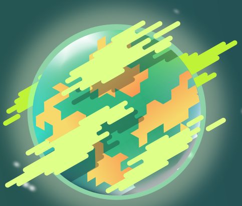
BEAM me up: A tale of Bounded Expansion Algorithms in Metagenomics



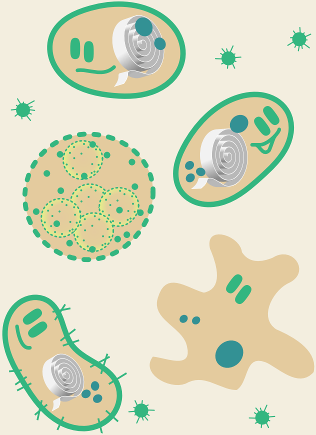
Blair D. Sullivan
NC State University
blair_sullivan@ncsu.edu

Workshop on Structural Sparsity, Logic and Algorithms

Part I

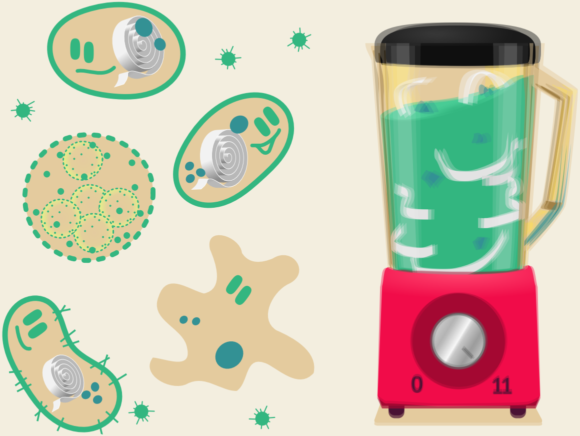
The scientific story

Metagenomics



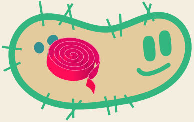
Metagenomics

Metagenomics

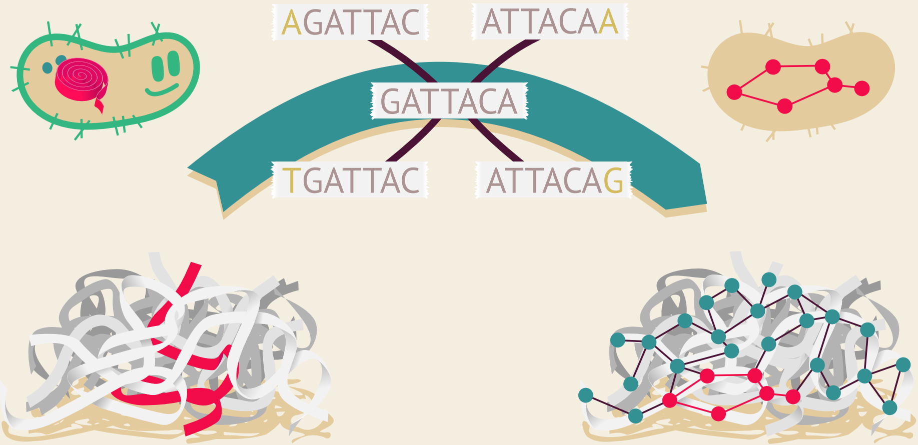


Metagenomics

De-Bruijn graphs



De-Bruijn graphs



AGATTAC

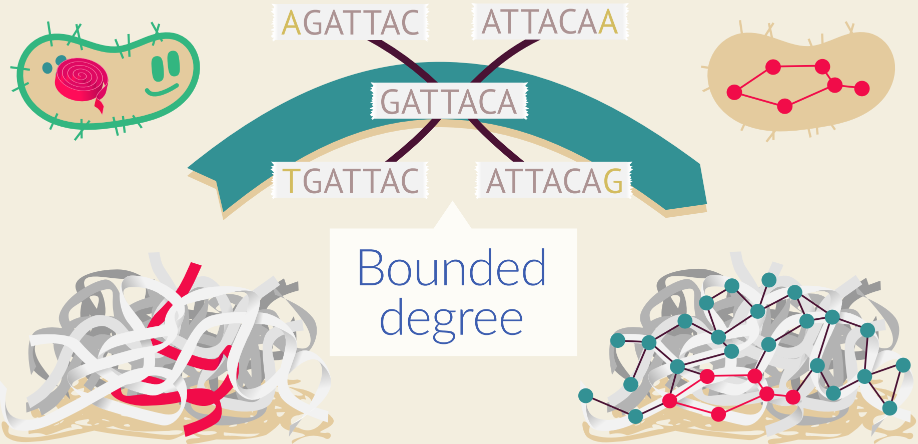
ATTACAA

GATTACA

TGATTAC

ATTACAG

De-Bruijn graphs



AGATTAC

ATTACAA

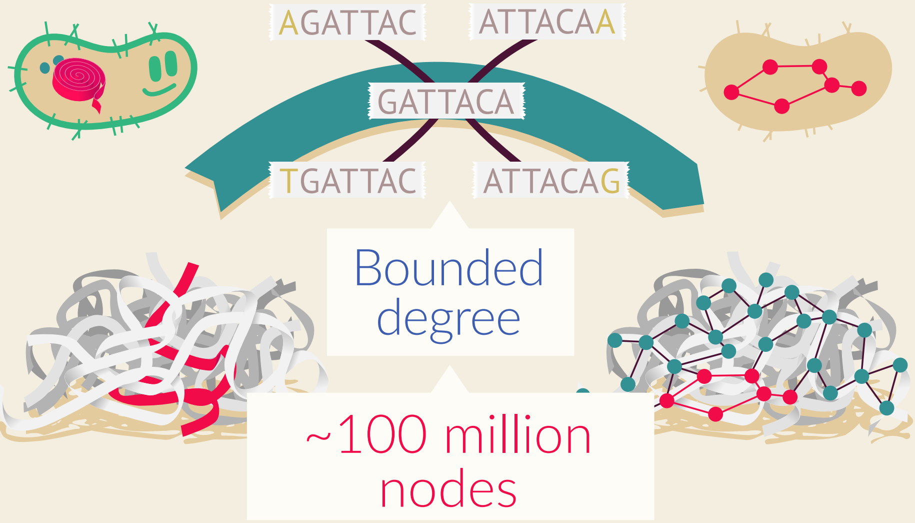
GATTACA

TGATTAC

ATTACAG

Bounded degree

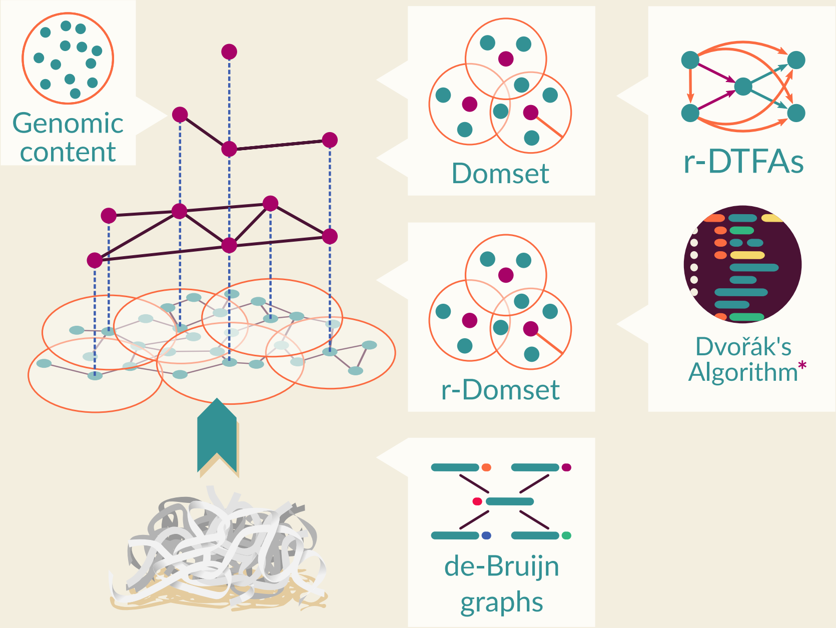
De-Bruijn graphs



Bounded degree

~100 million nodes

CATLAS Overview



Genomic content

Domset

r-DTFAs

r-Domset

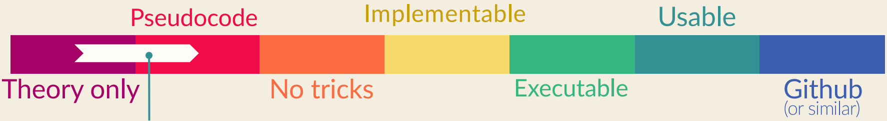
Dvořák's Algorithm*

de-Bruijn graphs

Engineering Dvořák's algorithm

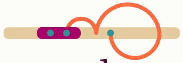
wcol_{2r}

Approximation is terrible in practice



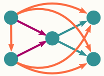
Dvořák Z. **Constant-factor approximation of the domination number in sparse graphs**.
European Journal of Combinatorics.
2013 Jul 31;34(5):833-40.

Engineering Dvořák's algorithm



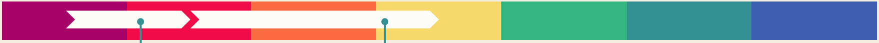
wcol_{2r}

Approximation is terrible in practice



$\Delta^{-}(\vec{G}_{2r})$

Approximation is terrible in practice



Pseudocode

Implementable

Usable

Theory only

No tricks

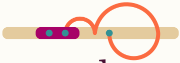
Executable

Github
(or similar)

Dvořák Z. **Constant-factor approximation of the domination number in sparse graphs.** European Journal of Combinatorics. 2013 Jul 31;34(5):833-40.

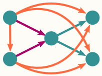
Reidl F. **Structural sparseness and complex networks.** (Doctoral dissertation, Dissertation, Aachen, Techn. Hochsch., 2015).

Engineering Dvořák's algorithm



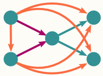
wcol_{2r}

Approximation is terrible in practise



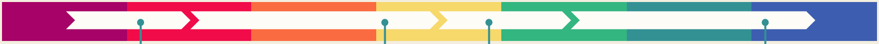
$\Delta^{-}(\vec{G}_{2r})$

Approximation is terrible in practise



$\Delta^{-}(\vec{G}_{r})$

Approximation is tunable (heuristic)



Pseudocode

Implementable

Usable

Theory only

No tricks

Executable

Github
(or similar)

Dvořák Z. **Constant-factor approximation of the domination number in sparse graphs.** European Journal of Combinatorics. 2013 Jul 31;34(5):833-40.

Reidl F. **Structural sparseness and complex networks.** (Doctoral dissertation, Dissertation, Aachen, Techn. Hochsch., 2015).

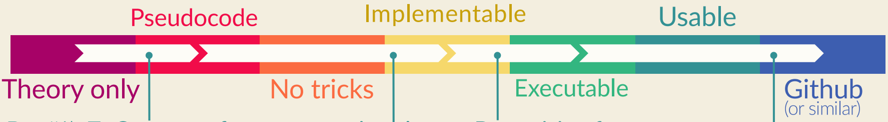
Barnraising for Data-Intensive Discovery at MDI Biological Laboratory in Salisbury Cove, Maine.

Ongoing collaboration with **Theory in Practice Group** (NCSU) and **Lab for Data Intensive Biology** (UC Davis)
https://github.com/spacegraphcats/spacegraphcats

Engineering Dvořák's algorithm

For input a graph G and integers r,t our algorithm computes an $2(t+2)\Delta^{-}(\vec{G}_{2r})\Delta^{-}(\vec{G}_{r})$ -approximate r-dominating set. Importantly, it computes only the rth dtf-augmentation $\vec{G}_r$.

Fudge-factor t: small t yield better approximation guarantee in the worst case, but larger dominating sets in practice!



Pseudocode

Implementable

Usable

Theory only

No tricks

Executable

Github
(or similar)

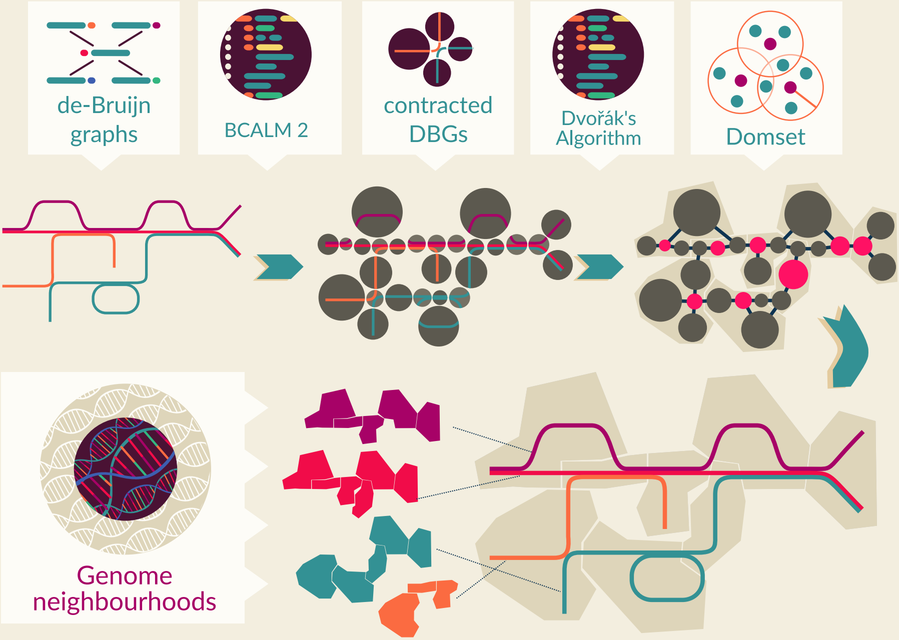
Dvořák Z. **Constant-factor approximation of the domination number in sparse graphs.** European Journal of Combinatorics. 2013 Jul 31;34(5):833-40.

Reidl F. **Structural sparseness and complex networks.** (Doctoral dissertation, Dissertation, Aachen, Techn. Hochsch., 2015).

Barnraising for Data-Intensive Discovery at MDI Biological Laboratory in Salisbury Cove, Maine.

Ongoing collaboration with **Theory in Practice Group** (NCSU) and **Lab for Data Intensive Biology** (UC Davis)
https://github.com/spacegraphcats/spacegraphcats

CATLAS-1 (spacegraphcats)



de-Bruijn graphs

BCALM 2

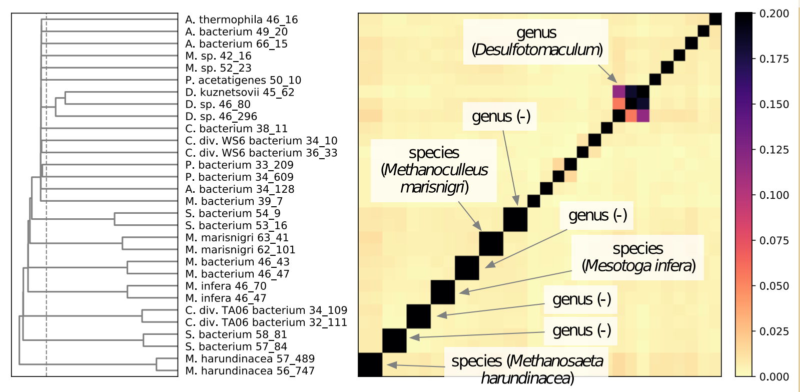
contracted DBGs

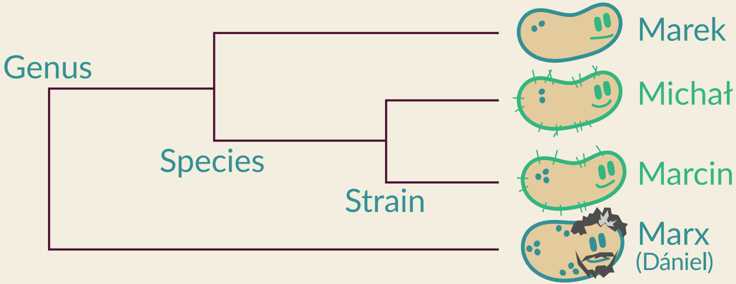
Dvořák's Algorithm

Domset

Genome neighbourhoods

CATLAS-1 Results





Genus

Species

Strain

Marek

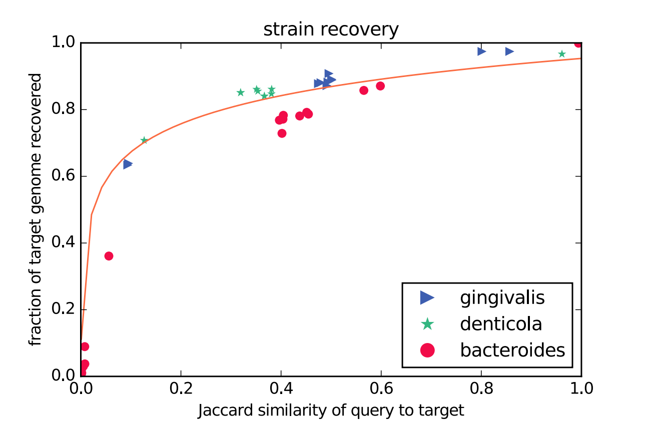
Michał

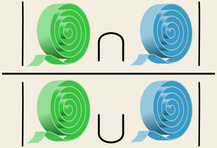
Marcin

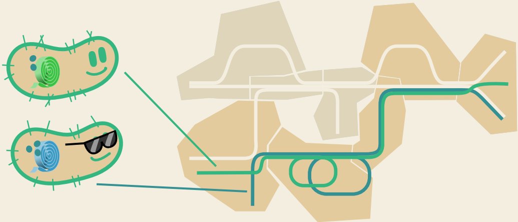
Marx

(Dániel)

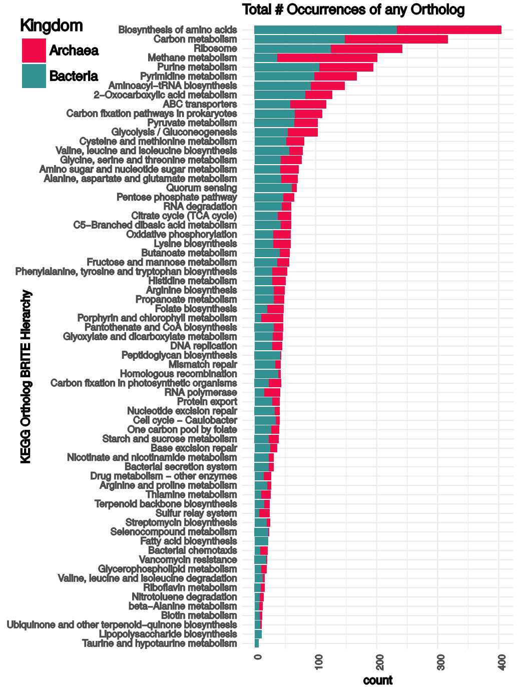
CATLAS-1 Results

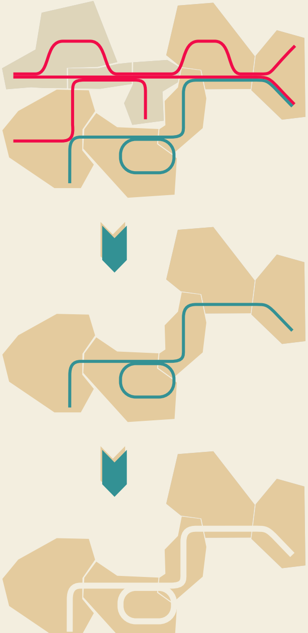






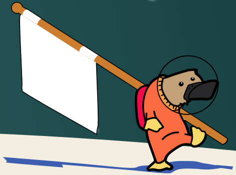
CATLAS-1 Results



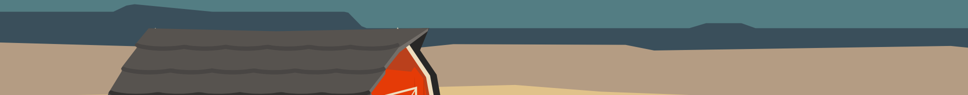


Part II

The *real* story



A long time ago, in a galaxy far, far away...



~~A long time ago, in a galaxy far, far away...~~

In May 2016, at a barn raising in rural Maine...

"Sparse Graph Cuts"

The barnraising crew

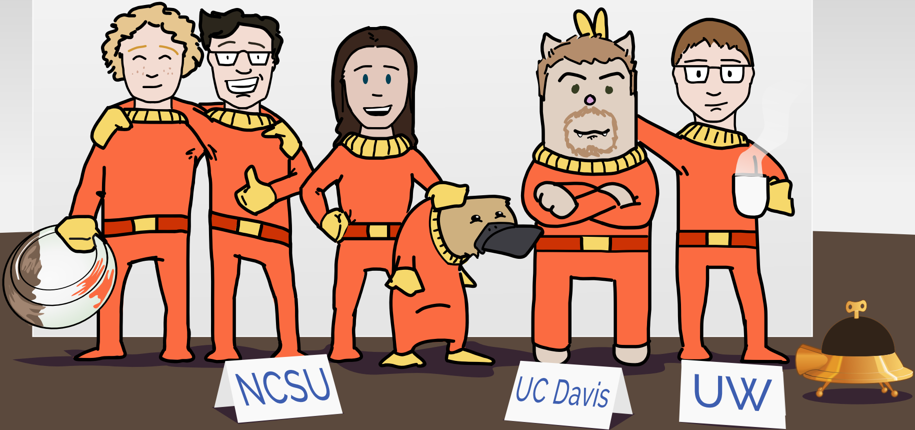
Felix Reidl

Mike O'Brien

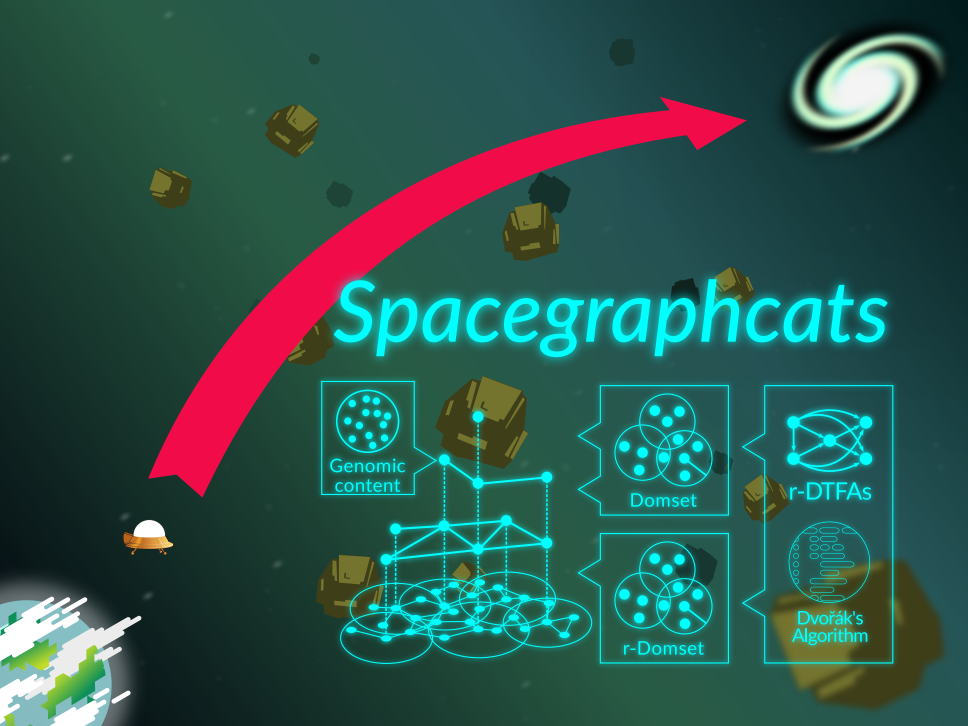
Blair Sullivan

Titus Brown

Dominik Moritz



Spacegraphcats



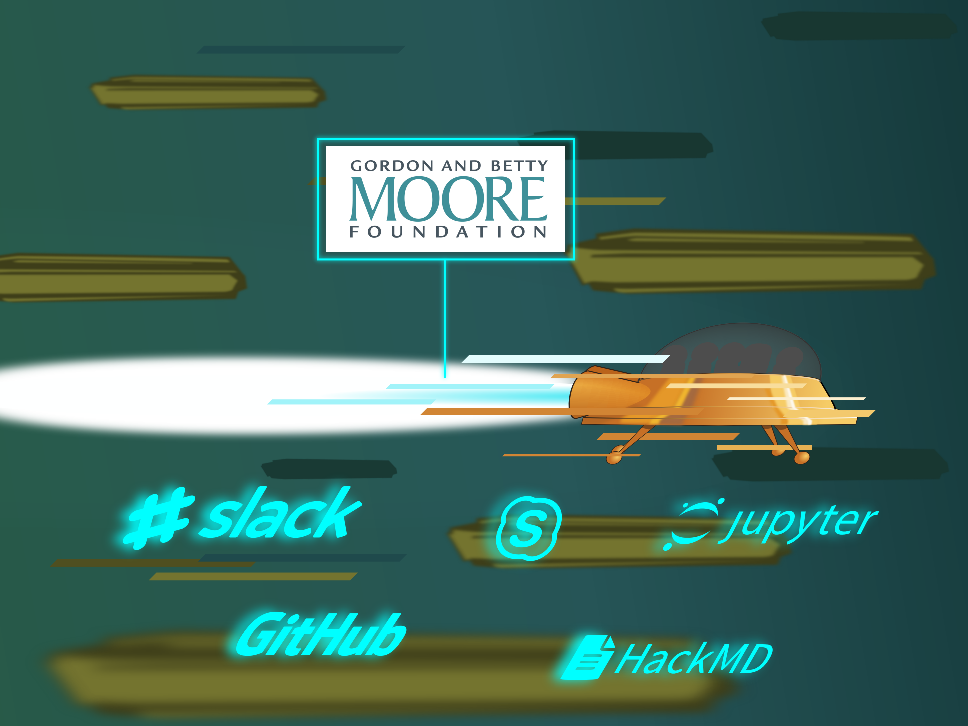
Genomic content

Domset

r-Domset

r-DTFAs

Dvořák's Algorithm


GORDON AND BETTY MOORE FOUNDATION
slack
jupyter
GitHub
HackMD

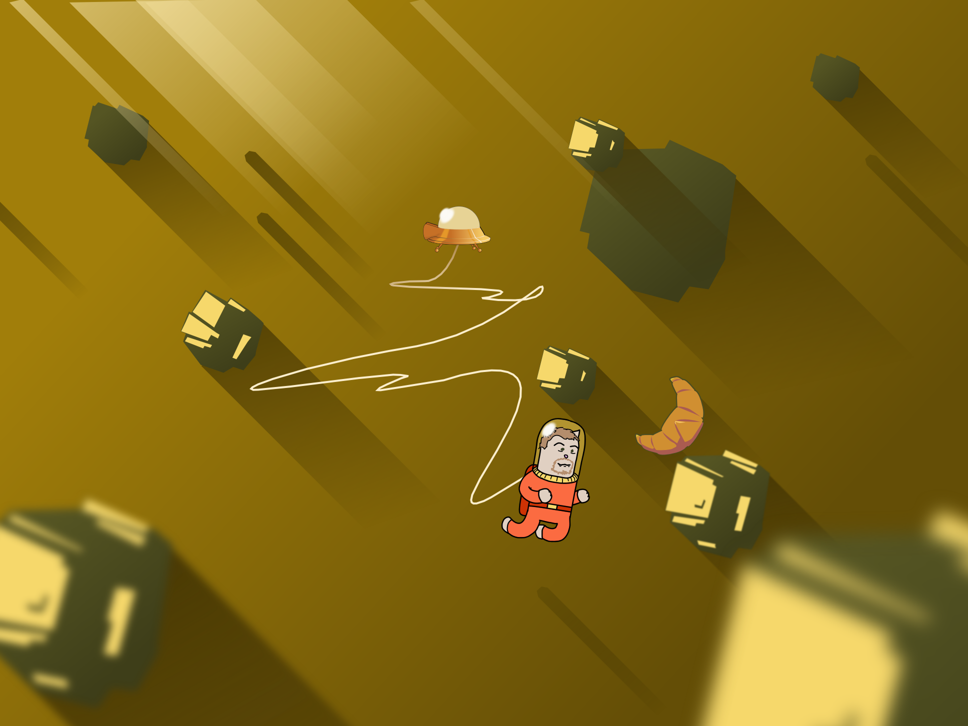
18 Months later.

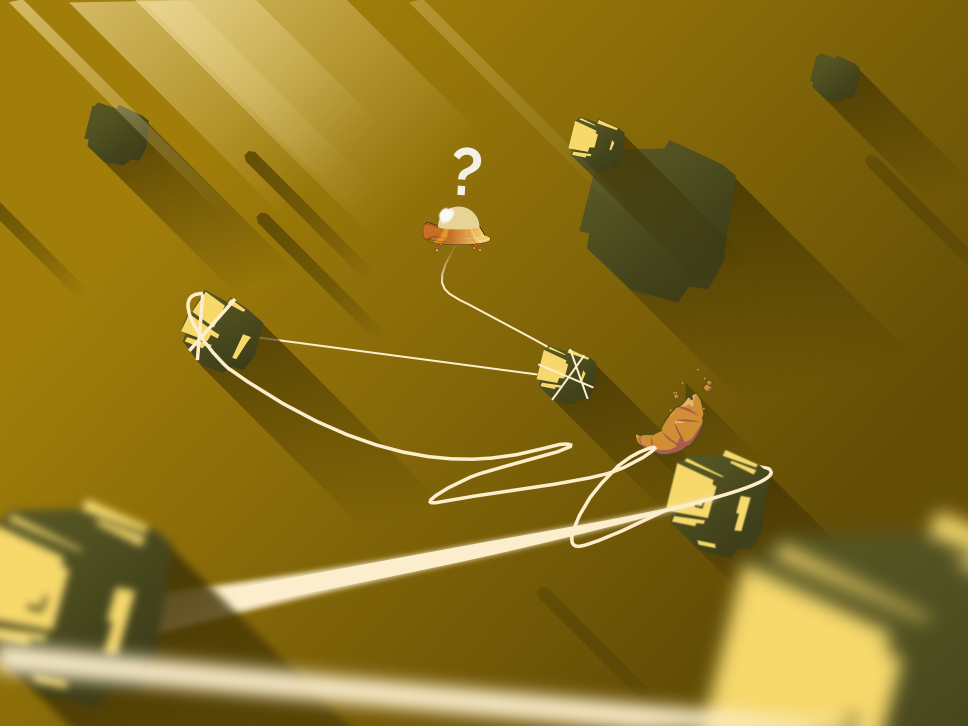
Translation
Method design
Course planning

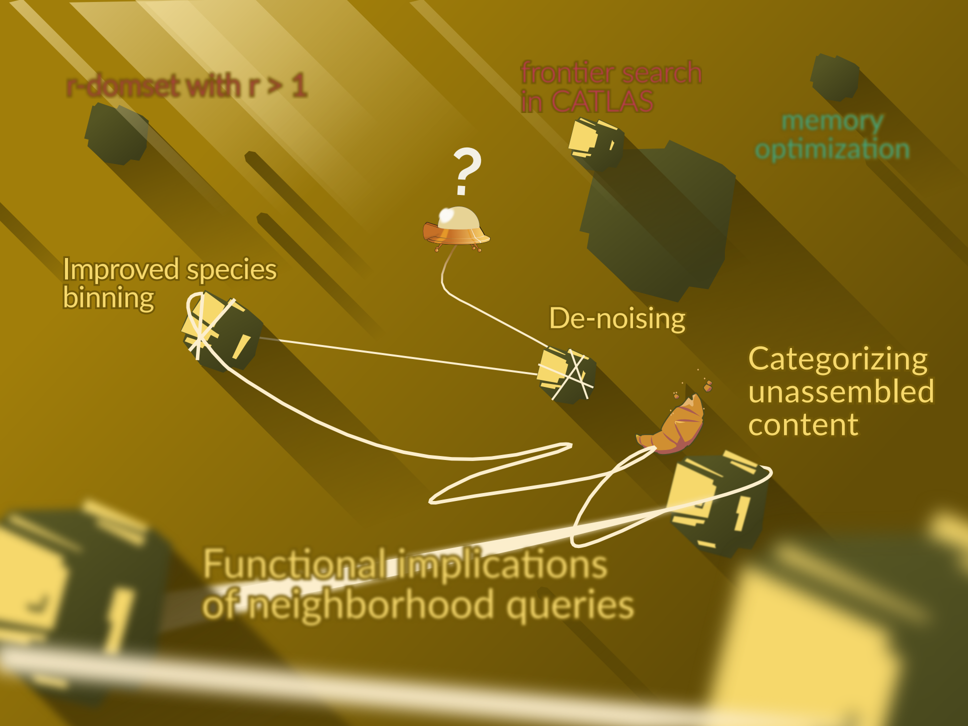
Data-wrangling
Target identification
Biological validation

Scalability
Feature development
Implementation

BIOLOGY!
WRONG!
BAD DATA!
BUG!
WRONG!
BUG!
BAD!
TOO SLOW!
BUG!
BUG!







r-domset with r > 1

frontier search in CATLAS

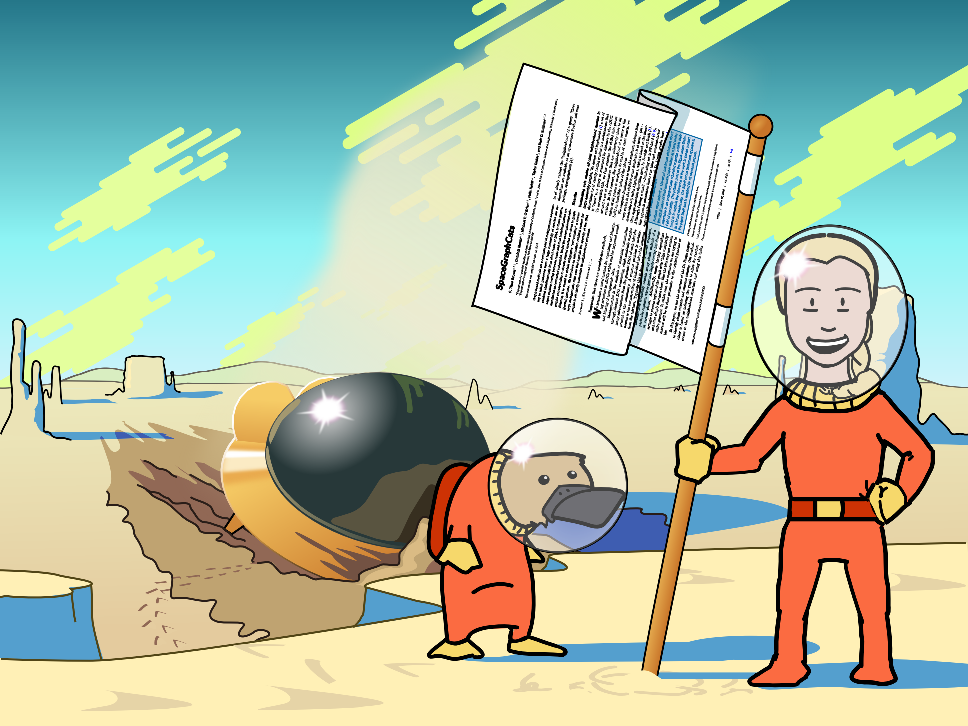
memory optimization

Improved species binning

De-noising

Categorizing unassembled content

Functional implications of neighborhood queries


SpaceGraphCats
C. Titus Brown, Dominik Moritz, Michael P. O'Brien, Felix Reidl, Taylor Reiter, and Blair D. Sullivan

PROJECT SPACEGRAPHCATS

Felix Reidl

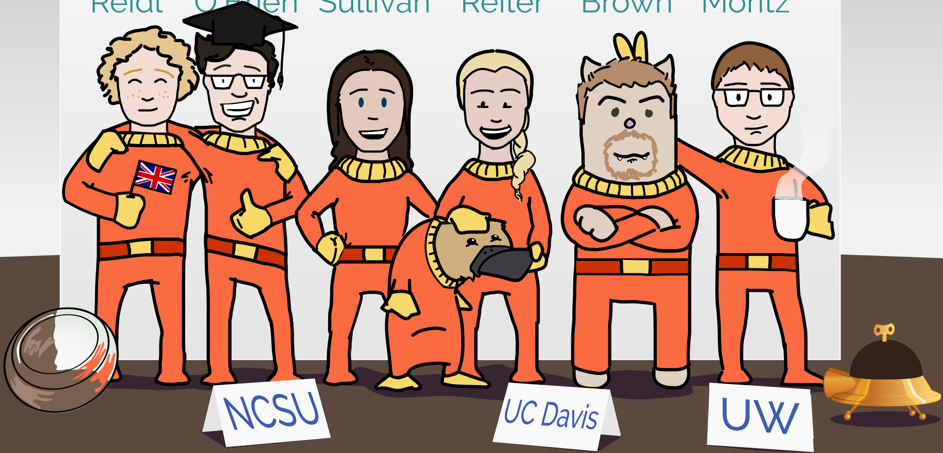
Mike O'Brien

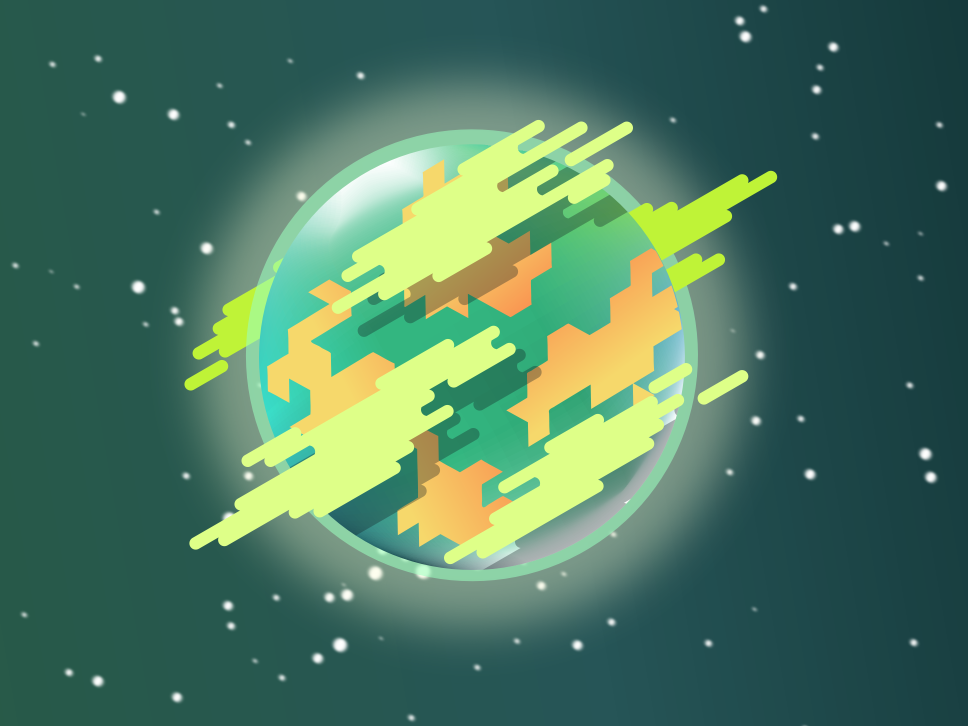
Blair Sullivan

Taylor Reiter

Titus Brown

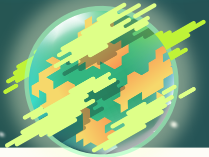
Dominik Moritz





> “The efficient computation of dominating sets will open up a whole new range of possibilities in bioinformatics.”

C. Titus Brown, Associate Professor at UC Davis

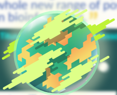


> “spacegraphcats will transform the way biologists interact with genome assemblies.
> It allows us to access previously discarded sequencing information thereby allowing more robust functional characterization.”

Taylor Reiter, his much more eloquent student

"The efficient computation of dominating sets will open up a whole new range of possibilities in bioi[illegible]"

C. Tit[illegible]essor at UC Davis

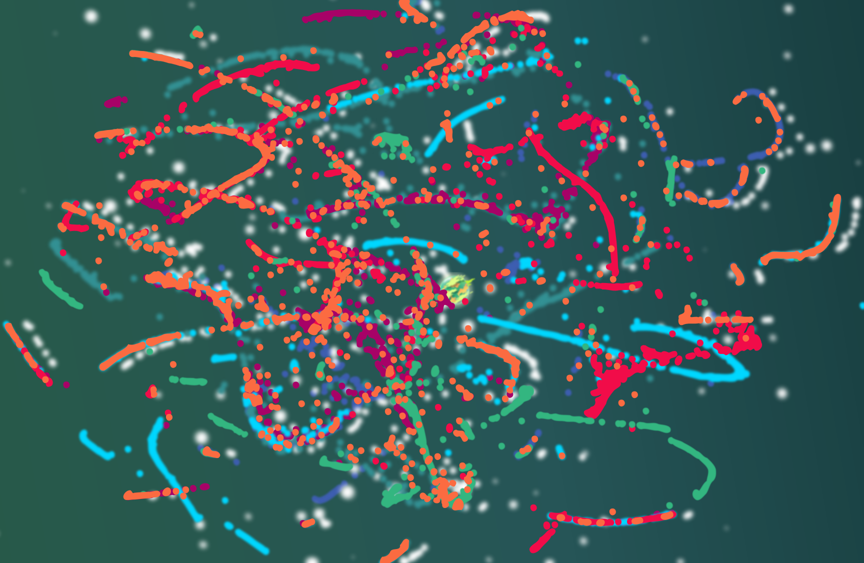


"spacegraphcats will transform the way biologists interact with genome assemblies.

It allows us to access previously discarded sequencing information thereby allowing more robust functional characterization."

Taylor Reiter, his much more eloquent student

SPACEGRAPHWORMS



Coming soon...