

# UNE CONTRIBUTION DU TEXT-MINING À LA CONNAISSANCE DU LANGAGE DES CACHALOTS

JOSÉ COCH : DASSAULT SYSTÈMES, LARA BERKENBAUM – OLIVIER ADAM : INSTITUT  
D'ALEMBERT, SORBONNE UNIVERSITÉ

EGC - ATELIER TEXTMINE

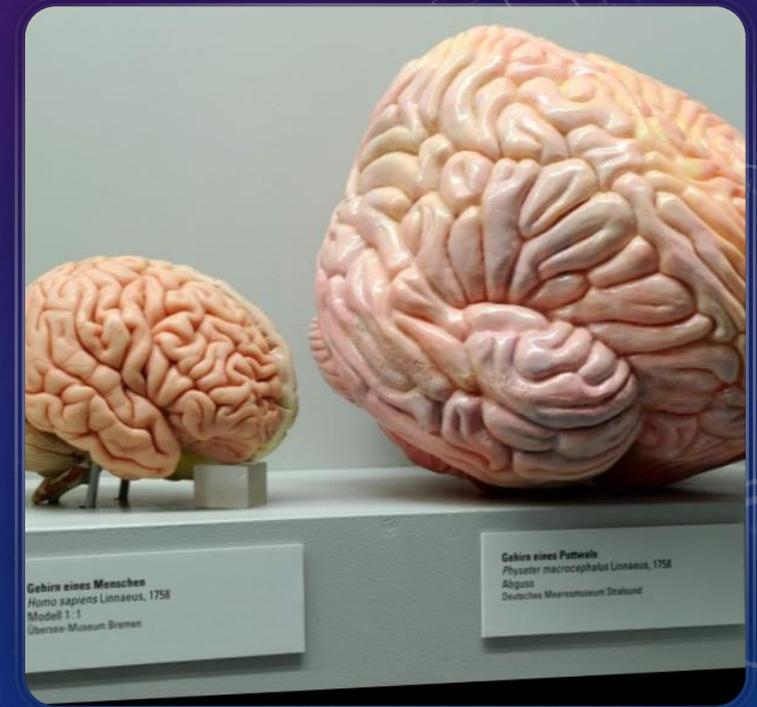
17 JANVIER 2023

# BRÈVE INTRODUCTION À LA COMMUNICATION DES CACHALOTS



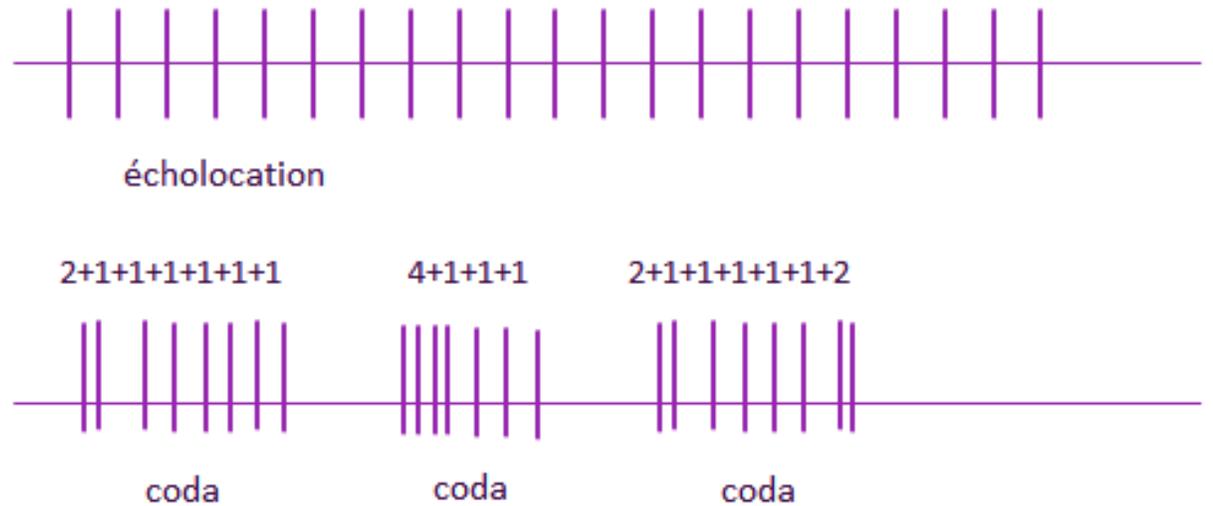
# CACHALOTS (*PHYSETER MACROCEPHALUS*)

- Espèce
  - Cétacés à dents (comme les Dauphins...)
  - Espèce protégée depuis 1985
  - Jusqu'à 20 mètres de long, sa tête représente 30% de sa longueur
  - Cerveau est le plus volumineux du règne animal
- Comportement social
  - **Clans** généralement composés de 20-30 individus ou plus (femelles adultes, jeunes de 0-4 ans, et immatures de 5 à 10 ans environ)
  - Les mâles devenant adultes quittent les clans et ont une vie solitaire sauf en période de reproduction



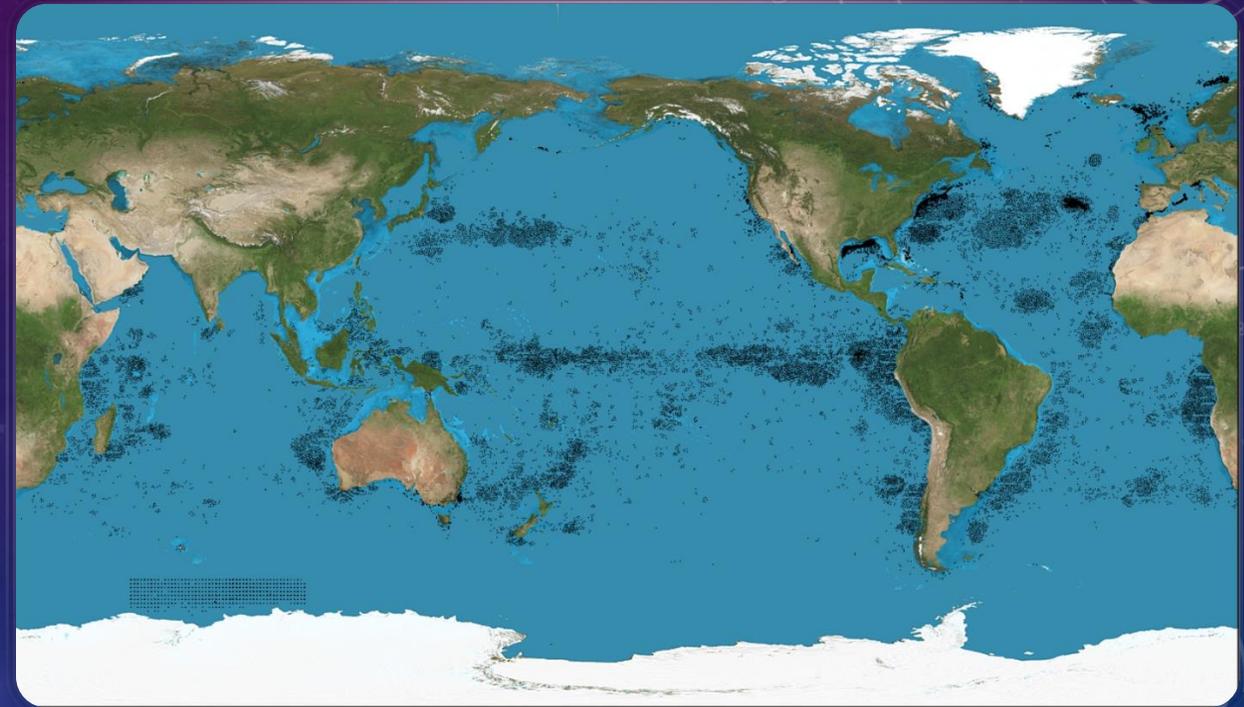
# ÉCHOLOCATION ET COMMUNICATION

- La communication des cachalots a la particularité d'être basée sur des **clics**.
- Il a été établi que les « messages » des cachalots sont composés de petites suites de clics relativement récurrentes, appelées « **coda** ».



# CLANS ET DIALECTES

- Un clan donné utilise toujours le même ensemble de codas (en général de 30 à 60)
- Chaque clan a des codas récurrentes typiques :
  - Un clan des Galapagos utilise très souvent les codas **2+2** et **2+4**
  - Nous allons voir que le clan de l'île Maurice utilise le plus souvent **2+1+1+1+1+1+1**
  - Un clan des Caraïbes utilise très souvent le **1+1+3** (absent des Galapagos et de l'île Maurice)
  - Un clan dans le Sanctuaire Pelagos (Méditerranée) utilise **3+1** dans 73% des cas
- On dit que chaque clan utilise son propre *dialecte*

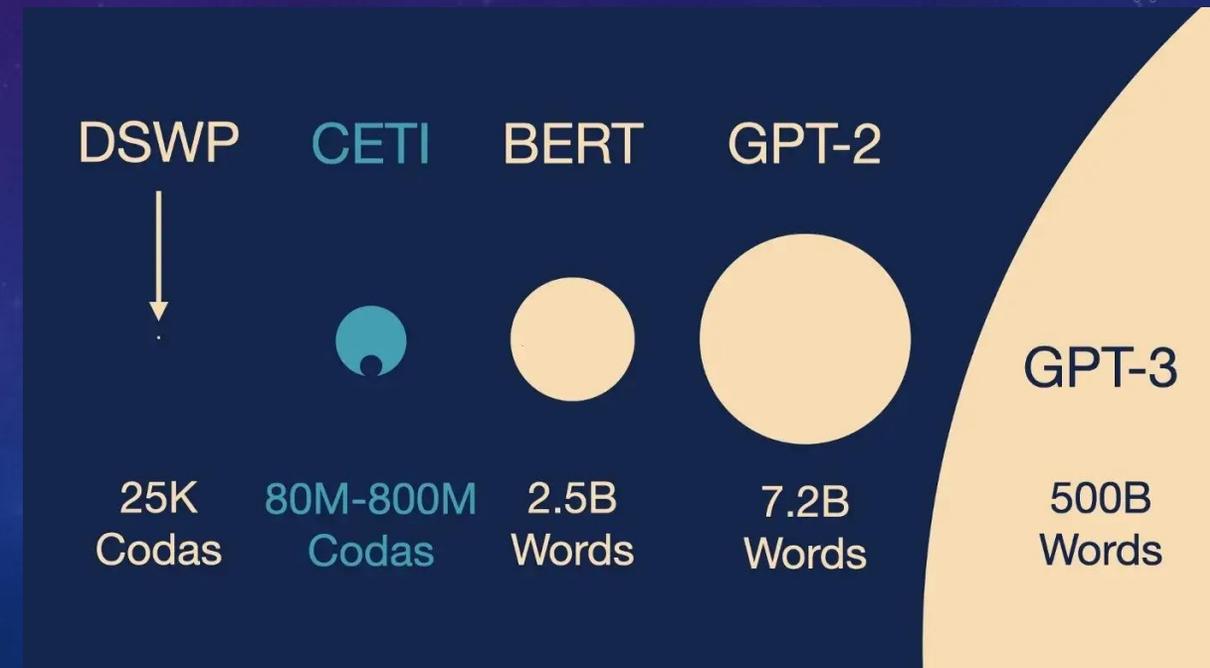


# PROJET CETI

(« CETACEAN TRANSLATION INITIATIVE »)



- Grandes avancées récentes sur le langage humain (« grands modèles »)
- L'appliquer aux interactions audio des cachalots
- **Méthodes d'automatisation d'enregistrement audio et reconnaissance des clics et codas**
- Objectif des dizaines voire des centaines de millions de codas
- De plus : des dizaines voire des centaines de dialectes
- **Cela prendra un certain temps**
- A notre connaissance : pas d'identification des individus, pas d'observation de contexte



# LE LOGICIEL PROXEM STUDIO



# PROXEM STUDIO

- Suite logicielle multisource et multilingue de Dassault Systèmes
  - pour la collecte, l'analyse et la visualisation de données textuelles
- Classification, sentiment analysis, extraction de connaissances, corrélations, tendances, signaux faibles...
- Conçue initialement pour l'Opinion Mining
- Depuis 5 ans d'autres types d'application qui prennent de l'importance
  - Text Mining sur corpus de documents longs (brevets, articles scientifiques, pages Web, contrats etc.)
  - Analyse de conversations commerciales (téléphoniques ou chat)
  - Développement / Enrichissement de Thésaurus basés sur des Corpus
  - Smart Indexation / Smart Search
  - ...
- Couverture 30 langues en « natif » et 100 en traduction automatique

# TRAITEMENTS ET PRINCIPAUX MODULES

- Traitement initial :
  - POS-tagging, NER
  - Détection des candidats termes
  - Espace vectoriel des mots du corpus
  - Indexation documentaire
- Annotate :
  - Faire le tri dans les suggestions du logiciel
  - Définir un thésaurus
  - Règles contextuelles
  - Analyse / Indexation sémantique
- Explore :
  - Moteur de recherche
  - Outils d'analyse
    - Histogrammes, treemaps, sentiment analysis, tagclouds, évolution temporelle, heatmaps...

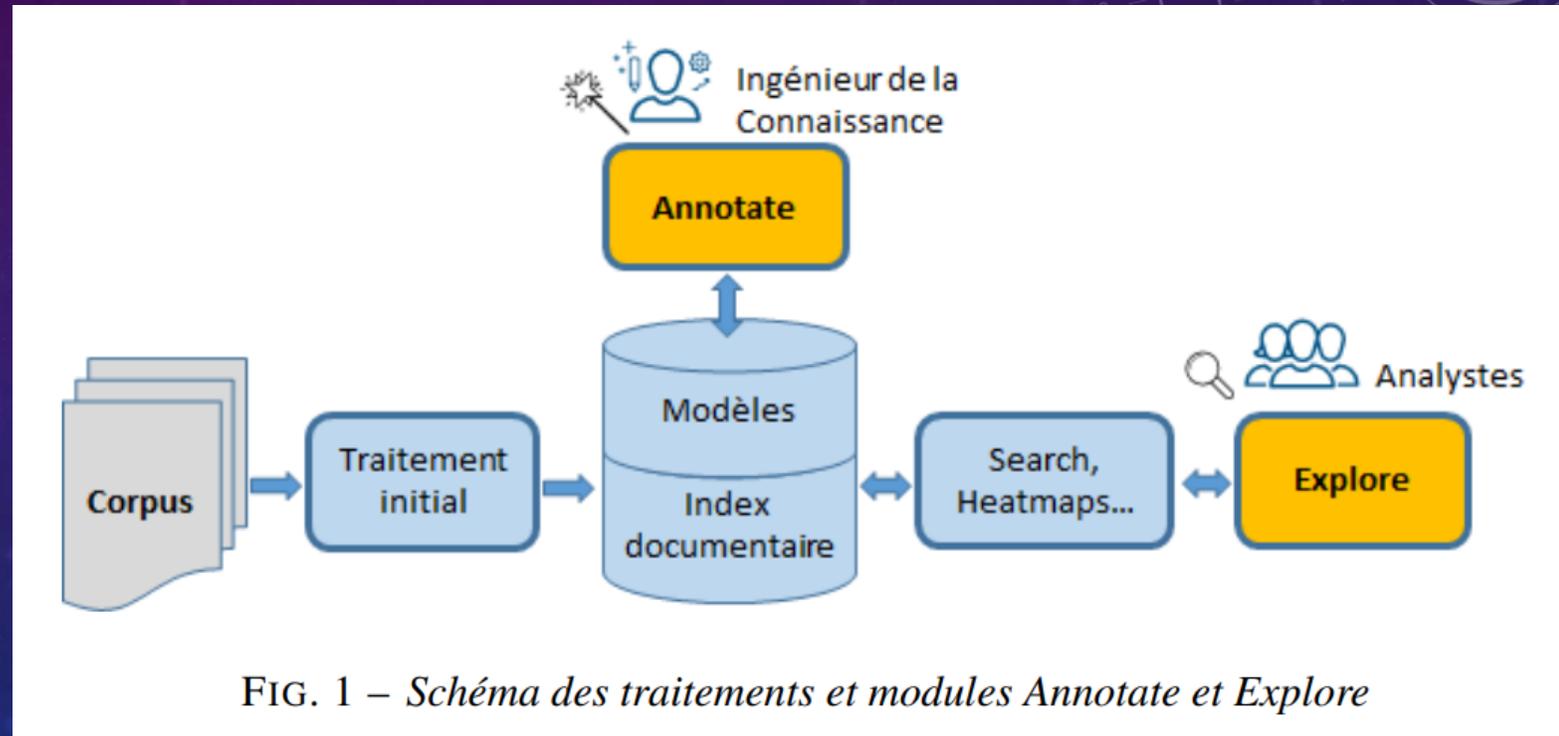


FIG. 1 – Schéma des traitements et modules Annotate et Explore

# Module Annotate

## Lexical fields



### Parameters

Minimum number of occurrences for simple words:

Recommended value : 0.5% of the current selection size

Minimum number of occurrences for phrases:

Recommended value : 0.1% of the current selection size

Number of first-level clusters:

Number of sub-clusters:

**Launch computation**

Search for a term in the suggestions 

- › no longer, told me, finally, and 169 others
- › be returning, be coming back, go back again, and 133 others
- › reasonably priced, reasonable prices, decent, and 118 others
- › valentine 's day, mother 's day, during restaurant week, and 114 others
- › new york city, top notch, manhattan skyline, and 106 others
- › striped bass, lamb burger, seafood risotto, and 102 others
- › comfortable, bar area, outdoor seating, and 90 others

?

Intention to return

Price/Ratio Quality-Price

Events/Dates

Named places

Food (Dishes/Ingredients)

Place context

What are you looking for? ⓘ

#Food



▸ Search options

### Concepts expansion

- SELECT ALL
- pork loin
- ricotta pancakes
- creamed spinach
- butternut squash
- burrata
- mushroom risotto
- greek sausage
- broccoli rabe
- pork belly
- crabmeat
- bone-in
- pork chop
- onion soup
- cobb
- rhubarb
- crab meat
- silky
- artichoke
- mofongo
- farfalle

Your corpus as filtered on

#Food ✕ ✕

2307 documents

Snippets

**Terms**

Display:  All  Queried  Significant  Multi-word

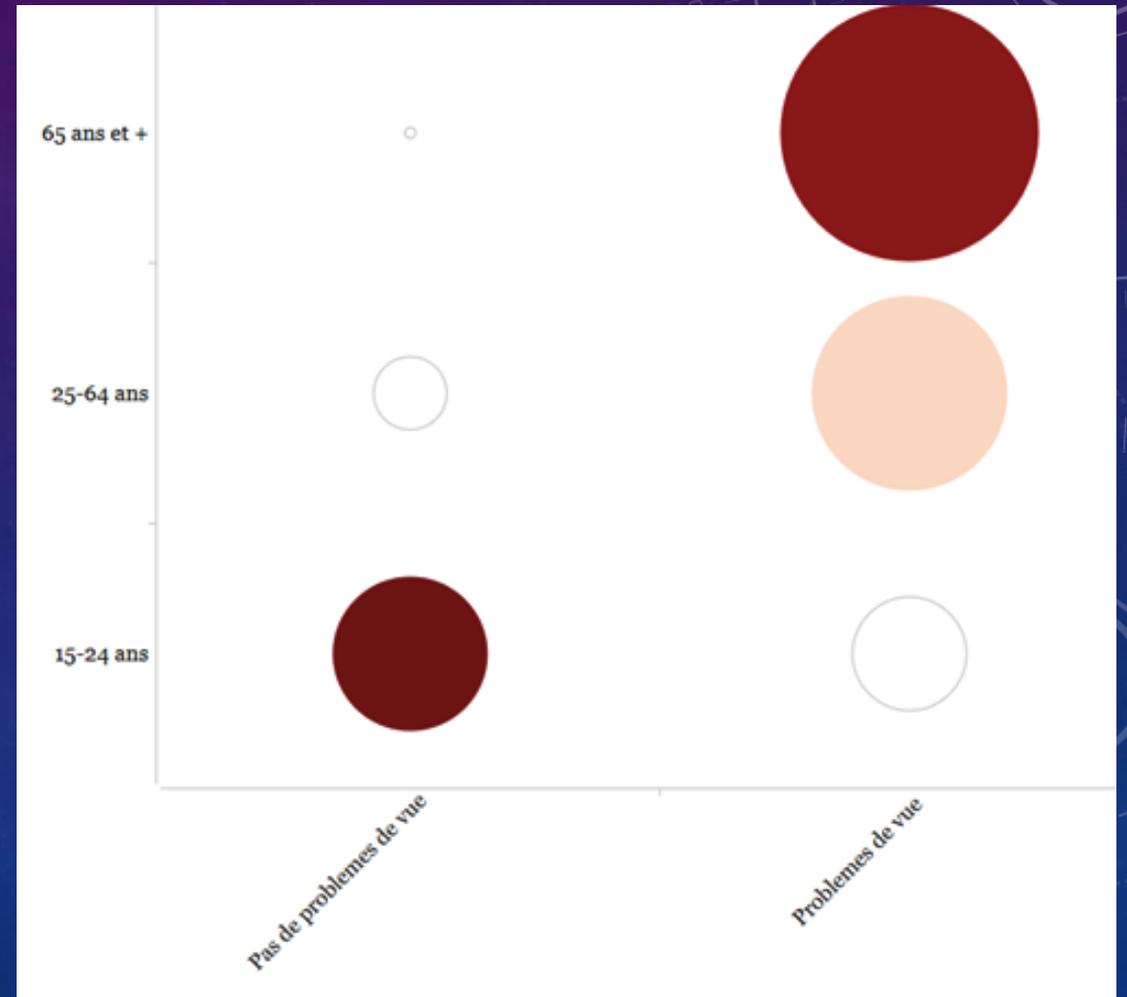
Filter:

+ salad	429	▼
+ fish	304	▼
+ steak	293	▼
+ cheese	212	▼
+ bread	208	▼
+ seafood	161	▼
+ shrimp	155	▼
+ salmon	150	▼
+ pasta	148	▼
+ chicken	147	▼
+ lamb	145	▼
+ sauce	137	▼
+ octopus	136	▼
+ burger	131	▼
+ meat	97	▼
+ mussels	88	▼

1 2 3 Next

## HEATMAPS OU CARTES DE CHALEUR / DE FRÉQUENTATION

- Croisement de deux axes d'analyse :
- La taille du rond représente le volume de documents associé
- L'intensité de la couleur du rond représente un score qui matérialise la corrélacion ou surreprésentation du croisement concerné
- La taille du rond ne représente pas la corrélation et de ce fait, souvent les *heatmaps* sont représentés avec une taille de rond artificiellement uniforme pour faciliter la visualisation des ronds les plus petits



# DONNÉES UTILISÉES



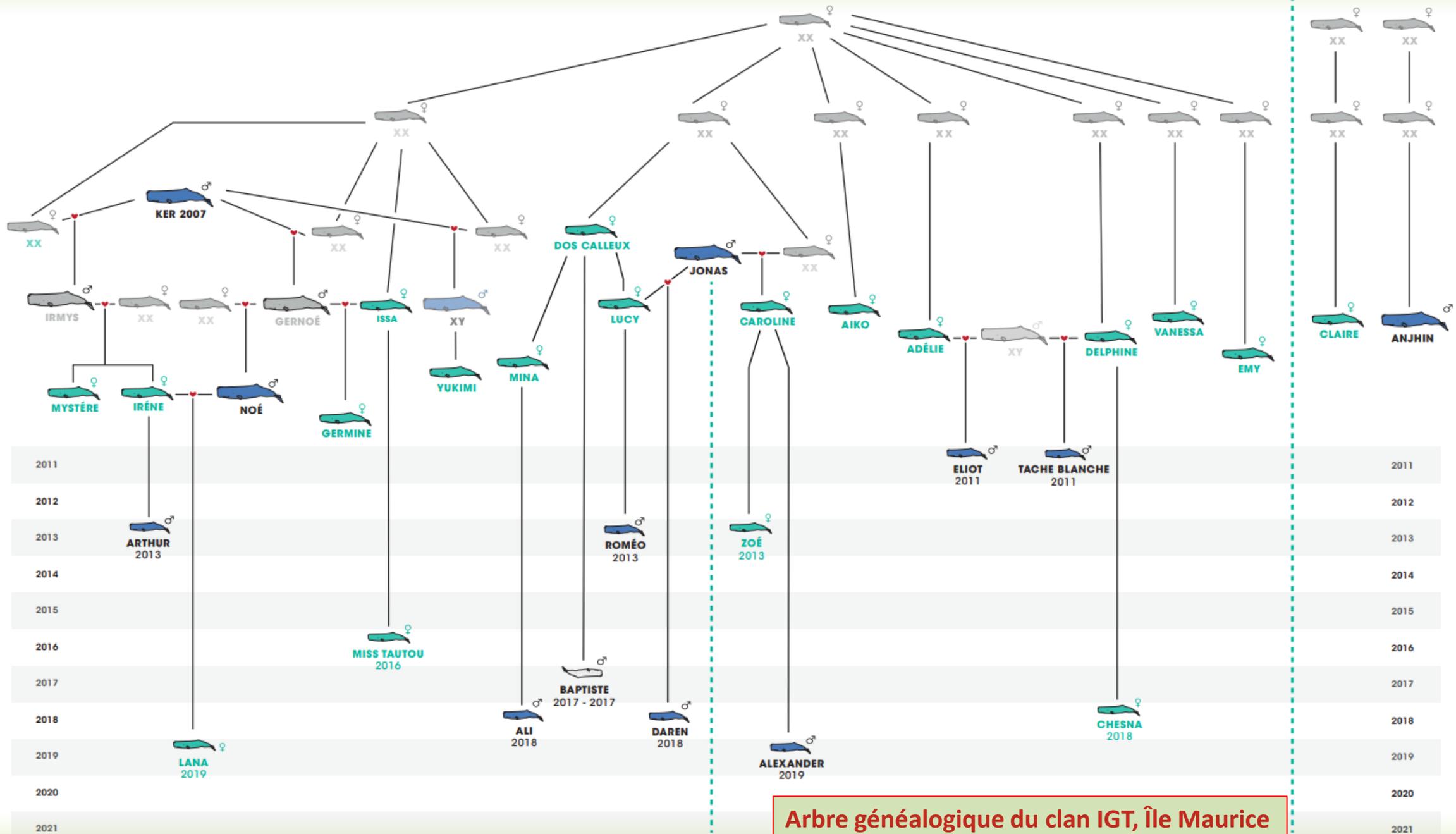
# OBJET ÉTUDIÉ

- Corpus de **138 transcriptions d'échanges audio** entre des cachalots membres du clan d'Irène Gueule Tordue ou « IGT » (Ile Maurice) représentant un total de 1622 codas
- **Données collectées par les associations Longitude 181 (François Sarano) et Un Océan de vie (René Heuzey) dans le cadre du projet « Maubidick »**
- L'enregistrement n'étant **pas stéréo** il n'a pas été possible aux transcripateurs d'identifier **l'émetteur de telle ou telle séquence sonore**, cependant **l'ensemble des participants à chaque échange a bien été identifié**
- Le corpus se présente sous forme de **fichier tabulé** avec un certain nombre d'informations (individus participants, date, heure, etc.) et le « texte » (les transcriptions des codas)
- Nous avons également eu accès à un arbre généalogique du clan d'IGT daté de janvier 2020, ce qui nous a permis de rentrer certaines **métadonnées** complémentaires



# EXTRAIT DE DONNÉES EN ENTRÉE (FICHER TABULÉ)

video	année	date	duree	nom indiv	début	codas	nbr_click
GOPR0867	2017	03/04/2017	00:00:56	Adelie_Eliot	00:00:21	2+1+1+1+1+1+1	8
GOPR0867	2017	03/04/2017	00:00:56	Adelie_Eliot	00:00:23	2+1+1+1+1+1	7
GOPR0867	2017	03/04/2017	00:00:56	Adelie_Eliot	00:00:27	2+1+1+1+1+1	7
GOPR0867	2017	03/04/2017	00:00:56	Adelie_Eliot	00:00:28	2+1+1+1+1+1+1+1	9
GOPR0867	2017	03/04/2017	00:00:56	Adelie_Eliot	00:00:30	2+1+1+1+1+1+1	8
GOPR0867	2017	03/04/2017	00:00:56	Adelie_Eliot	00:00:33	2+1+1+1+1+1	7
GOPR0867	2017	03/04/2017	00:00:56	Adelie_Eliot	00:00:37	3+1+1+1+1+1+1	9
GOPR0867	2017	03/04/2017	00:00:56	Adelie_Eliot	00:00:41	3+1+1+1+1+1+1	9
GOPR0867	2017	03/04/2017	00:00:56	Adelie_Eliot	00:00:46	3+1+1+1+1+1+1	9
GOPR0968	2017	06/04/2017	00:01:10	Vanessa_Zoe	00:00:40	5+1+1+1	8
GOPR0968	2017	06/04/2017	00:01:10	Vanessa_Zoe	00:00:41	2+1+1+1+1+1+4	11
GOPR0968	2017	06/04/2017	00:01:10	Vanessa_Zoe	00:00:44	2+1+1+1+1+1	7
GOPR0968	2017	06/04/2017	00:01:10	Vanessa_Zoe	00:00:51	2+1+1+1+1+1+1+1	9
GOPR0968	2017	06/04/2017	00:01:10	Vanessa_Zoe	00:00:56	2+1+1+1+1+1+1	8
GOPR0968	2017	06/04/2017	00:01:10	Vanessa_Zoe	00:01:08	5+1+1+1+1	9



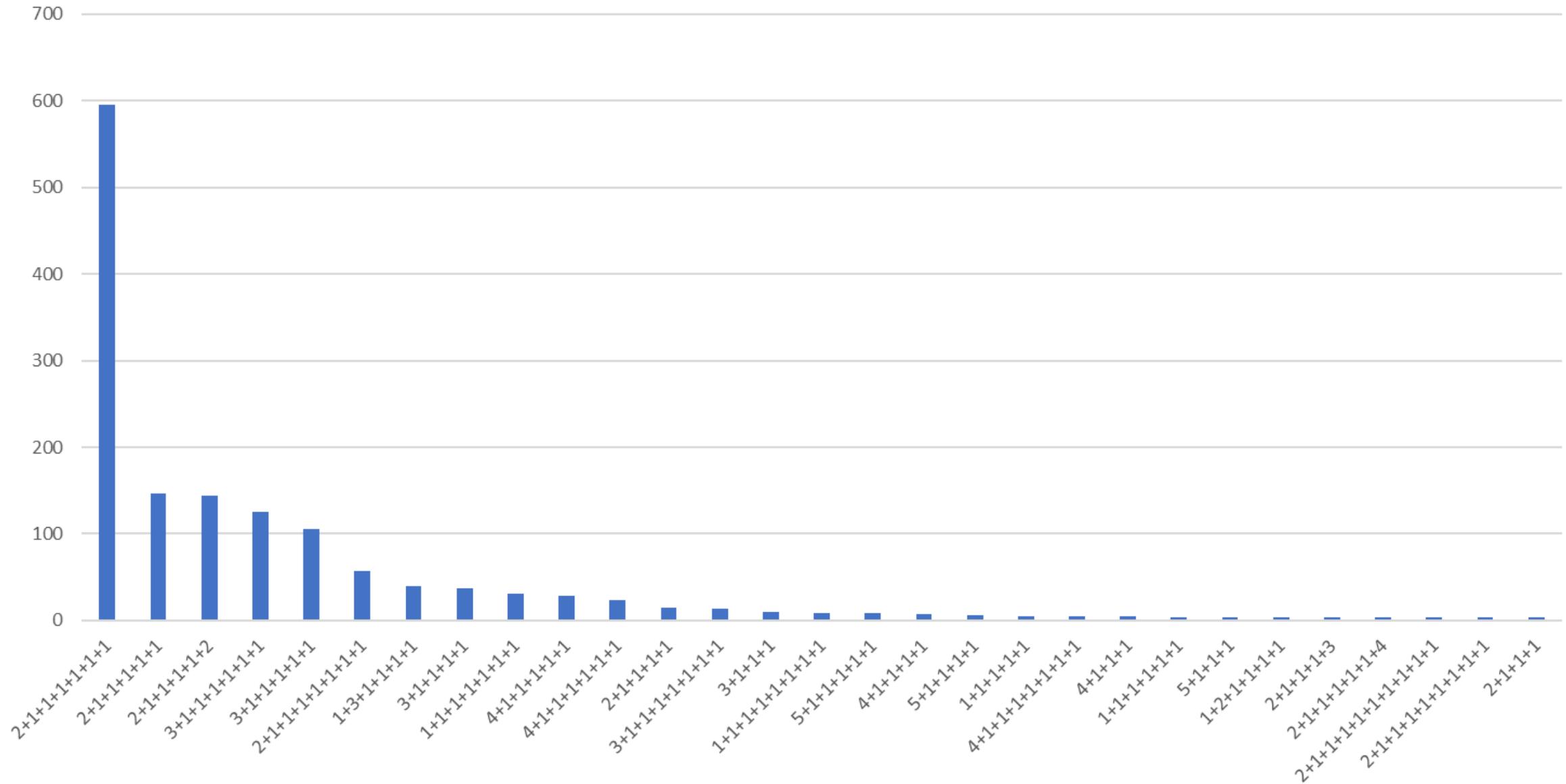
Arbre généalogique du clan IGT, Île Maurice

# CONSTATS PRÉALABLES

- On observe environ 80 codas différentes dans les 138 échanges ; cependant seulement **30 ont une fréquence supérieure à 2**. La coda la plus fréquente **2+1+1+1+1+1+1**, a une fréquence d'un peu moins de 600
- La très grande majorité des codas comportent **8 clics**. Suivent les codas à 7, 9 et 10 clics. Voir graphique.
- La slide suivante montre la fréquence de chacune des codas de la trentaine de codas les plus fréquentes



# Nombre d'occurrences des codas



# MÉTHODOLOGIE



# MÉTHODOLOGIE

- PRÉPARATION DES DONNÉES
  - FORMAT : “UNE CONVERSATION/ÉCHANGE PAR LIGNE” CONTRE “UN CODA PAR LIGNE”
  - NORMALISATION DES NOMS D’INDIVIDUS
  - AJOUT DE **MÉTADONNÉES** (SEXE DES INDIVIDUS, RELATIONS FAMILIALES, ET AUTRES) VENANT DE L’ARBRE GÉNÉALOGIQUE DU CLAN IGT
- TESTS DE CHARGEMENT ET MISE AU POINT DES DONNÉES
- PARAMÉTRAGE DE L’OUTIL, SÉLECTION DE MULTI-CODAS
  - ANALYSE DES PREMIERS RÉSULTATS
  - MISE AU POINT DU PARAMÉTRAGE ET DES MULTI-CODAS
  - (ITÉRATION)
- ANALYSE DES RÉSULTATS ET RÉDACTION D’UN PROJET DE RAPPORT

# DONNÉES NORMALISÉES ET ENRICHIES AVEC DES MÉTADONNÉES VENANT DE L'ARBRE GÉNÉALOGIQUE

video	annee	date	duree	nomindiv	type	famille	maman-enfant	tous males	toutes femelles	jeunes	tous jeunes	Immatures	tous immatures	tous adultes	age	tous jeunes immatures	individus	G1	G2	G3	début	conversation-aux	nbr codas	nbr indiv	nbr femelles
GOPR0867	2017	03/04/2017	00:00:56	Adelle_Eliot	maman-fils	maman-fils	1	0	0	FAUX	0	VRAI	0	FAUX	melange	0	Adelle Eliot	1			00:00:46	2+1+1+1+1+1+1 2+1+1+1+1+1 2+1+1+1+1+1 2+1+1+1+1+1+1+1 2+1+1+1+1+1 2+1+1+1+1+1 3+1+1+1+1+1+1 3+1+1+1+1+1+1 3+1+1+1+1+1+1	9	2	1

Mère-fils/fille

Tous mâles

Toutes femelles

Tous adultes

Tous jeunes  
immatures

Individus  
(norm.)

(Sous-)Groupes

# RÉSULTATS



# DÉTECTION DE MULTI-CODAS

- L'outil détecte un certain nombre de suites de deux et parfois trois, quatre ou même cinq codas dont la fréquence est significativement supérieure à celle à laquelle on s'attendrait si la répartition était au hasard. Dans les langues humaines, cette méthode permet par exemple de détecter les mots composés et les idiomes.
- Nous utilisons le mot « signe » pour faire référence aussi bien aux codas unitaires, qu'aux séquences multi-codas détectées par l'outil
- Parmi ces multi-codas il existe des répétitions de la même coda, et des suites de codas différentes
- La slide suivante montre les deux cas

# MULTI-CODAS RÉPÉTITIFS OU PAS AVEC LE NOMBRE D'ÉCHANGES DANS LESQUELLES ILS APPARAISSENT

Multi-coda répétitif	Nombre d'échanges
2+1+1+1+1+1 2+1+1+1+1+1	77
2+1+1+1+1+1 2+1+1+1+1+1 2+1+1+1+1+1	52
2+1+1+1+1+1 2+1+1+1+1+1 2+1+1+1+1+1 2+1+1+1+1+1	25
<u>2+1+1+1+1+2 2+1+1+1+1+2</u>	22
<u>2+1+1+1+1+1 2+1+1+1+1+1</u>	14
2+1+1+1+1+1+1 2+1+1+1+1+1+1 2+1+1+1+1+1+1 2+1+1+1+1+1+1	11
<u>3+1+1+1+1+1 3+1+1+1+1+1</u>	10
<u>1+3+1+1+1+1 1+3+1+1+1+1</u>	8

Multi-coda non répétitif	Nombre d'échanges
2+1+1+1+1+1 2+1+1+1+1+1+1	35
2+1+1+1+1+1+1 2+1+1+1+1+1	28
2+1+1+1+1+2 2+1+1+1+1+1+1	28
2+1+1+1+1+1+1 3+1+1+1+1+1+1	27
3+1+1+1+1+1+1 2+1+1+1+1+1+1	27
2+1+1+1+1+1+1 2+1+1+1+1+2	26
2+1+1+1+1+1+1 3+1+1+1+1+1	23
3+1+1+1+1+1 2+1+1+1+1+1+1	23
2+1+1+1+1+1+1+1 2+1+1+1+1+1+1	13
3+1+1+1+1 2+1+1+1+1+1+1	11
1+1+1+1+1+1+1 2+1+1+1+1+1+1	11
2+1+1+1+1+2 2+1+1+1+1+1	11
3+1+1+1+1+1 2+1+1+1+1+1+1 2+1+1+1+1+1+1 2+1+1+1+1+1+1	9
3+1+1+1+1+1+1 3+1+1+1+1+1	9
2+1+1+1+1+1 2+1+1+1+1+1+1 2+1+1+1+1+1+1 2+1+1+1+1+1+1	8
1+3+1+1+1+1 2+1+1+1+1+1+1	8
2+1+1+1+1+1+1 1+3+1+1+1+1	8
2+1+1+1+1+1+1 2+1+1+1+1+1+1 2+1+1+1+1+1+1 2+1+1+1+1+1+1	7
3+1+1+1+1+1+1 2+1+1+1+1+1+1 2+1+1+1+1+1+1 2+1+1+1+1+1+1	7
3+1+1+1+1+1+1 2+1+1+1+1+1+1+1	7
3+1+1+1+1+1+1 2+1+1+1+2	6
2+1+1+1+1+1+1 2+1+1+1+1+1+1 2+1+1+1+1+1+1 2+1+1+1+1+2	6
3+1+1+1+1+1+1 2+1+1+1+1+1	5
3+1+1+1+1+1+1 4+1+1+1+1+1+1	5
3+1+1+1+1+1+1 3+1+1+1+1+1+1+1	5

Le coda 2+1+1+1+1+1, le plus fréquent, donne lieu à plusieurs multi-codas répétitifs (de 2 à 5 fois). Les autres répétitions sont ici entourées en rouge.

Les 8 multi-codas non répétitifs les plus fréquents apparaissent dans les sur-représentations étudiées (voir plus loin les heatmaps).

La coda 1+1+1+1+1+1 2+1+1+1+1+1 a une fréquence 10 fois supérieure à la paire inverse 2+1+1+1+1+1 1+1+1+1+1+1

The background features a dark blue gradient with various data visualization elements. On the left, there is a bar chart with three groups of bars labeled Q1, Q2, and Q3. On the right, there is a circular gauge or dial with numerical markings from 80 to 210 and a needle pointing towards 180. The overall aesthetic is professional and data-driven.

# ANALYSE DES RÉSULTATS

CORRÉLATIONS - SURPONDÉRATIONS



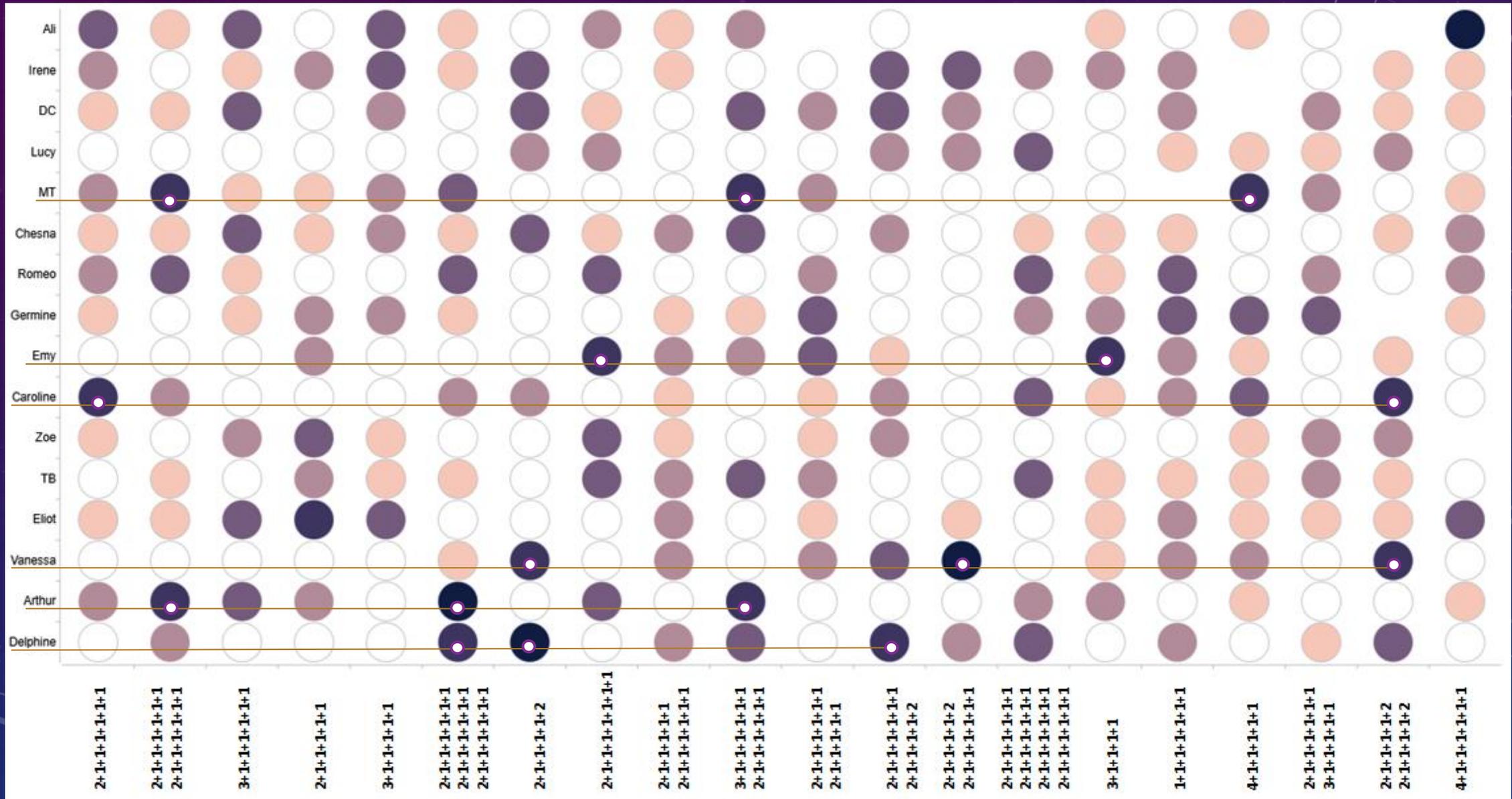
# MÈRE-FILS OU FILLE

- **2+1+1+1+1+2**
  - (effectif 23/45 avec une fréquence plus de deux fois supérieure que dans les autres conversations), suivie par
- **3+1+1+1+1+1+1**
  - (effectif 26/45, fréquence 27% supérieure),
- et les bi-codas
- **2+1+1+1+1+1+1 2+1+1+1+1+2** (effectif 14/45, fréquence 140% supérieure), et
- **2+1+1+1+1+2 2+1+1+1+1+2** (effectif 11/45, fréquence deux fois supérieure).
- L'ordre entre codas semble avoir une importance car si **2+1+1+1+1+1+1 2+1+1+1+1+2** est sur-représentée dans les conversations mère-fils/fille, ce n'est pas du tout le cas de la paire inverse **2+1+1+1+1+2 2+1+1+1+1+1+1** (effectif 9/45, fréquence presque identique voire un petit peu inférieure).





# Heatmap signes utilisés - individus



# DISCUSSION



# CONCLUSIONS - OBSERVATIONS

- Les clics ne sont pas du tout produits au hasard, mais dans des combinaisons assez contraintes dans un certain nombre relativement réduit de codas, propre à chaque clan
- Certaines suites de deux codas ou plus arrivent significativement plus souvent que si c'était par simple probabilité mathématique s'il s'agissait de phénomènes indépendants (phénomène des « multi-codas »)
- L'ordre des codas a clairement une importance dans certains cas
- Certaines codas ou combinaisons de codas sont clairement surpondérées dans certaines configurations : mère-fis/fille, tranches d'âge, individus
- D'autres surpondérations apparaissent mais elles sont à confirmer par l'analyse de corpus plus volumineux (mère-fils vs. mère-fille, frères ou sœurs, tous mâles, ...)
- Nous n'avons pas trouvé de phénomène ou sur-représentation particulière aux sous-groupes G1, G2, G3

# COMMUNICATION HUMAINE – COMMUNICATION ANIMALE

<b>Différences sur...</b>	<b>Communication humaine</b>	<b>Communication animale</b>
<b>Apprentissage</b>	nécessaire	non nécessaire
<b>Base</b>	unités arbitraires (sans lien entre le mot et l'objet à décrire)	unités iconiques (avec lien entre le signifié et le signifiant)
<b>Unités</b>	discrètes (divisibles et combinables)	graduées (messages indivisibles et non combinables)
<b>Ordre des unités, syntaxe</b>	syntaxe	superposition d'unités, l'ordre n'a pas d'importance pour le message
<b>Mutualité</b>	fréquente (les individus se répondent)	relativement rare
<b>Mensonges, divagations</b>	fréquents	cas sporadiques de mensonges et pas de divagation
<b>Polysémie</b>	prévalente (les mots peuvent avoir plusieurs sens)	monosémie
<b>Combinatoire/ Productivité/ Générativité</b>	combinatoire potentiellement illimitée	pas de combinatoire ou combinatoire limitée

# CLASSIFICATION DE LA COMMUNICATION DES CACHALOTS

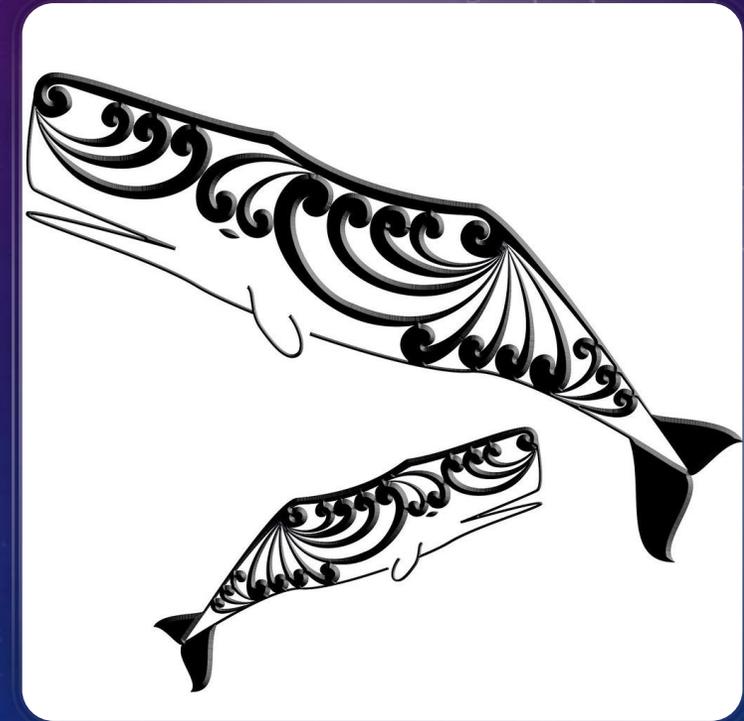
Différences sur...	Communication humaine	Communication animale	Cachalots
<b>Apprentissage</b>	nécessaire	non nécessaire	<b>apprentissage nécessaire</b>
<b>Base</b>	unités arbitraires (sans lien entre le mot et l'objet à décrire)	unités iconiques (avec lien entre le signifié et le signifiant)	<b>unités (codas) arbitraires</b>
<b>Unités</b>	discrètes (divisibles et combinables)	graduées (messages indivisibles et non combinables)	<b>discrètes</b>
<b>Ordre des unités, syntaxe</b>	syntaxe	superposition d'unités, l'ordre n'a pas d'importance pour le message	<b>l'ordre des unités semble avoir une importance</b>
<b>Mutualité</b>	fréquente (les individus se répondent)	relativement rare	<b>fréquente</b>
<b>Mensonges, divagations</b>	fréquents	cas sporadiques de mensonges et pas de divagation	on l'ignore pour l'instant
<b>Polysémie</b>	prévalente (les mots peuvent avoir plusieurs sens)	monosémie	on l'ignore ; cependant certaines codas sont utilisées dans des contextes différents, notamment dans des multi-codas différents
<b>Combinatoire/ Productivité/ Générativité</b>	combinatoire potentiellement illimitée	pas de combinatoire ou combinatoire limitée	<b>combinatoire potentiellement illimitée</b>

# SUITE DES TRAVAUX



# SUITE DES TRAVAUX

- Enregistrement stéréo et **identification des individus locuteurs**
- Corpus plus conséquent (**x 5**)
- Thèse de Lara Berkenbaum
- Description/**Classification des comportements**. En particulier :
  - Haptic/Active swimming/Approach/Others/Rest/Moving away
    - « Haptic » très majoritaire : sous-classement par type de contact



DES QUESTIONS ?



# ANNEXES

The background is a dark blue gradient with a field of small white stars. Overlaid on this are several technical diagrams. In the top right, there is a large circular diagram with concentric rings and radial lines, resembling a scale or a gauge, with numbers like 100, 120, 140, 160, 170, 180, 190, and 200. In the bottom right, there is a diagram with two concentric circles and a dashed outer ring, with arrows indicating a clockwise direction. In the bottom left, there is a diagram with a solid inner circle and a dashed outer ring, with an arrow indicating a counter-clockwise direction. In the top left, there is a simple circular diagram with a dashed line and an arrow.

# DÉTECTION DE TERMES MULTI-MOTS (MULTI-CODAS)

- Approche "PMI" pour « Pointwise mutual information »
- Probabilité co-occurrence (ou **suite**) de deux mots comparée à la probabilité théorique s'il s'agissait de phénomènes indépendants
- « pomme de terre »  $\neq$  « terre de pomme »
- Itération pour termes de trois mots ou plus (« pomme de terre nouvelle »)
- En cas de langue connue, prise en compte du POS-tagging et des mots vides

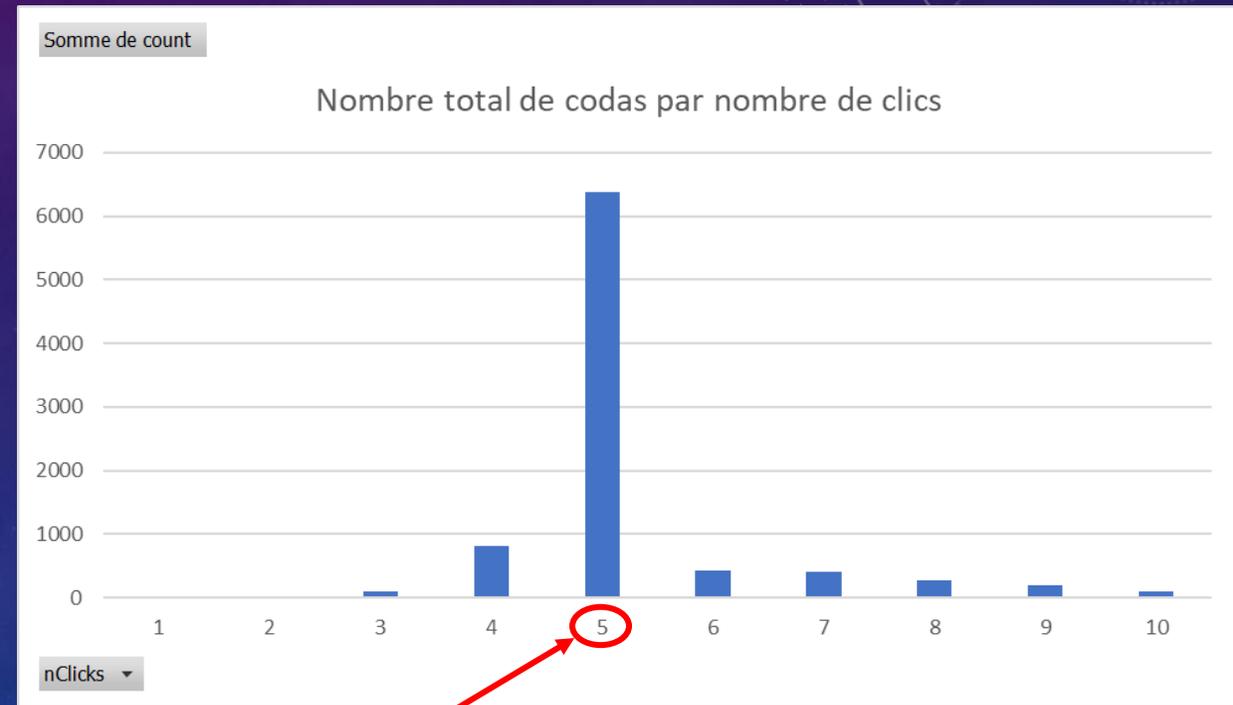
## REGROUPEMENT DE TERMES

- Espaces de vecteurs de mots et termes / word2vec
- Proximité des contextes : possible proximité des mots/termes (à valider)
- Variantes graphiques, morpho, synonymes, mot proches

# COMPARATIF CORPUS ILE MAURICE - CARAÏBES

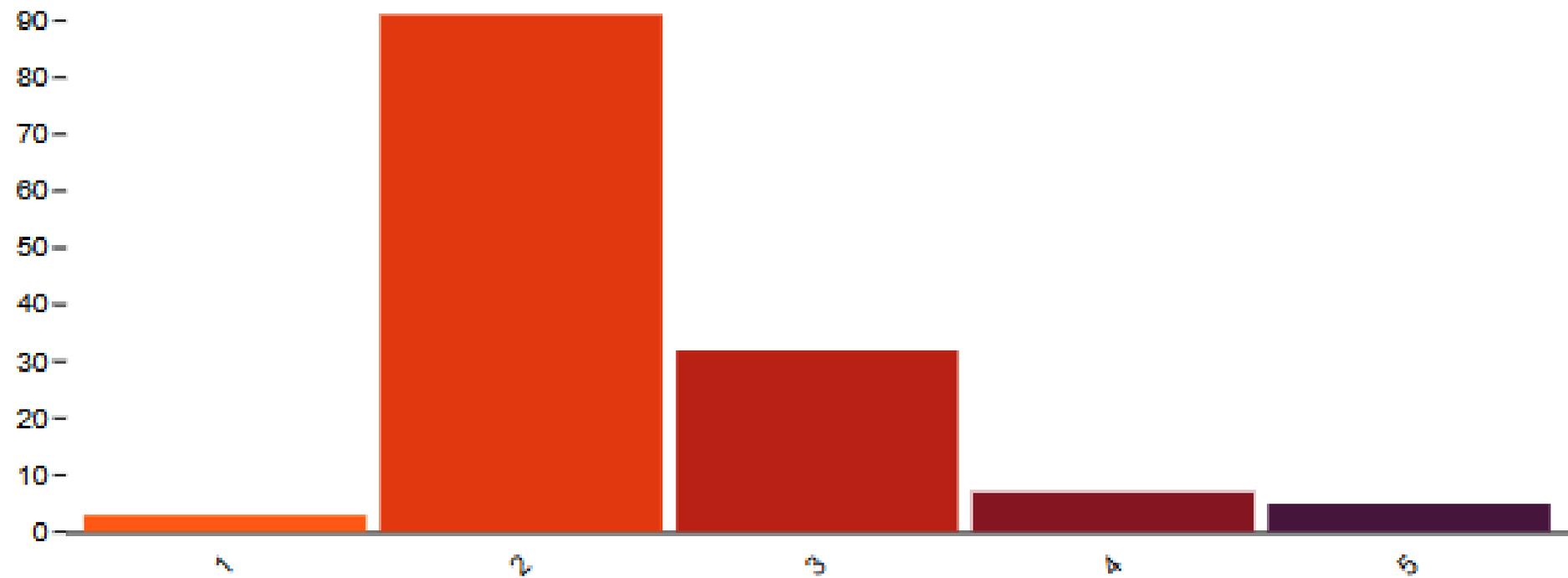


ILE MAURICE



CARAÏBES

## Nombre d'individus par échange



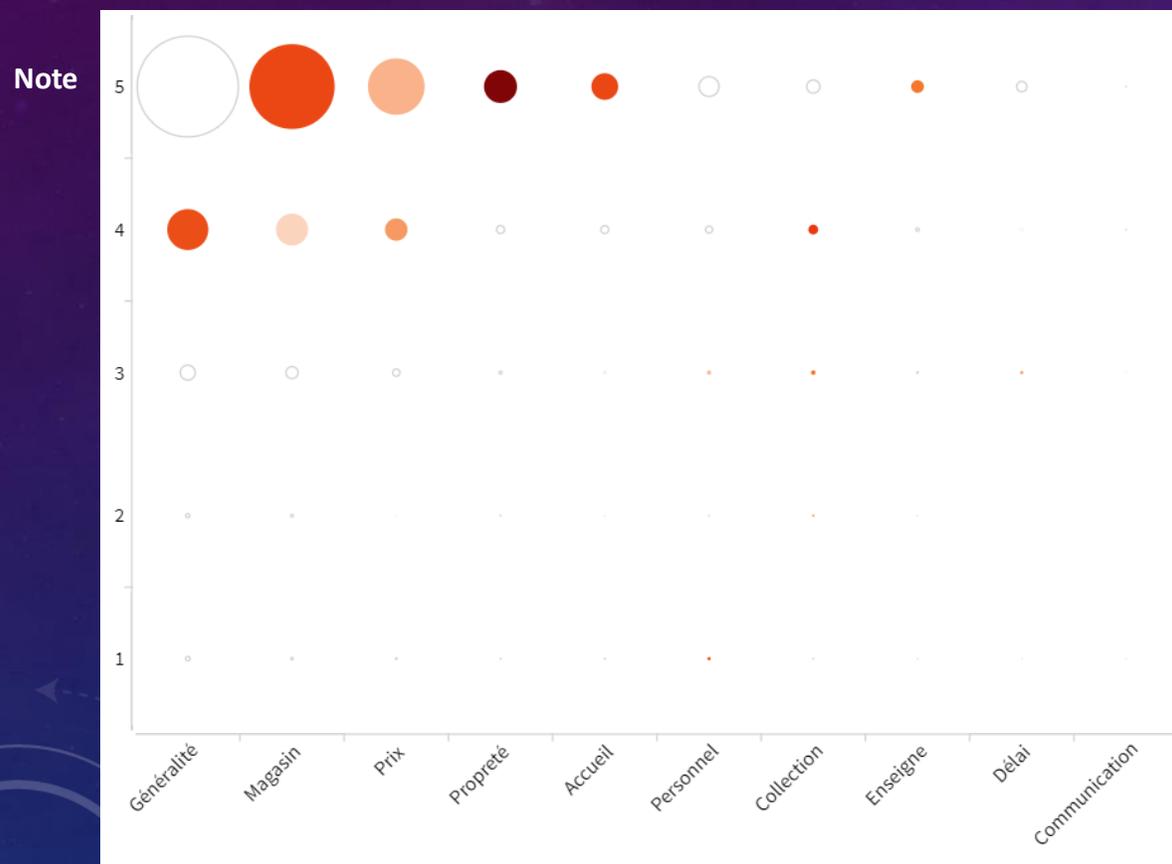


# CONCEPT DE SURREPRÉSENTATION - *HEATMAPS*

- Deux axes ou variables différents
- Comptage statistique des croisements entre les deux axes
- Détection de corrélations ou surreprésentations
  
- Exemple 1 : **Couleur des Yeux** versus **Problèmes de vue**
- Exemple 2 : **Age de la personne** versus **Problèmes de vue**



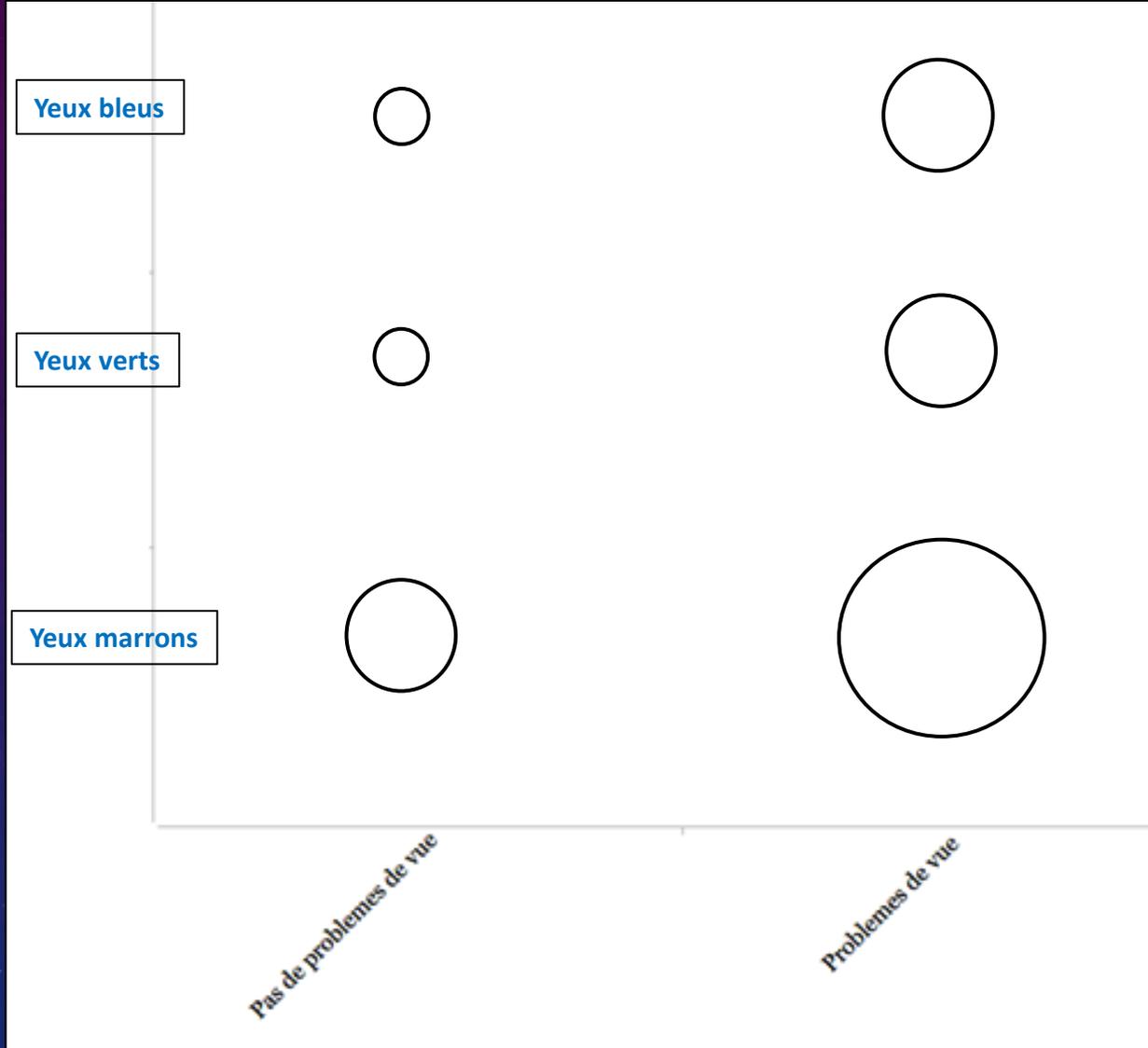
# PRINCIPES



- La **heatmap** ou « carte de chaleur » permet de croiser deux axes d'analyse :
  - Deux facettes (voir exemple ci-contre)
  - Une facette et l'axe du temps
  - Une facette avec elle-même
- La **taille** du rond représente le volume de documents associé  
 Dans l'exemple ci-contre, le plus gros rond est sur l'intersection entre « 5 » et « généralités ». Cela veut dire que les documents qui ont la note 5 et le concept « Généralité » sont plus nombreux que ceux qui ont d'autres notes et/ou d'autres concepts.
- L'intensité de la **couleur** du rond représente un score qui matérialise la **corrélation** ou surreprésentation du croisement concerné
- La taille du rond ne représente pas la corrélation et de ce fait, souvent les heatmaps sont représentés avec une taille de rond artificiellement uniforme pour faciliter la visualisation des ronds les plus petits



# EXEMPLE 1 : COULEUR DES YEUX VS PROBLÈMES DE VUE

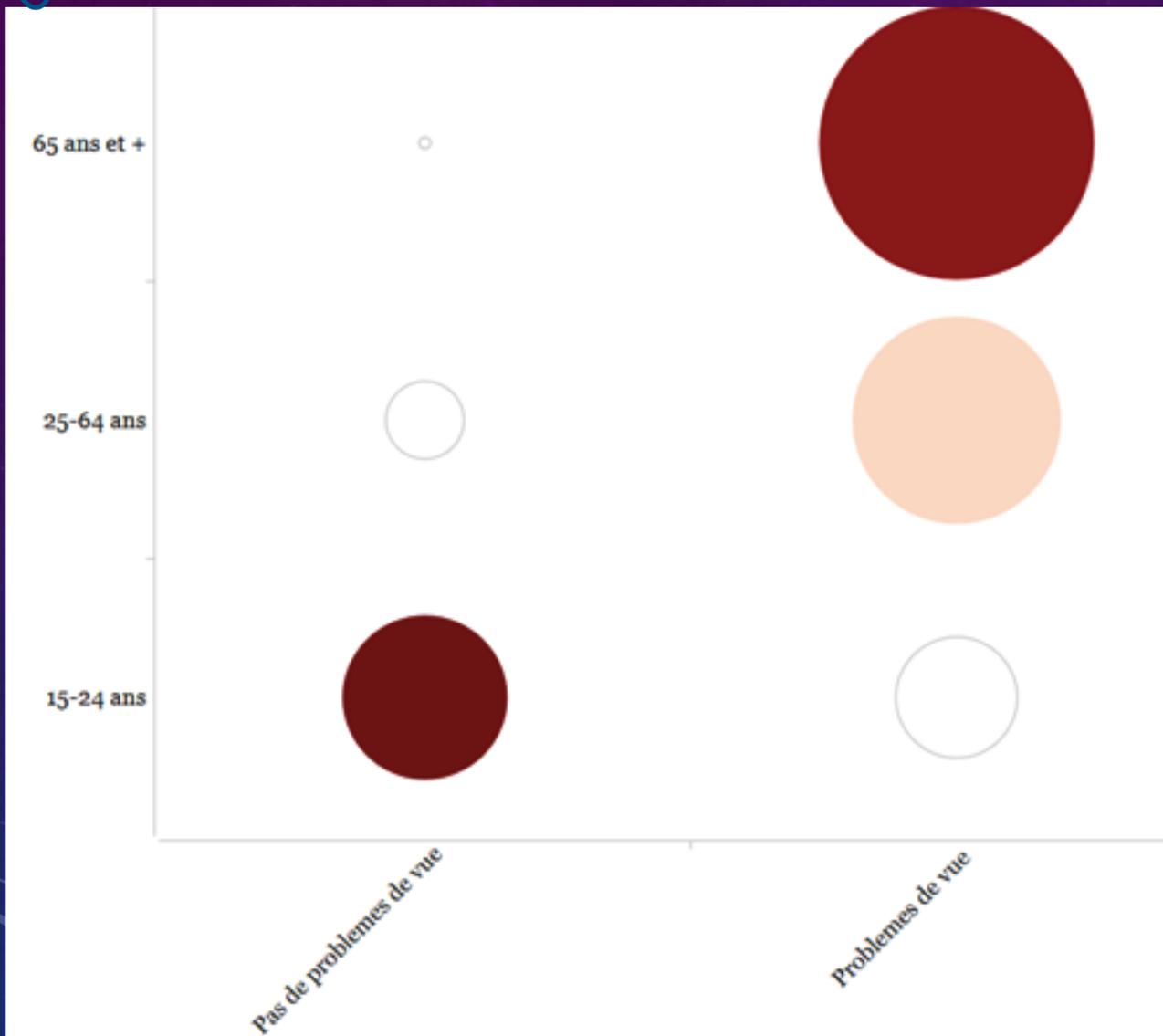


Il y a plus de personnes dont les yeux sont marrons que bleus et verts.

Mais pour chaque couleur, la proportion des personnes ayant des problèmes de vue est de x3 approximativement.

Il y a davantage de personnes aux yeux marron qui ont des problèmes de vue, cependant il n'y a pas de corrélation entre la couleur des yeux et les problèmes de vue.

## EXEMPLE 2 : AGE DE LA PERSONNE VS PROBLÈMES DE VUE



La proportion de personnes âgées qui ont des problèmes de vue est très grande (>95%).

La proportion de personnes entre 15 et 24 qui ont des problèmes de vue est plus réduite (<40%).

Il y a une surreprésentation des problèmes de vue chez les personnes âgées.

Inversement on peut remarquer qu'il y a une surreprésentation de l'absence de problèmes de vue chez les jeunes.

qt1t1t1t1t1  
dt1t1t1t1t1t1  
dt1t1t1t1t2  
trt1t1t1t1t1t1 qt1t1t1t1t1t1  
dt1t1t1t1t1 dt1t1t1t1t1t1t1  
qt1t1t1t1 trt1t1t1t1 ut1t1t1t1t1t1  
nd  
trt1t1t1t1t1  
trt1t1t1t1t1

Maman-enfant

Approche naïve : comparaison des fréquences simples

Selon cette approche, dt1t1t1t1t1 est une des plus fréquentes en présence de la relation maman-enfant

Mais : voir slide suivante

qt1t1t1t1t1  
dt1t1t1t1t1t1  
dt1t1t1t1t2  
trt1t1t1t1t1t1 qt1t1t1t1t1t1  
dt1t1t1t1t1 dt1t1t1t1  
qt1t1t1t1 trt1t1t1t1 ut1t1t1t1t1t1  
nd  
trt1t1t1t1t1  
trt1t1t1t1t1

Mais dt1t1t1t1t1 est aussi fréquent en cas d'absence de cette relation

Maman-enfant

trt1t1t1t1  
dt1t1t1t1t1 qt1t1t1t1t1t1 trt1t1t1t1 qt1t1t1t1t1  
dt1t1t1t1t1t1 trt1t1t1t1t1t1t1  
qt1t1t1t1t1 nd trt1t1t1t1t1t1  
dt1t1t1t1t1t1  
ut3t1t1t1t1 ut3t1t1t1t1  
dt1t1t1t1t1t1 dt1t1t1t1t1t1 dt1t1t1t1t1t2  
ut1t1t1t1t1t1 ut1t1t1t1t1t1t1 ct1t1t1t1t1 ut3t1t1t1t1  
dt1t1t1t1t1t2 dt1t1t1t1t1t2  
ut1t1t1t1t1t1 trt1t1t1t1t1 ct1t1t1t1t1 qt1t1t1t1

Pas Maman-enfant

Approche un peu moins naïve (mais fastidieuse) : comparaison des fréquences en positif et en négatif

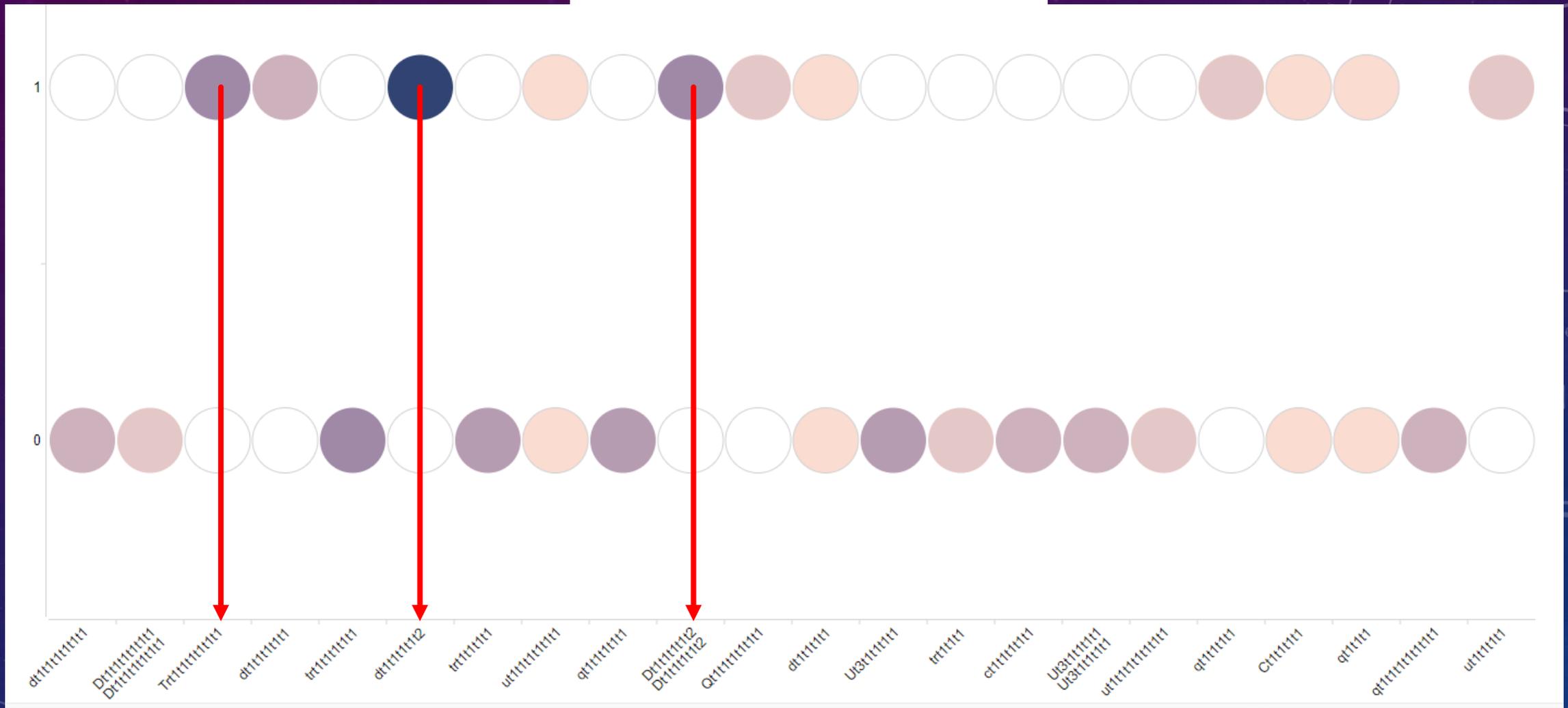
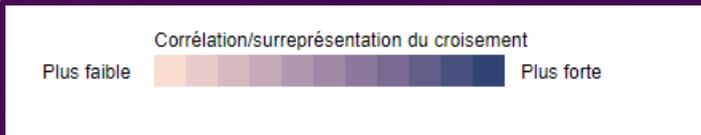
trt1t1t1t1t1t1

Maman-enfant

ut3t1t1t1t1t1  
dt1t1t1t1t1t1  
dt1t1t1t1t2  
ut1t1t1t1t1t1  
trt1t1t1t1t1

Pas Maman-enfant

# Maman-enfant



Approche plus « scientifique » : heatmaps/écart à la valeur théorique