Cross-Cutting Models of Lexical Semantics

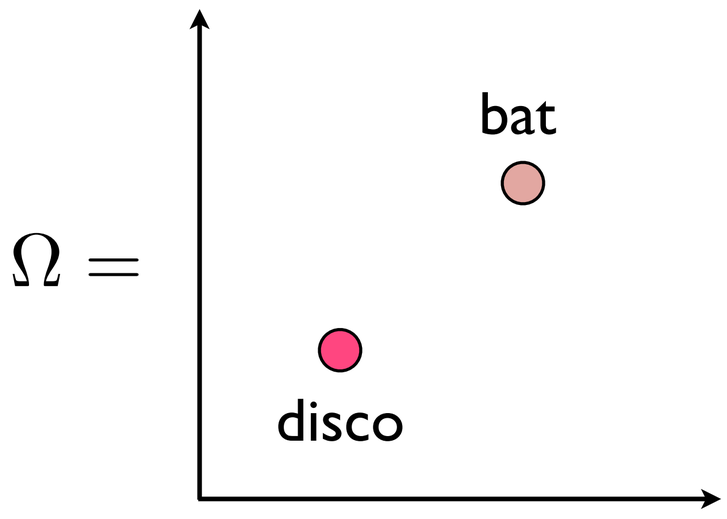
Joseph Reisinger and Raymond Mooney

Distributional Lexical Semantics

- Represent "meaning" as a point/vector in a high-dimensional space

- Word relatedness correlates with some distance metric

Almuhareb and Poesio (2004), Baroni and Lenci (2009), Bullinaria and Levy (2007), Erk (2007), Griffiths et al. (2007), Landauer and Dumais (1997), Moldovan (2006), Padó and Lapata (2007), Pantel and Pennacchiotti (2006), Sahlgren (2006), Turney and Pantel (2010)

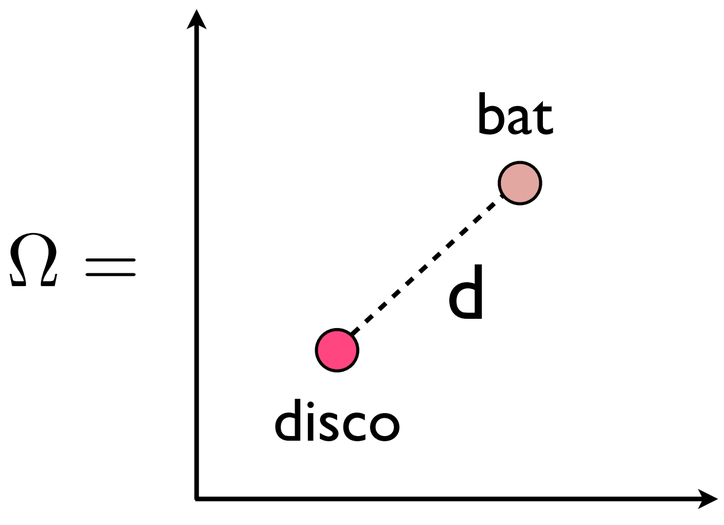


Distributional Lexical Semantics

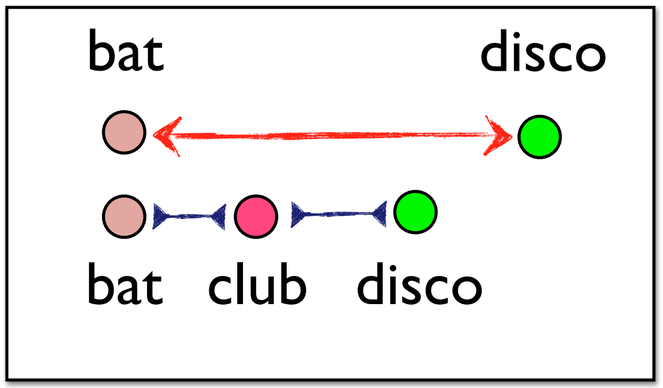
- Represent "meaning" as a point/vector in a high-dimensional space

- Word relatedness correlates with some distance metric

Almuhareb and Poesio (2004), Baroni and Lenci (2009), Bullinaria and Levy (2007), Erk (2007), Griffiths et al. (2007), Landauer and Dumais (1997), Moldovan (2006), Padó and Lapata (2007), Pantel and Pennacchiotti (2006), Sahlgren (2006), Turney and Pantel (2010)



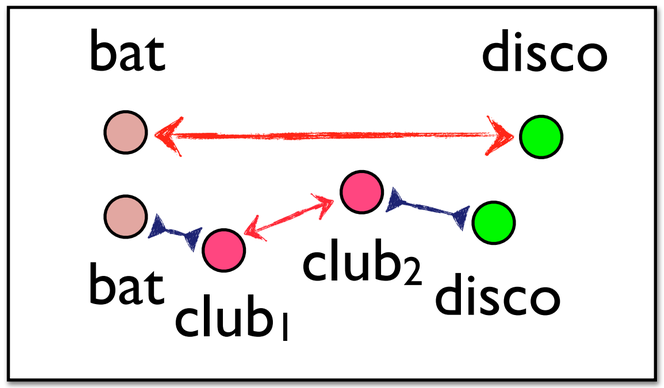
Distributional Lexical Semantics



"meaning violates the triangle inequality"

Tversky and Gati (1982), Griffiths et al. (2007)

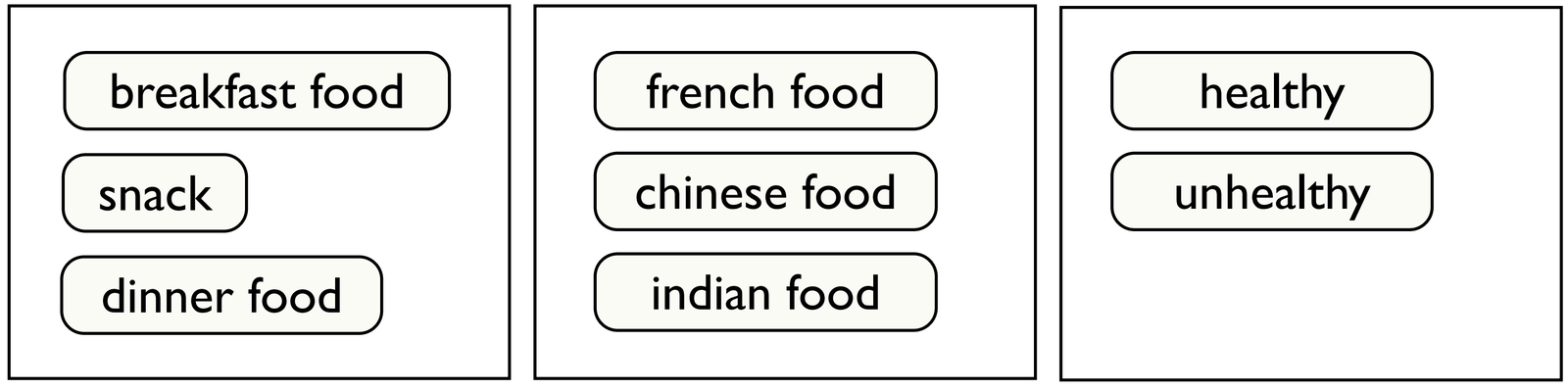
Distributional Lexical Semantics



"meaning violates the triangle inequality"
Tversky and Gati (1982), Griffiths et al. (2007)

- Address metric violations by learning word sense clusters / making use of local context
- Can we build a model that captures this directly?

Cross-cutting Concept Organization



- Human concept organization exhibits cross cutting structure
  Rosch, et al. (1976); Ross & Murphy (1999); Medin, et al. (2005); Shaftoe, et al. (2011)

- Each categorization system controls what kinds of generalizations (e.g. inferences) are valid.

- Do word usages exhibit similar cross-cutting?

  - Xue, Chen and Palmer (2006): sense disambiguation requires vastly different features for different polysemous verbs in Chinese.

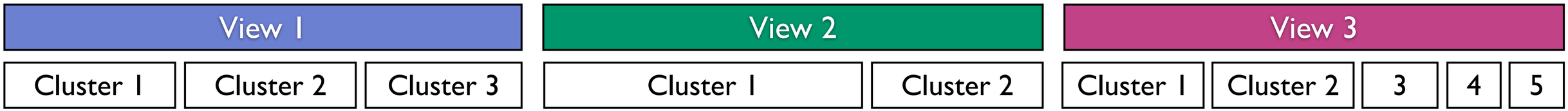
Multi View Multinomial Clustering

- There are many valid word clusterings, each capturing different aspects of syntax or topicality
- We introduce a model to explicitly capture multiple organizational systems
- Cross-cutting categorization / latent subspaces with separate, coherent clusterings
- Implement using LDA and DPMM primitives / Gibbs sampling

and is ___ **we are ___** **he is ___**	**and are ___** **which was ___** **who are ___**	
unwilling willing reluctant refusing glad	exceedingly sincerely logically justly appropriately	about because

brand new ___ **selection of ___** **___ for sale**	**results for ___** **the latest ___** **to buy ___**	
samsung panasonic toshiba sony epson	toyota nissan mercedes volvo audi	dunlop yokohama toyo uniroyal michelin

Multi View Multinomial Clustering Model



Data

Austin

History of Austin, Texas, University of Texas Medical Branch, 1993 Pacific hurricane season, Rutherford B. Hayes, List of pipeline accidents, List of Austin City Limits performers, Texas in the American Civil War, 6th Cavalry Regiment (United States)

___ texas homes, ___ law school, the citizens of ___, the ___ business directory, ___ police department, university in ___, ___ vacation rentals, the ___ parks and, by the ___ business journal, coming to ___, the ___ area, deals on ___ hotels

Betrayed

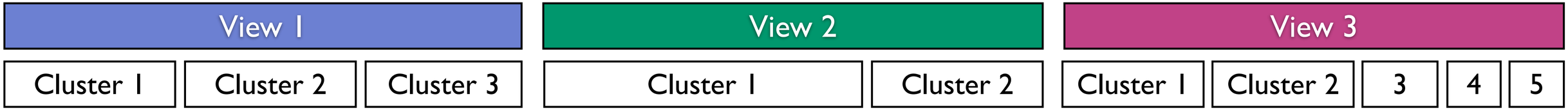
Survivor: The Amazon, Personal life of Marcus Tullius Cicero, Numb3rs, Huns, Rurouni Kenshin, Liberation of Paris, The Knightly Tale of Gologras and Gawain, Territories in The Pendragon Adventure, A Storm of Swords, Connor MacLeod, Paul Atreides

her manner ___, being ___ by their, ___ and murdered, ___ his weakness, she ___ him, ___ the secret, ___ by her husband, a voice that ___, who felt ___, ___ to the police, ___ their country, suspected of having ___, ___ the confidence, even when ___

Cat

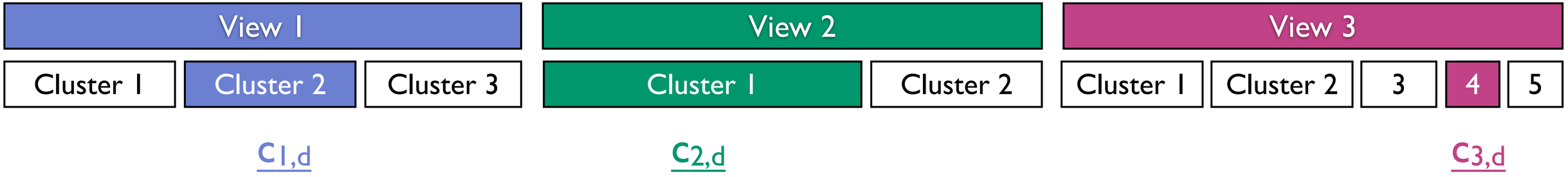
South China Tiger, Hybrid (biology), List of mammals of Cameroon, Cantonese cuisine, Pound Puppies, Wonder Pets, The Wizard of Oz (1902 stage play), Mee-Ow, Animal rights, Rickrolling, Mera (comics), Taboo food and drink, Tuna, Garfield: The Movie

ate the ___, have a ___ and a, the ___ and the mouse, the ___ who killed, ___ toys by, ___ in the city, ___ was diagnosed, crazy ___ lady, ___ of the month, protect your ___ from, new ___ food, and bought a ___, ___ or other animal, a sick ___,

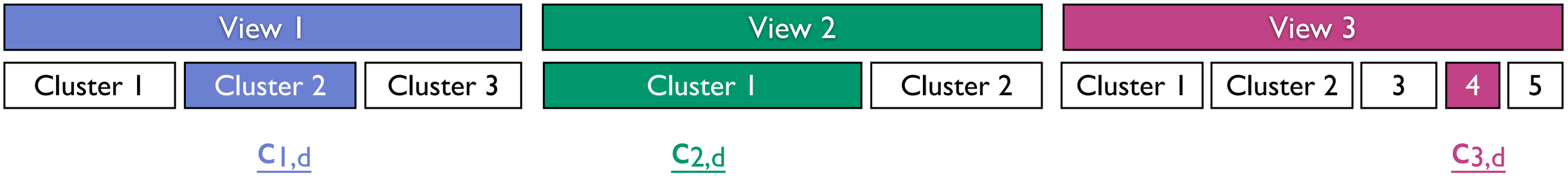

View 1
Cluster 1
Cluster 2
Cluster 3
View 2
Cluster 1
Cluster 2
View 3
Cluster 1
Cluster 2
3
4
5

View 1			View 2		View 3				
Cluster 1	Cluster 2	Cluster 3	Cluster 1	Cluster 2	Cluster 1	Cluster 2	3	4	5

Cat	South China Tiger, Hybrid (biology), List of mammals of Cameroon, Cantonese cuisine, Pound Puppies, Wonder Pets, The Wizard of Oz (1902 stage play), Mee-Ow, Animal rights, Rickrolling, Mera (comics), Taboo food and drink, Tuna, Garfield: The Movie ate the ___, have a ___ and a, the ___ and the mouse, the ___ who killed, ___ toys by, ___ in the city, ___ was diagnosed, crazy ___ lady, ___ of the month, protect your ___ from, new ___ food, and bought a ___, ___ or other animal, a sick ___,

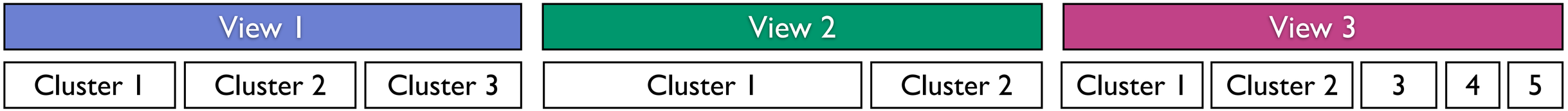


- Select a cluster assignment $c_{v,d}$ for d in each view v (DPMM) i.e. words are assigned to clusters within each view



- Select a cluster assignment $c_{v,d}$ for d in each view v (DPMM) i.e. words are assigned to clusters within each view

- Select a view v_f for each observed feature, and generate it from $c_{vf,d}$ (LDA) i.e. features distributed between views



Cat

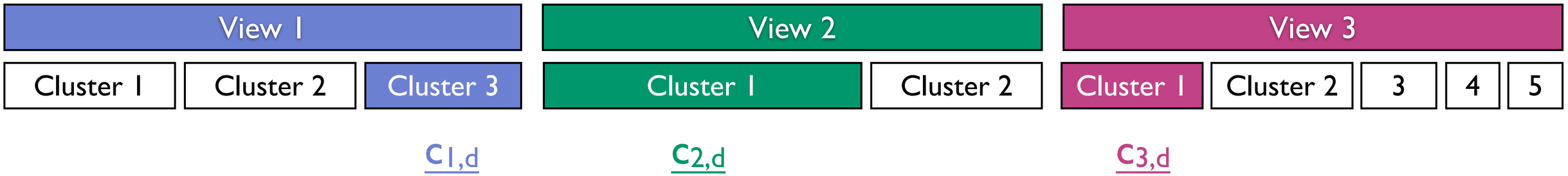
South China Tiger, Hybrid (biology), List of mammals of Cameroon, Cantonese cuisine, Pound Puppies, Wonder Pets, The Wizard of Oz (1902 stage play), Mee-Ow, Animal rights, Rickrolling, Mera (comics), Taboo food and drink, Tuna, Garfield: The Movie
ate the ___, have a ___ and a, the ___ and the mouse, the ___ who killed, ___ toys by, ___ in the city, ___ was diagnosed, crazy ___ lady, ___ of the month, protect your ___ from, new ___ food, and bought a ___, ___ or other animal, a sick ___,

Betrayed

Survivor: The Amazon, Personal life of Marcus Tullius Cicero, Numb3rs, Huns, Rurouni Kenshin, Liberation of Paris, The Knightly Tale of Gologras and Gawain, Territories in The Pendragon Adventure, A Storm of Swords, Connor MacLeod, Paul Atreides
her manner ___, being ___ by their, ___ and murdered, ___ his weakness, she ___ him, ___ the secret, ___ by her husband, a voice that ___, who felt ___, ___ to the police, ___ their country, suspected of having ___, ___ the confidence, even when ___

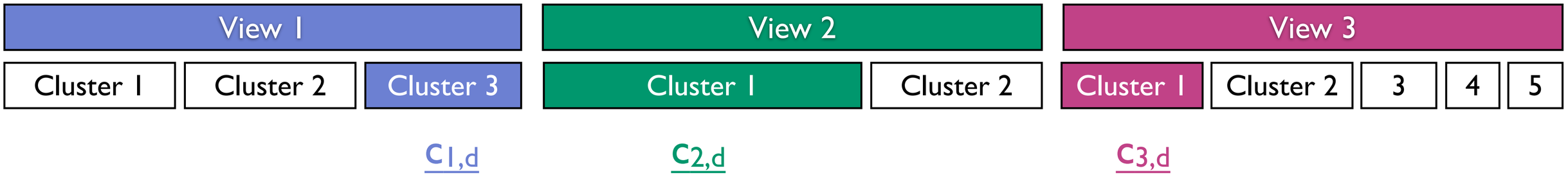
- Select a cluster assignment $c_{v,d}$ for d in each view v (DPMM) i.e. words are assigned to clusters within each view

- Select a view v_f for each observed feature, and generate it from $c_{vf,d}$ (LDA) i.e. features distributed between views



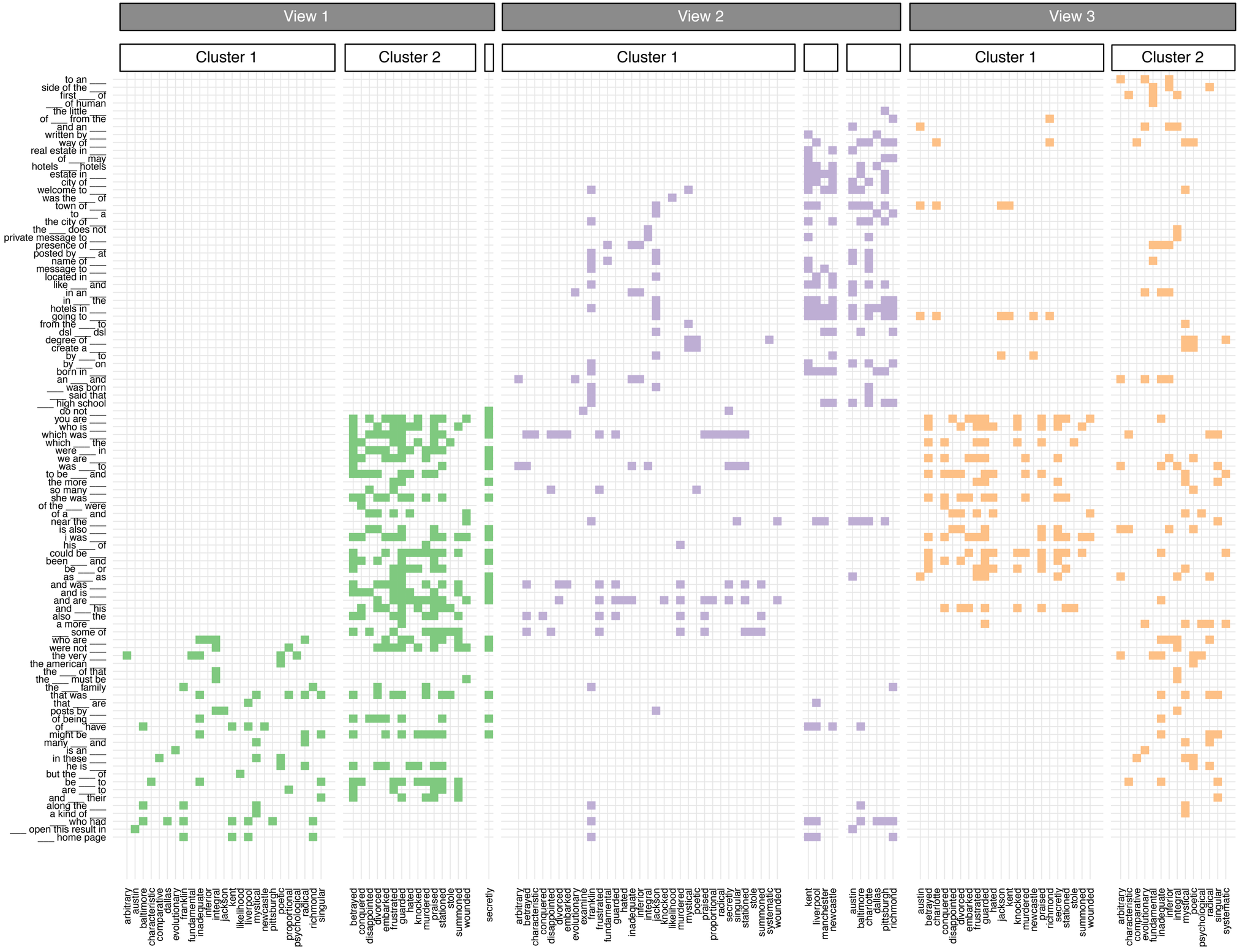
- Select a cluster assignment $c_{v,d}$ for d in each view v (DPMM) i.e. words are assigned to clusters within each view

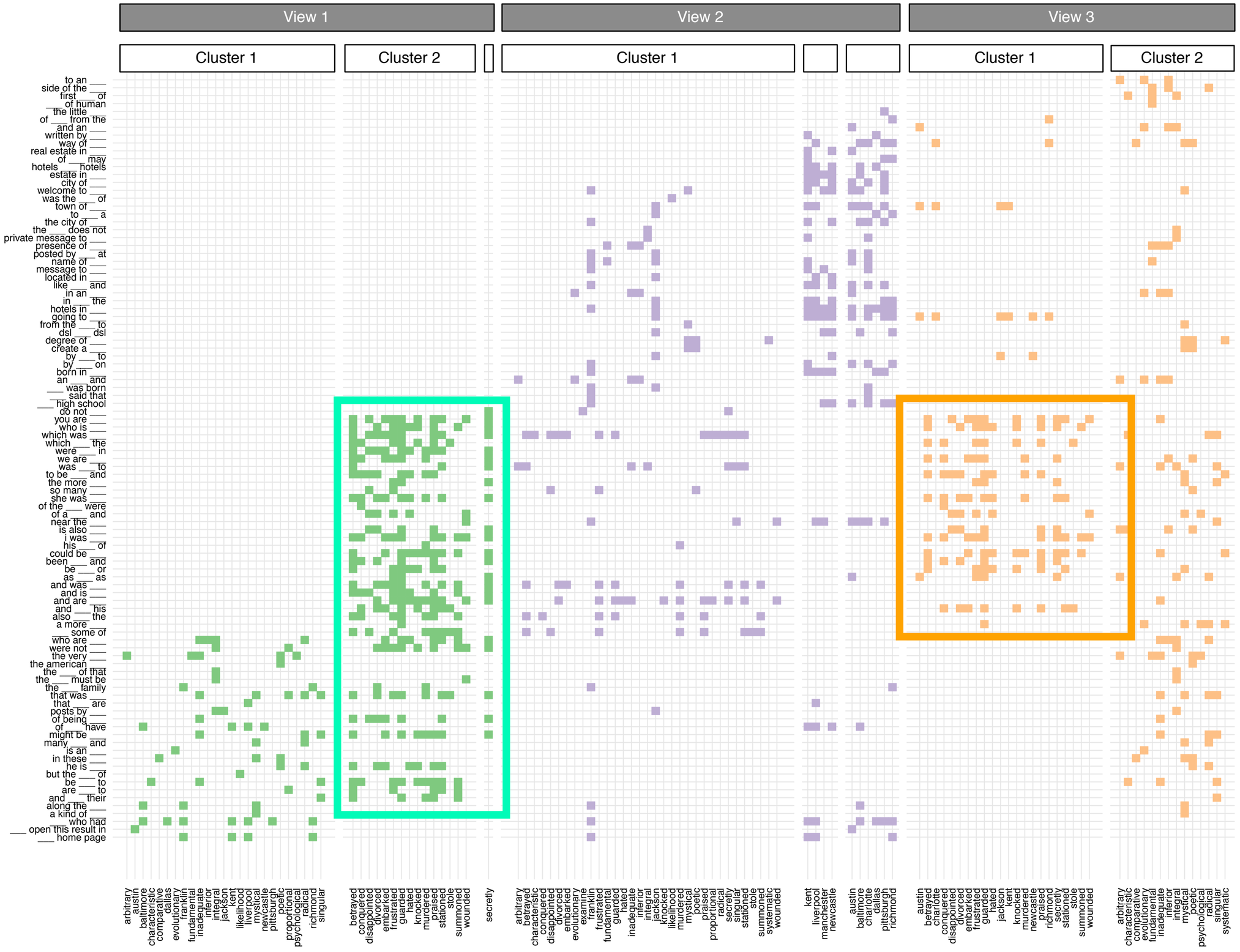
- Select a view v_f for each observed feature, and generate it from $c_{vf,d}$ (LDA) i.e. features distributed between views



- Select a cluster assignment $c_{v,d}$ for d in each view v (DPMM) i.e. words are assigned to clusters within each view

- Select a view v_f for each observed feature, and generate it from $c_{vf,d}$ (LDA) i.e. features distributed between views


View 1
View 2
View 3
Cluster 1
Cluster 2
Cluster 1
Cluster 1
Cluster 2
to an ___
side of the ___
first ___ of
___ of human
the little ___
of ___ from the
___ and an ___
written by ___
way of ___
real estate in ___
of ___ may
hotels ___ hotels
estate in ___
city of ___
welcome to ___
was the ___ of
town of ___
to ___ a
the city of ___
the ___ does not
private message to ___
presence of ___
posted by ___ at
name of ___
message to ___
located in ___
like ___ and
in an ___
in ___ the
hotels in ___
going to ___
from the ___ to
dsl ___ dsl
degree of ___
create a ___
by ___ to
by ___ on
born in ___
an ___ and
___ was born
___ said that
___ high school
do not ___
you are ___
who is ___
which was ___
which ___ the
were ___ in
we are ___
was ___ to
to be ___ and
the more ___
so many ___
she was ___
of the ___ were
of a ___ and
near the ___
is also ___
i was ___
his ___ of
could be ___
been ___ and
be ___ or
as ___ as
and was ___
and is ___
and are ___
and ___ his
also ___ the
a more ___
___ some of
who are ___
were not ___
the very ___
the american
the ___ of that
the ___ must be
the ___ family
that was ___
that ___ are
posts by ___
of being ___
of ___ have
might be ___
many ___ and
is an ___
in these ___
he is ___
but the ___ of
be ___ to
are ___ to
and ___ their
along the ___
a kind of ___
___ who had
___ open this result in
___ home page
arbitrary
austin
baltimore
characteristic
compatible
dallas
evolutionary
franklin
fundamental
inadequate
inferior
integral
jackson
kent
likelihood
liverpool
mystical
newcastle
pittsburgh
poetic
proportional
psychological
radical
richmond
singular
betrayed
conquered
disappointed
divorced
embarked
frustrated
guarded
hated
knocked
murdered
praised
stationed
stole
summoned
wounded
secretly
arbitrary
betrayed
characteristic
conquered
disappointed
divorced
embarked
evolutionary
examine
franklin
frustrated
fundamental
guarded
hated
inadequate
inferior
integral
jackson
knocked
likelihood
murdered
mystical
poetic
praised
proportional
radical
secretly
singular
stationed
stole
summoned
systematic
wounded
kent
liverpool
manchester
newcastle
austin
baltimore
charlotte
dallas
pittsburgh
richmond
austin
betrayed
charlotte
conquered
disappointed
divorced
embarked
frustrated
guarded
hated
jackson
kent
knocked
murdered
newcastle
praised
richmond
secretly
stationed
stole
summoned
wounded
arbitrary
characteristic
compatible
evolutionary
fundamental
inadequate
inferior
integral
mystical
poetic
psychological
radical
singular
systematic


View 1
View 2
View 3
Cluster 1
Cluster 2
Cluster 1
Cluster 1
Cluster 2
to an ___
side of the ___
first ___ of
___ of human
the little ___
of ___ from the
___ and an ___
written by ___
way of ___
real estate in ___
of ___ may
hotels ___ hotels
estate in ___
city of ___
welcome to ___
was the ___ of
town of ___
to ___ a
the city of ___
the ___ does not
private message to ___
presence of ___
posted by ___ at
name of ___
message to ___
located in ___
like ___ and
in an ___
in ___ the
hotels in ___
going to ___
from the ___ to
dsl ___ dsl
degree of ___
create a ___
by ___ to
by ___ on
born in ___
an ___ and
___ was born
___ said that
___ high school
do not ___
you are ___
who is ___
which was ___
which ___ the
were ___ in
we are ___
was ___ to
to be ___ and
the more ___
so many ___
she was ___
of the ___ were
of a ___ and
near the ___
is also ___
i was ___
his ___ of
could be ___
been ___ and
be ___ or
as ___ as
and was ___
and is ___
and are ___
and ___ his
also ___ the
a more ___
___ some of
who are ___
were not ___
the very ___
the american
the ___ of that
the ___ must be
the ___ family
that was ___
that ___ are
posts by ___
of being ___
of ___ have
might be ___
many ___ and
is an ___
in these ___
he is ___
but the ___ of
be ___ to
are ___ to
and ___ their
along the ___
a kind of ___
___ who had
___ open this result in
___ home page
arbitrary
austin
baltimore
characteristic
compatible
dallas
evolutionary
franklin
fundamental
inadequate
inferior
integral
jackson
kent
likelihood
liverpool
mystical
newcastle
pittsburgh
poetic
proportional
psychological
radical
richmond
singular
betrayed
conquered
disappointed
divorced
embarked
frustrated
guarded
hated
knocked
murdered
praised
stationed
stole
summoned
wounded
secretly
arbitrary
betrayed
characteristic
conquered
disappointed
divorced
embarked
evolutionary
examine
franklin
frustrated
fundamental
guarded
hated
inadequate
inferior
integral
jackson
knocked
likelihood
murdered
mystical
poetic
praised
proportional
radical
secretly
singular
stationed
stole
summoned
systematic
wounded
kent
liverpool
manchester
newcastle
austin
baltimore
charlotte
dallas
pittsburgh
richmond
austin
betrayed
charlotte
conquered
disappointed
divorced
embarked
frustrated
guarded
hated
jackson
kent
knocked
murdered
newcastle
praised
richmond
secretly
stationed
stole
summoned
wounded
arbitrary
characteristic
compatible
evolutionary
fundamental
inadequate
inferior
integral
mystical
poetic
psychological
radical
singular
systematic

Data

Austin	History of Austin, Texas, University of Texas Medical Branch, 1993 Pacific hurricane season, Rutherford B. Hayes, List of pipeline accidents, List of Austin City Limits performers, Texas in the American Civil War, 6th Cavalry Regiment (United States) ___ texas homes, ___ law school, the citizens of ___, the ___ business directory, ___ police department, university in ___, ___ vacation rentals, the ___ parks and, by the ___ business journal, coming to ___, the ___ area, deals on ___ hotels
Betrayed	Survivor: The Amazon, Personal life of Marcus Tullius Cicero, Numb3rs, Huns, Rurouni Kenshin, Liberation of Paris, The Knightly Tale of Gologras and Gawain, Territories in The Pendragon Adventure, A Storm of Swords, Connor MacLeod, Paul Atreides her manner ___, being ___ by their, ___ and murdered, ___ his weakness, she ___ him, ___ the secret, ___ by her husband, a voice that ___, who felt ___, ___ to the police, ___ their country, suspected of having ___, ___ the confidence, even when ___

- Word set: Top 43.7k words ranked by frequency in Wikipedia (ex top 1% as stop words)
- Syntax features: Contextual patterns from combined Google Web n-gram + Google Books n-gram corpus (3.5M features)
- Document features: Wikipedia article occurrence count (120k features)

Intrusion Task

word | context | document

humor
ingenuity
delight
advertisers
astonishment

- "Model-based" lexical semantics: read word similarity directly from the model
- Intruders are drawn from the top terms in other clusters
- More robust than asking for numeric similarity judgements
- Less inter-rater calibration required

Chang et. al (2009)

Intrusion Task

word	context	document
humor ingenuity delight advertisers astonishment		

- "Model-based" lexical semantics: read word similarity directly from the model
- Intruders are drawn from the top terms in other clusters
- More robust than asking for numeric similarity judgements
- Less inter-rater calibration required

Chang et. al (2009)

Intrusion Task

word	context	document
humor	____ is characterized	
ingenuity	symptoms of ____	
delight	cases of ____	
advertisers	in cases of ____	
astonishment	real estate in ____	

- "Model-based" lexical semantics: read word similarity directly from the model
- Intruders are drawn from the top terms in other clusters
- More robust than asking for numeric similarity judgements
- Less inter-rater calibration required

Chang et. al (2009)

Intrusion Task

word	context	document
humor	____ is characterized	
ingenuity	symptoms of ____	
delight	cases of ____	
advertisers	in cases of ____	
astonishment	real estate in ____	

- "Model-based" lexical semantics: read word similarity directly from the model
- Intruders are drawn from the top terms in other clusters
- More robust than asking for numeric similarity judgements
- Less inter-rater calibration required

Chang et. al (2009)

Intrusion Task

word	context	document
humor	____ is characterized	Puerto Rican cuisine
ingenuity	symptoms of ____	Greek cuisine
delight	cases of ____	ThinkPad
advertisers	in cases of ____	Palestinian cuisine
astonishment	real estate in ____	Field ration

- "Model-based" lexical semantics: read word similarity directly from the model
- Intruders are drawn from the top terms in other clusters
- More robust than asking for numeric similarity judgements
- Less inter-rater calibration required

Chang et. al (2009)

Intrusion Task

word	context	document
humor	____ is characterized	Puerto Rican cuisine
ingenuity	symptoms of ____	Greek cuisine
delight	cases of ____	ThinkPad
advertisers	in cases of ____	Palestinian cuisine
astonishment	real estate in ____	Field ration

- "Model-based" lexical semantics: read word similarity directly from the model
- Intruders are drawn from the top terms in other clusters
- More robust than asking for numeric similarity judgements
- Less inter-rater calibration required

Chang et. al (2009)

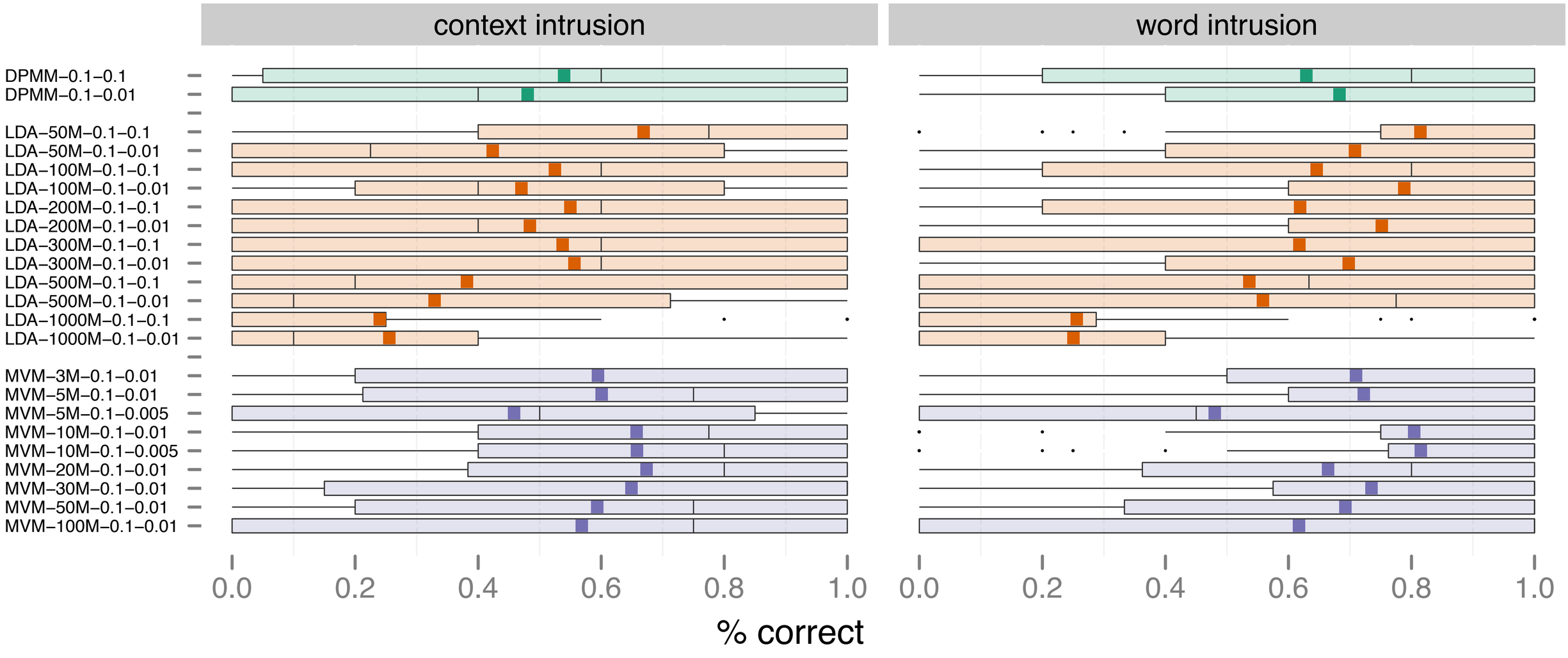
Evaluation

- Amazon Mechanical Turk
- 1256 unique raters (Country=US, >96% approval)
- 5.7k unique intrusion tasks at 5x duplication: ~30k evaluations total
- 2736 rejected
  - Per-user average time for <1.5s / question
  - Low-entropy answers
  - Low agreement

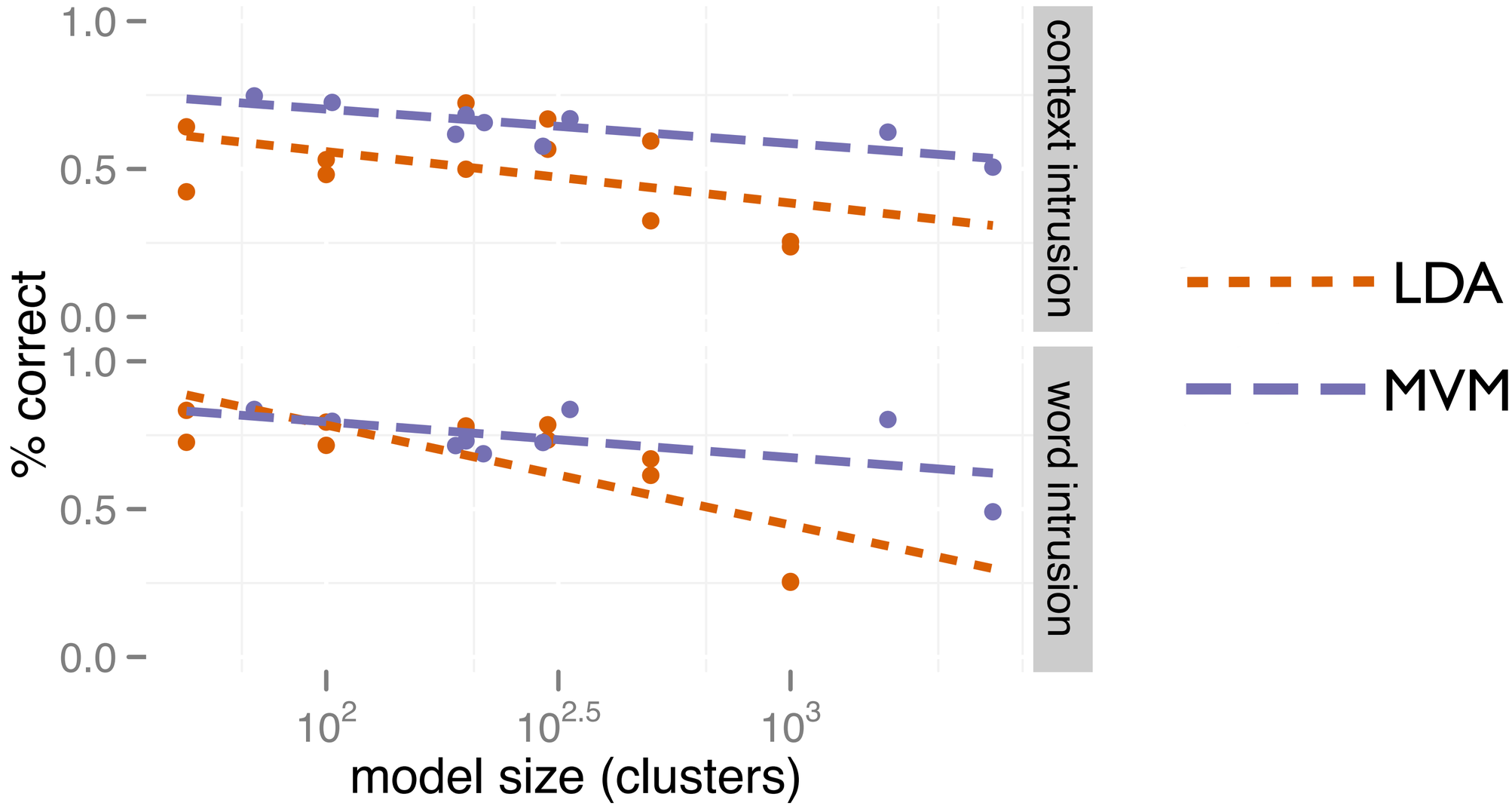
User Comments

U1	I just tried 30 of the what doesn't belong ones. They took about 30 seconds each due to thinking time so not worth it for me.
U2	I don't understand the fill in the blank ones to be honest. I just kinda pick one,since I don't know what's expected lol
U3	Your not filling in the blank just ignore the blank and think about how the words they show relate to each other and choose the one that relates least. Some have just words and no blanks.
U4	These seem very subjective to mw. i hope there isn't definite correct answers because some of them make me go [emoticon of head-scratching]
U5	I looked and have no idea. I guess I'm a word idiot because I don't see the relation between the words in the preview HIT - too scared to try any of these.
U6	I didn't dive in but I did more than I should have they were just too easy. Most of them I could tell what did not belong, some were pretty iffy though.

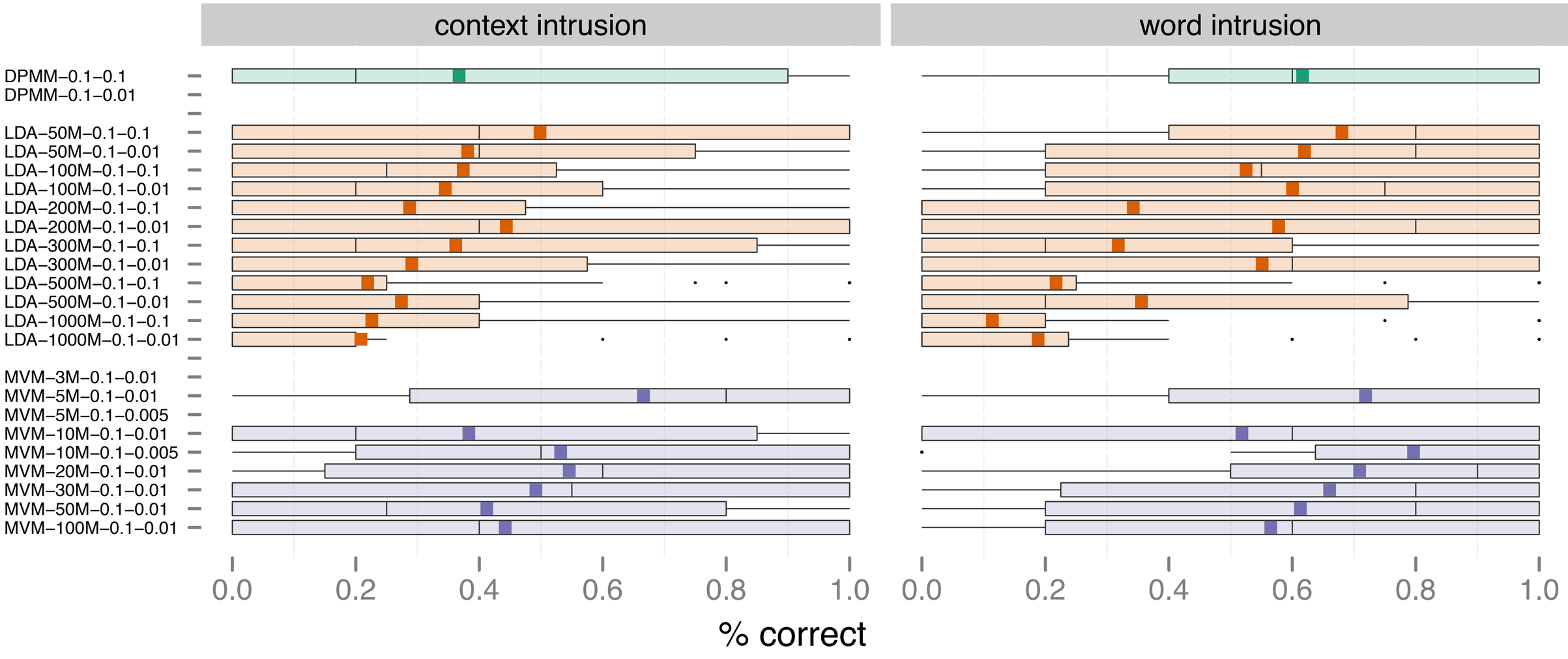
Syntax features only (freq>50; "common")



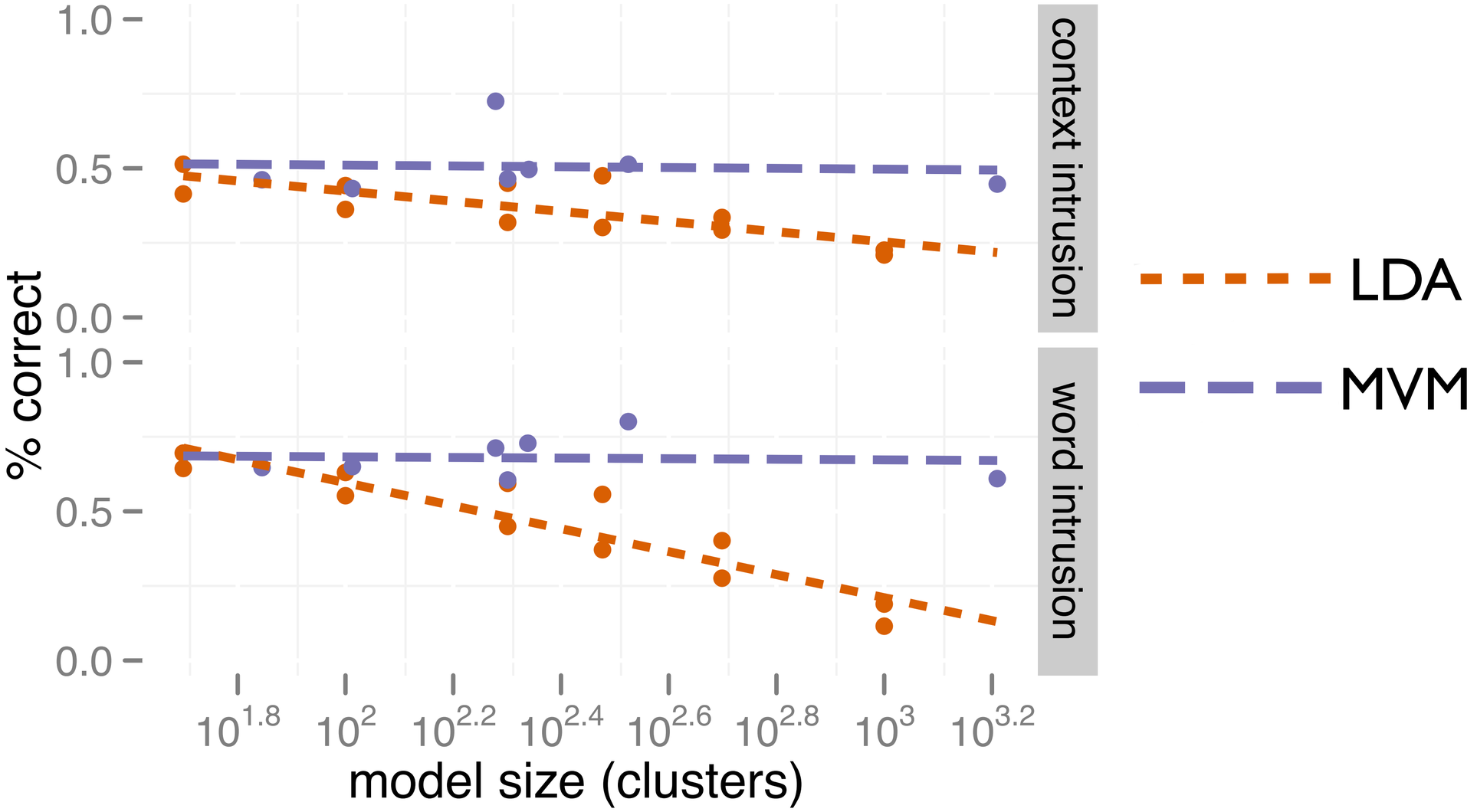
Syntax features only (freq>50; "common")



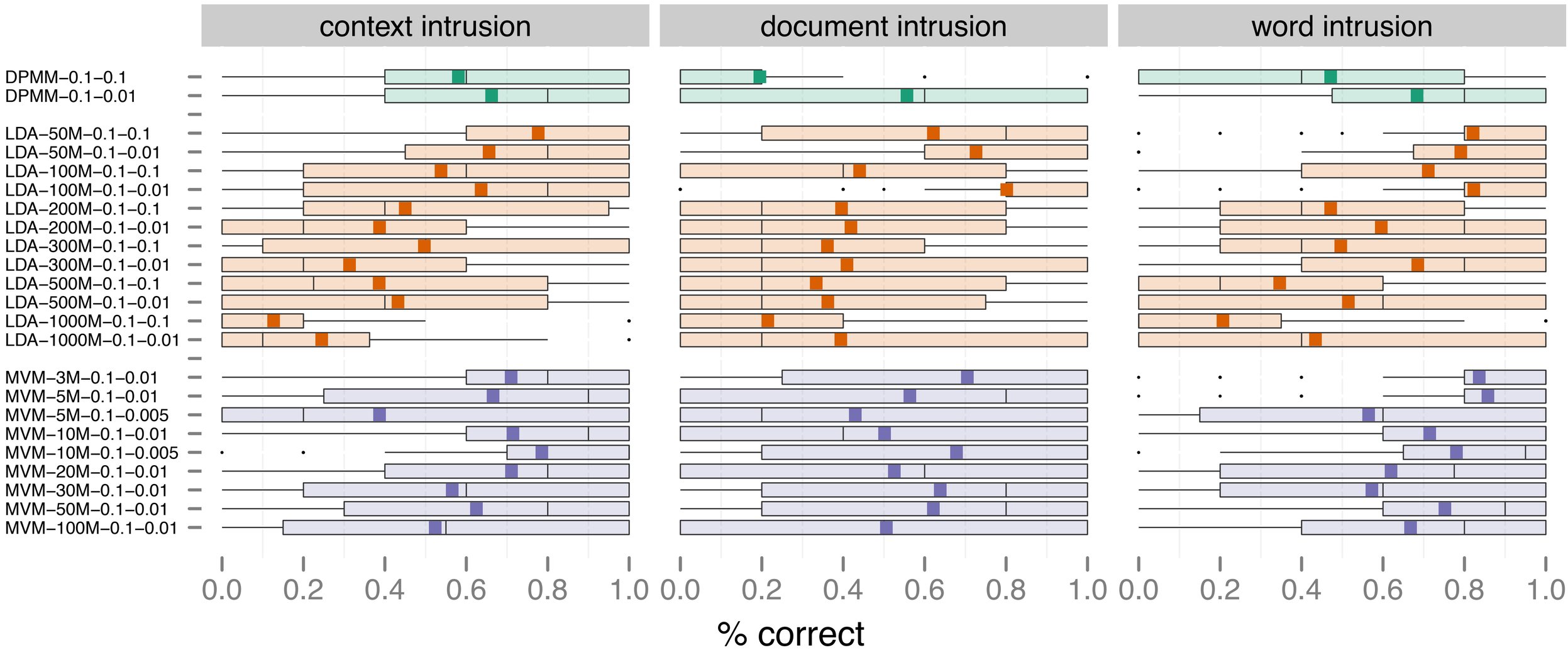
Syntax features only (freq < 50;“rare”)



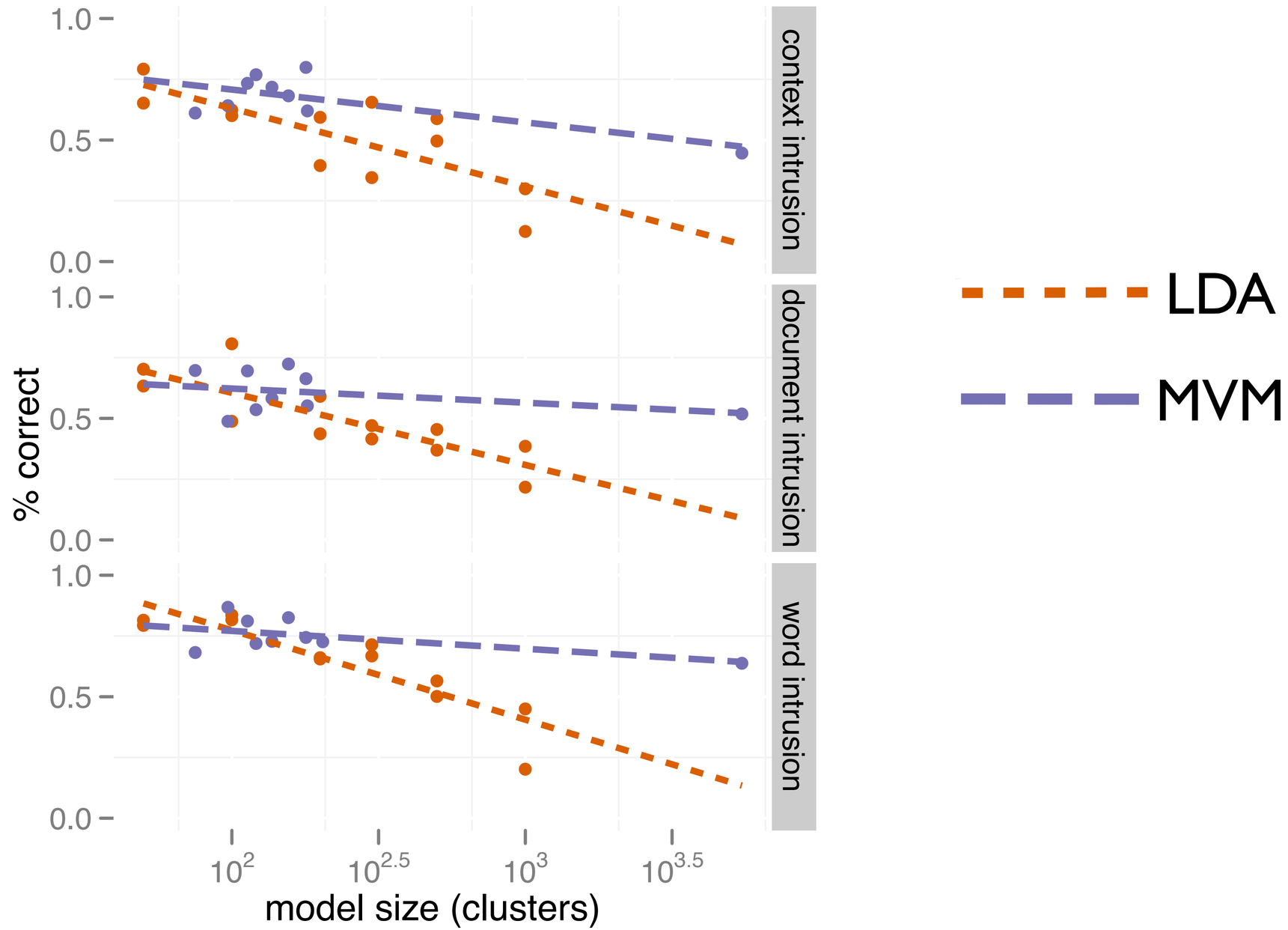
Syntax features only (freq < 50; "rare")



“Common” syntax features + document features



“Common” syntax features + document features



Conclusion

- Introduced a latent variable model accounting for cross-cutting / multiple clustering structure in word meaning

- Large-scale human evaluation of the semantic coherence of similarity predictions

- Significantly higher precision intrusion identification than related model-based approaches

  - Even for fine-grained clusterings