Inspire…Educate…Transform.

Architecting a Solution

Dr. K. V. Dakshinamurthy

President, International School of Engineering

ML or AI

Do you need an expert to solve this problem?
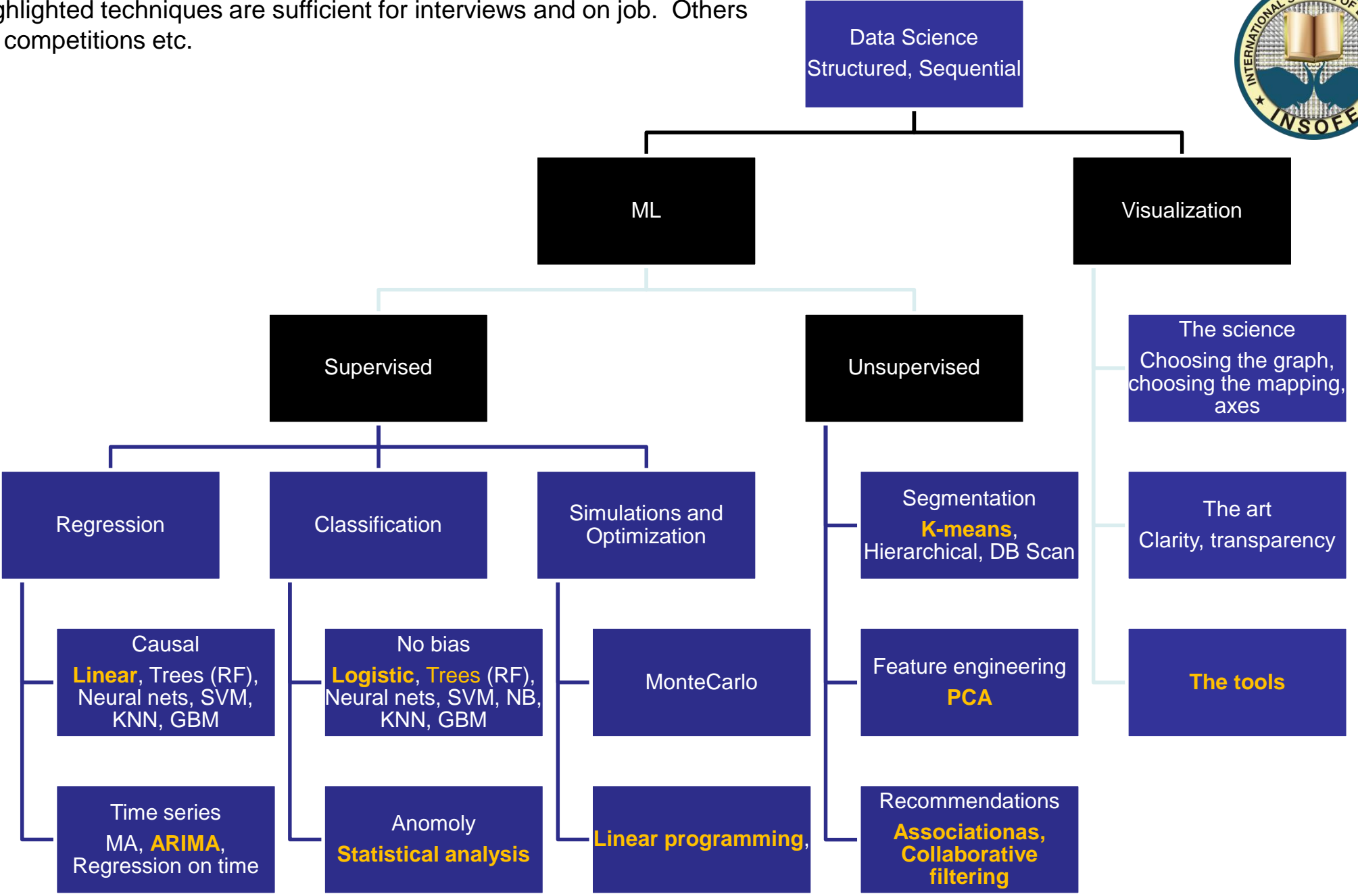
ML is the way to go

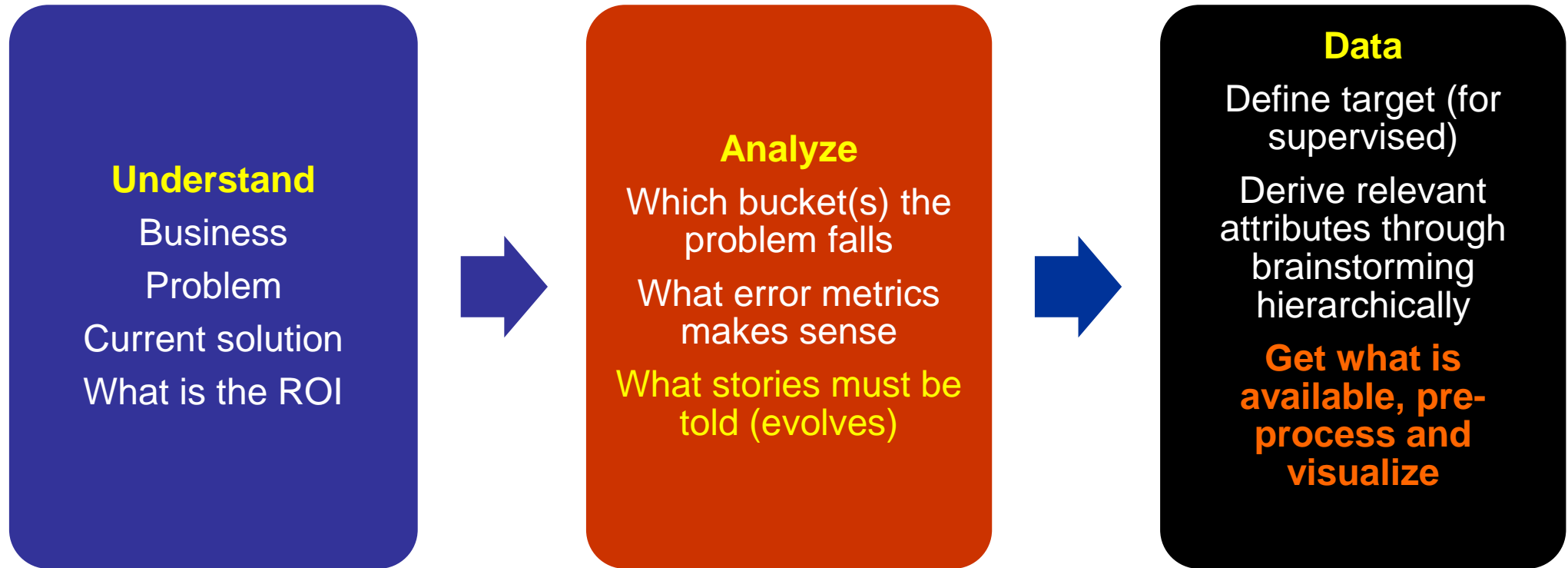Do you need lots of normal people to solve this problem?

AI is the way to go

ML

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

Data Science
Structured, Sequential

ML

Visualization

Supervised

Unsupervised

The science
Choosing the graph, choosing the mapping, axes

Regression

Classification

Simulations and Optimization

Segmentation
K-means, Hierarchical, DB Scan

The art
Clarity, transparency

Causal
Linear, Trees (RF), Neural nets, SVM, KNN, GBM

No bias
Logistic, Trees (RF), Neural nets, SVM, NB, KNN, GBM

MonteCarlo

Feature engineering
PCA

The tools

Time series
MA, **ARIMA**, Regression on time

Anomoly
Statistical analysis

Linear programming,

Recommendations
Associationas, Collaborative filtering

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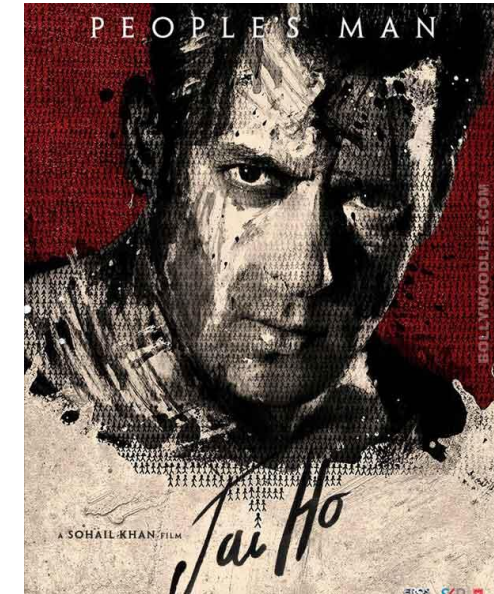
Try at least 3 models on 3 types of data sets with 3 sets of hyper parameters (27) before you pick the final choice.

Validation

Design validation strategy.

Divide the data into train, test and CV.

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

Gaining acceptance

Once upon a time... <sob>

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.
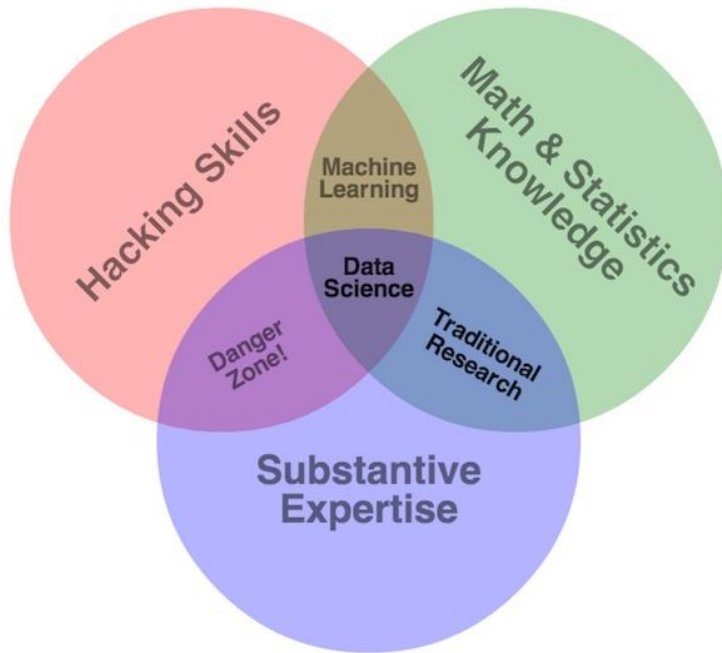
Explicability or Accuracy?

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

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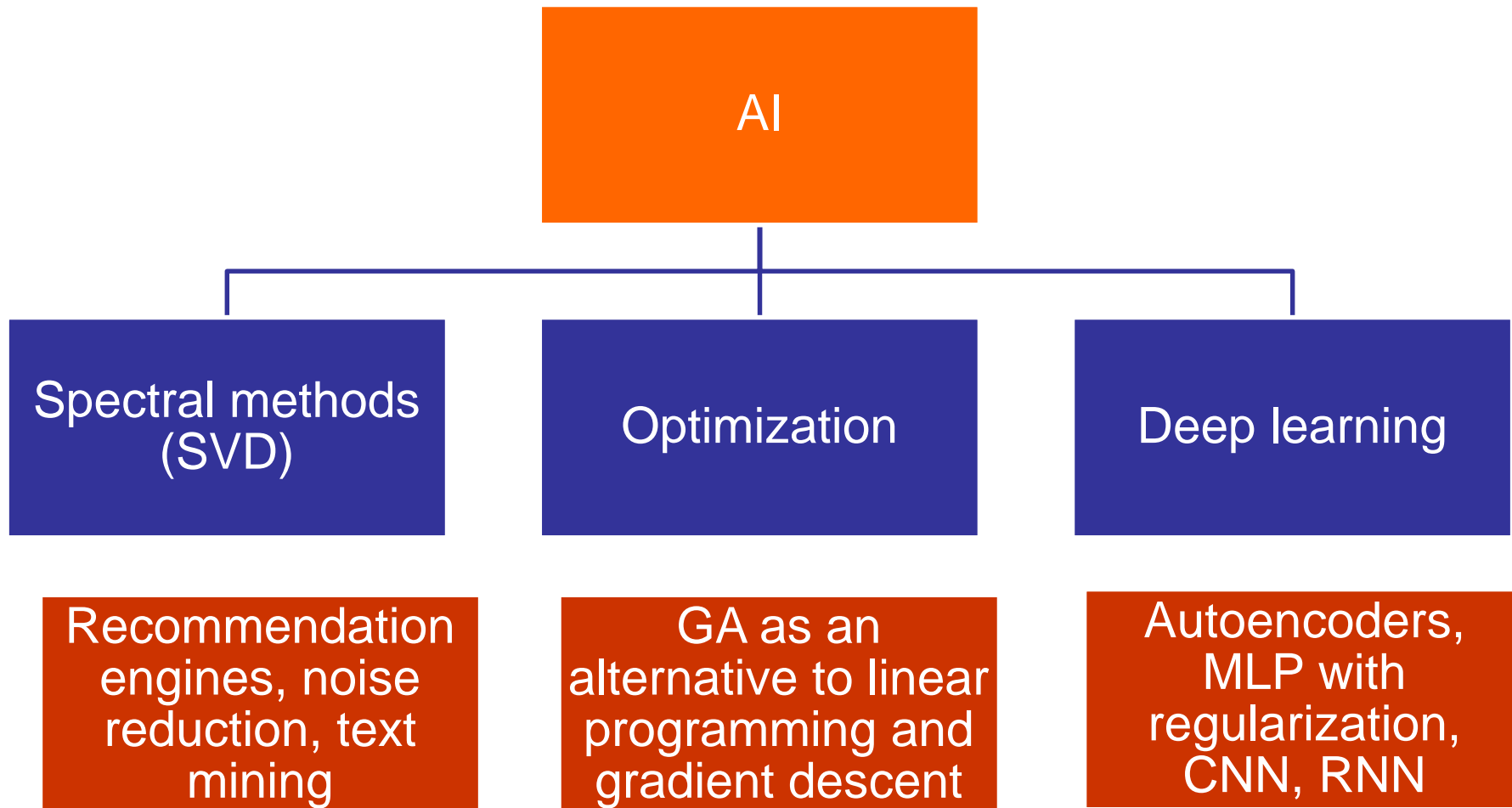In doubt? Do both

Skills at a glance

Data related	Math related	Business related
Excel, SQL, Hadoop (Hive/Pig/Spark)	Stat analysis, LR (Linear and Logistic), clustering	Visualize
Handle text	Linear programming,	Tell stories

Specifically for DL

AI

AI

Spectral methods (SVD)

Optimization

Deep learning

Recommendation engines, noise reduction, text mining

GA as an alternative to linear programming and gradient descent

Autoencoders, MLP with regularization, CNN, RNN

DL knows to use big data

DL
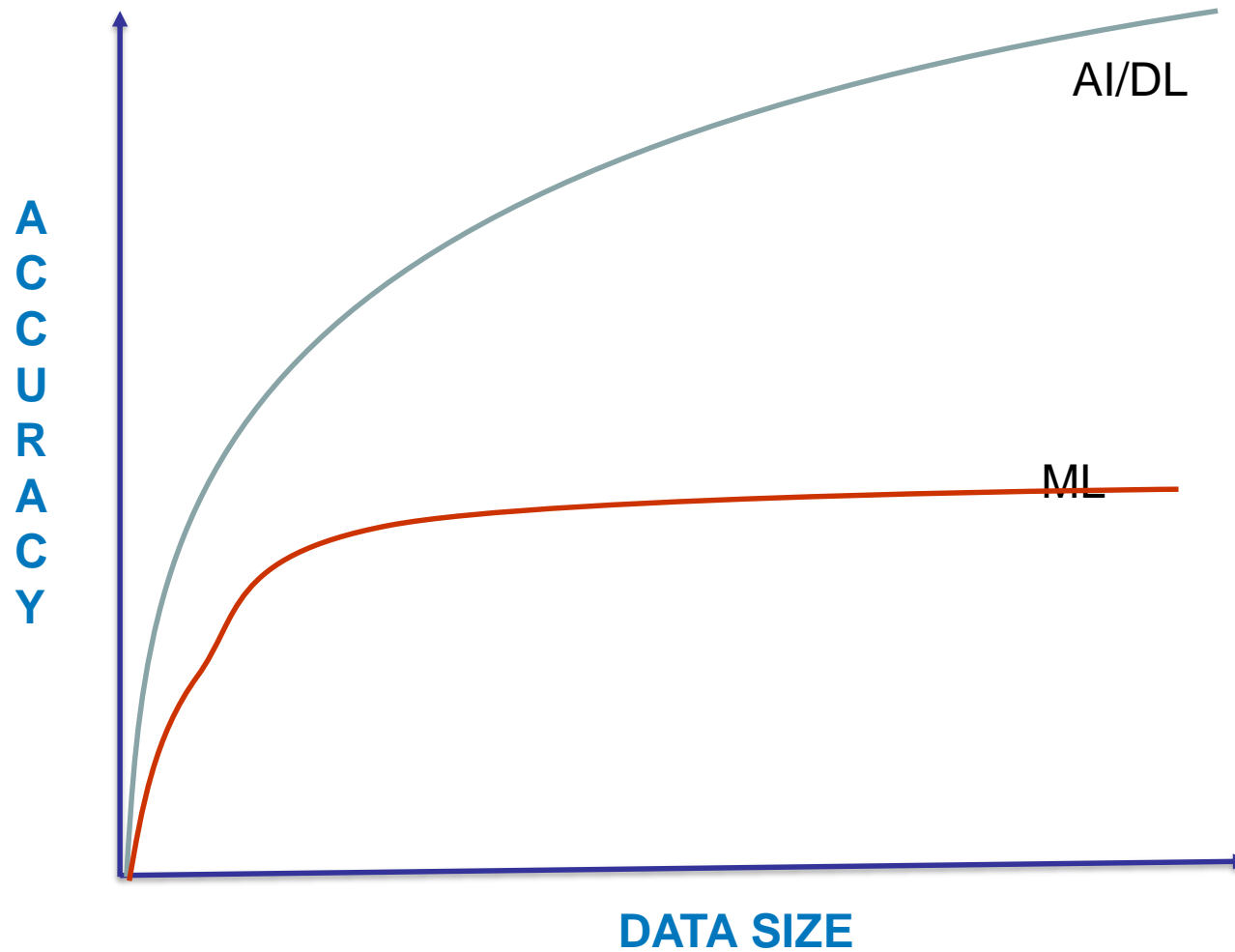
- Do not engineer features (your brain does not at least explicitly)

- Get large data sets, use large nets and use large regularizations
  - Hit bias and variance in one go
  - Things will work out quite well

DL: What data to train on

- ML spend a lot of time on creating just the right data

- AI/DL: Did you learn this task looking at slightly different data?
  - Then you can use that slightly different data
  - Did you learn to appreciate music of instrument and music of other instrument separately? Then music of one instrument can be used to train for the other

How to divide training and test

- Image recognition of a crack in a wheel captured during test time
  - Say, we have 1,000,000 images of cracks non-cracks downloaded from web. These are not perfect examples. But, have some relevance. Let us call it synthetic data
  - Let us say, we have 10000 cracks captured in the real scenario
  - How do we split?

Development and Test should be similar

- Option 1: Training: 950,000 synthesized data; Validation: 50,000 synthesized data, test: 10000 real images

- Option 2: Better: Training 1,000,000; Validation: 5000 real images, test 5000 real images

- Option 3: Best: SynthTraining: 950,000; SynthValid: 50,000; RealValid: 5000; RealTest: 5000

So, how to interpret?

- Collect the following numbers
  1. Human performance
  2. Synthetic Training
  3. Synthetic Validation
  4. Real Validation
  5. Real Testing

How do you treat

- 1 and 2: Bigger model, bias
- 2 and 3: Regularize, more data
- 3 and 4: Remove the training data that is very different from dev
- 4 and 5: More of dev data or a better architecture

# Time has come for comparing with humans

- Error analysis of bias variance effects
  - Training error: 8%
  - Dev error: 10%

  - Does not tell you whether it is bias or variance problem

  - But, if we know that human performance is 7.5%

How to get better

- Give yourself three months!
- Try three Kaggle/competition problems from ML or AI whose results are out
- Pick three solutions from the leader board
- Replicate as is and build two alternatives (one major change (model/additional) and one minor
- Participate in three fresh competitions and publicize!
- Accept the best of the 3 employers who come looking for you

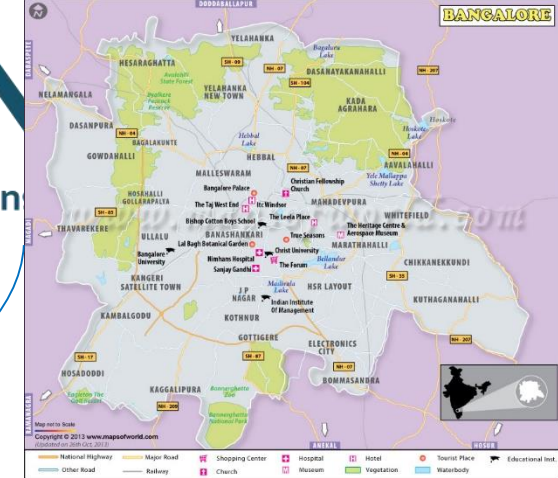# Software development versus data scientists (conflict)
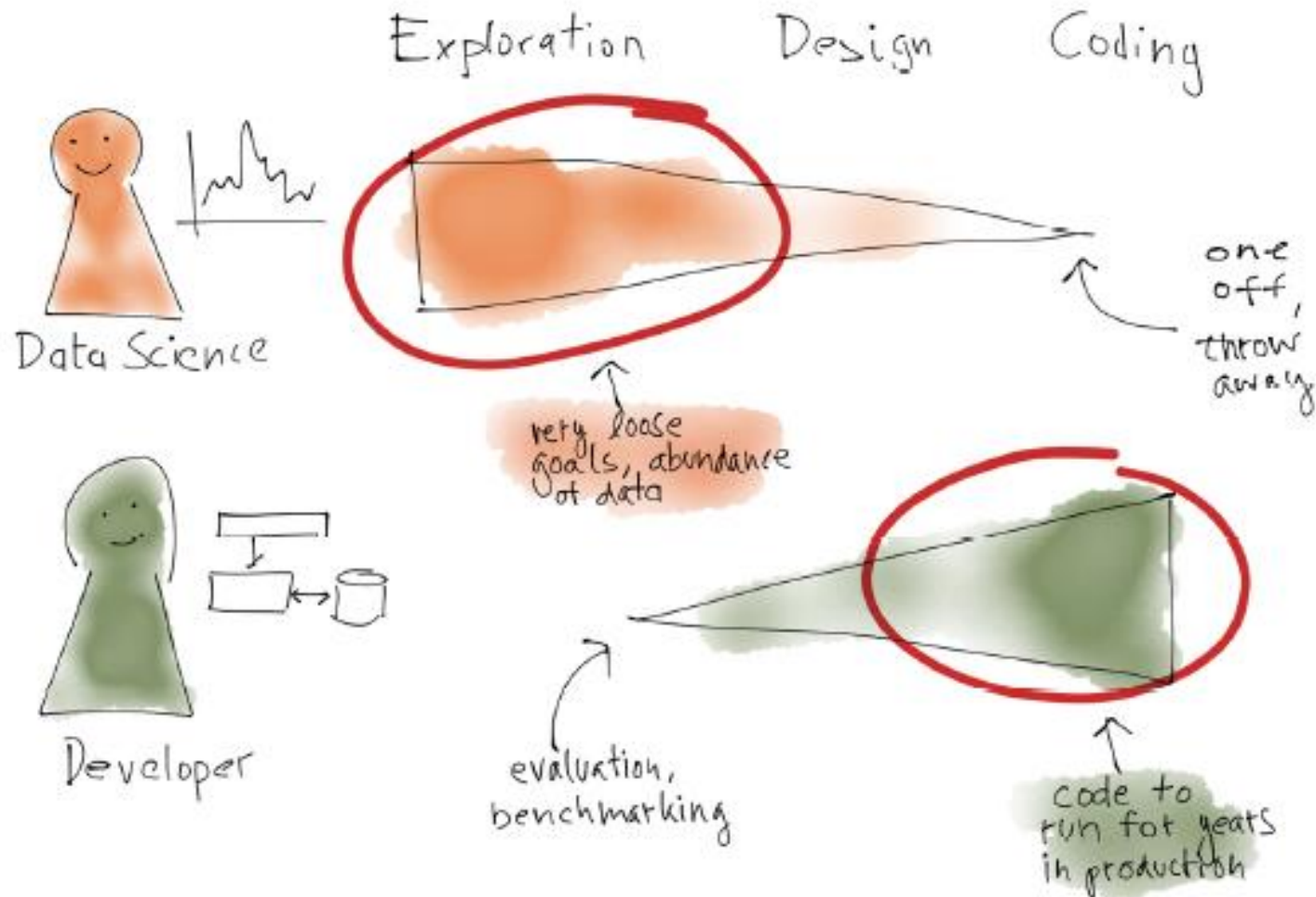
HYDERABAD

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

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

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

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.