

Statistics for Operational Data: Pitfalls and Good Practices

Audris Mockus

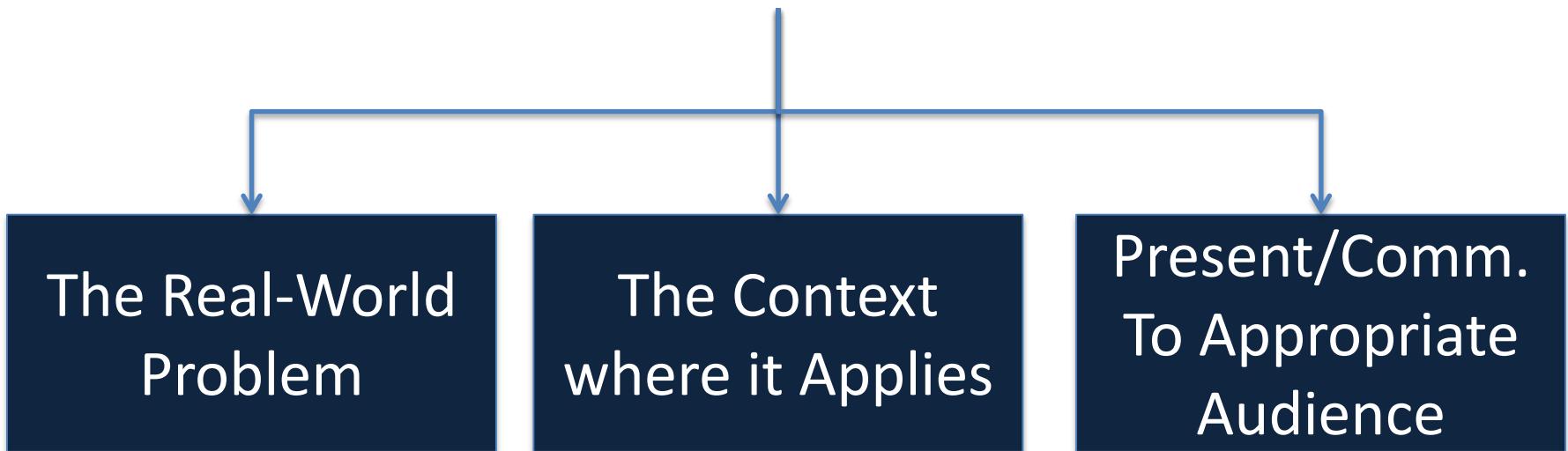
University of Tennessee

**Derived from Tutorials
at FSE'13 and ICSE'14**



Statistical thinking is much more than just p-values

Requires Non-Trivial
Understanding



Take Home

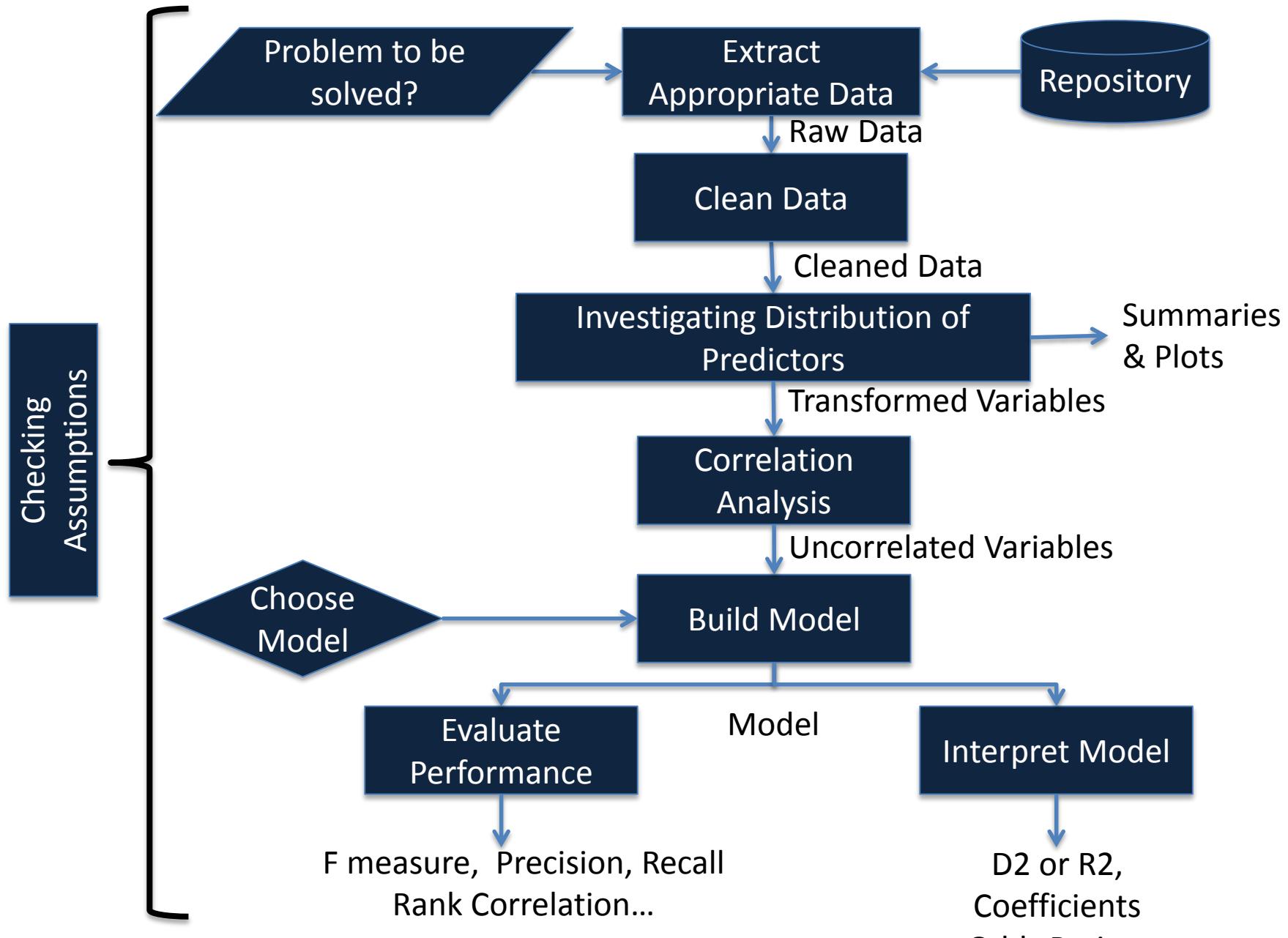


Steps to build a statistical
model for operational data

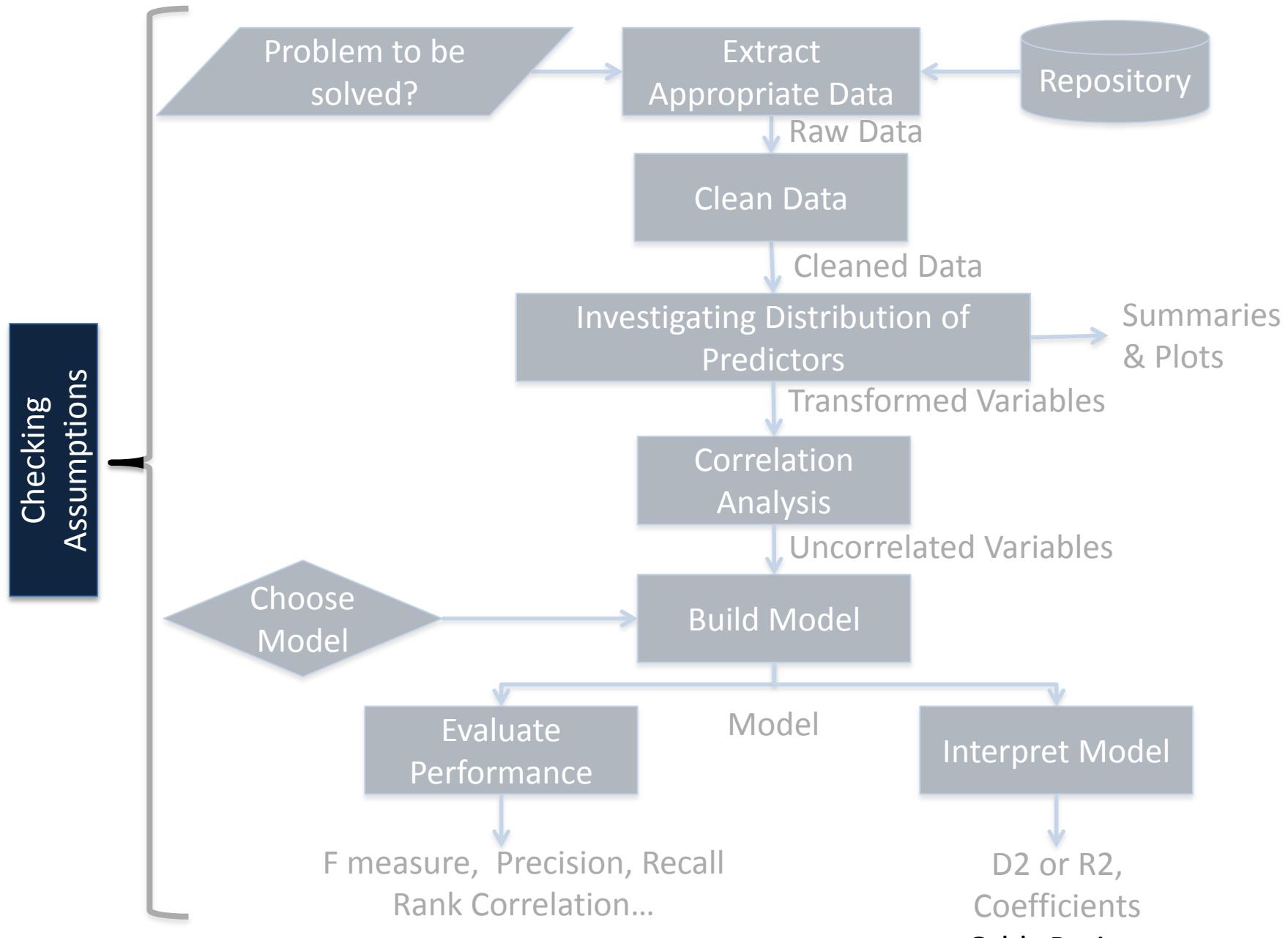


Sample R scripts for an
example problem

Workflow



Workflow



Assumptions

of LOC

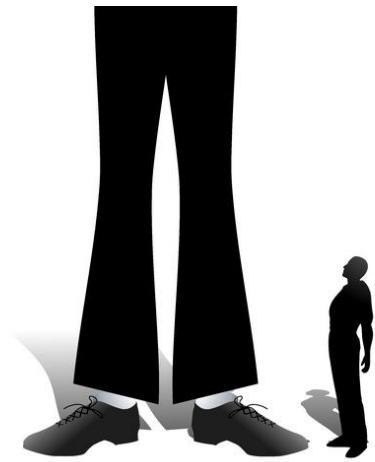


vs.

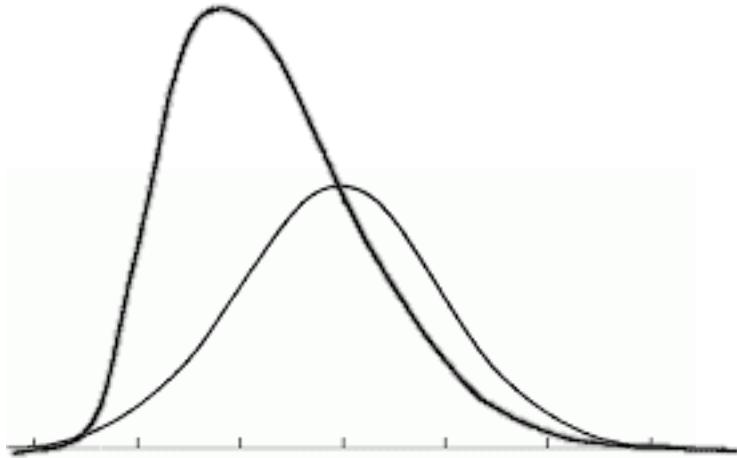
of Coffees



Relevant Data

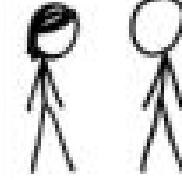


Representative Data

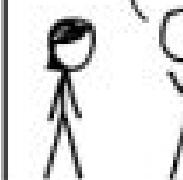


Model Assumptions

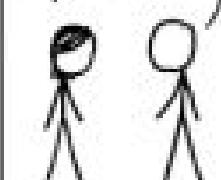
I USED TO THINK,
CORRELATION IMPLIED
CAUSATION.



THEN I TOOK A
STATISTICS CLASS.
NOW I DON'T.

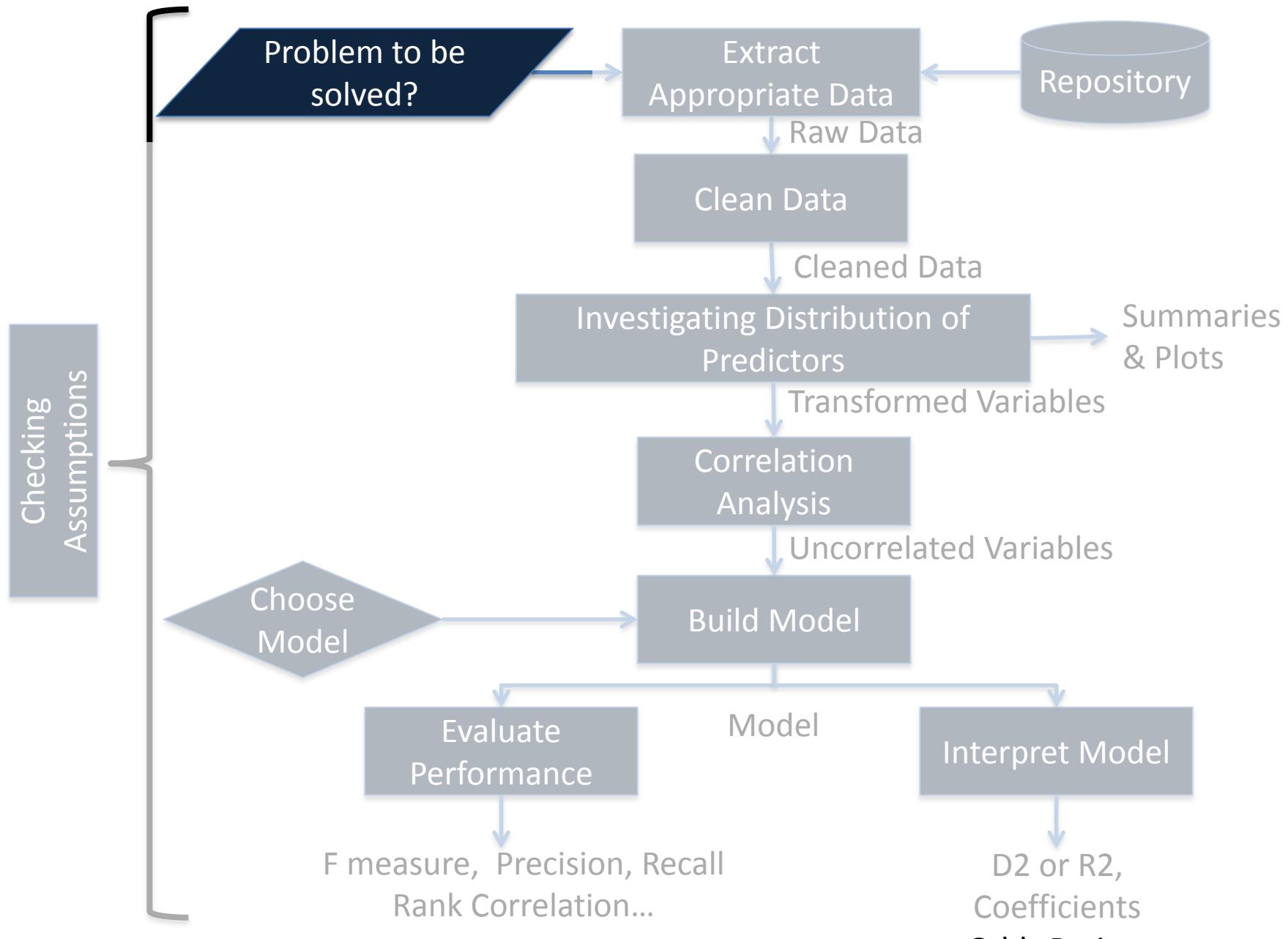


SOUNDS LIKE THE
CLASS HELPED.
WELL, MAYBE.



Correlation vs. Causation

Tutorial Workflow



Problem to be Solved

Test Case Prioritization



Tester

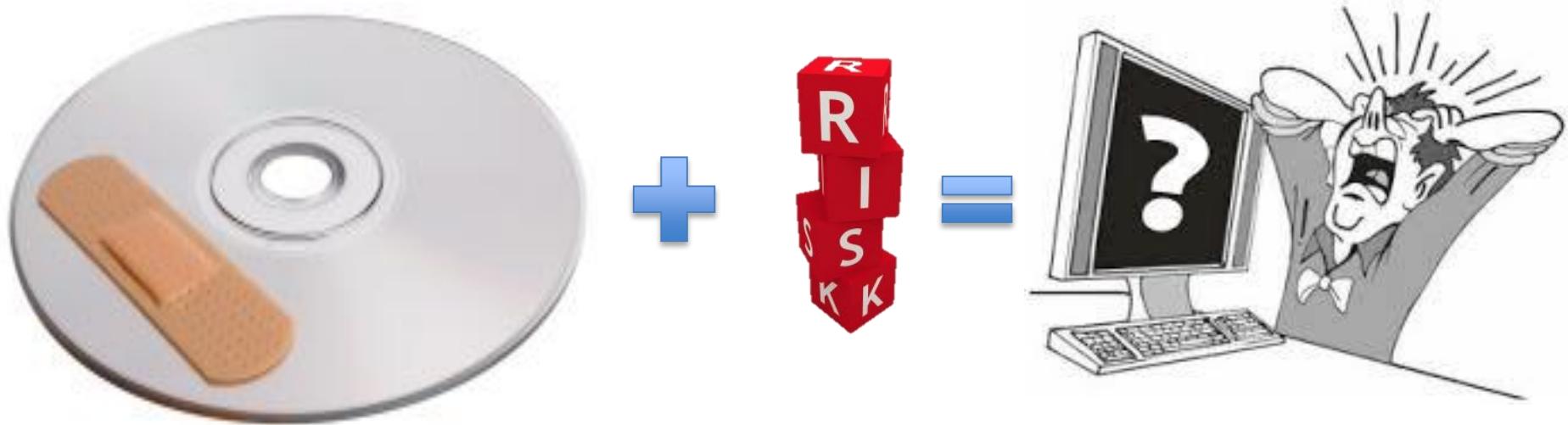
Metrics vs. Bugs



Product Manager

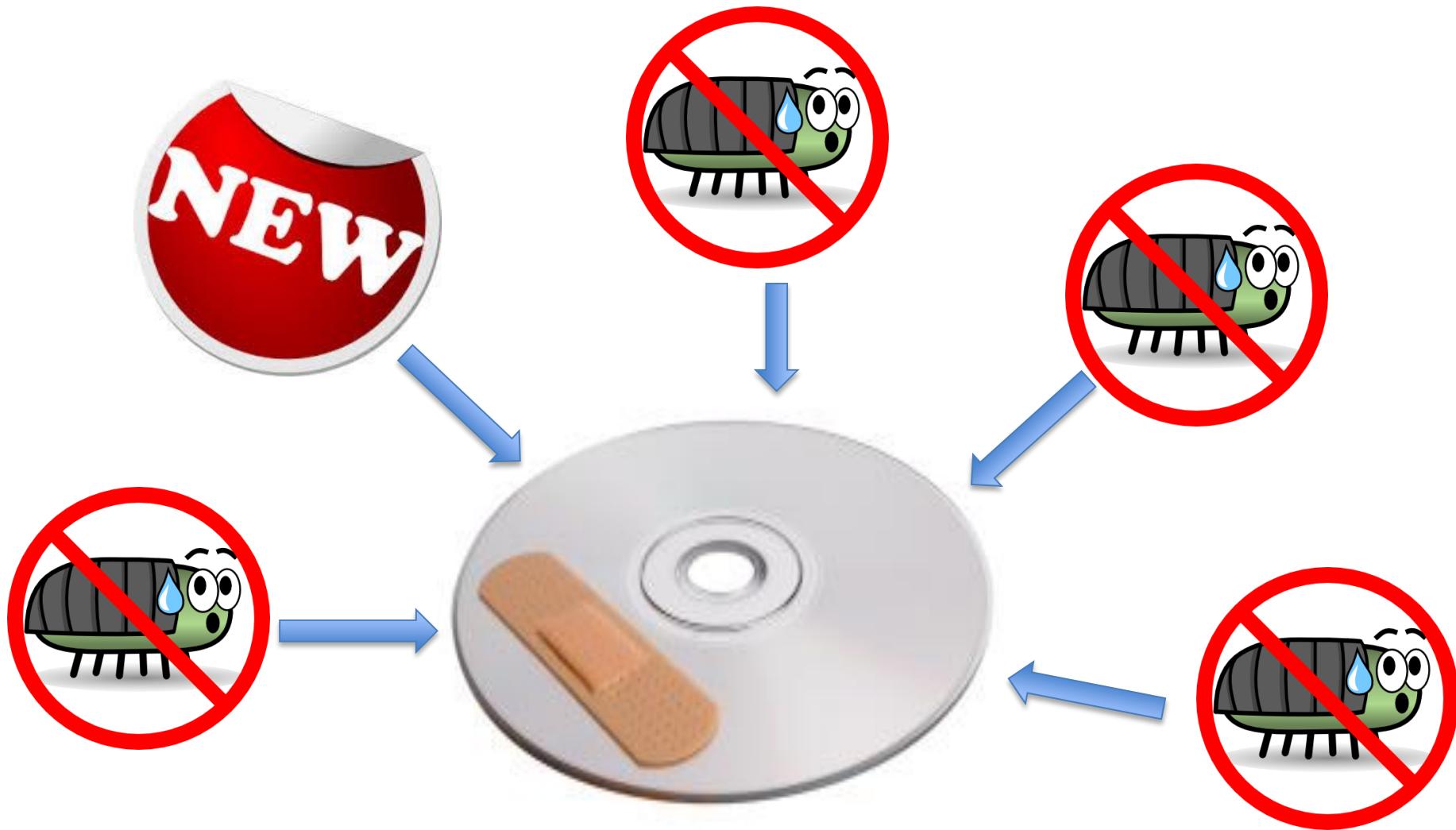
And many more... But Today -

Predicting Risk of Software Updates (Patches/MRs)

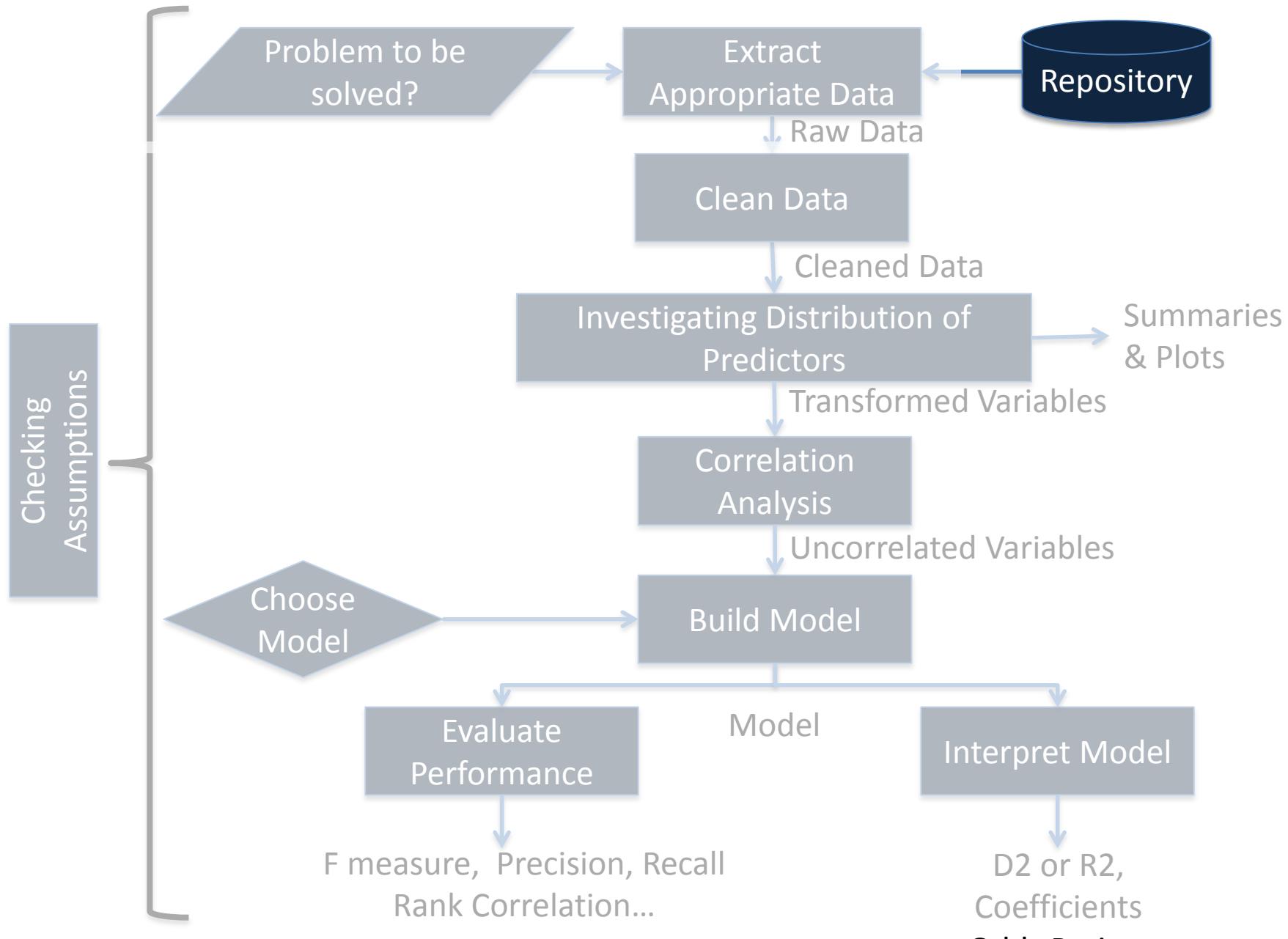


Customer = Developer

Each Patch consists of many Bug Fixes/Enhancements (Modification Requests or MRs)



Workflow



Lots of Rich Data Sources from Software Development

Code Repositories
SCCS

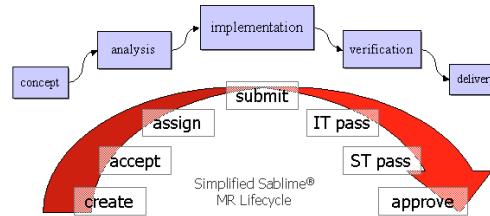


Commit Information:
Date, Change, Developer...

Issue Repository
ECMS/Sablime

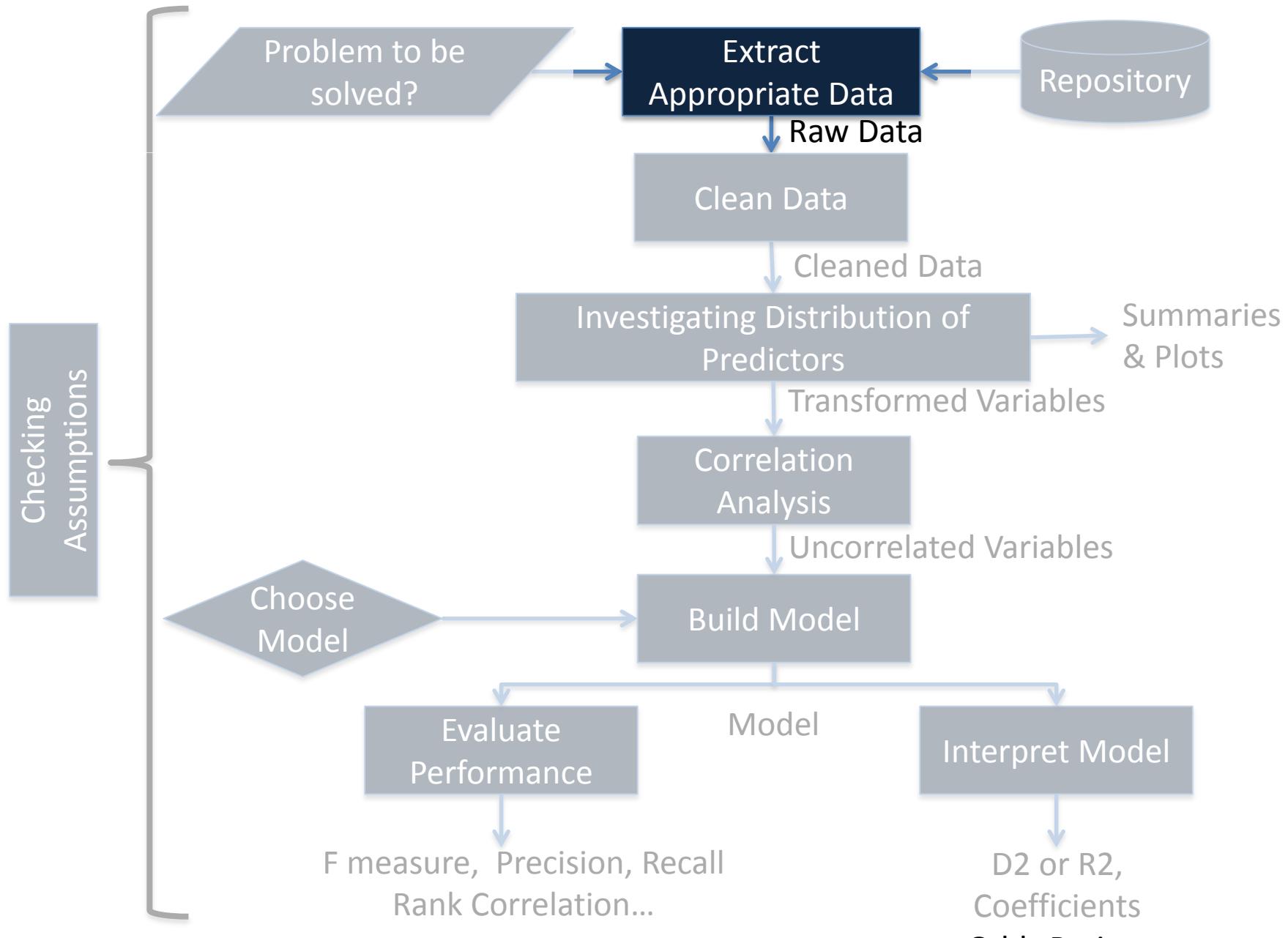


The Sablime® MR takes the idea from concept to customer.



Issue Information:
Date, Status, Assigned to...

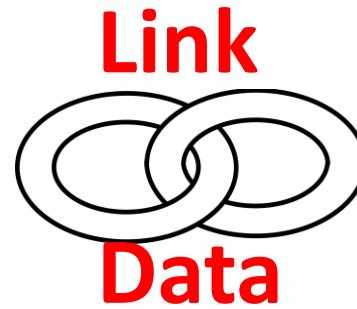
Workflow



Extract Raw Data Required for the Problem!



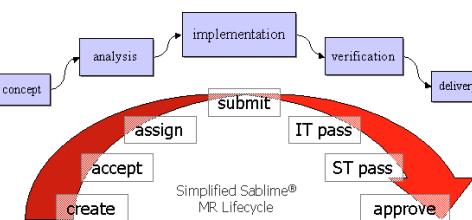
Code Repositories



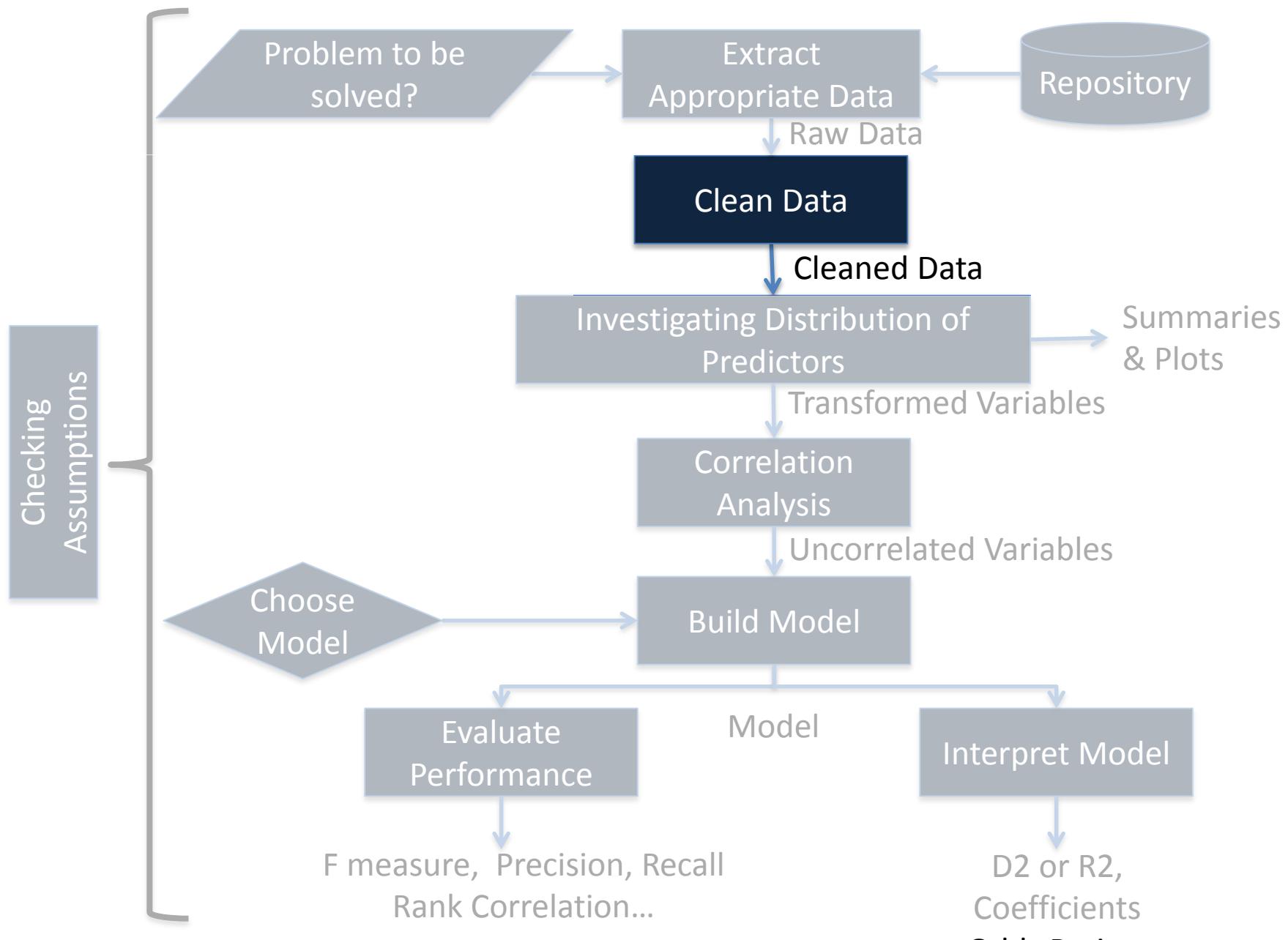
Issue Repository



The Sablime® MR takes the idea from concept to customer.



Workflow



Remove UNRELIABLE Data



Computer Generated



Unpopulated or
Unused Attributes



Irrelevant to Our
Question Data



Administrative
Changes



Other Low Quality
Data

Identifying possible causes of failure (Independent Variable/Predictors)

Formulate a hypothesis for each cause

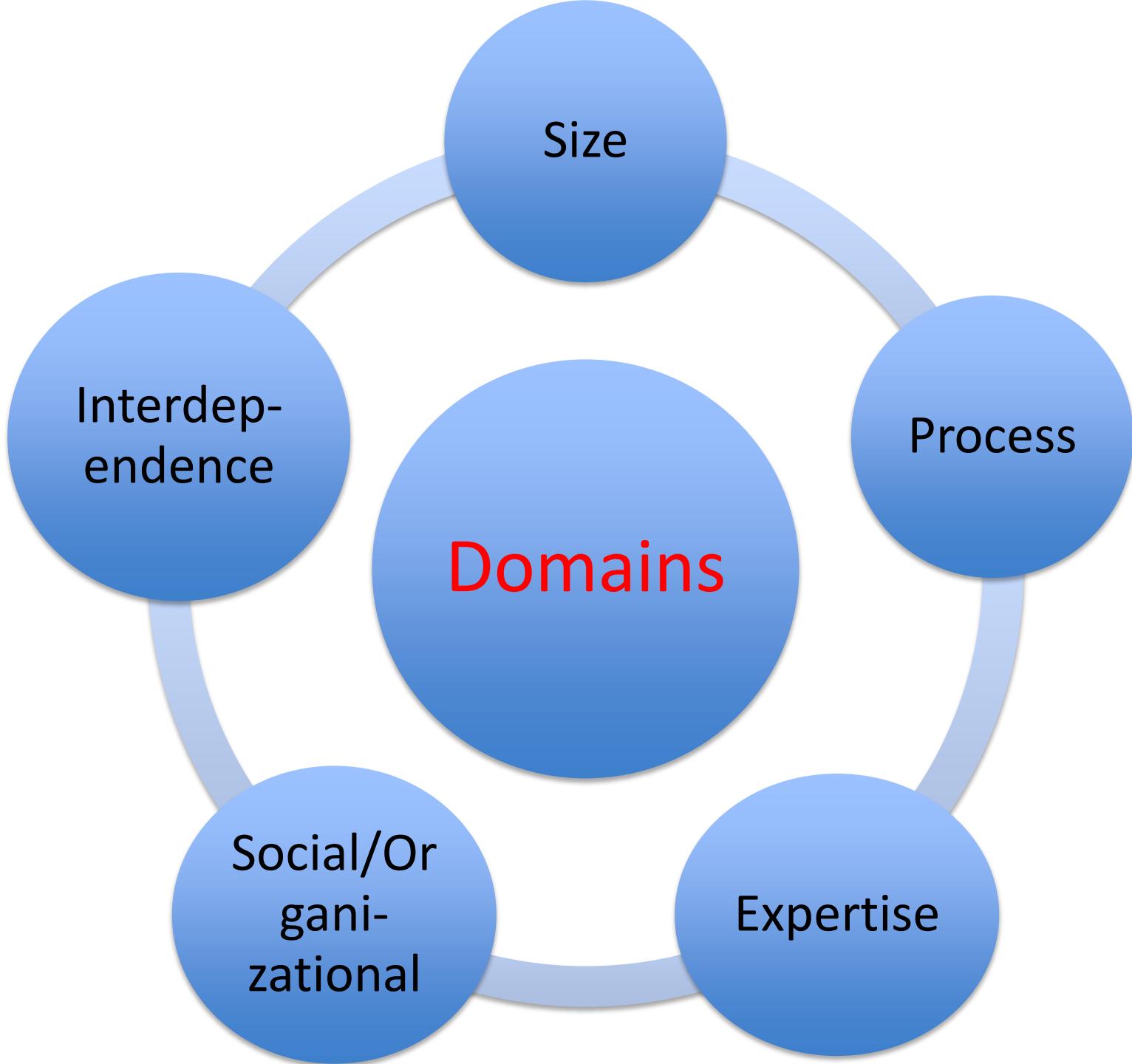
Pick interpretable hypothesis

Abandon ones that rely on poor quality data

Select several from each conceptual domain

Hypotheses in our example

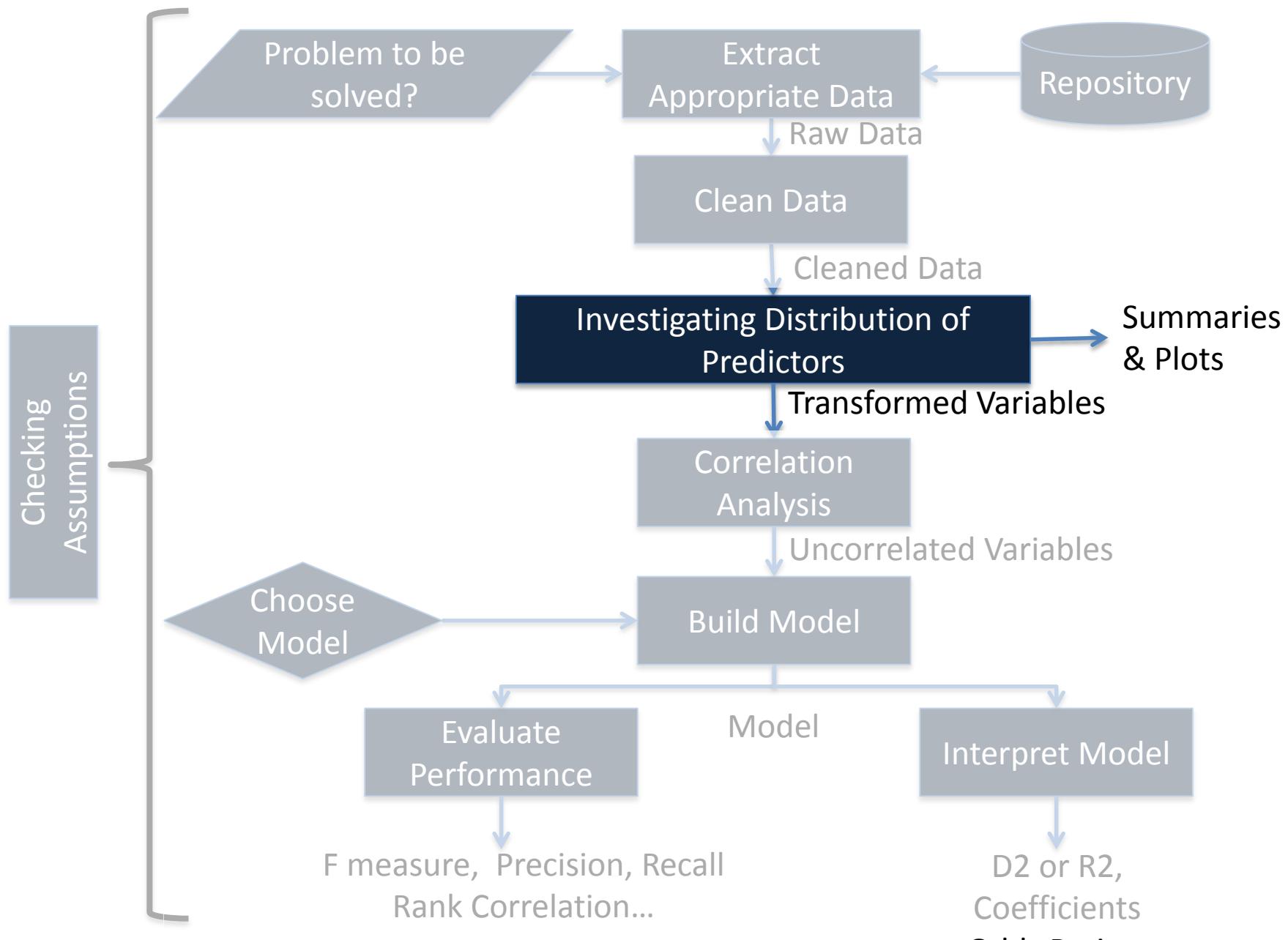
- H1:** More change and more code added is always more risky
- H2:** Diffusion (over code and organization) may cause dependency and coordination errors
- H3:** Long duration may indicate otherwise invisible problems
- H4:** Fixes for customer issues need to be delivered rapidly and may be more risky
- H5:** Individual's experience almost always matters
- H6:** Time of the MR (things tend to change over time)
- H7:** Social/organizational/interdependence measures



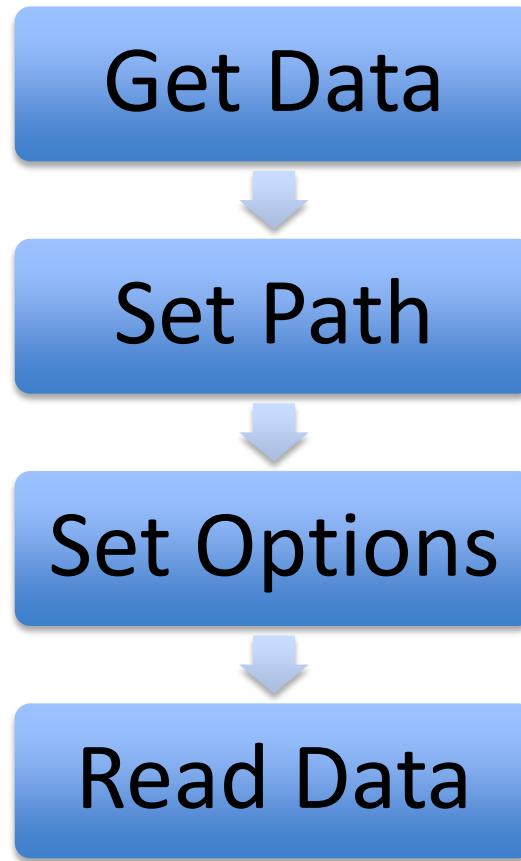
Final Set of Predictors

Type	Name	Description
Response	isBad	Did MR cause patch to fail?
Diffusion	NS	Number of subsystems touched
	NM	Number of modules touched.
	NF	Number of files touched.
	NLOGIN	Number of developers involved.
Size	LA	LOC added.
	LD	LOC deleted.
Diffusion and size	LT	LOC in the files touched by the change.
	NMR	Number of MRs.
Interval	ND	Number of deltas.
Purpose	INT	Time between the last and first delta.
Experience	FIX	Fix of a defect found in the field.
	EXP	Developer experience.
	REXP	Recent developer experience.
	SEXP	Developer experience on a subsystem.

Workflow



Read Data into R



Follow along the R script @

5 Number Summary of Predictors

```
summary (risk);
```

isBad	NS	LA
Min. :0.0000	Min. : 1.00	Min. : 0
1st Qu.:0.0000	1st Qu.: 1.00	1st Qu.: 13
Median :0.0000	Median : 1.00	Median : 47
Mean :0.0211	Mean : 1.65	Mean : 1289
3rd Qu.:0.0000	3rd Qu.: 2.00	3rd Qu.: 211
Max. :1.0000	Max. :18.00	Max. :1845795

Skewed data is not normally distributed!

Log Transform Predictors

```
summary (data[,c(1,2,8)]);
```

	1NS	1LA
Min.	:0.000	Min. : 0.00
1st Qu.	:0.000	1st Qu.: 2.64
Median	:0.000	Median : 3.87
Mean	:0.291	Mean : 4.13
3rd Qu.	:0.693	3rd Qu.: 5.36
Max.	:2.890	Max. :14.43

Nothing has normal distribution in SE

Log(Positive
continuous variables)

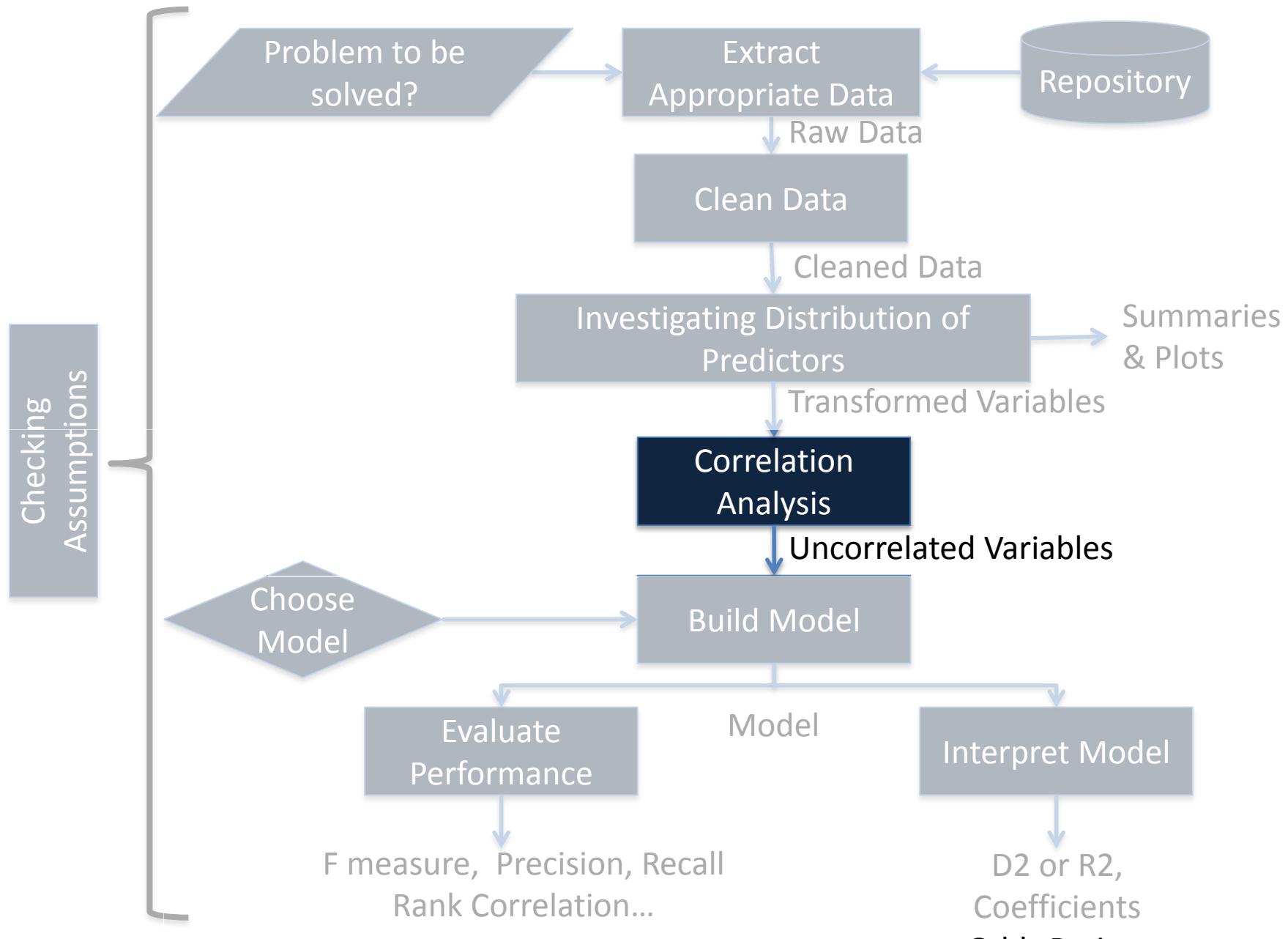
Log(Variabile + 1), if
variable can be 0

Transformation

Group infrequent
categories

Rank transform if still
highly skewed to
remove outliers

Tutorial Workflow



High correlation is typical in SE

```
cor(vars); # OK for normal distribution  
cor(vars,method="spearman"); #OK for any: uses ranks  
  
hiCor(data,.7); # check script for method definition
```

	lNS	lNM	lNF	lNMR	lND	lLA	lld	lLOC	lEXP	lREXP
lNS	1.000	0.7551	0.6501	0.652	0.577	0.512	0.477	0.489	-0.0272	-0.0191
lNM	0.755	1.0000	0.8561	0.743	0.744	0.653	0.591	0.605	-0.0092	0.0063
lNF	0.650	0.8561	1.0000	0.770	0.862	0.757	0.672	0.647	-0.0160	0.0031
lNMR	0.652	0.7432	0.7702	1.000	0.835	0.736	0.696	0.690	-0.0593	-0.0481
lND	0.577	0.7438	0.8615	0.835	1.000	0.868	0.819	0.772	-0.0536	-0.0311
lLA	0.512	0.6532	0.7571	0.736	0.868	1.000	0.776	0.707	-0.0314	-0.0182
lld	0.477	0.5907	0.6716	0.696	0.819	0.776	1.000	0.664	-0.0447	-0.0188
lLOC	0.489	0.6050	0.6472	0.690	0.772	0.707	0.664	1.000	-0.0591	-0.0457
lEXP	-0.027	-0.0092	-0.0160	-0.059	-0.054	-0.031	-0.045	-0.059	1.0000	0.9595
lREXP	-0.019	0.0063	0.0031	-0.048	-0.031	-0.018	-0.019	-0.046	0.9595	1.0000

At least one value > 0.7 in each Column

**But model does not work reliably
with High Correlation**



Select an Orthogonal Subset

Based on Problem and Audience

Use Common Sense to pick a variable



Examine Correlations



Examine Principal Components



Rgress each Predictor



Avoid automatic methods

Use PCA to understand the joint distribution of predictors

How many components are needed to explain for 70% of the variance?

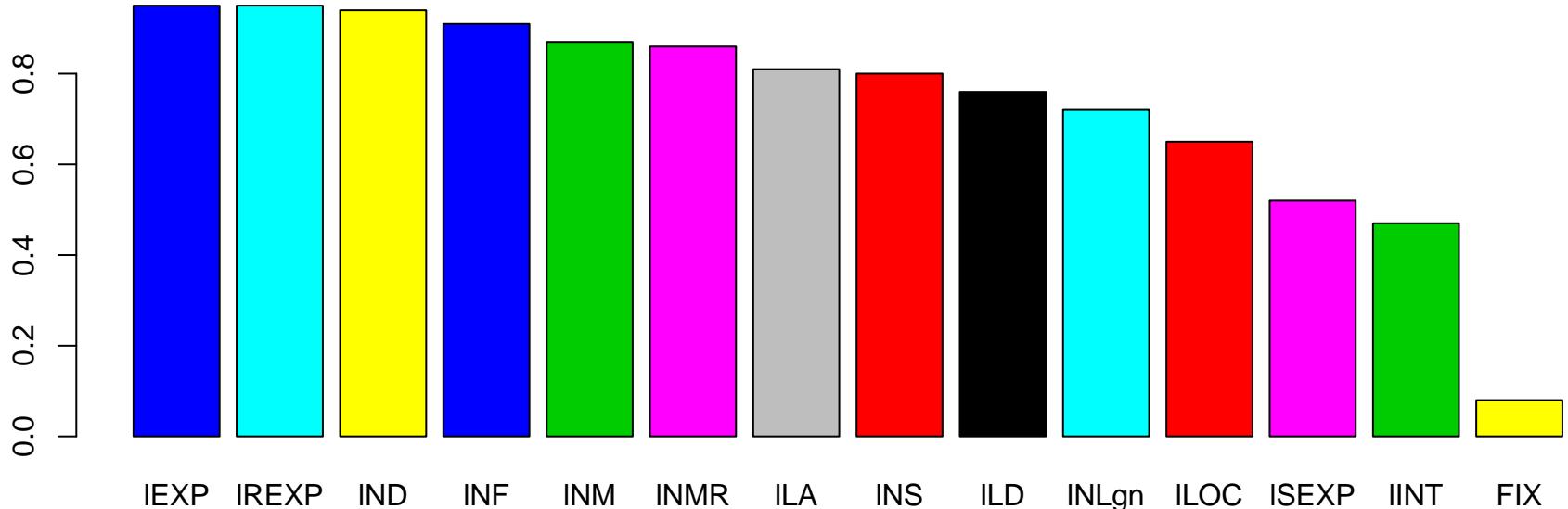
	1NM	1NF	1NLgn	1NMR	1ND	1LA	1LD
PC1	0.32	0.33	0.31	0.34	0.34	0.31	0.3
	1EXP	1REXP	1SEXP				
PC2	-0.65	-0.66	-0.33				



**Diffusion &
Size (PC1)**

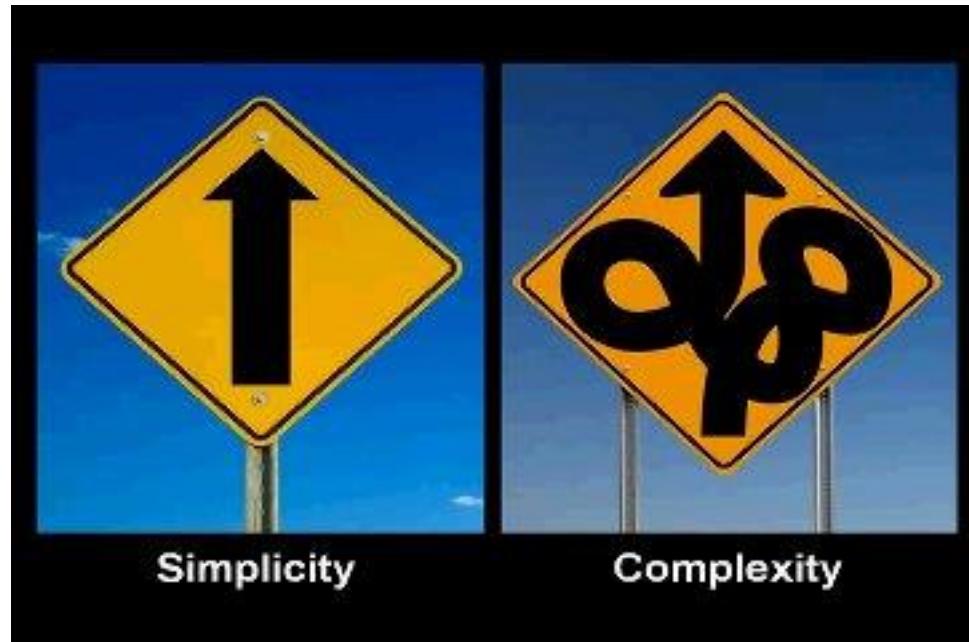
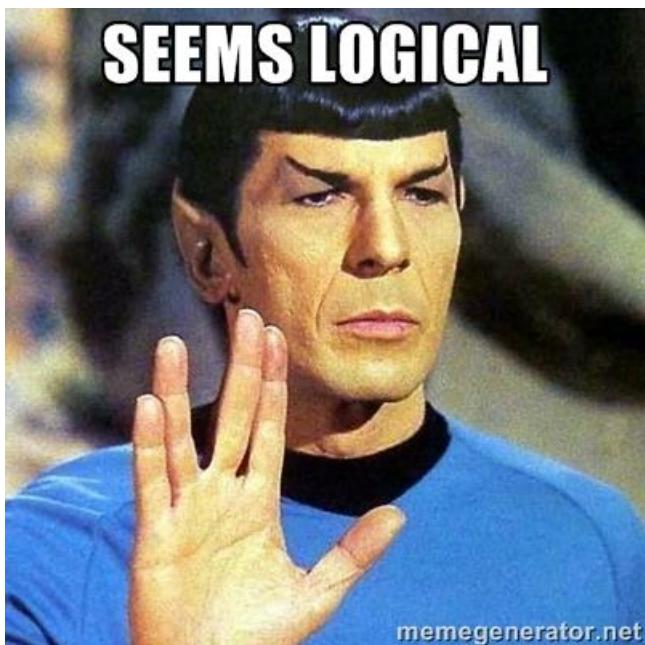
**Developer
Expertise (PC2)**

Eliminate most Predictable Predictor



Predictor with the highest adjR²

How to Pick Predictors



Remaining Variables < 0.85 (adjR²)

FIX, IINT, ISEXP, ILOC, INLgn, ILD, INA, INS



Pick Predictors

INS

IND, IEXP

FIX, IINT

Correlations Analysis

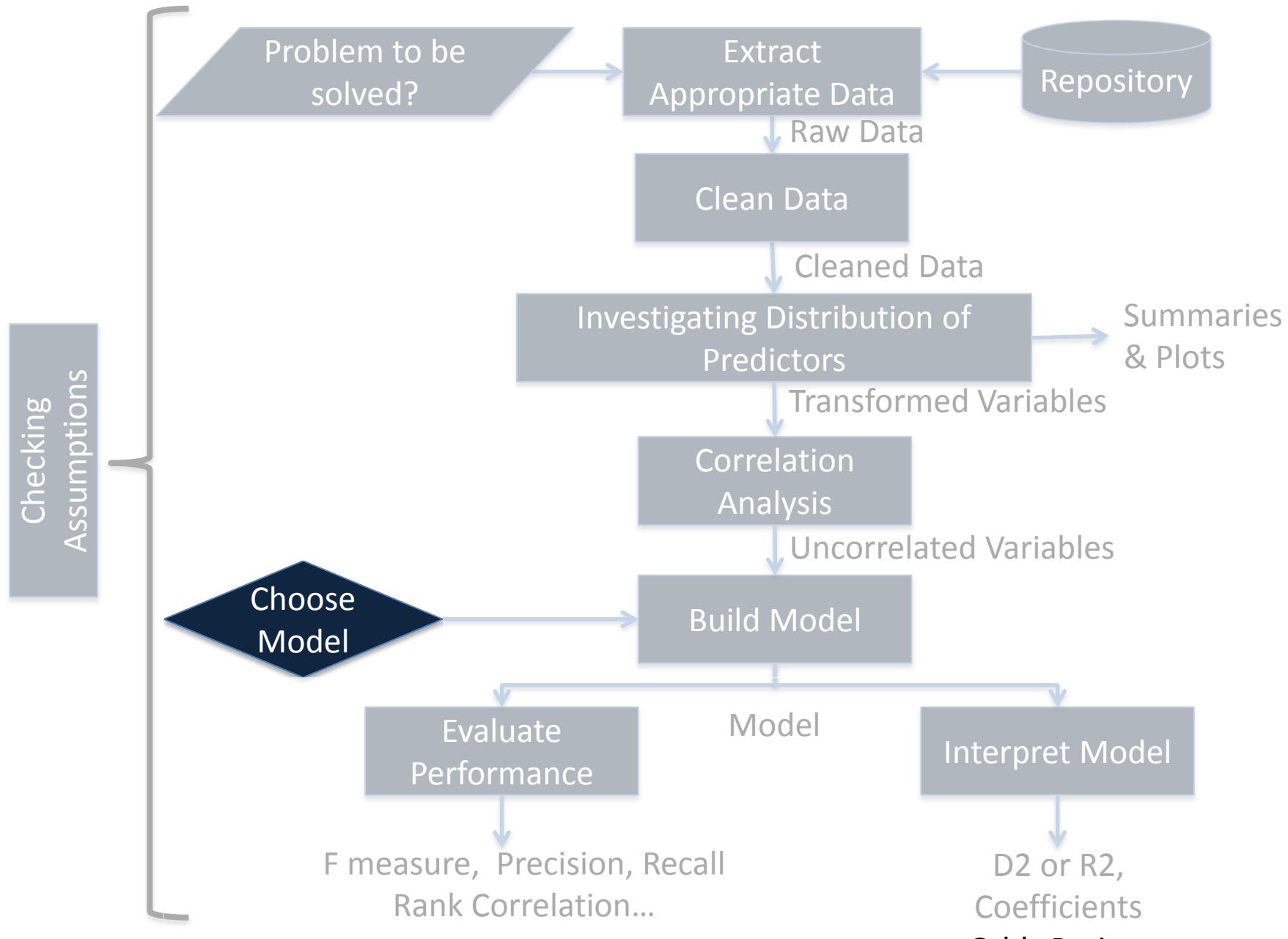
PCA Analysis

Regression Analysis



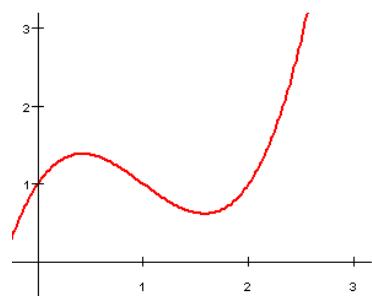
INS, IND, FIX, IEXP, IINT

Workflow



Select appropriate model

Continuous Function



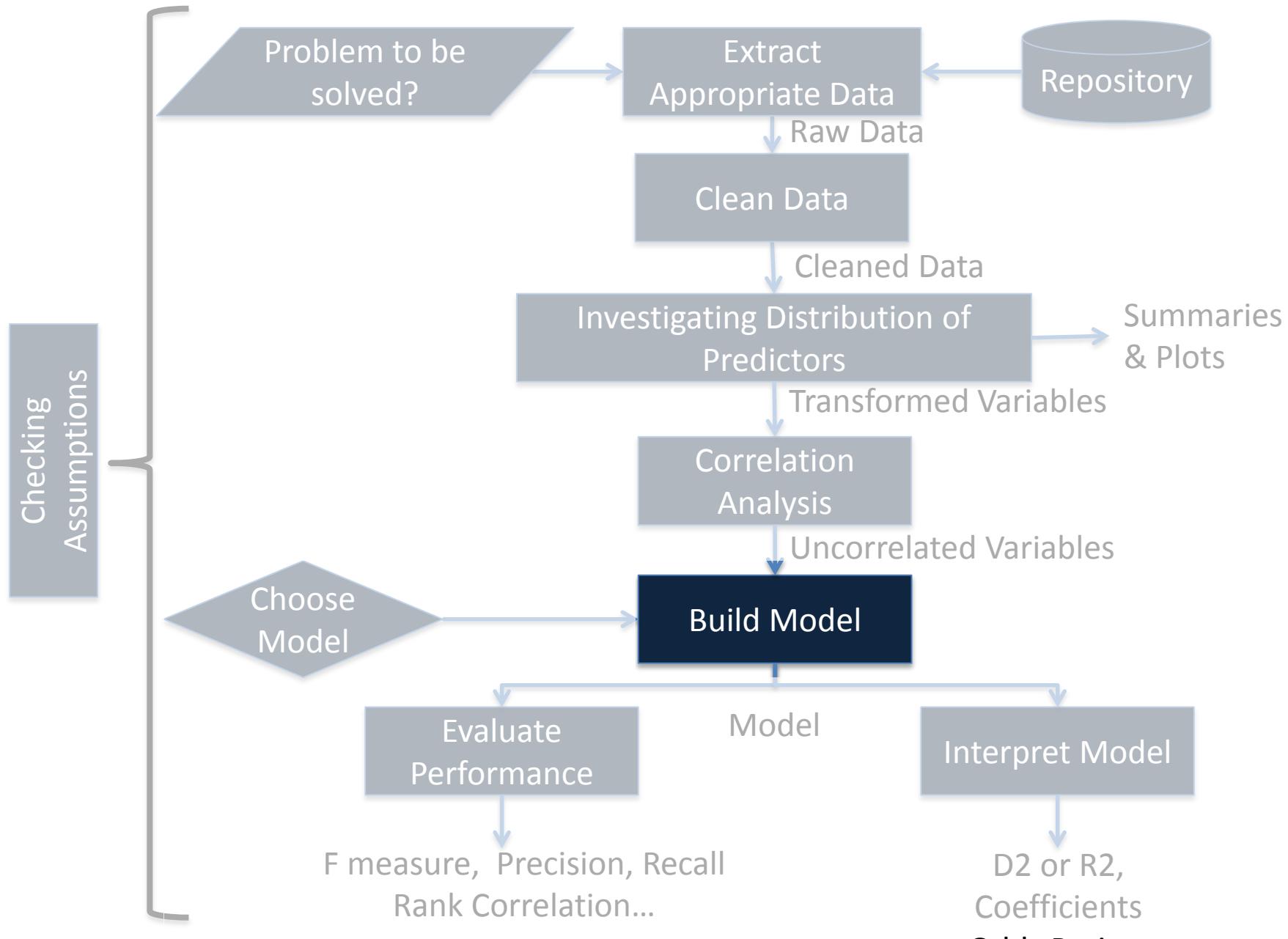
Dependent Variable



Linear Regression

Logistic Regression

Workflow



summary(mod)

Call:

```
glm(formula = isBad ~ lNS + lLA + FIX + lLOC + lINT + lEXP, family = binomial,  
    data = dataFit)
```

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-5.79427	0.45010	-12.87	< 2e-16 ***
lNS	0.35650	0.10268	3.47	0.00052 ***
lLA	0.17910	0.04751	3.77	0.00016 ***
FIX	0.53314	0.12943	4.12	3.8e-05 ***
lLOC	0.13590	0.04948	2.75	0.00603 **
lINT	0.01788	0.00925	1.93	0.05322 .
lEXP	-0.10757	0.03829	-2.81	0.00497 **

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

$$\text{Dev Explained} = \frac{(3011.5 - 2739.0)}{3011.5}$$

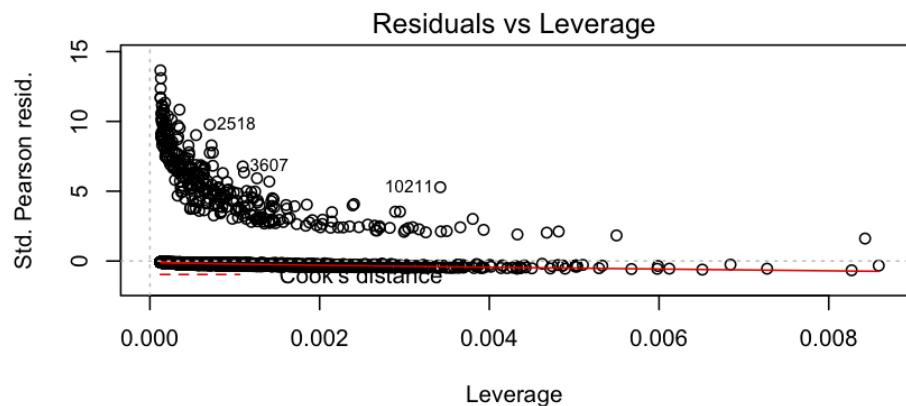
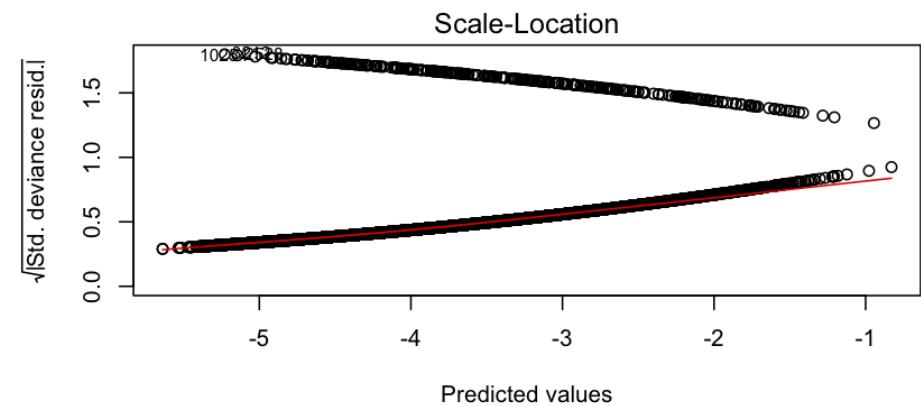
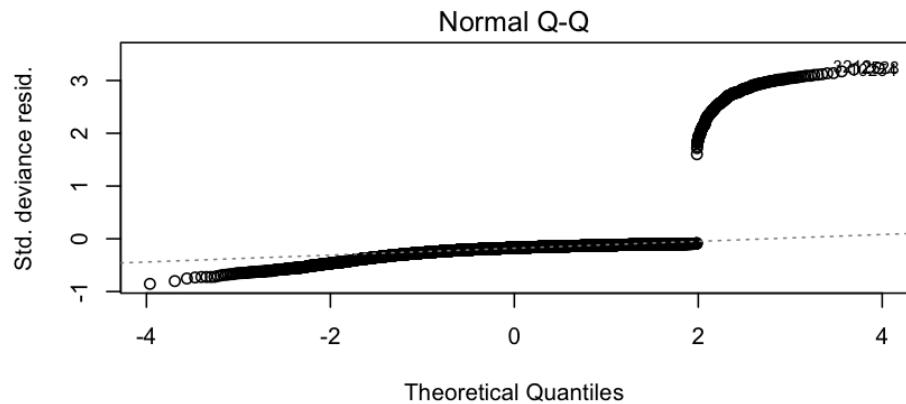
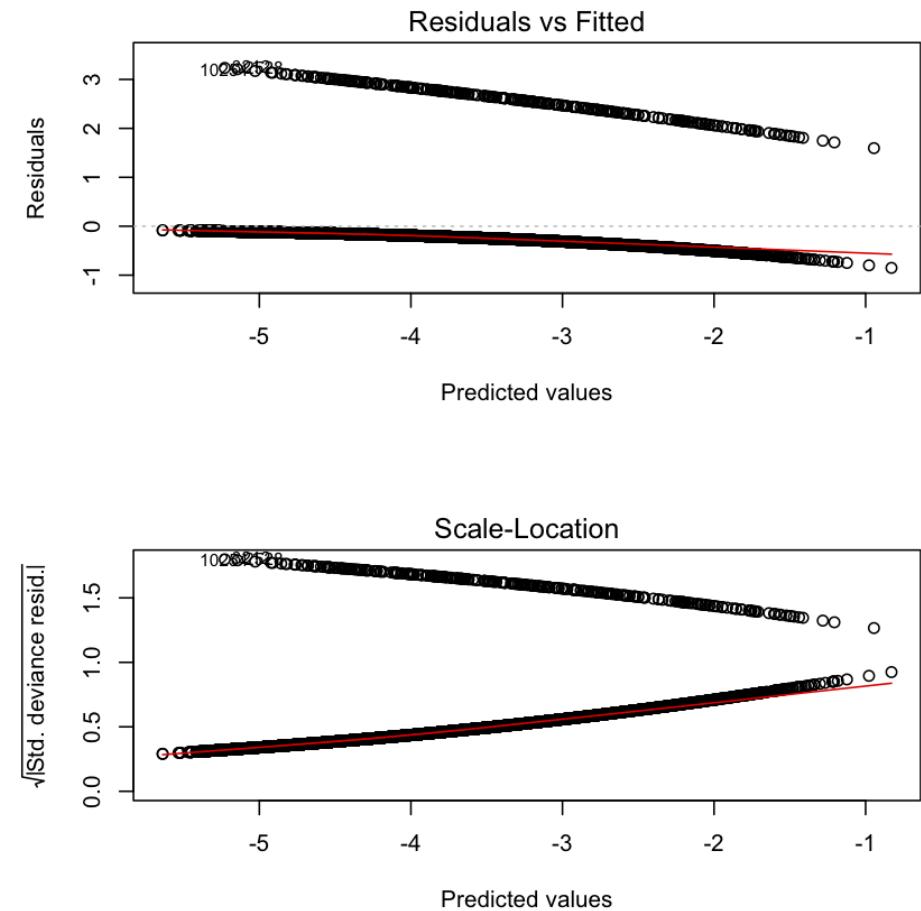
Null deviance: 3011.5 on 13480 degrees of freedom

Residual deviance: 2739.0 on 13474 degrees of freedom

AIC: 2753

Modeling Assumptions

Are there outliers?



Plots not very useful in Logistic Regression

Modeling Assumptions

Are correlations still a problem?

```
vif(mod)
```

```
1NS 1ND FIX 1EXP 1INT
```

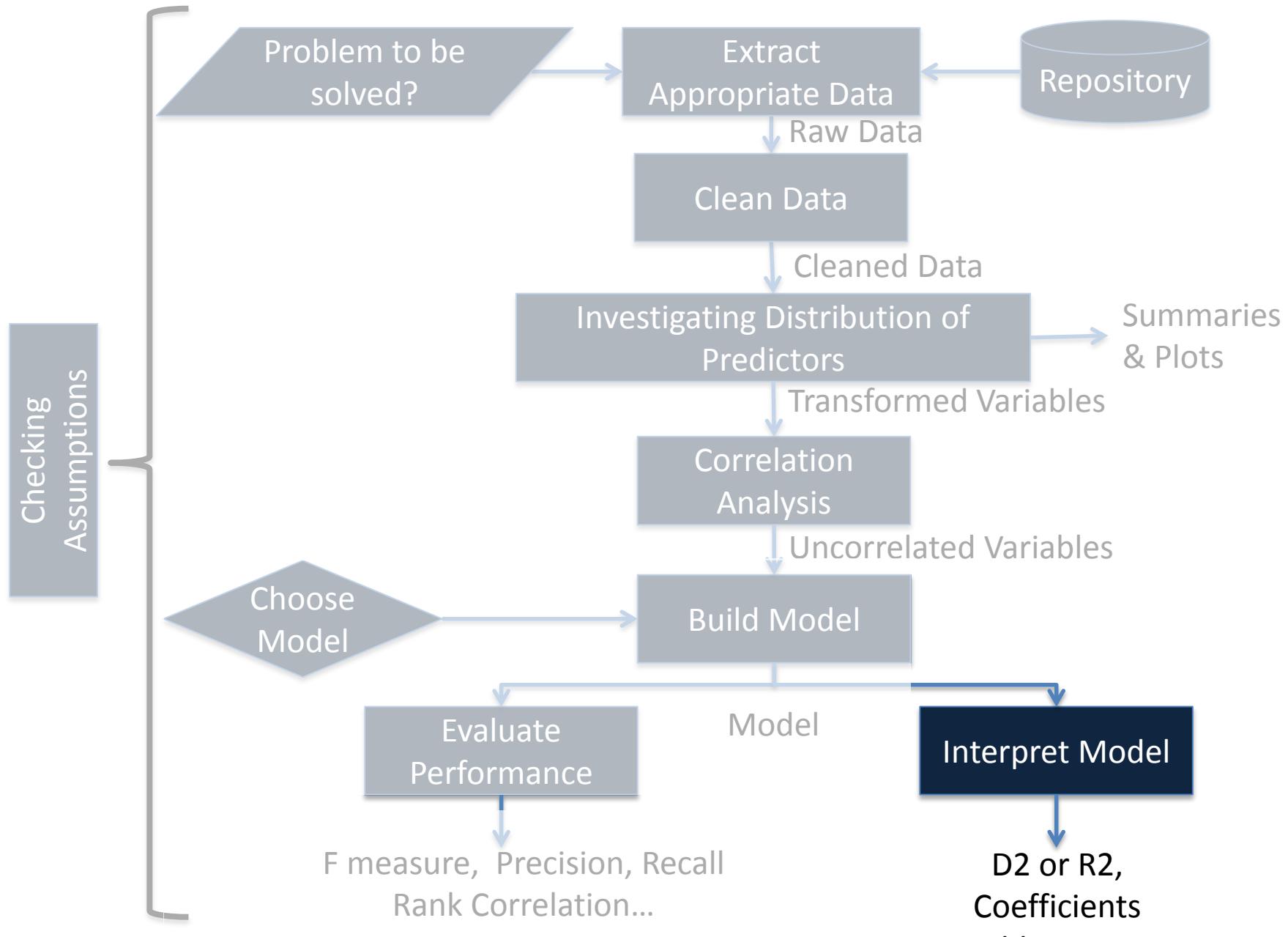
```
2.3 3.1 1.2 1.0 1.9
```

All Vif values < 4

```
#isBad ~ 1NS+1ND+FIX+1INT+1EXP+1LA+from
```



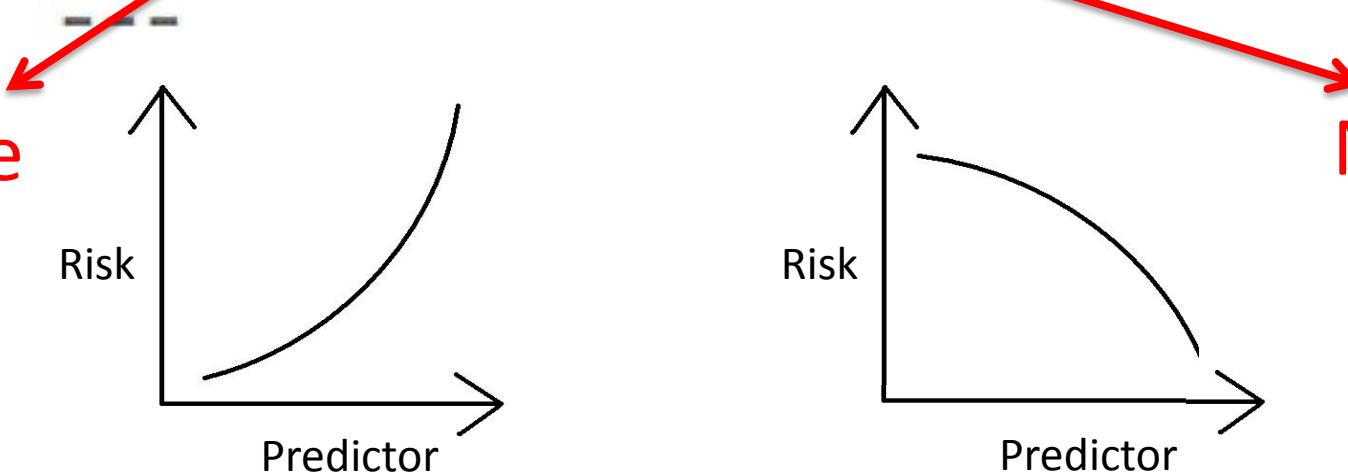
Workflow



Model Interpretation

Coefficients:

	Estimate	Std. Error
(Intercept)	-4.48210	0.28367
LN5	0.31994	0.11021
LN5	0.36503	0.05354
FIX	0.52993	0.12873
LEXP	-0.10508	0.03799
LINT	0.01847	0.00941



Positive
Coeff.

Negative
Coeff.

Using the p-Value

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)	
(Intercept)	-4.48210	0.28367	-15.80	< 2e-16	***
lNS	0.31994	0.11021	2.90	0.0037	**
lND	0.36503	0.05354	6.82	9.2e-12	***
FIX	0.52993	0.12873	4.12	3.8e-05	***
lEXP	-0.10508	0.03799	-2.77	0.0057	**
lINT	0.01847	0.00941	1.96	0.0496	*

Signif. codes:	0 ‘***’	0.001 ‘**’	0.01 ‘*’	0.05 ‘.’	0.1 ‘ ’ 1



If $\text{Pr}(>|z|) < 0.05$

Then the effect of the predictors are statistically significant.

Effect of Predictors on
Response Variables

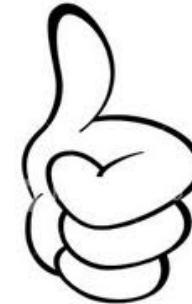
p-values

Large Dataset

Very small p-value

Use $\Pr(>|z|) < 0.001$
or less to consider significant

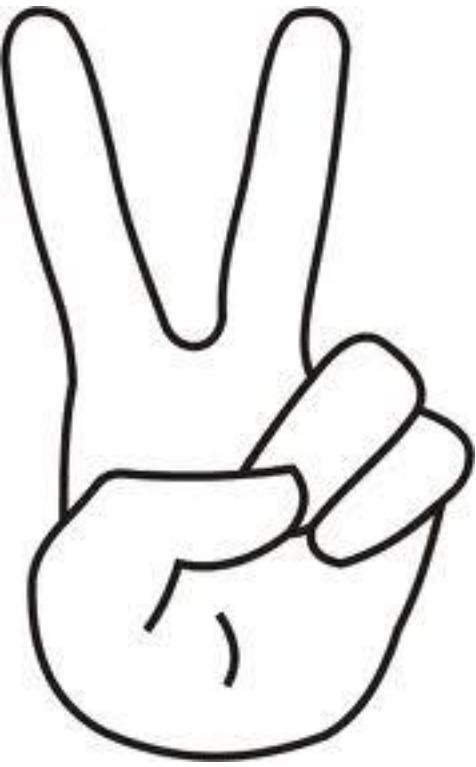
13481 MRs with 318 broken



Effective Sample Size



Using Odds ratio (effect size for logistic regression)



Effect of Predictors on
Response Variables



Hypothetical File
with median values

 INS by 1 unit

Then calculate risk of faulty patch

Measure the  in risk.

Using Anova Analysis

```
> anova(mod);
```

Analysis of Deviance Table

Model: binomial, link: logit

Response: isBad

Terms added sequentially (first to last)



Effect of Predictors on
Response Variables

	Df	Deviance	Resid. Df	Resid. Dev
NULL			13480	3012
LNS	1	166.1	13479	2845
LLA	1	65.1	13478	2780
FIX	1	17.4	13477	2763
LLOC	1	10.7	13476	2752
LINT	1	5.5	13475	2747
LEXP	1	7.7	13474	2739

**p-values from `anova(mod)` are
not p-values from `summary(mod)`**

summary(mod)

Is Regression Coeff \neq Zero

anova(mod)

Does a variable explains extra variance?

(in addition to variables already in the model)

Order of Variables in a Model Matters!

To see how much variance each variable explains if added last: use drop1

```
> drop1(mod)
```

Single term deletions

Model:

isBad ~ lNS + lLA + FIX + lLOC + lINT + lEXP

	Df	Deviance	AIC
<none>		2739	2753
lNS	1	2751	2763
lLA	1	2753	2765
FIX	1	2756	2768
lLOC	1	2747	2759
lINT	1	2743	2755
lEXP	1	2747	2759

In a linear regression typically want at least adjR²

Deviance explained in our Data

40%

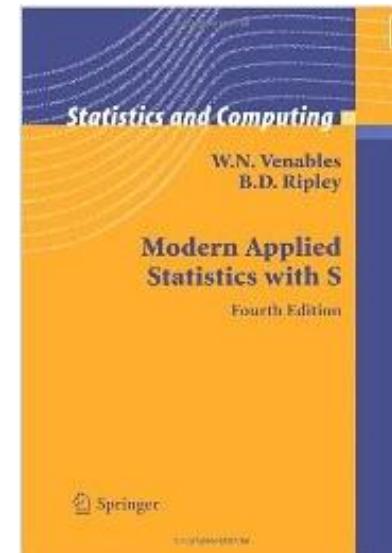
99%

Good for logistic regression and 2% average probability

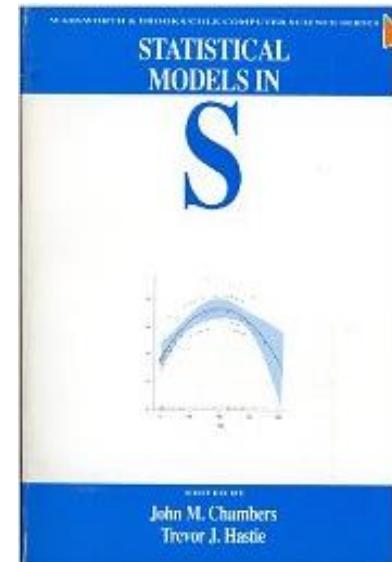
Low values of adjR² => model only partly reflects mechanism generating the data

Mass of online resources for R

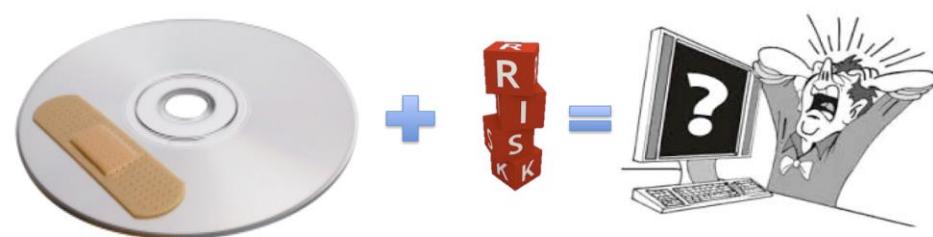
Modern Applied Statistics with S
(Statistics and Computing) [Paperback]
W.N. Venables (Author), B.D. Ripley



Statistical Models in S [Paperback] John
M. Chambers (Editor), Trevor J. Hastie
(Editor)



Predicting Risk of Software Updates (Patches)



Customer = Developer

Identifying possible causes of failure (Independent Variable/Predictors)

Formulate a hypothesis for each cause

Pick interpretable hypothesis

Abandon ones that rely on poor quality data

Select several from each conceptual domain

**Based on Problem and Audience
Use Common Sense to pick a variable**

Examine Correlations

Examine Principal Components

Regress each Predictor

Avoid automatic methods

Model Interpretation

Coefficients:	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-4.48210	0.28367	-15.80	< 2e-16 ***
LNS	0.31994	0.11021	2.90	0.0037 **
LND	0.36503	0.05354	6.82	9.2e-12 ***
FIX	0.52993	0.12873	4.12	3.8e-05 ***
LEXP	-0.10508	0.03799	-2.77	0.0057 **
LINT	0.01847	0.00941	1.96	0.0496 *

Signif. codes:	0 '***'	0.001 '**'	0.01 '*'	0.05 '.'
	0.1 ' '	1		

