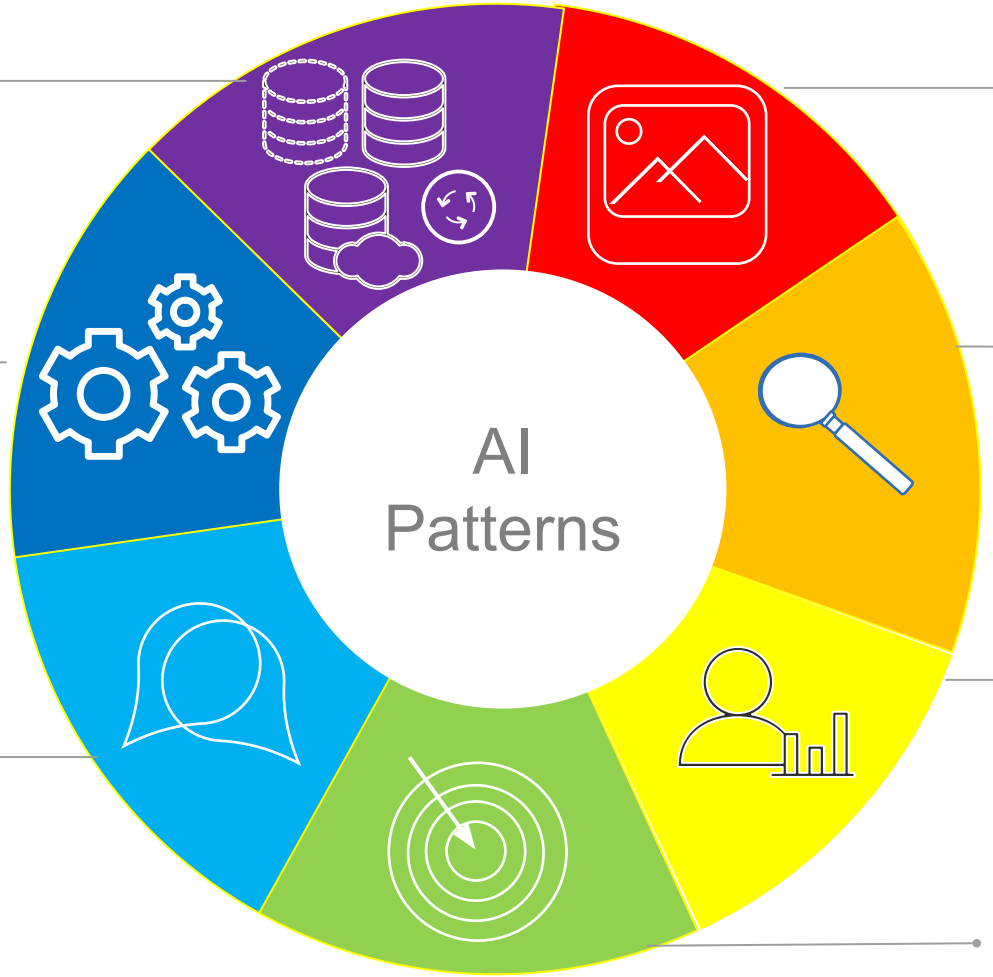
PERFORCE

AI/ML Data Megatrends and the Art of the Possible

Woody Evans

12n September 9, 2024

The 7 Major Patterns of Artificial Intelligence



Reasoning with Knowledge Structures

Recognition

Goal Driven Optimization

Event Detection

Interaction Support

Forecasting

(Hyper) Personalization

AI/ML Adoption Trends: Challenges

Data Availability & Quality

collection process, audits, hygiene, validation Techniques, not ensuring data security & privacy

Security/Privacy in Pipelines

Leakage, California AI Bill (SB 1047), EU Artificial Intelligence Act, AI Liability Directive, General Lack of AI Governance

Tech Adoption

No Vision, Lack of Skilled Talent, Infra & Integration Hard, Algo Complexity, Can't Scale, Hard to ID Biz Case

Ethical Concerns

Data Bias & Inaccuracy, Model Degradation, Hallucination, Interpretability, Privacy, Security, Safety, Reliability, Unexpected Outcomes

Cost to Implement

Cost to Acquire Data, Build AI Infra, Develop & Maintain & Catalog models, AI Ready Staff, Unclear ROIs/Hard to Project.

Culture

Resistance to Change, Organizational Inertia, Resistance to New Tech, The "We don't need AI" attitude.

AI/ML Adoption Trends: Emerging Data Challenges

Data Availability & Quality

collection process, audits, hygiene, validation Techniques, not ensuring data security & privacy

Security/Privacy in Pipelines

Leakage, California AI Bill (SB 1047), EU Artificial Intelligence Act, AI Liability Directive, General Lack of AI Governance

Tech Adoption

No Vision, Lack of Skilled Talent, Infra & Integration Hard, Algo Complexity, Can't Scale, Hard to ID Biz Case

Ethical Concerns

Data Bias & Inaccuracy, Model Degradation, Hallucination, Interpretability, Privacy, Security, Safety, Reliability, Unexpected Outcomes

Cost to Implement

Cost to Acquire Data, Build AI Infra, Develop & Maintain & Catalog models, AI Ready Staff, Unclear ROIs/Hard to Project.

Culture

Resistance to Change, Organizational Inertia, Resistance to New Tech, The "We don't need AI" attitude.

Data Wrangling: Finding/Avoiding the Achilles Heel

What is it?

Transforming, cleaning, standardizing, and enriching raw data into a format suitable for analysis with high quality and consistency.

We need to prepare data for AI modeling and analysis.

What's the Problem?

Collection is expensive.
Datasets get stale quickly, thus models do too.
Data cleansing has a goldilocks problem: Too aggressive and you can "overfit", too lax and you can "underfit".
Poor wrangling can amplify the bias of the original data.
Secure data via Synthetic data or masked data have pros & cons by use case and need.

Data collection is costly, freshness is crucial, cleansing is tricky, and bias lurks in wrangling.

Why is it hard?

- Continuous AI learning needs $$ data collection & cataloguing to repeat.
- AI amplifies bias and *isn't good* at finding its own bias.
- Data Quality issues are buried in complex algos, emerging late.
- Responding to AI's pace is a function of data velocity.
- Data utility and security are at odds no matter what you do: no free data lunch.

Humans often key to success: specific bias detection, issue analysis, & setting privacy budgets.

What can we do today?

- Automated Data Preparation
- DQ Assessment and Monitoring Tools
- Bias Detection and Mitigation Tools
- Data Management Platforms
- Highly Scalable Data Storage Platforms

Automate much of data prep, quality, bias detection, and dataset management and storage.

Data can be the Achilles' heel of AI. *When data is relevant, high-quality, complete, and unbiased, it produces high value results. But when it is noisy, limited, incomplete, or unfair, it produces bad decisions. Humans are still the arbiter for many of those tradeoffs.*

The Narrowing Scale Gap

What is it?

The deceasing performance, usability and/or applicability difference between LLMs and niche models.

We need AI to be more affordable and more in reach for the average company.

What's the Problem?

Large LLMs require immense computational resources.
Sharing data with a public LLM is dangerous but bias control & data security are on you if you roll your own.
Model innovation incredibly fast
Competitive frenzy: long time to market is the death knell.
LLM integration Talent scarce; Tuning talent even more so.

Consider how and why you adopt carefully; it's easy to get out of your depth, and the competition is brutal.

Why is it hard?

- Scale limits resources for most companies.
- In addition to leaking data, Public LLMs may learn from your data & compete!
- Bias control requires careful curation.
- Data Security is deceptively hard to get right with either Masked or Synthetic data.
- AI amplifies SW velocity.
- AI Talent is hard to grow and harder to find.

It's easy to be wrong about your bias, problem difficulty, how smart you are, and how fast to go.

What can we do today?

- Leverage pre-trained models, and fine-tuning.
- Use Private RAG with public LLMs.
- Avoid highly vulnerable data privacy solutions & use pro solutions like Data Masking.
- Adopt MLOps/LLMOps to improve AI SW velocity.
- Beyond the usual great workplace attractors, have a strong AI commitment and a roadmap to apply it.

We can find ways to limit cost, & data exposure and embrace methods and strategies leading to success

Google's famous memo: "We have no moat" is coming true. Small models are exploding and creating an LLM niche economy at incredible speed. But, just like DevOps translated to velocity, MLOps & now LLMOps are the keys to AI velocity at scale.

Explainable AI (XAI)

What is it?

AI that can explain its own decision-making process.

What's the Problem?

It's hard to trust what you can't inspect, and hard to alter what you don't understand, but both decision makers and regulators want this information. Trust requires transparency, and transparency is crucial for accountability.
These can cause decision hesitancy, resulting in missed opportunities and less efficiency and even reputational risk.

Why is it hard?

- Complex Models
- May violate privacy
- Conflicts with IP Protection
- Accuracy Tradeoffs
- It's Not Free
- No Framework; Only Principles
- Domain Specific
- Accountability Diffuse

What can we do today?

- Data Lineage
- Data Governance
- DevOps Like Model and Source Data Versioning

We need to know why AI makes Decisions, Recommendations, and Predictions

Without transparency we can't trust our decisions and regulators don't trust us.

But understanding that is years away, if it is even possible.

For now, we can control and catalogue datasets to provide rollback and accountability.

AI is non-linear – a change in an input doesn't always correspond to a change in output. Explanatory Techniques are being developed, but the processing is still largely "Black Box" Versioning models and source data is a powerful way to limit harm.

Responsible AI (RAI)

What is it?

A framework ensuring that AI systems are imagined, developed and deployed in a way that is ethical, transparent, fair, secure and accountable.

We need to AI to be ethical, transparent, and secure.

What's the Problem?

As organizations increasingly rely on AI decisions that impact our lives, they will have increased accountability for proper use and any bias, discrimination, or unfairness that may result from those decisions.

We will increasingly be held accountable for AI bias and decision impacts.

Why is it hard?

- Lack of Explainability
- Data itself can be biased
- No Standards
- Implementation requires highly specialized expertise

The landscape is fluid and non-standard, bias is often hard to identify/control, and finding experts is hard.

What can we do today?

- Feature Importance: Telling AI what's important in its decision making process.
- Data catalogues, lineage and history give maximum accountability.
- Rule Extraction: deriving human friendly rules from a model to explain a decision.
- Surrogate Models: Building simpler, easier to interpret models to approximate the more complex and study its behavior.

For now, we can use the best tools available, and keep tight tabs on our data.

Data cataloguing and quality is essential for responsible AI as it exists today; as models and massive datasets multiply, dataset lineage and version is tightly correlated to identifying, reducing, and controlling bias.

Data Dollar Diplomacy

What is it?

Strategic data management for optimal AI ROI, considering value, acquisition costs, storage, processing, and quality.

We need to understand the cost and value of AI better.

What's the Problem?

Debt: High cost, low value data can clog the AI pipeline.
Drift: the gap between training and real data, and the time between iterations, degrade models.
Dark: Unstructured data AI is not able to understand.
Dustbin: Poor Storage techniques and practices that kill ROI.
Dilemma: We're often unsure if our AI investment is paying off.

Time + Data + AI can breed inefficiency that's hard to pinpoint.

Why is it hard?

- Poor Dataset governance / training on the wrong data is often discovered when your model underperforms.
- Time kills data value.
- Natural Language Processing isn't a panacea for all context problems. There are still things too hard to solve.
- We have an "AI Tax Dilemma".

The signal often comes to late to avoid expenditure, and & AI ROI uncertainty exacerbates this.

What can we do today?

- Tools providing Dataset catalogues, lineage and history can govern dataset sprawl and avoid error.
- Leveraging NLP Advances to classify unstructured data.
- Invest in expert ROI tools to help us understand the value of AI to make better decisions about it.
- When considering AI Governance, Start with Data.

Tools can mitigate the issues, but data has to be a central topic for governance to get ROI under control.

The "AI Tax" dilemma is real – "we know we have to have Gen AI, and we know we have to pay for it, but we don't understand the value and so we will relegate our projects to the backwater."

The Looming Liability Landscape

What is it?

The growing tension between the need for efficient software pipelines and the imperative to protect us from catastrophic impacts caused by poor AI decisions.

We need to understand the risks AI poses and have a response that does not inhibit velocity.

What's the Problem?

As AI models provide decisions with larger reach, the risk and scale of societal impact goes up . Regulators want to curb behavior to limit negative & catastrophic social impact.

Data Quality, bias, governance, security, and privacy will require stricter governance since model makers will be held accountable.

Avoiding AI catastrophe and Staying out of regulatory sights requires significant attention to data.

Why is it hard?

- Risks are evolving faster than regulation and is being priced into the system.
- Globalization leaves worst actors out of the reach of regulation.
- Sacrificing Innovation for safety can be a competitive disadvantage vs. Less regulated environments.
- Unintended and unknown consequences.

Speed is against us, risks evolve quickly & and we must out innovate less restricted competitors

What can we do today?

- Explainable/Responsible AI
- Plan for Data sovereignty for both storage and processing including universal data privacy solutions that can be air gapped.
- Develop rapid data revocation protocols and rapid data reprovision for zero-day response.

Adding dataset agility and data sovereignty to explainable AI can mitigate significant risk for now.

Consumer behavior and retention isn't changing because of data breaches. It's priced into the system. **One exception**: when companies receive the attention of regulators. Investment is shifting to responding to regulators and away from mitigating churn,

AI/ML Data: 2024 insights from Perforce Experts

- Now Available – The 2024 State of Data Compliance and Security Report
- September 18 – AI/ML Data: 2024 Insights from Perforce Experts
- October 2 – 2024 Masking Insights: Revealed and Analyzed by the Delphix Experts


Thank You
PERFORCE