Bridging the Digital Divide from Orbit:

Using Data Engineering to Map Satellite Connectivity Impact on Human Development

Greg Sullivan

Washington University in St. Louis

DATA 5035: Data Engineering

Professor Paul Boal

February 2026

Executive Summary

Approximately 2.6 billion individuals around the world currently lack access to the internet, representing nearly one-third of the global population. Research demonstrates that connectivity is not merely convenient - it is transformational. A 10% increase in broadband penetration correlates with 0.77% to 1.38% GDP growth (International Telecommunication Union [ITU], 2020), while mobile connectivity drives measurable gains in employment and poverty reduction (Chiplunkar & Goldberg, 2024).

Low Earth Orbit (LEO) satellite constellations such as Starlink, OneWeb, and Amazon's Kuiper all promise to deliver high-speed internet anywhere on Earth. This project applies data engineering techniques to answer a critical question: ***Where will satellite broadband service have the greatest impact on human development?***

Personal Context: Why This Matters to Me

As Chief Information Officer (CIO) of Carnival Corporation, the world's largest cruise operator, I was responsible for satellite connectivity across our entire fleet of 100+ vessels and land-based properties in remote locations. I experienced firsthand how connectivity, or lack thereof, impacts the lives of people.

This is not an academic exercise for me. It reflects challenges I managed professionally for years. On many occasions I had the opportunity to travel by ship to countries with little to no internet connectivity. I noticed discernible differences in education and poverty levels, for example. Most of all, I was touched by the irony that guests and crew on our ships had better internet connectivity on our ship than those living on the shores we visited.

Another good example of the impact of limited connectivity is the case of our Alaska lodges. These are amazing properties that offer incredible experiences to our guests, and our crew enjoy the hard work during the summer season. However, with limited connectivity options, our employees had little to do in the few off-hours they had. They couldn't video call family back home, access online education, manage personal finances, or participate in the digital economy. For crew members on months-long contracts away from home, this isolation takes a real toll on morale and mental health. Connectivity is not a luxury - it's a lifeline.

Guests constantly complained about poor connectivity on our ships and at our Alaska lodges and hotels. This wasn't about scrolling social media - they couldn't share once-in-a-lifetime experiences with loved ones, handle urgent matters from home, access travel information or stay in touch with their office (we measured this and the results were astounding). In an era of seemingly ubiquitous connectivity, these gaps directly impacted guest satisfaction scores, crew wellbeing and brand perception.

Preliminary Finding: Early Data Supports the Thesis

In my prior classes I worked on projects (individually and as teams) where we presumed data sources were of high quality BEFORE we proposed and undertook a project. I've learned the hard way to do some early exploration of possible data sources in advance of any proposed project. So, I've already validated several candidate data sources and tested the core hypothesis. The results are compelling.

The Human Development Index (HDI) is a composite measure published by the United Nations Development Programme (UNDP) that captures three dimensions of human wellbeing: health (life expectancy), education (years of schooling), and standard of living as measured by gross national

income per capita (GNI). HDI values range from 0 to 1, with higher values indicating greater development. The UNDP classifies countries into four tiers: Low (below 0.550), Medium (0.550 to 0.699), High (0.700 to 0.799), and Very High (0.800 and above). By joining World Bank connectivity indicators (World Bank, 2024) with HDI data (UNDP, 2024), I identified 17 countries where less than 30% of the population has internet access AND human development falls in the Low category (below 0.550). These nations represent the "sweet spot" for satellite broadband impact.

Connectivity-Development Correlation Matrix

Connectivity	Development	Countries	Average Internet	Average HDI	Average GNI
Low (<30%)	**Low HDI (<0.55)**	17	18.7%	0.474	$2,052
Medium (30-69%)	Medium HDI	31	47.8%	0.617	$5,895
High (70%+)	High HDI (0.7+)	101	88.0%	0.847	$39,026

Key insight: Zero countries have high connectivity AND low HDI. The correlation is clear - and the opportunity is measurable.

Why Data Engineering?

This question cannot be answered by querying a single database or downloading one report. It requires **integrating heterogeneous data sources** that were never designed to work together. Here are a few examples:

- Satellite orbital data uses NORAD catalog numbers and two-line element sets
- World Bank uses ISO country codes with varying temporal coverage by indicator
- UNDP publishes country names that don't always match World Bank conventions
- Union of Concerned Scientists (UCS) Database tracks satellite operator ownership that must be reconciled with actual orbital data

Data engineering provides the discipline to ingest, validate, transform and join these sources into a coherent analytical foundation. Without proper Extract-Transform-Load (ETL) pipelines, data quality checks, and transformation logic, any analysis built on these sources would be unreliable.

Proposed Data Sources and Why They Matter

Each data source serves a specific analytical purpose. All four have been downloaded from trustworthy sources, loaded to Snowflake staging tables and validated.

Source	Data	Records	Why It Matters
World Bank	Development indicators	9,663	Provides connectivity rates and economic metrics by country over 10 years - essential for identifying gaps and trends
UNDP HDI	Human Development Index	208	Composite development measure (health, education, income) that quantifies human impact beyond GDP alone

CelesTrak	Starlink orbital data	9,548	Real-time constellation positions with inclination data to calculate coverage zones by latitude
UCS Database	All satellites	7,562	Operator ownership and purpose data enabling competitive analysis (Starlink vs OneWeb vs others)

Proposed Analytics

The preliminary finding demonstrates the connectivity-development correlation. The full project will integrate all four data sources (and any others we may find to be relevant) to explore the following analyses:

Analysis	Description	Data Sources
Coverage Gap Analysis	Map satellite coverage zones (based on inclination bands) against target countries. Do our 17 low-connectivity countries fall within Starlink's reach?	CelesTrak + World Bank
Constellation Comparison	Compare Starlink (9,548), OneWeb (589), and other operators by coverage area, growth trajectory, and target markets.	UCS + CelesTrak
Trend Analysis	Using 10 years of World Bank data: are target countries improving connectivity on their own, or stagnant? Where is satellite most needed vs. terrestrial buildout?	World Bank (time series)
Affordability Threshold	At what GDP per capita does satellite internet become viable? Which countries are approaching this threshold?	World Bank + UNDP
Impact Scoring Model	Rank countries by composite score: connectivity gap × development need × satellite coverage × affordability proximity.	All four sources
High-Latitude Case Study	Deep dive on Alaska/Arctic regions: how do 70° and polar-orbit satellites address guest experience, crew welfare, and operational safety gaps?	CelesTrak + UCS

Anticipated Challenges

Based on preliminary data validation, several challenges will require careful attention:

Entity Resolution: Country names are inconsistent across sources. World Bank uses "Côte d'Ivoire" while UNDP may use "Ivory Coast." Hong Kong appears as "Hong Kong, China (SAR)" in one source and "Hong Kong SAR, China" in another. We will need fuzzy matching or a master country crosswalk table.

Data Currency: The UCS Satellite Database (UCS, 2024) shows 3,950 Starlink satellites while CelesTrak shows 9,548 - a gap of approximately 5,600 satellites. UCS updates less frequently. We must document source currency and decide which to trust for different analyses.

Coverage Zone Calculation: Converting orbital inclination to geographic coverage requires assumptions about minimum elevation angles and constellation density. A 53° inclination satellite can serve latitudes up to approximately 53° - but actual coverage depends on the number of satellites and their spacing.

Temporal Alignment: World Bank data spans 2015-2023 with varying coverage by indicator. Some countries have gaps. HDI is a 2023 snapshot. Satellite data is near-real-time. Joining across different time horizons requires careful documentation of assumptions.

Sparse Data: Literacy rate (SE.ADT.LITR.ZS) has only 749 observations vs. 2,385 for life expectancy - many countries don't report annually. Analysis must account for missing data without introducing bias.

Scope Management and Risk Mitigation

This project is ambitious by design. Deliverables are prioritized to ensure a strong outcome regardless of time constraints. The following tiered approach ensures we deliver value even if complexity forces scope adjustments:

Priority	Deliverable	Rationale
Must Have	Integrated country-level dataset joining World Bank + UNDP HDI with validated DQ checks	Foundation for all analysis; already 80% complete from validation work
Must Have	Connectivity-development correlation analysis with target country identification	Core thesis validation; preliminary finding already demonstrates feasibility
Should Have	Satellite coverage mapping using inclination-based latitude bands (simplified model)	Adds satellite dimension; uses straightforward inclination to latitude relationship
Should Have	Constellation comparison (Starlink vs OneWeb) from UCS operator data	Competitive context; primarily aggregation and grouping - low complexity
Nice to Have	Precise geospatial coverage modeling with elevation angles and satellite density	Most complex component; can simplify to latitude bands if time-constrained
Nice to Have	Affordability threshold regression model	Advanced analytics; can present as descriptive statistics if modeling proves difficult

Key Risk Mitigation: If precise geospatial coverage modeling becomes too mathematically complex, we will use a simplified approach: satellites with inclination X° cover latitudes from -X° to +X°. This is technically an approximation but sufficient for identifying which target countries fall within constellation coverage zones. The simplified model still answers the core business question while avoiding orbital mechanics complexity that could derail the project timeline.

Team Approach

If selected as a team project, the work naturally divides into functional areas. Team size will determine whether these are individual roles or combined responsibilities:

Functional Area	Responsibilities	Deliverables
Data Ingestion	Build and maintain ETL pipelines; manage staging tables; handle incremental updates; document source refresh procedures	ETL scripts; staging table schemas; source documentation; refresh runbooks
Data Quality	Design DQ test suite; build entity resolution for country matching; manage schema changes; maintain data dictionary	DQ checks; country crosswalk table; data dictionary; lineage documentation
Analytics	Develop scoring model; perform coverage analysis; build analytical queries; document methodology	Impact scoring model; analytical SQL/Python; methodology documentation
Visualization	Consider Tableau dashboards and Flourish storytelling assets; build presentation materials	Interactive dashboard; presentation visuals; documentation

Delivery Methodology

Sprint Prioritization: Work will follow the Must Have → Should Have → Nice to Have priority tiers. Each week focuses on completing the highest-priority incomplete items before advancing. This ensures a demonstrable deliverable exists at any checkpoint.

Integration Testing: Before any analytical work begins, we will validate end-to-end data flow: staging tables → transformation → joined analytical tables → visualization layer. A simple "smoke test" query joining all four sources will confirm integration is working before building complex analytics on top.

Schema Evolution: Source data structures may change (World Bank adds indicators, CelesTrak adds fields). Staging tables will use flexible VARCHAR columns where appropriate, with transformation logic handling type casting. Schema changes may be documented in a log file, and downstream dependencies (views, dashboards) will be tested after any structural update.

Weekly Syncs: Brief coordination meetings will ensure integration points are aligned. This is particularly for handoffs between data ingestion, quality validation, analytics, and visualization. GitHub will be used for code management.

Data Engineering Lifecycle Alignment

This project maps directly to the data engineering lifecycle stages:

Lifecycle Stage	Project Application
Generation	Data originates from four external sources: World Bank API, UNDP downloads, CelesTrak API, and UCS database exports. Each has different generation patterns (API vs. file download) and update frequencies. Additional data sources may be identified, as well.
Storage	Snowflake serves as the cloud data warehouse. Staging tables hold raw ingested data; analytical tables will hold transformed, joined data ready for querying. Depending on team preference, Databricks may be used.
Ingestion	Python scripts, where possible, fetch from APIs and parse downloaded files. Pandas handles transformation from source formats (JSON, CSV, TSV, Excel) to data frames, which are then loaded via Snowpark.

Transformation	Wide-to-long pivots (World Bank), header cleanup (UNDP), orbital calculations (CelesTrak), column standardization (all). Entity resolution joins sources on country. Undoubtedly, others will be uncovered as the project work unfolds. DQ checks validate at each stage.
Serving	Final analytical tables support Tableau dashboards, Flourish animations and ad-hoc SQL queries. Pre-aggregated views optimize common access patterns (by country, by region, by connectivity tier).

Technical Stack

Platform: Snowflake (cloud data warehouse with Snowpark Python integration), but possibly Databricks (again, depending on team preference)

Languages: Python (ingestion, transformation), SQL (analysis, DQ checks)

Visualization: Tableau (interactive dashboards), Flourish (storytelling animations)

Version Control: GitHub (code, documentation, data dictionaries)

Expected Deliverables

1. **Integrated dataset** linking connectivity, development, and satellite coverage by country/region
2. **Impact scoring model** ranking countries by potential benefit from satellite broadband
3. **Interactive dashboard** allowing exploration of connectivity and human development relationships
4. **Human impact case study** demonstrating real-world, fact-based benefits of internet connectivity
5. **Technical documentation** of ETL pipelines, data quality framework, and analytical methodology

Conclusion

The preliminary analysis has already proven the thesis: connectivity and development are correlated, and we can identify specific countries where satellite broadband could be transformational. This project will build the data infrastructure to quantify that opportunity and communicate it compellingly.

The 17 target countries identified - representing hundreds of millions of people - are not statistics. They are communities where connectivity could mean access to education, healthcare information, economic opportunity, and connection to the wider world.

And closer to home, this analysis applies to places like Caribbean islands and Alaska where I witnessed the impact of connectivity gaps firsthand. When a crew member can finally video call their family after months away, or a ship captain receives real-time weather data that prevents a dangerous situation, or a guest can share their glacier experience with loved ones - that's what bridging the digital divide means in human terms.

Data engineering can help us understand where to focus that effort - and measure the impact when we do.

References

Chiplunkar, G., & Goldberg, P. K. (2024). The employment effects of mobile internet in developing countries (NBER Working Paper No. 30741). National Bureau of Economic Research. https://www.nber.org/papers/w30741

International Telecommunication Union. (2020). How broadband, digitization and ICT regulation impact the global economy. ITU Publications. https://www.itu.int/dms_pub/itu-d/opb/pref/D-PREF-EF.BDR-2020-PDF-E.pdf

Kelso, T. S. (2024). CelesTrak NORAD general perturbations element sets [Data set]. CelesTrak. https://celestrak.org/NORAD/elements/

Minges, M. (2015). Exploring the relationship between broadband and economic growth (World Development Report 2016 Background Paper). World Bank. https://documents.worldbank.org/en/publication/documents-reports/documentdetail/178701467988875888/

Union of Concerned Scientists. (2024). UCS satellite database [Data set]. https://www.ucsusa.org/resources/satellite-database

United Nations Development Programme. (2024). Human development index [Data set]. UNDP Human Development Report Office. https://hdr.undp.org/data-center/human-development-index

World Bank. (2024). World development indicators [Data set]. World Bank Group. https://databank.worldbank.org/source/world-development-indicators

Appendix

I took one quick look at a visualization based on the impact of internet access to underdeveloped countries. This was built from the above data and imported into Flourish. It provides an example of a possible visualization for these data sets and is intended not to draw a conclusion, but to show how visualizations may be approached. Values are 1-5, with 5 representing the highest possible impact to human development from the addition, or enhancement, of internet connectivity to that country.