Where Are We?

Where Are We?

The Realm of
MIDDLE EARTH
Miles
50 100 150 200
N
The Lost Realm of
ARNOR
ERIADOR
Carn Dûm
Here was of old the Witch-realm of Angmar
Ettenmoors
Fornost
Weather Hills
Weathertop (Amon Sûl)
RHUDAUR
Last Bridge
Bree
East-West Road
South Downs
CARDOLAN
THE SHIRE
Sarn Ford
Hobbiton
Nenuial Lake Evendim
North Downs
ERED LUIN
LINDON
Forlindon
Gulf of Lhûn
Grey Havens
Tower Hills
Harlond
Harlindon
Blue Mountains
Baranduin
MINHIRIATH
Tharbad
Nîn-in-Eilph (Swanfleet)
Glanduin
Gwathlo (Greyflood)
ENEDWAITH
Lond Daer
Dunland
Isengard
North-South Road
Fords of Isen
Angren (Isen)
Adorn
Helm's Deep
WESTFOLD
FANGORN
ERED NIMRAIS
Pinnath Gelin
Anfalas (Langstrand)
Andrast (Ras Morthil)

Where Are We?

We've covered supervised learning: the situation in which we have labeled data.

Where Are We?

We've covered supervised learning: the situation in which we have labeled data.

Now begin to think about documents that are not labelled but within which we expect some important, latent structure.

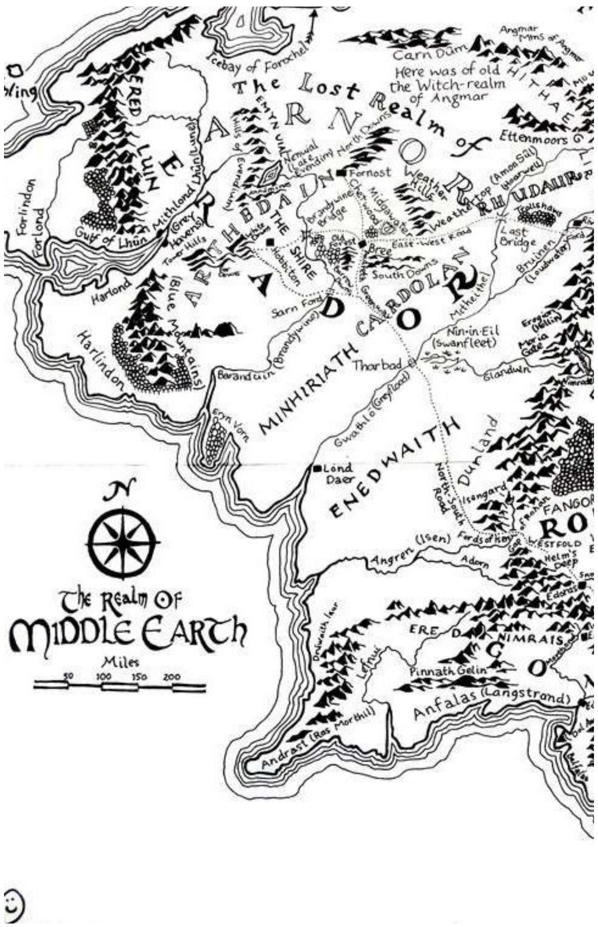
Where Are We?

We've covered supervised learning: the situation in which we have labeled data.

Now begin to think about documents that are not labelled but within which we expect some important, latent structure.

cover some fundamental techniques for accessing that structure

Where Are We?



We've covered supervised learning: the situation in which we have labeled data.

Now begin to think about documents that are not labelled but within which we expect some important, latent structure.

cover some fundamental techniques for accessing that structure

and demonstrate challenges that emerge in interpreting the results.

Terminology

Terminology

Unsupervised techniques:

Terminology

Unsupervised techniques: learning (hidden or latent) structure in unlabeled data.

e.g. PCA of legislators's votes:

Terminology

Unsupervised techniques: learning (hidden or latent) structure in unlabeled data.

e.g. PCA of legislators's votes: want to see how they are organized—

Terminology

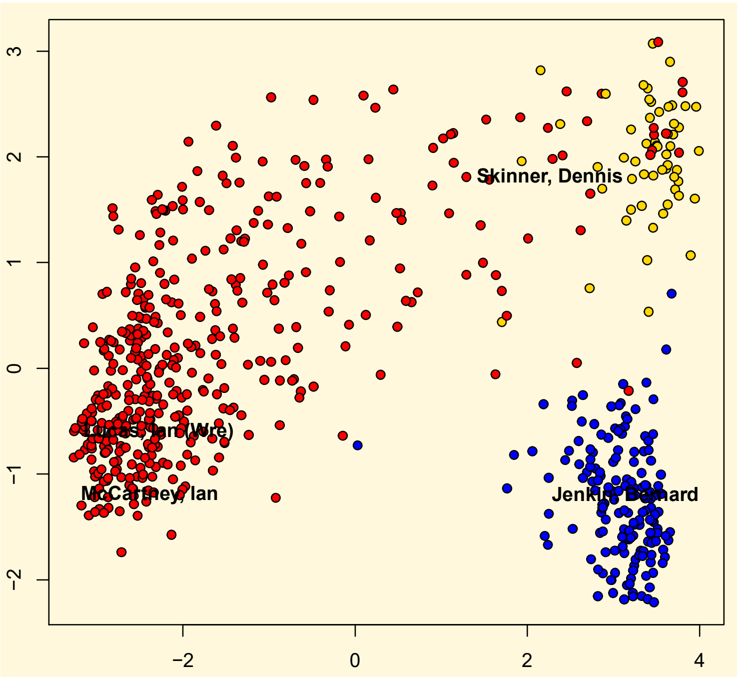
Unsupervised techniques: learning (hidden or latent) structure in unlabeled data.

e.g. PCA of legislators's votes: want to see how they are organized—by party? by ideology? by race?

Terminology

Unsupervised techniques: learning (hidden or latent) structure in unlabeled data.

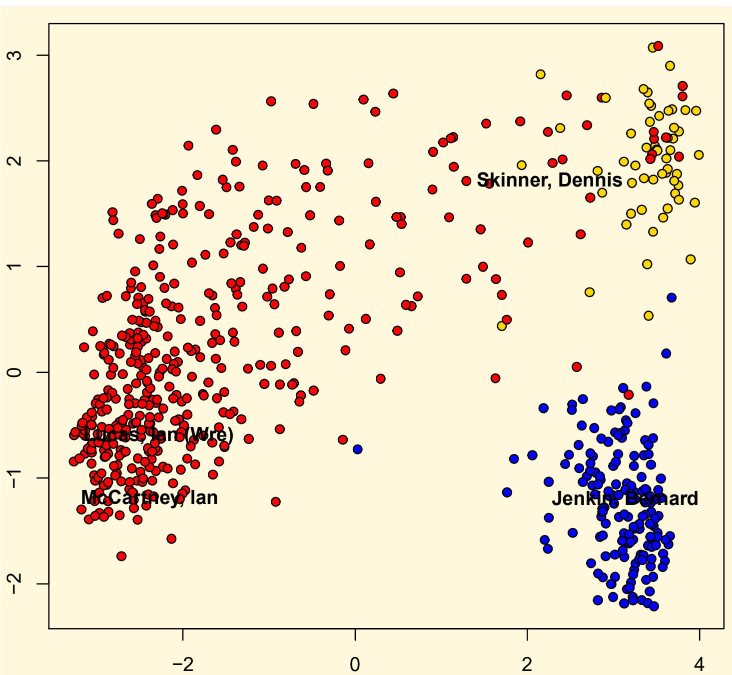
e.g. PCA of legislators's votes: want to see how they are organized—by party? by ideology? by race?



Terminology

Unsupervised techniques: learning (hidden or latent) structure in unlabeled data.

e.g. PCA of legislators's votes: want to see how they are organized—by party? by ideology? by race?

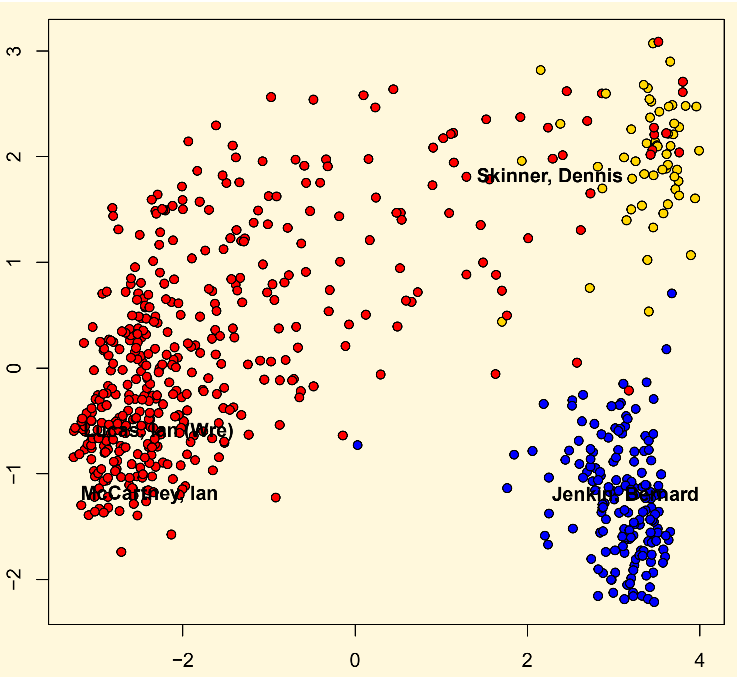


Supervised techniques:

Terminology

Unsupervised techniques: learning (hidden or latent) structure in unlabeled data.

e.g. PCA of legislators's votes: want to see how they are organized—by party? by ideology? by race?

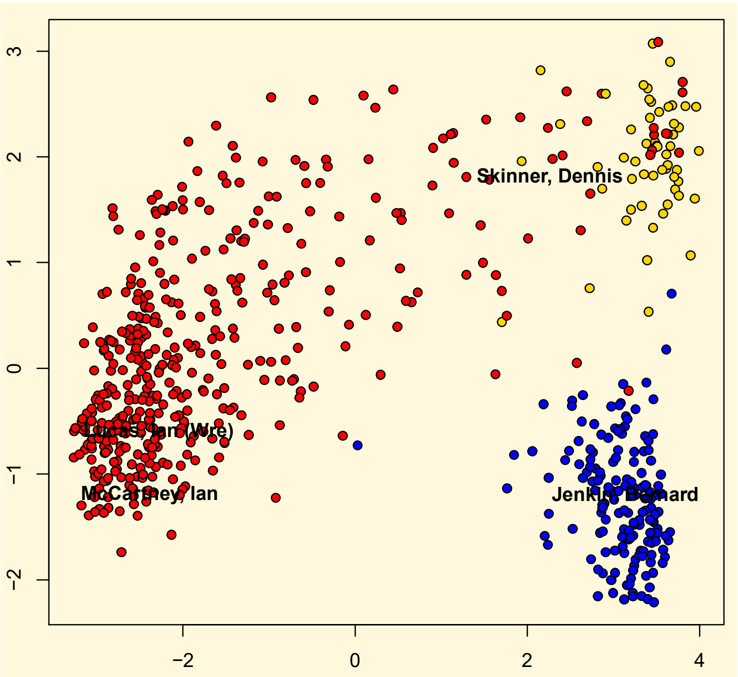


Supervised techniques: learning relationship between inputs and a labeled set of outputs.

Terminology

Unsupervised techniques: learning (hidden or latent) structure in unlabeled data.

e.g. PCA of legislators's votes: want to see how they are organized—by party? by ideology? by race?



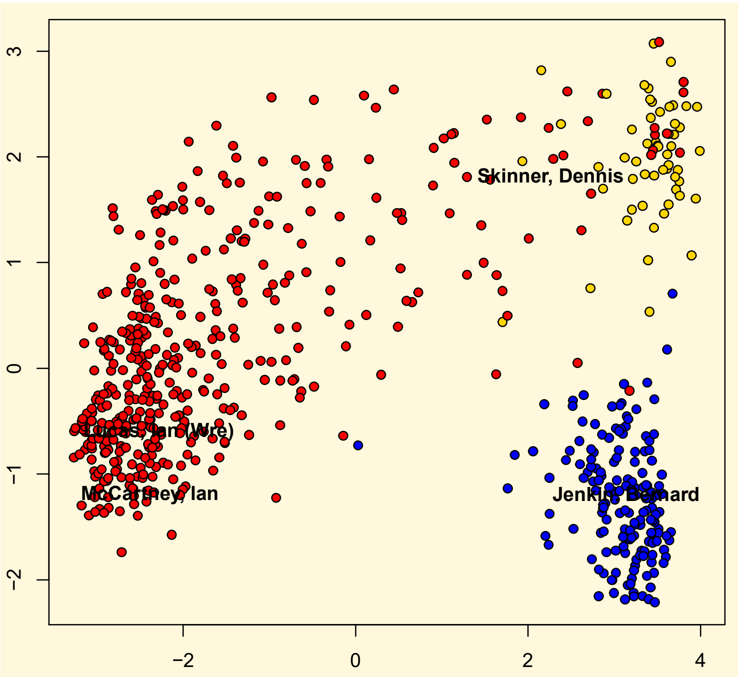
Supervised techniques: learning relationship between inputs and a labeled set of outputs.

e.g. opinion mining:

Terminology

Unsupervised techniques: learning (hidden or latent) structure in unlabeled data.

e.g. PCA of legislators's votes: want to see how they are organized—by party? by ideology? by race?



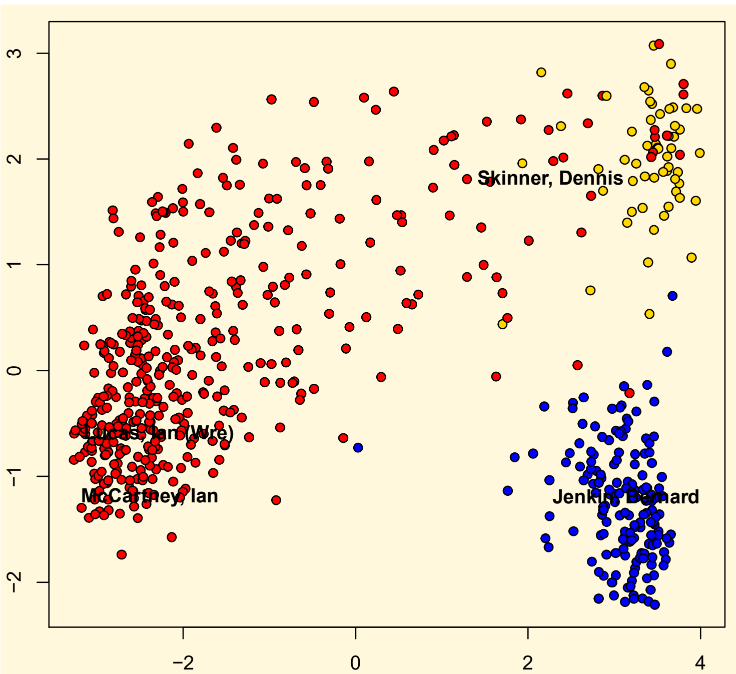
Supervised techniques: learning relationship between inputs and a labeled set of outputs.

e.g. opinion mining: what makes a critic like or dislike a movie ($y \in \{0, 1\}$)?

Terminology

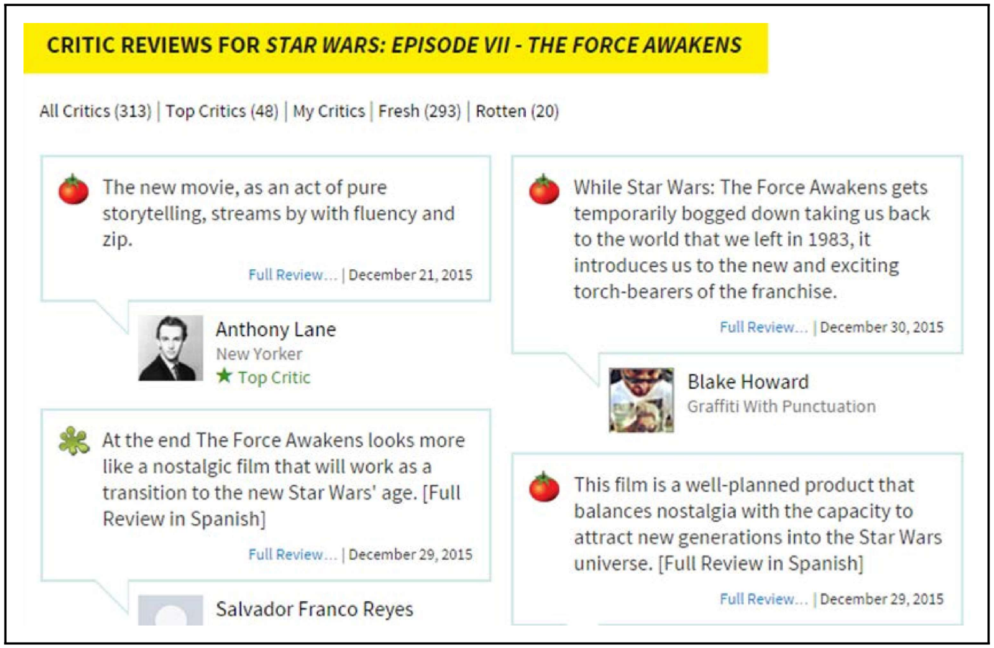
<u>Unsupervised</u> techniques: learning (hidden or latent) structure in unlabeled data.

e.g. PCA of legislators's votes: want to see how they are organized—by party? by ideology? by race?



<u>Supervised</u> techniques: learning relationship between inputs and a labeled set of outputs.

e.g. opinion mining: what makes a critic like or dislike a movie ($y \in \{0, 1\}$)?


CRITIC REVIEWS FOR *STAR WARS: EPISODE VII - THE FORCE AWAKENS*

All Critics (313) | Top Critics (48) | My Critics | Fresh (293) | Rotten (20)

The new movie, as an act of pure storytelling, streams by with fluency and zip.
Full Review... | December 21, 2015
Anthony Lane
New Yorker
★ Top Critic

While Star Wars: The Force Awakens gets temporarily bogged down taking us back to the world that we left in 1983, it introduces us to the new and exciting torch-bearers of the franchise.
Full Review... | December 30, 2015
Blake Howard
Graffiti With Punctuation

At the end The Force Awakens looks more like a nostalgic film that will work as a transition to the new Star Wars' age. [Full Review in Spanish]
Full Review... | December 29, 2015
Salvador Franco Reyes

This film is a well-planned product that balances nostalgia with the capacity to attract new generations into the Star Wars universe. [Full Review in Spanish]
Full Review... | December 29, 2015

Overview: Unsupervised Learning

Overview: Unsupervised Learning

Now our data—humans, documents, observations—are not pre-labeled in terms of some underlying concept.

Overview: Unsupervised Learning

Now our data—humans, documents, observations—are not pre-labeled in terms of some underlying concept.

e.g. while we know who gave a particular speech, we don't yet know what that speech 'represents' in terms of its latent properties,

Overview: Unsupervised Learning

Now our data—humans, documents, observations—are not pre-labeled in terms of some underlying concept.

e.g. while we know who gave a particular speech, we don't yet know what that speech 'represents' in terms of its latent properties, what 'kind' of speech it is,

Overview: Unsupervised Learning

Now our data—humans, documents, observations—are not pre-labeled in terms of some underlying concept.

e.g. while we know who gave a particular speech, we don't yet know what that speech 'represents' in terms of its latent properties, what 'kind' of speech it is, what 'topics' it covers,

Overview: Unsupervised Learning

Now our data—humans, documents, observations—are not pre-labeled in terms of some underlying concept.

e.g. while we know who gave a particular speech, we don't yet know what that speech 'represents' in terms of its latent properties, what 'kind' of speech it is, what 'topics' it covers, what speeches it is similar to conceptually, etc.

Overview: Unsupervised Learning

Now our data—humans, documents, observations—are not pre-labeled in terms of some underlying concept.

e.g. while we know who gave a particular speech, we don't yet know what that speech 'represents' in terms of its latent properties, what 'kind' of speech it is, what 'topics' it covers, what speeches it is similar to conceptually, etc.

Goal is to take the observations and find hidden structure and meaning in them.

Overview: Unsupervised Learning

Now our data—humans, documents, observations—are not pre-labeled in terms of some underlying concept.

e.g. while we know who gave a particular speech, we don't yet know what that speech 'represents' in terms of its latent properties, what 'kind' of speech it is, what 'topics' it covers, what speeches it is similar to conceptually, etc.

Goal is to take the observations and find hidden structure and meaning in them.

$\rightarrow$ look for (dis)similarities between documents (or observations):

Overview: Unsupervised Learning

Now our data—humans, documents, observations—are not pre-labeled in terms of some underlying concept.

e.g. while we know who gave a particular speech, we don't yet know what that speech 'represents' in terms of its latent properties, what 'kind' of speech it is, what 'topics' it covers, what speeches it is similar to conceptually, etc.

Goal is to take the observations and find hidden structure and meaning in them.

$\rightarrow$ look for (dis)similarities between documents (or observations): it will be up to us to *interpret* what the groups/dimensions/concepts represent after the technique has been used.

Overview: Unsupervised Learning

Now our data—humans, documents, observations—are not pre-labeled in terms of some underlying concept.

e.g. while we know who gave a particular speech, we don't yet know what that speech 'represents' in terms of its latent properties, what 'kind' of speech it is, what 'topics' it covers, what speeches it is similar to conceptually, etc.

Goal is to take the observations and find hidden structure and meaning in them.

$\rightarrow$ look for (dis)similarities between documents (or observations): it will be up to us to *interpret* what the groups/dimensions/concepts represent after the technique has been used.

So. . .

So. . .

in contrast to supervised approaches,

So. . .

in contrast to supervised approaches, we won't know 'how correct' the output is in a simple statistical sense

So. . .

in contrast to supervised approaches, we won't know 'how correct' the output is in a simple statistical sense

While there *are* ways to compare unsupervised models, we are generally judging utility/fit in a different way

So. . .

in contrast to supervised approaches, we won't know 'how correct' the output is in a simple statistical sense

While there *are* ways to compare unsupervised models, we are generally judging utility/fit in a different way

So, does it tell us something interesting/plausible?

So. . .

in contrast to supervised approaches, we won't know 'how correct' the output is in a simple statistical sense

While there *are* ways to compare unsupervised models, we are generally judging utility/fit in a different way

So, does it tell us something interesting/plausible? do somewhat similar (e.g. 2, 3, 4 clusters) specifications imply the same thing?

So. . .

in contrast to supervised approaches, we won't know 'how correct' the output is in a simple statistical sense

While there *are* ways to compare unsupervised models, we are generally judging utility/fit in a different way

So, does it tell us something interesting/plausible? do somewhat similar (e.g. 2, 3, 4 clusters) specifications imply the same thing?

(not "what is the recall/precision/accuracy?")

Motivating Problem

Motivating Problem

Have an $n \times p$ matrix,

Motivating Problem

Have an $n \times p$ matrix, want to summarize/analyze:

Motivating Problem

Have an $n \times p$ matrix, want to summarize/analyze: could be DTM.

Motivating Problem

Have an $n \times p$ matrix, want to summarize/analyze: could be DTM.

e.g. Political science: n legislators, p roll calls of interest, $n > p$

Motivating Problem

Have an $n \times p$ matrix, want to summarize/analyze: could be DTM.

e.g. Political science: n legislators, p roll calls of interest, $n > p$

Motivating Problem

Have an $n \times p$ matrix, want to summarize/analyze: could be DTM.

e.g. Political science: n legislators, p roll calls of interest, $n > p$

Name	Party	Vote 1	Vote 2	Vote 3	
Ainsworth, Peter (E S)	Con	NA	1	NA	...
Alexander, Douglas	Lab	NA	0	0	...
Allan, Richard	LD	1	0	1	...
Allen, Graham	Lab	0	0	0	...
Amess, David	Con	1	1	NA	...
					$\ddots$

Motivating Problem

Have an $n \times p$ matrix, want to summarize/analyze: could be DTM.

Motivating Problem

Have an $n \times p$ matrix, want to summarize/analyze: could be DTM.

e.g. Text: n speakers, p features in the speeches (often $p > n$ for text problems)

Motivating Problem

Have an $n \times p$ matrix, want to summarize/analyze: could be DTM.

e.g. Text: n speakers, p features in the speeches (often $p > n$ for text problems)

Name	Party	'cost'	'spend'	'tax'	
Ainsworth, Peter (E S)	Con	0.00	0.01	0.30	...
Alexander, Douglas	Lab	0.32	0.20	0.86	...
Allan, Richard	LD	0.99	0.82	0.61	...
Allen, Graham	Lab	0.52	0.86	0.34	...
Amess, David	Con	0.07	0.34	0.33	...
					$\ddots$

PCA: Introduction

PCA: Introduction

Possibly oldest multivariate technique (Pearson, 1901?)