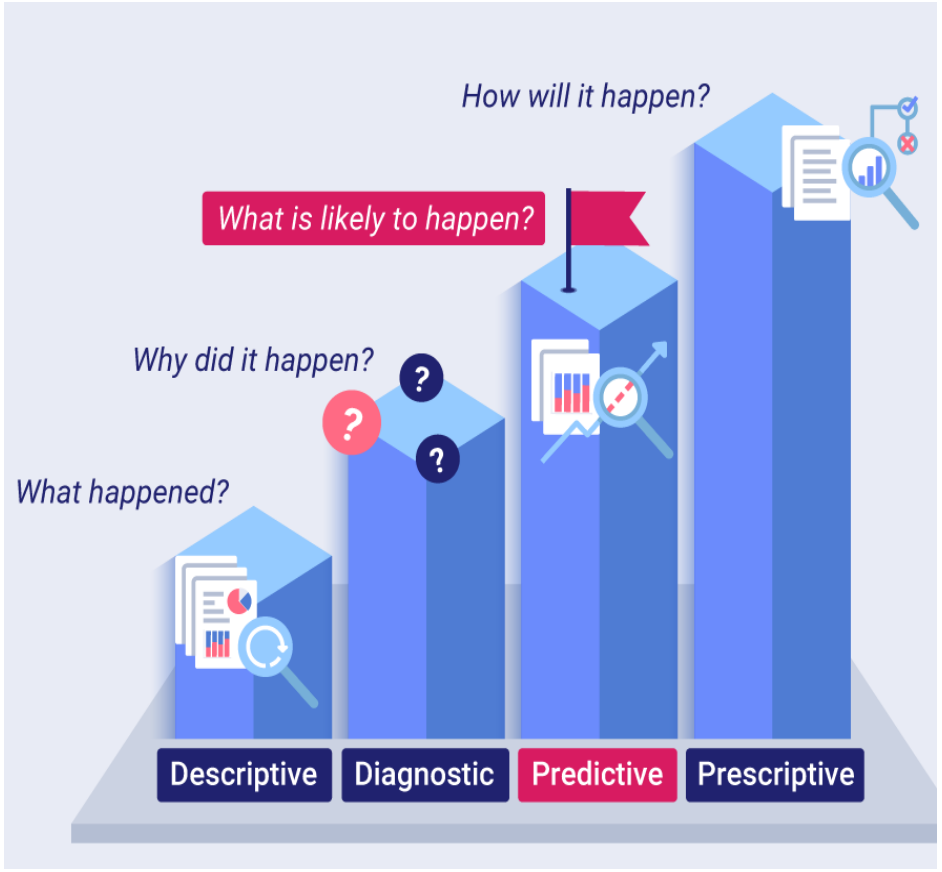
Data Driven Decision Making

Predictive Analysis

What is Predictive Analysis?



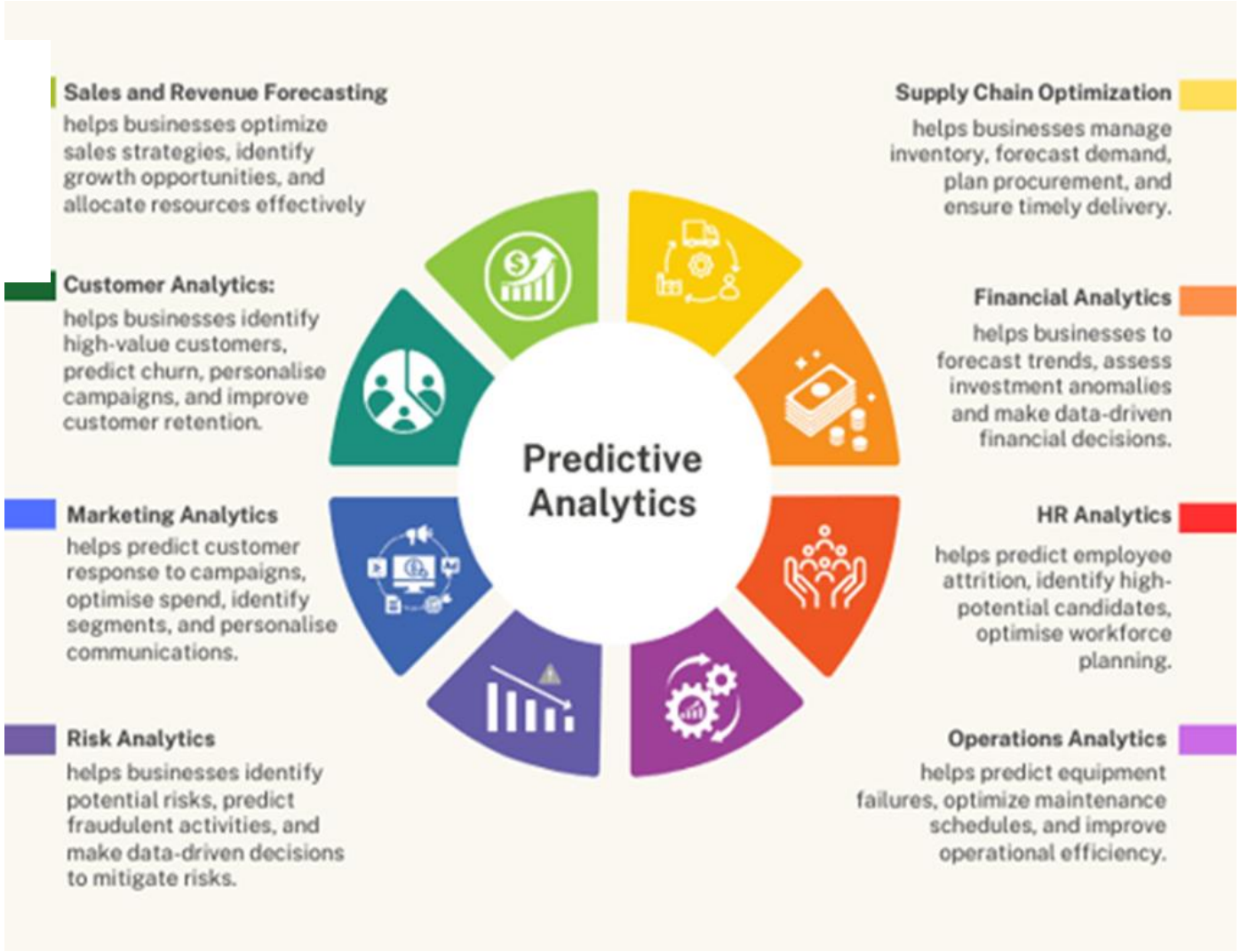
Predict the future by examining historical data, detecting patterns or relationships in these data, and then extrapolating these relationships forward in time.

Application of Predictive Analysis

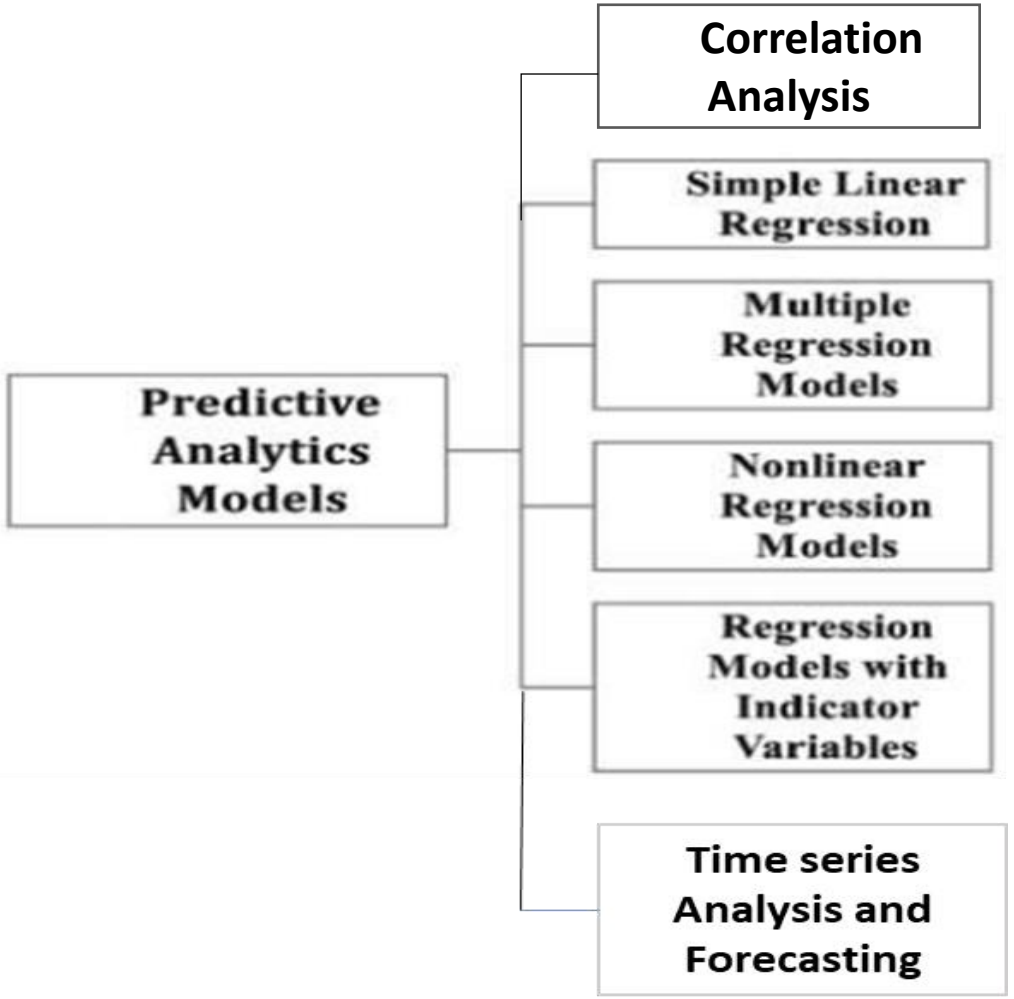
Where Predictive Analytics Is Having the Biggest Impact

by Jacob LaRiviere, Preston McAfee, Justin Rao, Vijay K. Narayanan and Walter Sun

- Predicting demand
- Improved pricing
- Predictive maintenance
- Diagnosis and treatment of diseases
- Distributed electricity generation to localized electricity demand



Predictive Analysis Models

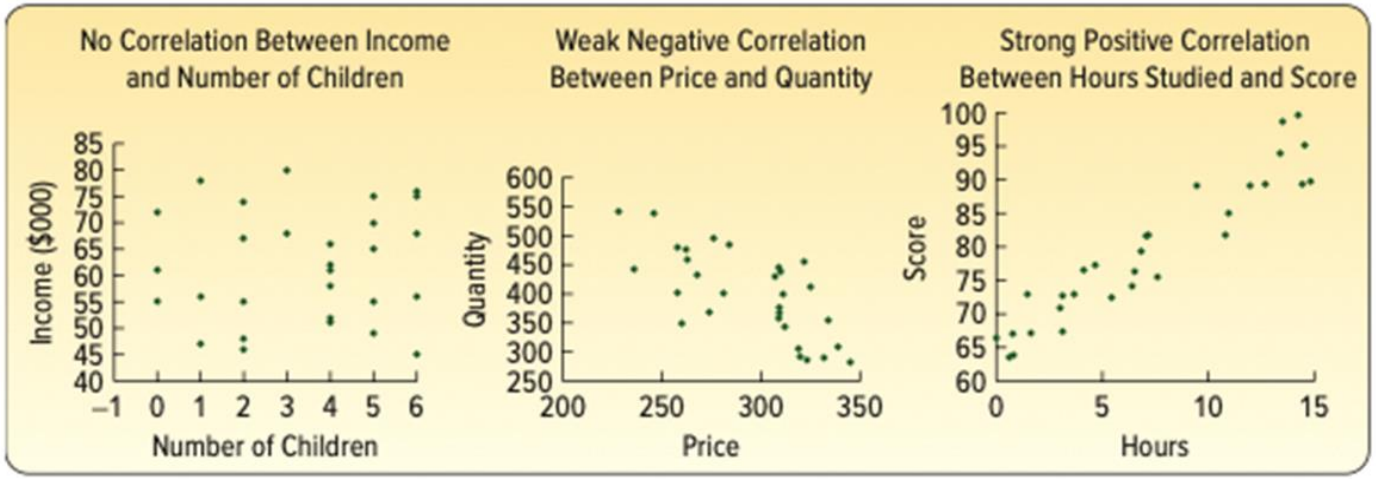


CORRELATION ANALYSIS

A group of techniques to measure the relationship between two variables.

Identified as r or ρ

- It ranges from −1 to 1.
- Near 0 indicates little linear association.
- Near 1 indicates a direct/positive linear association.
- Near −1 indicates an inverse/negative linear association.



Correlation Example

North American Copier Sales sells copiers to businesses of all sizes throughout the United States and Canada. The new national sales manager is preparing for an upcoming sales meeting and would like to impress upon the sales representatives the importance of making an extra sales call each day. She takes a random sample of 15 sales representatives and gathers information on the number of sales calls made last month and the number of copiers sold.

Sales Representative	Sales Calls	Copiers Sold
Brian Virost	96	41
Carlos Ramirez	40	41
Carol Saia	104	51
Greg Fish	128	60
Jeff Hall	164	61
Mark Reynolds	76	29
Meryl Rumsey	72	39
Mike Kiel	80	50
Ray Snarky	36	28
Rich Niles	84	43
Ron Broderick	180	70
Sal Spina	132	56
Soni Jones	120	45
Susan Welch	44	31
Tom Keller	84	30

Spurious Correlation

- If there is a strong association, we might be tempted assume a change in one variable causes a change in another variable.
- Spurious correlation: Strong correlation between variables that are not logically related to each other.

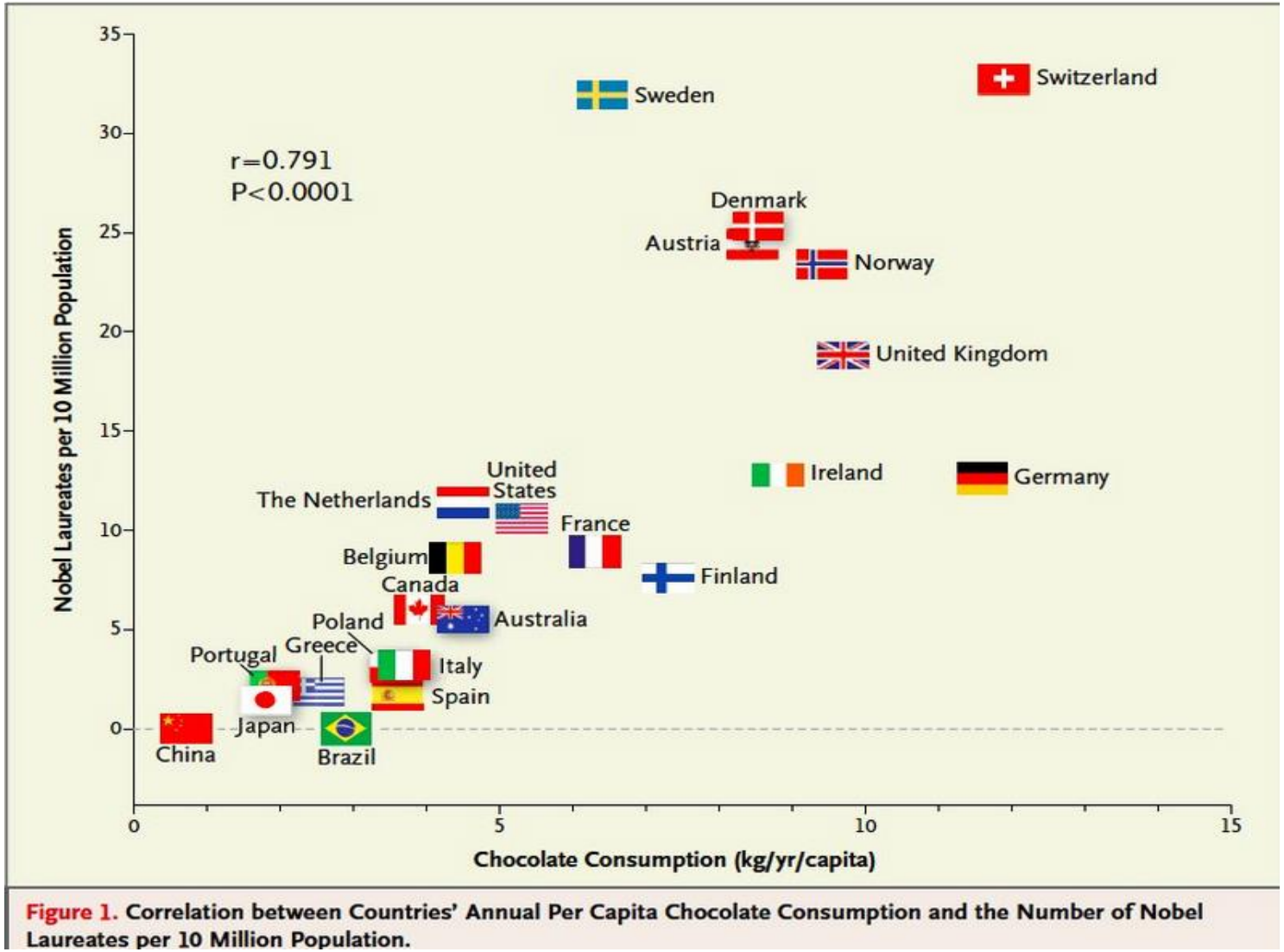
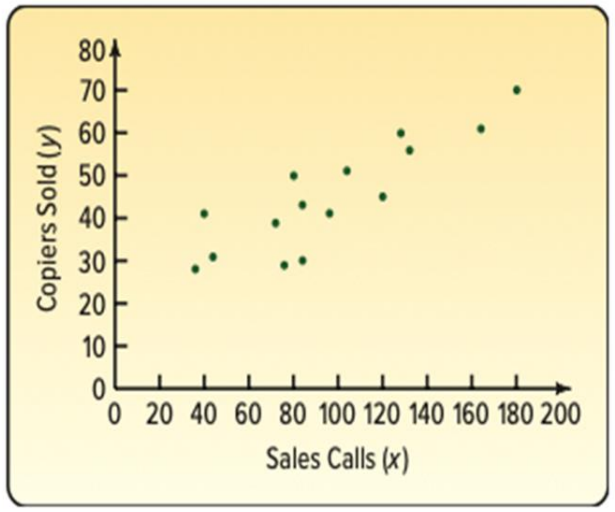


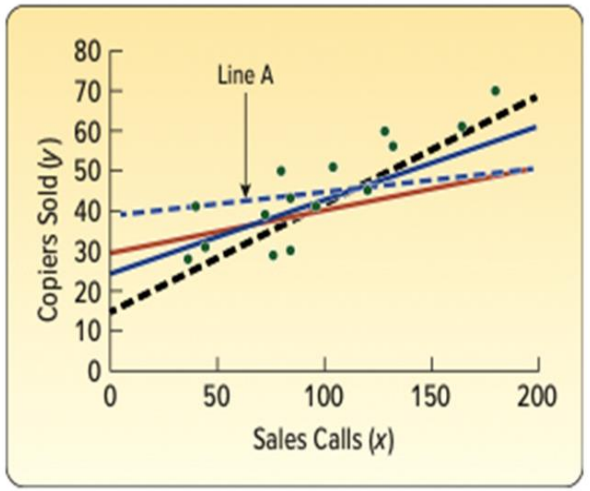
Figure 1. Correlation between Countries' Annual Per Capita Chocolate Consumption and the Number of Nobel Laureates per 10 Million Population.

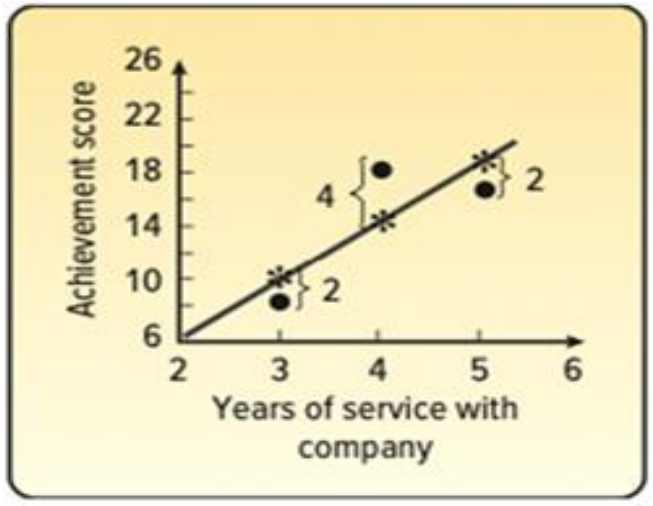
Regression Analysis

- It is another method to examine relationship between two variables.
- Provides more information than correlation.
- Allows estimating the value of the dependent variable (Y) for given a value of the independent variable (X).
- **Regression Equation** expresses the linear relationship between two variables.
- Use the data to position a line that best represents the relationship between X and Y.
- The ”best fitting” line is obtained by the least squares principle.

Best Fit Line (least square principle)







Regression Equation

General form of the linear regression equation:

$$\hat{y} = a + bx$$

$\hat{y}$ is the estimated value of y for a selected value of x.

- *a* is the constant or intercept.
- *b* is the slope of the fitted line.
- *x* is the value of the independent variable.

The values of *a* and *b* are given by:

$$b = r\left(\frac{s_y}{s_x}\right).$$

$$a = \overline{y} - b\overline{x}.$$

Example

Sales call vs copies sold

- What is the least squares equation?
- If a salesperson makes 100 calls, how many copiers do they expect to sale?

Simple Linear Regression Output

	A	B	C	D	E	F	G	H	I	J
1	**Sales Representive**	**Sales calls (x)**	**Copiers Sold (y)**		SUMMARY OUTPUT					
2	Brian Virost	96	41							
3	Carlos Ramirez	40	41		*Regression Statistics*					
4	Carol Saia	104	51		Multiple R	0.865				
5	Greg Fish	128	60		R Square	0.748				
6	Jeff Hall	164	61		Adjusted R Square	0.728				
7	Mark Reynolds	76	29		Standard Error	6.720				
8	Meryl Rumsey	72	39		Observations	15				
9	Mike Kiel	80	50							
10	Ray Snarsky	36	28		ANOVA					
11	Rich Niles	84	43			*df*	*SS*	*MS*	*F*	*Significance F*
12	Ron Broderick	180	70		Regression	1	1738.89	1738.89	38.5031	3.19277E-05
13	Sal Spina	132	56		Residual	13	587.11	45.1623		
14	Sani Jones	120	45		Total	14	2326			
15	Susan Welch	44	31							
16	Tom Keller	84	30			*Coefficients*	*Standard Error*	*t Stat*	*P-value*	
17					Intercept	19.9800	4.389675533	4.55159	0.00054	
18					Sales calls (x)	0.2606	0.042001817	6.20509	3.2E-05	

Multiple Linear Regression

Salsberry Realty sells homes along the East Coast of the United States. One of the questions most frequently asked by prospective buyers is: If we purchase this home, how much can we expect to pay to heat it during the winter? The analyst team at Salsberry has been asked to develop some guidelines regarding heating costs for single-family homes. Three variables are thought to relate to the heating costs:

- the mean daily outside temperature,
- the number of inches of insulation in the attic (roof), and
- the age in years of the furnace.

To investigate, Salsberry's research department selected a random sample of 20 recently sold homes. It determined the cost to heat each home last January, as well as the January outside temperature in the region, the number of inches of insulation in the attic, and the age of the furnace.

Home	Heating Cost ($)	Mean Outside Temperature (°F)	Attic Insulation (inches)	Age of Furnace (years)
1	$250	35	3	6
2	360	29	4	10
3	165	36	7	3
4	43	60	6	9
5	92	65	5	6
6	200	30	5	5
7	355	10	6	7
8	290	7	10	10
9	230	21	9	11
10	120	55	2	5
11	73	54	12	4
12	205	48	5	1
13	400	20	5	15
14	320	39	4	7
15	72	60	8	6
16	272	20	5	8
17	94	58	7	3
18	190	40	8	11
19	235	27	9	8
20	139	30	7	5

Non-linear Regression

- Regression analysis and the correlation coefficient require the relationship to be linear.
- But what if data is not linear?
  - Rescale one or both of the variables so the new relationship is linear.
  - Common transformations include.
  - The base 10 log, log(y).
  - The square root.
  - The reciprocal.
  - Square one or both variables.

Regression with indicator variables

- Indicator variables, also known as dummy variables, are binary variables (0 or 1) used to represent categorical variables in regression models. They effectively encode qualitative information into a format that quantitative models can understand.
- How to use them in regression models:
  - Create indicator variables: For a categorical variable with m categories, you'll create m-1 indicator variables.
  - Choose a reference category: One category is chosen as the reference (or baseline) and is not represented by an indicator variable.
  - Include in the model: The indicator variables are included as independent variables in your regression equation, alongside any continuous variables.
  - Interpret coefficients: The coefficients of the indicator variables represent the difference in the dependent variable's expected value compared to the reference category

Example of regression with indicators variable

- Suppose you are analyzing salary and want to include gender (male/female) as a factor. You would create one indicator variable, say "is_female" (1 if female, 0 if male).
- In your regression model: Salary = $\beta_0 + \beta_1 * is\_female + other\_variables$

- β_0 represents the average salary for males (the reference category).

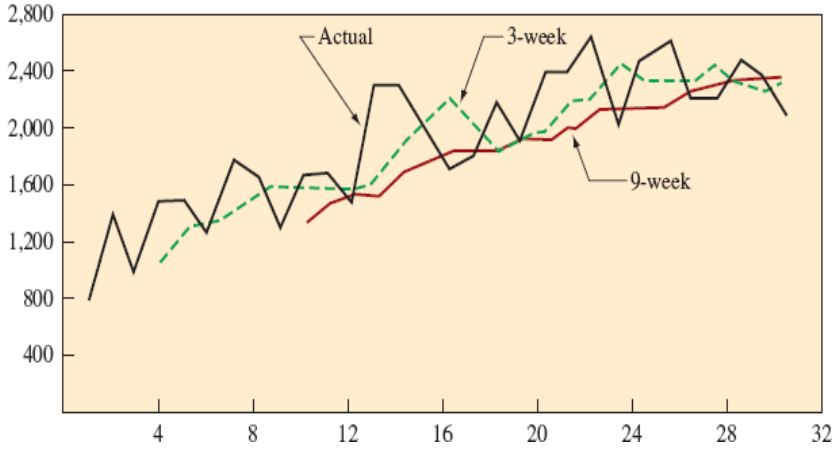
- β_1 represents the difference in average salary between females and males. If β_1 is negative, females on average earn less than males.

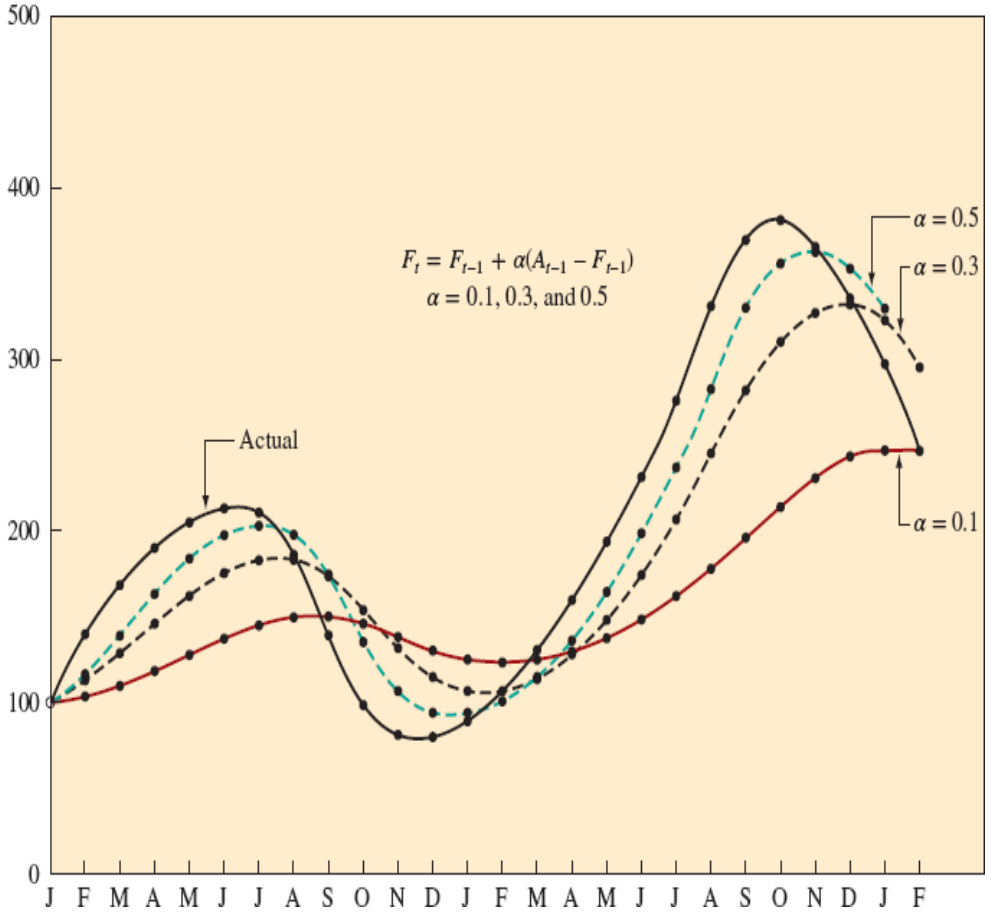
Time Series and Forecasting models

- Simple moving average
- Weighted moving average
- Simple exponential smoothing
- Exponential smoothing including trend
- Time Series Regression
- Causal Regression
- Trend with seasonality index

Time series forecasting example

Week	Demand	3-Week	9-Week	Week	Demand	3-Week	9-Week
1	800			16	1,700	2,200	1,811
2	1,400			17	1,800	2,000	1,800
3	1,000			18	2,200	1,833	1,811
4	1,500	1,067		19	1,900	1,900	1,911
5	1,500	1,300		20	2,400	1,967	1,933
6	1,300	1,333		21	2,400	2,167	2,011
7	1,800	1,433		22	2,600	2,233	2,111
8	1,700	1,533		23	2,000	2,467	2,144
9	1,300	1,600		24	2,500	2,333	2,111
10	1,700	1,600	1,367	25	2,600	2,367	2,167
11	1,700	1,567	1,467	26	2,200	2,367	2,267
12	1,500	1,567	1,500	27	2,200	2,433	2,311
13	2,300	1,633	1,556	28	2,500	2,333	2,311
14	2,300	1,833	1,644	29	2,400	2,300	2,378
15	2,000	2,033	1,733	30	2,100	2,367	2,378



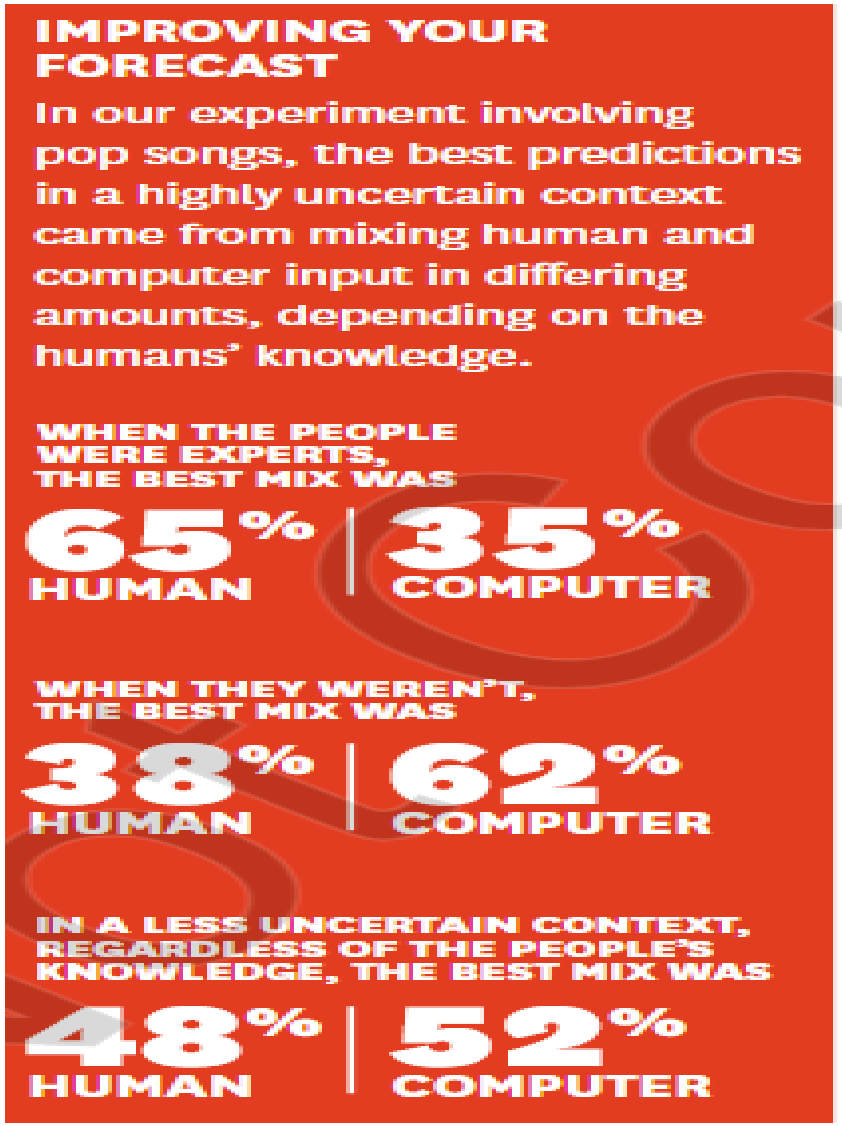


$$F_t = F_{t-1} + \alpha(A_{t-1} - F_{t-1})$$

Accuracy of data driven prediction

Depends on:

- Variables used (relevance)
- Quantity of data (adequacy)
- Quality of Data (accuracy and recentness and relevance)



Thank you