BIG DATA INSTITUTE
สถาบันข้อมูลขนาดใหญ่

Data Strategy for AI (& Cloud)

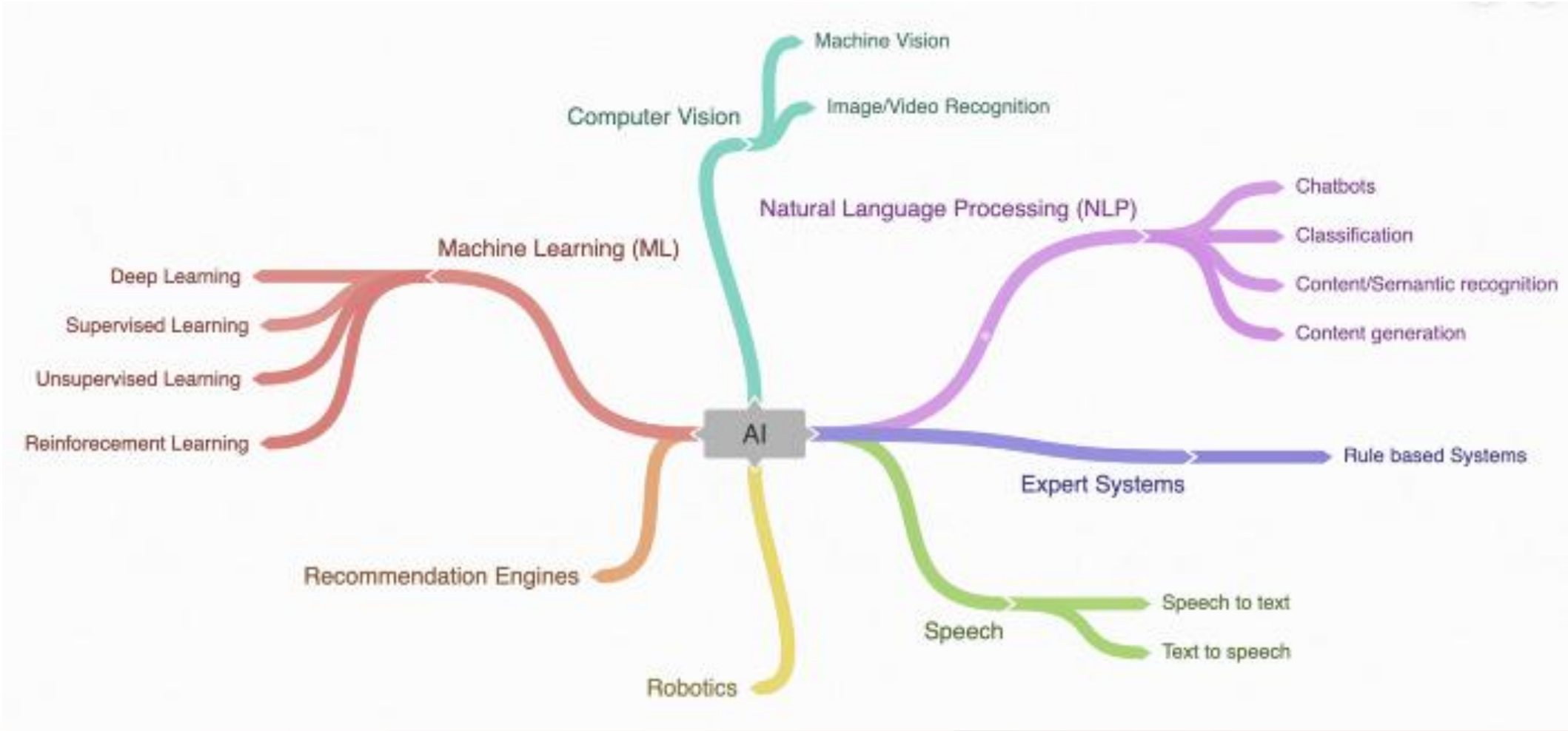
Patipan Prasertsom

Slide developed in parts by Dr. Apivadee Piyatumrong, Assoc. Prof. Tiranee Achalakul, and Chayasin Saetia with a lot of materials from AIGC-ETDA

ARTIFICIAL INTELLIGENCE (A.I.)

AI is already in our life

AI is a broad field



Figures from Bharath Kumar Hemachandran (https://medium.com/@bharathkumar.h/artificial-intelligence-101-tech-terms-explained-for-non-tech-folks-3190d7e711cd)

BIG DATA INSTITUTE

GENERATIVE AI

Generate new original content: text, images, videos and audio

- Generative AI learns to can create original utilizing patterns learned from training data
  - Content can be essays, solutions to problems, programming code, music score, realistic fakes, painting, and etc.
- Large Language Model (LLM) is a type of generative AI that excels in natural language tasks such as summarization, Q&A,etc.

GPT-3.5

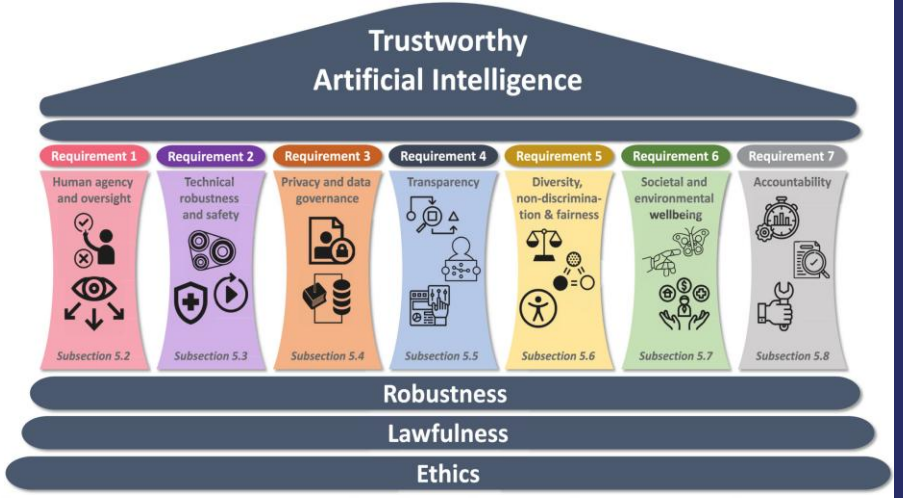
perplexity

ChatGPT

Gemini

Anthropic Claude AI

Copilot



Personalized AI
AI Automation
Agentic AI

2022

ChatGPT Launch

- Virality, "AI Boom
- Mass adoption of conversational AI
- AI in education, coding, content

2023

Competition & Integration

- Major AI releases from competitors
- AI integration wave
- Point concerns and regulation

2024–2025

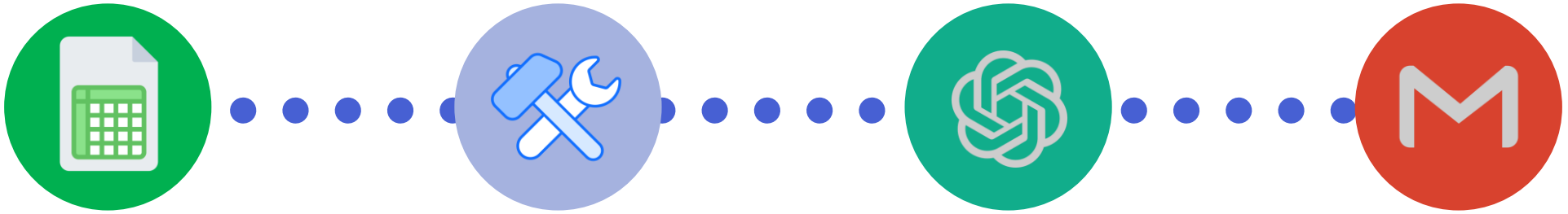
Multimodal & Personal AI

- Voice, image, video capabilities
- AI assistants and agents
- Reasoning and long-term memory

GPT-5
GPT-OSS-120b/20b (OpenW)
Claude Opus 4 / Sonnet 4
Gemini Deep Think
Llama 4 (not OpenW)
Gemma 3 (OpenW)
Magistral Small (reasoning)
QWEN 3

BIG DATA INSTITUTE

Examples of agentic workflows

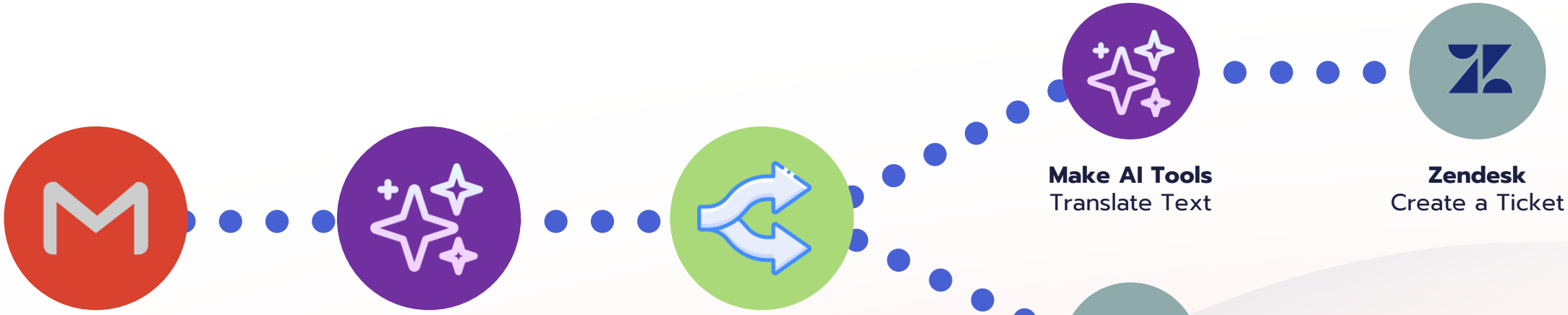


Get Customer Feedback
Search Rows

Aggregate All Feedback
Test aggregator

Analyze with ChatGPT
Create a completion (prompt) (GPT and o-series Models)

Send Insights via Email
Send an Email



New Support Email Received
Watch Emails

Make AI Tools
Identify Language

Router

Make AI Tools
Translate Text

Zendesk
Create a Ticket

Zendesk
Create a Ticket

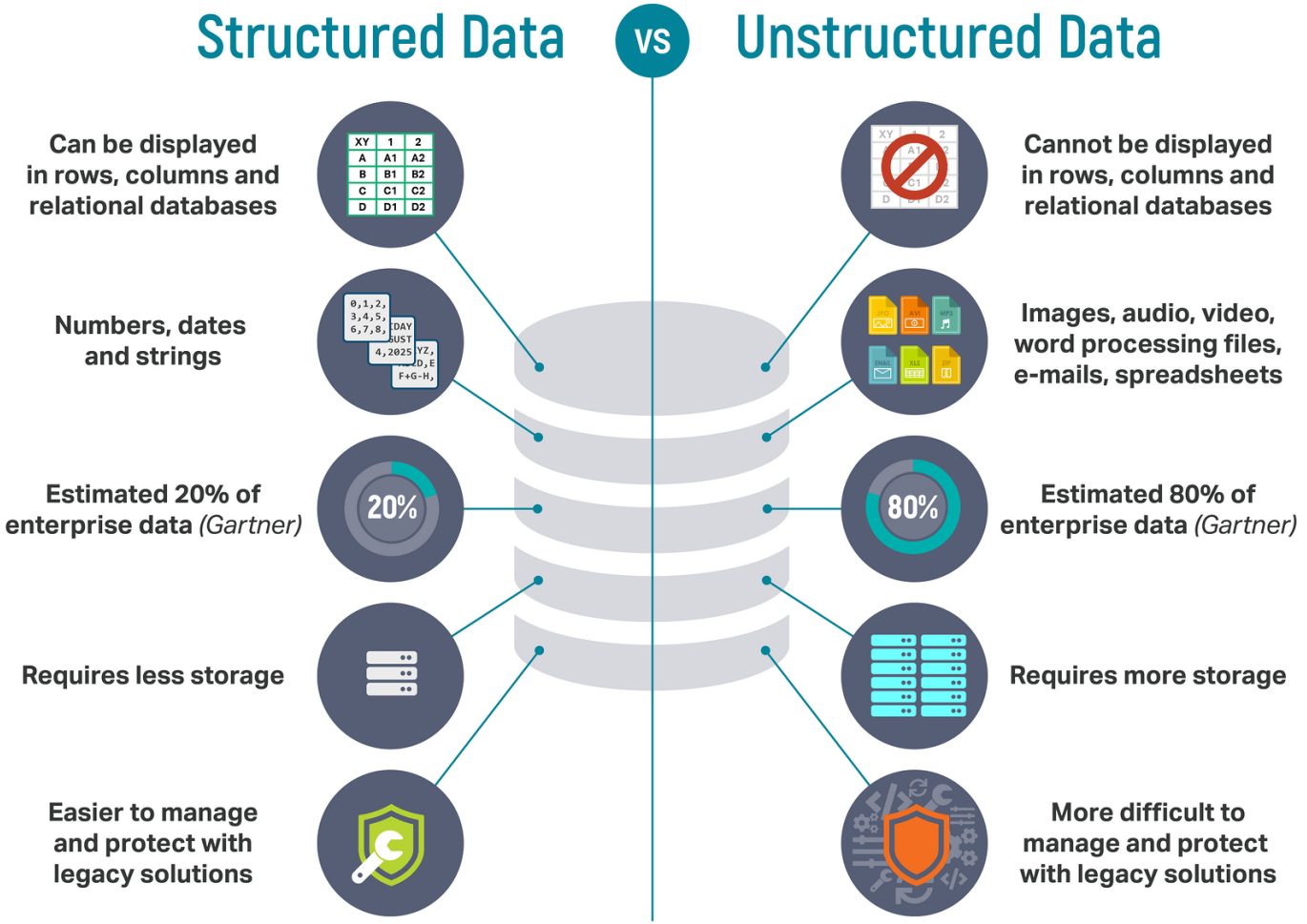
BIG DATA INSTITUTE

Data

(More on this later)

Mandatory Slide: Broad categories of collectible data

Suitable tasks:
- Classical ML tasks
  - Prediction
  - Classification
  - Clustering
  - Etc.
- Data retrieval
- Analytics tasks

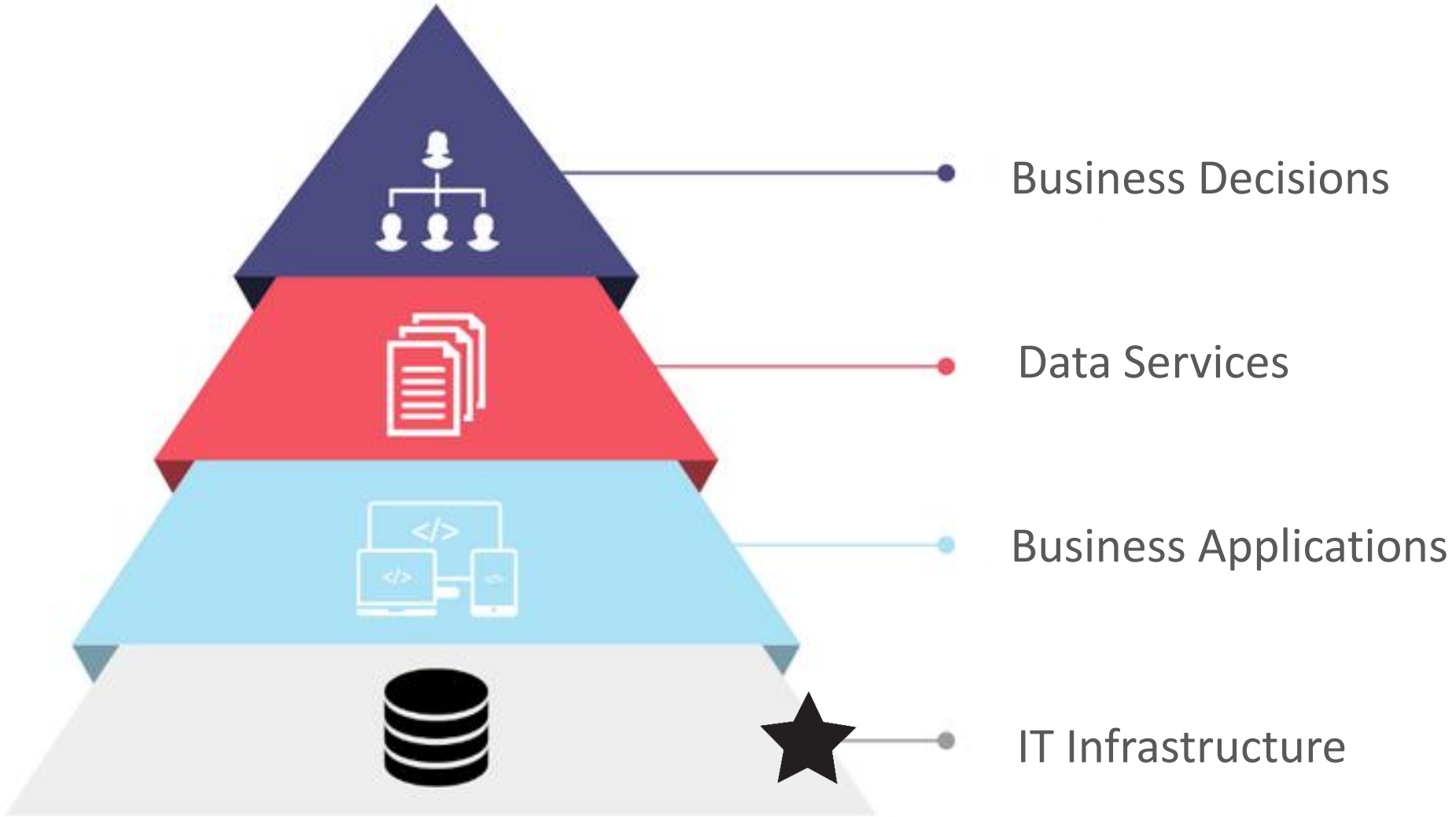


Suitable tasks:
- Natural Language Processing (NLP)
  - Text Analysis
  - Etc.
- Computer vision
- Generative AI and its applications

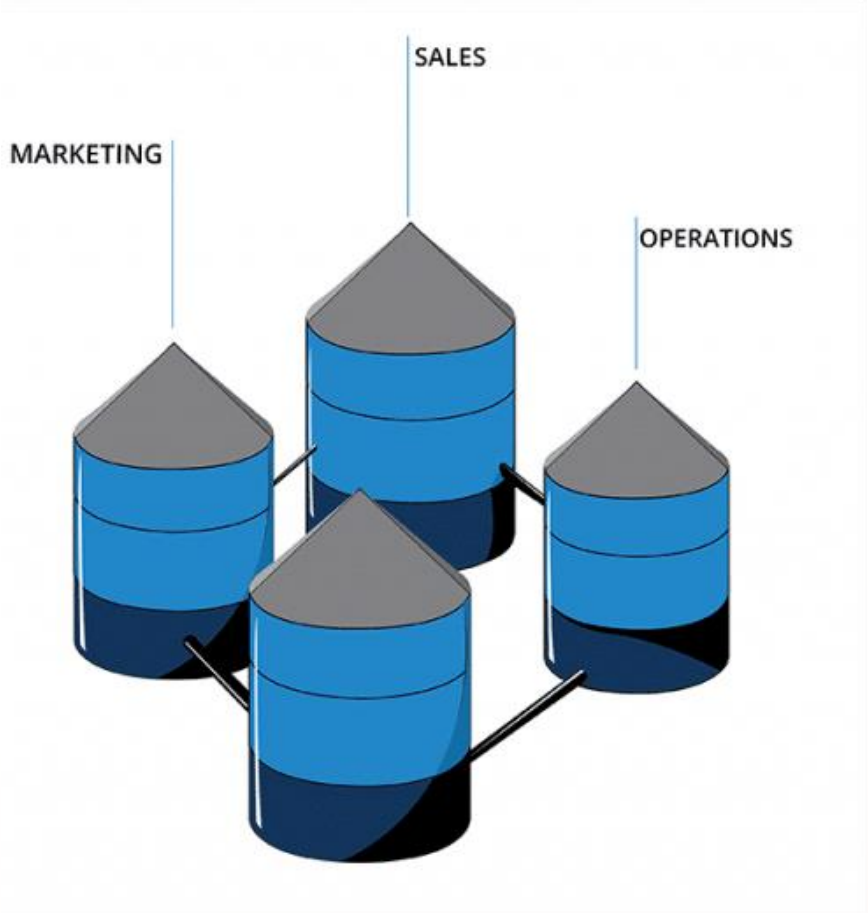
https://lawtomated.com/structured-data-vs-unstructured-data-what-are-they-and-why-care/

BIG DATA INSTITUTE

Strategy


Business Decisions
Data Services
Business Applications
IT Infrastructure

Common Issues: Data Silos

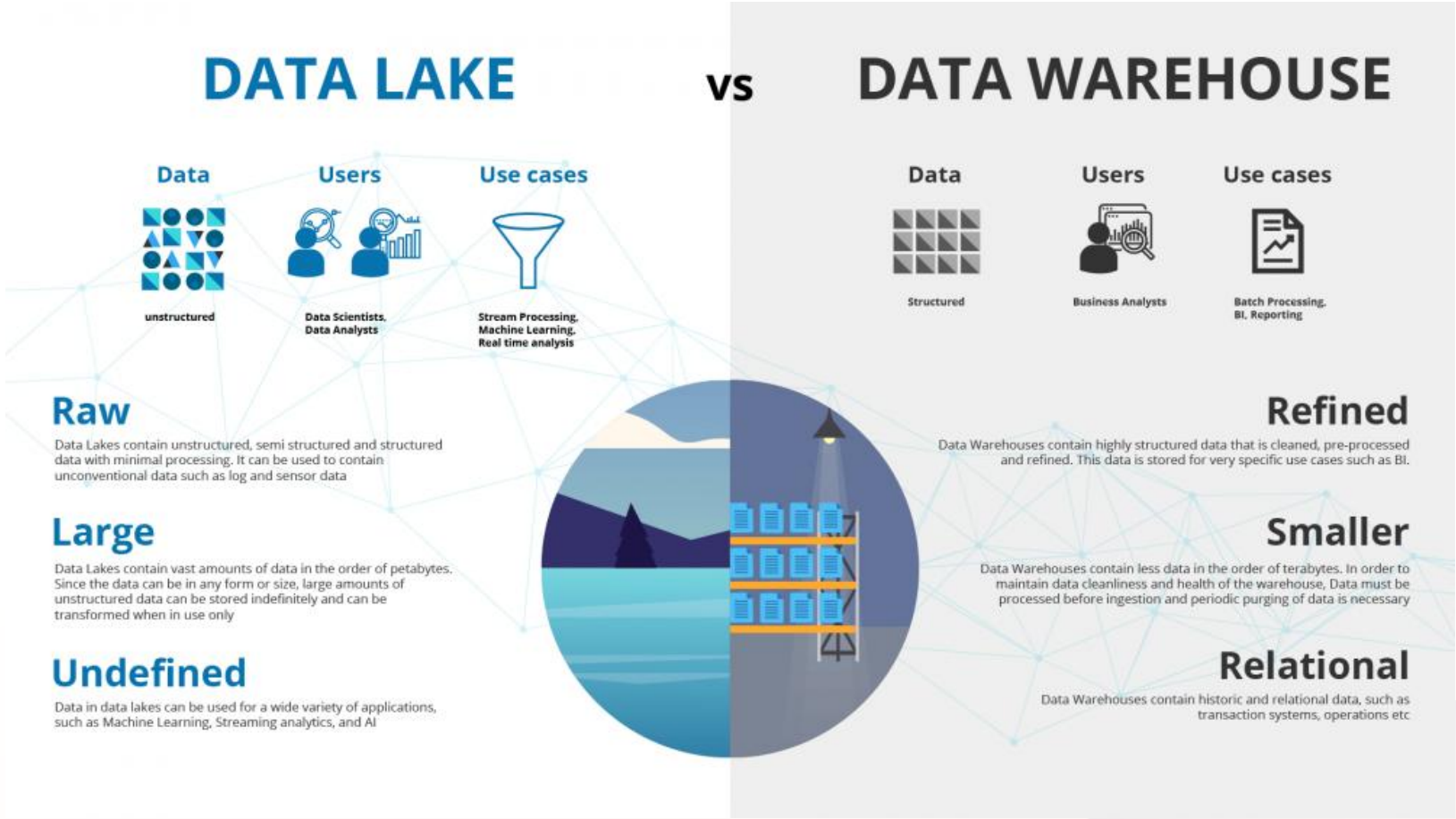


- Data Silos: separate data storages within an organization. Can result in issues such as:
  - Limited view of organization data
  - Data Inconsistency
  - Discourage collaborative work

- For AI/ML models to perform well, high quality data is needed (in preferably large amount)
  - There should be a centralized single source of truth for data.
  - This is also helpful for getting proper view of all available data

Figure: https://www.xtremepush.com/blog/what-is-the-impact-of-customer-data-silos-and-how-can-you-fix-the-problem/

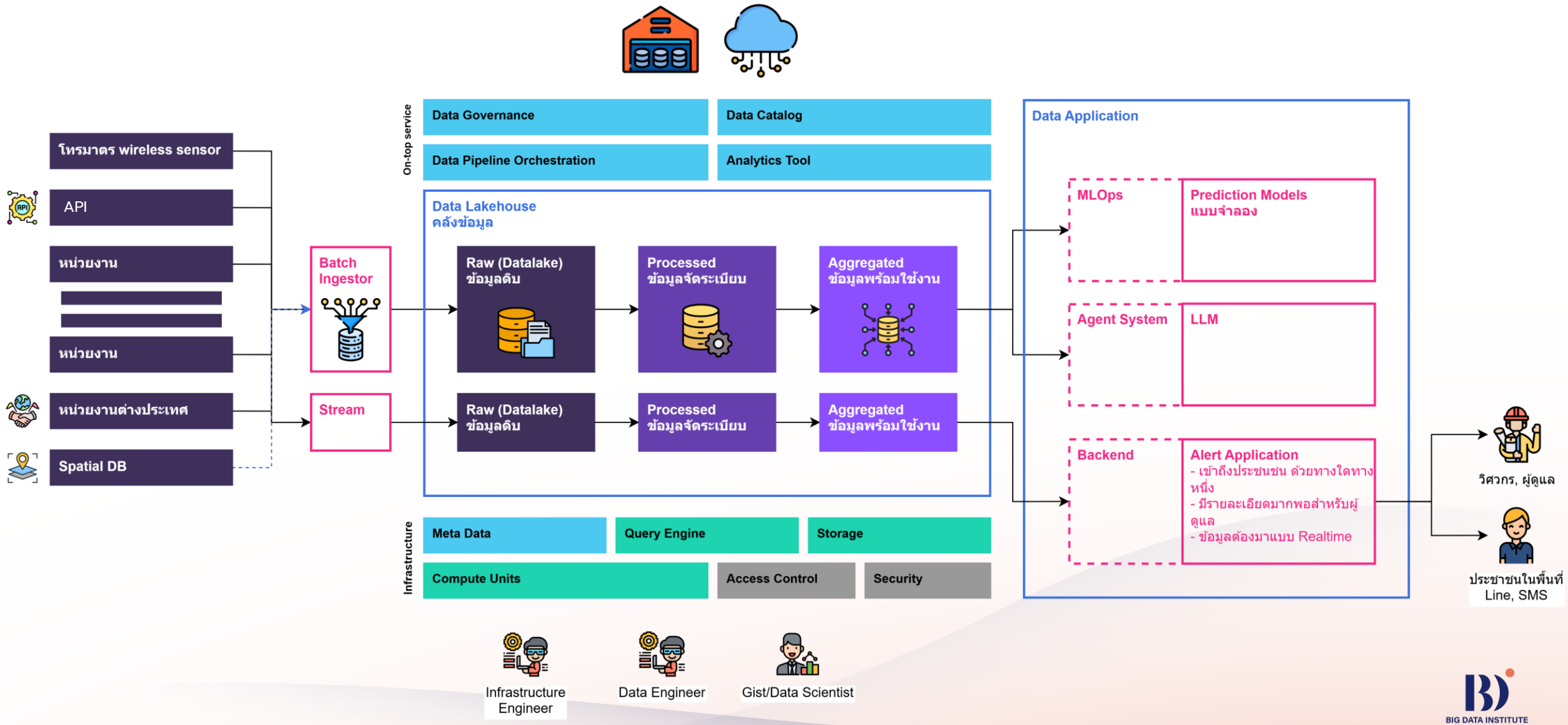
Common Solutions: Data Lake & Data Warehouse



Nowadays, though, organizations have multiple different use cases, requiring both type of storages

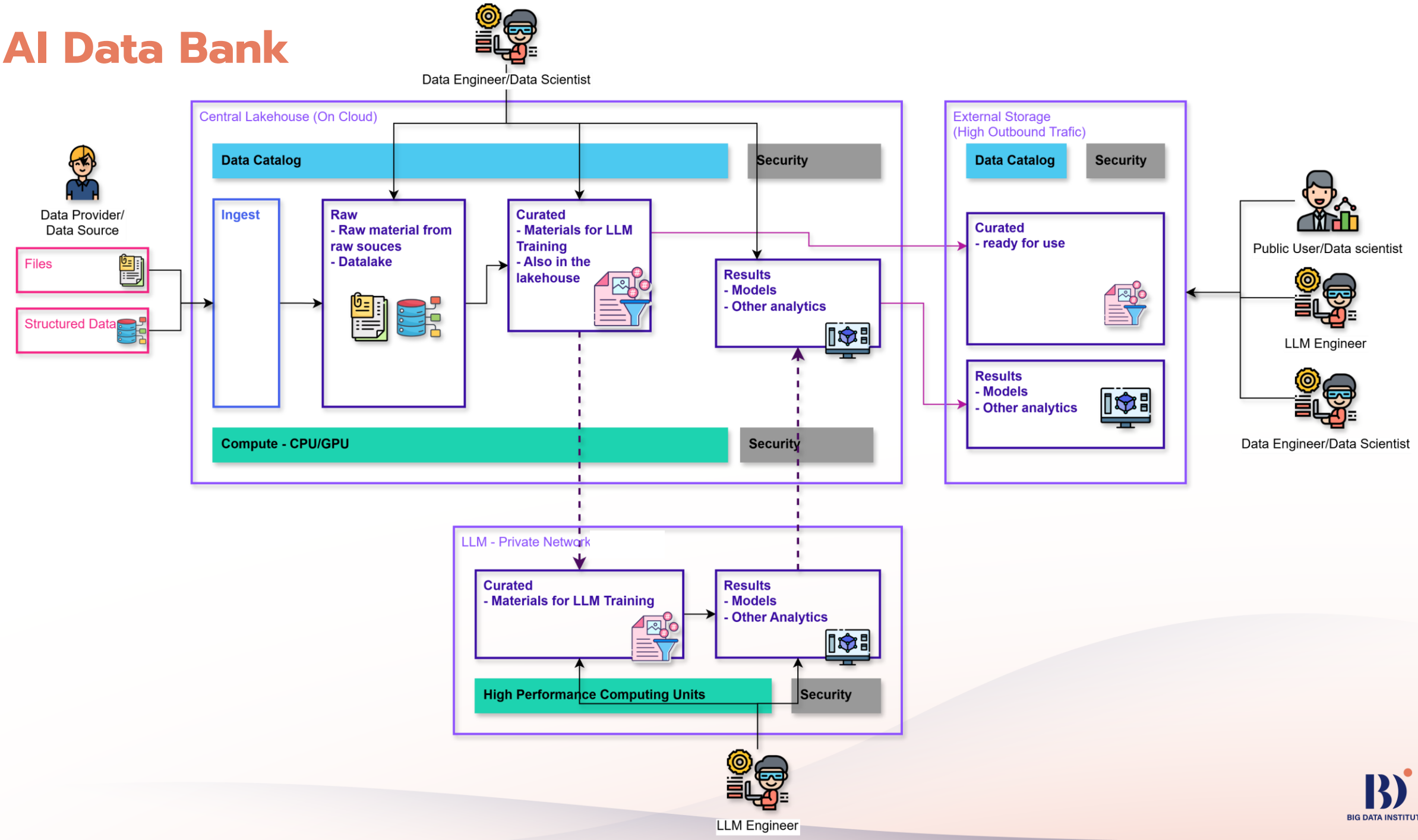
BIG DATA INSTITUTE

Data Lakehouse



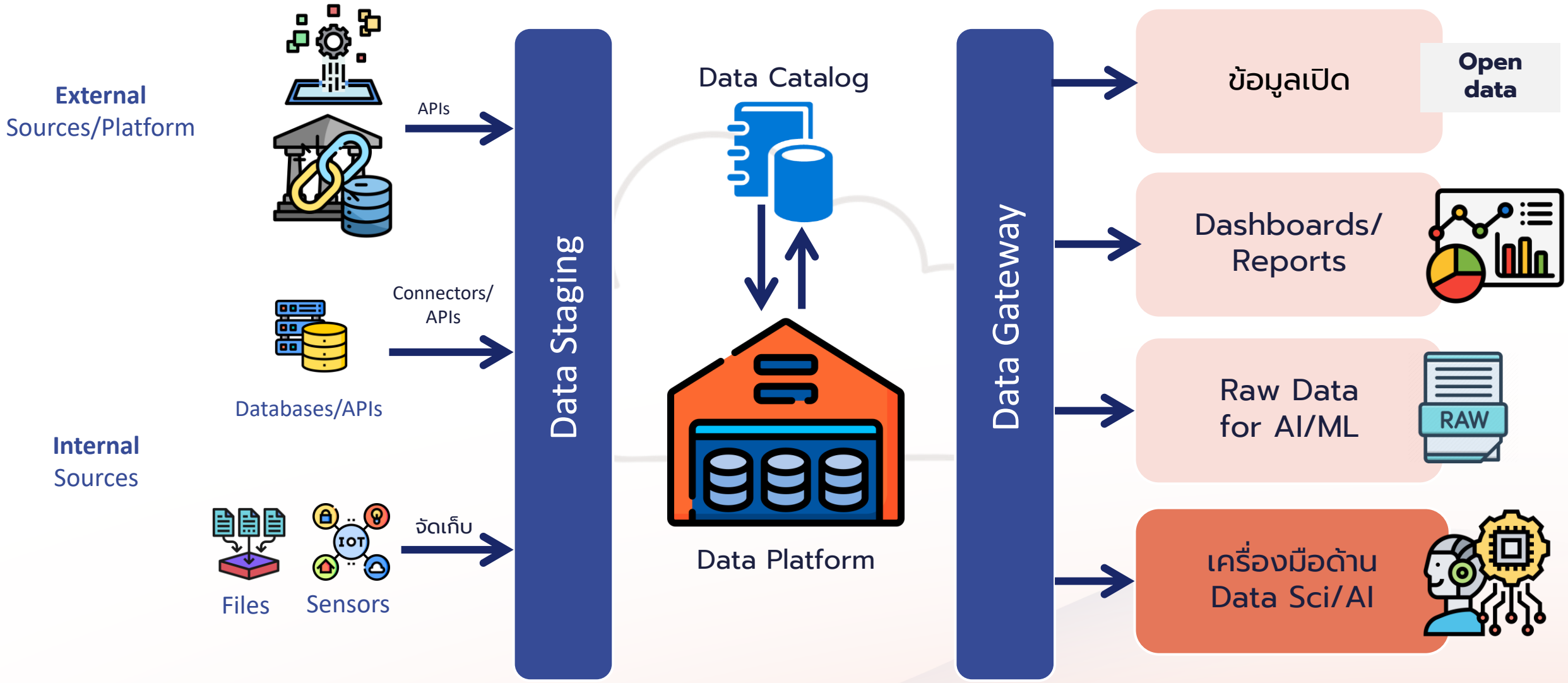
BIG DATA INSTITUTE

AI Data Bank



BIG DATA INSTITUTE

Overall view of data platform



BIG DATA INSTITUTE

Choices for deployment

- **On-premise (local)**
  - Pros: Control over data/resources. Customizable.
  - Cons: High operating/maintenance costs
- **Cloud**

Our topic

  - Pros: Low operation costs. Reliable. Scalable
  - Cons: Limited control. Require external connection
- **Hybrid**
  - Pros: Keep sensitive data locally. Cost saving
  - Cons: Complex to manage. High technical debts
- Need to decide
  - Which data goes on cloud?
  - How are we using the cloud?

BIG DATA INSTITUTE

Cloud service offerings

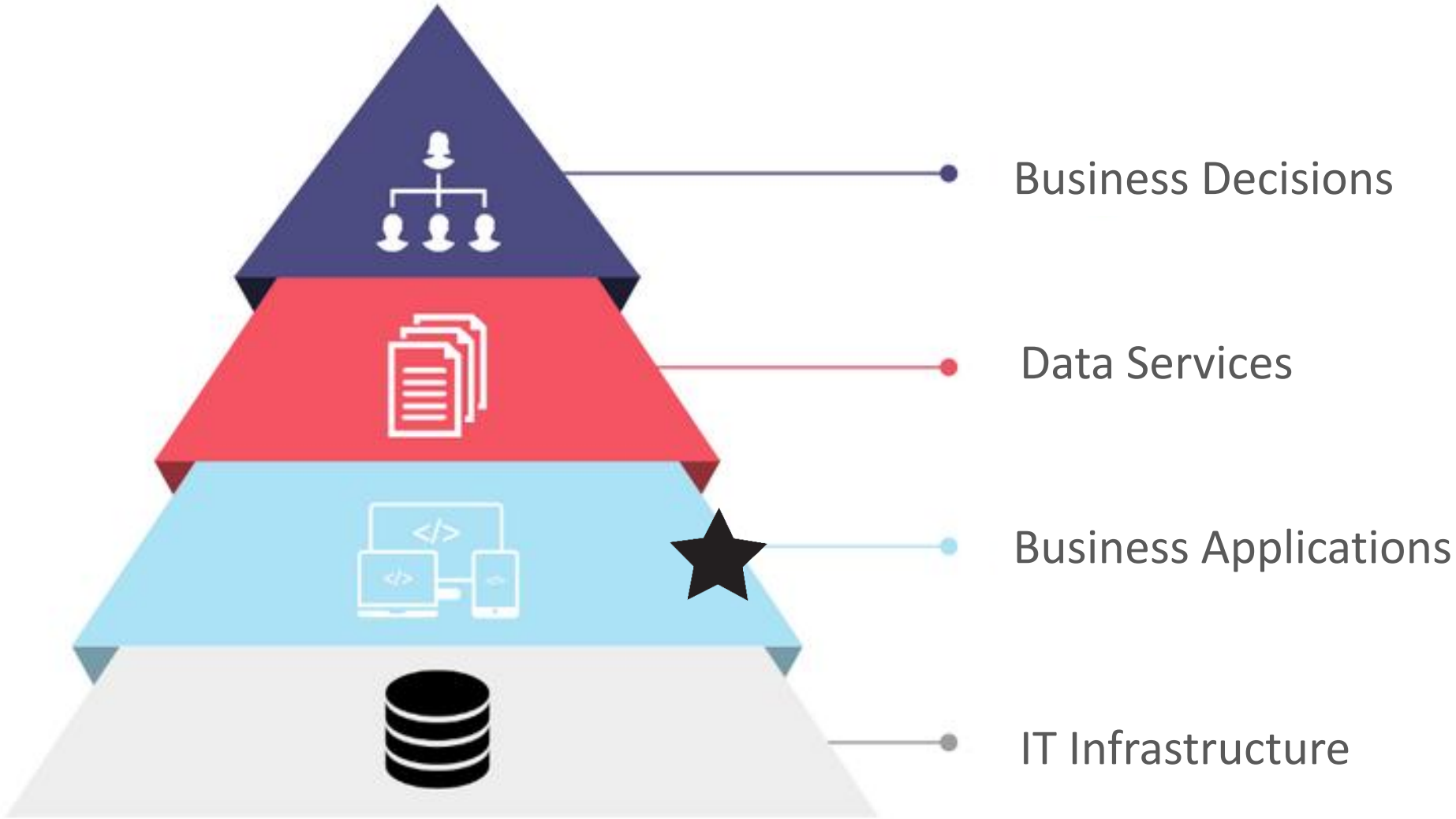
- **IaaS (Infrastructure as a Service)**
  - Provide “machines” (VM) as requested
  - Akins to lending private computers to users

- **PaaS (Platform as a Service)**
  - Provide “platform” to build applications/systems
  - Have pre-built components and tools available

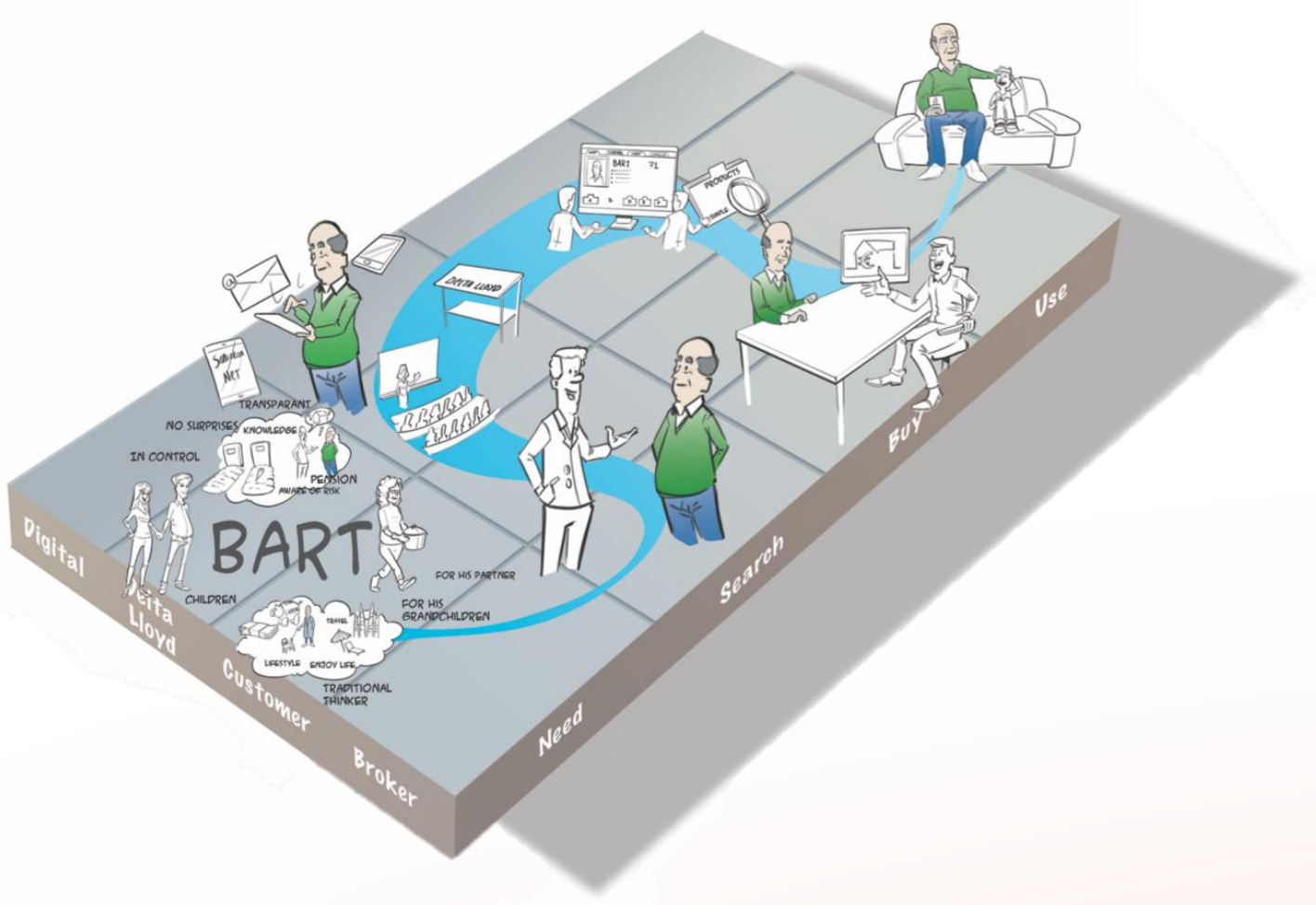
- **SaaS (Software as a Service)**
  - Provide ready-to-use “software”
  - User can access complete program via internet

Not much to talk about

BIG DATA INSTITUTE


Business Decisions
Data Services
Business Applications
IT Infrastructure

Business Journey



- It is almost never a good idea to blindly adopt technology
- Look for solutions to your problems, not the other way around
- Start from your pain points or things that should be improved
  - Look at journeys of your employees and customers

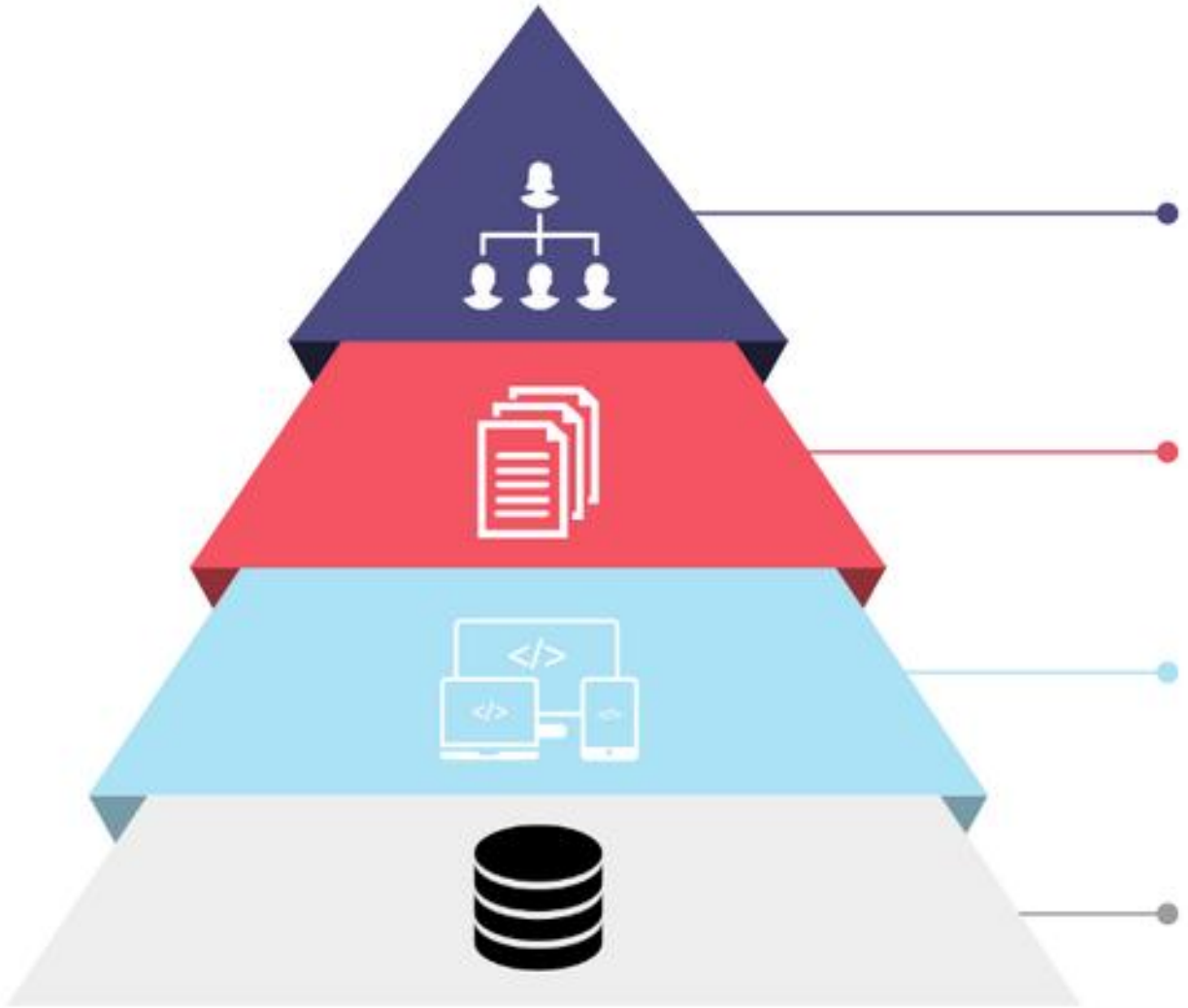
BIG DATA INSTITUTE

AI Use Cases Design

- ขั้นตอนการมองหาโอกาสในการนำ AI มาใช้งาน

1. เลือกกระบวนการทำงาน/ขั้นตอนปฏิบัติงานภายในองค์กร
2. วิเคราะห์ว่า AI สามารถเข้ามาช่วยปรับปรุงในขั้นตอนใดได้บ้าง
3. วิเคราะห์คุณค่าและความเป็นไปได้ในการนำ AI มาประยุกต์ใช้

BIG DATA INSTITUTE



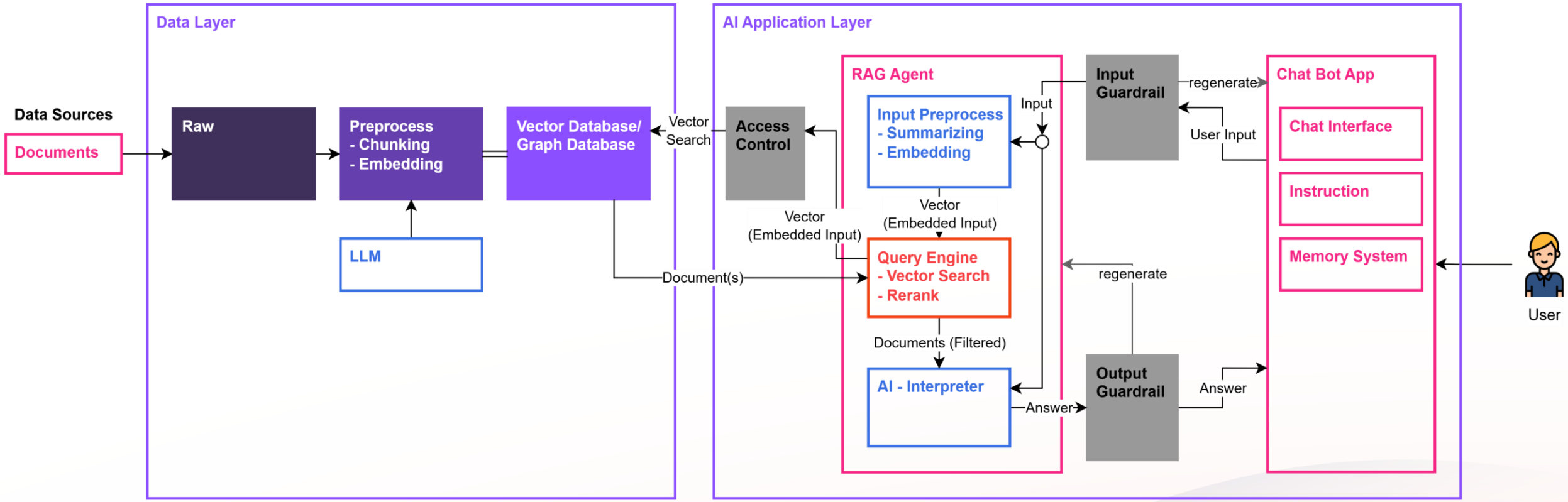
Business Decisions

Data Services

Business Applications

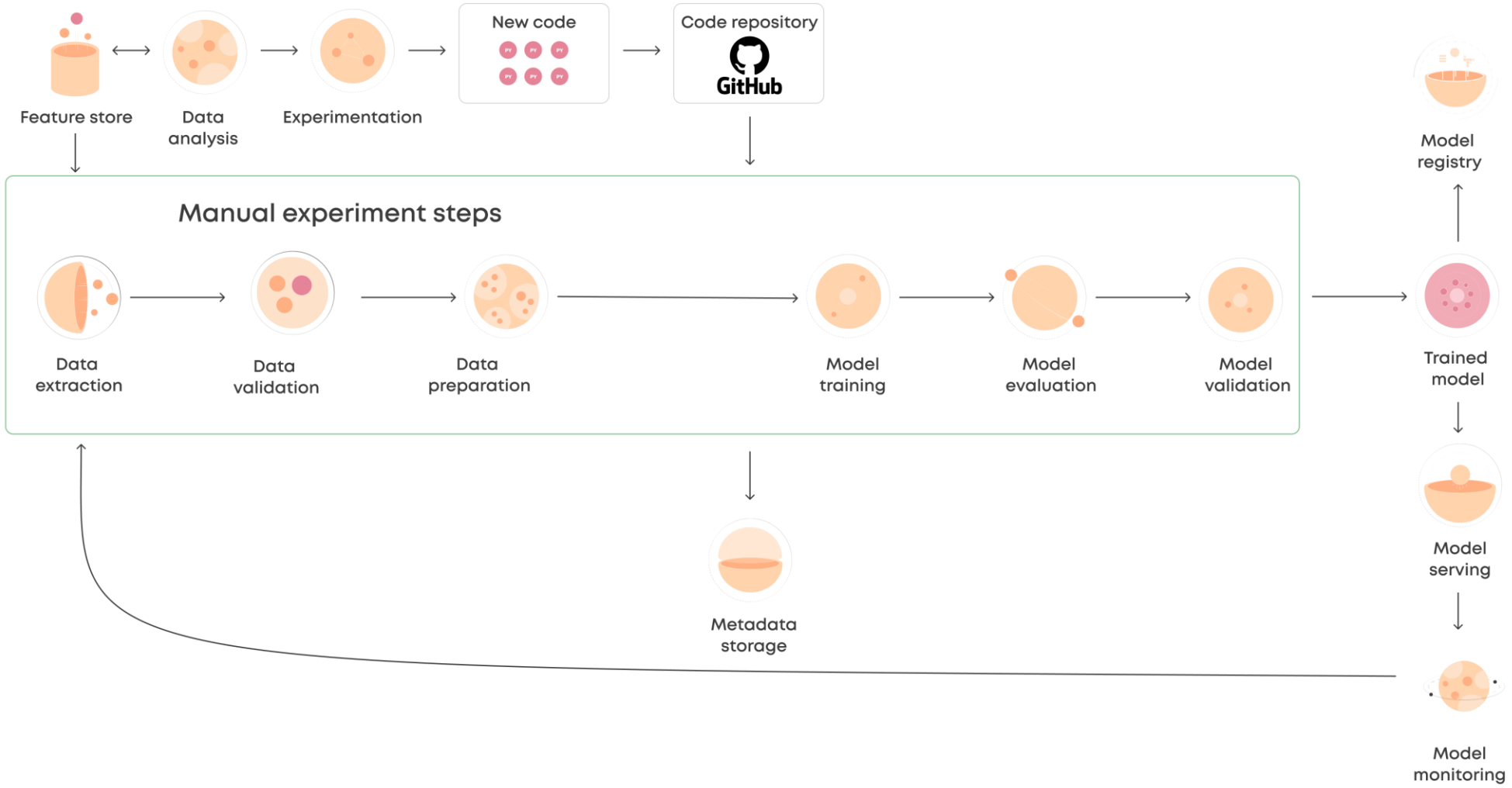
IT Infrastructure

General RAG Pipeline (document-based)



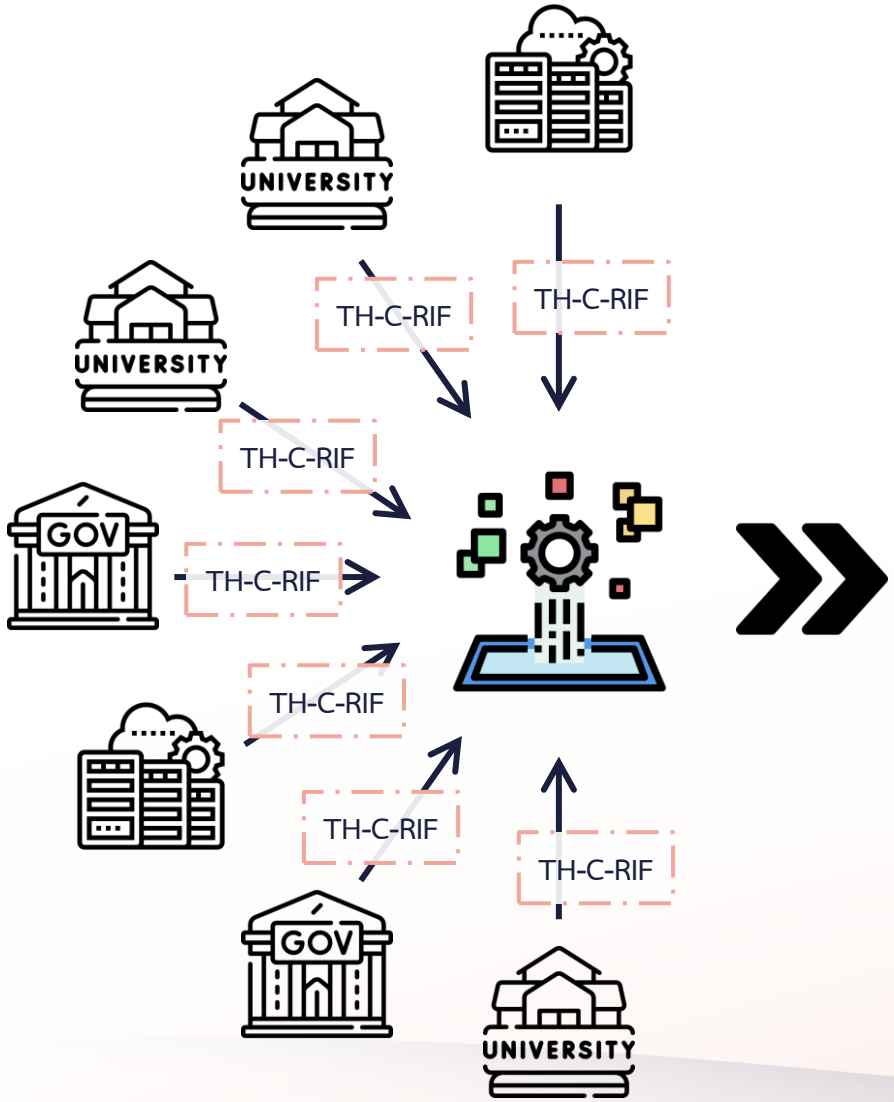
BIG DATA INSTITUTE

General MLOps Pipeline



https://valohai.com/machine-learning-pipeline/

Planning for scale – Common Data Standard



USECASE 1

Research Performance and Reporting ประกอบการกำหนดนโยบายและยุทธศาสตร์ วงน.

USECASE 2

Project Approval, Project Monitoring and Evaluation สำหรับการบริหารจัดการการให้ทุน

USECASE 3

One-Stop-Shop Research Information Service เพื่อให้บริการข้อมูลวงน.แบบเชื่อมโยงได้

USECASE อื่น ๆ

การวิเคราะห์ข้อมูลโดย Interactive Dashboard และ**แบบจำลอง**ทางคณิตศาสตร์

BIG DATA INSTITUTE

Data Governance

Data Governance : ธรรมาภิบาลข้อมูล

Accuracy *Completeness* *Timeliness* *Security* *Privacy* *Connectedness* *Worthiness*

Organization & Stewardship

Function of
- DG Committee
- Steward team
- Data Controller/Processor/User

Data management Policy

- Data Lifecycle
- Data Security & Privacy
- Data Quality Assurance
- Data Exchange

Audit

- Data Risk Management
- Law & Regulation Compliance
- Data Quality Audit

Building Knowledge & Awareness

- Program Coverage
- Measurement
- Ongoing

Data Management Policy : นโยบายการบริหารจัดการข้อมูล

Data Lifecycle

Create → Store → Use → Archive → Destroy

Data Exchange

Foundation

Data Catalog

- What data ?
- Who's the owner ?
- Search Tags
- Data sources

Data Security & Privacy

- Confidentiality
- Integrity
- Availability

Data Quality

- Accuracy
- Validity
- Timeliness
- Completeness
- Uniqueness
- Consistency

Data Governance

- Essentially, we want to be able to answer:
  - What data ?
  - Who owns it ?
  - How often is it updated?
  - Where's it from ?
- Along with being able to properly manage them

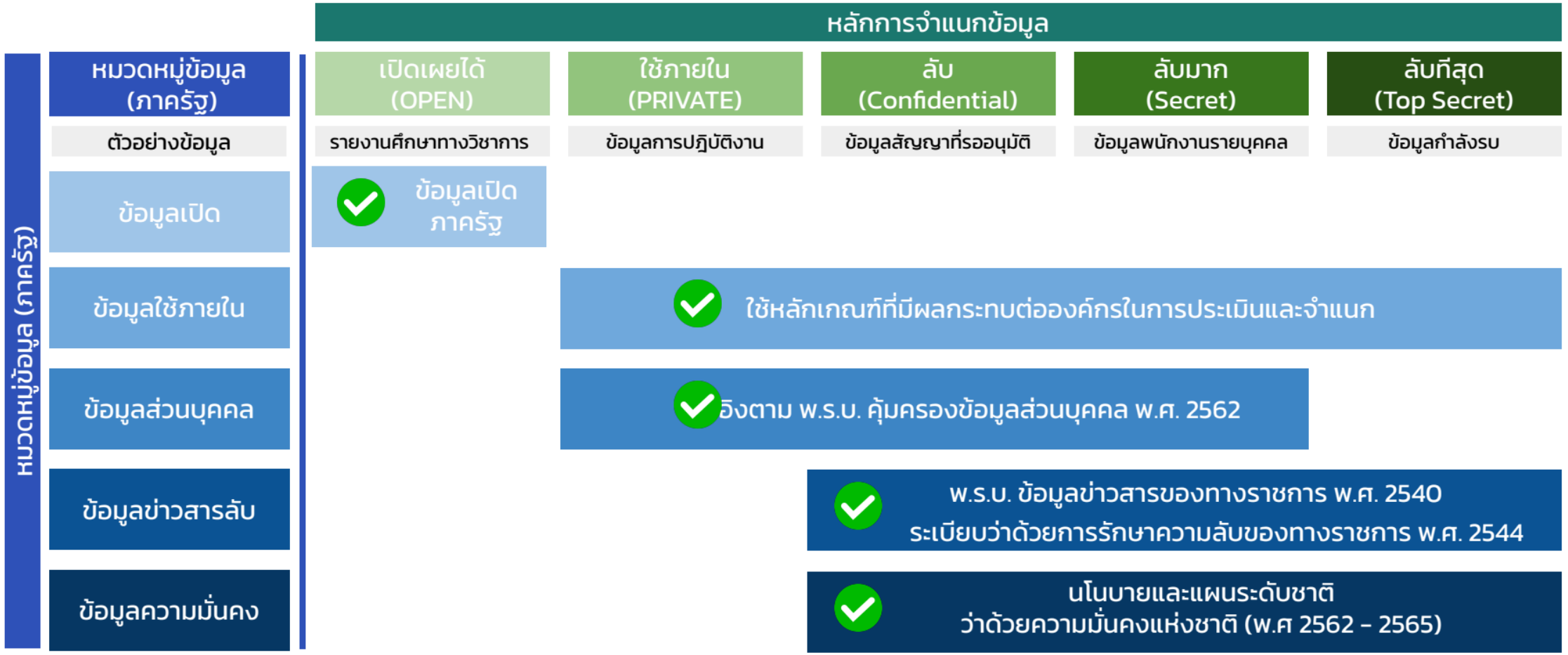
Security Control

Quality Control

Access Control

BIG DATA INSTITUTE

Data Classification



What about using/storing data on cloud?

BIG DATA INSTITUTE

Data Classification (2)

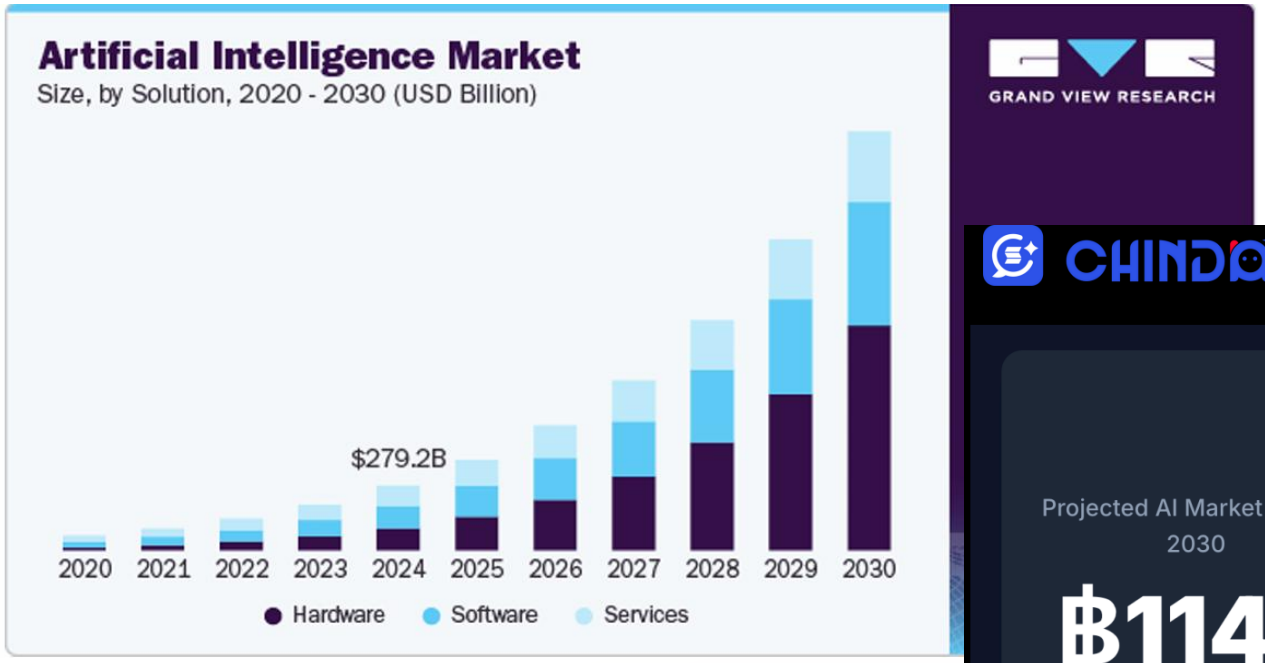
	ระดับความอ่อนไหวของข้อมูล (Sensitivity Level)	การเข้าถึง (Accessibility) (SAMPLE)	การจำแนกข้อมูล (Classification Label)	ตัวอย่างข้อมูล (Sample Type of Data)
1	Level 0 - None	ทุกคนสามารถเข้าถึงได้	เปิดเผยได้ (Public Used)	• ข้อมูล PR • ข้อมูล Website • ข้อมูลบริษัท
2	Level 1 - Low	พนักงานทั่วไป	ใช้ภายใน (Internal Used Only)	• ข้อมูลบริการ • ข้อมูลการตลาด • ข้อมูลจัดซื้อ
3	Level 2 - Medium	ระดับผู้บริหาร / IT / คณะทำงานธรรมภิบาลข้อมูล (ต้องมีการขอสิทธิในการเข้าถึงข้อมูล)	ลับ (Confidential)	• ข้อมูลการเงิน • ข้อมูลรายได้การขาย • ข้อมูล NDA / สัญญา
4	Level 3 - High	ระดับผู้บริหาร / IT / คณะทำงานธรรมภิบาลข้อมูล (ต้องมีการขอสิทธิในการเข้าถึงข้อมูล)	ลับมาก (Highly Confidential)	• ข้อมูลลูกค้า • ข้อมูลภายใน IT System • ข้อมูลพนักงานรายบุคคล
5	Level 4 - Extreme	เจ้าของข้อมูล (Data Owner) (ต้องมีการขอสิทธิในการเข้าถึงข้อมูล	ลับที่สุด (Restricted)	• ข้อมูลส่วนบุคคล • Credit Card • ข้อมูลส่วนบุคคลที่เป็นข้อมูลอ่อนไหว

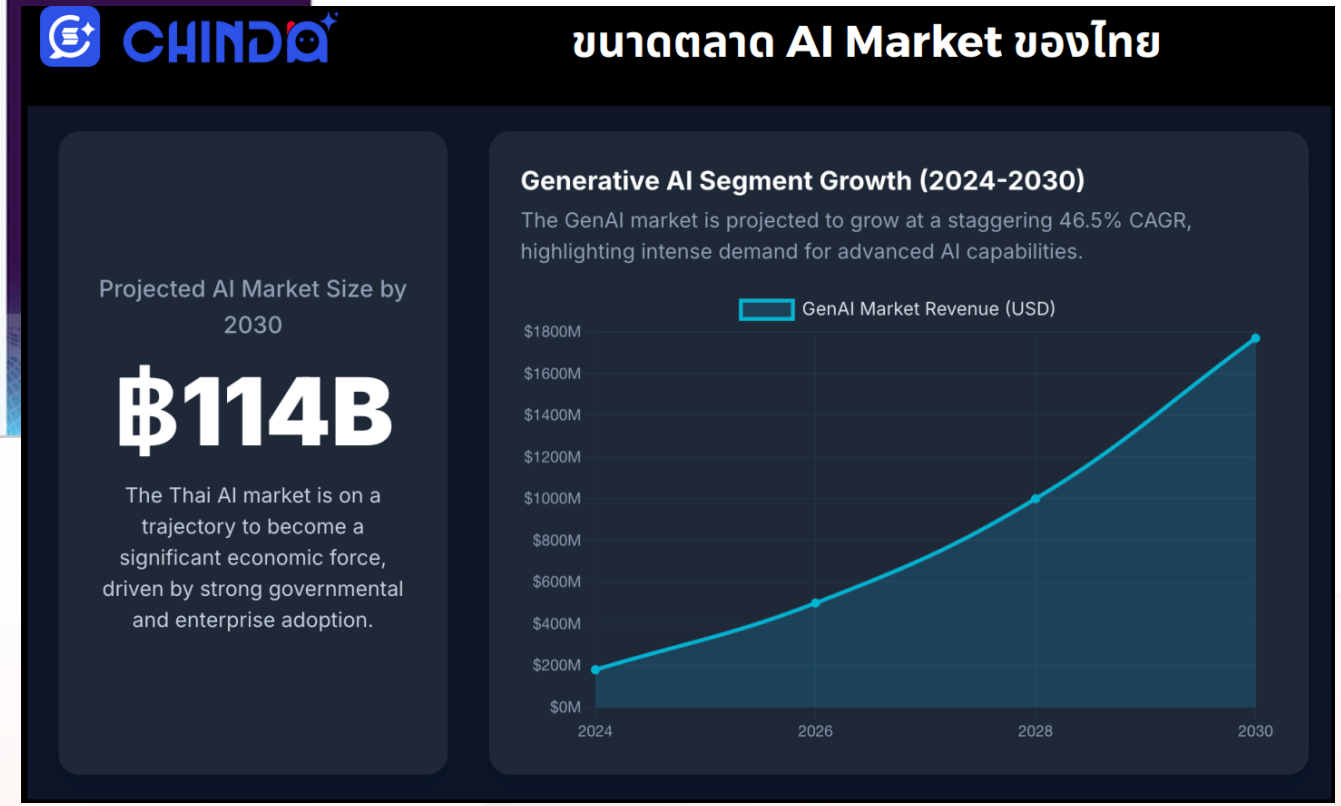
13

https://www.facebook.com/photo/?fbid=1162230310942450&set=pcb.1162236190941862

Data Governance For AI

Why do we particularly care about AI?





Source: Data Governance for AI by AIGC-ETDA, depa, iApp Technology

BIG DATA INSTITUTE

AI Governance

Competitiveness & Sustainability

Transparency & Accountability

Compliance with Laws, Ethics, Standard

Human-centric decision making

Security & Privacy

Equality & Fairness

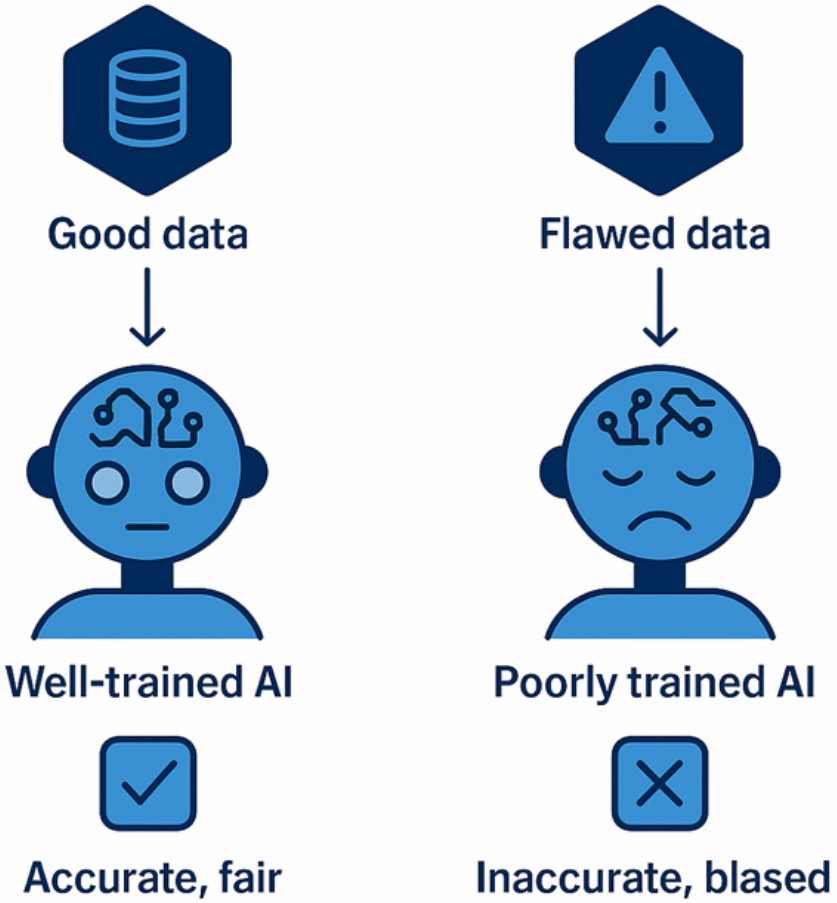
Reliability

Source: Data Governance for AI by AIGC-ETDA
images by Uniconlabs @flaticon

BIG DATA INSTITUTE

What role do data play in AI Governance?

เมื่อระบบ AI ได้รับ ข้อมูลที่มีคุณภาพ (Good Data) เช่น ข้อมูลที่ถูกต้อง ครบถ้วน และไม่ลำเอียง ส่งผลให้ AI ที่ถูกฝึกมีความสามารถในการวิเคราะห์ได้อย่าง ถูกต้อง (Accurate) และ เป็นธรรม (Fair)



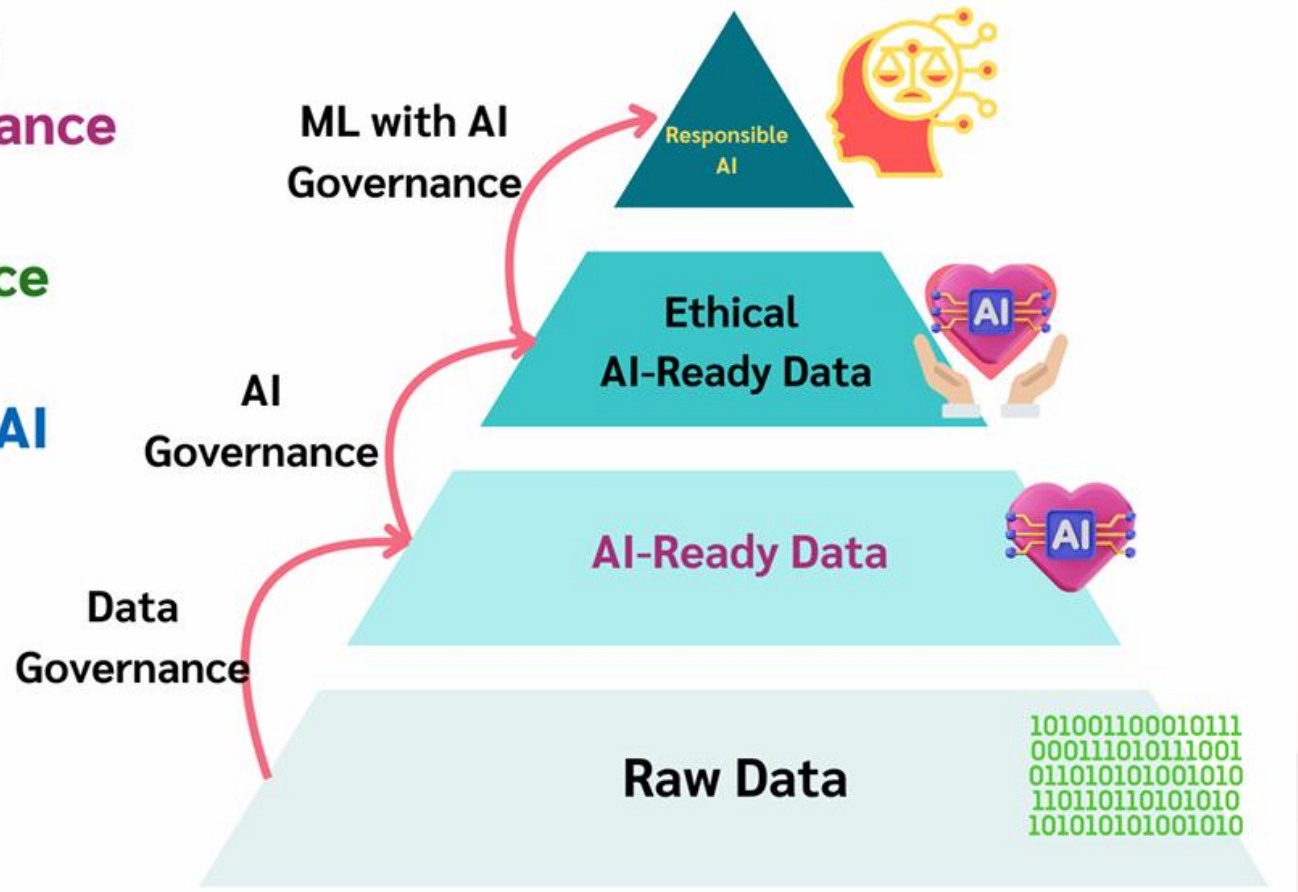
เมื่อระบบ AI ได้รับ ข้อมูลที่บกพร่อง (Flawed Data) เช่น ข้อมูลที่ล้าสมัย ขาดสมดุล หรือมีอคติ ส่งผลให้โมเดล AI มีแนวโน้ม ตัดสินใจผิดพลาด (Inaccurate) และ ไม่เป็นธรรม (Biased) ซึ่งอาจก่อให้เกิดผลกระทบทั้งเชิงนโยบาย เศรษฐกิจ และสิทธิมนุษยชน

Source: Data Governance for AI by AIGC-ETDA

BIG DATA INSTITUTE

Data Governance for AI

การบูรณาการ
Data Governance
และ
AI Governance
เพื่อมุ่งสู่
Responsible AI

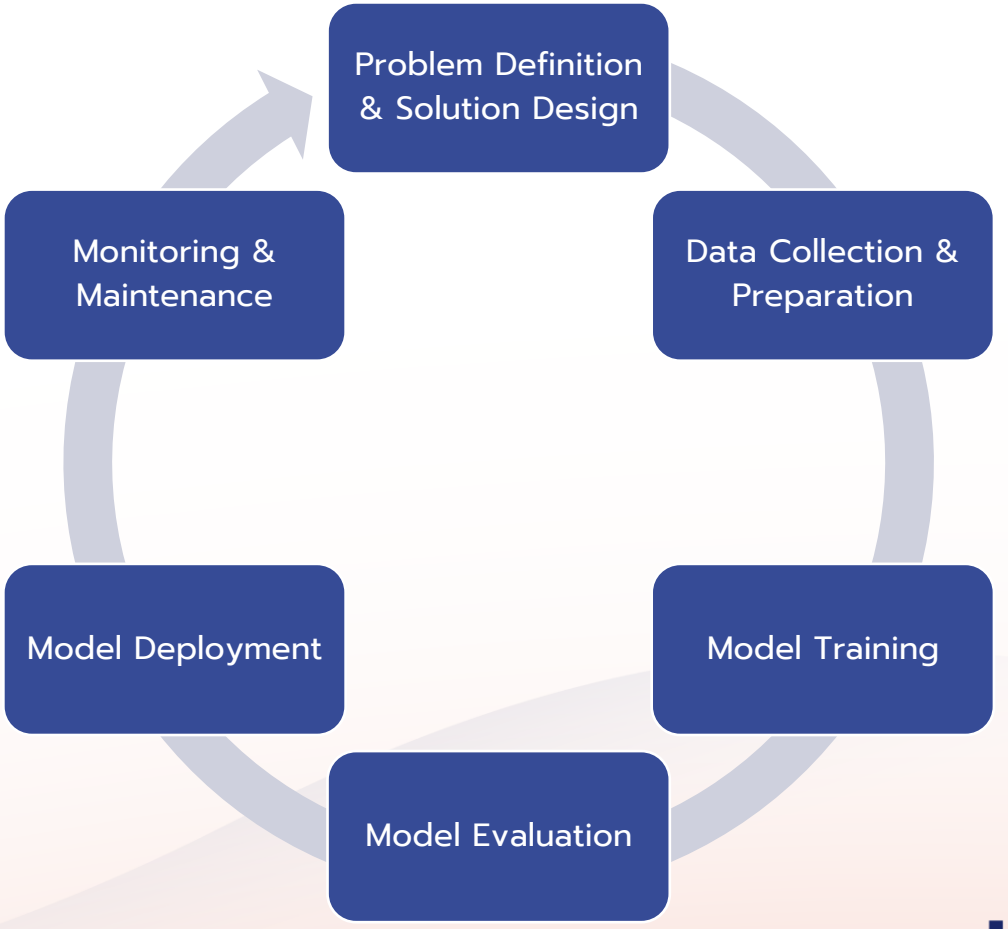


Source: Data Governance for AI by AIGC-ETDA

BIG DATA INSTITUTE

AI-Ready Data & AI Lifecycle

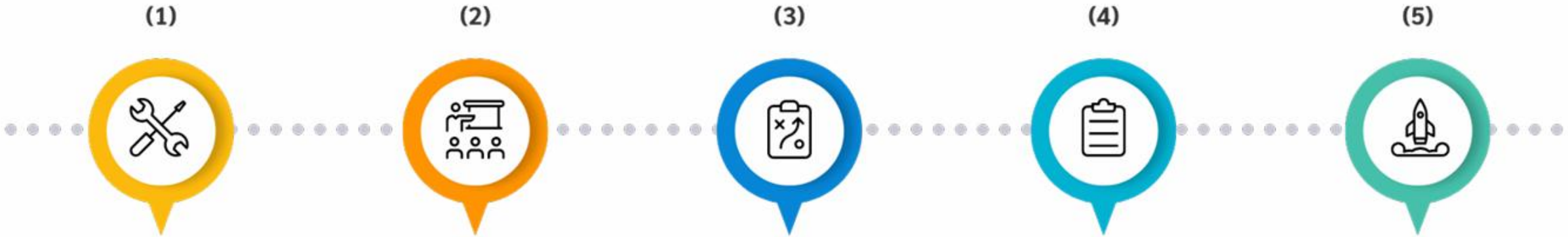
AI Lifecycle



Source: Gartner, Data Governance for AI by AIGC-ETDA

BIG DATA INSTITUTE

Bringing them together



- ควบคุมและตรวจสอบแหล่งที่มาของข้อมูลอย่างเคร่งครัด
- ขอความยินยอมจากเจ้าของข้อมูลอย่างเหมาะสม
- คำนึงถึงความหลากหลายของข้อมูลเพื่อหลีกเลี่ยงอคติในระบบ

- ตรวจสอบและทำความสะอาดอย่างเป็นระบบ
- จัดการกับค่าที่ขาดหาย (Missing Values) และดำเนินการระบุข้อมูล (Data Annotation) อย่างถูกต้องตามวัตถุประสงค์

- มีความสอดคล้องกับวัตถุประสงค์ของระบบ
- ได้รับการกลั่นกรองเพื่อหลีกเลี่ยงการสร้างอคติทางสถิติ (Statistical Bias) ที่อาจส่งผลกระทบต่อความเป็นธรรมของโมเดล

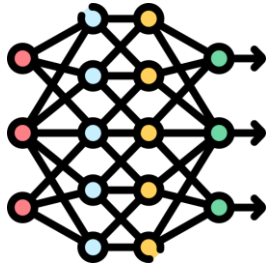
- ใช้ชุดข้อมูลใหม่ที่ไม่เคยผ่านการฝึกสอนมาก่อน เพื่อประเมินประสิทธิภาพ ความแม่นยำ และความเป็นธรรมของโมเดลอย่างรอบด้าน

- มีระบบในการติดตามและตรวจสอบข้อมูลที่นำเข้าสู่ระบบ AI
- ตรวจสอบผลลัพธ์ที่ได้จากการประมวลผล เพื่อให้สามารถประเมินความเสี่ยงและผลกระทบได้อย่างทันท่วงที

Source: Data Governance for AI by AIGC-ETDA

BIG DATA INSTITUTE

Data Preparation Guideline

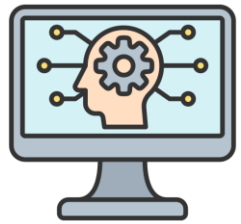


Original Model (from scratch)

- Use diverse datasets that covers all required scenario
- Perform **quality control**
- Contains **metadata** and **data lineage**
- Accurate Labeling of data

Open Source/ Pretrained Model

- Investigate **inherent bias**
- Use data that **reflect actual usage** for fine-tuning or testing
  - Transparent& Traceable
  - Do not drastically differ from original training data
  - Appropriate privacy handling
- Perform **fairness evaluation**



Off-the-shelf (service/software)

- Consider **Input appropriateness**
  - Sensitive data?
  - Personal data?
- **Validate outputs** before actual deployment
- Perform **risk assessment**

Source: Data Governance for AI by AIGC-ETDA
Images by Freepik, Konkapp, Nuricon from flaticon

BIG DATA INSTITUTE

Data Quality Criteria

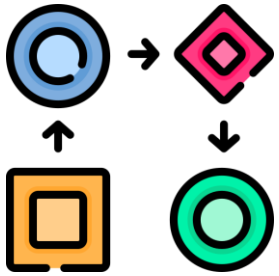
เกณฑ์	ความหมาย
Accuracy	ข้อมูลถูกต้อง เป็นไปตามความจริง
Consistency	ข้อมูลมีความสอดคล้อง ไม่ย้อนแย้งกันเอง
Availability	สามารถเข้าถึง/ใช้งานข้อมูลได้ทั้งในปัจจุบันและในอนาคต
Completeness	ข้อมูลมีความสมบูรณ์ ครบถ้วน
Conformance	ข้อมูลเป็นไปตามมาตรฐานที่เป็นที่ยอมรับ
Credibility	ข้อมูลได้มาจากแหล่งข้อมูลที่เชื่อถือได้
Processability	ข้อมูลสามารถถูกใช้งานและประมวลผลได้โดยระบบสารสนเทศ
Relevance	ข้อมูลมีจำนวนมากพอที่จะใช้งานได้อย่างมีประสิทธิผล
Timeliness	ข้อมูลมีความทันสมัย ใหม่พอที่จะใช้ประโยชน์ได้อย่างมีประสิทธิผล
Context	ข้อมูลเหมาะสมที่จะใช้งานในบริบทที่ต้องการ

BIG DATA INSTITUTE

4 Key Components of Data Governance for AI

Roles & Responsibilities

- Data Owner
- Data Steward
- Data Custodian
- AI Model Owner



Metadata & Data Lineage

- **Metadata:** data that explain the context of data
  - Format, owners, update freq., etc.
- **Data Lineage:** the process of tracking data over time
  - Origin, end point, transformation

Policy & Standards

- Purpose Limitation
- Consent Management
- Data Management Policy/Lifecycle

Risk Management

- Data Bias
- Privacy Risk
- Legal Risk
- Mitigation:
  - Bias Assessment
  - De-identification
  - Impact Assessment

Source: Data Governance for AI by AIGC-ETDA
Images by Freepik from flaticon

BIG DATA INSTITUTE

Example of Roles & Responsibility in AI Life Cycle

RACI CHART

AI Life Cycle
(Roles with Responsibilities)

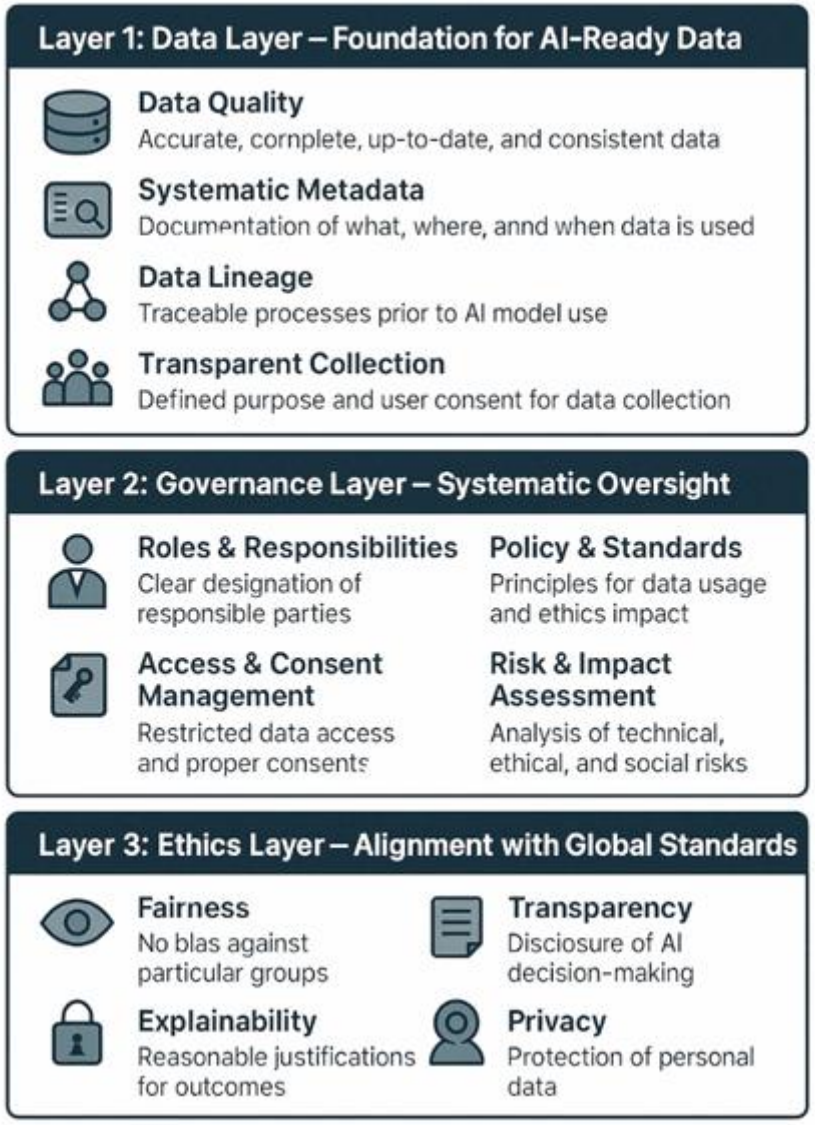
Roles	Collect	Prepare	Train	Monitor
Data Scientist	C	C	R	R
Data Engineer	R	R	A	C
Data Analyst	C	I	C	I
Project Manager	A	A	C	A

- R: Responsible
- A: Accountable
- C: Consulted
- I: Informed

Source: Data Governance for AI by AIGC-ETDA

BIG DATA INSTITUTE

AI-Ready Data Governance Framework



Source: Data Governance for AI by AIGC-ETDA

BIG DATA INSTITUTE

Thank You