Summer Working Group for Employer List Linking (SWELL)

Graton Gathright, Mark Kutzbach, Kristin Mccue, Erika McEntarfer,
Holly Monti, Kelly Trageser, Lars Vilhuber, Nada Wasi, Christopher Wignall

May 22, 2014

NCRN Meeting Spring 2014

Washington, DC

Summer Working Group for Employer List Linking (SWELL)

- Collaboration of researchers on four projects with shared methodology requirements
  - US Census Bureau
    - Social, Economic, and Housing Statistics Division (SEHSD)
    - Center for Economic Studies (CES)
  - Cornell and University of Michigan
  - Potentially consult with Rebecca Steorts (CMU) & Jerry Reiter (Duke)
- Working to develop tools for linking person-level survey responses to employer information in administrative records files using probabilistic record linkage

Payoff from linkages:

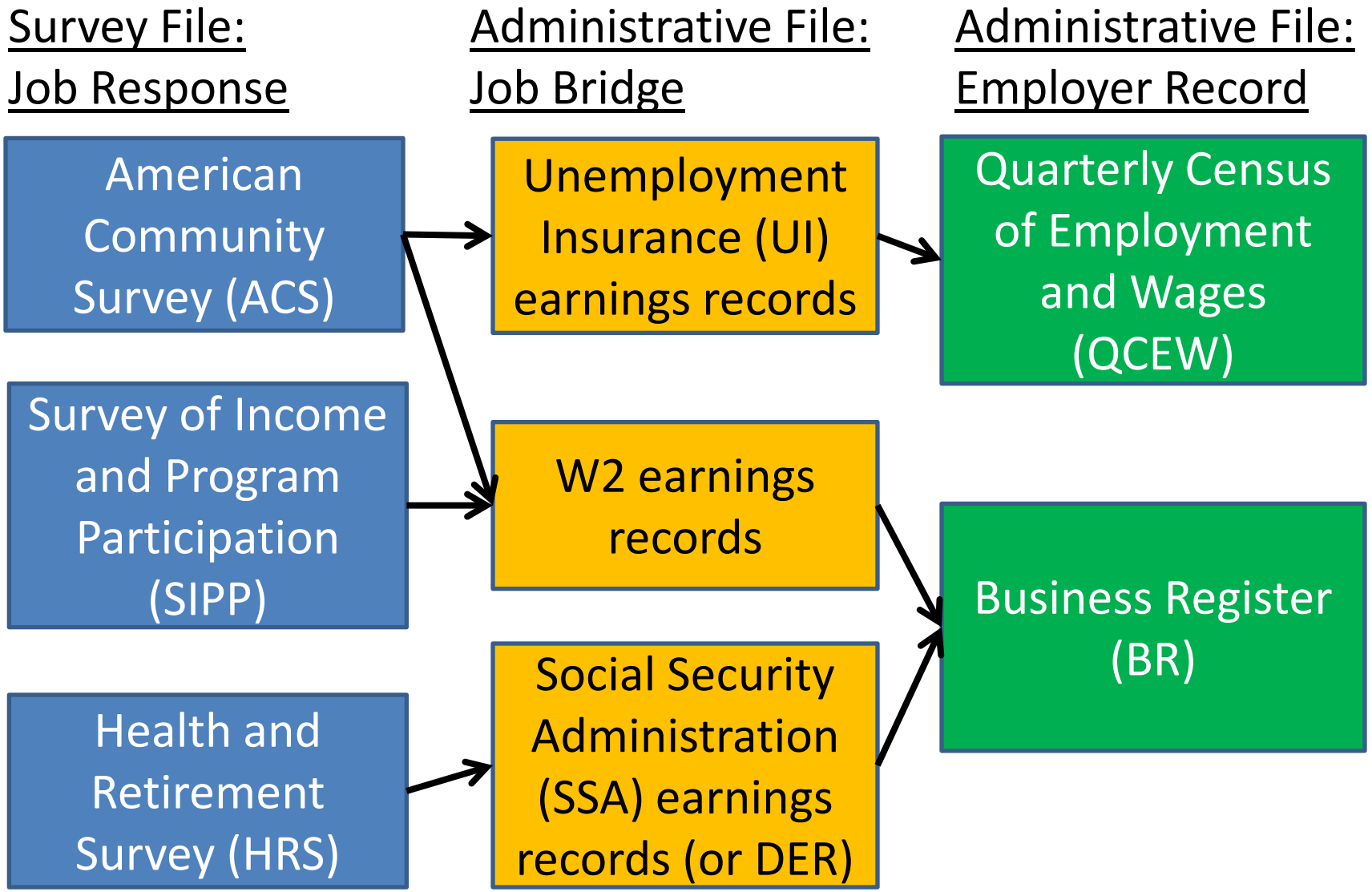
- Produce research-ready crosswalk between survey responses and administrative employer records
  - Quality metrics to help users assess the probability that a particular link is correct
- Compare self-reported vs. admin measures (e.g., location, earnings, firm size, industry, lay-offs)
  - Enhance data quality by improving edits and imputations
- Make improved/new measures available to users without increasing respondent burden
- Investigate new research questions that could not be answered by either dataset alone (e.g. new variables, longitudinal outcomes or histories)

Challenges	Solutions
How to narrow the list of candidates to a manageable set?	We use administrative records for blocking on job histories
How to measure the similarity of employer names (rather than person names)?	We develop a new standardizer/parser for business names
How to reflect the uncertainty of a match, with greater distinction than match/non-match?	Our clerical review trains the model to classify some records as a possible match and also reflects differences in reviewer assessments. We retain all matches and possible matches
How to maintain the match file, replicate results, or pass on learning to other groups?	We are producing a toolkit, testing it on 4 projects, and producing documentation

Presentation Outline

- Constituent projects and datasets
- Linking Methodology
  - Blocking strategy
  - Probabilistic record linkage
  - Standardizing and parsing
  - Comparators
  - Training set and clerical review
- Progress and current work
- Potential extensions

Data Linking Frameworks



Person-level survey responses

American Community Survey (ACS)

- ~ 3 million households, annual survey, cross-section
- Employment: job held last week (or no response)

Survey of Income and Program Participation (SIPP)

- ~ 14,000-36,700 households per wave, panel of 2.5-4 years
- Employment: jobs held in the past 4 months

Health and Retirement Study (HRS)

- ~30,000 respondents, age 50+, survey every 2 years (to death)
- Employment: current job if working, or last job held

Earnings record bridges

Longitudinal Employer Household Dynamics (LEHD)

- Quarterly earnings of jobs with employer UI reports (96% jobs)
- Data since 1990, includes states covering 90% jobs since 2001
- Includes state reported EIN of employer, or equivalent

W-2 Universe file

- Earnings information from W-2s only (no self-employment)
- Jobs where employer required to file W-2 reports with the IRS
- Includes EIN for each employer.

Detailed Earnings Record (DER)

- Extract from the SSA's Master Earnings File
- Includes earnings from W-2s and self-employment since 1978
- Includes EIN for each employer

Employer administrative records

Longitudinal Employer Household Dynamics (LEHD)

- Quarterly Census of Employment and Wages (QCEW), or ES-202
- Establishment employment, payroll, industry, location, ownership
- Contains multiple name fields: Legal, trade, worksite
- Earnings records (for most states) do NOT allocate workers to a specific establishment

Business Register (BR)

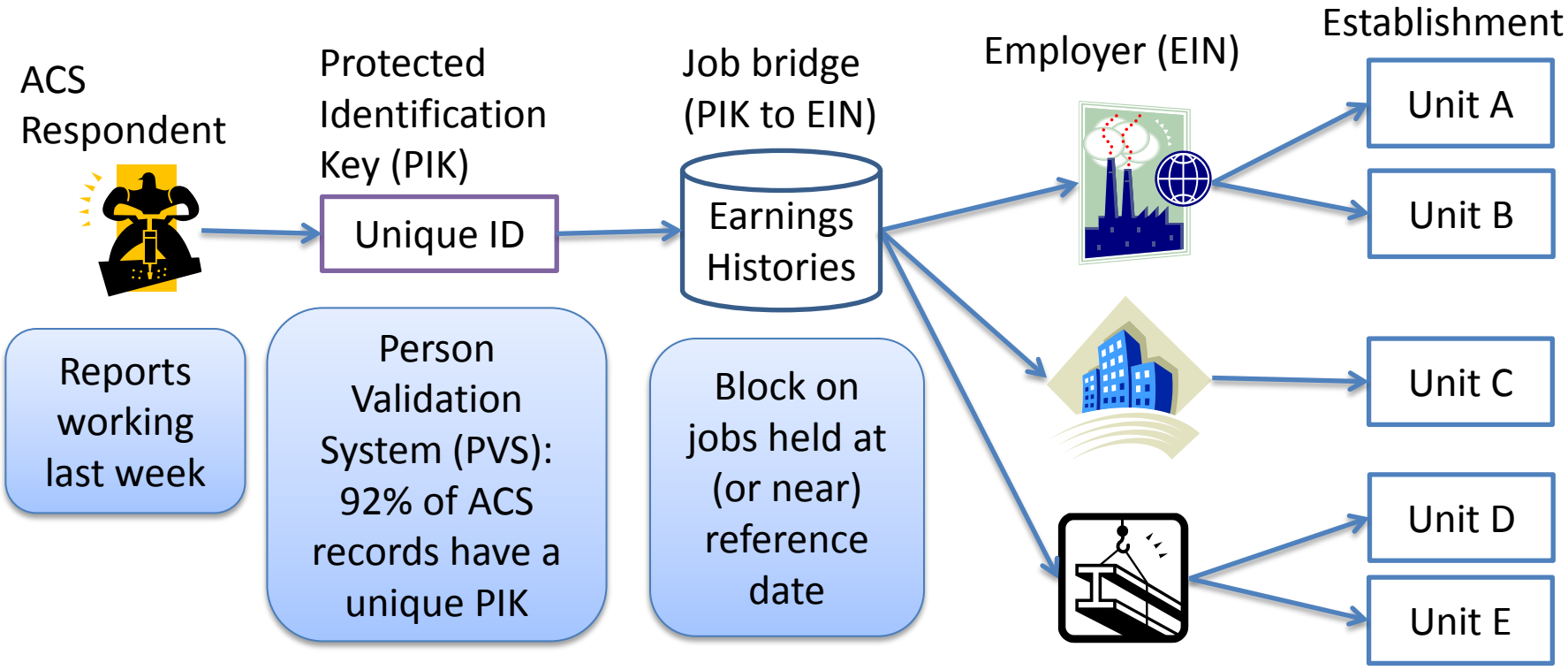
- Data since 1974, with ~7 million employer establishments
- Establishment employment, payroll, form of organization business location, organization type, industry
- Can be linked to other Census datasets containing more detailed business characteristics (Economic Censuses)
- Employer Identification Number (EIN) and Census firm and establishment identifiers

Record linkage procedures: overview

1. Pre-processing the two datasets to make sure their formats are consistent
   - Person and employer identifiers
2. For each job held by each respondent, narrow down their potential employer candidates using earnings history or EIN
   - (See following slide for example)
3. Retain a list of all candidate pairs of survey responses linked to administrative records (establishments)
   - For example, 3.4 million ACS respondents linked to 1 million LEHD employers and 3.7 million establishments result in 74 million pairs (for 2010)

Record linkage procedures: overview

Blocking Strategy: Example ACS/LEHD



Record linkage procedures: overview

- For each pair of a self-reported job and a potential candidate:
  - calculate agreement scores for each input field (e.g., name, address) based on a string/proximity comparator
  - Total scores of the pair is the sum of scores for each input fields weighted by their discriminating power.

- Fellegi & Sunter (1969) method - weights are derived from m and u probabilities
  - prob(field *k* agree| a pair is a true match) : “m probability”
  - prob(field *k* agree| a pair is unmatched) : “u probability”

Record linkage procedures: overview

- The pairs are classified into 3 regions based on matched scores (FS score) :

  match if FS score > upper-threshold

  non-match if FS score < lower-threshold

  uncertain if lower threshold < FS score < upper-threshold

  (clerical review)

- Unknown parameters: m and u probabilities for upper/lower thresholds
- The process typically involves multiple runs (passes), from more stringent to less stringent blocking requirements
- Classifications and FS score can be used in subsequent analyses. For example, analyses could restrict to the positive matches, or assign weights to records based on FS scores.

SWELL innovations

1. Develop or employ standardizer/parser for business names and addresses
2. Identify appropriate comparators for agreement of name and address fields
3. Calculate M and U probabilities, the upper/lower cutoffs based on clerical review of training set, using custom tool
4. Assemble SWELL toolkit for completing these steps and implementing FS

1. Standardizer/parser for business names and addresses

- This presentation focuses on a new standardizer for employer names

- For address standardizing
  - ACS/LEHD project is using Geocoded Address List (GAL) process based on a commercial software
  - SIPP : did not collect addresses in the past (plan for 2014 wave)
  - HRS : either use a customized tool or GAL (if available)

Pre-processing employer names

- Properly prepared data can lead to much higher quality matches

- The linking step relies on an approximate string comparator
  - can deal with small typos
  - cannot tell which words are not meaningful (e.g., THE, INC, LTD)
  - does not know acronym (e.g., CENTER = CTR)

- We are not aware of any "good" software available

  e.g., one not-so-good software changes "U S A" to "U South A"

Household survey database

Resp id	Employer name **
1	7-11
3	AT & T
4	KROGER
5	WAL-MART STORES, INC
6	EXTENDED STAY HOTEL
7	WLAMART
8	WALMART

Firm database

Firm id	Firm name
101	7-ELEVEN, INC
102	AT&T INC.
103	THE KROGER CO
104	WAL-MART STORES, INC.
105	DISH NETWORK CORPORATION
106	HVM L.L.C. D/B/A EXTENDED STAY HOTELS
107	PG INDUSTRIES ATTN JOHN SMITH
108	BB & T FKA COASTAL FEDERAL BANK

****ALL company names and addresses in this presentation are COMPLETELY artificial. No information from any survey or any administrative records was used in creating this document.**

stnd_compname: command to parse & standardizes company names

```
. stnd_compname varname, gen(newvar1, newvar2, newvar3,
                                    newvar4, newvar5)
```

Input : varname = name of a string variable containing company names

Output : newvar1 = official name
newvar2 = doing-business-as (DBA) name
newvar3 = formerly-known-as (FKA) name
newvar4 = entity type
newvar5 = attention name (normally a person name)

each component is standardized.

Optional inputs: patpath(directory of pattern files)
theme(public, pass-specific, or project specific)

Available in STATA and SAS*

*Ann Rodgers (U of Michigan) also contributes to the SAS program.

Example

```
. stnd_compname firm_name, gen(stn_name, stn_dbaname, stn_fkaname,
                               entitytype, attn_name)
```

firm name
7-ELEVEN, INC
AT&T INC.
DISH NETWORK CORPORATION
HVM L.L.C. D/B/A EXTENDED STAY HOTELS
THE KROGER CO
WAL-MART STORES, INC.
PG INDUSTRIES ATTN JOHN SMITH
BB & T FKA COASTAL FEDERAL BANK

stn_name	stn_dbaname	stn_fkaname	entitytype	attn_name
7 11			INC	
AT & T			INC	
DISH NETWORK			CORP	
HVM	EXTENDED STAY HOTELS		LLC	
KROGER			CO	
WAL MART STORES			INC	
PG IND				JOHN SMITH
BB & T		COASTAL FEDERAL BANK		

Customizing and updating pattern files

- **`stnd_compname`** is a wrapper of several subcommands.
- Each subcommand calls its associated CSV pattern file(s).

`stnd_compname's subcommands`	Description
parsing_namefield	parses company name without standardization
stnd_specialchar	standardizes special characters
stnd_entitytype	standardizes business entity types
stnd_commonwrd_name	standardizes words commonly appearing in company names
stnd_commonwrd_all	standardizes words commonly appearing in company names and addresses
stnd_NESW	standardizes directional words
stnd_numbers	standardizes numerals and their number equivalent
stnd_smallwords	standardizes small words e.g., conjunctions
parsing_entitytype	parses entity type from company name
agg_acronym	remove spaces between two or more one-letter words

Customizing and updating pattern files

- STATA & SAS programs call the same pattern files.
  These files are likely to be updated over time.

- Users may customize their own pattern files, but should be careful
  e.g., the sequence matters, expanding a word (E→ EAST) is risky.

Examples of pattern files (csv)

Key words used to parse (split) alternative names

1	DBA	DBA
2	D/B/A	DBA
3	D.B.A.	DBA
4	D B A	DBA
5	T/A	DBA
6	FKA	FKA
7	F/K/A	FKA
8	F.K.A.	FKA
9	F K A	FKA
10	FNA	FKA
11	F/N/A	FKA
12	F.N.A.	FKA
13	F N A	FKA
14	FORMERLY KNOWN AS	FKA
15	FORMERLY	FKA
16	AS SUCCESSOR TO	FKA
17	SUCCESSOR TO	FKA

Patterns to standardize some common words

6	CENTER	CTR
7	CENT ER	CTR
8	CNTRS	CTR
9	CNTR	CTR
10	FORTS	FT
11	FORT	FT
12	HEIGHTS	HTS
13	HEIGHT	HTS
14	HGTS	HTS
15	HGHTS	HTS
16	INT L	INTL
17	INTERNATIONAL	INTL
18	INTERNATL	INTL
19	I NTERNATL	INTL
20	INDUSTRIES	IND
21	INDS	IND
22	INDUSTRIAL	IND
23	INDL	IND
24	MNT	MT
25	MOUNT	MT
26	MOUNTAIN	MTN
27	MOUNTAINS	MTN

2. Name and address comparators

- Name
  - String distance: Damerau-Levenshtein, Jaccard, Q-grams, Monge-Elkan, SAS Data Quality
  - Jaro-Winkler string comparator
    - Employed in BigMatch for person names
  - Other string comparators appropriate for business names (suggestions welcome)
- Name components
  - One challenge is re-ordered names, partially missing names, entity types, and abbreviations
  - The standardizer/parser handles some of these, but flexible comparators may be necessary

Address comparators

- Rooftop match
- Distance (proximity)
  - Linear or non-linear
- Jurisdiction
  - Same Census Tract, ZIP code, City, County etc.
- Adjust for quality of geocoding?
  - Some addresses are only known to a ZIP code or county

3. Clerical review tool and training dataset

- Decisions required:
  - What info to use when scoring matches? Can reviewers use external knowledge?
  - What common rules to use for scoring as match, potential match, non-match? In what reasonable cases can reviewers disagree?
  - What match scores to capture (Characteristics/Establishment/Firm)?
  - How to select a review sample?

Review plan

- Goal is to review at least 1000 candidate pairs using ACS/LEHD data
- Each pair reviewed by two persons (may disagree)
- Reviewers evaluate the:
  - Overall establishment match
  - EIN level entity match
- Results used for calculating M and U
  - Fellegi-Sunter M and U estimation may use an empirical Bayes process to sample from reviews
- Same tool may be used for post processing evaluation or verification

Developing Training Set

- Pre-select sample of record pairs with wide range of agreement using arbitrary match rules

Sample distribution		Address score			
		Missing	Non-Match: Beyond Tract	Uncertain: Same Tract	Match: Rooftop
Name score		0	1	2	3
Missing	0	0%	0%	0%	0%
Non-Match: SASDQ<50	1	0%	33.3%	33.3%	33.3%
Uncertain: 50≤SASDQ<90	2	0%	33.3%	33.3%	33.3%
Match: 90≤SASDQ	3	0%	33.3%	33.3%	33.3%

Python Review Tool Layout
Example not from confidential files

Please score the match for these two establishments.

	ACS	LEHD
Name	Big Daddy's Restaurants	Asian Solutions
Address	1887 Gateway Road	106 Charter Street
	Portland, OR 97205	Fort Worth, KS 76102

Displays review pair with write-in response and candidate record

---

OTHER ESTABLISHMENTS

LEHD establishment 1 of 50	LEHD establishment 2 of 50
Mode O'Day	Quality Event Planner
1297 Brannon Avenue	2211 Hampton Meadows
Jacksonville, FL 32202	Ipswich, MA 01938

Set of additional candidate records for comparison

Please score the OVERALL ESTABLISHMENT match of the pair in the top section of the screen. Enter 'n' to view the next page of OTHER ESTABLISHMENTS.

SCORE DESCRIPTION
0 Missing
1 Inconsistent
2 Mostly consistent
3 Match

Reviewer responds: 0, 1, 2, or 3

4. SWELL toolkit

- Developing and testing SWELL tools on ACS/LEHD data
- Process is modular, and adaptable for project needs
- Components:
  - Standardizer/parsers
  - Clerical review tool
  - Fellegi-Sunter processing code including comparators
  - Documentation
- Once refined, tools will be portable to other projects
- M and U thresholds from ACS/LEHD clerical review may also be used as defaults (but may not be applicable if dataset is very different from ACS/LEHD)

Progress and Current Work

- Have working versions of basic components:
  - Standardizing/parsing code (SAS and Stata)
  - Probabilistic linking/workflow codes(SAS)
  - Clerical review tool (Python)

- Doing clerical review of a sample of pairs to develop a "truth set" for training Fellegi-Sunter thresholds

Potential extensions

- Social matching
  - Use networked name and address responses to supplement employer names or addresses
    - Colloquial names
    - Worksite locations
    - Public entities not reporting all establishments
- Reviewer variation in evaluation of training set
  - Reviewer fixed effects
  - Sampling from reviews to represent uncertainty

Thank you

- Contact:

Nada Wasi
nwasi@umich.edu

Mark Kutzbach
mark.j.kutzbach@census.gov

(we can put you in touch with any of the SWELL team)