1. Science: Project Motivation and Impact

This project, called **Imola (Intelligent Map recOgnition LAb),** aims to produce valuable historical geographic information to open new research opportunities related to large geographic areas and long time periods in the social, behavioral, and economic (SBE) sciences. The goal is to establish **a new large-scale database of historical road networks with corresponding data attributes (e.g., road types) and comprehensive uncertainty measures** from the US Geological Survey (USGS) topographic map series, which contains a total of 178,000 sheets between 1884 and 2006 (e.g., Figure 1)**.** The project will also produce case studies to demonstrate the potential of the new database in the SBE research.

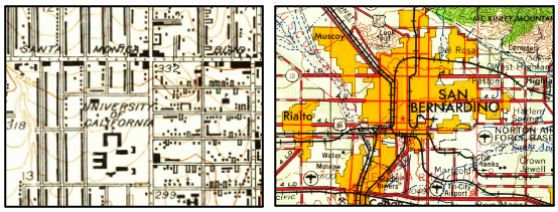

Figure 1. *Example historical USGS topographic maps.* ***Left:*** *Light-duty roads (black), Los Angeles, CA (1:24,000), 1931;* ***Right:*** *Urban areas (orange), San Bernardino, CA (1:250,000), 1966*

Longitudinal, detailed data on the states of landscapes and human activities at the continental scale are essential to answering critical questions in the SBE sciences. However, **for the periods before the 1970s (before Landsat satellite imagery), such data only exist on printed map sheets**. They were created, for example, by the USGS, which produced over 178,000 topographic map sheets between 1884 and 2006. Recently, the USGS National Geospatial Program (NGP) has scanned and made these historical paper maps publicly available. According to the USGS, in the US, these topographic maps "*portray both natural and manmade features. These maps show and name works of nature including mountains, valleys, plains, lakes, rivers, and vegetation. They also identify the principal works of man, such as roads, boundaries, transmission lines, and major buildings.*" **These maps have exceptional scientific, cultural, and societal value because they hold high-resolution data collected and mapped at high, well-documented accuracy standards (e.g., for data collection and consistency, cartographic techniques and quality control mechanisms).** They provide the best continuous, pre-satellite imagery for most natural and human landscape features in the US from their inception through the early 2000s and, thus, document 100+ years of detailed evolution of continental-scale landscapes, including landscape change, urban growth and sprawl, persistent underdevelopment of other urbanized areas, coastline erosion and modification, draining of swamps, growth and shrinkage of forest cover, and changes in river courses. **These topographic features depicted on historical maps can enable fundamentally new scientific research in the SBE (and natural) sciences related to large, national-scale areas and long time periods that would not be possible otherwise.** The potential of these historical USGS topographic maps has not been realized because the material has only recently become available and is stored as scanned images. While humans can easily understand these scanned images, systematic exploration of their contents requires **robust, efficient data extraction, derivation of uncertainty estimators, and conversion into a format that allows meaningful analysis in a geographic information system (GIS)** (see, e.g., [9, 10, 13, 38]).

This proposal will develop a new large-scale spatiotemporal database, called **US1884+**, of attributed long-term, national-scale detailed road networks from 1884 to 2006. **Historical road network data are of great interest, e.g., for urban planning and economic and transportation research; however, they are rarely available** [5]**.** US1884+ has the potential to unlock unique research opportunities as changes in road networks are determinants of the evolution of urban systems, socioeconomic and demographic trends, and land development; this kind of research remains severely limited without road databases covering large geographic extents and long time periods. To illustrate the potential need, searching Google Scholar using the keywords "historical maps", "change", and "analysis" returns about 25,300 results, of which around 21,480 appeared after 2001 (Figure 2). These results include top-ranked scientific journals such as Nature, Remote Sensing of Environment, Ecology, Global Change Biology, Landscape and Urban Planning, Landscape Ecology, Land Use Policy, Journal of Glaciology, Environment and Planning B, Forest Ecology and Management, Agriculture, Ecosystems &

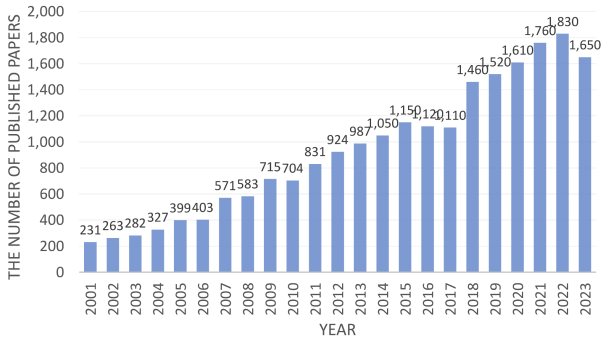

Figure 2. *The number of published papers per year using the search keywords "historical maps", "change," and "analysis" on Google Scholar*

Environment, Applied Geography, Journal of Coastal Research and River Research and Applications. **This list indicates the vast diversity of studies using historical maps for change analysis and describes an incredibly broad research community that will benefit from US1884+.** US1884+ will extend the temporal dimension of landscape research, such as reported in the above-listed publications, to the beginning of the 20th century for most parts of the US and thus open new forms of unseen data-intense research across a wide range of disciplines, including demography, history, geography, landscape ecology, and urban planning. We will demonstrate the potential of these data in a case study on comparative analysis of changes in the built-up areas, population density, and road networks.

2. Broader Impacts

The proposed project, called **Imola**, will benefit a broad interdisciplinary community that requires detailed, long-term geographic data in their studies. With the increasing investment in creating large digital archives of scanned maps (e.g., the newly released collection of Sanborn maps in the Library of Congress), **the size of the research community using historical map data also grows rapidly.** The PIs have close collaborations with most of these communities (e.g., from PI Leyk's previous NSF HNDS-I project on creating 200 years of historical settlement data for the US (HISDAC-US) [51], see Section 5) and will continue to reach out to additional disciplines to ensure that Imola addresses their needs. Through Imola, we will deliver four tangible and novel products.

1) US1884+: We will integrate our existing machine learning models [19, 21] and system [42] for digital map processing in a new system called DaVinci. We will use DaVinci to produce a new large database, called US1884+, of **GIS-accessible, multi-temporal spatial road network layers, indexed by map scale and edition year, from the historical USGS topographic map archive.** The data collection will include vector and raster layers representing road network features for years ranging from *1884 to 2006*. Each vector layer will link to a raster layer in which each raster cell represents the uncertainty of belonging to a certain type of road feature of interest. **The uncertainty measures will be derived from map metadata (e.g., rectification and co-registration errors) and the extraction process.** The raster representation is not limited to a pre-defined boundary and can be easily integrated with other data, such as the NSF-supported National Historical GIS (NHGIS) historical census data or USGS land-use and land-cover data. We will focus on processing the complete map series of scale 1:62,500, which covers long time periods for most places in the US. We will also process sampled map series of scales 1:24,000 and 1:250,000 to test for scalability and performance. We will model US1884+ with metadata with respect to the GeoBlacklight schema [24] and Open Geospatial Consortium (OGC) vocabularies and work with the community to adopt other metadata schemas. We will disseminate the US1884+ database through 1) the Imola project website, 2) Harvard Dataverse [27], and 3) the UMN U-Spatial (part of UMN's Research Computing, a center in the UMN's Research and Innovation Office (RIO). We will also make US1884+ readily usable with other historical data repositories, such as the NHGIS. NHGIS hosts all other IPUMS data projects, including IPUMS USA, which has enormous popularity and has been continuously developing for over 25 years. Providing the data in a readily usable way to the IPUMS NHGIS community will be a highly sustainable solution using one of the most well-known portals for open spatial, historical data. We will also publish the metadata through the OpenGeoMetadata repository [74], from which the raster products can be discovered by spatial repositories, such as Stanford EarthWorks [82]. As the only systematic geographic record spanning the nation for over 100 years, US1884+ will also serve as the contextual data for the CyberGIS communities [15] to train additional machine learning models for processing other US historical map collections.

(2) DaVinci: To further refine US1884+, we will develop new capabilities in DaVinci **for robust road network extraction from historical maps with significantly reduced needs for human-annotated training data.** DaVinci will also have impacts beyond map processing since the new capabilities will enable effective organization and incorporation of prior geographic knowledge for robust geospatial image understanding with uncertainty estimates. DaVinci will benefit the communities of map processing, remote sensing, and geospatial sciences. The source code will be released under the MIT License [85] on GitHub [25]. We will pack DaVinci in Docker containers, publicly available on DockerHub [16], and register their references with Zenodo [108].

3) Case Study: US1884+ will extend the temporal dimension of current landscape research to the beginning of the 20th century for most parts of the US and also serve as reference spatial data to help researchers study their historical spatial data collections (e.g., racial covenants in historical property records, see National Covenants Research Coalition [73]). Making a living national heritage accessible for analytical inquiries will benefit research in demography, landscape ecology, biogeography, natural hazards, history, and economics, where knowledge of fine-scale changes over time in, e.g., transportation networks, is essential to derive substantive interpretations. To foster partnerships and community development, **we will run a case study and work with community users to demonstrate the use of US1884+ for scientific studies, including comparative analyses of changes in built-up areas, transportation features, and population density.** This example case study, made possible through Imola, will showcase how inquiries in broad scientific fields can be substantively advanced to benefit society and how new knowledge can contribute to assessing human-environment interactions and their societal impact. Additional examples illustrating the future broader impact beyond the initial targets are:

- *National scale, fine-grained spatial population projections and demographic estimates* are essential to study neighborhood demographic change and human-environment relationships. The extracted map data (e.g., road networks) in combination with historical settlement data, such as HISDAC-US [51], can be used to spatially refine historical census data (in NHGIS) through advanced dasymetric modeling [67, 105] to inform population projections on a national scale.
- *Retroactive natural hazards (e.g., flood) risk assessment* can be advanced using historical human-made features such as roads and buildings in spatial proximity to hydrographic features linked to natural hazard databases to better understand changes in flooding-urbanization associations.
- *Long-term environmental exposure assessment* can be carried out at fine spatial scales to assess human exposure to proximate pollution sources (e.g., from heavy industry, traffic such as highways) over long periods. Linking such assessments to health (e.g., cancer) databases will enable public health researchers to conduct unprecedented inquiries to better understand relationships between health outcomes and long-term exposure histories.

In our case study, we will demonstrate the enormous potential and usability of Imola's data collections and tools for a new class of data-intensive, uncertainty-aware studies of landscape and population change. We will support the creation of digital projects highlighting the capabilities and features of US1884+ and encourage adoption by faculty, students, and members of the US library and archives communities, including data and digital librarians. We will collaborate on projects that combine extracted map data from USGS, Sanborn, and other historical maps with the census, environmental, and other data held by digital libraries and repositories.

(4) Project and community-contributed GIS tutorials with US1884+: We will leverage our case study to develop class projects, independent research opportunities, and hands-on workshops that showcase US1884+ in various disciplines and educational contexts. This includes developing, curating, and making available GIS tutorials that interface (e.g., filter and download) with US1884+ and can be used in a classroom setting. The documentation, training, and educational material will inform researchers and students about using US1884+ and will be available on Imola and open repositories, with each item indexed by Digital Object Identifiers (DOIs).

Outreach activities The project results will be disseminated through presentations and publications for community outreach at various online and face-to-face events. We will proactively reach out to community users at the Association of American Geographers (AAG) annual meetings. We will host research and educational sessions at AAG's annual meetings to engage community users and showcase scientific studies enabled by US1884+. We will also disseminate Imola through other community events, such as

the University Consortium for Geographic Information Science (UCGIS) meetings, to solicit feedback and foster community adoption. Finally, we will organize online hackathons and grand challenges every year to promote Imola and provide tutorials and quarterly online “office hours”.

Educational: The proposed work will promote interdisciplinary teaching and student training through direct project participation, courses taught by the PIs (GIScience at CU and UMN, Computer and Data Sciences at UMN), US1884+ users and their applied work, and interdisciplinary case studies. We will involve a postdoc and graduate and undergraduate students in this project (in spatial, computer, and data sciences and geography), and the research will provide thesis topics. Researchers and students will be highly exposed to teamwork and directly involved in interdisciplinary research design, analysis, and publishing processes. We will also use feedback from community users to create and revise lecture materials derived from Imola’s data collections and case studies. The annual sessions at AAG will include one dedicated to discussing educational materials and practices related to Imola. We will also actively seek out and involve students underrepresented in STEM disciplines, including Native American, Hispanic/Latino, and female students from traditionally less-funded schools (e.g., tribal colleges). The PIs have a successful track record of working with underrepresented groups and will continue to recruit students from these groups for the project.

3. Intellectual Merit

To enable new research opportunities in the social, behavioral, and economic sciences related to large geographic areas and long time periods, Imola presents the following novelties.

US1884+: US1884+ is a novel large-scale spatiotemporal database of attributed long-term, national-scale detailed road networks from 1884 to 2006, significantly extending the temporal dimension of landscape research across a wide range of disciplines, including demography, history, geography, landscape ecology, and urban planning to the beginning of the 20th century for most parts of the US.

Context-aware map processing framework: Imola’s DaVinci will build on our existing map processing work [19, 21, 42] to process large numbers of USGS map sheets efficiently for producing US1884+. Importantly, this effort will also develop a novel machine learning framework that effectively exploits spatial semantic relationships to enable fully automated, robust recognition in scanned maps with varying image conditions. Existing methods are limited because they require large amounts of training data, are sensitive to underlying image conditions, and do not effectively use *existing geographic data about things and their patterns in space* for robust and accurate recognition (e.g., [70, 72]).

Characterization of uncertainty: A critical component of this research is characterizing uncertainty in feature extraction and original data production. We will systematically characterize three types of uncertainty sources for US1884+: processing-related uncertainty (from the recognition techniques), production-related (inherent), and application-related uncertainty. This is a crucial novelty, enabling US1884+ (and other kinds of data extracted from maps) in meaningful studies since feature extraction results from machine learning models typically **only contain processing-related uncertainty** and cannot be used directly unless all types of uncertainty measures are fully characterized.

Case study and training materials: Imola’s database, case study, and training materials, including GIS tutorials interfacing with US1884+, enable domain scientists to extract and organize knowledge in evolving visual documents, such as a map series, opening new forms of data-intense, science-driven landscape research in a wide range of disciplines, and has the potential to open up new interdisciplinary research avenues. Our past and current collaborators include well-known senior researchers in historical cartography, ontology and geo-data representation, spatiotemporal data curation, demography and urban analysis, historical demography and historical geography, and CyberGIS. We will work with them to facilitate the use of Imola in their research and communities and initiate new collaborations.

4. Governance of Research Results

The Imola project website will be the main platform for the dissemination of the project products, which will include: US1884+, the extraction tools, DaVinci, sample data, user-published data, custom analytic workflows (e.g., GIS tutorials), training and educational materials, and information about the project goals and plans. Imola will be maintained for 5 years after the end of the project. All the software will be available via GitHub [25] with the MIT License. The GIS tutorials and the DaVinci containers used for

execution will be available via Zenodo [77, 108]. All Imola-related data, including the extracted spatiotemporal data collection, other sample data, and all secondary data derivatives from the case study, will be publicly available on Imola's project website, Harvard Dataverse [27], and UMN's U-Spatial, and their metadata deposited in OpenGeoMetadata [74], which represents a sustainable solution to provide wide access to the data after the expiry of the grant. During the project's lifetime, users will be encouraged to publish their data related to Imola in open repositories. The team will also collaborate with the USGS to perform an evaluation of the products for compatibility with National Map data.

5. PIs' Previous Contributions

Collaborative: The interdisciplinary research team includes a computer scientist (Chiang) and a geographer (Leyk) who have collaborated for more than 15 years on a wide range of recognition tasks in various historical map types. **PI Chiang** has been leading research efforts to automate the extraction and integration of geographic information from historical maps for nearly 20 years, with more than 60 peer-reviewed publications on related topics. His work has developed machine learning methods and systems to streamline the processing of historical maps, including georeferencing historical maps, extracting and linking geographic features, spotting text content, and analyzing large map series. In 2022, PI Chiang led a team that won first place in the Map Feature Extraction Competition in the Defense Advanced Research Projects Agency (DARPA) and USGS AI for Critical Mineral Assessment Competition [1]. **PI Leyk** has led research projects funded by the NSF on methodological work for data extraction and recognition from historical maps, data harmonization, uncertainty assessment, and spatial demography. He also led research on the production of spatio-temporal large-scale data infrastructure, funded by NSF HNDS-I, and is the designer of **the Historical Settlement Data Compilation for the US (HISDAC-US)** [51], which portrays the evolution of the built environment in the US over more than 200 years at fine spatial and temporal resolution. His research is highly interdisciplinary and collaborative around topics of natural hazards, urban change, and population projections, and has also received funding from other agencies, including the NIH and the European Commission. The team brings together unique and complementary expertise and experiences that qualify us to conduct the proposed research successfully.

Overview of Previous Contributions: Supported in part by a recent NSF project, *LinkedMap*, PIs Chiang and Leyk's groups have developed the ContextMP framework [58], which is a set of map processing algorithms tested on extracting road networks, railways, waterlines, building symbols, and wetland areas from a variety of historical maps from the USGS, Ordnance Survey, and other mapping agencies [10, 12, 18–21, 42, 59, 80, 97–100]. Below is an overview of the team's recent technologies in methods and systems for processing historical maps to generate valuable data, information, and knowledge. Further details can be found in the PIs' review paper that coined the term Digital Map Processing published in ACM Computing Survey [13], a vision paper that won the best visionary paper award at ACM SIGSPATIAL [8], a short book [10], and a recent book chapter [9].

One of the major challenges in extracting geographic features from scanned historical maps is the lack of training data. If the map is already georeferenced, overlaying an existing data source on the map can help provide annotations as training data. For example, Meta's Map with AI uses road networks from OpenStreetMap to collect over 100 million road annotations from satellite imagery. However, direct overlaying of external vector data on the map for annotation often led to misalignment and errors due to various data and quality inconsistencies between the two datasets. To address this challenge, we developed the Label Correction Algorithm (LCA) [21]. LCA optimizes the geometry of external vector data and the map's image features to minimize false annotations. Tested on the extraction of railroads and waterlines from USGS topographic maps, LCA significantly improved the quality of annotations and the results of geographic feature extraction compared to other methodologies.

In addition, **accurately detecting linear objects requires the detector to capture and combine image and spatial context.** The image context should encompass the cartographic symbols of the desired linear objects. For example, the black crosses, highlighted by red circles in the right panel of Figure 3, are representative cartographic symbols for railroads. In Figure 3, both roads and railroads are black lines. Capturing the black crosses as an image context is crucial to differentiate segments of railroads from roads and reduce false detection. While most of the existing approaches can effectively capture the image context (e.g., SIINet [83], CoANet [66], and Relationformer [81]), they typically ignore

the spatial context, which refers to the spatial relations among the detected image context in an image. For example, the detector must detect black(ish) pixels following an elongated area with repeated occurrences of the cross symbols to extract the railroads accurately. In preliminary work, we developed the Linear Object Detection Transformer (LDTR) to generate accurate vector graphs for linear objects from scanned map images [19]. **LDTR uses a multi-scale deformable attention mechanism to effectively capture complex spatial context, including the line's orientations, curvature, and topological relationships among multiple lines.** LDTR's attention mechanism learns representative image context to reduce false detection and explicitly encourages interactions among turning points in the detected line to extract accurate topology. **Our experimental results demonstrated that LDTR significantly improved connectivity by approximately 20% compared to state-of-the-art linear object detectors (Figure 4).** LDTR's superior results were the key to allowing our team to win the DARPA and USGS AI for Critical Mineral Assessment Competition in 2022.

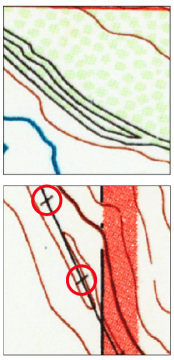

Figure 3. *Examples of a road (left) and railroad (right) in a USGS historical topographic map*

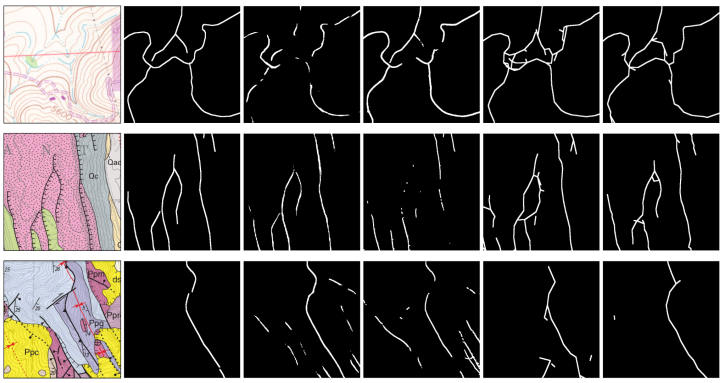

Figure 4. *LDTR's detection results.* ***Left to Right:*** *map images, ground truth, and the detection results from three baselines (SIINet [83], CoANet [66], Relationformer [81]) and LDTR.* ***Top to Bottom:*** *detection of waterlines, scarp lines, and fault lines.*

Extracting useful information from large numbers of scanned map images with large image dimensions also poses a significant challenge in executing machine learning models since most models are not designed to handle images over 10,000 by 10,000 pixels. In another line of our work, we focus on detecting and recognizing text from large numbers of scanned historical maps. Historical maps store a wealth of information, not only in geographic features but also in text labels and descriptions. Detecting, recognizing, and linking text in scanned historical maps required a unique approach since map text can have varying orientations, curvatures, spacings, font types, and font sizes. We developed SynthMap, a process that generates large amounts of synthetic historical-styled map images with text annotations, effectively addressing the lack of training data [61]. Text spotter models trained with SynthMap demonstrated significantly improved text detection and recognition performance on historical maps. **We built and released an open-source map processing system called mapKurator incorporating a pretrained text spotter using SynthMap and generated 90 million text labels from over 60,000 scanned historical maps to support searching these maps by their text content at the Stanford David Rumsey Map Center (Figure 5)** [42]**.** mapKurator slices a map image into smaller patches, predicts a bounding polygon for each map text instance, transcribes the text, merges the results, and converts the image coordinates of bounding polygons to geocoordinates, among other functions (Figure 6). Also, **mapKurator can adopt additional machine learning models (e.g., line detection) to streamline the feature extraction process from large numbers of scanned map sheets.**

Unlike the supervised geographic feature extraction methods discussed in the previous paragraphs (e.g., LDTR), we also developed fully unsupervised methods using modern geospatial contextual data to support training and sampling. For example, we have developed the Historical Road Network Extractor (Hironex) [100], publicly available as a Python tool. Hironex requires vector data of the modern road network (e.g., from OpenStreetMap available via OSMNx [2]), as well as a georeferenced historical map from the year *T*. Under the assumption that road symbols in the historical map

approximately spatially coincide with the modern road (if it existed in *T*), the tool samples the color values of the scanned map image near the modern road vector data and calculates the directional grey value variance perpendicular to the modern road axis. A linear object in the map, parallel to the modern road axis, will result in a high directional variance when such a perpendicular sample is collected. Thus, high variance at various sampling locations along the modern road axis can be interpreted as a proxy for existing road symbols in the underlying historical map (or as a higher likelihood that the road existed in *T*). The identified instances of high variance are registered for each modern road vector in the study area (Figure 7a,b). This method is fast and can be applied to large regions, as shown for the Bay area, which encompasses over 200,000 road segments (Figure 7c,d). Finally, a 1D-clustering method is applied to this continuous metric to identify the clusters of coinciding "historical" and "more recent" roads, which can be done for each point in time at which historical maps are available (Figure 7e).

Figure 5. *Search by Text-on-Maps on the David Rumsey Map Collection website. Searching "Peiping" using the 90 million text labels generated by our system with 50k maps in the David Rumsey Map Collection. Note that "Peiping" is the old name for Beijing, China.*

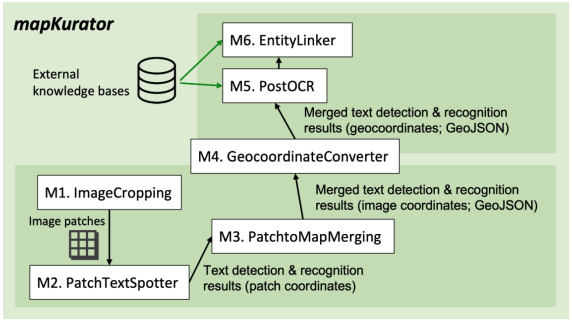

Figure 6. *mapKurator's capabilities include automatic processes of 1) detecting and recognizing text from maps and 2) linking map text to their corresponding entities in external knowledge bases and historical gazetteers. The automatically generated results are in the GeoJSON format.*

In summary, our past work has automated the information extraction process from historical maps, streamlined the conversion of these data into a structured, standard, readily usable format, and provided us valuable experience collaborating with domain scientists.

6. Information Technology

6.1 Imola Overview

We will build on our past efforts to develop Imola, including DaVinci, US1884+, and the case study. First, we will incorporate mapKurator [42] with LCA [21] and LDTR [19] in a new system called **DaVinci** to streamline the production of US1884+ **(Thrust I)**. This allows us to **start producing the first version of US1884+ at the very beginning of the proposed project** from the US Geological Survey (USGS) historical topographic map series composed of 178,000 sheets between 1884 and 2006. We have also already collected high-resolution, georeferenced copies of these map sheets from USGS. Each map sheet also contains metadata, including the map edition, map name, primary state, state list, county list, georeferencing information, map scale, published year, and original sheet dimensions (in inches). Second, Hironex (as well as LCA and LDTR) demonstrates how contextual information could be used to support the creation of **US1884+**. However, Hironex's performance highly depends on the original map condition and the scanning process since it operates solely in the image space. Supervised methods, like LCA and LDTR, can handle a range of map and image conditions if enough training data are available. We will also build on Hironex, LCA, and LDTR to develop new capabilities in **DaVinci**, which will integrate and exploit existing contextual data for robust and accurate map processing to refine US1884+ **(Thrust I).**

US1884+ will be a collection of georeferenced raster and vector layers. Each cell of the extracted raster layers will represent the uncertainty of whether it belongs to the feature of interest. The uncertainty measures are derived from map metadata (rectification and co-registration errors) and the extraction process **(Thrust II)**. Creating raster layers to represent geographic features with inherent uncertainty has

the benefit that the data are objective and can be easily integrated with other historical datasets such as historical census data (e.g., census summary tables or reporting area boundaries from NHGIS) or land-use and land-cover data (e.g., the USGS Geographic Information Retrieval and Analysis System, GIRAS). Also, the advantage of using raster data is that a road can be represented by an unbiased set of cells (preserving the approximate original shape), each of which can have its uncertainty estimate. If desired (e.g., to accommodate varying error tolerance levels), the user can apply specific uncertainty thresholds to each cell to produce a vector representation of the map feature. We will demonstrate the potential of these data in **an extensive, diverse case study** on comparative analysis of changes in the built-up areas, transportation, and population density **(Thrust III)**. In addition, we will generate a version of US1884+ in vector format to save time and space for disseminating the results. Below, we describe Imola's three novel thrusts.

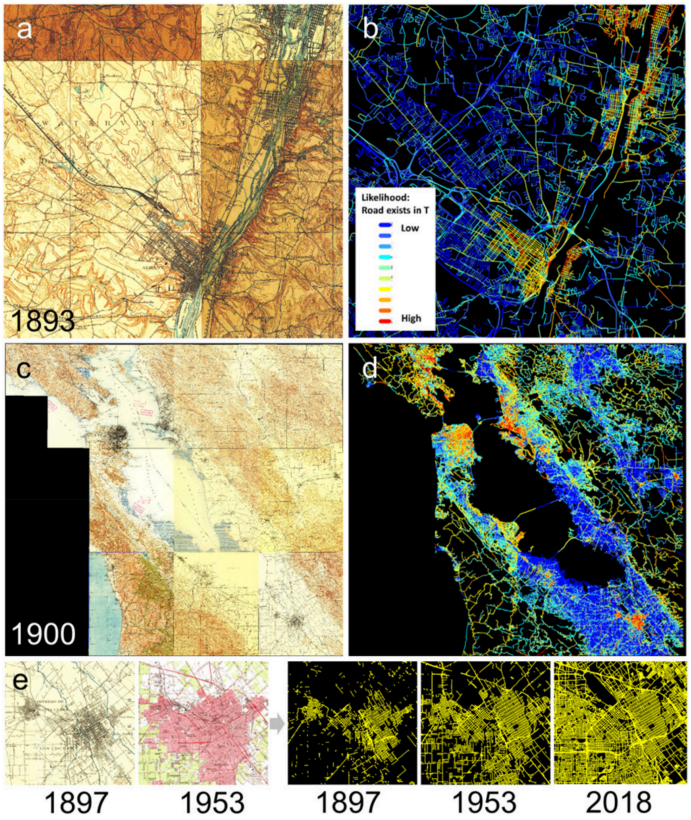

Figure 7. Illustrating the unsupervised extraction of the historical road networks: (a) map composite from 1893 covering the Albany (NY) region, (b) modern road network data for the same region attributed with a high likelihood of existence in 1893, (c) and (d) show a historical 10-map composite and the modern road network with likelihood attributes, respectively, for the Bay area (CA), (e) illustrating the multi-temporal application of the method, obtaining approximations of the historical road network for 1897 and 1953 for comparison to the 2018 road network data, shown for Santa Clara and San Jose (CA).

6.1 Thrust I: DaVinci - A Digital Map Processing System and New Capabilities

Background: Recent deep neural networks (DNNs) show promising results in many computer vision tasks using large amounts and varieties of annotated visual training samples, primarily available as natural images (e.g., scenic photos). Most state-of-the-art map processing approaches build on existing convolutional neural networks (CNNs) backbones and exploit diverse types of DNNs to tackle map-specific challenges (e.g., see [10, 13]), including handling complex map content, small target objects (e.g., less than 100 square pixels), and large search space. *However, these approaches often are not robust since they typically need to train the neural networks using out-of-domain, natural images and then fine-tune the networks using a small set of annotated visual samples* (e.g., see [10, 13, 23, 40]). Manual annotations would be required for processing every map because even different regions in a scanned map or scanned maps from the same series would suffer from varying issues of bleaching, blurring, and false coloring [41, 43], and the trained machine learning models are likely not generalizable due to the small size of manually created training data. Also, unsupervised methods, such as Hironex, only work in the color space and would not be robust to varying image conditions.

To tackle the training data scarcity challenge, new computer vision approaches based on visual transfer learning can *effectively employ existing knowledge bases for robust image understanding using*

reduced numbers and varieties of annotated visual samples (e.g., see [26, 60, 70–72]). **While successful, these new approaches do not fully exploit existing geographic data about things and their relations in space.** To overcome this challenge, the proposed work will leverage our ContextMP framework [58], including LCA, LDTR, and mapKurator, tested on extracting road networks, railways, waterlines, building symbols, and wetland areas from a variety of historical maps [10, 12, 18–21, 42, 59, 80, 97–100] to develop new capabilities in DaVinci. Methods in ContextMP are rooted in two conceptual insights. **First**, maps change gradually over time, and there is a dependence between map editions. **Second**, the features depicted on maps often have implicit "rules" regarding their spatial arrangement following the cartographic principles and publisher standards. ContextMP algorithms generally use contextual data to spatially constrain the area of interest and automatically collect graphics examples for training a computer vision model or directly extracting the desired geographic feature (e.g., LCA [21]). In the case of road extraction, ContextMP uses contextual road data to intelligently search for likely graphic examples of roads in a scanned map (e.g., linear objects having a homogenous color and similar shape to the nearby contextual data) to train a semantic segmentation model [21]. Also, using contextual housing data for building symbol extraction, our approach overcomes spatial shifts between map features and contextual data [97, 99] and shows accuracies between 80% and 95% (Figure 8). In addition, we demonstrated a ContextMP approach that automatically generates approximate geolocations for 500 scanned map images from the New York Public Library using LinkedGeoData as the contextual data [59]. A part of ContextMP is currently used in various research projects and open source and commercial products, including at the Alan Turing Institute (Living with Machines), USC Digital Library (Strabo Web) [31], Google AI (Kartta Labs) [84], Facebook (Spatial Computing Group), and CLS Risk Solutions [107].

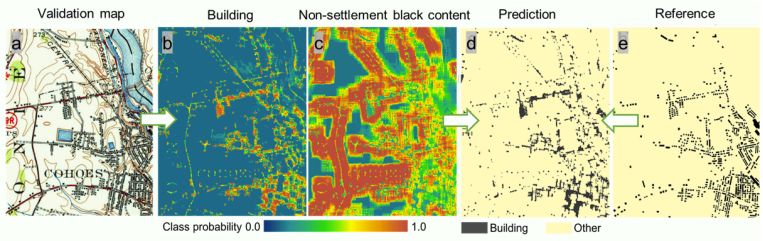

Figure 8. *Semantic segmentation on USGS topographic maps.* ***Left to right:*** *(a) Cohoes (NY), 1949, (b) building probability surface, (c) non-settlement probability surface, (d) semantic segmentation result, and (e) validation data used for accuracy assessment.*

Proposed Work: Task A. We will first integrate our current machine learning models, LCA [21] for training data generation and LDTR [19] for supervised linear object detection with mapKurator [42] to develop the first version of DaVinci to streamline the process of large dimensions and numbers of map images and start producing the first version of US1884+. LCA will use external vector data (e.g., the existing National Map vector layer) to generate training data automatically to train LDTR to detect the complete set of USGS road types. The mapKurator system is already well-documented and a proven map processing system used by other organizations (e.g., the David Rumsey Map Collection) for text spotting. Similar to ContextMP, we will start the production from the latest USGS historical topographic maps since their features are most similar to the external vector data and work our way back in time. US1884+ will include raster and vector layers of each road feature type from different map editions with map-level temporal (i.e., date of publication) and spatial (i.e., map sheet coordinates) meta-information. The raster layer will include uncertainty measures derived from the recognition process in DaVinci and a systematic locational offset estimate using cartographic principles and benchmark map features (Thrust II).

Proposed Work: Task B. The first version of US1884+ might suffer from missing road network features that do not appear often or can go through significant changes in close years (and hence with limited training data from LCA). These potential missing features are often the most interesting ones for SBE research. To refine US1884+, we propose to build new capabilities in DaVinci, which will be based on visual transfer learning to minimize the required training data for map processing, e.g., instance segmentation, using prior knowledge from contextual data. **The intuition is similar to how human researchers work on recognizing map features.** For example, if a researcher is trained to identify the visual appearance of paved roads on a map but not dirt roads (both are sub-types of light-duty roads), the researcher can tell that dirt roads have similar visual cues as paved roads but are not quite the same. The researcher can then determine that the objects can be *light-duty roads* but not paved roads. Further, by

inspecting the surrounding environment, one can find multiple common topological features in rural areas, and the objects in question are not very long and can curve significantly (compared to other types of light-duty roads), which fits the semantic description of dirt roads and hence determines the subtype.

Building the new capabilities in DaVinci includes solving two new challenges ***1)*** *building knowledge graphs from contextual data for capturing lexical, spatial, semantic, and domain prior knowledge and* ***2****) fusing multimodal prior knowledge and annotated visual samples for robust map processing using reduced numbers and varieties of annotated visual examples.* We will tackle the first challenge by building a **common-sense geographic knowledge graph** (CS-GKG) [35] from the available contextual data (e.g., the USGS National Map, OpenStreetMap, HISDAC-US). In CS-GKG, a *node* represents an entity type in the contextual data using *lexical embeddings* (e.g., generated from FastText [3, 36, 37]) of the entity type name (e.g., dirt roads). **The intuition is that if two lexical embeddings are close (e.g., gravel and dirt roads), their visual cues will also be similar (gravel and dirt road symbols are dashed lines).** Next, DaVinci will automatically generate *links* between node pairs to represent spatial semantic rules in the contextual data. These semantic rules describe the spatial relationships between geographic features. These rules will dictate changes in geometry or topological relations that are possible vs. those that are unlikely. Using the instance of roads, examples of semantic rules are (1) location-based rules (e.g., 'the majority of roads does not disappear and reappear'), (2) geometry-based rules (e.g., 'roads are unlikely to become shorter'), (3) topology-based rules (e.g., 'road features should intersect at the same or nearby location if both road features exist in both points in time' (Figure 9)), (4) spatial-integrity rules (e.g., 'bridges are usually connected to roads'), and (5) domain rules (i.e., 'the road supertype contains subtypes of light-duty roads and highways'). **Each rule will also be location, scale, and time-specific. When processing a map with known geocoordinates, scale, and time, DaVinci will always use the part of CS-GKG that is *closest in time and map scale* to the processed map to accommodate changes in image quality, feature generalization, and symbology across map editions and scales.** We will investigate methods to detect and quantify these rules from the contextual data automatically. For example, to capture the spatial-integrity rules, DaVinci could use recurrent neural networks (RNNs) or the recent advances in Transformer networks (e.g., [7, 17, 104, 106]) to learn a fixed-length embedding from varying lengths of local co-occurring other entity types.

Figure 9. *Historical 15-minute USGS topographic maps (left: circa 1950, middle: circa 1964) & USGS National Map data (right: circa 2012), Marina del Rey, CA*

Once the prior knowledge is organized, the new challenge is how to use it to facilitate robust and accurate feature extraction with reduced training data. We will investigate using graph-based neural networks (e.g., graph convolutional neural networks, GCNs) over CS-GKG to generate a *fused semantic embedding* representing the combined spatial, lexical, and semantic contexts for each node (entity type). We will also study effective graph representations to accommodate these semantic rules (e.g., multiple graphs in which each link represents one rule vs. one graph). In parallel, we will look into visual feature extraction methods for each CS-GKG node that has annotated visual samples (e.g., from LCA) using Transformer-based approaches, like our LDTR, or CNNs (e.g., Mask R-CNN [29] for instance segmentation). Finally, we will investigate and compare contrastive learning strategies for training semantic embedding fusion and visual feature extraction end-to-end so that the visual embeddings of an annotated visual example (e.g., a dirt road) are close to the example's fused semantic embeddings. For example, we will train multilayer perceptrons (MLPs) to map visual embeddings to their corresponding semantic embeddings. This way, the semantic embeddings would serve as the supervisor to encourage extracting novel image features to construct the visual embeddings (e.g., [70]). For example, since both "paved roads" and "dirt roads" belong to the "light-duty road" domain, their semantic embeddings will promote the extraction of shared image features between the two subtypes. The shared image features can then be used to identify light-duty roads without annotated visual samples. For inference, DaVinci will

identify candidate instances, extract and map their visual features, and use the closest semantic embedding as the detected object type. This way, the fused semantic embeddings representing prior knowledge will work together with visual cues for robust object detection and instance segmentation, similar to how a human researcher operates. We will build a complete production system in a reusable environment (e.g., containers) and register the images in Zenodo [77] (through its GitHub integration) for public discoverability and access, similar to what Chameleon Cloud does [6, 39].

Evaluation: We will use the standard correctness, completeness [28], and average path length similarity (APLS) [103] to evaluate the precision, coverage, and connectivity of detected lines, respectively. The correctness, completeness, and APLS calculation are between the detected road networks and ground truth. We will manually annotate the locations of desired road features from selected, diverse USGS topographic maps as ground truth (see Thrust II, process-related uncertainty).

6.2 Thrust II: Uncertainty Characterization

Background: Data extraction from many digital maps of varying graphical quality is an error-prone process resulting in complex forms of uncertainty. The team has experience analyzing uncertainty in historical map data, including a conceptual framework for uncertainty investigation in historical survey maps [47]. This framework systematically identifies sources of uncertainty to be addressed when working with spatial data extracted from such maps. The framework differentiates production-related (inherent in the map), processing-related (through data processing), and application-related (data use) uncertainty and outlines forms of uncertainty assessments at different analytical stages. For example, Leyk et al. [45, 46, 48] focus on quantifying the error in map processing tasks based on incorrectly detected objects, not single pixels, using stratified statistical sampling. The approach calculates various accuracy measures from the error matrix, such as *accuracy [68]*, *Kappa coefficient* of agreement [14], and *normalized mutual information* (NMI) criterion [22]. Accounting for chance agreement (Kappa) and biased class distributions (NMI, F-score) is important because map features can be large areas (vegetation), lines (roads), or small symbols (buildings). Map contents are susceptible to spatial inaccuracy due to the survey techniques or inadequate definitions. Our prior research to detect and quantify these rules from the contextual data automatically predicts such inherent uncertainty [55, 56].

Proposed Work: Mapmaking traditionally followed a set of cartographic rules regarding the spatial feature arrangement and symbol sizes, resulting in an ordered hierarchy of feature placement to control necessary displacements of some (lower order) map features. For example, in the USGS topographic maps, railroad lines are always mapped as the primary geographic feature with the highest spatial accuracy (i.e., a benchmark feature). Other features (in the order of roads, streams, buildings, etc.) may be displaced from their true geographic position, if necessary, to create a cartographically sound map. We will make use of this hierarchy in the assessment of locational offsets across map editions of the same map scale, stored in the same data collection for building US1884+. Starting with offsets between benchmark features (railroads) in different map editions will provide robust measures of offsets related to a lack of co-registration accuracy between editions because these features rarely change positions over time and are measured and mapped with high accuracy. Thus, these offsets are expected to be systematic. Once these errors can be quantified, the offsets between representations of other feature categories across editions can be measured and evaluated in relation to the benchmark offsets following the ordered hierarchy of feature placement. These offsets will be evaluated and used to inform the production of US1884+, e.g., define expected deviations between contextual layers and features in the map for extraction. These offsets will also be included in the metadata as they will be important for meaningful comparison in a GIS. They will also be critical for identifying semantically identical representations in different editions to describe changes in geographic features over time [32, 80]. These offset measures will be computed for a set of data collections for the target scale to collect a sample large enough to make reasonable assumptions for this map scale as well as for different map editions (i.e., it can be expected that such offsets are larger in older maps with lower co-registration accuracy and less consistent feature representations). While we will focus on evaluating the recognition techniques (processing-related uncertainty), we will also explore the assessment of production-related (inherent) and application-related uncertainties based on [47] during the case study.

Processing-related uncertainty arises through recognition (or other processing) errors (e.g., down-sampling processes in DNNs). Geographic validation data will be manually digitized from case

study maps, and accuracy measures will be calculated, such as completeness, correctness, quality, redundancy, and ALPS. We will evaluate the recognition for different map editions (i.e., levels of graphical quality) to assess both the robustness of the recognition results and the amount of required manual post-processing. We will compare the results to state-of-the-art map processing technologies (e.g., [11]) to determine baseline performance for our evaluation. We will also evaluate various DNN architectures to understand their impacts (e.g., from the convolutional processes) on processing-related uncertainty.

Production-related uncertainty, which is the uncertainty that is inherent in maps due to the survey measurements and the level of generalization or epistemological changes applied in the map-making process, can only be estimated if maps or other spatial datasets with higher accuracy are available [56]. We have identified several such US local maps and spatial datasets within our study areas for different points in time to perform the uncertainty assessment for several map editions. Since the recognition will target discrete representations, uncertainty can be assessed as completeness at the pixel level or at the object level (e.g., road features as well as road pixels that are missing in the USGS map but exist in the validation map) and deviation measures (e.g., distance or offset between the feature in the validation map and the feature in the USGS map). This validation will be done in relation to map scales to account for different degrees of the underlying feature generalization. **It is imperative to understand that our efforts here aim to extract the geographic features as they are represented in the original maps, not the meaning of epistemological changes in map representations and possible consequences for landscape analysis.** However, we will provide the basic data structure for researchers interested in the epistemology of map representations.

Application-related uncertainty may arise depending on the application for which the data is used. The general target application that will be facilitated through Imola is comparative analyses of spatial representations within larger areas and over long time periods. The two main sources of uncertainty are the semantic inconsistency of map contents (i.e., objects compared are underlying different definitions, which can be related to epistemological changes) as well as spatial deviations (i.e., spatial offsets) between features (of assumed identical semantics) from adjacent map pages and identical features from different map editions. We will demonstrate the assessment of uncertainty that occurs if the data are used for such purposes and focus on the spatial offset uncertainty that is directly tied to the computed measures of spatial offset between features across map editions. We will explore the nature of these spatial offsets and propagation effects in such applications. If they are highly systematic, the main cause may be a lack of map co-registration. However, if they are non-systematic, the cause may be due to changes in cartographic representations or enforced displacement, which will be explored by taking railroads as reference objects. In any case, knowledge of this type of uncertainty will be fundamental for applications of change analysis, where **uncertainty in observed changes must be differentiated from data uncertainty**. If sufficient historical validation maps can be found, we will compare changes as derived from the target maps (e.g., from USGS), respectively, and changes as derived from these local reference maps.

Evaluation: Similar to Thrust I, measures such as correctness, completeness [28], and average path length similarity (APLS) [103] will be applied to evaluate the precision, coverage, and connectivity of detected lines, respectively, and characterize the uncertainty. However, these measures will be calculated between the detected features and manually digitized road segments (process-related uncertainty), between roads in local reference maps and the roads in the USGS maps (production-related uncertainty), and between local reference maps and detected road features (application-related uncertainty).

6.3 Thrust III A Community-Driven Case Study on Uncertainty-Aware Land and Population Change Analysis

Background: This project needs to be understood not only as an innovative way to extract geographic information from large digital image archives. This project creates enormous potential and a large-scale data infrastructure from public-domain data for a new class of data-intensive studies of the effects of landscape change on human settlement and migration and vice versa. (Since the data will be derived from publicly available maps, there are no privacy or confidentiality issues about the data.) Just as conflicts and economic crises often drive large-scale migration and population changes, environmental shocks, like the Dust Bowl in the central US, have a profound effect on the population and landscape. Changes to the built environment, like infrastructure development (canals, rails, roads), profoundly affect

settlement patterns and ecosystems, but such data are chronically lacking for earlier points in time. US1884+ will document these changes and relationships and thus allow researchers to interrogate a set of novel questions of major societal significance. Understanding the effects of phenomena such as environmental shocks becomes increasingly important as the planet may enter a period of anthropogenic climate change. To date, the ability to study past landscapes and their evolution for this purpose is severely data-limited.

The extracted road layers have a unique data structure, and in our case study, we will demonstrate the utility of US1884+ for historical demographic and landscape research. This study is highly crucial to this project as we will also produce training material to aid others using the multi-temporal data collection. We will conduct uncertainty-aware spatiotemporal analysis of landscape changes over long time periods (120+ years) for selected study regions. While this is an intellectually meritorious activity in and of itself, we also hope to explore data fusion strategies with historical demographic data, such as the historical US census data provided through NHGIS going back to 1850, as well as historical settlement data such as HISDAC-US, to demonstrate the potential applications if different historical data types are integrated. Such studies can be used to investigate emerging research questions such as: "Using historical roads, built-up area and census data over 120+ years, what are the local, spatially refined patterns of change in the population distribution in the U.S.?" or: "What new insights can be found through identifying different types of suburban development (traditional vs. sprawling) using historical road and building data, to better understand the process of urbanization?" This case study will integrate the uncertainty evaluation for extracting the data (see **Thrust II**) and is intended to demonstrate the potential and usability of the produced data collections to inform and advise future users on working with the data and their uncertainty in complex applications.

Proposed Work: We will showcase **US1884+** in a case study with GIS tutorials (e.g., using the open-source QGIS) for querying, analyzing, and visualizing the extracted data to support a broad range of studies on uncertainty-aware land change analysis. We will use standard raster and vector format to store information about resources for the case study and US1884+ and to enable efficient spatial queries. The case study will be well documented, and training and educational materials will be developed to help a broad group of users and students, including non-experts in GIS, make the most out of **US1884+**.

The tutorial will be an effective way to demonstrate the potential of **US1884+** and how to work with **US1884+**, including GIS tools, data, and their uncertainty, in complex applications. We will select study regions representing different histories in landscape evolution and transitions driven by processes such as urbanization and its effects on rural and wild landscapes. For example, one promising study region is the I-95 megapolitan urban corridor stretching from Washington, D.C., to Boston, as it was continuously settled and underwent a transition from an agrarian to an industrial and finally a post-industrial urban conglomeration. In analyzing features of meaningful impact over the entire nation and long time periods (e.g., highways), we will leverage other historical data, such as our HISDAC-US over 200 years [51], to carry out unique assessments of changes in social and environmental systems in the US. We will conduct the different steps described below:

(a) We will model changes in feature density, e.g., highways, by computing kernel density functions for the extracted spatial layers from different points in time, creating raster layers in which each raster cell is assigned the density of the feature of interest within a given kernel. We will use these density surfaces to compute trajectories within, e.g., county boundaries or over regions to characterize change rates over time, providing a unique insight into urban and infrastructure development and the level and trends of landscape fragmentation. Such insights have the potential to advance pressing questions in urban analysis to characterize different types of suburban development (traditional vs. sprawl), land cover change, and landscape ecology.

(b) We will explore historical data fusion strategies to integrate our extracted layers as well as the road statistics from (a) with historical demographic data, such as the enumerated multi-temporal population summaries (e.g., [62]) as well as the HISDAC-US and its built environment attributes, to demonstrate the potential for multi-temporal demographic small area estimation. Historical road and built-up layers and environmental data can be used to spatially refine population data through methods of dasymetric refinement [67, 75, 105]. We have long-standing expertise in this field (e.g., [44, 49, 50, 76, 78, 79]) and will compute spatially refined patterns of change in the historical population distribution using those

integrated data products. Such refined change estimates will provide important new insights into demographic inquiry at fine granularity and also provide a better understanding of interactions between settlements and environmental systems.

Evaluation: We will use the case study and the developed tutorial to evaluate **US1884+** regarding usability and broader impacts (e.g., research publications, datasets published, and student mentoring results). Usability will be evaluated by soliciting feedback from the user communities and feedback collected during our dissemination and outreach activities. The feedback will be collected via 1-on-1 and user group discussions, via surveys after activities, and the Imola website.

7. Research Plan & Collaboration Mechanisms

Year 1: Build and deploy Imola with a preliminary version of US1884+ and data interface capability

- Evaluate and model location-, geometry-, and topology-based spatial semantic rules of the existing National Map vector layer and other sources (to be used as contextual data) for selected types of road network features (CU, UMN)
- Incorporate LCA and LDTR with mapKurator into DaVinci v.1 to perform large scale extraction of multiple road features in parallel from all map editions to **produce US1884+ v.1** (UMN, CU)
- Produce sample validation data across map editions and scales for target feature layers (UMN, CU)
- Perform a full evaluation on selected maps across editions for US1884+ v.1, including a first assessment of locational uncertainty, on the feature extraction results (CU, UMN)
- Build and test the initial new capabilities in DaVinci v.1 to exploit semantic descriptions within individual feature types to improve extraction accuracy and conduct full evaluation on selected maps across editions (UMN, CU)
- Release a prototype for multi-temporal data collections, US1884+ v.1, containing selected types of road network features, uncertainty assessments, and integrated metadata across map editions (UMN, CU)
- Establish protocols for data transfer to open repositories (e.g., Harvard Dataverse) and UMN U-Spatial, quality assessment & post-processing (CU, UMN)
- Build, deploy, and distribute Imola v.1 that includes GIS tutorials to access the USGS historical map archive (a total of 178,000 maps), metadata, tic point information and documentation, DaVinci, and US1884+ v.1 (UMN, CU)
- Seek feedback from community users. Start the uncertainty metric collection. (UMN, CU)

Year 2: Enhance Imola's functionalities with an improved version of US1884+ and insights from the case study and basic data analytics (integrate initial feedback)

- Create DaVinci v.2 to cover all spatial semantic rules more types of road network features (CU, UMN)
- Produce sample validation data across map editions and scales for target feature layers (UMN, CU)
- Use DaVinci v.2 to perform large scale extraction of multiple road features in parallel from all map editions to produce US1884+ v.2 (UMN, CU)
- Perform a full evaluation on selected maps across editions for US1884+ v.2, including an assessment of locational uncertainty, on the feature extraction results (CU, UMN)
- Develop GIS tutorials with case studies on landscape change with an assessment of uncertainty in the data versus the uncertainty inherent in the change of features across editions (CU, UMN)
- Build, deploy, and distribute Imola v.2, which includes additional GIS tutorials for data aggregation and other analytic integrations (e.g., with CyberGIS libraries), and enables users to submit their own case studies and tutorials to the project website (UMN)
- Seek feedback from community users (UMN, CU)

Year 3: Augment Imola's capabilities for providing the complete version of US1884+ and improved data extraction capabilities and finalize the case study and tutorials

- New, improved versions of DaVinci and case study based on community feedback (UMN, CU)
- Finish the iterative process between algorithm refinement and US1884+ quality assurance once given benchmarks are reached for accuracy and completeness (UMN, CU)
- The third production cycle for the complete set of 1:62.5K scale maps to create the final US1884+ and selective QA (UMN, CU); test production runs for 1:24K and 1:250K scale maps (UMN)
- Finalize US1884+ revisions and metadata corrections, including performing a full evaluation on selected maps across editions on the feature extraction results, with uncertainty assessment (CU, UMN)

- Finalize the case study and release GIS tutorial developed during the case study incorporating user feedback in previous years (CU, UMN)
- Submit project data to UMN and CU Digital Repositories, assuring that all pertinent software and documentation have been contributed to their respective repositories
- Finalize an Imola webpage and metadata features for data dissemination via open repositories (UMN)

Responsibilities: Chiang will be responsible for the overall research effort and will ensure that the work at UMN and CU Boulder is tightly integrated. He will also be primarily responsible for developing the recognition system, the road network extraction techniques, data production, and evaluation and will lead the other researchers in these efforts. Leyk will be responsible for uncertainty assessment, data integration, and the efforts related to the case study. He will oversee the efforts at CU Boulder and ensure information exchange and collaborative measures are in place and effective.

Meetings: Weekly video conferences: we plan to have weekly Zoom calls to discuss the various parts of the research effort and their coordination. We will have face-to-face annual meetings of the UMN and CU Boulder research groups and run them as one-day research retreats where we will discuss the current research and brainstorm on the next steps in the project. We will use annual conferences (e.g., AAG or ACM SIGSPATIAL) for these meetings.

8. Prior Results of NSF-Supported Research

Chiang is UMN Co-PI and **Leyk** is CU PI on a previous NSF IIS project. (a) **Grant information**. IIS #1564164, Funding: $904,000; Duration: 2016-2020. (b) **Project title**. III: Medium: Collaborative Research: Exploiting Context in Cartographic Evolutionary Documents to Extract and Build Linked Spatial-Temporal Datasets. (c) **Summary of results**. **Intellectual Merit:** We developed a framework that extracts, organizes, and links the knowledge found in map series [10, 80, 98, 99]. **Broader Impacts:** The produced map processing algorithms and systems have been deployed in open source projects at the Alan Turing Institute, Google, Facebook, and USC Digital Library to enable accurate extraction of metadata and geographic features from historical maps. A research paper presented at the 8th IAPR International Conference of Pattern Recognition Systems (ICPRS) won the Best Paper Award [97]. An undergraduate student working on this project won First Place at the ACM SIGSPATIAL Student Research Competition [57]. (d) **Publications**: [10, 80, 98–100]. (e) **Research products and their availability**. Software products are available on Git Hub and the project website [58].

Stefan Leyk is PI on a CMMI NSF project. (a) **Grant information**. CMMI#1924670, Funding: $450,000; Duration: 2019-2022. (b) **Project title**. The Creeping Disaster Along the Coast: Built Environment, Coastal Communities and Population Vulnerability to Sea Level Rise. (c) **Summary of results**. **Intellectual Merit**: The project explores ways to study the vulnerability of populations and the built environment to coastal flooding, advances the methodological and theoretical foundation of vulnerability analysis, and investigates the role of the built environment in mediating slow moving, persistent or creeping disasters. **Broader Impacts:** The project produces tools and disseminates data through HISDAC-US, critical for future disaster preparation and planning advised by NGO and industry partnerships and an expert advisory panel. This research received social and news media coverage, was featured in Science Advances and Nature Sustainability Community. (d) **Selected Publications:** [4, 5, 33, 34, 44, 51, 54, 69, 86, 95, 101]. (e) **Research products and their availability**: Several public HISDAC-US data products, e.g., [52, 53, 64, 65, 89–92, 96] were produced and are available for download from the Harvard Dataverse [30].

Stefan Leyk is PI on a BCS/ HNDS-I NSF project. (a) **Grant information**. BCS#2121976, Funding: $562,092; Duration: 2021-2024. (b) **Project title**. Collaborative Research: HNDS-I: Data Infrastructure for Research on Historical Settlement and Population Growth in the United States. (c) **Summary of results**. **Intellectual Merit**: The project produces new spatio-temporal data infrastructure of the built environment and population distribution over more than 150 years for most of the United States. The project develops efficient frameworks for spatial data integration and spatial refinement to process large amounts of data. **Broader Impacts:** The project produces tools and disseminates novel large-scale spatiotemporal data through the new edition of HISDAC-US and population grids over 150+ years. These data are critical for future research on the built environment, socio-environmental systems, and natural hazards risk assessments. (d) **Selected Publications:** [63, 87, 88, 93, 94, 102] (e) **Research products and their availability**: All public HISDAC-US data products v.2.0 and related data will become available for download from the Harvard Dataverse: e.g., [65, 89, 93].

References

[1] AI for Critical Mineral Assessment Competition: *https://criticalminerals.darpa.mil/The-Competition*. Accessed: 2024-01-26.

[2] Boeing, G. 2017. OSMnx: New methods for acquiring, constructing, analyzing, and visualizing complex street networks. *Computers, environment and urban systems*. 65, (Sep. 2017), 126–139. DOI:https://doi.org/10.1016/j.compenvurbsys.2017.05.004.

[3] Bojanowski, P., Grave, E., Joulin, A. and Mikolov, T. 2017. Enriching Word Vectors with Subword Information. *Transactions of the Association for Computational Linguistics*. 5, (2017), 135–146. DOI:https://doi.org/10.1162/tacl_a_00051.

[4] Braswell, A.E., Leyk, S., Connor, D.S. and Uhl, J.H. 2022. Creeping disaster along the U.S. coastline: Understanding exposure to sea level rise and hurricanes through historical development. *PloS one*. 17, 8 (Aug. 2022), e0269741. DOI:https://doi.org/10.1371/journal.pone.0269741.

[5] Burghardt, K., Uhl, J.H., Lerman, K. and Leyk, S. 2022. Road network evolution in the urban and rural United States since 1900. *Computers, environment and urban systems*. 95, (Jul. 2022), 101803. DOI:https://doi.org/10.1016/j.compenvurbsys.2022.101803.

[6] Chameleon Cloud: *http://www.chameleoncloud.org/*.

[7] Chen, J., Lu, Y., Yu, Q., Luo, X., Adeli, E., Wang, Y., Lu, L., Yuille, A.L. and Zhou, Y. 2021. TransUNet: Transformers Make Strong Encoders for Medical Image Segmentation. *arXiv [cs.CV]*.

[8] Chiang, Y.-Y. 2015. Querying Historical Maps As a Unified, Structured, and Linked Spatiotemporal Source: Vision Paper. *Proceedings of the 23rd SIGSPATIAL International Conference on Advances in Geographic Information Systems* (New York, NY, USA, Nov. 2015), 16:1–16:4.

[9] Chiang, Y.Y., Chen, M., Duan, W. and Kim, J. 2023. GeoAI for the Digitization of Historical Maps. *of Geospatial Artificial …*. (2023). DOI:https://doi.org/10.1201/9781003308423-11/geoai-digitization-historical-maps-yao-yi-chiang-muhao-chen-weiwei-duan-jina-kim-craig-knoblock-stefan-leyk-zekun-li-yijun-lin-min-namgung-basel-shbita-johannes-uhl.

[10] Chiang, Y.-Y., Duan, W., Leyk, S., Uhl, J.H. and Knoblock, C.A. 2020. *Using Historical Maps in Scientific Studies: Applications, Challenges, and Best Practices*. Springer, Cham.

[11] Chiang, Y.-Y. and Knoblock, C.A. 2013. A general approach for extracting road vector data from raster maps. *International Journal on Document Analysis and Recognition*. 16, 1 (2013), 55–81.

[12] Chiang, Y.Y. and Leyk, S. 2015. Exploiting online gazetteer for fully automatic extraction of cartographic symbols. *Proceedings of the 27th International Cartographic Conference ICC* (2015), 23–28.

[13] Chiang, Y.-Y., Leyk, S. and Knoblock, C.A. 2014. A survey of digital map processing techniques. *ACM Computing Surveys*. 47, 1 (2014), 1–44.

[14] Cohen, J. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*. 20, (1960), 37–46.

[15] CyberGISX: *https://cybergisxhub.cigi.illinois.edu/*. Accessed: 2024-01-26.

[16] DockerHub: *https://hub.docker.com/*. Accessed: 2024-01-26.

[17] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J. and Houlsby, N. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. *ICLR* (2021).

[18] Duan, W and Chiang, Y.-Y., and Leyk, S. and Uhl, J. and Knoblock, C. A. 2021. Guided Generative Models using Weak Supervision for Detecting Object Spatial Arrangement in Overhead Images. *2021 IEEE International Conference on Big Data (Big Data)* (2021).
[19] Duan, W., Chiang, Y.-Y. and Knoblock, C.A. 2023. LDTR: Linear Object Detection Transformer for Accurate Graph Generation by Learning the N-hop Connectivity Information. *In Preparation* (2023).
[20] Duan, W., Chiang, Y.-Y., Knoblock, C.A., Uhl, J.H. and Leyk, S. 2018. Automatic Generation of Precisely Delineated Geographic Features from Georeferenced Historical Maps Using Deep Learning. *Proceedings of the AutoCarto* (2018).
[21] Duan, W., Chiang, Y.-Y., Leyk, S., Uhl, J. and Knoblock, C.A. 2021. A Label Correction Algorithm Using Prior Information for Automatic and Accurate Geospatial Object Recognition. *2021 IEEE International Conference on Big Data (Big Data)* (2021).
[22] Forbes, A.D. 1995. Classification algorithm evaluation: five performance measures based on confusion matrices. *Journal of clinical monitoring and computing*. 11, (1995), 189–206.
[23] Frischknecht, S. and Kanani, E. 1998. Automatic interpretation of scanned topographic maps: A raster-based approach. *Graphics Recognition Algorithms and Systems*. K. Tombre and A. Chhabra, eds. 207–220.
[24] GeoBlacklight: *https://geoblacklight.org/*. Accessed: 2024-01-26.
[25] GitHub: *http://github.com*.
[26] Halilaj, L., Luettin, J., Monka, S., Henson, C. and Schmid, S. 2023. Knowledge Graph-Based Integration of Autonomous Driving Datasets. *International journal of semantic computing*. 17, 02 (Jun. 2023), 249–271. DOI:https://doi.org/10.1142/S1793351X23600048.
[27] Harvard Dataverse: *https://dataverse.harvard.edu/*. Accessed: 2021-11-25.
[28] Heipke, C., Mayer, H., Wiedemann, C. and Jamet, O. 1997. Evaluation of automatic road extraction. *International Archives of Photogrammetry and Remote Sensing*. 32, 3 (1997), 151–160.
[29] He, K., Gkioxari, G., Dollár, P. and Girshick, R. 2017. Mask r-cnn. *Proceedings of the IEEE international conference on computer vision* (2017), 2961–2969.
[30] HISDAC-US: Historical Settlement Data Compilation for the United States: *https://dataverse.harvard.edu/dataverse/hisdacus*. Accessed: 2024-01-26.
[31] Holmes-Wong, D. 2017. Unlocking Maps for Discovery and Other Purposes.
[32] Hornsby, K. and Egenhofer, M.J. 2000. Identity-based change: a foundation for spatio-temporal knowledge representation. *International journal of geographical information science: IJGIS*. 14, 3 (2000), 207–224.
[33] Iglesias, V., Braswell, A.E., Rossi, M.W., Joseph, M.B., McShane, C., Cattau, M., Koontz, M.J., McGlinchy, J., Nagy, R.C., Balch, J., Leyk, S. and Travis, W.R. 2021. Risky Development: Increasing Exposure to Natural Hazards in the United States. *Earth's future*. 9, 7 (Jul. 2021), e2020EF001795. DOI:https://doi.org/10.1029/2020EF001795.
[34] Iglesias, V., Stavros, N., Balch, J.K., Barrett, K., Cobian-Iñiguez, J., Hester, C., Kolden, C.A., Leyk, S., Chelsea Nagy, R., Reid, C.E., Wiedinmyer, C., Woolner, E. and Travis, W.R. 2022. Fires that matter: reconceptualizing fire risk to include interactions between humans and the natural environment. *Environmental research letters: ERL [Web site]*. 17, 4 (Mar. 2022), 045014. DOI:https://doi.org/10.1088/1748-9326/ac5c0c.
[35] Ilievski, F., Szekely, P. and Zhang, B. 2021. CSKG: The CommonSense Knowledge Graph. *Eighteenth Extended Semantic Web Conference - Resources Track* (2021).
[36] Joulin, A., Grave, E., Bojanowski, P., Douze, M., Jegou, H. and Mikolov, T. 2016. FastText.zip: Compressing text classification models. *https://openreview.net › forumhttps://openreview.net › forum*.
[37] Joulin, A., Grave, E., Bojanowski, P. and Mikolov, T. 2016. Bag of Tricks for Efficient Text Classification. *arXiv [cs.CL]*.
[38] Kang, Y., Gao, S. and Roth, R.E. Artificial intelligence studies in cartography: a review and

synthesis of methods, applications, and ethics. *Cartography and geographic information science*. 1–32. DOI:https://doi.org/10.1080/15230406.2023.2295943.
[39] Keahey, K., Riteau, P., Stanzione, D., Cockerill, T., Mambretti, J., Rad, P. and Ruth, P. 2019. Chameleon: a scalable production testbed for computer science research. *Contemporary High Performance Computing*. CRC Press. 123–148.
[40] Khotanzad, A. and Zink, E. 1996. Color paper map segmentation using eigenvector line-fitting. *Proceedings of the IEEE Southwest Symposium on Image Analysis and Interpretation* (1996), 190–194.
[41] Khotanzad, A. and Zink, E. 2003. Contour line and geographic feature extraction from USGS color topographical paper maps. *IEEE transactions on pattern analysis and machine intelligence*. 25, 1 (2003), 18–31.
[42] Kim, J., Li, Z., Lin, Y., Namgung, M., Jang, L. and Chiang, Y.-Y. 2023. The mapKurator System: A Complete Pipeline for Extracting and Linking Text from Historical Maps. *Proceedings of the 31st ACM International Conference on Advances in Geographic Information Systems* (New York, NY, USA, Dec. 2023), 1–4.
[43] Leyk, S. 2010. Segmentation of colour layers in historical maps based on hierarchical colour sampling. *Graphics Recognition, GREC 2009, Lecture Notes in Computer Science 6020*. O. J.-M., L. W., and L. J, eds. 231–241.
[44] Leyk, S. et al. 2019. The spatial allocation of population: a review of large-scale gridded population data products and their fitness for use. *Earth System Science Data*. 11, 3 (Sep. 2019), 1385–1409. DOI:https://doi.org/10.5194/essd-11-1385-2019.
[45] Leyk, S. and Boesch, R. 2010. Colors of the past: color image segmentation in historical topographic maps based on homogeneity. *GeoInformatica*. 14, 1 (2010), 1–21.
[46] Leyk, S. and Boesch, R. 2009. Extracting composite cartographic area features in low-quality maps. *Cartography and Geographical Information Science*. 36, 1 (2009), 71–79.
[47] Leyk, S., Boesch, R. and Weibel, R. 2005. A conceptual framework for uncertainty investigation in map-based land cover change modelling. *Transactions in GIS*. 9, 3 (2005), 291–322.
[48] Leyk, S., Boesch, R. and Weibel, R. 2006. Saliency and semantic processing - extracting forest cover from historical topographic maps. *Pattern recognition*. 39, 5 (2006), 953–968.
[49] Leyk, S., Nagle, N.N. and Buttenfield, B. 2013. Maximum entropy dasymetric modeling for demographic small area estimation under uncertainty. *Geographical analysis*. 45, 3 (2013), 285–306.
[50] Leyk, S., Ruther, M., Buttenfield, B.P., Nagle, N.N. and Stum, A.K. 2014. Modeling residential developed land in rural areas: A size-restricted approach using parcel data. *Applied geography* . 47, 1 (2014), 33–45.
[51] Leyk, S. and Uhl, J.H. 2018. HISDAC-US, historical settlement data compilation for the conterminous United States over 200 years. *Scientific data*. 5, (2018), 180175.
[52] Leyk, S. and Uhl, J.H. 2018. Historical built-up intensity layer series for the U.S. 1810 - 2015. Harvard Dataverse.
[53] Leyk, S. and Uhl, J.H. 2018. Historical settlement composite layer for the U.S. 1810 - 2015. Harvard Dataverse.
[54] Leyk, S., Uhl, J.H., Connor, D.S., Braswell, A.E., Mietkiewicz, N., Balch, J.K. and Gutmann, M. 2020. Two centuries of settlement and urban development in the United States. *Science advances*. 6, 23 (Jun. 2020), eaba2937. DOI:https://doi.org/10.1126/sciadv.aba2937.
[55] Leyk, S. and Zimmermann, N.E. 2004. A predictive uncertainty model for field-based survey maps using generalized linear models. *Proceedings of the 3rd International conference on Geographic Information Science (GIScience 2004)* (2004), 191–205.
[56] Leyk, S. and Zimmermann, N.E. 2007. Improving land change detection based on uncertain survey maps using fuzzy sets. *Landscape ecology*. 22, (2007), 257–272.
[57] Lin, H. and Chiang, Y.-Y. 2018. SRC: automatic extraction of phrase-level map labels from

historical maps. *SIGSPATIAL Special*. 9, 3 (Jan. 2018), 14–15. DOI:https://doi.org/10.1145/3178392.3178400.
[58] LinkedMap: *https://usc-isi-i2.github.io/linked-maps/*.
[59] Li, Z., Chiang, Y.Y., Tavakkol, S., Shbita, B. and Uhl, J.H. 2020. An Automatic Approach for Generating Rich, Linked Geo-Metadata from Historical Map Images. *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (2020).
[60] Li, Z., Gao, L., Gao, Y., Li, X. and Li, H. 2022. Zero-shot surface defect recognition with class knowledge graph. *Advanced Engineering Informatics*. 54, (Oct. 2022), 101813. DOI:https://doi.org/10.1016/j.aei.2022.101813.
[61] Li, Z., Guan, R., Yu, Q., Chiang, Y.-Y. and Knoblock, C.A. 2021. Synthetic Map Generation to Provide Unlimited Training Data for Historical Map Text Detection. *Proceedings of the 4th ACM SIGSPATIAL International Workshop on AI for Geographic Knowledge Discovery* (New York, NY, USA, Nov. 2021), 17–26.
[62] Logan, J.R., Xu, Z. and Stults, B. 2014. Interpolating U.S. Decennial Census Tract Data from as Early as 1970 to 2010: A Longtitudinal Tract Database. *The Professional geographer: the journal of the Association of American Geographers*. 66, 3 (Jul. 2014), 412–420. DOI:https://doi.org/10.1080/00330124.2014.905156.
[63] Mahood, A.L. et al. 2023. Ten simple rules for working with high resolution remote sensing data. *Peer Community Journal*. 3, (2023). DOI:https://doi.org/10.24072/pcjournal.223.
[64] Mc Shane, C., Uhl, J.H. and Leyk, S. 2021. Historical land use for the U.s. 1940-2015: Class counts. Harvard Dataverse.
[65] Mc Shane, C., Uhl, J.H. and Leyk, S. 2021. Historical land use for the U.s. 1940-2015: Major class. Harvard Dataverse.
[66] Mei, J., Li, R.-J., Gao, W. and Cheng, M.-M. 2021. CoANet: Connectivity Attention Network for Road Extraction From Satellite Imagery. *IEEE transactions on image processing: a publication of the IEEE Signal Processing Society*. 30, (Oct. 2021), 8540–8552. DOI:https://doi.org/10.1109/TIP.2021.3117076.
[67] Mennis, J. 2009. Dasymetric mapping for estimating population in small areas. *Geography Compass*. 3, 2 (2009), 727–745.
[68] Michie, D., Spiegelhalter, D. and Taylor, C. 1994. *Machine learning, neural and statistical classification*. Ellis Horwood.
[69] Mietkiewicz, N., Balch, J.K., Schoennagel, T., Leyk, S., St. Denis, L.A. and Bradley, B.A. 2020. In the line of fire: Consequences of human-ignited wildfires to homes in the U.s. (1992–2015). *Fire*. 3, 3 (Sep. 2020), 50. DOI:https://doi.org/10.3390/fire3030050.
[70] Monka, S., Halilaj, L. and Rettinger, A. A Survey on Visual Transfer Learning using Knowledge Graphs. *semantic-web-journal.net*.
[71] Monka, S., Halilaj, L. and Rettinger, A. 2022. Context-Driven Visual Object Recognition Based on Knowledge Graphs. *The Semantic Web – ISWC 2022* (2022), 142–160.
[72] Monka, S., Halilaj, L., Schmid, S. and Rettinger, A. 2021. ConTraKG: Contrastive-based Transfer Learning for Visual Object Recognition using Knowledge Graphs. *arXiv [cs.CV]*.
[73] National Covenants Research Coalition: *https://www.nationalcovenantsresearchcoalition.com/*. Accessed: 2024-01-26.
[74] OpenGeoMetadata: *https://opengeometadata.org/*. Accessed: 2024-01-26.
[75] Reibel, M. and Bufalino, M.E. 2005. Street-Weighted Interpolation Techniques for Demographic Count Estimation in Incompatible Zone Systems. *Environment & planning A*. 37, 1 (Jan. 2005), 127–139. DOI:https://doi.org/10.1068/a36202.
[76] Ruther, M., Leyk, S. and Buttenfield, B.P. 2015. Comparing the effects of an NLCD derived dasymetric refinement on estimation accuracies for multiple areal interpolation methods. *GIScience and Remote Sensing*. 52, 2 (2015), 158–178.
[77] van de Sandt, S., Nielsen, L.H., Ioannidis, A., Muench, A., Henneken, E., Accomazzi, A.,

Bigarella, C., Lopez, J.B.G. and Dallmeier-Tiessen, S. 2019. Practice meets Principle: Tracking Software and Data Citations to Zenodo DOIs. *arXiv [cs.DL]*.
[78] Schroeder, J.P. 2017. Hybrid areal interpolation of census counts from 2000 blocks to 2010 geographies. *Computers, environment and urban systems*. 62, (2017), 53–63.
[79] Schroeder, J.P. and Van Riper, D.C. 2013. Because Muncie's densities are not Manhattan's: Using geographical weighting in the expectation–maximization algorithm for areal interpolation. *Geographical analysis*. 45, 3 (2013), 216–237.
[80] Shbita, B., Knoblock, C.A., Duan, W., Chiang, Y.-Y., Uhl, J.H. and Leyk, S. 2020. Building Linked Spatio-Temporal Data from Vectorized Historical Maps. *The Semantic Web* (2020), 409–426.
[81] Shit, S., Koner, R., Wittmann, B., Paetzold, J., Ezhov, I., Li, H., Pan, J., Sharifzadeh, S., Kaissis, G., Tresp, V. and Menze, B. 2022. Relationformer: A Unified Framework for Image-to-Graph Generation. *Computer Vision – ECCV 2022* (2022), 422–439.
[82] Standford Earthworks: *https://earthworks.stanford.edu/*. Accessed: 2024-01-26.
[83] Tao, C., Qi, J., Li, Y., Wang, H. and Li, H. 2019. Spatial information inference net: Road extraction using road-specific contextual information. *ISPRS journal of photogrammetry and remote sensing: official publication of the International Society for Photogrammetry and Remote Sensing* . 158, (Dec. 2019), 155–166. DOI:https://doi.org/10.1016/j.isprsjprs.2019.10.001.
[84] Tavakkol, S., Chiang, Y.-Y., Waters, T., Han, F., Prasad, K. and Kiveris, R. 2019. Kartta Labs: Unrendering Historical Maps. *Proceedings of the 3rd ACM SIGSPATIAL International Workshop on AI for Geographic Knowledge Discovery* (New York, NY, USA, Nov. 2019), 48–51.
[85] The MIT License: *https://opensource.org/license/mit/*. Accessed: 2024-01-26.
[86] Uhl, J.H., Connor, D.S., Leyk, S. and Braswell, A.E. 2021. A century of decoupling size and structure of urban spaces in the United States. *Communications earth & environment*. 2, (Jan. 2021). DOI:https://doi.org/10.1038/s43247-020-00082-7.
[87] Uhl, J.H. and Leyk, S. 2022. A scale-sensitive framework for the spatially explicit accuracy assessment of binary built-up surface layers. *Remote sensing of environment*. 279, (Sep. 2022), 113117. DOI:https://doi.org/10.1016/j.rse.2022.113117.
[88] Uhl, J.H. and Leyk, S. 2022. Assessing the relationship between morphology and mapping accuracy of built-up areas derived from global human settlement data. *GIScience and Remote Sensing*. 59, 1 (Oct. 2022), 1722–1748. DOI:https://doi.org/10.1080/15481603.2022.2131192.
[89] Uhl, J.H. and Leyk, S. 2022. Historical building footprint area (BUFA) - gridded surfaces for the conterminous U.S. from 1900 to 2010. Harvard Dataverse.
[90] Uhl, J.H. and Leyk, S. 2020. Historical built-up areas (BUA) - gridded surfaces for the U.S. from 1810 to 2015. Harvard Dataverse.
[91] Uhl, J.H. and Leyk, S. 2020. Historical built-up property locations (BUPL) - gridded surfaces for the U.S. from 1810 to 2015. Harvard Dataverse.
[92] Uhl, J.H. and Leyk, S. 2020. Historical built-up property records (BUPR) - gridded surfaces for the U.S. from 1810 to 2015. Harvard Dataverse.
[93] Uhl, J.H. and Leyk, S. 2022. MTBF-33: A multi-temporal building footprint dataset for 33 counties in the United States (1900 - 2015). *Data in brief*. 43, (Aug. 2022), 108369. DOI:https://doi.org/10.1016/j.dib.2022.108369.
[94] Uhl, J.H. and Leyk, S. 2023. Spatially explicit accuracy assessment of deep learning-based, fine-resolution built-up land data in the United States. *International Journal of Applied Earth Observation and Geoinformation*. 123, (Sep. 2023). DOI:https://doi.org/10.1016/j.jag.2023.103469.
[95] Uhl, J.H. and Leyk, S. 2020. Towards a novel backdating strategy for creating built-up land time series data using contemporary spatial constraints. *Remote sensing of environment*.

238, (Mar. 2020), 111197. DOI:https://doi.org/10.1016/j.rse.2019.05.016.
[96] Uhl, J.H. and Leyk, S. 2020. Uncertainty surfaces accompanying the BUPR, BUPL, and BUA gridded surface series. Harvard Dataverse.
[97] Uhl, J.H., Leyk, S., Chiang, Y.-Y., Duan, W. and Knoblock, C.A. 2017. Extracting Human Settlement Footprint from Historical Topographic Map Series Using Context-Based Machine Learning. *Proceedings of the IAPR 8th International Conference on Pattern Recognition Systems* (Madrid, Spain, 2017).
[98] Uhl, J.H., Leyk, S., Chiang, Y.-Y., Duan, W. and Knoblock, C.A. 2018. Map Archive Mining: Visual-Analytical Approaches to Explore Large Historical Map Collections. *ISPRS International Journal of Geo-Information*. 7, 4 (2018), 148. DOI:https://doi.org/10.3390/ijgi7040148.
[99] Uhl, J.H., Leyk, S., Chiang, Y.-Y., Duan, W. and Knoblock, C.A. 2018. Spatializing Uncertainty in Image Segmentation Using Weakly Supervised Convolutional Neural Networks: A Case Study from Historical Map Processing. *IET Image Processing*. 12, 11 (2018), 2084–2091.
[100] Uhl, J.H., Leyk, S., Chiang, Y.-Y. and Knoblock, C.A. 2022. Towards the automated large-scale reconstruction of past road networks from historical maps. *Computers, environment and urban systems*. 94, (Jun. 2022). DOI:https://doi.org/10.1016/j.compenvurbsys.2022.101794.
[101] Uhl, J.H., Leyk, S., McShane, C.M., Braswell, A.E., Connor, D.S. and Balk, D. 2020. Fine-grained, spatio-temporal datasets measuring 200 years of land development in the United States. *Earth System Science Data Discussions*. (2020), 1–43. DOI:https://doi.org/10.5194/essd-2020-217.
[102] Uhl, J.H., Royé, D., Burghardt, K., Aldrey Vázquez, J.A., Borobio Sanchiz, M. and Leyk, S. 2023. HISDAC-ES: historical settlement data compilation for Spain (1900–2020). *Earth system science data*. 15, 10 (Oct. 2023), 4713–4747. DOI:https://doi.org/10.5194/essd-15-4713-2023.
[103] Van Etten, A., Lindenbaum, D. and Bacastow, T.M. 2018. SpaceNet: A Remote Sensing Dataset and Challenge Series. *arXiv [cs.CV]*.
[104] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł.U. and Polosukhin, I. 2017. Attention is All you Need. *Advances in Neural Information Processing Systems* (2017), 6000–6010.
[105] Wright, J.K. 1936. A method of mapping densities of population: with Cape Cod as an example. *Geographical review*. 26, 1 (1936), 103–110.
[106] Xie, Y., Zhang, J., Shen, C. and Xia, Y. 2021. CoTr: Efficiently Bridging CNN and Transformer for 3D Medical Image Segmentation. *MICCAI* (Mar. 2021).
[107] Yu, R., Luo, Z. and Chiang, Y.-Y. 2016. Recognizing Text On Historical Maps Using Maps From Multiple Time Periods. *Proceedings of the 23rd International Conference on Pattern Recognition* (2016).
[108] Zenodo: *https://zenodo.org/*. Accessed: 2024-01-26.