Data Mining for Data Modelers

Asoka Diggs

Data Scientist, Intel IT

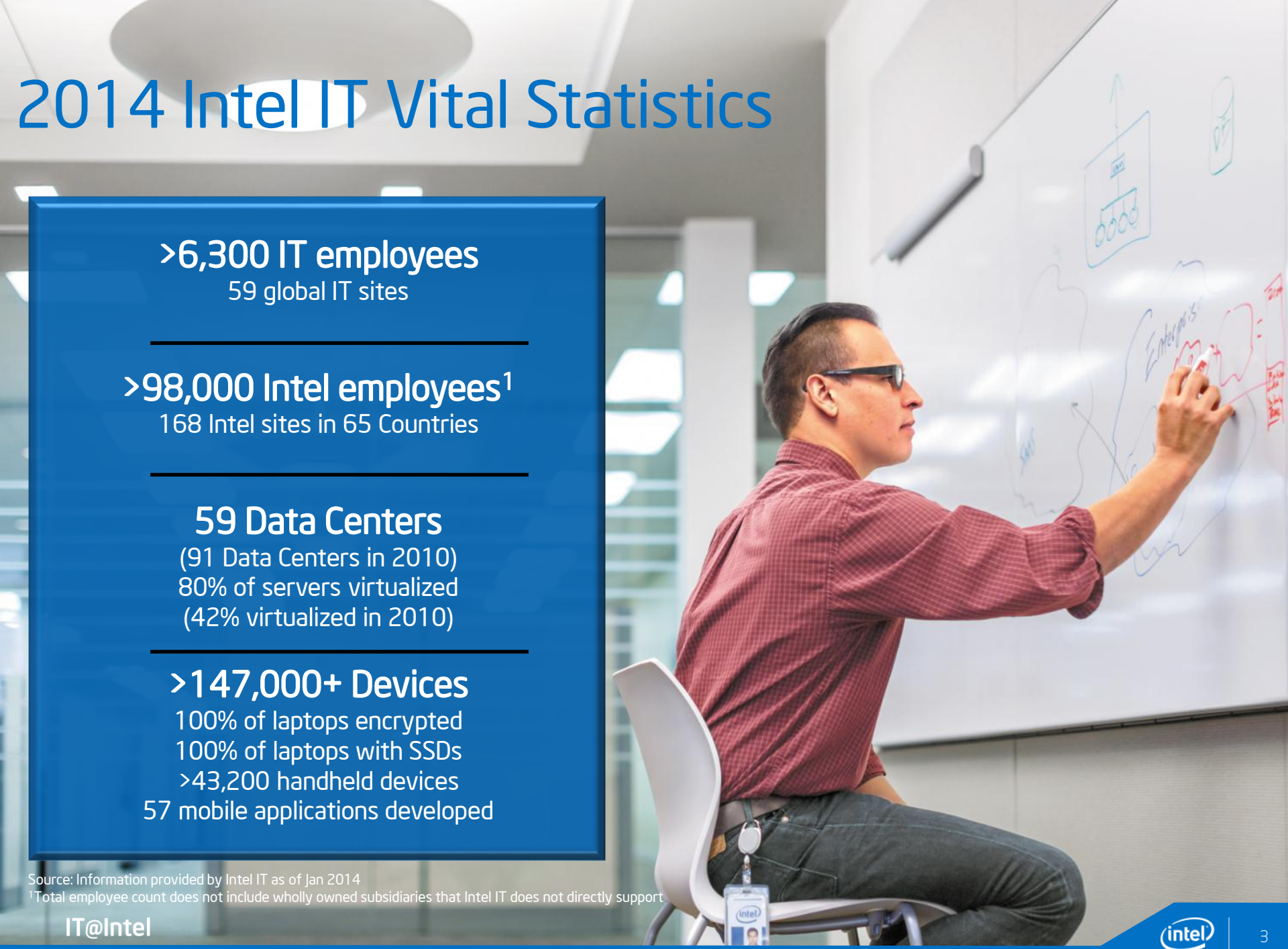
March 5, 2015

IT@Intel

Legal Notices

This presentation is for informational purposes only. INTEL MAKES NO WARRANTIES, EXPRESS OR IMPLIED, IN THIS SUMMARY.

2014 Intel IT Vital Statistics



>6,300 IT employees
59 global IT sites

>98,000 Intel employees[1]
168 Intel sites in 65 Countries

59 Data Centers
(91 Data Centers in 2010)
80% of servers virtualized
(42% virtualized in 2010)

>147,000+ Devices
100% of laptops encrypted
100% of laptops with SSDs
>43,200 handheld devices
57 mobile applications developed

Source: Information provided by Intel IT as of Jan 2014
[1]Total employee count does not include wholly owned subsidiaries that Intel IT does not directly support

Outline

Introduction and assumptions

Data Mining demo

Break

Overview of Data Mining concepts

Questions throughout

Resources and links for further study

My Background

DBA starting in late 90's

Various data management roles since then

- Entity-Relationship modeling
- ETL development
- Data integration
- Data architecture

MS Predictive Analytics started Fall 2011

- 9/11 classes completed

Designed and teach an internal Intro to Data Mining class

Assumptions

Audience is composed of experienced Data Management Professionals

You're interested in Predictive Analytics / Data Mining but haven't studied it

You're already convinced of the value of Predictive Analytics / Data Mining

- We won't particularly be covering the value here

What you will learn

Class is designed to teach a functional mental model of data mining

Definition: Mental Model

- A **mental model** is an explanation of someone's thought process about how something works in the real world
- acts as information filter – causes selective perception or perception of only selected parts of information

With the mental model, you'll be making your own connections between data mining and skills / experience you already have

Goal: Begin learning how to frame business problems as data mining problems

- Real goal: realize that you can learn to frame business problems as data mining problems with minimal effort
  - But more effort than a 3 hour demo version

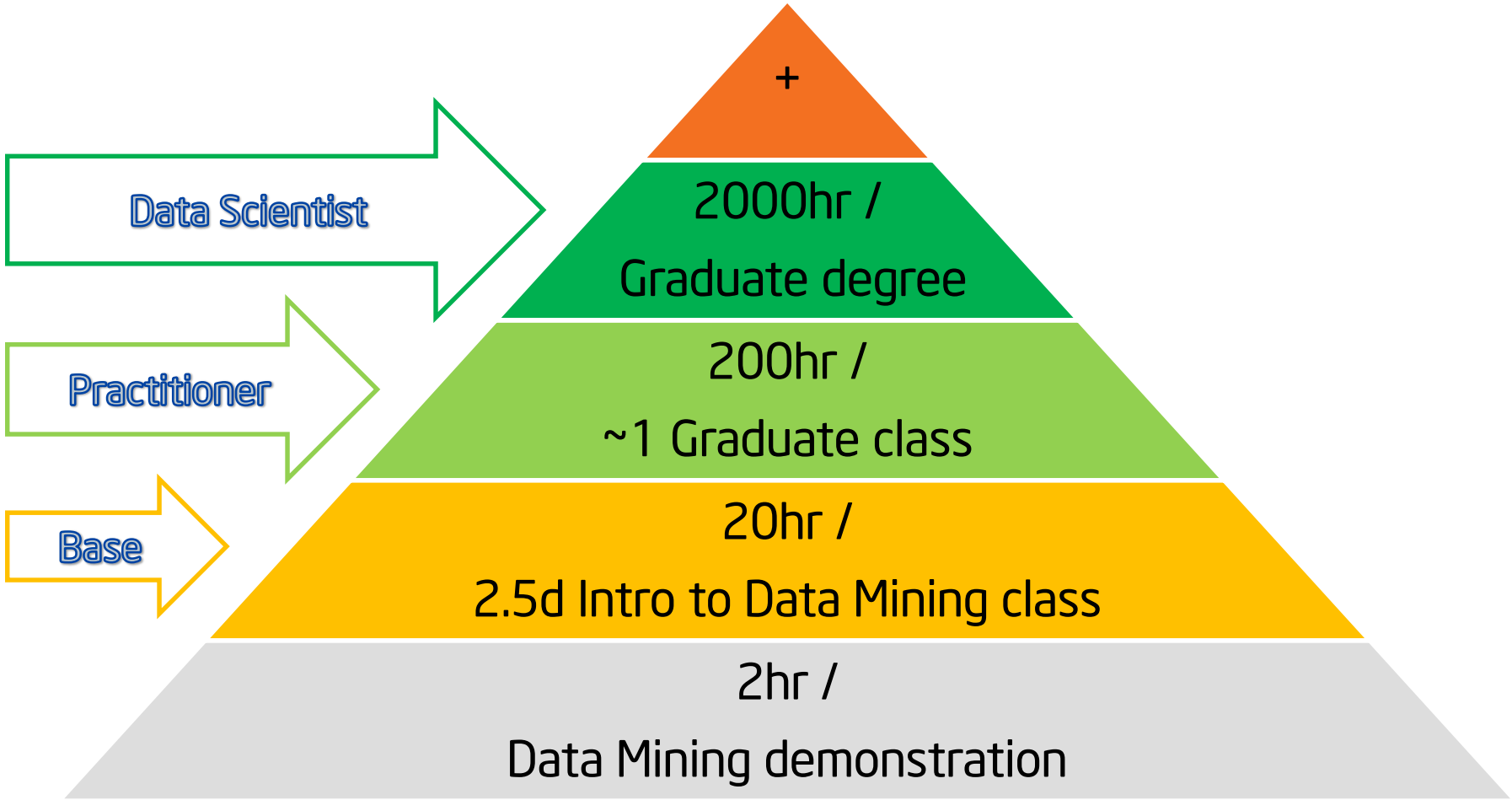
http://en.wikipedia.org/wiki/Mental_model

Demo

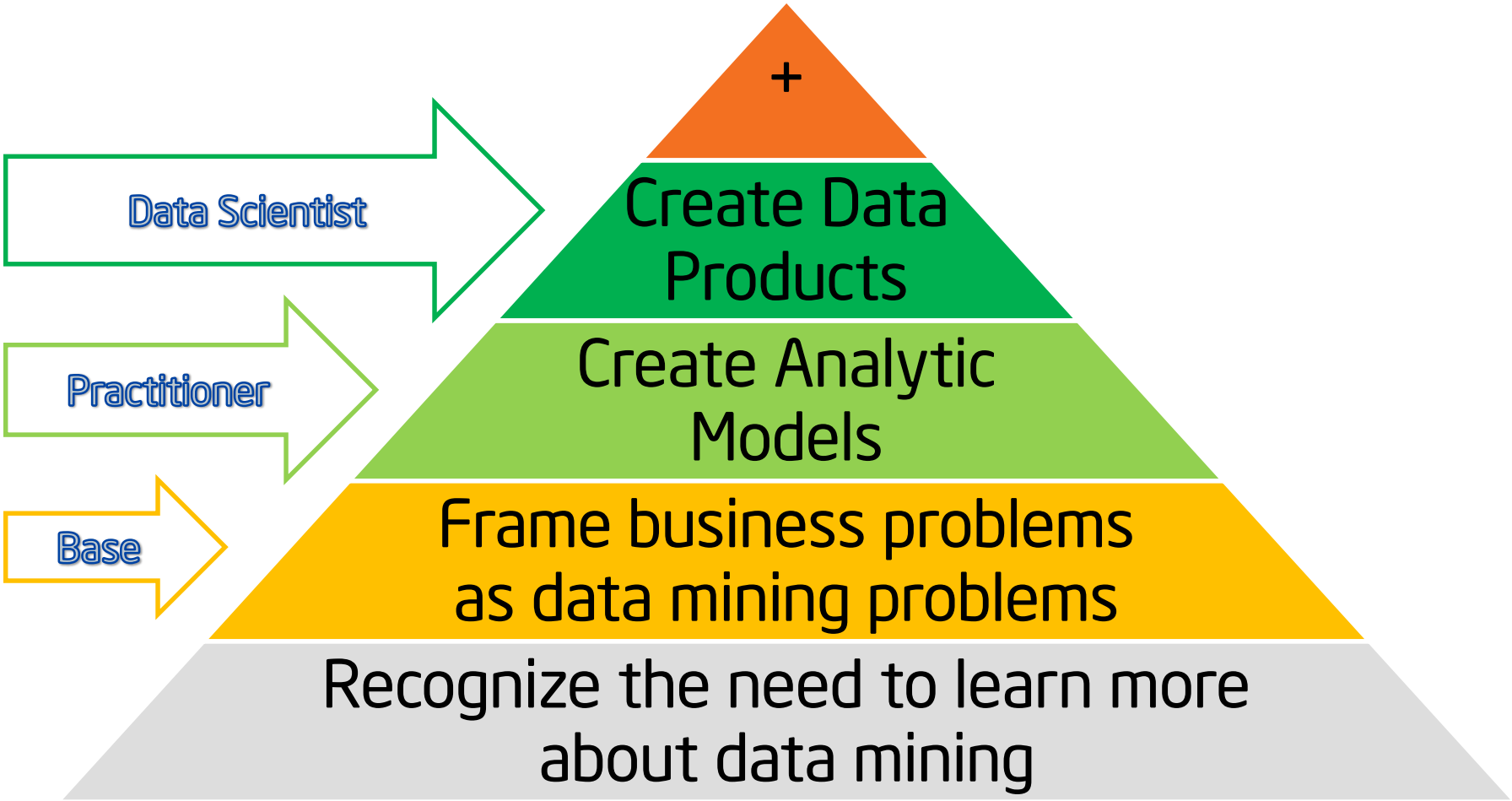
Decision Tree and

- Iris data set
- eReader Adoption data set

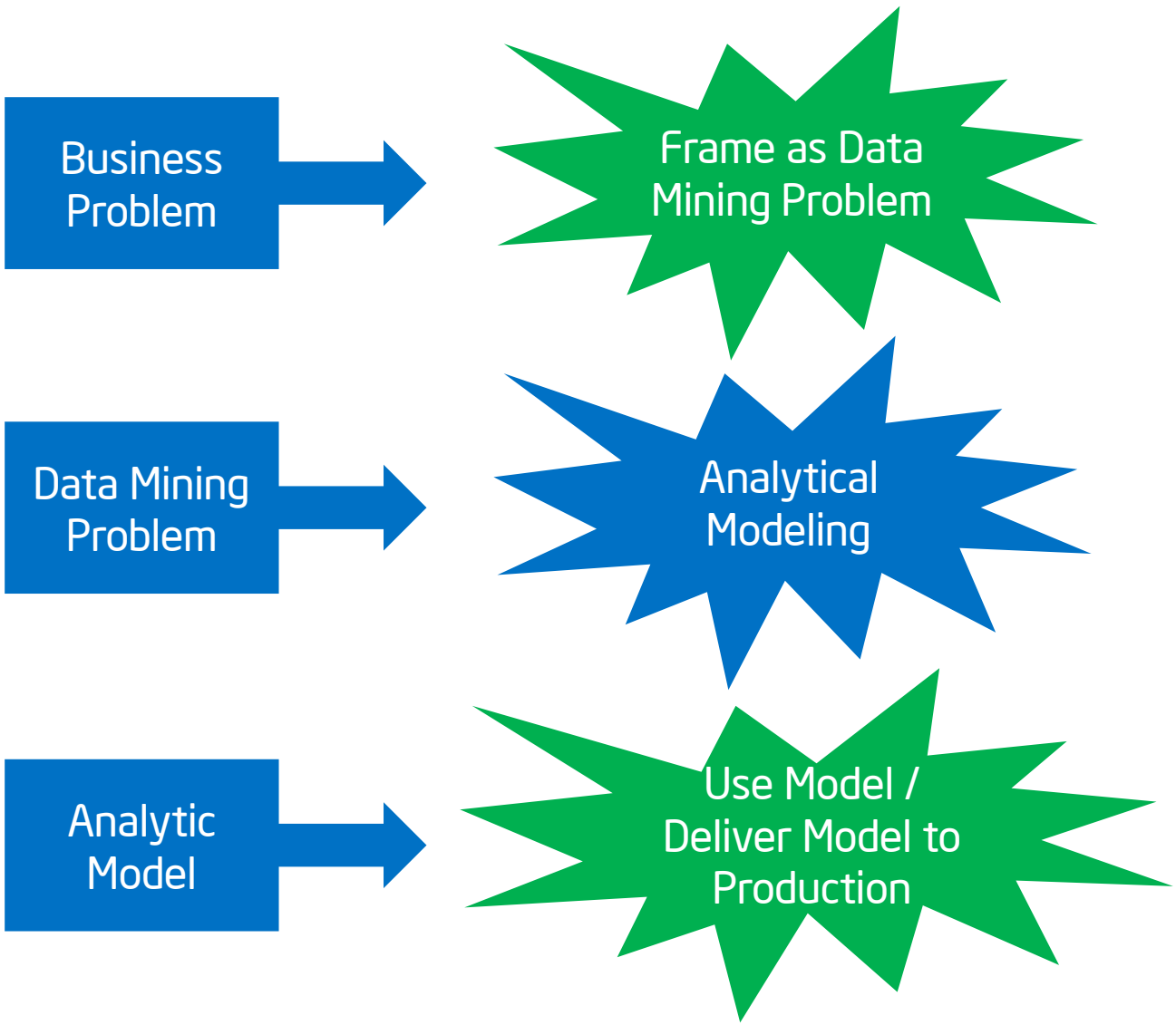
Effort to achieve skill level



Deliverable Expectations



A variation on CRiSP-DM



Formulate the data mining problem

Descriptive Analytics	Diagnostic Analytics	Predictive Analytics	Prescriptive Analytics	Pre-emptive Analytics
What happened?	Why did it happen?	What will happen?	What should I do?	What more can I do?
What is the attrition in the last 6 months? Which customers have I lost?	Why has the attrition rate increased in the last 6 months?	Which customers are most likely to leave? Which customers are most likely to leave if I don't contact them?	Which customers should I target for retention? What if ...?	What can I offer before the customer even realizes they have the need? Value-add?
Information	Insights		Decision	Action

Machine Learning (ML) input and output

Any ML algo will have certain kinds of data it can take as input

And will produce certain kinds of data as output

- Example: Logistic Regression requires numeric input (including binary variables), and generates a binary (1/0) output
- Data prep activity – converting data from an unusable form into a usable form

The nature of the data mining problem will shape the kind of answer needed (binary, category, number, ..)

- Which will start filtering the available techniques

Further filter the techniques based on the form of the data that is available

- And we will talk more about this

Supervised and Unsupervised learning

Machine Learning and predictive modeling falls into two buckets…

- Supervised or Directed (Classification)
  - **Presence of a response or outcome variable**
  - Finds relationships in the data
  - ~75% of the techniques
  - Generally more interesting and useful biz q's can be addressed
- Unsupervised or Undirected (Clustering)
  - **Absence of a response or outcome variable**
  - Finds patterns in the data
  - Not as many techniques
  - More easy to find patterns that aren't useful

Think of response variable as "human input" into the model training and validation process

The 1 Slide Intro to Data Mining class

Training Data

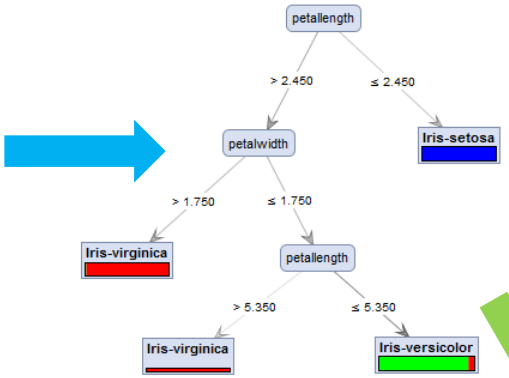
Build Analytic Model

Model Validation

Scoring Data

Scored Data

class	sepallength	sepalwidth	petallength	petalwidth
Iris-versicolor	5	2	3.500	1
Iris-versicolor	6	2.200	4	1
Iris-versicolor	6.200	2.200	4.500	1.500
Iris-virginica	6	2.200	5	1.500
Iris-setosa	4.500	2.300	1.300	0.300
Iris-versicolor	5.500	2.300	4	1.300
Iris-versicolor	6.300	2.300	4.400	1.300
Iris-versicolor	5	2.300	3.300	1



		TRUE			
		Iris-setosa	Iris-versicolor	Iris-virginica	class precision
PREDICTED	Iris-setosa	50	0	0	100.00%
	Iris-versicolor	0	47	6	88.68%
	Iris-virginica	0	3	44	93.62%
	class recall	100.00%	94.00%	88.00%	

sepallength	sepalwidth	petallength	petalwidth
5.700	4.400	1.500	0.400
5.500	4.200	1.400	0.200
5.200	4.100	1.500	0.100
5.800	4	1.200	0.200
5.400	3.900	1.700	0.400
5.400	3.900	1.300	0.400
5.700	3.800	1.700	0.300
5.100	3.800	1.500	0.300

class	sepallength	sepalwi...	petallength	petalwidth
Iris-setosa	5.700	4.400	1.500	0.400
Iris-setosa	5.500	4.200	1.400	0.200
Iris-setosa	5.200	4.100	1.500	0.100
Iris-setosa	5.800	4	1.200	0.200
Iris-setosa	5.400	3.900	1.700	0.400
Iris-setosa	5.400	3.900	1.300	0.400
Iris-setosa	5.700	3.800	1.700	0.300
Iris-setosa	5.100	3.800	1.500	0.300

Variable modeling types (aka data types):

- Discrete / Categorical / Qualitative (alphanumeric/text)
  - Nominal – no meaning in order
    - **Distinctness (= !=) ; Example: Name, favorite color**
    - Binary/Binomial variables - Represents presence or absence of something
      - » Not necessarily Black/White; Example: gender
  - Ordinal – order makes sense
    - **Order (< >) ; Example: Order ID, Pay Grade, low/medium/high**
- Continuous / Quantitative (numbers only)
  - Interval – sum/difference meaningful, but product/ratio not
    - **Addition (+ -) ; Example: temperature in Fahrenheit, time of day AM/PM**
  - Ratio – ratios make sense in addition to sum/difference
    - **Multiplication (* /) ; Example: absolute time, absolute temp, height, weight, wavelength**

Grain: Another way to think about data

Level of detail - What is an observation (row), an observation of?

- Think of the grain in a piece of wood
  - Example: Cumulative GPA vs. per course grade-point (transcript)

Has proven useful in reporting, data structure, formulating ML problems, etc..

- Can you identify the grain of an answer to the business question?
- What is the grain of the source data you are analyzing?
- Is the grain of the (data mining) analysis of the source data compatible with the grain of the business question?
  - Example: GPA is not a good grain to analyze "fairness" across faculty

Don't mix grain in a data set

- Or at least only mix grain consciously and intentionally

If you ever have an itchy feeling that something about the data is wrong, start with the grain

- Can you even identify what the grain is?

Data Terminology

Data Model

- Usually NOT an ERD (rather, a data mining model)
- I am trying to use ER Model and Analytic Model

Observation

- **Row**, example, case, tuple, sample, member

Variable

- **Column**, attribute, field, feature, parameter, dimension, metric, measure

Roles played by variables

- Label (dependent), attribute (independent), id

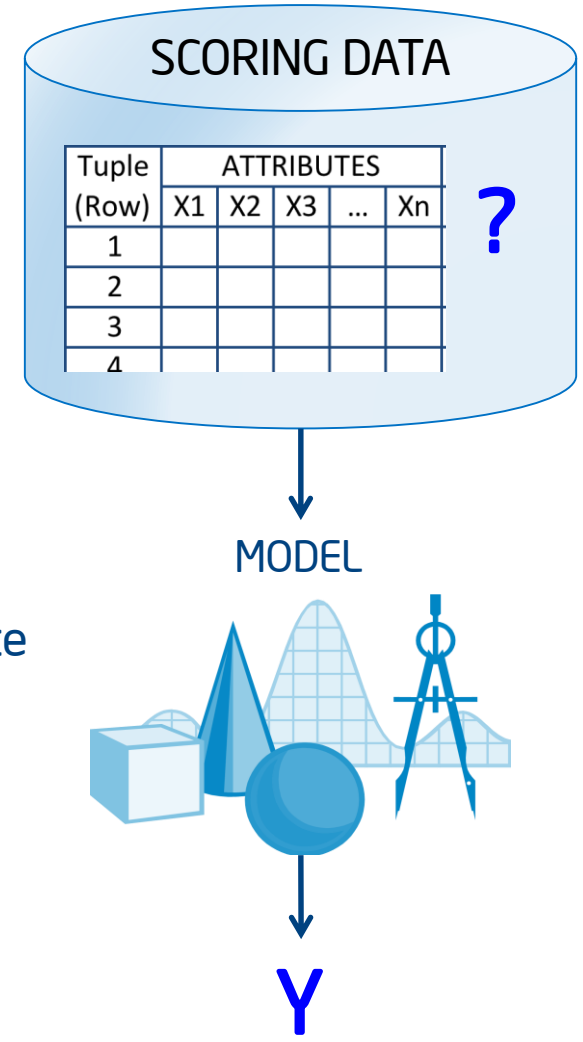
Emphasis in data mining is on de-normalized data

Tuple	ATTRIBUTES					LABEL
(Row)	X1	X2	X3	...	Xn	Y
1						
2						
3						
4						
5						
6						
.						
.						
.						
M						

Modeling Terminology

Data Set

- Training. Used to train the data (mining) model
  - The response is a known value
- Test. An out of sample (observations not part of the training set) used to validate the results of the model
  - The response is a known value
- Scoring. A set of observations where the response or outcome is unknown. We use the model created by the data mining method and the training set to Score or create values for the response variable
  - After scoring, the response is predicted
    - **NOT determined**



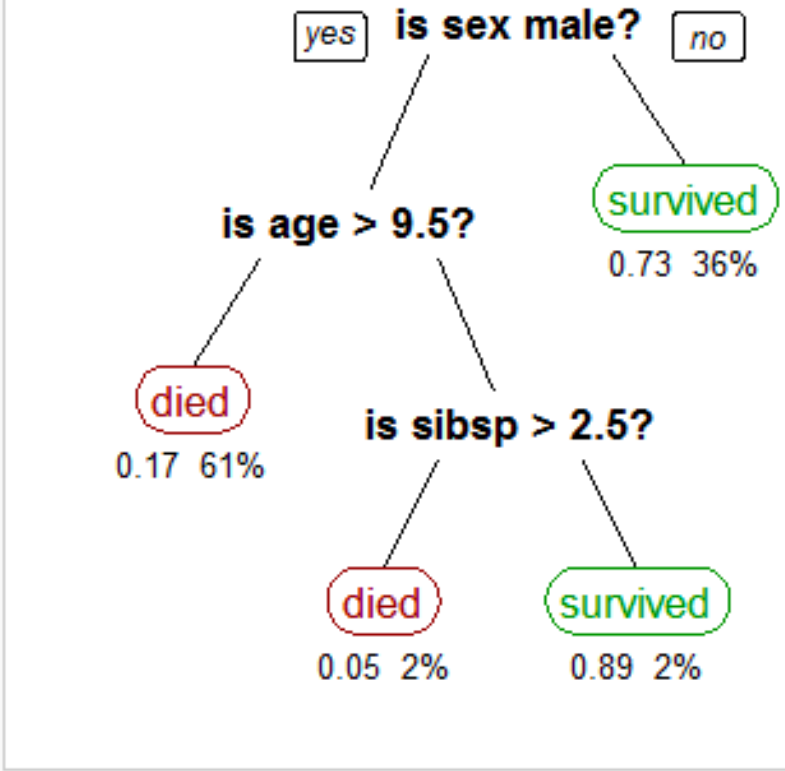
Decision Tree

Output / response / label: categorical (or continuous)

Input / observation / attribute: categorical or continuous

Model: (example below)

- **Very friendly for human interpretation**
- http://en.wikipedia.org/wiki/Decision_tree_learning
- http://en.wikipedia.org/wiki/C4.5_algorithm



A tree showing survival of passengers on the Titanic ("sibsp" is the number of spouses or siblings aboard). The figures under the leaves show the probability of survival and the percentage of observations in the leaf.

Model Validation and Testing

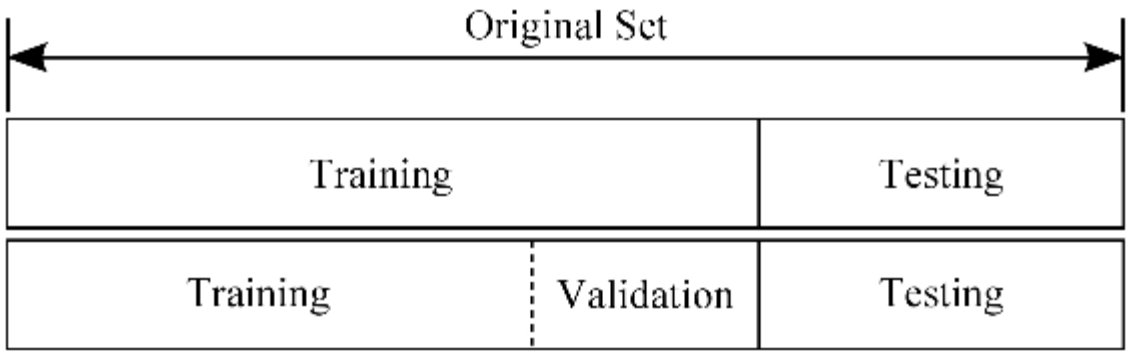


FIGURE 3.

Two and three way splitting

Training set – a set used for learning and estimating parameters of the model.

Validation set – a set used to evaluate the model, usually for model selection.

Testing set – a set of examples used to assess the predictive performance of the model.

Citation

Validation approaches

Most important part of model building

- How do you know the model you've built is reasonable?
  - This answers the technical question
  - Still need to answer the business question

70/30

- 70% of training data as an in-sample model building / training data set
- 30% of training data as an out-of-sample testing or validation set
  - Both training and testing data have a known value for the response variable
  - (Y has an actual value for all observations)
- Scoring data – the Y value is unknown (but we'd like to have an idea of what it is likely to be)

Confusion Matrix

We'll look at the model validation results together

- More reading on Precision and Recall
- http://en.wikipedia.org/wiki/Precision_and_recall

"In simple terms, high **precision** means that an algorithm returned substantially more relevant results than irrelevant, while high **recall** means that an algorithm returned most of the relevant results."

- Quality
- Quantity

Logistic Regression

Output / response: binary (1/0)

- Also polynomial; not covered in this book

Input / observation: numeric

- Can use categorical data after converting to numeric
  - Discretization
  - Binarization or Dummy variables

Model: logistic equation

- http://en.wikipedia.org/wiki/Logistic_regression

Conceptually:

- Event occurred / didn't occur
- We care / don't care
- True / false
- Black / not-Black

Absence of proof, is not proof of absence

Resources

Self-Study Resources

Textbook: Data Mining for the Masses, by Matthew North

Method:

- Download the textbook
- Follow the step-by-step instructions from the beginning to the end of the book
- By the time you finish the last chapter, you will have:
  - Analyzed 4 data sets using different supervised learning methods
  - Analyzed 2 more data sets using unsupervised learning methods
  - Used rudimentary text analytics on a data set
  - Incorporated model validation into at least 1 model

Our demo today has made use of 2 of these data sets and is drawn from this book

Additional Resources

Broadening Access to Advanced Analytics in the Enterprise, by Asoka Diggs & Christy Foulger

- And the subject of my Tuesday afternoon presentation "Democratization of Data Analytics"

Short video series on Text Analytics, by Neil McGuigan

Data Science Specialization through Coursera

- For those looking for an Intermediate level of knowledge of data science

The Signal and the Noise: Why So Many Predictions Fail – but Some Don't, by Nate Silver

Contact me at asoka.diggs@intel.com

IT@Intel

Sharing Intel IT Best Practices With the World

Learn more about Intel IT's Initiatives at www.intel.com/IT

Questions?

Thank You

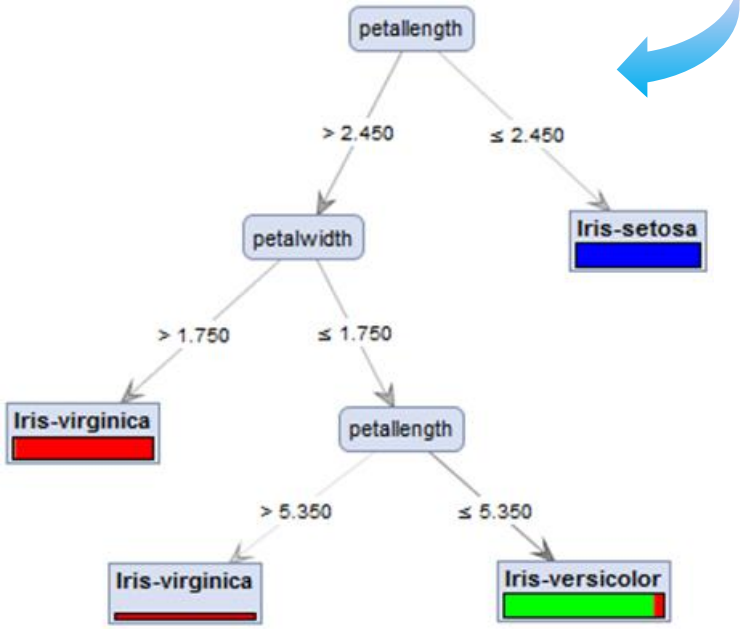
The 1 Slide Intro to Data Mining class

Training Data

class	sepallength	sepalwidth	petallength	petalwidth
Iris-versicolor	5	2	3.500	1
Iris-versicolor	6	2.200	4	1
Iris-versicolor	6.200	2.200	4.500	1.500
Iris-virginica	6	2.200	5	1.500
Iris-setosa	4.500	2.300	1.300	0.300
Iris-versicolor	5.500	2.300	4	1.300
Iris-versicolor	6.300	2.300	4.400	1.300
Iris-versicolor	5	2.300	3.300	1

The 1 Slide Intro to Data Mining class

class	sepallength	sepalwidth	petallength	petalwidth
Iris-versicolor	5	2	3.500	1
Iris-versicolor	6	2.200	4	1
Iris-versicolor	6.200	2.200	4.500	1.500
Iris-virginica	6	2.200	5	1.500
Iris-setosa	4.500	2.300	1.300	0.300
Iris-versicolor	5.500	2.300	4	1.300
Iris-versicolor	6.300	2.300	4.400	1.300
Iris-versicolor	5	2.300	3.300	1

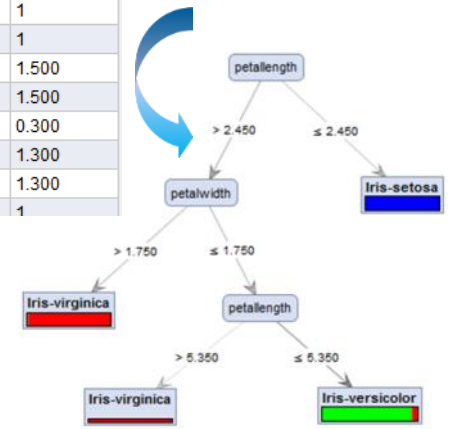


Training Data

Build Analytic Model

The 1 Slide Intro to Data Mining class

class	sepallength	sepalwidth	petallength	petalwidth
Iris-versicolor	5	2	3.500	1
Iris-versicolor	6	2.200	4	1
Iris-versicolor	6.200	2.200	4.500	1.500
Iris-virginica	6	2.200	5	1.500
Iris-setosa	4.500	2.300	1.300	0.300
Iris-versicolor	5.500	2.300	4	1.300
Iris-versicolor	6.300	2.300	4.400	1.300
Iris-versicolor	5	2.300	3.300	1



Training Data

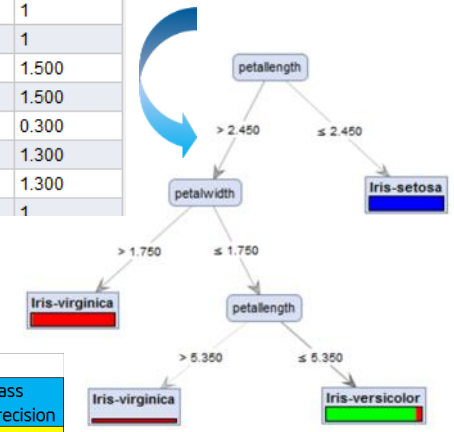
Build Analytic Model

Model Validation

		TRUE			
		Iris-setosa	Iris-versicolor	Iris-virginica	class precision
	Iris-setosa	50	0	0	100.00%
PREDICTED	Iris-versicolor	0	47	6	88.68%
	Iris-virginica	0	3	44	93.62%
	class recall	100.00%	94.00%	88.00%	

The 1 Slide Intro to Data Mining class

class	sepallength	sepalwidth	petallength	petalwidth
Iris-versicolor	5	2	3.500	1
Iris-versicolor	6	2.200	4	1
Iris-versicolor	6.200	2.200	4.500	1.500
Iris-virginica	6	2.200	5	1.500
Iris-setosa	4.500	2.300	1.300	0.300
Iris-versicolor	5.500	2.300	4	1.300
Iris-versicolor	6.300	2.300	4.400	1.300
Iris-versicolor	5	2.300	3.300	1



		TRUE			
		Iris-setosa	Iris-versicolor	Iris-virginica	class precision
	Iris-setosa	50	0	0	100.00%
PREDICTED	Iris-versicolor	0	47	6	88.68%
	Iris-virginica	0	3	44	93.62%
	class recall	100.00%	94.00%	88.00%	

sepallength	sepalwidth	petallength	petalwidth
5.700	4.400	1.500	0.400
5.500	4.200	1.400	0.200
5.200	4.100	1.500	0.100
5.800	4	1.200	0.200
5.400	3.900	1.700	0.400
5.400	3.900	1.300	0.400
5.700	3.800	1.700	0.300

Training Data

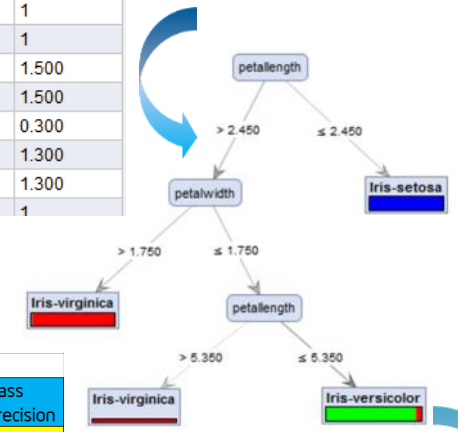
Build Analytic Model

Model Validation

Scoring Data

The 1 Slide Intro to Data Mining class

class	sepallength	sepalwidth	petallength	petalwidth
Iris-versicolor	5	2	3.500	1
Iris-versicolor	6	2.200	4	1
Iris-versicolor	6.200	2.200	4.500	1.500
Iris-virginica	6	2.200	5	1.500
Iris-setosa	4.500	2.300	1.300	0.300
Iris-versicolor	5.500	2.300	4	1.300
Iris-versicolor	6.300	2.300	4.400	1.300
Iris-versicolor	5	2.300	3.300	1



		TRUE			
		Iris-setosa	Iris-versicolor	Iris-virginica	class precision
	Iris-setosa	50	0	0	100.00%
PREDICTED	Iris-versicolor	0	47	6	88.68%
	Iris-virginica	0	3	44	93.62%
	class recall	100.00%	94.00%	88.00%	

sepallength	sepalwidth	petallength	petalwidth
5.700	4.400	1.500	0.400
5.500	4.200	1.400	0.200
5.200	4.100	1.500	0.100
5.800	4	1.200	0.200
5.400	3.900	1.700	0.400
5.400	3.900	1.300	0.400
5.700	3.800	1.700	0.300

class	sepallength	sepalwi...	petallength	petalwidth
Iris-setosa	5.700	4.400	1.500	0.400
Iris-setosa	5.500	4.200	1.400	0.200
Iris-setosa	5.200	4.100	1.500	0.100
Iris-setosa	5.800	4	1.200	0.200
Iris-setosa	5.400	3.900	1.700	0.400
Iris-setosa	5.400	3.900	1.300	0.400
Iris-setosa	5.700	3.800	1.700	0.300

Training Data

Build Analytic Model

Model Validation

Scoring Data

Scored Data