TEXT MINING FOR TAXONOMY CONSTRUCTION

Using Text Analytics for Term Discovery

Bob Kasenchak
Director of Business Development
Access Innovations
@taxobob
www.accessinn.com
bob_kasenchak@accessinn.com

STM Annual U.S. Conference 2016
Washington, D.C.
Text and Data Mining Panel

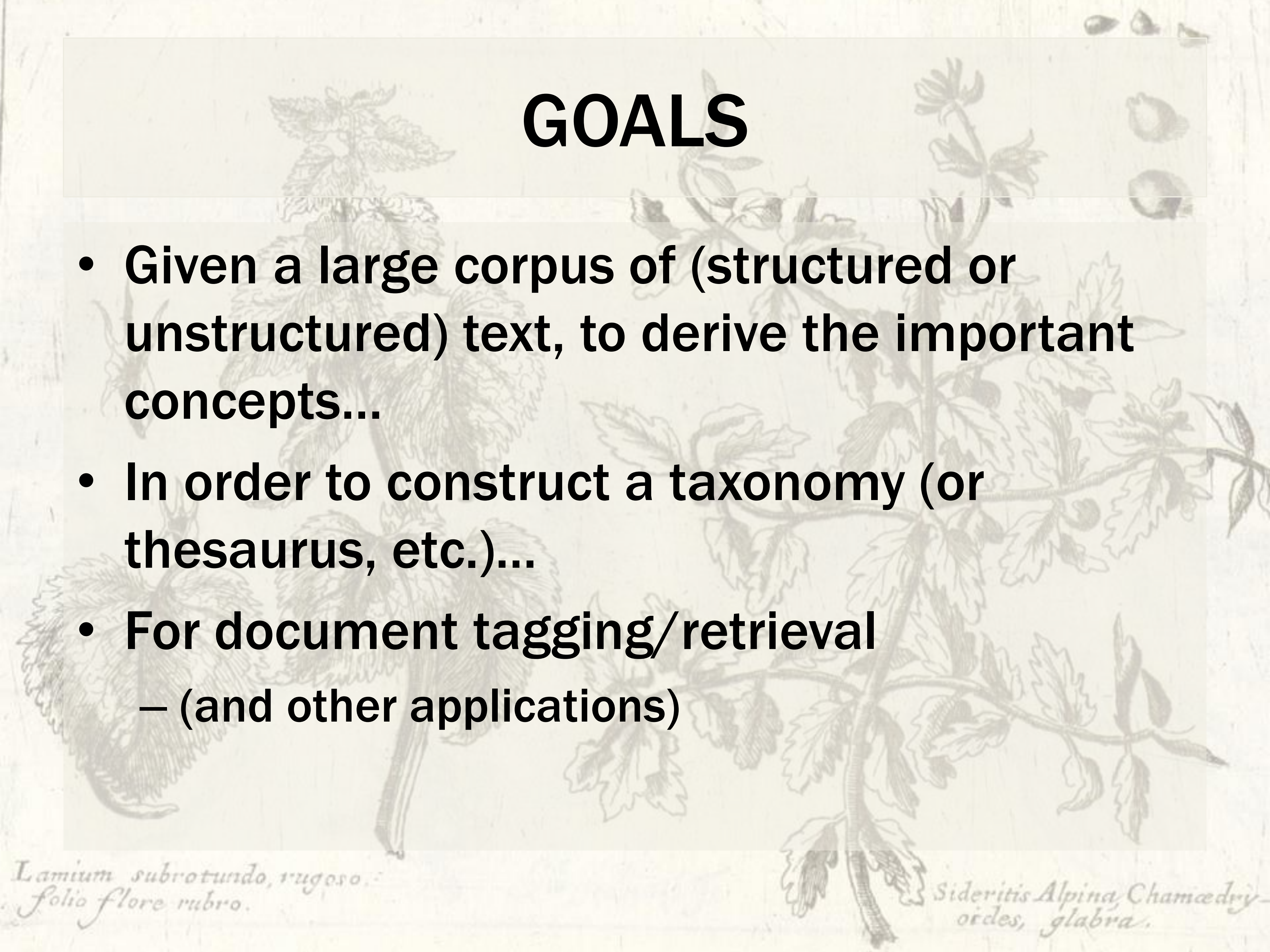
Taxonomists Hate Him!

Discover thousands of
terms using this
one WEIRD trick!

LEARN THE TRUTH NOW

GOALS

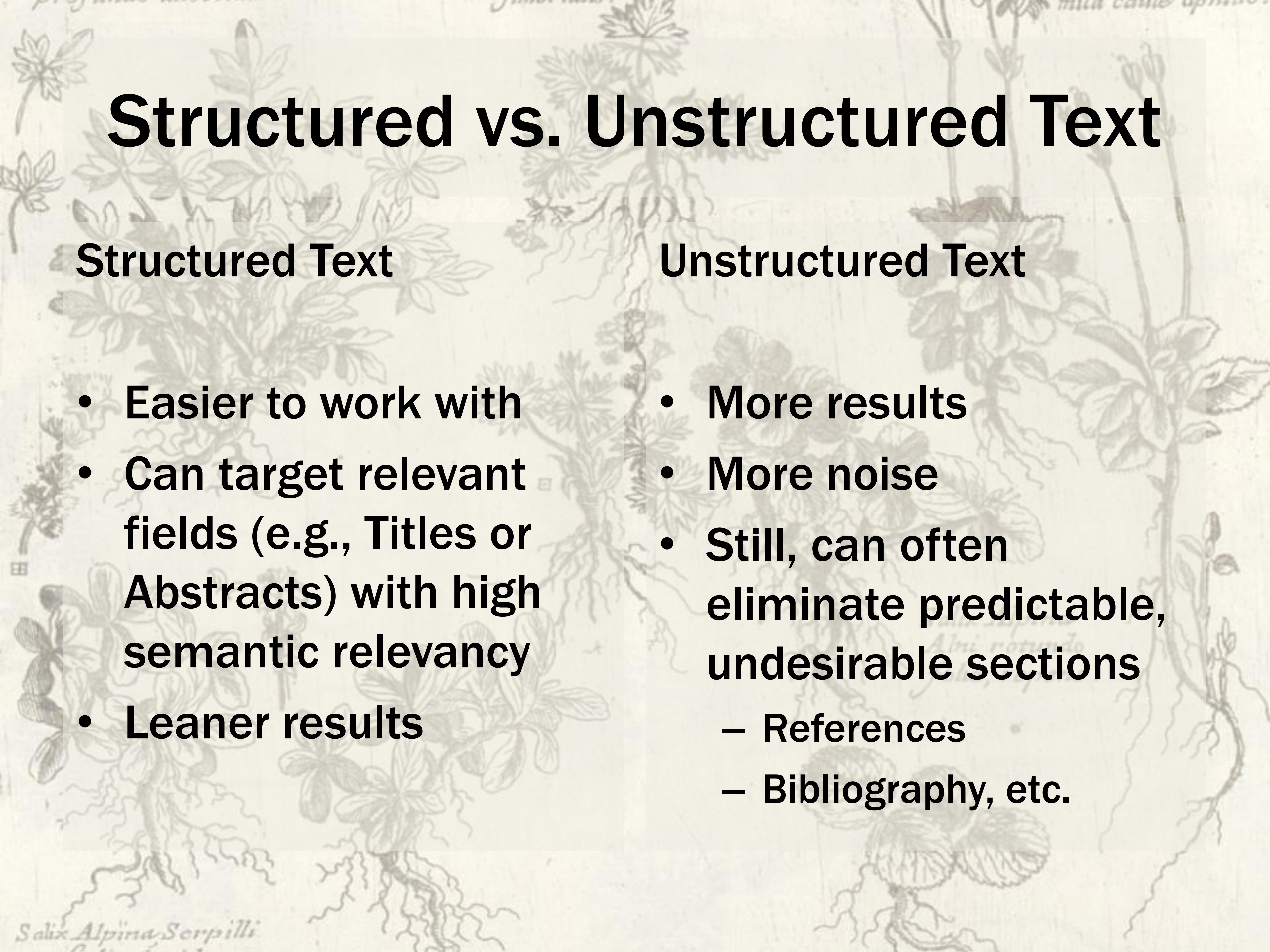
- Given a large corpus of (structured or unstructured) text, to derive the important concepts…
- In order to construct a taxonomy (or thesaurus, etc.)…
- For document tagging/retrieval
  - (and other applications)


Lamium subrotundo, rugoso. folio flore rubro.
Sideritis Alpina Chamædryoeides, glabra.

GOALS

angiosperms
magnoliids
monocots
commelinids
eudicots
core eudicots
rosids
eurosids I
eurosids II
asterids
euasterids I
euasterids II
Top Sites
In Rumania
Packed Houses To Hear Lecture
Goldwater Lecture Enthralls Audience At Heard Museum
Crowd At Heard Museum Hears Talk By Goldwater

Structured vs. Unstructured Text



Structured Text

- Easier to work with
- Can target relevant fields (e.g., Titles or Abstracts) with high semantic relevancy
- Leaner results

Unstructured Text

- More results
- More noise
- Still, can often eliminate predictable, undesirable sections
  - References
  - Bibliography, etc.

Methodology

I. Execute *n*-gram analysis
   I. …of titles? Abstracts? Full text?

II. Frequency sorting

III. Discard noise

IV. Cut off long tail

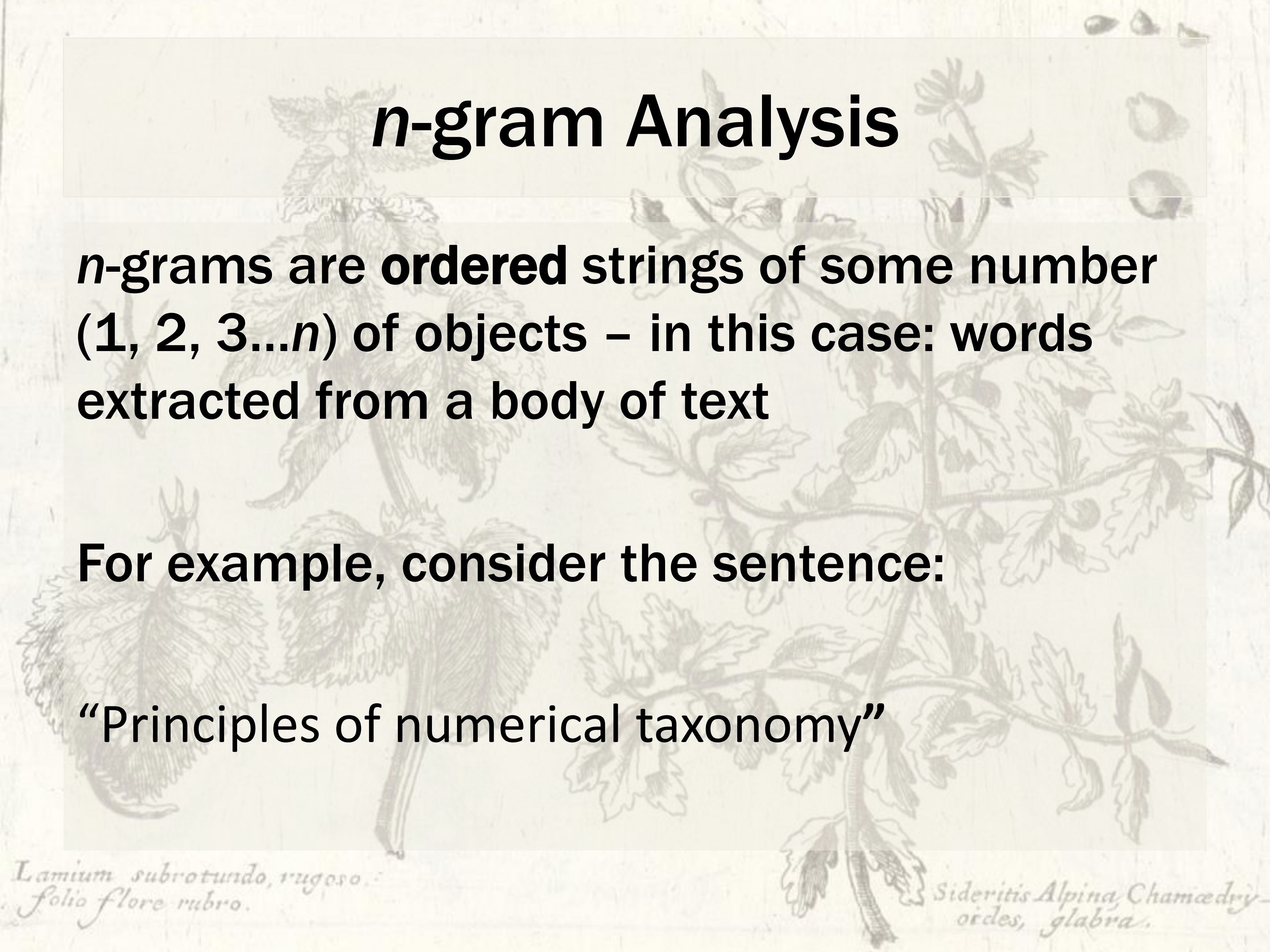
V. Human curation
   I. Identify relevant concepts (text strings)
   II. Remove conceptual duplicates
      I. Choose preferred version of concept/term
      II. Capture variants as synonyms/NPTs

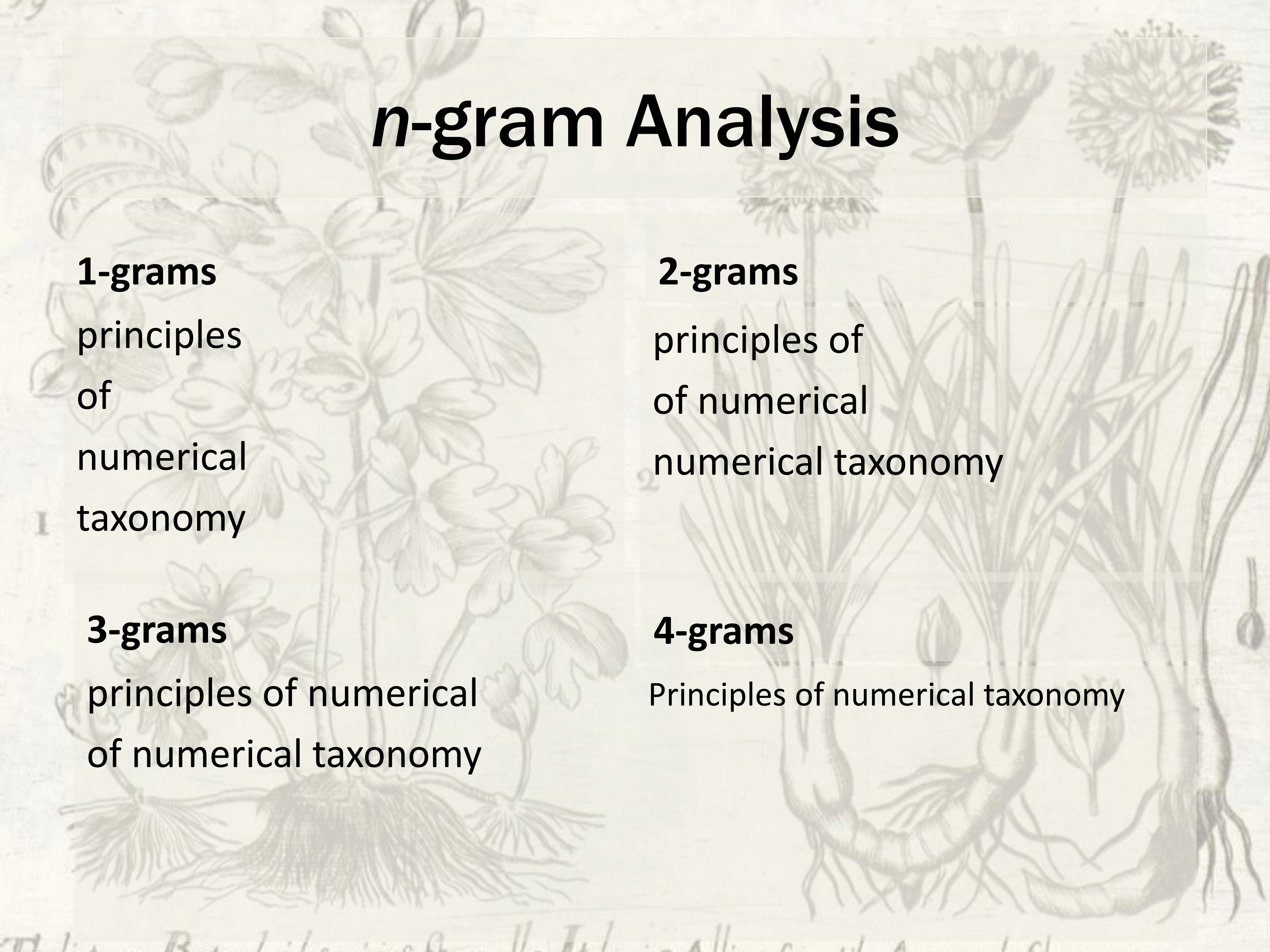
n-gram Analysis

n-grams are **ordered** strings of some number (1, 2, 3…*n*) of objects – in this case: **words** extracted from a body of text

For example, consider the sentence:

“Principles of numerical taxonomy”


Lamium subrotundo, rugoso: folio flore rubro.
Sideritis Alpina Chamædryordes, glabra.



n-gram Analysis

1-grams

principles

of

numerical

taxonomy

2-grams

principles of

of numerical

numerical taxonomy

3-grams

principles of numerical

of numerical taxonomy

4-grams

Principles of numerical taxonomy

n-gram Analysis

…but instead of a single sentence, we have, e.g.:

- 30,000 articles on healthcare
- 900,000 articles on physics (AIP)
- 65,000 standards (BSI)
- 285,000 patents (USA PO 2014 domestic filings)
- …or some other very large content corpus

n-gram Analysis

“unigrams”

Frequency	1-Grams
11415	health
9827	care
2767	medical
2264	patients
2254	quality
1955	medicare
1823	use
1521	new
1518	hospital
1480	patient
1480	insurance
1479	primary
1427	states
1369	costs
1284	spending
1265	united
1234	policy
1234	national
1219	medicaid
1196	study
1195	cost
1174	public
1153	more
1153	impact
1135	outcomes
1123	physician
1093	program
1084	disease
1075	drug

“bigrams”

Frequency	2-Grams
3041	health care
1136	primary care
977	united states
760	health insurance
450	health information
440	affordable care
436	electronic health
431	emergency department
406	systematic review
393	mental health
390	quality care
379	care act
376	public health
329	medical home
300	accountable care
287	comparative effectiveness
284	health reform
277	health spending
274	managed care
267	patientcentered medical
267	medical care
267	health system
262	information technology
251	health records
240	national health
235	health care:
226	quality improvement
226	nursing home
216	care system

“trigrams”

Frequency	3-grams
374	affordable care act
234	electronic health records
194	health information technology
190	health care system
185	health care spending
171	health care reform
169	type 2 diabetes
160	comparative effectiveness research
159	randomized controlled trial
158	patientcentered medical home
149	electronic health record
134	health care costs
133	accountable care organizations
130	primary care physicians
106	health insurance coverage
95	acute myocardial infarction
84	valuebased insurance design
84	medicare part d
80	health care delivery
76	emergency department visits
75	quality health care
74	coronary heart disease
70	under affordable care
70	protection affordable care
70	patient protection affordable
70	mental health care
70	health care quality
68	nursing home residents

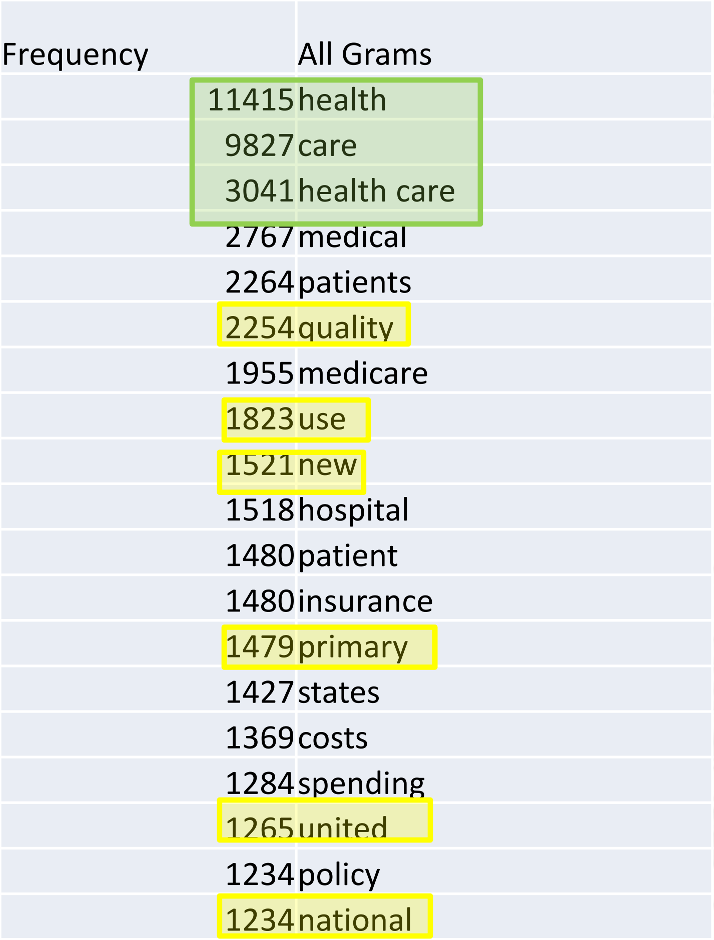
Frequency	4-grams
70	patient protection affordable care
66	under affordable care act
52	computerized physician order entry
49	use electronic health records
47	<p>the affordable care act
46	protection affordable care act
45	presidentâ€™s emergency plan aids
43	special health care needs
43	children special health care
38	regional variations medicare spending
38	implications regional variations medicare
37	triple aim: care health
37	emergency plan aids relief
37	aim: care health cost
36	childrenâ€™s health insurance program
36	affordable care act 2010
35	variations medicare spending part
33	centers medicare medicaid services
29	patientcentered outcomes research institute
28	use health information technology
28	preventive services task force
28	adoption electronic health records
27	quality health care delivered
27	health care delivered adults
27	electronic health record systems
27	delivered adults united states
27	care delivered adults united
26	content quality accessibility care

n-gram Analysis

Frequency	All Grams
11415	health
9827	care
3041	health care
2767	medical
2264	patients
2254	quality
1955	medicare
1823	use
1521	new
1518	hospital
1480	patient
1480	insurance
1479	primary
1427	states
1369	costs
1284	spending
1265	united
1234	policy
1234	national

n-grams	1,048,576
Highest frequency	11,415
Frequency=1	892,609
Frequency= <6	1,028,224
Frequency= <11	1,040,282
Frequency= <26	1,045,482
Frequency= >11	**8294**
Frequency= >26	**3094**

Curating the Raw Data



Frequency	All Grams
11415	health
9827	care
3041	health care
2767	medical
2264	patients
2254	quality
1955	medicare
1823	use
1521	new
1518	hospital
1480	patient
1480	insurance
1479	primary
1427	states
1369	costs
1284	spending
1265	united
1234	policy
1234	national

- Cut off long tail
- Target frequent, well-formed terms/concepts
- Discard noise
- Remove conceptual duplicates
  - – And compound fragments
- Identify relevant concepts; save duplicates as synonyms

Result is a list of candidates.

Curating the Raw Data

- 1-grams are not the most useful
  - – But require review, some good stuff there
- 2-, 3-, and 4- grams yield the most good terms
- 5-grams +: worth looking at, but low value
  - – “Single-photon emission computed tomography”

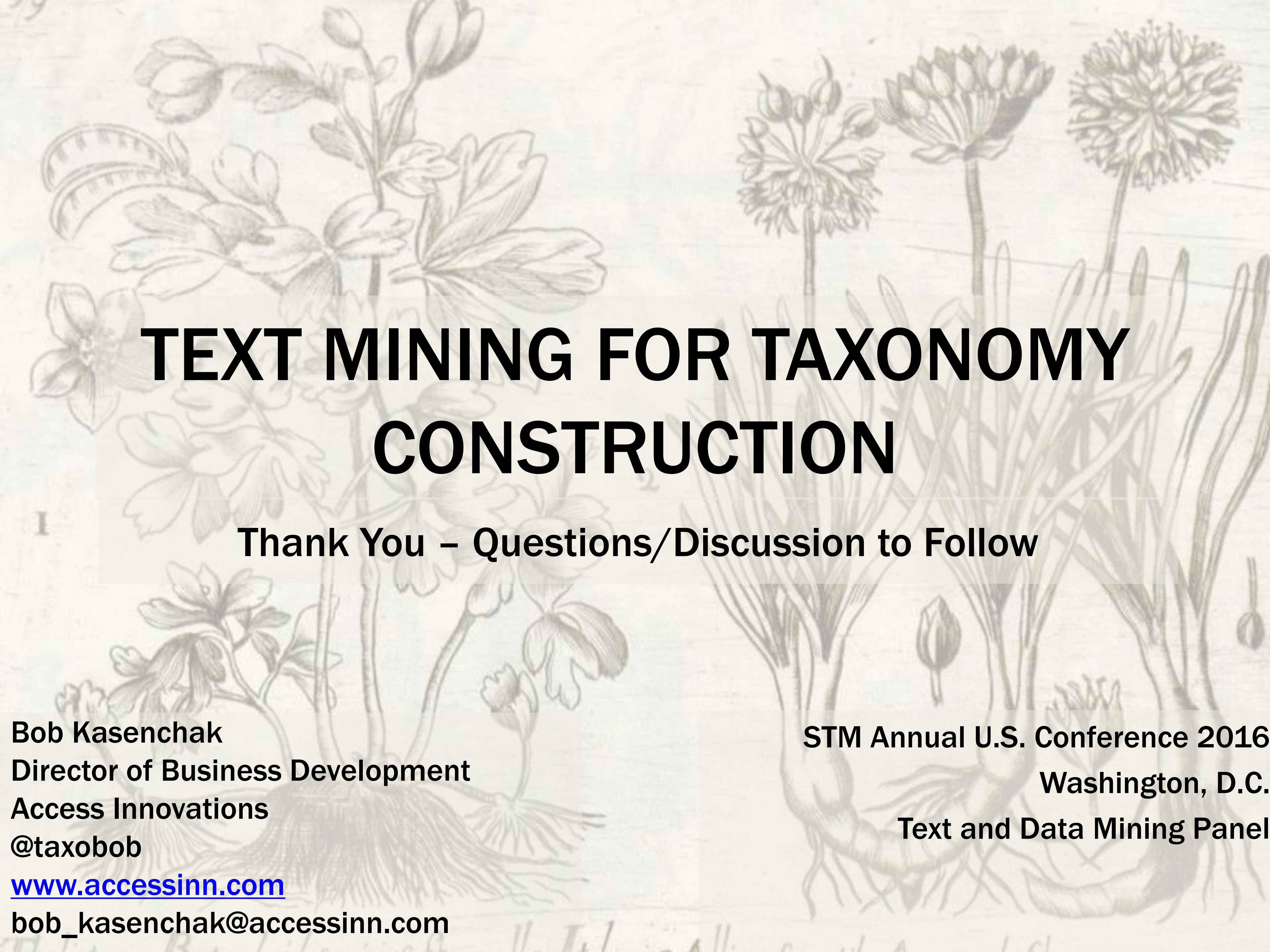
An SME should review the terms when you’re done, so don’t worry about getting the technical vocabulary 100% correct.

From Candidate Terms to Taxonomy

Result: list of candidate terms/concepts

Next steps:

- Resolve conceptual duplicates
  - Capture synonyms
- Build hierarchy
- Build out term records
- SME review



TEXT MINING FOR TAXONOMY CONSTRUCTION

Thank You – Questions/Discussion to Follow

Bob Kasenchak
Director of Business Development
Access Innovations
@taxobob
www.accessinn.com
bob_kasenchak@accessinn.com

STM Annual U.S. Conference 2016
Washington, D.C.
Text and Data Mining Panel