Data science per lo studio di ecosistemi complessi

Edoardo Pasolli

Dipartimento di Agraria

Università degli Studi di Napoli Federico II

Caffè scientifico di Agraria – Portici, Sala Cinese – 26 giugno 2019

edoardo.pasolli@unina.it

@epasolli

DIPARTIMENTO DI AGRARIA

Agenda

- Biografia in breve
- Introduzione alla *data science*
- Le mie linee di ricerca

edoardo.pasolli@unina.it

@epasolli

DIPARTIMENTO DI AGRARIA

Biografia in breve

- **Laurea in Ingegegneria delle Telecomunicazioni** @ Univ. di Trento, 2008

- **Dottorato in *Information and Communication Technology*** @ Univ. di Trento, 2011
  Tesi: *"Active learning methods for classification and regression problems"*

- **Postdoc** @ NASA Goddard Space Flight Center, 2012-13

- **Postdoc** @ Purdue University, 2013-14

edoardo.pasolli@unina.it

@epasolli

Biografia in breve

- **Assegnista di ricerca** & **Marie Skłodowska-Curie Individual Fellow** @ Univ. di Trento, 2014-18

- **RTD-B** @ Agraria, Da Dicembre 2018
  S.S.D.: ING-INF/03 (Telecomunicazioni)

Introduzione alla *data science*

- Una possibile definizione: *"Data science is an **interdisciplinary field** that uses **scientific** methods, processes, algorithms and systems to **extract knowledge** and insights **from data** in various forms“*

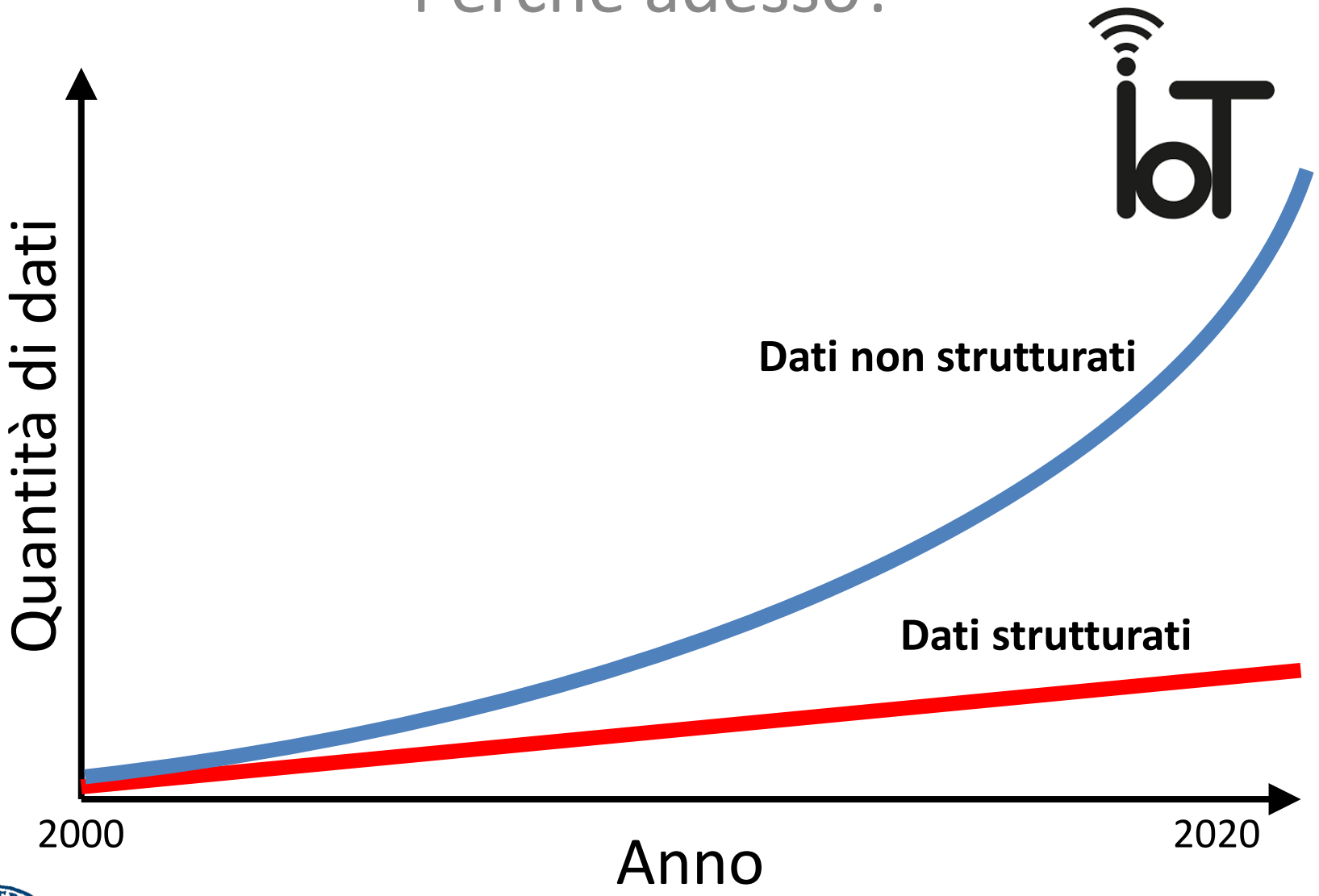
- In breve: *“The processes used to extract insights from data”*

edoardo.pasolli@unina.it

@epasolli

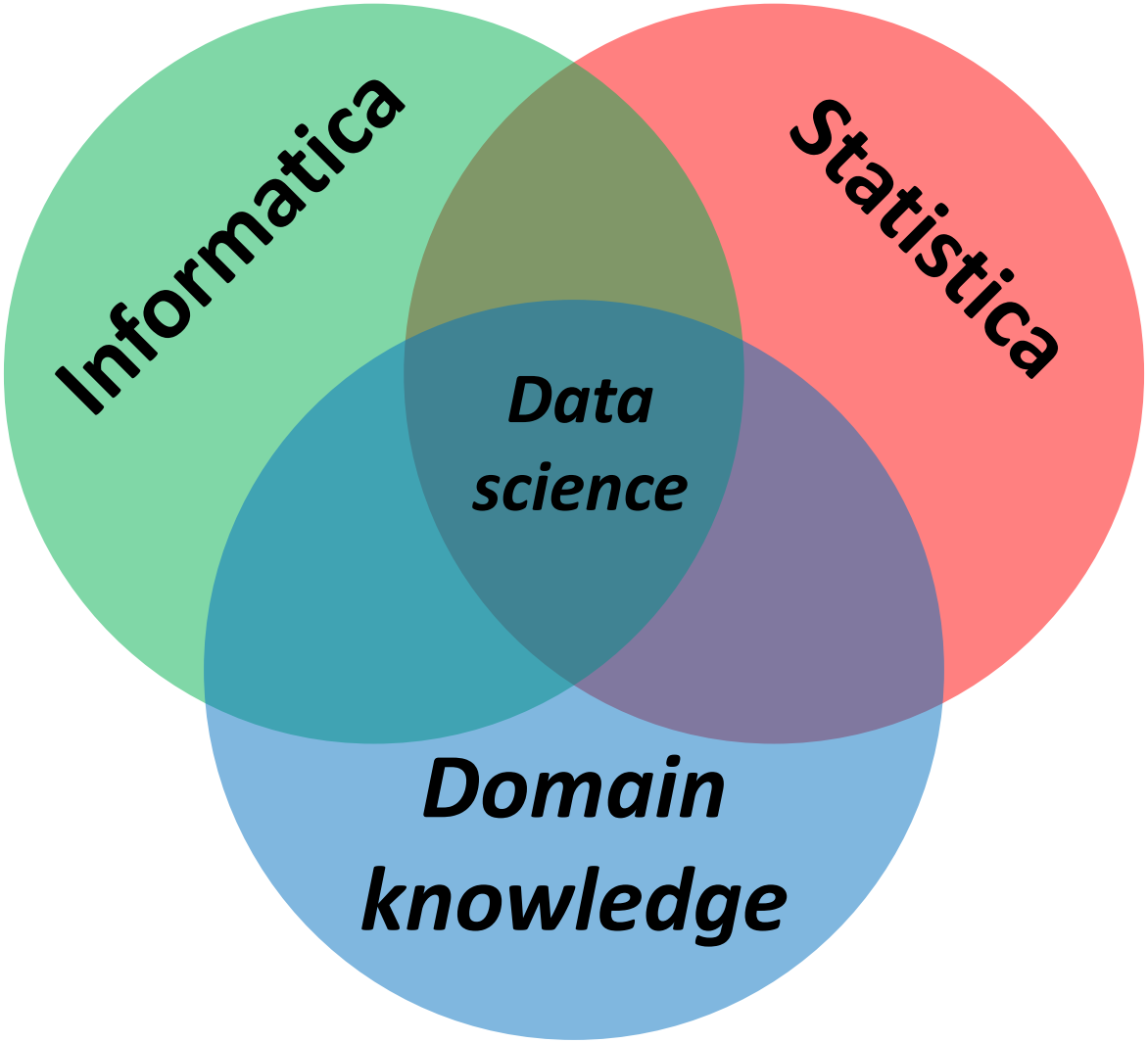
DIPARTIMENTO DI AGRARIA

Perchè adesso?



edoardo.pasolli@unina.it
@epasolli

Un ambito fortemente multidisciplinare



edoardo.pasolli@unina.it
@epasolli

DIPARTIMENTO DI AGRARIA

Il *data science lifecycle*

edoardo.pasolli@unina.it
@epasolli

DIPARTIMENTO DI AGRARIA

Garbage in = Garbage out

Dati **Analisi** **Risultato**

Dati **Analisi** **Risultato**

edoardo.pasolli@unina.it

@epasolli

DIPARTIMENTO DI AGRARIA

Le mie linee di ricerca

- Sviluppo di **algoritmi di *machine learning*** per **dati telerilevati**
  - Metodologie di *active learning*
  - Fusione di dati multi-sorgente

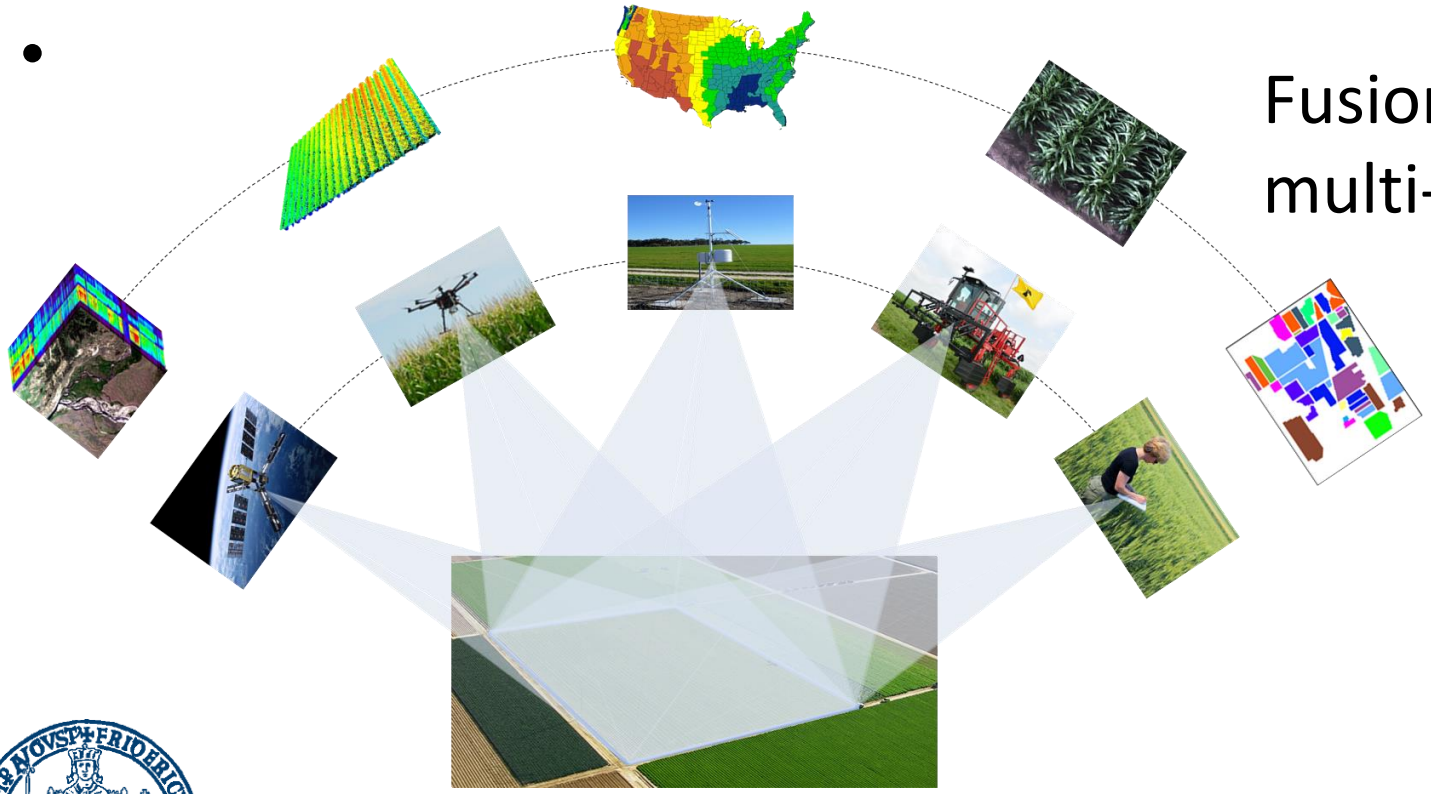
- Sviluppo ed applicazione di ***tool* computazionali** per il **microbioma** da **dati di metagenomica**
  - *Tools* per l'analisi a livello di ceppo
  - Microbioma umano... e da alimenti

edoardo.pasolli@unina.it

@epasolli

DIPARTIMENTO DI AGRARIA

Introduzione al telerilevamento

- Definizione: *"Remote sensing is the acquisition of information about an object or phenomenon without making physical contact with it"*
- 

Fusione di dati multi-sorgente

edoardo.pasolli@unina.it

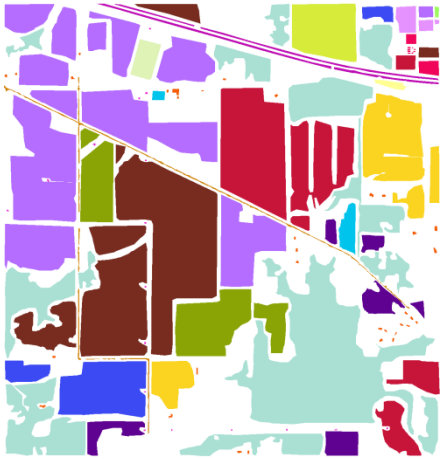
@epasolli

DIPARTIMENTO DI AGRARIA

Classificazione & Mappe di copertura del suolo

Input:
Dato telerilevato

Classificazione



Output:
Mappa di copertura del suolo
(*Semantic segmentation*)

- Lavorato principalmente con sensori ottici + fusione con dati LiDAR

Approccio di classificazione supervisionata

Input

Pre-processamento

Estrazione delle *features*

Classificazione

Output

Acquisizione *training set*

Operatore umano

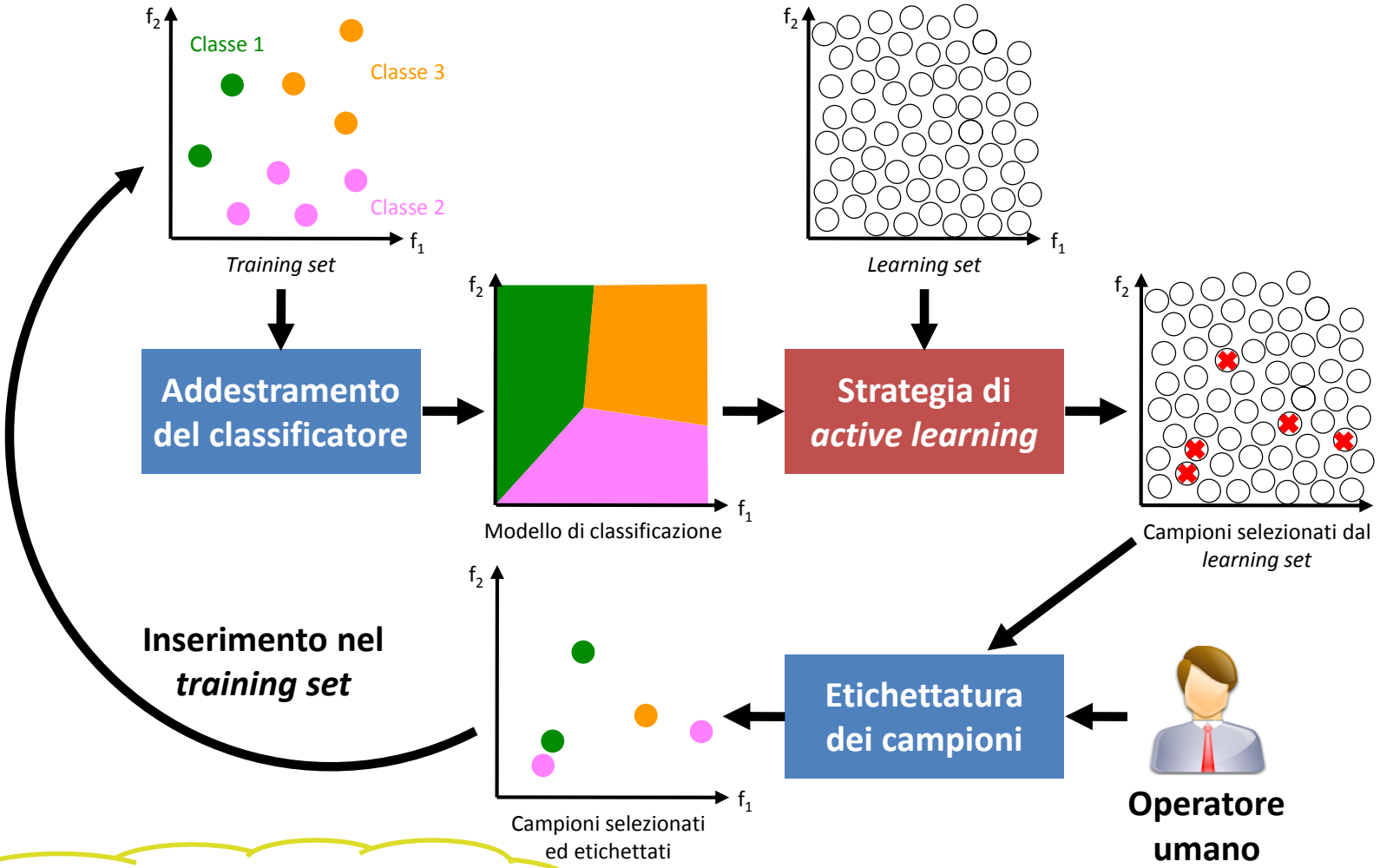
Il *training set* influenza l'accuratezza del modello

edoardo.pasolli@unina.it

@epasolli

DIPARTIMENTO DI AGRARIA

Active learning & Acquisizione *training set*



Obiettivo: Selezionare il *training set* migliore

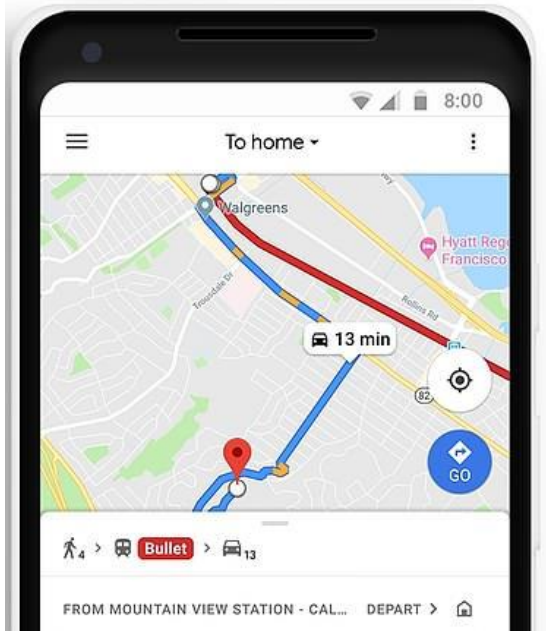
Human-in-the-loop

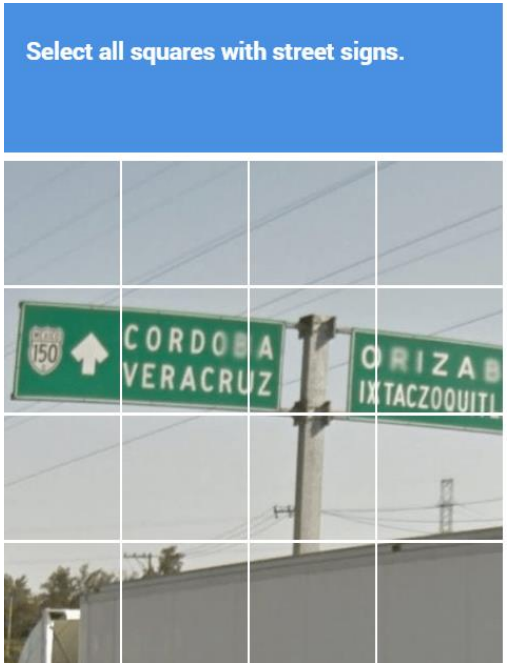
edoardo.pasolli@unina.it

@epasolli

DIPARTIMENTO DI AGRARIA

Human-in-the-loop, qualche esempio





edoardo.pasolli@unina.it
@epasolli

DIPARTIMENTO DI AGRARIA

Active learning & Mappe di copertura del suolo

Remote Sensing of Environment

Using active learning to adapt remote sensing image classifiers

D. Tuia [a,*], E. Pasolli [b], W.J. Emery [c]

IEEE GEOSCIENCE AND REMOTE SENSING LETTERS, VOL. 8, NO. 3, MAY 2011

Support Vector Machine Active Learning Through Significance Space Construction

Edoardo Pasolli, *Student Member, IEEE*, Farid Melgani, *Senior Member, IEEE*, and Yakoub Bazi, *Senior Member, IEEE*

IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING, VOL. 50, NO. 10, OCTOBER 2012

Active Learning Methods for Biophysical Parameter Estimation

Edoardo Pasolli, *Student Member, IEEE*, Farid Melgani, *Senior Member, IEEE*, Naif Alajlan, *Member, IEEE*, and Yakoub Bazi, *Senior Member, IEEE*

IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING, VOL. 51, NO. 6, JUNE 2013

Optical Image Classification: A Ground-Truth Design Framework

Edoardo Pasolli, *Member, IEEE*, Farid Melgani, *Senior Member, IEEE*, Naif Alajlan, *Member, IEEE*, and Nicola Conci, *Member, IEEE*

IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING, VOL. 52, NO. 4, APRIL 2014

SVM Active Learning Approach for Image Classification Using Spatial Information

Edoardo Pasolli, *Member, IEEE*, Farid Melgani, *Senior Member, IEEE*, Devis Tuia, *Member, IEEE*, Fabio Pacifici, *Senior Member, IEEE*, and William J. Emery, *Fellow, IEEE*

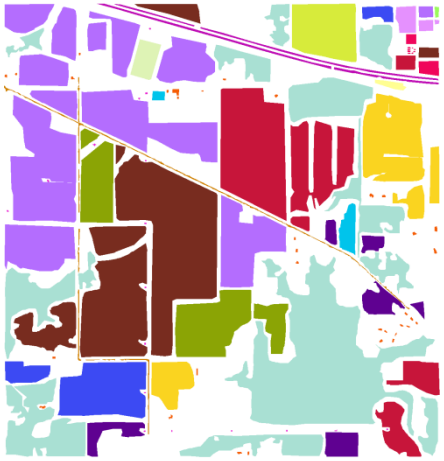
Active learning & Mappe di copertura del suolo

IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING, VOL. 54, NO. 4, APRIL 2016

Active-Metric Learning for Classification of Remotely Sensed Hyperspectral Images

Edoardo Pasolli, *Member, IEEE*, Hsiuhan Lexie Yang, *Member, IEEE*, and Melba M. Crawford, *Fellow, IEEE*

Dato telerilevato



Verità a terra

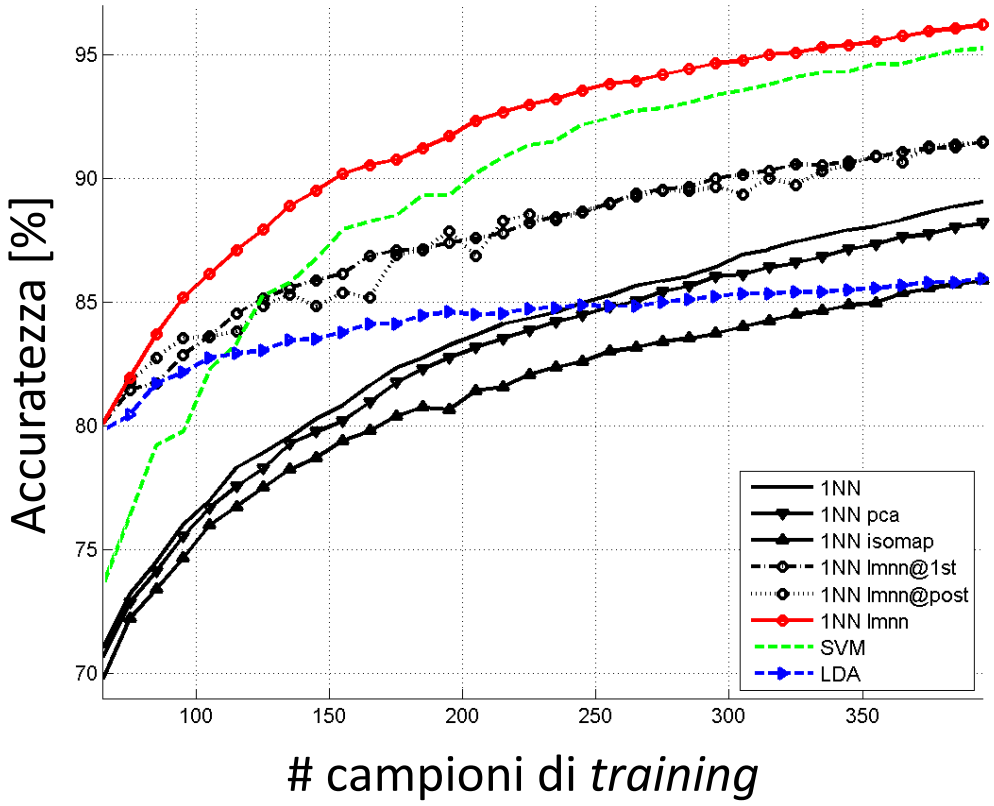
Dataset: New Indian Pine
Data di acquisione: Maggio 2010
Sensore: SpecTIR
Risoluzione spaziale: 2 m
Risoluzione spettrale: 5 nm
 # campioni etichettati: 1,094,132
 # classi tematiche: 19

edoardo.pasolli@unina.it
@epasolli

DIPARTIMENTO DI AGRARIA

Active learning & Mappe di copertura del suolo

Esempio di risultati di classificazione



Sviluppi futuri: ***Active + deep + transfer learning***

In collaborazione con prof. Antonia Maria Tulino @ DIETI

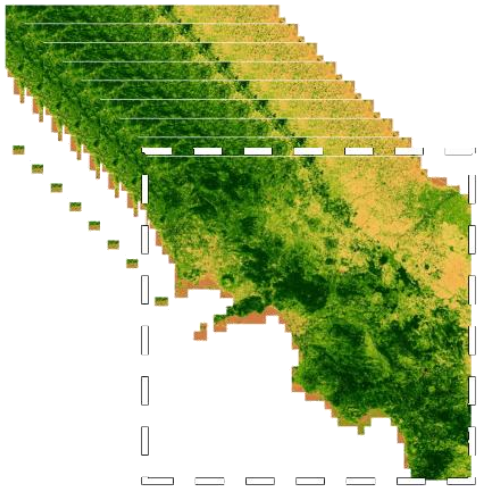
edoardo.pasolli@unina.it

@epasolli

DIPARTIMENTO DI AGRARIA

Identificazione di aree irrigate (in corso)

Con prof. Guido D'Urso



Dato multi temporale Landsat-8 & Sentinel-2 (2018)

Classe 1:
Bare soil

Classe 3:
Tree crop

Classe 2:
Herbaceous crop

Verità a terra
2.992 campioni

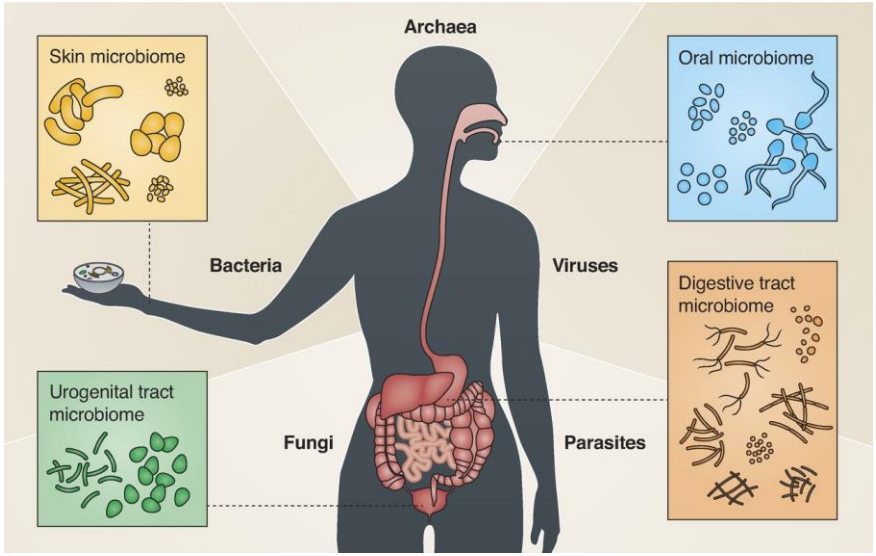
edoardo.pasolli@unina.it

@epasolli

DIPARTIMENTO DI AGRARIA

Introduzione al microbioma

"The ecological community of commensal, symbiotic, and pathogenic microorganisms living in the human body"



Structure, function and diversity of the healthy human microbiome nature

Role of the Microbiota in Immunity and Inflammation Cell

Potential of fecal microbiota for early-stage detection of colorectal cancer molecular systems biology

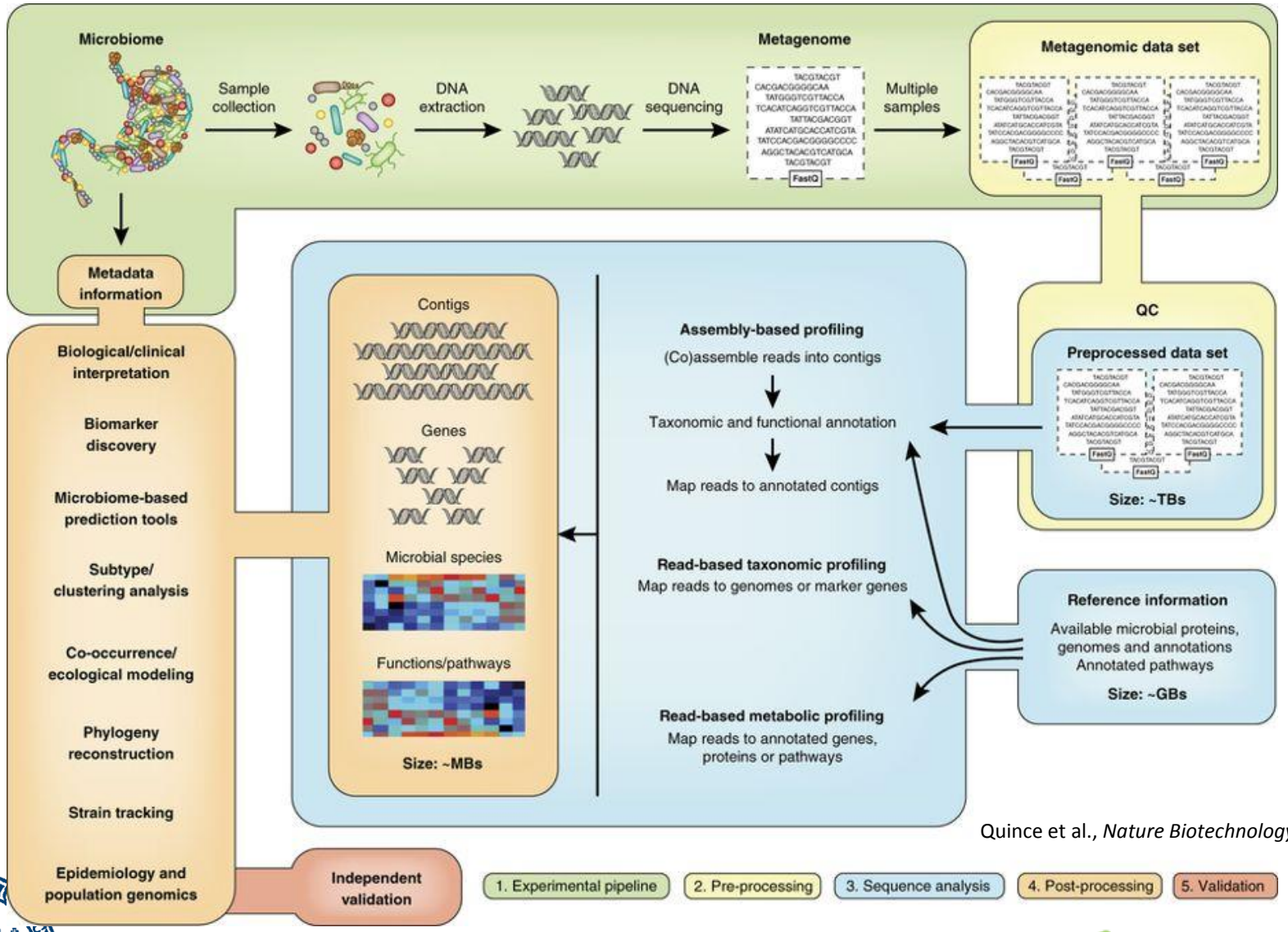
nature
Richness of human gut microbiome correlates with metabolic markers

A metagenome-wide association study of gut microbiota in type 2 diabetes nature

edoardo.pasolli@unina.it @epasolli

DIPARTIMENTO DI AGRARIA

Microbioma & Metagenomica



Quince et al., *Nature Biotechnology*, 2017

edoardo.pasolli@unina.it

@epasolli

Microbioma & Metagenomica

Da una prospettiva di *Data science*

Campioni biologici

Acquisizione dei dati

Analisi primaria
Generazione profili
Ricostruzione genomi

Analisi secondaria
Interpretazione biologica/clinica

Validazione

edoardo.pasolli@unina.it

@epasolli

DIPARTIMENTO DI AGRARIA

Microbioma & Metagenomica

Le mie linee di ricerca

Machine learning **per il microbioma**

Tools **per l'analisi a livello di ceppo**

edoardo.pasolli@unina.it

@epasolli

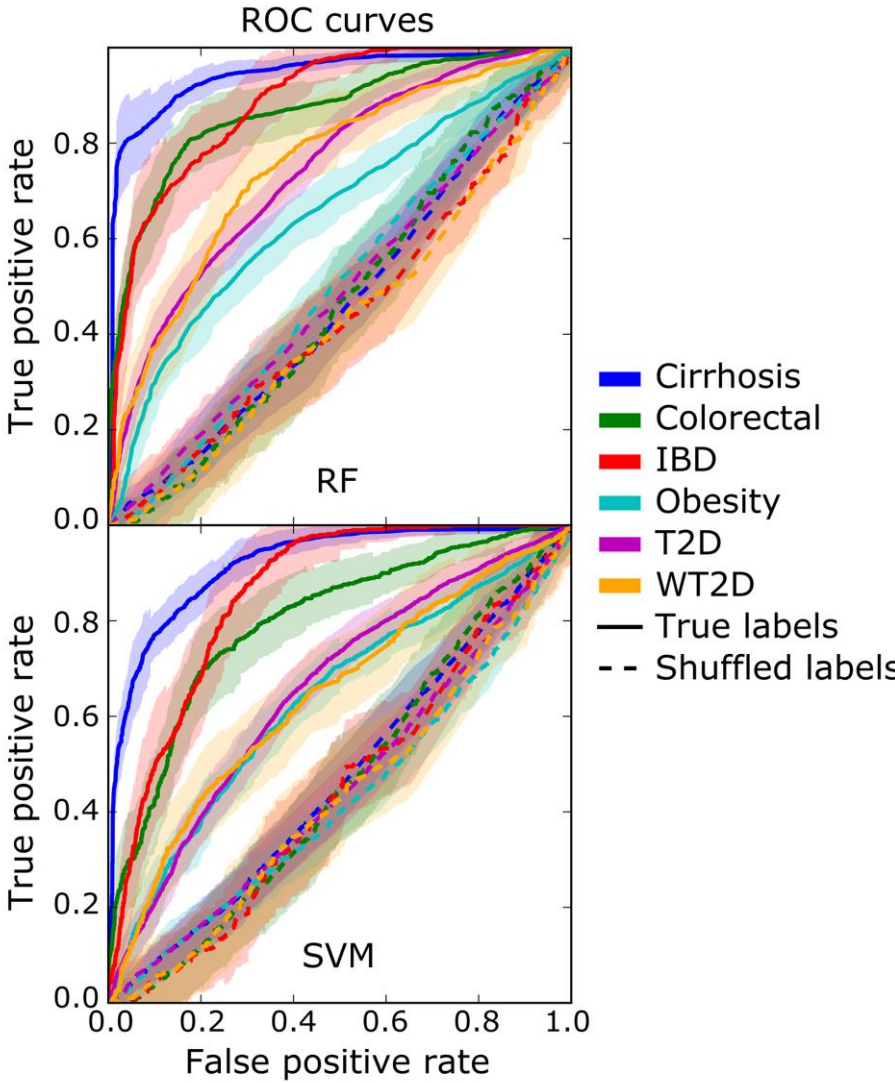
DIPARTIMENTO DI AGRARIA

Microbioma & Modelli predittivi

PLOS COMPUTATIONAL BIOLOGY

Machine Learning Meta-analysis of Large Metagenomic Datasets: Tools and Biological Insights

Edoardo Pasolli[1], Duy Tin Truong[1], Faizan Malik[2], Levi Waldron[2], Nicola Segata[1]*

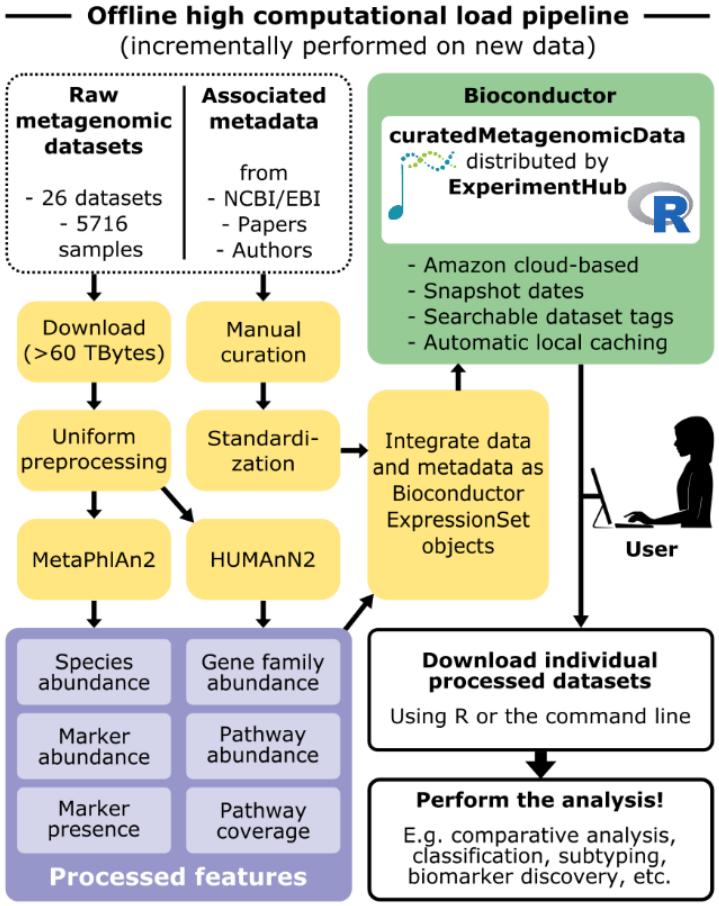


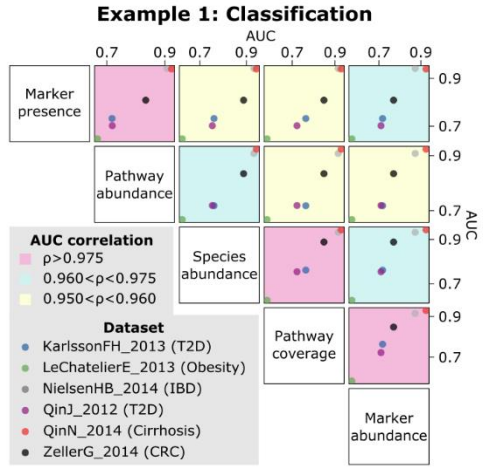
edoardo.pasolli@unina.it

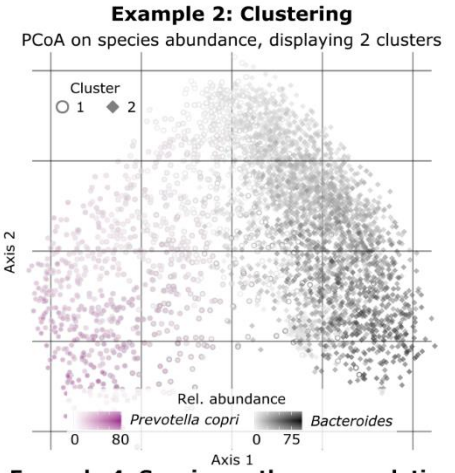
@epasolli

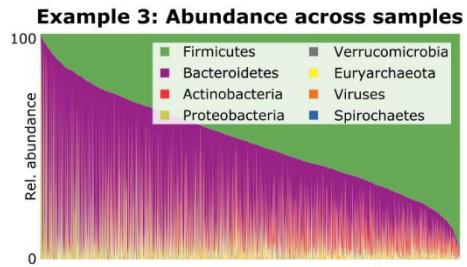
DIPARTIMENTO DI AGRARIA

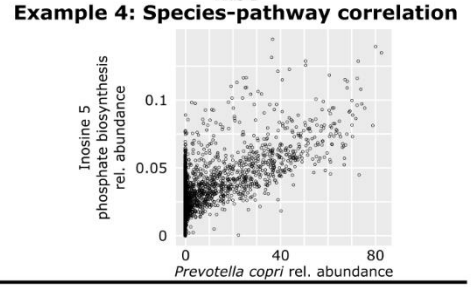
curatedMetagenomicData











User analyses

Bioconductor
OPEN SOURCE SOFTWARE FOR BIOINFORMATICS

edoardo.pasolli@unina.it
@epasolli

DIPARTIMENTO DI AGRARIA

Analisi a larga scala

Thomas*, Manghi*, et al.

nature medicine

Metagenomic analysis of colorectal cancer datasets identifies cross-cohort microbial diagnostic signatures and a link with choline degradation

Progetto di dottorato industriale

Modelli predittivi basati su dieta e microbioma per lo studio di malattie del cavo orale

edoardo.pasolli@unina.it

@epasolli

DIPARTIMENTO DI AGRARIA

Tools per l'analisi a livello di ceppo

Assunzione: Ciascun **ceppo** ha una combinazione unica di...

… geni dal pangenoma della specie
PanPhlAn

Strain-level microbial epidemiology and population genomics from shotgun metagenomics

Matthias Scholz[1,4], Doyle V Ward[2,4], Edoardo Pasolli[1,4], Thomas Tolio[1], Moreno Zolfo[1], Francesco Asnicar[1], Duy Tin Truong[1], Adrian Tett[1], Ardythe L Morrow[3] & Nicola Segata[1]

nature methods

… SNVs nel *core genome*
StrainPhlAn

Microbial strain-level population structure and genetic diversity from metagenomes

Duy Tin Truong,[1] Adrian Tett,[1] Edoardo Pasolli,[1] Curtis Huttenhower,[2,3] and Nicola Segata[1]

Genome Research

edoardo.pasolli@unina.it
@epasolli

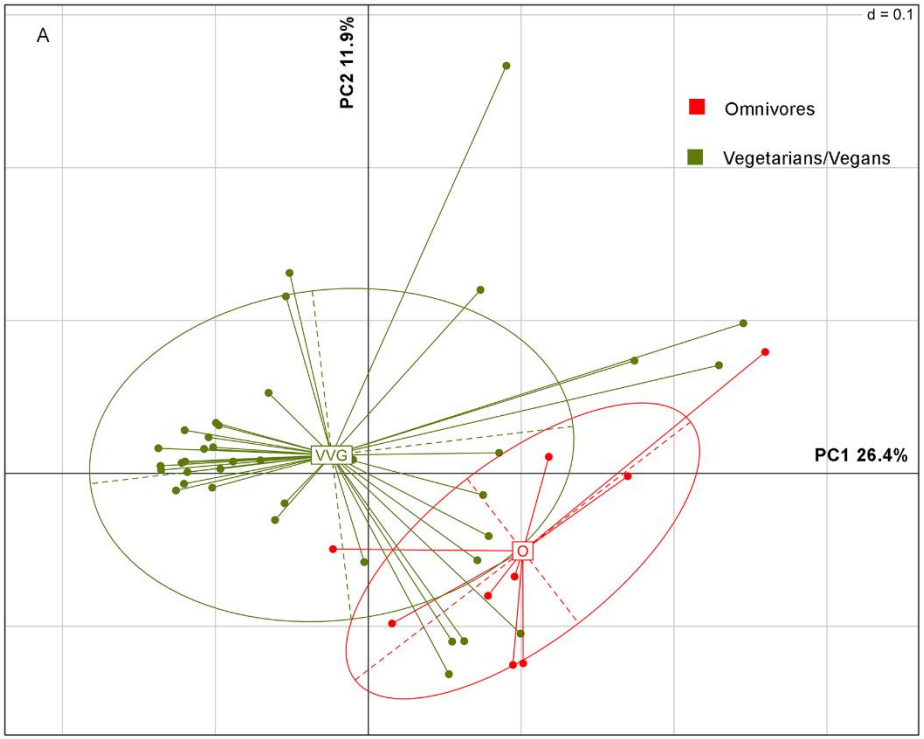
Dipartimento di Agraria

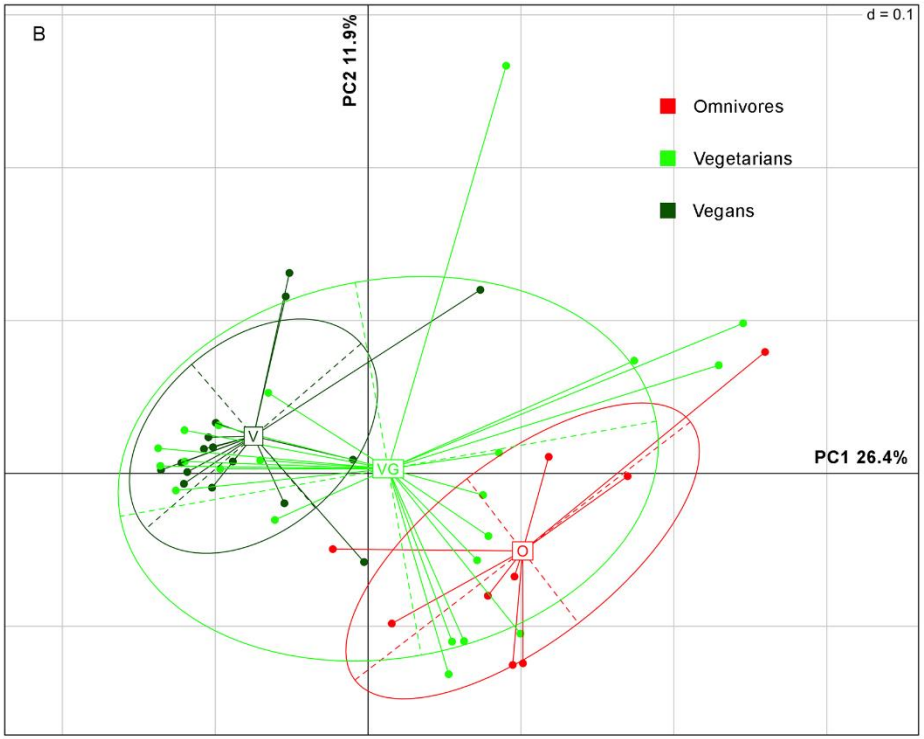
Prevotella copri & Dieta

Cell Host & Microbe

Distinct Genetic and Functional Traits of Human Intestinal *Prevotella copri* Strains Are Associated with Different Habitual Diets

Francesca De Filippis,[1,2] Edoardo Pasolli,[1,3] Adrian Tett,[3] Sonia Tarallo,[4] Alessio Naccarati,[4] Maria De Angelis,[5] Erasmo Neviani,[6] Luca Cocolin,[7] Marco Gobbetti,[8] Nicola Segata,[3] and Danilo Ercolini[1,2,9,*]





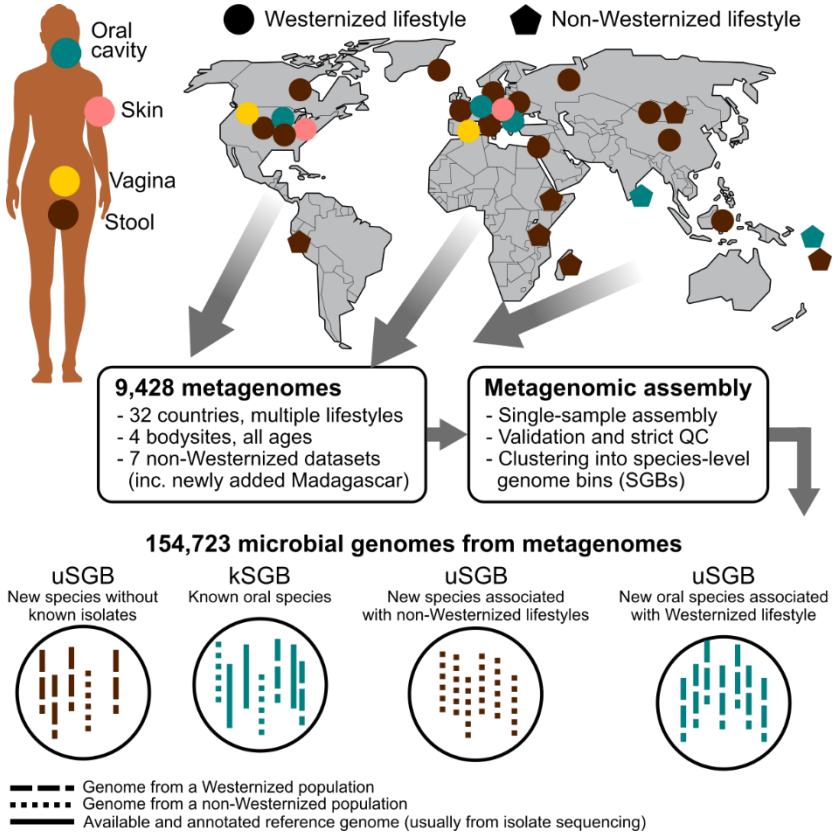
edoardo.pasolli@unina.it

@epasolli

Il più ricco catalogo del microbioma umano

Extensive Unexplored Human Microbiome Diversity Revealed by Over 150,000 Genomes from Metagenomes Spanning Age, Geography, and Lifestyle

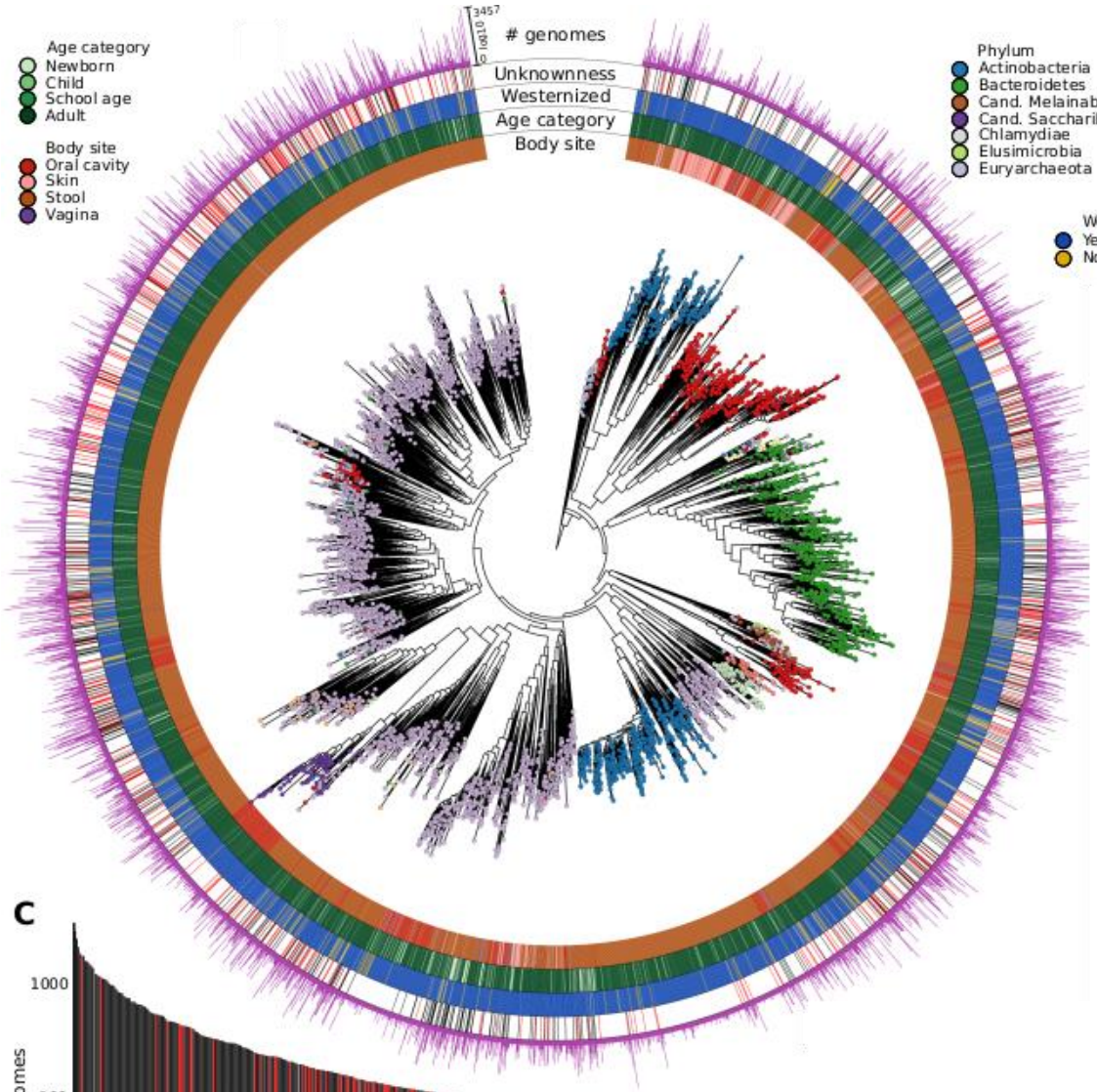
Edoardo Pasolli,[1] Francesco Asnicar,[1,8] Serena Manara,[1,8] Moreno Zolfo,[1,8] Nicolai Karcher,[1] Federica Armanini,[1] Francesco Beghini,[1] Paolo Manghi,[1] Adrian Tett,[1] Paolo Ghensi,[1] Maria Carmen Collado,[2] Benjamin L. Rice,[3] Casey DuLong,[4] Xochitl C. Morgan,[5] Christopher D. Golden,[4] Christopher Quince,[6] Curtis Huttenhower,[4,7] and Nicola Segata[1,9,*]



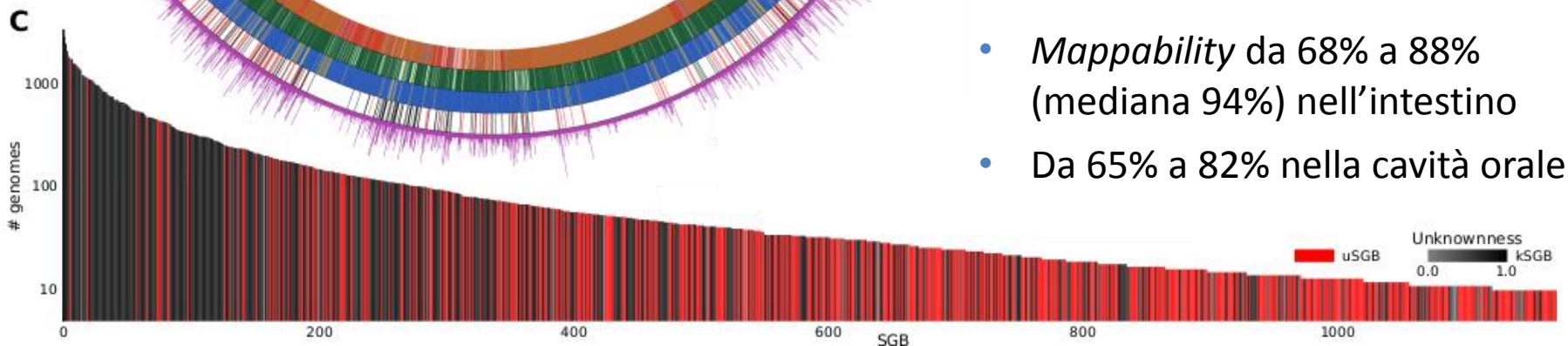
edoardo.pasolli@unina.it

@epasolli

Il più ricco catalogo del microbioma umano



- 154.723 genomi ricostruiti da 9.428 metagenomi
- Raddoppiato il numero di genomi disponibili
- Identificate 4,930 specie (77% nuove)
- Nuove specie prevalenti in popolazioni non-occidentalizzate
- *Mappability* da 68% a 88% (mediana 94%) nell'intestino
- Da 65% a 82% nella cavità orale



Lactic acid bacteria, bridging food and gut health

Batteri lattici sono altamente prevalenti negli alimenti fermentati. Ma quanto lo sono nell'intestino?

- Analisi condotta su 9.000 metagenomi umani
- *S. thermophilus* e *Lactobacillus* in popolazioni occidenzalizzate
- *Leuconostoc* e *Weissella* in popolazioni non-occidentalizzate

Batteri lattici negli alimenti e nell'intestino, come differiscono funzionalmente?

- Ricostruiti 650 nuovi genomi da 250 metagenomi da cibo
- Integrazione con il catalogo di genomi del microbioma umano
- Analisi funzionale in corso

master

In collaborazione con prof. Danilo Ercolini e dott. Francesca De Filippis

edoardo.pasolli@unina.it

@epasolli

DIPARTIMENTO DI AGRARIA

Prossimo appuntamento

18 Settembre 2019

Il valore della diversità

Francesco Caracciolo di Torchiarolo