Governments routinely collect data on the employment, wages, revenues, industry and location of nearly all businesses, often at both the establishment and firm levels.

– Data include taxpayer IDs and other identifiers

Governments use these data for tax compliance purposes, as inputs to NIPA tables, as sampling frames for business surveys, and as raw material in the creation of widely used economic statistics.
Large Business Databases

• With (much!) work, these raw data can be turned into core longitudinal research databases covering the universe of business firms and establishments
  – Establishments linked to parent firms
  – Both followed longitudinally (harder for firms)
  – In some cases, establishments also linked to their employees who are followed longitudinally.
Large Business Databases

• Data from other sources – one-off and recurring surveys, other administrative records, and proprietary data – are often merged with these core databases to investigate a specific issue.

• Advances in data storage and computing power greatly increase the scope for building these databases and using them as research tools.

• The next slide lists two core longitudinal U.S. business databases that emerge from different administrative record systems.
Two Large U.S. Databases

1. LBD: Longitudinal Business Database (Census)
   - Firms & Establishments, linked and tracked over time
   - Universal coverage, annual observations since 1976
   - Core variables: employment, payroll, sales/revenue (since 1994), industry, location, company name, IDs

2. BED: Business Employment Dynamics (BLS)
   - Firms (intra-state) & Establishments, linked and tracked over time
   - Universal coverage, quarterly since 1990
   - Core variables: employment, industry, location, IDs

The LBD and BED cover businesses with 1+ employees
Other Business Databases

• Other longitudinal business databases derive from panel surveys conducted by statistical agencies and from proprietary sources created for commercial purposes.

• Some of the most exciting research ideas on the margin combine core business micro datasets like the LBD and BED with smaller, targeted sources that contain rich information about particular aspects of business behavior and outcomes.
Selected U.S. Business Databases

3. LRD: A large, rotating panel of manufacturing plants with annual coverage back to 1972 and coverage in business census years before then
   – Large, rich set of measures
   – Draws on Annual Survey of Manufactures and once-every-five years Census of Manufactures
   – Heavily exploited in economics research

4. Manufacturing Energy Consumption Survey
   – Highly detailed data on energy-related inputs, expenditures, technologies and practices for roughly 15,000 manufacturing plants
   – Every 3 or 4 years since 1985
Selected U.S. Business Databases

5. Annual Surveys of Business for Retail, Wholesale and Manufacturing.

6. Various Censuses of Business
   - Extensive coverage of establishments and firms about once every five years, large set of measures
   - Coverage of Construction, FIRE, Retail, Wholesale, Manufacturing, Mining, Services and more. Only the Manufacturing data have been heavily worked.
   - First available year ranges from 1963 (Manufacturing) to 1987 or later, depending on sector.
Selected U.S. Business Databases

7. ILBD: LBD + revenue, industry and ownership for businesses with no employees. (Davis et al, 2009).

8. Job Openings and Labor Turnover Survey (JOLTS): BLS establishment-level survey of job vacancies and worker flows (quits, layoffs, hires, etc.)

9. CapitalIQ: Commercial platform, extensive data on financial characteristics of businesses and their sale (changes in ownership, capital structure, etc.)
   www.capitaliq.com/Main3/ourproducts_platform.asp

10. Dealogic: Another commercial platform with data on financial characteristics of businesses, business sales, etc. www.dealogic.com/
• Datasets 1 and 3-7 are products of the U.S. Census Bureau. They share common business identifiers, making it relatively easy to merge the establishment-level and firm-level data across databases. To get a fuller sense of the range of business micro data sets that reside within the U.S. Census system, peruse [www.census.gov/ces/dataprod...economicdata.html](http://www.census.gov/ces/dataprod...economicdata.html).

• Likewise, datasets 2 and 8 are products of the U.S. Bureau of Labor Statistics, and it is reasonably straightforward to merge them.
• When business datasets do not share common identifiers, they can be merged on the basis of business name and address, industry, size and other characteristics. This type of merging process can involve a great deal of work, and it typically yields match rates across datasets well below 100 percent.

• Fortunately, you can often build on previous efforts by others. For example, Davis et al. (2007) build on McCue and Jarmin (2005) to match Compustat data to firm-level records in the LBD through 2005.
Selected Longitudinal Employer-Worker Datasets

11. Longitudinal Employer Household Dynamics
   - LEHD micro data: quarterly earnings for workers, linked to employers and both followed longitudinally.
   - Data start in 1990 for selected states, with more states available in more recent years
   - See [www.census.gov/ces/dataproducts/lehddata.html](http://www.census.gov/ces/dataproducts/lehddata.html)

12. Social Security Administration (SSA) longitudinal employer worker data
   - Annual longitudinal data on the earnings of private sector employees from 1974 onwards, linked to employers who are also followed longitudinally
Selected Longitudinal Employer-Worker Datasets

13. German Social Security Data – Similar to LEHD and U.S. SSA data, but better in two respects: First, they allow for a relatively clean distinction between earnings and wages. Second, they allow for a much sharper dating of job-loss events and other employment transitions. Strong recent applications include:

- Schmieder et al. (2015) in an analysis of how unemployment insurance benefit extensions affect nonemployment durations and re-employment wages.
What Can You Do with These Datasets?

A. New measurement that informs our thinking about fundamental aspects of economic behavior and outcomes. Examples:

- Davis and Haltiwanger (2014) document a secular fall in the “fluidity” of U.S. labor markets and develop evidence on the implications for employment.
- Song et al. (2015) exploit longitudinal worker-employer linked data from the SSA to cast new light on the evolution of U.S. earnings inequality.
Empirical Regularities Uncovered by the Flow Approach to Labor Markets

- Gross job flows across employers are remarkably large – in good economic times and bad, in every market economy, in virtually every industry and sector
- Worker flows are larger yet
- Between-sector employment shifts account for a small share of job flows – idiosyncratic factors predominate
- Job and worker flows exhibit pronounced cyclical movements.
- Worker and job flows are tightly linked in the cross section and over time. (Davis, Faberman and Haltiwanger, 2012a)
B. Uncovering empirical regularities that challenge standard theory and guide new research.

Example: Davis, Faberman and Haltiwanger (2013) use JOLTS micro data to study vacancies, hires and vacancy yields in a large sample of U.S. employers.

- They show that the C-S behavior of vacancy yields is inconsistent with standard MP search theories.
- They develop compelling evidence that employers use other instruments, in addition to vacancy numbers, to vary their pace of hiring.
- Their evidence leads them to formulate a generalized matching function that accommodates a role for recruiting intensity (per vacancy) and outperforms the standard matching function, $H=M(U,V)$, in many ways.
Vacancy Yields and Establishment Growth Rates in the Cross Section

Does this strong positive relationship merely reflect a bigger flow of unobserved vacancies at more rapidly growing establishments?
A Model of Daily Hiring Dynamics

Daily laws of motion for flow of hires and vacancy stock:

\[ h_{s,t} = f_t v_{s-1,t} \]

\[ v_{s,t} = \left[ (1 - f_t)(1 - \delta_t) \right] v_{s-1,t} + \theta_t \]

- Where \( s \) indexes days, \( f_t \) is the daily job-filling rate in month \( t \), \( \delta_t \) is the rate at which unfilled vacancies lapse, and \( \theta_t \) is the daily flow of new vacancies.
Solving for the job-filling rate and vacancy flows

Use laws of motion to derive two equations relating end-of-month vacancy stock and hires flow during month, both observed, to two unknowns, \( \{f_t, \theta_t\} \).

\[ v_t = (1 - f_t - \delta_t + \delta_t f_t)^\tau v_{t-1} + \theta_t \sum_{s=1}^{\tau} (1 - f_t - \delta_t + \delta_t f_t)^{s-1} \]

\[ H_t = f_t v_{t-1} \sum_{s=1}^{\tau} (1 - f_t - \delta_t + \delta_t f_t)^{s-1} + f_t \theta_t \sum_{s=1}^{\tau} (\tau - s)(1 - f_t - \delta_t + \delta_t f_t)^{s-1} \]

Given data on \( \delta_t, v_t, v_{t-1}, H_t \), and a value for \( \tau \), solve numerically for \( f_t \) (daily job-filling rate) and \( \theta_t \) (daily flow of new vacancies).
Vacancy Flows and Job-Filling Rate Relationships to Employer Growth Rates

- Monthly Vacancy Flow Rate (Left Axis)
- Daily Job-Filling Rate (Right Axis)
- Daily Job-Filling Rate, Controlling for Establishment Fixed Effects
Vacancy Flows and Job-Filling Rate Relationships to Employer Growth Rates

Should be flat according to Mortensen-Pissarides models!
Is It Just “Lucky” Employers Growing Faster?

Stochastic nature of job filling induces a positive relationship between realized employment growth and job-filling rates at the establishment level.

• “Lucky” employers fill jobs faster and, as a result, grow faster.
• To quantify this effect, we simulate hires and employment growth at the establishment level for fitted values of $f$, $\theta$, $\delta$, and the distribution of vacancies, allowing parameters and vacancy distributions to vary freely by employer size class.

Result: Luck effect is much too small to explain the observed C-S relationship between job-filling rate and growth rate:

– Luck alone $\rightarrow$ job-filling rate rises by 2 percentage points in moving from 0% to 10% monthly growth rate.
– It rises by another 1 point in moving from 10 to 30%.
Partial Summary

• JOLTS micro data on hires and vacancies
• A simple model of daily hiring dynamics to identify the job-filling rate for vacancies
• Big CS variation in job-filling rates. Why?
  – Heterogeneity in the efficiency of search and matching
  – Scale economies (or diseconomies) in the hiring technology at the establishment or sectoral level
  – Employers use other instruments, in addition to vacancy numbers, to influence the pace of hiring.
Generalized Matching Function

• Generalized matching function (GMF) defined over unemployment, vacancies, and recruiting intensity per vacancy.
  – Combine micro-based evidence with GMF to identify the role of recruiting intensity per vacancy in the cross section
  – Build from micro evidence and GMF to construct a time-series index of recruiting intensity per vacancy
  – Interpret recent breakdown of standard MF as resulting, in part, from large movements in recruiting intensity per vacancy.
Fill Rate and Gross Hires Rate by Growth Rate Bin

Data points correspond to growth rate bins

Hires-Weighted Least Squares
Slope (s.e.) = 0.820 (0.006)
R-squared = 0.993

Note: The figure plots the relationship of the log daily job-filling rate to the log gross hires rate across growth rate bins in [-.3, .3] and the hires-weighted least squares regression fit of the bin-level data. Bin-level fill rates estimated from establishment-level data sorted into bins after removing mean establishment growth rates.
The GMF outperforms standard MF (SMF):

1. GMF accounts for CS behavior of job-filling rates. SMF does not.

2. GMF (as constrained by our recruiting intensity index) better accounts for movements over time in job-finding rates and job-filling rates.

3. GMF yields a more stable Beveridge Curve at national and regional levels.

4. Industry-level changes in fill rates, $v-u$ ratios, and recruiting intensity values during and after the 2008-09 recession satisfy restriction implied by the GMF. They violate restrictions implied by the SMF.
For a fuller development of these ideas, see my lecture slides on “Vacancies and Hiring, Establishment-Level Evidence and Aggregate Implications” at http://faculty.chicagobooth.edu/steven.davis/teaching.html, or read Davis, Faberman and Haltiwanger (2012b, 2013).
C. Identifying and correcting sample design flaws to produce better aggregate statistics.

Subtle flaws in sample design can lead to major errors in measurement and inference.

- **Remark**: A sample design suitable for estimating levels can be unsuitable for estimating changes or flows.

- **Example**: The sample design for the Job Openings and Labor Turnover Survey (JOLTS) under weighted tail mass in the cross-sectional growth rate density. As a result, published JOLTS statistics understated worker flows and job openings, and misstated the relative cyclicality of hires vs. separations and quits vs. layoffs.

- **Solution**: Reweight the sample to ensure that it replicates the employment growth rate density in the BED.
Rates of Hires, Quits, Layoffs and Job Openings in the Cross Section of Establishment Growth Rates, JOLTS Micro Data, 2001-2006

Cross-Sectional Densities of Establishment Growth Rates (Employment Weighted)

• Our adjusted statistics for hires and separations exceed the published statistics by about one third.

• Our adjusted layoff rate is more than 60 percent greater than the published layoff rate.

• Our adjustments significantly alter time-series properties as well.
  – Aggregate hires are 50 percent more variable than separations in published JOLTS statistics, as measured by the variance of quarterly rates, but 20 percent less variable according to our adjusted statistics.
  – Quarterly quit rates are more than twice as variable as layoffs in published statistics but equally variable according to our adjusted statistics.
• BLS revised its published JOLTS statistics after we circulated our paper. Their new statistics are much closer to our “adjusted estimates”.

• Do other important economic statistics suffer from similar measurement problems that distort estimated levels and time-series properties?
  – Investment expenditures, business borrowing?

• Perhaps, but it’s unclear (to me).
  – The errors introduced by flawed and ill-suited sample designs can be corrected ex post by adjusting survey-based estimates to match benchmark quantities.
  – However, suitable benchmarks are not always available or, when available, not always used well.
• For a very short and easy-to-digest discussion of the weaknesses in the JOLTS sample design, why they led to serious problems in the published JOLTS statistics, and how to correct for the sample design problems, I encourage you to read Davis (2010).

• While the basic idea behind the correction is quite simple, various challenges complicate the actual implementation. For the gory details, see Davis, Faberman, Haltiwanger and Rucker (2010).
D. Combining quasi-experimental variation and longitudinal business data to test theoretical predictions and estimate response magnitudes. There are many examples of this sort, and they differ widely in character. To be successful, this approach requires some ingenuity in recognizing a useful quasi experiment and in gathering or constructing data to merge into existing longitudinal business databases. If estimation of causal effects is the goal, then success also turns partly on whether the identification strategy is persuasive and yields enough power to recover reasonably precise estimates.
Examples

• Estimating the effects of private equity buyouts on employment, job reallocation, productivity and compensation per worker.
  – Davis et al. (2014) advance the literature on the effects of PE buyouts in several respects, but perhaps their most important innovation involves the simultaneous use of firm and establishment data to estimate within-firm reallocation effects.

• Identifying and quantifying spatial spillovers on the productivity of manufacturing plants.
  – Greenstone et al. (2010) compare outcomes at incumbent plants in counties that won contests for the opening of major new plants to outcomes in counties that lost.
• Greenstone, List and Syverson (2012) estimate the effects of air quality regulations on TFP at U.S. manufacturing plants. Their quasi-experimental variation: air quality regulations are more stringent in some counties (and for some types of plants) than others.

• How large were the employment effects of credit market disruptions in the wake of the Lehman bankruptcy?
  – Chodorow-Reich (2014) investigates this question using differences in lender health after the Lehman crisis as a source of variation in the availability of credit to borrowers.
• How do shocks to investment opportunities at one plant affect investment outcomes at other plants in the same firm? Does the answer turn on whether the firm is financially constrained?

  – Giroud and Mueller (2015) address these questions. They use the LRD to measure productivity, ownership, plant location, etc.
  – They use new airline routes that reduce travel time between the firm’s HQ and certain of its plants to obtain exogenous variation in plant-level investment opportunities. “[A] reduction in travel time makes it easier for HQ to monitor a plant, give advice, share knowledge, etc., raising the plant’s marginal productivity thus making investment in the (treated) plant more appealing.”
  – Clever recognition of an interesting, non-obvious quasi-experiment in the data.
• How sensitive is state-level employment to state-level taxes on business income? How much of the employment loss in one state (in response to a hike in its tax rate on business income) is offset by employment gains in other states?

• Giroud and Rauh (2015) address these questions, using the LBD and focusing on firms that operate in multiple states.

• They marshal and exploit much knowledge of business taxation and the nature of variation in business tax rates across states, time and form of business organizations. They use this knowledge to isolate several sources of state-specific time variation in the effective tax rate on corporate income (for businesses organized as “C” corporations) and in the personal tax rate (for unincorporated enterprises and businesses organized as pass-through entities).
E. Applications of employer-worker linked longitudinal datasets.

Perhaps the leading applications of these datasets thus far involve the study of earnings losses associated with job displacement. Jacobson, Lalonde and Sullivan (1993) provide an early and influential study that exploited employer-worker datasets and applied empirical techniques previously developed in research on program evaluation. More recent studies in the same mold include Couch and Placzek (2010), Von Wachter, Song and Manchester (2009), and Davis and von Wachter (2011).
Davis and von Wachter also show that workhorse search and matching models fail to explain the size and cyclicality of the present value earnings losses associated with job loss. Observed losses are several times larger than predicted by standard theory. Recent papers by Jung and Kuhn (2012), Kroliekowski (2015), Huckfeldt (2015), and Jarosch (2015) consider various modifications to search and matching models to address this issue. See my lecture slides “On the Consequences of Job Loss” at http://faculty.chicagobooth.edu/steven.davis/teaching.html for a fuller presentation.