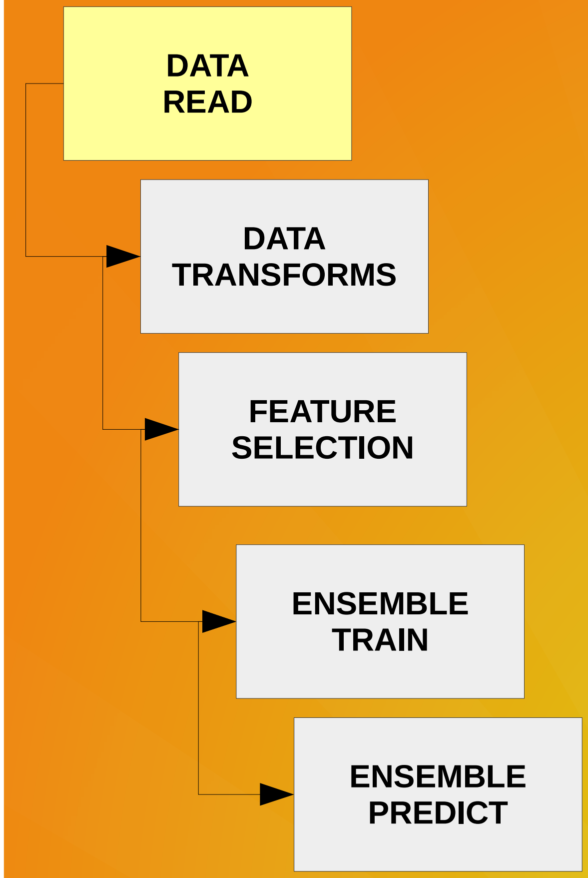
A Customizable Machine Learning Pipeline

Nelson R. Manohar

-

What is a Pipeline?



- A Data Flow
  - IN: data rows (e.g, Xt)
  - OUT: data rows (e.g, Yp)
- A Sequence of Steps
  - some required
  - some optional
- Customizable sequence
- Customizable steps
- Planned: each step can be done via a pub/sub compute-node

What is a Step?

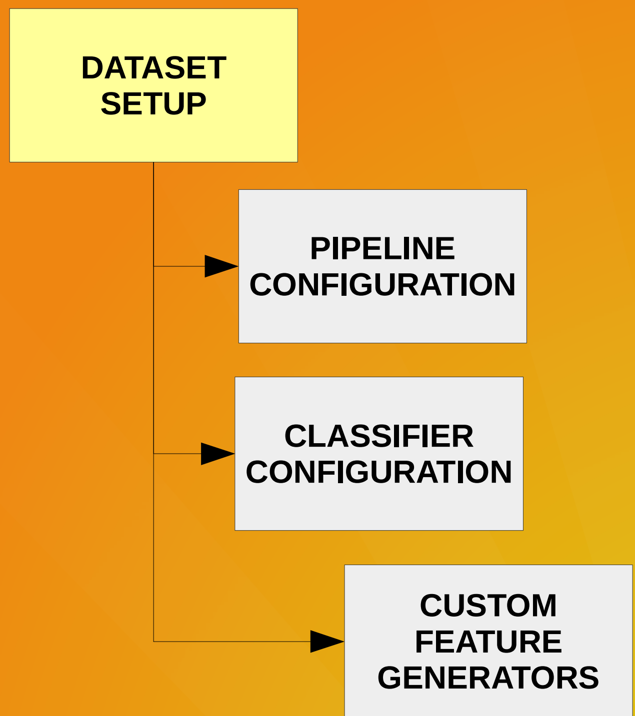
- data transformation function
  - data in: some features X, Y
  - data out: some features X', Y', findings
- some steps mandatory, some optional
- steps can be implemented:
  - currently, in python, scikit, pandas
  - planned, invoking steps implemented in bash, awk, R, java, MR

Predictive Pipeline

- Feat. Engineering
  - Overfitting Reduction
  - Feature Generation
  - Feature Selection
  - Feature Decorrelation
  - Feature Analytics

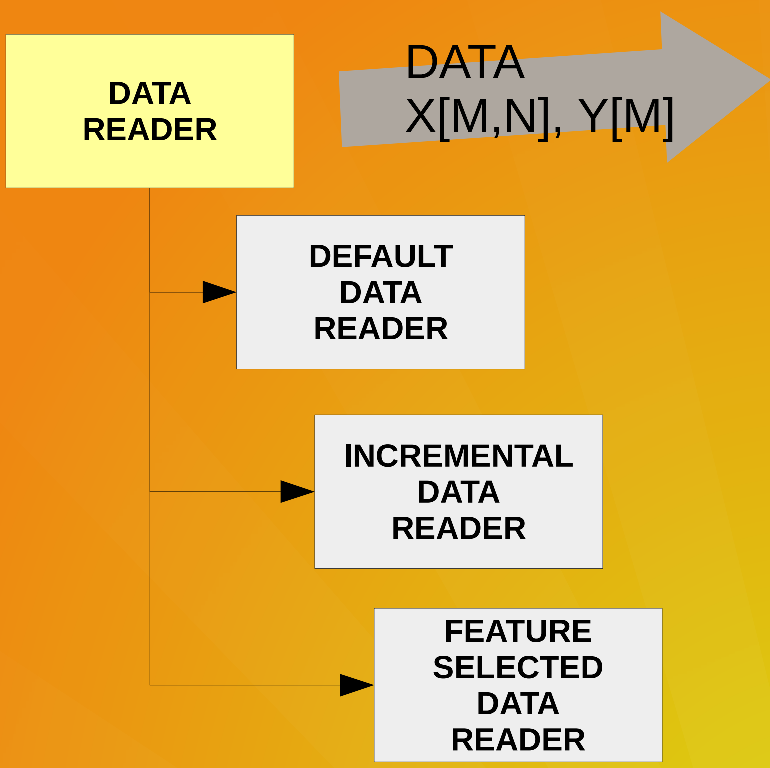
- Data Handling
  - Cleaning: NAs, Cuts, Factors
  - Scaling/Centering
  - Encoding
  - Partitioning
  - Augmentations
  - Reductions
  - Subsampling

Pipeline Setup



- User specified:
  - Dataset Spec
    - ID, X, Y, co-Y, Nas, sep
  - Pipeline Config
    - steps, options
  - Ensemble/Classifiers
    - clfs, options, cv
  - Feature Generator Hook
    - user-specified pre-processing for custom features

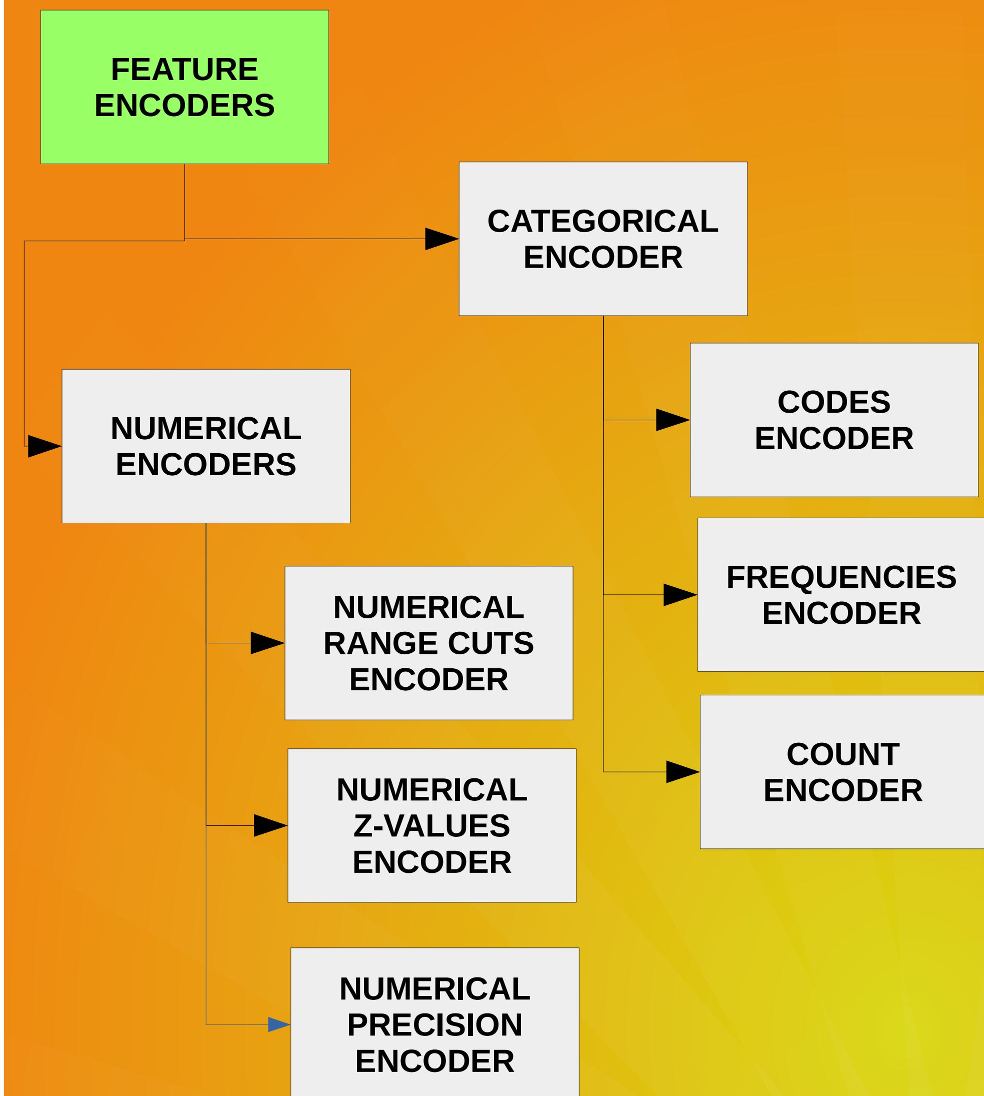
Dataset Load/Read



- User/System Choice:
  - Fit-all in memory (default)
  - Incremental reader (by features)
  - Preselected features readers
  - Chunk reader (in development)
  - Streaming reader (planned)

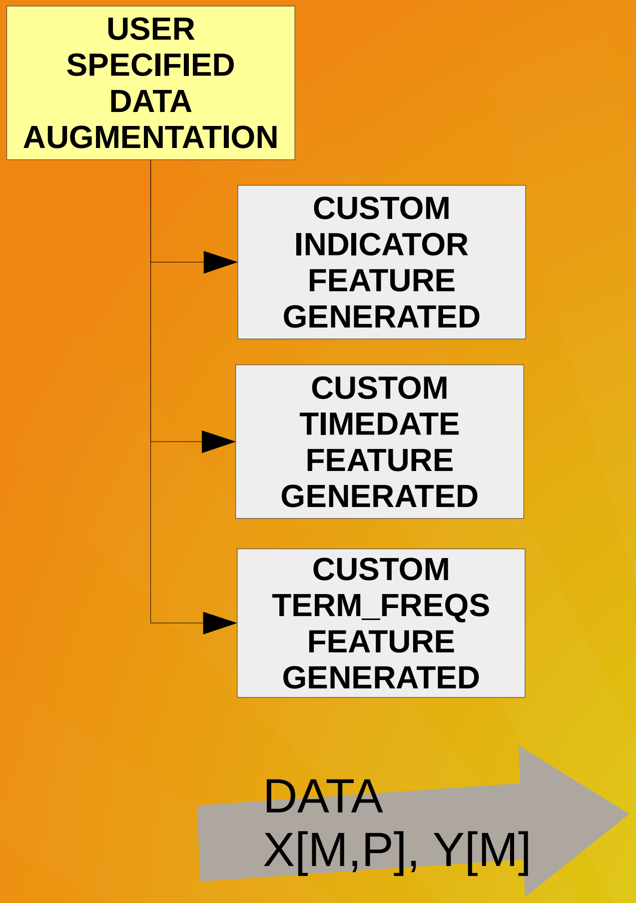
-

Data Encoding



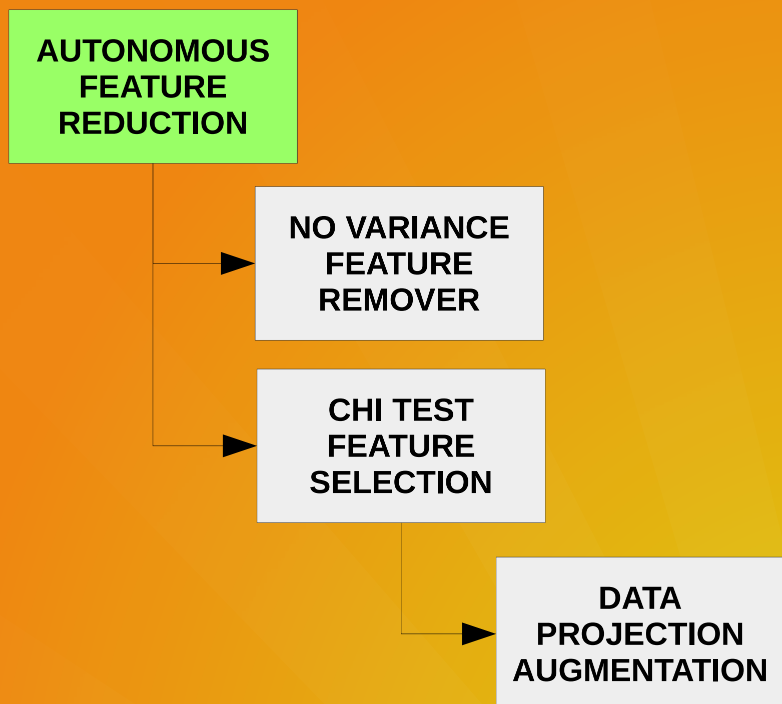
- Numerical encoders
  - Stabilize/condition numerical range
  - autonomously applied
- Categorical encoders,
  - statistical profiles of factors/categories
  - stabilize/condition numerical range

User-Gen'd Features



- Dataset augmentation hook allows
  - User-implemented features to be generated from existing dataset features
    - Examples: indicator variables, statistical profiles, td-idf of name fields, timedate parsing, additions, conditionals, etc.
  - Features subsequently stabilized/conditioned

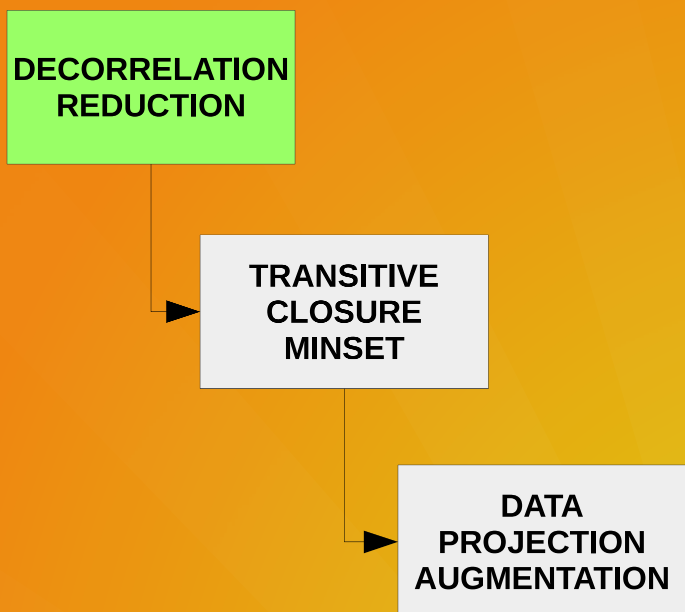
Feature Reduction



- Chisq Feature Significance
  - features numerically conditioned/stabilized
  - selects: All, top K, statistically significant, or any above threshold N
  - returns: ranked features

-

Feature Decorrelation



Transitive closure of features

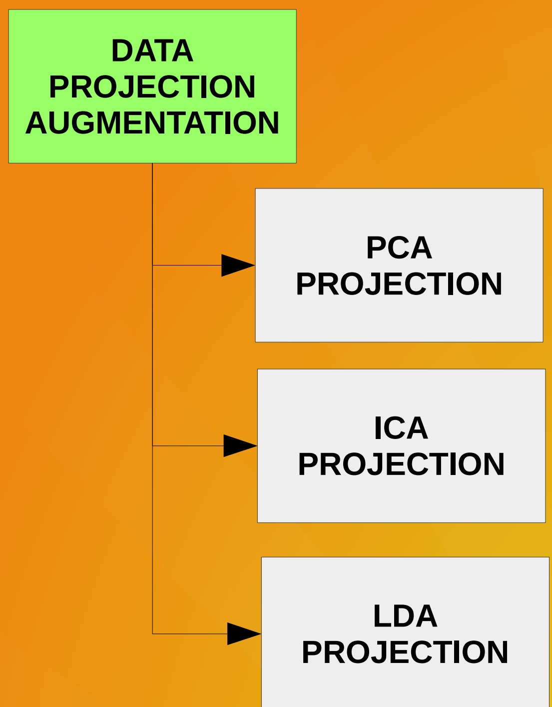
- keeps best rank feature given rank ordering by feature importance
- autonomous exploration of extent and degree
- for discarded features can
  - generate replacement projection
  - decorrelated replacement features

Output

```
--------------------------------------------------------------------------------
using subsample[wrt_vars] as is

FS:**    10    0                       PETBREED_POPULARITY   846    20903197.7787                 0.0
FS:**    11    1                      COATCOLOR_POPULARITY    72    3824192.20053                 0.0
FS:**     9    2                             AGE_IN_MONTHS    32    183468.329819                 0.0
FS:**    13    3                    COATPATTERN_POPULARITY    54    153410.786041                 0.0
FS:**    12    4                        PETNAME_POPULARITY    56     48118.986562                 0.0
FS:**    17    5                                 IS_INTACT     2    1156.17266483 2.82083934096e-244
FS:**    20    6                               IS_DOMESTIC     2    865.089920414 1.91015550786e-181
FS:**    18    7                                  IS_OLDER     2    604.943765259  2.0246443573e-125
FS:**    21    8                              IS_SHORTHAIR     2    508.493504916 1.05873638998e-104
FS:**    16    9                                  IS_TABBY     2    241.847898625   9.1960222577e-48
FS:**     2   10                           SexuponOutcome     6    92.9331863318  1.17872888383e-16
FS:**     7   11                                     MONTH    12    86.4379315863  2.45240381582e-15
FS:**     4   12                                     Breed   102    72.7992217887  1.35908585454e-12
FS:**     8   13                                      HOUR    31    48.3314261722  8.53782520074e-08
FS:**    19   14                                 IS_SUMMER     2    40.2944590438  2.82340684881e-06
FS:**    15   15                                  IS_MIXED     2    40.0748572092  3.10247750307e-06
FS:**     1   16                                AnimalType     2    32.2192501347   8.5084135566e-05
FS:**     6   17                                      YEAR     4    26.5811527577  0.000834834902178
FS:**    14   18                                 IS_FEMALE     2    23.1078365932   0.00322852061381
FS:**     3   19                            AgeuponOutcome    46    3.97366539031     0.859491606084
FS:**     5   20                                     Color   109   0.254900598553     0.999990069613
FS:**     0   21                                  DateTime    28  0.0131808229234     0.999999999922
--------------------------------------------------------------------------------
```

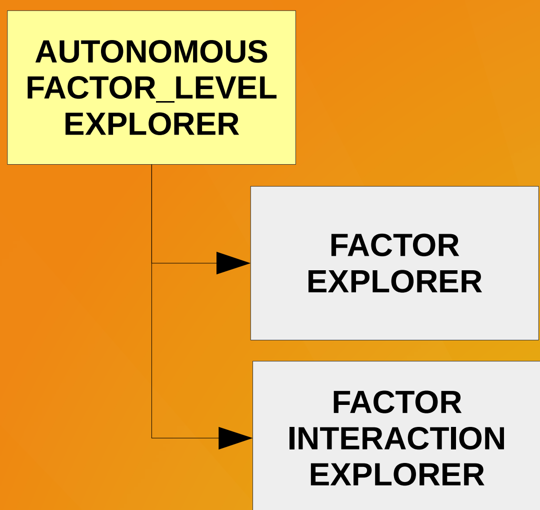
```
--------------------------------------------------------------------------------
CORRELATED COLUMNS SETS {0: [46], 1: [54], 3: [62], 4: [63, 76], 5: [49, 64], 6: [65], 7: [66], 8: [67, 90], 9: [84], 10: [45, 92], 12: [47, 98], 13:
[50, 101], 14: [51, 102], 15: [52, 103, 130], 16: [104], 17: [82, 125], 19: [55, 132], 20: [133], 21: [57, 134], 22: [58, 135], 23: [59], 24: [60, 136
, 164], 25: [137], 26: [83, 158], 27: [171], 29: [69, 166, 181], 30: [106, 182], 31: [187, 196], 32: [71, 172], 33: [85, 192], 34: [72, 167, 198, 202]
, 35: [73, 168, 205], 36: [74, 169], 37: [75, 170, 207, 211], 38: [86, 212], 39: [213, 219], 40: [87, 127, 217], 41: [78, 221], 42: [226, 230], 43: [8
8, 228], 44: [89, 232], 77: [173], 79: [174], 80: [175], 81: [176], 99: [122], 100: [49], 123: [173]}
--------------------------------------------------------------------------------
K: XTERM_AnimalType_BY_SexuponOutcome D:[ XTERM_AnimalType_BY_SexuponOutcome_BY_YEAR ];
K: XTERM_AnimalType_BY_MONTH D:[ XTERM_AnimalType_BY_MONTH_BY_YEAR ];
K: XTERM_AnimalType_BY_HOUR D:[ XTERM_AnimalType_BY_HOUR_BY_YEAR ];
K: XTERM_AnimalType_BY_PETBREED_POPULARITY D:[ XTERM_AnimalType_BY_PETBREED_POPULARITY_BY_YEAR XTERM_AnimalType_BY_IS_MIXED_BY_PETBREED_POPULARITY ];

K: XTERM_AnimalType_BY_PETNAME_POPULARITY D:[ XTERM_AnimalType_BY_PETNAME_POPULARITY_BY_SexuponOutcome XTERM_AnimalType_BY_PETNAME_POPULARITY_BY_YEAR
];
K: XTERM_AnimalType_BY_COATPATTERN_POPULARITY D:[ XTERM_AnimalType_BY_COATPATTERN_POPULARITY_BY_YEAR ];
K: XTERM_AnimalType_BY_COATCOLOR_POPULARITY D:[ XTERM_AnimalType_BY_COATCOLOR_POPULARITY_BY_YEAR ];
K: XTERM_AGE_IN_MONTHS_BY_AnimalType D:[ XTERM_AGE_IN_MONTHS_BY_AnimalType_BY_YEAR XTERM_AGE_IN_MONTHS_BY_AnimalType_BY_IS_OLDER ];
K: XTERM_AnimalType_BY_IS_OLDER D:[ XTERM_AnimalType_BY_IS_OLDER_BY_YEAR ];
K: XTERM_MONTH_BY_SexuponOutcome D:[ XTERM_AnimalType_BY_MONTH_BY_SexuponOutcome XTERM_MONTH_BY_SexuponOutcome_BY_YEAR ];
K: XTERM_HOUR_BY_SexuponOutcome D:[ XTERM_AnimalType_BY_HOUR_BY_SexuponOutcome XTERM_HOUR_BY_SexuponOutcome_BY_YEAR ];
```

FG: Projections



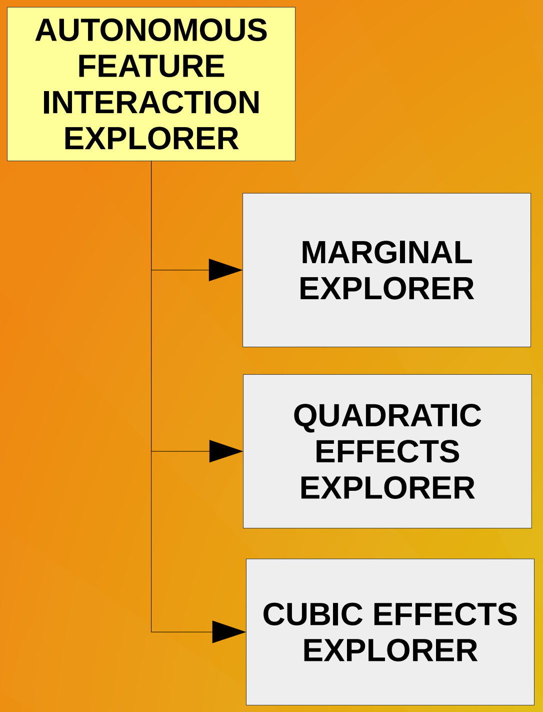
- Feature Projections generate features from features
  - Linear Discriminant of X features wrt Y
  - Independent Signal Components of X features wrt Y
  - Principal components of X features
  - Attempts to generate a pre-specified 0, 1, or upto N other features from N features
  - Features are automatically stabilized/condition

FG: Factor Levels



- Autonomously exploration
- Identifies factors and factor interactions with high levels of significance wrt to target variable Y
- Combinatorial exploration pruned via
  - random subsampling of factors,
  - pre-validation heuristics
  - feature selection ranking
  - timeout

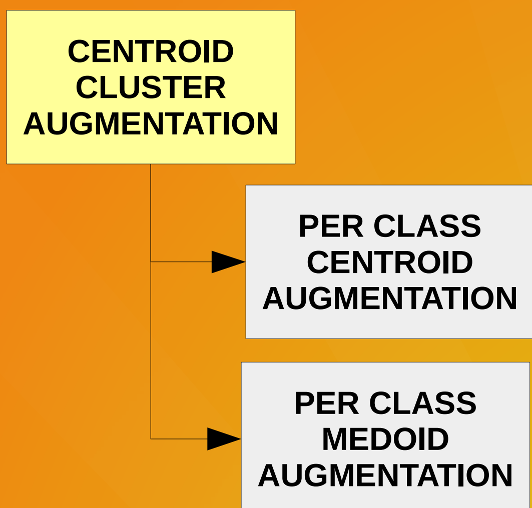
FG: Interaction Effects



- Marginals
  - Explores feature interactions
  - conditional statistical profiles (groupby)
  - Autonomously explored
  - Autonomously selected
  - Selected wrt target variable Y

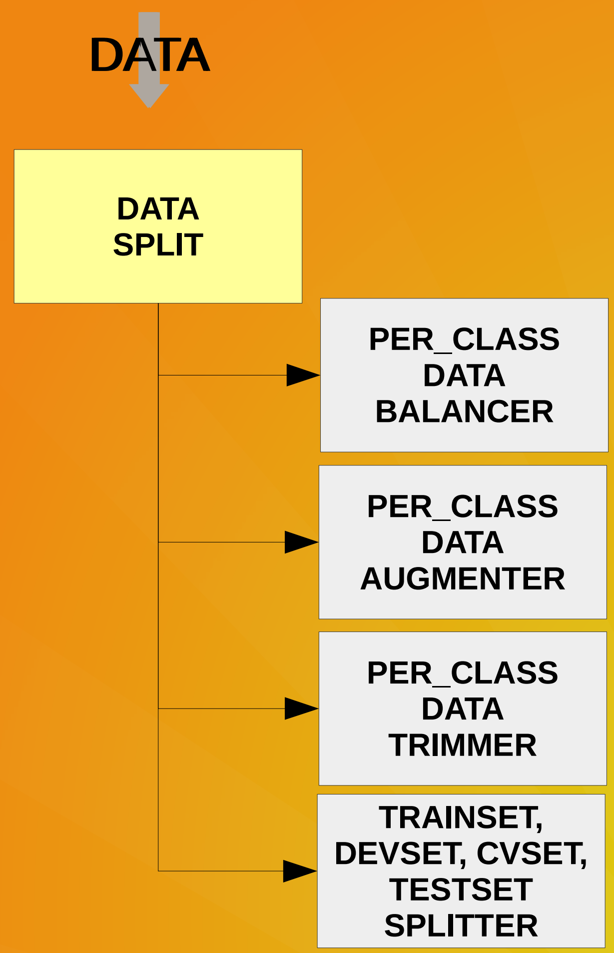
-

Data Augm/Reduction: Centroids



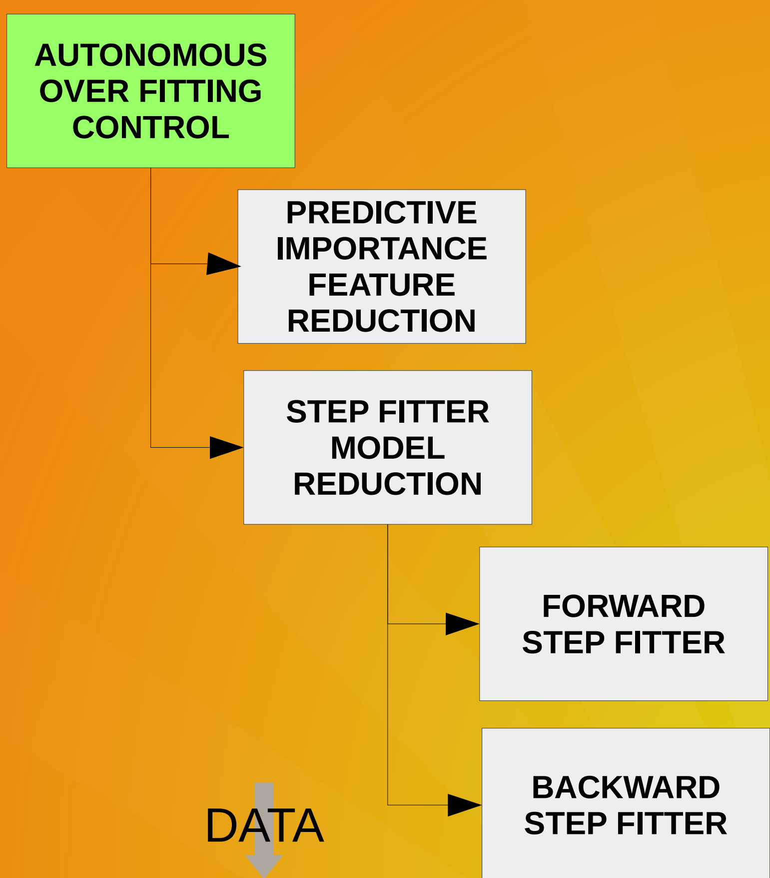
- Per-class random subsampling
  - Representative sampled
  - centroid (numerical)
  - medoid (categorical) for subsamples
- Used to
  - artificially augment dataset or
  - reduce dataset by deletion of the samples assoc. with a centroid

Dataset Partitioning



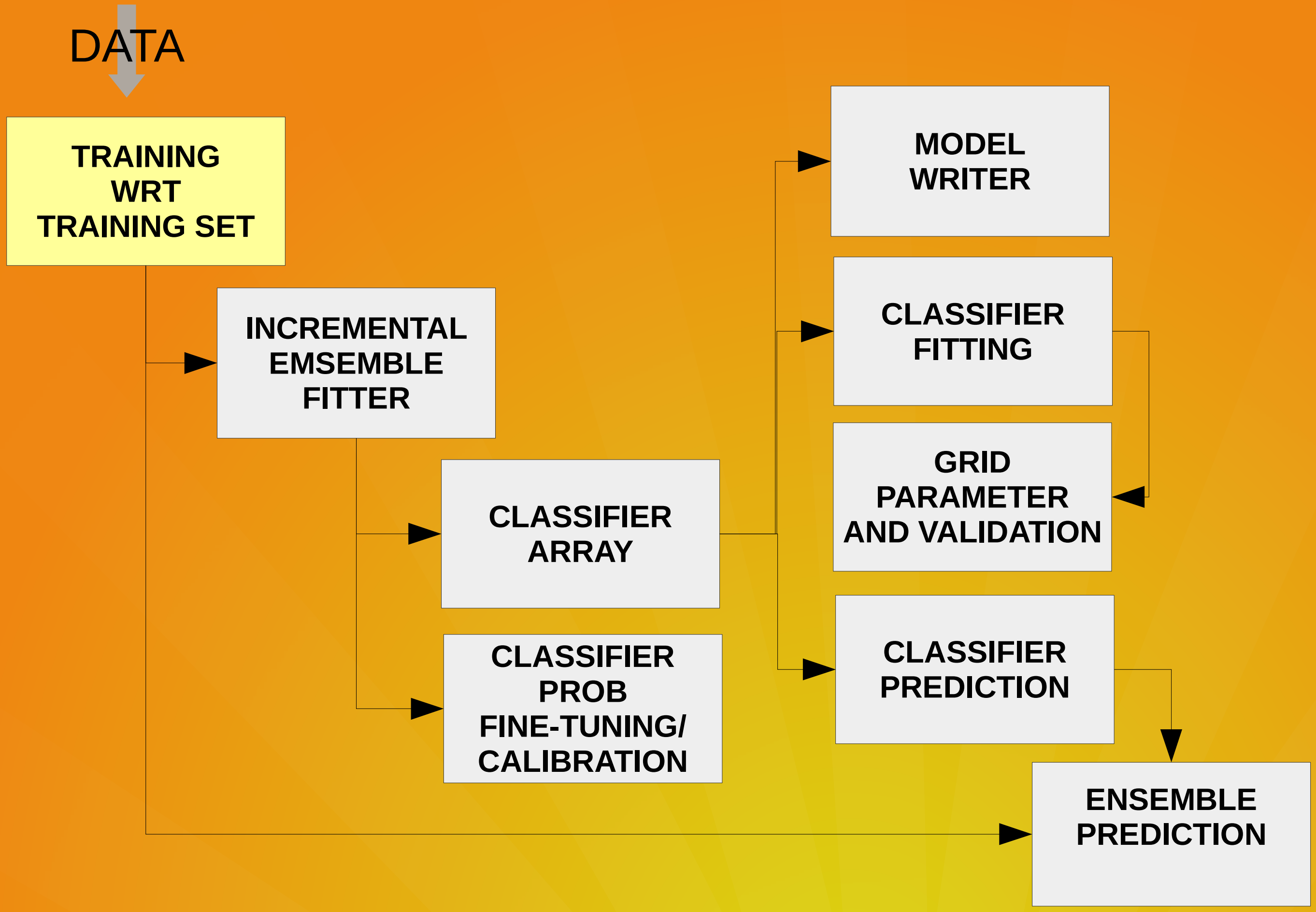
- Dataset partitioned into
  - Training set, dev set, cv set, and test set
  - Balancing of class sizes done wrt user-specified policies (percentage, augmentation, subsampling, etc)
  - Dev set used to fine-tune ensemble classifier parameters

Overfitting Control: Step Fitter



- Autonomous exploration via step-fitter of ranked features by importance
  - Forward stepper
  - Back stepper
  - Combo stepper
  - Maximum timeout
  - Warm-start **(fix code)** Gradient Boosting Classifie used to reduce training tim and reduce overfitting

Ensemble Training

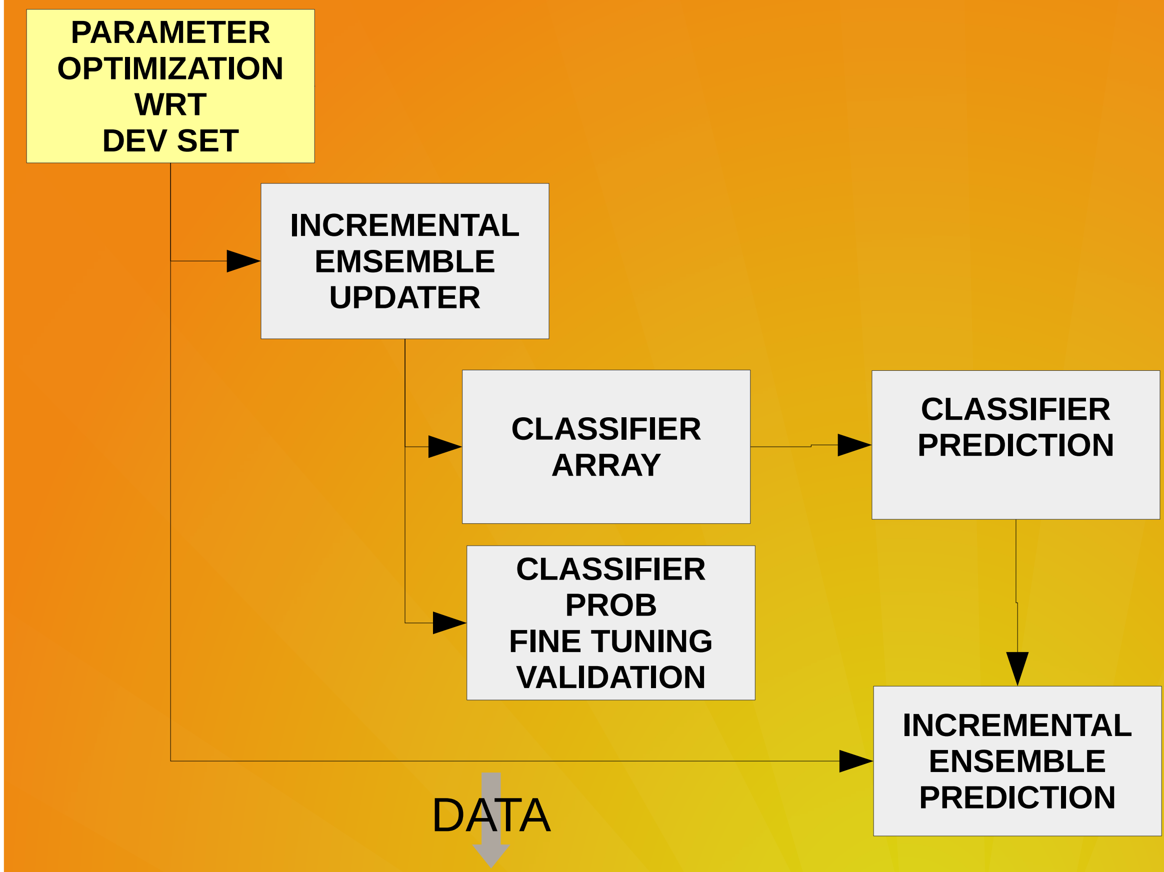


-

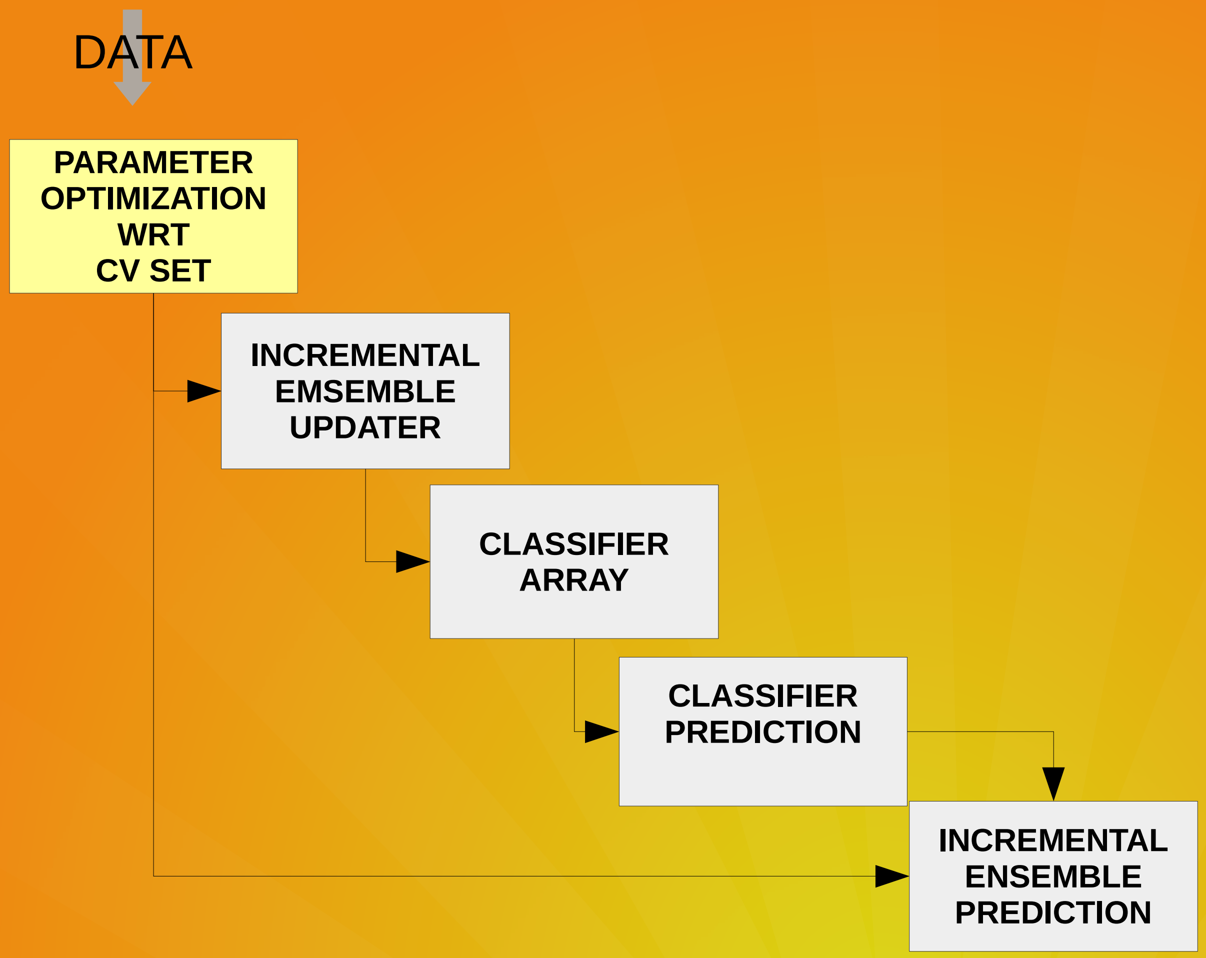
Ensembles

- Atop intra-classifier ensembles
  - Variants of same classifier allowed
  - Multiclass and Binary classifiers allowed
  - Ensemble classifiers allowed (such as Random Trees, Bagging, Boosting, etc)
  - Generative and discriminative classifiers can be mixed
- Weighted Voting ensemble (deprecated)
- Weighted Probability ensemble
  - Computes weighted average of selected predictors
  - Balances/conditioning classifier probabilities to 0.5
- Trained Probability Predictor ensemble
  - Trains meta classifier using predictions of ensemble classifiers
  - Predicts using meta classifier

Ensemble DevSet Optimization

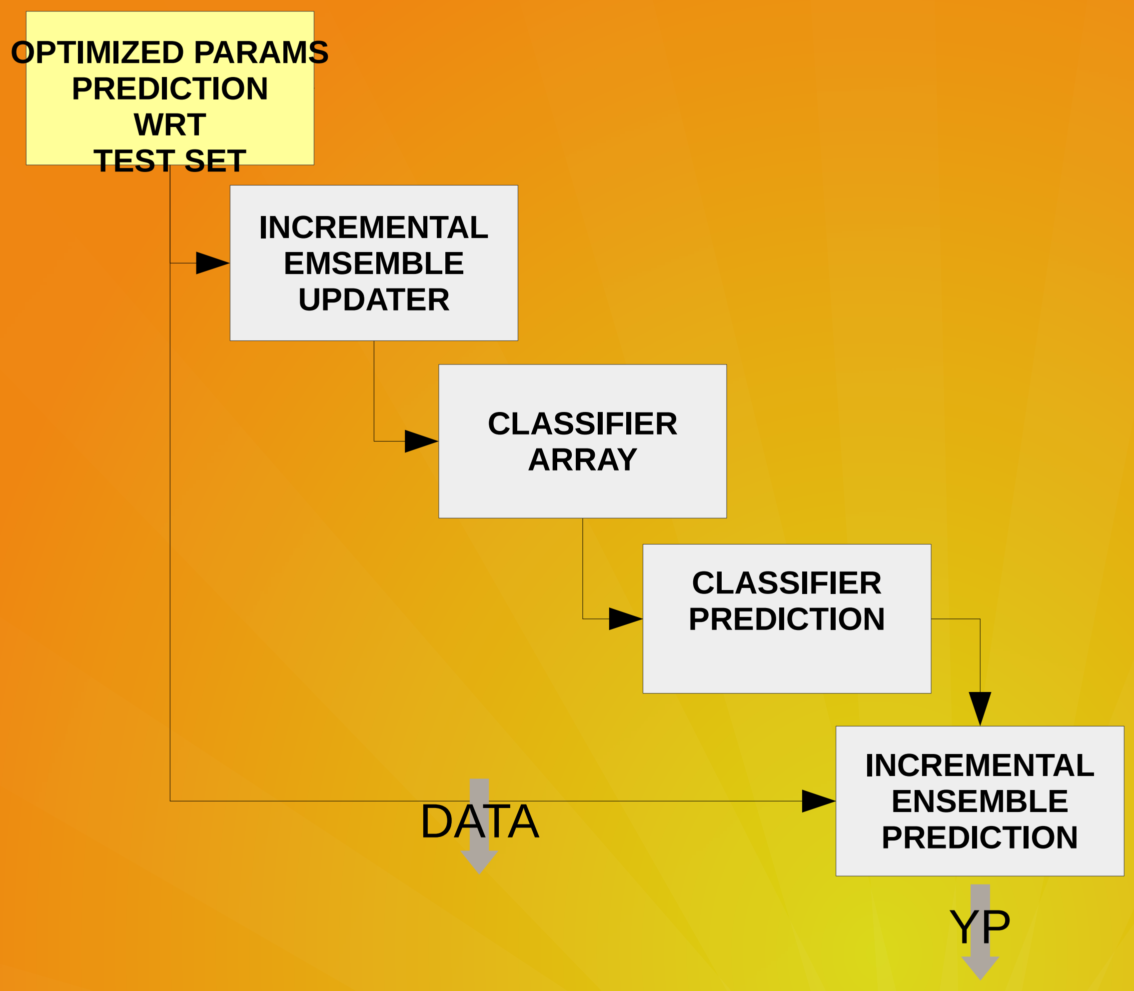


Ensemble Cross Validation



-

Ensemble: TestSet Prediction



Classification Performance

	Classification	MxN	Classifiers	LogLoss Accuracy
Santander	**Binary**	**130Kx300**	**RF, DT, GB**	
BNP	**Binary**	**150Kx130**	**ET, GB**	
Titanic	**Binary**	**1Kx11**		
Shelter	Multiclass			
S.F. Crime				
Avito				
Digits				

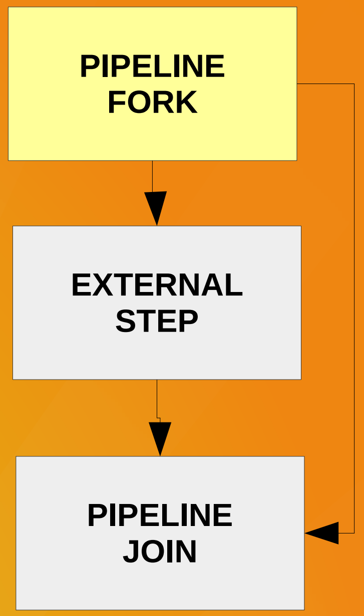
-

Future Work

- Configuration
  - Graphical “jobflow” style pipeline specification
- MR/Cluster
  - Steps to map reduce
  - Steps to n-cores or n-nodes
- Dashboard:
  - Per step: pipeline stats
  - Per step: data quality
  - Per step: predictive increase

Non-Native Step

- Via Pipeline Fork & Join
  - OUT: X, Y, f(), params
  - Bash fork f() process
  - Bash waits f() completion
  - F() generates X*, Y*, res
  - Pipeline reads X*, Y*, res
  - Pipeline resumes execution



MR-1

- MR implementation of certain pipeline steps
  - Already envisioned for subsequent implementation
  - Based on both simple (training and offline predictions) as well as streaming MR (classifier updates and production/online predictions)
- Clustering:
  - Currently, using representative-sample-KNN (centroids from subsamples)
  - Later:
    - chunks sent to reducers which generate local representative samples/local centroids
    - then combiners produce clustering of local centroids to generate global centroids

MR-2

- Encoders
  - Currently, encoder transforms learned on either training set or global dataset
  - Later,
    - Preliminary job selected chunks based on some criteria such as timestamp, id ordering, random subsampling
    - Chunks sent to first MR job reducers produce local/chunk statistical profiles for feature
    - MR combiners take local/chunk profiles and learn/yield global dataset transform
    - Second MR job applies learned transform to dataset chunks

MR-3

- Feature Selection/Decorrelation:
  - Currently, learned from subsampled slice of dataset
    - Planned, learned from multiple subsampled dataset slices
  - Later
    - Dataset chunks to MR reducers which produce chunk feature importances
    - combiners take chunk feature importances and learn/yield global/dataset feature importances

MR-4

- Classifiers:
  - Currently, ensemble, boosting, voting, and bagging classifiers (RT, DT, GB, etc)
    - trained using Random Patches and/or Random Spaces
  - Later,
    - First MR job produces local classifier
    - Combiners generate/grow ensemble classifier into pseudo-global classifier (see above Random Subspaces)
    - Second MR job's reducer job applies learner global ensemble classifier to data
    - Second MR job's combiner produces ordered predictions and stats