Science and
Technology
Facilities Council

Hartree Centre

Welcome

Science and Technology Facilities Council

Hartree Centre

INTELLECTUAL PROPERTY RIGHTS NOTICE:

Science and
Technology
Facilities Council

Hartree Centre

Big Data Week 12

Advanced data analysis & deep learning

Science and
Technology
Facilities Council

Hartree Centre

Outline

Combined batch and streaming analytics

- Batch processing and stream processing
- Combined analytics

Data modelling and data fusion

- Data models
- Describing complex data types
- Data fusion by unique key
- Statistical models and fused data

Break

Neural networks

Q&A

Science and
Technology
Facilities Council
Hartree Centre

Combined batch and streaming analytics

Batch and streaming analytics

Batch analytic

Big data volume problem

Uses an existing pot of data and runs a single set of calculations to build a model.
e.g. frequentist logistic regression using iteratively reweighted least-squares to predict probability of a patient getting a given disease.

+ Uses all the data that is available to build the model.
+ Can use expensive algorithms if the processing power is available.
- Only updates when the analytic is run over the data.

Streaming analytic

Big data velocity problem

Creates a model or state and continuously updates this as new data comes in.
e.g. Kalman filter for position and velocity of a satellite in space.

+ Updates and allows rapid response in changing circumstances.
+ Often more computationally efficient.
- Limited to models which can be continuously updated.

Streaming-batch analytic

Creates a model or state and updates this in blocks as batches of data come in.

Batch and streaming analytics

Streaming analytics are limited to models that can be continuously updated.

Some quantities can be hard to calculate in streams

e.g. median (needs a data sort or a specialised algorithm like the heap method)

5.44
↓

1.15 2.16 2.24 2.59 3.4 **3.98** 6.19 6.62 6.98 8.13 8.64

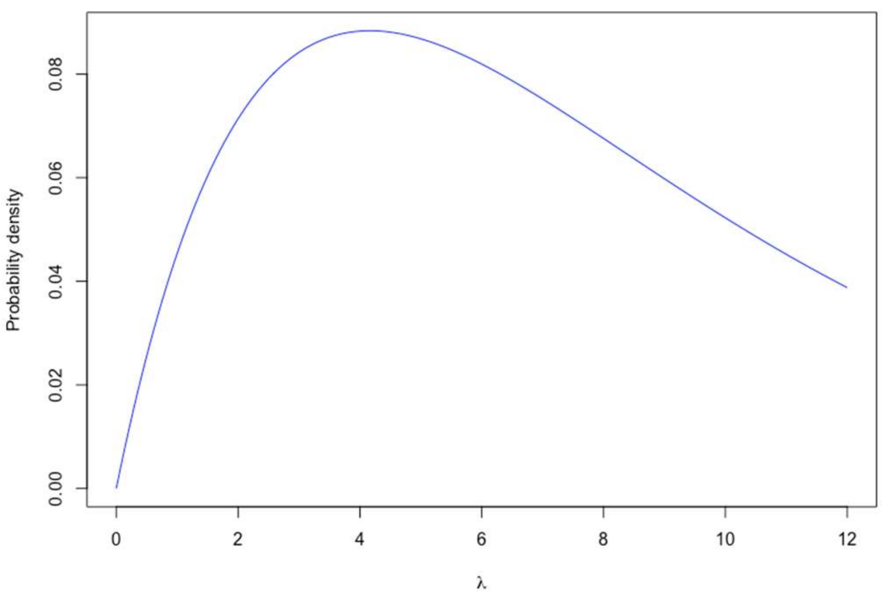
Can resort to an approximate method if you don’t need the exact answer

Batch and streaming analytics

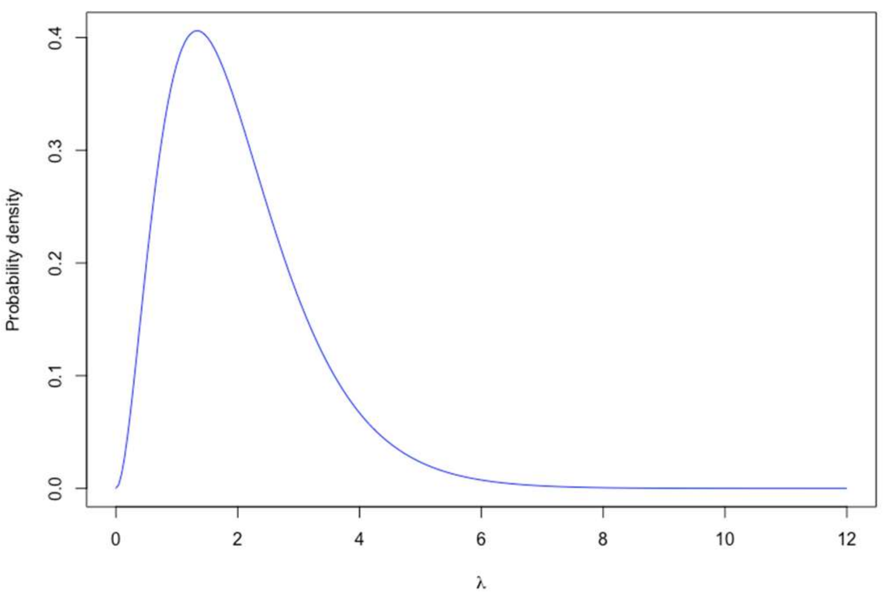
Some analytics naturally fit into streams

e.g. Bayesian updates with a conjugate prior

After 1 observation ΔT=0.15



After 2nd observation ΔT=1.26



$$\begin{array}{c} \alpha = 2 \\ \beta = 0.233 \end{array} \longrightarrow \begin{array}{c} \alpha' = \alpha + 1 \\ \beta' = \beta + \Delta T \end{array} \longrightarrow \begin{array}{c} \alpha = 3 \\ \beta = 1.493 \end{array}$$

Combined analytics

Can combine batch and streaming analytics together

e.g. build a batch model, and then stream updates into it

These models and systems of data analysis can become more and more complicated.

Physical considerations may also be important

e.g. edge devices reporting back to a central server over slow connections

Science and
Technology
Facilities Council

Hartree Centre

Data modelling and data fusion

Science and
Technology
Facilities Council

Hartree Centre

Data modelling

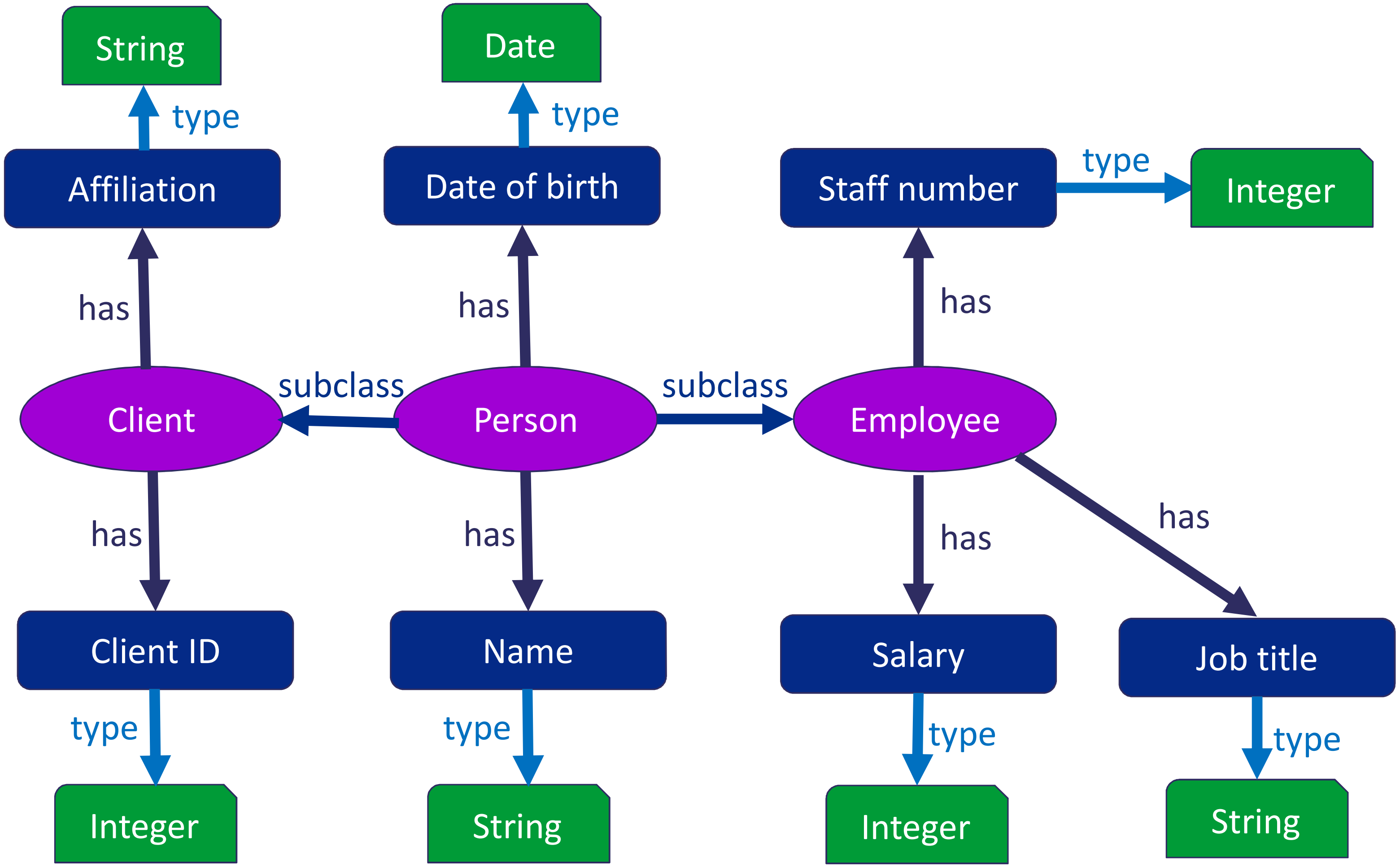
A *data model* is a model which can

- Organise elements of data
- Specify types of data (integer, string)
- Show how elements relate to one another
- Link to real-world definitions

Data model uses include:

- Validation of received data (types, ranges)
- Assist with building analytics
- Producing standards and agreed definitions of terms (e.g. W3C standards)
- Fusing disparate data sources

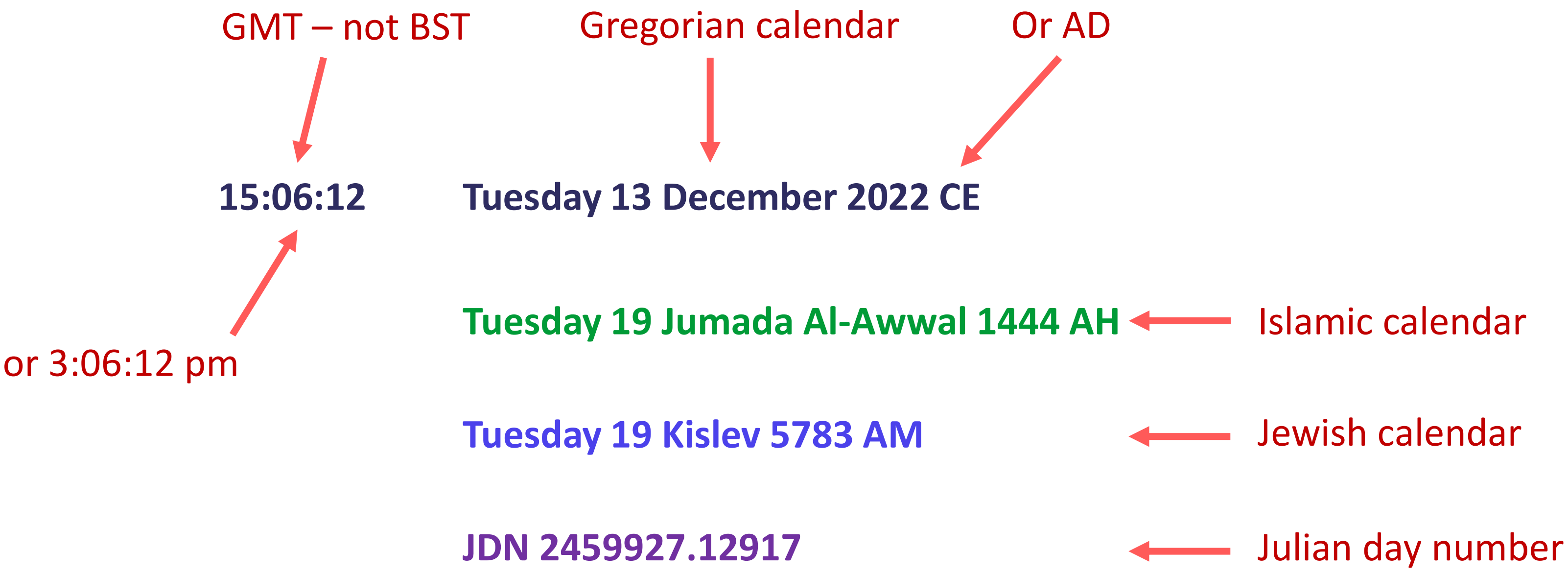
Data models



Science and Technology Facilities Council

Hartree Centre

Modelling a date and time



GMT – not BST

Gregorian calendar

Or AD

15:06:12

Tuesday 13 December 2022 CE

or 3:06:12 pm

Tuesday 19 Jumada Al-Awwal 1444 AH

Islamic calendar

Tuesday 19 Kislev 5783 AM

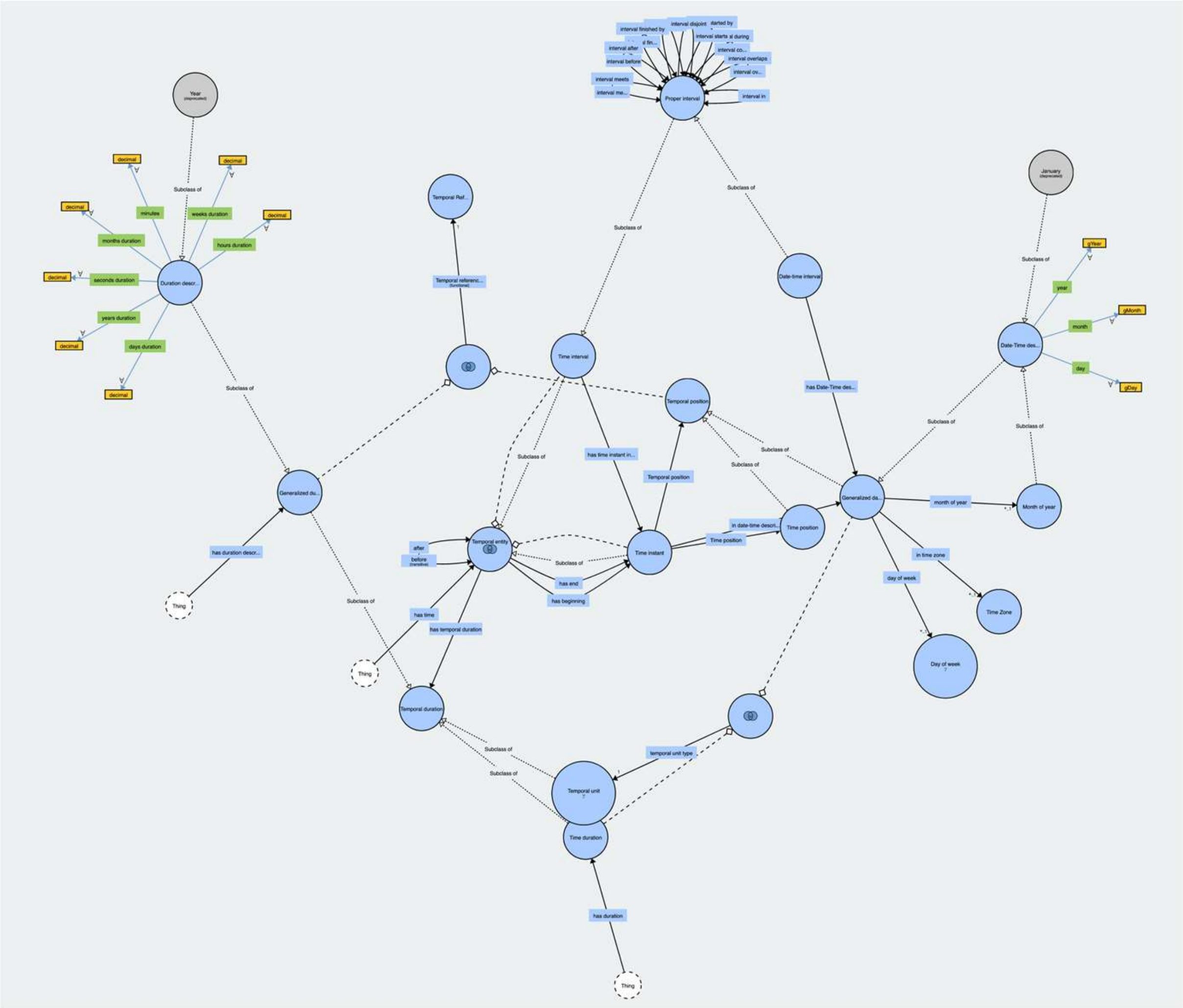
Jewish calendar

JDN 2459927.12917

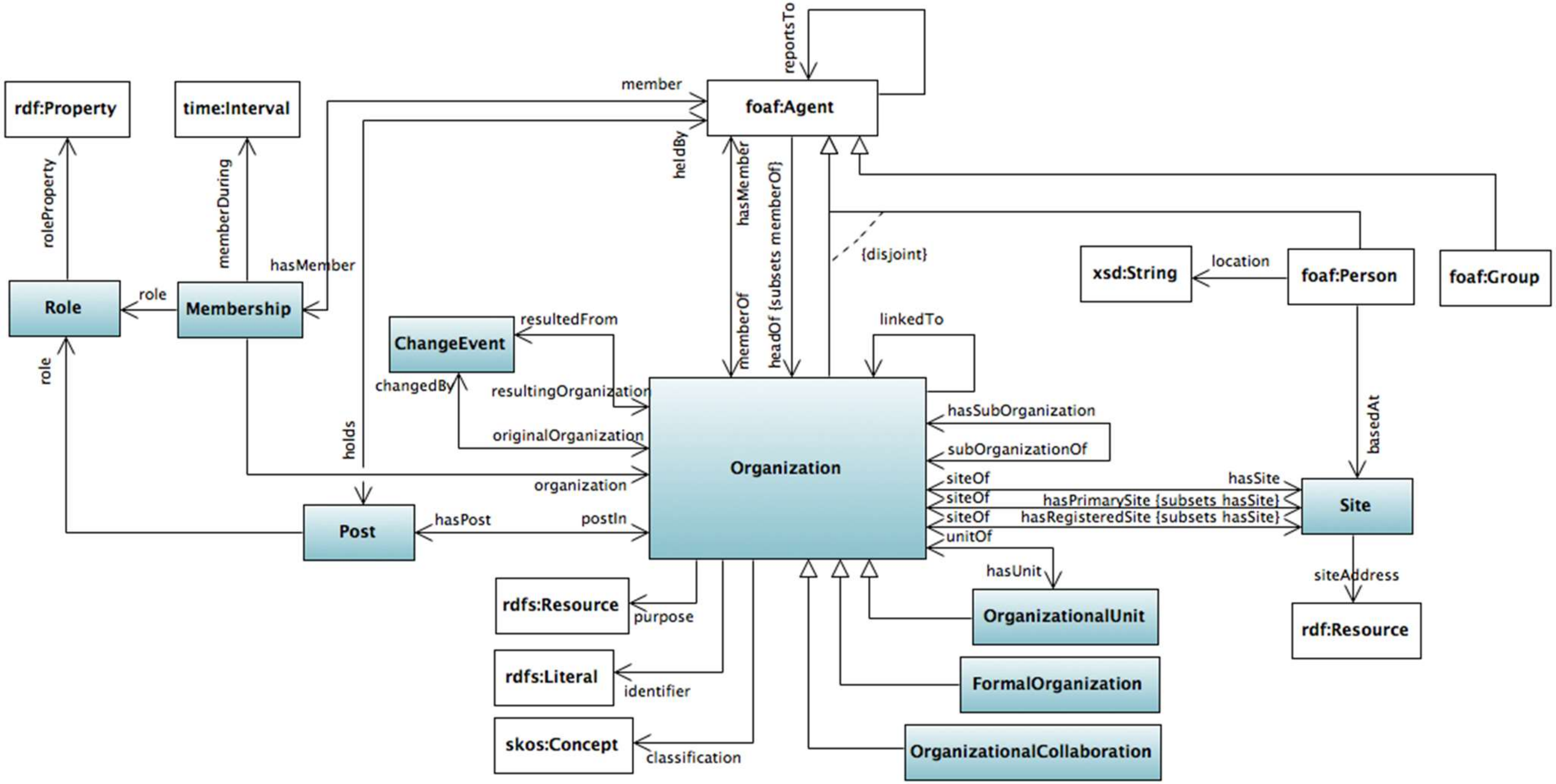
Julian day number

Data models

The complete OWL-Time ontology (https://www.w3.org/TR/owl-time/)



Data models



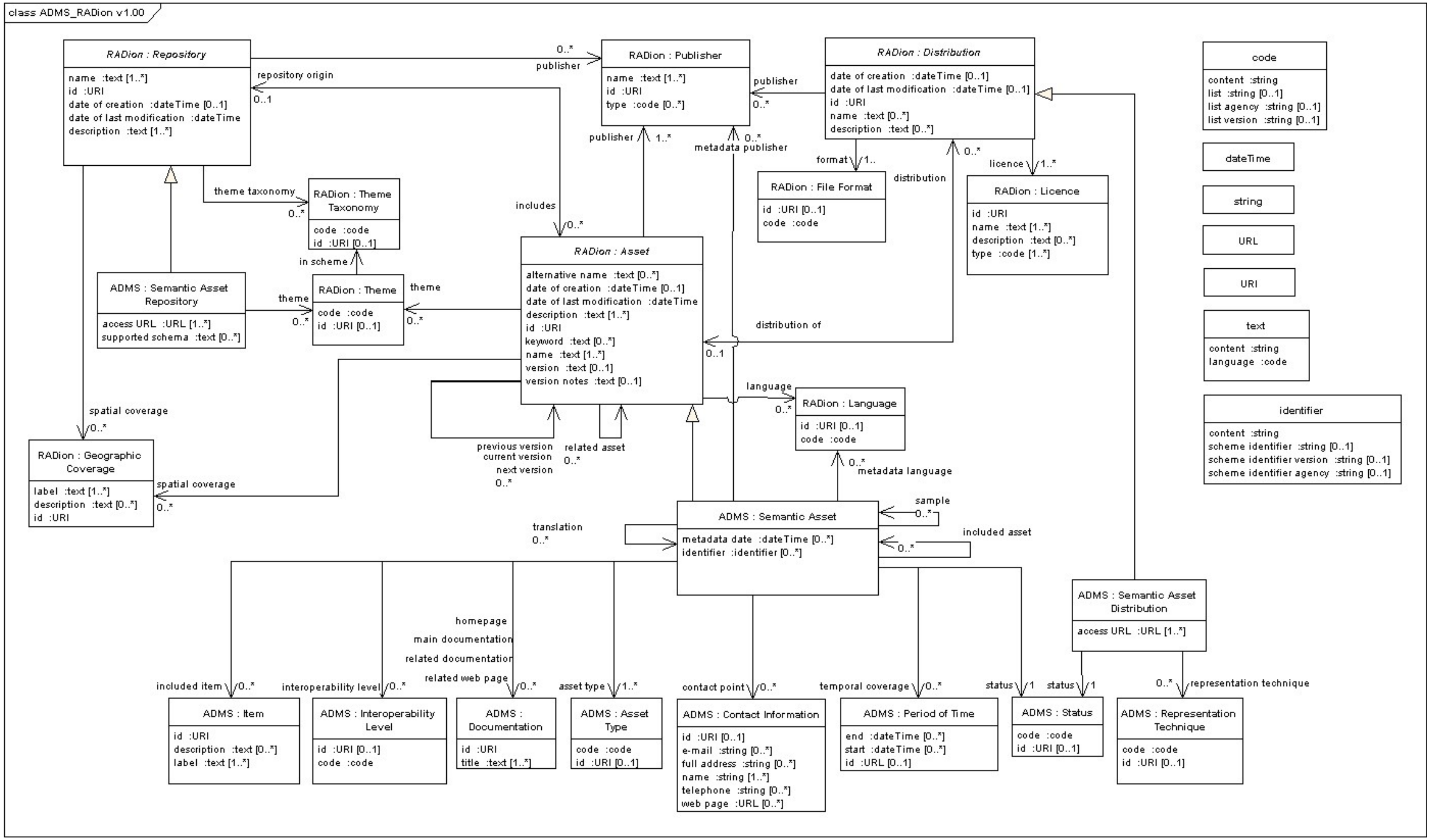
Science and Technology Facilities Council

Hartree Centre

Diagram of the Organization ontology: https://www.w3.org/TR/vocab-org/

Recording data models – UML

UML = Unified Modelling Language



class ADMS_RADion v1.00

RADion : Repository
name :text [1..*]
id :URI
date of creation :dateTime [0..1]
date of last modification :dateTime
description :text [1..*]

RADion : Publisher
name :text [1..*]
id :URI
type :code [0..*]

RADion : Distribution
date of creation :dateTime [0..1]
date of last modification :dateTime [0..1]
id :URI
name :text [0..*]
description :text [0..*]

code
content :string
list :string [0..1]
list agency :string [0..1]
list version :string [0..1]

dateTime

string

URL

URI

text
content :string
language :code

identifier
content :string
scheme identifier :string [0..1]
scheme identifier version :string [0..1]
scheme identifier agency :string [0..1]

RADion : Theme Taxonomy
code :code
id :URI [0..1]

RADion : File Format
id :URI [0..1]
code :code

RADion : Licence
id :URI
name :text [1..*]
description :text [0..*]
type :code [1..*]

ADMS : Semantic Asset Repository
access URL :URL [1..*]
supported schema :text [0..*]

RADion : Theme
code :code
id :URI [0..1]

RADion : Asset
alternative name :text [0..*]
date of creation :dateTime [0..1]
date of last modification :dateTime
description :text [1..*]
id :URI
keyword :text [0..*]
name :text [1..*]
version :text [0..1]
version notes :text [0..1]

RADion : Language
id :URI [0..1]
code :code

RADion : Geographic Coverage
label :text [1..*]
description :text [0..*]
id :URI

ADMS : Semantic Asset
metadata date :dateTime [0..*]
identifier :identifier [0..*]

ADMS : Semantic Asset Distribution
access URL :URL [1..*]

ADMS : Item
id :URI
description :text [0..*]
label :text [1..*]

ADMS : Interoperability Level
id :URI [0..1]
code :code

ADMS : Documentation
id :URI
title :text [1..*]

ADMS : Asset Type
code :code
id :URI [0..1]

ADMS : Contact Information
id :URI [0..1]
e-mail :string [0..*]
full address :string [0..*]
name :string [1..*]
telephone :string [0..*]
web page :URL [0..*]

ADMS : Period of Time
end :dateTime [0..*]
start :dateTime [0..*]
id :URL [0..1]

ADMS : Status
code :code
id :URI [0..1]

ADMS : Representation Technique
code :code
id :URI [0..1]

Asset Open Metadata Schema

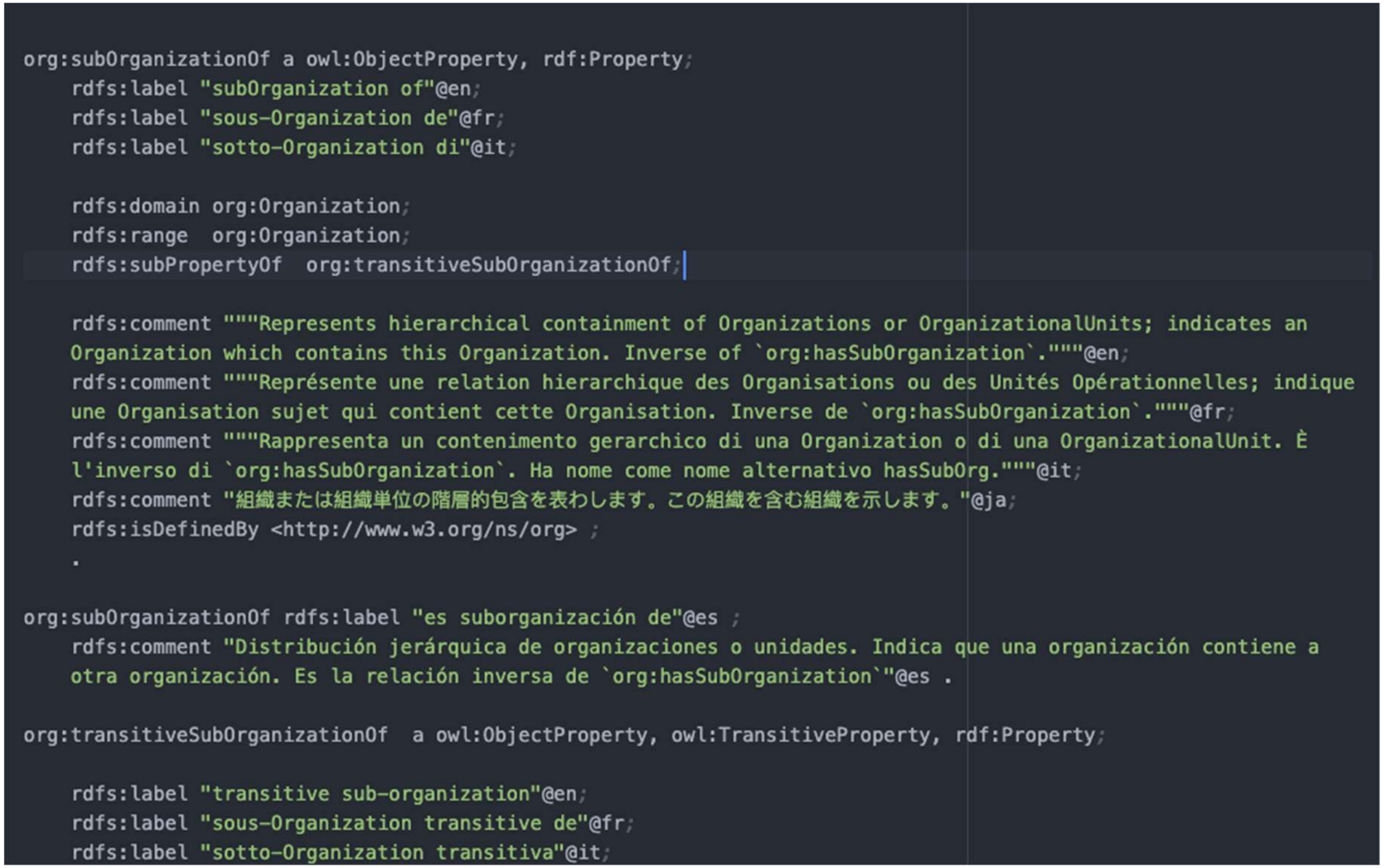
(from https://en.wikipedia.org/wiki/Asset_Description_Metadata_Schema)

Science and Technology Facilities Council

Hartree Centre

Recording data models – OWL

OWL = Web Ontology Language



```
org:subOrganizationOf a owl:ObjectProperty, rdf:Property;
    rdfs:label "subOrganization of"@en;
    rdfs:label "sous-Organization de"@fr;
    rdfs:label "sotto-Organization di"@it;

    rdfs:domain org:Organization;
    rdfs:range  org:Organization;
    rdfs:subPropertyOf  org:transitiveSubOrganizationOf;

    rdfs:comment """Represents hierarchical containment of Organizations or OrganizationalUnits; indicates an
    Organization which contains this Organization. Inverse of `org:hasSubOrganization`."""@en;
    rdfs:comment """Représente une relation hierarchique des Organisations ou des Unités Opérationnelles; indique
    une Organisation sujet qui contient cette Organisation. Inverse de `org:hasSubOrganization`."""@fr;
    rdfs:comment """Rappresenta un contenimento gerarchico di una Organization o di una OrganizationalUnit. È
    l'inverso di `org:hasSubOrganization`. Ha nome come nome alternativo hasSubOrg."""@it;
    rdfs:comment "組織または組織単位の階層的包含を表わします。この組織を含む組織を示します。"@ja;
    rdfs:isDefinedBy <http://www.w3.org/ns/org> ;
    .

org:subOrganizationOf rdfs:label "es suborganización de"@es ;
    rdfs:comment "Distribución jerárquica de organizaciones o unidades. Indica que una organización contiene a
    otra organización. Es la relación inversa de `org:hasSubOrganization`"@es .

org:transitiveSubOrganizationOf  a owl:ObjectProperty, owl:TransitiveProperty, rdf:Property;

    rdfs:label "transitive sub-organization"@en;
    rdfs:label "sous-Organization transitive de"@fr;
    rdfs:label "sotto-Organization transitiva"@it;
```

Example from the Organization ontology (https://www.w3.org/TR/vocab-org/) in Turtle

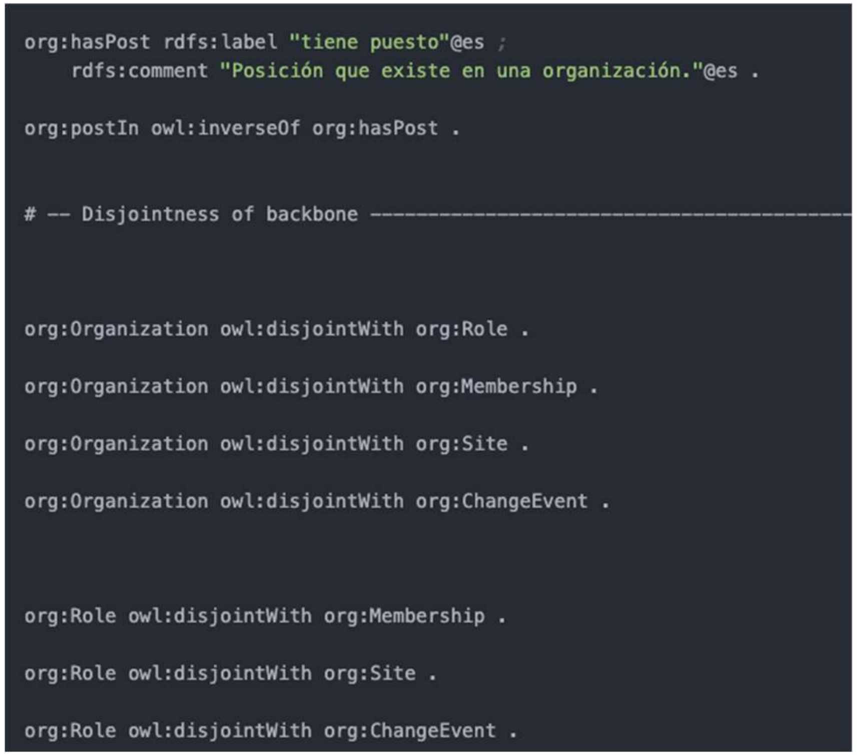
Science and Technology Facilities Council

Hartree Centre

Data models and ontologies

An *ontology* is a formal way to describe knowledge by a set of concepts and relations between those concepts.

- Closely related to data models
- Ontology descriptions (e.g. OWL2-DL) also include relations that allow automated logic



```
org:hasPost rdfs:label "tiene puesto"@es ;
    rdfs:comment "Posición que existe en una organización."@es .

org:postIn owl:inverseOf org:hasPost .


# -- Disjointness of backbone ------------------------------------------


org:Organization owl:disjointWith org:Role .

org:Organization owl:disjointWith org:Membership .

org:Organization owl:disjointWith org:Site .

org:Organization owl:disjointWith org:ChangeEvent .


org:Role owl:disjointWith org:Membership .

org:Role owl:disjointWith org:Site .

org:Role owl:disjointWith org:ChangeEvent .
```

More from the Organization ontology.

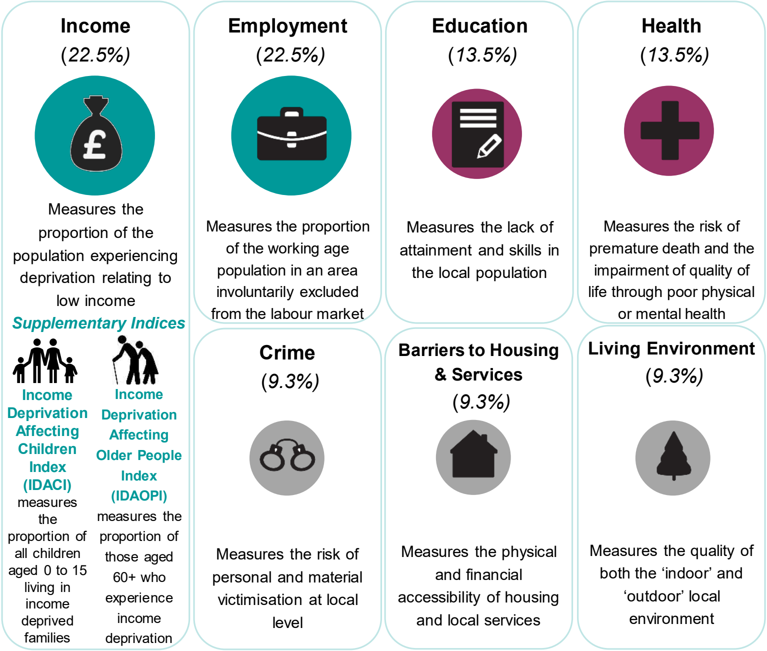
Science and Technology Facilities Council

Hartree Centre

Data fusion

Fusion is the combination of disparate data sets to give a more complete picture of a situation.
This is a big data *variety* problem.

Example: ONS Index of Multiple Deprivation[1]



Income (*22.5%*) Measures the proportion of the population experiencing deprivation relating to low income

Supplementary Indices

Income Deprivation Affecting Children Index (IDACI) measures the proportion of all children aged 0 to 15 living in income deprived families

Income Deprivation Affecting Older People Index (IDAOPI) measures the proportion of those aged 60+ who experience income deprivation

Employment (*22.5%*) Measures the proportion of the working age population in an area involuntarily excluded from the labour market

Education (*13.5%*) Measures the lack of attainment and skills in the local population

Health (*13.5%*) Measures the risk of premature death and the impairment of quality of life through poor physical or mental health

Crime (*9.3%*) Measures the risk of personal and material victimisation at local level

Barriers to Housing & Services (*9.3%*) Measures the physical and financial accessibility of housing and local services

Living Environment (*9.3%*) Measures the quality of both the 'indoor' and 'outdoor' local environment

Combines data from[2]:

- Department of Work & Pensions
- HMRC
- Home Office
- Department for Education
- Hospital Episode Statistics (NHS)
- NHS Digital
- 2011 Census
- …

[1] https://www.gov.uk/government/statistics/english-indices-of-deprivation-2019
[2] https://www.gov.uk/government/publications/english-indices-of-deprivation-2019-technical-report

Fusion by unique key

Unique key fusion: join data sets using a key which matches across them.

The key must be:

- Unique
- Consistent

Example: NHS Liverpool case study.

Social care data: **568a9123be19011d**, L14, 35, B17, physiotherapy, 2005-06, …

↑

LinkPseudo (hashed NHS number) unique key

↓

Primary care data: **568a9123be19011d**, L14, 33, C34, breaks and fractures, 2004-11, …

Examples of unique keys:

NHS number, National Insurance number, company registration number

Examples of things that are not unique keys:

Names, postcodes, local authority names

Science and
Technology
Facilities Council
Hartree Centre

Fusion by aggregated data

If a field is not unique, it may be possible to create a unique key by fusing several pieces of data

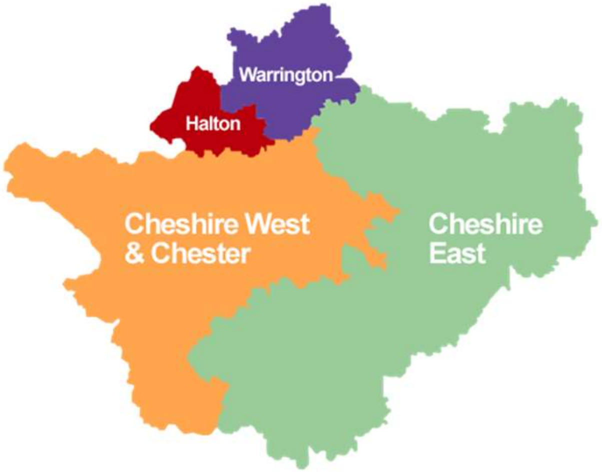
e.g	name	+	date of birth	+	place of birth
	probably not unique		more specific		even more specific

Also related to de-anonymisation: recovery of information hidden in an anonymised data set.

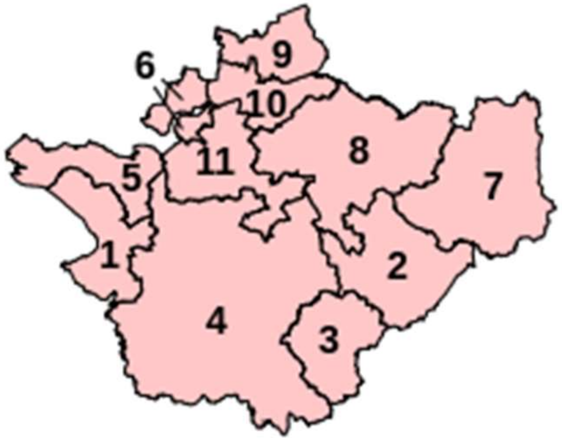
99.98% of Americans could be correctly identified using 15 demographic variables
(Rocher, Hendrix & de Montjoie, 2019, https://www.nature.com/articles/s41467-019-10933-3)

Data fusion

Sometimes descriptions can get quite complicated



Cheshire unitary authorities (2021)



Cheshire constituencies (2021)

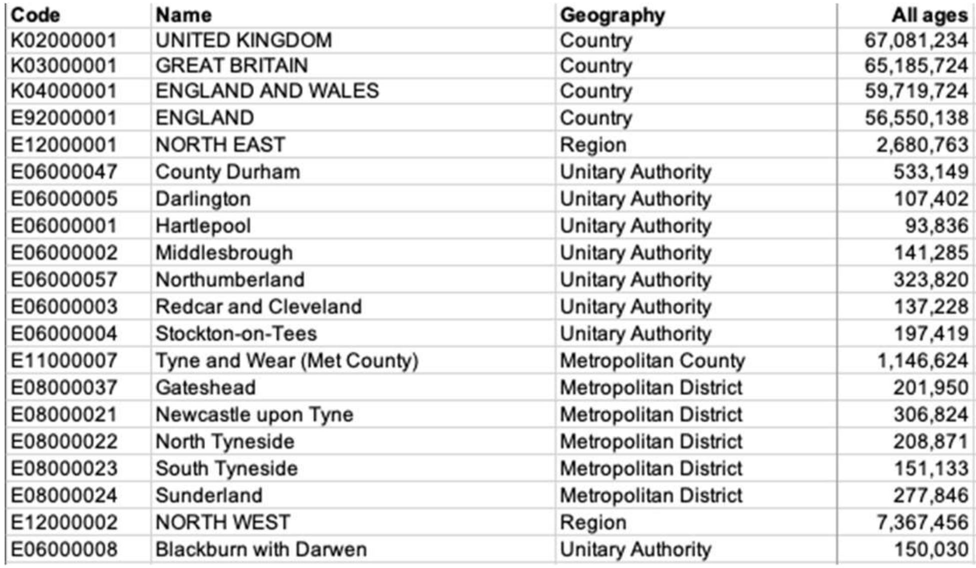
Cheshire boroughs (2008)

Science and
Technology
Facilities Council

Hartree Centre

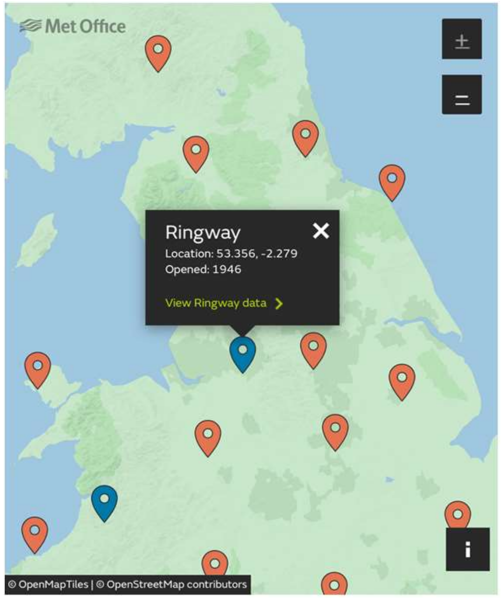
Data fusion

Sometimes data describes the same entities but presented in very different ways



Code	Name	Geography	All ages
K02000001	UNITED KINGDOM	Country	67,081,234
K03000001	GREAT BRITAIN	Country	65,185,724
K04000001	ENGLAND AND WALES	Country	59,719,724
E92000001	ENGLAND	Country	56,550,138
E12000001	NORTH EAST	Region	2,680,763
E06000047	County Durham	Unitary Authority	533,149
E06000005	Darlington	Unitary Authority	107,402
E06000001	Hartlepool	Unitary Authority	93,836
E06000002	Middlesbrough	Unitary Authority	141,285
E06000057	Northumberland	Unitary Authority	323,820
E06000003	Redcar and Cleveland	Unitary Authority	137,228
E06000004	Stockton-on-Tees	Unitary Authority	197,419
E11000007	Tyne and Wear (Met County)	Metropolitan County	1,146,624
E08000037	Gateshead	Metropolitan District	201,950
E08000021	Newcastle upon Tyne	Metropolitan District	306,824
E08000022	North Tyneside	Metropolitan District	208,871
E08000023	South Tyneside	Metropolitan District	151,133
E08000024	Sunderland	Metropolitan District	277,846
E12000002	NORTH WEST	Region	7,367,456
E06000008	Blackburn with Darwen	Unitary Authority	150,030

Local authority area populations



Ringway

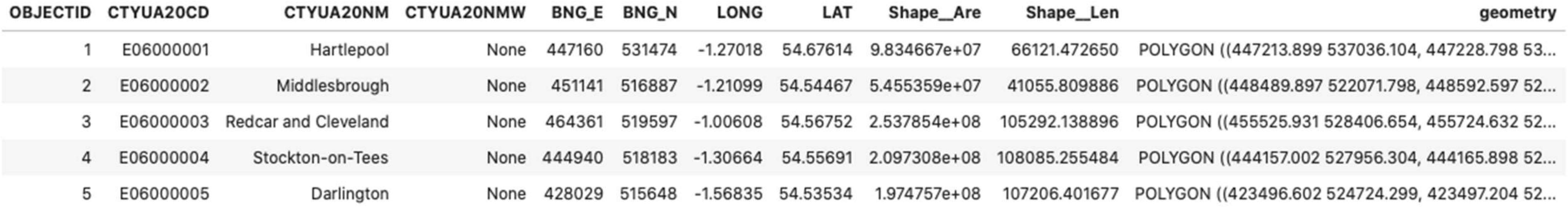
Location: 53.356, -2.279

Opened: 1946

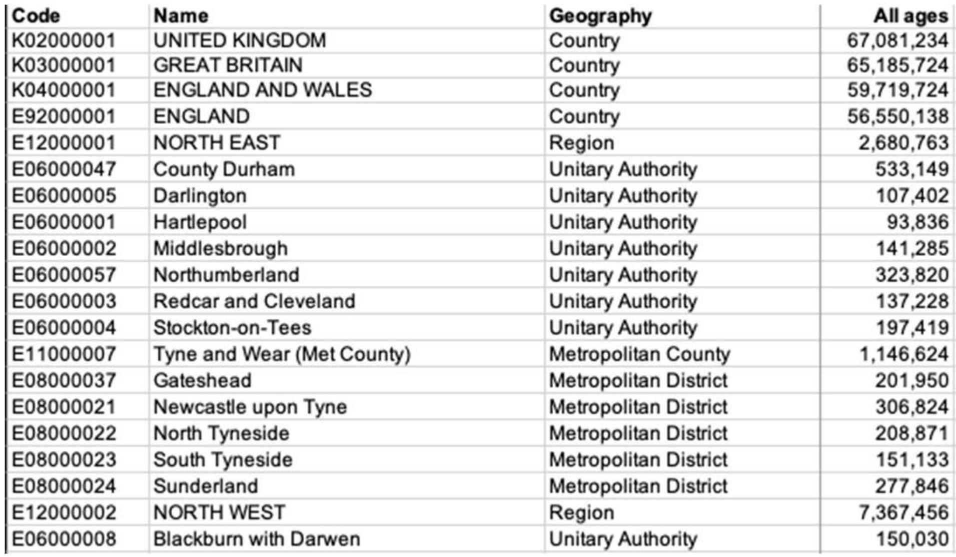
View Ringway data

Latitude and longitude

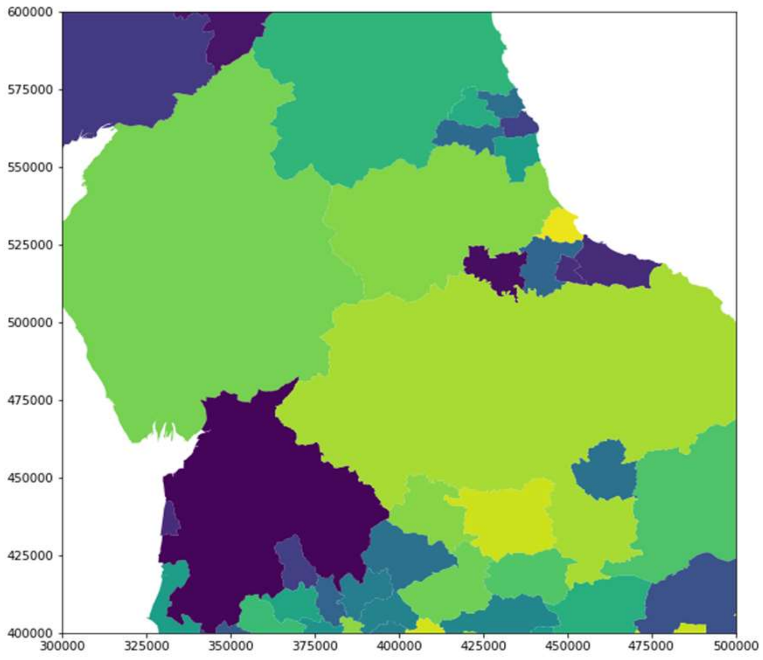
Data fusion



OBJECTID	CTYUA20CD	CTYUA20NM	CTYUA20NMW	BNG_E	BNG_N	LONG	LAT	Shape__Are	Shape__Len	geometry
1	E06000001	Hartlepool	None	447160	531474	-1.27018	54.67614	9.834667e+07	66121.472650	POLYGON ((447213.899 537036.104, 447228.798 53...
2	E06000002	Middlesbrough	None	451141	516887	-1.21099	54.54467	5.455359e+07	41055.809886	POLYGON ((448489.897 522071.798, 448592.597 52...
3	E06000003	Redcar and Cleveland	None	464361	519597	-1.00608	54.56752	2.537854e+08	105292.138896	POLYGON ((455525.931 528406.654, 455724.632 52...
4	E06000004	Stockton-on-Tees	None	444940	518183	-1.30664	54.55691	2.097308e+08	108085.255484	POLYGON ((444157.002 527956.304, 444165.898 52...
5	E06000005	Darlington	None	428029	515648	-1.56835	54.53534	1.974757e+08	107206.401677	POLYGON ((423496.602 524724.299, 423497.204 52...



Code	Name	Geography	All ages
K02000001	UNITED KINGDOM	Country	67,081,234
K03000001	GREAT BRITAIN	Country	65,185,724
K04000001	ENGLAND AND WALES	Country	59,719,724
E92000001	ENGLAND	Country	56,550,138
E12000001	NORTH EAST	Region	2,680,763
E06000047	County Durham	Unitary Authority	533,149
E06000005	Darlington	Unitary Authority	107,402
E06000001	Hartlepool	Unitary Authority	93,836
E06000002	Middlesbrough	Unitary Authority	141,285
E06000057	Northumberland	Unitary Authority	323,820
E06000003	Redcar and Cleveland	Unitary Authority	137,228
E06000004	Stockton-on-Tees	Unitary Authority	197,419
E11000007	Tyne and Wear (Met County)	Metropolitan County	1,146,624
E08000037	Gateshead	Metropolitan District	201,950
E08000021	Newcastle upon Tyne	Metropolitan District	306,824
E08000022	North Tyneside	Metropolitan District	208,871
E08000023	South Tyneside	Metropolitan District	151,133
E08000024	Sunderland	Metropolitan District	277,846
E12000002	NORTH WEST	Region	7,367,456
E06000008	Blackburn with Darwen	Unitary Authority	150,030



Science and Technology Facilities Council

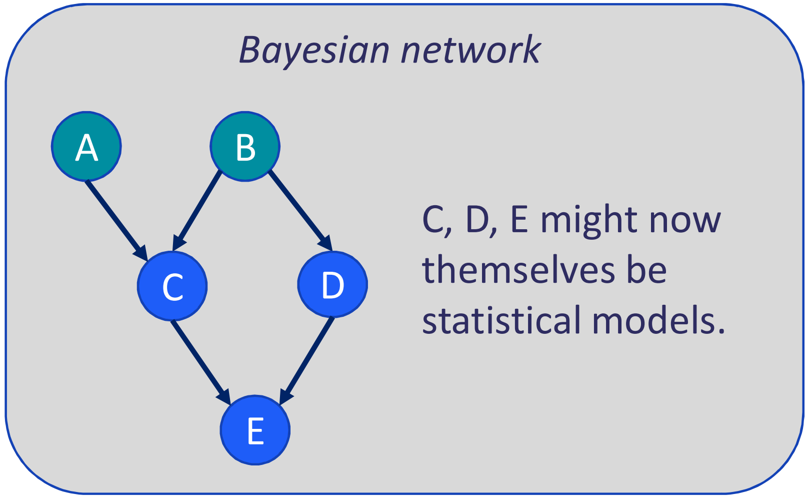
Hartree Centre

ONS: Population Estimates for UK, England and Wales, Scotland and Northern Ireland: Mid-2020

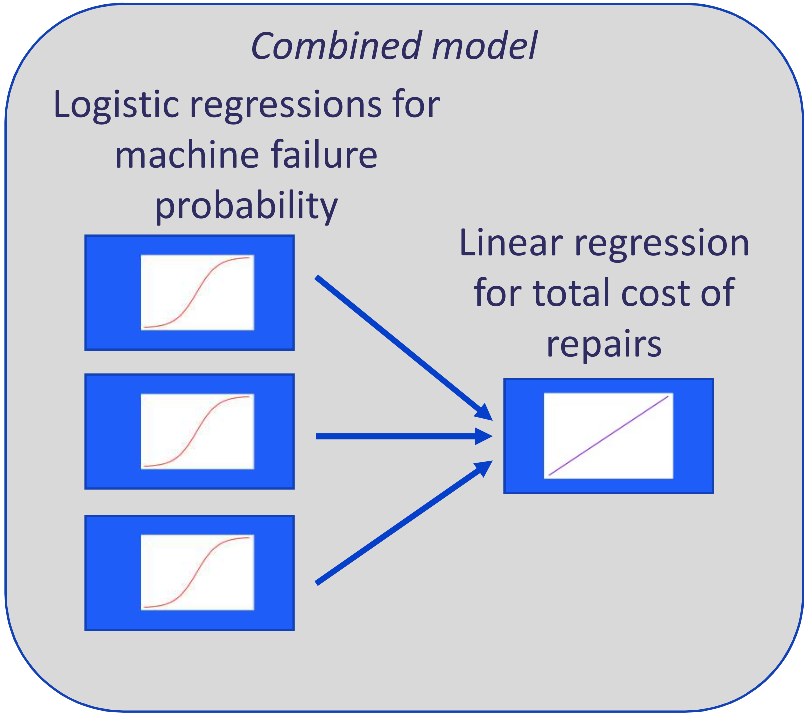
Statistical models with fused data

Using fused data

1. Build a single model (e.g. a random forest) using the whole data set.
   - k-prototypes clustering (Huang, 1997)[1]
2. Combine models derived from several data sets
   - Combined model
   - Bayesian network model



C, D, E might now themselves be statistical models.


Combined model
Logistic regressions for machine failure probability
Linear regression for total cost of repairs

Hartree Centre

[1] Huang, Zhexue. 'Clustering large data sets with mixed numeric and categorical values.' In *Proceedings of the 1st Pacific-Asia conference on knowledge discovery and data mining,(PAKDD)*, pp. 21-34. 1997.

Science and
Technology
Facilities Council

Hartree Centre

Questions?