



Automating ADaM Dataset Generation via Definitive Specification Writing with A Domain-Specific Language

@ PHUSE EU Connect 23

Kang Zhao

Biometrics, R&D China

AstraZeneca

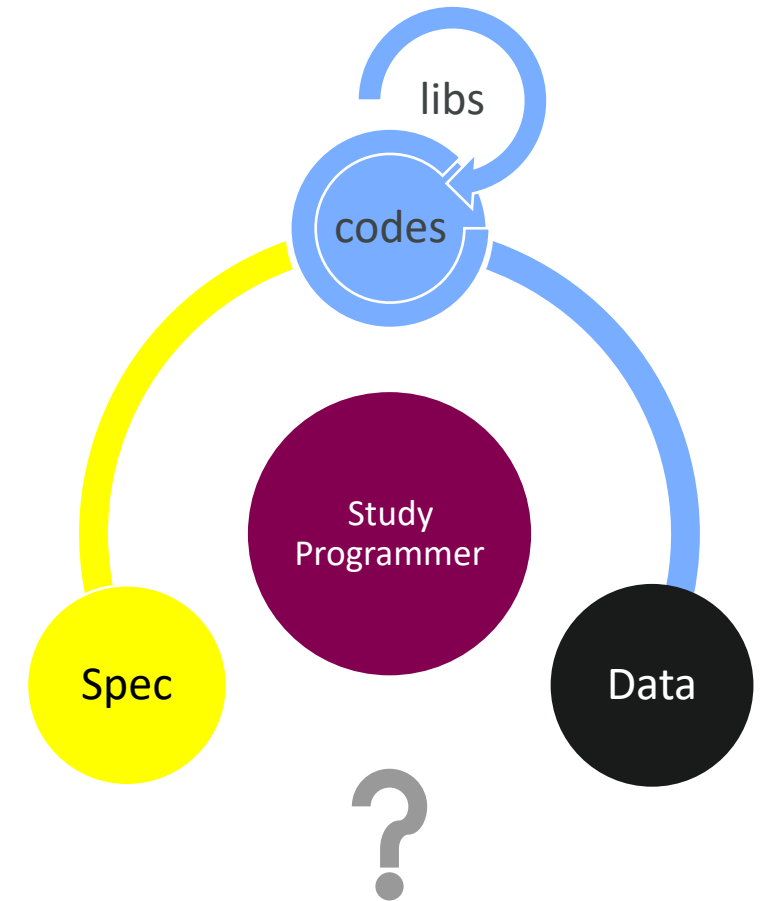
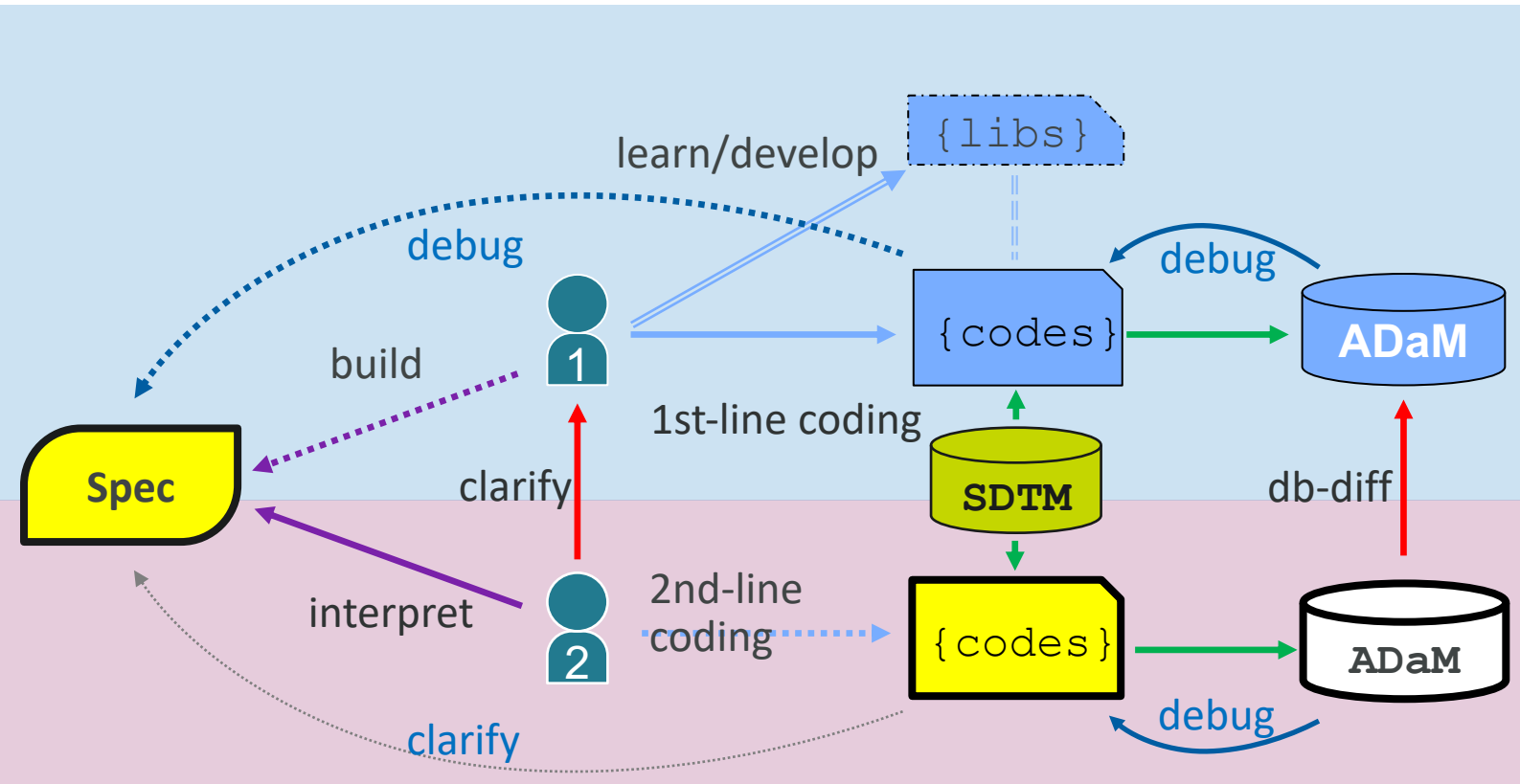


Agenda

1. ADaM dataset development process
2. Process optimizations and effects
3. A derivation rule example
4. Spec-writing, dependency analysis, and orchestration
5. Q&A



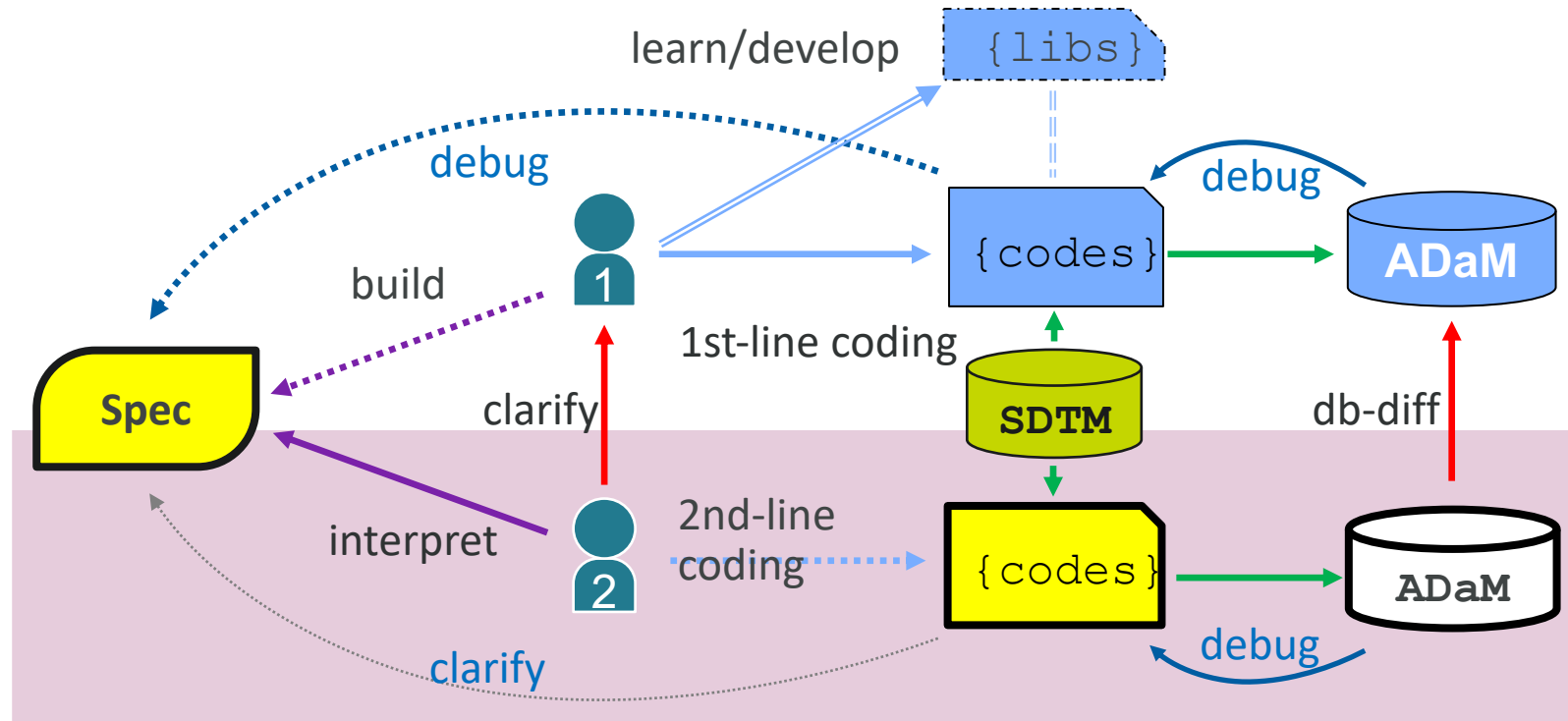
ADaM Dataset Development



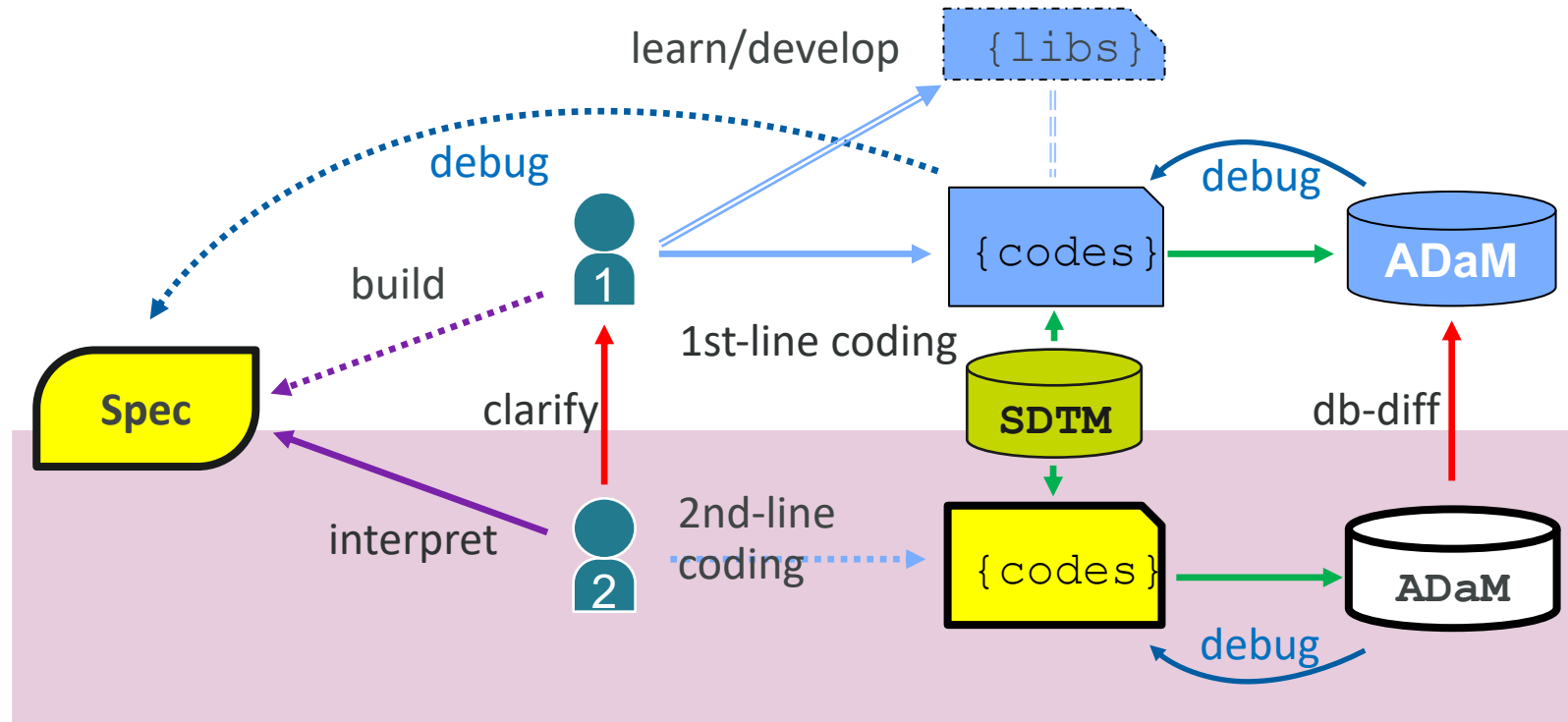
Can spec and data be directly linked



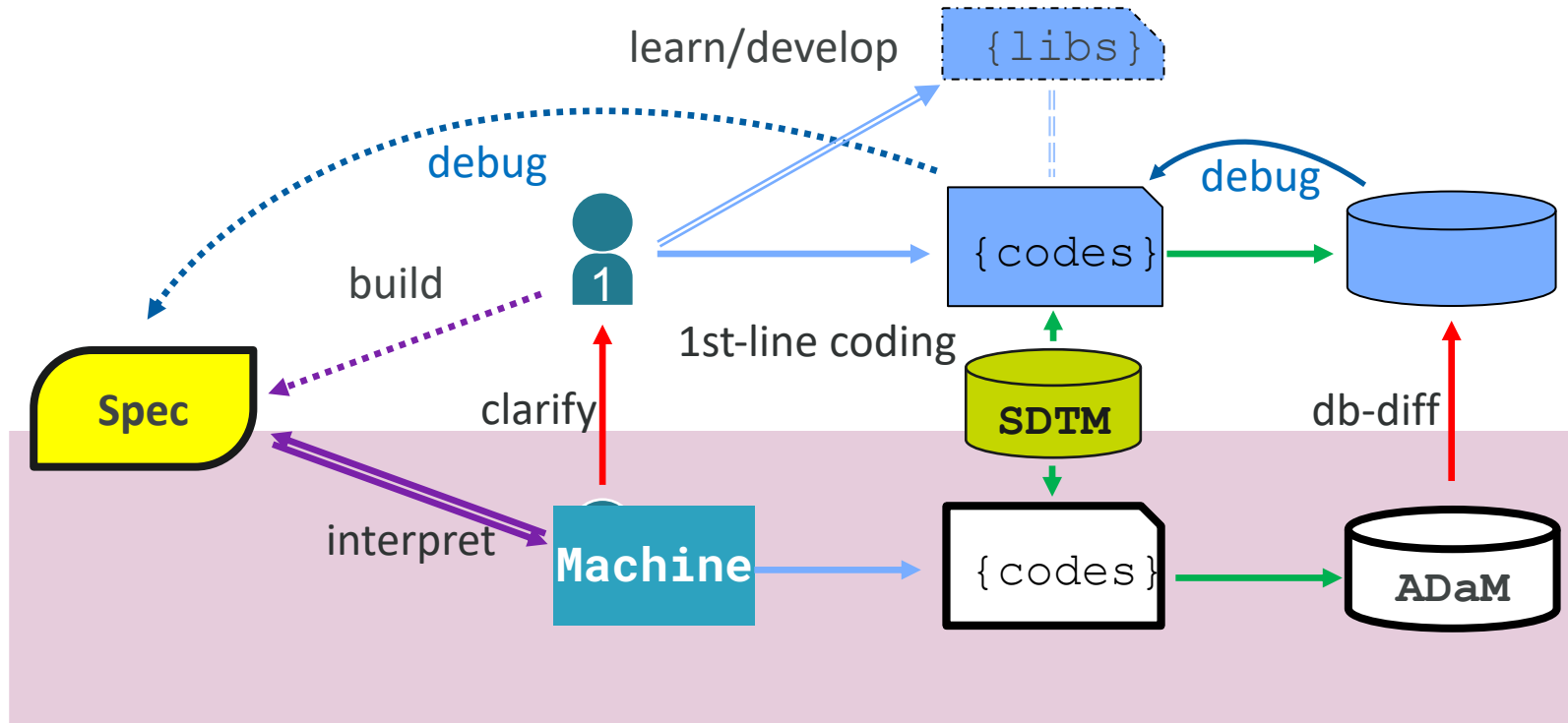
Change #1 We want to have spec of "good" quality



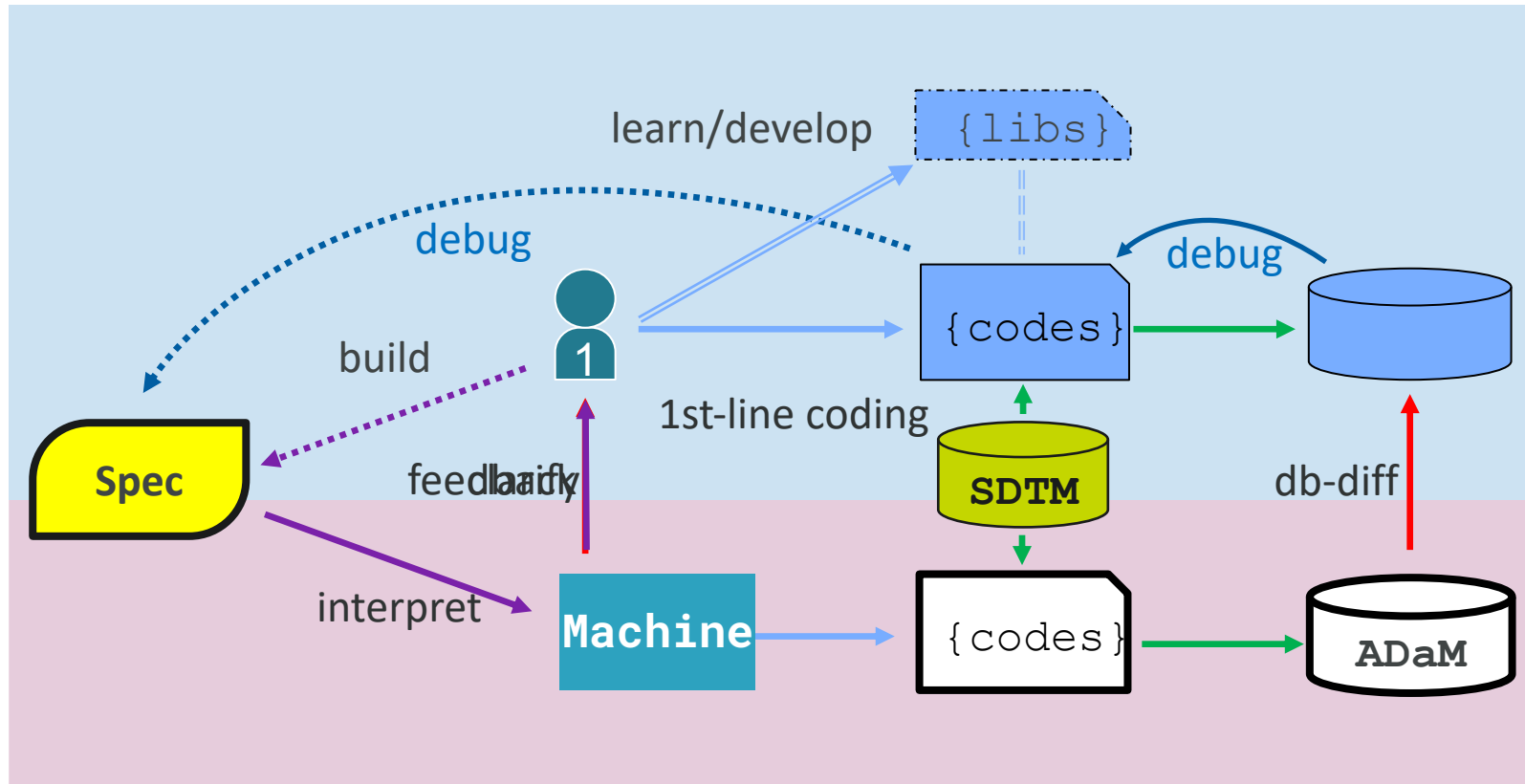
Change #2 We want to remove duplicated coding



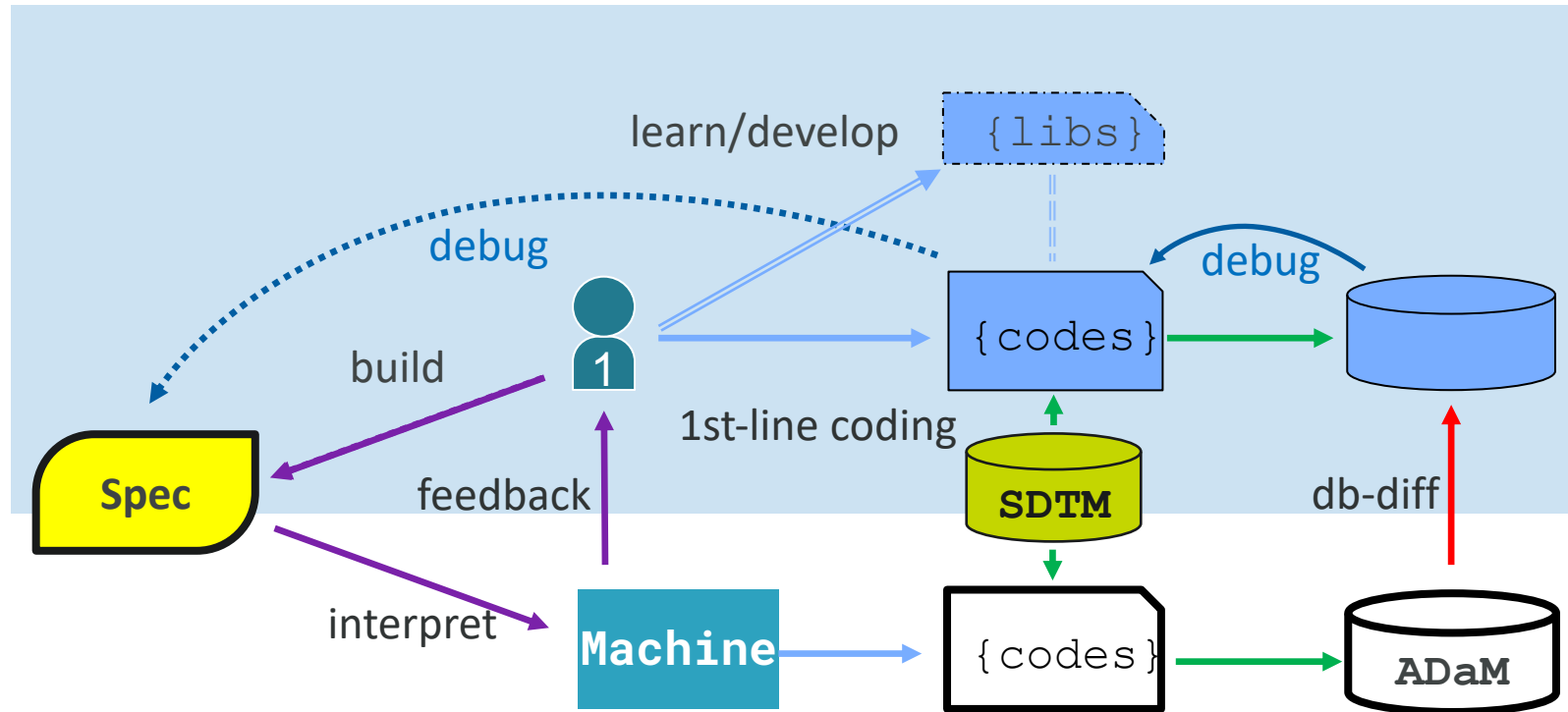
Change #3 Use spec as single source-of-truth



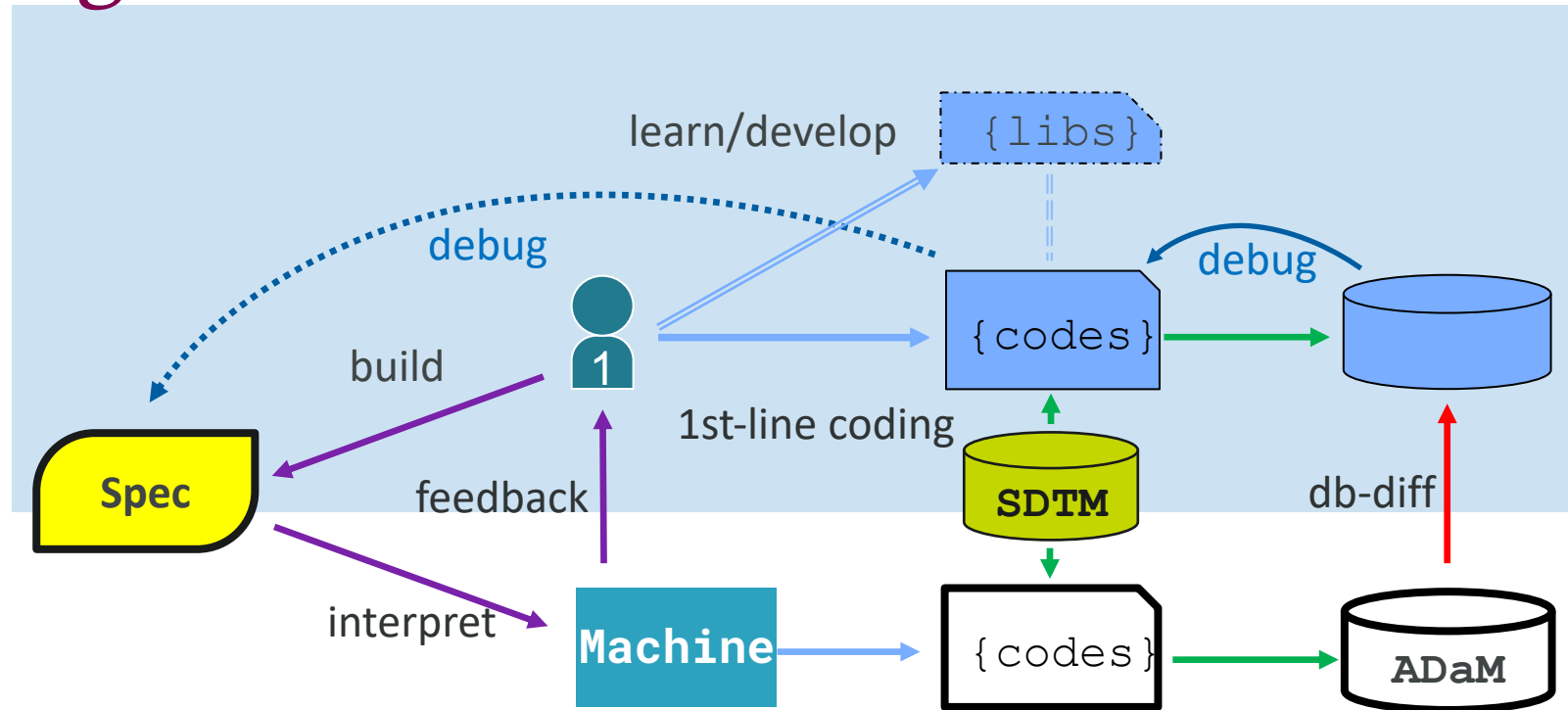
Effect #1 Feedback on the spec quality



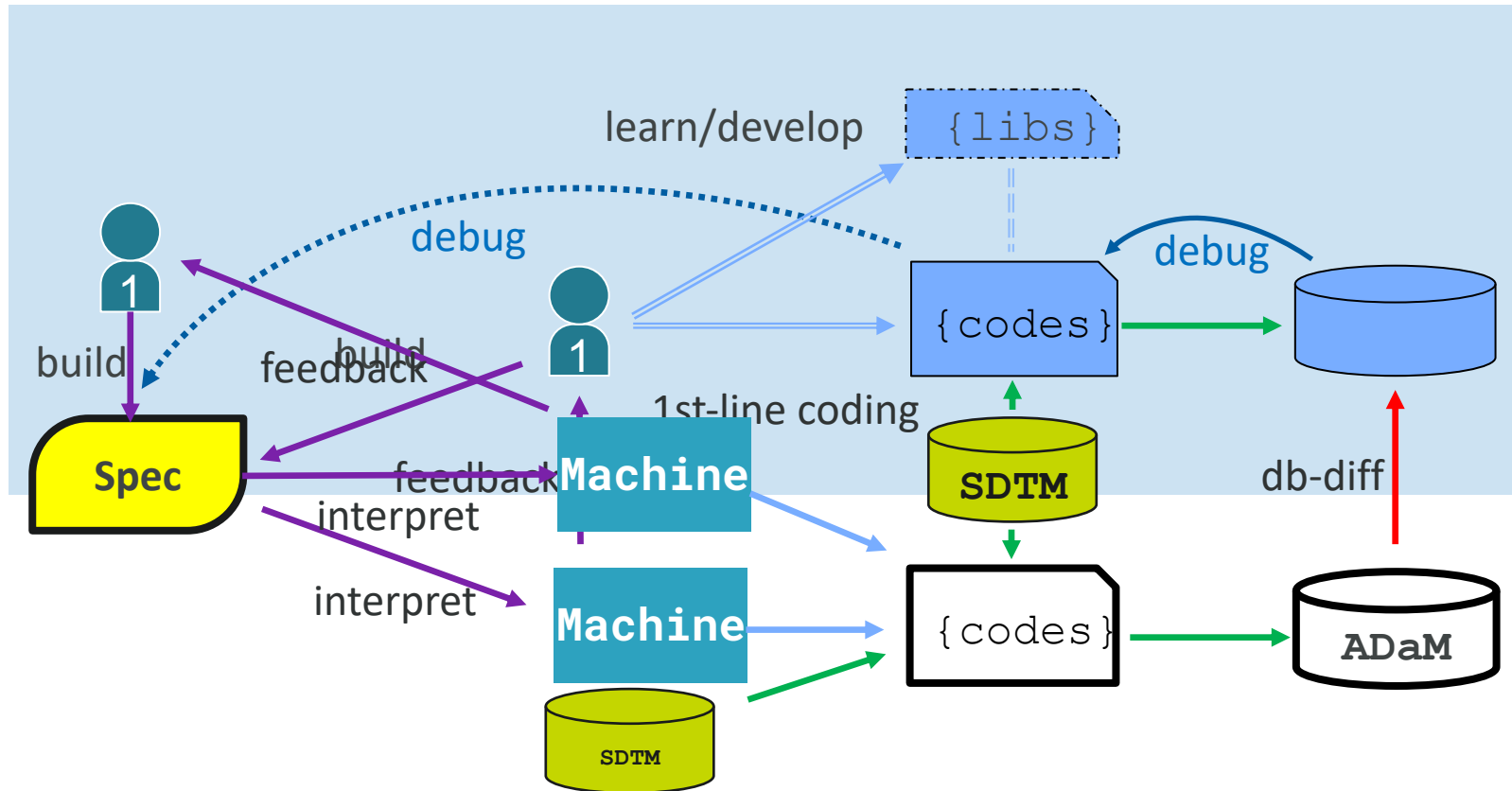
Change #4 More focused on spec-writing



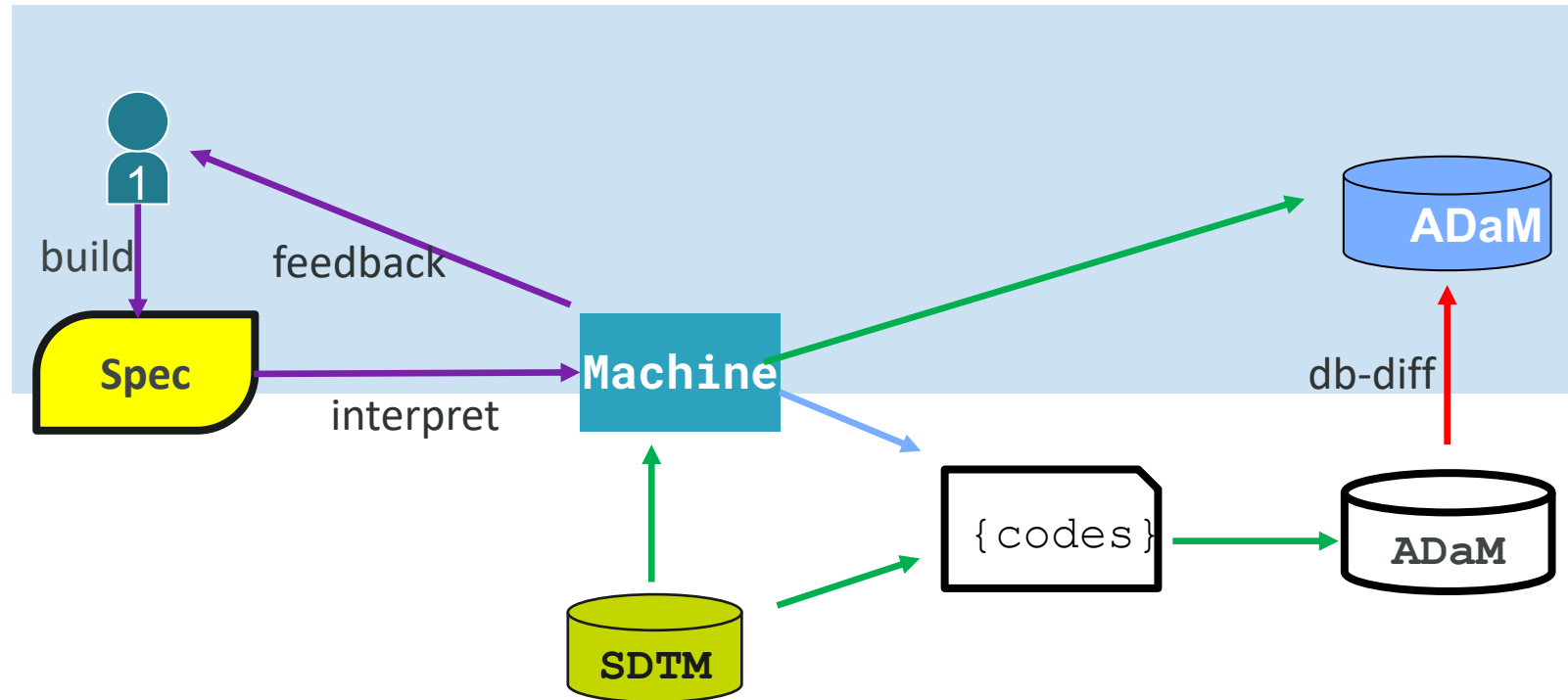
Effect #2 Making 1st-line coding a duplicated effort to spec-writing



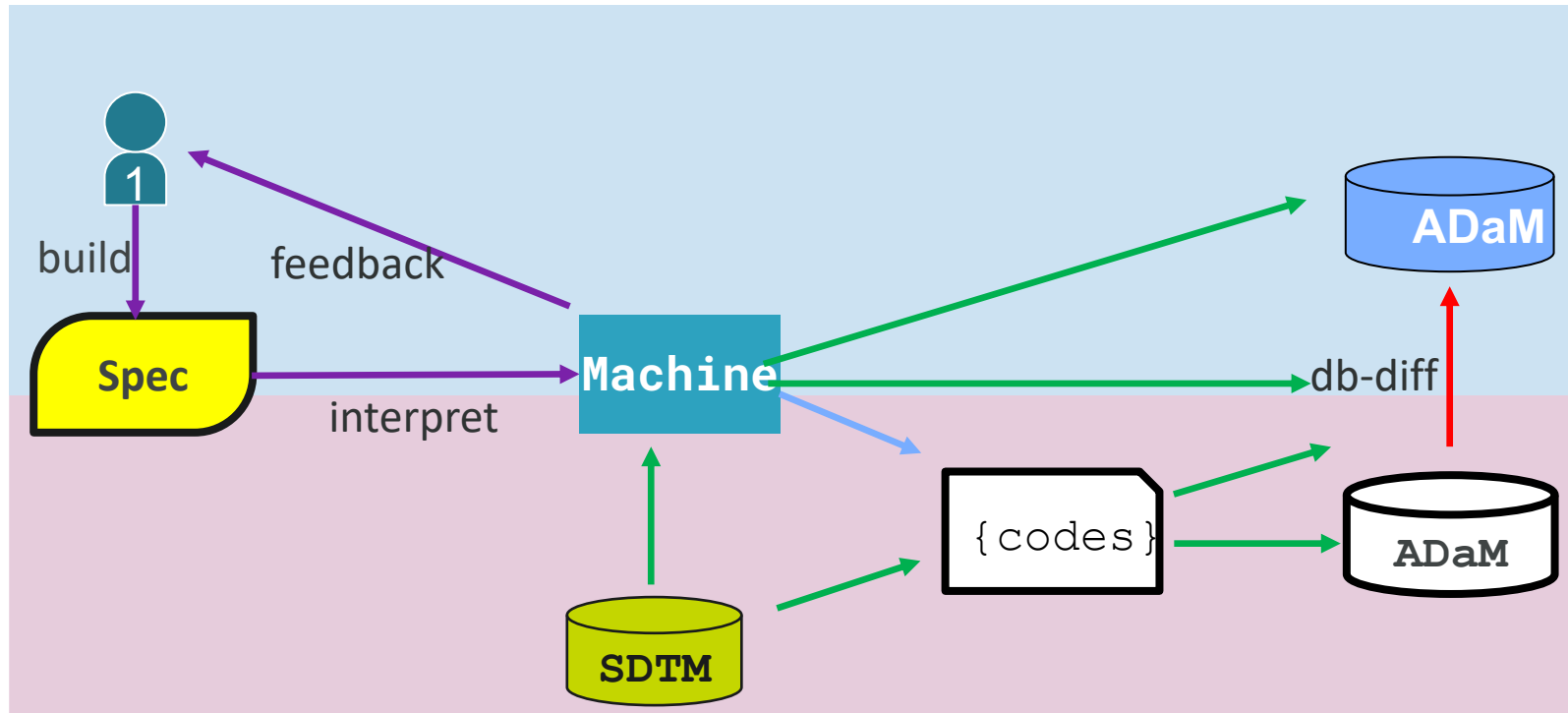
Change #5 Remove 1st-line coding



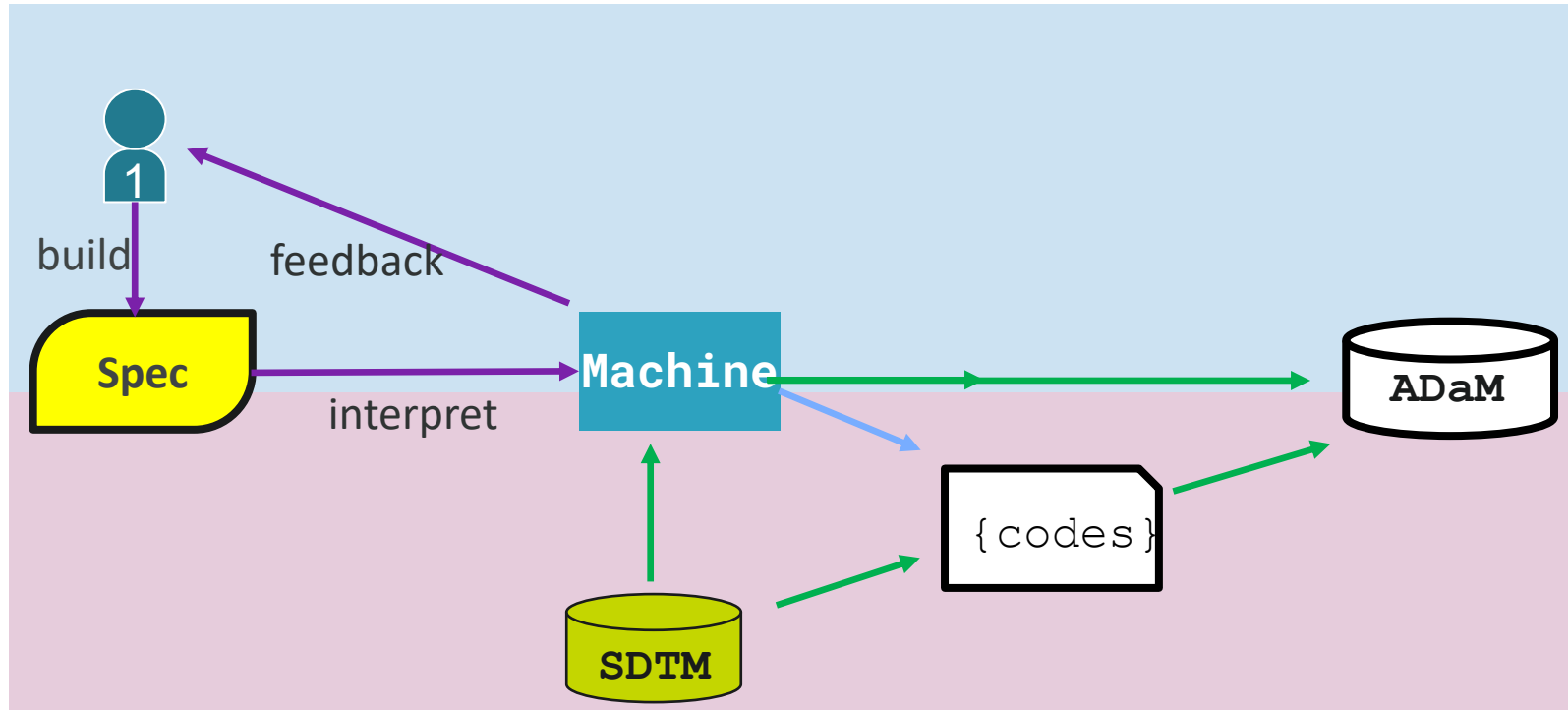
Change #6 Introducing datagen: from spec to dataset



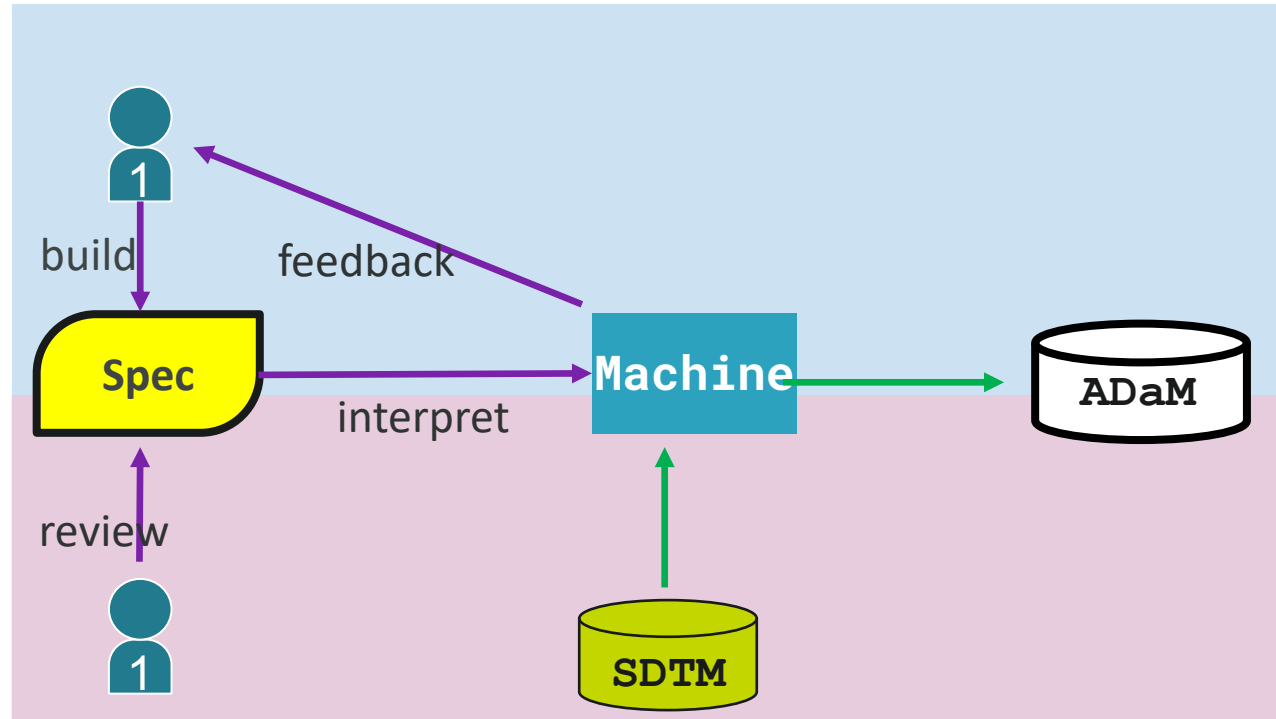
Effect #3 db-diff is internalized



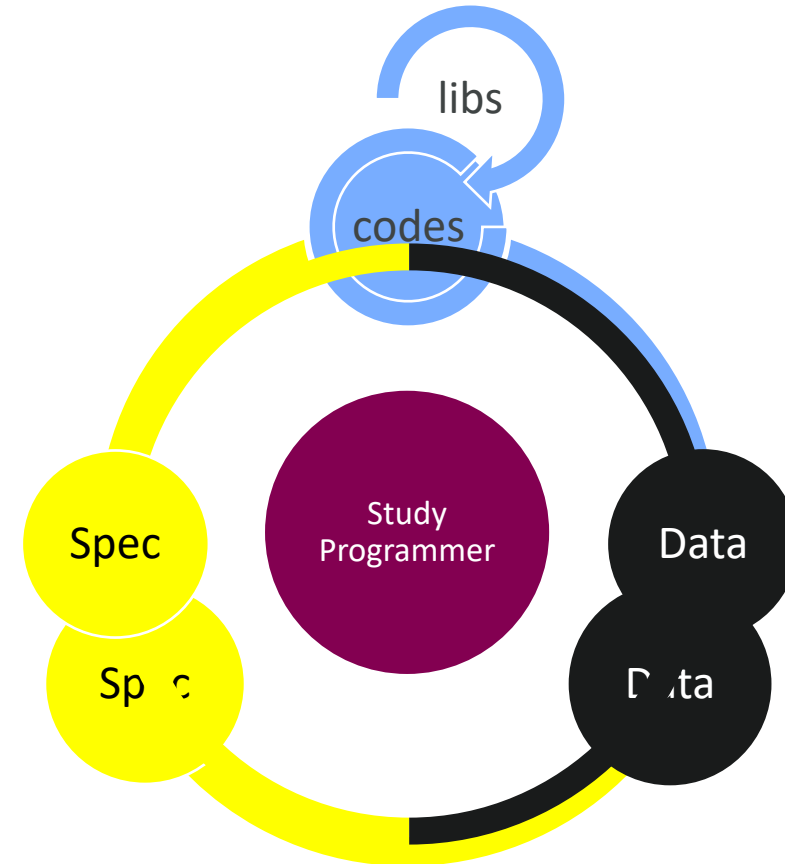
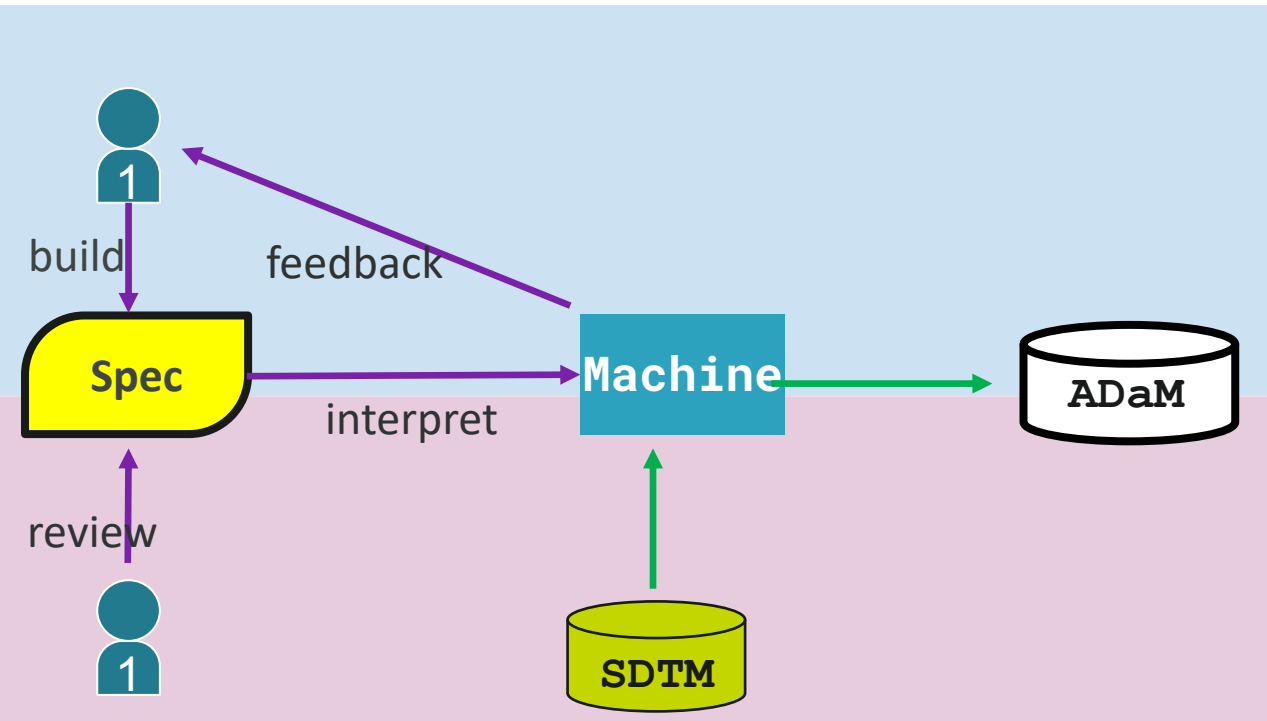
Effect #4 Separation of Concerns



Change #7 More QC on specs



Effect #5 Spec quality is improved



Example

Domain	Variable	Derivation
ADSL	DTHDTC	DM.DTHDTC
ADSL	DTHDTN	Numeric value of ADSL.DTHDTC. If DTHDTC is partial or missing when DM.DTHFL = 'Y', 1. Set to max(LSTALVDT+1, imputed date) if max(LSTALVDT+1, imputed date) <= DCO; 2. set to missing otherwise; Impute date as follows: For missing day only - using the 1st of the month; For missing day and month - using the 1st of January.

Table–1. Example derivation rule for a target variable

1	if DM.DTHFL = 'Y'
2	and the latest of
3	[ADSL.LSTALVDT + 1 day, ADSL.DTHDTC]
4	is not later than 2023-01-31,
5	then set to the latest of
6	[ADSL.LSTALVDT + 1 day, ADSL.DTHDTC]
7	else set to missing.
	Preprocessing Imputation: - if month of ADSL.DTHDTC is missing, then impute with January. - if day of ADSL.DTHDTC is missing, then impute with 01.

Table–2. A definitive way of expressing the derivation rule



Alternatives

```
1 if DM.DTHFL = 'Y'
2   and the latest of
3     [ADSL.LSTALVDT + 1 day, ADSL.DTHDTC]
4     is not later than 2023-01-31,
5 then set to the latest of
6     [ADSL.LSTALVDT + 1 day, ADSL.DTHDTC]
7 else set to missing.
```

Preprocessing Imputation:

- if month of ADSL.DTHDTC is missing, then impute with January.
- if day of ADSL.DTHDTC is missing, then impute with 01.

Table-2. A definitive way of expressing the derivation rule

```
1 if DM.DTHFL = 'Y'
2 then set to the latest of
3     [ADSL.LSTALVDT + 1 day, ADSL.DTHDTC].
4
5 if ADSL.DTHDTN is after 2023-01-31,
6 then set to missing.
```

Preprocessing Imputation:

- if month of ADSL.DTHDTC is missing, then impute with January.
- if day of ADSL.DTHDTC is missing, then impute with 01.

Table-3. An Alternative Derivation



Spec-writing assisted with syntax highlighting

Domain	Variable	Derivation
ADSL	DTHDTC	DM.DTHDTC
ADSL	DTHDTN	Numeric value of ADSL.DTHDTC. If DTHDTC is partial or missing when DM.DTHFL = 'Y', 1. Set to max(LSTALVDT+1, imputed date) if max(LSTALVDT+1, imputed date) <= DCO; 2. set to missing otherwise; Impute date as follows: For missing day only - using the 1st of the month; For missing day and month - using the 1st of January.

Table–1. Example derivation rule for a target variable

```
if DM.DTHFL = 'Y'
then set to the latest of
    [ADSL.LSTALVDT + 1 day, ADSL.DTHDTC].

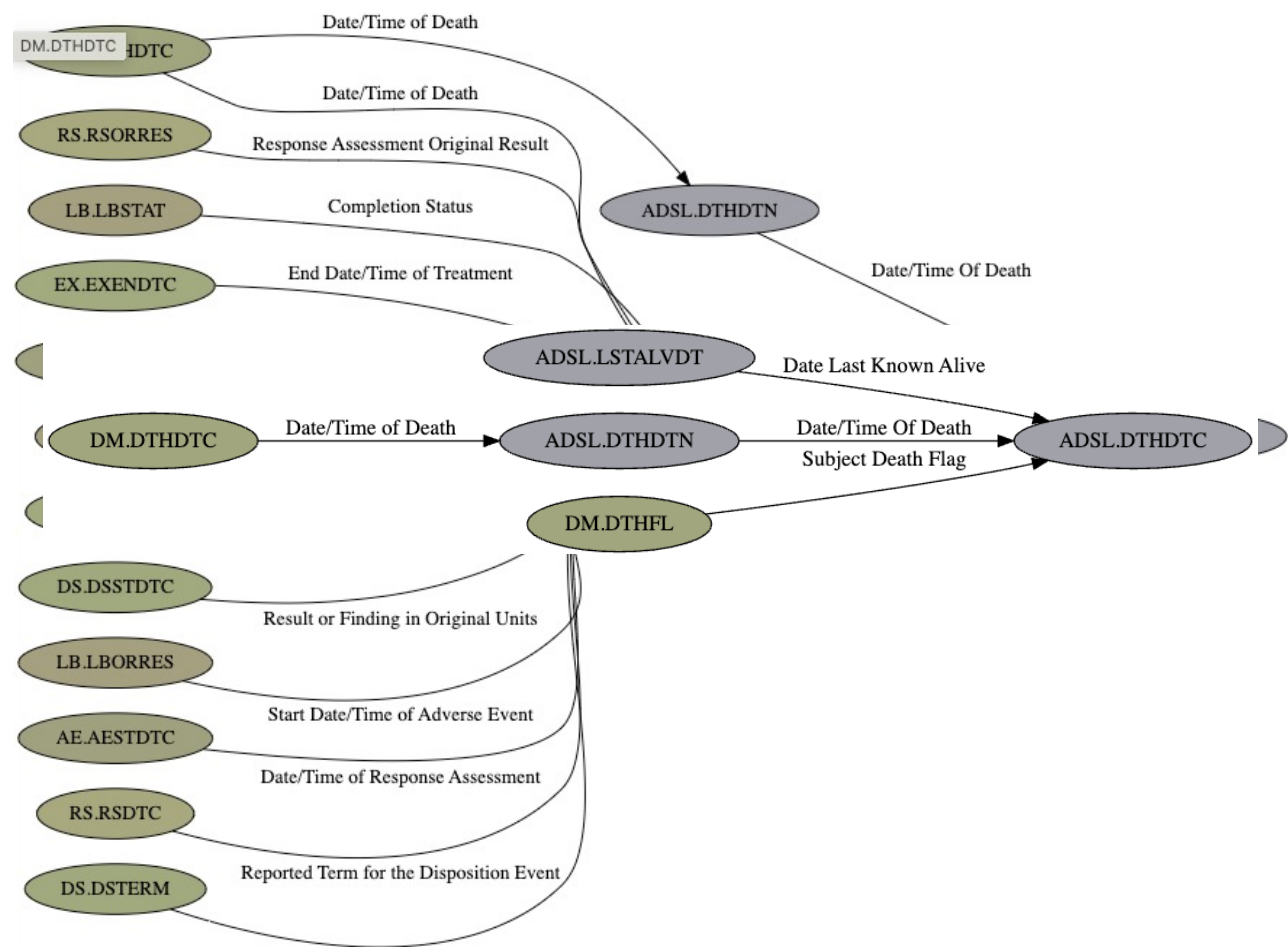
if ADSL.DTHDTN is after 2023-01-31,
then set to missing.

Preprocessing Imputation:
= if month of ADSL.DTHDTC is missing, then
  impute with January.
= if day of ADSL.DTHDTC is missing, then
  impute with 01.
```

Table–4. A definitive derivation rule with syntax highlighting



Dependency and Traceability



Figure–1. Backward tracing

```
1  if DM.DTHFL = 'Y'
2  then set to the latest of
3      [ADSL.LSTALVDT + 1 day, ADSL.DTHDTC].
4
5  if ADSL.DTHDTN is after 2023-01-31,
6  then set to missing.
```

Preprocessing Imputation:

- = if month of ADSL.DTHDTC is missing, then impute with January.
- = if day of ADSL.DTHDTC is missing, then impute with 01.

Table–4. A definitive derivation rule



Orchestration

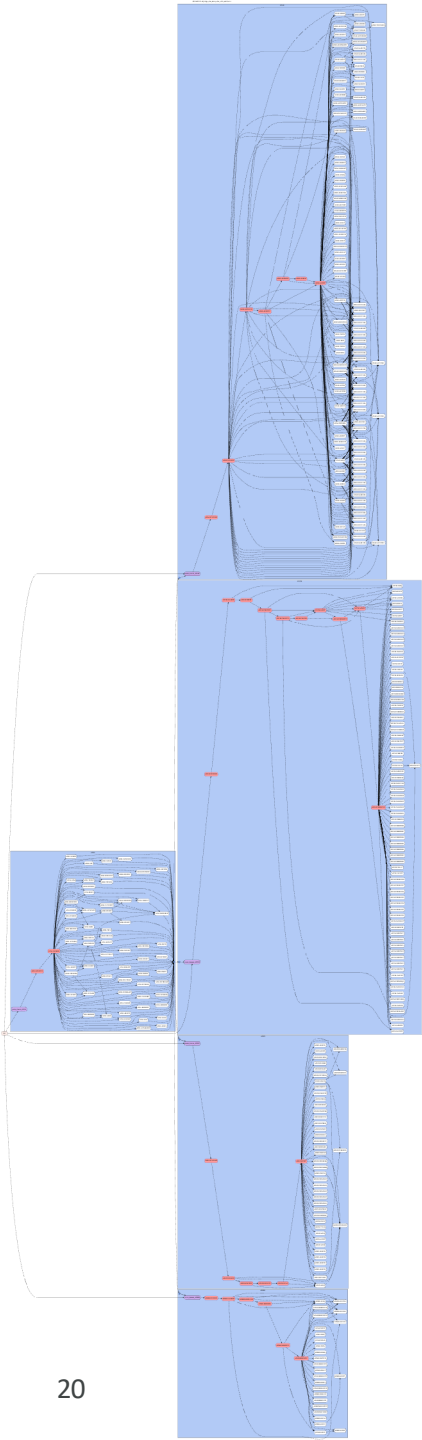


Figure-2.
Orchestrated
execution path
at domain level

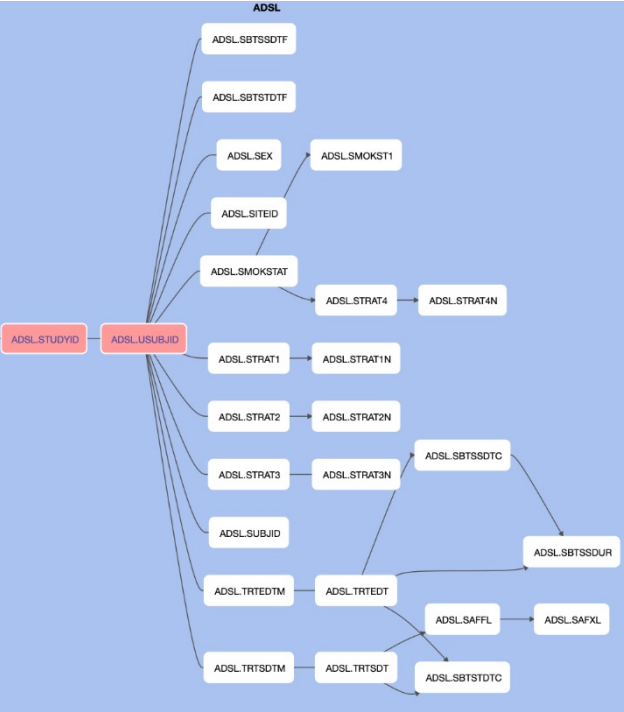


Figure-3.
Orchestrated
execution path
inside ADSL

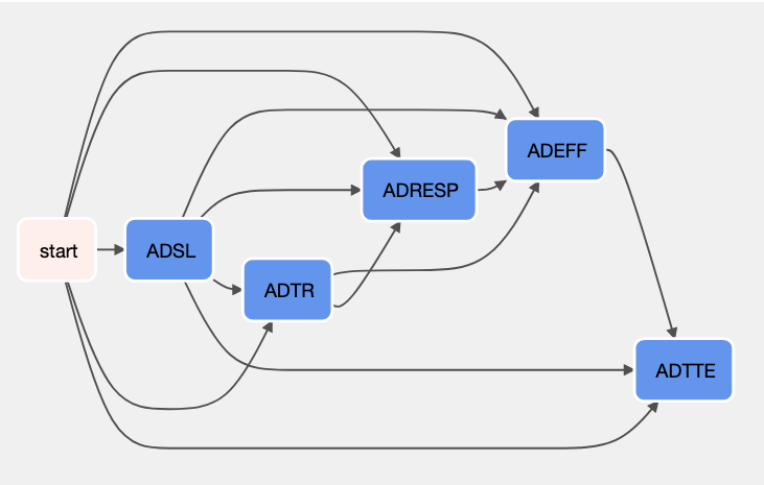


Figure-4.
Birdseye view
over domains

Paper Number DH12

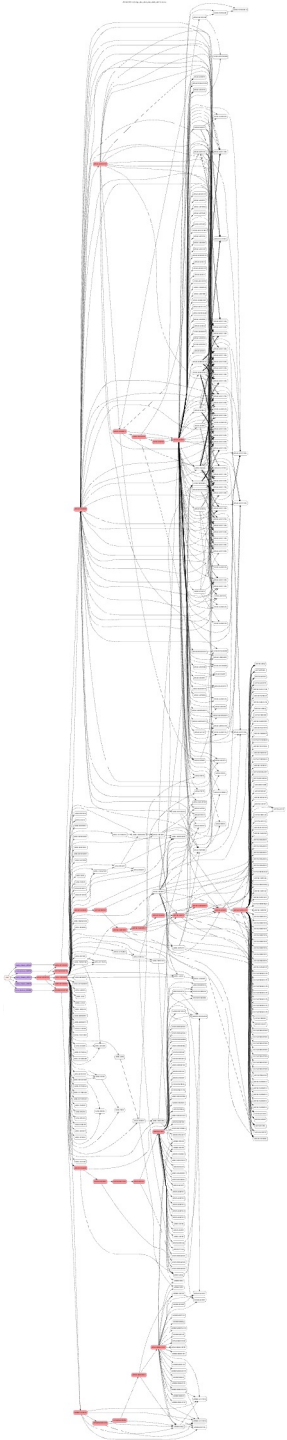


Figure-5.
Orchestrated
execution path
at variable level



Thank you!



Q&A

kang.zhao@astrazeneca.com

