



Karen Spärck Jones

Viviana Patti, Dipartimento di Informatica, University of Turin



di.unito.it

Turin, March 29 2021 - Tavola rotonda: Donne di co-Scienza

viviana.patti@unito.it

Karen Spärck Jones

(Regno Unito, 1935-2007)

- ❖ Percorsi non lineari e figure ibride

- ❖ **Formazione umanistica:** laurea in storia e filosofia

- ❖ **Ambiente di ricerca:**

Centro di Ricerca sul linguaggio diretto da Margaret Masterman: filosofa e linguista, allieva di Wittgenstein, pioniera degli studi in **linguistica computazionale**, **traduzione automatica**

- ❖ Wittgenstein's philosophy on language use

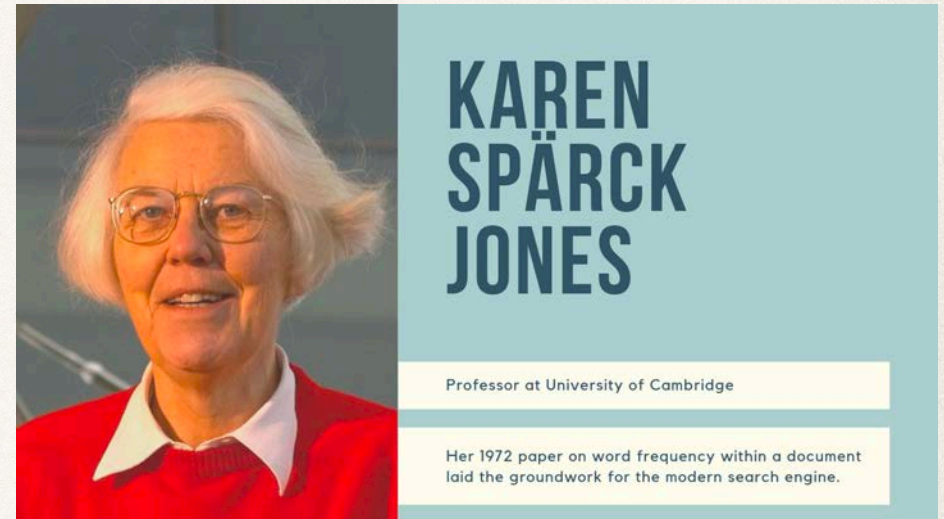
- ❖ **1964:** Tesi di dottorato: "Sinonimi e classificazione semantica": applicazione di tecniche statistiche all'analisi del linguaggio, intuisce l'**importanza delle risorse linguistiche**, un elemento fondamentale per la linguistica computazionale nell'epoca dei big data

- ❖ **1972:** contributo fondamentale: teoria statistica applicata al linguaggio concetto di **TF-IDF: term frequency- inverse document frequency**

- ❖ Funzione di peso che vent'anni dopo sarà alla base del **funzionamento dei primi motori di ricerca**

- ❖ **1999:** Full professor, Direttrice di Ricerca. Ricopre incarichi di prestigio e indirizzo in ACL. Premio a lei dedicato dalla comunità di Information Retrieval.

- ❖ Sposa Roger Needham ma mantiene il suo nome da nubile in ambito professionale:
"E' comunque una cosa buona da fare sempre, mantiene la tua identità, se sai cosa significa"



Karen Spärck Jones

Che cosa le dobbiamo

- ❖ **Linguistica Computazionale**

- ❖ summarization, natural language interfaces, semantica lessicale, sistemi di dialogo...

- ❖ **Traduttori automatici**

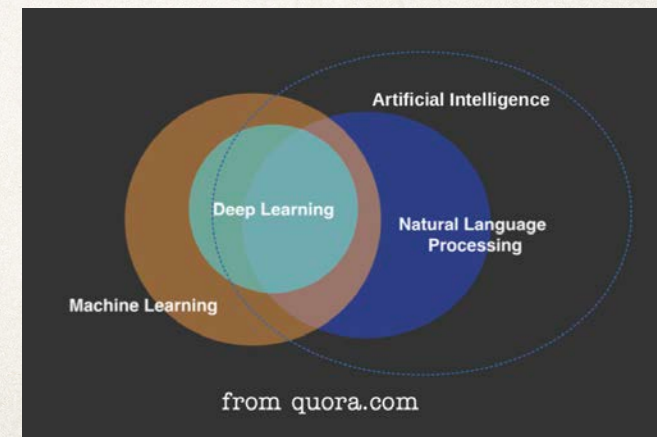
- ❖ **Motori di ricerca**



A composite image showing two search-related screenshots. The top screenshot is a Google search for "Karen Spärck Jones", displaying search filters, a grid of images, and search results including academic articles and a Wikipedia entry. The bottom screenshot is an AltaVista search interface from October 23, 1999, featuring a search bar, navigation links, and a "POWER SEARCH" section.

Che cosa le dobbiamo AI, NLP e CL

- ❖ Aree dell'Intelligenza Artificiale che si occupano specificamente del linguaggio umano si chiamano **Linguistica Computazionale (CL)** e **Trattamento Automatico del Linguaggio** (NLP Natural Language Processing)
- ❖ NLP include un'ampia varietà di compiti/task la cui soluzione si basa sulla comprensione del linguaggio umano.
- ❖ Esempi di task:
 - ❖ Machine translation
 - ❖ Part of Speech tagging (morphological analysis) and parsing (syntactic analysis)
 - ❖ Question answering
 - ❖ Sentiment analysis e opinion mining
 - ❖ Hate speech detection
 - ❖ **Information retrieval**
 - ❖ ...



Che cosa le dobbiamo

AI, NLP e CL

- ❖ Natural Language Processing (NLP) e Linguistica Computazionale (CL)
 - ❖ **Morfologia** → *how words are* **Sintassi** → *how words relate*
 - ❖ **Semantica** → *what words mean* **Pragmatica** → *what is the intention of the speaker* (contesto)
- ❖ Intuizione:
- ❖ Al pari di un essere umano, per poter comunicare in linguaggio umano un sistema di NLP deve
 - ❖ **Acquisire la conoscenza relativa al linguaggio**
 - ❖ **Rappresentare al suo interno tale conoscenza**
 - ❖ **Utilizzare** questa conoscenza in modo da **poter affrontare in modo automatico i vari task:**
- ❖ Ricevere domande e rispondere in modo pertinente, estrarre informazione per esempio sul sentiment o sull'odio, tradurre, cogliere l'intenzione ironica in un'espressione, **ritrovamento di documenti rilevanti per l'utente nel mare magnum dei documenti disponibili sul web**

Che cosa le dobbiamo AI, NLP e CL

- ❖ I sistemi più recenti **contengono poca conoscenza**, ma sono in grado di **apprenderla utilizzando algoritmi di apprendimento automatico supervisionato**, da esempi con uno sforzo limitato
- ❖ Si costruiscono quindi delle **grandi raccolte di esempi di uso del linguaggio**, in cui
 - ❖ **Regole, pattern ricorrenti e le irregolarità sono presenti** e possono essere utilizzate dai sistemi per fare astrazione e **generalizzare usando tecniche statistiche**
- ❖ **Deep Learning**
- ❖ **Profili multi-disciplinari fondamentali**



Karen Spärck Jones

Che cosa le dobbiamo

- Linguistica Computazionale

- Motori di ricerca

- 1995: AltaVista primo motore di ricerca

- Trovare risultati in poco tempo

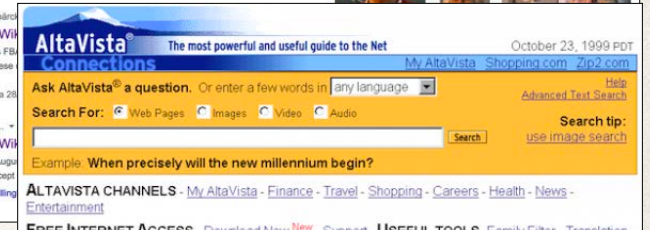
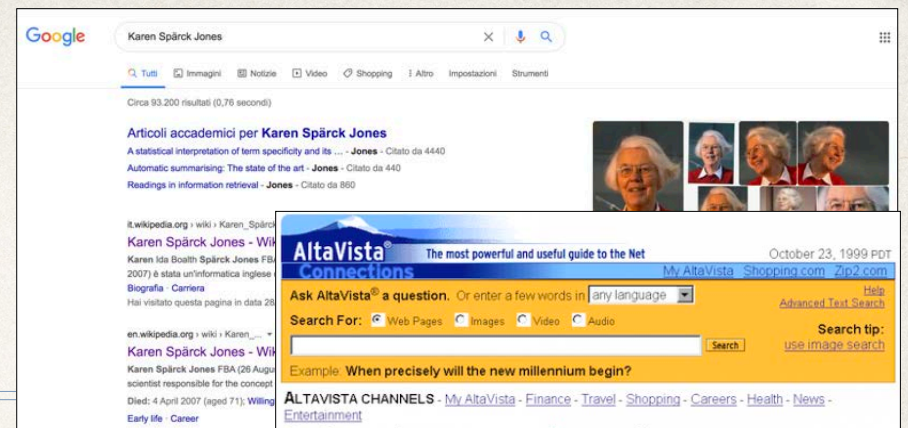
- 1972: contributo fondamentale: **teoria statistica applicata al linguaggio** concetto di TF-IDF: **term frequency- inverse document frequency**

John I. Tait, 2007

well as in computing and artificial intelligence more broadly. In IR, her greatest contribution was probably her invention (1972) of the concept of inverse document frequency (IDF), in which the importance of terms is weighted according to the proportion of documents in the corpus in which they occur; the intuition being that terms which occur in many documents are poor index terms. This is the partial basis of all weighting schemes adopted by widely used Internet search engines, and will undoubtedly be a lasting contribution to the field.

- Funzione di peso che vent'anni dopo sarà **alla base del funzionamento dei primi motori di ricerca**

- Sguardo pionieristico**: derivare informazione semantica su basi empiriche; importanza delle risorse linguistiche e di applicare la statistica all'analisi del linguaggio



108 CHAPTER 6 • VECTOR SEMANTICS AND EMBEDDINGS

Because of the large number of documents in many collections, this measure too is usually squashed by a log function. The resulting definition for inverse document frequency (idf) is thus

$$\text{idf} = \log_{10} \left(\frac{N}{df} \right) \quad (6.13)$$

Here are some idf values for some words in the Shakespeare corpus, ranging from extremely informative words which occur in only one play like *Romeo*, to those that occur in a few like *salad* or *Falstaff*, to those which are very common like *foot* or so common as to be completely non-discriminative since they occur in all 37 plays like *good* or *sweet*.³

Word	df	idf
Romeo	1	1.37
salad	2	1.27
Falstaff	4	0.967
forest	12	0.489
battle	21	0.246
we	34	0.037
foot	36	0.012
good	37	0
sweet	37	0

³idf The tf-idf weighted value $w_{i,j}$ for word i in document d thus combines term frequency $t_{i,j}$ (defined either by Eq. 6.11 or by Eq. 6.12) with idf from Eq. 6.13:

$$w_{i,j} = t_{i,j} \times \text{idf} \quad (6.14)$$

Fig. 6.9 applies tf-idf weighting to the Shakespeare term-document matrix in Fig. 6.2, using the tf equation Eq. 6.12. Note that the tf-idf values for the dimension corresponding to the word *good* have now all become 0, since this word appears in every document, the tf-idf algorithm leads it to be ignored. Similarly, the word *foot*, which appears in 36 out of the 37 plays, has a much lower weight.

	As You Like It	Twelfth Night	Julius Caesar	Henry V
battle	0.074	0	0.22	0.28
good	0	0	0	0
foot	0.019	0.021	0.0036	0.0083
we	0.009	0.044	0.018	0.022

FIGURE 6.9 A tf-idf weighted term-document matrix for four words in four Shakespeare plays, using the counts in Fig. 6.2. For example the 0.009 value for *we* in *As You Like It* is the product of $\text{tf} = \log_{10}(20 + 1) = 1.322$ and $\text{idf} = .037$. Note that the idf weighting has eliminated the importance of the ubiquitous word *good* and vastly reduced the impact of the almost-ubiquitous word *foot*.

The tf-idf weighting is the way for weighting co-occurrence matrices in information retrieval, but also plays a role in many other aspects of natural language processing. It's also a great baseline, the simple thing to try first. We'll look at other weightings like PPMI (Positive Pointwise Mutual Information) in Section 6.6.

³ *Sweet* was one of Shakespeare's favorite adjectives, a fact probably related to the increased use of sugar on European recipes around the turn of the 16th century (Lerch, 2014, p. 175).

Karen Spärck Jones

Che cosa le dobbiamo

- ❖ 1972: contributo fondamentale: teoria statistica applicata al linguaggio
- concetto di TF-IDF: term frequency- inverse document frequency

108 CHAPTER 6 • VECTOR SEMANTICS AND EMBEDDINGS

Because of the large number of documents in many collections, this measure too is usually squashed with a log function. The resulting definition for inverse document frequency (idf) is thus

$$\text{idf}_i = \log_{10} \left(\frac{N}{df_i} \right) \quad (6.13)$$

Here are some idf values for some words in the Shakespeare corpus, ranging from extremely informative words which occur in only one play like *Romeo*, to those that occur in a few like *salad* or *Falstaff*, to those which are very common like *fool* or so common as to be completely non-discriminative since they occur in all 37 plays like *good* or *sweet*.³

Word	df	idf
Romeo	1	1.57
salad	2	1.27
Falstaff	4	0.967
forest	12	0.489
battle	21	0.246
wit	34	0.037
fool	36	0.012
good	37	0
sweet	37	0

The tf-idf weighted value $w_{i,d}$ for word i in document d combines term frequency $tf_{i,d}$ (defined either by Eq. 6.11 or by Eq. 6.12) with idf from Eq. 6.13:

$$w_{i,d} = tf_{i,d} \times \text{idf}_i \quad (6.14)$$

Fig. 6.9 applies tf-idf weighting to the Shakespeare term-document matrix in Fig. 6.2, using the tf equation Eq. 6.12. Note that the tf-idf values for the dimension corresponding to the word *good* have now all become 0; since this word appears in every document, the tf-idf algorithm leads it to be ignored. Similarly, the word *fool*, which appears in 36 out of the 37 plays, has a much lower weight.

	As You Like It	Twelfth Night	Julius Caesar	Henry V
battle	0.054	0	0.22	0.26
good	0	0	0	0
fool	0.019	0.021	0.0036	0.0083
wit	0.009	0.044	0.016	0.022

Figure 6.9 A tf-idf weighted term-document matrix for four words in four Shakespeare plays, using the counts in Fig. 6.2. For example the 0.049 value for *wit* in *As You Like It* is the product of $tf = \log_{10}(20 + 1) = 1.322$ and $idf = 0.037$. Note that the idf weighting has eliminated the importance of the ubiquitous word *good* and vastly reduced the impact of the almost-obscure word *fool*.

The tf-idf weighting is the way for weighting co-occurrence matrices in information retrieval, but also plays a role in many other aspects of natural language processing. It's also a great baseline, the simple thing to try first. We'll look at other weightings like PPMI (Positive Pointwise Mutual Information) in Section 6.6.

³ Sweet was one of Shakespeare's favorite adjectives, a fact probably related to the increased use of sugar in European recipes around the turn of the 16th century (Charlady, 2014, p. 175).

A STATISTICAL INTERPRETATION OF TERM SPECIFICITY AND ITS APPLICATION IN RETRIEVAL

KAREN SPARCK JONES

University of Cambridge Computer Laboratory

The exhaustivity of document descriptions and the specificity of index terms are usually regarded as independent. It is suggested that specificity should be interpreted statistically, as a function of term use rather than of term meaning. The effects on retrieval of variations in term specificity are examined, experiments with three test collections showing in particular that frequently-occurring terms are required for good overall performance. It is argued that terms should be weighted according to collection frequency, so that matches on less frequent, more specific, terms are of greater value than matches on frequent terms. Results for the test collections show that considerable improvements in performance are obtained with this very simple procedure.

EXHAUSTIVITY AND SPECIFICITY

WE ARE FAMILIAR with the notions of exhaustivity and specificity: exhaustivity is a property of index descriptions, and specificity one of index terms. They are most clearly illustrated by a simple keyword or descriptor system. In this case the exhaustivity of a document description is the coverage of its various topics given by the terms assigned to it; and the specificity of an individual term is the level of detail at which a given concept is represented.

These features of a document retrieval system have been discussed by Cleverdon *et al.*⁴ and Lancaster,⁶ for example, and the effects of variation in either have been noted. For instance, if the exhaustivity of a document description is increased by the assignment of more terms, when the number of terms in the indexing vocabulary is constant, the chance of the document matching a request is increased. The idea of an optimum level of indexing exhaustivity for a given document collection then follows: the average number of descriptors per document should be adjusted so that, hopefully, the chances of requests matching relevant documents are maximized, while too many false drops are avoided. Exhaustivity obviously applies to requests too, and one function of a search strategy is to vary request exhaustivity. I shall be mainly concerned here, however, with document descriptions.

Specificity as characterized above is a semantic property of index terms: a term is more or less specific as its meaning is more or less detailed and precise. This is a natural view for anyone concerned with the construction of

More women in tech

- ❖ Nell'era digitale è imprescindibile: **more women working in tech**
- ❖ Immaginario collettivo: informatica è maschile
 - ❖ stereotipo del nerd, giovani maschi soli in una stanza chini su una tastiera
- ❖ **Impatto sociale**
 - ❖ Technology is pervasive and women must be active part of this digital revolution, also to be able to recognise, and counteract the misogynistic behaviours we are observing online
 - ❖ **Agenti conversazionali:** lo strano caso del chatbot Tay



More women in tech

❖ Rischio:

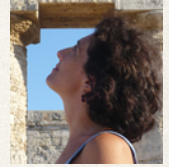
- ❖ Introdurre **bias** nei corpora e nei modelli addestrati sui dati
- ❖ Inclusive design & diversity
- ❖ “Credo anche che l’informatica sia intellettualmente affascinante, soprattutto se si ha intenzione di creare qualcosa che ancora non esiste”



Riferimenti

- ❖ **ACL Lifetime Achievement Award:**
Karen Spärck Jones (2005) *Some Points in a Time*
<https://www.aclweb.org/anthology/J05-1001.pdf>
- ❖ Karen Spärck Jones (1972). *A statistical interpretation of term specificity and its application in retrieval*. *Journal of Documentation*, 28:11–21
<https://doi.org/10.1108/00220410410560573>
- ❖ Karen Spärck Jones (March 2007) *Natural Language and the Information Layer: Professor Karen Spärck Jones' health did not permit her to receive the ACM Athena Award (at SIGIR 2007, 23–27 July 2007, Amsterdam) and BCS Lovelace Medal in person, so she prepared this video recorded talk instead*. Length 33 minutes, 16:9 format. Available here:
<https://www.cl.cam.ac.uk/misc/obituaries/sparck-jones/video/>
- ❖ John Tait (2007). *Obituary. Karen Spärck Jones*. *Comput. Linguistics* 33(3): 289-291
<https://www.aclweb.org/anthology/J07-3001.pdf>
- ❖ Daniel Jurafsky & James H. Martin (2020) *Speech and Language Processing*. Draft of December 30, 2020. Chapter 6.
<https://web.stanford.edu/~jurafsky/slp3/6.pdf>
- ❖ Cinzia Ballezio, Giovanna Giordano (2019) *L'informatica al femminile. Storie sconosciute di donne che hanno cambiato il mondo*
<https://informaticalfemminile.it/>

Who am I?



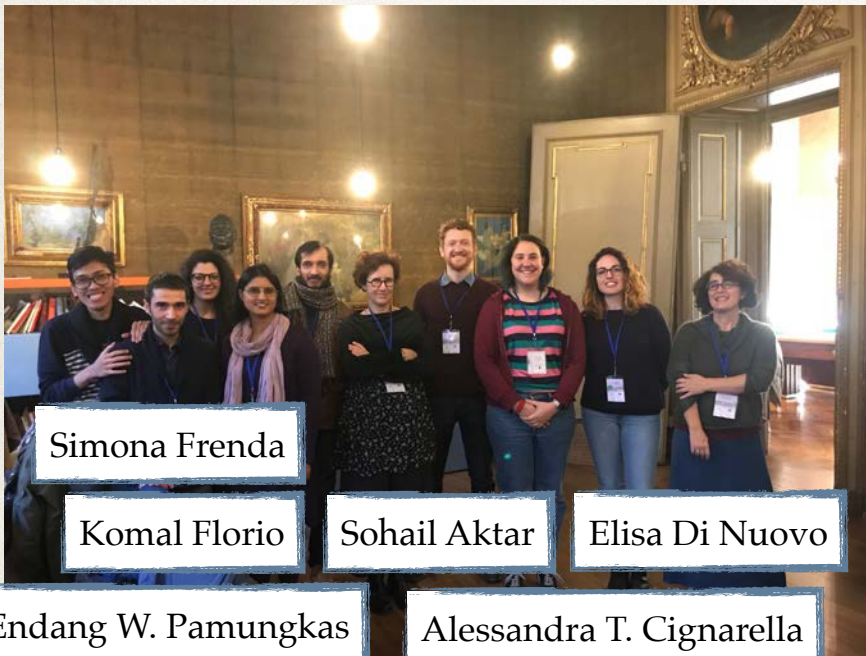
- ❖ Msc Computer Science in **Philosophy**
 - ❖ **Multidisciplinary curriculum (analysis of language and computer science):** philosophy of language, computational linguistics, cognitive science, programming, artificial intelligence.
- ❖ PhD in Computer Science
- ❖ Associate Professor at Computer Science Department, University of Turin
- ❖ co-founding member of the **Logic, Language and Cognition center**, University of Turin (Philosophy, Computer Science, Psychology)
- ❖ Past member of **Guarantee Committee** at UniTo
- ❖ <https://www.unito.it/persona/vpatti>



di.unito.it



The multi-disciplinary team



Simona Frenda

Komal Florio

Sohail Aktar

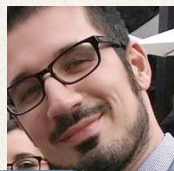
Elisa Di Nuovo

Endang W. Pamungkas

Alessandra T. Cignarella



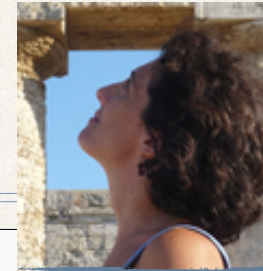
Fabio POLETTO



Marco STRANISCI



Valerio BASILE



Viviana PATTI



Cristina BOSCO



Content Centered Computing

ABOUT PEOPLE PROJECTS MISCELLANEOUS

About

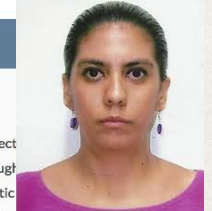
Group Content-Centered Computing topics are inter- and trans-discipline enterprises, where the language is the modality that defines the field of the item, and the medium is the storage and representation of the item. Group CCC is engaged in research projects concerning the following:

- languages: spoken natural language, dramatic writing, etc.
- domains: railway transportation, etc.
- media: text, audio, visual, etc.

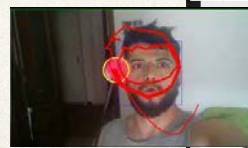
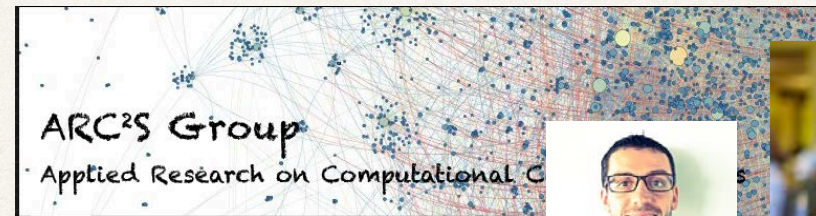
language, graphic, iconic and diagrammatic annotation language; in the social and human sciences, contemporary art production, one can explore the current



Manuela SANGUINETTI



Delia Irazu HERNANDEZ FARIAS



Arthur CAPOZZI



Mirko LAI



Rossano SCHIFANELLA



Giancarlo RUFFO

March 2013
January 2013

Thank you!



viviana.patti@unito.it
<https://www.unito.it/persona/vpatti>

CHIEN-SHIUNG WU

Shanghai 31 maggio 1912

New York 16 febbraio 1997

Simonetta Marcello, Dipartimento di Fisica
Donne di co-Scienza 29 marzo 2021
Public Engagement Lab
Università degli Studi di Torino



Eroe Coraggioso



1912 Chien-Shiung Wu nasce nel villaggio di Liuhe vicino a Shanghai

Il suo nome: Eroe Coraggioso

Il padre Wu Zhong-Yi era un Rivoluzionario e un Educatore

Studia alla Scuola elementare per bambine fondata dal padre e poi lontano da casa nella Scuola SuZhou per ragazze

La sua Bisnonna, la Matriarca della Famiglia, vuole che studi

1936 Si laurea all'Università Nazionale Centrale di Nanchino e parte alla volta degli Stati Uniti con la sua amica chimica Duo-Ron-Fen per fare un PhD all'Università del Michigan incoraggiata da Gu Jing-Wei, una sua professoressa di fisica



Al lavoro con i Nobel



1936 Arriva a San Francisco e decide di fermarsi a fare il PhD all'Università di Berkeley

Studia sotto la guida di Ernest Lawrence, che progetta e realizza il primo Ciclotrone, Emilio Segrè e Robert Oppenheimer

1940 Consegue il PhD con una tesi di Fisica Nucleare

Lavora intensamente in laboratorio fino a tardi la sera

Diventa esperta nella Fissione Nucleare

Le sue Radici



Mantiene un forte legame con il suo Paese e si fa mandare le stoffe dalla Cina per confezionare i suoi abiti col colletto alto, che indossa per tutta la vita, anche al lavoro sotto un camice bianco

A Berkeley diventa famosa per la sua bravura e la sua eleganza

Non rinuncia alla sua identità e mantiene il suo nome che nessuno riesce a pronunciare. E tutti la chiamano *Miss Wu*.

Con la sua amica Ursula Schaefer cerca sempre di andare a mangiare qualcosa di cinese

Non immagina che potrà tornare in Cina solo nel 1973

1942 Sposa Luke Yuan nella casa di Robert Millikan

CHIEN-SHIUNG WU



nuclear physicist

Crediti Massive Science

Chiedi a Miss Wu

Si trasferisce a Princeton dove lavora il marito e per un po' insegna alle ragazze dello Smith College e poi come Istruttrice a Princeton

Oppenheimer la chiama "l'Autorità nella Fissione Nucleare"

Quando Enrico Fermi, che nel 1939 ha iniziato il Progetto Manhattan, non riesce a risolvere un problema che blocca la fissione nucleare, Segrè suggerisce di "chiedere a Miss Wu" e lei lo risolve.

Ma non viene chiamata a lavorare al Progetto perché Donna e Asiatica

Solo nel 1944 viene chiamata a far parte del Progetto Manhattan

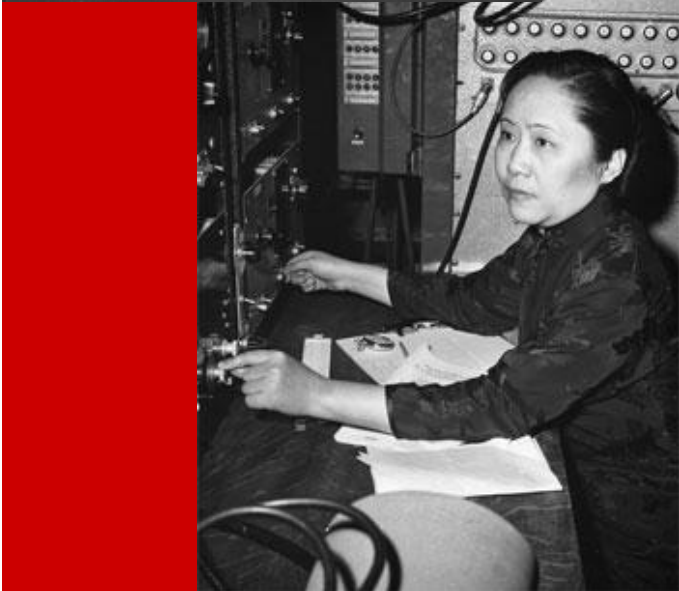
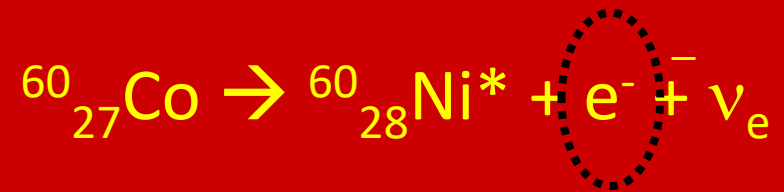
Dopo la guerra le viene offerto un lavoro alla Columbia University dove può fare ricerca e diventa esperta del decadimento beta

L'Esperimento di Wu

1956 I fisici teorici Tsung-Dao Lee e Chen-Ning Yang ipotizzano che la Simmetria di Parità possa essere violata nelle interazioni deboli e chiedono a Wu di fare un esperimento per dimostrarlo

È una sfida, molti fisici non vogliono cimentarsi perché non credono che la Simmetria di Parità possa essere violata

Wu mette insieme un team di esperti di criogenia per realizzare l'esperimento, misurando il decadimento beta del **Co-60** a una temperatura vicina allo Zero assoluto. Lavorano sei mesi senza sosta per preparare tutto



CHIEN-SHIUNG WU



parity violation — manhattan project — beta decay

CHIEN-SHIUNG WU

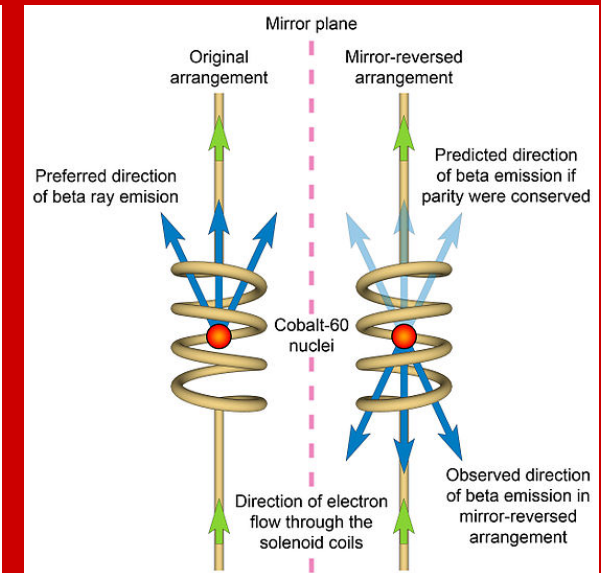
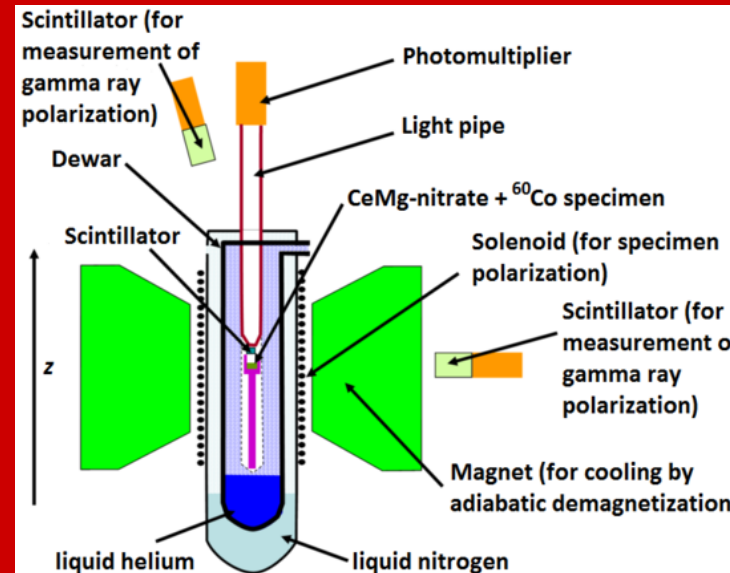
[Crediti marchforscience.org](http://Crediti.marchforscience.org)

L'Esperimento di Wu

1957 Wu dimostra che l'idea di Lee e Yang è corretta

La Simmetria di Parità è Violata nelle Interazioni deboli

1957 Lee e Yang vincono il Premio Nobel dal quale viene esclusa Chien-Schiung Wu



La Prima Donna



1943 Istruttrice all'Università di Princeton

1952 Professoressa associata alla Columbia University

1958 Professoressa ordinaria alla Columbia University

1958 Riceve il dottorato onorario all'Università di Princeton

1964 Vince il Premio Comstock della National Academy of Sciences

1975 Diventa Presidente dell'American Physical Society

1978 Unica Donna a ricevere il **Premio Wolf in Fisica** in Israele

1990 Prima Scienziata vivente ad avere un Asteroide col suo nome 2752

Chien-Shiung Wu

Dal 1958 al 1975 riceve DIECI dottorati onorari (Harvard, Yale, ...)

Il discorso al MIT

1964 Discorso a un Simposio del MIT sulla discriminazione di Genere

«Vorrei sapere se gli atomi e i nuclei, o i simboli matematici, le molecole del DNA preferiscano avere un trattamento maschile o femminile»

1990 Esprime pubblicamente la sua disapprovazione per il sessismo nell'Accademia statunitense dichiarando "Le donne in America sono veramente oppresse"



1997 Muore e le sue ceneri vengono portate in Cina e sepolte nel cortile della Scuola Ming fondata dal padre, che aveva frequentato da piccola

Quale eredità ha lasciato ?



L'11 febbraio 2021 Giornata Mondiale delle Donne nella Scienza
Negli USA è stato emesso un francobollo dedicato
a Chien-Shiung Wu per ricordarla

Chien-Shiung Wu è un esempio di quello che una donna può
realizzare quando le viene data l'opportunità di avere successo

-Non fermarci davanti agli ostacoli e alle discriminazioni

-Non dubitare delle nostre capacità come Donne

La Forza Nascosta Scienziante nella Fisica e nella Storia



Spettacolo teatrale

di Elena Ruzza e Gabriella Bordin
con Elena Ruzza e Fé Avouglan

Idea e cura di Anna Ceresole,
Nora De Marco,
Emiliana Losma,
Simonetta Marcello,
Nadia Pastrone e Rita Spada

<http://laforzanascosta.to.infn.it>

Al lavoro con i Nobel



Wu (second from right) in Berkeley with J. Robert Oppenheimer (third from right), her advisor, Emilio Segrè (fourth from right) and good friend Margaret Lewis (extreme left).