Artificial Intelligence (AI)-Enabled Analytics: A Brief Overview and An Open-Source Tools Inventory

Dr. Sagar Samtani and Dr. Hsinchun Chen

The inputs from Ben Ampel, Charlie DeVries, Reza Ebrahimi, Ben Lazarine, Amy Lin, Agrim Sachdeva, Steven Ullman, and Hongyi Zhu are gratefully acknowledged.

Outline

- Background of AI-enabled Analytics
- An Open-Source Tools Inventory for AI-enabled Analytics
  1. Data Collection and Aggregation
  2. Data Extraction and Representation
  3. Analytics
  4. Visualization and Presentation
  5. Other Selected Resources
- Conclusion
- Summary

A Brief Background of AI-enabled Analytics

- Artificial Intelligence (AI) has rapidly emerged as a key disruptive technology of the 21^{st} century.

- AI has shown significant promise in various application areas, including robotics, game playing, drones, self-driving cars, and others.

- Increasingly, many organizations are seeking to identify how AI can help analyze their structured (e.g., transactions) and unstructured data (e.g., text, sensor signals).

- This interest is giving rise to an emerging field of AI-enabled analytics.
    - An abstracted approach to conducting AI-enabled analytics is presented in Figure 1.

A Brief Background of AI-enabled Analytics

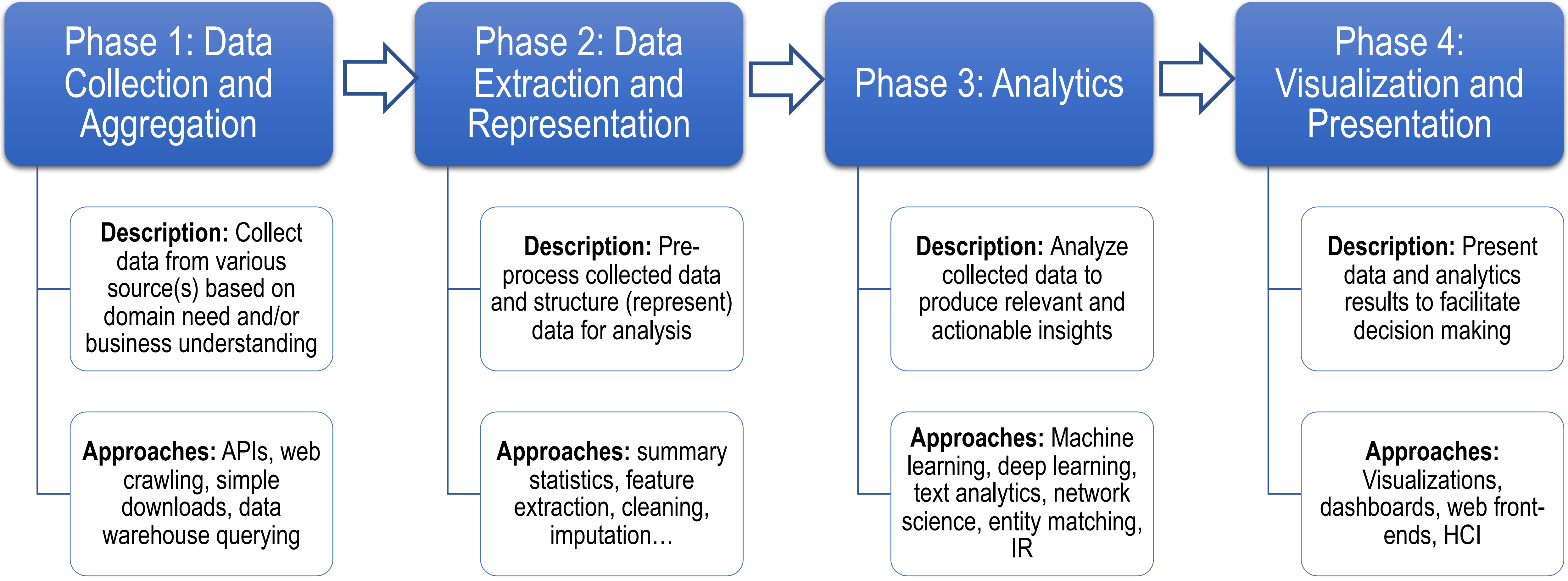

Figure 1. An Abstracted (Domain-agnostic) Approach to Conducting AI-enabled Analytics

A Brief Background of AI-enabled Analytics

- This process can be executed for different domains, including BI&A (Chen et al. 2012), cybersecurity (Samtani et al. 2020), and privacy (Samtani et al. 2021).

- Most successful implementations of AI-enabled analytics are based on a strong understanding of the domain or business being studied.

- Understanding of domain or business can be based on:
  - How domain experts or professionals execute their tasks (e.g., workflows)
  - Regulations and statutes e.g., HIPAA, GDPR, CCPA
  - Extant frameworks e.g., cybersecurity risk management frameworks
  - …

An Open-Source Tools Inventory for AI-enabled Analytics

- Executing each phase of the AI-enabled analytics requires a set of tools.

- Therefore, a review of prevailing tools was conducted. The review was conducted based on the following key criteria:
    1. Tools should be open-source, rather than paid (to help with cost management).
    2. Tools should be interoperable with Python and with SQL or NoSQL.

- Tools are organized based on each major AI-enabled analytics phase.
    - For each identified tool, a brief description and a link to the tool are provided.
    - **Important!** Some tools can perform multiple functions e.g., both extract and analyze.

An Open-Source Tools Inventory for AI-enabled Analytics – Data Collection and Aggregation

- Data collection and aggregation is focused on collecting data that could be used for subsequent analysis. This phase comprises tools (Table 1) for:
  - **Collection:** Mechanisms to access and collect (crawl, download) the data sources.
  - **Storage:** Storing data for users to query and to serve as a backend for web portals.

Task	Tool/Package Name(s)	Description	Documentation
Collection	Scrapy	Package for incremental web crawlers	https://scrapy.org/
	JSON	Package for parsing JSON data from APIs	https://docs.python.org/3/library/json.html
	BeautifulSoup	Package for general web crawling	https://pypi.org/project/beautifulsoup4/
	Google BigQuery	Queriable data warehouse of public datasets	https://cloud.google.com/bigquery
	Paramiko	SSH connection to extract data from VMs	https://www.paramiko.org/
Storage	MySQL	Relational database	https://www.mysql.com/
	Pickle	Storing ML/DL models	https://docs.python.org/3/library/pickle.html
	MongoDB	NoSQL database	https://www.mongodb.com/
	Elasticsearch	NoSQL database for storing documents	https://www.elastic.co/
	Neo4j	NoSQL database for storing graph data	https://neo4j.com/v2/
	Hadoop	Framework that allows distributed storage	https://hadoop.apache.org

Table 1. Open-Source Tools for Data Collection and Aggregation

An Open-Source Tools Inventory for AI-enabled Analytics – Data Extraction and Representation

- Since collected data is rarely in a format that can be directly analyzed, relevant data of interest (based on business/domain needs) should be:
  - **Extracted** from its original, raw format and cleaned (pre-processed) to remove noise.
  - **Represented** in a data structure (e.g., vector, graph, grid) suitable for the targeted analytics

- Common tasks include producing summary statistics, imputation, deduplication, cleaning, annotation, and many others. Prevailing tools appear in Table 2.

Tool/Package Name(s)*	Description	Documentation
re	Support regular expression matching	https://docs.python.org/3/library/re.html
numpy	Creation and operations on multi-dimensional numeric arrays	https://numpy.org
Data Analysis Baseline Library	Common ML pre-processing tasks	https://amueller.github.io/dabl/dev/
Pandas	Formatting and structure data inputs from varying data sources	https://pandas.pydata.org/
SideTable	Advanced data-wrangling with Python	https://pbpython.com/sidetable.html
Pigeon	Interface to rapidly annotate unlabeled data	https://github.com/agermanidis/pigeon

Table 2. Open-Source Tools for Data Extraction and Representation

An Open-Source Tools Inventory for AI-enabled Analytics – Analytics

- The analytics phase is the heart of conducting AI-enabled analytics.

- In this set of slides, six sets of analytics are covered:
    1. **Conventional Machine Learning (ML):** Approaches that learn from feature vectors.
    2. **Deep Learning (DL):** Approaches that learn from data structures (e.g., grids, sequences).
    3. **Text Analytics:** Techniques that aim to extract insights from unstructured text data.
    4. **Network Science:** Approaches that analyze graph or tree-structured data.
    5. **Information Retrieval (IR) and Entity Resolution (ER):** Techniques that link multiple sources of data (for retrieval or resolution).
    6. **Emerging Learning Paradigms:** Specialized approaches for learning from data beyond the classical supervised learning and unsupervised learning perspectives.

- Although not comprehensive of *all* analytics approaches, the listed categories represent some of the most popular and prevailing at the time of this writing (2022).
    - A (very) brief summary of the underlying concepts for each analytics procedure is provided.

An Open-Source Tools Inventory for AI-enabled Analytics – Analytics (Conventional ML)

- Conventional ML techniques and tasks have historically been the most closely associated with AI-enabled analytics.

- Conventional ML can be broadly categorized into:
  - **Supervised learning:** Aims to predict an output variable based on a set of input (independent) variables (features).
    - **Process:** gold-standard dataset development → feature extraction → model (e.g., SVM) selection and training → model evaluation (e.g., hold-out, CV, performance measurement via accuracy, precision, recall, F1) → model tuning
  - **Unsupervised learning:** Aims to find the “natural” relationships (e.g., partitions, associations) of data instances within a dataset.
    - **Common approaches:** clustering (hierarchical, partitional), association rule mining.

An Open-Source Tools Inventory for AI-enabled Analytics – Analytics (Conventional ML)

- Three major categories of conventional ML tools exist (Table 3):
    1. ML packages that include a comprehensive set of ML algorithms and procedures.
    2. GUI-based ML workflows that allow users to conduct ML in a drag-and-drop fashion
    3. AutoML tools that automate aspects of the conventional ML process (e.g., tuning parameters).

- Each tool provides a suite of conventional ML algorithms and mechanisms to evaluate the performance of ML algorithms.

Category	Tool/Package Name(s)	Brief Description	Documentation
ML packages	Scikit-learn	Basic ML algorithm implementation and evaluation	https://scikit-learn.org/stable/
	Spark	Unified analytics engine for large-scale data processing	https://spark.apache.org
GUI-based ML workflows	RapidMiner	GUI-based, general purpose ML toolkits for creating workflows	https://rapidminer.com/
	WEKA		https://www.cs.waikato.ac.nz/ml/weka/
AutoML	TPOT	Sklearn-based AutoML feature selection and model selection	https://epistasislab.github.io/tpot/
	HyperOpt	Sklearn-based AutoML ML hyperparameter tuner	https://hyperopt.github.io/hyperopt-sklearn

Table 3. Open-Source Tools Conventional ML-based Analytics

An Open-Source Tools Inventory for AI-enabled Analytics – Analytics (Deep Learning)

- DL has rapidly emerged as an approach to automatically extract multiple levels of features (representations, embeddings) from “raw” data. DL comprises:
  - **Data encoding** structures the raw data into a format (e.g., grid) for a DL model to learn from.
  - **Basic processing units (architectures)** such as ANN, CNN, RNN, and GNN that operate on the data encoding.
  - **Architecture extensions** (e.g., attention, highway, bidirectional processing) to improve the model’s capacity to learn from the data encoding.
  - **Learning paradigm** (e.g., supervised, unsupervised, adversarial) that defines *how* the model learns from the data encoding.

- Many DL approaches are deployed using supervised learning or unsupervised learning paradigms and therefore follow evaluation approaches as conventional ML.
  - However, they are also used with many emerging learning paradigms (see Table 8).

An Open-Source Tools Inventory for AI-enabled Analytics – Analytics (Deep Learning)

- A summary of prevailing open-source tools for conducting DL-based analytics is presented in Table 4.

- Some key takeaways of the tools include:
  - Keras offers some of the most user-friendly approaches to executing basic DL with supervised, unsupervised, adversarial, or transfer learning.
  - PyTorch is excellent for customizing DL models (e.g., loss) for specific applications.
  - Huggingface and SimpleTransformers provide access to large pre-trained models (e.g., BERT, GPT) as well as emerging architectures, namely, transformers.

Tool/Package Name(s)	Brief Description	Documentation
Pytorch	Advanced Python package for customizable deep learning	https://pytorch.org/
Keras	Basic package with standard DL algorithms	https://keras.io/
fastai	Various tools and resources for DL	https://www.fast.ai/
Huggingface	Large repository of pre-trained language models (e.g., BERT)	https://github.com/huggingface
SimpleTransformers	Barebones implementation of pre-trained language models	https://github.com/ThilinaRajapakse/simpletransformers

Table 4. Open-Source Tools for DL-based Analytics

An Open-Source Tools Inventory for AI-enabled Analytics – Analytics (Text Analytics)

- Many modern data sources, especially for BI&A, are text-based.
- Three major categories of tools for text analytics exist (Table 5).
  1. **Multi-purpose general text analytics** that supports common text analytics
  2. **Specialized text analytics** for particular types of text analytics tasks (e.g., NER, PoS)
  3. **Multi-lingual analytics** to support analysis of text in non-English languages.

Category	Tool/Package Name(s)	Brief Description	Documentation
Multi-purpose text analytics	NLTK	Python package for symbolic and statistical NLP	https://www.nltk.org/
	Spacy	Industrial strength, large-scale information extraction and NLP	https://spacy.io/
Specialized text analytics	Flair	PyTorch extension for NER, PoS, and custom embeddings	https://github.com/flairNLP/flair
	T-NER	Pre-trained language models for NER	https://github.com/asahi417/tner
	Gensim	Package for basic word embeddings and topic modelling	https://radimrehurek.com/gensim/
Multi-lingual analytics	Textflint	Unified multi-lingual robustness evaluation toolkit for NLP	https://github.com/textflint/textflint
	Polyglot	NLP pipeline for multi-lingual analysis (supports 196 languages)	https://polyglot.readthedocs.io/en/latest/
	Stanza	Python package from Stanford for multi-lingual analysis	https://stanfordnlp.github.io/stanza/

Table 5. Open-Source Tools for Text Analytics

An Open-Source Tools Inventory for AI-enabled Analytics – Analytics (Network Science)

- Many contexts can be represented as a network (e.g., graph) that captures relationships (edges) between different entities (nodes).

- In recent years, network science been conducted with two major categories of tasks (Table 6)
  1. **Graph Construction and Analysis** (1) represents a graph and (2) extracts graph-level properties (e.g., density, diameter), node-level statistics (e.g., centralities), and structures (e.g., communities).
  2. **Graph Embedding** techniques aim to project various components of a graph (e.g., nodes, edges, etc.) into a low-dimensional space to facilitate downstream analysis (e.g., classification, propagation).

Category	Tool/Package Name(s)	Brief Description	Documentation
Graph Construction and Analysis	Networkx	Python package for basic network science tasks	https://networkx.github.io/
	igraph	Package for extensive (non-DL) network science	https://igraph.org/python/
Graph Embeddings	stellargraph	Graph embedding package with common graph embedding methods	https://github.com/stellargraph/stellargraph
	PyG	PyTorch-based library to develop custom Graph Neural Network	https://pytorch-geometric.readthedocs.io/
	Deep Graph Library	Python package for deep learning on graphs	https://www.dgl.ai/

Table 6. Open-Source Tools for Network Science Based on Task Category

An Open-Source Tools Inventory for AI-enabled Analytics – Analytics (IR and ER)

- Many organizations have access to multiple modalities or sources of data.
  - Linking data instances across these different sources is of growing interest.

- Increasingly, two major approaches (major open-source tools summarized in Table 7) are being leveraged for multi-modal analysis, particularly linking:
  - **Information Retrieval** tasks such as Q&A systems, short text matching, search engines, etc. often aim to retrieve an entity (e.g., document) based on a key (e.g., query).
  - **Entity resolution** seeks to resolve different data instances that refer to the same entity.

Category	Tool/Package Name(s)	Brief Description	Documentation
Information Retrieval	MatchZoo	Short text matching and deep structured semantic modeling	https://ntmc-community.github.io/
	Pyserini	Implementations of non-DL-based IR algorithms	https://github.com/castorini/pyserini
	OpenMatch	Algorithms for document matching	https://github.com/thunlp/OpenMatch
Entity Resolution	RecordLinkage	Supports non-DL-based entity resolution	https://github.com/J535D165/recordlinkage

Table 7. Open-Source Tools for Information Retrieval and Entity Resolution

An Open-Source Tools Inventory for AI-enabled Analytics – Analytics (Emerging Learning Paradigms)

- Many extant analytics are based in supervised learning or unsupervised learning.

- However, an increasing body of work, especially in DL, is leveraging learning paradigms that go beyond this dichotomy and more closely emulate a human's learning.
  1. **Transfer Learning and Knowledge Distillation** transfer or distill knowledge between models.
  2. **Reinforcement Learning** has an "agent" to learn from an environment using feedback from its actions.
  3. **Self-Supervised Learning** aims to obtain supervisory signals (labels) from the data itself by leveraging the underlying structure of the data (to generate labels) during the model training process.

Category	Tool/Package Name(s)	Brief Description	Documentation
Transfer Learning and Knowledge Distillation	TLlib	Transfer learning library built on PyTorch	https://github.com/thuml/Transfer-Learning-Library
	KD Awesome List	A list of open-source repositories for knowledge distillation	https://github.com/FLHonker/Awesome-Knowledge-Distillation
Reinforcement Learning	OpenAI Gym	Provides pre-built environments to execute RL methods	https://gym.openai.com
	Coach	Offers various RL agents and algorithms	https://github.com/IntelLabs/coach
Self-Supervised Learning	VISSL	Self-supervised learning from images	https://vissl.ai/
	Graph SSL Awesome List	Supports self-supervised learning on graphs	https://github.com/LirongWu/awesome-graph-self-supervised-learning

Table 8. Open-Source Tools for Emerging Learning Paradigms

An Open-Source Tools Inventory for AI-enabled Analytics – Visualization and Presentation

- Visualizations and web-based user interfaces (UIs) can help end-users realize the full potential of insights extracted from AI-enabled analytics.
  - Can help facilitate effective decision-making processes and improve AI trust.

- Visualizations can also enable A/B tests or user evaluations e.g., usability, ease of use, usefulness, validation of algorithm results, task completion, etc.

- A summary of prevailing visualization and web front-end tools is presented in Table 9.

An Open-Source Tools Inventory for AI-enabled Analytics – Visualization and Presentation

Task	Tool/Package Name(s)	Brief Description	Documentation
Visualization	Seaborn	Basic Python-based statistical visualization package	https://seaborn.pydata.org/
	Plotly	Advanced Python-based visualization package for ML/DL	https://plotly.com/
	TensorBoard	TensorFlow's visualization toolkit, works with PyTorch	https://www.tensorflow.org/tensorboard
Web front-end	Streamlit	Python package for rapid prototyping of DL/ML-based systems	https://www.streamlit.io/
	Django	Python-based web application technologies	https://www.djangoproject.com/
	Gradio	Python package for rapid DL/ML model demonstrations	https://gradio.app/
	Plotly Dash	Framework to build ML and data science web applications	https://github.com/plotly/dash
	Netlify	Hosting and serverless webapps with GitHub integrations	https://www.netlify.com/
	Hugo/Hugon	Rapid static site generator	https://gohugo.io/

Table 9. Open-Source Tools for Visualization and Presentation

- **Key Takeaways:**
  - Visualization tools are available to directly visualize (1) the raw, collected data and/or (2) the outputs of an analytics (e.g., DL) procedure.
  - Most web-front end technologies are relying on serverless architectures to help facilitate rapid prototyping and development without employing extensive server stacks (e.g., XAMPP).

An Open-Source Tools Inventory for AI-enabled Analytics – Other Resources

- Since AI is rapidly evolving, it is very important to keep abreast of recent developments to help maximize the value of AI-enabled analytics.

- Three key areas can be monitored to identify the “latest” approaches that could be leveraged for AI-enabled analytics.
  1. **Foundational AI conferences** that offer thoughts on theoretical or fundamental AI.
  2. **Applied AI conferences** that employ or adapt AI for specific application areas.
  3. **Non-peer reviewed public materials** that provide code examples, new applications, tutorials, courses, etc. related to various aspects of AI

An Open-Source Tools Inventory for AI-enabled Analytics – Other Resources

Category	Selected Conference/Platform	Focus of Conference and Description of Resource	Size	Primary Audience(s)
Foundational AI Conferences	NeurIPs	ML and computational neuroscience with topical workshops	1,900 papers in 2020	Academics and industry
	ICML	Fundamental ML methodologies with topical workshops	1,088 papers in 2020	Academics and industry
	ICLR	Constructing and processing representations for ML	860 papers in 2021	Academics and industry
Applied AI Conferences	ACM KDD and IEEE ICDM	Applied ML and data mining conferences with topical workshops	~3K papers in 2020	Academics and industry
	ACM CIKM	Knowledge and information management with topical workshops	1,367 papers in 2020	Academics and industry
	AAAI	Conference focused on promoting	1,594 papers in 2020	Academics and industry
	ICCV and CVPR	Applied and fundamental computer vision tasks	~2,400 in 2020	Academics and industry
	ICPPT	A prevailing conference for quantum computing research	Not Listed	Academics and industry
	RAAI and IEEE Robotic Computing	Prevailing conferences for applied and fundamental robotics	Not Listed	Academics and industry
	Open Data Science Conference	AI thought leadership for various application areas	5K+ attendees	Industry
Non-Peer Reviewed Public Materials	ArXiv	Preprint server with published and unpublished work	~2K+ AI pre-prints posted daily	Academic
	Machine Learning Mastery	Online tutorial website for ML with sample code and e-books	1K+ tutorials	Academic, industry, students
	Stack Overflow	Question-Answer site for code related queries	50M questions	
	Papers with Code	Directory of academic AI papers with public code bases	197,327 papers	Academic
	University level courses	MOOCs, publicly accessible courses	Varies	Students
	Companies with open-sourced AI	Companies that use AI that provide their code bases (e.g., Elastic)	Varies	Industry

Table 10. Summary of Other Selected Resources for AI-enabled Analytics

An Open-Source Tools Inventory for AI-enabled Analytics – Other Resources

- Documenting an AI-enabled analytics process is essential to maintaining good progress.

- Common mechanisms include:
  - **IDE's and Package Management:** PyCharm, Jupyter, Anaconda Navigator
  - **Code repositories:** GitHub, Stack Overflow
  - **Communication Software:** Slack, Zoom, Skype, Teams, Outlook
  - **Citation Management:** PaperPile (with plugins), Google Scholar
  - **Note Management and Collaboration:** Confluence, Notability, Evernote
  - **Public presence:** Google Scholar profile, DBLP, Semantic Scholar, personal website

- Keeping these up to date can help you quickly develop a suite of resources to rapidly advance processes and help onboard new members quickly!

Summary

- AI-enabled analytics is a rapidly growing area of modern AI.
  - Has shown promise in high-impact applications (e.g., BI&A, cybersecurity, privacy).
- The AI-enabled analytics process includes (1) Data Collection and Aggregation, (2) Data Extraction and Representation, (3) Analytics, and (4) Visualization and Presentation.
  - Process is based on careful domain/business understanding.
- In this set of slides, a review of prevailing open-source tools for each phase of the AI-enabled analytics process.
  - These slides reflect tools as of April 2022 and will be updated in the future.

References

- Chen, H., Chiang, R. H. L., and Storey, V. C. 2012. “Business Intelligence and Analytics: From Big Data To Big Impact,” *Management Information Systems Quarterly* (36:4), pp. 1165–1188. (https://doi.org/10.1145/2463676.2463712).
- Samtani, S., Kantarcioglu, M., and Chen, H. 2020. “Trailblazing the Artificial Intelligence for Cybersecurity Discipline: A Multi-Disciplinary Research Roadmap,” *ACM Trans. Manage. Inf. Syst.* (11:4), New York, NY, USA: Association for Computing Machinery, pp. 1–19. (https://dl.acm.org/doi/abs/10.1145/3430360).
- Samtani, S., Kantarcioglu, M., and Chen, H. 2021. “A Multi-Disciplinary Perspective for Conducting Artificial Intelligence-Enabled Privacy Analytics: Connecting Data, Algorithms, and Systems,” *ACM Trans. Manage. Inf. Syst.* (12:1), New York, NY, USA: Association for Computing Machinery, pp. 1–18. (https://doi.org/10.1145/3447507).