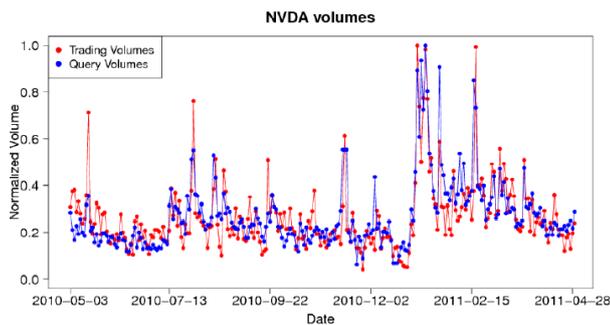


“A data scientist is someone who is better at statistics than any software engineer and better at software engineering than any statistician”

Prevedere il mercato

Il traffico nei motori di ricerca può essere usato per tracciare e, talvolta, anticipare le dinamiche di fenomeni sociali. Per esempio, il volume di scambi commerciali giornalieri è correlato con il volume di interrogazioni giornaliere effettuate relativamente a tali scambi. In particolare, in molti casi l'analisi delle interrogazioni consente di prevedere, con almeno un giorno di anticipo, i picchi degli scambi commerciali e, quindi, di identificare, attraverso l'analisi dell'attività degli utenti sul web, situazioni di allarme relativi al rischio finanziario sistemico.



Prevedere i risultati elettorali

Nel 2012 Nate Silver ha usato un modello predittivo, con il quale ha non solo previsto correttamente i risultati elettorali in 49 dei 50 stati degli Stati Uniti d'America, ma ha anche previsto in modo accurato quale candidato avrebbe vinto in tutti e 50 gli stati. Nel 2016 lo stesso modello è risultato essere quello che dava più possibilità a Donald Trump di vincere le elezioni (sia pure minori di quelle di Hilary Clinton).



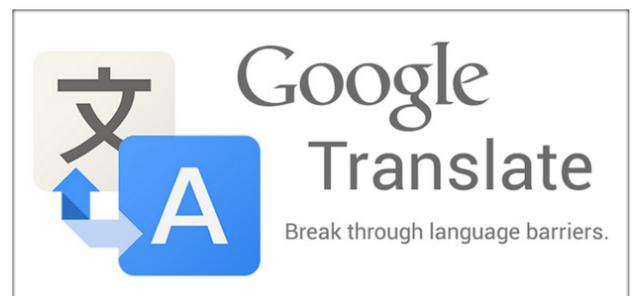
Sulla strada di HAL 9000

Nel 2016 un sistema basato sull'intelligenza artificiale, chiamato LipNet, è stato in grado di leggere il labiale in modo più efficace di professionisti di tale attività. Una rete neurale opportunamente addestrata ha imparato a identificare le variazioni nella forma della bocca e a collegare queste variazioni all'interpretazione di ciò che veniva detto. Il sistema è stato in grado di identificare in modo corretto il 93,4% delle parole, mentre gli esseri umani ne hanno identificate solamente il 52,3%.



Google Translate

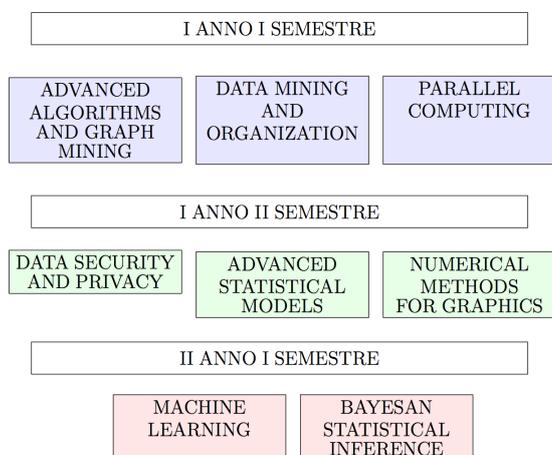
Il sistema di traduzione di Google utilizza un approccio chiamato di *traduzione automatica statistica*. Tale approccio si basa sulla convinzione che una solida base per sviluppare un sistema di traduzione automatica in grado di tradurre da una lingua ad un'altra, debba avere almeno un milione di corrispondenze di parole della prima lingua con l'altra, ed il significato di almeno un miliardo di parole per ognuna delle due lingue. I modelli statistici di questi dati vengono poi utilizzati per creare la traduzione.



Il corso di laurea magistrale in Informatica (curriculum Data Science) combina l'esperienza di quattro aree disciplinari dell'Università degli Studi di Firenze: informatica, ingegneria dell'informazione, matematica e statistica.

Insegnamenti fondamentali

- ❑ **Advanced Algorithmic Methods and Graph Mining:** per imparare tecniche algoritmiche per l'analisi dei dati, con particolare attenzione a insieme giganteschi di dati e agli aspetti sia teorici che pratici.
- ❑ **Advanced Statistical Models:** per conoscere i principali modelli statistici e imparare a selezionare quello più appropriato in base agli obiettivi di ricerca e ai dati disponibili.
- ❑ **Bayesian Statistical Inference:** per imparare le problematiche connesse alla inferenza Bayesiana, che è alla base dell'apprendimento automatico.
- ❑ **Data Mining and Organization:** per imparare gli algoritmi di data mining per la ricerca di regolarità e pattern nei dati e le strutture dati per la loro organizzazione.
- ❑ **Data Security and Privacy:** per imparare i concetti e i metodi fondamentali della crittografia classica e di quella moderna.
- ❑ **Machine Learning:** per imparare gli algoritmi di base e avanzati per l'apprendimento statistico, le basi della teoria dell'apprendimento computazionale e a progettare soluzioni a problemi reali.
- ❑ **Numerical Methods for Graphics:** per acquisire quelle competenze del computer aided geometric design utili all'implementazione e all'uso di algoritmi specifici per la visualizzazione di dati.
- ❑ **Parallel Computing:** per imparare le tecniche di programmazione parallela e ad alte prestazioni.

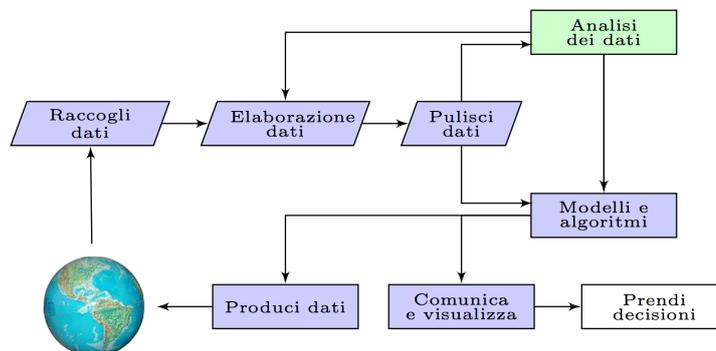


Il paradigma *big data*, con il suo impulso alla crescita di produttività e all'innovazione, è destinato a diventare uno dei fattori chiave della competizione a livello globale. La *data science* è un campo di studi interdisciplinare, che ha per oggetto i metodi scientifici, le procedure e i sistemi per estrarre conoscenza e comprensione da grandi quantità di dati, sia strutturati che non, nel rispetto della privacy degli individui. A tale scopo, la data science impiega teorie e metodi provenienti da vari campi dell'informatica, della statistica e della matematica.

Un *data scientist* mette a frutto le proprie capacità analitiche per:

- ❑ individuare fonti di dati e campionarle;
- ❑ organizzare e gestire in maniera efficiente grandi quantità di dati tenendo conto dei vincoli imposti da software, hardware e banda di comunicazione;
- ❑ costruire modelli matematici per analizzare regolarità e *pattern* nascosti nei dati, o anche apprendere da essi;
- ❑ garantire che raccolta, trasmissione e analisi dei dati siano condotte senza rischi per la privacy;
- ❑ creare visualizzazioni che aiutino la comprensione dei dati;
- ❑ presentare e comunicare la conoscenza ricavata dai dati.

Il processo della data science



Nella società dell'informazione, quella di data scientist sta emergendo naturalmente come una delle professioni più ricercate. Secondo uno studio spesso citato (McKinsey Global Institute, 2011), solo negli USA nel 2018 la domanda di data scientist potrebbe eccedere di 1,5 milioni la disponibilità effettiva di tali figure.

Il laureato magistrale in Informatica (curriculum Data Science) sarà in possesso delle competenze per rivolgersi direttamente ad aziende che, sul territorio o in ambito globale, operano nel campo dell'analisi di mercato e della *business intelligence*; ad istituzioni che per statuto mantengono e trattano grandi quantità di dati (medici, finanziari, di censo, e così via); a imprese piccole o grandi che per la gestione delle loro attività si affidano a sistemi informativi complessi.