Michal Rosen-Zvi
Director, AI for Accelerated HC&LS Discovery, IBM Research
Professor, Faculty of Medicine, The Hebrew University

VISUM Summer School, 2022

Generative Models for Healthcare and Life Sciences

Outline

- ❑ Motivation - at the verge of a new era of accelrated discoveries (10mins)

- ❑ Data types, from text to molecules – a quick intro/reminder (20 mins)
  - ❑ Coding text
  - ❑ Coding (small) molecules
  - ❑ Coding protiens

- ❑ Generative models – gentle introduction (20 mins)
  - ❑ Introduction: discreminative vs generative
  - ❑ Latent Dirichlet Allocation models – LDA models

- ❑ Generative models – Deep generative modesl (40 mins)
  - ❑ Variational Autoencoders – VAE models
  - ❑ Bidirectional Encoder Representations from Transformers– BERT
  - ❑ Reinforcement learning framing – RL

- ❑ Break

- ❑ Hands-on session (45 mins)

Motivation

- It takes on average 10–15 years and USD 1.5–2.0 billion to bring a new drug to market
- A study that identified 21,143 compounds in clinical trials found that the overall success rate was as low as 6.2%
- More than 30% of clinical trials failed in Humans because of their toxicity
- Bigger datasets of molecules are made available with an increase level of thoroughness regarding attributes of molecules; presents a huge opportunity; datasets include coverage of the following types of molecules:
  - Proteins – large biomolecules and macromolecules that comprise one or more long chains of amino acid residues
  - Druglike small molecules – a molecule with attributes such as solubility in both water and fat, potency at the biological target and small molecular weight (vast majority of drugs on the market are small molecules - have molecular weights between 200 and 600 Daltons, which is 1.7e-24 gram)

About me: an IBMer in the past 15 years and an adjunct professor at the Hebrew University since 2022

Overview of IBM Research Innovation Engine

- IBM Research has a portfolio compromised of activities and cutting-edge technologies spanning across exploratory science and translational technologies aimed at biomarker and molecule discovery
- Over the last 20 years, key partnerships, top talent, and external eminence have been cornerstones

Key Partnerships

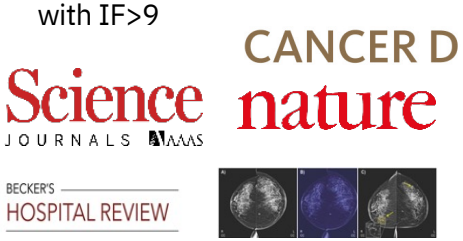
- Various partner engagement models to gain access to data, studies, validation & funding
- Research grants supporting innovation in US, Europe and Africa

Top-Tier Talent

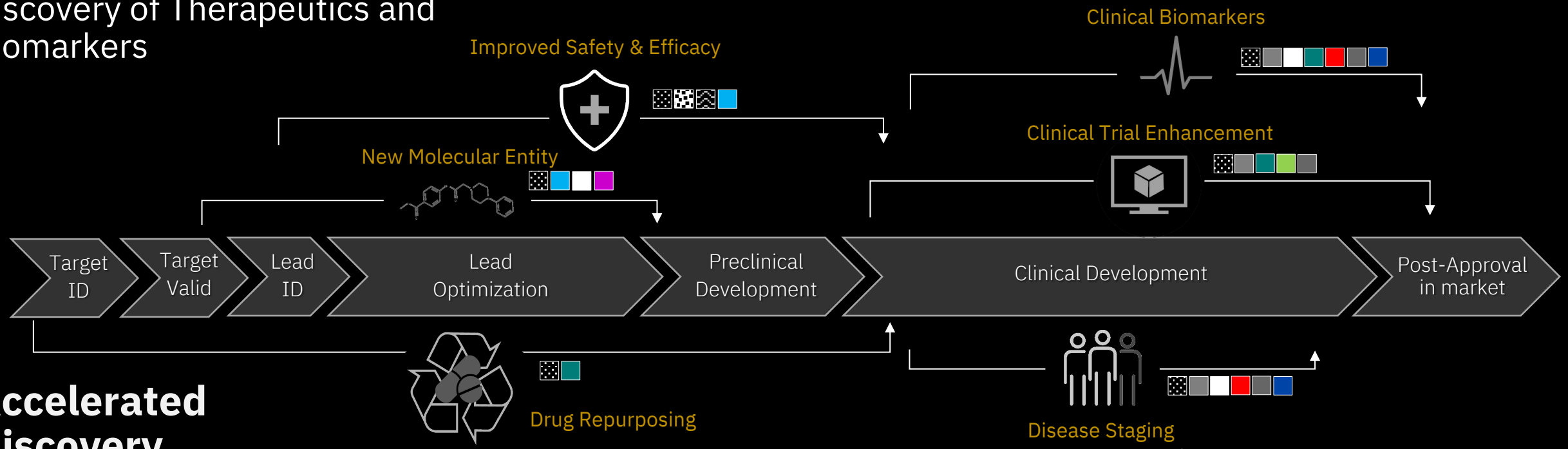
- 13+ average patents awarded per researcher
- 12+ Fellows of professional societies, among Research leaders alone

External Eminence

- 153 PubMed-indexed publications in 2021
- 33 of these publications in 2021 are in venues with IF>9



Motivations Accelerating Discovery of Therapeutics and Biomarkers



Accelerated Discovery Enablers

Accelerator Technologies for Therapeutics and Biomarkers

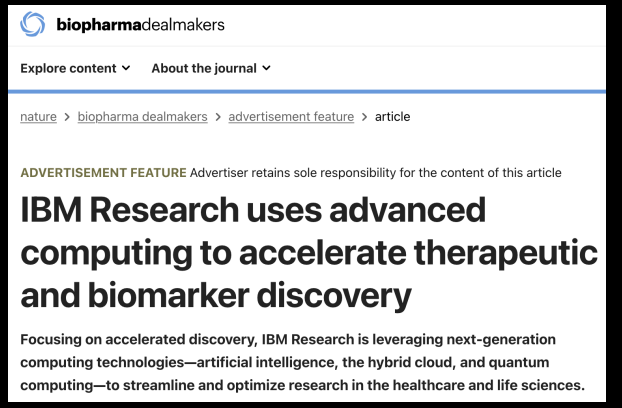
- Causal Inference
- Histopathology Graph Models
- Molecule Generation
- Digital Health
- Disease Progression Modeling
- Multimodal Data Fusion

Discovery Technology Foundations

- Knowledge Extraction
- AI-Enriched Simulation
- Generative Models

Core Technologies

- AI
- HPC
- Hybrid Cloud
- Quantum



biopharmadealmakers

Explore content · About the journal

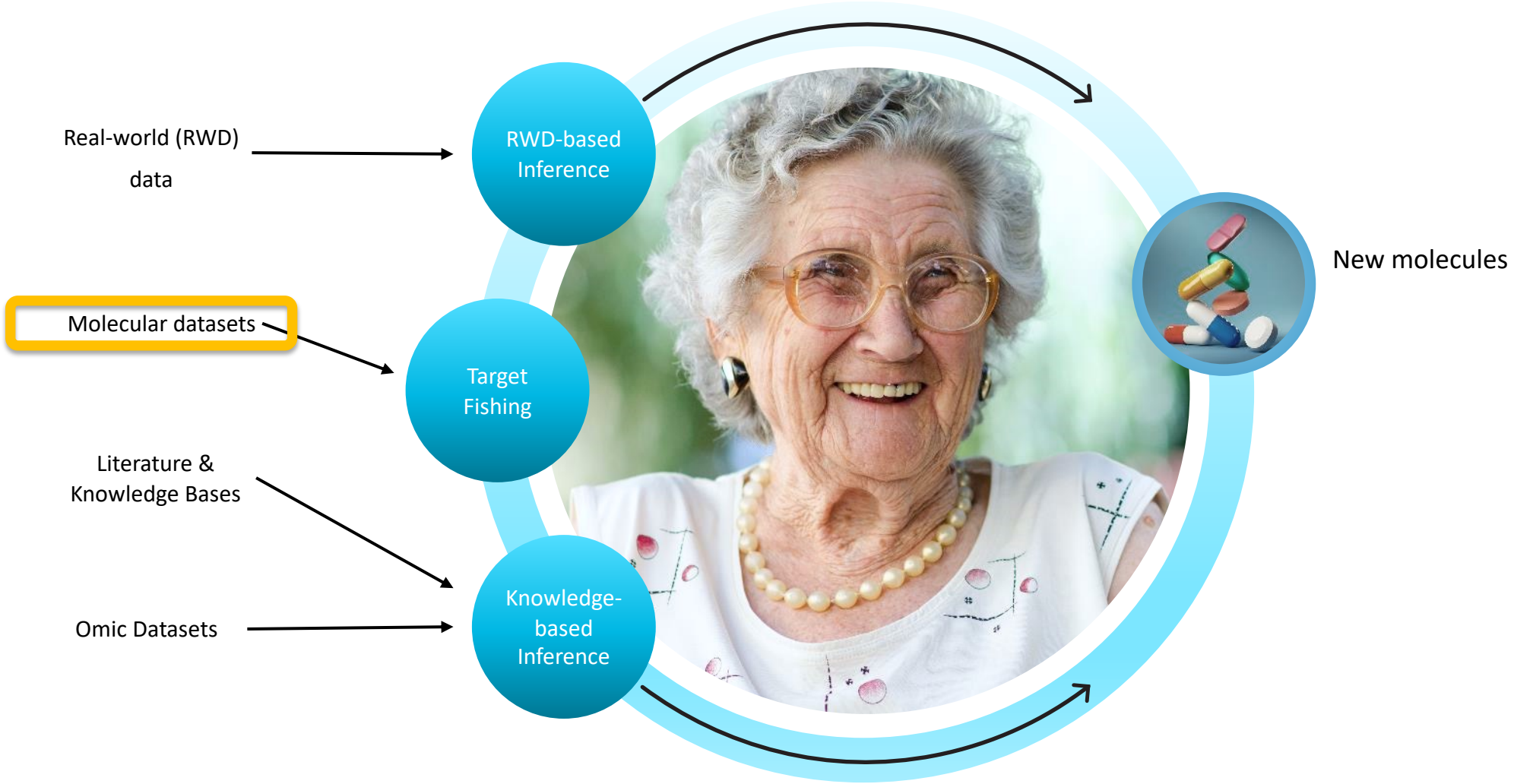
nature › biopharma dealmakers › advertisement feature › article

ADVERTISEMENT FEATURE Advertiser retains sole responsibility for the content of this article

IBM Research uses advanced computing to accelerate therapeutic and biomarker discovery

Focusing on accelerated discovery, IBM Research is leveraging next-generation computing technologies—artificial intelligence, the hybrid cloud, and quantum computing—to streamline and optimize research in the healthcare and life sciences.

Drug discovery



Outline

- ❑ Motivation - at the verge of a new era of accelrated discoveries (10mins)

- ❑ Data types, from text to molecules – a quick intro/reminder (20 mins)
  - ❑ Coding text
  - ❑ Coding (small) molecules
  - ❑ Coding protiens

- ❑ Generative models – gentle introduction (20 mins)
  - ❑ Introduction: discreminative vs generative
  - ❑ Latent Dirichlet Allocation models – LDA models

- ❑ Generative models – Deep generative modesl (40 mins)
  - ❑ Variational Autoencoders – VAE models
  - ❑ Bidirectional Encoder Representations from Transformers– BERT
  - ❑ Reinforcement learning framing – RL

- ❑ Hands-on session (45 mins)

Coding text: aiming at a standraized codifiable representation of text

"The goal is to develop a conditional generative model that can be queried with a protein and returns novel ligands that have high binding affinities to the target and, as a secondary objective, low toxicity. This goal is achieved by employing two predictive models. One model for binding affinity and another model for toxicity. Both models are employed as reward functions for the conditional generator."

Sentence Tokenization

The goal is to develop a conditional generative model that can be queried with a protein and returns novel ligands that have high binding affinities to the target and, as a secondary objective, low toxicity.

This goal is achieved by employing two predictive models.

One model for binding affinity and another model for toxicity. Both models are employed as reward functions for the conditional generator.

Word Tokenization

[The, goal, is, to, develop, a, conditional, generative, model, that, can, be, queried, with, a, protein, and, returns, novel, ligands, that, have, high, binding, affinities, to, the, target, and, as, a, secondary, objective, low, toxicity] …

Stemming and lemminzation

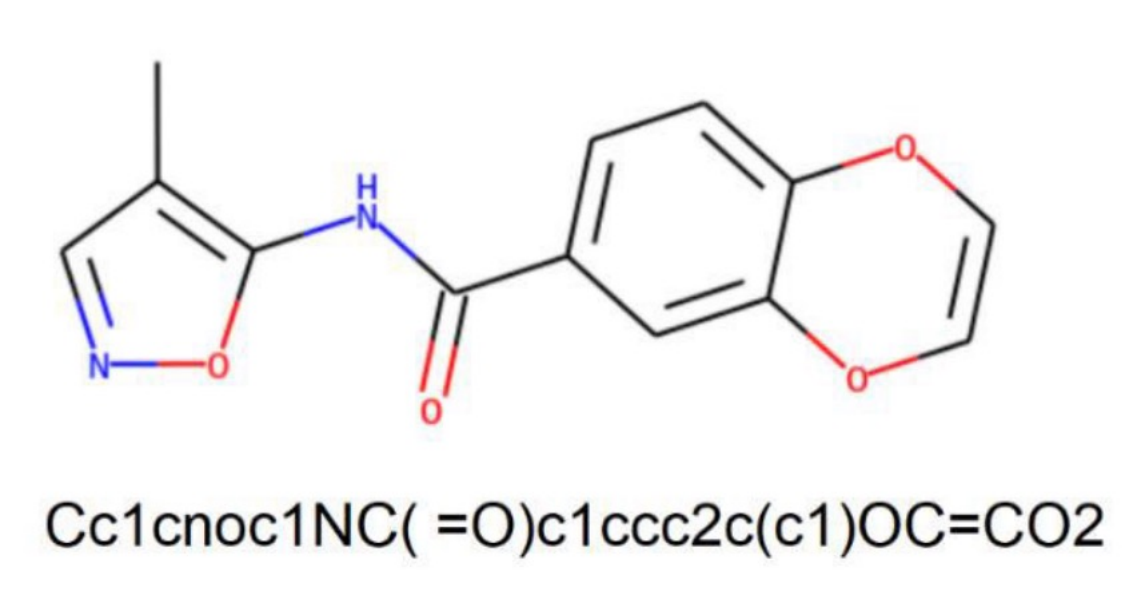
[The, goal, is, to, develop, a, condition, generative, model, that, can, be, query, with, a, protein, and, return, novel, ligand, that, have, high, bind, affinity, to, the, target, and, as, a, secondary, objective, low, toxicity] ...

Coding molecules: aiming at a standraized codifiable representation of a molecule (1/3)

Morphine

Coding molecules: aiming at a standraized codifiable representation of a molecule (2/3)

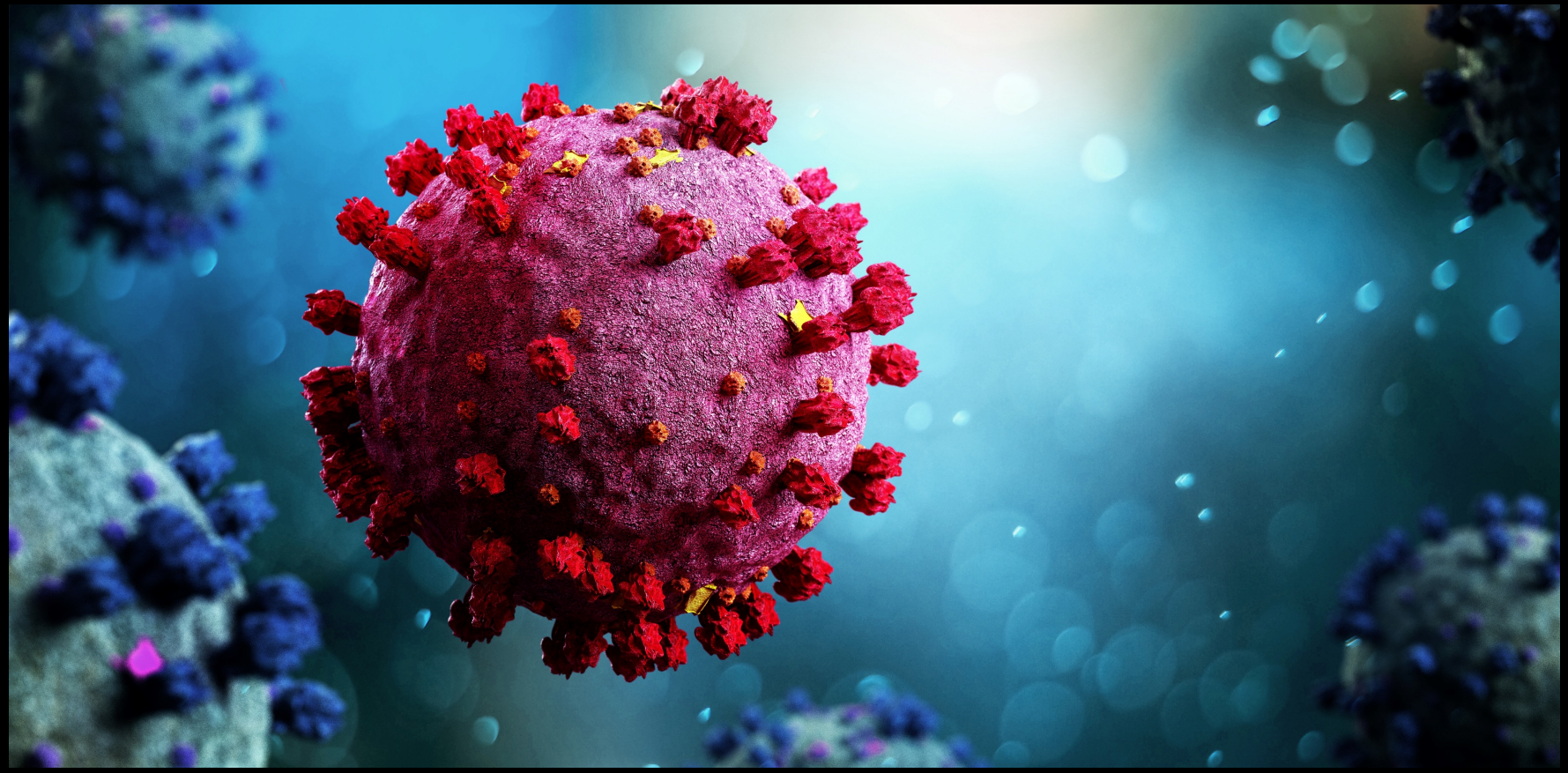
- A string notation - Simplified Molecular Input Line Entry System (SMILES)
- 2D graphs



Cc1cnoc1NC(=O)c1ccc2c(c1)OC=CO2

Coding molecules: Protiens(3/3)

large biomolecules and macromolecules that comprise one or more long chains of amino acid residues
The SARS-CoV-2 genome encodes 29 proteins. Each plays a specific role in the virus lifecycle
Example: Spike glycoprotein (S) has a length of 1273 amino acids, and mass of 141,178(Da) The virus uses its S to bind its receptor, and mediate host cells membrane fusion and virus entry.



Outline

- ❑ Motivation - at the verge of a new era of accelrated discoveries (10mins)

- ❑ Data types, from text to molecules – a quick intro/reminder (20 mins)
  - ❑ Coding text
  - ❑ Coding (small) molecules
  - ❑ Coding protiens

- ❑ Generative models – gentle introduction (20 mins)
  - ❑ Introduction: discreminative vs generative
  - ❑ Latent Dirichlet Allocation models – LDA models

- ❑ Generative models – Deep generative modesl (40 mins)
  - ❑ Variational Autoencoders – VAE models
  - ❑ Bidirectional Encoder Representations from Transformers– BERT
  - ❑ Reinforcement learning framing – RL

- ❑ Break

- ❑ Hands-on session (45 mins)

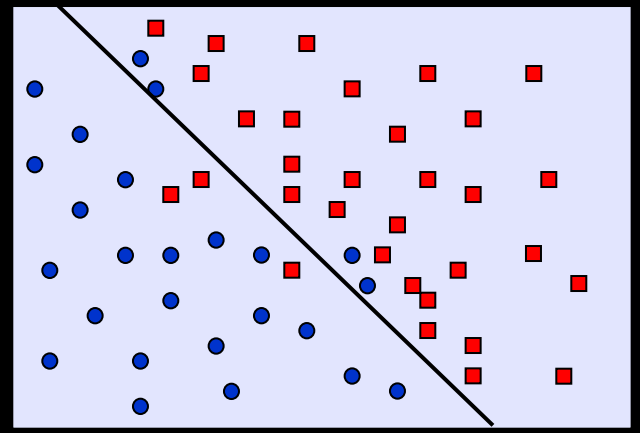
Classification problem: Discriminative approach vs Generative approach (1/2)

Input:

a set X of samples

A set Y of labels.

A training dataset S = $\{(x_1,y_1), (x_2,y_2), (x_3,y_3), \ldots, (x_m,y_m)\}$



Output:

A hypothesis (prediction rule) h: X → Y

Can be used for prediction on new samples from X

Learning algorithm: selects a good hypothesis from a predefined hypotheses class H

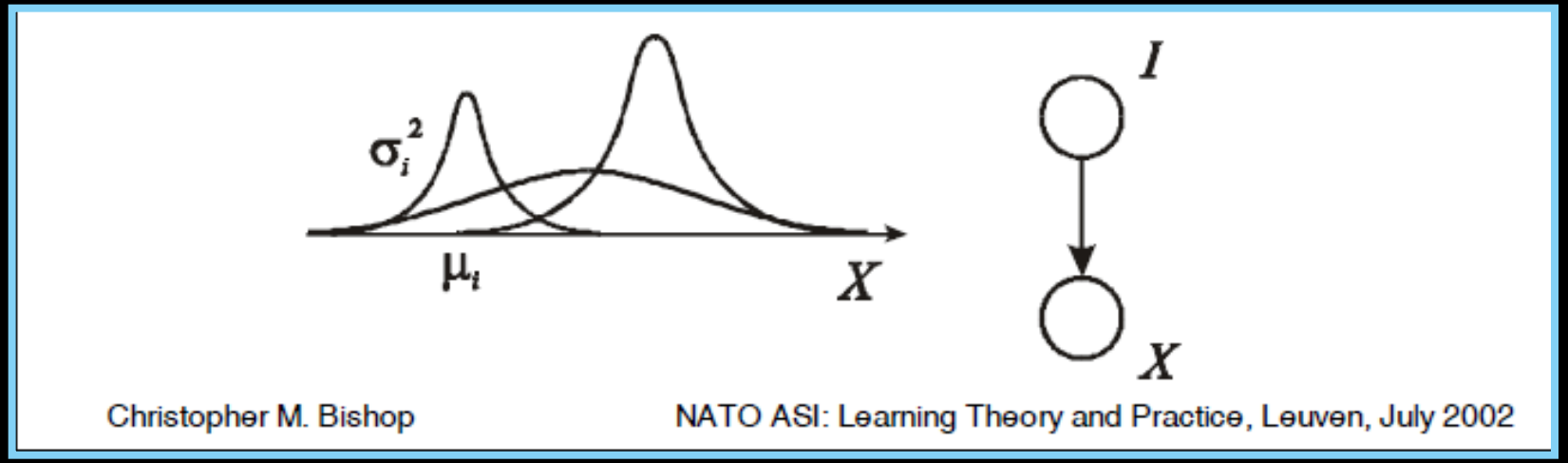
A ***loss function*** $L(h(x), y)$ is a measure of the ***classification quality***

Example: binary classification and the ***0-1 loss***: $L(h(x), y) = I(h(x) \neq y)$

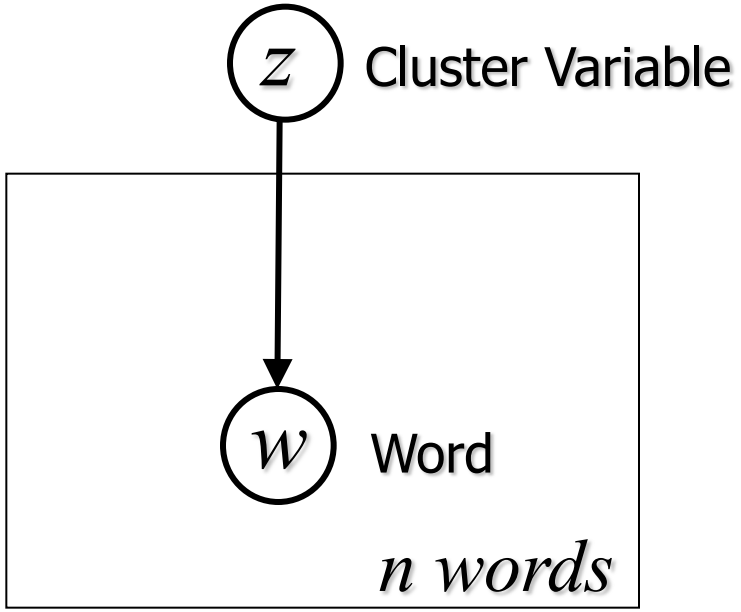
Classification problem: Example: Mixture of Gaussians (2/2)

$P(Class=i)=\pi_i$
$P(X|class=i)=Normal(\mu_i,\sigma_i)$
$P(X)=\Sigma_i\, P(Class=i)\, P(X|class=i) = \Sigma_i\, \pi_i\, Normal(\mu_i,\sigma_i)$

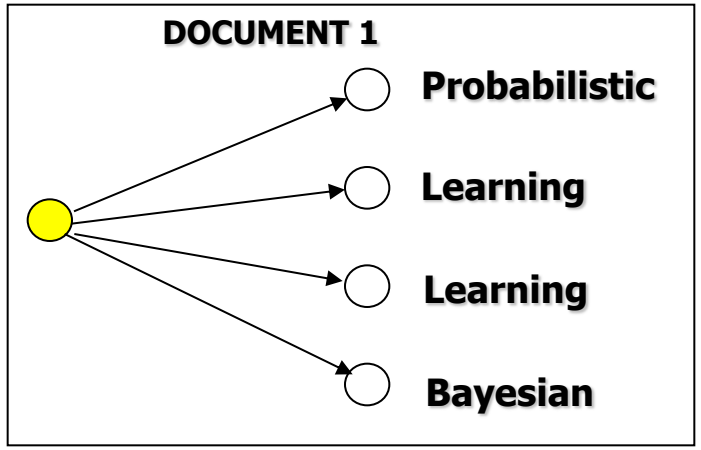
Data representation includes iid samples from a latent variable. Aiming at learning the parameters, inferring the classes and generating new examples

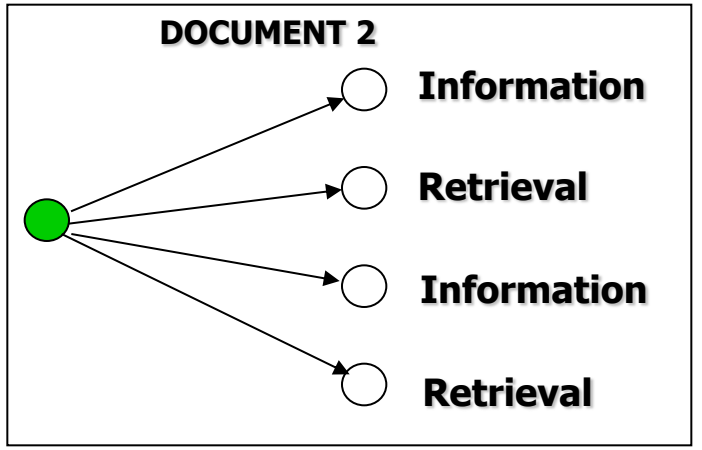


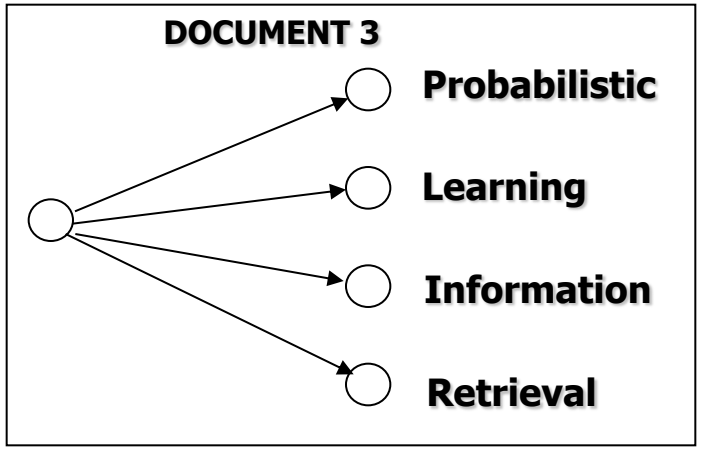
Graphical models 1/3



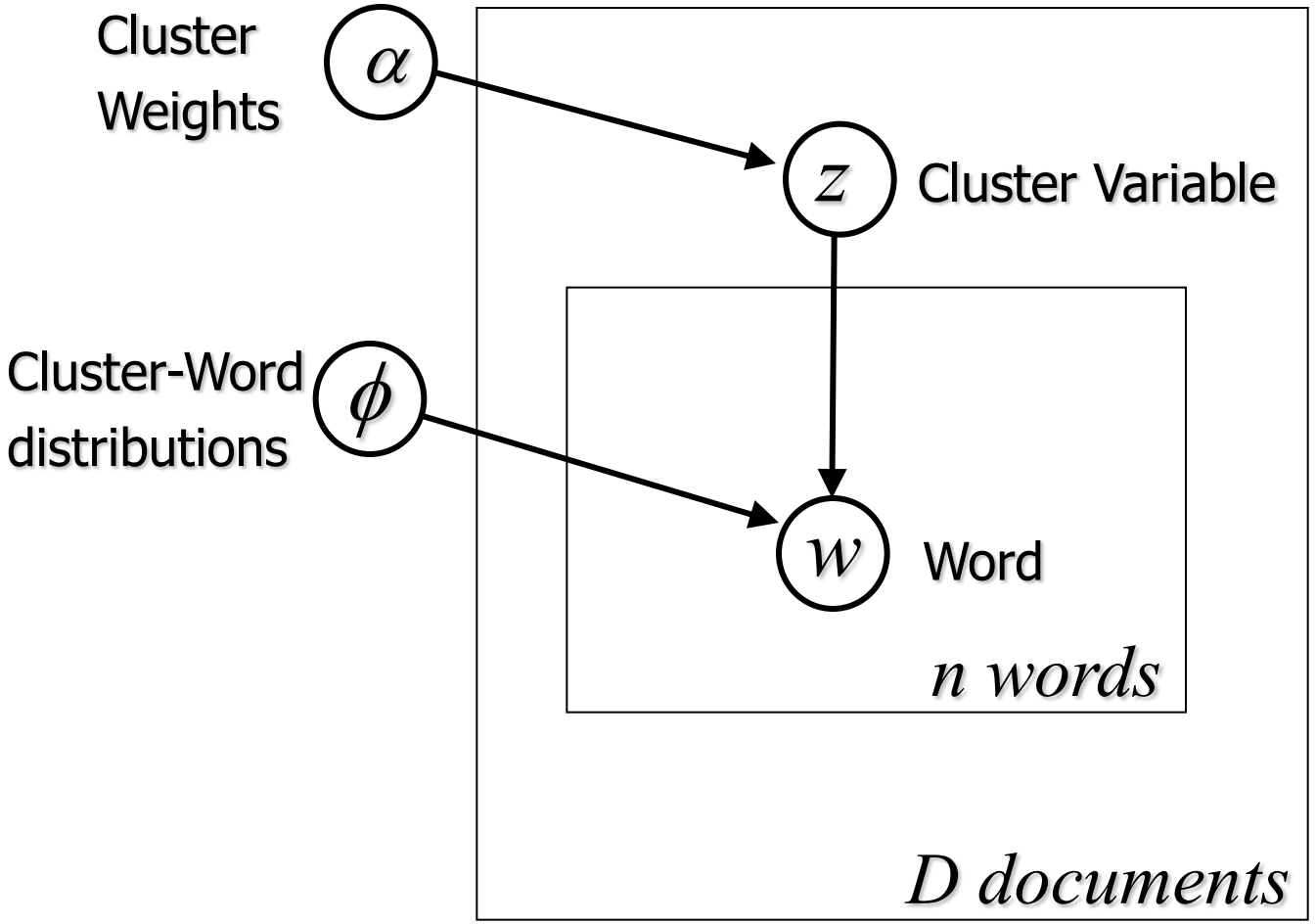
Cluster Models



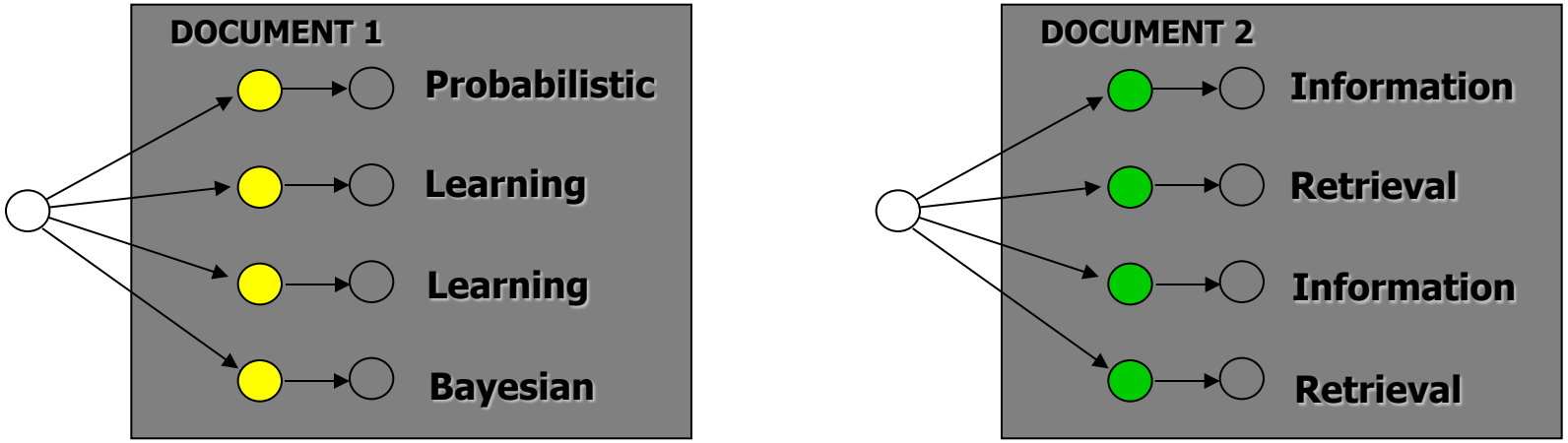




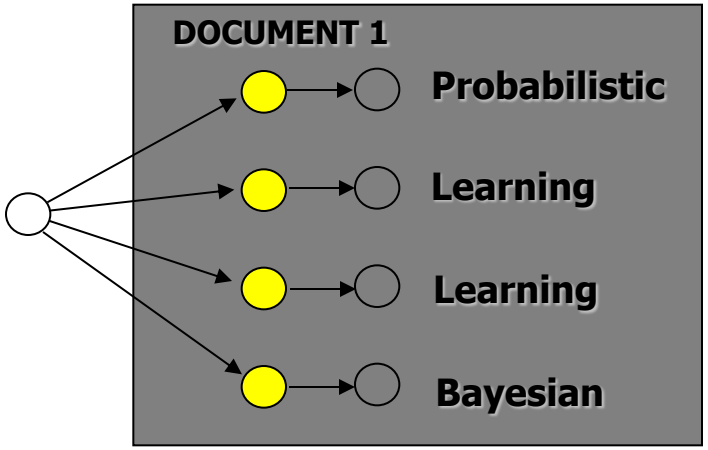
Graphical Model

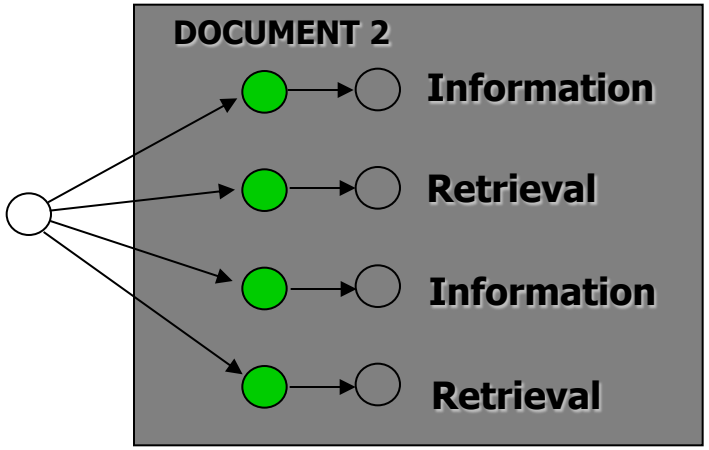


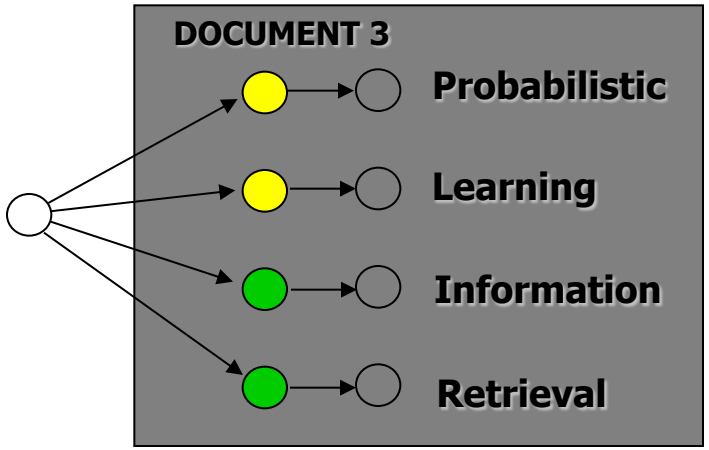
Topic Models



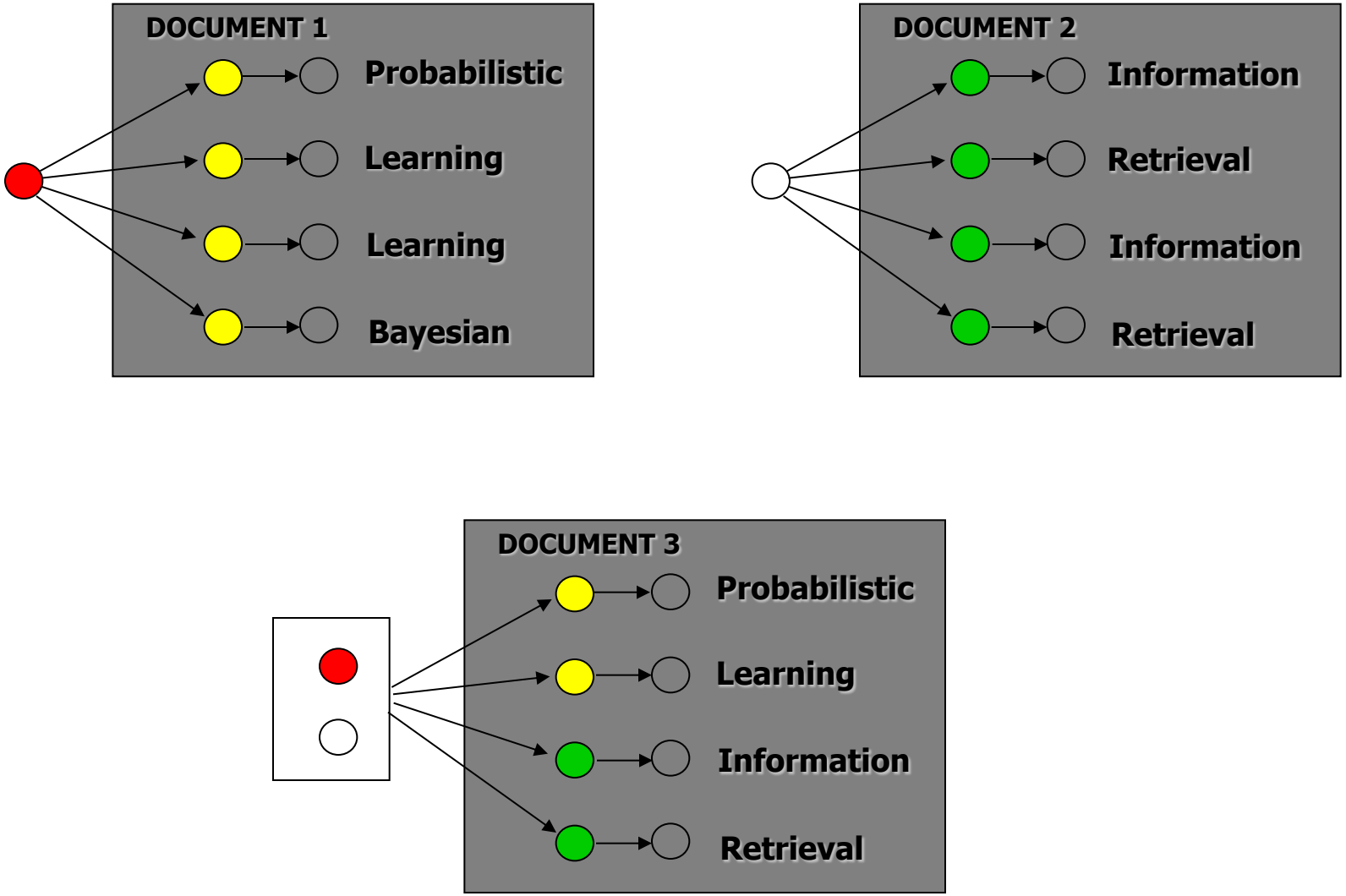
Topic Models

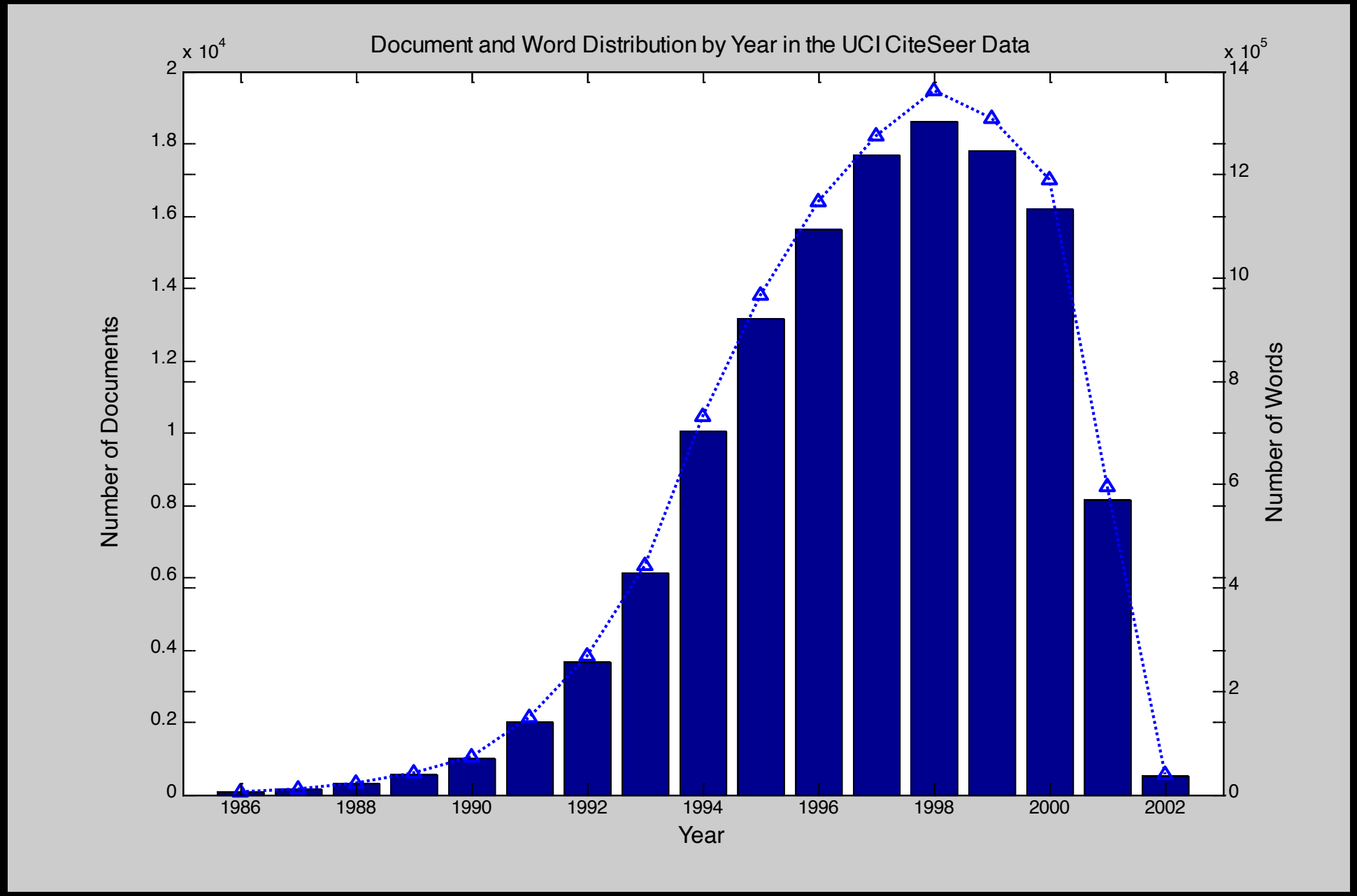


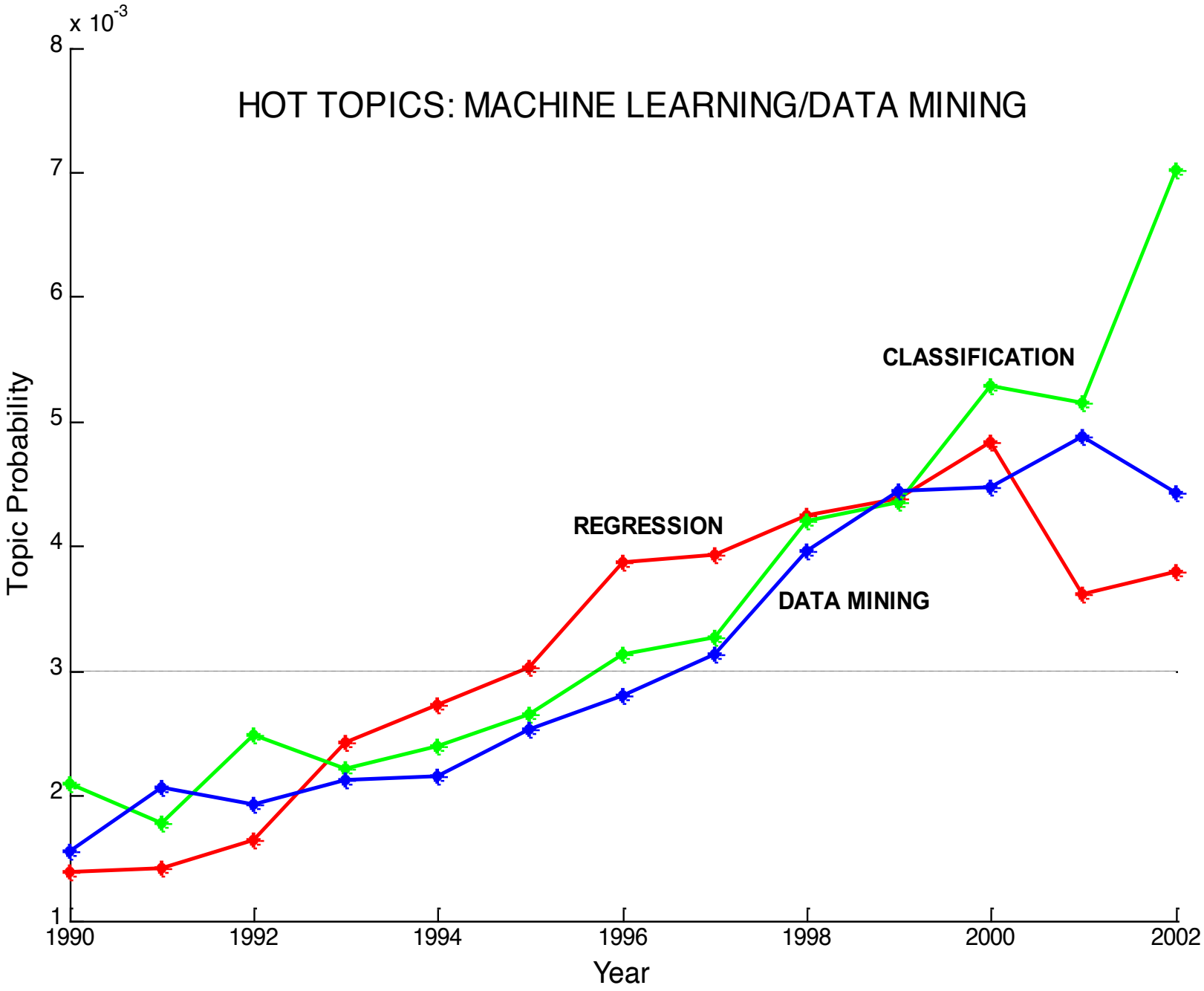


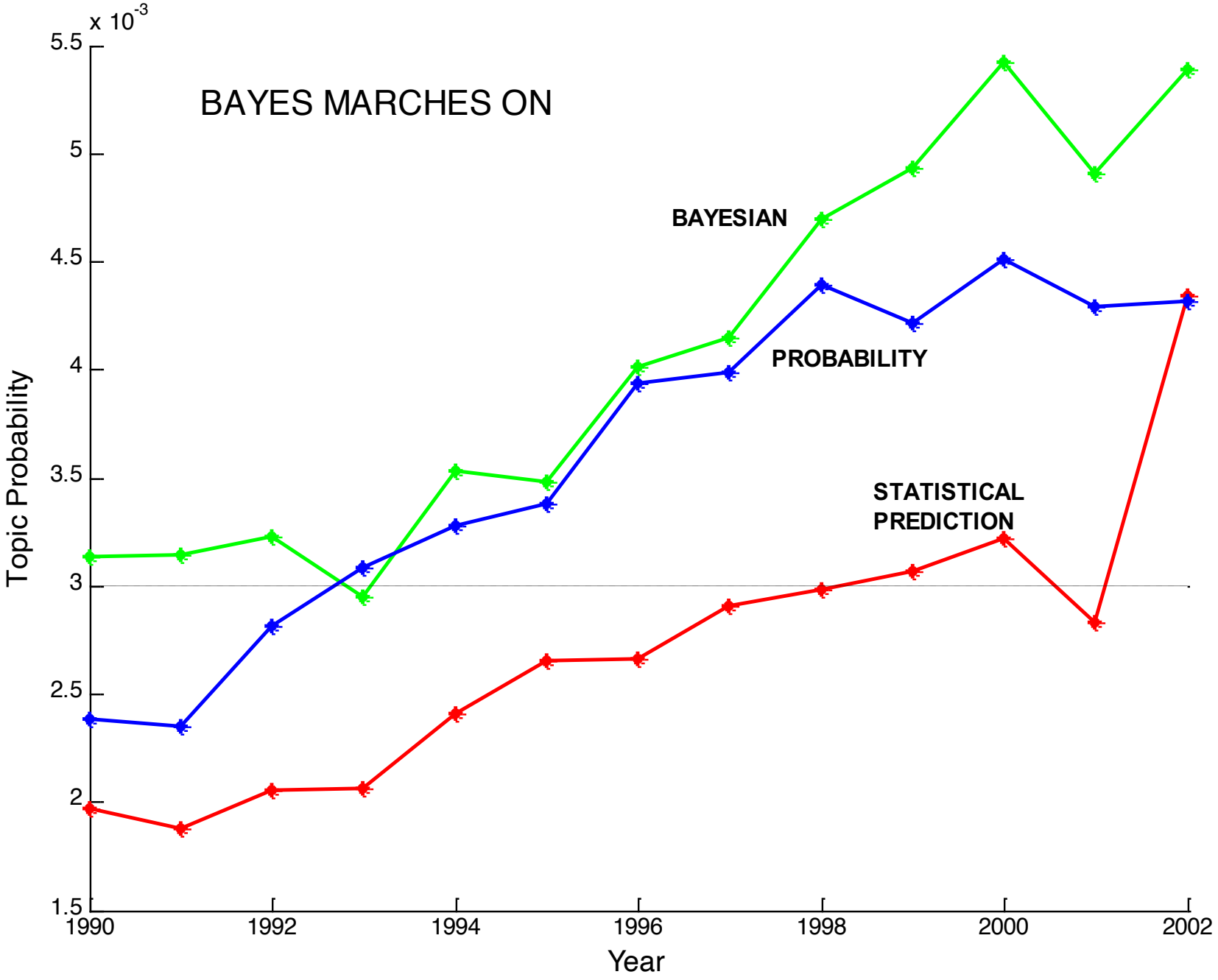


Author-Topic Models




Document and Word Distribution by Year in the UCI CiteSeer Data
x 10^4
x 10^5
Number of Documents
Number of Words
Year
0
0.2
0.4
0.6
0.8
1
1.2
1.4
1.6
1.8
2
0
2
4
6
8
10
12
14
1986
1988
1990
1992
1994
1996
1998
2000
2002


HOT TOPICS: MACHINE LEARNING/DATA MINING
x 10^-3
8
7
6
5
4
3
2
1
Topic Probability
1990
1992
1994
1996
1998
2000
2002
Year
CLASSIFICATION
REGRESSION
DATA MINING


BAYES MARCHES ON
x 10^-3
5.5
5
4.5
4
3.5
3
2.5
2
1.5
Topic Probability
1990
1992
1994
1996
1998
2000
2002
Year
BAYESIAN
PROBABILITY
STATISTICAL
PREDICTION

Outline

- ❑ Motivation - at the verge of a new era of accelrated discoveries (10mins)

- ❑ Data types, from text to molecules – a quick intro/reminder (20 mins)
  - ❑ Coding text
  - ❑ Coding (small) molecules
  - ❑ Coding protiens

- ❑ Generative models – gentle introduction (20 mins)
  - ❑ Introduction: discreminative vs generative
  - ❑ Latent Dirichlet Allocation models – LDA models

- ❑ Generative models – Deep generative modesl (40 mins)
  - ❑ Variational Autoencoders – VAE models
  - ❑ Bidirectional Encoder Representations from Transformers– BERT
  - ❑ Reinforcement learning framing – RL

- ❑ Break

- ❑ Hands-on session (45 mins)

Variational autoencoder: mathematical foundation (1/2)

Given a set of examples

$$X = \{x_i\}_{i=1}^{N}$$

and a latent variable Z generate

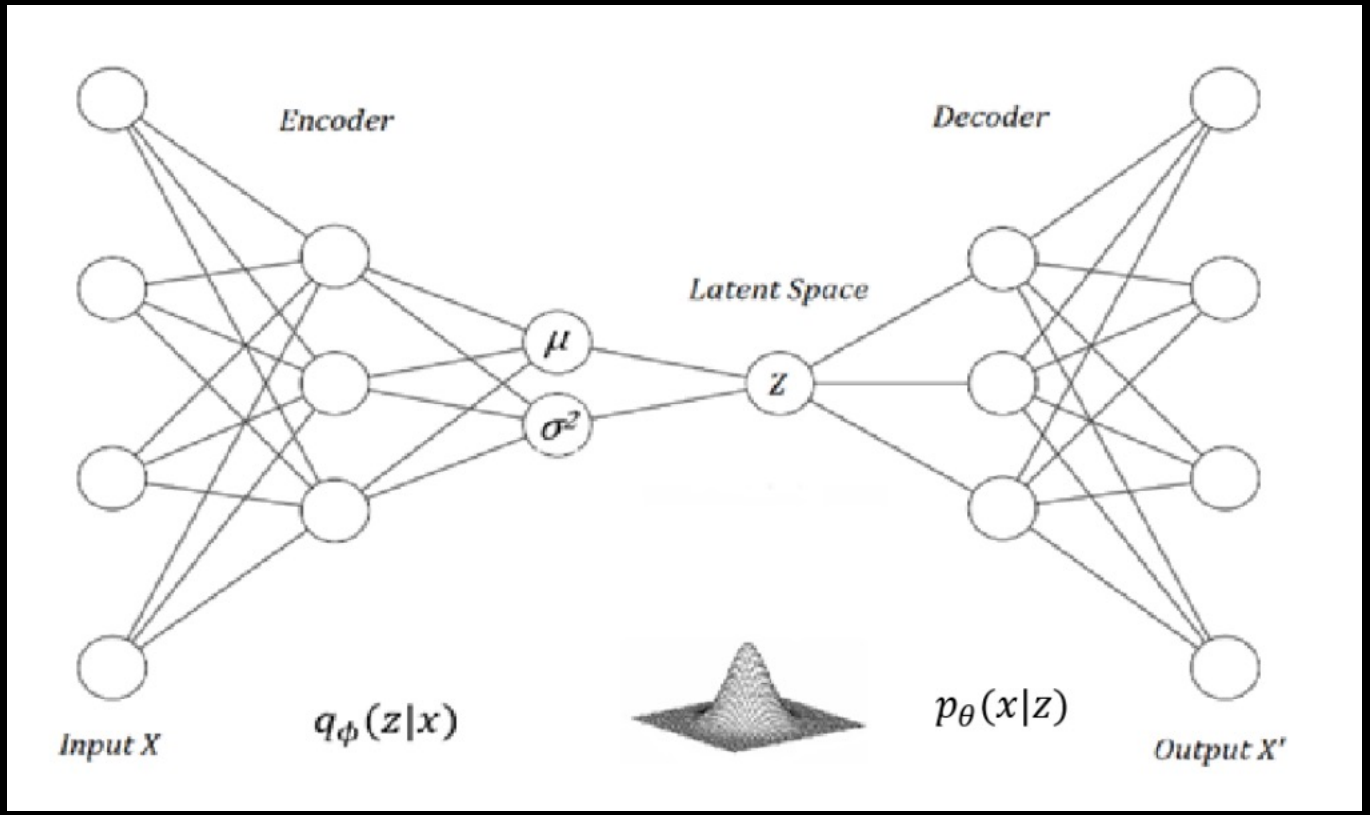
$$X' = \{x_i\}_{i=1}^{N}$$

Learns an approximation of the posterior

$$p_\theta(z|x) = p_\theta(x|z)p_\theta(z)/p_\theta(x)$$

and benefits from a generation process

$$p_\theta(x) = \int p_\theta(z)p_\theta(x|z)dz$$



R. Wei, A. Mahmood: Recent Advances in VAEs With Representation Learning for Biomedical Informatics: A Survey

Variational autoencoder: mathematical foundation (2/2)

$$D_{KL}\left[q\left(Z|X\right)\|p\left(Z|X\right)\right]$$

$$= \sum_{Z} q\left(Z|X\right)\log\left[\frac{q\left(Z|X\right)}{p\left(Z|X\right)}\right] \quad (1)$$

$$= \mathrm{E}\left[\log\left[\frac{q\left(Z|X\right)}{p\left(Z|X\right)}\right]\right] = \mathrm{E}\left[\log\left[q\left(Z|X\right)\right] - log[p\left(Z|X\right)]\right] \quad (2)$$

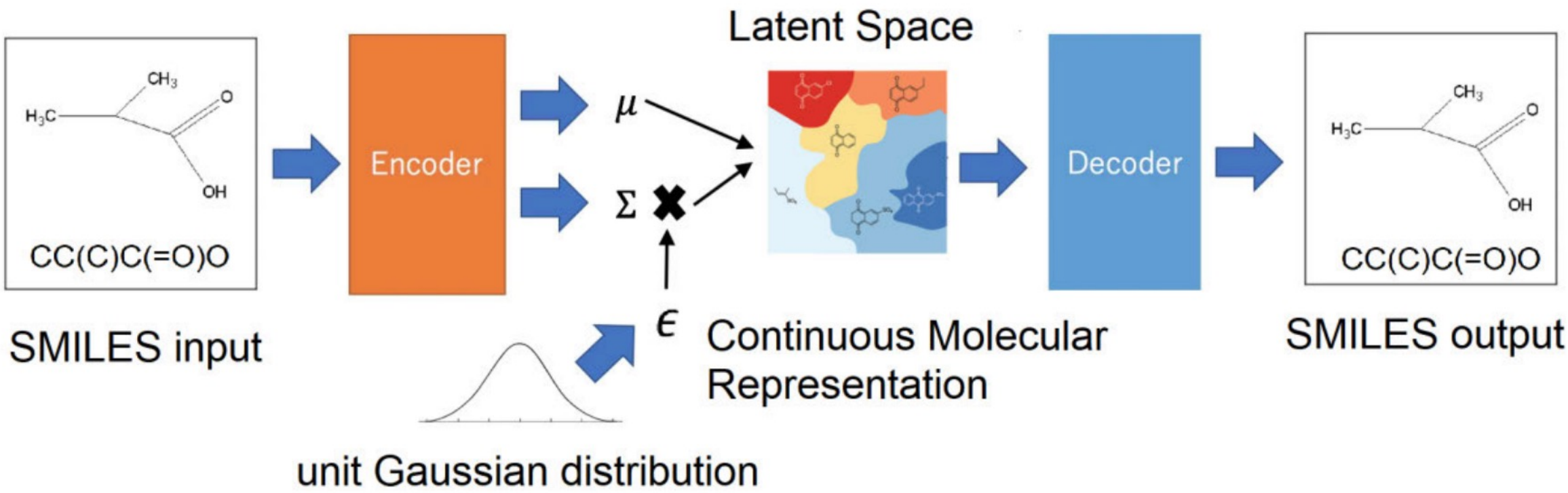
Since D_{KL} is always positive, we can conclude that:

$$\log p\left(X\right) \geq E[\log p\left(X|Z\right)] - D_{KL}\left[q\left(Z|X\right)\|p\left(Z\right)\right] \quad (3)$$

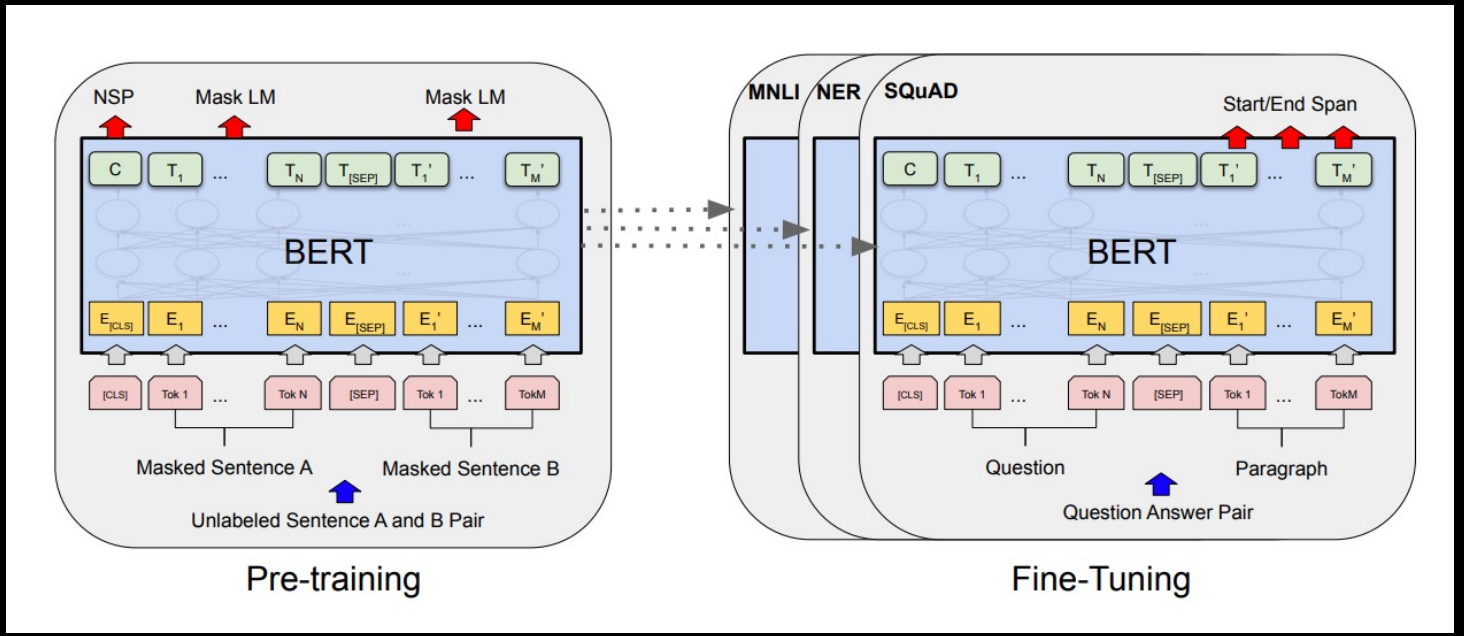
Eq. 3 is the Evidence Lower Bound (ELBO) method and serves as the loss function
q(Z|X) is a modeler choice – in what follows we assume multivariant unit Gaussian

R. Wei, A. Mahmood: Recent Advances in VAEs With Representation Learning for Biomedical Informatics: A Survey

Variational autoencoders for molecules:



Bidirectional Encoder Representations from Transformers



Given a sequence of words, processed from raw text, assume latent variables can embed the inforamtion successfully. Devote a unique seperator between senences [SEP] and at the beginning of each sequence a special classification token [CLS].

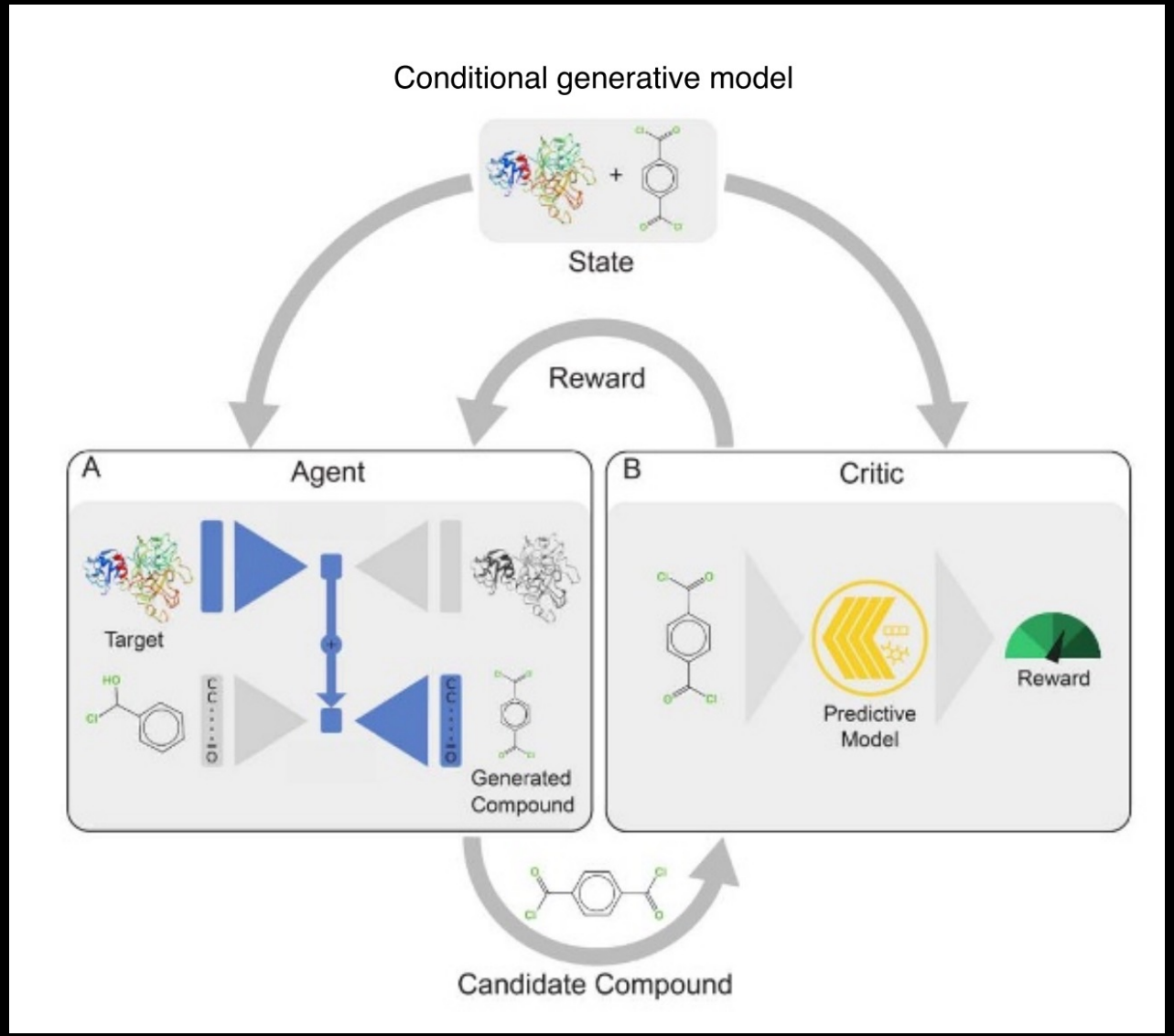
Designed to provide answers on two main tasks:

- Mask some percentage of the input tokens at random, and then predict those masked tokens “masked LM” (MLM),
- **Next Sentence Prediction (NSP)** Many important downstream tasks such as Question Answering and Natural Language Inference are based on understanding the relationship between two sentences

BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding

Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova

Reinforcement learning with generative models: Exploration-exploitation in the latent space (1/3)



Mach. Learn.: Sci. Technol. **2** (2021) 025024

Reinforcement learning with generative models: rewarding on molecules toxcity and bingding affinity (2/3)

Given a protien p and molecule m, the reward is

$$R(p,m) := \mathrm{Aff}(p,m) - \gamma \cdot \mathrm{Tox}(m)$$

Using VAE for molecules and protiens separately

$$\Theta_{\mathrm{Prot}} : \mathcal{P} \to \mathcal{P} \text{ and } \Theta_{\mathrm{Mol}} : \mathcal{M} \to \mathcal{M}$$

$$m = G_{\Theta}(p) = \Theta_{\mathrm{Mol}}^{\mathrm{Dec}}(\Theta_{\mathrm{Prot}}^{\mathrm{Enc}}(p))$$

Data type	Data source	Repr.	Model
Toxicity $\Phi_{\mathrm{Tox}} : \mathcal{M} \to \hat{y}_{\mathrm{Tox}}$	Tox21 database	Array of 12 binary indicators Moiecules: SMILE	11765 training Molecules; a multiscale conv. attention model
Proteins and molecules $\Phi_{\mathrm{Aff}} : \mathcal{P} \times \mathcal{M} \to [0,1]$	BindingDB	SMILE Proteins: AA seuqence	2273 726 samples; bimodal NN based on multiscale conv. attention model
Moclecules $\Theta_{\mathrm{Mol}} : \mathcal{M} \to \mathcal{M}$	ChEMBL	SELFIES strings, a robust adaption of SMILES	1576 904 bioactive compounds; VAE
Proteins $\Theta_{\mathrm{Prot}} : \mathcal{P} \to \mathcal{P}$	UniProt	TAPE (BERT like algorithm) pre-trained - 110 M parameters	404 552 proteins; VAE

Mach. Learn.: Sci. Technol. **2** (2021) 025024

Reinforcement learning with generative models: Generating new molecules and assisting them (3/3)

Given a protien p genearte a new molecule

$$\mathcal{P} \rightarrow Z \rightarrow \mathcal{M}$$

The conditional generator

$$G_\Theta : [\Theta^{\text{Dec}}_{\text{Mol}} \circ \Theta^{\text{Enc}}_{\text{Prot}}]$$

Metric for assessment includes

- Quantitative Estimate of Druglikeness (QED)
- Synthetic accessibility score
- Molecular weight

And more

Key results

- demonstrate the feasibility of swift chemical synthesis of molecules with potential antiviral
- antiviral candidate designed against the host protein (ACE2)

Mach. Learn.: Sci. Technol. **2** (2021) 025024

Next: a break & hands-on session

Case Study: Deep Generative Models for COVID 19 Drug Discovery

We will use the GT4SD framework to generate molecules to match a COVID-19 protein.

Compare two Variational Autoencoders:

- **PaccMannRL**: VAE is trained using reinforcement learning.
- **PaccMannGP**: Gaussian Processes used to search for a molecule in the latent space of the VAE.