8 March 2017

ISGC 2017
Academia Sinica
Taipei, Taiwan

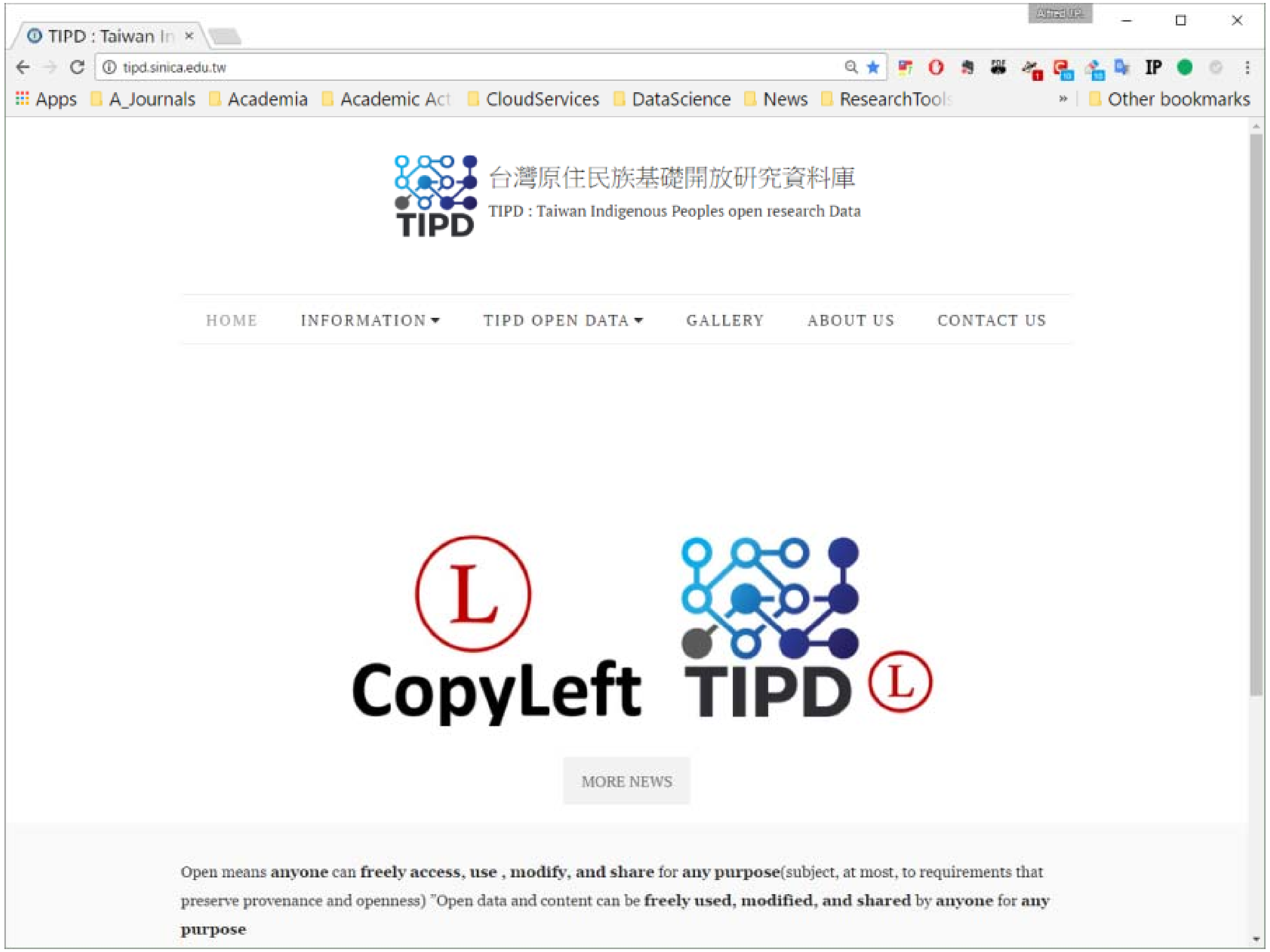
Data Science as a Foundation Toward Open Data and Open Science: The Case of Taiwan Indigenous Peoples open research Data (TIPD)

Ji-Ping Lin (RCHSS, Academia Sinica, Taiwan; email: jplin@sinica.edu.tw)

Acknowledgements: The research acknowledges financial & administrative supports from Council of Indigenous Peoples, Academia Sinica, Ministry of Science & Technology.

1. What TIPD Is & Its Aims

- **TIPD open data:** the research constructs the HDI of TIPs based on TIPD (Taiwan Indigenous Peoples open research Data, see http://TIPD.sinica.edu.tw;

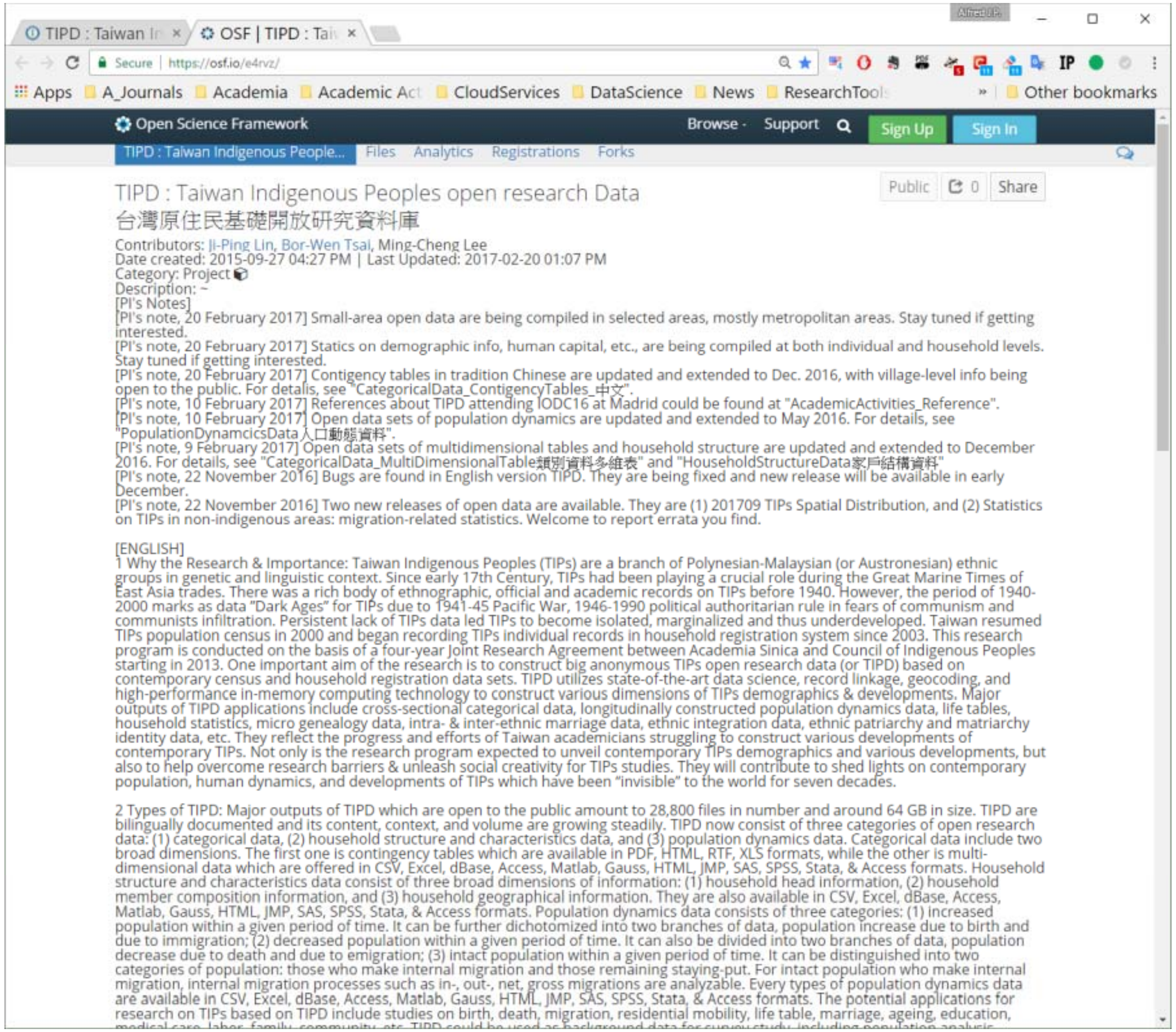


台灣原住民族基礎開放研究資料庫
TIPD : Taiwan Indigenous Peoples open research Data

Open means **anyone** can **freely access, use , modify, and share** for **any purpose**(subject, at most, to requirements that preserve provenance and openness) "Open data and content can be **freely used, modified, and shared** by **anyone** for **any purpose**

1. What TIPD Is & Its Aims (cont'd)

- Repository site of TIPD: Nature-recommend Open Science Framework @ https://osf.io/e4rvz/.



TIPD : Taiwan Indigenous Peoples open research Data
台灣原住民基礎開放研究資料庫

Contributors: Ji-Ping Lin, Bor-Wen Tsai, Ming-Cheng Lee
Date created: 2015-09-27 04:27 PM | Last Updated: 2017-02-20 01:07 PM
Category: Project
Description: ~
[PI's Notes]
[PI's note, 20 February 2017] Small-area open data are being compiled in selected areas, mostly metropolitan areas. Stay tuned if getting interested.
[PI's note, 20 February 2017] Statics on demographic info, human capital, etc., are being compiled at both individual and household levels. Stay tuned if getting interested.
[PI's note, 20 February 2017] Contigency tables in tradition Chinese are updated and extended to Dec. 2016, with village-level info being open to the public. For detalis, see "CategoricalData_ContigencyTables_中文".
[PI's note, 10 February 2017] References about TIPD attending IODC16 at Madrid could be found at "AcademicActivities_Reference".
[PI's note, 10 February 2017] Open data sets of population dynamics are updated and extended to May 2016. For details, see "PopulationDynamcicsData人口動態資料".
[PI's note, 9 February 2017] Open data sets of multidimensional tables and household structure are updated and extended to December 2016. For details, see "CategoricalData_MultiDimensionalTable類別資料多維表" and "HouseholdStructureData家戶結構資料"
[PI's note, 22 November 2016] Bugs are found in English version TIPD. They are being fixed and new release will be available in early December.
[PI's note, 22 November 2016] Two new releases of open data are available. They are (1) 201709 TIPs Spatial Distribution, and (2) Statistics on TIPs in non-indigenous areas: migration-related statistics. Welcome to report errata you find.

[ENGLISH]
1 Why the Research & Importance: Taiwan Indigenous Peoples (TIPs) are a branch of Polynesian-Malaysian (or Austronesian) ethnic groups in genetic and linguistic context. Since early 17th Century, TIPs had been playing a crucial role during the Great Marine Times of East Asia trades. There was a rich body of ethnographic, official and academic records on TIPs before 1940. However, the period of 1940-2000 marks as data "Dark Ages" for TIPs due to 1941-45 Pacific War, 1946-1990 political authoritarian rule in fears of communism and communists infiltration. Persistent lack of TIPs data led TIPs to become isolated, marginalized and thus underdeveloped. Taiwan resumed TIPs population census in 2000 and began recording TIPs individual records in household registration system since 2003. This research program is conducted on the basis of a four-year Joint Research Agreement between Academia Sinica and Council of Indigenous Peoples starting in 2013. One important aim of the research is to construct big anonymous TIPs open research data (or TIPD) based on contemporary census and household registration data sets. TIPD utilizes state-of-the-art data science, record linkage, geocoding, and high-performance in-memory computing technology to construct various dimensions of TIPs demographics & developments. Major outputs of TIPD applications include cross-sectional categorical data, longitudinally constructed population dynamics data, life tables, household statistics, micro genealogy data, intra- & inter-ethnic marriage data, ethnic integration data, ethnic patriarchy and matriarchy identity data, etc. They reflect the progress and efforts of Taiwan academicians struggling to construct various developments of contemporary TIPs. Not only is the research program expected to unveil contemporary TIPs demographics and various developments, but also to help overcome research barriers & unleash social creativity for TIPs studies. They will contribute to shed lights on contemporary population, human dynamics, and developments of TIPs which have been "invisible" to the world for seven decades.

2 Types of TIPD: Major outputs of TIPD which are open to the public amount to 28,800 files in number and around 64 GB in size. TIPD are bilingually documented and its content, context, and volume are growing steadily. TIPD now consist of three categories of open research data: (1) categorical data, (2) household structure and characteristics data, and (3) population dynamics data. Categorical data include two broad dimensions. The first one is contingency tables which are available in PDF, HTML, RTF, XLS formats, while the other is multi-dimensional data which are offered in CSV, Excel, dBase, Access, Matlab, Gauss, HTML, JMP, SAS, SPSS, Stata, & Access formats. Household structure and characteristics data consist of three broad dimensions of information: (1) household head information, (2) household member composition information, and (3) household geographical information. They are also available in CSV, Excel, dBase, Access, Matlab, Gauss, HTML, JMP, SAS, SPSS, Stata, & Access formats. Population dynamics data consists of three categories: (1) increased population within a given period of time. It can be further dichotomized into two branches of data, population increase due to birth and due to immigration; (2) decreased population within a given period of time. It can also be divided into two branches of data, population decrease due to death and due to emigration; (3) intact population within a given period of time. It can be distinguished into two categories of population: those who make internal migration and those remaining staying-put. For intact population who make internal migration, internal migration processes such as in-, out-, net, gross migrations are analyzable. Every types of population dynamics data are available in CSV, Excel, dBase, Access, Matlab, Gauss, HTML, JMP, SAS, SPSS, Stata, & Access formats. The potential applications for research on TIPs based on TIPD include studies on birth, death, migration, residential mobility, life table, marriage, ageing, education, medical care, labor, family, community, etc. TIPD could be used as background data for survey study, including population analysis

1. What TIPD is & Its Aims (cont’d)

- ■ **Why TIPD is designed as open data**
- ✓ Open sources as an effective ways of collective wisdom & improvement;
- ✓ The main goal of open sources is freethat serves as the role of unleashing creativity;
- ✓ Open does not mean “at the costs of sacrificing privacy, confidentiality, and ethics”, rather it promotes transparency and thus security.

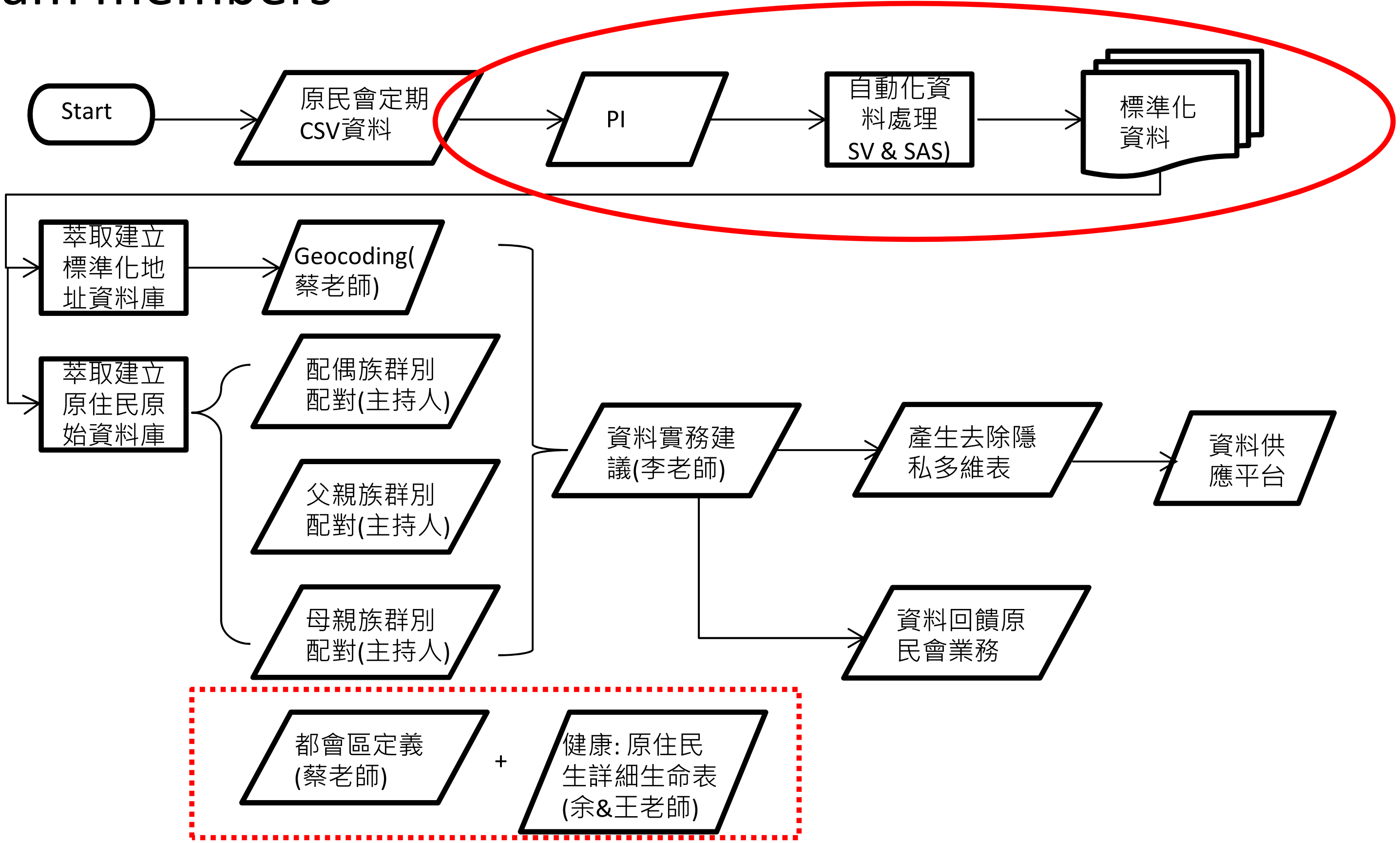
e.g.

1. The deaths & rebirths of IBM, Microsoft, & Apple…etc.
2. WWW and Linux etc…

1. What TIPD Is & Its Aims (cont'd)

■ Why TIPD is designed as open data

- ✓ To overcome in-house data lab restrictions
- ✓ To enhance data analyses efficiency and flexibility for team members



Start

原民會定期 CSV資料

PI

自動化資料處理 SV & SAS)

標準化資料

萃取建立標準化地址資料庫

Geocoding(蔡老師)

萃取建立原住民原始資料庫

配偶族群別配對(主持人)

父親族群別配對(主持人)

母親族群別配對(主持人)

資料實務建議(李老師)

產生去除隱私多維表

資料供應平台

資料回饋原民會業務

都會區定義(蔡老師)

+

健康: 原住民生詳細生命表(余&王老師)

1. What TIPD is & Its Aims (cont'd)

- If TIPD is designed as open data for research team members, why not make it open to the public?
- Thus, data sets of TIPD are placed on Open Science Framework.
- Its aims:

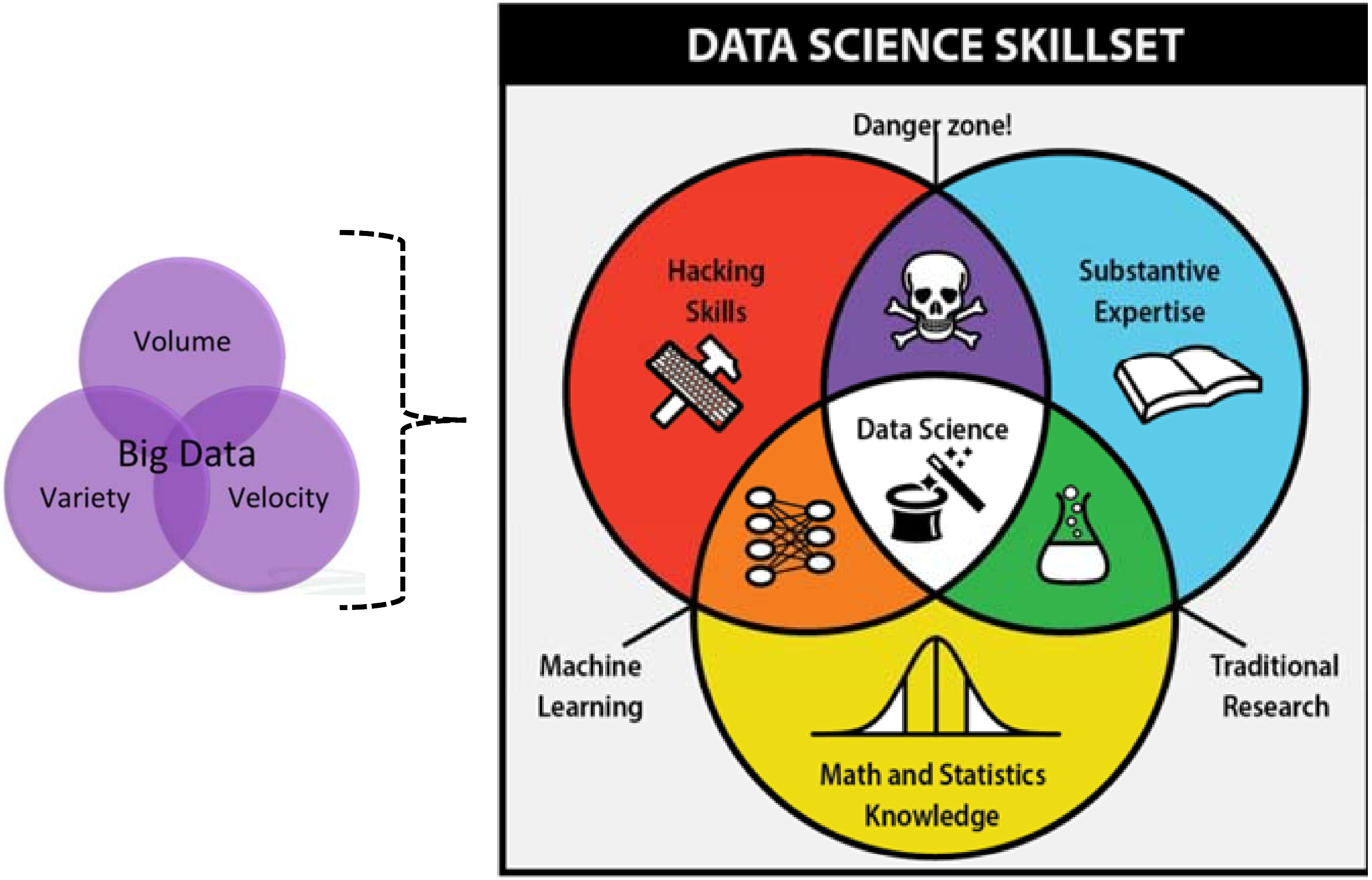
CopyLeft (L)

2. Contents and Context of TIPD

- ■ **Principles of constructing TIPD: being friendly & easy access & ease of use for "ordinary people"**
- ■ **Types of data in TIPD**
  - ✓ Cross-sectional multi-dimensional time-series data sets;
  - ✓ Longitudinal multi-dimensional data sets
  - ✓ Household structure and characteristics data are cross-sectional multi-dimensional time-series data sets
  - ✓ Population dynamics data
  - ✓ Data formats: they are available in PDF, HTML, RTF, XLS formats, while the other is multi-dimensional data which are offered in CSV, Excel, dBase, Access, Matlab, Gauss, HTML, JMP, SAS, SPSS, Stata, & Access formats.

2. Contents and Context of TIPD

- **Data Science** as foundation of constructing TIPD



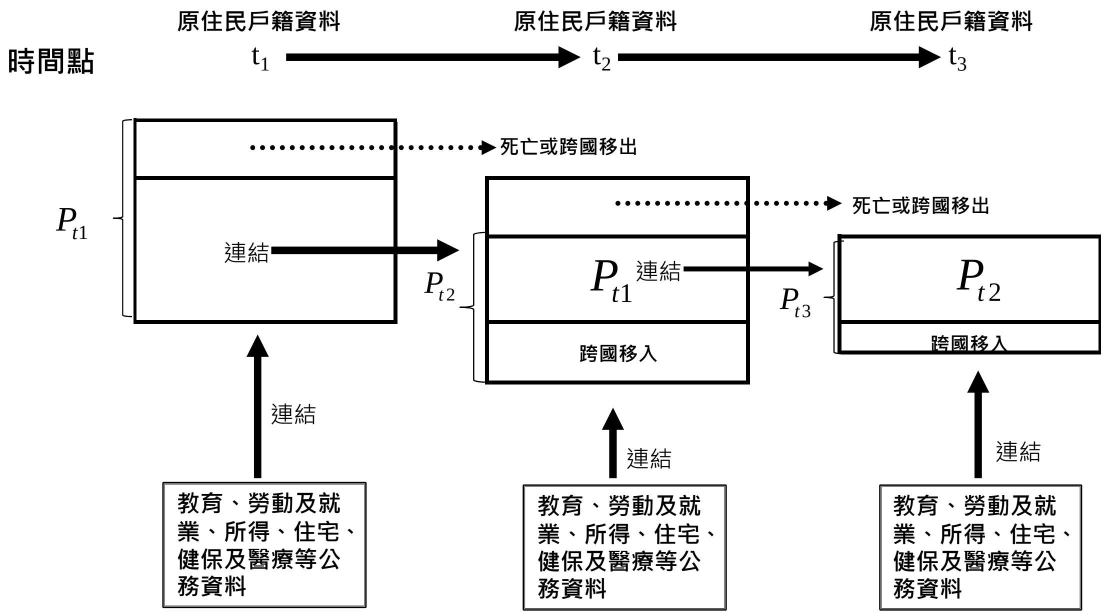
Source: O'Neil and Schutt 2013

2. Contents and Context of TIPD (cont'd)

✓Release of advanced TIPD open data sets in late 2015

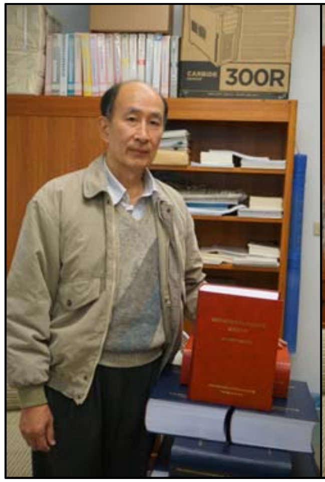
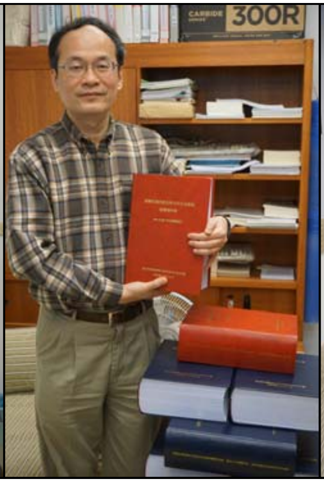
- ✓ Construct population dynamics data
  1. Pop'n of increase: comprising of "birth" & "immigration"
  2. Pop'n of decrease: comprising of "death" & "emigration"
  3. Pop'n of intact: comprising of "staying-put" & "internal migrants"
- ✓ Distinguish "death" from "emigration" records from data on "pop'n of decrease"

原住民基礎生活發展資料庫：人口及公務資料整合及動態結構

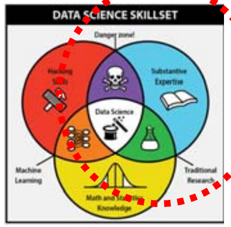


2. Contents and Context of TIPD (cont’d)

✓Debut of TIPD V1.0 in late 2014

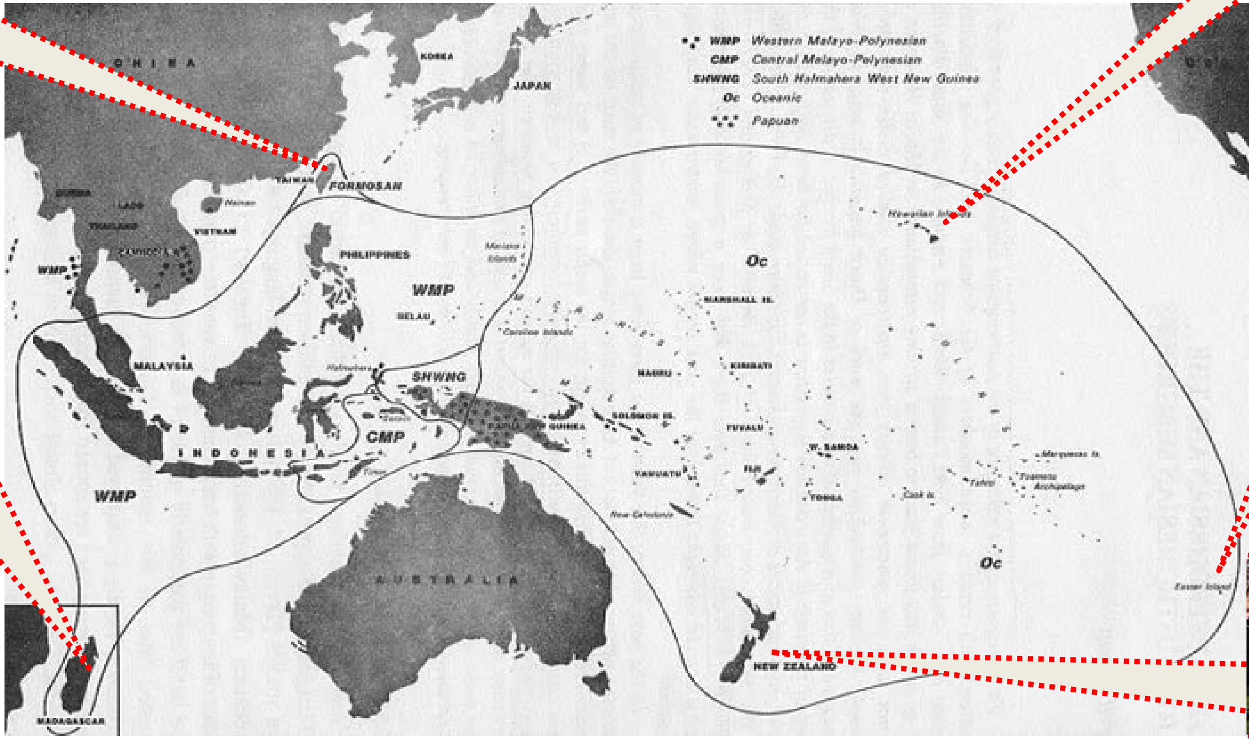



3. Substantive Expertise

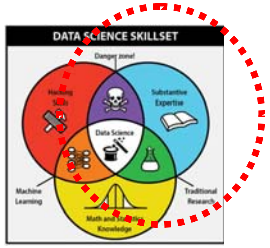


- Taiwan Indigenous peoples are a branch of Polynesian-Malaysian (or Austronesian) ethnic groups in genetic and linguistic context, whose ancestors have been living in Taiwan 8,000 years before the influx of Chinese immigrants in the 17th century.

Fig 1, Geographic Distribution of the Austronesians

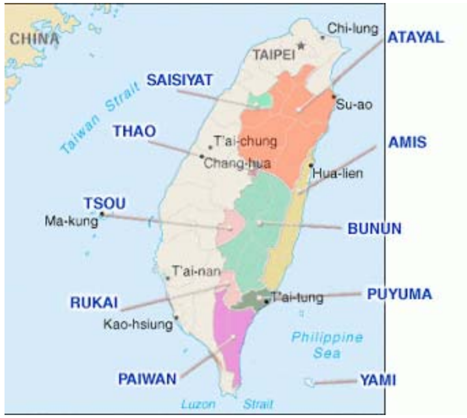


Source: *http://www.taiwandna.com/AborigineAustronesia.jpg*



3. Substantive Expertise (cont'd)

- ## A Look at TIPs
  (Taiwan Indigenous Peoples)



Amis

Bunun

Seediq

Tsou

Sakizaya

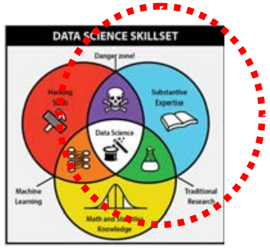
Paiwan

Rukai

Kavalan

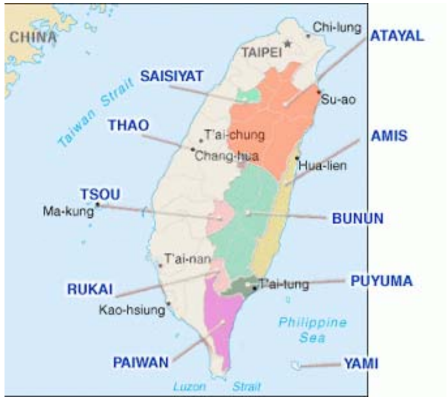
Source: http://thetaiwanphotographer.com/

3. Substantive Expertise (cont'd)



■ A Look at TIPs

(Taiwan Indigenous Peoples)



Saisiyat

Dao (Yami)

Truku

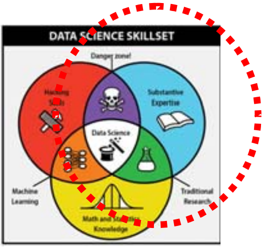
Puyuma

Thao

Dao (Yami)

Source: http://thetaiwanphotographer.com/

3. Substantive Expertise (cont'd)



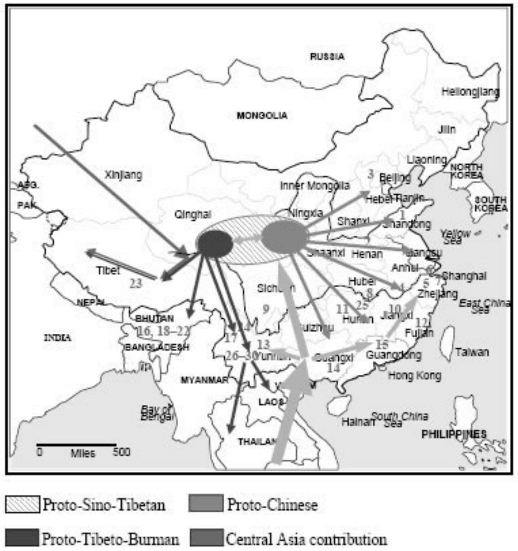
- Various Aspects of TIPS like linguistic system & culture infrastrure **don't support** "Traditional Wisdoms":

  e.g.,

  1) **Law of Geographic Proximity**
  2) **Zipf Power Law**

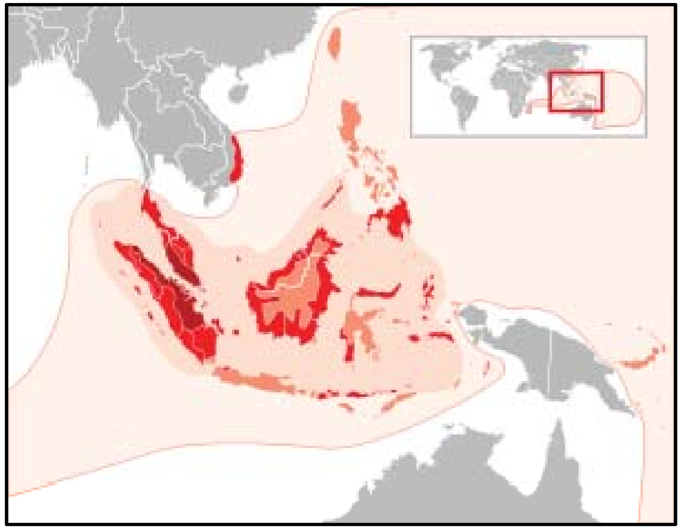
  e.g. Formosan languages are branch of Austronesian linguistic system, but are irrelevant to Tibetan-Han linguistic system.

Tibetan-Han languages



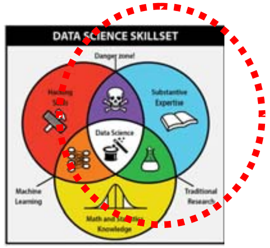
Source: http://historum.com/asian-history/77013-sino-tibetan-languages.html

Austanesian languages

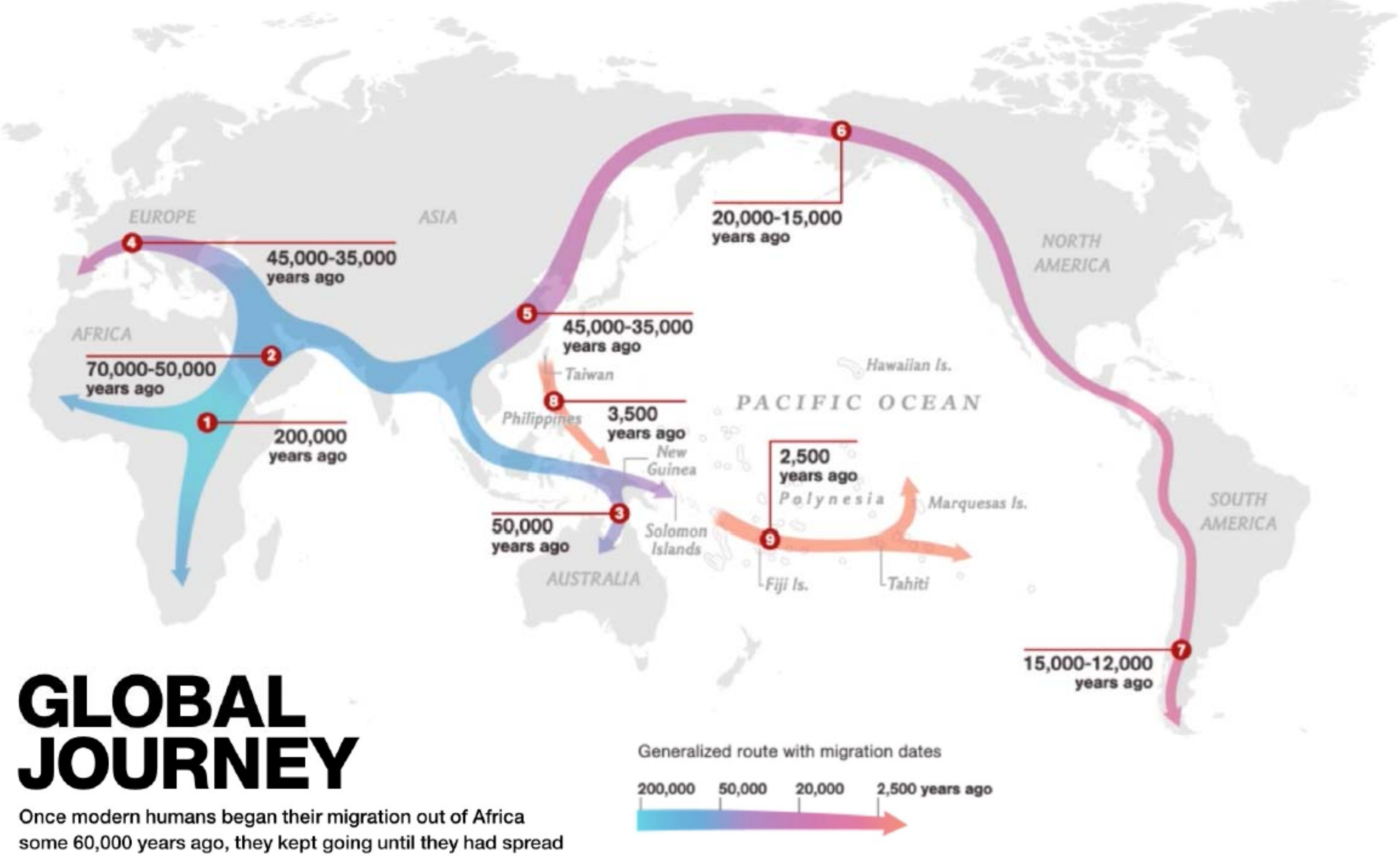


Source: https://en.wikipedia.org/wiki/Austronesian_languages

3. Substantive Expertise (cont'd)



- Global Journey of Modern Humans starting at 60,000 Years ago

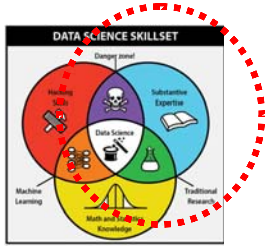


GLOBAL JOURNEY

Once modern humans began their migration out of Africa some 60,000 years ago, they kept going until they had spread to all corners of the Earth. How far and fast they went depended on climate, the pressures of population, and the invention of boats and other technologies. Less tangible qualities also sped their footsteps: imagination, adaptability, and an innate curiosity about what lay over the next hill.

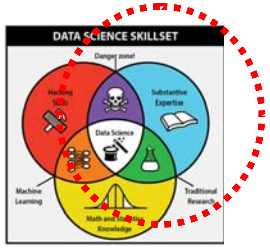
CREDITS i

3. Substantive Expertise (cont'd)



- There was a rich body of ethnographic, official and academic records on TIPs before 1940.
- However, the period of **1940-2000 marks as data "Dark Ages" for TIPs** due to 1941-45 Pacific War and 1946-1990 KMT authoritarian rule.
- Persistent lack of TIPs data led TIPs to become isolated and marginalized and thus underdeveloped.

3. Substantive Expertise (cont'd)



- Historical data: the past four centuries, Taiwan indigenous peoples experienced problems like political suppression, economic deprivation, social exclusion, and cultural sustainability in the face of a series of colonizing Dutch, Japanese, and Chinese regimes.

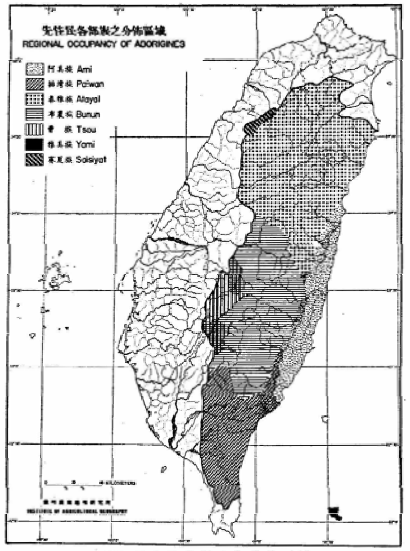


圖39. 高山族各部族之分佈區域

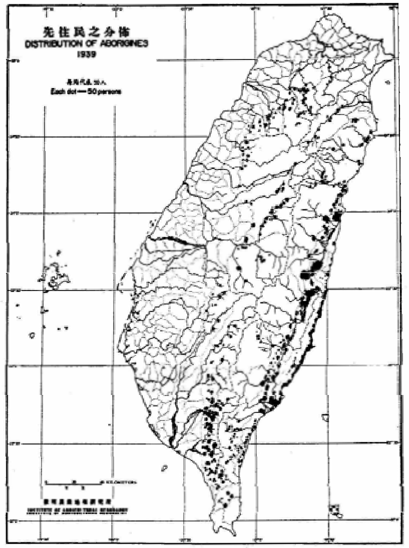

先住民之分佈
DISTRIBUTION OF ABORIGINES
1939
Each dot — 50 persons
KILOMETERS
INSTITUTE OF AGRICULTURAL GEOGRAPHY

圖40. 高山族人口之分佈，1939年

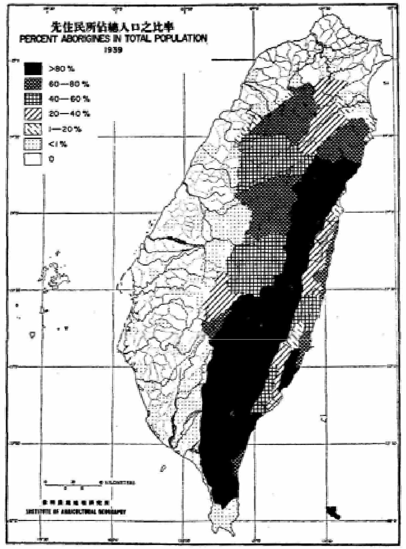

先住民所佔總人口之比率
PERCENT ABORIGINES IN TOTAL POPULATION
1939
>80 %
60—80 %
40—60 %
20—40 %
1—20 %
<1 %
0
KILOMETERS
INSTITUTE OF AGRICULTURAL GEOGRAPHY

圖41. 高山族人口佔總人口之%，1939年

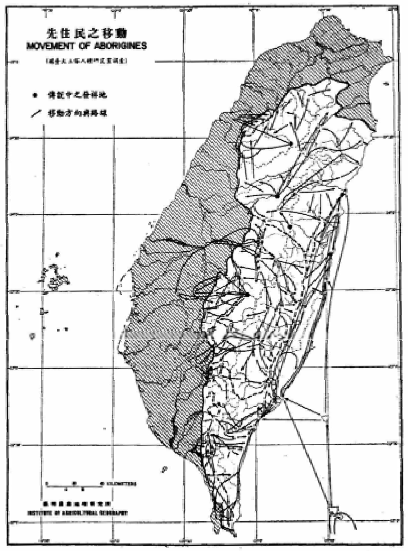
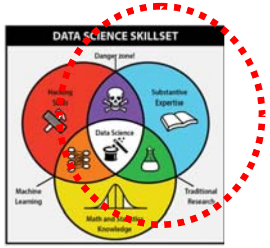


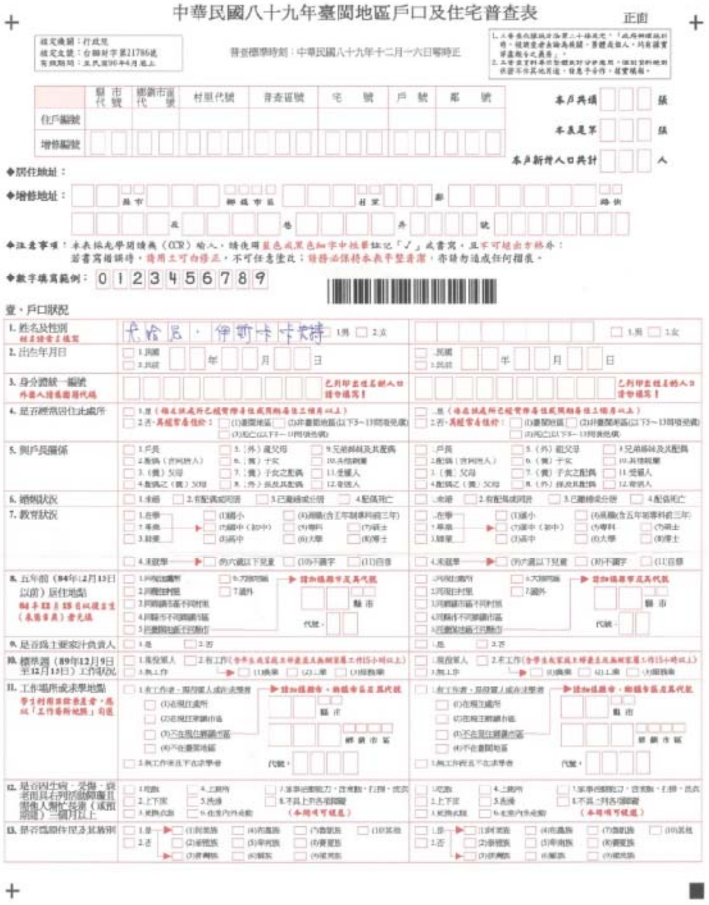
圖42. 高山族之移動

3. Substantive Expertise (cont'd)



✓ **Administration Household Data:**

- 2003: the **onset of ethnicity registration** on Household Registration System
- 2006-2009: Academia Sinica's **"Indigenous Population Survey"** research program, with 2000 Pop'n Census & 2007 household registration archive serving as basis for pre-survey pop'n analyses & sampling design;
- 2009-now: improving quality of ethnicity registration on Household Registration System


中華民國八十九年臺閩地區戶口及住宅普查表　正面

核定機關：行政院
核定文號：台89財字第21786號
有效期間：至民國90年4月底止

普查標準時刻：中華民國八十九年十二月十六日零時正

	縣市代號	鄉鎮市區代號	村里代號	普查區號	宅號	戶號	鄰號
住戶編號							
增修編號							

本戶共填 □□□ 張
本表是第 □□□ 張
本戶新增人口共計 □□□ 人

◆居住地址：
◆增修地址：縣市　鄉鎮市區　村里　鄰　路街　段　巷　弄　號

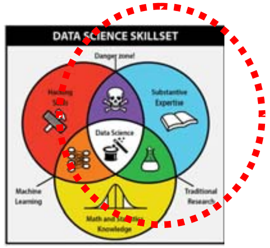
◆注意事項：本表採光學閱讀機（OCR）輸入，請使用藍色或黑色細字中性筆劃記「✓」或書寫，且不可超出方格外；若書寫錯誤時，請用立可白修正，不可任意塗改；請務必保持本表平整清潔，亦請勿造成任何摺痕。

◆數字填寫範例：0 1 2 3 4 5 6 7 8 9

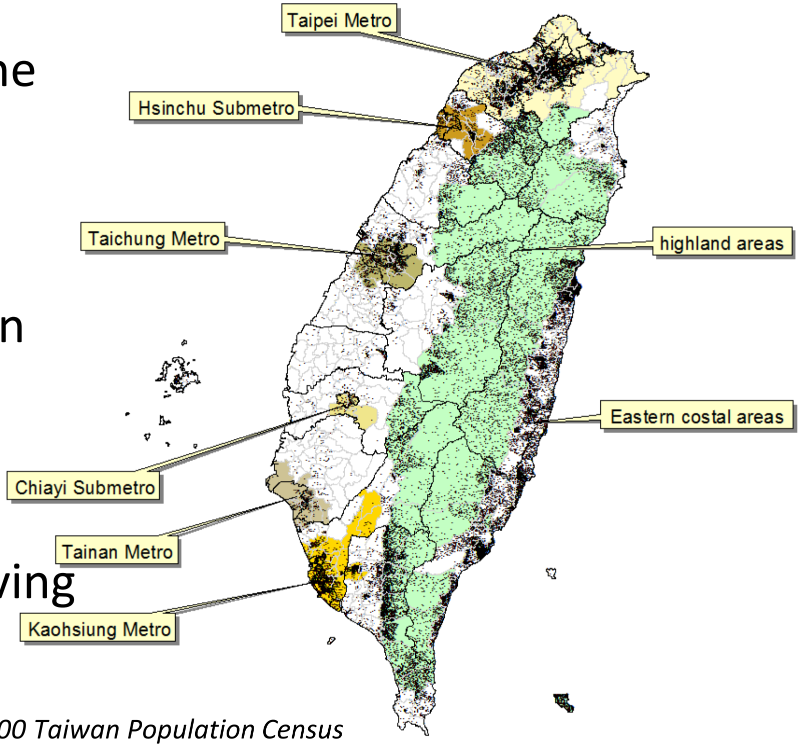
壹、戶口狀況

1. 姓名及性別　尤哈尼・伊斯卡卡夫特　□1.男 □2.女
2. 出生年月日　□1.民國 □2.民前　年　月　日
3. 身分證統一編號
4. 是否經常居住此處所
5. 與戶長關係
6. 婚姻狀況
7. 教育程度
8. 五年前（84年12月16日以前）居住地點
9. 是否為主要家計負責人
10. 標準週（89年12月9日至12月15日）工作狀況
11. 工作場所或求學地點
12. [illegible]
13. [illegible]

3. Substantive Expertise (cont'd)

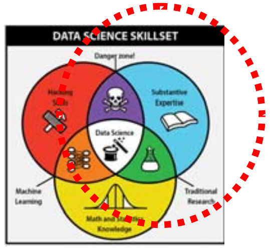


- TIPs share to total Taiwan population is of only 2.3%, the importance of research on TIPs lies in the following facts. Based on the author previous co-authored studies on the internal migration of TIPs, TIPs are characterized by four features in terms of population distribution and migration:
    1. geographically segregated population distribution,
    2. very migratory and mostly rural-to-urban migration,
    3. periphery of metropolitan areas serving as main destination choice for TIPs rural-to-urban migrants;
    4. weak ability of TIPs migrants to make onward migration and mostly choose return migration, once repeat migration occurs (see Map 1).



Source: 2000 Taiwan Population Census

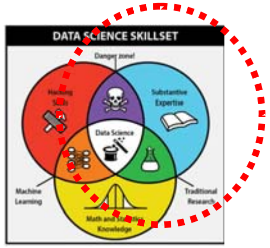
3. Substantive Expertise (cont'd)



- Contemporary Taiwan Indigenous peoples are ethnic minority. Similar to the situations of ethnic minority in the world, they are associated with higher unemployment, lower incomes, poorer health, shorter life span, etc.; e.g.,
  - ✓ Relative to non-indigenous peoples in terms of life expectancy in 2012,
  1. TIPS are 8.7 years shorter in general,
  2. 10.09 years shorter for males,
  3. 7.36 years shorter for females

- Although TIPs share to total Taiwan population is of only 2.3%, the importance of research on TIPs lies in the following facts.

3. Substantive Expertise (cont'd)



- Distribution Characteristics of Ethnic TIPs as Community Indicator & Social Embeddedness & structure of social network.

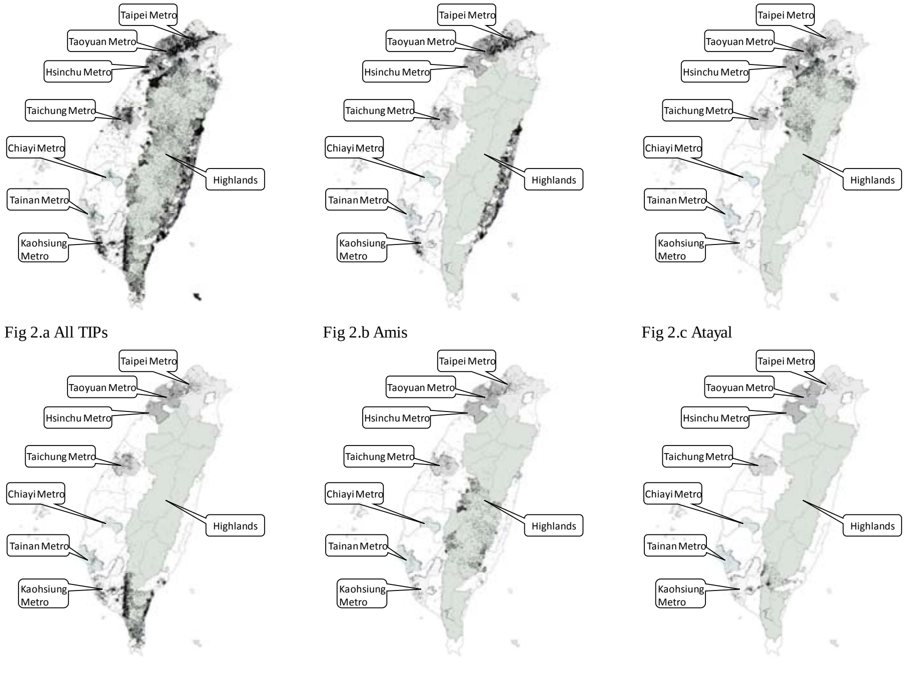


Fig 2.a All TIPs

Fig 2.b Amis

Fig 2.c Atayal

Fig 2.d Paiwan

Fig 2.e Bunun

Fig 2.f Rukai

Figure 2 Spatial population distribution of Taiwan indigenous peoples (TIPs) by ethnicity

Note 1 dot = 10 persons & figures are mapped by the author based on the 2013 year end of TIPs household registration data.

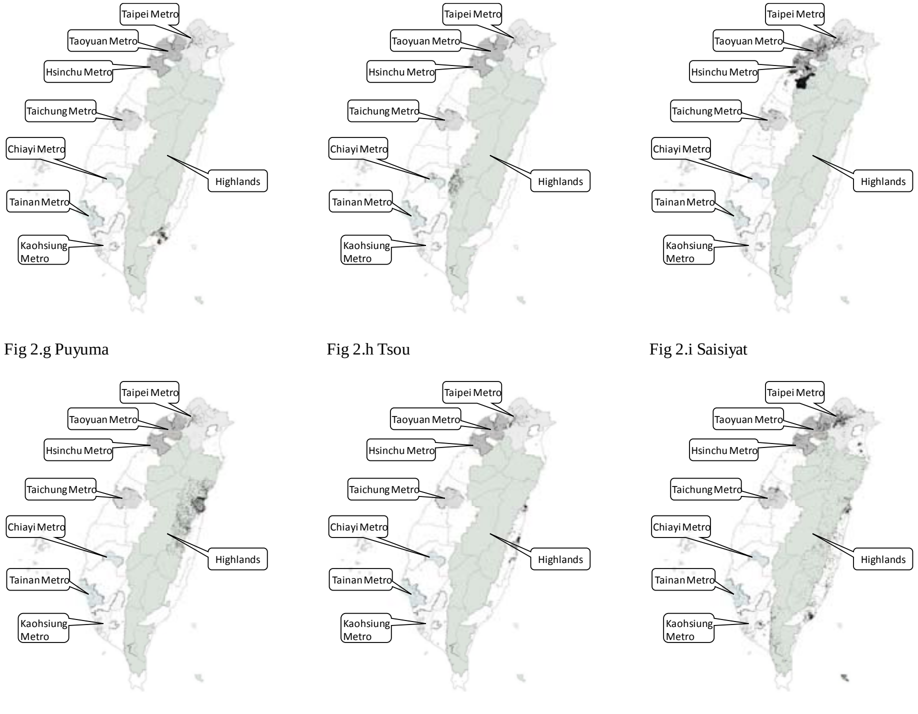


Fig 2.g Puyuma

Fig 2.h Tsou

Fig 2.i Saisiyat

Fig 2.j Truku

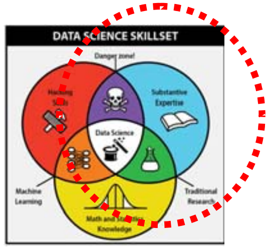
Fig 2.k Kavalan

Fig 2.l all others

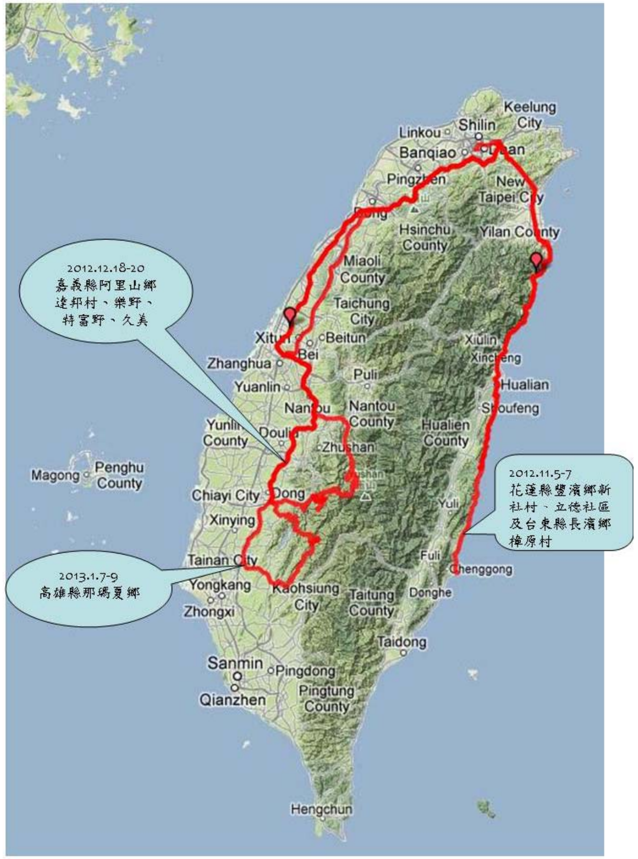
Figure 2 (cont'd) Spatial population distribution of Taiwan indigenous peoples (TIPs) by ethnicity

Note 1 dot = 10 persons & figures are mapped by the author based on the 2013 year end of TIPs household registration data.

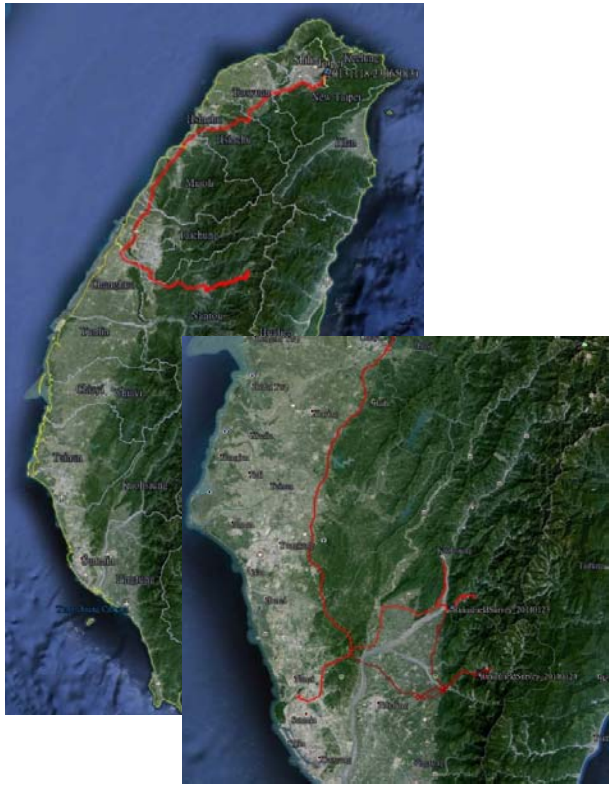
3. Substantive Expertise (cont'd)



- Formosan endanger languages surveys: 2012-2015

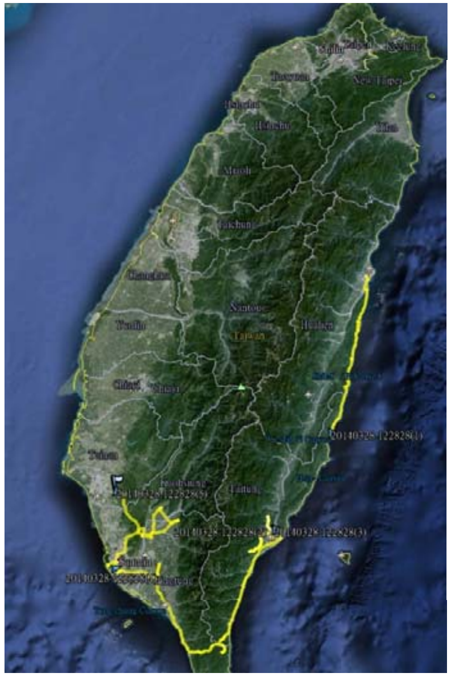


20131118-21 Seediq field survey

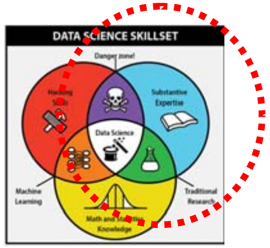


20140123-25 Rukai field survey

20140328-29 Puyuma field survey

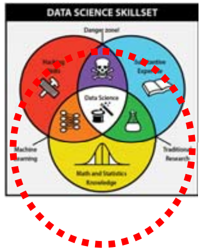


3. Substantive Expertise (cont'd)



- Formosan endanger languages surveys: 2012-2015
  - ✓ Survey, face-to-face interviews, ethnography study… etc
  - ✓ Collect more than 30,000 photos, 400 video & audio records

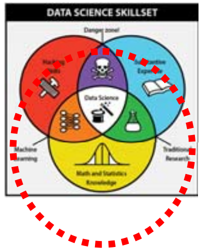
4. Methodology



■ The data

1. administrative data: Taiwan Household Registration Data (THRD)
2. THRD data sets are archived for the study on a monthly base, with the archived time point being the last day of each month
3. Information in micro data sets of THRD: Household ID, Time of data creation, PIN, name, spouse name, parents' names, education, age, marital status, address, birth place, mobility…

4. Methodology (cont’)

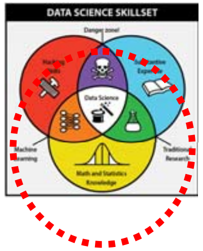


■ Methods used to **overcome legal & ethic issues:**

1. Giving up in-house data lab mode
2. Distributed storage of raw data + centralized data integration as the main methodology
3. Basic concepts of distributed data storage & centralized data integration:

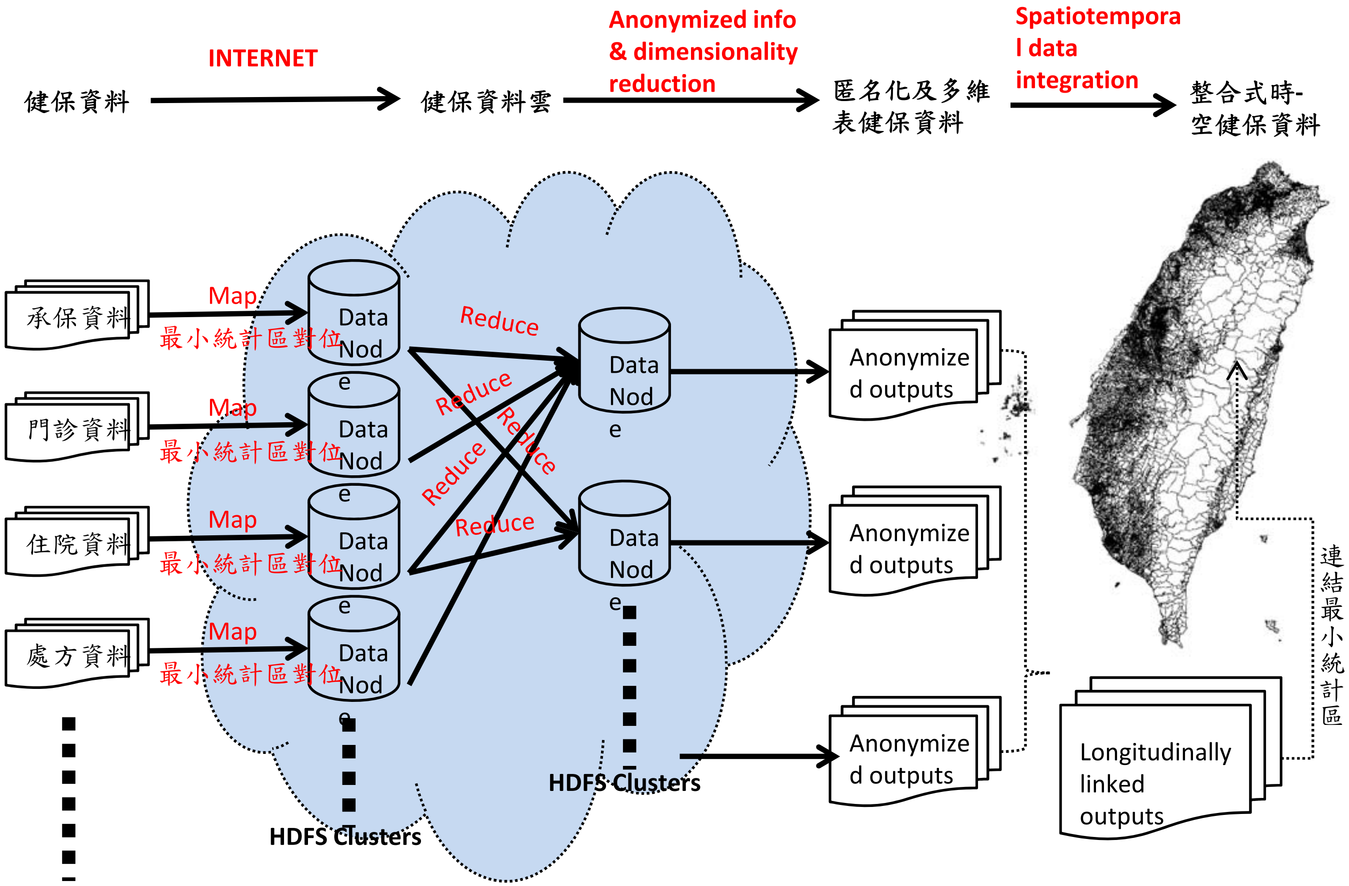
   (explain more about this concept here....)

4. Methodology (cont’)



■ Methods used to overcome legal & ethic issues:

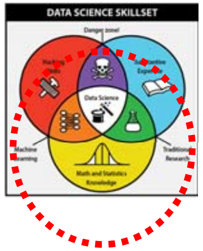
✓ Distributed Computing & Storage Network: the first tool that was considered to use at the beginning of research: Appache Hadoop (open sourced version of Google GDFD+MapReduce)



4. Methodology (cont')

■ Methods used to overcome legal & ethic issues:

✓ Construction of conventional "old-school" multi-dimensional tables is adopted as means for "distributed data storage" and "centralized data integration"



An cheap but effective way to preserve source data information & protect privacy

(1) Source data

Source data:

Individual ID	Sex (1: male; 2: female)	Age (years of age)
1	1	6
2	2	14
3	2	48
4	2	69
5	1	24
6	2	38
7	1	42
8	1	56
9	2	20
10	1	19

(2) Contingency table

Table: Frequency counts by Sex & Age

Age	Sex		
	Male (as of 1)	Female (as of 2)	Total
0-15	1	1	2
16-30	1	1	2
31-45	2	2	4
45-65	0	1	1
65+	0	1	1
Total	4	6	10

(3) Categories in contingency table

Table: Assignment of categories

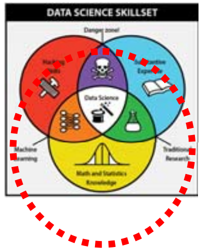
Age (B)		Sex (A)	
		A1=1	A2=2
		Male (as of 1)	Female (as of 2)
B1=1	0-15	(A1, B1)	(A2, B1)
B2=2	16-30	(A1, B2)	(A2, B2)
B3=3	31-45	(A1, B3)	(A2, B3)
B4=4	45-65	(A1, B4)	(A2, B4)
B5=5	65+	(A1, B5)	(A2, B5)

(4) Multidimensional tables

Format of Multidimensional Table Data

Sex	Age	Frequency as weight
1	1	1
1	2	1
1	3	2
1	4	0
1	5	0
2	1	1
2	2	1
2	3	2
2	4	1
2	5	1

4. Methodology (cont')



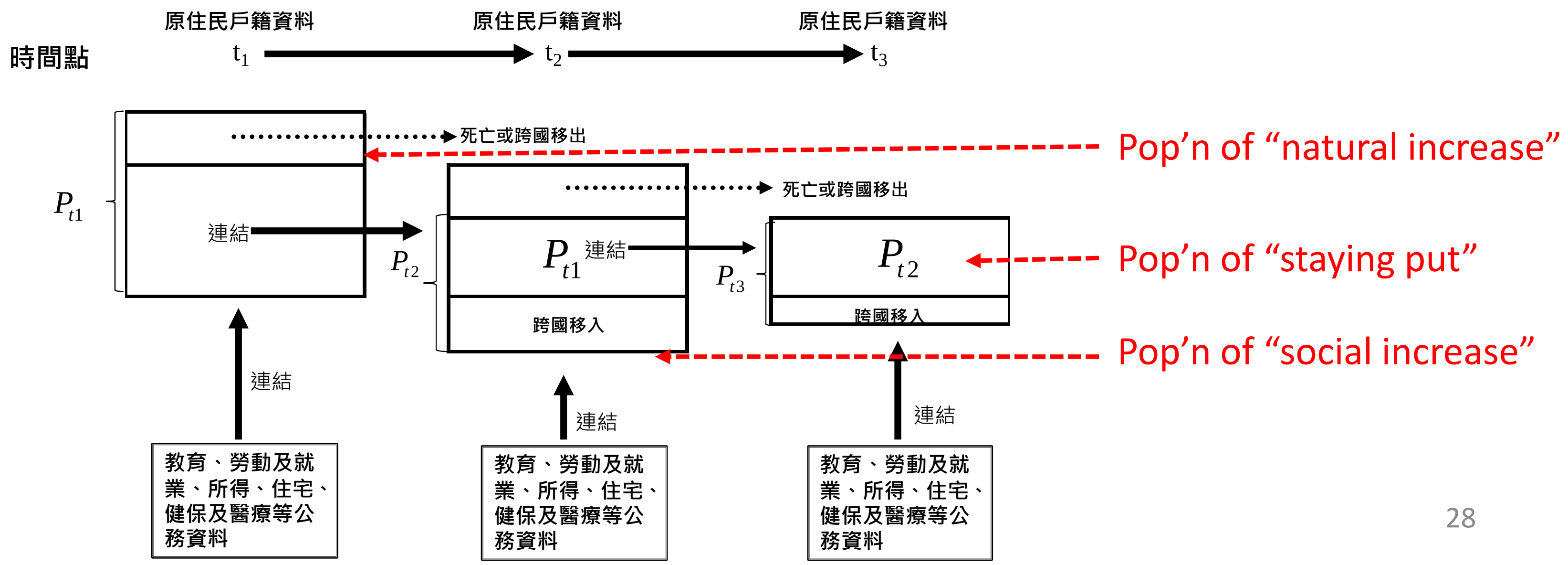
- ■ Data model

✓ Construct population dynamics data

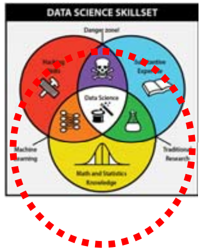
1. Pop'n of increase: comprising of "birth" & "immigration"
2. Pop'n of decrease: comprising of "death" & "emigration"
3. Pop'n of intact: comprising of "staying-put" & "internal migra

✓ Distinguish "death" from "emigration" records from data on "pop' decrease"

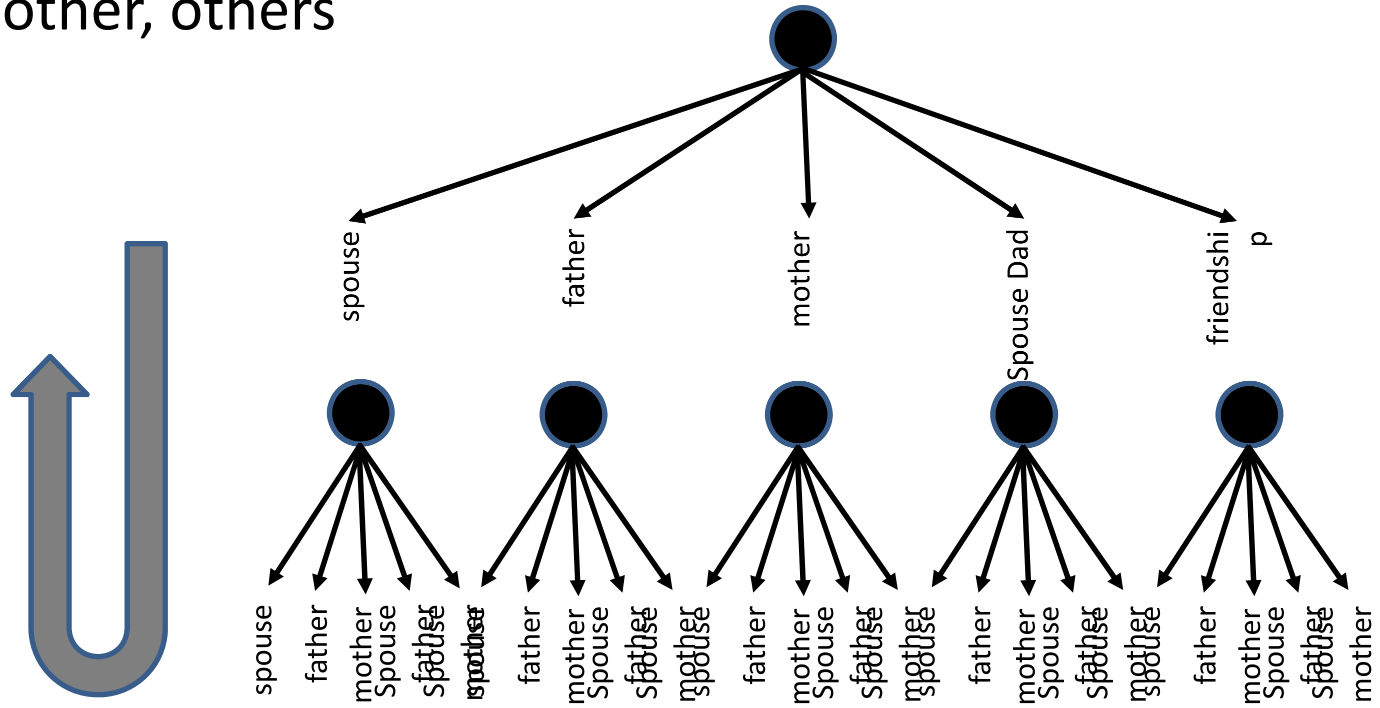
原住民基礎生活發展資料庫：人口及公務資料整合及動態結構


原住民戶籍資料
原住民戶籍資料
原住民戶籍資料
時間點
t_1
t_2
t_3
死亡或跨國移出
死亡或跨國移出
Pop'n of "natural increase"
P_{t1}
連結
P_{t2}
P_{t1}
連結
P_{t3}
P_{t2}
Pop'n of "staying put"
跨國移入
跨國移入
Pop'n of "social increase"
連結
連結
連結
教育、勞動及就業、所得、住宅、健保及醫療等公務資料
教育、勞動及就業、所得、住宅、健保及醫療等公務資料
教育、勞動及就業、所得、住宅、健保及醫療等公務資料

4. Methodology (cont'd)

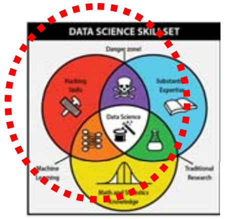


- ■ Data model
- ✓ Genealogy: Construction of Micro Kinship & Friendship Network
- ✓ Construction of kinship/friendship network includes: father, mother, spouse, spouse father, spouse mother, others



Recursively build-up process (see source code)

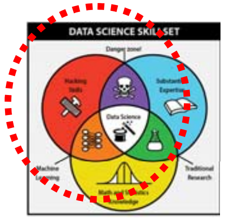
4. Methodology (cont'd)



- ■ Hacking skills & methods

- ✓ Current implementation strategy: fully utilize the advantages of 64-bit digital infrastructures to perform high-performance computing
- ✓ Current implementation digital infrastructure: hardware environment(Supermicro A7X9-7f mobo + dual Intel Xeon E5-2680v2 + 256GB ECC DDR3 1600 + 80GB RAM disk + RAID0 of 2*1TB SATA3 Micron Crucial MX200 SSD + nVidia GTX Titan…)
- ✓ **No longer a "dream machine" for individual researcher (100,000 USD in 2013 → 10,000 USD in 2015 → 5,000 USD now).**

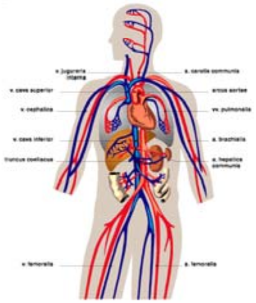
4. Methodology (cont'd)



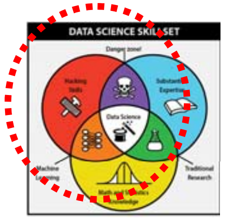
- Hacking skills & methods

✓ Hardware: genealogy computing methods: matching involves thousands of billion matching in TIPD accumulated data bank; to accelerate computing, we use: In-memory computing to achieve genealogy computing by overclocking digital hardware (1) CPUs & (2) IO bus & (3) DRAM.

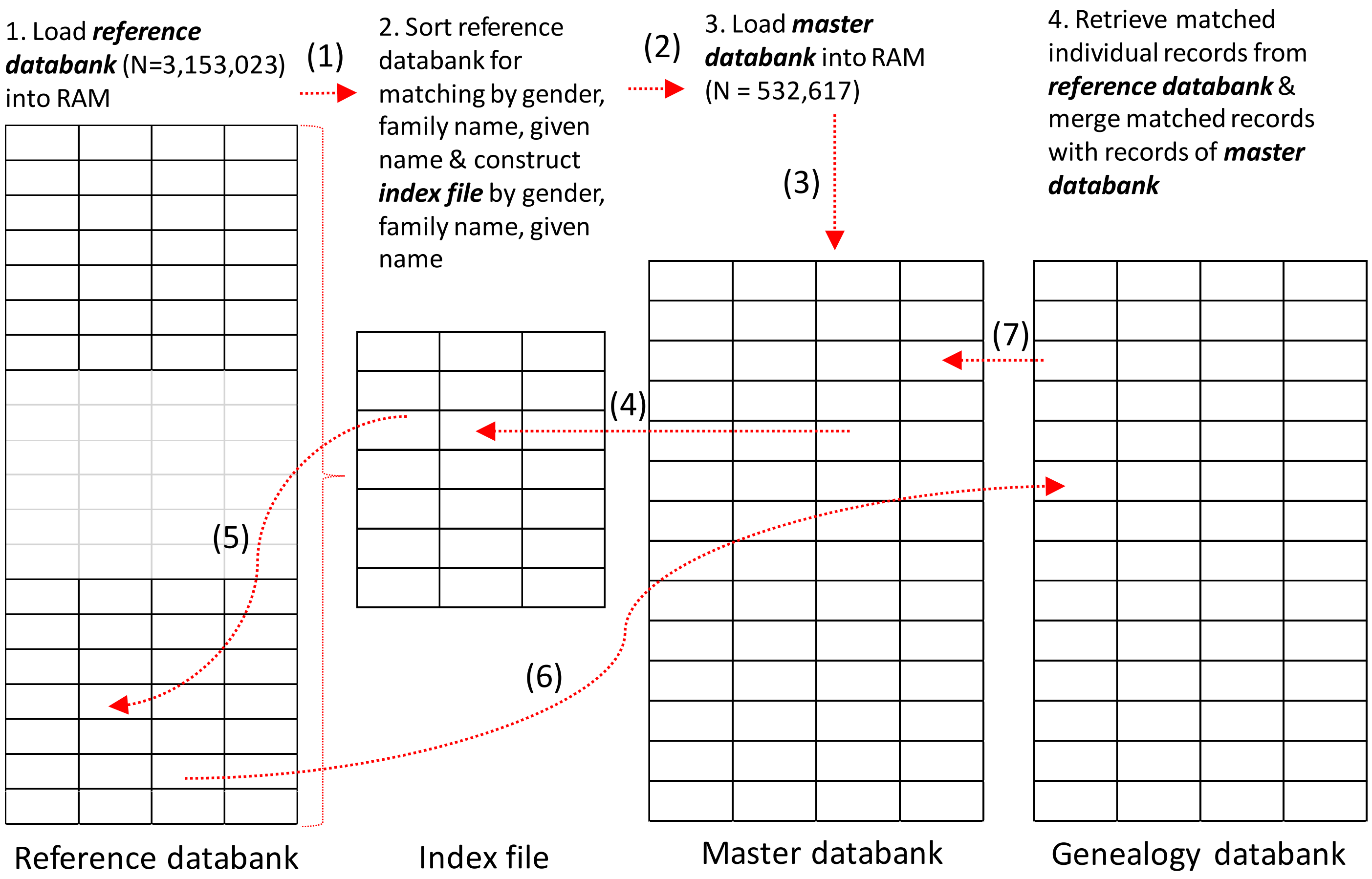
CPUs overclocking + I/O bus overclocking+ DRAM overclocking



4. Methodology (cont'd)



- ■ Hacking skills & methods
- ✓ Manipulation of High Performance Computing (HPC), e.g., Matching process of constructing micro genealogy



1. Load ***reference databank*** (N=3,153,023) into RAM

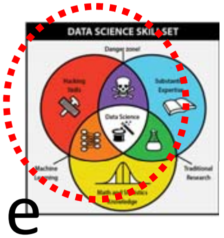
2. Sort reference databank for matching by gender, family name, given name & construct ***index file*** by gender, family name, given name

3. Load ***master databank*** into RAM (N = 532,617)

4. Retrieve matched individual records from ***reference databank*** & merge matched records with records of ***master databank***

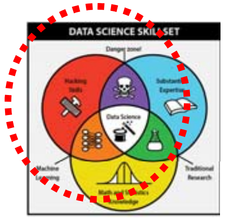
Figure 1 procedures of record matching using in-memory computing

5. What We Earn in Return?



- Reorganizing raw data as open data to overcome legal & ethic issues boosts academic & crowd sourcing (civil) research, e.g., Taiwan indigenous peoples study and international cooperation.
- To allow us to enrich data through the process of data integration methodology, making longitudinally linked administrative data less expensive and more efficient, e.g., population dynamics data, birth & data & migration processes…
- To allow us to do what was not able to do before, e.g., micro genealogy, identity, ethnic marriage pattern

5. What We Earn in Return?



- For examples, ethnic identity, ethnic marriage practice and social cohesion:

Table 1 Marriage practice and category of ethnic identity formation

Type of ethnic marriage practice			Type of ethnic identify formation	
Ethnic marriage Practice	Endogamy	Intra-ethnic endogamy	Mono-ethnic identity	
			Unspecified ethnic identity	
		Inter-ethnic endogamy	Multi-ethnic identity	Patrilineal ethnic identity
				Matrilineal ethnic identity
			Unspecified ethnic identity	
	Exogamy		Multi-ethnic identity	Patrilineal ethnic identity
				Matrilineal ethnic identity
			Unspecified ethnic identity	

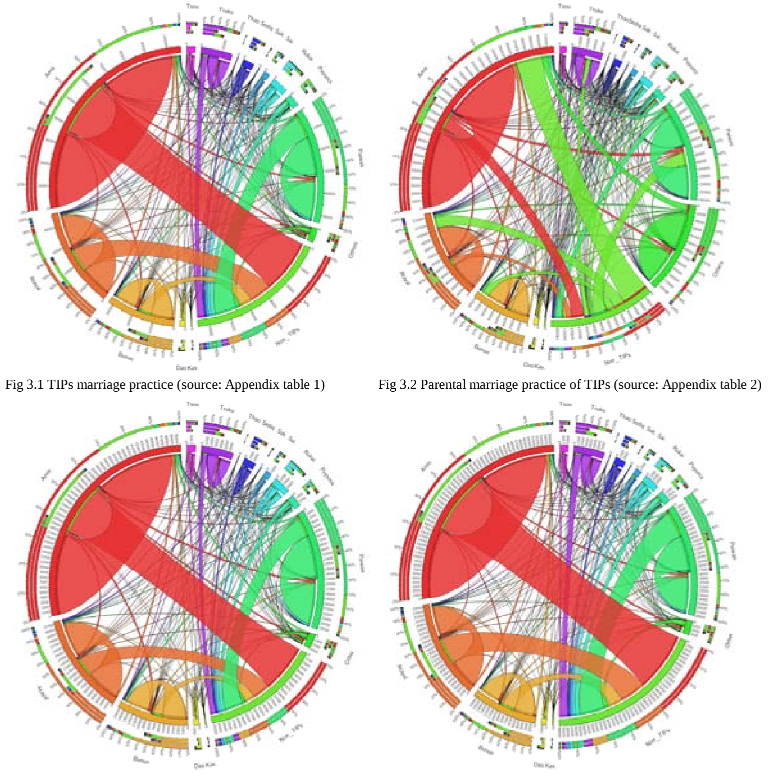


Fig 3.1 TIPs marriage practice (source: Appendix table 1)

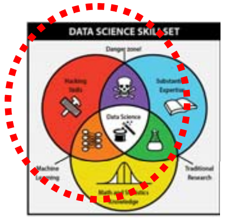
Fig 3.2 Parental marriage practice of TIPs (source: Appendix table 2)

Fig 3.3 Male TIPs marriage practice (source: Appendix table 3)

Fig 3.4 Female TIPs marriage practice (source: Appendix table 4)

Figure 3 TIPs marriage practice in circular layout by ethnic groups

6. What Challenges Ahead of Open Data?

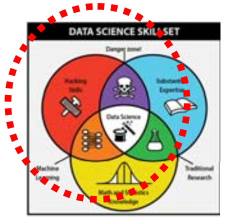


✓ The 22 Skills of a Data Scientist

(source: DataScienceCentral @ www.datasciencecentral.com)

1. Back-End Programming (ex: Assembly/C C++/Pascal Delphi/JAVA) DC, DD
2. Algorithms (ex: computational complexity, CS theory) DD,DR
3. Big and Distributed Data (ex: Hadoop, Map/Reduce) DB, DC, DD
4. Structured Data (ex: SQL, JSON, XML) DC, DD
5. Unstructured Data (ex: noSQL, text mining) DC, DD
6. Data Manipulation (ex: regexes, R, SAS, web scraping) DC, DR
7. Web Programming (ex: JavaScript, HTML, CSS) DC, DD
8. Systems Administration (ex: *nix, DBA, cloud tech.) DC, DD
9. Math (ex: linear algebra, real analysis, calculus) DD,DR
10. Optimization (ex: linear, integer, convex, global) DD, DR
11. Science (ex: experimental design, technical writing/publishing) DC, DR
12. Classical Statistics (ex: general linear model, ANOVA) DB, DC, DR
13. Bayesian/Monte-Carlo Statistics (ex: MCMC, BUGS) DD, DR
14. Machine Learning (ex: decision trees, neural nets, SVM, clustering) DC, DD
15. Temporal Statistics (ex: forecasting, time-series analysis) DC, DR
16. Spatial Statistics (ex: geographic covariates, GIS) DC, DR
17. Graphical Models (ex: social networks, Bayes networks) DD, DR
18. Simulation (ex: discrete, agent-based, continuous) DD,DR
19. Visualisation (ex: statistical graphics, mapping, web-based data?viz) DC, DR
20. Business (ex: management, business development, budgeting) DB
21. Surveys and Marketing (ex: multinomial modeling) DC, DR
22. Product Development (ex: design, project management) DB

6. What Challenges Ahead of Open Data?

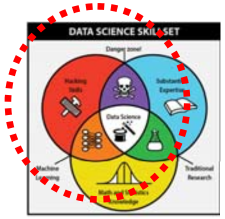


- ✓ Challenges from computing

According to National Research Council. 2013. ***Frontiers in Massive Data Analysis.*** Washington, D.C.: The National Academies Press, future challenges include:

1. Dealing with highly distributed data sources,
2. Tracking data provenance, from data generation through data preparation,
3. Validating data,
4. Coping with sampling biases and heterogeneity,
5. Working with different data formats and structures,
6. Ensuring data integrity, data security,
7. Enabling data discovery, integration, sharing,
8. Developing algorithms that exploit parallel and distributed architectures,
9. Developing methods for visualizing massive data,
10. Developing scalable and incremental algorithms, and
11. Coping with the need for real-time analysis and decision-making.

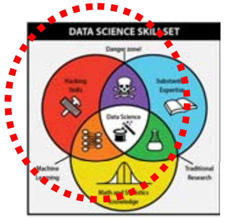
6. What Challenges Ahead of Open Data?



✓ Challenges from manipulation of digital infrastructure

1. To work in massive data analysis will require experience with massive data and with computational infrastructure that permits the real problems associated with massive data to be revealed,
2. There are computational constraints that arise within any particular problem domain that help to determine,
3. the specialized algorithmic strategy to be employed. Most work in the past has focused on a setting that involves a single processor with the entire data set fitting in random access memory (RAM).
4. Additional important settings for which algorithms are needed include the following:
    1) The streaming setting, in which data arrive in quick succession, and only a subset can be stored;
    2) The disk-based setting, in which the data are too large to store in RAM but fit on one machine’s disk;
    3) The distributed setting, in which the data are distributed over multiple machines’ RAMs or disks; and
    4) The multi-threaded setting, in which the data lie on one machine having multiple processors that share RAM.

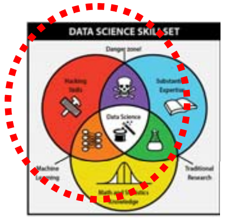
6. What Challenges Ahead of Open Data?



✓ Manipulation of digital infrastructure: an example

1. CPU Instruction Set : processor's built-in code;
2. CPU On-Board Level-2 (L2) Cache: enables the CPU to access repeatedly used data directly from its own on-board memory, rather than repeatedly requesting it from the system RAM. L2 Cache is very critical to applications such as games, video editing, and 3-D applications such as CAD/CAM programs. It's less important for activities such as web surfing, email, and word processing;
3. CPU Clock Speed : a measure of how many instructions the processor can execute in one second (like speed limit on a highway);
4. CPU Bandwidth : measured in bits, the bandwidth determines how much information the processor can process in one instruction (like the number of lanes on a highway);
5. Front Side Bus (FSB)/QPI/DMI… Speed: The FSB/QPI/DMI is the interface between the processor and the system memory. The CPU‘s FSB speed determines the maximum speed at which it can transfer data to the rest of the system;
6. Motherboard chipset/controller clock speed, and RAM speed
7. Heat and Heat Dissipation
8. Operating System and Application softwares

5. Concluding Remarks



Potential Contributions of TIPD

- **From "Close" to "Open":** the research on TIPD contributes to shed lights on contemporary geography of Taiwan Indigenous Peoples and human dynamics which have *been "invisible" to the world for seven decades;*

- **From "Elite" to "Ordinary":** based on data science & household register records, the constructed open data sets reduce tech-barriers for researchers interested in indigenous population studies;

- **From "Local" to "Global":** English beta version of TIPD are open to international academic communities in December 2015, aiming to promote further value-added data enrichment through crowd-sourcing collaboration for international comparative studies.

- **From "Macro" to "Micro":** e.g., micro social network data will be reorganized in categorized open data format & open to the public this year.