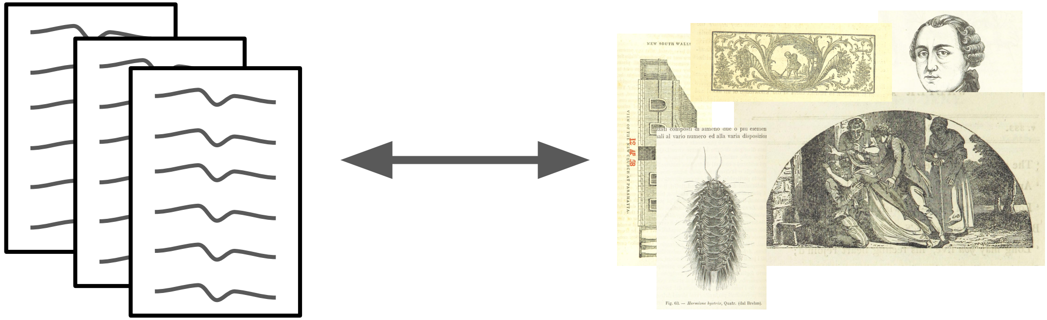
Grounding Images from a Digital Library in their Textual Contexts

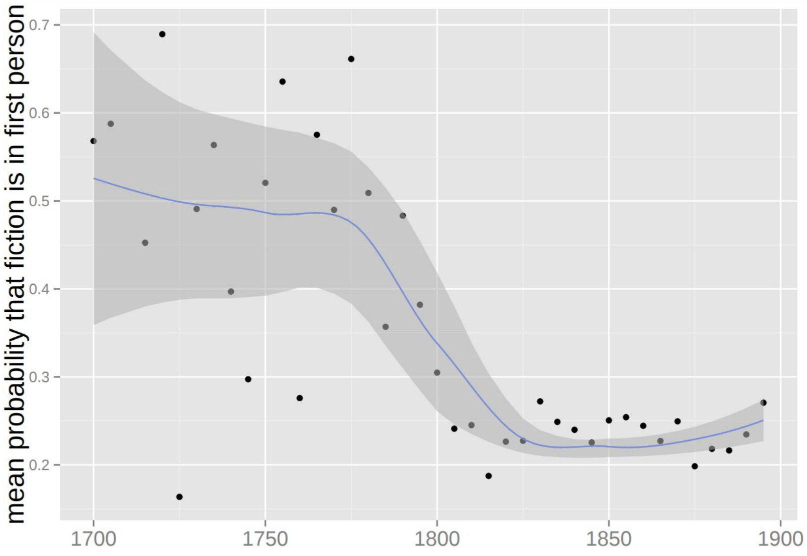


Jack Hessel
joint work with David Mimno and Lillian Lee

- Similarities between text data and image data
- Why should we want to model text and images jointly?
- Computer vision and "why digital libraries?"
- The dataset/experiments
- Are concrete things easier to learn?

- **Similarities between text data and image data**
- Why should we want to model text and images jointly?
- Computer vision and "why digital libraries?"
- The dataset/experiments
- Are concrete things easier to learn?

Using Text Data in the Digital Humanities

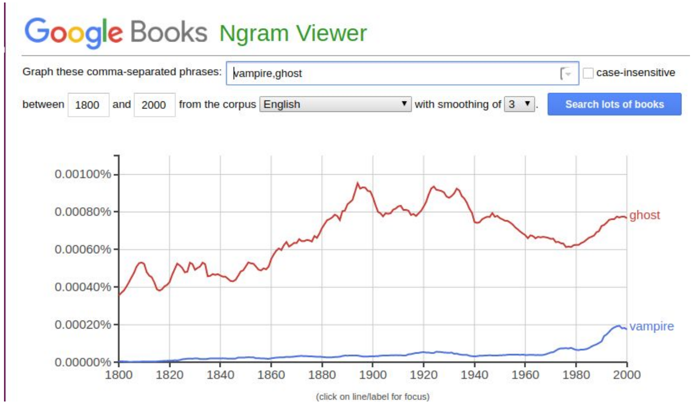


[Underwood et al. 2013]

There exist lots of text tools for Digital Humanists

MALLET

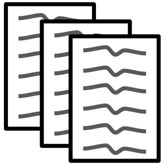
machine learning for language toolkit



Textalyser Results

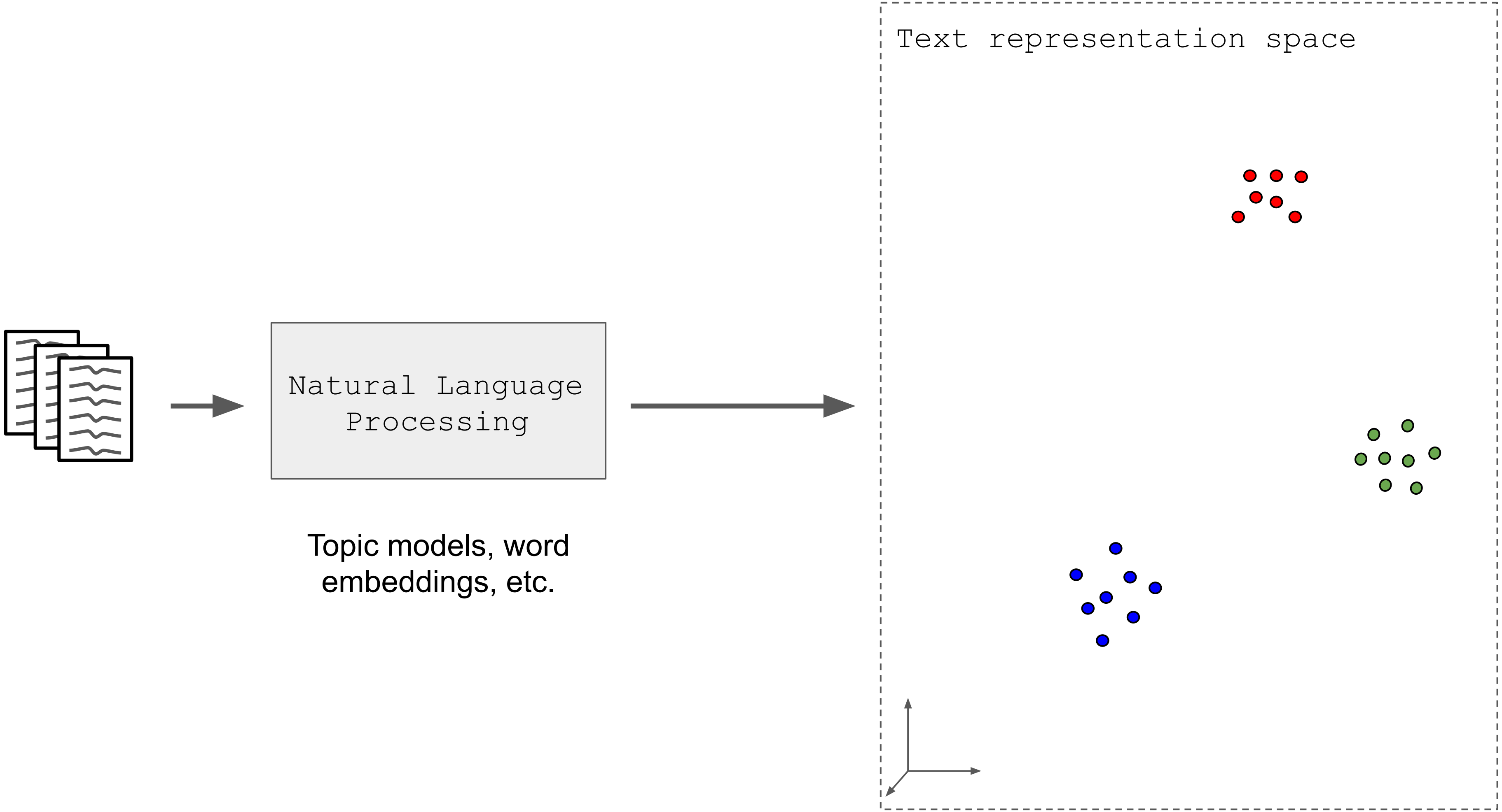
The complete results, incuding compexity factor, and other features

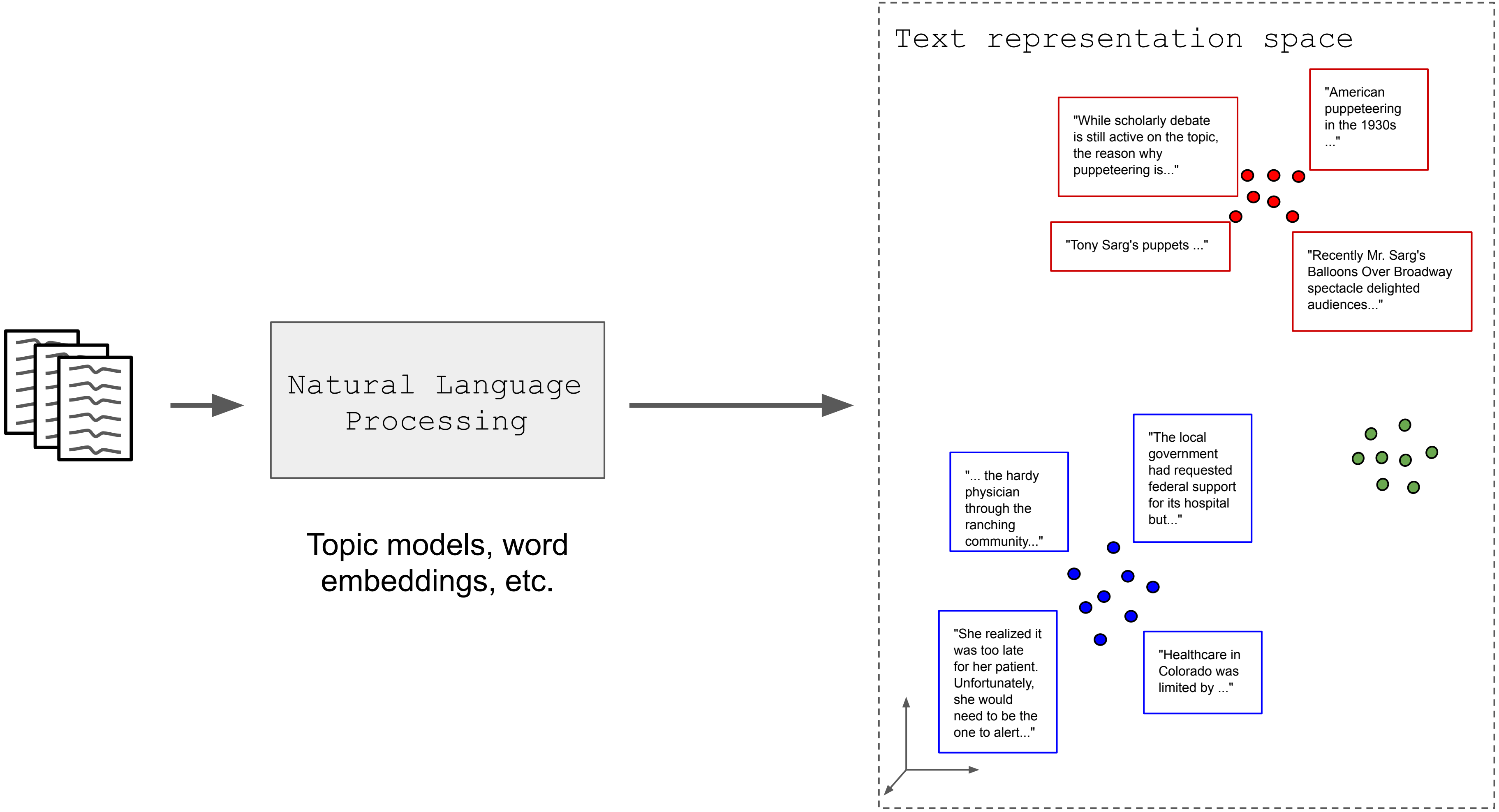
Total word count :	42
Number of different words :	37
Complexity factor (Lexical Density) :	88.1%
Readability (Gunning-Fog Index) : *(6-easy 20-hard)*	9.5
Total number of characters :	449
Number of characters without spaces :	278
Average Syllables per Word :	1.71




Natural Language Processing

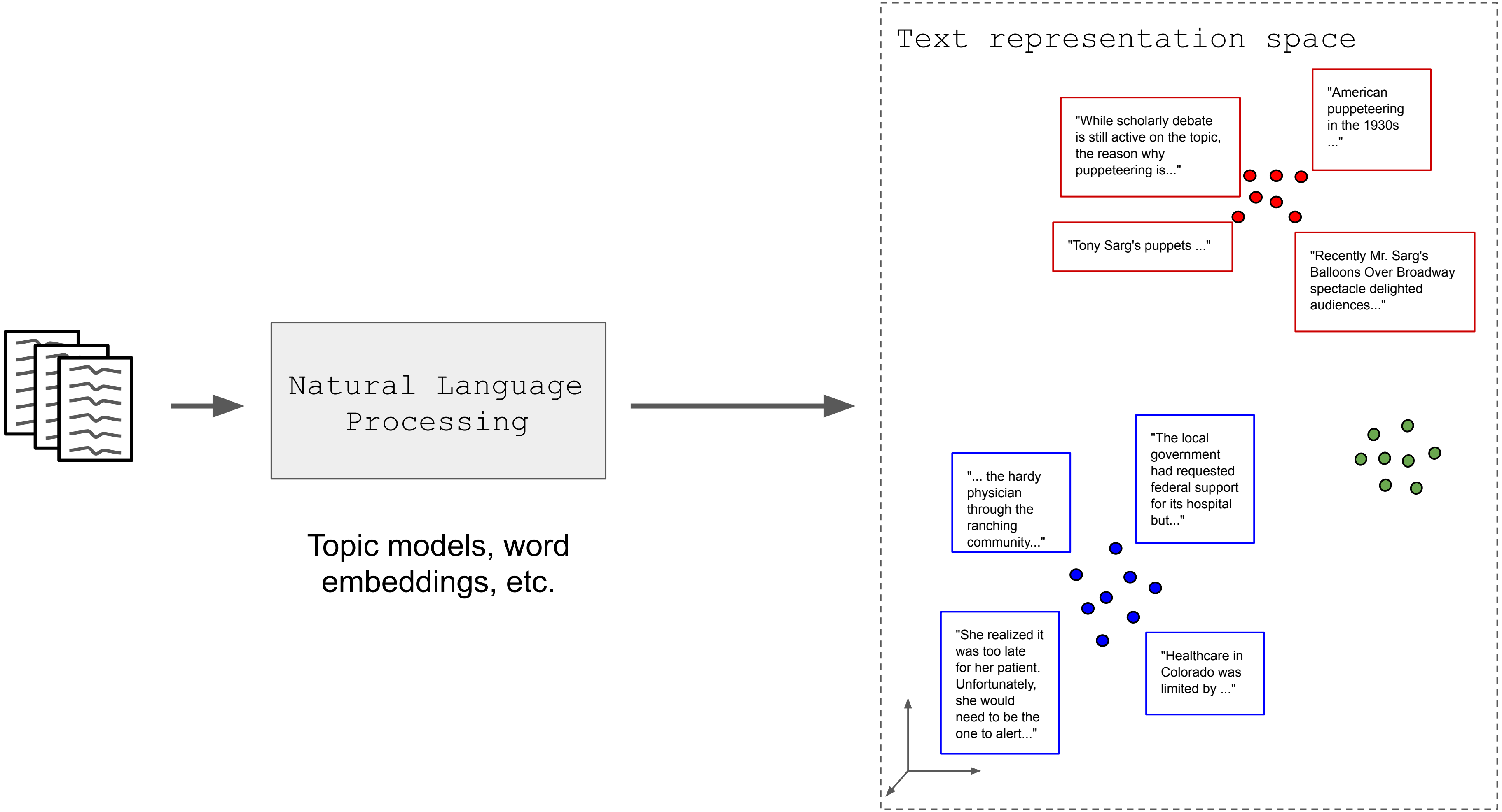
Topic models, word embeddings, etc.

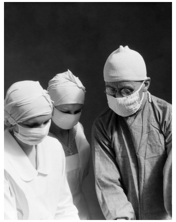

Natural Language Processing
Topic models, word embeddings, etc.
Text representation space

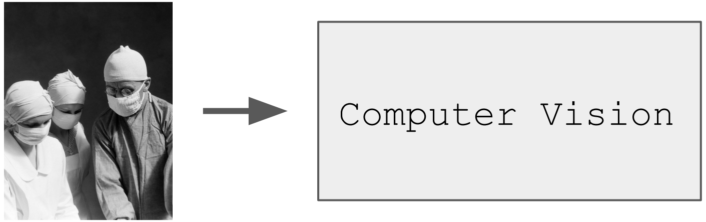


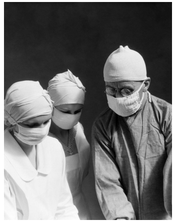
Natural Language Processing
Topic models, word embeddings, etc.
Text representation space
"While scholarly debate is still active on the topic, the reason why puppeteering is..."
"American puppeteering in the 1930s ..."
"Tony Sarg's puppets ..."
"Recently Mr. Sarg's Balloons Over Broadway spectacle delighted audiences..."
"... the hardy physician through the ranching community..."
"The local government had requested federal support for its hospital but..."
"She realized it was too late for her patient. Unfortunately, she would need to be the one to alert..."
"Healthcare in Colorado was limited by ..."

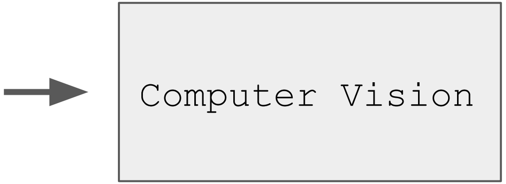
[Hamilton et al. 2016]


Natural Language Processing
Topic models, word embeddings, etc.
Text representation space
"While scholarly debate is still active on the topic, the reason why puppeteering is..."
"American puppeteering in the 1930s ..."
"Tony Sarg's puppets ..."
"Recently Mr. Sarg's Balloons Over Broadway spectacle delighted audiences..."
"... the hardy physician through the ranching community..."
"The local government had requested federal support for its hospital but..."
"She realized it was too late for her patient. Unfortunately, she would need to be the one to alert..."
"Healthcare in Colorado was limited by ..."

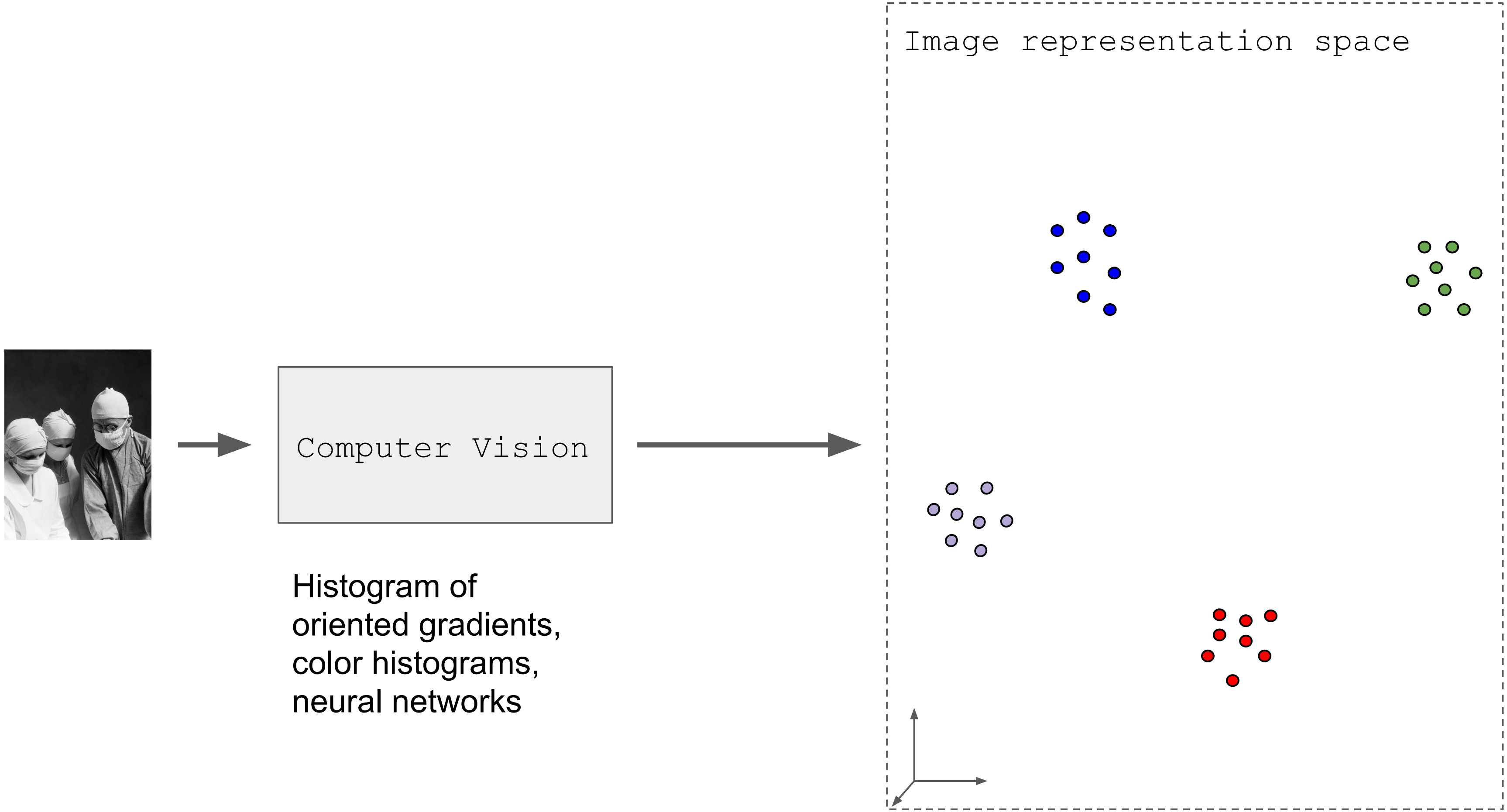


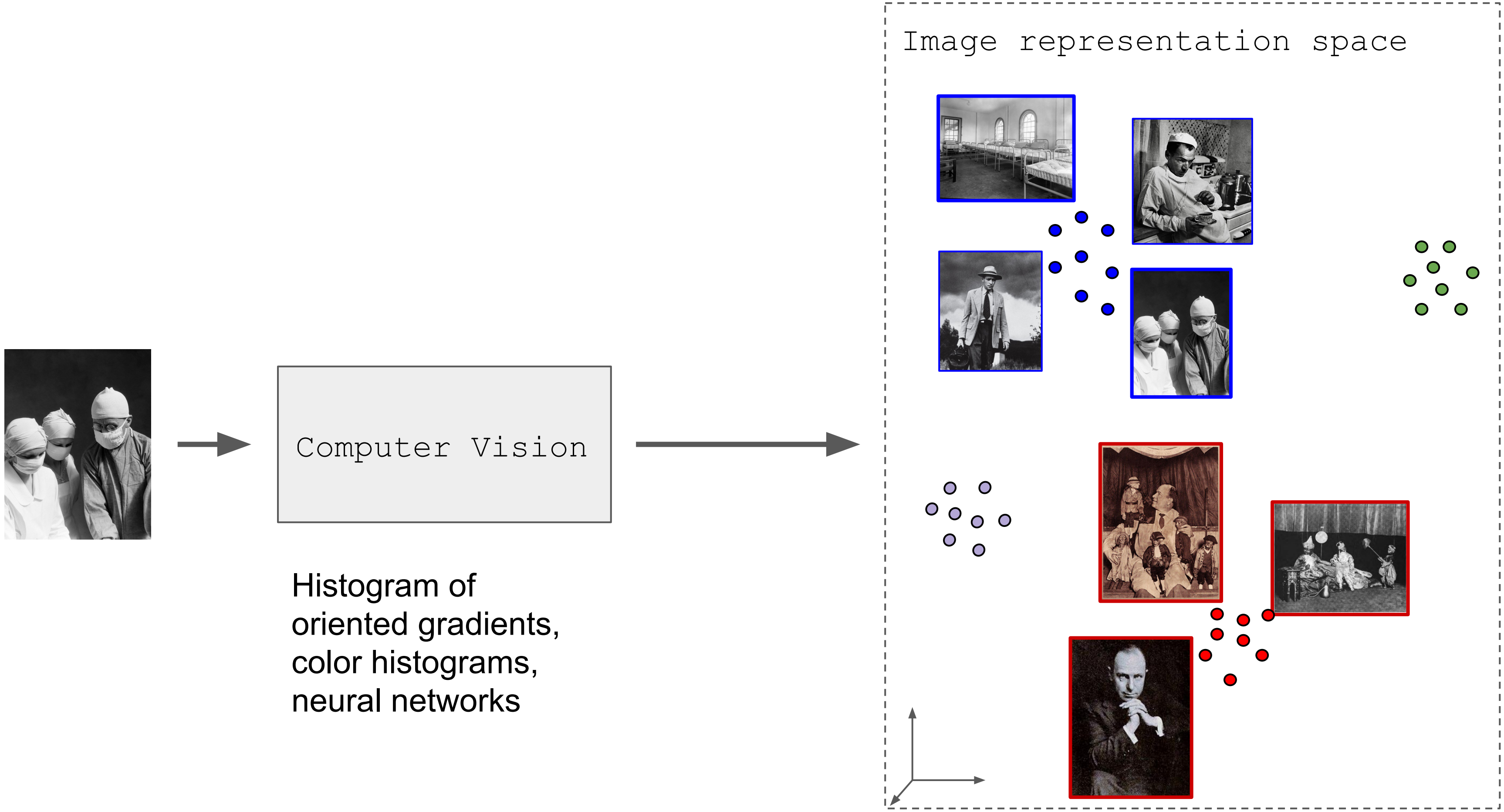

Computer Vision





Histogram of
oriented gradients,
color histograms,
neural networks


Computer Vision
Histogram of oriented gradients, color histograms, neural networks
Image representation space


Computer Vision
Histogram of oriented gradients, color histograms, neural networks
Image representation space

Levels of abstraction in computer vision

Raw pixels

Concepts

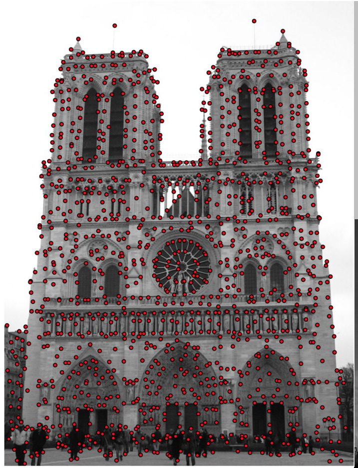
Higher level
"understanding"

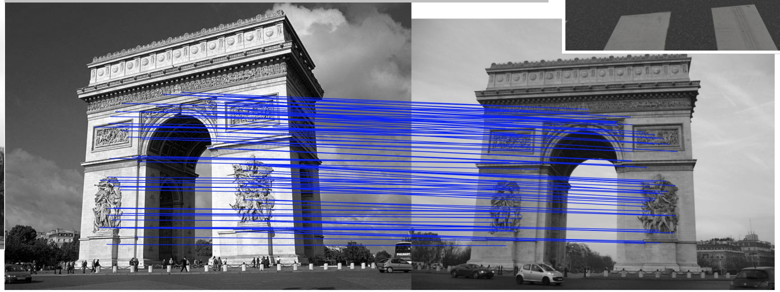
Levels of abstraction in computer vision

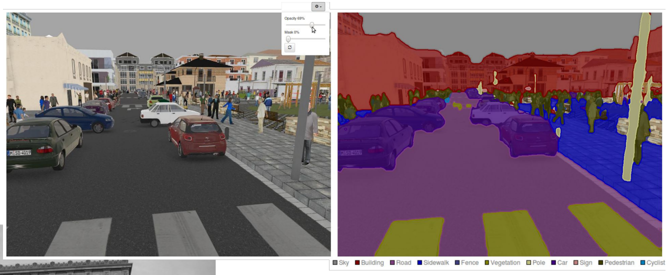
Raw pixels

Concepts

Higher level "understanding"





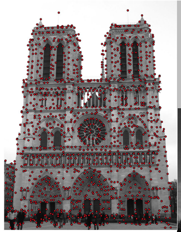


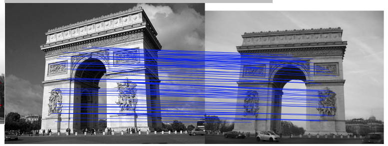
Levels of abstraction in computer vision

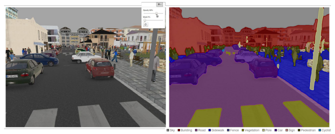
Raw pixels

Concepts

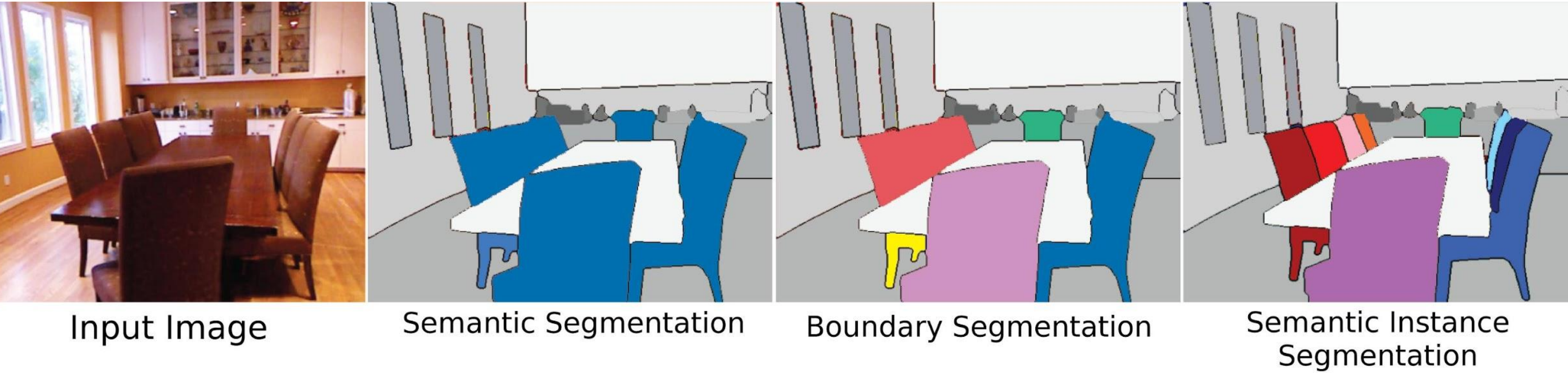
Higher level
"understanding"







Levels of abstraction in computer vision

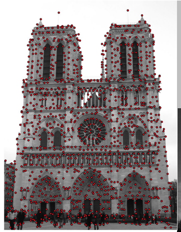


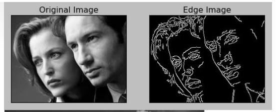
Levels of abstraction in computer vision

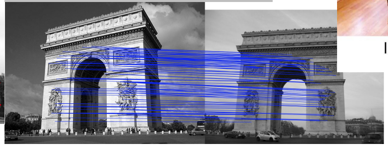
Raw pixels

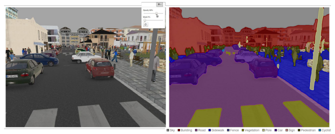
Concepts

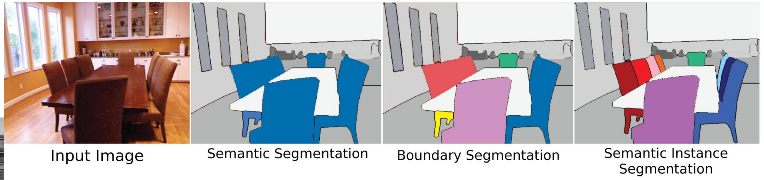
Higher level
"understanding"









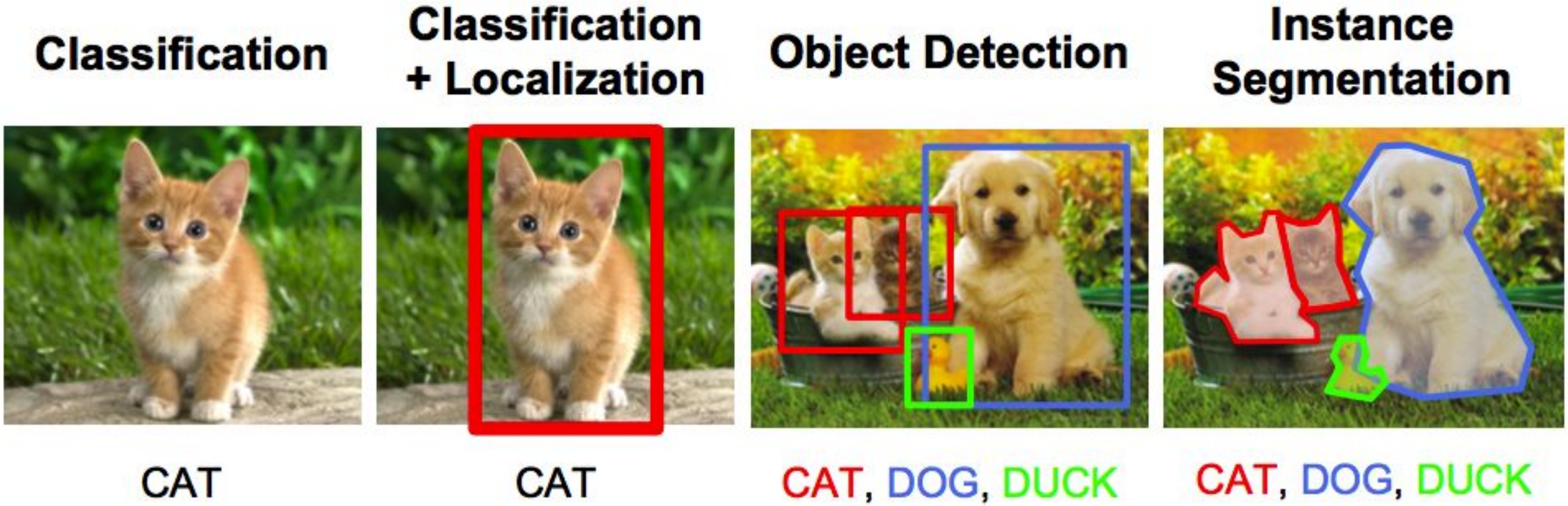


Levels of abstraction in computer vision

Raw pixels

Concepts

Higher level "understanding"

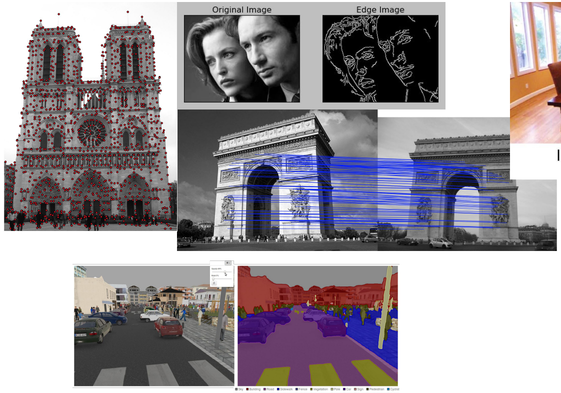


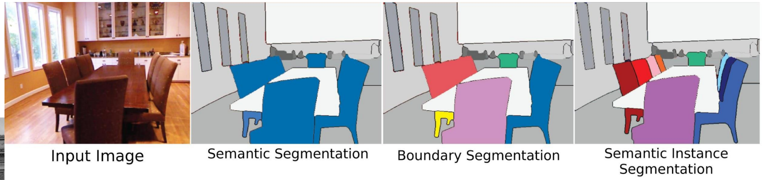
Levels of abstraction in computer vision

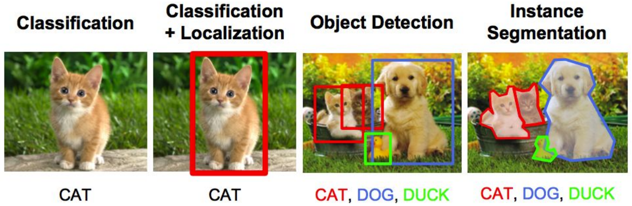
Raw pixels

Concepts

Higher level
"understanding"





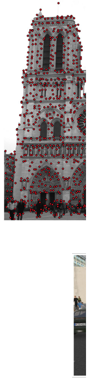


Levels of abstraction in computer vision

Raw pixels

Concepts

Higher level
"understanding"



"man in black shirt is playing guitar."

"construction worker in orange safety vest is working on road."

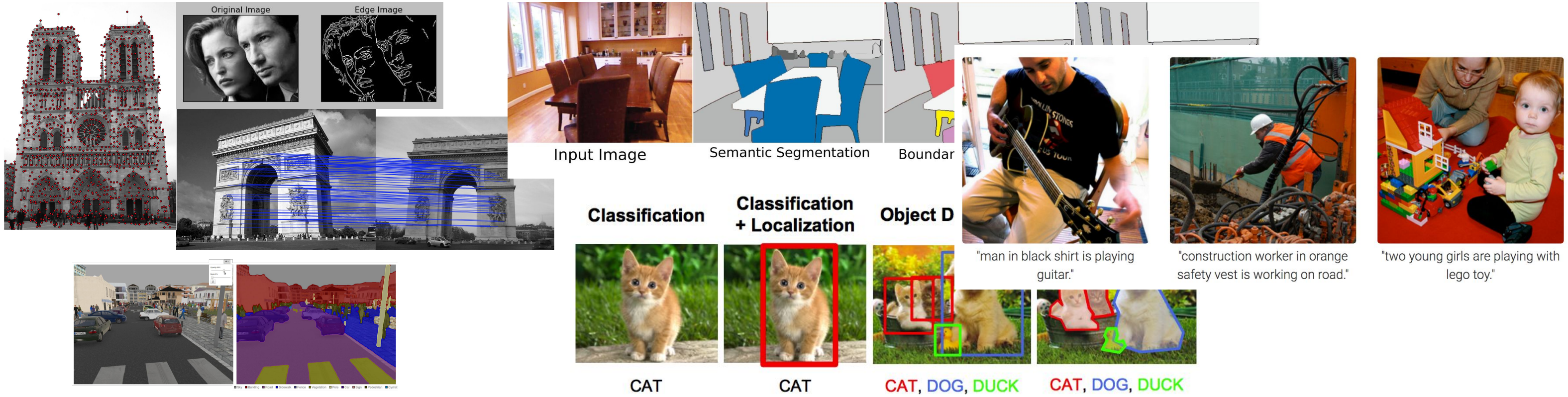
"two young girls are playing with lego toy."

Levels of abstraction in computer vision

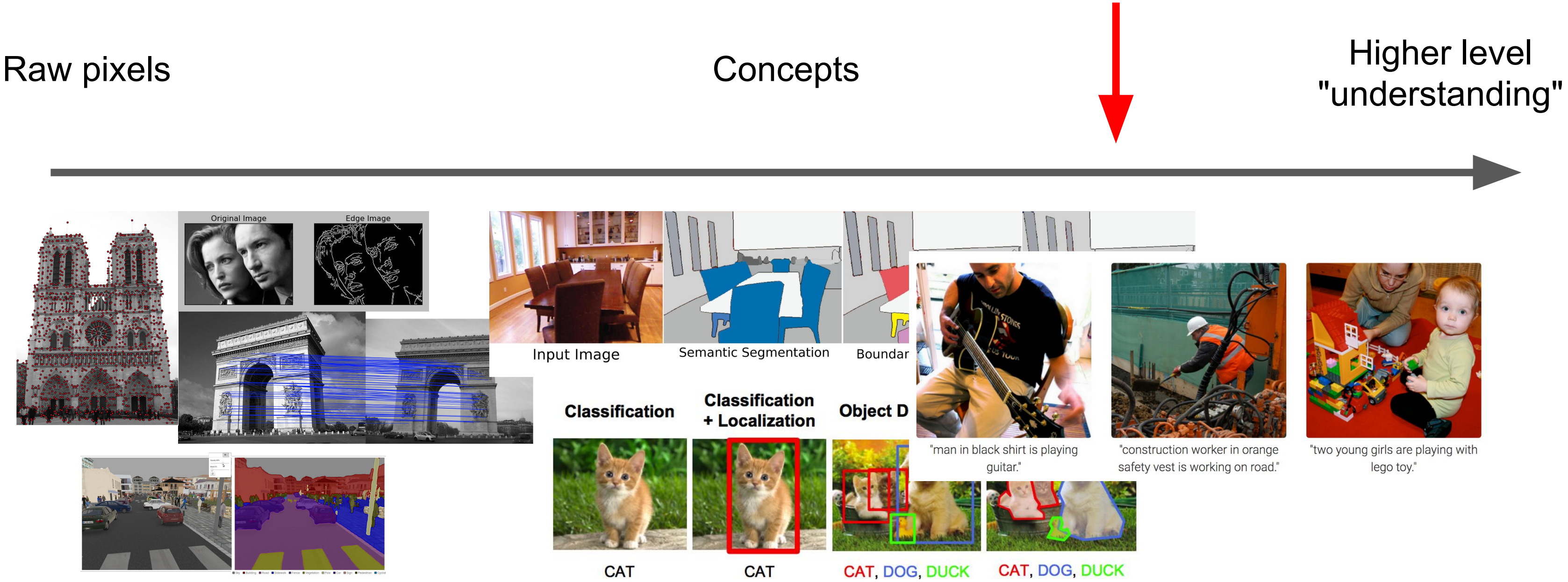
Raw pixels

Concepts

Higher level "understanding"



Levels of abstraction in computer vision



Levels of abstraction in computer vision

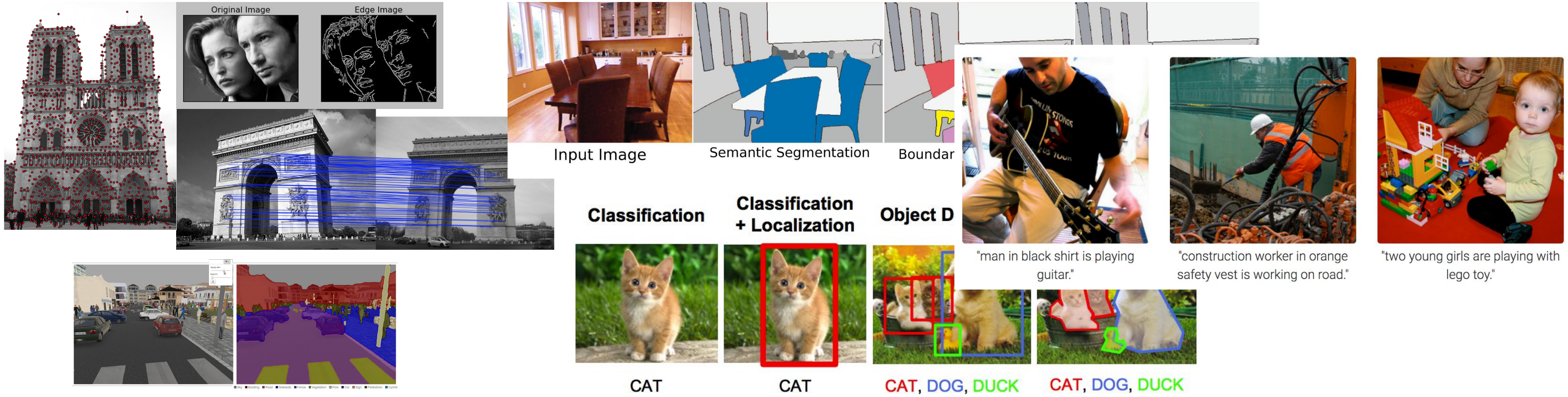
Easier

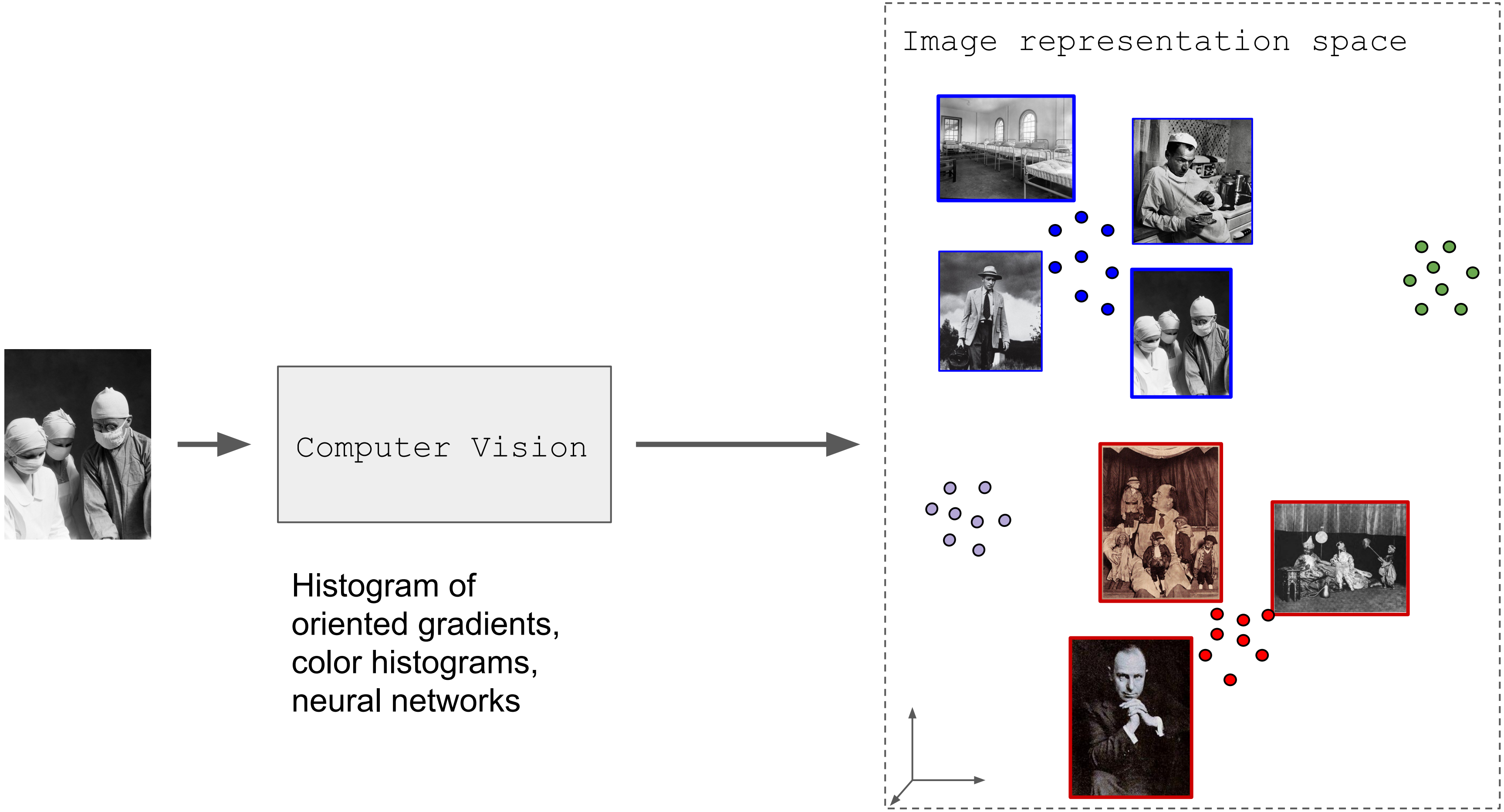
Hard

Raw pixels

Concepts

Higher level "understanding"




Computer Vision
Histogram of oriented gradients, color histograms, neural networks
Image representation space

Previous Work with Images in DH

Langmead et al. 2017.

Wevers and Lonij, 2017.

Previous Work with Images in DH

(a) Cities and Towns

(b) Homes and Living Conditions

(c) Intellectual and Creative Activity

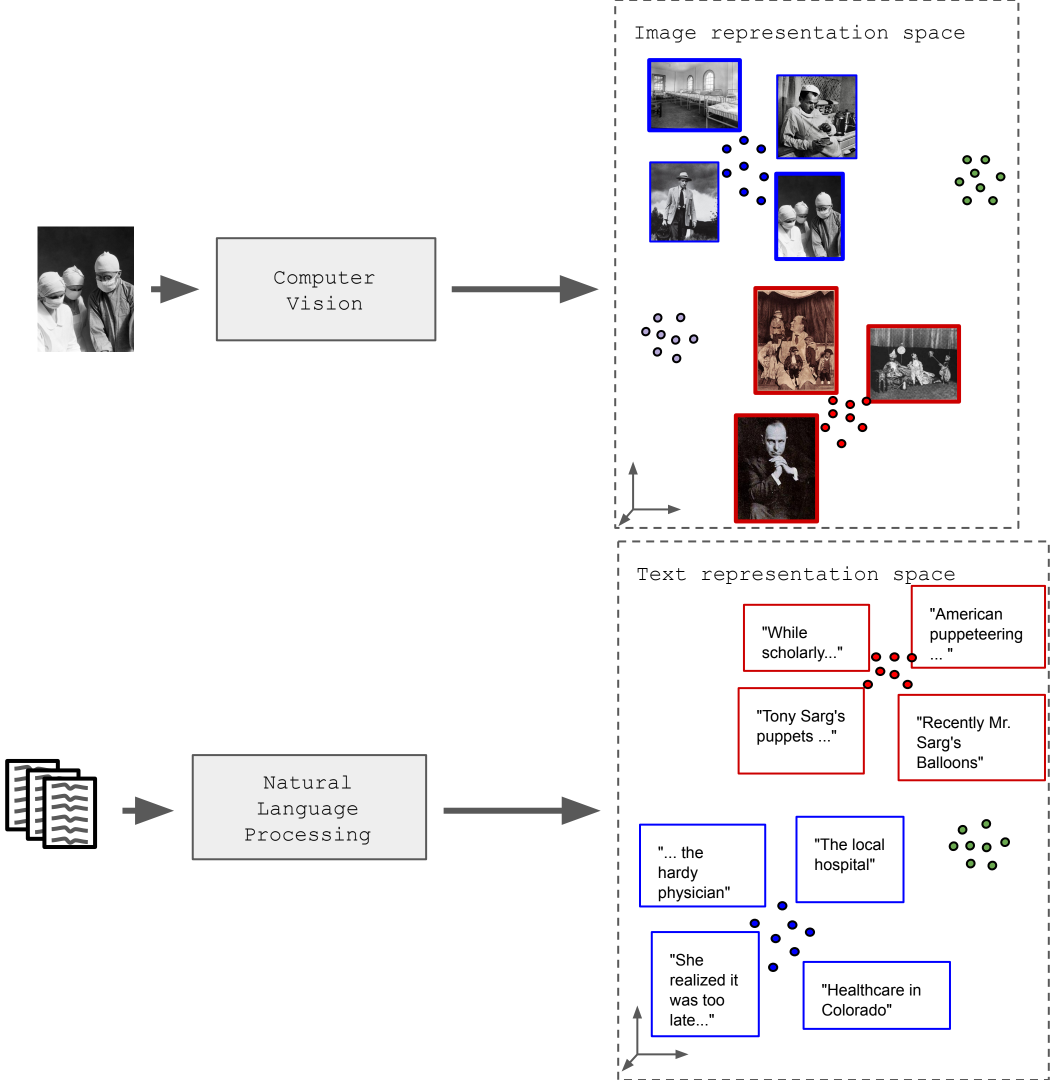
Taylor et al. 2017

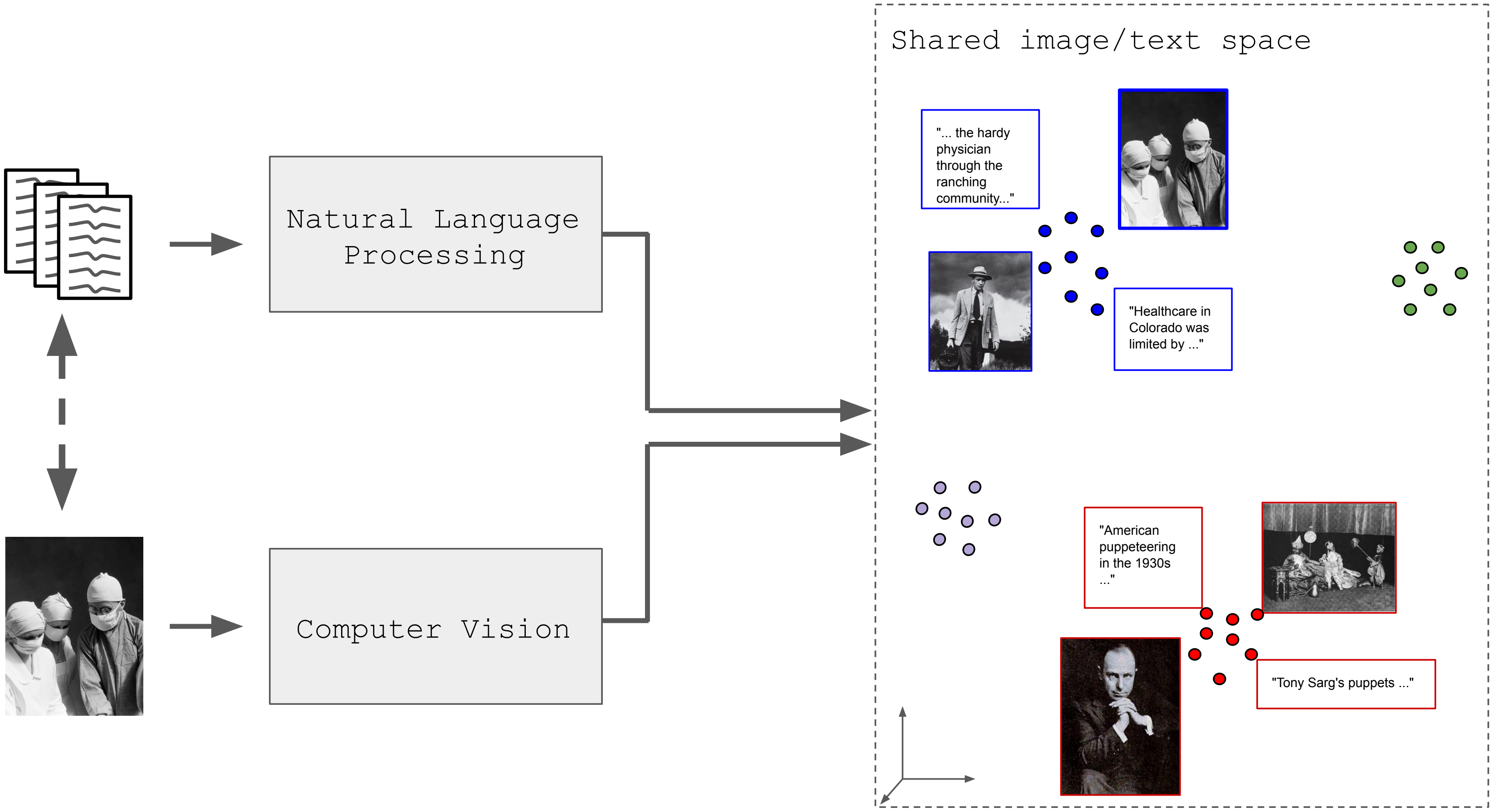
Do there exist a lot of image analysis tools for DH?

How do you take a first-pass look at a set of images from a data perspective?

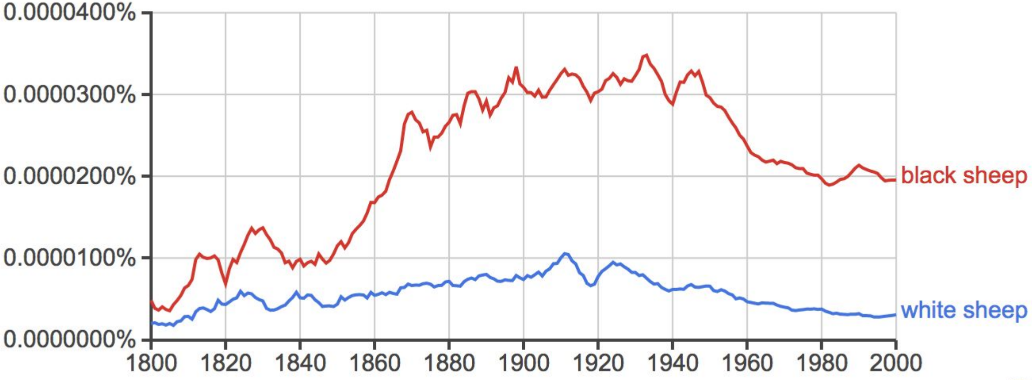
Do there exist a lot of image analysis tools for DH?

How do you take a first-pass look at a set of images from a data perspective?


Computer Vision
Image representation space
Natural Language Processing
Text representation space
"While scholarly..."
"American puppeteering ... "
"Tony Sarg's puppets ..."
"Recently Mr. Sarg's Balloons"
"... the hardy physician"
"The local hospital"
"She realized it was too late..."
"Healthcare in Colorado"


Natural Language Processing
Computer Vision
Shared image/text space
"... the hardy physician through the ranching community..."
"Healthcare in Colorado was limited by ..."
"American puppeteering in the 1930s ..."
"Tony Sarg's puppets ..."

- Similarities between text data and image data
- **Why should we want to model text and images jointly?**
- Computer vision and "why digital libraries?"
- The dataset/experiments
- Are concrete things easier to learn?


0.0000400%
0.0000300%
0.0000200%
0.0000100%
0.0000000%
1800
1820
1840
1860
1880
1900
1920
1940
1960
1980
2000
black sheep
white sheep

NASA

NASA

NASA

NASA

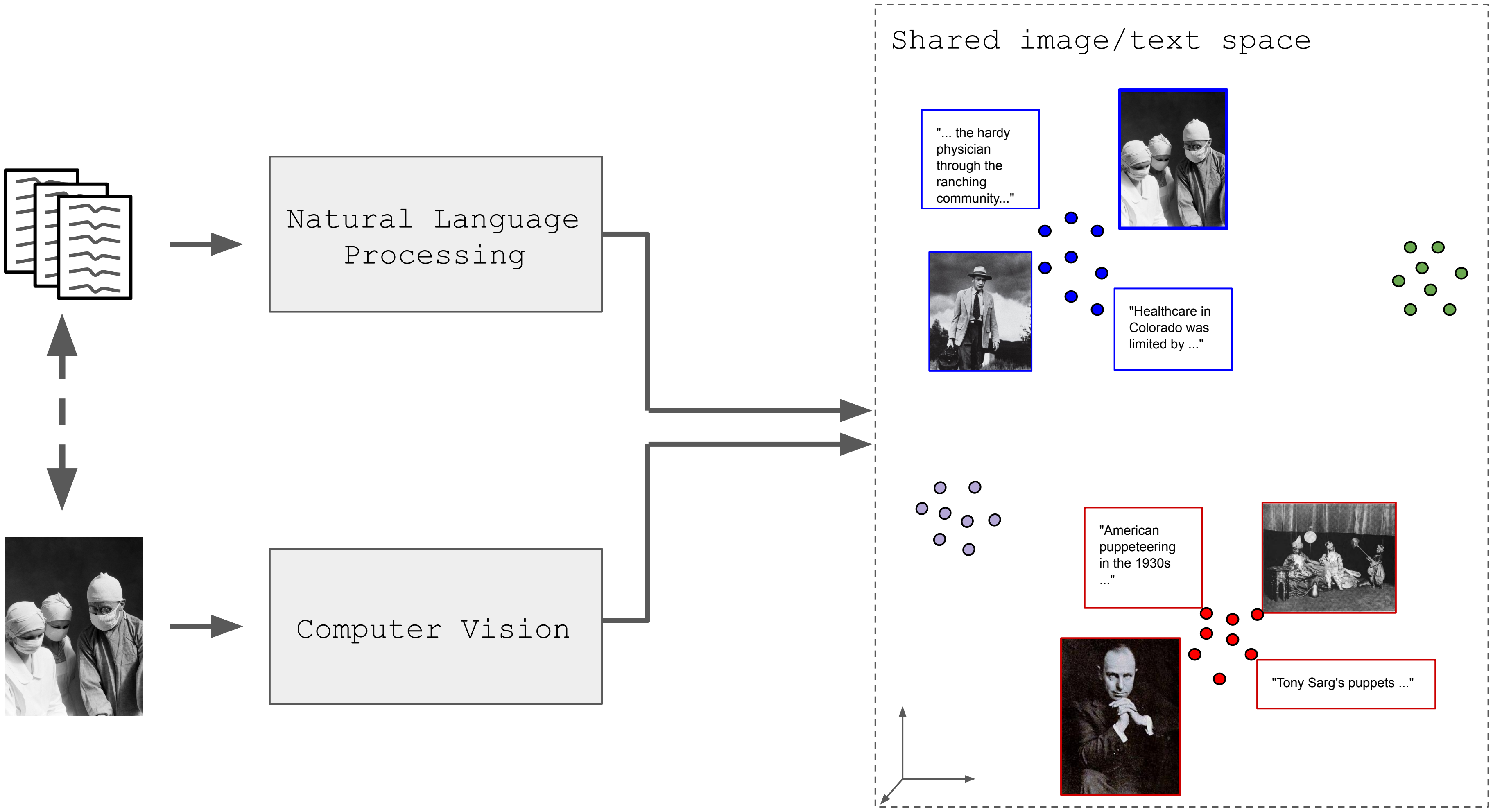
NASA

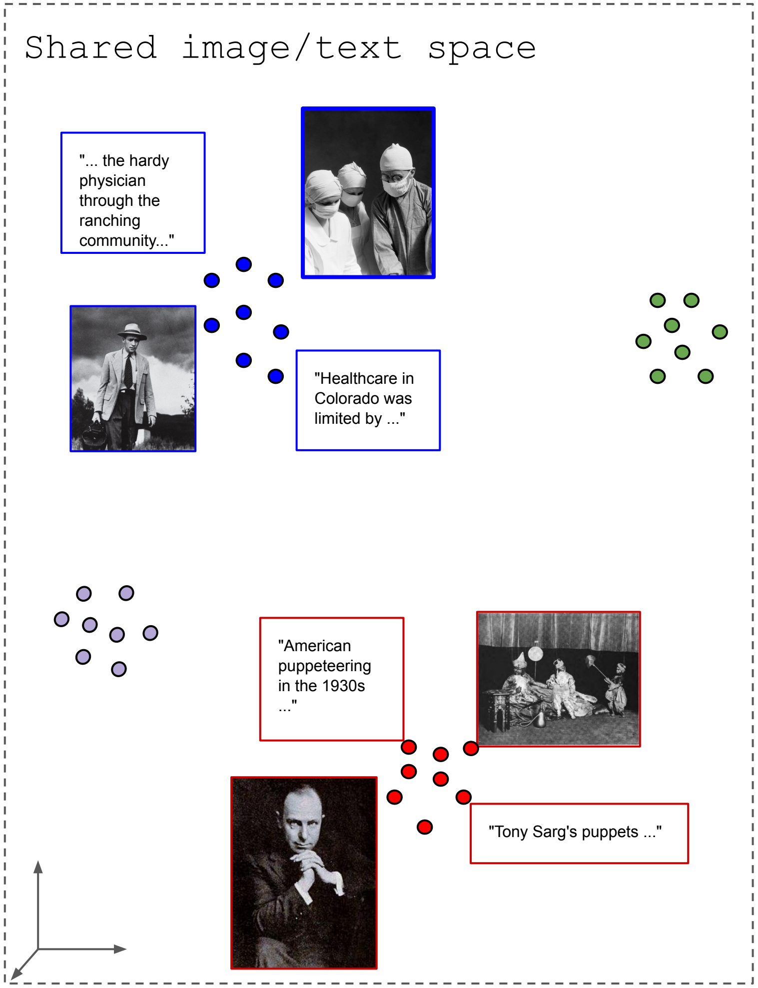
NASA

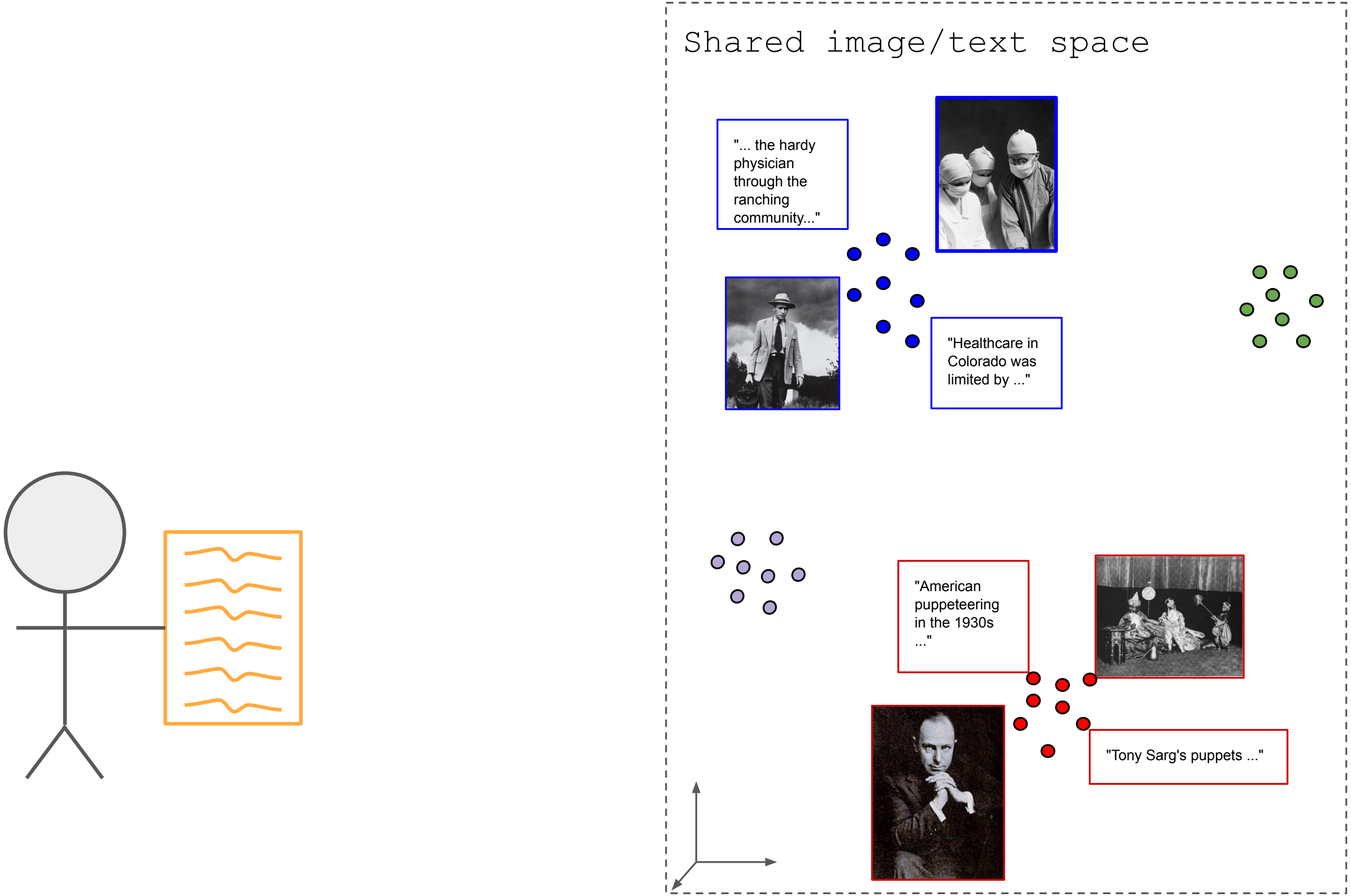
Democrats,
Dreamers,
Healthcare

UK Parliament,
Brexit,
EU

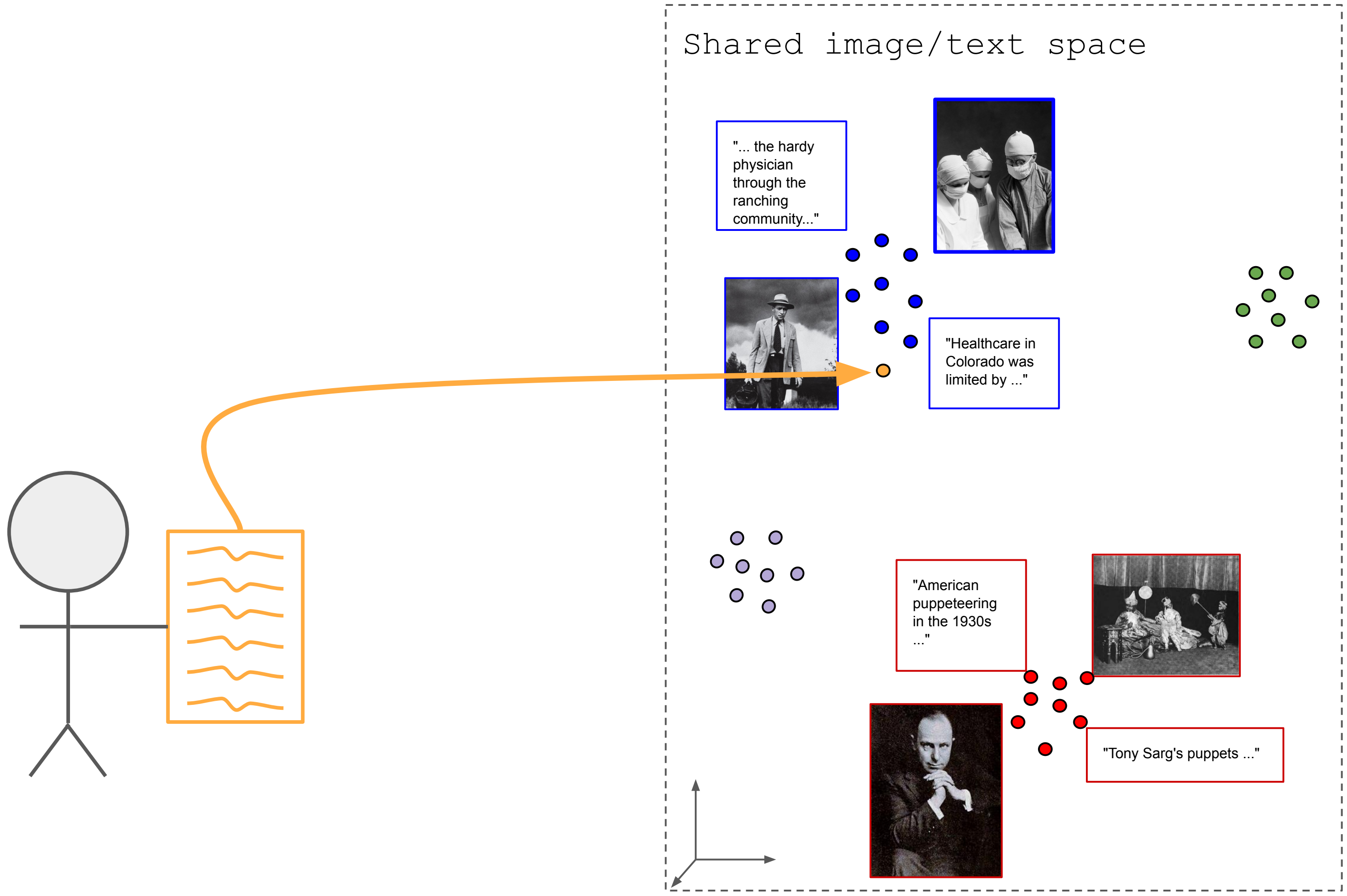
Republicans,
The Wall,
Handouts

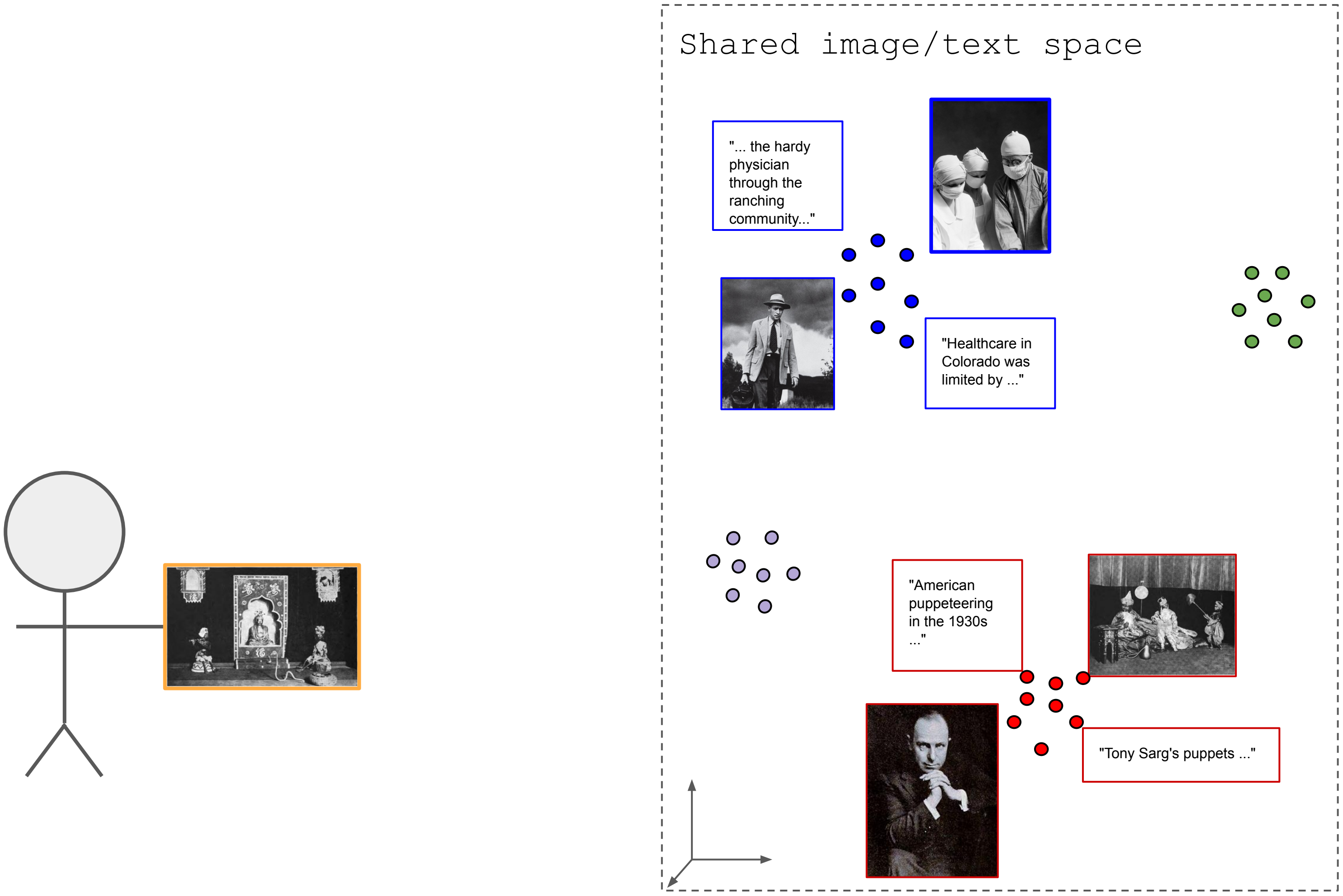

Natural Language Processing
Computer Vision
Shared image/text space
"... the hardy physician through the ranching community..."
"Healthcare in Colorado was limited by ..."
"American puppeteering in the 1930s ..."
"Tony Sarg's puppets ..."

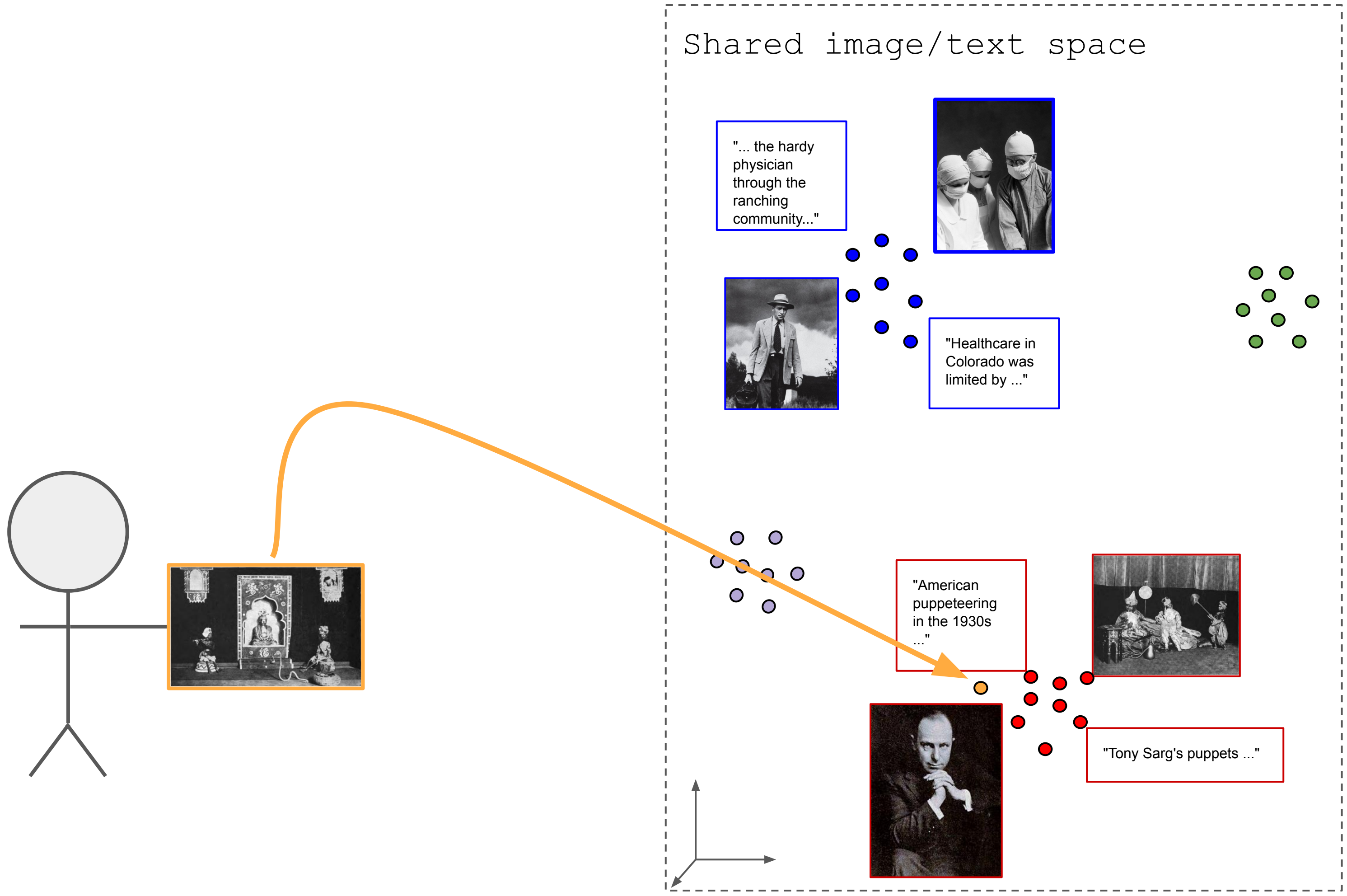

Shared image/text space
"... the hardy physician through the ranching community..."
"Healthcare in Colorado was limited by ..."
"American puppeteering in the 1930s ..."
"Tony Sarg's puppets ..."

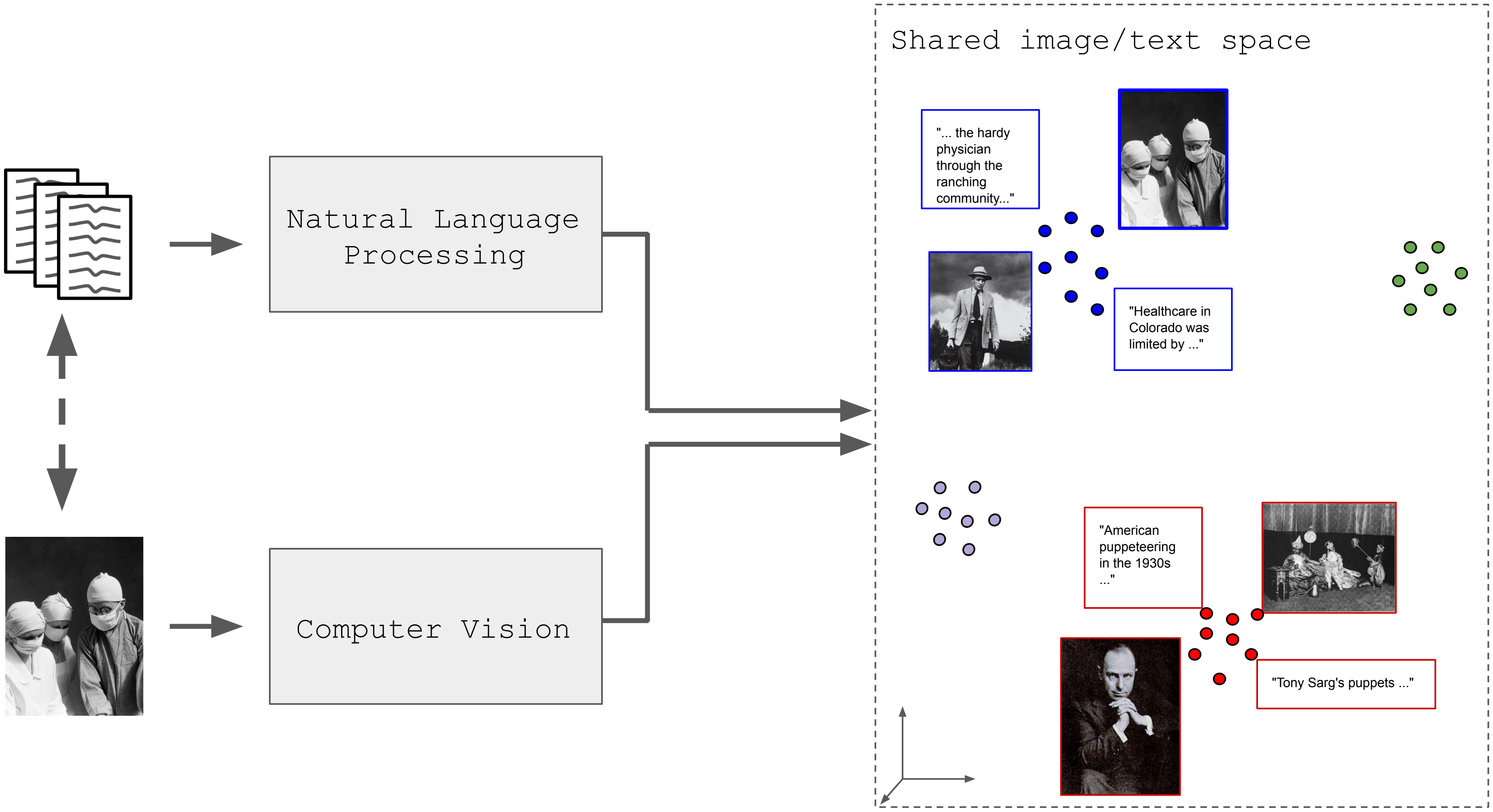


Shared image/text space
"... the hardy physician through the ranching community..."
"Healthcare in Colorado was limited by ..."
"American puppeteering in the 1930s ..."
"Tony Sarg's puppets ..."


Shared image/text space
"... the hardy physician through the ranching community..."
"Healthcare in Colorado was limited by ..."
"American puppeteering in the 1930s ..."
"Tony Sarg's puppets ..."


Shared image/text space
"... the hardy physician through the ranching community..."
"Healthcare in Colorado was limited by ..."
"American puppeteering in the 1930s ..."
"Tony Sarg's puppets ..."


Shared image/text space
"... the hardy physician through the ranching community..."
"Healthcare in Colorado was limited by ..."
"American puppeteering in the 1930s ..."
"Tony Sarg's puppets ..."



Natural Language Processing
Computer Vision
Shared image/text space
"... the hardy physician through the ranching community..."
"Healthcare in Colorado was limited by ..."
"American puppeteering in the 1930s ..."
"Tony Sarg's puppets ..."

A few caveats:

A few caveats:

This is an exploratory, pilot study with aspirations on the machine learning side.

A few caveats:

This is an exploratory, pilot study with aspirations on the machine learning side.

Can computer vision tools be applied here at all?

Does multimodal learning make sense to apply here?

Is the issue of compounding noise insurmountable?

Can organize images/text in an unsupervised fashion?

- Similarities between text data and image data
- Why should we want to model text and images jointly?
- **Computer vision and "why digital libraries?"**
- The dataset/experiments
- Are concrete things easier to learn?

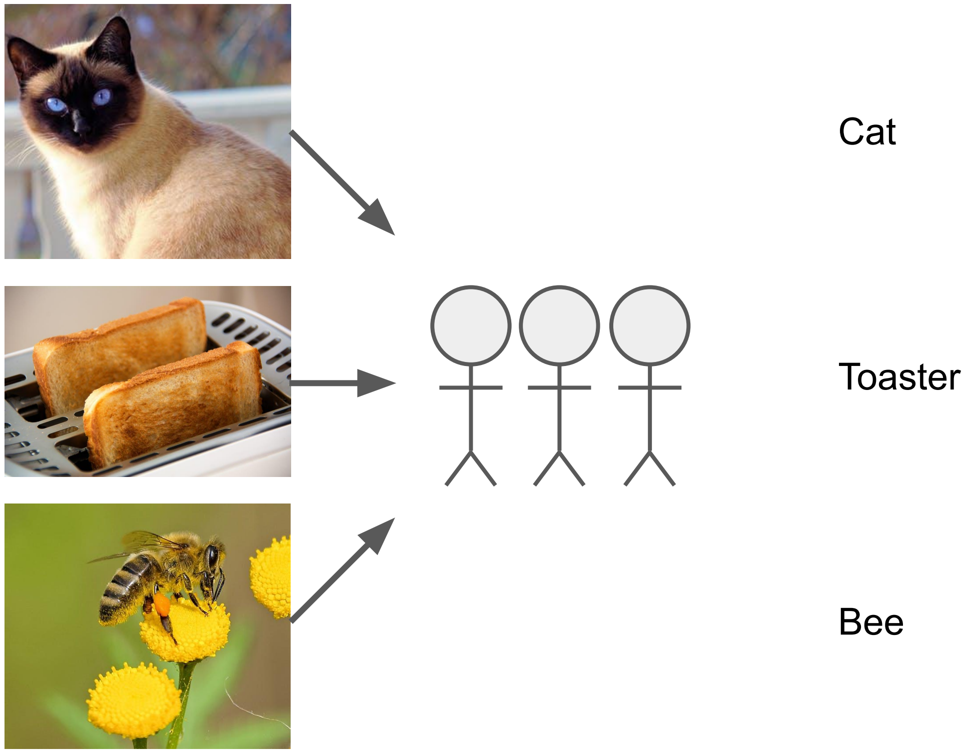
A brief aside into computer vision...

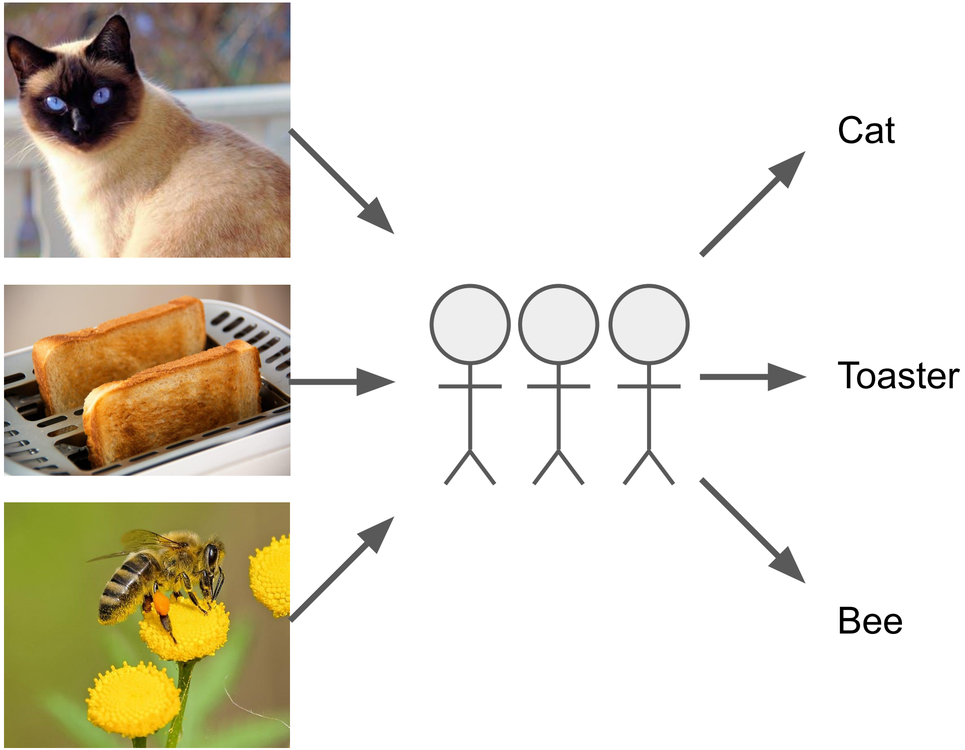


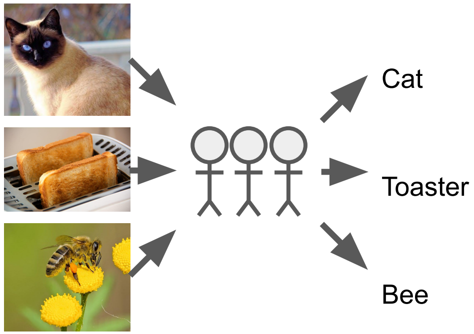
Cat

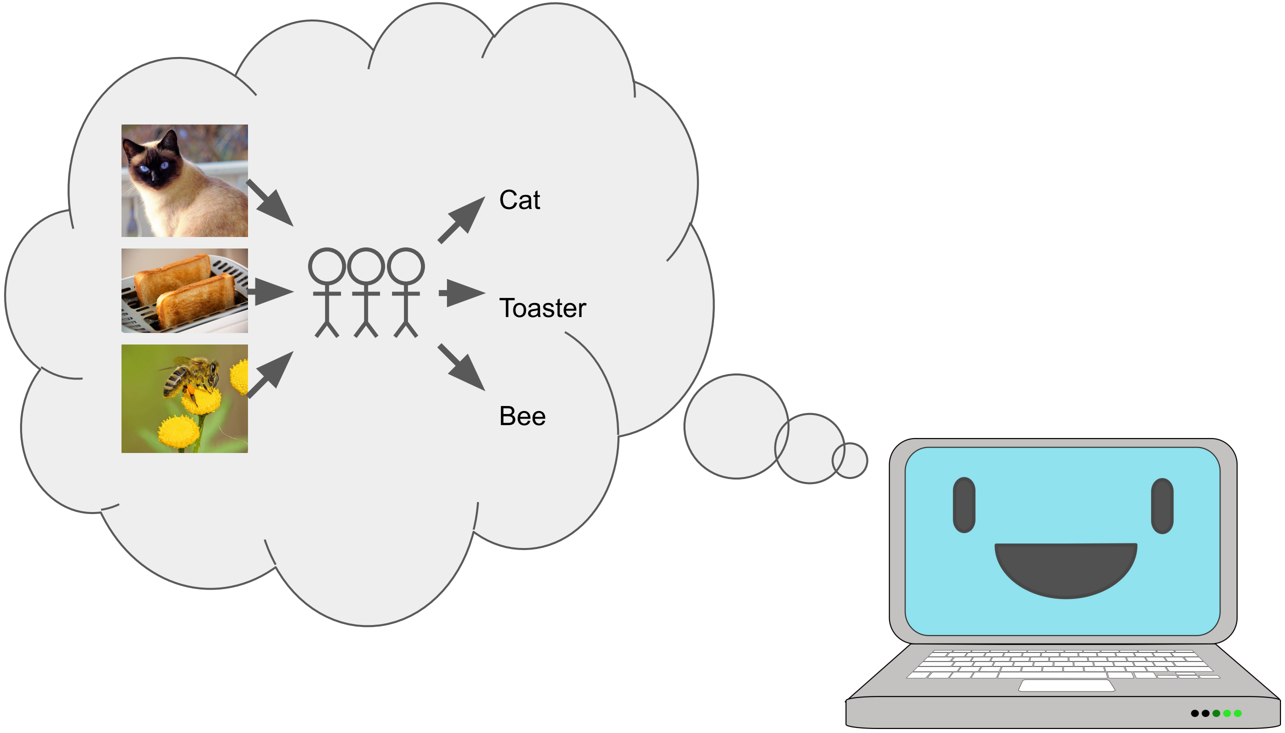
Toaster

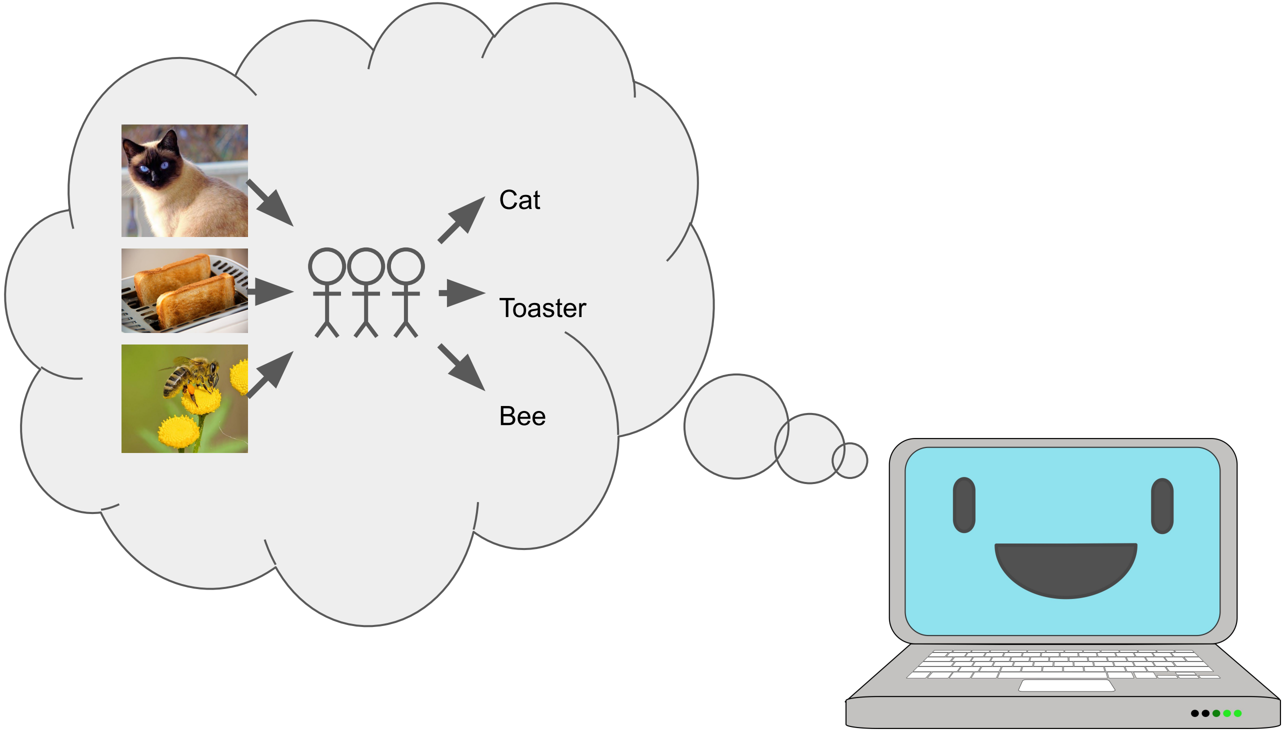
Bee


Cat
Toaster
Bee

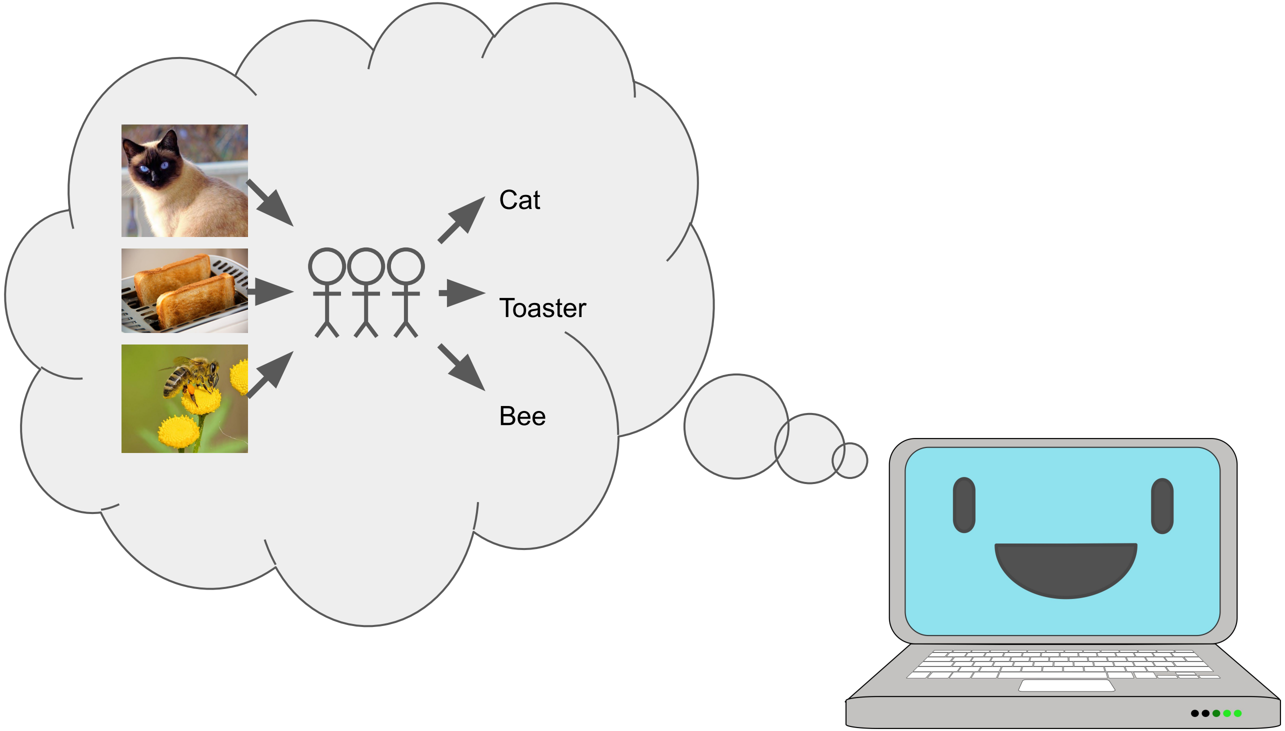

Cat
Toaster
Bee

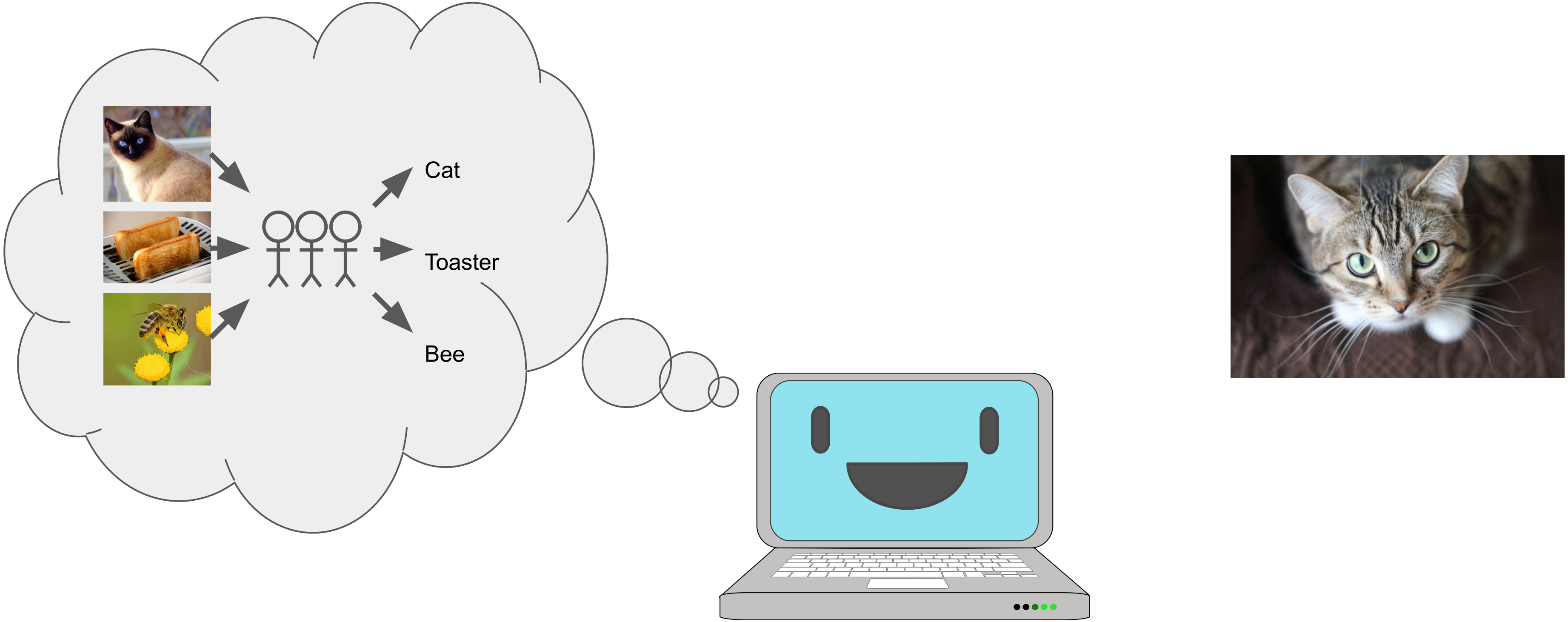

Cat
Toaster
Bee


Cat
Toaster
Bee

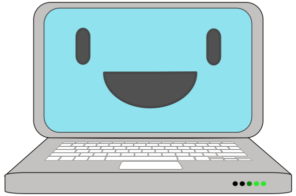


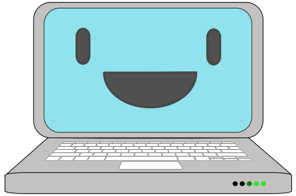
I can mimic this behavior!


Cat
Toaster
Bee



I am 99.6% sure this is a photo of a cat.

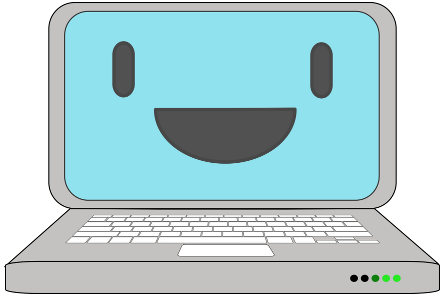




Object detection
=/=
image understanding

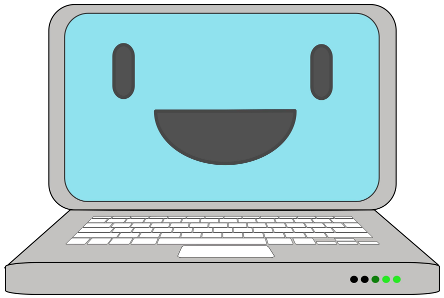
[Karpathy 2012]

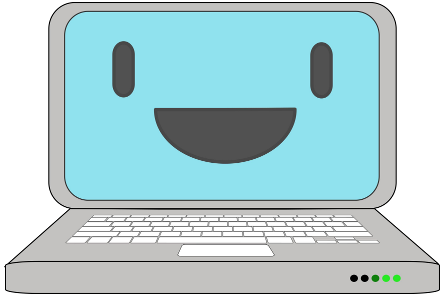
[Karpathy 2012]



I am 86.6% sure this photo has a person in it.

[Karpathy 2012]



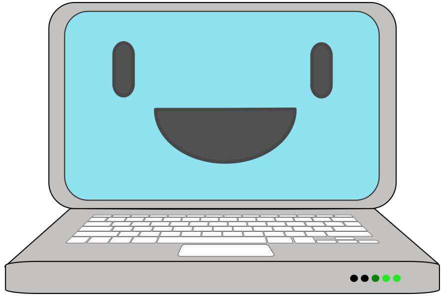


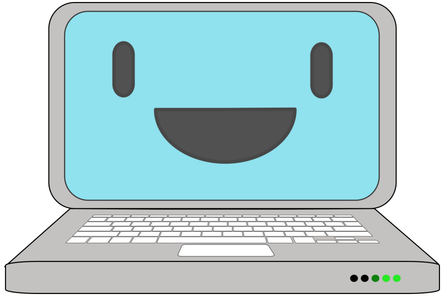
clarifai

PREDICTED CONCEPT	PROBABILITY
people	0.999
adult	0.996
man	0.988
one	0.984
wear	0.981
military	0.974
war	0.969
two	0.958
soldier	0.947

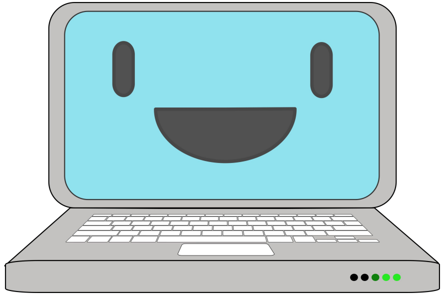
I think it's a person riding a horse in a field and he seems 😐.

How can we expect a computer vision algorithm to learn without fuller context?



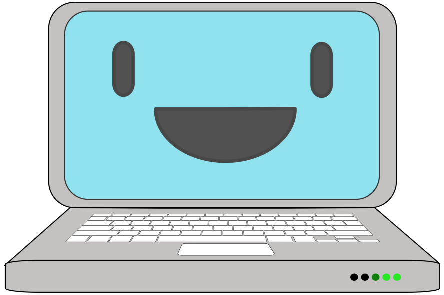


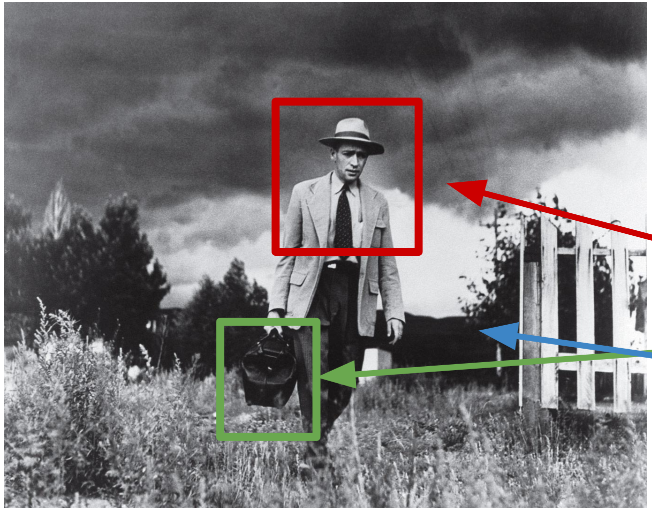
[Smith 1945. "Country Doctor." Subject: Dr. Ernest Ceriani]



Although lauded for his war photography, W. Eugene Smith left his most enduring mark with a series of midcentury photo essays for LIFE magazine. The Wichita, Kans.–born photographer spent weeks immersing himself in his subjects' lives, from a South Carolina nurse-midwife to the residents of a Spanish village. His aim was to see the world from the perspective of his subjects—and to compel viewers to do the same. "I do not seek to possess my subject but rather to give myself to it," he said of his approach. Nowhere was this clearer than in his landmark photo essay "Country Doctor." Smith spent 23 days with Dr. Ernest Ceriani in and around Kremmling, Colo., trailing the hardy physician through the ranching community of 2,000 souls beneath the Rocky Mountains. He watched him tend to infants, deliver injections in the backseats of cars, develop his own x-rays, treat a man with a heart attack and then phone a priest to give last rites. By digging so deeply into his assignment, Smith created a singular, starkly intimate glimpse into the life of a remarkable man. It became not only the most influential photo essay in history but the aspirational template for the form.

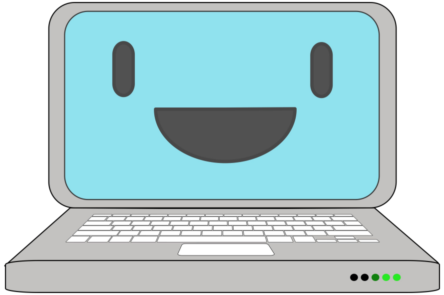
[Smith 1945. "Country Doctor." Subject: Dr. Ernest Ceriani; Annotation from Time 100 Most Influential Photographs]

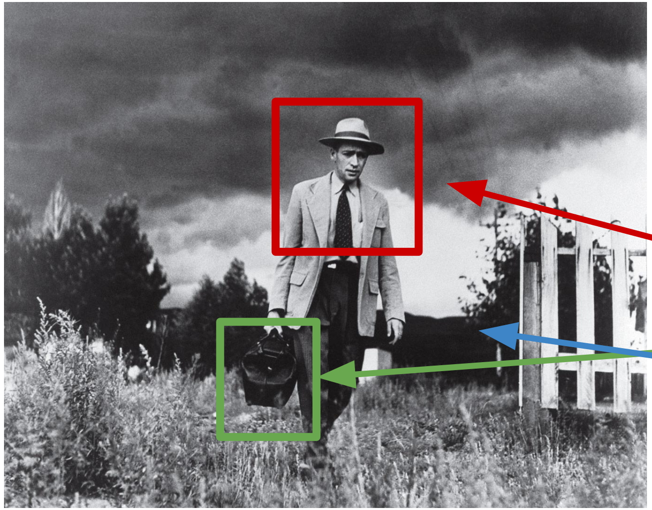




Although lauded for his war photography, W. Eugene Smith left his most enduring mark with a series of midcentury photo essays for LIFE magazine. The Wichita, Kans.–born photographer spent weeks immersing himself in his subjects’ lives, from a South Carolina nurse-midwife to the residents of a Spanish village. His aim was to see the world from the perspective of his subjects—and to compel viewers to do the same. “I do not seek to possess my subject but rather to give myself to it,” he said of his approach. Nowhere was this clearer than in his landmark photo essay “Country **Doctor**.” Smith spent 23 days with **Dr. Ernest Ceriani** in and around Kremmling, Colo., trailing the hardy physician through the ranching community of 2,000 souls beneath the **Rocky Mountains**. He watched him tend to infants, deliver injections in the backseats of cars, develop his own x-rays, treat a man with a heart attack and then phone a priest to give last rites. By digging so deeply into his assignment, Smith created a singular, starkly intimate glimpse into the life of a remarkable man. It became not only the most influential photo essay in history but the aspirational template for the form.

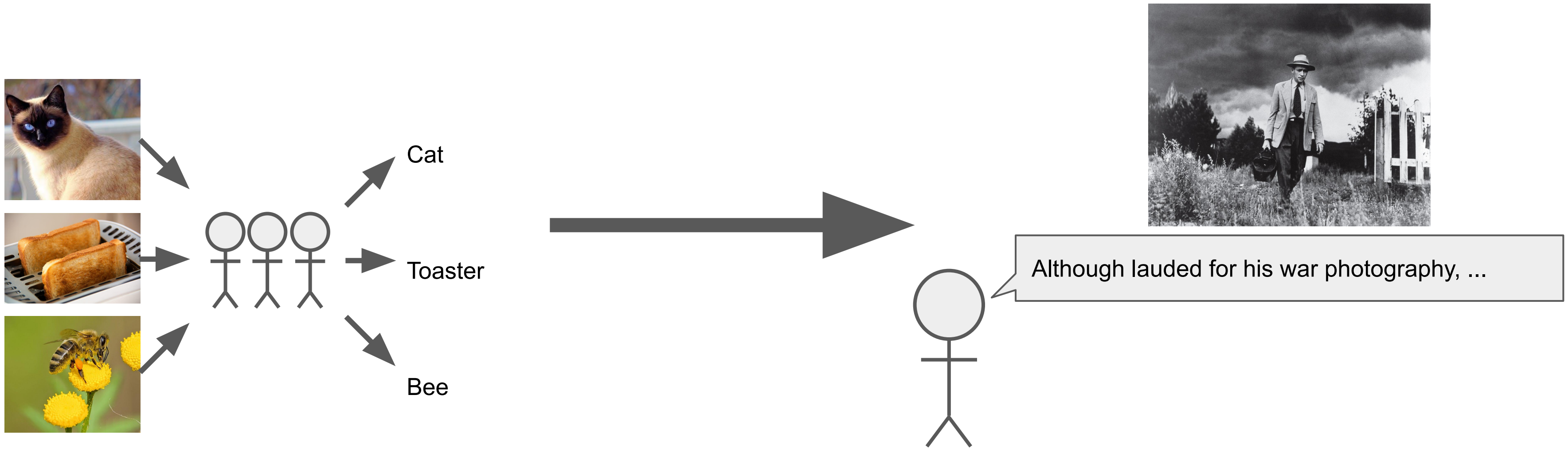
[Smith 1945. "Country Doctor." Subject: Dr. Ernest Ceriani; Annotation from Time 100 Most Influential Photographs]

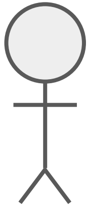




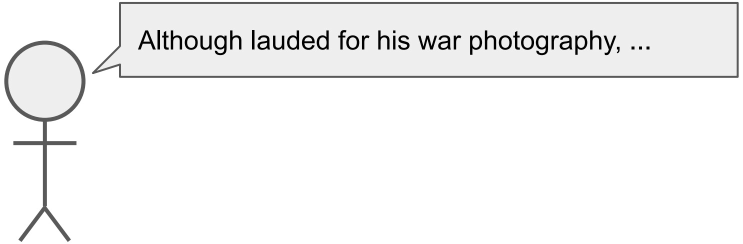
Although lauded for his war photography, W. Eugene Smith left his most enduring mark with a series of midcentury photo essays for LIFE magazine. The Wichita, Kans.–born photographer spent weeks immersing himself in his subjects' lives, from a South Carolina nurse-midwife to the residents of a Spanish village. His aim was to see the world from the perspective of his subjects—and to compel viewers to do the same. "I do not seek to possess my subject but rather to give myself to it," he said of his approach. Nowhere was this clearer than in his landmark photo essay "Country ***Doctor****." Smith spent 23 days with* ***Dr. Ernest Ceriani*** *in and around Kremmling, Colo., trailing the hardy physician through the ranching community of 2,000 souls beneath the* ***Rocky Mountains****. He watched him tend to infants, deliver injections in the backseats of cars, develop his own x-rays, treat a man with a heart attack and then phone a priest to give last rites. By digging so deeply into his assignment, Smith created a singular, starkly intimate glimpse into the life of a remarkable man. It became not only the most influential photo essay in history but the aspirational template for the form.*

[Smith 1945. "Country Doctor." Subject: Dr. Ernest Ceriani; Annotation from Time 100 Most Influential Photographs]


Cat
Toaster
Bee
Although lauded for his war photography, ...



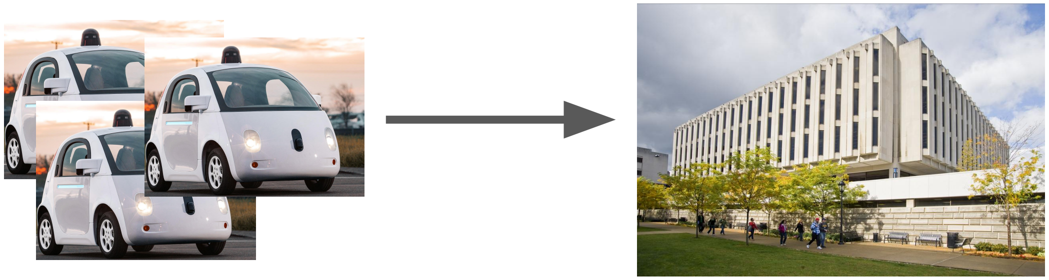
Although lauded for his war photography, ...

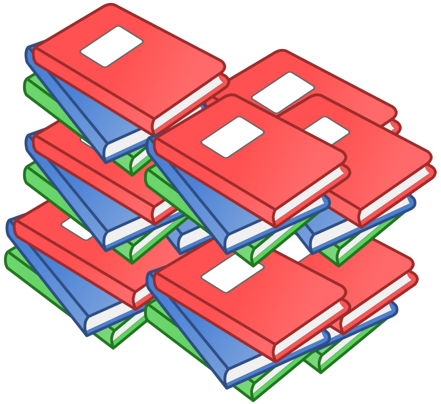


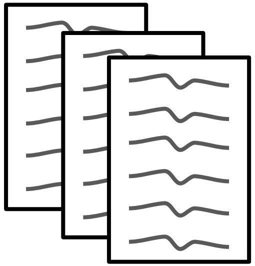
1. Address interesting digital humanities questions
2. Contains images associated with text
3. Contains lots of image/text pairs (preferably 100K+)

HATHI
TRUST

Google books

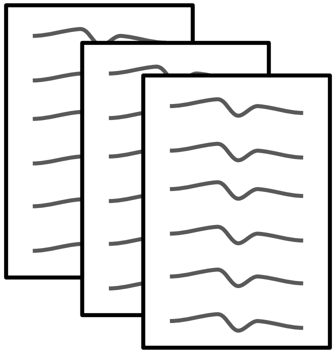


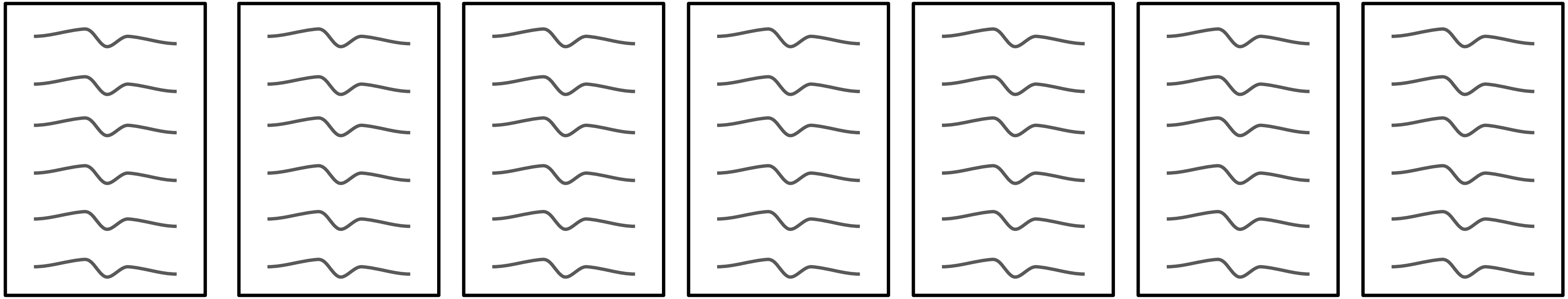




OCR

Evidence to support arguments in the digital humanities, and new questions.

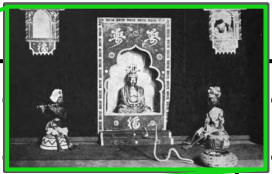




Two scenes from *A Stolen Beauty and the Great Jewel*, as acted by Tony Sarg's puppets.

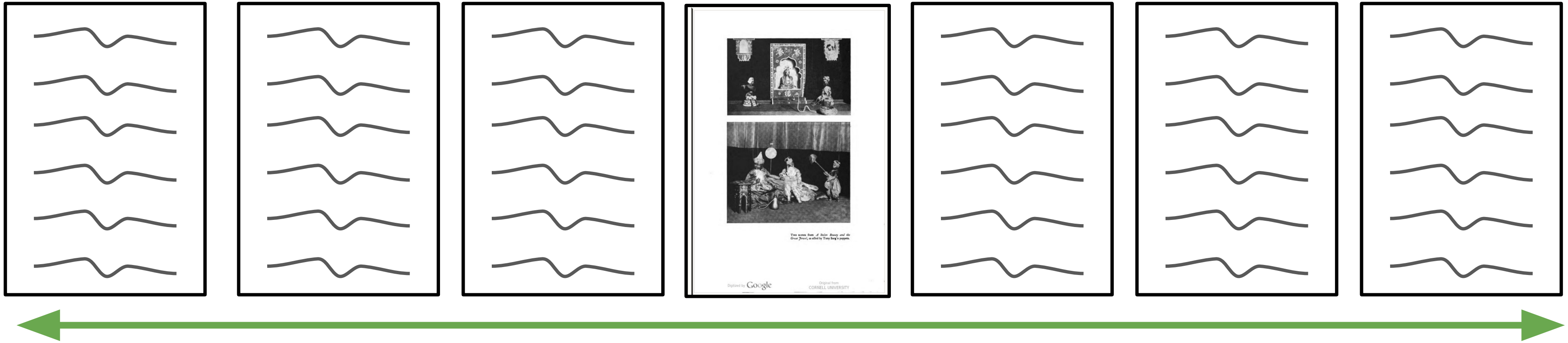
Two scenes from *A Stolen Beauty and the Great Jewel*, as acted by Tony Sarg's puppets.

Two scenes from *A Stolen Beauty and the Great Jewel*, as acted by Tony Sarg's puppets.

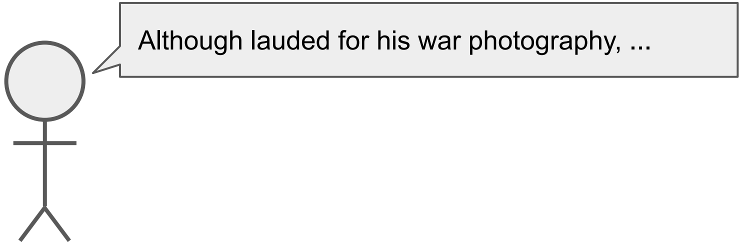


Two scenes from *A Stolen Beauty and the Great Jewel*, as acted by Tony Sarg's puppets.

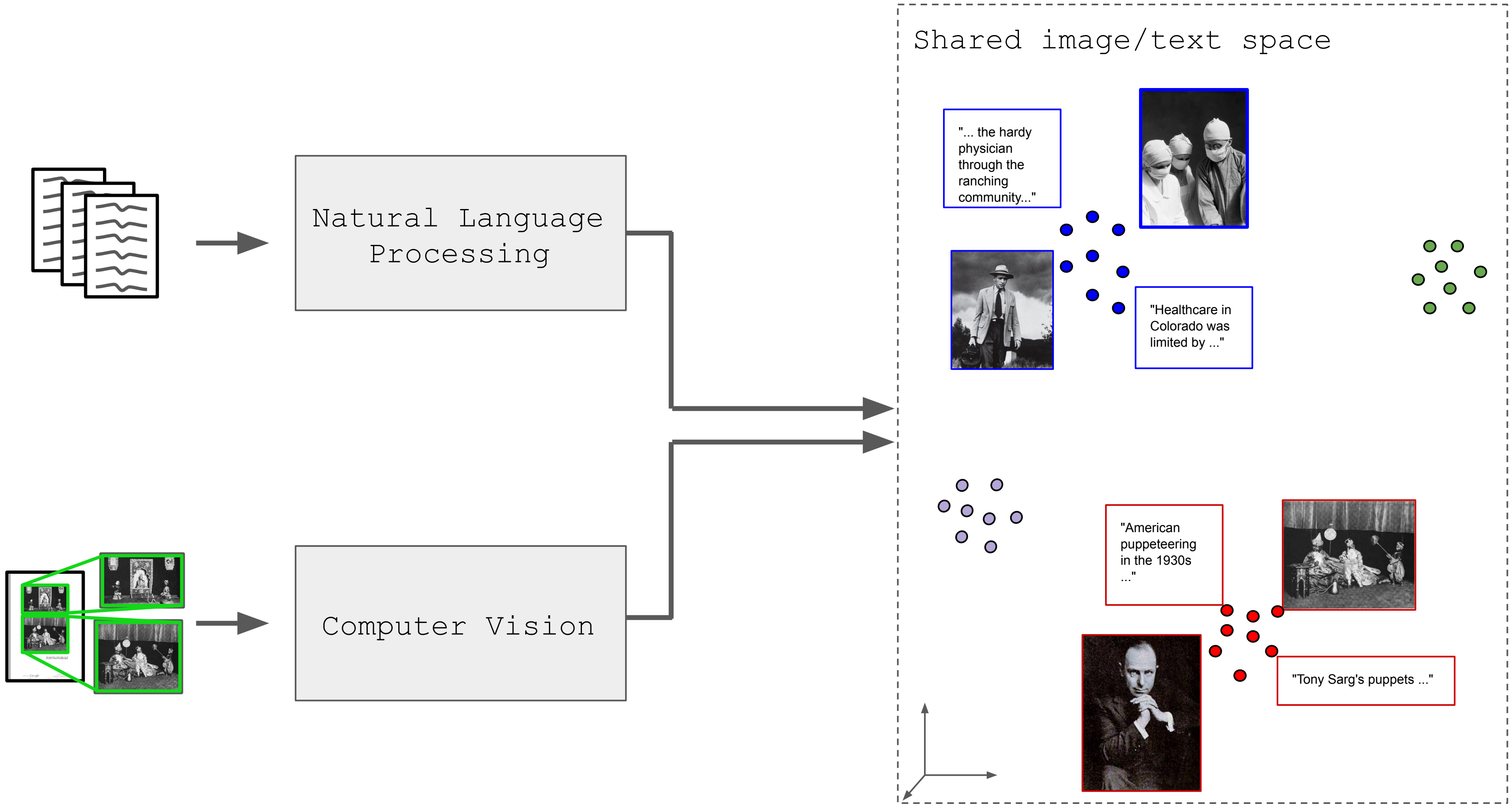
Two scenes from *A Stolen Beauty and the Great Jewel*, as acted by Tony Sarg's puppets.



Can this surrounding text provide adequate context for visual understanding?



1. Address interesting digital humanities questions
2. Contains images associated with text
3. Contains lots of image/text pairs (preferably 100K+)


Natural Language Processing
Computer Vision
Shared image/text space
"... the hardy physician through the ranching community..."
"Healthcare in Colorado was limited by ..."
"American puppeteering in the 1930s ..."
"Tony Sarg's puppets ..."

- Similarities between text data and image data
- Why should we want to model text and images jointly?
- Computer vision and "why digital libraries?"
- **The dataset/experiments**
- Are concrete things easier to learn?

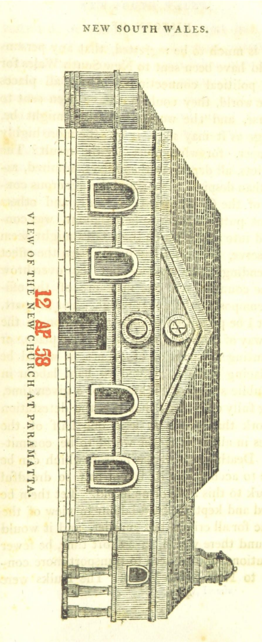
British Library Dataset

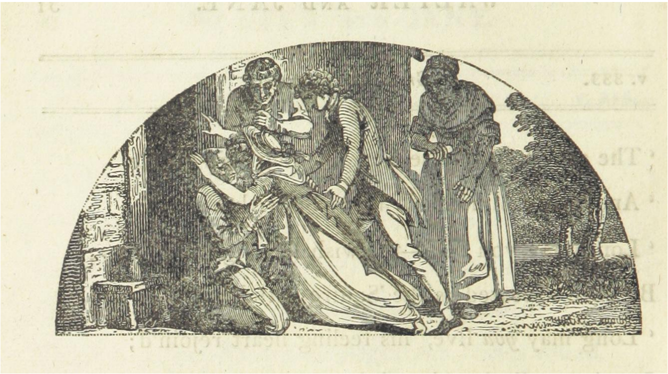
- Released by British Library to the public domain

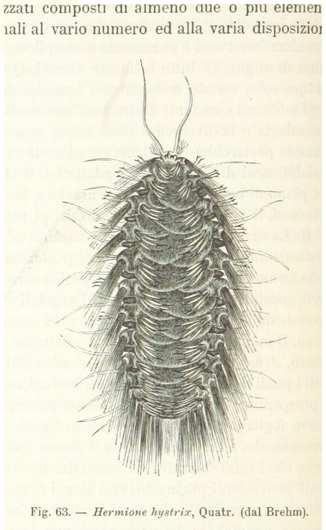
- 49,455 digitised books (65,227 volumes) largely from the 19th Century

- 405K images associated with text in +/-3 pages with a mean of 2.3K tokens. We use only use books that are in english

Example "medium" images







Example "plate" images

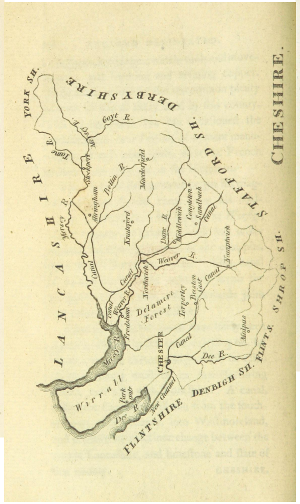
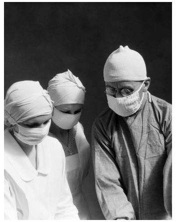
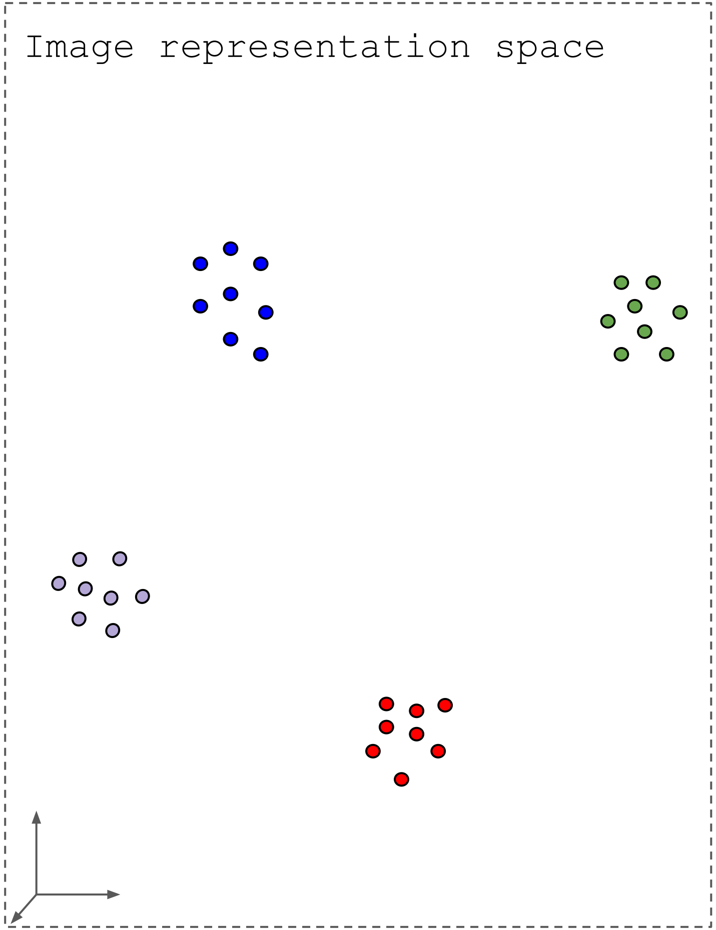
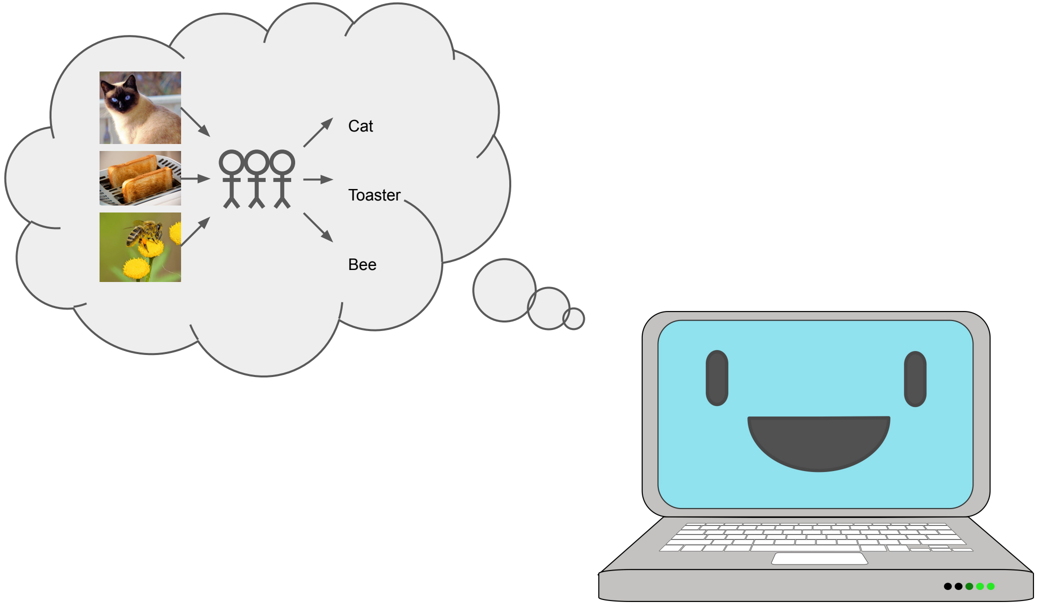


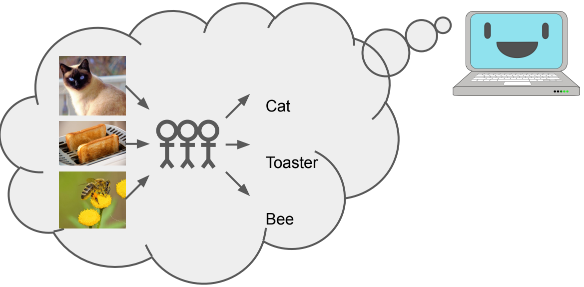
Image Models

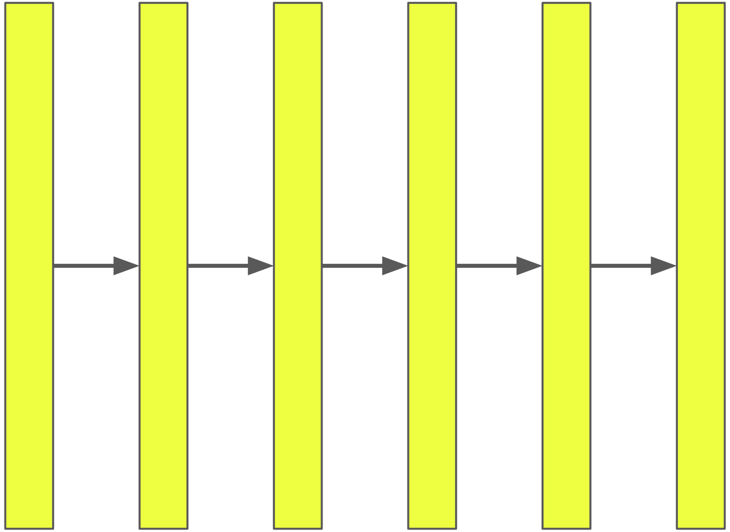


Computer Vision



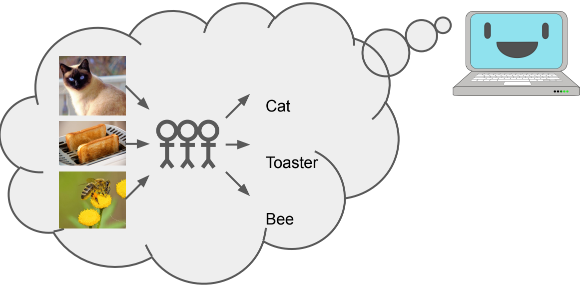

Cat
Toaster
Bee





50 layers of convolutions, dropout, batch normalization, residual connections...

[He et al. 2015]



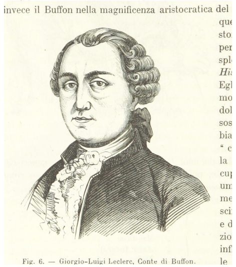
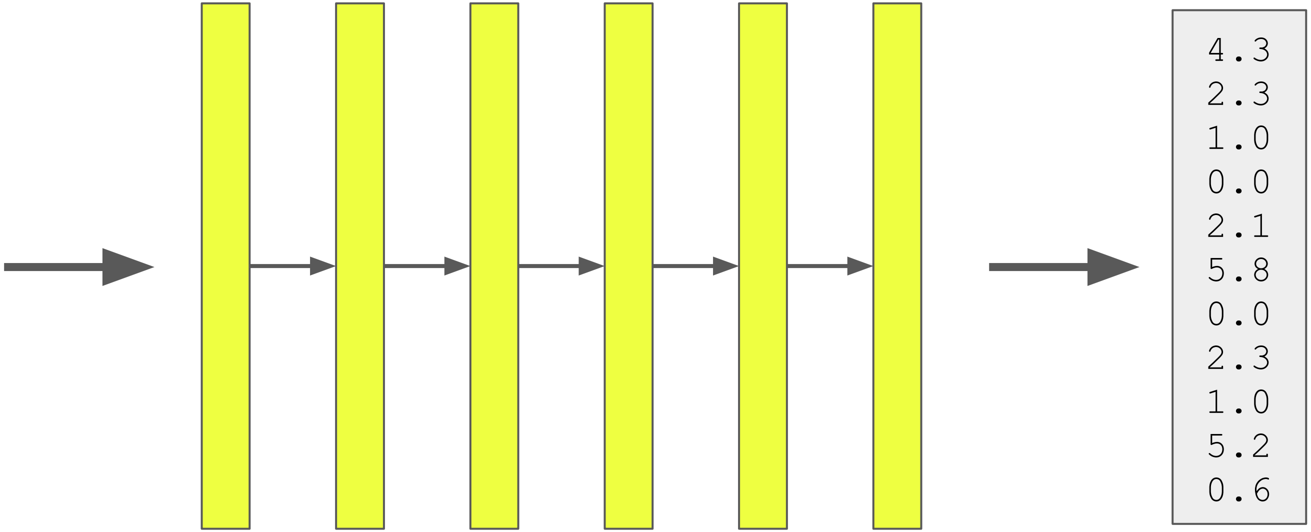
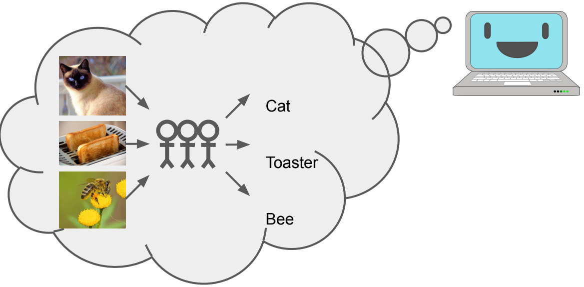

invece il Buffon nella magnificenza aristocratica del

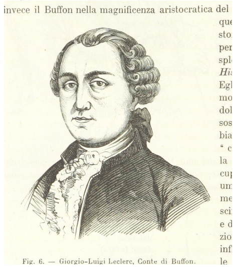
Fig. 6. — Giorgio-Luigi Leclerc, Conte di Buffon.



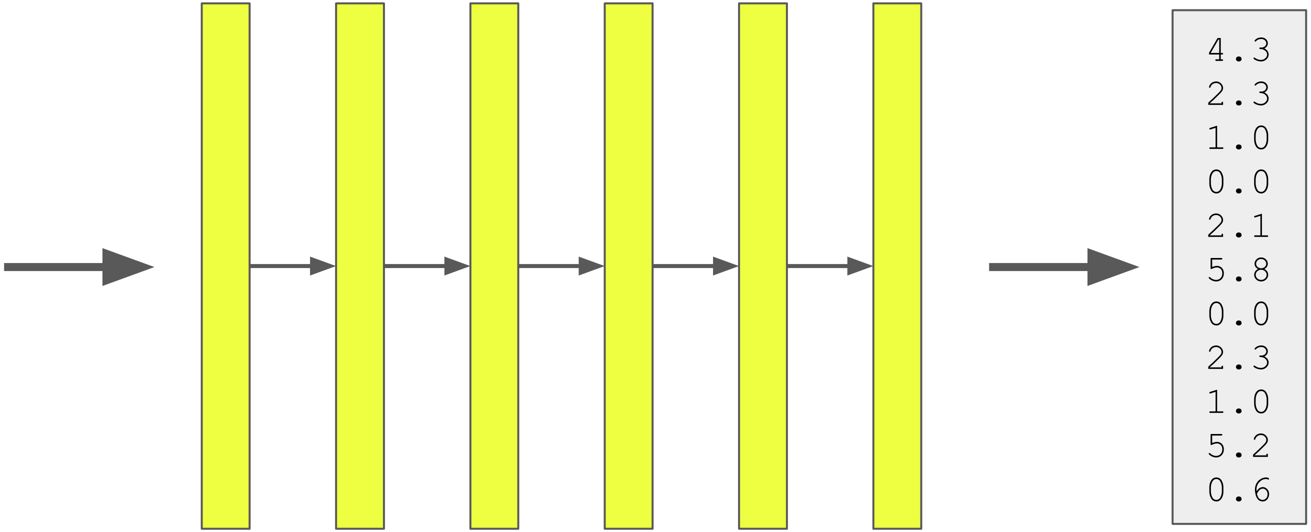
50 layers of convolutions, dropout, batch normalization, residual connections...

[He et al. 2015]





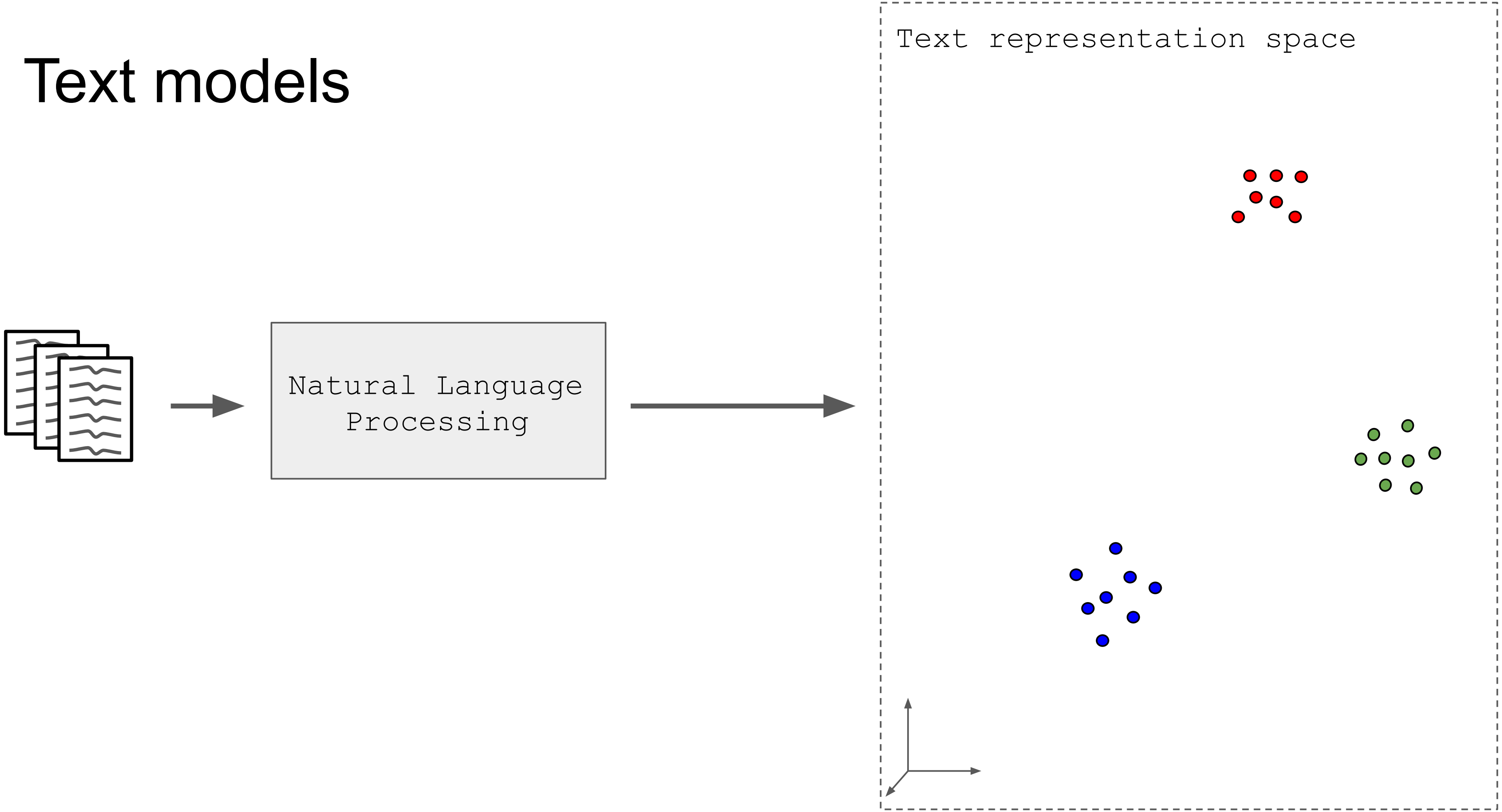
2096
dimensions



50 layers of convolutions, dropout, batch normalization, residual connections...

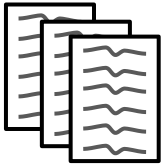
[He et al. 2015]

Text models



Text models

Text models



Natural Language Processing

Base Methods

- unigram vectors (uni)
- tfidf vectors (tfidf)

man	woman	doctor	hill	tree	axe	happy	sad	red	mean	dog	cat
0	1	1	1	1	0	1	0	0	0	1	0

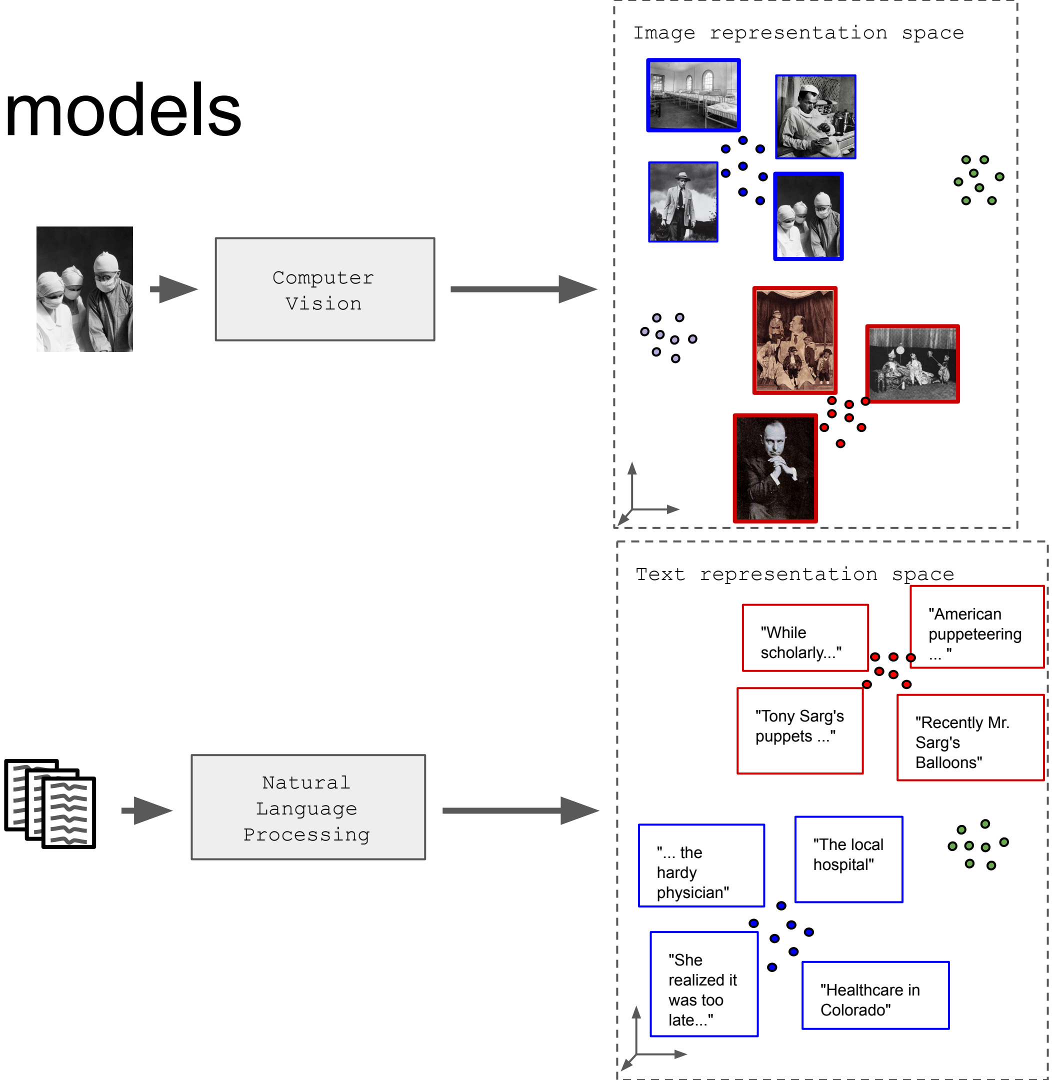
Clustering Methods

- Latent Dirichlet Allocation (LDA)
- Paragraph Vector (PV)
- Bi-term Topic Model (BTM)

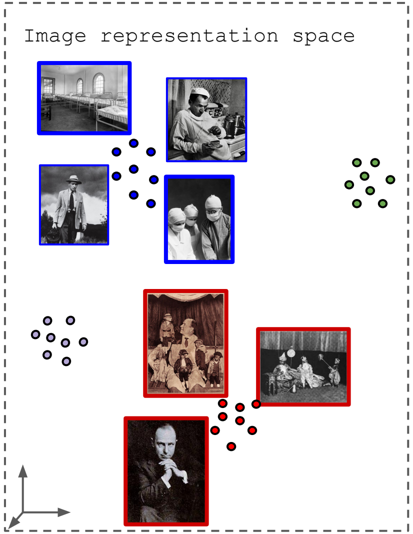
[Blei et al. 2003; Yan et al. 2013; Le and Mikolov 2014]

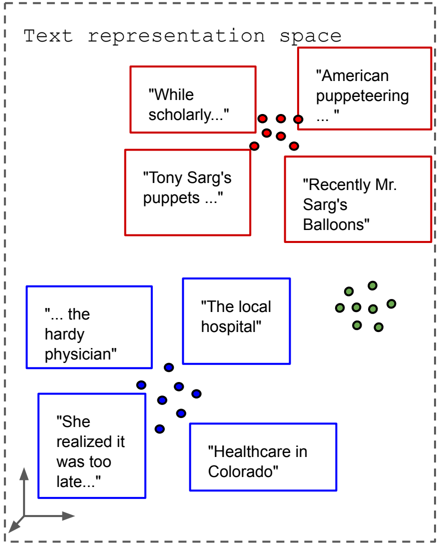
Alignment models

Alignment models

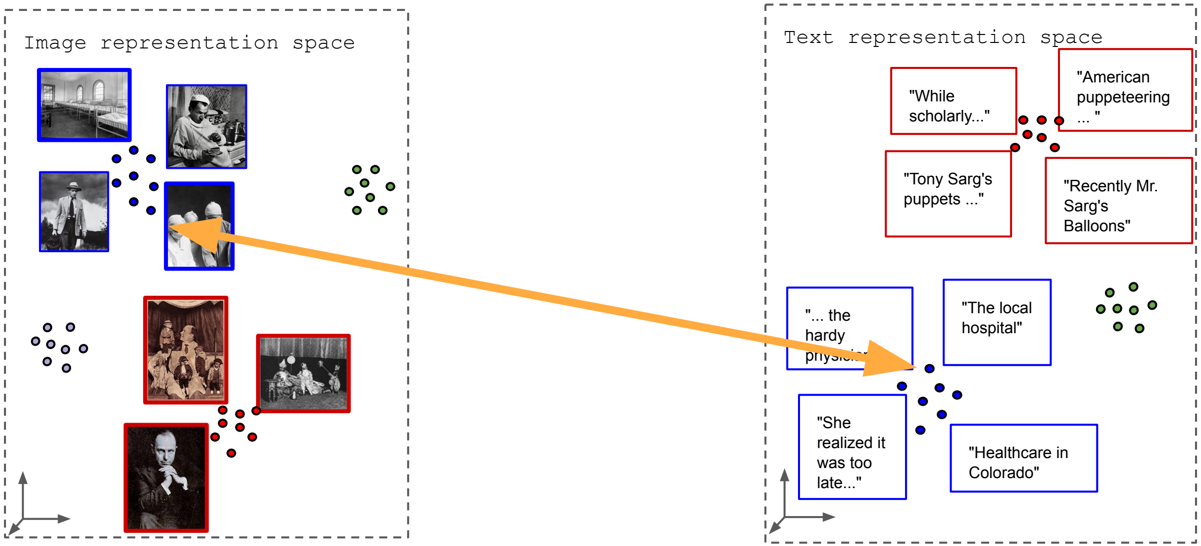


Alignment models

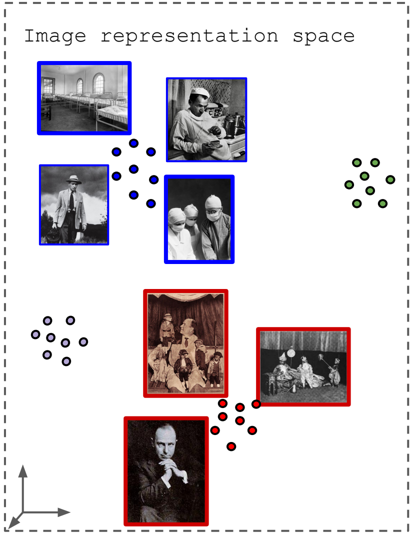


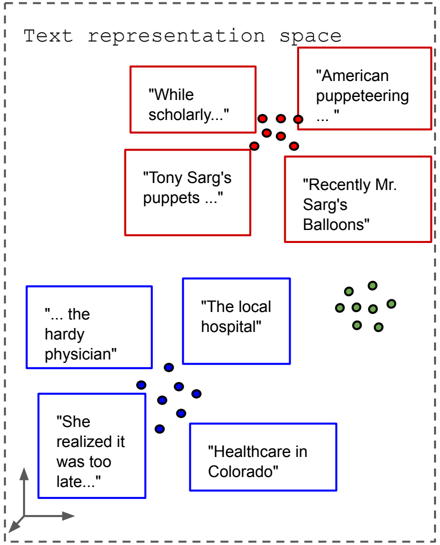


Alignment models

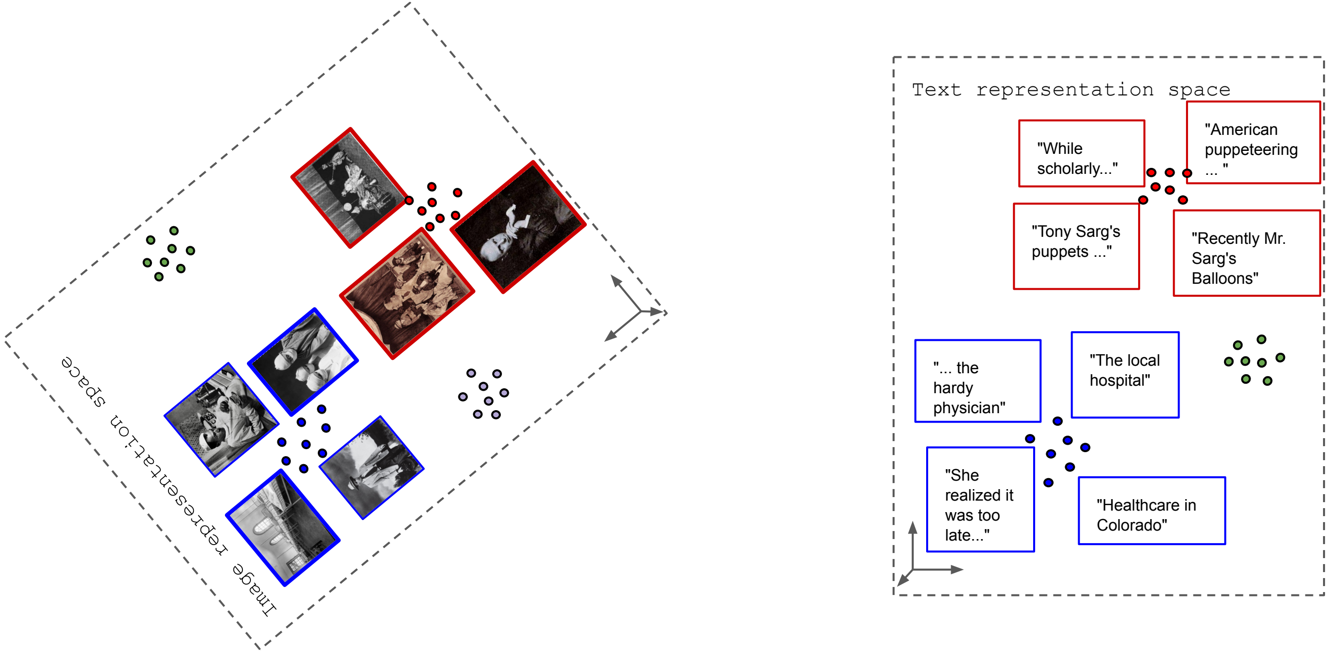


Alignment models

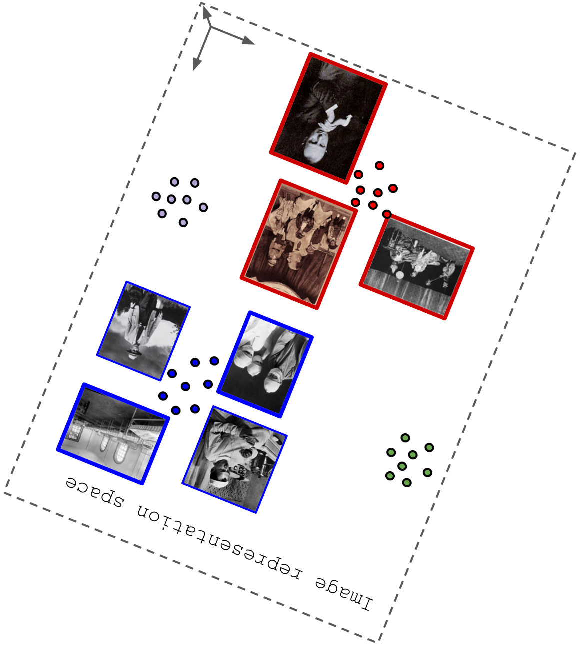


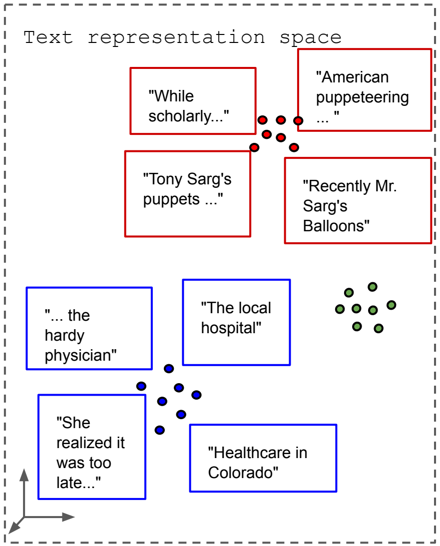


Alignment models

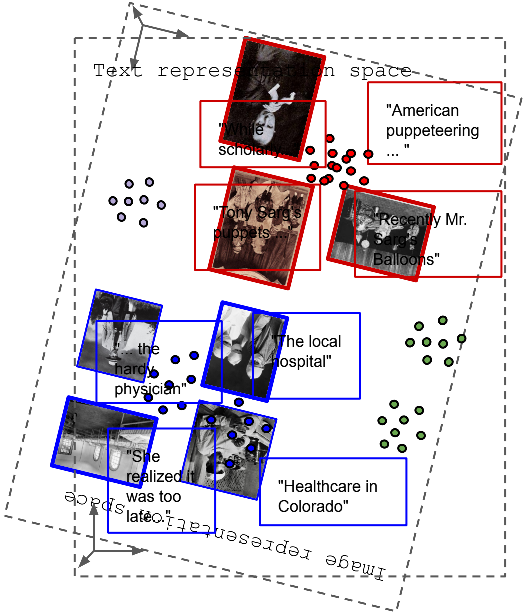


Alignment models





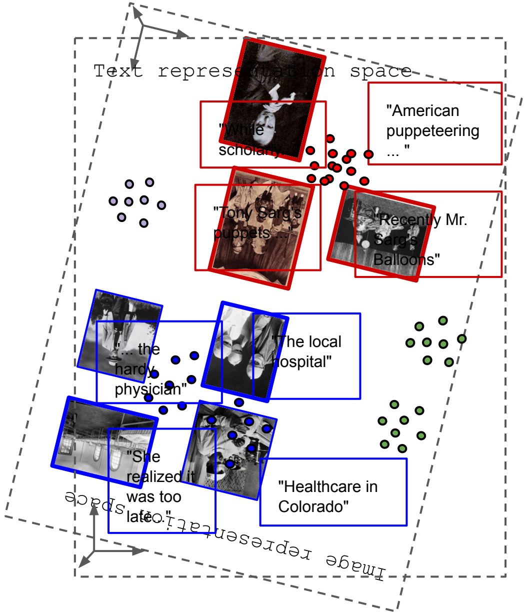
Alignment models



Nonparametric baseline (NP), Least-Squares (LS)
Negative Sampling (NS), Deep Canonical Correlation Analysis (DCCA)

[Hodosh et al. 2013; Andrew et al. 2013; Kiros et al. 2015]

Alignment models



Nonparametric baseline (NP), Least-Squares (LS)
Negative Sampling (NS), Deep Canonical Correlation Analysis (DCCA)

[Hodosh et al. 2013; Andrew et al. 2013; Kiros et al. 2015]

NASA

NASA

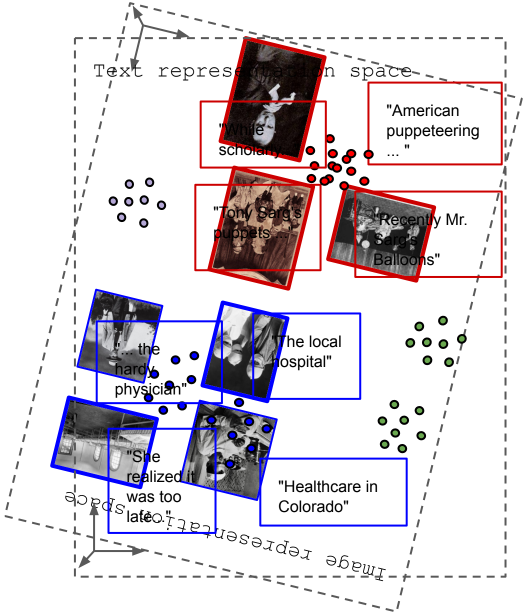
NASA

Democrats,
Dreamers,
Healthcare

UK Parliament,
Brexit,
EU

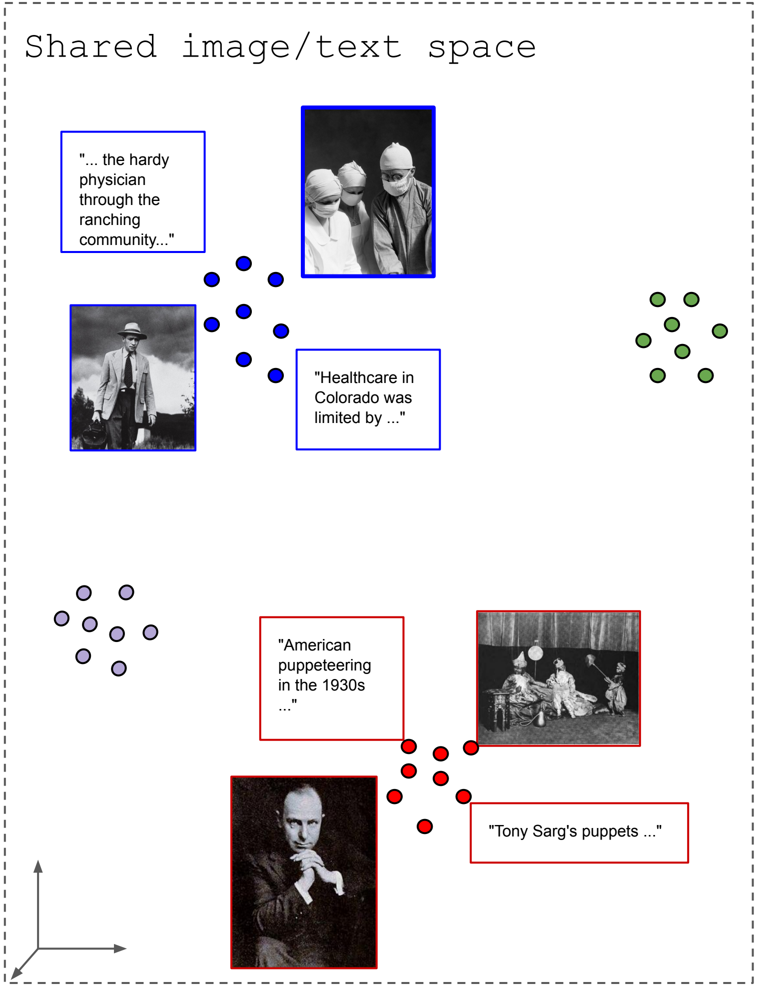
Republicans,
The Wall,
Handouts

Alignment models

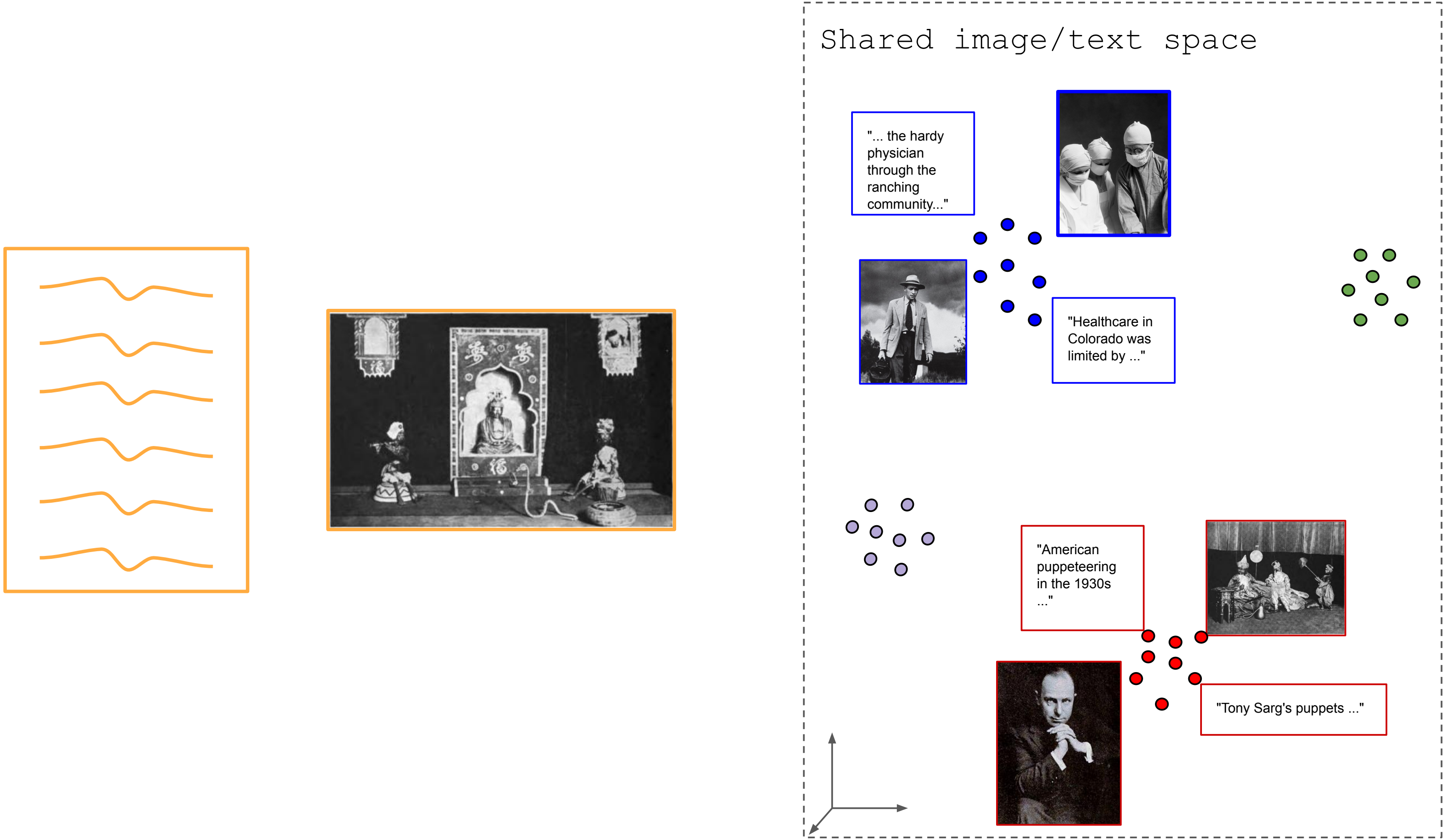


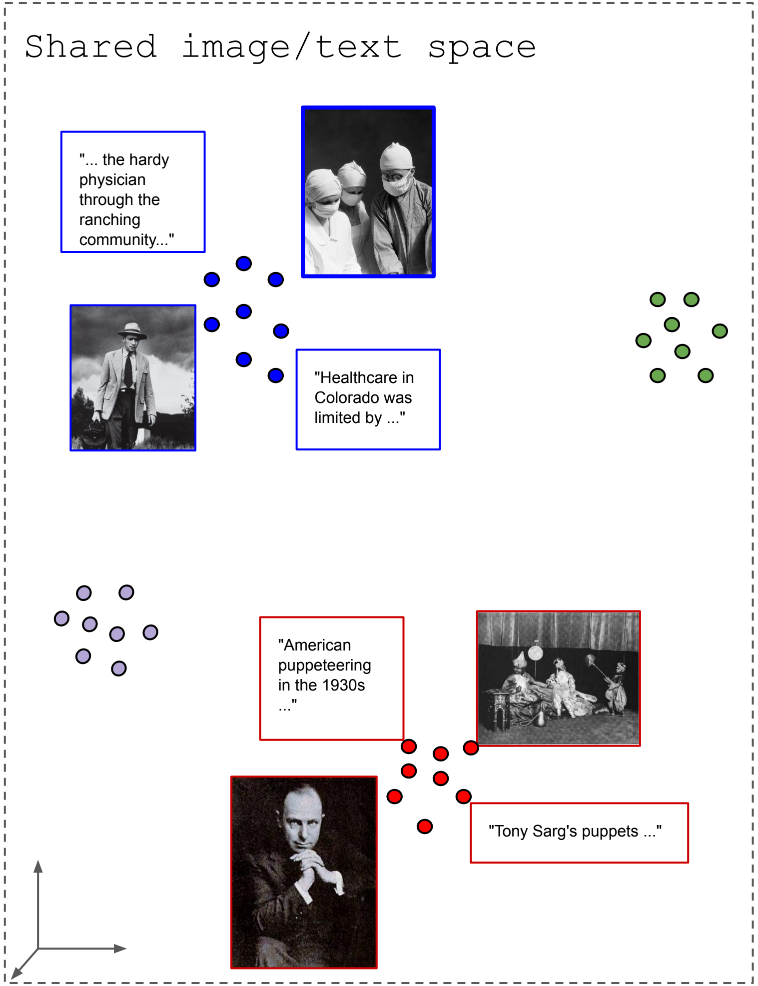
Nonparametric baseline (NP), Least-Squares (LS)
Negative Sampling (NS), Deep Canonical Correlation Analysis (DCCA)

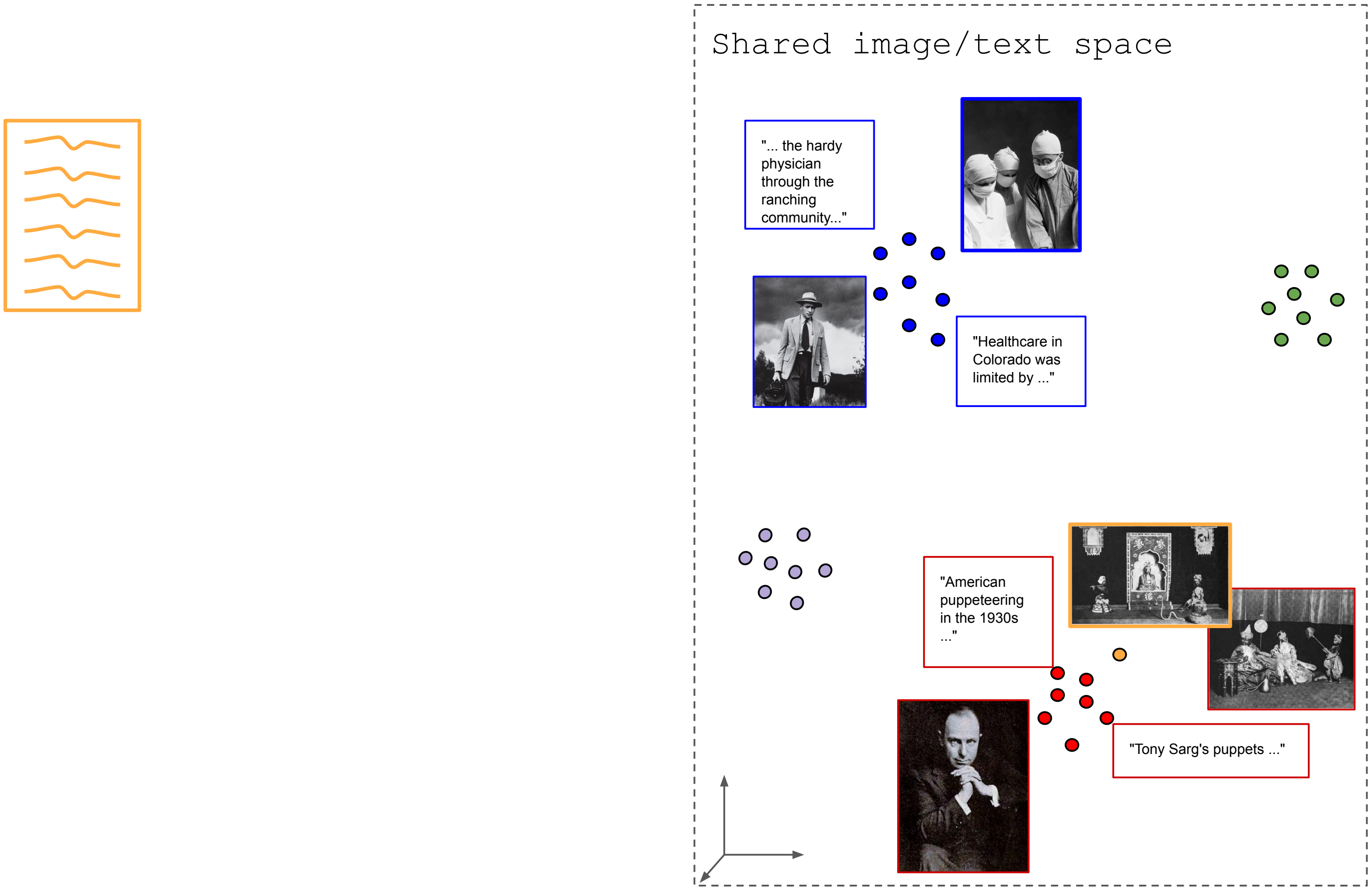
[Hodosh et al. 2013; Andrew et al. 2013; Kiros et al. 2015]



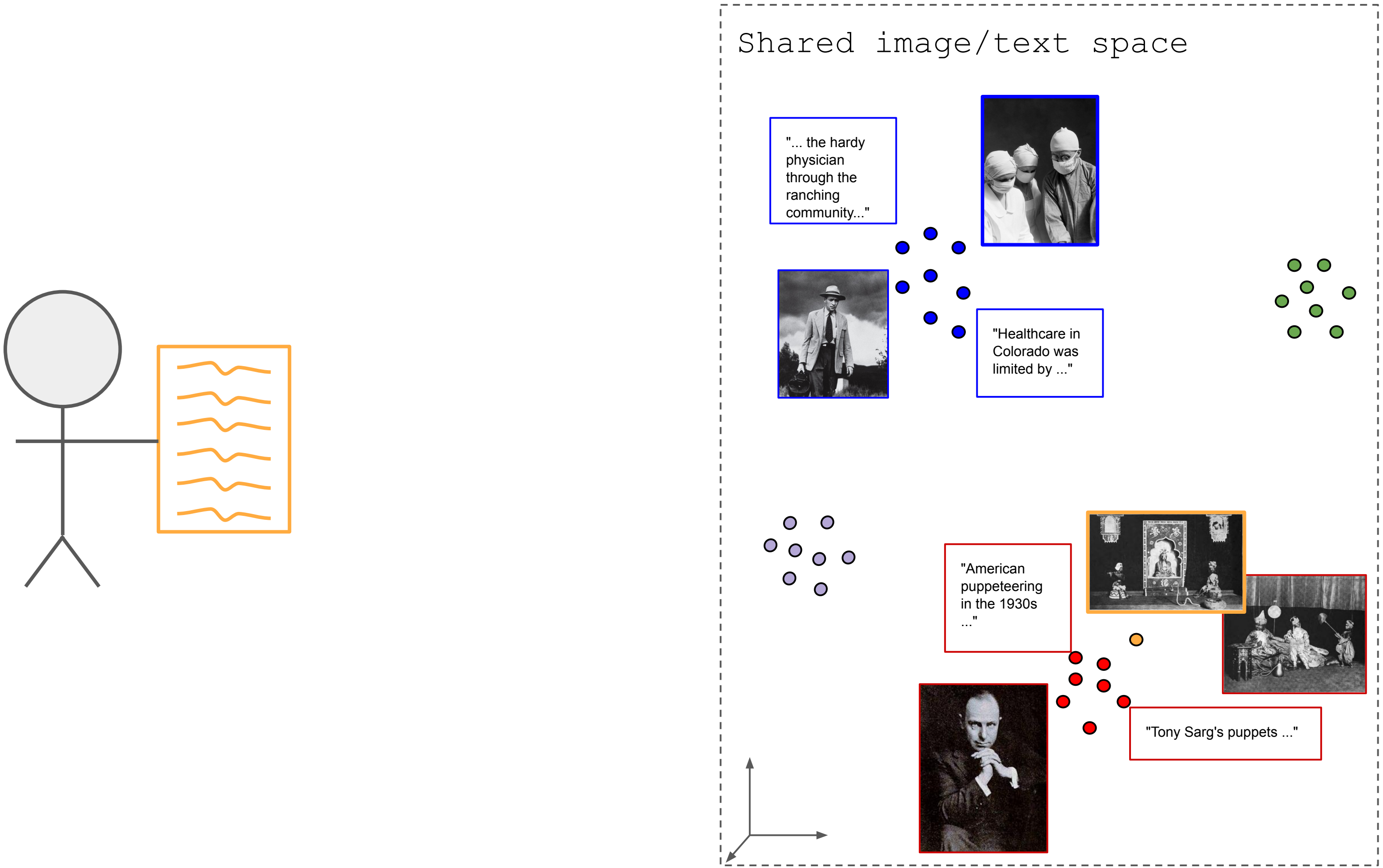
Shared image/text space
"... the hardy physician through the ranching community..."
"Healthcare in Colorado was limited by ..."
"American puppeteering in the 1930s ..."
"Tony Sarg's puppets ..."

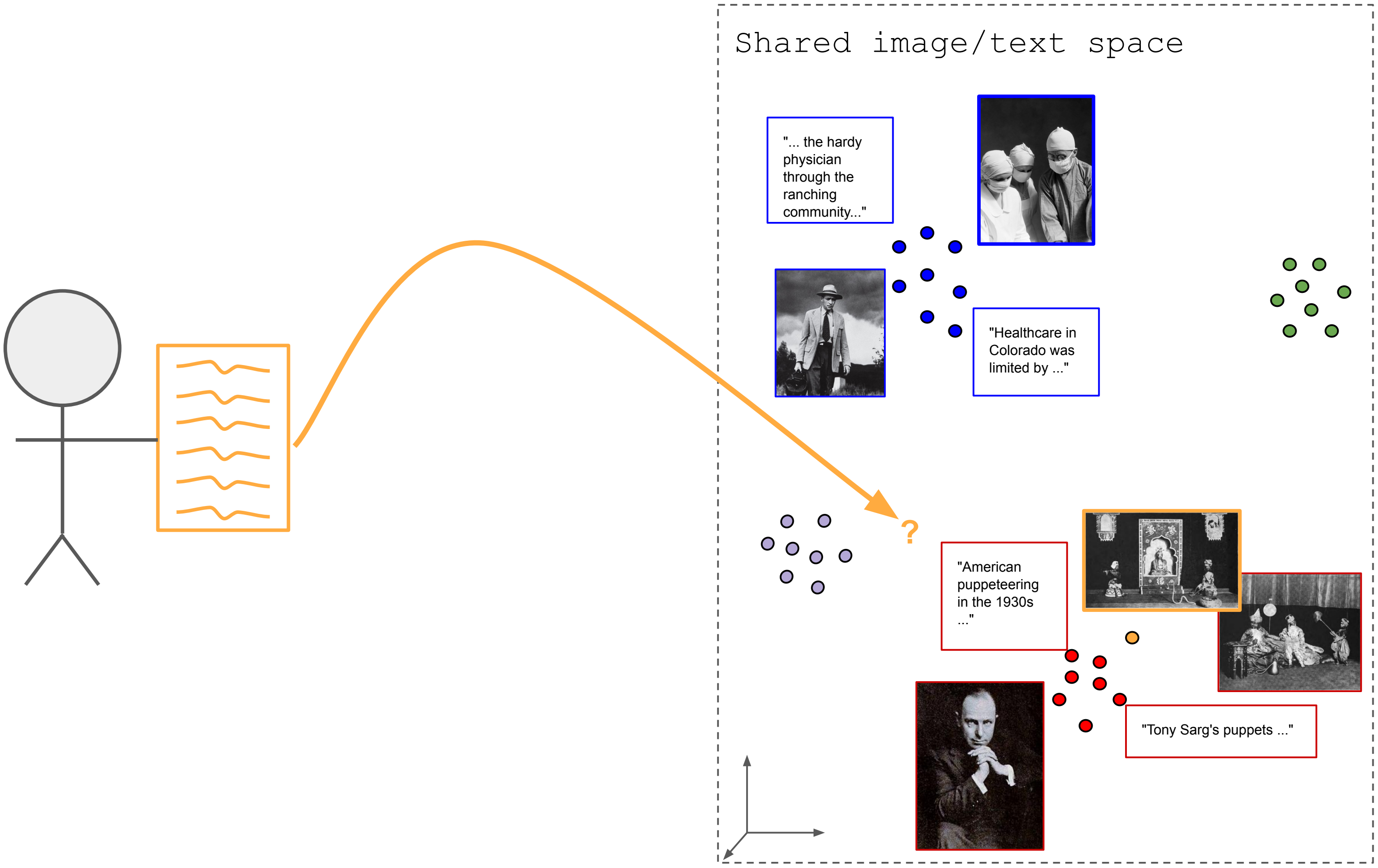

Shared image/text space
"... the hardy physician through the ranching community..."
"Healthcare in Colorado was limited by ..."
"American puppeteering in the 1930s ..."
"Tony Sarg's puppets ..."

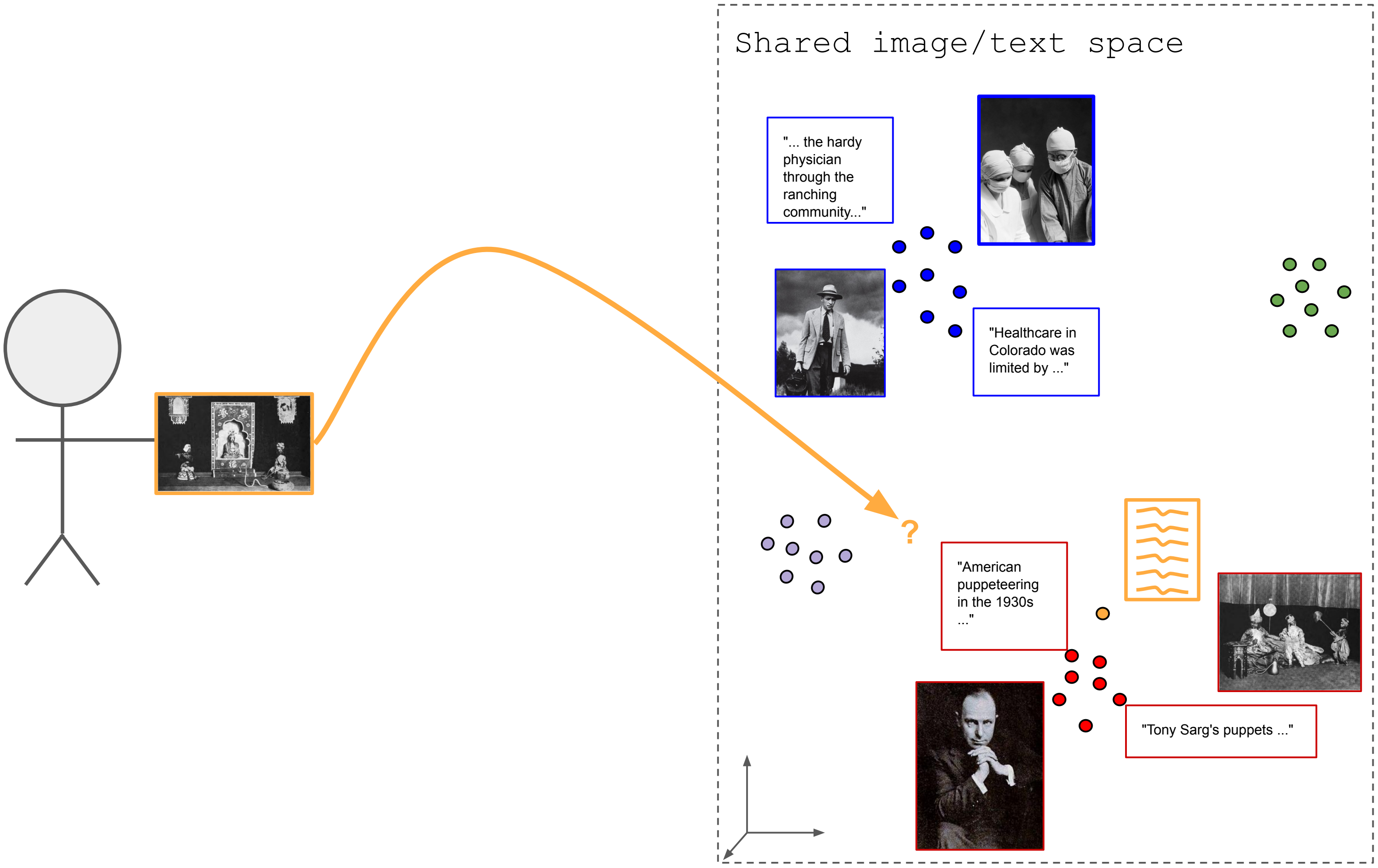

Shared image/text space
"... the hardy physician through the ranching community..."
"Healthcare in Colorado was limited by ..."
"American puppeteering in the 1930s ..."
"Tony Sarg's puppets ..."



Shared image/text space
"... the hardy physician through the ranching community..."
"Healthcare in Colorado was limited by ..."
"American puppeteering in the 1930s ..."
"Tony Sarg's puppets ..."


Shared image/text space
"... the hardy physician through the ranching community..."
"Healthcare in Colorado was limited by ..."
"American puppeteering in the 1930s ..."
"Tony Sarg's puppets ..."


Shared image/text space
"... the hardy physician through the ranching community..."
"Healthcare in Colorado was limited by ..."
?
"American puppeteering in the 1930s ..."
"Tony Sarg's puppets ..."


Shared image/text space
"... the hardy physician through the ranching community..."
"Healthcare in Colorado was limited by ..."
?
"American puppeteering in the 1930s ..."
"Tony Sarg's puppets ..."

Does it work?

Does it work?

1.0 = Random Guessing
100.0 = Perfect retrieval
Higher is better

Does it work?

1.0 = Random Guessing
100.0 = Perfect retrieval
Higher is better

Alignment Algorithms

NLP Algorithms

	NP	LS	NS	DCCA
BTM	6.7	7.3	7.2	9.5
LDA	10.2	**17.1**	13.8	**16.4**
PV	12.6	14.1	14.1	**17.8**
uni	11.0	13.2	12.4	15.6
tfidf	10.9	15.1	13.5	15.5

Does it work?

1.0 = Random Guessing
100.0 = Perfect retrieval
Higher is better

Alignment Algorithms

NLP Algorithms

	NP	LS	NS	DCCA
BTM	6.7	7.3	7.2	9.5
LDA	10.2	**17.1**	13.8	**16.4**
PV	12.6	14.1	14.1	**17.8**
uni	11.0	13.2	12.4	15.6
tfidf	10.9	15.1	13.5	15.5

Baseline for modeling modalities independently

Does it work?

1.0 = Random Guessing
100.0 = Perfect retrieval
Higher is better

Alignment Algorithms

NLP Algorithms

	NP	LS	NS	DCCA
BTM	6.7	7.3	7.2	9.5
LDA	10.2	**17.1**	13.8	**16.4**
PV	12.6	14.1	14.1	**17.8**
uni	11.0	13.2	12.4	15.6
tfidf	10.9	15.1	13.5	15.5

Does it work?

Some good examples

11.6% iron furnace acid gas fig process copper air heat fire
7.6% made end side long iron wood hand work round cut
4.0% gold ore mill water stamp solution battery silver tailings ores
3.7% steam fig cylinder engine shaft water pressure pump valve inch
3.5% great time present part form generally large found number small

-These furnaces should be deeper than the prece

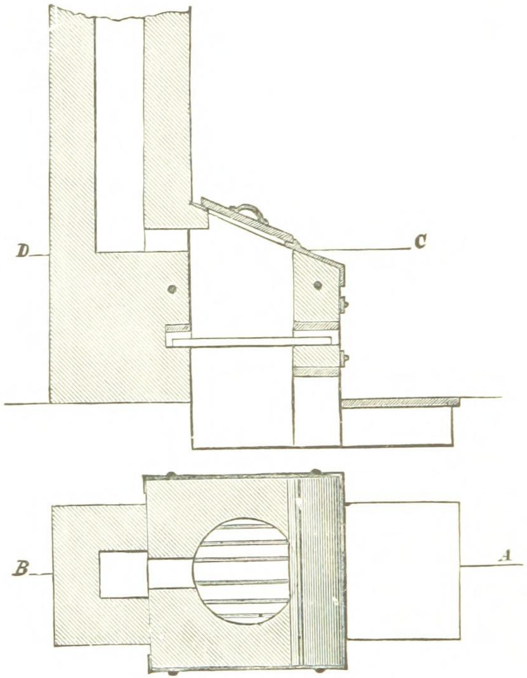
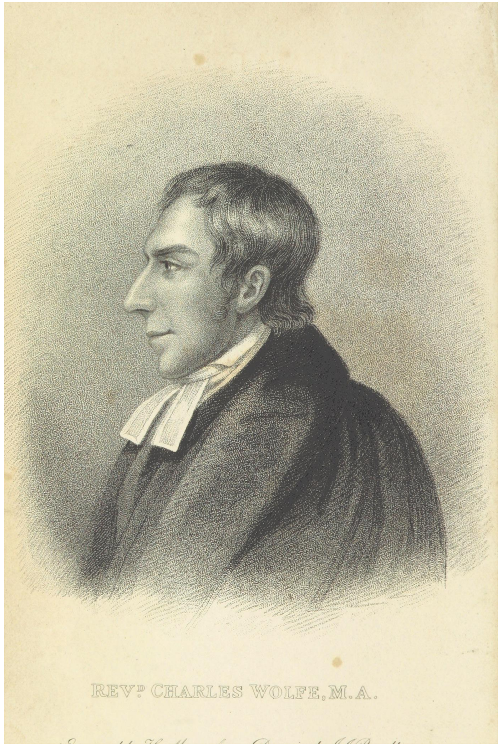


FIG. 5.
Scale ½-inch to the foot.

Predicted:
11.6% iron furnace acid gas fig process copper air heat fire
7.6% made end side long iron wood hand work round cut
4.0% gold ore mill water stamp solution battery silver tailings ores
3.7% steam fig cylinder engine shaft water pressure pump valve inch
3.5% great time present part form generally large found number small

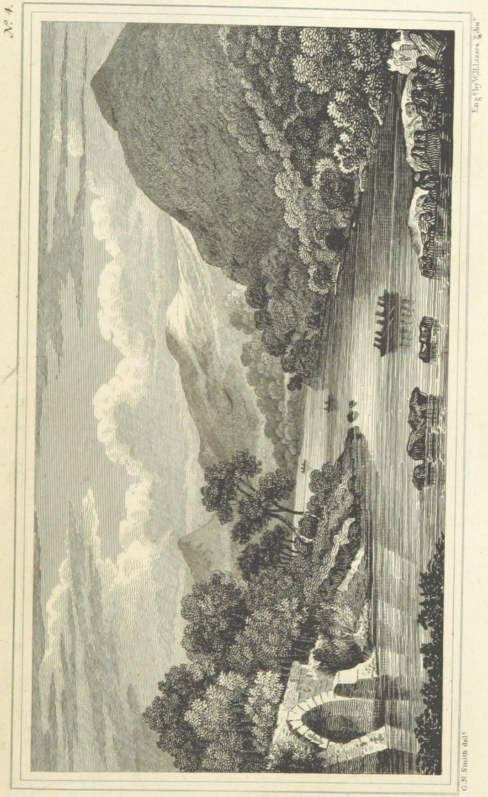
True:
49.1% iron furnace acid gas fig process copper air heat fire
12.2% made end side long iron wood hand work round cut
11.5% gold ore mill water stamp solution battery silver tailings ores
7.6% made point line plan case general found work position time
4.0% steam fig cylinder engine shaft water pressure pump valve inch



Predicted:
11.1% london street author printed vol edition illustrations john volumes reserved
5.2% history work published book author edition account present volume historical
3.7% life men man great good character day world fact nature
3.7% time made work make great means place found con purpose
3.3% poet poems works poem poetry poets english edited poetical life

True:
21.3% london street author printed vol edition illustrations john volumes reserved
12.5% history work published book author edition account present volume historical
12.5% john esq william sir thomas george james rev robert henry
7.1% time made work make great means place found con purpose
6.8% found stone stones remains bones ancient discovered age roman relics

No. 4.



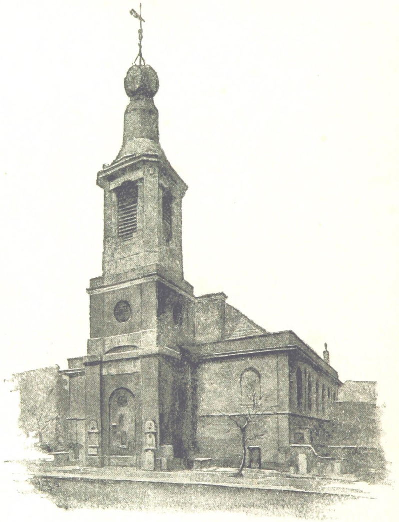
THE MIDDLE LAKE FROM BRICKEEN ISLAND.

Predicted:
5.2% view beautiful scenery place picturesque fine village town situated beauty
4.5% rock rocks mountain feet wild scene deep water waters mountains
3.9% thy eye nature till ring mind oft vain tis pride
3.7% london street author printed vol edition illustrations john volumes reserved
3.7% ofthe part con account tion country pro present general state

True:
25.6% view beautiful scenery place picturesque fine village town situated beauty
10.7% rock rocks mountain feet wild scene deep water waters mountains
7.1% trees village green hill country long wood hills road forest
5.7% great time present part form generally large found number small
5.2% ofthe part con account tion country pro present general state

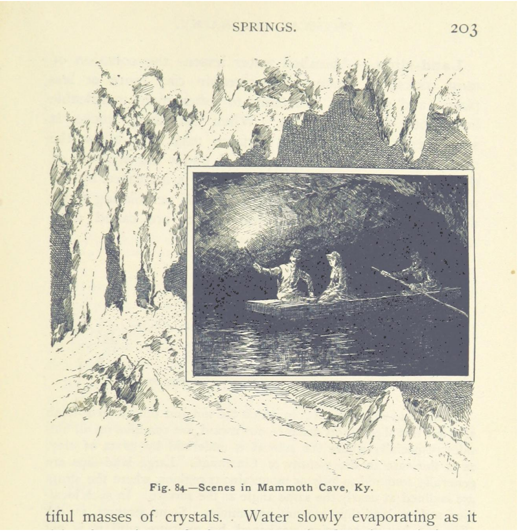
Does it work?

Some not-so-good examples



Predicted:
4.8% church south nave north tower window side windows chancel east
3.4% sweet love song day summer flowers heart fair bright green
3.3% building marble great columns palace front beautiful architecture feet centre
2.8% london street author printed vol edition illustrations john volumes reserved
2.6% church rev chapel minister congregation pastor sunday meeting worship school

True:
37.8% history work published book author edition account present volume historical
12.1% london street house lane westminster square great inn paul thames
9.3% time made work make great means place found con purpose
7.1% oxford court windsor hampton thames college richmond queen house surrey
6.2% great time present part form generally large found number small



Predicted:
2.4% life men man great good character day world fact nature
2.3% thy eye nature till ring mind oft vain tis pride
2.2% time made work make great means place found con purpose
2.2% rocks beds limestone strata sandstone clay rock geological geology formation
2.1% rock rocks mountain feet wild scene deep water waters mountains

True:
38.9% water sand waters feet spring stream springs great surface salt
11.2% rocks beds limestone strata sandstone clay rock geological geology formatio
7.9% iron furnace acid gas fig process copper air heat fire
7.3% great time present part form generally large found number small
6.7% cave rock caves cavern entrance caverns roof grotto floor rocks

Predicted:
5.2% view beautiful scenery place picturesque fine village town situated beauty
2.8% rock rocks mountain feet wild scene deep water waters mountains
2.1% time made work make great means place found con purpose
2.1% great time present part form generally large found number small
2.0% life men man great good character day world fact nature

True:
30.0% york hull yorkshire ripon bolton abbey scarborough hall north leeds
9.7% valley mountain mountains hills miles feet range road plain great
6.6% view beautiful scenery place picturesque fine village town situated beauty
2.8% town city houses inhabitants miles place streets towns built large
2.8% power state conduct act manner death fate mind length evil

WIKIPEDIA
The Free Encyclopedia

COCO
Common Objects in Context

flickr

WIKIPEDIA
The Free Encyclopedia

COCO
Common Objects in Context

flickr

Ice hockey

From Wikipedia, the free encyclopedia

For other uses, see Ice hockey (disambiguation).

Ice hockey is a contact team sport played on ice, usually in a rink, in which two teams of skaters use their sticks to shoot a vulcanized rubber puck into their opponent's net to score points. The sport is known to be fast-paced and

Ice hockey

The Toronto Maple Leafs (white) defend their goal against the Washington Capitals (red) during the 2016–17 NHL season.

Highest governing body	International Ice Hockey Federation
First played	19th century Canada

WIKIPEDIA
The Free Encyclopedia

COCO
Common Objects in Context

flickr

Ice hockey

From Wikipedia, the free encyclopedia

For other uses, see Ice hockey (disambiguation).

Ice hockey is a contact team sport played on ice, usually in a rink, in which two teams of skaters use their sticks to shoot a vulcanized rubber puck into their opponent's net to score points. The sport is known to be fast-paced and

Ice hockey

The Toronto Maple Leafs (white) defend their goal against the Washington Capitals (red) during the 2016–17 NHL season.

Highest governing body	International Ice Hockey Federation
First played	19th century Canada

The man at bat readies to swing at the pitch while the umpire looks on.

WIKIPEDIA
The Free Encyclopedia

COCO
Common Objects in Context

flickr

Ice hockey

From Wikipedia, the free encyclopedia

For other uses, see Ice hockey (disambiguation).

Ice hockey is a contact team sport played on ice, usually in a rink, in which two teams of skaters use their sticks to shoot a vulcanized rubber puck into their opponent's net to score points. The sport is known to be fast-paced and

Ice hockey

The Toronto Maple Leafs (white) defend their goal against the Washington Capitals (red) during the 2016–17 NHL season.

Highest governing body	International Ice Hockey Federation
First played	19th century Canada

The man at bat readies to swing at the pitch while the umpire looks on.

beach sun warm 2017 islandday vacation motivationmonday

WIKIPEDIA
The Free Encyclopedia

COCO
Common Objects in Context

flickr

WIKIPEDIA
The Free Encyclopedia

COCO
Common Objects in Context

flickr

1.0 = Random Guessing
100.0 = Perfect retrieval
Higher is better

Recall that british library performance was ~17

1.0 = Random Guessing
100.0 = Perfect retrieval
Higher is better

Recall that british library performance was ~17

Wikipedia The Free Encyclopedia

	NP	LS	NS	DCCA
BTM	14.1	19.8	20.7	27.9
LDA	19.8	36.2	33.1	**37.0**
PV	22.0	30.8	29.4	**37.1**
uni	17.3	29.3	30.2	36.3
tfidf	18.1	35.2	33.2	**38.7**

COCO Common Objects in Context

	NP	LS	NS	DCCA
BTM	27.3	39.9	52.5	58.6
LDA	23.2	51.6	51.9	51.8
PV	14.1	28.4	25.7	33.5
uni	28.7	**74.6**	72.5	**75.0**
tfidf	32.9	74.0	74.1	**74.9**

flickr

	NP	LS	NS	DCCA
BTM	23.9	19.1	31.0	32.4
LDA	18.4	32.2	34.4	34.7
PV	13.9	21.3	20.0	26.6
uni	34.7	**62.5**	**62.0**	59.6
tfidf	35.1	61.6	**63.9**	60.2

1.0 = Random Guessing
100.0 = Perfect retrieval
Higher is better

BRITISH LIBRARY

vs.

flickr

COCO
Common Objects in Context

WIKIPEDIA
The Free Encyclopedia

- Similarities between text data and image data
- Why should we want to model text and images jointly?
- Computer vision and why digital libraries
- The dataset/experiments
- **Are concrete things easier to learn?**

Easy to learn concept vs hard to learn concept?

Easy to learn concept vs hard to learn concept?

"Performance advantages of [multi-modal approaches] over language-only models have been clearly established when models are required to learn concrete noun concepts."

- *Hill and Korhonen 2014*

Easy to learn concept vs hard to learn concept?

*"Performance advantages of [multi-modal approaches] over language-only models have been clearly established when models are required to learn **concrete noun concepts**."*

- *Hill and Korhonen 2014*

The cat is in the grass.

This cat is enjoying the sun.

The cat is in the grass.

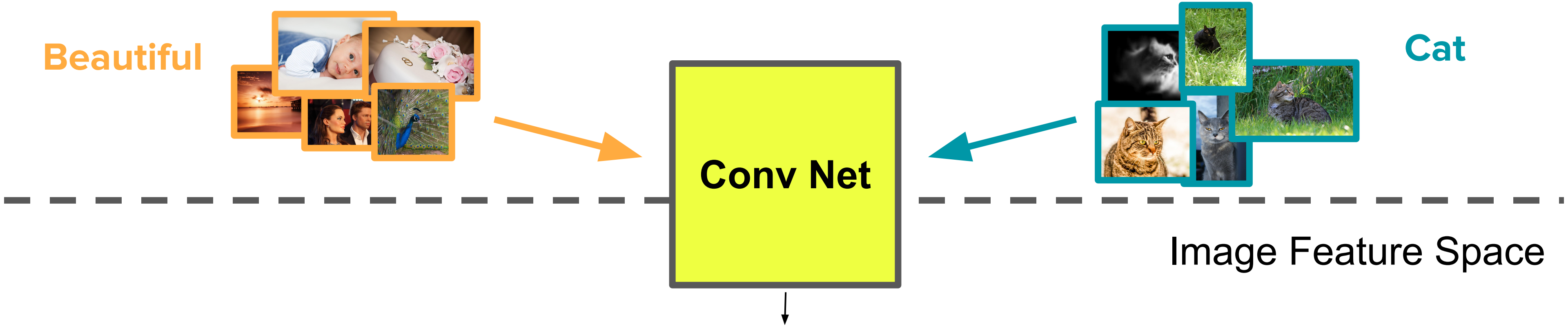
This cat is enjoying the sun.

This is a beautiful baby.

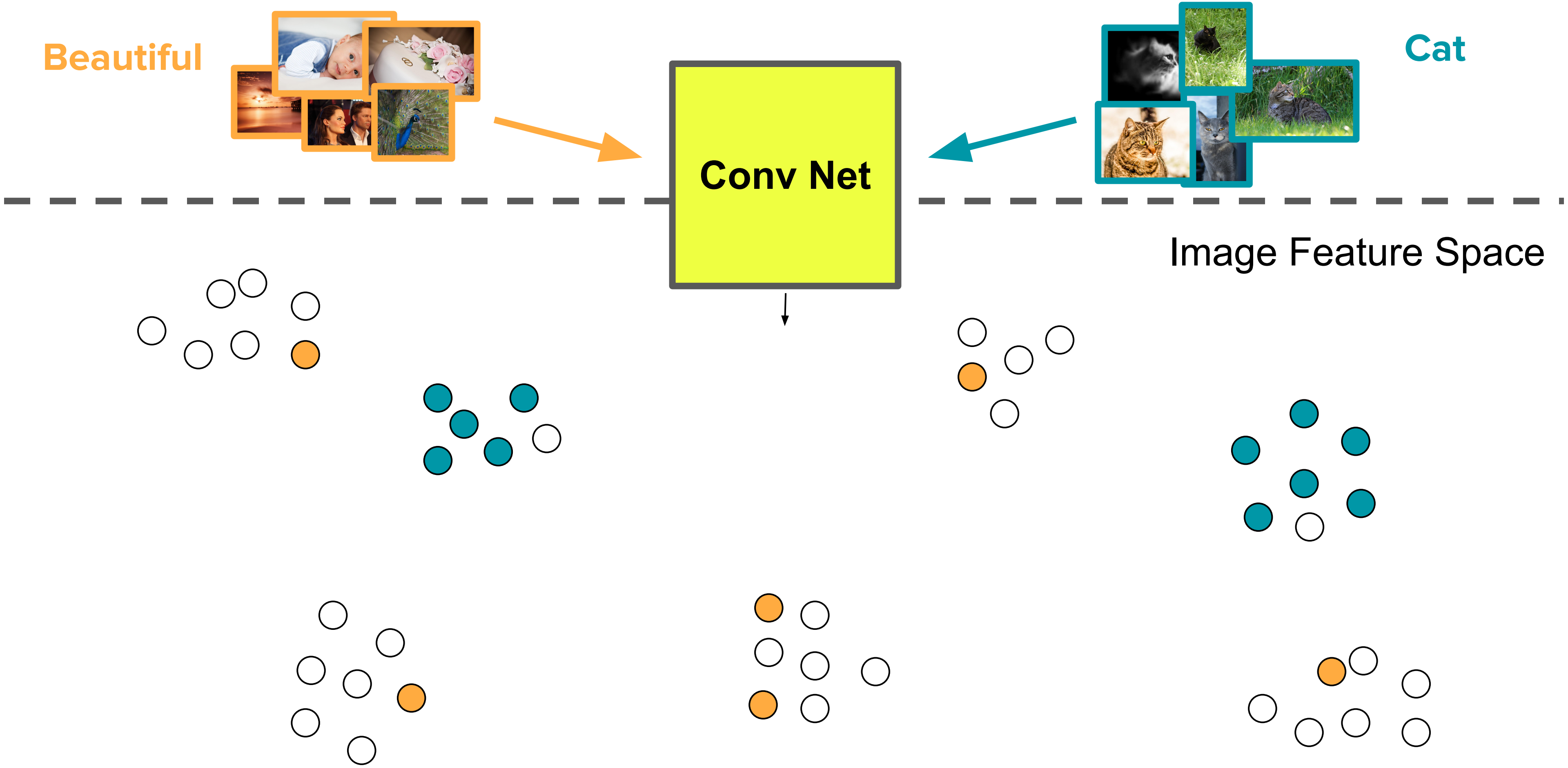
The sunset is beautiful.

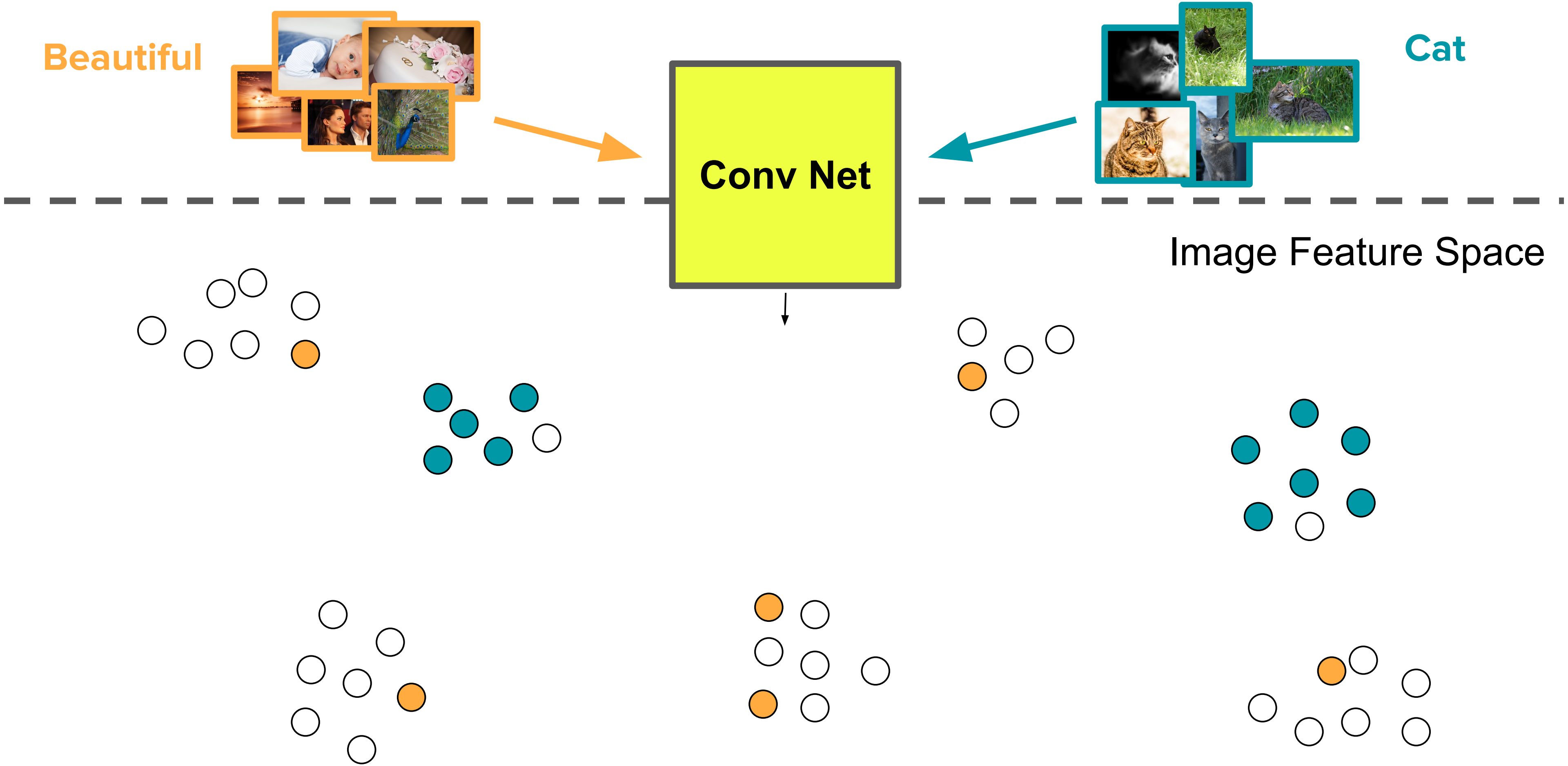
Cat

Beautiful

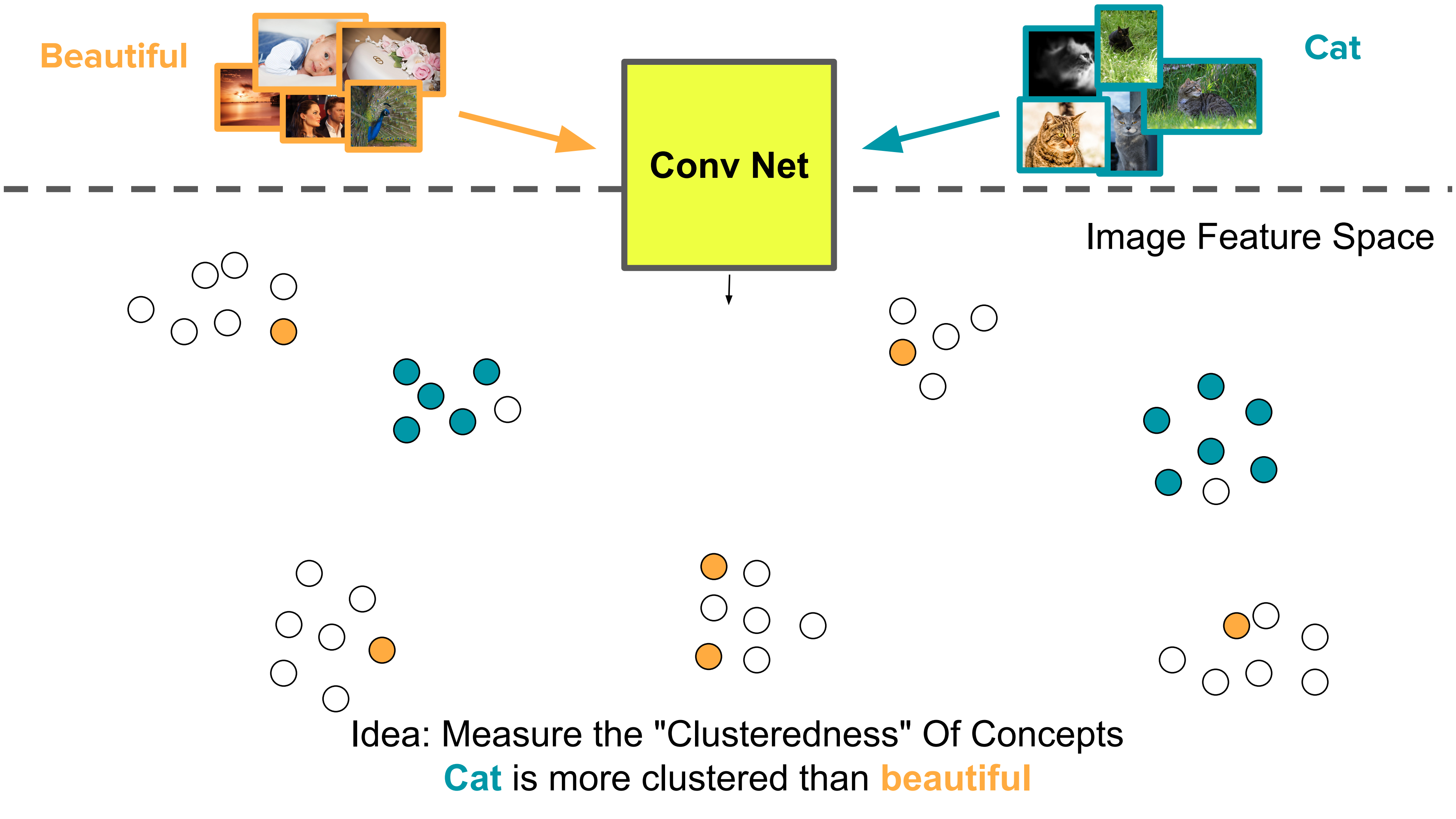


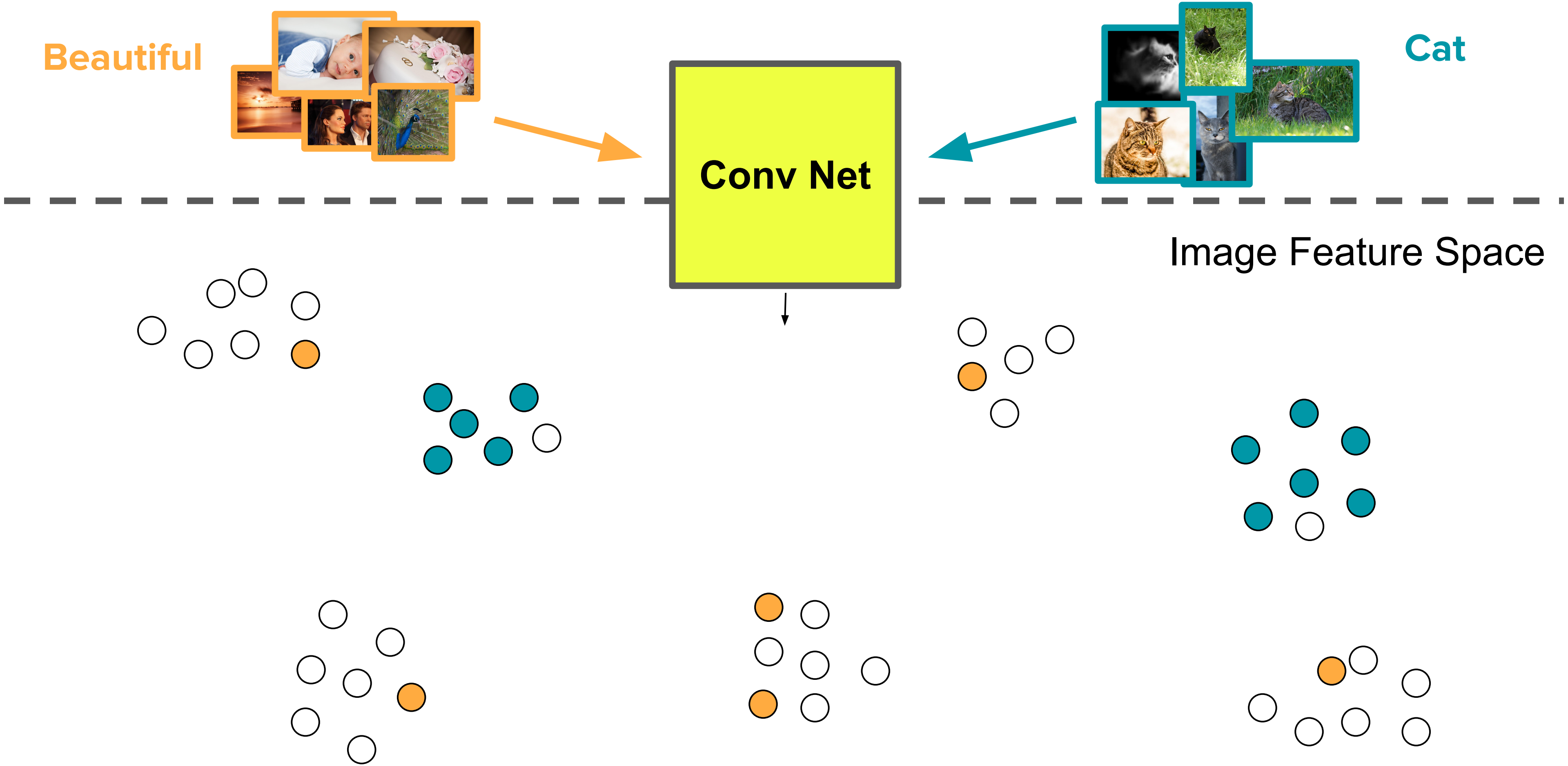
Beautiful
Cat
Conv Net
Image Feature Space


Beautiful
Cat
Conv Net
Image Feature Space



Idea: Measure the "Clusteredness" Of Concepts


Beautiful
Cat
Conv Net
Image Feature Space
Idea: Measure the "Clusteredness" Of Concepts
Cat is more clustered than beautiful



Idea: Measure the "Clusteredness" Of Concepts

Cat is more *concrete* than beautiful

COCO Results

COCO Results

The man at bat readies to swing at the pitch while the umpire looks on.

COCO Results

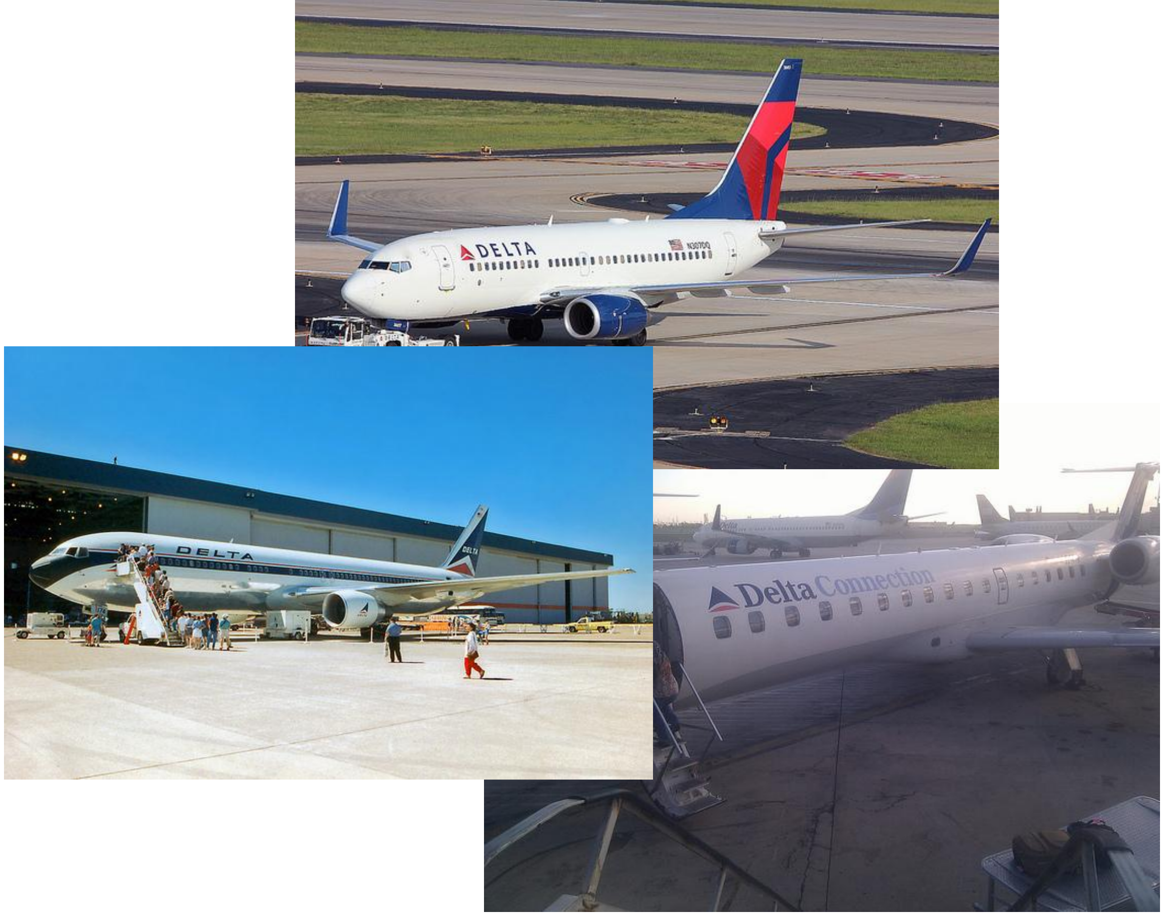
Most concrete

wok	315.595
hummingbird	291.804
vane	290.037
racer	269.043
grizzly	229.274
equestrian	219.894
taxiing	205.410
unripe	201.733
siamese	199.024
delta	195.618
kiteboarding	192.459
airways	183.971
compartments	182.015
burners	180.553
stocked	177.472
spire	177.396
tulips	173.850
ben	171.936

COCO Results

Most concrete

wok	315.595
hummingbird	291.804
vane	290.037
racer	269.043
grizzly	229.274
equestrian	219.894
taxiing	205.410
unripe	201.733
siamese	199.024
delta	**195.618**
kiteboarding	192.459
airways	183.971
compartments	182.015
burners	180.553
stocked	177.472
spire	177.396
tulips	173.850
ben	171.936



COCO Results

Most concrete

wok	315.595
hummingbird	291.804
vane	290.037
racer	269.043
grizzly	229.274
equestrian	219.894
taxiing	205.410
unripe	**201.733**
siamese	199.024
delta	195.618
kiteboarding	192.459
airways	183.971
compartments	182.015
burners	180.553
stocked	177.472
spire	177.396
tulips	173.850
ben	171.936

COCO Results

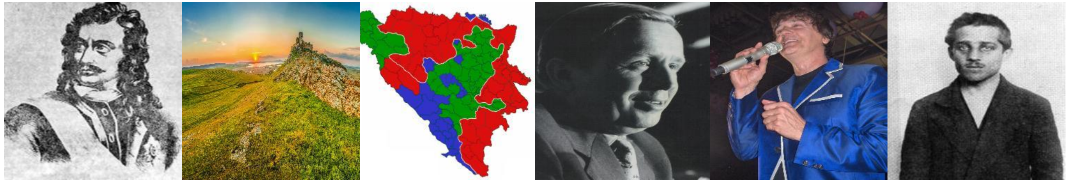
Most concrete

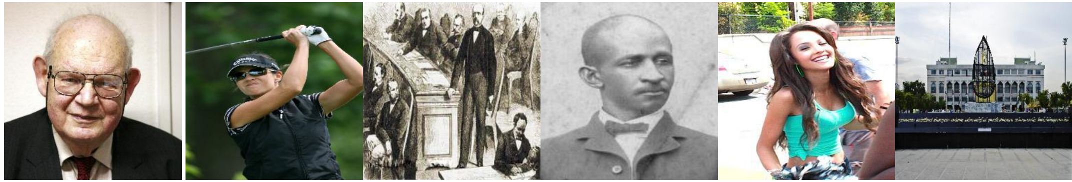
wok	315.595
hummingbird	291.804
vane	290.037
racer	269.043
grizzly	229.274
equestrian	219.894
taxiing	205.410
unripe	201.733
siamese	199.024
delta	195.618
kiteboarding	192.459
airways	183.971
compartments	182.015
burners	180.553
stocked	177.472
spire	177.396
tulips	173.850
ben	171.936

COCO Results

Most concrete		Somewhat concrete		Not concrete	
wok	315.595	motorcycle	10.291	side	1.770
hummingbird	291.804	fun	10.267	while	1.752
vane	290.037	including	10.262	other	1.745
racer	269.043	lays	10.232	sits	1.741
grizzly	229.274	fish	10.184	for	1.730
equestrian	219.894	goes	10.161	behind	1.709
taxiing	205.410	blurry	10.147	his	1.638
unripe	201.733	helmet	10.137	as	1.637
siamese	199.024	itself	10.128	image	1.620
delta	195.618	umbrellas	10.108	holding	1.619
kiteboarding	192.459	teddy	10.060	this	1.602
airways	183.971	bar	10.055	picture	1.589
compartments	182.015	fancy	10.053	couple	1.585
burners	180.553	sticks	10.050	from	1.569
stocked	177.472	himself	10.038	large	1.568
spire	177.396	take	10.016	person	1.561
tulips	173.850	steps	10.014	looking	1.502
ben	171.936	attempting	9.986	out	1.494

Top Words for Topic	Topic Concreteness	Top Images for Topic
open round world final won lost tournament tennis match tour sets defeated win title year player doubles championship grand masters	63.9	
game games nintendo super released version mario video wii console sonic sega arcade series boy japan	4.32	

Top Words for Topic	Topic Concreteness	Top Images for Topic
portuguese brazil wine brazilian portugal rio wines paulo grape janeiro lisbon grapes region state porto made santos joo brazil's vineyards	3.59	
hungarian serbia serbian hungary romanian romania yugoslavia croatia croatian bulgarian bulgaria bosnia albanian albania	3.14	
company million business group companies corporation acquired billion sold announced company's largest owned	2.58	

Top Words for Topic	Topic Concreteness	Top Images for Topic
police fire people killed officers shot incident found reported day time died officer report death emergency shooting car injured area attack safety members	1.36	
property contract law copyright legal land patent rights act estate party case person parties common owner agreement liability	1.33	
term word common list names called form refer meaning include generally number referred considered terms	1.11	

Disclaimer: in the paper we have...

Disclaimer: in the paper we have...

... correlations with human judgements

... confirmations that concreteness is not simply measuring frequency

... fuller definitions of how concreteness is computed

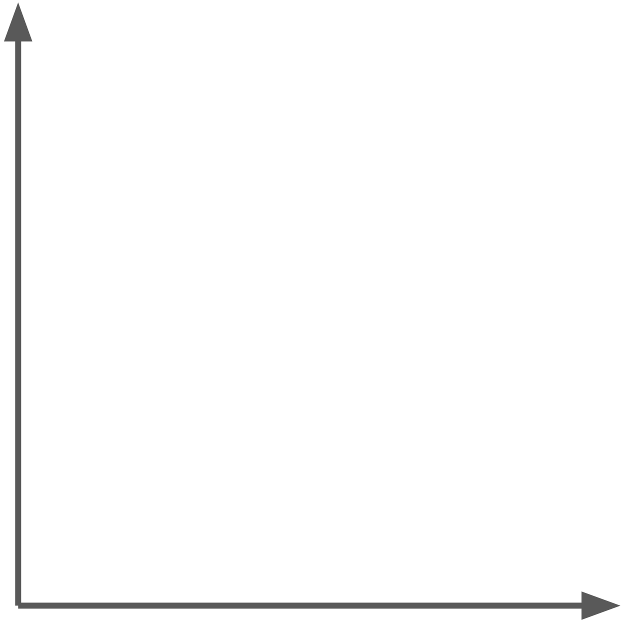
... additional experiments using a second image model

What about the British Library Set?

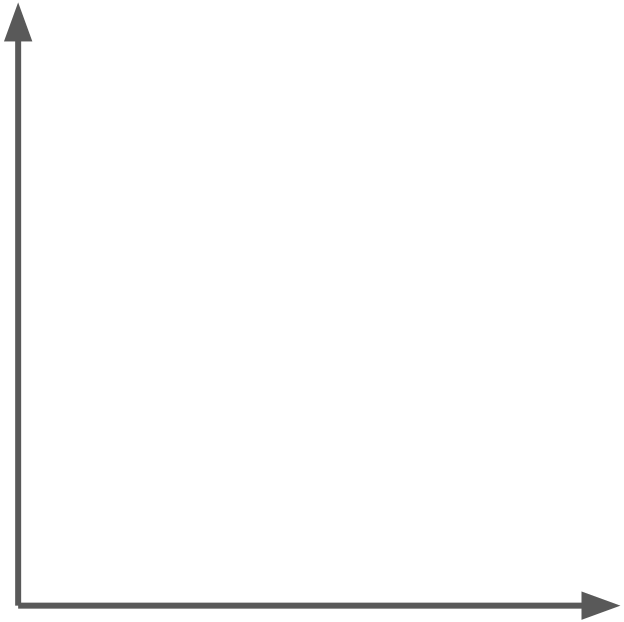
What about the British Library Set?

... harder to interpret; quite correlated with volume occurrence structure

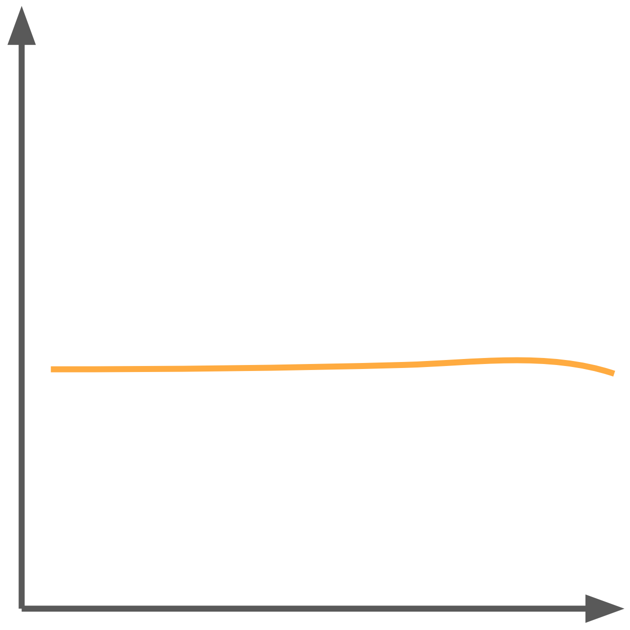
Are concrete concepts more learnable?



Are concrete concepts more learnable?

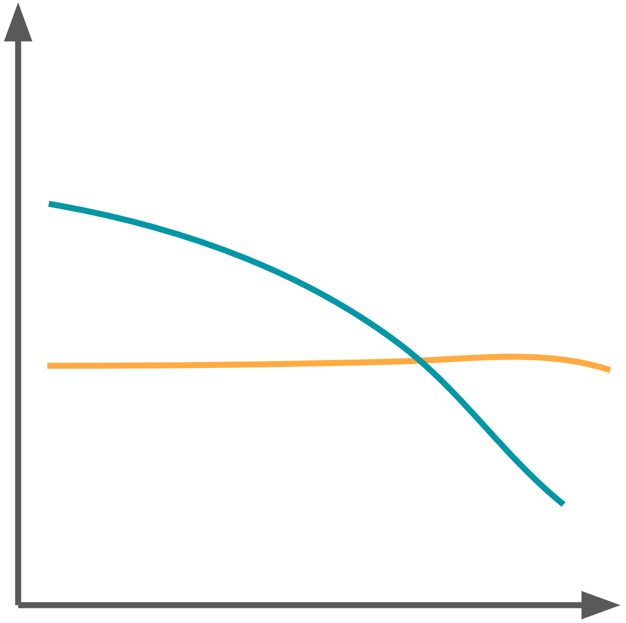


Are concrete concepts more learnable?



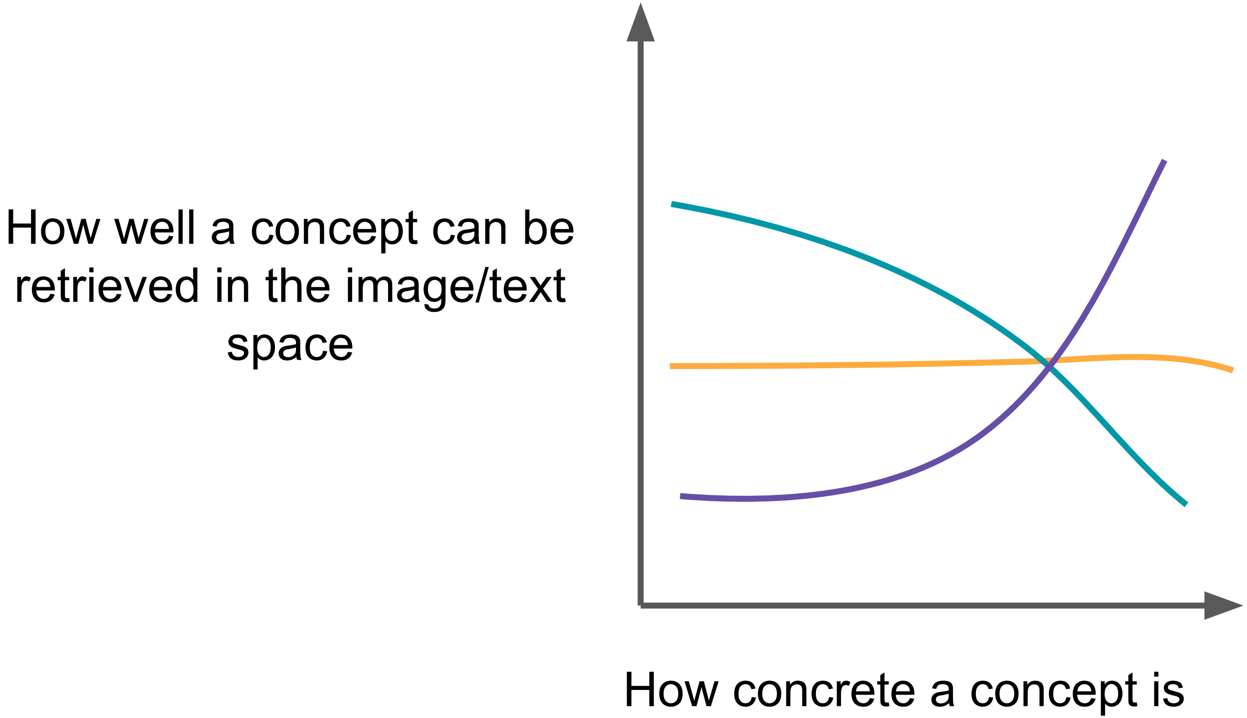
Are concrete concepts more learnable?

How well a concept can be retrieved in the image/text space

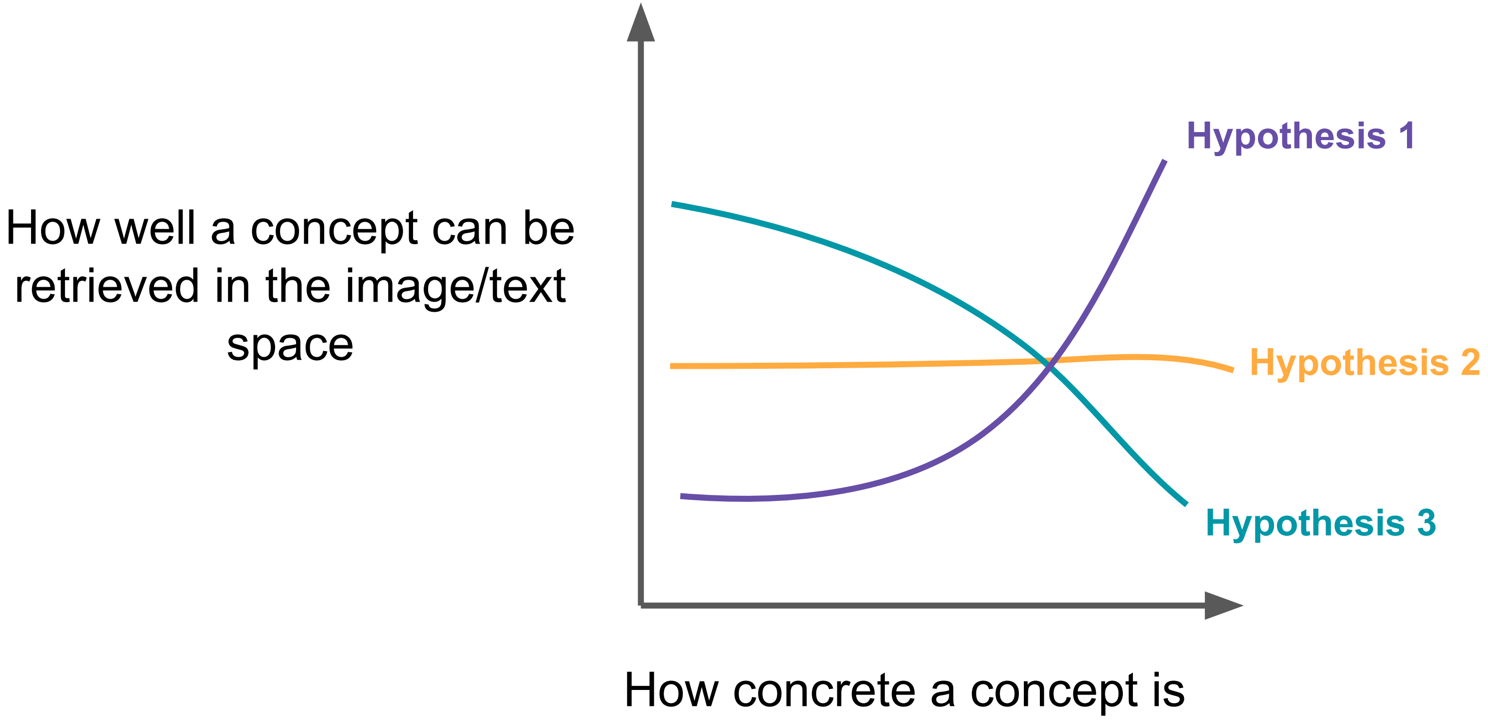


How concrete a concept is

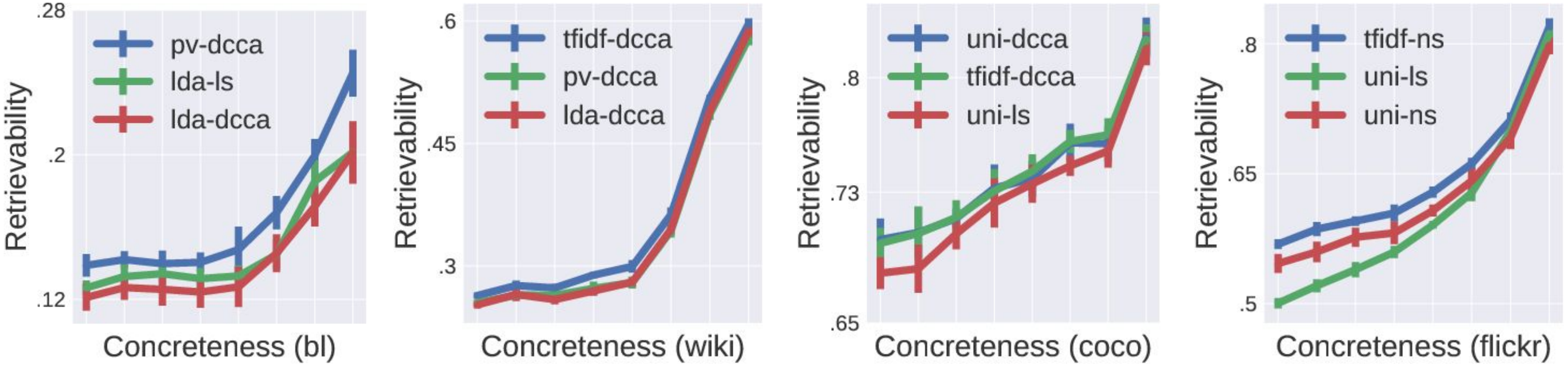
Are concrete concepts more learnable?



Are concrete concepts more learnable?



Are concrete concepts more learnable?



- Similarities between text data and image data
- Why should we want to model text and images jointly?
- Computer vision and why digital libraries
- The dataset/experiments
- Are concrete things easier to learn?

Three takeaways:

Three takeaways:

1. Computer vision and large image sets are increasingly available for digital humanists.

Three takeaways:

1. Computer vision and large image sets are increasingly available for digital humanists.

NASA

NASA

NASA

2. Multimodal modeling can be advantageous, and enable new types of search.

Three takeaways:

1. Computer vision and large image sets are increasingly available for digital humanists.

NASA

NASA

NASA

2. Multimodal modeling can be advantageous, and enable new types of search.

3. Some concepts are less concrete than others, and those are generally more difficult to learn.

Cat

Beautiful