

LAB Apside Brest

Projet Veille des marchés

Besoin client

Vue de l'activité compétences, formations, demandes, offres et placements

Besoin métier

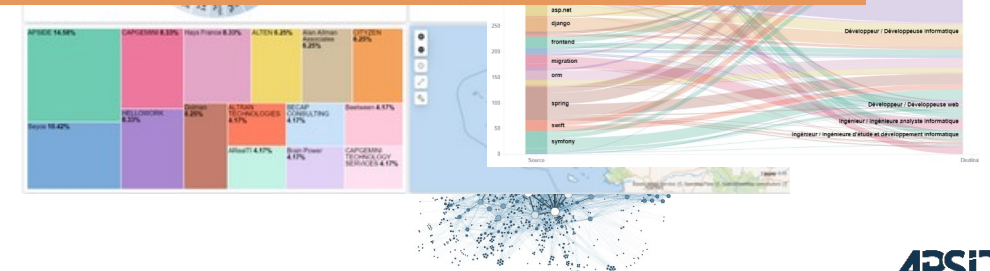
Suivre la donnée du graphe jusqu'à la donnée brute
Suivre le code au long des données, mécaniser la conduite du code par les données

Idée porteuse

Le FAIR dans un ETL « maison / open-source » : traçabilité totale de la donnée
Mécanisation de l'introspection et du marquage du code et de ses usages

Innovation

Navigation synoptique dans le code et les données
Sélection multi points de vue à N dimensions



Sommaire

Clefs et éléments de data science

- Mots clefs - Glossaire
- Principes structurants
- Utilité, cibles et domaines d'applications

Le projet concret

- Veille des marchés
- Cycle de production de tableaux de bord
- Architecture applicative
- Collecte et système de production de données
- Principes de l'UI :
 - Le synoptique et la loupe
 - Design

Quelques exemples

- Normalisation et catégorisation vs filiales sur les données missions d'ApsidePASS
- Algo et méthode pour une approche synoptique
- Visualisation par Sankey (graphe Fluviale)

UI Data Sciences en quelques planches

- Interface principale (environnement de navigation et d'édition)
- Sélecteur de données
- Les menus et les types de données

Plannings

- Roadmap stages et POC
- Cycle AGILE R&D et Data Sciences

Mots clefs et notions de data science

Glossaire

Abstraction	réduction de la complexité par isolation et schématisation pour une généralisation simplifiée facilitant calcul et raisonnement
Agrégation	technique de constitution de données de synthèse par catégorisation et par approximation (ex : de 10 données on crée une données de valeur/poids 10)
Audit	technique d'analyse et processus d'étude des systèmes (sociaux, mécaniques et calculatoires) par leurs composantes, rôles, ressources, activités, processus, dynamiques et individus
Biais	caractéristique d'un point de vue focalisé et d'un raisonnement raccourci édulant, contournant ou résumant une partie du cheminement et réduisant le processus d'analyse
Capteur	moyen matériel d'observation d'un changement d'état dans un contexte physique ou virtuel
Cartographie	techniques d'écriture de représentations symboliques destinées à décrire et enrichir un contexte (ex : carte des risques, carte marine, carte des rôles)
Catégorisation	classement par critères plus ou moins arbitraires permettant de réduire la complexité d'un ensemble de mots/éléments
Complexité	notion s'évaluant par le nombre d'éléments, d'interactions entre éléments, de formes d'éléments et de formes d'interactions constituant un système
Convergence	phénomène de transformation réduisant la complexité d'un système en une synthèse stable et couvrante de l'ensemble de ses parties actives et porteuses
Convolution	phénomène de transformation schématique appliquant un motif simplifié à une partie d'un graphe complexe
Data science	domaine et techniques de collecte d'information, de manipulation, de production et d'inférence incluant l'apprentissage automatique et les systèmes experts
Décisionnel	techniques d'aide à la prise de décision, depuis la collecte et l'analyse jusqu'au tri et à la priorisation voire à l'action automatisée
Données fossiles	données protégées principalement liées au fonctionnement comme les listes de références incluant des échantillonnages pour réévaluation du système
Émergence	technique détournant l'adage « le tout est plus grand que la somme de ses parties » avec l'approximation et les biais visuel (reconnaissance d'un schéma, une forme...)
Émergence visuelle	processus de mise en évidence plus ou moins contrôlée de schémas/formes visuels par la synthèse, les effets de seuil et l'association cf. théorie de Gestalt
ETL vs ELT	système d'import de données par extraction, transformation et chargement la distinction entre ETL et ELT est où se place la charge de calcul et si les données sont répliquées
Étymologie	science de l'origine des mots, leur histoire, leur construction et leur évolution
FAIR data	principe fondateur structurant d'interopérabilité (Findable, Accessible, Interoperable, Reusable)
Géographie	techniques de sciences sociales décrivant le rapport des sociétés à leurs espaces
Géomatique	techniques de calcul et de représentation des données géographiques
GED – GEIDE	système de gestion électronique de documents, indexation, suivi de transformation, gestion des processus sur les documents, alerte de changement d'état, automatisation
GIS/SIG	système d'information géographique
Glissement	technique d'application progressive d'un calcul d'agrégation durant le parcours d'un graphe, transformation sémantique détachant le sens d'un mot de son étymologie
Graphe alluvial	(CosmoGraph, Sankey) représentation graphique répartissant une liste dans une autre et montrant les volumes de la première dans la seconde
Habitus	manière d'être, de raisonner ou de se comporter quasi involontaire voir inconscient lié à la représentation de soi ou d'un rôle à la conformité à un modèle « l'habit fait le moine »
Immersion	technique d'observation participative progressive
Indicateur	source de données ou élément physique révélant un changement d'état dans son milieu
Intuition	conclusion immédiate ou processus de conclusion sans réflexion ou sans réflexion apparente ou consciente
Matrice risque	représentation graphique organisée en tableau à deux dimensions avec un axe gravité et un axe probabilité
Répartition	technique de fragmentation ou de distribution de données globales sur un découpage en sous ensembles selon des critères liés à l'origine et à l'usage des données
Sémantique	étude des mots par le sens, l'origine, la proximité, l'évolution
Supervision	système de suivi et de pilotage de système d'information et d'architecture matérielle
Synonymie	proximité de sens avec éventuellement plus de précision ou de contextualisation
Synoptique	représentation (graphique) simplifiée d'un arbre couvrant le graphe de relations et d'interactions
Tuilage	voisinage incluant un recouvrement des bordures mitoyennes (ex : pour réduire les déformations de référentiel de projection ; ex : pour traiter les points de bordure)
Voisinage	relation de proximité (physique, sémantique, interactionnelle ou structurelle) ex : un carreau a 8 voisins sur un quadrillage

Paradigmes DATA en data science

Essentiels

Données brutes inaltérables	exigence de conservation pour l'analyse à rebours, la preuve et pour une réutilisation
Traçabilité	toutes les étapes les plus insignifiantes sont identifiées et fournissent le cheminement synoptique
Évaluation et Qualité	chaque étape fournit une évaluation de précision, de temps de calcul, d'échantillonnage, de perte, de biais

FAIR

<https://dorum.fr/enjeux-benefices/principes-fair/>

Facile à trouver :

- Déposer les données dans un entrepôt
- Attribuer un identifiant unique et pérenne aux données
- Décrire les données par des métadonnées riches

Accessible :

- Définir les conditions d'accès aux données
- Si possible, rendre les données accessibles librement
- Si les données doivent rester en accès restreint, rendre accessibles les métadonnées pour signaler l'existence des données

Interopérable :

- Privilégier des formats ouverts ou largement utilisés
- Partager le code source du logiciel nécessaire pour lire, traiter, analyser les données s'il a été développé en interne
- Privilégier les standards de métadonnées et les vocabulaires standards
- Indiquer des liens vers d'autres ressources comparables et explorées (autres données, publication...)

Réutilisable :

- Associer une licence de diffusion aux jeux de données
- Associer de la documentation pour décrire les données de façon détaillée, les contextualiser, les rendre compréhensibles...

Clefs et éléments de data science

A quoi sert la data science, pourquoi faire et pour qui ?

Elle sert à : prioriser, synthétiser

- Prioriser et simplifier les choix
- Synthétiser et relier les effets et les perspectives
- Identifier les singularités et les classes de population
- Analyser les évolutions et rapprocher des cas similaires
- Approximer les impacts et leurs portées

Pourquoi faire : aide à la prise de conscience et ciblage

- Observation et prise de recul sur les faits et les données
- Se détacher des biais et habitus contextuels
- Déterminer des cibles d'analyse et d'audit
- Modéliser pour optimiser et accompagner

Pour qui :

- Tous ceux qui doivent décider, optimiser et prioriser à tous niveaux

Data science : de la supervision au décisionnel

Acquisition de données en temps réel, événementiel ou régulier (sans distinction absolue)

Analyse à rebours ou prédictive

Suivi des flux, détection et prévision d'engorgement

Observation des phénomènes d'interaction, d'effet de bord, de singularité

Cibles et clients de la veille des marchés

Les cibles (données directement utiles) :

- Sociétés et parmi elles nos clients, les prospects, les concurrents, les sociétés de secteurs proches (Marc, François, Clémence)
- Les métiers, rôles, technologies, méthodes, compétences (Clémence, François)
- Les offres d'emplois, les offres de formations et certification, les missions (Clémence, Marc)

Les clients :

- Utilisateurs :
 - décideurs
 - consultants (affinage des cv, orientation et formation permanente)
- Contributeurs :
 - Design de communication sur les tableaux de bord : ce qu'on montre et comment (Clémence, François)
 - Design d'évolution sur les fonctionnalités et les cibles : ce qu'on fait émerger et ce qu'on prend en compte (Clémence, François, Goulven)
 - Design géomatique et statistique : la symbolique, la portée, le rapprochement visuel, les seuils de visibilité et de population (Marc, François)
- Exploitation :
 - Veille sur les données, les catégorisations, les algorithmes et les calculs (Clémence, François Marc)

Veille des marchés

L'objectif : un **tableau de bord**

sur **une analyse continue** des données territoriales
pour **orienter la prise de décision**

Comment : en cycle (donnée → calcul → présentation)

pose des objectifs :

cible de collecte

besoin décisionnel

Algo/R&D analyse/calcul/design

conduite d'émergence

priorisation et réduction des choix

Livrable : Présentation / objets / listes courtes

Analyse des objectifs :

Adhésion **vs** rejet (**métier final**)

Evolution suivante **vs** Reprise de cycle

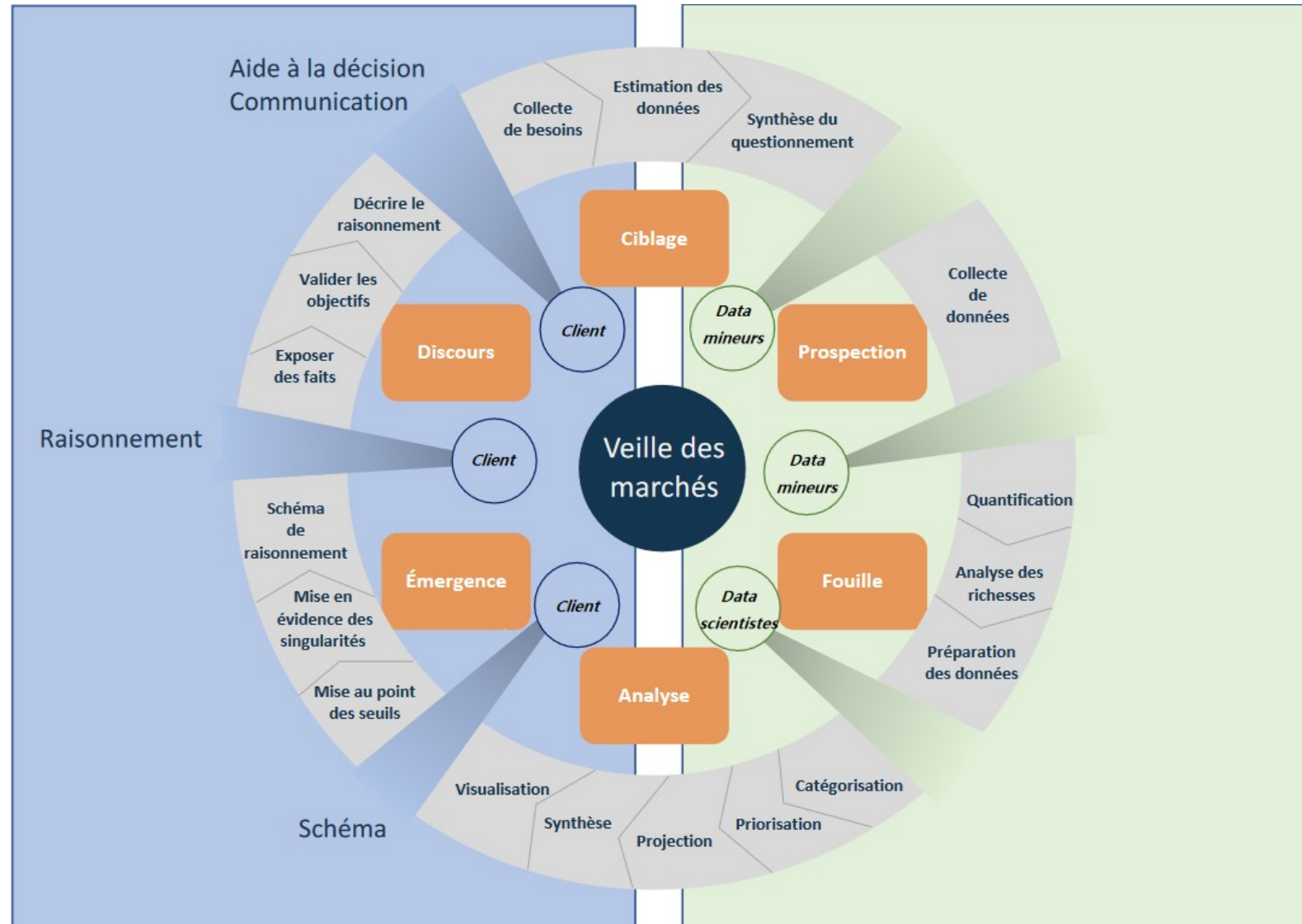
Les cibles



Moteur de tableau de bord

Data Science et décisionnelle

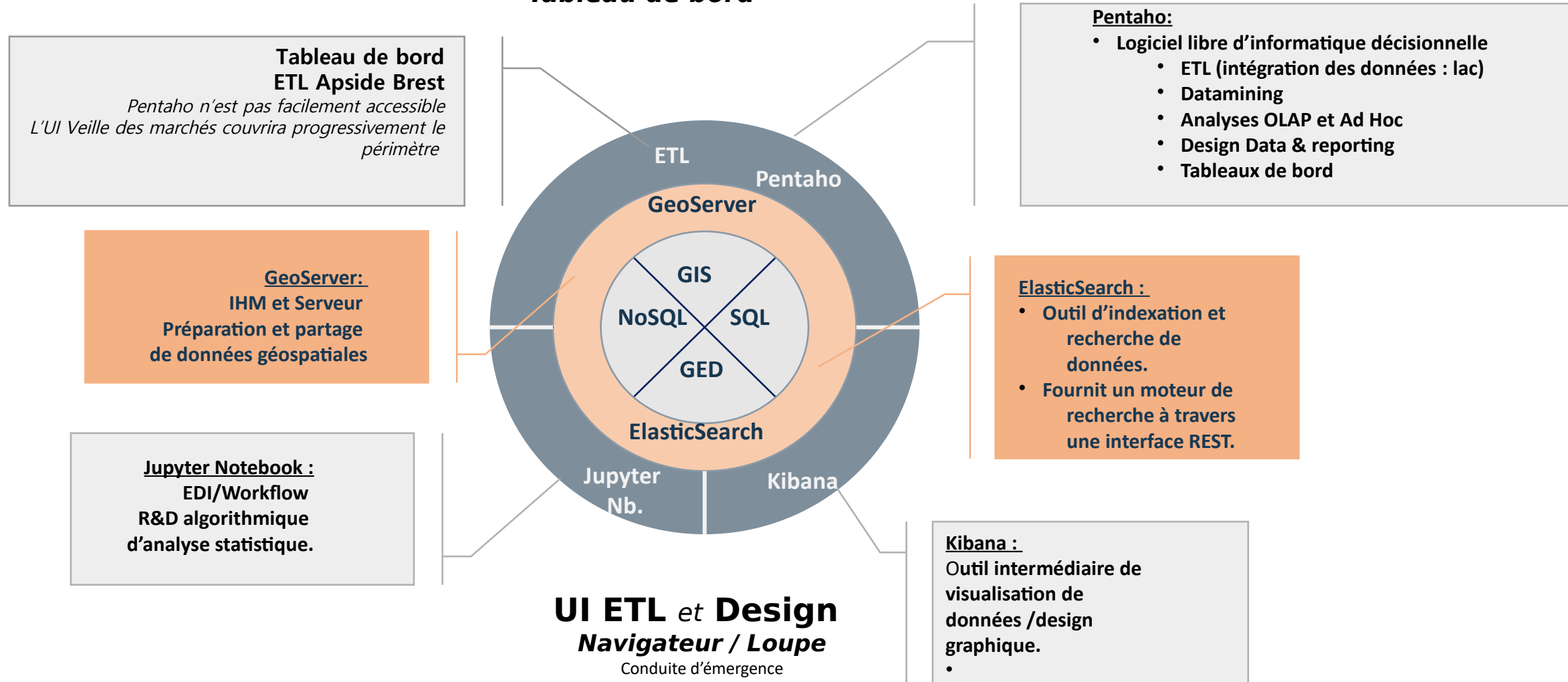
la méthode les rôles



Mise en place d'un tableau de bord

Architecture applicative

UI Veille des marchés Tableau de bord



Architecture du lien synoptique

Scénarisation Automatisation

des sélections et des transformations

Besoins :

- suivre/construire les liens (synoptique) depuis la donnée jusqu'à l'objet du tableau de bord
- permettre le passage de l'UI de la donnée à l'algo (code) et vis versa
- guider graphiquement la sélection
 - d'une donnée
 - d'un algo/code
- mémoriser en continue
 - des données sélectionnées depuis le code source (introspection dynamique~debug)
 - des données ciblées ou source du genre url, api, base, fichier, collection...

Solution α proposée : semi automatisation de l'extraction des marqueurs (mots clef et expressions)

- imposer en douceur des bonnes pratiques par des automates d'analyse « git/hook »:
 - tests unitaires, fonctionnels
 - commentaires utilisables automatiquement (références, tests/démo...)
 - classes et méthodes introspectives
- mécaniser les scripts du gestionnaire de versions (git/hook)
- mécaniser la GED en parallèle au gestionnaire de sources (lier le dépôt)
- construire progressivement un outillage interactif de gestion et de supervision des marqueurs

Automatisations :

- robot d'analyse des environnements jupyter, du git, des log sgbd...
- cibler les références aux :
 - données (match sur champs pré indexés, noms de variables...)
 - sources (API, DB-URI,...)

Semi automatisation :

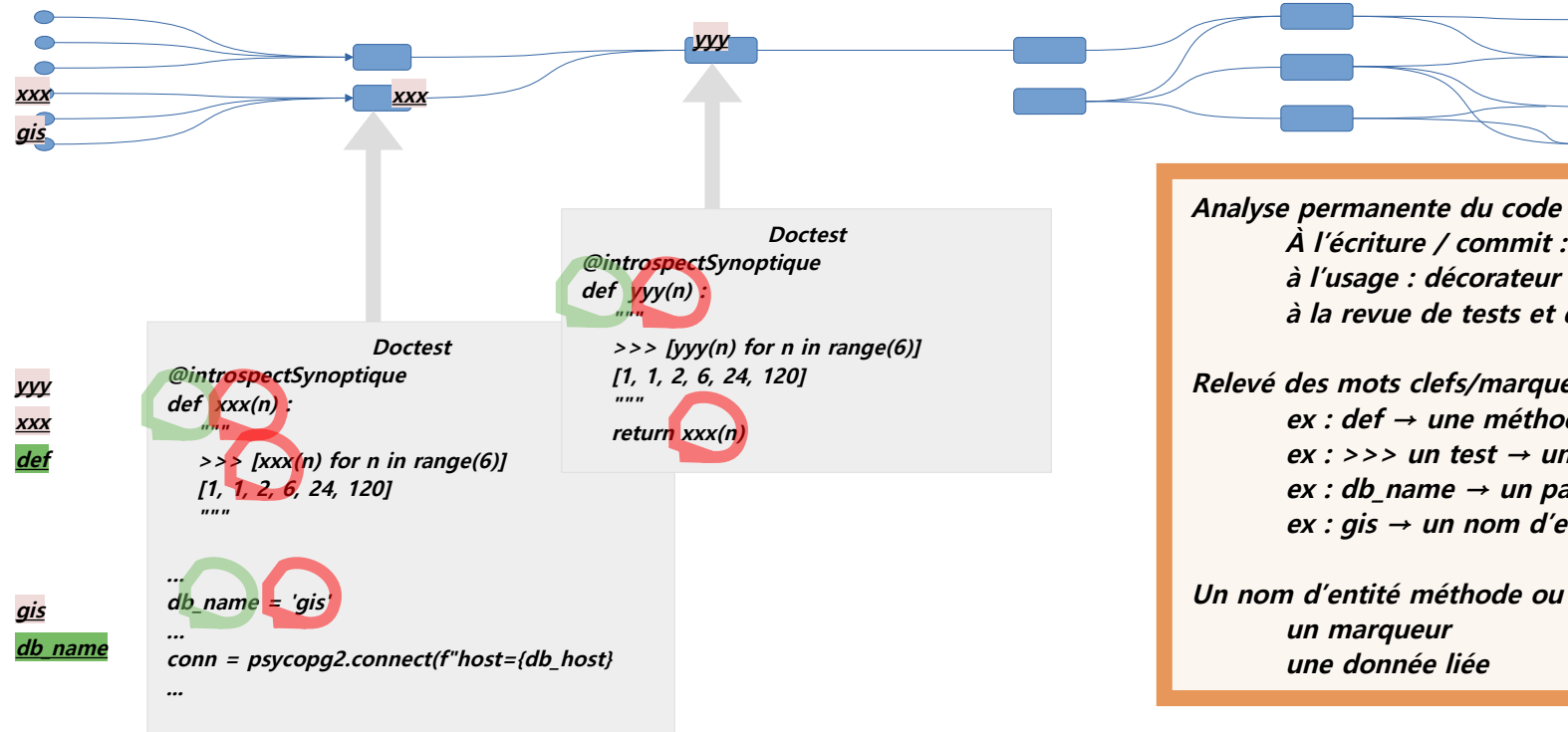
- utilisation massive des doctest/docstring python et des marqueurs

Architecture du lien synoptique

Scénarisation
Automatisation

Doctest
Docstring
Décorateurs

Exemple d'extraction et de liaison synoptique de données



Analyse permanente du code :

À l'écriture / commit : hook de git en local ou api-jupiter

à l'usage : décorateur introspectSynoptique

à la revue de tests et documentation auto : sphinx et doctest

Relevé des mots clefs/marqueurs et de leur contenu variable

ex : def → une méthode

ex : >>> un test → un appel à la méthode

ex : db_name → un param défini en lib ou dans un autre code

ex : gis → un nom d'entité

Un nom d'entité méthode ou base de données... devient :

un marqueur

une donnée liée

Remarque de Marc :

Ça va faire un paquet de données → « la vie des données »

À étudier selon les principes de persistance, protection, vieillissement

JupyterAPI / Git

Architecture logicielle / Projets

Data Analyse Factory

UI

Sous projets front

API

API-Géomatique

Synoptique

Sélecteur

dataViz

UIDataDriven

Arbre couvrant de dépendance/relation allant de la donnée brute jusqu'à la présentation synthétique et symbolique

Outil de recherche, de filtrage et de sélection des données pour toutes les étapes de manipulation des données (nœud éditable du synoptique)

Synthèses visuelles présentant soit l'hétérogénéité et la répartition des données sélectionnées, soit une vision unifiée de ces mêmes données

Outil paramétrable d'analyse automatique et de proposition de visualisation des données

Veille des marchés

Introspection Collecte

Sous projets Back

CollecteurGIT

pyIntrospection

API

API-Géomatique

Analyseur de code branché sur le GIT

- Relevé des mots clés
- Couverture de test et de documentation
- branchement du code d'introspection
- Découpage algorithmique semi automatique
- Revue de convention de nommage

Bibliothèque et automates d'introspection

- API du Hook/GIT pour insérer le code d'introspection
- Décorateurs chargés de relever / analyser les flux de paramètres et de données

Squelette / framework

Data Science Back

Ordonnanceur/Cron

Gestionnaire de tâches

- Régulières
- Déclenchées
- Auto déclenchées

Data Science UI

légende

Groupe (gitlab)

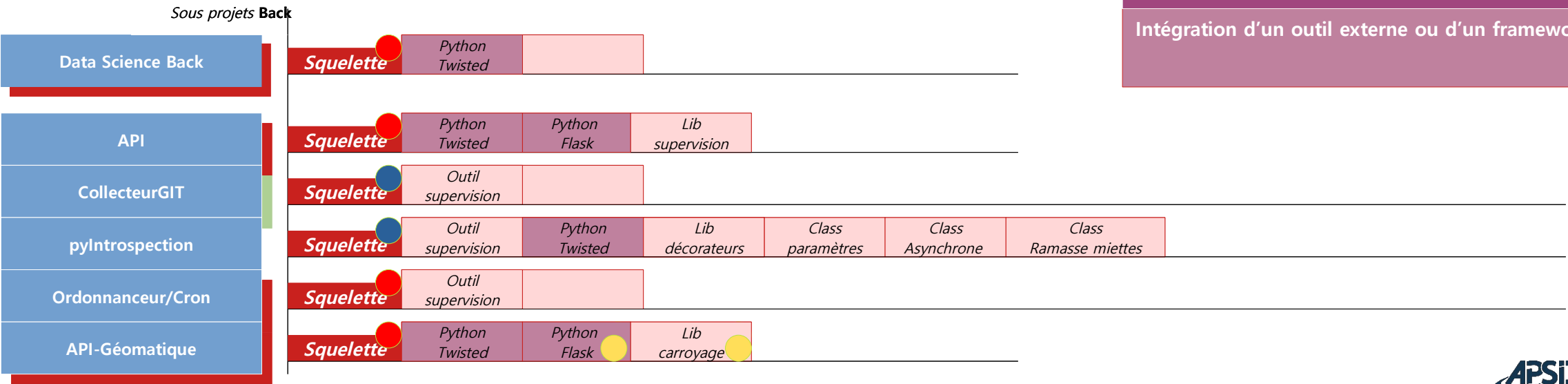
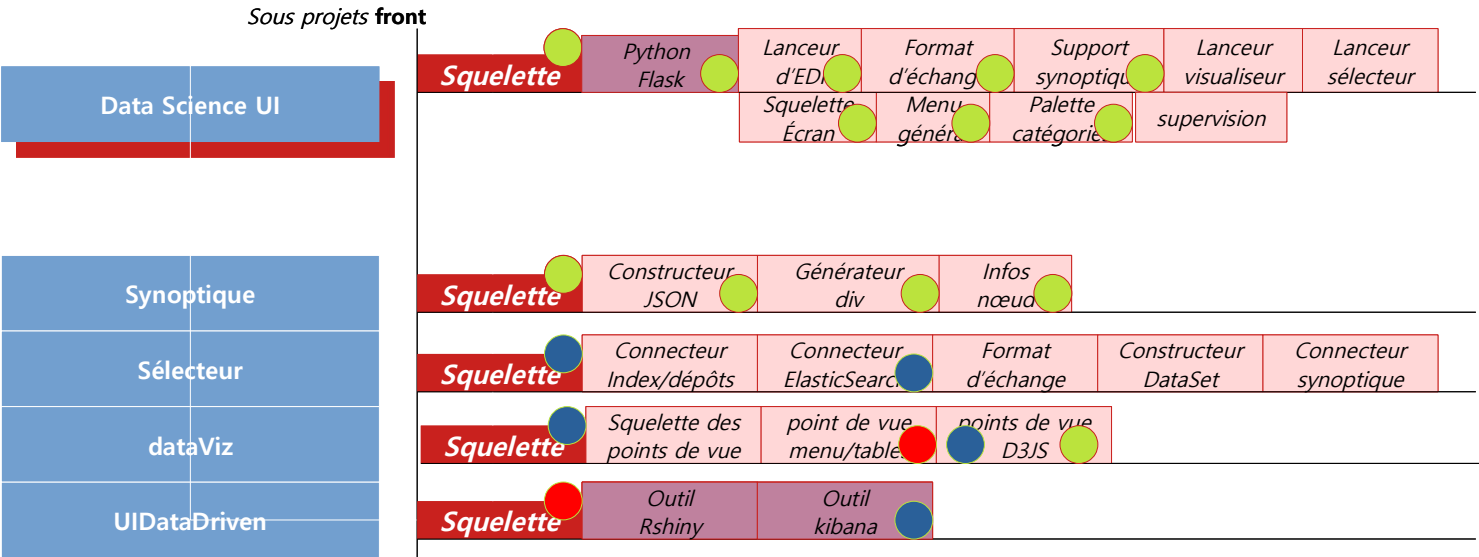
Projet (gitlab)

Appli

Scripts

lib

Architecture logicielle / Projets / Détails



Squelette : organisation/automatisation du projet
Infrastructure du service et ou de l'application

Traité en POC par Goulven

Traité en POC par Marc

Traité par Clémence

Traité par François

Intégration d'un outil externe ou d'un framework

Veille des marchés

UI CLIENT

Navigateur / Loupe

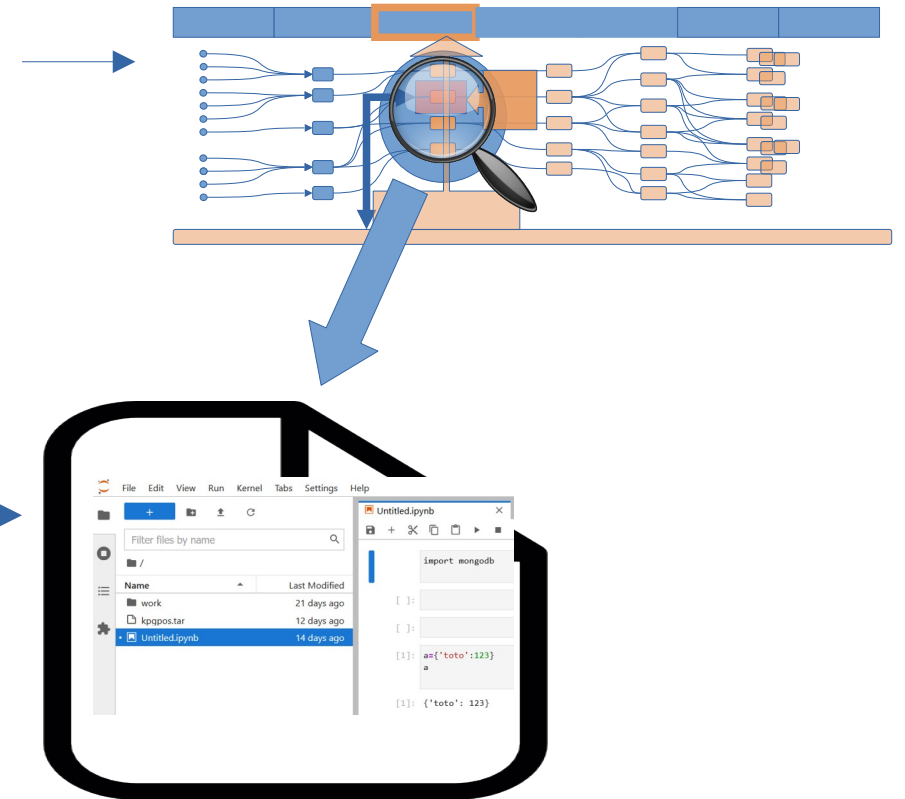
Vision synoptique des données de bout en bout (brute... intermédiaire... publication)

Menu - Loupe

Synoptique des données de la source brute à la présentation
Graphe des données et transformations liées
Accès/interface vers le code
Sélection horizontale de l'étape de transformation
Sélection verticale de l'algorithme calcul-agrégation-projection

Page

IHM contextuelle
éditeur de code
gestionnaire de données
gestion de contenu



Principes

Portail sur les données, le code et les outils spécialisés (jupyter notebook, kibana, GED, ETL...)

- Un menu fixe : les métiers/étapes (données, catégorisation...)
- Un menu synoptique de navigation dans les objets traités/produits
- Une page IDE spécialisée commandée et incrustée :
 - L'UI va manipuler l'IDE par son entrée API et présenter tout ou partie de son interface dans un cadre principal (objet, frame, iframe...)

Squelette

Portail sur les données, le code et les outils spécialisés (jupyter notebook, kibana, GED, ETL...)

- Un menu fixe : les métiers/étapes (données, catégorisation...)
- Un menu synoptique de navigation dans les objets traités/produits (niveau de finition et de fonctionnalités graduels : n°0 listes/tableaux n°<n> 3djs, cosmograph...)
- Une page IDE spécialisée commandée et incrustée :
 - L'UI va manipuler l'IDE par son entrée API et présenter tout ou partie de son interface dans un cadre principal (objet, frame, iframe...)

Formats

- Synoptique : construit sur un json décrivant les liens et les nœuds :
 - Liens amonts : ce que le nœud courant utilise (variables, contenus, méthodes) ex : la catégorie « filiale » utilise le champ company et les données de la table company
 - Liens avals : ce qui utilise le nœud courant ex : company est utilisé par la méthode makeCompany et par la catégorie filiale
 - La portée visible du synoptique est volontairement limitée à ce qu'il est possible de visualiser et d'utiliser sur l'écran IHM
- Page IDE :
 - Objet HTML5 construit et commandé par le synoptique et son gestionnaire
 - Elle est paramétrée selon la nature du nœud sélectionné et encapsule le bon outil positionné sur la bonne étape de présentation et d'édition, ex : makeCompany sur l'éditeur python avec le fichier .py à la ligne de la méthode makeCompany

Data Science

géomatique

Exemples de processus géomatique

Exploration fonctionnelle

Question initiale : répartition d'une donnée individuelle globalisée sur un territoire
exemple => covid.vaccin, covid.grave, covid.deces % de population

Besoin : Préparation d'un support de répartition des données
Cible géomatique : échelle individuelle/familiale/logement

Les données sont associées géographiquement à la commune au département ou à la nation
Elles sont caractérisées par des éléments de contextualisation (niveau de vie, niveau de santé, famille...)

Éléments de solution : Cadastre, logement

Cadastre : découpage territorial fin à l'échelle individu/famille/propriété
Logement : sous découpage individu/famille/habitation

Éléments de sous répartition (précisions)

Coef familiale, niveau de vie, type d'habitat, comorbidité, usages à risque

```
for nom_col in noms_col:
    gdf_dvf[nom_col] = convert_to_float(gdf_dvf[nom_col])

surface_habitable = gdf_dvf[noms_col].sum(axis=1)
gdf_dvf['nb_hab_estime'] = (surface_habitable/1000)*densite_pop_brest
```

```
[198]: choropleth_map(gdf_dvf, column_name='nb_hab_estime')
```

Colour column min value 27.9204, max value 1395.6849551999999, mean value 182.553269787
Hexagon cell count: 31

```
[198]:
```



En cours : à revoir sur le thème

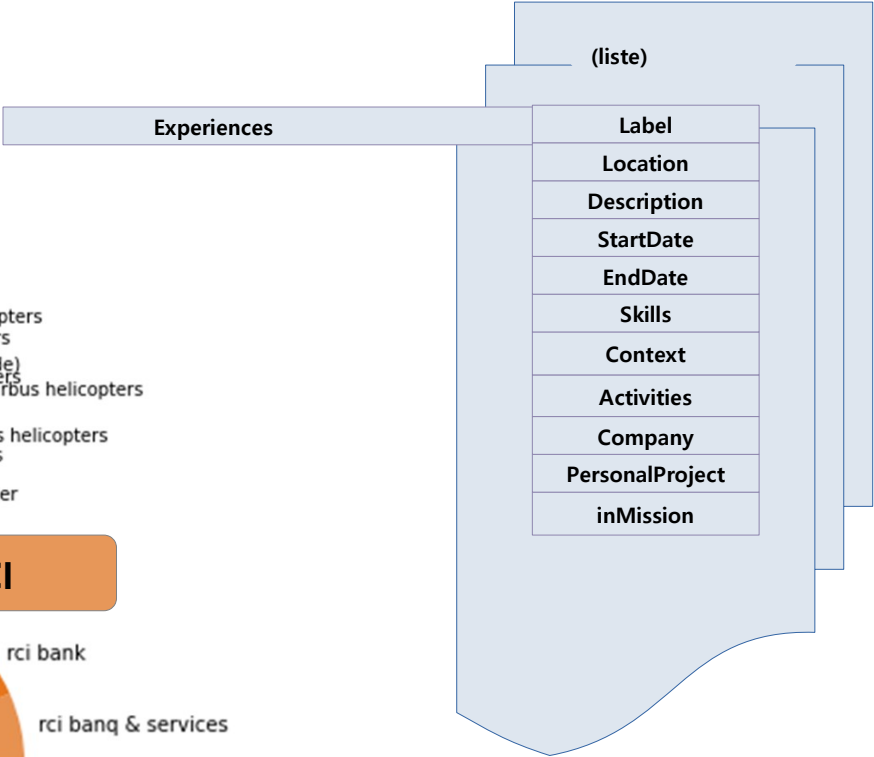
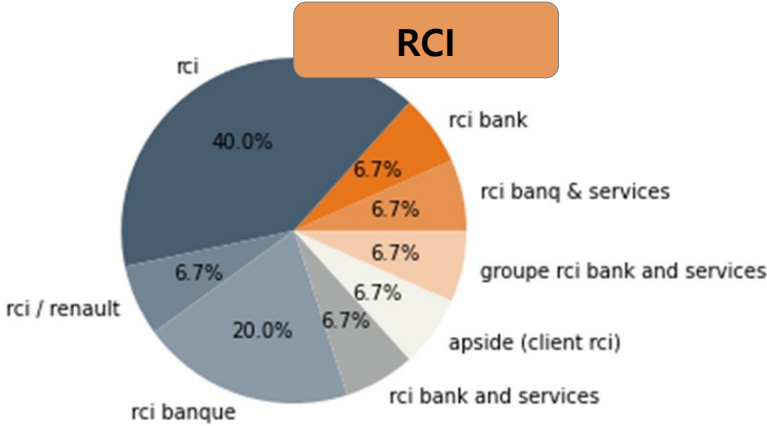
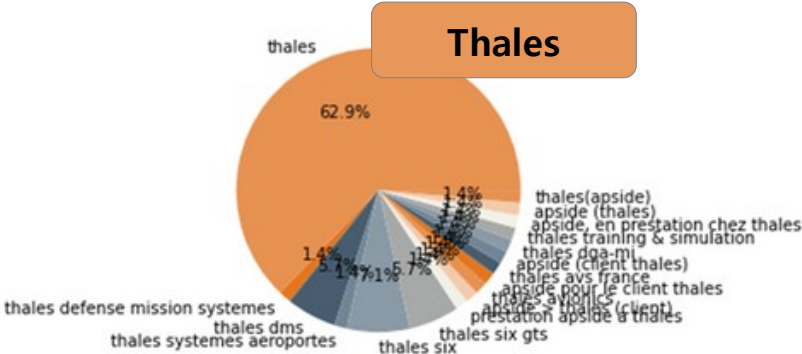
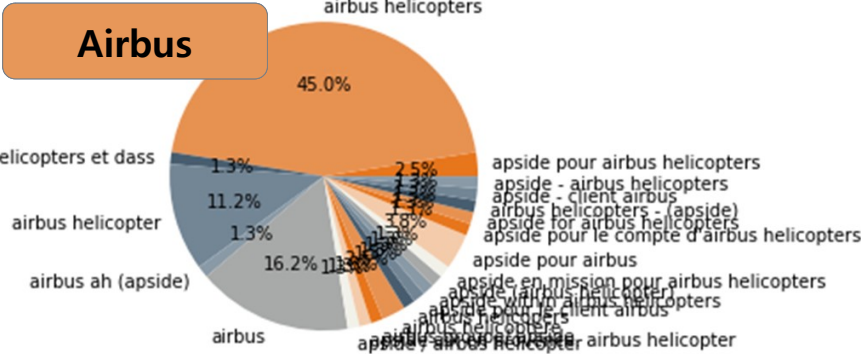
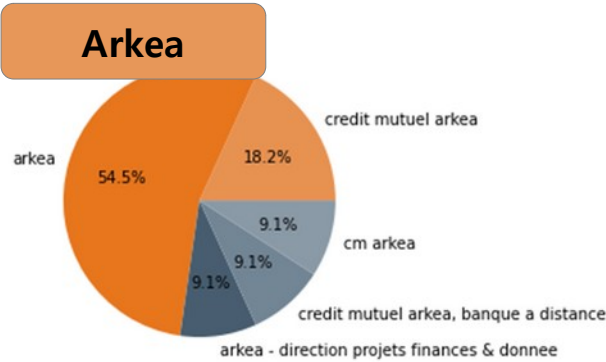
Data Science

prémisses

Micro étude des données APSIDE-PASS

diversité de saisie du nom de client dans les dossiers de compétence

@CClémente

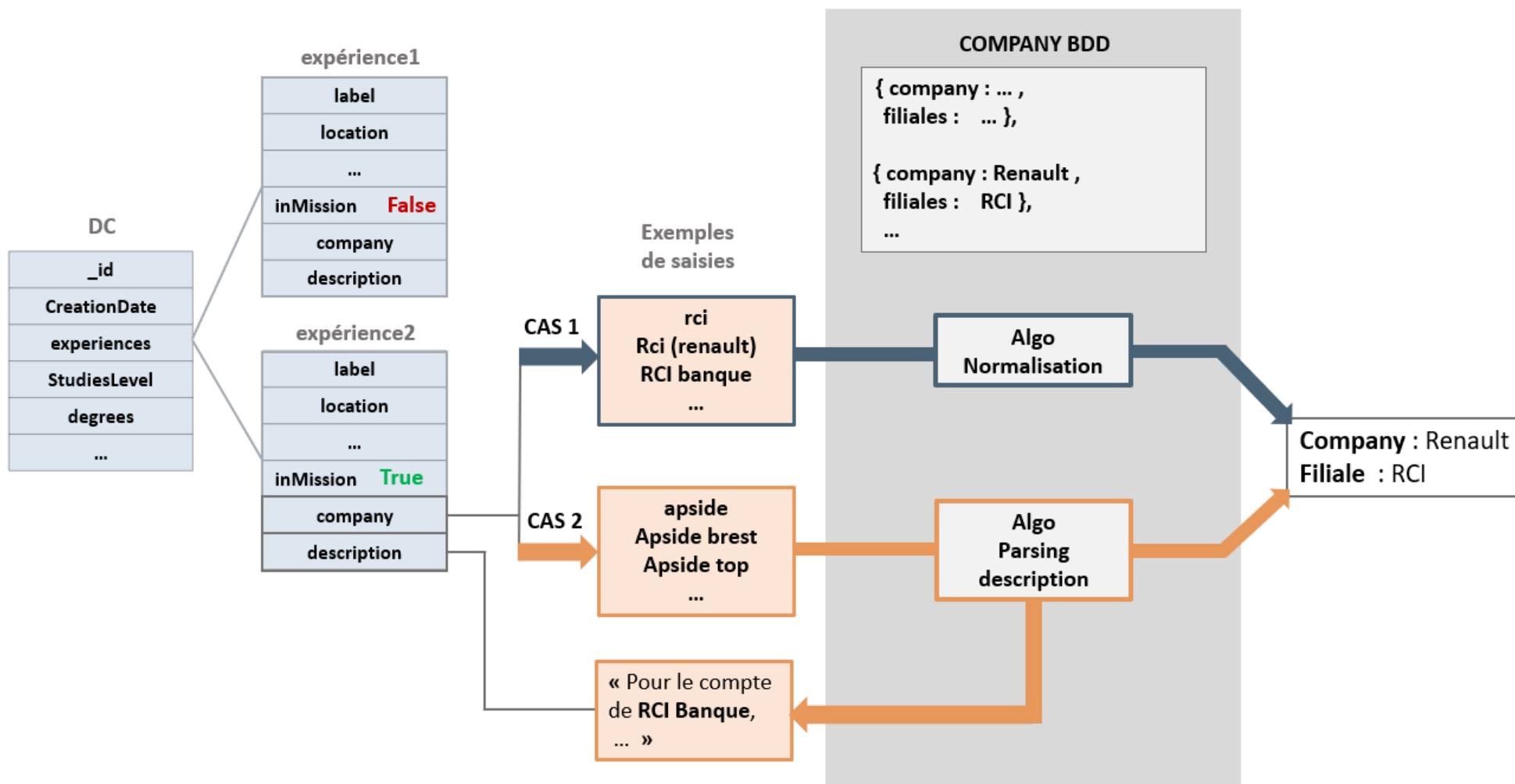


Data Science

prémisses

Micro algo de synthèse de compagnies APSIDE-PASS

@CClémente

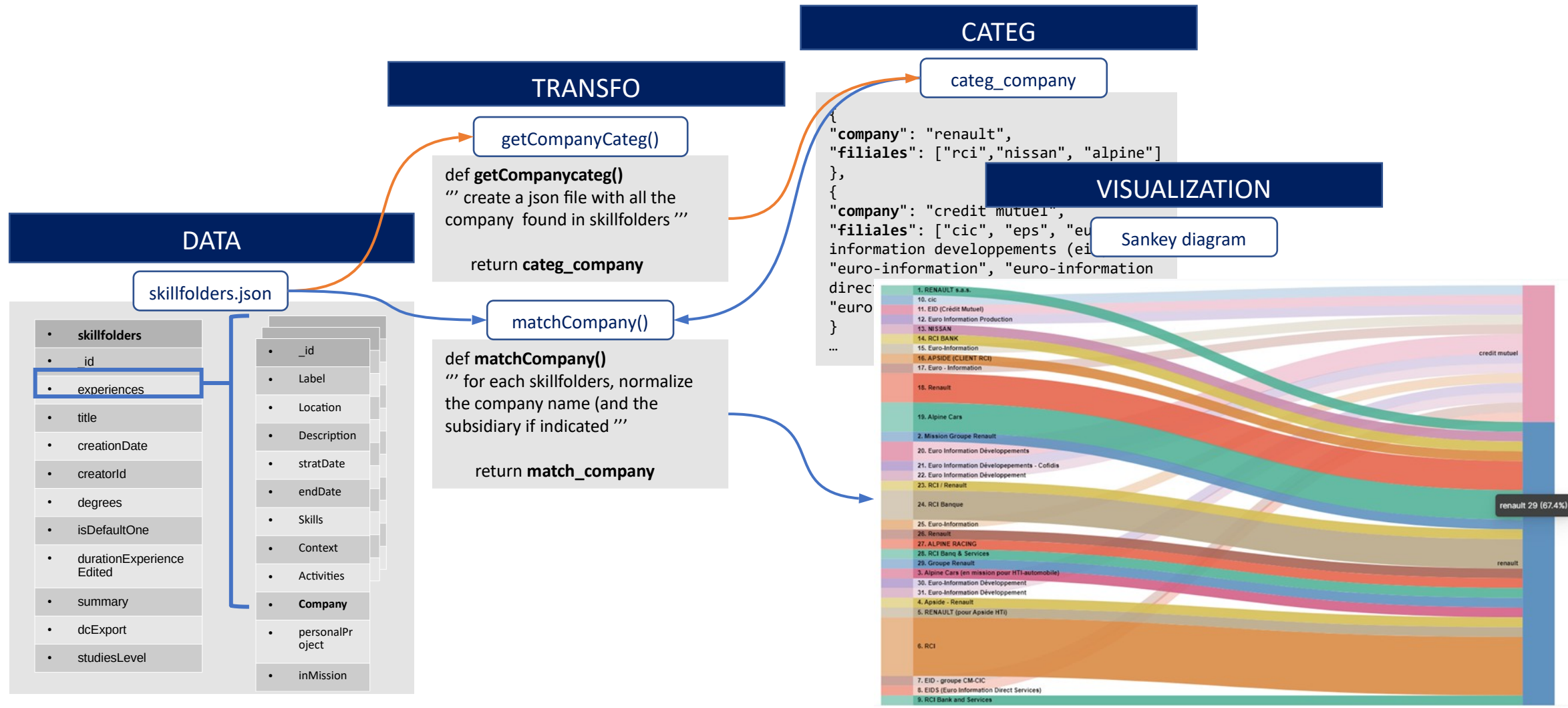


Data Science

prémisses

Visualisation
@CCLémence

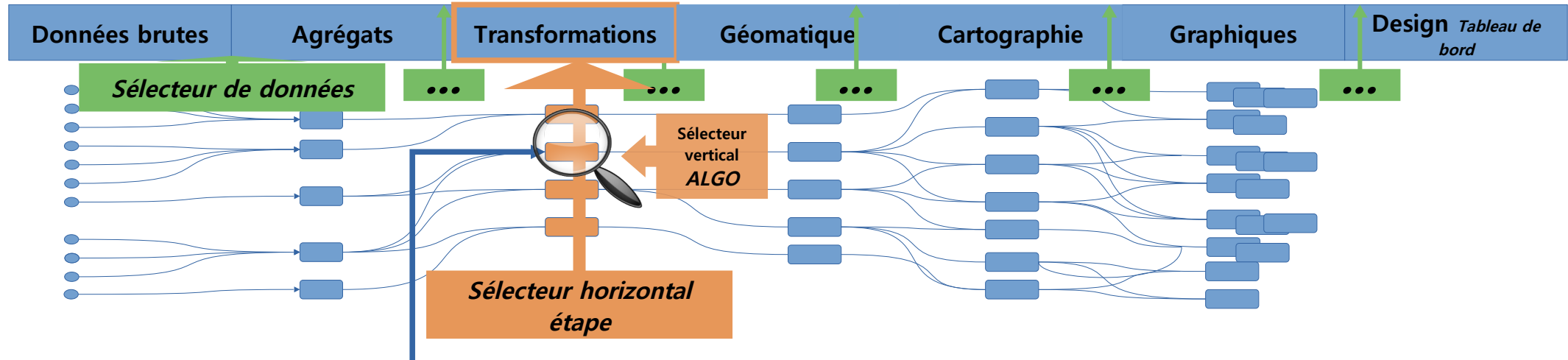
sociétés/clients vs filiales



Veille des marchés

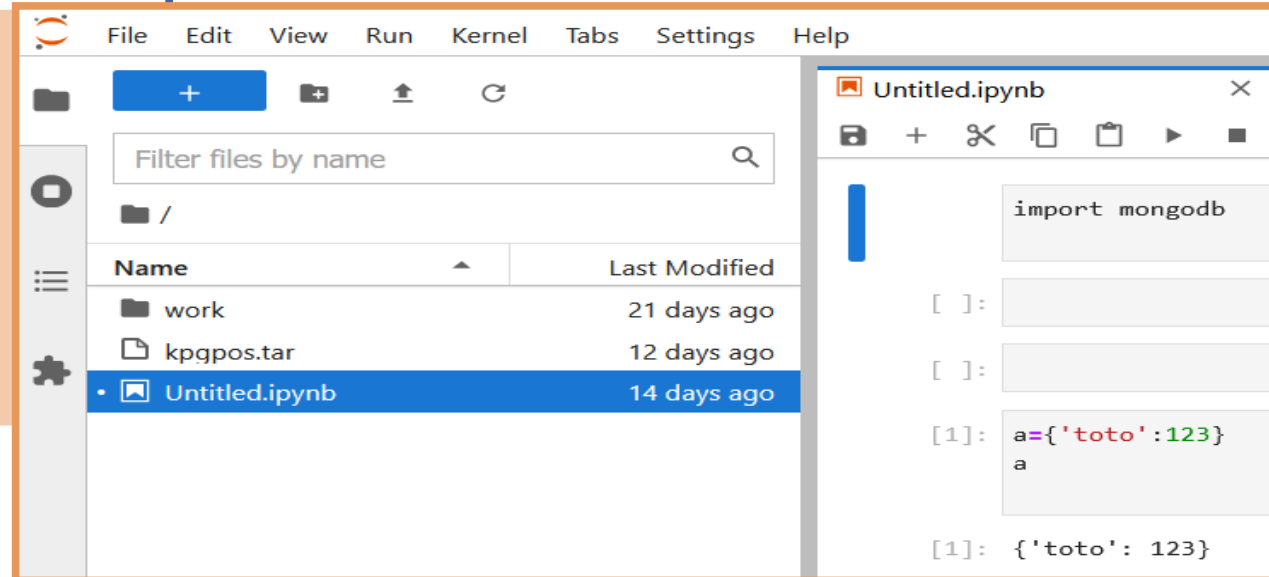
UI DataSci. *Navigateur / Loupe*

intégration des données



UI Agrégation
(précédente)

UI Transformation
(courante)



UI Géomatique
(suivante)

Tableaux de bord

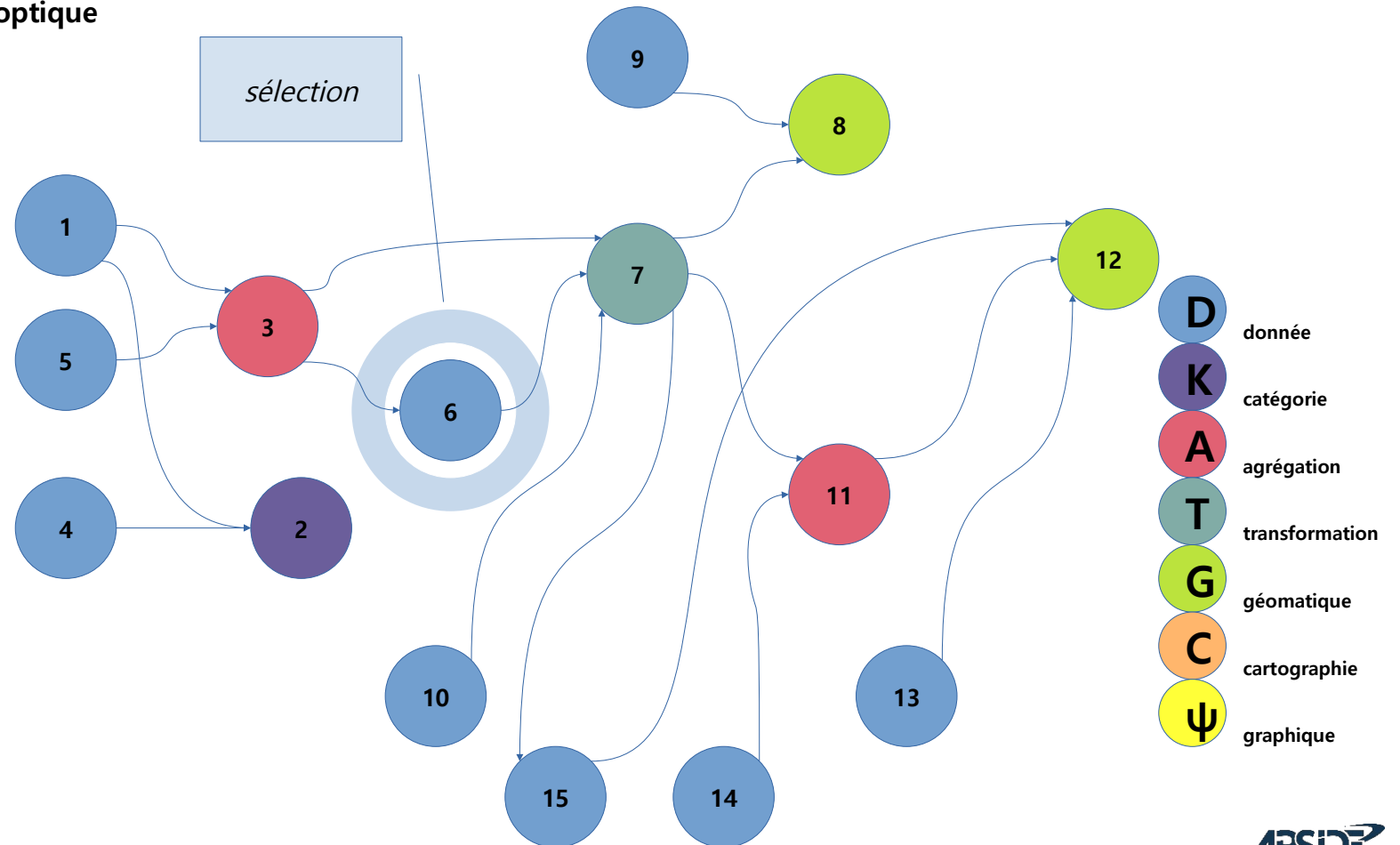
Veille des marchés

Synoptique JSON design driven

Exemple : formats/représentations

JSON transmis par le back pour construire le synoptique

```
[
  {"id":1,"nom":"1""categ":"D","infos": "", "parent": [], "child": [2,3]},
  {"id":2,"nom":"2""categ":"K","infos": "", "parent": [1,4], "child": []},
  {"id":3,"nom":"3""categ":"D","infos": "", "parent": [1,5], "child": [6]},
  {"id":4,"nom":"4""categ":"D","infos": "", "parent": [], "child": [2]},
  {"id":5,"nom":"5""categ":"D","infos": "", "parent": [], "child": [3]},
  {"id":6,"nom":"6""categ":"D","infos": "", "parent": [3], "child": [7]},
  {"id":7,"nom":"7""categ":"T","infos": "", "parent": [3,6,10], "child": [8,11,15]},
  {"id":8,"nom":"8""categ":"G","infos": "", "parent": [7,9], "child": []},
  {"id":9,"nom":"9""categ":"D","infos": "", "parent": [], "child": [8]},
  {"id":10,"nom":"10""categ":"D","infos": "", "parent": [], "child": [7]},
  {"id":11,"nom":"11""categ":"A","infos": "", "parent": [7,14], "child": [12]},
  {"id":12,"nom":"12""categ":"G","infos": "", "parent": [11,13,15], "child": []},
  {"id":13,"nom":"13""categ":"D","infos": "", "parent": [], "child": [12]},
  {"id":14,"nom":"14""categ":"D","infos": "", "parent": [], "child": [11]},
  {"id":15,"nom":"15""categ":"D","infos": "", "parent": [], "child": [12]}
]
```



Veille des marchés

Besoin

Trouver, identifier, évaluer, comparer les données
 ... les ensembles (set) de données,
 ... les flux (disponibilité, fréquence, volumes...)

Le sélecteur de données

Explications et structuration des données marquées (produites ou identifiées)

Impact - portée

Durant tout le cycle des données et au-delà
 - ajout pour un calcul, un graphique, une liste
 - ajout d'une donnée/set comme paramètre

Service

Qualifier et suivre la vie des données
 - suivi de fourniture régulière vers le tableau de bord
 - durée de vie, usages, obsolescence, contextes, fournisseurs, usagers

Vue par Clémence sur <https://www.micropole.com/fr-fr/mapping-micropole/index.html>



Principe de ce sélecteur circulaire

Sélection excentrique

centre vers périphérie
 ressources vers donnée

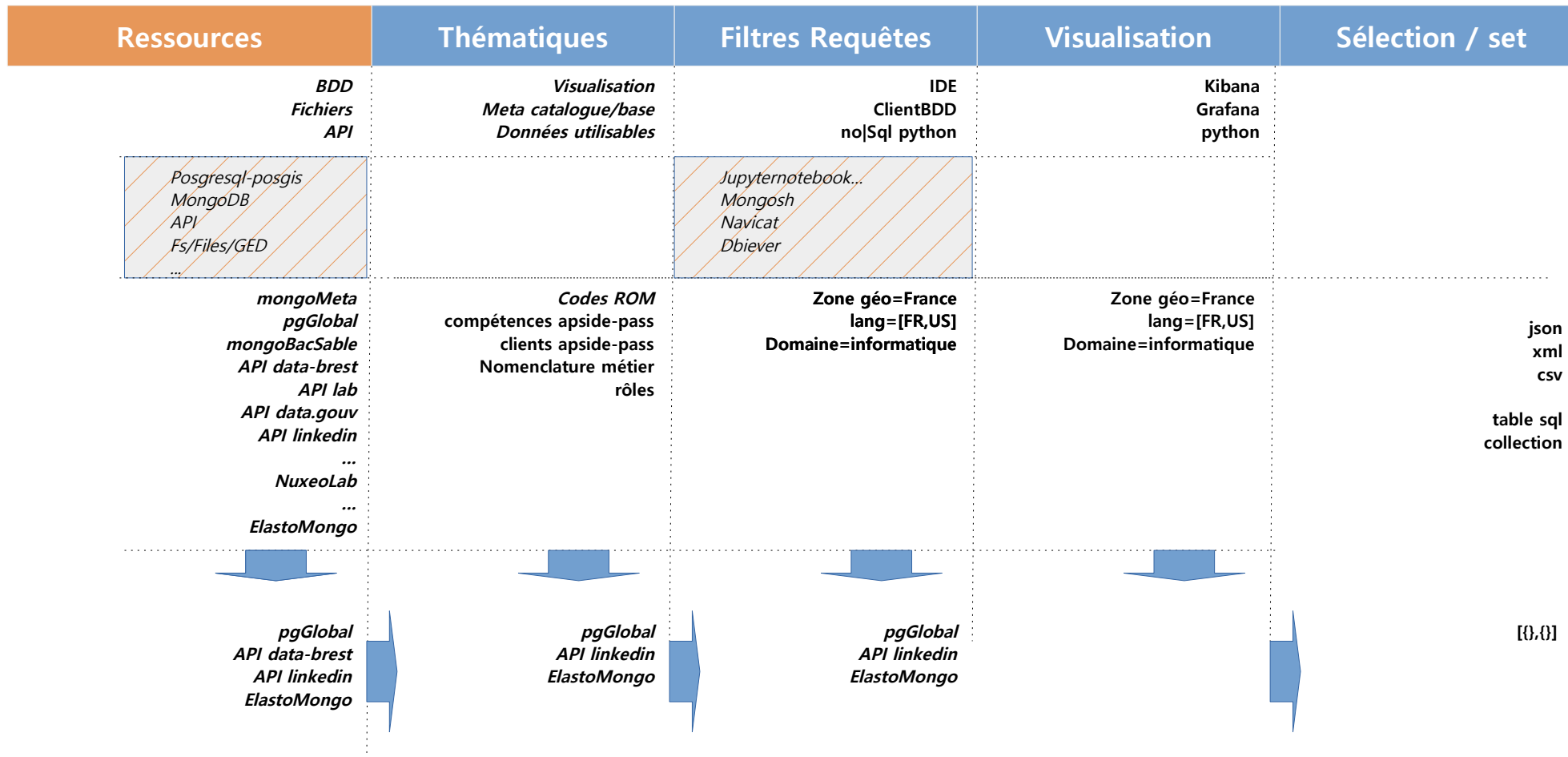
- au centre 0 : bdd, api, fichiers...
- cercle 1 : thèmes, catégories...
- cercle 2 : ensembles, « set »
- cercle 3 : les données retenues => ensemble

Veille des marchés

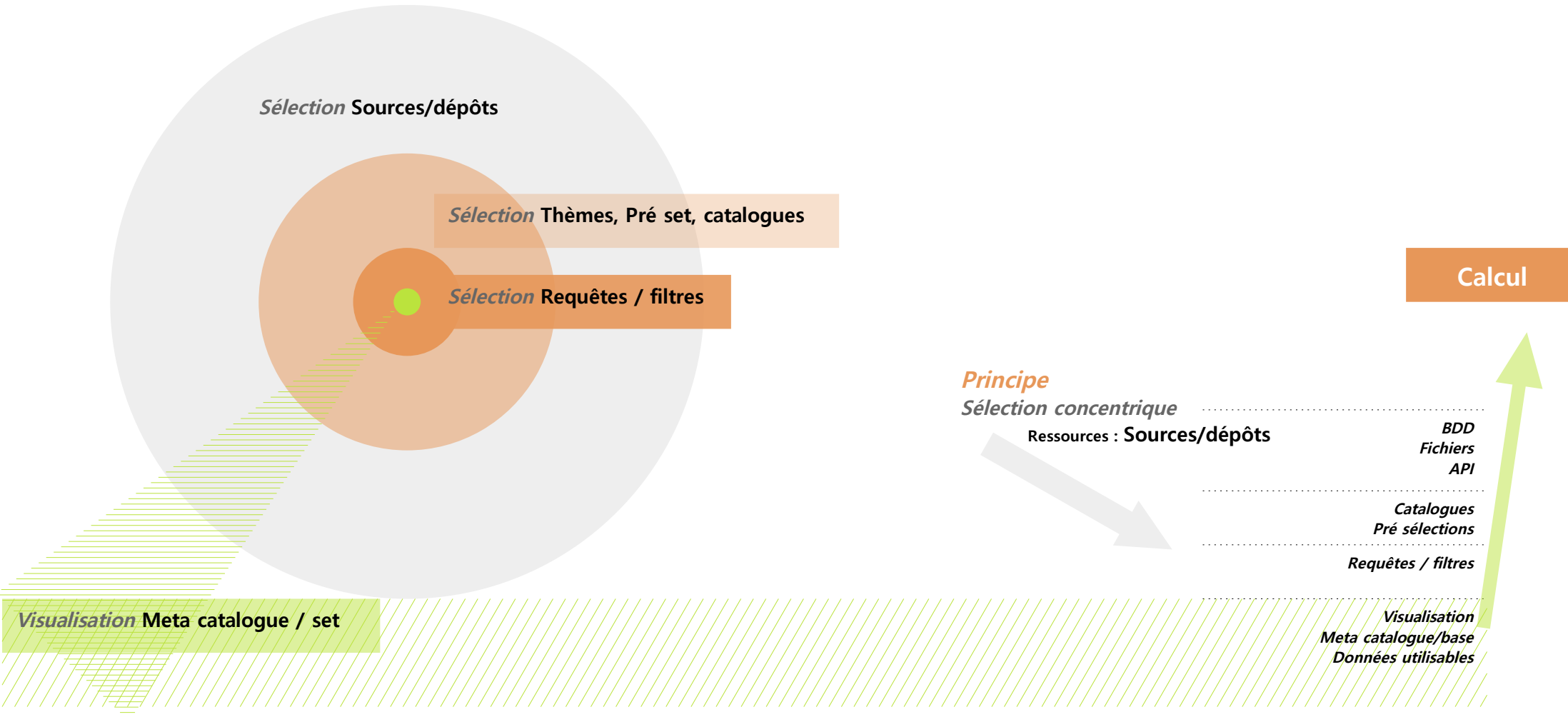
Le sélecteur de données

Outil (toutes étapes) permettant de composer un ensemble de données

Vision classique par niveaux/couches : multi sélection progressive



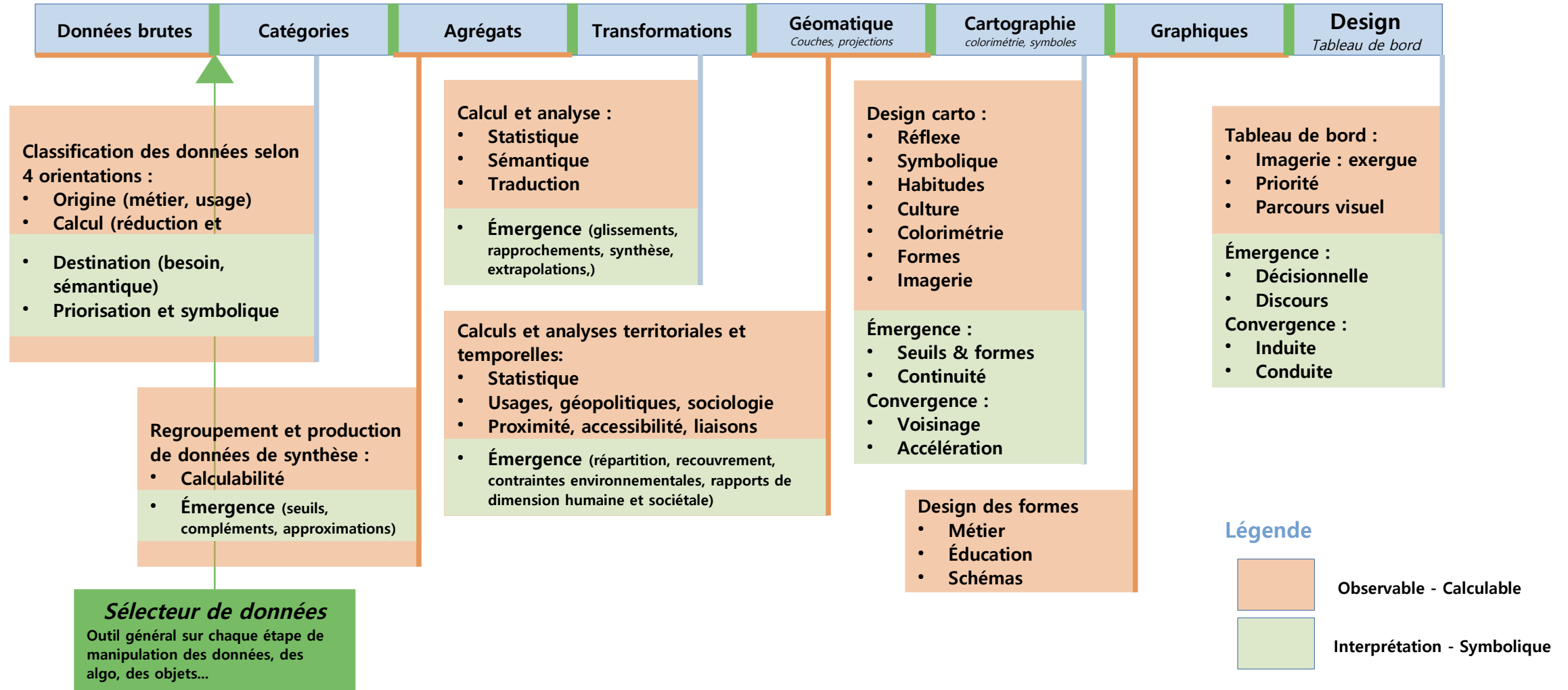
Vision gouvernail : multi sélections concentrique



Veille des marchés

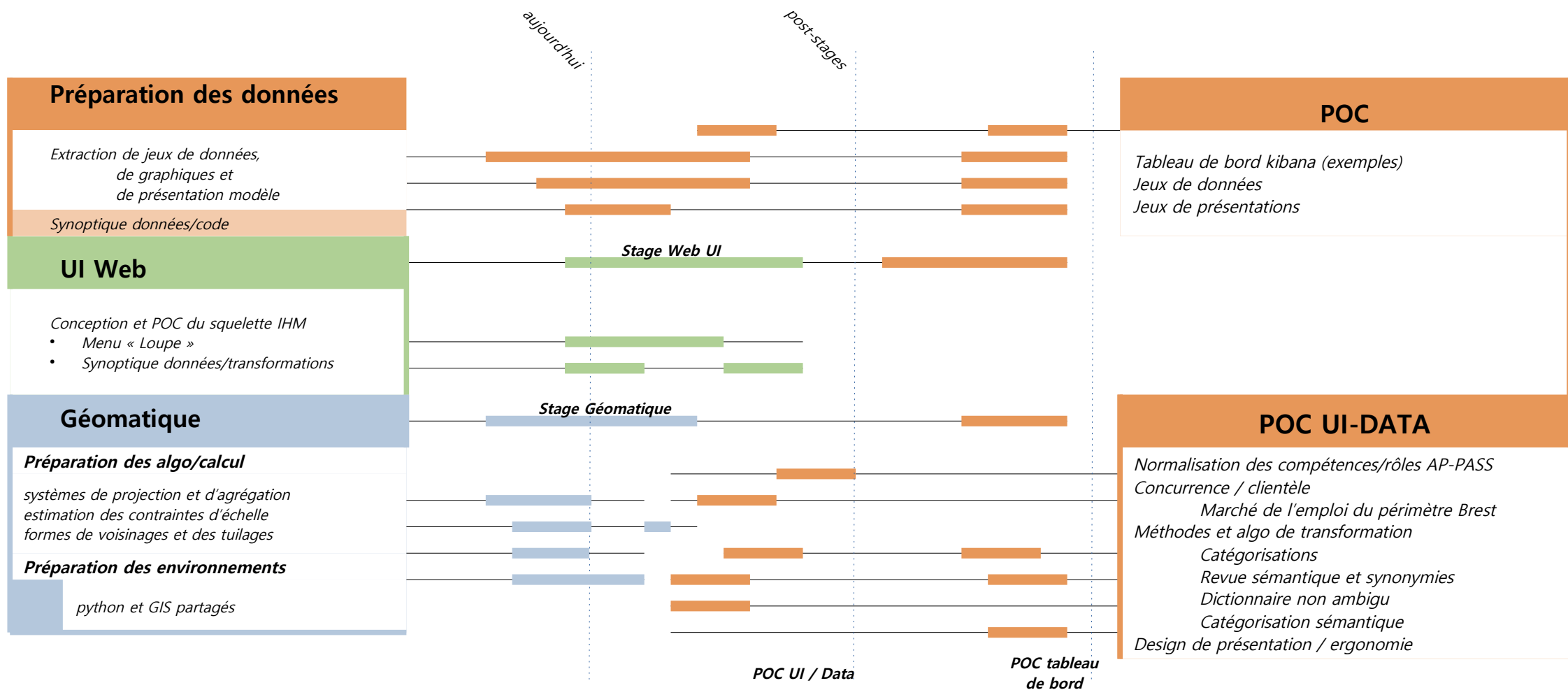
Les étapes au menu

Explications et distinctions



Roadmap Stages et POC

Data Science : Veille des Marchés



Mise en place d'un tableau de bord

Principe d'Avancement étapes et cycles

