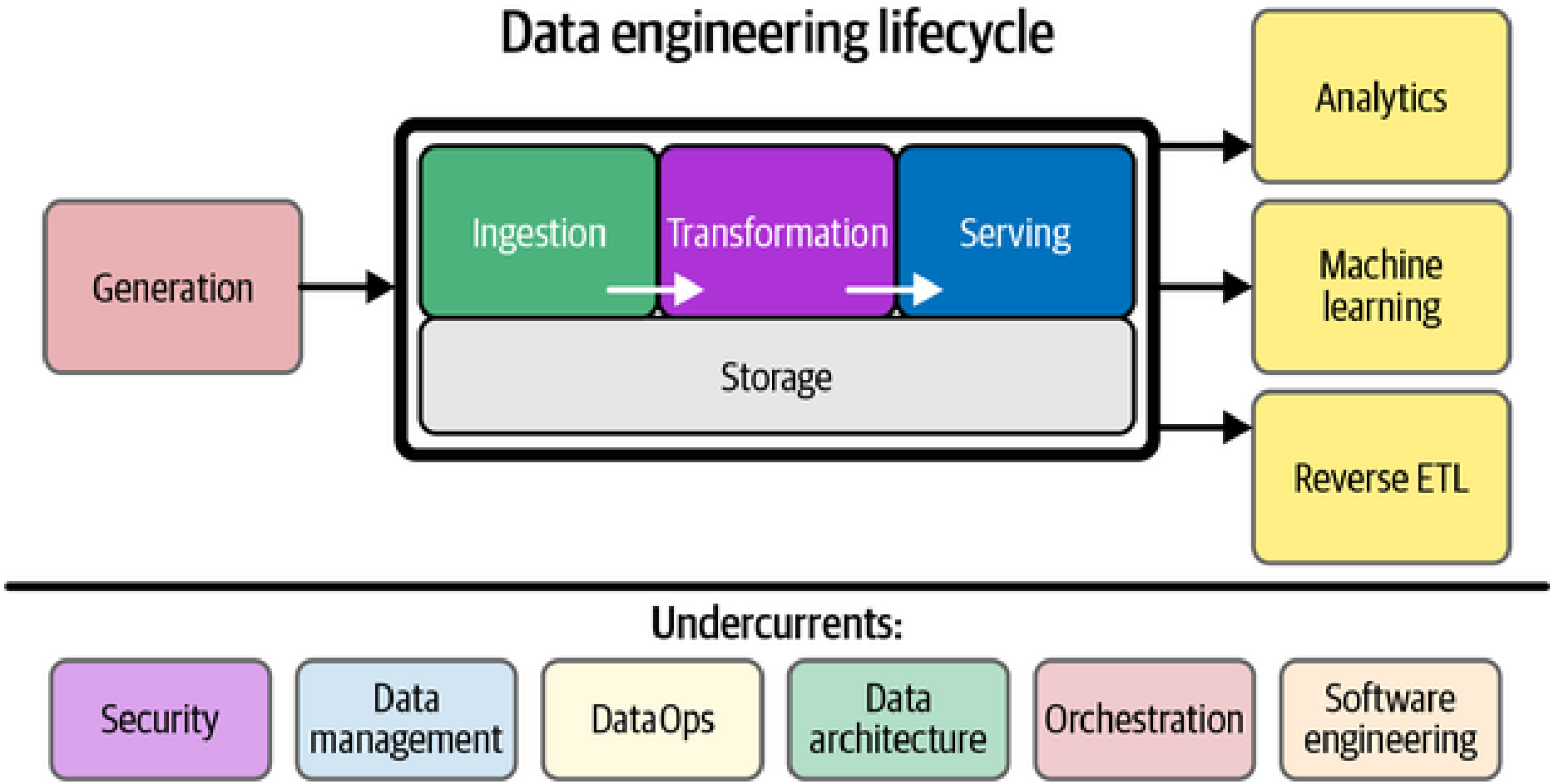
[639] Data Generation

Meenakshi Syamkumar

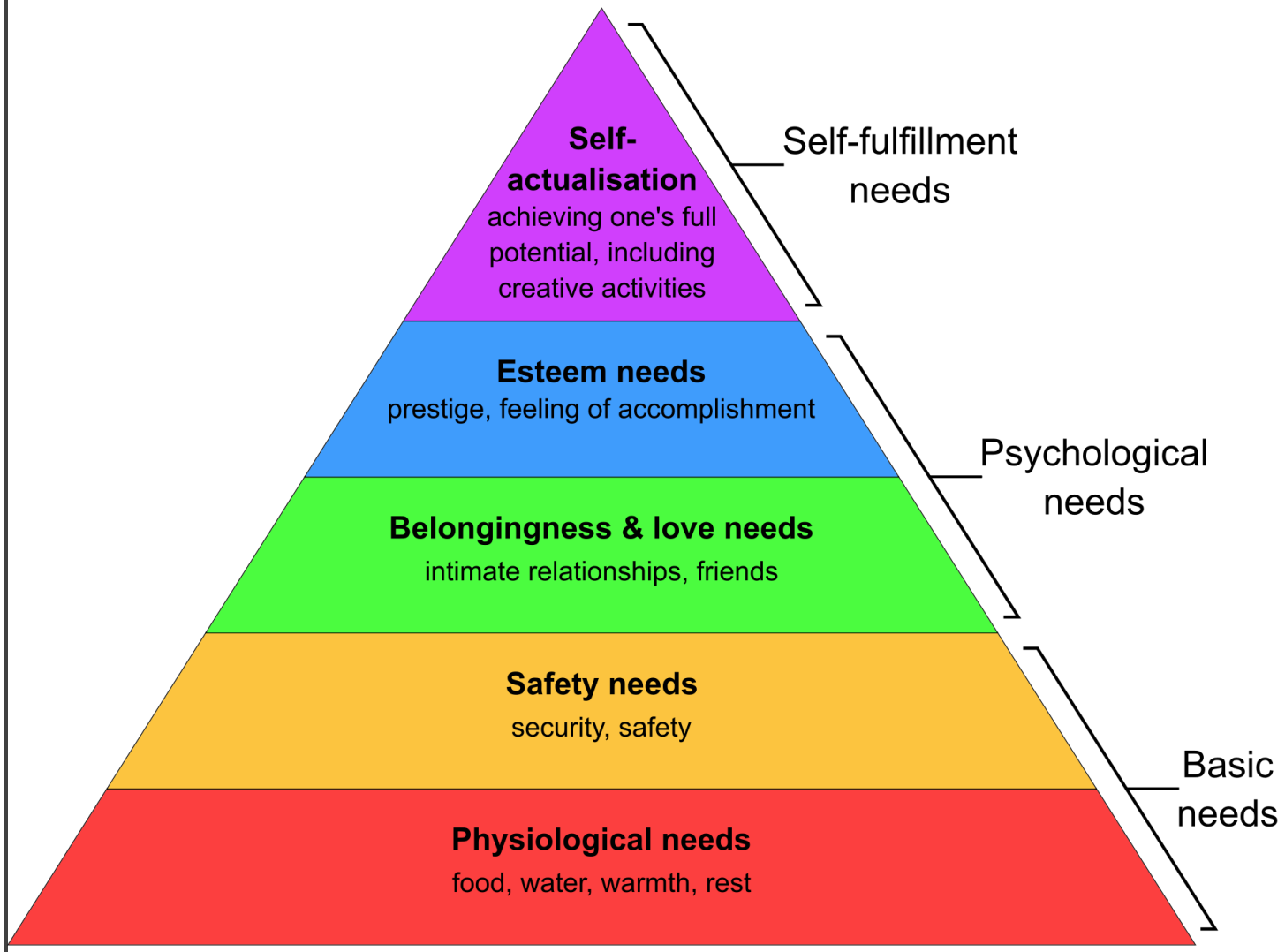
Learning Objectives

Understand	the data engineering lifecycle
Recognize	various operations source system patterns

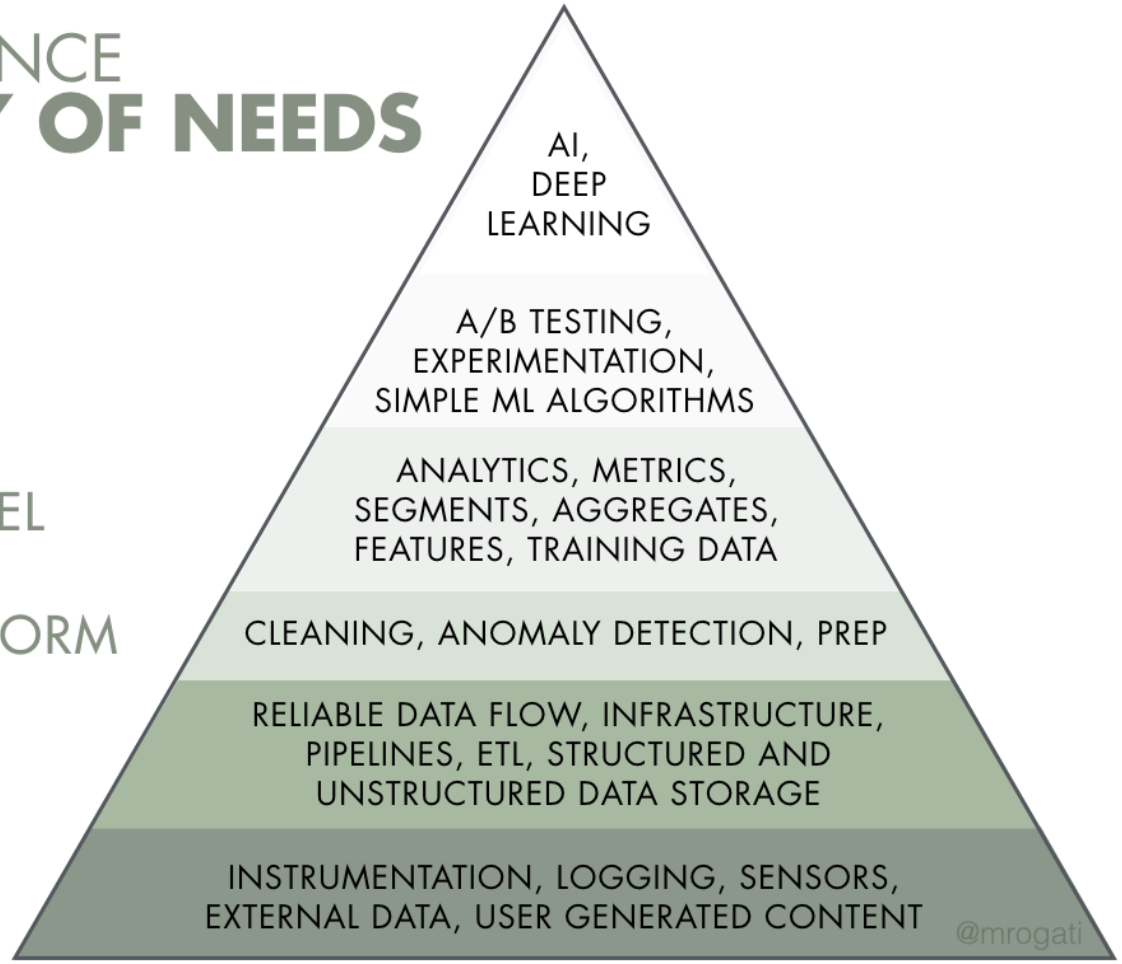
Data Engineering Lifecycle



Maslow's Hierarchy of Needs



THE DATA SCIENCE **HIERARCHY OF NEEDS**



LEARN/OPTIMIZE

AGGREGATE/LABEL

EXPLORE/TRANSFORM

MOVE/STORE

COLLECT

Source Systems

How is data created?

- Data: unorganized, context-less collection of facts and figures
- Two types of data:
  - Analog: vocal speech, sign language, writing on paper, playing an instrument
  - Digital:
    - Converting analog data into digital form (ex: speech-to-text apps)
    - Native product of a digital system (ex: credit card transaction)

Source Systems

- Files
  - Structured: Excel, CSV
  - Semi-structured: JSON, HTML, XML
  - Unstructured: TXT
- Application Programming Interfaces (APIs)
  - Standard way of exchanging data between systems
- Application databases (OLTP Systems)
  - Stores state of an application (ex: account balances for bank accounts)
  - Online Transaction Processing (OLTP) aka transactional databases efficiently read and write individual records at a high rate
  - Transaction: ACID principle

Source Systems

- Online Analytical Processing System (OLAP)
  - Large analytical queries
  - Ex: columnar databases
- Change Data Capture (CDC)
  - Method for extracting change event in data – insert, update, delete
  - Create an event stream for downstream processing
- Logs
  - Information about events that occur in systems ex: traffic and usage patterns on a web server
  - Sources: OS, applications, servers, containers, networks, IoT devices
  - Who? What happened? When?

Source Systems

- Database Logs
  - Write-ahead logs; binary files stored in specific database-native format
  - Crucial role in database guarantees and recoverability
- CRUD (Create, Read, Update, and Delete)
  - Basic tenet: data must be created before being used. After creation, data can be read and updated. Finally, data needs to be destroyed.
- Insert-Only
  - Rather than updating records, new records get inserted with a timestamp whenever we try to make changes.
  - Ex: Customer address update
  - Maintains a database log directly in the table.

Source Systems

- Messages and streams
  - *Message*: raw data communicated across two or more systems (discrete, singular signals)
  - *Message queue*: message is sent from publisher to consumer; once message is received and action is taken, message gets removed
  - Event-streams: ordered append-only log of records that are persisted over a long time
- Types of Time
  - Event generation time
  - Even ingestion time
  - Event processing time

Data Management Implications

- Data Governance
- Data Quality
- Schema
- Master Data Management
- Privacy and Ethics
- Regulatory