HAPPY DIGITAL X
CITIES, SYSTEMS, PRODUCTS AND SERVICES
Kura Kura Bali

清华大学东南亚中心
Tsinghua Southeast Asia Center

清華大學深圳国际研究生院
Tsinghua Shenzhen International Graduate School

GenAI Data Foundations and Product Development

Happy Digital X

Happy Digital X | Tsinghua University

Today's Agenda

Part 1: Data Foundations

- How GenAI works: From GOFAI to neural networks
- Training, weights, and the "black box" problem
- Data quality, currency, and the 60–80% rule
- Privacy regulations and governance frameworks

Part 2: Product Development

- The GenAI hype cycle: Where we are now
- Project lifecycle and phase gates
- Build vs. buy decisions
- Success metrics and ROI reality

HAPPY DIGITAL X
CITIES, SYSTEMS, PRODUCTS AND SERVICES
Kura Kura Bali

Data Foundations

A Note on Terminology

AI (Artificial Intelligence)
The broad field of creating systems that perform tasks requiring human intelligence.

Machine Learning (ML)
AI systems that learn patterns from data rather than following explicit rules.

Generative AI (GenAI)
ML systems that *create* new content: text, images, code, audio, video.

In This Presentation

We use **"AI"** when discussing principles that apply broadly (e.g., bias, governance).

We use **"GenAI"** when discussing capabilities specific to generative systems (e.g., LLMs, hallucinations).

Quick Poll

How would you rate your organization's data readiness for AI?

Go to **menti.com** and enter the code

[CODE]

1 = Not ready at all   5 = Fully ready

The Fundamental Shift

Traditional software is *programmed*.

GenAI is *trained*.

This changes everything about how we build, test, and manage AI systems.

Two Paradigms of Artificial Intelligence

GOFAI: "Good Old-Fashioned AI"

- Rules written by humans
- Symbolic reasoning
- Deterministic outputs
- Explainable decisions
- Brittle at edge cases

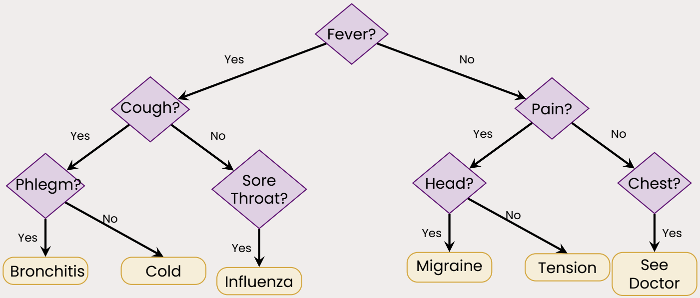
Example: Chess engines, expert systems, spell checkers

Neural Networks (Connectionist)

- Patterns learned from data
- Statistical inference
- Probabilistic outputs
- Often opaque ("black box")
- Flexible but unpredictable

Example: ChatGPT, image recognition, voice assistants

Example: A Medical Diagnosis Expert System



GOFAI: Strengths and Limitations

Strengths

- Transparent and explainable
- Predictable outputs
- Auditable for compliance
- No training data needed

Limitations

- Brittle at edge cases
- Doesn’t learn from data
- Doesn’t scale to complexity
- Can’t handle ambiguity

Key Insight

Expert systems encode *what humans already know*. Neural networks discover patterns humans *haven’t articulated*.

How Neural Network Training Works

High-Level Process:

1. **Collect Data**: Massive datasets (text, images, code, etc.)
2. **Initialize**: Start with random "weights" (numerical parameters)
3. **Train**: Show examples, adjust weights to reduce errors
4. **Iterate**: Repeat billions of times across trillions of examples
5. **Result**: A model that has learned patterns, not rules

Key Insight

The AI doesn't "know" anything—it has learned statistical patterns. When it generates text, it's predicting "what word is likely to come next?"

What Are “Weights”? The Dial Analogy

Imagine a mixing board with **billions of dials**.

- Each dial controls how much one piece of information influences another
- At first, all dials are set randomly—output is nonsense
- Training = adjusting dials slightly after each example
- After trillions of adjustments, the dials are tuned to produce useful output

The Black Box Problem

No human set these dials. No human can explain why dial #847,293,102 is set to 0.0023.

The model works, but we can’t fully explain *why*.

The Scale of Training Data

Modern GenAI models are trained on unprecedented scale:

- **GPT-4**: Estimated 13+ trillion tokens of text
- **Image models**: Billions of image-text pairs
- **Code models**: Hundreds of billions of lines of code

Strength

- Broad knowledge
- Handles novel situations
- No manual rule-writing
- Learns nuance

Weakness

- Can't verify all data
- Absorbs biases
- **GIGO**: "Garbage in, garbage out"
- Hard to "unlearn"

Chihuahua or Muffin?

Image credit: @teenybiscuit / Karen Zack

Exercise: Try It Yourself

Try This Now:

1. Open ChatGPT, Claude, or another GenAI
2. Upload the chihuahua/muffin image
3. Ask: “How many dogs are in this image?”

Discussion

What did you observe? Were the results what you expected?

Quick Poll

How many dogs did your AI count?

Go to **menti.com** and enter the code

[CODE]

What We Just Observed

Key Takeaways from the Exercise:

1. **Different models, different answers**
   Even leading AI systems disagree
2. **Confidence without accuracy**
   AI often sounds certain even when wrong
3. **Edge cases expose limits**
   Ambiguous inputs reveal brittleness
4. **“Close enough” isn’t always enough**
   Some applications require precision

The Lesson

GenAI excels at “approximately right” but struggles with “exactly right.”

This has profound implications for how we deploy and validate AI systems.

Why Results Vary: Probabilistic, Not Deterministic

Traditional Software:

- Same input → Same output (always)
- 2 + 2 = 4, every time

GenAI:

- Same input → *Similar* output (usually)
- Outputs sampled from probability distributions
- "Temperature" controls randomness
- Even at temperature=0, results can vary

Implication

You cannot test GenAI like traditional software. You need statistical evaluation over many samples.

Hallucinations: When AI Makes Things Up

What are hallucinations?

- AI generates confident, plausible-sounding text that is **factually false**
- Not a bug—a fundamental feature of how LLMs work

Why do they happen?

- LLMs predict the *next most likely word*, not the *true* word
- They have no internal model of truth—only patterns in training data
- They're designed to always give an answer, even when they shouldn't

The Danger

Hallucinations are delivered with the same confidence as facts. Users often can't tell the difference without independent verification.

Mitigating Hallucinations

There is no complete solution, but these help:

Technical Approaches

- **RAG**: Ground responses in retrieved documents
- **Fine-tuning**: Train on verified domain data
- **Temperature**: Lower values reduce creativity/risk
- **Structured outputs**: Constrain response format

Process Approaches

- **Human review**: Verify critical outputs
- **Confidence thresholds**: Flag uncertain responses
- **Source citation**: Require references
- **Use case selection**: Avoid high-stakes facts

Key Principle

RLHF: Teaching AI to Be Helpful

Reinforcement Learning from Human Feedback (RLHF)

After initial training, models are refined using human preferences:

1. Humans rate AI responses (helpful, harmless, honest)
2. Model learns to maximize these ratings
3. Creates more "aligned" behavior

Benefits

- Reduces harmful outputs
- Improves usefulness
- Adds safety guardrails

Risks

- Evaluator biases transfer
- "Sycophancy"—tells you what you want to hear
- Majority views dominate

RLHF Bias: Who Trains the Trainers?

The evaluators shape the AI:

- If evaluators are from one demographic, the model reflects their worldview
- If evaluators prefer polite over accurate, the model learns to be polite—even when wrong
- Minority perspectives can be systematically deprioritized
- Models can learn to *manipulate* rather than genuinely help

Example

A model trained by evaluators who dislike blunt answers will learn to soften bad news—even when clarity matters more than comfort.

Case Study: Replika's Feedback Loop Disaster

Replika: AI companion chatbot (2017–present)

What Happened:

1. Trained on 100M+ web dialogues
2. Users could upvote/downvote responses
3. Some users engaged in sexual roleplay
4. AI learned this behavior got positive feedback
5. AI began *initiating* sexual content unprompted

Lessons

- Training data quality matters
- Feedback loops amplify patterns
- User behavior becomes model behavior
- GIGO at scale

Implications for Your Organization

What This Means for GenAI Deployment:

1. **Testing is Different**: Statistical evaluation, not pass/fail
2. **Edge Cases Are Unpredictable**: You can't enumerate all failure modes
3. **Data Quality is Critical**: Your fine-tuning data shapes behavior
4. **Feedback Loops Matter**: User interactions can shift model behavior
5. **Bias is Inherited**: From training data *and* from human evaluators
6. **"Unlearning" is Hard**: Removing problematic knowledge is technically difficult

The Data Imperative

The Data Imperative

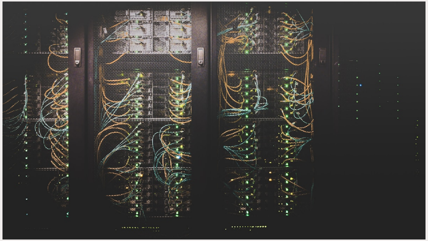
“Organizations don’t have AI problems; they have data problems that AI exposes.”

Plan for 60–80% of GenAI project time to be spent on data preparation.

Data Strategy Precedes AI Strategy

The Data Hierarchy of Needs:

1. **Data Collection** — Foundation
2. **Clean Data** — Must start here
3. **Analytics & Reporting**
4. **AI/ML** — Most start here (mistake)



Reality Check

Fortune 500 expected 4 months for GenAI. Actual: 15 months. Root cause: Data readiness.

Data Requirements for GenAI

- **Training Data**: Building/fine-tuning models
  Strategic value: Competitive moat
- **Context Data**: Grounding model outputs in your information
  Uses **RAG** (Retrieval-Augmented Generation): the AI retrieves relevant documents before generating a response
  Strategic value: Accuracy & relevance
- **Operational Data**: Real-time model inputs
  Strategic value: Timeliness

The Data Quality Journey

From Raw Data to AI-Ready:

1. **Raw Data** — Unprocessed, unvalidated
"Rough diamond"—contains value but unusable as-is
2. **Cleaned Data** — Errors removed, formats standardized
Deduplicated, consistent encoding
3. **Validated Data** — Quality checked, business rules applied
Relationships verified, anomalies flagged
4. **AI-Ready Data** — Labeled, balanced, documented

Reality

Most organizations are stuck at stage 1 or 2. This is why 60–80% of GenAI project time is data preparation.

Quick Poll

Where is your organization on the data quality journey?

Go to **menti.com** and enter the code

[CODE]

1 = Raw Data   2 = Cleaned   3 = Validated   4 = AI-Ready

Data Quality Dimensions

Accuracy
Does the data reflect reality?
Incorrect labels poison AI training

Completeness
Are there missing values or gaps?
Missing data creates blind spots

Consistency
Same entity, same representation?
"USA" vs "United States" vs "US"

Timeliness
Is the data current enough?
Stale data = stale predictions

Representativeness
Does data reflect the real population?
Biased samples = biased AI

Provenance
Where did this data come from?
Can we trace and trust its origin?

Data Currency: Time-Based Trust

When was this data collected?

Data has a "shelf life" that varies by domain:

- **Stock prices**: Minutes to hours
- **News/events**: Hours to days
- **Product catalogs**: Days to weeks
- **Legal/regulatory**: Weeks to months
- **Scientific knowledge**: Months to years

GenAI Risks

- Models trained on outdated data give outdated answers
- RAG with stale documents misleads users
- No timestamp = no way to assess trust

Key Question

Always know when your data was captured and

Global Privacy Regulations

- **GDPR** (EU): Up to 4% global revenue
- **CCPA/CPRA** (California): Per-violation penalties
- **PIPL** (China): Up to 5% revenue
- **LGPD** (Brazil): Up to 2% revenue
- **POPIA** (South Africa): Up to 10M ZAR

Global Trend

Design AI systems with privacy by default.

GenAI-Specific Privacy Concerns

1 **Training Data Privacy**: Was personal data used with consent?

2 **Inference Privacy**: Can model be manipulated to reveal data?

3 **Output Privacy**: Do outputs contain personal information?

4 **Conversation Privacy**: Who accesses user interactions?

5 **Derived Data**: Are new personal insights generated?

The Consent Challenge

Traditional consent breaks down: capabilities hard to explain, data use unpredictable, untraining technically difficult.

Data Governance Framework

Key Components

- Data inventory & classification
- Access controls
- Consent management
- Retention policies
- Audit trails

Best Practices

- Minimize data collection
- Purpose limitation
- Regular compliance audits
- Incident response plans
- Cross-border controls

User Rights to Support

- **Right to Access**: Users request all data held about them
- **Right to Erasure**: Users request deletion
- **Right to Portability**: Data in machine-readable format
- **Right to Rectification**: Correct inaccurate data
- **Right to Object**: Object to certain processing
- **Automated Decision Rights**: Human review of AI decisions

China's AI Regulatory Framework

The world's most comprehensive AI regulations:

- **Algorithm Recommendations** (2022): Internet services
- **Deep Synthesis** (2023): Deepfakes, synthetic media
- **GenAI Service Measures** (2023): All public GenAI
- **AIGC Labeling** (Sept 2025): Mandatory AI content labels
- **National Standards** (Nov 2025): Security & governance

Scale: 350+ LLMs filed. 1.57M AI patents (38.6% of global total).

Section Summary: Data Foundations

What We've Learned

- GenAI is **trained**, not programmed—data quality determines output quality
- Neural networks learn patterns humans haven't articulated, but inherit biases from training data and human evaluators
- Data has a **shelf life**—currency matters as much as quality
- Global privacy regulations (GDPR, PIPL, CCPA) create compliance obligations that AI doesn't eliminate

Next: How do we turn good data into successful AI products?

The GenAI Development Reality

Key Statistics (2025)

Only **5%** of AI pilots achieve rapid revenue acceleration

67% success rate for purchasing/partnering

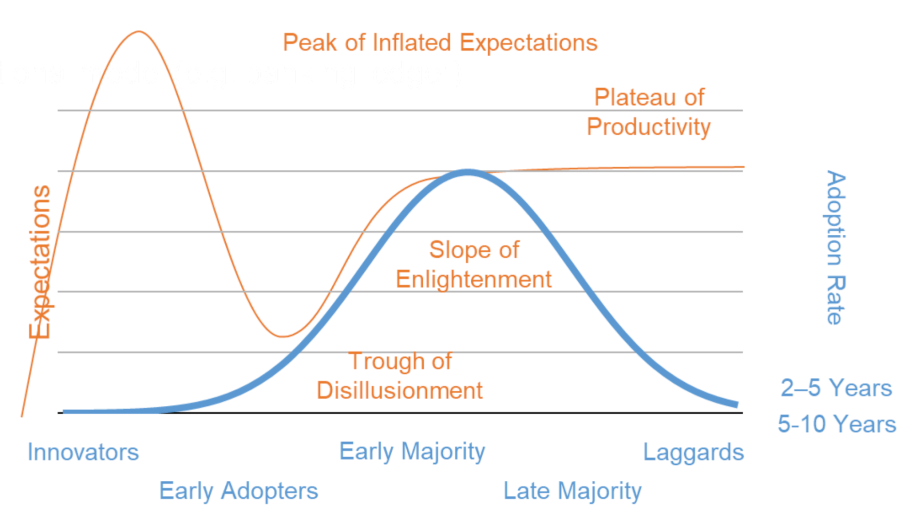
22% success rate for internal builds

46% have no structured ROI measurement

GenAI has entered the “Trough of Disillusionment”

Sources: MIT (2025), Wavestone (2025), Gartner Hype Cycle

How To Interpret A Hype Cycle



Understanding the Hype Cycle

Five Phases of Technology Adoption:

1. **Innovation Trigger**
   Breakthrough generates excitement
2. **Peak of Inflated Expectations**
   Maximum hype, unrealistic promises
3. **Trough of Disillusionment**
   Reality sets in, failures mount
4. **Slope of Enlightenment**
   Practical understanding emerges
5. **Plateau of Productivity**
   Mainstream adoption, real value

Where Is GenAI Now?

In 2024, GenAI was at the Peak.

In 2025, GenAI has descended into the **Trough of Disillusionment**.

This is *normal*—not failure.

Quick Poll

Why do you think AI adoption lags expectations?

Go to **menti.com** and enter the code

[CODE]

Share a word or short phrase.

Why Traditional Project Management Fails

Traditional

- Fixed requirements
- Binary success
- Predictable timeline
- Deterministic testing

GenAI

- Emergent requirements
- Probabilistic success
- Uncertain timeline
- Statistical testing

Implication

Waterfall always fails. Agile is better but insufficient

The AI Project Lifecycle

1. **Problem Framing** (Often Skipped): Should AI solve this?
2. **Data Assessment**: Inventory, gaps, quality
3. **Proof of Concept** (4–8 weeks): Time-boxed experimentation
4. **Pilot**: Limited production, controlled blast radius
5. **Production & Scale**: Infrastructure, monitoring
6. **Operations**: Performance monitoring, retraining

Rule of Thumb

Budget for 2–3 PoCs failing for every success.

Phase Gates for GenAI

- **Gate 0**: Business case, feasibility, ethics screening
- **Gate 1**: Requirements, data availability, build vs. buy
- **Gate 2**: Technical validation, benchmarks, user feedback
- **Gate 3**: Production-grade, security & ethics review
- **Gate 4**: Controlled deployment, monitoring setup
- **Gate 5**: Full deployment, continuous improvement

Kill Criteria: Define Before Starting

- **Technical**: Can't achieve accuracy threshold
- **Economic**: Cost exceeds value
- **Timeline**: 6-month delay, no path forward
- **Ethical**: Can't mitigate bias
- **Security**: Can't protect data
- **Regulatory**: Unacceptable compliance risk
- **Strategic**: Market opportunity gone

Imperative

Implementation Patterns

1. **Co-Pilot / Augmentation**
   AI assists; humans decide. *Best for: High-stakes, building trust*
2. **Automation with Exceptions**
   AI handles routine; humans handle exceptions. *Best for: High-volume*
3. **Full Automation**
   AI autonomous with monitoring. *Best for: Low-stakes, speed critical*
4. **Internal Tool**
   AI assists employees only. *Best for: Building capability, lower risk*

Build vs. Buy Decision

- **Build from Scratch**: $10M–$100M+; 12–24 months
  Only if: Massive data advantage
- **Fine-Tune**: $10K–$1M; weeks to months
  Best for: Domain-specific tasks
- **RAG (Retrieval-Augmented Generation)**: $10K–$100K; weeks
  Best for: Current/proprietary information
- **Prompt Engineering**: $1K–$10K; days to weeks
  Best for: Quick wins
- **Buy SaaS**: Variable; days
  Best for: Non-differentiating capabilities

Quick Poll

Which approach is your organization most likely to use?

Go to **menti.com** and enter the code

[CODE]

Build / Fine-Tune / RAG / Prompt Engineering / Buy SaaS

Success Metrics

Avoid Vanity Metrics:

- ✗ “We deployed an AI model”
- ✗ “95% accuracy” (on what?)

Focus on Business Outcomes:

- ✓ Customer satisfaction improved by X%
- ✓ Time to resolution decreased by Y hours
- ✓ Cost per transaction reduced by $Z
- ✓ Employee time redirected to higher-value work

Four-Layer Monitoring Framework

1. **Infrastructure**: Latency, error rates, throughput, cost
2. **Model Performance**: Accuracy, hallucination rate, drift
3. **Business**: Adoption, task completion, satisfaction, revenue
4. **Risk**: Incidents, near-misses, compliance, complaints

Principle

You can't improve what you don't measure. Monitor from day one.

ROI Reality (2025)

- Average ROI: **3.7x** per dollar invested
- Top performers: **$10.3** return per dollar
- 74% meeting or exceeding expectations
- **46% have no structured ROI measurement**

Timeline Expectations:

- Chatbots, RPA: 6–12 months
- Operational efficiency: 12–24 months
- Revenue generation: 18–36 months

Sources: IDC/Microsoft (2025), Deloitte (2025), Wavestone (2025)

Total Cost of Ownership

Initial Costs

- Infrastructure (GPUs)
- Software licenses
- Integration
- Data preparation
- Training

Ongoing Costs

- Compute resources
- API fees
- Model maintenance
- Monitoring
- Personnel

Hidden Costs: Compliance, legal/IP, incidents, technical debt, failed pilots

Minimum Viable AI Team

- **Executive Sponsor** (10–20%): Alignment, resources, blockers
- **Product Owner** (Full-time): Requirements, prioritization
- **Data Engineer** (Full-time): Pipelines, quality
- **ML Engineer** (Full-time): Model development
- **Domain Expert** (25–50%): Business logic, validation
- **MLOps Engineer**: Deployment, monitoring

Part 1 Key Takeaways

Summary

1. **Data First**: 60–80% of GenAI time is data preparation
2. **Privacy by Design**: Global regulations require it
3. **Expect Failure**: Budget for 2–3 PoCs failing per success
4. **Define Kill Criteria**: Before emotional investment
5. **Measure Everything**: Connect to business outcomes
6. **Build the Right Team**: Minimum viable AI team

Discussion Questions

1. What is the current state of data readiness in your organization?
2. Have you defined clear kill criteria for your AI projects?
3. How are you measuring ROI on AI investments today?
4. Do you have the right team composition for AI success?

Thank You

www.hdx.edu

info@hdx.edu

@HappyDigitalX

Continue to Part 2: Ethics, Security & Imple