Russian Security State
GOVT-5519/IPOL-3519/REES-5519/SEST-6763
Lecture 02. Backgrounder.
Collecting Data on Russia

Yuri M. Zhukov
Associate Professor
Georgetown University

September 9, 2025

Today's objectives

1. *Clarify* core concepts and uses of data
2. *Illuminate* persistent challenges in Russian data collection
3. *Equip* you with tools to find and handle open-source data

What are data?

Data (plural): organized collections of observations or measurements (e.g., official government statistics, crowd-sourced battlefield reports, social media posts, photo albums, public opinion polls, maps, scores)

We use data to answer questions and inform decision-making.

Examples of data applications:

1. Science (test hypotheses, predictive modeling, experiments, surveys)
2. Military (tracking enemy movements, battle damage assessments)
3. Intelligence (analyzing imagery, intercepting communications)
4. Law Enforcement (documenting crimes, arrests, prosecutions)
5. Human Rights (investigating abuses and rights violations)
6. Medicine (clinical trials, tracking patient vital signs)
7. Public Health (tracking epidemics, mental health impact studies)
8. Industry (consumer behavior analysis, forecasting, market research)
9. Sports (recruitment, performance analytics, broadcasting)
10. Entertainment (content creation, audience analytics, anti-piracy)

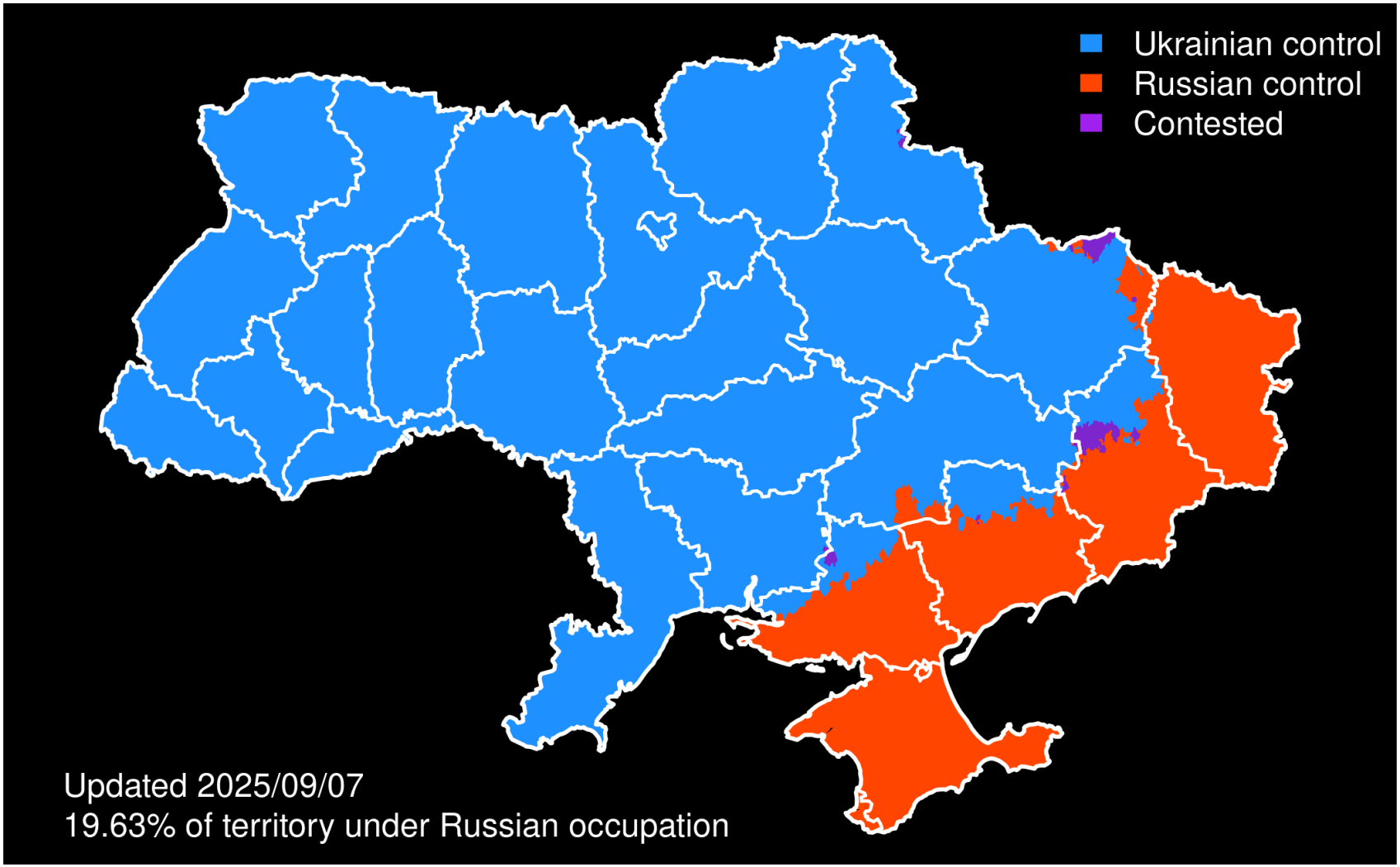

Figure 1: Territorial control in Ukraine (today)

	geonameid	date	status_wiki	status_boost	status_dsm	status_isw	status
	<num>	<int>	<char>	<char>	<char>	<char>	<char>
1:	461727	20221231	RU	RU	RU	RU	RU
2:	467852	20221231	RU	RU	RU	RU	RU
3:	468196	20221231	RU	RU	RU	RU	RU
4:	477085	20221231	RU	RU	RU	RU	RU
5:	485524	20221231	RU	RU	RU	RU	RU

33137:	12434460	20221231	UA	UA	UA	UA	UA
33138:	12434461	20221231	UA	UA	UA	UA	UA
33139:	12434462	20221231	UA	UA	UA	UA	UA
33140:	12434463	20221231	UA	UA	UA	UA	UA
33141:	12435968	20221231	UA	UA	UA	UA	UA

>

Figure 2: Territorial control data extract for New Year's Eve, 2022

This table (and the preceding map) are from VIINA, a near-real time multi-source event data system tracking the Russian-Ukrainian War. Available here: `https://github.com/zhukovyuri/VIINA`

This table is a daily extract from a **panel dataset**, where the same towns and villages (indexed by `geonameid`) are observed at multiple time points (`date`), enabling analysis of temporal dynamics and spatial differences.

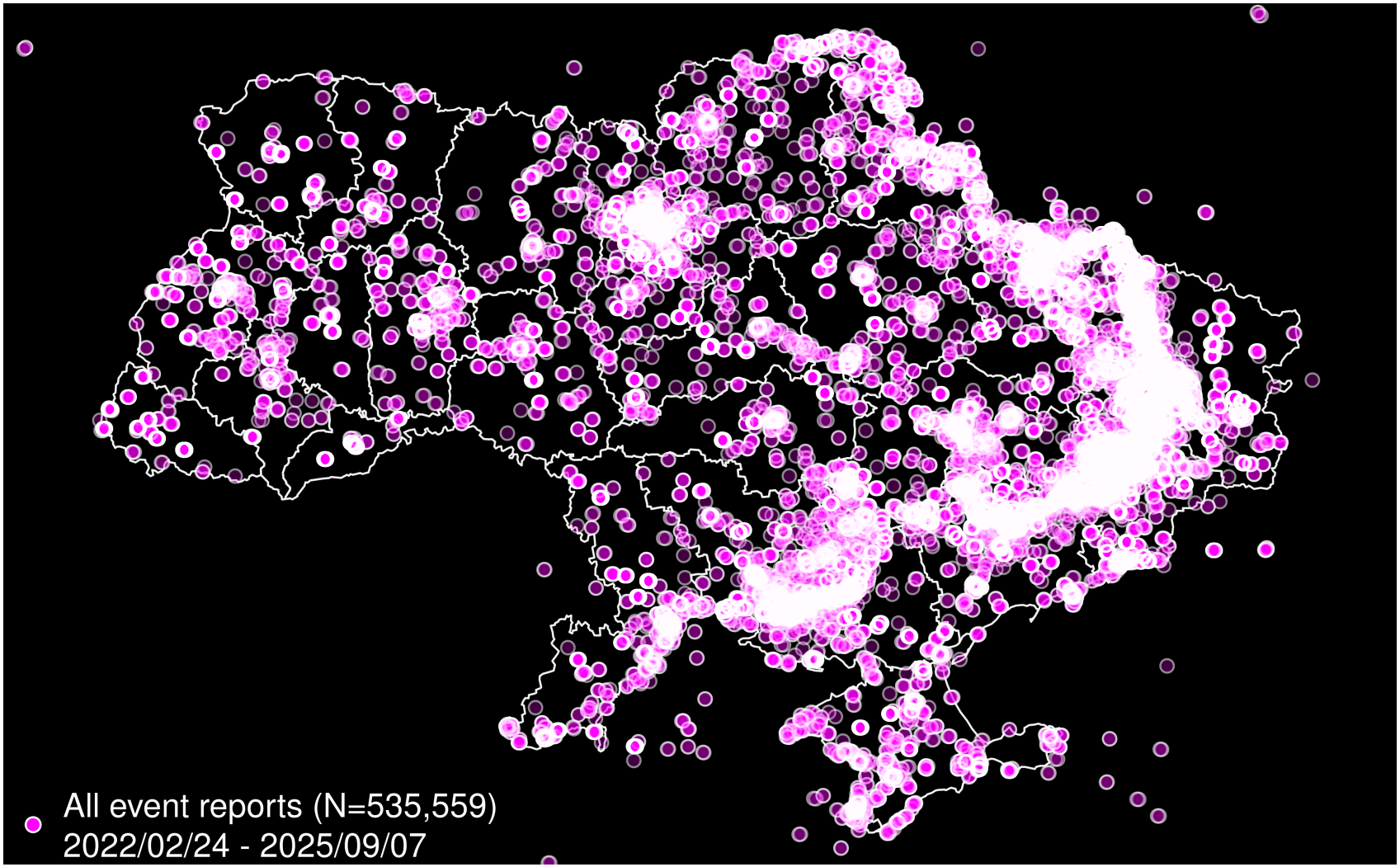

Figure 3: War-related events in Ukraine (2/24/2022 – today)

	event_id	date	longitude	latitude	source	text_10
	<int>	<char>	<num>	<num>	<char>	<char>
1:	44494	20221231	30.73169	46.48421	interfaxua	В Одесі на...
2:	44495	20221231	34.12376	51.10624	interfaxua	Ворог у п'...
3:	44496	20221231	34.09941	44.95363	interfaxua	У Криму ро...
4:	44497	20221231	32.61496	46.64032	interfaxua	Завершено ...
5:	44498	20221231	37.67529	47.95508	interfaxua	Окупанти в...

693:	1788710	20221231	37.24820	47.77905	liveuamap	На Новопав...
694:	1788726	20221231	37.90972	49.47945	liveuamap	Українські...
695:	1788733	20221231	31.29865	51.49101	liveuamap	Російська ...
696:	1788735	20221231	33.52291	44.60105	liveuamap	У Севастоп...
697:	1788736	20221231	30.49585	50.38876	liveuamap	Безпілотни...

Figure 4: Event data extract for New Year's Eve, 2022

This extract is from VIINA's **event dataset**, where the each row is the location, timing, attributes of a single incident, with source info.

VIINA is an example of an open-source data project.

Open-source data: information that is unclassified and non-proprietary, accessible through public channels without special permissions/clearances

Examples:

- Government publications (official statistics, administrative records, legislative documents, press briefings, vote counts, court cases)
- Media and journalism (news reports, investigative pieces, editorials)
- Social media content (posts, comments, user-generated content)
- Commercial information (stock prices, revenues, contracts)
- Geospatial data (satellite imagery, maps, location-based info)
- Leaked materials: (whistleblower disclosures, document dumps)

Raw data: original, unprocessed information (e.g., images, webpages, books, transcripts), requiring some cleaning or transformation before use.

Processed data: raw information after it has been cleaned, organized and stored for efficient retrieval, interpretation, and analysis.

Storage options for processed data:

1. Delimited text (`csv`, `json`, `xml`) (simple, portable text files; easy for basic storage and transfer; can open/edit them in Excel/GoogleDocs)
2. Relational databases (structured tables stored in systems like MySQL or PostgreSQL; support complex queries and relationships)
3. Cloud storage (stored offsite on platforms like AWS S3 or Google Cloud; scalable and accessible from anywhere)
4. Object storage (data stored as discrete objects with metadata in a flat namespace; optimal for unstructured, large-scale datasets)
5. NoSQL databases (flexible, schema-less storage, like document, key-value, or graph DBs, designed for big, rapidly evolving data)

We will be working with **delimited text** files only in this class.

Data on Russia

The 1937 All-Soviet Census

- First population census since launch of Stalin's 5-year plans, collectivization (famines, purges)
- Counted 162 million people, 18M below official projections
- Results were never published
- NKVD arrested and executed Census directors and statisticians (and next 3 heads of Central Statistical Administration)
- New census in 1939! Now with "corrected" (inflated) numbers.

Discussion

How might censored or falsified data affect government decision-making and public policy?

Figure 5: Deceitful numbers

Data use in Stalin's USSR

1. *Tracking agricultural output*
   a) Examples: crop yield, livestock counts, grain production
   b) Uses: plan collectivization, allocate resources, set quotas
2. *Secret police databases*
   a) Examples: arrests, surveillance, denunciations, purges
   b) Uses: population control, identify "enemies", set quotas
3. *Military mobilization and planning*
   a) Examples: population, conscription, procurement, casualties
   b) Uses: organize armed forces, plan operations, manage logistics
4. *Economic central planning*
   a) Examples: production stats, labor force, input-output tables
   b) Uses: distribute resources, set development targets, quotas
5. *Demographic surveillance and social control*
   a) Examples: population size, birth and death rates, residence
   b) Uses: monitor migration, compliance with internal passports

Why collecting data on Russia is hard

How the Kremlin keeps its secrets

1. Over-classification of records
2. Falsification/manipulation of data
3. Censorship (official and self)
4. Restricted access to documents
5. Punishment of whistleblowers
6. Funding restrictions on media, survey firms (e.g. foreign agent laws)
7. Blocking/surveillance of communications
8. Re-classification of archival materials

Figure 6: Don't chatter

How to find open-source information on Russia

Data type 1: **Government statistics**

- National accounts, census data, labor statistics (e.g., GDP, pop. density, unemployment rates)
- *Common processing steps:* aggregating or matching raw data to geographic units and time periods; tabulation, validation, statistical adjustments

Examples:

1. *Rosstat (State Statistics Service)*: official demographic, economic stats (`eng.rosstat.gov.ru`)
2. *Central Electoral Commission*: elections, candidates (`cikrf.ru`)
3. *Demoscope Weekly*: demographic indicators from Soviet period (`demoscope.ru`)

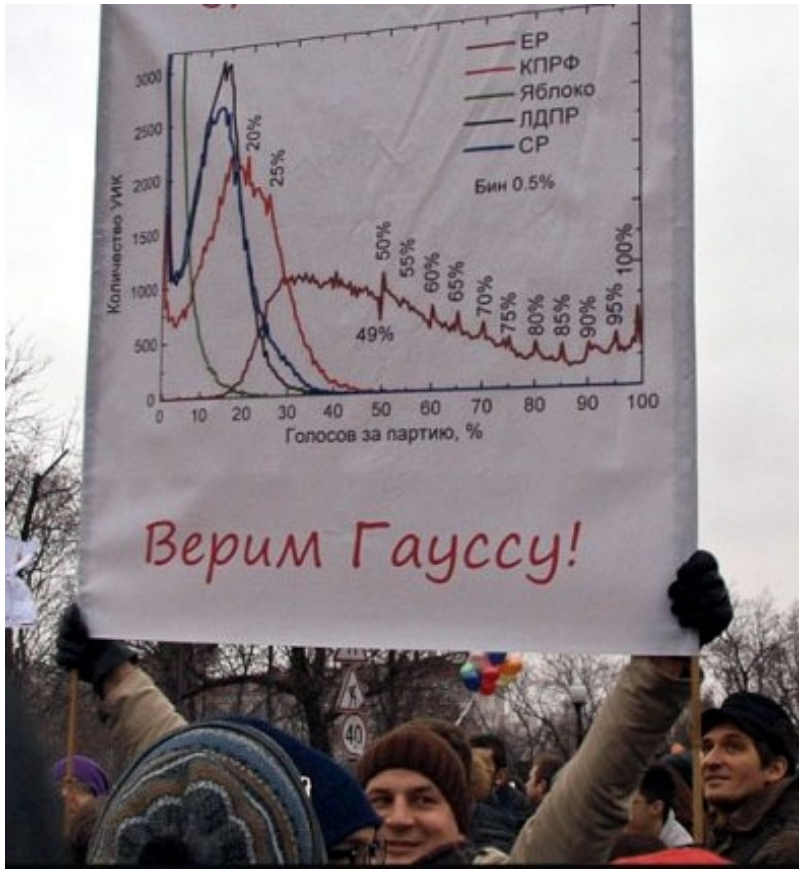

Figure 7: Not a bell curve

Data type 2: **Administrative records**

- Transaction or status records at individual or case level (e.g., personnel files, court cases, tax filings, arrest records, passports)
- *Common processing steps:* record linkage across source systems, de-duplication, anonymization of personally identifiable info

Examples:

1. *Pamyat Naroda (Memory of the People)*: WWII military service records, awards, burials, unit histories, operational documents (`pamyat-naroda.ru`)
2. *OVD-Info*: political detentions, administrative and criminal cases (`ovd.info/en`)

Figure 8: Another data point

Data type 3: **Public opinion surveys**

- Structured survey responses from sampled individuals (e.g., polls, social attitudes, approval ratings)
- *Common processing steps:* sample weighting, population estimation, aggregation, cleaning

Reliability of survey data in Russia

- Only 1 major non-govt pollster ↓
- Non-response rates low, but close to Western standards
- Self-censorship and preference falsification are pervasive

Examples:

1. *Levada Center*: monthly omnibus surveys, regular polling reports, analytical pubs (`levada.ru`)

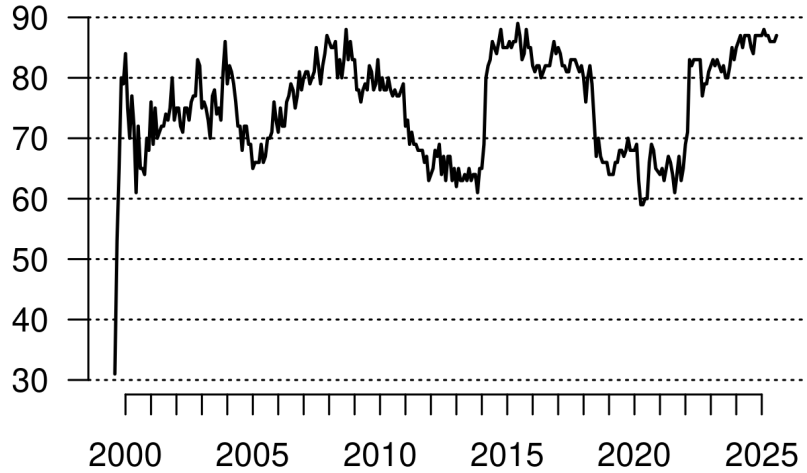

Figure 9: Putin approval (Levada)

Data type 4: **Text data**

- Unstructured or semi-structured textual content (e.g., archival documents, news articles, transcripts, social media posts)
- *Common processing steps:* Natural Language Processing techniques, translation, scaling, classification, topic modeling

Examples:

1. *University library e-resources*: EastView, Jane's, LexisNexis
2. *Books*: `militera.lib.ru`
3. *Documents*: `soldat.ru`
4. *Government*: `kremlin.ru`
5. *Social Media*: Telegram channels, social media news aggregators

Бородино
Михаил Юрьевич Лермонтов (1837)
— Скажи-ка, дядя, ведь не даром
Москва, спаленная пожаром,
Французу отдана?
Ведь были ж схватки боевые,
Да, говорят, еще какие!
Недаром помнит вся Россия
Про день Бородина!

Figure 10: Poems can be data

Data type 5: **Geospatial data**

- Location-specific numeric or imagery data (e.g., event coordinates, boundaries, roads, satellite images, gazetteers)
- *Common processing steps:* georeferencing, coordinate transformation, spatial joins, image analysis

Examples:

1. *Open data portals for cities*: Moscow (`data.mos.ru`), St. Petersburg (`gov.spb.ru`)
2. *Historical scanned maps*: `davidrumsey.com`, `tinyurl.com/28a7shm5`
3. *General GIS data links*: `freegisdata.rtwilson.com`

Figure 11: For official use only

Data type 6: **Non-geographic images**

- Rasterized images or vector graphics (e.g., photos, diagrams, blueprints, artwork)
- *Common processing steps:* computer vision, deep learning for object detection, classification, feature extraction

Examples:

1. *University library*: Angelica Image Database, Perry Photography Collection, GU Art Collection
2. *Image search engines*: DuckDuckGo, Google, Yandex

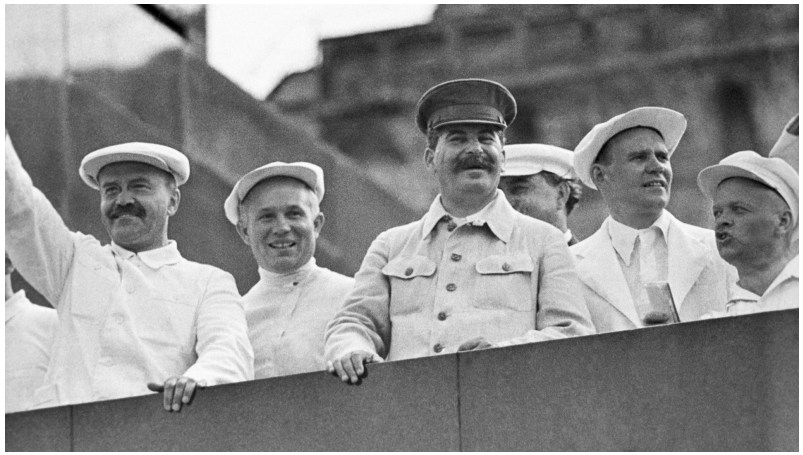

Figure 12: Who's who

Data type 7: **The dark side**

- OSINT from leaked/purchased data (e.g., financial files, private comms, personal identifiers)
- *Common processing steps:* strict ethical/legal review

In academic research and teaching, we *cannot* use such data due to privacy, consent, and institutional restrictions

Examples:

1. *Bellingcat*: investigative journalism (`bellingcat.com`)
2. *WikiLeaks*: pro-Russian cutout

Figure 13: Geolocated Buk-332

Data best practices

1. *Variable naming*: use "`snake_case`" for maximum compatibility (e.g., `putin_approval`, `region_name`, `year`, `month`)
2. *File format*: save processed data in delimited text format ('.csv')
3. *Create codebook*: variable names, descriptions, sources ('.pdf')
4. *Keep things organized!*

```
project_folder/
|-- data/
    |-- raw/
    |   |-- levada_survey_2025.csv
    |   |-- rosstat_demographics.xlsx
    |-- processed/
    |   |-- combined_data.csv
    |-- documentation/
        |-- codebook.pdf
```

NEXT MEETING

Economic Foundations: Land, Labor and Serfdom (Th, Sep. 11)

- the "origin story" of Russian autocracy, imperial expansion
- things to consider:
  - what incentives led Russia to adopt institution of serfdom
  - parallels and differences between forced labor practices in Russia vs. Western Europe vs. United States
  - why did the Russian state ultimately dismantle this institution?