Cloud based data science for FHIR and genomics data

Presenter

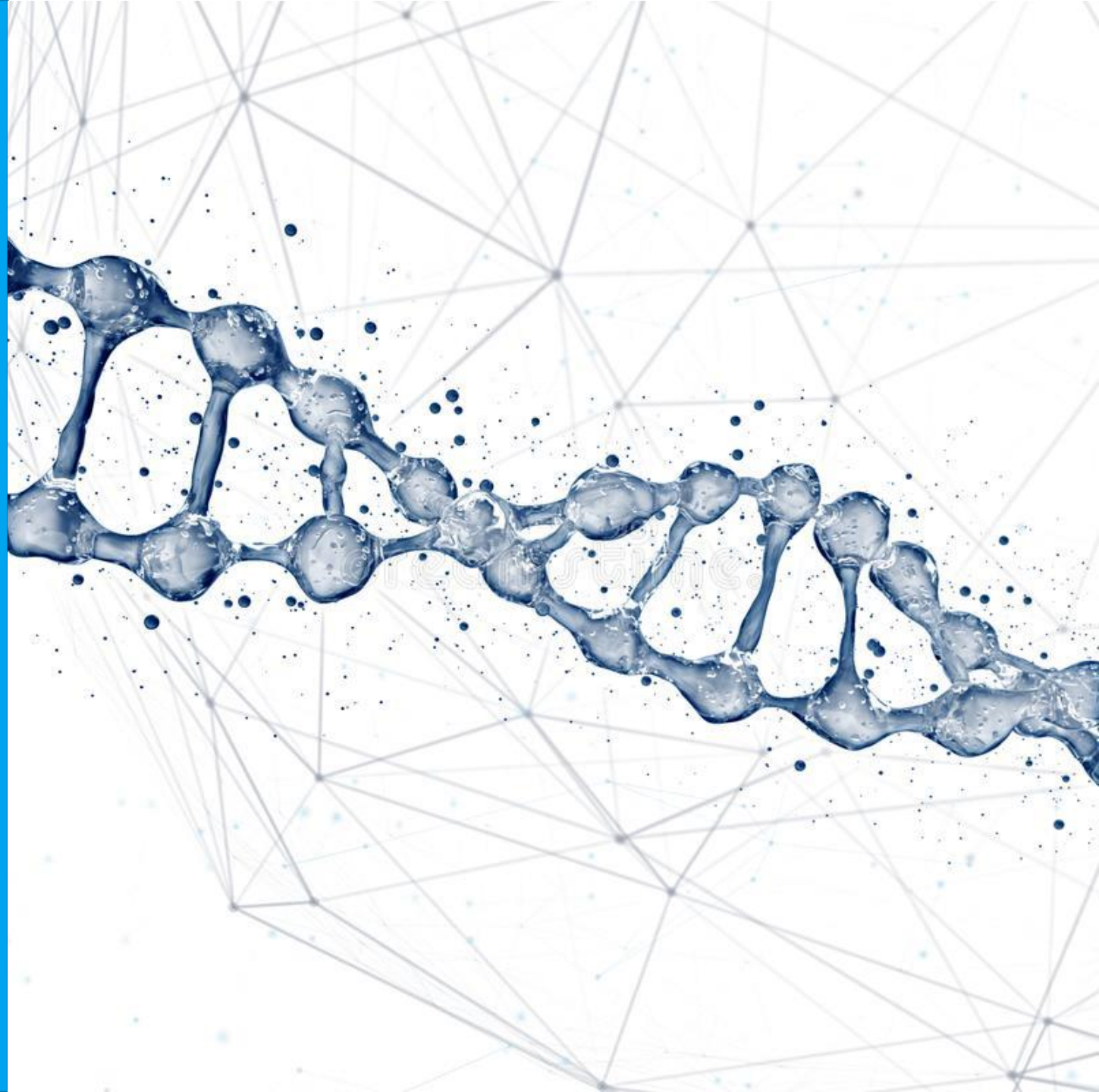
Tim Dunn

Mentor

Erdal Cosgun

Co-Mentors

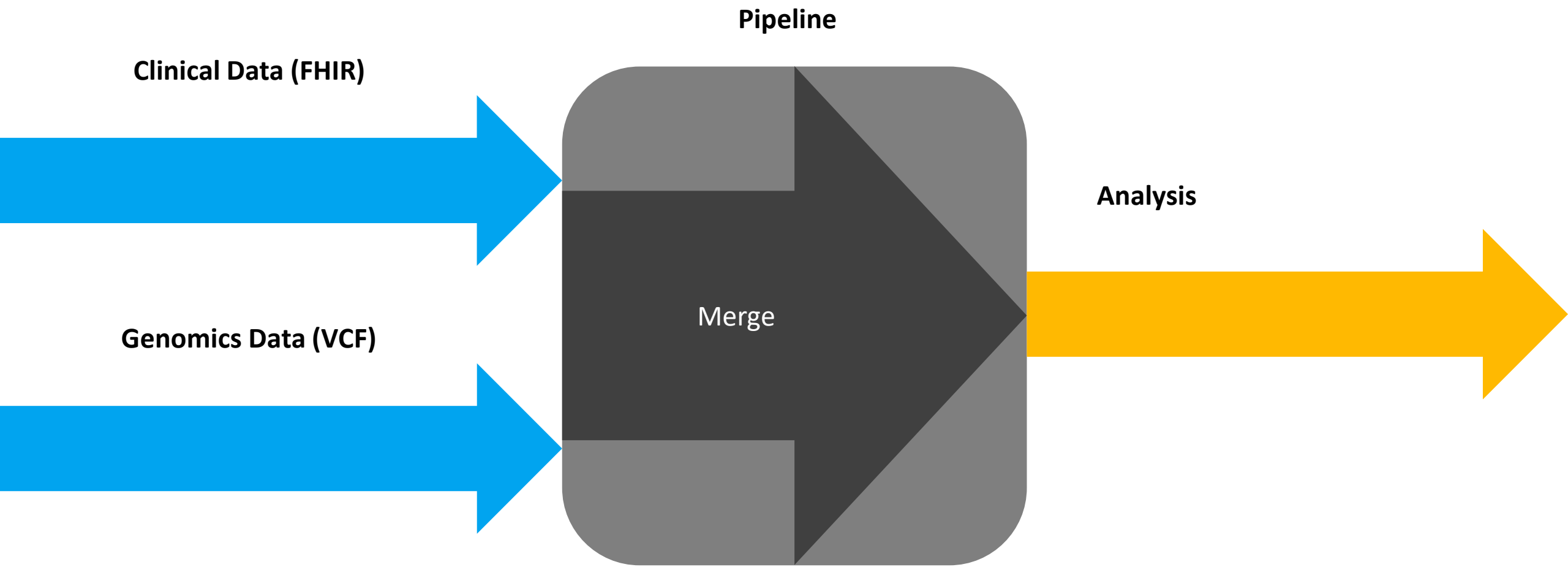
Venkat Malladi and Olesya Melnichenko



Outline

1. Introduction
2. Methods
3. Results
4. Conclusion

Project Overview



Fast Healthcare Interoperability Resource (FHIR)

- Standard framework for storing clinical data
- ~140 defined Resource types

https://www.hl7.org/fhir/

Fast Healthcare Interoperability Resource (FHIR)

Name	Flags	Card.	Type
Patient	N		DomainResource
identifier	Σ	0..*	Identifier
active	?! Σ	0..1	boolean
name	Σ	0..*	HumanName
telecom	Σ	0..*	ContactPoint
gender	Σ	0..1	code
birthDate	Σ	0..1	date
deceased[x]	?! Σ	0..1	
deceasedBoolean			boolean
deceasedDateTime			dateTime
address	Σ	0..*	Address
maritalStatus		0..1	CodeableConcept
multipleBirth[x]		0..1	
multipleBirthBoolean			boolean
multipleBirthInteger			integer
photo		0..*	Attachment

Fast Healthcare Interoperability Resource (FHIR)


```
'resource': {'resourceType': 'Patient',
    'id': 'b5f1da11-3826-4821-bb84-dd72294c9a4c',
    'meta': {'versionId': '1',
     'lastUpdated': '2022-07-13T19:23:23.981+00:00',
     'profile': ['http://hl7.org/fhir/us/core/StructureDefinition/us-core-patient']},
    'text': {'status': 'generated',
     'div': '<div xmlns="http://www.w3.org/1999/xhtml">Generated by Synthea</div>'},
    'extension': [{'extension': [{'url': 'ombCategory',
        'valueCoding': {'system': 'urn:oid:2.16.840.1.113883.6.238',
         'code': '2106-3',
         'display': 'White'}},
       {'url': 'text', 'valueString': 'White'}],
      'url': 'http://hl7.org/fhir/us/core/StructureDefinition/us-core-race'},
     {'extension': [{'url': 'ombCategory',
        'valueCoding': {'system': 'urn:oid:2.16.840.1.113883.6.238',
         'code': '2135-2',
         'display': 'Hispanic or Latino'}},
       {'url': 'text', 'valueString': 'Hispanic or Latino'}],
      'url': 'http://hl7.org/fhir/us/core/StructureDefinition/us-core-ethnicity'},
     {'url': 'http://hl7.org/fhir/StructureDefinition/patient-mothersMaidenName',
      'valueString': 'Julia241 Luna60'},
     {'url': 'http://hl7.org/fhir/us/core/StructureDefinition/us-core-birthsex',
      'valueCode': 'M'},
     {'url': 'http://hl7.org/fhir/StructureDefinition/patient-birthPlace',
      'valueAddress': {'city': 'Portsmouth',
       'state': 'Saint John Parish',
       'country': 'DM'}},
        ...
```

Genome Sequencing

Reference:

ACGTCCATGGACATATATGAGGCC…

Reads:

ACGTCGAT

TCGATGGACA

CATATGAGGC

Variant Call Format (VCF)

Reference:

ACGTCCATGGACATATATGAGGCC…

Alternate:

ACGTCGATGGACA TATGAGGCC…

Variant Call Format (VCF)

Reference:

ACGTCCATGGACATATATGAGGCC…

Alternate:

ACGTCGATGGACA TATGAGGCC…

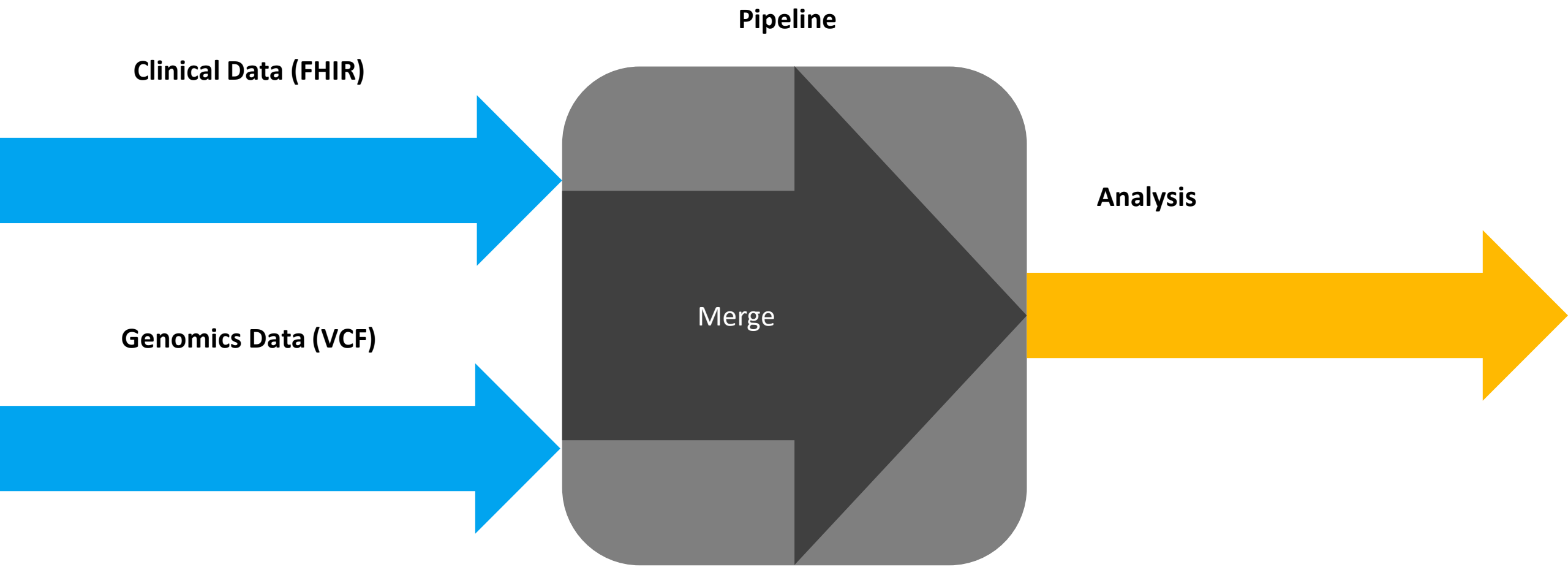
VCF:

POS	REF	ALT
6	C	G
13	ATA	A

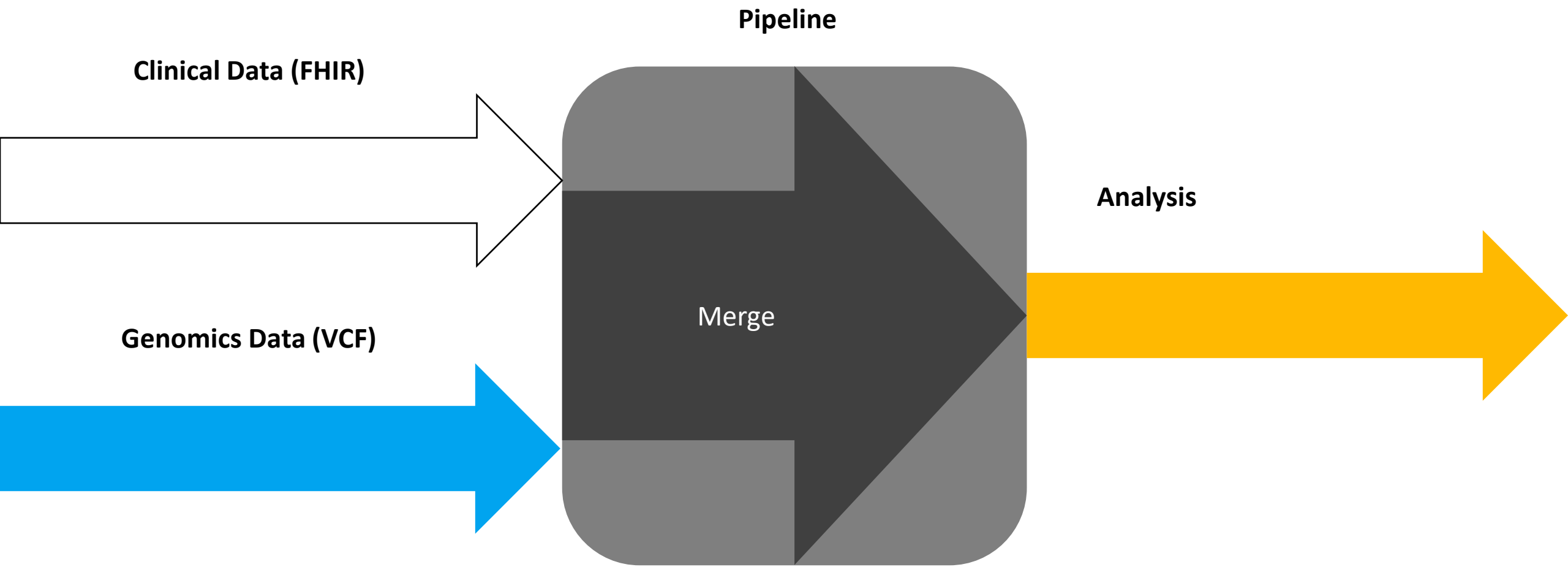
Variant Call Format (VCF)

```
##fileformat=VCFv4.3
##fileDate=20210115
##source=VariantCallerScript
##reference=file:///references/HG38.fasta
##contig=<ID=20,length=62435964,assembly=B36,species="Homo sapiens">
##phasing=partial
##INFO=<ID=DP,Number=1,Type=Integer,Description="Total Depth">
##INFO=<ID=AF,Number=A,Type=Float,Description="Allele Frequency">
##INFO=<ID=AA,Number=1,Type=String,Description="Ancestral Allele">
##INFO=<ID=DB,Number=0,Type=Flag,Description="dbSNP membership, build 129">
##FILTER=<ID=q10,Description="Quality below 10">
##FILTER=<ID=s50,Description="Less than 50% of samples have data">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
##FORMAT=<ID=HQ,Number=2,Type=Integer,Description="Haplotype Quality">
#CHROM POS     ID  REF ALT    QUAL FILTER INFO              FORMAT   NA00001     NA00002
20     14370   rs6 G   A      29   PASS   DP=14;DB;         GT:DP:HQ 0|0:1:51,51 1|0:8:51,51
20     17330   .   T   A       3   q10    DP=11;AF=0.017 GT:DP:HQ 0|0:3:58,50 0|1:5:65,3
20     1110696 rs5 A   G,T    67   PASS   DP=10;AA=T;DB  GT:DP:HQ 1|2:6:23,27 2|1:0:18,2
20     1230237     .   T .    47   PASS   DP=13;AA=T      GT:DP:HQ 0|0:7:56,60 0|0:4:51,51
20     1234567 ms1 GTC G,GTCT 50   PASS   DP=9;AA=G       GT:DP    0/1:4       0/2:2
```

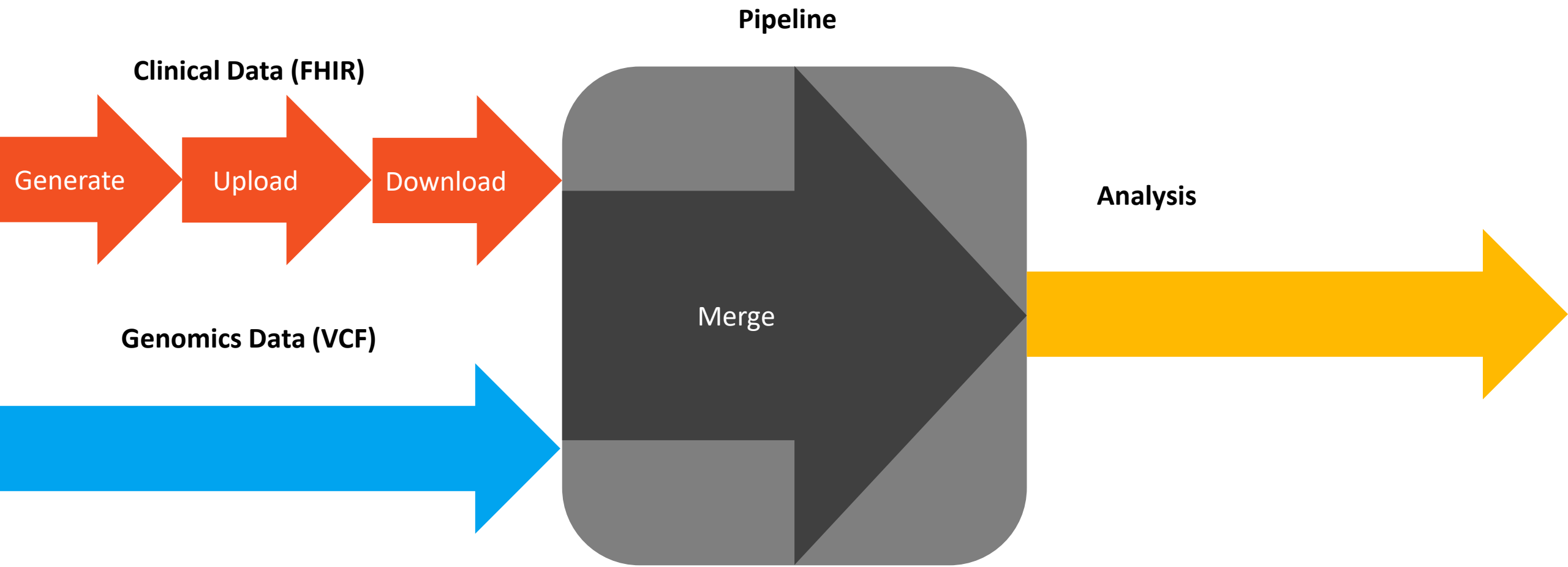
Project Overview



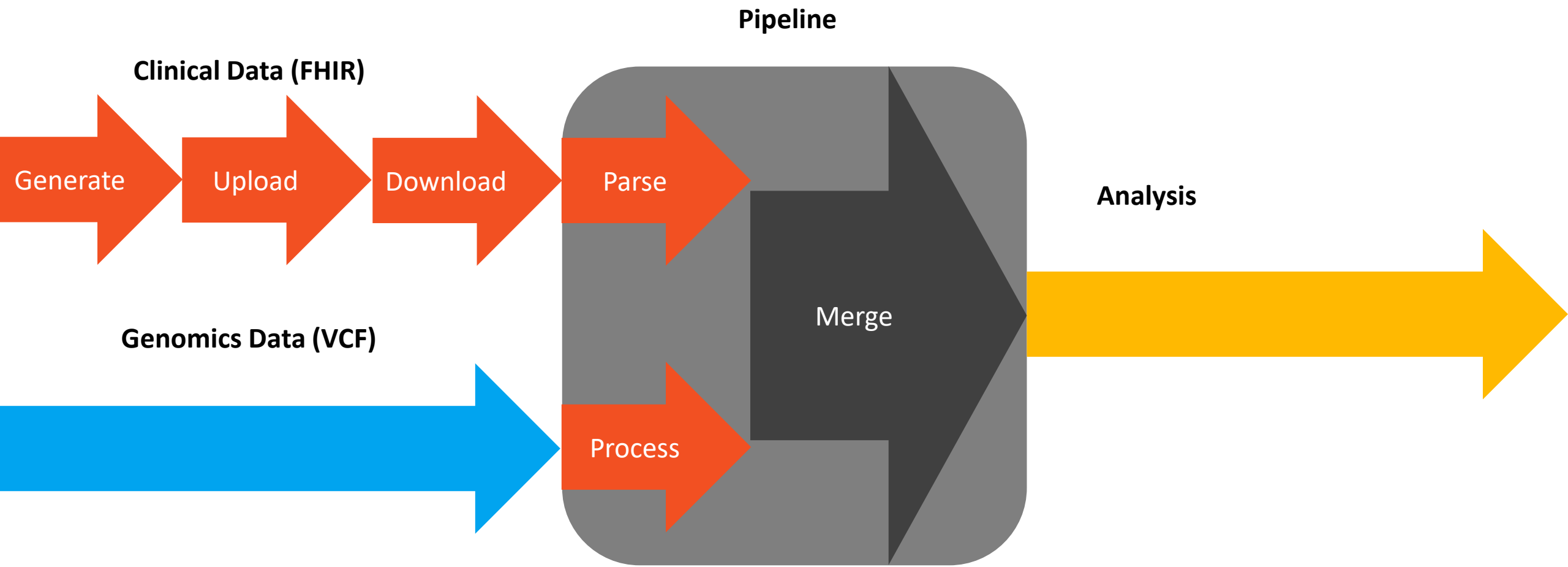
Project Overview



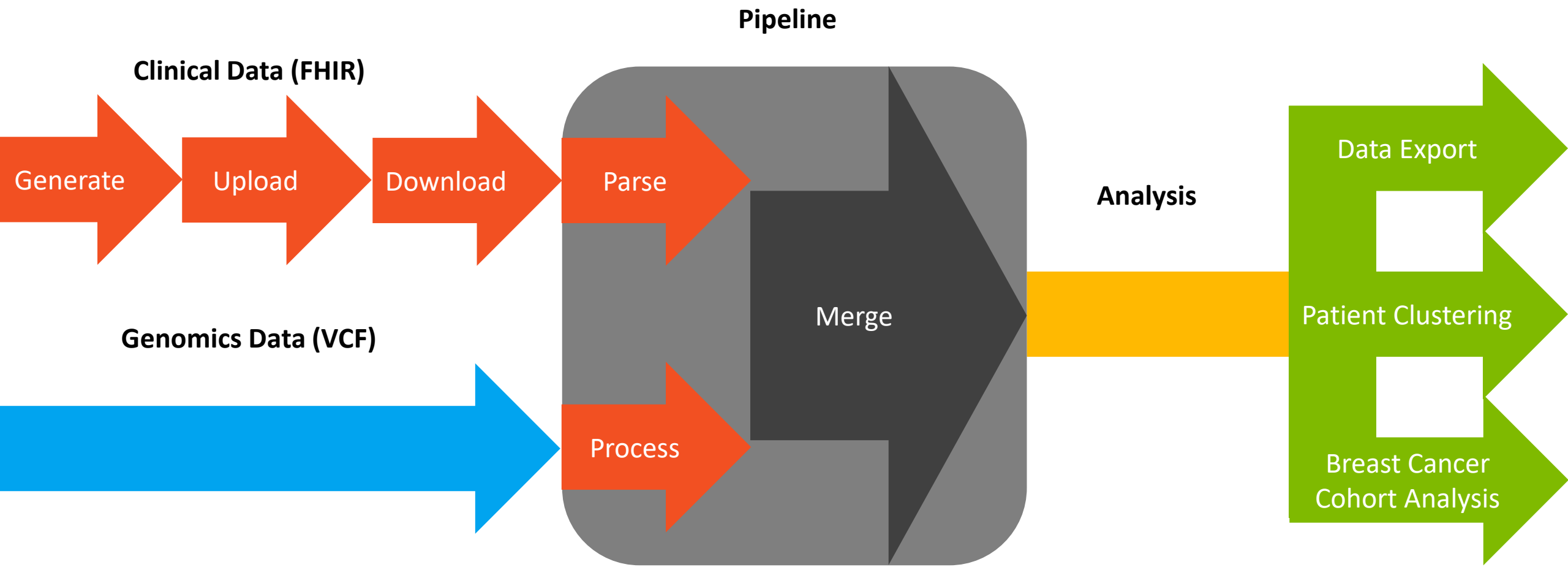
Project Overview



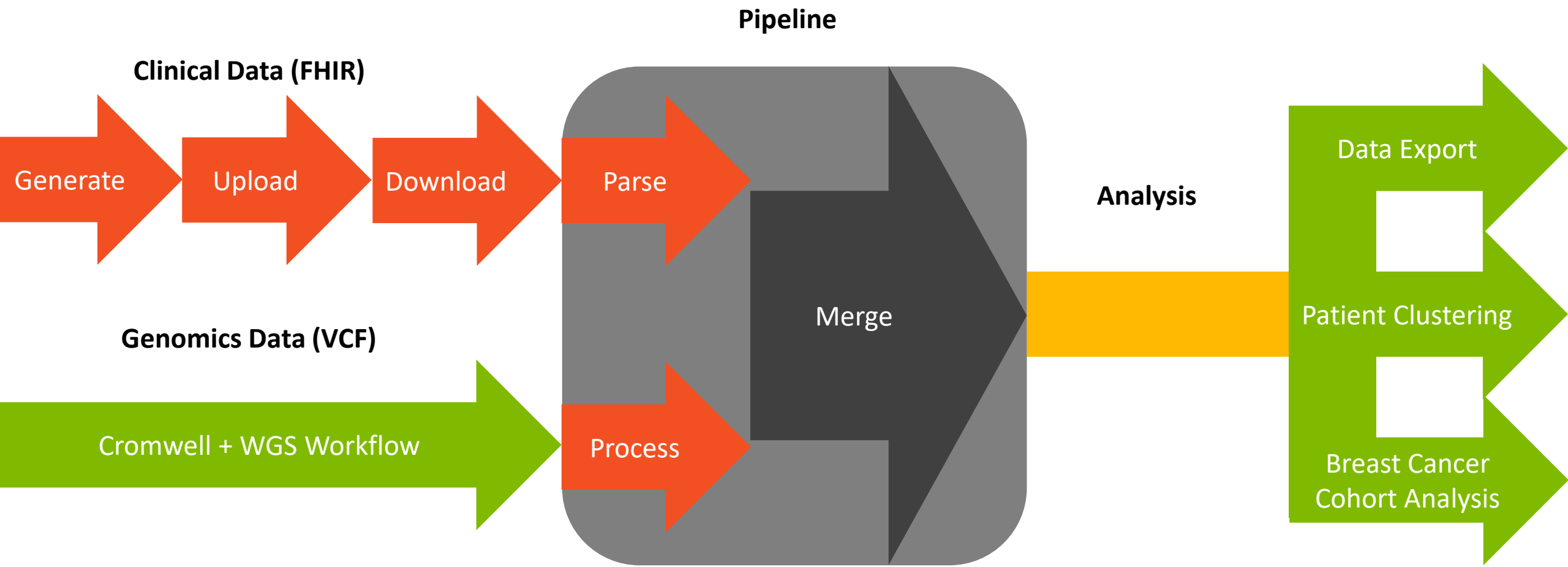
Project Overview



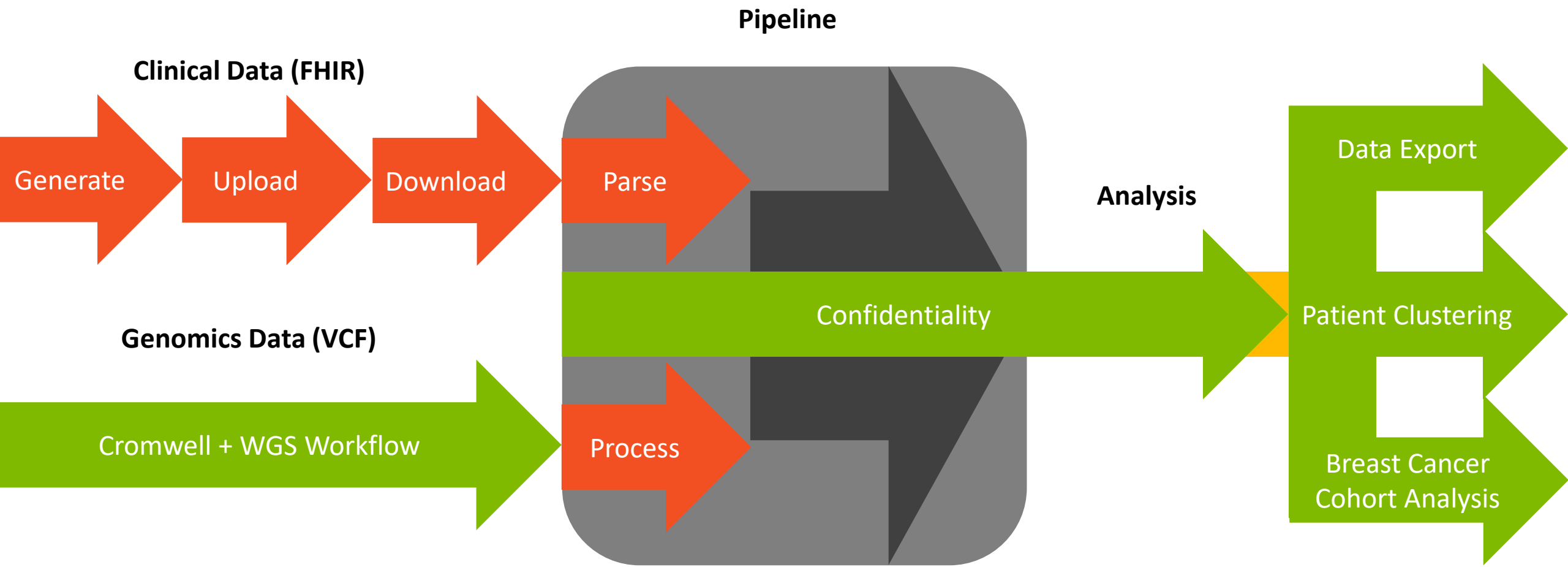
Project Overview



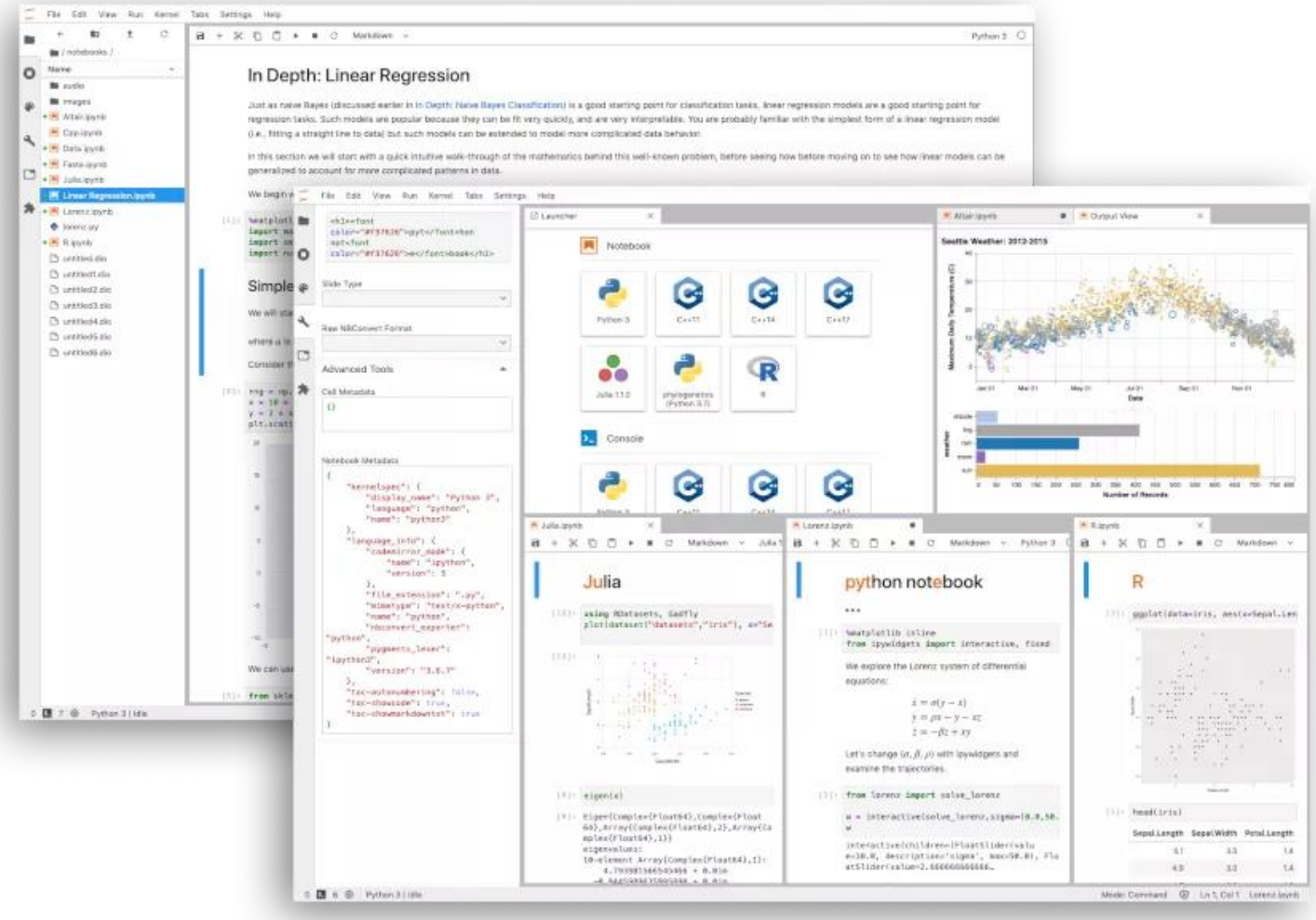
Project Overview



Project Overview

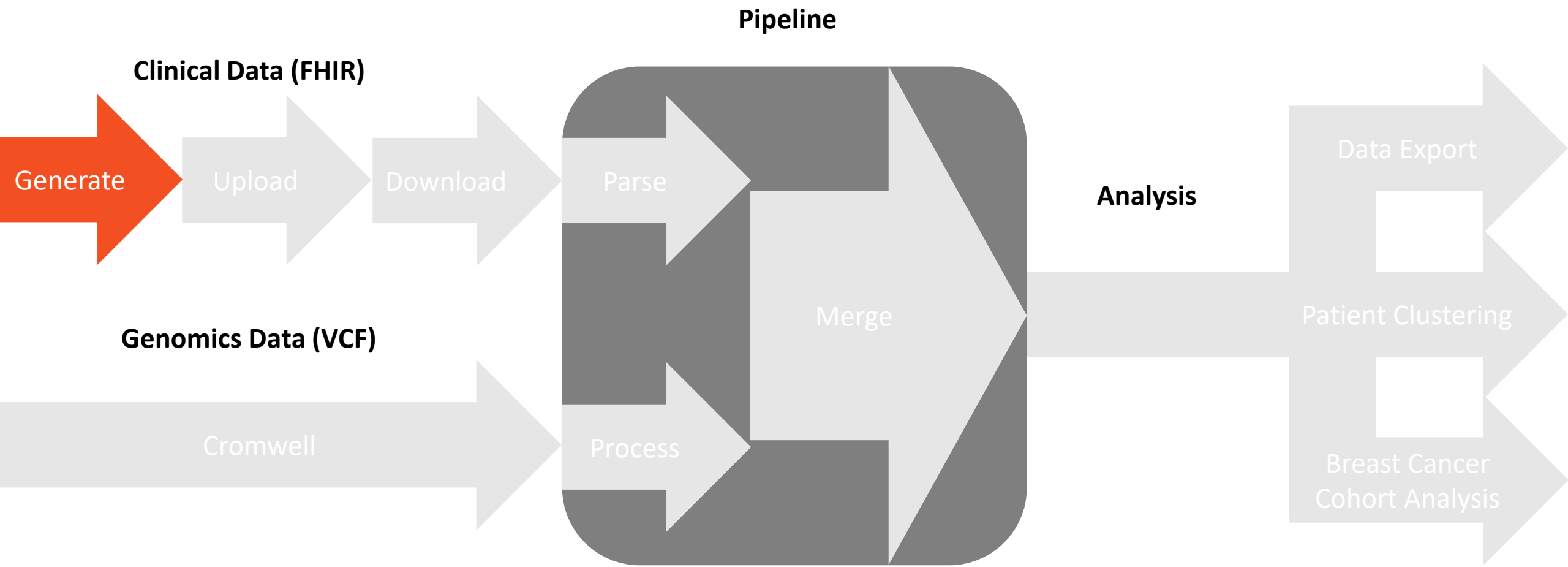


Environment: JupyterLab and Jupyter Notebook

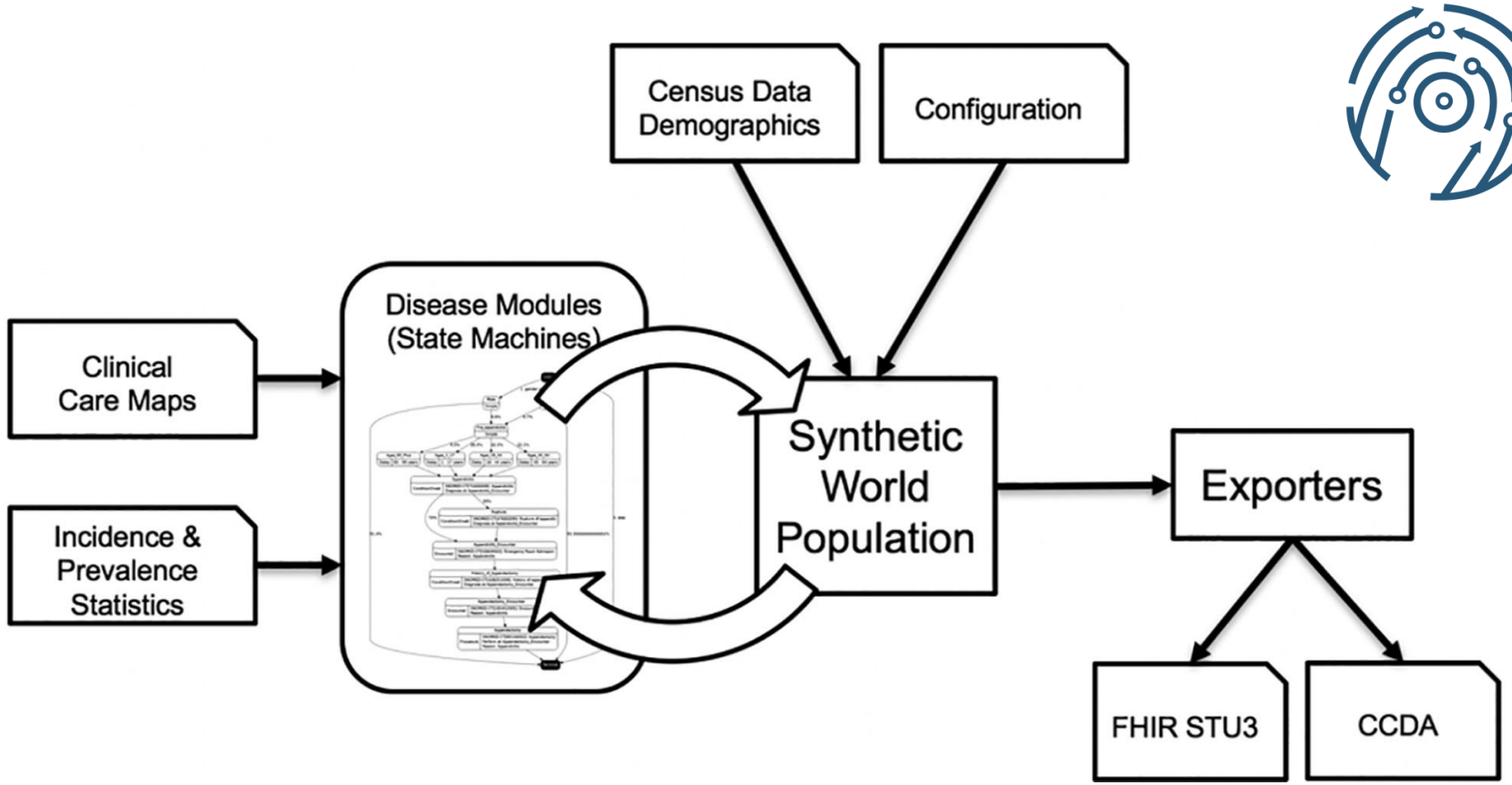


https://jupyter.org/

Project Overview



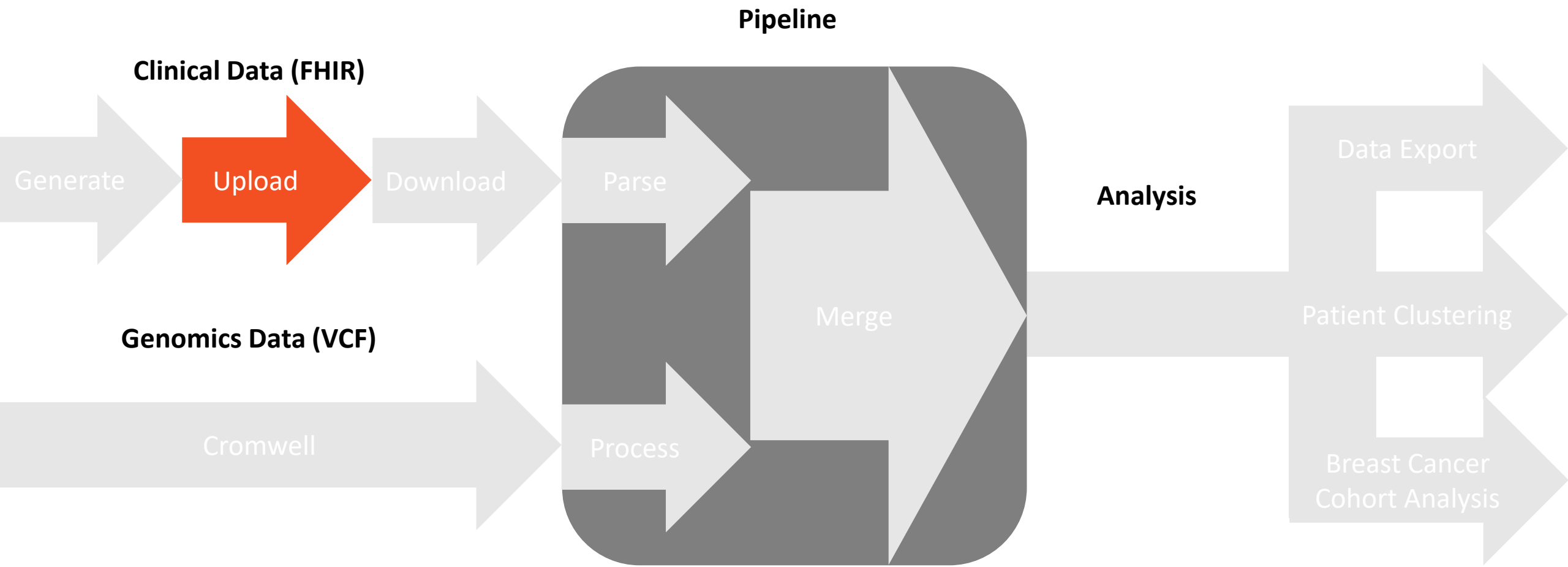
Synthea: Generate Synthetic FHIR Data



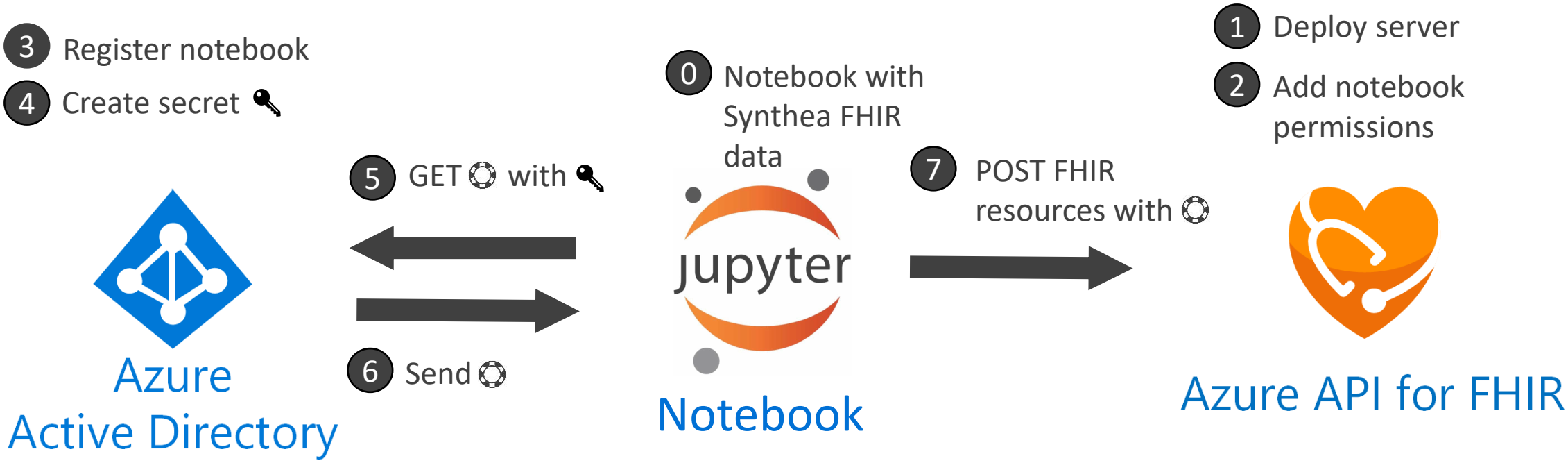
Synthea

Walonoski, Jason, et al. "Synthea: An approach, method, and software mechanism for generating synthetic patients and the synthetic electronic health care record." *Journal of the American Medical Informatics Association* 25.3 (2018): 230-238.

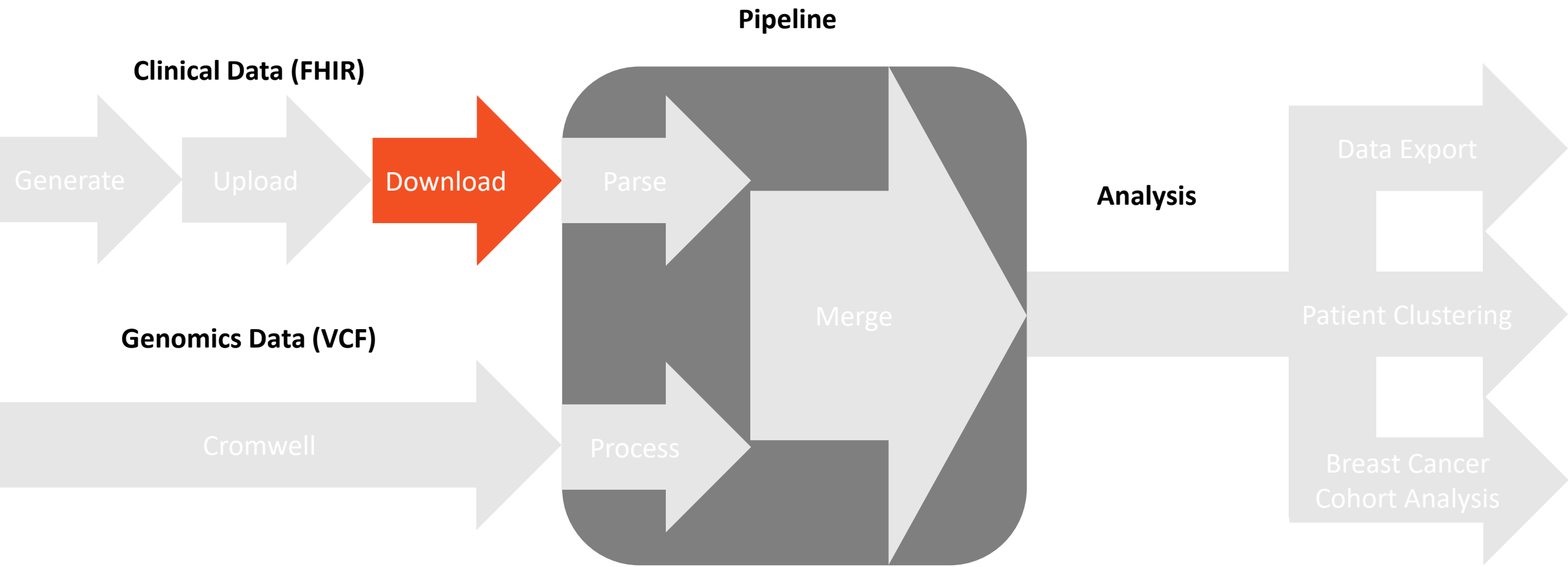
Project Overview



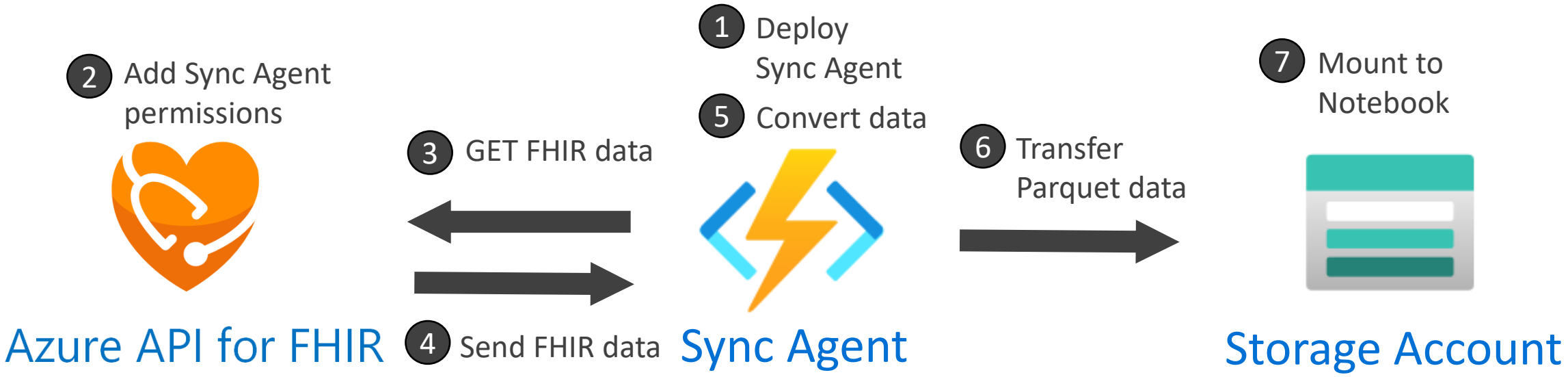
Azure API for FHIR: Upload FHIR Data



Project Overview

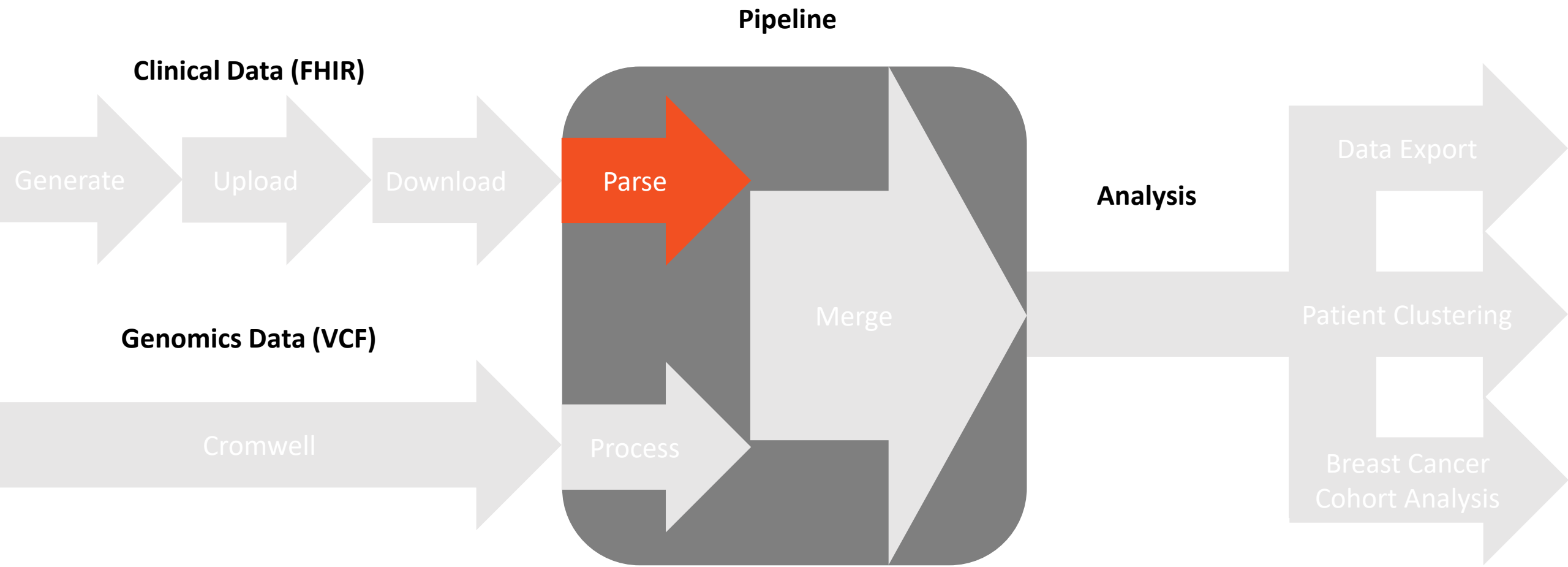


Sync Agent: Download FHIR Data

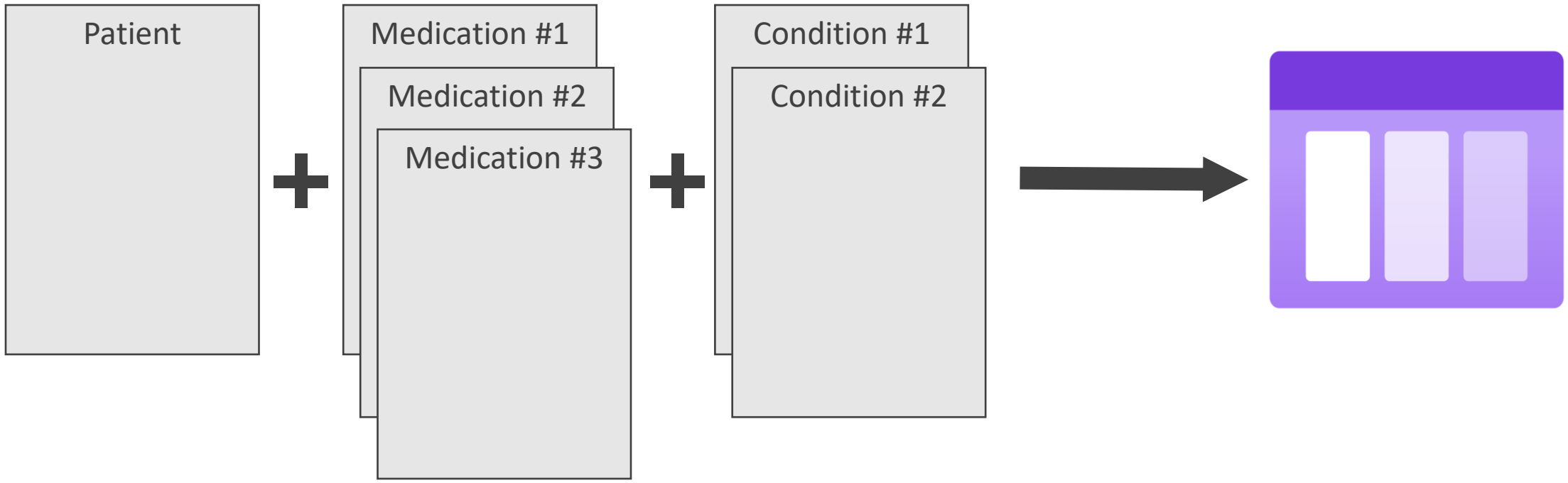


https://github.com/microsoft/FHIR-Analytics-Pipelines/blob/main/FhirToDataLake/docs/Deployment.md

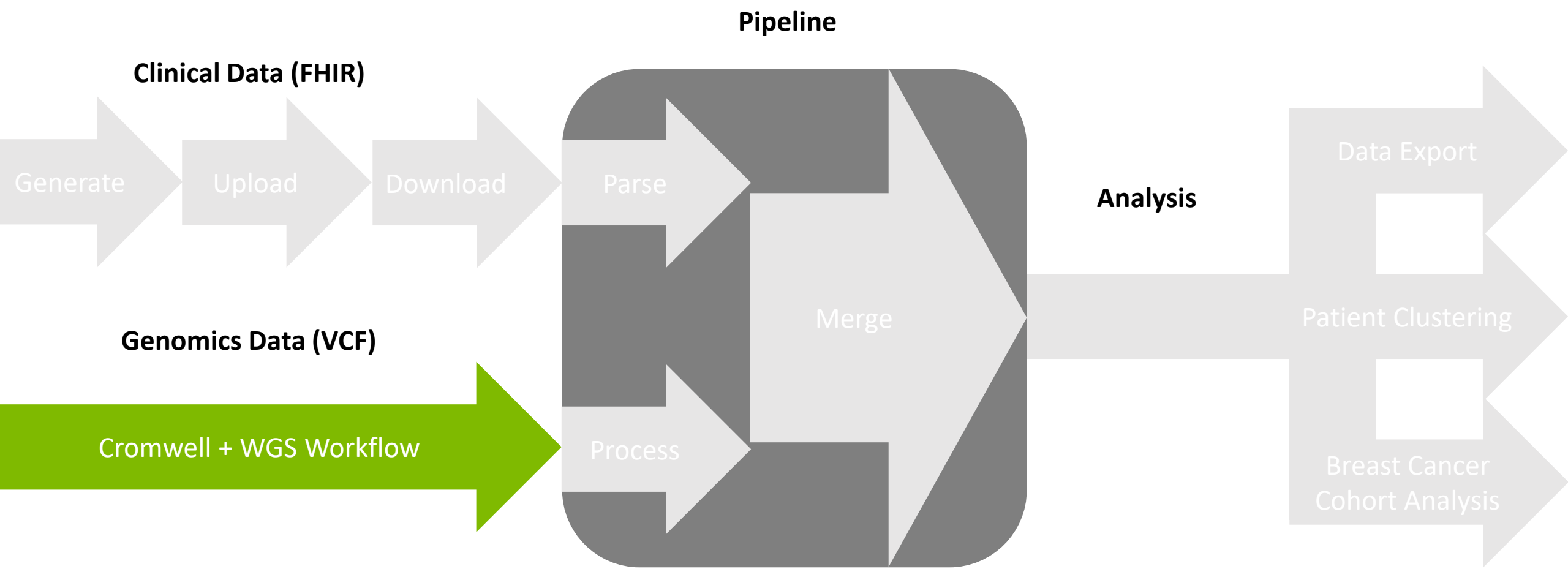
Project Overview



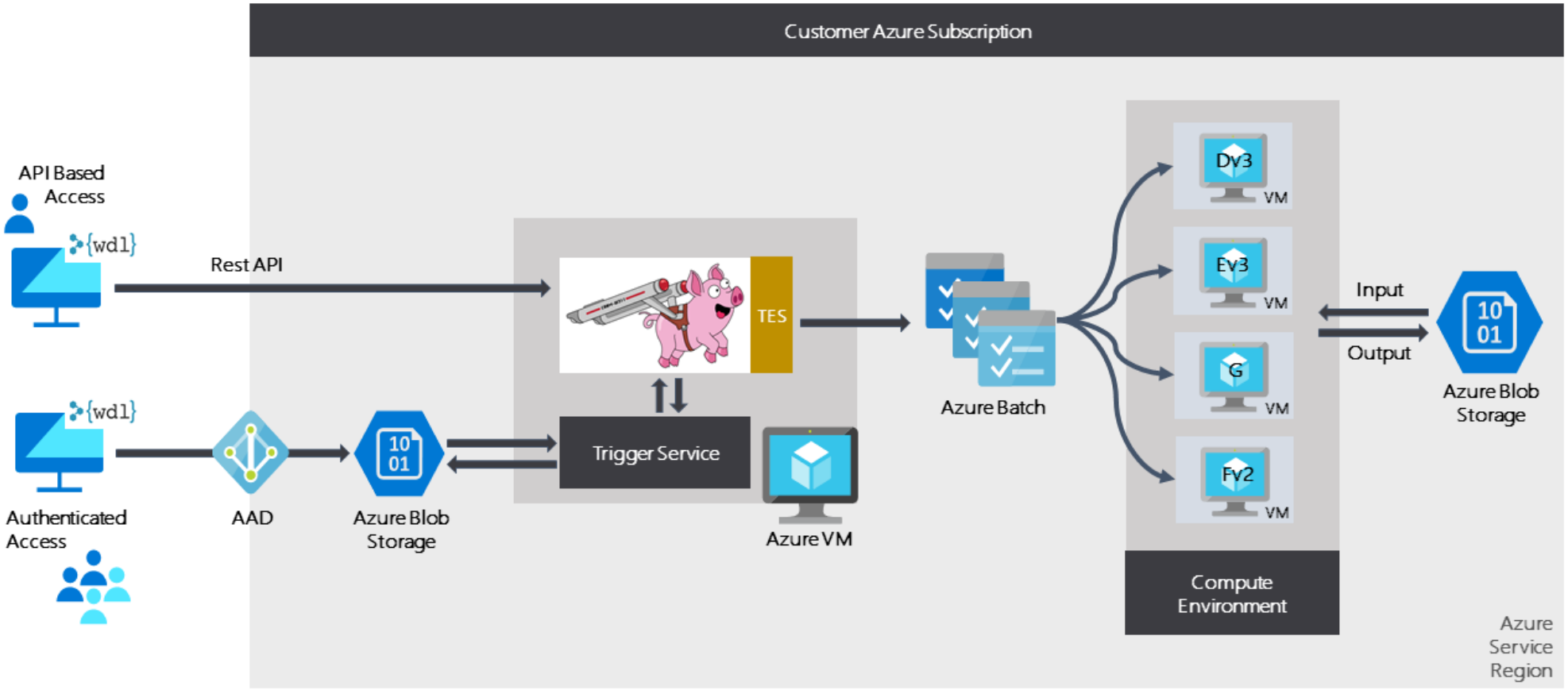
Jupyter NB: Parse FHIR Data



Project Overview

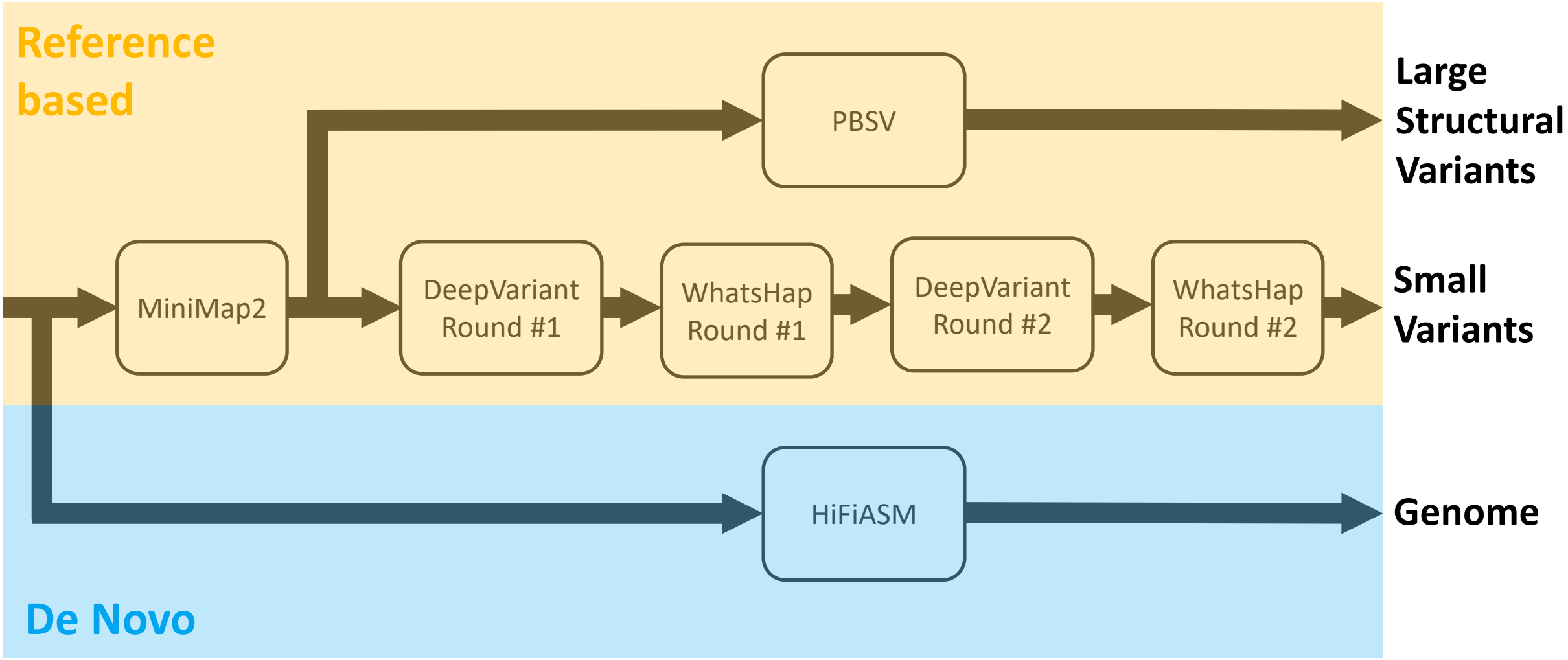


Cromwell on Azure: Overview

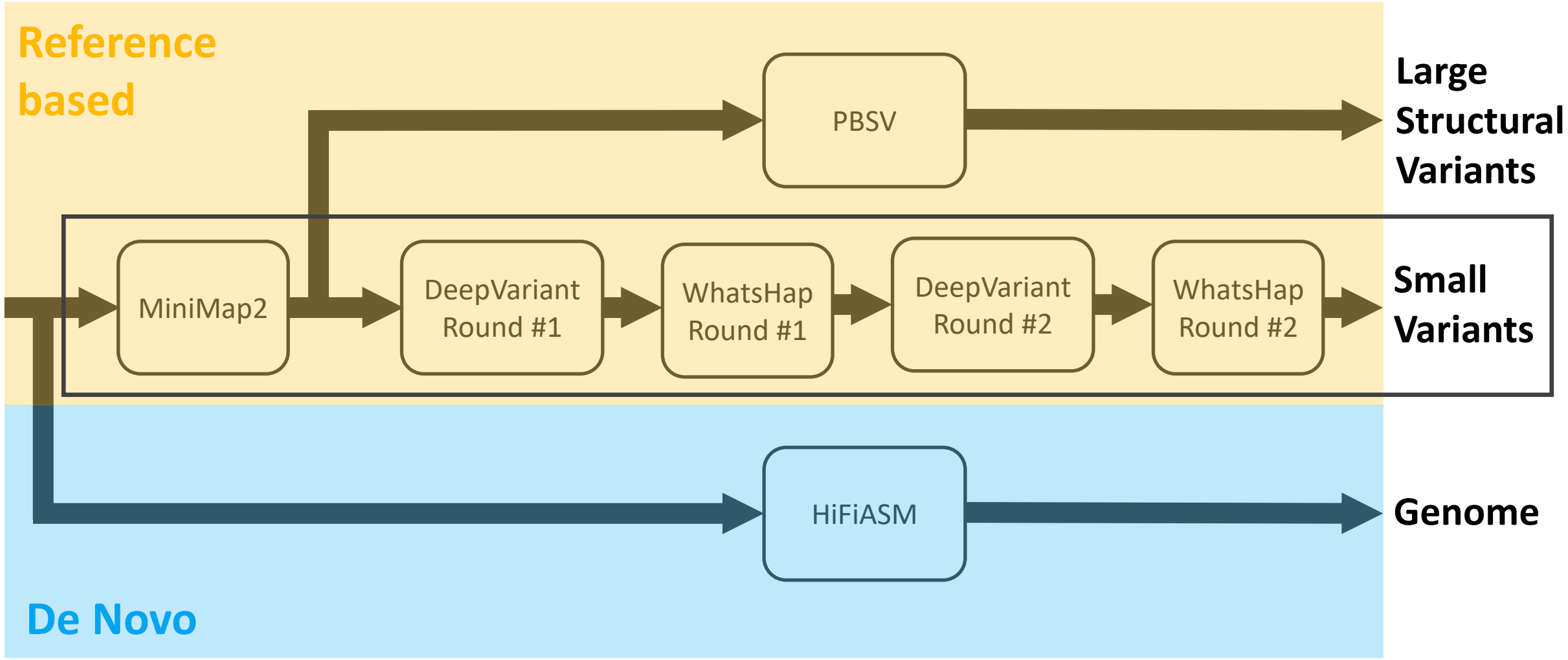


Microsoft “Cromwell on Azure” Github page: https://github.com/microsoft/CromwellOnAzure

Cromwell on Azure: PacBio Human WGS Workflow



Cromwell on Azure: PacBio Human WGS Workflow



Cromwell on Azure: Changes

https://github.com/PacificBiosciences/pb-human-wgs-workflow-wdl/pull/49

TimD1 added 7 commits 4 days ago

`Added missing 'sample' workflow inputs.` ··· `8810520`

- added 'sample_trial.run_jellyfish' to 'trial.singleton.inputs.json'
- added 'sample_trial.score_matrix', 'sample_trial.tg_bed', and 'sample_trial.tg_list' to 'reference.trial.inputs.json'

`Removed 'sample_trial.last_reference' from 'trial.singleton.inputs.json'` ··· `6c2766c`

- 'sample_trial.last_reference' was defined twice
- second definition is in 'reference.trial.inputs.json'

`Fixed nesting of 'sample_trial.last_reference'.` ··· `dd28b67`

- based on usage in 'sample.trial.wdl', shouldn't be in nested `Array`s

`Added '--ignore-read-groups' for round 1 WhatsHap.` ··· `b4f10d8`

- this fixes "no read groups in common" error between VCF and BAM
- this fix appears to break downstream `Picard MergeSamFiles` with error "program record with group id whatshap already"

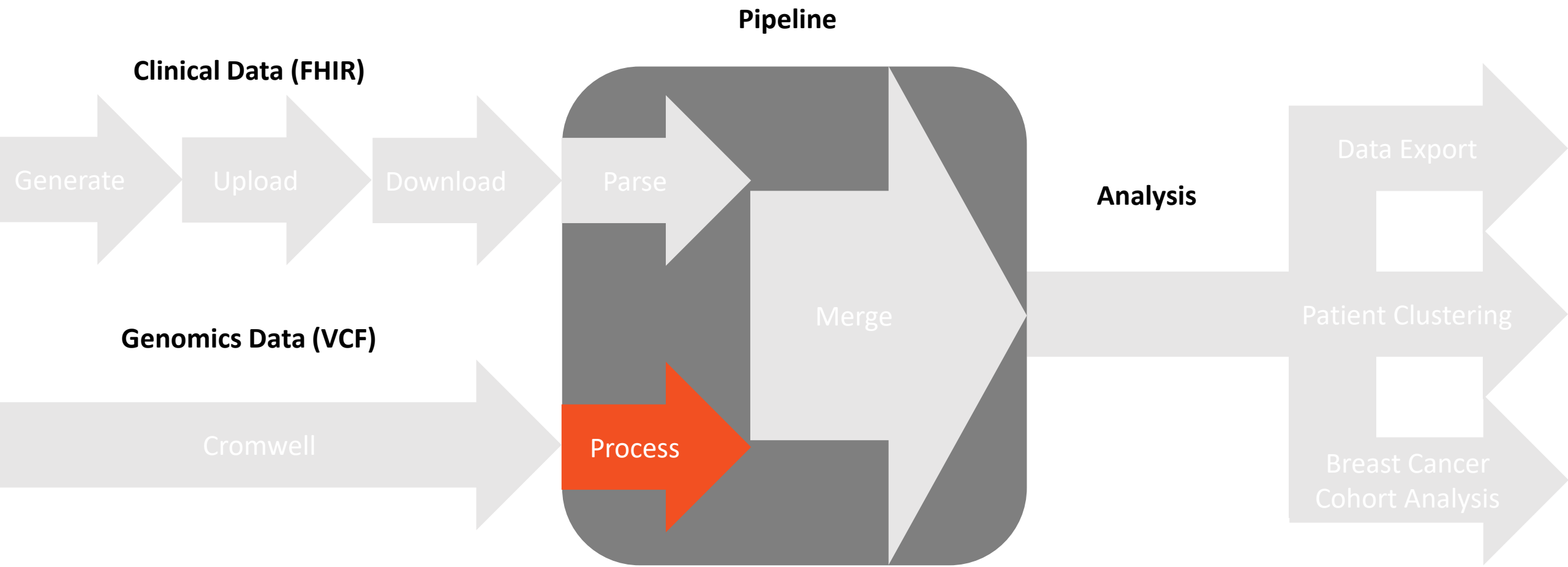
`Replaced Picard MergeSamFiles w samtools merge` ··· `6a8c2b2`

- Fixes "program record with group id whatshap already" error

`Added helpful 'rename_urls' script` ··· `4151099`

- renames all WDL Github URLs for easier development, between two URLs: local for testing, and remote for committing
- just run ./rename_urls for usage information

Project Overview



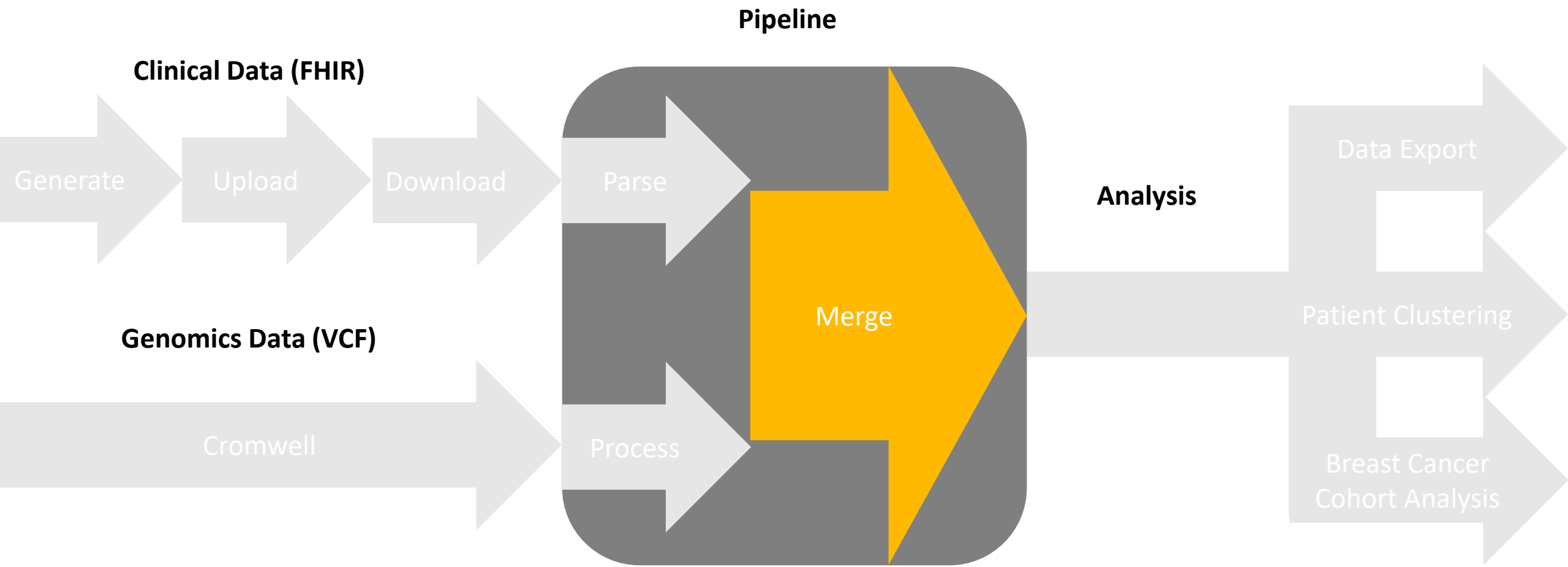
Process VCF Data: BCFTools and Pandas

- Split multi-allelic sites
- Linkage Disequilibrium (LD) pruning
- Merge VCFs to single joint VCF
- Export to TSV
- Filter sites
  - depth, position, quality, allele frequency…
- Impute missing fields

CHROM	POS	REF	ALT
chr20	1232	A	T,C

CHROM	POS	REF	ALT
chr20	1232	A	T
chr20	1232	A	C

Project Overview



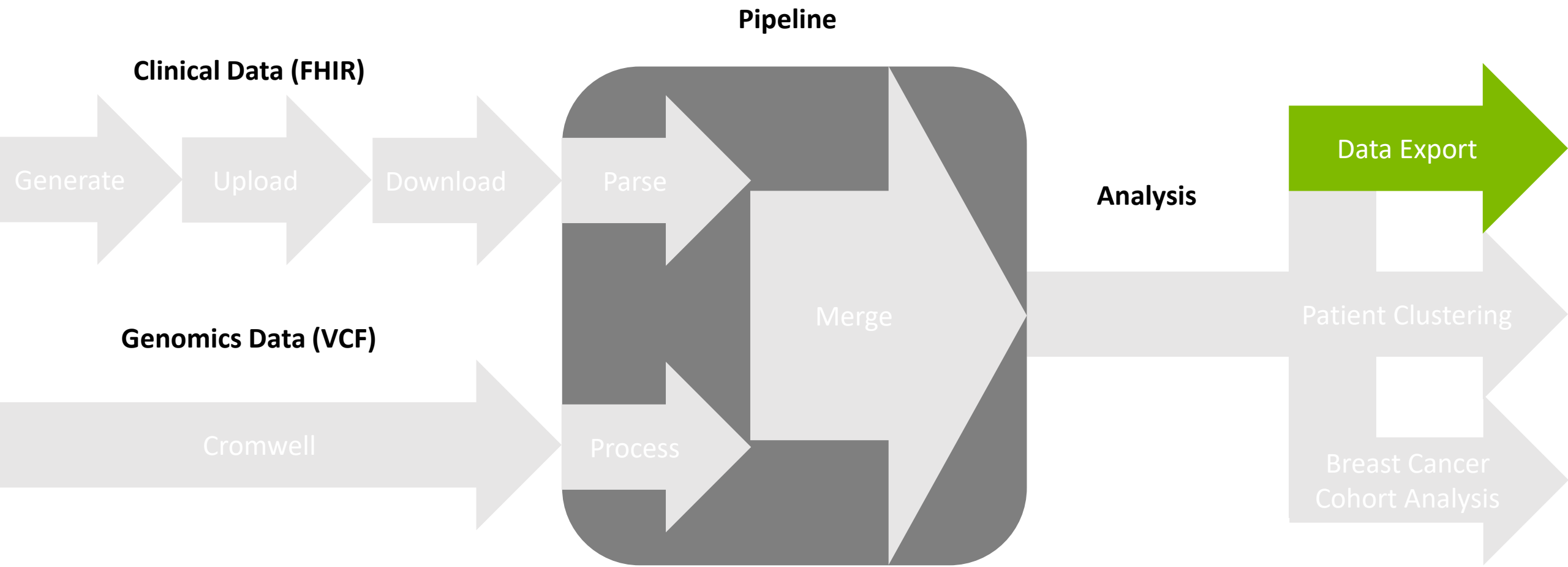
Merge VCF and FHIR Data: Pandas

Clinical: FHIR

Genomic: VCF

	city	state	country	gender	dead	age	newest_med_code	...	49:GT	49:AF_0	49:AF_1	49:PL_0/0	49:PL_0/1	49:PL_1/1	49:DP
0	Boston	MA	US	female	False	22.009037	748856	...	0/0	1	0	0	8	8	7
1	Amesbury	MA	US	male	False	64.588092	310798	...	0/0	1	0	0	3	3	7
2	Boston	MA	US	female	False	17.316361	1367439	...	0/1	0.285714	0.714286	29	0	23	7
3	Yarmouth	MA	US	male	True	69.918688	896209	...	0/0	1	0	0	4	4	5
4	Attleboro	MA	US	female	False	25.182206	751905	...	0/0	1	0	0	9	9	15

Project Overview



Application #1: Data Exportation

Overview

- Perform joint queries of VCF/FHIR data in Azure Synapse Analytics

Motivation

- Simplicity
- Requires conversion of VCF/FHIR to tabular formats
- Demonstrates database integration

Application #1: Data Exportation

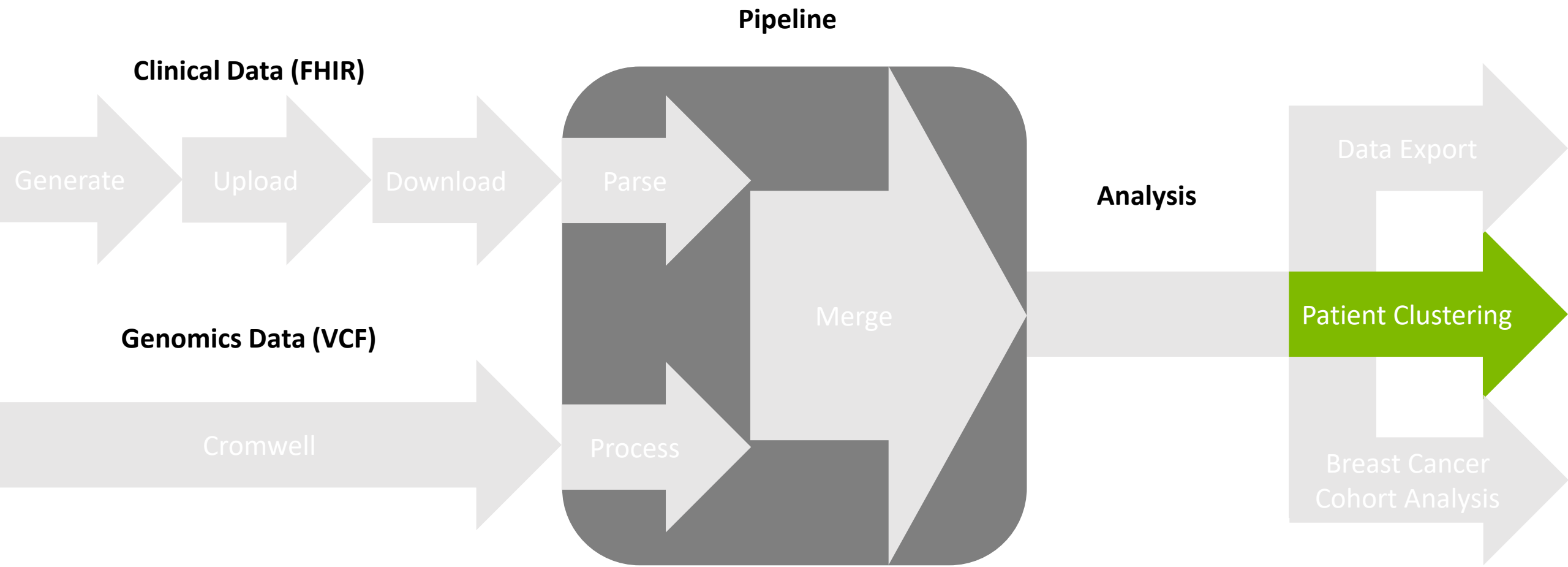
```
1  %%sql
2  SELECT * from fhir_table, pacbio_table
3  WHERE fhir_table.id='f6338d55-c6b5-41dc-9479-97348c418d60' AND pacbio_table.CHROM="chr1" AND pacbio_table.POS < 11000
```

[9] ✓ - Command executed in 10 sec 647 ms on 10:57:38 AM, 7/07/22

View: Table | Chart ⟼ Export results

CHROM	POS	TYPE	REF	ALT
chr1	10108	INDEL	C	CT
chr1	10622	SNP	T	G
chr1	10623	SNP	T	C
chr1	10626	INDEL	A	AGGCGCAG
chr1	10627	INDEL	A	AG
chr1	10884	SNP	C	G
chr1	10897	SNP	T	A
chr1	10927	SNP	A	G
chr1	10931	SNP	C	T
chr1	10934	SNP	C	T

Project Overview



Application #2: Patient Clustering

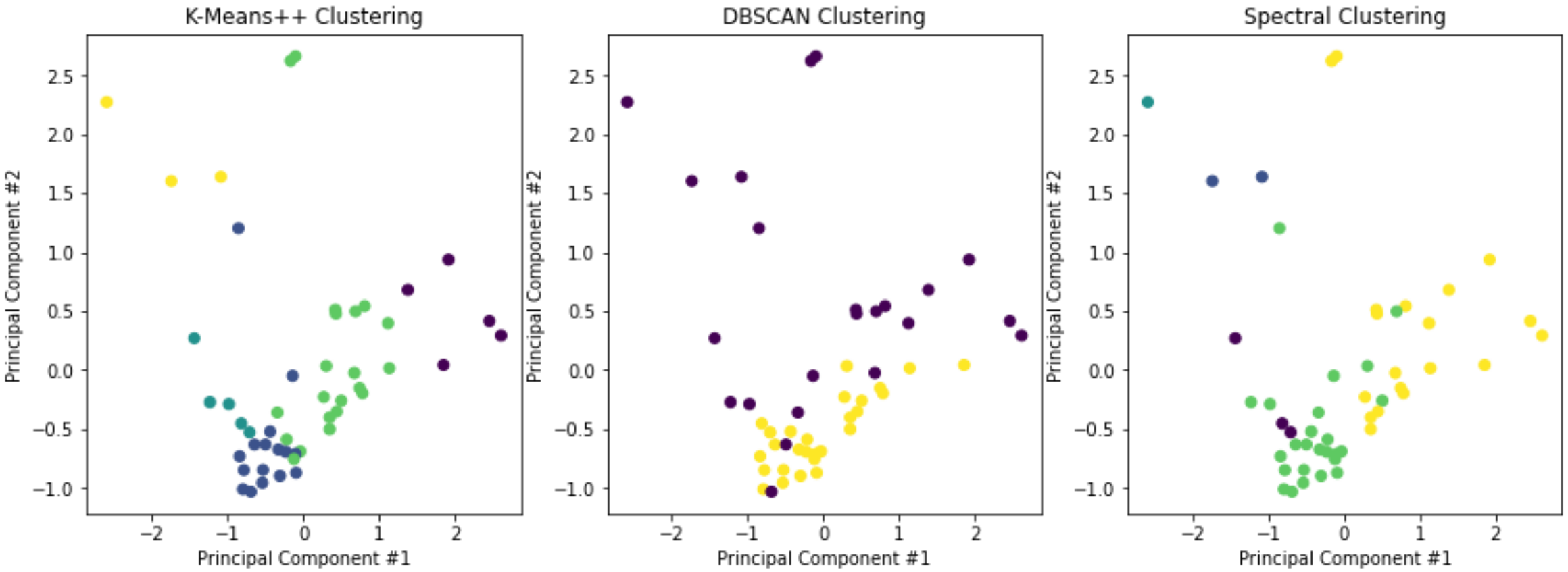
Overview

- Cluster merged patient FHIR/VCF data

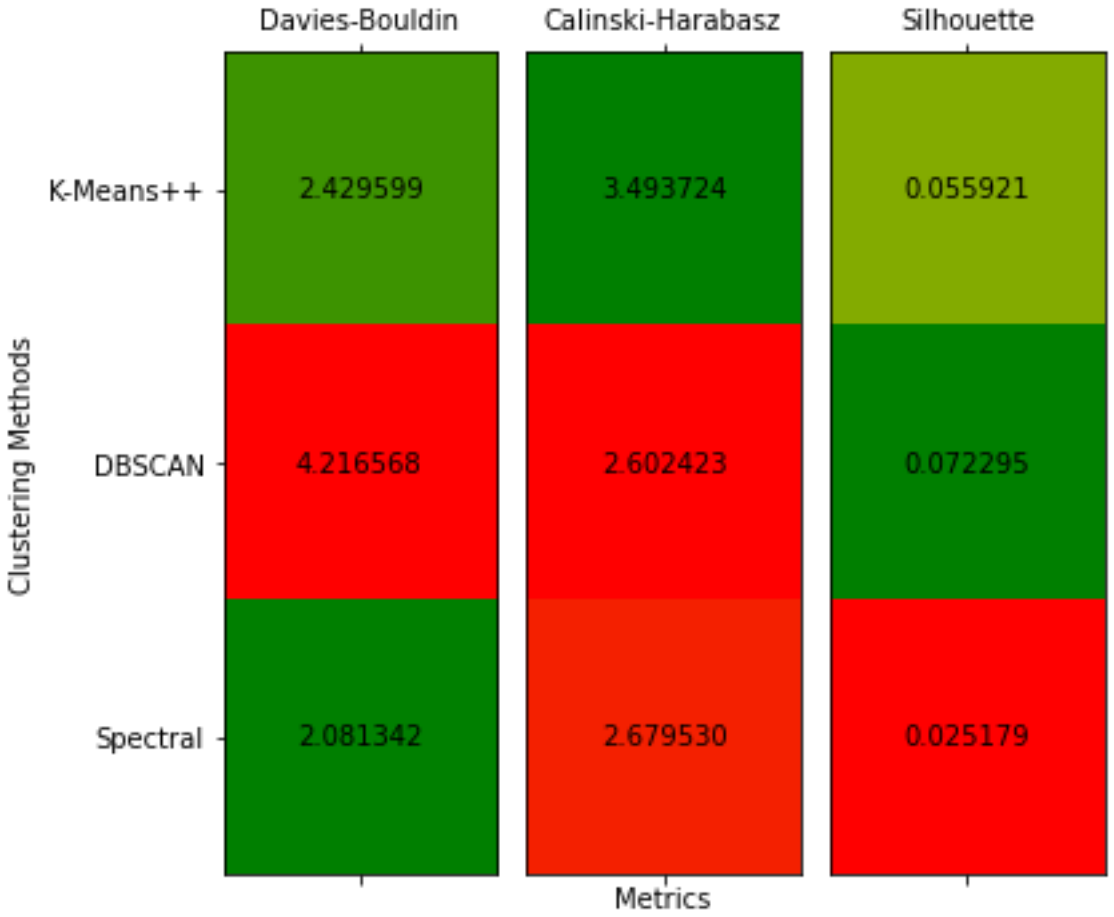
Motivation

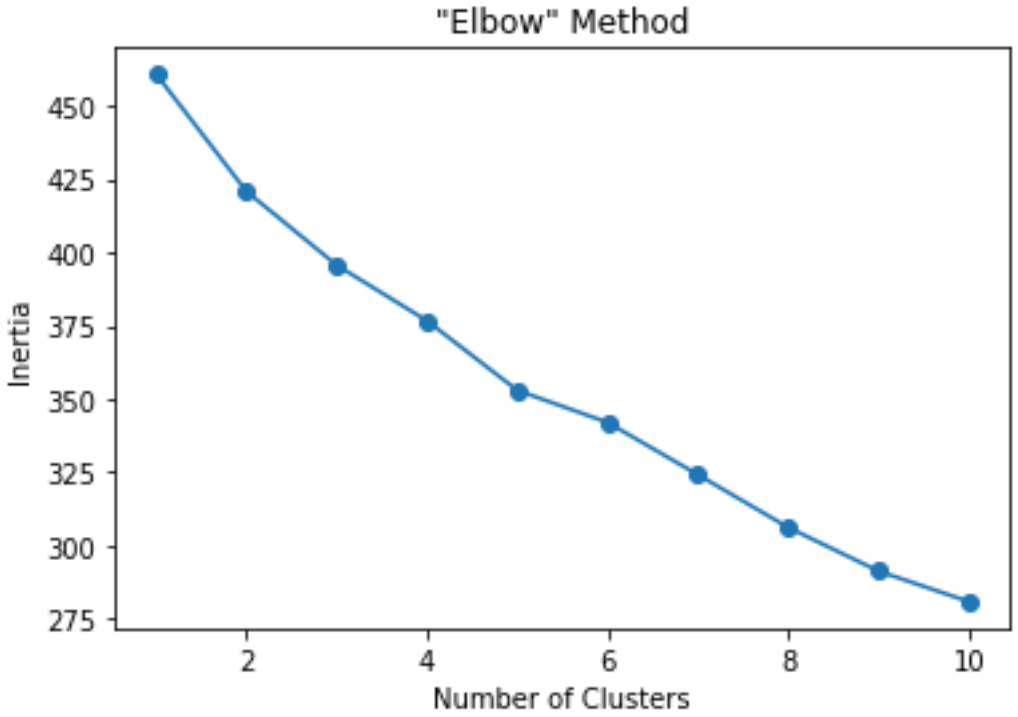
- Example machine learning application
- Demonstrates integration with popular machine learning libraries
- Requires merging FHIR/VCF data into pandas DataFrame

Application #2: Patient Clustering

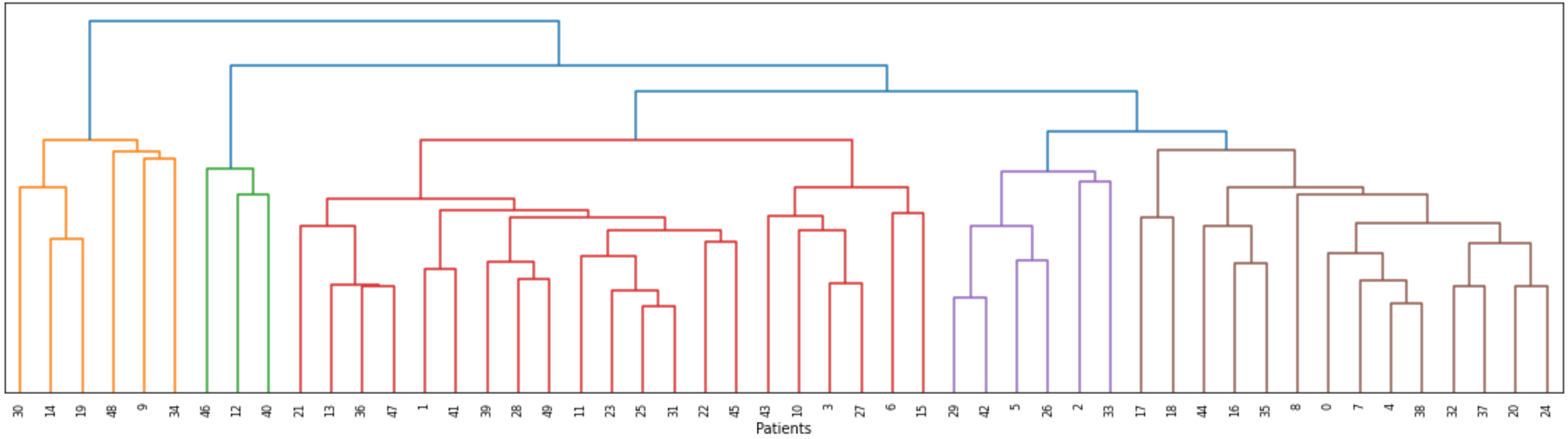


Application #2: Patient Clustering

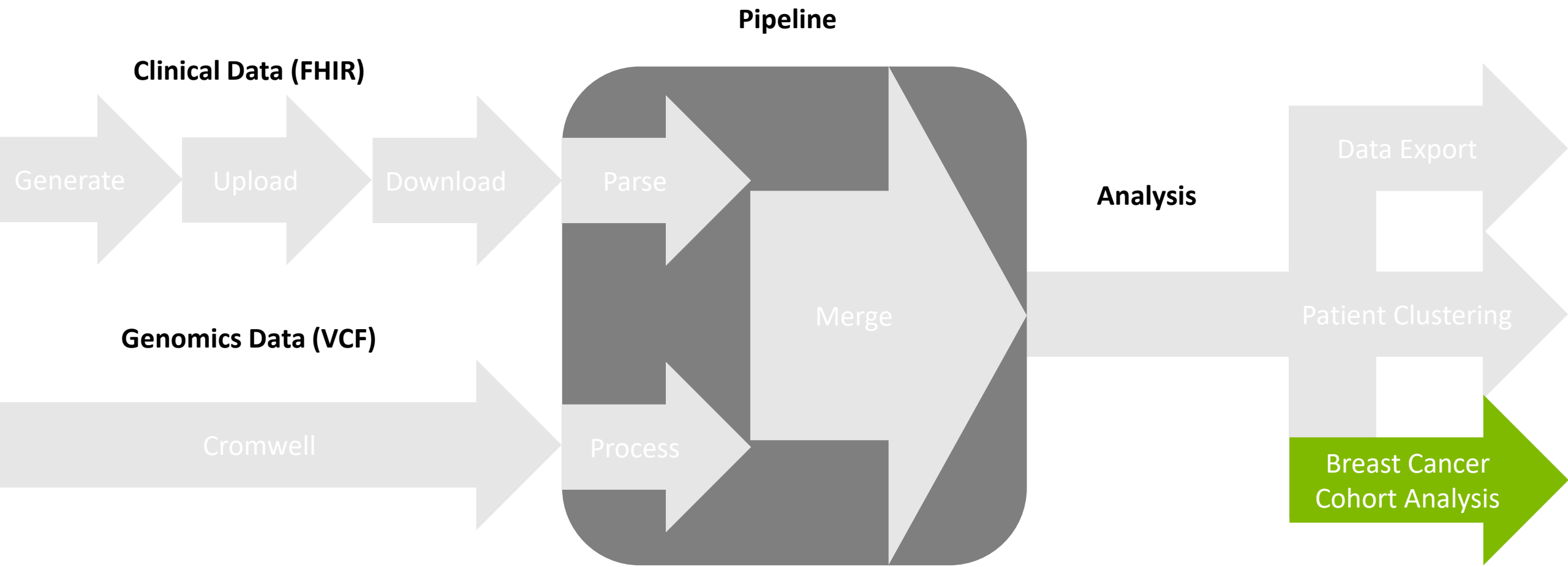




Application #2: Patient Clustering Dendrogram



Project Overview



Application #3: Breast Cancer Cohort Analysis

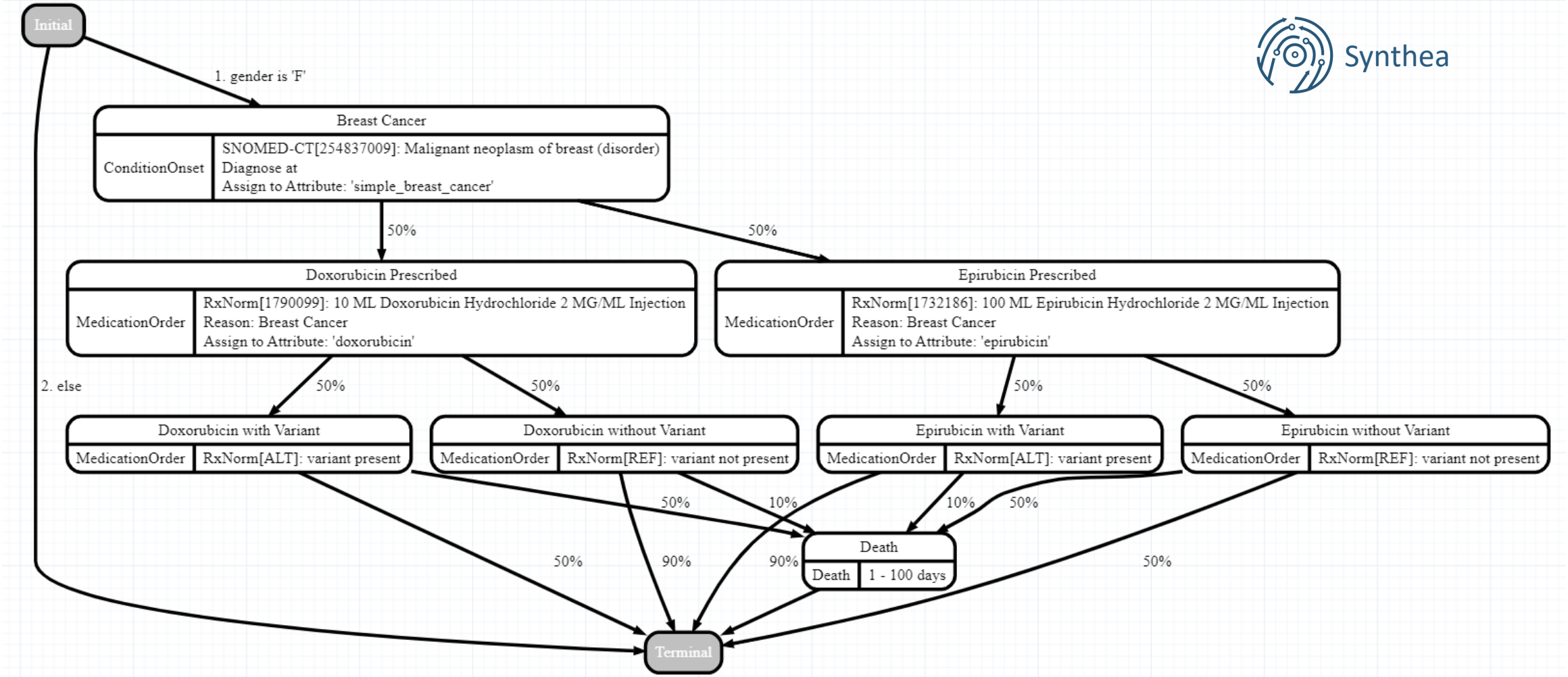
Overview

- Evaluate breast cancer patient outcomes, as a function of selected medication (FHIR) and presence of a specific variant (VCF)

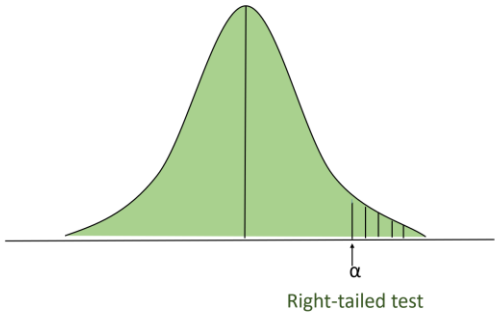
Motivation

- A pharmacogenetic study requiring combining genomic and clinical data
- Test development of custom Synthea modules for unique patient cohorts
- Demonstrate integration of Azure confidential compute on full application

Application #3: Breast Cancer Cohort Analysis

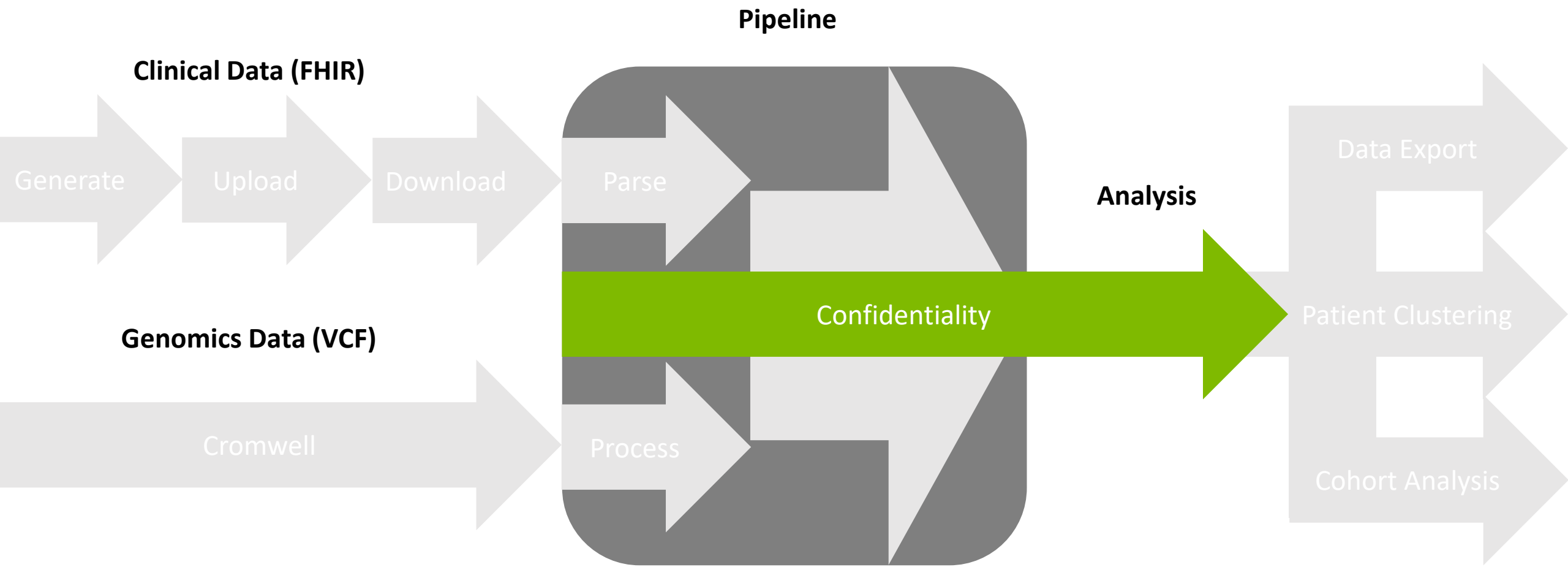


Application #3: Breast Cancer Cohort Analysis



There is a 40.683% chance Doxorubicin improves survivability for patients WITH Variant.
There is a 100.000% chance Doxorubicin improves survivability for patients WITHOUT Variant.
There is a 99.999% chance Epirubicin improves survivability for patients WITH Variant.
There is a 89.082% chance Epirubicin improves survivability for patients WITHOUT Variant.

Project Overview

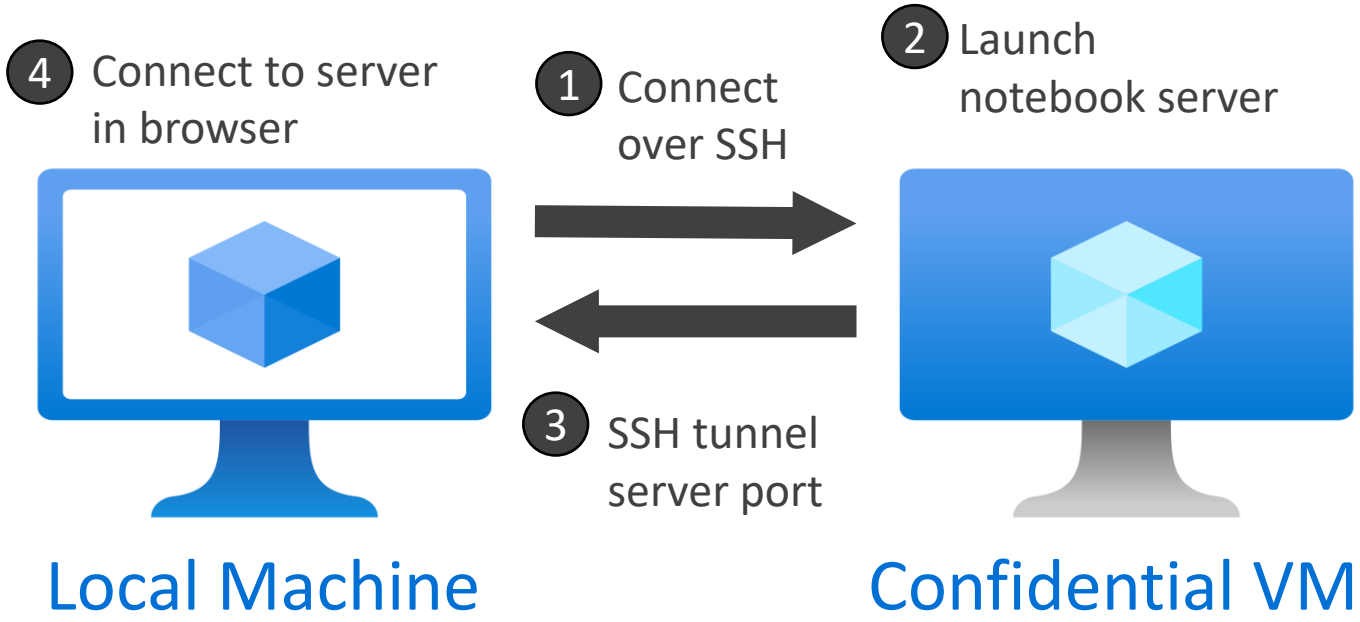


Confidential Computing: Azure VM Image

- Protection against persistent/advanced threats
- Offers confidentiality and integrity guarantees
- Trusted Launch
  - Secure Boot
  - virtual Trusted Platform Module (vTPM)
- Virtualization-based Security
- Hardware Enclaves
  - Intel Software Guard Extensions (SGX)
  - AMD Secure Encrypted Virtualization (SEV-SNP)

Confidential Computing: Pipeline

- Plan to release custom VM image



Blog Posts: Healthcare and Life Sciences Blog

App #1: Data Export

Convert Synthetic FHIR and PacBio VCF Data to parquet and...

Erdal Cosgun on Jul 21 2022 02:18 PM

Convert synthetic FHIR and PacBio data to parquet for further tertiary analysis!

App #2: Clustering

Data Science for Merged FHIR and PacBio VCF Data on Azure...

Erdal Cosgun on Aug 09 2022 02:20 PM

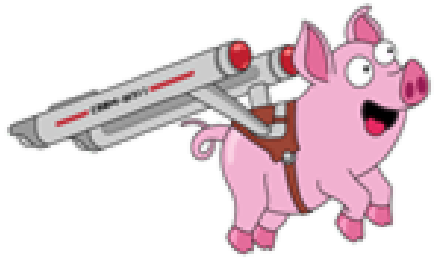
How to use data science for merged FHIR and Long Read Genomics sequencing data?

App #3: Cohort Analysis

Upcoming
Sept. 2022

Personal Learnings

- Azure
- Synapse
- FHIR
- Cromwell
- Synthea
- Microsoft!



Questions?

Thanks for listening!

After the internship, I'll be resuming my PhD in Ann Arbor, Michigan.

timdunn@umich.edu