

Exploring the Potential of Generative Models for Evaluative Tasks: IEG Experiments

Estelle Raimondo, PhD : Program Manager

Harsh Anuj: Data Scientist

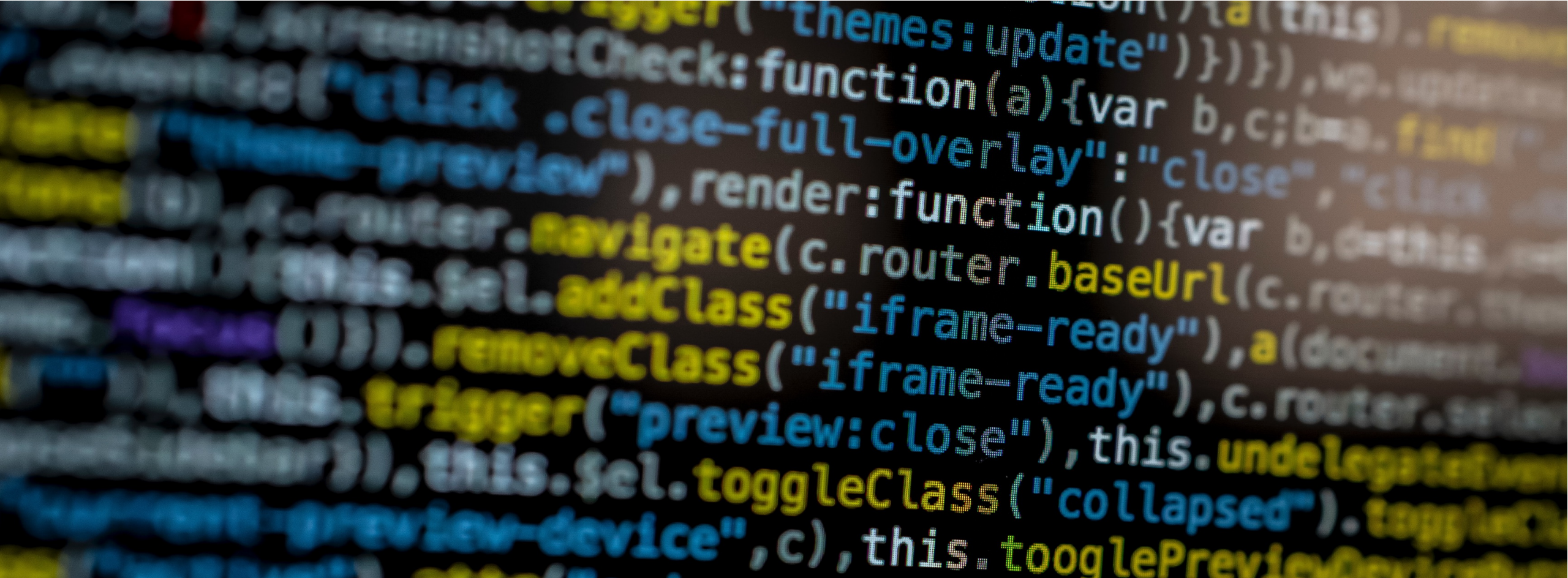
Methods Advisory Function

October, 2023



IEG
INDEPENDENT
EVALUATION GROUP

WORLD BANK GROUP
World Bank • IFC • MIGA



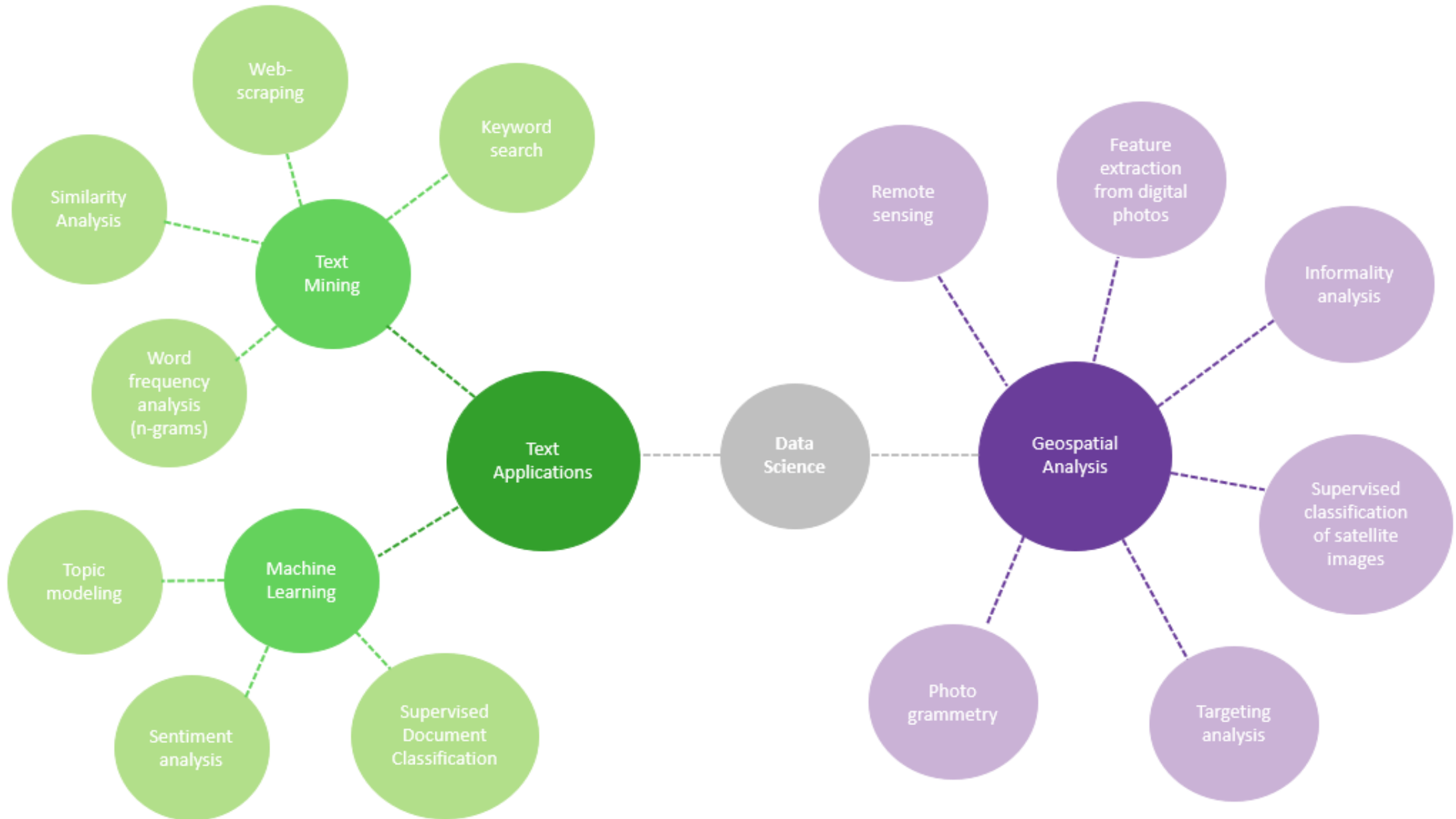
Introduction



- **Generative AI** is a set of **deep-learning** models that can create outputs that mimic the data they were trained on
- A common application is **Large Language Model (LLM)** such as ChatGPT, which can generate text and code on most topics.
- IEG has used **discriminative AI** for several years to extract meaning from text and images. They model decision rules but do not generate new text.

Typical applications & workflow

IEG has incorporated data science and AI applications



Text as data



RAP 22: NLP & sentiment analysis to identify factors of success –failure and risks associated with private sector investments



RAP 21: NLP to classify objectives and indicators by type (input, output, outcome, high-level outcome)



Undernutrition evaluation: NLP to classify text based on theory of change



Doing Business: NLP to identify complex portfolio and sources for structured literature review



Ukraine CPE: Sentiment analysis on media and social media data

Image as data



Mozambique CPE: Geospatial analysis to assess relevance of targeting based on needs



Morocco CPE: Supervised classification of satellite images to assess climate resilience outcome of irrigation interventions



Tanzania CPE: Feature extraction to assess impact of Rapid Bus Transit intervention

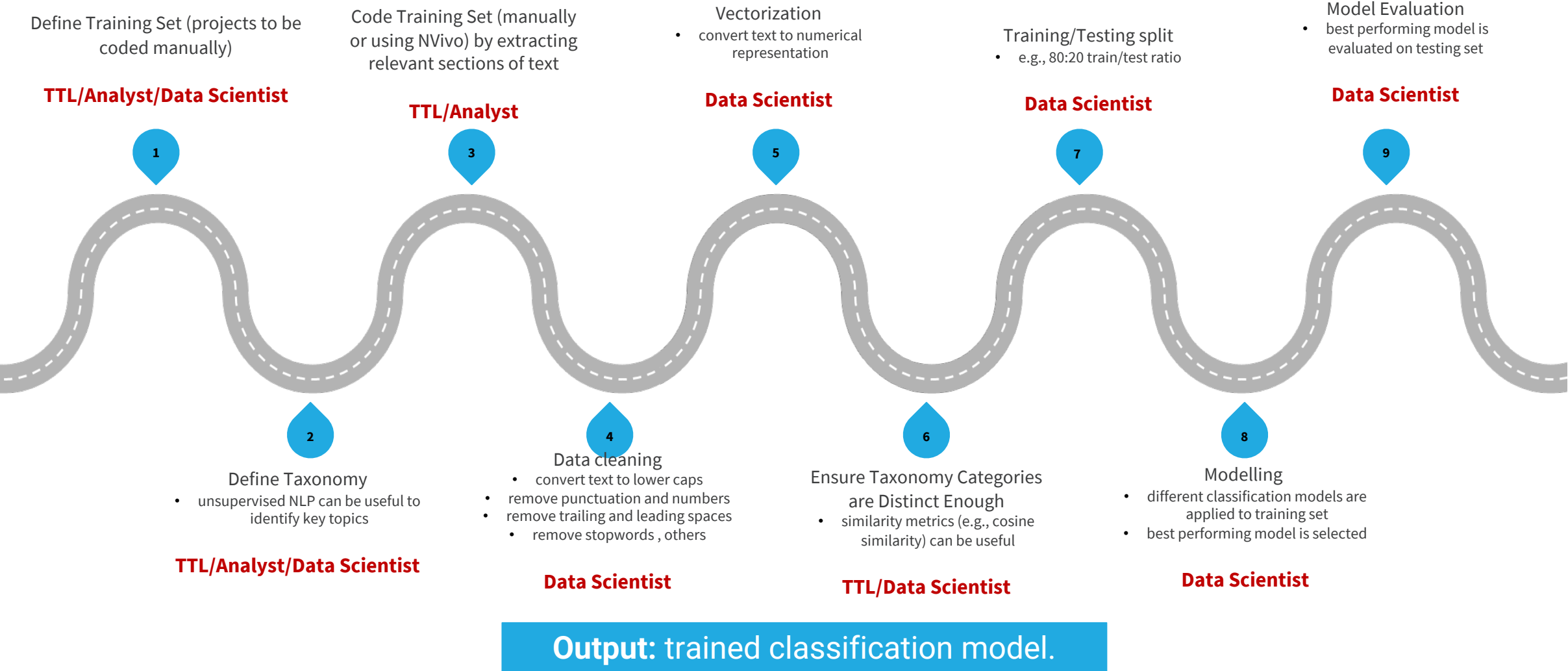


Blue Economy: Remote sensing and computer vision object detection to assess health of mangrove and coral

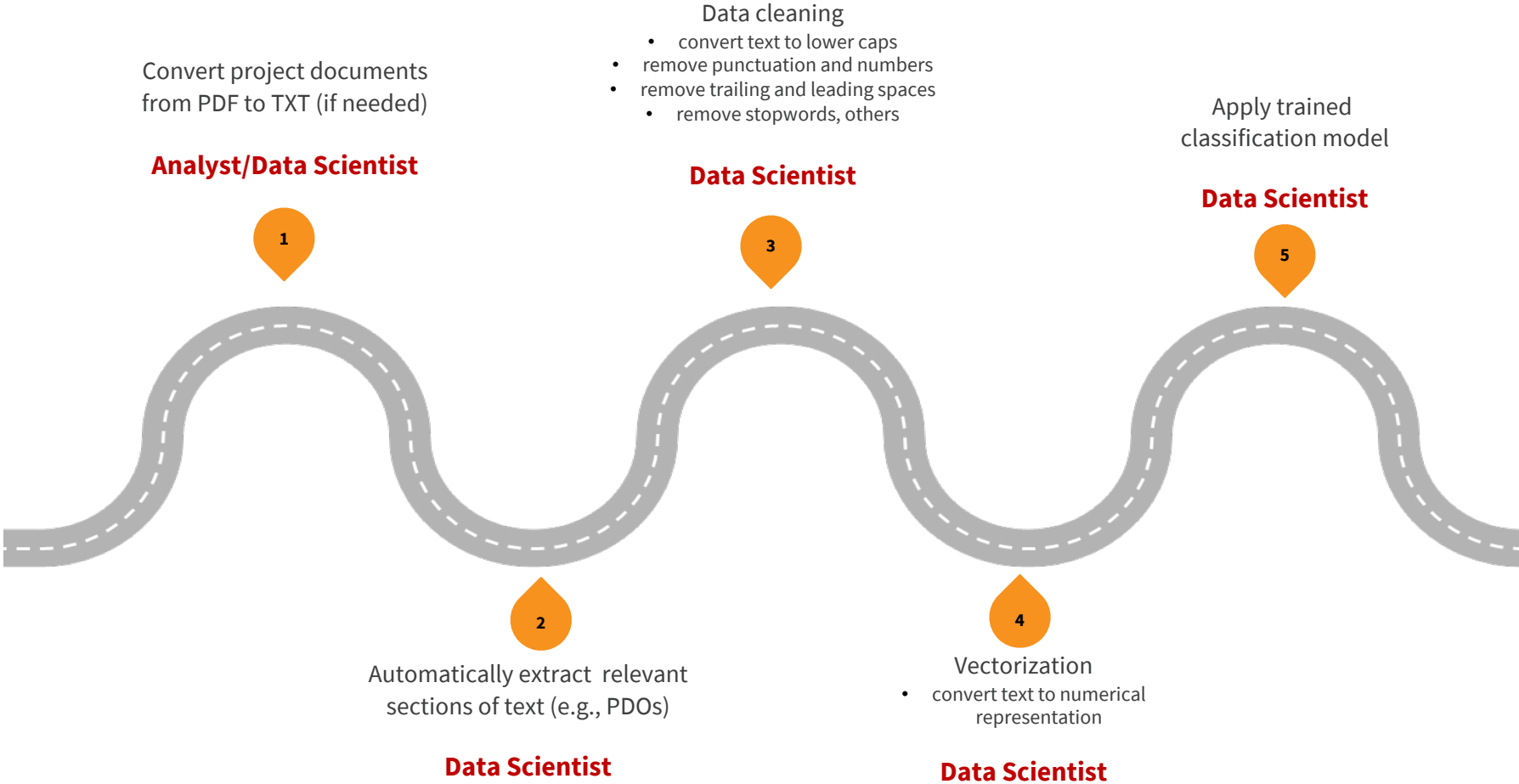


Urban Spatial Growth: Image segmentation of urban landscapes to assess economic and spatial impacts of interventions

Conventional Machine Learning Workflow: Training a Classification Model



Conventional Machine Learning Workflow: Applying a Classification Model to Uncoded Data





Experiments

Promises of Generative AI for the evaluation practice

- Speed
- Breadth of resources
- Quality
- Provide new insight
- Enhanced capabilities

Perils of Generative AI for the evaluation practice

- Ethics
- Biases
- Safety and security
- Truthfulness
- Transparency



Experiments' Design

- Awareness of risks and limitations
- Only publicly available input
- Ability to compare with existing outputs
- Testing across potential users (TTLs, analysts, data scientists)
- Testing various LLMs
- Close scrutiny on the output
- Some level of probing and interaction

Typology of use cases

For Data Scientists

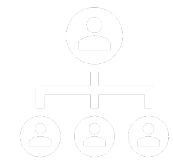
- Leverage LLM for NLP
- Expand or enhance NLP in evaluation
- Require specialized skills and knowledge (e.g., programming, using API, ML, etc.)

For Analysts

- Capacity to interpret and verify the output
- Ability to use OpenAI playground, manipulate plugins
- Capacity for prompt engineering

For TTLs

- No specialized knowledge of data science
- Tasks can be accomplished in natural Language
- Use of chatbot primarily





BARD AI



mAI Knowledge

mAI knowledge
WBG GPT

mAI knowledge **BETA**
WBG GPT

Simplifies access to curated World Bank knowledge through AI-powered discovery and interactions. Explore the collections below or create your own.

Need a custom collection?

You can create your own collection in minutes. Try it out!



[GET STARTED >](#)

mAI projects **BETA**
WBG GPT

Access to nearly 300,000 World Bank project documents through AI-powered discovery and interactions

[Try Now >](#)

mAI documents and reports **BETA**
WBG GPT

Access to over 500,000 official documents and reports through AI-powered discovery and interactions

[Try Now >](#)

Time for Q&A



Fulfilled Promises

Writing and explaining code

Experiment

- We asked ChatGPT to:
 - **generate Python code** to complete a text pre-processing pipeline on the training set used for an evaluation.
 - **provide R code** to replicate the multivariate regression analysis conducted for an evaluation.
 - add **detailed comments** to and provide a **high-level summary** of an R script used to bulk download WB documents.

Findings

- ChatGPT **produced correct code** that was suitable for the task as long as **detailed instructions** were included as part of the prompt.
- Code **performed well** and allowed to replicate the results of the study (including the plots).
- **Several iterations** with ChatGPT were needed to fix some issues in the initial outputs.
- It also provided **excellent responses** to the code explaining tasks.

Recommendation

- ChatGPT can be very useful for analysts and data scientists for writing code for standard tasks such as preprocessing text data.
- It is critical to:
 - **be very specific** on the prompts
 - **understand what the correct output should look like** (to identify any potential oversights in the output).
- Data scientists and analysts can further explore the use of LLMs and chatbots to **make their or colleague's code more understandable**.



Writing and explaining code for standardized tasks

Experiment

- We asked ChatGPT to **generate Python code** to complete a text pre-processing pipeline on the training set used for RAP 2023.
- We also asked ChatGPT to:
 - a. **add comments** to R script which bulk downloads WDRs from D&R API;
 - b. provide a high-level **summary of the code**; and
 - c. provide a **detailed description** of a user-defined function within the code.



Findings

- ChatGPT **produced correct code** that was suitable for the task as long as **detailed instructions** were included as part of the prompt.
- **Several iterations** with ChatGPT were needed to fix some issues in the initial outputs.
- It also provided excellent responses to the code explaining tasks.



Recommendation

- ChatGPT can be very useful for analysts and data scientists for writing code for standard tasks such as processing text data.
- It is critical to:
 - **be very specific** on the prompts
 - **understand what the correct output should look like** (to identify any potential oversights in the output).
- Data scientists and analysts can further explore the use of LLMs and chatbots to **make their code more understandable**
- Data scientists/analysts can find this application useful **to better understand code written by other colleagues**.



Analysis

Data Scientist

Analyst

TTLs

Conducting a simple classification

Experiment

- We asked ChatGPT, GPT's API, and Mai to **classify text data** (PDOs) as to whether or not they are related to disaster risk reduction.
- Manually coded data was used to evaluate the accuracy of the results.

Findings

- **ChatGPT and GPT's API** performed this task very well (**>76% accuracy**).
- In comparison, **Mai** had a significantly lower performance (**~57%**).
- ChatGPT and Mai could only process ~25 entries at the time, while the API processed the complete dataset.

Recommendation

- ChatGPT and GPT's API can be useful tools for simple classification tasks (e.g., binary classification).
- **At least a subset of the output should be manually validated** to ensure the model is functioning as expected.



Analysis

Data Scientist

Analyst

TTLs

Conducting sentiment analysis

Experiment

- We asked ChatGPT and GPT's API to **provide the sentiment (positive, neutral, or negative)** of input sentences.
- The manually coded training set from RAP 2023 was used to evaluate the accuracy of GPT's predictions.

Findings

- **GPT** achieved **very high accuracy (~95%)** - which was expected.
- ChatGPT could only process ~50 entries at time, while the API was able to process the complete dataset.
- **ChatGPT** started to **"hallucinate"** new sentences.

Recommendation

- GPT's API is a **solid option** for sentiment analysis.
- ChatGPT can be useful to classify a limited number of sentences. It is advisable to:
 - start a new window for each prompt
 - ensure that the output includes the sentences provided as input.



Post-Analysis

Data Scientist

Analyst

TTLs

Summarizing individual documents

Experiment

- We asked ChatGPT to summarize a recently published Country Program Evaluation report from IEG. The report was close to 200 pages long.

Finding

- It produced a well-written and accurate high-level summary of the document based on its key topics.

Recommendation

- Chatbots can be used to summarize a document that the user is deeply familiar with, so as to save time and effort in formulating and typing a first draft of the summary.
- Summaries of individual documents can also be leveraged for evaluative synthesis tasks.



Unfulfilled Promises

Pre-Analysis

Data Scientist

Analyst

TTLs

Generating images for geospatial analysis (data augmentation)

Experiment

- We asked DALL-E to **generate urban images** in a style similar to that of Bathore, Albania.
- Two variations of this experiment were conducted: (i) generation of images from text prompt, and (ii) generation of images from image upload.

Findings

- **Only 4 images** per prompt are produced.
- **Hard to evaluate** the similarity between the generated images and real images.

Recommendation

- **Not recommended** for data augmentation..



Post-Analysis

Data Scientist

Analyst

TTLs

Conducting a literature review

Experiment

- We asked ChatGPT and Mai to *"please write a short literature review of the advantages and challenges of the use of Doing Business Indicators"*.

Finding

- ChatGPT provided a much longer and detailed response (approximately double in length) than Mai.
- Not possible to ascertain the veracity of the response (e.g., hallucinations of references).

	chatGPT	mAI
Advantages	Competitive Analysis	Comparative Analysis
	Reform Incentives	Transparency
	Investor Confidence	Incentive for reforms
	Policy Monitoring	Investor Confidence
Challenges	Simplified Assessments	Limited Scope
	Data Limitations	Methodological Limitations
	Incomplete Scope	Lack of Context
	Incentive for Data Manipulation	Political Bias

Recommendation

- These models can have some use for obtaining some background knowledge on a specific topic, but caution needs to be exercised to verify the outputs through reliable sources. They can't be used to generate content to be used directly in reviews



Post-Analysis

Data Scientist

Analyst

TTLs

Conducting an evaluation synthesis (e.g., EIN)

Experiment

- We asked ChatGPT to ingest the text from 6 PPARs and generate the text for an evaluative synthesis based on this evidence base.

Finding

- While the writing was very good and some of the high-level messages were appropriate, it fabricated evidence not present in the evidence base, thus eroding trust in the entire exercise.

Recommendation

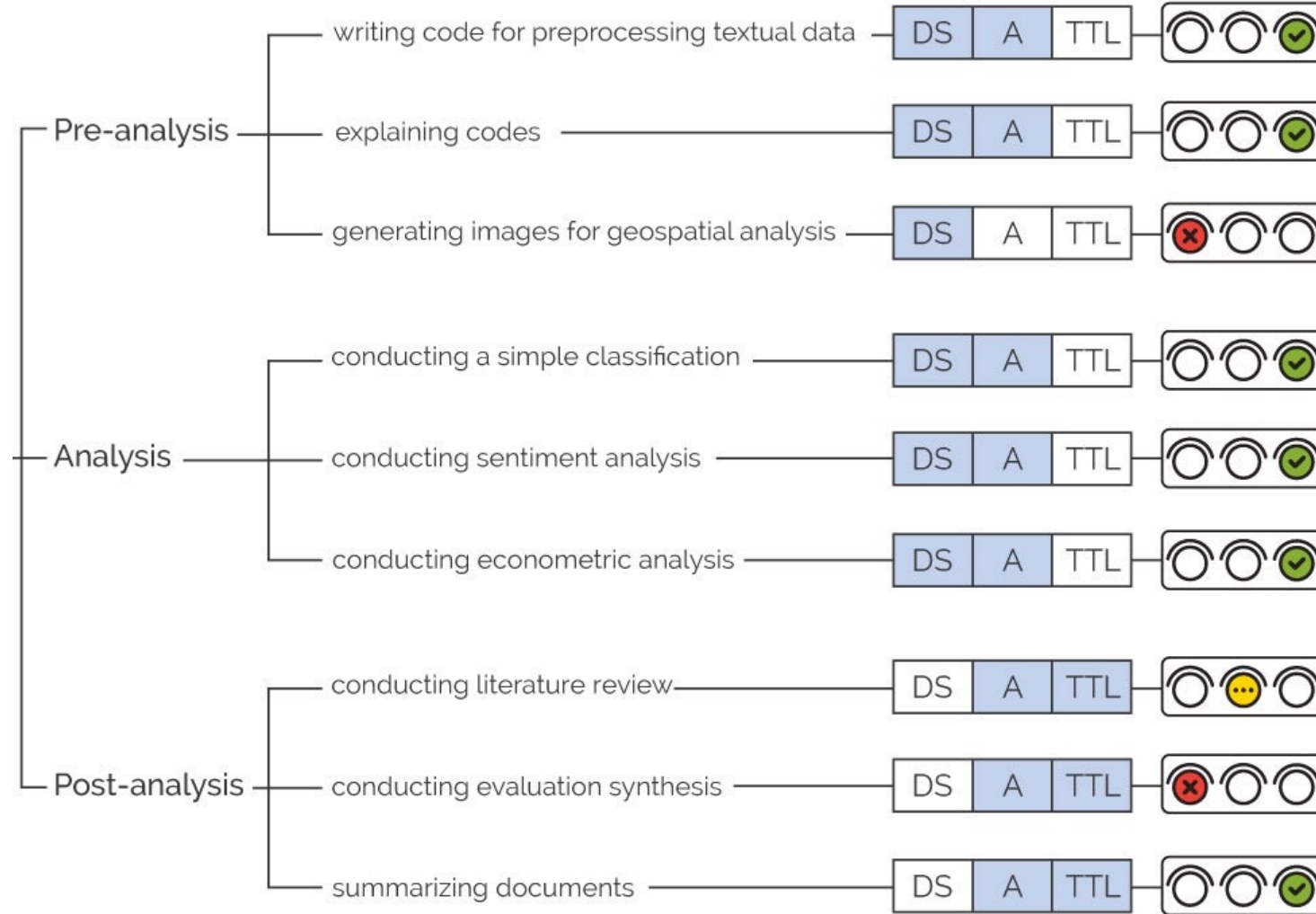
- Chatbots should not be used for synthesizing evidence from multiple sources.





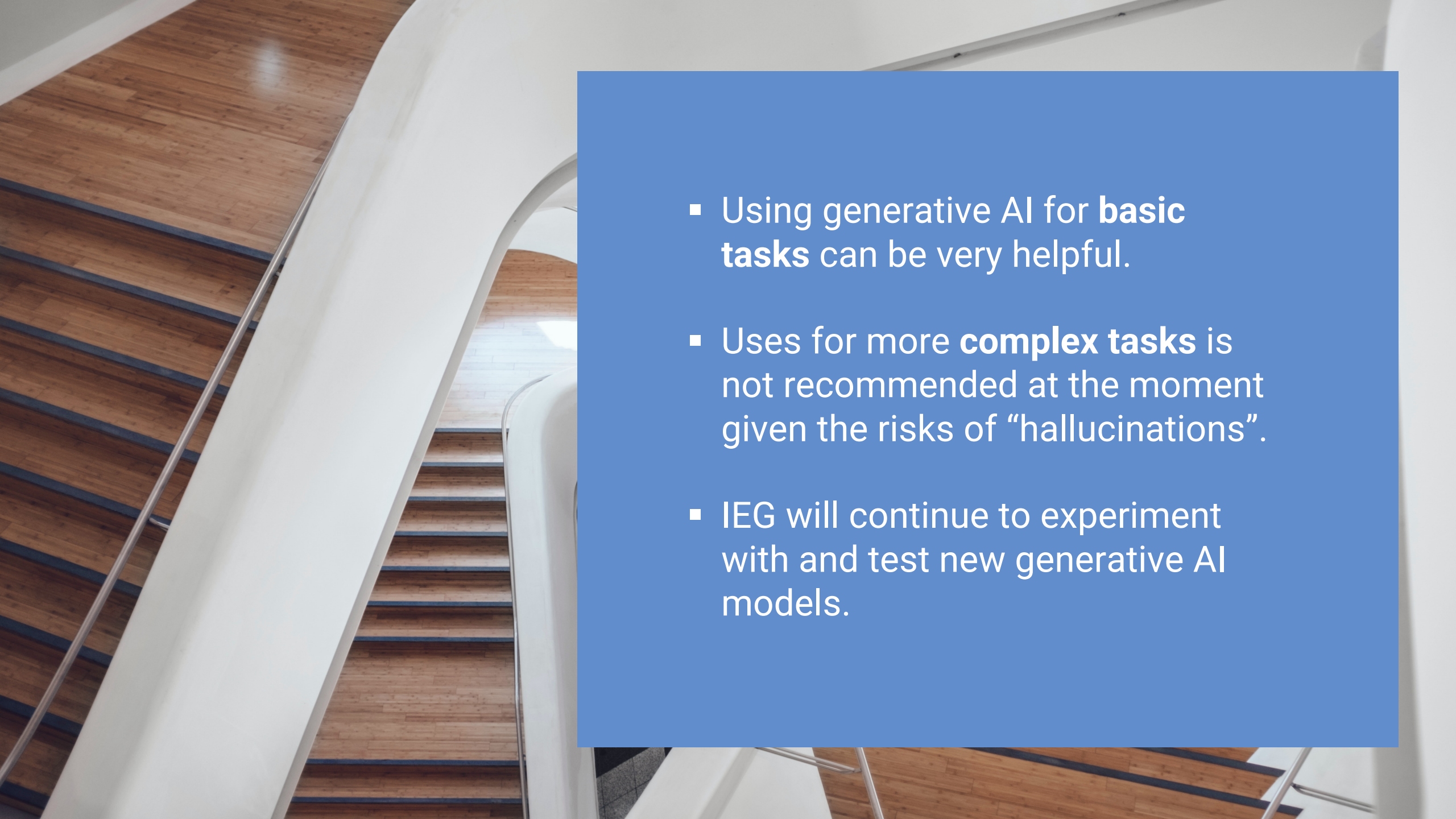
Future Directions

Recommended Uses for Evaluation Practice



Source: Independent Evaluation Group.

Note: A = analyst; DS = data scientist; TTL = task team leader.

- 
- Using generative AI for **basic tasks** can be very helpful.
 - Uses for more **complex tasks** is not recommended at the moment given the risks of “hallucinations”.
 - IEG will continue to experiment with and test new generative AI models.

Thank You!

<http://ieg.worldbankgroup.org>