Our dataset build started with collecting information from the Edge.org on all of the conversations and annual questions. We built a program that downloaded the information from the website, including the year, title, link to, and type of the conversation, as well as the text itself and who said it. Two independent coders then coded gender of the contributors based on their profile picture on Edge.org, or, if that was not available, pictures and pronouns on other reputable websites. We then manually collected information on the job title, workplace, and PhD by finding CVs, university webpages, news articles, personal websites, and Linked-In profiles. We wrote a program to collect the US News and World Report International Rankings and the Shanghai Rankings and manually gathered the rankings from the National US News and World Report Rankings. We then ran the text through the LIWC program. Finally, we calculated the rest of the variables (such as male contributors, previous contributions, etc.) based on the data we already had collected.

The descriptions below include the name in the full version of the dataset and the shortened name used in the dataset for older software.

- **Conversation level**:
  - **Year**: the year when it took place
  - **Title:** the title of the conversation. For example: "What Scientific Idea Is Ready For Retirement?"
  - **Link:** a link to the conversation
  - **Type:** 1 for annual question, 2 for conversation
    - Edge does an **annual question** every year; some examples are "what scientific idea is ready for retirement?" and "What will change everything?" People then write in with their answers. So all of the text is written and asynchronous
    - What Edge refers to as a **conversation** can actually be multiple things. Some of these are written essays by a single person, some are transcripts of a speech, and some are transcripts of a conversation (either between two or more guests or an interview).
  - **ThreadID (ThrdID):** a unique identifier for each conversation/annual question (between two or more people)
  - **MaleContributions (Mcontr):** the **number of times** a man speaks in a specific conversation, it **does not** always equal the number of unique men in a conversation (see below)
  - **FemaleContributions (Fcontr):** the **number of times** a woman speaks in a specific conversation, it **does not** always equal the number of unique women in a conversation (see below)
  - **FemaleParticipation (Fpart):** simply femalecontributors/(number of total contributions); the percentage of comments that are made by a woman
  - **NumberAuthors (NumAut):**
    - For the annual questions, this equals 0; because the website is the author of the question, everyone is considered commentators
    - Otherwise, this is the totle number of times people contribute to the main body of the text, rather than people who just comment. For example, in http://edge.org/conversation/how-democracy-works-or-why-perfect-elections-should-all-end-in-ties, there are multiple people commenting on

the post, but W. Daniel Hillis is the only author and only speaks once (as it is an essay). So NumberAuthors is "1." If two people each spoke five times in a dialogue, NumberAuthors would be "10."

- **DebateSize (DebSiz):** number of text pieces in a conversation; it is the sum of female and male contributions
- **Live:** whether the text piece was transcribed or written; it is 0 if it is written (either an essay or a comment on a piece) and 1 if it was part of a live conversation or speech that was later transcribed**.** Here are the types of text and how they would be classified:
  - **A single author essay** (live = 0 because it is written): http://edge.org/conversation/the-evolved-self-management-system
  - **A single author speech** (live = 1 because it was spoken and later transcribed): http://edge.org/conversation/cities-as-gardens
  - **A live conversation**, either between multiple people or in an interview format (live = 1 because it was spoken and later transcribed): http://edge.org/conversation/japan-inc-meets-the-digerati
  - **Online Comments** on any of the three types above (live = 0 because it was written)
  - **The annual question (Type = 1):** live = 0 because these were all written and submitted.
- **UniqueContributors (UContr)**: **UniqueMaleContributors** + **UniqueFemaleContributors**
- **UniqueMaleContributors (UMContr)**: the number of unique male contributors
- **UniqueFemaleContributors (UFContr):** the number of unique female contributors
- **UniqueFemaleParticipation (UFPar)**: the percentage of unique female participants; **UniqueFemaleContributors** divided by **UniqueContributors**

- **Participant Level**
  - **Id:** the unique identifier of the contributor
  - **Id_num**: the unique identifier of the contributor as text (this is typically the format of first name_last name)
  - **Role**: Either author (=1) or commentator (=2)
  - **Name**: name of the commentator
  - **TwoAuthors (TwoAutrs):** some of the edge comments are written by two people. In this case, we duplicated the row and kept the text level and conversation level information the same and had one author per row. This variable is 1 if this text was written by two people and 0 otherwise.
  - **Female**: the commentator is male = 0, the commentator is female = 1
  - **Male**: the commentator is female = 0; the commentator is male = 1
  - **Academic (Acad):** 1 = the person is in academia, 0 = they are not
  - **Limited_Information (LimInfo):** equals 1 if we could only find limited information about the person (e.g. they commented in 2013 but we only have their job title from 2012), 0 otherwise
  - **Job_Title (JobT)**: The job title of the commentator

- **Job_Title_S (JobTS):** This is a simplified list of job titles (e.g. we have "Eugene Higgs Professor" in Job.Title but "Chaired Professor" in Job.Title.Collapsed)
    - Chaired Professor
    - Professor
    - Associate Professor
    - Assistant Professor
    - Non-Tenure-Track Faculty
    - Postdoctoral Researcher
    - Graduate Student
    - Academic Leadership (Dean, Vice President, etc.)
    - Researcher
    - Artist/Author/Editor/Writer
    - Director
    - Founder
    - Other
    - Top Management and Founder
    - Top Management
    - Entrepreneur
    - Not Available
- **Job_Title_S_num (JobTSn):** Job_Title_S as numbers instead of text
- **Department (Dept):** what academic department someone is in
- **Department_S (DeptS):** a simplified version of all the departments (e.g. while John Smith's Department is "Experimental Physics," his Department_S is "Physics")
    - Physics (Phy)
    - Anthropology (Ant)
    - Earth Sciences (ES)
    - Biology (Bio)
    - Psychology (Psych)
    - Journalism, media studies and communication (JMS)
    - Medicine (Med)
    - Philosophy (Phil)
    - Space Sciences (SS)
    - Linguistics (Lin)
    - Computer Sciences (CS)
    - Engineering (Eng)
    - Arts (Arts)
    - Business/Management (Bus)
    - Environmental Studies and Forestry (ESF)
    - Sociology (Soc)
    - Mathematics (Math)
    - Asian Studies (AS)
    - Education (Educ)
    - Political Science (PS)
    - Economics (Econ)
    - Systems Science (Sys)

        - History (Hist)
        - Music (Musc)
        - Chemistry (Chem)
        - Archeology (Arch)
        - Architecture and Design (ArchD)
        - Law (Law)
        - Zoology (Zoo)
        - Literature (Lit)
        - Divinity (Div)
    - **Department_S_num (DeptSn):** Department_S as numbers instead of text
    - **Discipline (Disc)**: this groups academic departments into disciplines
        - Natural Sciences (NS)
        - Social Sciences (SocS)
        - Professions (Prof)
        - Humanities (Hum)
        - Formal Sciences (FS)
    - **Workplace (Workpl):** where someone works; some people are self-employed
    - **HavePhD (PhD):** equals 1 if they have a phd, 0 otherwise. It is 1 even if someone earns a phd after they comment (e.g. John Doe comments in 2000 and earns his PhD in 2012; his comment in 2000 will still have HavePhD = 1)
    - **PhD_Field (PhDF):** what field people got their PhD in
    - **PhD_Year (PhDY):** what year they got their PhD
    - **PreviousContributions (PrContr)**: how many times **before this year** they have made contributions. So if John Doe only talked three times in one conversation in 2012 and one time each in two conversations in 2014 (and never made any other comments), this will be 0 for his comment in 2012 and 3 for both his comments in 2014.
    - **ContributionsThisYear (ContrTY):** how many times they contributed this year; even if they only participated in one conversation, if they spoke 40 times in that conversation, this variable will be 40.
    - **ThreadsThisYear (ThrTY):** how many threads they participated in this year; thus if John spoke in two threads in 2014, one twenty times and one once, this would equal 2 in 2014, while **ContributionsThisYear** would equal 21 for 2014.
    - **PreviousThreads (PreThrd):** how many threads they participated in **before this year**. So, if John contributed for the first time twice in one thread in 2000, once each in two different threads in 2004, and once in 2014, this would be 0 for 2000, 1 for 2004, and 3 for 2014 (and for **PreviousContributions** it would be 0 for 2000, 2 for 2004, and 4 for 2014).
    - **AuthorandCommentator (AutAndCom):**: if, for the same piece, someone is both an author and a commentator, this is 1 for that person for that piece; otherwise it is 0
    - **PhD_Institution (PhDI)**: what school they got their PhD
    - **Years_from_PhD (YfPhD):** how many years at the time of the comment since they earned their PhD; this is just Year - PhD.Year. This can be negative because people may have earned their phd years after they make a comment

- **PhD_Institution_SR (PhDISr):** The Shanghai Rankings of their PhD Institution; this is only for people who received their PhDs from institutions that are ranked by Shanghai. Shanghai ranks only between 500 and 510 universities worldwide each year and also bins their rankings after a certain point, in different ways for different years (e.g. a university may be ranked as 301-352).
- **PhD_Institution_SR_Bin (PhDISrB)**:
  - 1 = university was ranked between 1 and 50
  - 2 = university was ranked between 51 and 100
  - 3 = university was ranked between 101 and 150
  - 4 = university was ranked between 151 and 200
  - 5 = university was ranked between 201 and 300
  - 6 = university was ranked between 301 and 400
  - 7= university was ranked between 401 and 510
- **Workplace_SR (WorkSr):** The Shanghai Rankings of their workplace; this is only for academics and academic institutions that are ranked by Shanghai (see **PhD_Institution_SR** for more information)
- **Workplace_SR_Bin (WorkSrB):**
  - 1 = university was ranked between 1 and 50
  - 2 = university was ranked between 51 and 100
  - 3 = university was ranked between 101 and 150
  - 4 = university was ranked between 151 and 200
  - 5 = university was ranked between 201 and 300
  - 6 = university was ranked between 301 and 400
  - 7= university was ranked between 401 and 510
- **SR_Ranking_Dif (SrRDif):** The difference between the binned Shanghai Ranking University of their workplace and the binned Shanghai Ranking of their PhD; a positive ranking means that they work at a place that has a higher ranking than where they got their PhD
- **PhD_Institution_US_IR (PhDIR):** The US News and World Report created an international ranking system in 2014 to rank the top 500 universities. Thus, even if a comment was made in 1999, if they have a PhD from Carnegie Mellon, this ranking will be Carnegie Mellon's ranking in the 2014 report
- **PhD_Institution_US_IR_Bin (PhDIRB):**
  - 1 = university was ranked between 1 and 50
  - 2 = university was ranked between 51 and 100
  - 3 = university was ranked between 101 and 150
  - 4 = university was ranked between 151 and 200
  - 5 = university was ranked between 201 and 250
  - 6 = university was ranked between 251 and 300
  - 7= university was ranked between 301 and 350
  - 8 = university was ranked between 351 and 400
  - 9 = university was ranked between 401 and 450
  - 10 = university was ranked between 451 and 500
- **Workplace_US_IR (WorkIR):** See **PhD_Institution_US_IR**
- **Workplace_US_IR_Bin (WorkIRB):**
  - 1 = university was ranked between 1 and 50

        - 2 = university was ranked between 51 and 100
        - 3 = university was ranked between 101 and 150
        - 4 = university was ranked between 151 and 200
        - 5 = university was ranked between 201 and 250
        - 6 = university was ranked between 251 and 300
        - 7= university was ranked between 301 and 350
        - 8 = university was ranked between 351 and 400
        - 9 = university was ranked between 401 and 450
        - 10 = university was ranked between 451 and 500
    - **USA_I_Ranking_Dif (IRDif):** the difference between the rank of someone's workplace and the rank of their PhD Institution (as ranked by US News and World Report International Rankings). If this is positive, it means they're working at an institution ranked higher than their PhD Institution.
    - **PhD_Institution_US (PhDIUS):** The ranking of their PhD Institution by USA News and World Report; this is **only** for US institutions and only for a limited number of them. Different numbers of school were ranked in different years; for example, 129 schools were ranked in 2005, while only 51 were ranked in 2003. These only go from 2003-2014.
    - **PhD_Institution_US_Bin (PhDIUSB):**
        - 1 = university was ranked between 1-5
        - 2 = university was ranked between 6-10
        - 3 = university was ranked between 11-25
        - 4 = university was ranked between 26-50
        - 5 = university was ranked between 51-100
        - 6 = university was ranked between 101-150
        - 7 = university was ranked between 151-200
    - **Workplace_US (WorkUS):** The ranking of their workplace by USA News and World Report; this is **only** for US institutions and only for a limited number of them. Different numbers of school were ranked in different years; for example, 129 schools were ranked in 2005, while only 51 were ranked in 2003. These only go from 2003-2014.
    - **Workplace_US_Bin (WorkUSB):**
        - 1 = university was ranked between 1-5
        - 2 = university was ranked between 6-10
        - 3 = university was ranked between 11-25
        - 4 = university was ranked between 26-50
        - 5 = university was ranked between 51-100
        - 6 = university was ranked between 101-150
        - 7 = university was ranked between 151-200
    - **USA_Ranking_Dif (USRDif):** the difference between the rank of someone's workplace and the rank of their PhD Institution (as ranked by US News and World Report Rankings). If this is positive, it means they're working at an institution ranked higher than their PhD Institution.
    - **Total_Citations (TotCit):** the total number of citations they have received, including that year and all previous years (it's citations.year + previouscitations)

- **H_Index (Hind):** this is their h-index in **2014;** a scholar has an index of *h* if they have published *h* papers each of which has been cited in other papers at least *h* times
- **i10_index (iTenIn):** how many papers in **2014** they had authored that has more than 10 citations; this is only for Google Scholar pages. As the GS pages only have an i10 index from 2014, even if the comment was from 1999, the i10 index is from 2014
- **Citations_Year (CitY):** how many citations they received this year; this is only for Google Scholar pages, so not all academics have this
- **Citations_Cumulative (CitCum):** how many citations they have received in this year and previous years; this is only for Google Scholar pages, so not all academics have this
- **AcademicHierarchyStrict (AcaHier):**
  - 1 = Graduate Student
  - 2 = Postdoctoral
  - 3 = Assistant Professor
  - 4 = Associate Professor
  - 5 = Professor
  - 6 = Chaired Professor
- **PreviousCitations (PreCit):** the number of citations they have received in all of the previous years
- **ContributionsbyAuthor (ContrAut):** the number of contributions by this author in this conversation
- **Dummy variables for Discipline**
- **Dummy variables for department_S**

- **Text-Level**
  - **Order:** The order of the text pieces. This is **meaningless** for Annual Questions.
  - **Text:** the text of the conversation
  - **Number_Characters:** number of characters in the text piece
  - **LIWC variables (see www.liwc.net/descriptiontable1.php)**