. Martin

Examined Balize Sept. 27th 1832

Correct.

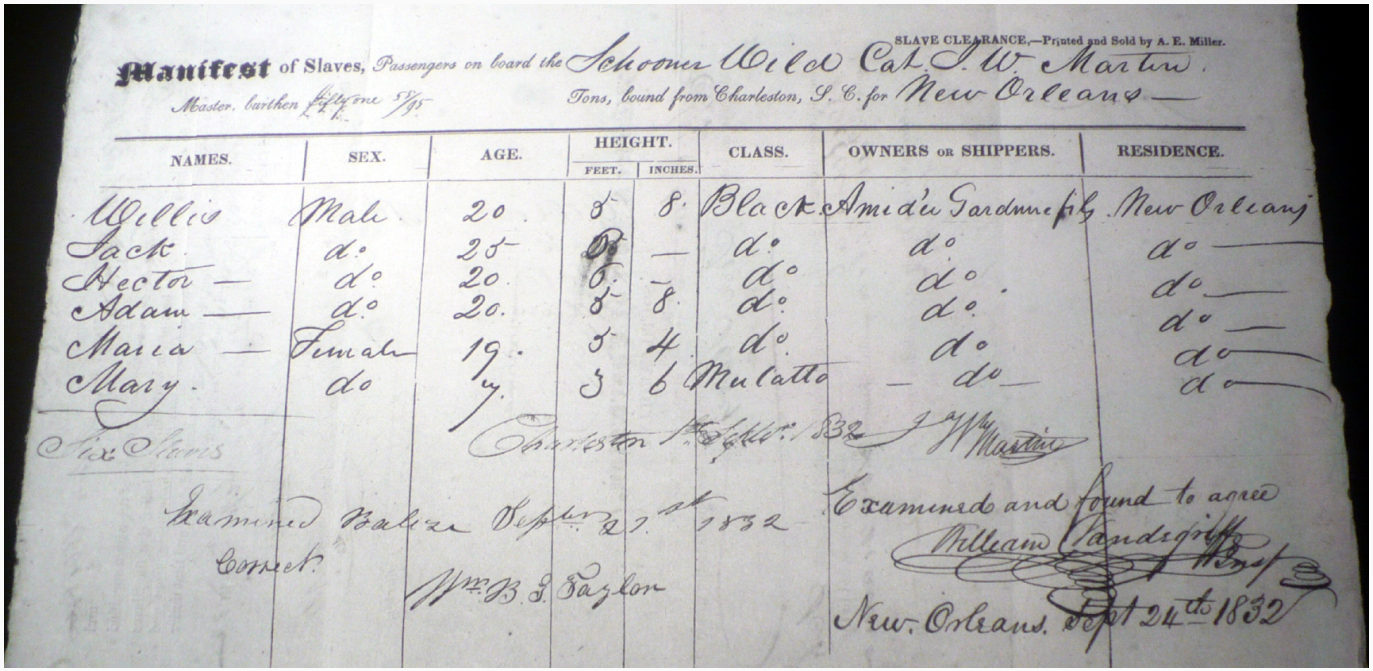
Wm B. J. Taylor

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, barthen Fifty one 51 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	Amidu Gardner & Co	New Orleans
Jack	do	25	5	—	do	do	do
Hector	do	20	5	8	do	do	do
Adam	do	20	5	8	do	do	do
Maria	Female	19	5	4	do	do	do
Mary	do	4	3	6	Mulatto	do	do

Six Slaves

Charleston Sept. 22 1832 J. W. Martin

Examined Baker Sept. 27 1832

Examined and found to agree

William Vandegrift

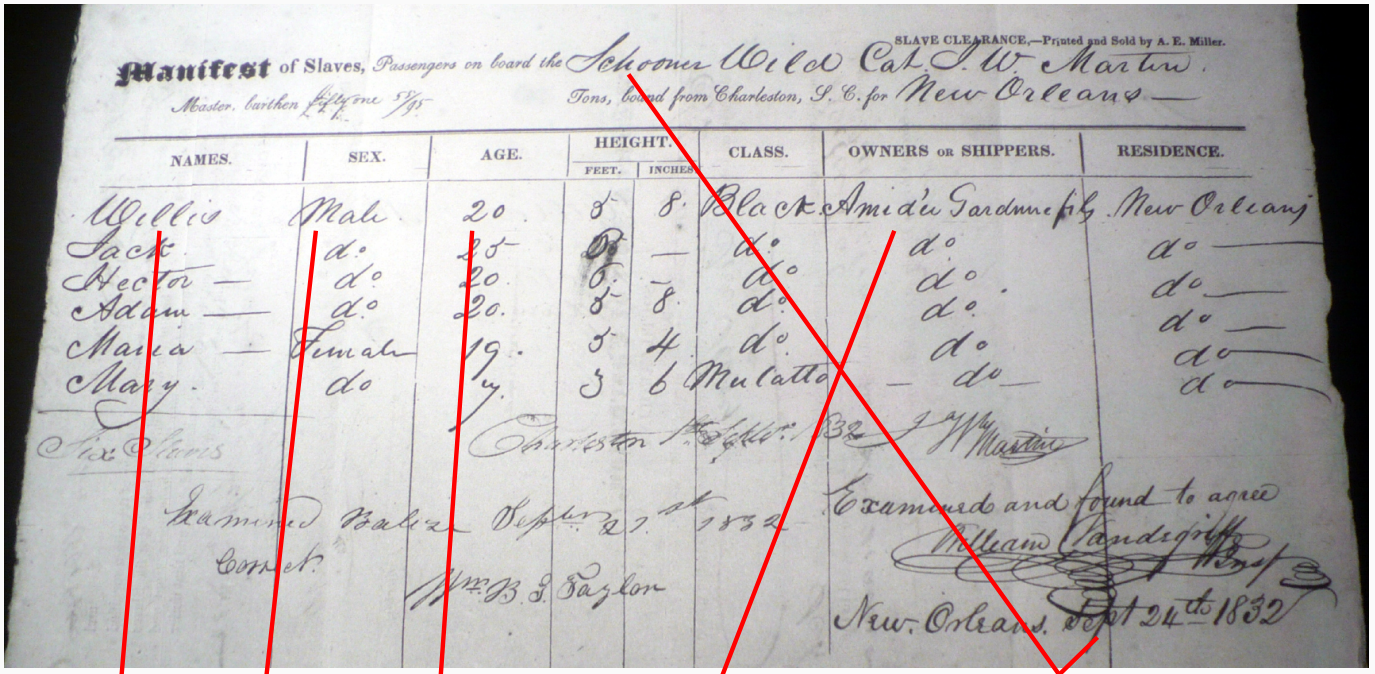
Correct.

Wm B. J. Taylor

New Orleans Sept 24 1832

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

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, barthen Fifty one 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	Amidu Gardner & Co	New Orleans
Jack	do	25	5	—	do	do	do
Hector	do	20	5	8	do	do	do
Adam	do	20	5	8	do	do	do
Maria	Female	19	5	4	do	do	do
Mary	do	4	3	6	Mulatto	do	do

Six Slaves

Charleston S.C. Oct. 1832 J. W. Martin

Examined Baker Septr 27th 1832

Examined and found to agree

William Vandegrift

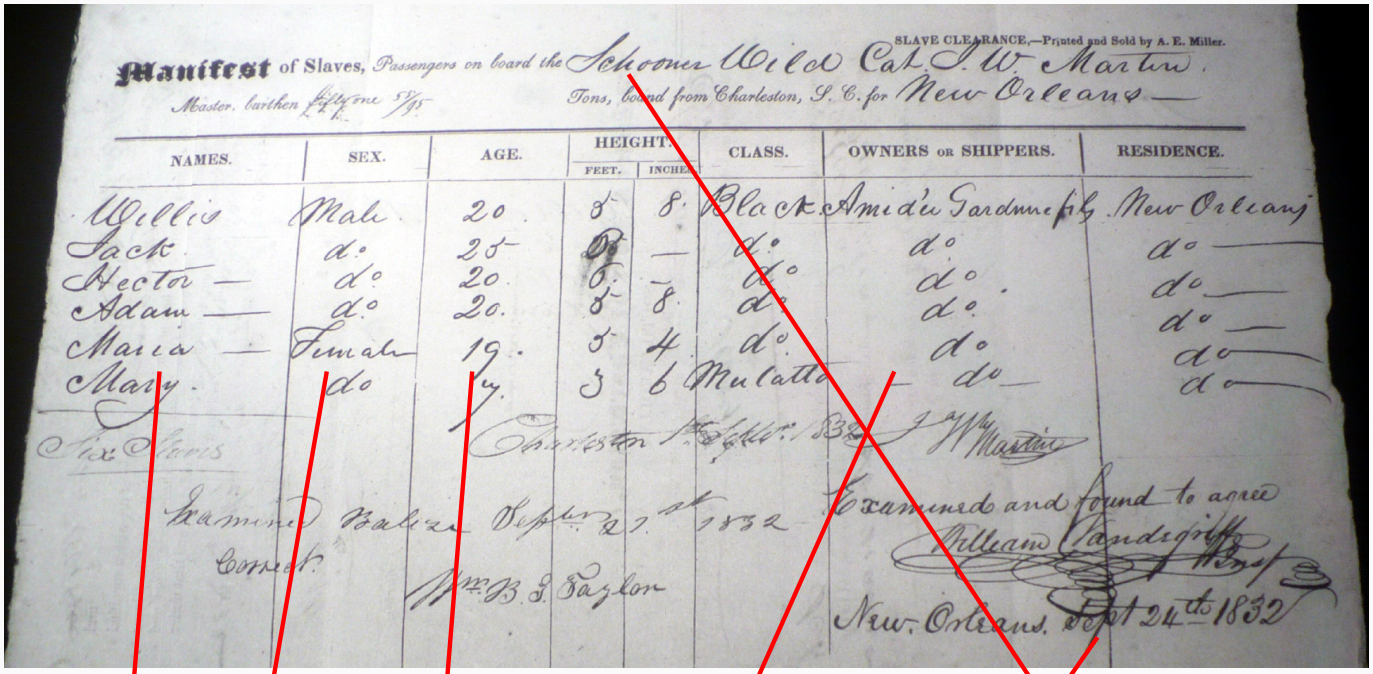
Consert

Wm B. J. Taylor

New Orleans Sept 24th 1832

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

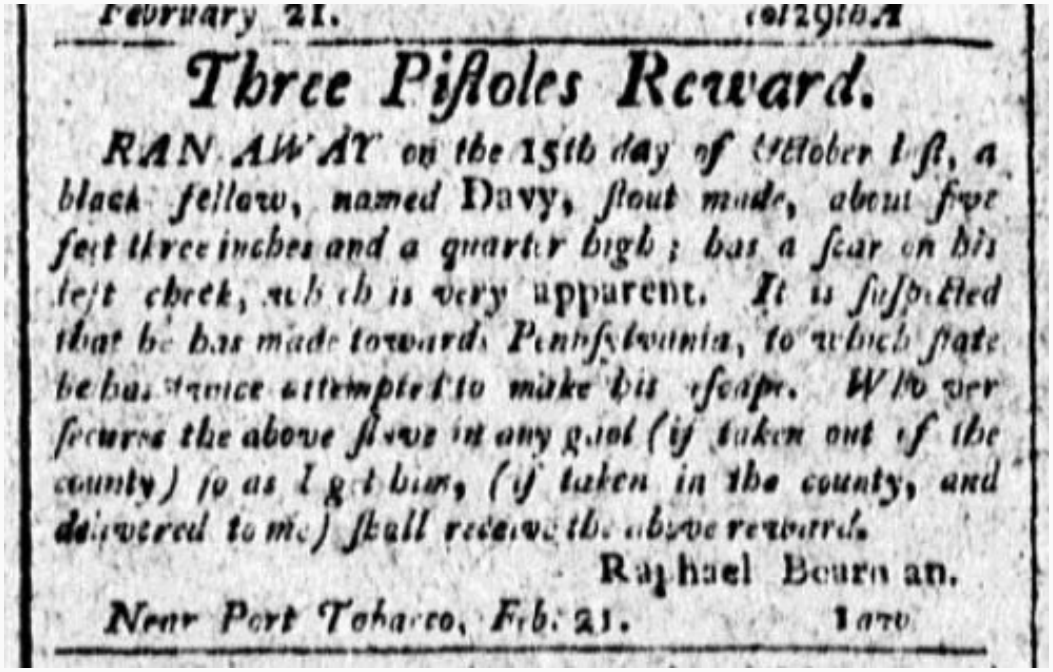
Shipping manifests



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. [illegible]

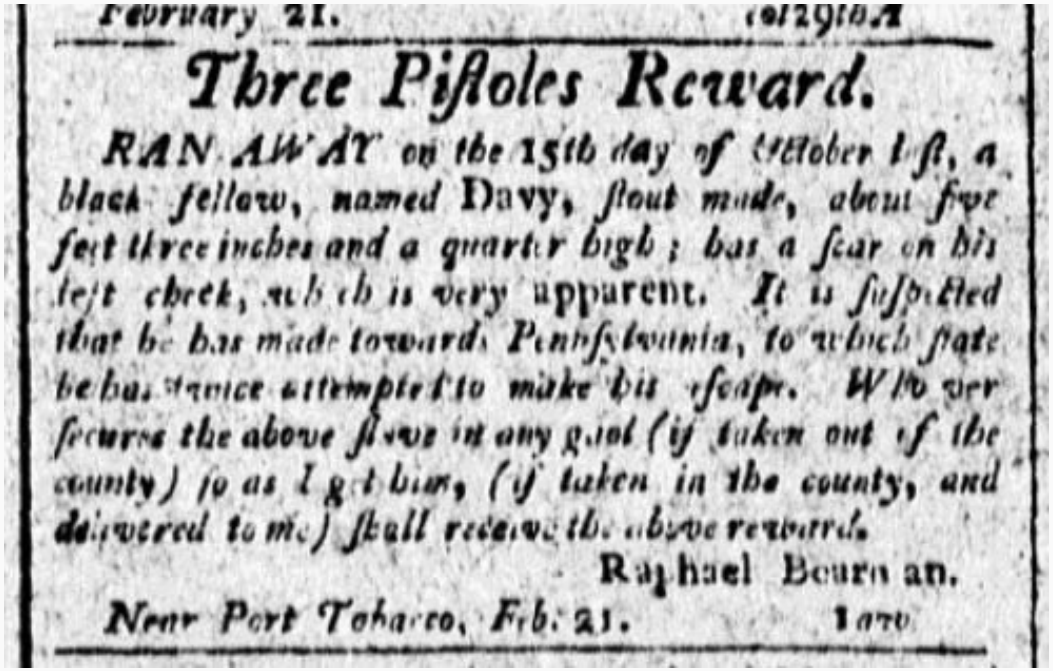
Three Piſtoles Reward.

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 goal (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 Bourn an.

Near Port Tobacco, Feb. 21. 1770

Fugitive notices



February 21. 1795.

Three Piſtoles Reward.

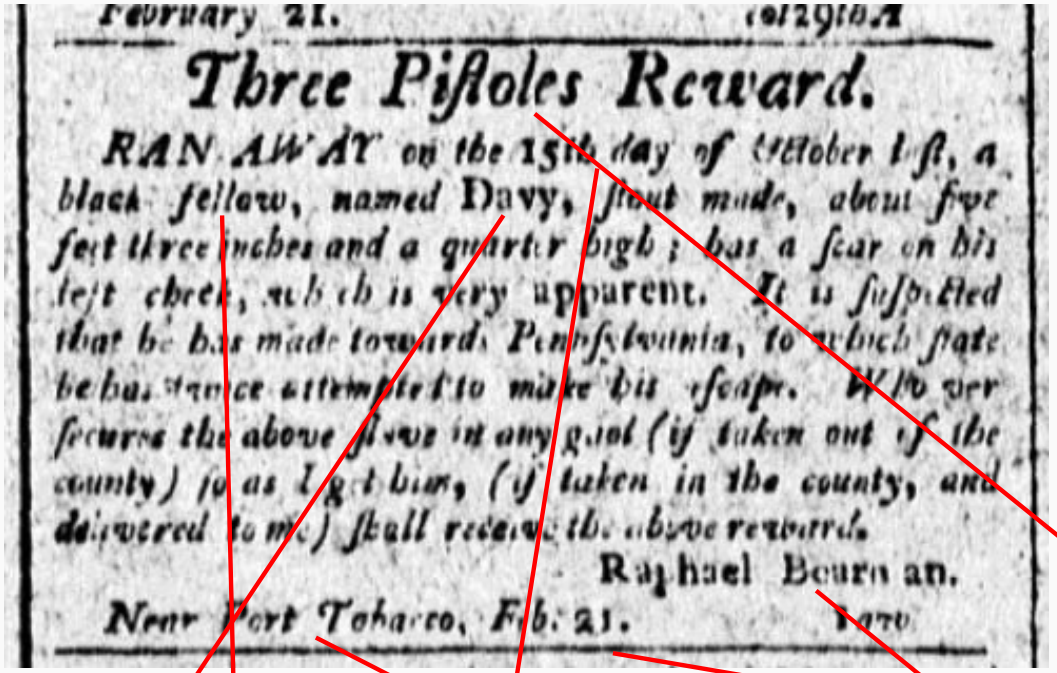
RAN AWAY on the 15th day of October laſt, a black fellow, named Davy, ſtout made, about five feet three inches and a quarter high; has a ſcar on his left cheek, which is very apparent. It is ſuſpected that he has made towards Pennſylvania, to which ſtate he has twice attempted to make his eſcape. Whoever ſecures the above ſlave in any goal (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 Bourn an.

Near Port Tobacco, Feb. 21. 1795

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 once 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. 1796

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
- 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”

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”
- Missing and underspecified entities
  - *Majority* of slaves have no last name
  - Can’t tell if two “John”s 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
- Map out trade “ecosystems” of sellers, shippers, owners, etc
- Reconstruct slave families when there are no last names
- Determine what drove valuation in transactions and rewards

Entities, field types, and relations

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
...	...

Entities, field types, and relations

Numeric

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

Categorical

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

Text

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
...	...

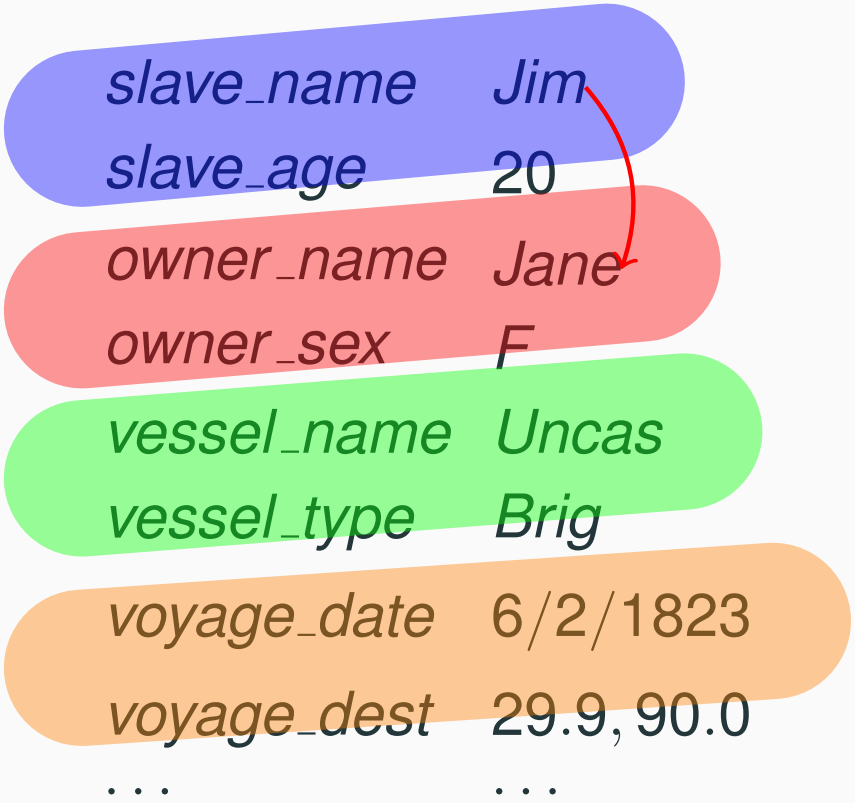
Entities, field types, and relations

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
...	...

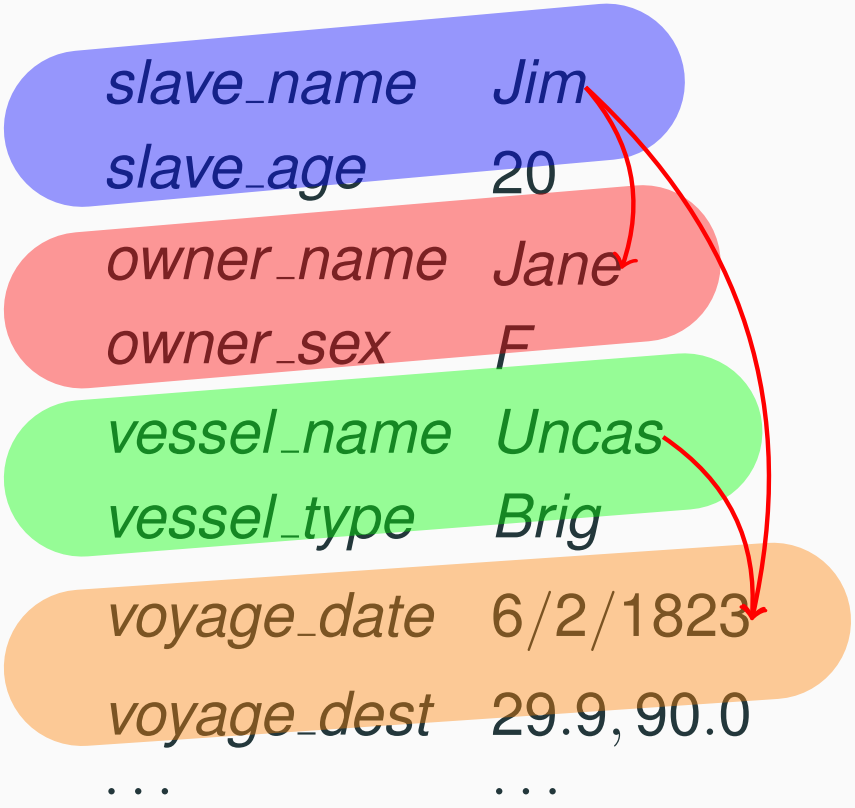
Entities, field types, and relations

Slave-to-owner



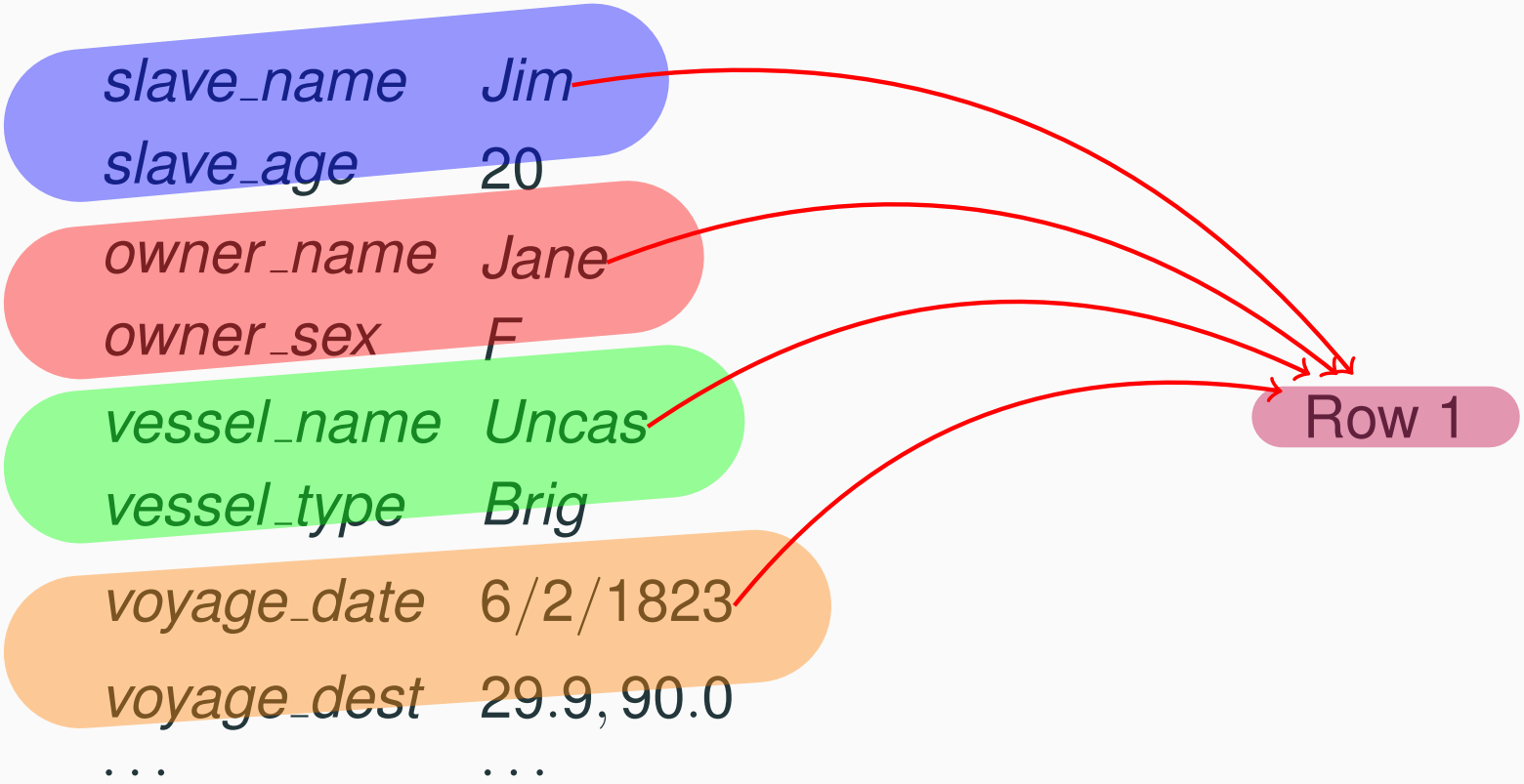
Entities, field types, and relations

Vessel-to-voyage, slave-to-voyage

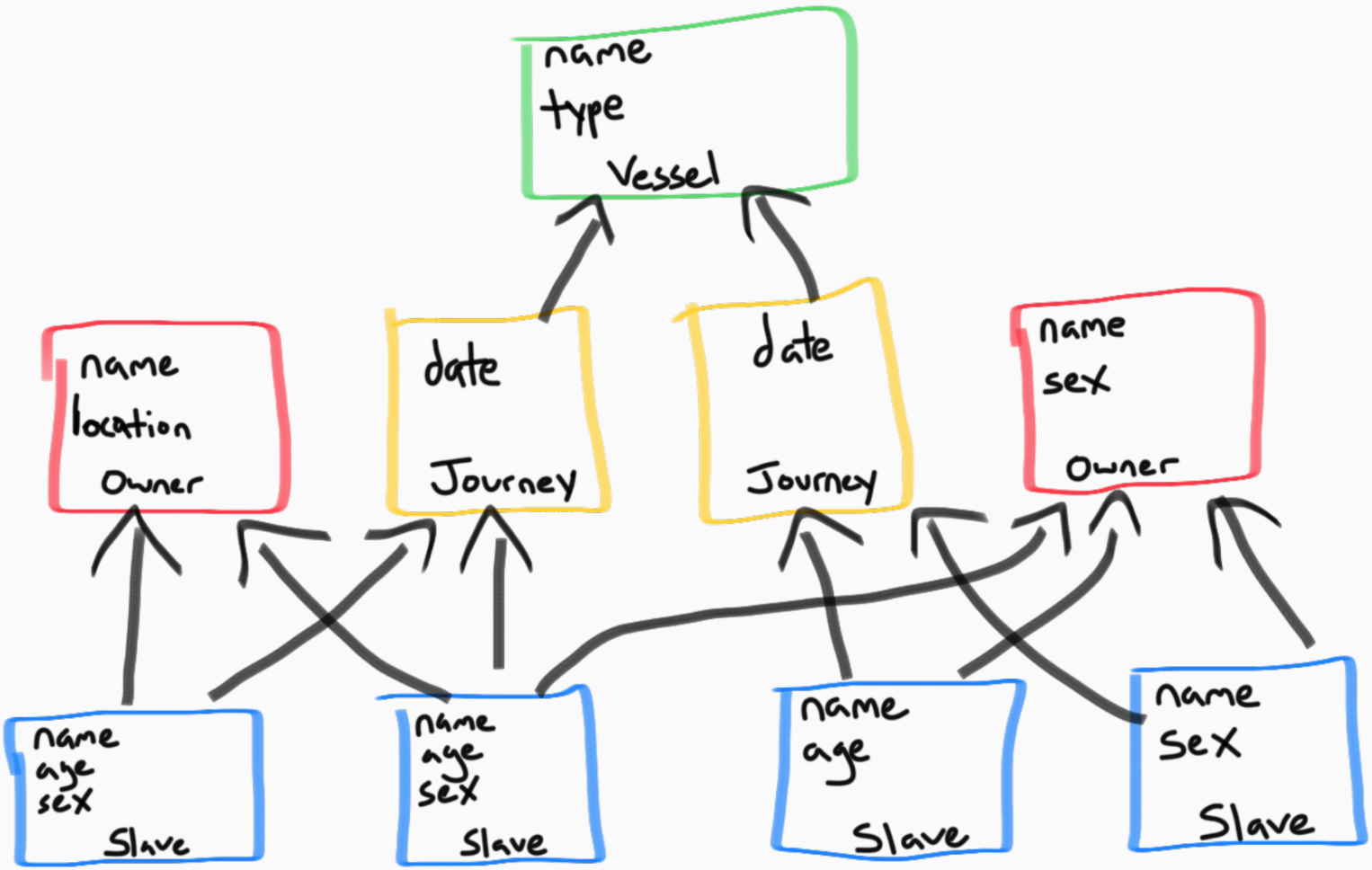


Entities, field types, and relations

Fewer assumptions



One graph


name
type
Vessel
name
location
Owner
date
Journey
date
Journey
name
sex
Owner
name
age
sex
Slave
name
age
sex
Slave
name
age
Slave
name
sex
Slave

Train Starcoder. . .

Looking at the most-similar pairs by entity-type, some trends emerge:

Mistranscriptions

Baltiomre	$\Longleftrightarrow$	Baltimore
William Wiliams	$\Longleftrightarrow$	William Williams

Semantically-equivalent variants

George Y. Kelso	$\Leftrightarrow$	Kelso & Ferguson
New Orleans	$\Leftrightarrow$	Louisiana

Same slave transported multiple times[1]

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

[1] Many more instances of e.g. “John” following this superficial pattern

Ongoing work

Benefits of defining these APIs

- Scaling: easy to add new studies
- Research: new architectures to implement and import
- Visualization: auto-generate figures based on schemas
- Interaction: well-defined interface to explore ML output
- **Remains grounded in the particular scholarly domain**

Entertaining America: pre-Vaudeville theater

Data

- Derived from local newspaper ads
- Actors, troupes, towns, performances
- Geo-coded town maps

Questions/Goals

- Location of performances (within a town)
- How performances are chosen and described

Chaucer's (and Gower's) Metrical Voice

Data

- Poetry (e.g. Canterbury Tales) with (partial) stress annotations
- Entities are lines, stanzas, chapters, tales

Questions/Goals

- Train Starcoder as words2stresses, lines2rhymes
- Stresses as features for authorship
- *Does the traditional scansion obscure interesting properties of the text?*

Financial and demographic history

Studies

- Tax records of Medieval Paris
- Employment in early modern London
- Economy of the Caribbean Colonies

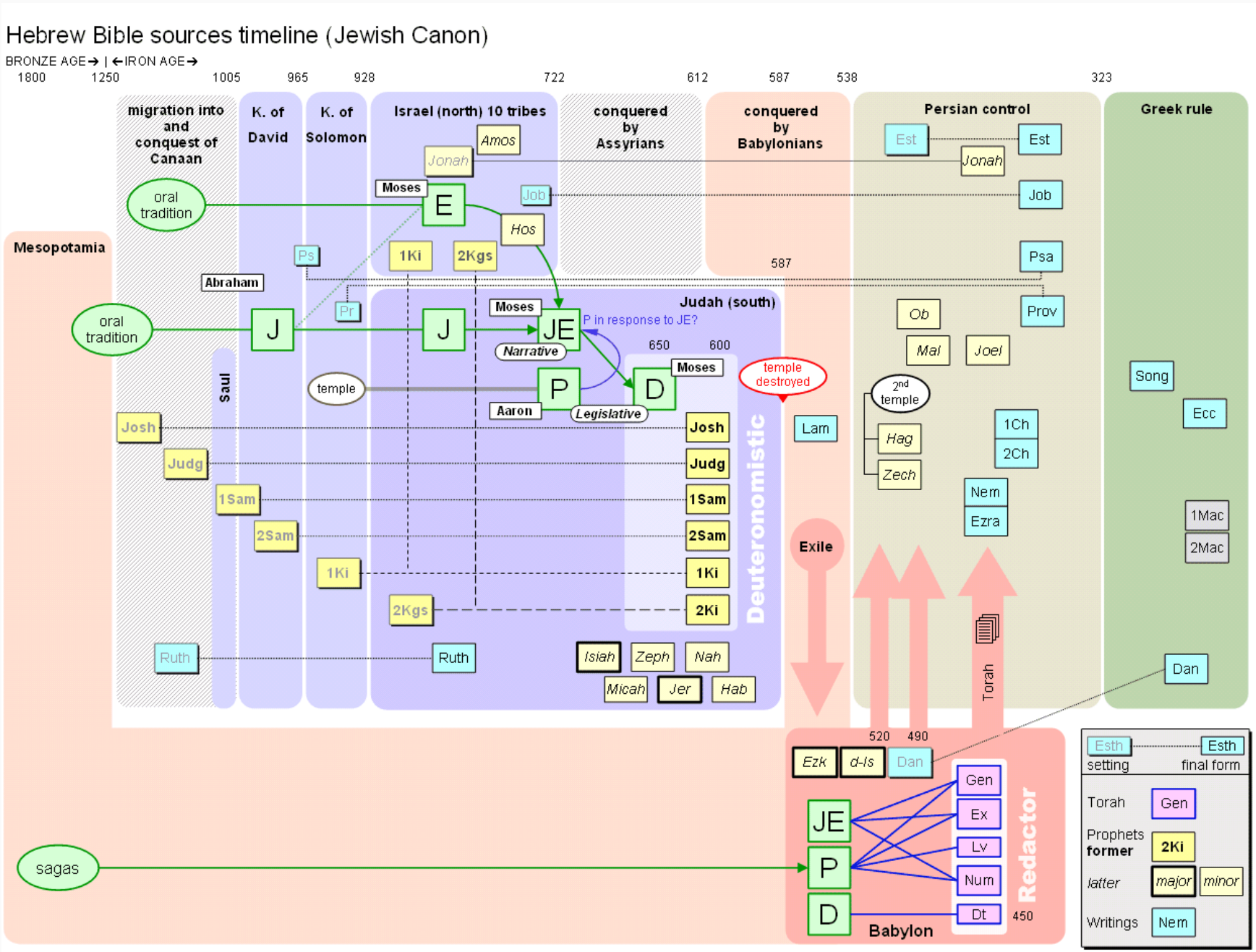
Questions/Goals

- Impetus to treat dates and coordinates as first-class field types
- Differential look at economic policies

Cuneiform in the Ancient Near East

- Traditionally, inscriptions are grouped by physical object-type
- We also have geographic location, transcription, ruler …
- What is the relationship between these properties?
- Is the traditional view a useful distinction, and how?

Compositional forces on foundational cultural texts



Some familiar NLP tasks

Language ID

- Communication network
- Users, messages, and languages
- Capture how multi-lingual users tweet and follow

Sentiment analysis

- Parse trees
- Words (leaves), nodes (constituents), polarities
- Capture how sentiment composes

Also, compare with GraphSAGE, KB embedding techniques, etc