PROFESSOR BENJAMIN NOBLE (UCSD)

DISCOVERING HIDDEN TOPICS

POLITICS • THE D.C. BRIEF

Why Joe Biden Sounded So Conservative in His State of the Union

6 MINUTE READ

BY PHILIP ELLIOTT

MARCH 2, 2022 3:10 PM EST

Coming from a Democratic President, Joe Biden's first State of the Union sure had some moments that felt downright Republican. It may give Washington a hint about the months ahead.

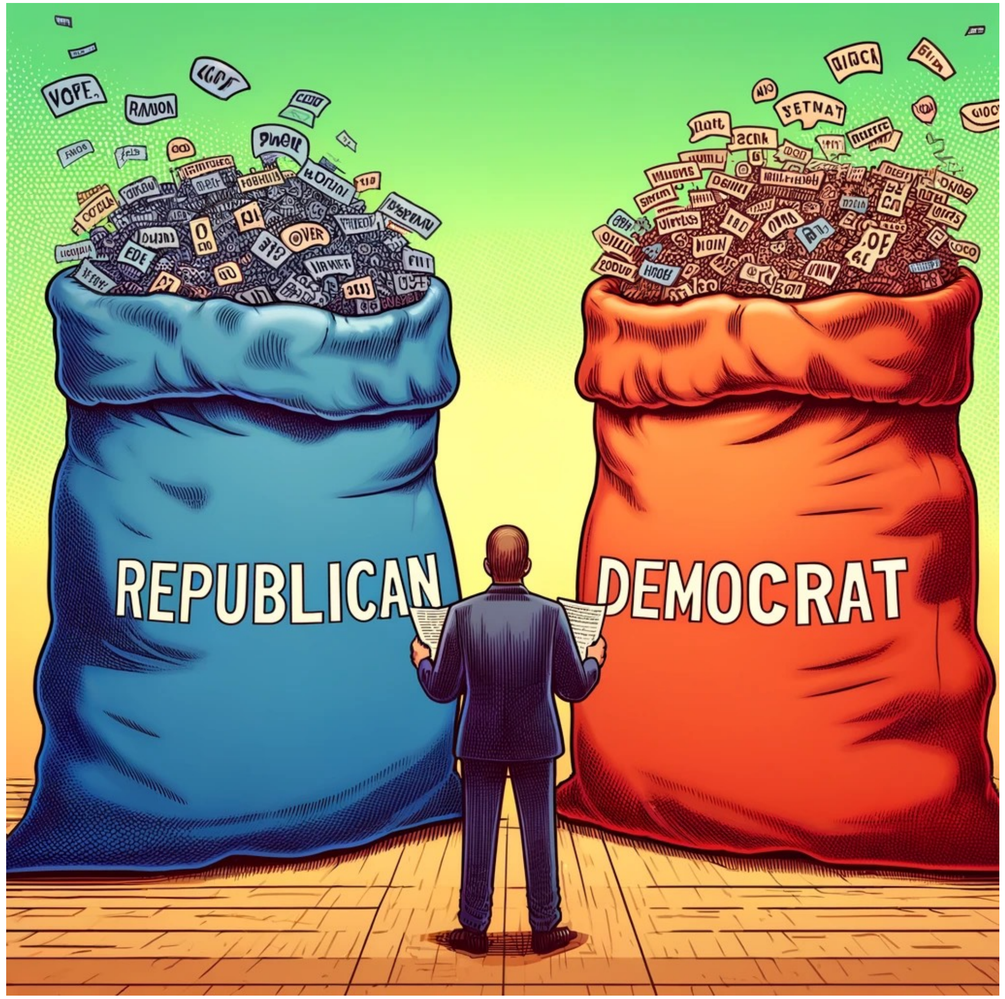
The riffs about funding—not defunding—the police. A made-in-America manufacturing agenda. Deficit reduction. A rousing, if slightly vengeful, call to beat back Moscow's march westward with Cold War belligerence. If someone claimed Peggy Noonan, Michael Gerson or David Frum had huddled with Biden in his private office to discuss themes, it wouldn't stretch the imagination.

via Time

	References “deficit”	Percent of party’s emails
Democrat	3,039	4.9%
Republican	6,179	6.1%

Data from the DC Inbox Project

A PROBABILISTIC MODEL OF LANGUAGE



WRITING WITH THE CATEGORICAL DISTRIBUTION

- Posit a three word vocabulary: cat, dog, fish; posit only one word per document.
- We can represent each word, and also each document in a “one-hot” encoding:
  - $\text{cat} = (1,0,0)$
  - $\text{dog} = (0,1,0)$
  - $\text{fish} = (0,0,1)$
- Each document is a draw from a **categorical distribution**, $W_i \sim \text{Categorical}(\mu)$.
  - μ is the probability of drawing each type, e.g., $\mu = (0.5,.0.25,0.25)$

THE CATEGORICAL DISTRIBUTION

- We can write the probability mass function as:

$$p(W_i \mid \mu) = \prod_{j=1}^{J} \mu_j^{w_{ij}}.$$

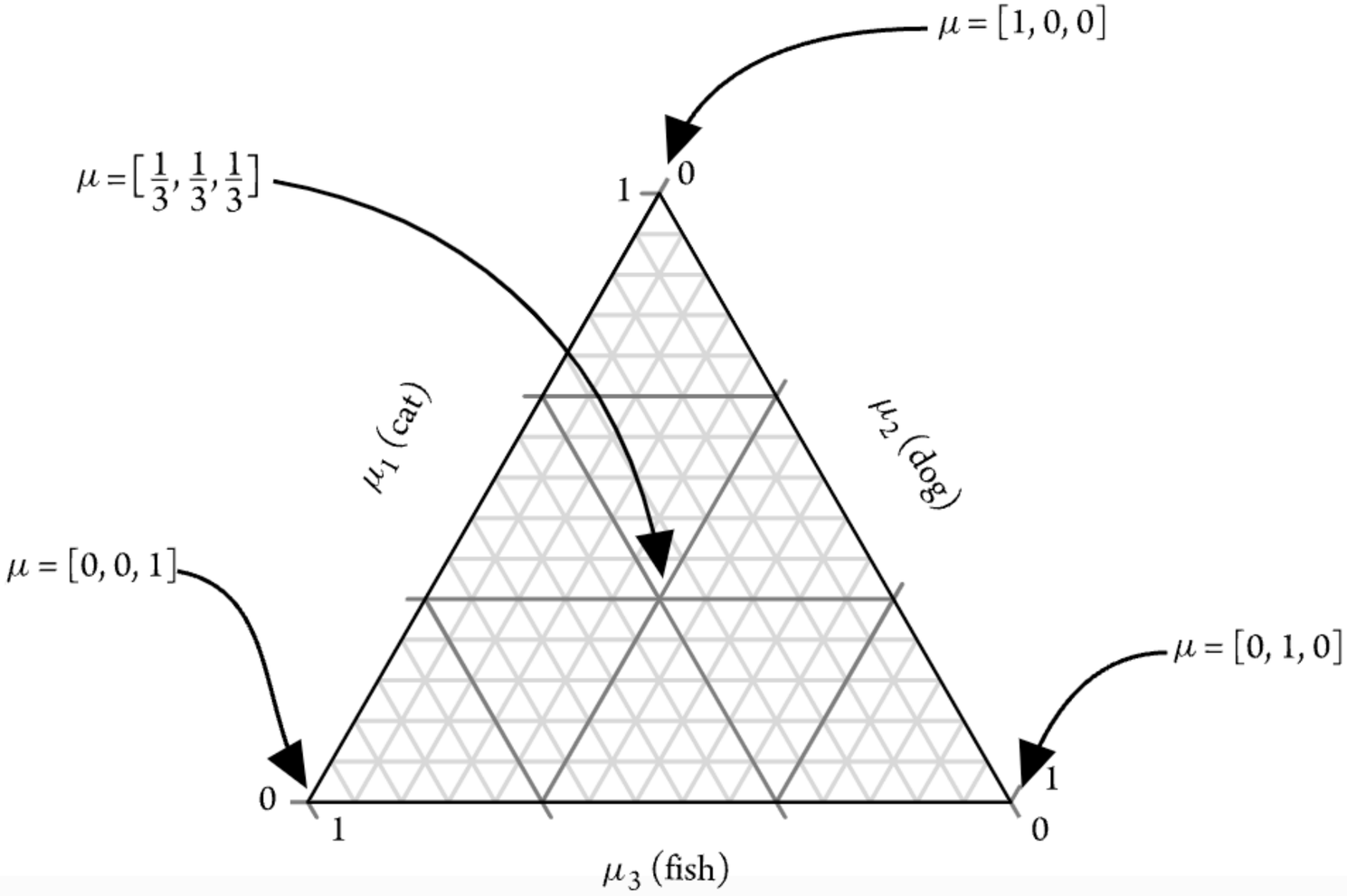
- An illustrative example: probability of observing the document “cat” is:
  - $p(\text{cat} \mid \mu) = \mu_1^1 \times \mu_2^0 \times \mu_3^0 = 0.5^1 \times 0.25^0 \times 0.25^0 = 0.5$

THE MULTINOMIAL DISTRIBUTION

- Now each document is a draw from a **multinomial distribution** where $W_i \sim \text{Multinomial}(M_i, \mu)$.

$$p(W_i \mid \mu) = \frac{M!}{\Pi_{j=1}^{J} W_{ij}!} \prod_{j=1}^{J} \mu_j^{W_{ij}}$$

- What is the probability of the document: fish, cat, fish?
  - Vector encoding: (1, 0, 2)
  - $p(\text{fish,cat,fish} \mid \mu) = \frac{3!}{(1!)(0!)(2!)}(0.5)^1(0.25)^0(0.25)^2 = 0.09375$


$\mu = [1, 0, 0]$
$\mu = [\frac{1}{3}, \frac{1}{3}, \frac{1}{3}]$
$\mu = [0, 0, 1]$
$\mu = [0, 1, 0]$
1
0
μ_1 (cat)
μ_2 (dog)
μ_3 (fish)

REVERSING THE PROCESS

- Suppose we discover a corpus that contains 1,000 total words with a unique vocabulary of 18 pets: *cat, dog, fish, bird, hamster, rabbit, horse, snake, lizard, turtle, frog, parrot, guinea pig, goldfish, betta fish, mouse, chinchilla, hedgehog.*
- We can estimate $\hat{\mu}_j = \frac{W_{ij}}{M_i}$.
  - E.g., *dog* appears 63 times, so $\mu_{\text{dog}} = 63/1000 = 0.063$.
  - Repeat for all pets...
    - $\mu = (0.067, 0.063, 0.049, 0.050, 0.067, 0.053, 0.063, 0.057, 0.056, 0.038, 0.047, 0.057, 0.051, 0.051, 0.059, 0.052, 0.062, 0.058)$.

REVERSING THE PROCESS

- We can generate 15 word documents by drawing words from this probability distribution.
  - "betta_fish cat bird horse guinea_pig goldfish chinchilla dog frog turtle snake lizard fish rabbit hedgehog"
  - "rabbit hedgehog goldfish parrot horse lizard cat snake chinchilla mouse betta_fish hamster dog fish bird"
  - "betta_fish guinea_pig fish snake dog hamster goldfish lizard chinchilla hedgehog frog parrot horse bird cat"

APPLYING A PROBABILISTIC LANGUAGE MODEL

- **Language model:** a model that assigns a probability to observing a particular sequence of tokens given a set of parameters.
  - Application: authorship of the Federalist Papers, Mostellar and Wallace (1963).

via The National Constitution Center

	"by"	"man"	"upon"	Word count
Hamilton	859	102	374	1335
Jay	82	0	1	83
Madison	474	17	7	498
???	15	2	0	17

$$W_H \sim \text{Multinomial}(1335, \mu_H) \text{ where } \mu_h = (\frac{859}{1335}, \frac{102}{1335}, \frac{374}{1335}) = (0.64, 0.08, 0.28)$$

$$W_J \sim \text{Multinomial}(83, \mu_J) = (0.99, 0.0, 0.01)$$

$$W_M \sim \text{Multinomial}(498, \mu_M) = (0.95, 0.035, 0.015)$$

APPLYING THE MULTINOMIAL DISTRIBUTION

- $p(W_? \mid \hat{\mu}_H) = \frac{17!}{(15!)(2!)(0!)}(0.64)^{15}(0.08)^{2}(0.28)^{0} = 0.001$
- $p(W_? \mid \hat{\mu}_J) = 0$
- $p(W_? \mid \hat{\mu}_M) = 0.077$
- Jay never used the word “man” in our data, which creates a mathematical quirk.
  - Laplace smoothing: adding noise to the data to smooth out our distribution (e.g., increase all counts by one).
  - We can incorporate this “pseudo-data” into our model using the Dirichlet distribution…

THE DIRICHLET DISTRIBUTION (WE'RE ALMOST DONE...I PROMISE!)

- We can write the full data generating process as:
  - Sample word probabilities for each author: $\mu_k \sim Dirichlet(\alpha)$ for $k \in 1,2,3$
    - Where α is the same length as size of μ (e.g., here three for the three words in our vocabulary) and each $\alpha_j > 0$. This corresponds to the amount of noise added as in Laplace smoothing.
  - Stack the resulting vectors: $\mu = [\mu_1, \mu_2, \mu_3]$
  - Sample text using author probabilities: $p(W_i \mid \mu, \pi_i) \sim \text{Multinomial}(M_i, \mu\pi_i)$

THE DIRICHLET DISTRIBUTION (WE'RE ALMOST DONE...I PROMISE!)

- Suppose we define $\mu_k \sim \text{Dirichlet}(\alpha(1,1,1))$ and $W_i \,|\, \mu, \pi_i \sim \text{Multinomial}(M_i, \mu\pi_i)$.
- Bayesian magic happens...
- Then, $\mu_k \,|\, \alpha, W_i \sim \text{Dirichlet}(W_i + \alpha)$ and $\mathbb{E}[\mu_k \,|\, \alpha, W_i] = \dfrac{W_i + \alpha}{\Sigma_{j=1}^{J}(W_{ij} + \alpha_j)}$.
- For Jay, the probability of "man" is: $\hat{\mu}_{J,man} = \dfrac{0+1}{82+0+1+(1+1+1)} = 0.01$.

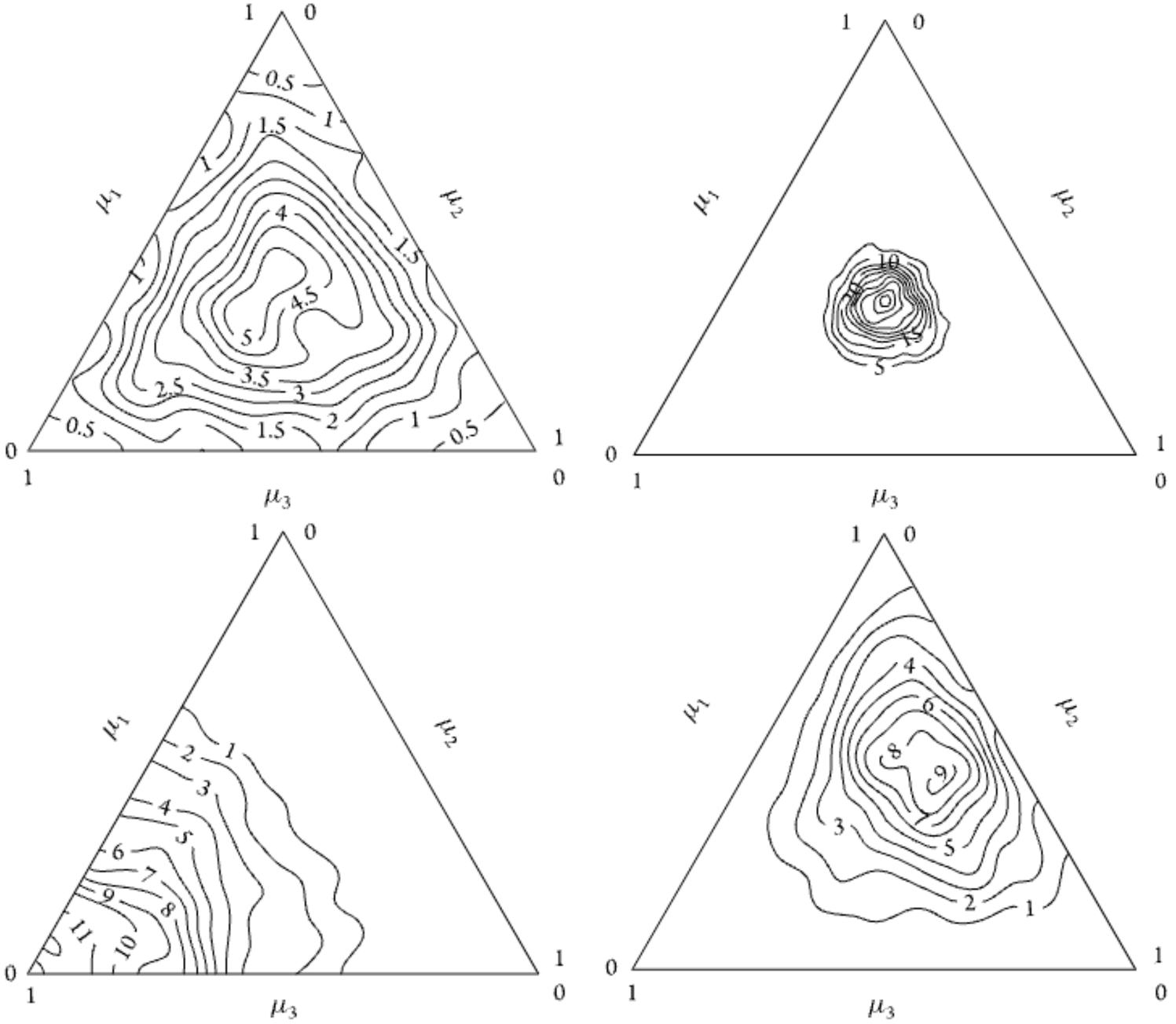


Figure 6.5. Contours of draws from a Dirichlet distribution with four different values of α. Top left: α=(2,2,2), Top right: α=(20,20,20), Bottom left: α=(1.2,1.2,4), Bottom right: α=(4,3,2). Note that as α0 gets large the values become more concentrated while having a particular value of αj be larger than the others pushes the draws into that corner of the simplex.

WHAT WE'VE BEEN BUILDING TOWARD: LATENT DIRICHLET ALLOCATION (LDA)

via Brookings

via UN Photo/Loey Felipe

HOW YOU (PROBABLY) WRITE AN ARTICLE

via DALL-E

HOW LDA *THINKS* YOU WRITE AN ARTICLE

- First, choose a number of words in your article (e.g., 50), using the Poisson distribution.
- Choose a mixture of topics for your document according to some fixed number of possible topics, K, and a Dirichlet distribution over topics (e.g., two topics; 1/3 food, 2/3 cute animals).
- Choose each word by…
  - Selecting a topic according the multinomial distribution using the mixture above.
  - Conditional on the topic, select a word according to a multinomial distribution specific to that topic (e.g., the distribution has 30% banana, 20% broccoli, etc).
  - Repeat 50 times.

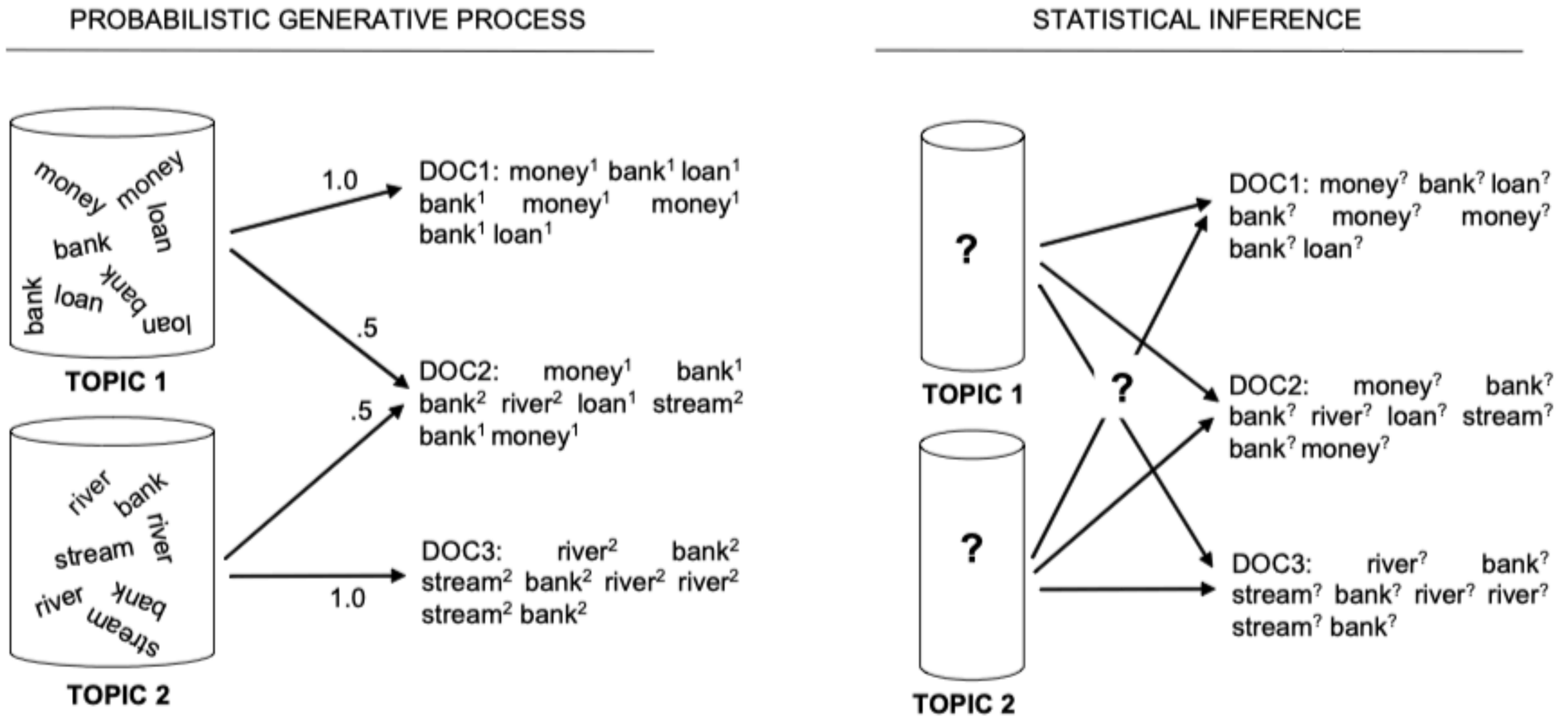


Figure 2. Illustration of the generative process and the problem of statistical inference underlying topic models

Steyvers and Griffiths (2007)

BUT HOW...?

- Short answer: magic.
- Long answer: we will do this “by hand” momentarily.
- Medium answer:
  - Start by randomly assigning each word to a topic.
  - Remove word_1 in document_1. Then, consider how strongly the document is associated with topic k and how and strongly w is associated with topic k.
  - Reassign word_1 to document_1 based on this new probability.
  - Repeat...

THE OUTPUT

- Remember…we are assuming a model of writing where for each document we…
  - Select a distribution of topics (call this θ).
  - From that distribution, choose a topic and then a word from that topic (call this ϕ).
- LDA produces an estimate of θ (document-topic proportions) and ϕ (topic-word proportions).

AN EXAMPLE...

- We want to uncover topics in the State of the Union Addresses.
  - We suppose $k = 20$.
  - We use standard pre-processing steps we used last lecture.
  - We have a total of 48,538 documents, 6,690 features and a dfm that is 99.75% sparse.

```
Topic 3 Top Words:
  Highest Prob: world, peac, war, forc, secur, defens, nation
  FREX: soviet, alli, weapon, aggress, threat, communist, terrorist
  Lift: adversari, airlift, align, angola, anthrax, arabia, buildup
  Score: world, soviet, nuclear, peac, weapon, terrorist, alli

 Topic 7 Top Words:
   Highest Prob: year, tax, economi, million, increas, budget, cost
   FREX: tax, budget, billion, cut, percent, spend, unemploy
   Lift: 1-year, after-tax, bracket, breadwinn, businessmen, bust, demobil
   Score: billion, budget, tax, percent, spend, million, deficit

  Topic 4 Top Words:
    Highest Prob: america, american, one, year, know, children, time
    FREX: thank, god, dream, tell, teacher, young, children
    Lift: son, spoke, walk, 100th, 200th, abraham, alic
    Score: tonight, america, children, school, today, thank, dream
```

hadiya's parents, nate and cleo, are in this chamber tonight, along with more than two dozen americans whose lives have been torn apart by gun violence. they deserve a vote. they deserve a vote. gabby giffords deserves a vote. the families of newtown deserve a vote. the families of aurora deserve a vote. the families of oak creek and tucson and blacksburg, and the countless other communities ripped open by gun violence -- they deserve a simple vote. they deserve a simple vote.

---

and so, tonight i'm going to ask something of every one of you. now, let me start with my generation, with the grandparents out there. you are our living link to the past. tell your grandchildren the story of struggles waged at home and abroad, of sacrifices freely made for freedom's sake. and tell them your own story as well, because every american has a story to tell.

---

on easter sunday of 2008, chris was out on patrol in baghdad when his bradley fighting vehicle was hit by a roadside bomb. that night, he made the ultimate sacrifice for our country. sergeant hake now rests in eternal glory in arlington, and his wife kelli is in the gallery tonight, joined by their son, who is now a 13-year-old and doing very, very well. to kelli and gage: chris will live in our hearts forever. he is looking down on you now. thank you. thank you very much. thank you both very much.

---

it lives on in the 8-year old boy in louisiana, who just sent me his allowance and asked if i would give it to the people of haiti. and it lives on in all the americans who've dropped everything to go some place they've never been and pull people they've never known from rubble, prompting chants of "u.s.a.! u.s.a.! u.s.a!" when another life was saved.

```
          V1         V2          V3          V4         V5         V6          V7
1 0.008769929 0.06922240 0.013821975 0.033857636 0.05999856 0.01930193 0.022367730
2 0.038170589 0.02324019 0.014794990 0.006534922 0.16409903 0.01612312 0.001096551
3 0.014863619 0.03331877 0.009035167 0.008894801 0.28300886 0.02672657 0.002479766
          V8          V9         V10         V11         V12          V13         V14
1 0.02704743 0.009455040 0.023614072 0.02702485 0.050424985 0.036744453 0.01731663
2 0.07904832 0.023389755 0.011475310 0.01238968 0.006965039 0.001136387 0.12448173
3 0.02099831 0.006006602 0.007178702 0.01420767 0.010628119 0.005273977 0.19234638
         V15        V16         V17         V18         V19         V20
1 0.01828501 0.04938809 0.42772417 0.01337487 0.04622609 0.02603416
2 0.12749588 0.03098440 0.01636553 0.20375442 0.01455091 0.08390325
3 0.02625240 0.04284532 0.08530092 0.07991099 0.04825554 0.08246751
```

SOME CONSIDERATIONS

- Validation.
- The researcher has to choose K (the number of topics).
- Short documents versus long documents.
- One model is better than many models.
- Adding covariates.
- Extensions galore!

via DALL-E