CHOOSING A DIGITAL METHOD

DATA MINING

@idah_Ku
@kalanicraig

August 31, 2017
START WITH YOUR RESEARCH

What's your research question?
At what stage is your research?
What's the scale of your research?
DATA MINING = PATTERN MATCHING IN A&H

- Identify similarities in a large dataset
  - Overarching trends across a whole set
  - Association within sets
  - Categorization of new items being added to a set

- Identify differences
  - Outliers and anomalies

- Combine the two
  - Clusters of similar groups with indications of outlier groups
Spies in Cuba in the late 19th century
A joint project with Prof. Arlene Diaz, Associate Professor of History, Indiana University--
Bloomington

AN EXAMPLE
List of Topics—Pinkerton “Operative” Reports

1. day give merriweather noteworthy world report february people working higgins
2. police military crowd censo central parque autonomy blanco secret appeared
3. bryan garcia told journal cesneros evangelina good insurrection scoop wasdin
4. gomez valle herald gonzalez months maximo spiritus sancti family pepper
5. johnstone rodriguez pardo public dunning weyer manolo porto maccreey excitement
6. havana congost palace autonomy junta connection train important speaking protection
7. sullivan battle russell piece man important guns part cement fort
8. maine left officers capst signeebee officer mr navy holman raff
9. herald boat nichols explained write man tuesday ward disguised taking
10. coup confidence lee mckinley general called york tobacco desert
11. bryan lewis halsted alfonso masso expected warship begin oriente calling
12. explosion americans naval maine theory lie barton clara torpedo asked
13. caldwell afternoon supervisee sargus make night koop grande island reich
14. told key west send wednesday mrs world chief tug ago
15. government war night friends sunday morrow coming brother casares article
16. garsia blanco cuba decker evening government small article money marina
17. make party instructions believed gentlemen gave giving reason twenty account
18. states united york knew deal ago lot high fellow intention
19. dr maran spanish betancourt write ahead talked spoke mr red
20. information bay ship present spaniards at bruner ponty mine finding
21. father chidwick mages chicago flag list royal cemetery chaplain secretary
22. la newspapers discuss soldiers thing seovell made blanco pando situation
23. redding young marix man senator matanzas maccreey confidence passengers wreck
24. evening work sir de minister operative spanish aboard things contiinuance
25. cuban morning long mail correspondent insurgent chat pretty pay york
26. halsted king reconcentrados washington cuban give hour congressman question fellows
27. left means afternoon la washington subject excellency spanish operative lake
28. business president received meaning washington page copy answer bring march
29. caldwell told garcia january pepper and sail room calle dupuy
30. american herald york paper mr interpret citizen aide printed ricardo
31. nea dock day men spaniards san juan nuz beard couple
32. time night case place doctor left english coming good england
33. spanish journal cuba papers find authorities consult time british learned
34. hotel bryson people thing letter delorme inglaterra letters de se
35. put showed back truth telegram big dr luncheon taking statement
36. general lee made volunteers journal troops good large means boy
37. scovel world account yesterday wrote morning laborantes interview speaks things

List of Topics—Sylvester Scovel’s NY World articles (newspaper belonged to Pulitzer)

1. made men crew back beach minutes full hands gussie fire point started felt light head
2. american battle left ashore hour began texas shill afternoon struck killed cabanans force
3. attack firing
4. boat triton coast tug landing shore close night land high white north miles yesterday port
5. gen men shafter samaly army jully city fire cavalry infantry surrender spainards artillery tural
6. make
7. havana world despatch cable steamer west correspondent special key harbor war news april
8. saturday monady
9. maine forward explosion side big wreck divers port blown part magazines bottom starboard
10. water intact
11. gurs big ships havana battery sampson shot batteries fleet yards gun morro range clara bay
12. man americans men told la full marine papers discipline duty correspondents true paper face
13. truth
14. capt captain officer officers lieut commander chief leave ensign anthony margrove quickly
15. charge correspondents lieutenant
16. spanish march staff board thought present reports instead made question war condition finally
17. facts months
18. signeebee divers work maine capt government navy yesterday alfonso recovered inside
19. investigation offer mines morrow
20. good brought long orders reached answer passed black cuba thirty carried arrival gave order
21. great time days give day war authorities danger strong streets life regular worked splendid delay
22. cuba states united people gomez cuban government mr years brave food relief consular
town thousand
23. magazine ship powder torpedo inch ten mine exploded case room fact force hundred experts
reserve
24. lee gen general blanco consult state official palace washington volunteers called clock staff
congreso department
25. spanish admiral miles hall expected june saniglio morning flag warships sight entrance lines
cervera knight
26. day court naval maine inquiry report board wreck testimony work members navy evidence
wrecking international
27. spanish found position bow feet small ram twenty anchor line discovered heavy proper turret
show
28. night boats city water explosion washington iron wounded pieces deck hand idea sailors
29. twenty number
30. spanish cubans men spaniards forces troops soldiers fight put ago food small field easily
31. world key west correspondent calived time hours week find information yesterday friday
32. officials proved commission
33. american spain havana censor press red barker president barton received citizens consulate

1. question span cuba island work favor prominent long subject responsible
2. camo found house rebel days afterward guards order quarters civil
3. sugar factory cane situation fields time estate agune ordered large
4. insurgents made insurgent fires pac country district reports authorities guerrillas
5. made manner determined attention follow post order rates aware remember
6. spanish young allowed search incident recall presence statement roll woman
7. country people order cattle property food provinces remained rebels homes
8. road half men train machetes taking brought expected caesed charge
9. make money day informed matter care making pay paid evidenced
10. cubans great scene mind military campaign sides completely attack sufficient
11. del pinar province habana río maceo santa clara san rebel
12. men soldiers place hills left full places orders ignorant put
13. smiths return country trip columnist guides miles leave sand
14. town small houses night fire fort cross garrison palma party
15. maceo spanish trocha troops part attack head forces staff body
16. habana public press head city spanish news hear chicago night
17. de destroyed cuba vietanzas real class partially burned los local
18. present saa de mirabal requested carry point weeks efforts navy
19. time story told felt friend clear finally intende toward find
20. herald gian don turned words answer general started shoot prisoner
21. spaniards machete hand impediments cut cavalry retreat stand abandoned quick
22. correspondents news stories leading newspaper imposed great letters laboratories chapter
23. york story world west fake cuba key florida junta miles
24. men col command regiment armed district twenty received arrested ia
25. american truth case americans facts correspondent starvation authorities owing june
26. spanish front direction artillery move short moved feet quickly volley
27. gomez man thought rogue chief soldier fellow etsy asked back
28. insurgent march chief chiefs dizaz command long called times held
29. united government cuba congress president true resolution washington call law
30. cubans cases bermudez conduct rebels volunteers basic death committed campos
31. letter correspondent journal york race world papers time action days
32. mr president superintend halstead winsomd slated parents high book separation
33. cubans spanish part failed battle government attempt scene purpose live
34. states war state give people find article times warfare stop
35. insurgents rebels poor work entered winsomd burning clothes unfortunate stealing
36. gomez chief nevel rank character pass generals met superior concluded

List of Topics—George Clark Musgrave, Under Three Flags in Cuba (1899)

1. general city santiago ship shatter shatter surrender july harbor navy
2. wyler insurgents cubans thousands standing lines attempted prisoners brutal duty
3. father young prison miss girl prisoner evangelina seized obtained lady
4. hour hand fade guard la face dozen imminent ordered sounded
5. fight carried turned started poured time told opened shocks worn
6. habana district province women trocha cities camaguey columns guerrillas cattle
7. cubans negro orders return liberty cane remained jos foreign planters
8. se time left mrs friend years friends wife or passed
9. lieutenant brave man soldiers struck held formed face charge cry
10. mr english story action visited house consul money finally police
11. day days fever siboney santa clara landing died railroad passing
12. army lack troops excellent tampa provided meat military beef fresh
13. horse gave light rose rear valley distant burst black faces
14. back received good light finally soldiers formed scene hold rapidly
15. rebels spaniards weeks bodies months past sugar frequently terror family
16. colonel shot young son head place brother means molina rebellion
17. cubans men proved field volunteers held home armed number campos
18. cuban island spanish government officials revolution independence reforms secretary administration
19. gomez small maceo del insurgent command end force rio death
20. dead rebel day man met exhausted save stores ride passed
21. fire enemy cavalry artillery spaniards advance road fired position bullets
22. country party marched on air wealthy blood glorious route silence
23. hundred thousand cuban long americans gained commander patriots city point
24. president long autonomy civil or made conditions span matusan assembly
25. city strongly, lay ladies blue showed turned amid quarter gold
26. field town captured side spanish knew capture escape effect garrison
27. spanish span great general late political power base situation rule
28. habana life starvation prominent daughters grave drew class scenes clear
29. military made heart pass door received state greater letter approach
30. colonel attempt sick court absolutely fear spanish thirty constant secure
31. white sea water sun rode rifles narrow coast don grew
32. men officers general captain line staff position rough brigade american
Embajada de España en Washington DC, 143 to 144
Embajada de España en Washington DC, 149 to 150
Embajada de España en Washington DC, 140 to 142
Embajada de España en Washington DC, 145 to 146
Embajada de España en Washington DC, 139 to 140
Embajada de España en Washington DC, 147 to 148
Embajada de España en Washington DC, 133 to 132
Embajada de España en Washington DC, 146
Embajada de España en Washington DC, 131 to 132
Embajada de España en Washington DC, 155 to 157
Embajada de España en Washington DC, 156 to 158
Embajada de España en Washington DC, 151 to 152
Embajada de España en Washington DC, 150 to 151
Embajada de España en Washington DC, 009 to 011
Embajada de España en Washington DC, 007 to 008
Embajada de España en Washington DC, 004 to 005
Embajada de España en Washington DC, 006 to 007
Embajada de España en Washington DC, 011 to 012
Embajada de España en Washington DC, 015 to 016
Embajada de España en Washington DC, 010 to 011
Embajada de España en Washington DC, 014 to 015
Embajada de España en Washington DC, 013 to 014
Embajada de España en Washington DC, 019 to 021
Embajada de España en Washington DC, 023 to 024
Embajada de España en Washington DC, 021 to 022

Spanish Minister
Sylvester Scovel

Embajada de España en Washington DC, 072 to 073
Embajada de España en Washington DC, 076 to 078
Embajada de España en Washington DC, 074 to 075
Embajada de España en Washington DC, 071 to 072
Embajada de España en Washington DC, 070 to 071
Embajada de España en Washington DC, 069 to 071
Embajada de España en Washington DC, 068 to 069
Embajada de España en Washington DC, 067
Embajada de España en Washington DC, 066 to 068
Embajada de España en Washington DC, 065 to 066
Embajada de España en Washington DC, 064 to 065
Embajada de España en Washington DC, 063 to 064
Embajada de España en Washington DC, 062 to 063
Embajada de España en Washington DC, 061 to 062
Embajada de España en Washington DC, 060 to 061
Embajada de España en Washington DC, 059 to 060
Embajada de España en Washington DC, 058 to 059
Embajada de España en Washington DC, 057 to 058
Embajada de España en Washington DC, 056 to 057
Embajada de España en Washington DC, 055 to 056
Embajada de España en Washington DC, 054 to 055
Embajada de España en Washington DC, 053 to 054
Embajada de España en Washington DC, 052 to 053
Embajada de España en Washington DC, 051 to 052
Embajada de España en Washington DC, 050 to 051
Embajada de España en Washington DC, 049 to 050
Embajada de España en Washington DC, 048 to 049
Embajada de España en Washington DC, 047 to 048
Embajada de España en Washington DC, 046 to 047
Embajada de España en Washington DC, 045 to 046
Embajada de España en Washington DC, 044 to 045
Embajada de España en Washington DC, 043 to 044
Embajada de España en Washington DC, 042 to 043
Embajada de España en Washington DC, 041 to 042
Embajada de España en Washington DC, 040 to 041
Embajada de España en Washington DC, 039 to 040
Embajada de España en Washington DC, 038 to 039
Embajada de España en Washington DC, 037 to 038
Embajada de España en Washington DC, 036 to 037
Embajada de España en Washington DC, 035 to 036
Embajada de España en Washington DC, 034 to 035
Embajada de España en Washington DC, 033 to 034
Embajada de España en Washington DC, 032 to 033
Embajada de España en Washington DC, 031 to 032
Embajada de España en Washington DC, 030 to 031
Embajada de España en Washington DC, 029 to 030
Embajada de España en Washington DC, 028 to 029
Embajada de España en Washington DC, 027 to 028
Embajada de España en Washington DC, 026 to 027
Embajada de España en Washington DC, 025 to 026
Embajada de España en Washington DC, 024 to 025
Embajada de España en Washington DC, 023 to 024
Embajada de España en Washington DC, 022 to 023
Embajada de España en Washington DC, 021 to 022
Embajada de España en Washington DC, 020 to 021
Embajada de España en Washington DC, 019 to 020
Embajada de España en Washington DC, 018 to 019
Embajada de España en Washington DC, 017 to 018
Embajada de España en Washington DC, 016 to 017
Embajada de España en Washington DC, 015 to 016
Embajada de España en Washington DC, 014 to 015
Embajada de España en Washington DC, 013 to 014
Embajada de España en Washington DC, 012 to 013
Embajada de España en Washington DC, 011 to 012
Embajada de España en Washington DC, 010 to 011
Embajada de España en Washington DC, 009 to 010
Embajada de España en Washington DC, 008 to 009
Embajada de España en Washington DC, 007 to 008
Embajada de España en Washington DC, 006 to 007
Embajada de España en Washington DC, 005 to 006
Embajada de España en Washington DC, 004 to 005
Embajada de España en Washington DC, 003 to 004
Embajada de España en Washington DC, 002 to 003
Embajada de España en Washington DC, 001 to 002
Embajada de España en Washington DC, 000 to 001

Robert Pinkerton
D.C. Office

Edward S. Gaylor

On Whistler, a very talented in el
reco de un instrucciones, de con
persono con el Piden al princi
mentada una de los dias de la
ora y tra la comisión directora
El World y para cuba a

ARCHIVAL PROOF
DATA MINING VOCABULARY

- Computational processes
  - Machine learning
  - Training corpus
  - Distribution analysis
  - NLP (natural language processing)

- Modifying documents for input
  - Corpus
  - Chunks
  - Tokens

- Metadata
  - Just that. Metadata.
HOW DOES DATA MINING WORK?

An example of distribution analysis in topic modeling: Latent Dirichlet Analysis (LDA)

1. Randomly assign each word in the corpus to a topic
2. Look at Word 1 in Topic A from Document X. Assume all other words in Topic A and Document X are properly distributed.
   2. Do the same calculations for all topics with Word 1 and Document X. With each topic, create a probability curve by comparing probabilities with the probability created using Word 1/Topic A/Document X.
3. If the probability curve is more certain for one of these topics, move Word 1 into the new topic. If not, leave it in Topic A.
3. Repeat ad nauseam.
DIFFERENT KINDS OF TEXT MINING

- **Word clouds**
  - Basic in-classroom use

- **N-grams**
  - One element of corpus linguistics.
  - Basic in-classroom use or simple change-over-time analysis

- **Corpus linguistics**
  - Advanced in-classroom use
  - N-grams with more flexibility
  - Concordance comparison

- **Comparing smaller differences in corpora (keyness)**

- **Topic modeling**
  - Advanced in-classroom use
  - Comparing large trends in corpora

- **Adding natural-language processing**
  - Part-of-speech tagging
  - Dependency parsing
  - Named Entity Recognition (NER)
## DATA MINING + ANOTHER DIGITAL METHOD

<table>
<thead>
<tr>
<th>Topic modeling only</th>
<th>Text analysis with maps</th>
</tr>
</thead>
<tbody>
<tr>
<td>For comparing language use between subjects or across time</td>
<td>For connecting ideas to movement in space</td>
</tr>
<tr>
<td>• How will you demonstrate specific instances of the broad trends?</td>
<td>• How will you demonstrate the connection of topics to space?</td>
</tr>
<tr>
<td>• Why are trends important?</td>
<td>• Why is space important?</td>
</tr>
<tr>
<td>• How will you handle exceptions?</td>
<td>• How will you show change over time?</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Topic modeling and corpus linguistics</th>
<th>Text analysis with networks</th>
</tr>
</thead>
<tbody>
<tr>
<td>For comparing language trends to language outliers</td>
<td>For connecting ideas to people</td>
</tr>
<tr>
<td>• How will you show trends vs outliers?</td>
<td>• How will you demonstrate the connection of topics to people?</td>
</tr>
<tr>
<td>• Why are outliers important?</td>
<td>• Is it people’s use of the topics or the topics’ influence on people that matters? (or both?)</td>
</tr>
<tr>
<td>• How will you show change over time?</td>
<td>• How will you deal with people in the network who aren’t in the topic model?</td>
</tr>
</tbody>
</table>
DIFFERENT KINDS OF IMAGE MINING

- Histograms
- Outlines
- Textures
LIMITATIONS AND ERRORS IN DATA MINING

- Be cognizant of your question as you gather data.
  - Data lies (if it’s collected in inappropriate ways). What’s in your set?
  - Data from NY and LA between 1920-1940 and 2000-2010 can’t be used to make claims about patterns in the US between 1900 and 2010.

- Be aware of the limits of data mining.
  - You need computer readable files.

- Be aware of what you can munge.*
  - If your transcription/OCR always turns the word “like” into the word “lilce”, consistency in errors is still consistency. (Natural Language Processing and facial recognition are exceptions)

* Actually a technical term
COMBINING IMAGE AND TEXT MINING

- Can you turn images into text?
  - Use metadata from images in text mining
  - Turn histogram/shape/textures into discrete image tokens

- Can you look at how images are embedded in text?
  - Change image into token and look at co-occurences of “image” token with other linguistic tells

- Can you look at how images influence text?
  - Combine the first two approaches and look at which linguistic tells appear when certain histogramatic/shape/texture patterns are present
A PRACTICAL ANALOG EXAMPLE

Topic modeling by hand
END WITH YOUR RESEARCH

What method(s) look useful?
What patterns do you think are there?
What data issues will you run into?
<table>
<thead>
<tr>
<th>Event</th>
<th>Date</th>
<th>Time</th>
<th>Series &amp; URL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mapping</td>
<td>12:30pm, Sept 7</td>
<td>CartoShop Mapping Series</td>
<td><a href="http://go.iu.edu/1Gz1">http://go.iu.edu/1Gz1</a></td>
</tr>
<tr>
<td>Network Analysis</td>
<td>10:30am, Sept 14</td>
<td>Digital Tools and Visualization Methods for Humanities series</td>
<td><a href="http://go.iu.edu/1Gz0">http://go.iu.edu/1Gz0</a></td>
</tr>
<tr>
<td>Making Digital Objects</td>
<td>Noon, Sept 21</td>
<td>Digital Tools and Visualization Methods for Humanities series</td>
<td><a href="http://go.iu.edu/1Gz0">http://go.iu.edu/1Gz0</a></td>
</tr>
</tbody>
</table>

Thursdays in August & September  
Wells Library, room E174
OTHER TOPIC MODELING SOFTWARE PACKAGES

- Serendip: http://vep.cs.wisc.edu/serendip/#serendipHero
- MALLET command line: http://mallet.cs.umass.edu
- Using R and MALLET together: http://www.themacrooscope.org/?page_id=67 (The Historian’s Macroscope is a great resource for many dighist tasks)