**Day 1:**

**Motivation:**

We are drowning in data, but starving for knowledge!

• Data = raw information

• Knowledge = set of patterns or models underlying the data

**Define:**

Hypothesis: pre-existing data repositories contain a lot of potentially important information

Mission of ML: ﬁnd it

**ML: automatic extraction of valid, novel, useful and comprehensible knowledge (rules, regularities, patterns, constraints, models, ...) from arbitrary sets of data**

**Tasks:**

Our main tasks (depending on data set & problem)

• Classiﬁcation

• Clustering

• Regression

• Sequence Discovery

• Association Rule Mining

• Outlier Detection • ...

**Day 2:**

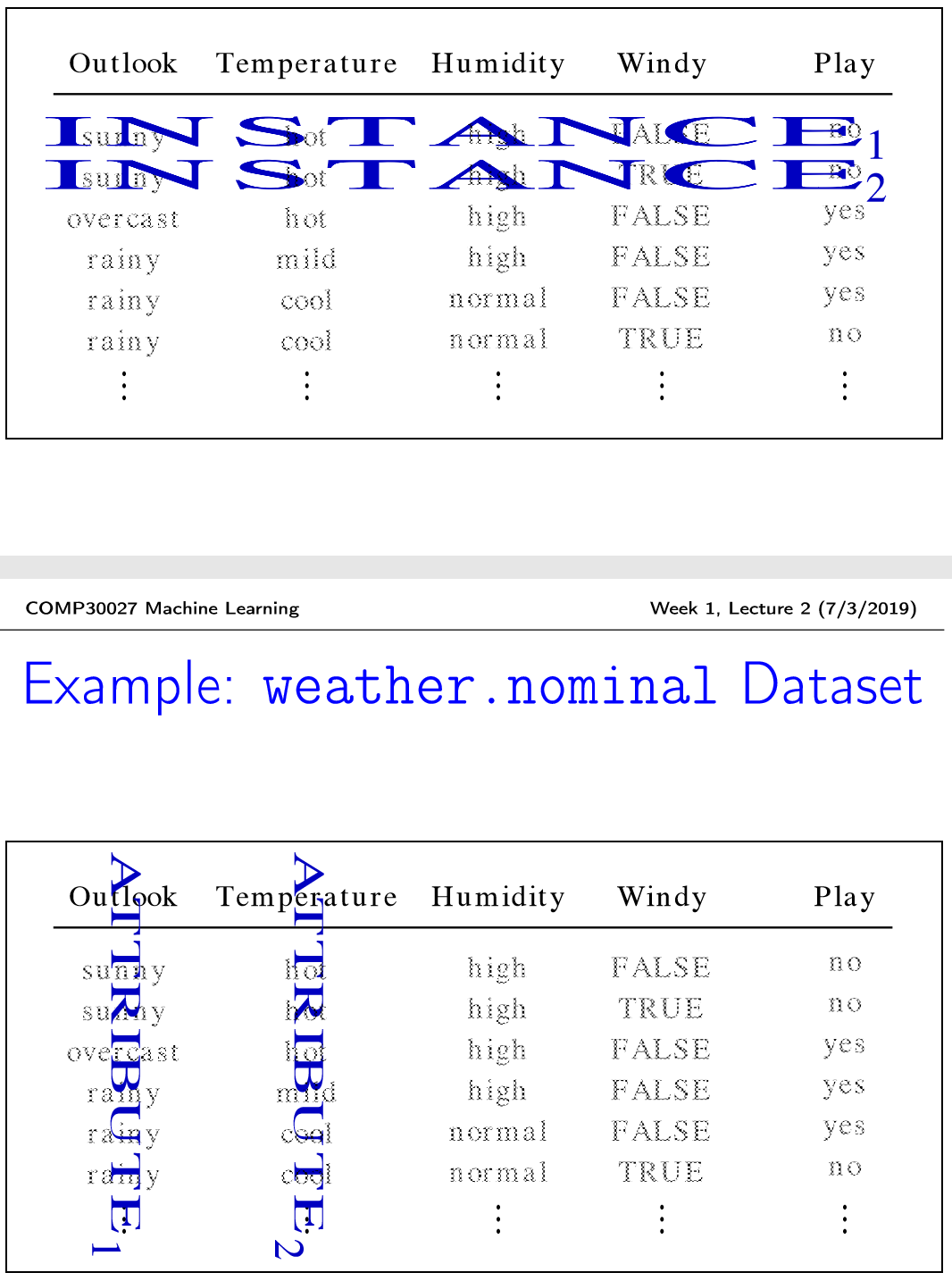
**Basics/Terminology:**

The input to a machine learning system consists of:

• Instances: the individual, independent examples of a concept also known as exemplars

• Attributes: measuring aspects of an instance also known as features

• Concepts: things that we aim to learn generally in the form of labels or classes



**What’s a Concept?**

• Styles of “concepts” that we aim to learn:

• Classiﬁcation learning: predicting a discrete class

• Clustering: grouping similar instances into clusters

• Regression: predicting a numeric quantity

• Association learning: detecting associations between attribute values

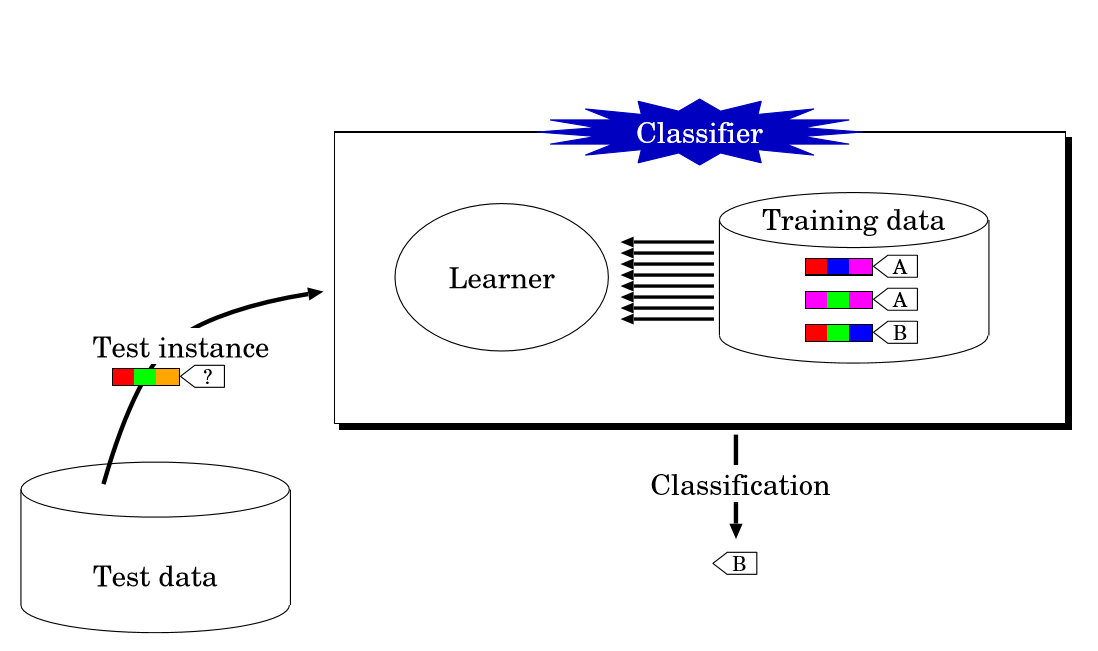
**Classiﬁcation Learning**

• Scheme is provided with actual outcome or class

• The learning algorithm is provided with a set of classiﬁed training data

• Measure success on “held-out” data for which class labels are known (test data)

• Classiﬁcation learning is supervised



**Clustering**

• Finding groups of items that are similar

• Clustering is unsupervised — the learner operates without a set of labelled training data

• The class of an example is not known ... or at least, not given to the classiﬁer

• Success often measured subjectively; evaluation is problematic

**A Word on Supervision**

• Supervised methods have prior knowledge of a closed set of classes and set out to discover and categorize new instances according to those classes

• Unsupervised methods:

• dynamically discover the “classes” (implicitly derived from grouping of instances) in the process of categorising the instances [STRONG] ...

OR • categorise instances as certain labels without the aid of pre-classiﬁed data [WEAK]

**Regression**

• Classiﬁcation learning, but class is continuous **(numeric prediction)**

• Learning is supervised

• Why is this distinct from Classiﬁcation?

• In Classiﬁcation, we can exhaustively enumerate all possible labels for a given instance; a correct prediction entails mapping an instance to the label which is truly correct

• In Regression, inﬁnitely many labels are possible, we cannot conceivably enumerate them; a “correct” prediction is when the numeric value is acceptably close to the true value

**Association Learning**

• Detect “useful” patterns, associations, correlations, or causal structures among sets of items or objects in dataset

• “Good” pattern: combination of attribute values where the presence of one (or more) value(s) suggests that one (or more) other value(s) will also be attested for numerous instances in the dataset

• Any kind of structure is considered interesting, and no a priori sense of what we hope to predict; unsupervised; evaluation is problematic • Potentially many, many association rules

Example:

1. humidity=normal windy=FALSE ==> play=yes

2. temperature=cool ==> humidity=normal

**Instance Topology**

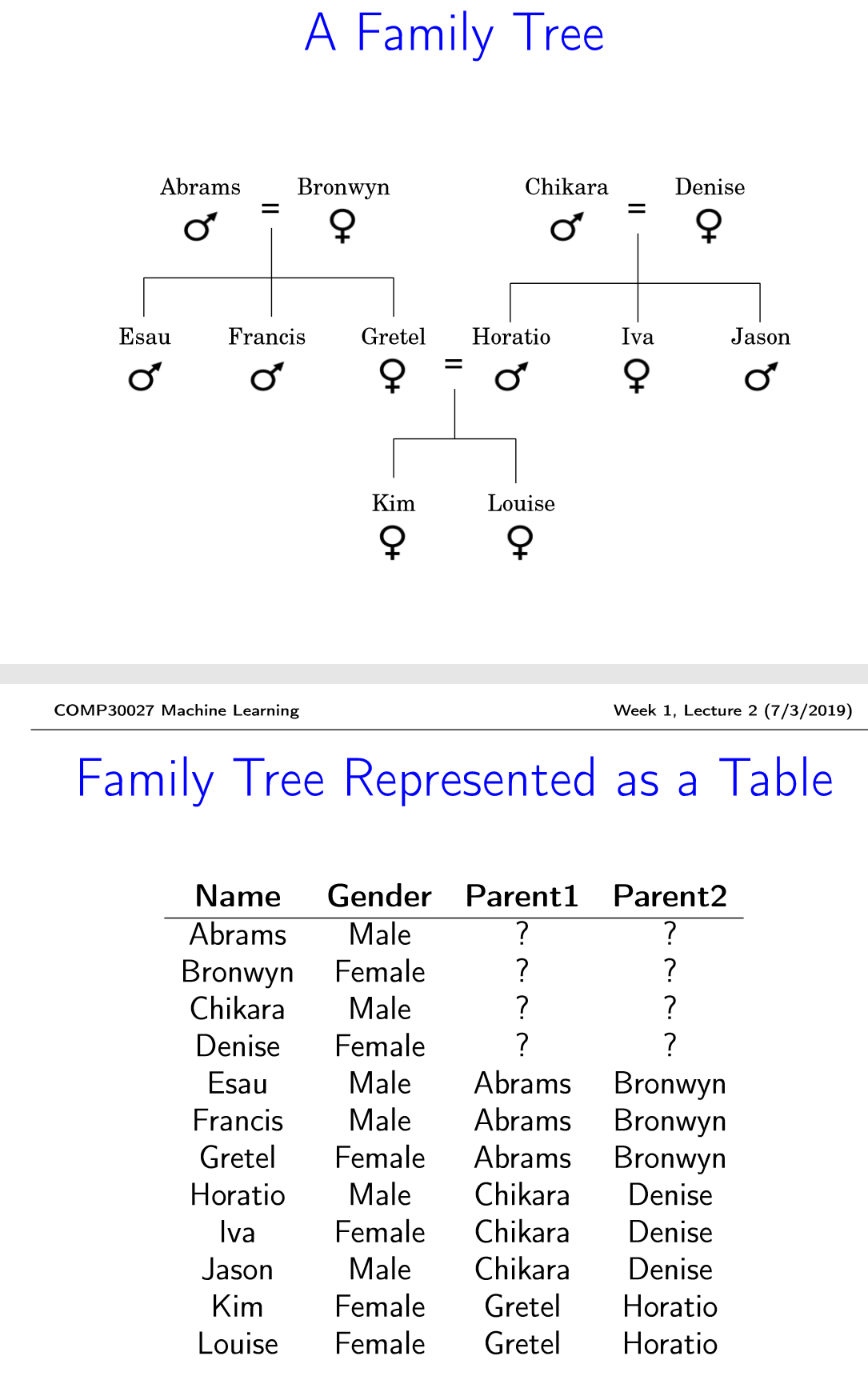
• Instances characterized as “feature vectors”, deﬁned by a predetermined set of attributes

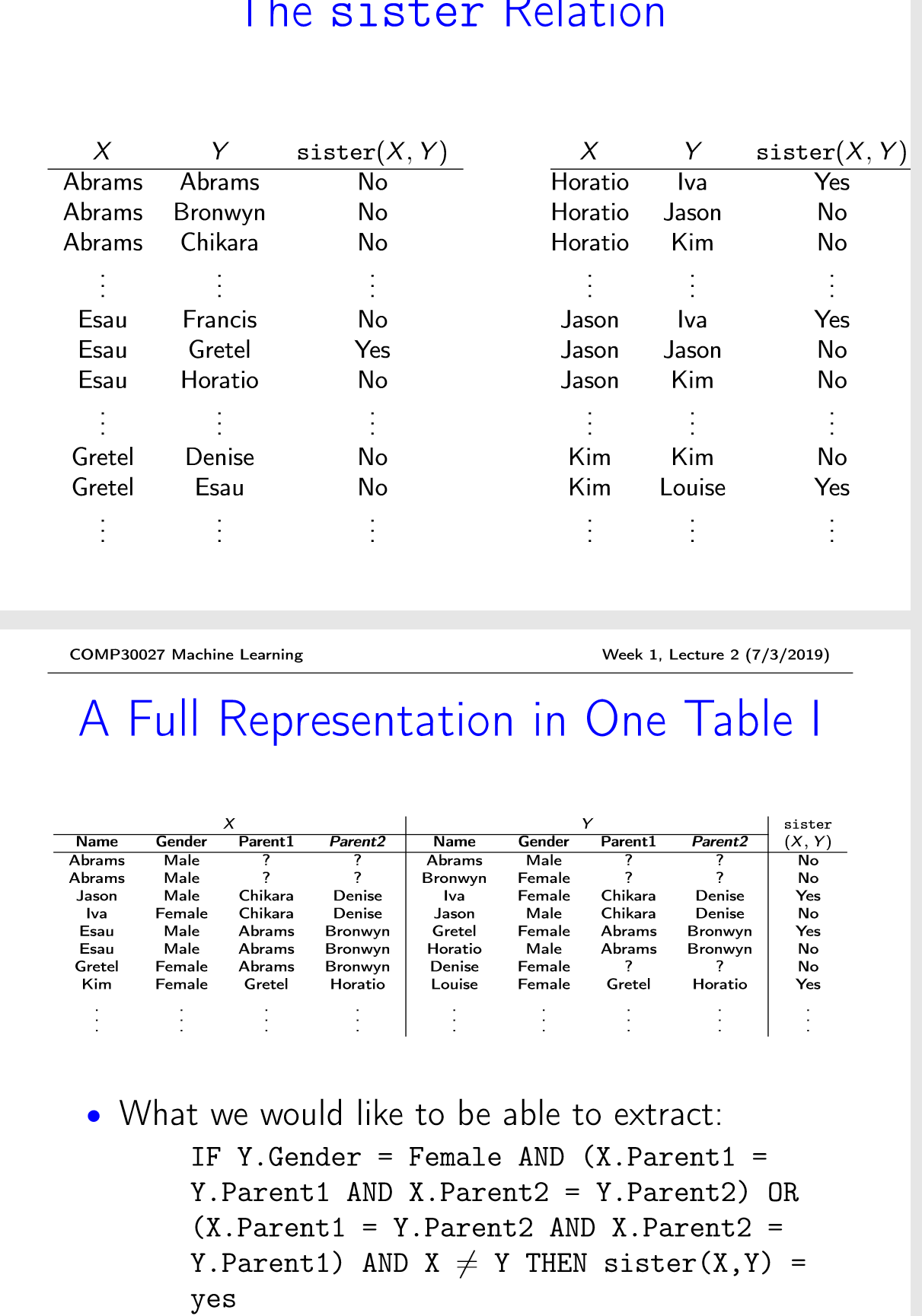
• Input to learning scheme: set of instances/dataset

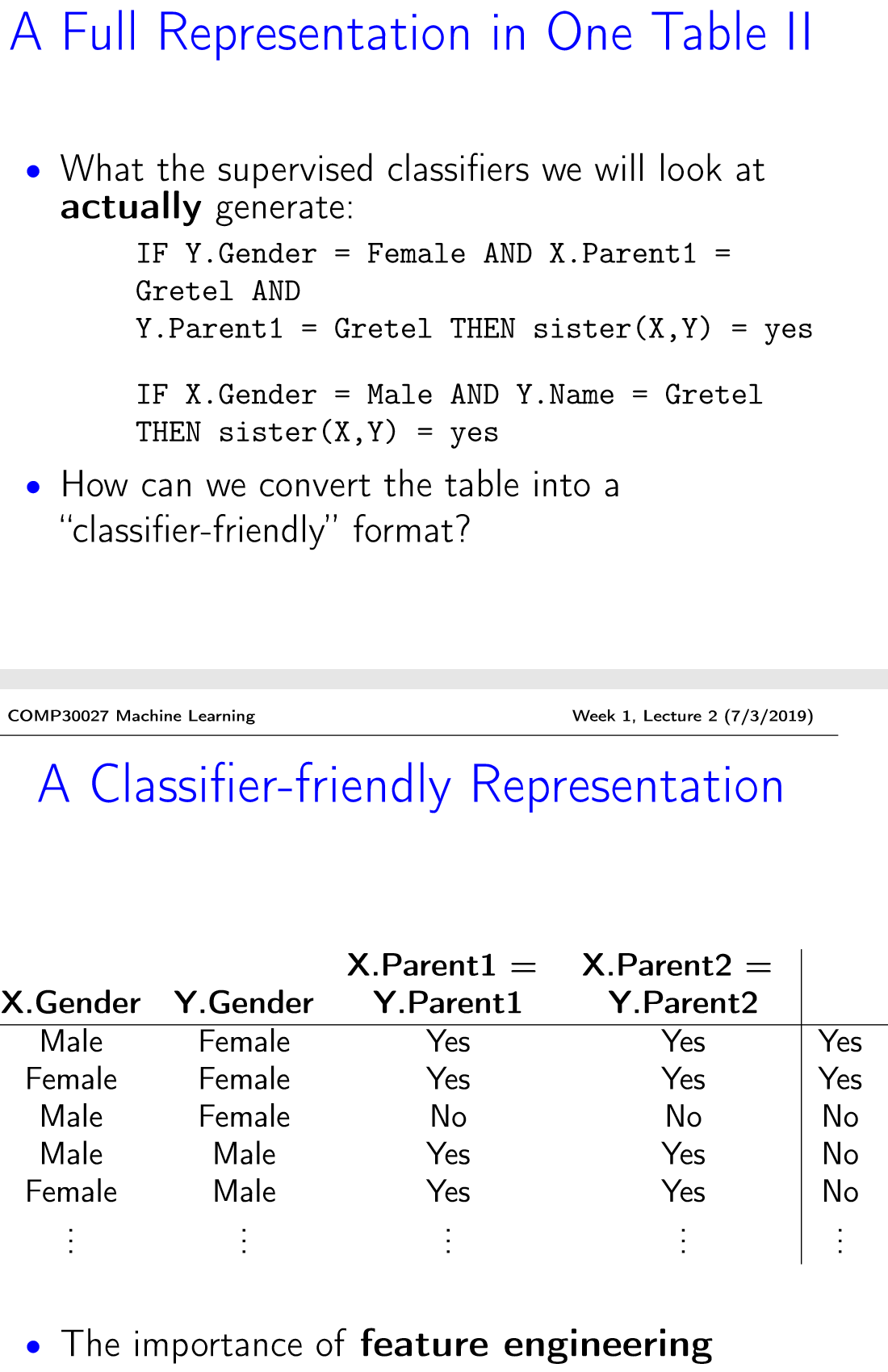
• Flat ﬁle representation

• No relationships between objects

• No explicit relationship between attributes







**What’s in an Attribute?**

• Each instance is described by a ﬁxed feature vector

• Possible attribute types (levels of measurement):

nominal

ordinal

continuous

**Nominal Quantities**

• Values are distinct symbols (e.g. {sunny,overcast,rainy})

• values themselves serve only as labels or names

• Also called categorical, or discrete (NB. “discrete” implies an order which tends not to exist) • Special case: dichotomy (“Boolean” attribute)

• No relation is implied among nominal values (no ordering or distance measure), and only equality tests can be performed

**Ordinal Quantities**

• An explicit order is imposed on the values (e.g. {hot,mild,cool} where hot > mild > cool)

• No distance between values deﬁned and addition and subtraction don’t make sense

• Example rule: temperature < hot →play = yes • Distinction between nominal and ordinal not always clear (e.g. outlook)

**Continuous Quantities**

• Continuous quantities are real-valued attributes with a well-deﬁned zero point and no explicit upper bound

• Example: attribute distance • Distance between an object and itself is zero

• All mathematical operations are allowed (of which addition, subtraction, scalar multiplication are most salient, but other operations are relevant in some contexts)

**Transforming attributes to Boolean is one commonly-used work-around**

**Day 3**

**ML in Reality:**

**Preparing the Input**

• Problem: diﬀerent data sources (e.g. sales department, customer billing department, ...)

• Diﬀerences: styles of record keeping, conventions, time periods, data aggregation, primary keys, errors

• Data must be assembled, integrated, cleaned up

• Data warehouse: consistent point of access • External data/storage may be required • Critical: type and level of data aggregation

**Sample Representation: ARFF**

@relation weather

@attribute outlook {sunny, overcast, rainy}

@attribute temperature real

@attribute humidity real

@attribute windy {TRUE, FALSE}

@attribute play {yes, no}

@data

sunny,85,85,FALSE,no

sunny,80,90,TRUE,no

overcast,83,86,FALSE,yes

rainy,70,96,FALSE,yes

**Missing Values**

• The number of attributes may vary in practice

• missing values

• inter-dependent attributes

• Frequently indicated by out-of-range entries

• Types: unknown, unrecorded, irrelevant

• Reasons: • malfunctioning equipment • changes in experimental design • collation of diﬀerent datasets • measurement not possible

• Missing value may have signiﬁcance in itself (e.g. missing test in a medical examination) • Most schemes assume that is not the case →missing may need to be coded discretely

**Inaccurate Values**

• Cause: a given data mining application is often not known at the time logging is set up

• Result: errors and omissions that don’t aﬀect original purpose of data (e.g. age of customer)

• Typographical errors in nominal attributes values need to be checked for consistency

• Typographical and **measurement errors** in numeric attributes →outliers need to be identiﬁed

• Errors may be deliberate (e.g. wrong post codes)

**Getting to Know the Data**

• Simple visualization tools are very useful

• Nominal attributes: **histograms** (distribution consistent with background knowledge?)

• Numeric attributes: **scatter plots** (any obvious outliers?)

• 2-D and 3-D plots show dependencies

• Need to consult domain experts • Too much data to inspect? Take a sample! • You can never know your data too well

**Day 4**