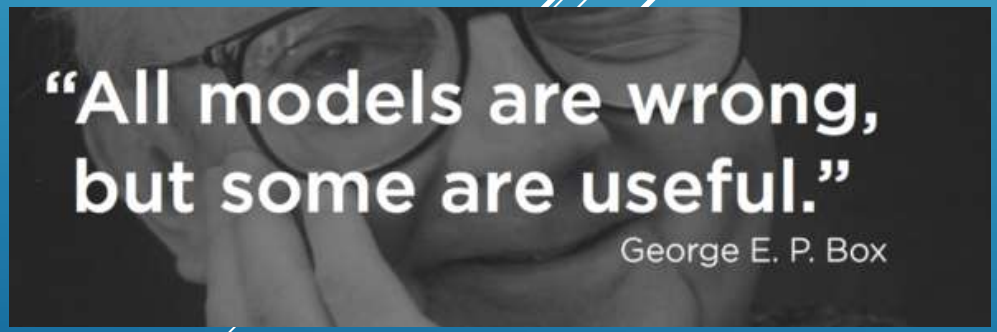
MODELS

Statistical view

Next week – Raman is OOO – no class
We will pick up with
- bias/variance
- hold-out

before moving on to Logistic



1. generate a list (x1) containing 1 through 50
2. randomly select 13 numbers of that list and assign to a second list (x2)
3. use lapply to square the 13 numbers in the second list and assign it to 3rd list
4. create a data.frame with these two lists x1, and x2
5. create a matrix with this data.frame, m1
6. transpose this matrix, m2
7. multiply m1 and m2 and create a m3
8. create a data.frame with m3
9. create a random list of 13 characters from a through z
10. make a word containing anywhere from 3 to 7 characters from these 13 characters
11. make a list of 23 such words

Practice – 01

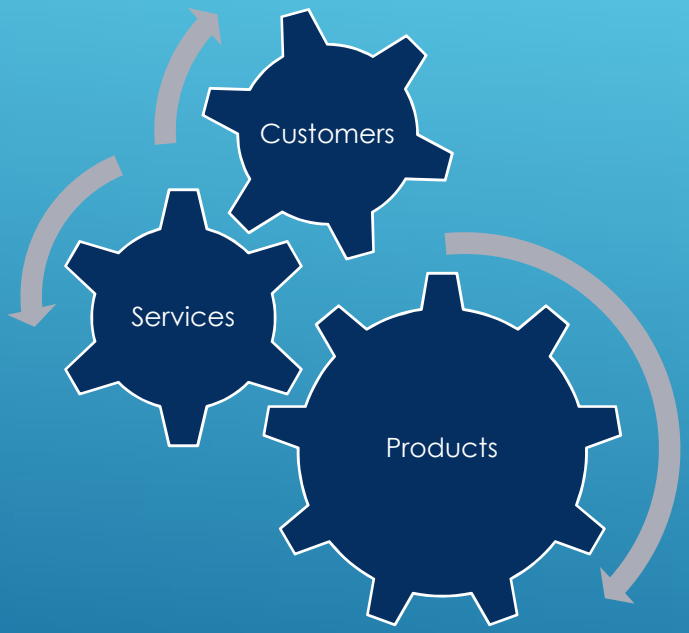
```
require(ggplot2)
require(vcd)

ggplot(data=Arthritis, aes(x=Treatment))+geom_bar(aes(fill=Improved))


table(Arthritis[Arthritis$Treatment=="Treated",]$Improved)
table(Arthritis[Arthritis$Treatment=="Placebo",]$Improved)

ggplot(data=Arthritis, aes(x=Treatment))+geom_bar()
#geoms are objects we draw bar,line,point
#aes() maps variables to aesthetic attributes (color,shape,x-axis,y-axis) of
#the geoms, things that we draw, such as lines,points,bars

ggplot(data=Arthritis,aes(x=Sex))+geom_bar()+coord_flip() #horizontal, aes (x=Sex)
```

```
ggplot(data=Arthritis, aes(x=Treatment))+geom_bar(aes(fill=Improved),position="fill")
ggplot(data=Arthritis, aes(x=Treatment))+geom_bar(color="blue")
ggplot(data=Arthritis, aes(x=Treatment))+geom_bar(color="black",fill='blue')

x<-1:20
y<-sample(1:1000,20,replace=F)

df<-data.frame(x=x,y=y)

ggplot(data=df, aes(x=x,y=y))+geom_point()
ggplot(data=df, aes(x=x,y=y))+geom_point()+geom_smooth()
ggplot(data=df, aes(x=x,y=y))+geom_line()
ggplot(data=df, aes(x=x,y=y))+geom_line()+geom_smooth()
```

OTHER USEFUL R RESOURCES

REFERENCE→ https://r-dir.com/community/forums.html
RSTUDIO
http://web.cs.ucla.edu/~gulzar/rstudio/basic-tutorial.html
SHINY
https://shiny.rstudio.com/tutorial/written-tutorial/lesson2/
Rmarkdown

Real World Data Source
http://www.asdfree.com/
Good book by Garret Grolemund and Hadley Wickham
http://r4ds.had.co.nz/



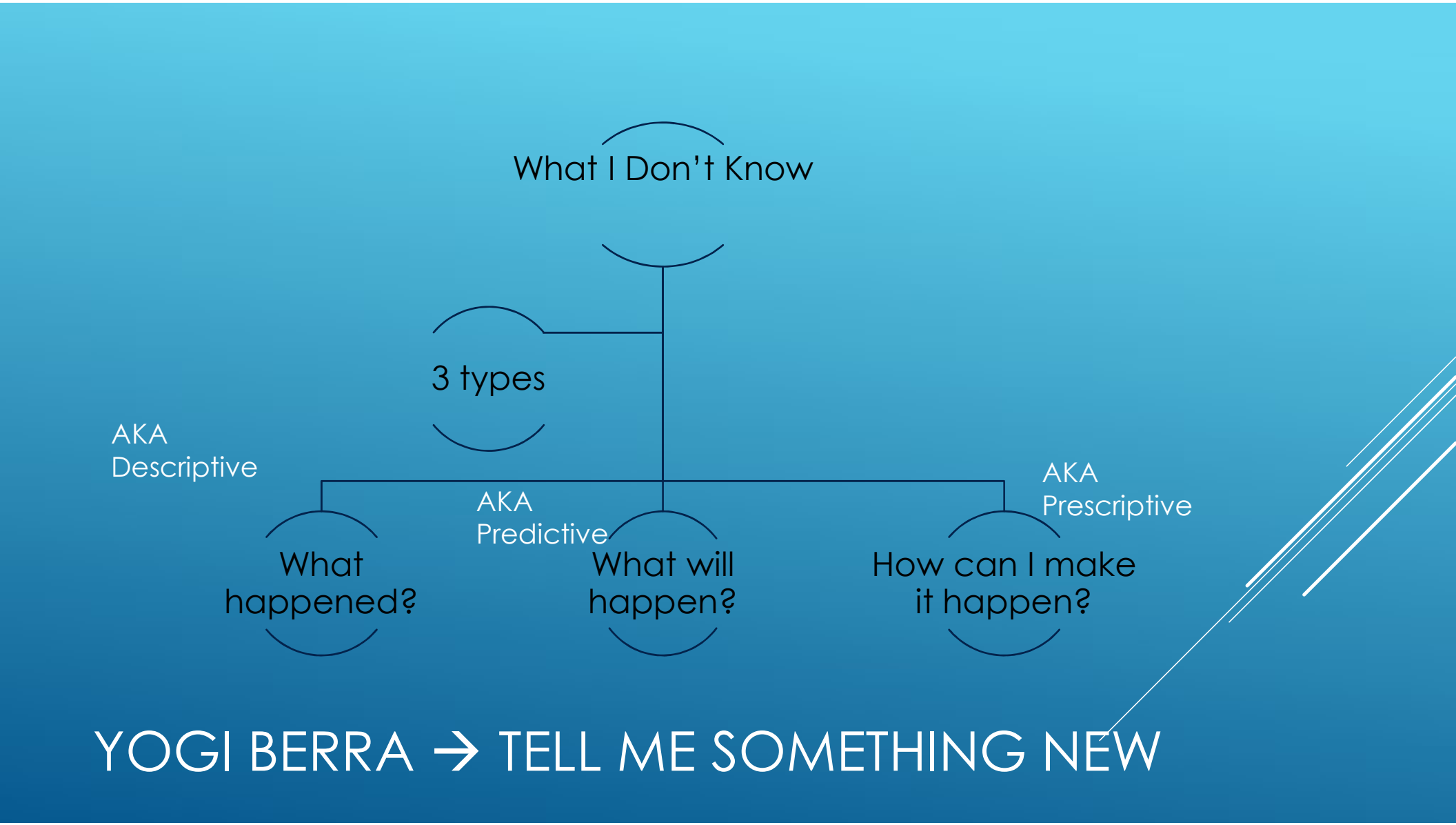
Normal operations generate data.

What can be done with data to improve customer experience, product quality, discovering new markets and invent new products resulting in more revenue.

And, similarly, reduce cost, eliminating waste, optimizing processes.

Profits = Revenue - costs

SESSION 03: TOWARD DATA MINING: DATA



YOGI BERRA → TELL ME SOMETHING NEW

It is not efficient to convey all of the data. A Model is a concise representation of the data.
Error is the amount by which observed Data differs from the Model output
And so we want error to be as low as possible,closer to zero.
Adding additional explanatory terms we can seek to minimize the error
but doing so makes the model more complicated. Complexity vs Accuracy/Quality
Our goal is to come up with the
Simplest Model that "sufficiently" accurately captures the data (error → zero) – aka principle of *parsimony*

MINING DATA

MEETUP LINGO

Data is a set of observations. Each observation is defined by a set of attributes. Attributes are also known as variables. Some variables are not dependent on any other variable. Some variables are determined by other variables.
The purpose of the Model is to determine the dependent given the independent.

Let Y_i be the dependent (output) and X_i be the independent (input) and let f(x) be our model.

$$Y = f(x) + \varepsilon$$

f(x) can be anything before we define the form of the f(x), let us develop the Vocabulary → X are also known as coVariates, predictors, features, regressors in addition to attributes/variables.

Under certain assumptions, given a set of observations we can determine f(x)

MODELS

The simplest form of f(x) results when Y is a linear function of x and the domain of Y is a real (numerical) as shown below.

$Y_1 = \beta_1 x_1 + \beta_0 + \varepsilon$

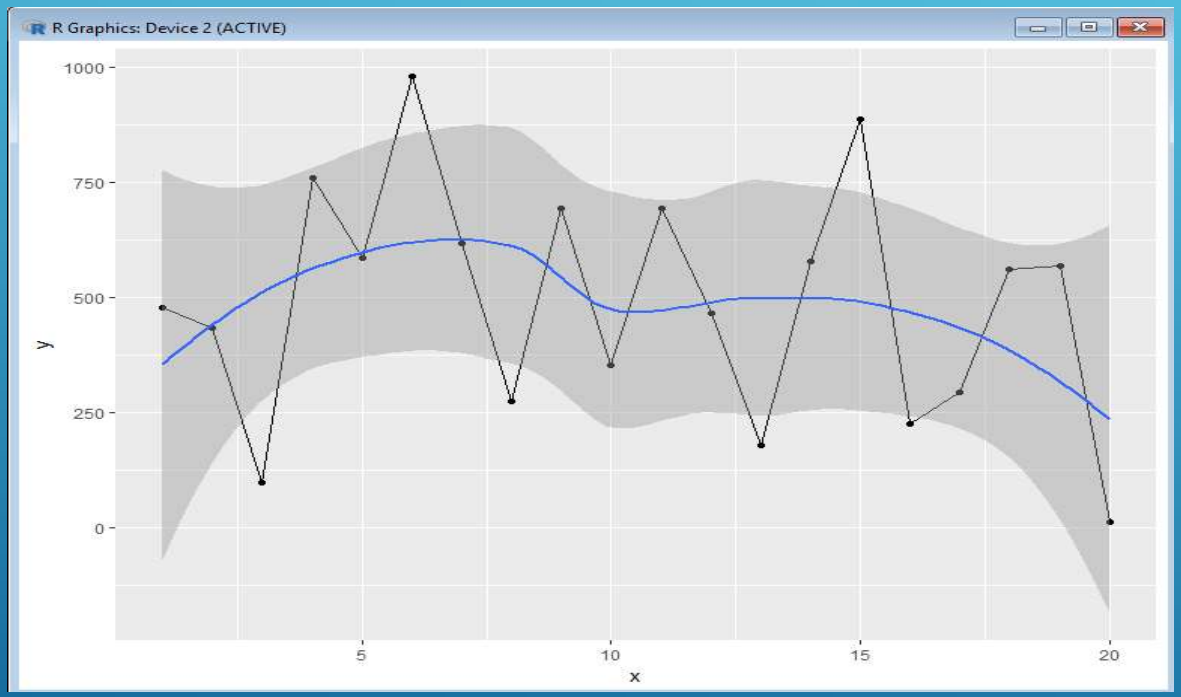
Here ε are random noise and are known as residuals. When MEAN ε is != 0, will manifest as bias. The βs are known as the parameters.

Our objective is to estimate the βs. Thus they are also referred to as estimates.

This is known as linear regression and is valid when the residuals are
-- i.i.d (independent and identical distribution), are unbiased
-- following a standard Normal distribution N(0,1) mean→zero, variance→1
-- absence of heteroskedasticity (constant variance)
-- uncorrelated errors

https://www.quora.com/What-are-the-validation-techniques-of-linear-regression-model
http://people.duke.edu/~rnau/testing.htm#homoscedasticity

```
> ggplot(data=df,aes(x=x,y=y))+geom_point()+geom_line()+geom_smooth()
`geom_smooth()` using method = 'loess'
> df$x
 [1]  1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20
> df$y
 [1] 478 433  98 760 585 982 619 275 694 353 695 465 178 578 888 224 293 562 570  11
>
```



LET US MAKE UP SOME DATA

df$y
[1] 478 433 98 760 585 982 619 275 694 353 695 465 178 578 888 224 293 562 570 11

y<-sample(1:1000,20,replace=F)
> x<-1:20
> df<-data.frame(x,y)

```
> summary(lmdf)

Call:
lm(formula = y ~ x, data = df)

Residuals:
    Min      1Q  Median      3Q     Max
-454.40 -157.59   20.52  166.23  455.74

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  578.542    120.458   4.803 0.000143 ***
x             -8.714     10.056  -0.867 0.397604
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 259.3 on 18 degrees of freedom
Multiple R-squared:  0.04004,   Adjusted R-squared:  -0.01329
F-statistic: 0.7509 on 1 and 18 DF,  p-value: 0.3976
```

Estimates
b_o = 578.542
b_1 = -8.714

R-squared—proportion of variance explained by the model

P-value for the slope b_1 is above 0.05
The Null hypothesis cannot be rejected.

Null for linear regression is that there is no Relationship.

RUNNING THE REGRESSION/

CHECKING THE ASSUMPTIONS: SHAPIRO

```
> resid<-residuals(lmdf)
> stats::shapiro.test(resid)

        Shapiro-Wilk normality test

data:  resid
W = 0.97097, p-value = 0.7752
```

```
> summary(lmdf)

Call:
lm(formula = y ~ x, data = df)

Residuals:
    Min      1Q  Median      3Q     Max
-454.40 -157.59   20.52  166.23  455.74

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  578.542    120.458   4.803 0.000143 ***
x             -8.714     10.056  -0.867 0.397604
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 259.3 on 18 degrees of freedom
Multiple R-squared:  0.04004,   Adjusted R-squared:  -0.01329
F-statistic: 0.7509 on 1 and 18 DF,  p-value: 0.3976
```

p>0.05 => cannot reject the NULL

NULL for Shapiro is that the sample comes from a population that has a normal distribution.

So this sample could be from a population that has a normal distribution. Population does Have a NORMAL distribution. Particular samples may not.

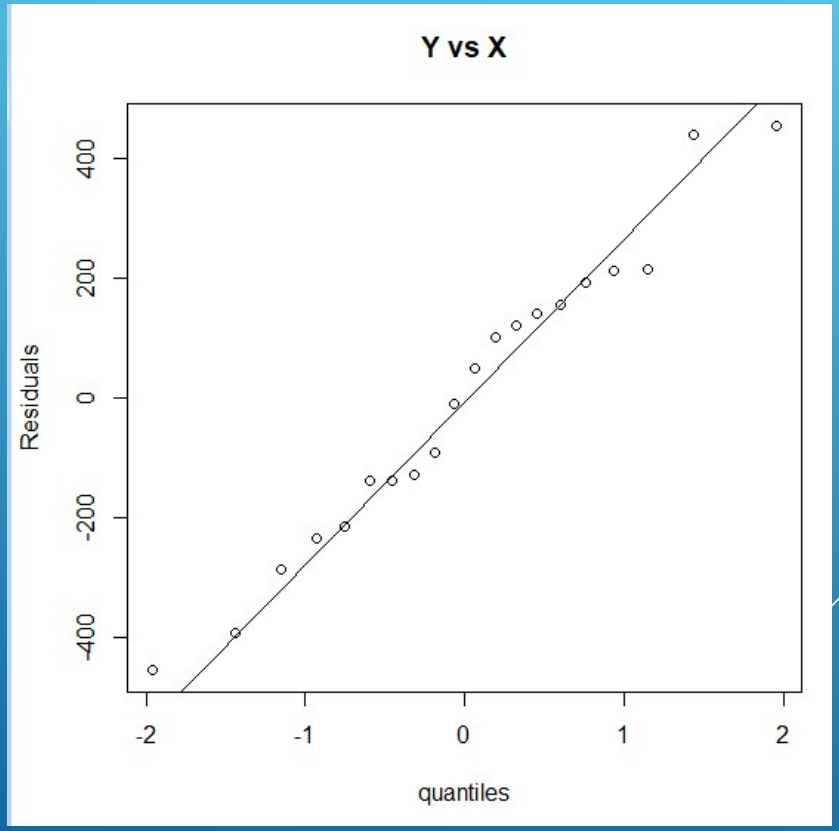
https://stats.stackexchange.com/questions/15696/interpretation-of-shapiro-wilk-test

```
> qqnorm(resid,main="Y vs X", xlab="quantiles",ylab="Residuals",plot.it=TRUE)
> qqline(resid,distribution=qnorm,probs=c(0.05,0.95),qtype=4)
```

The resides must lie on the line…Deviations represent deviations from Normality.

resid<-residuals(lmdf)

Run qqnorm and qqplot
As above…



NORMALITY TEST

BREUSCH PAGAN TEST – QUANTITATIVE TEST FOR CONSTANT VARIANCE

https://www.queryxchange.com/q/20_239060/interpretation-of-breusch-pagan-test-bptest-in-r/
https://stats.stackexchange.com/questions/239060/interpretation-of-breusch-pagan-test-bptest-in-r
https://stats.stackexchange.com/questions/239060/interpretation-of-breusch-pagan-test-bptest-in-r

```
> lmtest::bptest(lmdf)

        studentized Breusch-Pagan test

data:  lmdf
BP = 0.0029947, df = 1, p-value = 0.9564

> lmtest::bgtest(lmdf)

        Breusch-Godfrey test for serial correlation of order up to 1

data:  lmdf
LM test = 0.51569, df = 1, p-value = 0.4727
```

Here since p-value is > 0.05 we cannot reject the NULL.

NULL for bptest is constant variance

So we can infer that the data DOES NOT suffer from heteroskedascity.

TEST FOR AUTO CORRELATON

https://en.wikipedia.org/wiki/Breusch%E2%80%93Godfrey_test
http://www.staff.city.ac.uk/d.hristova/Slides8.pdf

```
> lmtest::bptest(lmdf)

        studentized Breusch-Pagan test

data:  lmdf
BP = 0.0029947, df = 1, p-value = 0.9564

> lmtest::bgtest(lmdf)

        Breusch-Godfrey test for serial correlation of order up to 1

data:  lmdf
LM test = 0.51569, df = 1, p-value = 0.4727
```

P-value is 0.47 > 0.05

Therefore we cannot reject the nULL at a 0.05 sig. level

The nULL for BG is that there is no auto-correlation.