Processing data from a Web API - a step by step guide

Damon Douglas / damondouglas@

BEAM SUMMIT

September 4-5, 2024

Sunnyvale, CA. USA

Is there an example?

Yes:

https://beam.apache.org/documentation/io/built-in/webapis/

And

https://github.com/apache/beam/tree/master/examples/java/webapis

BEAM SUMMIT

Prerequisites

Understand Beam Programming Guide:

- Basics - PTransform, PCollection
- DoFn - ProcessElement, Setup, Teardown

https://beam.apache.org/documentation/programming-guide

BEAM SUMMIT

Web API processing with Beam challenging

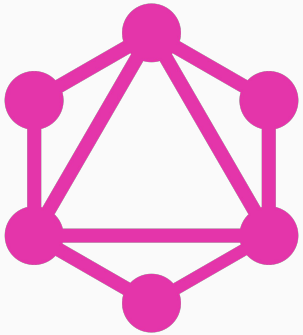
- Beam designed for large scale parallelized workloads

- (most) Web APIs designed for application workloads

Web API landscape

gRPC

OPENAPI INITIATIVE



GraphQL

BEAM SUMMIT

Web API processing concerns

- Rate limits / Quota
- Client / Server contract adherence
- API SDK serializability
- Beam Schema compatibility
- Limited Data per API endpoint

BEAM SUMMIT

Talk agenda

1. Describe Use Case
2. Simple HTTP GET
3. Chain API requests and responses
4. API Client Setup/Teardown
5. Additional features - caching and throttling
6. Conclusion

BEAM SUMMIT

Describe Use Case

BEAM SUMMIT

Trade off selecting a use case

Trade off

Irrelevant
Surprising when applied

Too much domain knowledge

What do we want to achieve?

Data sources: National Institute of Health (NIH) Web APIs

- RxClass
- RxNorm
- DailyMed

Final sink: Gemini Vertex AI API

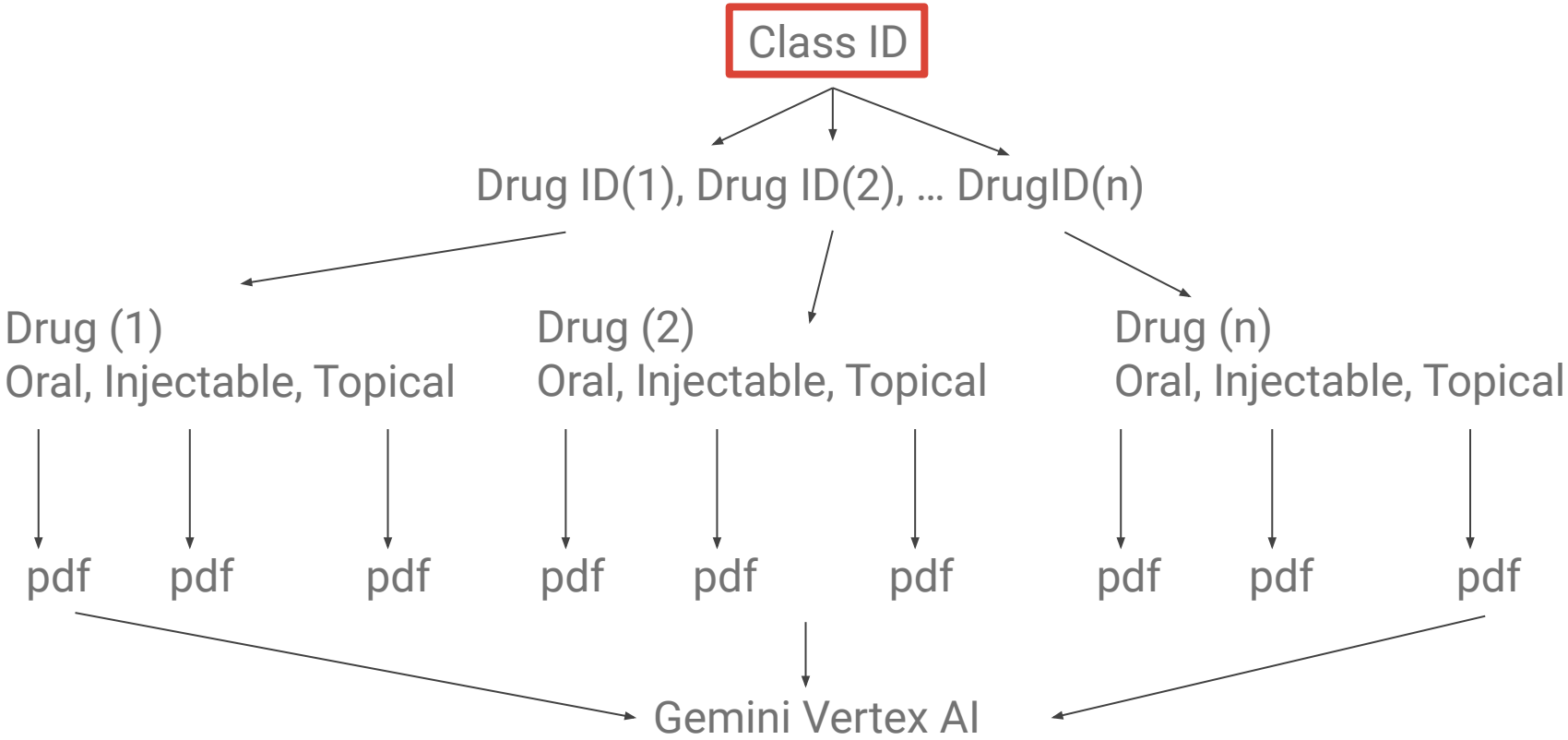
BEAM SUMMIT

Why NIH Web APIs useful for this talk?

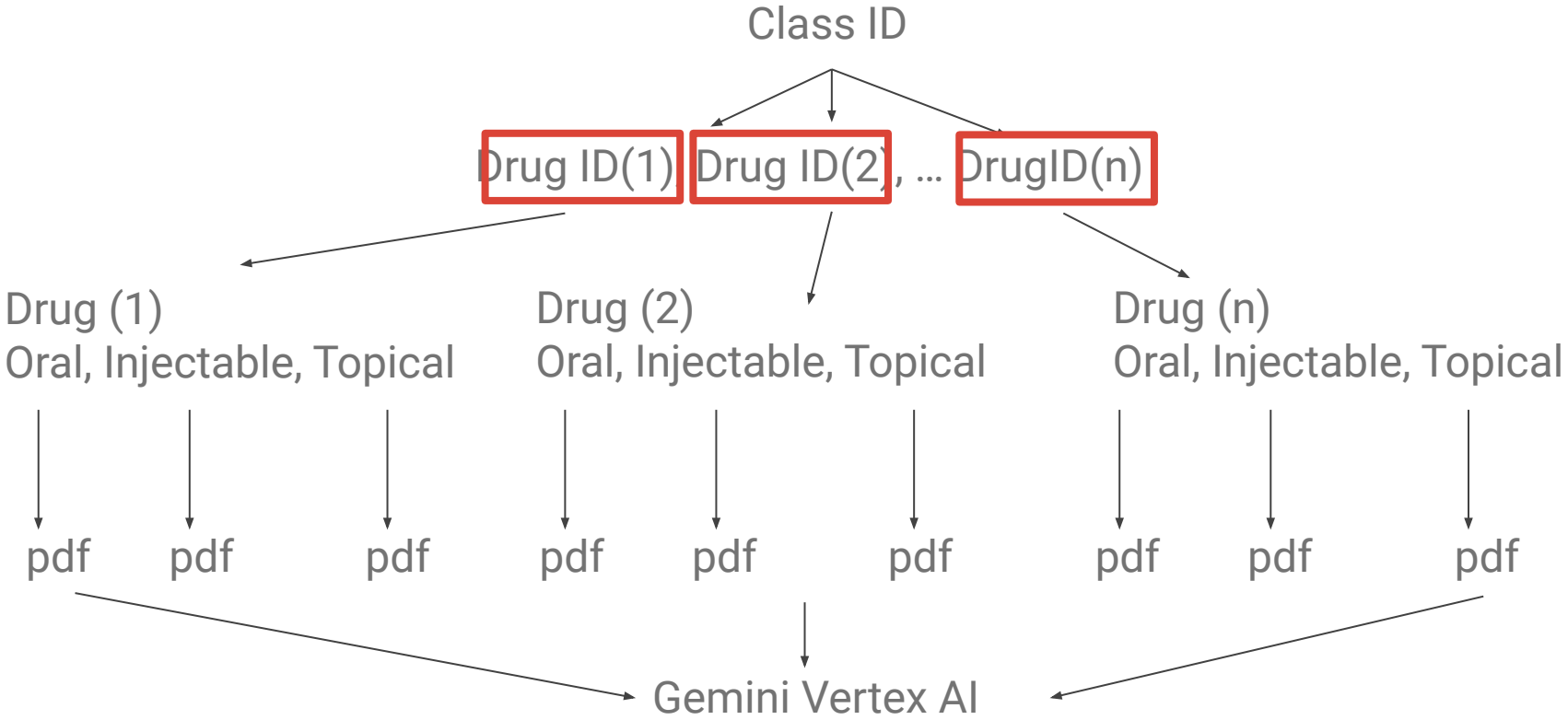
- Simple to invoke single endpoints
- Some complexity to relate responses between endpoints
- No SDK

BEAM SUMMIT

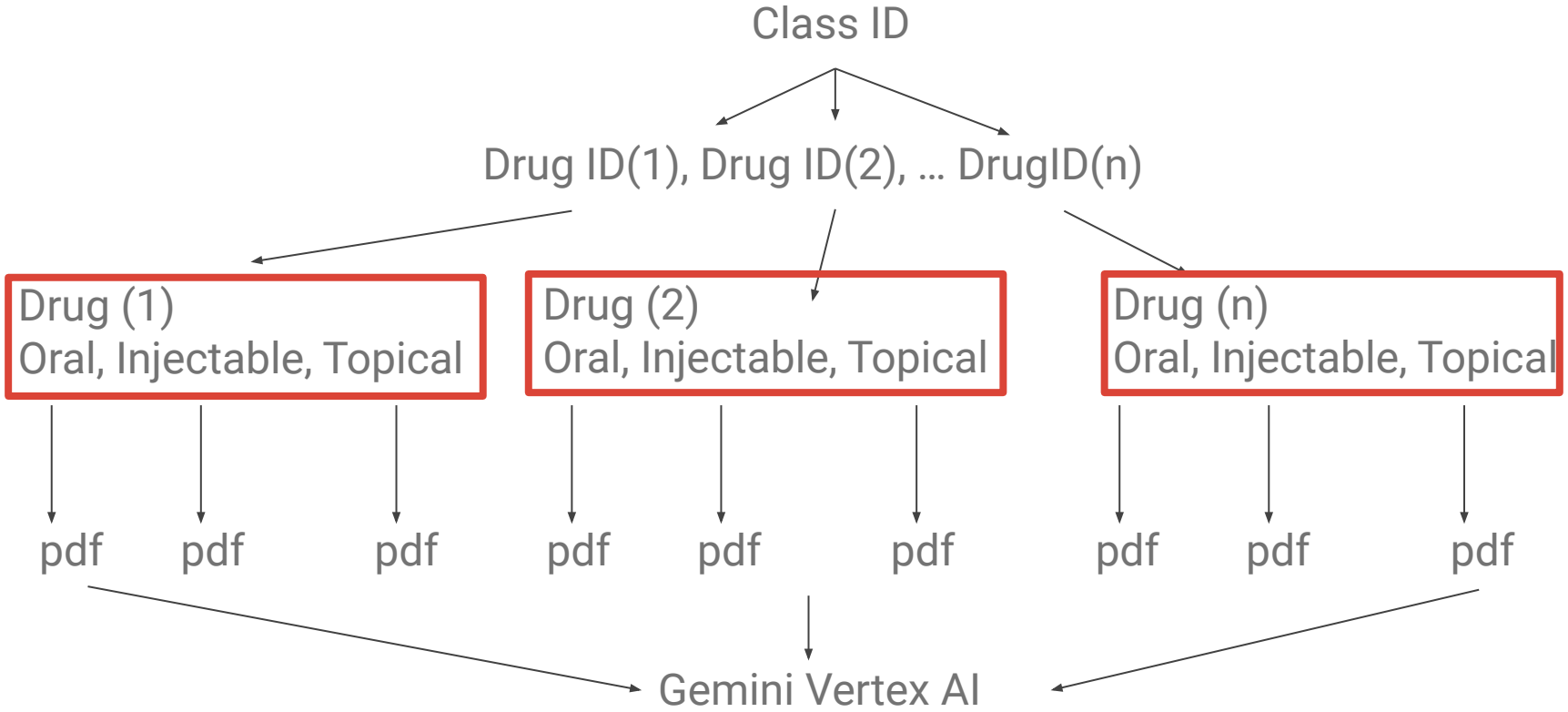
The Pipeline



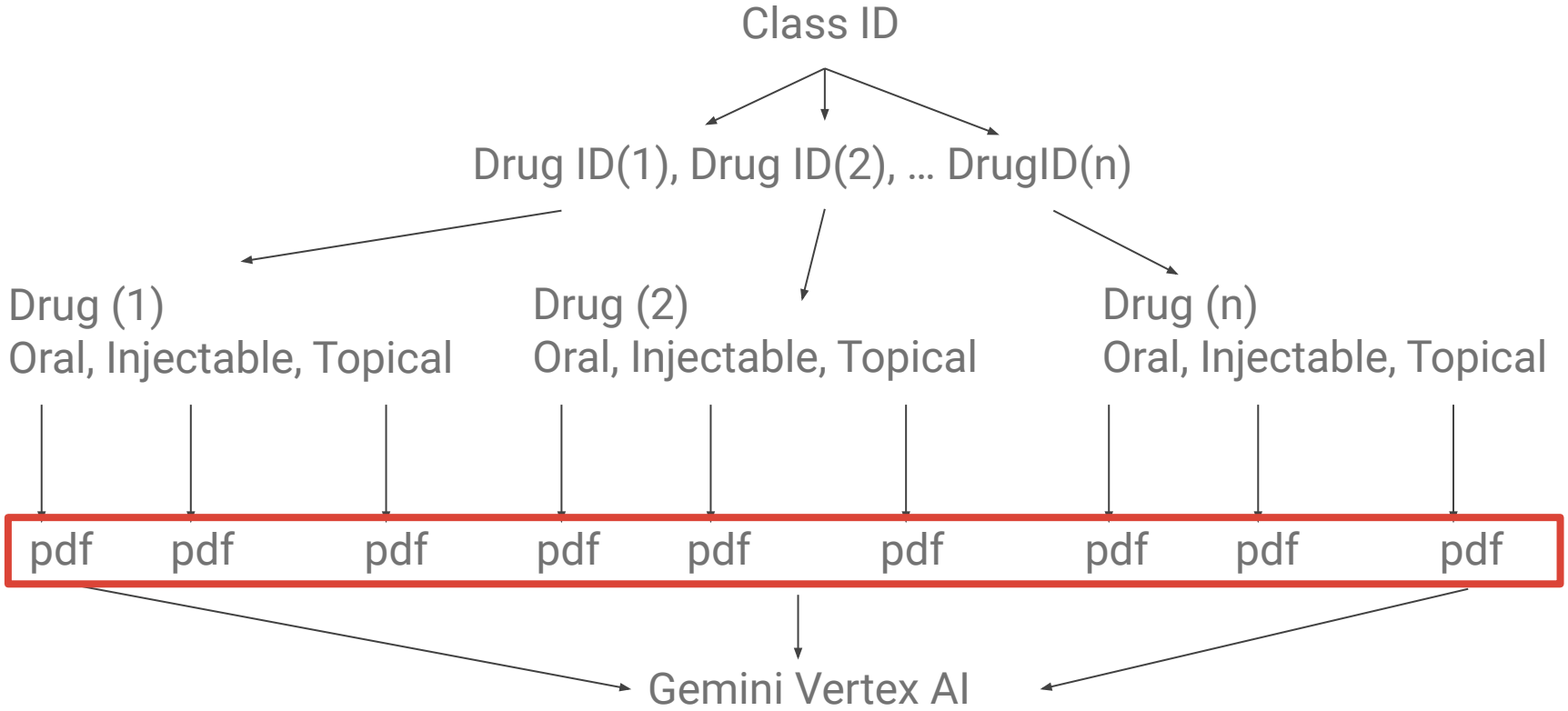
The Pipeline



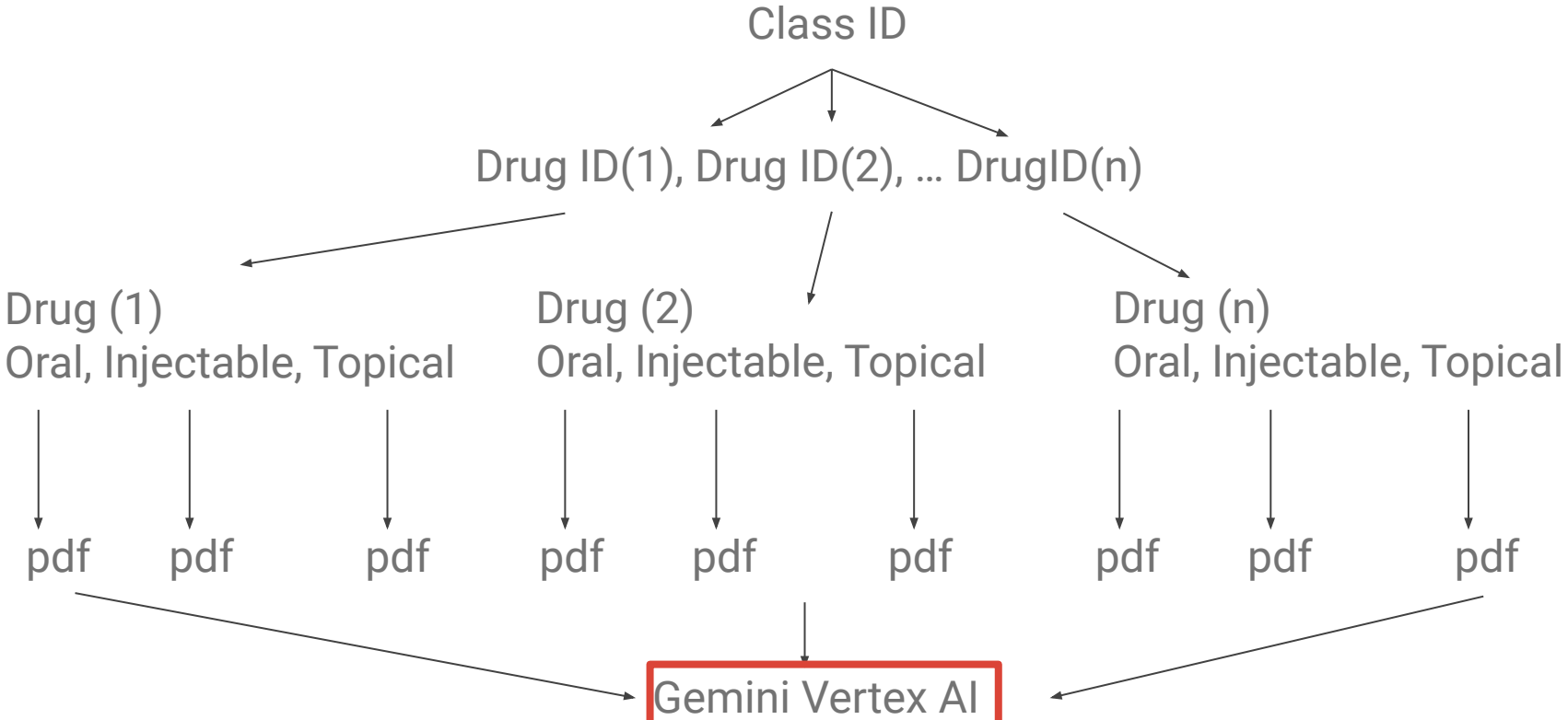
The Pipeline



The Pipeline



The Pipeline

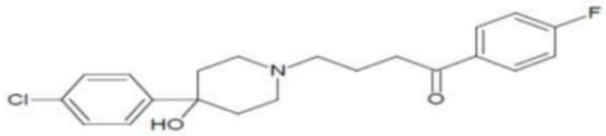


HALOPERIDOL- haloperidol tablet
Lifestar Pharma LLC

Haloperidol Tablets, USP
Rx Only

DESCRIPTION

Haloperidol, USP is the first of the butyrophenone series of major tranquilizers. The chemical designation is 4-[4-(*p*-chloro-phenyl)-4-hydroxypiperidino]-4'-fluorobutyrophenone and it has the following structural formula:



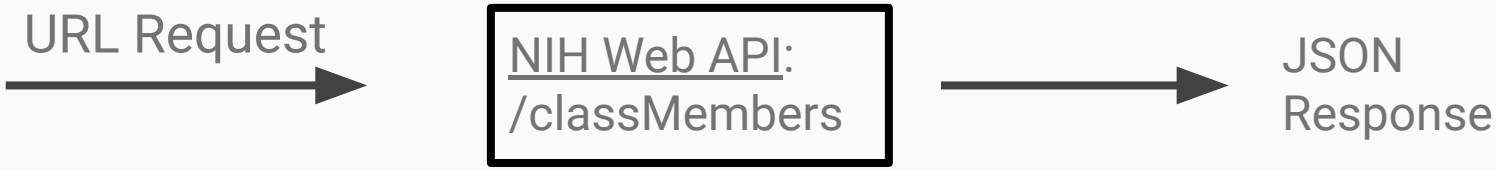
$C_{21}H_{23}ClFNO_2$ 375.86

Haloperidol, USP is supplied as tablets for oral administration containing 0.5 mg, 1 mg, 2 mg, 5 mg, 10 mg or 20 mg of haloperidol, USP and contains the following inactive ingredients: colloidal silicon dioxide, lactose monohydrate, magnesium stearate, microcrystalline cellulose and pregelatinized starch (corn). In addition, the 1 mg, 2 mg, 5 mg, and 10 mg tablets also contains D&C Yellow #10 Aluminum Lake. The 10 mg and 20 mg tablets also contains FD&C Blue #1 Aluminum Lake, 2 mg and 5 mg tablets also contains FD&C Red #40 Aluminum Lake, and 1 mg, 2 mg, 5 mg and 20 mg tablets also contains FD&C Yellow #6 Aluminum Lake.

Simple HTTP GET

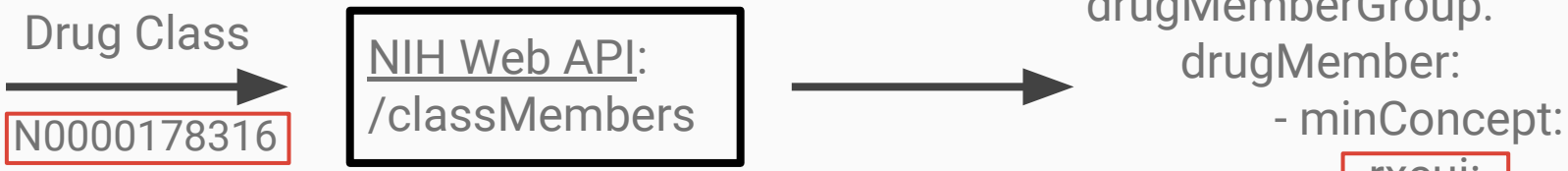
BEAM SUMMIT

Simple HTTP GET



https://lhncbc.nlm.nih.gov/RxNav/APIs/api-RxClass.getClassMembers.html

Simple HTTP GET



Point of this slide is to show a common problem with Web APIs:

- We start with one ID i.e. N0000178316
- The data we want, **rxcui**, is deep in the payload

https://lhncbc.nlm.nih.gov/RxNav/APIs/api-RxClass.getClassMembers.html

Simple HTTP GET

Tasks:

1. Model the response payload
2. Implement an interface

BEAM SUMMIT

Simple HTTP GET - 1. Model response

```
@DefaultSchema(AutoValueSchema.class)
@AutoValue
abstract static class GetClassMembersResponse implements Serializable {

  abstract DrugMemberGroup getDrugMemberGroup();

  @AutoValue.Builder
  abstract static class Builder {
    abstract Builder setDrugMemberGroup(DrugMemberGroup drugMemberGroup);

    abstract GetClassMembersResponse build();
  }
}
```

Simple HTTP GET - 2. Implement interface

```
interface Caller<RequestT, ResponseT> {

    ResponseT call(RequestT request);

}
```

See Caller interface Javadoc:
https://beam.apache.org/releases/javadoc/current/org/apache/beam/io/requestresponse/Caller.html

BEAM SUMMIT

Simple HTTP GET - 2. Implement interface

```
class HttpGet

implements Caller<String, String> {

    String call(String url) {
        return buildGetRequest(url)
                .execute()
                .parseAsString();
    }
```

Simple HTTP GET - 3. Bring it all together

```
PCollection<GetClassMembersResponse> expand(PCollection<String> input) {

    Result<String, String> result =
        input.apply(
            RequestResponseIO.of(new HttpGet(), StringUTF8Coder.of());

    result.getFailures().apply( // some dead letter )
```

https://beam.apache.org/releases/javadoc/current/org/apache/beam/io/requestresponse/RequestResponseIO.html

BEAM SUMMIT

Simple HTTP GET - 3. Bring it all together

```
PCollection<GetClassMembersResponse> expand(PCollection<String> input) {

    return result
            .getOutput()

            .apply(jsonToRow) // Using JsonPayloadSerializerProvider

            .apply(fromRow);   // Using AutoValueSchemaProvider
```

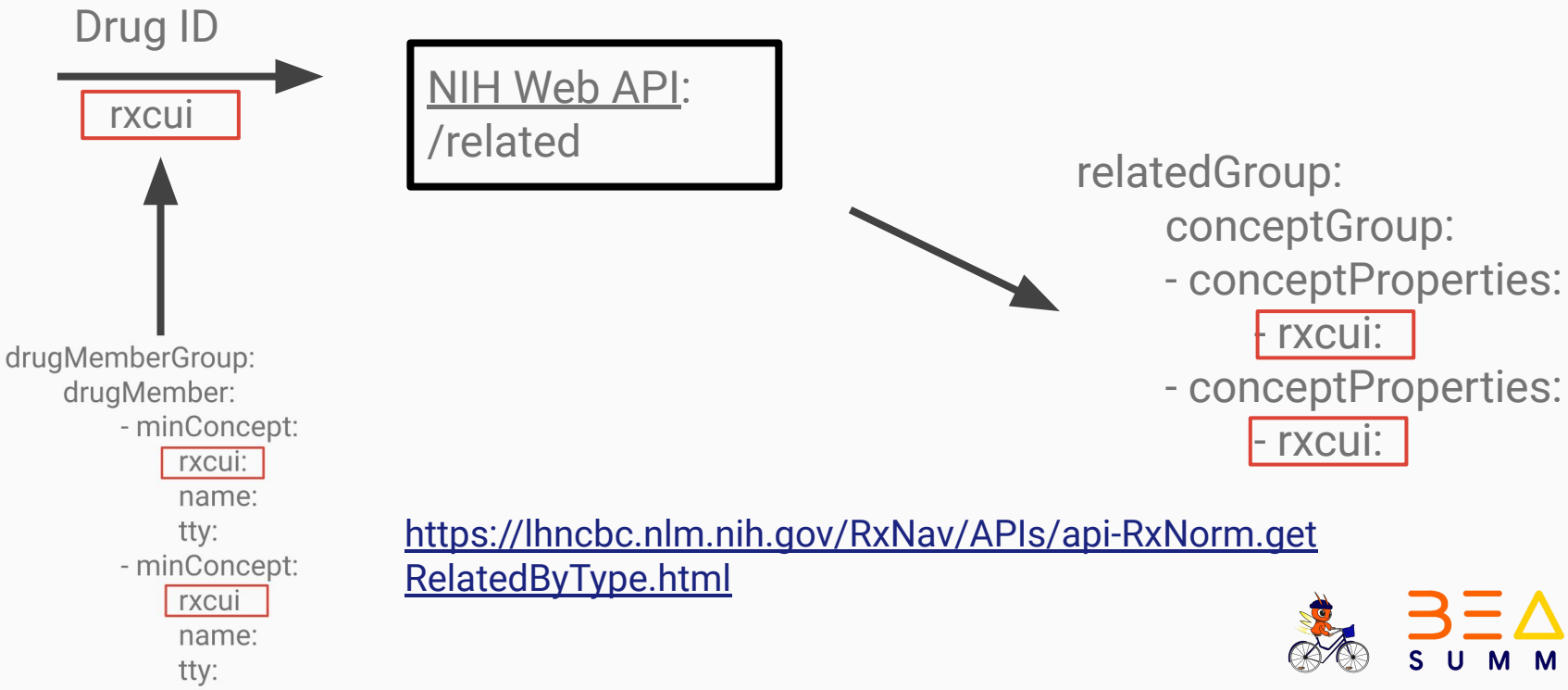
https://beam.apache.org/releases/javadoc/current/org/apache/beam/io/requestresponse/RequestResponseIO.html

BEAM SUMMIT

Chain API requests and responses

BEAM SUMMIT

Chain API requests and responses



https://lhncbc.nlm.nih.gov/RxNav/APIs/api-RxNorm.getRelatedByType.html

Chain API requests and responses

Tasks:

1. Model the response payload
2. Reuse the interface

BEAM SUMMIT

Chain API requests and responses

```
@DefaultSchema(AutoValueSchema.class)
@AutoValue
abstract static class GetSemanticClinicalDrugGroupResponse implements Serializable {
  abstract RelatedGroup getRelatedGroup();

  @AutoValue.Builder
  abstract static class Builder {
    abstract Builder setRelatedGroup(RelatedGroup relatedGroup);

    abstract GetSemanticClinicalDrugGroupResponse build();
  }
}
```

Chain API requests and responses

```
@DefaultSchema(AutoValueSchema.class)
@AutoValue
abstract static class ConceptProperties implements Serializable {
  abstract String getRxcui();

  @AutoValue.Builder
  abstract static class Builder {
    abstract Builder setRxcui(String rxcui);

    abstract ConceptProperties build();
  }
}
```

API Client Setup/Teardown

BEAM SUMMIT

API Client Setup/Teardown

```
interface SetupTeardown {

    void setup();

    void teardown();

}
```

See Javadoc:
https://beam.apache.org/releases/javadoc/current/org/apache/beam/io/requestresponse/SetupTeardown.html

BEAM SUMMIT

API Client Setup/Teardown

```
class CreateContextCache implements
        Caller<CreateCachedContentRequest, CacheContent> {

            CacheContent call(CreateCachedContentRequest request) { }

            void setup() { }

            void teardown() { }

}
```

API Client Setup/Teardown

```
private transient GenAiCacheServiceClient client;

void setup() {
      client = GenAiCacheServiceClient.create(settings);
}

void teardown() {
      client.close();
}
```

BEAM SUMMIT

Additional features - caching and throttling, error handling, etc.

BEAM SUMMIT

Additional features - caching (optional)

```
Cache.Pair<String, String> cache;

RequestResponseIO
    .of(new HttpGet(), coder)
    .withCache(cache);
```

https://beam.apache.org/releases/javadoc/current/org/apache/beam/io/requestresponse/RequestResponseIO.html#withCache-org.apache.beam.io.requestresponse.Cache.Pair-

BEAM SUMMIT

Additional features - caching (optional)

```
Cache.Pair<String, String> cache
    = Cache
        .usingRedis(
            uri,
            requestCoder,
            responseCoder,
            expiry
);
```

https://beam.apache.org/releases/javadoc/current/org/apache/beam/io/requestresponse/Cache.html#usingRedis-java.net.URI-org.apache.beam.sdk.coders.Coder-org.apache.beam.sdk.coders.Coder-org.joda.time.Duration-

BEAM SUMMIT

Additional features - throttling (defaulted)

Handling Overload

Written by Alejandro Forero Cuervo

Edited by Sarah Chavis

Client request rejection probability

$$\max(0, \frac{\text{requests} - K \times \text{accepts}}{\text{requests} + 1})$$

https://sre.google/sre-book/handling-overload/

BEAM SUMMIT

Additional features - error handling

```
interface Caller<RequestT, ResponseT> {
    ResponseT call(RequestT request)
        throws UserCodeExecutionException;
}
```

See Javadoc:
https://beam.apache.org/releases/javadoc/current/org/apache/beam/io/requestresponse/UserCodeExecutionException.html

BEAM SUMMIT

Additional features - error handling

```
Result<RequestT, ResponseT> result;

PCollection<ApiIOError> failures
      = result.getFailures();
```

https://beam.apache.org/releases/javadoc/current/org/apache/beam/io/requestresponse/Result.html

BEAM SUMMIT

Additional features - error handling

Modifier and Type	Method and Description
`abstract java.lang.String`	**`getMessage()`** The `Exception` message.
`abstract Instant`	**`getObservedTimestamp()`** The observed timestamp of the error.
`abstract java.lang.String`	**`getRequestAsString()`** The string representation of the request associated with the error.
`abstract java.lang.String`	**`getStackTrace()`** The `Exception` stack trace.

https://beam.apache.org/releases/javadoc/current/org/apache/beam/io/requestresponse/ApiIOError.html

BEAM SUMMIT

Additional features - signal repeat with backoff

Extends UserCodeExecutionException:

- UserCodeQuotaException
- UserCodeRemoteSystemException
- UserCodeTimeoutException

See Javadoc:
https://beam.apache.org/releases/javadoc/current/org/apache/beam/io/requestresponse/UserCodeExecutionException.html

BEAM SUMMIT

Additional features - configure timeout

```
DEFAULT_TIMEOUT =
    Duration.standardSeconds(30L);

RequestResponseIO
    .withTimeout(duration);
```

BEAM SUMMIT

Conclusion

BEAM SUMMIT

How did Beam help?

- Reduced resulting pdf redundancy
- Reduced pdf input from 1,400 to 27
- Gemini AI Token Count:
  1,039,678 out of 2,097,152 limit

BEAM SUMMIT

“Generate a table comparing adverse events”

Adverse Effect	More Likely with:	Less Likely with:	Comments
Extrapyramidal Symptoms (EPS)	High potency typicals (e.g., haloperidol), risperidone (at higher doses)	Atypicals like quetiapine, clozapine, aripiprazole	EPS risk often dose-dependent; can be acute lead to tardive dyskinesia
Weight Gain	Olanzapine, clozapine	Aripiprazole, lurasidone, ziprasidone (some studies)	Weight gain varies significantly between individuals; monitoring crucial
Hyperglycemia/Diabetes	Olanzapine, clozapine	Aripiprazole, lurasidone	Regular blood glucose monitoring essential, especially in patients with risk factors
Dyslipidemia	Olanzapine, clozapine	Aripiprazole, lurasidone, ziprasidone	Fasting lipid profile monitoring recommended
QT Prolongation	Ziprasidone, iloperidone, thioridazine	Most atypicals (except ziprasidone, iloperidone)	QTc prolongation increases risk of serious arrhythmia; careful ECG monitoring with certain agents
Sedation	Clozapine, olanzapine, quetiapine, low potency typicals	Aripiprazole, lurasidone, risperidone	Sedation can vary greatly by individual and dose
Anticholinergic Effects	Typical antipsychotics	Most atypicals	Anticholinergic effects can lead to constipation, dry mouth, blurred vision, etc

BEAM SUMMIT