KDD cleaning integrtn selectn transfrmn mining pattern evaln knowldg presntn **Data Types** db, dw, trnscnl **Pattern Types** concept dscrpn charctrzn discrmn, freq patterns – assoc corrlns, classfcn regressn, clstr analys, outlier anls **Intrstng** Objctv support, confdnc, Subj Unexpectd, actionbl **Tech** Stats, m/c lrng, IR, db dw **Challengs** Uintrcn, methodlgy, db types, society, effcncy scalablty

Att typs Numeric(*interval, ratio scaled*) binary ordinl nominal discrete, continuous **statscl desc** Cntrl tendncy-*mean mdn mode*, disprsn-*range,quartiles,boxplots,variance stdev* **graphic displys** qplot,qqplot,histogram,scatterplot **data visulzn** pixl orientd, geom proj-*scatter plot matrix,parallel coordinates* icon based-*chernoff stick figurs* hierarchial-*wrlds within w,treemaps* complx-*tag clouds* **sim n dissim** – data matrix, dissim matrix. Nominal, Bin-*symm, asymm*,Numeric-*euclidn mnhattn minkowski supremum (chebyshev)* ,ordinal , mixd ,sparse-*cosine similarity*

Data cleaning- missing vals *ignore, fill manually, mean/medn , class m/m, const, most probbl*, noisy data – *binning, regression, outlier analysis,* cleaning process **Integrtn** Entity identificn prob, redundancy, correltn, covar **–** *nominal X^2 numeric- pearsons correln, covariance coeff* tuple duplicn, data val conflict **Data reductn** Dimensionality Reductn-*wavelet, PCA, sttribute subset selcn(fwd, backward, combintn, dec tree)* Numerosity Reductn-*Parametric(regress, log linear*), *Non parametric(hist – eq width, eq freq, clustrng, sampling- SRSWOR SRSWR clustr stratified, data cube aggrgn)*Transformn and discretizn – *smoothing, attri constructn, aggregtn, normalizn(min max, zscore, deciml scling), discretizn (clustr, dec tree, correln), concept hierar genertn nominal data (experts, portion, set of attribts, partial set)*

DW models- enterprise wide, data marts, virtual **Data cube n OLAP** star snowflake, opertns (*distri, algebraic, holistic*) olap oprns*(roll up drill dwn slice n dice pivot)***DW Des** – top dwn, b up, combind **Cube computn** Indexng(bitmap, join) server(olap, rolap, holap) **Att orientd inductn** Data Characterizn *(data focusing, att removal- higher concepts = other atts, cant generalize, att generalizn – apply genrlzn operator*), data comparisons *(data colltcn – 2 classes, att relevance analysis – remove irrelevant, synchro generlzn - bring atts in both classes to same level, result presentn)*

Cube materialzn full iceberg closd shell **Computatn methds** multi way array, BUC, start cubing, shell frags **Adv cube queries** sampling cubes(conf interval - boosting - intra cube expn, inter) ranking cubes **Multi Dimnl analysis** predctn cubes, multi feature, except based, discovery driven (selfexp, inexp, pathexp)

sup=P(AUB), conf=P(B|A). closed=no subset hs same sup.
Mining meth Apriori, pattern growth, vertical data format. close n max pruning strategies. **P eval methods**lift=P(AUB)/PA.PB all_conf=sup(AUB)/ max(supa, supb)= min(P(A|B),P(B|A)) , max_conf=max() , Kulc=avg.
cosine=root(.)

Roadmap: Pattern Diversty-*Basic(freq, closed, max, rare, neg) , abstrctn-(single,multi level), num dimn(single,multi),val type (bool,quanti), constraint(c-based, approx., compressd, near match, topk, redundancy aware top k)* , App type-*Features (freq, seq, structural) , app domain, data analysis usage(patt based classificn, p basd clustering, recommender sys)*
Multilevel, mdimn ML:uniform, reduced, grp based support MD:inter, intra, hybrid. **Quant attr**: discretize based on predef concept hier/ data distri. Datacube, clustering basd, stat theory
Rare n neg.