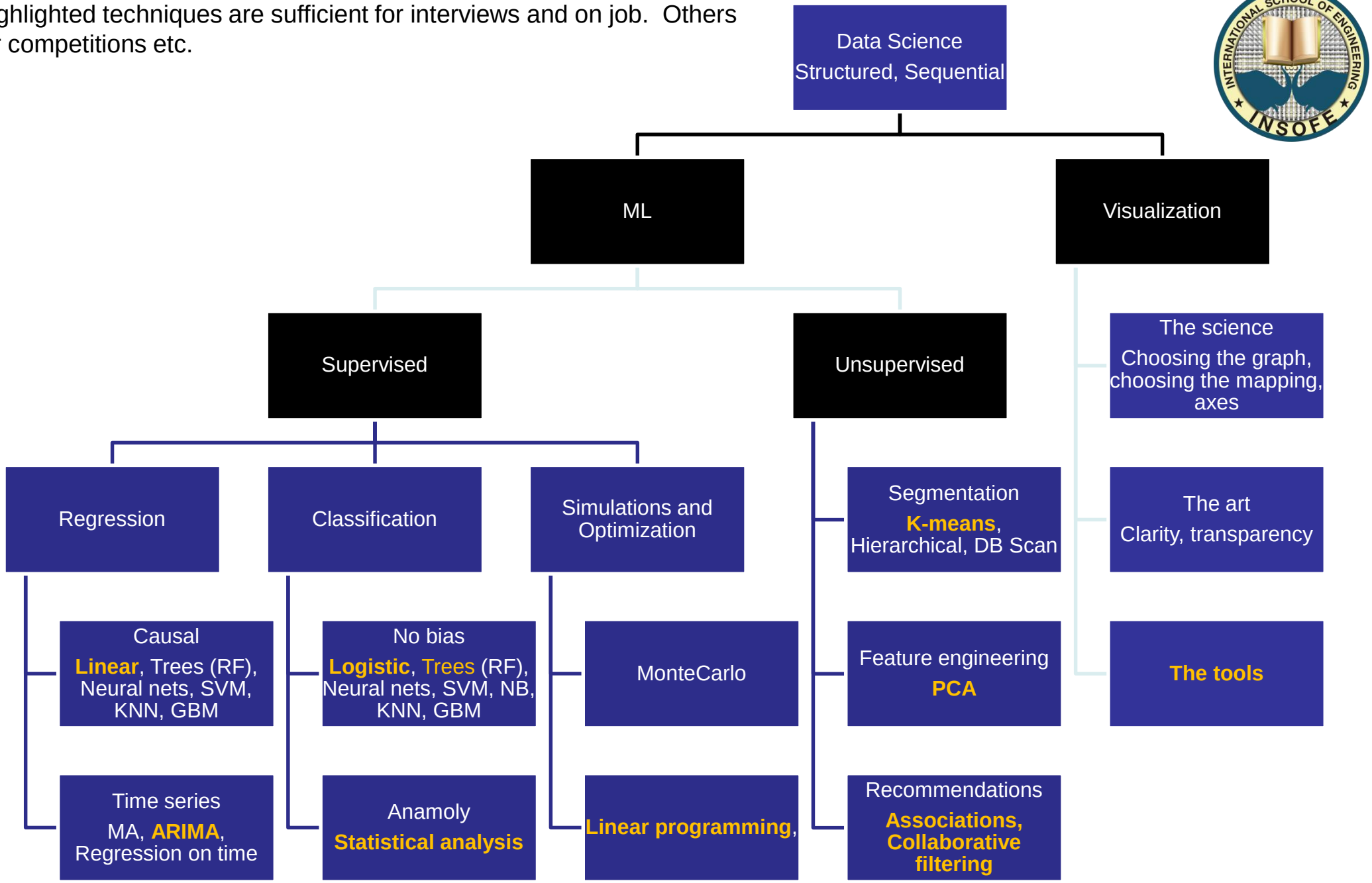
Inspire…Educate…Transform.

Architecting ML Solutions

Dr. Manish Gupta
Principal Applied Scientist, Microsoft
Adjunct Professor, IIIT Hyderabad

Highlighted techniques are sufficient for interviews and on job. Others for competitions etc.



The process

Understand

Business Problem

Current solution

What is the ROI

Analyze

Which bucket(s) the problem falls

What error metrics makes sense

What stories must be told (evolves)

Data

Define target (for supervised)

Derive relevant attributes through brainstorming hierarchically

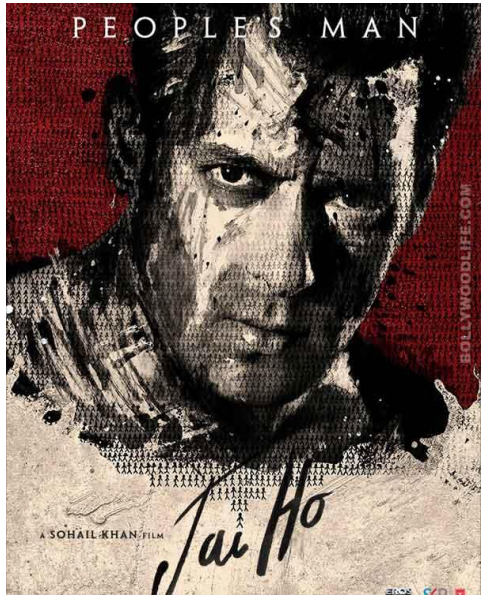
Get what is available, pre-process and visualize

Engineering

Anybody can build a model. You should build the most efficient and accurate

Engineer features (transform data, PCA it. Create smaller number of classes for categorical). Did you derive at least 3 additional attributes? Can you do better?

Keep experimenting with various hyper parameters; Regularize



Try at least 3 models on 3 types of data sets with 3 sets of hyper parameters (27) before you pick the final choice.

Validation

Design validation strategy.

Divide the data into train, test and validation.

Plot the error metric for all data for all models and pick the best

Gaining acceptance

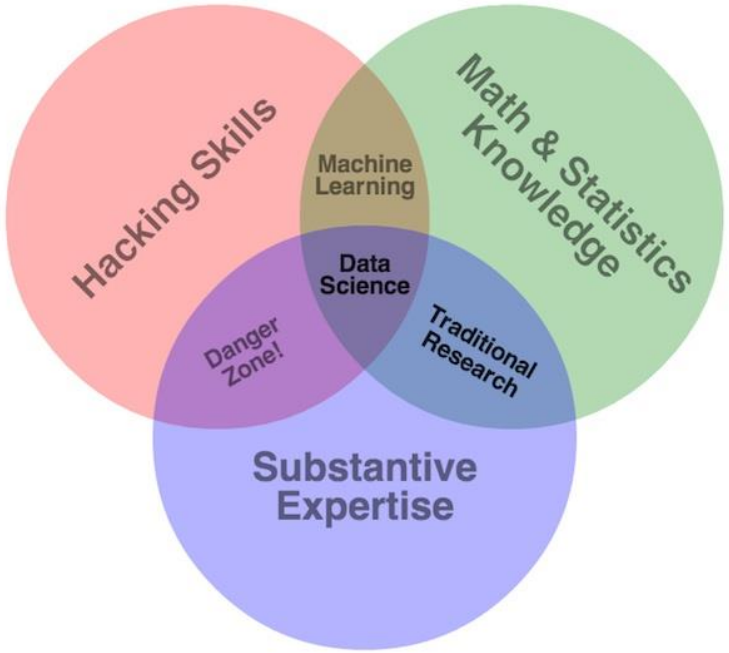
Tell stories from the beginning. Every slide you write must entertain, challenge. Work closely with the client. This tells you what they really are excited with.

Explicability or Accuracy?

Do they want highest possible accuracy?
Go for RF, SVM etc.

Do they want to understand and get high level Patterns? Try LR, DT
In doubt? Do both

Skills at a glance



Data related

Excel, SQL, Hadoop (Hive/Pig/Spark)

Handle text

Math related

Stat analysis, LR (Linear and Logistic), clustering

Linear programming,

Business related

Visualize

Tell stories

What is the Machine Learning Process?

A person will find job in Microsoft or not ?

Develop a binary classifier to predict

Details of folks who got selected for MS and who DID NOT make it

- # years of job experience
- previous companies
- keywords in resume

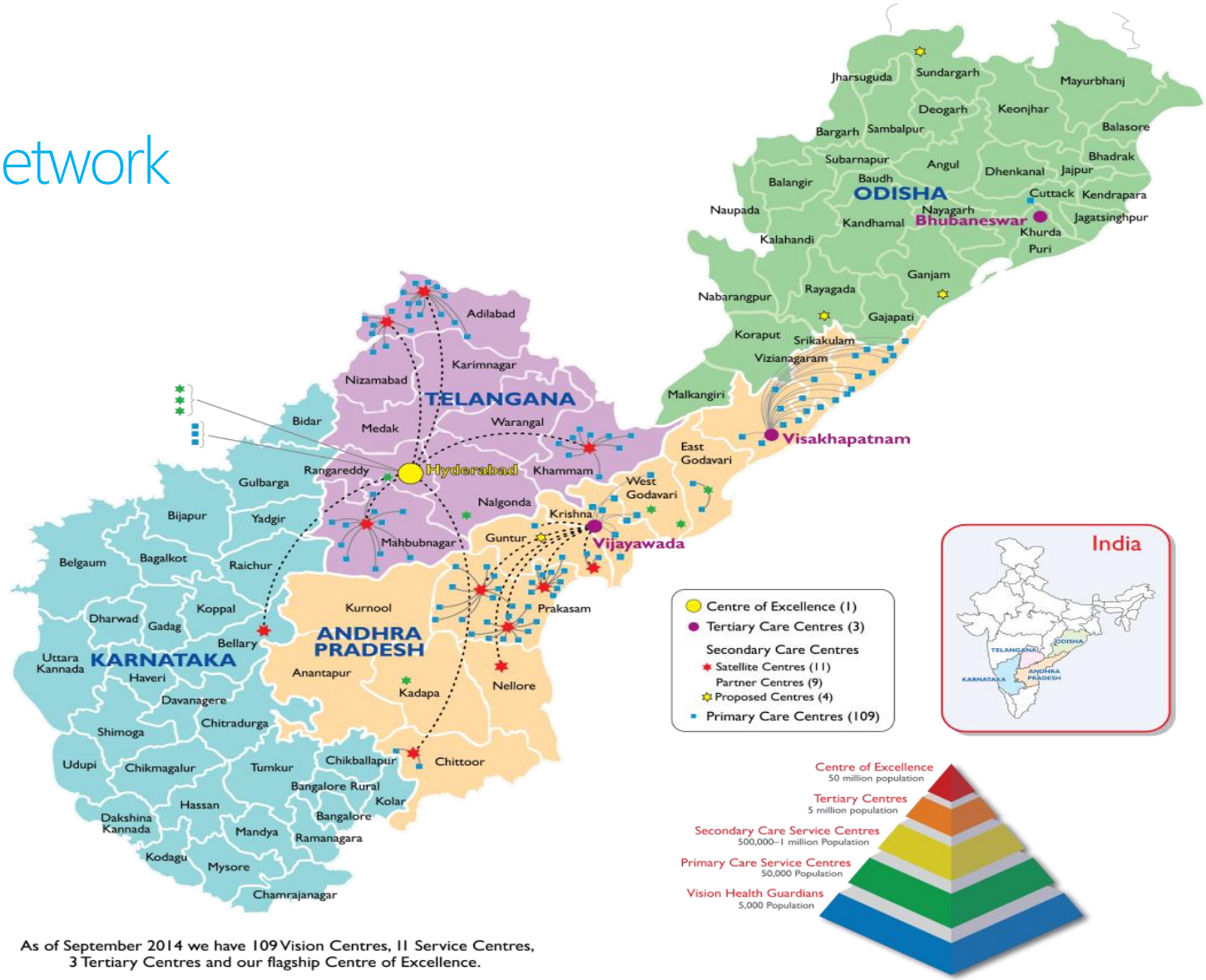
- Decision Trees
- SVM
- Random Forest
- Neural Networks
- Logistic Regression

- Precision
- Recall
- AUC
- Accuracy
- F1

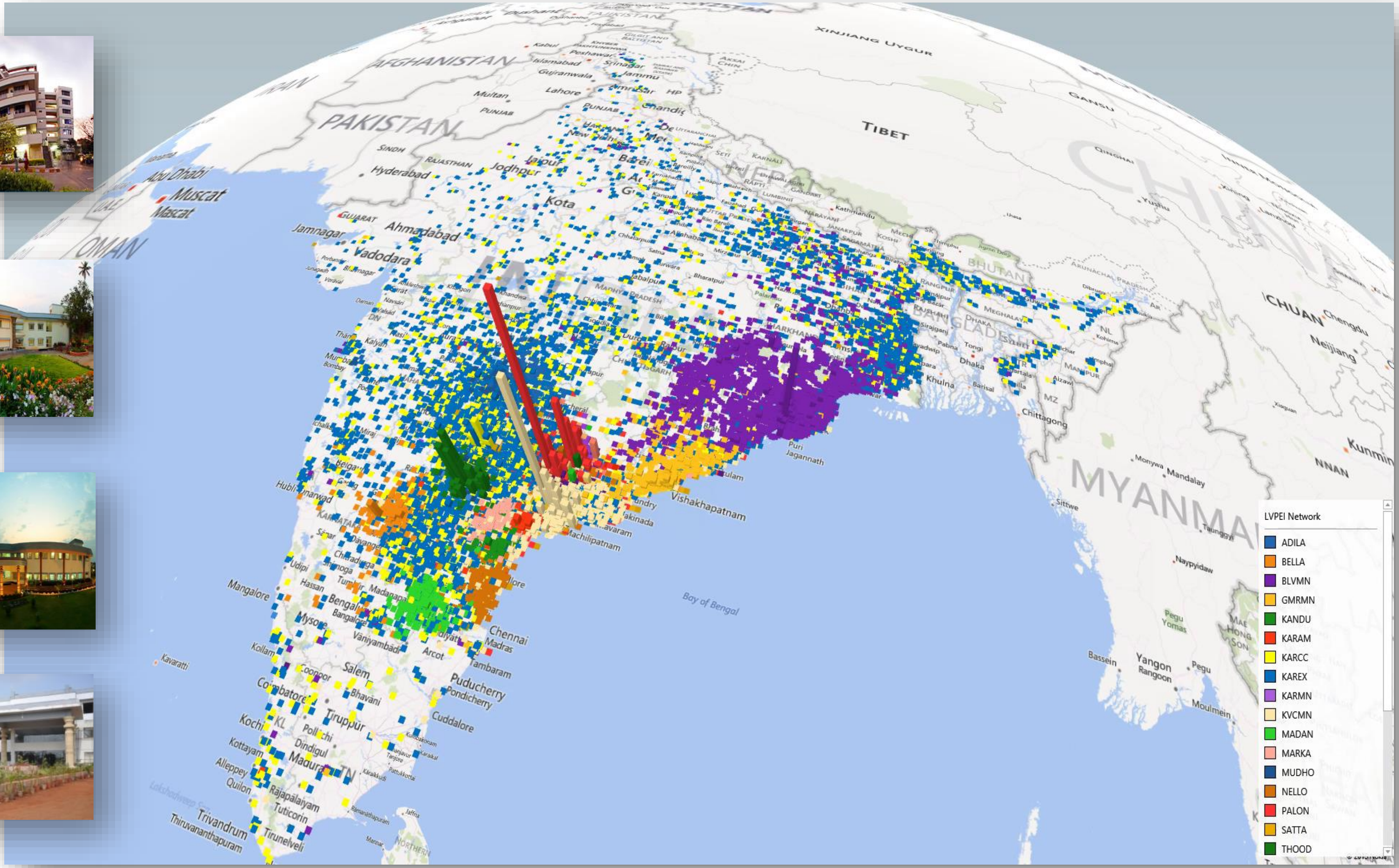
Practical Issues when applying Machine Learning Models

- Supervised, semi-supervised, unsupervised algorithms.
- Overfitting: Too much specialization on training data.
- Understanding data and acquiring clean data is the key.
- Various data transformations like normalization, discretization could be useful.
- It is important to handle missing values appropriately.
- If the data is imbalanced, it is important to oversample the minority or undersample the majority class.
- If certain mis-classification errors are more important than others, cost-sensitive learning should be done.
- Many times, selecting a set of good predictors as features is useful compared to throwing all possible features at the model.

LVPEI Network

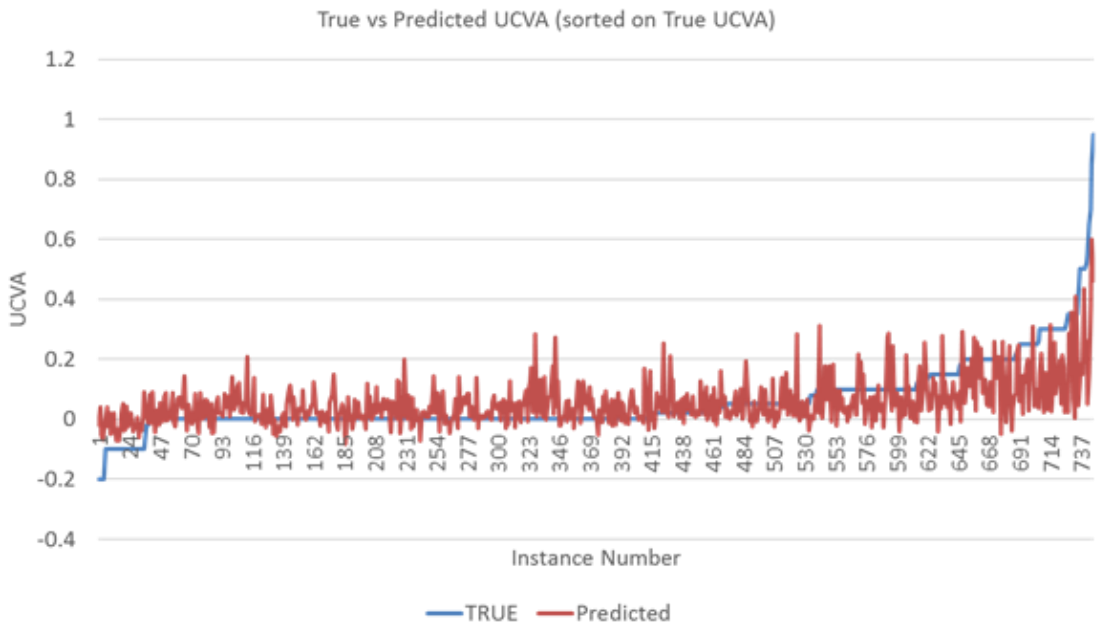


As of September 2014 we have 109 Vision Centres, 11 Service Centres, 3 Tertiary Centres and our flagship Centre of Excellence.


LVPEI Network
ADILA
BELLA
BLVMN
GMRMN
KANDU
KARAM
KARCC
KAREX
KARMN
KVCMN
MADAN
MARKA
MUDHO
NELLO
PALON
SATTA
THOOD
AFGHANISTAN
PAKISTAN
TIBET
MYANMAR
OMAN
Bay of Bengal

Machine Learning

Predict the LASIK refractive surgery outcomes



Prediction of Post-Operation Eye Number for Lasik Surgeries

- Given pre-surgery data about 404 patients
- Train a machine learning model to predict what would be the new "eye number" (Uncorrected Visual Acuity or UCVA) 1 day/1 week after the surgery
- Features: Gender, age, UCVA, Near vision, BCVA with glasses, Sphere, Cylinder, Axis, Spherical equivalent, Slit lamp, IOP, Retina, Topography machine, AR sphere, AR cylinder, AR Axis, Preop Corneal Thickness-Thinnest, steep-K, Flat-K, Axis@Flat K
- Overall we can predict the right "eye number" UCVA with an L1 error of 0.07 (±0.0123) for day 1 and 0.06 (±0.0085) for week 1 after the surgery.

Results Summary

- Missing values were replaced by average value for column for numeric features, and maximum value for column for categorical features.
- Categorical features were converted to numeric features by mapping them to consecutive integers.

10 fold cross validation over 747 instances that had the post-surgery UCVA for day 1 after surgery

Model	L1(avg)	L2(avg)	RMS(avg)
Linear Regression	0.0771 (0.0084)	0.0136 (0.0065)	0.1155 (0.026)
Poisson Regression	0.0744 (0.008)	0.0128 (0.0049)	0.1108 (0.0217)
Boosted Decision Trees Regression	**0.0695 (0.0123)**	**0.0108 (0.0051)**	**0.1024 (0.0234)**
Neural Network Regression	0.082 (0.0077)	0.0142 (0.0068)	0.1179 (0.027)

Day 1 UCVA

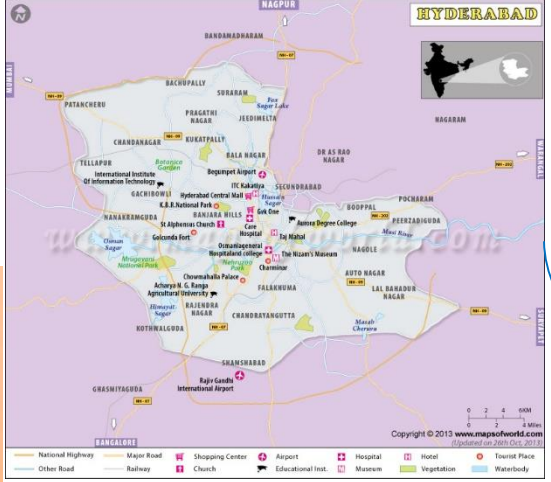
10 fold cross validation over 622 instances that had the post-surgery UCVA for week 1 after surgery

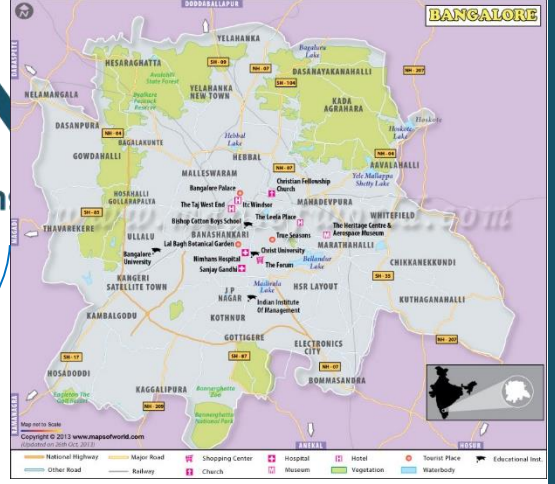
Model	L1(avg)	L2(avg)	RMS(avg)
Linear Regression	0.0657 (0.0112)	0.0138 (0.0068)	0.1138 (0.0303)
Poisson Regression	0.0636 (0.0118)	0.0132 (0.0107)	0.1102 (0.0374)
Boosted Decision Trees Regression	**0.0623 (0.0085)**	**0.0111 (0.005)**	**0.1008 (0.0229)**
Neural Network Regression	0.0673 (0.0121)	0.0149 (0.0091)	0.1169 (0.0342)

Week 1 UCVA

Practical Issues

- Interpreting Data and ensuring data quality
- Data privacy
- Safe data sharing
- Delays and formatting issues
- Domain understanding
- Deployment
- User acceptance
- Charging customers
- Data Size





HYDERABAD

Office and Classrooms

Plot 63/A, Floors 1&2, Road # 13, Film Nagar,
Jubilee Hills, Hyderabad - 500 033
+91-9701685511 (Individuals)
+91-9618483483 (Corporates)

BENGALURU

Office

Incubex, #728, Grace Platina, 4th Floor, CMH Road,
Indira Nagar, 1st Stage, Bengaluru – 560038
+91-9502334561 (Individuals)
+91-9502799088 (Corporates)

Classroom

KnowledgeHut Solutions Pvt. Ltd., Reliable Plaza,
Jakkasandra Main Road, Teacher's Colony, 14th Main
Road, Sector – 5, HSR Layout, Bengaluru - 560102

Social Media

Web:	http://www.insofe.edu.in
Facebook:	https://www.facebook.com/insofe
Twitter:	https://twitter.com/Insofeedu
YouTube:	http://www.youtube.com/InsofeVideos
SlideShare:	http://www.slideshare.net/INSOFE
LinkedIn:	http://www.linkedin.com/company/international-school-of-engineering

This presentation may contain references to findings of various reports available in the public domain. INSOFE makes no representation as to their accuracy or that the organization subscribes to those findings.