Quality AI Requires Quality Data

Lower Risk, Improve Results, Avoid Embarrassment!

Edward Pollack
Microsoft Data Platform MVP

Our Sponsors

Microsoft

PURE STORAGE

ADASTRA

snowflake®

avanade

Simple Talk

Pythian
love your data®

LUCID
Data Hub

TDPC
It's all about data
Toronto Data Professionals Community (TDPC)

SQL Saturday
(#1093)

Community Support

Toronto Data Professionals Community (TDPC), one of the largest data professional's community in Toronto, host monthly event which offers interactive learning built by community and guided by trusted data experts.

TDPC Event Partners

Toronto Data Professionals Community (TDPC)

Ed Pollack

Microsoft Data Platform MVP

Published author of:

- Dynamic SQL: Applications, Performance, and Security, 2nd Edition
- Analytics Optimization with Columnstore Indexes in SQL Server
- Expert Performance Indexing in SQL Server, 4th Edition
- Published in Expert T-SQL Functions in SQL Server, 3rd Edition

Author on Simple Talk.

Organizes:

- SQL Saturday Albany 2024
- SQL Saturday New York City (Tentatively: May 10th)
- Future Data Driven
- Capital Area SQL Server User Group

Speaker at many data events

Find me on: LinkedIn

SQLSATURDAY

transfinder

Simple Talk

Valued Author

Making a Difference through Books and Video

Apress®

redgate

COMMUNITY AMBASSADOR 2024

MVP

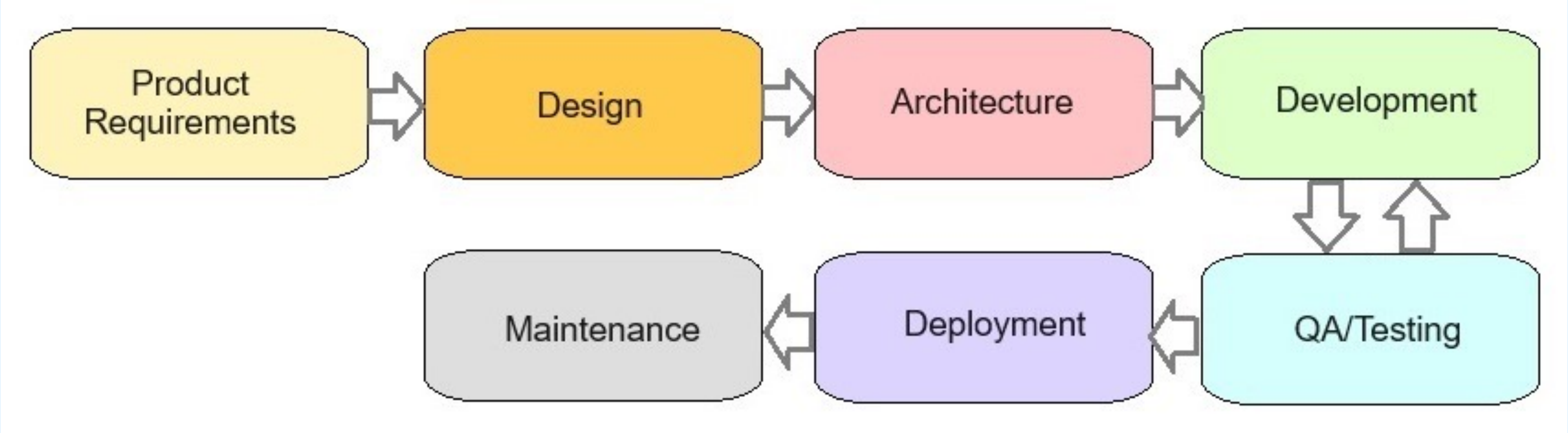
Microsoft® Most Valuable Professional

Agenda

- Why does data quality matter?
- Best practices to improve data quality.
- What are frequent data mistakes made in AI?
- Bring-it-all-together!

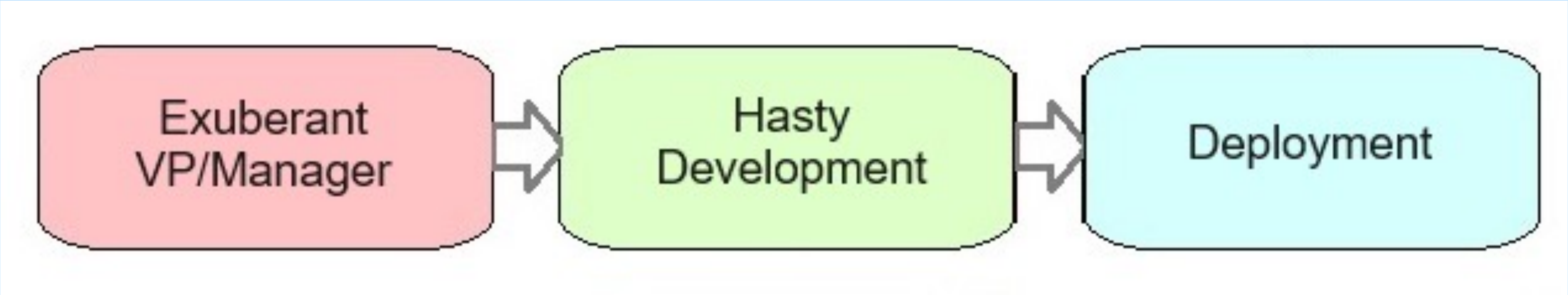
Software Development Life Cycle

(Sometimes)



AI Development Life Cycle

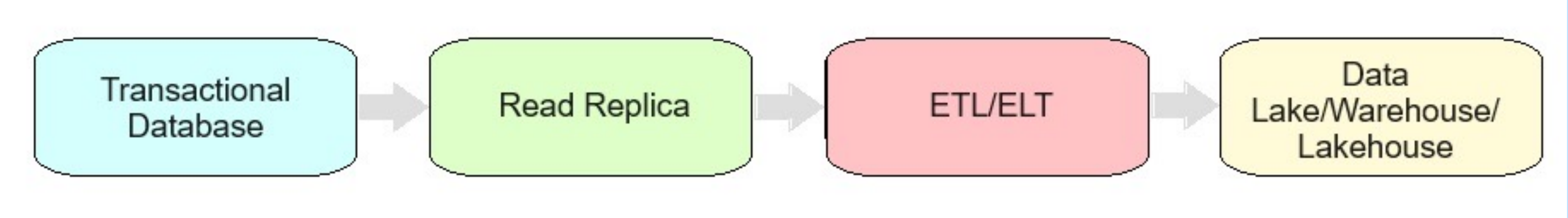
(Sometimes)



Exuberant VP/Manager → Hasty Development → Deployment

Challenges of Data & AI

- Data grows and evolves over time
- Data may be copied/moved/transformed many times prior to ML/AI
- Existing data quality is inherited by downstream processes
- AI processes are often quite authoritative



Transactional Database

Read Replica

ETL/ELT

Data Lake/Warehouse/ Lakehouse

How can we prevent bad data?

Validating OLTP/App/Edge Data

- This is *bad application data*!
- Its data journey *begins here*.
- Bad data from here will persist forever.

Validating OLTP/App/Edge Data

- Application constraints/restrictions
- Routine validation processes
- Unique indexes/constraints
- Foreign keys
- Check constraints
- Many of the above!

Validation (OLAP/Report/Analytic Data)

- Validate data after movement:
  - Data size (row count, byte count, etc…)
  - Validate values (uniqueness, NULL? invalid values?)
  - Missing data?
  - Duplicate data?
  - Edge-cases?

Validate BEFORE training models/RAG!

Validation (Releases)

- When code changes, validate impacted data
- Back up any data-to-be-modified!
- Without QA, existing data/validation may become incomplete/incorrect.

Names/Data Types Matter!

- Poorly named data elements can trick AI into making bad decisions.
- Poorly typed data can confuse AI.
- Check with original data source, if needed.

Integer named "Invoice"? What is it?
Datetime named "EntryTime"? Is it date/time or time?
Column named "IsDeleted": Should AI use this data?

Note: Training Data vs. RAG Data

- Both are important for a scalable AI system
- Both can experience bad data
- Bad training data = misbehaving model
- Bad RAG data = incorrect responses

How do we cheat bad data?

Prompt Engineering

- Improves AI interactions and output
- Delineates purpose
- Ensures relevance
- Refines inputs/outputs

Cannot prompt your way out of bad data!

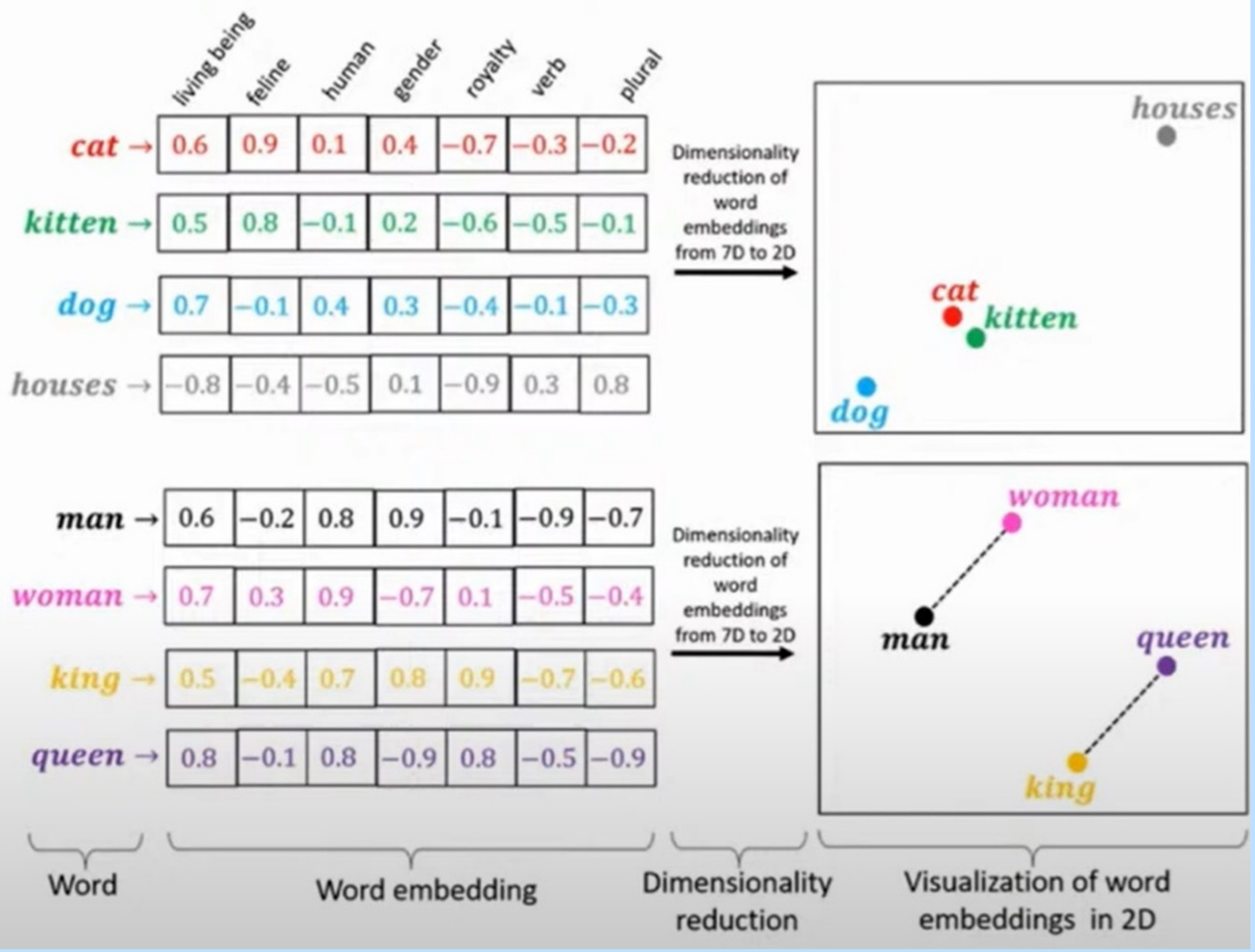
RAG

- Train model on one data set, use current/updated/relevant data for responses.
- Cannot be used to “fix” bad training
- Fixing/updating bad RAG data is not hard.

Semantic Search

- Breaks data into chunks, creating mathematical associations of similarity.
- Bad data will create bad associations that are hard to find and fix.
- Vectorization detail can be used, if needed, to reverse engineer bad results.

Semantic Search



	living being	feline	human	gender	royalty	verb	plural
cat →	0.6	0.9	0.1	0.4	−0.7	−0.3	−0.2
kitten →	0.5	0.8	−0.1	0.2	−0.6	−0.5	−0.1
dog →	0.7	−0.1	0.4	0.3	−0.4	−0.1	−0.3
houses →	−0.8	−0.4	−0.5	0.1	−0.9	0.3	0.8
man →	0.6	−0.2	0.8	0.9	−0.1	−0.9	−0.7
woman →	0.7	0.3	0.9	−0.7	0.1	−0.5	−0.4
king →	0.5	−0.4	0.7	0.8	0.9	−0.7	−0.6
queen →	0.8	−0.1	0.8	−0.9	0.8	−0.5	−0.9

Dimensionality reduction of word embeddings from 7D to 2D
houses
cat
kitten
dog
Dimensionality reduction of word embeddings from 7D to 2D
woman
man
queen
king
Word
Word embedding
Dimensionality reduction
Visualization of word embeddings in 2D

Fine-Tuning

- Allows a model to be tailored to a more specific use-case.
- Requires significant time/effort to implement.
- Is NOT a solution to bad data anywhere else.

Synthetic Data

- Artificially-generated data
- Mimics real-world data
  - Same mathematical properties
  - Different information
  - Can remove PII/protected data
- Hard to generate without bias/replication/bad data
- Need to prove that new data is ***valid*** AND ***unique***.
- Cannot dilute bad data with good synthetic data
- Must be validated.

Unlearning Data Can Harm Models

- Unlearning involves forcing a model to forget specific information.
    - PII
    - Bad data
    - Copyrighted material
- Current unlearning methods are not mature enough to manage data loss without retraining.

Intelligent Capture

- Using existing unstructured data:
  - Reads
  - Interprets
  - Generates Insights
  - Writes new data
- Bad data is magnified via this process!

Test carefully before implementing

Feedback Loops

- AI can create/modify insights, data, responses, and content
- Beware this new data becoming part of existing data
- Is this intentional!?
- Feedback loops can amplify some results or diminish others.

Use caution when adding new data to existing data sets!

Conclusion

- Bad data is most easily resolved at its source.
- AI model manipulation is not a substitute for good data.
- Carefully test models and ensure that invalid responses are identified and resolved by finding their origin.

Questions? Thank You!

Find me here:

- Ed Pollack | LinkedIn
- Edward Pollack | Most Valuable Professionals
- https://sessionize.com/edward-pollack/

Find my content here:

- EdwardPollack (Ed Pollack) (github.com)
- Edward Pollack, Author at Simple Talk (red-gate.com)
- Ed Pollack, Author at SQL Shack - articles about database auditing, server performance, data recovery, and more