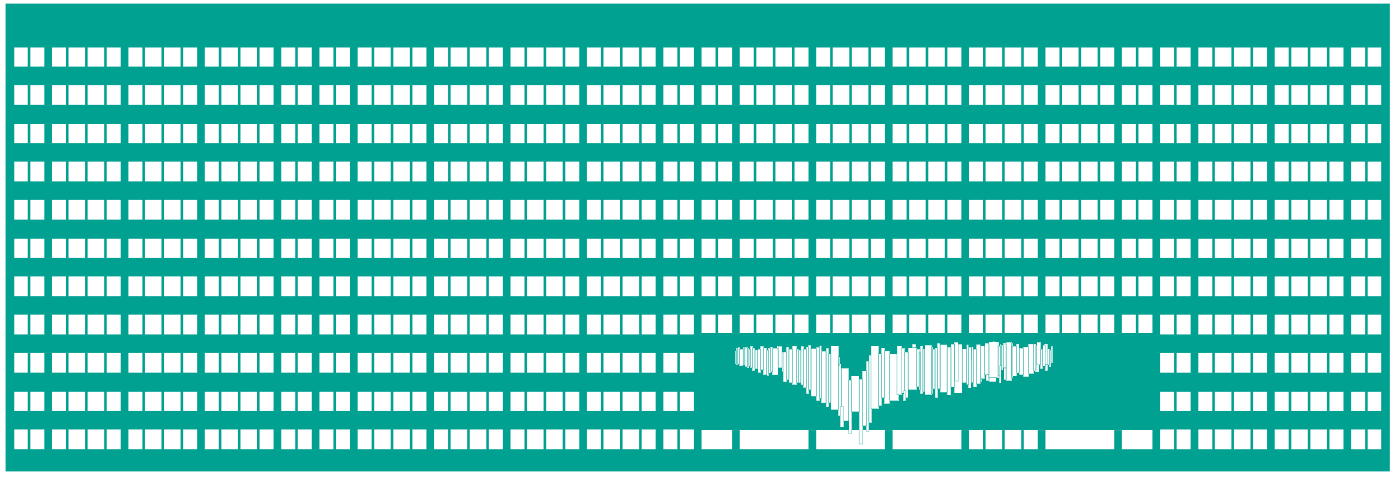
VŠB TECHNICKÁ UNIVERZITA OSTRAVA

VSB TECHNICAL UNIVERSITY OF OSTRAVA



www.vsb.cz

Komprese stromových struktur

Semestrální projekt

Marek Beran

VŠB – Technická univerzita Ostrava

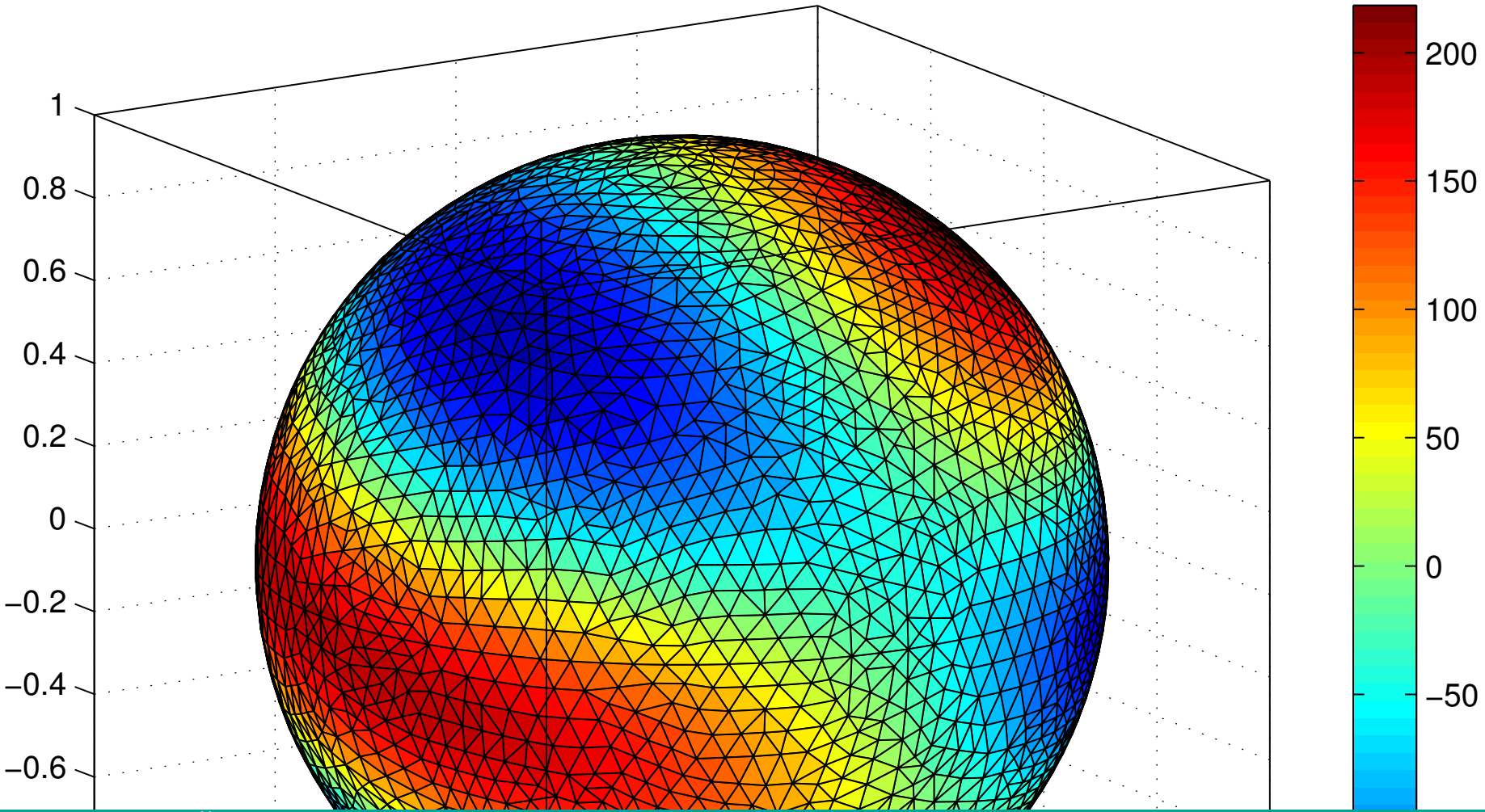
marek.beran.st@vsb.cz

23. května 2025

VŠB TECHNICKÁ UNIVERZITA OSTRAVA | FAKULTA ELEKTROTECHNIKY A INFORMATIKY

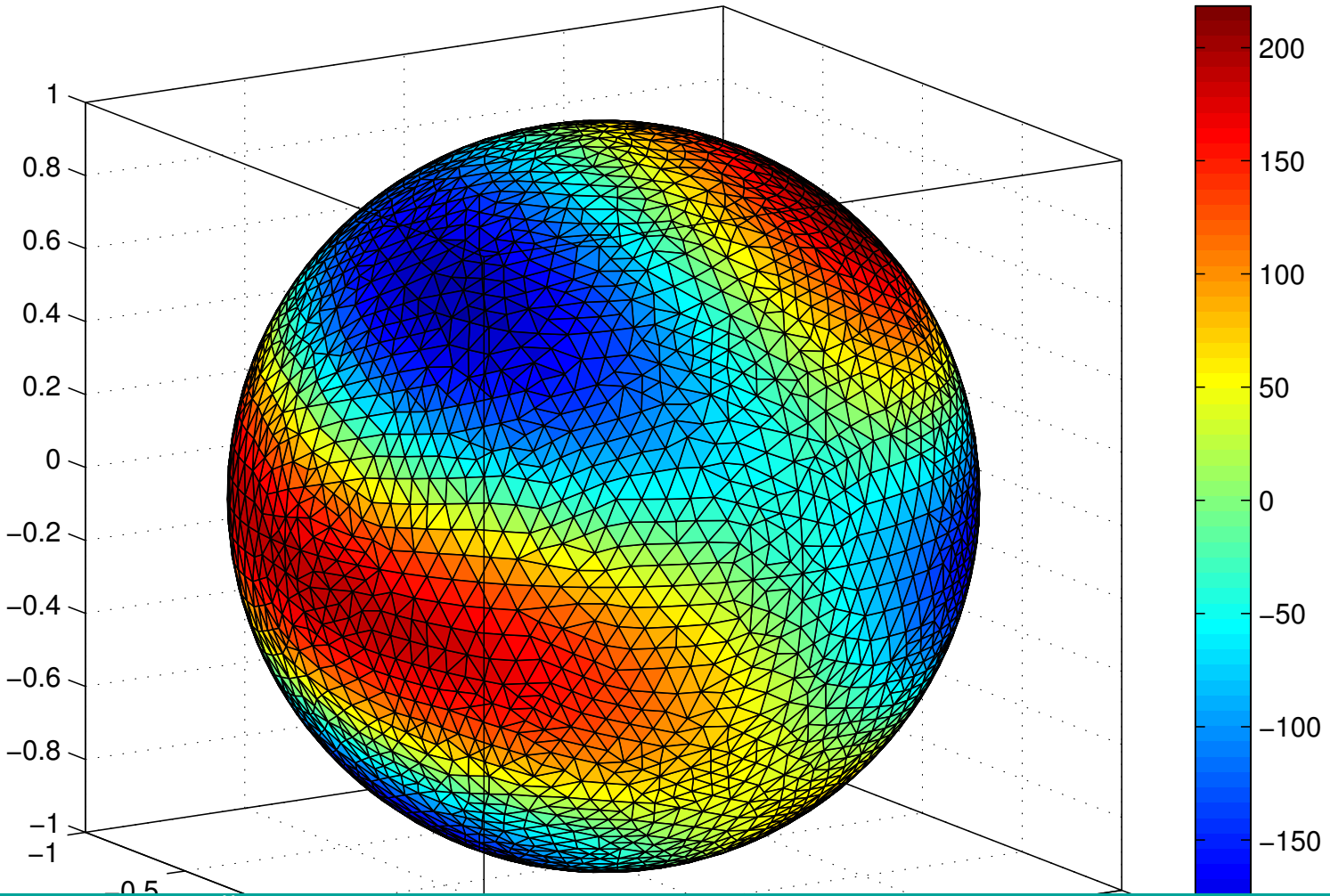
Proč komprimovat stromy?

- Stromy = přirozený výstup syntaktické analýzy
- Text obsahuje opakující se vzorce
- Hypotéza: komprese možná díky opakovatelnosti



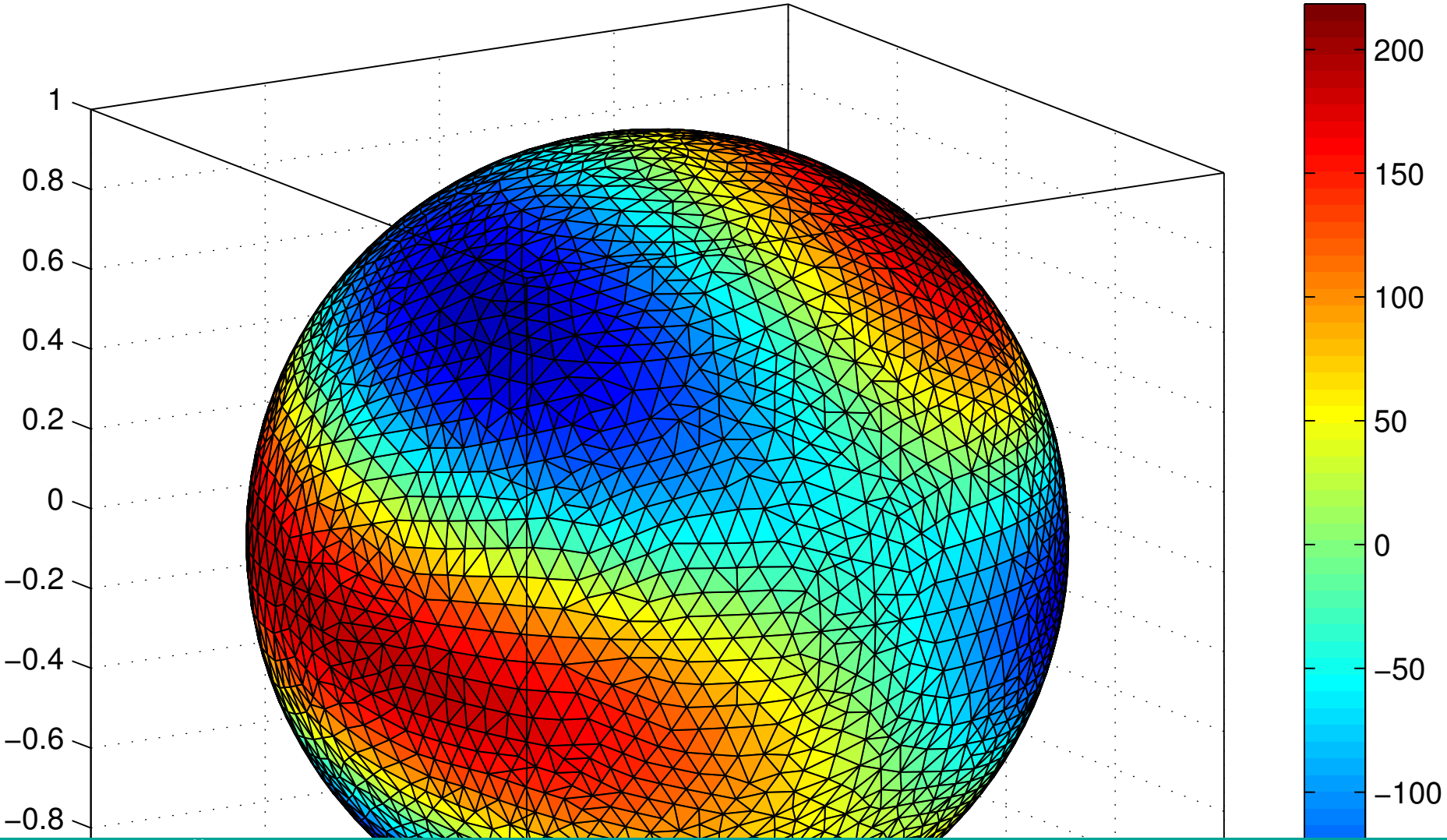
Cíle projektu

- Komprimovat syntaktické stromy
- Porovnat různé metody
- Vyhodnotit efektivitu



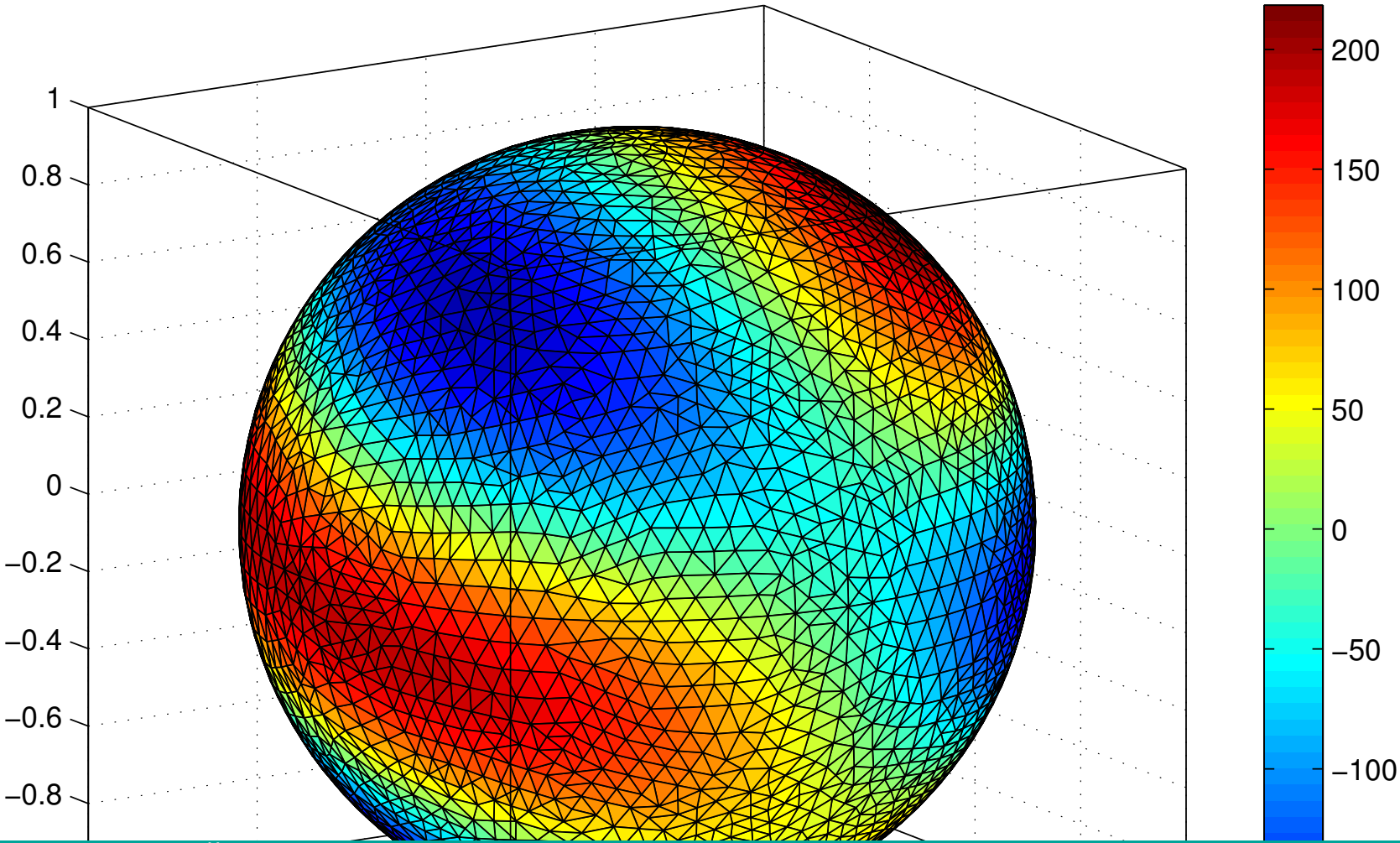
Z textu na strom

- MorphoDiTa → lemmatizace
- UDPipe → syntaktický strom



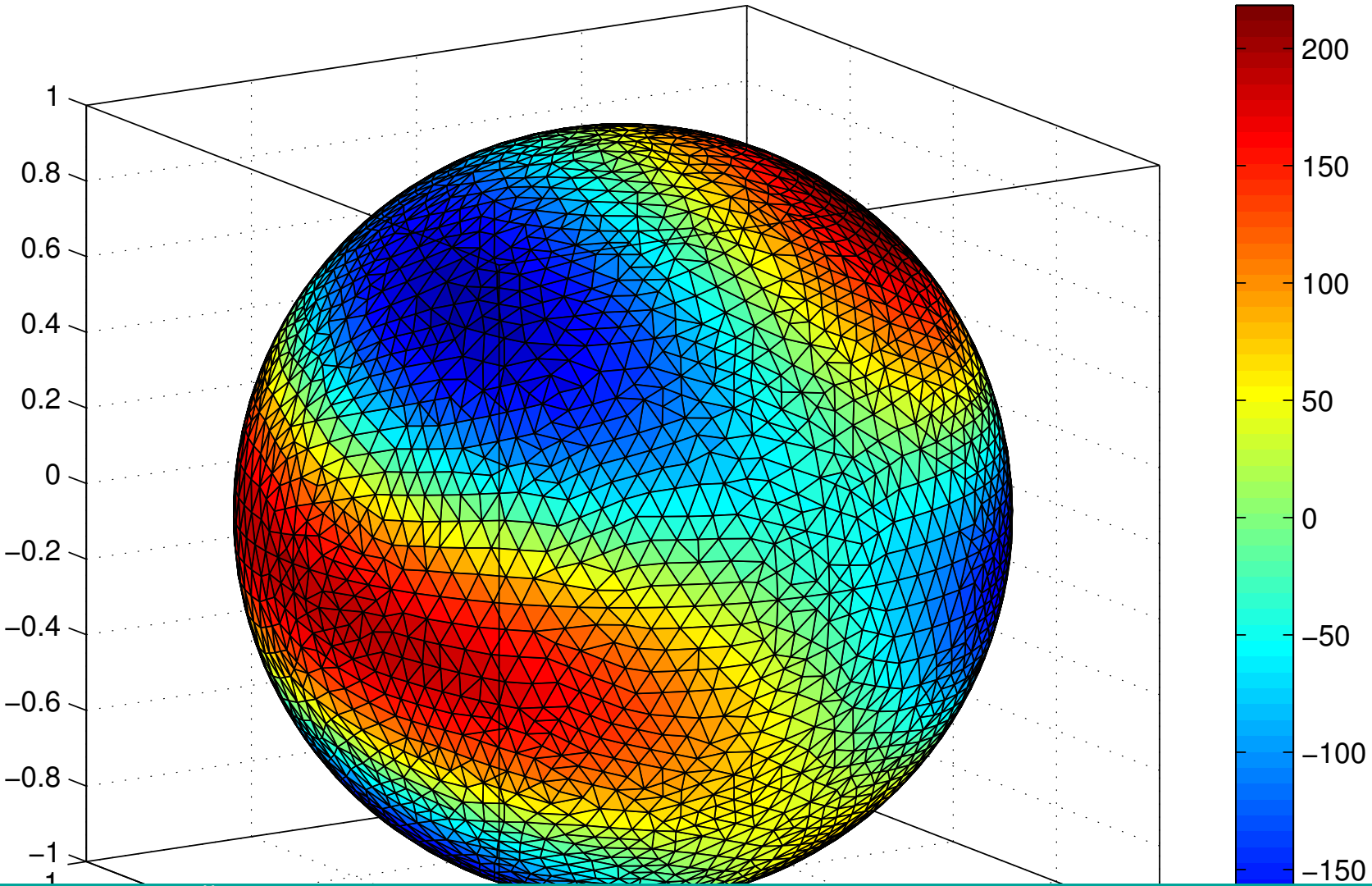
Struktura knihovny

- Vzor Pipes and Filters
- Moduly pro analýzu, kompresi, dekompresi
- Konzolová aplikace pro testování



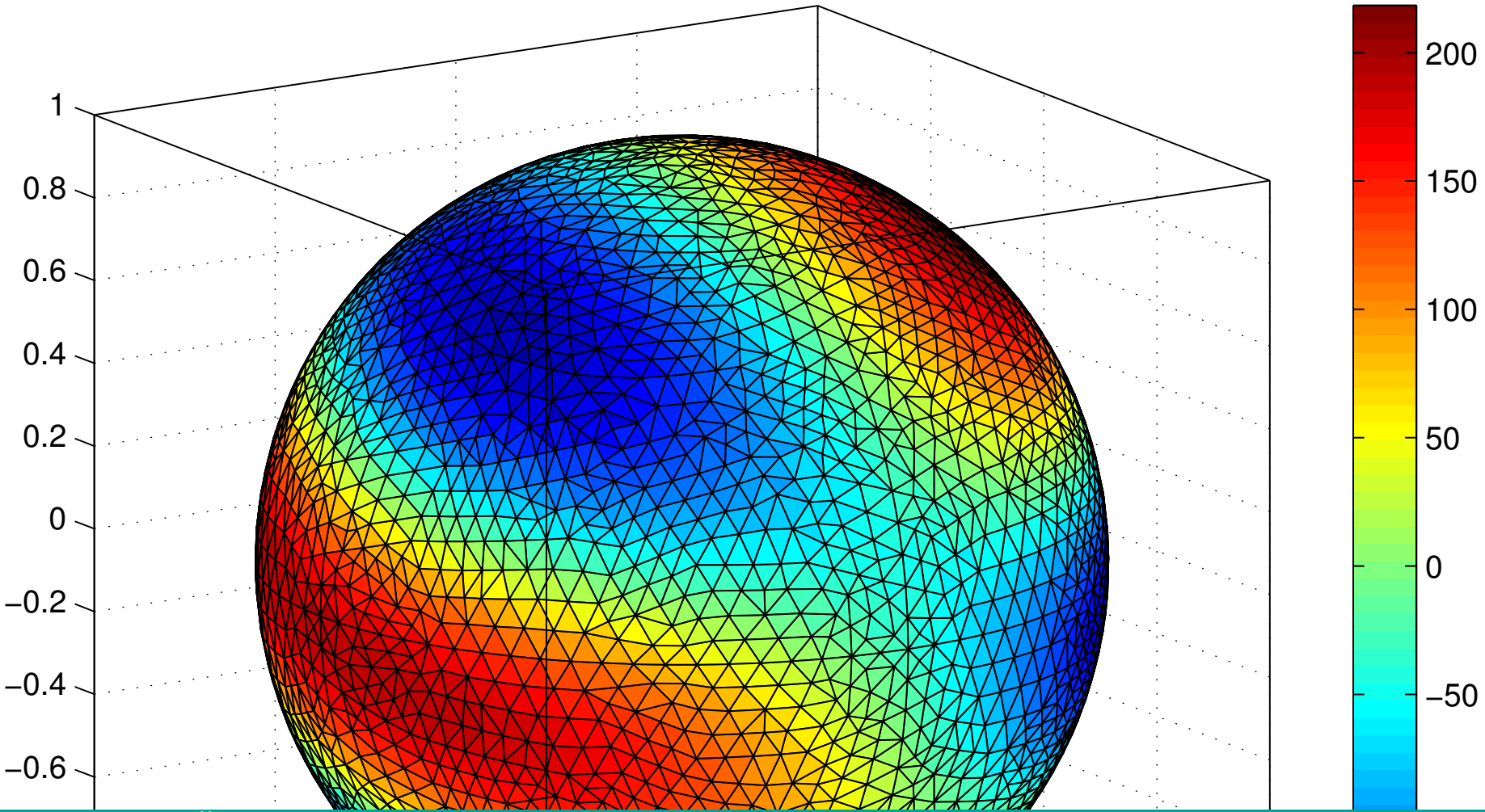
Pipeline scénáře

- Komprese: Text → Strom → Komprese
- Dekompres: Načtení → Dekomprese → Ověření



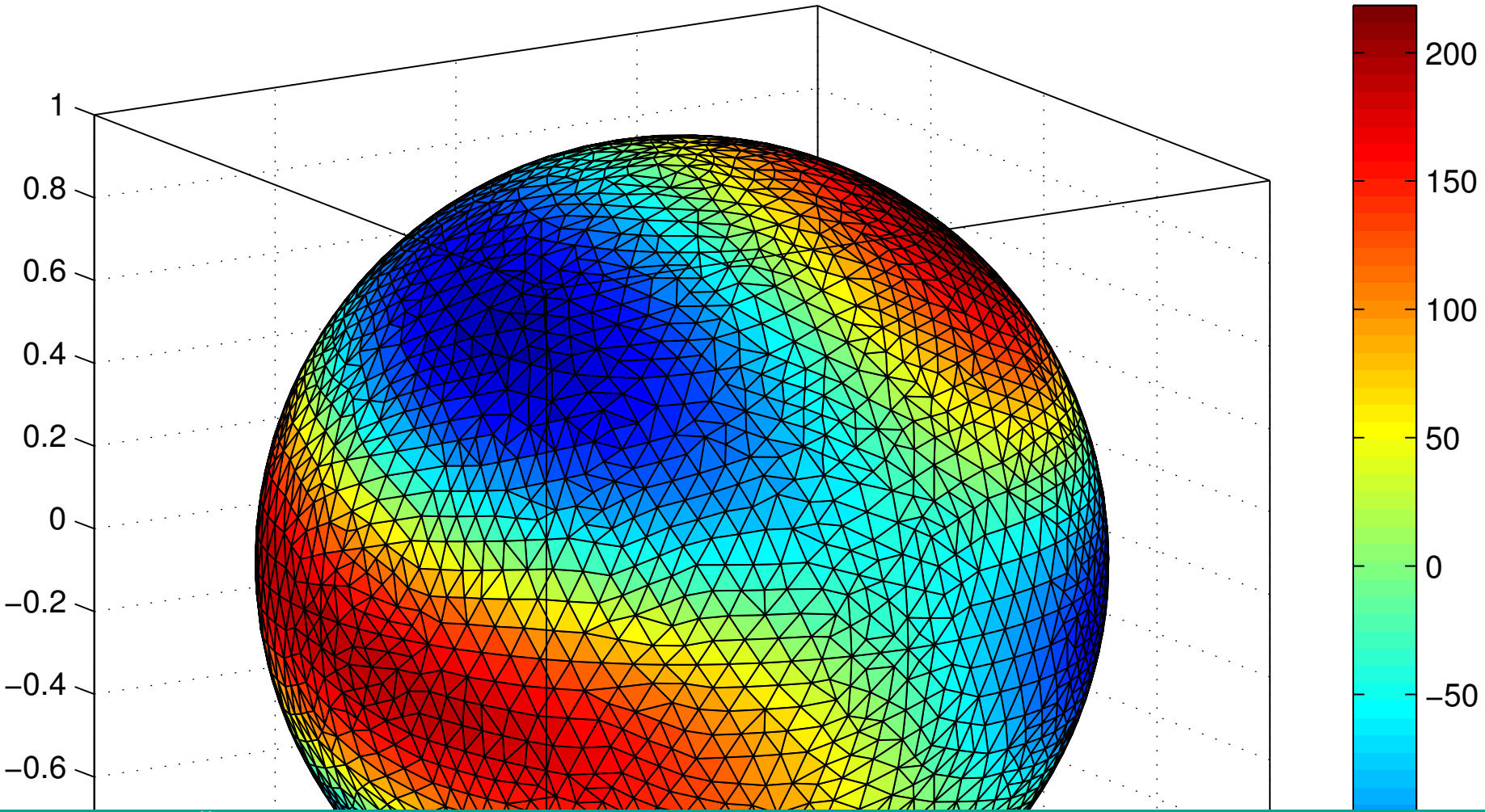
Algoritmus RePair

- Hledání opakujících se párů
- Nahrazení neterminály
- Vytvoření gramatiky



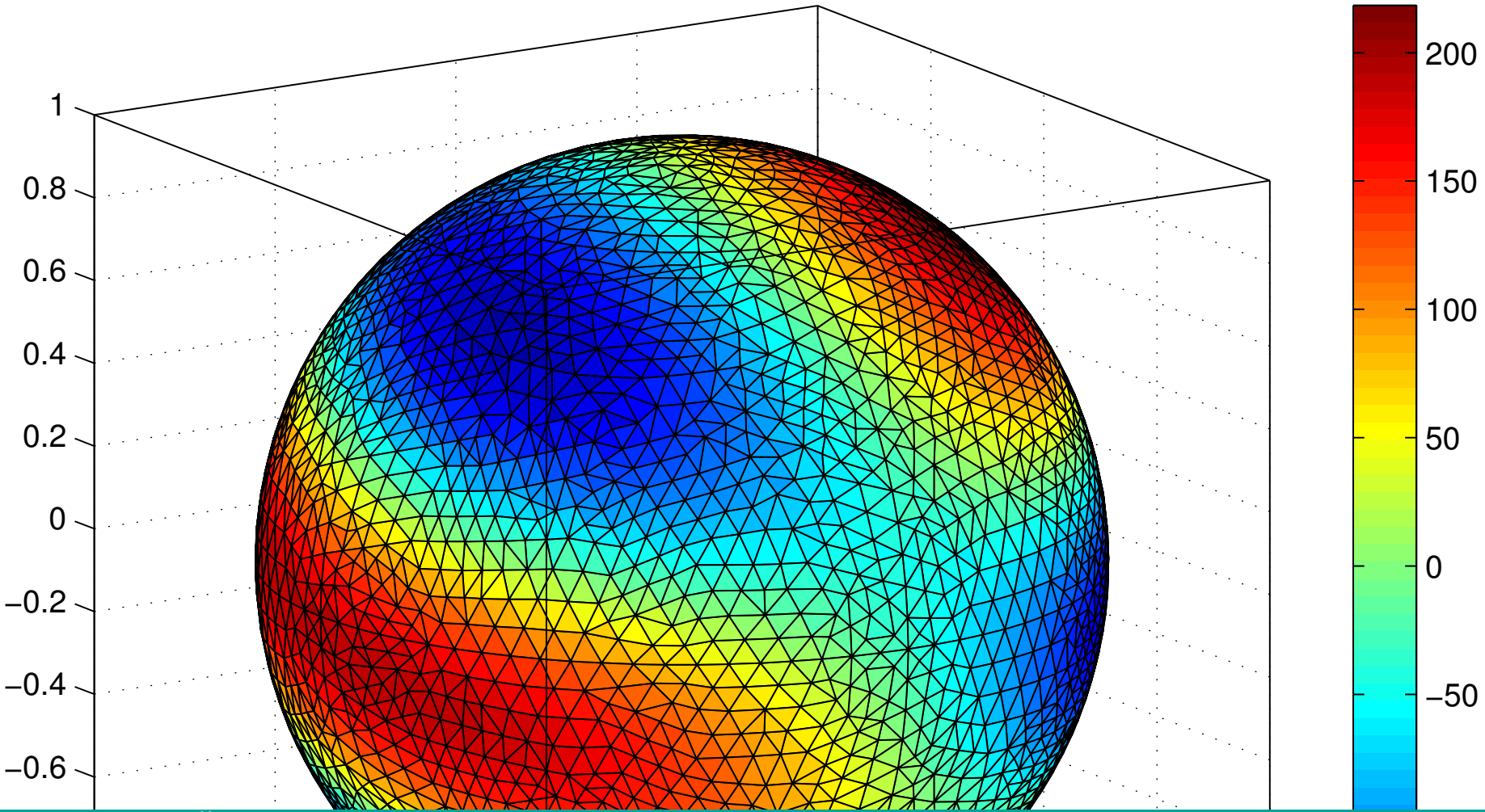
TreeRePair – bez linearizace

- Hledání opakujících se podstromů
- Hodnocení podle četnosti a velikosti
- Nahrazení pravidly



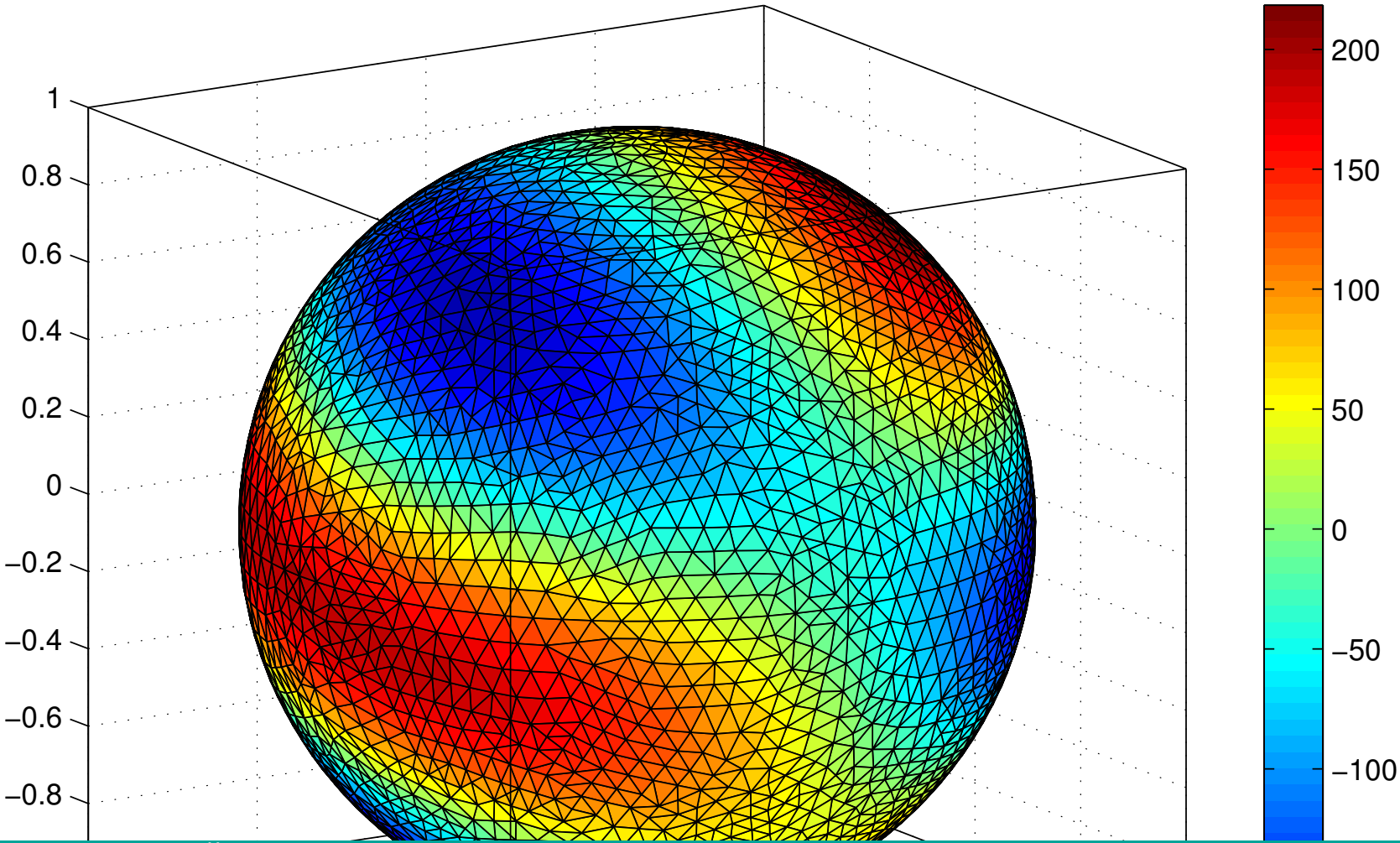
Linearizace + RePair

- Preorder průchod → sekvence
- RePair → pravidla
- Vylepšení: n-gramy, kontext



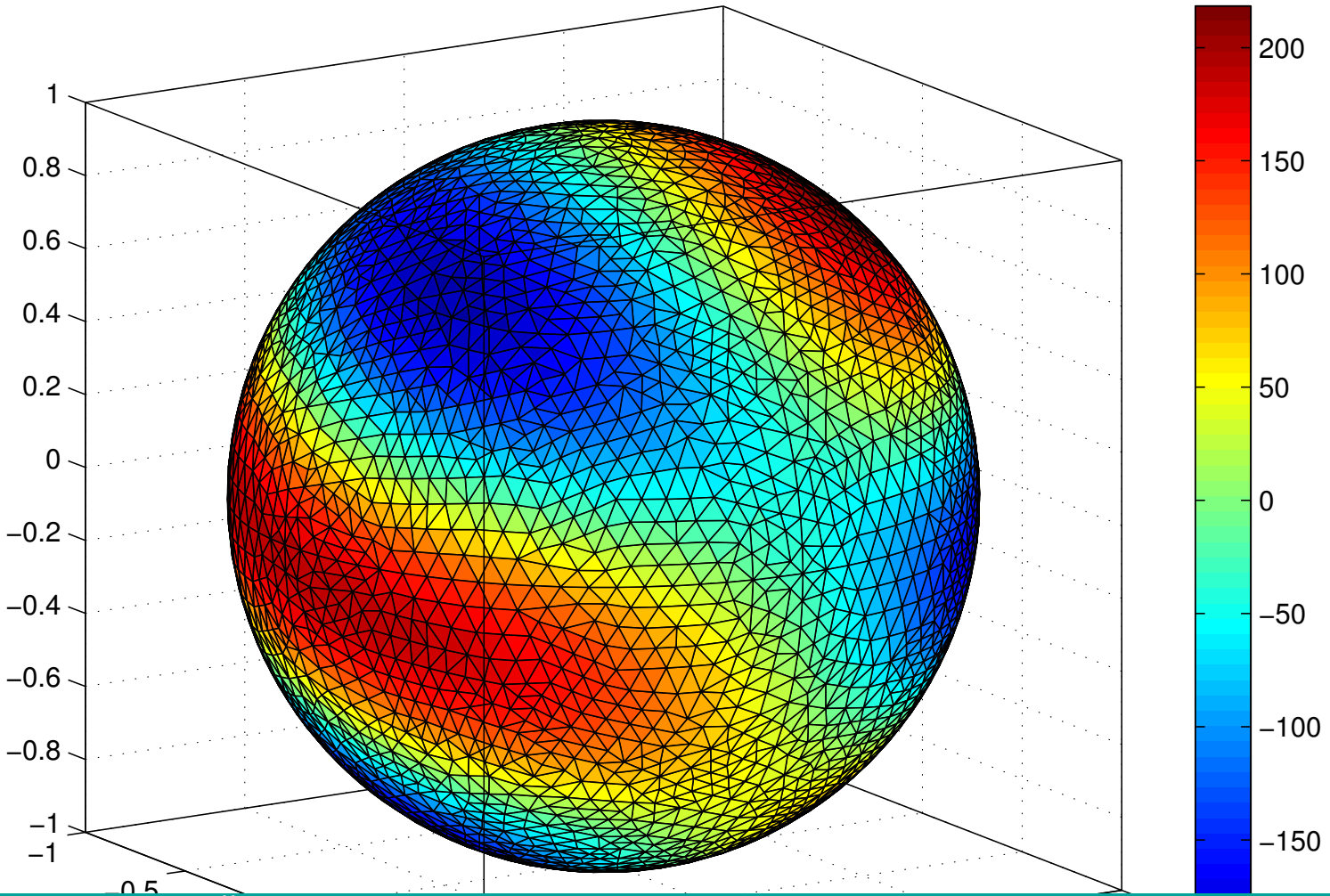
Shrnutí výsledků

- Linearizace + RePair: nejlepší průměrné výsledky
- TreeRePair: náročnější, méně efektivní
- Skutečná komprese: cca 9



Závěr a výhled

- Komprese stromů není triviální
- Optimalizace výrazně pomáhají
- Možnosti do budoucna: paralelizace, adaptivní metody



Děkuji za pozornost

Marek Beran

VŠB – Technická univerzita Ostrava

marek.beran.st@vsb.cz

23. května 2025

VŠB TECHNICKÁ UNIVERZITA OSTRAVA | FAKULTA ELEKTROTECHNIKY A INFORMATIKY