Digital Humanities of/by/for "East Asia"

Asanobu KITAMOTO

Center for Open Data in the Humanities

National Institute of Informatics

http://codh.rois.ac.jp/

Digital Humanities

- **Humanities**: study of human society and culture, such as languages, literature, philosophy, geography, history, religion, art and musicology (Wikipedia).
- **Digital Humanities**: humanities research of digital tools, by digital tools, for digital tools.
- Annual “Digital Humanities” conference attracts 500-900 participants.

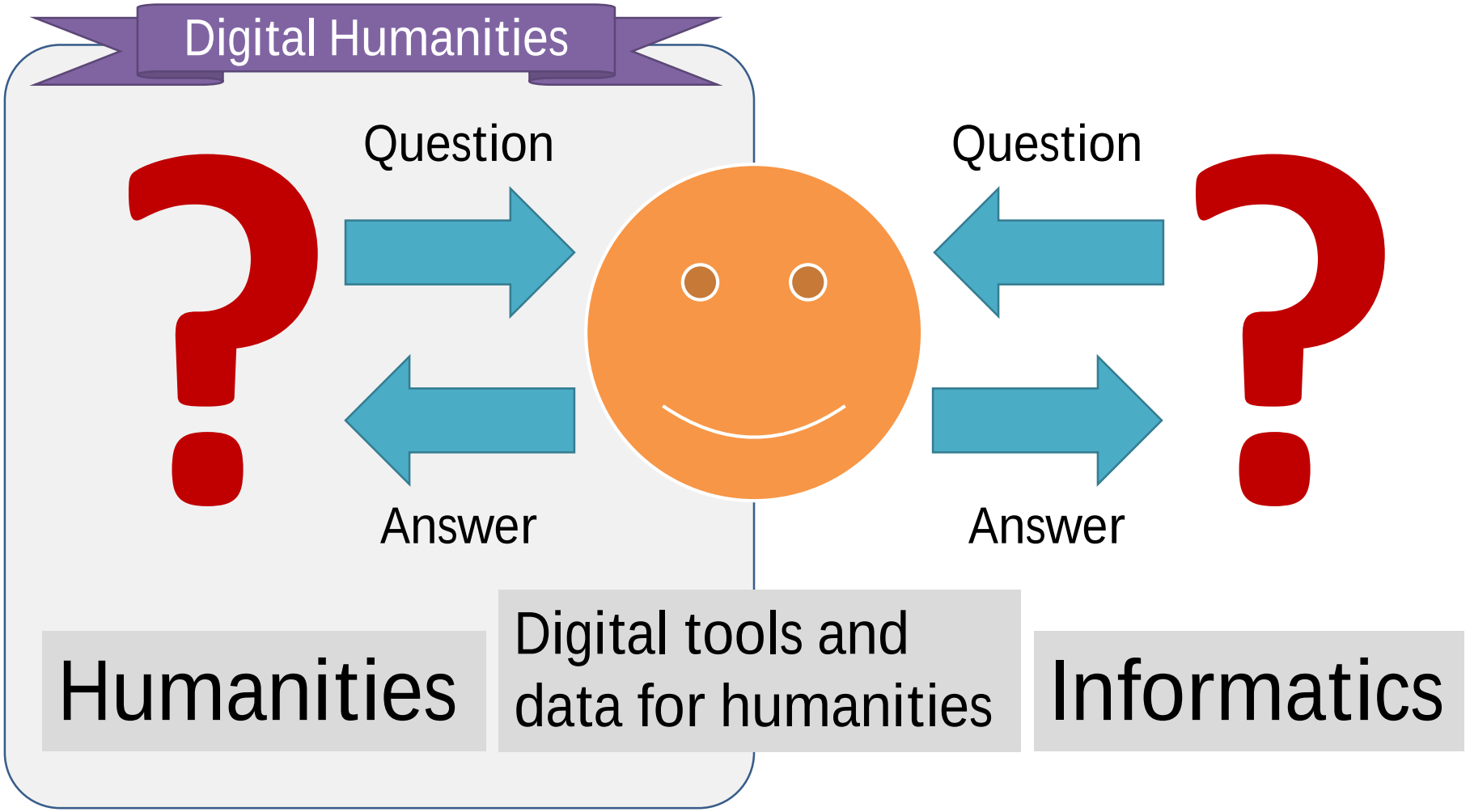
About CODH

http://codh.rois.ac.jp/

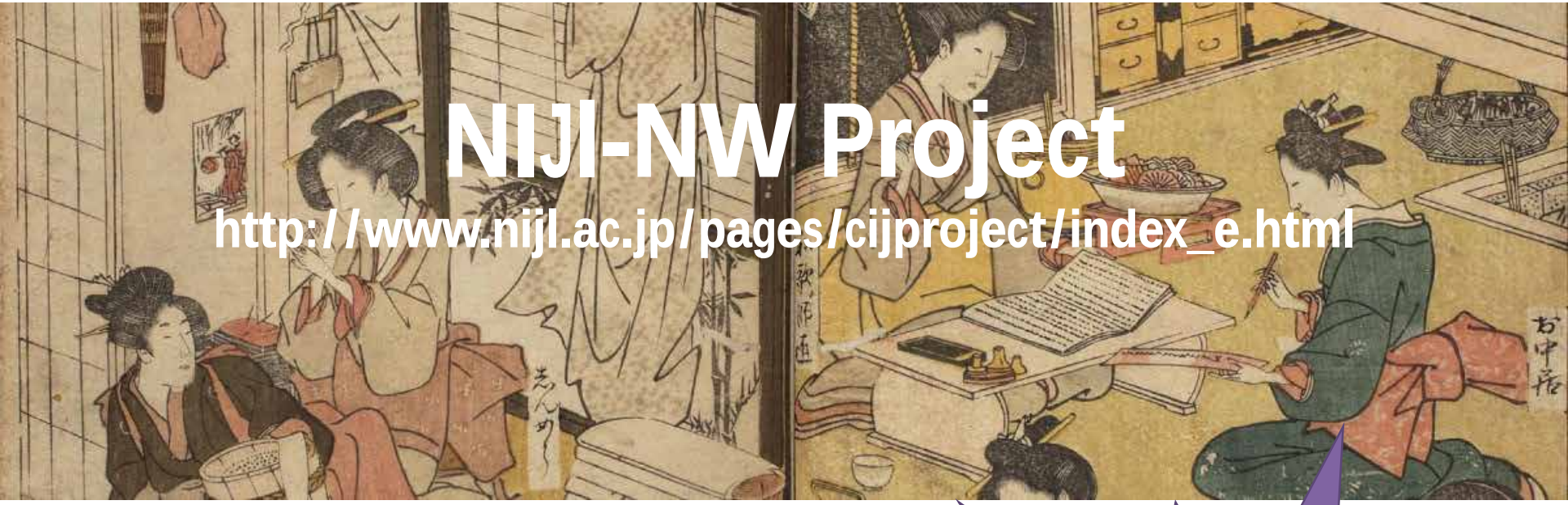
- April 1, 2017: **Center for Open Data in the Humanities (CODH)** was officially launched.
- Computer scientists and statisticians work with humanities scholars.

1. **Innovate humanities research by informatics and statistics technology.**
2. Innovate informatics and statistics research by humanities (big) data.

Research Questions and Answers



Digital Humanities of “East Asia”



NIJL-NW Project

http://www.nijl.ac.jp/pages/cijproject/index_e.html

It was decided to convert approximately **300 thousand** **"Pre-modern Japanese Books"** into image data to be amalgamated with the bibliographic data base to produce the "Database of Pre-modern Japanese Books."

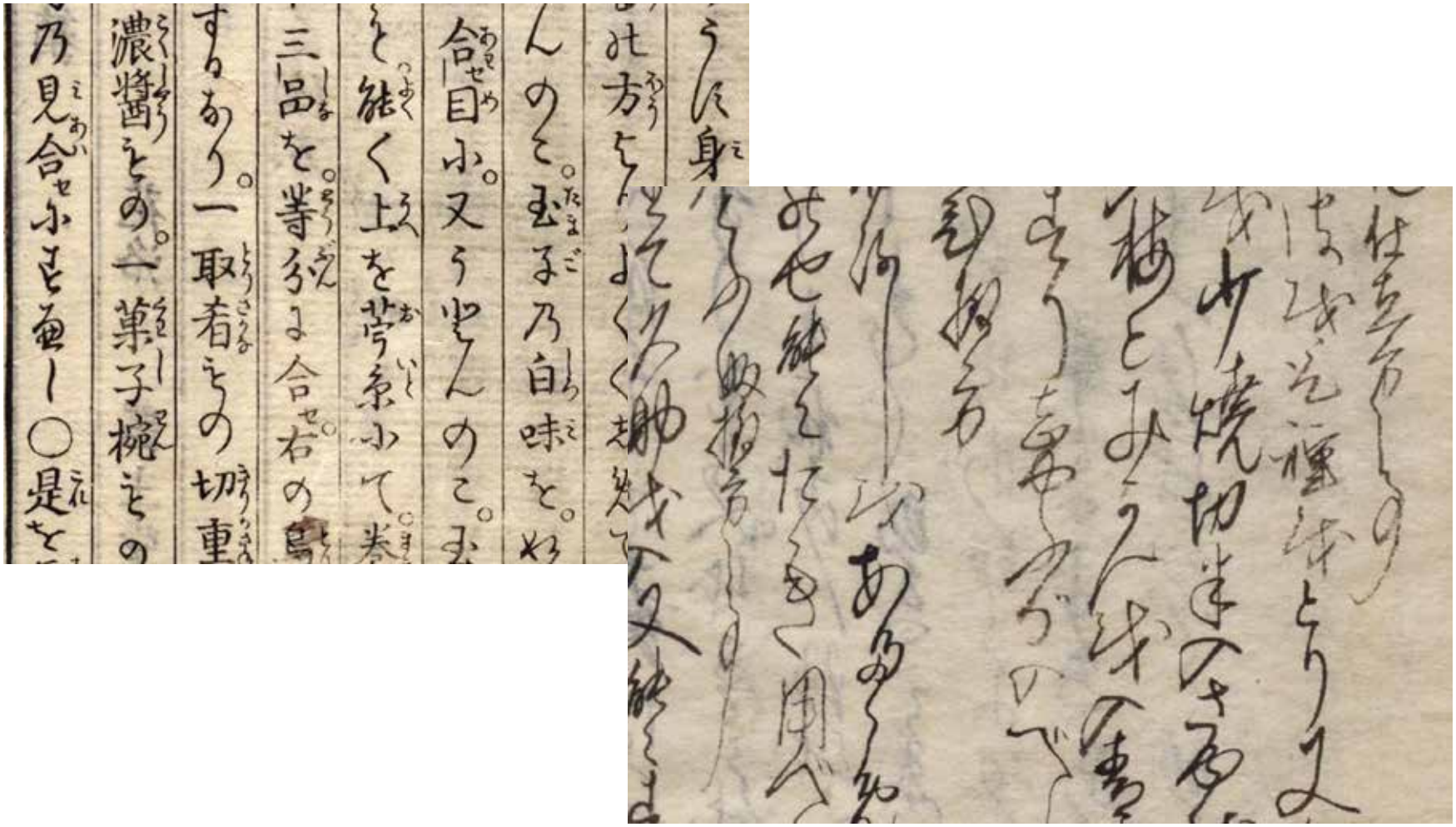
Research on Japanese culture finally entered into the big data era...

Open Data for Scholars

http://codh.rois.ac.jp/pmjt/

Pre-Modern Japanese Text Dataset (from NIJL)

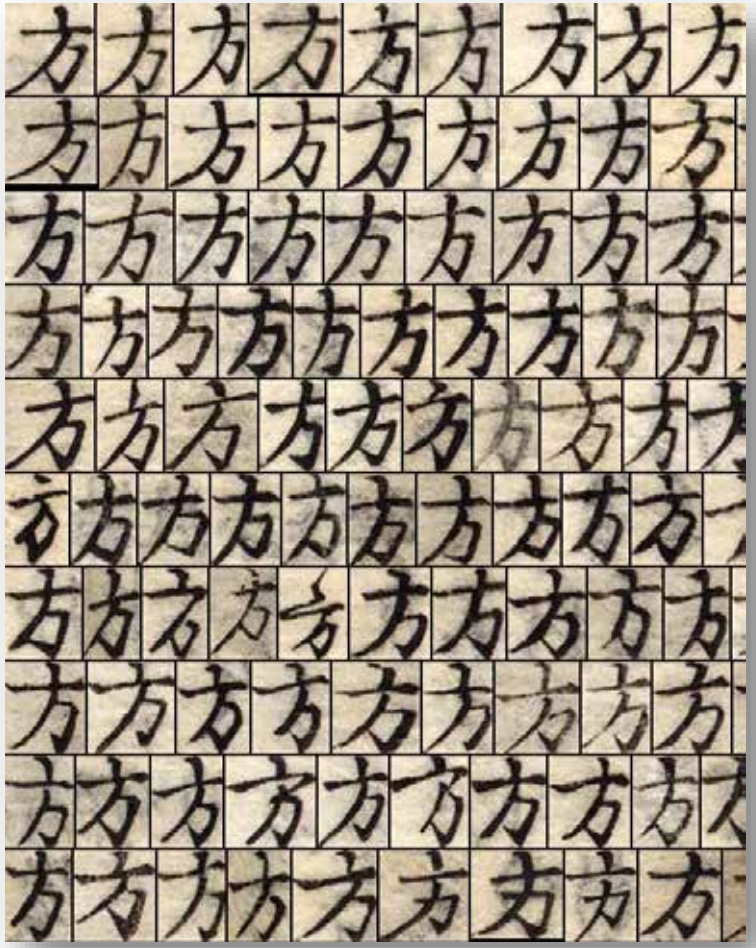
The Problem of Characters



Japanese Old Books Forgotten

1. Japan had a very active publishing industry that produced many books every year.
2. Even native Japanese are not good readers of books published just 150 years ago!
3. Characters and writing styles have changed, and people rely on modern translation.
4. **Big imbalance between books and readers = unique situation in the world!**

Kuzushi-ji Challenge!



http://codh.rois.ac.jp/char-shape/

- A dataset of 403,242 characters with 3,999 character types was released as open data.
- Hold a competition to develop machine learning algorithms for Kuzushi-ji OCR.

Results of Competition

The IAPR Best Paper Award of HIP 2017 is presented to

Hung Tuan Nguyen, Nam Tuan Ly, Kha Cong Nguyen, Cuong Tuan Nguyen and Masaki Nakagawa

for their paper entitled

Attempts to recognize anomalously deformed Kana in Japanese historical documents

We congratulate the authors for their outstanding work, which is breaking ground for the challenging problem of Kana recognition in historical Japanese documents, thoroughly investigates state-of-the-art methods at different recognition levels, and promotes the use of open data for historical document imaging and processing.

Andreas Fischer, Angelika Garz, Kengo Terasawa, and Bill Barrett
IAPR Best Paper Award Committee

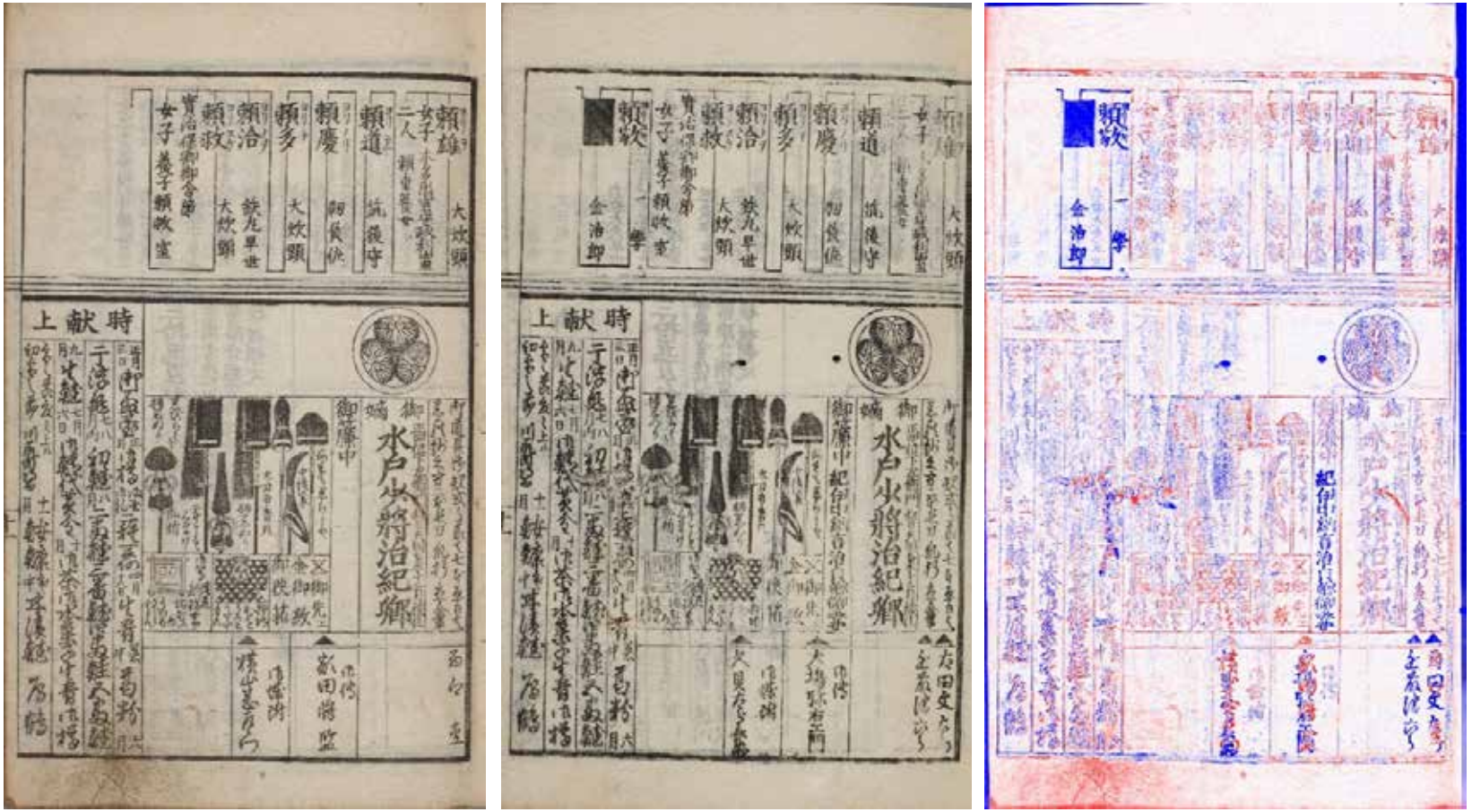
http://events.unifr.ch/hip2017/

- The winner was a Vietnamese student team in a Japanese university.
- 2nd competition in 2018? Keep track of the progress and share knowledge within the community.

Crowd-Sourced Transcription

Crowd sourcing platform transcribed about 3.5 million characters in one year.

Differential Reading



Left: “Bukan” in 1789, Middle: “Bukan” in 1791, Right: Comparison of versions.

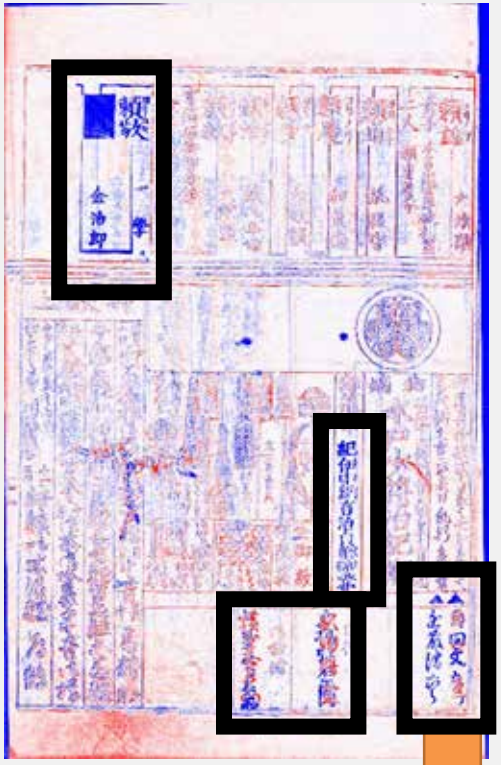
Time-Series Historical Sources

- Bukan: directory of state king families and bureaucrats of the central government in the Edo period (1603-1868) .
- Time-series publications for 100 to 200 years with a peak frequency of a few times in a month.
- 381 versions of Bukan will be released as open data.

Human-Machine Specialization

Many versions of "Bukan"

Comparison with the base version



Change detection

Machine's Role

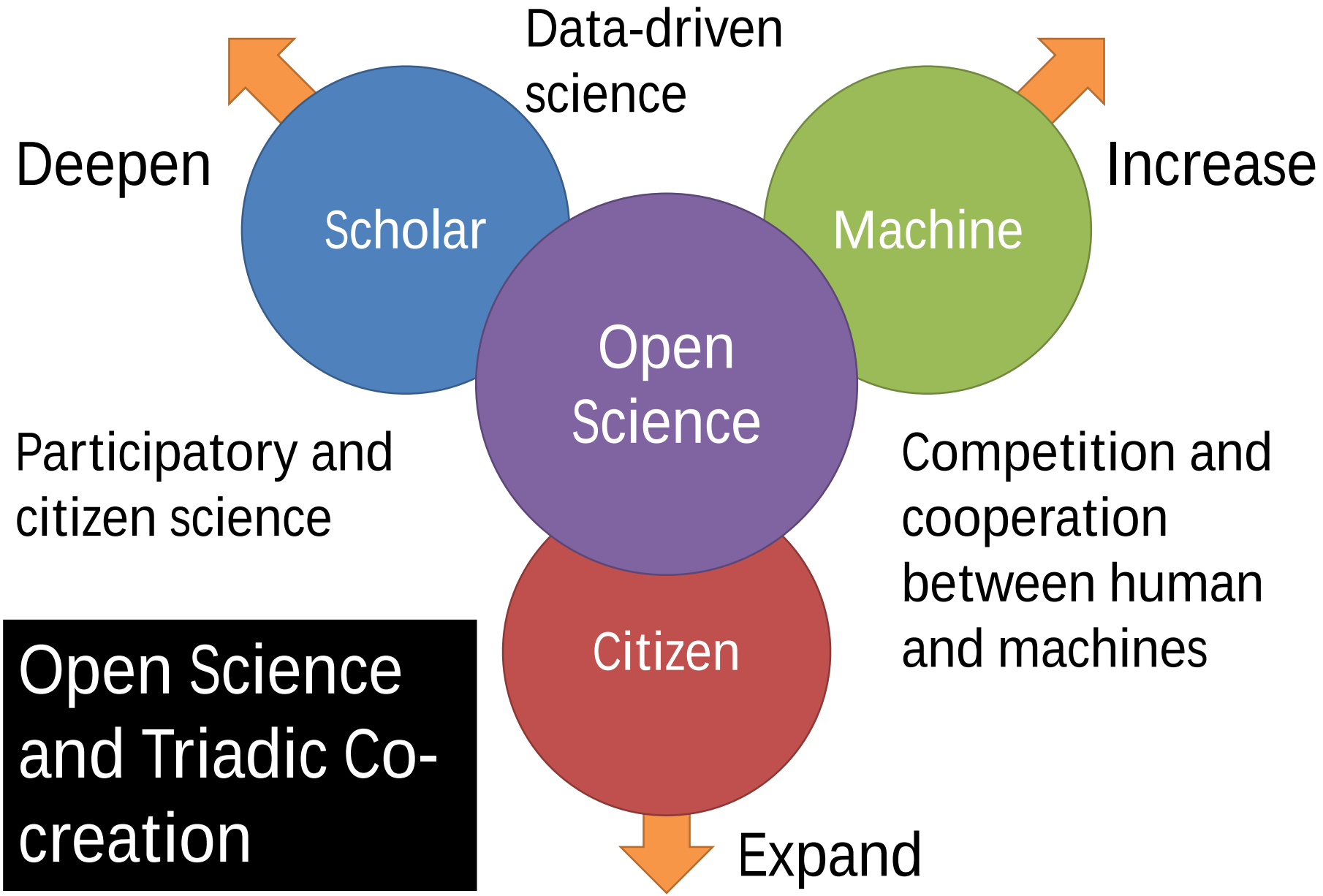
Traditional

Whole transcription

Cost Reduction

Differential transcription

Human 's Role



Open Science and Triadic Co-creation

Historical Big Data Challenge

Personal or family diaries

Weather, earthquake, ..
Events, accidents,..
Human relationships, ..

Official records

Economies, industry, ..
Organization and promotion ..

Travelogues and reports

Places and regions,..
Human mobility ..

Transcribed or OCRed

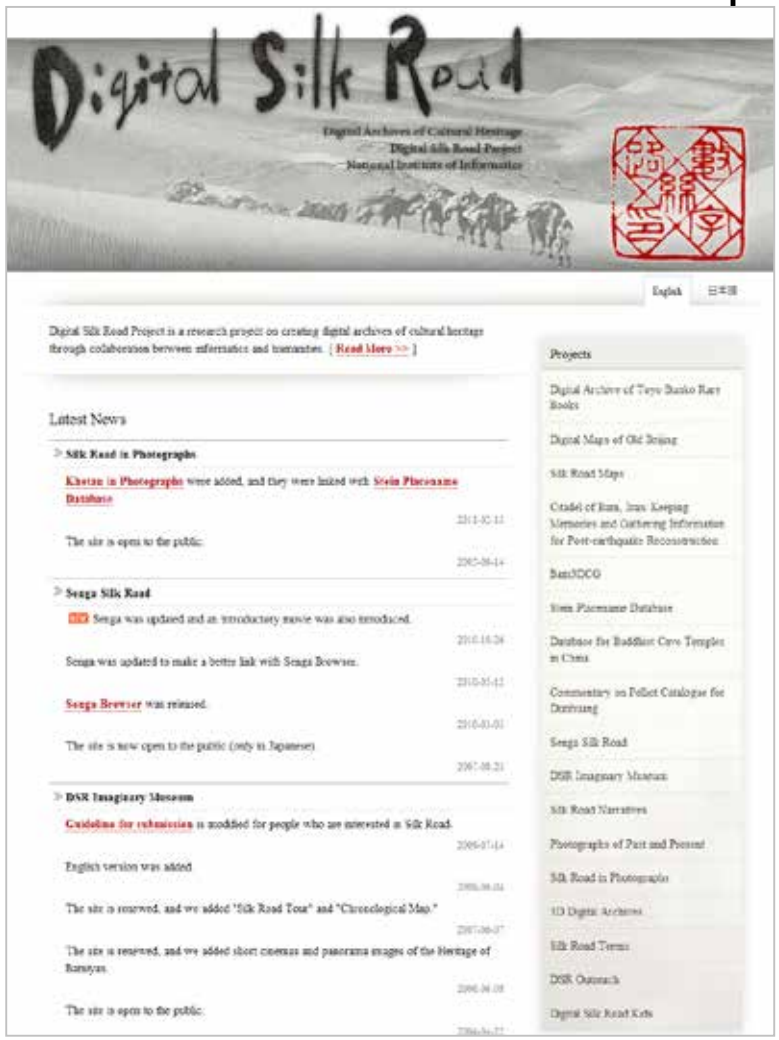
Filtered

Structured

Re-analyze the historical world and society.

Digital Silk Road

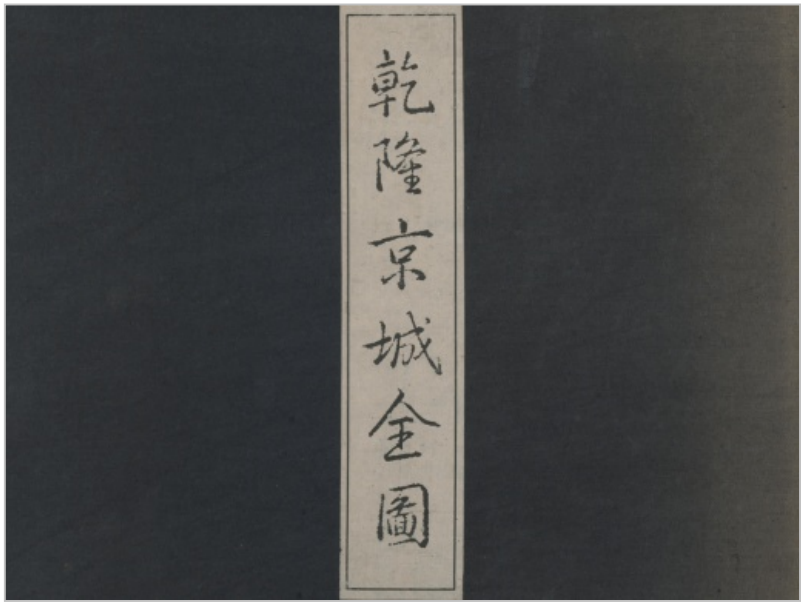
http://dsr.nii.ac.jp/

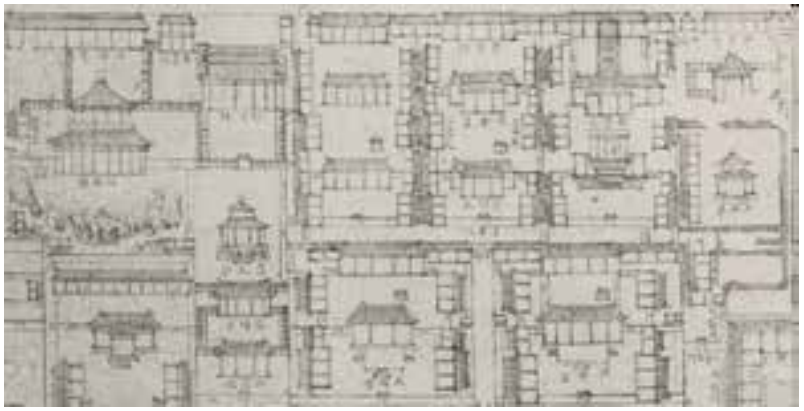


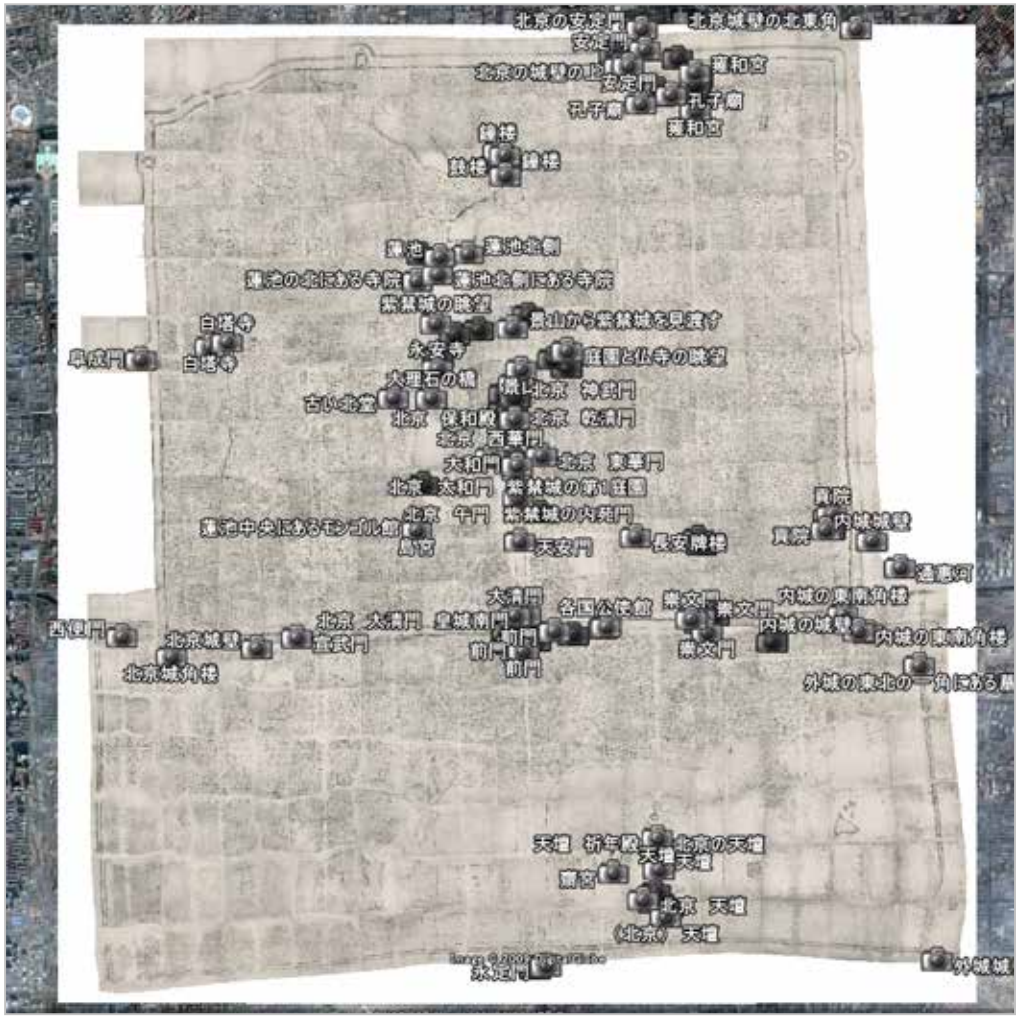
- Started in 2001 with UNESCO to make the digital archive of Silk Road.
- Digitized books, maps, photographs, essays, databases, 3D models, and many others are open on the Web.

Qianlong Map (Beijing)

http://dsr.nii.ac.jp/beijing-maps/

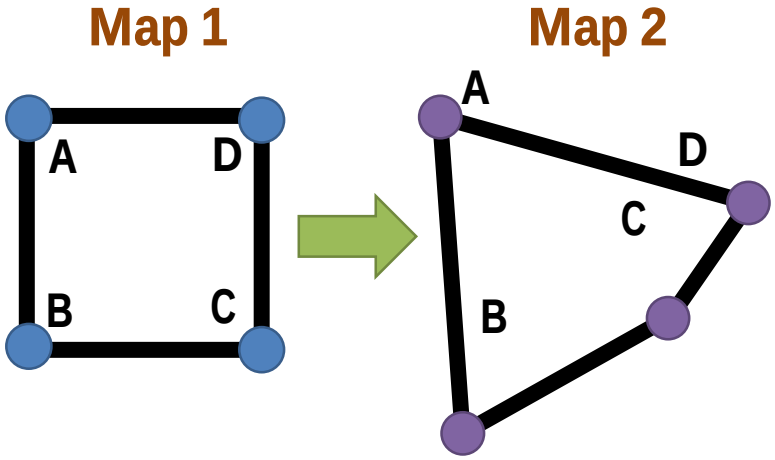






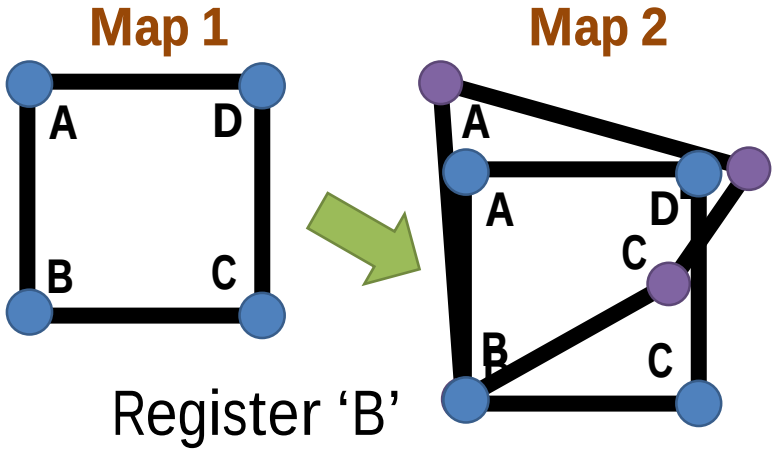
Map Registration Methods

Geometric Correction



- All points are registered.
- Shapes are distorted.

Interactive Georeferencing

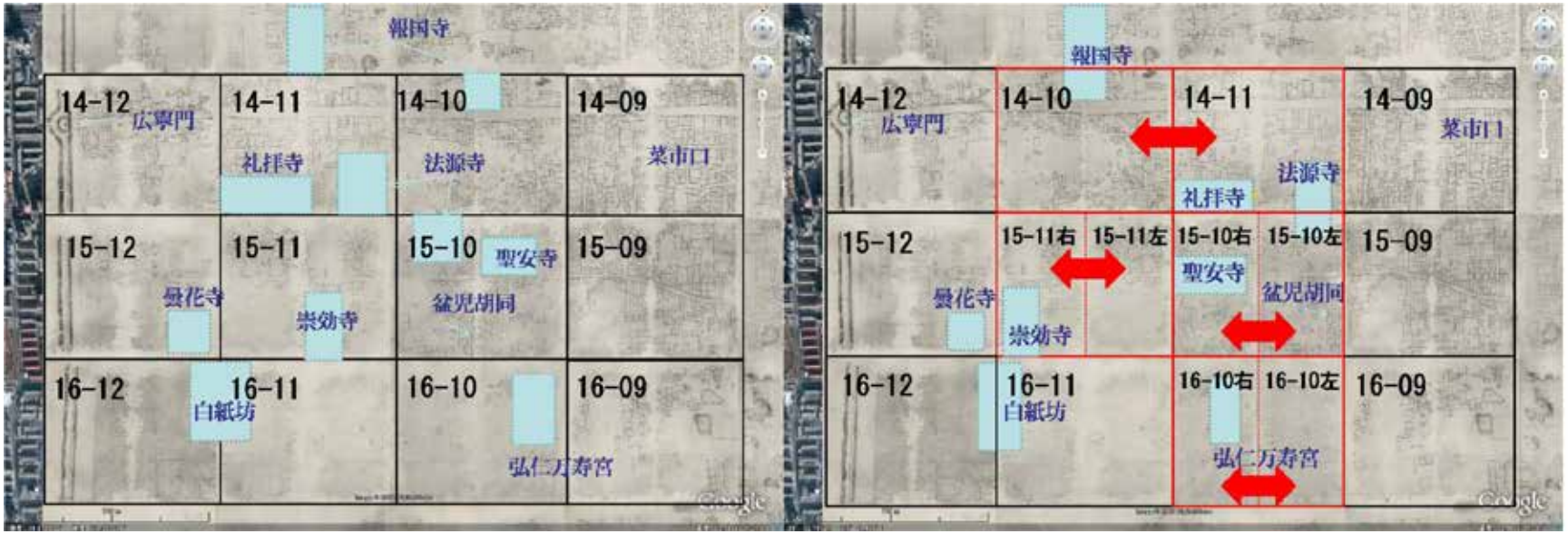


- Single point is registered (but no other points).
- Shapes are not distorted.

Massive Geometric Correction

- **A map of Beijing created around 1750.**
- **Huge:** W13m x H14m
- **Fragmented:** 203 sheets
- **Massive:** 29 billion pixels
- **Geometric correction:** grid structure should be preserved using control lines in addition to points.

Question and Discovery

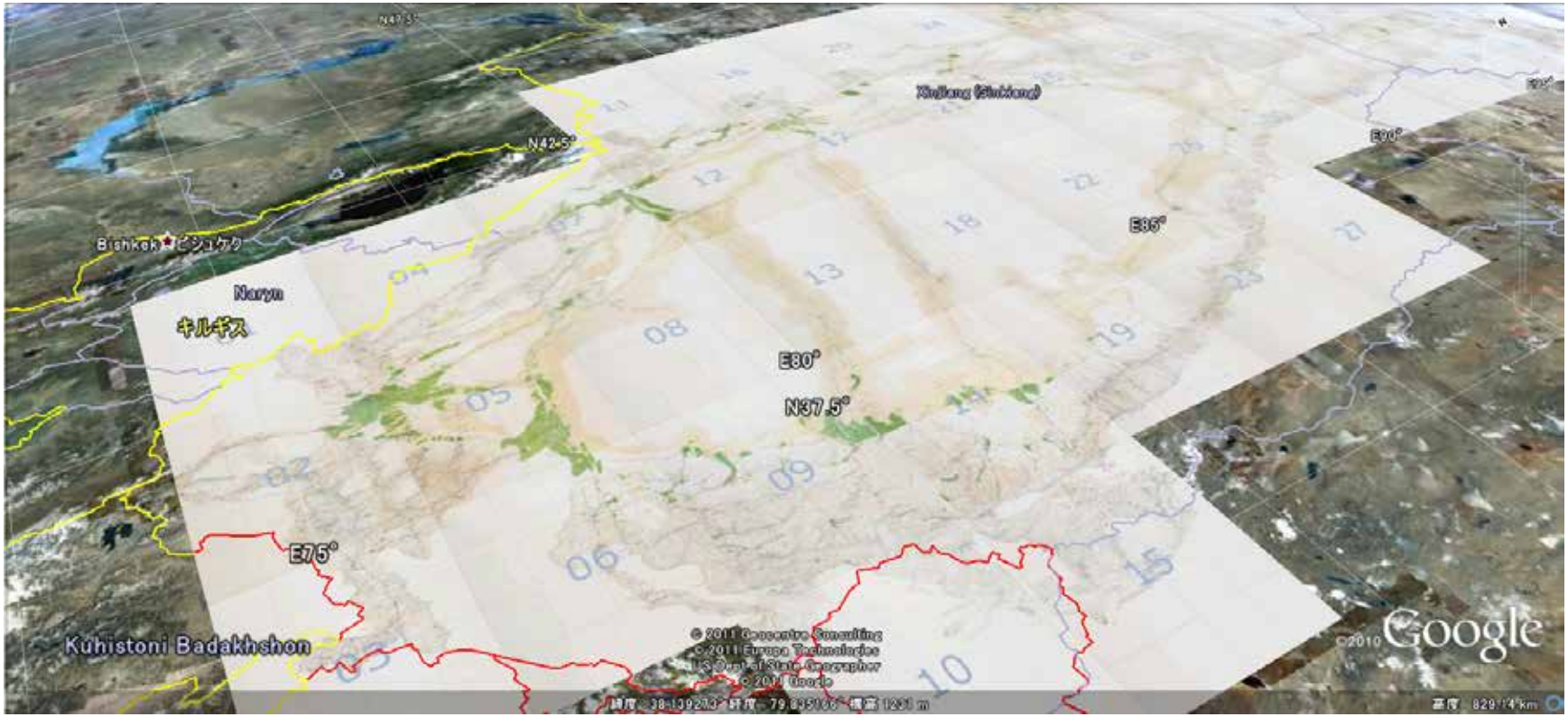


- **Question:** some areas of the map do not match well with the current city. Why?
- **Discovery**: because five sheets are mis-arranged due to improper reconstruction in the past.

Stein Map (Silk Road)

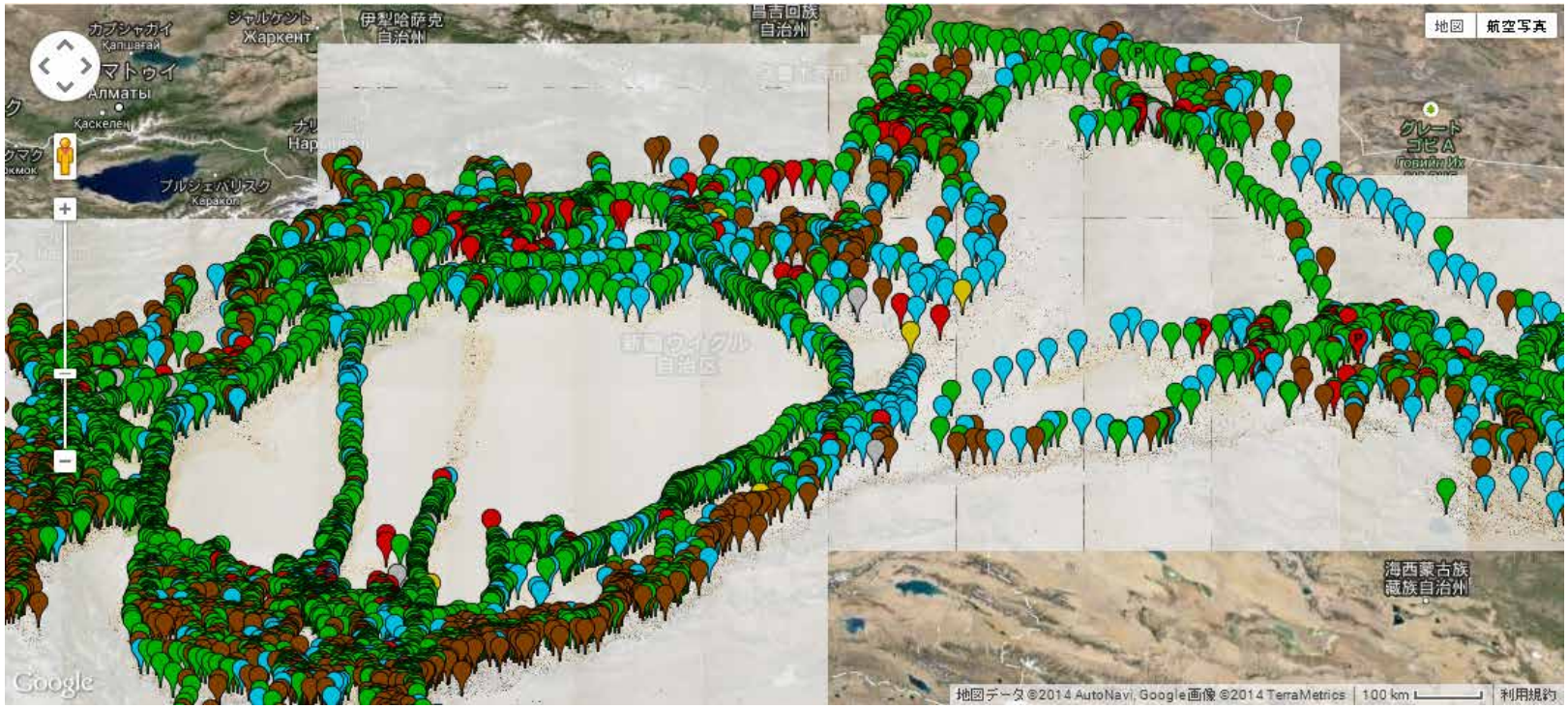
http://dsr.nii.ac.jp/geography/

A map considered as trustworthy caused the problem of missing ruins due to inaccuracy.



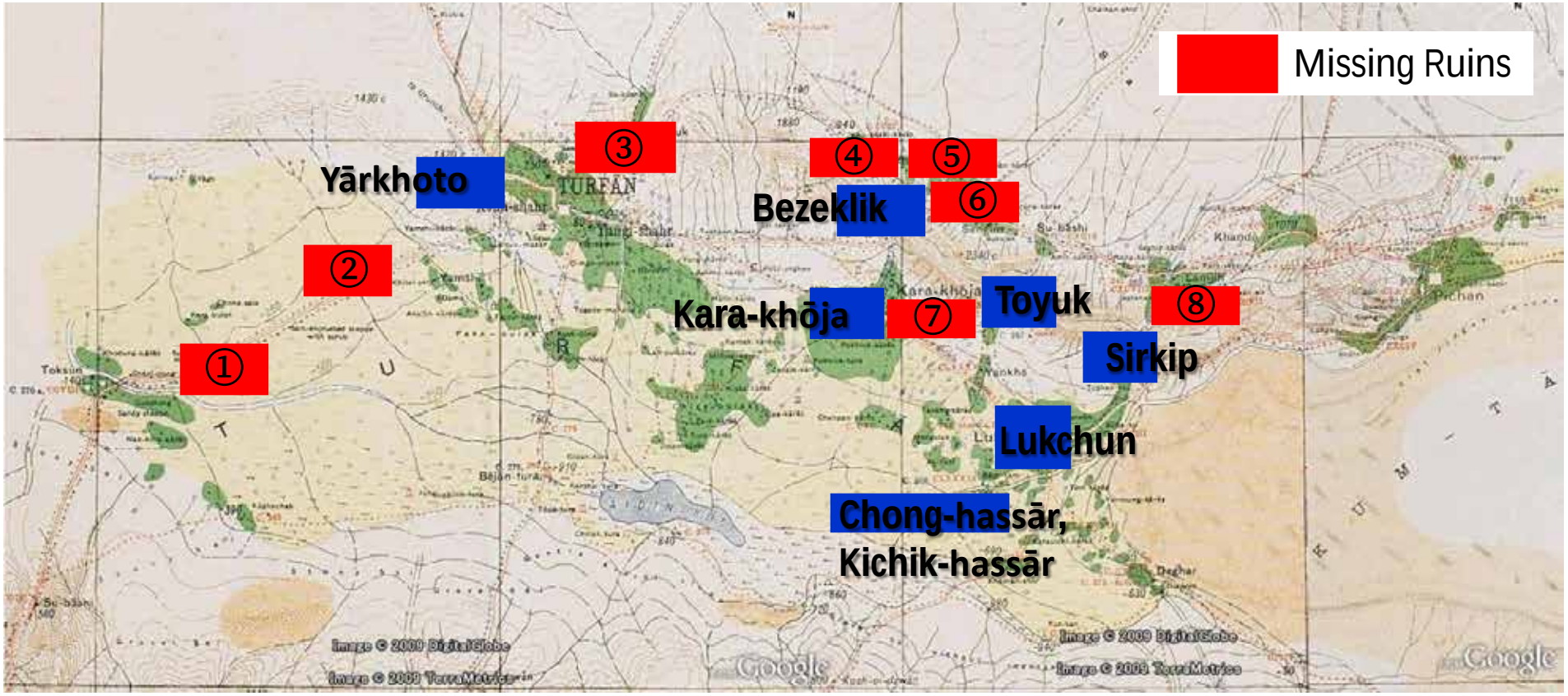
Stein Placename Database

http://dsr.nii.ac.jp/digital-maps/stein/place-names/map/



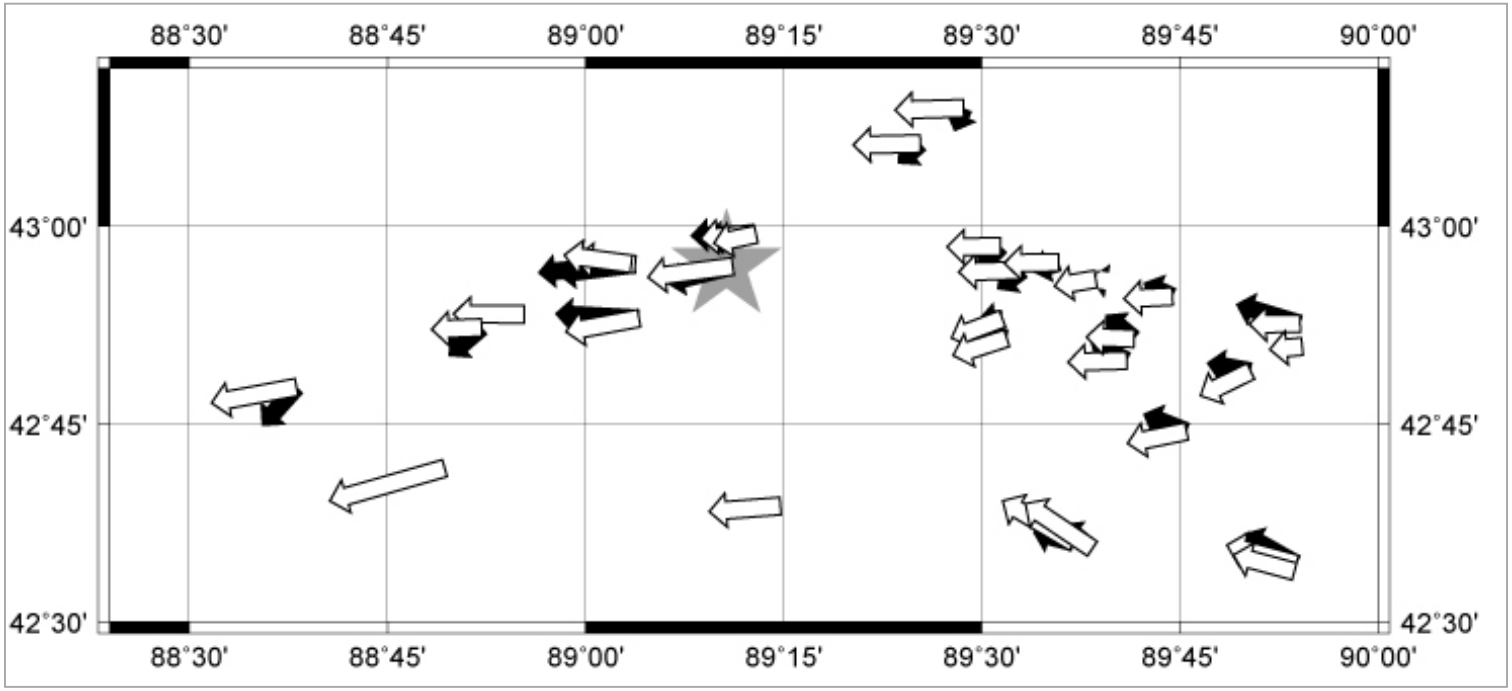
6303 place names in “Innermost Asia” maps are located on the old map, with current photographs.

Question: “Missing” Ruins?



① Oi-tam, ruined fort
② Bögan-tura
③ Buluyuk (Shipang, Sassik-bulak, Kazma)
④ Murtuk-ruins
⑤ Yoghan-tura
⑥ Chikkan-köl
⑦ Bedaulat’s town, Bēsh-kāwuk, Kosh-gumbaz
⑧ Yutōgh

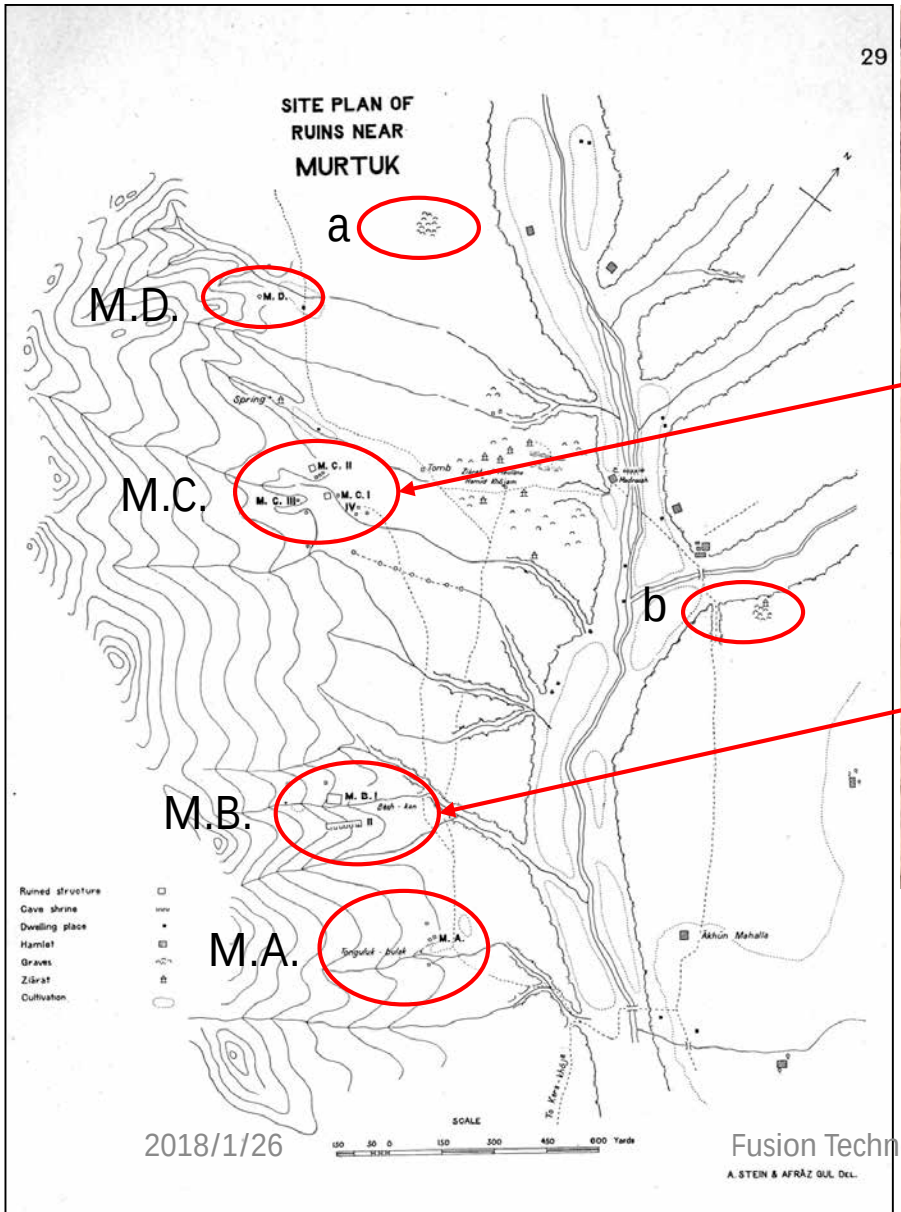
Error Distribution in Turfan

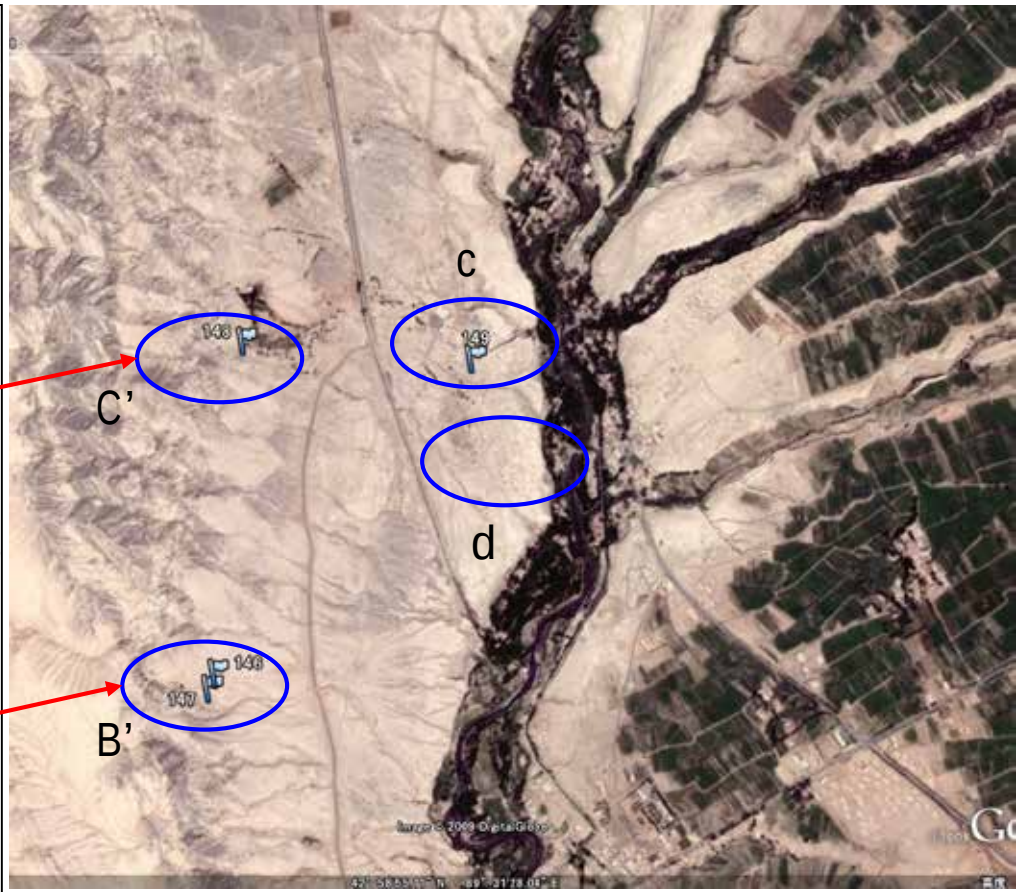


Error Distribution in Turfan Basin / White: Innermost Asia / Black: Serindia

Some ruins were reported by 20th expeditions, but are missing in recent survey reports.

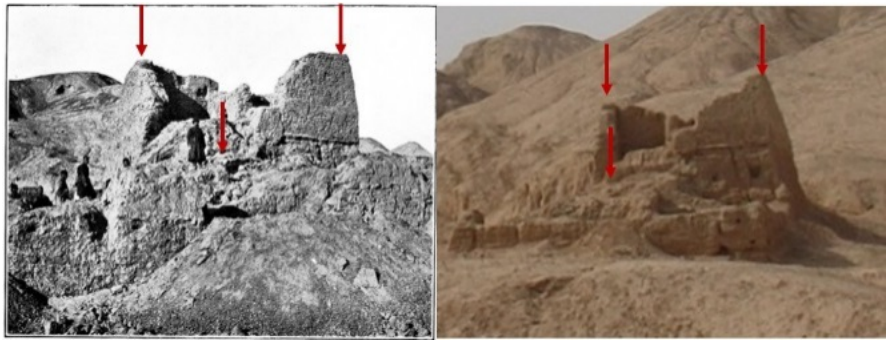
Matching Entities between Maps

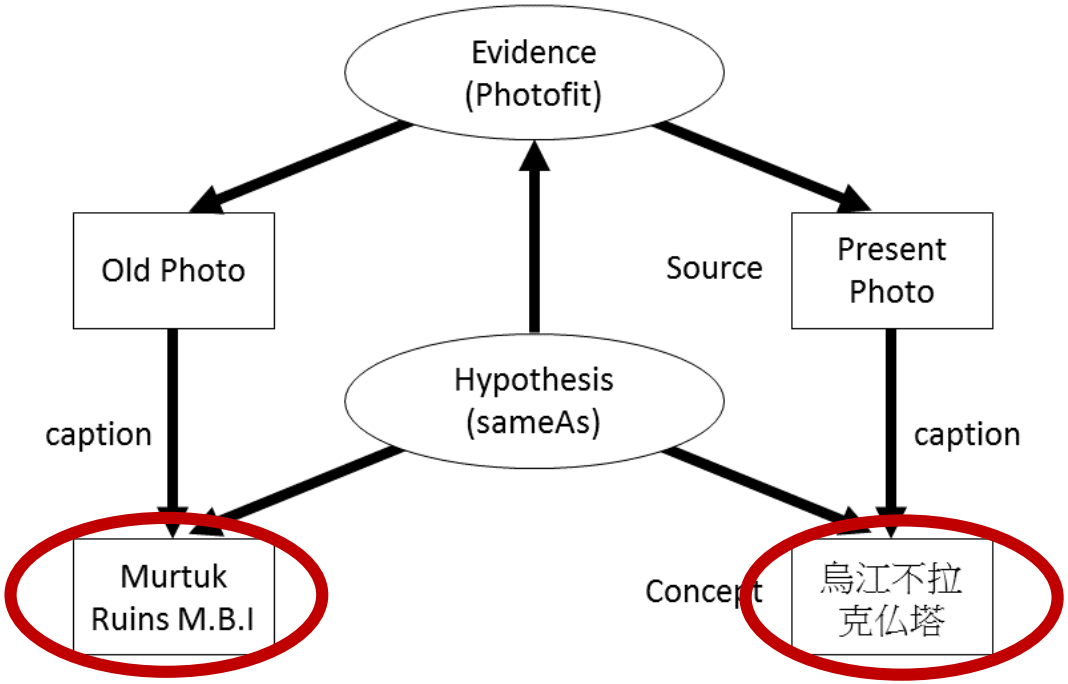




Stein's map and satellite images for the same area. Each source reports different ruins due to different conceptualization.

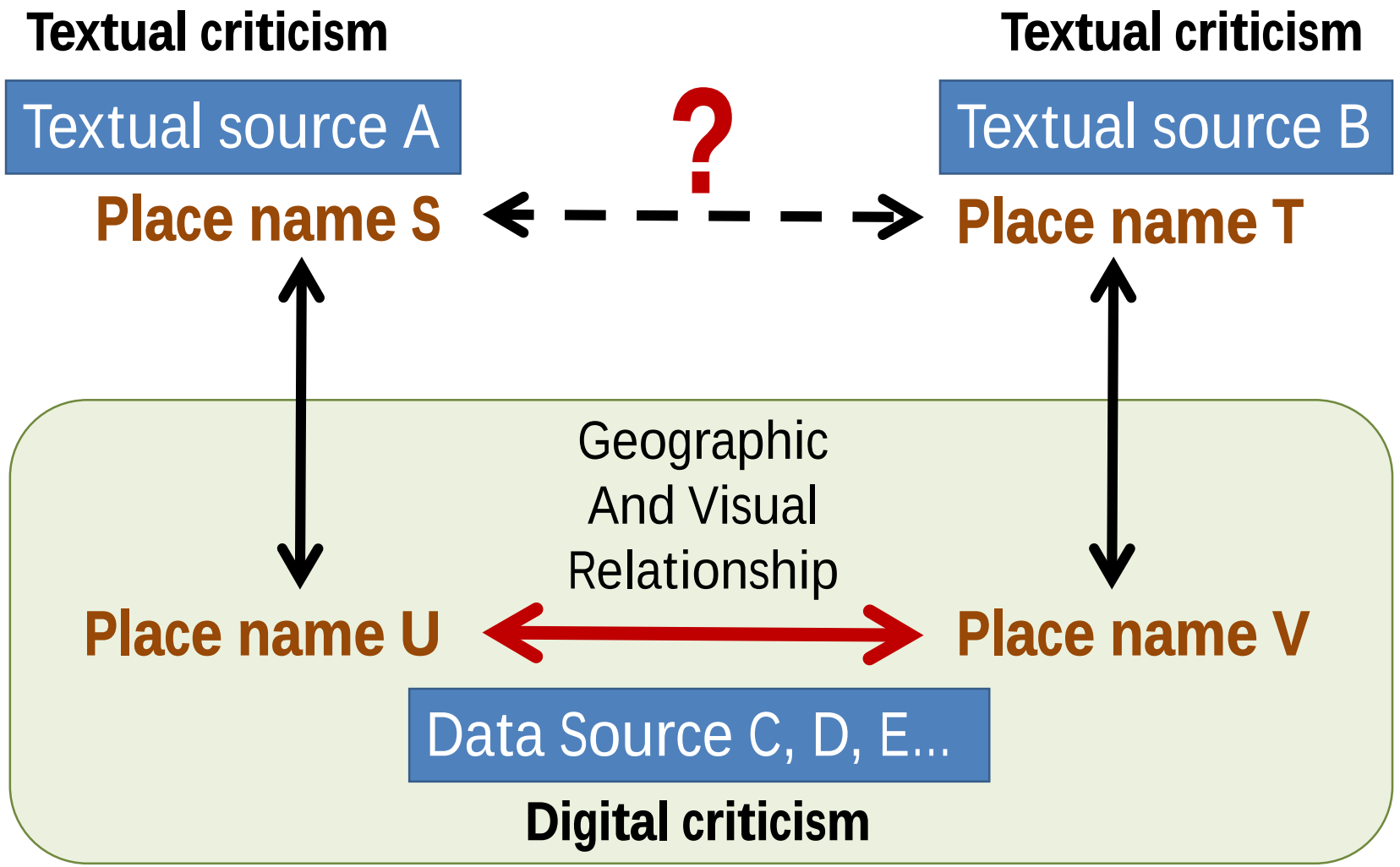
Murtuk Ruins？烏江不拉克仏塔





Question: A ruin described as “Murtuk Ruins” in an old book, and a ruin described as 「烏江不拉克仏塔」 in a new book is the same ruin or not?

Linking Entities Across Sources



Digital Humanities by “East Asia”

Digital Humanities Activities

- Japanese Association of Digital Humanities was established in **2011**.
- IPSJ SIG Computers and the Humanities was established in **1989**.
- Association for Computers and the Humanities (US) was established in **1978**.
- Association for Literary and Linguistic Computing (Europe) was established in **1973**.

Europe and North America

- Excellent collection of cultural heritage promises continued tasks for a long time.
- Technological barrier is not a problem for some humanities scholars.
- Funding from EU is strong. Support from US gov. is limited, but US has many strong foundations to support humanities research.
- Strong communities for old and young scholars to enhance the network of people.

Emerging East Asia

- **Japan in 2011 and Taiwan in 2018?** Expanding steadily across the region.
- Humanities data: already rich in China (electronic text of classics, Dunhuang cave images) and other countries.
- Community in East Asia: Not only country and region, but also technology divides the community of scholars.

Digital Humanities for “East Asia”

Historical Complexities

- **“History is written by the winners**.” There is no “unbiased” history, and we cannot write a history that satisfies all.
- Artificial intelligence (AI) cannot solve this problem!! AI simply learns from the training data given by the winners.
- We should start from sharing data, then sharing interpretations, and finally sharing better mutual understanding.

Archive of North China Railway

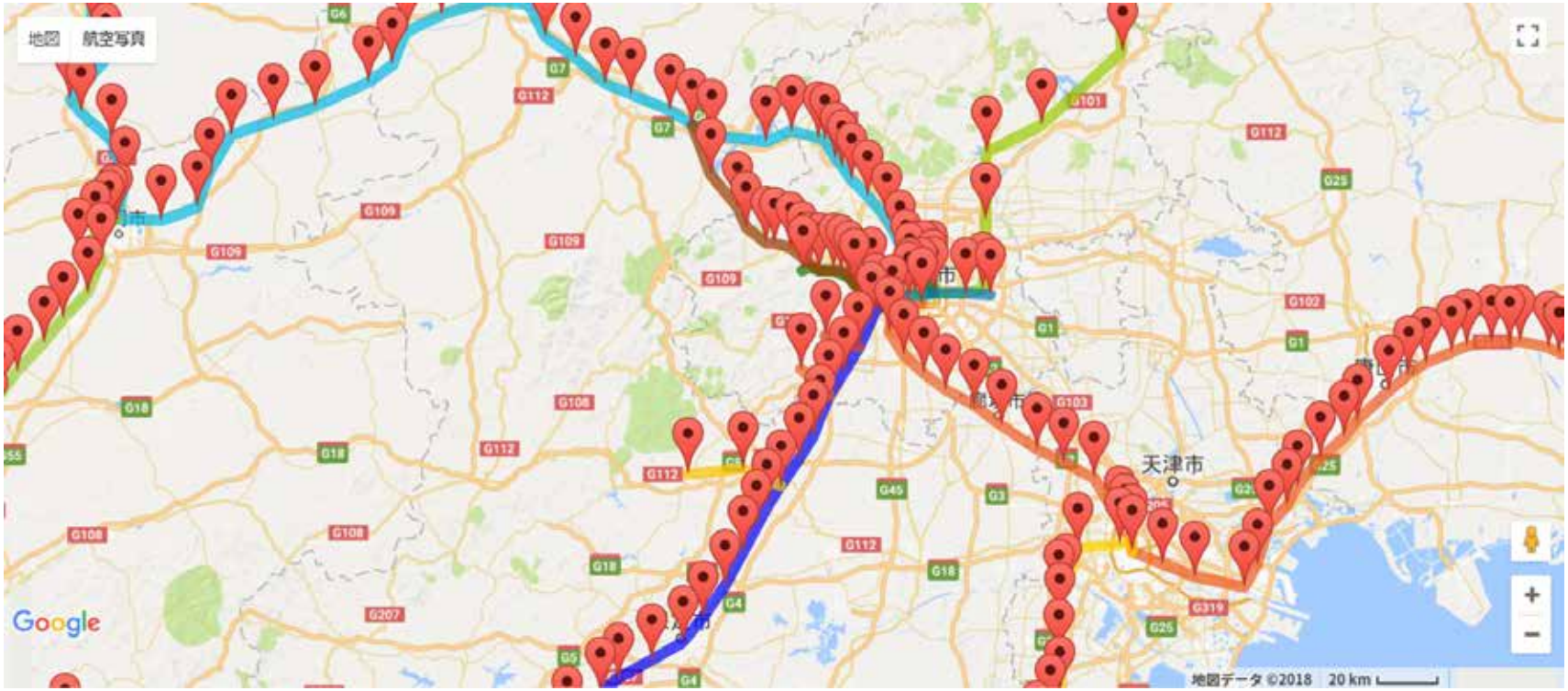
Collaborator: Institute for Research in Humanities, Kyoto University, Center for Southeast Asian Studies, Kyoto University, and Toyo University.

http://codh.rois.ac.jp/north-china-railway/

Photographs in the Wartime

- 35,000+ stock photo taken between 1937 and 1945 by **North China Railway**, a Japanese transportation company.
- The main purpose was propaganda, but it contains richer information.
- Photographs record not only the railway construction, but also people's daily lives, cultural heritage, and so on.

NCR Railway Map



NCR railway map was created to understand relationship between photographs and stations.

Bottom-Up Sharing Process

- **History** is controversial because it involves values.
- **Fact** is also controversial because there may be “alternative facts.”
- **Photographs** may be controversial because they allow multiple interpretations.
- But, multiple interpretations of photographs and data are starting points for discussion.

Toward Opening Data

- **Exhibition in a museum (Dec. 2016)** : Small part of the stock photo was exhibited.
- **Book (Nov. 2016)** : More photos and discussion could be purchased.
- **Review (Ongoing) :** Most photos are under review to avoid unnecessary controversy.
- **Release (2018?)**: All stock photos and context information will be publicly released on a new Website.

Summary

Summary

- Digital Humanities is an emerging field to "read" cultural heritage in a new way using digital tools and algorithms.
- Japanese and Chinese (Silk Road) projects were introduced to demonstrate techniques and discoveries in digital humanities.
- East Asia is a diverse and complex region, but sharing data may enhance mutual understanding between the people.

Acknowledgment and Links

- PMJT datasets and PMJT character datasets are collaboration with National Institute of Japanese Literature, especially Prof. Kazuaki Yamamoto.
- Digital Silk Road is collaboration with Toyo Bunko, and Dr. Yoko Nishimura, Toyo University.
- Center for Open Data in the Humanities
  - http://codh.rois.ac.jp/
- Digital Silk Road
  - http://dsr.nii.ac.jp/
- Kuzushi-ji Challenge
  - http://codh.rois.ac.jp/old-char-challenge/