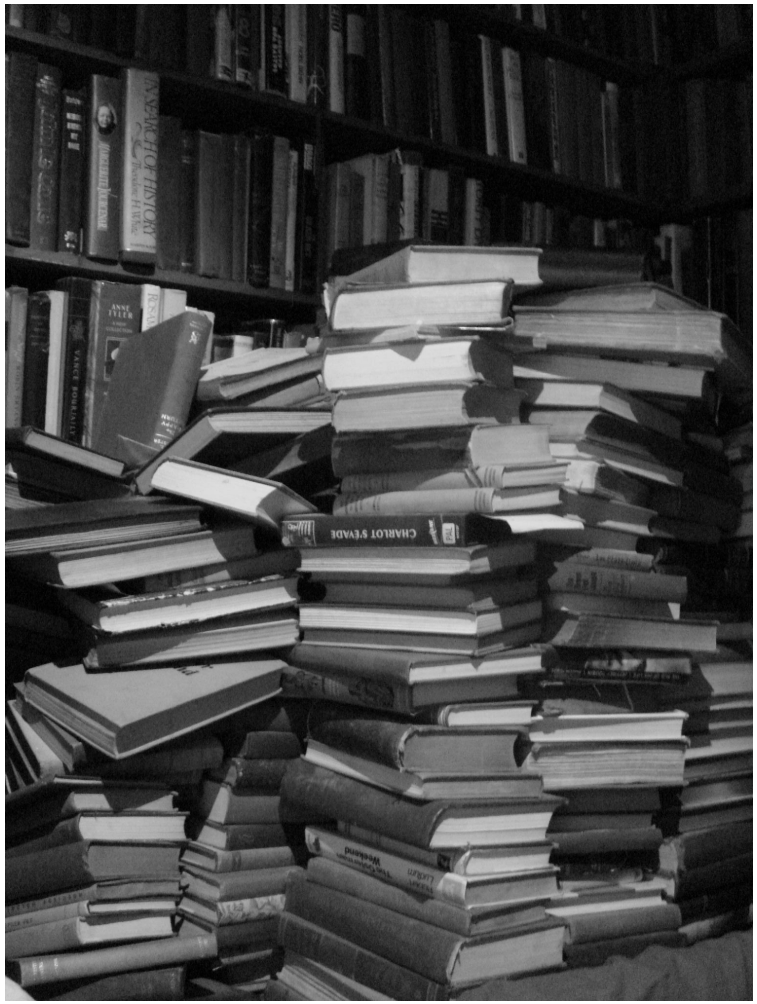
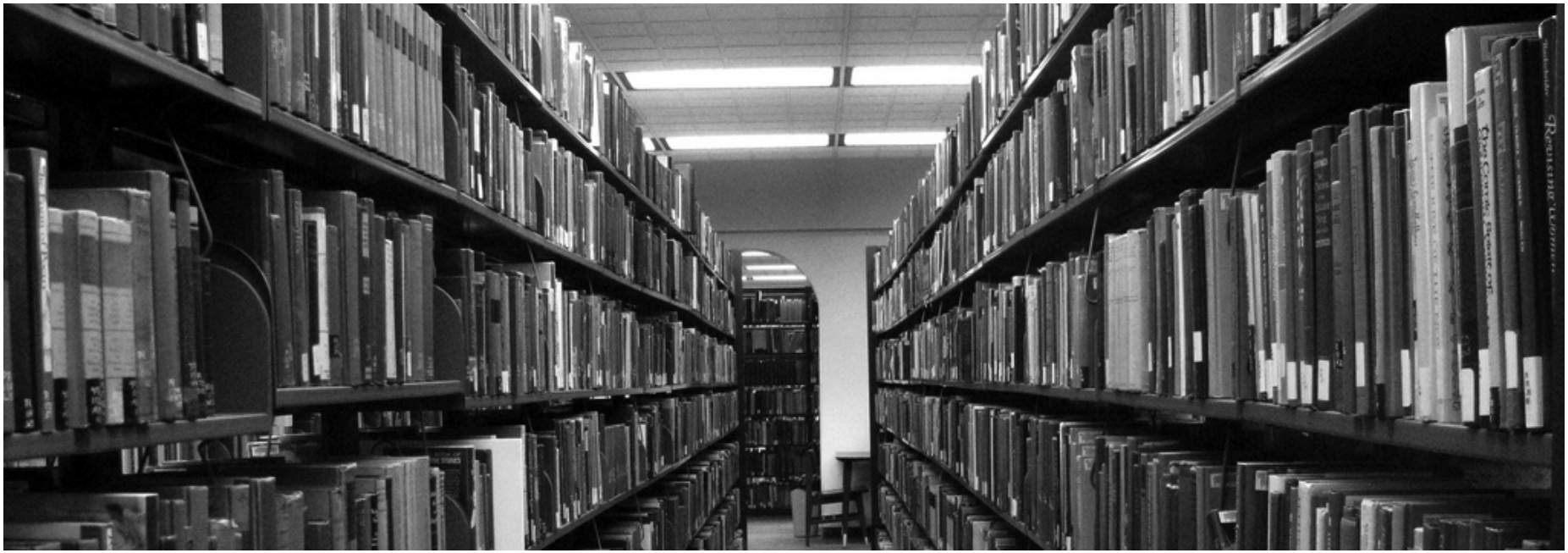
Topic Models for Understanding History

David M. Blei
Columbia University

joint work with Allison Chaney, Hanna Wallach, and Matt Connelly

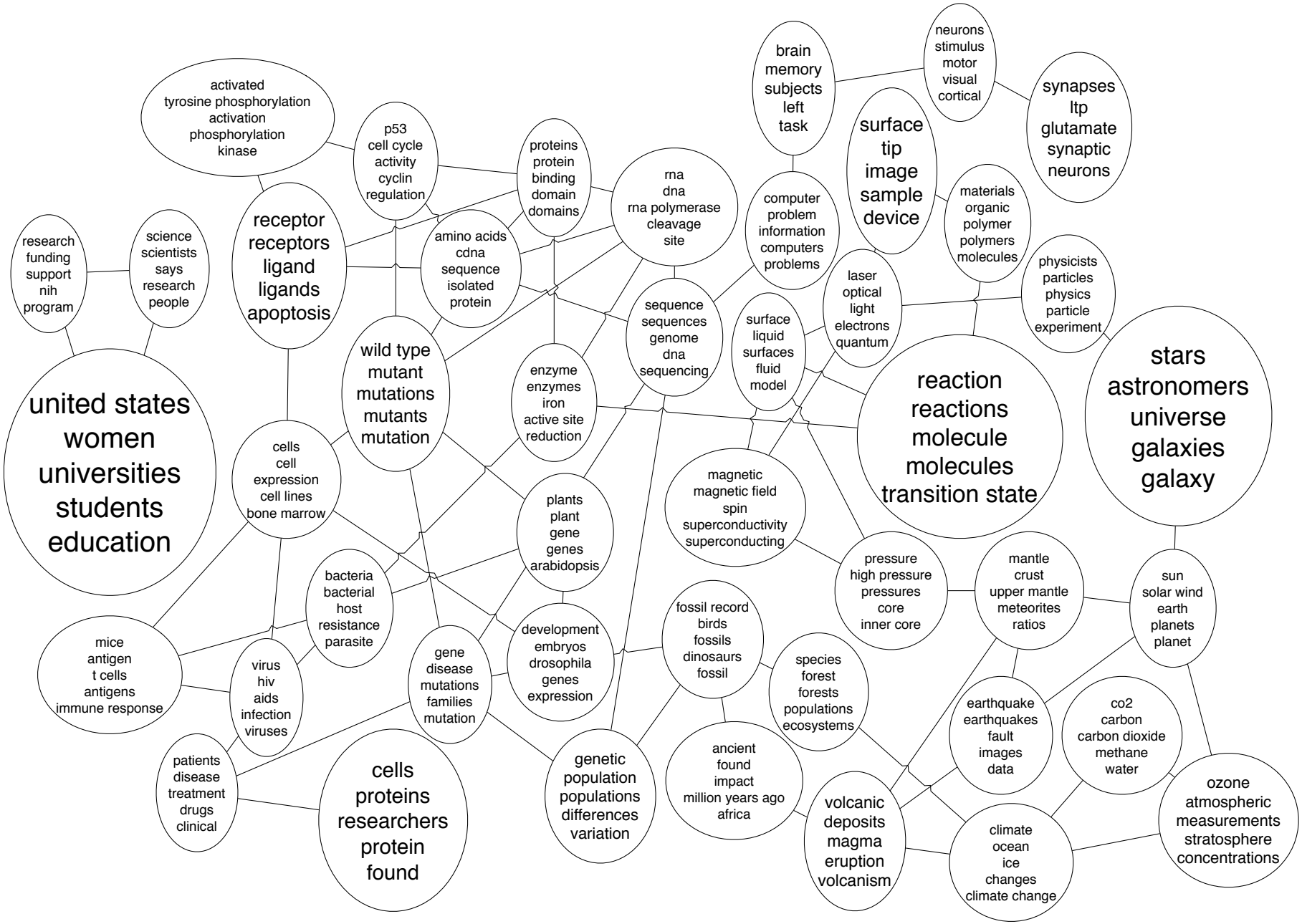


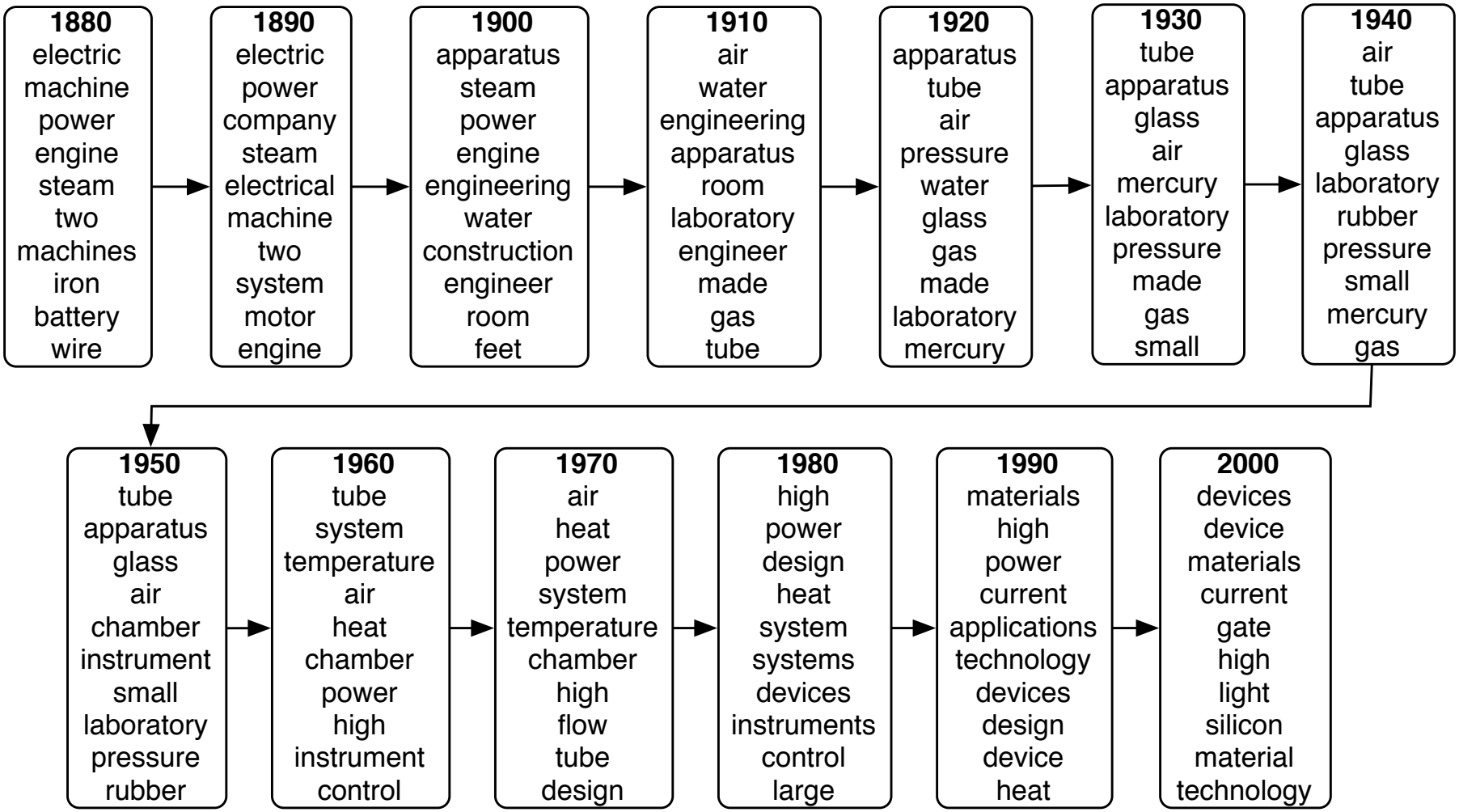
- **ORGANIZE**
- **VISUALIZE**
- **SUMMARIZE**
- **SEARCH**
- **PREDICT**
- **UNDERSTAND**

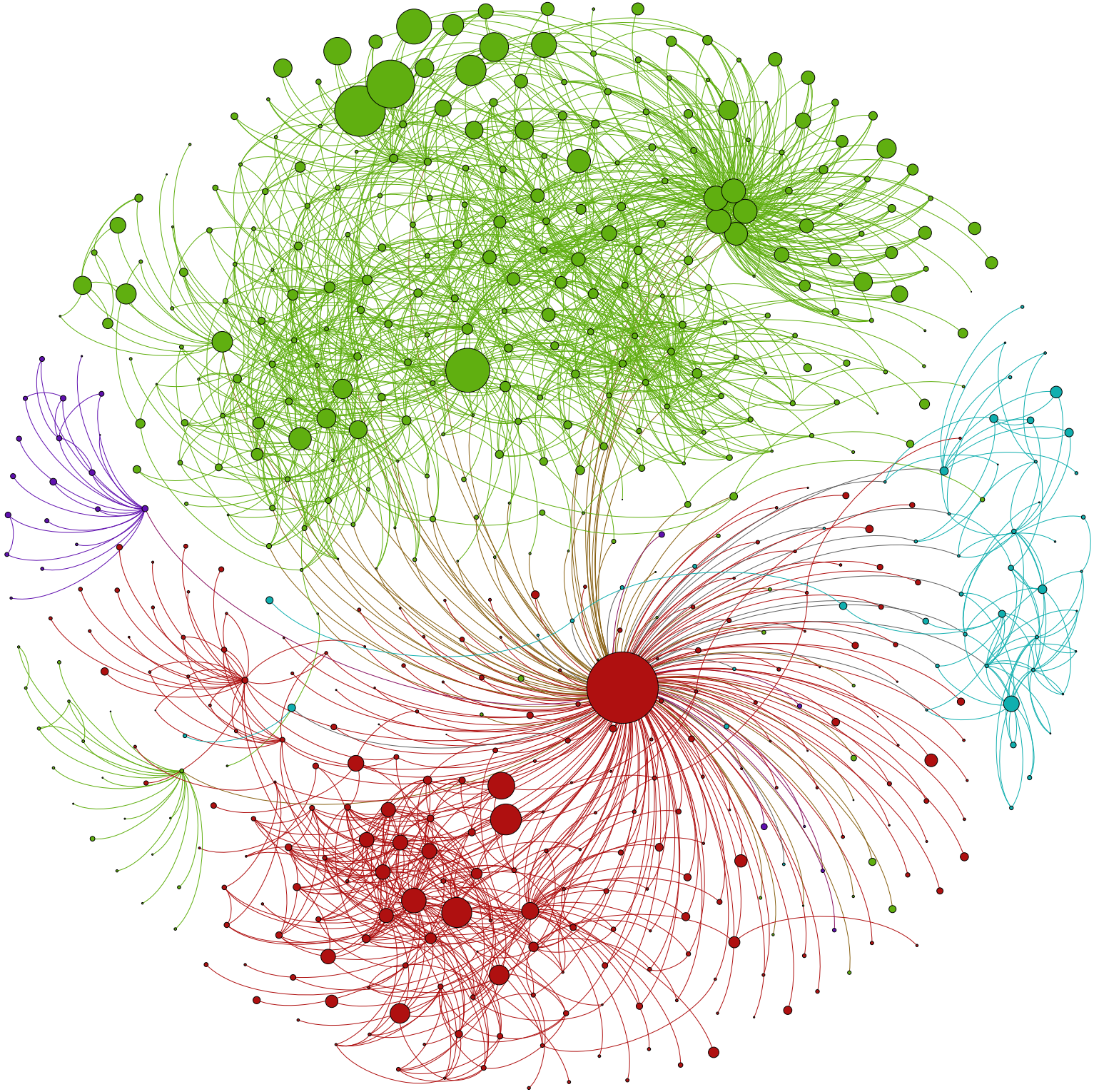


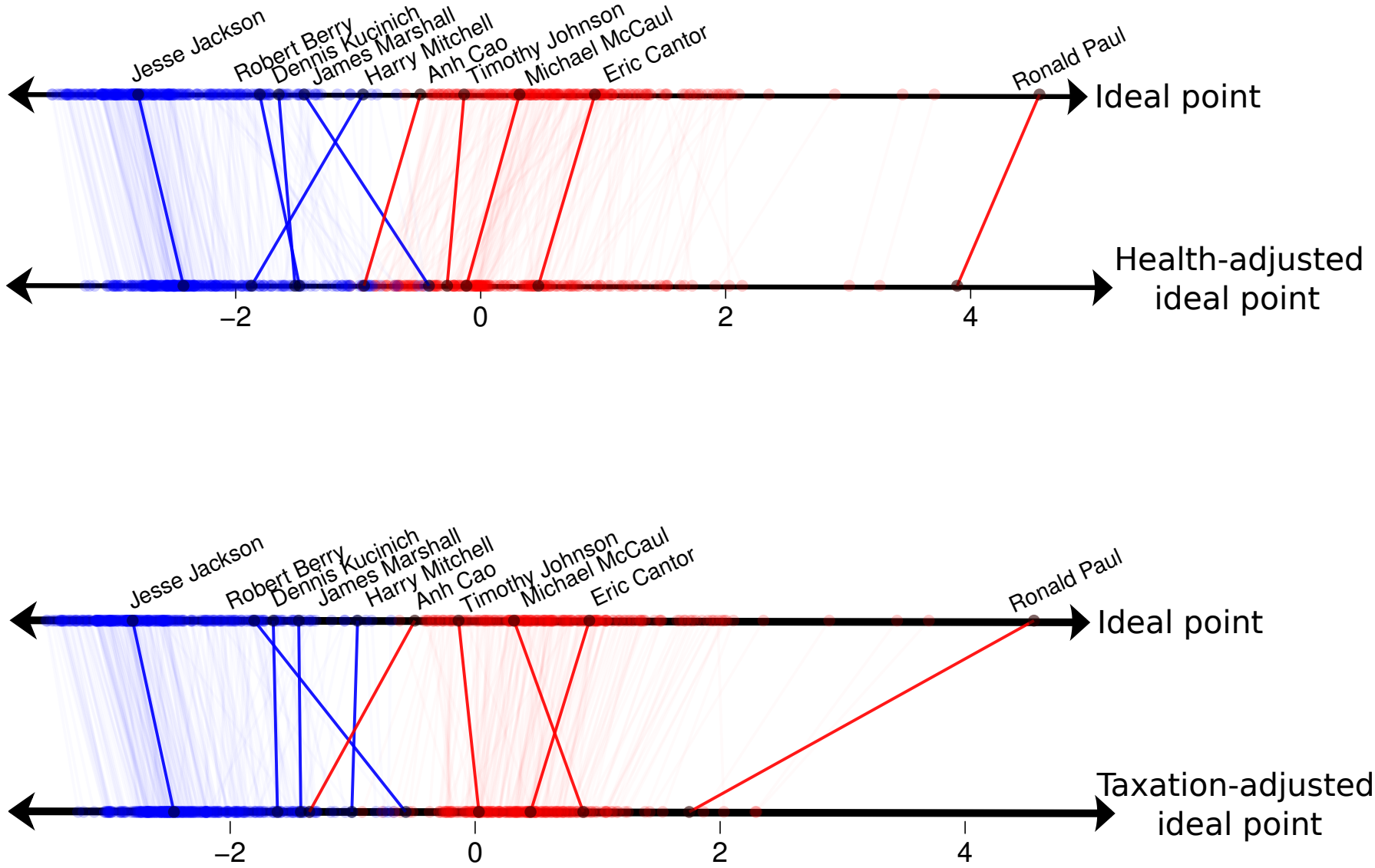
TOPIC MODELING

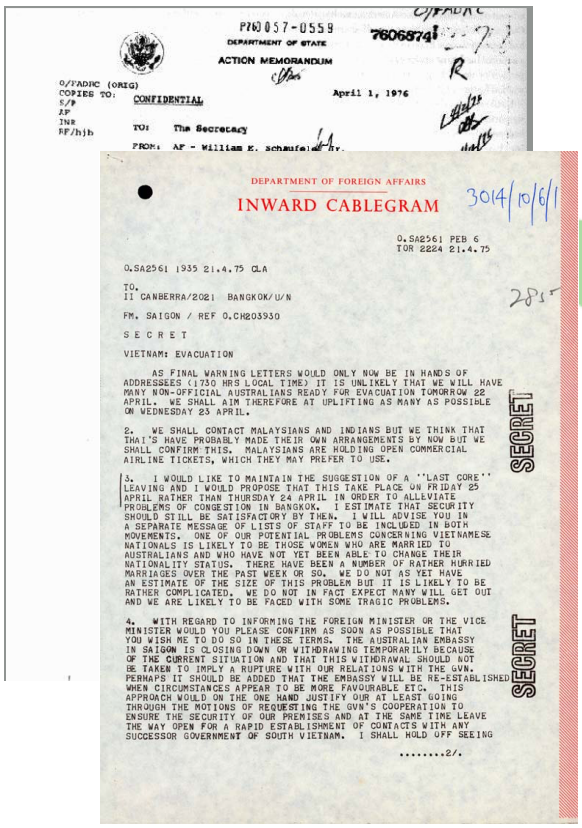
1. **Discover** the thematic structure
2. **Annotate** the documents
3. **Use** the annotations to visualize, organize, summarize, **...**


activated
tyrosine phosphorylation
activation
phosphorylation
kinase
p53
cell cycle
activity
cyclin
regulation
proteins
protein
binding
domain
domains
brain
memory
subjects
left
task
neurons
stimulus
motor
visual
cortical
synapses
ltp
glutamate
synaptic
neurons
surface
tip
image
sample
device
rna
dna
rna polymerase
cleavage
site
computer
problem
information
computers
problems
materials
organic
polymer
polymers
molecules
research
funding
support
nih
program
science
scientists
says
research
people
receptor
receptors
ligand
ligands
apoptosis
amino acids
cdna
sequence
isolated
protein
laser
optical
light
electrons
quantum
physicists
particles
physics
particle
experiment
wild type
mutant
mutations
mutants
mutation
sequence
sequences
genome
dna
sequencing
surface
liquid
surfaces
fluid
model
enzyme
enzymes
iron
active site
reduction
reaction
reactions
molecule
molecules
transition state
stars
astronomers
universe
galaxies
galaxy
united states
women
universities
students
education
cells
cell
expression
cell lines
bone marrow
plants
plant
gene
genes
arabidopsis
magnetic
magnetic field
spin
superconductivity
superconducting
pressure
high pressure
pressures
core
inner core
mantle
crust
upper mantle
meteorites
ratios
sun
solar wind
earth
planets
planet
bacteria
bacterial
host
resistance
parasite
mice
antigen
t cells
antigens
immune response
virus
hiv
aids
infection
viruses
gene
disease
mutations
families
mutation
development
embryos
drosophila
genes
expression
fossil record
birds
fossils
dinosaurs
fossil
species
forest
forests
populations
ecosystems
earthquake
earthquakes
fault
images
data
co2
carbon
carbon dioxide
methane
water
patients
disease
treatment
drugs
clinical
cells
proteins
researchers
protein
found
genetic
population
populations
differences
variation
ancient
found
impact
million years ago
africa
volcanic
deposits
magma
eruption
volcanism
climate
ocean
ice
changes
climate change
ozone
atmospheric
measurements
stratosphere
concentrations


1880
electric
machine
power
engine
steam
two
machines
iron
battery
wire
1890
electric
power
company
steam
electrical
machine
two
system
motor
engine
1900
apparatus
steam
power
engine
engineering
water
construction
engineer
room
feet
1910
air
water
engineering
apparatus
room
laboratory
engineer
made
gas
tube
1920
apparatus
tube
air
pressure
water
glass
gas
made
laboratory
mercury
1930
tube
apparatus
glass
air
mercury
laboratory
pressure
made
gas
small
1940
air
tube
apparatus
glass
laboratory
rubber
pressure
small
mercury
gas
1950
tube
apparatus
glass
air
chamber
instrument
small
laboratory
pressure
rubber
1960
tube
system
temperature
air
heat
chamber
power
high
instrument
control
1970
air
heat
power
system
temperature
chamber
high
flow
tube
design
1980
high
power
design
heat
system
systems
devices
instruments
control
large
1990
materials
high
power
current
applications
technology
devices
design
device
heat
2000
devices
device
materials
current
gate
high
light
silicon
material
technology

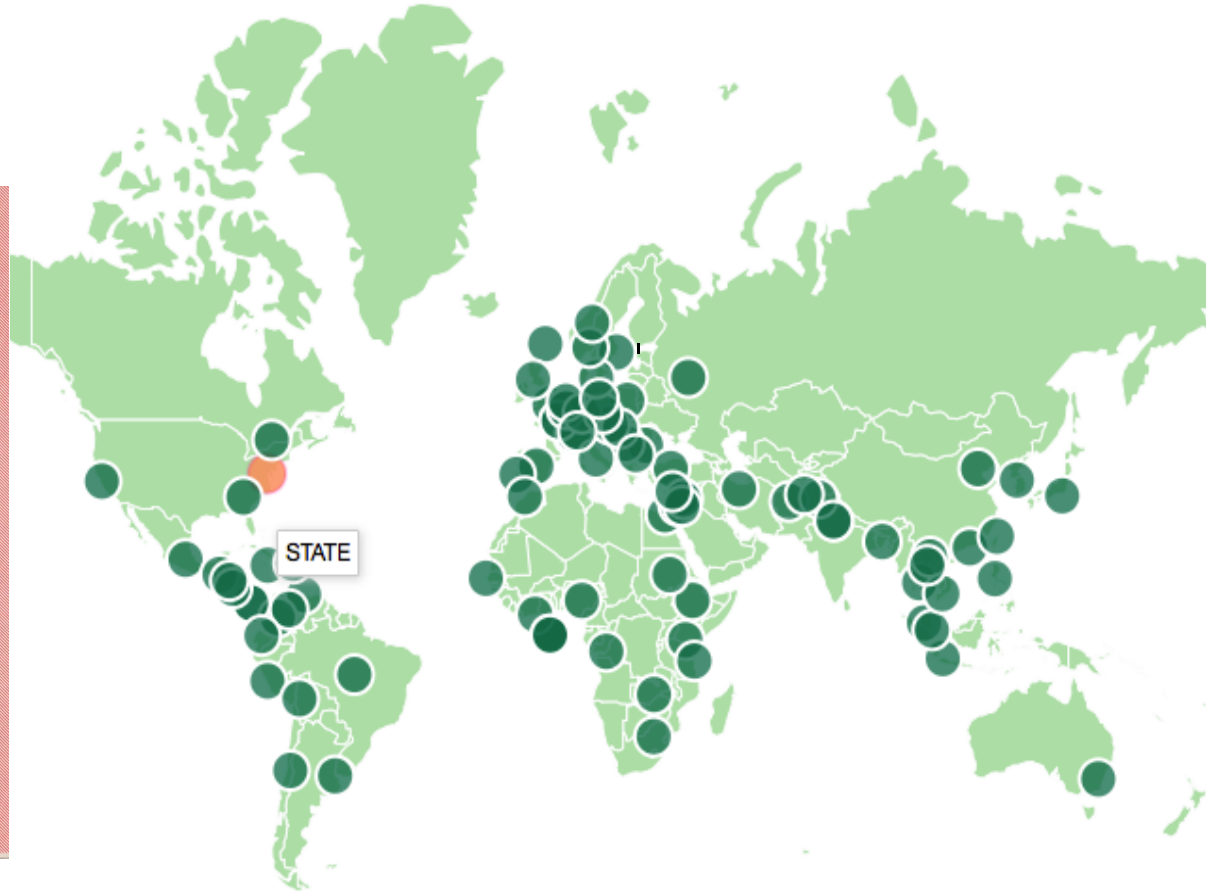



Jesse Jackson
Robert Berry
Dennis Kucinich
James Marshall
Harry Mitchell
Anh Cao
Timothy Johnson
Michael McCaul
Eric Cantor
Ronald Paul
Ideal point
Health-adjusted ideal point
−2
0
2
4
Jesse Jackson
Robert Berry
Dennis Kucinich
James Marshall
Harry Mitchell
Anh Cao
Timothy Johnson
Michael McCaul
Eric Cantor
Ronald Paul
Ideal point
Taxation-adjusted ideal point
−2
0
2
4



P760057-0559
7606574
DEPARTMENT OF STATE
ACTION MEMORANDUM
O/FADRC (ORIG)
COPIES TO: CONFIDENTIAL
S/P
AF
INR
RF/hjb
April 1, 1976
TO: The Secretary
FROM: AF - William E. Schaufele, Jr.

DEPARTMENT OF FOREIGN AFFAIRS
INWARD CABLEGRAM
3014/10/6/1
O.SA2561 PEB 6
TOR 2224 21.4.75
O.SA2561 1935 21.4.75 CLA
TO.
II CANBERRA/2021 BANGKOK/U/N
285
FM. SAIGON / REF O.CH203930
S E C R E T
VIETNAM: EVACUATION
AS FINAL WARNING LETTERS WOULD ONLY NOW BE IN HANDS OF ADDRESSEES (1730 HRS LOCAL TIME) IT IS UNLIKELY THAT WE WILL HAVE MANY NON-OFFICIAL AUSTRALIANS READY FOR EVACUATION TOMORROW 22 APRIL. WE SHALL AIM THEREFORE AT UPLIFTING AS MANY AS POSSIBLE ON WEDNESDAY 23 APRIL.
2. WE SHALL CONTACT MALAYSIANS AND INDIANS BUT WE THINK THAT THAI'S HAVE PROBABLY MADE THEIR OWN ARRANGEMENTS BY NOW BUT WE SHALL CONFIRM THIS. MALAYSIANS ARE HOLDING OPEN COMMERCIAL AIRLINE TICKETS, WHICH THEY MAY PREFER TO USE.
3. I WOULD LIKE TO MAINTAIN THE SUGGESTION OF A "LAST CORE" LEAVING AND I WOULD PROPOSE THAT THIS TAKE PLACE ON FRIDAY 25 APRIL RATHER THAN THURSDAY 24 APRIL IN ORDER TO ALLEVIATE PROBLEMS OF CONGESTION IN BANGKOK. I ESTIMATE THAT SECURITY SHOULD STILL BE SATISFACTORY BY THEN. I WILL ADVISE YOU IN A SEPARATE MESSAGE OF LISTS OF STAFF TO BE INCLUDED IN BOTH MOVEMENTS. ONE OF OUR POTENTIAL PROBLEMS CONCERNING VIETNAMESE NATIONALS IS LIKELY TO BE THOSE WOMEN WHO ARE MARRIED TO AUSTRALIANS AND WHO HAVE NOT YET BEEN ABLE TO CHANGE THEIR NATIONALITY STATUS. THERE HAVE BEEN A NUMBER OF RATHER HURRIED MARRIAGES OVER THE PAST WEEK OR SO. WE DO NOT AS YET HAVE AN ESTIMATE OF THE SIZE OF THIS PROBLEM BUT IT IS LIKELY TO BE RATHER COMPLICATED. WE DO NOT IN FACT EXPECT MANY WILL GET OUT AND WE ARE LIKELY TO BE FACED WITH SOME TRAGIC PROBLEMS.
4. WITH REGARD TO INFORMING THE FOREIGN MINISTER OR THE VICE MINISTER WOULD YOU PLEASE CONFIRM AS SOON AS POSSIBLE THAT YOU WISH ME TO DO SO IN THESE TERMS. THE AUSTRALIAN EMBASSY IN SAIGON IS CLOSING DOWN OR WITHDRAWING TEMPORARILY BECAUSE OF THE CURRENT SITUATION AND THAT THIS WITHDRAWAL SHOULD NOT BE TAKEN TO IMPLY A RUPTURE WITH OUR RELATIONS WITH THE GVN. PERHAPS IT SHOULD BE ADDED THAT THE EMBASSY WILL BE RE-ESTABLISHED WHEN CIRCUMSTANCES APPEAR TO BE MORE FAVOURABLE ETC. THIS APPROACH WOULD ON THE ONE HAND JUSTIFY OUR AT LEAST GOING THROUGH THE MOTIONS OF REQUESTING THE GVN'S COOPERATION TO ENSURE THE SECURITY OF OUR PREMISES AND AT THE SAME TIME LEAVE THE WAY OPEN FOR A RAPID ESTABLISHMENT OF CONTACTS WITH ANY SUCCESSOR GOVERNMENT OF SOUTH VIETNAM. I SHALL HOLD OFF SEEING
.....2/.
SECRET
SECRET



- Historians want to identify important events from primary sources.
- Example: Embassies send cables to each other during the 1970s
- Goal: Use topic models to discover **events** in this data set

This talk

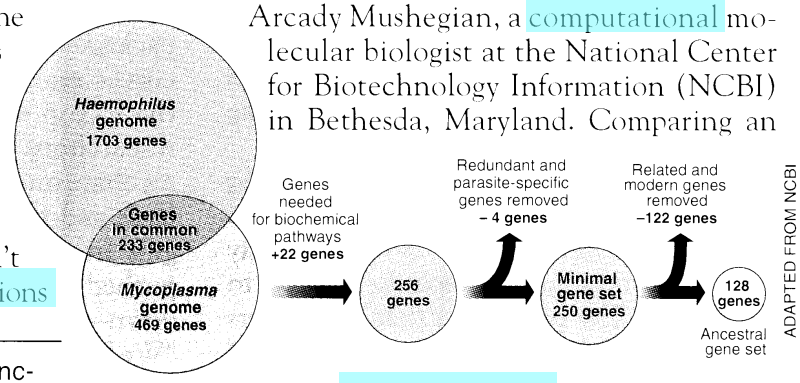
1. Introduction to topic modeling
2. Topic models for understanding history
3. The bigger picture: Using probability models to solve problems with data

Introduction to Topic Modeling

Seeking Life's Bare (Genetic) Necessities

COLD SPRING HARBOR, NEW YORK—How many genes does an organism need to survive? Last week at the genome meeting here,* two genome researchers with radically different approaches presented complementary views of the basic genes needed for life. One research team, using computer analyses to compare known genomes, concluded that today's organisms can be sustained with just 250 genes, and that the earliest life forms required a mere 128 genes. The other researcher mapped genes in a simple parasite and estimated that for this organism, 800 genes are plenty to do the job—but that anything short of 100 wouldn't be enough.

Although the numbers don't match precisely, those predictions "are not all that far apart," especially in comparison to the 75,000 genes in the human genome, notes Siv Andersson of Uppsala University in Sweden, who arrived at the 800 number. But coming up with a consensus answer may be more than just a genetic numbers game, particularly as more and more genomes are completely mapped and sequenced. "It may be a way of organizing any newly sequenced genome," explains Arcady Mushegian, a computational molecular biologist at the National Center for Biotechnology Information (NCBI) in Bethesda, Maryland. Comparing an



Stripping down. Computer analysis yields an estimate of the minimum modern and ancient genomes.

* Genome Mapping and Sequencing, Cold Spring Harbor, New York, May 8 to 12.

Documents exhibit multiple topics.

Topics

gene	0.04
dna	0.02
genetic	0.01
...	

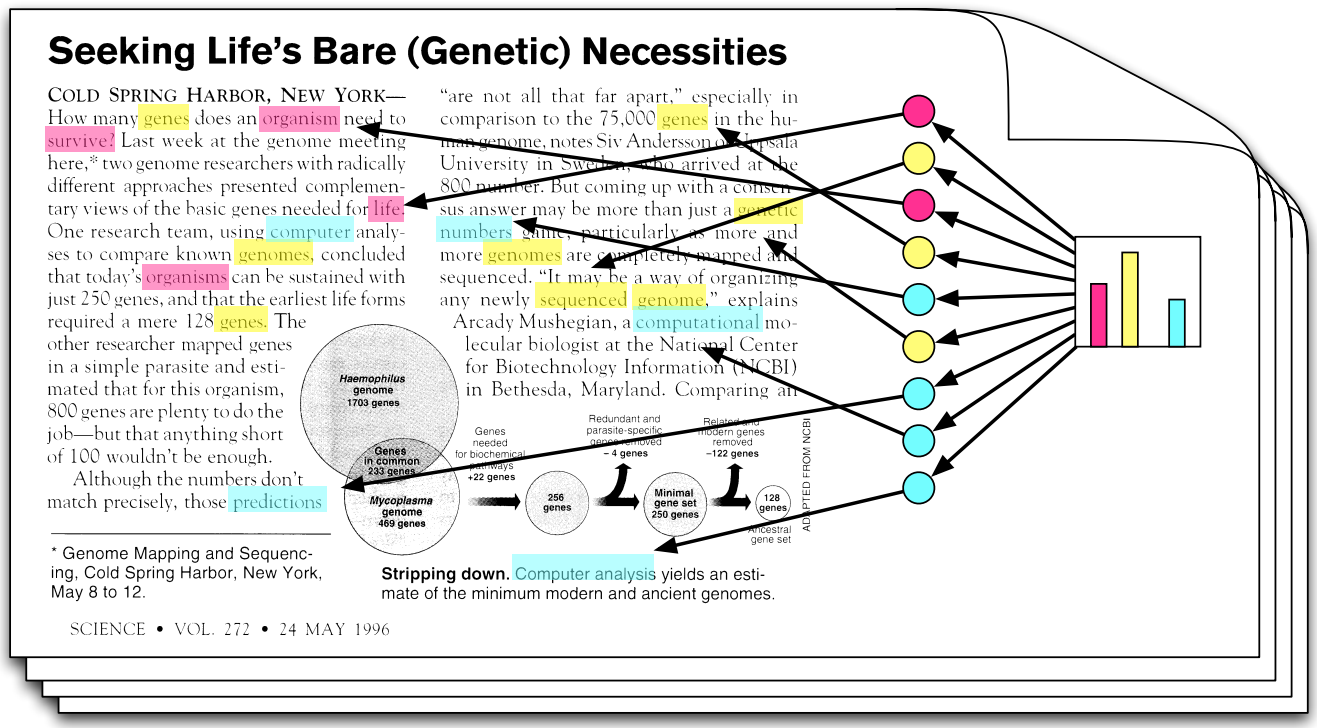
life	0.02
evolve	0.01
organism	0.01
...	

brain	0.04
neuron	0.02
nerve	0.01
...	

data	0.02
number	0.02
computer	0.01
...	

Documents

Topic proportions and assignments



Seeking Life's Bare (Genetic) Necessities

COLD SPRING HARBOR, NEW YORK—How many genes does an organism need to survive? Last week at the genome meeting here,* two genome researchers with radically different approaches presented complementary views of the basic genes needed for life. One research team, using computer analyses to compare known genomes, concluded that today's organisms can be sustained with just 250 genes, and that the earliest life forms required a mere 128 genes. The other researcher mapped genes in a simple parasite and estimated that for this organism, 800 genes are plenty to do the job—but that anything short of 100 wouldn't be enough.

Although the numbers don't match precisely, those predictions are not all that far apart," especially in comparison to the 75,000 genes in the human genome, notes Siv Andersson of Uppsala University in Sweden, who arrived at the 800 number. But coming up with a consensus answer may be more than just a genetic numbers game, particularly as more and more genomes are completely mapped and sequenced. "It may be a way of organizing any newly sequenced genome," explains Arcady Mushegian, a computational molecular biologist at the National Center for Biotechnology Information (NCBI) in Bethesda, Maryland. Comparing an

* Genome Mapping and Sequencing, Cold Spring Harbor, New York, May 8 to 12.

Stripping down. Computer analysis yields an estimate of the minimum modern and ancient genomes.

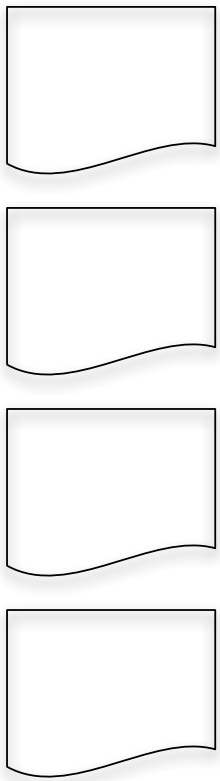
SCIENCE • VOL. 272 • 24 MAY 1996

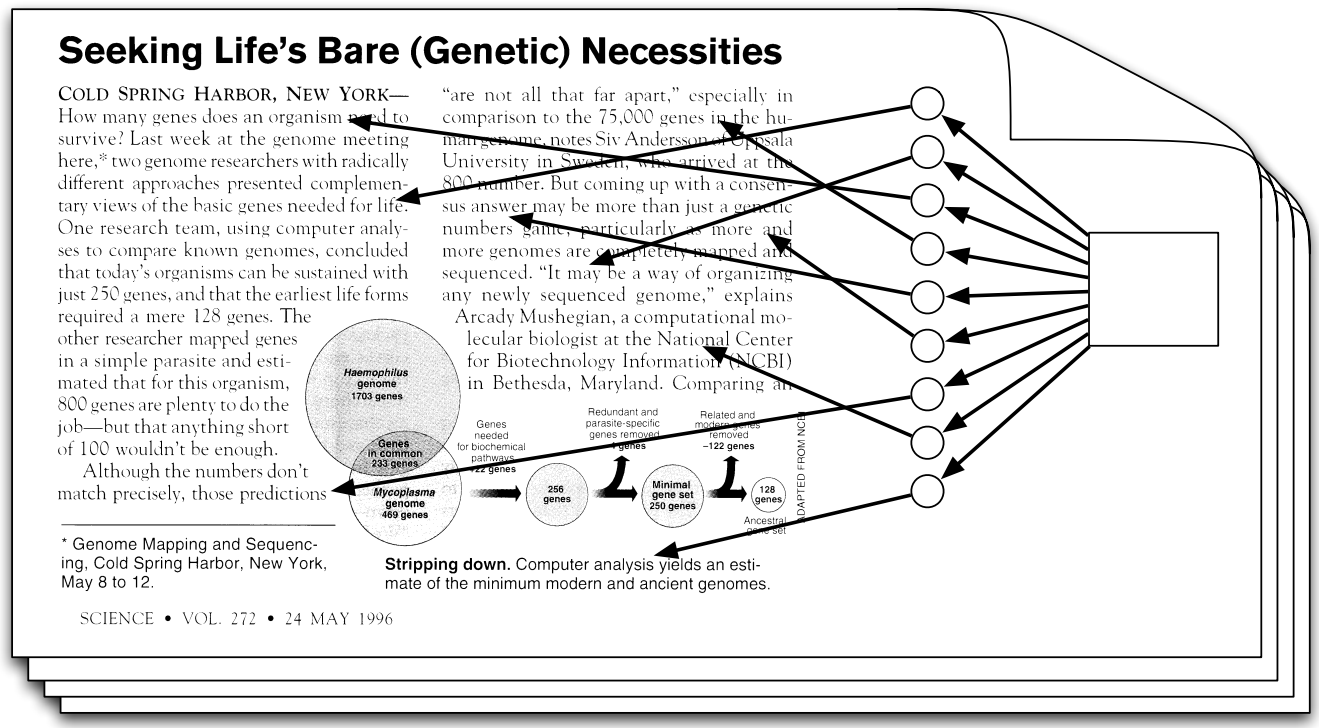
Latent Dirichlet Allocation

Topics

Documents

Topic proportions and assignments





Seeking Life's Bare (Genetic) Necessities

COLD SPRING HARBOR, NEW YORK—How many genes does an organism need to survive? Last week at the genome meeting here,* two genome researchers with radically different approaches presented complementary views of the basic genes needed for life. One research team, using computer analyses to compare known genomes, concluded that today's organisms can be sustained with just 250 genes, and that the earliest life forms required a mere 128 genes. The other researcher mapped genes in a simple parasite and estimated that for this organism, 800 genes are plenty to do the job—but that anything short of 100 wouldn't be enough.

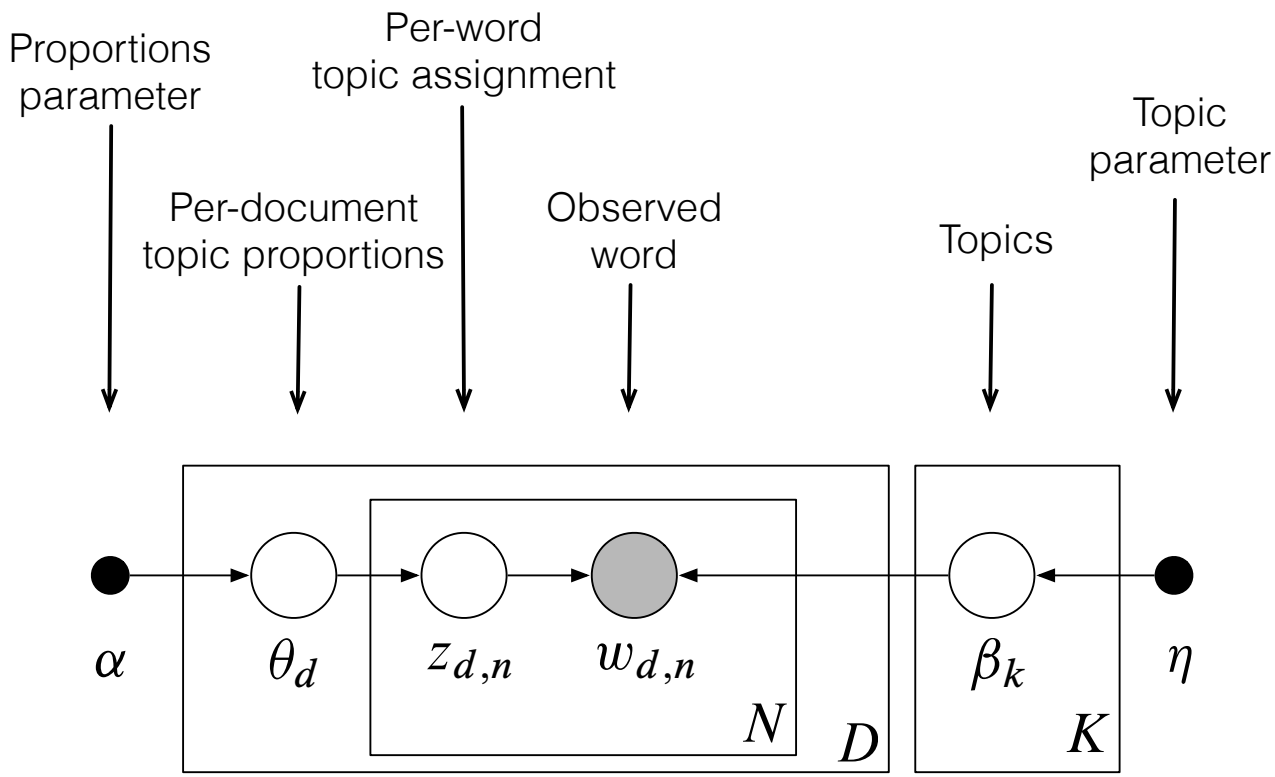
Although the numbers don't match precisely, those predictions "are not all that far apart," especially in comparison to the 75,000 genes in the human genome, notes Siv Andersson of Uppsala University in Sweden, who arrived at the 800 number. But coming up with a consensus answer may be more than just a genetic numbers game, particularly as more and more genomes are completely mapped and sequenced. "It may be a way of organizing any newly sequenced genome," explains Arcady Mushegian, a computational molecular biologist at the National Center for Biotechnology Information (NCBI) in Bethesda, Maryland. Comparing an

* Genome Mapping and Sequencing, Cold Spring Harbor, New York, May 8 to 12.

Stripping down. Computer analysis yields an estimate of the minimum modern and ancient genomes.

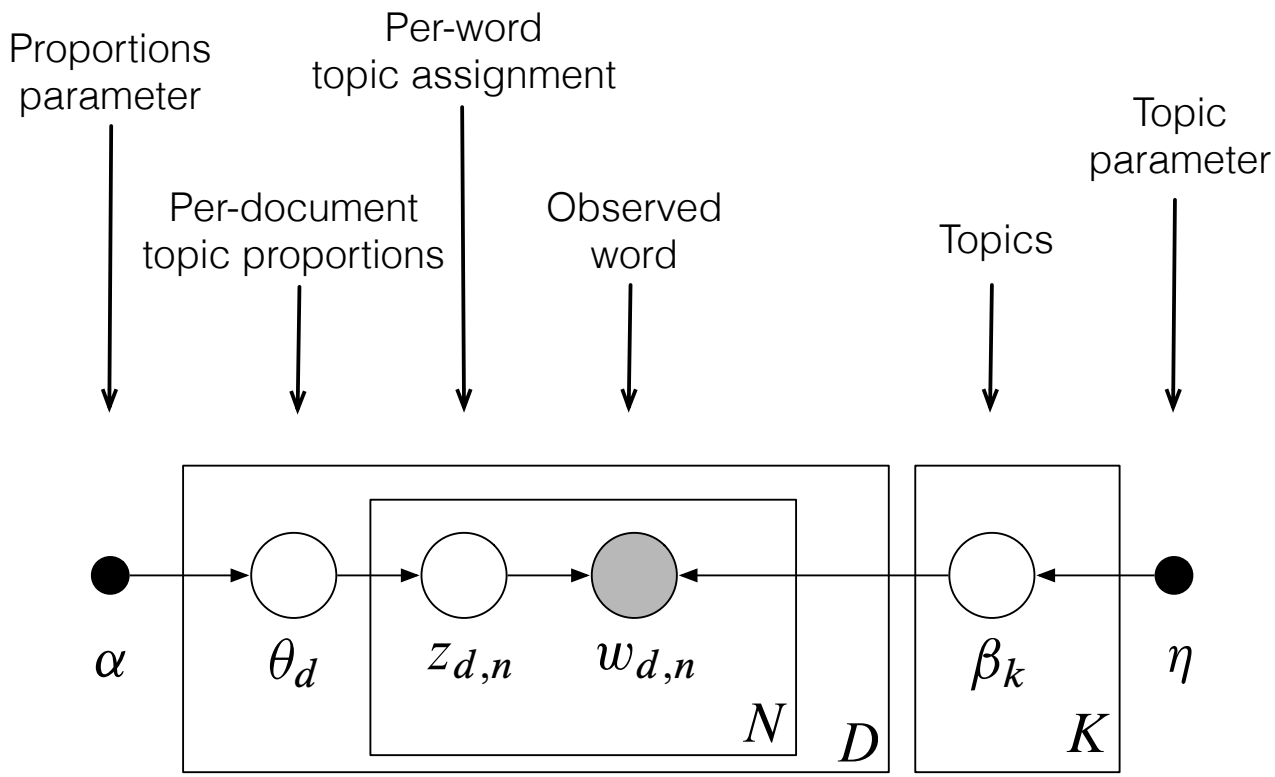
SCIENCE • VOL. 272 • 24 MAY 1996

Latent Dirichlet Allocation



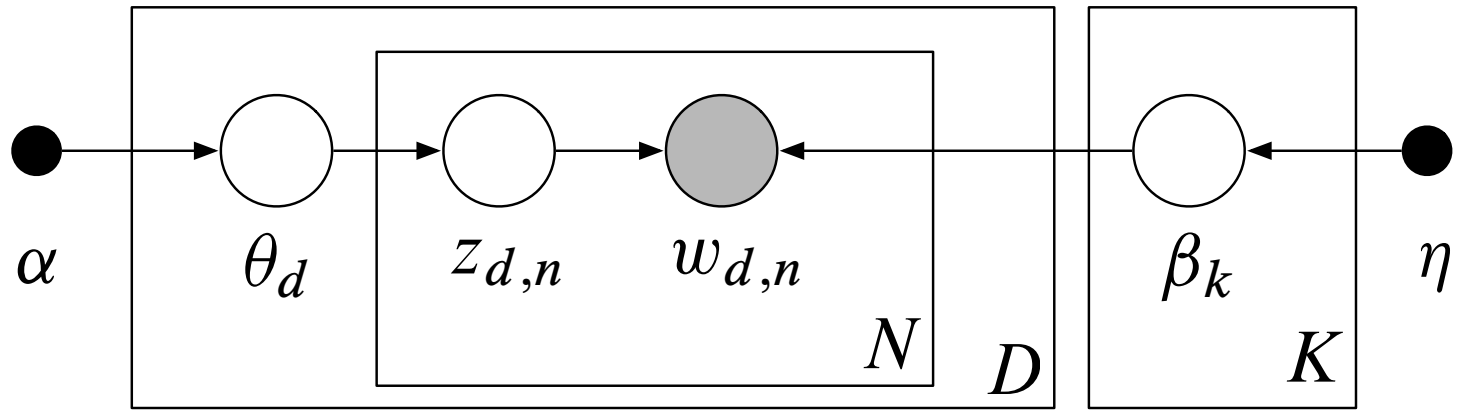
LDA as a graphical model

- Nodes are random variables; edges indicate dependence.
- Shaded nodes are observed; unshaded nodes are hidden.
- Plates indicate replicated variables.

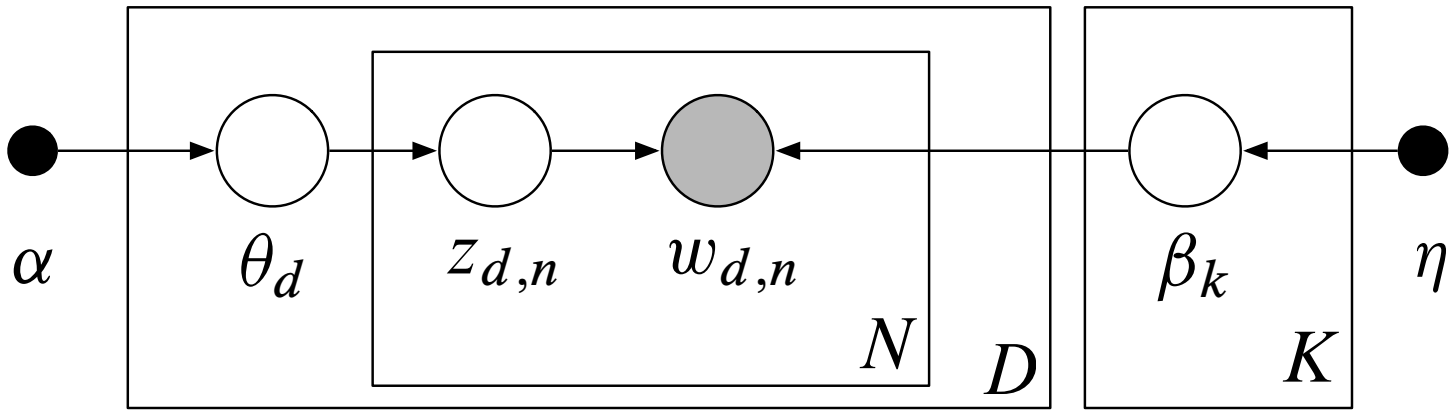


LDA as a graphical model

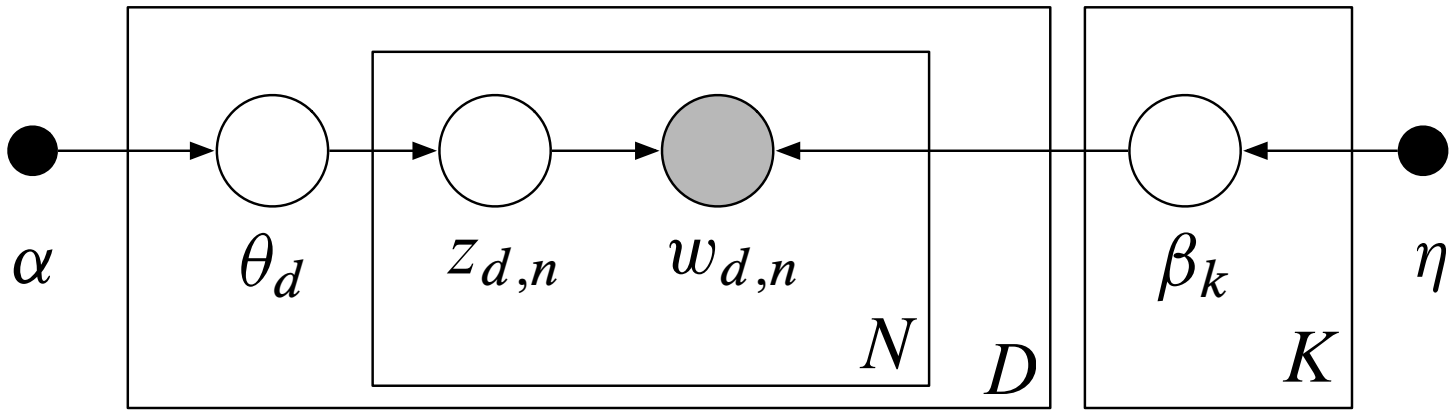
- Encodes independence assumptions about the variables
- Defines a factorization of the joint probability distribution
- Connects to algorithms for computing with data



- The joint defines a posterior, $p(\theta, z, \beta \mid w)$.
- From a collection of documents, infer
  - Per-word topic assignment $z_{d,n}$
  - Per-document topic proportions θ_d
  - Per-corpus topic distributions β_k
- Then use posterior expectations to perform the task at hand: information retrieval, document similarity, exploration, and others.



- Mean field variational methods (Blei et al., 2001, 2003)
- Expectation propagation (Minka and Lafferty, 2002)
- Collapsed Gibbs sampling (Griffiths and Steyvers, 2002)
- Distributed sampling (Newman et al., 2008; Ahmed et al., 2012)
- Collapsed variational inference (Teh et al., 2006)
- Stochastic inference (Hoffman et al., 2010, 2013; Mimno et al., 2012)
- Factorization inference (Arora et al., 2012; Anandkumar et al., 2012)



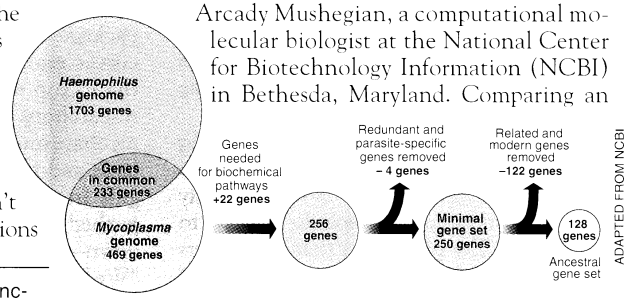
- LDA in R
- GenSim [https://radimrehurek.com/gensim]
- Mallet [http://mallet.cs.umass.edu]
- Vowpal Wabbit [http://hunch.net/∼vw/]
- Apache Spark [http://spark.apache.org/]
- SciKit Learn [http://scikit-learn.org/]

- **Data**: The OCR'ed collection of *Science* from 1990–2000
  - 17K documents
  - 11M words
  - 20K unique terms (stop words and rare words removed)
- **Model**: 100-topic LDA model using variational inference.

Seeking Life's Bare (Genetic) Necessities

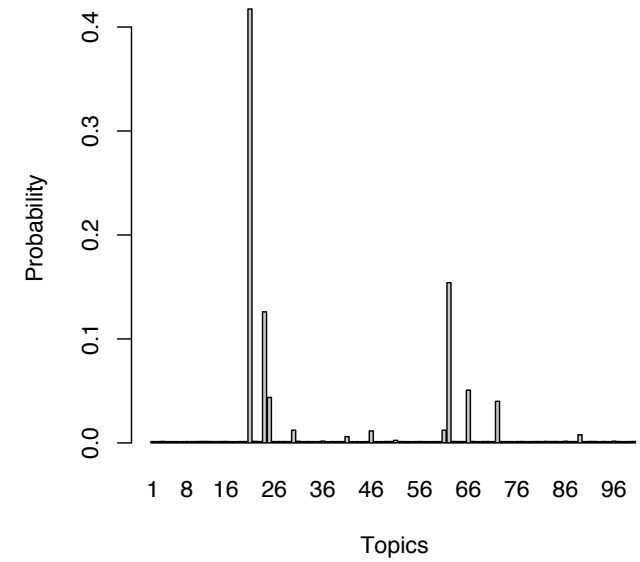
COLD SPRING HARBOR, NEW YORK—How many genes does an organism need to survive? Last week at the genome meeting here,* two genome researchers with radically different approaches presented complementary views of the basic genes needed for life. One research team, using computer analyses to compare known genomes, concluded that today's organisms can be sustained with just 250 genes, and that the earliest life forms required a mere 128 genes. The other researcher mapped genes in a simple parasite and estimated that for this organism, 800 genes are plenty to do the job—but that anything short of 100 wouldn't be enough.

Although the numbers don't match precisely, those predictions "are not all that far apart," especially in comparison to the 75,000 genes in the human genome, notes Siv Andersson of Uppsala University in Sweden, who arrived at the 800 number. But coming up with a consensus answer may be more than just a genetic numbers game, particularly as more and more genomes are completely mapped and sequenced. "It may be a way of organizing any newly sequenced genome," explains Arcady Mushegian, a computational molecular biologist at the National Center for Biotechnology Information (NCBI) in Bethesda, Maryland. Comparing an



Stripping down. Computer analysis yields an estimate of the minimum modern and ancient genomes.

* Genome Mapping and Sequencing, Cold Spring Harbor, New York, May 8 to 12.



human	evolution	disease	computer
genome	evolutionary	host	models
dna	species	bacteria	information
genetic	organisms	diseases	data
genes	life	resistance	computers
sequence	origin	bacterial	system
gene	biology	new	network
molecular	groups	strains	systems
sequencing	phylogenetic	control	model
map	living	infectious	parallel
information	diversity	malaria	methods
genetics	group	parasite	networks
mapping	new	parasites	software
project	two	united	new
sequences	common	tuberculosis	simulations

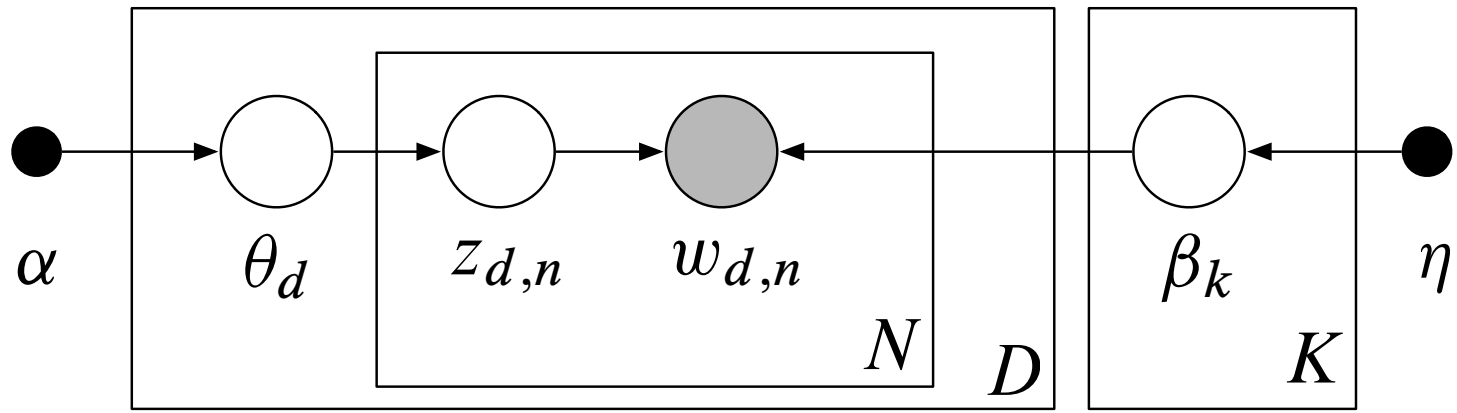
1	2	3	4	5
Game	Life	Film	Book	Wine
Season	Know	Movie	Life	Street
Team	School	Show	Books	Hotel
Coach	Street	Life	Novel	House
Play	Man	Television	Story	Room
Points	Family	Films	Man	Night
Games	Says	Director	Author	Place
Giants	House	Man	House	Restaurant
Second	Children	Story	War	Park
Players	Night	Says	Children	Garden

6	7	8	9	10
Bush	Building	Won	Yankees	Government
Campaign	Street	Team	Game	War
Clinton	Square	Second	Mets	Military
Republican	Housing	Race	Season	Officials
House	House	Round	Run	Iraq
Party	Buildings	Cup	League	Forces
Democratic	Development	Open	Baseball	Iraqi
Political	Space	Game	Team	Army
Democrats	Percent	Play	Games	Troops
Senator	Real	Win	Hit	Soldiers

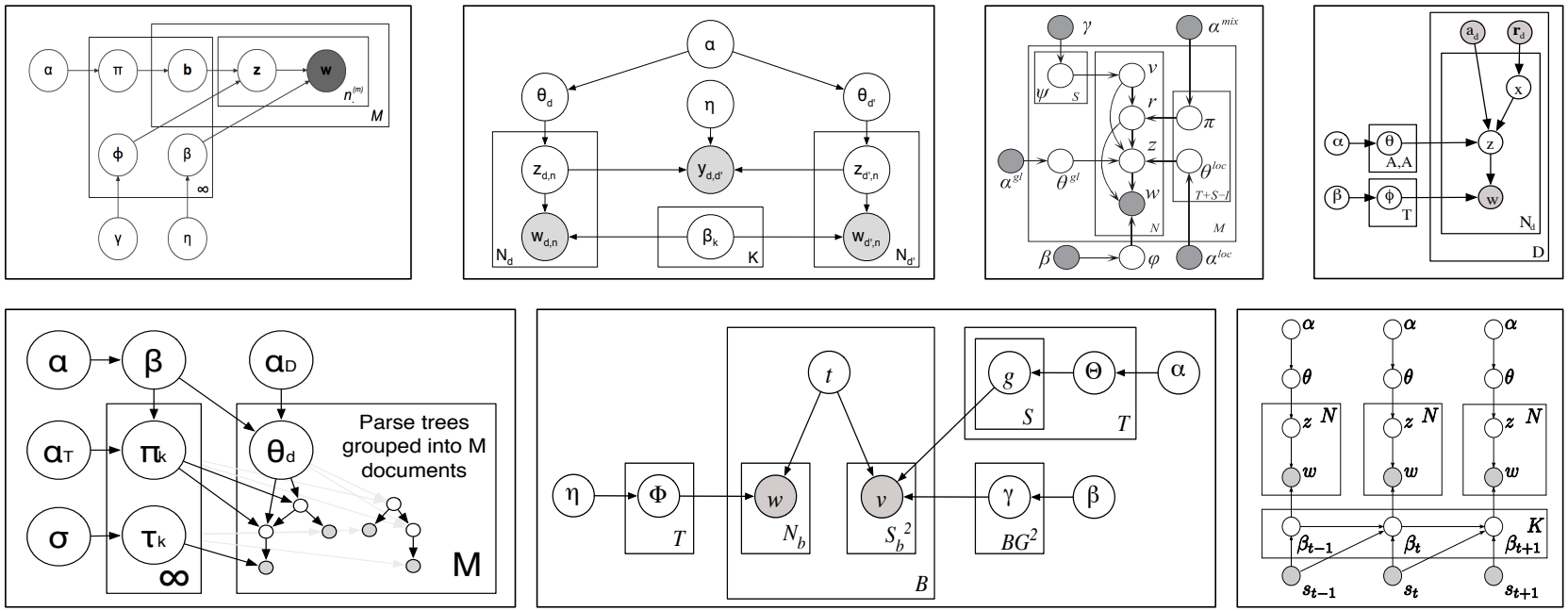
11	12	13	14	15
Children	Stock	Church	Art	Police
School	Percent	War	Museum	Yesterday
Women	Companies	Women	Show	Man
Family	Fund	Life	Gallery	Officer
Parents	Market	Black	Works	Officers
Child	Bank	Political	Artists	Case
Life	Investors	Catholic	Street	Found
Says	Funds	Government	Artist	Charged
Help	Financial	Jewish	Paintings	Street
Mother	Business	Pope	Exhibition	Shot

How does LDA "work"?

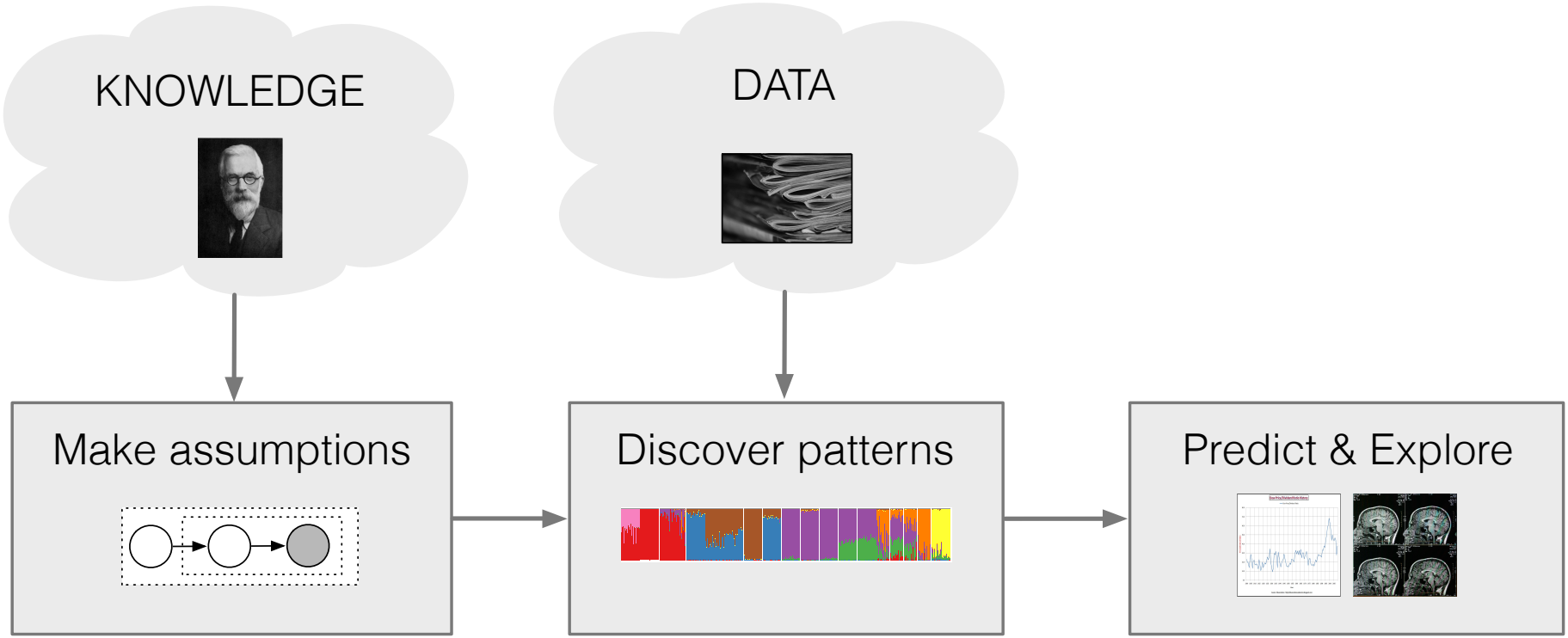
- LDA trades off two goals.
  1. In each **document**, allocate its words to **few topics**.
  2. In each **topic**, assign high probability to **few terms**.
- These goals are at odds.
  - Putting a document in a single topic makes #2 hard:
    All of its words must have probability under that topic.
  - Putting very few words in each topic makes #1 hard:
    To cover a document's words, it must assign many topics to it.
- Trading off these goals finds groups of tightly co-occurring words.



- Summary: LDA discovers themes through posterior inference.
- Other perspectives
  - Latent semantic analysis [Deerwester et al., 1990; Hofmann, 1999]
  - A mixed-membership model [Erosheva, 2004]
  - PCA and matrix factorization [Jakulin and Buntine, 2002]
  - Was independently invented for genetics [Pritchard et al., 2000]

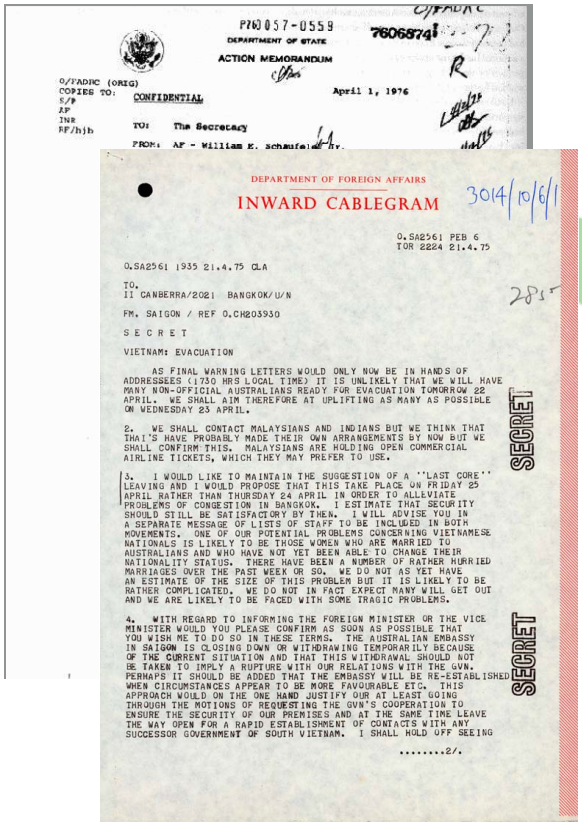


- Organizing and finding patterns in text is important in the sciences, humanities, industry, and culture.
- LDA is a simple building block that enables many applications. Topic modeling is an active field of research.
- Algorithmic improvements let us fit models to massive data.



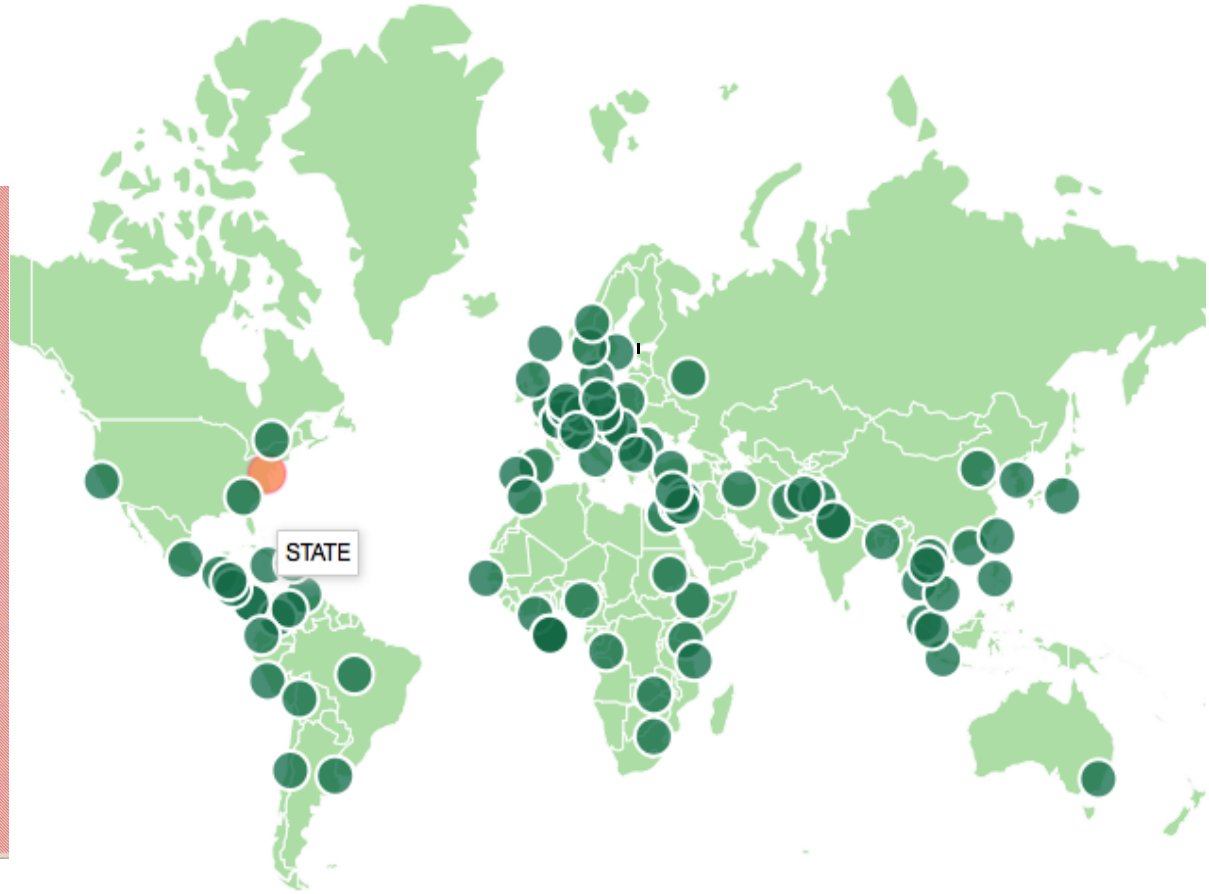
- Case study in **text analysis with probability models**
- Topic modeling research
  - develops new models.
  - develops new inference algorithms.
  - develops new applications, visualizations, tools.

Topic Models for Understanding History



P760057-0559
7606574
DEPARTMENT OF STATE
ACTION MEMORANDUM
O/FADRC (ORIG)
COPIES TO: CONFIDENTIAL
S/P
AF
INR
RF/hjb
April 1, 1976
TO: The Secretary
FROM: AF - William E. Schaufele, Jr.

DEPARTMENT OF FOREIGN AFFAIRS
INWARD CABLEGRAM
3014/10/6/1
O.SA2561 PEB 6
TOR 2224 21.4.75
O.SA2561 1935 21.4.75 CLA
TO.
II CANBERRA/2021 BANGKOK/U/N
FM. SAIGON / REF O.CH203930
S E C R E T
VIETNAM: EVACUATION
AS FINAL WARNING LETTERS WOULD ONLY NOW BE IN HANDS OF ADDRESSEES (1730 HRS LOCAL TIME) IT IS UNLIKELY THAT WE WILL HAVE MANY NON-OFFICIAL AUSTRALIANS READY FOR EVACUATION TOMORROW 22 APRIL. WE SHALL AIM THEREFORE AT UPLIFTING AS MANY AS POSSIBLE ON WEDNESDAY 23 APRIL.
2. WE SHALL CONTACT MALAYSIANS AND INDIANS BUT WE THINK THAT THAI'S HAVE PROBABLY MADE THEIR OWN ARRANGEMENTS BY NOW BUT WE SHALL CONFIRM THIS. MALAYSIANS ARE HOLDING OPEN COMMERCIAL AIRLINE TICKETS, WHICH THEY MAY PREFER TO USE.
3. I WOULD LIKE TO MAINTAIN THE SUGGESTION OF A "LAST CORE" LEAVING AND I WOULD PROPOSE THAT THIS TAKE PLACE ON FRIDAY 25 APRIL RATHER THAN THURSDAY 24 APRIL IN ORDER TO ALLEVIATE PROBLEMS OF CONGESTION IN BANGKOK. I ESTIMATE THAT SECURITY SHOULD STILL BE SATISFACTORY BY THEN. I WILL ADVISE YOU IN A SEPARATE MESSAGE OF LISTS OF STAFF TO BE INCLUDED IN BOTH MOVEMENTS. ONE OF OUR POTENTIAL PROBLEMS CONCERNING VIETNAMESE NATIONALS IS LIKELY TO BE THOSE WOMEN WHO ARE MARRIED TO AUSTRALIANS AND WHO HAVE NOT YET BEEN ABLE TO CHANGE THEIR NATIONALITY STATUS. THERE HAVE BEEN A NUMBER OF RATHER HURRIED MARRIAGES OVER THE PAST WEEK OR SO. WE DO NOT AS YET HAVE AN ESTIMATE OF THE SIZE OF THIS PROBLEM BUT IT IS LIKELY TO BE RATHER COMPLICATED. WE DO NOT IN FACT EXPECT MANY WILL GET OUT AND WE ARE LIKELY TO BE FACED WITH SOME TRAGIC PROBLEMS.
4. WITH REGARD TO INFORMING THE FOREIGN MINISTER OR THE VICE MINISTER WOULD YOU PLEASE CONFIRM AS SOON AS POSSIBLE THAT YOU WISH ME TO DO SO IN THESE TERMS. THE AUSTRALIAN EMBASSY IN SAIGON IS CLOSING DOWN OR WITHDRAWING TEMPORARILY BECAUSE OF THE CURRENT SITUATION AND THAT THIS WITHDRAWAL SHOULD NOT BE TAKEN TO IMPLY A RUPTURE WITH OUR RELATIONS WITH THE GVN. PERHAPS IT SHOULD BE ADDED THAT THE EMBASSY WILL BE RE-ESTABLISHED WHEN CIRCUMSTANCES APPEAR TO BE MORE FAVOURABLE ETC. THIS APPROACH WOULD ON THE ONE HAND JUSTIFY OUR AT LEAST GOING THROUGH THE MOTIONS OF REQUESTING THE GVN'S COOPERATION TO ENSURE THE SECURITY OF OUR PREMISES AND AT THE SAME TIME LEAVE THE WAY OPEN FOR A RAPID ESTABLISHMENT OF CONTACTS WITH ANY SUCCESSOR GOVERNMENT OF SOUTH VIETNAM. I SHALL HOLD OFF SEEING
.....2/.
SECRET
SECRET



- Historians want to identify important events from primary sources.
- Example: Embassies send cables to each other during the 1970s
- Goal: Use topic models to find events in this data set

GUARD, MARINE, DETACHMENT, PERSONNEL, MSG, STAFF, COORDINATE, REPAIR, AREA, OPERATIONAL

AFRICAN, MALIAN, AFRICA, OAU, GOM, IVORY, NIGERIA, IVORIAN, ANGOLA, AID

ART, BAND, ILO, BICENTENNIAL, SINGER, JAZZ, CELEBRATION, DANCE, STRING, WIND

DOLLAR, DEPOSIT, FUND, ESTIMATE, REGISTRY, IRAQ, SPEC, POUND, COST, BANK

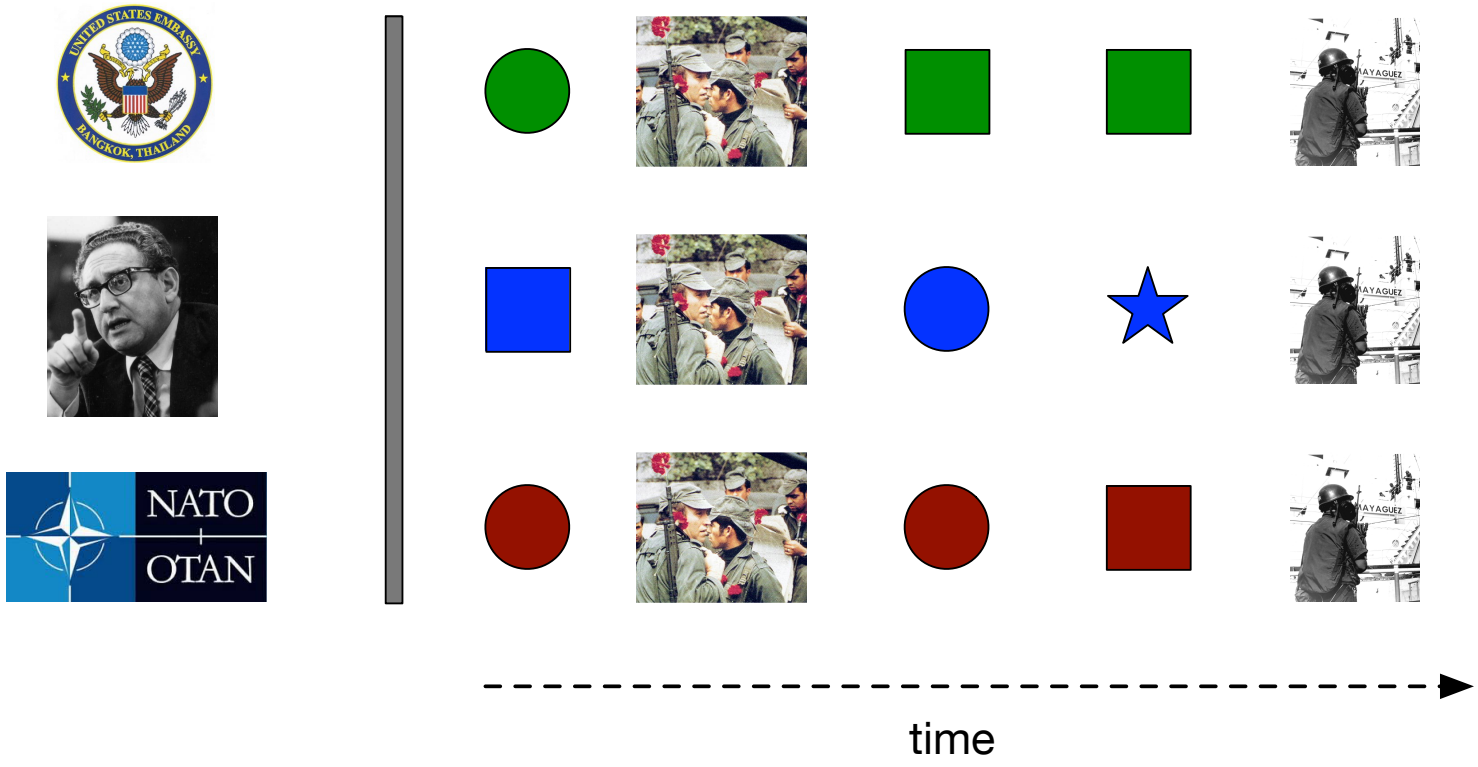
PROGRAM, TRAVEL, NOMINATE, VISIT, PLAN, ARRIVAL, IVP, HOTEL, UNIVERSITY, ITINERARY

YEAR, TOTAL, DOLLAR, BALANCE, LATIN, FOLLOW, CREDIT, STATE, INCREASE, AMOUNT

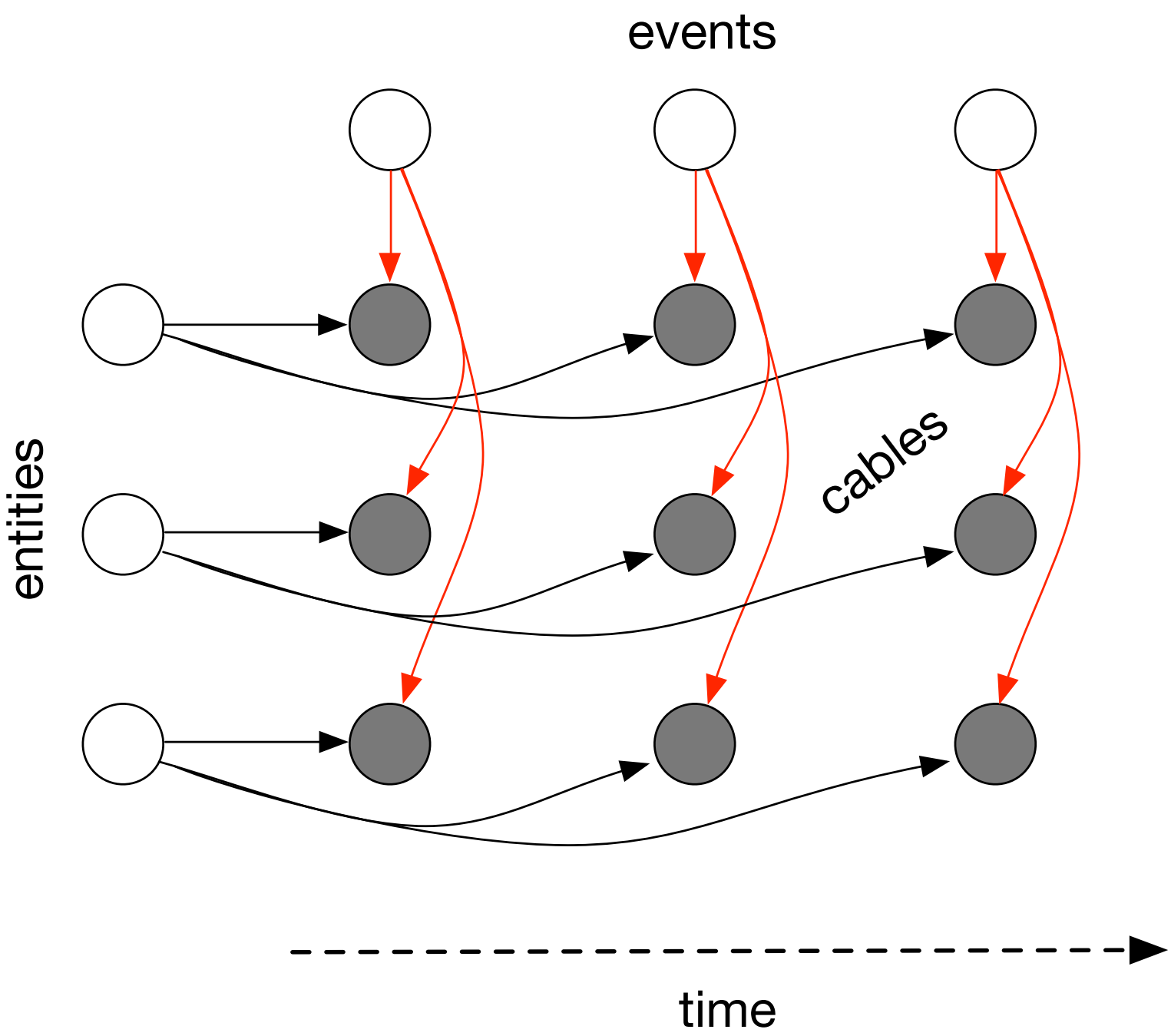
COMPANY, PROJECT, OIL, MILLION, DRILL, CONTRACT, CONSTRUCTION, WATER, KUWAIT, LOCAL

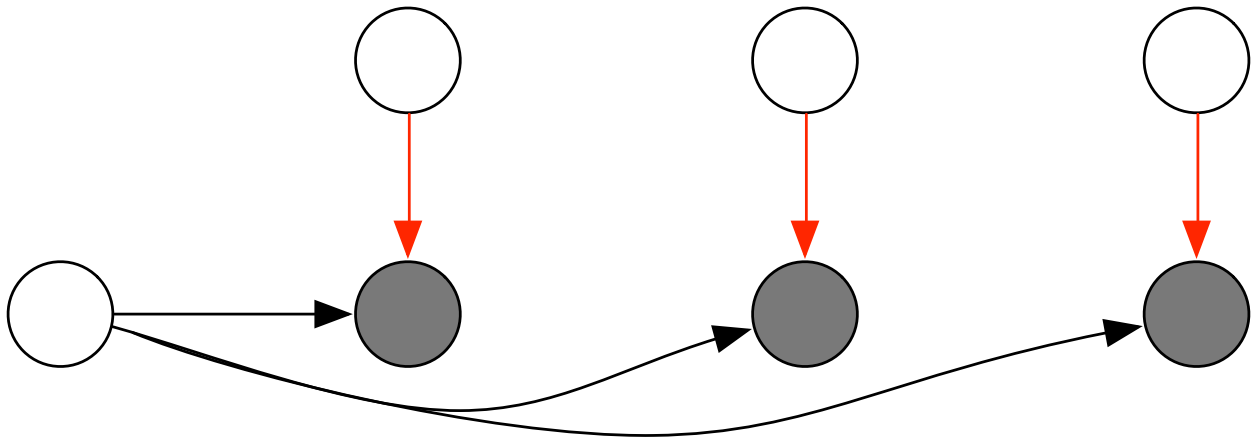
PAYMENT, MAKE, DRAFT, PRIVILEGE, REFTEL, FINAL, RUN, IMMUNITY, FUND, VIEW

Topic models (by themselves) are a start. But they don't identify events.



- Embassies typically discuss their *usual business*
- When a cable is about an **event**:
  - It diverges from the usual business of the sender
  - Multiple embassies discuss it
- *Usual business* is framed in terms of topics;
  events are framed in terms of words.


events
entities
cables
time

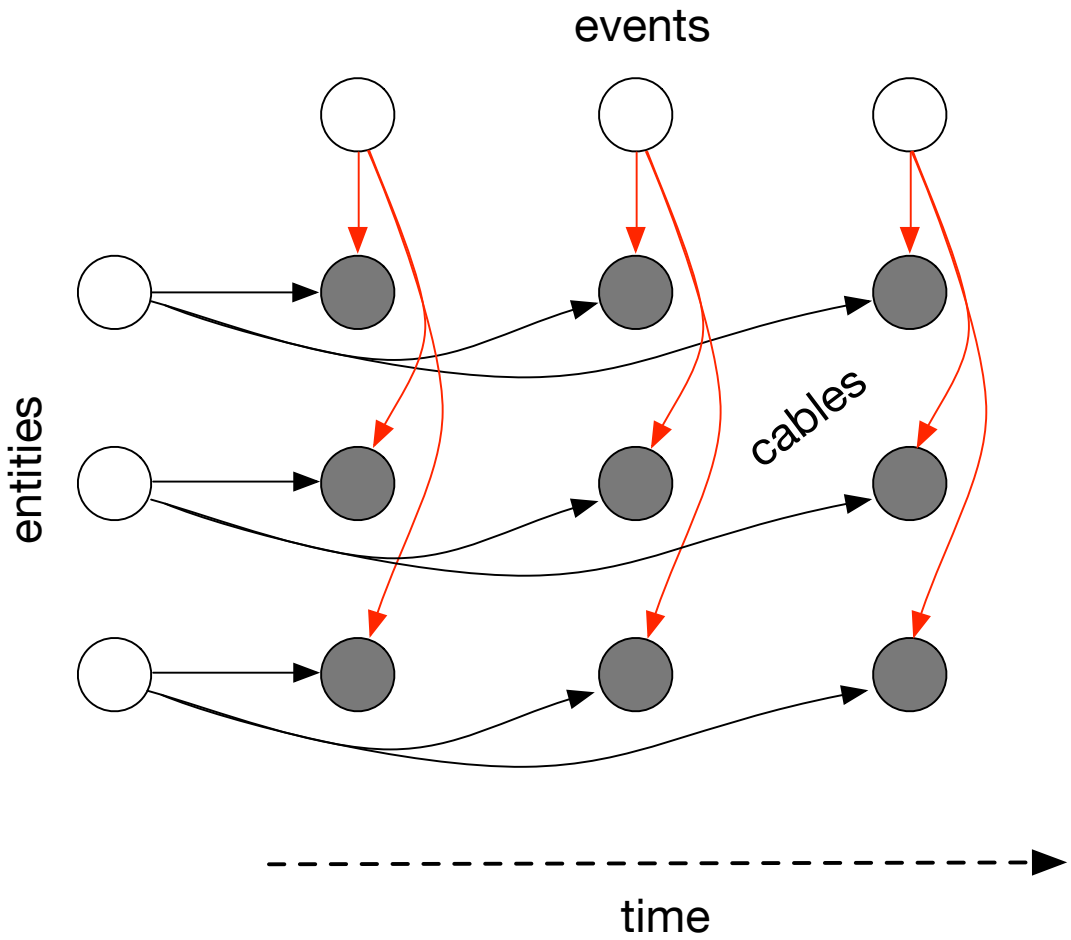


Hidden variables

- Topics
- Event description (per week)
- Topic description (per entity)
- Topic strength, event strength (per cable)

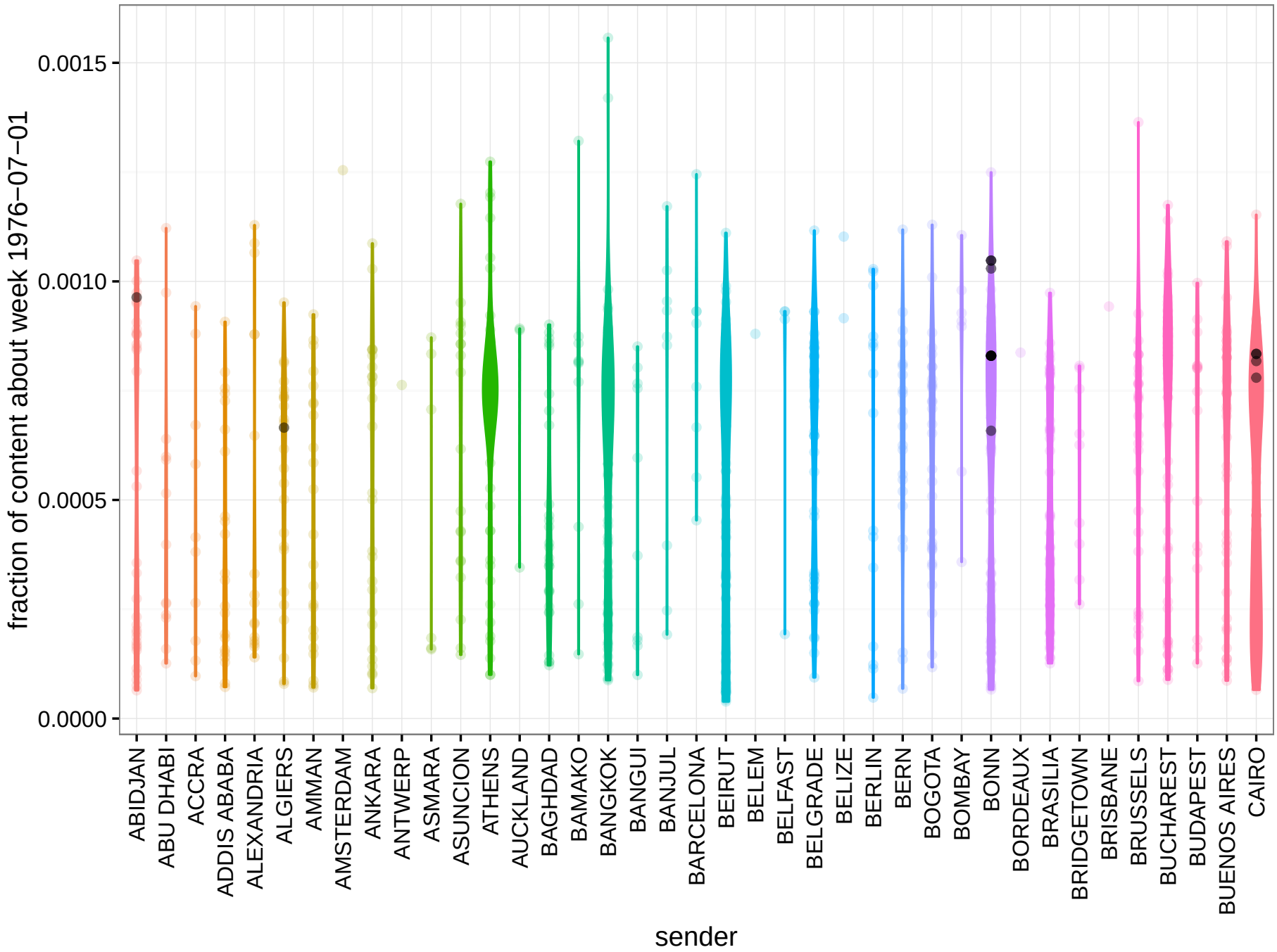
Observed variables

- Cables (per-week, per-entity)



To find events:

- Calculate the posterior of the hidden variables given the observed variables
- Examine cables where the event strength is high


fraction of content about week 1976–07–01
0.0015
0.0010
0.0005
0.0000
ABIDJAN
ABU DHABI
ACCRA
ADDIS ABABA
ALEXANDRIA
ALGIERS
AMMAN
AMSTERDAM
ANKARA
ANTWERP
ASMARA
ASUNCION
ATHENS
AUCKLAND
BAGHDAD
BAMAKO
BANGKOK
BANGUI
BANJUL
BARCELONA
BEIRUT
BELEM
BELFAST
BELGRADE
BELIZE
BERLIN
BERN
BOGOTA
BOMBAY
BONN
BORDEAUX
BRASILIA
BRIDGETOWN
BRISBANE
BRUSSELS
BUCHAREST
BUDAPEST
BUENOS AIRES
CAIRO
sender

1. FRG INFORMED EMBASSY AT 2130 LOCAL THAT FRENCH GOVERNMENT HAS PASSED IT THE FOLLOWING MESSAGE: A. THE HIJACKERS REJECT ANY EXCHANGE OUTSIDE OF ENTEBBE AIRPORT. B. THE EXCHANGE MUST TAKE PLACE UNDER THE SUPERVISION OF AMIN (OR OTHER HIGH UGANDAN OFFICIAL), TWO FRENCH REPS, AND SOMALI AMBASSADOR. C. HIJACKERS NOT PREPARED TO DISTINGUISH BETWEEN THE PRISONERS HELD IN ISRAEL AND THOSE HELD IN OTHER COUNTRIES. D. ALL HOSTAGES MUST BE EXCHANGED AGAINST ALL THE PRISONERS. E. THE HIJACKERS EXPECT AN ANSWER FROM ALL FOUR COUNTRIES HOLDING PRISONERS, NOT ONLY ISRAEL. F. THE HIJACKERS REFER TO THE COMPLETE LIST OF 53 "COMRADES." G. THE HIJACKERS INSIST AGAIN ON A PACKAGE DEAL: 53 "COMRADES" AGAINST ALL THE HOSTAGES AT ENTEBBE AIRPORT. 2. THE FRENCH PASSED A SECOND MESSAGE, THIS ONE FROM AMIN: AMIN TOLD THE FRENCH AMBASSADOR THAT HE EXPECTS ALL FOUR COUNTRIES TO COMMUNICATE TO HIM THE FLIGHT NUMBERS AND ETA OF ALL AIRCRAFT BRINGING PRISONERS TO UGANDA BEFORE THE END OF THE ULTIMATUM. (HE DID NOT SPECIFY AN HOUR.) 3. THE FRG CRISIS CENTER TOLD US THE FRG IS CONSULTING WITH THE OTHER GOVERNMENTS INVOLVED AT THE HIGHEST LEVEL. IT HAS NOT RPT NOT REACHED A DECISION ON RELEASE OF PRISONERS. HILLENBRAND

1976-07-03 | BONN | STATE | AIR FRANCE HIJACKING

1. THE FRENCH HAVE TOLD THE FRG FOREIGN OFFICE THAT THE ISRAELIS ARE STILL WILLING TO RELEASE "CERTAIN, BUT NOT ALL, PRISONERS FOR ALL HOSTAGES" AND ARE READY TO DISCUSS THE MODALITIES FOR SUCH AN EXCHANGE. BUT THEY HAVE SAID THAT THEY WILL NOT DELIVER THEIR PRISONERS TO UGANDA, BECAUSE THE ISRAELIS DO NOT TRUST AMIN. 2. KENYA CONTINUES TO DENY THAT IT IS HOLDING ANY PALESTINIANS. 3. THE GERMANS UNDERSTAND THAT THE SWISS WILL GO ALONG WITH WHATEVER THE GERMANS DECIDE. 4. BUT, ACCORDING TO THE FOREIGN OFFICE ASSISTANT DIRECTOR FOR EAST AFRICA, THE GERMANS HAVE NOT REACHED A DECISION. THE CABINET WILL MEET AT 3:00 PM, JULY 3, IN A FURTHER EFFORT TO RESOLVE THE DILEMMA. HILLENBRAND

1976-07-02 | BERN | STATE | AIR FRANCE HIJACKING

1. SWISS FOREIGN OFFICE (KAUFFMAN) INFORMS US THAT ISRAELIS HAD PASSED LIST OF 12-15 IMPRISONED TERRORISTS TO FRENCH GOVT WHOM THEY WERE PREPARED TO RELEASE IN EXCHANGE FOR RELEASE OF ISRAELIS IN KAMPALA. 2. KAUFFMAN SAID THAT GOS IS PREPARED TO REMAIN FIRM, BUT WOULD FOLLOW LEAD OF OTHER COUNTRIES INVOLVED IF THEIR RESOLVE WEAKENS. SWISS HAVE HAD NO DIRECT NEWS FROM TEL AVIV AND ARE CONCERNED THAT ISRAELIS MAY BE PREPARING TO GIVE IN TO HIJACKERS' DEMANDS WITHOUT FULL CONSULTATIONS WITH OTHER COUNTRIES HOLDING TERRORISTS DEMANDED BY HIJACKERS. DAVIS

1976-07-06 | BELGRADE | STATE | REACTION TO ISRAELI RAID

1. SEPARATE USINFO TELEGRAM REPORTS "POLITIKA" EDITORIAL ENTITLED "STATE TERRORISM", WHICH CONDEMNS ISRAELI RAID ON ENTEBBE AIRPORT AS WELL AS ORIGINAL HIJACKING OF PLANE. HIJACKING INITIALLY PRODUCED NO OUTCRY HERE. EDITORIAL AND NEGATIVE PRESS BEING GIVEN TO ISRAELI RAID ARE FURTHER EXAMPLES OF DOUBLE STANDARD WHICH YUGOSLAVS MAINTAIN TOWARDS TERRORISM. SILBERMAN

The way it really happened~The tense, action~packed story of the raid that startled the world.

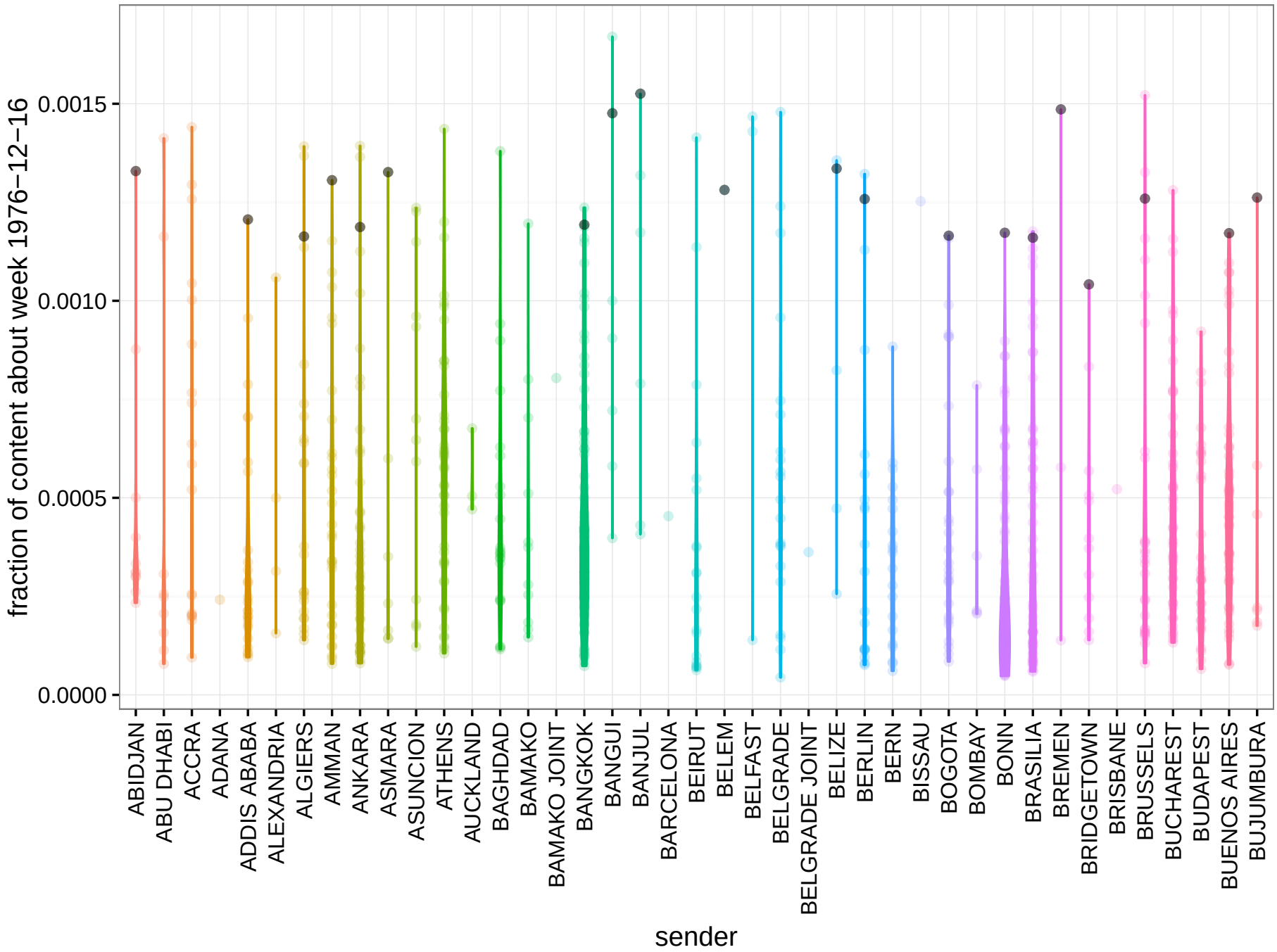
"OPERATION THUNDERBOLT" A

An INTER-OCEAN film "OPERATION THUNDERBOLT" A Golan-Globus Film of a G.S. Films Production
starring YEHORAM GAON · ASSAF DAYAN · KLAUS KINSKY SYBIL DANNING · ORI LEVY · ARIK LAVI and MARK HEATH as Idi Amin
Produced by MENAHEM GOLAN and YORAM GLOBUS Music by DOV SELTZER Directed by MENAHEM GOLAN Eastmancolour Distributed by EMI Films Limited

EMI

FROM THURS. OCT. 20th **ABC 1** **Shaftesbury Ave** Tel: 836 8861 Licensed Bar

AND AT SELECTED **ABC** AND OTHER LEADING CINEMAS FROM OCT. 27th. SEE LOCAL PRESS FOR DETAILS


fraction of content about week 1976−12−16
0.0015
0.0010
0.0005
0.0000
ABIDJAN
ABU DHABI
ACCRA
ADANA
ADDIS ABABA
ALEXANDRIA
ALGIERS
AMMAN
ANKARA
ASMARA
ASUNCION
ATHENS
AUCKLAND
BAGHDAD
BAMAKO
BAMAKO JOINT
BANGKOK
BANGUI
BANJUL
BARCELONA
BEIRUT
BELEM
BELFAST
BELGRADE
BELGRADE JOINT
BELIZE
BERLIN
BERN
BISSAU
BOGOTA
BOMBAY
BONN
BRASILIA
BREMEN
BRIDGETOWN
BRISBANE
BRUSSELS
BUCHAREST
BUDAPEST
BUENOS AIRES
BUJUMBURA
sender

- 1976-12-21 | BANJUL | STATE | GAMING DEVICES
  THERE ARE NO GAMING DEVICES IN EMBASSY FACILITIES.WYGANT

- 1976-12-22 | BREMEN | STATE | GAMING DEVICES
  NO RPT NO GAMING DEVICES OF ANY TYPE AT AMCONSUL BREMEN. LONGMYER

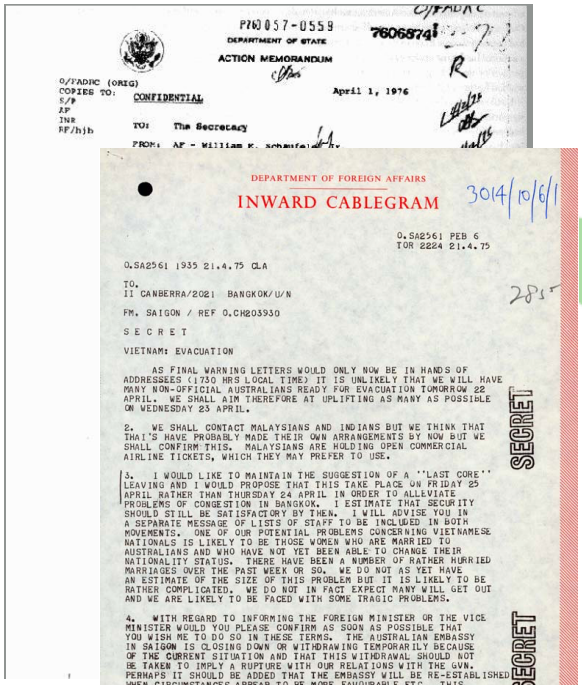
- 1976-12-21 | BANGUI | STATE | GAMING DEVICES
  THERE ARE NO RPT NO GAMING DEVICES IN EMBASSY FACILITIES. QUAINTON

- 1976-12-22 | BELIZE | STATE | GAMING DEVICES
  NEGATIVE RESPONSE. WALSH

- 1976-12-21 | ABIDJAN | STATE | GAMING DEVICES
  FOR: LEAMON R. HUNT, DEPUTY ASSISTANT SECRETARY FOR OPERATIONS NO GAMING DEVICES EXIST AT THIS POST. STEARNS

- 1976-12-21 | ASMARA | STATE | GAMING DEVICES
  CONSULATE GENERAL ASMARA NEITHER OPERATES OR OWNS ANY GAMING DEVICES. WAUCHOPE

- 1976-12-21 | AMMAN | STATE | GAMING DEVICES
  EMBASSY AMMAN SUBMITS NEGATIVE REPORT ON AVAILABILITY OF GAMING DEVICES WITHIN MISSION. PICKERING



P760057-0559
DEPARTMENT OF STATE
ACTION MEMORANDUM
7606874
April 1, 1976
CONFIDENTIAL
TO: The Secretary

DEPARTMENT OF FOREIGN AFFAIRS
INWARD CABLEGRAM
3014/10/6/1

O.SA2561 PEB 6
TOR 2224 21.4.75

O.SA2561 1935 21.4.75 CLA

TO.
II CANBERRA/2021 BANGKOK/U/N

FM. SAIGON / REF O.CH203930

S E C R E T

VIETNAM: EVACUATION

AS FINAL WARNING LETTERS WOULD ONLY NOW BE IN HANDS OF ADDRESSEES (1730 HRS LOCAL TIME) IT IS UNLIKELY THAT WE WILL HAVE MANY NON-OFFICIAL AUSTRALIANS READY FOR EVACUATION TOMORROW 22 APRIL. WE SHALL AIM THEREFORE AT UPLIFTING AS MANY AS POSSIBLE ON WEDNESDAY 23 APRIL.

2. WE SHALL CONTACT MALAYSIANS AND INDIANS BUT WE THINK THAT THAI'S HAVE PROBABLY MADE THEIR OWN ARRANGEMENTS BY NOW BUT WE SHALL CONFIRM THIS. MALAYSIANS ARE HOLDING OPEN COMMERCIAL AIRLINE TICKETS, WHICH THEY MAY PREFER TO USE.

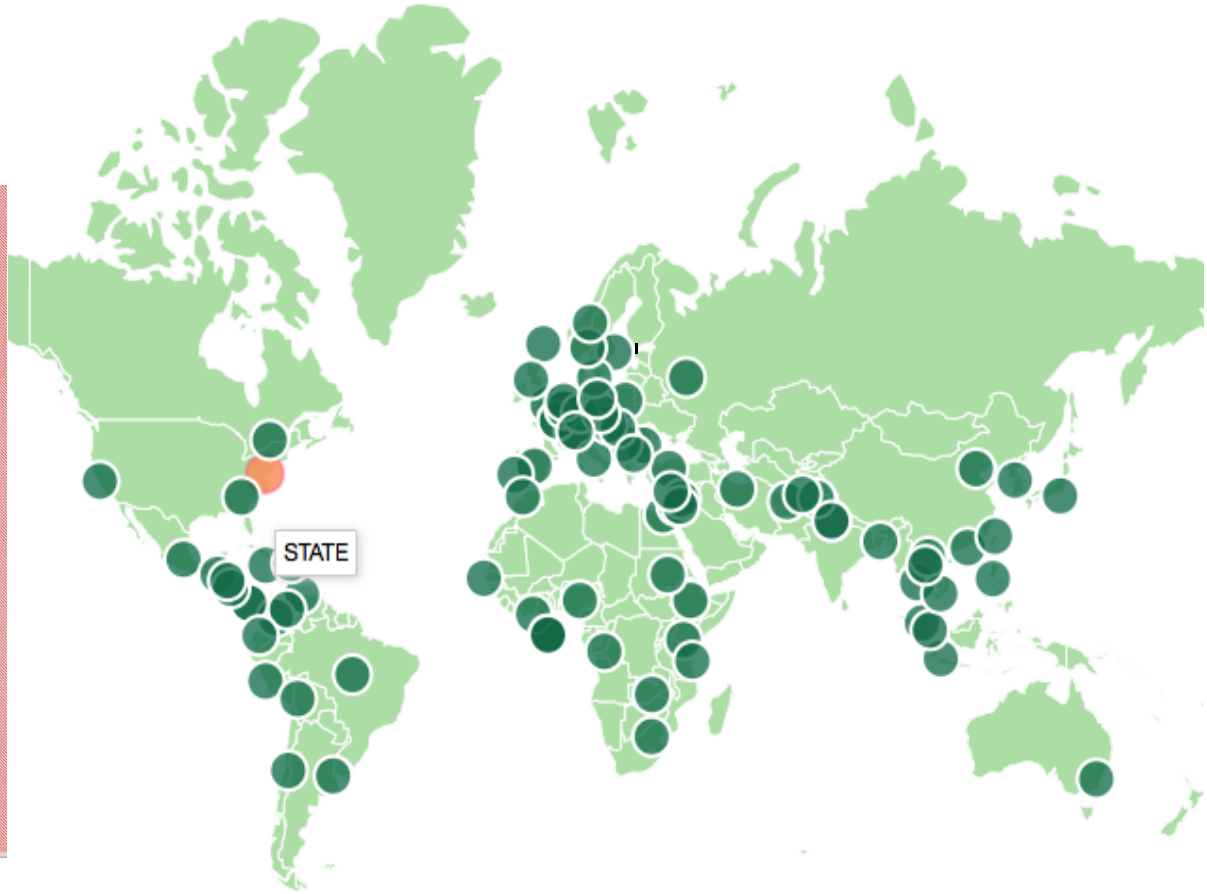
3. I WOULD LIKE TO MAINTAIN THE SUGGESTION OF A "LAST CORE" LEAVING AND I WOULD PROPOSE THAT THIS TAKE PLACE ON FRIDAY 25 APRIL RATHER THAN THURSDAY 24 APRIL IN ORDER TO ALLEVIATE PROBLEMS OF CONGESTION IN BANGKOK. I ESTIMATE THAT SECURITY SHOULD STILL BE SATISFACTORY BY THEN. I WILL ADVISE YOU IN A SEPARATE MESSAGE OF LISTS OF STAFF TO BE INCLUDED IN BOTH MOVEMENTS. ONE OF OUR POTENTIAL PROBLEMS CONCERNING VIETNAMESE NATIONALS IS LIKELY TO BE THOSE WOMEN WHO ARE MARRIED TO AUSTRALIANS AND WHO HAVE NOT YET BEEN ABLE TO CHANGE THEIR NATIONALITY STATUS. THERE HAVE BEEN A NUMBER OF RATHER HURRIED MARRIAGES OVER THE PAST WEEK OR SO. WE DO NOT AS YET HAVE AN ESTIMATE OF THE SIZE OF THIS PROBLEM BUT IT IS LIKELY TO BE RATHER COMPLICATED. WE DO NOT IN FACT EXPECT MANY WILL GET OUT AND WE ARE LIKELY TO BE FACED WITH SOME TRAGIC PROBLEMS.

4. WITH REGARD TO INFORMING THE FOREIGN MINISTER OR THE VICE MINISTER WOULD YOU PLEASE CONFIRM AS SOON AS POSSIBLE THAT YOU WISH ME TO DO SO IN THESE TERMS. THE AUSTRALIAN EMBASSY IN SAIGON IS CLOSING DOWN OR WITHDRAWING TEMPORARILY BECAUSE OF THE CURRENT SITUATION AND THAT THIS WITHDRAWAL SHOULD NOT BE TAKEN TO IMPLY A RUPTURE WITH OUR RELATIONS WITH THE GVN. PERHAPS IT SHOULD BE ADDED THAT THE EMBASSY WILL BE RE-ESTABLISHED WHEN CIRCUMSTANCES APPEAR TO BE MORE FAVOURABLE ETC. THIS APPROACH WOULD ON THE ONE HAND JUSTIFY OUR AT LEAST GOING THROUGH THE MOTIONS OF REQUESTING THE GVN'S COOPERATION TO ENSURE THE SECURITY OF OUR PREMISES AND AT THE SAME TIME LEAVE THE WAY OPEN FOR A RAPID ESTABLISHMENT OF CONTACTS WITH ANY SUCCESSOR GOVERNMENT OF SOUTH VIETNAM. I SHALL HOLD OFF SEEING

.....2/.

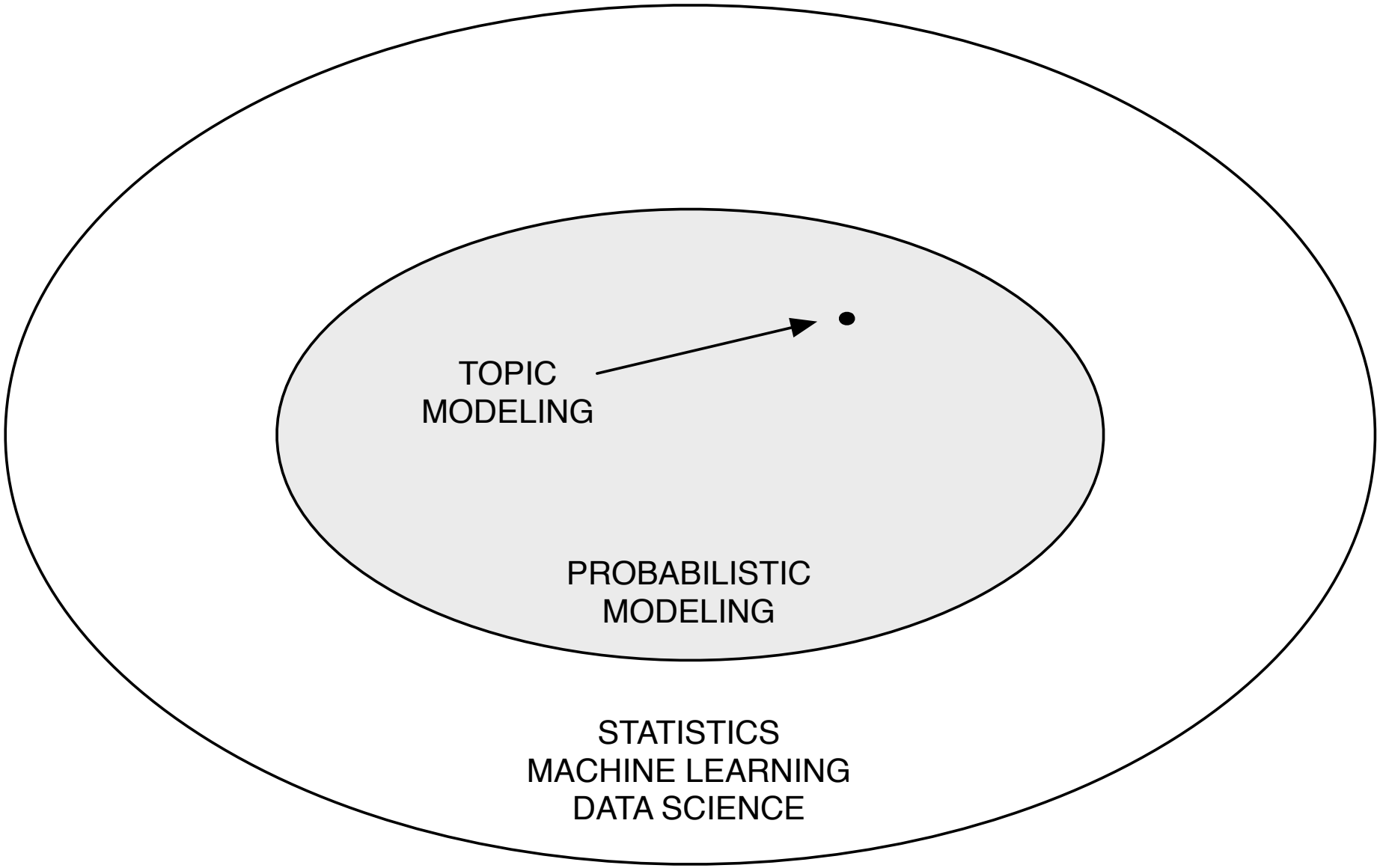
SECRET

SECRET

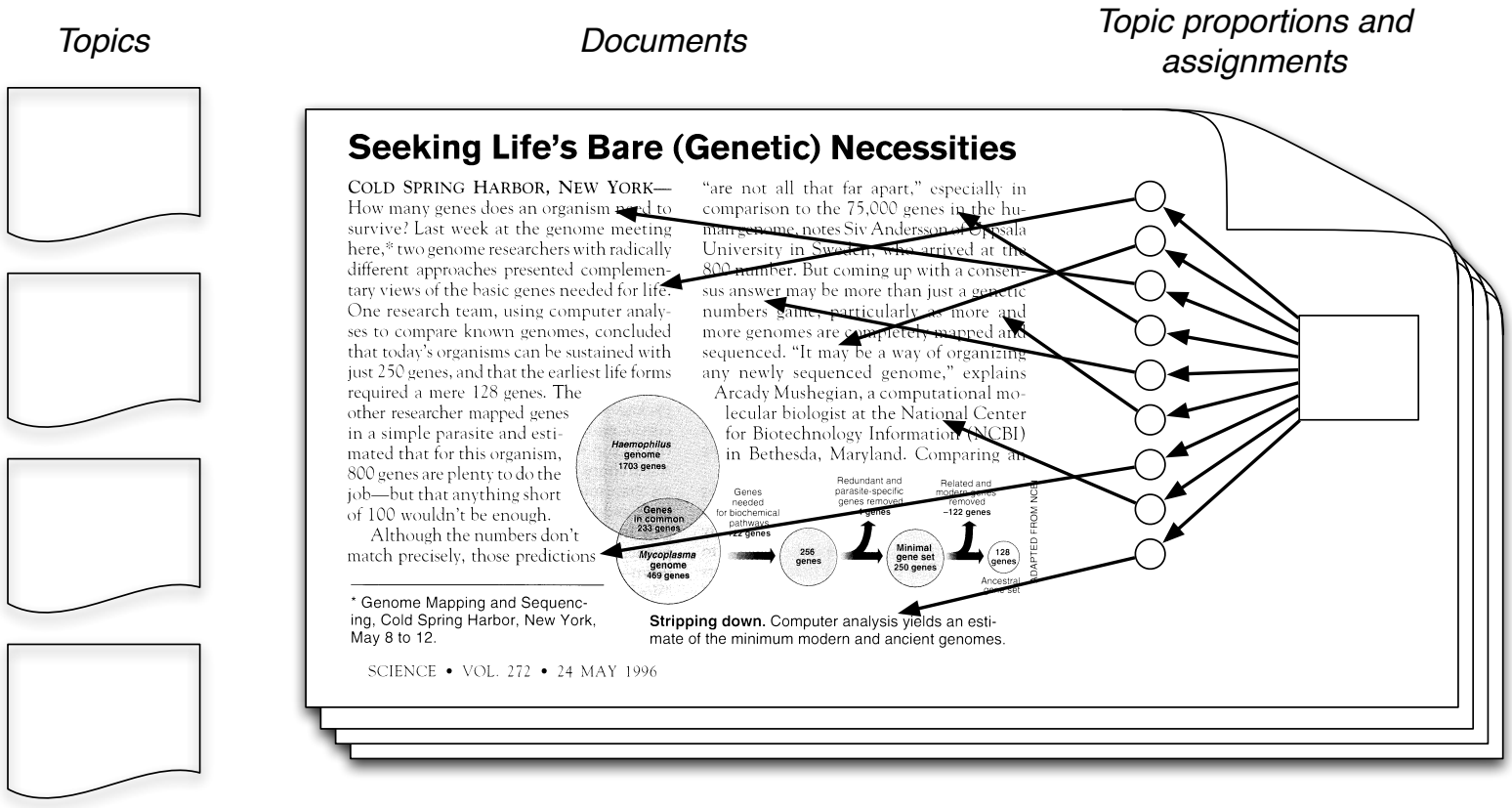


- Topic models can help us find events
- Extensions:
  - Network characteristics
  - Better characterize an event for fewer “false positives”
  - Word embeddings
  - Autocorrelated time series

Discussion: Modern Probabilistic Modeling


TOPIC
MODELING
PROBABILISTIC
MODELING
STATISTICS
MACHINE LEARNING
DATA SCIENCE

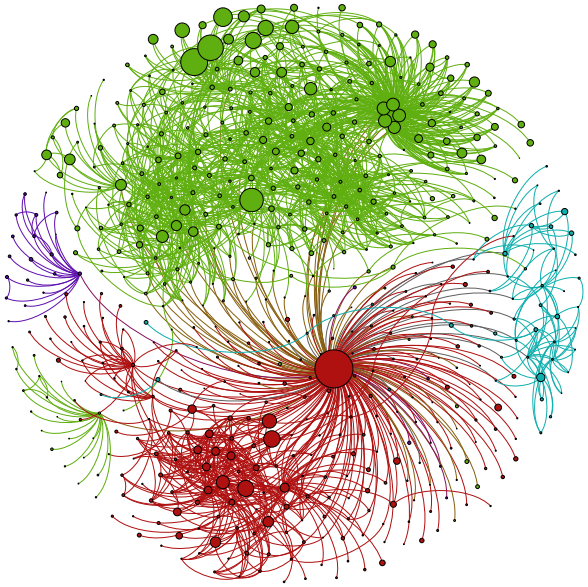
I. Assume our data come from a model with hidden patterns at work

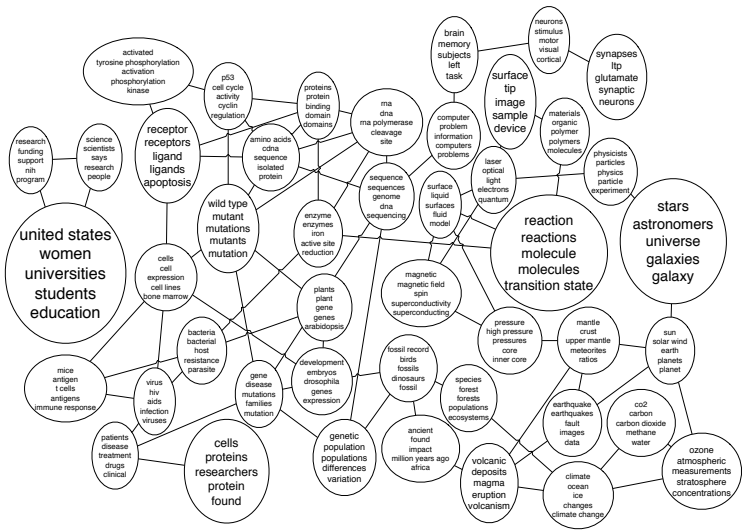


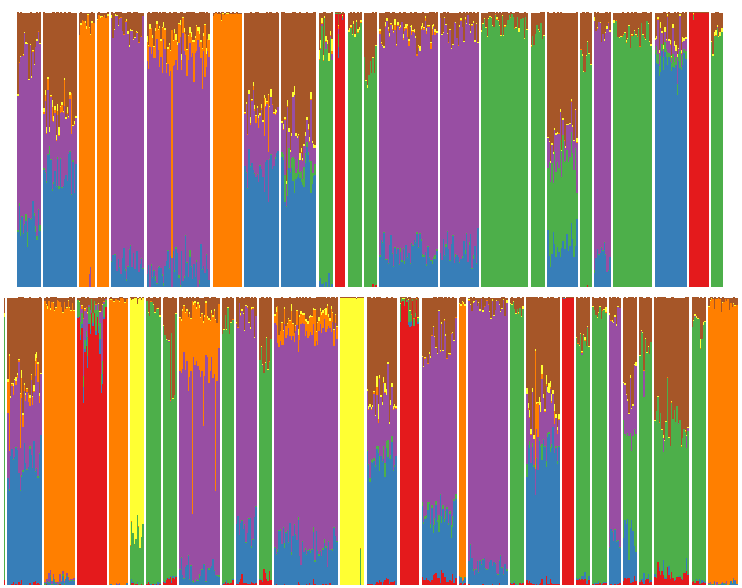
II. Discover those patterns from data

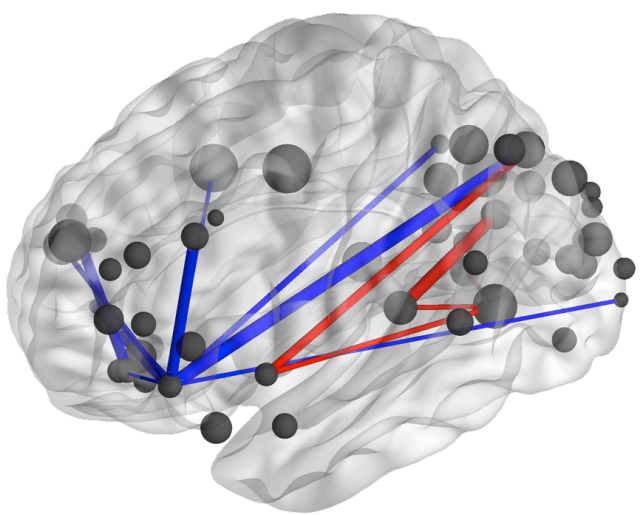
$$\nu^* = \arg\max_{\nu} \mathbb{E}_q \left[\log p(x, z, \beta \mid \alpha)\right] + \mathbb{H}\left[q(z, \beta \mid \nu)\right]$$

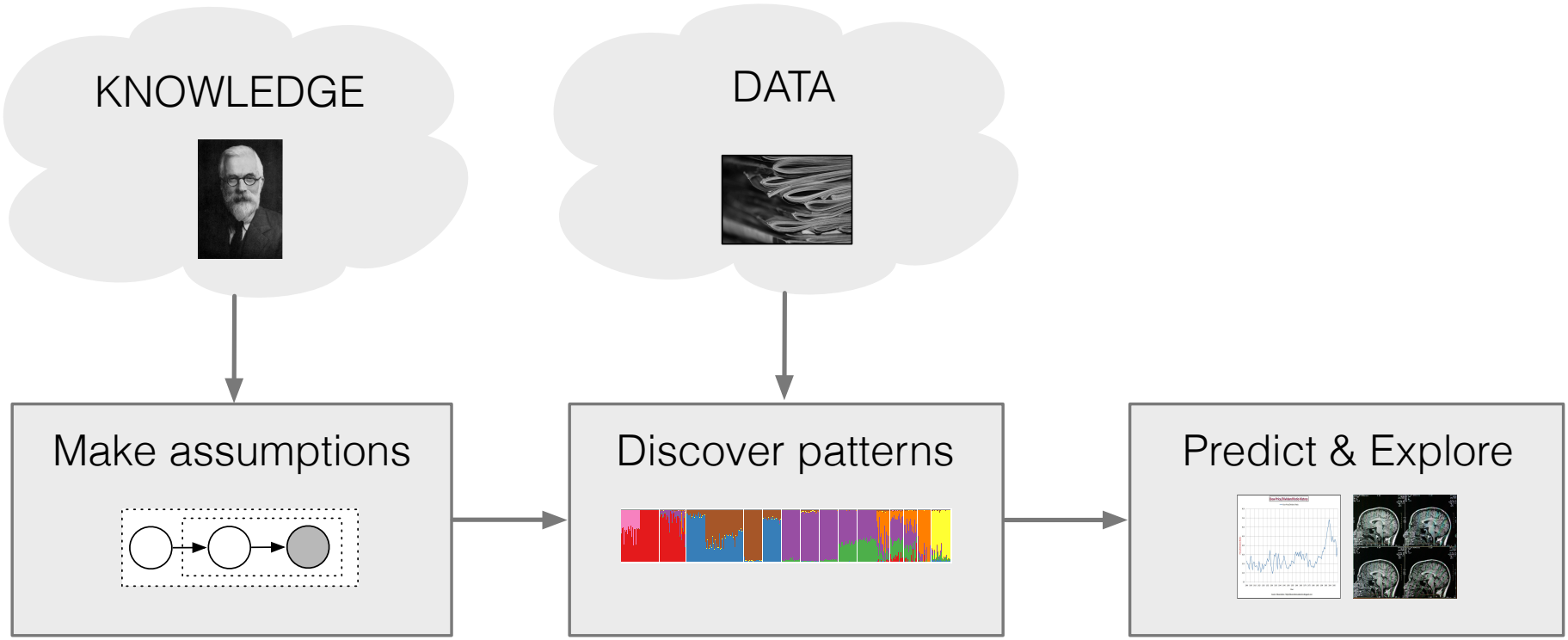
III. Use the discovered patterns to predict about and explore the data





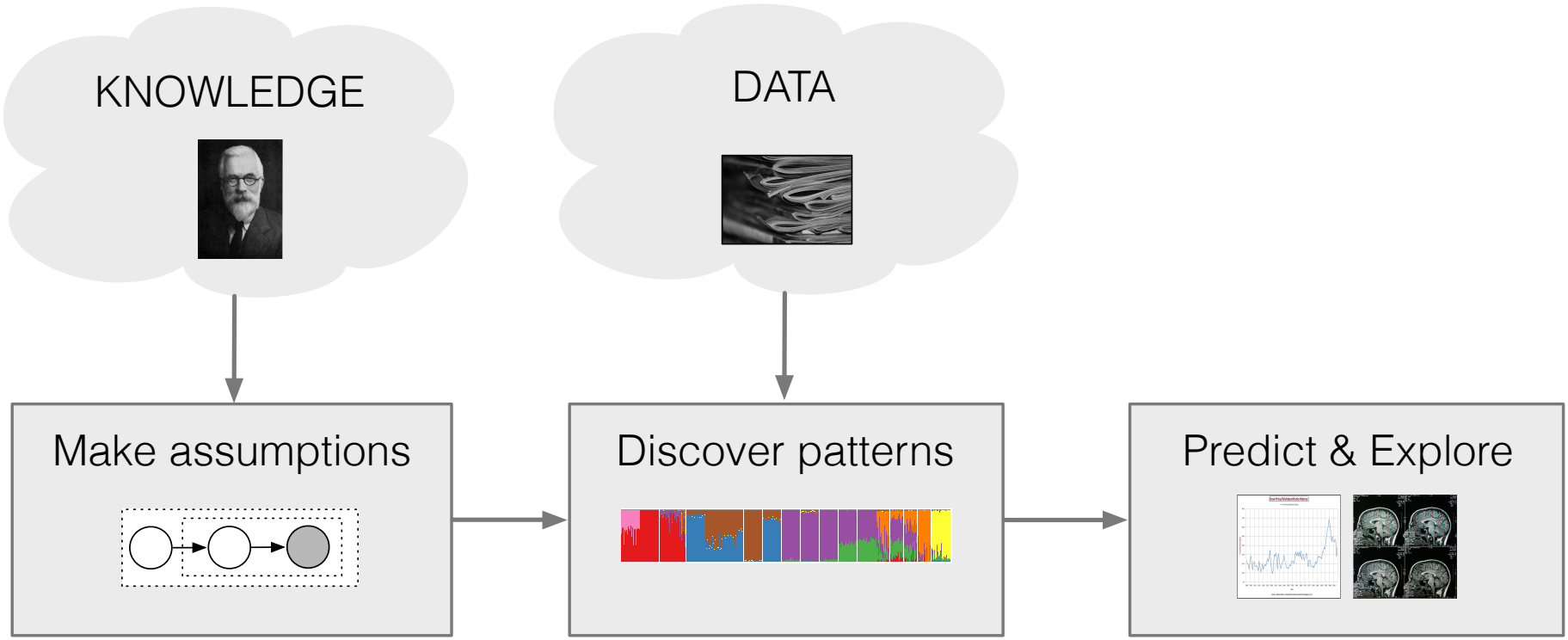






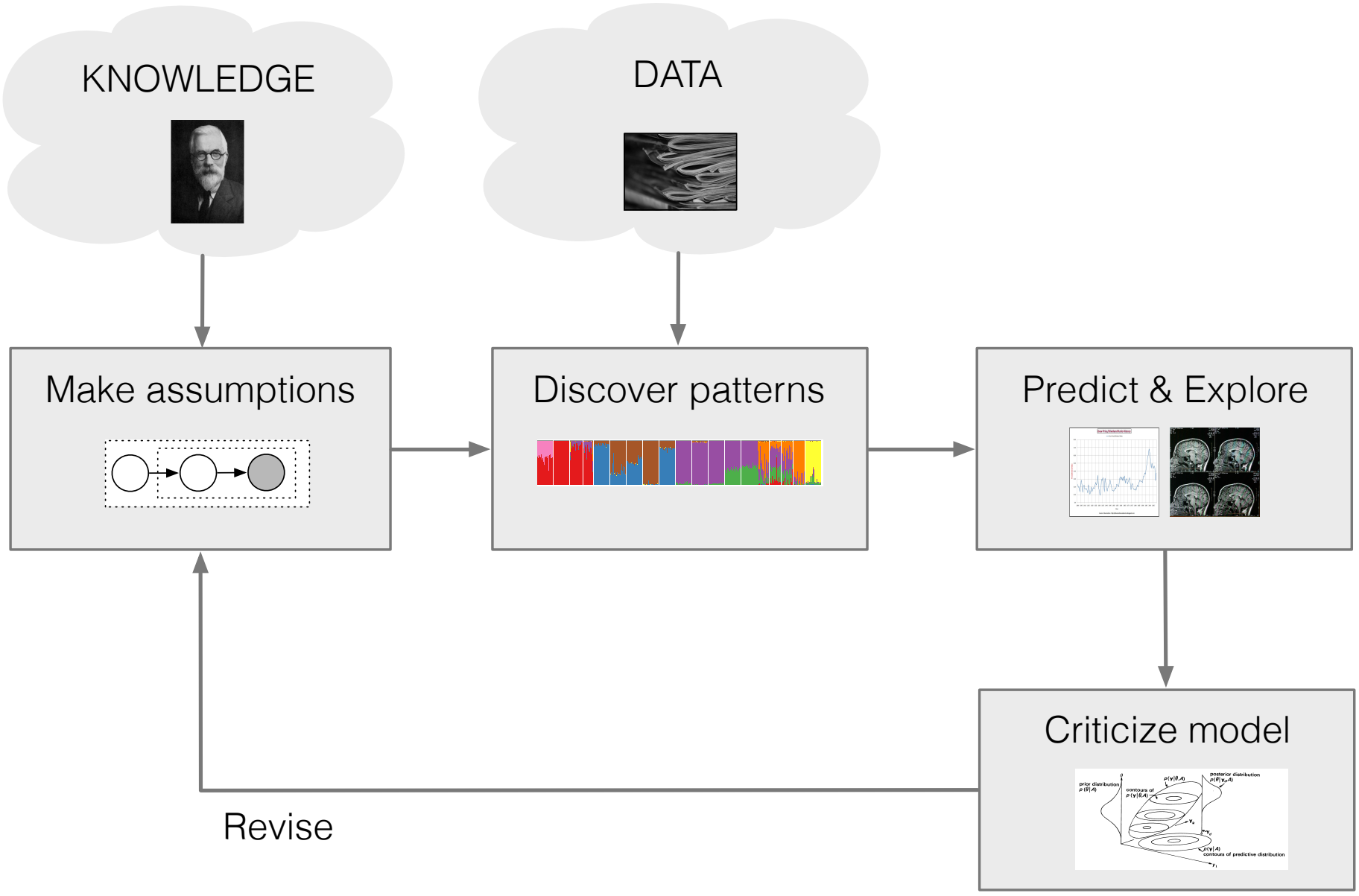
Our perspective:

- Customized data analysis is important to many fields.
- This pipeline separates assumptions, computation, application.
- It facilitates solving data science problems.



What we need in probabilistic ML:

- **Flexible** and **expressive** components for building models
- **Scalable** and **generic** inference algorithms
- **Easy to use** software to stretch probabilistic modeling into new areas


KNOWLEDGE
DATA
Make assumptions
Discover patterns
Predict & Explore
Criticize model
Revise

We should seek out unfamiliar summaries of observational material, and establish their useful properties... And still more novelty can come from finding, and evading, still deeper lying constraints.

(John Tukey, *The Future of Data Analysis*, 1962)