what is data science

1) hacking skills

2) hacking skills and substantiative expertise software engineer

3) traditional research - hypothesis testing

4 V's of Big Data

1) volume

2) veracity

3) velocity

4) value

- wide data - k > n - ols fails

note: be careful while concluding causality.

- genome data, is an example of wide data.

- not all parameters are associated with the dependent variable.

4) adversarial machine leaarning

project information

1) identify the problem, insight about the data.

2) define the deliverable.

- actions on the deliverable. ( number of robberies for the chief to allocate police officers)

3) define success of the deliverables.

- prevent robberies , look for similar data ( 10% for the cincinnati)

4) data resources and limitations.

-privacy - health care data

- technical limitations -

- data understanding,

- form team

5) create plan

6)

- feasibility test of project.

------------------------data preparation

- data security clearance

1) getting acces to the data.

2) data summary.

3) summarizing data

4) groupin data -age

visualization.

----------deployment

1) where it is an api, map

----prediction vs causality

feature extraction

1) features are combinations of raw variables.

raw variables in columns.

-raw data - make a copy of the raw data.

1) for cpd, the data needs to be checkpointed at each step. so that data is not modified.

-----------regulaizations

1) feature to be included or weighted in the final model.

2) pruning -

3) shrinkage -

4) overfitting

5) cross validation

- out of sample performance of the model.

- train and test split for the time series dataset.

programming language

1) R - Bell labs

- interpreted language.

- dynamical language - variable type need not be mentioned.

final research project.

1) git - version control

- versions of the project.

--------------R

1) <-

- assignment operator.

- c operator is used for combining vectors

- : is useful for the range.

- matrix function is useful for creating matrix structure of a vector.