I529: Machine Learning in Bioinformatics (B659 Topics in Artificial Intelligence)

Yuzhen Ye
School of Informatics, Computing and Engineering
Indiana University, Bloomington
Spring 2019

Overview

- **Prerequisites** I519 or equivalent knowledge in bioinformatics.
- **Grading:**
  - Combined assignments (35%), Final exam (20%), Paper presentation (10%), Class Project (35%)
  - Attendance will be considered in borderline cases.
- **Recommended textbook**:
  - Richard Durbin, Sean R. Eddy, Anders Krogh, and Graeme Mitchison, Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids , Cambridge University Press, 1999, (BSA)
- Online book: Neural Networks and Deep Learning (http://neuralnetworksanddeeplearning.com)

Work with large amounts of biological data

- Efficient information storage and management
- Extraction of useful information from these data
  - development of tools and methods capable of transforming heterogeneous data into biological knowledge about the underlying mechanism.

A few definitions

- Machine learning
  - A computer program is said to **learn** from experience *E* with respect to some class of **task**s *T* **and performance measure** *P*, if its performance at tasks in *T*, as measured by *P*, improves with experience *E* (T. Mitchell)
  - There are many different machine learning algorithms: supervised (classification) vs unsupervised (clustering); discriminative vs generative
- Probabilistic models
  - A model means a system that simulates the object under consideration
  - A probabilistic model is one that produces different outcomes with different probabilities (BSA)

Examples of learning problems

- A checkers learning problem
  - Task *T*: playing checkers
  - Performance measure *P*: percent of games won against opponents
  - Training experience *E*: playing practice games against itself.
- Handwriting recognition

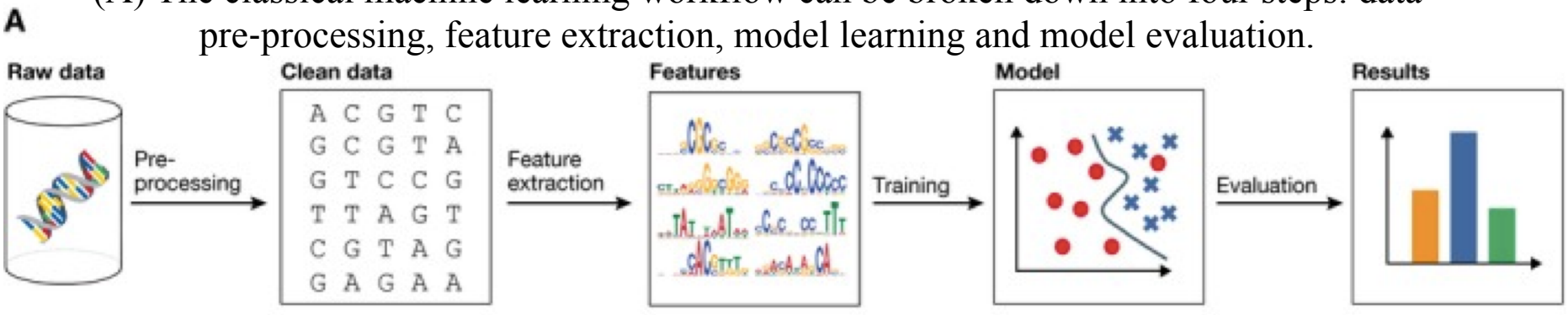
- Prediction of the viability of a cancer cell line when exposed to a chosen drug

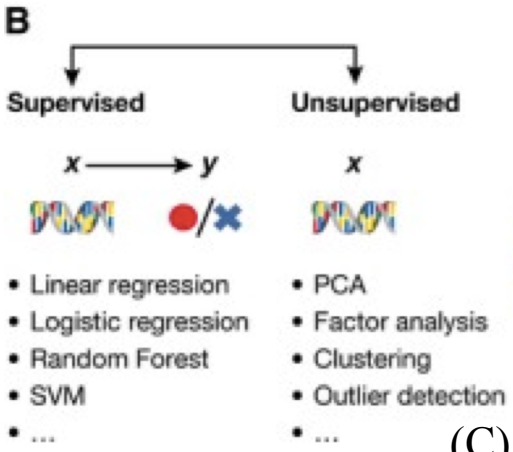
Designing/choosing a learning system

- Choosing the training experience (training data)
  - Does the training experience provide direct or indirect feedback regarding the choices made by the performance system
  - How well the training experience represents the distribution of examples?
- Choosing the target function
  - E.g., in checkers learning, *ChooseMove* is an obvious function, which accepts as input any legal board and produces as output some move from the set of legal moves. An alternative target function *V* (which will be easier to learn) assigns a numerical score to any given board state. So if the learning system can learn such a function *V*, then it can use it to select the best move from any current board position. When the target function is not efficiently computable, then the goal of learning is to discover an operational description of *V,* denoted as $\hat{V}$.
- Choosing a **representation** for the target function
  - E.g., $\hat{V}(b) = w_0 + w_1 x_1 + w_2 x_2 + w_3 x_3 + w_4 x_4 + w_5 x5 + w_6 x_6$
  - Where w_0 through w_6 are numerical coefficients (weights) to be chosen by the learning algorithm; x1 is the number of black pieces on the board, etc.
- Choosing an **algorithm** to find the weights

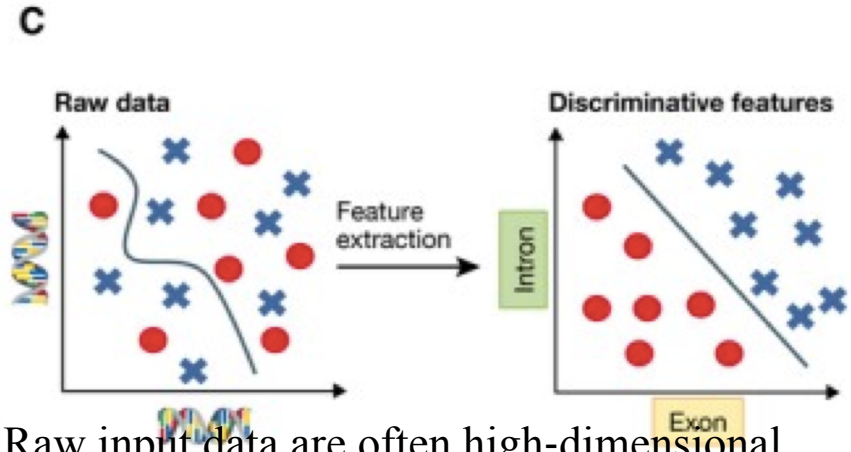
Machine learning and representation learning

(A) The classical machine learning workflow can be broken down into four steps: data pre-processing, feature extraction, model learning and model evaluation.

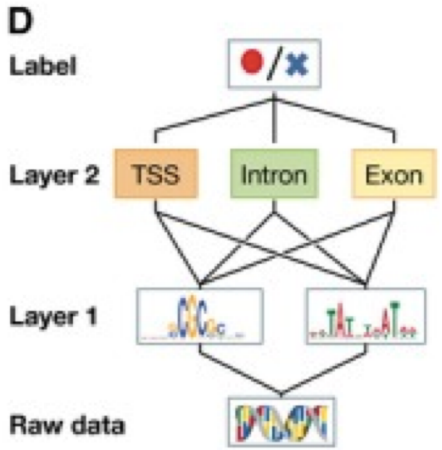




(B) Supervised machine learning methods relate input features x to an output label y.



(C) Raw input data are often high-dimensional and related to the corresponding label in a complicated way (left plot). Alternatively, higher-level features extracted using a deep model may be able to better discriminate between classes (right plot).



(D) Deep networks use a hierarchical structure to learn increasingly abstract feature representations from the raw data.

Ref: Mol Syst Biol. 2016 Jul; 12(7): 878.

An example of learning and prediction in bioinformatics

- Prediction of the viability of a cancer cell line when exposed to a chosen drug

The input features (x): somatic sequence variants of the cell line, chemical make-up of the drug and its concentration

Output label y: measured viability (output label y)

Learning: use (x, y) pairs to train a support vector machine, a random forest classifier, etc.

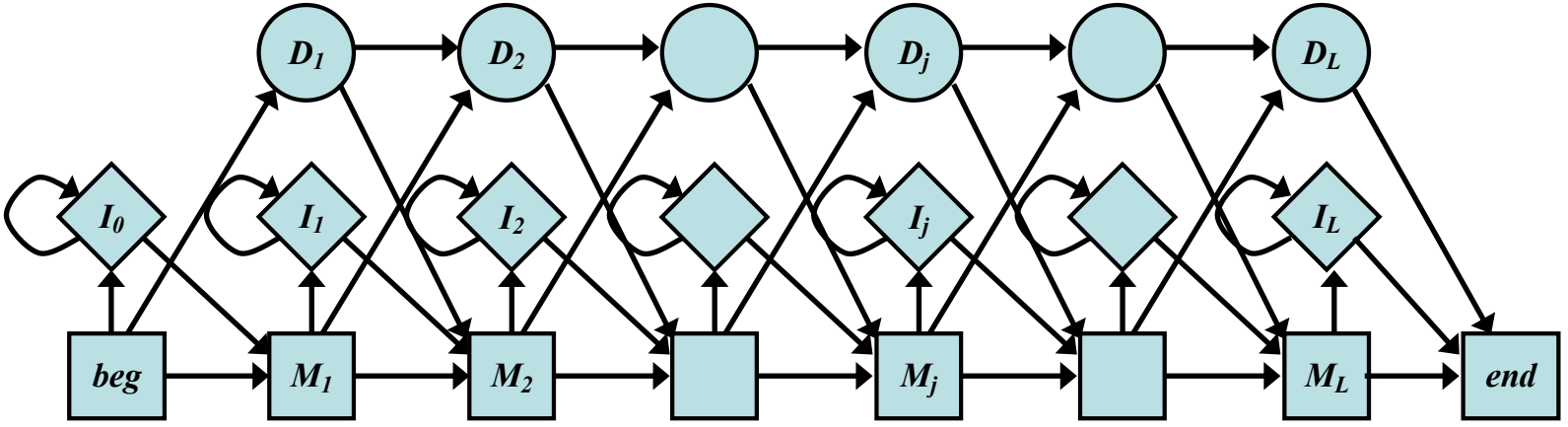
Prediction: given a new cell line (sample x*), the learnt function predicts its survival (output label y*) by calculating f(x*)

DNN and Deep learning

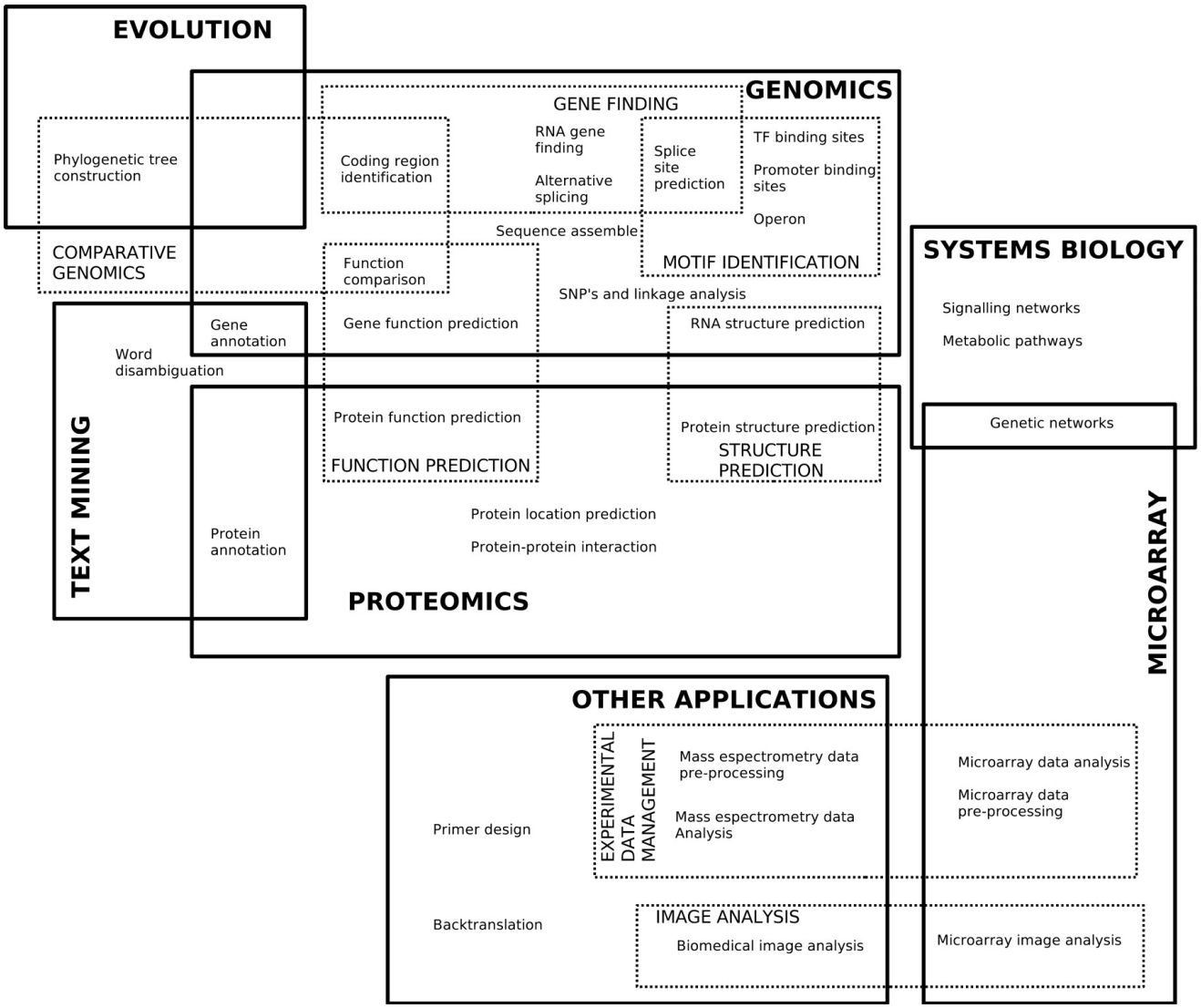
- A deep neural network (DNN) takes raw data at the lowest (input) layer and transforms them into increasingly abstract feature representations by successively combining outputs from the preceding layer in a **data-driven** manner
- Deep learning is now one of the most active fields in machine learning and has been shown to improve performance in image and speech recognition, and natural language understanding
- And hopefully in computational biology; Deep learning has been applied in regulatory genomics and biological image analysis
- Ref: Deep learning for computational biology. Mol Syst Biol. 2016 Jul; 12(7): 878.

An example of using probabilistic models: protein function annotation

- Each known protein family is presented as a HMM (a HMM is a probabilistic model that can generate different sequences with different probabilities)
- Given a new protein sequence, does it belong to one of known families (which model has the best chance of producing this sequence)?



Classification of the topics where ML methods are applied



Larrañaga P et al. Brief Bioinform 2006;7:86-112

Briefings in Bioinformatics

Types of data sets

- Record data
  - For the most basic form of record data, there is no explicit relationship among records or data fields, and every record (object) has the same set of attributes
- Graph-based data
  - Data with relationships among objects: the data objects are the nodes, and the relationships among objects are captured by the links between objects.
  - Data with objects that are graphs: Protein structures; small molecules
- Ordered data
  - Sequential data
  - Sequence data
  - Time series data
  - Spatial data

Data compression

- String compression
  - It is becoming a new challenging problem in Bioinformatics, because of the rapid development of sequencing techniques!!
  - Need developments of bioinformatics tools that work on compressed data
    - Most current bioinformatics tools only work on only uncompressed sequence data
    - So need to expand the data before using

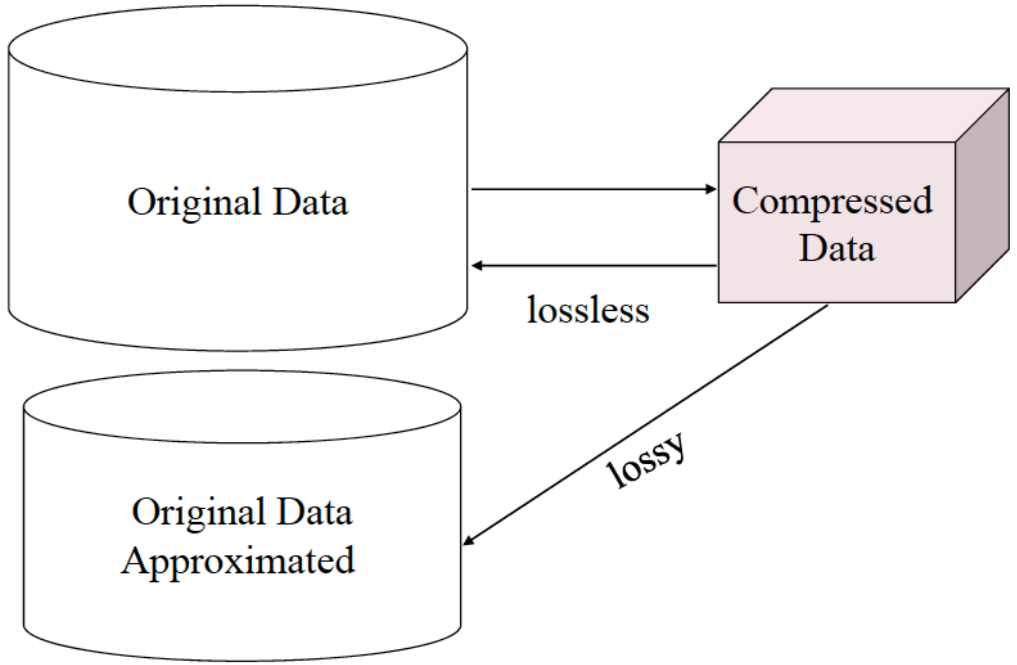
"Algorithms that compute directly on compressed genomic data allow analyses to keep pace with data generation"

Compressive Genomics, Nature Biotechnology 30, 627–630, 2012

Lossy or lossless compression

Lossless compression reduces bits by identifying and eliminating statistical redundancy. No information is lost in lossless compression.

Lossy compression reduces bits by identifying marginally important information and removing it.



Knowing your data

- Descriptive data summarization
  - Boxplot
  - Histogram
  - Quantile & Quantile-quantile plot
  - Scatter plot & Loess curve (add a smooth curve to scatter plot)
- Data preprocess/cleaning
  - Data can be incomplete, noisy, and inconsistent
  - No quality data, no quality mining
- Handle missing data
  - Exclude records with missing feature
  - Fill in manually
  - Fill in automatically (e.g., “UNK”, mean, most probable value).

Genomics and beyond

- Genomics
  - Study of the genomes of organisms
  - Problems: Gene finding; CpG island; motif finding
- Epigenomics
  - Study of the epigenomes; the epigenome of an organisms is the complete set of epigenetic modifications on its genetic material
  - “The cells in a multicellular organism have nominally identical DNA sequences.., yet maintain different terminal phenotypes. This **nongenetic** cellular memory, which records developmental and environmental cues.., is the basis of epi-(above)–genetics. “ (Science, *330 no. 6004 p. 611, 2010*)
  - Problems: chromotin state decoding; identification of combinatorial epigenetic regulation patterns; identification of DNA methylation states
- Metagenomics
  - Study of the genomic sequences of an entire microbial community
  - Problems: classification of 16S rRNA reads and shotgun sequences
- Systems biology
  - Systems biology is the study of systems of biological components, which may be molecules, cells, organisms or entire species.
  - Problems: integration of inhomogeneous data for function prediction and for the inference of cellular pathways

Focus of I529

- Machine learning approaches (supervised and unsupervised approaches) and recent advances (esp. deep networks).
- Probabilistic models (Markov models, Hidden Markov models, and Bayesian networks) for biological sequence analysis and systems biology.
- Biological problems & applications of the ML approaches to genomics and beyond (e.g., metagenomics, and epigenomics)