NUMANTIC SOLUTIONS

UNLOCKING THE POWER OF DATA: HOW TO TREAT DATA AS A STRATEGIC ASSET

NATHAN MORONEY @ 25NTC: ROOM 327 – 1:15 TO 1:45 PM

AUTHORS

Nathan Moroney (Presenter)

Stephen Godfrey (Co-Author)

AGENDA

- **Data as a Strategic Asset**
  - A session on data at a technology conference?
    - Already have been several data related presentations at 25NTC!
  - Data assets
  - Organizational considerations
  - Technical considerations
- **Custom Datasets**
  - What
  - How
  - Why
- **Examples**
  - K-12 program dashboard
  - Environmental field work
  - Federal Register summarization
  - California community colleges policy chat tool
- **Summary and Questions**

NUMANTIC SOLUTIONS

NUMANTIC SOLUTIONS

Numantic Solutions LLC
Palo Alto, CA
numanticsolutions.com
Founded 2024

From wrangling to inference - we can help you get more out of your data. We work with clients in social impact and commercial segments to design, build and use their data environments.

- Data Set Creation & Cleaning
- Cloud Workflows & Databases
- Machine learning & AI
- UI & Data Experience
- Contact Us With Questions

Data as Strategic Asset

It's not just about efficiency

NUMANTIC SOLUTIONS

STRATEGIC DATA ASSETS

Data that builds organizational credibility, supports missions and enables customization. This data also enhances decision making, aligns with expertise and provides independent value. It is increasingly a basis for differentiating AI / Machine Larning.

DATA AS A STRATEGIC ASSET

- The case for strategic data
  - Motivation to build strategic data assets
- Getting started with strategic data
  - Approaches to creating these assets
- Establishing a data program
  - Considerations from a program level
- A personal worked example
  - Open Patent Gazette
  - A weekly data set relating to US Patents
- End result is a custom data set (next section of talk)
  - Custom data is often strategic data

NUMANTIC SOLUTIONS

THE CASE FOR STRATEGIC DATA

- Defensibility
  - Proprietary datasets remain unique assets even in a world of open-source tools and artificial intelligence.
- Affordability
  - Advances in open-source technology and access to public data make data collection more cost-effective than ever.
- Leadership
  - Organizations with well-structured data can lead their industries and set new benchmarks.
- Persuasiveness
  - Data-driven arguments carry more weight in advocacy and decision-making.
- Operational Efficiency
  - Data fuels automation and intelligent workflows, reducing inefficiencies and cutting costs.

GETTING STARTED WITH STRATEGIC DATA

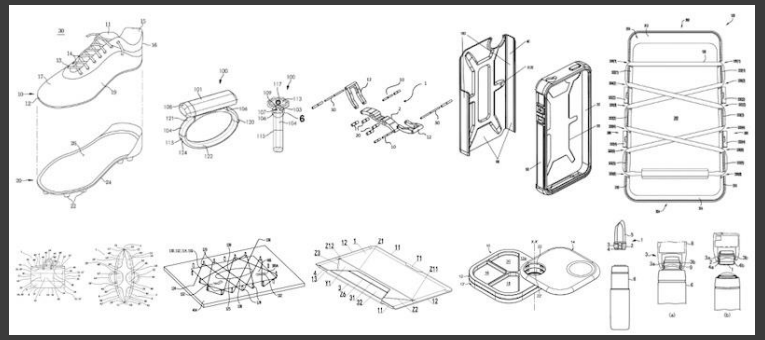
- Define Objectives (But Stay Flexible)
  - While having clear goals helps, data exploration itself often reveals new opportunities. The process should be iterative, refining objectives as insights emerge.
- Leverage Internal Data
  - Organizations should first assess their existing knowledge base—documents, emails, internal reports—to identify valuable information.
- Explore Public Data Sources
  - Government, university, and nonprofit datasets can provide valuable, cost-effective insights.
- Consider Private Data Vendors
  - Purchasing raw data from specialized providers may be worthwhile for certain use cases.
- Develop Proprietary Data Collection Methods
  - If no existing data meets an organization's needs, direct data collection (e.g., surveys, sensors, web scraping) may be necessary.

ESTABLISHING A STRATEGIC DATA PROGRAM

- Allocate Resources
  - Budgeting for data collection, storage, and analysis ensures long-term success.
- Define Ownership
  - Assign responsibility for data management to dedicated teams or roles.
- Implement Data Governance
  - Simple but effective policies on data security, storage, and access control enhance reliability.
- Make Technology Decisions
  - Once strategic objectives are clear, selecting appropriate data management tools becomes much easier.
- Understand Data Frequency and Life Cycle
- Plan Forward
  - Once the primary deliverables are achieved, what are the maintenance, support and possible future extensions
- Manage Risks
  - Identify and communicate any associated risks and manage expectations

PERSONAL EXAMPLE

A tool to convert weekly web publication from the US Patent Office to a single spreadsheet



OPEN GAZETTE

- Case For : Affordability
  - Single developer part time sufficient for prototype
- Getting Started : Public Data
  - Published to the public domain
  - Approximately 5,000 documents a week
- Data Program : Technology Decisions
  - Local python application
  - Zipped data directly to CSV file
- Data Program : Plan Forward
  - A future cloud-based version would be nice

https://github.com/NumanticSolutions/OpenGazette

NUMANTIC SOLUTIONS

Custom Data Sets

Another way to differentiate your efforts

NUMANTIC SOLUTIONS

EXAMPLES TEMPLATE

For the following four examples consider the : what, how and why

CUSTOM DATASETS

- What
  - Data exists, is generated or can be created for a wide range of domains
  - Spreadsheets, databases, repositories, web sites, sensors, government agencies, surveys, commercial sources, interviews, ...
- How
  - Identify existing resources
  - Complement or augment these resources
  - Field work or experiments to generate
  - Targeted gap filling efforts
- Why
  - Unique data can complement organizational strengths and capacities
  - It doesn't have to be a substantial investment

NUMANTIC SOLUTIONS

Examples of Data Assets

Specific use cases to illustrate benefits

NUMANTIC SOLUTIONS

DATA STAKEHOLDERS

Dashboard example included multiple stakeholders for the custom dataset. Details will vary, but as much as possible making data available as soon as possible allows early feedback and provides visibility for the program objectives.

K-12 PROGRAM DASHBOARD

- What
  - State level databases were collected into a data warehouse
  - School administrators completed a targeted survey relating to social and emotional learning programs
  - Additional national databases and resources were identified
- How
  - All databases hosted in a single cloud environment
  - Survey run via a service and resulting data added to cloud environment
  - Devise aggregation metrics on a per school basis via a web application
- Why
  - Administrators provided with a overview and specific guidance on future action plans
  - Possible future point of integration for surveys

NUMANTIC SOLUTIONS

DATA FILTERING

Custom data sets also include specifying and processing to match the data to the objectives.

Workflow to achieve reduce the raw environmental scans by factor of 100.

ENVIRONMENTAL FIELD WORK

- What
  - Characterizing the nature and changes in a given geography benefits from hands on data collection
  - This can include heterogeneous data sources and types of data
  - Density and scale of modern sensors offer to opportunities for analysis, but also challenges for management
- How
  - Explore cloud integration options
  - Understand workflow requirements and budget
  - Strategically implement data processing steps
- Why
  - Improve scope and timeliness of enviromental data
  - Enable new areas of analysis and publication

NUMANTIC SOLUTIONS

DERIVED DATA

Collected data can be post-processed to generate derived data sets.

Summarization and style revision useful examples (patent summaries can be generated in a style that is less legal).

FEDERAL REGISTER SUMMARIZATION

- What
  - Monday through Friday, the Federal Register publishes 50 to 100 documents relating to US government agencies
  - Publications provided in a range of formats and many lack a summary or abstract
  - Keeping up with this rapidly changing collection is both important and time consuming
- How
  - Automated document retrieval via API
  - Concise summarization via AI
- Why
  - Use targeted summaries to engage prospective clients and partners
  - Provides an example of data-differentaited AI

NUMANTIC SOLUTIONS

WORKED EXAMPLE

An open implementation that provides a more detailed end-to-end example.

Recent project by Numantic Solutions, not a product so can be discussed in more detail than might otherwise be possible.

CA COMMUNITY COLLEGE POLICY ASSISTANT

- What
  - We were interested in creating a policy assistant for the California Community Colleges as a chat bot
  - Data and documents exist, but are spread across multiple organizations and web sites
  - No integrated index or interface to this data
- How
  - Crawl relevant web sites for documents and data
  - Generate a vector database in cloud
  - Use vector database for RAG or Retrieval Augmented Generative pipeline
  - Develop a Streamlit application for user feedback
- Why
  - Assist in understanding current policy considerations for CA Community Colleges
  - End-to-end example of custom data set as strategic asset

NUMANTIC SOLUTIONS

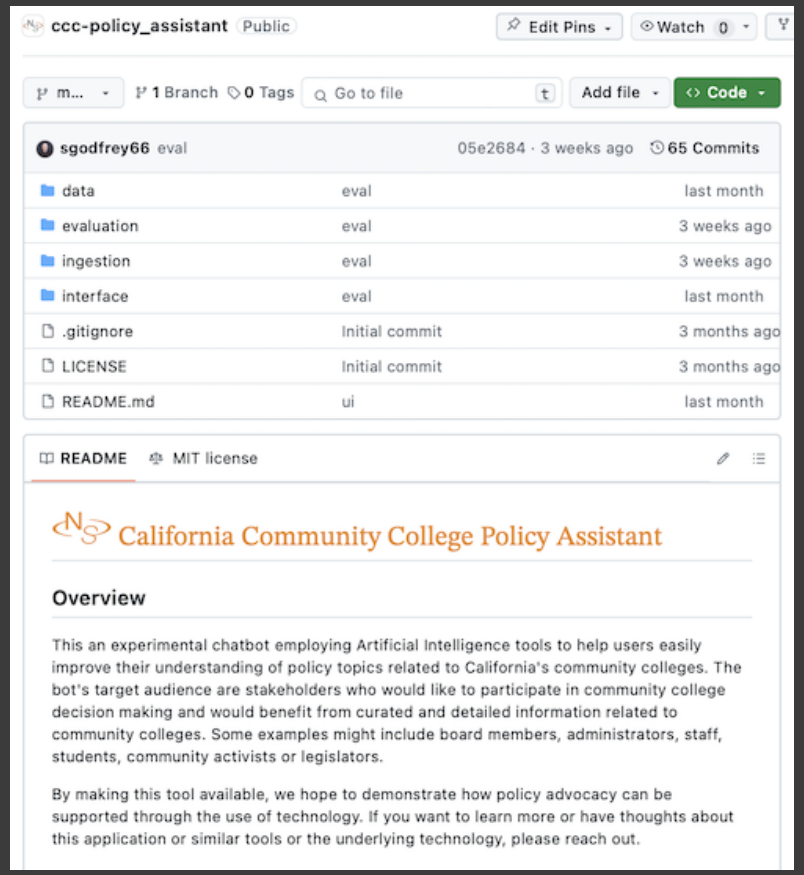
WEB APPLICATION

Chat interface hosted in the cloud, includes RAG or retrieval augmented generation.

Questions are supplemented with results of searching a custom dataset.

CA COMMUNITY COLLEGE POLICY ASSISTANT

PROJECT REPOSITORY



CA COMMUNITY COLLEGE POLICY ASSISTANT

- GitHub
  - https://github.com/NumanticSolutions/ccc-policy_assistant
- Overview
  - "This an experimental chatbot employing Artificial Intelligence tools to help users easily improve their understanding of policy topics related to California's community colleges. The bot's target audience are stakeholders who would like to participate in community college decision making and would benefit from curated and detailed information related to community colleges. Some examples might include board members, administrators, staff, students, community activists or legislators.:

NUMANTIC SOLUTIONS

CUSTOM DATASETS

A dozen (and growing) set of sources of documents and tabular data relating to CA Community Colleges

CA COMMUNITY COLLEGE POLICY ASSITANT

- Legislative Analyst's Office
- Community College League of California
- California Community Colleges
- California Community College Association for Occupational Education
- National Center for Education Statistics (NCES) – IPEDS
- Community College Research Center (CCRC) - Columbia University
- American Association of Community Colleges (AACC)
- U.S. Department of Education - College Scorecard
- National Student Clearinghouse Research Center
- Wikipedia
- The Aspen Institute - College Excellence Program
- Education Commission of the States (ECS)
- Integrated Postsecondary Education Data System (IPEDS) - National Center for Education Statistics (NCES)

NUMANTIC SOLUTIONS

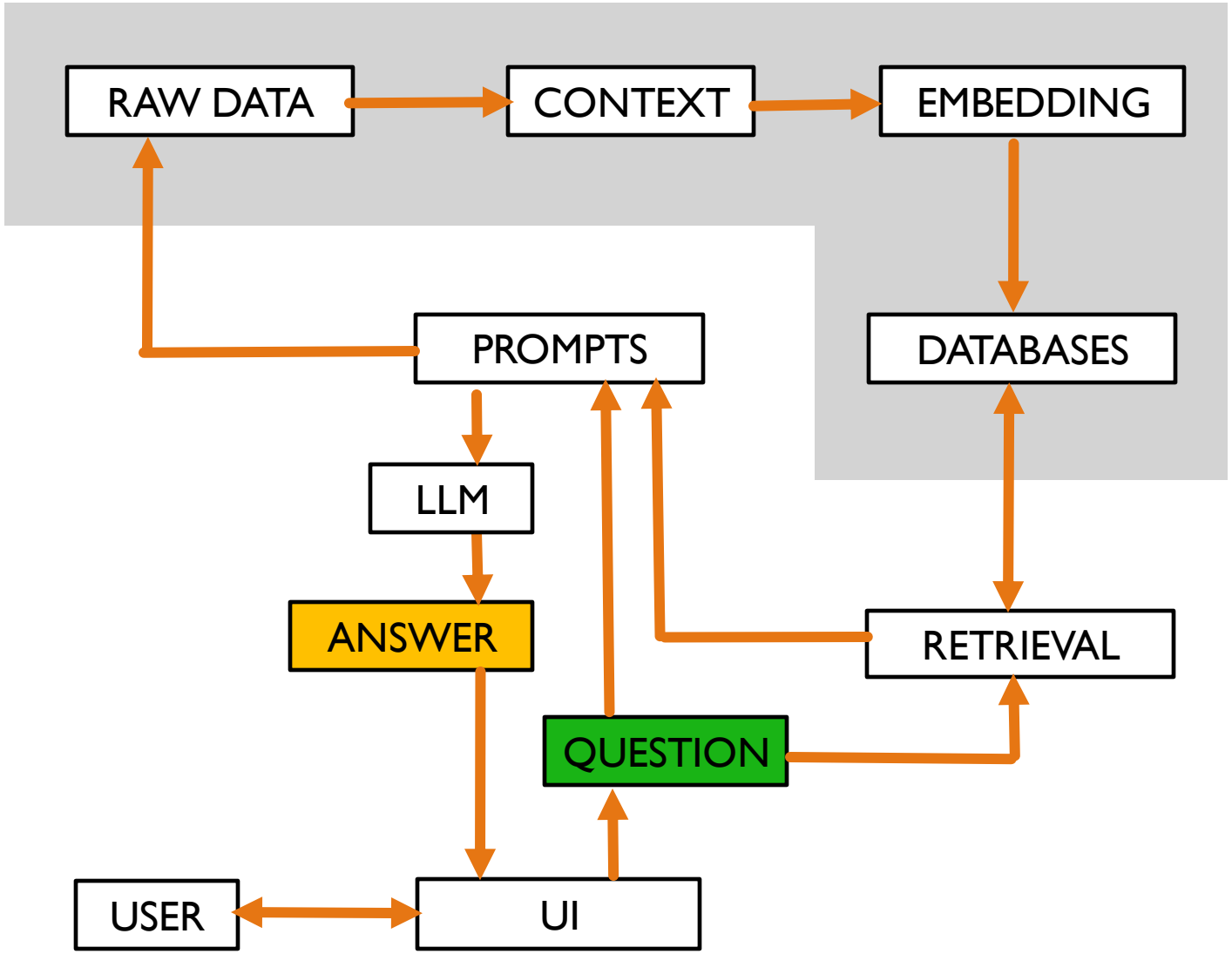
COLOR CODING

Gray region are steps run when adding new data to the assistant.

Green box is user question input to the system.

Orange box is the answer to the users question.

ASSISTANT PIPELINE



NUMANTIC SOLUTIONS

SPECIFIC ITEM

Technical details of the current implementation

Key in-progress component : testing, using a mix of LangChain and cutom code (next slide)

ASSISTANT COMPONENTS

- Technological Components
  - Coding Language: Python
  - RAG Framework: LangChain
  - Vector Database: Chroma
  - Embedding Model: Text-embedding-004
  - AI Large language Model (LLM): Gemini-1.5-pro
  - Document & Vector Storage: Google Cloud Storage
  - User Interface: Google Cloud Run and Streamlit

NUMANTIC SOLUTIONS

EXAMPLE OF CHAT REPLAY

- Using custom data set with cloud LLM effectively eliminated certain types of hallucination compared to a local only LLM without custom data
- Exercise of crafting domain questions (and answers) is its own pedagogy task

California Community Colleges : Policy-Assistant

A sampling of previously asked questions and corresponding answers.

SUMMARIZE HOW COMMUNITY COLLEGE PROGRAMS CAN BE ACCESSED FROM CALIFORNIA STATE PRISONS.

All 116 California Community Colleges partner with the California Department of Corrections and Rehabilitation (CDCR) to offer courses inside state prisons. These programs primarily focus on helping incarcerated individuals earn their GED or high school diploma and provide vocational training. This access to education aims to equip individuals with the skills and knowledge needed for successful reintegration into society upon release.

Summary and Questions?

Data can differentiate expertise, applications and AI

NUMANTIC SOLUTIONS