TOPIC MODELING
WITH STRUCTURED PRIORS
FOR TEXT-DRIVEN SCIENCE

MICHAEL J. PAUL
JOHNS HOPKINS UNIVERSITY

University of Colorado, Boulder | February 27, 2015
What color is the dress? black and blue, white and gold

BuzzFeed @BuzzFeed · 6h
A look at the United States right now
TEXT AS DATA

Google

Twitter

Facebook

Foursquare

Bing

Yelp

World

People
TEXT AS DATA
TEXT AS DATA

natural language processing (NLP)
TEXT AS DATA

Text-driven science
TEXT AS DATA

- Computational social science
- Computational journalism
- Crisis informatics
- Public health informatics
- Computational epidemiology

text-driven science
TEXT AS DATA

- Computational social science
- Computational journalism
- Crisis informatics
- Public health informatics
- Computational epidemiology

text-driven science
Let’s look at some examples!
Important public health task: disease surveillance

Many health agencies do flu monitoring
- But reports have a delay of ~2 weeks
Tracking the spread of influenza through tweets:


Air pollution in China

Public health tasks:
• Measure pollution levels
• Identify health effects
• Understand public response
Monitoring air pollution through social media:

Relationship between pollution levels and weibos

New trends in drug use

- Record numbers of new drugs recently
- Health officials can be years behind

Reports: Miami 'zombie' attacker may have been using 'bath salts'

A naked man who chewed off the face of another man in what is being called a zombie-like attack may have been under the influence of "bath salts," a drug referred to as the new LSD, according to reports from CNN affiliates in Miami.
Analyzing online forums:


Understanding healthcare quality from online reviews:

Text from reviews is significantly predictive of external measures of healthcare quality.

Paul, Wallace, Dredze (2013) Analyzing online doctor ratings with a joint topic-sentiment model. AAAI Workshop on Expanding the Boundaries of Health Informatics Using AI.

Many other applications:

- Air pollution in Chinese social media
  

- Health decision-making in search logs
  

- Public opinion in Twitter on public health issues:
  - Gun control
  - Vaccination
  - Smoking
  
  Benton, Paul, Hancock, Dredze (under review) A joint model of topic and perspective in social media.
TEXT-DRIVEN PUBLIC HEALTH
TEXT AS DATA
Structure of language:

- Morphology/strings
- Syntax/grammar
- Discourse/speech acts
- Topics/concepts


Structure of language:

- Morphology/strings

- Syntax/grammar

- Discourse/speech acts

- Topics/concepts


Paul, Dredze (2011) You are what you tweet: Analyzing Twitter for public health. 5th International Conference on Weblogs and Social Media (ICWSM).

A topic model is a statistical model of text

- We pretend that our data (text) are the output of a probabilistic process that generates data
sick
sore
throat
feel
fever
flu
...

allergies
nose
eyes
allergy
allergic
sneezing
...

watch
watching
tv
killing
movie
seen
...

class
school
read
test
doing
finish
...
sick
sore
throat
feel
fever
flu
...

allergies
nose
eyes
allergy
allergic
sneezing
...

watch
watching
tv
killing
movie
seen
...

class
school
read
test
doing
finish
...

TOPIC MODELING
I’ve had the flu and fever all week :( staying home from school and watching a lot of tv
I've had the flu and fever all week 😞 staying home from school and watching a lot of tv...
I've had the flu and fever all week:( staying home from school and watching a lot of tv
I've had the flu and fever all week :-( staying home from school and watching a lot of tv.
I've had the flu and fever all week:( staying home from school and watching a lot of tv.
I've had the flu and fever all week:( staying home from school and watching a lot of tv.
I've had the flu and fever all week :( staying home from school and watching a lot of tv.
I've had the flu and fever all week :(. Staying home from school and watching a lot of tv.
I've had the flu and fever all week :(

staying home from school and watching a lot of tv

sick sore throat feel fever flu

allergies nose eyes allergy allergic sneezing

watch watching tv killing movie seen

class school read test doing finish

Michael Paul @mjp39 · Jan 24

fever

watching
I've had the flu and fever all week :( staying home from school and watching a lot of tv.
I've had the flu and fever all week :( staying home from school and watching a lot of tv
I’ve had the flu and fever all week :( staying home from school and watching a lot of tv
I’ve had the flu and fever all week :( staying home from school and watching a lot of tv
I’ve had the flu and fever all week :( staying home from school and watching a lot of tv
I’ve had the flu and fever all week :( staying home from school and watching a lot of tv
I've had the flu and fever all week :( staying home from school and watching a lot of tv
Our imaginary process also needs to generate all these distributions
Our imaginary process also needs to generate all these distributions

- We need a distribution over distributions
  - Called a **prior** distribution
Dirichlet(\(\rho \times \times \))
TOPIC MODELING

PRIORS

Dirichlet( )
TOPIC MODELING

sick
sore
throat
feel
fever
flu
...

allergies
nose
eyes
allergy
allergic
sneezing
...

watch
watching
tv
killing
movie
seen
...

class
school
read
test
doing
finish
...

Dirichlet( )
sick  sore  throat  feel  fever  flu

allergies  nose  eyes  allergy  allergic  sneezing

watch  watching  tv  killing  movie  seen

class  school  read  test  doing  finish

Dirichlet()
I’ve had the flu and fever all week :( staying home from school and watching a lot of tv
Latent Dirichlet Allocation (LDA)
Blei, Ng, Jordan 2003

The topic and word distributions have Dirichlet priors
Latent Dirichlet Allocation (LDA)
Blei, Ng, Jordan 2003

The topic and word distributions have Dirichlet priors

Standard topic models are often insufficient for particular applications
• We need richer structure
TOPIC MODELING

APPLICATIONS

sick
sore
throat
feel
fever
flu
...

allergies
nose
eyes
allergy
allergic
sneezing
...

watch
watching
tv
killing
movie
seen
...

class
school
read
test
doing
finish
...

Paul, Dredze (2011) You are what you tweet: Analyzing Twitter for public health. 5th International Conference on Weblogs and Social Media (ICWSM).

TOPIC MODELING

About health issues

sick
sore
throat
feel
fever
flu
...  

ALLERGIES

allergies
nose
eyes
allergy
allergic
sneezing
...  

Irrelevant to health

watch
watching
tv
killing
movie
seen
...  

class
school
read
test
doing
finish
...
In general:
Topics can be organized in ways that are more interpretable to users.
Understanding healthcare quality from online reviews:
Topics in online doctor reviews:

- best
- years
- caring
- care
- patients
- patient
- recommend
- family
- time
- staff
- great
- helpful
- feel
- questions
- office
- friendly
- office
- time
- appointment
- rude
- staff
- room
- didn’t
- wait
Topics in online doctor reviews:

Both have **positive sentiment**

- best
  - years
  - caring
  - care
  - patients
  - patient
  - recommend
  - family

- time
  - staff
  - great
  - helpful
  - feel
  - questions
  - office
  - friendly

- office
  - time
  - appointment
  - rude
  - staff
  - room
  - didn’t
  - wait

Both about **staff/office issues**
Topics in online doctor reviews:

<table>
<thead>
<tr>
<th></th>
<th>Staff/Office</th>
<th>Personality</th>
<th>Surgery</th>
</tr>
</thead>
<tbody>
<tr>
<td>Positive</td>
<td>time</td>
<td>best</td>
<td>surgery</td>
</tr>
<tr>
<td></td>
<td>staff</td>
<td>years</td>
<td>first</td>
</tr>
<tr>
<td></td>
<td>great</td>
<td>caring</td>
<td>son</td>
</tr>
<tr>
<td></td>
<td>helpful</td>
<td>care</td>
<td>life</td>
</tr>
<tr>
<td></td>
<td>feel</td>
<td>patients</td>
<td>surgeon</td>
</tr>
<tr>
<td></td>
<td>questions</td>
<td>patient</td>
<td>daughter</td>
</tr>
<tr>
<td></td>
<td>office</td>
<td>recommend</td>
<td>recommend</td>
</tr>
<tr>
<td></td>
<td>friendly</td>
<td>family</td>
<td>thank</td>
</tr>
<tr>
<td>Negative</td>
<td>office</td>
<td>care</td>
<td>pain</td>
</tr>
<tr>
<td></td>
<td>time</td>
<td>medical</td>
<td>told</td>
</tr>
<tr>
<td></td>
<td>appointment</td>
<td>patients</td>
<td>went</td>
</tr>
<tr>
<td></td>
<td>rude</td>
<td>doesn’t</td>
<td>said</td>
</tr>
<tr>
<td></td>
<td>staff</td>
<td>help</td>
<td>surgery</td>
</tr>
<tr>
<td></td>
<td>room</td>
<td>know</td>
<td>later</td>
</tr>
<tr>
<td></td>
<td>didn’t wait</td>
<td>don’t</td>
<td>didn’t</td>
</tr>
<tr>
<td></td>
<td></td>
<td>problem</td>
<td>months</td>
</tr>
</tbody>
</table>
A multi-dimensional topic model
Word distributions are grouped into different concepts
  • e.g. sentiment and aspect

Analyzing online drug forums:

Reports: Miami 'zombie' attacker may have been using 'bath salts'

A naked man who chewed off the face of another man in what is being called a zombie-like attack may have been under the influence of "bath salts," a drug referred to as the new LSD, according to reports from CNN affiliates in Miami.
3-dimensional model:

- Drug type
- Route of administration (i.e. method of intake)
- Aspect

<table>
<thead>
<tr>
<th>Drug (22 total)</th>
<th>Route</th>
<th>Aspect</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alcohol</td>
<td>Injection</td>
<td>Chemistry</td>
</tr>
<tr>
<td>Amphetamine</td>
<td>Oral</td>
<td>Culture</td>
</tr>
<tr>
<td>Cannabis</td>
<td>Smoking</td>
<td>Effects</td>
</tr>
<tr>
<td>Cocaine</td>
<td>Snorting</td>
<td>Health</td>
</tr>
<tr>
<td>...</td>
<td></td>
<td>Usage</td>
</tr>
<tr>
<td>Salvia</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tobacco</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Joint model with 3 factors:

- Drug type
- Route of administration (i.e. method of intake)
- Aspect

Each “topic” is a triple such as:

(Cocaine, Snorting, Health)  (Cocaine, Snorting, Usage)

- coke
- line
- lines
- nose
- small
- cut

- nose
- pain
- damage
- blood
- cocaine
- problem
Suppose we want to model: \((\text{Marijuana, Oral, Chemistry})\)
<table>
<thead>
<tr>
<th>Marijuana</th>
<th>Oral</th>
<th>Chemistry</th>
</tr>
</thead>
<tbody>
<tr>
<td>weed</td>
<td>capsules</td>
<td>solvent</td>
</tr>
<tr>
<td>cannabis</td>
<td>consumes</td>
<td>extraction</td>
</tr>
<tr>
<td>thc</td>
<td>toast</td>
<td>evaporate</td>
</tr>
<tr>
<td>marijuana</td>
<td>stomach</td>
<td>evaporated</td>
</tr>
<tr>
<td>stoned</td>
<td>chewing</td>
<td>solvents</td>
</tr>
<tr>
<td>bowl</td>
<td>ambien</td>
<td>evaporated</td>
</tr>
<tr>
<td>bud</td>
<td>digestion</td>
<td>solvents</td>
</tr>
<tr>
<td>joint</td>
<td>juice</td>
<td>evaporation</td>
</tr>
<tr>
<td>blunt</td>
<td>absorbed</td>
<td>yield</td>
</tr>
<tr>
<td>herb</td>
<td>ingestion</td>
<td>chloride</td>
</tr>
<tr>
<td>bong</td>
<td>meal</td>
<td>alkaloids</td>
</tr>
<tr>
<td>pot</td>
<td>tiredness</td>
<td>tek</td>
</tr>
<tr>
<td>sativa</td>
<td>chew</td>
<td>compounds</td>
</tr>
<tr>
<td>blaze</td>
<td>juices</td>
<td>evaporating</td>
</tr>
<tr>
<td>indica</td>
<td>gelatin</td>
<td>atom</td>
</tr>
<tr>
<td>smoking</td>
<td>fruit</td>
<td>aromatic</td>
</tr>
<tr>
<td>blunts</td>
<td>yogurt</td>
<td>non-polar</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>purified</td>
</tr>
<tr>
<td></td>
<td></td>
<td>jar</td>
</tr>
<tr>
<td></td>
<td></td>
<td>....</td>
</tr>
</tbody>
</table>
Marijuana

weeds
cannabis
thc
marijuana
stoned
bowl
bud
joint
blunt
herb
bong
pot
sativa
blaze
indica
smoking
blunts
...

Oral

capsules
consumes
toast
stomach
chewing
ambien
digestion
juice
absorbed
ingestion
meal
tiredness
chew
juices
gelatin
yogurt
fruit
...

Chemistry

solvent
extraction
evaporate
evaporated
solvents
evaporation
yield
chloride
alkaloids
tek
compounds
evaporating
atom
aromatic
non-polar
purified
jar
...

\( \exp( \text{...} ) \)
Dirichlet distribution

- thc
- method
- extraction
- plant
- material
- cannabis
- simple
- coffee
- oil
- contains
- jar
- dried
- process
- dry
- water
- extract
- results
- ...

Factors:

- LDA
- Drug Discussions
word distribution for the triple:

\[
\text{Marijuana (Oral Chemistry)}
\]

\[
\text{oil, water, butter, thc, weed, hash, cannabis, alcohol, make, milk, high, marijuana, add, ... mixture, hours, try, brownies}
\]

\[
\sim \text{Dirichletlet}(\text{thc, method, extraction, plant, material, cannabis, simple, coffee, oil, contains, jar, dried, process, dry, water, extract, results, ...})
\]
FACTORIAL LDA

DRUG DISCUSSIONS

~ Dirichlet( )

oil
water
butter
thc
weed
hash
cannabis
alcohol
make
milk
high
marijuana
add
...
mixture
hours
try
brownies

thc
method
extraction
plant
material
cannabis
simple
coffee
oil
contains
jar
dried
process
dry
water
extract
results
...

Stop
word distribution for the triple: (Marijuana Oral Chemistry)

- oil
- water
- butter
- thc
- weed
- hash
- cannabis
- alcohol
- make
- milk
- high
- marijuana
- add
- ...
- mixture
- hours
- try
- brownies

~ Dirichlet(

- thc
- method
- extraction
- plant
- material
- cannabis
- simple
- coffee
- oil
- contains
- jar
- dried
- process
- dry
- water
- extract
- results
- ...

FACTORIAL LDA

DRUG DISCUSSIONS
We can use this model to extract specific information about new drugs

- e.g. dosage, desired effects, negative effects

“What is the dosage when taking mephedrone orally?”
We can use this model to extract specific information about new drugs

- e.g. dosage, desired effects, negative effects

“What is the dosage when taking mephedrone orally?”

Mephedrone
Oral Usage
We can use this model to extract specific information about new drugs

- e.g. dosage, desired effects, negative effects

“What is the dosage when taking mephedrone orally?”

If it is [someone who isn’t you]’s first time using Mephedrone [someone who isn’t me] recommends a 100mg oral dose on an empty stomach.
We can use this model to extract specific information about new drugs

- e.g. dosage, desired effects, negative effects

“What is the dosage when taking mephedrone orally?”

If it is [someone who isn’t you]’s first time using Mephedrone [someone who isn’t me] recommends a 100mg oral dose on an empty stomach.

It is recommended by users that Mephedrone be taken on an empty stomach. Doses usually vary between 100mg – 1g.
Word distributions are factored into multiple dimensions

Each topic prior is informed by parameters for each dimension

<table>
<thead>
<tr>
<th></th>
<th>Staff/Office</th>
<th>Personality</th>
<th>Surgery</th>
</tr>
</thead>
<tbody>
<tr>
<td>Positive</td>
<td>time</td>
<td>best</td>
<td>surgery</td>
</tr>
<tr>
<td></td>
<td>staff</td>
<td>years</td>
<td>first</td>
</tr>
<tr>
<td></td>
<td>great</td>
<td>caring</td>
<td>son</td>
</tr>
<tr>
<td></td>
<td>helpful</td>
<td>care</td>
<td>life</td>
</tr>
<tr>
<td></td>
<td>feel</td>
<td>patients</td>
<td>surgeon</td>
</tr>
<tr>
<td></td>
<td>questions</td>
<td>patient</td>
<td>daughter</td>
</tr>
<tr>
<td></td>
<td>office</td>
<td>recommend</td>
<td>recommend</td>
</tr>
<tr>
<td></td>
<td>friendly</td>
<td>family</td>
<td>thank</td>
</tr>
<tr>
<td>Negative</td>
<td>office</td>
<td>care</td>
<td>pain</td>
</tr>
<tr>
<td></td>
<td>time</td>
<td>medical</td>
<td>told</td>
</tr>
<tr>
<td></td>
<td>appointment</td>
<td>patients</td>
<td>went</td>
</tr>
<tr>
<td></td>
<td>rude</td>
<td>doesn’t</td>
<td>said</td>
</tr>
<tr>
<td></td>
<td>staff</td>
<td>help</td>
<td>surgery</td>
</tr>
<tr>
<td></td>
<td>room</td>
<td>know</td>
<td>later</td>
</tr>
<tr>
<td></td>
<td>didn’t</td>
<td>don’t</td>
<td>didn’t</td>
</tr>
<tr>
<td></td>
<td>wait</td>
<td>problem</td>
<td>months</td>
</tr>
</tbody>
</table>
Topics:

- oil
- water
- butter
- THC
- weed
- cannabis
- hash
- cannabis
- alcohol
- make
- milk

Components:

- weed
- cannabis
- THC
- marijuana
- stoned
- bowl
- bud
- joint
- blunt
- herb
- capsules
- consumes
- toast
- stomach
- chewing
- ambien
- digestion
- juice
- absorbed
- ingestion
- solvent
- extraction
- evaporate
- evaporated
- solvents
- evaporation
- yield
- chloride
- alkaloids
- tek
TOPIC STRUCTURES  FACTORORIZATION

Diagram illustrating the relationship between topic structures and factorization.
Generalization:
Components don’t need to be factored into groups. They can feed into topics in many different ways!
TOPIC STRUCTURES

- Tree

- Directed acyclic graph (DAG)

- Weighted DAG
“Surgery”
surgery
pain
got
surgeon
told
procedure
months
performed
removed
left
fix
said
later
years

“Family”
dr
best
children
years
kids
cares
hes
care
old
daughter
child
husband
family
pediatrician
trust

Dr. best children years kids cares hes care old daughter child husband family pediatrician trust

dr
life
thank
saved
god
husband
heart
cancer
years
helped
 doctors
hospital
father
man
able
**Structured-prior topic models**

A family of topic models in which the Dirichlet priors are functions of underlying components

**SPRITE**

SPRITE generalizes many existing topic models:

<table>
<thead>
<tr>
<th>Model</th>
<th>Document priors</th>
<th>Topic priors</th>
</tr>
</thead>
<tbody>
<tr>
<td>LDA</td>
<td>Single component</td>
<td>Single component</td>
</tr>
<tr>
<td>SCTM</td>
<td>Single component</td>
<td>Sparse binary $\beta$</td>
</tr>
<tr>
<td>SAGE</td>
<td>Single component</td>
<td>Sparse $\omega$</td>
</tr>
<tr>
<td>FLDA</td>
<td>Binary $\delta$ is transpose of $\beta$</td>
<td>Factored binary $\beta$</td>
</tr>
<tr>
<td>PAM</td>
<td>$\alpha$ are supertopic weights</td>
<td>Single component</td>
</tr>
<tr>
<td>DMR</td>
<td>$\alpha$ are feature values</td>
<td>Single component</td>
</tr>
</tbody>
</table>
The priors over word distributions are weighted combinations of components:

$$\tilde{\phi}_{iv} = \exp\left(\sum_{c=1}^{C(\phi)} \beta_{ic} \omega_{cv}\right)$$

$$\phi_i \sim \text{Dirichlet}(\tilde{\phi}_i)$$

distribution over words in $i$th topic
The priors over word distributions are weighted combinations of components:

\[
\exp(\beta_{11} \omega_1 + \beta_{12} \omega_2 + \beta_{13} \omega_3 + \beta_{14} \omega_4 + \beta_{15} \omega_5)
\]
The priors over word distributions are weighted combinations of components:

\[ \exp(\beta_{21} \times \omega_1 + \beta_{22} \times \omega_2 + \beta_{23} \times \omega_3 + \beta_{24} \times \omega_4 + \beta_{25} \times \omega_5) \]
The priors over word distributions are weighted combinations of components:

We can induce different structures by constraining the values of $\beta$. 
The priors over word distributions are weighted combinations of components:
The priors over word distributions are weighted combinations of components:

\[ \exp(1 \times \omega_1 + 0 \times \omega_2 + 0 \times \omega_3 + 0 \times \omega_4 + 0 \times \omega_5) \]
The priors over word distributions are weighted combinations of components:

$$\exp(0 \times \omega_1 + 0 \times \omega_2 + 1 \times \omega_3 + 0 \times \omega_4 + 0 \times \omega_5)$$
The priors over word distributions are weighted combinations of components:

**Tree:** Each topic’s $\beta$ vector is zero in all but one component
The priors over word distributions are weighted combinations of components:

**Factorization:** Like a tree, but a nonzero component in each factor
• Organizes topics in a variety of useful ways
  • Can be tailored toward different applications
• Generalizes many topic models
  • While opening up new possibilities
• Allows practitioners to make sense of big text data
  • Can drive new scientific research
MOVING FORWARD
MOVING FORWARD BETTER MODELS

Text data → NLP → External data

Modeling → Validation
MOVING FORWARD BETTER MODELS

Text data → NLP → External data

Modeling

NLP

External data

Modeling
Challenges with big models:

• Spurious correlations between text and datasets

Modeling flu prevalence in Twitter:
Challenges with big models:

- Spurious correlations between text and datasets

Modeling flu prevalence in Twitter:

Covariance prior:

- flu
- sick
- swine
- shot
- cancer
- fever
- h1n1
- #beatcancer
- better
- getting
- home
- halloween
- breast

- sick
- flu
- swine
- better
- shot
- getting
- cancer
- home
- hope
- fever
- feel
- feeling
- h1n1
Challenges with big models:

- Spurious correlations between different datasets
Challenges with big models:

- Spurious correlations between different datasets

- Need models with domain expertise
  - Opportunities for interactive machine learning
Challenges with big models:

- **Solutions**: language structure and human feedback
MOVING FORWARD
What happens when this article gets published?
What happens when this article gets published?

Most cancers are due to bad luck, not lifestyle, researchers say. 
time.com/3651785/cancer... ...Anyone fancy a smoke?

Most cancer just 'bad luck'? 
time.com/3651785/cancer... So I'm going for checkup w/ my magic eightball today. Just saved copay
**Variation in cancer risk among tissues can be explained by the number of stem cell divisions**

Cristian Tomasetti and Bert Vogelstein

Some tissues give rise to human cancers millions of times more often than other tissues. Although this has been recognized for more than a century, it has never been explained. Here, we show that the lifetime risk of cancers of many different sites is closely correlated (0.99) with the total number of divisions of the normal self-renewing cells maintaining that tissue's homeostasis. These results suggest that only a third of the variation in cancer risk among tissues is attributable to genetic factors or inherent predispositions. The majority is due to “bad luck,” that is, random mutations arising during DNA replication in normal, noncancerous stem cells. This is important not only for understanding the disease but also for designing strategies to limit the mortality it causes.

---

**Bad Luck of Random Mutations Plays Predominant Role in Cancer, Study Shows**

Statistical modeling links cancer risk with number of stem cell divisions

**Release Date:** January 1, 2015

**Addendum to news release added Jan. 7, 2015**

Johns Hopkins Medicine is gratified by the response and discussion generated by Cristian Tomasetti and Bert Vogelstein’s research paper, “Variation in cancer risk among tissues can be explained by the number of stem cell divisions,” published in *Science* on Jan. 2, 2015, and a news release describing the work, “Bad Luck of Random Mutations Plays Predominant Role in Cancer, Study Shows.” Cancer is driven by a number of factors and causes, and concepts related to calculating risk are complex and often subject to debate. To facilitate the ongoing discussion, and to address the many thoughtful questions their research stimulated, the two scientists have provided the following answers to frequently asked questions.

Is there an analogy that can help put the results of your research in perspective?

Getting cancer could be compared to getting into a car accident. Our results would be equivalent to showing a high correlation between length of trip and getting into an accident.

---

**Most Types of Cancer Just ‘Bad Luck,’ Researchers Say**

Helen Regan @helenregen 1/2 Jan 1, 2015

Two thirds of cancers could be explained as biological misfortune

Researchers have found that bad luck plays a major role in determining most types of cancer, rather than genetics or risky lifestyle choices such as smoking.

The results, published in the journal *Science* on Thursday, found that random DNA mutations that arise in the body when stem cells divide into various tissues cause two thirds of cancers.
Variation in cancer risk among tissues can be explained by the number of stem cell divisions

Cristian Tomasetti and Bert Vogelstein

Some tissue types give rise to human cancers millions of times more often than other tissue types. Although this has been recognized for more than a century, it has never been explained. Now, we show that the lifetime risk of cancers of many different types is closely correlated (Spearman’s ρ) with the total number of divisions of the normal self-renewing cells maintaining that tissue’s homeostasis. These results suggest that only a third of the variation in cancer risk among tissues can be explained by cell divisions in that tissue, and that most of the remaining variation may be explained by cell divisions in other tissues.

Bad Luck of Random Mutations Plays Predominant Role in Cancer, Study Shows

—Statistical modeling links cancer risk with number of stem cell divisions

Release Date: January 1, 2015
Addendum to news release added Jan. 7, 2015

Johns Hopkins Medicine is required by the responses and discussion generated by this story. A link to the abstract of the paper, “Variation in cancer risk by the number of stem cell divisions,” published in the journal. Bad Luck of Cell Division in Cancer, Study Shows. Cancer is driven by the number of cell divisions, and the number of cell divisions in a lifetime is complex. To facilitate the ongoing discussion, and to address some scientific questions, the two scientists have frequently asked questions.

Top put the results of your research in perspective? Am I getting into a car accident? Our results would be a rare event, specifically the length of time spent in a car accident and the number of other accidents.

After examining 31 cancer types, researchers found 22 were from mutations in stem cells that could not be prevented.
There’s no end to exciting questions we can ask of big, open data

We need methods to link what people are saying on the web with real-world trends

This requires advancements at the intersection of language processing and data science
Flu:
• David Broniatowski
• Andrea Dugas
• Nicholas Generous
• Alex Lamb
• Michael Smith

Air pollution:
• Shiliang Wang
• Angie Chen
• Brian Schwartz

Doctor reviews:
• Byron Wallace
• Urmimala Sarkar
• Thomas Trikalinos

Drug forums:
• Meg Chisolm
• Matthew Johnson
• Ryan Vandrey

Advisors: Mark Dredze, Jason Eisner
Funding: Microsoft Research, NSF, JHU Dean’s office
THANK YOU

QUESTIONS?
“What opinions are people tweeting about gun rights?”
Sprite can model this!

We created a structured prior that incorporates geographic data about gun ownership.

- Certain topics have a higher/lower prior depending on whether a tweet is from a high/low gun ownership state.

Benton, Paul, Hancock, Dredze. *A structured model of topic and perspective in social media.* Submitted to *ICWSM.*
Sprite can model this!

Associated with: Low gun ownership

- violence
culture
less
sense
problem
makes
world
country
commom
america’s

- children
  kids
  likely
times
  murdered
giving
  nothing
  stand
  sorry
  insane

Associated with: High gun ownership

- nra
  war
  even
  keep
  murder
  members
  liberal
  fear
  government
  call

- teachers
  school
  armed
  schools
  carry
  god
  kids
  teacher
  protect
  security
This topic is associated with high gun ownership:

- guns
- ’merica
- truck
- shoot
- deer
- hunting
- day
- beer
- season
- friends
This topic is associated with high gun ownership:

- guns
- ’merica
- truck
- shoot
- deer
- hunting
- day
- beer
- season
- friends

Probably also associated with:
- Population density
- Political affiliation
- …
MOVING FORWARD
Prior for triple \((i,j,k)\):

\[
\hat{\phi}_{(i,j,k)v} = \exp(\omega^{\text{drug}}_{iv} + \omega^{\text{route}}_{jv} + \omega^{\text{aspect}}_{kv})
\]

\[
\phi_{(i,j,k)} \sim \text{Dirichlet}(\hat{\phi}_{(i,j,k)})
\]

distribution over words for this triple

In general, prior for tuple \(t\):

\[
\hat{\phi}_{tv} = \exp(\sum_{k=1}^{K} \omega^{(k)}_{t_kv})
\]

\[
\phi_{t} \sim \text{Dirichlet}(\hat{\phi}_{t})
\]
$\beta_i$ has one value of 1 in each factor; 0 elsewhere

Recall (FLDA):

\[
\tilde{\phi}_{iv} = \exp\left(\sum_{c=1}^{C(\phi)} \beta_{ic} \omega_{cv}\right)
\]

\[
\tilde{\theta}_{mi} = \exp\left(\sum_{c=1}^{C(\theta)} \alpha_{mc} \delta_{ci}\right)
\]

\[
\delta_c \text{ is transpose of } \beta_i
\]

\[
\tilde{\phi}_{\tilde{i}v} = \exp\left(\sum_{k=1}^{K} \omega_{i_kv}\right)
\]

\[
\tilde{\theta}_{\tilde{m}i} = \exp\left(\sum_{k=1}^{K} \alpha_{m_ki}\right)
\]
\[ \phi_{iv} = \exp(\sum_{c=1}^{C(\phi)} \beta_{ic} \omega_{cv}) \]
\[ \theta_{mi} = \exp(\sum_{c=1}^{C(\theta)} \alpha_{mc} \delta_{ci}) \]

\[ \tilde{\phi}_{iv} = \exp(\sum_{c=1}^{C(\phi)} \beta_{ic} \omega_{cv}) \]
\[ \tilde{\theta}_{mi} = \exp(\sum_{c=1}^{C(\theta)} \alpha_{mc} \delta_{ci}) \]

\[ \tilde{\phi}_{iv} = \exp(\sum_{k=1}^{K} \omega_{ikv}) \]
\[ \tilde{\theta}_{mi} = \exp(\sum_{k=1}^{K} \alpha_{mi_k}) \]

\( \beta_i \) has one value of 1 in each factor; 0 elsewhere

\( \delta_c \) is transpose of \( \beta_i \)

Recall (FLDA):
Suppose we want to model how perspective influences topics

- e.g. certain topics are “pro” or “anti” gun control

A single-component SPRITE model:

$$\tilde{\phi}_{kv} = \exp(r_k \omega_v)$$

- The $v$th word’s perspective association
- The $k$th topic’s perspective association
Suppose we want to model how perspective influences topics

• e.g. certain topics are “pro” or “anti” gun control

A single-component SPRITE model:

\[
\tilde{\phi}_{kv} = \exp(r_k \omega_v)
\]

The \(v\)th word’s perspective association

\[
\tilde{\Theta}_{mk} = \exp(\alpha_m r_k)
\]

The \(k\)th topic’s perspective association

The \(m\)th document’s perspective association

A positive \(r_k\) means:

• Words with positive \(\omega_k\) are more likely in topic \(k\)
• Topic \(k\) is more likely in documents with positive \(\alpha_m\)
Suppose we want to model how perspective influences topics

- e.g. certain topics are “pro” or “anti” gun control

A single-component SPRITE model:

\[ \tilde{\theta}_{mk} = \exp(\alpha_m r_k) \]

Incorporating soft supervision: \( \alpha_m \sim \mathcal{N}(s_m, \sigma^2) \)

Supervision \( s_m \) is a function of:

- Survey data (% gun ownership in each state)
- Hashtags (#GunControlNow vs #NoGunControl)
Previous section: linking text to population data
Another idea: linking text to individual data
Previous section: linking text to population data

Another idea: linking text to individual data

Extraversion:

Introversion:

Schwartz et al. (2013) Personality, gender, and age in the language of social media: the open-vocabulary approach. *PLOS ONE.*
Previous section: linking text to population data

Another idea: linking text to individual data

- New territory: clinical records
How does the web influence medical decision-making?
