

# BIG DATA –AINEISTOT, TIEDON LOUHINTA JA LAAJAT KIELIMALLIT TUTKIJAN NÄKÖKULMASTA

Mila Oiva

Kulttuurihistorian dosentti, Turun yliopisto

1.10.2024 – Tekoälyä ja tiedon louhintaa –webinaari

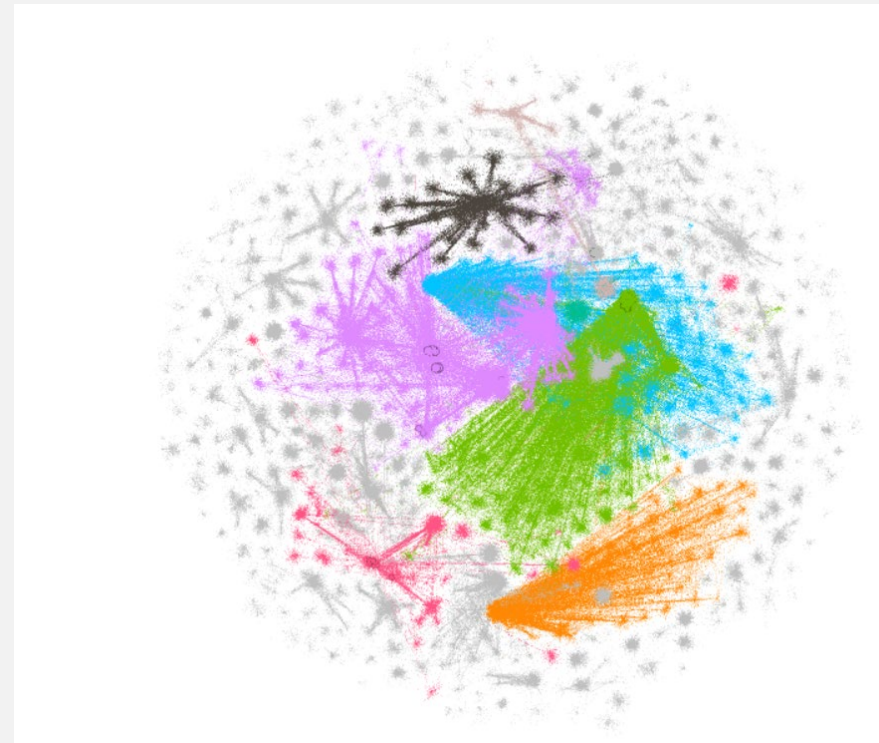
Suomen tieteellinen kirjastoseura, kirjastojuridiikan työryhmä

## MITÄ TUTKIN?

- Kulttuurihistoria
- Tutkimuksen kohteena miten maailmankuvat ja tieto muodostuvat ylirajaisesti
  - Neuvostoliiton ja Puolan välinen kulutustuotekauppa 1950- ja 1960-luvulla
  - Uutisten kansainvälinen liikkuminen 1800-luvun lehdistössä
  - Pseudohistorialliset narratiivit venäjänkielisessä internetissä
  - Neuvostoliittolaiset ja virolaiset uutiselokuvat 1920-luvulta 1990-luvulle

# MITEN TUTKIN?

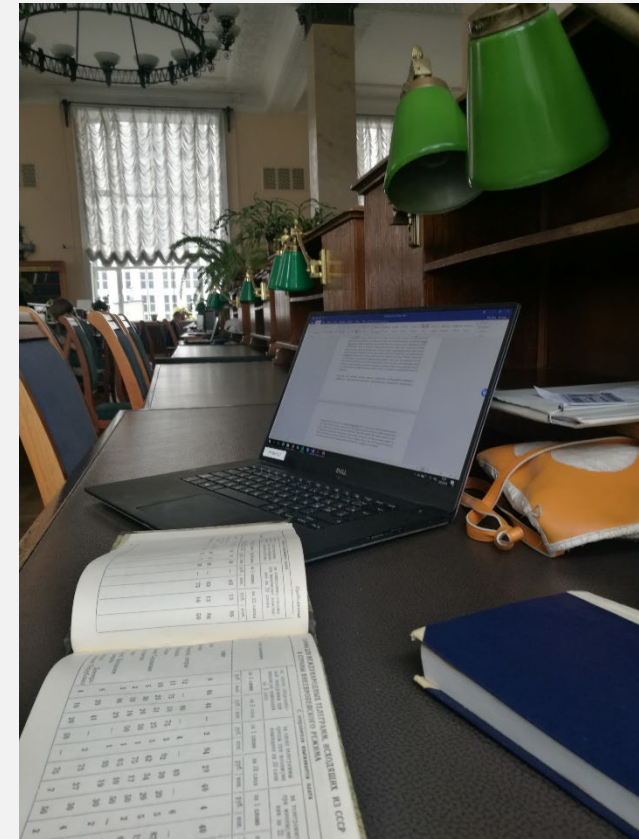
- laadullinen analyysi
- tekstin uudelleenkäyttö
- verkostanalyysi
- moniulotteiset vektorit
- laajat kielimallit



Venäjänkielisten keskiaikaa käsittelevien verkkotekstien hyperlinkkiverkosto, pääklusterit eri väreillä

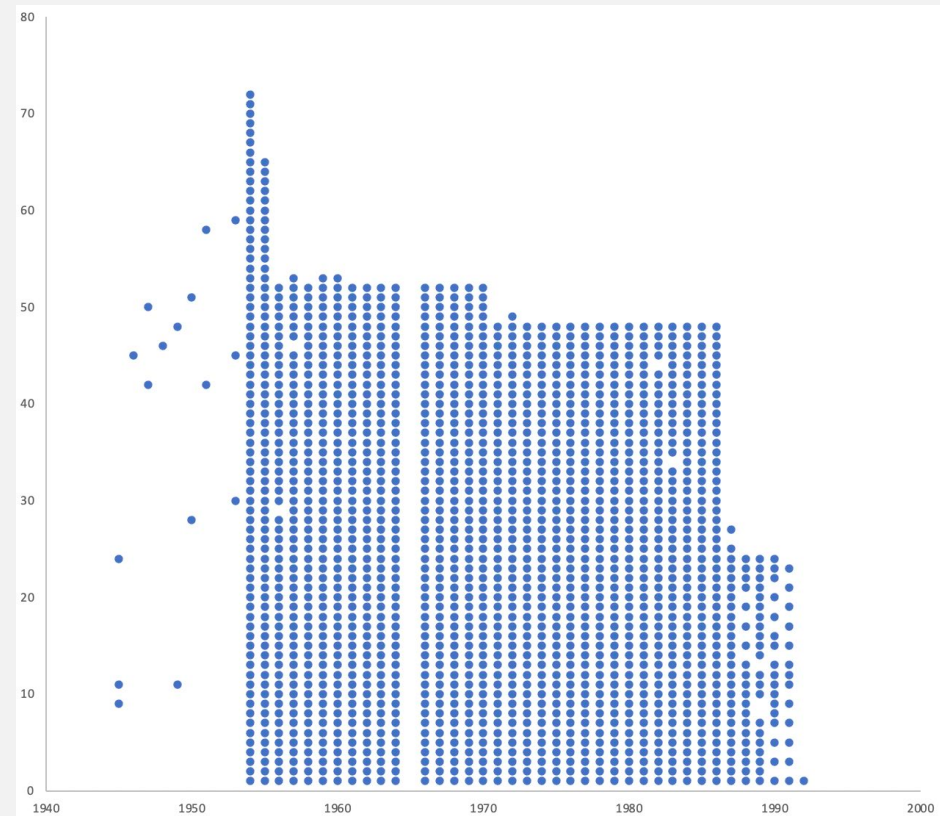
# AINEISTOT

- Arkistoaineistot
- Muistitietohaastattelut
- Verkkotekstit (blogit ja foorumkeskustelut)
- Digitaaliset sanomalehdet (suomalaiset; EastView: Pravda & Izvestija; Integrum)
- Uutiselokuvat (videot ja metadata)



# AINEISTOISSA TÄRKEÄÄ

- Saavutettavuus data dumpina (API)
- Laatu
- Mahdollisimmat yhtenäinen otos
- Tieto siitä kuka ne on tuottanut ja missä, ja miten ovat valikoituneet kokoelmaan



'Novosti dnia' –uutiselokuvasarjan saatavilla olevat julkaisut  
1944-1993

## SEURAAVAKSI

- Audiovisuaalisten aineistojen saatavuus
- Laajat kielimallit kulttuurin tutkimuksessa

# AUDIOVISUAALISTEN AINEISTOJEN SAATAVUUS

# AUDIOVISUAALISET AINEISTOT

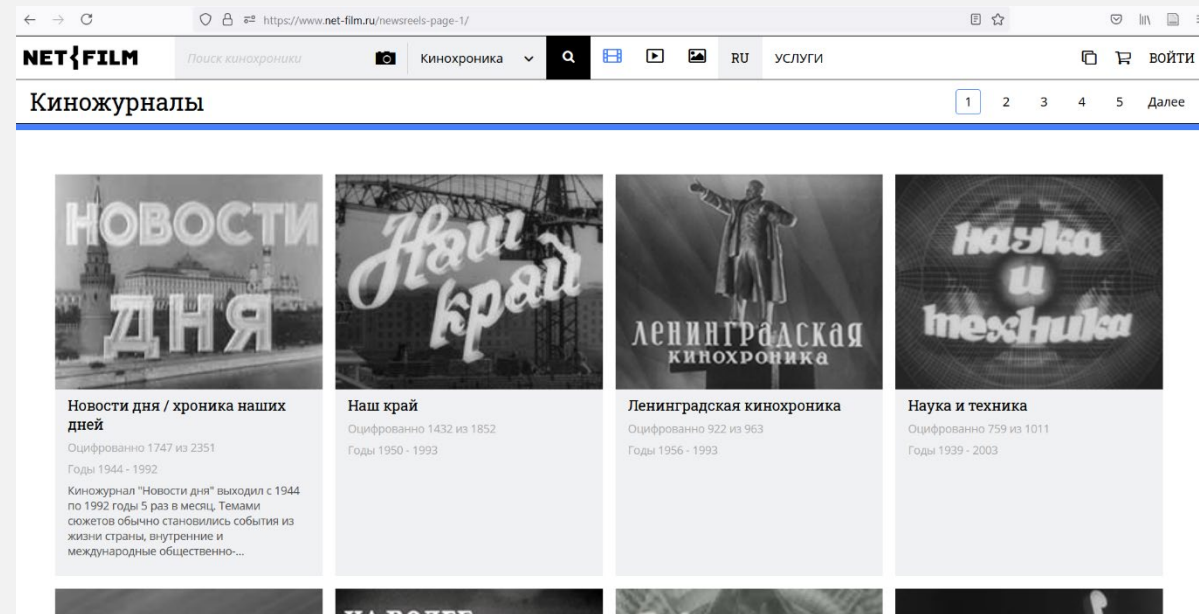
- Audiovisuaalisia kulttuuriperintöaineistoja digitoidaan paljon
- Niiden saatavuus suurina aineistokokonaisuuksina usein hankalaa tekijänoikeuskysymysten vuoksi
- Tämä on hidastanut audiovisuaalisten aineistojen tietokoneavusteisten tutkimusmenetelmien kehittymistä
- -> vaikea ymmärtää historiallista ja nykyistä yllirajaista tietoympäristöä jossa audiovisuaalisuuden merkitys on kasvava

# VENÄJÄNKIELISET JA VIROLAISET UUTISELOKUVAT



# VENÄJÄNKIELISET AINEISTOT

- Noin 10 000 uutiselokuvaa metatietoineen, yli 70 sarjaa
- Lupa hankkia ja käyttää tutkimukseen
- Aineistojen skreippaus
- Sodan alettua ei voida sopia aineistojen julkaisemisesta



# VIRONKIELISET AINEISTOT

- Positiivinen suhtautuminen digitoitujen aineistojen tutkimuskäyttöön
- Lisenssi- ja aineistonjakosopimukset Viron elokuvainsituutin, Viron elokuvatietokannan, ja Viron kansallisarkiston (elokuva-osasto) kanssa
- Neuvottelut 6 kk (projektin kesto 12 kk)

FILMIARHIIVI INFOSÜSTEEM

General Film Sound

Search

eesi | english | русский

Back

Filmoteegi materjal

| Title    | Genre | Language | In       | Out      | Duration | Producer     | Date          | Colour/Sound | Format | Osade arv |
|----------|-------|----------|----------|----------|----------|--------------|---------------|--------------|--------|-----------|
| chronide |       |          | 01:43:15 | 01:49:39 | 00:06:24 | Estonia Film | [1924]-[1927] | b/w, silent  | 35mm   | 1         |

User copy 102 BETACAM-SP, 9 VHS

Annotation

1. Vaated 1. Mai (esimene mai) rongkäigule Tallinnas. Demonstrandid Kaarli puistest ja Estonia puistest (teatri esine). Ratsapolitseinik hobusel. Töölisorganisatsioonide esindajad lippude ja loosungitega mererannas (Stroomi?) kõnekoosolekut pidamas.

2. Vaated marsevatele kaitseväe videosaade Tallinna tänaval. Viru tänav, ratsaväeosad, orkester, Pikk tänav, pealtvaatajad. Filmi lõpuosas toimub kaitseväelaste ja kaitsejõudude autasustamine Kadrioru lossi ees.

00:00:05:17

Suurenda

| Nr | Piece                                                | Language | Duration | Producer     | Authors |
|----|------------------------------------------------------|----------|----------|--------------|---------|
| 1  | [Töölisorganisatsioonide rongkäik 1. mail Tallinnas] |          | 00:03:12 | Estonia Film |         |
| 2  | [Kaitseväe ja kaitsejõudude paraad Tallinnas]        |          | 00:03:14 | Estonia Film |         |

Archive no. 14 Filmistuudio Estonia Film

# AINEISTONJAKOSOPIMUKSET TUTKIJAN NÄKÖKULMASTA

- Yhteistyö muissa yliopistoissa työskentelevien tutkijoiden kanssa -> oikeus lisensoida eteenpäin tutkimusryhmän sisällä
- Tutkijan työn liikkuvuus vs. sopimus yliopistokohtainen

## Sculpting Time with Computers Workshop

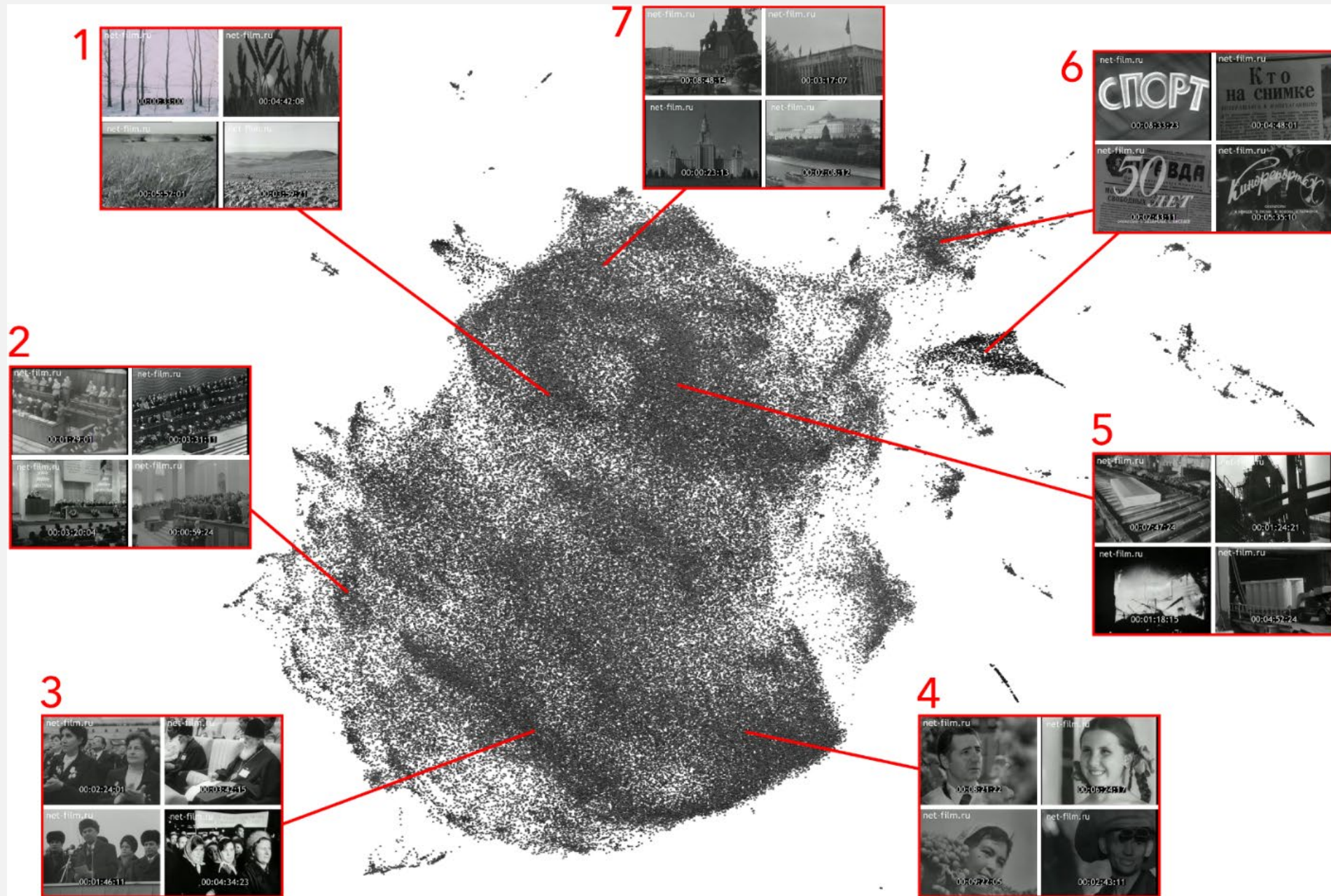
Daniel Chávez Heras



Kuva kesällä 2023 King's Collegessa järjestetystä "Sculpting Time with Computers" työpajasta

## AINEISTONJAKOSOPIMUKSET TUTKIJAN NÄKÖKULMASTA

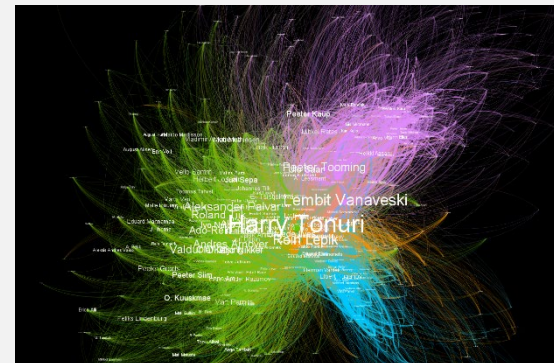
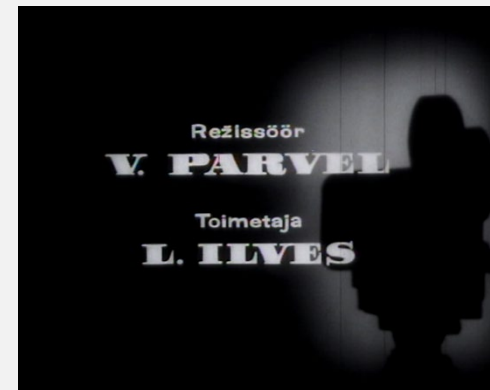
- Aineisto saatiin käyttöön ilman lissenssimaksuja, mutta sopimukseen kirjattiin mitä luovuttajaorganisaatiot hyötyvät (näkyvyys ja tietoisuuden lisääminen)
  - Tutkimus on epävarmaa -> vaikea määritellä etukäteen mitä aineiston osia tulee käyttämään ja miten
- > ”tutkimuskäyttö” määritelty laajassa mielessä: sisältää analyysin lisäksi käytön opetuksessa, esitelmissä, verkossa ja aineiston pohjalta luodut visualisoinnit



Oiva, Mila, Ksenia Mukhina, Vejune Zemaityte, Andres Karjus, Mikhail Tamm, Tillmann Ohm, Mark Mets, et al. "A Framework for the Analysis of Historical Newsreels." *Humanities and Social Sciences Communications* 11, no. 1 (April 25, 2024): 1–15. <https://doi.org/10.1057/s41599-024-02886-w>.

# HENKILÖTIEDOT JA EETTISET KYSYMYKSET

- Aineisto julkiseksi luotu ja julkisesti saatavilla
- Tekijätietojen esiintuominen ja työn tunnustaminen tärkeää
- Uudet menetelmät paljastavat uusia asioita ja esiin nousee kysymys tietosuojasta: työskentely ”Neuvostoliiton propaganda-järjestelmän hyväksi” tai ”nationalismin edistämiseksi Viron neuvostotasavallassa”



Yllä: Nõukogude Eesti 31/1966 uutiselokuvan tekijätietoja; alla: kaikkien uutiselokuvan tekijöiden yhteistyöverkosto 1927-1997

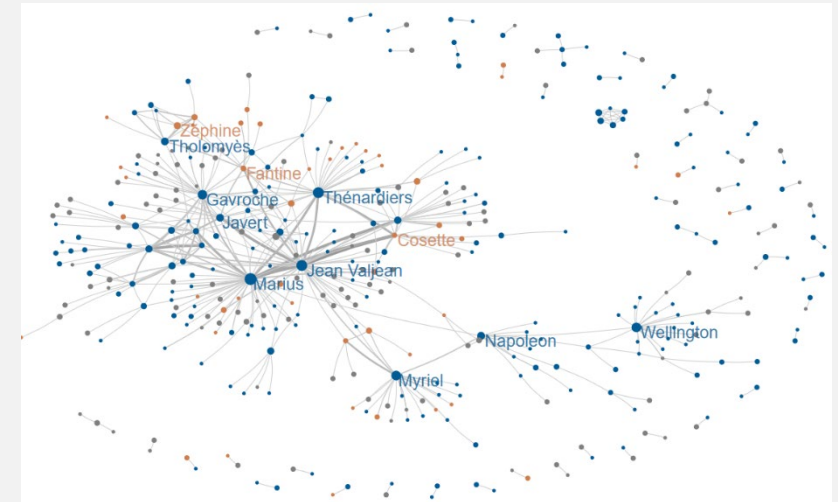
# LAAJAT KIELIMALLIT KULTTUURIN TUTKIMUKSESSA

## LAAJAT KIELIMALLIT KULTTUURIN TUTKIMUKSESSA

- Laajat kielimallit peilaavat niiden opettamiseen käytettyä aineistoa -> mukana vääristymiä koska kulttuurissa vääristymiä
- Olennaista käyttää ihmisen ohjauksessa ja verifioimana siihen missä ovat hyviä
- Laajentavat mahdollisuuksia kategorisoida laajoja teksti- ja kuva-aineistoja

# LAAJAT KIELIMALLIT KULTTUURIN TUTKIMUKSESSA

- Esimerkiksi tekstin teeman tai sentimentin määrittely, sanan semanttisen merkityksen tunnistaminen, tapahtuman syyn tunnistaminen, genren tunnistus, jne.
- Jo aiemmin mahdollista raskaan koneoppimisen kautta, mutta nyt helpommin käytettävissä promptien kautta
- Karjus, Andres. “Machine-Assisted Mixed Methods: Augmenting Humanities and Social Sciences with Artificial Intelligence.” arXiv, September 24, 2023. <https://doi.org/10.48550/arXiv.2309.14379>.



ChatGPT:n avulla ”Kurjat” –kirjasta tunnistetut hahmot, niiden sukupuoli ja yhteisesiintyminen romaanissa.

# ESIMERKKI I: UUTISELOKUVIEN UUTISPÄTKIEN KATEGORISOINTI METATIETOJEN AVULLA

| Story_ID | Story_Outline                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                 | Story_Topic_Category       | chatGPT     |
|----------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----------------------------|-------------|
| 12336_01 | 1. Кинообзор достижений, событий, итогов 1970 года. 45: Празднование 100-летия со дня рождения В. И. Ленина в г. Ульяновске. 70: Автоматический самоходный "Луноход-1", созданный советскими учеными, на Луне. 95: Производство стали в цехе металлургического завода. 120: Уборка зерновых. 145: Вид гидроэлектростанции. 169: Производство автомобилей в цехе автозавода. 194: Делегаты XVI-го съезда ВЛКСМ. 219: Интервью новаторов производства рабочих Г. Ивановской, Д. Никитина, фигуристов И. Родниной и А. Уланова (синхронно). 244: | progress of the USSR       | progress    |
| 12336_02 | 2. Фабрика елочных украшений в Высоковске. 294: Предновогодняя продажа игрушек в магазинах.                                                                                                                                                                                                                                                                                                                                                                                                                                                   | industrial production      | culture     |
| 12336_03 | 3. Производственные процессы в цехе Куйбышевской шоколадной фабрики "Россия".                                                                                                                                                                                                                                                                                                                                                                                                                                                                 | industrial production      | industrial  |
| 12336_04 | 4. Новое здание кукольного театра в Москве. 418: Дети на кукольном спектакле.                                                                                                                                                                                                                                                                                                                                                                                                                                                                 | culture                    | culture     |
| 12336_05 | 5. Новогодний праздник в Москве. 493: Новогодняя елка для детей. 518: Новогодний молодежный бал.                                                                                                                                                                                                                                                                                                                                                                                                                                              | festivities                | festivities |
| 12337_01 | 1. Собрание представителей трудящихся Москвы, посвященное принятию социалистических обязательств в честь XXIV съезда КПСС. Выступает В. В. Гришин.                                                                                                                                                                                                                                                                                                                                                                                            | domestic political meeting | political   |
|          | 2. Производственные процессы в цехах завода "Москабель", включившегося в социалистическое соревнование за повышение эффективности труда. 108: Митинг строителей Капчагайской                                                                                                                                                                                                                                                                                                                                                                  |                            |             |

# ESIMERKKI I: UUTISELOKUVIEN UUTISPÄTKIEN KATEGORISOINTI METATIETOJEN AVULLA

This is the classification prompt:

Classify this USSR news Text with a topic Tag. ONLY use one of the Tags defined in this list:

Politics = use for political events and messages, presidiums, Communist Party, Komsomol; communism, socialism, Leninism, political leaders and premiers; foreign governments and politics; international relations; but do NOT use for wars or industry news.

Military = military, national defense, wars, battles; but NOT politics.

Science = scientific and industrial progress and construction; space and aviation; technological advancements and innovations.

Social = social issues and lifestyle, education, students and schools, family, health, leisure, arts, culture, religion, ceremonies.

Disasters = disasters, fires, weather warnings; but NOT wars.

Sports = sports events; sports results and scores; athletic performances.

Farm = domestic USSR agriculture, farming, hunting, kolkhozes; but NOT foreign news, NOT economics.

Industry = USSR economy, domestic business, economic trends, economy plans; industry, mining; industry workers, brigades; but NOT politics, NOT international news, NOT agriculture, NOT tourism.

Misc = use this for any other topics, only if no Tag above fits well; or if topic is unclear.

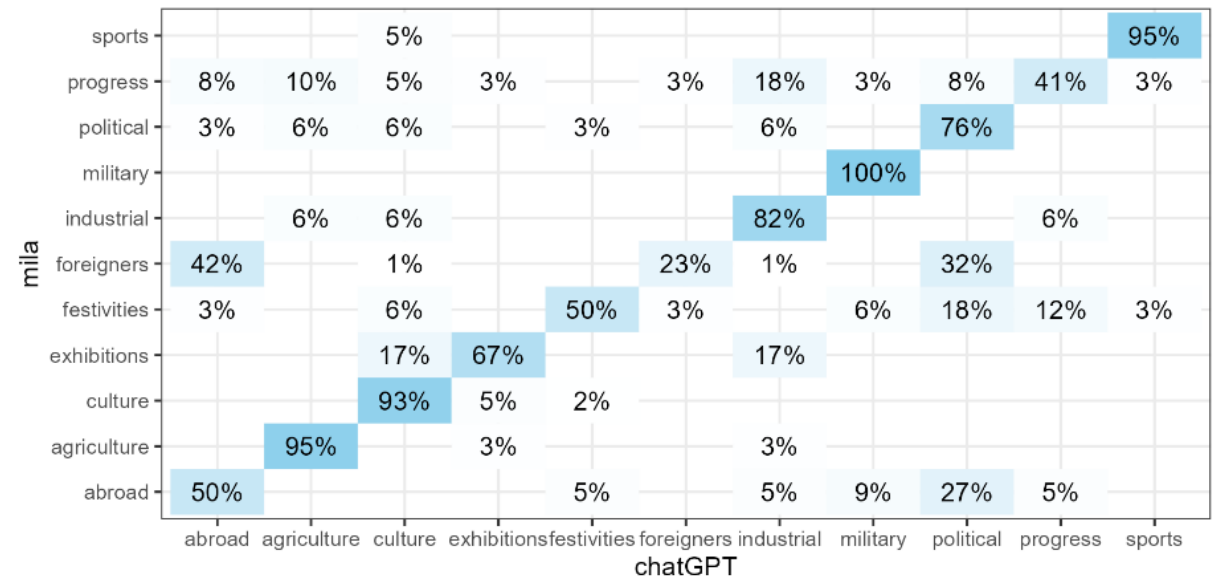
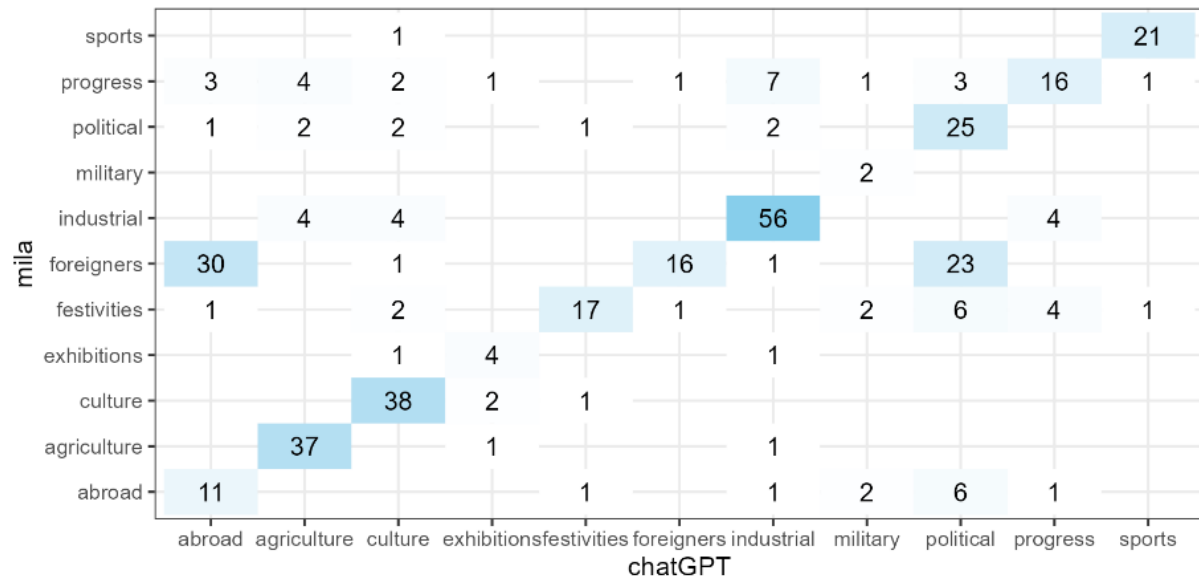
Classify text: [followed by the synopsis text]

This is the additional exploratory prompt:

You are a cultural historian and Soviet history expert. Analyze this corpus of Soviet era newsreels synopses (one per line, unrelated texts). Come up with a set of the most frequent general content categories that could be used to describe the themes and topics in a newsreels dataset like this. Give me 10 topics, where the last is a Misc topic which subsumes all the remaining less frequent topics. Output the topics as a comma-separated list of keywords or short key phrases, in English. Do not expand the keywords or comment, just give a list. The corpus:

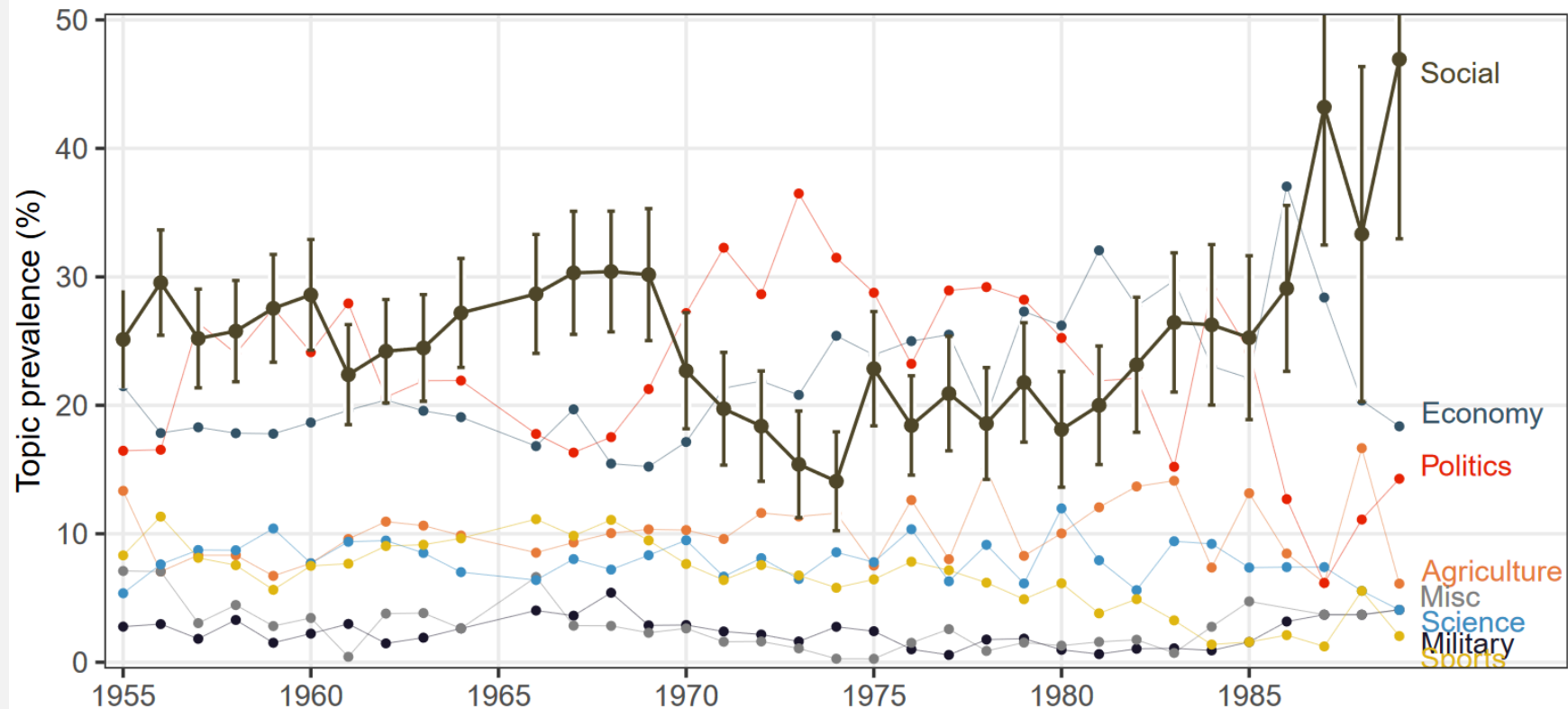
[followed by a sample from the corpus]

# ESIMERKKI I: UUTISELOKUVIEN UUTISPÄTKIEN KATEGORISOINTI METATIETOJEN AVULLA



# ESIMERKKI I: UUTISELOKUVIEN UUTISPÄTKIEN KATEGORISOINTI METATIETOJEN AVULLA

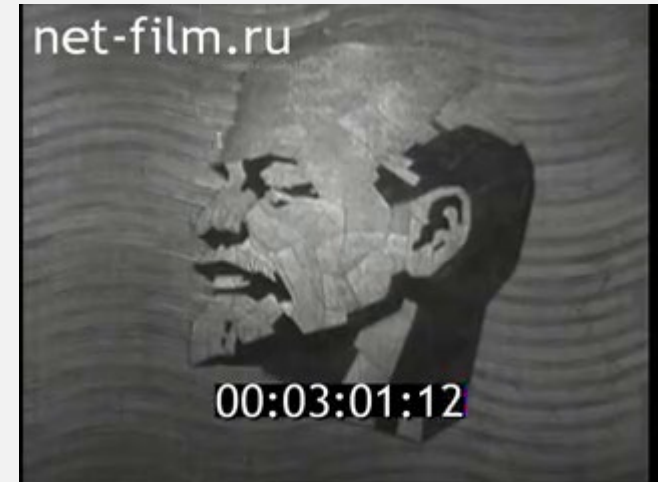
(A) Predicted topics in Soviet newsreels 1955–1989



## ESIMERKKI 2: LENINISMI UUTISELOKUVISSA

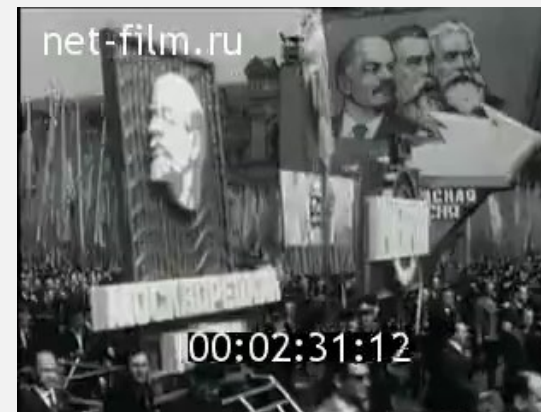
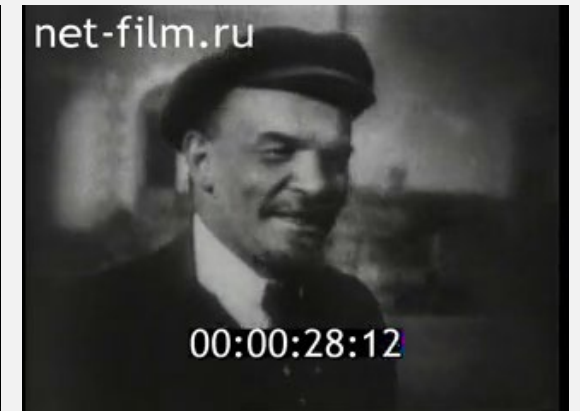
”Quantifying Leninism?” on työn alla oleva tutkimus, jossa

- Testataan kuinka eri kielimallisovelluksia voi käyttää tunnistamaan tunnettuja historiallisia henkilöitä visuaalisesta aineistosta
- Visuaalisen kulttuurin kvantifioimisen rajojen etsiminen

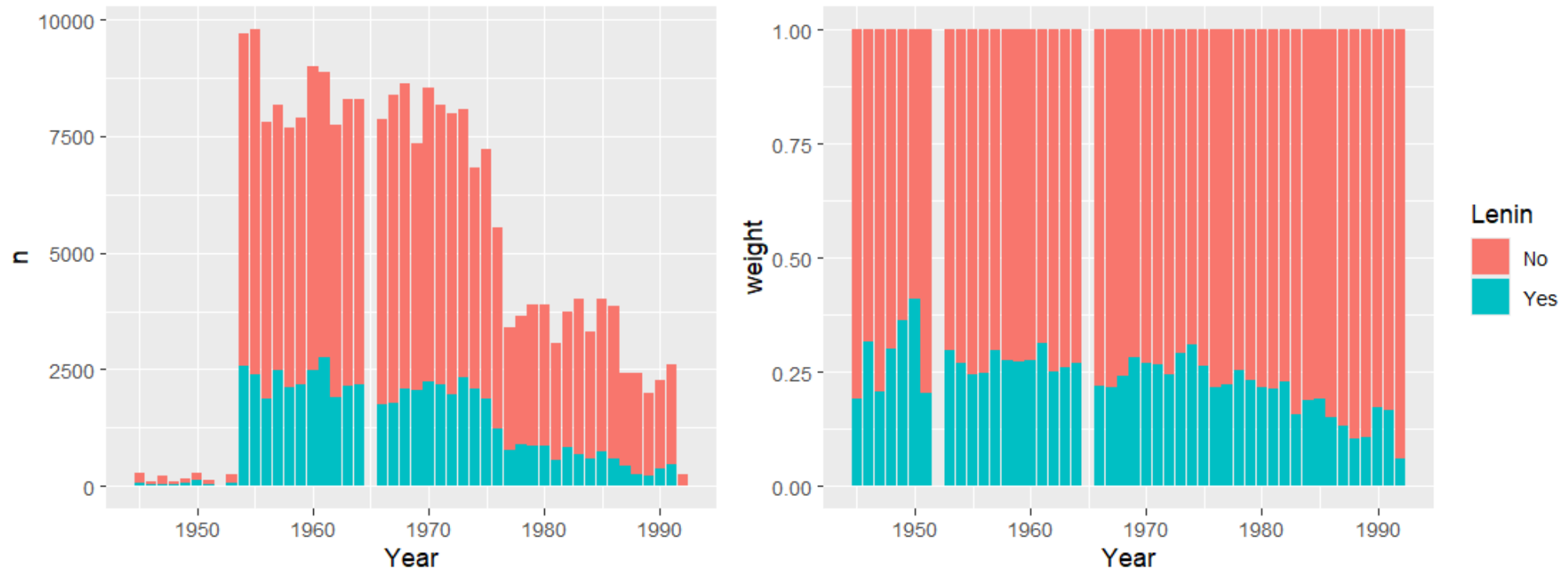


## ESIMERKKI 2: LENINISMI UUTISELOKUVISSA

- Tunnistetaan elävä Lenin ja Lenin taiteessa, (Lenin muumio)
- CLIP, Llava I.5 ja GPT-4 Vision vs. Milan annotointi -> 97% osuvuus



## ESIMERKKI 2: LENINISMI UUTISELOKUVISSA



Alustavia tuloksia: vasemmalla Lenin-kuvien määrä ja oikealla niiden prosenttiosuus vuositasolla.

# LOPUKSI

Tutkijan näkökulmasta olennaista

- aineistojen mahdollisimman helppo saatavuus (sekä toimivat hakutoiminnot laadullista että API määrällistä aineiston hakua varten)
- mahdollisimman laaja käyttöoikeus (myös laajat kielimallit)
- Tarvitaan keskustelua uusien menetelmien kautta saatavien tulosten julkistamistavoista, laajojen kielimallien käytön ekologisista ja yhteiskunnallisista vaikutuksista, ja vaihtoehdoista suurten ylikansallisten yhtiöiden työkaluille

# KIITOS!

[milaoiv@utu.fi](mailto:milaoiv@utu.fi)

Lisenssi: CC BY-SA