

Census Scientific Advisory Committee Re-engineering Census' Construction Data Programs

Stephanie Studds, Division Chief, Economic Indicators Division

September 18, 2020

Shape
your future
START HERE >

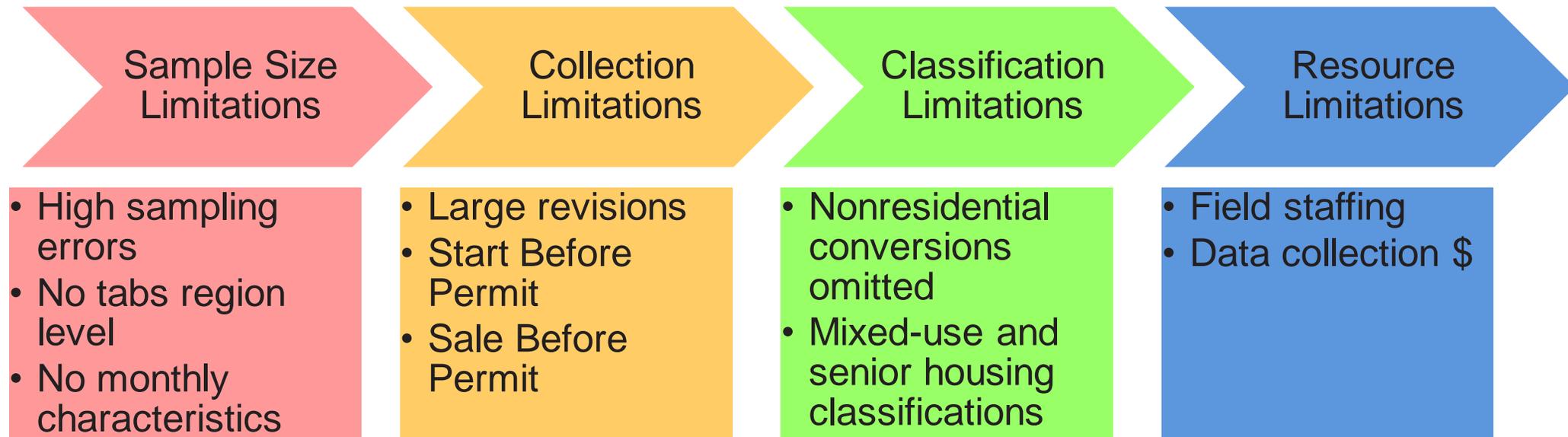
United States[®]
Census
2020

Current Programs

- Building Permits Survey (BPS)
- Survey of Construction (SOC)
 - Includes HUD-sponsored data on new residential sales, completions, and characteristics
- Construction Spending (also called Value Put in Place or VIP)
 - Construction Progress Reporting Surveys
 - Residential Remodeling data from the Consumer Expenditure Survey
- Rental Housing Finance Survey (sponsored by HUD)

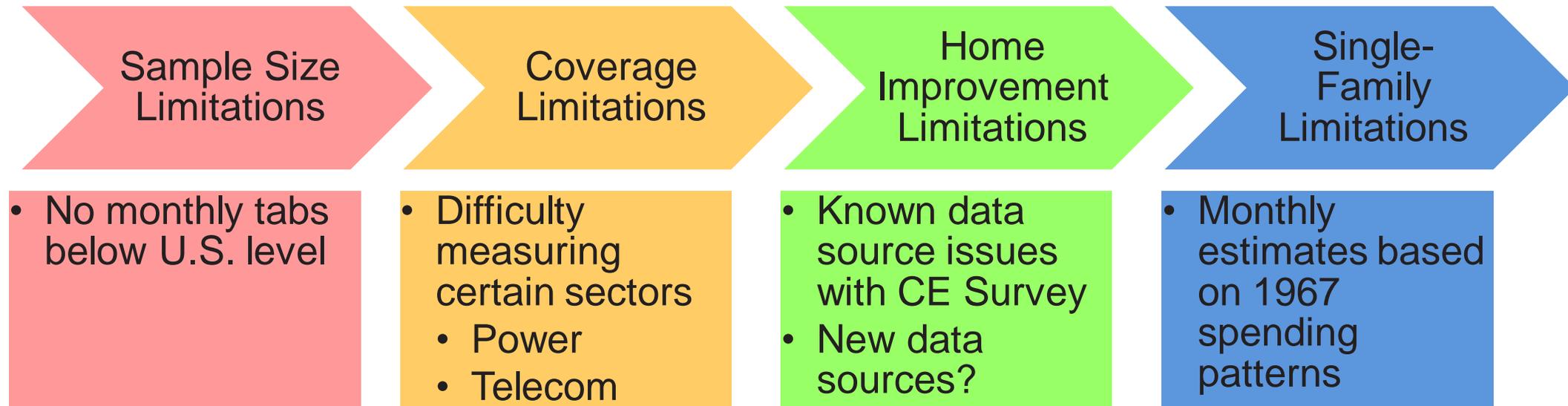
What we have heard...

... about New Residential Construction and Sales

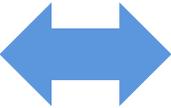


What we have heard...

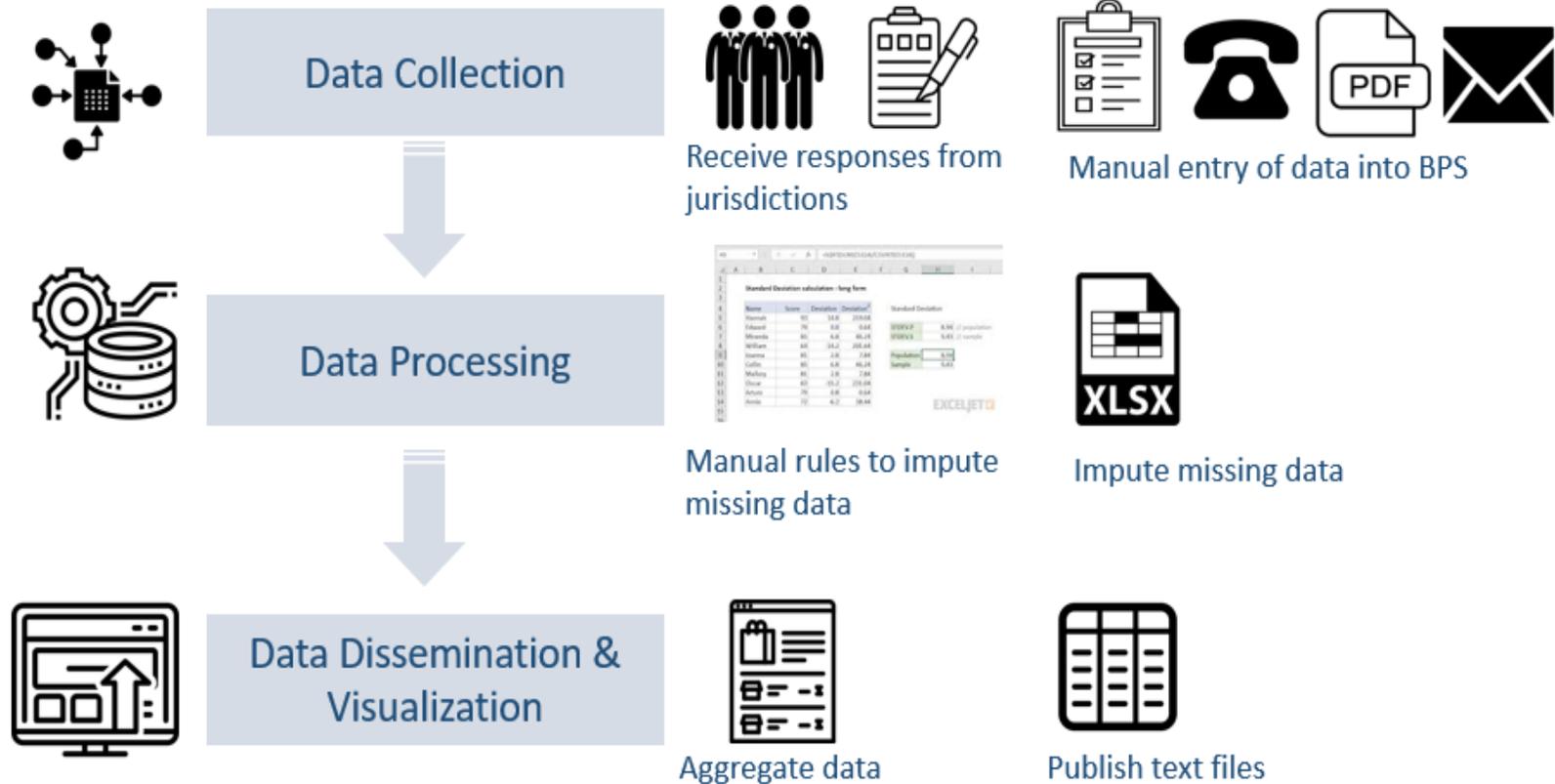
... about Construction Spending



Goals and Measures of Success

- Create new and improved products
 - Products on repair and improvements
 - Consider demographic and economic measures due to sector shocks such as weather events and financial crisis
 - Use alternative data sources
 - Accelerate delivery of data products
 - Increase organization flexibility and operational Efficiency
 - Innovate business processes and develop new methodologies for data ingestion, analysis, and dissemination
- 
- Reduction in the need for field collection
 - Reduction in respondent burden
 - Remain cost neutral across the construction programs
 - Protect the core value of the indicator and prevent its reconstruction by external entities

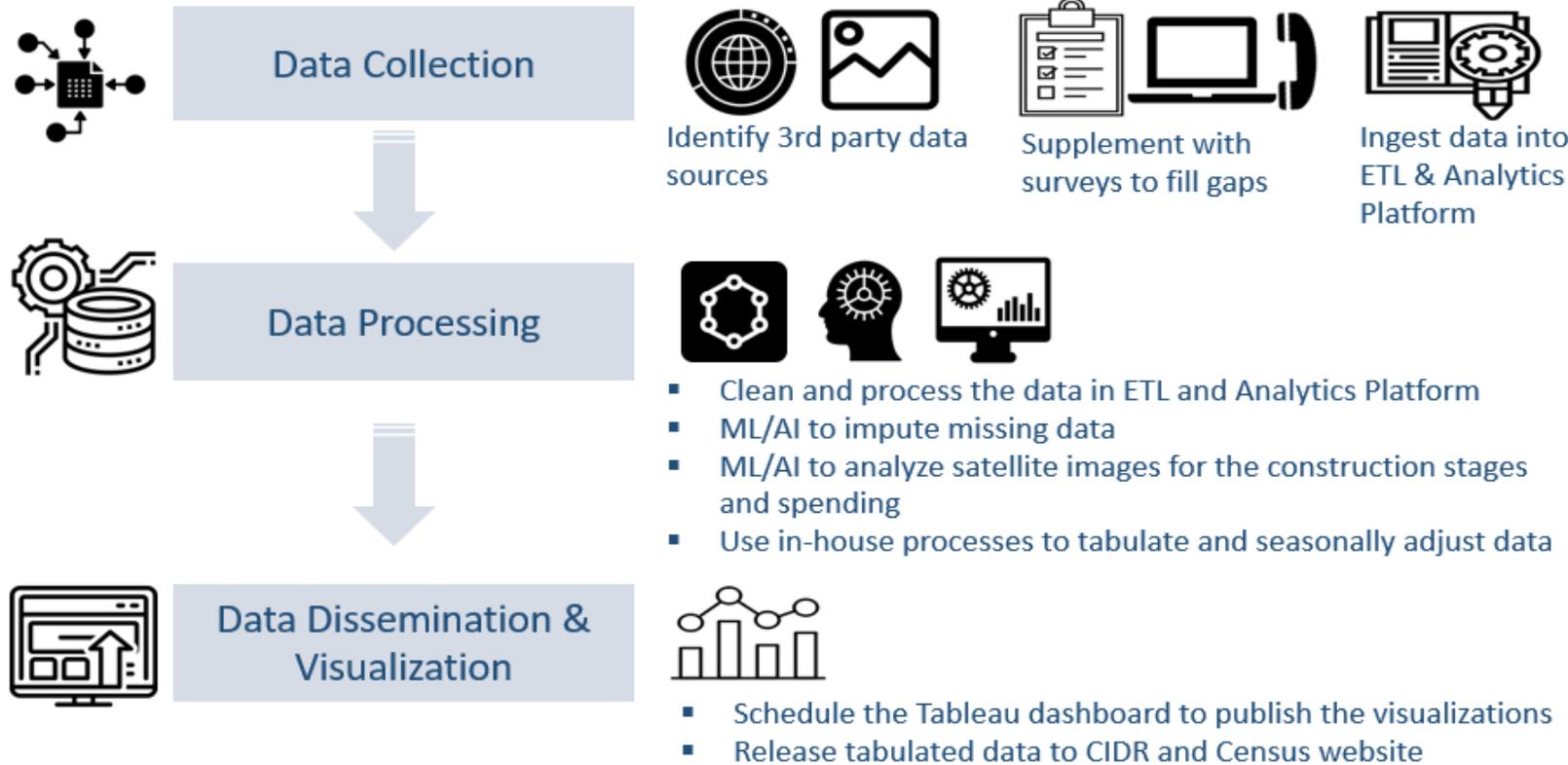
Current State



Challenges

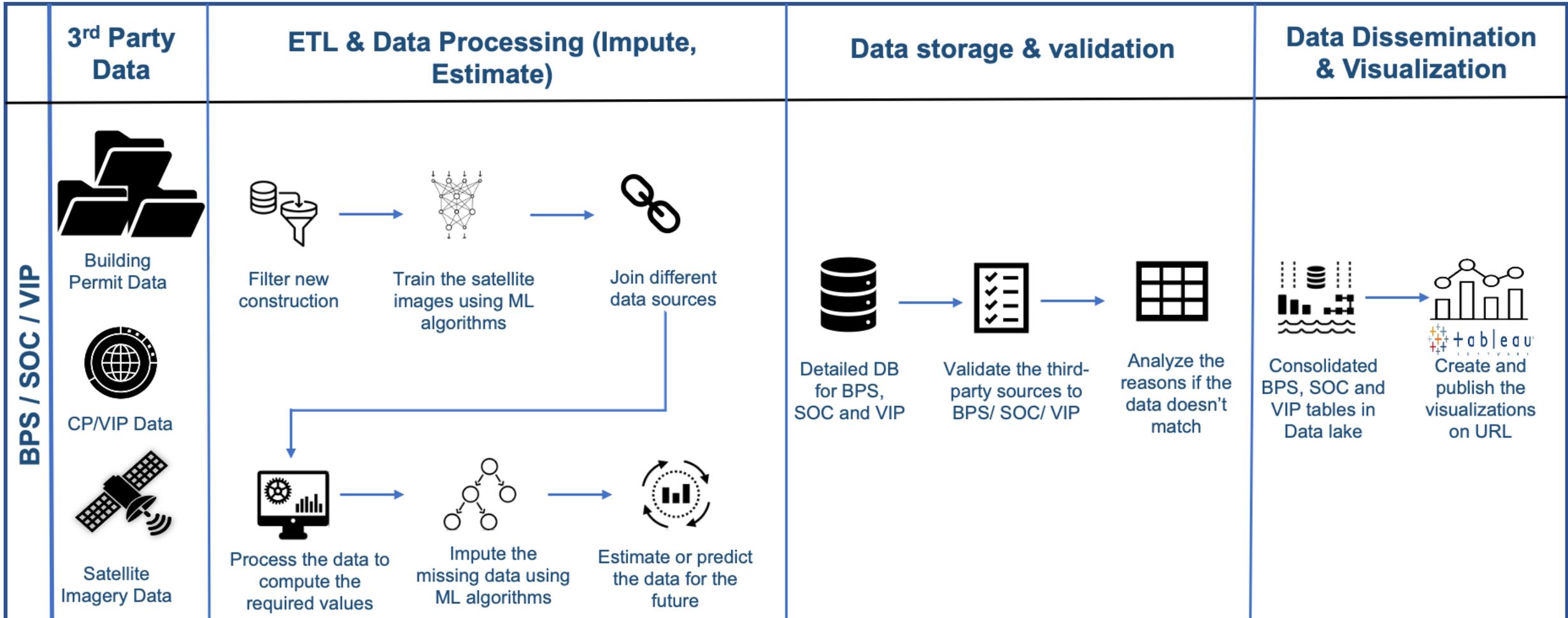
- Every program has its own data structure and system
- Relies heavily on componentized processing and dissemination
- Census staff waits for data to arrive at fixed points in time

Future State



- ### Benefits
- Near real-time data collection and processing
 - All encompassing data environment based on data lake architecture and enterprise level data ingestion
 - Analytics platform sits on Python codebase
 - Census data scientists will control the codebase
 - Ability to create and accelerate release of new data products

Methodology



Building Permit



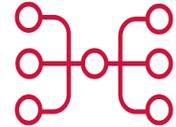
Identify features from BPS



Select critical features

- Number of Units
- Number of Buildings
- Permit Valuation

which needs evaluation from the vendor data for the Jurisdictions from TX, WA, OH, GA, CO



Map Jurisdictions from Vendors to Census BPS



Filter data for New Residential Construction



Categorize construction type:

- Cabins
- Duplexes & Twin Homes
- Single Family Homes
- Apartments & Condos
- Mixed Use
- Other Residential Structures



Evaluate Number of Units, Number of Buildings and Permit Valuation by construction type

- Formulate rules
- Validate rules with the math stats team
- Apply rules

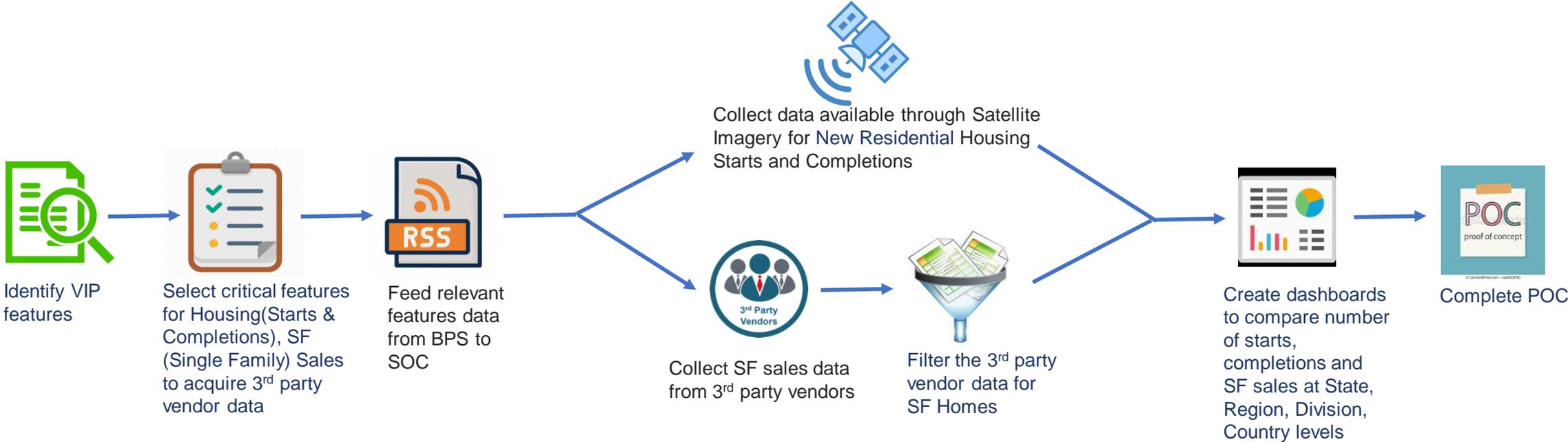


Create dashboards to compare Number of Units, Number of Buildings, Permit Valuation by Jurisdiction & Period

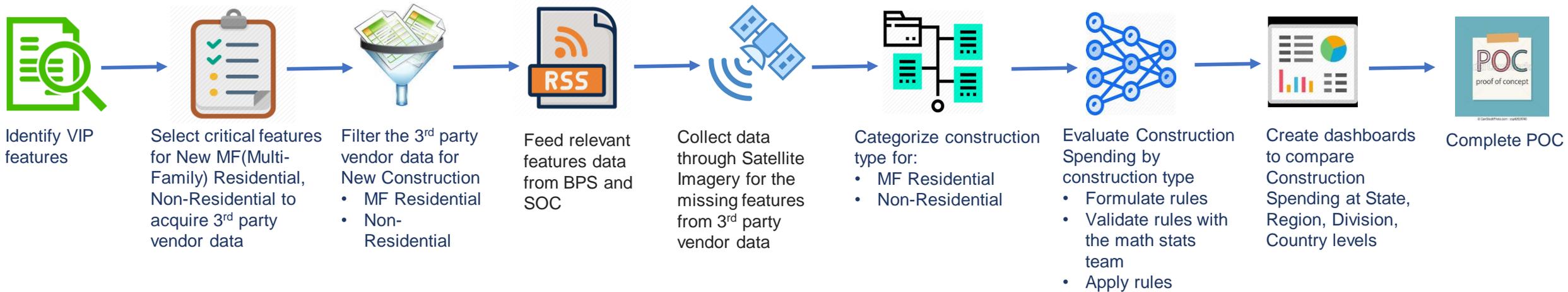


Complete POC

Construction Reengineering (SOC)

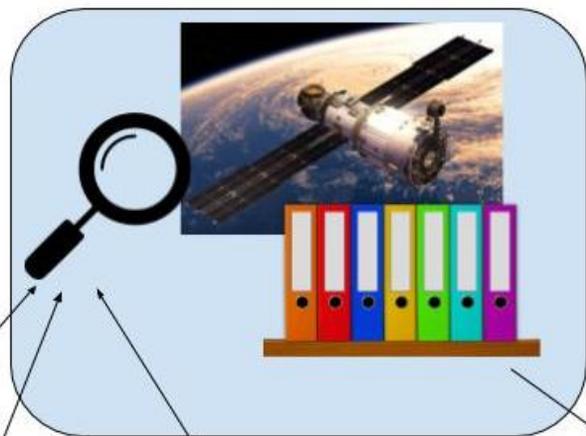


Construction Reengineering (CP/VIP)



From Permit Data to Satellite Images

SEARCHING PROCESS



INPUT

Filtered Permit Data

<i>permit_id,</i>	<i>lat,</i>	<i>lon,</i>	<i>permit_date,</i>
#111,	35.233232,	-156.9211,	02/24/2017,
#112,	32.43232,	-198.3332,	06/13/2017,
#113,	29.3232,	-102.9877,	09/06/2017,
#114,	44.113999,	-166.5671,	10/14/2017,
....

OUTPUT

Satellite Images

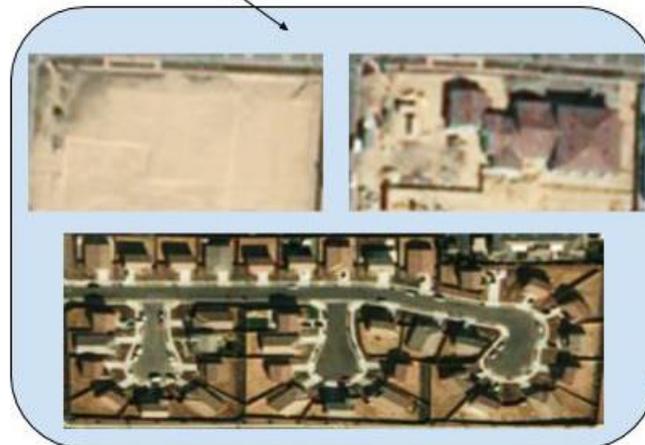


Image Resolution: 0.5m, 0.7m

Time range for searching Images
(Based on [Census statistics](#)):

- **Pre-constructions**
Images from 31 days to 1 year before the permit date.
- **Starts**
Images from 30 days to 4 months after permit date.
- **Completions**
Images from 9 months to 1 year after permit date.

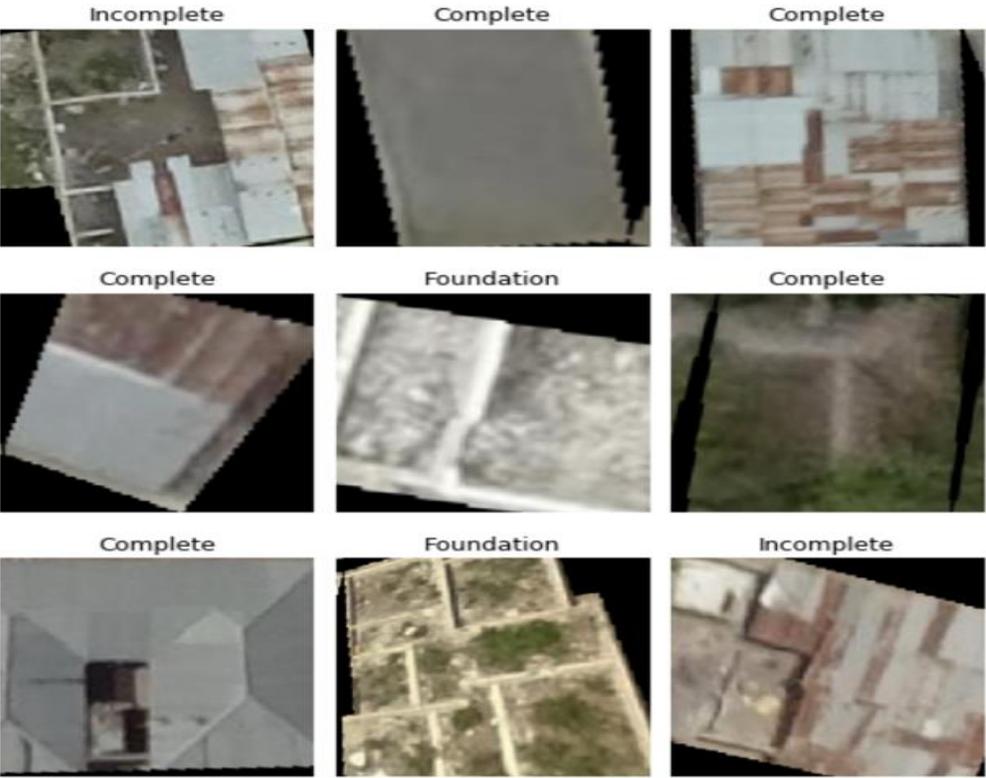
Model Training

Our team collaborated with StatsCanada to select the **most efficient methodology and tools for image classification.**

Two Convolutional Neural Networks (**CNN's**) were created based on different Deep Learning frameworks: Tensorflow and Pytorch.

Training was done using the satellite images of **pre-constructions, construction starts, and completions** derived from the permit data.

```
1 data.show_batch(rows=3, figsize=(7,8))
```



```
1 learn = cnn_learner(data, models.resnet34, metrics=error_rate)
```

```
1 learn.fit_one_cycle(6)
```

epoch	train_loss	valid_loss	error_rate	time
0/6			0.00%	00:00<00:00
2/38			5.26%	00:21<06:20 2.2095

Model Testing

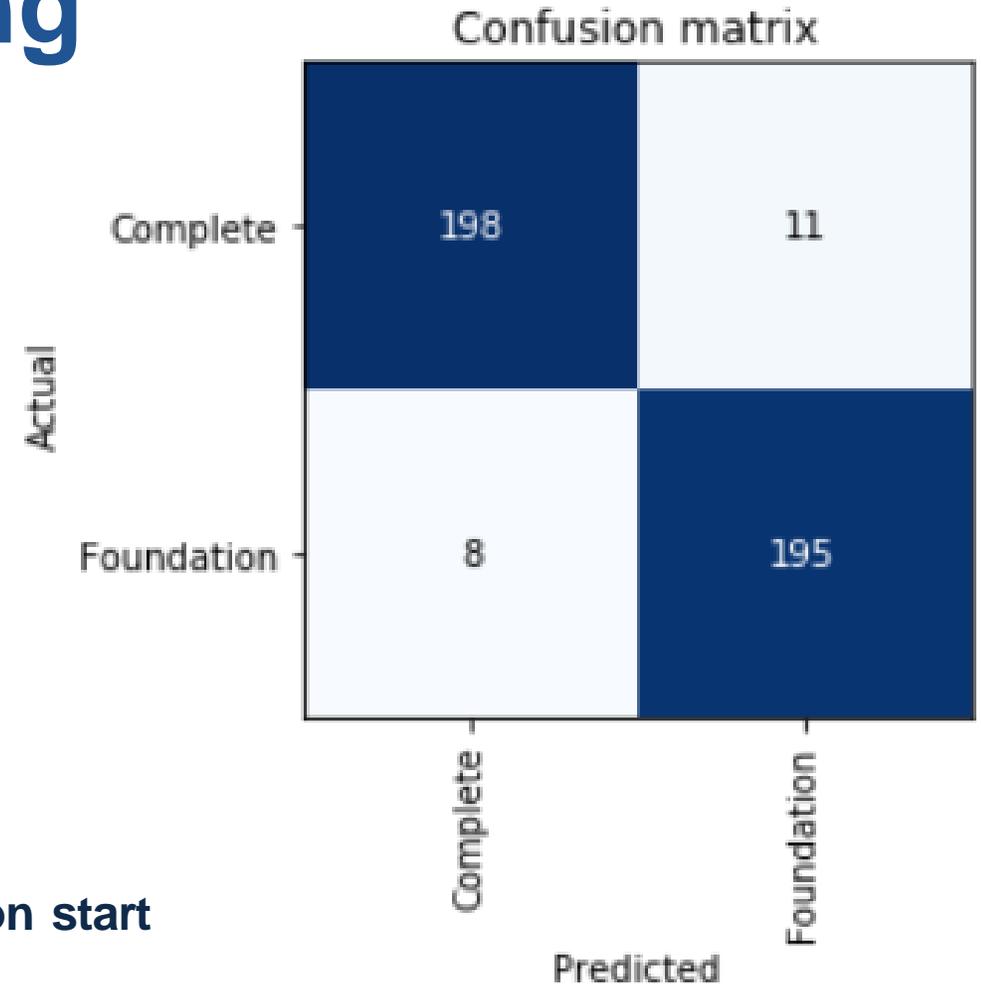
Current Accuracy - 95.5%

```
1 learn = cnn_learner(data, models.resnet34, metrics=error_rate)
```

```
1 learn.fit_one_cycle(10)
```

epoch	train_loss	valid_loss	error_rate	time
9	0.126794	0.114064	0.046117	03:06

Model accurately classifies any satellite image as a construction start (foundation) or as a completion



Hunting for Construction Starts

We will task the satellites monthly in areas that do not require building permits, “hunting” for Construction Starts: 59 county segments, approximately 12,683 sq miles

Next, we feed the tasked images to the Hunting Mode model in order to identify construction starts



Construction Change Detection

- **Change Detection** algorithm will be used to analyze the image changes in between construction stages

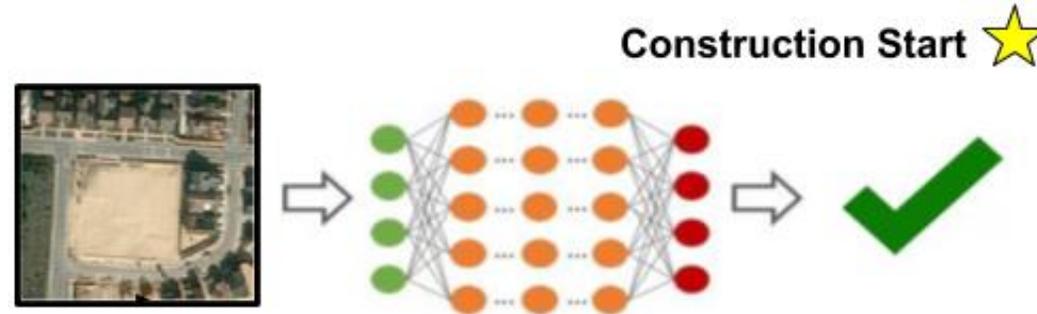
Starts



Completions



Construction Tracking



- Continuously task locations tagged as “starts” monthly and use change detection between the images to estimate Construction Spending and Completion Date



Construction Start ★

Learning from Others

- **CSAC Working Group**
- **Geography Division**
- **Statistics Canada**
- **Population Division**
- **Housing Statistics Users Group (HSUG)**
- **National Association of Home Builders (NAHB)**
- **Bureau of Economic Analysis (BEA)**
- **U.S. Department of Housing and Urban Development (HUD)**
- **Associated General Contractors of America (AGC)**

Questions to the Committee

- Can you discuss opportunities to communicate this effort to industry leaders and other principal stakeholders?
- What challenges do you see with the project, including time series breaks, methodological differences, blending data, etc.?
- What current and/or new products would you prioritize for the project? What industry data needs are most critical?
- Are there any additional data sources that we should look into?
- Are there additional modeling techniques we should investigate?