Semantic Analysis: Topic Modeling

Natural Language Processing

Types of Topic Modeling

- **Canonical**—Match a preestablished list of topics for our domain.

Types of Topic Modeling

- **Canonical**—Match a preestablished list of topics for our domain.

- **Organic**—Discover the “natural” topics of a corpus.

Types of Topic Modeling

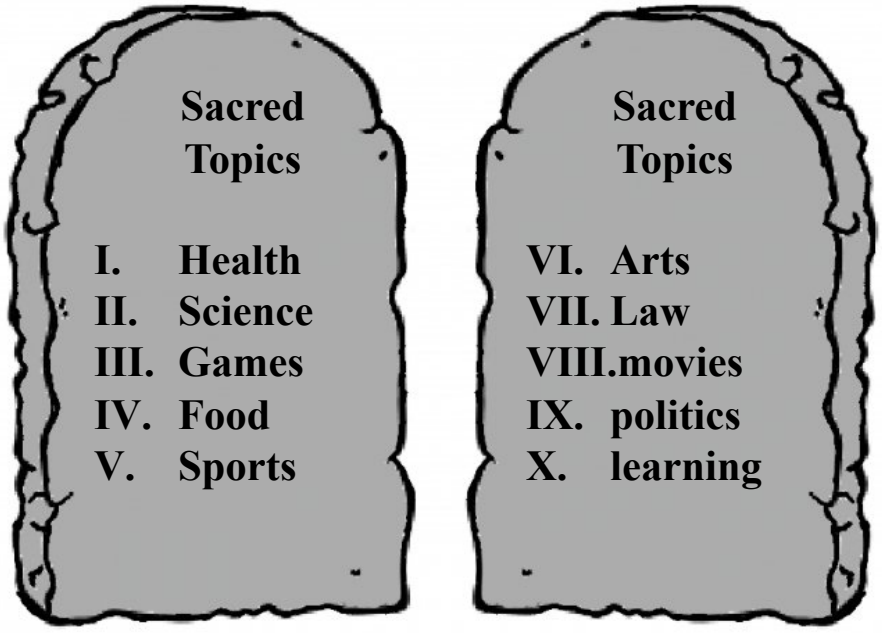
- **Canonical**—Match a preestablished list of topics for our domain.
- **Organic**—Discover the "natural" topics of a corpus.
- **Entity-centric**—Topics are strongly related to sets of NEs that may change over time.

Types of Topic Modeling: Examples

Canonical—Match an established list of topics for our domain.

The topic list comes from an authoritative source, such as:

- The Library of Congress
- The Vatican
- The Thomson Bible Chain Reference
- An anointed topic expert, e.g., some ~~overpaid tool~~ esteemed gentleman who works at a Big Publisher.



Types of Topic Modeling: Examples

Organic—The ideal is to let topics bubble up from out of the "lake" of unstructured documents.

Types of Topic Modeling: Examples

Entity-centric—Topics are strongly related to sets of NEs that may change over time.

- Topics are worth little unless tied to NEs.
  NEs are worth little unless tied to topics.

AI Community Bias

And the winner is…organic!

Organic is the type of topic modeling that most folks in AI want to do, rather than canonical or entity-centric.

Why?

AI Community Bias

And the winner is…organic!

Organic is the type of topic modeling that most folks in AI want to do, rather than canonical or entity-centric. Why?

Because it's the only kind that is nothing but "crunching big data with statistics" (the go-to tactic of current-day AI).

- And yes, it is indeed a very good thing to know how to do. For many jobs in NLP, it would be considered "table stakes."

DataScience@SMU

Organic Topic Modeling

Approaches to Organic Topic Modeling

So let's do it! We will look at three ways to implement organic topic modeling:

- LSA—latent semantic analysis
- LDA—latent Dirichlet allocation*
- NMF—non-negative matrix factorization

*Not to be confused with *linear discriminant analysis*

LSA—Latent Semantic Analysis

- Intuitively, LSA-based topic modeling tries to find groups of words associated with the largest variances between documents in the corpus.
- It answers the question “What small groups of words, when found in documents, predict those documents being very different overall from the rest of the corpus?”
- Because of this, it is unlikely that topics will share keywords (compared to what LDA produces).

LSA Intuitively

LSA starts with a large term-document matrix (usually populated with tf-idf values), but after assigning words to topics, it then uses a topic-to-topic matrix to determine what will maximize the variance between documents.

Topics that don’t do a good job of separating documents are adjusted until they do so.

LSA Intuitively

Topics that don't do a good job of separating documents are adjusted until they do so.

Topic A vs. Topic B viewed by how many shared documents there are

LSA Intuitively

Imagine we have food menus from a variety of restaurants.

Suppose that in these menus, whenever we see “fish” we usually see “chips, and vice versa.

It means having both those as features for a topic doesn’t help that topic to separate any restaurant menus.

Mr. Chips Fish Diner

FRESH CUT FRIES

Eat In • Take Out • Delivery
Family Pack • Kids Menu • Daily Special

519-751-1141

129 Colborne St. W., Brantford

Tues–Sat 11am–8pm • Sunday 4pm–8pm

We use vegetable oil which is low in saturated fats and 100% cholesterol free!

http://mrchipsfish.goldbook.ca

Appetizers

Fisherman’s Combo 9.50
3 scallops, 3 shrimp, clams
Hand Battered Shrimp (8) ... 9.50
Scallops (8) 8.95
Clam Strips 8.95
Mozzarella Sticks (7) 9.95
“FRESH” Battered Mushrooms 5.45
Battered Vegetables 7.50
Clam Chowder Bowl 4.50
Jalapeno Shrimp (12) 8.95
Calamari ¼ lb 8.95
Jalapeno Poppers (6) 8.95
Onion Rings 4.95

Fish & Chips

Alaskan Pollock (2pcs) 8.95
Extra piece 4.95
Halibut (4-5oz/2pcs) 16.95
Extra Piece 8.95
Haddock (2pcs) 11.95
Extra piece 6.75
Fish Only 10.95
Cod (4–6oz/1pc) 10.95
Fish Only 9.95
Perch (from Lake Erie) 14.95
Extra piece 3.00
Fish Bits 8.95

Fish & Chip Family Combo

10 Pc Alaskan Pollock
& 4 Large Chips 36.95
10 Pc Halibut
& 4 Large Chips 72.95
10 Pc Haddock
& 4 Large Chips 48.95

Entrees

Seafood Platter
(2pcs Halibut dinner, clams, scallops, shrimps) 23.95
Shrimp (8) & Chips 10.95
Jalapeno Shrimp (12) & Chips 9.95
Calamari ¼ lb & Chips 9.95
Chicken Fingers & Chips 9.45
Extra piece 3.35
Chicken Wings & Chips 9.50
Extra piece 1.75
Scallops & Chips (8) 9.50

Smaller Portions

Halibut & Chips (1pc) 10.25
Alaskan Pollock & Chips (1pc) 6.95
Perch & Chips (3pcs) 8.45
Haddock & Chips (1pc) 8.25
Extra piece Haddock 6.75
Shrimp (4) & Chips 7.25
Chicken Fingers & Chips (2pcs) 7.25
Chicken Wings & Chips 6.50

Mr. Chips Kids Menu

Dixi Dog & Chips 6.95
Fish & Chips 6.95
Chicken Fingers & Chips 6.95
Chicken Nuggets & Chips 6.50
Fish Bits & Chips 6.95

To The Side

Fresh Cut Fries
Small 2.95
Large 4.75
Extra Large 6.95
Mashed Potatoes 3.75
Poutine 6.95
Homemade Onion Rings 5.45
Dixi Dog 4.95
Coleslaw ..... small 1.50
..... large 2.25
Gravy ..... small 1.30
..... large 2.25
Bread 1.00
(white or whole wheat)

Halibut Tuesdays
2 Can Dine For $33.99
2 Pcs Halibut & Chips, Drink & Coleslaw

Join us for
All You Can Eat Wednesdays only 11.95
Eat In Only

Daily Lunch Special $6.95
2 Small Pcs of Pollock & Chips
Eat In or Take Out • 11am – 2pm

FRESH CUT FRIES

LSA Intuitively

Further suppose that the word “onions” is in the vast majority of the menus.

Obviously it doesn’t help much in separating menus from one another.

In this way you can see that we intuitively can find a subset of features, much smaller than the global set, that serves best to maximize variance from one topic to the next.

Joe’s Burger Shack

1. Blah blah with **onions** $5
2. Blah and onions and blah $6

Jose’s Shrimp Shack

1. Blah and diced onions $4
2. Blah blah with green **onions** and blah $7

Giuseppe’s Pasta Shack

1. Blah and stewed onions $8
2. Blah marinated **onions** $3

LSA Intuitively

LSA is built for systemically making these discoveries for us, organizing the salient features into a manageable number of vectors.

When LSA is executed, we can take its top *k* singular vectors to get “latent topics” that achieve maximum variance.

Philosophy of Organic Topic Modeling

Yes, there's a bit of philosophy to this.

All organic topic modeling operates on two fundamental assumptions:

1. What ordinarily separates documents in a collection are their *topical differences,* and
2. A topic can be fairly represented as a *mixture of words.*

DataScience@SMU

Topic Modeling with LDA

LDA—Latent Dirichlet Allocation

Intuitively, LDA construes topics as groups of words that have high co-occurrences among different documents in the corpus.

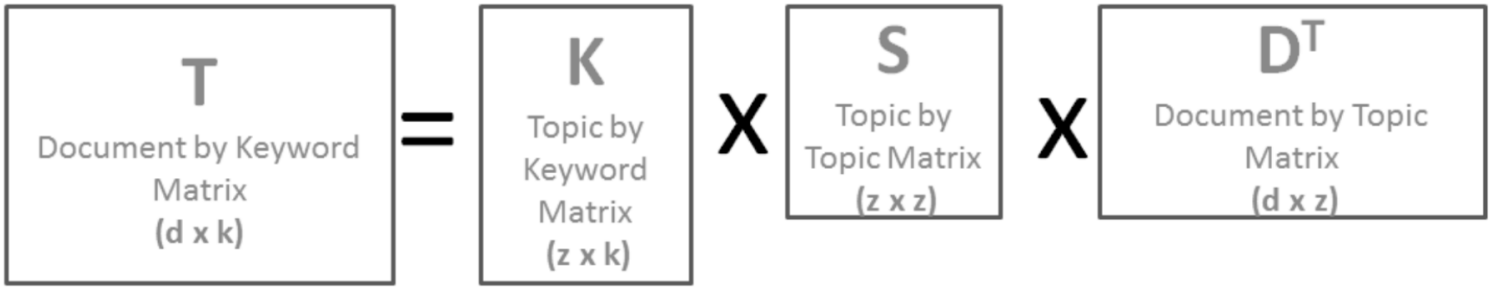
Different topics can share keywords when those keywords co-occur frequently enough across the different topics.

Peter Gustav Lejeune Dirichlet

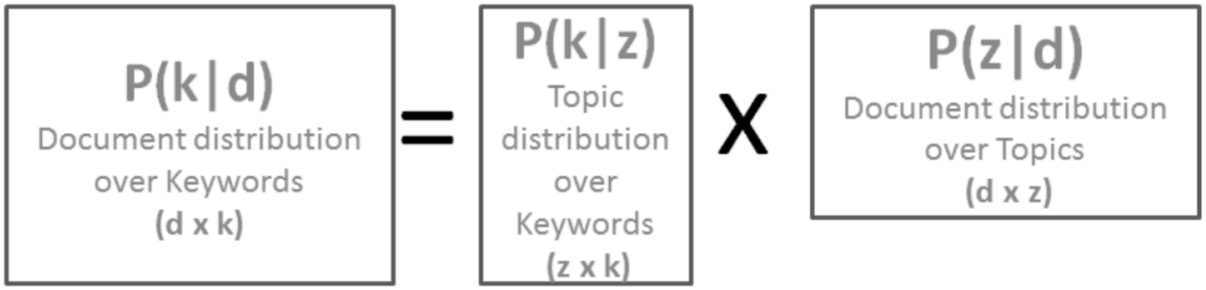
Comparing LSA and LDA

LDA is akin to LSA except in the way that it utilizes probability distributions over words, rather than a topic-topic matrix, to guide the assignment of words to topics.

LSA



LDA



Bergamaschi & Po 2015

LDA Intuitively

To oversimplify, LDA initially assigns a random topic to each word in every document, then iterates the following:

1. For every word *z* in every document *d*, compute:

   P(z|d) = proportion of words in document *d,* assigned topic *z*

   P(k|z) = proportion of assignments to topic *z* among all docs having word *k*

2. Now reassign word *k* in document *d* to whatever topic *t* gives the highest *P(z|d) x P(k|z)* from all topics that you can substitute for *k.*

Now repeat steps these steps, again and again.

LDA Intuitively

1. For every word *z* in every document *d*, compute:

   P(k|d) = proportion of words in document *d,* assigned topic *k*

   P(k|z) = proportion of docs having word *z,* assigned topic k
2. Now reassign word *z* in document *d* to a whatever topic *t* gives the highest *P(k|d) x P(k|z)* from all topics that you can substitute for *k*

Now repeat these steps, again and again.

With each iteration, words assigned to the same topic start to be ones that “belong together” and do not appear random.

This makes sense because of step 2. Words tend to be reassigned to the same topics as other words they co-occur with.

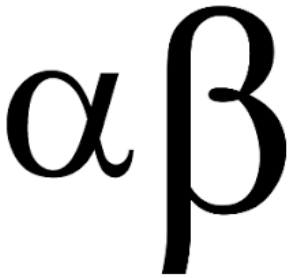
Concentration Parameters in LDA

There are two very important *concentration parameters* in LDA:

- The *alpha* parameter
  - A high alpha-value means that each document is likely to contain a mixture of many topics.
  - A low alpha value means it is more likely that a document may contain just a few topics, or only one.
- The *beta* parameter
  - Similarly, a high beta-value means that each topic is likely to contain a mixture of many words.
  - A low value means that a topic is more likely to contain just a few words.

Concentration Parameters in LDA

As a consequence of the foregoing, in practice:



- A high alpha-value will lead to documents being more similar in terms of what topics they contain.
- A high beta-value will similarly lead to topics being more similar in terms of what words they contain.

Pros and Cons of LSA vs. LDA

From anecdotal experience…

Model	Tends to Separate Word Senses?	Time to Train	Time to Run	Typical Number of Topics for Optimal Results
LSA	N	½x	2x	300-500
LDA	Y	1x	1x	30-50

Which is more accurate? You'll hear opposite opinions, usually from practitioners with different applications.

NMF—Non-Negative Matrix Factorization

- For all practical purposes, you can view NMF as a version of LDA in which the parameters have been tweaked to enforce a sparse number of topics.
- Another way to say it, is that NMF naturally produces sparse representations.

DataScience@SMU

Topic Modeling with NMF

NMF—Non-Negative Matrix Factorization

- The inherent sparseness of NMF means it's not the best solution for finding lots of topics in long documents, but it is well suited to handling projects where all the documents are very short.
- If you try to set a large number of topics for NMF, it tends to produce nonsensical topics (awkward combinations of words with little semantic coherence).

NMF vs. LDA

- NMF is usually much cheaper in computation than LDA
- NMF works better “out-of-the-box” on noisy texts (does not require as much tuning as LDA, in such cases).
- NMF often works better on very small corpora than either LSA or LDA.

NMF vs. LDA

- Producing fewer topics per document, as NMF does, is thought to be more *psychologically* accurate, i.e., more similar to how humans judge the topicality of documents.
- LDA tends to produce more usable results than NMF when setting for a high number of topics, which is appropriate when processing a lot of long documents.

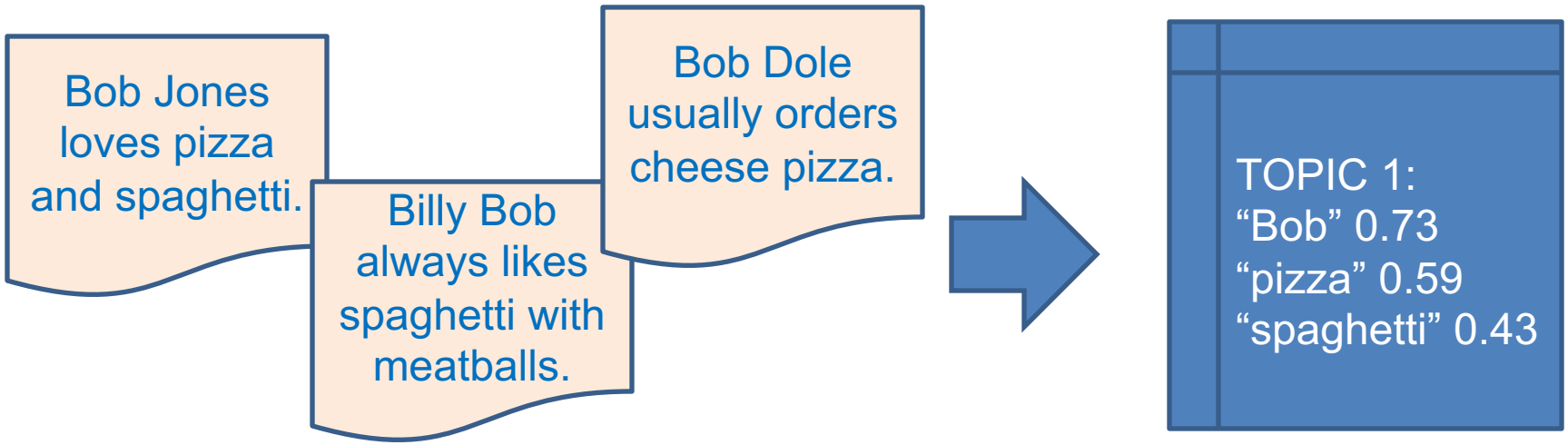
DataScience@SMU

Working with Organic Topic Models

Topic Coherence and Statistical Interference

It can be unlucky coincidence that certain mixtures of words are sufficiently associated in a corpus to manifest themselves as topics in the output of LSA, LDA, or NMF.

This phenomenon is a form of *statistical interference.* A simplified example:



Topic Coherence and Statistical Interference

To test for this, we measure the high-scoring outputted topics to see if they have internal coherence, semantically.

The higher the topic coherence, the lower the statistical interference.

Trying for too many topics, or too few, can cause this sort of interference, as can adjusting the alpha/beta parameters in LDA. So it’s important to measure coherence and experiment with the parameters of your system.

Troubleshooting Your Organic Topic Modeler

Problem: Outputted topics mostly consist of *the, a, an, in*, etc.

Solution: Don’t forget that you should have normalized the text by removing stop words.

Troubleshooting Your Organic Topic Modeler

Problem: Outputted topics mostly consist of *the, a, an, in*, etc.

Solution: Don't forget that you should have normalized the text by removing stop words.

Problem: After removing stop words, all my topics still look too much the same, e.g., they all have "football, games, tickets..."

Solution: Reduce the *alpha* concentration parameter in LDA, and/or normalize your text again by adding high document-frequency words to your stop word list.

Troubleshooting Your Organic Topic Modeler

Problem: Outputted topics mostly consist of *the, a, an, in*, etc.

Solution: Don't forget that you should have normalized the text by removing stop words.

Problem: After removing stop words, all my topics still look too much the same, e.g., they all have "football, games, tickets..."

Solution: Reduce the *alpha* concentration parameter in LDA, and/or normalize your text again by adding high document-frequency words to your stop word list.

Problem: I don't know the right number of topics to set.

Solution: Do UAT ("user-acceptance testing"). Refine your use case with a stakeholder and show them a few samples of early, representative results at a few different topic settings.
"The customer is always right."

Applications of Topic Modeling

One big area of application is that of automated recommenders:

- Movie recommender
- News article recommender
- Book recommender
- Dating-website match recommender

…and anything else where there's a large collection of unstructured texts to be matched on a small number of texts (or just one).

DataScience@SMU

Canonical Topic Modeling

Canonical Topics

“canonical”

[kuh-**non**-i-kuh l]

1. pertaining to, established by, or **conforming to a canon**
2. included in the canon of the Bible
3. **authorized; recognized; accepted**: as in *canonical works*

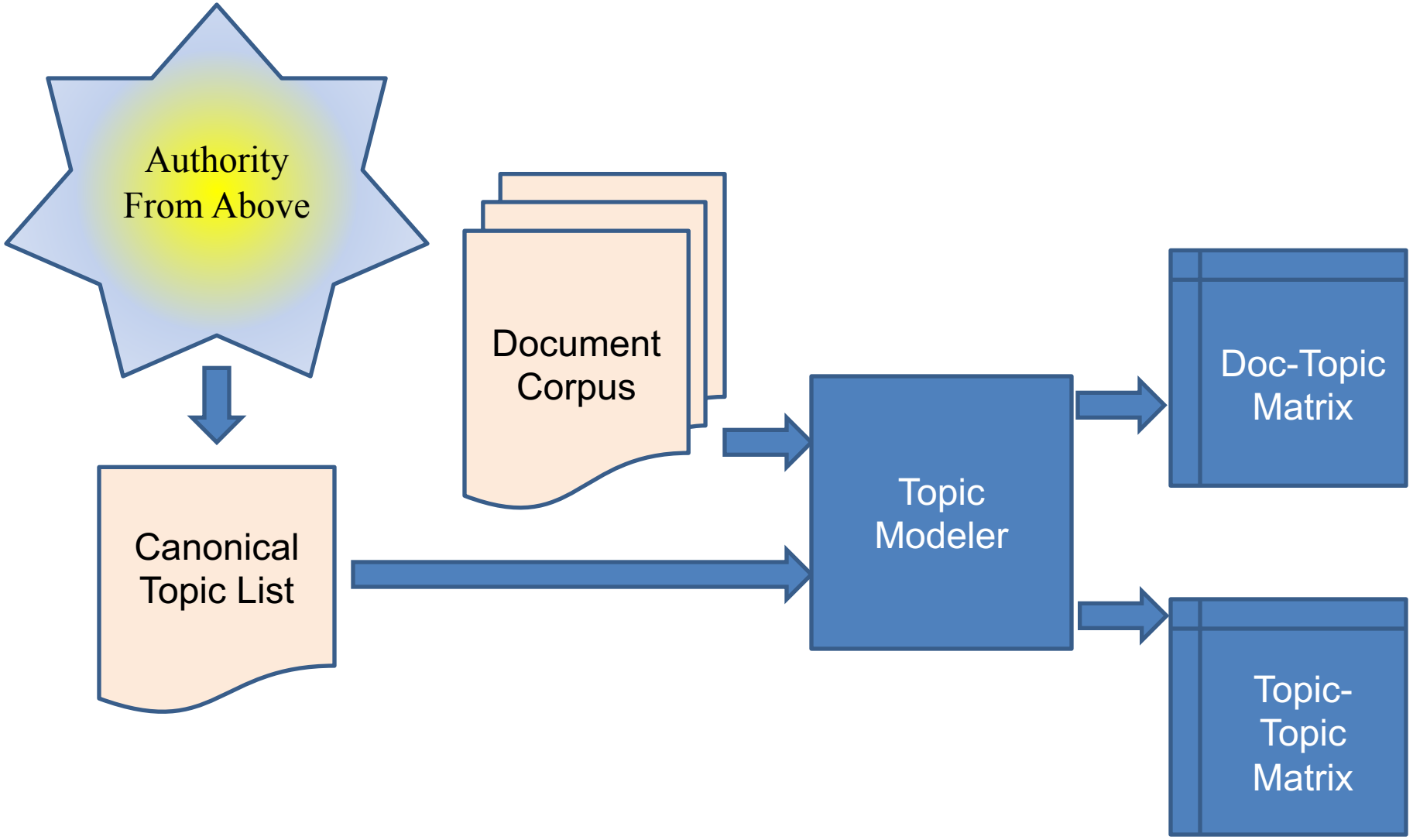
“canon”

[**kan**-uh n]

…

11. **a catalog or list**, as of the saints acknowledged by the Church

Canonical Topic Modeling



Canonical Topic Modeling

Isn’t this just classification?

No.

Because the goal is to know any case where a topic is materially treated in a document, even if the document is not primarily “about” that topic.

Canonical Topic Modeling

Then isn't this just keyword tagging?

No, because:

(a) the goal is to treat a topic as a mixture of words, not simply one word

(b) the goal is also to perform corpus-based intertopic modeling

Canonical Topic Modeling

So what is it?

The goal of canonical topic modeling is to

determine a *subset of canonical topics*

that are *materially treated* in a given corpus,

showing which topics are *contextually related* in that corpus.

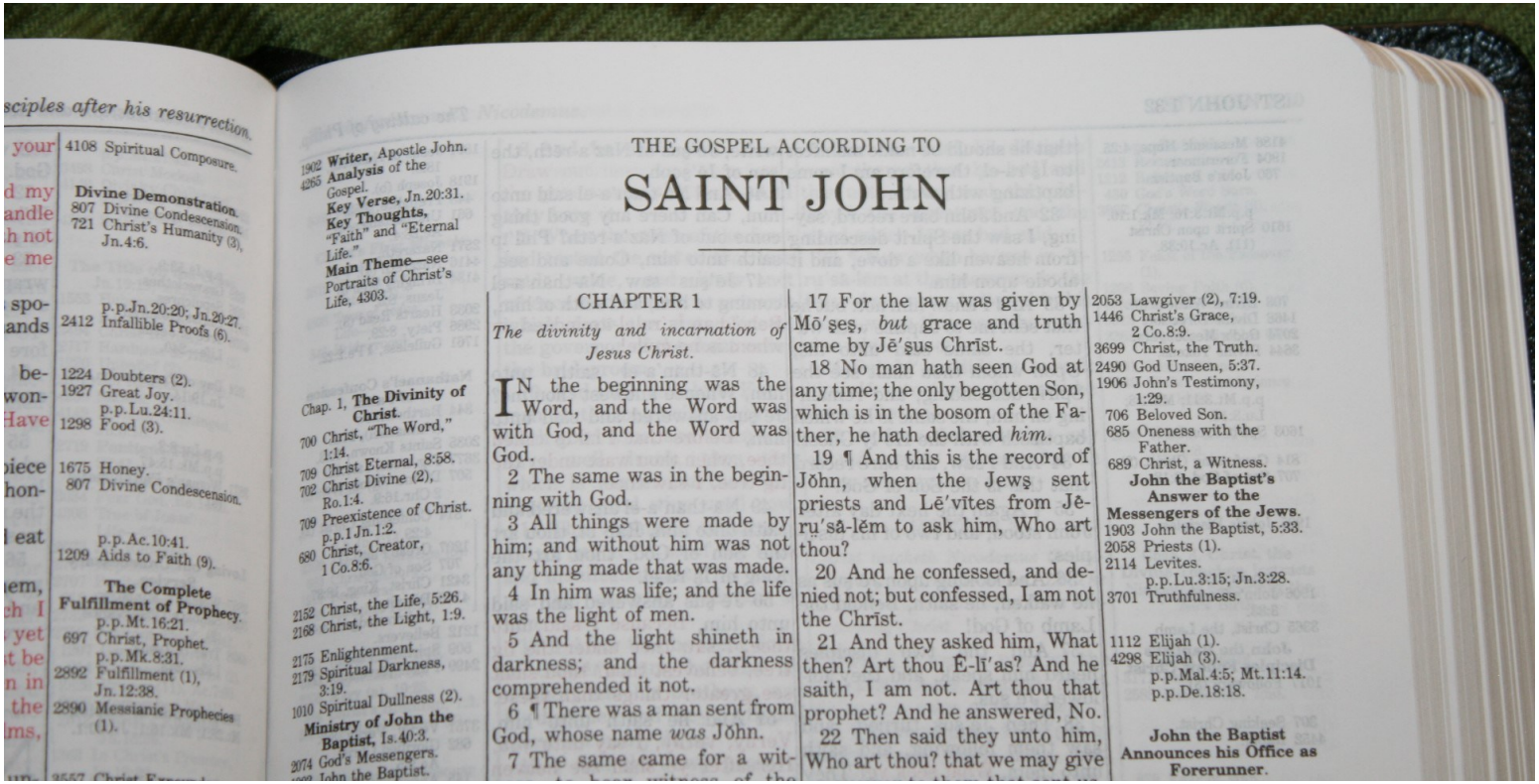
The driving concern is usually to enable topic-driven exploration of the corpus by end users.

Canonical Topic Modeling

- So if we're given "injuries" as a canonical topic in a sports domain, our primary task is to discover the mixture of words or phrases that constitute this topic across all documents, e.g., "concussion," "hyperextension," "sprain," "collapsed," etc.
- Secondarily, we want our model to provide the contextual topics in documents that have the "injuries" topic, e.g., "rehab," "surgery," etc.

Canonical Topic Modeling: Examples

Some of the best examples are those made by hand, e.g., by literary critics or biblical scholars.



sciples after his resurrection.

your
d my
andle
h not
e me
spo-
ands
be-
won-
Have
piece
hon-
l eat
hem,
ch I
yet
t be
n in
the
lms,

4108 Spiritual Composure.

Divine Demonstration.
807 Divine Condescension.
721 Christ's Humanity (3), Jn.4:6.

p.p.Jn.20:20; Jn.20:27.
2412 Infallible Proofs (6).

1224 Doubters (2).
1927 Great Joy.
p.p.Lu.24:11.
1298 Food (3).

1675 Honey.
807 Divine Condescension.

p.p.Ac.10:41.
1209 Aids to Faith (9).

The Complete Fulfillment of Prophecy.
p.p.Mt.16:21.
697 Christ, Prophet.
p.p.Mk.8:31.
2892 Fulfillment (1), Jn.12:38.
2890 Messianic Prophecies (1).

1902 **Writer,** Apostle John.
4265 **Analysis** of the Gospel.
Key Verse, Jn.20:31.
Key Thoughts, "Faith" and "Eternal Life."
Main Theme—see Portraits of Christ's Life, 4303.

Chap. 1, **The Divinity of Christ.**
700 Christ, "The Word," 1:14.
709 Christ Eternal, 8:58.
702 Christ Divine (2), Ro.1:4.
709 Preexistence of Christ. p.p.1 Jn.1:2.
680 Christ, Creator, 1 Co.8:6.

2152 Christ, the Life, 5:26.
2168 Christ, the Light, 1:9.

2175 Enlightenment.
2179 Spiritual Darkness, 3:19.
1010 Spiritual Dullness (2).
Ministry of John the Baptist, Is.40:3.
2074 God's Messengers.

THE GOSPEL ACCORDING TO

SAINT JOHN

CHAPTER 1

The divinity and incarnation of Jesus Christ.

IN the beginning was the Word, and the Word was with God, and the Word was God.

2 The same was in the beginning with God.

3 All things were made by him; and without him was not any thing made that was made.

4 In him was life; and the life was the light of men.

5 And the light shineth in darkness; and the darkness comprehended it not.

6 ¶ There was a man sent from God, whose name *was* Jōhn.

7 The same came for a wit-

17 For the law was given by Mō′şeş, *but* grace and truth came by Jē′şus Chrīst.

18 No man hath seen God at any time; the only begotten Son, which is in the bosom of the Father, he hath declared *him*.

19 ¶ And this is the record of Jōhn, when the Jews sent priests and Lē′vītes from Jē-ru′sā-lēm to ask him, Who art thou?

20 And he confessed, and denied not; but confessed, I am not the Chrīst.

21 And they asked him, What then? Art thou Ē-lī′as? And he saith, I am not. Art thou that prophet? And he answered, No.

22 Then said they unto him, Who art thou? that we may give

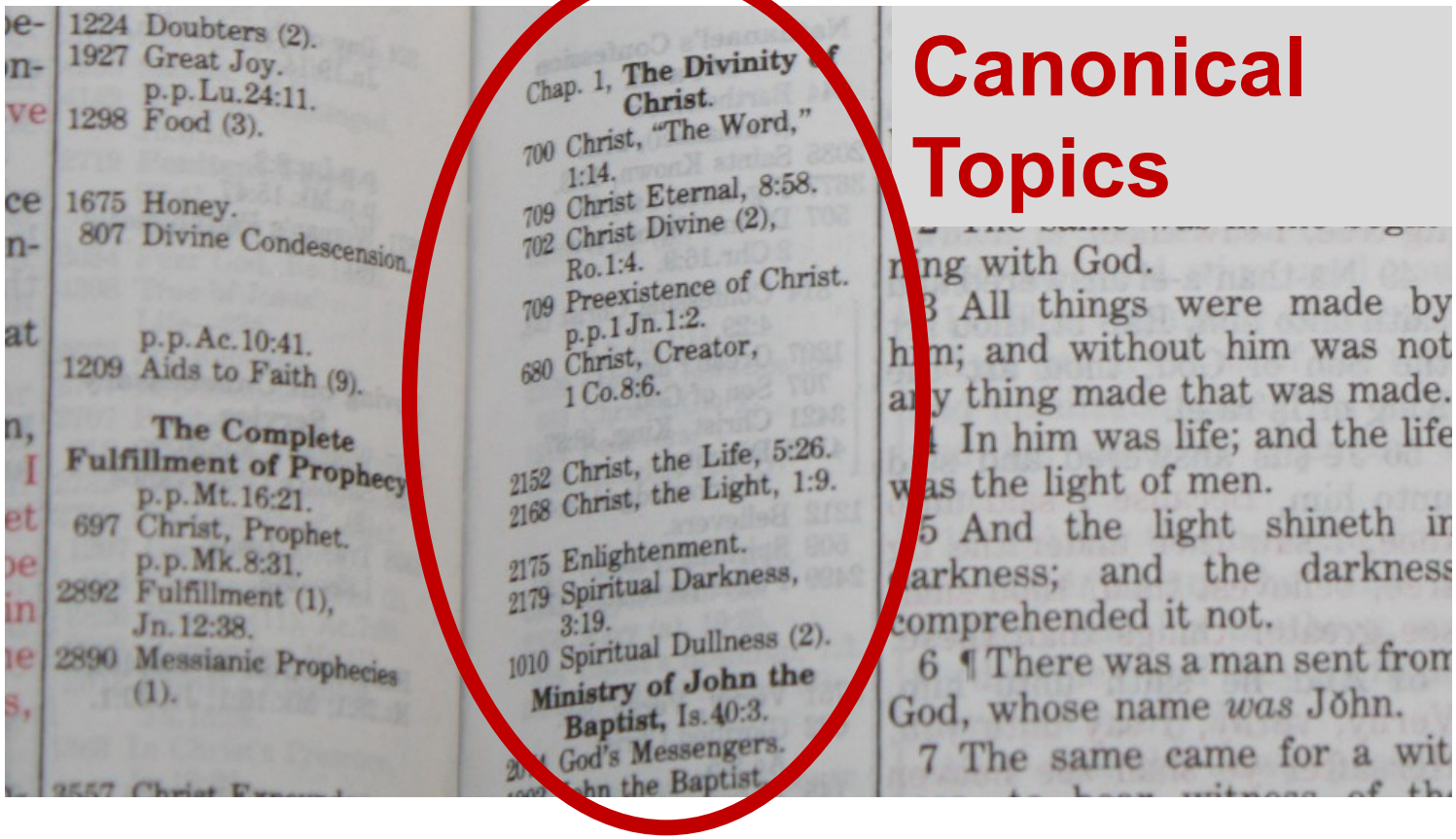
2053 Lawgiver (2), 7:19.
1446 Christ's Grace, 2 Co.8:9.
3699 Christ, the Truth.
2490 God Unseen, 5:37.
1906 John's Testimony, 1:29.
706 Beloved Son.
685 Oneness with the Father.
689 Christ, a Witness.
John the Baptist's Answer to the Messengers of the Jews.
1903 John the Baptist, 5:33.
2058 Priests (1).
2114 Levites. p.p.Lu.3:15; Jn.3:28.
3701 Truthfulness.

1112 Elijah (1).
4298 Elijah (3). p.p.Mal.4:5; Mt.11:14. p.p.De.18:18.

John the Baptist Announces his Office as Forerunner.

Canonical Topic Modeling: Examples

Some of the best examples are those made by hand, e.g., by literary critics or biblical scholars.



Canonical Topic Modeling: Examples

A good index in the back of a book has references to pages and very ample “and…” listings as well as very ample “c.f.” listings.

DataScience@SMU

Approaches to Canonical Topic Modeling

Canonical Topic Modeling

There are two basic options how to proceed:

1. Constrain the organic topic model to the canonical list of topics.
2. Use an IR approach, leveraging the canonical topic list to build queries, and analyze the hits.

Canonical Topic Modeling

First approach: constrain the organic topic model to the canonical list of topics.

There are, of course, risks.

- Your fancy LDA machine might not assemble topic elements in a way that your Authority From Above deems acceptable/useable.
- It also might not surface weakly represented topics that the Authority From Above still cares about despite their paucity.

Canonical Topic Modeling

Second approach—Use a semantic IR architecture, and leverage the canonical topic list to build queries.

- When we see that two topic words used as queries share a lot of the same hits, it shows us that those topics are related—this is the ***extensional*** approach.
- Topics that have close semantic distance are also related—this is the ***intensional*** approach.

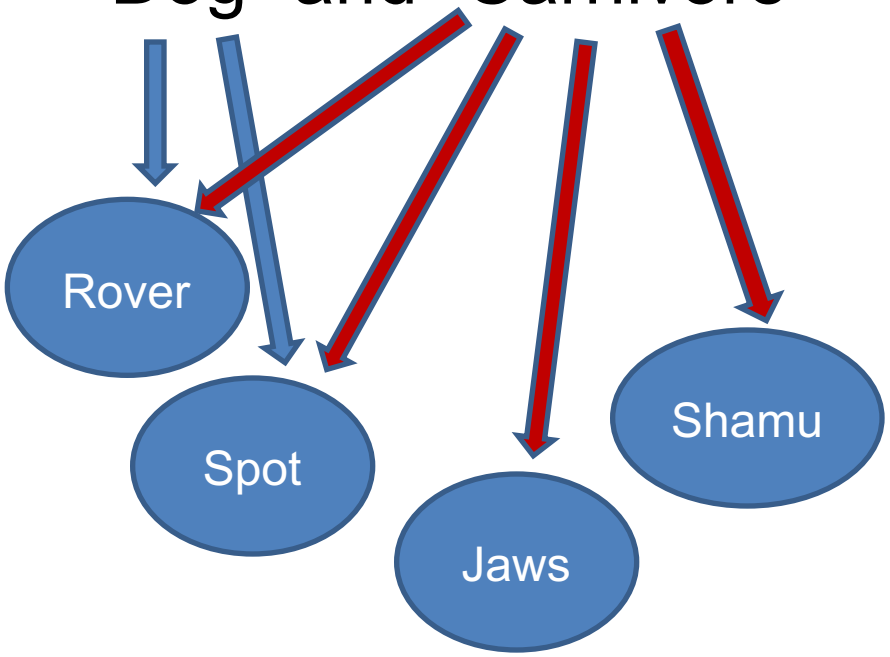
Extension vs. Intension

- Concepts are extensionally related when they extend to some of the same referents in the world.
- Concepts are intensionally related when their meanings (definitions) overlap.

Extension vs. Intension

Extensional Relation

"Dog" and "Carnivore"



Intensional Relation

"Dog (def.): quadrupedal terrestrial **carnivorous animal**"

"**Carnivore** (def.): meat-eating **animal**"

DataScience@SMU

Semantic Analysis: Entity-Centric Topic Modeling

Natural Language Processing

Entity-Centric Topic Modeling

In real life, your mission statement for topic modeling often starts from little more than this:

Entity-Centric Topic Modeling

- Your mission is to model whatever topics are strongly related to a set of named entities in a domain.
  - Examples are websites that aggregate news on one of these: NFL, MLB, NBA, NHL, FIFA, etc.
  - Also news aggregators for movies, TV series, music groups, etc.

For convenience we'll just say "entities" going forward, to mean "named entities."

Entity-Centric Topic Modeling

You're asked to organize all the

- teams, countries, shows, leagues, years, seasons, franchises, etc.,
- by whomever the heads of state, actors, directors, coaches, players, agents, singers, drummers, etc., are
- as well as discovering the big topics, such as treaties, contracts, hiring, firing, renewals, sanctions, events, tickets, endorsements, etc.

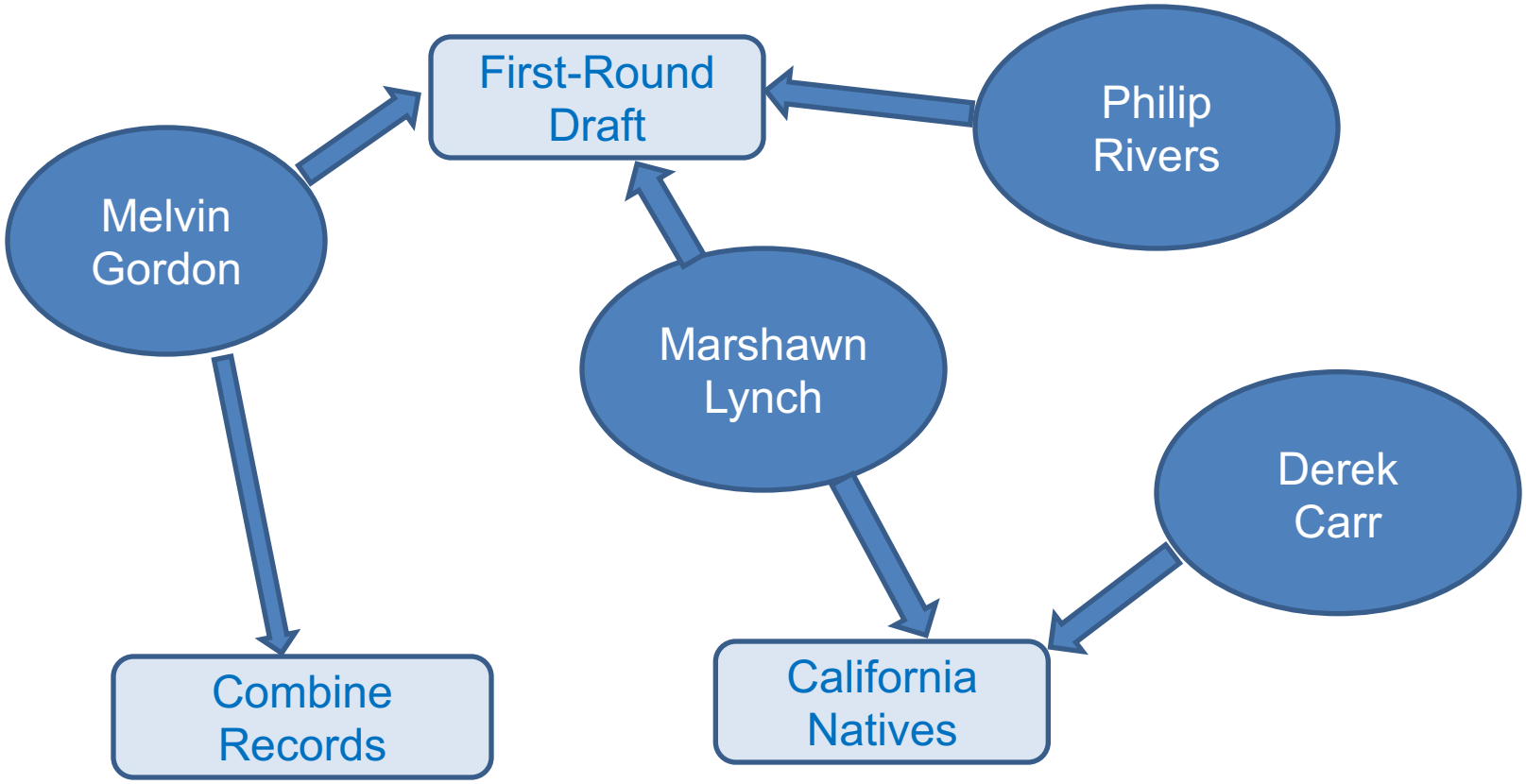
Entity-Centric Topic Modeling

You are given this…

Quarterback	Lynch	Rivers	Carr	Mahomes
Halfback	Booker	Gordon	Lynch	Hunt
Fullback	Janovich	Watt	Smith	Sherman

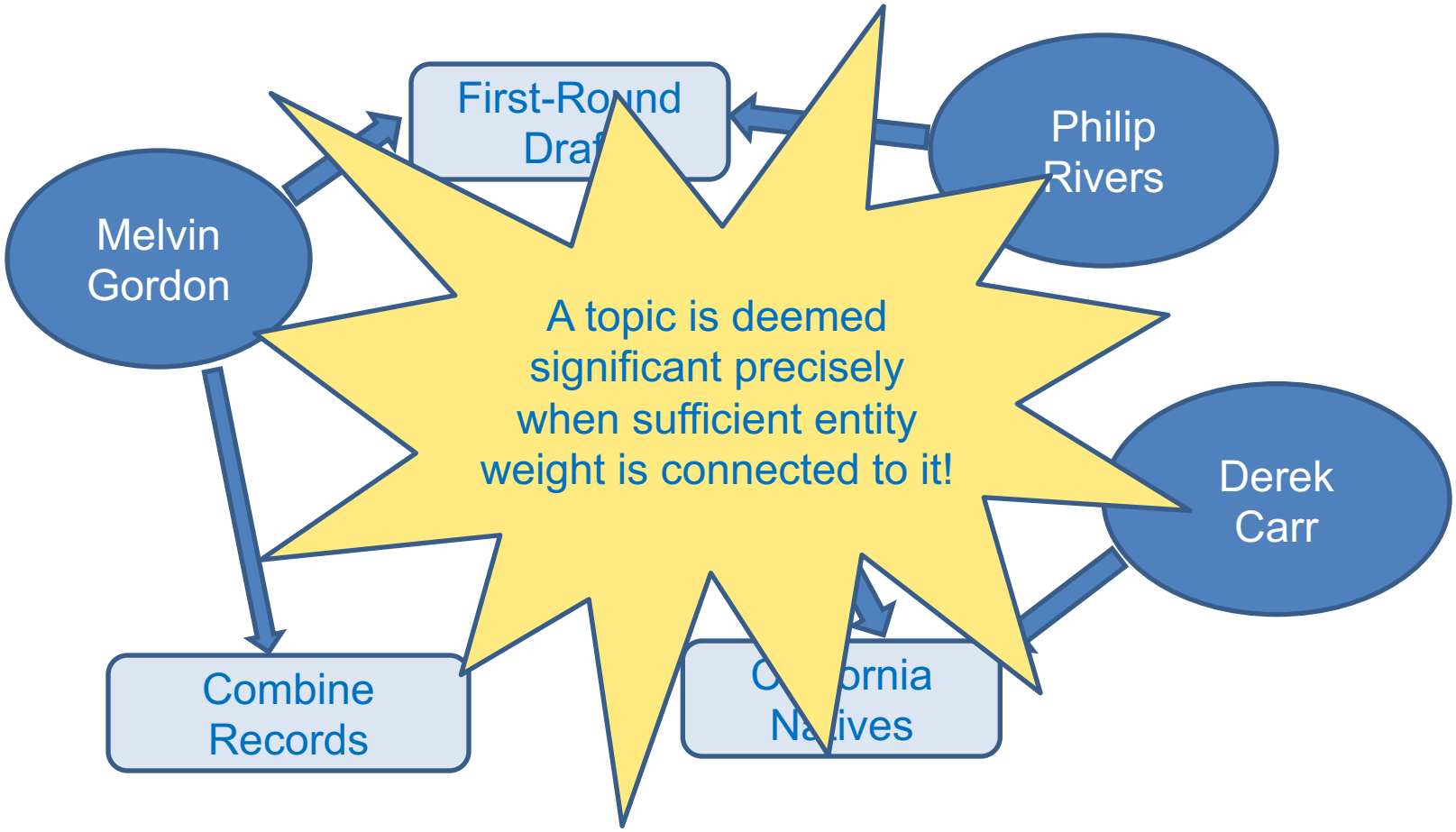
Entity-Centric Topic Modeling

And are asked to use it to create this…



Entity-Centric Topic Modeling

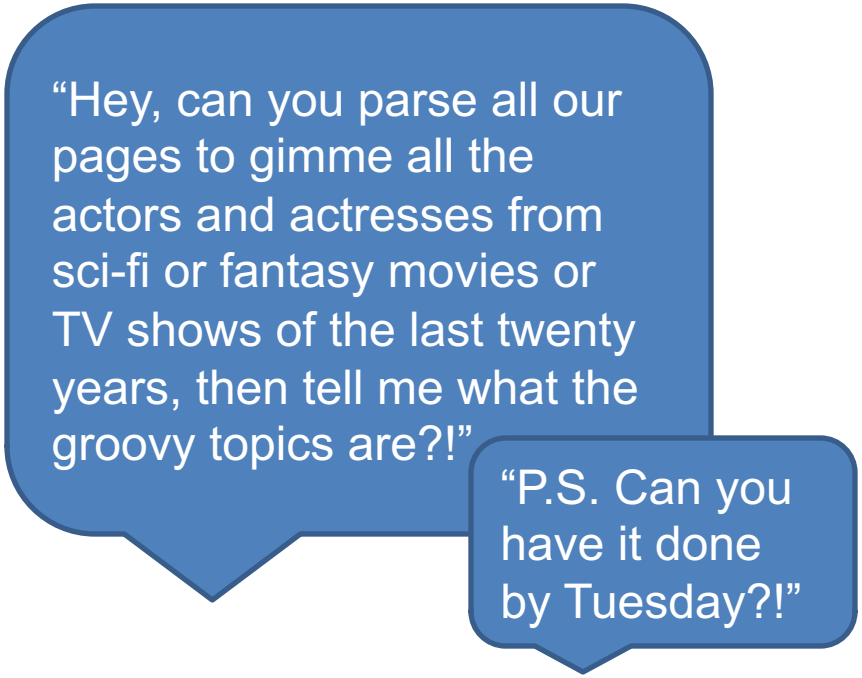
The pivotal principle of the entire model….



Two Ways of Establishing the NE List

The business owner will determine the NEs in one of two ways:

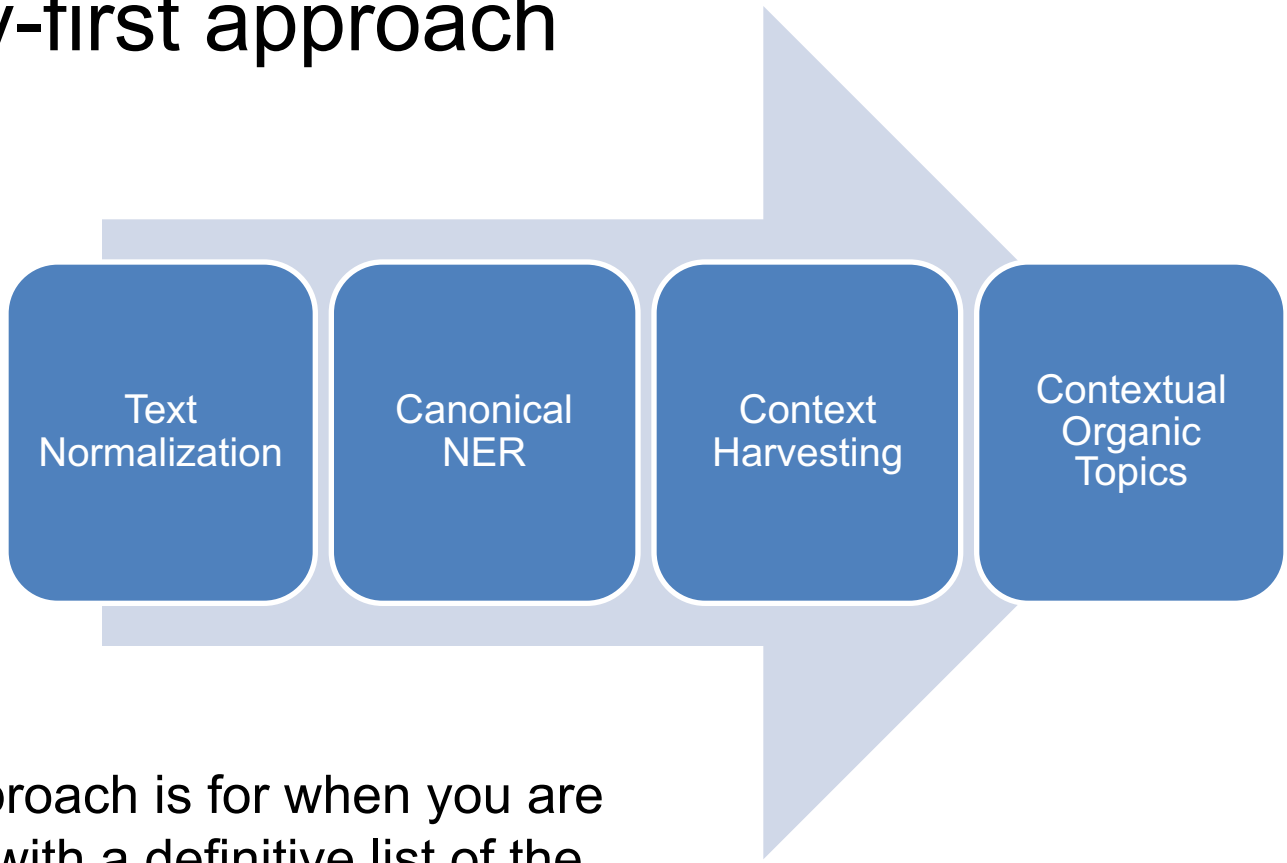
- By reference—they are simply listed (nice!)
- By description—a universal quantifier is applied to a predicate (more difficult!)



Corey C. Comey, COO of CooCoo4ComiCon.com

Entity-Centric Topic Modeling Process

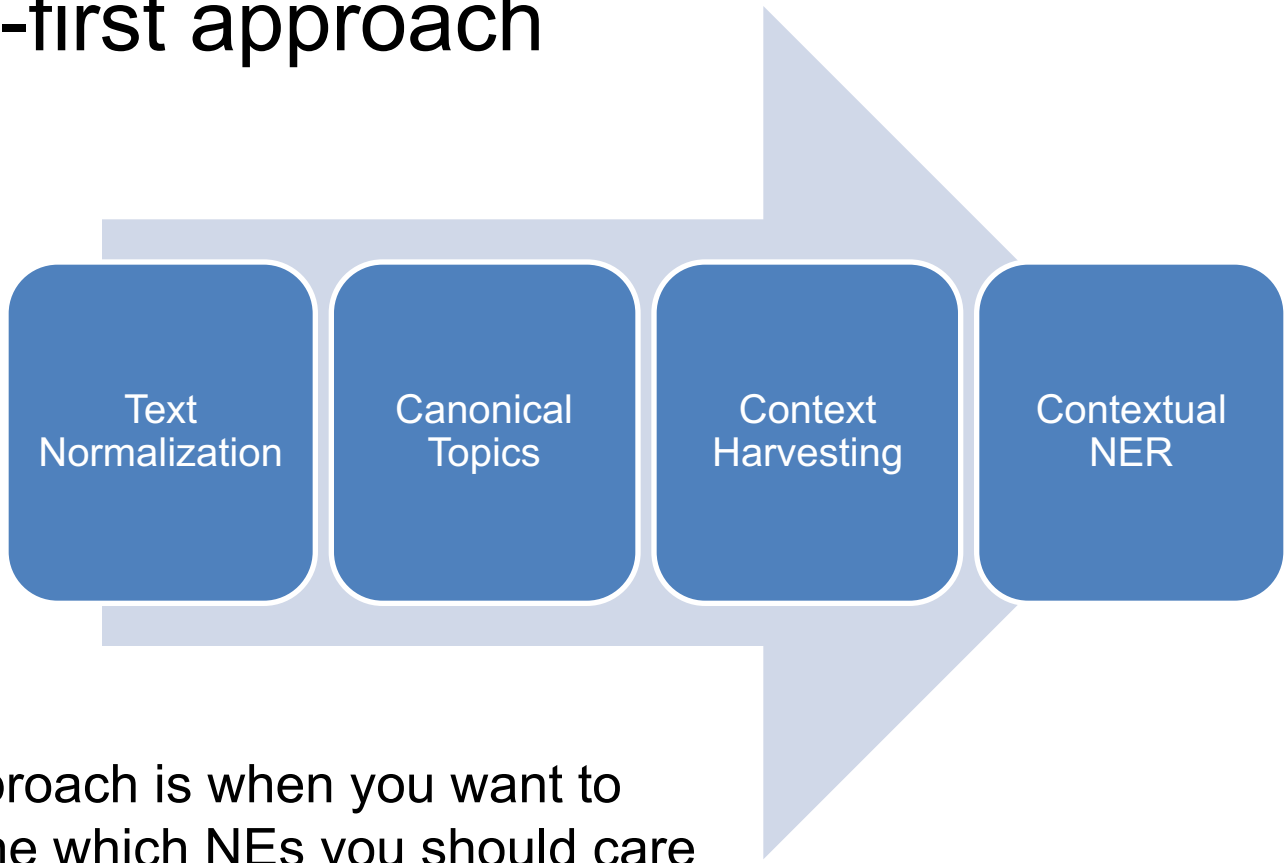
Entity-first approach



This approach is for when you are starting with a definitive list of the NEs you care about.

Entity-Centric Topic Modeling Process

Topic-first approach



This approach is when you want to determine which NEs you should care about, based on their association with canonical topics.

Entity-Centric Topic Modeling

The two approaches can be combined. Your marching orders from your business owner might be:

1. “Here’s my list of definitive topics; go find the people associated with them.”
2. “But now, please find other topics that are significant (useful for differentiating those people).”

Don’t Be Afraid of Curation

- All the approaches to automated topic modeling can produce less-perfect-than-human results.
- You don’t have to accomplish 100% automation!
- Just get the job 90% accomplished by AI, then let human curators “cover the last mile”—this is actually a huge win!

Don’t Be Afraid of Curation

“But if I depend on curation, then I’m not a purist anymore! I wouldn’t have a pure statistical system that can be baselined against other systems, so I won’t get a publication of the kind that all my professors had me read in school…”

But you know what you’ll probably get instead?
A raise and a promotion in a company whose management now sees you as a *problem solver*.

DataScience@SMU