Data Quality: The Backbone of Effective Analytics and AI

Adam Kinard
Refinery Data Scientist at bp

Technology

Refinery Data Scientist

Adam Kinard
Data Scientist
BP—Cherry Point Refinery
Blaine, WA

- Use **Data Science Toolkit** to solve **High Impact work** not meant for product teams
- Deliver quick insights, Accelerate Root-cause analyses, Automate workflows
- Help refinery **identify valuable** Data Science **use cases**
- Combine **data science expertise** with **downstream business knowledge** for efficient and effective solutions

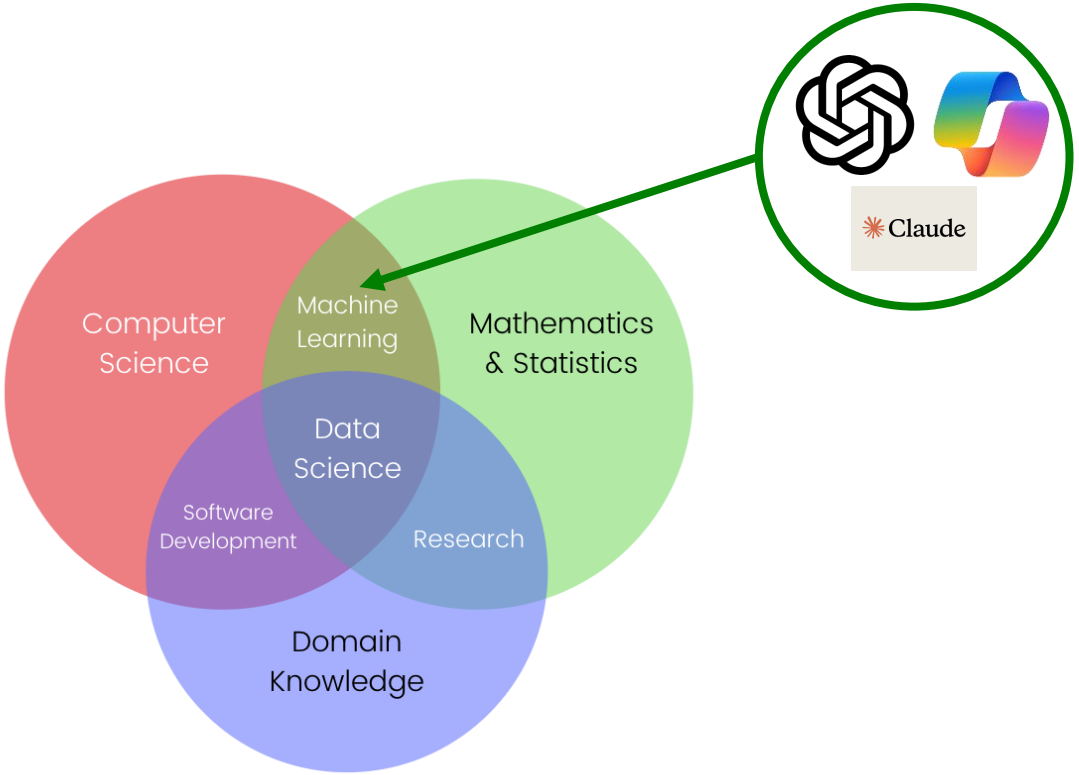
Technology

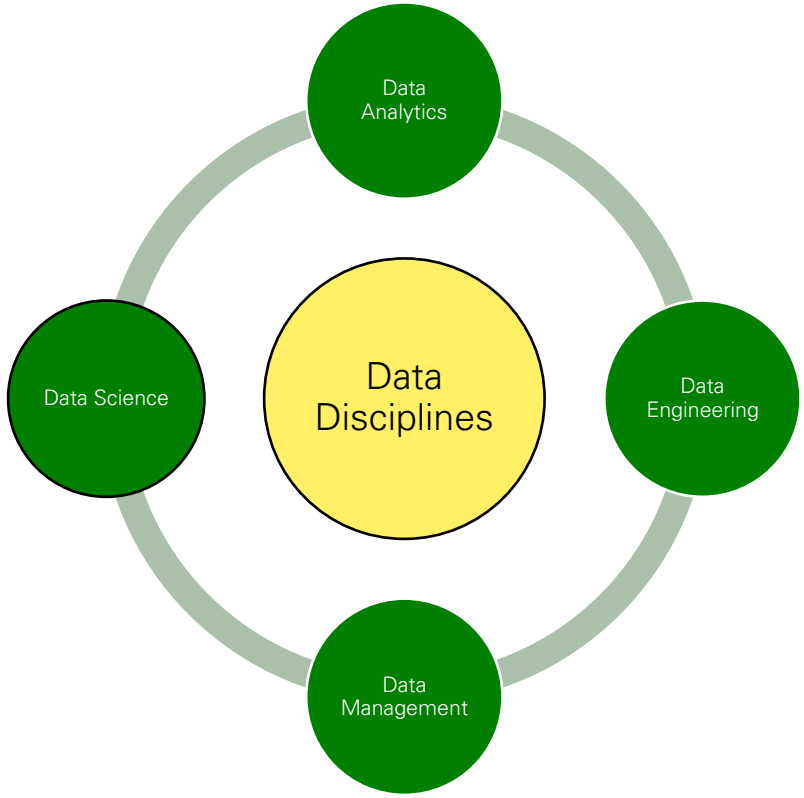
Project Examples

- Insights and Correlations from Process Unit Feed Rates, temperatures, and pressures
- Large Language Models (LLMs) trained on Turnaround Reports, Equipment History, Reference Material
- Predictive Models for Product % Yields and Quality
- Automated Reports on Health and Safety Incident Data

What is Data Science

- Mixture of statistics, computer science, information science, and domain expertise to analyze, interpret and gain knowledge and insights from structured and unstructured data
- These actionable data insights are used to make informed decisions





Technology

“

A Model is only as good as its Data.

”

What is Data Quality?

- Accuracy
- Timeliness
- Reliability
- Relevance
- Completeness
- Quantity
- Accessibility
- Consistency

What is Data Quality: A Basketball Example

What is a Player's Shooting% Over a Season?

Accuracy

- A Player shot 6/10 3-Pointers, but the system is showing 10/10
- Overestimation of a player's ability could occur

Timeliness

- Stats from the previous year carry over into the stat sheet.
- Real-time feedback is unavailable

Reliability

- Stats are recorded from a distracted spectator, instead of the official scorekeeper.
- Unreliable data can introduce bias

Relevance

- The player's shoe color being recorded.
- Irrelevant data can clutter a model and dilute meaningful patterns

What is Data Quality: A Basketball Example

What is a Player's Shooting% Over a Season?

Completeness

- All games track FGs and FTs, but only a few games track 3PTs.
- Empty values or inconsistent sampling removes model confidence

Quantity

- There are 82 games over a season, but the system uses data from just 10.
- Fewer instances results in less confidence

Accessibility

- Your scorekeeper only allows special access to stats, or they are stored on one analyst's computer.
- Valuable insights are delayed or lost entirely

Consistency

- Some games Free Throws are counted and some games they aren't.
- Inconsistent comparisons can flaw results

Importance of Data Quality

- The foundation upon which successful data science and AI projects are built
- Enables accurate insights, efficient processes, trustworthy results, compliance, and scalability
- Understanding the type of data collected, and being organized in how it's handled leads to effective Model Utilization!

Improving Data Quality

Key components

Clear Data Management Goals

- File naming conventions
- Folder structures and locations
- Objectives from data storage and leveraging

Data Governance Framework

- Quality: How confident are we in the values of our data?
- Security: How sensitive is the data?
- Compliance: Are we adhering to all relevant regulations and standards

Identify any Data Gaps

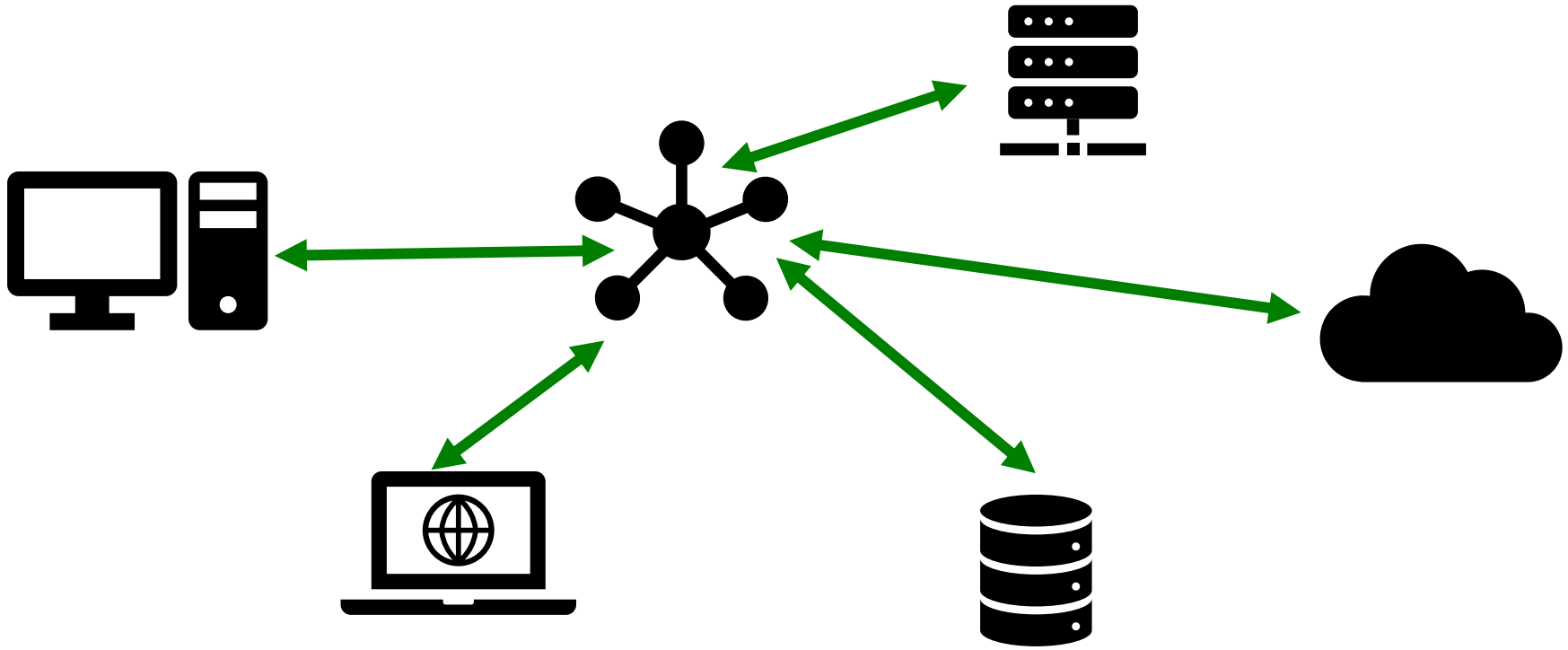
- Are there useful insights we cannot realize because we aren't recording the right data?
- Frequency of data Collection
- Different mediums of collection

Data Audits & Compliance Checks

- Are we adhering to all relevant regulations & standards
- Routine Checks of Data Quality

Data Engineering

- The practice of designing and building systems for the collection, storage and analysis of data at scale
- Crucial for enabling subsequent data analysis and data science activities
- Data Engineers create infrastructure that allows organizations to process and utilize large datasets to gain real-time insights



Prompt Engineering

- The process of designing and refining prompts to guide generative AI models, such as LLMs, to produce desired outputs
- Crafting **specific, clear, and contextually rich inputs** that the AI can interpret effectively to generate accurate and relevant responses

Tips:

- Break a task down into multiple sections or smaller answers
- Provide context of a particular data source and examples to improve LLM understanding and result
- Ask for details and justification so output will have reasoning for its answer

Key Takeaways

- **Data Quality** is paramount to enable an organizations effective utilization of Data Science and AI
  - Accuracy, Timeliness, Reliability, Quantity, Relevance, Completeness, Accessibility, Consistently

- A Machine Learning Model is only as good as its Data!

- An organization needs a **Data Governance Framework**:
  - Quality, Security, Compliance
  - Clear **Data Management** Goals
  - **Data Engineering** and **Data Science teams** will save time and money for your organization

- **Prompt Engineering** skills are increasingly valuable in the modern world

Thank You

Adam Kinard
Refinery Data Scientist at bp

Technology