

CSEP 546
Data Mining
Machine Learning

Instructor: Pedro Domingos

Logistics

- **Instructor:** Pedro Domingos
 - Email: pedrod@cs
 - Office: CSE 648
 - Office hours: Mondays 5:30-6:20
- **TAs:** Kenton Lee, Alon Milchgrub
 - Email: kentonl@cs, alonmil@cs
 - Office: CSE TBD
 - Office hours: Mondays 5:30-6:20
- **Web:** www.cs.washington.edu/csep546

Evaluation

- Four assignments (25% each)
 - Handed out on weeks 2, 4, 6 and 8
 - Due two weeks later
 - Mix of:
 - Implementing machine learning algorithms
 - Applying them to real datasets (e.g.: clickstream mining, recommender systems, spam filtering)
 - Exercises

Source Materials

- T. Mitchell, ***Machine Learning***, McGraw-Hill (*Required*)
- R. Duda, P. Hart & D. Stork, ***Pattern Classification*** (2nd ed.), Wiley (*Required*)
- P. Domingos, ***The Master Algorithm***, Basic Books (*Recommended*)
- Papers

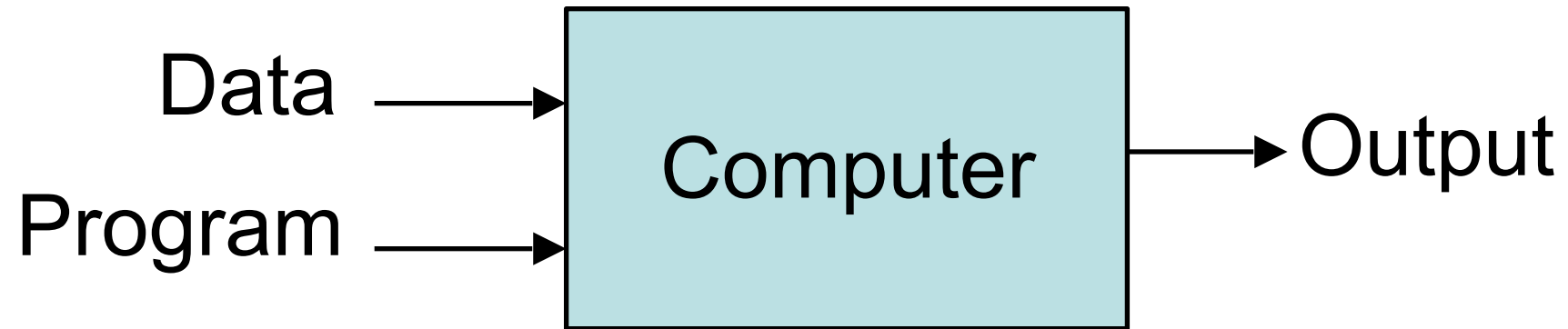
A Few Quotes

- “A breakthrough in machine learning would be worth ten Microsofts” (Bill Gates, Founder, Microsoft)
- “Machine learning is the next Internet” (Tony Tether, Director, DARPA)
- “Machine learning is the hot new thing” (John Hennessy, President, Stanford)
- “Machine learning is Google’s top priority” (Eric Schmidt, Chairman, Alphabet)
- “Machine learning is Microsoft Research’s largest investment area” (Peter Lee, Head, Microsoft Research)
- “Machine learning is the single most important technology trend” (Steve Jurvetson, Partner, Draper Fisher Jurvetson)
- “‘Data scientist’ is the hottest job title in Silicon Valley” (Tim O’Reilly, Founder, O’Reilly Media)

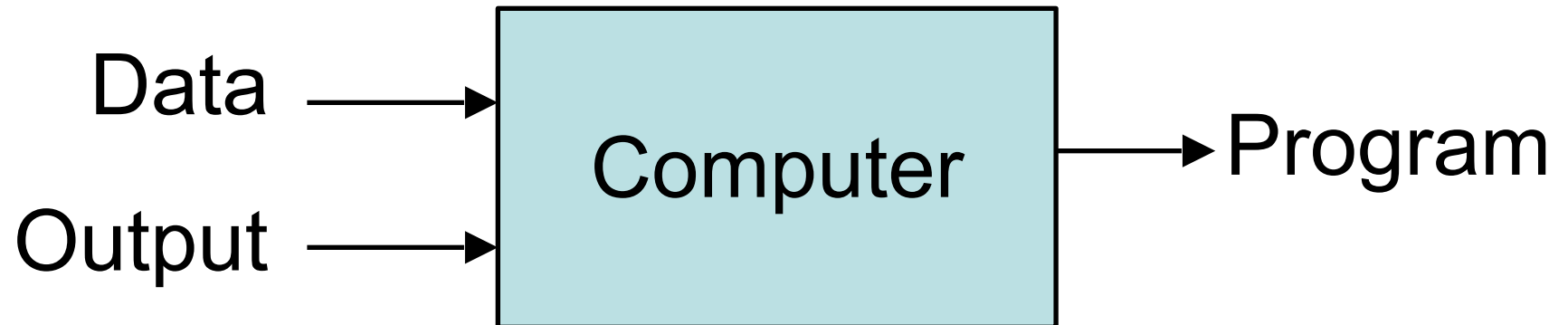
So What Is Machine Learning?

- Automating automation
- Getting computers to program themselves
- Writing software is the bottleneck
- Let the data do the work instead!

Traditional Programming



Machine Learning



Magic?

No, more like gardening

- **Seeds** = Algorithms
- **Nutrients** = Data
- **Gardener** = You
- **Plants** = Programs



Sample Applications

- Web search
- Computational biology
- Finance
- E-commerce
- Space exploration
- Robotics
- Information extraction
- Social networks
- Debugging
- [Your favorite area]

ML in a Nutshell

- Tens of thousands of machine learning algorithms
- Hundreds new every year
- Every machine learning algorithm has three components:
 - **Representation**
 - **Evaluation**
 - **Optimization**

Representation

- Decision trees
- Sets of rules / Logic programs
- Instances
- Graphical models (Bayes/Markov nets)
- Neural networks
- Support vector machines
- Model ensembles
- Etc.

Evaluation

- Accuracy
- Precision and recall
- Squared error
- Likelihood
- Posterior probability
- Cost / Utility
- Margin
- Entropy
- K-L divergence
- Etc.

Optimization

- Combinatorial optimization
 - E.g.: Greedy search
- Convex optimization
 - E.g.: Gradient descent
- Constrained optimization
 - E.g.: Linear programming

Types of Learning

- **Supervised (inductive) learning**
 - Training data includes desired outputs
- **Unsupervised learning**
 - Training data does not include desired outputs
- **Semi-supervised learning**
 - Training data includes a few desired outputs
- **Reinforcement learning**
 - Rewards from sequence of actions

Inductive Learning

- **Given** examples of a function $(X, F(X))$
- **Predict** function $F(X)$ for new examples X
 - Discrete $F(X)$: Classification
 - Continuous $F(X)$: Regression
 - $F(X) = \text{Probability}(X)$: Probability estimation

What We'll Cover

- **Supervised learning**
 - Decision tree induction
 - Rule induction
 - Instance-based learning
 - Bayesian learning
 - Neural networks
 - Support vector machines
 - Model ensembles
 - Learning theory
- **Unsupervised learning**
 - Clustering
 - Dimensionality reduction

ML in Practice

- Understanding domain, prior knowledge, and goals
- Data integration, selection, cleaning, pre-processing, etc.
- Learning models
- Interpreting results
- Consolidating and deploying discovered knowledge
- Loop