

Data Mining and Machine Learning
CompSci 760

Lecture 21

**Discovering Qualitative
and Quantitative Laws**

Descriptive Knowledge in Science

We can divide descriptive knowledge in the sciences into three broad categories:

- *Taxonomic knowledge*, which defines classes of entities and organizes them in a hierarchy;
- *Qualitative laws*, which state generalizations about classes of entities and about relations among them; and
- *Quantitative laws*, which specify numeric equations that relate attributes/variables of these classes.

Numeric laws always occur in a qualitative context, which in turn depends on a taxonomy, even if not stated explicitly.

Different processes underlie discovery of these knowledge types.

Qualitative Descriptive Laws

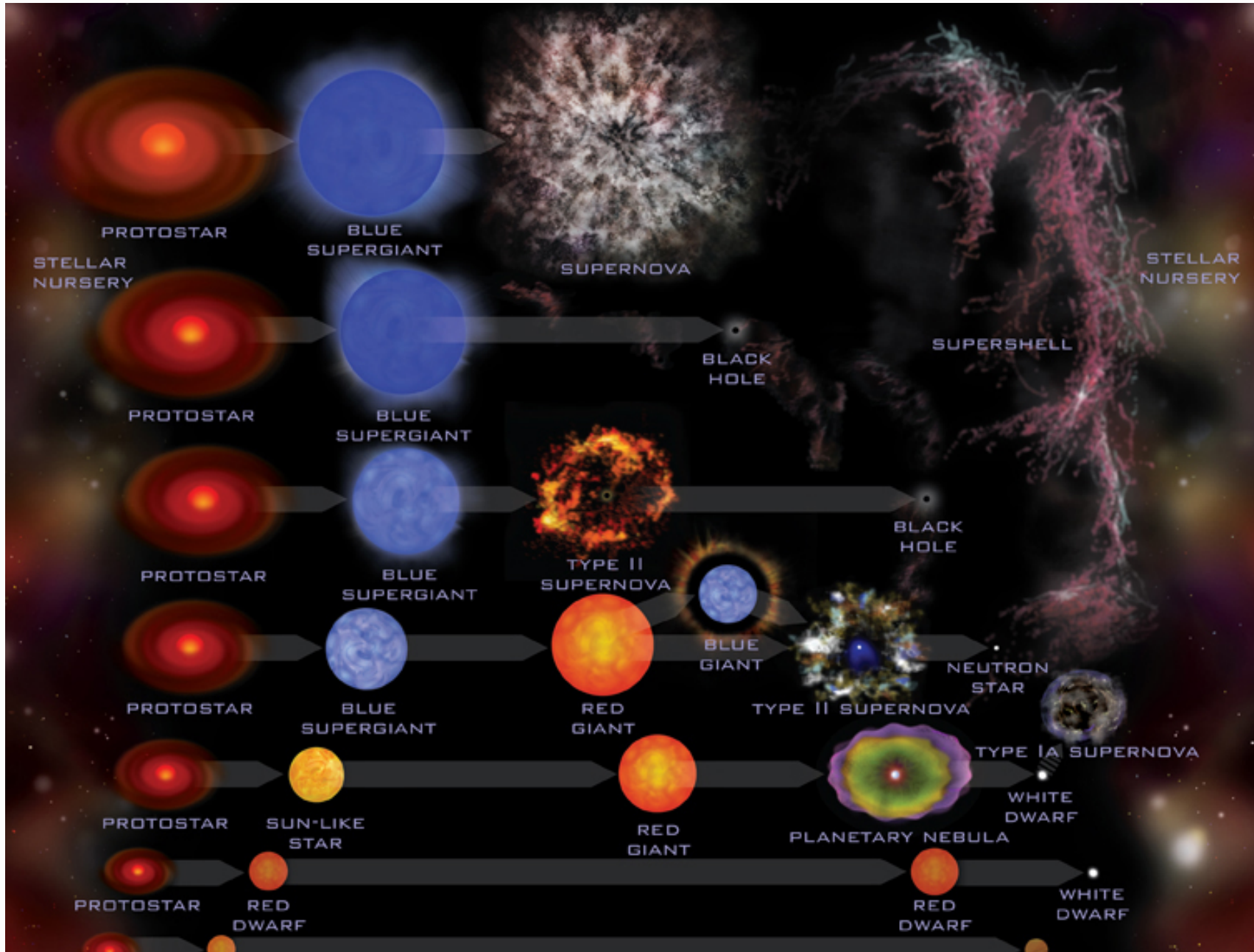
A scientific *law* states some regularity that occurs among instances of categories and/or attributes.

A *qualitative* law takes the form of some symbolic relation or rule, such as:

- Some planets follow retrograde trajectories.
- Zooplankton feed on phytoplankton.
- Acids react with alkalis to form salts.
- Chemicals with certain features produce cancer.
- American families include two parents and their children.

These qualitative laws *describe* recurring relationships, but they do not *explain* them.

Qualitative Laws About Stars



Finding Qualitative Laws

Statement of the task:

- *Given*: Qualitative data about objects or events that occur in the world.
- *Find*: General relations that hold among these classes of items that predict future behavior.

Historical examples:

- Qualitative motions of stars through the night and year
- Recurring reactions between chemicals and chemical groups
- Predator/grazing relations among animals and plants
- Regularities in observed tracks of elementary particles

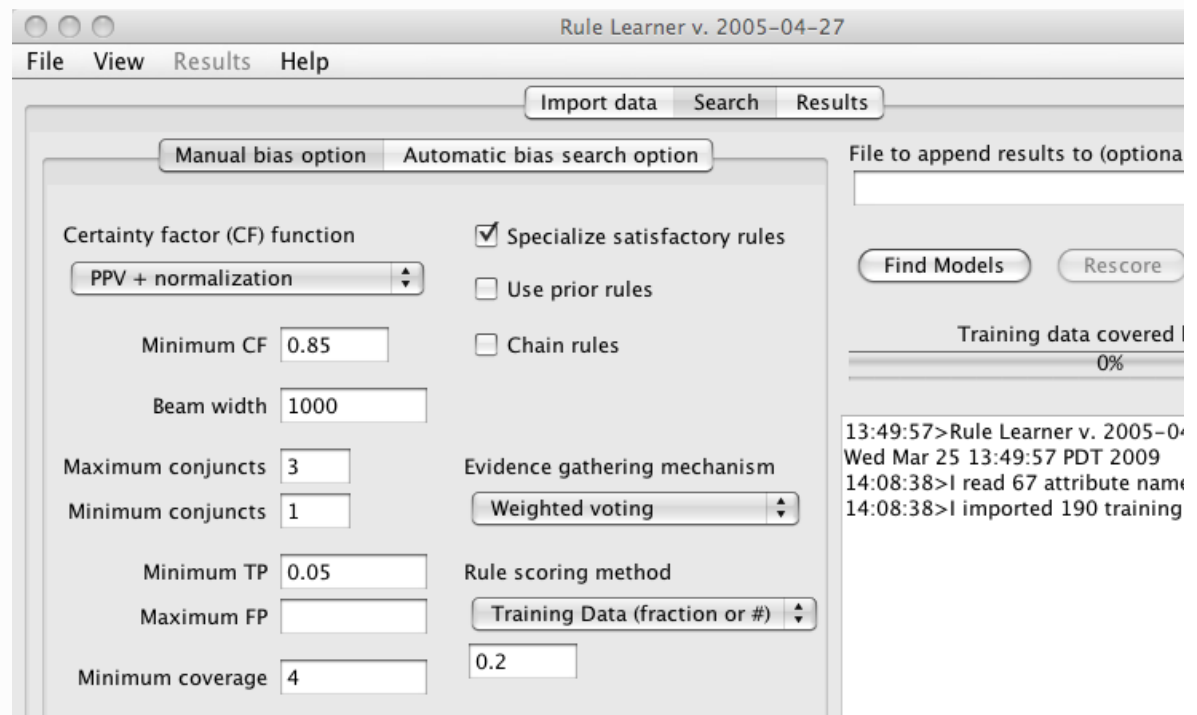
The RL System

One system of this sort – Lee et al.’s RL (1998) – adapts rule induction to find qualitative relations.

Each rule states that, if certain conditions hold for an entity or situation, then the class variable has a certain value.

As input, RL takes a classified training data and details about:

- a hierarchy over attributes’ values
- constraints among properties in rules
- min. rule accuracy
- max. rule properties



Applications of RL

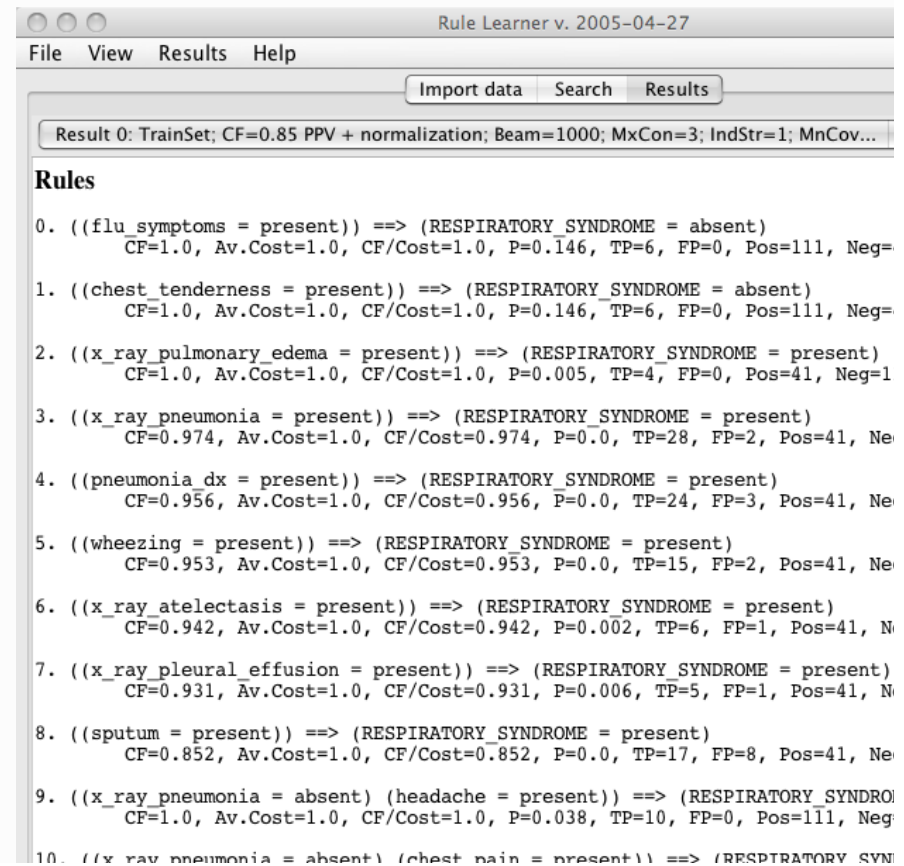
As an example, consider the discovery of law-like relations that link symptoms to disease.

The data are patient findings and the target is a set of rules that together, predict ailments (e.g., lower respiratory syndrome).

Each such rule has a numeric measure of support in the data.

RL has been applied to:

- classifying respiratory syndromes
- identifying carcinogens
- predicting when crystals will form



```
Rule Learner v. 2005-04-27
File View Results Help
Import data Search Results
Result 0: TrainSet; CF=0.85 PPV + normalization; Beam=1000; MxCon=3; IndStr=1; MnCov...
Rules
0. ((flu_symptoms = present)) ==> (RESPIRATORY_SYNDROME = absent)
   CF=1.0, Av.Cost=1.0, CF/Cost=1.0, P=0.146, TP=6, FP=0, Pos=111, Neg=
1. ((chest_tenderness = present)) ==> (RESPIRATORY_SYNDROME = absent)
   CF=1.0, Av.Cost=1.0, CF/Cost=1.0, P=0.146, TP=6, FP=0, Pos=111, Neg=
2. ((x_ray_pulmonary_edema = present)) ==> (RESPIRATORY_SYNDROME = present)
   CF=1.0, Av.Cost=1.0, CF/Cost=1.0, P=0.005, TP=4, FP=0, Pos=41, Neg=1
3. ((x_ray_pneumonia = present)) ==> (RESPIRATORY_SYNDROME = present)
   CF=0.974, Av.Cost=1.0, CF/Cost=0.974, P=0.0, TP=28, FP=2, Pos=41, Ne
4. ((pneumonia_dx = present)) ==> (RESPIRATORY_SYNDROME = present)
   CF=0.956, Av.Cost=1.0, CF/Cost=0.956, P=0.0, TP=24, FP=3, Pos=41, Ne
5. ((wheezing = present)) ==> (RESPIRATORY_SYNDROME = present)
   CF=0.953, Av.Cost=1.0, CF/Cost=0.953, P=0.0, TP=15, FP=2, Pos=41, Ne
6. ((x_ray_atelectasis = present)) ==> (RESPIRATORY_SYNDROME = present)
   CF=0.942, Av.Cost=1.0, CF/Cost=0.942, P=0.002, TP=6, FP=1, Pos=41, N
7. ((x_ray_pleural_effusion = present)) ==> (RESPIRATORY_SYNDROME = present)
   CF=0.931, Av.Cost=1.0, CF/Cost=0.931, P=0.006, TP=5, FP=1, Pos=41, N
8. ((sputum = present)) ==> (RESPIRATORY_SYNDROME = present)
   CF=0.852, Av.Cost=1.0, CF/Cost=0.852, P=0.0, TP=17, FP=8, Pos=41, Ne
9. ((x_ray_pneumonia = absent) (headache = present)) ==> (RESPIRATORY SYNDRO
   CF=1.0, Av.Cost=1.0, CF/Cost=1.0, P=0.038, TP=10, FP=0, Pos=111, Neg
10. ((x_ray_pneumonia = absent) (chest_pain = present)) ==> (RESPIRATORY SYN
```

The Glauber System

Glauber (Langley et al., 1987) rediscovers qualitative regularities from the history of chemistry; it takes as input:

- Qualitative features of substances, such as their taste;
- Chemical reactions in which the substances participate.

As output, the system produces two linked forms of knowledge:

- *Classes* of substances that have similar properties
- Qualitative *laws* that summarize relations among classes

Thus, Glauber integrates, in a primitive way, taxonomy formation and law discovery.

Glauber on Acids, Alkalis, and Salts

Initial state:

(reacts in {HCl NaOH} out {NaCl})

(reacts in {HCl KOH} out {KCl})

(reacts in {HNO₃ NaOH} out {NaNO₃})

(reacts in {HNO₃ KOH} out {KNO₃}) ...

(has object {HCl} taste {sour})

(has object {NaOH} taste {bitter})

(has object {NaCl} taste {salty}) ...

Final state:

salt: {NaCl, KCl, ...} acid: {HCl, HNO₃, ...} ...

\forall salt (has object {salt} taste {salty})

\forall acid (has object {acid} taste {sour})

\forall alkali (has object {alkali} taste {bitter})

\forall acid \forall alkali \exists salt (reacts in {acid alkali} out {salt})

Additional Work on Qualitative Discovery

Other systems for discovering qualitative laws include:

- AM (Lenat, 1977) in number theory
- Meta-Dendral (Buchanan & Mitchell, 1978) in org. chemist
- IDS (Nordhausen & Langley, 1993) in physics
- PROGOL (Muggleton et al., 1996, 1998) in biochemistry
- HR (Colton, 1997) in various branches of mathematics

Much of this work has aimed at novel discoveries rather than a reconstruction of historical ones.

There has been substantial work on qualitative discovery in the field of molecular biology.

Quantitative Descriptive Laws

A *quantitative* law takes the form of some numeric relationship typically stated as an equation.

Quantitative scientific laws fall into two broad categories:

- *Algebraic equations*, such as Coulomb's law ($F = kQ_1Q_2/D^2$) and the ideal gas law ($PV = aNT + bN$); and
- *Differential equations* that describe dynamic behavior, such as population growth over time ($dP/dt = kP$).

Any quantitative law always occurs within the context of some qualitative law, even if the former is not stated explicitly.

What is a Causal Relation?

We can define causality in abstract but unambiguous terms; we will say that variable X *causally influences* variable Y if:

- a change in X 's value results in a change to Y 's value
- provided that other variables are held constant

Note that this definition of causality does not mention:

- that X is the only causal influence on Y
- the directionality of this influence
- the functional form of the causal relation

Many quantitative scientific laws describe causal relationships but not all have this character.

Discovering Algebraic Laws

Statement of the task:

- *Given*: Quantitative measurements about objects or events in the world.
- *Find*: Numeric relations that hold among variables that describe these items and that predict future behavior.

Historical examples:

- Kepler's three laws of planetary motion
- Archimedes' principle of displacement in water
- Black's law relating specific heat, mass, and temperature
- Proust's and Gay-Lussac's laws of definite proportions

Regression Equations

Perhaps the most basic type of numeric law involves regression equations, which include:

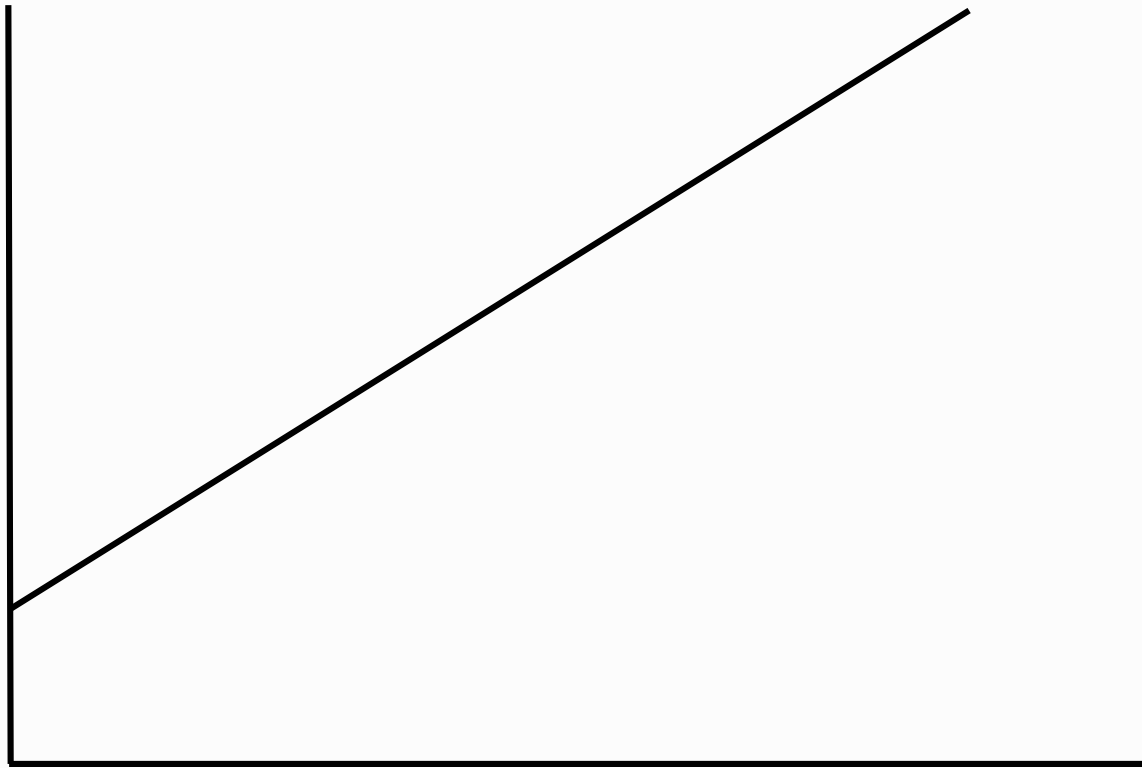
- univariate linear equations (e.g., $D = a \times I + b$)
- complex univariate equations (e.g., $D = a \times I^2 + b \times I + c$)
- multivariate linear equations ($D = a \times I_1 + b \times I_2 + \dots j \times I_N +$

These are all designed to model *static* situations; we will delay discussion of dynamic situations until later.

Not all regression equations have a causal interpretation, even though it is tempting to interpret them this way.

Visualizing Regression Equations

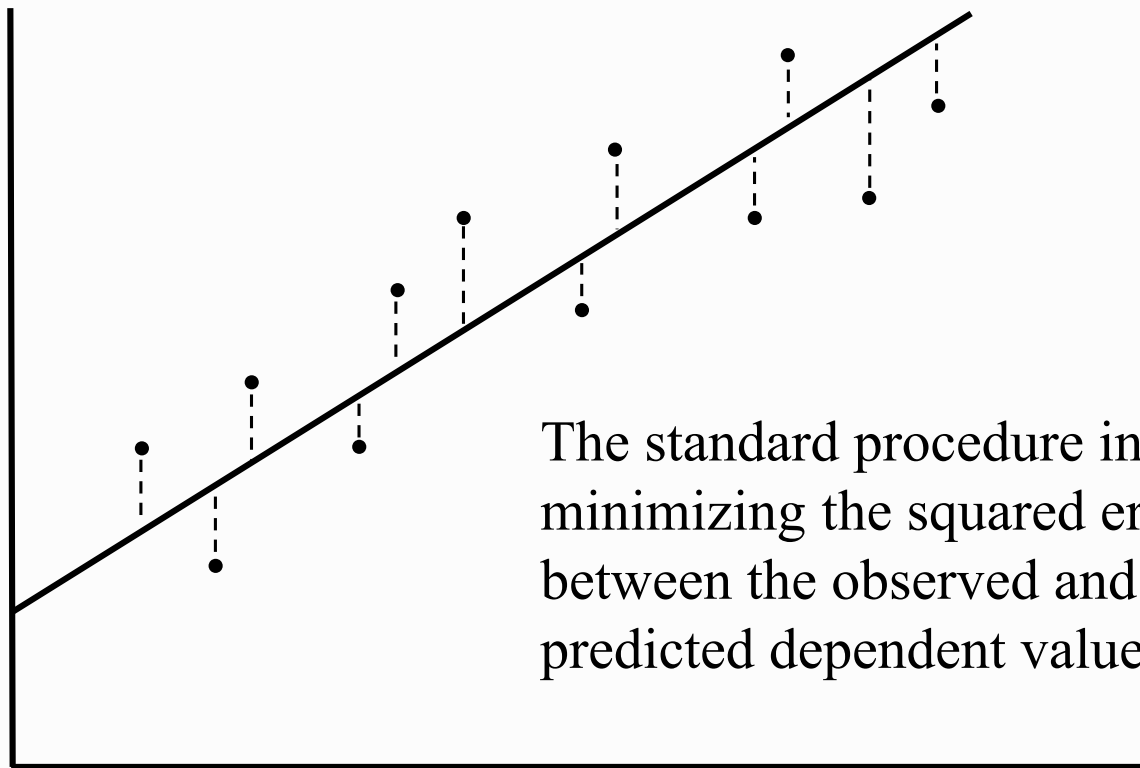
We can easily visualize the relationship $D = a \times I + b$ specified in a univariate linear model.



In this graph, a is the *slope* of the line and b is the *intercept*; either parameter may be positive or negative.

Fitting Linear Equations to Data

There exist well-established methods for determining the best parameters in a linear equation for a given data set.



However, even multivariate linear equations are very restrictive in the kinds of laws they can specify.

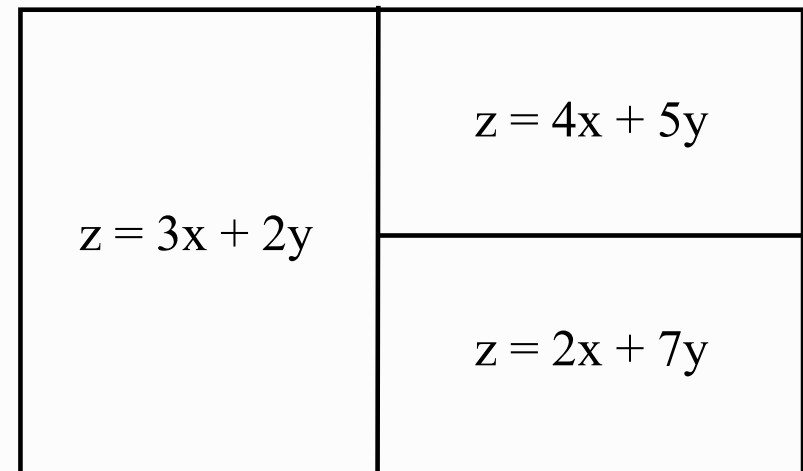
Inducing Regression Trees

One extension to linear regression involves adapting methods of decision-tree induction.

For instance, a regression tree has the structure of a decision tree but stores multivariate linear equations at terminal nodes.

These comprise a set of complementary numeric laws that hold under different conditions.

A regression tree partitions the instance space into mutually exclusive regions, with a different linear equation used to predict the dependent variable in each.



Quinlan's (2001) Cubist system embeds linear regression with

Complex Multivariate Models

However, even regression trees are quite different from many quantitative relations in the sciences.

These laws often have more complicated multivariate forms:

- The ideal gas law / $V = (aNT + bN) / P$
- Coulomb's law / $F = kQ_1Q_2/D^2$
- Black's law / $T_f = (c_1m_1T_1 + c_2m_2T_2) / (c_1m_1 + c_2m_2)$
- Snell's law of refraction / $\sin r = c_2 / (c_2 \sin i)$

Many of these were discovered through systematic experiment that varied one attribute at a time.

The Bacon System

Bacon (Langley, 1981) rediscovers quantitative laws of this type from the history of physics; it takes as input:

- A set of dependent and independent variables
- Measurements from a set of controlled ‘experiments’
- Heuristics to guide its search for candidate laws

The system finds simple laws that relate one dependent variable to one independent variable.

Bacon’s heuristics are stated as symbolic rules that specify the conditions for defining new terms.

For instance, if X decreases with Y , then define a term $Z = X/Y$

Bacon on Kepler's Third Law

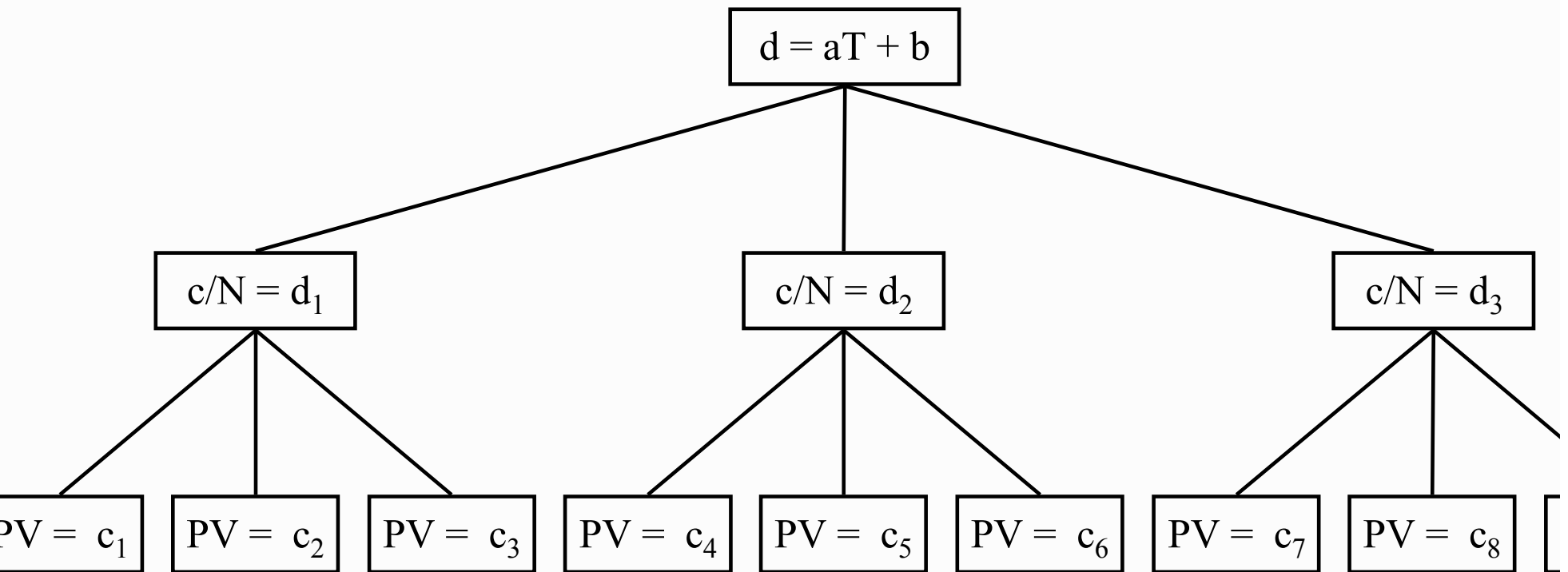
Bacon carries out heuristic search, through a space of numerical terms, looking for constants and linear relations.

moon	d	p	d/p	d ² /p	d ³ /p ²
A	5.67	1.77	3.20	18.15	58.15
B	8.67	3.57	2.43	21.04	51.06
C	14.00	7.16	1.96	27.40	53.61
D	24.67	16.69	1.48	36.46	53.89

This example shows the system's progression from primitive variables (distance and period of Jupiter's moons) to a complex term that has a nearly constant value.

Rediscovery of the Ideal Gas Law

Bacon rediscovers the ideal gas law, $PV = aNT + bN$, in three stages, each at a different level of description.



Parameters for laws at one level become dependent variables in laws at the next level, enabling discovery of complex relations

Some Laws Discovered by Bacon

Basic numeric relations:

- Ideal gas law $PV = aNT + bN$
- Kepler's third law $D^3 = [(A - k) / t]^2 = j$
- Coulomb's law $FD^2 / Q_1Q_2 = c$
- Ohm's law $TD^2 / (LI - rI) = r$

Relations with *intrinsic properties*:

- Snell's law of refraction $\sin I / \sin R = n_1 / n_2$
- Archimedes' law $C = V + i$
- Momentum conservation $m_1V_1 = m_2V_2$
- Black's specific heat law $c_1m_1T_1 + c_2m_2T_2 = (c_1m_1 + c_2m_2)T$

The RF6 Algorithm

Saito and Nakano (2000) describe RF6, a discovery system that

1. Creates a multilayer neural network that links predictive with predicted variables using additive and product units.
2. Invokes the BPQ algorithm to search through the weight space defined by this network.
3. Transforms the resulting network into a polynomial equation of the form $y = \sum c_i \prod x_j^{d_{ij}}$.

They have shown this approach can discover an impressive class of numeric equations from noisy data.

Their results counter the common assumption that neural network methods produce uninterpretable structures.

Discovering Dynamic Laws

Statement of the task:

- *Given*: Quantitative measurements about objects or events as they change over time.
- *Find*: Numeric equations that describe the variables' dynamic behavior and that predict future trajectories.

Examples of dynamic laws include:

- Population change
- Predator-prey ecosystems
- Power generation and use
- Biochemical kinetics

Difference Equations

The simplest form of dynamic quantitative model involves a single *difference equation* or *recurrence equation*.

This takes a form similar to an algebraic equation, but specifies some attribute's *change* as the dependent variable.

For example, consider the simple difference equation:

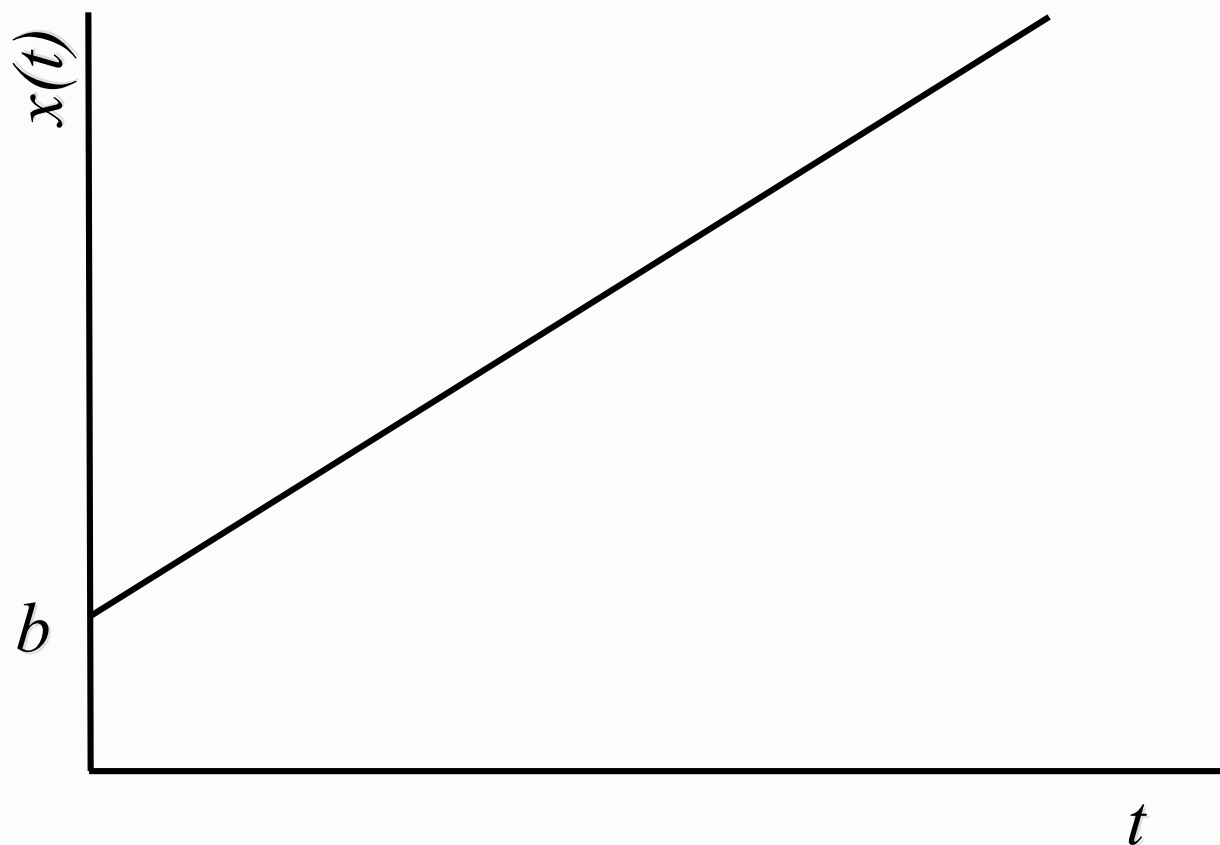
$$x(t + 1) - x(t) = a ,$$

which describes a system with a single variable, x , whose value increases by the constant value a on each time step.

Difference equations assume that time passes in fixed step size which makes them easy to interpret.

Visualizing Difference Equations

We can visualize the relationship $x(t + 1) - x(t) = a$ by plotting how the variable x changes over time.



In this graph, a is the *slope* of the line and b is the *intercept*, which is determined by the initial value of x .

Ordinary Differential Equations

A closely related form of dynamic quantitative model involves *ordinary differential equation*.

This takes a form similar to a difference equation, but specifies an attribute's *derivative* as the dependent variable.

The analog of the earlier difference equation would be:

$$dx/dt = a ,$$

which describes a system with a single variable, x , whose value increases at the constant rate a .

Differential equations assume continuous time, which can lead to different results in regions that involve high rates of change.

Solving Difference/Differential Equations

To use a difference/differential equation for prediction, we must *solve* for the dependent variable as a function of time.

Some equations can be solved analytically using mathematical techniques, but many require *simulation*.

- Because they assume discrete time, simulation of difference equations is straightforward.
- Because they assume continuous time (and space), simulation of differential equations is more challenging.

The field of *numerical analysis* emphasizes the latter, selecting step size and minimizing round-off error.

We will assume that such techniques are given and focus on their use in simulating dynamic systems.

The LAGRAMGE System

LAGRAMGE is a discovery program that finds quantitative laws that describe dynamical systems; it takes as input:

- time series observations for a set of variables
- an indication of which variables are dependent
- a context-free grammar that specifies acceptable equations

The system uses depth-limited search to induce one algebraic or differential equation for each dependent variable.

Estimating parameters involves simulating candidate structures.

Todorovski and Dzeroski (1997) have applied LAGRAMGE to ecosystem dynamics, hydrodynamics, and other areas.

Temporal Laws of Ecological Behavior

(Todorovski & Dzeroski, 1997)

Input:

time	phyt	zoo	phosp	temp
time ₁	phyt ₁	zoo ₁	phosp ₁	temp ₁
time ₂	phyt ₂	zoo ₂	phosp ₂	temp ₂
⋮	⋮	⋮	⋮	⋮
time _m	phyt _m	zoo _m	phosp _m	temp _m

Input:

a context-free grammar of domain constraints

Output:

$$\dot{phyt} = c_1 \cdot phyt \cdot \frac{phosp}{c_2 + phosp} - c_3 \cdot phyt$$

Additional Work on Numeric Discovery

Other systems for discovering numeric laws include:

- ABACUS (Falkenhainer, 1985) and ARC (Moulet, 1992)
- Fahrenheit (Zytkow, Zhu, & Hussam, 1990)
- COPER (Kokar, 1986) and E* (Schaffer, 1990)
- IDS (Nordhausen & Langley, 1990)
- Hume (Gordon & Sleeman, 1992)
- DST (Murata, Mizutani, & Shimura, 1994)
- SSF (Washio & Motoda, 1997)
- Genetic programming (Koza, 2001)

These rely on different methods but share a concern with finding explicit mathematical laws from data.

Summary Remarks

There has been a long history of work on computational scientific discovery of descriptive laws, including:

- Qualitative relations stated as rules or logical statements
- Quantitative laws stated as numeric equations

Such systems produce an important type of knowledge, but they

- can lead to shallow interpretations of data;
- typically avoid statements of causality or process;
- make little contact with a discipline's theoretical content

In future lectures, we will discuss systems that move beyond description to provide *explanations*.