Intelligent Agents
Probabilistic Models for Sequential Structures

Prof. Dr. Ralf Möller
Universität zu Lübeck
Institut für Informationssystem

Motivation: Part Of Speech Tagging

- Annotate each word in a sentence with a part-of-speech (POS) tags.
- Lowest level of syntactic analysis.

John saw the saw and decided to take it to the table.
NNP VBD DT NN CC VBD TO VB PRP IN DT NN

- Useful for subsequent syntactic parsing and word sense disambiguation

- Topic modeling as discussed before could be extended to better consider POS tags

Information Extraction

- Identify phrases in language that refer to specific types of entities and relations in text.
- Named entity recognition is the task of identifying names of people, places, organizations, etc. in text.

  people organizations places

  – Michael Dell is the CEO of Dell Computer Corporation and lives in Austin Texas.
- Extract pieces of information relevant to a specific application, e.g. used car ads:

  make model year mileage price

  – For sale, 2002 Toyota Prius, 20,000 mi, $15K or best offer. Available starting July 30, 2006.

Semantic Role Labeling

- For each clause, determine the semantic role played by each noun phrase that is an argument to the verb.

  agent patient source destination instrument

  – John drove Mary from Austin to Dallas in his Toyota Prius.

  – The hammer broke the window.

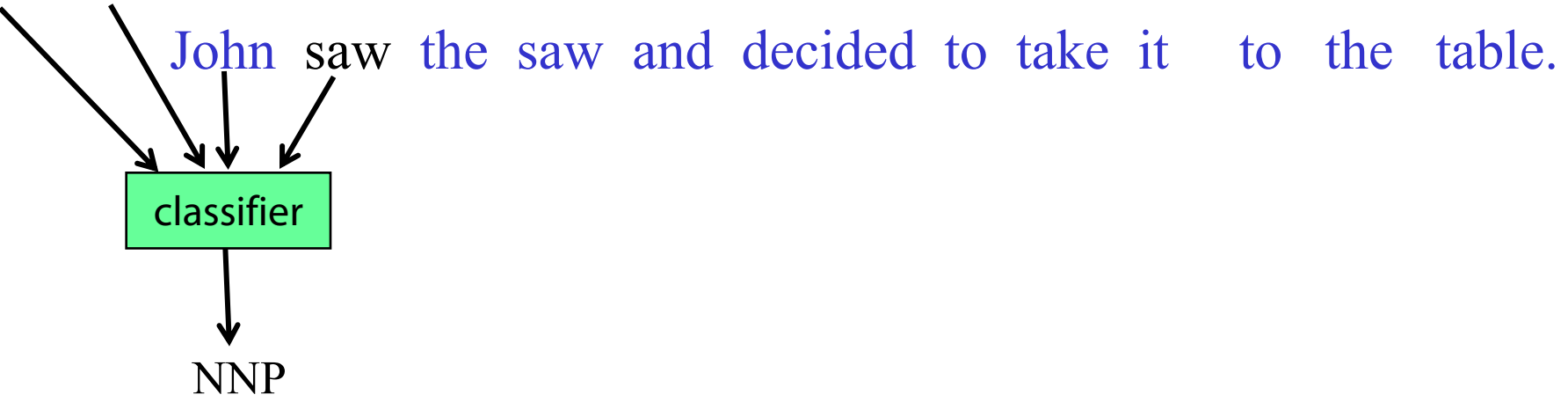
- Also referred to a "case role analysis," "thematic analysis," and "shallow semantic parsing"

Sequence Labeling as Classification

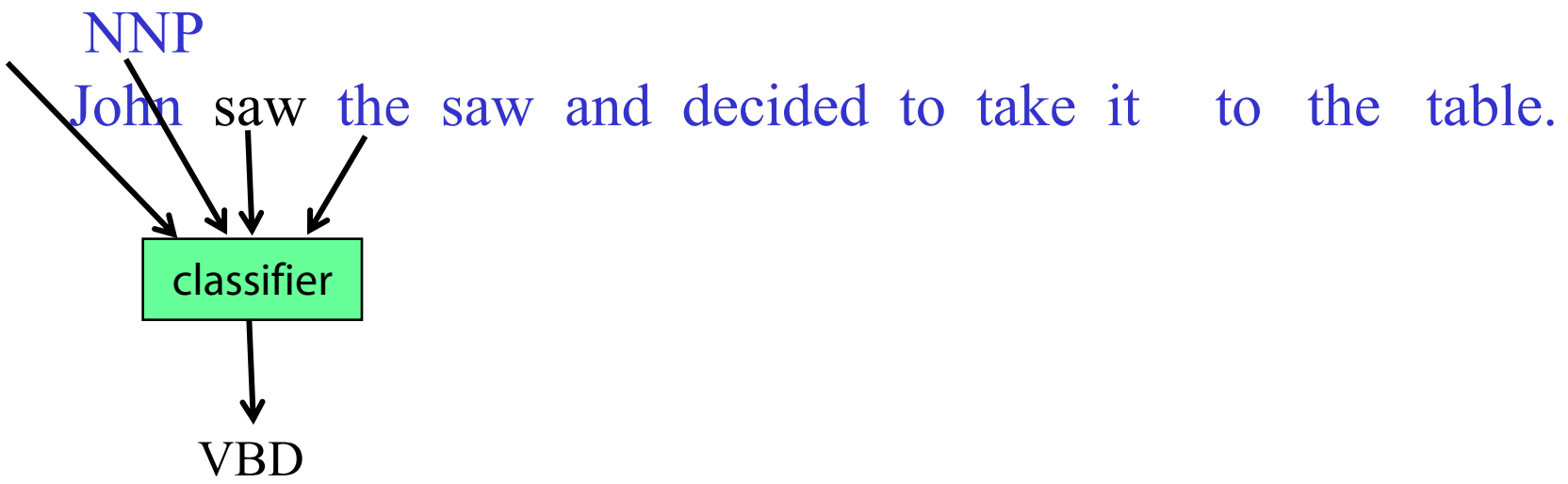
Using Outputs as Inputs

- Better input features are usually the categories of the surrounding tokens, but these are not available yet.
- Can use category of either the preceding or succeeding tokens by going forward or back and using previous output.

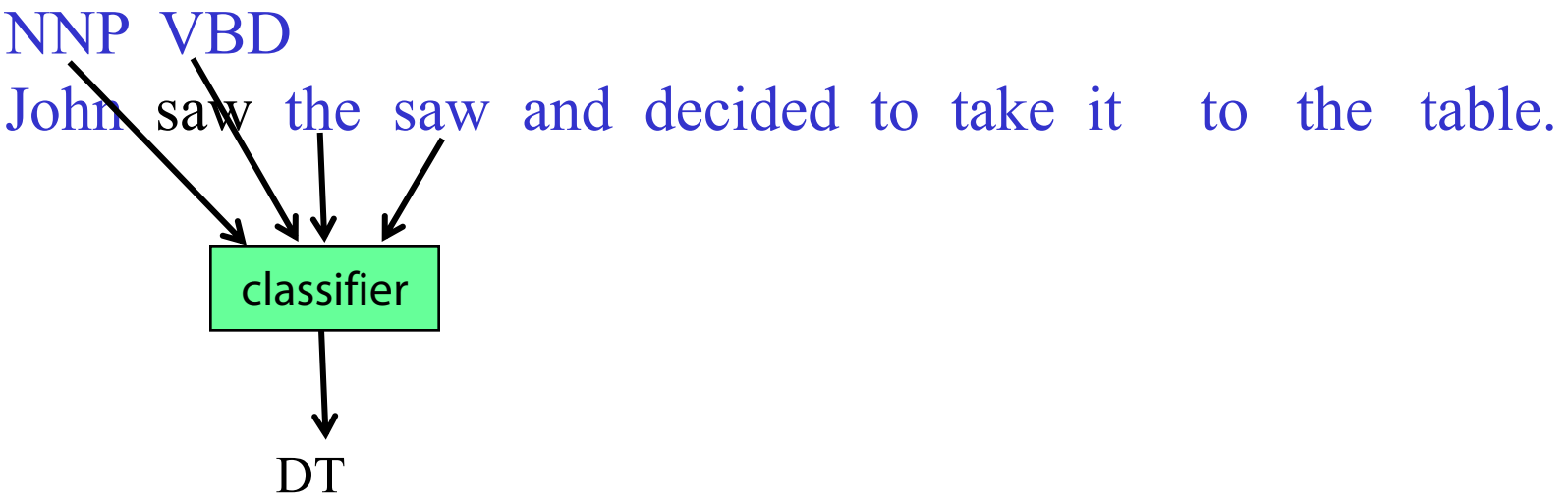
Forward Classification



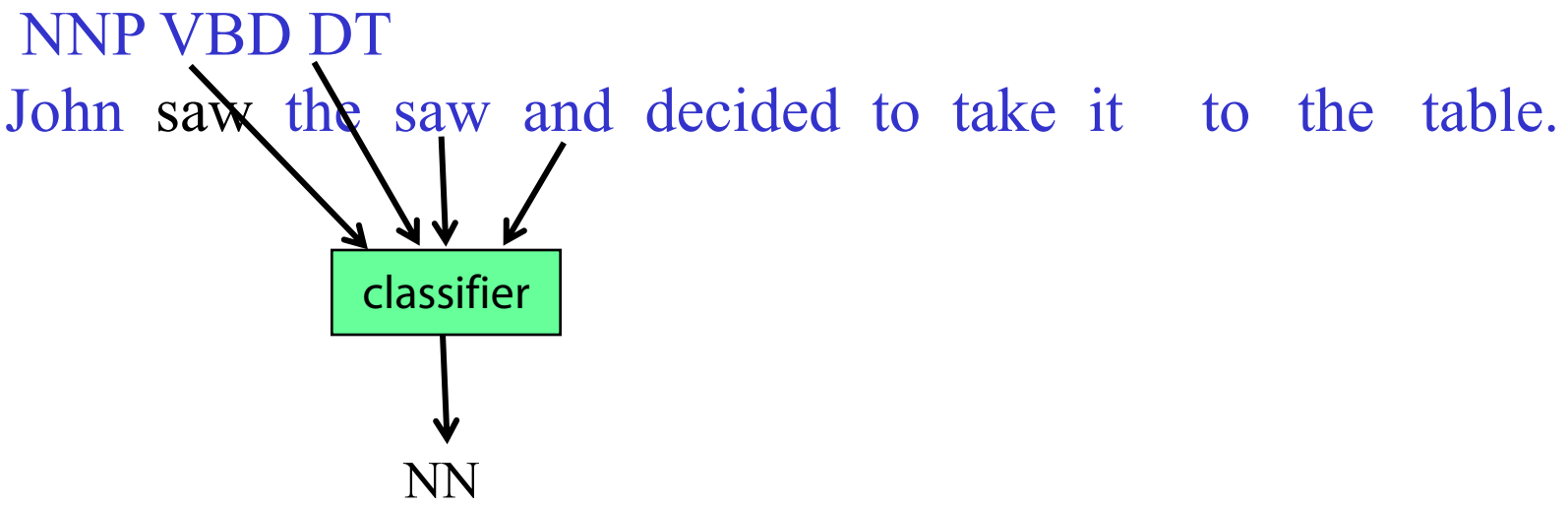
Forward Classification



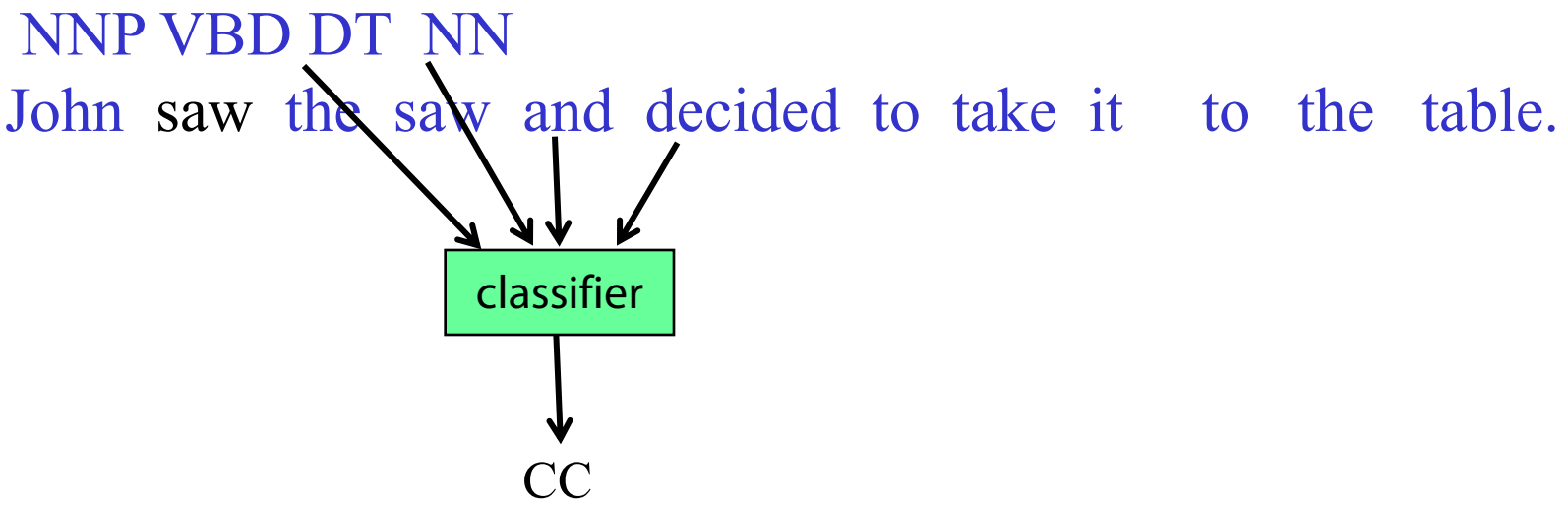
Forward Classification



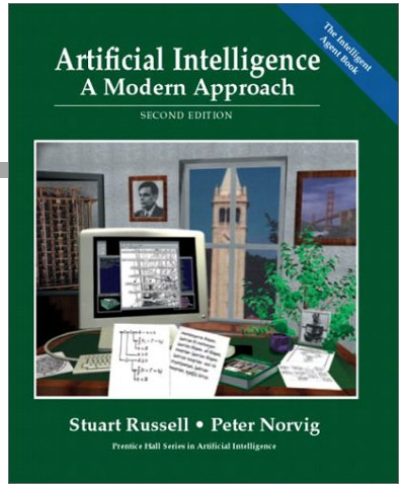
Forward Classification

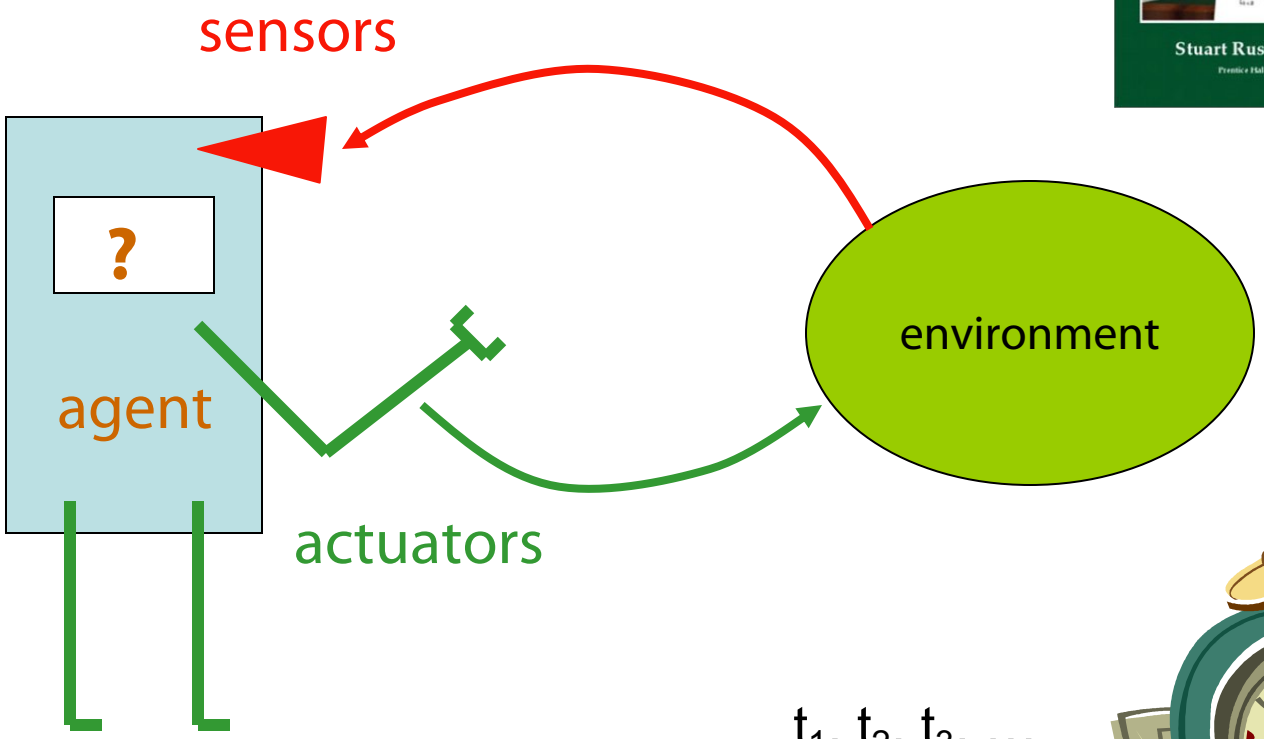


Forward Classification



More general perspective…





$t_1, t_2, t_3, \ldots$

Time and Uncertainty

- The world changes, we need to track and predict it
- Examples: diabetes management, traffic monitoring
- Uncertainty is everywhere
- Need temporal probabilistic graphical models
- Basic idea: copy state and evidence variables for each time step
- $\mathbf{X}_t$ – set of unobservable state variables at time t
  - e.g., BloodSugar_t, StomachContents_t
- $\mathbf{E}_t$ – set of evidence variables at time t
  - e.g., $\text{MeasuredBloodSugar}_t$, PulseRate_t, FoodEaten_t
- Assumes discrete time steps

States and Observations

- Process of change viewed as series of snapshots, each describing the state of the world at a particular time
- Time slice involves a set of random variables indexed by t:
  - the set of unobservable state variables $\mathbf{X}_t$
  - the set of observable evidence variable $\mathbf{E}_t$
- The observation at time t is $\mathbf{E}_t = \mathbf{e}_t$ for some set of values e_t
- The notation $\mathbf{X}_{a:b}$ denotes the set of variables from $\mathbf{X}_a$ to $\mathbf{X}_b$

Dynamic Bayesian Networks

- How can we model dynamic situations with a Bayesian network?

- Example: Is it raining today?

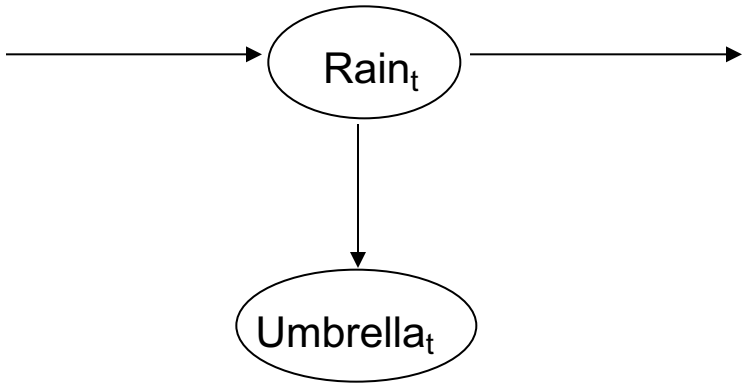
$$X_t = \{R_t\}$$

$$E_t = \{U_t\}$$

⇨ next step: specify dependencies among the variables.

The term "dynamic" means we are modeling a dynamic system, not that the network structure changes over time.

Example



DBN - Representation

- Problem: all previous random variables could have an influence on those of the current timestamp

  1. Necessity to specify an unbounded number of conditional probability tables, one for each variable in each slice,
  2. Each one might involve an unbounded number of parents.

- Solution:

  1. Assume that changes in the world state are caused by a stationary process (unmoving process over time).

$P(U_t / Parent(U_t))$ is the same for all t

Stationary Process/Markov Assumption

- Markov Assumption: X_t depends on some parent X_is
- First-order Markov process:

$$P(X_t|X_{0:t-1}) = P(X_t|X_{t-1})$$

Transition Model

  - kth order: depends on previous k time steps
- Sensor Markov assumption:

$$P(E_t|X_{0:t}, E_{0:t-1}) = P(E_t|X_t)$$

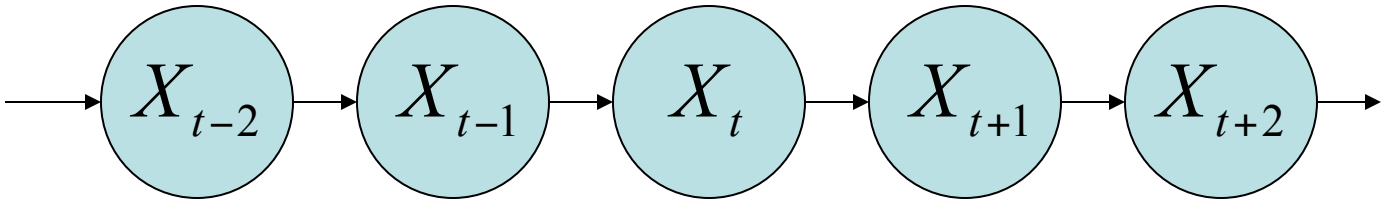
Sensor Model

- Assume stationary process: transition model:
  - $P(X_t|X_{t-1})$ and sensor model $P(E_t|X_t)$ are the same for all t
  - Changes in the world state governed by laws not changing over time

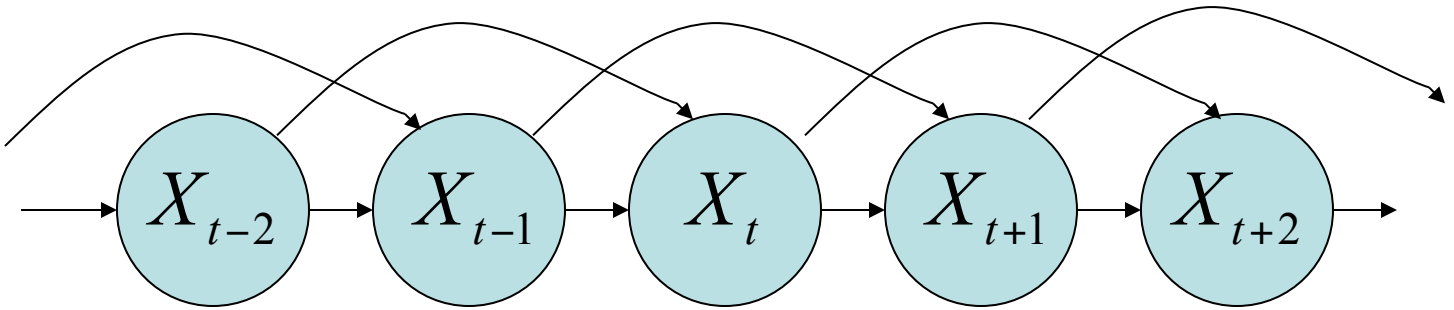
Dynamic Bayesian Networks

- There are two possible fixes if the approximation is too inaccurate:
  - Increasing the order of the Markov process model. For example, adding $Rain_{t-2}$ as a parent of $Rain_t$, which might give slightly more accurate predictions.
  - Increasing the set of state variables. For example, adding $Season_t$ to allow to incorporate historical records of rainy seasons, or adding $Temperature_t$, $Humidity_t$ and $Pressure_t$ to allow to use a physical model of rainy conditions.

Dynamic Bayesian Network

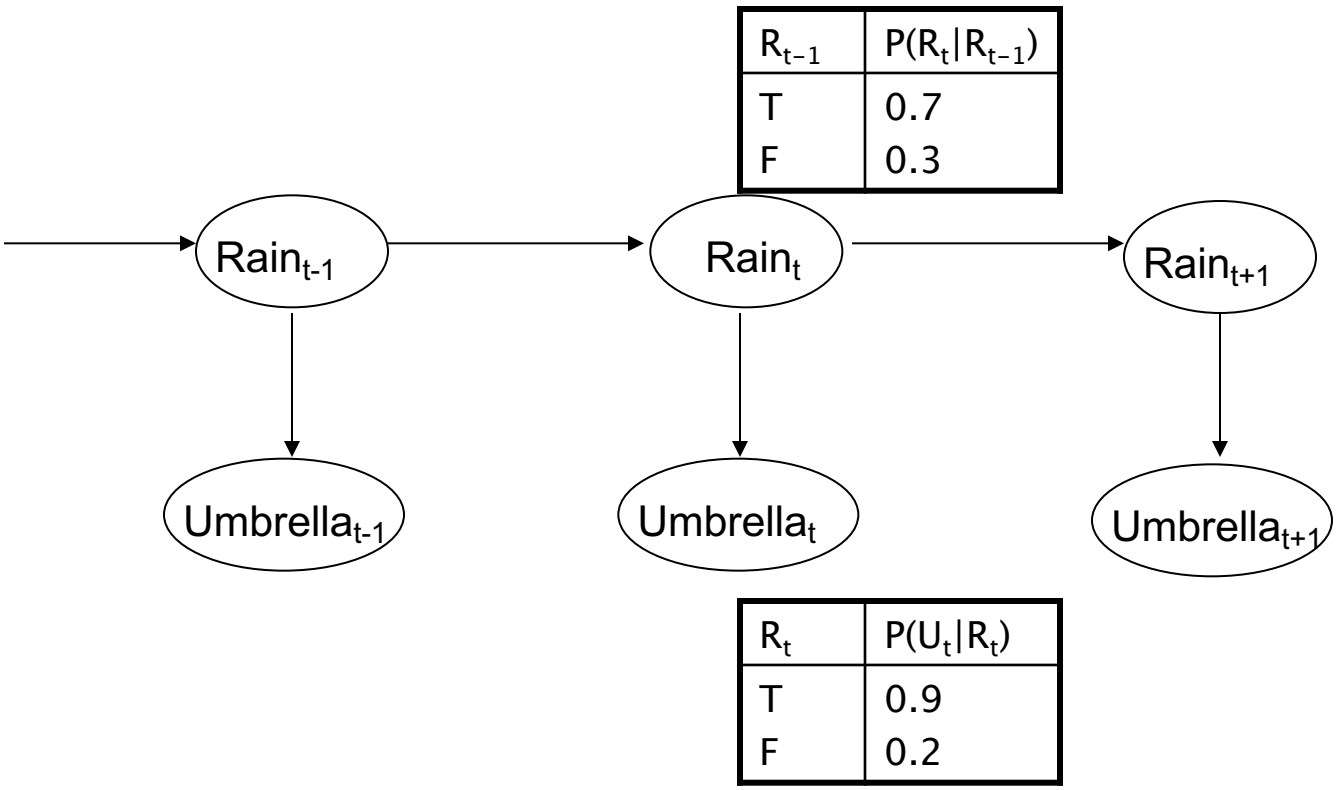


Bayesian network structure corresponding to a first-order of Markov process with state defined by the variables Xt.



A second order of Markov process

Example


R_{t-1} | $P(R_t|R_{t-1})$
T | 0.7
F | 0.3
$Rain_{t-1}$
$Rain_t$
$Rain_{t+1}$
$Umbrella_{t-1}$
$Umbrella_t$
$Umbrella_{t+1}$
R_t | $P(U_t|R_t)$
T | 0.9
F | 0.2

Complete Joint Distribution: Markov-1

- Given:
  - Transition model: $P(X_t|X_{t-1})$
  - Sensor model: $P(E_t|X_t)$
  - Prior probability: $P(X_0)$

- Then we can specify complete joint distribution:

$$P(X_0, X_1, \ldots, X_t, E_1, \ldots, E_t) = P(X_0) \prod_{i=1}^{t} P(X_i \mid X_{i-1}) P(E_i \mid X_i)$$

Inference Tasks

- Filtering: What is the probability that it is raining today, given all the umbrella observations up through today?
- Prediction: What is the probability that it will rain the day after tomorrow, given all the umbrella observations up through today?
- Smoothing: What is the probability that it rained yesterday, given all the umbrella observations through today?
- Most likely explanation / most probable explanation: if the umbrella appeared the first three days but not on the fourth, what is the most likely weather sequence to produce these umbrella sightings?

DBN – Basic Inference

- Filtering or Monitoring:

Compute the belief state - the posterior distribution over the *current* state, given all evidence to date.

$$P(X_t / e_{1:t})$$

Filtering is what a rational agent needs to do in order to keep track of the current state so that the rational decisions can be made.

DBN – Basic Inference

- Filtering cont.

Given the results of filtering up to time *t*, one can easily compute the result for *t+1* from the new evidence e_{t+1}

$$P(X_{t+1} / e_{1:t+1}) = f(e_{t+1}, P(X_t / e_{1:t+1}))$$ (for some function *f*)

$$= P(X_{t+1} / e_{1:t}, e_{t+1})$$ (dividing up the evidence)

$$= \alpha P(e_{t+1} / X_{t+1}, e_{1:t}) P(X_{t+1} / e_{1:t})$$ (using Bayes' Theorem)

$$= \alpha P(e_{t+1} / X_{t+1}) P(X_{t+1} / e_{1:t})$$ (by the Markov property of evidence)

α is a normalizing constant used to make probabilities sum up to 1.

Bayes Rule

$$P(A \mid B) = P(A, B) / P(B)$$

$$P(A,B) = P(A \mid B)\, P(B) = P(B \mid A)\, P(A) = P(B, A)$$

Application of Bayes Rule

$$
\begin{aligned}
P(A \mid B, C) &= P(A, B, C) / P(B, C) \\
&= P(C, A, B) / P(B, C) \\
&= P(C \mid A, B)\, P(A, B) / P(B, C) \\
&= P(C \mid A, B)\, P(A \mid B)\, P(B) / (P(C \mid B)\, P(B)) \\
&= \alpha\, P(C \mid A, B)\, P(A \mid B)
\end{aligned}
$$

$$P(X_{t+1} / e_{1:t}, e_{t+1}) = \alpha P(e_{t+1} / X_{t+1}, e_{1:t}) P(X_{t+1} / e_{1:t})$$

DBN – Basic Inference

- Filtering cont.

Given the results of filtering up to time *t*, one can easily compute the result for *t+1* from the new evidence e_{t+1}

$$P(X_{t+1} / e_{1:t+1}) = f(e_{t+1}, P(X_t / e_{1:t+1}))$$ (for some function *f*)

$$= P(X_{t+1} / e_{1:t}, e_{t+1})$$ (dividing up the evidence)

$$= \alpha P(e_{t+1} / X_{t+1}, e_{1:t}) P(X_{t+1} / e_{1:t})$$ (using Bayes' Theorem)

$$= \alpha P(e_{t+1} / X_{t+1}) P(X_{t+1} / e_{1:t})$$ (by the Markov property of evidence)

α is a normalizing constant used to make probabilities sum up to 1.

Application of Bayes Rule

$$
\begin{aligned}
P(A \mid B) &= \Sigma_c P(A, c \mid B) \\
&= \Sigma_c P(A, c, B) / P(B) \\
&= \Sigma_c P(A \mid c, B)\, P(c, B) / P(B) \\
&= \Sigma_c P(A \mid c, B)\, P(c \mid B)\, P(B) / P(B) \\
&= \Sigma_c P(A \mid c, B)\, P(c \mid B)
\end{aligned}
$$

$$
P(X_{t+1} / e_{1:t}) = \sum_{X_t} P(X_{t+1} / x_t, e_{1:t}) P(x_t / e_{1:t})
$$

DBN – Basic Inference

- Filtering cont.

The second term $P(X_{t+1} / e_{1:t})$ represents a one-step prediction of the next step, and the first term $P(e_{t+1} / X_{t+1})$ updates this with the new evidence.

Now we obtain the one-step prediction for the next step by conditioning on the current state Xt:

$$P(X_{t+1} / e_{1:t+1}) = \alpha P(e_{t+1} / X_{t+1}) \sum_{X_t} P(X_{t+1} / x_t, e_{1:t}) P(x_t / e_{1:t})$$

$$= \alpha P(e_{t+1} / X_{t+1}) \sum_{X_t} P(X_{t+1} / \boldsymbol{x_t}) P(x_t / e_{1:t})$$

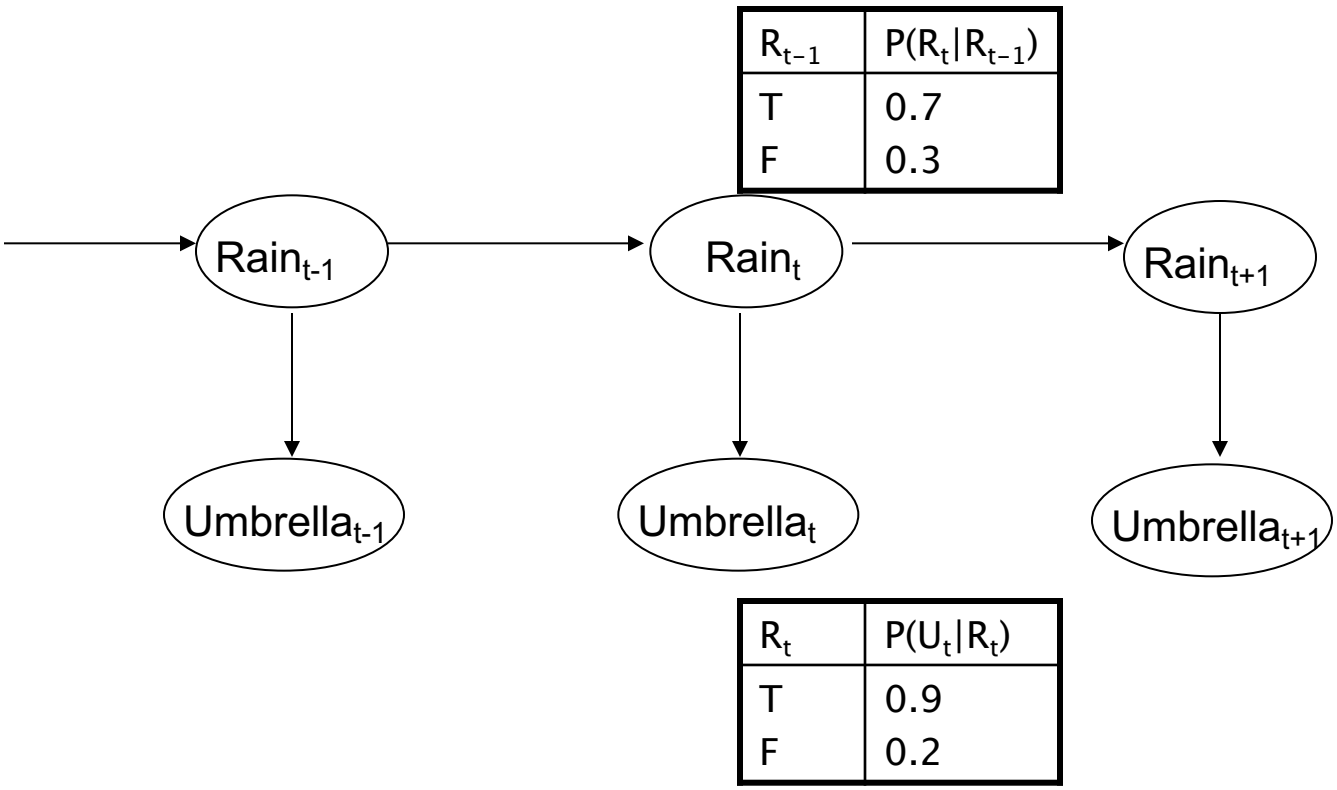
(using the Markov property)

Forward Messages

$\mathbf{f}_{1:t+1} = \text{FORWARD}(\mathbf{f}_{1:t}, \mathbf{e}_{t+1})$ where $\mathbf{f}_{1:t} = \mathbf{P}(\mathbf{X}_t | \mathbf{e}_{1:t})$

Time and space **constant** (independent of t)

Example $P(Rain_0) = (0.5\ 0.5)^T$



R_{t-1}	$P(R_t \mid R_{t-1})$
T	0.7
F	0.3

R_t	$P(U_t \mid R_t)$
T	0.9
F	0.2

DBN – Basic Inference

Illustration for two steps in the umbrella example:

- On day 1, the umbrella appears, so U1=true. The prediction from t=0 to t=1 is

$$P(R_1) = \sum_{r_0} P(R_1 / r_0) P(r_0)$$

and updating it with the evidence for t=1 gives

$$P(R_1 / u_1) = \alpha P(u_1 / R_1) P(R_1)$$

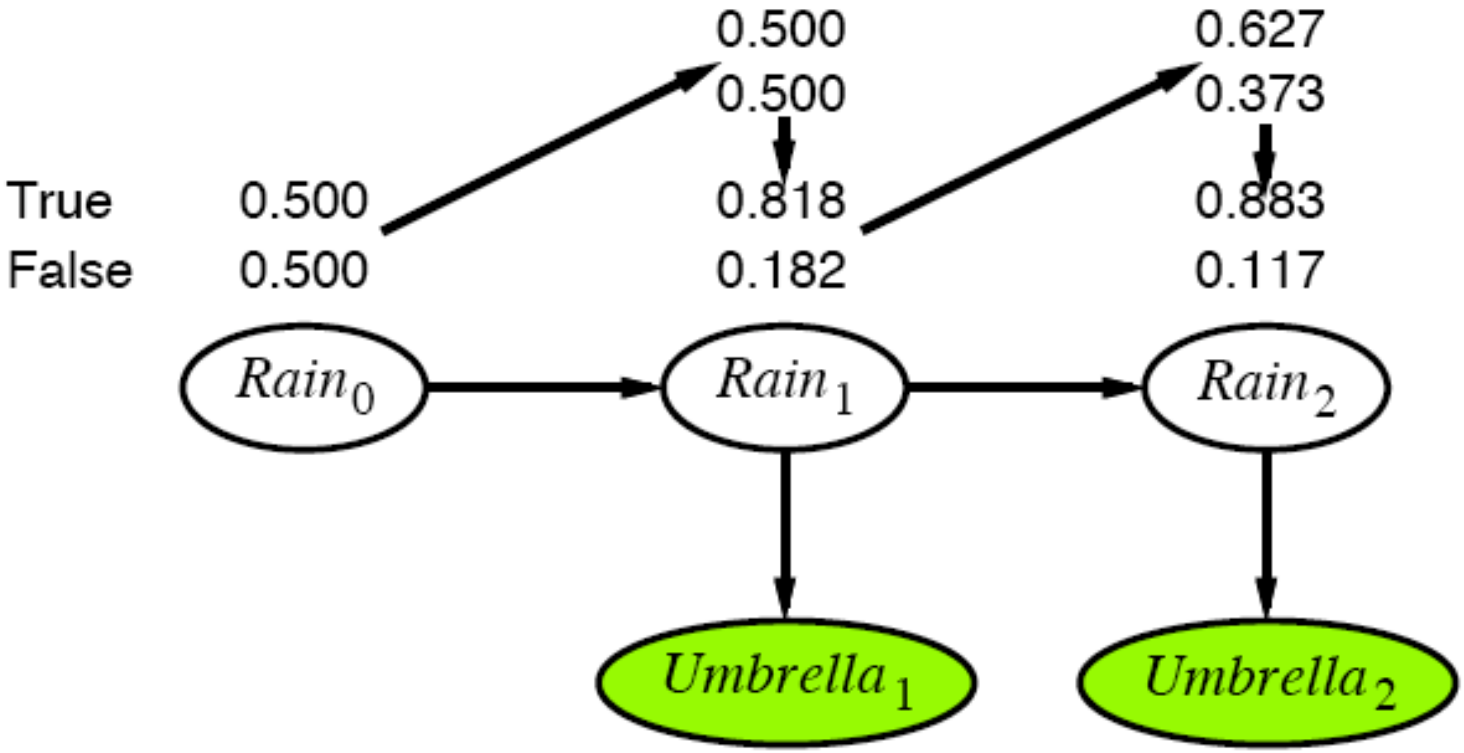
- On day 2, the umbrella appears, so U2=true. The prediction from t=1 to t=2 is

$$P(R_2 / u_1) = \sum_{r_1} P(R_2 / r_1) P(r_1 / u_1)$$

and updating it with the evidence for t=2 gives

$$P(R_2 / u_1, u_2) = \alpha P(u_2 / R_2) P(R_2 / u_1)$$

Example cntd.



DBN – Basic Inference

- Prediction:

  Compute the posterior distribution over the *future* state, given all evidence to date.

$$P(X_{t+k} / e_{1:t})$$

for some k>0

  The task of prediction can be seen simply as filtering without the addition of new evidence.

DBN – Basic Inference

- Smoothing or hindsight:

  Compute the posterior distribution over the *past* state, given all evidence up to the present.

$$P(X_k / e_{1:t})$$

for some k such that $0 \leq k < t$.

Hindsight provides a better estimate of the state than was available at the time, because it incorporates more evidence.

Smoothing

Divide evidence $\mathbf{e}_{1:t}$ into $\mathbf{e}_{1:k}$, $\mathbf{e}_{k+1:t}$:

$$\begin{aligned}\mathbf{P}(\mathbf{X}_k|\mathbf{e}_{1:t}) &= \mathbf{P}(\mathbf{X}_k|\mathbf{e}_{1:k},\mathbf{e}_{k+1:t}) \\ &= \alpha\mathbf{P}(\mathbf{X}_k|\mathbf{e}_{1:k})\mathbf{P}(\mathbf{e}_{k+1:t}|\mathbf{X}_k,\mathbf{e}_{1:k}) \\ &= \alpha\mathbf{P}(\mathbf{X}_k|\mathbf{e}_{1:k})\mathbf{P}(\mathbf{e}_{k+1:t}|\mathbf{X}_k) \\ &= \alpha\mathbf{f}_{1:k}\mathbf{b}_{k+1:t}\end{aligned}$$

Backward message computed by a backwards recursion:

$$\begin{aligned}\mathbf{P}(\mathbf{e}_{k+1:t}|\mathbf{X}_k) &= \textstyle\sum_{\mathbf{x}_{k+1}}\mathbf{P}(\mathbf{e}_{k+1:t}|\mathbf{X}_k,\mathbf{x}_{k+1})\mathbf{P}(\mathbf{x}_{k+1}|\mathbf{X}_k) \\ &= \textstyle\sum_{\mathbf{x}_{k+1}}P(\mathbf{e}_{k+1:t}|\mathbf{x}_{k+1})\mathbf{P}(\mathbf{x}_{k+1}|\mathbf{X}_k) \\ &= \textstyle\sum_{\mathbf{x}_{k+1}}P(\mathbf{e}_{k+1}|\mathbf{x}_{k+1})P(\mathbf{e}_{k+2:t}|\mathbf{x}_{k+1})\mathbf{P}(\mathbf{x}_{k+1}|\mathbf{X}_k)\end{aligned}$$

Forward–backward algorithm: cache forward messages along the way
Time linear in t (polytree inference), space $O(t|\mathbf{f}|)$

Application of Bayes Rule

$$
\begin{aligned}
P(A \mid B, C) &= P(A, B, C) / P(B, C) \\
&= P(C, A, B) / P(B, C) \\
&= P(C \mid A, B)\, P(A, B) / P(B, C) \\
&= P(C \mid A, B)\, P(A \mid B)\, P(B) / (P(C \mid B)\, P(B)) \\
&= \alpha\, P(C \mid A, B)\, P(A \mid B)
\end{aligned}
$$

$$\mathbf{P}(\mathbf{X}_k|\mathbf{e}_{1:k}, \mathbf{e}_{k+1:t}) = \alpha\mathbf{P}(\mathbf{X}_k|\mathbf{e}_{1:k})\mathbf{P}(\mathbf{e}_{k+1:t}|\mathbf{X}_k, \mathbf{e}_{1:k})$$

Smoothing

Divide evidence $\mathbf{e}_{1:t}$ into $\mathbf{e}_{1:k}$, $\mathbf{e}_{k+1:t}$:

$$\begin{aligned}\mathbf{P}(\mathbf{X}_k|\mathbf{e}_{1:t}) &= \mathbf{P}(\mathbf{X}_k|\mathbf{e}_{1:k},\mathbf{e}_{k+1:t}) \\ &= \alpha\mathbf{P}(\mathbf{X}_k|\mathbf{e}_{1:k})\mathbf{P}(\mathbf{e}_{k+1:t}|\mathbf{X}_k,\mathbf{e}_{1:k}) \\ &= \alpha\mathbf{P}(\mathbf{X}_k|\mathbf{e}_{1:k})\mathbf{P}(\mathbf{e}_{k+1:t}|\mathbf{X}_k) \\ &= \alpha\mathbf{f}_{1:k}\mathbf{b}_{k+1:t}\end{aligned}$$

Backward message computed by a backwards recursion:

$$\begin{aligned}\mathbf{P}(\mathbf{e}_{k+1:t}|\mathbf{X}_k) &= \textstyle\sum_{\mathbf{x}_{k+1}}\mathbf{P}(\mathbf{e}_{k+1:t}|\mathbf{X}_k,\mathbf{x}_{k+1})\mathbf{P}(\mathbf{x}_{k+1}|\mathbf{X}_k) \\ &= \textstyle\sum_{\mathbf{x}_{k+1}}P(\mathbf{e}_{k+1:t}|\mathbf{x}_{k+1})\mathbf{P}(\mathbf{x}_{k+1}|\mathbf{X}_k) \\ &= \textstyle\sum_{\mathbf{x}_{k+1}}P(\mathbf{e}_{k+1}|\mathbf{x}_{k+1})P(\mathbf{e}_{k+2:t}|\mathbf{x}_{k+1})\mathbf{P}(\mathbf{x}_{k+1}|\mathbf{X}_k)\end{aligned}$$

Forward–backward algorithm: cache forward messages along the way
Time linear in t (polytree inference), space $O(t|\mathbf{f}|)$

Application of Bayes Rule

$$
\begin{aligned}
P(A \mid B) &= \Sigma_c P(A, c \mid B) \\
&= \Sigma_c P(A, c, B) / P(B) \\
&= \Sigma_c P(A \mid c, B)\, P(c, B) / P(B) \\
&= \Sigma_c P(A \mid c, B)\, P(c \mid B)\, P(B) / P(B) \\
&= \Sigma_c P(A \mid c, B)\, P(c \mid B)
\end{aligned}
$$

$$\mathbf{P}(\mathbf{e}_{k+1:t}|\mathbf{X}_k) = \sum_{\mathbf{x}_{k+1}} \mathbf{P}(\mathbf{e}_{k+1:t}|\mathbf{X}_k, \mathbf{x}_{k+1}) \mathbf{P}(\mathbf{x}_{k+1}|\mathbf{X}_k)$$

Smoothing

Divide evidence $\mathbf{e}_{1:t}$ into $\mathbf{e}_{1:k}$, $\mathbf{e}_{k+1:t}$:

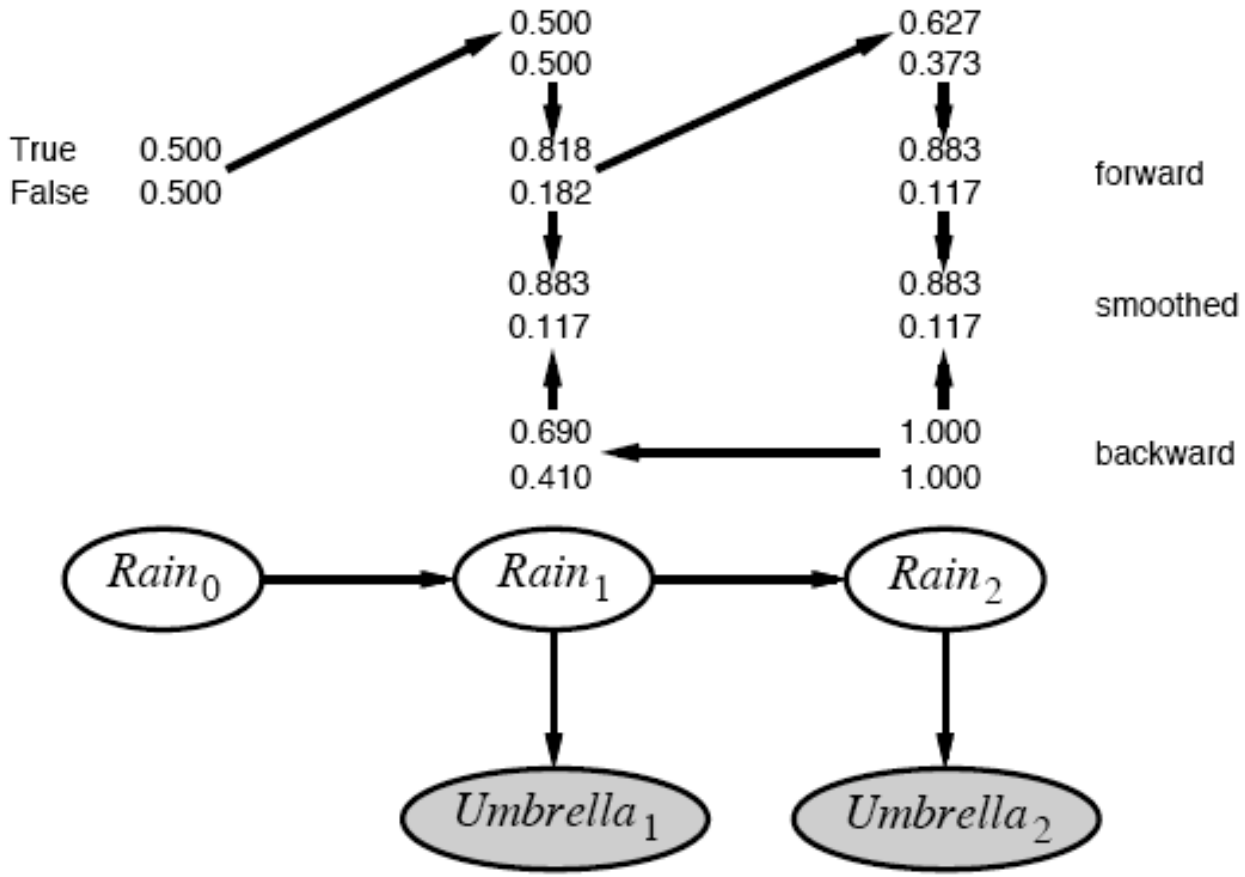
$$\begin{aligned}\mathbf{P}(\mathbf{X}_k|\mathbf{e}_{1:t}) &= \mathbf{P}(\mathbf{X}_k|\mathbf{e}_{1:k},\mathbf{e}_{k+1:t}) \\ &= \alpha\mathbf{P}(\mathbf{X}_k|\mathbf{e}_{1:k})\mathbf{P}(\mathbf{e}_{k+1:t}|\mathbf{X}_k,\mathbf{e}_{1:k}) \\ &= \alpha\mathbf{P}(\mathbf{X}_k|\mathbf{e}_{1:k})\mathbf{P}(\mathbf{e}_{k+1:t}|\mathbf{X}_k) \\ &= \alpha\mathbf{f}_{1:k}\mathbf{b}_{k+1:t}\end{aligned}$$

Backward message computed by a backwards recursion:

$$\begin{aligned}\mathbf{P}(\mathbf{e}_{k+1:t}|\mathbf{X}_k) &= \textstyle\sum_{\mathbf{x}_{k+1}}\mathbf{P}(\mathbf{e}_{k+1:t}|\mathbf{X}_k,\mathbf{x}_{k+1})\mathbf{P}(\mathbf{x}_{k+1}|\mathbf{X}_k) \\ &= \textstyle\sum_{\mathbf{x}_{k+1}}P(\mathbf{e}_{k+1:t}|\mathbf{x}_{k+1})\mathbf{P}(\mathbf{x}_{k+1}|\mathbf{X}_k) \\ &= \textstyle\sum_{\mathbf{x}_{k+1}}P(\mathbf{e}_{k+1}|\mathbf{x}_{k+1})P(\mathbf{e}_{k+2:t}|\mathbf{x}_{k+1})\mathbf{P}(\mathbf{x}_{k+1}|\mathbf{X}_k)\end{aligned}$$

Forward–backward algorithm: cache forward messages along the way
Time linear in t (polytree inference), space $O(t|\mathbf{f}|)$

Example contd.



DBN – Basic Inference

- Filtering cont.

The second term $P(X_{t+1} / e_{1:t})$ represents a one-step prediction of the next step, and the first term $P(e_{t+1} / X_{t+1})$ updates this with the new evidence.

Now we obtain the one-step prediction for the next step by conditioning on the current state Xt:

$$P(X_{t+1} / e_{1:t+1}) = \alpha P(e_{t+1} / X_{t+1}) \sum_{X_t} P(X_{t+1} / x_t, e_{1:t}) P(x_t / e_{1:t})$$

$$= \alpha P(e_{t+1} / X_{t+1}) \sum_{X_t} P(X_{t+1} / \boldsymbol{x_t}) P(x_t / e_{1:t})$$

(using the Markov property)

DBN – Basic Inference

- Most likely explanation:

Compute the sequence of states that is most likely to have generated a given sequence of observation.

$$\arg\max_{x_{1:t}} P(X_{1:t} \mid e_{1:t})$$

Algorithms for this task are useful in many applications, including, e.g., speech recognition.

Most-likely explanation

Most likely sequence $\neq$ sequence of most likely states!!!!

Most likely path to each $\mathbf{x}_{t+1}$
= most likely path to **some** $\mathbf{x}_t$ plus one more step

$$\max_{\mathbf{x}_1 \ldots \mathbf{x}_t} \mathbf{P}(\mathbf{x}_1, \ldots, \mathbf{x}_t, \mathbf{X}_{t+1} | \mathbf{e}_{1:t+1})$$
$$= \mathbf{P}(\mathbf{e}_{t+1} | \mathbf{X}_{t+1}) \max_{\mathbf{x}_t} \left(\mathbf{P}(\mathbf{X}_{t+1} | \mathbf{x}_t) \max_{\mathbf{x}_1 \ldots \mathbf{x}_{t-1}} P(\mathbf{x}_1, \ldots, \mathbf{x}_{t-1}, \mathbf{x}_t | \mathbf{e}_{1:t}) \right)$$

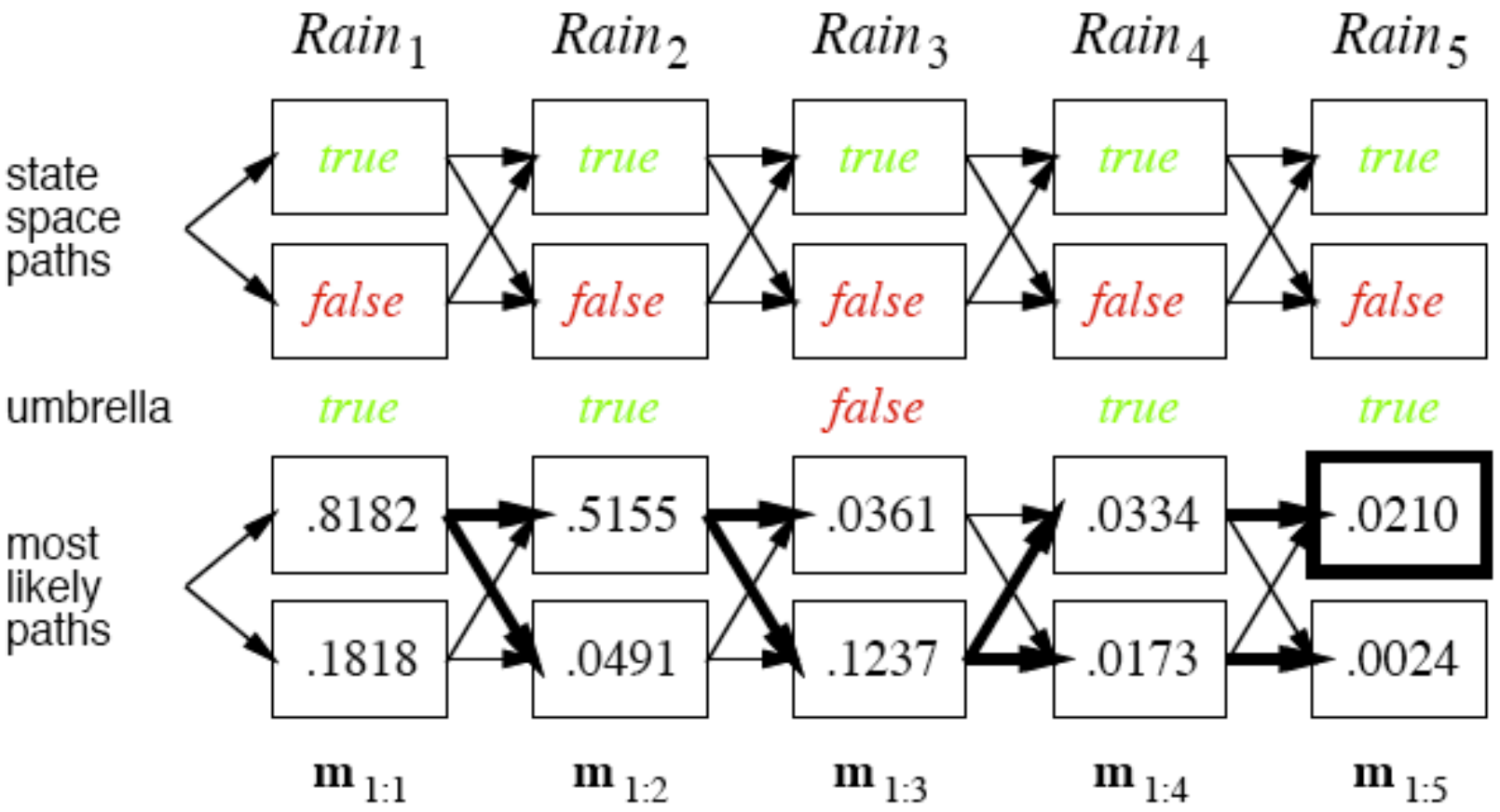
Identical to filtering, except $\mathbf{f}_{1:t}$ replaced by

$$\mathbf{m}_{1:t} = \max_{\mathbf{x}_1 \ldots \mathbf{x}_{t-1}} \mathbf{P}(\mathbf{x}_1, \ldots, \mathbf{x}_{t-1}, \mathbf{X}_t | \mathbf{e}_{1:t}),$$

I.e., $\mathbf{m}_{1:t}(i)$ gives the probability of the most likely path to state i.
Update has sum replaced by max, giving the Viterbi algorithm:

$$\mathbf{m}_{1:t+1} = \mathbf{P}(\mathbf{e}_{t+1} | \mathbf{X}_{t+1}) \max_{\mathbf{x}_t} (\mathbf{P}(\mathbf{X}_{t+1} | \mathbf{x}_t) \mathbf{m}_{1:t})$$

Rain/Umbrella Example



Hidden Markov Model (HMM)

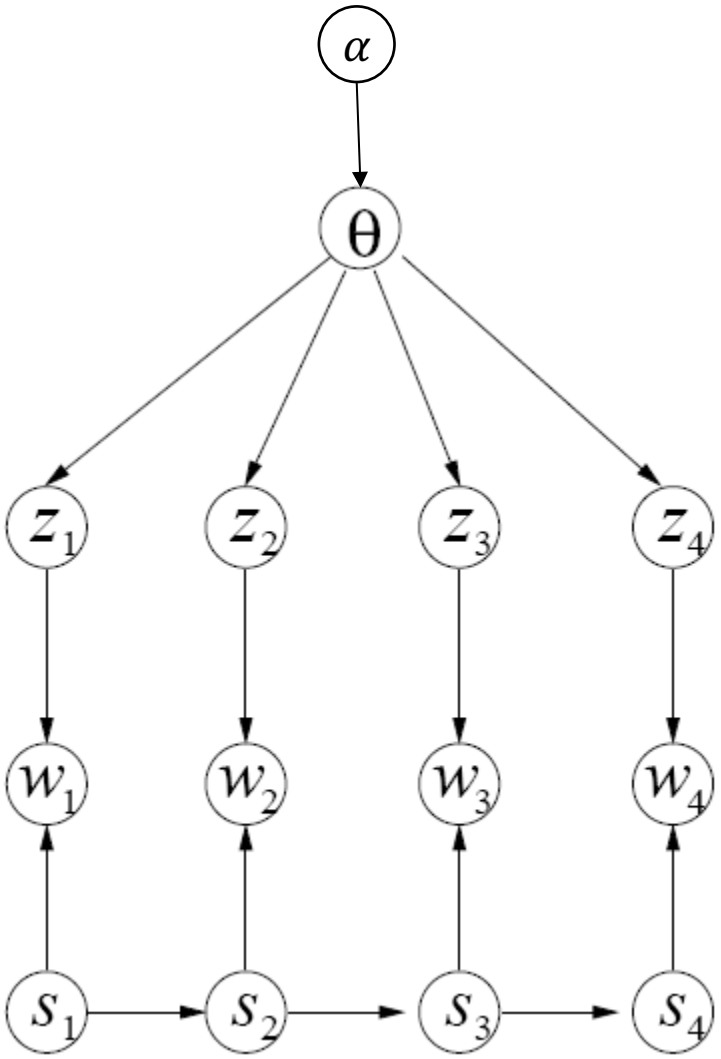
Consider special case of a dynamic Bayesian Network:

- Use vector of independent state variables $\mathbf{X}_t$
- Use vector of independent evidence variables $\mathbf{E}_t$
- This was already used in the rain-umbrella example
- For high-dimensional vectors the transition and sensor models become quite complex: $O(d^2)$ space

NB:

- In a general dynamic Bayesian network, state variables are not necessarily independent
- Even evidence variable might be dependent on one another (naïve Bayes does not work)

How to Incorporate Context into LDA?



LDA

HMM (e.g. for POS tagging)

Dynamic Topic Models

- In LDA the order of documents does not matter
- Not appropriate for sequential corpora (e.g., that span hundreds of years)
- Further, we may want to track how language changes over time
- Let the topics *drift* in a sequence.

1789

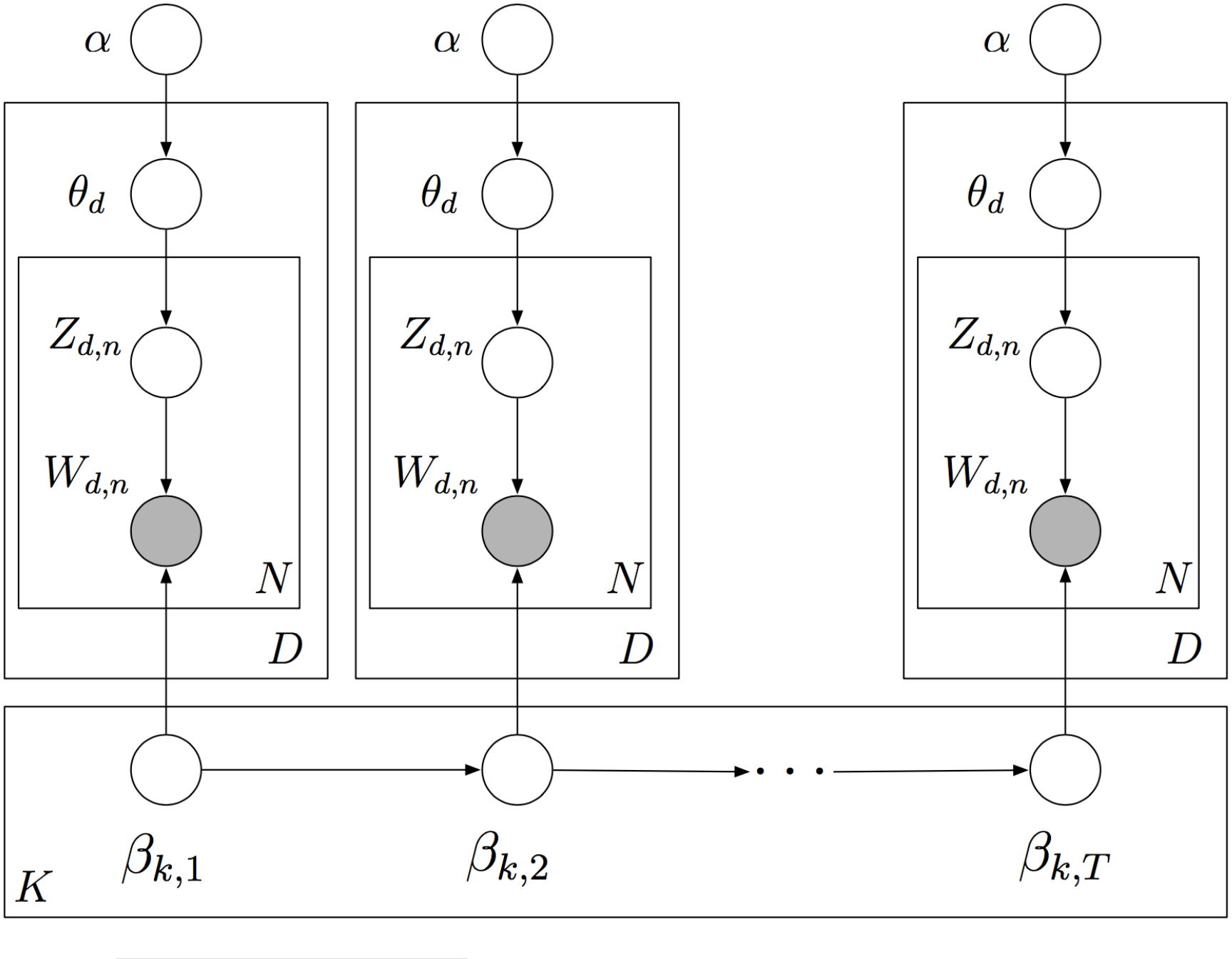
My fellow citizens: I stand here today humbled by the task before us, grateful for the trust you have bestowed, mindful of the sacrifices borne by our ancestors...

Inaugural addresses

2009

AMONG the vicissitudes incident to life no event could have filled me with greater anxieties than that of which the notification was transmitted by your order...

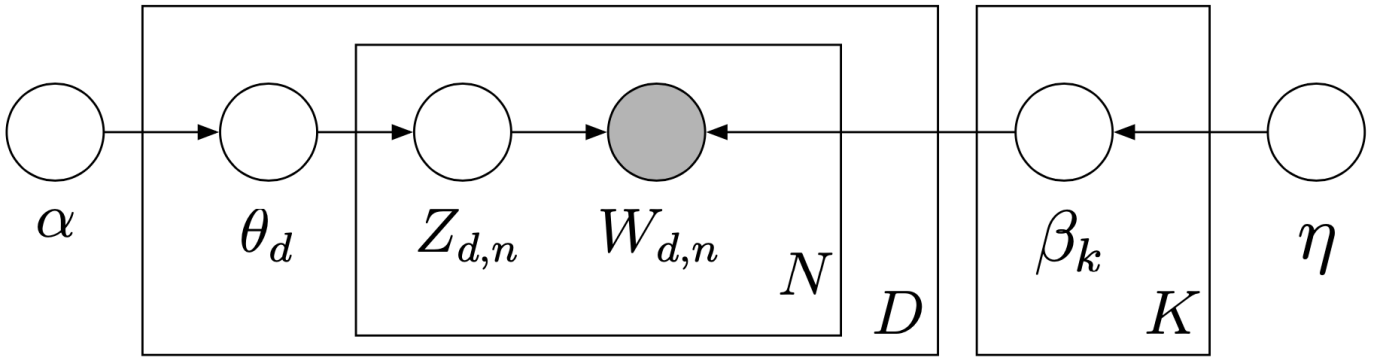
David M. Blei and John D. Lafferty. Dynamic topic models. In Proc. ICML '06. pp. 113-120. **2006**.


α
α
α
θ_d
θ_d
θ_d
$Z_{d,n}$
$Z_{d,n}$
$Z_{d,n}$
$W_{d,n}$
$W_{d,n}$
$W_{d,n}$
N
N
N
D
D
D
$\beta_{k,1}$
$\beta_{k,2}$
$\beta_{k,T}$
K
Topics drift through time

Intelligent Agents
Probabilistic Models for Sequential Structures

Prof. Dr. Ralf Möller
Universität zu Lübeck
Institut für Informationssystem

LDA



- LDA is a simple topic model.
- It can be used to find topics that describe a corpus.
- Each document exhibits multiple topics.
- How can we build on this simple model of text?

Using and Embedding LDA

- LDA model used to infer posterior distribution
  - $P(Z \mid wd)$
- Based on Z one can find and rank related documents
  - Infer $P(Relevant_d \mid Z_q)$ for d being the documents in a repository and q being the query document
  - Previously introduced models for information retrieval can be extended with topic information
  - Works for books, articles, images, videos, and other media
- LDA can be embedded in more complicated models
  - Model further intuitions about the structure of texts
    - Links, citations ("relational" topic models), …

LDA

- Traditional topic modeling (e.g., LDA):
  - Interested in *meaning*
  - Remove most syntactic words (e.g., stopwords)
  - Discard much of the structure, and all order information that the original author intended
  - Concerned about long-range topic dependencies rather document structure

- Not always easy to decide which words to remove
  - Keep only nouns? Example: saw vs. saw

HMM

- For natural language text: POS tagging
  - The standardized nature of grammar means that it stays fairly constant across different contexts

- HMMs are useful for segmenting text documents into different classes of words, regardless of meaning
  - ✓ For example, all nouns can be grouped together because they play the same role in different passages/documents.
  - ✗ Syntactic dependencies last at most for a sentence

Combining Syntax and Semantics: HMM-LDA

- All words (both syntactic and semantic) exhibit short range dependencies.
- Only content (semantic) words exhibit long range semantic dependencies.
- This leads to the HMM-LDA
- HMM-LDA is a composite model, in which an HMM decides the parts of speech, and a topic model (LDA) extracts topics from only those words which are deemed semantic

Thomas L. Griffiths, Mark Steyvers, David M. Blei, and Joshua B. Tenenbaum. Integrating topics and syntax. In *Proc. of* NIPS'04, pp. 537-544. **2004**.

Generative Process 1

Definitions

Words $\mathbf{w} = \{w_1, \ldots, w_n\}$ form document d where each word w_i is one of **W** words

Topic assignments $\mathbf{z} = \{z_1, \ldots z_n\}$ for each word, where each z_i taking one of **T** topics

Class assignments $\mathbf{c} = \{c_1, \ldots, c_n\}$ for each word, where each c_i taking one of **C** word classes

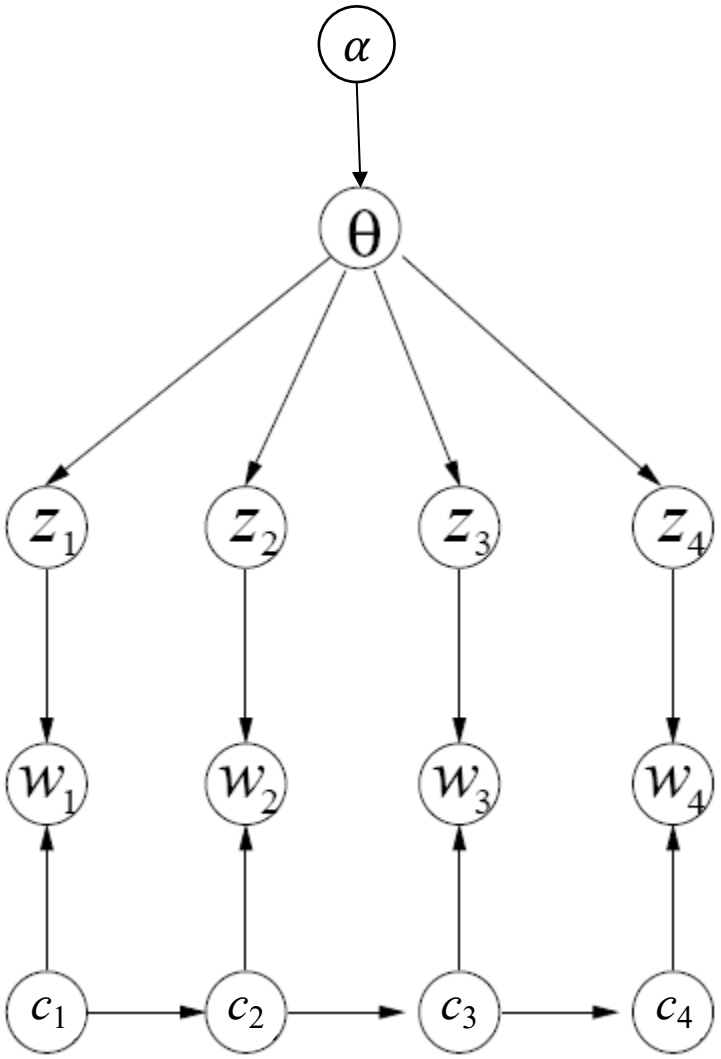
$\theta^{(d)}$ Multinomial distribution over topics for document *d*

$\phi^{(z)}$ Multinomial distribution over ***semantic*** words for topic indicated by *z*.

$\phi^{(c)}$ Multinomial distribution over ***non-semantic*** words for class indicated by *class c*.

$\pi^{(c_{i-1})}$ Transition probability from c_{i-1} to c_i

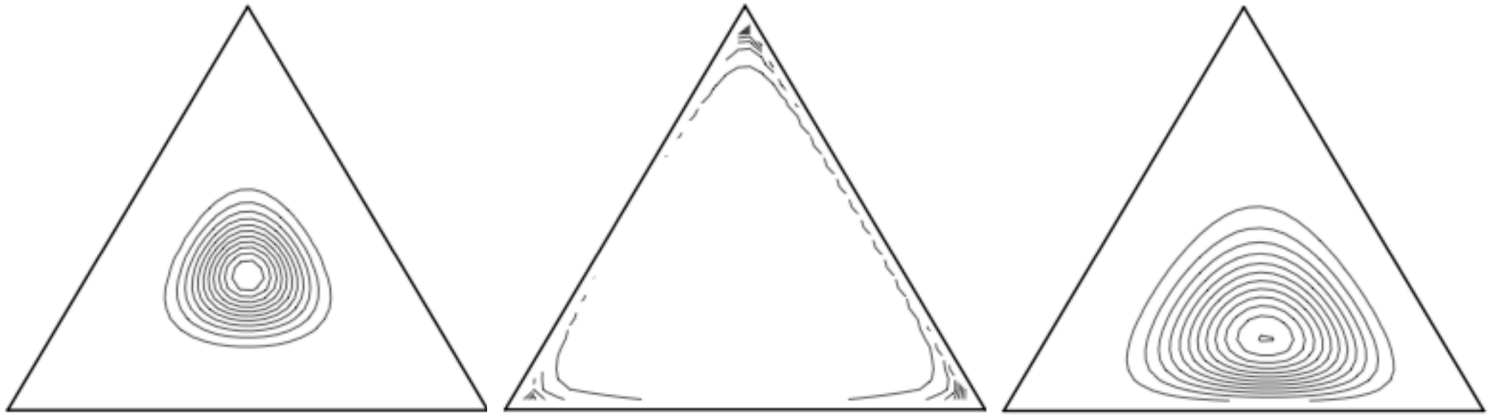
How to Incorporate Context into LDA?



LDA

HMM (e.g. for POS tagging)

Dirichlet Distribution



- The Dirichlet is a distribution on the simplex, positive vectors that sum to 1.
- It assumes that components are nearly independent.
- In real data, an article about *fossil fuels* is more likely to also be about *geology* than about *genetics*.

Generative Process 2

$\theta^{(d)} \sim \text{Dirichlet}(\alpha)$

$\phi^{(z)} \sim \text{Dirichlet}(\beta)$

$\pi^{(c)} \sim \text{Dirichlet}(\gamma)$ Where $\pi^{(c)}$ is the row of the transition matrix indicated by c.

$\phi^{(c)} \sim \text{Dirichlet}(\delta)$

For document *d*

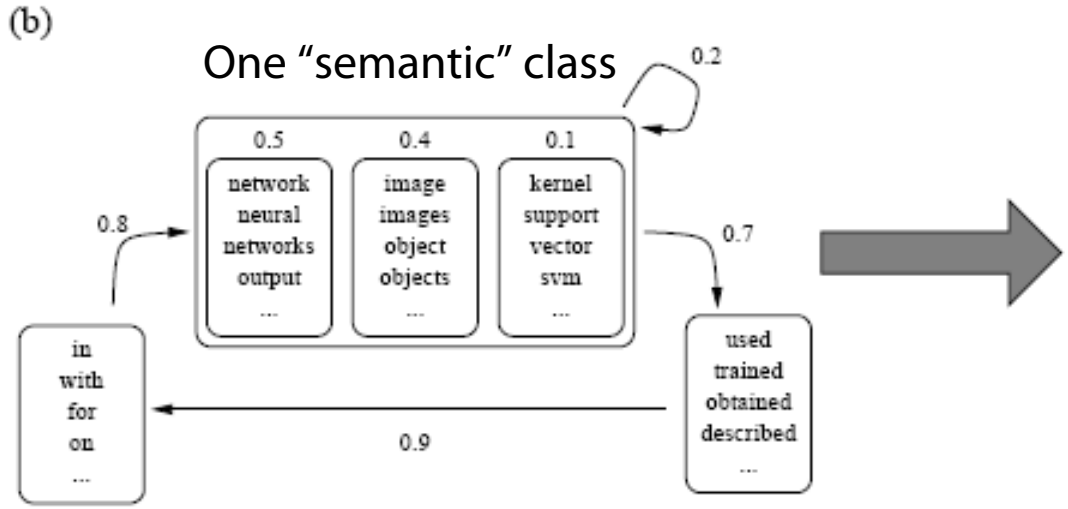
Draw topic distribution →

1. Sample $\theta^{(d)}$ from a Dirichlet(α) prior
2. For each word w_i in document d
   - (a) Draw z_i from $\theta^{(d)}$ ← Draw a topic for word i
   - (b) Draw c_i from $\pi^{(c_{i-1})}$ ← Draw a class for word i from transition matrix
   - (c) If $c_i = 1$, then draw w_i from $\phi^{(z_i)}$, else draw w_i from $\phi^{(c_i)}$

Semantic class

Draw a semantic word **OR** Draw a syntactic word

Simplified Example



network used for images

image obtained with kernel

output described with objects

neural network trained with svm images

Preposition class

Verb class

LDA-HMM: Summary

- HMM-LDA is a composite topic model
  - Long range semantic dependencies
  - Short-range syntactic dependencies

- Quite competitive with traditional HMM POS tagger
- Outperforms LDA when stop-words and punctuation are not removed

Dynamic Topic Models

- In LDA the order of documents does not matter
- Not appropriate for sequential corpora (e.g., that span hundreds of years)
- Further, we may want to track how language changes over time
- Let the topics *drift* in a sequence.

1789

My fellow citizens: I stand here today humbled by the task before us, grateful for the trust you have bestowed, mindful of the sacrifices borne by our ancestors...

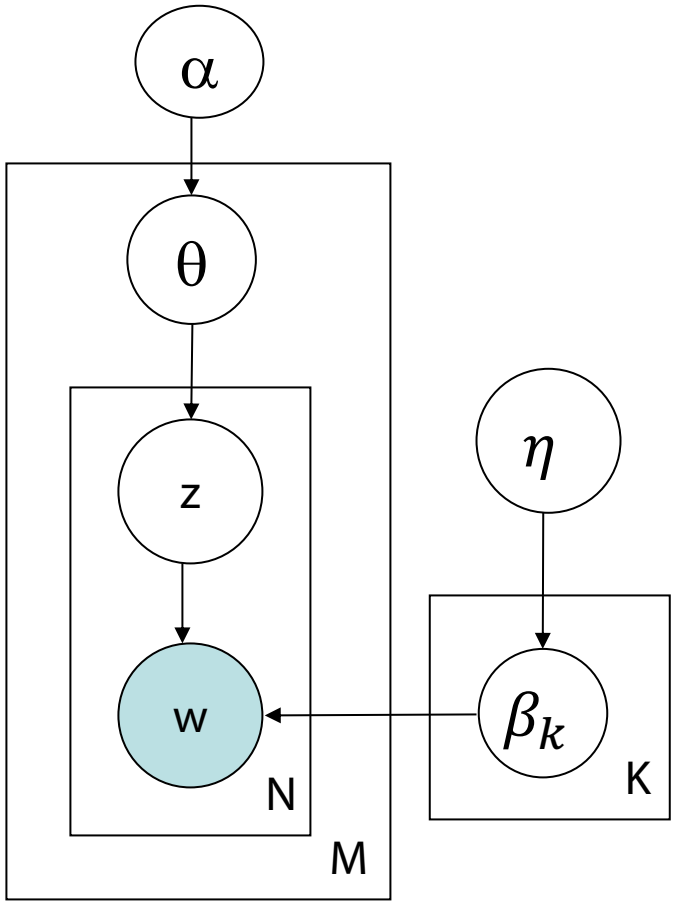
Inaugural addresses

2009

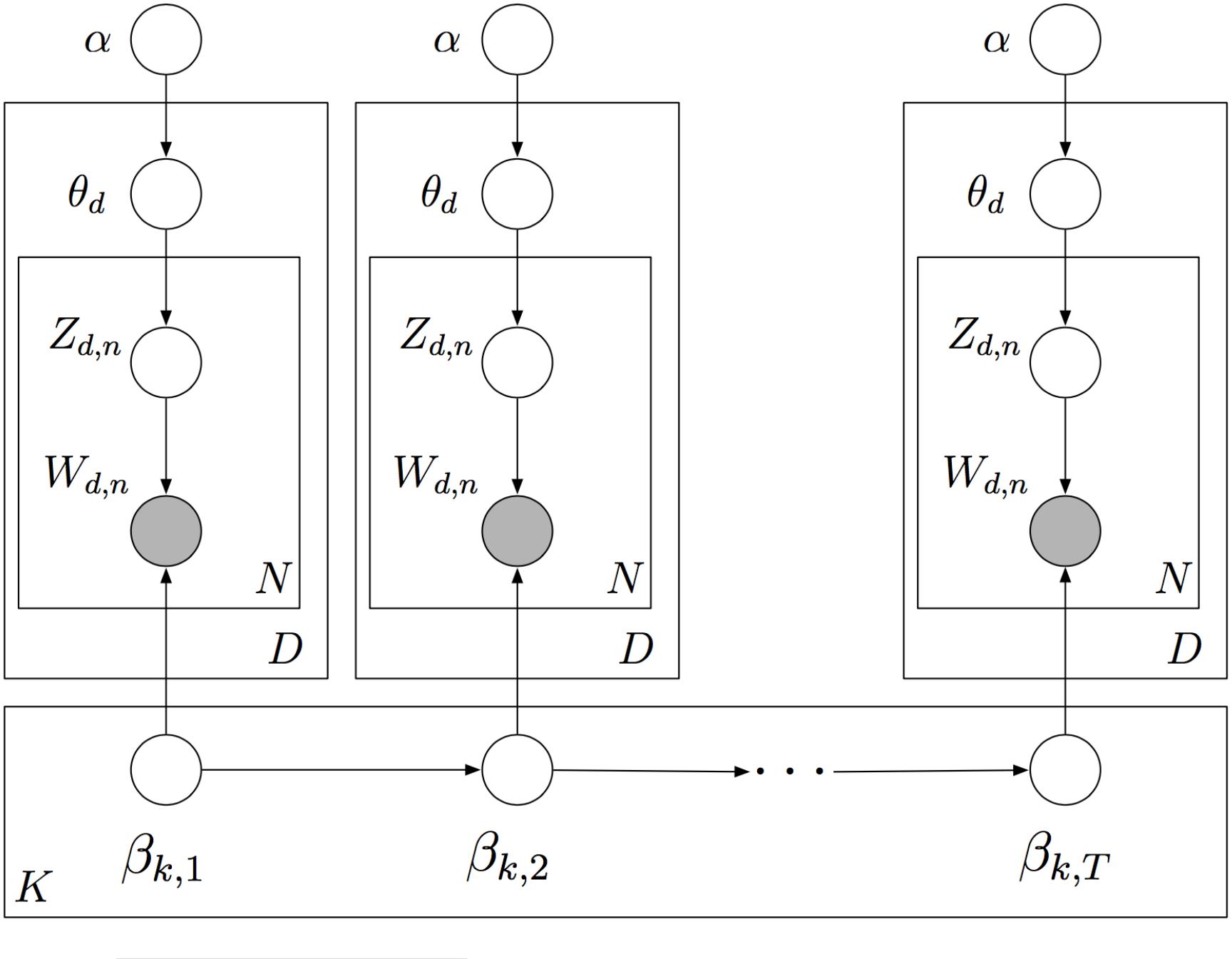
AMONG the vicissitudes incident to life no event could have filled me with greater anxieties than that of which the notification was transmitted by your order...

David M. Blei and John D. Lafferty. Dynamic topic models. In Proc. ICML '06. pp. 113-120. **2006**.

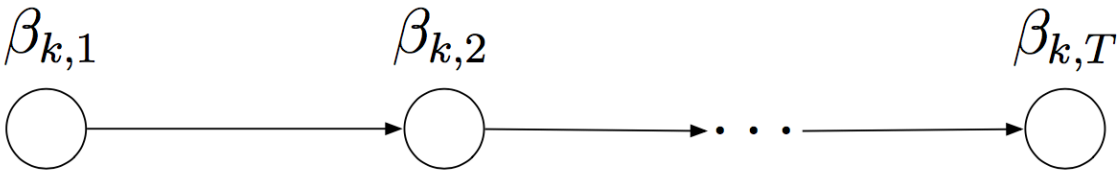
Recap: Smoothed LDA Model



- Give a different word distribution to each topic
  - β is $K \times V$ matrix (V vocabulary size), each row denotes word distribution of a topic
- For each document d
  - Choose $\theta_d \sim$ Dirichlet(α)
  - Choose $\beta_k \sim$ Dirichlet(η)
  - For each position $i = 1, \ldots, N_d$
    - Generate a topic $z_i \sim$ Mult($\cdot \mid \theta_d$)
    - Generate a word $w_i \sim$ Mult($\cdot \mid z_i, \beta_k$)


α
α
α
θ_d
θ_d
θ_d
$Z_{d,n}$
$Z_{d,n}$
$Z_{d,n}$
$W_{d,n}$
$W_{d,n}$
$W_{d,n}$
N
N
N
D
D
D
$\beta_{k,1}$
$\beta_{k,2}$
$\beta_{k,T}$
K
Topics drift through time

Dynamic Topic Models

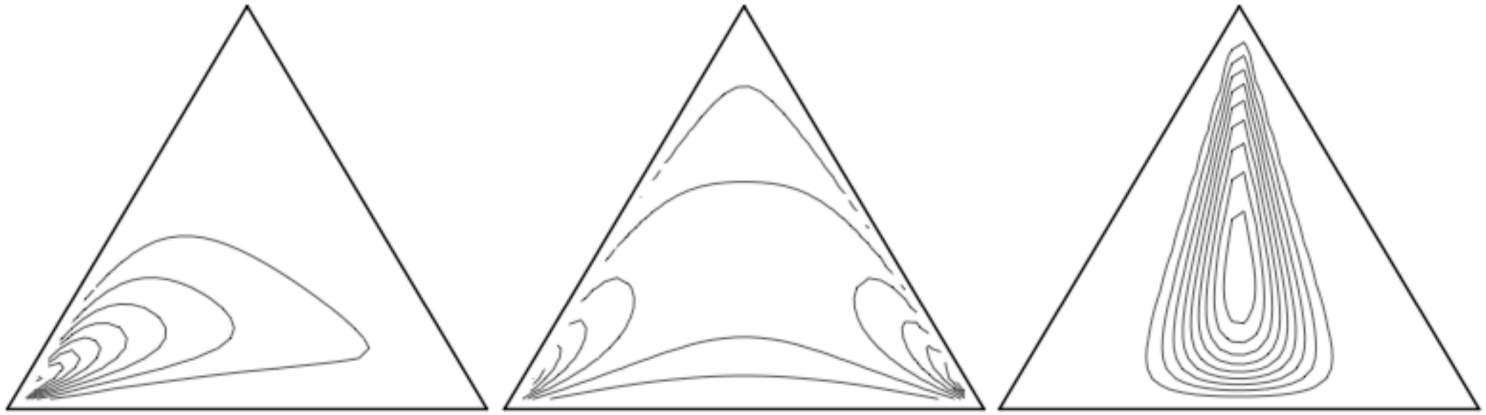


- Use a logit normal distribution to model topics evolving over time
- Embed it in a state-space model on the log of the topic distribution

$$\beta_{t,k} \mid \beta_{t-1,k} \sim \mathcal{N}(\beta_{t-1,k}, I\sigma^2)$$
$$p(w \mid \beta_{t,k}) \propto \exp\{\beta_{t,k}\}$$

- Lets us make inferences about sequences of documents

Logistic Normal Distribution



- The **logistic normal** is a distribution on the simplex that can model dependence between components (Aitchison, 1980).
- The log of the parameters of the multinomial are drawn from a multivariate Gaussian distribution,

$$\begin{aligned} X &\sim \mathcal{N}_K(\mu, \Sigma) \\ \theta_i &\propto \exp\{x_i\}. \end{aligned}$$

Logit Normal Distribution

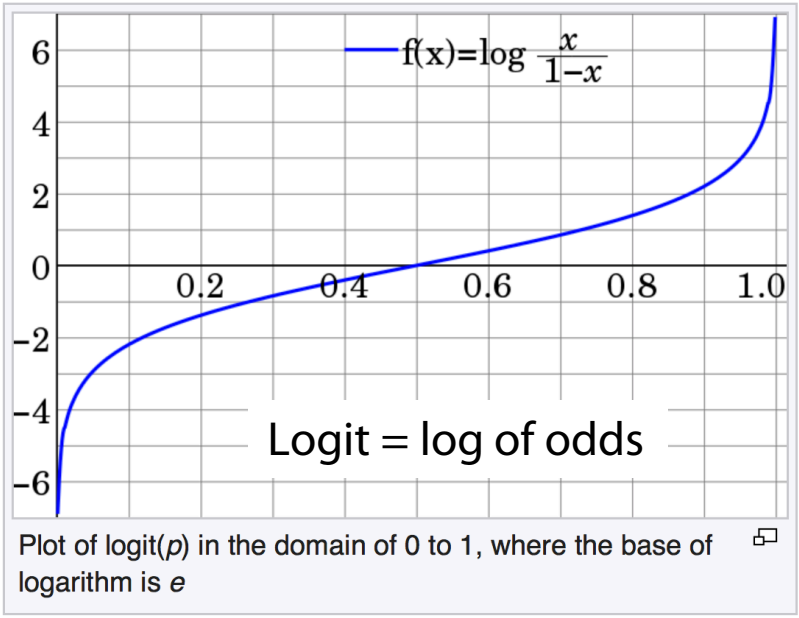
Normal Distribution

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

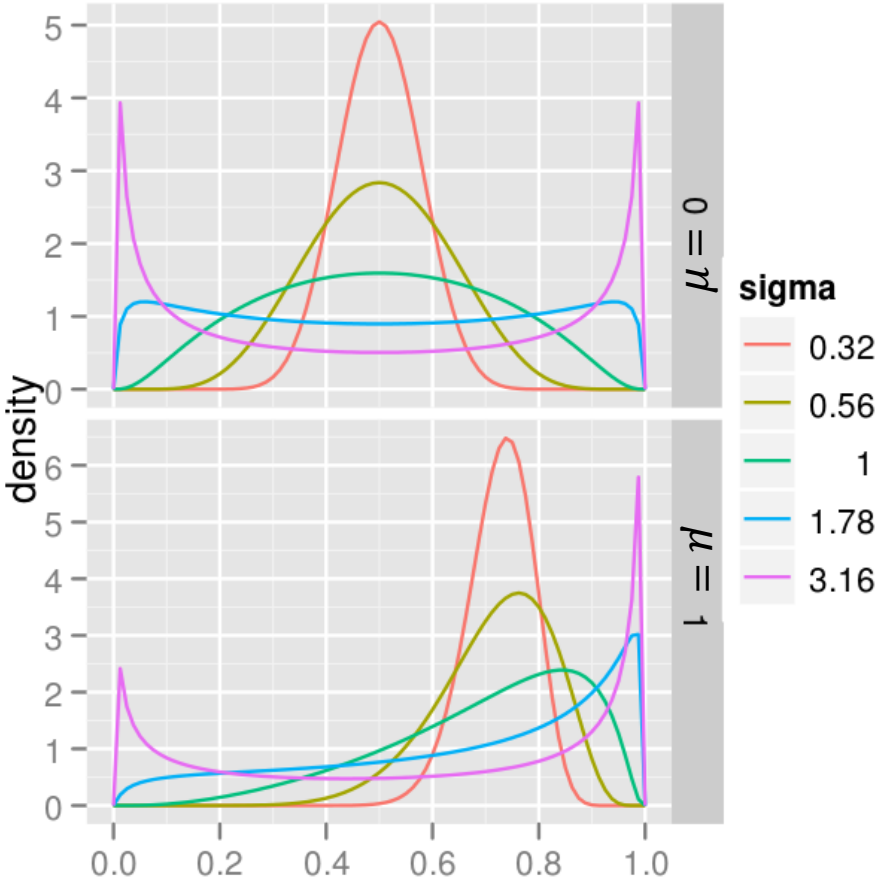
The probability density function (PDF) of a logit-normal distribution, for $0 \le x \le 1$, is:

$$f_X(x;\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} \frac{1}{x(1-x)} e^{-\frac{(\operatorname{logit}(x)-\mu)^2}{2\sigma^2}}$$

where μ and σ are the mean and standard deviation of the variable's logit (by definition, the variable's logit is normally distributed).



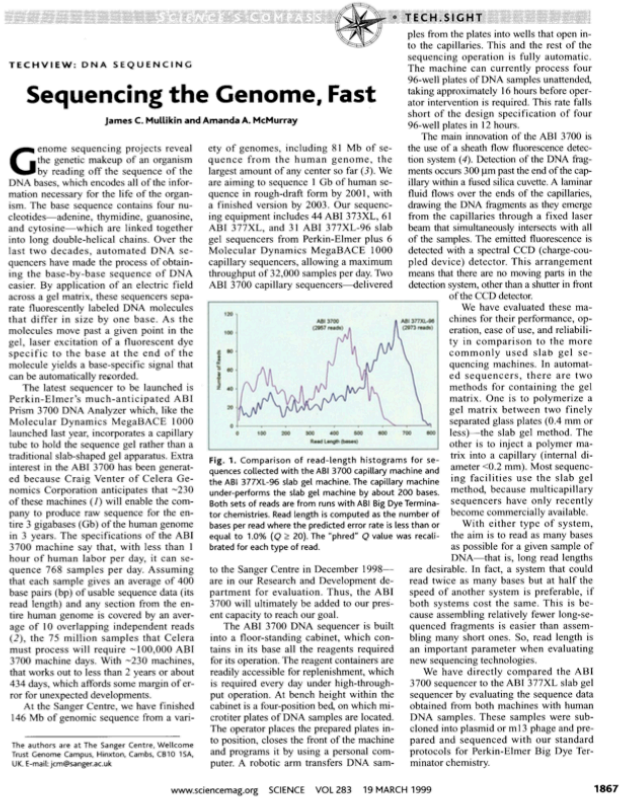
Plot of logit(p) in the domain of 0 to 1, where the base of logarithm is e



[Wikipedia]

Dynamic Topic Models

Original article



SCIENCE'S COMPASS • TECH.SIGHT

TECHVIEW: DNA SEQUENCING

Sequencing the Genome, Fast

James C. Mullikin and Amanda A. McMurray

Genome sequencing projects reveal the genetic makeup of an organism by reading off the sequence of the DNA bases, which encodes all of the information necessary for the life of the organism. The base sequence contains four nucleotides—adenine, thymidine, guanosine, and cytosine—which are linked together into long double-helical chains. Over the last two decades, automated DNA sequencers have made the process of obtaining the base-by-base sequence of DNA easier. By application of an electric field across a gel matrix, these sequencers separate fluorescently labeled DNA molecules that differ in size by one base. As the molecules move past a given point in the gel, laser excitation of a fluorescent dye specific to the base at the end of the molecule yields a base-specific signal that can be automatically recorded.

The latest sequencer to be launched is Perkin-Elmer's much-anticipated ABI Prism 3700 DNA Analyzer which, like the Molecular Dynamics MegaBACE 1000 launched last year, incorporates a capillary tube to hold the sequence gel rather than a traditional slab-shaped gel apparatus. Extra interest in the ABI 3700 has been generated because Craig Venter of Celera Genomics Corporation anticipates that ~230 of these machines (*1*) will enable the company to produce raw sequence for the entire 3 gigabases (Gb) of the human genome in 3 years. The specifications of the ABI 3700 machine say that, with less than 1 hour of human labor per day, it can sequence 768 samples per day. Assuming that each sample gives an average of 400 base pairs (bp) of usable sequence data (its read length) and any section from the entire human genome is covered by an average of 10 overlapping independent reads (*2*), the 75 million samples that Celera must process will require ~100,000 ABI 3700 machine days. With ~230 machines, that works out to less than 2 years or about 434 days, which affords some margin of error for unexpected developments.

At the Sanger Centre, we have finished 146 Mb of genomic sequence from a variety of genomes, including 81 Mb of sequence from the human genome, the largest amount of any center so far (*3*). We are aiming to sequence 1 Gb of human sequence in rough-draft form by 2001, with a finished version by 2003. Our sequencing equipment includes 44 ABI 373XL, 61 ABI 377XL, and 31 ABI 377XL-96 slab gel sequencers from Perkin-Elmer plus 6 Molecular Dynamics MegaBACE 1000 capillary sequencers, allowing a maximum throughput of 32,000 samples per day. Two ABI 3700 capillary sequencers—delivered to the Sanger Centre in December 1998—are in our Research and Development department for evaluation. Thus, the ABI 3700 will ultimately be added to our present capacity to reach our goal.

Fig. 1. Comparison of read-length histograms for sequences collected with the ABI 3700 capillary machine and the ABI 377XL-96 slab gel machine. The capillary machine underperforms the slab gel machine by about 200 bases. Both sets of reads are from runs with ABI Big Dye Terminator chemistries. Read length is computed as the number of bases per read where the predicted error rate is less than or equal to 1.0% ($Q \geq 20$). The "phred" Q value was recalibrated for each type of read.

The ABI 3700 DNA sequencer is built into a floor-standing cabinet, which contains in its base all the reagents required for its operation. The reagent containers are readily accessible for replenishment, which is required every day under high-throughput operation. At bench height within the cabinet is a four-position bed, on which microtiter plates of DNA samples are located. The operator places the prepared plates into position, closes the front of the machine and programs it by using a personal computer. A robotic arm transfers DNA samples from the plates into wells that open into the capillaries. This and the rest of the sequencing operation is fully automatic. The machine can currently process four 96-well plates of DNA samples unattended, taking approximately 16 hours before operator intervention is required. This rate falls short of the design specification of four 96-well plates in 12 hours.

The main innovation of the ABI 3700 is the use of a sheath flow fluorescence detection system (*4*). Detection of the DNA fragments occurs 300 μm past the end of the capillary within a fused silica cuvette. A laminar fluid flows over the ends of the capillaries, drawing the DNA fragments as they emerge from the capillaries through a fixed laser beam that simultaneously intersects with all of the samples. The emitted fluorescence is detected with a spectral CCD (charge-coupled device) detector. This arrangement means that there are no moving parts in the detection system, other than a shutter in front of the CCD detector.

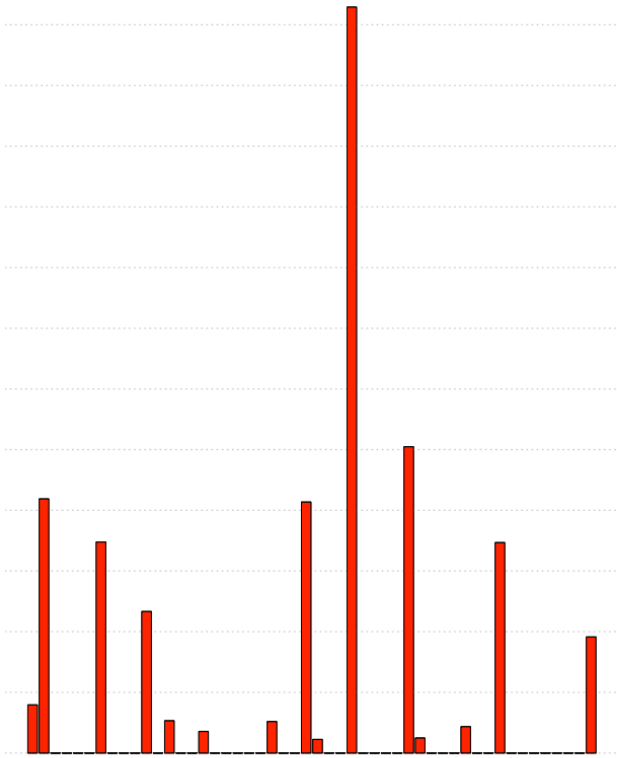
We have evaluated these machines for their performance, operation, ease of use, and reliability in comparison to the more commonly used slab gel sequencing machines. In automated sequencers, there are two methods for containing the gel matrix. One is to polymerize a gel matrix between two finely separated glass plates (0.4 mm or less)—the slab gel method. The other is to inject a polymer matrix into a capillary (internal diameter <0.2 mm). Most sequencing facilities use the slab gel method, because multicapillary sequencers have only recently become commercially available.

With either type of system, the aim is to read as many bases as possible for a given sample of DNA—that is, long read lengths are desirable. In fact, a system that could read twice as many bases but at half the speed of another system is preferable, if both systems cost the same. This is because assembling relatively fewer long-sequenced fragments is easier than assembling many short ones. So, read length is an important parameter when evaluating new sequencing technologies.

We have directly compared the ABI 3700 sequencer to the ABI 377XL slab gel sequencer by evaluating the sequence data obtained from both machines with human DNA samples. These samples were subcloned into plasmid or m13 phage and prepared and sequenced with our standard protocols for Perkin-Elmer Big Dye Terminator chemistry.

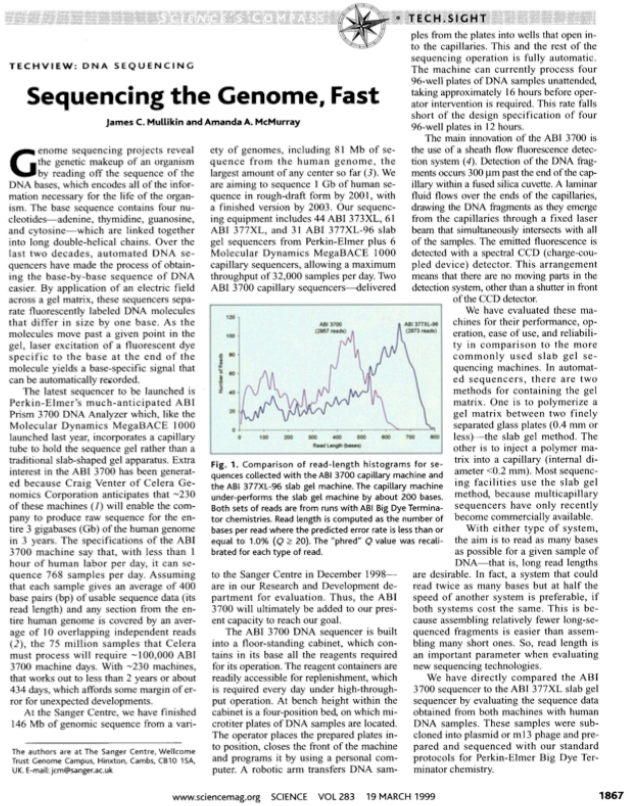
The authors are at The Sanger Centre, Wellcome Trust Genome Campus, Hinxton, Cambs, CB10 1SA, UK. E-mail: jcm@sanger.ac.uk

Topic proportions



Dynamic Topic Models

Original article



SCIENCE'S COMPASS

TECH.SIGHT

TECHVIEW: DNA SEQUENCING

Sequencing the Genome, Fast

James C. Mullikin and Amanda A. McMurray

Genome sequencing projects reveal the genetic makeup of an organism by reading off the sequence of the DNA bases, which encodes all of the information necessary for the life of the organism. The base sequence contains four nucleotides—adenine, thymidine, guanosine, and cytosine—which are linked together into long double-helical chains. Over the last two decades, automated DNA sequencers have made the process of obtaining the base-by-base sequence of DNA easier. By application of an electric field across a gel matrix, these sequencers separate fluorescently labeled DNA molecules that differ in size by one base. As the molecules move past a given point in the gel, laser excitation of a fluorescent dye specific to the base at the end of the molecule yields a base-specific signal that can be automatically recorded.

The latest sequencer to be launched is Perkin-Elmer's much-anticipated ABI Prism 3700 DNA Analyzer which, like the Molecular Dynamics MegaBACE 1000 launched last year, incorporates a capillary tube to hold the sequence gel rather than a traditional slab-shaped gel apparatus. Extra interest in the ABI 3700 has been generated because Craig Venter of Celera Genomics Corporation anticipates that ~230 of these machines (*1*) will enable the company to produce raw sequence for the entire 3 gigabases (Gb) of the human genome in 3 years. The specifications of the ABI 3700 machine say that, with less than 1 hour of human labor per day, it can sequence 768 samples per day. Assuming that each sample gives an average of 400 base pairs (bp) of usable sequence data (its read length) and any section from the entire human genome is covered by an average of 10 overlapping independent reads (*2*), the 75 million samples that Celera must process will require ~100,000 ABI 3700 machine days. With ~230 machines, that works out to less than 2 years or about 434 days, which affords some margin of error for unexpected developments.

At the Sanger Centre, we have finished 146 Mb of genomic sequence from a variety of genomes, including 81 Mb of sequence from the human genome, the largest amount of any center so far (*3*). We are aiming to sequence 1 Gb of human sequence in rough-draft form by 2001, with a finished version by 2003. Our sequencing equipment includes 44 ABI 373XL, 61 ABI 377XL, and 31 ABI 377XL-96 slab gel sequencers from Perkin-Elmer plus 6 Molecular Dynamics MegaBACE 1000 capillary sequencers, allowing a maximum throughput of 32,000 samples per day. Two ABI 3700 capillary sequencers—delivered to the Sanger Centre in December 1998—are in our Research and Development department for evaluation. Thus, the ABI 3700 will ultimately be added to our present capacity to reach our goal.

Fig. 1. Comparison of read-length histograms for sequences collected with the ABI 3700 capillary machine and the ABI 377XL-96 slab gel machine. The capillary machine under-performs the slab gel machine by about 200 bases. Both sets of reads are from runs with ABI Big Dye Terminator chemistries. Read length is computed as the number of bases per read where the predicted error rate is less than or equal to 1.0% ($Q \geq 20$). The "phred" Q value was recalibrated for each type of read.

The ABI 3700 DNA sequencer is built into a floor-standing cabinet, which contains in its base all the reagents required for its operation. The reagent containers are readily accessible for replenishment, which is required every day under high-throughput operation. At bench height within the cabinet is a four-position bed, on which microtiter plates of DNA samples are located. The operator places the prepared plates into position, closes the front of the machine and programs it by using a personal computer. A robotic arm transfers DNA samples from the plates into wells that open into to the capillaries. This and the rest of the sequencing operation is fully automatic. The machine can currently process four 96-well plates of DNA samples unattended, taking approximately 16 hours before operator intervention is required. This rate falls short of the design specification of four 96-well plates in 12 hours.

The main innovation of the ABI 3700 is the use of a sheath flow fluorescence detection system (*4*). Detection of the DNA fragments occurs 300 μm past the end of the capillary within a fused silica cuvette. A laminar fluid flows over the ends of the capillaries, drawing the DNA fragments as they emerge from the capillaries through a fixed laser beam that simultaneously intersects with all of the samples. The emitted fluorescence is detected with a spectral CCD (charge-coupled device) detector. This arrangement means that there are no moving parts in the detection system, other than a shutter in front of the CCD detector.

We have evaluated these machines for their performance, operation, ease of use, and reliability in comparison to the more commonly used slab gel sequencing machines. In automated sequencers, there are two methods for containing the gel matrix. One is to polymerize a gel matrix between two finely separated glass plates (0.4 mm or less)—the slab gel method. The other is to inject a polymer matrix into a capillary (internal diameter ~0.2 mm). Most sequencing facilities use the slab gel method, because multicapillary sequencers have only recently become commercially available.

With either type of system, the aim is to read as many bases as possible for a given sample of DNA—that is, long read lengths are desirable. In fact, a system that could read twice as many bases but at half the speed of another system is preferable, if both systems cost the same. This is because assembling relatively fewer long-sequenced fragments is easier than assembling many short ones. So, read length is an important parameter when evaluating new sequencing technologies.

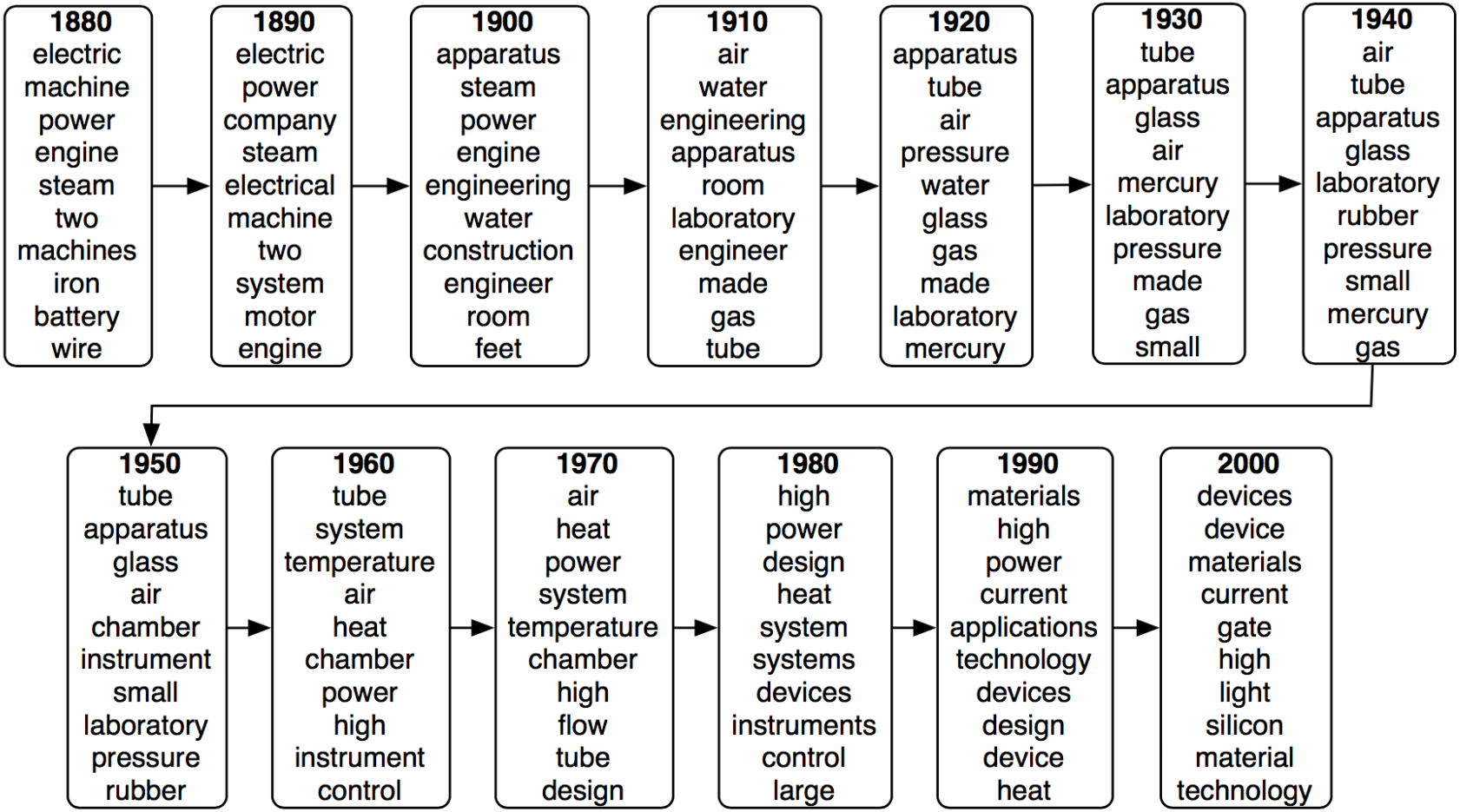
We have directly compared the ABI 3700 sequencer to the ABI 377XL slab gel sequencer by evaluating the sequence data obtained from both machines with human DNA samples. These samples were subcloned into plasmid or m13 phage and prepared and sequenced with our standard protocols for Perkin-Elmer Big Dye Terminator chemistry.

The authors are at The Sanger Centre, Wellcome Trust Genome Campus, Hinxton, Cambs, CB10 1SA, UK. E-mail: jcm@sanger.ac.uk

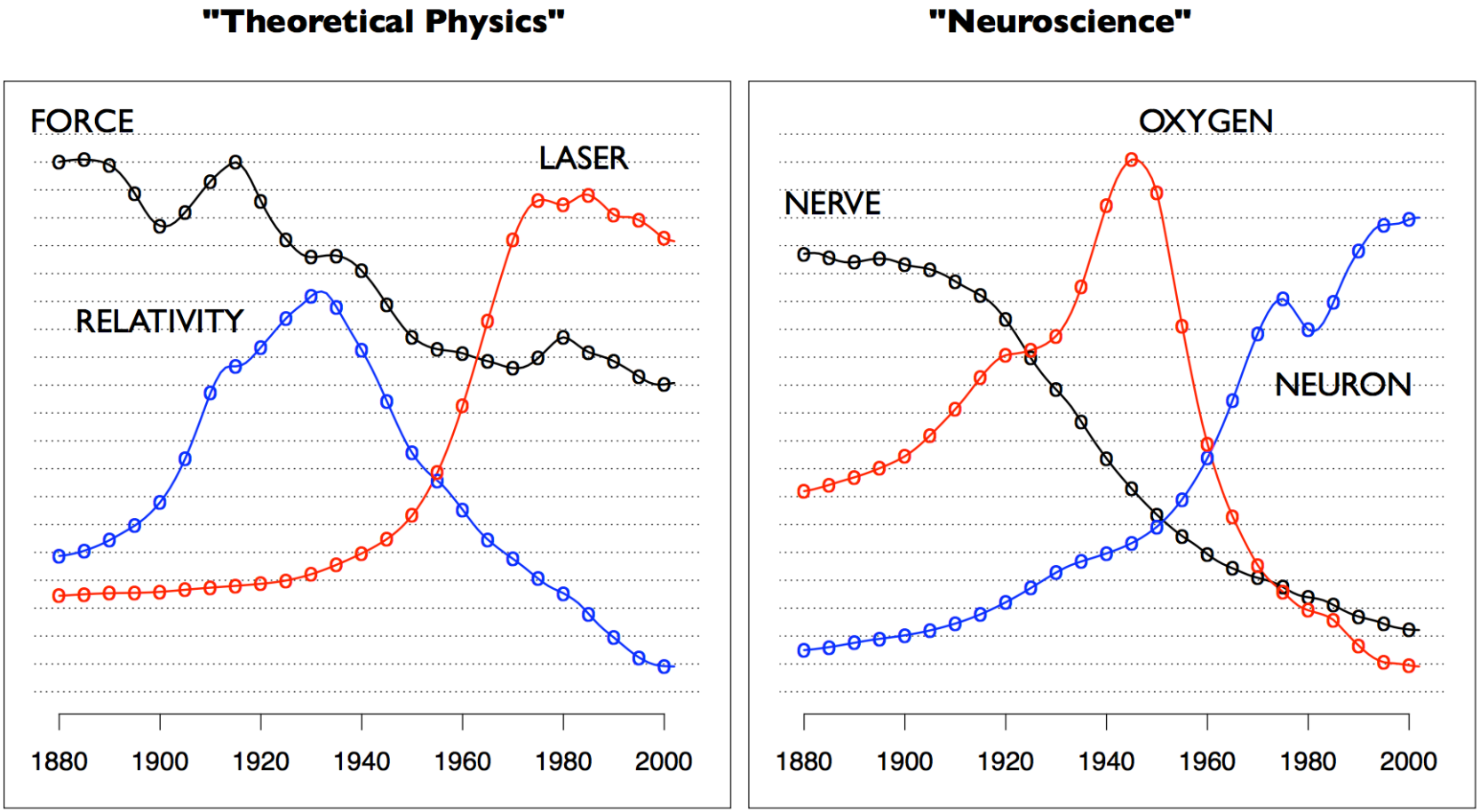
Most likely words from top topics

sequence	devices	data
genome	device	information
genes	materials	network
sequences	current	web
human	high	computer
gene	gate	language
dna	light	networks
sequencing	silicon	time
chromosome	material	software
regions	technology	system
analysis	electrical	words
data	fiber	algorithm
genomic	power	number
number	based	internet

Dynamic Topic Models



Dynamic Topic Models



David M. Blei and John D. Lafferty. Dynamic topic models. In Proc. ICML '06. pp. 113-120. **2006**.

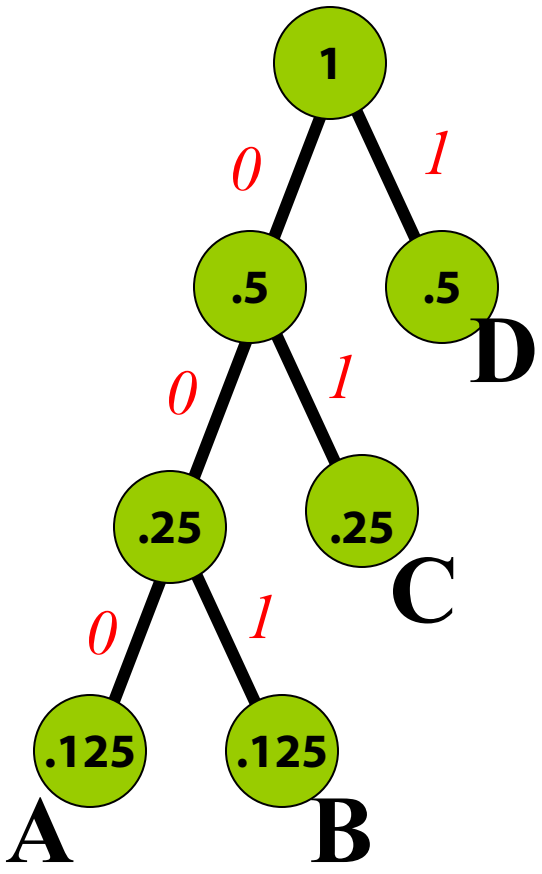
Dynamic Topic Models

- Understand developments
- Distributions of topics over time
- Discretization of time might be a problem
  - Runtime increases dramatically
  - Continuous dynamic topic models

- Many applications
  - E.g., comparison of science areas, analysis of scientific work
- How can we compare distributions?

Wang, Chong; Blei, David; Heckerman, David. "Continuous Time Dynamic Topic Models". *Proceedings of ICML*'08, **2008**.

Recap: Huffman code example



M	code	length	prob	Exp. len
A	000	3	0,125	0,375
B	001	3	0,125	0,375
C	01	2	0,250	0,500
D	1	1	0,500	0,500
average message length				1,750

If we need to send many messages (A,B,C or D) and they have this probability distribution and we use this code, then over time, the average bits/message should approach 1.75

Recap: Information Theory Background

- Assume that you need to send messages from a repertoire of n messages
- If there are n equally probable possible messages, then the probability p of each is 1/n or n = 1/p
- Information (number of bits) conveyed by a message is log(n) = log(1/p)= -log(p)
- E.g., if there are 16 messages, then log(16) = 4 and we need 4 bits to identify/send each message.
- In general, if we are given a probability distribution

  $P = (p_1, p_2, .., p_n)$
- Information conveyed by distribution (aka entropy of P) is:

  $I(P) = -(p_1*\log(p_1) + p_2*\log(p_2) + .. + p_n*\log(p_n))$

  $= -\sum_i p_i*\log(p_i) = \sum_i p_i*\log(1/p_i)$

The KL Divergence

- The *cross-entropy*, or *Kullback-Leibler divergence*, between two distributions $\mathbf{p}$ and $\mathbf{q}$ measures the *expected information gain* (reduction in average number of bits per event) due to replacing the "wrong" distribution $\mathbf{q}$ with the "right" distribution $\mathbf{p}$:

$$D^{KL}(\mathbf{p}, \mathbf{q}) \equiv \sum_i p_i(\ln(1/q_i) - \ln(1/p_i)) = \mathbf{E}_{\mathbf{p}}[\ln(\mathbf{p}/\mathbf{q})]$$

- Not symmetric

Hellinger Distance

- The *Hellinger distance* is a symmetric measure of distance between two distributions that is popular in machine learning applications:

$$D^{HEL}(\mathbf{p}, \mathbf{q}) \equiv \|\sqrt{\mathbf{p}} - \sqrt{\mathbf{q}}\|_2 = \left(\sum_{j=1}^{n} \left(\sqrt{p_j} - \sqrt{q_j} \right)^2 \right)^{1/2}$$

$$\in [0, \sqrt{2}]$$

- Sometimes value should be in [0, 1]

For two discrete probability distributions $P = (p_1, \ldots, p_k)$ and $Q = (q_1, \ldots, q_k)$, their Hellinger distance is defined as

$$H(P, Q) = \frac{1}{\sqrt{2}} \sqrt{\sum_{i=1}^{k} (\sqrt{p_i} - \sqrt{q_i})^2},$$

[Wikipedia]

Dynamic Topic Models

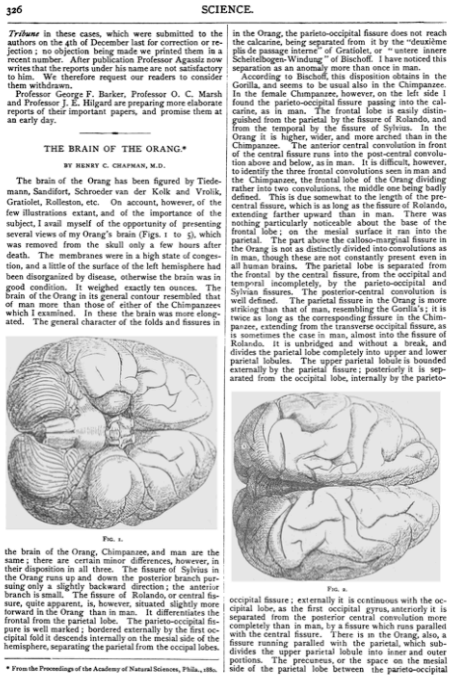
- **Time-corrected similarity** shows a new way of using the posterior
- Consider the expected Hellinger distance between the topic proportions of two documents,

$$d_{ij} = \mathrm{E}\left[\sum_{k=1}^{K}(\sqrt{\theta_{i,k}} - \sqrt{\theta_{j,k}})^2 \,|\, \mathbf{w}_i, \mathbf{w}_j\right]$$

- Uses the latent structure to define similarity
- Time has been factored out because the topics associated to the components are different from year to year
- Similarity of documents based only on topic proportions

Dynamic Topic Models

The Brain of the Orang (1880)



326 SCIENCE.

Tribune in these cases, which were submitted to the authors on the 4th of December last for correction or rejection; no objection being made we printed them in a recent number. After publication Professor Agassiz now writes that the reports under his name are not satisfactory to him. We therefore request our readers to consider them withdrawn.

Professor George F. Barker, Professor O. C. Marsh and Professor J. E. Hilgard are preparing more elaborate reports of their important papers, and promise them at an early day.

THE BRAIN OF THE ORANG.*

BY HENRY C. CHAPMAN, M.D.

The brain of the Orang has been figured by Tiedemann, Sandifort, Schroeder van der Kolk and Vrolik, Gratiolet, Rolleston, etc. On account, however, of the few illustrations extant, and of the importance of the subject, I avail myself of the opportunity of presenting several views of my Orang's brain (Figs. 1 to 5), which was removed from the skull only a few hours after death. The membranes were in a high state of congestion, and a little of the surface of the left hemisphere had been disorganized by disease, otherwise the brain was in good condition. It weighed exactly ten ounces. The brain of the Orang in its general contour resembled that of man more than those of either of the Chimpanzees which I examined. In these the brain was more elongated. The general character of the folds and fissures in

FIG. 1.

the brain of the Orang, Chimpanzee, and man are the same; there are certain minor differences, however, in their disposition in all three. The fissure of Sylvius in the Orang runs up and down the posterior branch pursuing only a slightly backward direction; the anterior branch is small. The fissure of Rolando, or central fissure, quite apparent, is, however, situated slightly more forward in the Orang than in man. It differentiates the frontal from the parietal lobe. The parieto-occipital fissure is well marked; bordered externally by the first occipital fold it descends internally on the mesial side of the hemisphere, separating the parietal from the occipal lobes.

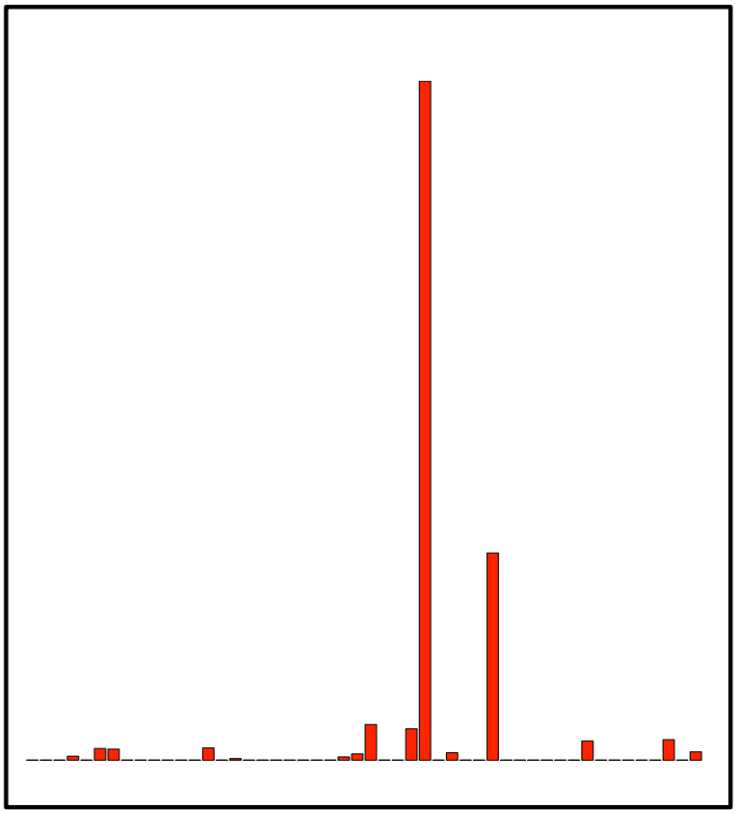
* From the Proceedings of the Academy of Natural Sciences, Phila., 1880.

in the Orang, the parieto-occipital fissure does not reach the calcarine, being separated from it by the "deuxième plis de passage interne" of Gratiolet, or "untere innere Scheitelbogen-Windung" of Bischoff. I have noticed this separation as an anomaly more than once in man.

According to Bischoff, this disposition obtains in the Gorilla, and seems to be usual also in the Chimpanzee. In the female Chimpanzee, however, on the left side I found the parieto-occipital fissure passing into the calcarine, as in man. The frontal lobe is easily distinguished from the parietal by the fissure of Rolando, and from the temporal by the fissure of Sylvius. In the Orang it is higher, wider, and more arched than in the Chimpanzee. The anterior central convolution in front of the central fissure runs into the post-central convolution above and below, as in man. It is difficult, however, to identify the three frontal convolutions seen in man and the Chimpanzee, the frontal lobe of the Orang dividing rather into two convolutions, the middle one being badly defined. This is due somewhat to the length of the pre-central fissure, which is as long as the fissure of Rolando, extending farther upward than in man. There was nothing particularly noticeable about the base of the frontal lobe; on the mesial surface it ran into the parietal. The part above the calloso-marginal fissure in the Orang is not as distinctly divided into convolutions as in man, though these are not constantly present even in all human brains. The parietal lobe is separated from the frontal by the central fissure, from the occipital and temporal incompletely, by the parieto-occipital and Sylvian fissures. The posterior-central convolution is well defined. The parietal fissure in the Orang is more striking than that of man, resembling the Gorilla's; it is twice as long as the corresponding fissure in the Chimpanzee, extending from the transverse occipital fissure, as is sometimes the case in man, almost into the fissure of Rolando. It is unbridged and without a break, and divides the parietal lobe completely into upper and lower parietal lobules. The upper parietal lobule is bounded externally by the parietal fissure; posteriorly it is separated from the occipital lobe, internally by the parieto-

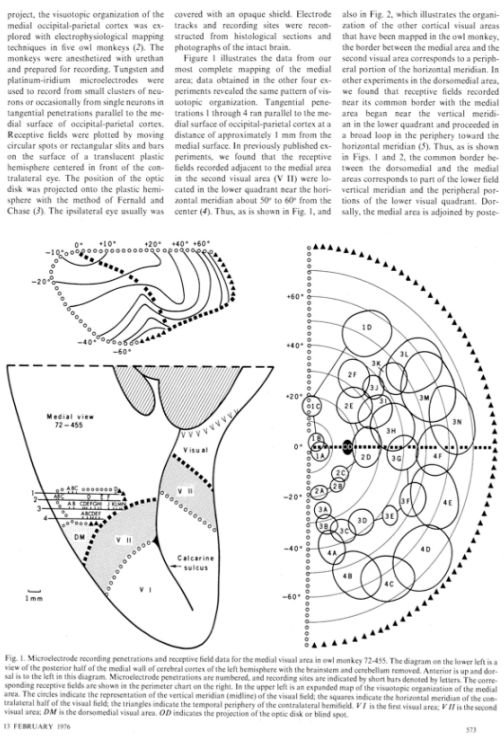
FIG. 2.

occipital fissure; externally it is continuous with the occipital lobe, as the first occipital gyrus, anteriorly it is separated from the posterior central convolution more completely than in man, by a fissure which runs parallel with the central fissure. There is in the Orang, also, a fissure running parallel with the parietal, which subdivides the upper parietal lobule into inner and outer portions. The precuneus, or the space on the mesial side of the parietal lobe between the parieto-occipital



Dynamic Topic Models

Representation of the Visual Field on the Medial Wall of Occipital-Parietal Cortex in the Owl Monkey (1976)



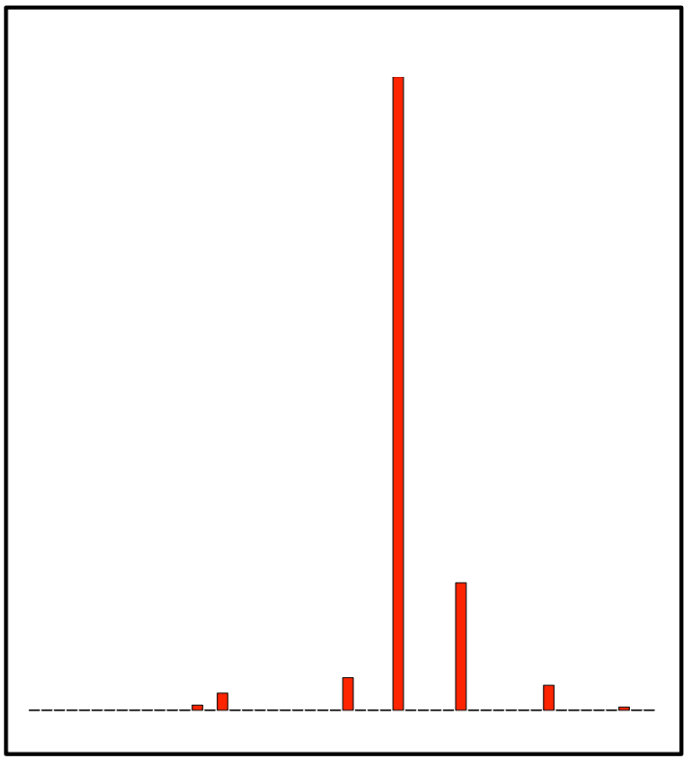
project, the visuotopic organization of the medial occipital-parietal cortex was explored with electrophysiological mapping techniques in five owl monkeys (*2*). The monkeys were anesthetized with urethan and prepared for recording. Tungsten and platinum-iridium microelectrodes were used to record from small clusters of neurons or occasionally from single neurons in tangential penetrations parallel to the medial surface of occipital-parietal cortex. Receptive fields were plotted by moving circular spots or rectangular slits and bars on the surface of a translucent plastic hemisphere centered in front of the contralateral eye. The position of the optic disk was projected onto the plastic hemisphere with the method of Fernald and Chase (*3*). The ipsilateral eye usually was covered with an opaque shield. Electrode tracks and recording sites were reconstructed from histological sections and photographs of the intact brain.

Figure 1 illustrates the data from our most complete mapping of the medial area; data obtained in the other four experiments revealed the same pattern of visuotopic organization. Tangential penetrations 1 through 4 ran parallel to the medial surface of occipital-parietal cortex at a distance of approximately 1 mm from the medial surface. In previously published experiments, we found that the receptive fields recorded adjacent to the medial area in the second visual area (*V II*) were located in the lower quadrant near the horizontal meridian about 50° to 60° from the center (*4*). Thus, as is shown in Fig. 1, and also in Fig. 2, which illustrates the organization of the other cortical visual areas that have been mapped in the owl monkey, the border between the medial area and the second visual area corresponds to a peripheral portion of the horizontal meridian. In other experiments in the dorsomedial area, we found that receptive fields recorded near its common border with the medial area began near the vertical meridian in the lower quadrant and proceeded in a broad loop in the periphery toward the horizontal meridian (*5*). Thus, as is shown in Figs. 1 and 2, the common border between the dorsomedial and the medial areas corresponds to part of the lower field vertical meridian and the peripheral portions of the lower visual quadrant. Dorsally, the medial area is adjoined by poste-

Fig. 1. Microelectrode recording penetrations and receptive field data for the medial visual area in owl monkey 72-455. The diagram on the lower left is a view of the posterior half of the medial wall of cerebral cortex of the left hemisphere with the brainstem and cerebellum removed. Anterior is up and dorsal is to the left in this diagram. Microelectrode penetrations are numbered, and recording sites are indicated by short bars denoted by letters. The corresponding receptive fields are shown in the perimeter chart on the right. In the upper left is an expanded map of the visuotopic organization of the medial area. The circles indicate the representation of the vertical meridian (midline) of the visual field; the squares indicate the horizontal meridian of the contralateral half of the visual field; the triangles indicate the temporal periphery of the contralateral hemifield. *V I* is the first visual area; *V II* is the second visual area; *DM* is the dorsomedial visual area. *OD* indicates the projection of the optic disk or blind spot.

13 FEBRUARY 1976

573



Dynamic Topic Models: Summary

- Can model changes of topics (= word distributions) in corpora over time
- Uses HMM as a technique for modeling temporal influences
- As a by-product we have discussed techniques for comparing distributions