Machine Learning for Finance

Neal Parikh

Cornell University

Spring 2018

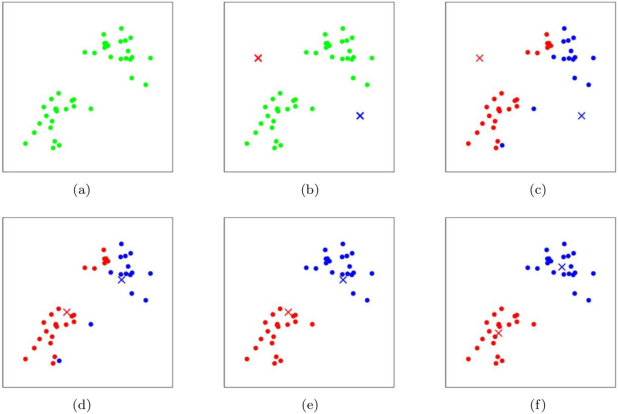
k-means

k-means

given $\mathcal{D} = \{x_1, \ldots, x_N\}$, $x_i \in \mathbf{R}^n$, group data into a few 'clusters'

1. randomly initialize cluster centroids $\mu_1, \ldots, \mu_k \in \mathbf{R}^n$
2. repeat until convergence
   1. find cluster assignment for x_i
   $$c_i := \operatorname*{argmin}_j \|x_i - \mu_j\|_2^2$$
   2. recompute cluster centroids using these assignments
   $$\mu_j := \frac{\sum_{i=1}^N [c_i = j] x_i}{\sum_{i=1}^N [c_i = j]}$$

k-means



Alternating minimization

- k-means can also be viewed as alternating minimization on the (biconvex) 'distortion function'

$$J(c, \mu) = \sum_{i=1}^{N} \|x_i - \mu_{c_i}\|_2^2$$

- results dependent on initialization, so do random restarts and pick one with lowest distortion
- can also derive k-means as a limit of a probabilistic model

Mixture models and the EM algorithm

Mixture of Gaussians

- probabilistic model for clustering / density estimation
- consider data $\mathcal{D} = \{x_1, \ldots, x_N\}$
- generative model

$$
\begin{aligned}
z &\sim \mathrm{Multinomial}(\phi) \\
x \,|\, z = k &\sim \mathrm{N}(\mu_k, \Sigma_k)
\end{aligned}
$$

- *i.e.*, each x_i generated by sampling a **unobserved** (hidden, latent) $z_i \in [K]$ and then drawing x_i from the corresponding Gaussian
- presence of these latent variables is the key new wrinkle
- model parameters are ϕ, μ_k, Σ_k

Maximum likelihood estimation

- model parameters are ϕ, μ_k, Σ_k
- as usual, write down likelihood for $w = (\phi, \mu_k, \Sigma_k)$

$$\begin{aligned}\ell(w) &= \sum_{i=1}^{N} \log p(x_i; w) \\ &= \sum_{i=1}^{N} \log \sum_{z_i=1}^{K} p(x_i \mid z_i) p(z_i)\end{aligned}$$

- this function is *nonconvex* due to sum over values of z_i

Maximum likelihood estimation

- if z_i were known, problem is easy and becomes

$$\ell(w) = \sum_{i=1}^{N} \log p(x_i \mid z_i) + \sum_{i=1}^{N} \log p(z_i)$$

- maximizing with respect to ϕ, μ, Σ gives

$$\phi_j = \frac{1}{N} \sum_{i=1}^{N} [z_i = j], \quad \mu_j = \frac{\sum_{i=1}^{N} [z_i = j] x_i}{\sum_{i=1}^{N} [z_i = j]}$$

  similar expression for Σ

- *i.e.*, if z_i were known, nearly identical to maximum likelihood estimates in GDA (with z_i's as class labels)

EM algorithm

- **idea**: iteratively guess the z_i and then use formulas above:

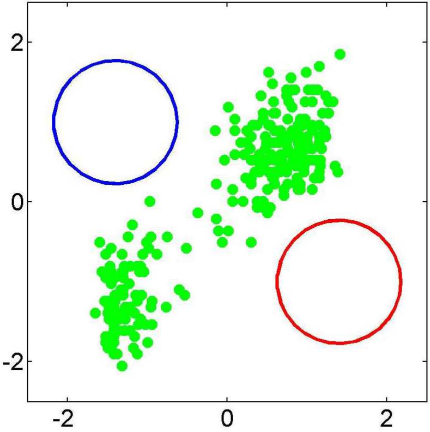
  1. E-step: compute $\rho_{ij} = p(z_i = j \,|\, x_i; \theta, \mu, \Sigma)$
  2. M-step: use formulas above with ρ_{ij} in place of $[z_i = j]$

- E-step computes posterior probability of z_i's, given data and current setting of parameters; 'soft guesses' for values of z_i

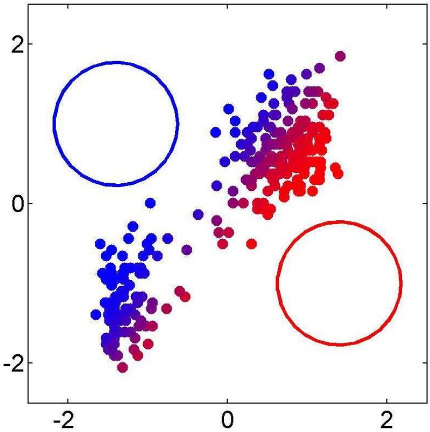
- M-step is maximum likelihood estimation, but there is uncertainty around the value of the z_i and that's incorporated in estimates

- a 'soft' version of k-means in this context

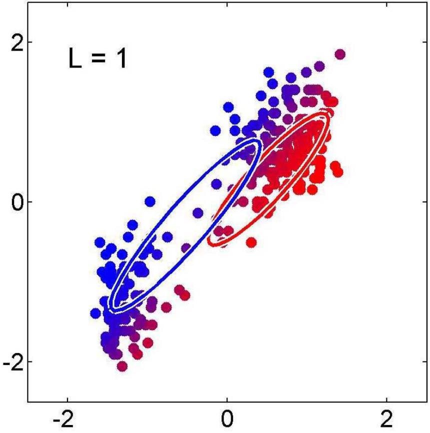
Gaussian mixture model



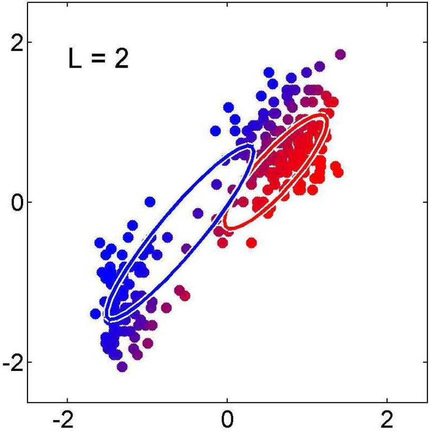
Gaussian mixture model



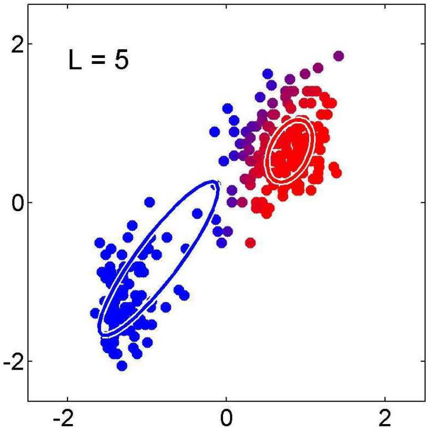
Gaussian mixture model



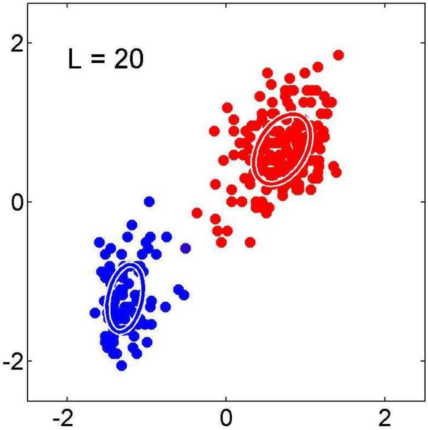
Gaussian mixture model



Gaussian mixture model



Gaussian mixture model



EM algorithm

- in general, EM algorithm is standard approach to maximum likelihood estimation with latent variable models
- data $\mathcal{D} = \{x_1, \ldots, x_N\}$
- want to fit model $p(x, z)$ with z hidden
- likelihood is given by

$$\ell(w) = \sum_{i=1}^{N} \log p(x; w) = \sum_{i=1}^{N} \log \sum_{z} p(x, z; w)$$

- often the case that maximum likelihood estimation of x would be easy if z were known, so alternate the two steps

EM algorithm

- EM algorithm can be motivated and analyzed in various ways
- iteratively lower bound ℓ, then maximize that lower bound
- for each i, let q_i be a distribution over z's

$$\begin{aligned}
\sum_{i=1}^{N} \log p(x_i) &= \sum_{i=1}^{N} \log \sum_{z_i} p(x_i, z_i) \\
&= \sum_{i=1}^{N} \log \sum_{z_i} q_i(z_i) \frac{p(x_i, z_i)}{q_i(z_i)} \\
&\geq \sum_{i=1}^{N} \sum_{z_i} q_i(z_i) \log \frac{p(x_i, z_i)}{q_i(z_i)}
\end{aligned}$$

by Jensen's inequality

EM algorithm

- previous formula gives lower bound for *any* q_i; ideally, have the lower bound be tight (inequality holds with equality) for current value of w

- can show that this is the case when $q_i(z_i) = p(z_i \mid x_i; w)$; it suffices that $q_i(z_i) \propto p(x_i, z_i; w)$, so

$$\begin{aligned} q_i(z_i) &= \frac{p(x_i, z_i; w)}{\sum_z p(x_i, z; w)} \\ &= \frac{p(x_i, z_i; w)}{p(x_i; w)} \\ &= p(z_i \mid x_i; w) \end{aligned}$$

- E-step (above): obtain lower bound (has form of an expectation) on ℓ

- M-step: maximize this lower bound with respect to w

EM algorithm

- can show that this algorithm converges because it monotonically improves the log likelihood
- *i.e.*, can show $\ell(w^k) < \ell(w^{k+1})$

EM algorithm

- EM algorithm can also be viewed as coordinate ascent on

$$J(q, w) = \sum_{i=1}^{N} \sum_{z_i} q_i(z_i) \log \frac{p(x_i, z_i; w)}{q_i(z_i)}$$

- E-step: maximization with respect to q

- M-step: maximization with respect to w

Factor analysis

- fitting Gaussian mixture model to data $x_1, \ldots, x_N \in \mathbf{R}^n$ assumes enough data ($N \gg n$) to discern this structure

- if $n \gg N$, cannot even fit a single Gaussian

- here, data points span low-dimensional subspace of $\mathbf{R}^n$, so MLEs of the parameters result in degenerate Gaussian (singular covariance matrix) that puts all mass in affine space spanned by the data

- consider models that explicitly handle low rank structure

Factor analysis

- consider generative model $p(x,z)$ given by

$$\begin{aligned} z &\sim \mathrm{N}(0, I) \\ x \,|\, z &\sim \mathrm{N}(\mu + \Lambda z, \Psi) \end{aligned}$$

  where $\mu \in \mathbf{R}^n$, $\Lambda \in \mathbf{R}^{n \times k}$, $\Psi \in \mathbf{R}^{n \times n}$ diagonal

- x observed, z latent

- low dimensional structure: $k < n$, *i.e.*, data is generated by affine transformation of k-dimensional Gaussian (plus noise)

Factor analysis

- $p(x, z)$ is Gaussian, and need to find its mean and covariance from the generative model

- ideally, would want to maximize log (marginal) likelihood of data, using marginal distribution of x, but this function is hard to optimize

- so, use EM

  - E-step: compute $q_i(z_i) = p(z_i \mid x_i)$ (also Gaussian)
  - M-step: maximize lower bound

$$\sum_{i=1}^{N} \int_{z_i} q_i(z_i) \log \frac{p(x_i, z_i)}{q_i(z_i)} \, dz_i$$

- involves some messy algebra, but can obtain closed form solutions for all these subproblems (matrix computations)

Principal components analysis

Dimensionality reduction

- model data $x \in \mathbf{R}^n$ as approximately lying in some k-dimensional subspace, with $k \ll n$
- has many different use cases
  - data compression
  - data visualization
  - noise reduction
  - preprocessing for supervised learning
  - feature discovery
  - structure discovery

Principal components analysis

- let $\mathcal{D} = \{x_1, \ldots, x_N\}$, with $x_i \in \mathbf{R}^n$, $n < N$
- rescale data to have mean zero and unit variance
  1. replace x_i with $x_i - (1/N)\sum_i x_i$
  2. replace x_i^j with x_i^j/σ_j, where $\sigma_j = (1/N)\sum_i (x_i^j)^2$

Principal components analysis

- several ways to motivate PCA
- select directions on which to project points to maximize variance
- compute top k eigenvectors of empirical covariance matrix
- pick k-dimensional basis so approximation error of projecting data onto it is minimized

Principal components analysis

- given $\mathcal{D}$, find unit vector u such that projection of $\mathcal{D}$ onto direction u has maximum variance
- length of projection of x onto u is $x^T u$, so solve

$$\begin{array}{ll} \text{maximize} & (1/N)\sum_{i=1}^N (x_i^T u)^2 \\ \text{subject to} & \|u\|_2 = 1 \end{array}$$

- objective can be rewritten as quadratic form $u^T \Sigma u$, where

$$\Sigma = \frac{1}{N}\sum_{i=1}^N x_i x_i^T$$

is **empirical covariance matrix** of (preprocessed) $\mathcal{D}$

- so solution of problem above is computing principal eigenvector of Σ

Principal components analysis

- in general, find top k eigenvectors $u_1, \ldots, u_k$ of Σ
- these give an orthonormal basis for $\mathbf{R}^k$
- compute rank k approximation to x_i as

$$y_i = (u_1^T x_i, \ldots, u_k^T x_i)$$

- choice of u_i maximizes $\sum_i \|y_i\|_2^2$

Topic models

Topic models

- topic models: methods for automatically organizing, understanding, searching, and summarizing large electronic archives
  - discover hidden themes that pervade the collection
  - annotate documents with those themes
  - use annotations to organize, summarize, and search texts
- unsupervised generative latent variable models of document structure
- originally introduced by Blei, Ng, and Jordan (2003); much subsequent work by Blei and collaborators, among many others

Topic models

- idea: documents composed of multiple topics
- each topic is a distribution over words
- each document is a mixture of corpus-wide topics
- each word is drawn from one of these topics

Latent Dirichlet allocation

- generative model $p(\theta, z, w \mid \alpha, \beta)$

$$
\begin{aligned}
\theta &\sim \mathrm{Dirichlet}(\alpha) \\
z_n &\sim \mathrm{Multinomial}(\theta), \quad n = 1, \ldots, N \\
w_n &\sim \mathrm{Multinomial}(\beta_{z_n}), \quad n = 1, \ldots, N
\end{aligned}
$$

- estimate parameters by, *e.g.*, maximizing log-likelihood

$$\ell(\alpha, \beta) = \sum_{d=1}^{D} \log p(\mathbf{w} \mid \alpha, \beta)$$

where $\mathbf{w}_1, \ldots, \mathbf{w}_D$ are documents (training set)

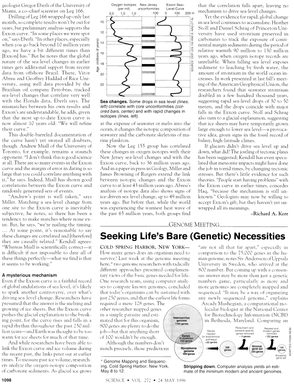
- want to compute posterior of latent variables

- conceptually, use EM (but need approximations here)

A 100 topic model of Science 1980-2000

sound	quantum	brain	computer	ice
speech	laser	memory	data	climate
acoustic	light	human	information	ocean
language	optical	visual	problem	sea
sounds	electron	cognitive	computers	temperature
stars	research	materials	fossil	volcanic
universe	national	organic	species	years
galaxies	science	molecules	evolution	fig
astronomers	new	molecular	birds	deposits
star	funding	polymer	evolutionary	rocks

Topic proportions in documents



geologist Gregor Eberli of the University of Miami, a co–chief scientist on Leg 166.

Drilling of Leg 166 wrapped up only last month, so complete results won't be out for years, but preliminary analysis supports the Exxon curve. "In some places we were spot on," says Eberli. "In other places, especially when you go back beyond 10 million years ago, we have a bit different times than [Exxon] has." But he notes that the global nature of the sea-level changes in earlier times gets additional support from recent data from offshore Brazil. There, Vitor Abreu and Geoffrey Haddad of Rice University, using well data provided by the Brazilian oil company Petrobras, tracked sea-level changes that correlate very well with the Florida data, Eberli says. The mismatches between his own results and Exxon's are understandable, he adds, given that the most up-to-date Exxon curve is now almost 10 years old: "We will refine their curve."

This double-barreled documentation of the curve hasn't yet swayed all doubters, though. Andrew Miall of the University of Toronto, for example, remains a staunch opponent. "I don't think this is good science at all. There are so many events in the Exxon curve and the margin of error in dating is so large that you could correlate anything with it," he says. Indeed, Miall has shown good correlations between the Exxon curve and randomly generated sets of events.

"Andrew's point is well taken," says Miller. Matching a sea-level change from one site to the Exxon curve is inevitably subjective, he notes, so there has been a tendency to make matches where none exist. But, he says, "we're nailing the timing. … At some point, it's reasonable to say these changes are correlated and [therefore] they are causally related." Kendall agrees: "Whereas Miall is scientifically correct—it is difficult if not impossible to date all of these things perfectly—what we find is that it seems to be working."

A mysterious mechanism

Even if the Exxon curve is a faithful record of global undulations of sea level, it's likely to spark another controversy, over what's driving sea-level change. Researchers have presumed that the answer is the melting and growing of ice sheets. But the Exxon curve pushes the glacial explanation to the breaking point, for the curve rises and falls in a rapid rhythm throughout the past 250 million years—and Earth was thought to be too warm for ice sheets for much of that time.

And while researchers have been able to link the Exxon curve and ice volume during the recent past, the links peter out at earlier times. To measure past ice volume, researchers analyze the oxygen-isotope composition of carbonate sediments. As glacial ice grows at the expense of seawater or melts into the ocean, it changes the isotopic composition of seawater and the carbonate skeletons of marine plankton.

Sea changes. Some drops in sea level *(lines, left)* correlate with core unconformities *(colored bars, center)* and with rapid changes in isotopes *(lines, left)*.

Now the Leg 155 group has correlated these changes in oxygen isotopes with their New Jersey sea-level changes and with the Exxon curve, back to 36 million years ago. And in a paper in press in *Geology*, Miller and James Browning of Rutgers extend the link between isotopic changes and the Exxon curve to at least 43 million years ago. Abreu's analysis of isotope data also shows signs of ice-driven sea-level change, up to 49 million years ago. But before that, while the world was experiencing the warmest heat wave of the past 65 million years, both groups find that the correlation falls apart, leaving no mechanism to drive sea-level changes.

Yet the evidence for rapid, global change in sea level continues to accumulate. Heather Stoll and Daniel Schrag of Princeton University have used strontium preserved in carbonates to track the exposure of continental margin sediments during the period of relative warmth 90 million to 130 million years ago, when oxygen isotope records are unreliable. When falling sea level exposes sediment to leaching by fresh water, the amount of strontium in the world ocean increases. In work presented at last fall's meeting of the American Geophysical Union, the researchers found that seawater strontium doubled in a few hundred thousand years, suggesting rapid sea-level drops of 30 to 50 meters, and the drops coincide with major falls in the Exxon curve. Stoll and Schrag also turn to a glacial explanation, suggesting that ice sheets may have temporarily grown large enough to lower sea level—a provocative idea, given signs in the fossil record of balmy, high-latitude climes.

If glaciers didn't drive sea level up and down, what did? The jostling of tectonic plates has been suggested; Kendall has even speculated that meteorite impacts might have done the job in torrid times, by changing tectonic stresses. But there's little evidence for such theories. "People start having problems" with the Exxon curve in earlier times, concedes Haq, "because the mechanism is still unknown." Geologists may now be willing to accept Exxon's gift, but they haven't yet unwrapped all its meanings.

–Richard A. Kerr

GENOME MEETING

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.

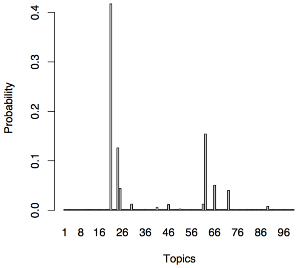
Although the numbers don't match precisely, those predictions "are not all that far apart," especially in comparison to the 75,000 genes in the human genome, notes Siv Andersson of Uppsala University in Sweden, who arrived at the 800 number. But coming up with a consensus answer may be more than just 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.

1098

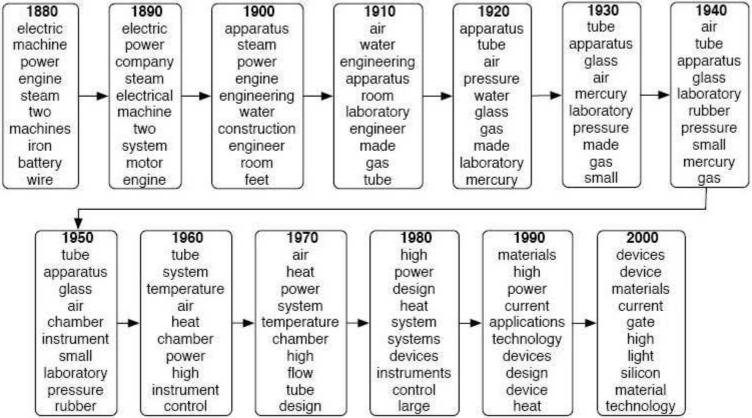
SCIENCE • VOL. 272 • 24 MAY 1996



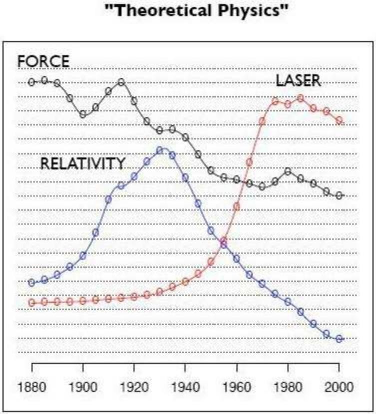
Variants and applications

- finding similar documents
- measuring scholarly impact (detect influential articles)
- discover evolution of topics over time
- discover correlations between topics
- annotate images with captions
- characterizing political decisions
- organize and browse large document collections

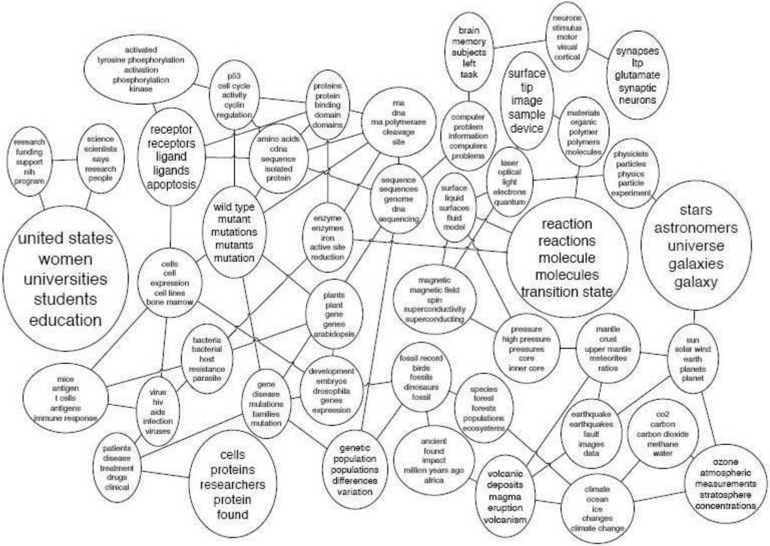
Model Evolution of Topics over Time



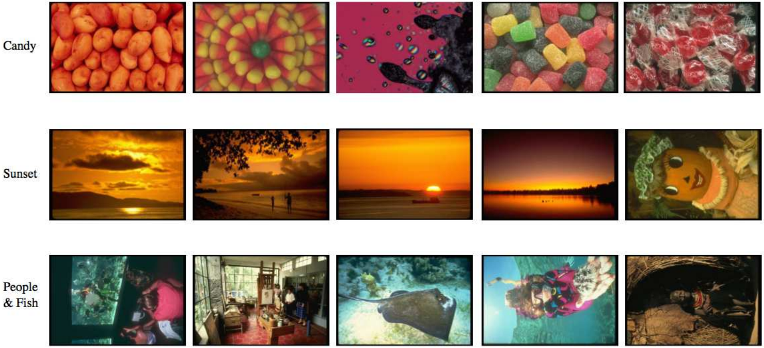
Visualizing Trends Within Topics



Model Connections Between Topics



Matching Words and Pictures



Matching Words and Pictures

True caption
market people
Corr–LDA
people market pattern textile display

True caption
scotland water
Corr–LDA
scotland water flowers hills tree

True caption
bridge sky water
Corr–LDA
sky water buildings people mountain

True caption
sky tree water
Corr–LDA
tree water sky people buildings

True caption
birds tree
Corr–LDA
birds nest leaves branch tree

True caption
fish reefs water
Corr–LDA
fish water ocean tree coral

True caption
mountain sky tree water
Corr–LDA
sky water tree mountain people

True caption
clouds jet plane
Corr–LDA
sky plane jet mountain clouds