Language Models

Hongning Wang

CS@UVa

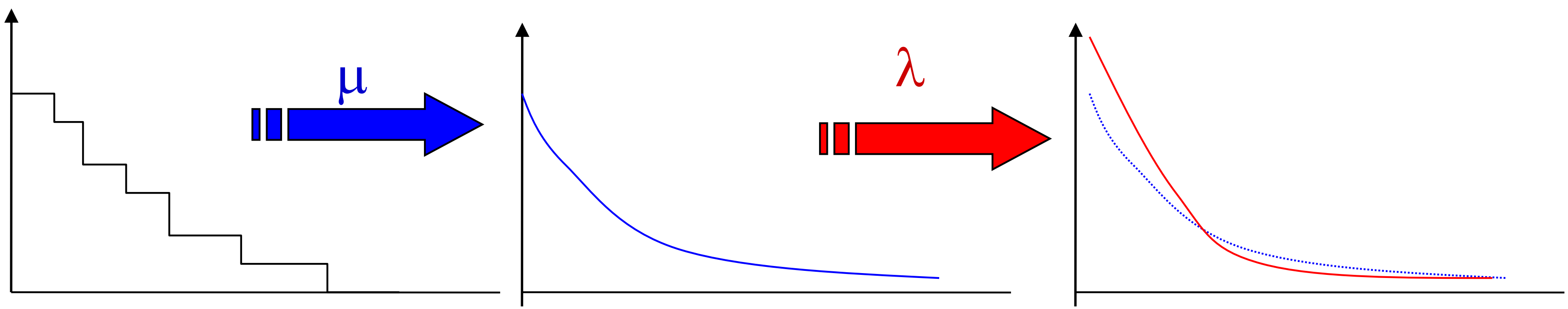
Two-stage smoothing [Zhai & Lafferty 02]

Stage-1

-Explain unseen words
-Dirichlet prior (Bayesian)

Stage-2

-Explain noise in query
-2-component mixture



Collection LM

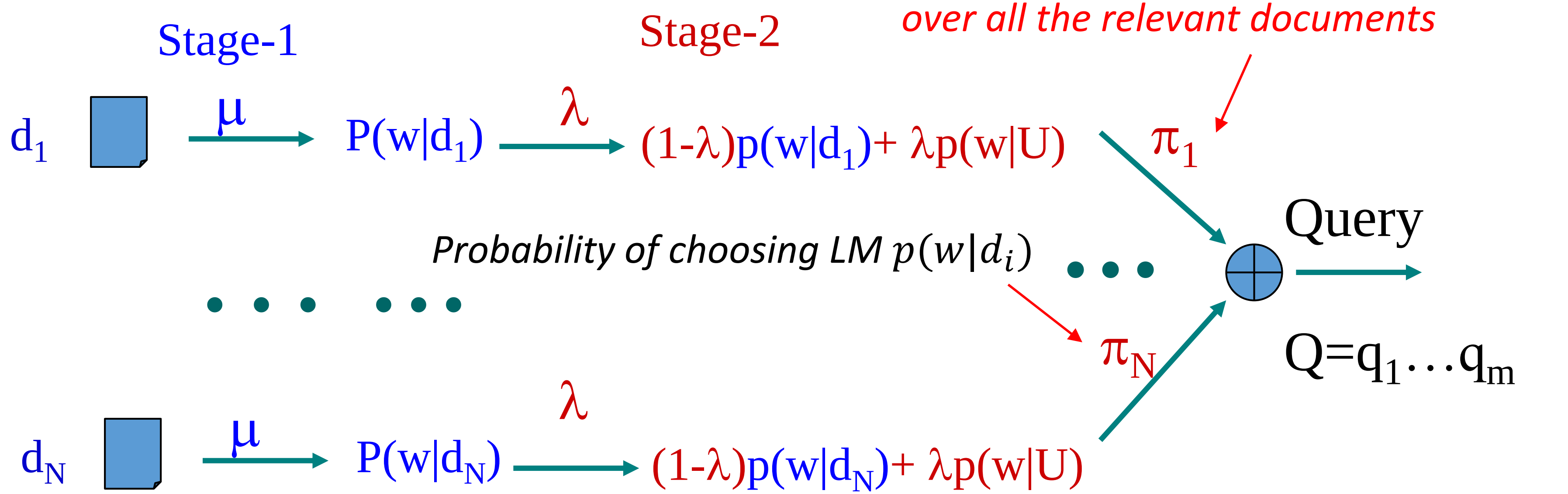
$$P(w|d) = (1-\lambda)\frac{c(w,d) + \mu p(w|C)}{|d| + \mu} + \lambda p(w|U)$$

User background model
Can be approximated by p(w|C)

Estimating λ using EM algorithm

[Zhai & Lafferty 02]

Consider query generation as a mixture over all the relevant documents



$$p(Q \mid \lambda, U) = \sum_{i=1}^{N} \pi_i \prod_{j=1}^{m} ((1-\lambda)p(q_j \mid d_i) + \lambda p(q_j \mid U))$$

$$\hat{\lambda} = \underset{\lambda}{\operatorname{argmax}}\ p(Q \mid \lambda, U)$$

Estimated in stage-1

$$p(q_j \mid d_i) = \frac{c(q_j, d_i) + \hat{\mu} p(q_j \mid C)}{|d_i| + \hat{\mu}}$$

Expectation-Maximization (EM) algorithm for estimating λ and $\{\pi_i\}_{i=1}^{N}$

Introduction to EM

- Parameter estimation
  - All data is observable
    - Maximum likelihood method
    - Optimize the analytic form of $L(\theta) = \log p(X|\theta)$
  - Missing/unobservable data
    - Data: X (observed) + H(hidden)
    - Likelihood: $L(\theta) = \log \int p(X, \mathrm{H}|\theta) \mathrm{dH}$
    - Approximate it!

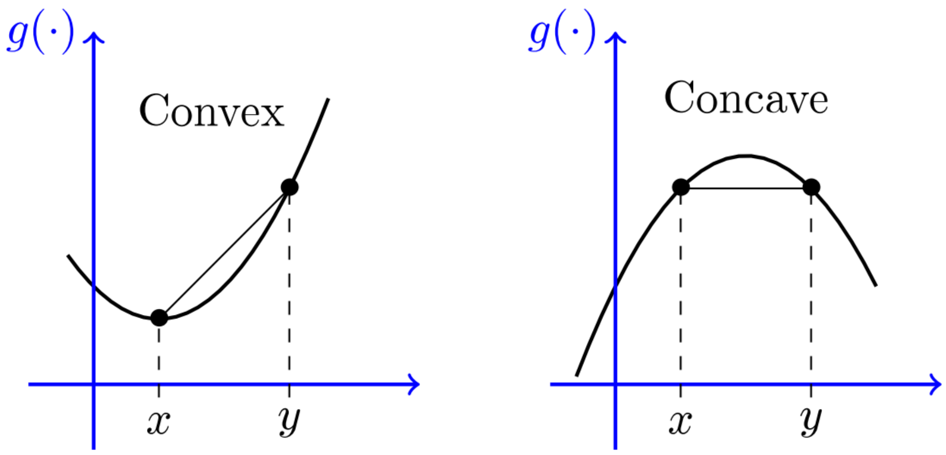
Most of cases are intractable

We have missing dat~~e~~.

Background knowledge

- Jensen's inequality
  - For any convex function $f(x)$ and positive weights λ,

$$f\left(\sum_i \lambda_i x_i\right) \leq \sum_i \lambda_i f(x_i) \qquad \sum_i \lambda_i = 1$$



Expectation Maximization

- Maximize data likelihood function by pushing the lower bound
  - $L(\theta) = \log \int p(X, \mathrm{H}|\theta) \mathrm{dH} = \log \int \frac{q(H)p(X, \mathrm{H}|\theta)}{q(H)} \mathrm{dH}$

Proposal distributions for $q(H)$

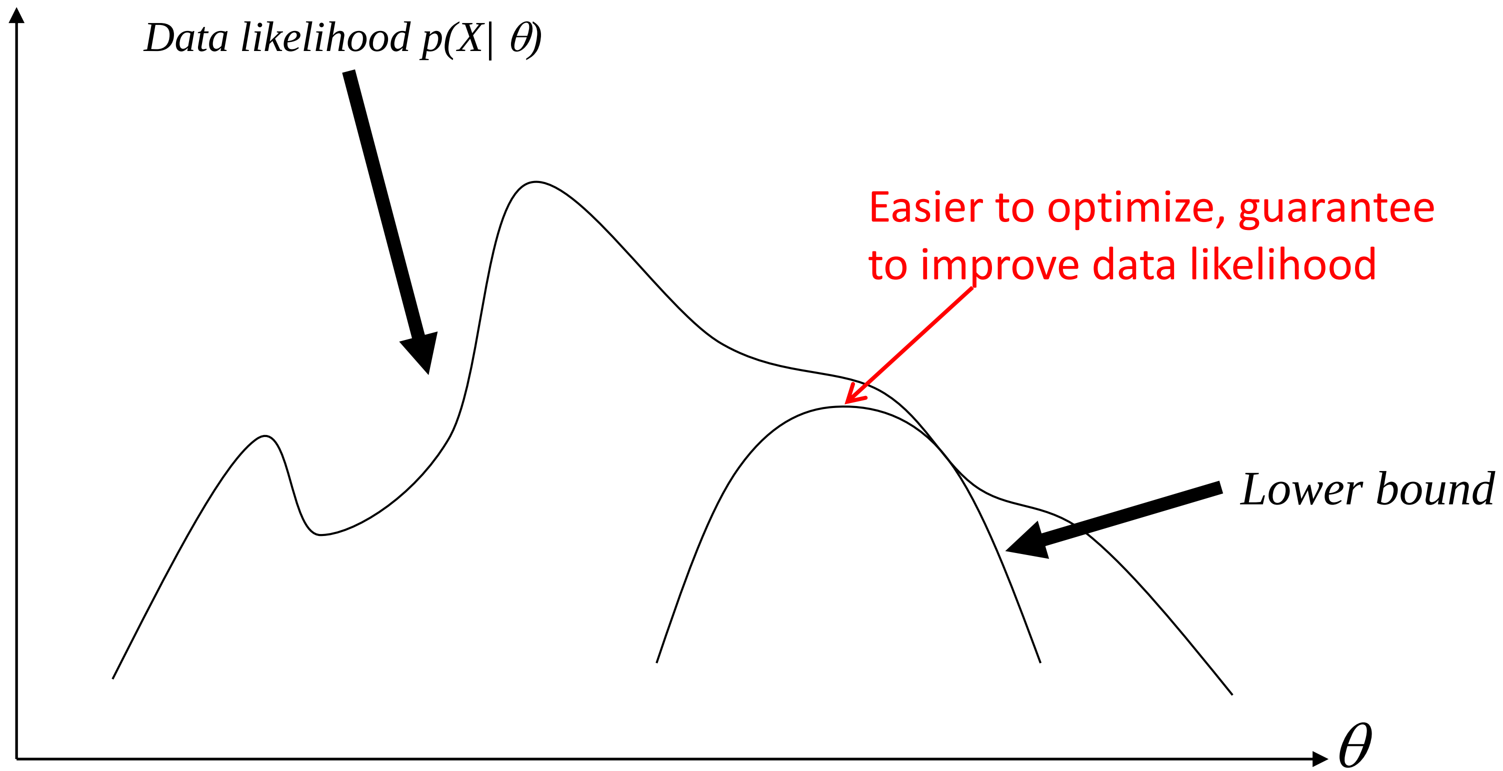
Jensen's inequality
$f(E[x]) \geq E[f(x)]$

$$\geq \int q(H) \log p(X, H|\theta)\, dH - \int q(H) \log q(H)\, dH$$

Lower bound: easier to compute, many good properties!

Components we need to tune when optimizing $\boldsymbol{L(\theta)}$: $\boldsymbol{q(H)}$ and $\boldsymbol{\theta}$!

Intuitive understanding of EM



Expectation Maximization (cont)

- Optimize the lower bound with respect to $q(H)$
  - $L(\theta) = \log \int p(X, \mathrm{H}|\theta)\mathrm{dH} = \log \int \frac{q(H)p(X, \mathrm{H}|\theta)}{q(H)} \mathrm{dH}$

Jensen's inequality
$f(E[x]) \geq E[f(x)]$

$$\geq \int q(H) \log p(X, H|\theta)\, dH - \int q(H) \log q(H)\, dH$$

$$= \int q(H)[\log p(H|X, \theta) + \log p(X|\theta)] dH - \int q(H) \log q(H)\, dH$$

$$= \int q(H) \log \frac{p(H|X, \theta)}{q(H)} dH + \log p(X|\theta)$$

KL-divergence between $q(H)$ and $p(H|X, \theta)$

Constant with respect to $q(H)$

Expectation Maximization (cont)

- Optimize the lower bound with respect to $q(H)$
  - $L(\theta) \geq KL\big(q(H)||p(H|X,\theta)\big) + L(\theta)$
  - KL-divergence is non-negative, and equals to zero iff $q(H) = p(H|X,\theta)$
  - A step further: when $q(H) = p(H|X,\theta)$, we will get $L(\theta) \geq \mathrm{L}(\theta)$, i.e., the lower bound is tight!
  - Other choice of $q(H)$ cannot lead to this tight bound, but might reduce computational complexity
  - **Note**: calculation of $q(H)$ is based on current θ

Expectation Maximization (cont)

- Optimize the lower bound with respect to $q(H)$
  - Optimal solution: $\mathrm{q}(\mathrm{H}) = p(H|X, \theta^t)$

Posterior distribution of H given current model θ^t

Expectation Maximization (cont)

- Optimize the lower bound with respect to θ
  - $L(\theta) \geq \int p(H|X,\theta^t) \log p(X,H|\theta)\, dH - \cancel{\int p(H|X,\theta^t) \log p(H|X,\theta^t)\, dH}$ ← Constant w.r.t. θ
  - $\theta^{t+1} = argmax_\theta \int p\left(H|X,\theta^t\right) \log p(X,H|\theta)\, dH$

    $= argmax_\theta E_{H|X,\theta^t}[\log p(X,H|\theta)]$

Expectation of complete data likelihood

Expectation Maximization

- EM tries to iteratively maximize likelihood
  - "Complete" likelihood: $L^c(\theta) = \log p(X, H| \theta)$
  - Starting from an initial guess $\theta^{(0)}$,
    1. **E-step**: compute the expectation of the complete likelihood

$$Q(\theta; \theta^t) = \mathrm{E}_{H|X,\theta^t}[L^c(\theta)] = \int p(H|X, \theta^t) \log \mathrm{p}(\mathrm{X}, \mathrm{H}|\theta^t) \mathrm{dH}$$

    2. **M-step**: compute $\theta^{(t+1)}$ by maximizing the Q-function

$$\theta^{t+1} = argmax_\theta Q(\theta; \theta^t)$$

Key step!

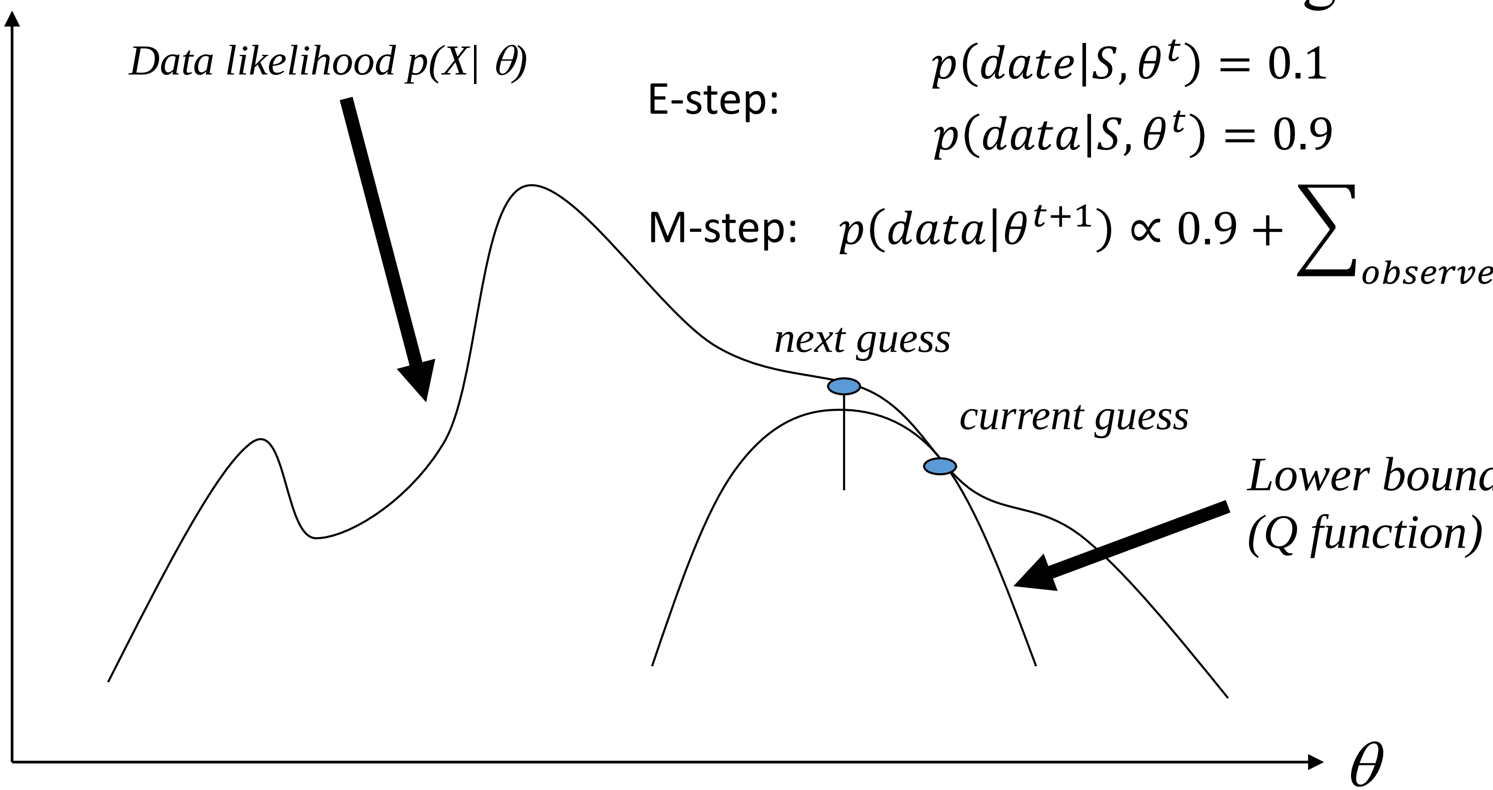
Intuitive understanding of EM

S=We have missing dat~~e~~.

E-step: $p(date|S,\theta^t) = 0.1$

$p(data|S,\theta^t) = 0.9$

M-step: $p(data|\theta^{t+1}) \propto 0.9 + \sum_{observed} 1$



E-step = computing the lower bound

M-step = maximizing the lower bound

Convergence guarantee

- Proof of EM

$$\log p(X|\theta) = \log p(H,X|\theta) - \log p(H|X,\theta)$$

Taking expectation with respect to $p(H|X,\theta^t)$ of both sides:

$$\log p(X|\theta) = \int p(H|X,\theta^t) \log p(H,X|\theta) dH - \int p(H|X,\theta^t) \log p(H|X,\theta) dH$$

$$\log p(X|\theta) = Q(\theta;\theta^t) + \underline{H(\theta;\theta^t)}$$ Cross-entropy

Then the change of log data likelihood between EM iteration is:

$$\log p(X|\theta) - \log p(X|\theta^t) = Q(\theta;\theta^t) + H(\theta;\theta^t) - Q(\theta^t;\theta^t) - H(\theta^t;\theta^t)$$

By Jensen's inequality, we know $H(\theta;\theta^t) \geq H(\theta^t;\theta^t)$, that means

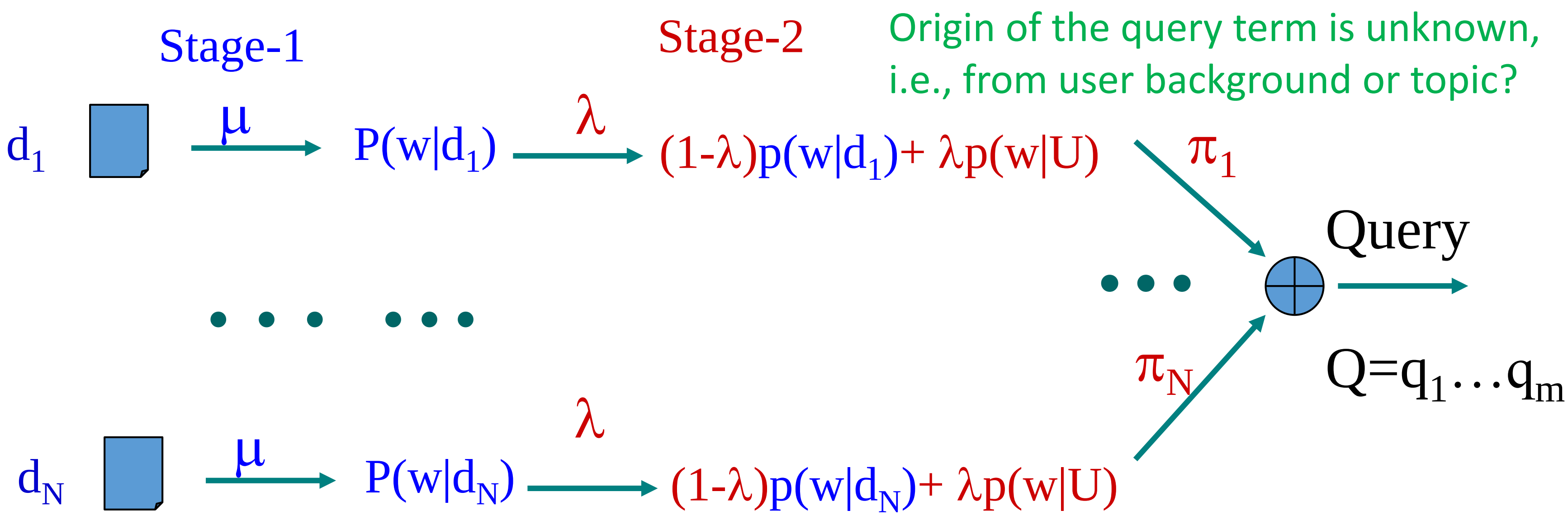
$$\log p(X|\theta) - \log p(X|\theta^t) \geq Q(\theta;\theta^t) - Q(\theta^t;\theta^t) \underline{\geq 0}$$

M-step guarantee this

What is not guaranteed

- Global optimal is not guaranteed!
  - Likelihood: $L(\theta) = \log \int p(X, \mathrm{H}|\theta) \mathrm{dH}$ is non-convex in most of case
  - EM boils down to a <u>greedy</u> algorithm
    - Alternative ascent
- Generalized EM
  - E-step: $\hat{q}(H) = \mathrm{argmin}_{q(H)} KL\big(q(H)||p(H|X,\theta)\big)$
  - M-step: choose θ that improves $Q(\theta; \theta^t)$

Estimating λ using Mixture Model [Zhai & Lafferty 02]



$$p(Q \mid \lambda, U) = \sum_{i=1}^{N} \pi_i \prod_{j=1}^{m} ((1-\lambda)p(q_j \mid d_i) + \lambda p(q_j \mid U))$$

$$\hat{\lambda} = \underset{\lambda}{\operatorname{argmax}}\ p(Q \mid \lambda, U)$$

Estimated in stage-1

$$p(q_j \mid d_i) = \frac{c(q_j, d_i) + \hat{\mu} p(q_j \mid C)}{|d_i| + \hat{\mu}}$$

Expectation-Maximization (EM) algorithm for estimating λ and $\{\pi_i\}_{i=1}^{N}$

Variants of basic LM approach

- Different smoothing strategies
  - Hidden Markov Models (essentially linear interpolation) [Miller et al. 99]
  - Smoothing with an IDF-like reference model [Hiemstra & Kraaij 99]
    - Performance tends to be similar to the basic LM approach
  - Many other possibilities for smoothing [Chen & Goodman 98]
- Different priors
  - Link information as prior leads to significant improvement of Web entry page retrieval performance [Kraaij et al. 02]
  - Time as prior [Li & Croft 03]
  - PageRank as prior [Kurland & Lee 05]
- Passage retrieval [Liu & Croft 02]

Improving language models

- Capturing limited dependencies
  - Bigrams/Trigrams [Song & Croft 99]
  - Grammatical dependency [Nallapati & Allan 02, Srikanth & Srihari 03, Gao et al. 04]
  - Generally insignificant improvement as compared with other extensions such as feedback
- Full Bayesian query likelihood [Zaragoza et al. 03]
  - Performance similar to the basic LM approach
- Translation model for p(Q|D,R) [Berger & Lafferty 99, Jin et al. 02,Cao et al. 05]
  - Address polesemy and synonyms
  - Improves over the basic LM methods, but computationally expensive
- Cluster-based smoothing/scoring [Liu & Croft 04, Kurland & Lee 04,Tao et al. 06]
  - Improves over the basic LM, but computationally expensive
- Parsimonious LMs [Hiemstra et al. 04]
  - Using a mixture model to “factor out” non-discriminative words

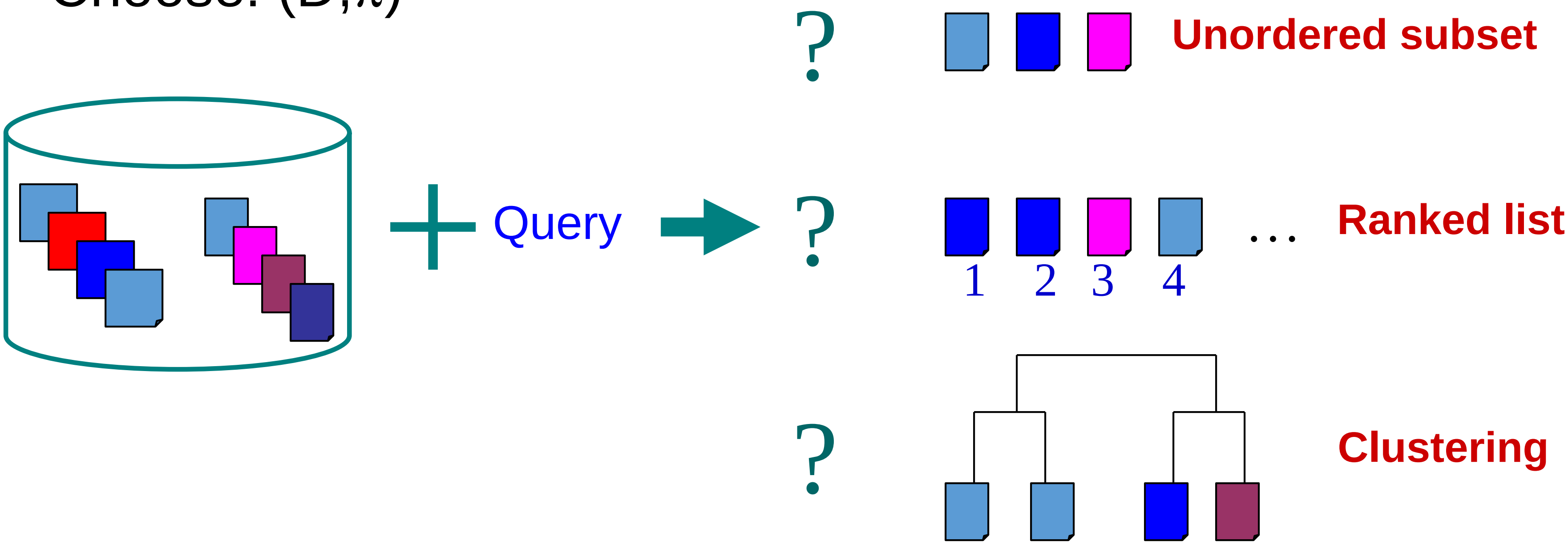
A unified framework for IR: Risk Minimization

- Long-standing IR Challenges
  - Improve IR theory
    - Develop theoretically sound and empirically effective models
    - Go beyond the limited traditional notion of relevance (independent, topical relevance)
  - Improve IR practice
    - Optimize retrieval parameters automatically
- Language models are promising tools …
  - How can we systematically exploit LMs in IR?
  - Can LMs offer anything hard/impossible to achieve in traditional IR?

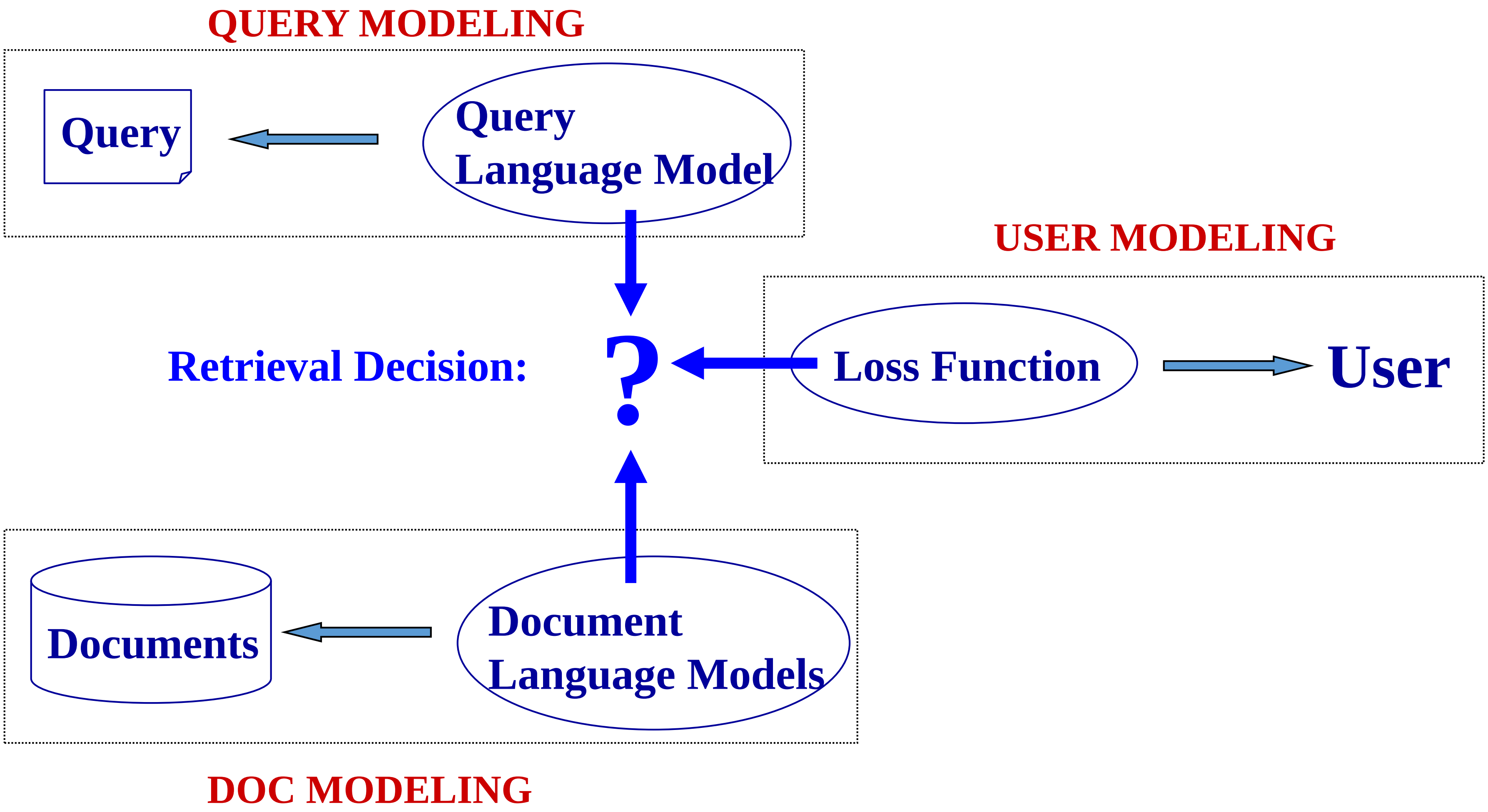
Idea 1: Retrieval as decision-making
(A more general notion of relevance)

Given a query,
- Which documents should be selected? (D)
- How should these docs be presented to the user? (π)

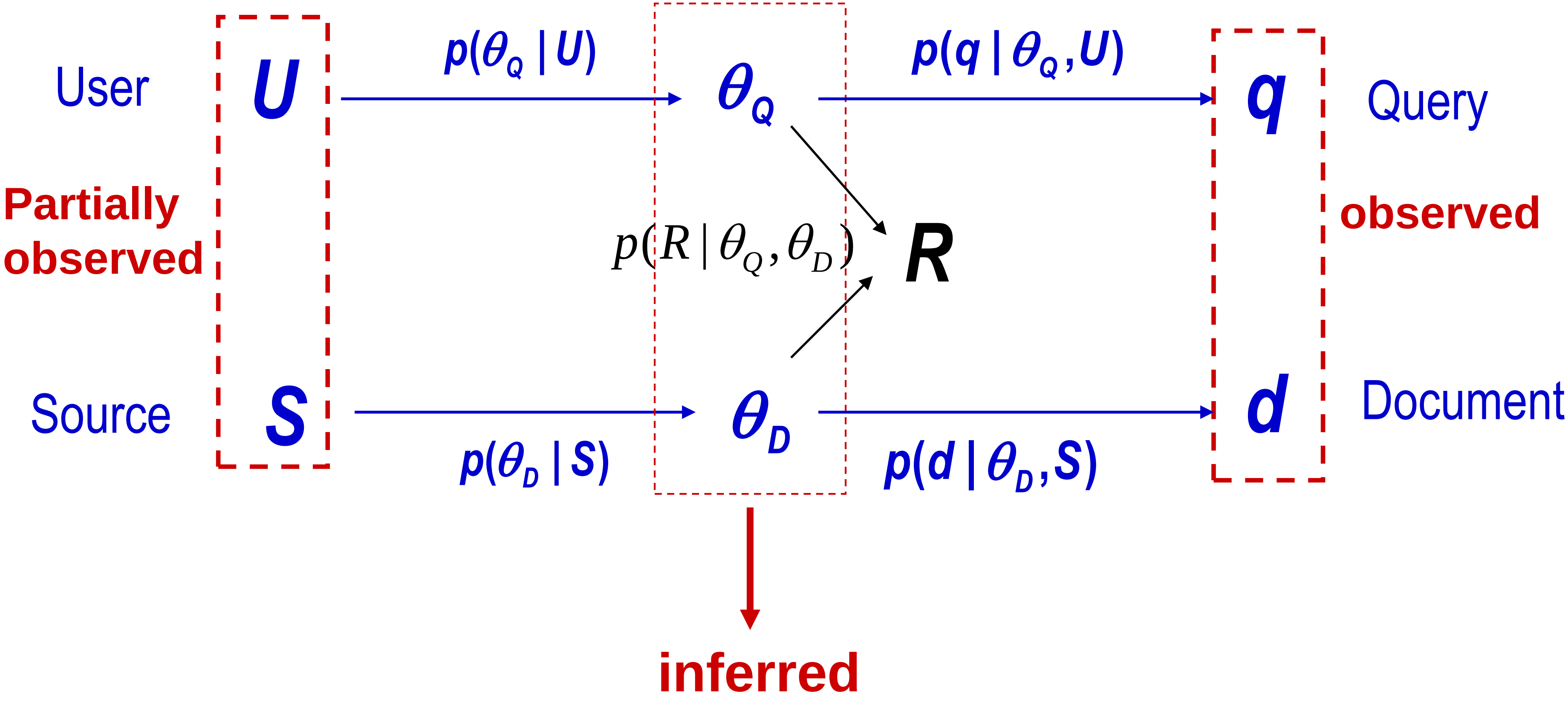
Choose: (D,π)



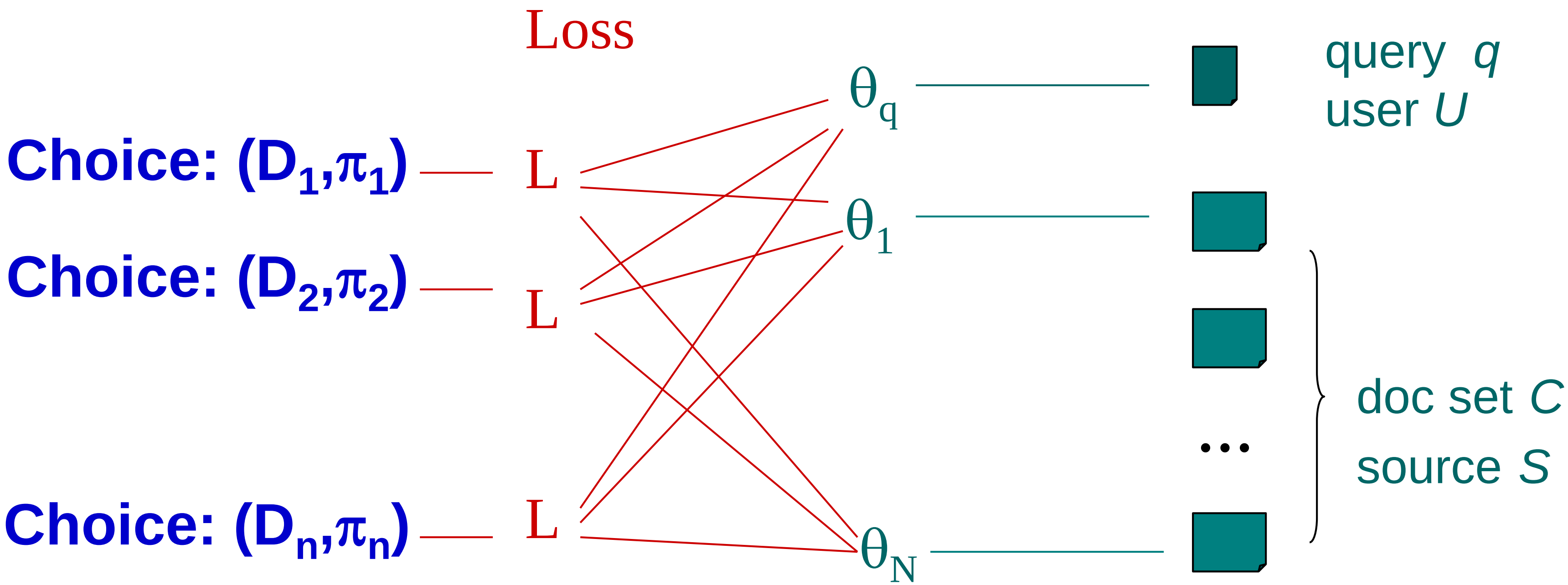
Idea 2: Systematic language modeling



Generative model of document & query
[Lafferty & Zhai 01b]



Applying Bayesian Decision Theory
[Lafferty & Zhai 01b, Zhai 02, Zhai & Lafferty 06]



$$(D^*,\pi^*) = \underset{D,\pi}{\arg\min} \int_{\Theta} \underbrace{L(D,\pi,\theta)}_{\text{loss}}\, \underbrace{p(\theta}_{\text{hidden}} \mid \underbrace{q,U,C,S)}_{\text{observed}}\, d\theta$$

RISK MINIMIZATION

Bayes risk for choice (D, π)

Special cases

- Set-based models (choose D) → **Boolean model**
- Ranking models (choose π)
  - Independent loss
    - Relevance-based loss → **Probabilistic relevance model**, **Generative Relevance Theory**
    - Distance-based loss → **Vector-space Model**, **Two-stage LM**, **KL-divergence model**
  - Dependent loss
    - Maximum Margin Relevance loss
    - Maximal Diverse Relevance loss → **Subtopic retrieval model**

Optimal ranking for independent loss

$$\pi^* = \arg\min_{\pi} \int_{\Theta} L(\pi,\theta)\, p(\theta \mid q, U, C, \vec{S})\, d\theta$$

Decision space = {rankings}

Sequential browsing

Independent loss

s_i is the probability that the user would stop reading after seeing the top i documents

Independent risk
= independent scoring

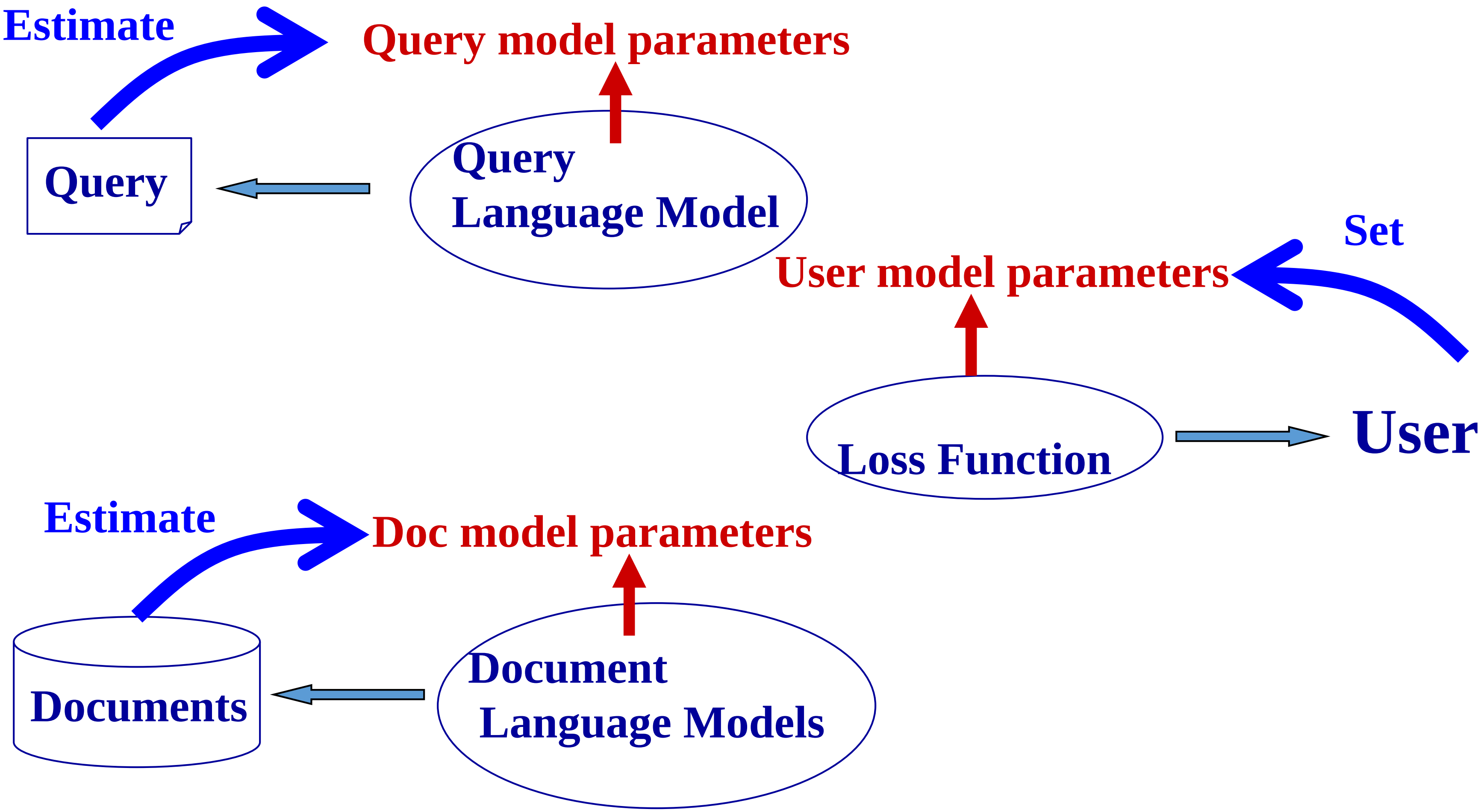
“Risk ranking principle”

[Zhai 02]

Automatic parameter tuning

- Retrieval parameters are needed to
  - model different user preferences
  - customize a retrieval model to specific queries and documents
- Retrieval parameters in traditional models
  - EXTERNAL to the model, hard to interpret
  - Parameters are introduced heuristically to implement “intuition”
  - No principles to quantify them, must set empirically through many experiments
  - Still no guarantee for new queries/documents
- Language models make it possible to estimate parameters…

Parameter setting in risk minimization



Summary of risk minimization

- Risk minimization is a general probabilistic retrieval framework
  - Retrieval as a decision problem (=risk min.)
  - Separate/flexible language models for queries and docs
- Advantages
  - A unified framework for existing models
  - Automatic parameter tuning due to LMs
  - Allows for modeling complex retrieval tasks
- Lots of potential for exploring LMs…

What you should know

- EM algorithm
- Risk minimization framework