



Artificial Intelligence & Machine Learning: Building the Right Architectural Foundation

Donna Burbank
Global Data Strategy, Ltd.

July 27, 2023



Donna Burbank



Donna is a recognised industry expert in data management with over 25 years of experience in data strategy, data governance, data modeling, metadata management, and enterprise architecture. Her background is multi-faceted across consulting, product development, product management, brand strategy, marketing, and business leadership.

She is currently the Managing Director at Global Data Strategy, Ltd., an international data management consulting company that specializes in the alignment of business drivers with data-centric technology.

In past roles, she has served in key brand strategy and product management roles at CA Technologies and Embarcadero Technologies for several of the leading data management products in the market.

As an active contributor to the data management community, she is a long time DAMA International member, contributor to the DMBOK 2.0, Past President and Advisor to the DAMA Rocky Mountain chapter, and was awarded the Excellence in Data Management Award from DAMA International.

She has worked with dozens of Fortune 500 companies worldwide in the Americas, Europe, Asia, and Africa and speaks regularly at industry conferences. She has co-authored several books and is a regular contributor to industry publications. She can be reached at donna.burbank@globaldatastrategy.com
Donna is based in Boulder, Colorado, US.



DATAVERSITY Data Architecture Strategies

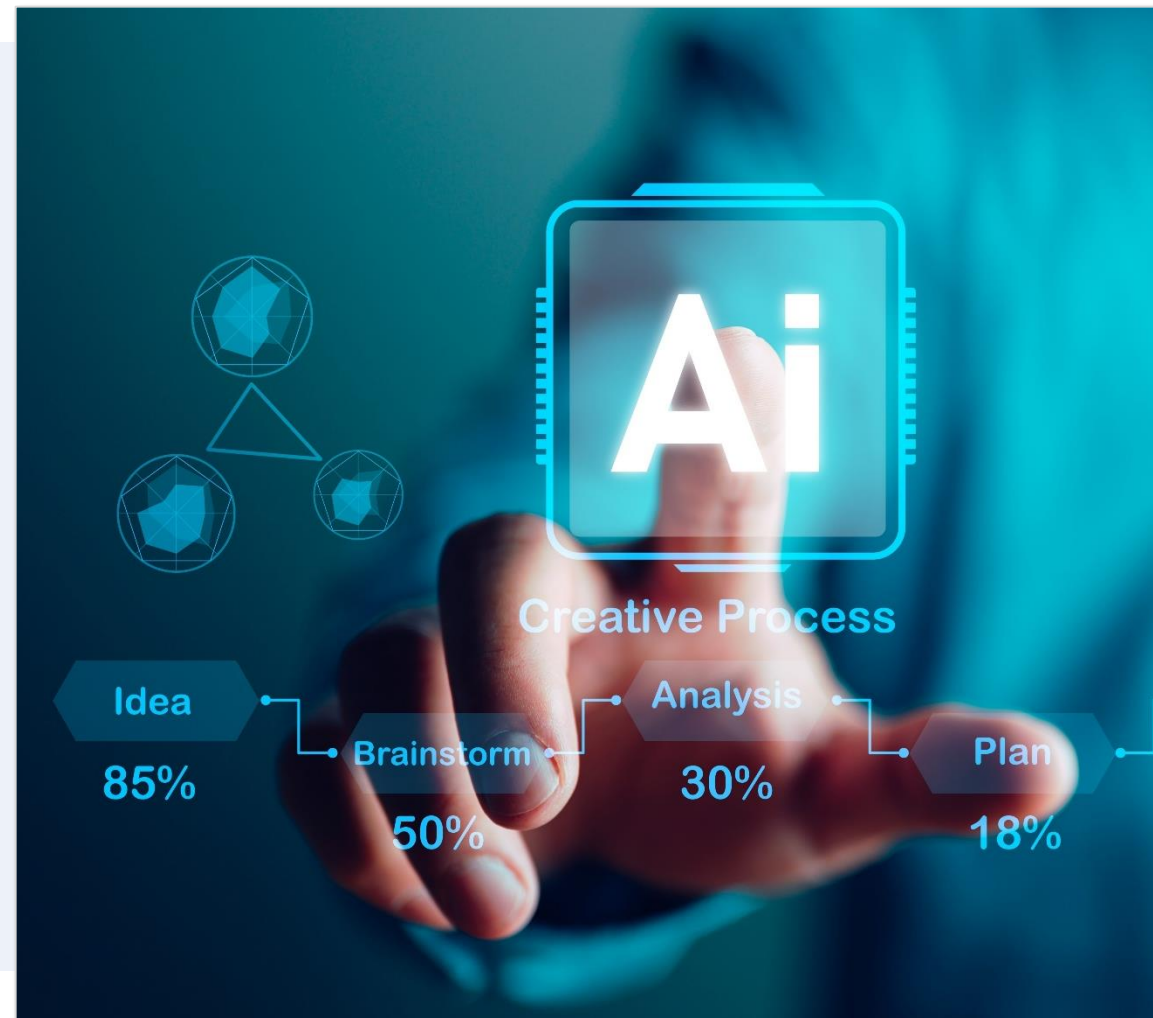
This Year's Lineup

- **January** Emerging Trends in Data Architecture – What's the Next Big Thing?
- **February** Building a Data Strategy - Practical Steps for Aligning with Business Goals
- **March** Data Mesh or Data Mess? Separating the Reality from the Hype
- **April** Master Data Management - Aligning Data, Process, and Governance
- **May** How do Data Governance & Data Architecture Support Each Other?
- **June** Why You Need Data Management – Getting Executive Buy-In
- **July** Artificial Intelligence and Machine Learning – Building the Right Architectural Foundation
- **August** Data Quality Best Practices (with Nigel Turner)
- **September** Best Practices in Metadata Management
- **October** Designing Data for Business Intelligence & Analytics – Where the Star Schema Fits in a Modern Data Architecture
- **December** Enterprise Architecture vs. Data Architecture



What We'll Cover Today

- Artificial intelligence (AI) and machine learning (ML) are increasing in popularity as more organizations are looking to become more data-driven.
- To support strong AI/ML models and algorithms, accurate and timely data is needed, supported by a strong Data Architecture.
- This webinar discussed how to create a robust Data Architecture for AI and ML that takes both business and technology needs into consideration.



AI – Risk to Humanity?

The New York Times

A.I. Poses 'Risk of Extinction,' Industry Leaders Warn



Risks from Artificial Intelligence

San Francisco Chronicle

Yes, AI poses an extinction risk to humanity. And not just for the obvious reasons



- From 2001: A Space Odyssey (1968)

AI: Boon for Humanity?

**The
Guardian**

**Five ways AI could improve the world:
'We can cure all diseases, stabilise our
climate, halt poverty'**

Forbes

**Artificial Intelligence For
Good: How AI Is Helping
Humanity**

FORTUNE

**3 reasons why VC billionaire Marc Andreessen
thinks 'A.I. is quite possibly the most important—and
best—thing our civilization has ever created'**

andreessen.
horowitz It's time to build

Why AI Will Save the World

by Marc Andreessen

“Any sufficiently advanced technology is indistinguishable from magic”.

- Arthur C. Clarke, 1962



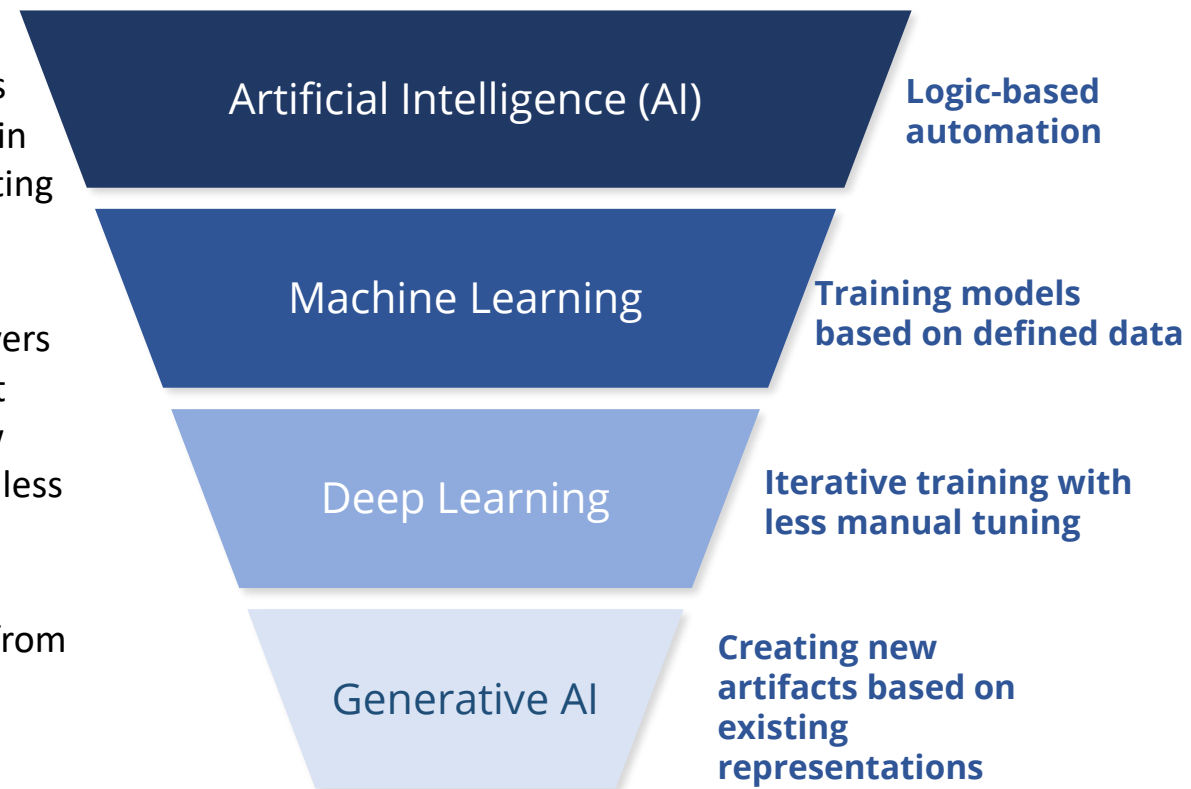
Let's Start with Some Definitions

Artificial intelligence (AI) applies advanced analysis and logic-based techniques, including machine learning, to interpret events, support and automate decisions, and take actions.

Advanced **Machine Learning** algorithms are composed of many technologies (such as deep learning, neural networks and natural language processing), used in unsupervised and supervised learning, that operate guided by lessons from existing information.

Deep learning is a variant of machine learning algorithms. It uses multiple layers to solve problems by extracting knowledge from raw data, and transforming it at every level. These layers incrementally obtain higher-level features from the raw data, allowing the solution of more complex problems with higher accuracy and less manual tuning.

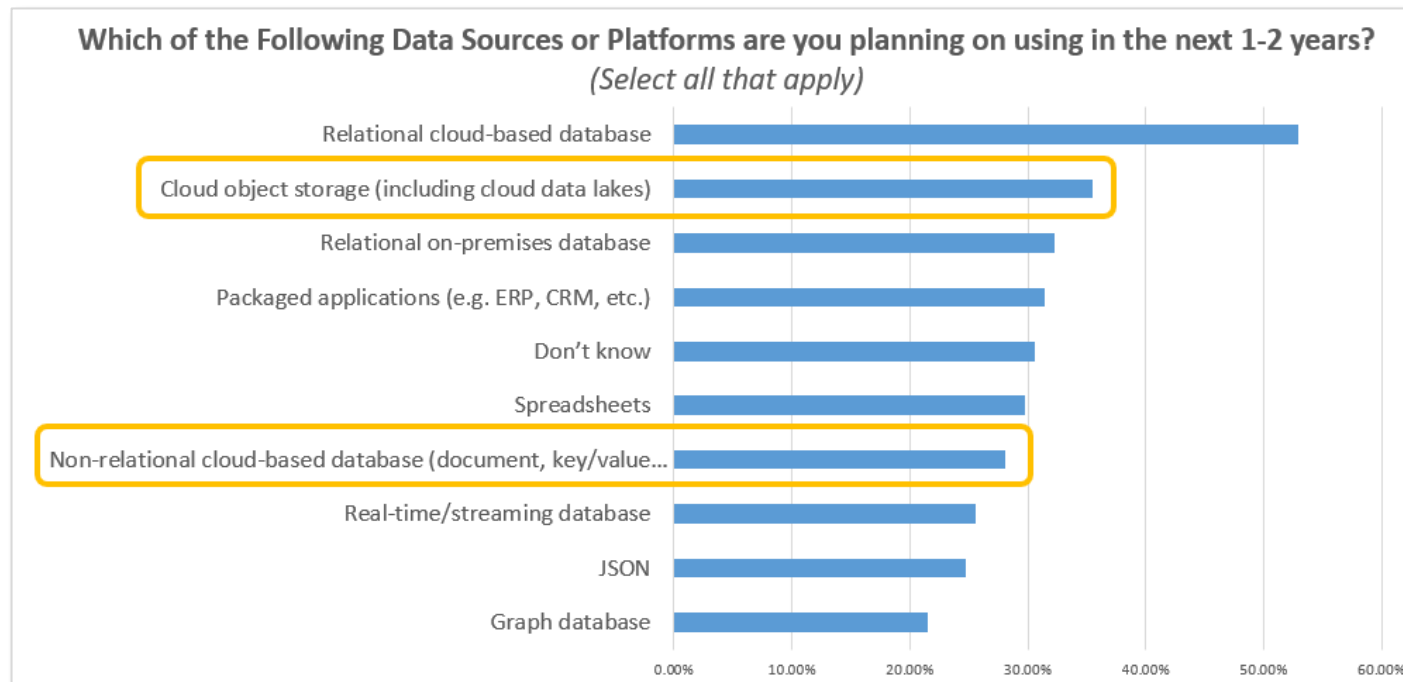
Generative AI refers to AI techniques that learn a representation of artifacts from data, and use it to generate brand-new, unique artifacts that resemble but don't repeat the original data. These artifacts can serve benign or nefarious purposes. Generative AI can produce totally novel content (including text, images, video, audio, structures), computer code, synthetic data, workflows and models of physical objects. Generative AI also can be used in art, drug discovery or material design.



AI – Why Now?

Artificial Intelligence & Machine Learning are not new

- AI and ML concepts have arguably been around since the 1950s.
- Many of us learned these concepts in the “old days” at university



From Trends in Data Management, 2022, DATAVERSITY, by Donna Burbank and Keith Foote

Improvements in computing power and processing have allowed us to harness the power of AI

- **Scale & Volume** of Data Storage
- **Computing Power & Processing Speed**
 - **CPUs** driving Machine Learning
 - **GPUs** driving Generative AI:

A GPU/Graphics processing unit, a specialized processor originally designed to accelerate graphics rendering. GPUs can process many pieces of data simultaneously

AI / Machine Learning Basics

Some common basic steps for AI/machine learning

Gather the Data

- What factors do I want to focus on?
- Where will I source the data to train my model?
- What is the volume of the data set?
- Etc.

Prepare the Data

- Analyze/Visualize the data to understand patterns, relationships, etc.
- Is it a realistic mix of factors?
- Randomize the order.
- Etc.

Choose the Model

- What model is the best fit for the scenario at hand?, e.g.
 - Linear Regression
 - Logistic Regression
 - Naïve Bayes
 - Random Forest
 - Etc.

Train the Model

- Initialize parameter values & run the model with those values.
- Compare model's predictions with expected output
- Adjust the values to have more correct results.



Evaluate & Tune

- Run the model against data it has never seen.
- Compare to desired result and tune parameters as needed.

Quality Data is the Foundation for AI

Data Quality & AI Myths – Heard in the Real World

“Data Quality isn’t important for Machine Learning & Data Science – We’ll make up for lack of quality with higher volumes of data.”

“We don’t need to worry about governance and security at this point – it’s just sandbox data.”

Myths

Quality Data is the Foundation for AI

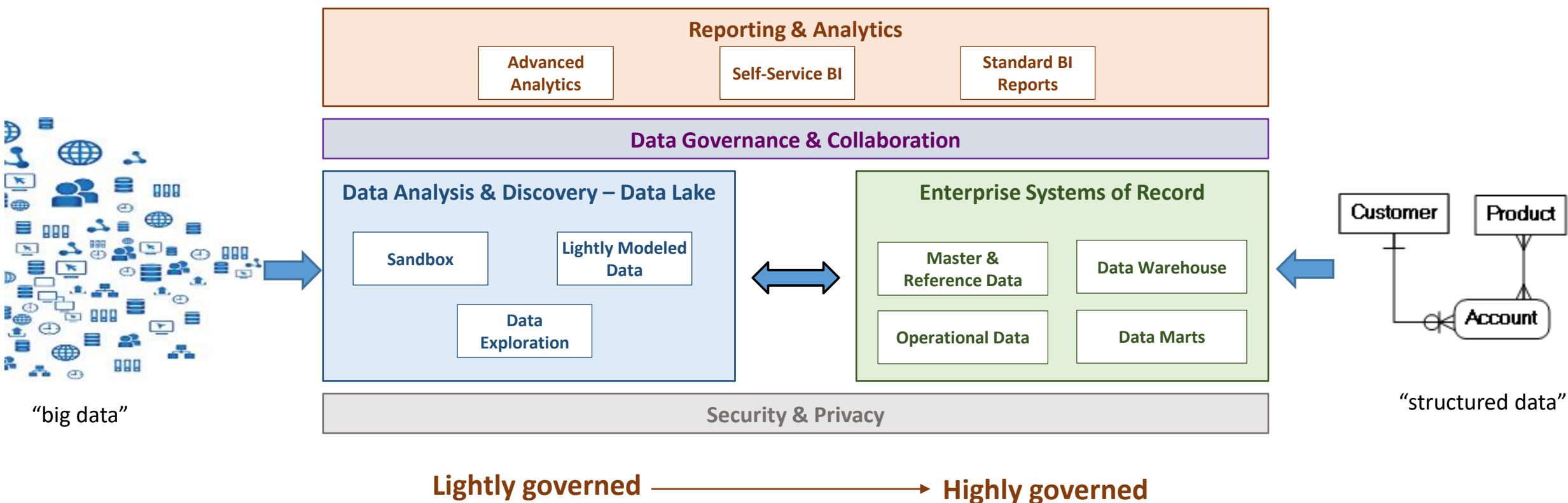
Truth

Data Scientists typically spend 80% of their time cleaning data.

Source: Forbes

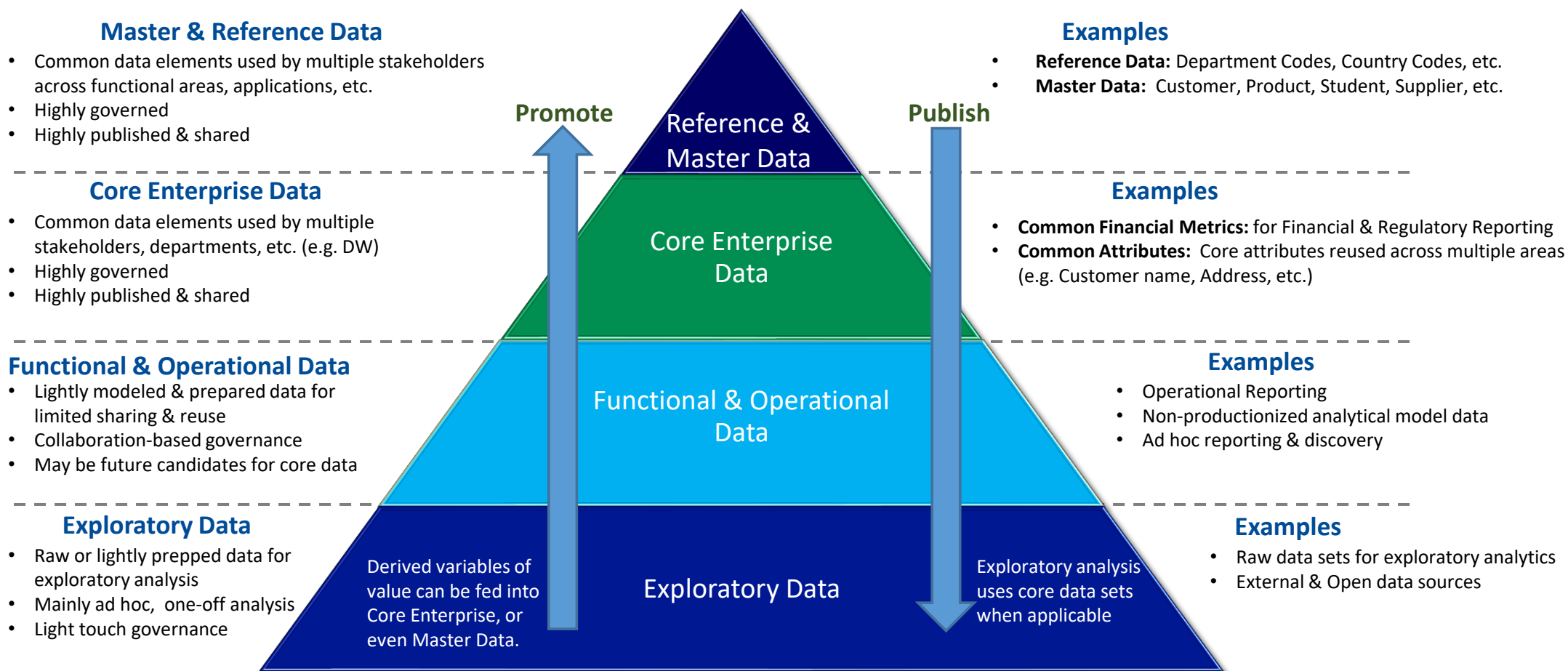
Provide an Integrated Data Architecture Ecosystem

- A modern data architecture provides zones for exploration & discovery
- ...combined with trusted, vetted data sets
- ... with a layer of governance and security underpinning each.



Implement “Just Enough” Data Governance - Allow for Iteration & Discovery

- Know **what to manage closely** and **what to leave alone**
- **The more the data is shared across & beyond the organization, the more formal governance needs to be**



Different Data Modeling & Storage Patterns Exist for AI/ML

Third Normal Form is the only way to go!!

I store everything in arrays!!



Both

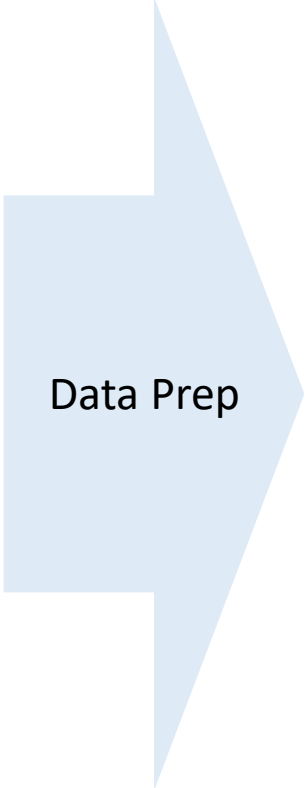


And

Organizing/Storing Data for Machine Learning

Storing Data for Operations or Reporting

- Relational
- Dimensional
- Key-Value
- JSON, ML
- Document Store
- Property Graph
- Spreadsheet (!)
- Etc.



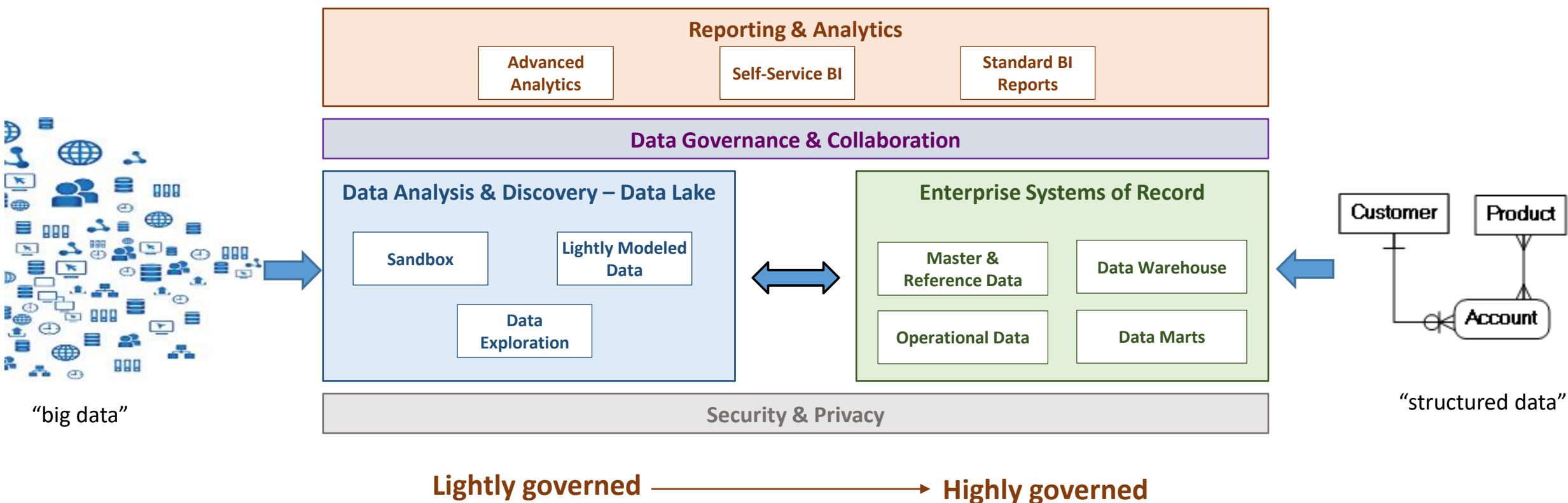
Data Prep

How Algorithms Work with Data

- Arrays
- Stacks
- Queues
- Trees
- Graphs
- Flattened Tables
- Etc.

Provide an Integrated Data Architecture Ecosystem

- A modern data architecture provides zones for exploration & discovery
- ...combined with trusted, vetted data sets
- ... with a layer of governance and security underpinning each.



Applications & Use Cases for AI/ML



Machines Learn Like Humans Do

In many ways, computers learn the same way we do

- Computer algorithms can “learn” just like humans do.



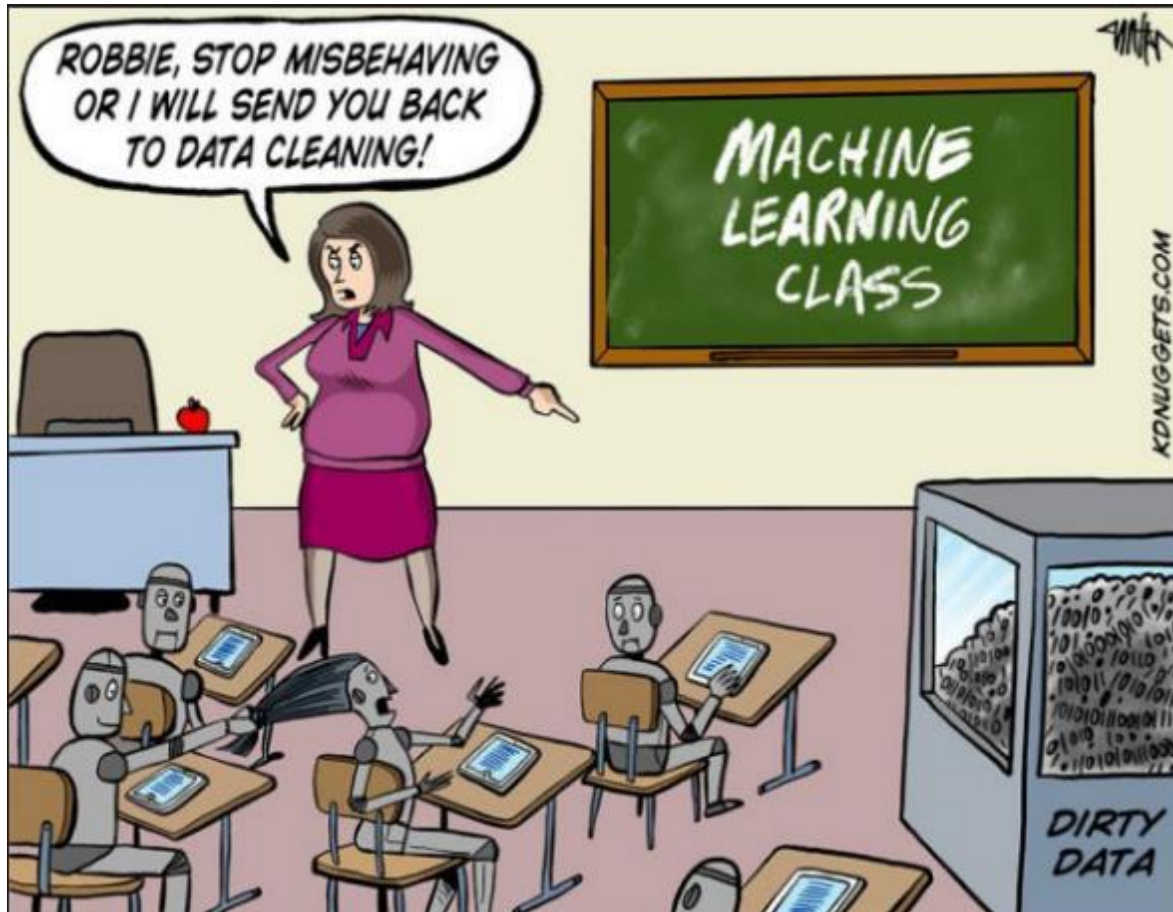
Dog?



This is a **DOG**, Johnny!
Look at the **DOG**!



Use Case: Machine Learning & Metadata Discovery



Source kdnuggets.com

- Machine Learning offers ways to automate tedious tasks that may have been done manually before:
 - e.g. Data Mapping
 - SSN -> Field1_SSN
 - SSN -> Soc_Num
 - Etc.
 - Machine Learning Pattern Matching
 - NNN-NN-NNNN -> Field_X follows this pattern, it must be a SSN
- There is a place for both methods:
 - Sometimes you want to define specific mapping rules
 - Sometimes you want a pattern-matching, discovery-style approach.

Machine Learning & Metadata Discovery

SSN?



917-98-2765

This is a **SSN**, Johnny!
Look at the **SSN**!

043-27-1975

013-11-1238

978-65-1239

097-27-9875

285-58-2945

111-11-1111

Machines Learn Like Humans Do

In many ways, computers learn the same way we do

- In many ways, we “learn” conditions responses to typical questions or situations.



Mine?

What do you say, Marco?

You say THANK YOU!

You say THANK YOU, Marco!

Marco! Say THANK YOU!

Machines Learn Like Humans Do

In many ways, computers learn the same way we do

- Most of us generally improve over time...



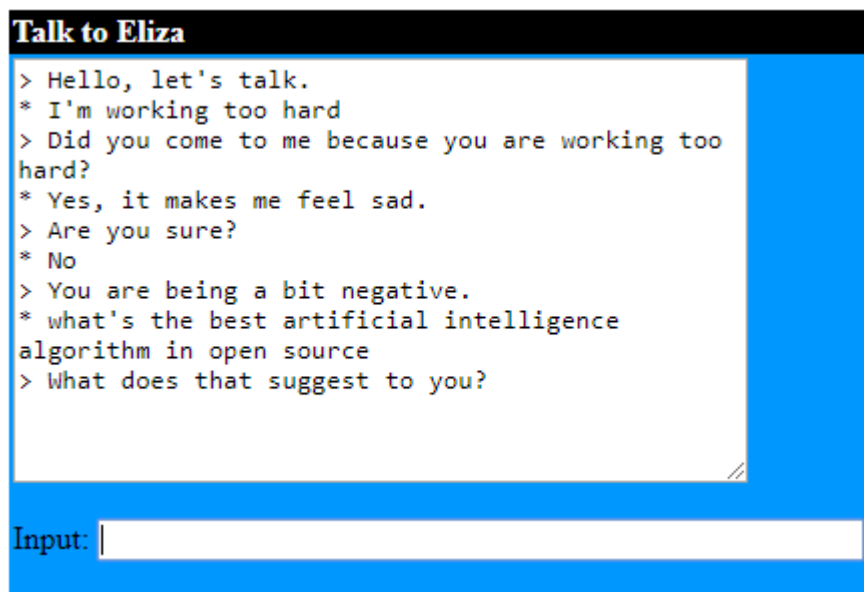
I'm fine, and you?

How are you?

Chat Bots

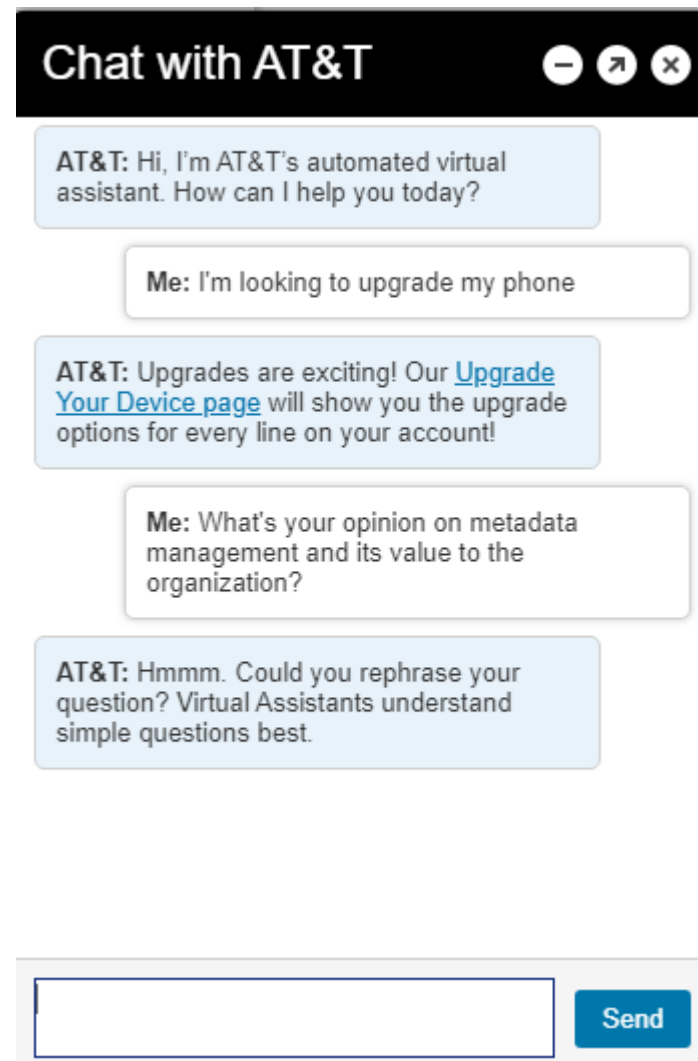
Automating common questions

- Chat bots are a common way to provide automated answers to common questions.



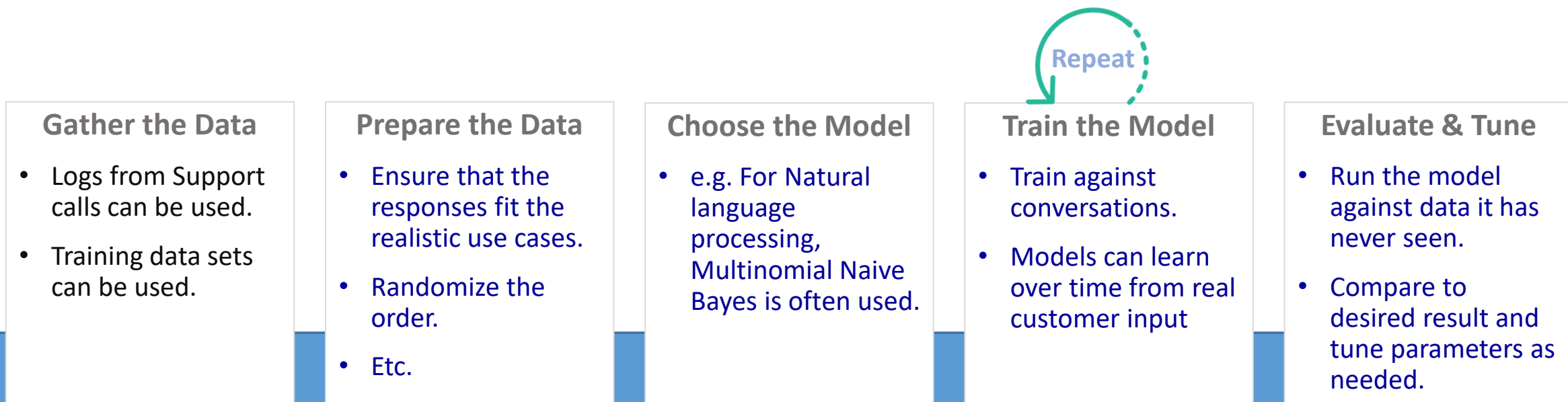
Eliza is still learning! Please let me know your experience with the computer therapist, and anything you might want to see improved.

<https://www.cyberpsych.org/eliza/>



Chat Bot Basics

Some common basic steps for building Chat bots



Quality Data is the Foundation for Chat Bots

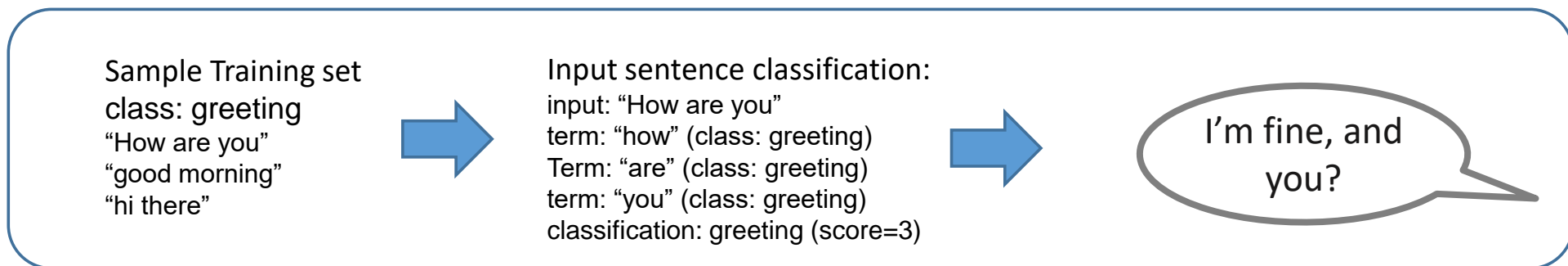


Image Recognition

Identifying patterns

- By now, most of us have seen the Muffin or Chihuahua graphic



Image Recognition

- Labelled data sets can help with training algorithms.

The screenshot shows the ImageNet website interface. At the top, there is a search bar with the text 'IMAGENET' and a 'SEARCH' button. Below the search bar, there are navigation links for 'Home', 'About', 'Explore', and 'Download'. The main content area displays the search results for 'Chihuahua'. The title 'Chihuahua' is followed by a description: 'An old breed of tiny short-haired dog with protruding eyes from Mexico held to antedate Aztec civilization'. To the right of the description, there are statistics: '1750 pictures', '68.81% Popularity Percentile', and a 'Wordnet IDs' icon. Below the description, there is a 'Numbers in subjects' section with a tree view of hierarchical categories. The categories include 'ImageNet 2011 Fall Release (32326)', 'plant, flora, plant life (4486)', 'geological formation, formation (175)', 'natural object (1112)', 'sport, athletics (176)', 'artifact, artefact (10504)', 'fungus (308)', 'person, individual, someone, somebody, mortal, soul (5970)', 'animal, animate being, beast, brute, creature, fauna (3998)', 'invertebrate (766)', 'homeotherm, homootherm, homotherm (0)', 'work animal (4)', 'darter (0)', 'survivor (0)', 'range animal (0)', 'creepy-crawly (0)', 'domestic animal, domesticated animal (213)', 'domestic cat, house cat, Felis domesticus, Felis catus (10)', 'dog, domestic dog, Canis familiaris (189)', 'pooch, doggie, doggy, berker, bow-wow (0)', 'hunting dog (101)', 'dalmatian, coach dog, carriage dog (1)', 'cut, mongrel, mull (2)', 'corgi, Welsh corgi (2)', 'Mexican hairless (0)', 'lapdog (0)', 'Newfoundland, Newfoundland dog (0)', and 'poodle, poodle dog (4)'. To the right of the tree view, there are three tabs: 'Treemap Visualization', 'Images of the Synset', and 'Downloads'. The 'Images of the Synset' tab is active, showing a grid of 20 small images of Chihuahuas in various poses and settings.

Photo from www.image-net.org/



Photo from aws.amazon.com/rekognition/

Real-World Use Cases for Image Recognition

Auto-Organizing your Image Library

Vacation Photos



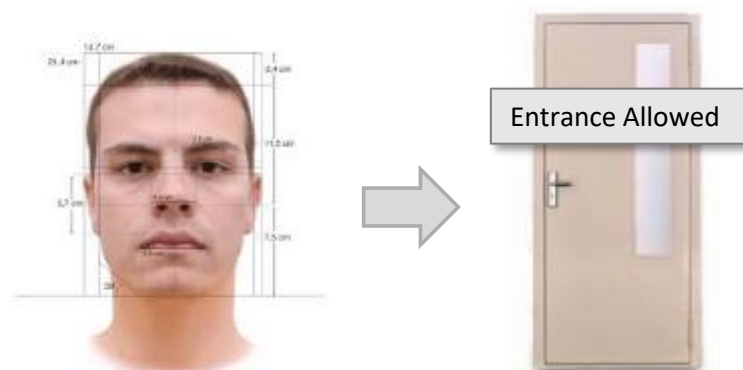
Machine & Factory Maintenance



=

Part Number
PHY18374EU

Facial Recognition for Office Security



Etc! New use cases
constantly emerging.

Artificial Intelligence & Data Quality

- Amazon.com's Recommendation Engine uses Artificial Intelligence
 - Based on analyzing data from shopping trends
 - Combined with product master data

Customers Who Bought This Item Also Bought Page 1 of 8

Product Image	Product Title	Rating	Price
	Tesla Women's Lightweight Active Performance Full-zip Hoodie Jacket WK24	★★★★★ 1	\$19.98
	Ultra Comfort KS Tech Lightweight Hoodie for Men & Women -TRY RISK	★★★★★ 35	\$22.99 Prime
	Tesla Men's Lightweight HyperDri Running Shorts With Pockets MTP07	★★★★☆ 127	\$9.98 - \$13.98
	Tesla Men's Thermal Coldgear Compression Long Sleeve T Shirts R34	★★★★☆ 162	\$9.98 - \$19.98
	Tesla Men's Cool Dry Compression Baselayer Pants Legging Shorts Tights P16	★★★★☆ 496	\$12.98 - \$19.98
	Hanes Adult Nano Full Zip Hood	★★★★☆ 123	\$14.19 - \$49.12
	Tesla Men's Cool Dry Compression Baselayer Mock Long Sleeve T Shirts T11	★★★★☆ 195	\$9.98 - \$14.98

Customer Purchasing Patterns

Product Master Data

Artificial Intelligence & Data Quality

AI is only as good as the underlying data

- Artificial Intelligence is based on evaluating data sets. If those data sets are faulty or of poor quality, your AI results will be flawed.
 - Especially if the data sets are small

Customers Who Bought This Item Also Bought



1 X Melitta - 3.5" Disc
Coffee Filter
★★★★★ 714
\$4.70 ✓ Prime

Don't Forget the Business Value

Just because you “can” doesn’t mean it’s effective.



Jac Rayner @GirlFromBlupo · Apr 6

Dear Amazon, I bought a **toilet seat** because I needed one. Necessity, not desire. I do not collect them. I am not a **toilet seat** addict. No matter how temptingly you email me, I'm not going to think, oh go on then, **just one more toilet seat**, I'll treat myself.

1.9K 73K 403K



Liz Rice @lizrice · Jun 1

Boeing took a look at my profile, thought “now there’s a woman in the market for a military sub”, and promoted me this tweet 🙄



Boeing Defense @BoeingDefense

Designed for long endurance operations and multi-mission capability, Echo Voyager adds depth to your intelligence. See how: bit.ly/2J2InLp

Governance & Metadata for Machine Learning/AI



Source: David Robinson, Data Scientist at Stack Overflow

- With Machine Learning (& Data Science), not **only the data needs to be governed with documented metadata, but the models and algorithms** themselves must be documented as well.
 - What data are we using and why?
 - What algorithms are being used and what is the logic?

Ethics

Think before you code

- Ethics are a key consideration in the usage of Artificial Intelligence, i.e.
 - Just because we *can*, does it mean we *should*?
- Some considerations
 - **Privacy** – consideration of consumers’ rights
 - **Errors** – how do we ensure a correct result (e.g. self-driving cars, decision algorithms)
 - **Job Loss** – will this replace human staff? Is that a concern?
 - **Bias** – do the training sets and algorithms promote inherent bias?
 - **Security** – can data sets or algorithms be hacked by nefarious sources?
 - **Control** – is there a risk of losing control over the algorithm and its results?
 - **The “Creep Factor”** – perhaps it’s not illegal or doesn’t break official privacy rules, but does it “feel right”? Would I want to be the consumer in this scenario?
 - Etc.



Computers Can “Learn” Bias

Consider this fact in selecting your training data sets

Doctor

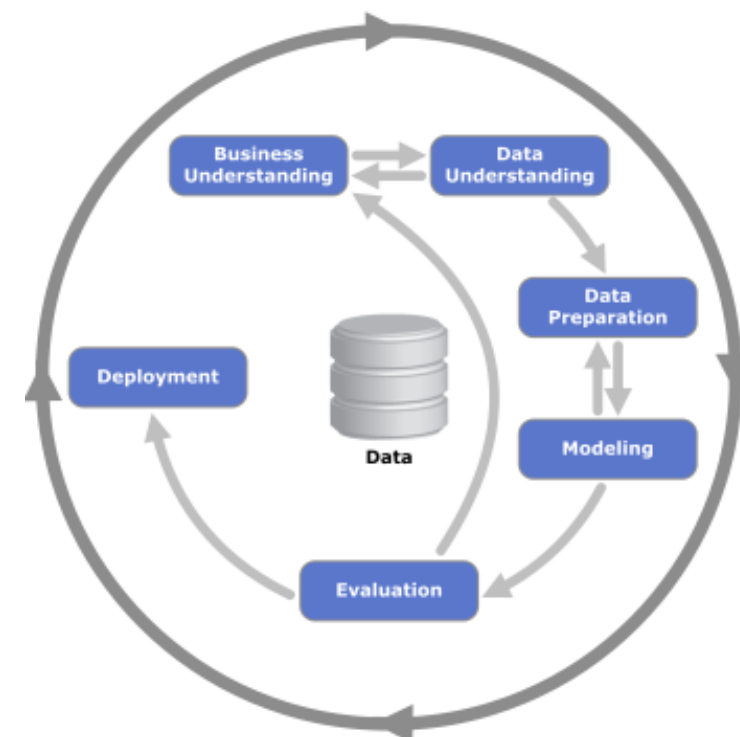
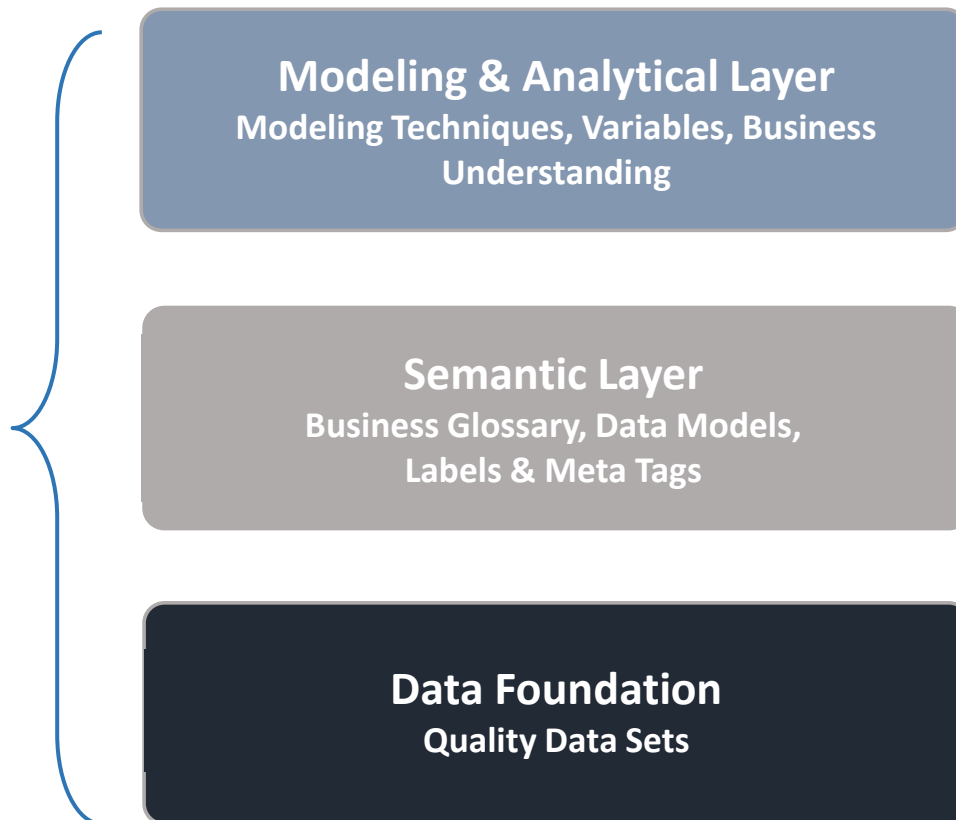


Doctor



Data Governance is Critical for AI

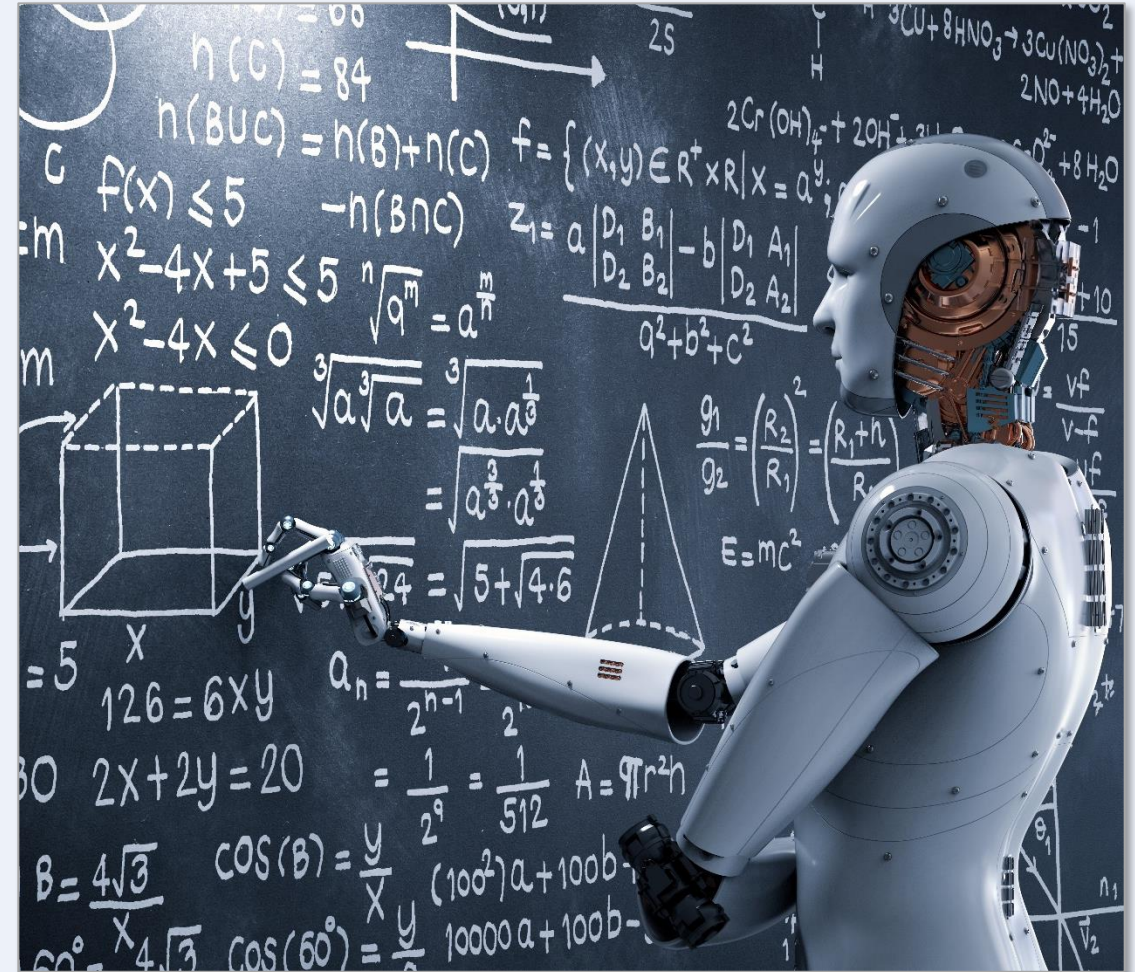
Governance is important at a number of layers in the AI ecosystem – from the data to the algorithms.



The Crisp Methodology is one methodology for governing analytical modelling.
Credit to Data Science Central

Summary

- AI/ML is growing in popularity as storage & compute capabilities increase and business opportunities grow.
- Trusted data sets for AI/ML depend on a mix of both traditional, governed data sets and more exploratory, higher-volume data sets.
- Data Governance is critical for AI at all layers: from storage, to meaning, to the analytic models themselves.



DATAVERSITY Data Architecture Strategies

Join Us Next Month

- **January** Emerging Trends in Data Architecture – What’s the Next Big Thing?
- **February** Building a Data Strategy - Practical Steps for Aligning with Business Goals
- **March** Data Mesh or Data Mess? Separating the Reality from the Hype
- **April** Master Data Management - Aligning Data, Process, and Governance
- **May** How do Data Governance & Data Architecture Support Each Other?
- **June** Why You Need Data Management – Getting Executive Buy-In
- **July** Artificial Intelligence and Machine Learning – Building the Right Architectural Foundation
- **August** Data Quality Best Practices (with Nigel Turner)
- **September** Best Practices in Metadata Management
- **October** Designing Data for Business Intelligence & Analytics – Where the Star Schema Fits in a Modern Data Architecture
- **December** Enterprise Architecture vs. Data Architecture



Who We Are: Business-Focused Data Strategy

Maximize the Organizational Value of Your Data Investment



In today's business environment, showing **rapid time to value** for any technical investment is critical.

But technology and data can be complex. At Global Data Strategy, **we help demystify technical complexity** to help you:

- Demonstrate the ROI and **business value of data** to your management
- Build a data strategy **at your pace to match your unique culture** and organizational style.
- Create an **actionable roadmap for “quick wins”**, which building towards a long-term scalable architecture.

Global Data Strategy's shares experience from some of the largest international organizations scaled to the pace of your unique team.

Global Data Strategy has worked with organizations globally in the following industries:

Finance · Retail · Social Services · Health Care · Education · Manufacturing
· Government · Public Utilities · Construction · Media & Entertainment ·
Insurance and more



Thoughts? Ideas?
Questions?