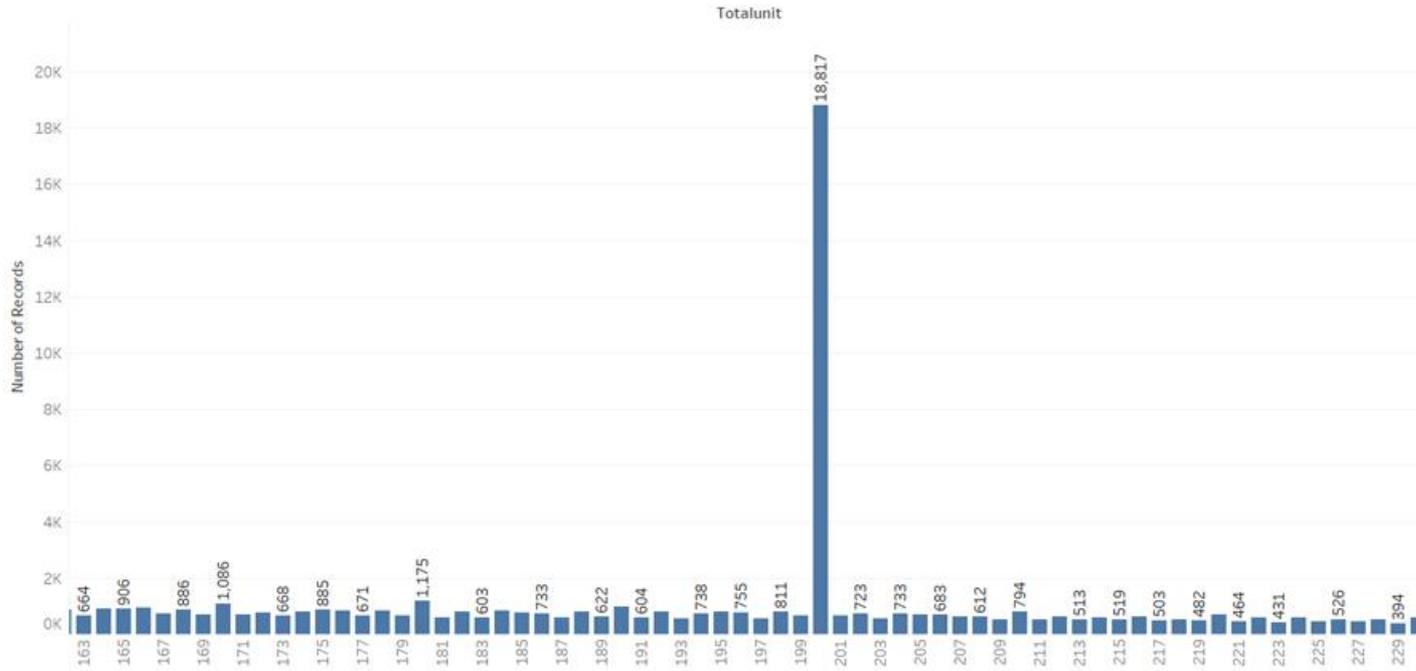


Auditing in Digital World

Gaurav Rai

Director, Centre for Data Management and Analytics

Number of bills energy consumptionwise

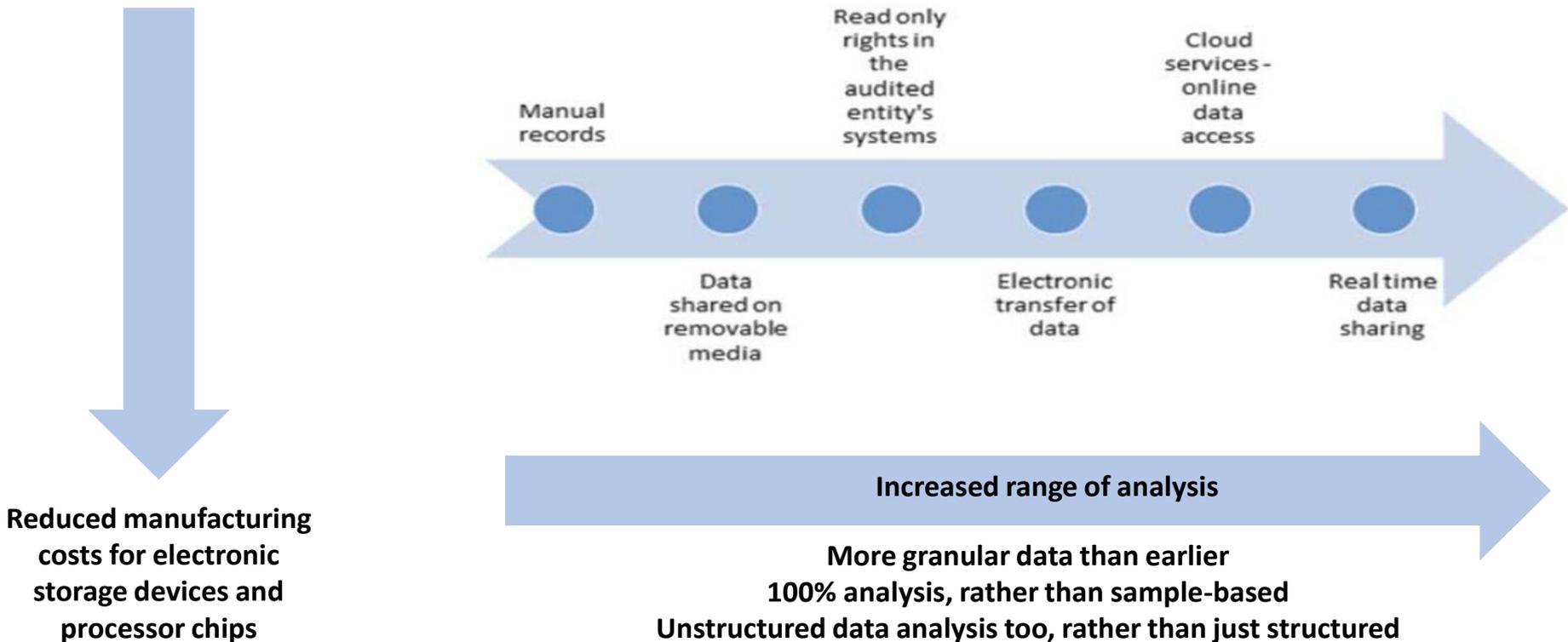


Electricity Billing Data

- 18817 consumers are having a monthly electricity consumption of 200 units (which is more than 20 times the average of consumers in other consumption units)

Why will Data analytics likely be required by SAIs?

Increasing ease, degree of access to data from audited entities and range of analysis





Use of Data Analytics

- ❑ For deriving insights from granular/ transaction-level data planning and detection of anomalies/ outliers to improve risk-based audit and focused audit execution
- ❑ Data analytics helps in analyzing 100% of electronic transactions rather than just a sample
- ❑ Field audit of a sample of transactions flagged by data analysis for validation/elaborate examination is generally required - in situations of poor data quality (incomplete/unreliable data) and in cases where 100% end to end computerization does not exist (i.e. where some records/ approvals/ documentation is manual)

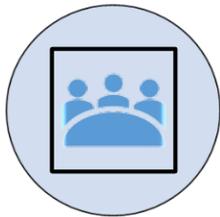
Re-usability of data analytics model with value additions each year – saves replicability of work every year in different country/regional offices –

Most of the IT Audits by IAAD will involve data analysis

For substantive testing of IT application controls, essentially to validate whether business rules have been adequately mapped

Setting-Up & Functions

Set up in 2016; Data Analytics Groups set up in Field Offices



Support to field audit offices in data restoration, preparation, and modelling.

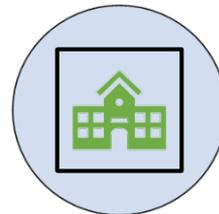
Verification of data analysis findings in audit reports including GIS based findings Working on certain

important data analytics projects which assists in flagship schemes/department data driven audit

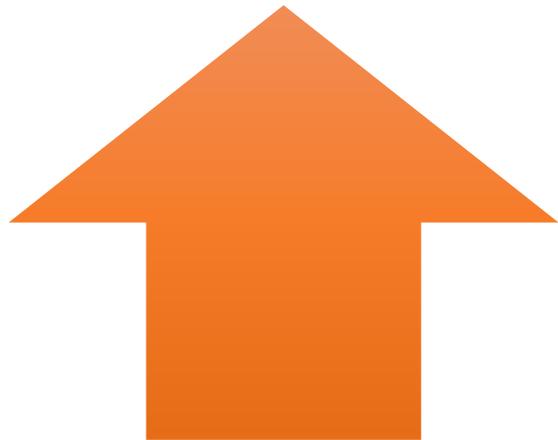


Getting access to centrally available data periodically and make it available to field offices on demand/ on an ongoing basis.

Capacity building activities – Advanced courses as well as supporting iCISA, NAAA, RTIs, Bilateral trainings



IT Audit Vs IT based Audit



IT audits

- Audit of IT system
- Data Analysis is significant component
- Checking of controls

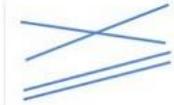
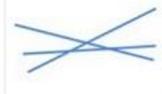
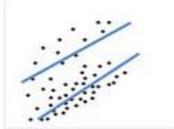


IT based audit

- After digital transformation nearly every PA/thematic/compliance audit is IT based
- Data Analytics may both form part of audit planning as well as audit execution.
- Data analytics should be followed by field validation.

ANALYTICAL PATTERNS

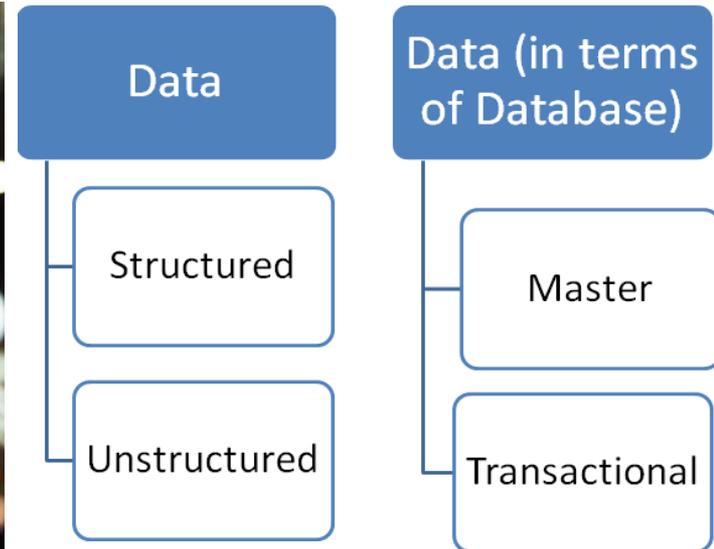
- Basic analytical patterns that we identify when looking at a visual

Pattern	Example	Pattern	Example
High, low and in between		Non-intersecting and intersecting	
Going up, going down and remaining flat		Symmetrical and skewed	
Steep and gradual		Wide and narrow	
Steady and fluctuating		Clusters and gaps	
Random and repeating		Tightly and loosely distributed	
Straight and curved		Normal and abnormal	



सत्यमेव जयते
Truth Alone Triumphs
in Public Interest

Types of Data





TYPES OF DATA

STRUCTURED

- Predefined format with defined data types
- Relational databases and Data warehouses
- Schema-dependent
- Estimated 20% of data

SEMI-STRUCTURED

- It has a self-describing structure that contains tags or attributes to separate various entities within data
- Can be coerced into useful and easy-to-leverage table formats
- Ex: XML, JSON

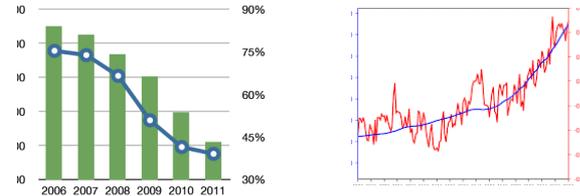
UNSTRUCTURED

- Raw/native format with varied data types
- NoSQL databases, Data warehouses, Data lakes
- Not schema-dependent
- Estimated 80% of data

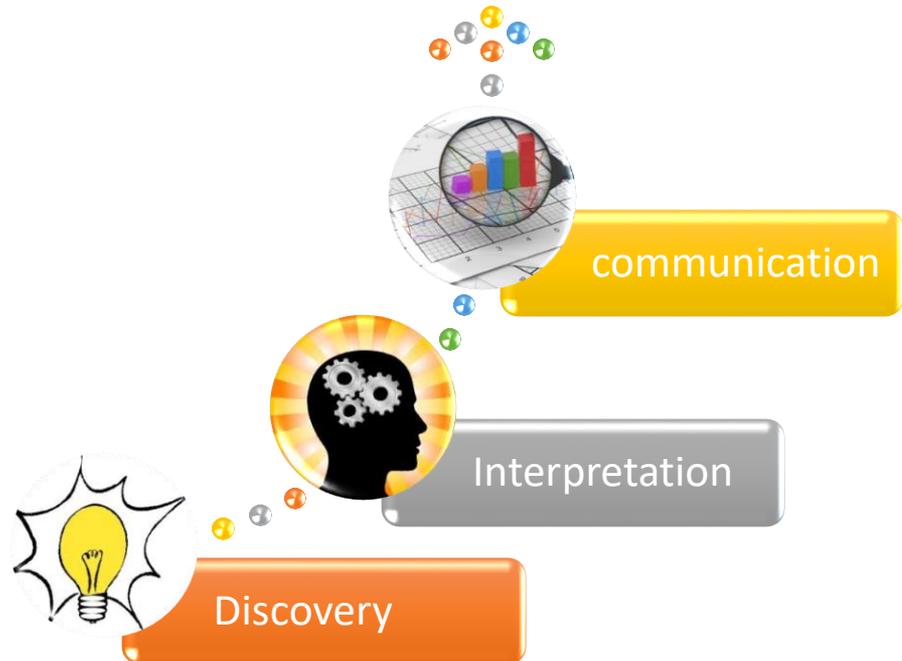
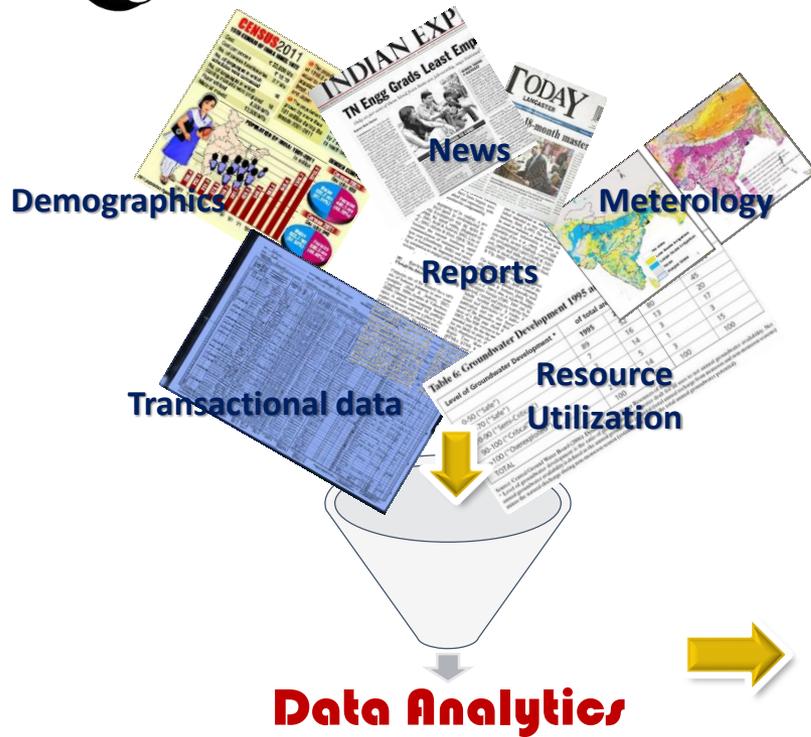


लोकहितार्थं सत्यमिच्छा
Dedicated to Truth in Public Interest

Meaningful patterns in data



What is data analytics?





Data Analysis Vs. Data Analytics

**Data
Analysis**

Narrow Scope

Case specific

Unidirectional

Querying for
exceptions

Rooted in the
past

**Data
Analytics**

Wide Scope

Open to Insights

Multidirectional

Mining for patterns

Futuristic



लोकहितार्थं सत्यनिष्ठा
Dedicated to Truth in Public Interest

Analytics Framework



Descriptive : What happened?

- Comprehensive, accurate and live data
- Effective visualisation



Diagnostic: Why it happened ?

- Ability to drill down the root cause
- Ability to isolate all confounding information



Predictive : What is likely to happen ?

- Predicting specific results using past data.
- Decisions are automated using algorithm and technology.



Prescriptive : What do I need to do ?

- Recommended actions and strategies to take
- Applying advance analytic techniques to give recommendations

*"Torture the data, and it will
confess to anything."*

*Ronald Coase, winner of the
Nobel Prize in Economics*



Big Data Analytics tools used in SAI INDIA

- Tableau
- KNIME
- Idea
- SQL
- R
- Python
- Power BI

Functionalities handled

- Handling of Big Data
- Data Cleaning
- Handling Missing Values
- Graphs & Charts
- Sampling
- Statistical Methods
- Machine Learning and Artificial Intelligence
- Table Joins/Relations
- Data Modelling
- Geo Data
- Audit Trail and Work flow
- Importing ERP Reports/Report Reader
- SAP ERP Connector



Comparative features of tools used

Functionalities	Tableau	Knime	Idea	SQL	R	Python	Power BI
Handling of Big Data	✓	✓	✓	✓	✓	✓	✓
Data Cleaning		✓			✓	✓	✓
Handling Missing Values		✓	✓		✓	✓	✓
Graphs & Charts	✓		✓		✓	✓	✓
Sampling		✓	✓		✓	✓	
Statistical Methods		✓	✓	✓	✓	✓	
Machine Learning and AI					✓	✓	
Table Joins/Relations	✓	✓	✓	✓			
Data Modelling	✓				✓	✓	✓
Geo Data	✓				✓	✓	✓
Audit Trail and Work flow		✓	✓		✓	✓	
UI based Tool	✓	✓	✓				✓

Common data analysis one should know

Appending/Merging Data Sets

Aggregation or Summarization

Duplicate Beneficiaries

Sampling Methods

General Statistics

Developing Heat Maps/Choropleths

Using Google earth to plot coordinates

Data Preparation-Financial Year

Joining of Datasets

Data Cleaning - Validating Date Fields

IP Address to Location

Aging analysis

VPN - Usage and Applications

Use of Fuzzy Matching

Gap detection

Data Analytics in Field Audit Offices



Data Collection

Datadump instead of queries/tables

(Not MIS data)

Unmasked PII

UID Token (72 digit alphanumeric key) – 1-1 mapping for that auditee/Aadhar user agency



Analytics

Data entry methods used by auditee

Replicability

Open Source Tools (R, Python, KNIME, etc.)

Advanced analytics

Reverse Geocoding

Geo Mapping

IP location look up technology

Third Party data sets

Pattern matching (regex)

Network of Duplicates



Field Validation

Field audit of sample cases is required, e.g.

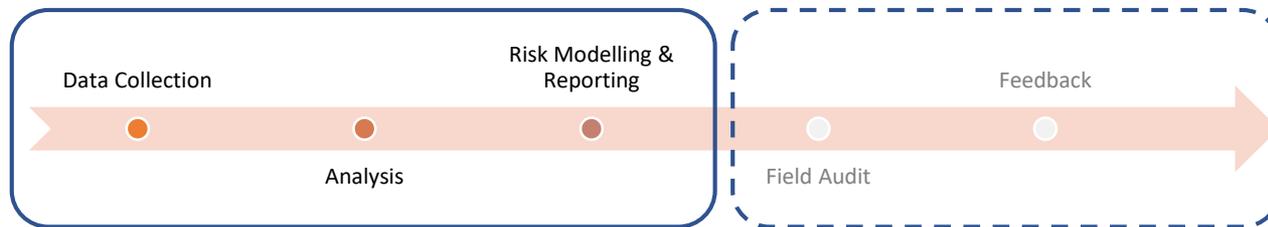
Reasons for high number of approvals in odd business hours

Reasons for pendency/delay

Reasons for non-matching of geo-locations



Data Driven Audits



Data Collection and preparation– First and most important step in data driven audits. Includes both system and data flow understanding. In India, data quality is a major issue for which we have to put 30% - 50% of time.

Analysis – Data Analytics done after interaction with IAAD domain experts, auditee and going through literature for targeted as well as exploratory analysis.

Risk Modelling – Moving away from random sampling, using exception analysis, outlier detection etc. in data analytics, we are developing risk models for selection of high-risk entities and transactions.

Typical areas for data driven audits

- Social Sector Schemes
 - e.g. MGNREGA; PMAY; NSAP; PM Kisan Samman Nidhi; Ayushman Bharat etc.
- Central Departments (Taxation), Railways (Ticketing) etc.
- Central & State PSUs
 - ERP systems audit, Operational activities; financial reporting
 - Electricity Boards/Companies
- Finance Management Systems
 - Public Financial Management System (PFMS), Integrated Financial Management & Information Systems (IFMIS),
- State Receipts
 - Registration and Stamp Duty
 - Motor Vehicle Department
- E-Procurement/ e-Tendering
- Other e-governance systems; Local Bodies



लोकहितार्थ सत्यनिष्ठा
Dedicated to Truth in Public Interest

SIX FUNDAMENTAL DESIGN PRINCIPLES

- Show comparisons
- Show causality
- Use multivariate data
- Completely integrate modes
- Establish credibility
- Focus on content



Professor at Yale University - Political science,
Statistics, and Computer Science

1983: The Visual Display of Quantitative Information
1990: Visual Explanations: Images and Quantities,
Evidence and Narrative
1990: Envisioning Information
2006: Beautiful Evidence



Current Infrastructure

Available

- Cloud storage (100 TB on NIC cloud)
- On Premise Storage 50 TB NAS
- On-premise 6 TB Server
- Software/tools – a mix of proprietary (tableau) as well as open-source (R, python, Knime, etc.)

New Infrastructure being deployed

- New high compute laptops
- Two high-end computing machines desktops
- Two additional servers – with Red Hat Enterprise (RHEL) and Windows licenses
- Software/tools – Tableau extensions



What kind of datasets are we looking into?

Flat files

Relational Database

Big data tables



Objective of Analysis

Exploratory Analysis

Statistical Analysis

Exception Analysis

Data visualisation, Reports



As a rule of thumb, no one analytics tool is superior than the other. It depends on the use case.

Selection of Analytics tool

Challenges & Response..

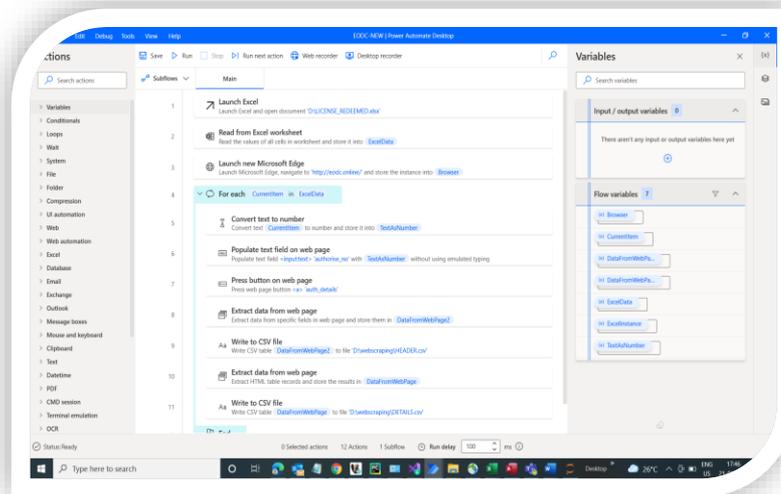
Challenges	How SAI INDIA overcame the challenges
IT systems in Governments not ensuring end-to-end computerization	<ul style="list-style-type: none"> ▪ Phase I audit primarily focusing on data analysis and Phase II field audit to validate findings of Phase I to ensure any data discrepancy or incompleteness.
Diversity of IT systems , data sources and their non-integration	<ul style="list-style-type: none"> ▪ Technical infrastructure and human resource to handle various tools/techniques ▪ Documentation, interaction with technical team of auditee, access to data dictionary.
Lack of skills	<ul style="list-style-type: none"> ▪ Capacity building through RTIs, ICISA & CDMA ▪ Technical support through CDMA and consultants
Timely Access to databases	<ul style="list-style-type: none"> ▪ Agreements with Ministry to provide periodic data. ▪ Senior Management intervention wherever required.
Data privacy	<ul style="list-style-type: none"> ▪ Role based access ▪ Masking of private data as per Acts & Rules
Data Completeness and reliability	<ul style="list-style-type: none"> ▪ Control Totals ▪ Written assurance from Ministry ▪ Matching figures from MIS portal

Alternative Data & Audit



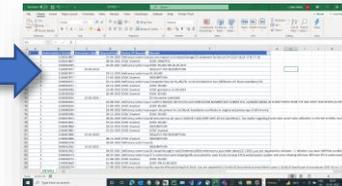
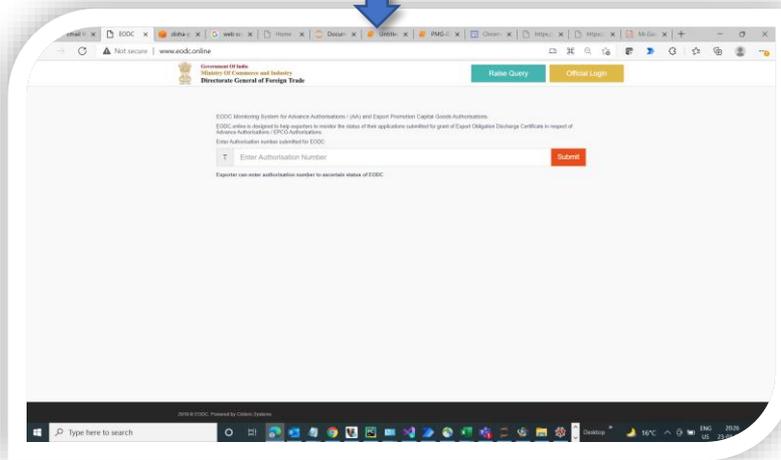
लोकहितार्थं सत्यमिच्छा
Dedicated to Truth in Public Interest

Web scraping –
Power Automate as
Web scraping tool.
We also use Python
extensively for web
scrapping.



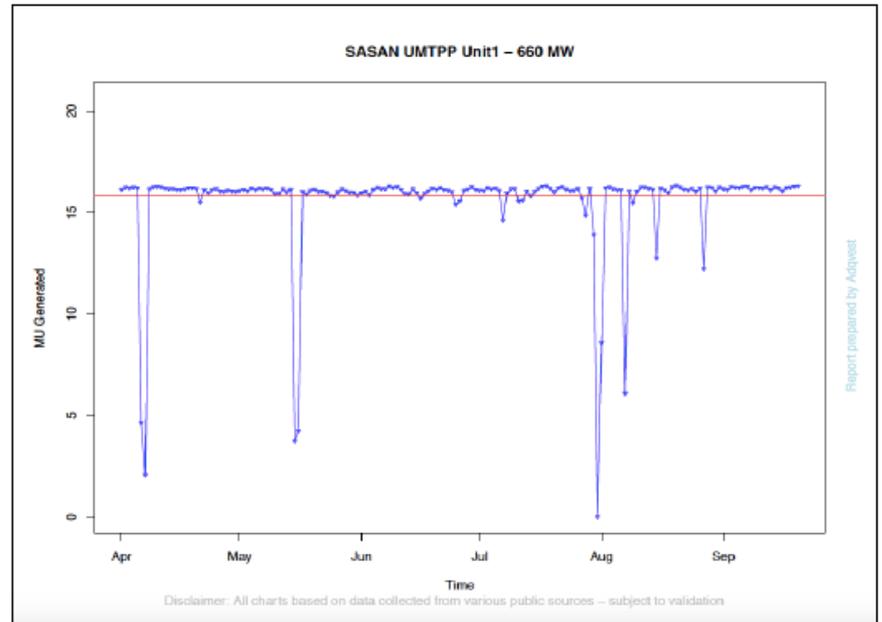
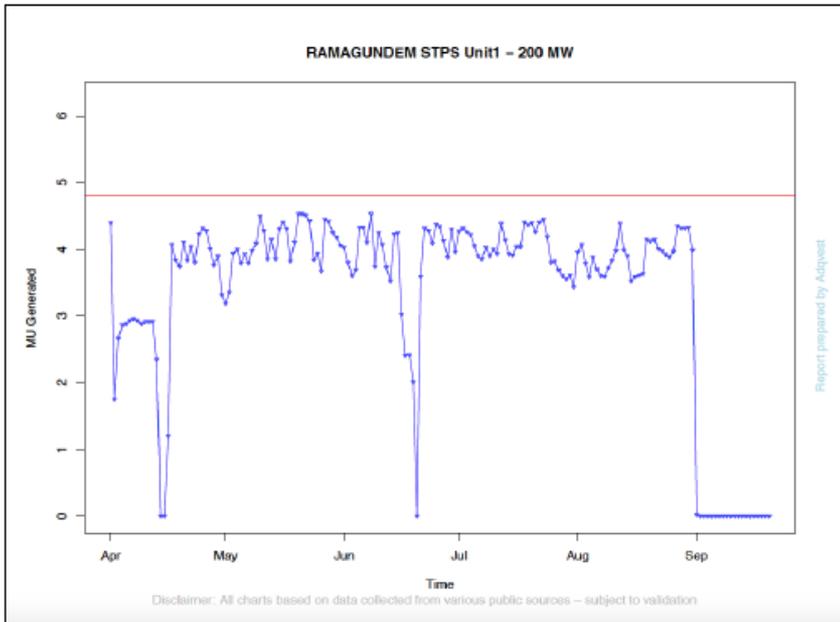
INPUT
License_Redeemed
.xlsx
(containing license numbers)

OUTPUT
Header.csv
(containing license
authorisation details)
Details.csv
(containing license workflow
details)



Example – Use of publicly available data for generation of leads in audit (CEA)

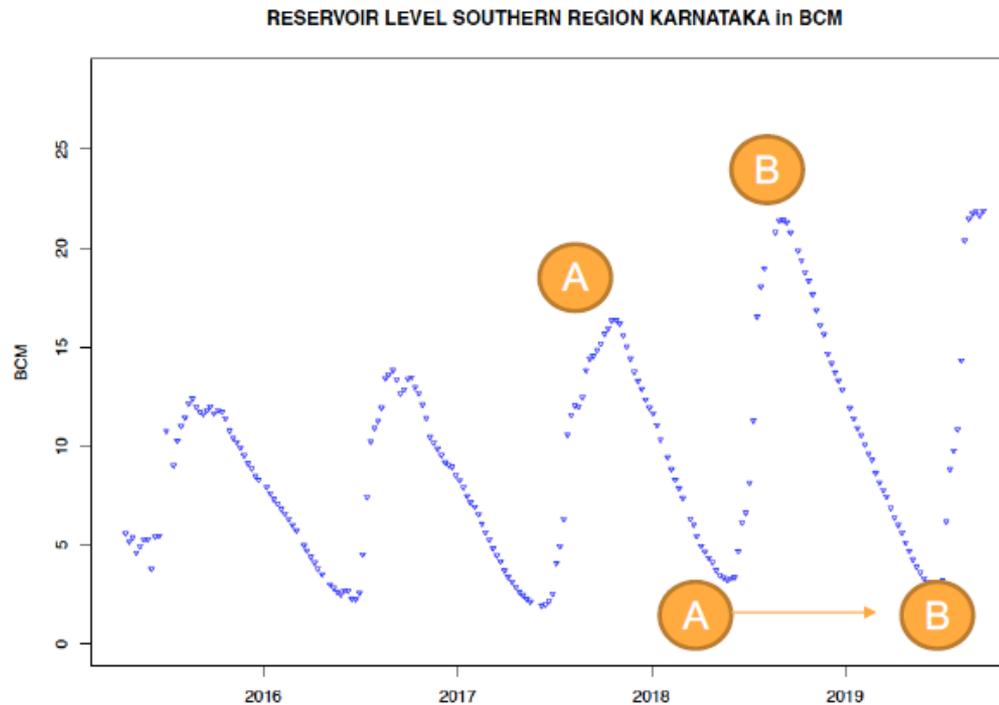
Comparing performance of TPS - Ramagundem TPS vs Sasan TPS



In the last 6 months, Sasan TPS has performed far better than NTPC Ramagundam

Another Example

Using weekly published reservoir data to identify usage patterns



Regardless of the reservoirs level, the zero level at depletion remains the same



लोकहितार्थं सत्यमिच्छा
Dedicated to Truth in Public Interest

Audit Findings/Insights from Data Analysis

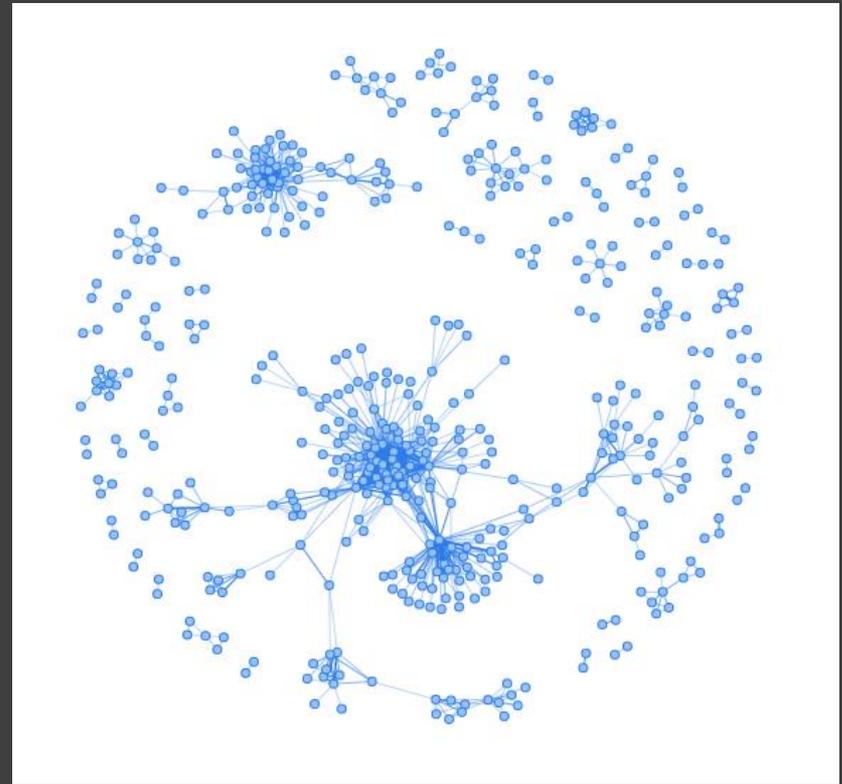
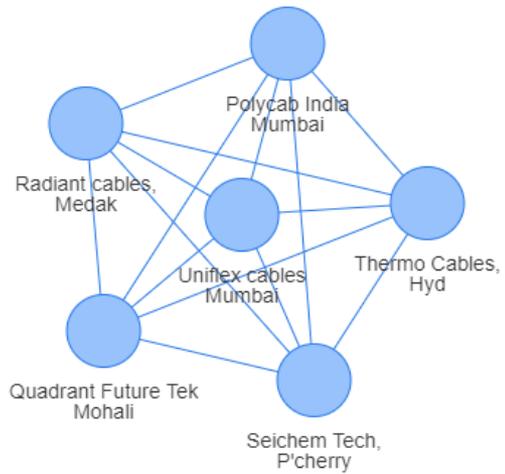
Examples

Image Analytics

Python was used to identify
duplicates/similar images used for
multiple beneficiaries

Differentiation between metadata and
scanned images

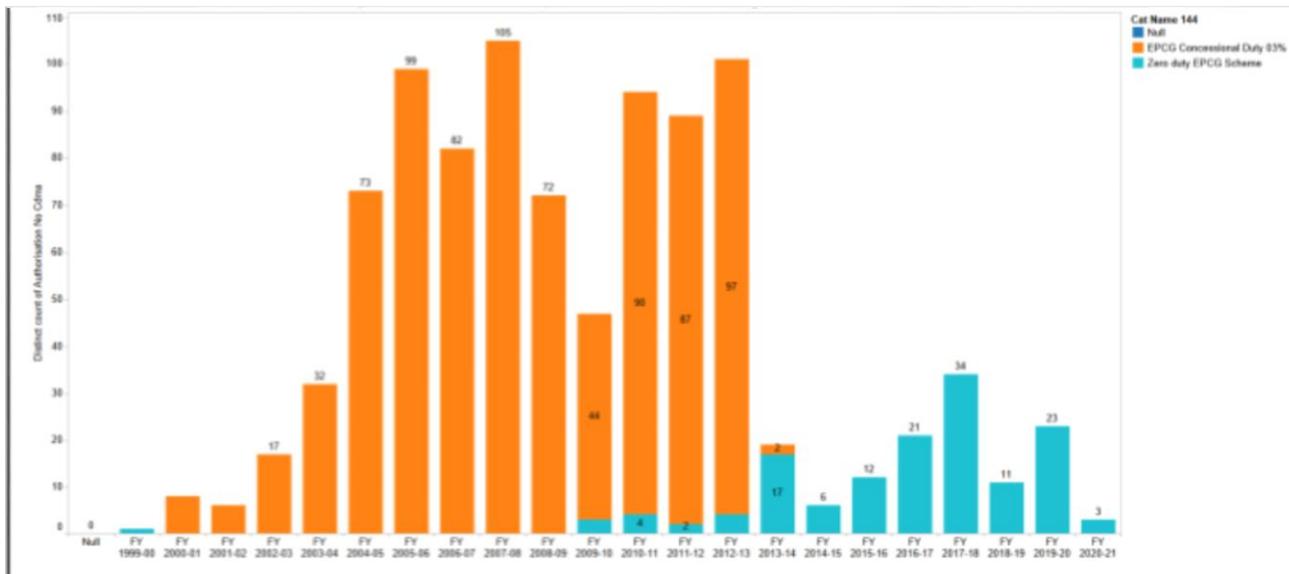
Essential for false inclusion in beneficiary
universe



Network Analysis



Cross checking with VAHAN data revealed that Vehicles imported under EPCG Licenses were not registered with VAHAN- possibility that vehicles were either not imported or may be re-sold



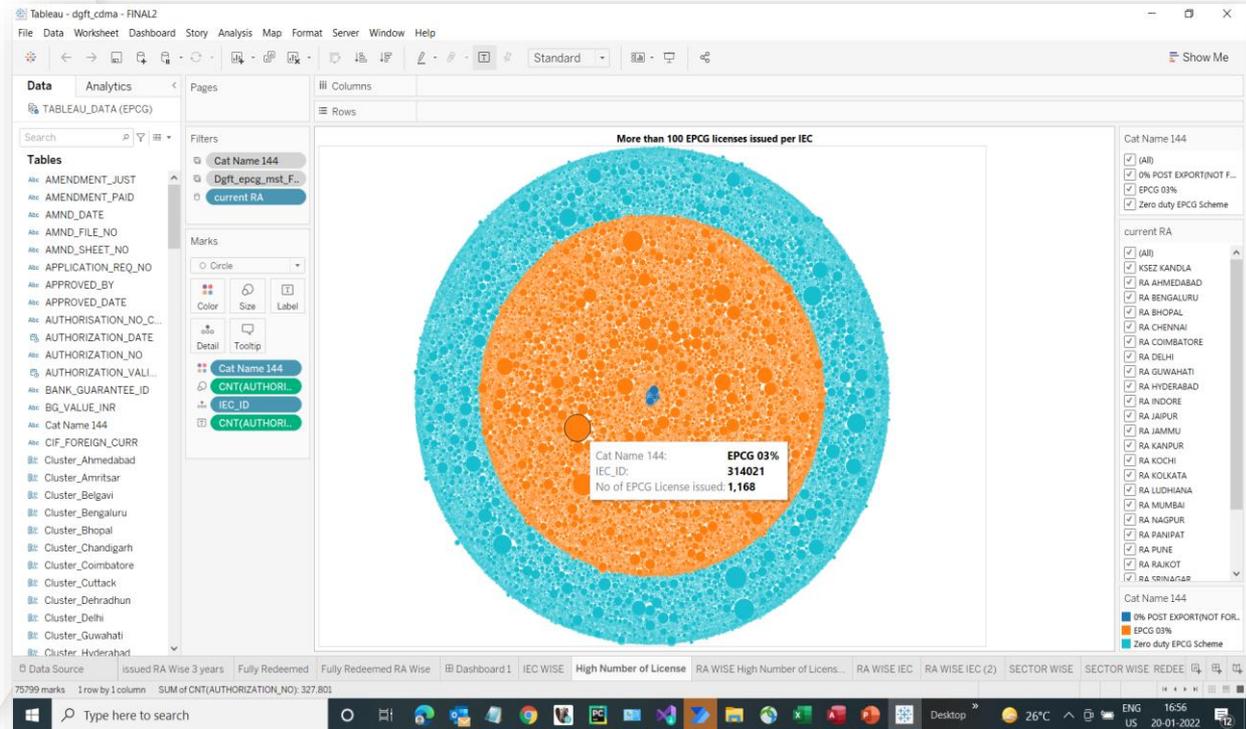


Data visualization using tableau

The bubble chart shows the number of licenses issued. The size of the bubble indicates the number of licenses.

High number of licenses issued to a single entity

(Detection using bubble chart in tableau)





लोकहितार्थं सत्यनिष्ठा
Dedicated to Truth in Public Interest

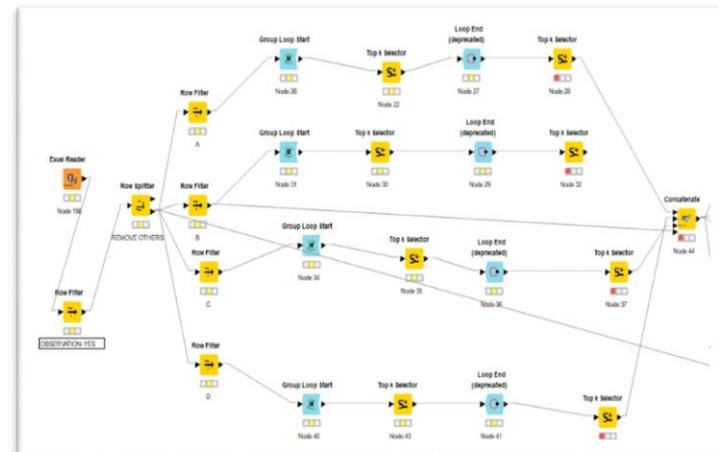
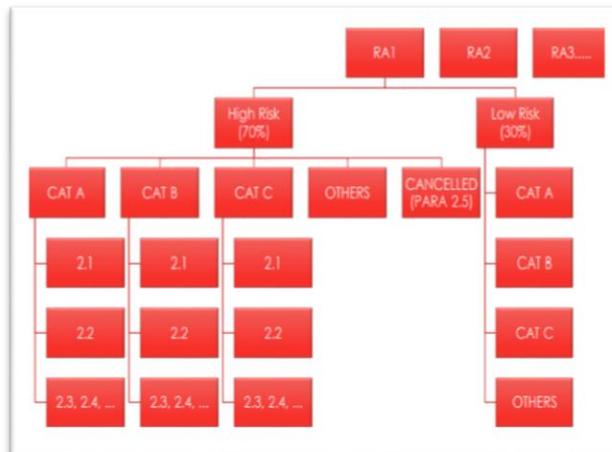


Risk Model for selection of samples for field audit was prepared in KNIME - EPCG

Category for Sampling of cases for Audit

implemented

KNIME Model for selection of cases



RA1, RA2, RA3 – various Regional Authority selected for audit
CAT A – Issued cases, B-EODC cases, C- Non-Redeemed cases
2.1, 2.2, 2.3 – Insights identified in Analysis

Figure No. 3.4: Operation of poclains in Neyvasal quarry

8 June 2015 -7 Poclains

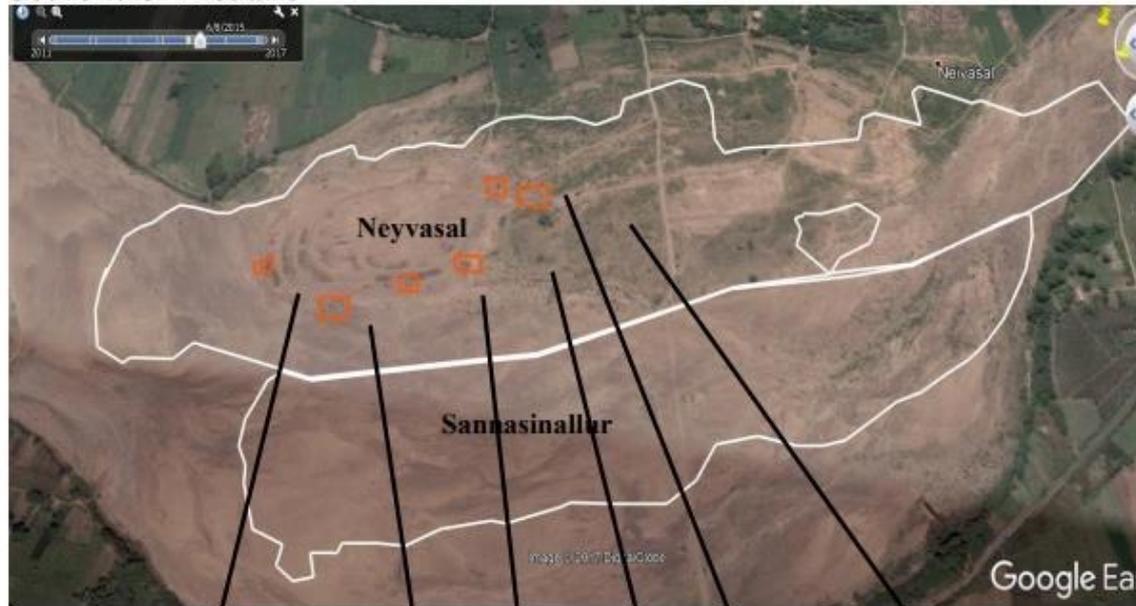


Figure No. 3.7: 3D image showing height difference 5.90 m



Figure No. 3.8: 3D image showing height difference 6.23 m



Figure No. 3.6 UAV Ortho Image as on - 03 November 2017



Ending Remarks



Technology provides opportunities for SAI to enhance their performance and ensure continuity of operations in emergencies.



In the new normal times **SAI could look to build its off-site auditing capability by utilizing technology** and suitably trained human resource to use the technology along with the domain expertise of auditing a particular sector.



Audit work, used to be conducted by auditors on-site manually can be **made impactful by off-site data analyzing teams**. SAI could benefit from conducting data analytical research by reducing field work.



An audit tool by itself intended for off-site use will not help in good audit but a **combination of the tool, domain (subject matter) expertise, IT team and infrastructure** to access the data could make a positive change.