KAPSARC
مركز الملك عبدالله للدراسات والبحوث البترولية
King Abdullah Petroleum Studies and Research Center

Value of data transparency

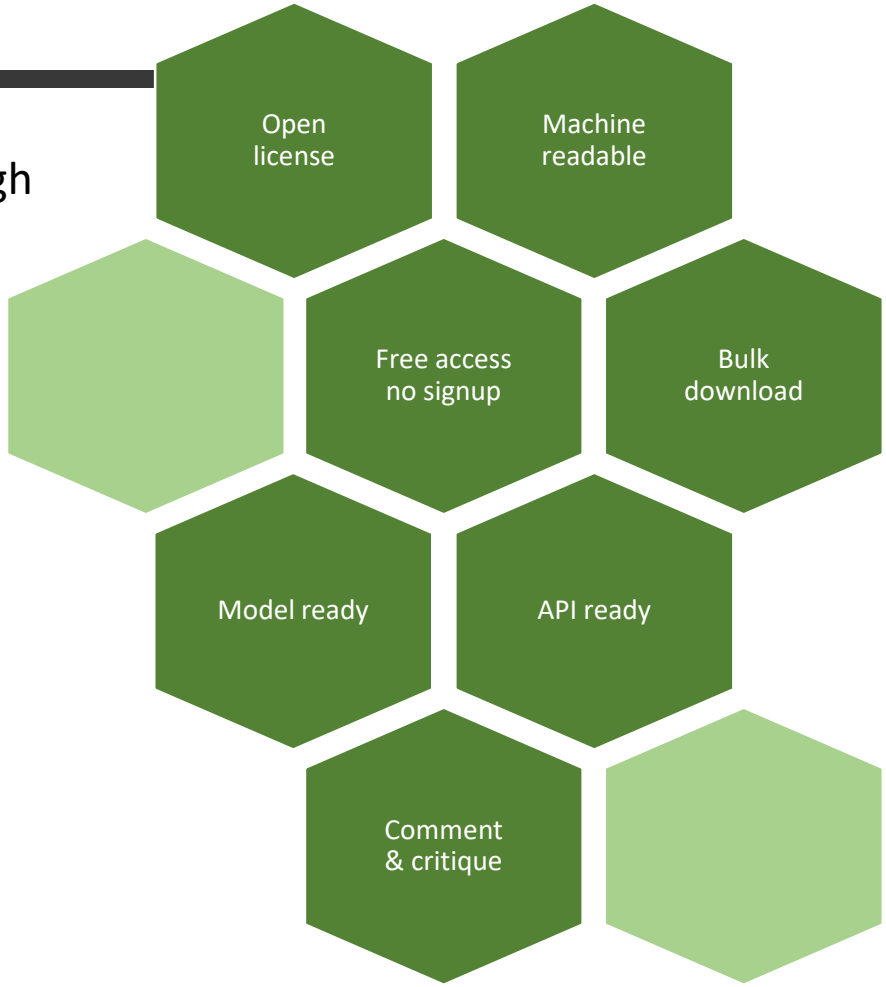
from an energy research perspective

Amar Amarnath,
Head of Data Management
Oct 2019

Is data transparent?

Meaningful insights are only possible with high quality data

- ❑ Is openly licensed?
- ❑ Is machine readable?
- ❑ Is free of charge?
- ❑ Is available in bulk?
- ❑ Is up-to-date?
- ❑ Is online?
- ❑ Is digital?
- ❑ Is available?
- ❑ Is publicly available?



To what extent have new JODI data dissemination tools been effective? What are the views of data JODI world databases redistributors after two years?

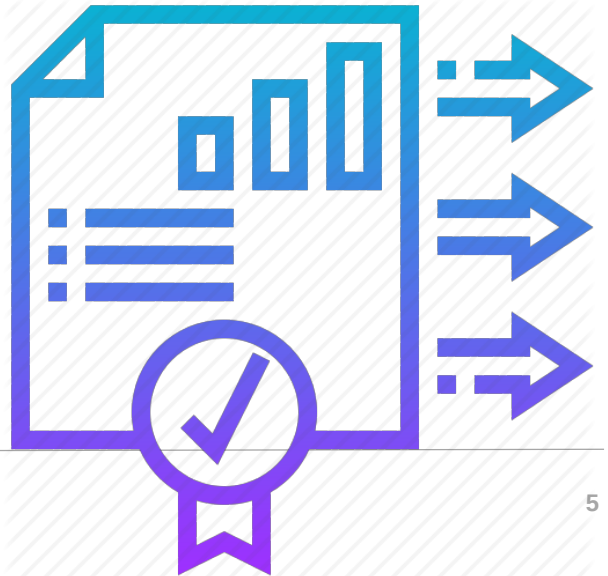
- 60K searches and 600 downloads from KAPSARC portal
- API access, as artificial intelligence and new predictions models advancing, structured machine readable data for insights is key
- Make data social, so we can capture users comments

How can the needs and benefits of energy data transparency be more effectively communicated?

- Users to subscribe for alerts
- Mobile and API first, browser native vs download software (20/20 beyond)
- Allow users to request for new dataset

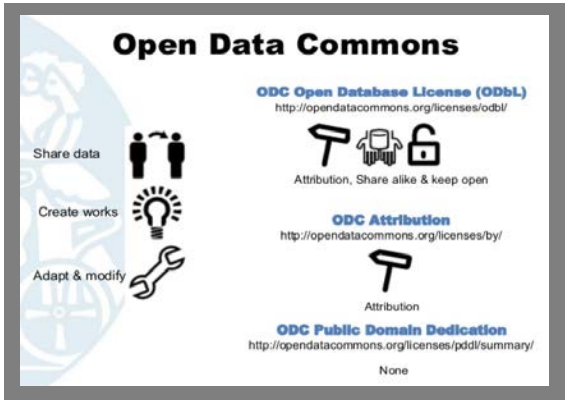
How can JODI further improve the confidence of data users on the data quality of JODI data?

- Report quality metrics
  - Correctness Validating data accuracy relative to external reference points
  - Currency Delivering new and updated content in a timely manner
  - Completeness Providing the right data attributes and analysis
  - Consistency Standardized identities, definitions and content across datasets
- Meta-data completeness and data recipes



How can JODI improve granularity and data context notes to improve transparency and its value?

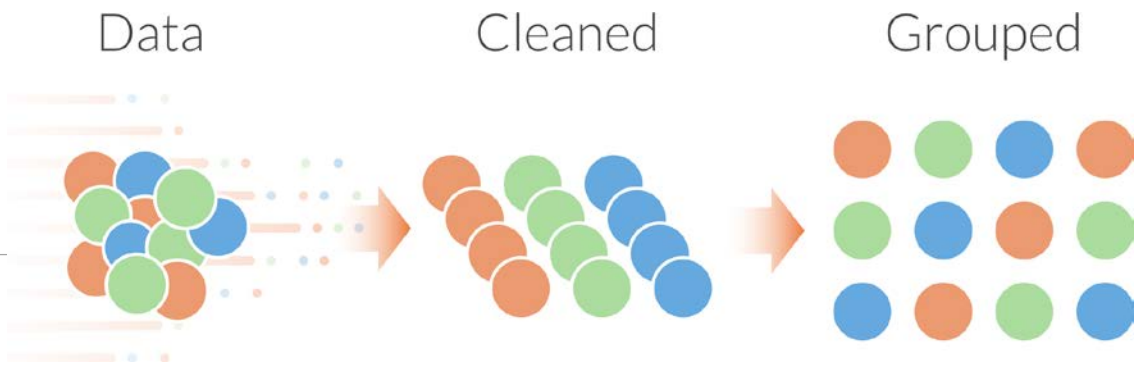
- Great having monthly granular data
  - Modelers need sector and regional granularity
  - Future will pivot on API/IOT data streams
- Make data social – allow discussion threads on data
- Add easy to share data web links and web widget
- Specify “data use” standards; share, create, adapt, attribute
  - Creative commons has 7 types
    - Public domain, Open data license (attribute, share alike, no commercial, no derivative)
  - Open data commons has 3 types
    - Odbl - attribute, share alike, keep open
    - Odc – attribute
    - Opdd - none



How can the uniqueness of the JODI data reporting mechanism be emphasized?

- Report, rank and highlight data quality
- Highlight the data recopies and transformation process

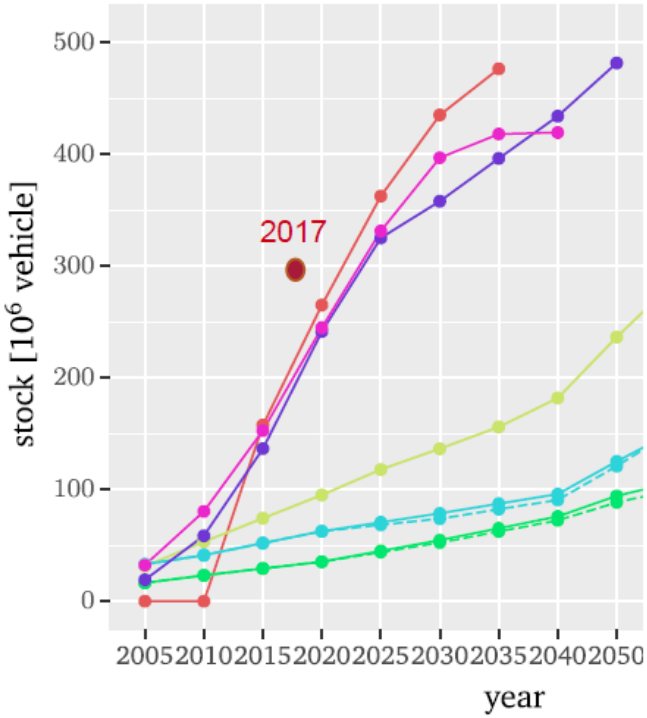
  “Raw data -> Match -> Identify -> Relationships -> Analysis -> Forecast -> Insight -> Action”
- List energy models and organizations that use Jodi data, call for modelers to add the tag “powered by JODI data”
- Data value messaging and training data wranglers to improve quality



How can efforts of data providers be effectively recognized?

- It’s a common practice for researchers to list author name in papers/articles, promote and recognize data champions as “data wrangler”
- Document, share and market - use case and benefits of data.
- Data champions can see the line-of-sight importance of their output to policy makers, market, academia and research

Researcher challenge example #1 : Trust on research output



- Huge uncertainty about China: China's LDV stock
- Will there be 90 million cars or 500 million cars in China by 2050?
- Large implications on vehicle stock, fuel prices, urban planning, refineries investment, etc.

Researcher challenge example #2: Variation in our feedstack - data

How Much Do People Travel? How Certain Are We?

Thousand KMs/capita/year, all modes

	Australia	Brazil	China	U.S.
Energy Modeler 1	26.2	5.4	5.0	23.1
Energy Modeler 2	21.7	8.3	8.4	26.6
Energy Modeler 3	33.7	6.9		15.2
Energy Modeler 4	43.5	5.4	4.5	27.7
Energy Modeler 5		4.2	6.3	19.7
Energy Modeler 6	17.5	8.4	6.5	26.8

Researcher challenge example #3: Takes too long, missing monthly

Challenges Faced and Action Taken (I)

It took at least 3 months to finalise the energy consumption and CO_2 data that involved numerous email/telephone exchanges with the admin staff and the people employed working on the data due to the following:

- Historical end points were incorrect for a number of series
- Some of the series were interpolated from annual figures and the utilised procedures had produced negative values
- Missing values introduced in error
- No documentation of data sources, definitions and procedures (including interpolation method and factors used to constuct CO_2 emissions)
  - In the end we had to drop CO_2 emissions because it was simply constructed as a weighted average of the consumption data provided and hence the two were highly correlated

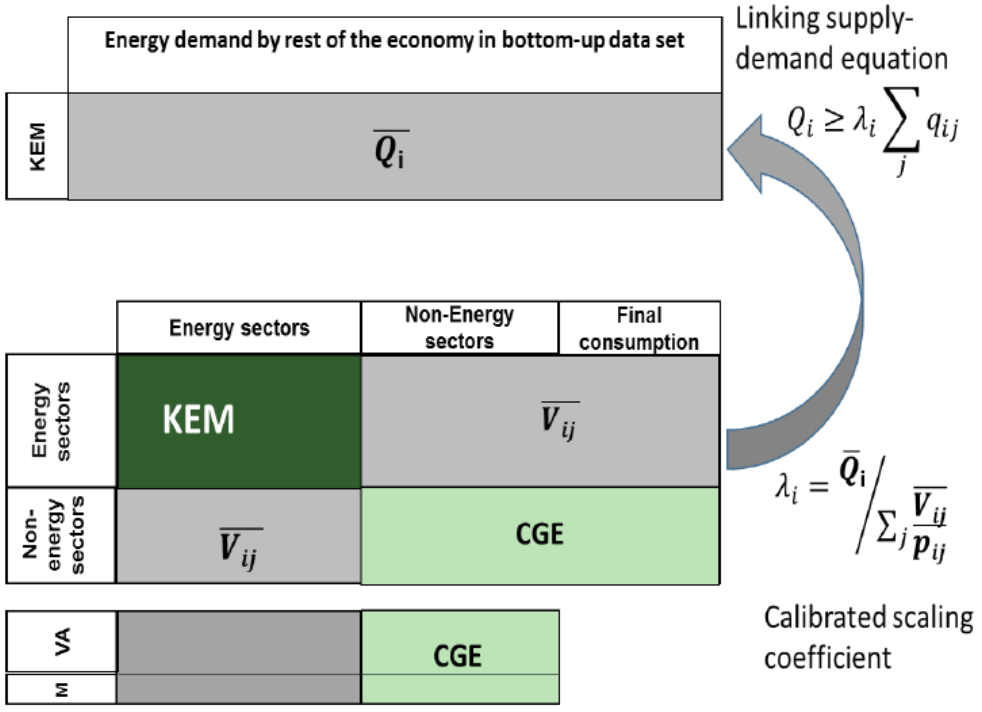
Challenges Faced and Action Taken (II)

- For energy prices quarterly coverage was very poor (Datastream) or included many gaps and a lot of missing series (International Energy Agency)
  - Had to use world energy prices instead of country-specific prices
- The energy consumption data was not temperature adjusted
  - Spent 3 months to download and construct degree days from temperature data at the quarterly frequency
  - KAPSARC's global degree days database for energy-related applications covering 147 countries over a period of several decades could have been useful for our purposes, which I recently became aware of

Researcher challenge example #4: Sector level data, granularity

KEM – G: Data challenges

- **Data availability:**
  - Social Accounting matrix
  - Disaggregated data (energy consumption, sectors)
- **Inconsistencies between data sets**
  - But top-down and bottom-up data don't always match…
- **Data management**
  - Update the calibration year
  - Linking the model to other energy models (KEM, KEM-GCC,…)



Dealing with the data differences: Example of domestic energy demand

Researcher challenge example #5: monthly data please!

- Large-scale administrative data sets and proprietary private sector data can greatly improve the way we measure, track, and describe economic activity.
- Models are constructed under severe constraints of limited availability of data, mostly compiled annually

Serious reduction in degrees of freedom and diagnostic testing.

- Limited policy relevance of model results due to lack of disaggregated monthly, quarterly data
- Lack of disaggregated regional data
- The choice for appropriate lag length is a challenging due to the shortness of annual time series and the relative absence of monthly and quarterly data for most macro-variables.

Modeling Process

- **Developing a New Model**
  - Model Objective
  - Data Availability
- **Improving and Maintaining**
  - Backward compatibility
- **Archiving**
  - Data and model archiving
  - Version controlling and annotations

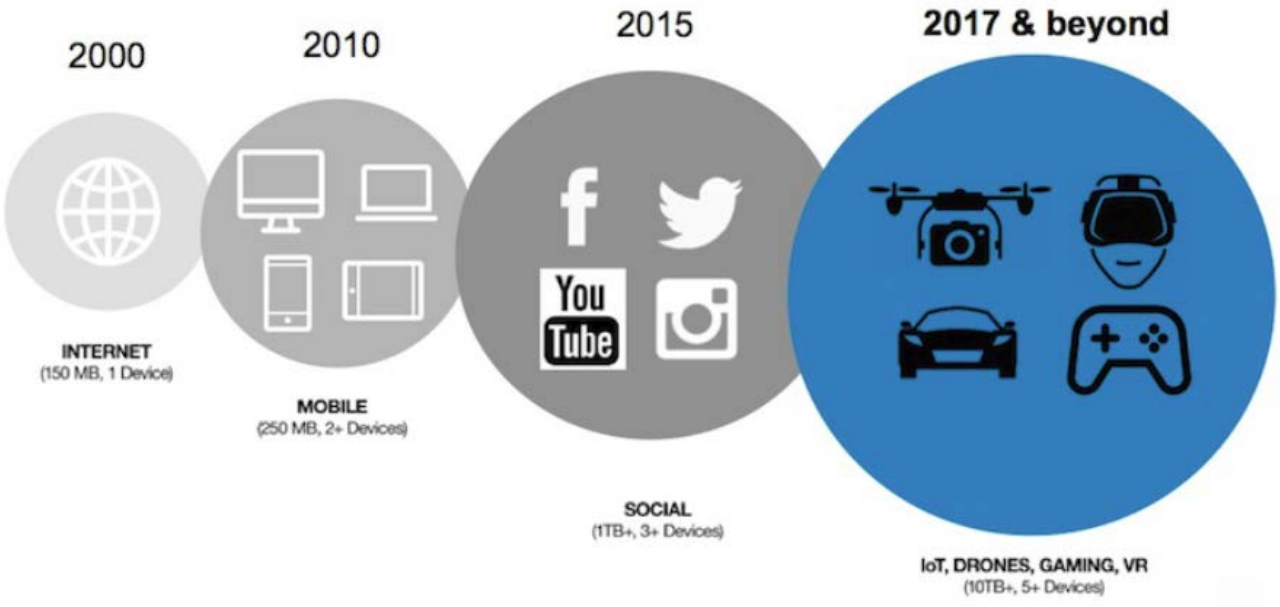
Data Issues for New Model

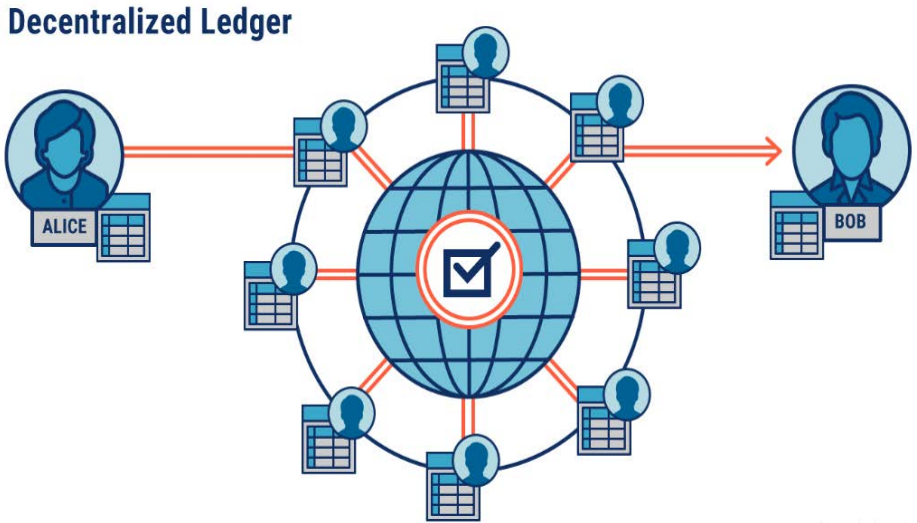
- **Availability**
- **Compatibility with the model usage**
- **Data generation**
- **Discrepancy**

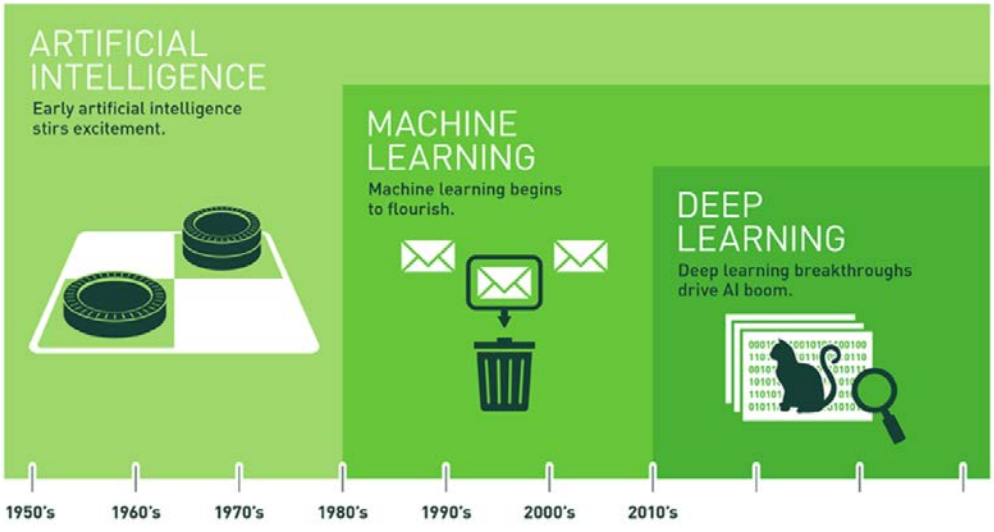
Data Issues for Model Improvements and maintenance

- **Model Maintenance**
  - Updating data
  - Adding time series
- **Model Improvement**
  - Replacing inconsistent and generated data
  - Regenerating for new features
- **Model Archiving**
  - Versioning and descriptors

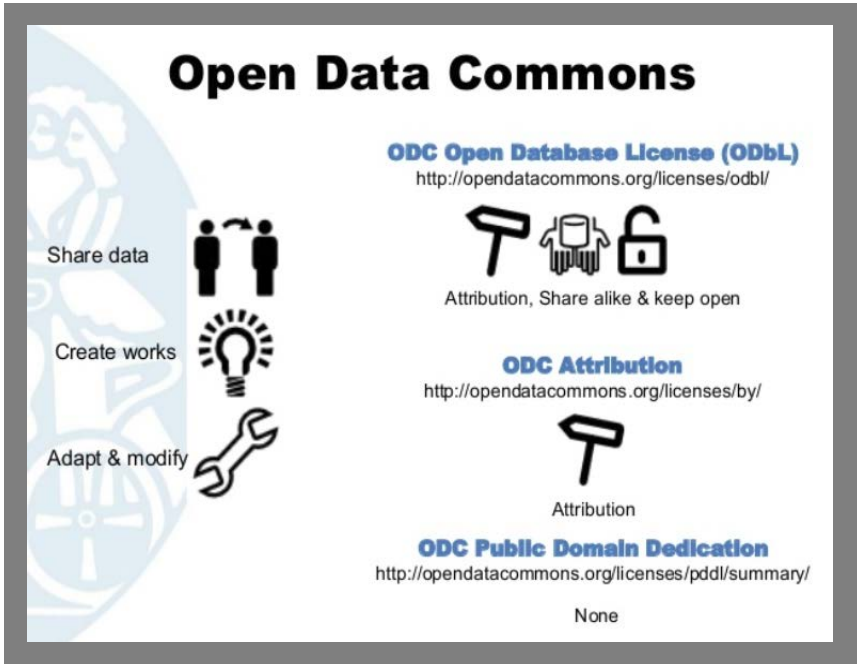
A minute into future where data and prediction landscape…

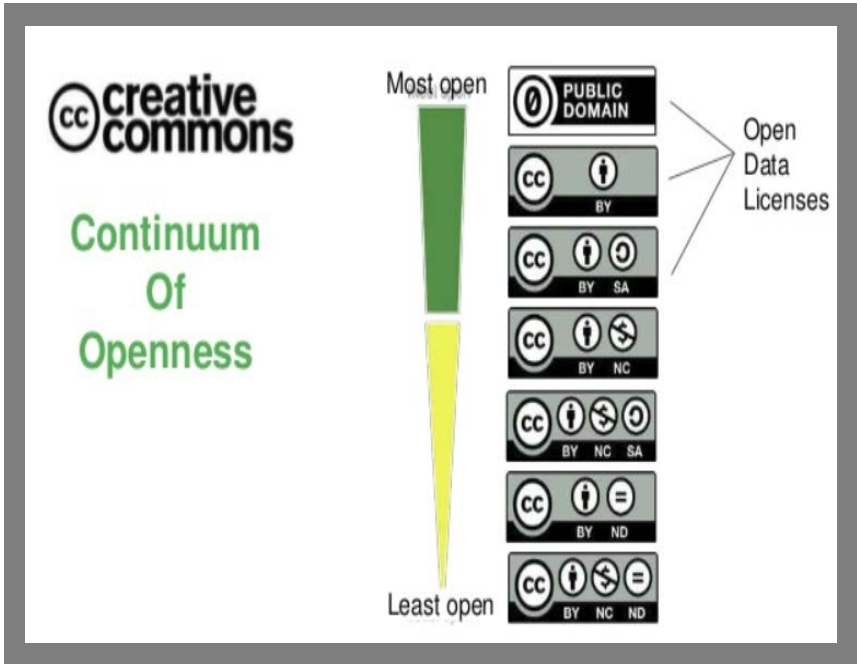






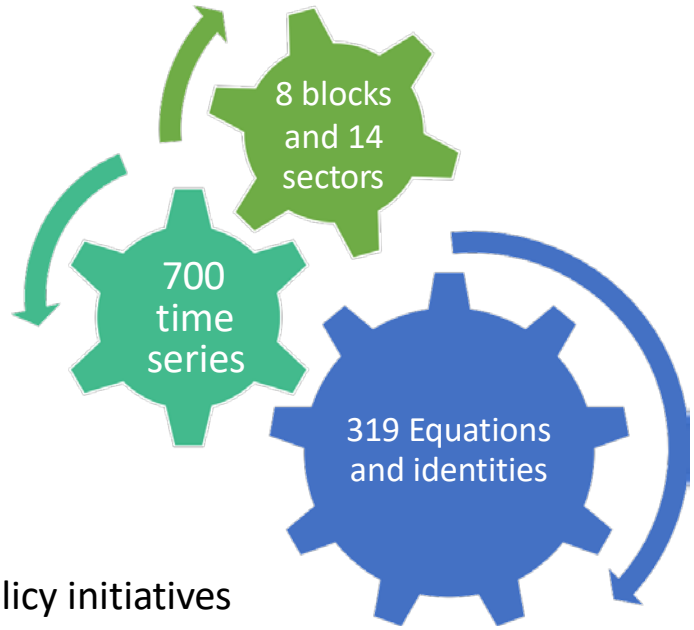
Data owners need to specify “data use” standards, soon technology will aid





Data Challenge – Use case KAPSARC Global Energy Macro-econometric Model

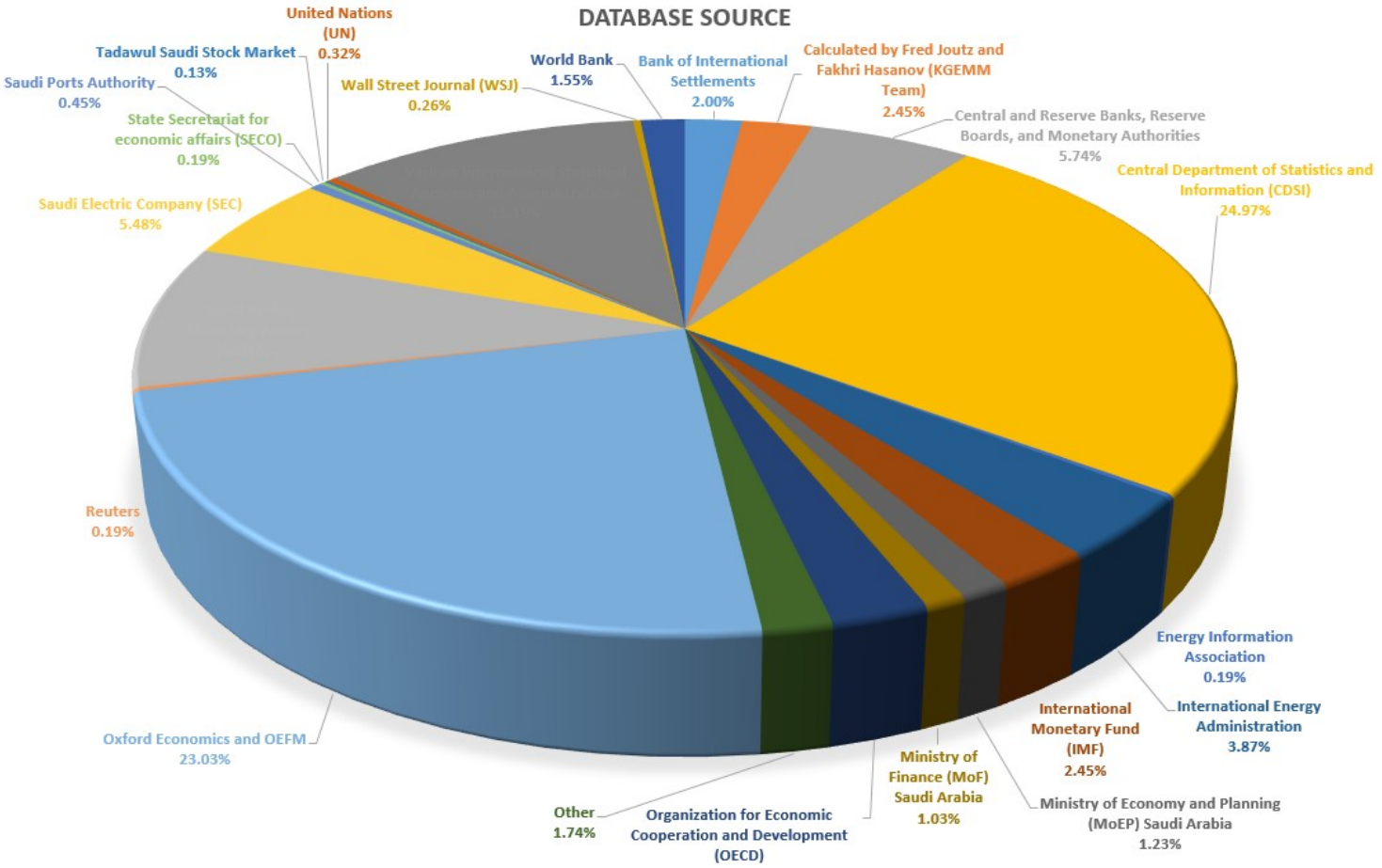
- **Objective**
  - A policy analysis tool examining the impacts of policy decisions and the interaction between the global economy and macroeconomic energy environment of Saudi Arabia.



- **Use**
  - To evaluate the impacts of different policy initiatives and macroeconomic targets
  - Analyzing and forecasting behaviors of energy and macroeconomic indicators

Models' data sources

*There are more than **700** time series aggregated and disaggregated energy and economic variables from **22** sources as of 2019*



Researcher Data Challenges

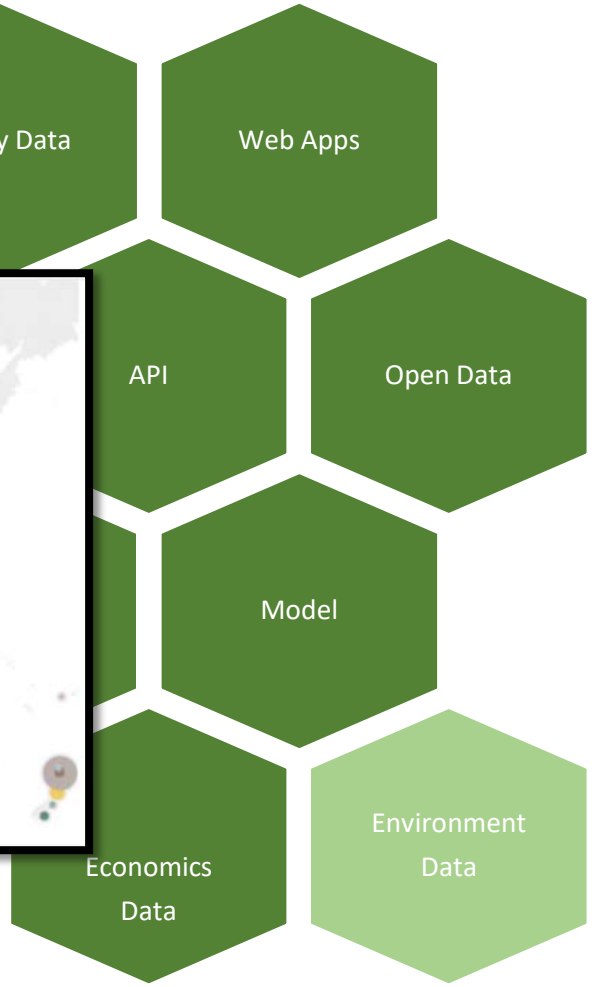
- Data revision without alerts
- Archive old data not publically available
- Unavailability of disaggregated data
  - Energy consumption and prices by sector and customer segments
  - Sectoral investment, sectoral employment by nation and gender

Consequences without official statistical data

- Out of date representation of the country economy
- Absence/improper of granular relationships in the economy
- Unable to conduct short-term policy analysis and projections
- Unable to represent the economy in a meaningful way

Energy Information Management - Data, Web Apps, Models

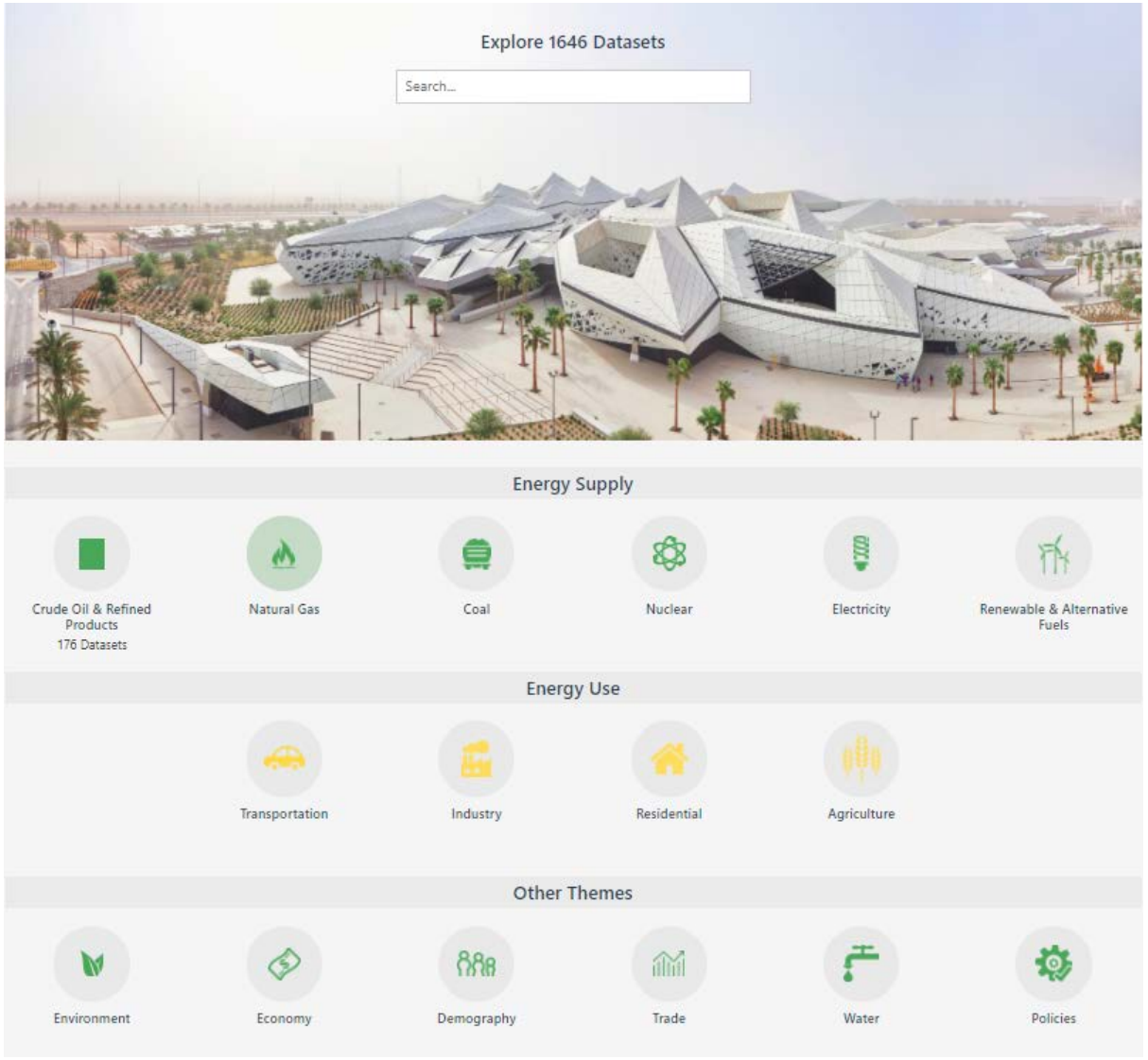
- Develop web portal hosting “data, models and algorithms”
- Acquire, wrangle and delineate data from models
- Aggregate and make machine readable open data to world
- Publish data insights showcasing changes and trends
- Develop web applications to showcase models

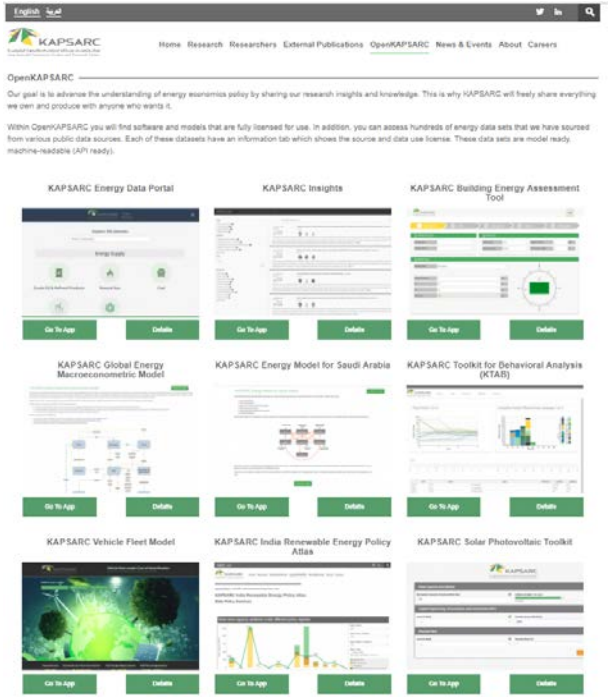


- 50M records, 3K indicators, 10 years time series data
- 45 days average data currency from 170 data sources
- 6000 searches and 40 downloads per day

OPEN KAPSARC www.kapsarc.org https://datasource.kapsarc.org

features data and web applications to run scenarios on KAPSARC models.





KAPSARC OpenTools

English العربية Subscribe

KAPSARC
مركز الملك عبدالله للدراسات والبحوث البترولية
King Abdullah Petroleum Studies and Research Center

Home Researchers Published Articles OpenKAPSARC News & Events About Careers

OpenKAPSARC

The OpenKAPSARC platform provides free access to KAPSARC's data and research tools, to advance the understanding of energy economics and environment policy. It contains over 1000 machine-readable (API ready), model-ready datasets, 3000 indicators and over 30 million records.

Web Apps | Developer Tools | Data Insight

KAPSARC Energy Data Portal

Go To App | Details

KAPSARC Insights

Go To App | Details

KAPSARC Maritime Transport Analysis Framework

Go To App

KAPSARC Building Energy Assessment Tool

Go To App | Details

KAPSARC Global Energy Macroeconometric Model

Go To App | Details

KAPSARC Energy Model for Saudi Arabia

Go To App | Details

Researcher data challenges aligned to model and insight development

Data quality

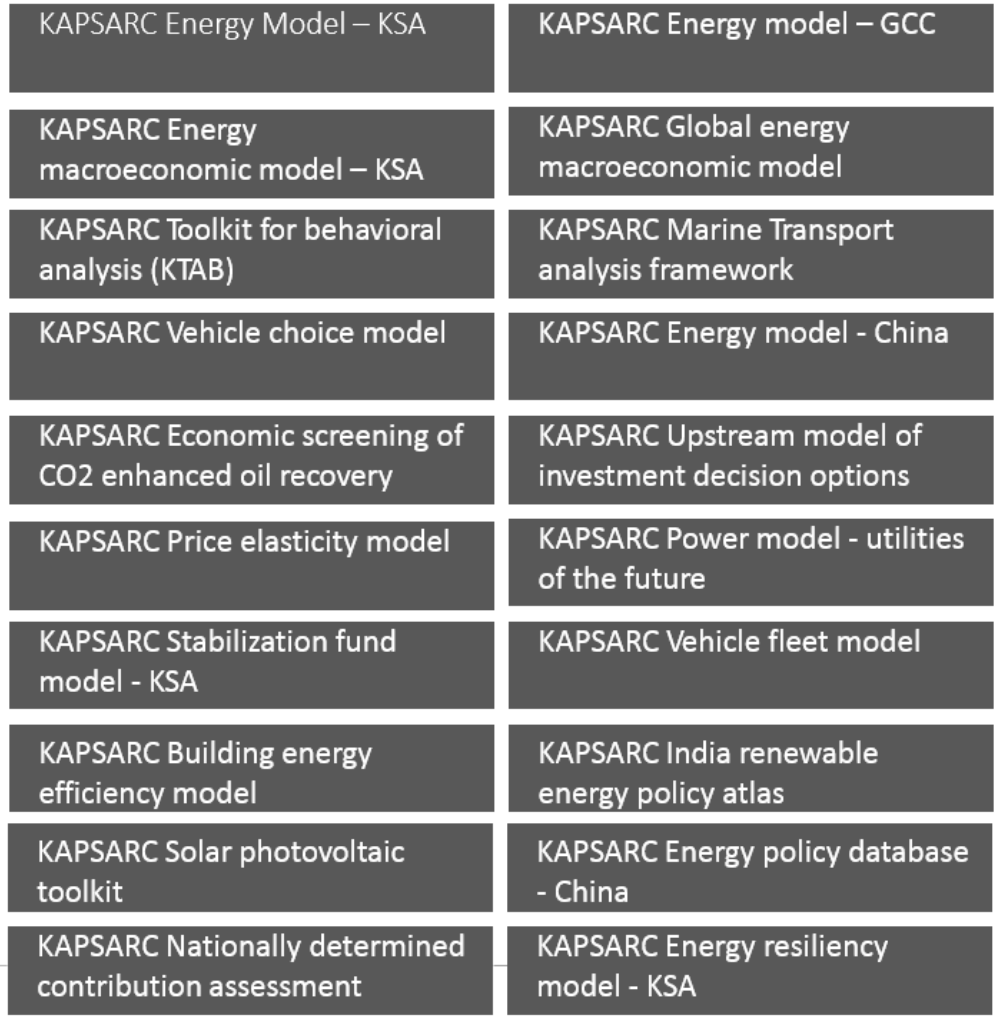
- Currency (monthly PLEASE!)
- Completeness
  - Coverage
  - Temporal
  - Granularity (disaggregated regional data)
- Consistency
- Correctness

Data trust improves when data is

- Traceable back to OFFICIAL source
- Used, reviewed, commented and critiqued
- Relevant to researcher context
- Interpretable across systems
- Definitions are standard, meta data published

Security

- Public, Restricted. Confidential



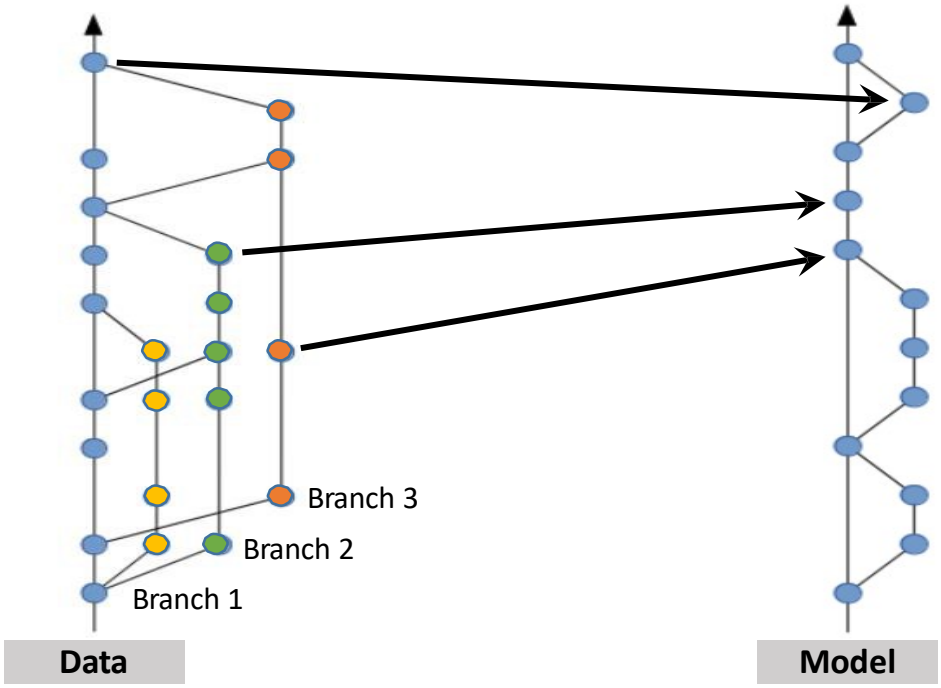
Granularity is key, better data, better insights

Regional Source	International Source	Not available

92% Annual, 6% Quarterly
70% Data 2015 – 2019

KAPSARC Model Data Editor

Addressing challenge of data versions



1.Seed Data Automated

2. Edit Data

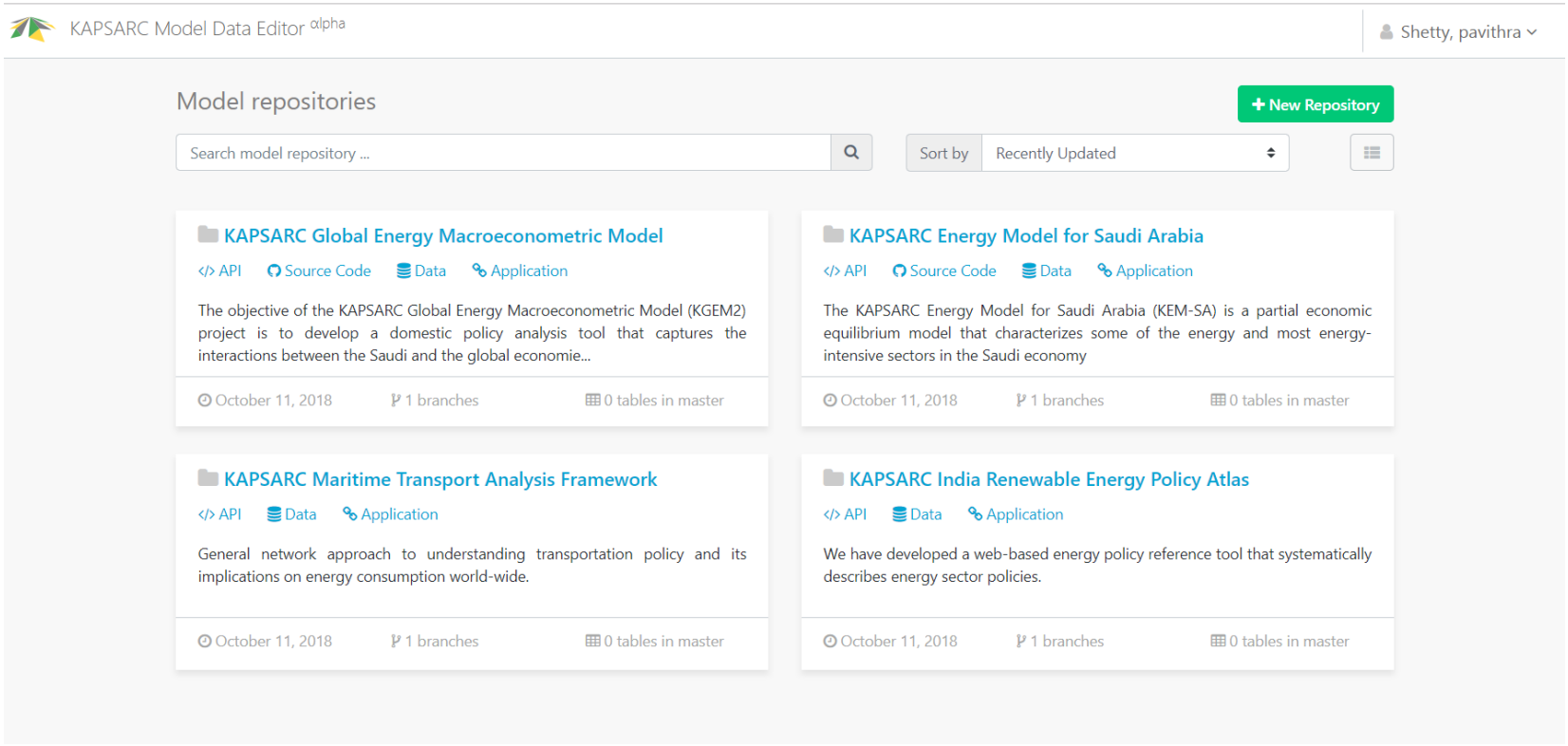
3. Version Alert

4. Collaborate

5. API Calls Hist.

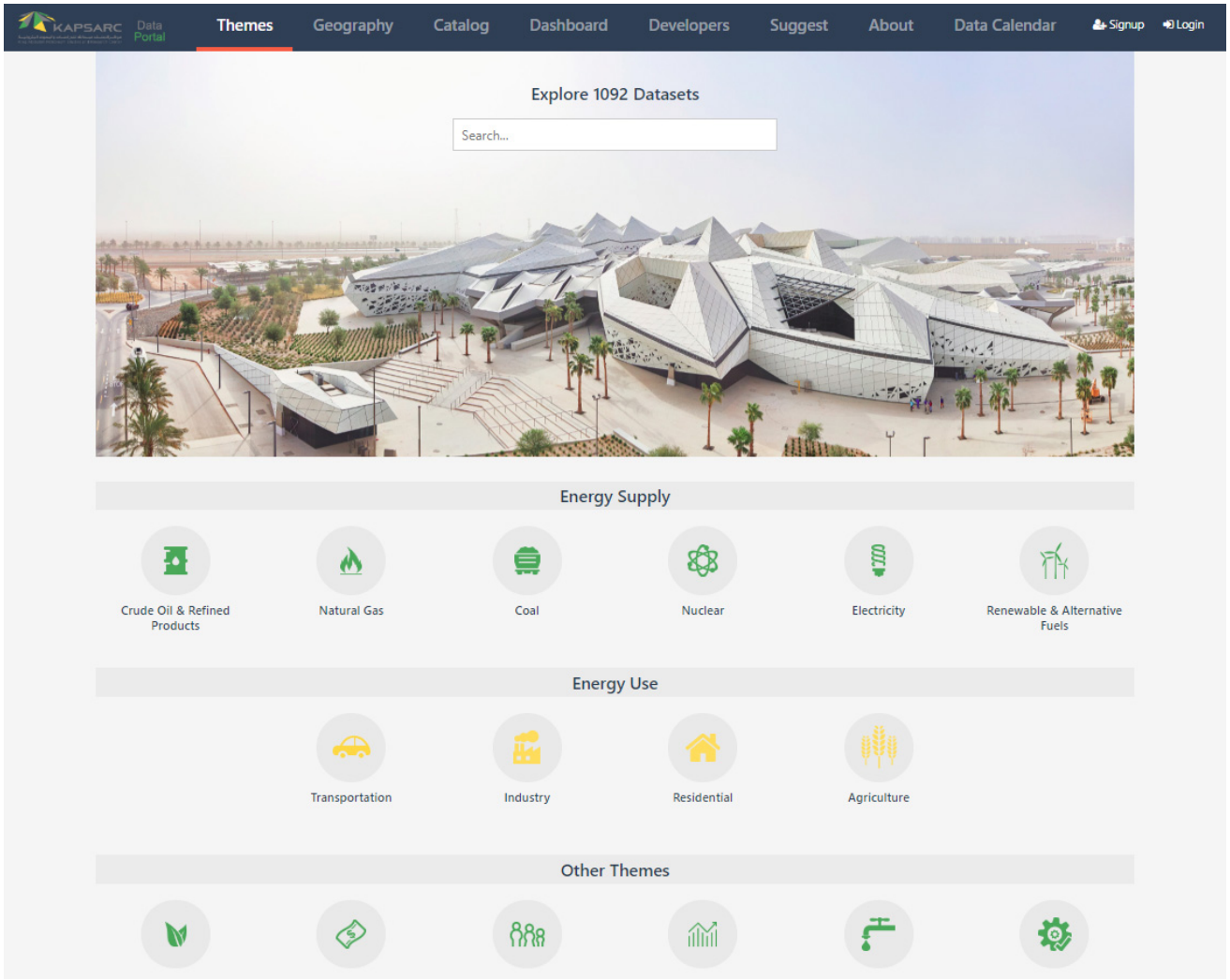
KAPSARC DataHub for Modelers

Open source portal to manage models' data and call via APIs (Application Programming Interface)



KAPSARC OpenData

In 2018 datasource.kapsarc.org served over 2,000,000 searches and 15000 downloads in from 100 countries



1.2K Datasets
machine readable

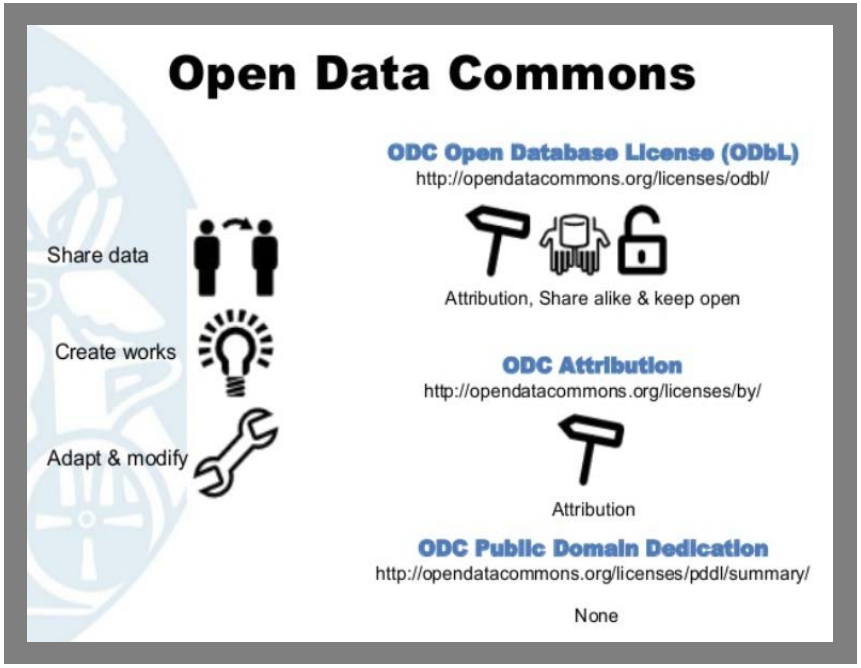
50M Records

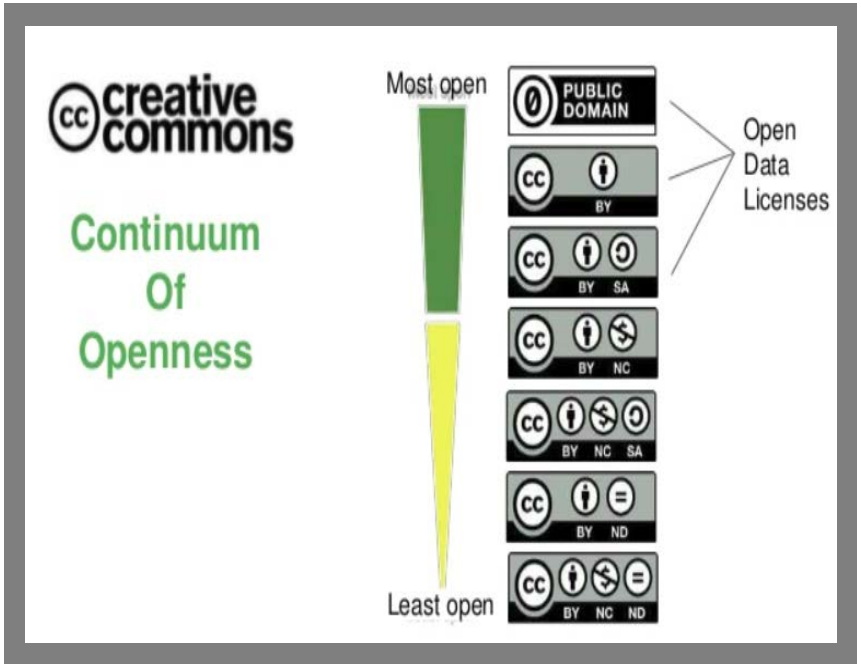
45 Days Currency

170 Data Sources

60 GCC Sources

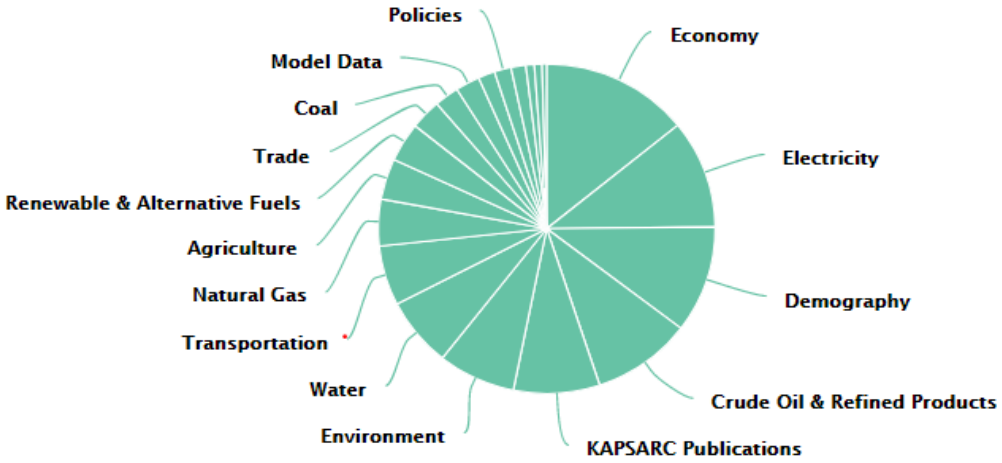
Data owners need to specify “data use” standards, soon technology will aid





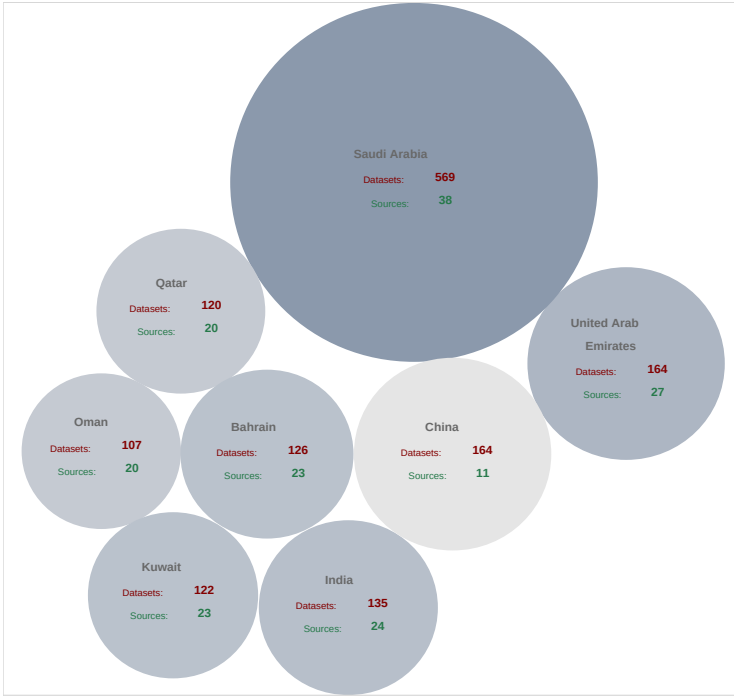
Datasets and themes

- **50 million records**
- **1700 Datasets**
  1000 public - 700 restricted
- **150 Sources**
  Identifying key sources from GCC, China and India
- **16 Themes**
  3 categories



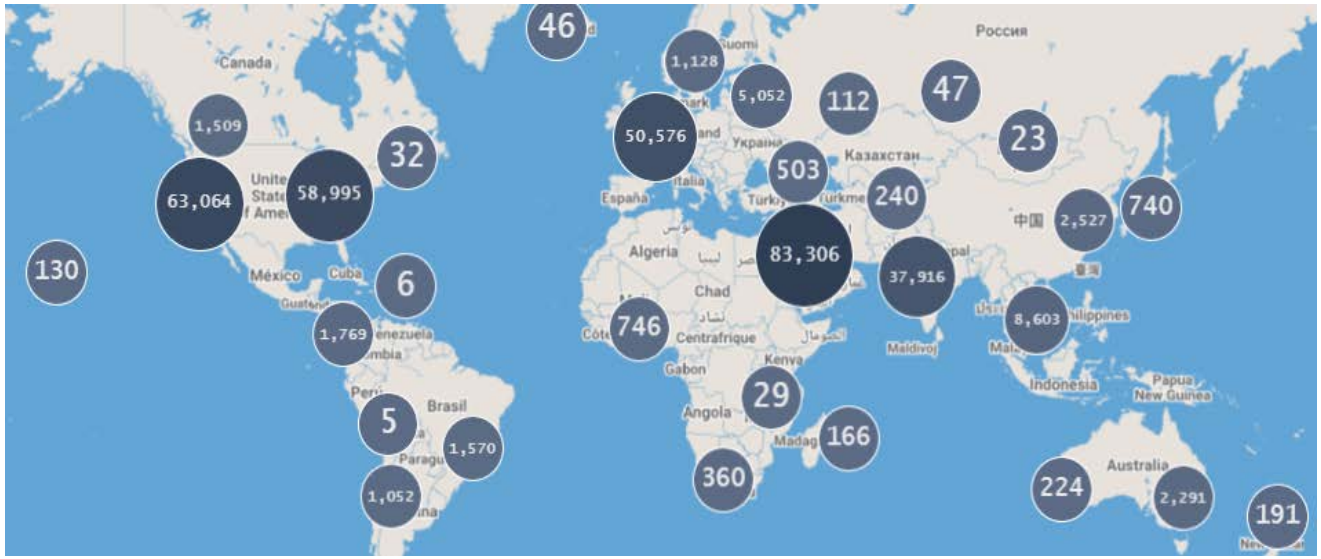
Top countries data coverage

- **GCC** 60+ data sources 1200+ datasets
  - **Saudi Arabia** 35+ sources
  - 560+ datasets
- **India and China** 35+ data sources 290+ datasets



Key take away

- Collaborate on granular data availability to advance research insights
- High frequency data, currency, machine readable data with auto alerts
- Extend data, model, insights via apps to improve understanding of energy economics



Let's collaborate on open

Data

Models

Tools

Insights