UCIrvine | The Paul Merage School of Business

2022.11.24

BANA 273 Session 1

Introduction

Data Mining and Business Intelligence

Prof. Vibhanshu (Vibs) Abhishek

The Paul Merage School of Business

University of California, Irvine

Agenda

- Introduction
  - Instructor and TA
- Course Logistics
- Data Mining Examples
- SQL

About the Instructor

- Undergraduate degree in Computer Sc & Engr
  - Indian Institute of Technology, Kanpur
- Masters in Statistics
  - University of Pennsylvania, The Wharton School
- Ph.D in Operations and Information Management
  - University of Pennsylvania, The Wharton School
- Assistant Professor at Carnegie Mellon University
- Just joined UCI ☺

Teaching

- 273 Business Intelligence for Analytical Decision Making – BANA/FEMBA
- 173 Business Intelligence for Analytical Decision Making – UG
- Digital Marketing Analytics
- Managing Disruptive Technologies

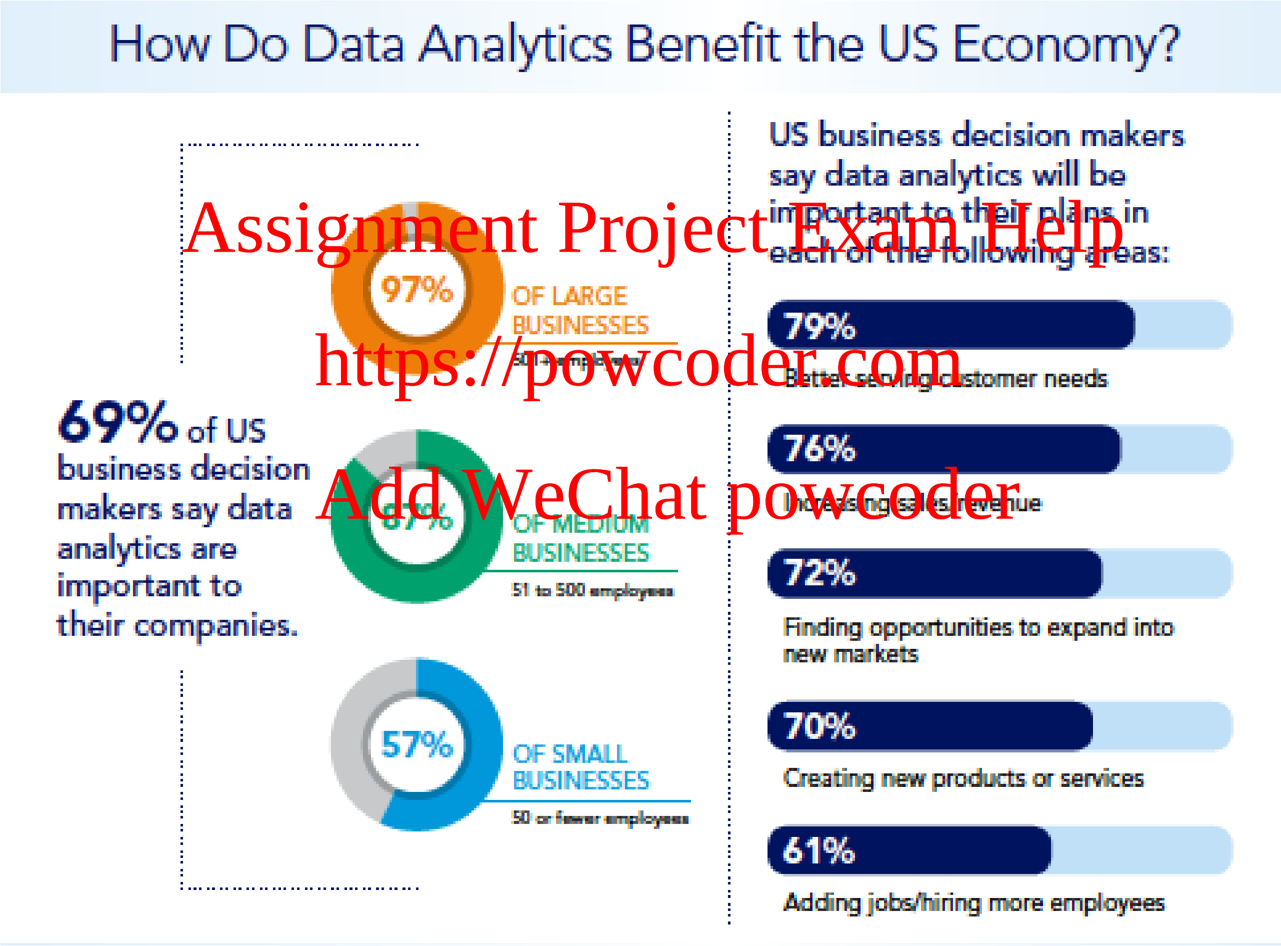
Research

- Effect of emerging technologies on consumers, firms and markets
- Digital Marketing, Retail, Mobility

Some Consulting Projects

- Sequoia Capital
- Adobe
- McKinsey
- LEGO
- Kaplan
- One Championship
- Rivigo

BSA Survey on Analytics



Other than work …

- Startups
- Photography
- Travel
- Anvi

Course Topics

- Getting data from Database: Structured Query Language
- RFM Analysis, Excel Pivot Tables
- Basic Probability & Information Theory
- Preparing Data, Testing and Validation
- Data Mining Methods
  - Classification Using Naïve Bayes
  - Classification using Decision Trees
  - Association Rule Mining
  - Clustering

Who should take this class

- Required core class for MSBA students
- No prior technical knowledge is assumed
- Aptitude for learning technical concepts
  - Algorithms
  - Software (MS Excel, Weka)
  - Understand business concepts
- Course is recommended for students interested in understanding the techniques and applications of data mining and acquiring hands-on skills for making intelligent business decisions in data-rich organizations.

Guidelines for this course

- No cell phone or laptop usage in-class unless required for class

- UCI Office of Academic Integrity: Do not
  - Post class materials on non-UCI websites
  - Do not share individual assignments
- Learn about US businesses, language and culture
- Groups for class project

Motivation: Business Value of Customer Data

- What can a retail store do with shopping basket data?

How to increase customer viewership through tune in ads?

KINGS
KINGS
NAGAW...LD.COM
MANTO
HALEO

© ONE Championship

PUBLIC CARRIER

How can big data be used to reduce fuel consumption?

© Prof. V Abhishek,

How data is utilized is key

- Accurate, timely intelligence
- Rapid deployment
- Effective tactics
- Relentless follow-up and assessment

© Prof. V Choudhary,

2022.11.24

Business Intelligence & Data Mining

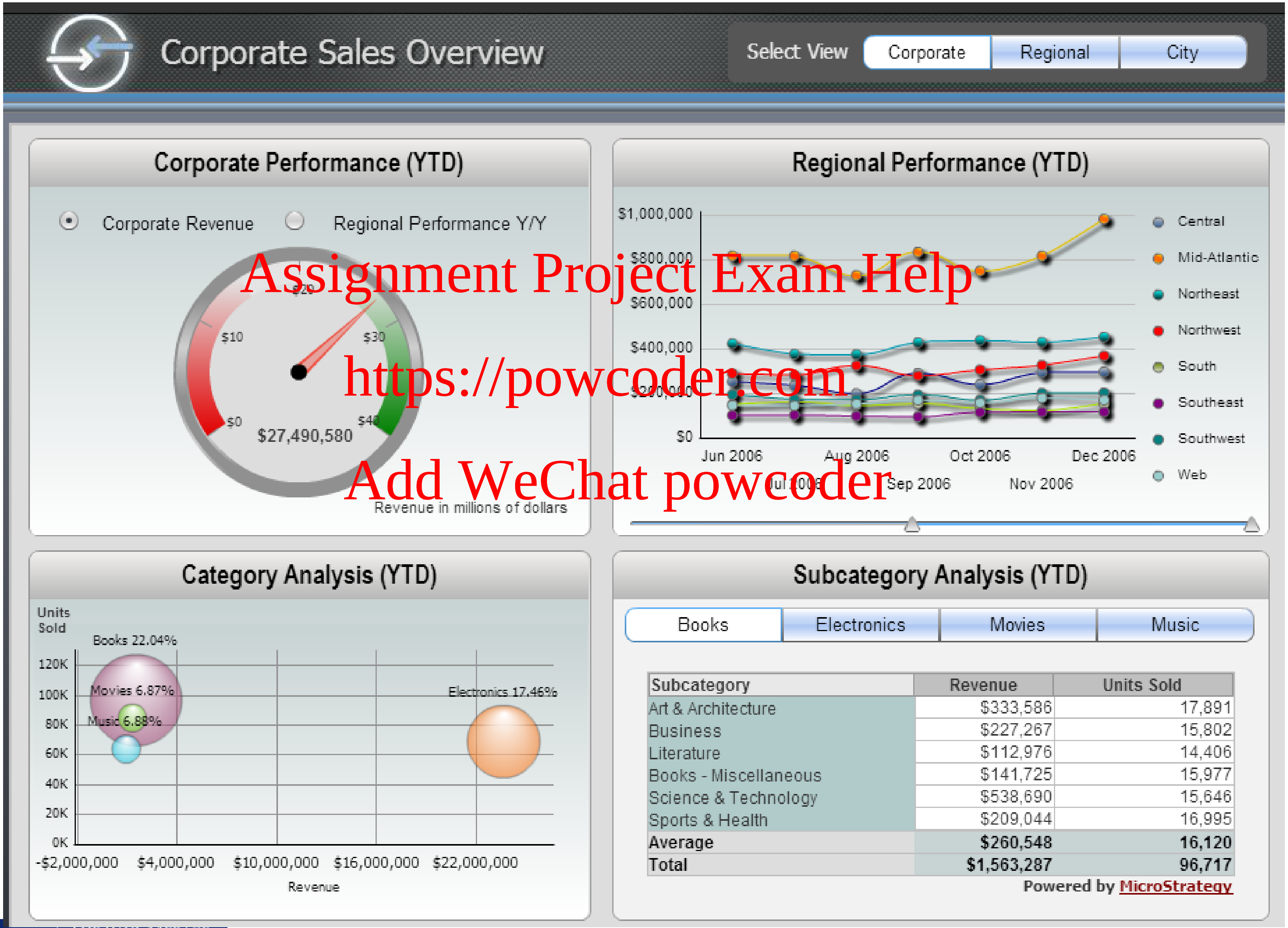
Business Intelligence: Components

- Reporting
  - Tooth-paste sales in September?
  - Cosmetics purchased last quarter?

Business Intelligence

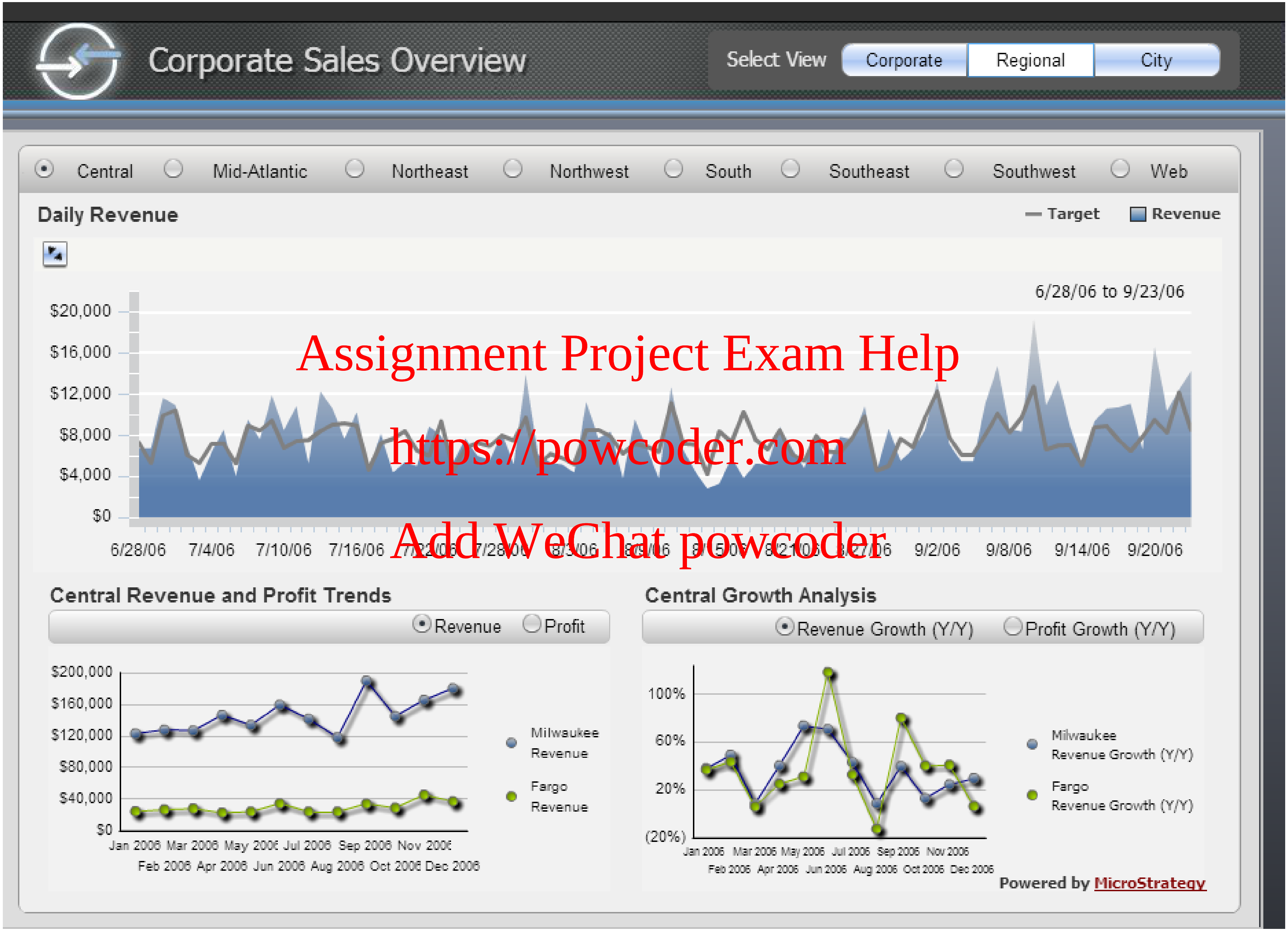
- Reporting
  - Tooth-paste sales in September?
  - Cosmetics purchased last quarter?
- Dashboards
  - Easy access to performance metrics
  - Ability to drill down into the data

Example of Dashboard from MicroStrategy


Corporate Sales Overview
Select View
Corporate
Regional
City
Corporate Performance (YTD)
● Corporate Revenue
○ Regional Performance Y/Y
$0
$10
$20
$30
$40
$27,490,580
Revenue in millions of dollars
Regional Performance (YTD)
$1,000,000
$800,000
$600,000
$400,000
$200,000
$0
Jun 2006
Jul 2006
Aug 2006
Sep 2006
Oct 2006
Nov 2006
Dec 2006
Central
Mid-Atlantic
Northeast
Northwest
South
Southeast
Southwest
Web
Category Analysis (YTD)
Units Sold
Books 22.04%
Movies 6.87%
Music 6.88%
Electronics 17.46%
120K
100K
80K
60K
40K
20K
0K
-$2,000,000
$4,000,000
$10,000,000
$16,000,000
$22,000,000
Revenue
Subcategory Analysis (YTD)
Books
Electronics
Movies
Music

Subcategory	Revenue	Units Sold
Art & Architecture	$333,586	17,891
Business	$227,267	15,802
Literature	$112,976	14,406
Books - Miscellaneous	$141,725	15,977
Science & Technology	$538,690	15,646
Sports & Health	$209,044	16,995
Average	**$260,548**	**16,120**
Total	**$1,563,287**	**96,717**

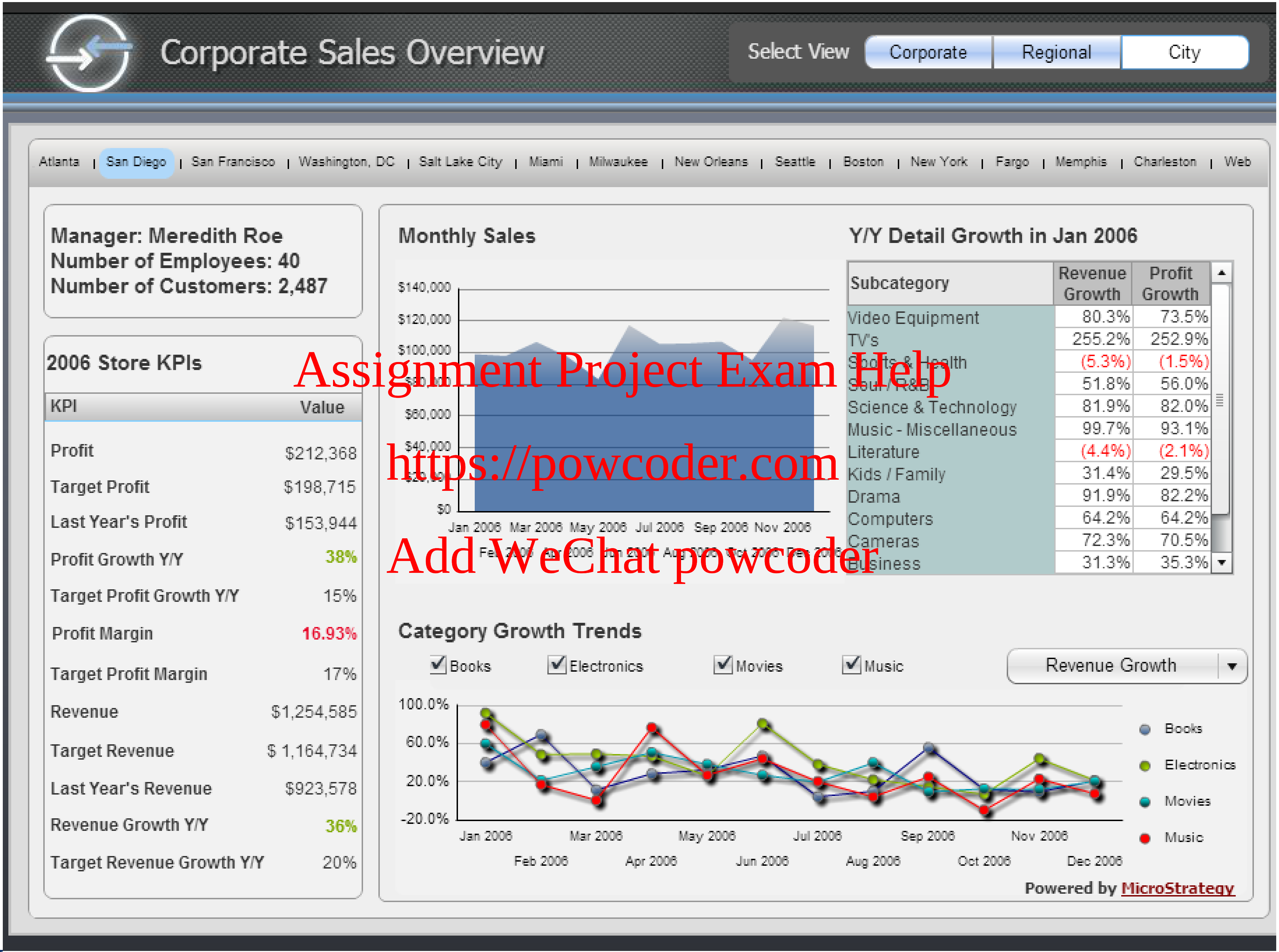
Powered by MicroStrategy


Corporate Sales Overview
Select View
Corporate
Regional
City
Daily Revenue
Target
Revenue
6/28/06 to 9/23/06
$20,000
$16,000
$12,000
$8,000
$4,000
$0
6/28/06
7/4/06
7/10/06
7/16/06
7/22/06
7/28/06
8/3/06
8/9/06
8/15/06
8/21/06
8/27/06
9/2/06
9/8/06
9/14/06
9/20/06
Central Revenue and Profit Trends
$200,000
$160,000
$120,000
$80,000
$40,000
$0
Jan 2006
Feb 2006
Mar 2006
Apr 2006
May 2006
Jun 2006
Jul 2006
Aug 2006
Sep 2006
Oct 2006
Nov 2006
Dec 2006
Milwaukee Revenue
Fargo Revenue
Central Growth Analysis
100%
60%
20%
(20%)
Jan 2006
Feb 2006
Mar 2006
Apr 2006
May 2006
Jun 2006
Jul 2006
Aug 2006
Sep 2006
Oct 2006
Nov 2006
Dec 2006
Milwaukee Revenue Growth (Y/Y)
Fargo Revenue Growth (Y/Y)
Powered by MicroStrategy

● Central ○ Mid-Atlantic ○ Northeast ○ Northwest ○ South ○ Southeast ○ Southwest ○ Web

● Revenue ○ Profit

● Revenue Growth (Y/Y) ○ Profit Growth (Y/Y)


Corporate Sales Overview
Select View
Corporate
Regional
City
Atlanta | San Diego | San Francisco | Washington, DC | Salt Lake City | Miami | Milwaukee | New Orleans | Seattle | Boston | New York | Fargo | Memphis | Charleston | Web

Manager: Meredith Roe
Number of Employees: 40
Number of Customers: 2,487

2006 Store KPIs

KPI	Value
Profit	$212,368
Target Profit	$198,715
Last Year's Profit	$153,944
Profit Growth Y/Y	38%
Target Profit Growth Y/Y	15%
Profit Margin	16.93%
Target Profit Margin	17%
Revenue	$1,254,585
Target Revenue	$ 1,164,734
Last Year's Revenue	$923,578
Revenue Growth Y/Y	36%
Target Revenue Growth Y/Y	20%

Monthly Sales

$140,000
$120,000
$100,000
$80,000
$60,000
$40,000
$20,000
$0
Jan 2006
Feb 2006
Mar 2006
Apr 2006
May 2006
Jun 2006
Jul 2006
Aug 2006
Sep 2006
Oct 2006
Nov 2006
Dec 2006

Y/Y Detail Growth in Jan 2006

Subcategory	Revenue Growth	Profit Growth
Video Equipment	80.3%	73.5%
TV's	255.2%	252.9%
Sports & Health	(5.3%)	(1.5%)
Soul / R&B	51.8%	56.0%
Science & Technology	81.9%	82.0%
Music - Miscellaneous	99.7%	93.1%
Literature	(4.4%)	(2.1%)
Kids / Family	31.4%	29.5%
Drama	91.9%	82.2%
Computers	64.2%	64.2%
Cameras	72.3%	70.5%
Business	31.3%	35.3%

Category Growth Trends

☑ Books ☑ Electronics ☑ Movies ☑ Music

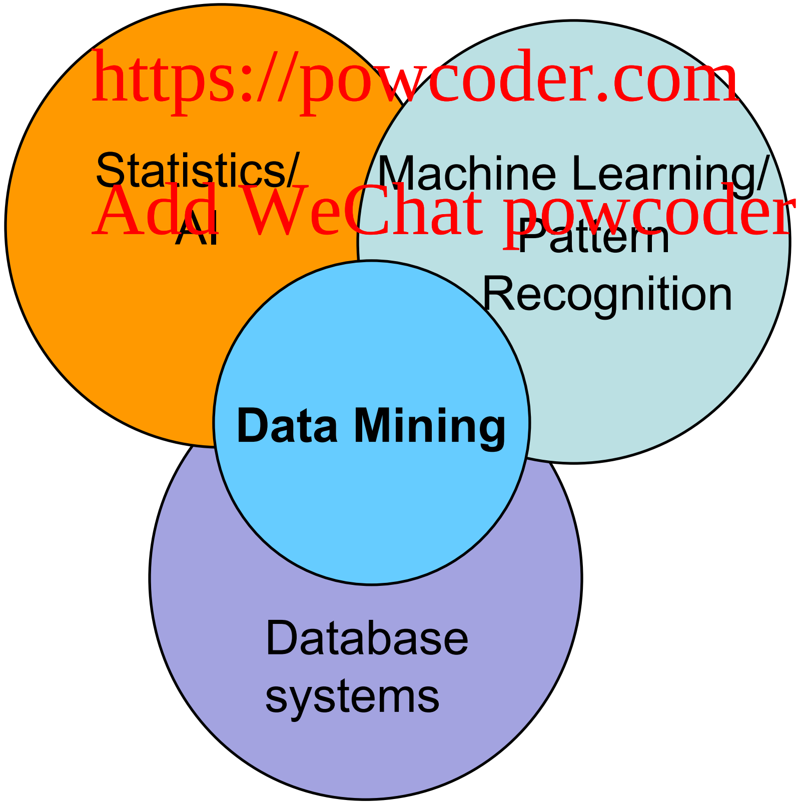
Revenue Growth
100.0%
60.0%
20.0%
-20.0%
Jan 2006
Feb 2006
Mar 2006
Apr 2006
May 2006
Jun 2006
Jul 2006
Aug 2006
Sep 2006
Oct 2006
Nov 2006
Dec 2006
Books
Electronics
Movies
Music
Powered by MicroStrategy

Business Intelligence

- Reporting
  - Tooth-paste sales in September?
  - Cosmetics purchased last quarter?
- Dashboards
  - Easy access to performance metrics
  - Ability to drill down into the data
- Data Mining

What Is Data Mining?

Data mining is a ***process*** *of discovering and interpreting (unknown) patterns in data to solve a business problem.*



Statistics/ AI

Machine Learning/ Pattern Recognition

Data Mining

Database systems

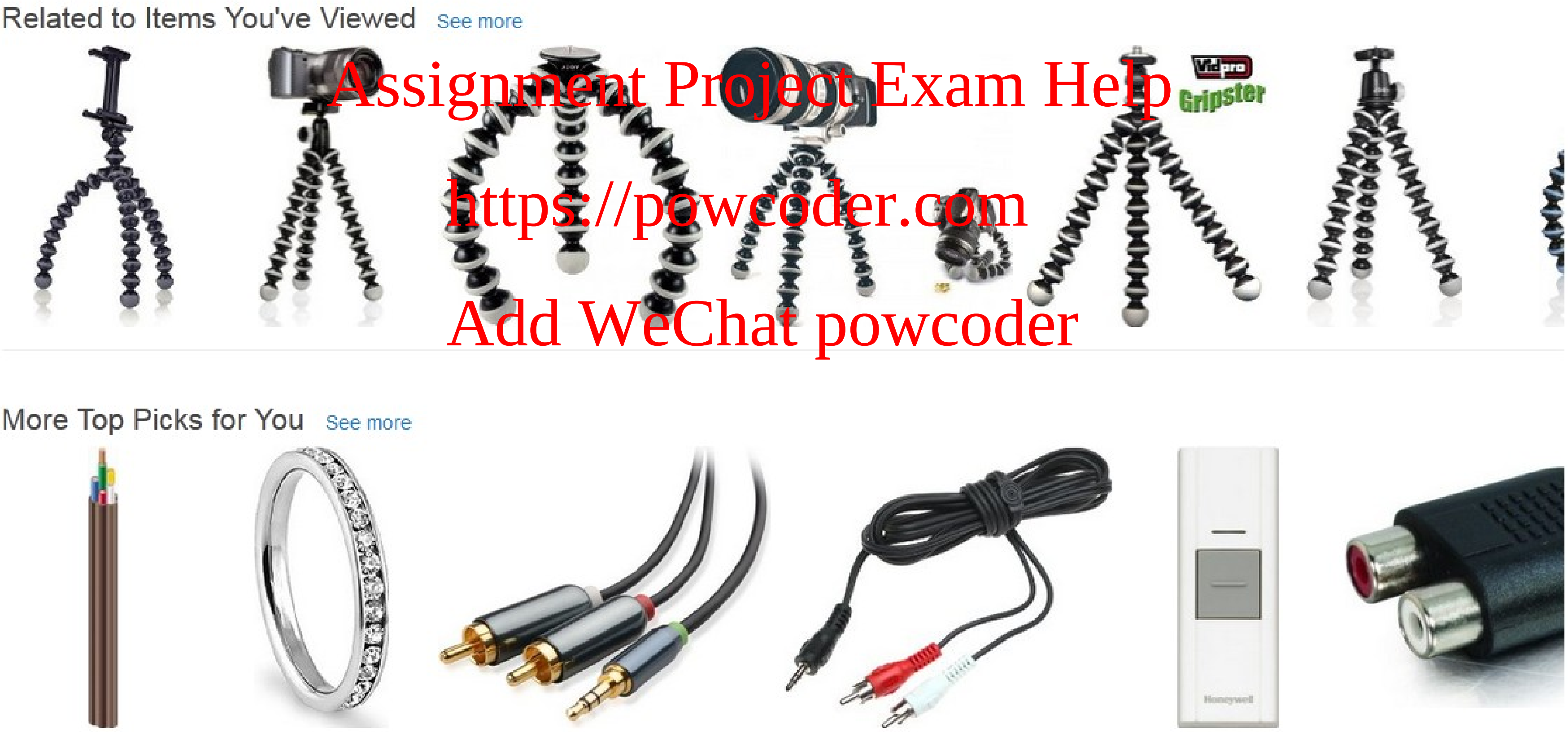
Data Sources for Data Mining

- Public and private data
  - Demographic data
  - Customer data
  - Transactions data
  - Product data
- Accounting and marketing data
  - Web logs (click-stream data)
  - Application server-logs
  - Third party server-logs
- Data integration is an important challenge

Who Uses Data Mining?

amazon.com

Amazon’s Personalized Web Pages



Related to Items You've Viewed See more

More Top Picks for You See more

Vidpro
Gripster
Honeywell

Who Uses Data Mining?

- Firms with corporate offices in OC

Yum!

Experian

A world of insight

INGRAM MICRO®

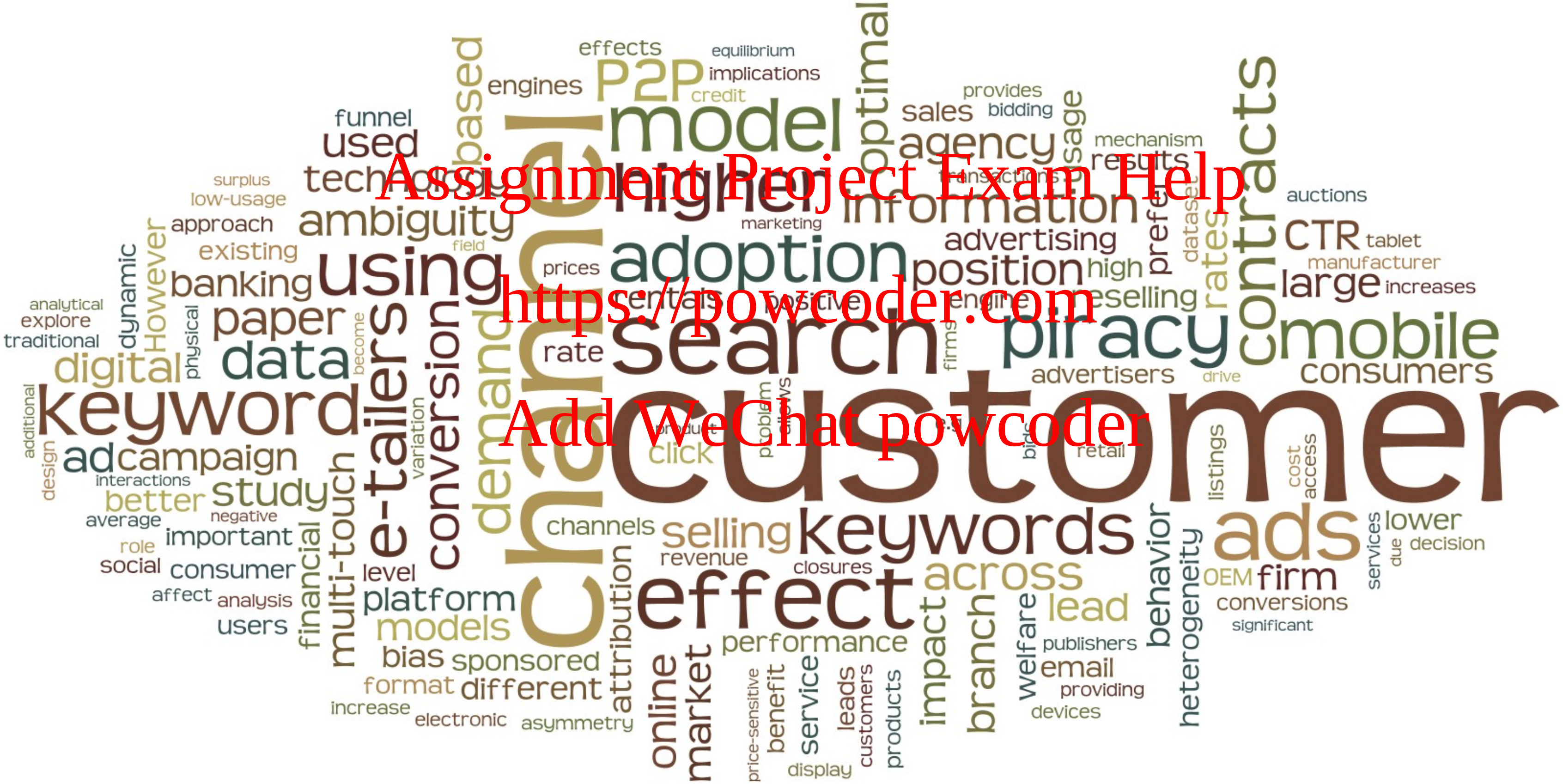
houzz

Data Mining Methods...

- Classification
- Clustering
- Association Rule Discovery
- Decision Trees

- Visualization
- Collaborative filtering
- Text Mining
- Neural networks and Deep learning

Examples of Visualizing Data: Word Cloud


effects
equilibrium
implications
P2P
credit
engines
based
optimal
provides
bidding
sales
usage
funnel
used
model
agency
mechanism
results
surplus
low-usage
technology
higher
transactions
information
auctions
approach
ambiguity
marketing
advertising
CTR
tablet
existing
field
channel
dataset
preference
rates
contracts
manufacturer
dynamic
However
banking
using
prices
adoption
position
high
large
increases
analytical
explore
traditional
physical
paper
become
potential
positive
engine
reselling
data
rate
search
piracy
mobile
digital
firms
advertisers
drive
consumers
additional
keyword
e-tailers
conversion
demand
product
problem
allows
customer
design
ad
campaign
variation
click
bids
retail
listings
cost
access
interactions
study
multi-touch
better
negative
channels
selling
keywords
behavior
ads
lower
average
services
due
decision
role
important
financial
revenue
closures
heterogeneity
social
consumer
level
across
OEM
firm
affect
analysis
platform
attribution
effect
lead
conversions
users
models
impact
branch
welfare
significant
bias
sponsored
performance
publishers
format
different
online
market
email
increase
price-sensitive
benefit
service
leads
customers
products
providing
electronic
asymmetry
display
devices

Classification: Definition

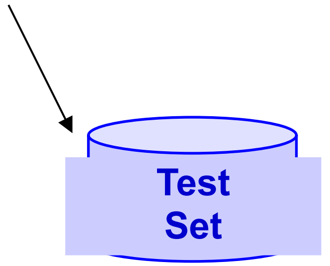
- Given a collection of records (*training set*)
  - Each record contains a set of *attributes*, one of the attributes is the *class*.
- Find a *model* for class attribute as a function of the values of other attributes.
- Goal: previously unseen records should be assigned a class as accurately as possible.

Classification Example

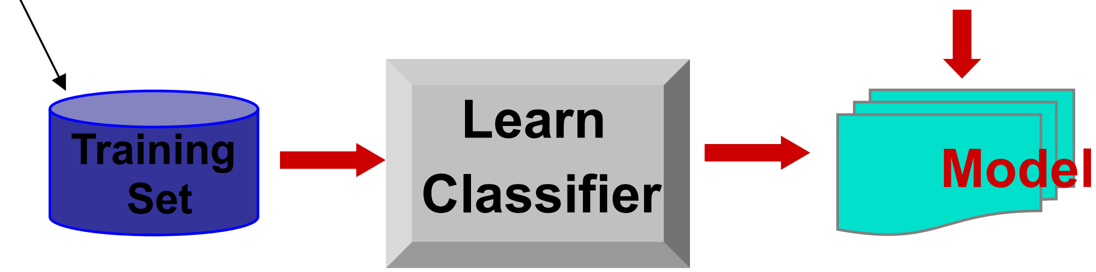
categorical categorical continuous class

Tid	**Refund**	**Marital Status**	**Taxable Income**	**Cheat**
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Refund	**Marital Status**	**Taxable Income**	**Cheat**
No	Single	75K	?
Yes	Married	50K	?
No	Married	150K	?
Yes	Divorced	90K	?
No	Single	40K	?
No	Married	80K	?



Test Set



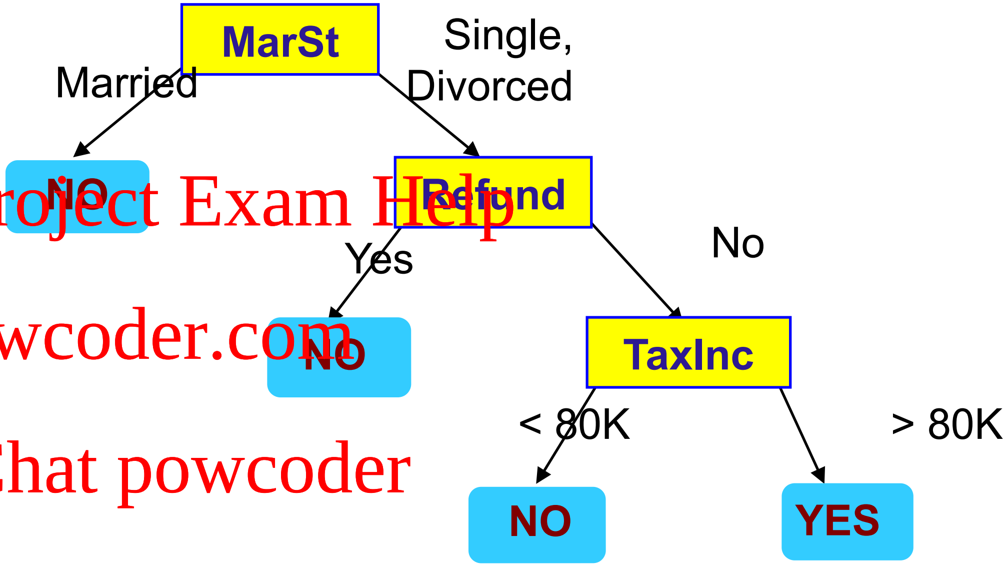
Training Set

Learn Classifier

Model

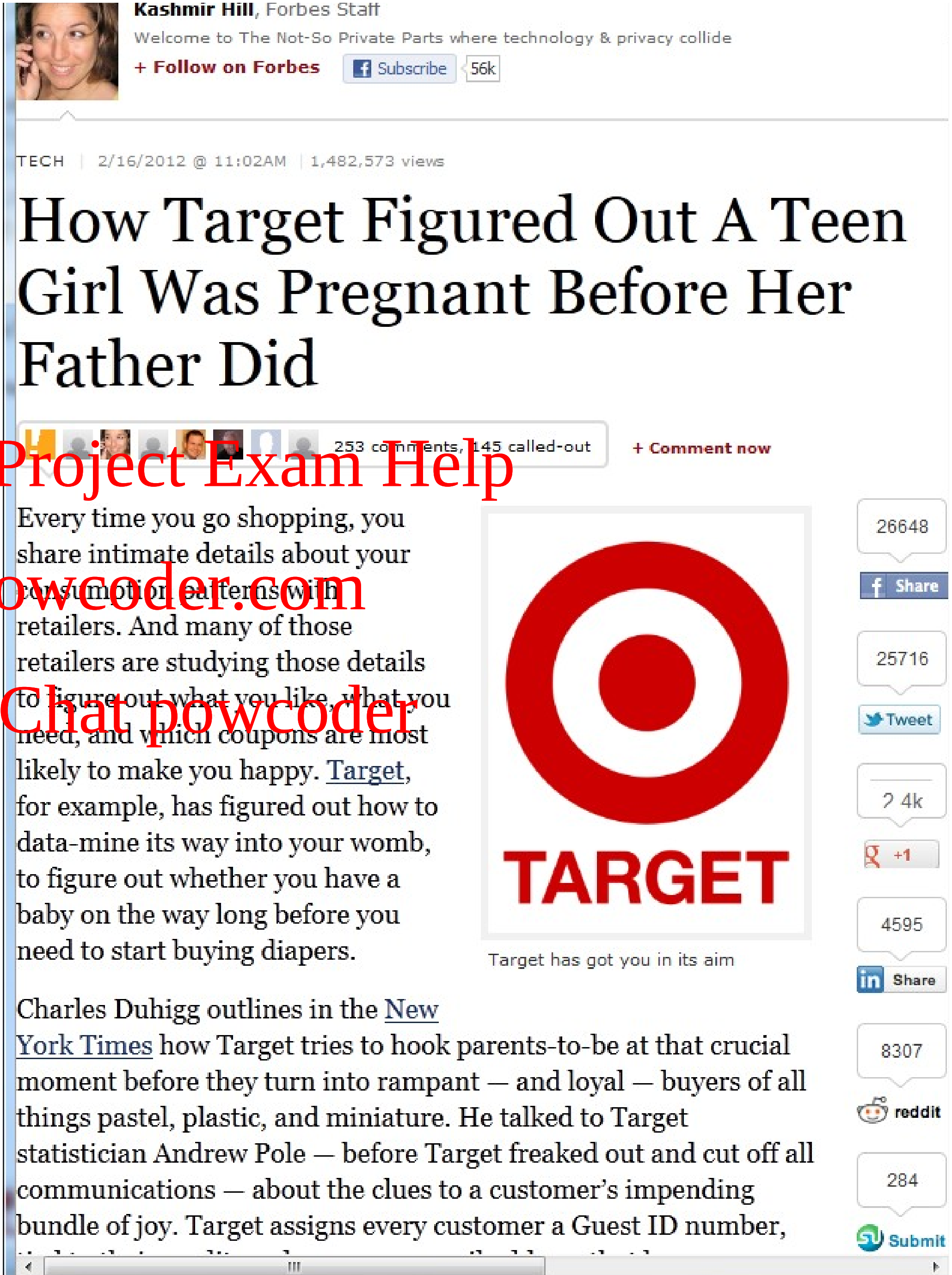
Example of Decision Tree

	categorical	categorical	continuous	class
Tid	**Refund**	**Marital Status**	**Taxable Income**	**Cheat**
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes


MarSt
Married
Single, Divorced
NO
Refund
Yes
No
NO
TaxInc
< 80K
> 80K
NO
YES

(11, Yes, Divorced, 115K)

(12, No, Single, 150K)



Kashmir Hill, Forbes Staff

Welcome to The Not-So Private Parts where technology & privacy collide

TECH | 2/16/2012 @ 11:02AM | 1,482,573 views

How Target Figured Out A Teen Girl Was Pregnant Before Her Father Did

253 comments, 145 called-out

Every time you go shopping, you share intimate details about your consumption patterns with retailers. And many of those retailers are studying those details to figure out what you like, what you need, and which coupons are most likely to make you happy. Target, for example, has figured out how to data-mine its way into your womb, to figure out whether you have a baby on the way long before you need to start buying diapers.

Target has got you in its aim

Charles Duhigg outlines in the New York Times how Target tries to hook parents-to-be at that crucial moment before they turn into rampant — and loyal — buyers of all things pastel, plastic, and miniature. He talked to Target statistician Andrew Pole — before Target freaked out and cut off all communications — about the clues to a customer's impending bundle of joy. Target assigns every customer a Guest ID number,

Other Applications for Classification

- Direct Marketing – LL Bean
  - Goal: Reduce cost of mailing by *targeting* a set of consumers likely to buy a new cell-phone product.

Other Applications for Classification

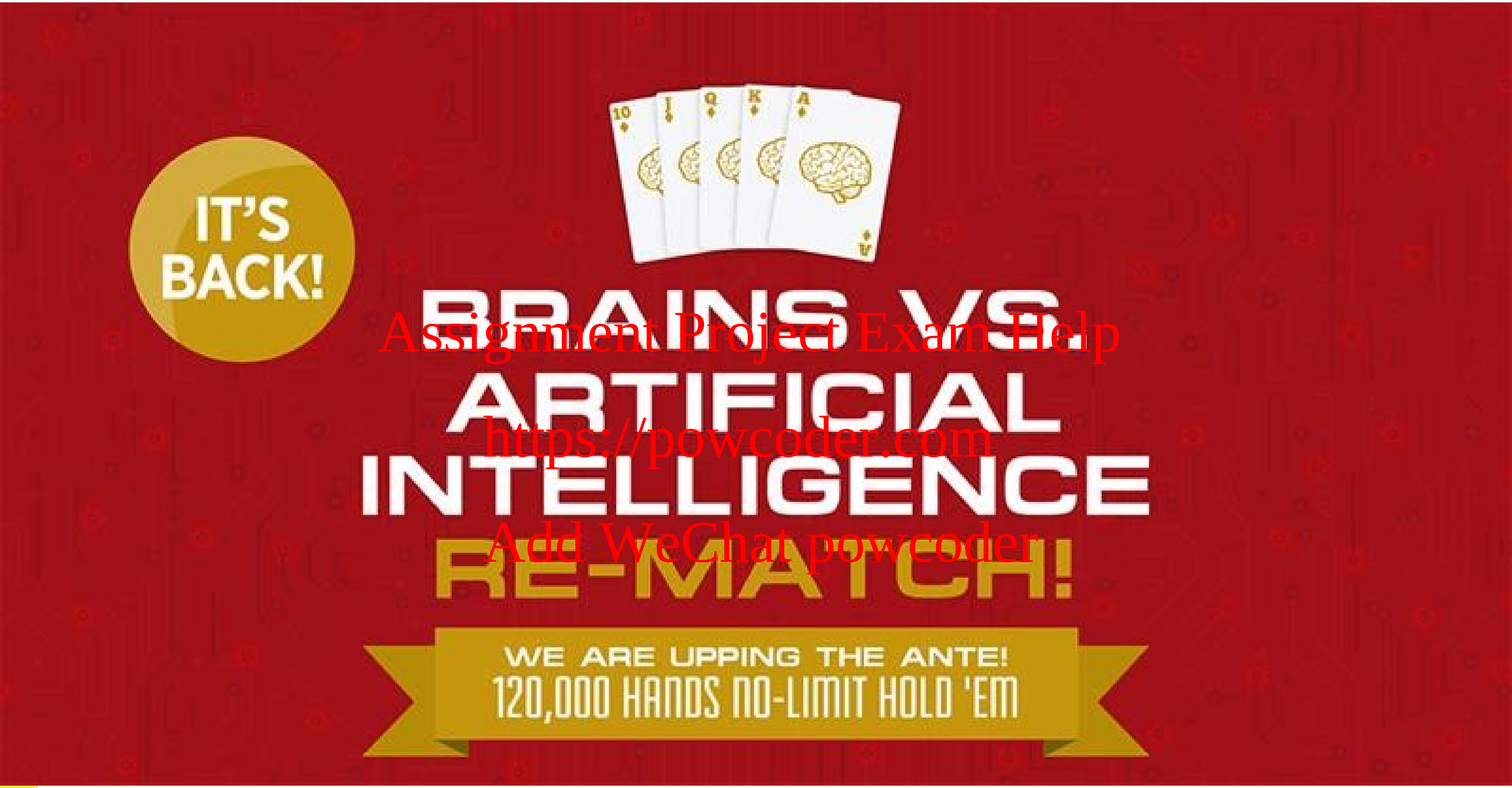
- Customer Attrition/Churn at AT&T
  - 16% in 2013
  - Goal: To predict whether a customer is likely to be lost to a competitor.

Who is Watson?

IBM Watson: Final Jeopardy!

https://www.youtube.com/watch?v=lI-M7O_bRNg

0 to 0:25; 1:10 to 2:03; 3:10 to 4:27; 6:30 to 7:40

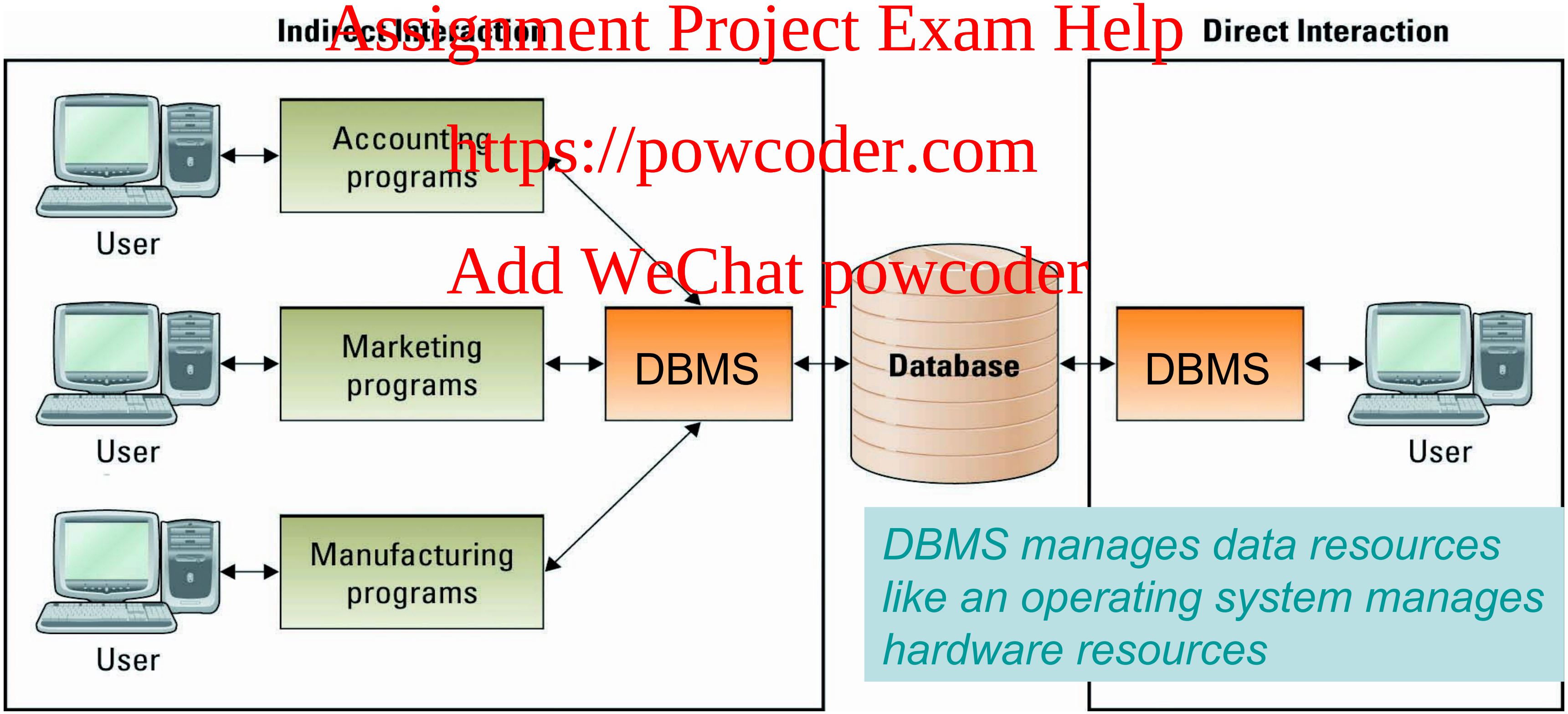

IT'S BACK!
BRAINS VS. ARTIFICIAL INTELLIGENCE
RE-MATCH!
WE ARE UPPING THE ANTE!
120,000 HANDS NO-LIMIT HOLD 'EM

https://www.riverscasino.com/pittsburgh/BrainsVsAI

Structured Query Language

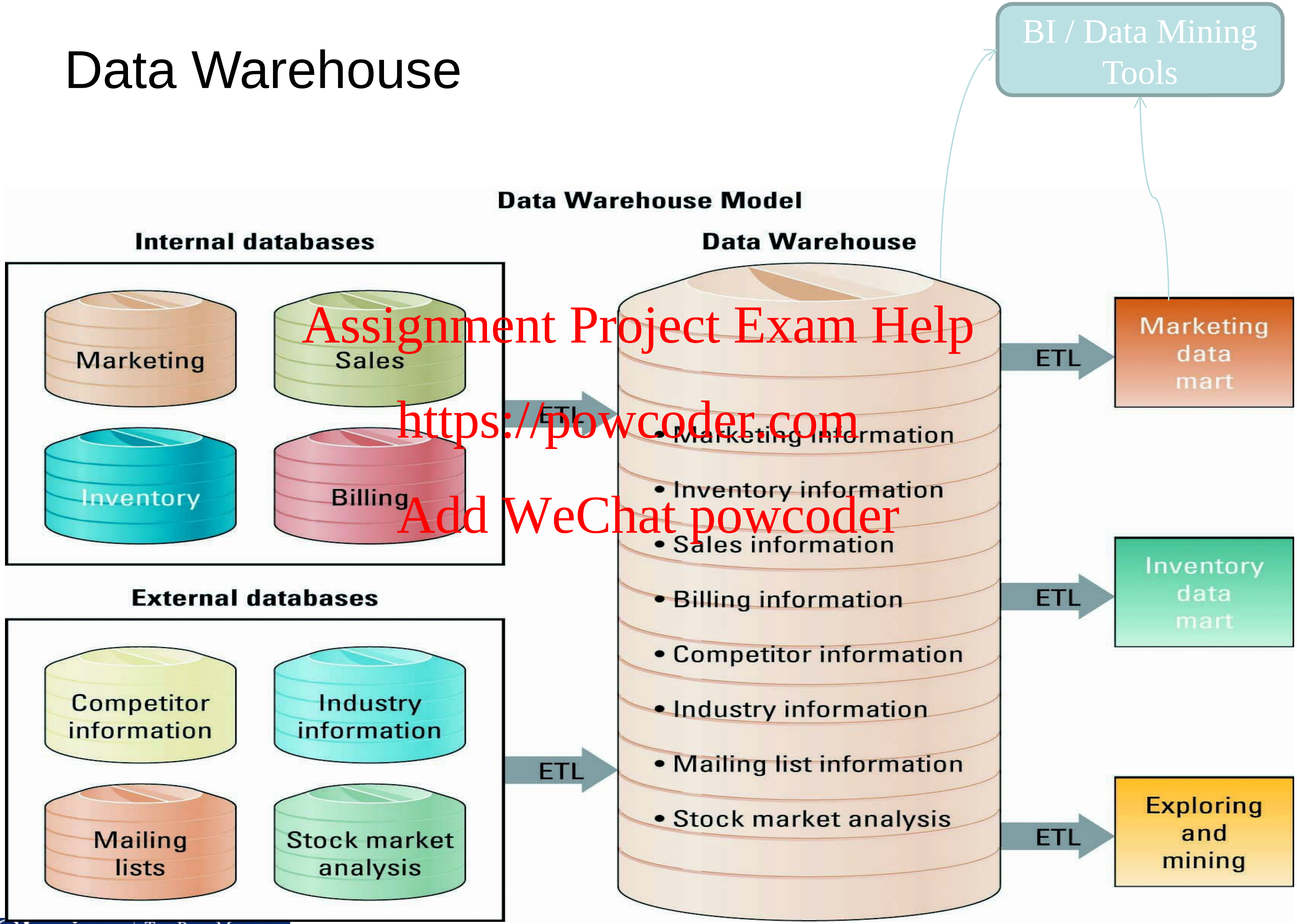
Database Management Systems *(DBMS)*

- Software through which users and application programs interact with a database


Indirect Interaction
Direct Interaction
User
Accounting programs
User
Marketing programs
User
Manufacturing programs
DBMS
Database
DBMS
User

DBMS manages data resources like an operating system manages hardware resources

Data Warehouse


BI / Data Mining Tools
Data Warehouse Model
Internal databases
Marketing
Sales
Inventory
Billing
External databases
Competitor information
Industry information
Mailing lists
Stock market analysis
ETL
ETL
Data Warehouse
• Marketing information
• Inventory information
• Sales information
• Billing information
• Competitor information
• Industry information
• Mailing list information
• Stock market analysis
ETL
ETL
ETL
Marketing data mart
Inventory data mart
Exploring and mining

SQL Example

PC

PCModel
Speed
RAM
HD
CD
Price

Product

Maker
Model
Type

Laptop

LaptopModel
Speed
RAM
HD
Screen
Price

Printer

PrinterModel
Color
Type
Price

Primary Key
and
Foreign Key

Sample data for table PC

PCmodel	speed	ram	hd	cd	price
1001	133	16	1.6	6x	1595
1002	120	16	1.6	6x	1399
1003	166	24	2.5	6x	1899
1004	166	32	2.5	8x	1999
1005	166	16	2	8x	1999
1006	200	32	3.1	8x	2099
1007	200	32	3.2	8x	2349
1008	180	32	2	8x	3699
1009	200	32	2.5	8x	2599
1010	160	16	1.2	8x	1495

Sample data for PRODUCT

maker	model	type
A	1001	pc
A	1002	pc
A	1003	pc
B	1004	pc
B	1006	pc
B	3002	printer
B	3004	printer
C	1005	pc
C	1007	pc
D	1008	pc
D	1009	pc
D	1010	pc
D	2001	laptop
D	2002	laptop
D	2003	laptop
D	3001	printer
D	3003	printer
E	2004	laptop
E	2008	laptop
F	2005	laptop
G	2006	laptop
G	2007	laptop
H	3005	printer
I	3006	printer

Sample data for table LAPTOP					
LaptopModel	**speed**	**ram**	**hd**	**screen**	**price**
2001	100	20	1.1	9.5	1999
2002	117	12	0.75	11.3	2499
2003	117	32	1	10.4	3599
2004	133	16	1.1	11.2	3499
2005	133	16	1	11.3	2599
2006	120	8	0.81	12.1	1999
2007	150	16	1.35	12.1	4799
2008	120	16	1.1	12.1	2099

Sample data for table PRINTER			
PrinterModel	**color**	**type**	**price**
3001	TRUE	ink jet	275
3002	TRUE	ink jet	269
3003	FALSE	laser	829
3004	FALSE	laser	879
3005	FALSE	ink jet	180
3006	TRUE	dry	470

The SELECT Statement

- Used for queries on single or multiple tables
  Clauses of the SELECT statement:
  - SELECT - List the columns (and expressions) that should be returned from the query
  - FROM - Indicate the table(s) or view(s) from which data will be obtained
  - WHERE - Indicate the conditions under which a row will be included in the result
    Optional clauses:
    - GROUP BY - Indicate categorization of results
    - HAVING - Indicate the conditions under which a category (group) will be included
    - ORDER BY - Sorts the result according to specified criteria

SQL Query

Printer Table

printermodel	color	type	price
3004	FALSE	laser	879
3005	FALSE	ink jet	180
3006	TRUE	dry	470

- SELECT <attribute list> FROM <table list>;

- SELECT printermodel, color FROM printer;

- SELECT <attribute list> FROM <table list> WHERE <condition>;
- SELECT printermodel, color FROM printer WHERE price<300;

Aggregation Operators

- SUM
  - SELECT SUM(price) FROM PC;
- AVG
  - SELECT AVG(price) FROM PC WHERE speed>800;
- MIN
  - SELECT MIN(price) FROM PC where ram=256;
- MAX
  - SELECT MAX(price) FROM PC WHERE cd=‘32x’;
- COUNT
  - SELECT COUNT(model) FROM PC;
- Keyword DISTINCT (not an aggregation operator)
  - SELECT DISTINCT type FROM printer;

Practice

PRODUCT(maker, **model**, type)
PC(**PCmodel**, speed,ram, hd, cd, price)
LAPTOP(**Laptopmodel**,speed,ram,hd,screen,price)
PRINTER(**Printermodel**,color,type,price)

- List the model and speed of PCs with price<2000.
- How many laptops are there with screen size>11.
- What is average price of laptops with ram=20.
- List all attributes of all color printers

Practice

Book(ISBN, Title, Publisher ID, Author, Cost, price)

Publisher(Publisher ID, Name, Location)

1. List the publishers and their location.
2. List all book titles from publisher with ID = 11
3. How many books exist in the store. (Assuming each book has one entry in the book table)
4. What is the total cost of the books?
5. List all distinct book titles
6. What is the expected profit per book for books published by Publisher ID 222?

Print the speed and price of PCs that cost less than $2000

A: Select speed, price from printer where price < 2000

B: Select speed, price from laptop where price < 2000

C: Select speed from PC where price < 2000

D: Select speed, price from product where price < 2000

E: None of the above

Print the price of the cheapest laptop

A: Select price from printer

B: Select min(price) from pc

C: Select min(price) from laptop

D: Select cheapest(price) from laptop

E: None of the above

Online help for learning SQL

- Tutorials:
- http://sqlzoo.net/wiki/SQL_Tutorial

- Practice with tables that are already populated with data:
- http://www.w3schools.com/sql/trysql.asp?filename=trysql_select_all

Joins - Queries From Multiple Tables

- Some queries require combining information from two or more tables.
- Example: I want the Model #, Manufacturer and Speed of all PCs. ‘maker’ is in the ‘product’ table, but ‘speed’ is in the ‘pc’ table.

- Example:

```
SELECT product.model,
       product.maker, pc.speed
FROM   product, pc
WHERE  product.model=pc.pcmodel;
```

Cartesian Product

- This helps to understand what a query from two tables does.

- Set theory concept - the Cartesian product of 2 sets: R and S is a set of pairs formed by choosing the an element from R and another from S.

- Example: R={a,b,c}, S={b,e}
  - R X S ={ a.b, a.e, b.b, b.e, c.b, c.e }

Cartesian Product Example

UCI_Students

SSN	L Name	F Name
000-00-0001	Doe	John
000-00-0002	Doe	Jason
000-00-0003	Smith	Jack

Football_Players

SSN	Team
000-00-0001	Chargers
111-11-1112	AntEaters
000-00-0003	49ers

- SELECT * FROM UCI_Students, Football_Players;
- SELECT UCI_Students.SSN, Football_players.SSN
  FROM UCI_Students, Football_Players;
- How can we get names of UCI students who also play football?

PRODUCT(maker, **model**, type)
PC(**PCmodel**, speed,ram, hd, cd, price)
LAPTOP(**Laptopmodel**,speed,ram,hd,screen,price)
PRINTER(**Printermodel**,color,type,price)

- List the model #, price of laptops made by maker ‘C’.

- List the model, type and price of color printers from maker ‘G’

- List model, speed and price for PCs and Laptops that have the same price. (Do not JOIN on *model*)

© Prof. V Choudhary,

Practice

Book(ISBN, Title, Publisher ID, Author, Cost, price)

Publisher(Publisher ID, Name, Location)

1. List all books by the publisher with name 'Wiley and Sons'
2. List all book titles and publisher names
3. List the ISBN and the name of publisher where the title of the book is the same as the location of the publisher.

© Prof. V Choudhary,

Practice 2

Student(Student_ID, LastName, FirstName, Phone, Street, City, Zip)

Registration(**Student_ID, Course_Number, Term**, Reg_date, Score)

Course(Number, Term, Name, Description, Room_Number, Bldg_Name, **Faculty_ID**)

Faculty(Faculty_ID, Last_Name, First_Name, phone, Office_Room, Office_Bldg)

Foreign Keys in **Bold**

© Prof. V Choudhary,

Next Session

- SQL Continued

© Prof. V Choudhary,