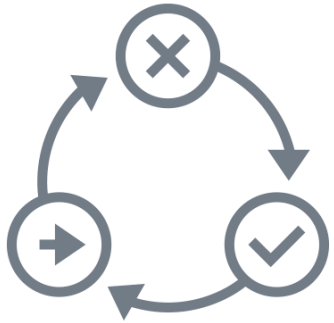
Analytics in India a few trends – field notes from a layman

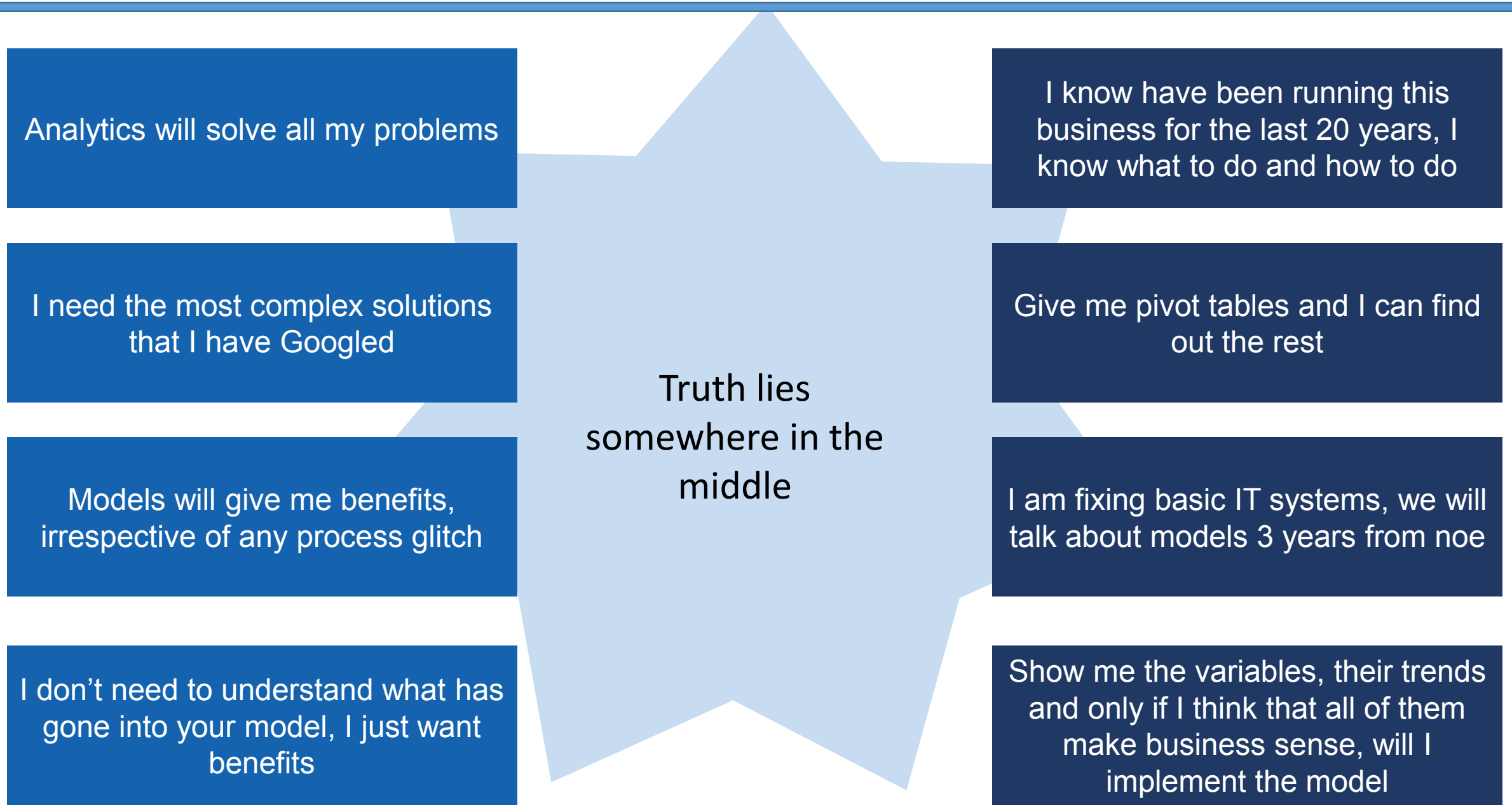
Hindol Basu

India presents its own unique challenges in leveraging the power of data

Focus on quantified impact leading to top and bottom line results	Focus on data quality and data management	Balance between intuitive feel of models and higher predictive power	Communicating analytics across the organization	Adopt a “test & learn” approach
				

1. Start with KPIs and understand how they translate to financials and what are the benchmarks
2. Identify how benefits will be measured – test vs. control, pre vs. post, champion vs. challenger
3. Identify the change management requirement and manage the softer aspects of implementation
4. Start with glass box methodologies – so that business users have confidence
5. Move from glass box to black box methods (if needed)

The limitations of a data driven approach is often not understood properly



These gives rise to 4 key challenges

1. Difficulty in implementing relatively complex models (neural network, tree ensemble etc.)
2. Data quality issues, sparse data and lack of data
3. Change management required to push implementation
4. Sound methodology of tracking, attributing and demonstrating financial benefits

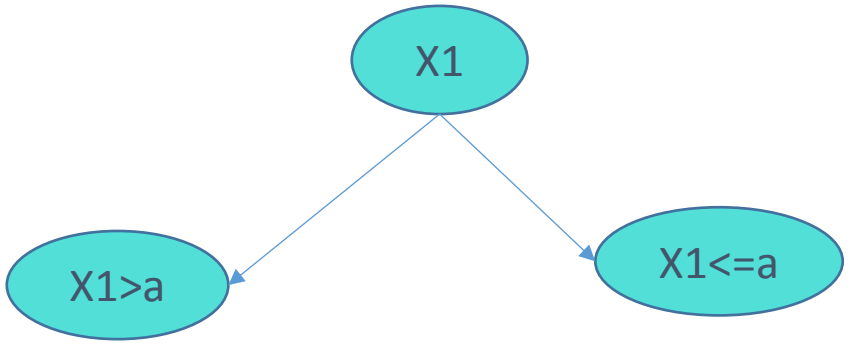
Difficulty in implementing complex models

Models with no or very limited interactions

$$\frac{d\,(\text{Prediction})}{d\,X1} = K \text{ (irrespective of X2....Xn)}$$

$$\frac{d\,(\text{Prediction})}{d\,X1} = KX2 \text{ (irrespective of X3.....Xn)}$$

Simple decision trees



Complex interactions

$$\left.\frac{\delta\,(\text{Prediction})}{\delta\,X1}\right| X2....Xn = K$$

Complex interactions

- Use intuitive models in cases where current processes requires subject matter expert judgement
- Use complex modes in cases where only human intelligence needs to be automated
- Create sensitivity analysis to help unbox complex models
- Leverage intelligent methods of segmentation, tree interaction variables etc. to build power of interaction within simple models

Data quality – data is our best friend, poor data quality is our worst enemy

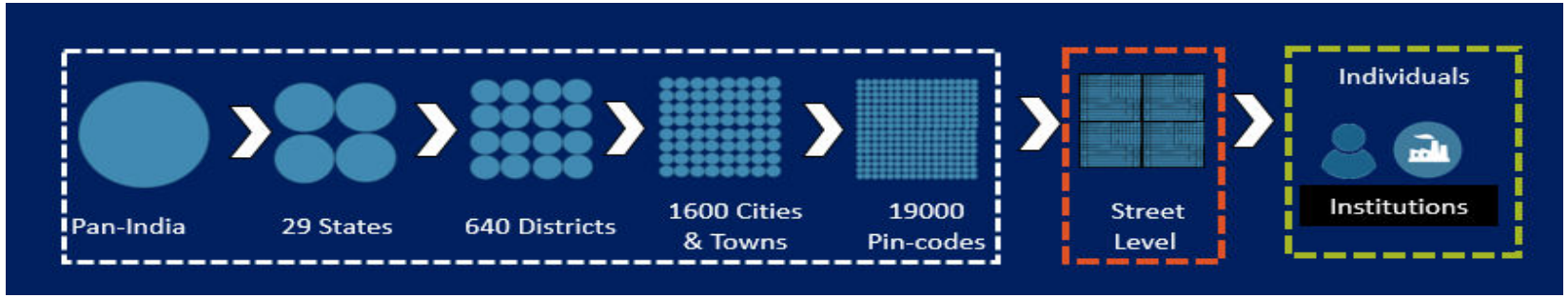
Models with no or very limited interactions

- Missing value
- Erroneous values
- Errors creeping into the data during summarization and storage in warehouse
- Lack of unique ID for the entity of interest (customer, supplier etc.)
- Lack of data capture
- Lack of history
- Issues in human data capture
- Lack of a proper data owner who can help

Models with no or very limited interactions

- Unique entity creation (unique customer, household, supplier etc.)
  - Particularly critical for non-financial services (in the absence of Aadhar, PAN etc.)
  - Fuzzy matching of name, address
  - Issues of multiple phone, email
- Value correlation analysis
- Identify if missings are part of data generation process
- Analytical methodologies for sparse data handling
  - Expectation maximisation
  - MCMC
- Non-human captured data: clickstreams, sensors etc. tend to have far lesser issues
- Plan for future data capture
- Takes about 75%-80% of project timeline

Key Initiative-1: Aggregate multiple sources of public external data



- Census
- NSO
- Credit Bureau
- Vehicle registration
- Crop prices
- Ministry of Corp Affairs
- Map data
- …. and many more

Developing and managing a complex external data at various levels of aggregation

- Proprietary surveys
- Customer data
- Augmented (off-us) customer data from partners, affiliates
- Lists from marketing activities

- Most organizations view external data only for the purpose of solving a specific problem
  - Store/branch location
  - Demand forecasting at micro geo level
  - Distribution planning
  - And others
- There may be a lack of concerted efforts in creating a comprehensive economic data repository by companies
- Indiucus and Nielsen may be a good example

Key Initiative-2: Sensor data for manufacturing

Operations managers have been using traditional control charts for ages

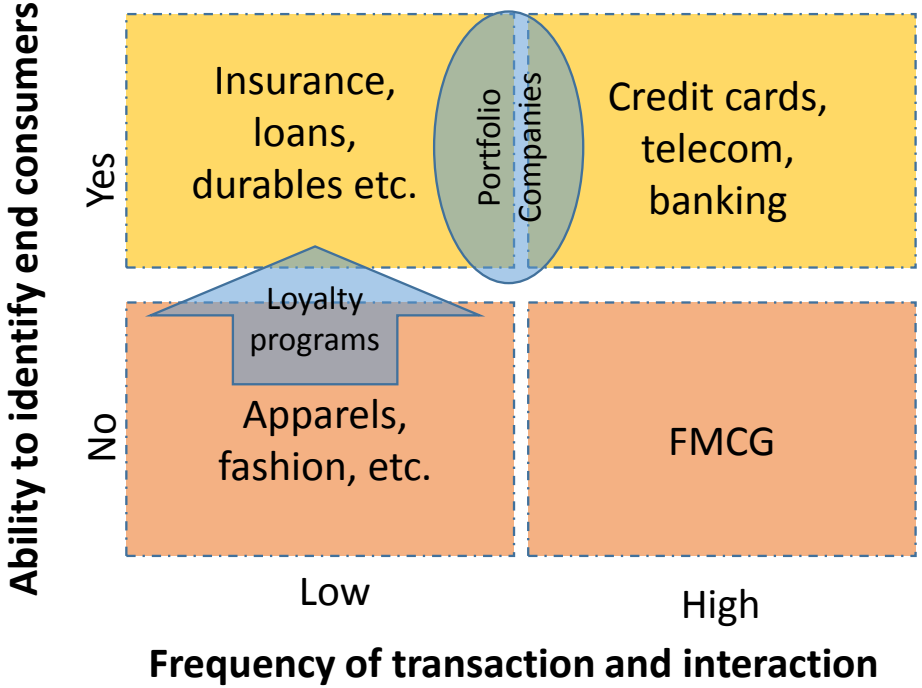
- Key sensors have been part of most equipment
- Univariate approach using only max and min
- Multi variate anomaly detection
- Availability of low cost of sensors

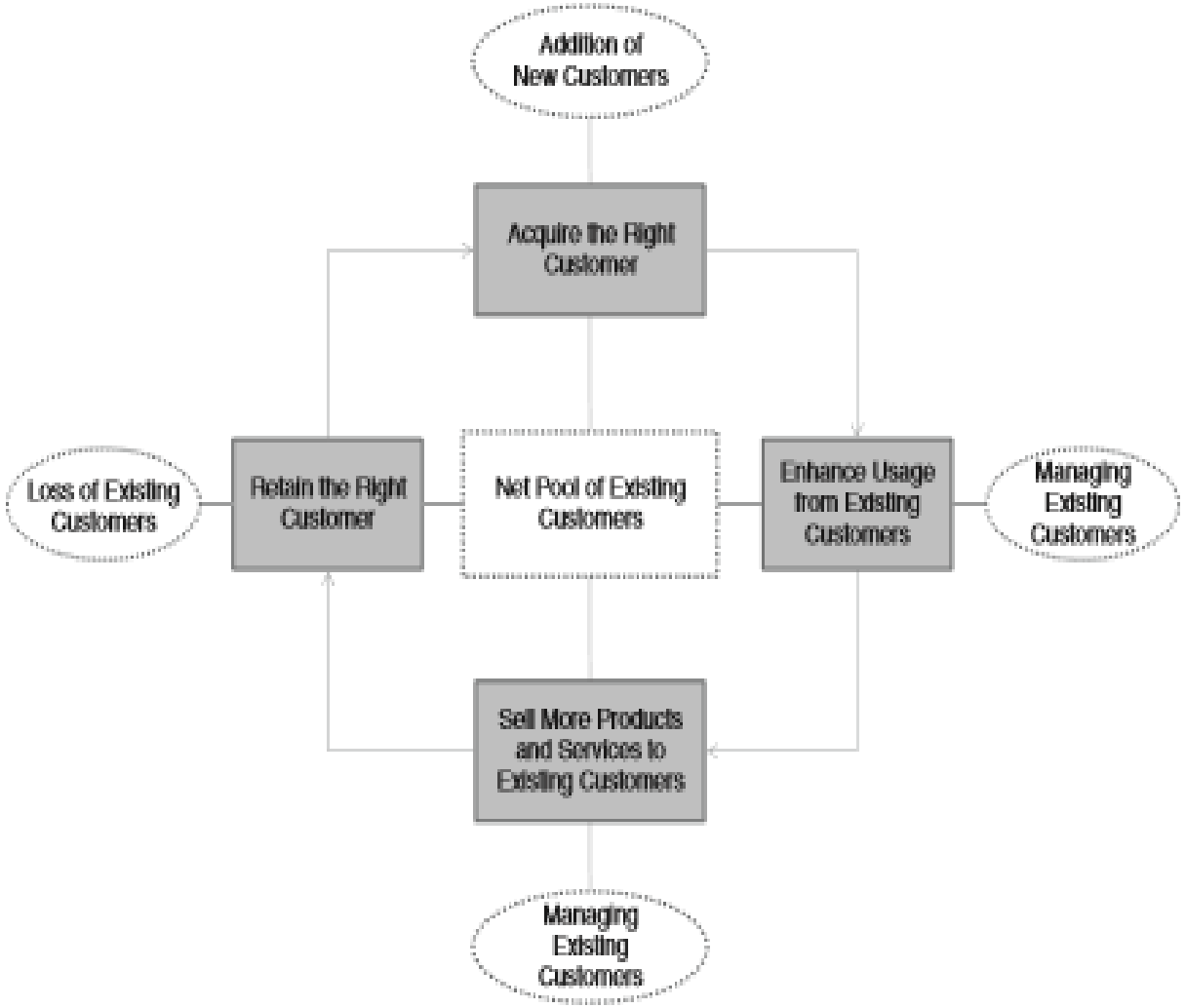
- Key use cases
  - Predictive Asset Management
  - Predictive Quality Management
- Unsupervised learning for anomaly detection
- Supervised learning for identifying quality issues, equipment failure
- Thresholding is another key problem
- Type-1 and Type-2 errors have different costs associated. Hence, meta cost analysis is important for thresholding
- The data is structured but of high frequency in update; hence, specific data engineering is important
- Linking operations technology to ERP, CRM etc.

Key Initiative-3: consumer companies trying to bring efficiency of marketing

Solution Landscape

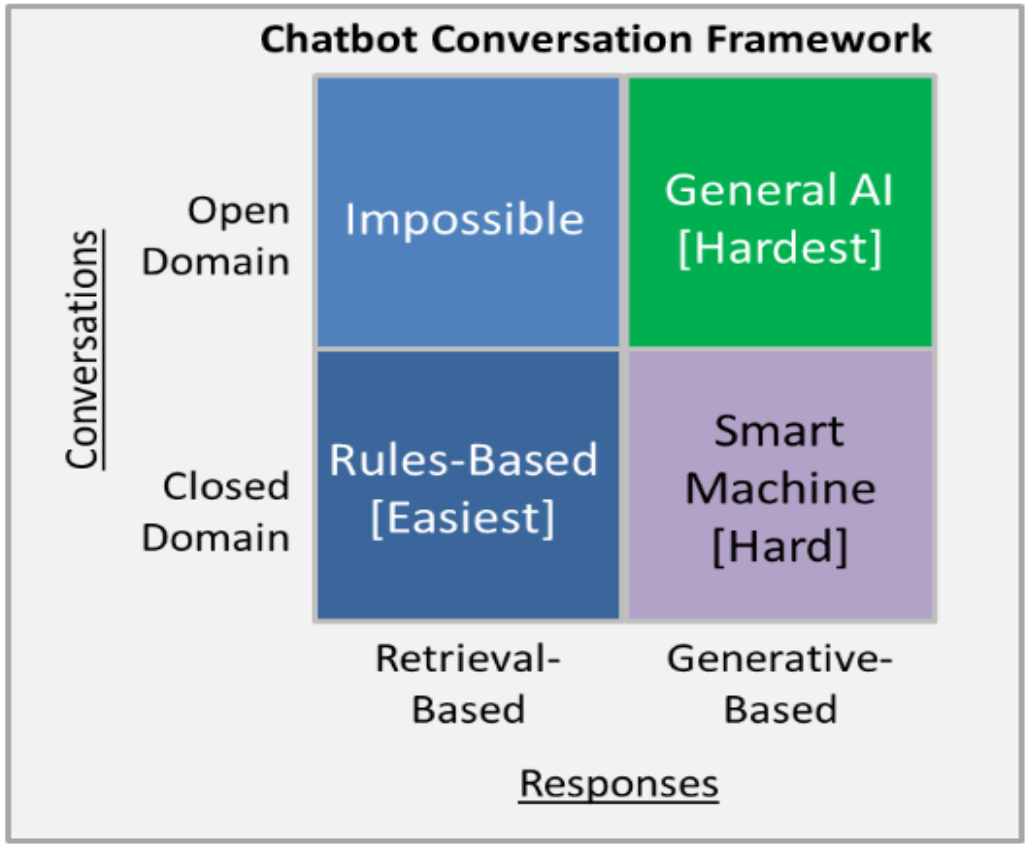
1. 360 degree customer view – data integration and unique ID creation
2. Micro segmentation – use demogs, transaction, interaction and others (including channel preference)
3. Targeting models for cross sell, retention and reactivation
4. Optimized communication to maximise revenue





Managing the curse of false positives in churn models

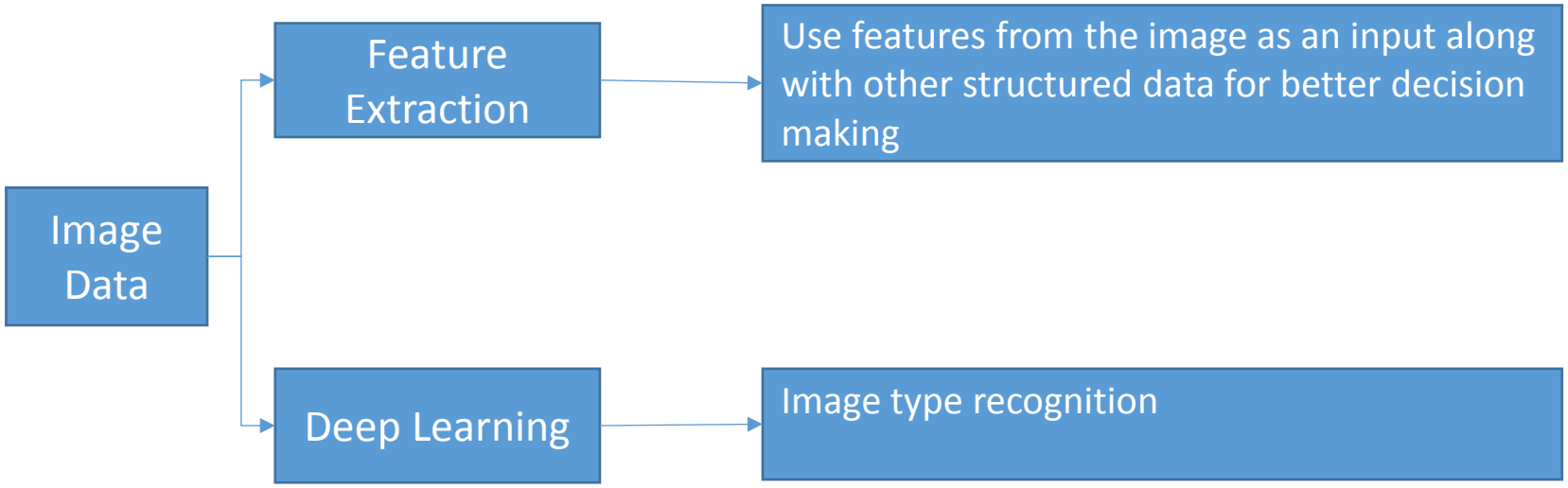
Key Initiative-4: Automated customer interaction



- **Retrieval bots**: Retrieval bots do not generate any response on their own, they pick up one of the pre-defined responses from the thousands of available responses
- **Conversational bots**: are a sub set of retrieval bots, that uses business rules derived from expected flow of conversations
- **Generative bots**: these bots actually generate a response given the question
- **Importance of the context**: Specific retrieval rules (models) should be created for specific context so as to identify the right response
- **Usage of paid frameworks (e.g. Microsoft bot framework)**: Tends to be very expensive in the long run. Loss of IP

Key Initiative-5: Explore new data sources

- Images
- Scanned documents
- Map data
- Satellite image data
- Video data
- Wifi Latching data



- Applications of satellite images for agricultural applications

- Physical click stream
- Movement within store
- Store fraud

Key Initiative-6: GPS data for logistics management

- Availability of real time GPS data has very interesting applications for India
  - Avoiding toll roads
  - Pilferage
  - Idle time
  - And many more
- Stochastic nature of lead times
- Semi real time re-routing
- Driver behaviour monitoring
- Plugging in telematics for failure prediction

Key Initiative-7: Small data

Big Data	+	Small Data	=	Value
• Social media • Customer data • External lists and consortium data • Example: face book likes		• Sample survey • Customer survey • Example: preferred lifestyle		• Model • Example: predict lifestyle based on likes
X		Y		f(X)

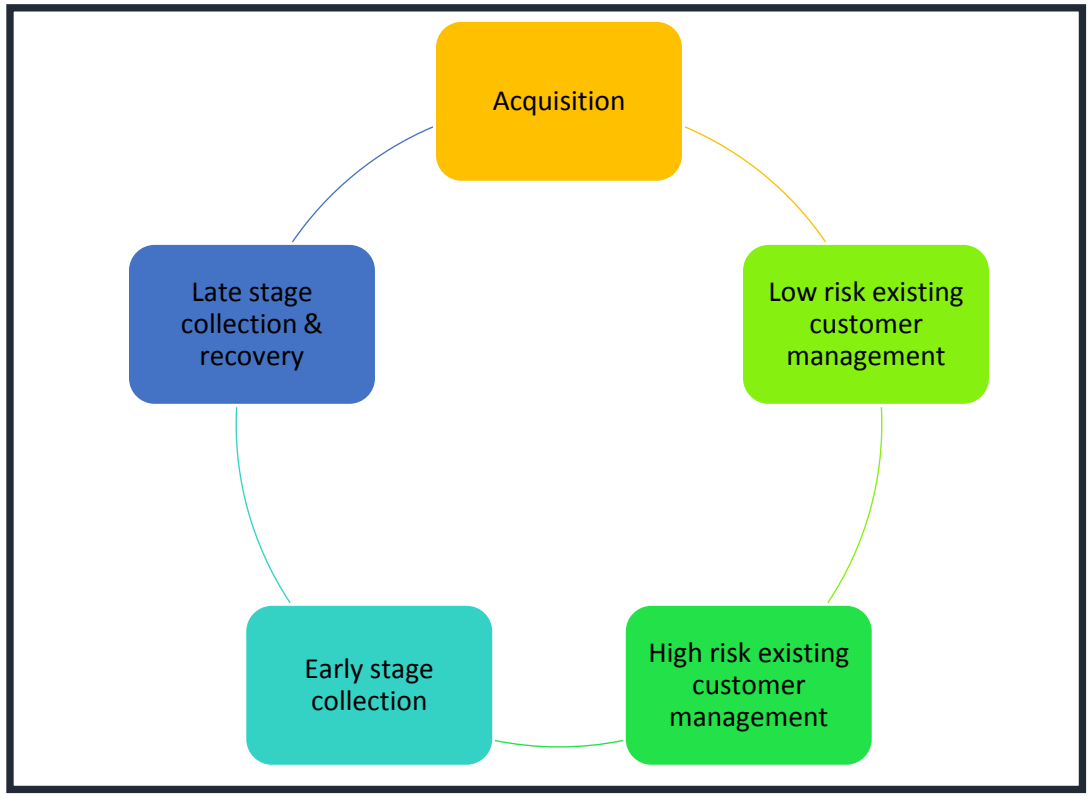
Key Initiative-8: Risk estimation

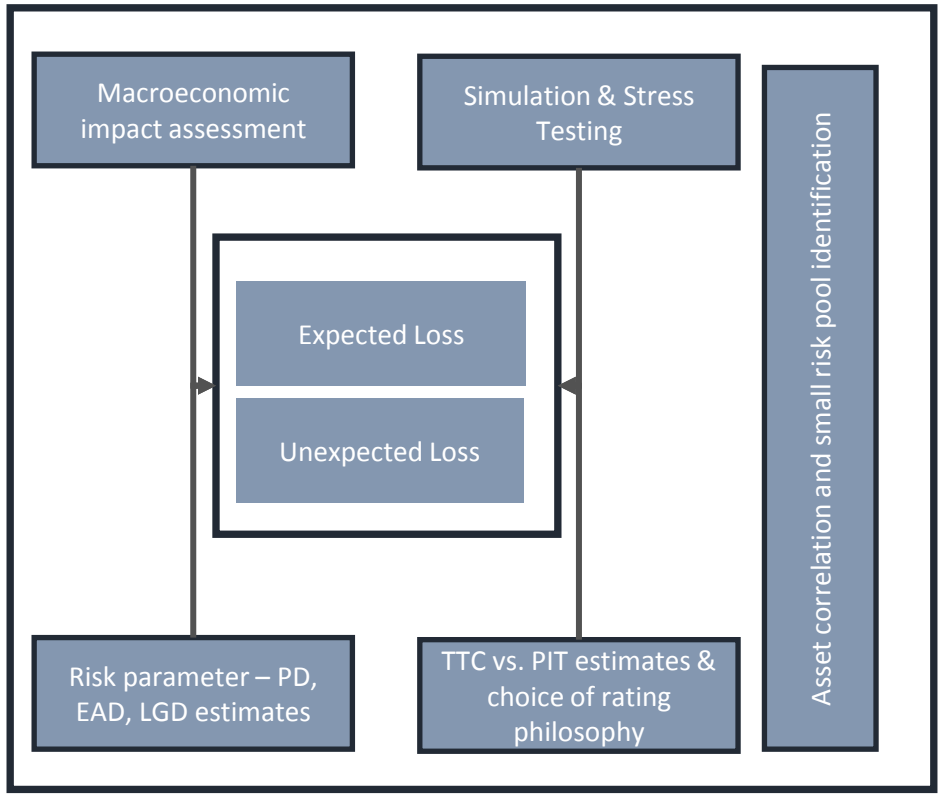
Credit risk management across customer life cycle

Provisioning & Capital management

Acquisitions	Existing Customer Management	Collections and Recovery	Provisioning and P&L Management	Capital Compensation and Balance Sheet Management Identification of risk pools

- Data
- Reports
- Business Analysis
- Predictive Analytics
- Optimization
- Governance and Tracking





Key Initiative-8: Risk estimation (consumer and SMEs)

Risk distribution - SMEs

- Lack of financial statement
- Performance definition – grouping facilities, entities, customer group to account for risk spreading across groups
- Alternate source of data – EPF submission, bill payment, etc.