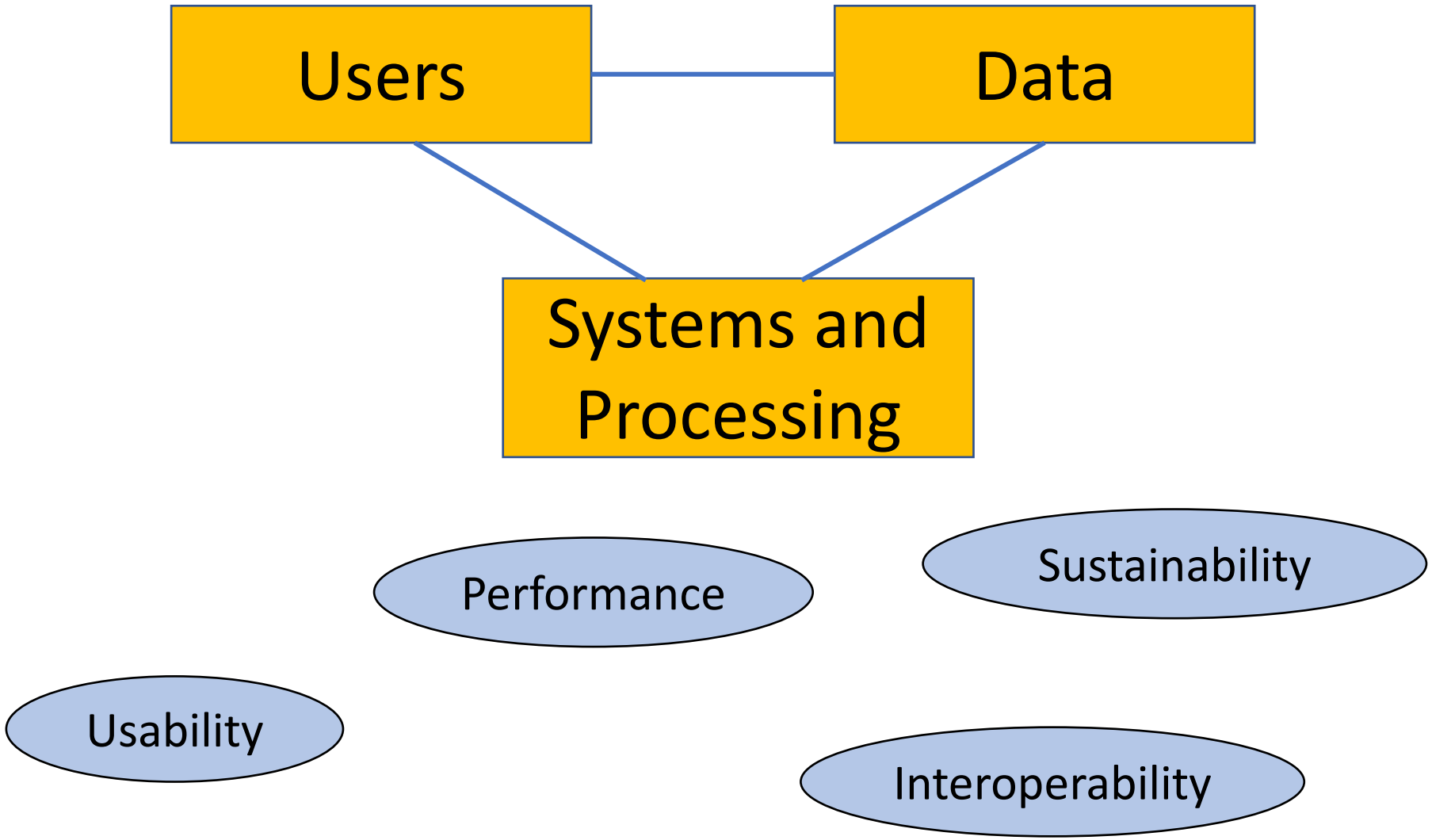
# Big Data Platforms
# (the Data Engineer’s job)

# Challenges: The Big Picture



# The Data

- Variety of Data Models
  - Relational, semi-structured, graph, multi-dimensional, text,
  - Social, semantic, streaming
- Data Integration
  - Virtual / physical
  - Distributed / central
- Data transformation
- Declarative and programmatic interfaces
- *Secondary data*
  - Meta-information for data integratin
  - Index etc.
  - output of analytics / transformation
- Lineage / provenance
- Sustainable data repository (long term preservation)
- Fair data principles: findable, accessible, interoperable and reusable

# Systems and processing

- Declarative query evaluation over diverse data
- Data management vs. data analysis
  - Function vs. data shipping
  - Streaming data
- Non-functional / performance
  - Distribution
  - Scalability
  - Efficiency
    - Management of secondary data structures
  - Scheduling
    - Real-time vs. batch
  - Trade-off: resource usage vs. goodness of results
  - Deployment and understanding infrastructure
    - Cloud, central, controlled distributed
  - Optimization techniques
    - Per component, holistic
- Tool management
  - Linking tools to make workflows possible
  - Reusability
  - Reproducability (lineage helps)
  - Sustainability (Long-term maintenance)

# Users and usability

- The “right” query language
  - Raising the level of abstraction
  - Hiding complexity
  - Providing more expressiveness
- Transparency
  - Make result understandable (not a black box)
- Tools to simplify the customization to application domains
- Tools to help applications to choose the right configuration (data models, libraries, algorithms, ….)
- Training – there are so many tools around

# Intersection with other themes

- Privacy / Security
  - Support contextual information
    - Allows to specify rules
    - Platform needs then mechanisms to enforce rules
  - Lineage / provenance help
  - Consent withdrawal
- Trustable and usable data
  - Feedback loop with platforms
    - Platforms tells theme 1 that it needs more / less / better data
  - Ask also for trustable and usable models
  - Ask also for prep that helps with privacy and security management
  - Provenance / Lineage
  - Integration and transformation – where is it exactly?

# Intersection with other themes

- Analytics
  - EVERYWHERE
  - Tightly intertwined
    - Produces more information that needs to be stored
    - Data vs. function shipping
    - Workflow support
    - interoperability
  - DM for ML and ML for DM
- Visualization and dissemination
  - Interlinks with all aspects of users / usability

# The tough one

- SOOO many existing tools / platforms
    - Training
    - Linkage
    - Usage and reuse
    - Maintenance
- Transformative: simplify the ecosystem / system complexity