# Recovering Historical Data from Text

leveraging LLMs for social impact

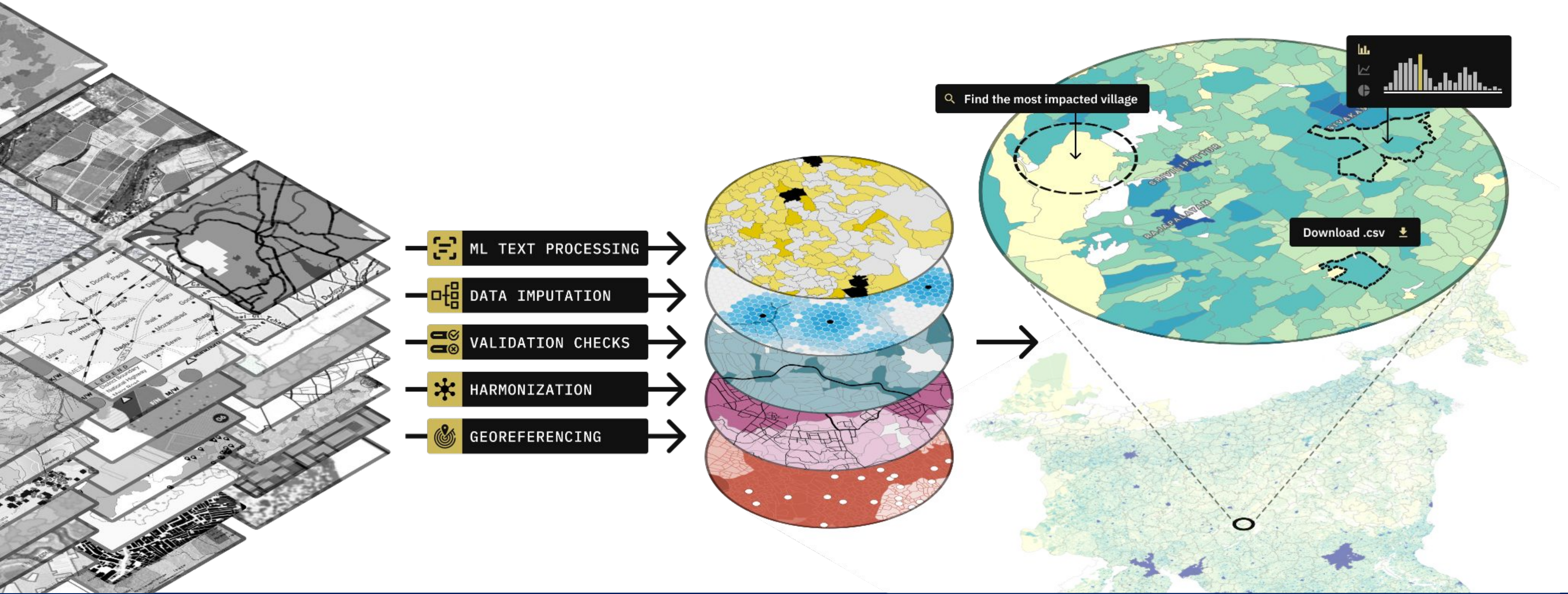
**Toby Lunt**
Anisha Grover

March 2024

Development Data Lab

# Development Data Lab

**DDL's mission** is to make India's data speak to each other, empowering policymakers, the private sector, and civil society.



ML TEXT PROCESSING

DATA IMPUTATION

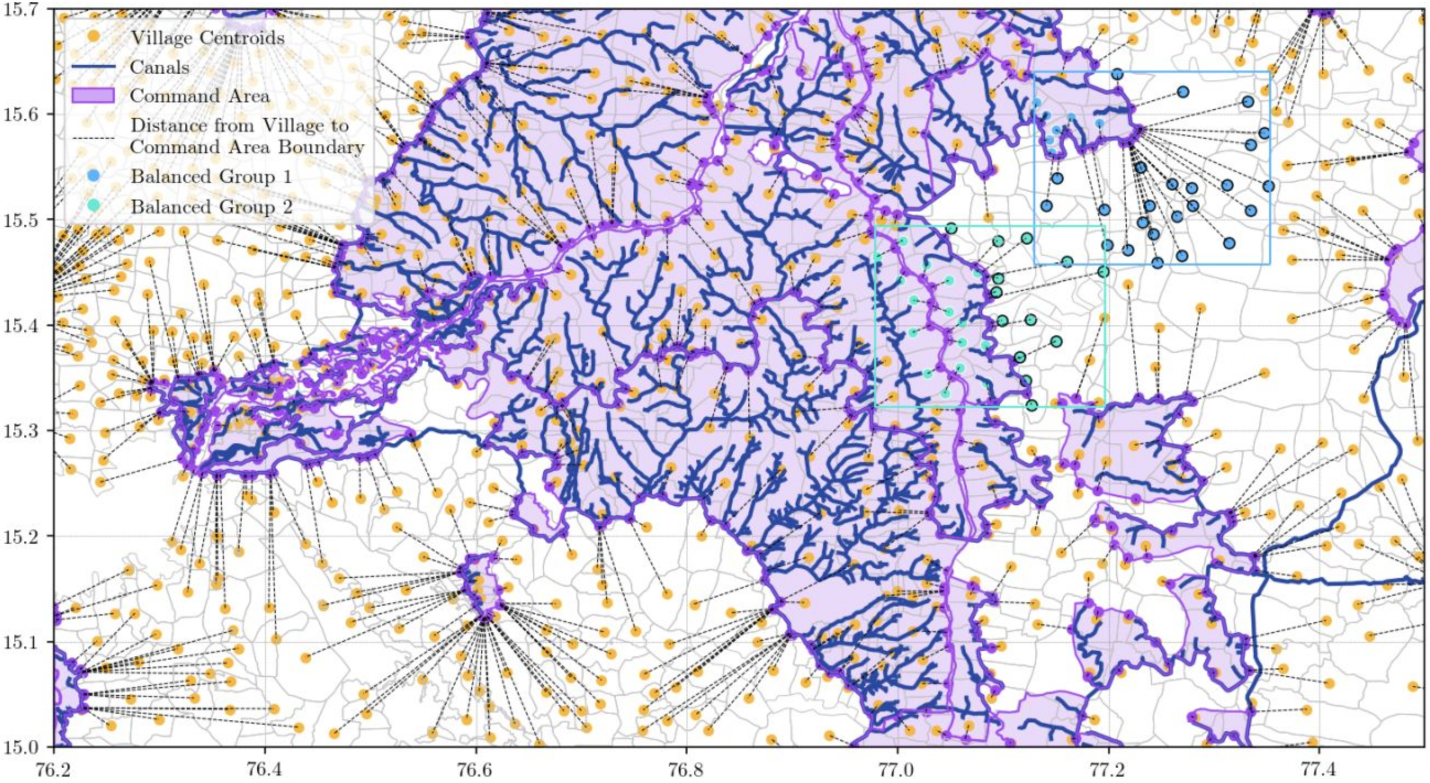
VALIDATION CHECKS

HARMONIZATION

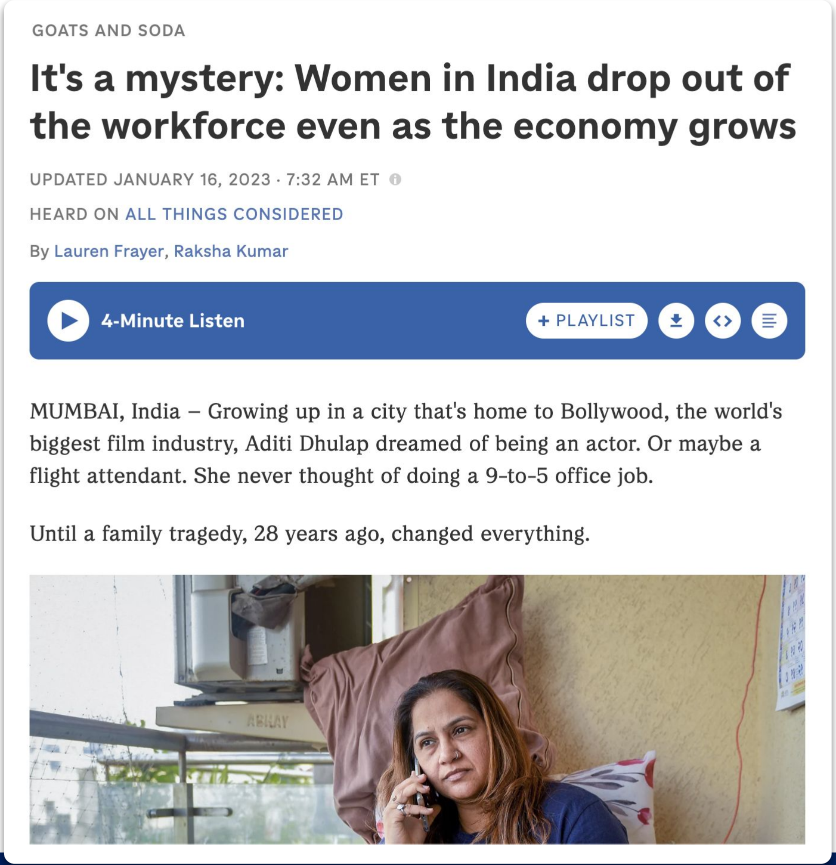
GEOREFERENCING

# Data work example

Insight depends on interoperability, even in social science



# Today: ChatGPT + 150 years of cultural norms



GOATS AND SODA

## It's a mystery: Women in India drop out of the workforce even as the economy grows

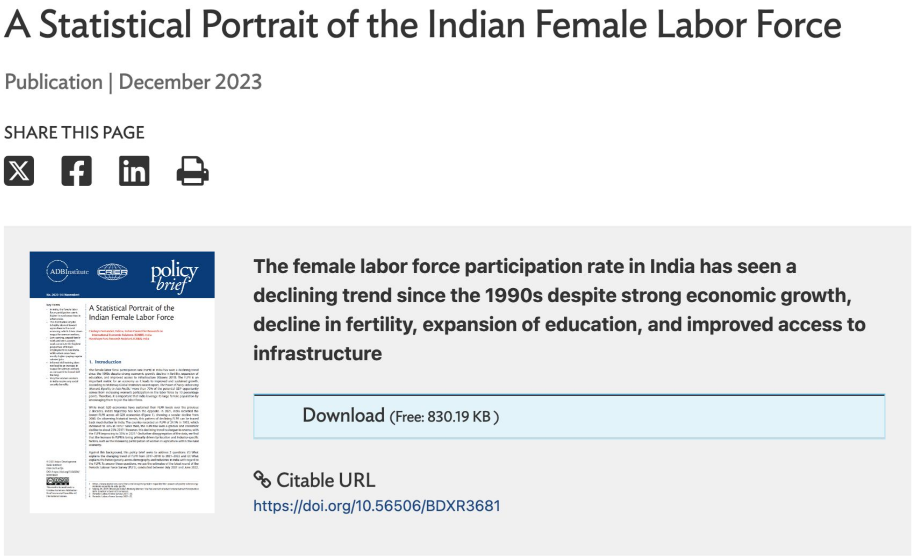
UPDATED JANUARY 16, 2023 · 7:32 AM ET

HEARD ON ALL THINGS CONSIDERED

By Lauren Frayer, Raksha Kumar

MUMBAI, India – Growing up in a city that's home to Bollywood, the world's biggest film industry, Aditi Dhulap dreamed of being an actor. Or maybe a flight attendant. She never thought of doing a 9-to-5 office job.

Until a family tragedy, 28 years ago, changed everything.



## A Statistical Portrait of the Indian Female Labor Force

Publication | December 2023

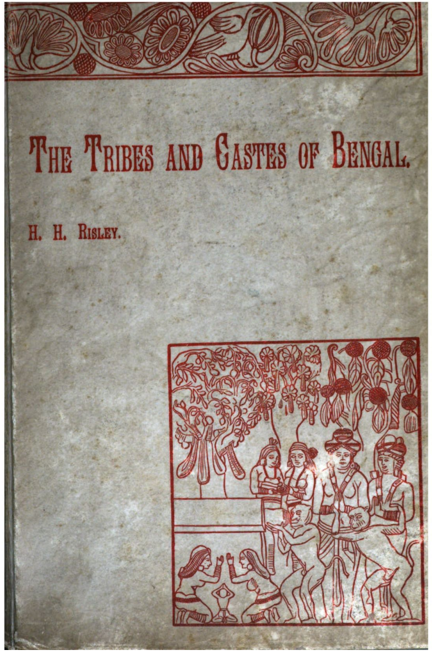
SHARE THIS PAGE

**The female labor force participation rate in India has seen a declining trend since the 1990s despite strong economic growth, decline in fertility, expansion of education, and improved access to infrastructure**

Citable URL

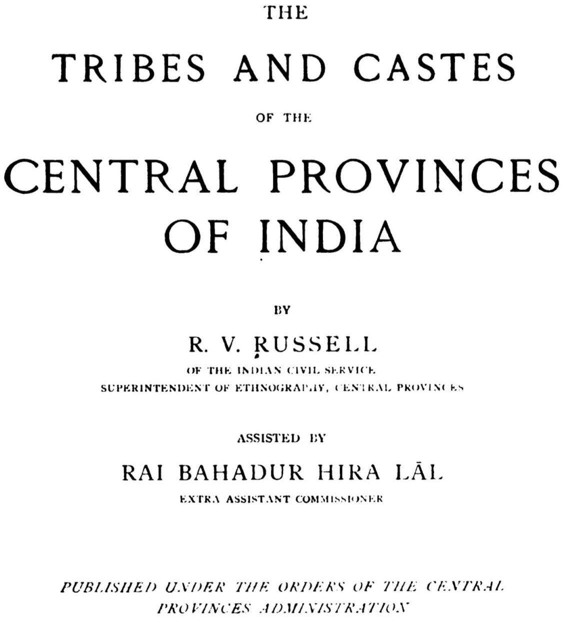
https://doi.org/10.56506/BDXR3681

# Today: ChatGPT + 150 years of cultural norms



THE TRIBES AND CASTES OF BENGAL.

H. H. RISLEY.



THE

TRIBES AND CASTES

OF THE

CENTRAL PROVINCES
OF INDIA

BY

R. V. RUSSELL

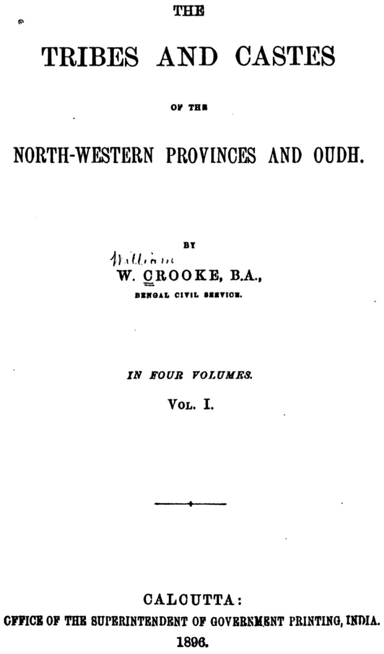
OF THE INDIAN CIVIL SERVICE
SUPERINTENDENT OF ETHNOGRAPHY, CENTRAL PROVINCES

ASSISTED BY

RAI BAHADUR HIRA LĀL

EXTRA ASSISTANT COMMISSIONER

*PUBLISHED UNDER THE ORDERS OF THE CENTRAL
PROVINCES ADMINISTRATION*



THE

TRIBES AND CASTES

OF THE

NORTH-WESTERN PROVINCES AND OUDH.

BY

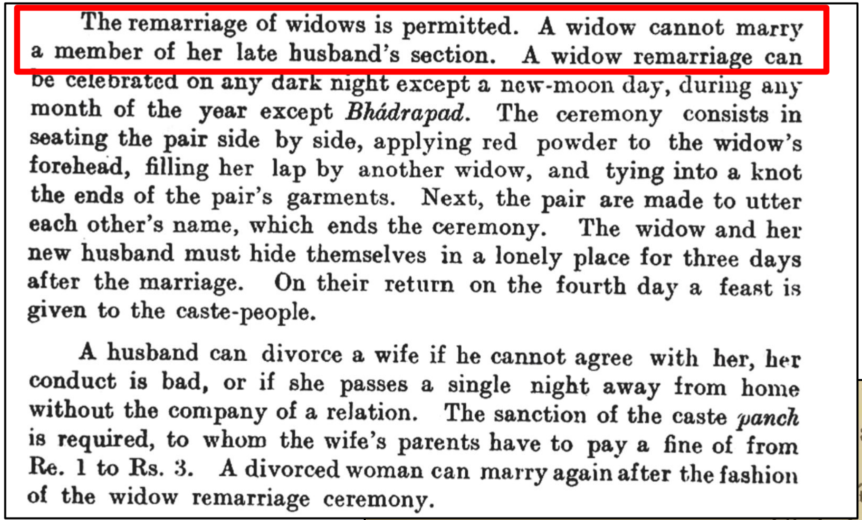
William
W. CROOKE, B.A.,

BENGAL CIVIL SERVICE.

*IN FOUR VOLUMES.*

VOL. I.

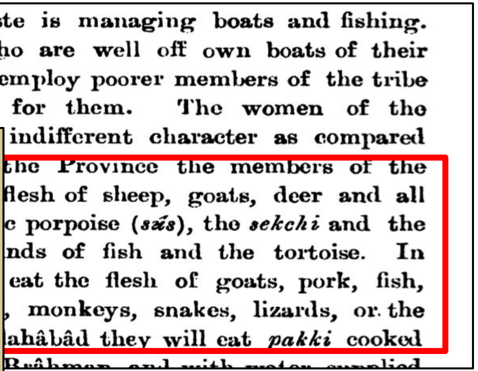
CALCUTTA:
OFFICE OF THE SUPERINTENDENT OF GOVERNMENT PRINTING, INDIA.
1896.



The remarriage of widows is permitted. A widow cannot marry a member of her late husband's section. A widow remarriage can be celebrated on any dark night except a new-moon day, during any month of the year except *Bhádrapad*. The ceremony consists in seating the pair side by side, applying red powder to the widow's forehead, filling her lap by another widow, and tying into a knot the ends of the pair's garments. Next, the pair are made to utter each other's name, which ends the ceremony. The widow and her new husband must hide themselves in a lonely place for three days after the marriage. On their return on the fourth day a feast is given to the caste-people.

A husband can divorce a wife if he cannot agree with her, her conduct is bad, or if she passes a single night away from home without the company of a relation. The sanction of the caste *panch* is required, to whom the wife's parents have to pay a fine of from Re. 1 to Rs. 3. A divorced woman can marry again after the fashion of the widow remarriage ceremony.

**Vaidu Community**
Enthoven 1922 Bombay
Volume 3



8. The business of the caste is managing boats and fishing. Those who are well off own boats of their own and employ poorer members of the tribe to work for them. The women of the

Occupation and social status.

indifferent character as compared

the Province the members of the

flesh of sheep, goats, deer and all

c porpoise (*sús*), the *sekchi* and the

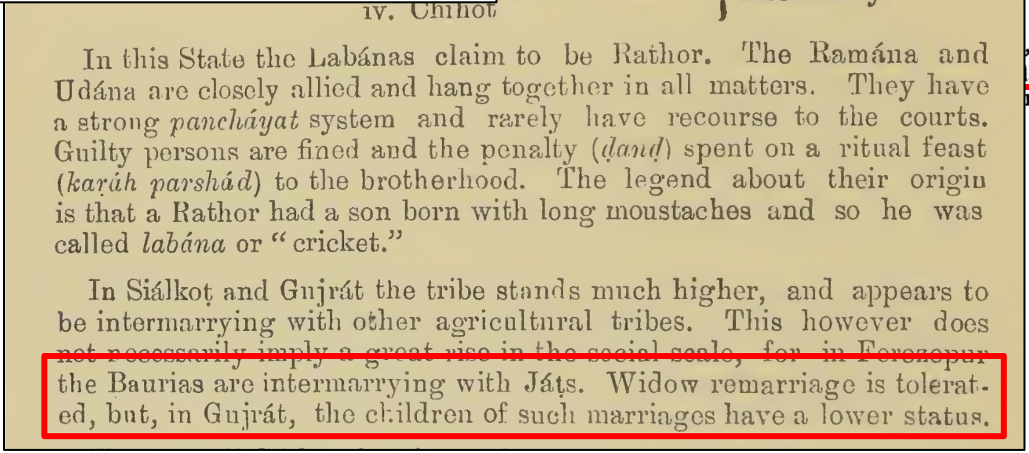
nds of fish and the tortoise. In

eat the flesh of goats, pork, fish,

, monkeys, snakes, lizards, or the

ahâbâd they will eat *pakki* cooked

**Mallah Community**
Crooke 1896 North West
Province Volume 3



thus grouped :—

a

} do not intermarry.

́wat or Gharnot } intermarry.

iv. Chihot

In this State the Labánas claim to be Rathor. The Ramána and Udána are closely allied and hang together in all matters. They have a strong *pancháyat* system and rarely have recourse to the courts. Guilty persons are fined and the penalty (*dand*) spent on a ritual feast (*karáh parshád*) to the brotherhood. The legend about their origin is that a Rathor had a son born with long moustaches and so he was called *labána* or "cricket."

In Siálkoṭ and Gujrát the tribe stands much higher, and appears to be intermarrying with other agricultural tribes. This however does not necessarily imply a great rise in the social scale, for in Ferozepur the Baurias are intermarrying with Jáṭs. Widow remarriage is tolerated, but, in Gujrát, the children of such marriages have a lower status.

**Labanas Community**
Rose 1919 Punjab
Volume 3

# Basic overview of the work

OCR ➡️ Chunking ➡️ Norms queries

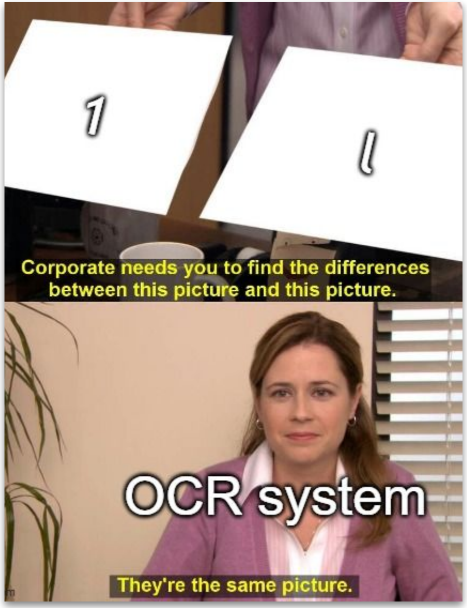
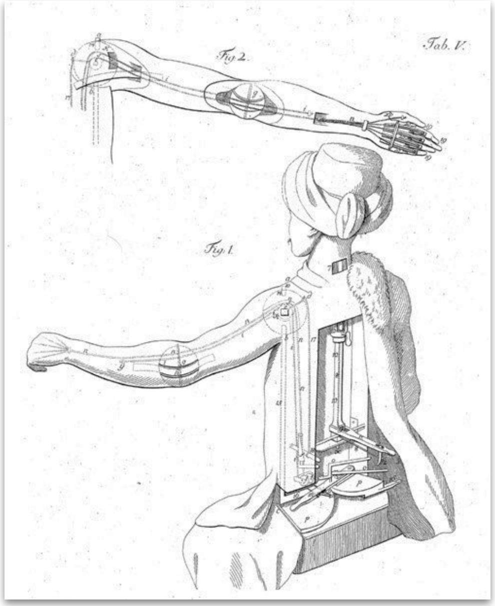
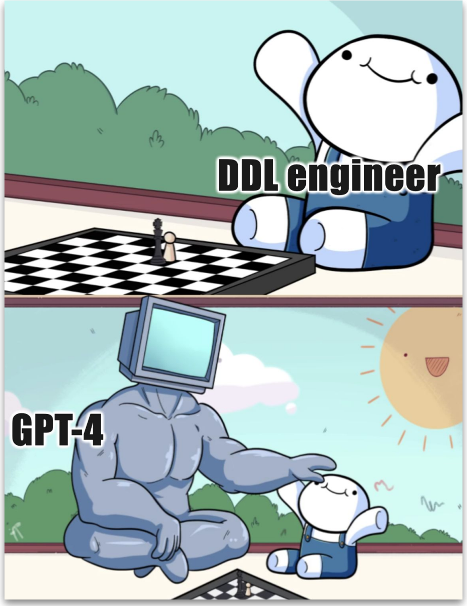

1
l
Corporate needs you to find the differences between this picture and this picture.
OCR system
They're the same picture.

Image credit:
https://medium.com/poatek/understanding-ocr-your-entertaining-guide-to-optical-character-recognition-5891d31159c6





They eat the flesh of goats, hogs, fowls, rats, and fish. They drink toddy to excess. They eat at the hands of all castes except Mahars, Chamars, Lingayats, Jains and Halleers. No caste eats with them. They rank below the cultivating classes and above the impure castes.

**<<<<<AGHORI**

AGHORI.-A sect of Shiva worshippers.

**<<<<<AGLE**

AGLE.-A synonym for Agri.

**<<<<<AGRIS**

AGRIS, also known as Agle and Kherpdtii, numbered 211,176 at the Census of 1901. They are principally found in Théng (83,733), Koléba (113,115) and the State of Janjira (9,617).

There are no exogamous sub-divisions above families having the same surname and observing common mourning. The following is a list of such families, kuls or gotras :-

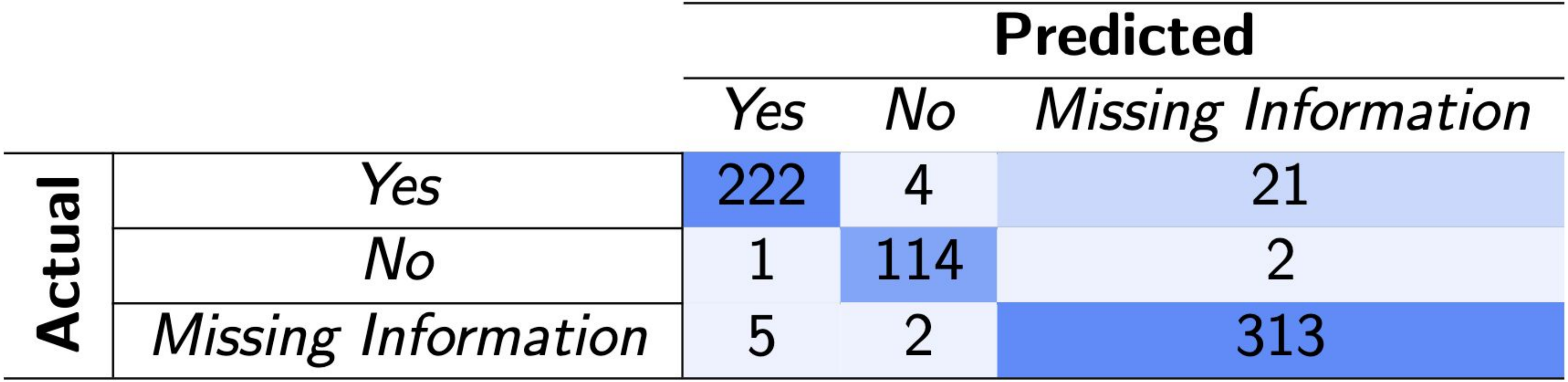
(1) Bhoir. (29) Joshi.
(2) Chaudhari. (30) Mobhile.
(3) Chavan, (31) Kharik.

```
{
  "name": "norm_response",
  "description": "Function that combines the short answer to user query and the quoted lines of text providing evidence for the short answer into a paragraph.",
  "parameters": {
    "type": "object",
    "properties": {
      "short_answer": {
        "type": "string",
        "enum": [
          "yes",
          "no",
          "no information"
        ],
        "description": "The answer to the question inferred from the text provided in the prompt."
      },
      "text_quote": {
        "name": "string",
        "type": "string",
        "description": "Lines quoted from the corpus that were used to frame the answer."
      }
    },
    "required": [
      "short_answer",
      "text_quote"
    ]
  }
}
```

```
{
  "id": "chatcmpl-91ulDQKG2IhmLR09Ui6UHSph7folF",
  "object": "chat.completion",
  "created": 1710243543,
  "model": "gpt-4-0125-preview",
  "choices": [
    {
      "index": 0,
      "message": {
        "role": "assistant",
        "content": null,
        "function_call": {
          "name": "norm_response",
          "arguments": "{\"short_answer\":\"yes\",\"text_quote\":\"Children are betrothed at any age, and
married between eight and fifteen. Every year on one fixed day Ahir marriages take place.\\n\\nIn the case of
girls, both infant and adult marriages are in vogue. The adult marriage is mostly prevalent among the poorer
classes, and those who can afford it marry their daughters as infants at the age of nine or ten.\"}"
        }
      },
      "logprobs": null,
      "finish_reason": "stop"
    }
  ],
  "usage": {
    "prompt_tokens": 12060,
    "completion_tokens": 86,
    "total_tokens": 12146
  },
  "system_fingerprint": "fp_c121a3f431"
}
```

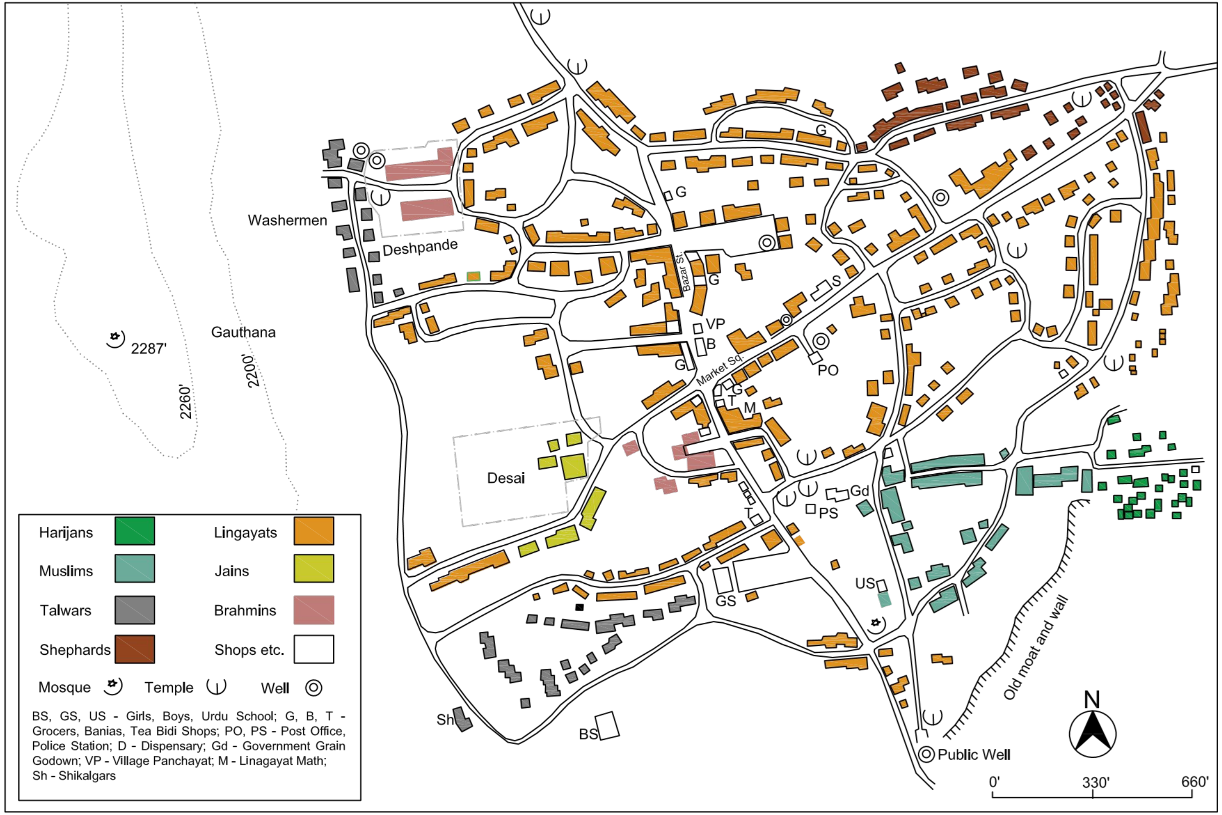
# Early results

Table: Confusion Matrix for GPT's Response vs. Actual Norm



| | | **Predicted** | | |
|---|---|---|---|---|
| | | *Yes* | *No* | *Missing Information* |
| **Actual** | *Yes* | 222 | 4 | 21 |
| | *No* | 1 | 114 | 2 |
| | *Missing Information* | 5 | 2 | 313 |

# More results?


Washermen
Deshpande
Gauthana
2287'
2260'
2200'
Desai
Bazar St.
Market Sq.
VP
B
G
S
T
M
PO
PS
Gd
US
GS
BS
Sh
Public Well
Old moat and wall
N
0'
330'
660'
Harijans
Muslims
Talwars
Shephards
Lingayats
Jains
Brahmins
Shops etc.
Mosque
Temple
Well
BS, GS, US - Girls, Boys, Urdu School; G, B, T - Grocers, Banias, Tea Bidi Shops; PO, PS - Post Office, Police Station; D - Dispensary; Gd - Government Grain Godown; VP - Village Panchayat; M - Linagayat Math; Sh - Shikalgars

**Source:** Spate and Learmonth, 1954 via Bharathi et al, 2020

# Next steps: validation

## Manual spot checks

Scale library of tests that evaluate specific types of error at very high confidence

Develop basic infrastructure to facilitate easy comparisons across runs

## Response evaluation

Internal consistency: do the excerpts agree with the binary responses?

Validation of response evaluation LLM with separate test set

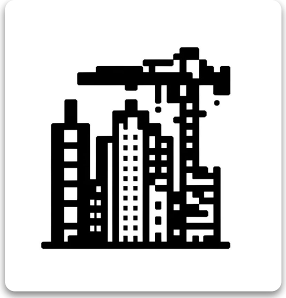
## Ambiguous cases

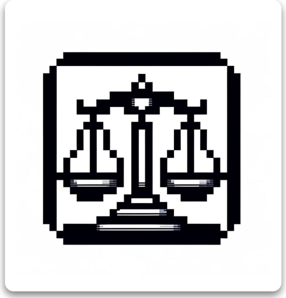
Develop confidence scoring measure to flag instances requiring manual review

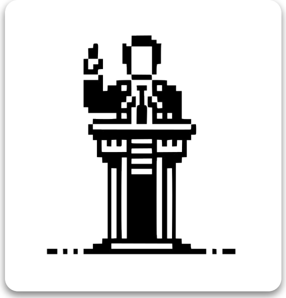
# Where we go from here

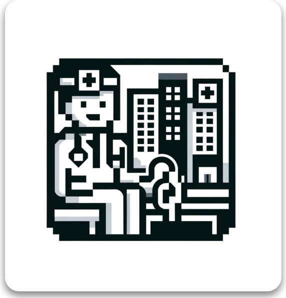
**We think these techniques will generalize well** to additional datasets that will help us answer important but historically intractable questions:

- How does land regulation affect urban development?
- Bias in the indian judiciary
- Political capture of the judicial system
- Integrating private sector data to create a real-time economic tracker
- Medical records - respiratory illness against surface mines and power generation







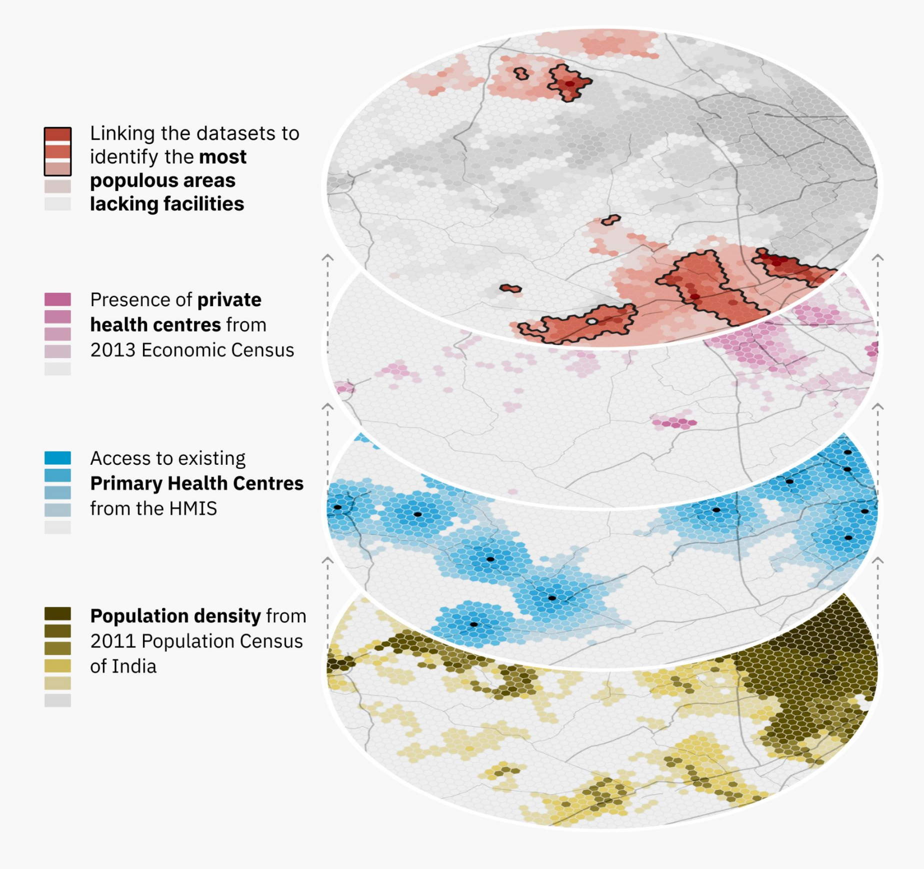


Development Data Lab

# Thank you!

**Tobias Lunt**

lunt@devdatalab.org



Linking the datasets to identify the **most populous areas lacking facilities**

Presence of **private health centres** from 2013 Economic Census

Access to existing **Primary Health Centres** from the HMIS

**Population density** from 2011 Population Census of India

# Appendix slides

```
# GPT API call
enthoven_vol_1_ahirs = openai.ChatCompletion.create(
    model=GPT_MODEL,
    messages=[
        {"role": "system", "content": system_message}, # general
        {"role": "user", "content": input_message},    # specific
    ],
    max_tokens=max_response_tokens,
    temperature=model_temperature,
    functions=input_function,
    function_call={
        "name": "norm_response" # function calling
    },
)
```

# Challenges

- Domain-specific
    - Interpreting the text! Lots of ambiguity and confusion the writing
    - Subgroups with conflicting norms
    - Subgroups exist in a complex, hierarchical network that differs across books
    - The same norm can be described in multiple ways by different authors
    - Many subgroups don’t have text that actually describes norms - exclusion criterion
    - Interpreting absence of a norm for a group - ’missing information’ or a negative answer?

- Domain-agnostic
    - Evaluating accuracy
    - Greediness of relevant excerpts returned by GPT
    - To RAG or not to RAG?
    - Nondeterministic model output

# Where we go from here

There are many examples of unstructured text that are already being processed with less sophisticated methods (such as deep neural networks) in current economics research, from judicial decisions to political speeches and corporate filings (see Ash and Hansen, 2023, for a review of these methods and applications).

Land markets, zoning, and urban development

- landlord tenant disputes
- eminent domain / takings / acquisitions
- SFI (square foot index)
- land ceiling acts cases

We will be developing three studies investigating a) the role of criminal politicians in influencing judicial outcomes in India, b) the relationship between land governance and urban growth, and c) extending prior work on in-group bias in India's lower judiciary by mining the text of judicial decisions of 80 million criminal and civil cases filed between 2010-18.

- The first component of our study will examine patterns in sentencing, undertrial incarceration, and bail decisions, using data parsed from the text of judicial decisions.
- Second, we will examine the relationship between the frequency and pace of resolution of land disputes and urban development. Land regulation in Indian cities is famously restrictive, and land disputes that drag on for years and years are thought to be a major hindrance to development. This in turn prevents firms from growing, and prevents new housing from being built, making it challenging for rural people to find opportunities in the more dynamic cities. In addition to creating new data describing the legal barriers to land development in Indian cities, we will match case outcomes directly to individual properties to study whether land development is more likely to take place when cases are resolved. By calculating the extent of hindered land development, we can measure one of the key economic costs of India’s slow judiciary.
- The third component of our study will examine whether influential litigants get treated impartially under the law. Specifically, we will identify cases involving politicians—MLAs and MPs—who have been charged under India’s criminal codes. Over a quarter of India’s politicians have open criminal charges; how those charges  are treated under the law has not been previously studied.