Analysis of Software Data: Basic Good Practices

2018 Midwest Big Data School

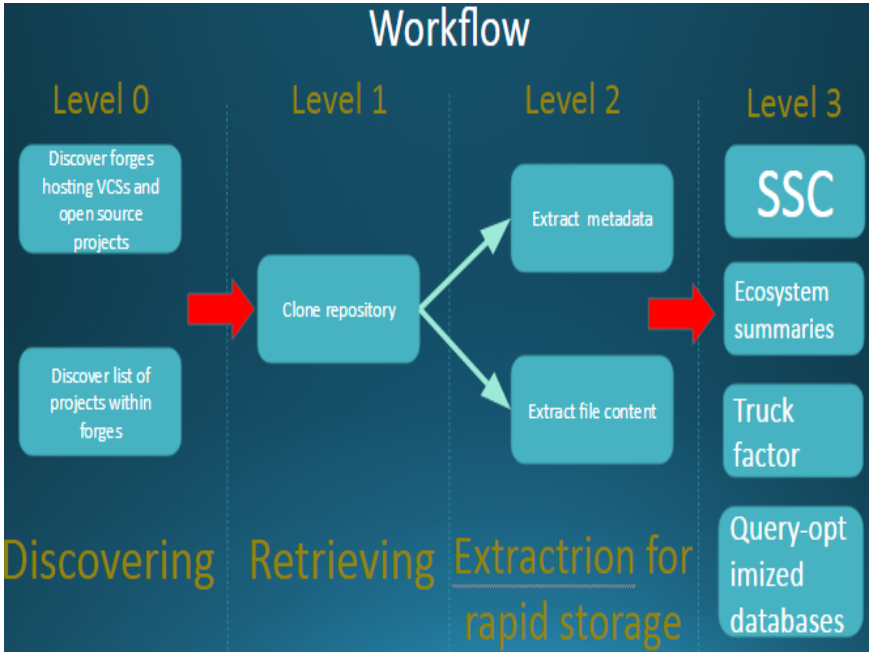
Audris Mockus

University of Tennessee

Based on Tutorials

at FSE'13 and ICSE'14

Class CS445/545 in Fall 1[4-7]



How to use notebook:
Use Docker Container

- Clone data + notebook

```
git clone https://github.com/fdac17/MWBigData
```

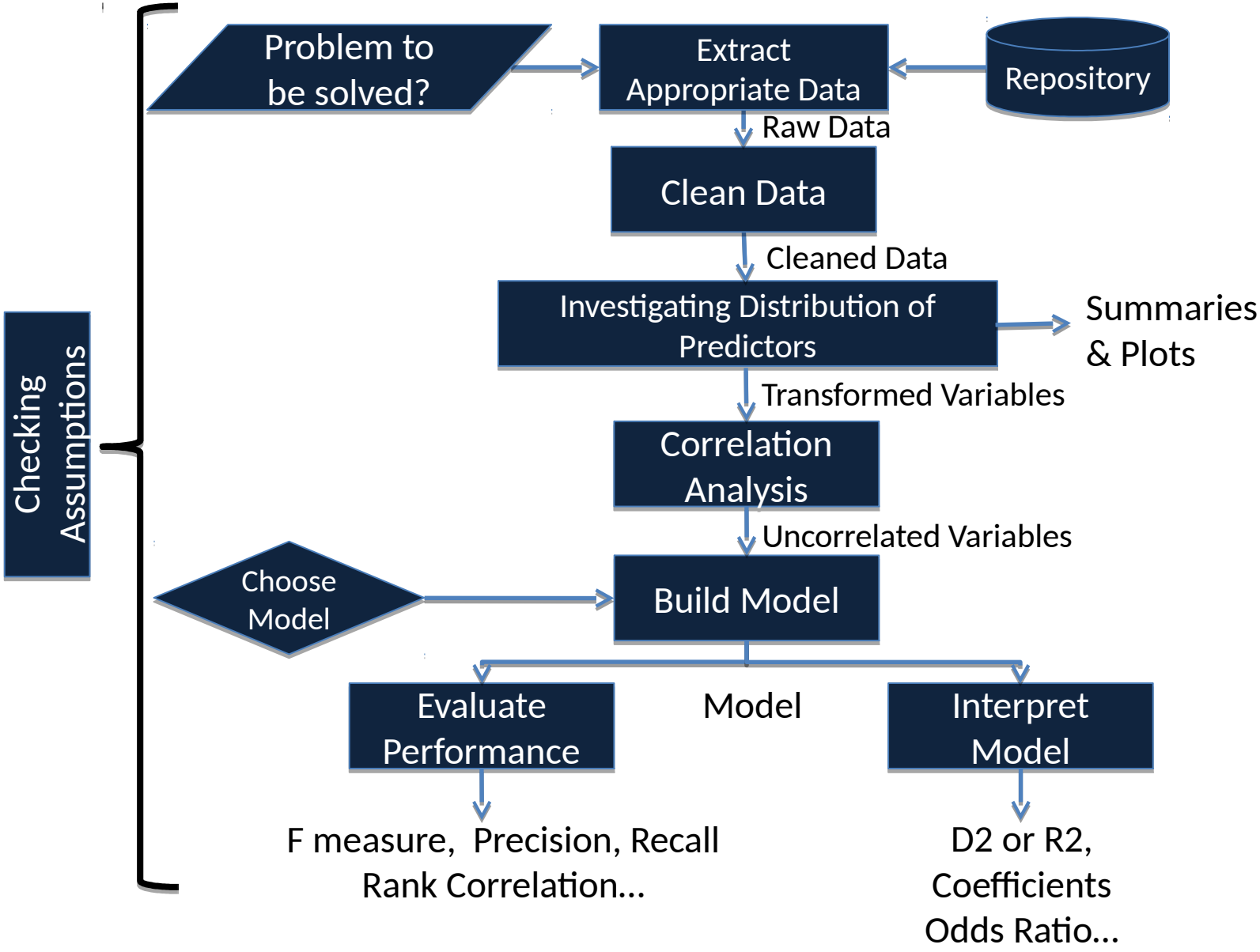
- Install docker engine
- Get docker container
- Run docker container

```
docker pull audris/jupyter-r
docker run -v $(pwd):/home/jovyan -w /home/jovyan -p 8888:8888 \
   audris/jupyter-r  /bin/startDef.sh jovyan
```

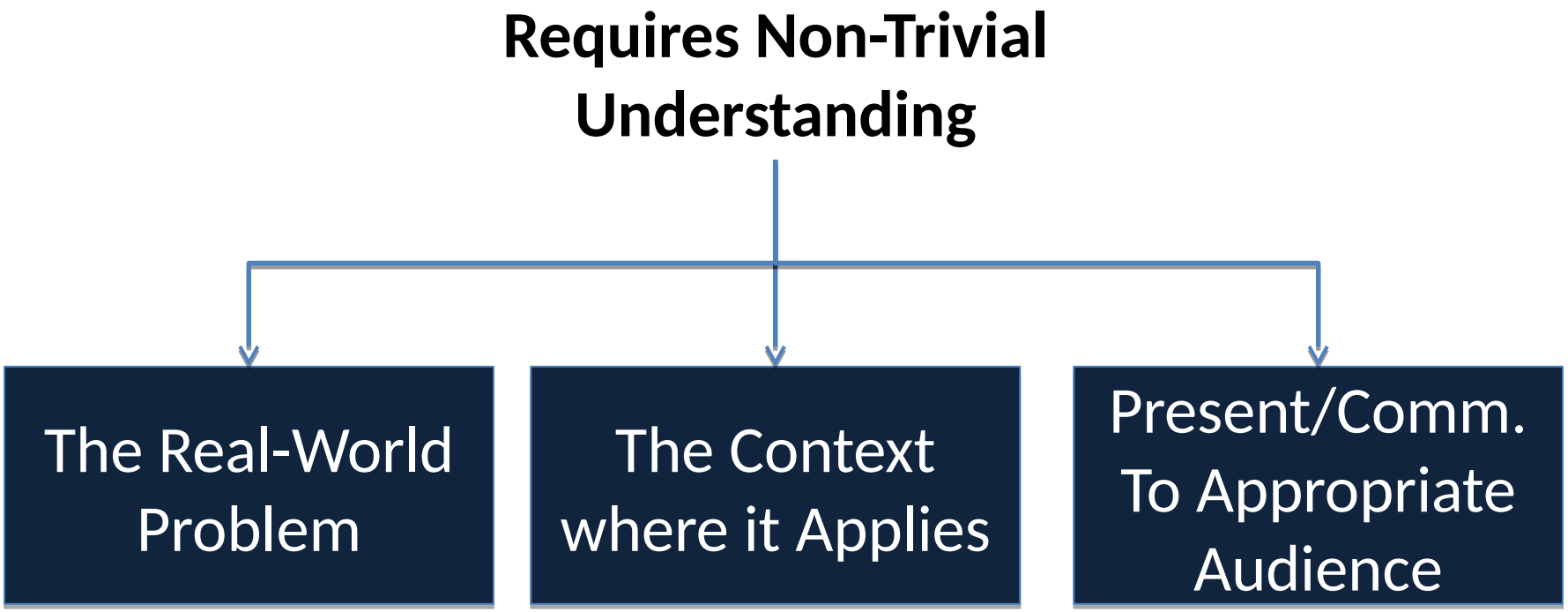
- Enter localhost:8888 into your browser url

And open Tutorial Jupyter notebook: MWBigDataNotebook.ipynb

Workflow


Problem to be solved?
Extract Appropriate Data
Repository
Raw Data
Clean Data
Cleaned Data
Investigating Distribution of Predictors
Summaries & Plots
Transformed Variables
Correlation Analysis
Uncorrelated Variables
Checking Assumptions
Choose Model
Build Model
Evaluate Performance
Model
Interpret Model
F measure, Precision, Recall Rank Correlation...
D2 or R2, Coefficients Odds Ratio...

Statistical thinking is much more than just p-values

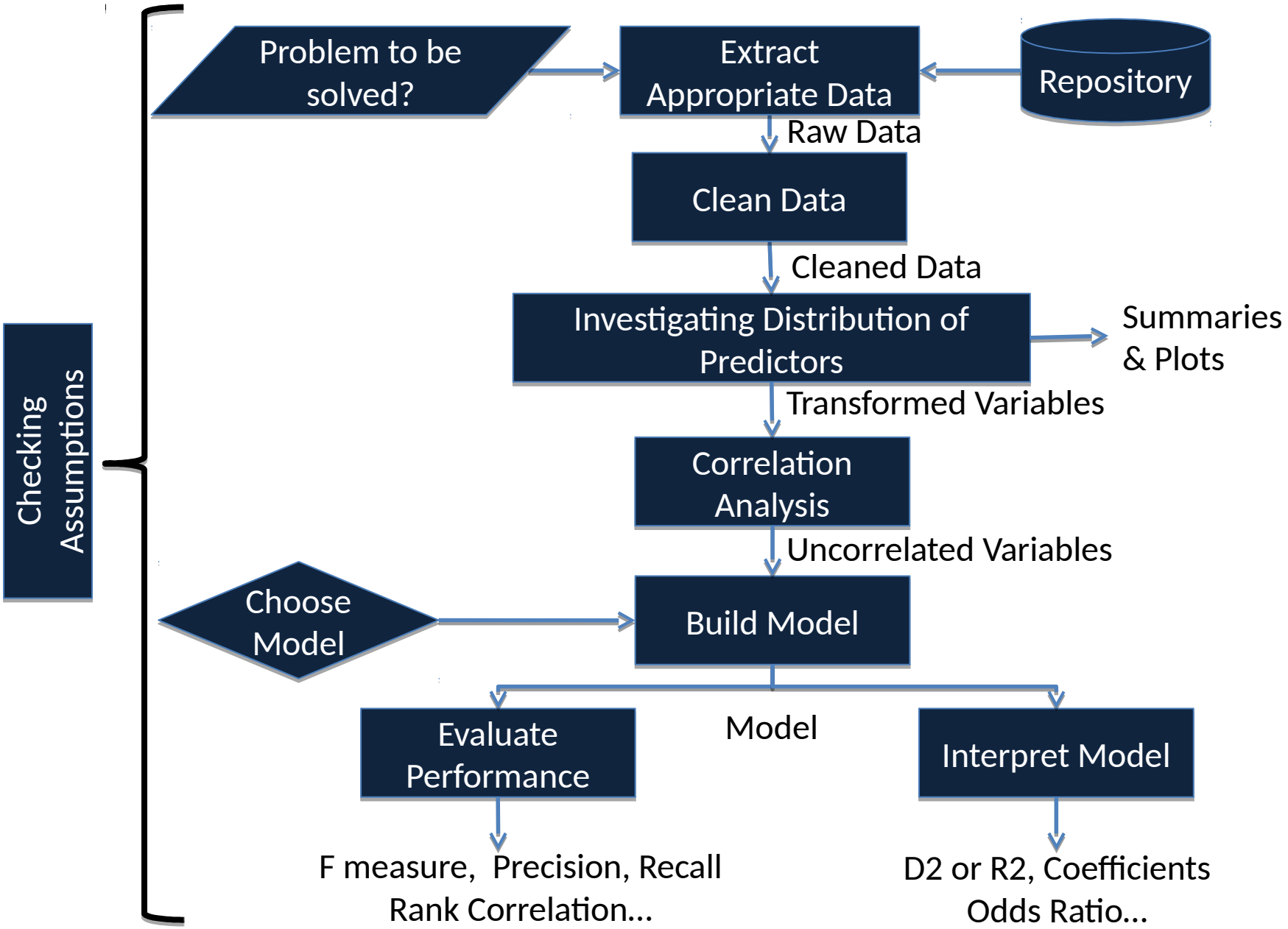


Take Home

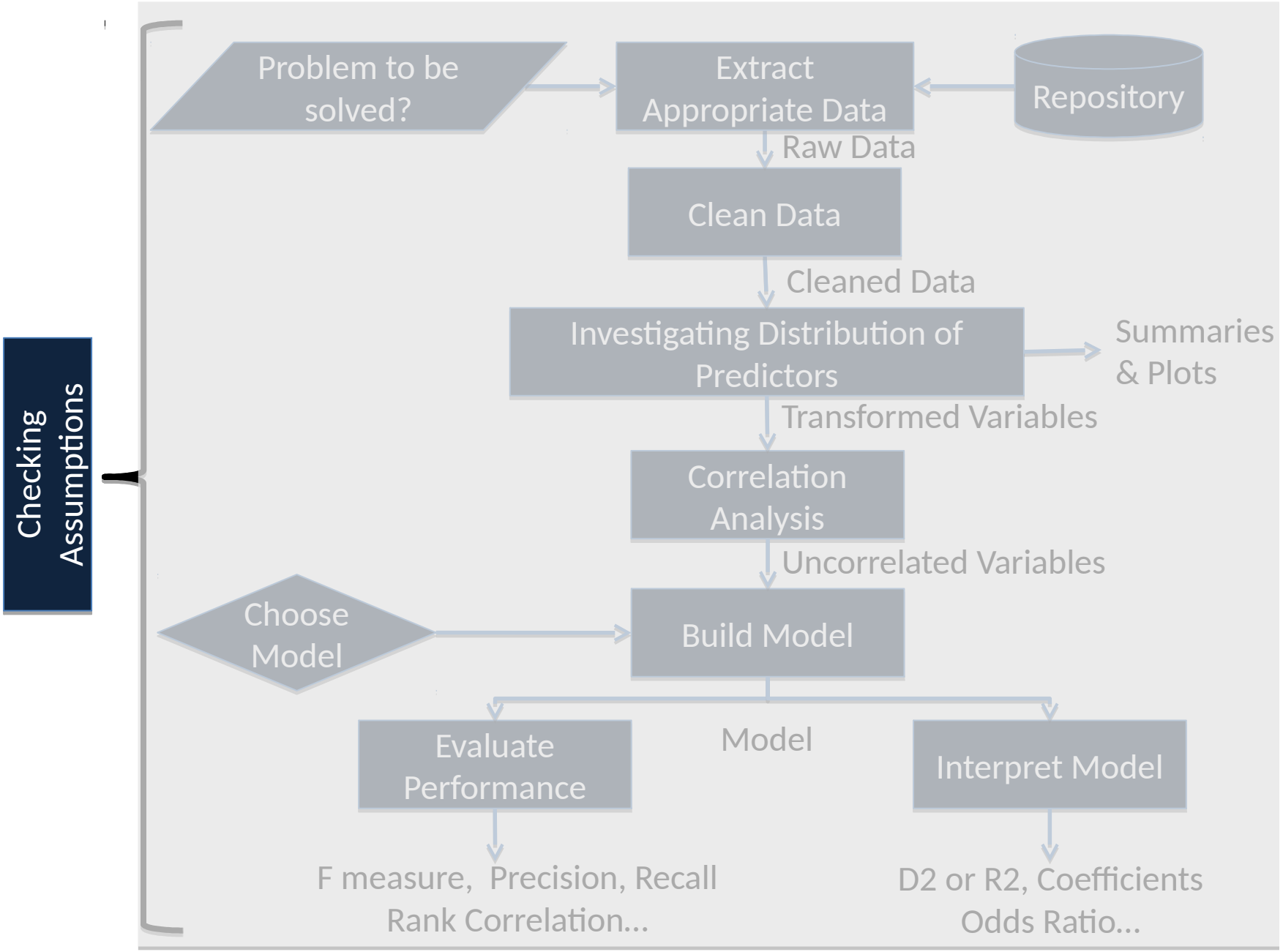
Steps to build a statistical model for operational data

Sample R scripts for an example problem

Workflow


Problem to be solved?
Extract Appropriate Data
Repository
Raw Data
Clean Data
Cleaned Data
Investigating Distribution of Predictors
Summaries & Plots
Transformed Variables
Correlation Analysis
Uncorrelated Variables
Choose Model
Build Model
Model
Evaluate Performance
Interpret Model
F measure, Precision, Recall
Rank Correlation...
D2 or R2, Coefficients
Odds Ratio...
Checking Assumptions

Workflow


Problem to be solved?
Extract Appropriate Data
Repository
Raw Data
Clean Data
Cleaned Data
Investigating Distribution of Predictors
Summaries & Plots
Transformed Variables
Correlation Analysis
Uncorrelated Variables
Choose Model
Build Model
Model
Evaluate Performance
Interpret Model
F measure, Precision, Recall Rank Correlation...
D2 or R2, Coefficients Odds Ratio...
Checking Assumptions

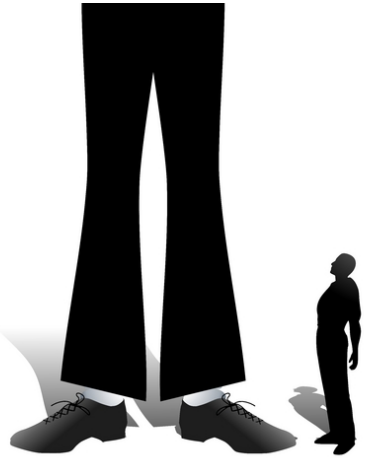
Assumptions

of LOC

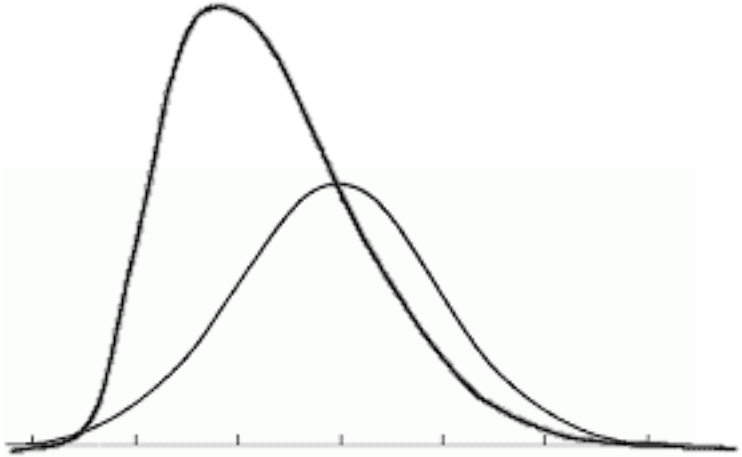
vs.

# of Coffees

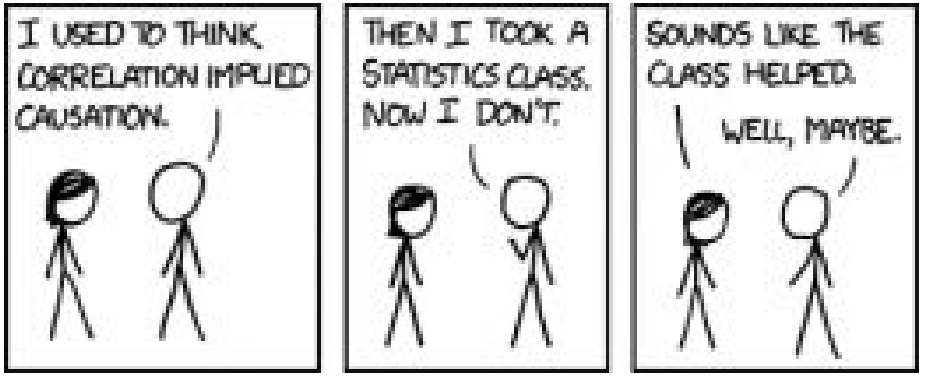
Relevant Data



Representative Data

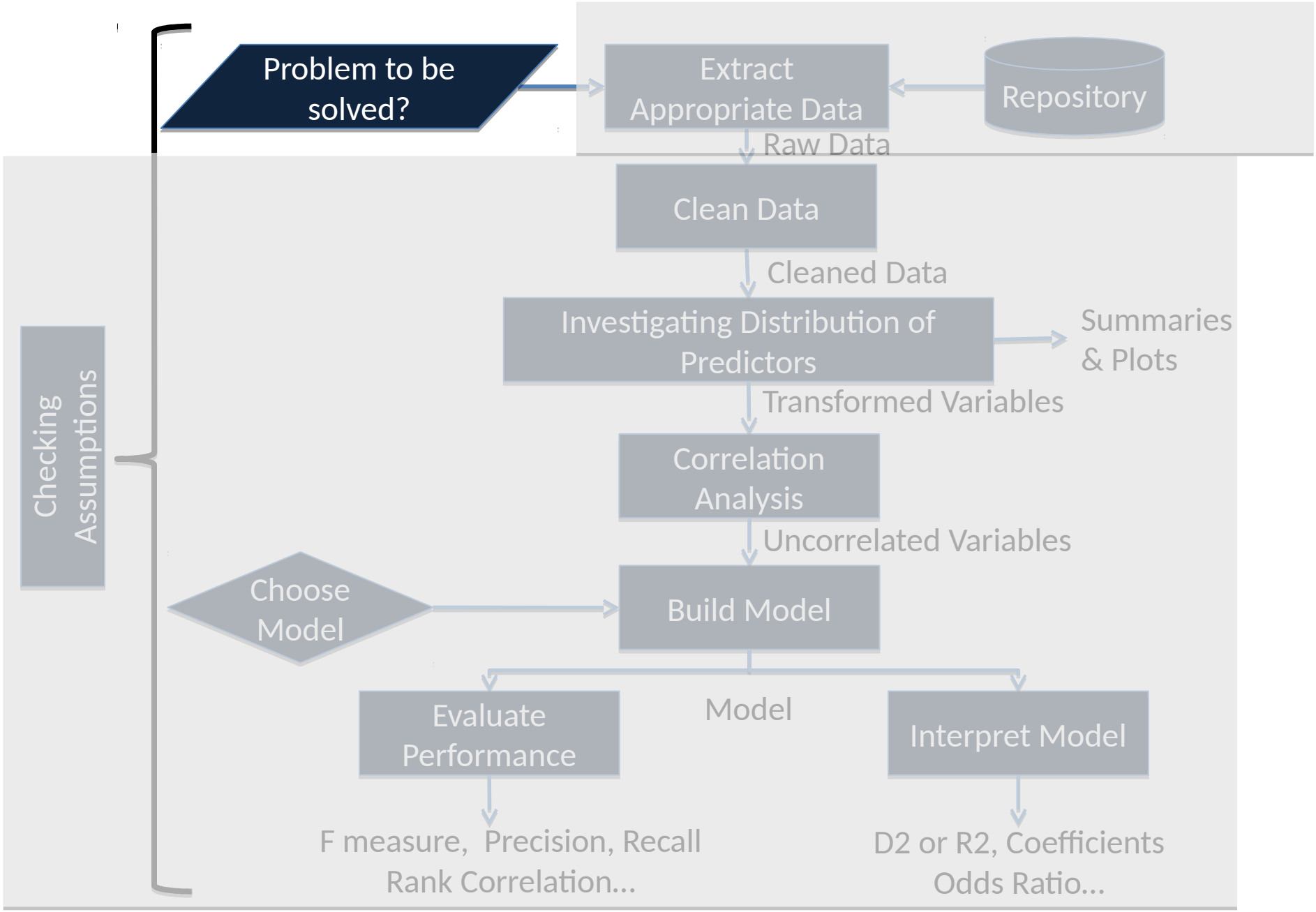


Model Assumptions



Correlation vs. Causation

Tutorial Workflow



Problem to be Solved

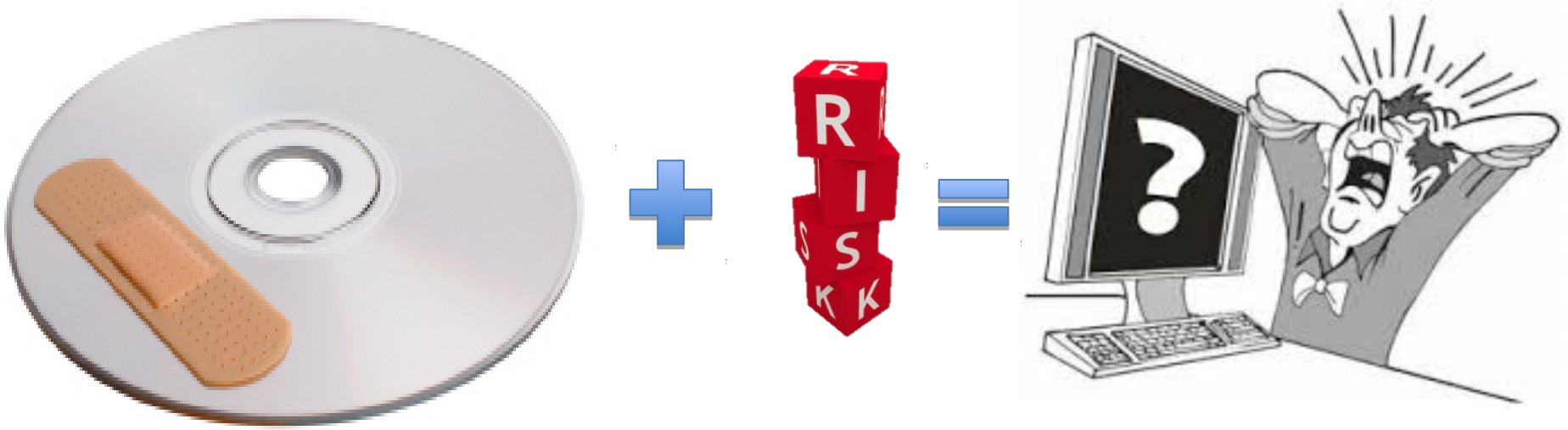
Test Case Prioritization

Tester

Metrics vs. Bugs

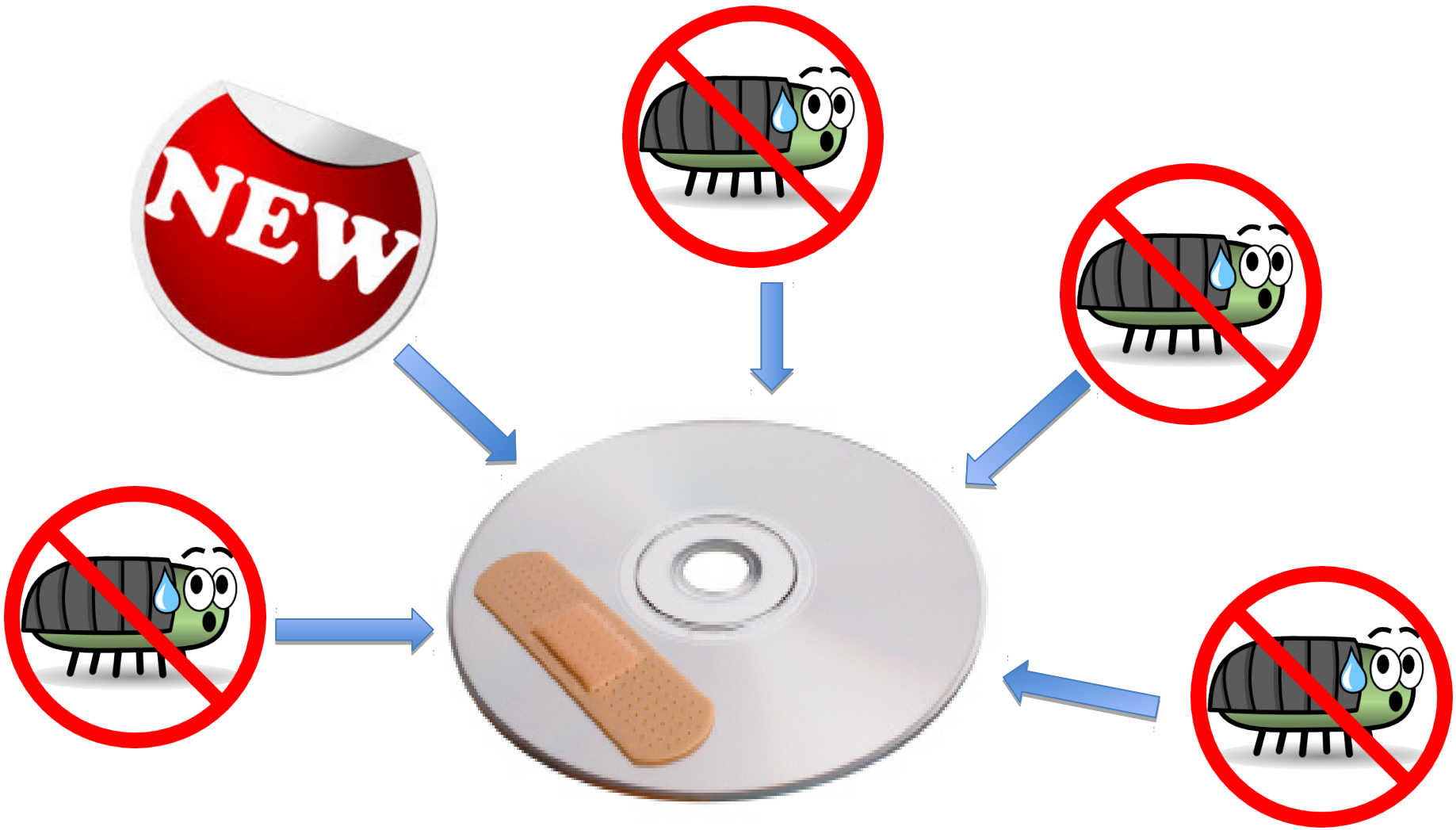
Product Manager

Predicting Risk of Software Updates (Patches/MRs)

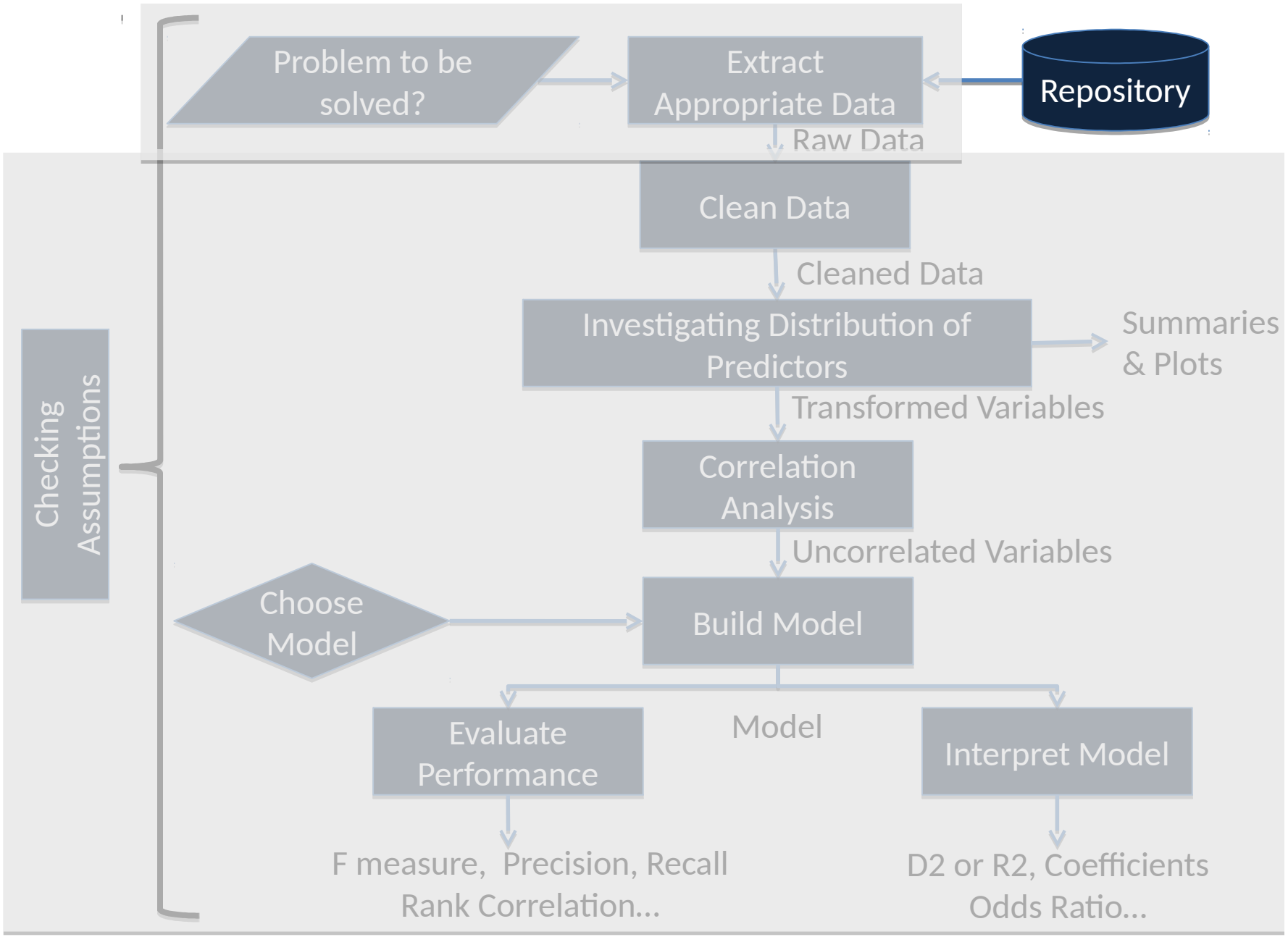


Customer = Developer

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



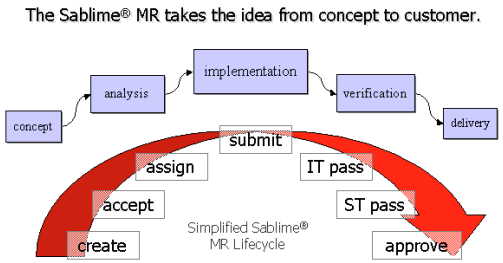
Workflow


Problem to be solved?
Extract Appropriate Data
Repository
Raw Data
Clean Data
Cleaned Data
Investigating Distribution of Predictors
Summaries & Plots
Transformed Variables
Correlation Analysis
Uncorrelated Variables
Checking Assumptions
Choose Model
Build Model
Model
Evaluate Performance
Interpret Model
F measure, Precision, Recall
Rank Correlation...
D2 or R2, Coefficients
Odds Ratio...

Lots of Rich Data Sources from Software Development

Code Repositories
SCCS

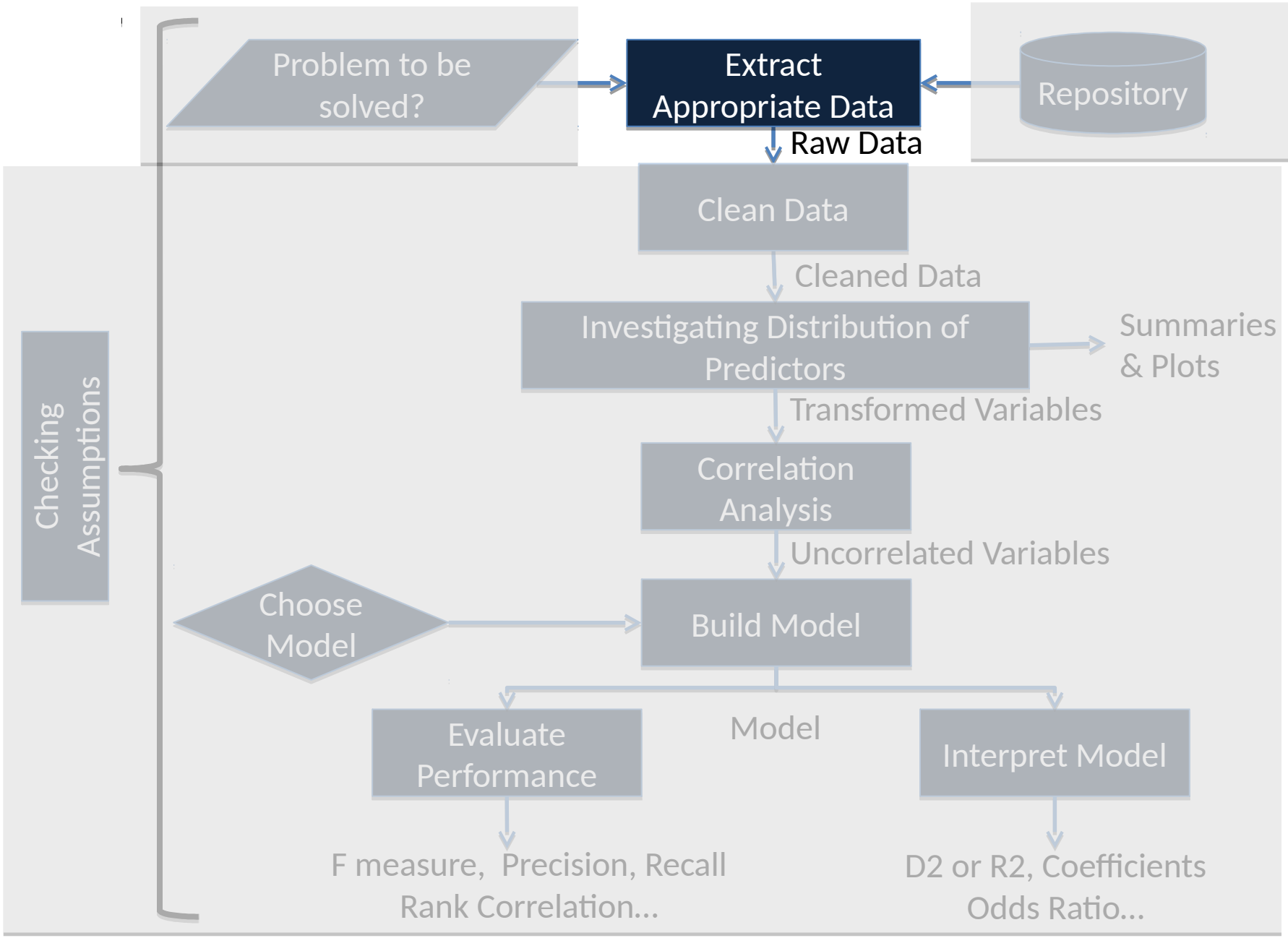
Issue Repository
ECMS/Sablime



Commit Information:
Date, Change, Developer…

Issue Information:
Date, Status, Assigned to…

Workflow


Problem to be solved?
Extract Appropriate Data
Repository
Raw Data
Clean Data
Cleaned Data
Investigating Distribution of Predictors
Summaries & Plots
Transformed Variables
Correlation Analysis
Uncorrelated Variables
Checking Assumptions
Choose Model
Build Model
Model
Evaluate Performance
Interpret Model
F measure, Precision, Recall
Rank Correlation...
D2 or R2, Coefficients
Odds Ratio...

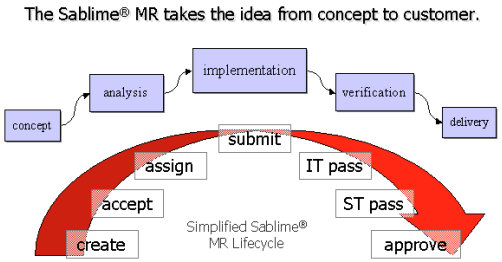
Extract Raw Data Required for the Problem!

Code Repositories

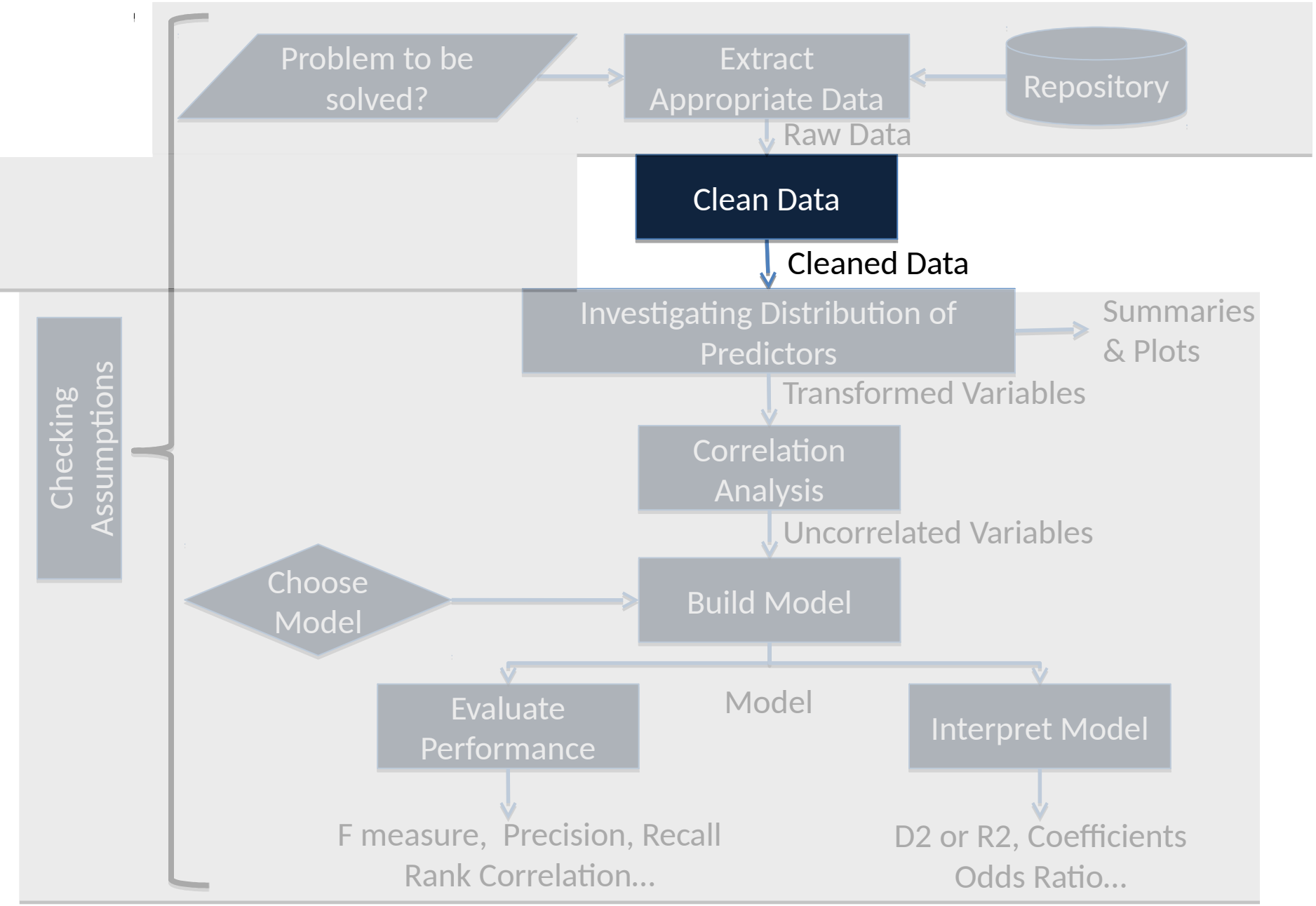
Link

Data

Issue Repository

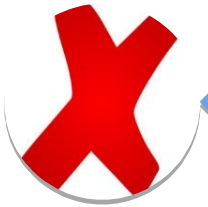


Workflow


Problem to be solved?
Extract Appropriate Data
Repository
Raw Data
Clean Data
Cleaned Data
Investigating Distribution of Predictors
Summaries & Plots
Transformed Variables
Correlation Analysis
Uncorrelated Variables
Checking Assumptions
Choose Model
Build Model
Model
Evaluate Performance
Interpret Model
F measure, Precision, Recall
Rank Correlation...
D2 or R2, Coefficients
Odds Ratio...

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

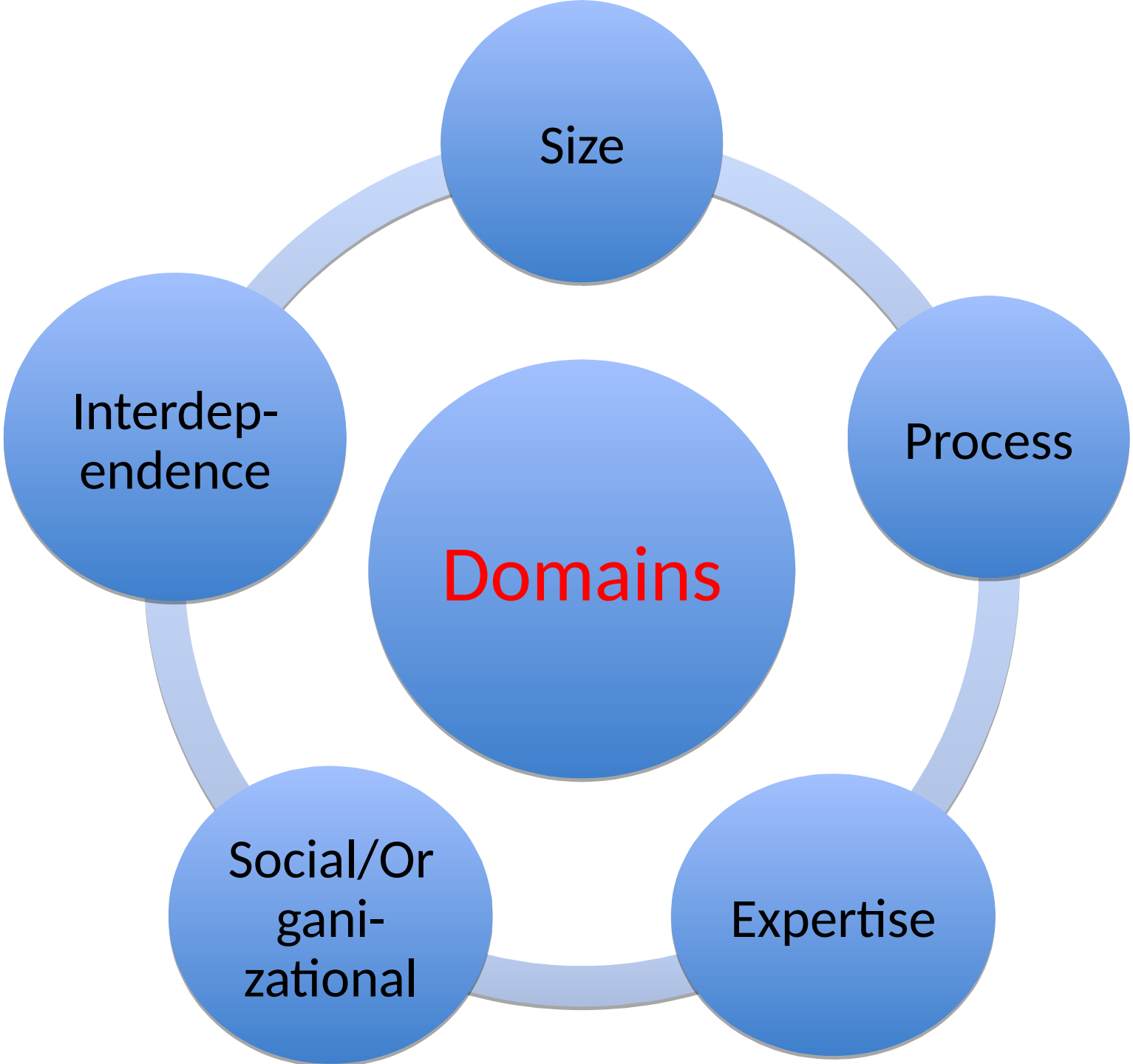
H3: Long duration may indicate otherwise invisible problems

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)

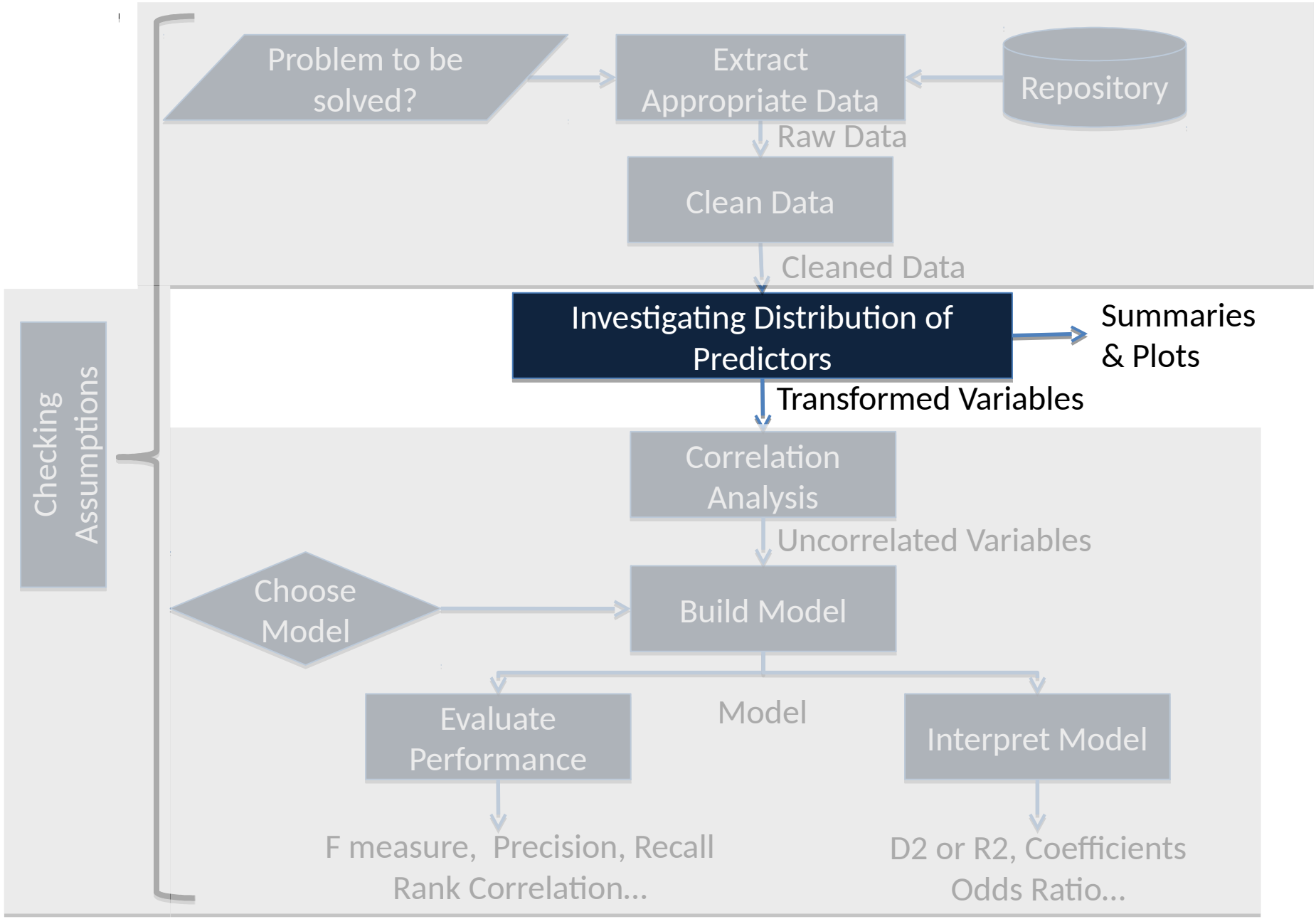
H7: Social/organizational/interdependence measures


Size
Process
Expertise
Social/Or gani- zational
Interdep- endence
Domains

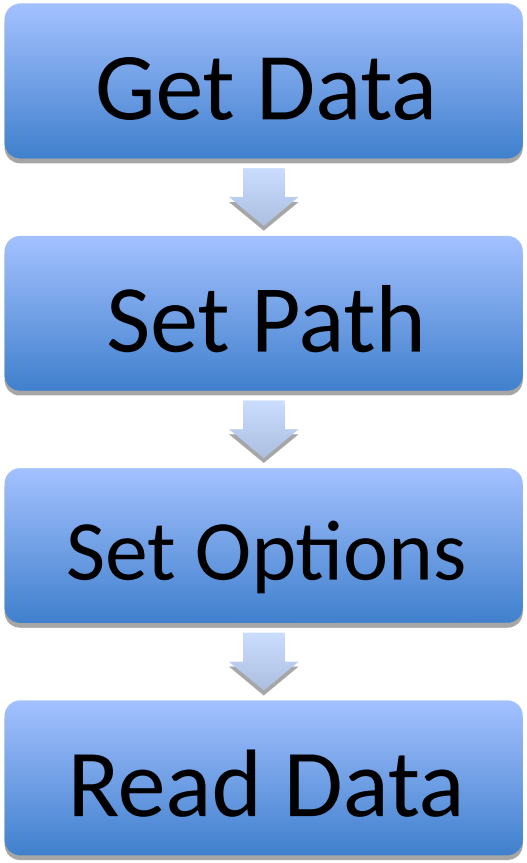
Final Set of Predictors

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.
	LA	LOC added.
Size	LD	LOC deleted.
	LT	LOC in the files touched by the change.
Diffusion and size	NMR	Number of MRs.
	ND	Number of deltas.
Interval	INT	Time between the last and first delta.
Purpose	FIX	Fix of a defect found in the field.
Experience	EXP	Developer experience.
	REXP	Recent developer experience.
	SEXP	Developer experience on a subsystem.

Workflow


Problem to be solved?
Extract Appropriate Data
Repository
Raw Data
Clean Data
Cleaned Data
Investigating Distribution of Predictors
Summaries & Plots
Transformed Variables
Checking Assumptions
Correlation Analysis
Uncorrelated Variables
Choose Model
Build Model
Model
Evaluate Performance
Interpret Model
F measure, Precision, Recall
Rank Correlation...
D2 or R2, Coefficients
Odds Ratio...

Read Data into R



Follow along the R script @

5 Number Summary of Predictors

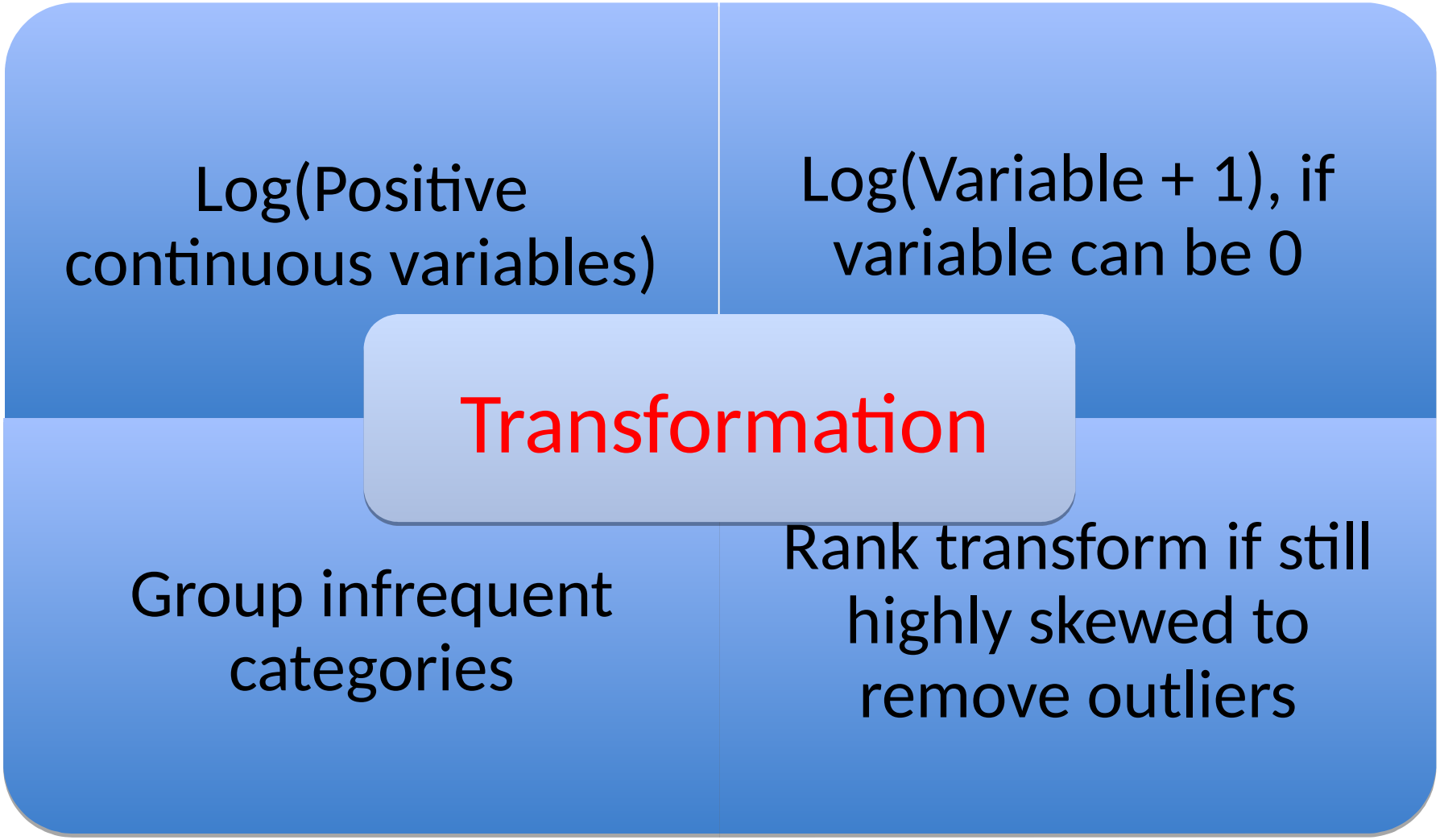
```
summary (risk);
     isBad                NS                LA
 Min.   :0.0000   Min.   : 1.00   Min.   :      0
 1st Qu.:0.0000   1st Qu.: 1.00   1st Qu.:     13
 Median :0.0000   Median : 1.00   Median :     47
 Mean   :0.0211   Mean   : 1.65   Mean   :   1289
 3rd Qu.:0.0000   3rd Qu.: 2.00  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

Skewed data is not normally distributed!

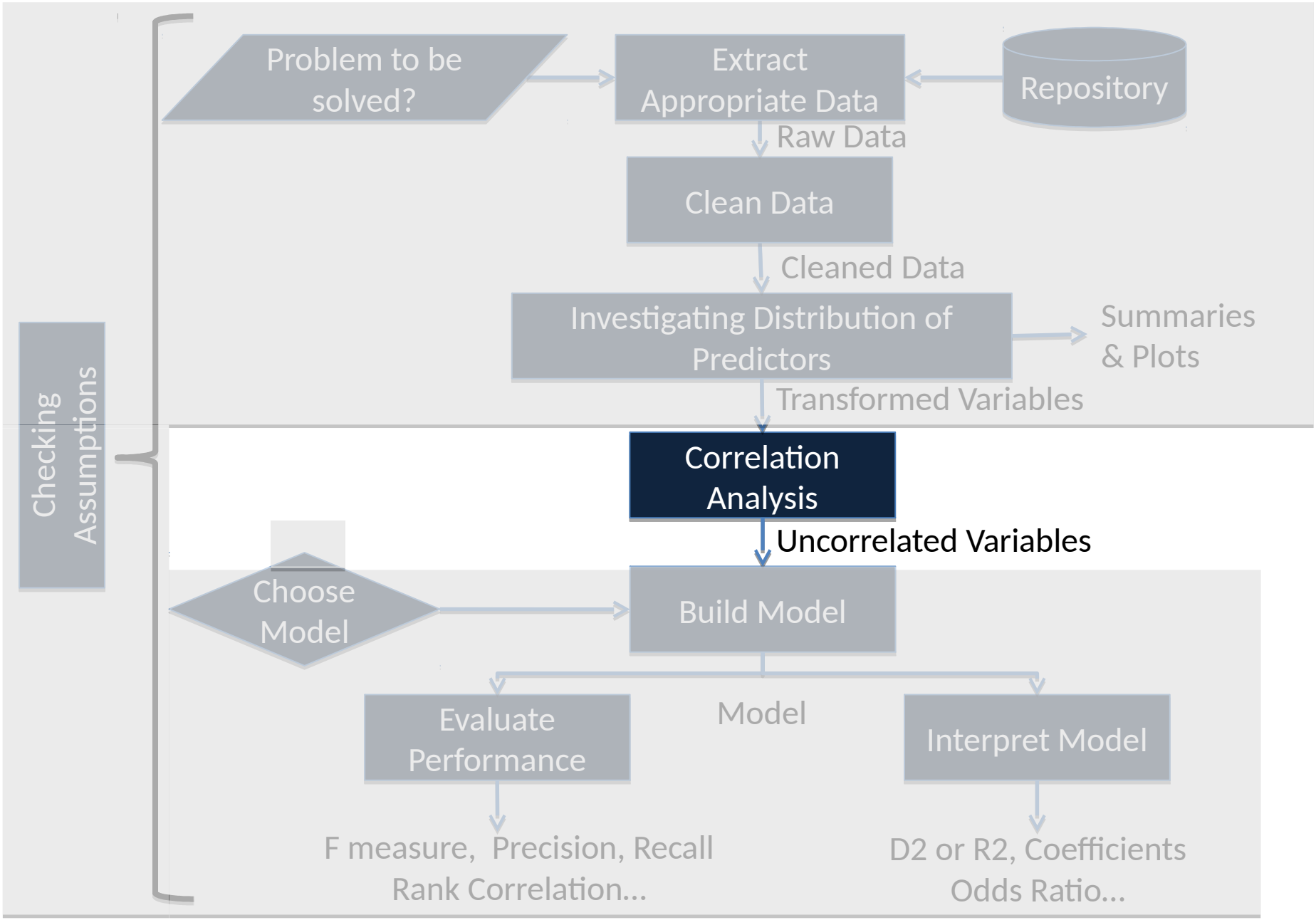
Log Transform Predictors

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



Tutorial Workflow



High correlation is typical in SE

cor(vars); # OK for normal distrubution
cor(vars,method="spearman"); #OK for any: uses ranks

hiCor(data,.7); # check script for method definition

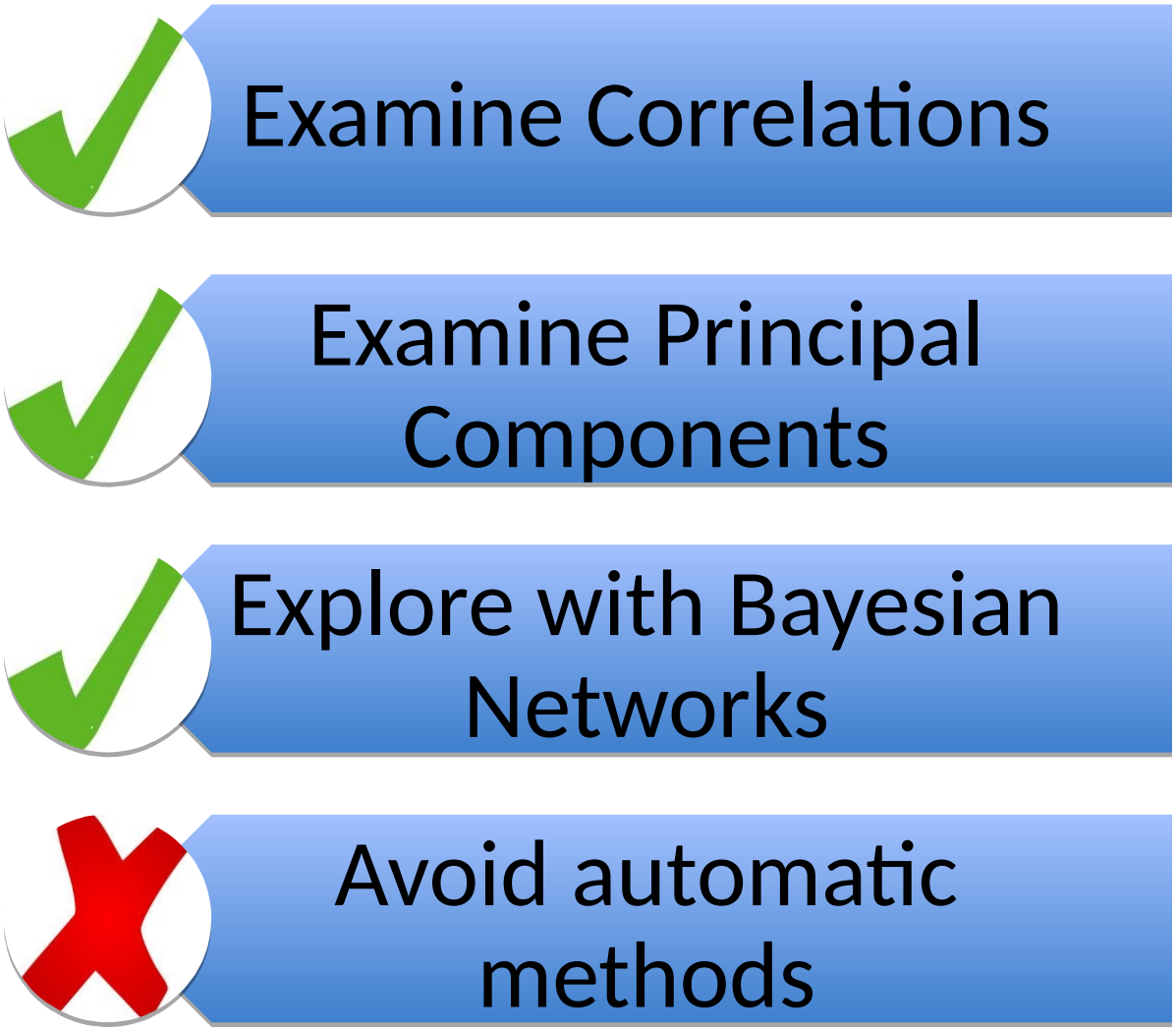
	lNS	lNM	lNF	lNMR	lND	lLA	lLD	lLOC	lEXP	lREXP
lNS	1.000	0.7551	0.6501	0.652	0.577	0.512	0.477	0.489	-0.0272	-0.0191
lNM	0.755	1.0000	0.8561	0.743	0.744	0.653	0.591	0.605	-0.0092	0.0063
lNF	0.650	0.8561	1.0000	0.770	0.862	0.757	0.672	0.647	-0.0160	0.0031
lNMR	0.652	0.7432	0.7702	1.000	0.835	0.736	0.696	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

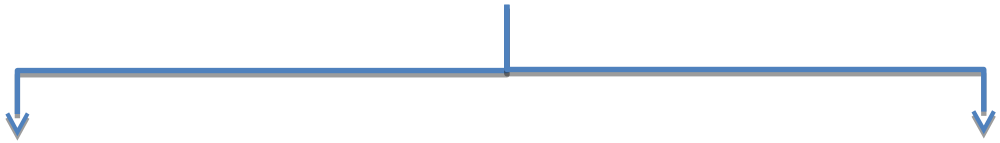
Based on Problem and Audience
Use Common Sense to pick a variable



Use PCA to understand the joint distribution of predictors

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

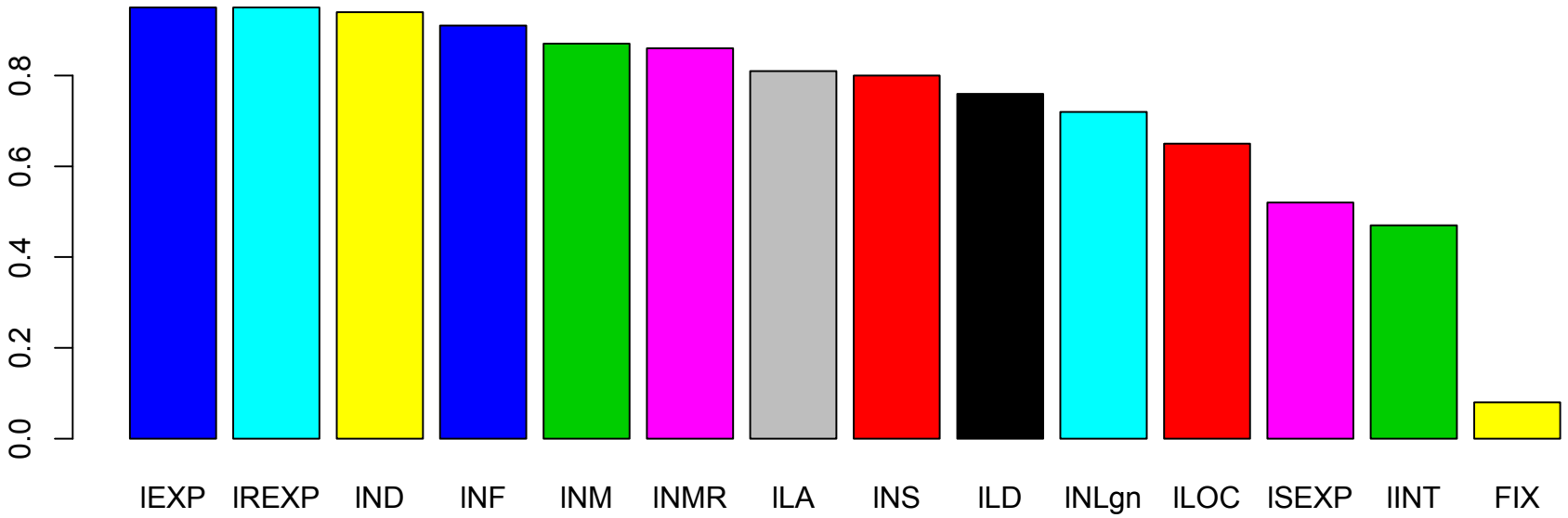
```
      lNM  lNF lNLgn lNMR  lND  lLA lLD
PC1 0.32 0.33  0.31 0.34 0.34 0.31 0.3
     lEXP lREXP lSEXP
PC2 -0.65 -0.66 -0.33
```



Diffusion & Size (PC1)

Developer Expertise (PC2)

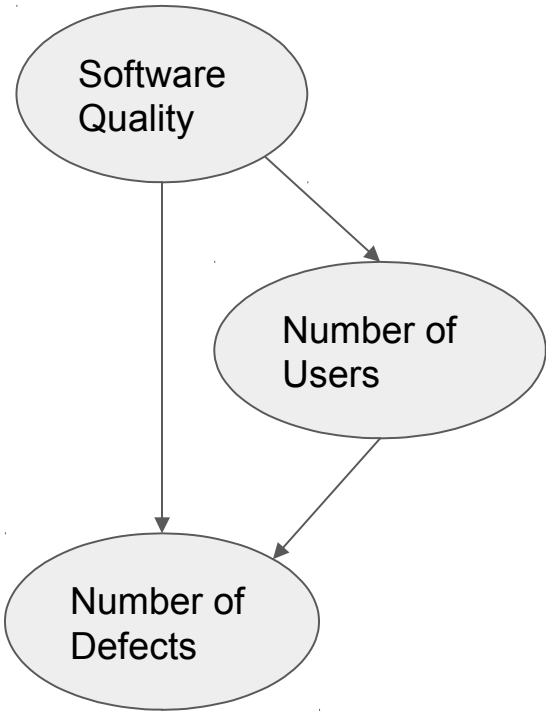
Eliminate most Predictable Predictor



Predictor with the highest adjR^2

Bayesian Networks in SE Context

- BN - a PGM - a DAG -
  - A set of variables and their conditional dependencies
- Two major applications:
  - Variable Selection
  - Joint probability distribution.
- Bayesian Networks for causal inference. (Pearl, J. (2009). *Causality*. Cambridge university press.)

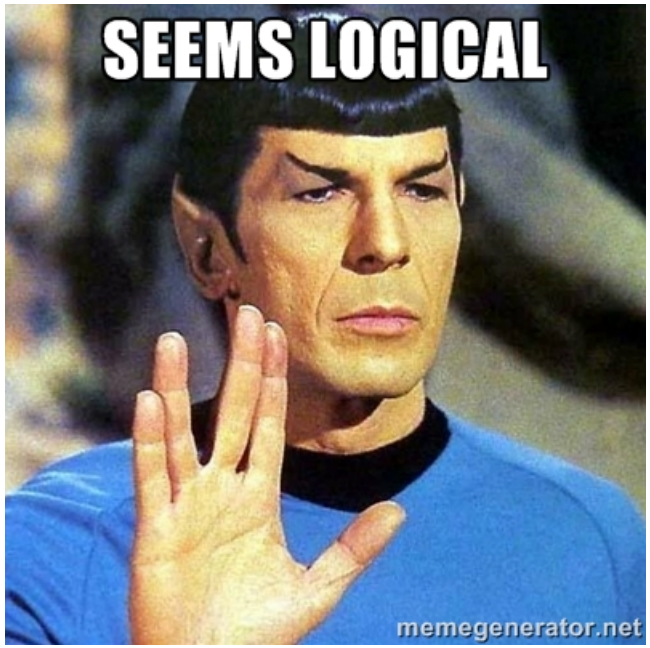


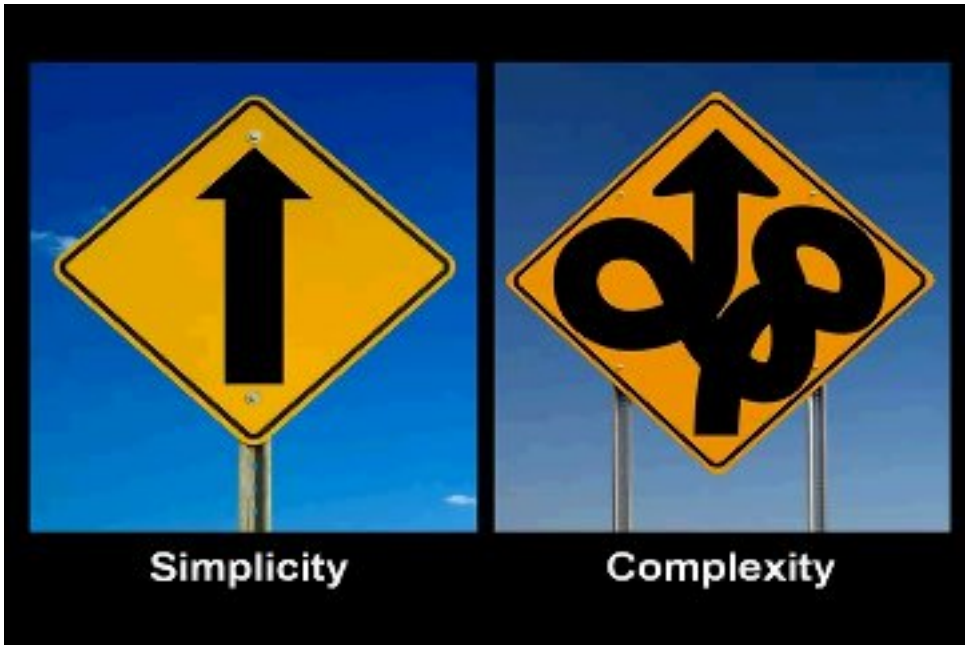
Using BNs to Adress Correlation

Choice of variables is often subjective

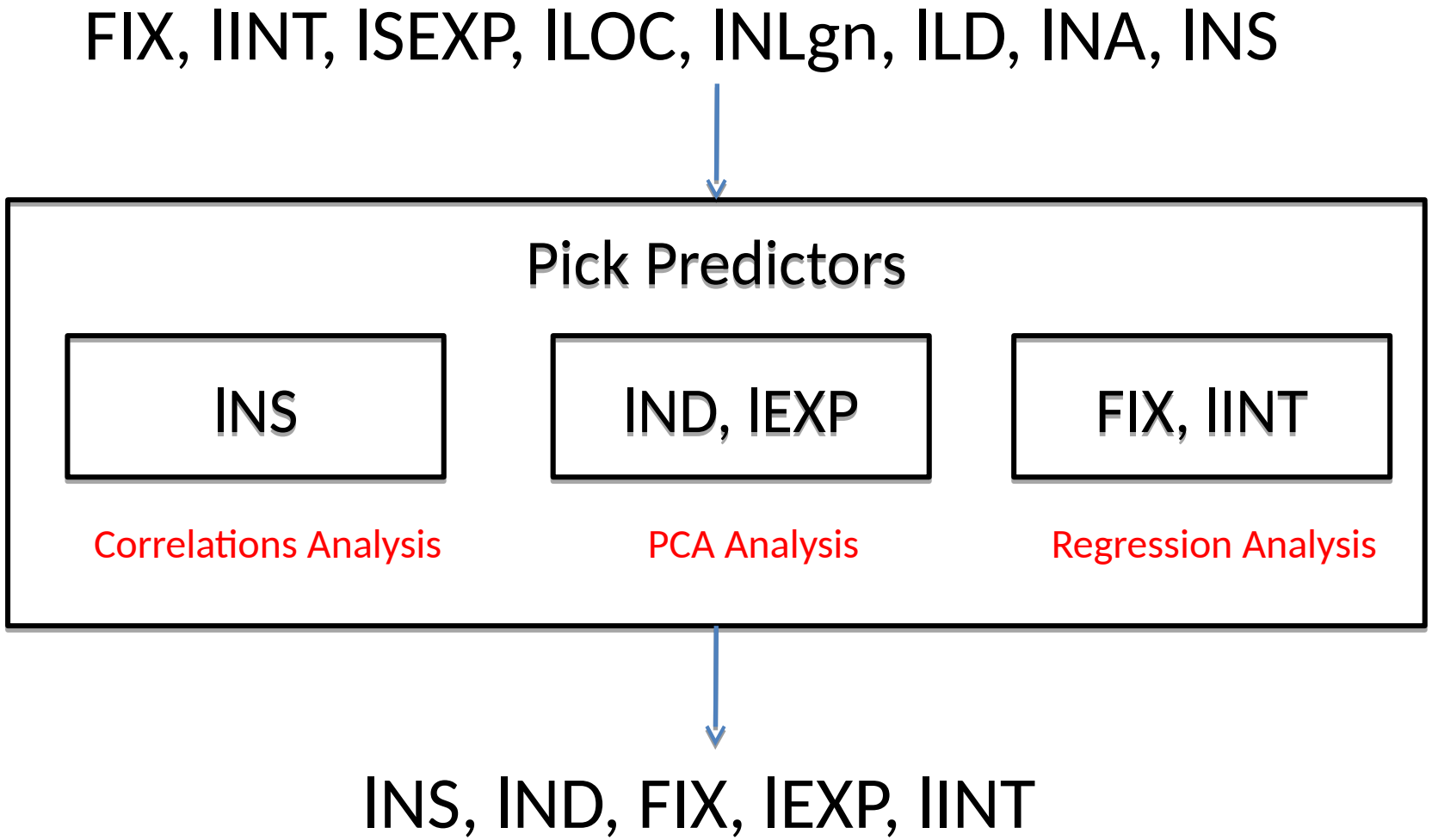
- Regression:
  - response depends on **all** predictors
- BN Structure Search:
  - No predictors/responses
  - Result: correlated variables are connected
  - Value: can identify additional relationships

How to Pick Predictors

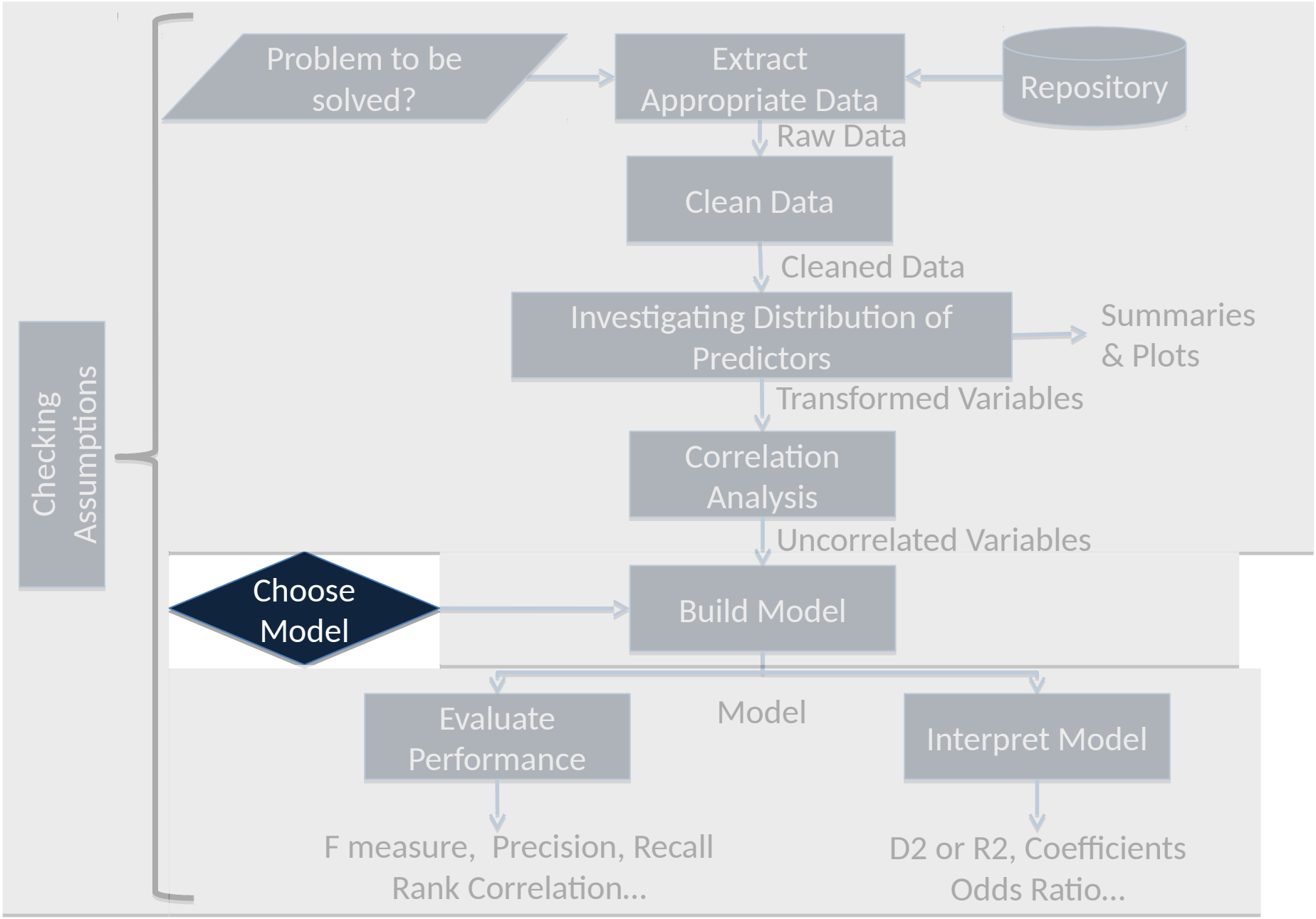




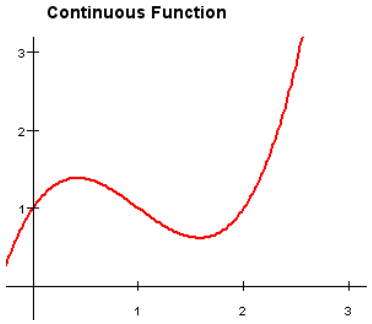
Remaining Variables < 0.85 (adjR^2)



Workflow


Problem to be solved?
Extract Appropriate Data
Repository
Raw Data
Clean Data
Cleaned Data
Investigating Distribution of Predictors
Summaries & Plots
Transformed Variables
Correlation Analysis
Uncorrelated Variables
Checking Assumptions
Choose Model
Build Model
Model
Evaluate Performance
Interpret Model
F measure, Precision, Recall
Rank Correlation...
D2 or R2, Coefficients
Odds Ratio...

Select appropriate model

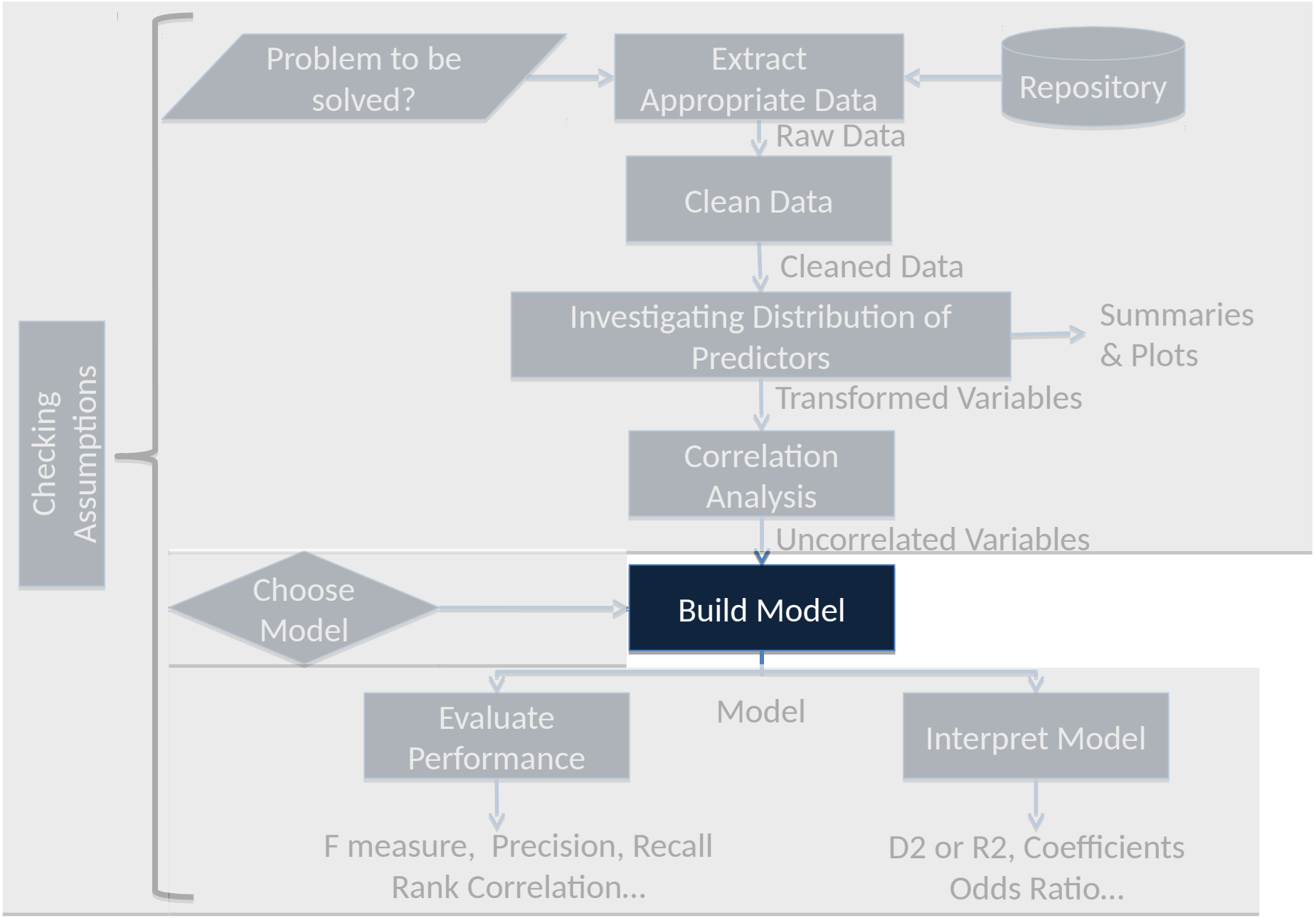


Dependent Variable

Linear Regression

Logistic Regression

Workflow

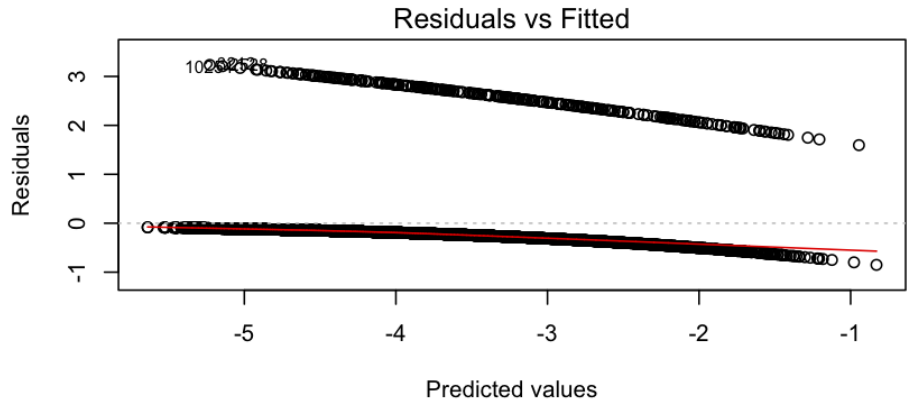

Problem to be solved?
Extract Appropriate Data
Repository
Raw Data
Clean Data
Cleaned Data
Investigating Distribution of Predictors
Summaries & Plots
Transformed Variables
Correlation Analysis
Uncorrelated Variables
Checking Assumptions
Choose Model
Build Model
Model
Evaluate Performance
Interpret Model
F measure, Precision, Recall
Rank Correlation...
D2 or R2, Coefficients
Odds Ratio...

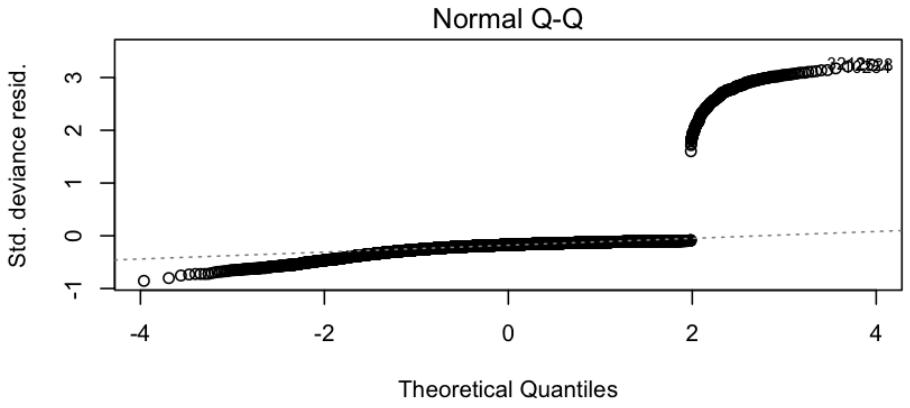
summary(mod)

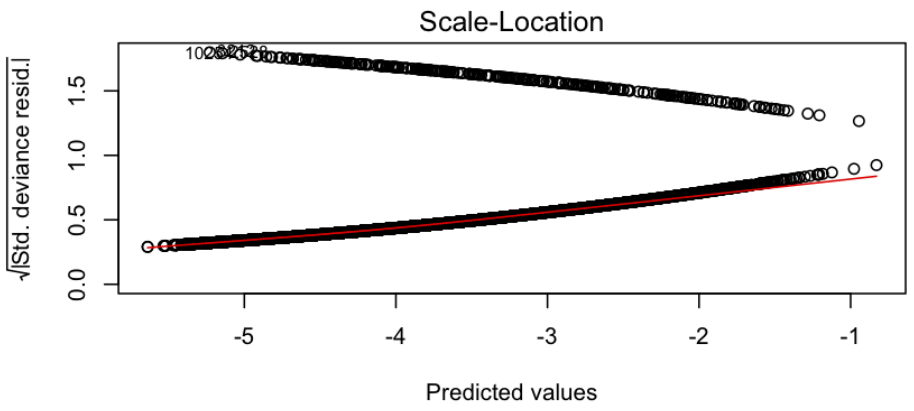
```
Call:
glm(formula = isBad ~ lNS + lLA + FIX + lLOC + lINT + lEXP, family = binomial,
    data = dataFit)

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

    Null deviance: 3011.5  on 13480  degrees of freedom
Residual deviance: 2739.0  on 13474  degrees of freedom
AIC: 2753
```

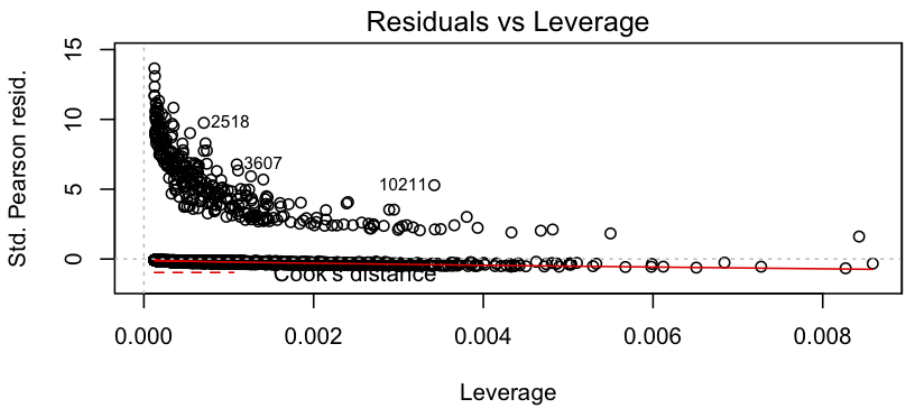
Dev Explained = (3011.5 – 2739.0) / 3011.5

Modeling Assumptions
Are there outliers?









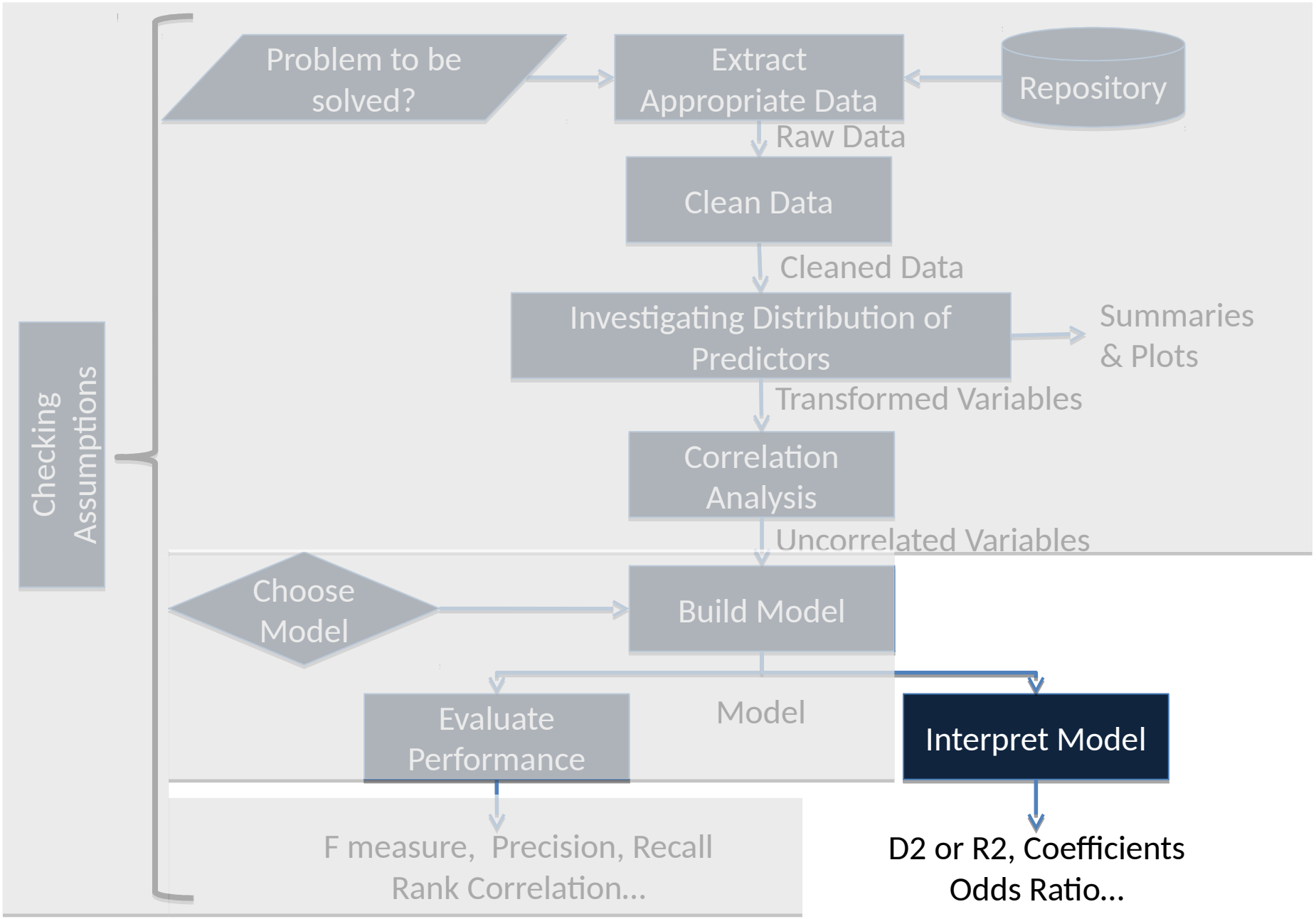
Plots not very useful in Logistic Regression

Modeling Assumptions
Are correlations still a problem?

```
vif(mod)
lNS   lND   FIX  lEXP  lINT
 2.3   3.1   1.2   1.0   1.9
#isBad ~ lNS+lND+FIX+lINT+lEXP+lLA+from
```

All Vif values < 4

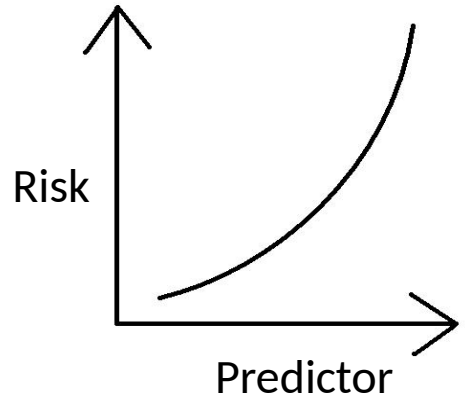
Workflow


Problem to be solved?
Extract Appropriate Data
Repository
Raw Data
Clean Data
Cleaned Data
Investigating Distribution of Predictors
Summaries & Plots
Transformed Variables
Correlation Analysis
Uncorrelated Variables
Checking Assumptions
Choose Model
Build Model
Model
Evaluate Performance
Interpret Model
F measure, Precision, Recall
Rank Correlation...
D2 or R2, Coefficients
Odds Ratio...

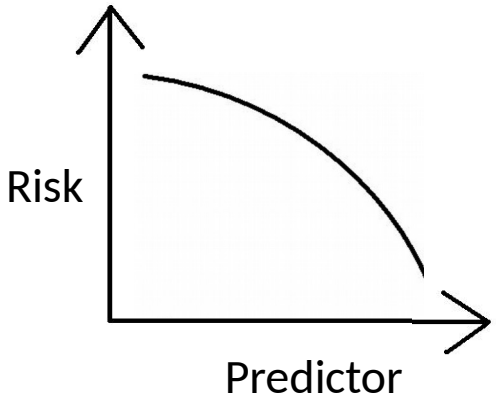
Model Interpretation

```
Coefficients:
             Estimate Std. Error
(Intercept) -4.48210     0.28367
lNS          0.31994     0.11021
lND          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

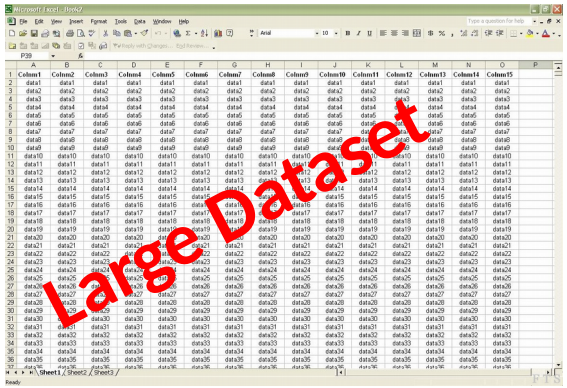
```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -4.48210    0.28367  -15.80  < 2e-16 ***
lNS          0.31994    0.11021    2.90   0.0037 **
lND          0.36503    0.05354    6.82  9.2e-12 ***
FIX          0.52993    0.12873    4.12  3.8e-05 ***
lEXP        -0.10508    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 Pr(>|z|) < 0.05

Then the effect of the predictors are statistically significant.

Effect of Predictors on Response Variables

p-values



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)

Hypothetical File
with median values

Effect of Predictors on Response Variables

↑ INS by 1 unit

Then calculate risk of faulty patch

Measure the ↑ in risk.

Using Anova Analysis

Effect of Predictors on Response Variables

```
> anova(mod);
Analysis of Deviance Table

Model: binomial, link: logit

Response: isBad

Terms added sequentially (first to last)


     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 ≠ 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 adjR^2

Deviance explained in our Data

Good for logistic regression and 2% average probability

Low values of adjR^2 => model only partly reflects mechanism generating the data

Extract Raw Data Required for the Problem!

[SVN] svn log -v $--$non-interactive PRJ

[GIT] git --git-dir=PRJ log $--$numstat -M -C $--$diff-filter=ACMR
$--$full-history \\
$--$pretty=tformat:"STARTOFTHECOMMIT\%n\%H;\%T;\%P;\%an;\%ae;\%at;\%cn;\%ce;\%ct;\%s"