Des environnements de calcul sécurisés pour les chercheurs et datascientists

Gouvernance des données, protection des données et centre de calcul

23 janvier 2024

Kamel Gadouche, directeur du Centre d'Accès Sécurisé aux Données, kamel.gadouche@casd.eu


gouvernance
usages
conclusion
Données
calculs
Sécurité
intersectoriel

Données

Secure use file

Données les plus détaillées possibles

FPR ou Scientific use files

Données à très faible risque de réidentification

Open data ou public use files

Données anonymisées (risque théoriquement nul de réidentification)

Fichier brut

nom	prénom	rue	ville	département	région	Tel	salaire	poste	Siret employeur
Dupont	Paul	Albert Camus	Orsay	91	IDF	6025212	45072	Gardien	15624857954
Durant	Julie	des sentiers	Palaiseau	91	IDF	6959849	58233	Dirigeant	59864785668
Morel	Pierre	Marie Curie	Serre	78	IDF	6446827	23392	Ouvrier	95684759666

Fichier « Secure Use File »

Pseudonymiser avec risque élevé → CASD

Ilôts (Iris)	ville	département	région	salaire	poste	Siret employeur
0103	Orsay	91	IDF	45072	Gardien	15624857954
0141	Palaiseau	91	IDF	58233	Dirigeant	59864785668
0102	Serre	78	IDF	23392	Ouvrier	95684759666

Fichier FPR « Scientific use file »

Anonymiser avec faible risque → Diffusion très restreinte

département	région	salaire	poste	Siret employeur
91	IDF	45072	Gardien	15624857954
91	IDF	58233	Dirigeant	59864785668
78	IDF	23392	Ouvrier	95684759666

Fichier open data « Public use file »

Anonymiser avec risque nul → Diffusion publique

région	Salaire (K€)	CSP
IDF	40-50	Employés
IDF	50-60	Chefs entreprise
IDF	20-30	Ouvrier

Gouvernance

Loi 51-711

articles

1 le SSP = Insee + SSM
Statistiques issues des données d'enquêtes et administratives

3 Les données transmises ne peuvent faire l'objet d'aucune communication de la part du service dépositaire.

6 Transmission si décision de l'administration des archives, prise après avis du comité du secret statistique et relative à une demande effectuée à des fins de statistique publique ou de recherche scientifique ou historique.

6 Les renseignements ne peuvent en aucun cas être utilisés **à des fins de contrôle fiscal ou de répression économique**.

6bis Les bénéficiaires des communications de données prises après avis du comité du secret statistique s'engagent à ne communiquer ces données à quiconque.

7bis Obligation de transmission de données administrative à des fins d'établissements de statistiques avec interdiction de les céder pour le service dépositaire à un autre organisme.

7ter Le comité du secret statistique peut autoriser l'accès aux données issues du 7bis pour des besoins de recherche scientifique ou de réalisation d'étude.

Définition du
Service Statistique Public

Gouvernance des données :
- Conseil national de l'information statistique
- Autorité de la statistique publique

Le service statistique public (SSP) – Art. 1

Insee

SSMs

DARES

Agri-SSP

Données brutes
Art. 1

Article 3bis

Entreprises

Collecte

Article 1bis

Citoyens

Enquête

Enquête

Produisent

Statistiques et études

Diffusent en respectant le secret statistique (7bis)

Citoyens, média, pouvoirs publics, …

Autres organismes publics

DGFip

IR

cnav

DSN

Agriculture

RICA

Personnel soumis au secret professionnel

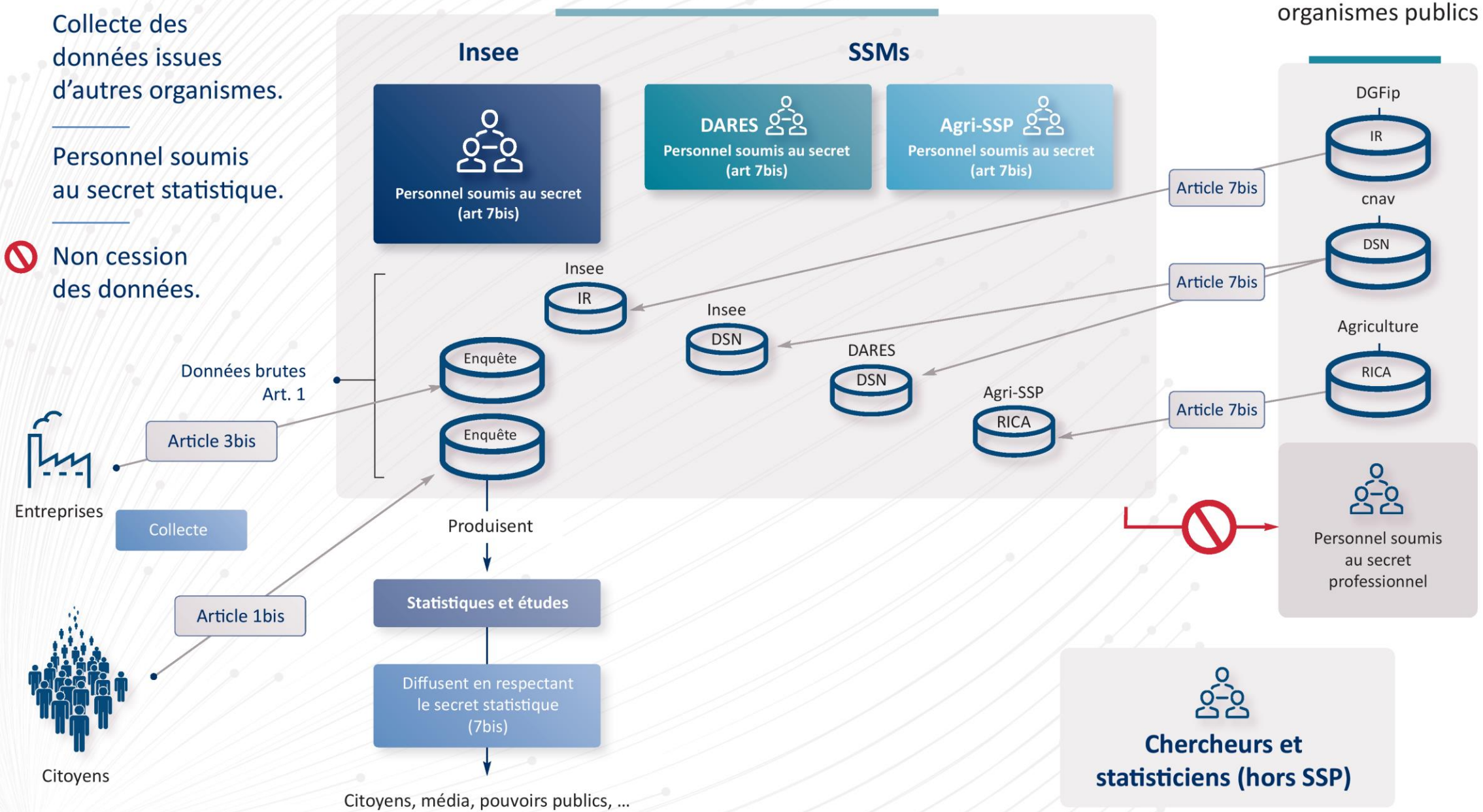
Chercheurs et statisticiens (hors SSP)

Le service statistique public (SSP) – Art. 1

Collecte des données issues d'autres organismes.

Personnel soumis au secret statistique.

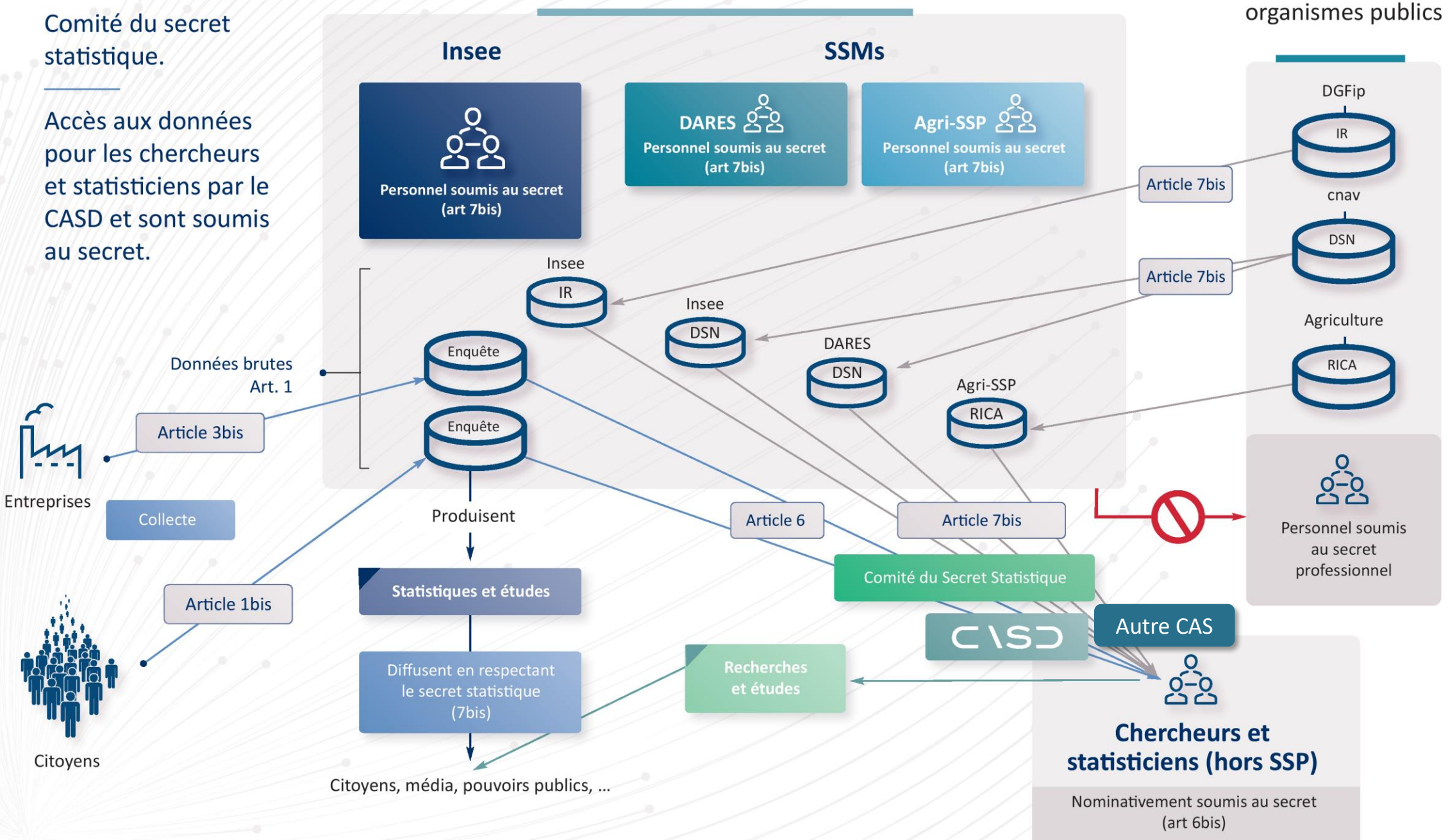
Non cession des données.



Le service statistique public (SSP) – Art. 1

Comité du secret statistique.

Accès aux données pour les chercheurs et statisticiens par le CASD et sont soumis au secret.



Le CASD en bref

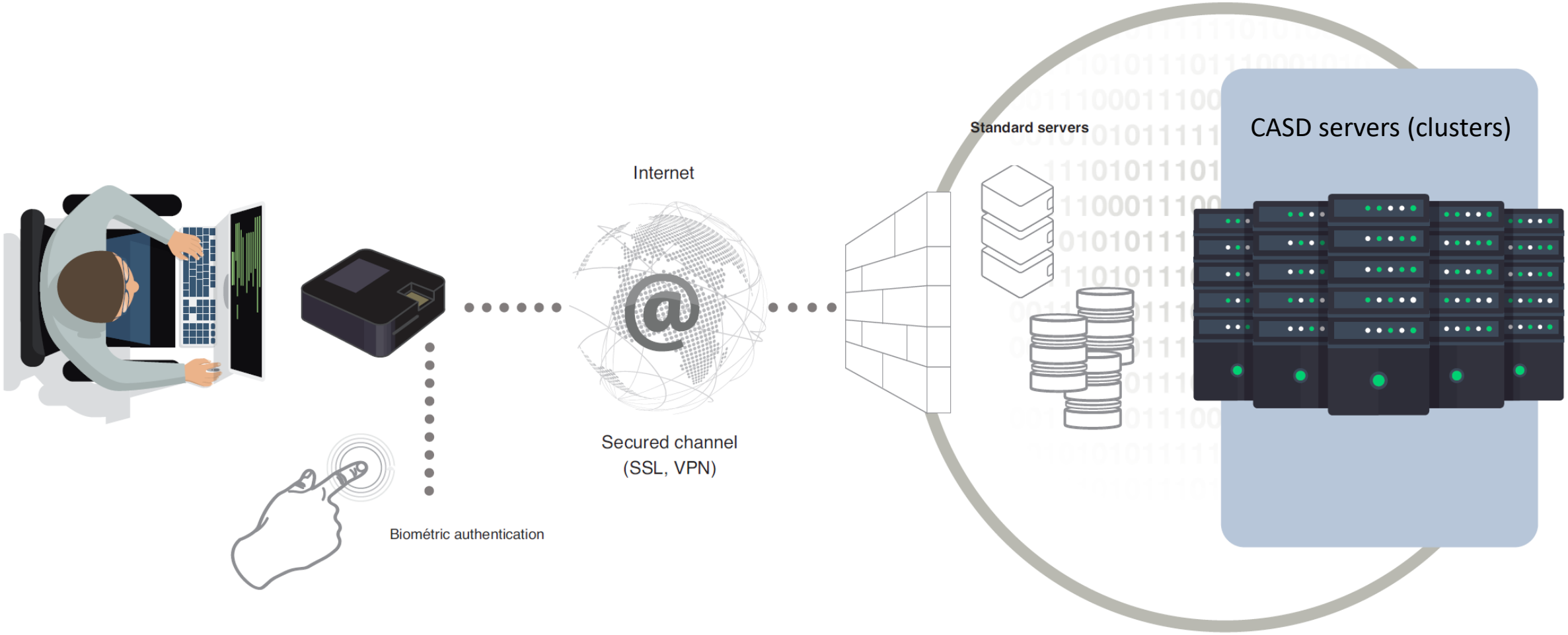
- Groupement d'intérêt public (GIP), consortium regroupant l'Insee, le Genes, la Banque de France, le CNRS, HEC Paris et l'Ecole polytechnique
  - A but non lucratif
  - Facturation pour la recherche à marge négative (contributions des membres)
- 30 personnes
  - Service IT et Datascience
  - Service Data Management
  - Service Project Management
- 500+ sources de données mises à disposition
  - Plusieurs centaines de To de données
- 1000 institutions utilisatrices

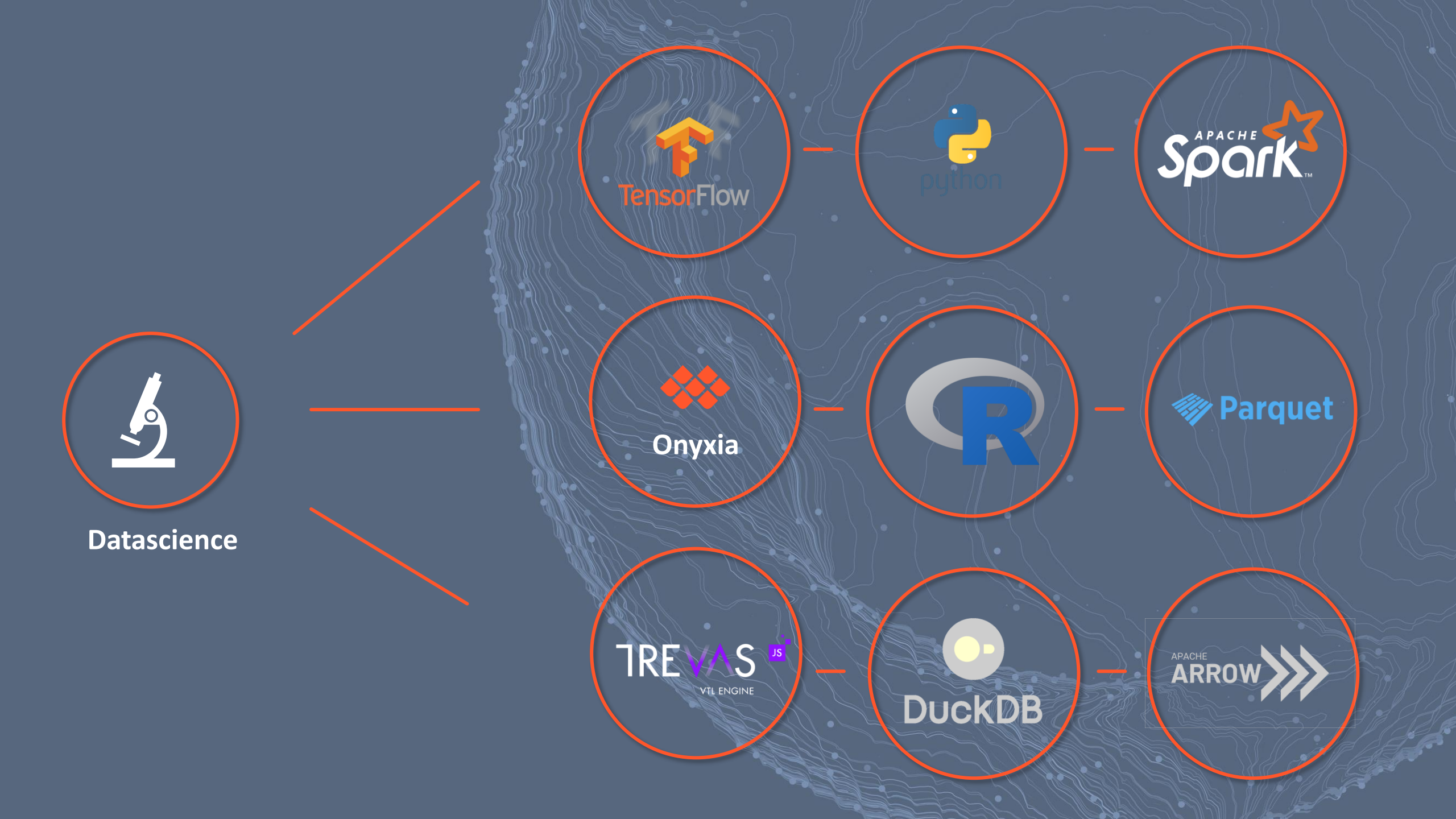
Offre

Environnement de calcul pour la data science

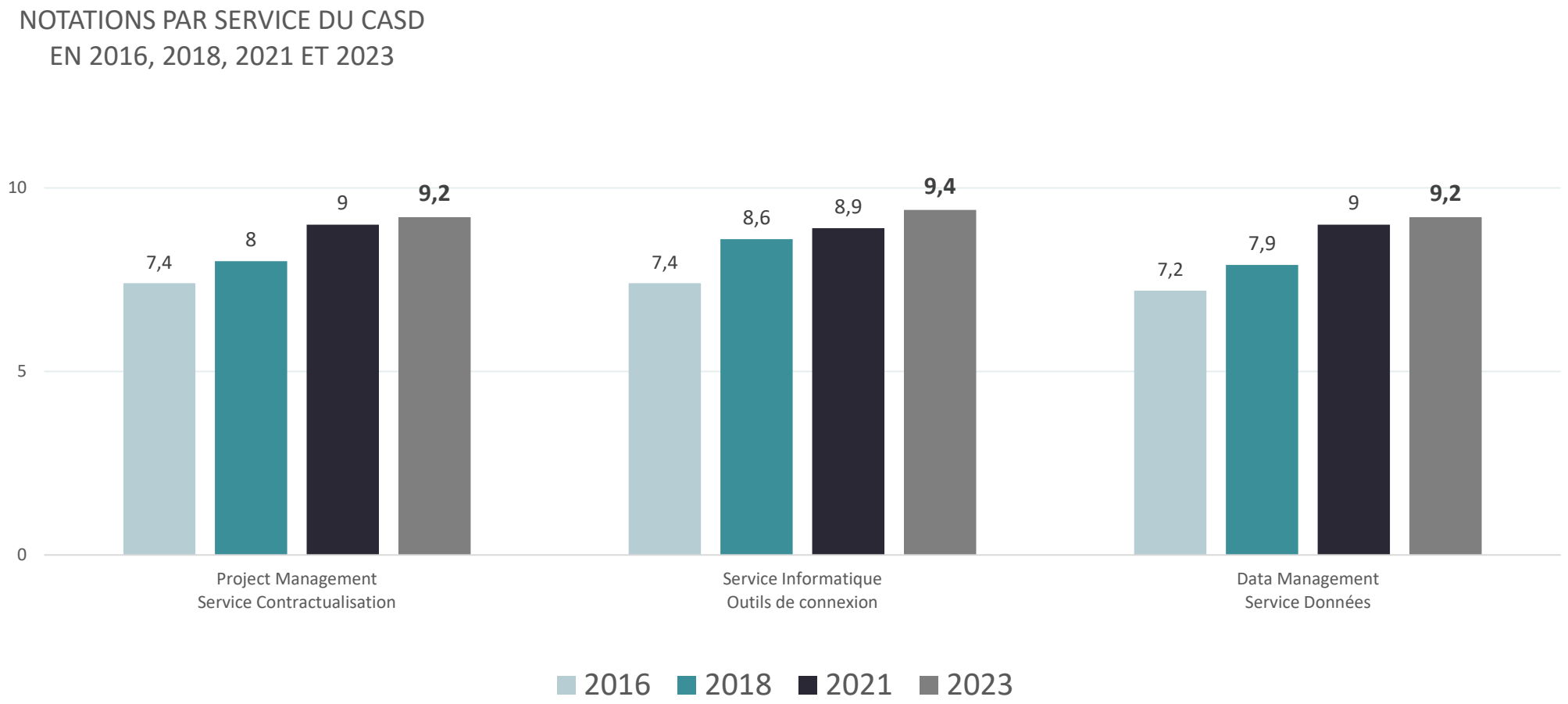
- Puissance de calcul paramétrable : 1 serveur à des clusters de serveurs
- Large offre logicielle : plus de 80 logiciels en catalogue avec notamment les packages pour R, python et stata
- Support technique, sauvegarde, localisation multisite
- Contrôle des sorties, documentation, accompagnement pour les démarches, appariements

usages


Internet
Standard servers
CASD servers (clusters)
Secured channel
(SSL, VPN)
Biométric authentication


Datascience
TensorFlow
python
APACHE Spark
Onyxia
R
Parquet
TREVAS
VTL ENGINE
JS
DuckDB
APACHE ARROW

Enquêtes de satisfaction



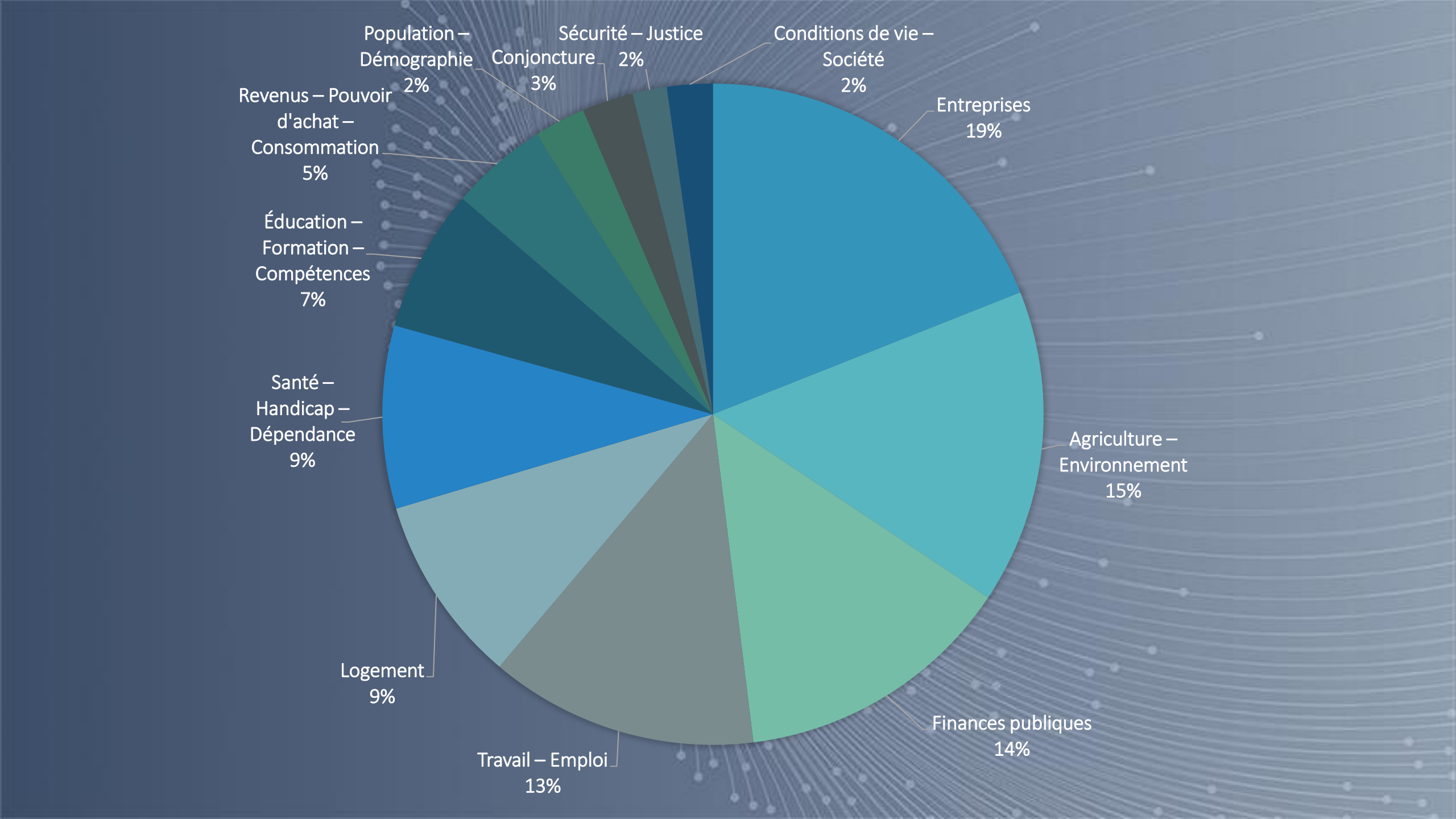
Sécurité

Conformité RGPD par le Bureau Veritas et autorisation CNIL (n°2014-369)

Certifications ISO 27001 et ISO 27701 (RGPD)
Certification « hébergeur de données de santé », homologué SNDS, EDS

Audits techniques réguliers de sécurité réalisés

Données


Population – Démographie 2%
Sécurité – Justice 2%
Conjoncture 3%
Conditions de vie – Société 2%
Revenus – Pouvoir d'achat – Consommation 5%
Entreprises 19%
Éducation – Formation – Compétences 7%
Santé – Handicap – Dépendance 9%
Agriculture – Environnement 15%
Logement 9%
Finances publiques 14%
Travail – Emploi 13%

Intersectoriel

Plusieurs dizaines de producteurs dans de multiples domaines : insee, fiscalité, santé, travail, enseignement supérieur et recherche, agriculture, justice, sécurité intérieure, banque, etc.

Une procédure d’accès unifiée au bénéfice des chercheurs et datascientists, un atout essentiel

2/3 des 700 projets menés sur le CASD s’appuient sur des données de plusieurs producteurs

10% des projets concernent des études localisées dans un ou plusieurs territoires (Conseil régionaux …)

Conclusions

Un appui initial et continu des producteurs et du CNIS, en particulier le comité du label.
Des progrès considérables permis à la gouvernance mise en place au fil des années.

Un dispositif intersectoriel qui découle de l'organisation de la gouvernance de la statistique publique et permet de prévenir les silos.

Un dispositif donc assez unique au monde à cette échelle qui s'explique en grande partie par la mise en place de lieux d'échange, de concertation et de décision : le comité du secret statistique, le CNIS, l'ASP, l'INSEE en tant que coordinateur…