Database assisted state machine learning

Introducing

Hielke Walinga
Master thesis project

Algorithms group, TU Delft

Sicco Verwer, responsible professor
Robert Baumgartner, supervising PhD candidate

TOC

- Relevance and context
- Changing machine learning
- Learning state machines
- Algorithm
- Experiments
- Results
- Q&A

1.

Relevance and context

Reverse engineering software systems

Using state machines to understand software from logs

- State machines are a good model for software systems
- Complex software systems produce logs
- These logs can be used to infer how the software works
- These logs are often found in databases, such as Splunk

2.

Changing machine learning

Learning models from very big data

Too much data

- Data does not fit in memory
- Data does not fit on one computer
- Data is often very similar

Solutions for too much data

- Sample the data → Some information inevitably lost
- Batch the data → How to make batches?
- Stream the data → Cannot go back to previous data

Often multiple passes (epochs) of the data needed

Only need an informative sample

- Much data is often the same
- A much smaller subset is often enough
- Also known as a “characteristic sample” for state machines

Learning from a database

- Save your data to a database
- Ask relevant data from the database
- Data can be spread over multiple machines

Learning *state machines* from a database

- Fits already very well in this field:
  - The database is the system under learn
  - *Active learning*

- Depending on the indexing, allows for clever queries

- Log data might already be saved in a database (splunk)

3.

State machine learning

Learning from an incomplete teacher

L#: Partially building the state machine

- Maintaining current hypothesis and observations as a tree
- Making partial hypothesis by state merging
- Allows intuitive analysis halfway to guide the search

Frits Vaandrager, Bharat Garhewal, Jurriaan Rot, and Thorsten Wißmann. A New Approach for Active Automata Learning Based on Apartness (2022)

4.

DAALder algorithm

Database Assisted Automaton Learning

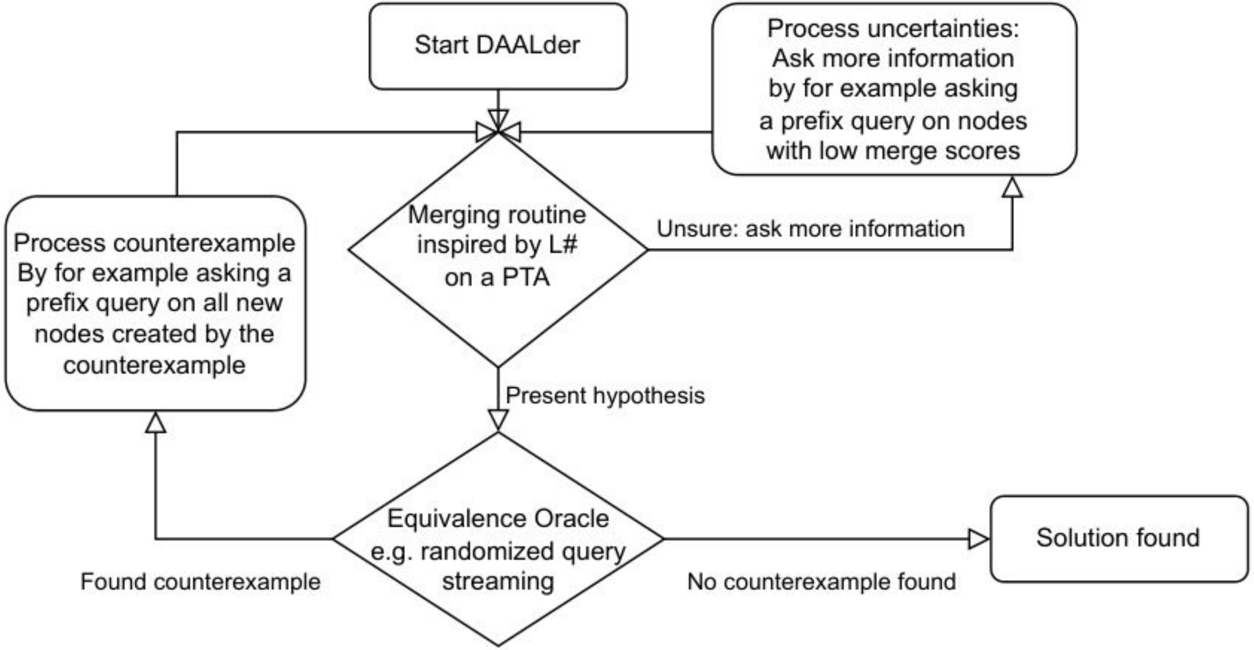
DAALder Algorithm

- Maintain a partial hypothesis as a tree

- Perform state merging

- If during state merging, more information is needed, ask
  - For example: *Prefix queries*

DAALder Algorithm


Start DAALder
Process uncertainties: Ask more information by for example asking a prefix query on nodes with low merge scores
Process counterexample By for example asking a prefix query on all new nodes created by the counterexample
Merging routine inspired by L# on a PTA
Unsure: ask more information
Present hypothesis
Equivalence Oracle e.g. randomized query streaming
Found counterexample
No counterexample found
Solution found

5.

Implementation details

FlexFringe, PostGreSQL

Implementation details

- Flexfringe:
  - State machine learning framework in C++
  - Easy access to many different merging routines
- PostGreSQL:
  - Mostly out of convenience
- SP-GiST indexing:
  - Very similar to a PTA

6.

Experiments and results

Experiments

- Randomized state machines
- Data size doubled each test from 625 to 40960000

- Random sampling:
  - Uniform
  - Non-uniform

- Compared with:
  - EDSM
  - iMAT

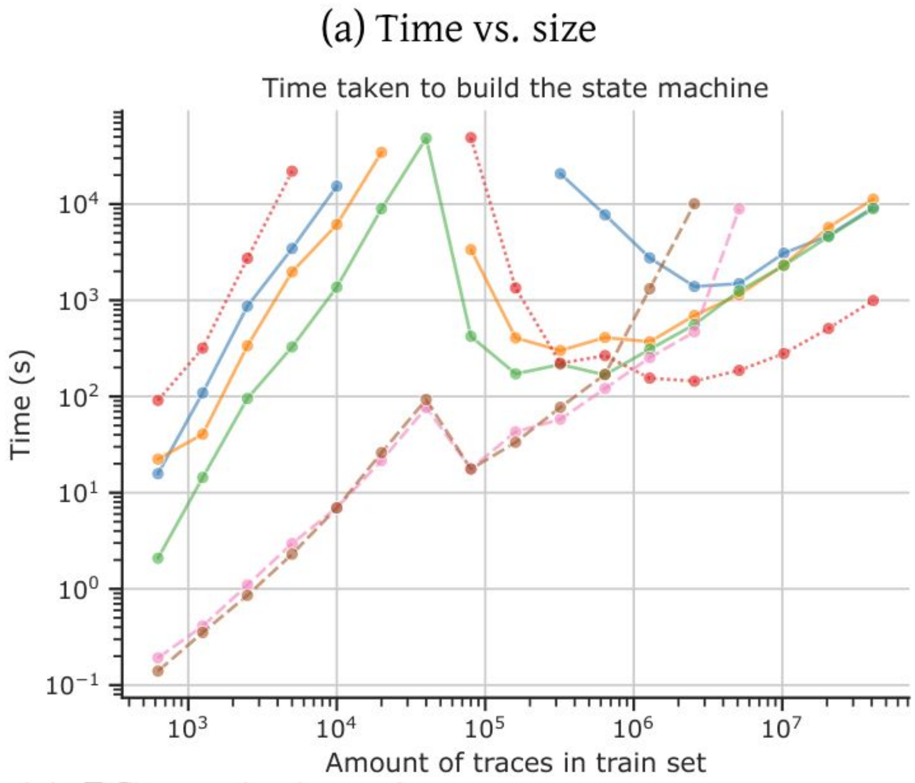
https://github.com/hwalinga/FSM-learning

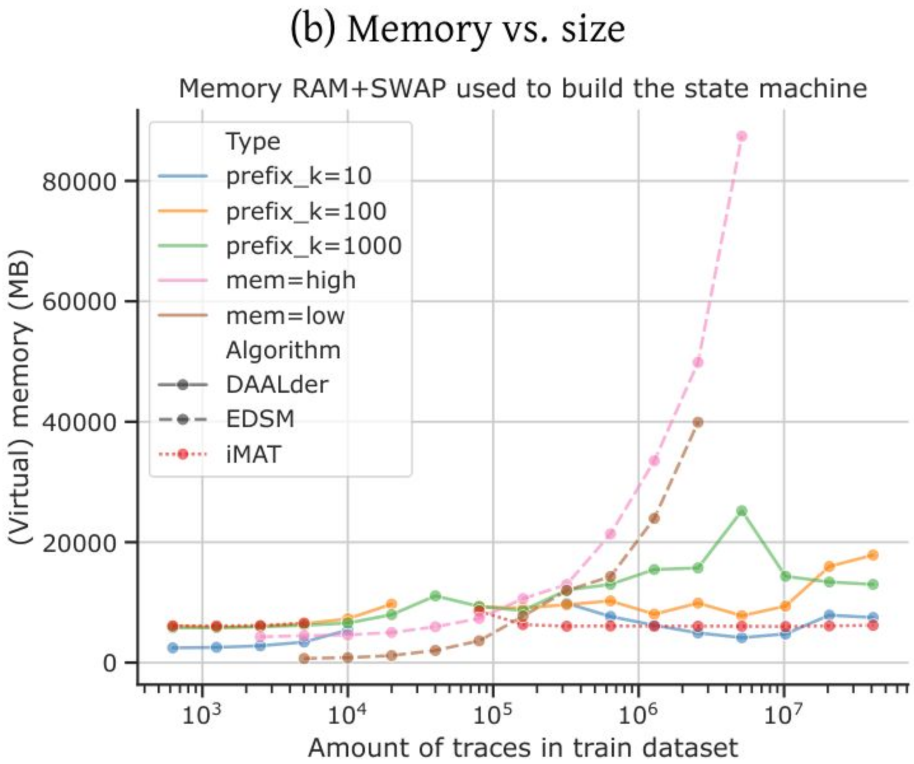
Georgios Giantamidis, Stavros Tripakis, and Stylianos Basagiannis. Learning Moore machines from input–output traces. (2021)

Olga Grinchtein, Martin Leucker, and Nir Piterman. Inferring Network Invariants Automatically (2006)

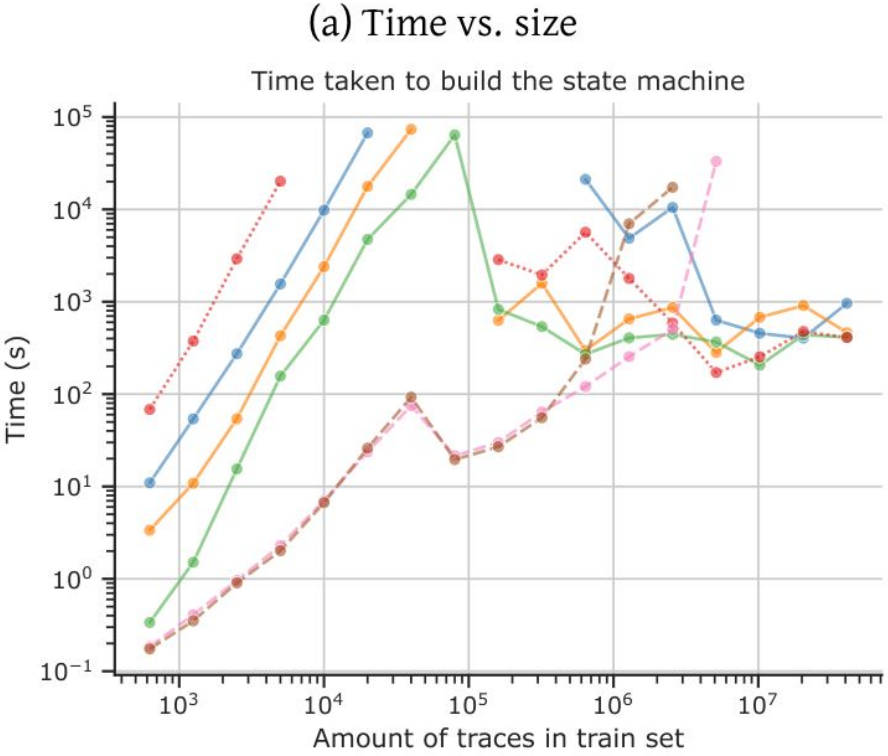
Mark Moeller et al. Automata Learning with an Incomplete Teacher. (2023)

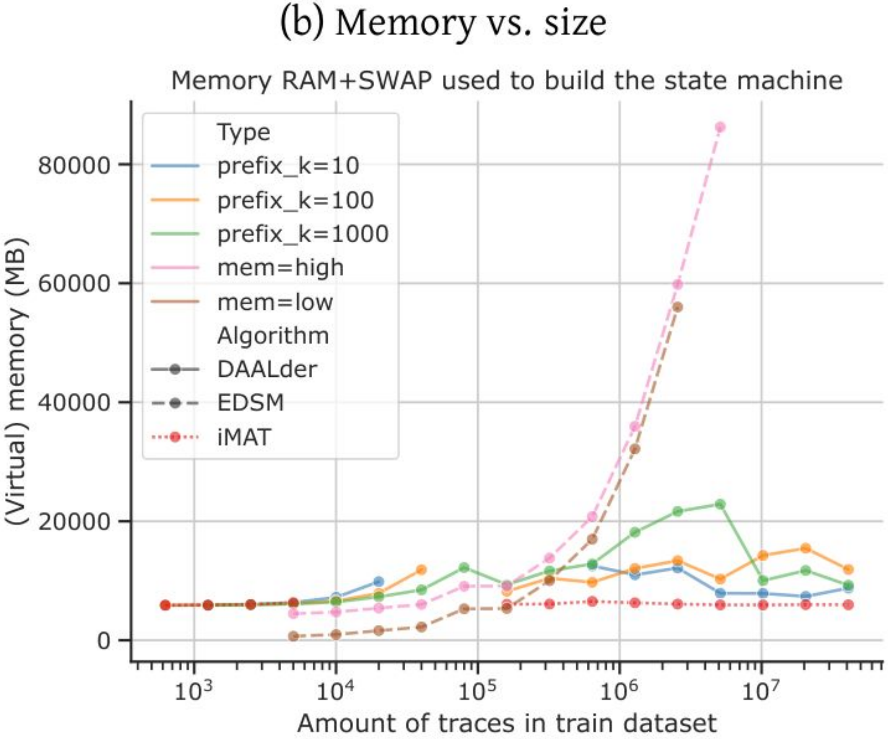
Measuring performance: Uniform data


(a) Time vs. size


(b) Memory vs. size

Measuring performance: Non-uniform data





7.

Final remarks

Discussion

- DAALder only works well for large datasets
- DAALder seems more useful with more sparsity in the data
- *I expect that iMAT performs worse with bigger alphabet*

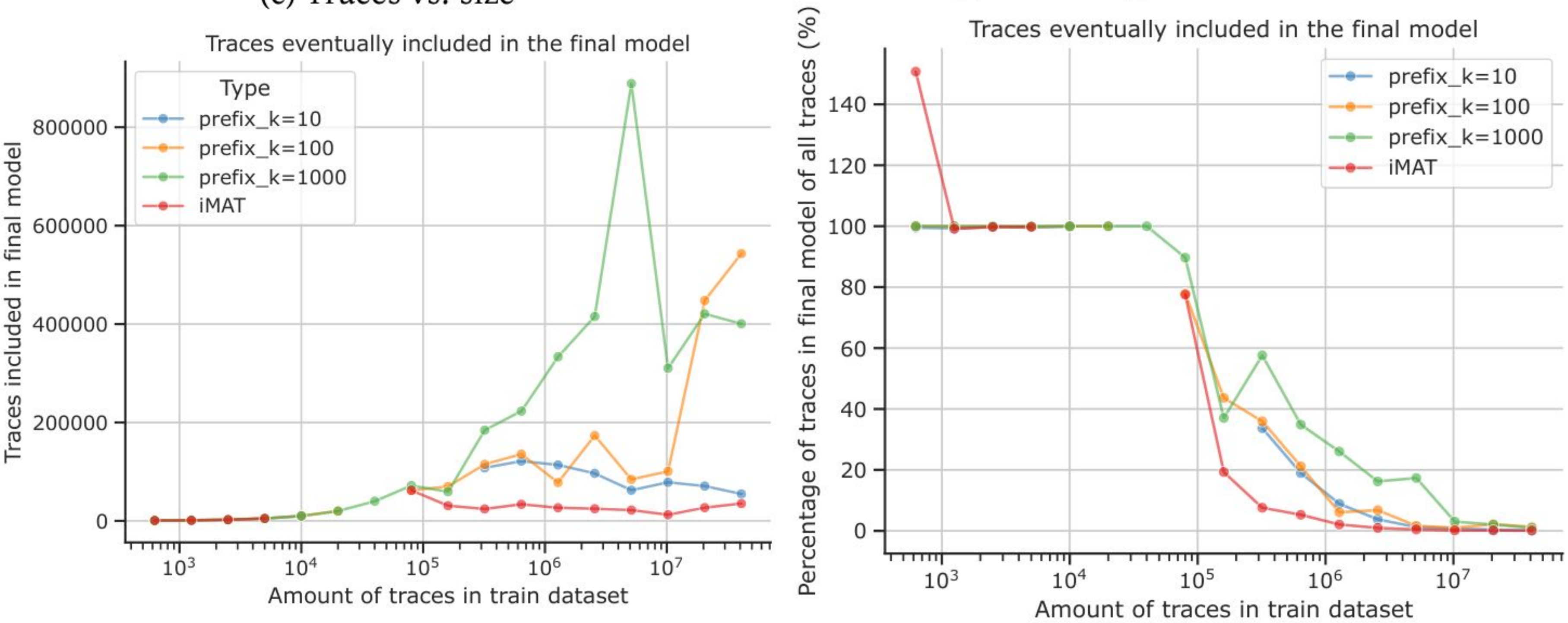
Future work

- Improvements:
  - Different queries/indexes
  - Better heuristics on guidance what to ask
- More future work:
  - Incorporate more information sources
  - Learn the most informative sources
  - Learn a strategy

Q&A

Implementation available
(branch: Publlications/learnaut24):
https://github.com/tudelft-cda-lab/FlexFringe
Thesis: https://hielkewalinga.nl/uploads/thesis.pdf
Questions:
hielkewalinga@gmail.com
S.E.Verwer@tudelft.nl

Measuring performance: Uniform data

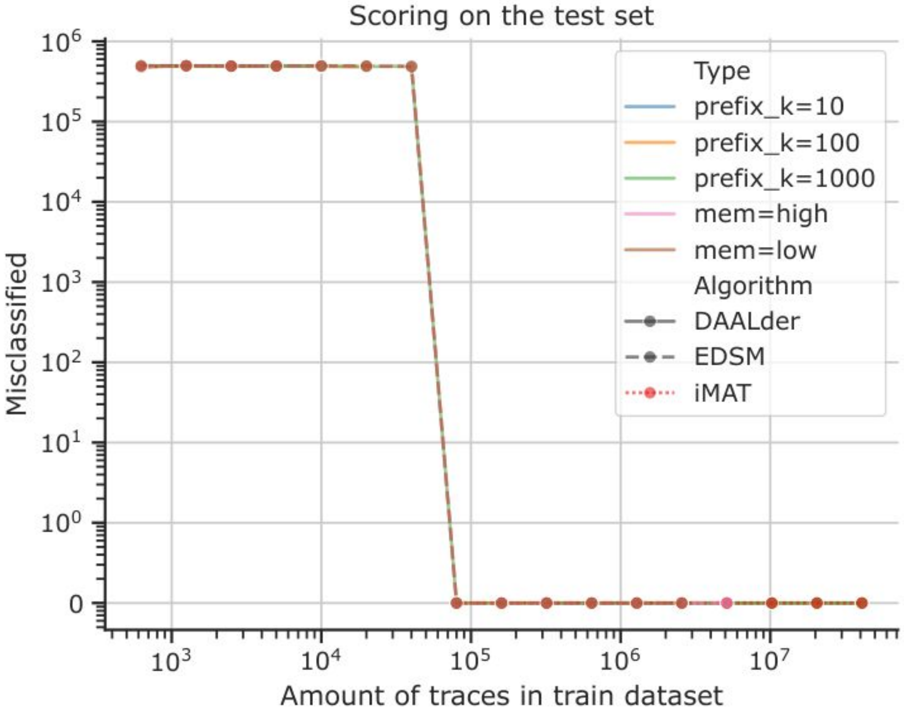


(c) Traces vs. size

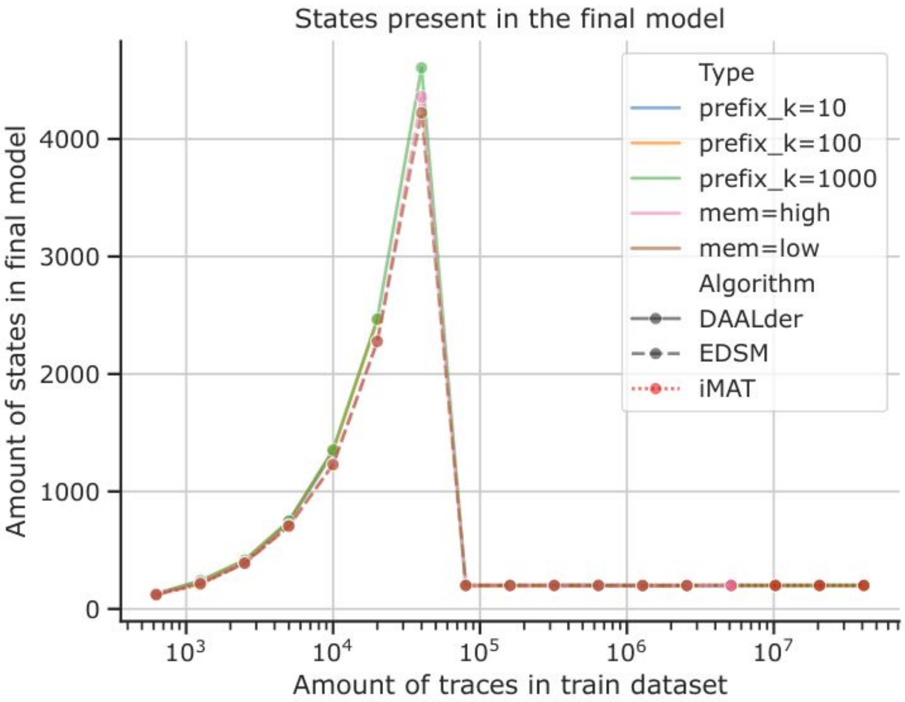
(d) Percentage of traces vs. size

Measuring performance: Uniform data

(e) Accuracy vs. size

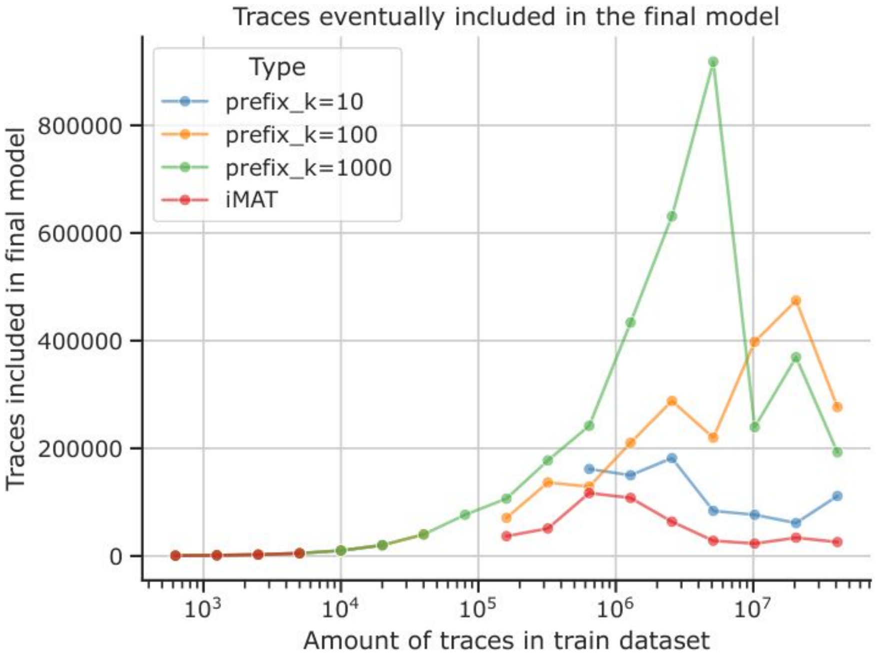


(f) States vs. size

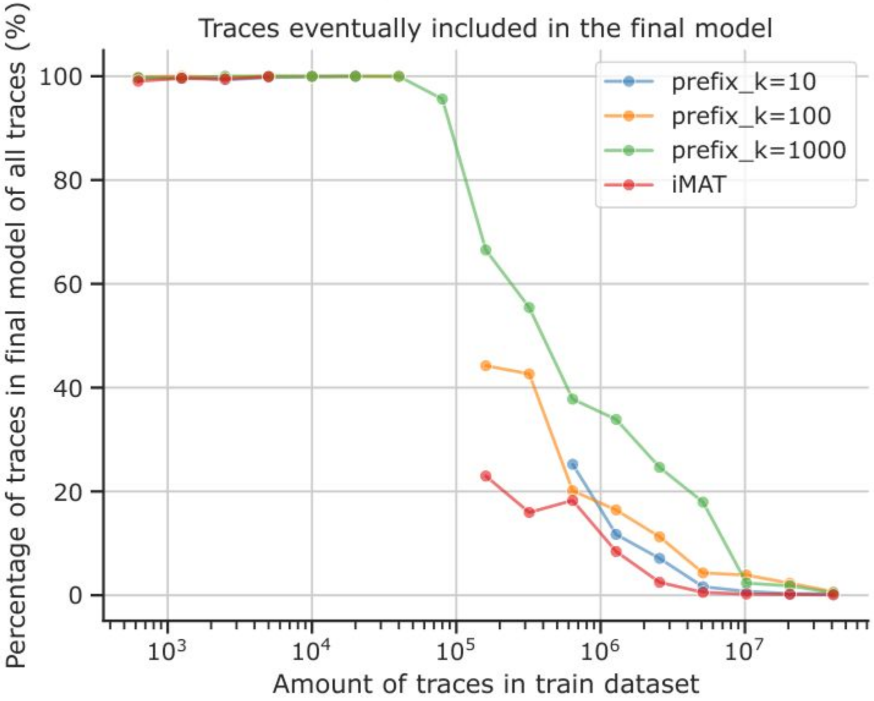


Measuring performance: Non-uniform data

(c) Traces vs. size

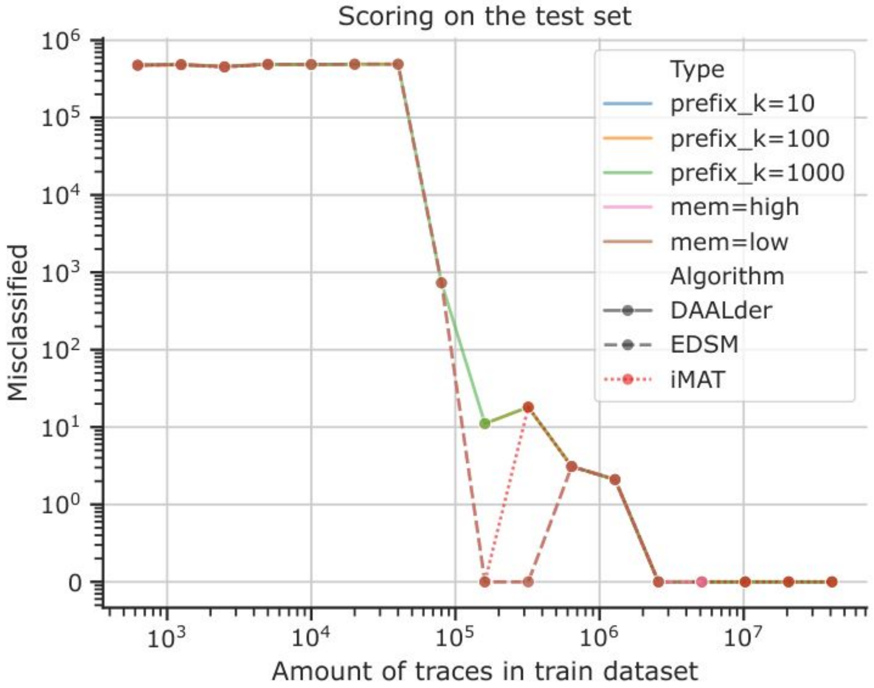


(d) Percentage of traces vs. size

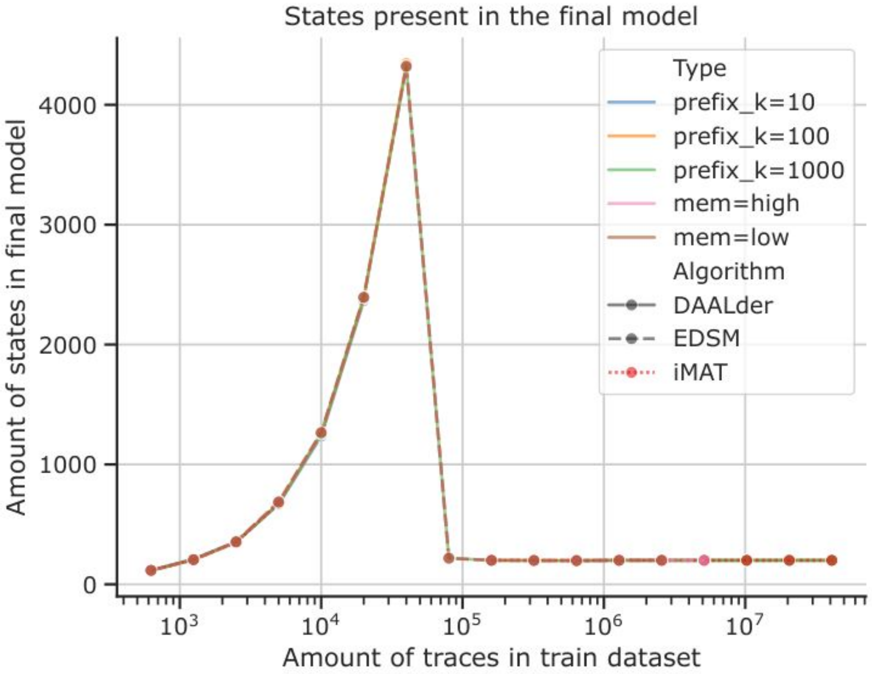


Measuring performance: Non-uniform data

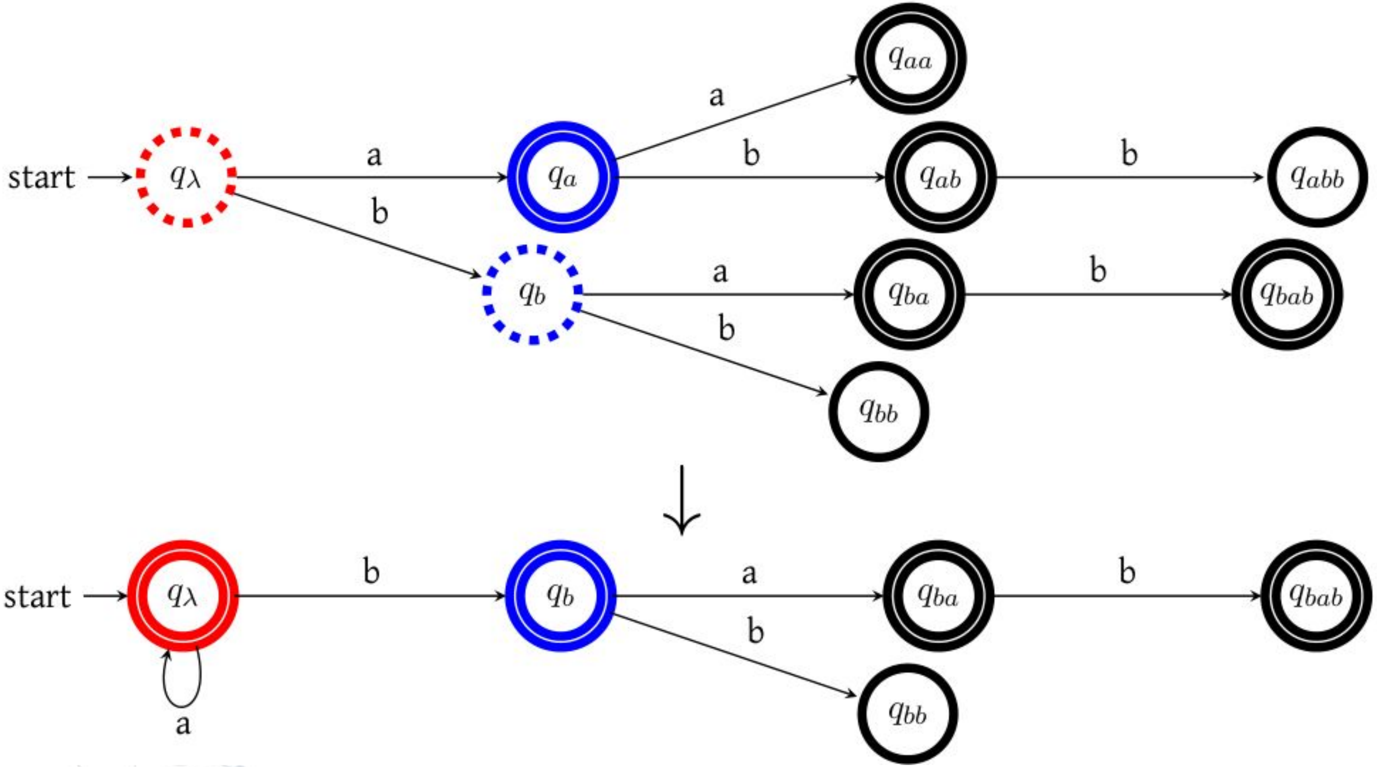
(e) Accuracy vs. size



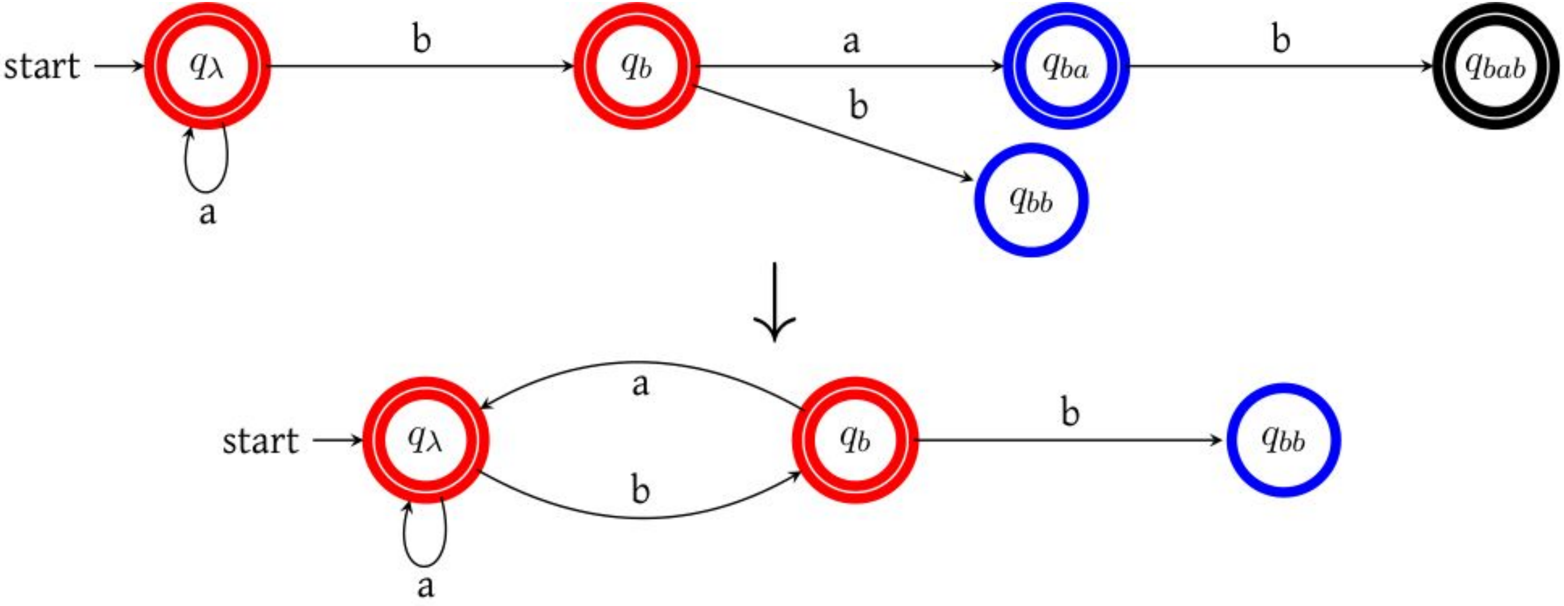
(f) States vs. size



Concrete example



Concrete example



Software models

- Input → Output
- Analyzable
- Understandable
- Useful reduction

2.

State machines

A simple model of computation

State machines

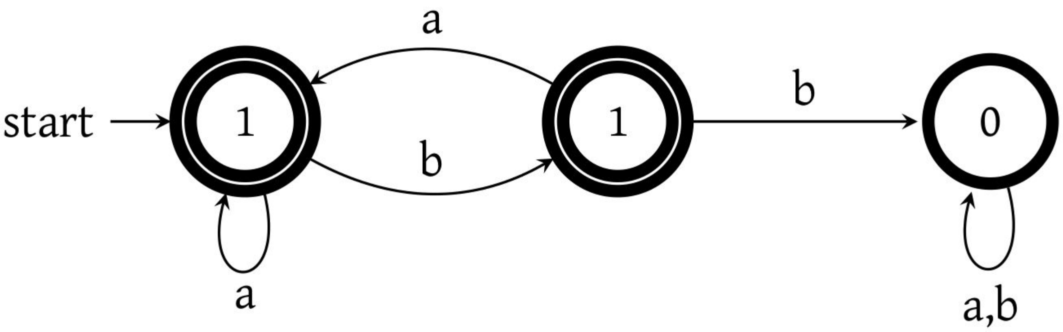
- Directed graph
- Nodes and edges
- Traverse different paths
- Last node gives the output symbol

abba -> 0

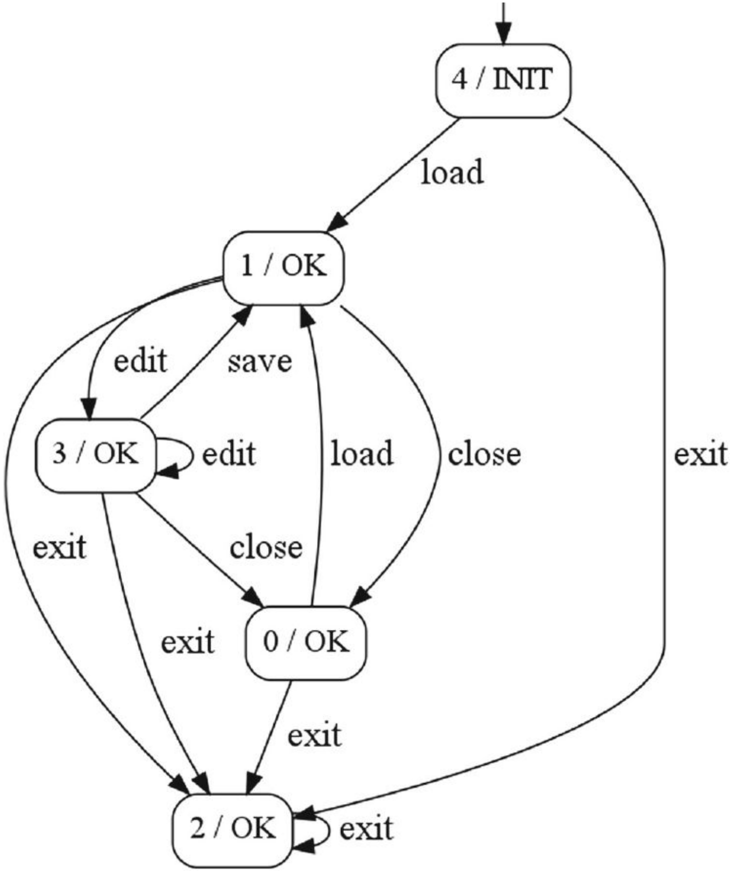
bab -> 1

babb -> 0

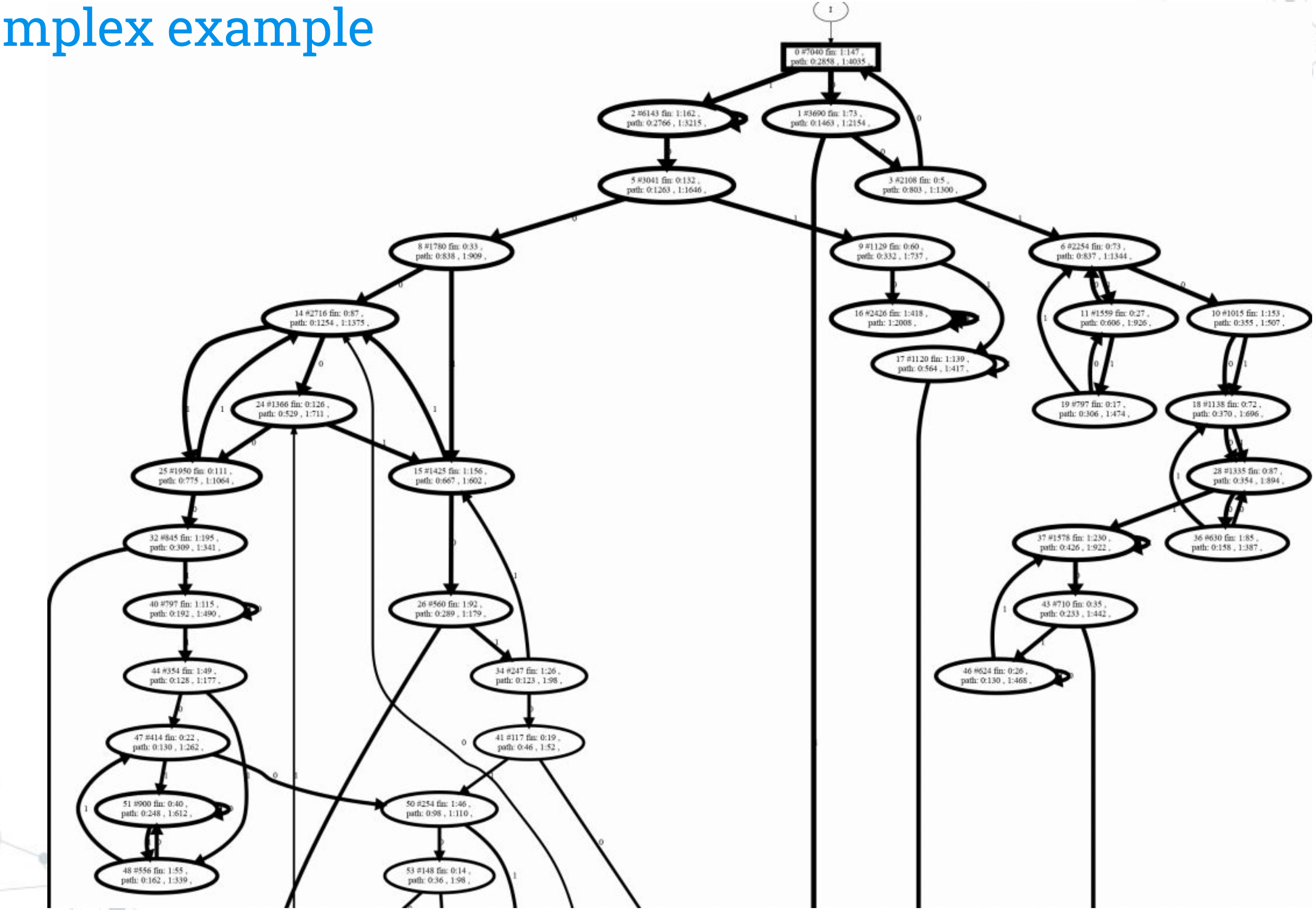
ab -> 1



A simple editing program

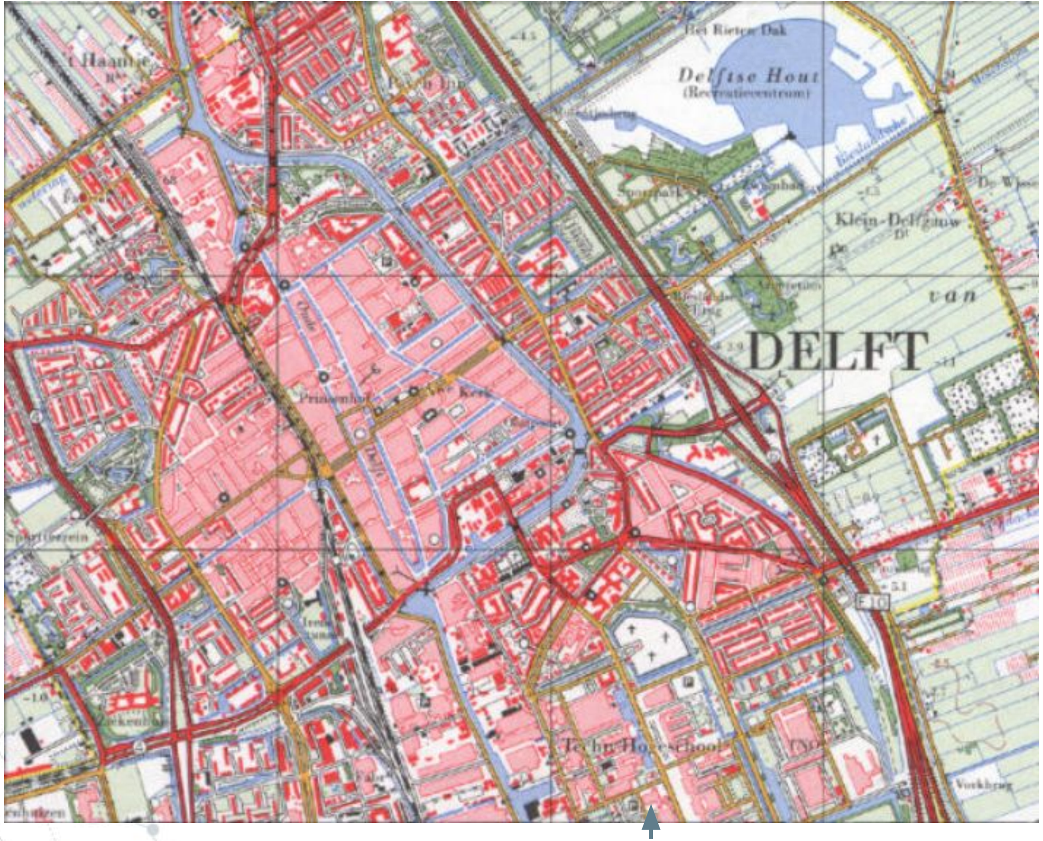


More complex example



1
0 #7040 fin: 1:147 , path: 0:2858 , 1:4035 ,
2 #6143 fin: 1:162 , path: 0:2766 , 1:3215 ,
1 #3690 fin: 1:73 , path: 0:1463 , 1:2154 ,
5 #3041 fin: 0:132 , path: 0:1263 , 1:1646 ,
3 #2108 fin: 0:5 , path: 0:803 , 1:1300 ,
8 #1780 fin: 0:33 , path: 0:838 , 1:909 ,
9 #1129 fin: 0:60 , path: 0:332 , 1:737 ,
6 #2254 fin: 0:73 , path: 0:837 , 1:1344 ,
14 #2716 fin: 0:87 , path: 0:1254 , 1:1375 ,
16 #2426 fin: 1:418 , path: 1:2008 ,
11 #1559 fin: 0:27 , path: 0:606 , 1:926 ,
10 #1015 fin: 1:153 , path: 0:355 , 1:507 ,
17 #1120 fin: 1:139 , path: 0:564 , 1:417 ,
24 #1366 fin: 0:126 , path: 0:529 , 1:711 ,
19 #797 fin: 0:17 , path: 0:306 , 1:474 ,
18 #1138 fin: 0:72 , path: 0:370 , 1:696 ,
25 #1950 fin: 0:111 , path: 0:775 , 1:1064 ,
15 #1425 fin: 1:156 , path: 0:667 , 1:602 ,
28 #1335 fin: 0:87 , path: 0:354 , 1:894 ,
32 #845 fin: 1:195 , path: 0:309 , 1:341 ,
37 #1578 fin: 1:230 , path: 0:426 , 1:922 ,
36 #630 fin: 1:85 , path: 0:158 , 1:387 ,
40 #797 fin: 1:115 , path: 0:192 , 1:490 ,
26 #560 fin: 1:92 , path: 0:289 , 1:179 ,
43 #710 fin: 0:35 , path: 0:233 , 1:442 ,
44 #354 fin: 1:49 , path: 0:128 , 1:177 ,
34 #247 fin: 1:26 , path: 0:123 , 1:98 ,
46 #624 fin: 0:26 , path: 0:130 , 1:468 ,
47 #414 fin: 0:22 , path: 0:130 , 1:262 ,
41 #117 fin: 0:19 , path: 0:46 , 1:52 ,
51 #900 fin: 0:40 , path: 0:248 , 1:612 ,
50 #254 fin: 1:46 , path: 0:98 , 1:110 ,
48 #556 fin: 1:55 , path: 0:162 , 1:339 ,
53 #148 fin: 0:14 , path: 0:36 , 1:98 ,

A state machine is like a map



start

2.

State machine learning

From software system to model

State machine learning

- Active learning
  - *Actively* probing a software system to find the model
  - Needs the system present
- Passive learning
  - Learn the state machine from a collection of input-output
  - Learning from *log-data*
  - Requires a lot of data

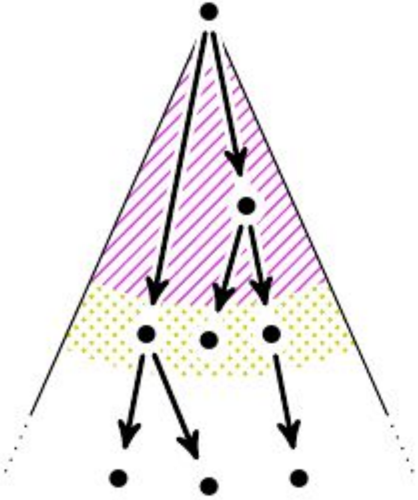
Back to the map analogy

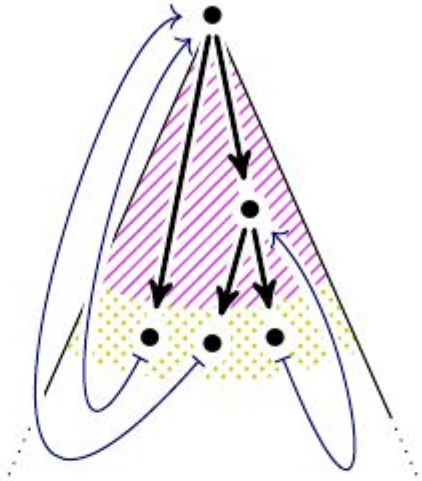
- Active learning
  - Sending out people one by one
- Passive learning
  - Asking X people what they have seen

How does it work: State merging

- “Places” with exactly the same future paths are likely the same
- Merging these iteratively creates the final state machine
- Can use either tables are trees to hold this information

A sketch of the idea





Problem of passive learning: too much data

- Log-data sets can be very big
- Conventional state merging algorithms are not sufficient
- Current solution: Don’t use all data

3.

Databases

When you have too much data

Some computer infrastructure analogies

- RAM memory
  - Your workbench of data
  - A giant blackboard
- Disk
  - Lots of cabinets
  - Every cabinet is a “disk page”
- Database
  - Cabinets but ordered

4.

Research question

How can we learn a state machine from a large set of data using a database?

Solution: Combine active and passive learning

- Save data to a database
- Use *active* learning techniques to extract data
- Design mechanisms to quickly answer questions by database

- **Problem:** Active learning assumes complete information
- Thus, use *passive* learning techniques to learn state machines from this extracted data.

DAALder

First hit is a DAALder

DAALder: Database-Assisted Automaton Learner

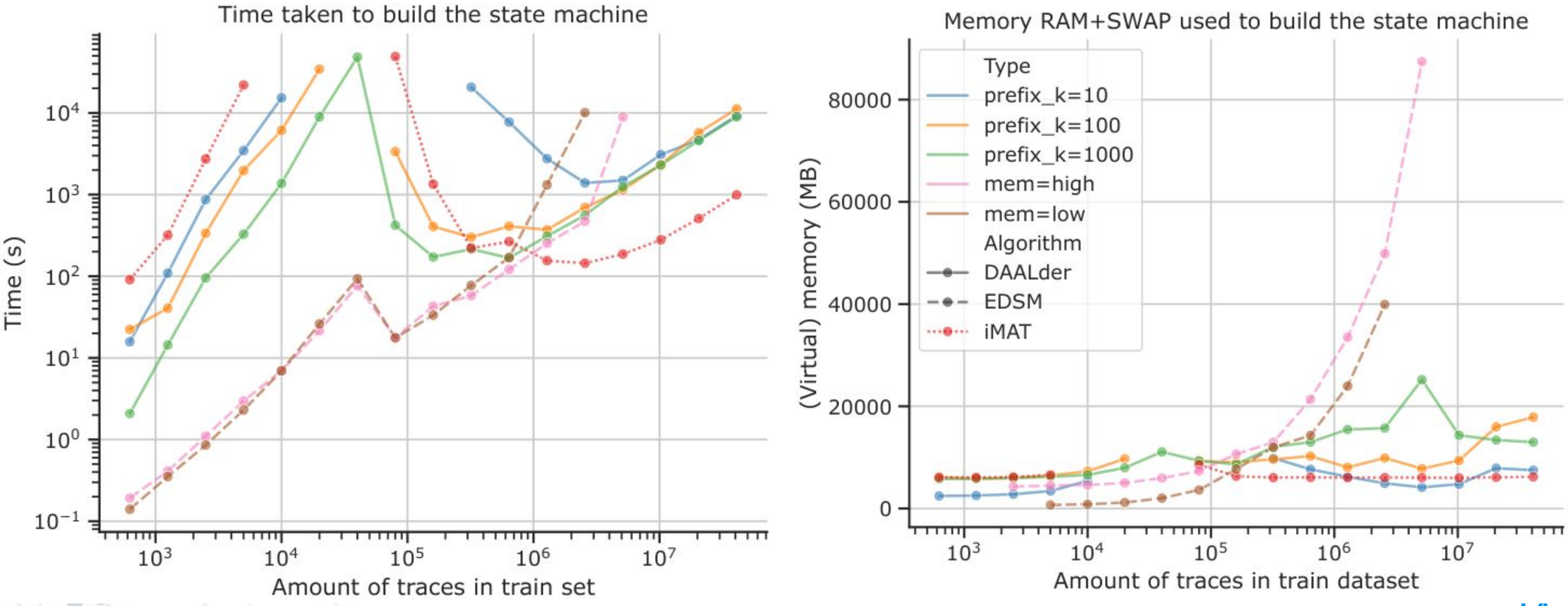
- Ask data from database and save in tree
- Perform state merging

- If not enough information → ask for more information
- Uses state merging heuristics to ask for what

Measuring performance

- Learning on artificial data
  - Input and output alphabet of size 2
  - Uniform
  - Non-uniform
- Different algorithms
  - Conventional passive learning: EDSM
  - Slightly modified active learning: iMAT
  - DAALder with different hyperparameters for exploration

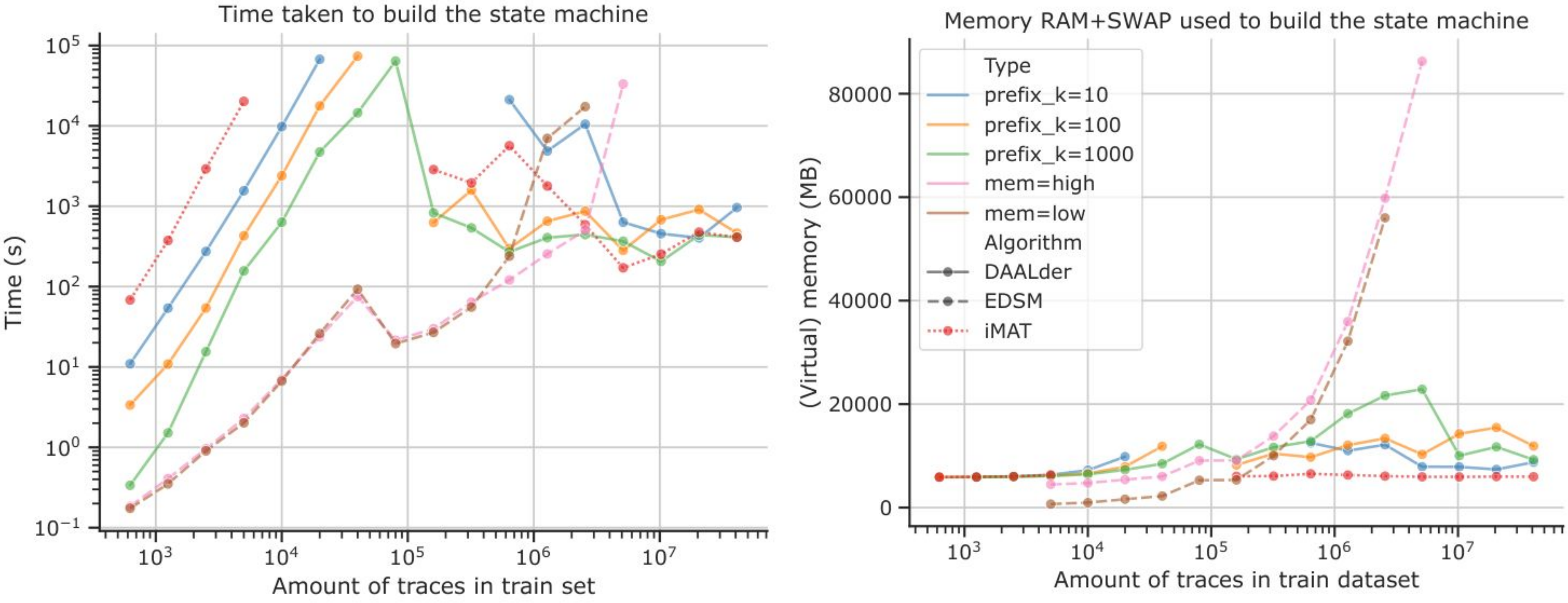
Measuring performance: Uniform data



(a) Time vs. size

(b) Memory vs. size

Measuring performance: Non-uniform data



(a) Time vs. size

(b) Memory vs. size

Discussion

- DAALder only works well for large datasets
- DAALder seems more useful when there is more sparsity in the data

Conclusion and future work

- More research is needed for better heuristics and performance on different datasets
- More future work:
  - Incorporate more information
  - How do we exactly learn from bigger datasets
  - What information to include

Questions?

Instructions for use

EDIT IN GOOGLE SLIDES

Click on the button under the presentation preview that says "Use as Google Slides Theme".

You will get a copy of this document on your Google Drive and will be able to edit, add or delete slides.

You have to be signed in to your Google account.

EDIT IN POWERPOINT®

Click on the button under the presentation preview that says "Download as PowerPoint template". You will get a .pptx file that you can edit in PowerPoint.

Remember to download and install the fonts used in this presentation (you’ll find the links to the font files needed in the Presentation design slide)

More info on how to use this template at www.slidescarnival.com/help-use-presentation-template

Hello!

I am Jayden Smith

I am here because I love to give presentations.

You can find me at:

@username

1.

Transition headline

Let’s start with the first set of slides

*Quotations are commonly printed as a **means of inspiration** and to invoke philosophical thoughts from the reader.*

This is a slide title

- Here you have a list of items
- And some text
- But remember not to overload your slides with content

Your audience will listen to you or read the content, but won’t do both.

You can also split your content

White

Is the color of milk and fresh snow, the color produced by the combination of all the colors of the visible spectrum.

Black

Is the color of ebony and of outer space. It has been the symbolic color of elegance, solemnity and authority.

In two or three columns

Yellow

Is the color of gold, butter and ripe lemons. In the spectrum of visible light, yellow is found between green and orange.

Blue

Is the colour of the clear sky and the deep sea. It is located between violet and green on the optical spectrum.

Red

Is the color of blood, and because of this it has historically been associated with sacrifice, danger and courage.

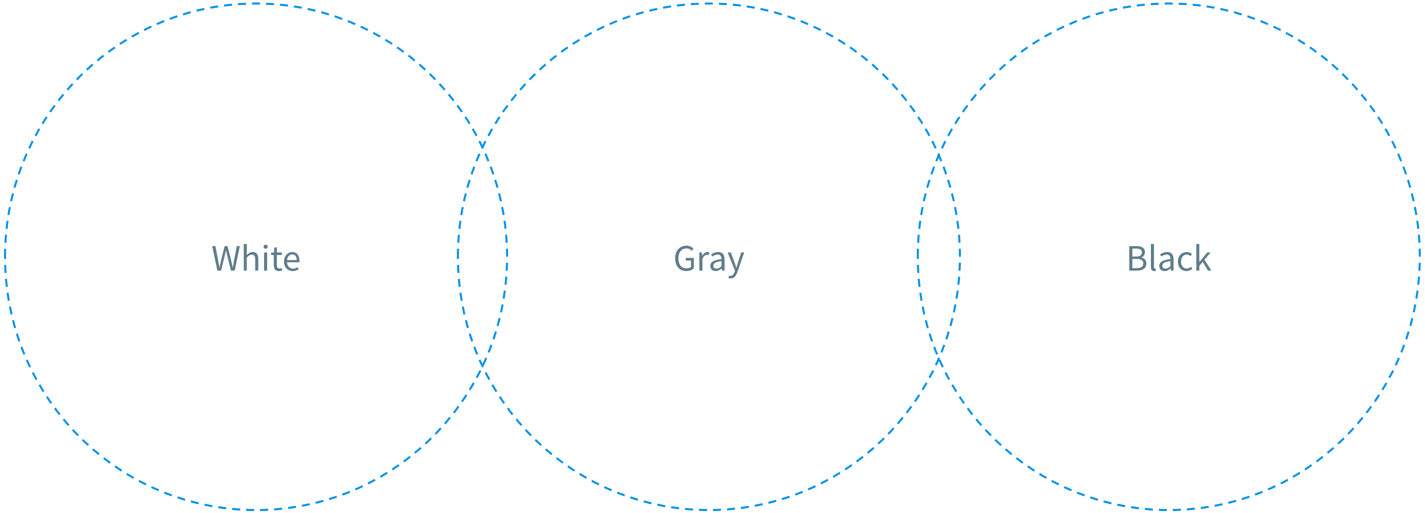
A picture is worth a thousand words

A complex idea can be conveyed with just a single still image, namely making it possible to absorb large amounts of data quickly.

Want big impact?

Use big image.

Use charts to explain your ideas



White

Gray

Black

Or diagrams to explain complex ideas

Example text.

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Nam venenatis nisi at nisl tempor, et luctus diam lobortis. Nulla sit amet metus consequat velit iaculis tempor.

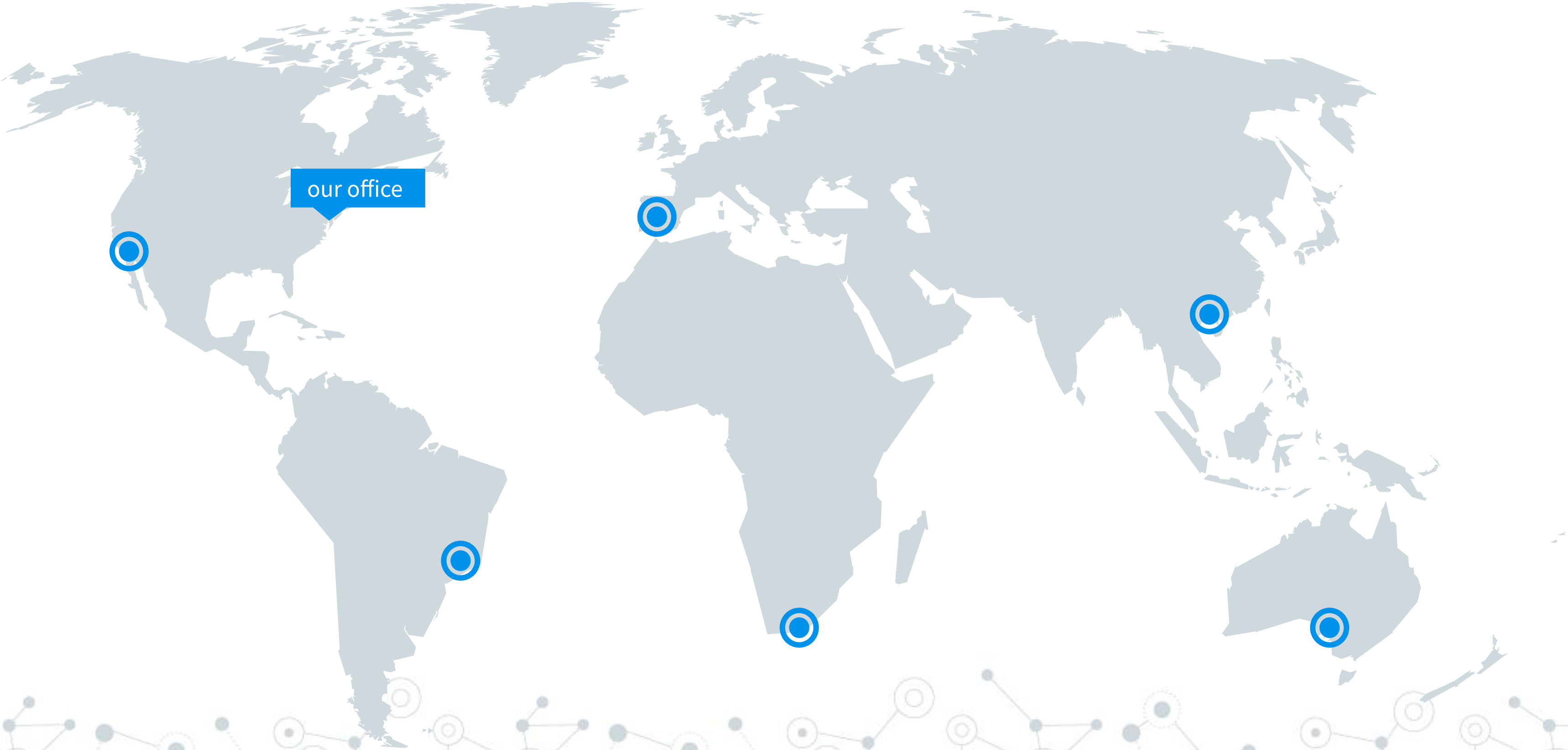
Example text.

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Nam venenatis nisi at nisl tempor, et luctus diam lobortis. Nulla sit amet metus consequat velit iaculis tempor.

And tables to compare data

	A	B	C
Yellow	**10**	**20**	**7**
Blue	**30**	**15**	**10**
Orange	**5**	**24**	**16**

Maps



89,526,124

Whoa! That’s a big number, aren’t you proud?

Presentation design

This presentations uses the following typographies and colors:

- Titles: **Roboto Slab**
- Body copy: **Source Sans Pro**

Download for free at:

https://www.fontsquirrel.com/fonts/roboto-slab

https://www.fontsquirrel.com/fonts/source-sans-pro

You don't need to keep this slide in your presentation. It's only here to serve you as a design guide if you need to create new slides or download the fonts to edit the presentation in PowerPoint®

89,526,124$

That’s a lot of money

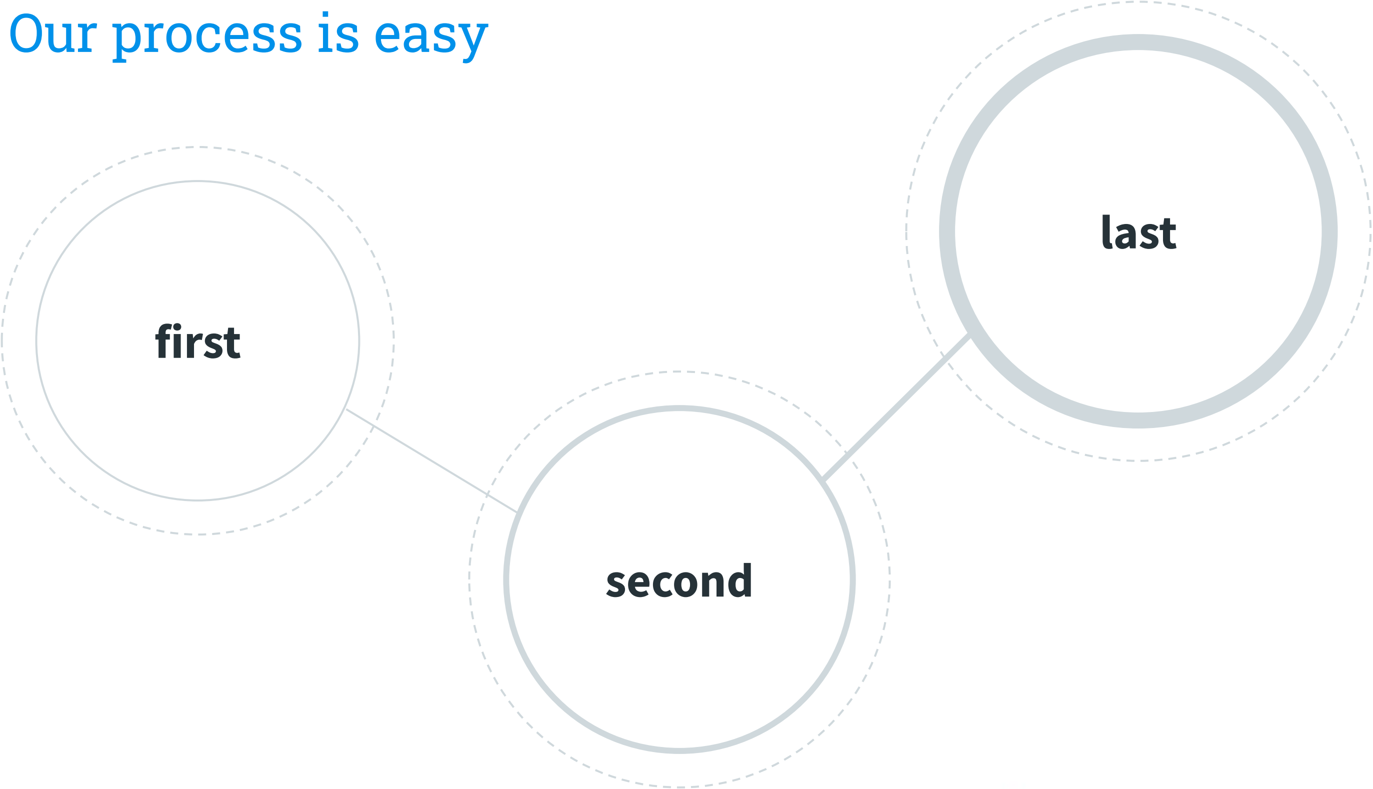
185,244 users

And a lot of users

100%

Total success!

Our process is easy



Let’s review some concepts

Yellow

Is the color of gold, butter and ripe lemons. In the spectrum of visible light, yellow is found between green and orange.

Blue

Is the colour of the clear sky and the deep sea. It is located between violet and green on the optical spectrum.

Red

Is the color of blood, and because of this it has historically been associated with sacrifice, danger and courage.

Yellow

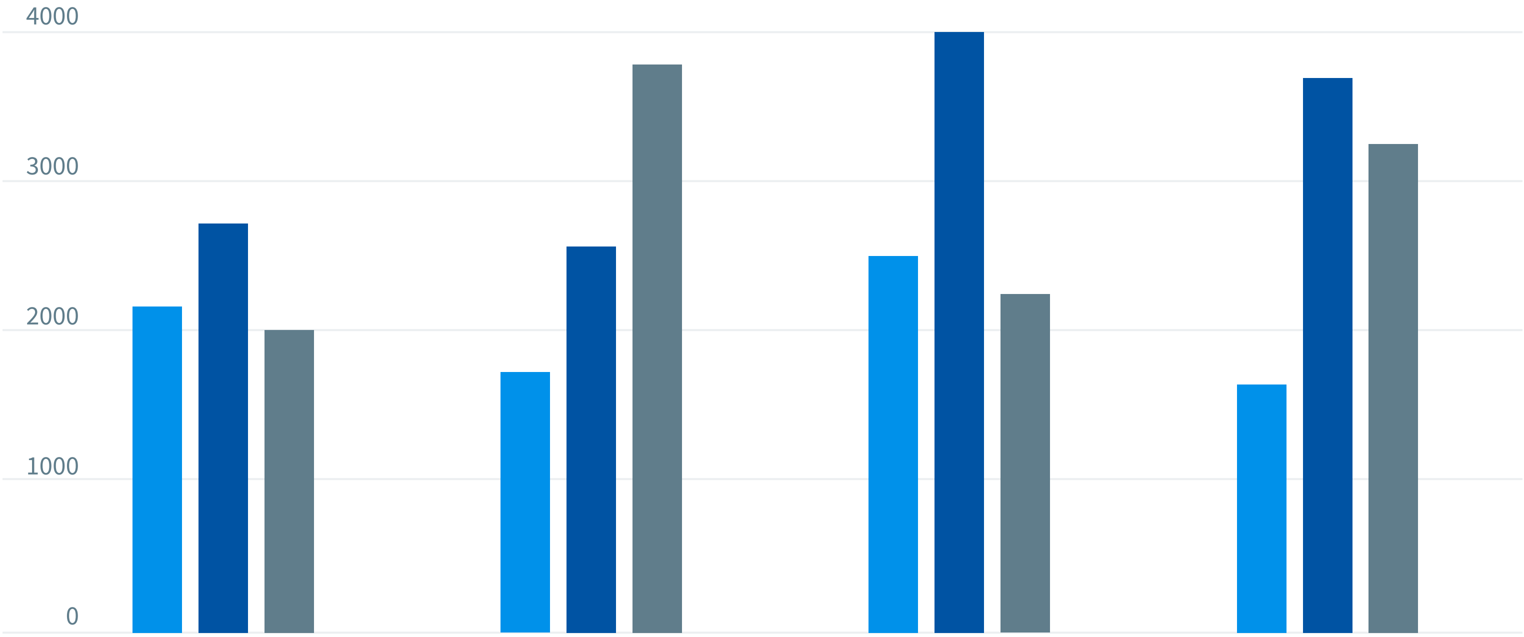
Is the color of gold, butter and ripe lemons. In the spectrum of visible light, yellow is found between green and orange.

Blue

Is the colour of the clear sky and the deep sea. It is located between violet and green on the optical spectrum.

Red

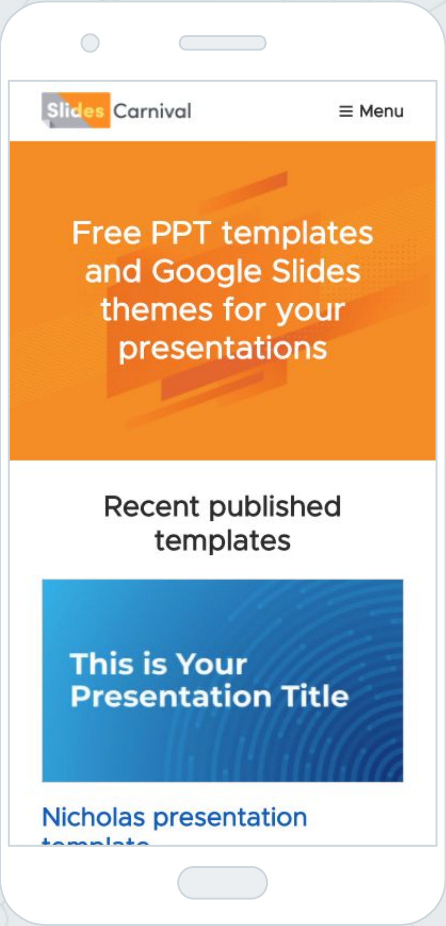
Is the color of blood, and because of this it has historically been associated with sacrifice, danger and courage.



You can insert graphs from Excel or Google Sheets

Mobile project

Show and explain your web, app or software projects using these gadget templates.



Free PPT templates and Google Slides themes for your presentations

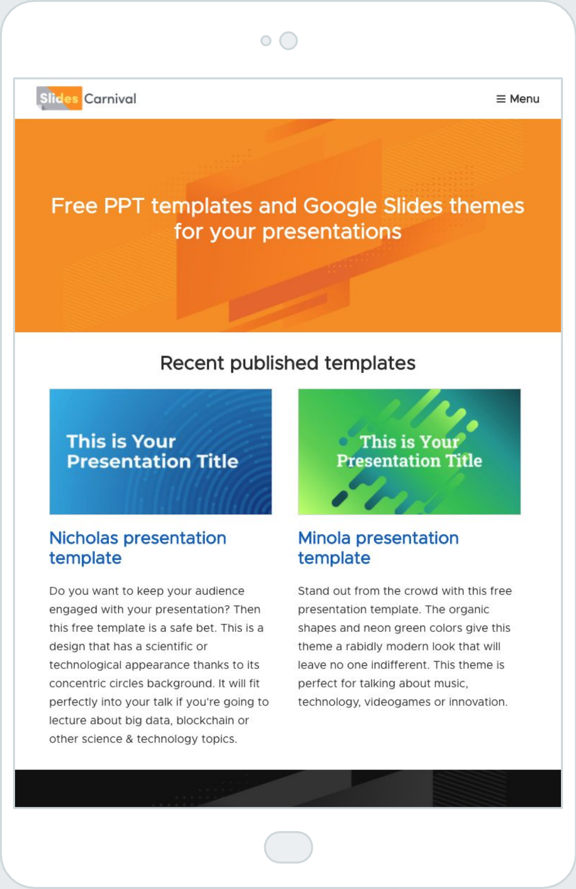
Recent published templates

This is Your Presentation Title

Nicholas presentation template

Tablet project

Show and explain your web, app or software projects using these gadget templates.


Slides Carnival
≡ Menu
This is Your Presentation Title
This is Your Presentation Title

Free PPT templates and Google Slides themes for your presentations

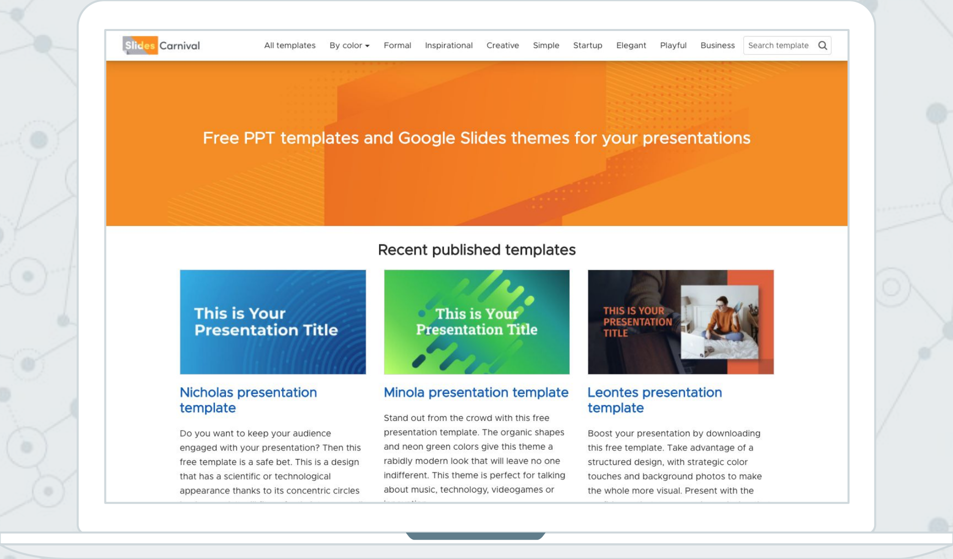
Recent published templates

Nicholas presentation template

Do you want to keep your audience engaged with your presentation? Then this free template is a safe bet. This is a design that has a scientific or technological appearance thanks to its concentric circles background. It will fit perfectly into your talk if you're going to lecture about big data, blockchain or other science & technology topics.

Minola presentation template

Stand out from the crowd with this free presentation template. The organic shapes and neon green colors give this theme a rabidly modern look that will leave no one indifferent. This theme is perfect for talking about music, technology, videogames or innovation.



Free PPT templates and Google Slides themes for your presentations

Recent published templates

Nicholas presentation template

Do you want to keep your audience engaged with your presentation? Then this free template is a safe bet. This is a design that has a scientific or technological appearance thanks to its concentric circles

Minola presentation template

Stand out from the crowd with this free presentation template. The organic shapes and neon green colors give this theme a rabidly modern look that will leave no one indifferent. This theme is perfect for talking about music, technology, videogames or

Leontes presentation template

Boost your presentation by downloading this free template. Take advantage of a structured design, with strategic color touches and background photos to make the whole more visual. Present with the

Desktop project

Show and explain your web, app or software projects using these gadget templates.

Thanks!

Any questions?

You can find me at:

@username & user@mail.me

Credits

Special thanks to all the people who made and released these awesome resources for free:

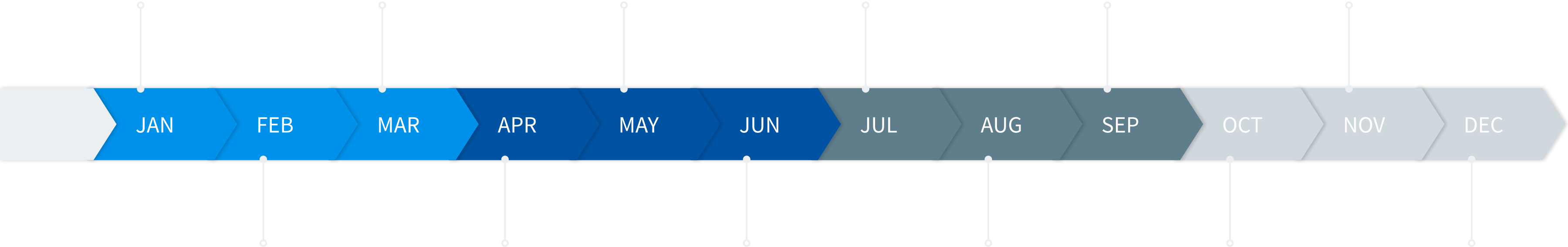
- Presentation template by SlidesCarnival
- Photographs by Unsplash

2.

Extra Resources

For Business Plans, Marketing Plans, Project Proposals, Lessons, etc

Timeline



JAN
Blue is the colour of the clear sky and the deep sea

FEB
Yellow is the color of gold, butter and ripe lemons

MAR
Red is the colour of danger and courage

APR
White is the color of milk and fresh snow

MAY
Black is the color of ebony and of outer space

JUN
Blue is the colour of the clear sky and the deep sea

JUL
Yellow is the color of gold, butter and ripe lemons

AUG
Red is the colour of danger and courage

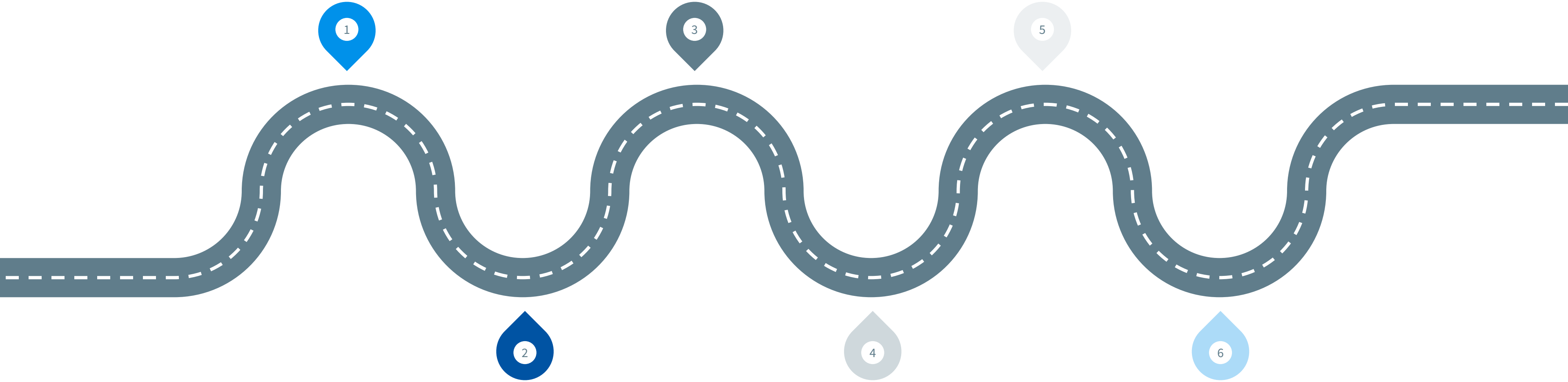
SEP
White is the color of milk and fresh snow

OCT
Black is the color of ebony and of outer space

NOV
Blue is the colour of the clear sky and the deep sea

DEC
Yellow is the color of gold, butter and ripe lemons

Roadmap



Blue is the colour of the clear sky and the deep sea

Yellow is the color of gold, butter and ripe lemons

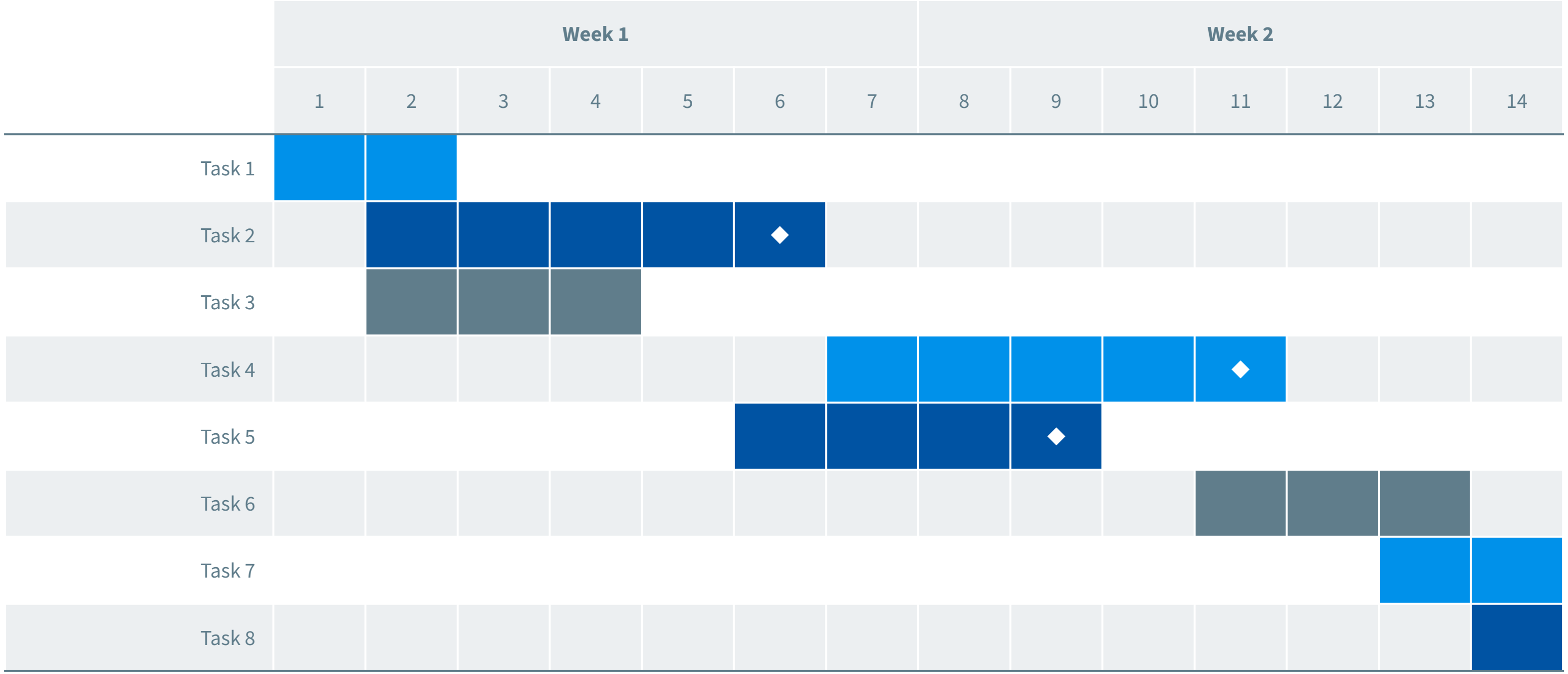
Red is the colour of danger and courage

White is the color of milk and fresh snow

Black is the color of ebony and of outer space

Blue is the colour of the clear sky and the deep sea

Gantt chart



	Week 1							Week 2						
	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Task 1	🟦	🟦												
Task 2		■	■	■	■	◆								
Task 3		▨	▨	▨										
Task 4							🟦	🟦	🟦	🟦	◆			
Task 5						■	■	■	◆					
Task 6											▨	▨	▨	
Task 7													🟦	🟦
Task 8														■

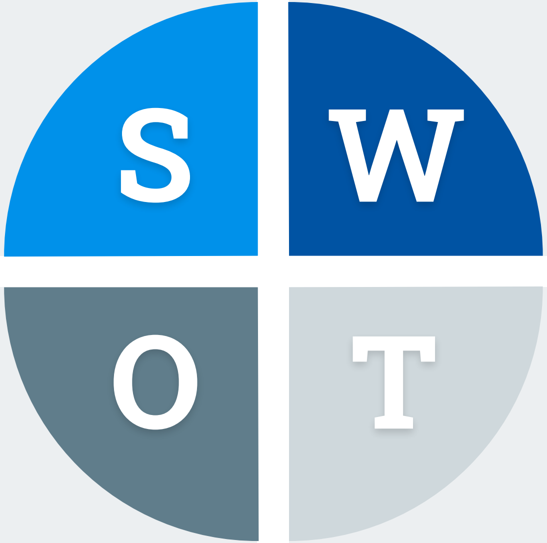
SWOT Analysis

STRENGTHS

Blue is the colour of the clear sky and the deep sea

WEAKNESSES

Yellow is the color of gold, butter and ripe lemons



Black is the color of ebony and of outer space

OPPORTUNITIES

White is the color of milk and fresh snow

THREATS

Business Model Canvas

Key Partners

Insert your content

Key Activities

Insert your content

Key Resources

Insert your content

Value Propositions

Insert your content

Customer Relationships

Insert your content

Channels

Insert your content

Customer Segments

Insert your content

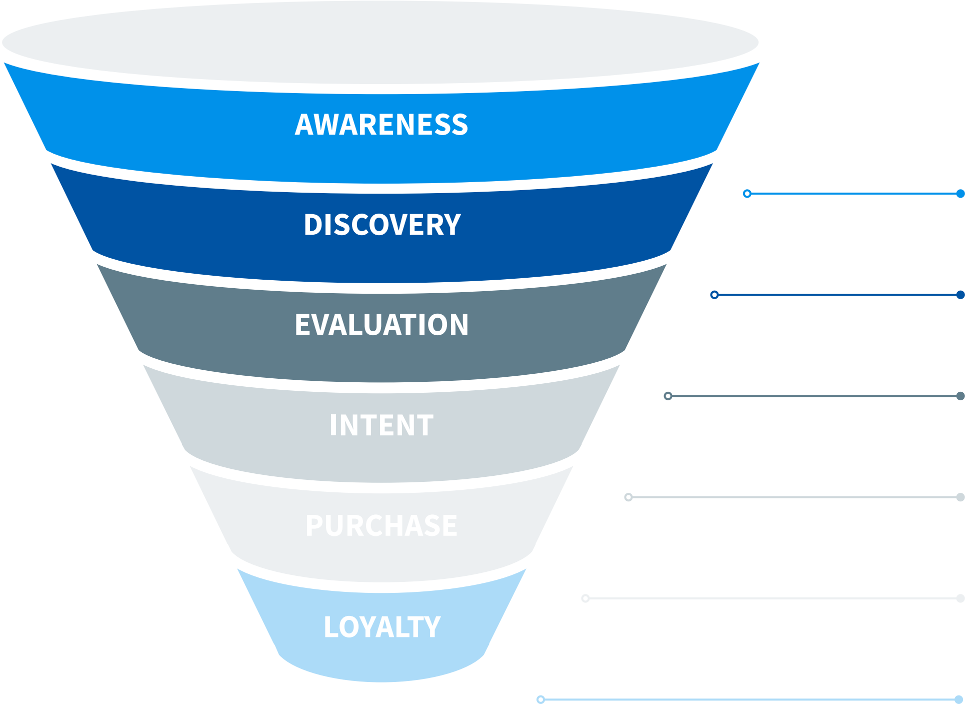
Cost Structure

Insert your content

Revenue Streams

Insert your content

Funnel



Insert your content

Insert your content

Insert your content

Insert your content

Insert your content

Insert your content

Team Presentation

Imani Jackson

JOB TITLE

Blue is the colour of the clear sky and the deep sea

Marcos Galán

JOB TITLE

Blue is the colour of the clear sky and the deep sea

Ixchel Valdía

JOB TITLE

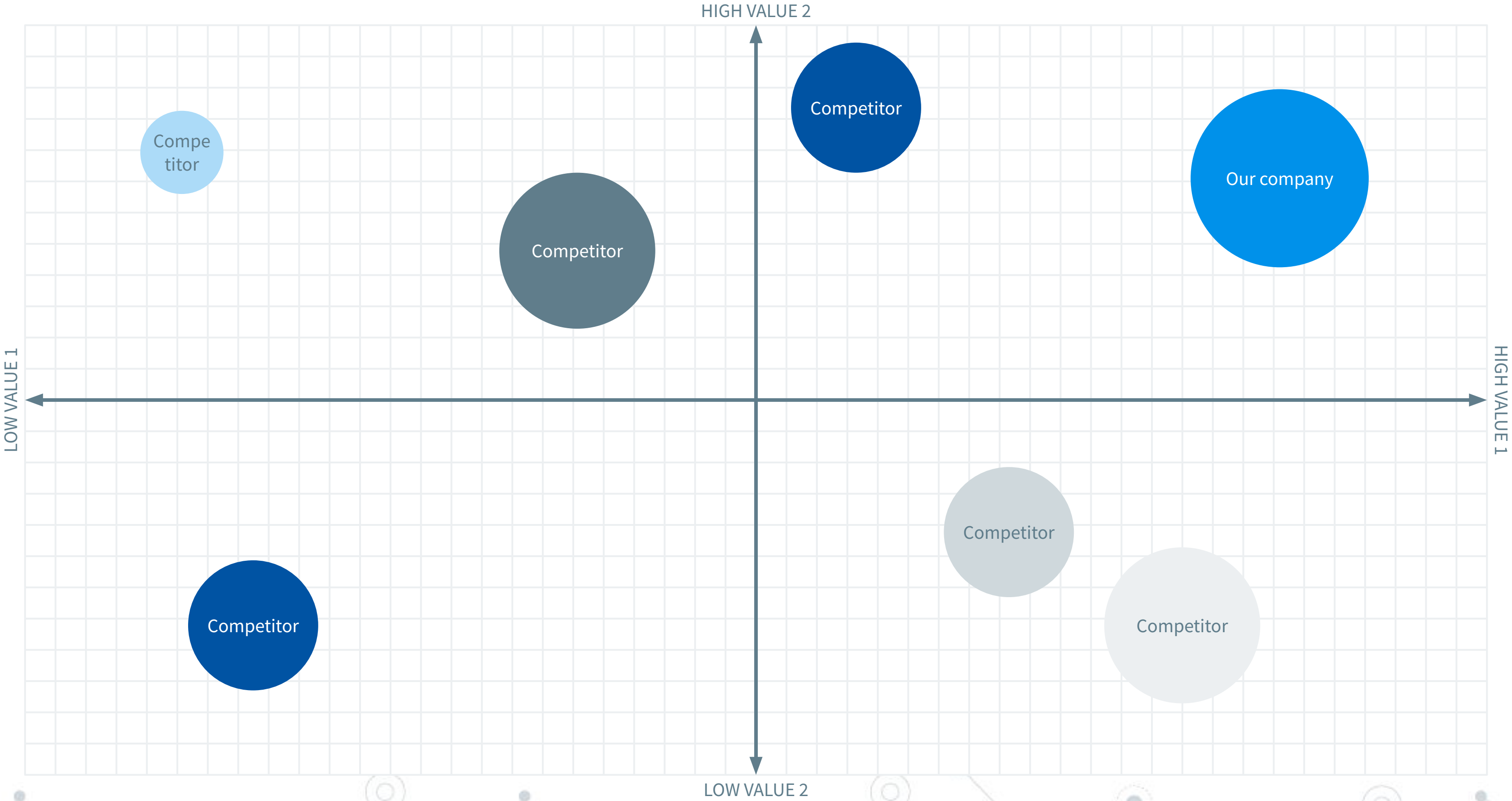
Blue is the colour of the clear sky and the deep sea

Nils Årud

JOB TITLE

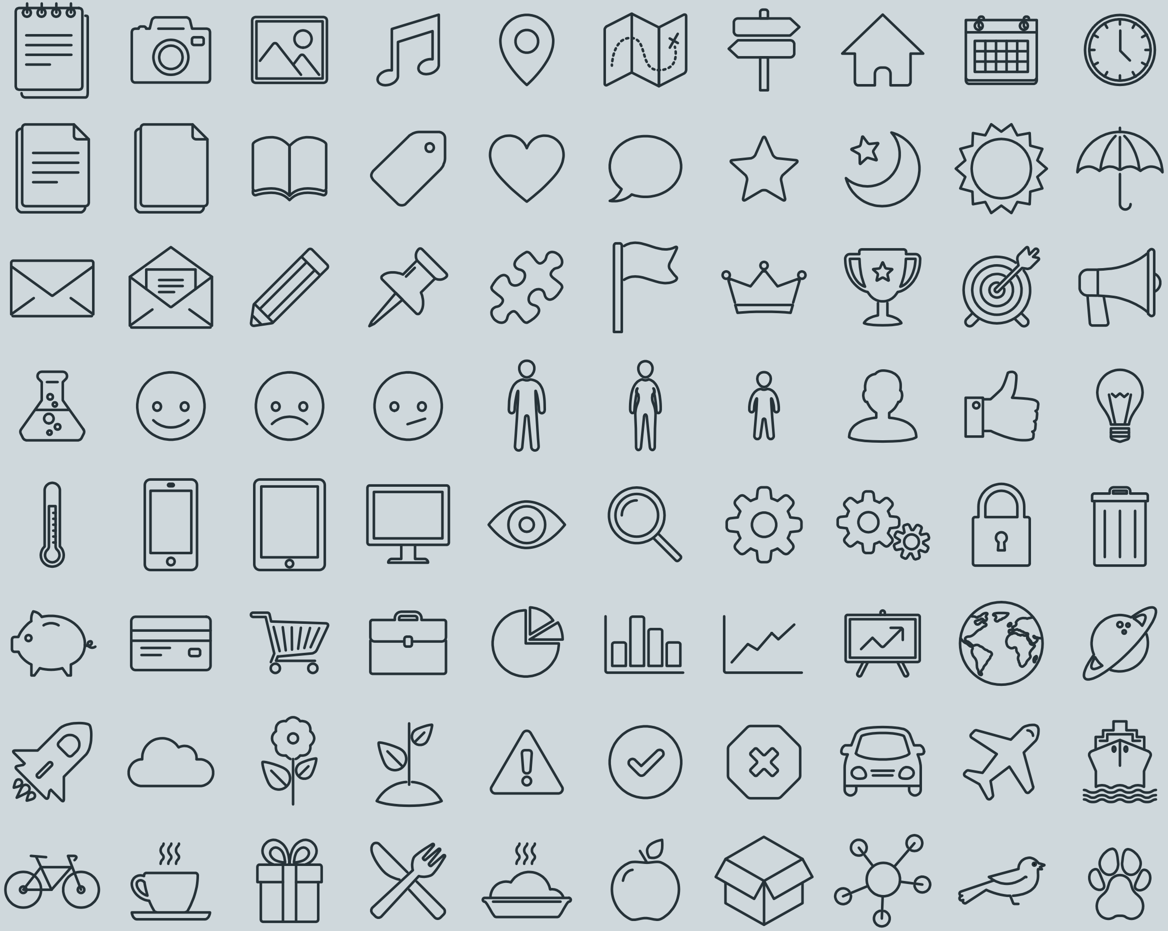
Blue is the colour of the clear sky and the deep sea

Competitor Matrix



Weekly Planner

	SUNDAY	MONDAY	TUESDAY	WEDNESDAY	THURSDAY	FRIDAY	SATURDAY
09:00 - 09:45	Task	Task	Task	Task	Task	Task	Task
10:00 - 10:45	Task	Task	Task	Task	Task	Task	Task
11:00 - 11:45	Task	Task	Task	Task	Task	Task	Task
12:00 - 13:15	✔ Free time	✔ Free time	✔ Free time	✔ Free time	✔ Free time	✔ Free time	✔ Free time
13:30 - 14:15	Task	Task	Task	Task	Task	Task	Task
14:30 - 15:15	Task	Task	Task	Task	Task	Task	Task
15:30 - 16:15	Task	Task	Task	Task	Task	Task	Task



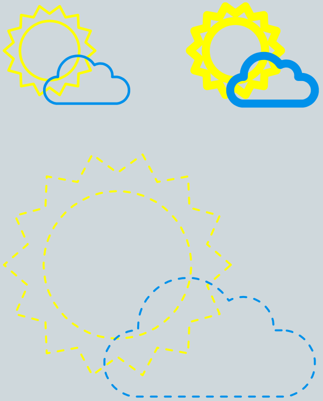
SlidesCarnival icons are editable shapes.

This means that you can:

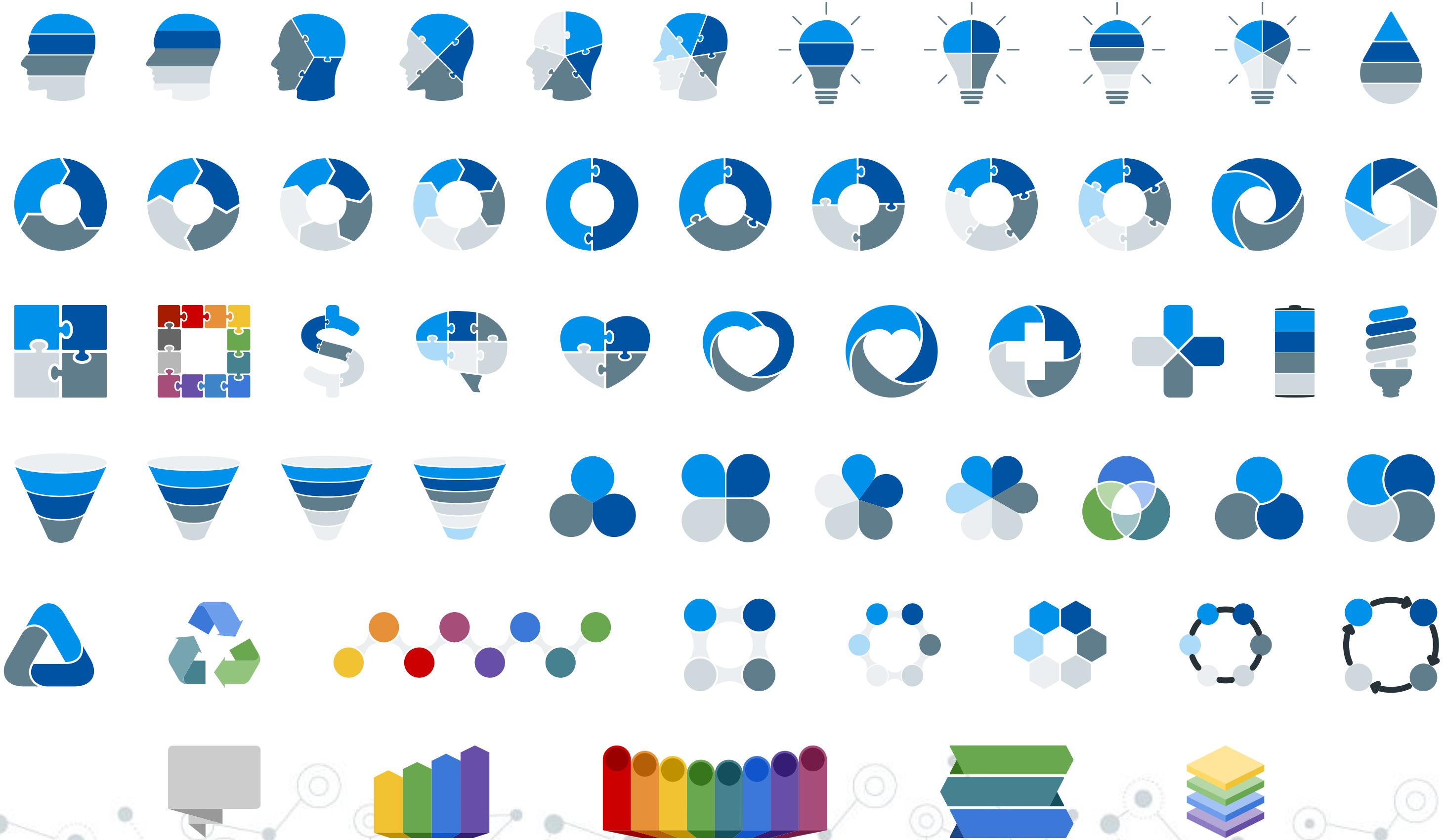
- Resize them without losing quality.
- Change line color, width and style.

Isn’t that nice? :)

Examples:



Diagrams and infographics



You can also use any emoji as an icon!
And of course it resizes without losing quality.
How? Follow Google instructions https://twitter.com/googledocs/status/730087240156643328

and many more...