INFORMATION
ENGINEERING
ASSOCIATES

DAMA-CHICAGO, JUNE 15,2016

Predictive Analytics:
A Statistical Primer for Data Modelers

Presenter:
Bob Conway
Information Engineering Associates
Bob.Conway@InfoEngAssc.com
303-885-4811

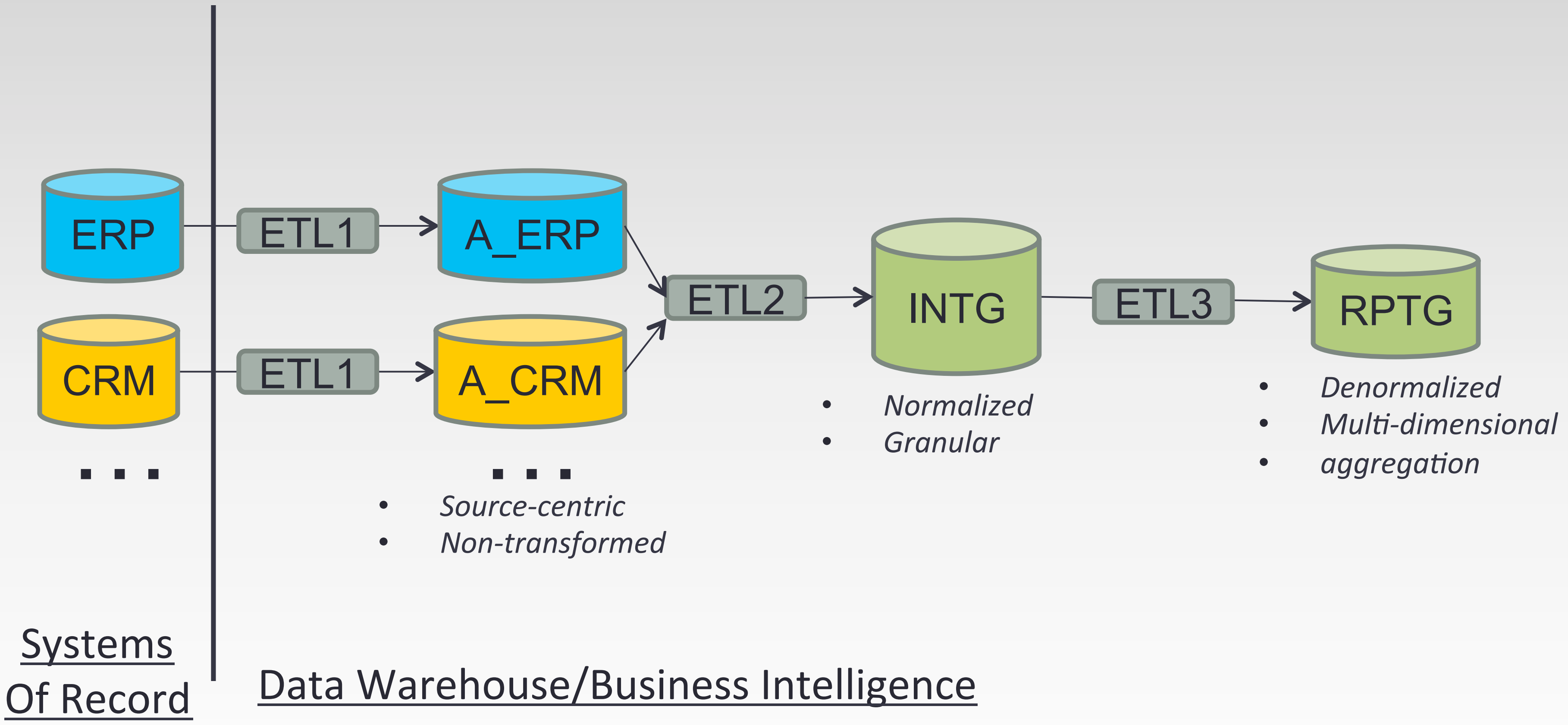
Copyright 2014© Information Engineering Associates. All Rights Reserved.

INFORMATION
ENGINEERING
ASSOCIATES

Predictive Analytics: Agenda

- **What** – Contrast to Descriptive Analytics
- **Why** – Value Proposition for Predictive Analytics
- **How** – Statistical basis of Predictive Analytics
- **Getting Started** with Predictive Analytics
- **Q&A**

‘Traditional’ DW/BI Architecture



ERP

CRM

. . .

ETL1

ETL1

A_ERP

A_CRM

. . .

- *Source-centric*
- *Non-transformed*

ETL2

INTG

- *Normalized*
- *Granular*

ETL3

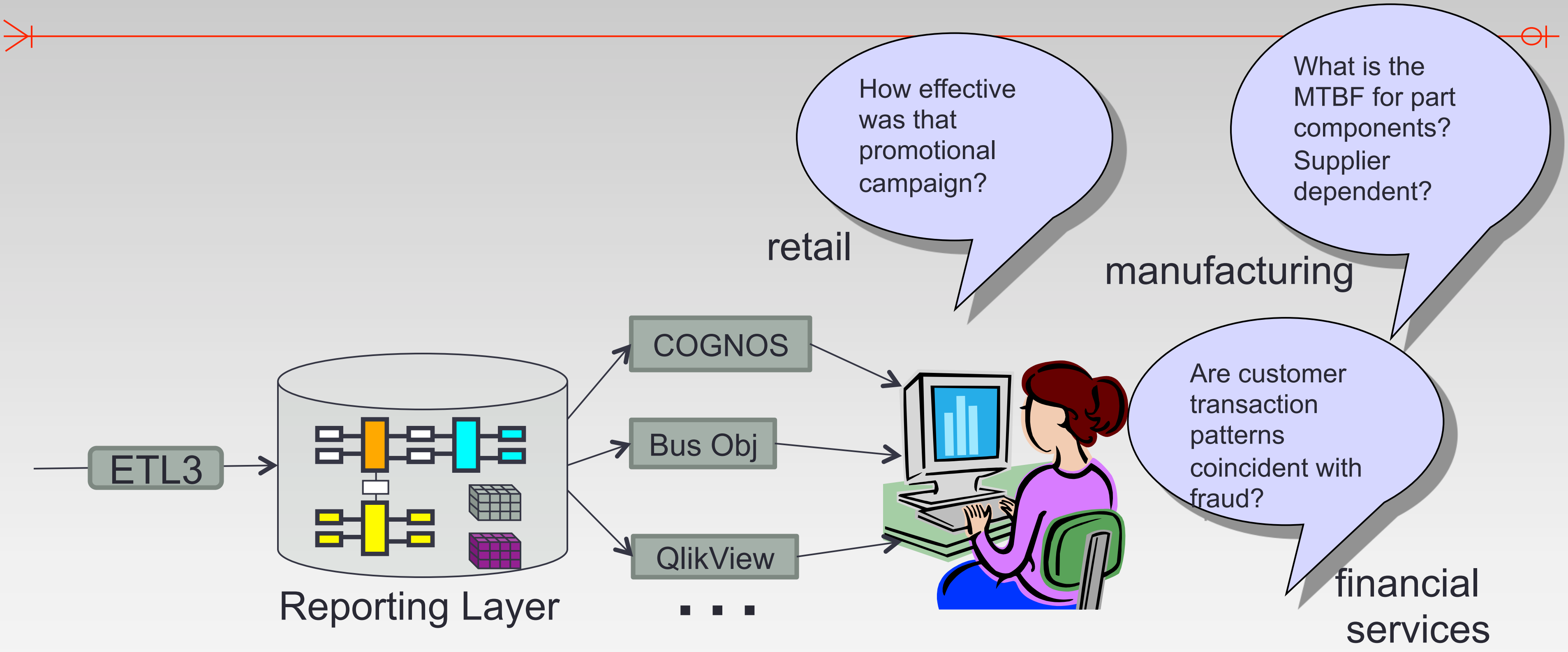
RPTG

- *Denormalized*
- *Multi-dimensional*
- *aggregation*

Systems
Of Record

Data Warehouse/Business Intelligence

Descriptive Analytics (OLAP)



ETL3

Reporting Layer

COGNOS

Bus Obj

QlikView

. . .

retail

How effective was that promotional campaign?

manufacturing

What is the MTBF for part components? Supplier dependent?

financial services

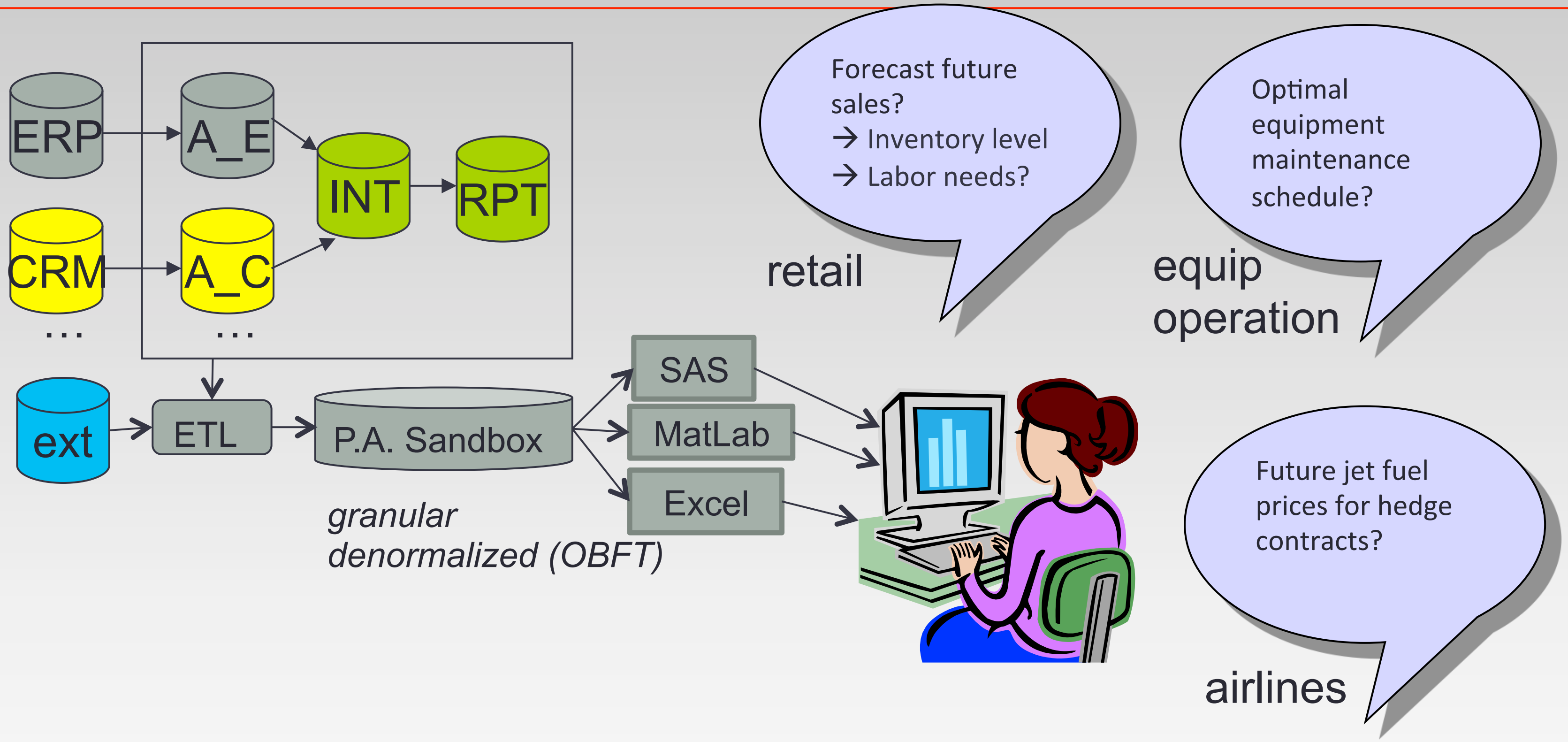
Are customer transaction patterns coincident with fraud?

Descriptive Analytics** – What **has** happened**?

Trends, patterns, exceptions in historic data

Monitor/Control, business process improvement

Predictive Analytics



ERP

CRM

…

A_E

A_C

…

INT

RPT

ext

ETL

P.A. Sandbox

granular
denormalized (OBFT)

SAS

MatLab

Excel

retail

Forecast future sales?
→ Inventory level
→ Labor needs?

equip operation

Optimal equipment maintenance schedule?

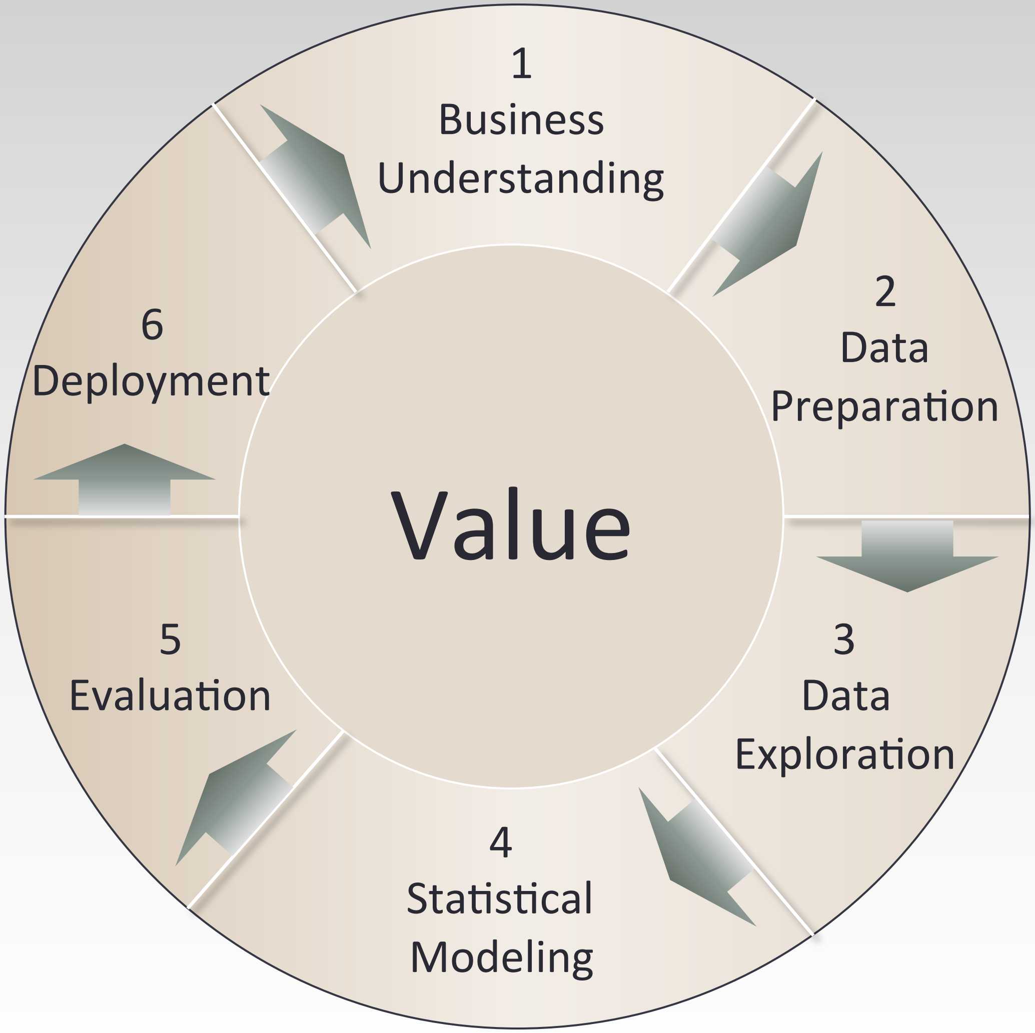
airlines

Future jet fuel prices for hedge contracts?

***Predictive Analytics –** What **will** happen?*

Advanced statistics →mathematical models
Forecast future state/behavior

Predictive Analytics Lifecycle



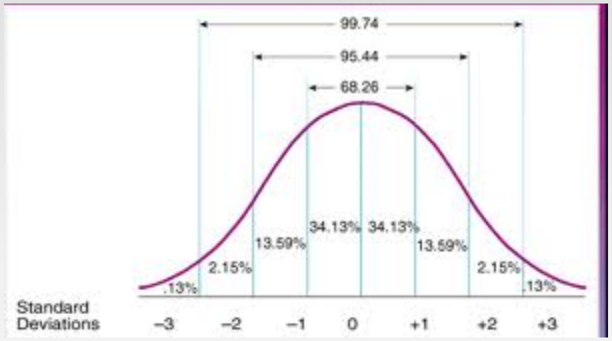
Copyright 2014© Information Engineering Associates. All Rights Reserved.

Statistics 101

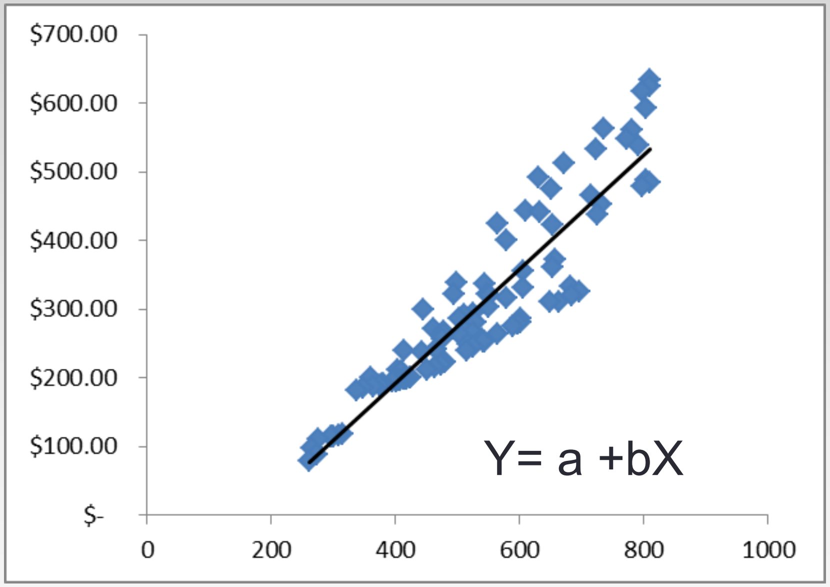
Mean (average), $\underline{X} = \Sigma (X_i)/N$

Variance, $S_x^{2} {}^{=} \Sigma (X_i – \underline{X})^2 / (N-1) = s_x^{2}$

Standard Deviation, $s_x = \surd s_x^{2}$



Simple Linear Regression



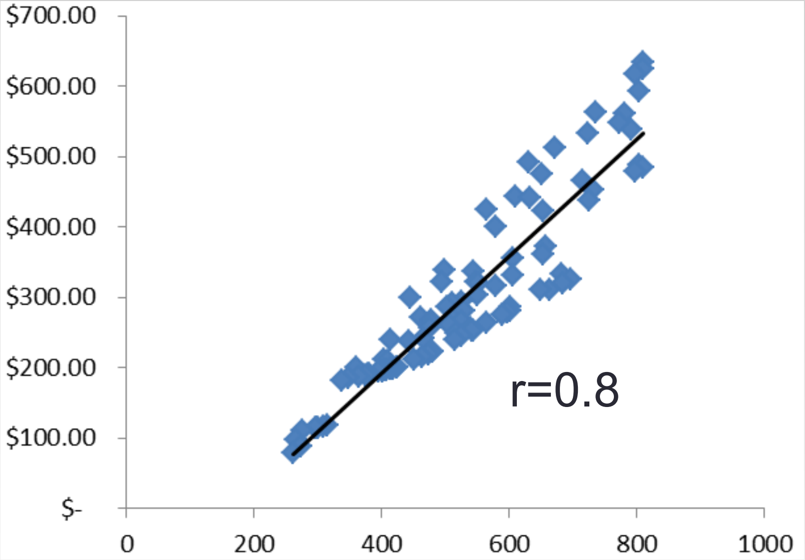
Covariance, $S_{xy}^2 = \Sigma(X_i - \underline{X})\ (Y_i - \underline{Y})/\ (N\text{-}1) = s_{xy}^2$

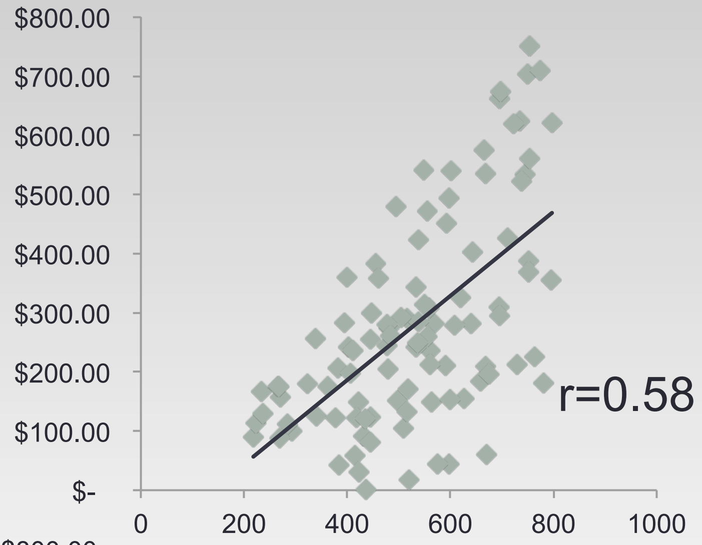
Slope, $b = s_{xy}\ /s_x$

Intercept, $a = \underline{Y} - b\underline{X}$

Best Fit: **minimizes variance** between **predicted** values and **observed** values

Simple Correlation





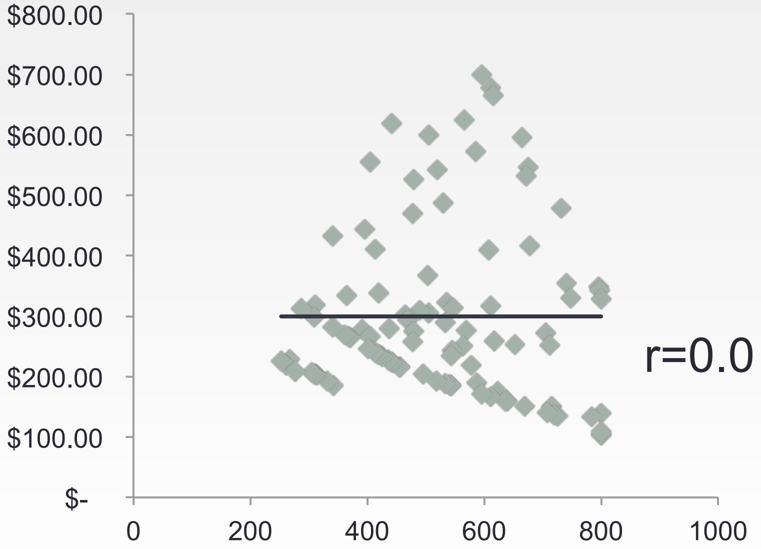
Correlation Coefficient, $r = s_{xy} / s_x s_y$

$-1 <= r <= +1$

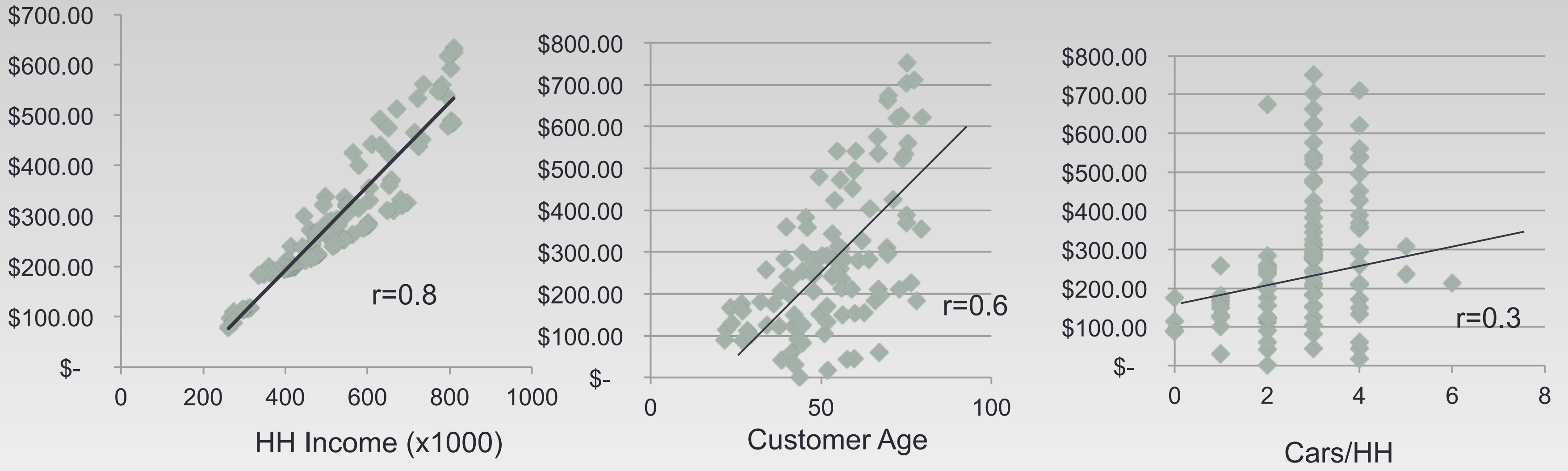
Correlation =/=> Cause and Effect

$0 <= r^2 <= +1$

r^2 ==> fraction of Y variance attributable to X



Multiple Correlation & Regression



Mercedes Sales = a + b(Income) + c(Age) + d(HH Cars)

$R^2 = 0.89$

Variable	Sales	Income	Age	Cars
Sales	1.0	0.8	0.6	0.3
Income		1.0	0.7	0.4
Age			1.0	0.3
Cars				1.0

Partial Correlation

Original Correlation Matrix

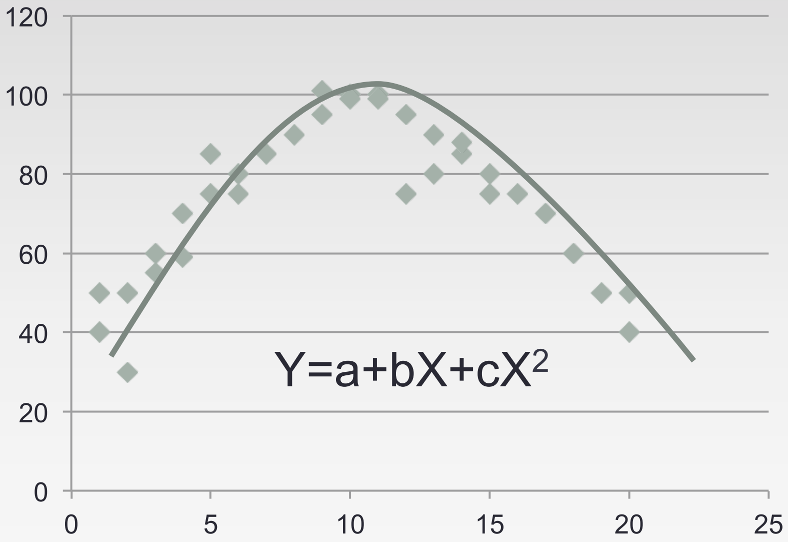
Variable	**Sales**	**Income**	**Age**	**Cars**
Sales	1.0	0.8	0.6	0.3
Income		1.0	0.7	0.4
Age			1.0	0.3
Cars				1.0

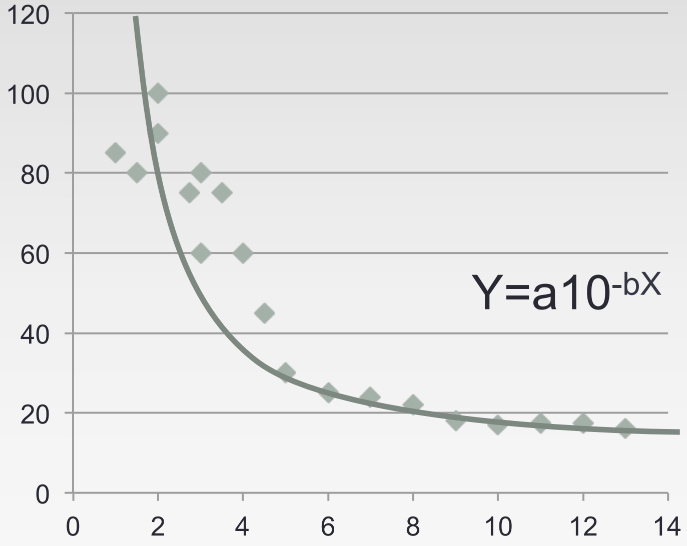
Stepwise (Partial) Correlation Matrix

Variable	**Sales**	**Income**	**Age***	**Cars***
Sales	1.0	0.79	-0.48	0.09
Income		1.0		
Age			1.0	
Cars				1.0

Mercedes Sales = a’ + b’(HH Income) - c’(Age)

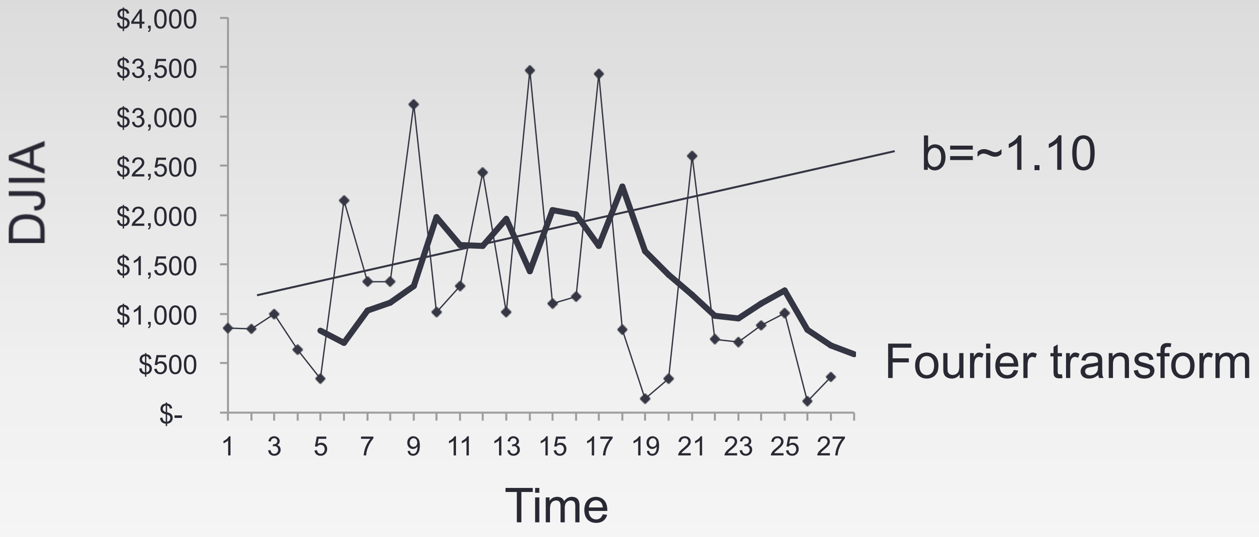
Nonlinear Regression





Does the mathematical model make business sense?

Periodic (Cyclic) Models



$$DJIA = a + bT + cSIN(T_c - T) + dSIN(T_d - T) + \ldots$$

Getting Started Suggestions

1. Bone up on statistics/predictive modeling
   - www.coursera.org – free, online classes
   - Books – “Predictive Analytics” by Conrad Carlberg
   - Leverage in-house methods/tools expertise
2. Engage a user-partner/business problem
   - Evaluate current forecast process and impact
   - Explore the data – metrics, sensitivity variables
   - Prototype a model – test it against actual data
3. Estimate business impact ($) – what if scenarios, precasting
4. Get started with simple tools (Excel with advance stat plug-ins)
5. Parallel trials with current process and proposed predictive model
   - Continuous monitoring and refinement

Questions

Copyright 2014© Information Engineering Associates. All Rights Reserved.